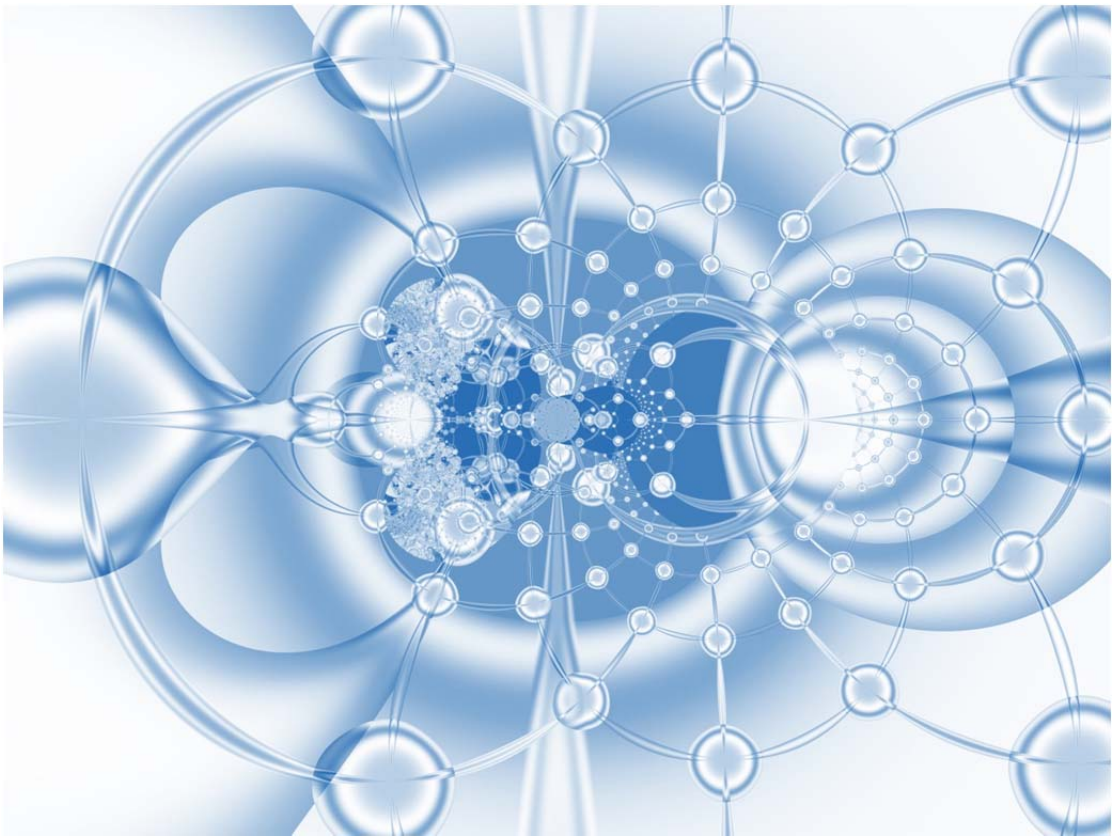
Inter-X: Resilience of the Internet Interconnection Ecosystem

Full Report – April 2011



enisa
European Network and Information Security Agency

About ENISA

The European Network and Information Security Agency (ENISA) is an EU agency created to advance the functioning of the internal market. ENISA is a centre of expertise for the European Member States and European institutions in network and information security, giving advice and recommendations and acting as a switchboard for information on good practices. Moreover, the agency facilitates contacts between European institutions, the Member States, and private business and industry actors. Internet: *http://www.enisa.europa.eu/*

Acknowledgments:

While compiling this report, we talked extensively over a period of many months to a large number of technical and managerial staff at communications service providers, vendors, and service users. Many of our sources requested that we not acknowledge their contribution. Nonetheless we thank them all here. ENISA would like to express its gratitude to the stakeholders that provided input to the survey.

Editor: Panagiotis Trimintzios, ENISA

Authors:

Chris Hall, Highwayman Associates

Richard Clayton, Cambridge University

Ross Anderson, Cambridge University

Evangelos Ouzounis, ENISA

Contact

For more information about this study, please contact:

Dr. Panagiotis Trimintzios

panagiotis.trimintzios@enisa.europa.eu

Internet: http://www.enisa.europa.eu/act/res

18th April 2011 (b)

Table of Contents

Executive Summary and Introduction

Executive Summary

The Internet has so far been extremely resilient. Even major disasters, such as 9/11 and Hurricane Katrina, have had only a local impact. Technical failures have lasted only a few hours, and congestion has had a sustained effect only where the infrastructure is inadequate. The low cost and general reliability of communications over the Internet have led more and more systems to depend on it; we are now at the point where a systemic failure would not just disrupt email and the web, but cause significant problems for other utilities, transport, finance, healthcare and the economy generally. So the continued resilience of the Internet is critical to the functioning of modern societies, and hence it is right and proper to examine whether the mechanisms that have such an excellent track record in providing a resilient Internet are likely to continue to be as effective in the future.

The focus of this report is the 'Internet interconnection ecosystem'. This holds together all the networks that make up the Internet. The ecosystem is complex and has many interdependent layers. This system of connections between networks occupies a space between and beyond those networks and its operation is governed by their collective self-interest – the Internet has no central Network Operation Centre, staffed with technicians who can leap into action when trouble occurs. The open and decentralised organisation that is the very essence of the ecosystem is essential to the success and resilience of the Internet. Yet there are a number of concerns.

First, the Internet is vulnerable to various kinds of common mode technical failures where systems are disrupted in many places simultaneously; service could be substantially disrupted by failures of other utilities, particularly the electricity supply; a flu pandemic could cause the people on whose work it depends to stay at home, just as demand for home working by others was peaking; and finally, because of its open nature, the Internet is at risk of intentionally disruptive attacks.

Second, there are concerns about sustainability of the current business models. Internet service is cheap, and becoming rapidly cheaper, because the costs of service provision are mostly fixed costs; the marginal costs are low, so competition forces prices ever downwards. Some of the largest operators – the 'Tier 1' transit providers – are losing substantial amounts of money, and it is not clear how future capital investment will be financed. There is a risk that consolidation might reduce the current twenty-odd providers to a handful, at which point they would start to acquire pricing power and the regulation of transit service provision might become necessary as in other concentrated industries.

Third, dependability and economics interact in potentially pernicious ways. Most of the things that service providers can do to make the Internet more resilient, from having excess capacity to route filtering, benefit other providers much more than the firm that pays for them, leading to a potential 'tragedy of the commons'. Similarly, security mechanisms that would help reduce the likelihood and the impact of malice, error and mischance are not implemented because no-one has found a way to roll them out that gives sufficiently incremental and sufficiently local benefit.

Fourth, there is remarkably little reliable information about the size and shape of the Internet infrastructure or its daily operation. This hinders any attempt to assess its resilience in general and the analysis of the true impact of incidents in particular. The opacity also hinders research and

development of improved protocols, systems and practices by making it hard to know what the issues really are and harder yet to test proposed solutions.

So there may be significant troubles ahead which could present a real threat to economic and social welfare and lead to pressure for regulators to act. Yet despite the origin of the Internet in DARPA-funded research, the more recent history of government interaction with the Internet has been unhappy. Various governments have made ham-fisted attempts to impose censorship or surveillance, while others have defended local telecommunications monopolies or have propped up other industries that were disrupted by the Internet. As a result, Internet service providers, whose good will is essential for effective regulation, have little confidence in the likely effectiveness of state action, and many would expect it to make things worse.

Any policy should therefore proceed with caution. At this stage, there are four types of activity that can be useful at the European (and indeed the global) level.

The first is to understand failures better, so that all may learn the lessons. This means consistent, thorough, investigation of major outages and the publication of the findings. It also means understanding the nature of success better, by supporting long term measurement of network performance, and by sustaining research in network performance.

The second is to fund key research in topics such as inter-domain routing – with an emphasis not just on the design of security mechanisms, but also on traffic engineering, traffic redirection and prioritisation, especially during a crisis, and developing an understanding of how solutions are to be deployed in the real world.

The third is to promote good practices. Diverse service provision can be encouraged by explicit terms in public sector contracts, and by auditing practices that draw attention to reliance on systems that lack diversity. There is also a useful role in promoting the independent testing of equipment and protocols.

The fourth is public engagement. Greater transparency may help Internet users to be more discerning customers, creating incentives for improvement, and the public should be engaged in discussions on potentially controversial issues such as traffic prioritisation in an emergency. And finally, Private Public Partnerships (PPPs) of relevant stakeholders, operators, vendors, public actors etc is important for self-regulation. In this way even if regulation of the Internet interconnection system is ever needed after many years, policy makers will be able to make informed decisions leading to effective policies.

The objective of these activities should be to ensure that when global problems do arise, the decision and policy makers have a clear understanding of the problems and of the options for action.

There are local regulatory actions that Europe can encourage where needed. Poor telecommunications regulation can lead to the consolidation of local service provision so that cities have fewer independent infrastructures; and in countries that are recipients of EU aid, telecommunications monopolies often deepen the digital divide.

The aim of all these activities should be to ensure that the Internet is ubiquitous and resilient, with service provided by multiple independent competing firms who have the incentives to provide a prudent level of capacity not just for fair weather, but for when the storms arrive.

Introduction to the Report

This study looks at the resilience of the Internet interconnection ecosystem. The Internet is a network of networks, and the interconnection ecosystem is the collection of layered systems that holds it together. The interconnection ecosystem is the core of the Internet, providing the basic function of reaching anywhere from everywhere.

The Executive Summary above provides an abstract of the report's subject and broad recommendations. The rest of the report is in four parts:

Part I Summary and Recommendations

This contains a more extended examination of the subject and a discussion of our recommendations in detail, followed by the recommendations themselves.

This part of the report is based on the parts which follow.

Part II State of the Art Review

This includes a detailed description of the Internet's routing mechanisms and analysis of their robustness at the technical, economic and policy levels.

The material in this part supports the analysis presented in Part I, and sets out to explain how and why the issues and challenges the report identifies come about.

Part III Report on the Consultation

As part of the study a broad range of stakeholders were consulted. This part reports on the consultation and summarises the results.

Part IV Bibliography and Appendices

There is an extensive bibliography and summaries of the financial statements of some of the major transit providers.

There is a Summary Report for the study. That report is Part I of this Full Report, along with the Executive Summary.

This revised version of the report replaces the version published in December 2010.

PART I – Summary and Recommendations

Introduction to the Summary and Recommendations

This report looks at the resilience of the Internet interconnection ecosystem – how it may be assessed, and maintained or improved. In the State of the Art Review (Part II) and in the Consultation (Part III) a number of issues and challenges to the system's resilience and its assessment are identified.

Section 1 of this part of the report is a summary of the issues and challenges. It is intended to be read as an introduction to the recommendations, to give the background and the rationale for them. It serves also as an introduction to the rest of the report. Expert readers may wish to read the Report on the Consultation (Part III) before proceeding any further with this part.

Our recommendations are given in Section 2.

Note: this part of the Full Report is reproduced in the Summary Report.

1 Summary

The Internet has been pretty reliable so far, having recovered rapidly from most known incidents. The effects of natural disasters such as Hurricane Katrina, terrorist attacks such as 9/11 and assorted technical failures have all been limited in time and space. However it does appear likely that the Internet could suffer systemic failure, leading perhaps to local failures and system-wide congestion, in some circumstances including:

- A regional failure of the physical infrastructure on which it depends (such as the bulk power transmission system) or the human infrastructure needed to maintain it (for example if pandemic flu causes millions of people to stay at home out of fear of infection).
- Cascading technical failures, of which some of the more likely near-term scenarios relate to the imminent changeover from IPv4 to IPv6; common-mode failures involving updates to popular makes of router (or PC) may also fall under this heading.
- A coordinated attack in which a capable opponent disrupts the BGP fabric by broadcasting thousands of bogus routes, either via a large AS or from a large number of compromised routers.

There is evidence that implementations of the Border Gateway Protocol (BGP) are surprisingly fragile. There is evidence that some concentrations of infrastructure are vulnerable and significant disruption can be caused by localised failure. There is evidence that the health of the interconnection system as a whole is not high among the concerns of the networks that make up that system – by and large each network strives to provide a service which is reliable, most of the time, at minimum achievable cost. The economics do not favour high dependability as there is no incentive for anyone to provide the extra capacity that would be needed to deal with large-scale failures.

To date, we have been far from an equilibrium: the rapid growth in capacity has masked a multitude of sins and errors. However, as the Internet matures, as more and more of the world's optical fibre is lit, and as companies jostle for advantage, the dynamics may change.

There may well not be any immediate cause for concern about the resilience of the Internet interconnection ecosystem, but there is cause for concern about the lack of good information about how it works and how well it might work if something went very badly wrong.

This section proceeds as follows:

- in Section 1.1 the challenges posed by the sheer scale and complexity of the Internet interconnection system are discussed.
- the nature of resilience and the difficulty of assessing it are discussed in Section 1.2.
- Section 1.3 discusses the information that we do not have, and how that limits our ability to address the issue of resilience, among other things.
- resilience and efficiency are antipathetic, which raises the challenges given in Section 1.4.
- the problems posed by the reliability of equipment, and the possibility for systemic failure are covered in Section 1.5.
- Section 1.6 examines the value of Service Level Agreements in the context of the interconnection system.

- all parts of the Internet must be able to reach all other parts, so 'reachability' is a key objective. However, being able to reach a destination does not guarantee that traffic will flow to and from there effectively and that expected levels of performance will be met. Section 1.7 discusses the challenges, with particular reference to the behaviour of the system if some event has disabled parts of it.
- every year the price of transit goes down, and every year people feel it must level off. The reason to believe that the price will tend to zero, and the challenges that poses are discussed in Section 1.8.
- the rise of the Content Delivery Networks (CDNs) and the effect on the interconnection system is discussed in Section 1.9.
- Section 1.10 tackles the insecurity of BGP.
- in Section 1.11 the value of disaster recovery exercises ("war games") is examined.
- a number of issues are related; tackling them would benefit everybody, but addressing them also costs each network more than they gain individually. This is discussed in Section 1.12.
- the contentious subject of regulation is raised in Section 1.13.

In the following, references of the form **[C:xx]** refer to general points made in the consultation, while those of the form **[Q:xx]** refer to quotations from the consultation which made a particular, or a particularly apposite, point. The report on the consultation is in Part III, below.

1.1 Scale and Complexity

The Internet is very big and very complicated **[C:1]**.

The interconnection system we call the Internet comprises some 37,000 'Autonomous Systems' or ASes (ISPs or similar entities) and 350,000 blocks of addresses (addressable groups of machines), spread around the world – as of March 2011 (see Section 3).

This enormous scale means that it is hard to conceive of an external event which would affect more than a relatively small fraction of the system – as far as the Internet is concerned, a large earthquake or major hurricane is, essentially, a little local difficulty. However, the failure of a small fraction of the system may still have a significant impact on a great many people. When considering the resilience of this system it is necessary to consider not only the global issues, but a large number of separate, but interconnected, local issues.

The complexity of the system is partly related to its sheer scale, and the number of interconnections between ASes. This is compounded by a number of factors.

- Modelling the interconnection system is hard because we only have partial views of it and because it has a number of layers, each with its own properties and interacting with other layers. For example, the connections between ASes use many different physical networks, often provided by third parties, which are themselves large and complicated. Resilience depends on the diversity of interconnections, which in turn depends on physical diversity – which can be an illusion, and is often unknown **[C:7]**.

  While it is possible to discover part of the 'AS-level topology' of the Internet (which ASes are interconnected), from a resilience perspective, it would be more valuable to know the 'router-

level topology', (the number, location, capacity, traffic levels etc. of the actual connections between ASes) **[C:2]**. If we want to estimate how traffic might move around when connections fail, we also need to know about the 'routing layer' (what routes the routers have learned from each other) so we can estimate what routes would be lost when given connections failed, and what routes would be used instead **[C:3]**. That also touches on 'routing policy' (the way each AS decides which routes it will prefer) and the 'traffic layer' [where end-user traffic is going to and from]. This is perhaps the most important layer, but very little is known about it on a global scale.

- The interconnection system depends on other complex and interdependent systems. The routers, the links between them, the sites they are housed in, and all the other infrastructure that the interconnection system depends on, themselves depend on other systems – notably electricity supply – and those systems depend in turn on the Internet. **[C:8]**, **[Q:3]** and **[Q:17]**.
- The interconnection ecosystem is self-organising and highly decentralised. The decision whether to interconnect is made independently by the ASes, driven by their need to be able to reach, and be reachable from, the entire Internet. The same holds at lower levels: the administrators of an AS configure their routers to implement their routing policy, then the routers select and use routes. But different routers in the same AS may select different routes for a given destination, so even the administrators may not know, a priori, what path traffic will take.
- The interconnection ecosystem is dynamic and constantly changing. Its shape changes all the time, as new connections are made, or existing connections fail or are removed. At the corporate level, transit providers come and go, organisations merge, and so on. At the industry level, the recent rise of the content delivery networks (CDNs) changed the pattern of interconnections.
- The patterns of use are also constantly evolving. The rise of the CDNs also changed the distribution of traffic; and while peer-to-peer (P2P) traffic became a large proportion of total traffic in the early-to-mid 2000s, now video traffic of various kinds is coming to dominate both in terms of volume and in terms of growth.
- The Internet is continuing to grow. In fact, just about everything about it continues to grow: the number of ASes, the number of routes, the number of interconnections, the volume of traffic, etc.

The scale and complexity of the system make it hard to grasp. Resilience is itself a slippery concept, so the resilience of the interconnection system is non-trivial to define – let alone measure!

This study attempts to provide some insight by describing the workings of the system and what we know about its resilience.

1.2 The Nature of Resilience

There is a vast literature on reliability where engineers study the failure rates of components, the prevalence of bugs in software, and the effects of wear, maintenance etc.; the aim being to design machines or systems with a known rate of failure in predictable operating conditions [**1**]. Robustness relates to designing systems to withstand overloads, environmental stresses and other insults, for example by specifying equipment to be significantly stronger than is needed for normal operation. In traditional engineering, resilience was the ability of a material to absorb energy under

stress and release it later. In modern systems thinking, it means the opposite of 'brittleness' and refers to the ability of a system or organisation to adapt and recover from a serious failure, or more generally to its ability to survive in the face of threats, including the prevention or mitigation of unsafe, hazardous or detrimental conditions that threaten its existence [**2**]. In the longer term, it can also mean evolvability: the ability of a system to adapt gradually as its environment changes – an idea borrowed from systems biology [**3**] [**4**].

Resilience of a system is defined as *the ability to provide and maintain an acceptable level of service in the face of various faults and challenges to normal operation*[1]. That is the ability to adapt itself to recover from a serious failure, or more generally to its ability to survive in the face of threats. A given event may have some impact on a system and hence some immediate impact on the service it offers. The system will then recover, service levels will improve and at some time full service and the system will be restored.

Resilience therefore refers both to failure recovery at the micro level, as when the Internet recovers from the failure of a router so quickly that users perceive a connection failure of perhaps a few seconds (if they notice anything at all); through coping with a mid-size incident, as when ISPs provided extra routes in the hours immediately after the 9/11 terrorist attacks by running fibres across collocation centres; to disaster recovery at the strategic level, where we might plan for the next San Francisco earthquake or for a malware compromise of thousands of routers. In each case the desired outcome is that the system should continue to provide service in the event of some part of it failing, with service degrading gracefully if the failure is large.

There are thus two edge cases of resilience:

1. the ability of the system to cope with small local events – such as equipment failures – and reconfigure itself essentially automatically and over a time scale of seconds to minutes. This enables the Internet to cope with day-to-day events with little or no effect on service – it is reliable. This is what most network engineers think of as resilience.
2. the ability of a system to cope with and recover from a major event, such as a large natural disaster or a capable attack, on a time scale of hours to days or even longer. This type of resilience includes, first, the ability of the system to continue to offer some service in the immediate aftermath, and second, the ability to repair and rebuild thereafter. The key words here are 'adapt' and 'recover'. This 'disaster recovery' is what civil authorities tend to think of as resilience.

This study is interested in the resilience of the ecosystem in the face of events which have medium to high impact and which have a correspondingly medium to low probability. It is thus biased toward the second of these cases.

Robustness is an important aspect of resilience. A robust system will have the ability to resist assaults and insults, so that whatever some event is throwing at it, it will be unaffected, and no resilient response is required. While resilience is to do with coping with the impact of events,

[1] *following: James P.G. Sterbenz, David Hutchison, Egemen K. Çetinkaya, Abdul Jabbar, Justin P. Rohrer, Marcus Schöller and Paul Smith: "Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines", Computer Networks, Volume 54, Issue 8, 1 June 2010, Pages 1245-1265, Resilient and Survivable networks.*

robustness is to do with reducing the impact in the first place. The two overlap, and from the users' perspective these are fine distinctions; what the user wants is for the system to be predictably dependable.

Resilience is context-specific. Robustness can be sensibly defined only in respect of specified attacks or failures, and in the same way resilience also makes sense only in the context of recovery from specified events, or in the face of a set of possible challenges of known probability. We call bad events of known probability 'risk', but there is a separate problem of 'uncertainty' where we do not know enough about possible future bad events to assign them a probability at all. In the face of uncertainty, it is difficult to assess a combination of intermediate levels of service and recovery/restoration times, especially when what is acceptable may vary depending on the nature and scale of the event. **[C:5]**

Moreover, no good metrics are available to actually assess the performance of the Internet or its interconnection system. This makes it harder still to specify acceptable levels of service. For the Internet the problem is compounded by its scale and complexity (see above) and by lack of information (see below), which make it hard to construct a model which might be used to attach numbers to resilience. It is even hard to assess what impact a given single event might have – an earthquake in San Francisco of a given severity may have a predictable impact on the physical infrastructure, but that needs to be translated into its effect on each network, and hence the effect on the interconnection system.

Given these difficulties (and there are many more), service providers commonly fall back on measures that improve resilience in general terms, in the hope that this will improve their response to future challenges. This qualitative approach runs into difficulty when the cost of an improvement must be justified on much more restricted criteria. For the Internet as a whole, the cost justification of investment in resilience is an even harder case to make.

1.3 The Lack of Information

Each of the ASes that make up the Internet each has a Network Operation Centre (NOC), charged with monitoring the health of the AS's network and instigating action when problems occur. There is no NOC for the Internet.

In fact it is worse than that. ASes understand their own networks but know little about anyone else's. At every level of the interconnection system, there is little global information available, and what is available is incomplete and of unknown accuracy. In particular:

- there is no map of physical connections – their location, capacity, etc.;
- there is no map of traffic and traffic volume;
- there is no map of the interconnections between ASes – what routes they offer each other.

The Internet interconnection system is, essentially, opaque. This opacity hampers the research and development communities in their attempts to understand the workings of the Internet, and to develop and test improvements; it makes the study and modelling of complex emergent properties such as resilience harder still. **[C:2]**, **[Q:1]** and **[Q:2]**.

The lack of information has a number of causes:

- **Complexity and scale.** To map the networks of fibre around the world might be a tractable problem. Over those physical fibres run many different logical connections, each of which will carry network traffic for numerous providers, which in turn support yet more providers' networks and circuits – rapidly multiplying up the combinations and permutations of overlapping use of the underlying fibre. Furthermore, not all those things are fixed – providers reroute existing networks and circuits as they extend or adapt their networks. To keep track, meticulous record keeping is required, but even within a single AS it is not always achieved. At a global level, measuring traffic volumes would be an immense undertaking, given the sheer number of connections between networks.
- **The information hiding properties of the routing system.** When trying to map connections by probing the system from the outside, each probe will reveal something about the path between two points in the Internet at the time of the probe. But the probe reveals little about what other paths may exist at other times, or what path might be taken if any part of the usual path is not working, or what the performance of those other paths might be.
- **Security concerns.** Mapping the physical layer is thought to be an invitation to people with bad intentions to improve their target selection so those maps that do exist are seldom shared.
- **The cost of storing and processing the data.** If there was complete information, there would be a very great deal of it, and more would be generated every minute. Storing it and processing it into a usable form would be a major engineering task.
- **Commercial sensitivity.** Information about whether, how and where networks connect to each other is deemed commercially sensitive by some. Information about traffic volumes is quite generally seen as commercially sensitive. Because of this, some advocate powerful incentives to disclose information, and possibly in anonymised and aggregated form. **[C:23]**
- **Critical information is not collected in the first place, or not kept up to date.** Information gathering and maintenance costs money, so there must be some real use for it before a network will bother to gather it or strive to keep it up to date. The Internet Routing Registries (IRRs) are potentially excellent resources, but are not necessarily up to date, complete or accurate, because the information seldom has operational significance (and may in any case be deemed commercially sensitive).
- **Lack of good metrics.** While there are some well-known metrics for the performance of connections between two points in a network, there are none for a network as a whole or, indeed, a network of networks. ENISA has already started working in this direction, looking at resilience metrics from a holistic point of view[2].

The poor state of information reflects not only the difficulty of finding or collecting data, but also the lack of good ways to process and use it even if one had it.

[2] http://www.enisa.europa.eu/act/res/other-areas/metrics

1.3.1 Incidents as a Source of Information

Small incidents occur every day, and larger ones every now and then. Given the lack of information about the interconnection system, the results of these natural experiments tell us much of what we know about its resilience. **[C:4]**. For example, we know the following.

- It is straightforward to divert traffic away from its proper destination by announcing invalid routes. The well-known incident in February 2008 in which YouTube stopped working for a few hours is one example; see Section 5.6.4. More publicity, and political concern, was raised by a 2010 incident in which China Telecom advertised a number of invalid routes, effectively hijacking 15% of Internet addresses for 18 minutes; see Section 5.6.9.
- Latent bugs in BGP implementations can disrupt the system. Most recently, in August 2010, an experiment that sent an unusual (but entirely legal) form of route announcement triggered a bug in some routers, causing their *neighbours* to terminate BGP sessions, and for many routes to be lost. The effects of this incident lasted less than two hours; see Section 5.6.5.
- In some parts of the world a small number of cable systems are critical. Undersea cables near Alexandria in Egypt were cut in December 2008. Interestingly, three cable systems were affected at the same time, and two of those systems had been affected similarly in January/February of that year. This seriously affected traffic for perhaps two weeks. See Section 5.6.6.
- The system is critically dependent on electrical power. A large power outage in Brazil in November 2009 caused significant disruption, though it lasted only four and a half hours; see Section 5.6.8. Interestingly, previous blackouts in Brazil had been attributed to 'hackers', suggesting that these incidents are examples of the risk of inter-dependent networks. This particular conspiracy theory has been refuted.
- The ecosystem can work well in a crisis. The analysis of the effect of the destruction at the World Trade Centre in New York on 11th September 2001 shows that the system worked well at the time, and in the days thereafter, even though large cables under the buildings were cut and other facilities were destroyed or damaged. Generally, Internet services performed better than the telephone system (fixed and mobile). See Section 5.6.10.

These sorts of incident are well known. However, hard information about the exact causes and effects is hard to come by – much is anecdotal and incomplete, while some is speculative or simply apocryphal. Valuable information is being lost. The report *"The Internet under Crisis Conditions: Learning from September 11"*, [**5**] is a model of clarity; but even there the authors warn:

> *"... While the committee is confident in its assessment that the events of September 11 had little effect on the Internet as a whole ..., the precision with which analysts can measure the impact of such events is limited by a lack of relevant data."*

1.4 Resilience and Efficiency

There are fundamental tensions between resilience and efficiency. **[Q:5]** Resilience requires spare capacity and duplication of resources, and systems which are loosely coupled (made up of largely independent sub-systems) are more resilient than tightly coupled systems whose components depend more on each other. But improving the efficiency of a system generally means eliminating excess capacity and redundant resources.

A more diverse system is generally a more resilient one, but diversity adds to cost and complexity. Diversity of connections is most efficiently achieved using infrastructure whose cost is shared by many operators, but collective-action problems can undermine the resilience gain **[C:7] [Q:9]**. It is efficient to avoid duplication of effort in the development of software and equipment, and efficient to exploit economies of scale in its manufacture, but this reduces the diversity of equipment used **[C:9]**. It is efficient for the entire Internet to depend on one protocol for its routing, but this creates a single point of failure. Setting up and maintaining multiple, diverse, separate connections to other networks costs time and effort and creates extra complexity to be managed **[C:6]**.

The Internet is a loosely coupled collection of independently managed networks. However, at its core there are a few very large networks, each of which strives to be as efficient as possible both internally and in its connections to other networks. So it is an open question whether the actual structure of the Internet is as resilient as its architecture would suggest. In the past it has been remarkably resilient, and it has continued to perform as it has evolved from a tiny network connecting a handful of research facilities into the global infrastructure that connects billions today. However, as in other areas, past performance is no guarantee of future results.

1.5 Resilience and Equipment

A particular concern for the interconnection system is the possibility of an internal technical problem that could have a systemic effect. The imminent changeover to IPv6 will provide a high-stress environment in which such a problem could be more likely to manifest itself, and the most likely proximate cause of such a problem is bugs in BGP implementations, which could be serious given the small number of equipment vendors for this kind of equipment. **[C:9]** There have been a number of incidents in which large numbers of routers across the entire Internet have been affected by the same problem, something unprecedented and unexpected which exposes a bug in the software, and occasionally in the specification of BGP.

No software is free from bugs, but the universal dependence on BGP makes bugs there more serious. ISPs may test equipment before buying and deploying it, but those tests concentrate on issues directly affecting the ISP, such as the performance of the equipment and its ability to support the required services. Manufacturers test their equipment as part of their development process. But the interests of both ISPs and manufacturers are for the equipment to work well under normal circumstances. Individual ISPs cannot afford to do exhaustive testing of low-probability scenarios for the benefit of the Internet at large. The manufacturers for their part balance the effort and time spent testing against their customers' demands for new and useful features, new and faster routers and less expensive software. Also of concern is how secure routers and routing protocols are against deliberate attempts to disrupt or suborn them.

A number of respondents to the consultation felt that money spent on testing equipment and protocols would be money well spent. **[C:10]**

1.6 Service Level Agreements (SLAs) and 'Best Efforts'

In any market in which the buyer has difficulty in establishing the relative value of different sellers' offerings, it is common for sellers to offer guarantees to support their claims to quality. Service Level Agreements (SLAs) perform that function in the interconnection ecosystem. From a resilience perspective, it would be nice to see ISPs offering SLAs that covered not just their own networks but the interconnection system too, and customers preferring to buy service with such SLAs.

Unfortunately, SLAs for Internet access in general are hard, and for transit service are of doubtful value **[C:20]**. In particular, where an operator offers an SLA, it does not extend beyond the borders of their network **[C:19]**; so whatever their guarantees are, they do not cover the interconnection system – the part between the borders of all networks.

The SLAs offered to end-customers by their ISPs reflect the SLAs that ISPs obtain from their transit providers and peers. The standard SLAs offered to end-customers may be published, but the SLAs offered between networks may be part of contracts that are kept confidential. Given how little such SLAs are generally thought to cover, it is an open question how much information is being hidden here – but it is another aspect of the general lack of information about the ecosystem at all levels. (The consultation asked specifically about inter-provider agreements, see section 9, question 8.)

Providers do not attempt to guarantee anything beyond their borders because they cannot. Any such guarantee would require a back-to-back system of contracts between networks so that liability for a failure to perform would be borne by the failing network. That system of contracts does not exist, not least because the Internet is not designed to guarantee performance. It is fundamental to the current Internet architecture that packets are delivered on a 'best efforts' basis, that is, the network will do its best but it does not guarantee anything. The Internet leaves the hard work of maintaining a connection to the end-points of the connection – the 'end-to-end' principle. The Transmission Control Protocol (TCP), which carries most Internet traffic apart from delay-sensitive traffic, will reduce demand if it detects congestion – it is designed to adapt to the available capacity, not to guarantee some level of performance.

The other difficulty with SLAs is what can and what should be measured. For a single connection between *a* and *b* it is clear what can be measured, but it is not clear what level of performance could be guaranteed, or by whom. Consider a connection from *a* in one network to *b* in another network, which traverses four other networks and the connections between them:

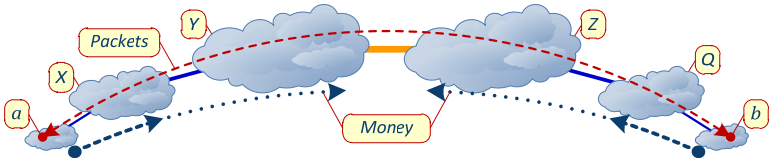


Figure 1: Connection between *a* and *b*

All these networks are independent, and have their own SLAs, each extending only as far as their borders. If we follow the money, *a* is paying directly and indirectly for packets to and from the connection between networks *Y* and *Z*. Similarly, *b* is paying for packets to and from the mid-point on the other side. If network *Q* has low standards, or is having a bad day, to whom does *a* complain? Network *X* has a contract with *a*'s network, and offers an SLA, but that does not extend beyond *X*. Network *Y* has a contract with *X*, with a different SLA, but even if *X* complained to *Y* about its customer's problem we have come to the end of the money trail: *Y* cannot hold *Z* to account for the performance of *Q*. Suppose *a* were to demand a strong SLA from their provider: *X* certainly has no way of imposing some standard of service on *Q*, and simply cannot offer to make any guarantee.

Even if it were possible to establish an end-to-end SLA for this connection, and pin liability on the failing network, there are hundreds of thousands of paths between *a*'s network and the rest of the Internet. The problem is intractable. So whatever value SLAs have, they do not offer a contractual framework through which customers can influence the resilience of the interconnection system, even if they wanted to. In addition, few customers understand the issue, or care to do anything about it.

Generally the Internet is remarkably reliable, so customers' principal interest in choosing a supplier is price – possibly moderated by the suppliers' reputation. **[C:18]**

1.7 Reachability, Traffic and Performance

While end-users care about traffic and performance, the basic mechanism of the interconnection system – BGP – only understands reachability **[Q:11]**. Its function is to provide a way for every network to reach every other network, and for traffic to flow across the Internet from one network to another. All ASes (the ISPs and other networks that make up the Internet) speak BGP to each other, and reachability information spreads across the 'BGP mesh' of connections between them. BGP is the heart of the interconnection system, so its many deficiencies are a problem. **[Q:16]**

The problems with the protocol itself include:

- there is no mechanism to verify that the routing information distributed by BGP is valid. In principle traffic to any destination can be diverted – so traffic can be disrupted, modified, examined or all three. These security issues are discussed separately in Section 1.10.
- there is no mechanism in BGP to convey capacity information – so BGP cannot help reconfigure the interconnection system to avoid congestion. **[Q:12]** When a route fails, BGP will find another route to maintain reachability, but that route may not have sufficient capacity for the traffic it now receives.
- the mechanisms in BGP which may be used to direct traffic away from congestion in other networks – 'inter-domain traffic engineering' – are strictly limited.
- when things change BGP can be slow to settle down ('converge') to a new, stable state. **[C:12]**
- the ability of BGP to cope or cope well under extreme conditions is not assured.

End-users expect to be able to reach every part of the Internet, so reachability is essential. But they also expect to be able to move data to and from whatever destination they choose, so they expect their connection with that destination to perform well. As BGP knows nothing about traffic, capacity or performance, network operators must use other means to meet end-users' expectations. When something in the Internet changes, BGP will change the routes used to ensure continuing reachability, but it is up to the network operators to ensure that the result will perform adequately, and take other steps if it does not.

Service quality in a 'best efforts' network is all to do with avoiding congestion, for which it is necessary to ensure that there is always sufficient capacity. The most effective way to do that is to maintain enough spare capacity to absorb the usual short-term variations in traffic and provide some safety margin. Additional spare capacity may be maintained to allow time (weeks or months, perhaps) for new capacity to be installed to cater for long-term growth of traffic. Maintaining spare capacity in this way is known as 'over-provisioning'; it is key to day-to-day service quality and to the resilience of the interconnection system.

Each operator constantly monitors its network for signs of congestion and will make adjustments to relieve any short-term issues. In general the pattern of traffic in a network of any size is stable from day to day and month to month. An operator will also monitor their network for long-term trends in traffic. The management of capacity is generally done on the basis of history, experience and rules of thumb, supported by systems for gathering and processing the available data. The levels of spare

capacity in any network will depend on many things, including how the operator chooses to balance the cost of spare capacity against the risk of congestion.

A key point here is that capacity is managed on the basis of actual traffic and the usual day-to-day events, with some margin for contingencies and growth. Capacity is not managed on the basis of what might happen if some unusual event causes a lot of traffic to shift from one network to another. If an event has a major impact on the interconnection system, then the amount of spare capacity within and between networks will determine the likelihood of systemic congestion. So each individual network's degree of over-provisioning makes some contribution to the resilience of the whole – though it is hard to say to what extent.

If an event disables some part of the Internet, BGP will work to ensure that reachability is maintained, but the new paths may have less capacity than the usual ones, which may result in congestion. For many applications, notably web-browsing, the effect is to slow things down, but not stop them working. More difficulties arise with any sort of data that is affected by reduced throughput or increased delay, such as VoIP and streaming video. Congestion may stop these applications working satisfactorily, or at all.

The important distinction between reachability and traffic is illustrated by considering what appears to be a simple metric for the state of the Internet: the percentage of known destinations that are reachable from most of the Internet at any given moment. This metric may be used to gauge the impact of a BGP failure, or of the failure of some critical fibre, or any other widely felt event. But while the significance of, say, 10% of known destinations becoming unreachable is obviously extremely high for the 10% cut off, it may not be terribly significant for the rest of the Internet. We would prefer to know the amount, and possibly the value, of traffic that is affected. If the 10% cut off accounts for a large proportion of the remaining 90%'s traffic, the impact could be significant. So when talking about the resilience of the system, what is an 'acceptable level' of the 'best efforts' service? Are we aiming at having email work 95% of the time to 95% of destinations, or streaming video work 99.99% of the time to 99.99% of destinations? The answer will have an enormous effect on the spare capacity needed! Each extra order of magnitude improvement (say from 99% to 99.9%) could cost an order of magnitude more money; yet the benefits of service quality are unevenly distributed. For example, a pensioner who uses the Internet to chat to grandchildren once a week may be happy with 99% or even 90%, while a company providing a cloud-based business service may need 99.99% or more.

1.7.1 Traffic Prioritisation

In a crisis it is common for access to some resources to be restricted, to shed demand and free up capacity. For telephony a traditional approach is to give emergency services priority. But restricting phone service to 'obvious' emergency workers such as doctors is unsatisfactory. Modern medical practice depends on team working and can be crippled if nurses are cut off; and many patients who depend on home monitoring may have to be hospitalised if communications fail.

If capacity is lost in a disaster and parts of the system are congested, then all users of the congested parts will suffer a reduction in service, and some types of traffic (notably VoIP) may stop working effectively. If some types, sources or destinations of traffic are deemed to be important, and so should be given priority in a crisis, then serious thought needs to be given to how to identify priority traffic, how the prioritisation is to be implemented and how turning that prioritisation on and off fits into other disaster planning. **[Q:19]**

It is not entirely straightforward to identify different types of traffic. So an alternative approach may be to prioritise by source or destination. It may be tempting to consider services such as Facebook or YouTube as essentially trivial, and YouTube uses a lot of bandwidth. However, in a crisis keeping in contact using Facebook may be a priority for many. Moreover, shutting down YouTube in a crisis – thereby preventing the free reporting of events – would require solid justification. On the other hand, rate limiting ordinary users, irrespective of traffic type, may appear fair, but could affect essential VoIP use, and cutting off peer-to-peer traffic could be seen as censorship.

So it is inappropriate for ISPs to decide to discriminate between different sorts of traffic, or between customers of the same type (although premium customers at premium rates might expect to get better performance in a crisis). **[Q:21]** It is not even clear that ISPs are, in general, capable of prioritising some traffic on any given basis. So, if some traffic should be prioritised in a crisis, who will make the call, and will anyone be ready to act when they do?

It is clear that this challenge entails both technical and policy aspects. The former are related mainly to the mechanisms that should exist in network equipment to support traffic prioritisation. The latter refer mainly to the policies that specify what traffic should be given priority. It is very important to tackle both aspects of the problem.

1.7.2 Traffic Engineering

'Traffic Engineering' is the jargon term for adjusting a network so that traffic flows are improved. In a crisis that would mean shifting traffic away from congested paths. This is less controversial than traffic prioritisation, but no less difficult.

When some event creates congestion in some part(s) of the interconnection system it would be convenient if networks could redirect some traffic away from the congested parts. When a network is damaged its operators will work to relieve congestion within their network by doing internal traffic engineering, adding temporary capacity, repairing things, and so on. One of the strengths of the Internet is that each operator will be working independently to recover its own network as quickly and efficiently as possible.

Where a network's users are affected by congestion in other networks, the simplest strategy is to wait until those networks recover. This may leave spare capacity in other networks unused, so is not the optimum strategy for the system as a whole. However, there are two problems with trying to coordinate action:

1. there is no way of telling where the spare capacity in the system is;
2. BGP provides very limited means to influence traffic in other operators' networks.

In effect, if networks attempt to redirect traffic they are blundering around in the dark, attempting to make adjustments to a delicate instrument with a hammer. Their attempts to redirect traffic may create congestion elsewhere, which may cause more networks to try to move traffic around. It is possible to imagine a situation in which many networks are chasing each other creating waves of congestion and routing changes as they do, like the waves of congestion that pass along roads which are near their carrying capacity.

With luck, if a network cannot handle the traffic it is sent and pushes it away to other networks, it will be diverted towards spare capacity elsewhere. Given enough time the system would adapt to a new distribution of capacity, and a new distribution of traffic. It is impossible to say how much time

would be required; it would depend on the severity of the capacity loss, but it could be days or even weeks.

Strategic local action will not necessarily lead to a socially optimal equilibrium, though, as the incentives may be perverse. Since any SLA will stop at the edge of its network, a transit provider may wish to engineer traffic away from its network in order to meet its SLAs for traffic within its network. The result may still be congestion, somewhere, but the SLA is still met.

1.7.3 Routing in a Crisis

Experience shows that in a crisis the interconnection system can quite quickly create new paths between networks to provide interim connections and extra capacity – for example, in the aftermath of the '9/11' attack, as discussed above.

The interconnection ecosystem has often responded in this way with many people improvising, and working with the people they know personally. **[C:13]** This is related to traffic engineering, to the extent that it addresses the problem by adding extra connections to which traffic can be moved. The response of the system might be improved and speeded up if there were more preparation for this form, and perhaps other forms, of cooperation in a crisis. **[C:14]**

In the end, if there is insufficient capacity in a crisis, then no amount of traffic engineering or manual reconfiguration will fit a quart of traffic into a pint of capacity. In extreme cases some form of prioritisation would be needed.

1.8 Is Transit a Viable Business?

The provision of transit – the service of carrying traffic to every possible destination – is a key part of the interconnection system, but it may not be a sustainable business in the near future.

Nobody doubts that the cost of transit has fallen fast, or that it is a commodity business, except where there is little or no competition. In the US, over the last ten to fifteen years transit prices have fallen at rate of around 40% per annum – which results in a 99% drop over a ten year period. In other parts of the world prices started higher, but as infrastructure has developed, and transit networks have extended to into new markets, those prices have fallen – for example, prices in London are now scarcely distinguishable from those in New York.

Where there is effective competition, the price of transit falls, and consumers benefit. In a competitive market, price tends towards the *marginal* cost of production. The *total* cost of production has fallen sharply, as innovation reduces the cost of the underlying technologies and with increasing economies of scale. Yet every year industry insiders feel that surely nobody can make money at today's prices, and that there must soon be a levelling off. So far there has been no levelling off, though the rate at which prices fall may be diminishing.

The reason is simple: the *marginal* cost of production for transit service is generally *zero*. At any given moment there will be a number of transit providers with spare capacity: first, network capacity comes in lumps, so each time capacity is added the increment will generally exceed the immediate need; second, networks are generally over-provisioned, so there is always some spare capacity – though eating into that may increase the risk of congestion, perhaps reducing service quality at busy times or when things go wrong.

The logic of this market is that the price for transit will tend towards zero. So it is unclear how pure transit providers could recoup their capital investment. The logic of the market would appear to favour consolidation until the handful of firms left standing acquire market power.

At a practical level, the provision of transit may be undertaken not to make profits, but to offset some of the cost of being an Internet network. For some networks the decision to offer transit at the market price may be increasingly a strategic rather than a commercial decision. Another significant factor is the recent and continuing increase in video traffic and the related rise in the amount of traffic delivered by the Content Delivery Networks (CDNs, see below). This means that the continued reduction in the unit price for transit is not being matched by an increase in transit traffic, so transit providers' revenues are decreasing.

The acknowledged market leader, Level 3, lost $2.9 billion in 2005-2008 and a further $0.6 billion in 2009, and another $0.6 billion in 2010. It is not possible to say what contribution their transit business made to this; industry insiders note that Level 3 did not go through bankruptcy as many others did, and would make a small profit if it were not for the cost of servicing its debt. However, the industry as a whole is losing large amounts of money (we summarise some of the major providers' financial statements in Appendix II).

1.9 The Rise of the Content Delivery Networks

Over the past four years or so, more and more traffic has been delivered by Content Delivery Networks (CDNs). Their rise has been rapid and has changed the interconnection landscape, concentrating a large proportion of Internet traffic into a small number of networks. This shift has been driven by both cost and quality considerations. With the growth of video content, of ever richer web-sites, and of cloud applications, it makes sense to place copies of popular data closer to the end users who fetch it. This has a number of benefits:

- local connections perform better than remote ones – giving quicker response and faster transfers.
- costs are reduced because the data is not being repeatedly transported over large distances – saving on transit costs. However, the key motivation for the customers of CDNs is not to reduce the cost of delivery, but to ensure quality and consistency of delivery – which is particularly important for the delivery of video streams;
- the data are replicated, stored in and delivered from a number of locations – improving resilience.

This has moved traffic away from transit providers to peering connections between the CDNs and the end-user's ISP. In some cases content is distributed to servers within the ISP's own network, bypassing the interconnection system altogether.

One CDN claims to deliver some 20% of all Internet traffic. Since the traffic being delivered is the sort which is expected to grow most quickly in the coming years, this implies that an increasing proportion of traffic is being delivered locally, and a reducing proportion of traffic is being carried (over long distances) by the transit providers.

Another effect of this is to add traffic at the Internet Exchange Points (IXPs), which are the obvious way for the CDNs to connect to local ISPs. This adds value to the IXP – particularly welcome for the smaller IXPs, which have been threatened by the ever falling cost of transit (eating into the cost

advantage of connecting to the IXP) and the falling cost of connecting to remote (larger) IXPs (where there is more opportunity to pick up traffic).

There is a positive effect on resilience, and a negative one. The positive side is that systems serving users in one region are independent of those serving users in other regions, so a lot of traffic becomes less dependent on long distance transit services. On the negative side, CDNs are now carrying so much traffic that if a large one were to fail, transit providers could not meet the added demand, and some services would be degraded. CDNs also concentrate ever more infrastructure in places where there is already a lot of it. If parts of some local infrastructure fail for any reason, will there be sufficient other capacity to fall back on?

Finally, it is possible to count a couple of dozen CDNs quite quickly, but it appears that perhaps two or three are dominant. Some of the large transit providers have entered the business, either with their own infrastructure or in partnership with an existing CDN. There are obvious economies of scale in the CDN business, and there is now a significant investment barrier to entry. The state of this market in a few years' time is impossible to predict, but network effects tend to favour a few, very large, players. These players are very likely to end up handling over half the Internet's traffic by volume.

1.10 The "Insecurity" of BGP

A fundamental problem with BGP is that there is no mechanism to verify that the routing information it distributes is valid. In principle traffic to any destination can be diverted – so traffic can be disrupted, modified, examined or all three. **[C:11]** The effect of this is felt on a regular basis when some network manages to announce large numbers of routes for addresses that belong to other networks; this can divert traffic into what is effectively a black hole. Such incidents are quite quickly dealt with by network operators, and disruption can be limited to a few hours, at most. It is worth remembering that the operational layer is part of the ecosystem, and not all problems require technical solutions.

The great fear is that this insecurity might be exploited as a means to deliberately disrupt the Internet, or parts of it. There is also a frequently expressed concern that route hijacking might be used to listen in on traffic, though this can be hard to do in practice.

Configuring BGP routers to filter out invalid routes, or only accept valid ones, is encouraged as best practice. However, as discussed in Section 3.1.11, where it is practical (at the edges of the Internet) it does not make much difference, until most networks do it. Where it would make most difference (in the larger transit providers) it is not really practical because the information on which to base route filters is incomplete and the tools available to manage and implement filters at that scale are inadequate. **[Q:13]**

More secure forms of BGP, in which routing information can be cryptographically verified, depend on there being a mechanism to verify the 'ownership' of blocks of IP addresses, or to verify that the AS which claims to be the origin of a block of IP addresses is entitled to make that claim. The notion of title to blocks of IP addresses turns out not to be as straightforward as might be expected. However, some progress is now being made, under the name RPKI (Resource Public Key Infrastructure). The RPKI initiative should allow ASes to ignore announcements where the origin is invalid – that is, where some AS is attempting to use IP addresses it is not entitled to use. This is an important step forward, and might tackle over 90% of 'fat finger' problems (outages caused by mistakes rather than deliberate attempts to disrupt). **[Q:14]**

But the cost of RPKI is significant. Every AS must take steps to document their title to their IP addresses, and that title must be registered and attested to by the Internet Registries. Then, every AS must extend their infrastructure to check the route announcements they receive against the register. What is more, the problem that RPKI tackles is, so far, largely a nuisance not a disaster. When some network manages to announce some routes it should not, this is noticed and fixed quite quickly, if it matters. Sometimes a network announces IP addresses nobody else is using – generally they are up to no good, but this does not actually disrupt the interconnection system. So the incentive to do something about the problem is weak, although the number of such incidents is expected to rise when IPv4 addresses are exhausted in late 2011.

Further, a route may pass the checks supported by RPKI, and still be invalid. A network can announce routes for a block of IP addresses, complete with a valid origin, but do so only to disrupt or interfere with the traffic (apparently) on its way to its destination. The S-BGP extensions to BGP (first published in 1997) address the issue more completely, and there have been other proposals since; however, they make technical assumptions about routing (traffic greed and valley-free customer preferences) that don't hold in today's Internet. Details of a new initiative, BGPSEC, were announced in March 2011. The aim is that this should lead to IETF standards by 2013 and deployed code in routers thereafter.

During the standardisation process in 2011-2013 a key issue will be security economics. ASes see the cost of BGP security as high, and the benefit essentially zero until it is very widely deployed. Ideally, implementation and deployment strategies will give local, incremental benefit, coupled with incentives for early adopters. One possible mechanism is for governments to use their purchasing power to bootstrap early adoption; another is for routers to prefer signed routes. Technical issues that must be studied during the standardisation phase include whether more secure BGP might, in fact, be bad for resilience (as was pointed out in the consultation, **[Q:15]**). Adding cryptography to a system can make it brittle. The reason is that when recovering from an event, new and possibly temporary routes may be distributed in order to replace lost routes, and if the unusual routes are rejected because they do not have the necessary credentials, then recovery will be harder. Finally, BGPSEC will not be a silver bullet, there are many threats, but it should tackle about half the things that can go wrong after RPKI has dealt with origin validation.

To sum up, most of the time BGP works wonderfully well, but there is plenty of scope to make it more secure and more robust. However, individual networks will get little direct benefit from an improved BGP, despite the significant cost. We will probably need some new incentive to persuade networks to invest in more secure BGP, or a proposal for securing BGP that gives local benefits from incremental deployment. **[Q:20]**

1.11 Cyber Exercises on Interconnection Resilience

The practical approach to assessing the resilience of the interconnection system is to run large-scale exercises in which plausible scenarios are tested. **[C:16]** Exercises can test both operational and technical aspects as well as procedural, policy, structural and communication aspects.

Such exercises have a number of advantages and benefits.

- They start with real world issues. These exercises are not cheap, so there is an incentive to be realistic: planners consider what really are the sorts of event that the system is expected to face.

- They can identify some dependencies on physical infrastructure. By requiring the participants to consider the effects of some infrastructure failure, an exercise may reveal previously unknown dependencies.
- They can identify cross-system dependencies. For example, how well can network operations centres communicate if the phone network fails, or how well can field repairs proceed if the mobile phone network is unavailable? **[Q:17]**
- They exercise disaster recovery systems and procedures. This is generally a good learning experience for everybody involved, particularly as otherwise crisis management is generally ad hoc. **[C:15]**

Such scenario testing has been done at a national level and found to be valuable[3]. Something at a larger scale has also been proved to be valuable.

On 4th November 2010 the European Member States organised the first pan-European cyber exercise, called CYBER EUROPE 2010, which was facilitated by ENISA. The final evaluation report published by ENISA[4] proves the importance of such exercises and calls for future actions based on the lessons learned.

1.12 The "Tragedy of the Commons"

The resilience of the Internet interconnection system benefits everyone, but an individual network will not in general gain a net benefit if it increases its costs in order to contribute to the resilience of the whole. **[C:21]**

This manifests itself in a number of ways.

- In Section 1.10 above, we discussed the various proposals for more secure forms of BGP, from S-BGP in 1997 to BGPSEC in 2011, none of which have so far been deployed (see Section 3.1.12). There is little demand for something which is going to be difficult to implement and whose direct benefit is limited.
- There exists best practice for filtering BGP route announcements, which, if universally applied, would reduce instances of invalid routes being propagated by BGP and disrupting the system (see Section 3.1.11). But these recommendations are difficult to implement and mostly benefit other networks, so are not often implemented.
- There is an IETF BCP[5] [**6**] for filtering packets, to reduce 'address spoofing', which would mitigate denial of service attacks (see Section 5.8.3). These recommendations also mostly benefit others, so are not often implemented.
- A smaller global routing table would reduce the load on all BGP routers in the Internet, and leave more capacity to deal with unusual events. Nevertheless, the routing table is as about

[3] *Good Practice Guide on National Cyber Exercises,* ENISA Technical Report, *2009. Available at: http://www.enisa.europa.eu/act/res/policies/good-practices-1/exercises*

[4] *CYBER EUROPE 2010-Evaluation Report,* ENISA Report 2011. *Available (after 15/04/2011) at: http://www.enisa.europa.eu/act/res/*

[5] *An Internet Engineering Task Force (IETF) Best Common Practice (BCP) is as official as it gets in the Internet.*

75% bigger than it needs to be, because some networks announce extra routes to reduce their own costs (see Section 3.1.9). Other networks could resist this by ignoring the extra routes, but that would cost time and effort to configure their routers, and would most likely be seen by their customers as a service failure (not as a noble act of public service).

- The system is still ill-prepared for IPv6, despite the now imminent (circa Q3 2011) exhaustion of IPv4 address space. **[Q:10]**

It is in the clear interest of each network to ensure that in normal circumstances 'best efforts' means a high level of service, by adjusting interconnections and routing policy – each network has customers to serve and a reputation to maintain **[C:17]**. Normal circumstances include the usual day-to-day failures and small incidents **[Q:7]**.

The central issue is that the security and resilience of the interconnection system is an externality as far as the networks that comprise it are concerned. It is not clear is that there is any incentive for network operators to put significant effort into considering the resilience of the interconnection system under extraordinary circumstances. **[Q:18]**

1.13 Regulation

Regulation is viewed with apprehension by the Internet community. Studies such as this are seen as stalking horses for regulatory interference, which is generally thought likely to be harmful. **[C:22]** Despite having its origins in a project funded by DARPA, a US government agency, the Internet has developed since then in an environment that is largely free from regulation. There have been many local attempts at regulatory intervention, most of which are seen as harmful.

- The governments of many less developed countries attempt to censor the Internet, with varying degrees of success. The 'Great Firewall of China' is much discussed, but many other states practice online censorship to a greater or lesser extent. It is not just that censorship itself is contrary to the mores of the Internet community – whose culture is greatly influenced by California, the home of many developers, vendors and service companies. Attempts at censorship can cause collateral damage, as when Pakistan advertised routes for YouTube in an attempt to censor it within their borders, and instead made it unavailable on much of the Internet for several hours.
- Where poor regulation leads to a lack of competition, access to the Internet is limited and relatively expensive. In many less developed countries, a local telecommunications monopoly restricts wireline broadband access to urban elites, forcing the majority to rely on mobile access. However the problem is more subtle than 'regulation bad, no regulation good'. In a number of US cities, the diversity of broadband access is falling; cities that used to have three independent infrastructures (say from a phone company, a cable company and an electricity company) may find themselves over time with two, or even just one. In better-regulated developed countries (such as much of Europe) local loop unbundling yields price competition at least, thus mitigating access costs, even if physical diversity is harder. Finally, few countries impose a universal service provision on service providers; its lack can lead to a 'digital divide' between populated areas with broadband provision, and rural areas without.
- There has been continued controversy over surveillance for law-enforcement and intelligence purposes. In the 'Crypto Wars' on the 1990s, the Clinton administration tried to control cryptography, which the industry saw as threatening not just privacy but the growth of e-commerce and other online services. The Clinton administration passed the

Communications Assistance for Law Enforcement Act (CALEA) in 1994 mandating the cooperation of telecommunications carriers in wiretapping phone calls. The EU has a Data Retention Directive that is up for revision in 2011 and there is interest both in the UK and the USA in how wiretapping should be updated for an age not only of VoIP but also of diverse messaging systems. This creates conflicts of interest with customers, raises issues of human rights, and leads to arguments about payment and subsidy.

- Governments which worry about Critical National Infrastructure may treat Internet regulation as a matter of National Security, introducing degrees of secrecy and shadowy organisations, which does nothing to dispel concerns about motivation – not helped by a tendency to talk about the problem in apocalyptic terms[6].

Whatever the motivation, government policies are often formulated with insufficient scientific and technical input. They often manage to appear clueless, and in some cases make things worse. This study is an attempt to help alleviate this problem.

This study has identified a number of areas where the market does not appear to provide incentives to maintain the resilience of the interconnection system at a socially optimal level. However, any attempt to tackle any of the issues by regulation is hampered by a number of factors:

- the lack of good information about the state and behaviour of the system. It is hard to determine how material a given issue may be. It is hard to determine what effect a given initiative is likely to have – good or bad.
- the scale and complexity of the system. Scale may make local initiatives ineffective, while complexity means that it is hard to predict how the system will respond or adapt to a given initiative.
- the dynamic nature of the system. CDNs have been around for many years, but their emergence as a major component of the Internet is relatively recent; it is testament to the system's ability to adapt quickly (in this case, to the popularity of streamed video).

Up until now, the lack of incentives to provide resilience (and in particular to provide excess capacity) has been relatively unimportant: the Internet has been growing so rapidly that it has been very far from equilibrium, with a huge endowment of surplus capacity during the dotcom boom and significant capacity enhancements since then. This cannot go on forever.

One caveat: we must point out that the privatisation, liberalisation and restructuring of utilities worldwide has led to institutional fragmentation in a number of critical infrastructure industries that could in theory suffer degradation of reliability and resilience for the same general microeconomic reasons we discuss in the context of the Internet. Yet studies of the electricity, water and telecomms industries in a number of countries have failed to find a reliability deficit thus far [**7**]. In practice, utilities have managed to cope by a combination of anticipatory risk management and Public-Private Partnerships (PPPs). However it is sometimes necessary for government to act as a 'lender of last resort'. If a router fails, we can fall back on another router, but if a market fails – as with the California electricity market – there is no fall-back other than the state.

[6] *See [**236**] UK Government, Cabinet Office Factsheet 18: Cyber Security. And for the popular perception of what government thinks see [**237**] "Fight Cyber War Before Planes Fall Out of Sky".*

In conclusion, it may be some time before regulatory action is called for to protect the resilience of the Internet, but it may well be time to start thinking about what might be involved. Regulating a new technology is hard; an initiative designed to improve today's system may be irrelevant to tomorrow's, or, worse, stifle competition and innovation. For example, the railways steadily improved their efficiency from their inception in the 1840s until regulation started in the late nineteenth century, after which their efficiency declined steadily until competition from road freight arrived in the 1940s [**8**].

The prudent course of action for policy makers today is to start working to understand the Internet interconnection ecosystem. The most important package of work is to increase transparency, by supporting consistent, thorough, investigation of major outages and the publication of the findings, and by supporting long-term measurement of network performance. The second package we recommend is to fund key research in topics such as distributed intrusion detection and the design of security mechanisms with practical paths to deployment, and the third is to promote good practice, to encourage diverse service provision and to promote the testing of equipment. The fourth package includes the preparation and relationship-building through a series of PPPs for resilience. Modest and constructive engagement of this kind will enable regulators to build relationships with industry stakeholders and leave everyone in a much better position to avoid, or delay, difficult and uninformed regulation. Regulatory intervention must after all be evidence-based; and while there is evidence of a number of issues, the workings of this huge, complex and dynamic system are so poorly understood that there is not yet enough evidence on which to base major regulatory intervention with sufficient confidence.

2 Recommendations

Our recommendations come in four groups. The first group is aimed at understanding failures better, so that all may learn the lessons.

Recommendation 1 Incident Investigation

An independent body should thoroughly investigate all major incidents and report publicly on the causes, effects and lessons to be learned. Incident correlation and analysis may lead to assessment and forecast models. The appropriate framework should be the result of a consultation with the industry and the appropriate regulatory authorities. Incident investigation might be undertaken by an industry association, by a national regulator or by a body at the European level, such as ENISA. The last option would require funding to support the work, and, perhaps, powers to obtain information from operators – under suitable safeguards to protect commercially sensitive information. The implementation of Article 13a of the recent EU Telecom Package[7] may provide a model for this.

Recommendation 2 Data Collection of Network Performance Measurements

Europe should promote and support consistent, long-term and comprehensive data collection of network performance measurements. At present some real-time monitoring is done by companies such as ArborNet and Renesys, and some more is done by academic projects – which tend to languish once their funding runs out. This patchwork is insufficient. There should be sustainable funding to support the long-term collection, processing, storage and publication of performance data. This also has a network management / law enforcement angle in that real-time monitoring of the system could help detect unusual route announcements and other undesirable activity.

The second group of recommendations aims at securing funding for research in topics related to resilience – with an emphasis not just on the design of security mechanisms, but on developing an understanding of how solutions can be deployed in the real world.

Recommendation 3 Research into Resilience Metrics and Measurement Frameworks

Europe should sponsor research into better ways to measure and understand the performance and resilience of huge, multi-layered networks. This is the research aspect of the second recommendation; once that provides access to good data, the data should help clever people to come up with better metrics.

[7] *Directive 2002/21/EC of the European Parliament and of the Council, of 7 March 2002, on a common regulatory framework for electronic communications networks and services (Framework Directive), as amended by Directive 2009/140/EC and Regulation 544/2009.*

Recommendation 4 Development and Deployment of Secure Inter-domain Routing

Europe should support the development of effective, practical mechanisms which have enough incentives for deployment. This may mean mechanisms that give local benefit to the firms that deploy them, even where deployment is incremental; it may require technical mechanisms to be supplemented by policy tools such as the use of public-sector purchasing power, subsidies, liability shifts, or other kinds of regulation.

Recommendation 5 Research into AS Incentives that Improve Resilience

Europe should support research into economic and legal mechanisms to increase the resilience of the Internet. Perhaps a system of contracts can be constructed to secure the interconnection system, starting with the connections between the major transit providers and spreading from the core to the edges. Alternatively, researchers might consider whether liability rules might have a similar effect. If the failure of a specific type of router caused loss of Internet service leading to damage and loss of life, the Product Liability Directive 85/374/EC would already let victims sue the vendor; but there is no such provision relating to the failure of a transit provider.

The third group of recommendations aims at promoting good practice.

Recommendation 6 Promotion and Sharing of Good Practice on Internet Interconnections

Europe should sponsor and promote good practice in network management. Where good practice exists its adoption may be hampered by practical and economic issues. The public sector may be able to help, but it is not enough to declare for motherhood and apple pie! It can contribute various incentives, such as through its considerable purchasing power. For that to be effective, purchasers need a way to tell good service. The first three of our recommendations can help, but there are some direct measures of quality too. Such information sharing should include modest and constructive engagement of industry stakeholders with public sector in relationship-building strategic dialogue and decisions through a series of PPPs for resilience.

Recommendation 7 Independent Testing of Equipment and Protocols

Public bodies at national or European-level should sponsor the independent testing of routing equipment and protocols. The risk of systemic failure would be reduced by independent testing of equipment and protocols, looking particularly for how well these perform in unusual circumstances, and whether they can be disrupted, suborned, overloaded or corrupted.

Recommendation 8 Conduct Regular Cyber Exercises on the Interconnection Infrastructure

The consultation noted that these are effective in improving resilience at local and national levels. The efforts at this level should continue in all countries in Europe as ‘we are as weak as the weakest link’. ENISA will support the national efforts. In addition regular pan-European exercises should be organised by European Member States in order to test and improve European-wide contingency plans (measures, procedures and structures). These large scale exercises will provide an umbrella for a number of useful activities, such as investigating what extra preparation might be required to

provide more routes in a crisis; thus effectively becoming part of improving the pan European cyber preparedness and contingency plans.

The final group of recommendations aims at engaging policymakers, customers and the public.

Recommendation 9 Transit Market Failure

It is possible that the current twenty-odd largest transit providers might consolidate down to a handful, in which case they might start to exercise market power and need to be regulated like any other concentrated industry. If this were to happen just as the industry uses up the last of its endowment of dark fibre from the dotcom boom, then prices might rise sharply. European policymakers should start the conversation about what to do then. Action might involve not just a number of European agencies but also national regulatory authorities. Recommendations 1, 2, 3, and 5 will prepare the ground technically so that policy makers will not be working entirely in the dark, but we also need political preparation.

Recommendation 10 Traffic Prioritisation

If, in a crisis, some traffic is to be given priority, and other traffic is to suffer discrimination, then the basis for this choice requires public debate, and mechanisms to achieve it need to be developed. Given the number of interests seeking to censor the Internet for various reasons, any decisions on prioritisation will have to be taken openly and transparently, or public confidence will be lost.

Recommendation 11 Greater Transparency – Towards a Resilience Certification Scheme

Finally, transparency is not just about openness in taking decisions on regulation or on emergency procedures. It would greatly help resilience if end-users and corporate customers could be educated to understand the issues and send the right market signals. In the long term efforts, including ENISA's, should focus on what mechanisms can be developed to give them the means to make more informed choices. This might involve combining the outputs from recommendations 2, 3, 5, 6 and 7 into a 'quality certification mark' scheme. Such scheme may prove an important tool to drive the market incentives towards enhancing the resilience of the networks and more generally of the interconnection ecosystem.

PART II – the State of the Art Review

Introduction to the State of the Art Review

The Internet connects a large number of independent networks, which cooperate to ensure that each network's users can reach every other network's users – directly or indirectly. The heart of the Internet, the Internet interconnection system, is a fabric of connections between these networks. The Internet's resilience is hugely important to us all; it depends on the resilience not just of the interconnection system, but of the component networks, particularly those large networks that provide services to smaller ones.

The interconnection ecosystem consists of:

- all the physical infrastructure that supports the networks and the links between them;
- the higher level network of connections between the independent component networks;
- the operational and commercial arrangements between them;
- the system of economic and other incentives that drive the whole system.

Within this ecosystem, each network acts in its own perceived best interests. The interests of the ecosystem as a whole are, essentially, a matter for the 'invisible hand'. The equilibrium (if we can call it that) arises from the behaviour of tens of thousands of independent networks, each seeking to maximise its own profits.

The question "Is the interconnection system resilient?" appears trivial. It is resilient by design – designed to "withstand nuclear attack"[8]. It is resilient in practice, for example, in the aftermath of '9/11'. Its decentralised structure means that it is both hard to attack and unlikely to all fail at the same time.

The question that motivates this study, however, is "How resilient is it?" In particular we consider:

- how it might cope with events with medium to high impact, which have corresponding medium to low probability;
- how its resilience may be assessed, assured and/or improved;
- what may influence its resilience in the long term.

We take a European perspective but, as with anything to do with the Internet, the context is clearly global. We exclude the day-to-day running of the ecosystem and individual networks. We also exclude the resilience of end-user connections to their ISPs.

This state of the art review proceeds as follows:

[8] *This is in fact apocryphal. In the Internet Society's "A Brief History of the Internet" **[230]** it is noted that: "It was from the RAND study that the false rumor started claiming that the ARPANET was somehow related to building a network resistant to nuclear war." See also **[231]**.*

- In Section 3 we describe the Internet interconnection ecosystem in some detail, identifying the key components and building up, layer by layer, an understanding of the system and how it works.
- In Section 4 we examine what we mean by resilience, how that may be assessed, and some general approaches to improving resilience.
- In Section 5 we draw the parts together and consider the resilience of the ecosystem, concentrating mainly on the practical issues.
- In Section 6 we look at the wider issues which shape and guide the system, and which may influence the resilience of the ecosystem in the medium and long term.
- In Section 7 we explore whether, given all of the above, there is cause for concern, and if so, why.

3 The Internet Interconnection Ecosystem

The Internet today (April 2011) comprises some 37,000 independent networks, all connected to each other. How this works and how it is maintained is key to assessing, maintaining and improving its resilience. This depends on two factors.

First, the resilience of the Internet as a whole depends on each network being resilient – from its end users to its interconnections with other networks. That is under the control of each network, individually and independently.

Second, it depends on the connections between networks, direct and indirect. Each direct connection between two networks is a bilateral and generally private arrangement, under the shared control of the two networks. Even the largest networks are only connected to a fraction of the total. Traffic between most pairs of networks, does not pass directly between them, but crosses other networks from source to destination. These indirect connections are underpinned by a system of incentives and bilateral agreements – formal and informal.

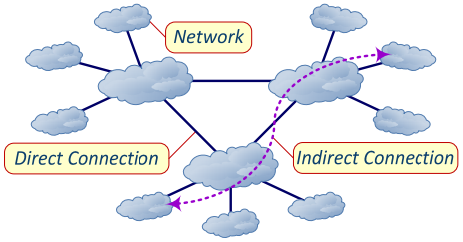


Figure 2: Direct and Indirect Connections

The system of direct and indirect connections between networks, and the incentives and agreements that underpin those are, together, the Internet Interconnection Ecosystem. This section identifies and describes its components, and the relationships between them, which are divided into the layers shown opposite. The bottom two layers, the Physical and Network Layers, contain the networks and the connections between them. The Operational Layer contains the people and systems that build and run the networks. The Commercial Layer contains the web of agreements between networks, driven by their business needs. The Economics Layer contains the economic incentives and drivers, and the Regulatory Layer any regulation governing the entire ecosystem.

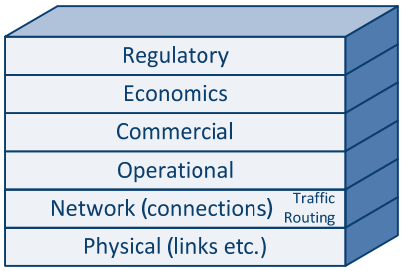


Figure 3: Many Layered Ecosystem

This section proceeds as follows:

- Section 3.1 describes the network layer which implements the direct and indirect connections between networks, arranging for every network to be able to reach every other network, and for data to be able to travel from anywhere to everywhere in the Internet.
- The physical layer – described in Section 3.2 – covers the equipment and the links on which the network layer depends – not forgetting the sites, the electricity supply, etc. on which they all depend.
- Operational and commercial arrangements between independent networks for the exchange of traffic are discussed next in Section 3.3.
- Section 3.5 gives a description of the different roles played by different classes of independent networks.

- The economic incentives in the system and how these drive the Internet Interconnection Ecosystem are discussed in Section 3.6.
- Section 3.7 discusses how the economic incentives and contractual relationships result in independent networks assuming some responsibility for resilience.
- The difficulties in actually mapping the Ecosystem are described in Section 3.8.
- The companies providing the physical transportation links have great difficulty extracting value from their networks other than by charging the market price for transport. This Problem of Value is discussed in Section 3.9.
- Regulation, discussed in Section 3.10, is the final layer of the Internet Interconnect Ecosystem.
- Finally, Section 3.11 is a summary of the Ecosystem.

For a good review of the complexity of Internet interconnections see [**9**]. In [**10**] the authors survey the evolution of the Internet Ecosystem over the ten years to 2008.

Scale and Growth

The Internet is very big. As of April 2011, it comprises some 37,000 independent networks which use some 350,000 distinct blocks of addresses, according to the 'CIDR Report' [**11**]. The CIDR Report web-site can provide graphs for the numbers of networks and address blocks over the last twenty years, as follows:

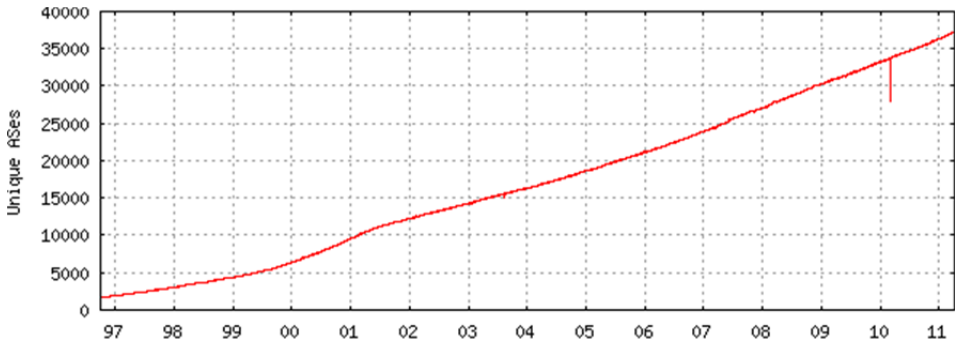


Figure 4: Number of Independent Networks, 1997 to Apr-2011 – Source: CIDR Report

In Figure 4 we see that in the last 14 years the number of networks in Internet has grown from ~2,000 to ~37,000. In the boom years of 1997-2001, 10,000 new networks connected to the Internet, and the growth rate was ~55% per annum, compound for those 4 years. Growth more recently appears approximately linear: and was about 10% in 2010 with about 3,300 new independent networks added.

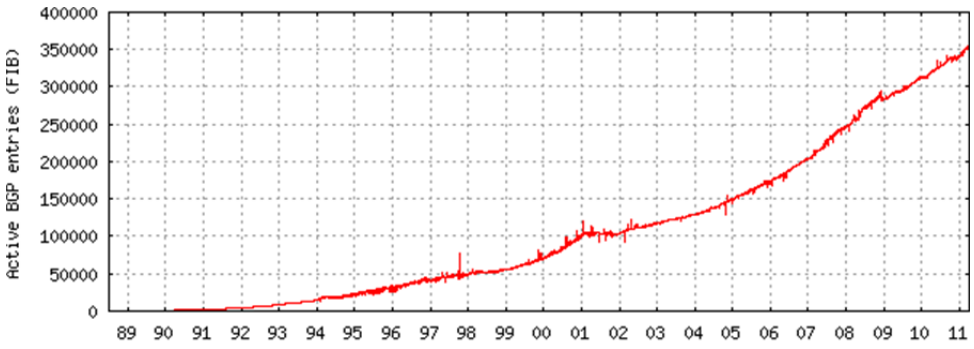


Figure 5: Number of Address Blocks, 1989 to Apr-2011 – Source: CIDR Report

In Figure 5 we see that in the last 22 years the number of address blocks in the Internet has grown from more or less zero to some 355,000 (April 2011). In the boom years of 1997-2001 the growth was about 25% per annum, compound, reaching ~100,000 at the end of 2001. From the beginning of 2002 to the end of 2008 the growth was approximately 16% per annum, compound. In the last couple of years this appears to have slowed to about 8% per annum, compound. The number of address blocks is (rather roughly) related to the number of addresses in use, and hence the number of machines connected by the Internet.

Total traffic is difficult to establish. The Minnesota Internet Traffic Studies (MINTS) web site [**12**] provides some figures for traffic growth for 2003-2009. Using their figures, we get:

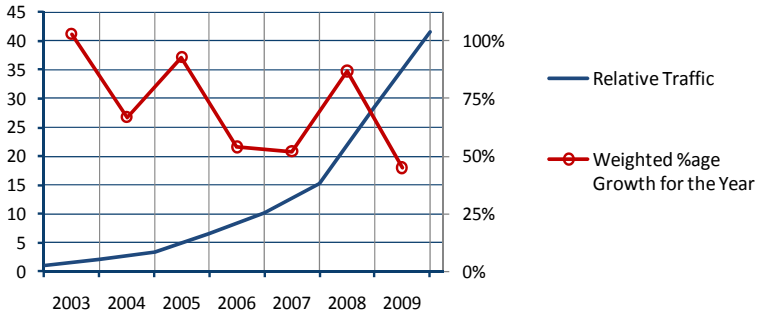


Figure 6: Internet Traffic Growth 2003-2009 – Source: MINTS

which shows that at the end of 2009 traffic was about 33 times greater than at the beginning of 2003 – which over that period is growth at 64% per annum. But, the percentage growth in each year is rather variable, between ~100% during 2003, ~50% in 2007, and just 45% in 2009[9]. In [**13**] the authors estimate that traffic between May 2008 and 2009 grew at ~45% – using similar methodology to MINTS, from data collected by the Arbor Networks "ATLAS Internet Observatory"[10] – but also give a separate estimate of 35%-45%.

Projections referred to in this review come from a Cisco study "Cisco Visual Networking Index" [**14**] of June-2010. In "*Table 3. Global IP Traffic 2009-2014*", a 34% CAGR (Compound Annual Growth

[9] *The MINTS data for 2009 is not as good as for earlier years (see http://www.dtc.umn.edu/mints/news/news_22.html), but their estimate for traffic growth in that year is 40%-50%.*

[10] *http://www.arbornetworks.com/en/atlas.html*

Rate) is projected for "Internet" traffic – that is "*all IP traffic that crosses an Internet backbone*". This is lower than the recent ~45%, but we use the 34% figure as the more conservative.

This is all some way from the 'well known' doubling of traffic every three or four months of the glory days. But it was never true [**15**]; in the late 1990s Internet backbone traffic roughly doubled year on year.

According to the Euro-IX[11] 2010 Annual Report [**16**], traffic across all European IXPs grew by 63% in 2010. This is higher than the general trend of 34%-45%, and appears to support the general view that traffic is shifting to the CDNs, and that CDNs connecting at IXPs to deliver traffic.

On 'networks', 'connections' and 'links'

An internet network is a logical construct, created by the routers in the network and the links between those routers. As, for example:

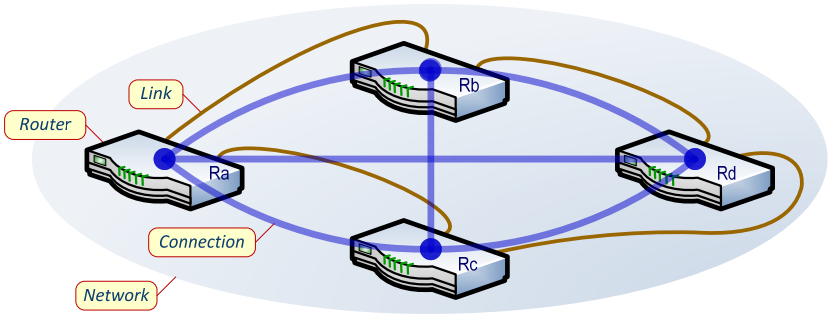


Figure 7: Network, Connection, Router and Link

where the four routers that make up the network have links, shown in brown, running between them, but not all the routers are directly linked together. The network administrator has configured the routers to establish network layer connections, shown in blue, between all the routers – so router Ra can reach Rd across the network, even though there is no link between them. The network is the set of connections between the routers, and the routers forward data across the network over the links.

When we say 'connection' we will be referring to connections at the network layer, either within or between networks.

When we say 'link' we will be referring to the links between routers (or other equipment). A link may be as simple as a glass fibre cable, or as complicated as a virtual circuit carried across any number of different sorts of network.

3.1 The Network Layer

The Internet is organised as independent 'Autonomous Systems' (ASes). Generally a network operator's network appears as a single AS, though a small number of operators use more than one.

[11] *Euro-IX is the European Internet Exchange Association: http://www.euro-ix.net/*

Each AS has a unique number, so networks in the Internet are often referred to by their AS number (ASN).

Each AS is, essentially, the home for a number of blocks of Internet addresses. The objective of every AS is:

1. to acquire a way for it to reach every other block of Internet addresses, and
2. to ensure that every other AS can acquire a way to reach its blocks of Internet addresses,

so that data can travel from the AS to everywhere else, and from everywhere else to the AS.

There are two aspects to the network layer: routing and traffic. Routing is how all the ASes acquire a way to reach each other – so we talk of 'routes' and 'reachability'. Traffic is the flow of data between ASes. Traffic flows along the routes that each AS acquires and chooses to use (where it has a choice). How well each flow of traffic is handled depends on the effectiveness of the route chosen for it. One of the key issues with the interconnection system is that the effectiveness of the available routes is not taken into account by the automatic mechanisms for choosing routes.

The complexity of the network layer is a lot to do with the sheer scale of the system: hundreds of thousands of address blocks; tens of thousands of ASes; many hundreds of thousands of direct connections between ASes; billions of indirect connections between ASes. The mechanisms within the network layer are, deliberately, as simple as possible, but they are also subtle, and interact in interesting ways. Another key issue is the difficulty of knowing in any detail what the system is doing, and the greater difficulty of predicting how it will respond to change.

Above the network layer is the operational layer. Each AS's administrators manage their network and its connections to other networks. In the following discussion, we will frequently touch on how the AS's administrators can and do configure their routers to influence the behaviour of the network layer. A further key issue is that, at the network layer, each AS's administrators have limited means to influence how other ASes handle traffic to and from it.

3.1.1 Autonomous Systems and Blocks of Internet Addresses

The function of the Internet is to carry packets of data between independent networks. To do that, those networks all use the Internet Protocol (IP). An IP Address is used to identify the destination of a packet, so it is important that the address is unique across the entire Internet. There are two versions of IP, IPv4 and IPv6. In IPv4, addresses are 32 bit values, which limits the number of addresses to about four billion. In IPv6, addresses are 128-bit values, giving an effectively unlimited number of addresses.

Originally an IP address had two parts to it: the first part identified the network (the 'network number') and the second part identified something within the network. This scheme divided all the available IPv4 addresses (the address space) into a number of fixed size blocks of addresses, and each network received one of these blocks with its built-in network number. When there were a dozen networks, and a few dozen more were envisaged, it was possible to allocate address blocks containing millions of addresses – an apparently inexhaustible number. However, as the Internet grew it was obvious that this scheme would run out of unallocated address blocks. So, the scheme

was changed so that addresses were allocated in relatively small blocks as and when needed. This meant that a network could end up using IP addresses with different 'network number's[12].

An Autonomous System (AS) Number uniquely identifies an independent network – since the 'network number' part of an IP Address no longer does. The 'Regional Internet Registries'[13] (RIRs) allocate blocks of IP addresses on request to ASes. They also allocate AS numbers. The Internet Assigned Numbers Authority (IANA) manages the IP address and AS number space, and allocates blocks of those to the RIRs as required[14].

The basic organisation of the Internet is a number of ASes, each of which uses a collection of unique blocks of IP addresses. So the Internet is roughly like this:

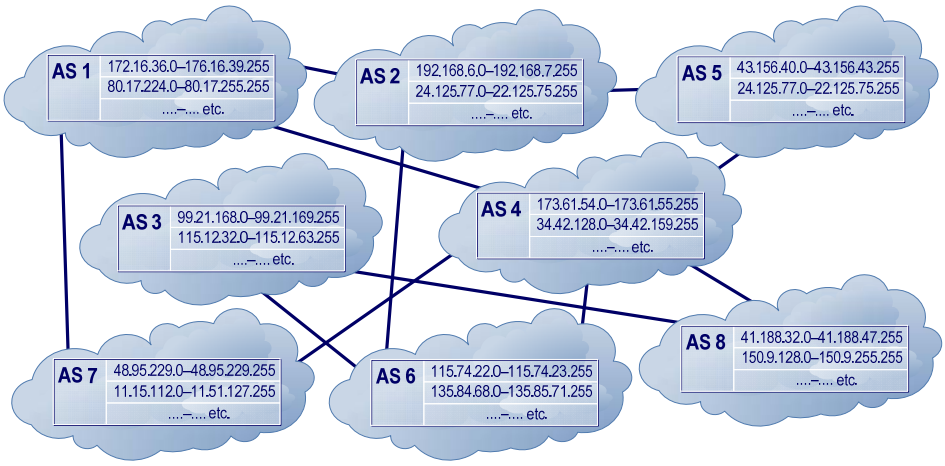


Figure 8: The Internet – Basic Organisation

where this shows a few ASes, and for each one the first couple of IP address blocks the AS uses. AS1, for example, needs all other ASes to be able to send packets to its address blocks (172.16.36.0–172.16.39.255, 80.17.224.0–80.17.255.255, etc.) and needs to be able to send packets to all other address blocks in all the other ASes.

Where an AS uses a given IP address block the AS is known as the 'origin' for the block and the addresses it contains. The collection of address blocks that an AS originates are also known as that AS's 'own' address blocks. The AS will allocate addresses from its own address blocks for its own use (for routers and other equipment that make up the network), and also for its users to use. Some of an AS's users will be paying customers, but the addresses they use are still referred to as part of the AS's own addresses. (Later we will see that some customers of an AS can have their own addresses,

[12] *The original IP address scheme had three classes of network, A, B and C. This division of address space was scrapped in favour of 'Classless Inter-Domain Routing' in 1993. (Domain is alternative name for an AS.) See RFCs 1517 [**216**], 1518 [**217**], 1519 [**218**] (now 4632 [**219**]) and 1520 [**232**].*

[13] *AfriNIC in Africa, APNIC in the Asia-Pacific region, ARIN in North America, LACNIC in Latin American and the Caribbean, and RIPE in Europe. These bodies allocate addresses to ensure that a given address is not allocated twice. They also apply some policies to avoid wasting IPv4 address space, and to have some consistency in IPv6 space.*

[14] *The last available blocks of IPv4 space were allocated by IANA on 3rd February 2011. The policy issues around the management of IP address space are outside the scope of this report. However, now that there are no more free IPv4 addresses, the management of those addresses may move from managing how they are allocated to managing a possible market in the existing addresses [**241**].*

and those are known as 'customer addresses', to distinguish them from the AS's own addresses. Customers using some of the AS's own addresses will be referred to as 'direct customers' where the distinction matters.)

Some other related pieces of jargon are worth covering at this point. A 'route' is a way to reach a given block of addresses. Each AS needs (at least) one route for every block of addresses used by every other AS. When we say AS in this context, what we really mean is the 'routers' that make up the AS. A router has two related functions, 'routing' and 'forwarding'. Routing is the process of exchanging and using the information required to acquire (or learn) and distribute routes – which is done by talking to other routers, using various routing protocols. Forwarding is the process of sending a packet towards its destination (using the routes learned).

3.1.2 What the Network Mechanisms Guarantee – Nothing

The basic mechanism for carrying data across the Internet is the Internet Protocol (IP). When an IP packet is given to the Internet for delivery to some address, the basic mechanisms guarantee nothing. In particular, it is not guaranteed that a packet will arrive:

- at all;
- in one piece;
- in a timely fashion;
- before a later packet or after an earlier one.

Furthermore, it is not guaranteed that a packet will:

- arrive at its intended destination;
- travel by any given path across the network;
- not be duplicated at any point along the path.

And when a packet arrives, it is not guaranteed to:

- be from the address it says it is from;
- contain the data originally sent.

This is not because the designers of the system were stupid, but because making the network guarantee any or all of these would have added too much cost or complexity.

Most of the time the network does deliver packets reliably and in a timely fashion. This is partly because the network is as simple as it can be, and partly because, simple though it is, a lot of work goes into keeping it running. However, the basic mechanisms are not designed to deal with the extreme cases; they are not designed to provide that last '1%' which would mean complete reliability or perfect service.

3.1.3 The Distribution and Use of Routing Information – BGP

Each AS directly connects to one or more others. At each end of an inter-AS connection is a router, and those routers use the Border Gateway Protocol (BGP) to communicate with each other. In a BGP conversation each AS announces to the other that it can reach some blocks of Internet addresses, and gives an indication of how they are reachable.

If AS4321 is the origin for the block of addresses 10.0.0.0–10.0.0.255, then there could be a chain of connections as shown:

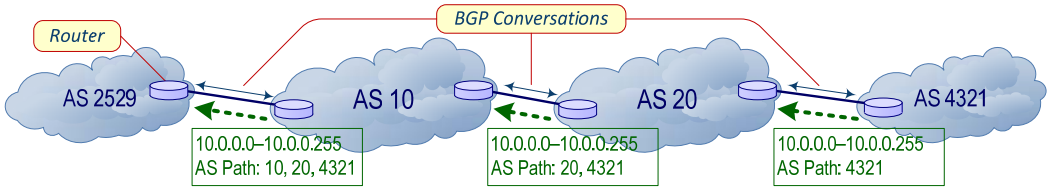


Figure 9: Announcing and Learning a Route

in which AS2529 learns a route to 10.0.0.0–10.0.0.255 from AS10, which it learned from AS20, to whom it was announced by AS4321. The act of announcing a route makes known the existence of the block of addresses, and promises that traffic to that destination can be sent this way. So in this example, AS10 is promising AS2529 that it can and will carry traffic towards 10.0.0.0–10.0.0.255 – and that promise is based on the promise from AS20, and so on.

Carrying traffic has to be paid for, and we will later look at why ASes choose to connect, what routes they will choose to announce to each other, and hence what traffic the connection may carry. As far as the underlying BGP mechanics are concerned, it is entirely up to the administrator of an AS what it announces to whom. This is one of the reasons that the resilience of BGP is complex: the availability of routes is a matter of policy as well as technology, so both economics and regulation can get in the way.

Note that part of the information that BGP carries for each route is the 'AS Path'. As shown in Figure 9, when one AS announces a route to another, it adds its own AS number to the AS Path, placing it at the front ('prepending' it). This gives, for the route in question, the chain of ASes that packets will pass through on their way to the destination – in the order that packets to the destination will encounter them. The last AS number in the AS Path is the origin of the route and the home of the address block. Analysing the AS Paths in collections of routes may be used to discover which ASes are connected to each other.

The AS Path is a vital part of a route. Its primary function is to prevent 'routing loops' – whenever an AS receives a route whose AS Path contains its own AS number, it discards it, because the route passes through itself; if it were to use it then a routing loop would be created. Its secondary function is as a measure how good a route is; the shorter the AS Path the better. The way in which each AS adds its number to the AS Path, as routes are passed from one AS to another, is specified in the BGP standard [**17**]. However, the AS Path need bear little or no relation to the actual path[15]; in particular, when packets are forwarded there is no mechanism to check that the path followed passes through the expected ASes.

[15] *Indeed, apart from the danger of routing loops, nothing will go wrong if a BGP router replaces the AS Path in route by a work of fiction.*

From AS2529's perspective, it would be wise to have at least one alternative way to reach all destinations. So, suppose both AS2529 and AS4321 are also connected to AS1, thus:

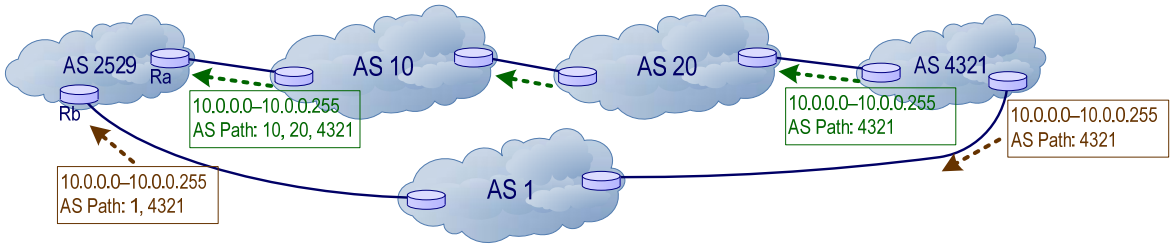


Figure 10: Learning an Alternative Route

Now AS2529 has two ways to reach 10.0.0.0-10.0.0.255, one route known to router Ra and the other known to router Rb. Where it has more than one route for a destination, a router must select and then use just one.

To illustrate how an AS uses the routes it learns, Figure 11 shows four of AS2529's routers which are connected and distribute routes to each other. Routers Ra and Rb also connect to routers in other ASes, so are generally known as 'Border Routers'. Routers Rx and Ry are internal to the AS, and learn how to reach the outside world from the border routers. The figure shows the routes that each router has learned for the address block 10.0.0.0–10.0.0.255, and from whom each was learned. It also shows how each router has ranked the available routes, and has selected the one it considers to be the best:

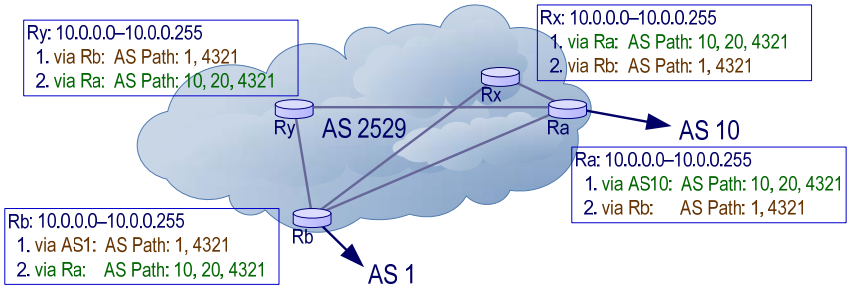


Figure 11: Selection of Routes within an AS

Router Ra chooses the route via AS10, because that is the quickest way to get traffic out of AS2529 and on its way to its destination. Similarly Router Rb chooses the route via AS1. Routers Ra and Rb tell all the others in AS2529 the routes they have selected. So, Ra and Rb are both aware of the two available routes, but in this scenario neither changes their preferred route. Routers Rx and Ry learn a different route from each of Ra and Rb, and each must choose one. In this scenario, Rx and Ry choose the route which takes traffic the shortest distance across AS2529.

The AS's border routers speak BGP to each other and exchange routing information with the other ASes they are connected to. No other protocol is used – the Internet interconnection system is, essentially, BGP. BGP is usually also used within an AS to exchange information about routes to everywhere outside the AS. There are small differences between BGP used externally and internally, and where it matters the two variants of BGP are called eBGP (external BGP) and iBGP (internal BGP). (In the jargon, an AS may also be referred to as a 'domain'. So that eBGP is spoken 'inter-domain', and iBGP 'intra-domain'.)

Note that there are two important pieces of information missing from BGP:

1. there is no information about the capacity of each route[16].
2. there is no information about the quality of each route. A route which has fewer ASes in the AS Path (a shorter AS Path) may be implicitly less tenuous, but there is certainly no explicit information about physical path length or any other quality metric.

The administrator of each AS must find other means to arrange for effective routes for every destination (every address block). With hundreds of thousands of distinct address blocks in the Internet it is not really feasible to attempt to manage the route to each one actively.

The example above illustrates how each router in an AS makes an independent decision on which route to select for each address block, and those decisions depend not only on the information distributed by BGP but also on information distributed by other routing protocols. Recall that Rx's decision between the route via Ra and the one via Rb was based on Ra being the quicker way out – that information is likely to be provided by the routing protocol(s) the AS uses internally (its intra-domain routing protocols). So the complexity of managing routes is multiplied up by the number of routers involved.

The administrator's task is further complicated by commercial considerations. BGP offers many ways in which a router can be configured to implement administrative 'policy' – to the extent that, as observed in [**18**], BGP is *"…a protocol weighed down with a huge number of mechanisms that can overlap and conflict in various unpredictable ways."* Network administrators must configure all their routers to implement a consistent set of policies across their network.

In our example, AS2529 has chosen to select routes which use the quickest path out, to minimise its internal network cost, which means that Rx and Ra are selecting a route with a longer AS Path. The administrator could decide to select for shortest AS Path – the default option – in which case all the routers would select the route via AS1. Alternatively, AS2529 might choose to avoid paths via AS1 on cost grounds. There are an indefinite number of policies an AS might wish to implement, and many mechanisms with which to attempt to implement those.

The complexity of the problem is summarised in [**19**], *"…because BGP route selection is distributed, indirectly controlled by configurable policies, and influenced by complex interaction with intra-domain routing protocols, operators cannot predict how a particular BGP configuration would behave in practice."*

The saving grace is that much of the time tomorrow's traffic patterns are similar to yesterday's (see [**20**], which examines the day-to-day effects of routing changes, and observes that the effects are small) so a configuration that works can evolve over time, possibly by trial and error. The corollary, however, is that even when everything is working well, the complexity of the system means that operators have a limited understanding of the routing decisions their routers are making, and effectively no way of predicting the effect of any change in the routes it learns from other ASes.

[16] *BGP has enough to do without trying to carry this information. The available capacity on a given link changes constantly. Trying to adjust routing to switch to better paths has a tendency to oscillate [**239**], but is a topic of research [**242**]*

We have looked at how AS2529 reaches AS4321. For that to be useful, AS4321 must also be able to reach AS2529. This is not something that AS2529 can arrange, but it arises naturally because AS4321 (in common with all other ASes) must acquire a route to every possible Internet destination, which will include those in AS2529. A conversation between a computer in AS2529 and a computer in AS4321 may appear symmetrical, but in fact the path in one direction is entirely independent of the path in the reverse direction.

When two routers speak BGP to each other the conversation is known as a 'BGP Session'. When a BGP session starts the two routers announce to each other all of the routes they know and that their administrators allow them to announce. Thereafter, further routes will be announced when they become available, and previously announced routes will be 'withdrawn' when they become unavailable. If something goes wrong, the BGP session will 'drop', and that implicitly withdraws all the routes the two ends have learned.

3.1.4 Rerouting – Adjusting to Changes

The interaction between rerouting and resilience is important. Using the same illustration as Figure 10, suppose the link between AS2529 and AS1 fails:

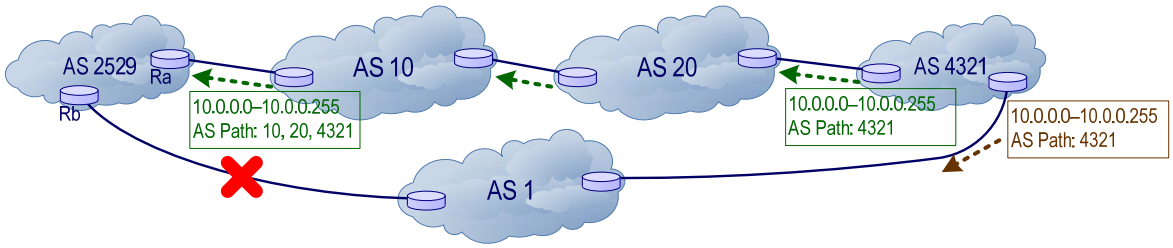


Figure 12: Rerouting in the Event of a Failure

then the route via AS1 will be withdrawn by Router Rb – it will tell all the routers in AS2529 that the route no longer exists, and each router will have to reconsider their selection of route for 10.0.0.0–10.0.0.255.

This is all as it should be, but there is no real way of knowing in advance how traffic patterns will respond, or whether there is adequate capacity on the routes that will now be selected. As the number of connections to other ASes increases, so the combinations of alternative routes increase, and the problem becomes still less tractable.

In this simple example, after the failure of a connection it is not party to, AS10 will end up carrying all the traffic between AS2529 and AS4321. It is an open question whether AS10 will have the necessary spare capacity. Certainly there is no information available to AS10 to tell it what spare capacity it needs to deal with possible network failures in other networks. The same is true for AS20 and for all the routers and connections on the path between AS2529 and AS4321 which are also now being given extra traffic. If AS2529 has other routes for destinations in AS4321, then traffic which was going via AS10 will be spread in some way across those routes. This may or may not spread the extra traffic more thinly – it is essentially impossible to predict where the traffic will go or whether there will be sufficient capacity to accommodate it.

Furthermore, the effect of the failure on packets from AS4321 to AS2529 may be quite different from the effect on packets going the other way.

It may come as a surprise, but it is nevertheless true, that:

1. the basic mechanism for routing in the Internet interconnection system, BGP, will adjust to changes in the network to maintain the ability of networks to reach each other, but is no help when it comes to maintaining the required capacity or quality of connections between networks.
2. because each router makes an independent decision, and hides alternative paths, it is almost always impossible to predict the effect of a given change.

The illustrations given above are extremely simple. As more connections and routers are added, the combinations and permutations build up quickly; the system becomes steadily more complicated and less tractable. Many properties of a network are related to the square of its size – doubling the number of objects makes the network four times as complicated. So, compared to our toy example with 4 routers, a small network with 32 routers is 64 times as complicated; a larger network with 128 routers is over 1,000 times as complicated.

3.1.5 The 'Global Routing Table'

Each AS is the origin for one or more blocks of Internet addresses, and must also acquire at least one route for every block of Internet addresses. A complete set of routes is known as a 'Global Routing Table'. Currently the global routing table contains some 340,000 address blocks from some 36,000 ASes.

The function of BGP is to distribute routing information so that every AS can have a complete global routing table. Although it is sometimes referred to in the singular, it is important to remember that every router that speaks BGP has its own version of the global routing table. All BGP routers which have a complete set of routes will have the same collection of Internet address blocks as each other (generally speaking), but each router has its own particular collection of routes for those address blocks.

To illustrate this, consider a trivial internet with just four ASes, each of which is the origin of a single block of addresses:

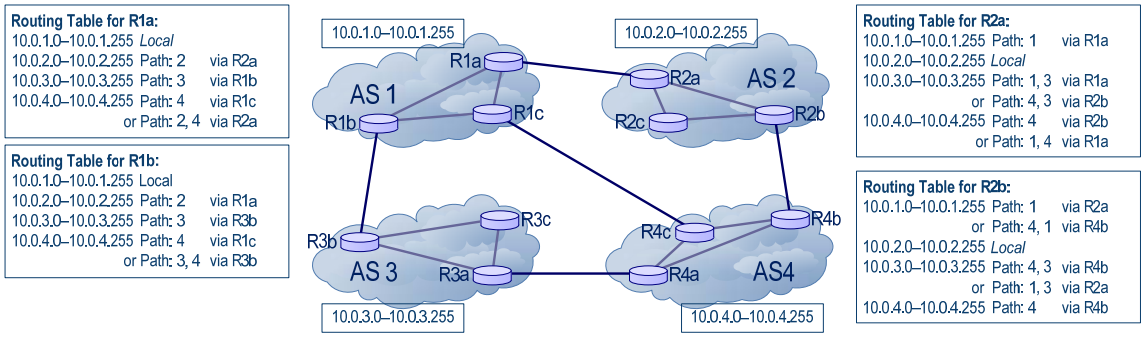


Figure 13: 'Global Routing Table'

The diagram shows the Global Routing Table as seen by four of the routers. Each has an entry for all blocks of addresses in the Internet, but the routes are different. The tables also show that for some blocks of addresses the routers have alternative, less preferable paths.

Note that in this example all ASes are announcing all the routes they have to each other. This means that all the ASes will carry each others' traffic across themselves; for example, AS1 will carry traffic

from AS2 across itself to both AS3 and AS4 and back again. We will see in Section 3.4.5 below that this is called a 'mutual transit' arrangement, and is very rare in the real world. However artificial, this simple example does illustrate the basic mechanics.

This trivial internet illustrates some of the issues discussed above:

- routers R2a and R2b use different routes for 10.0.3.0–10.0.3.255 (in AS3). This illustrates the effect of each router making separate decisions, even routers within an AS. Each router in AS2 which connects to R2a and R2b will choose to use one of the two routes they offer, probably by choosing the closer of R2a and R2b – so R2c would probably choose the route via R2a.
- router R1a could reach 10.0.4.0–10.0.4.255 (in AS4) via R1b, R3b and so on, but because R1b has chosen the route via R1c, it cannot tell R1a about the alternative path. This illustrates the effect of information hiding properties of BGP.
- router R1a has announced to R2a that it can reach 10.0.3.0–10.0.3.255 (in AS3), and R2a has decided that is its best way of reaching that block of addresses, so R2a will forward any packets for any of those addresses to R1a. This illustrates the dual nature of an announcement: it both carries information about how to reach the address block, and is an undertaking to carry traffic towards it.

Looking at what happens if the link between R1b and R3b fails is also instructive. Before the link fails the routing table entries for 10.0.3.0–10.0.3.255 (in AS1) are:

Routing Table for R1a (fragment):	Routing Table for R1b (fragment):	Routing Table for R1c (fragment):
10.0.3.0–10.0.3.255 Path: 3 via R1b	10.0.3.0–10.0.3.255 Path: 3 via R3b	10.0.3.0–10.0.3.255 Path: 3 via R1b or Path: 4, 3 via R4c

When the link to R3b fails, router R1b can no longer reach 10.0.3.0–10.0.3.255, so any packets that it receives for those addresses will be lost because they are forwarded to a link that is not working, or discarded because the link is known not to work. Also, because R1a and R1b are sending packets for 10.0.3.0–10.0.3.255 to R1b, they will be lost. So the position is:

Routing Table for R1a (fragment):	Routing Table for R1b (fragment):	Routing Table for R1c (fragment):
10.0.3.0–10.0.3.255 Path: 3 via R1b	10.0.3.0–10.0.3.255 xxx xxx	10.0.3.0–10.0.3.255 Path: 3 via R1b or Path: 4, 3 via R4c

Sooner or later router R1b will realise that the link is down, and tell R1a and R1c the bad news. How long it will take for R1b to realise depends on the type of link and the capabilities of the router, but generally the worst case will be 90 seconds. When the news reaches them, R1a will no longer have a route for 10.0.3.0–10.0.3.255, but R1c is able to start using the alternative route it has available. Now packets that reach R1c for 10.0.3.0–10.0.3.255 have a working route, but those that reach R1a or R1b still have nowhere to go. So the state is:

Routing Table for R1a (fragment):	Routing Table for R1b (fragment):	Routing Table for R1c (fragment):
10.0.3.0–10.0.3.255 xxx xxx	10.0.3.0–10.0.3.255 xxx xxx	10.0.3.0–10.0.3.255 Path: 4, 3 via R4c

Finally, R1c tells both R1a and R1b the good news, so all routers again have a working route for 10.0.3.0–10.0.3.255, so:

Routing Table for R1a (fragment):	Routing Table for R1b (fragment):	Routing Table for R1c (fragment):
10.0.3.0–10.0.3.255 Path: 4, 3 via R1c	10.0.3.0–10.0.3.255 Path: 4, 3 via R1c	10.0.3.0–10.0.3.255 Path: 4, 3 via R4c

In general, the process of withdrawing and announcing routes continues until all routers have selected a route they are satisfied with – that is, the BGP routing system has 'converged'. In this case it only took two or three steps, so that once the link is known to have failed it would not take long to adjust to the failure. In a more complicated system it can take many, many steps to achieve convergence.

Furthermore, it has been found to improve the overall stability of the BGP routing system if successive route changes for a given address block are deliberately delayed. So, when a router has announced a route for a given block of addresses, it will not announce a different route for at least 30 seconds[17].

This touches on the ability of the BGP mesh – the vast network of interconnected BGP routers across the entire Internet – to converge. In practice it does, but this is not assured [**21**]. [**22**] discusses this, with particular reference to deliberately delaying repeated route changes. Understanding how well the BGP mesh may behave when challenged [**23**] [**24**] [**25**], and improving its behaviour remains a research topic [**26**]. In [**27**] the authors investigate the effect of transient route changes on the BGP mesh, and observe *"The convergence time of the interdomain routing protocol, BGP, can last as long as 30 minutes. Yet, routing behaviour during BGP route convergence is poorly understood."* In [**28**] the authors observed that small routing changes can create multiple 20 second bursts of packet loss; [**29**] notes some routing changes causing loss of connection for more than 300 seconds; and [**30**] report an average path failover time of 3 minutes, and some up to 15 minutes. The time taken by the BGP mesh to adjust to routing changes has an impact on VoIP, as discussed in [**31**] (while [**32**] discusses the impact of link failures and rerouting within a large AS).

3.1.6 Policy and Route Announcements

In our example, trivial internet, some packets between AS2 and AS3 are travelling across AS1. In the jargon AS1 is providing some 'transit'.

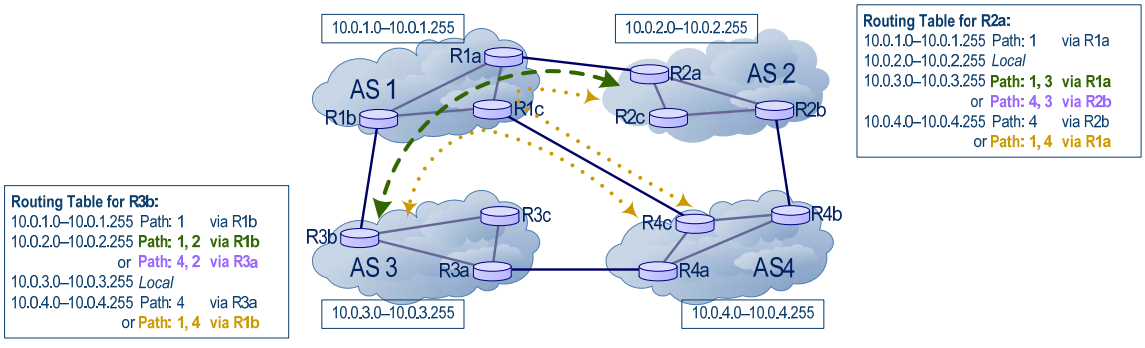


Figure 14: Trivial Internet – Complete Cooperation

AS2 and AS3 benefit from this arrangement, because their users can reach each other, even though there is no direct connection between their networks (as shown by the green, dashed line). AS1 is also providing alternative paths (shown by the dotted lines) between AS2 and AS4 and between AS3

[17] *The 30 second delay applies to announcements of routes between ASes, within an AS the delay is 5 seconds – or at least those are the recommended delays, network administrators may set other delays where they feel that improves things. This is the 'Minimum Route Announcement Interval' (MRAI).*

and AS4, which can be used if a link fails. AS1 does not directly benefit from this arrangement, indeed it probably incurs some cost carrying this traffic. (AS4 is providing a similar service, but that is not shown on the diagram in the interests of simplicity.)

In this example we have assumed that all the ASes announce all their routes to each other. So they are providing transit to each other on a mutual basis, though AS1 and AS4 will only make use of the facility if the R1c–R4c link fails. Traffic can transit AS1 between AS2 and AS3 because its R1a has announced 10.0.3.0–10.0.3.255 (AS3's addresses) to R2a, and R1b has announced 10.0.2.0–10.0.2.255 (AS2's addresses) to R3b – that is all there is to it.

Policy decisions also impact resilience. Let us suppose that AS1 decides it no longer wishes to carry traffic that does not start or terminate in AS1. To implement this change of 'policy' it simply changes its router configuration to only announce its own addresses to the other ASes. The position would then be:

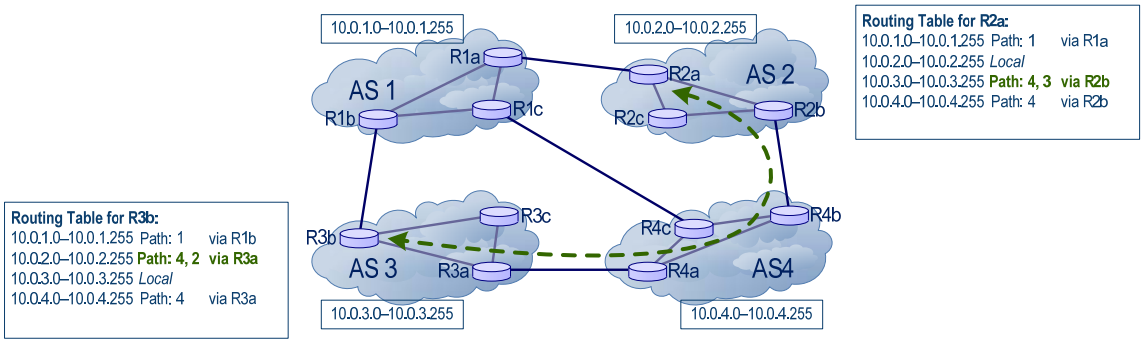


Figure 15: Trivial Internet – Incomplete Cooperation

Now AS2 and AS3 can only reach each other via AS4, so if either of the links R2b–R4b or R3a–R4a fail, then they would be disconnected from each other, even though there would be a viable connection via AS1 – because AS1, as a matter of policy has decided not to carry their transit traffic. Similarly AS4 can only reach AS2 via R2b–R4b, and AS3 via R3a–R4a.

Note that AS1's policy is implemented by changes to what it announces to whom. Note also that a change of policy does not require cooperation from the other ASes – indeed they may be unpleasantly surprised to find that all traffic between AS2 and AS3 is now going via AS4. How well all the links along that path will cope with the extra traffic is a separate issue. The system is obviously less resilient, but from AS1's perspective, it is no longer carrying the cost of traffic between AS2 and AS3. In this trivial example it is clear to AS1 that it has adversely affected the system, in the rather larger, real world it may not be at all obvious. Since AS1 has changed its policy, the other ASes may reconsider theirs. If they do nothing, they will all continue to provide alternative paths for AS1 in the event of connections to AS1 failing, which does benefit each of them as well as AS1, though AS1 no longer contributes.

For AS2 and AS3 the key questions are: will AS4 continue to be willing to provide transit, and should they do something to make the connection between themselves more resilient?

Even in a trivial illustration like this there are interesting combinations and permutations of policy. Policies are implemented independently by each AS – each independent Internet company – by configuring their routers to announce some routes but not others, and possibly announce different sets of routes to different ASes. While policies are implemented independently, they will be

influenced by the need to cooperate with other ASes so that each can reach all the others, and may be influenced by a desire for some level of resilience.

3.1.7 Information Hiding

The purpose of BGP is to distribute information about routes and to exchange undertakings to forward packets. In the process, however, it also hides information.

As described above, for each destination:

1. each router may learn a number of routes, but will select one for use and only advertise the one it selects. So, while BGP is distributing information between routers and ASes, it is also hiding all the information about routes that are available, but not currently selected – each router only advertises the route it has selected and will use to forward packets along.
2. different routers make their own selection, and the route selected may be different in different parts of the AS.

Among many other complexities, these two issues alone make it impossible (or at least very difficult) to discover the topology of the Internet and map how ASes are interconnected. See [**33**], which discusses what may be achieved.

One method commonly used to attempt to map the connections between ASes is to examine the AS Paths in data collected by 'Route Collectors'. A route collector will connect to border routers in a number of ASes, and collect from each one the routes it has selected – so the routes collected are only a partial view. Indeed, if a route collector connected to a different router in a given AS it could well receive a different partial view. Hence the maps created tend to be incomplete, even if the results from a number of route collectors are combined. There are two sets of route collectors which have good, publicly available data: the University of Oregon Route Views Project[18], which has data from 2001 onwards, and the RIPE RIS Project[19], which has data from 1999 onwards. Most studies which use route collector data use these data sets, the 'Cyclops' AS-level connectivity laboratory [**34**], for example. 'BGP Beacons' [**35**] are used in conjuction with route collectors to measure the propagation of route changes across the system.

Another method uses traceroute probes from many places in the Internet, and examines the results to determine where packets are passed between networks. The traces can also provide some information about the length and quality of paths. However, within an AS there may be multiple active routes to the same destination, so a traceroute probe's result will depend on where it starts in the AS. Moreover, within an AS there may be any number of inactive routes, which are hidden from traceroute probes. A further problem with traceroute probing is that it can say something about the path between *a* and *b*, but the path from *b* to *a* is independent and probably different.

The 'The Cooperative Association for Internet Data Analysis'[20] (CAIDA) ran the 'skitter'[21] data collection system for ten years up to February 2008, and now runs the 'Archipelago Measurement

[18] *http://www.routeviews.org/*
[19] *http://www.ripe.net/ris/*
[20] *http://www.caida.org/home/about/*
[21] *http://www.caida.org/tools/measurement/skitter/*

Infrastructure'[22] (Ark for short). These use traceroute probing to collect data on the paths across the Internet seen from a range of locations. The 'DIMES Project'[23] uses traceroute probes from volunteers' machines, thereby hoping to get widespread coverage of the Internet, and one of its applications is described in [**36**]. In [**37**] the authors describe the care required to obtain good data using traceroute probes. In [**38**] the limitations of what can be discovered from the available data are described, and the paper concludes: *"…we demonstrated the infeasibility to obtain a complete AS-level topology through the current data collection efforts…"*. [**39**] describes the findings of a workshop on Internet topology, one of whose observations is that the *"lack of comprehensive and high-quality topological and traffic data represents a serious obstacle…"*.

3.1.8 Traffic Engineering – Making the Best of What is Available

As we have seen, in 3.1.3 above, each router in an AS must choose which route to use to send packets to a given destination. However, as noted above, it is not practical to make detailed decisions for all possible destinations. Furthermore, once packets have left the AS they are entirely in the hands of other ASes, who will make their own decisions about forwarding.

The path for packets coming from other ASes is determined by decisions made by the AS at the far end, and all the ASes in between. The receiving AS has no direct say in the matter, but it can attempt to influence other ASes' decisions. The length of the AS Path is a standard mechanism to use when ranking routes, and BGP allows the AS Path to be padded to increase its length, which is conventionally done by adding the origin AS number one or more times.

Adapting the illustration in Figure 10, let us suppose that, AS4321 would prefer traffic coming to it to not arrive via AS1, perhaps for cost reasons. In an attempt to do this it can pad the AS Path in routes announced to AS1, as shown:

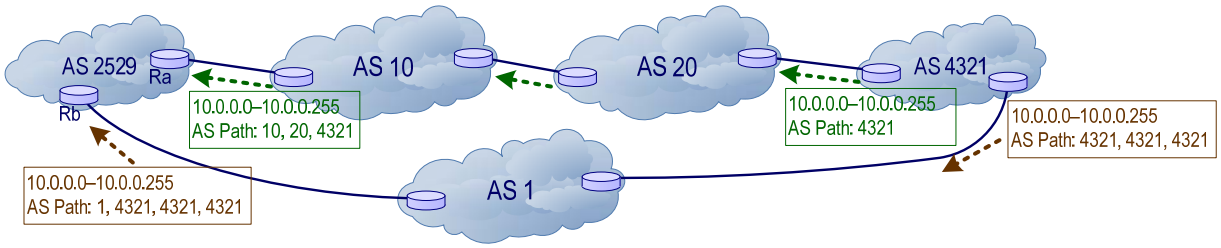


Figure 16: AS Path 'Padding'

Assuming that AS2529 does the conventional thing and takes notice of the AS Path length, then AS4321 will have achieved its goal, at least for traffic from AS2529.

Unfortunately, this is the only tool that an AS has to try to influence the path that packets take en route to it, but it is not a precise tool for a number of reasons:

a. it affects all ASes that see more than one route to the AS – it is not possible to influence traffic from specific places;

[22] *http://www.caida.org/projects/ark/*

[23] *http://www.netdimes.org/*

b. it only affects ASes that see more than one route to the AS – and we have seen how each BGP router only passes on some of information it has (specifically, only one route per destination address block);

c. BGP routers are not required to take notice of the AS Path length – though many do;

d. it is a blunt instrument – many AS Paths are relatively short, between 3 and 4 ASes long, so padding the path even by 2 or 3 ASes can make a significant difference.

Each AS has very limited ability to manage how its traffic makes its way across the interconnection system [**40**] [**41**] [**42**]. Most of the time network administrators adjust the capacity of their interconnections to fit the way that traffic is naturally distributed across the available routes, rather than trying to push traffic around to make it fit the capacity they have. That works because, in bulk, traffic is reasonably stable, so once a working arrangement of interconnections has been achieved, maintaining capacity is an incremental process.

If there is a sudden change in the available routes, and traffic redistributes itself onto paths which do not have the required capacity – creating congestion – network operators have limited means to try to make the traffic fit the available capacity, and it may be a while before capacity can be adjusted. And it is hard to know what to prepare for because it is hard to predict how the distribution of traffic might be affected by a given failure, and there are many, many possible failures.

3.1.9 Deaggregation – the Unacceptable Face of Traffic Engineering

One particular form of traffic engineering is worth looking at because of the effect it has on the interconnection system.

Suppose AS64500 is the home for the address block 10.0.0.0–10.0.1.255, and connects to AS10 and AS20. The usual thing would be for AS64500 to announce this block of addresses to all the ASes it connects to, but in this case it splits the address block in two, and announces both halves (10.0.00–10.0.0.255 and 10.0.1.0–10.0.1.255):

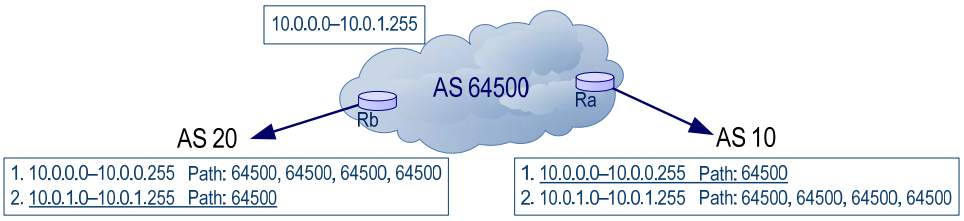


Figure 17: Deaggregation

where the AS Paths are padded by repeating the AS number, so that ASes that have a choice of routes to 10.0.0.0–10.0.0.255 will tend to use the path which reaches AS64500 via AS10. Similarly, packets for 10.0.1.0–10.0.1.255 will tend to use the path via AS20. Subject to the significant limitations discussed above (Section 3.1.8), this will control how incoming packets arrive at AS64500. AS64500 has complete control over how packets leave, so some degree of traffic engineering is achieved. Note that AS64500 is here announcing two address blocks to the world, where it could announce just the one. Splitting an address block in this way is known as 'deaggregation', and is generally frowned upon, as will be explained shortly.

As far as AS64500 is concerned, there are significant private benefits. This AS may wish to (roughly) balance its traffic across AS10 and AS20, which deaggregation will achieve if the two halves of their address block generate roughly the same amount of traffic. Alternatively, its network may be

arranged so that it saves AS64500 money to have traffic delivered in this way; perhaps its network is geographically dispersed, so for users close to router Ra it is cheaper if their traffic comes and goes via Ra.

There is another way in which AS64500 can use deaggregation to achieve their traffic engineering goals, as shown:

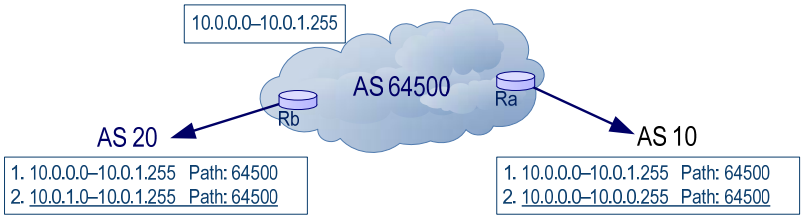


Figure 18: Deaggregation using 'More Specific' Routes

which uses the more powerful 'more specific' route mechanism. In this case AS64500 announces the entire address block, 10.0.0.0–10.0.1.255, to both AS10 and AS20 – in the usual way. It also announces one half of that address block to AS10 and the other half to AS20. Those extra announcements each contain what is known as a 'more specific' route (the underlined routes in the diagram). A more specific route refers to a block of addresses that is part of a larger block announced in another route. Wherever an AS has a choice it must use the most specific route available for any given address – this is an absolute requirement of the BGP protocol – a more specific route will always be given priority over a less specific one. So this traffic engineering method is more effective than the first, but note that AS64500 is here announcing three address blocks to the world, where it could announce just the one.

Deaggregation is frowned on because although it provides private benefit, it increases the costs of all other ASes. In particular, it increases the size of the Global Routing Table. Information about every single block of addresses that an AS announces must be transported across the entire Internet, and every single BGP router in every single AS has to process this information and store it. The local benefit to the deaggregator has created global costs. Where an AS deaggregates an address block for its own convenience, it is polluting the commons.

3.1.10 'Hot Potato Routing'

In the discussion above it was assumed that where there was a choice of route to an address outside the AS, that routers will choose the route that sends packets out of the AS as quickly as possible – which generally means forwarding it to the closest usable connection to another AS. This is known as 'hot potato routing'.

Left to its own devices, BGP will do hot potato routing. In fact, to do anything else – including so-called 'cold potato routing' – requires extra effort, and may require some measure of explicit cooperation between ASes. Traffic engineering across multiple ASes is a research topic; among the issues are: what information is required, how that would be distributed, how all the ASes involved could be satisfied that the result was 'fair' (or in their interests, at least), and so on [**43**].

Although hot potato routing affects the distribution of an AS's traffic generally [**44**], it is a particular issue when two ASes have more than one connection with each other. Consider two large networks which connect to each other in Newark NJ, Moscow and Palermo:

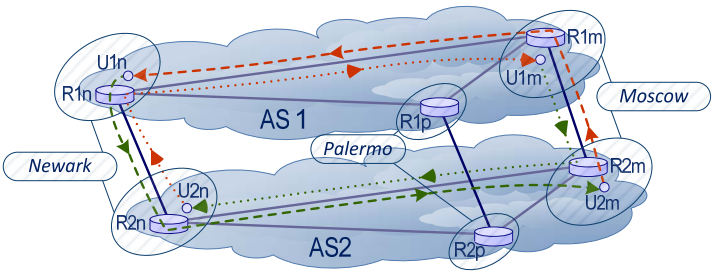


Figure 19: Hot Potato Routing

which shows the two networks AS1 and AS2, separated vertically. The connections in Newark (R1n–R2n), in Palermo (R1p–R2p) and Moscow (R1m–R2m) are local to those cities.

The diagram also shows an AS1 user in Newark, U1n, an AS2 user in Moscow, U2m, and the paths which packets will take between the two (the dashed lines). This is hot potato routing in action; AS1 sends packets destined for anywhere in AS2 to the nearest available connection with AS2, and vice-versa:

- packets from U1n to U2m go via the link R1n–R2n, and then from Newark to Moscow on AS2's network;
- packets from U2m to U1n go via the link R2m–R1m, and then from Moscow to Newark on AS1's network.

This is not entirely asymmetrical: as shown in the diagram, a conversation between U2n and U1m is the mirror image of the U1n–U2m one (as shown by the dotted lines), so that, all things being equal, the load on the two networks is the same, in both directions, between Newark and Moscow.

Not shown on the diagram are conversations U1n–U2n or U1m–U2m which are rather simpler, and do not require long-haul carriage. Note, however, that the packets from AS1 are carried by AS2's long-haul network, at AS2's cost, and vice versa.

Hot potato routing may appear peculiar, but is reasonable. When AS1 has a packet to forward to one of AS2's addresses, it does not know where in AS2's network it is destined for, so the only thing it can do is pass the packet as quickly as possible to AS2 – there is not much point carrying packets from Newark to Moscow, only to have the other network carry them all the way back. For AS1 to be able to exchange traffic at a point closer to the final destination, AS2 would have to provide a lot of information about its network, and keep it up to date, and AS1 would have to configure its routers to use this extra information. Anything other than hot potato routing requires extra work.

What has been described so far is what will happen if the two ASes announce their address blocks in the same way to each other at all connection points – which is the usual thing to do. Suppose, however that AS2 would prefer AS1 to carry more of the long-haul traffic. To do this piece of traffic engineering, AS2 could announce a more specific route from R2n to R1n, where that route included all its users who are local to Newark. Similarly it could announce a different, more specific route from R2m to R1m, where that route included all its users who are local to Moscow, and similarly in Palermo. The effect of this is shown:

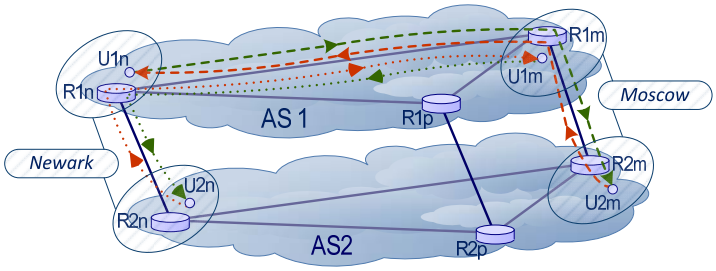


Figure 20: One Sided Hot Potato Routing

Whether AS1 would be happy with this arrangement, in which AS2 is avoiding carrying traffic on its long-haul network, is unclear. Suffice it to say that ASes which connect in more than one place often specify that 'consistent announcements' are expected – i.e. not the above.

3.1.11 BGP Insecurity and Route Filtering

In a connection between ASes, each AS decides what routes it will announce to the other, and we have seen how that affects what paths traffic may take through the resulting network. An AS is not required to use all the routes it receives – it may choose to filter out some routes.

One of the issues with BGP is that while it distributes information about routes across the entire Internet, it does not provide any means to verify that the information it carries is valid. This causes a number of problems:

a. mistakes of one sort or another can propagate across the system and disrupt it. On a number of occasions some AS has mistakenly announced that it can carry traffic to all parts of the Internet, when it cannot[24]. This diverts some traffic which disappears into a bottomless pit. The failure persists until the AS in question fixes the mistake, or other ASes add route filters to their routers to discard the mistaken announcements.

b. unused blocks of Internet addresses can be announced by ASes who have no right to use them. This 'hijacking' of addresses is usually done by people who might otherwise have difficulty obtaining legitimate addresses, so are generally up to no good. The address blocks hijacked may be from unallocated address space, or from allocated space which is not being used.

c. blocks of Internet addresses can be announced with the intention of diverting or intercepting traffic. Announcing 'more specific' routes for somebody else's address blocks is a good way of diverting traffic.

[24] *One of the earliest recorded instances of this is the now infamous AS7007 incident, covered in section 5.8.2, below.*

d. the possibility that somebody might deliberately set out to disrupt the system. Like many systems the Interconnection system is vulnerable to an "inside job"; the above tricks (and others) could be used by malicious routers that set out to disrupt the routing fabric, perhaps after being taken over by an attacker.

These problems could be mitigated if ASes routinely filtered the routes they received and rejected any that should not be announced. Unfortunately there is no easy way to establish which routes should not have been announced. Consider this fragment of the Internet:

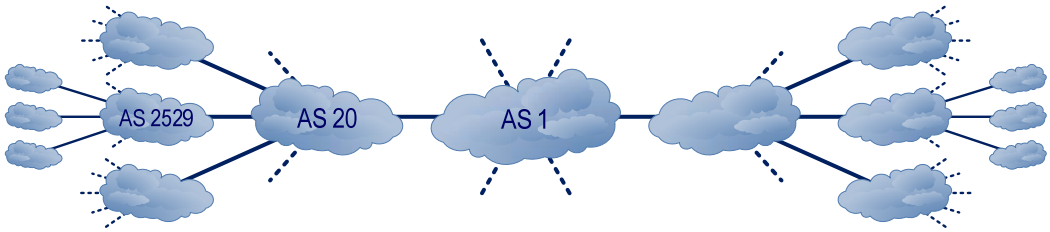


Figure 21: Route Filtering Problem

It is practical for AS2529 to filter the routes it receives from the smaller ASes it connects to, which are at the edge of the system, and only announce their own routes. For AS20 to filter the routes it receives from AS2529, it would need to know all about all the ASes which AS2529 connects to and announces routes for; and the same for all the other ASes that AS20 receives routes from. Clearly AS1 has an even less tractable problem. Looking the other way, AS20 is receiving routes from AS1 for everything it knows, which is 340,000-odd routes, gathered, first, second, third, etc. hand – AS20 has no practical means of knowing what is valid. Similarly AS2529 has no practical means of filtering the routes it receives from AS20, and so on.

More straightforwardly, when a route is received by an AS's border routers, the first AS in the AS Path should be the AS the route is coming from. Checking this, and filtering out any routes that fail the check, is a simple way of avoiding some invalid routes – or, at least, of ensuring that if an AS is passing on counterfeit routes, then it cannot do so anonymously.

3.1.12 More Secure BGP and RPKI

Any BGP router can announce any route it likes and BGP offers no means for the receiver of the route to distinguish a valid from an invalid route. There are other security issues with BGP which are covered in RFC4272 [**45**]; for a survey of BGP security issues see [**46**]. The U.S. Department of Homeland Security (DHS) has an initiative "Secure Protocols for the Routing Infrastructure (SPRI)", and the roadmap [**47**] covers the issues, including the issue of adoption and deployment of more secure protocols. The DHS is supporting the BGPSEC work (see below).

On the face of it, BGP is shockingly insecure. The willingness to trust the validity of the routes it distributes appears to leave the system open to all sorts of potential abuse. Nevertheless, BGP works well most of the time. Further, to seriously disrupt the BGP mesh would require a well placed rogue AS, or a conspiracy among medium size ASes, or a software attack that took over a number of routers. Experience with occasional configuration mistakes suggests that once an attempt to disrupt the mesh were detected, the mesh would quite quickly filter out the disruptive routes, probably by disconnecting the disruptive ASes.

There are two well known proposals for more secure forms of BGP, but neither is close to deployment:

a. 'Secure BGP' (S-BGP), dating back to 1997 [**48**] [**49**].

b. 'Secure Origin BGP' (soBGP), dating back to 2002 [**50**] [**51**]
(the last IETF draft expired in Oct-2004)

most recently there is also:

c. BGPSEC, the first IETF drafts for which were published on 7th March 2011 [**52**]

All these proposals use a Public Key Infrastructure to support digital signatures that can be applied to route announcements. When a BGP router receives a route it can check the signatures, and so verify that the route is valid. BGPSEC is based on the proposed RPKI.

The starting point for these schemes is a large table containing every valid Internet address block and the AS which is entitled to originate each one. Somehow every AS must acquire a copy of this table, and be sure of its validity. This is even trickier than it sounds, and no such verifiable table presently exists. If it did, then armed with this information an AS could configure its routers to examine the address block and the origin AS in every route it receives, and reject any that did not appear in the table. This would enable the following to be rejected:

a. attempts to hijack unused blocks of addresses.

b. attempts to announce another AS's address blocks, or parts thereof.

which would be a step forward, but leaves the problem that a bogus route for an address block could still be manufactured, and would be accepted so long as the true origin AS is given. (This falls short of the ability to deal with Cyber Warfare or State Sponsored Cyber Terrorism, in which one might suppose that numbers of ASes, or subverted routers, would conspire to disrupt the BGP mesh.)

At present the 'Resource Public Key Infrastructure' (RPKI) initiative aims, essentially, to build the table specifying which address blocks an AS is entitled to originate [**53**]. The RPKI includes a 'Repository' [**54**] which contains, amongst other things, Route Origination Authorizations (ROAs), which say that AS 'y' is entitled to originate address block 'y'. When the RPKI is implemented, ASes will be able to use it, as described above, to filter out a class of invalid routes.

The next requirement for a more secure BGP is the ability to verify that the AS Path is a valid path to the destination, without any additions or deletions. In BGPSEC, therefore, when an AS announces a route, the announcing AS authorises the receiving AS to pass on the route. This uses 'BGPSEC_Path_Signatures' attributes which link every route received back to its origin, as shown:

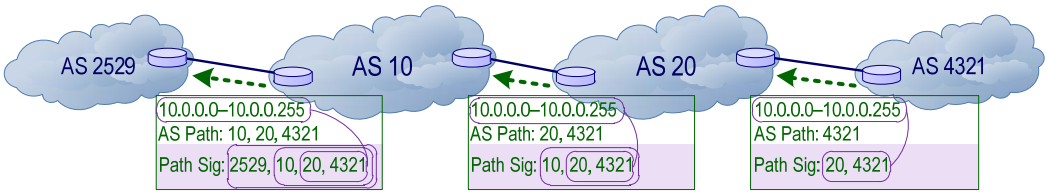


Figure 22: BGPSEC and 'BGPSEC_Path_Signatures'

where the originator, AS4321, creates a BGPSEC_Path_Signatures attribute to say that the AS to whom it is announcing the route, AS20, is authorized to use and pass on the route. The signature also covers the address. When AS20 announces the route to AS10, it adds its signature to say that AS10 is

authorised to use and pass on the route, and so on. In the diagram each layer of signature is shown, enclosing the previous one. When an AS receives a route, it can check each level of signature in turn and trace the path back to the origin. The threats that BGPSEC is designed to address are discussed in [**55**]. Using the RPKI ROAs, the receiving AS can also verify that AS4321 is entitled to originate 10.0.0.0-10.0.0.255.

The use of BGPSEC (and RPKI) protects against an AS forging a route to a given destination, or inserting itself into the path to a given destination. What it does not do is guarantee that all ASes in the AS_Path are trustworthy, or that an AS will, in fact, forward packets to the next AS in the path. From AS4321's perspective, it must trust that AS20 will only distribute its routes to trustworthy ASes, but it cannot control to whom AS20 passes its routes. Similarly, AS20 must trust AS30, and so on. Should some AS prove untrustworthy, then all other ASes could ignore routes which include an untrustworthy one.

BGPSEC is similar to S-BGP. Among the difficulties thought to exist with S-BGP is the significant extra work that each BGP router would have to do to create and verify all the RAs – indeed, some proposals to implement S-BGP involve an extra device placed next to each existing BGP router, to provide the required extra muscle. There are real concerns that BGP already has enough to deal with, so fear of the extra overhead of S-BGP has been an effective deterrent, and S-BGP has not progressed. In [**56**] the authors report between 46% and 230% increase in convergence times with S_BGP and an 11-fold increase in memory requirement.

The inventors of soBGP aimed to improve the security of BGP, but with a lot less overhead. In the soBGP scheme an AS publishes an 'AS Policy', which can specify various properties of the AS and of routes originating from the AS. An AS can publish which ASes it is connected to: so, in the example above AS4321 would declare that it is connected to AS20, so a route with an AS Path …, 20, 4321 is valid to that extent. If AS20 declares its connection with AS10, then AS2529 can validate the route in the example. An AS may also publish ASes which are not expected to appear in the AS Path for any route to the AS, which could be used to identify bogus routes created by an AS up to no good.

An advantage of soBGP is that little of the work has to be done by the BGP router. All the work to establish what can be accepted as valid can be done somewhere else within each AS. The AS's BGP routers can then be configured to filter out invalid routes, using the usual BGP facilities. This still means that the BGP router is doing more work, just not as much as with S-BGP.

The disadvantage of soBGP is that it offers rather less security than S-BGP, while still representing a great deal of extra effort. So soBGP has not progressed either.

The deployment of BGPSEC is considered in [**57**]. The distribution of the extra attributes in every BGP message, the handling of the RPKI, the checking of signatures, and so on, represent a significant overhead and may require hardware as well as software upgrades to the routers that will speak BGPSEC. It is envisaged that 'origin validation', using just RPKI, might start to be deployed in the next two to five years, and BGPSEC might start to be deployed towards the end of that time.

Other proposed schemes for securing BGP include:

a. Interdomain Route Validation (IRV) (2003) [**58**]

   This takes a rather different approach, and does not attempt to extend BGP itself or require a comprehensive table of all known addresses. The essence of the scheme is that each AS runs an IRV server which other ASes can reach, securely. When a BGP router receives a route announcement it would send a copy of it to its local (within the AS) IRV server. That IRV

server would examine the AS Path, contact the IRV servers for each AS in the path, and ask them to verify the route. This has not been implemented.

b. Secure Path Vector (SPV) (2004) [**59**]

This is strongly related to S-BGP, the principal difference being that it uses less computationally intensive cryptography, in order to reduce the overhead of validating routes. The effectiveness of the security is in doubt [**60**]. This has not been implemented.

c. Pretty Secure BGP (psBGP) (2005) [**61**]

This is similar to S-BGP, but to avoid the need for a comprehensive table of all known addresses, it lets each AS attest to its own addresses. If RPKI comes to fruition, there will be the comprehensive table. This has not been implemented.

d. Pretty Good BGP (PGBGP) (2008) [**62**]

This takes a very different approach. Observing that most routes today are the same as they were yesterday, or have been seen before, any unrecognised route may be suspect, and is put "on probation" – most route hijackings last for less than 24 hours, so the probationary period need not be a long one. This has not been implemented.

For a discussion of how secure secure routing protocols are see [**63**]. The question of how readily more secure versions of BGP might be adopted is addressed in [**64**].

3.1.13 Source Address 'Spoofing'

Every IP packet carries the destination IP address (to which the packet should be forwarded) and the source address (whence it came, and to which replies are to be sent). The destination address is vital to the delivery of the packet. The source address is not, and is there for the receiver of the packet only; it is of no interest to the network at all.

There is no good reason to send out a packet with a source address other than the actual source address, and the only reasons to do so are bad ones. Invalid or 'spoof' source addresses are used mostly in 'Denial of Service' (DoS) attacks, either because the spoofed address is part of the attack, or simply to hide the source of the attack, or to confuse attempts to deal with the attack.

The fact that the network does not look at the source address helps keep the work that the network must do to an absolute minimum. It would be nice if the network could guarantee that a packet really came from where it says it came from, but this is a function the two ends can do, and probably do better given that IP addresses are not proofs of identity of the parties.

As a packet enters the network it is possible to check that the source address is valid. As packets are passed from one AS to another the receiving AS could check that the source address of every packet is consistent with the routes that the sending AS has announced. So some filtering of packets would help, but in general, it is hard to imagine how the network could firmly guarantee the source address of every packet – see Section 5.8.4 below.

3.1.14 Quality of Service, Congestion and 'Over Provisioning'

Quality of service between two points across a network has a number of dimensions:

a. the time taken for a packet to travel between the two points. This will depend on many things. At bottom there is the physical distance between the two points, which depends on the layout

of the network, and the path taken by packets. Along a stretch of glass fibre packets travel at approximately 200km/millisecond, so by the shortest possible route a packet would take 12 milliseconds to travel from London to Athens. Each router the packet travels through must receive the complete packet, decide what to do with it, place it in a queue for sending and then send it – which will add some delay, more if the outgoing link is busy and the queue is long.

b. how consistent the time taken for each packet is. In the best case all packets will take the same time to travel from point to point. If a packet encounters congestion, then delays will be introduced as the packet waits for earlier packets to be sent along the congested link. Congestion will show up as variable packet transmission times, also known as 'jitter'.

c. how reliably packets are delivered. In the best case all packets are delivered, and in the order they were sent. If a packet encounters severe congestion it may be discarded. If the network is reorganising itself, for example in the event of failure, then the path taken by a later packet may be more effective than the path taken by an earlier one, so packets arrive out of order.

Essentially two things affect quality of service of a network, capacity and stability. An empty network that maintains consistent paths will provide the maximum possible quality of service. A network may add extra links to reduce the network distance between some points or make other changes to increase that maximum quality of service, but at any given moment, capacity and stability are key.

Stability is generally to do with the response of the network to failures. When a failure occurs, the time taken to detect and adjust to the failure, and how service is affected during the detection and adjustment phases and thereafter, all affect stability. The acme is a network that detects and adjusts to failures with no effect on service. Conversely, a very unstable network is one that is frequently affected by failures, or which takes a long time to detect and adjust to them, or both.

In the absence of failures, insufficient capacity causes congestion, which reduces the quality of service. So the management of a network is in large part the management of capacity. The demand on a network varies from day to day and week to week, but the peak demand from month to month is reasonably stable. Most networks manage their capacity so that there is a margin above the usual, long-term peak demand, to cope to some extent with the unexpected – that is, they over-provision their networks.

When failures occur, and capacity is temporarily lost, then congestion may occur and service will suffer until repairs are made. Over-provisioning plays a part in maintaining spare capacity to mitigate the effect of failure.

Over-provisioning is also a straightforward way to support delay sensitive traffic – where there is no congestion, there is no excess delay. On large backbone links the degree of overprovioning required for this purpose is modest [**65**].

Network quality of service is hard to measure. For connections between two points it can be relatively straightforward[25], and we might consider the quality of service for a network to be the average of the quality of service between all pairs of points; but that is not only intractable, but also

[25] *There is an IETF 'Framework for IP Performance Metrics (IPPM)', and a number of RFCs defining a number of metrics in that framework, including: RFC2330 [**245**], RFC2678 [**244**], RFC2679 [**246**], RFC2680 [**247**], RFC2681 [**248**], RFC3357 [**249**], RFC3393 [**250**] and RFC5136 [**251**].*

assumes that every possible connection is of equal value – though there is research in this area [**66**]. Also, averaging measures tend to hide significant issues. In short, a complete measure of network quality of service is impractical, and any practical measure is incomplete.

3.1.15 'Best Efforts' and Quality of Service

The standard of service offered by the Internet is 'best efforts'. That is, the network does its best to deliver packets, but does not offer any guarantee that packets will be delivered, or when, or that packets will arrive in the order sent, or almost anything else. This absence of guarantees makes the Internet less expensive than other kinds of network – and in fact makes it possible. And given stability and capacity, a best-efforts network performs very well. If those properties could be guaranteed, then nothing more than best efforts would be required.

Where a stronger guarantee is required, it is possible to mark packets for preferential service, and to configure routers to take notice of 'differentiated services' (or 'type of service') markings. Both of these require extra work, in particular routers will not take any notice of the markings unless told to do so. When routers do take notice, they give some priority to marked packets when there is a queue of packets waiting to be sent across a given link. In the absence of congestion, there will be no queue, so the markings make no difference; but where there is congestion, marked packets will be less affected. It is also possible to send marked packets along different routes, which may use special circuits with extra redundancy.

This is often what people mean when they talk of Quality of Service (QoS), though technically it is 'Differentiated Quality of Service', or 'DiffServ'. There are also mechanisms to support some quality of service for individual connections, for example a connection carrying a video stream, which reserve a certain amount of bandwidth along the path taken by the connection – this is known as 'IntServ'.

These QoS mechanisms are implemented within some operators' networks, particularly for customer 'Virtual Private Networks' (VPNs) and for transport of 'Voice over IP' (VoIP). In some cases operators will support these mechanisms for particular traffic between networks, notably again Virtual Private Networks.

These QoS mechanisms are not implemented in the "open Internet". The reason for that is pretty straightforward: there is no mechanism for verifying what level of service a packet is entitled to. If all ASes honoured "priority" markings, then users could so mark their important or time critical packets, and avoid those packets being affected by congestion. Unfortunately, there is no mechanism to verify that a user is using this facility appropriately – indeed, no mechanism to prevent all users marking all their traffic "top priority" (which for them it may well be).

Perhaps a fundamental omission in the current Internet architecture is any means to pass payment along with each packet – though there are no good proposals for how such a thing might be achieved. If there were such a mechanism then each AS which a priority packet passed through could collect a toll, and handle the packet appropriately, and the user would pay for every priority packet. Research problems include: how prices would be negotiated across multiple ASes; how return traffic would be paid for; and how dropped packets would be accounted for.

3.1.16 Congestion and the Transmission Control Protocol (TCP)

As discussed above, the basic mechanisms for ensuring that all parts of the Internet can reach all other parts take no account of how well any two parts are connected, and in particular what capacity there is available.

A founding tenet of the Internet architecture is that the network should do the absolute bare minimum necessary to maintain the ability to forward packets. That problem is hard enough. Anything that the end-points of a conversation across the Internet need which the network does not supply, the end points must supply.

This 'end-to-end' principle is exemplified by the Transmission Control Protocol (TCP). Most conversations across the Internet are carried by TCP connections between each end of the conversation. TCP deals with the issues caused by the fact that the network offers no guarantee whether or how packets are delivered.

Among the issues that TCP deals with is congestion. When a TCP connection detects (or suspects) congestion is affecting it, it reduces the load that it is placing on the network. All TCP connections which cross a congested part of the network should do this [**67**]. Although they are operating independently, the effect is that they cooperate to reduce the congestion[26]. TCP is not only making up for the things that the network cannot do, it is also adapting the load on the network to make the most of what it can do, but no more. Increasingly, however, network operators are addressing congestion in the network – for a discussion of the evolution of congestion mechanisms see [**68**].

It is important to note that TCP cannot do anything about network delays; it cannot cause packets to be delivered in a given time or at a given rate. For delay or rate sensitive traffic, if the network is congested, it is congested.

3.1.17 Traffic Aggregation and Capacity Management

Individual sources of traffic can be extremely variable. In the core of a network and where it connects to other networks, high capacity links carry the aggregate of traffic from thousands of different sources. Aggregate traffic levels are remarkably stable. Traffic varies across the day, and across the days of the week. Traffic varies from month to month across the year. These cycles are relatively slow, and a network operator can make reasonable predictions about, say, next month's peak traffic. Overlaid on these patterns of use are longer term growth (almost invariably, growth) trends, caused by increasing consumption per end user, or increasing numbers of users, and so on.

[26] *Conforming to the standard is voluntary. If an implementation of TCP does not 'play by the rules' then it can grab more of the available bandwidth, effectively elbowing more 'polite' implementations out of the way.* ***[223]*** *reports that perhaps 5% of TCP conversations were not playing by their sample of backbone traffic. The authors suggest that this may increase. This is a potential 'tragedy of the commons' issue, though* ***[238]*** *suggests this is not the case..*

The following shows a typical weekly cycle of traffic on connections between networks:

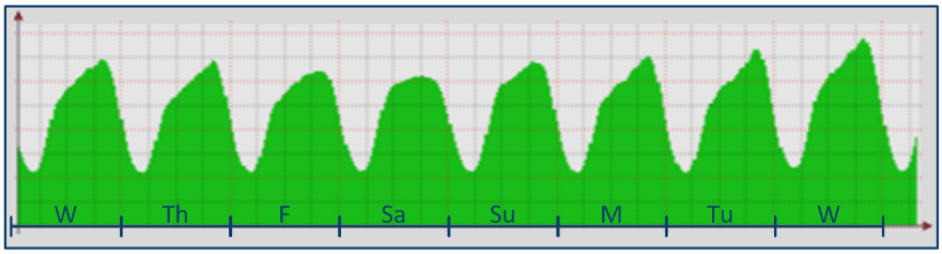


Figure 23: Typical Weekly Traffic Cycle

The lowest demand is around 04:00-05:00 each day, and demand peaks at around 20:00. In this example the day with the lowest peak demand is a Saturday. This graph shows eight days, and we can see that the most recent Wednesday's peak traffic is about 10% higher than the previous Wednesday's (which is quite a lot higher, perhaps a symptom of something unusual).

Note that the issue is the peak traffic. Average traffic, which in this example is about 60% of the peak, is not the issue, nor is it the number of bytes transferred over the week. To avoid congestion the network operator must deal with the peak traffic. To avoid congestion during busy periods each link in the network needs sufficient capacity to cover its peak demand.

Note also that congestion occurs on individual links in a network; it does not occur across the entire network uniformly. The network has no automatic means to move traffic around to relieve congestion on particular links. When their monitoring systems indicate that some link or links are congested, the network operator may be able to reconfigure their network to make different routing decisions and hence change the distribution of traffic – though this is not an exact science.

Congestion on some links in a network affects only the traffic routed to use those links, which may be a small percentage of the total. Any change to a network carries a risk of error or unintended consequences – bearing in mind the difficulty of predicting the effect of a routing change – which must be balanced against the likely gain in service quality for some small percentage of total traffic.

Capacity costs money, so the management of capacity is an essential part of managing a network. Since aggregated traffic is reasonably stable, an operator can base its capacity on history and its view of any underlying growth factors. When deciding on suitable capacity for a given link, the operator will add some margin to the projected peak demand, to:

a. absorb possible short bursts of traffic.
b. allow for some variation – such as that between successive Wednesdays in Figure 23, above.
c. allow time to respond should the speed of underlying growth have been underestimated.
d. generally provide for the unknown.

The general approach means that there is usually some spare capacity to absorb a 'reasonable fluctuation' in demand [**69**]. Fluctuation in demand may be caused by changes in end-user behaviour – for example, a mass audience for an online sports event – or by failure of some part of the network. What is deemed a 'reasonable fluctuation' will depend on the network operator, who will take a view on what should be absorbed without it creating congestion. This is not an exact science, and acceptable levels of spare capacity will be based on experience and rules of thumb in addition to the operator's traffic measurements [**70**].

Since it is hard to predict both fluctuations and failures, the network operator must expect to tolerate occasional periods of congestion, at least until either the surge in demand abates or some other steps to re-route traffic can be taken. Operators can move traffic around in their own network to make best use of capacity but, as noted above, traffic engineering at the interconnection system level is difficult. At least demand is not inflexible, as TCP will reduce demand when it encounters congestion.

Some people argue that Internet traffic is 'self-similar', which is to say that at any scale of traffic the degree of variation in traffic volume will be the same. An individual source of traffic can be very, very variable. So this would imply that many sources of traffic, even when aggregated together, can also be very variable [**71**]. In practice, aggregating traffic appears to reduce the variability, or there are other effects which smooth out the peaks – perhaps TCP's response to congestion, or at the measurement of traffic by averaging bits transmitted over five minute periods.

3.1.18 Local vs Global – Traffic vs Reachability

One obvious metric for how well the interconnection system is working is the proportion of all possible addresses which are reachable at any given moment. It is not easy to measure this for the entire system, but in any case it is not that relevant to each user of the system. While each user of the system expects to be able to reach everywhere, their use is dominated by a tiny minority of possible destinations. So while global reachability is important, there may be many, different, local perspectives on what is significant. Being able to reach a given destination has some value. In the absence of any other measure, that value can be gauged by the amount of traffic to and from that destination. Again, there may be many, different, local perspectives on what destinations have what value.

We may think of reachability as a 'static' view of the system while the amount of traffic and its distribution are a 'dynamic' view of it. (Of course, destinations come and go, and routes to destinations come and go, but compared to traffic, these are static.) At any given moment, it is the traffic that matters, since that reflects what users and customers actually want. Sadly, traffic and its distribution are more difficult to measure than reachability.

A large proportion of traffic is 'local', that is, it passes between places that are relatively local to each other. Language and national boundaries have some influence. Also, we tend to interact more with people and organisations which are local to us and less with those which are further away. There are some Internet hyper-giants which attract traffic on a global scale, but those tend to have a local presence, so tend to reinforce locality rather than dilute it.

There is not a lot of good data on Internet traffic or even types of traffic [**72**]. There is not much data on the locality of Internet traffic, and we are in any case leaving the definition of what is local somewhat open. However, estimates for the proportion of local traffic vary between 60% and 90%. In recent years there has been a lot of concern about the volume of Peer-to-Peer (P2P) traffic, and its locality – particularly with some ISPs identifying 60% or more of their traffic as P2P. Now, however, video traffic is emerging as the dominant type of traffic by volume, and that is increasingly delivered by local systems – notably local servers operated by content delivery networks. This means that the proportion of traffic which is local is likely to increase.

There are two apparent conundrums here. The Internet is Global, but most traffic is Local. The Internet routing mechanisms are all to do with Reachability, but users care more about Traffic.

3.1.19 On 'Connectivity'

The notion of 'connectivity' is as important as it is vague. Network people talk of improving connectivity, or of a network being well-connected, as generally qualitative but desirable qualities.

As a rule, the more directly connected two networks are, the better. There are a number of reasons for this:

a. a connection between two machines 20 milliseconds apart will generally perform better than one between machines that are 200 milliseconds apart. The shorter the connection, the more responsive it will be, and the more likely to deliver data quickly.

b. short links are cheaper than long distance ones, so short links are more likely to be high speed and less likely to be heavily loaded – so likely to perform better.

c. the more routers a connection passes through, the more links it depends on and the more queues each packet will sit in – so the fewer routers, the less likely the connection is to encounter problems.

d. the more ASes a connection passes through, the more likely it is to pass through more routers, plus the more organisations the path depends on.

The acme, therefore, is for networks to be directly connected. The next best is to be connected via a single transit provider, and so on.

The recent rise and rise of the content delivery networks, which we will discuss later, is related to this. By placing copies of web sites as close as possible to as many users as possible, those web sites gain the advantages of being more directly connected to their users. This improves the quality of web sites which deliver large volumes of traffic, particularly throughput- and delay-sensitive traffic – notably video traffic. It also improves quality where it is important for the site to respond quickly – as in 'cloud computing'.

The quality of a network's connections affects the quality of its connectivity. That has many dimensions, including capacity, loading, reliability, resilience, etc. Given the difficulty of measuring any of these things, only a gross notion of quality can be taken into account. Mainly, when we speak of improving a given network's connectivity, we mean improving its connections with the networks with which it exchanges a material amount of traffic. Given that much traffic is local, a practical way of improving a network's connectivity is to improve its connections with local networks.

When we speak of a network as being well-connected, we probably mean that it has good, preferably direct, connections to a wide range of destinations – though we would probably discount destinations which account for a trivial amount of traffic. So, a network that is well-connected in Europe, for example, one would expect to be directly connected to all the large networks in Europe, most of the medium size ones and generally at most one network removed from any network of consequence – this is, after all, a qualitative measure.

3.1.20 Key Points

The mesh of BGP routers does an astonishing job of distributing routes across the entire Internet. Without any central organisation, tens of thousands of networks learn about each other and can exchange traffic. But it is limited, and those limitations affect how we consider the resilience of the system:

a. it is hard for any AS to know a great deal about how traffic from it reaches its destination or to try to influence that. An AS will track volumes of traffic across its connections to other ASes, but will not have a finer grain view;

b. it is even harder for any AS to control how traffic is sent to it.

c. the complexity and information hiding properties of the BGP mesh mean that it is hard or impossible to know what it is doing. It is hard for an AS to predict where traffic will move to when it makes changes to its own network, or when others make changes and the routes available to the AS change. It is even hard to discover the topology of the BGP mesh.

d. experience with BGP has shown that most of the time it is better to react relatively slowly to change – the global mesh of BGP routers is more likely to remain stable that way. The downside is that it can take a while for the effect of some large scale change to be detected and for a given AS to make the necessary adjustments to its route selections. While this is going on, packets will continue to be forwarded on the basis of out of date route selections, and may well not reach their destination. This hiatus may be measured in minutes or tens of minutes, which for some kinds of traffic is inconvenient, but for other kinds of traffic it may be extremely disruptive.

e. BGP is concerned only with reachability – that is it tells an AS that a given destination can be reached, but does not say how much traffic a given route can carry, or how well it will be carried. Each AS must monitor and maintain the quality of its connections to the rest of the Internet by some other means, possibly requiring manual intervention (and limited by the difficulty of controlling traffic). If there is a major change to the routes available to an AS, it may take a long time (hours or days) to assess and adjust to the capacity and quality of the remaining routes.

f. BGP is not secure. It does not provide any means to verify that the routes which it distributes are valid. Proposals that address this issue are complex and add complexity and cost. An authoritative table of what address blocks an AS is entitled to use is essential to any validation scheme, but is not the complete solution. The RPKI initiative will provide that table.

g. capacity is key and capacity is managed by tracking demand. The practical way to manage a network is to manage capacity to meet the expected month to month demand, where that expectation is based on previous demand. From a resilience perspective, it is important to note that this does not take into account any unusual demand which might be created in the event of an extraordinary shift of traffic caused by a major event.

All events which negatively affect the Internet will, in essence, disable some quantity of equipment and some number of connections across and between one or more ASes. The effect of that may be broken down as follows:

a. static: the loss of some routes to some destinations. The BGP mesh will adjust to this automatically, though not instantly. While the mesh adjusts there will be some disruption to traffic. The more distinct routes there are to a destination, the less likely it is that a given event

will affect all routes to that destination. (The problem of whether the infrastructure which supports those routes is also distinct will be discussed later.)

b. dynamic: the loss of some capacity between some destinations. The effect of this will depend on whether the routes onto which traffic is diverted have sufficient spare capacity to cope with the new demand without becoming overloaded. Most networks maintain a margin of spare capacity to allow for routine fluctuations in demand. Whether any given event will exceed this pool of spare capacity is essentially impossible to predict. It is worth remembering that when congestion is detected TCP will reduce the demands it makes on the network. For many Internet applications this means that an event which creates overload in some parts of the network may be detected only as a performance reduction, and not as a complete breakdown; unless the overload is so severe that TCP can no longer cope. Congestion will, however, adversely affect Internet applications which are time and/or capacity critical, for example Voice-over-IP or any form of real-time audio or video service.

The difficulties of measuring and controlling the behaviour of BGP and the resulting Internet wide fabric of routes, suggests that its resilience can only (as a practical matter) be considered on a probabilistic basis.

3.2 The Physical and Link Layers

The physical and link layers underpin the interconnection system. The network layer discussed above is overlaid on the physical layer. The higher layers of the system, discussed below, sit on top of the network layer.

The physical layer includes the following:

a. routers and other equipment.

b. sites for that equipment complete with:

   i. reliable electrical power – a key element in the resilience of the system (and which may be interdependent with the Internet);

   ii. reliable cooling – also dependent on electrical power;

   iii. physical security – note that colocation sites house many different operators' equipment and are where those operators can interconnect.

   A site may also be known as a 'Point of Presence' or 'PoP'.

c. networks of fibre and other cabling, which includes:

   i. the ducting and other physical infrastructure that protects the cabling;

   ii. cabling within sites – particularly between operators in a given site;

   iii. cabling between sites – from metropolitan networks within a city to continental and inter-continental cables.

From a resilience perspective, some of this physical infrastructure is concentrated in relatively small areas, so that single failures can have a significant impact. A single fibre cable will comprise many fibres, each capable of carrying hundreds of Gbits/sec, and a single cut anywhere along the length of the cable is enough to stop it working. In some places many fibre cables are laid side by side in

conduits. Some undersea cable systems are particularly vulnerable; in some parts of the world there are relatively few of them and they converge into some surprisingly small areas.

The link layer provides the connections between routers within an AS, and the interconnections between ASes. There are two forms of links between ASes:

1. direct links between a router in one AS and a router in the other;
2. indirect links, notably via an Internet Exchange Point (IXP).

The physical and link layers are complex systems in their own right.

The link layer starts with physical links – generally fibre links – between various sorts of equipment. Over those physical links may ride many (anything from 16 to 160) independent wavelengths, each providing up to 10-40Gbit/sec. The capacity of each wavelength may be divided up into many separate circuits. Those circuits may then be part of quite independent networks, and those networks may themselves support circuits which are components of other networks.

As soon as we step up from the network of physical cables we enter a many layered system of networks, each of which provides circuits which are the links in the next network layer above. At each level there will be multiple operators using facilities provided by the operator of the lower level, and to build their network each operator will use facilities from many other operators. So: one operator may own a network of fibre cables in some metropolitan area, and sell fibres to other operators; those operators may light wavelengths, and sell either entire wavelengths or space on one wavelength to yet other operators; those other operators may build networks on top of that, and sell virtual circuits across those networks; and so on. Many apparently separate links may be dependent on one physical cable.

From a resilience perspective the physical layer is critical, and most incident scenarios start with the failure, disabling or destruction of some part of the physical layer. One of the key difficulties in assessing resilience is assessing, first, what effect a possible event would have on the physical infrastructure, and then, second, translating that into the effect on the link layer, and then, third, translating that into the effect on the interconnection layer.

3.2.1 Direct Links

There are many different kinds of direct link, each with its own resilience properties. In general, the closer the two routers are to each other, the less the link will cost to set up and maintain.

Direct links are of three basic types:

a. direct fibre: the simplest direct link is a glass-fibre pair running between the two routers, with no intervening equipment. Such links are generally possible where the two routers are in the same building, or in areas where there is plenty of fibre available between buildings (usually close to each other). These links have a consistent and guaranteed minimum and maximum capacity. They fail if something fails in either router or if the physical link is cut.

b. data circuit: the next simplest link is some form of data circuit between the two routers. Such links may cover any distance, and will be used where a direct fibre link is either impossible or uneconomic. These links also have a consistent and guaranteed minimum and maximum capacity. They will also fail if something fails in either router. The possibility of failure of the data circuit between the routers depends on the nature of the circuit. In some cases, the possibility of failure is increased simply because there is more equipment involved. In other

cases, and probably at extra cost, the data circuit may have some redundancy built in, and hence offer greater reliability than a direct fibre link.

c. virtual circuit: the most complicated link is some form of network supporting a virtual circuit between the two routers. Such links may cover also cover any distance, and are generally more flexible and less expensive than direct data circuits. These links do not offer consistent and guaranteed capacity. Part of the reason that these are less expensive than a data circuit is that their underlying transport mechanism is another network – either an Ethernet (or other Layer 2 network) or an IP or IP/MPLS network. The service provider may offer a guarantee of a certain minimum capacity for, say, 97.5% of the time. But if the provider's network becomes busy, for example in the event of a major network incident elsewhere, all users may experience degradation of the service.

Whatever form the link takes, it will have a fixed cost (a virtual circuit may, in addition have some usage based charges). The volume of traffic flowing between the two ASes must justify that cost.

These are direct links as far as the network layer – BGP and the rest of the interconnection system – is concerned. Data and virtual circuits may, as previously discussed, be provided by a complex and itself many layered system of fibre, wavelength and other networks.

3.2.2 Indirect Links – Internet Exchange Points

Most indirect links are Internet Exchange Points. Some providers of metropolitan area networks offer interconnection between their customers in addition to their other services, so operate a form of IXP on the side.

An Internet Exchange Point is, essentially, a switch. Many ASes may connect to the IXP, which requires a direct link (of any of the kinds described above) between a router in each AS and the IXP. Once connected to the IXP, an AS may link across it to any or all of the other connected ASes. The advantage is straightforward. A single link to the IXP allows an AS to exchange traffic with many other ASes. This may not only be a cost saving, but may allow some ASes to connect to each other when the cost of a direct link would be prohibitive. [**73**]

IXPs offer significant economies of scale. An IXP which attracts many ASes and carries a lot of traffic will tend to attract more ASes and more traffic. The downside is that ASes' links to the IXP and the IXP itself become a potentially serious point of failure. The larger IXPs allow (and encourage) ASes to make more than one link to the IXP, and take steps to ensure that the IXP itself is resilient.

Most IXPs are, as the name suggests, points; that is, they comprise a certain amount of equipment (typically Ethernet switches) in a single site. Users of the IXP must connect to the IXP, which does not have any network of its own – so there is a clear demarcation between the IXP and the networks that connect to it. The larger IXPs have, however, expanded to have equipment in several sites, with links between, to form an extended "point". This benefits the users of the IXP, because it makes it easier for more ASes to connect. There are also resilience benefits, where ASes connect in more than one site.

When an IXP extends to a new site, the cost of that extension increases the cost of the IXP for all its users. However, the extension is seen as mostly benefiting the (initially perhaps small numbers of) new users who connect at the new site – because they could previously have connected to the IXP, but at greater cost. This limits the IXPs expansion and partly accounts for IXPs continuing to be points or locally 'smeared' points.

Europe is particularly rich in IXPs. A local IXP is, obviously, the best place to exchange local traffic, from both cost and performance perspectives. Language and other national factors mean that a good proportion of traffic that arises locally will terminate locally.

It is effective to have local copies of web sites which must respond rapidly if they are to provide the required level of service, or which deliver large amounts of traffic, or both. Placing such local copies of a web site close to an IXP, and using the IXP to connect to ISPs locally, is an obvious strategy.

3.2.3 Clusters and Clustering

Because inter-AS connections are generally cheaper the shorter they are, ASes cluster together. These clusters generally appear in major cities, usually growing up around pre-existing telecommunications centres. These clusters attract investment in colocation sites and in fibre networks between those sites. This reduces the cost of locating infrastructure in the cluster, which attracts more ASes, and so on. As clusters develop around the world, they attract investment in networks connecting those clusters.

IXPs and clusters of sites are strongly related, and often develop in parallel. Colocation providers often promote, and may create, an IXP to add value to their site. Within a cluster one IXP will be dominant (or unique) or there will be a small number of IXPs of approximately equal size. This is because an IXP benefits from strong network effects to the point that it may become a natural monopoly – traffic attracts traffic. This is a further reason for IXPs continuing to be 'points'. The larger IXPs can attract connections from a surprising distance away.

The effect of all this is that the Internet, at least from an interconnection perspective, looks like a number of clusters of ASes (typically in major cities), where within each cluster interconnections are quite dense, and those clusters live on top of various networks carrying traffic between them. From a resilience perspective, these clusters have the advantage of fostering dense and diverse interconnection between ASes. However, they have the disadvantage of concentrating possibly vulnerable infrastructure in relatively small areas (perhaps a few square kilometres).

3.3 The Operational Layer

The operational layer consists of the people, processes, systems and equipment that monitor, manage and maintain networks.

The operational layer includes the management of the network and physical layers, as described above. The operational layer is also involved in the relationships between ASes, which is described below.

Each AS has its 'Network Operations Centre' (NOC), which runs its network and its connections to other ASes. The functions of the NOC include:

a. monitoring and measurement: the AS must know how well its network is working, everything else depends on this.

b. dealing with equipment and circuit failures.

c. network management: the day-to-day adjustments needed to maintain service.

d. capacity management: the longer time-scale changes needed to keep pace with demand and maintain service quality.

Congestion is the key issue that the each AS must deal with at the operational layer. As discussed in Section 3.1 above, the avoidance of congestion is key to service quality, and that is not done by the routing mechanisms at the network layer.

Where congestion occurs within an AS, it will move traffic around to relieve the problem, using whatever suitable spare capacity it has. Within its own network an AS will have a number of tools at its disposal to achieve this – subject, of course, to there being spare capacity on hand.

Where congestion occurs on an AS's direct connection to another AS, what the AS can do depends on whether outbound or inbound traffic is affected. For outbound traffic the AS can treat the problem as an internal one, and move traffic around to other suitable connections to other AS(es). For inbound traffic the AS has to treat it as congestion outside itself.

Where congestion outside an AS affects traffic to and from it, the problem is more difficult, because of the very limited tools at the disposal of the AS to influence traffic outside itself. If the location of the congestion can be identified, then the NOC may contact the responsible NOC and try get them to resolve the issue. This is likely to be easier if the congestion is in a directly connected AS. Of course, it is to be expected that wherever the congestion is, the AS in question will be working to relieve it.

The objective of capacity management is to avoid congestion under normal circumstances.

As will be seen later, the operational layer is a key part of the resilience of the interconnection system. Where some event creates congestion, the operational layer must deal with that. In a large scale event the operational layer will manage the recovery of service and the restoration of networks.

All the individual NOCs, across the interconnection system, strive all day, every day, to keep their networks running and free from congestion. While there is no coordination of the operational layer, it operates coherently to a common goal.

3.4 The Operational and Commercial Layers – Peering and Transit

So far we have covered the mechanics of how routes are distributed between ASes, and how ASes are physically connected to each other. Now we look at why two ASes choose to connect, what those connections do, and how that is paid for. (For a historical perspective on this see [**74**], [**75**] [**76**] and [**77**].)

Every AS is run by an independent organisation[27]. Each AS must arrange for itself a way to reach the entire Internet, and ensure that the entire Internet has a way to reach it – and ensure that it has access to enough capacity to satisfy its users and customers. It does that by connecting to a number of other ASes and entering into arrangements with them to exchange routes and traffic. There is an operational aspect to this: how each connection is configured and what it will do. There is a commercial aspect to this: what the connection and the resulting traffic is worth and how those paid for, which may influence whether a connection is established in the first place.

[27] *Except for the very small number of cases where an organisation has more than one AS, in which case a small group of technically separate ASes behave in a coordinated fashion, more or less like a single AS, at the operational and commercial levels.*

Not all ASes are equal. Some are small Internet Service Providers (ISPs) with small networks, serving customers in, for example, a small town or group of towns. Some are larger ISPs with larger networks serving customers in, say, a country or a part of a country. Some are still larger networks which operate on a multi-national or regional scale. There are a few networks that are more or less global – though even the largest networks are usually stronger in some regions than in others. It is generally accepted that the scale of ASes follows a power-law. So there are many small ASes and a few very large ones.

Every connection between two ASes is a bilateral arrangement. When AS1 connects to AS2 it has to decide:

a. what routes to announce to AS2. When it announces a route to AS2, AS1 is undertaking to carry traffic to that destination. So, if it announces routes learned from other ASes, it is undertaking to carry traffic from AS2 across itself to those other ASes.

b. what to do with routes received from AS2. If it announces those routes to other ASes, then it is undertaking to carry traffic from those ASes across itself to AS2.

Similarly AS2 must decide what to announce to AS1 and what to do with routes learned from AS1.

Neither AS is going to carry traffic that is not paid for. Each AS is home to a number of blocks of IP addresses; routes to those addresses are known as the AS's 'own routes', most of which will be for the AS's users, including its 'direct customers'. An AS may also have customers, other ASes, with their own blocks of IP addresses; routes to those addresses are known as the AS's 'customer routes'. Traffic to or from an AS's own routes and its customer routes is paid for by the AS's users and customers – indeed, the entire function of an AS is to carry traffic to and from those routes.

There are two basic arrangements that two ASes can come to:

1. a 'Peering' arrangement. In any form of peering arrangement the parties exchange traffic destined only for their own users and customers, including transit customers. The traffic is paid for by the two parties' users and customers – it is part of the 'Internet access' that they are buying.

2. a 'Transit' arrangement. In any form of transit arrangement one of the parties, the transit provider, will carry traffic to and from other parts of the Internet – not just to and from its own users and customers. The transit provider will charge for this service; on what terms is discussed below.

In [9] peering is described as a 'horizontal' relationship, and transit as a 'vertical' one.

3.4.1 Peering Arrangements

Peering is the simpler arrangement between ASes, in which they will both:

1. announce their own and their customer routes to each other. This means that each peer is announcing routes for all the addresses it is paid to carry traffic for, where: (a) 'own routes' means routes to all the address blocks within the AS, which will include addresses used by its direct customers; and (b) 'customer routes' means routes learned from transit customer ASes.

2. not announce the routes they learn from each other to any other AS, other than their customers. If an AS announced one peer's routes to another peer, then the AS would be providing a free connection between the two peers – in effect a form of free transit. Similarly,

if an AS announced a peer's routes to its transit providers, it would be providing free transit to that peer.

This is a generally symmetrical and mutually beneficial arrangement. All traffic using the connection is destined for addresses within the receiving AS or its customers. No traffic is destined to go any further.

Apart from the very largest networks (the Tier 1 networks, discussed in 3.5.2 below), all networks must buy some transit; peering has quality advantages, but is not essential; one transit provider is enough (a bare minimum) to reach the rest of the Internet.

At an IXP ASes will, generally, peer with each other. Each AS bears its own costs in reaching the IXP, and its share of the costs of maintaining the IXP. Peering at an IXP is generally called 'public peering' (except where the IXP itself is somehow private). If the ASes connect directly, i.e. not at an IXP, they will come to some arrangement about who pays what to cover the cost of the link between them. Peering where ASes connect directly is generally called 'private peering' (and the terms on which they connect, and indeed the existence of the connection at all, may be considered confidential by the parties).

The economics of peering are such that when smaller ASes do peer, it will generally be at an IXP. To justify the expense of a direct peering connection the two ASes need to be exchanging a significant amount of traffic with each other. For most ASes a direct peering connection will be contemplated only if there is no suitable IXP, or one of the ASes does not wish to connect to a suitable IXP. For connections between the largest ASes, however, private peering is the norm. Technically, a peering arrangement in which neither AS charges the other is a 'settlement-free peering arrangement'. But because few peering arrangements are otherwise, 'settlement-free' is implied by the term 'peering', unless otherwise qualified.

The fact that an AS does not advertise a peer's routes to any other peer is often referred to in the literature as the 'no valley' rule, which is illustrated here:

Figure 24: The 'No Valley' Rule

where AS2529 peers with AS4321 and AS5417, and all have various transit providers. Traffic between any of AS4321's users and customers and AS5417's users and customers will travel all the way to the top, via the various transit providers, and back down again – following the green, dashed path. AS2529 could carry this traffic, and would if it announced its peers' routes to each other, creating the dotted line path. But AS2529 would want to be compensated for providing this service, and in the absence of any other arrangement, it will apply the 'no valley' rule.

When a peering connection is set up, each AS will configure its BGP routers to select routes learned from the peer, in preference to routes learned from any form of transit provider. A peering connection is the best way to reach the peer, but in any case there is no point paying a transit provider to carry traffic that can be exchanged at no cost directly with the peer! On the other hand,

an AS will prefer routes learned from a transit customer – why pay a transit provider, or give the traffic to a peer, when a customer is there to pay for it?

3.4.2 Paid Peering Arrangements

Paid peering is unusual. The two parties peer as above, but one pays the other for the privilege. The motivation for this is discussed later in Section 3.6.6 below.

3.4.3 Transit Arrangements

In a transit arrangement one party, the transit provider, undertakes to carry traffic across itself (hence 'transit') to and from the entire Internet, on behalf of the other party, the transit customer. For this service the transit provider will invariably make a charge.

In terms of route announcements, the transit provider will:

1. announce all the routes it knows to the transit customer;
2. announce all the routes it learns from the transit customer to every AS it connects to.

and it will undertake that these announcements provide access to and from the entire Internet. This is, technically, 'Full Transit', but most transit arrangements are full transit arrangements, so 'Full' is implied by the term 'Transit'. 'Partial Transit', where access to just some part of the Internet is provided, is unusual but is covered below.

The transit customer will:

1. announce their own and their customer routes to the transit provider;
2. not announce the routes it learns from the transit provider to any other AS, other than their customers – because if they did they would be providing free transit.

Note that the business of being an ISP is essentially to sell access to the Internet. In a transit arrangement one AS sells access to the Internet to another AS. Transit customer will carry all of the costs of the link between the ASes. Generally this will be a direct connection of some kind (though occasionally an IXP may be used.)

In addition, the transit customer will be charged for the traffic that the transit provider carries on its behalf. Generally that charge depends on the minimum amount of traffic the customer has committed to and the actual traffic during each month – usually measured as the 95th percentile of traffic samples taken every five minutes in each month [**78**]. This means that the cost of transit depends, effectively, on the peak demand in a given month. By ignoring the top 5% of the samples, the provider is ignoring the busiest 36 hours in each month. (This can be especially handy if some network failure diverts unusual amounts of traffic to a transit provider.)

Conceptually, what the customer is paying is a retainer to cover the minimum committed capacity across the provider's network. If the customer exceeds the minimum committed capacity, the provider charges for the extra traffic, and may or may not offer much in the way of a guarantee of actually being able to deliver the excess. Transit cost is almost, but not quite, a fixed cost for an agreed transit capacity. Provided the customer guesses the correct level for the minimum commitment, and does not exceed that level, the cost is effectively fixed. If the customer exceeds the minimum level in a given month, then there will be an extra charge.

When a transit connection is set up, the transit provider will configure its BGP routers to select routes learned from the transit customer, in preference to routes learned from anywhere else. Since the customer pays for the traffic, there is no reason to want to use another route. This preference means that where two transit providers peer with each other, they will not send each other traffic for any customers they share.

3.4.4 Partial Transit Arrangements

A partial transit arrangement is similar to full transit, except that the transit provider is offering access to part of the Internet, not the entire Internet. For example, where an AS is particularly strong in one region it may offer partial transit to that region. Suppose an AS is well-connected in Europe, so has many customers across Europe, and peers with many other networks across Europe. It can offer partial transit, transit just to its customers and peers, and incur little extra cost.

If that AS offered full transit, it would have to buy in transit to the rest of the world from other ASes, on which margins may be slim. Since it costs the partial transit supplier very little to source the traffic it sells as partial transit, it can be supplied at lower cost than full transit. So a partial transit customer may improve their connections to a given region and reduce their costs. Furthermore, the more partial transit customers that the provider can acquire, the better its connectivity, and the more attractive its partial transit offering becomes.

3.4.5 Mutual Transit Arrangements

Mutual transit is theoretically possible. In a mutual transit arrangement each AS would:

1. announce all the routes they know to the other;
2. announce all the other's own and customer routes to every other AS it connects to, including and especially its transit providers.

The effect of this would be that each AS gains access through the other to the other's transit and peering connections. So if the two ASes have different transit providers, their combined transit supply is more diverse. Also, the two ASes combined peering connections are likely to be more extensive and diverse.

It would be unlikely for mutual transit to be used in normal circumstances, but might be enabled in a crisis where the extra diversity it offers could improve connectivity for both parties. The parties would need to recover any extra transit costs they each incurred. The difficulty in establishing those costs, and the generally low probability of needing to invoke a mutual transit arrangement, mean that mutual transit is more theoretical than actual. However, such arrangements could be beneficial in an emergency.

Where an operator organises their network as more than one AS, those ASes will effectively provide mutual transit to each other. In the literature these are referred to as 'sibling' ASes. The unusual sibling relationship – neither simple peers nor transit customer/provider – causes difficulty when trying to establish how ASes are interconnected by observing their behaviour from the outside.

3.4.6 How ASes Interconnect

The vast majority of connections between ASes are either (settlement-free) peering or (full) transit.

Most ASes in the Internet buy transit to reach most destinations. Where an AS buys transit it will generally buy from two or more distinct suppliers. This is called 'multi-homing' [**79**]. If one of the transit connections fails, then the remaining transit connection(s) must carry the traffic. By buying from distinct suppliers the AS hopes to avoid having more than one connection fail at the same time.

The extent to which an AS will peer with other ASes varies, depending mostly on the scale of the AS. A very small AS may not find the extra fixed costs of peering are justified – though if there is a local IXP the AS would have to be very small indeed. The very largest ASes must peer with each other, but may not peer with smaller ASes. The commercial imperatives which govern this are covered in Section 3.6 below. Where ASes peer with each other, they must be prepared for any failure of the peering connection. In some cases, that can be one or more separate peering connections. For most ASes, however, their transit arrangements are implicitly their back-up connections to their peers.

Where an AS is connected to an IXP, it is possible for the AS to exchange a significant proportion of its traffic at the IXP. This is an incentive for the users of the IXP to have more than one connection to the IXP. It is also an incentive for those users to ensure that the IXP is as resilient as possible. Some ASes will peer with each other at more than one IXP, in order to improve the resilience of their peering arrangements.

3.4.7 Formal and Informal Arrangements

In any inter-AS arrangement the parties must ensure not only that the required routes are announced but also that any resulting traffic is properly looked after – which means ensuring that the inter-AS connection is monitored and managed, and that there is adequate capacity where it is needed.

In a *peering arrangement* both parties have duties to their respective users and customers, so looking after the traffic and the inter-AS connection is in both their interests. When the connection is set up the parties will establish operational procedures to deal with any problems with the connection, including traffic growing beyond its capacity. At an IXP most of that is already taken care of when connecting to the IXP, and the parties have little direct operational contact.

Many peering arrangements are entered into without a formal contract between the parties, particularly when peering at an IXP. This is for two basic reasons: first, the self-interest of both parties is a good enough incentive to keep the arrangement running properly; second, in many cases, particularly at an IXP, a peering connection is essentially optional, and may carry a relatively small amount of traffic, so the effort of entering into a formal contract is unjustified. In this context a requirement for a formal contract would most likely deter the parties from peering.

Where ASes connect directly there may be more formality, not least because the cost of the connection has to be covered, but also because there is likely to be significant traffic involved (otherwise there would be no justification for the direct connection). As will be discussed later, peering arrangements between the largest networks are likely to be formal, and are strongly driven by commercial and competitive issues.

In a *transit arrangement* (full or partial) there will be a contract, since the transit provider is providing a paid-for service. From the transit customer's perspective, all traffic on the transit connection is to or from itself or its customers, so its incentive to look after the traffic is clear. From the transit provider's perspective, all traffic on the transit connection is a paying customer's traffic, so its incentive should also be clear – though, of course, each customer may be one among many. A formal contract will cover the physical link, which may be provided by a third party. A formal

contract will cover the transit service. The contracts will include some Service Level Agreement and specify the operational arrangements to maintain the connection and the transit service.

3.4.8 Service Level Agreements

Service Level Agreements for peering arrangements are rare, since most peering arrangements are informal.

Service Level Agreements (SLAs) for transit are interesting, to the extent of how little is covered. A common SLA measure is the availability of the service, generally expressed as a minimum percentage of each calendar month for which the service is expected to be available. At 99% that measure allows for some 7.2 hours down-time per month. Further, availability generally means that the router at the transit provider's end of the connection is accessible, and is announcing routes. This does not necessarily guarantee that the router is capable of effectively carrying traffic at all times it is nominally available.

The SLA may specify some performance measures for the provider's network, but these rarely specify performance measures directly related to the customer's actual traffic. It may also specify a maximum acceptable traffic level on the transit connection.

The SLA will not cover anything beyond the borders of the transit provider's network. Despite the fact that the transit provider is selling a service which is nominally access to and from the entire Internet, the SLA provisions (such as they are) extend no further than the transit provider's own network. This may be disappointing, but is not entirely unreasonable. Clearly, once traffic leaves the transit provider's network it is no longer under its control. For the transit provider to be in position to offer guarantees for that traffic, they would need binding contracts with all other networks, who in turn would need binding contracts with the networks they connect to, and so on. Such contracts do not exist, so as a practical matter, transit providers' SLAs stop at the edge of their network.

3.5 The Sum of the Parts

So far we have looked at the components of Internet interconnect, now we look at how those fit together to form the Internet Interconnect Ecosystem. To do this we will identify some general classes of AS, and look at how they connect to each other to form the Internet.

It is common to see ASes, particularly transit providers, classified in 'tiers'. The top tier, Tier 1, are the Global ISPs, the major transit providers. Tier 1 providers sell transit to regional or national Tier 2 ISPs, who in turn sell transit to local Tier 3 ISPs. The notional tier organisation is discussed further below. At the edge of the interconnection system are ASes who do not provide transit to other ASes, these are 'stub' ASes. Stub ASes buy transit from any tier of AS, depending on availability, accessibility and cost. The conventional model is described in [**80**].

There is a wide range of different types and scales of AS, which differ in the way they tend to interconnect and in their significance in the interconnection system. The following general classes of AS, and their position in the 'Tiers Model', may be identified:

1. 'multi-homed' organisations ..... stub ASes of various sizes

   These are organisations which are not in the business of selling Internet access to other organisations, so are not counted as ISPs. These are, nevertheless, separate ASes, generally because they want the added resilience of buying Internet access – i.e. transit – from more than one ISP. The academic networks are a special case. Where they sell Internet access it is only to

academic institutions, so they are not general ISPs. On the other hand, they do peer with other networks, unlike commercial organisations.

2. small ISPs ..... stub ASes – occasionally Tier 3

   Generally local or specialist ISPs.

3. medium size ISPs ..... Tier 3 or stub ASes

   These include the national or, for larger countries, regional scale ISPs.

4. incumbent operators..... Tier 2 ASes

   Included in this class would be BT, DT, FT, NTT, Telecom Italia, Telefonica, and so on.

5. large ISPs (multi-national or regional)..... Tier 2 ASes

   Included in this class would be Colt, Cable & Wireless, Comcast, Interoute, PCCW, Reach and so on.

6. large content delivery networks (CDNs) ..... not really part of the tier model

   Google is an example of a large content provider that runs its own network and operates distributed facilities to deliver its content. Akamai and Limelight Networks are examples of large content delivery networks, providing service to many third part content providers. From the perspective of the interconnection system it does not matter whether a CDN is delivering its own content, or delivering third party content. Where the distinction matters it will be made clear.

   In the last few years the amount of traffic being delivered by the CDNs is changing the balance of the transit and peering system – a significant change which we will return to a number of times below.

7. global ISPs – major transit providers..... Tier 1 ASes and some others

   Included in this class would be Abovenet, Cogent, Level 3, Savvis, Sprint, Tinet, Verizon Business, and so on.

Each of these classes is described in more detail in the following sections. Of course this is quite a broad classification, and some networks behave partly as one class, and partly as another. The notions of small and medium are not absolute; a small ISP in one place may be a medium size ISP in another. In any of the classes there are significant differences in scale.

The role played by the largest networks leads some to refer to them as Network Service Providers (NSPs) rather than Internet Service Providers.

It is generally felt that the sizes of ASes follow some power law, so that there are a small number of very large ASes, and a large number of very small ones. There are approximately 36,000 ASes in the current Internet, of which it is estimated 80% are stub ASes, and around 20 may be counted as major transit providers. This suggests:

Level	Number of ASes
Major Transit Provider	20
Large ISPs	800
Medium Size ISP	6,400
Stub ASes	28,800

These numbers are very, very approximate, but give a sense of the distribution of the scales of ASes.

There are perhaps 30 content delivery networks of any size, but the top 3 or 4 are thought to dominate in traffic volume terms. One of the top third party CDNs claims to carry 20% of all Internet traffic. In the absence of good traffic data, this is essentially qualitative information.

Almost every AS has to buy at least some transit. Peering carries extra costs in establishing and maintaining peering connections. Those costs have to be met either by reducing the cost of obtaining the same traffic via the AS's transit connections, or by an improvement in quality or a combination of the two. Quality in this context may be the quality for exchanges of traffic between the ASes, or improvements in reliability or resilience. Peering from the ISPs' (excluding the very large ISPs) and CDNs' perspective is discussed in [**81**].

3.5.1 'Eyeball' and Content

ASes may also be classified according to their predominant sort of customer. The most significant division is into access and content ASes – this is discussed further, below. On the access side, some ASes may cater mostly for domestic customers, others for business customers, others for academic users, and so on – these distinctions affect the type of traffic to and from the AS, but are at quite a fine level of detail.

Apart from 'Peer-to-Peer' (P2P) traffic, most traffic is from web-sites to end-users. With increasing amounts of video content being delivered either as part of a web-page or via a web-site, an increasing proportion of traffic is from web-sites to end-users.

An ISP whose business is, predominantly, selling Internet access to end-users is known as an 'eyeball' ISP. Since accessing a web-site generates only a little traffic to the web-site compared to the traffic coming from the site, an eyeball network's overall traffic inbound is a lot greater than its traffic outbound.

A content ISP, on the other hand, is one whose business is, predominantly, selling Internet access to organisations who run web-sites – usually selling hosting services for those web-sites as well. A content ISP's traffic outbound is a lot greater than its traffic inbound. The content delivery networks are at the extreme end of the content network spectrum.

Some ISPs provide a range of services, so may manage to have similar inbound and outbound traffic volumes. The significance of whether a network's traffic is roughly balanced, or not, is covered in 3.6.1 below.

3.5.2 The 'Traditional' Tier Structure

The 'traditional' tier structure has Tier 1 networks at the top of the hierarchy. The distinguishing characteristic of a Tier 1 network is not how big it is, but that it does not require transit in order to reach the entire Internet – however, being a Tier 1 network and being a global network pretty much go together. This means that at the top, or the centre, of the Internet are the Tier 1 networks who all peer with each other.

Below the Tier 1 ASes are the Tier 2 networks. These are generally large networks who take transit from Tier 1 networks. It is likely that the Tier 2 networks will peer with each other. Some Tier 2 networks will peer with some of the Tier 1 ones.

Below the Tier 2 ASes are the Tier 3 ASes, who take transit from the Tier 2 networks. It is likely that Tier 3 ASes will peer with each other, and they may peer with some Tier 2 ASes. Whether Tier 3 networks provide transit to others will depend on the scale and location of the network. Anything deeper than Tier 3 is not really worth considering – generally anything beyond Tier 3 is a stub network.

For an imaginary Internet with just three Tier 1 ASes, AS1, AS2 and AS3, the strict tiered hierarchy can be pictured as shown:

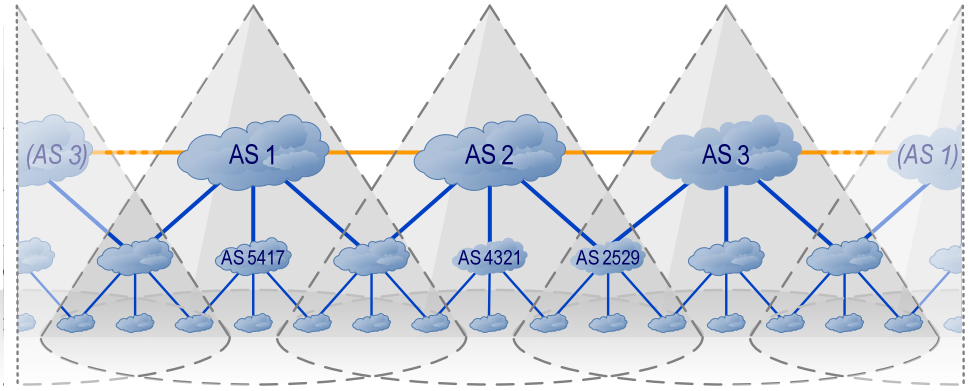


Figure 25: ISP Tiers and Customer Cones

where this diagram should be imagined wrapped around a cylinder, so that AS1 and AS3 are next to each other, just as they are both next to AS2. The Tier 1 ASes all peer with each other, as shown by the horizontal, orange lines. The Tier 1 ASes provide transit to the Tier 2 ASes, who in turn provide transit to the Tier 2 ASes – as shown by the blue lines. Some of the Tier 2 and 3 ASes are shown as 'single-homed' – that is they only have one transit provider – which does happen, but, at Tier 2 at least, perhaps not to the degree suggested by the diagram.

The 'customer cone' or just 'cone' of an AS is all the ASes that it provides transit to, and the cone of all those ASes, and so on – as shown above. The diagram also shows how the cones of two ASes overlap. Not shown, but not to be forgotten, are each AS's users and direct customers, who are a vital part of an AS's cone, as are all the users and direct customers of the ASes in its cone. An important subset of an AS's cone is its 'exclusive cone', that is the part of its cone which cannot be reached via any other AS. The exclusive cone comprises the ASes users and direct customers, and the cones of any single-homed AS customers, and their single-homed customers.

Because the Tier 1 networks all peer with each other, every Tier 1 AS can reach every other AS.

If the peering connection between AS1 and AS2 stops working (for whatever reason) then the exclusive cones of those ASes will be cut off from each other. That includes AS1's and AS2's users and direct customers, and, in this illustration, AS5417 and AS4321 (and their single-homed customers) will no longer be able to reach each other.

Tier 1 providers will peer with each other in multiple locations, so the connection between them should not fail. However, just occasionally one Tier 1 AS will 'de-peer' another, for commercial reasons, which stops the peering connection and cuts some fraction of the Internet off from another.

Another way of visualising the overlapping codes is shown opposite, where the various segments represent parts of the cones of three ASes showing where they overlap. The piece in the centre represents where the cones of all three ASes overlap.

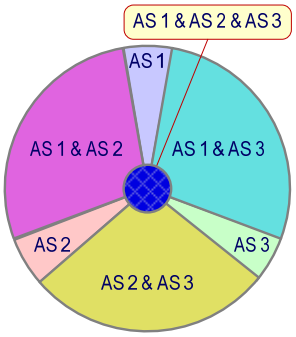


Figure 26: Cone Overlaps

In passing, this raises an interesting question: how to compare degrees of 'connectedness'? Some part of the Internet is in AS1's unique cone, but how do we compare that to, say, AS2's unique cone? One approach is to count the number of Internet addresses covered. However, not all addresses are used, and not all addresses are of equal value. Another approach is to consider how much traffic goes to and from the addresses in question. But that assumes traffic volume is a measure of its value – and starts to depend on where the measure is being taken from (a Chinese language video-on-demand site might deliver a lot of traffic, but that will be in limited demand outside the Chinese speaking world).

The somewhat abstract tiers model is related to a simple physical view of networks and their interconnections. The following supposes the three Tier 1 networks, serving three distinct regions:

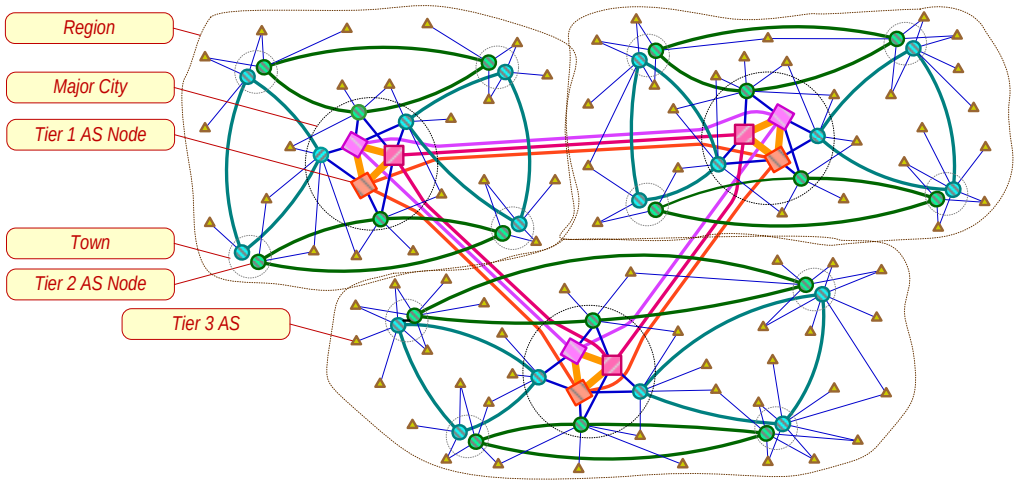


Figure 27: Simple Tiered Internet – Map

The Tier 1, major transit providers have a presence (a 'node') in each region, in the region's major city, and their network links those cities. In each city the Tier 1 networks peer with each other, so they are multiply connected to each other. The major transit providers do not have much, if any, presence outside the major city in each region.

Below Tier 1, the illustration shows a number of Tier 2 providers, each with a network within a region running between the region's major city and some of the larger towns. Then there are the Tier 3 ASes taking transit from the Tier 2 networks, and servicing their parts of the region.

Although this is a simple illustration, it does capture some of the essence of the real world, in which the major transit providers do have, in some places, similar networks. It is not significant that the number of networks and the number of regions are the same, three is just enough to give the flavour of the thing, without being impossibly complicated.

Of course this simple model is incomplete:

a. the hierarchy is not rigid in the real world – smaller ASes connect to the major transit providers directly, as well as, or instead of, taking transit from a Tier 2 provider. An AS that can reach one or more Tier 1 providers might as well buy direct. It may be possible that a Tier 2 provider can buy transit at a lower price than a smaller AS – simply because of volume pricing – and is willing to pass on a lower price. However, with ever lower transit prices, the opportunity and the absolute saving are reducing. The notion of a Tier 2 provider is increasingly less relevant.

b. it does not show the peering connections between Tier 1 and Tier2 ASes at all, or the part that the IXPs play in that – mostly because the number of connections make the diagram impossible. However, one would expect to see an IXP in each of the major cities. One would expect the Tier 2 ASes to connect to that IXP and peer with each other, and some of the Tier 2 ASes might peer privately. Finally, one would expect as many of the Tier 3 providers as could afford it to also connect to the IXP and peer with each other and as many Tier 2 providers as possible.

c. it does not include the content delivery networks – the tier model is an old one, and pre-dates the content delivery networks as they are today. One would expect the content delivery networks to have a presence at least in each of the major cities, and to peer at the local IXP and also, perhaps, privately with some of the larger ISPs.

d. it does not show the stub ASes, multi-homed enterprise networks, or users and direct customers of any of the ASes. These are all at the edge of the interconnection system. How they connect affects how reliably they can reach the rest of the Internet, and how reliably they may be reached. But this does not really affect the system as a whole.

e. an AS may, and many do, connect to more than two transit providers. Connecting to three or more transit providers reduces the impact of any single transit connection failing, and spreads the AS's traffic more thinly.

The strict definition of a Tier 1 Provider is a network which can reach the entire Internet without buying transit. To become a Tier 1 Provider an AS must, therefore, persuade all other Tier 1 providers to peer, and to remain a Tier 1 Provider an AS must retain those peering connections. A number of very large networks are not by this strict definition Tier 1, they, apparently, peer with most but not all Tier 1 providers. However, this does not prevent them being major transit providers. So when we talk about major transit providers, we include the Tier 1 networks and some others too.

3.5.3 'Multi-Homed' Organisation

In this category is any network which is an AS, but which is not in the business of selling Internet access to others – so these are stub ASes and not ISPs. Most organisations buy Internet access from a single ISP, so are, as far as the Internet is concerned, part of their ISP's AS, using some of the AS's own address space. These organisations are 'single-homed'. For resilience some may connect to their ISP from and/or to more than one location, but from the interconnection system's perspective they are still single-homed.

Where an organisation wishes to take Internet access – i.e. transit – from more than one ISP they are (generally) ASes in their own right. These are the 'multi-homed' organisations. Depending on the

scale and location of the multi-homed organisation they may connect to their transit providers locally, nationally or internationally.

Large multi-homed organisations differ from smaller ones only in the number and distribution of places where they connect to their transit providers, and perhaps in their traffic volumes. Apart from not selling transit to third parties, these networks differ from an ISP in not peering with other ASes. The large academic networks are mostly like other multi-homed organisations, in that they do not have external customers (at least, not beyond their academic users and possibly mutual arrangements with other academic networks), but they do seek peering with other ASes.

3.5.4 Small ISPs

For a small ISP it is likely that minimising the cost of connecting to its customers will be essential, and that will mean having facilities which are local to those customers. Where a local ISP offers hosting as well as access services, local facilities may be a unique selling point.

Most large transit providers have facilities in major metropolitan clusters of network sites. So the small, local ISP may:

a. buy transit from medium size ISP(s) to whom they can connect close to their local facilities;

b. set up long distance connections to transit providers in the nearest cluster;

c. establish a minimal presence in the nearest cluster, connected back to their local facilities. Once connected to the cluster it can there connect to large ISPs or major transit providers, and perhaps an IXP.

It is not essential for a small ISP to buy transit from more than one provider, but since their business is selling access to the Internet, it would be wise to have diverse access. In common with all other ISPs, these will peer with other ASes where possible and cost effective.

3.5.5 Medium Size ISPs

A medium size ISP is one which has a national or, in larger countries, regional reach. It will have facilities in one or more clusters of sites, and connect, within those clusters, to transit providers, to IXP(s) and, perhaps, directly to some peers.

The larger of these ISPs may extend connections to a few other site clusters outside their main area of operation, where that reduces the cost of traffic or improves quality or resilience. Given enough traffic, for example, it may be cost effective to connect to a distant IXP, where they can pick up more peers and more traffic, as well as offering some backup if the local IXP fails.

As noted above, these ISPs may provide transit to smaller ISPs and to multi-homed enterprises, as well as providing services – Internet access and/or hosting – to their users and direct customers.

3.5.6 Incumbent Operators

Within their own territory an incumbent operator looks like a medium size ISP, except that its network is generally more extensive than ordinary ISPs in the territory, and they are likely to be the largest or among the largest in the territory. For historical reasons the incumbent operators generally have significant multi-national or global infrastructure – partly to distribute their own traffic widely themselves, and partly to support their transit business.

3.5.7 Large ISPs

As the extent of a network increases, so its density tends to reduce. A regional ISP, for example a pan-European ISP, may have network connecting facilities in the major metropolitan clusters in each country it serves, but little beyond that.

This scale of ISP will connect to transit providers, IXPs across their region and directly with their larger peers. Where it needs to connect to customers beyond the clusters it appears in, it will buy access services from national ISPs or the incumbent operators.

These ISPs will provide transit (notably in places not reached by the global ISPs), and may offer partial transit. They will also have direct customers for Internet access and/or hosting.

3.5.8 Large Content Delivery Networks (CDNs)

A few large content providers run their own content delivery network for their own use. Most content delivery networks exist to sell their services to third party content providers.

One purpose of a CDN is to avoid repeatedly transmitting the same data over long distances. It is more cost effective to keep multiple copies of the data spread around the world, so that it can be delivered locally. The other purpose is to improve the quality of delivery – local delivery gives better and more consistent throughput and response times. IXPs are perfect places for CDNs to be connected, reducing the cost both of delivering and accessing the content. Indeed, the presence of a few of the largest CDNs at a given IXP can be an excellent reason to connect to that IXP.

These networks are, in effect, by-passing the transit providers – hence the cost savings. Since they are now delivering a significant proportion of total Internet traffic, they are a significant and increasingly significant part of the interconnect system.

3.5.9 Global ISPs – Major Transit Providers

A global ISP may be particularly strong in one or two regions, but have a presence in all parts of the world. These networks will be present in the usual metropolitan clusters, but outside their 'home' regions, perhaps only the locally major clusters. As noted above, the local or national incumbent operators may also be global ISPs outside their home territories. A regional ISP may also be a global ISP outside its home region.

These are the primary suppliers of transit across the world. When we talk of major transit providers, these are the networks we are referring to.

3.5.10 The Pattern of Interconnections

Looking at the physical infrastructure we have:

a. clusters of sites. There are smaller and larger clusters. We might identify them as:
    i. small clusters: at which local and national operators are likely to be present, and perhaps some regional operators (in areas where their network is particularly dense);
    ii. medium size clusters: at which local, national and regional operators are likely to be present, and perhaps some global operators (in areas where their network is particularly dense);

iii. large clusters: at which all but the smaller local operators are likely to be present, including many global operators.

Sites in a cluster can be network operators' own sites, or sites purpose built to be sub-let to many operators. (Network operators may sub-let parts of their own sites.) A cluster concentrates a lot of infrastructure in a relatively small area. Within the cluster there will be even more concentrated pockets of infrastructure. On the other hand, a cluster enables greater numbers of connections between networks than would be possible otherwise. From a resilience perspective the effect is mixed.

b. networks of fibre running within and between sites in a cluster. When building a fibre network the cables which are laid contain many pairs of fibres. The fibre network operator will sell the use of individual fibre pairs to different network operators. Those operators may use the entire capacity of the fibre pair, or may in turn sell part of that capacity to further operators, and so on.

   The existence of the cluster encourages the building of more fibre infrastructure, which reduces the cost of local links, which enables more interconnections and draws more networks to the cluster, and so on. In terms of cost this is wonderfully efficient. In terms of resilience, a lot of apparently independent networks can be dependent on the same run of fibre cable, which if cut could have a major impact.

c. Internet Exchange Points – the IXPs. In medium and large clusters one can expect to find an IXP, and in some cases more than one.

d. networks of fibre running between those clusters. More precisely, networks of fibre where each link is a run of fibre cable between a site in one cluster and a site in another. The existence of clusters encourages the building of more fibre infrastructure between those clusters, which reduces the cost of links between clusters, which encourages more interconnections and draws more networks to those clusters, and so on. Again, in terms of resilience, any given cable run will carry many different operators, so again a single cut can have a major impact.

It is worth noting that cost efficiency tends to have a negative effect on resilience.

That physical infrastructure supports the different classes of ASes, which connect to each other in the ways described, above. Unlike abstract map in Figure 27, which depicted the relationship between various scales of transit provider, the following diagram shows peering connections, IXPs and the CDNs, which are key parts of the interconnection system:

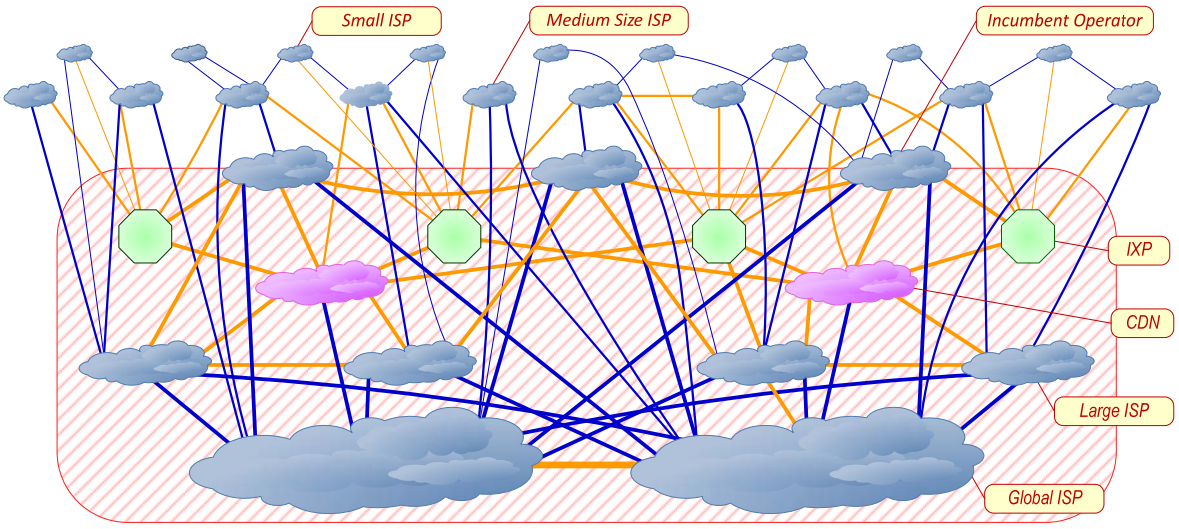


Figure 28: The System of Connections

where the clouds represent various sizes of AS, the amber lines peering connections, and the blue lines transit connections – where the larger AS is the transit provider. Working from the bottom of the diagram, it shows the global ISPs (the major transit providers), then the large ISPs, then other ISPs and enterprises. As the diagram suggests, most ASes depend directly or indirectly on the major transit providers. The IXPs facilitate peering amongst ASes of a range of sizes and the CDNs. Some ASes peer directly with each other and/or with the CDNs. Some smaller ASes may sell transit to other ASes.

The diagram is busy, but it seeks to show the common relationships between the various classes of AS. What the diagram cannot show is:

- how many more smaller ASes there are compared to the larger ones;
- more than a token number of connections per AS – in particular the amount of peering between smaller and medium size ISPs;
- multiple connections between the larger ASes – which improve resilience;
- multi-homed organisations – which would appear as various sizes of stub AS;

Obviously one of the difficulties when considering the resilience of this system is the sheer number of ASes and the number of connections between them. However, we can identify a core part of the system, which carries the majority of Internet traffic, and that is shown as the shaded part of Figure 28. The Internet does not formally have a backbone, but this core appears much like a backbone. This 'virtual backbone' comprises a relatively small number of regional and global ASes, along with the CDNs and the IXPs. This view of the system may offer a tractable model when considering its resilience.

3.5.11 Relative Scale of ASes

CAIDA maintain a ranking of ASes according to the size of their customer cones [**82**], and [**83**] describes the methodology. The significance of the customer cone is that it is the proportion of the Internet that a transit provider can reach on its own and via its customers, without sending traffic to a peer or a transit provider (of which it is a customer). The following uses the CAIDA data, as of mid-November 2010:

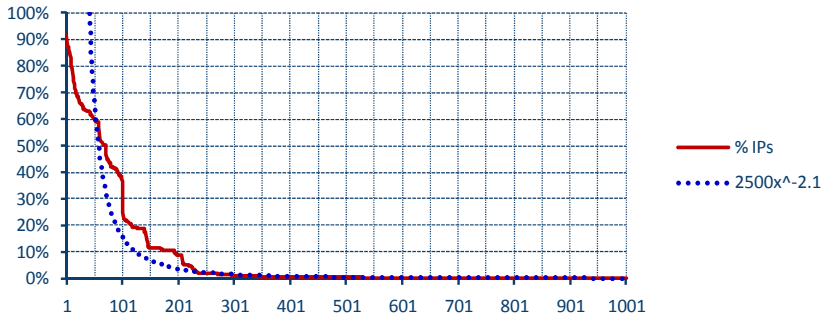


Figure 29: Top 1001 ASes by IPv4 Addresses in Customer Cone – Source CAIDA AS Rankings

which shows the percentage of all allocated IPv4 addresses which the top 1,001 ASes have in their cone. There are a further 35,000 or so ASes in the tail to the right of this. This means that each of the top 10 can reach 79% or more of the Internet as customers (or customers of customers, etc.), each of the top 20 can reach 68% or more, the top 70 50% or more, and so on. This supports the view that the 'core' of the interconnection system consists of a relatively small number of networks. (The power-law $2500\ x^{(-2.1)}$ is not a good fit, but is suggestive.)

The CAIDA data also provides the "AS Degree", which for the largest ASes is effectively the number of direct customer ASes each has:

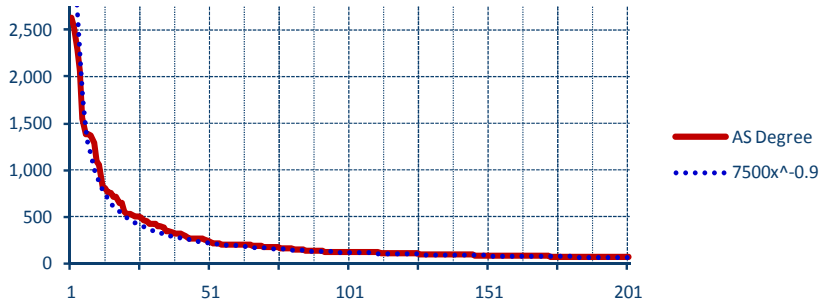


Figure 30: 'AS Degree' for the Top 201 ASes – Source: CAIDA AS Rankings

which shows that even the very largest networks connect to less than 8% of all ASes. The power-law $7500\ x^{(-0.9)}$ has a suggestive fit.

However, the CAIDA data also demonstrates the issue of information hiding. For AS2529 (which the authors of this report know well) the CAIDA data shows only 2 peers, which is something of an underestimate, since AS2529 peers extensively at LINX. AS2529 does not, however, connect to the

RIS Route Collector, so its peering connections are less likely to be visible[28]. The AS degree being shown here appears to be missing what we would expect it to miss, namely most of the peering amongst Tier 2 and Tier 3 networks.

This view of the interconnection system is flawed because it assumes that every IP address is of equal importance to everyone. If the ranking had something to do with traffic it might overcome that issue, except that would require multiple rankings, because different traffic matters to different people. Renesys have their own measurement systems and provide information to their customers about how well those customers' connections across the Internet are working. They also occasionally publish their rankings for the top few networks. Their methodology also counts how much of the address space each network reaches via customer-provider connections, but they have a weighting scheme for address space. Their methodology is described in [**84**], in which they also note that "*globally valid, representative traffic data is non-existent*".

For the top 21 networks according to CAIDA and the top 13 networks according to Renesys are:

Rank			Network		Rating		
By IP	By AS	Renesys	AS	Name	IP %age	AS %age	Renesys
1	1	1	3356	Level 3	92%	93%	100
2	6	3	1239	Sprint	88%	86%	62
3	7	13	209	Qwest	87%	86%	20
4	5	2	3549	Global Crossing	87%	87%	64
5	2	9	7018	AT&T	86%	89%	36
6	3	4	701	Verizon	85%	89%	42
7	4	12	174	Cogent	85%	87%	23
8	9	-	4323	tw telecom	83%	81%	-
9	8	-	6939	Hurricane Electric	80%	81%	-
10	10	6	1299	Telia	79%	81%	41
11	11	7	2914	NTT	78%	80%	40
12	12	8	6453	Tata	76%	78%	37
13	16	10	3257	Tinet	74%	73%	34
14	13	5	3561	Savvis	73%	77%	41
15	14	-	9002	ReTN	72%	75%	-
16	15	-	1273	C&W	71%	75%	-
17	18	-	6461	AboveNet	71%	71%	-
18	17	-	19151	WV FIBER	70%	71%	-
19	23	-	3320	Deutche Telekom	69%	67%	-
20	20	-	2828	XO	68%	67%	-
21	19	-	3491	PCCW	68%	70%	-

Table 1: Top 21 Networks in Nov-2010 – Source CAIDA & Renesys

This list is ranked by the percentage of IPv4 space in the network's customer cone as measured by CAIDA – the same metric that was used in Figure 29 above. The second 'CAIDA' rank is by the number of ASes in the network's customer cone, shown here as a percentage of all known ASes. The 'Renesys' rank is for the top 13 networks according to their methodology, from July-2010 [**85**]. The

[28] *AS2529 will not announce its peering connections to its transit providers or to its peers, it will announce them to its transit customers. AS2529 has a limited number of ISP transit customers, and they would be unlikely to connect to a RIS Route Collector. Even if one did, the route collector would only see AS2529 peering with AS8426 (say) if that was the customer's selected route for any address block in AS8426. It has already been noted that efforts to understand the AS level topology of the interconnection system tend to miss peering connections – this is an example.*

'Renesys' rank is for the top 13 networks according to their methodology, from July-2010 [**85**]. China Telecom is ranked 11th by Renesys but is not placed in the CAIDA top 21.

The CAIDA and Renesys ratings in Table 1 are normalised to 100 for the number 1 network. It is interesting that the two ratings are rather different, as shown:

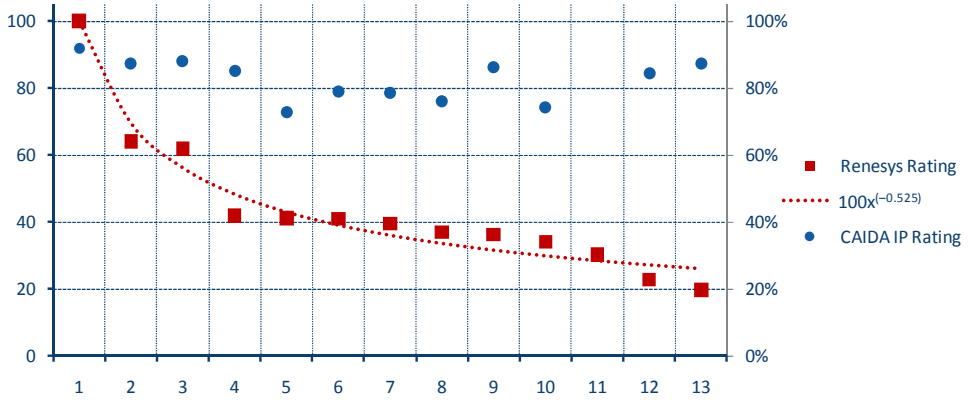


Figure 31: Renesys and CAIDA Ratings for the Renesys Top 13 – Source CAIDA and Rensys

The Renesys ratings suggest that even among the top 13 networks there is a significant difference in scale, while the CAIDA ratings suggest there is not a great deal to choose between these networks. The available data is interesting, but inscrutable.

In [**13**] Labovitz et al provide figures for all Internet inter-domain traffic share. For 2007 and 2009 the shares for the top 10 providers are:

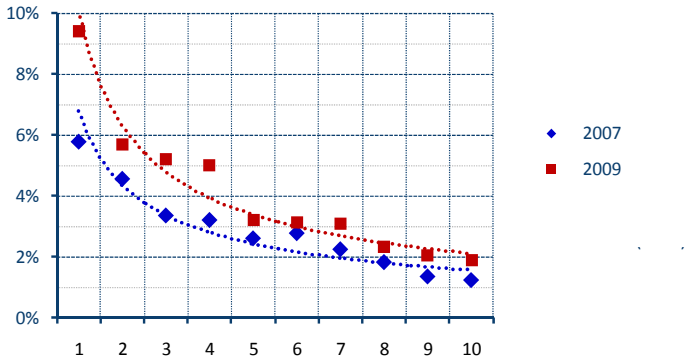


Figure 32: Top 10 by Percentage of all Internet Inter-Domain Traffic – Source: Labovitz et al

The two key points are that: first, the total for the top 10 in 2007 was 29% of all traffic, while in 2009 it was 41%.; second, that where the top 10 in 2007 were all major transit providers, in 2009 number 3 was a CDN. This underlines the shift towards the CDNs but also points to an increasing concentration of traffic in the larger networks. (The dotted lines are power-laws – but with this much data no great significance is attached to that.)

Labovitz et all also observe that in July 2007 the top 150 ASes were the origin of ~30% of all traffic, but in July 2009 the top 150 ASes were the origin of more than 50% of all traffic, again pointing to a concentration of traffic in a smaller number of ASes.

3.6 The Driving Force – the Commercial Imperatives

There is no central coordination of Internet interconnection. All ASes operate independently and in their own interests. But all ASes must establish a way to reach all other ASes and exchange traffic with any and all destinations their users and customers choose. This common incentive creates coherent activity which delivers the Internet, to everyone's mutual advantage.

All relationships between ASes are bilateral, each one entered into to further the commercial interests of the two parties. The commercial interests of the two parties are to maximise the difference between revenues and cost. An AS's revenues are derived from its users and customers. Its costs are incurred building and running its network, which includes the cost of all interconnections and any transit service it needs – recalling that the major component of transit cost is traffic volume related.

Peering arrangements are mutually beneficial and offer the most direct, and hence most effective, connections. However, for the vast majority of ASes it is only cost effective to peer locally. IXPs improve the economics of local peering.

In a peering connection the traffic exchanged is either to or from a user or customer, so the benefit appears to be the same for both, and one might expect ASes to peer with each other wherever possible. However, where the ASes are of markedly different scale:

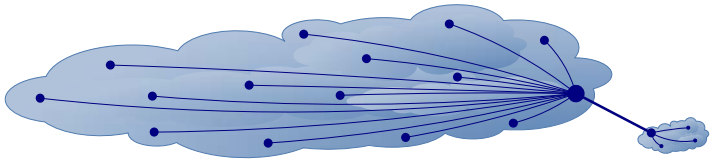


Figure 33: ASes of Markedly Different Scale

it is clear which one would be doing more work, and carrying more cost. Indeed, in a peering relationship there is the unspoken assumption that the two ASes are in some sense roughly equal (hence the name!), so that the arrangement brings roughly equal costs and benefits to each of them. This does not mean that peering only occurs where the ASes are strictly equal. In practice ASes will peer if they both believe that the benefit to themselves outweighs any perceived (or, indeed, actual) inequality in costs or benefits.

A transit provider will not peer with another AS if it feels it has a realistic hope of selling transit to that AS, or would compromise its negotiating position with similar customers or possible customers. This means that all but the very few Tier 1 networks must buy transit in order to reach the entire Internet. In many parts of the world there are several possible transit providers to choose from, and those transit providers compete fiercely with each other for business. So the price of transit capacity has fallen dramatically over the years, and continues to fall.

3.6.1 Commercial Imperatives and the Major Transit Providers

The large, global transit providers see the Internet rather differently to other ISPs. The objective for a transit provider is to sell transit to as many ASes as possible – the more ASes that are customers, the more of the Internet the transit provider is being paid to reach. To be a transit provider, however, it is necessary to reach the rest of the Internet, and to do so at minimum cost.

All large transit providers are in a similar position, in that they all have some proportion of the Internet as customers, and the rest they must reach via other networks – so one way or another

these networks must connect. The question is, will that connection be a peering connection, a transit connection, a paid peering connection or some other concoction? How this actually works is shrouded in confidentiality, but is said to resemble poker as much as it resembles business – see [**86**] for some insight into the process.

As discussed above, the largest transit providers are generally 'Tier 1' networks, and so peer with each other. In a world with five Tier 1 networks, consider the position of a major transit supplier (AS20) which is not, strictly, a Tier 1 network:

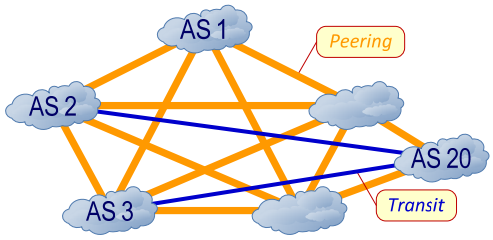


Figure 34: Near Tier 1

In this example AS20 is directly connected to all the Tier 1 ASes, except AS1, peering with all except AS2 and AS3, with whom it has transit arrangements. Those transit arrangements allow AS20 to reach AS1. From the perspective of the rest of the Internet, AS20 can be a perfectly good major transit provider. Traffic to and from AS1 (and its users and customers) passes through one extra AS, compared to transit bought from one of the Tier 1 providers, but that need not be significant.

Suppose that AS20 wishes to become a Tier 1 network. To do that it must, first, establish a peering arrangement with AS1. From its current position AS20 might persuade AS1 that it has nothing to lose by peering with AS20 – for AS1 their mutual traffic goes via existing peering connections, so from its perspective a peering arrangement would not reduce its revenue; in fact it could increase it! Because AS20 is connected to all the other Tier 1 providers, the only traffic which passes between AS20 and AS1 will be for AS1's unique cone. This means that where AS1 shares a customer with any of the other Tier 1 ASes, the other ASes will carry that customer's traffic to and from AS20. If AS20 were to connect directly to AS1, then some of that traffic could go via AS1, increasing AS1's revenue from its customers (and decreasing other ASes' revenues). Clearly AS20 is not going to buy transit from AS1, but it might enter into a paid peering arrangement (see 3.6.6 below). And peering with AS20 would take some revenue away from AS2 and AS3.

AS20 must then convert the transit arrangements with AS2 and AS3 into peering (paid or otherwise). Once AS20 peers with AS1, the connections AS20-AS2 and AS20-AS3 will carry the same traffic as if they were peering connections. AS20 could threaten to terminate the transit arrangements, individually or together, leaving AS2 and AS3 to worry about either increasing the other's revenue (if only one transit connection is actually terminated) or losing connectivity to AS20 (and its users and customers). Finally, AS20 must, at all times, avoid upsetting its peering arrangements with all the other Tier 1 ASes.

How these things actually work themselves out is shrouded in commercial confidentiality. Note that at no time is the resilience of the interconnection system a major consideration, even though connections at this level are a key component of that system. If AS20 succeeds in becoming a Tier 1 network, the traffic between AS1 and AS20 will flow over new, direct connections between the two, reducing the dependence on AS2 and AS3, and thereby improving the resilience of the system.

The driver for AS20 in all this is to reduce its transit costs to zero, at the cost of AS2 and AS3 losing some revenue. AS1, AS2 and AS3 will see some reduced load on their peering interconnections, which may be of benefit by delaying the need to upgrade.

In the discussion of 'hot-potato routing' (Section 3.1.10 above), it was observed that the traffic between two large ASes takes the form shown:

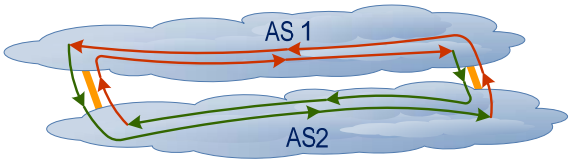


Figure 35: Hot-Potato Routing and Tier 1 Networks

where the long-haul traffic in AS2 is packets from AS1, and vice versa.

If the traffic to and from the ASes is roughly equal, then this arrangement is equitable. Suppose, however, that AS2 sends AS1 much more traffic than AS1 sends to AS2:

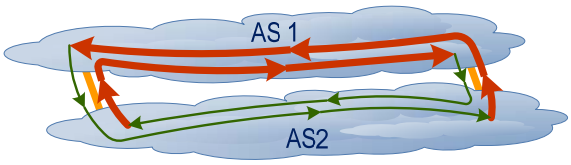


Figure 36: Hot-Potato Routing and 'Unbalanced' Tier 1 Networks

Now AS1 is doing all the heavy lifting – the cost of carrying traffic between the two ASes' mutual customers is largely being borne by AS1, which would be a factor in the commercial negotiation between the ASes. In principle the two ASes could arrange to do cold-potato routing, though then the other AS is doing all the heavy lifting, which may not suit either; and in any case cold-potato routing is, at best, hard work. We have seen that traffic engineering between ASes is difficult, and this is another aspect of the problem.

It is apparently common for peering agreements between the major transit providers to be formal, because they are of significant commercial value to the parties. These agreements apparently often specify a maximum allowable disparity in traffic flows. We say apparently because these agreements tend to be treated as trade secrets.

Occasionally a large network will reassess its relationships with its peers and decide that one of them no longer qualifies as a peer. When this happens it is generally said to be because, for whatever reason, the traffic between the peers is no longer sufficiently balanced. As part of the negotiation that follows, one party may de-peer the other, unilaterally, which leads to users and single-homed customers of both networks being cut off from each other. The matter may then be settled by whose users and customers complain most effectively.

Where two large networks peer it is in their interests to connect to each other in multiple locations, because that minimises the distance that traffic has to travel across their respective networks. Suppose two ASes are both present in some area and connect there. Traffic for their users and customers in that area would use the local connection, so:

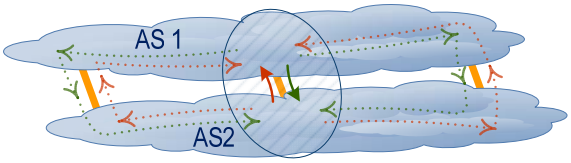


Figure 37: Value of Local Connections

where the area and local connection in question are circled in the middle of the diagram. If the local connection did not exist, then traffic would have to flow over the dotted lines in one direction or the other. The local connection saves both networks money. It will also improve the quality of the connectivity between the two networks, by reducing the distance travelled by some traffic.

In this case, the self interest of the two ASes improves the resilience of their interconnection, and hence the resilience of the interconnection system.

3.6.2 Commercial Imperative and the Small and Medium Size ISP

An ordinary ISP must buy transit from (preferably) two (or more) providers. It is then in business, and can provide Internet access and/or hosting services to its users and customers. The cost of delivering traffic is set by the cost of its transit connections, which is set by the cost of the connections to the chosen transit suppliers, and their charges for transit service.

The ISP must ensure that it has sufficient transit capacity to meet its users' and customers' day-to-day needs, and must also ensure that its service can survive one of its transit connections failing. The amount of spare capacity the ISP will pay for will depend on its view of the commercial risk of congestion occurring in the event of failure, and how long that congestion is likely to last. The risk is mostly to its reputation, since its SLAs will not cover congestion beyond the borders of the ISP.

With two transit providers the ISP could arrange for each connection to be able to carry 100% of its traffic[29], so that the failure of one should hardly affect the ISPs users and customers. But in normal running only 50% of the total capacity would be in use.

With three transit providers the ISP might be safe with 66% of its total capacity in use. But that assumes, first, that traffic divides equally between the three providers and, second, that when one connection fails, the traffic will divide equally across the remaining two. Neither of these assumptions is likely to be true in practice. Furthermore, as discussed above, only when a connection actually fails will the ISP discover how traffic will be redistributed.

[29] *Noting that 95th percentile charging helps here: if one connection fails and all the traffic falls onto the second provider, then provided the connection is fixed within 36 hours, the extra traffic will not increase the second provider's bill for that month.*

Any peering connections that an ISP can make may improve the quality of its connectivity, but its cost will be compared with the cost of transit (even if traffic exchanged with peers is thought to be worth more).

The low cost of peering at a local IXP is an important factor in most ASes' peering calculations. At an IXP, ASes of quite different scales may peer with each other. It costs essentially nothing to add another peer at an IXP, so no cost justification is required. Where it has nothing to lose, a larger AS may peer with smaller ones to improve its connectivity with local ISPs. (Noting that the larger AS will consider it has something to lose if peering with a given small AS compromises its ability to sell transit.)

Local traffic is a large part of an AS's total traffic, so peering with local ASes is an effective way of improving connectivity.

3.6.3 Commercial Imperative and the Large ISP

The large ISP must also buy at least some transit, in much the same way as the small and medium size ISPs. Some of the large ISPs may be able to obtain peering with some major transit providers. They are likely to peer amongst themselves to reduce their own transit costs – since they are unlikely to sell each other transit. Enough traffic may be exchanged between two such ASes to justify a private, direct peering link.

Whether the large ISPs will peer with any medium size ones will depend on all the factors discussed above, but with particular consideration being paid to the question of compromising the ISPs ability to sell transit.

3.6.4 Commercial Imperative and the Content Delivery Network

The content delivery networks (CDNs) have a somewhat different commercial imperative to the ISPs, of whatever size. First, the CDNs aim to minimise their transit costs. They can do this by placing copies of their content in facilities near to significant bodies of users, such as at local IXPs, where the CDN can peer with the local ISPs. This bypasses the transit providers, and reduces the distance the traffic has to be carried; everyone wins, except the transit providers.

However, the CDNs are also driven by quality. By placing their facilities close to the end users, and peering directly with as many end users' ISPs as possible, the CDN aims to maximise the quality of their service.

The service provided by the third party CDNs to their customers, the content providers, is not a direct substitute for transit. The CDNs maintain facilities in sites all over the world. There are obvious economies of scale here, and the CDNs can provide service more cheaply than its customers could achieve on their own. The challenge for the CDNs is to maintain margins and avoid having their prices tied to the price of transit.

3.6.5 Peering Policy

Whether one AS will peer with another depends entirely on how each AS perceives its self interest. Each AS will have a 'Peering Policy', which it may or may not publish. The policy may state criteria which a putative peer must meet in order to qualify as a peer or qualify to be considered as a peer. There is a wide range of peering policies, but the following are common:

a. 'Open Peering Policy' – the AS will peer with all comers (subject to the availability of capacity). This policy commonly applies at IXPs where the AS is already present. No AS is going to spend money on a direct peering connection, or a new IXP connection, without careful consideration. Unless an AS's connection to the IXP is fully utilized, it costs nothing to add another peer at the IXP, and will save something on transit. So, this is an obvious policy for ASes which have little or no expectation of turning potential peers into transit customers. Smaller ISPs have every reason to operate an open peering policy, as do CDNs.

b. 'Restricted Peering Policy' – the AS will peer with ASes it feels like peering with. The AS will consider each request for peering on its merits (as understood by the AS when the request is made). Note that the expectation here is that the AS will be receiving petitions from other ASes for a peering arrangement, and there is the implication that the petitioning ASes will be lower down the pecking order.

c. 'Subject to Criteria' – the AS will (probably) peer with ASes which meet the stated criteria. This is effectively a restricted policy, except with a published basis for judging the merits of a request. However, there is often a sense that the stated criteria are designed to ensure that a very small number of ASes will qualify. Of course, the criteria are not necessarily exhaustive and may be open to interpretation.

d. 'Closed Peering Policy' – which suggests the AS is not interested in entertaining peering requests. This may mean that the AS simply has no interest in peering at all, or that peering is 'by invitation only'.

An AS's peering policy may well vary from place to place. So an AS which is strong in one region may have a restricted policy there, but a more open policy elsewhere. In particular, an ISP may wish to apply a different policy at an IXP to ASes from outside its home region than it does to ASes from within that region.

This follows the classification in [**87**]. Peering policies are covered in more detail in [**88**], which covers the sorts of requirements that some networks expect peers to meet.

3.6.6 Paid Peering

So far, we have discussed common or garden 'settlement free' peering – there is a rarer form: Paid Peering, which is the same as ordinary peering, except one party pays the other, and may cover the cost of the link.

In the world of the Tier 1 and near Tier 1 networks, paid peering may be a mechanism to achieve, or hold on to, peering connections with other Tier 1 networks. This may also apply in the world of the Tier 2 and near Tier 2 networks. These are somewhat special cases, and there are not many of these networks in the world. Whatever the arrangements are, they are shrouded in confidentiality, and we will put those to one side for this discussion. More generally, it seems reasonable to expect paid peering to be a standard way of interconnecting, allowing ASes of different scales to connect directly,

and compensating the larger ISP for carrying a greater part of the costs of the traffic. But paid peering is rare, but on the increase [**89**].

The problem is that the alternative to paid peering is transit. In practice, where ISPs are of sufficiently different scales that paid peering might be an option, the larger ISP will consider transit to also be an option (except, perhaps, in the rarefied world of the largest ISPs.) A medium or small ISP might wish to peer with a large or global ISP, but the latter are transit providers.

The great majority of ISPs are medium and small ISPs, and they peer with other local ISPs, generally at an IXP, where the arrangement is a straightforward benefit to both parties. ISPs of quite widely varying scales will peer at an IXP, except where one ISP has expectations of being a transit provider to the other ISP, or ones like it.

The benefit to the smaller ISP in a proposed paid peering arrangement is obvious: it would expect to get a material proportion of its traffic at lower cost, and more directly, than it does via its existing transit arrangements. There would be no point proposing paid peering if the traffic were not substantial: the absolute saving would not be great, and the cost of the arrangement would probably exceed it. But the incentive on the larger ISP to consider a paid peering proposal is very limited. The amount of traffic involved will be small compared to the ISPs total traffic, so any improvement due to the direct connection with the proposed peer will not be material. The overriding consideration is the larger ISP's negotiating position on transit. If it accepted a paid peering proposal from one potential transit customer, it could compromise its position in the market. Whatever small revenue it might get from a paid peering arrangement must be weighed against the immediate and future possible loss of full transit revenue. So paid peering is unheard of where transit is a possible alternative.

For the CDNs, where transit not a factor, we see a different dynamic. The CDNs have an obvious interest in peering with as many 'eyeball' networks as possible, and indeed vice versa. From a traffic perspective, the decision to peer is straightforward, both parties save money. However, for the eyeball networks the traffic is only part of their total, where for the CDNs it is all of their traffic. For the CDNs, quality is also an issue. The CDNs revenues depend on providing a quality service to end users. A large eyeball network may look the CDN in the eye, suggest that they be paid to provide direct access to their customers, and see who blinks first. In [**90**] it is reported that in such negotiations the CDNs are paying for peering.

Paid peering raises the issue of "Network Neutrality". If an eyeball network is paid to accept some traffic, will it give special treatment or preference to that traffic across its network? If so, is that an issue? Is it more of an issue if the eyeball network excludes or degrades similar traffic that is not paying the special toll?

For a long time ISPs have looked at some web services, particularly those which generate a lot of traffic, as free-loaders: these businesses are seen to make money by dumping more traffic in the ISPs' networks, without making any contribution to the cost of upgrading those networks to cope. At the same time, market pressures prevent the ISPs from recovering these increased costs from their customers. So, for some, paid peering with CDNs is finally a mechanism to, in their view, redress the balance.

3.6.7 The Value of Traffic

In a peering relationship it is generally held that the traffic between the two ASes should be roughly the same in both directions. For peering between very large networks, we have seen above how that holds, at least for straightforward hot-potato routing.

But in general, from a cost perspective, it does not matter if a link is full in one direction and empty in the other – the cost is dictated by the peak traffic, no matter which direction it is going in.

The deeper issue is that traffic, and particularly relative traffic volume, has no discernable value. When a customer in one AS visits a web-site hosted in another, then relatively small amounts of data are sent to the web-site and relatively large amounts are sent in return. So the traffic is unbalanced, but:

a. suppose the web-site is a information service: we might assume that the traffic from the web-site to the user is more valuable, as well as being the greater volume. Indeed, in the telephony world the user might have paid for the call, and in some cases the service might be provided on a premium rate number;

b. suppose the web site is an online shop: now we might assume that the traffic from the user to the web site is the more valuable – it may, in fact, carry payment information! In the telephony world, the shop might well use an 0800 number.

In the telephony world, the direction of the value flow is implicit in the way the call is handled, and that explicitly establishes the direction of the flow of money. But on the Internet, all 'calls' look the same. It is simply impossible to ascribe meaningful value to any individual packet, and impossible to ascribe value to packets in one direction or the other, or to the volume of traffic in one direction or the other.

This is discussed in [**9**] and [**91**] in the context of the negotiation between potential peers.

3.6.8 Metcalfe's Law

Suppose we have a network which connects *n* users. That network supports approximately $\frac{n^2}{2}$ connections. Generalising wildly, let us therefore say that the value of a network is proportional to the number of users squared – this is known as Metcalfe's law[30].

We can use this to look at the issue of whether one network will consider peering with another.

Taking two networks of differing size, they are individually 'worth' n^2 and N^2, and connected together the result is worth $(N + n)^2$. If $N = 2{,}000$ and $n = 100$ then separately the networks are worth 4.01 million something-or-others, but together 4.41 million. The problem is not whether the combined network is worth significantly more, it is the relative improvement from the two networks' points of view. In this case the larger network sees an ~10% improvement, the small network sees

[30] *Andrew Odlyzko and others have made the case for the value of the network being proportional to n(log n), not n² [**220**]. This is still super-linear, so the basic argument holds, but the detail changes.*

an ~440-fold improvement! Of course, as the difference in scale increases so does the disparity in benefit[31].

The reluctance of larger ASes to peer with rather smaller ASes is, perhaps, understandable in these terms. The keenness of the smaller AS to peer appears equally understandable.

The larger AS will, no doubt, happily offer transit, at a price. From its perspective, adding another small peer makes little difference, so there is no incentive to forgo a possible transit customer. The smaller AS will probably buy transit from an AS that the larger either peers with or sells transit to (directly or indirectly), so peering offers no cost advantage to the larger AS. Furthermore, there is every incentive not to invite every small AS to request peering.

The same analysis for two large networks points in a different direction. Taking For N = 2,000 and n = 1,000, the two networks see a 2.3- and a 9.0-fold improvement; for N = 2,000 and n = 1,500, it is 3.1- and 5.4- fold. So, even though one benefits more, both benefit substantially[32].

3.7 Responsibility and Resilience

Returning to the earlier small scale internet, assume that AS2529 buys transit from AS10 and AS1, and AS4321 buys transit from AS20 and AS1. Assuming all the transit providers peer with each other, we have:

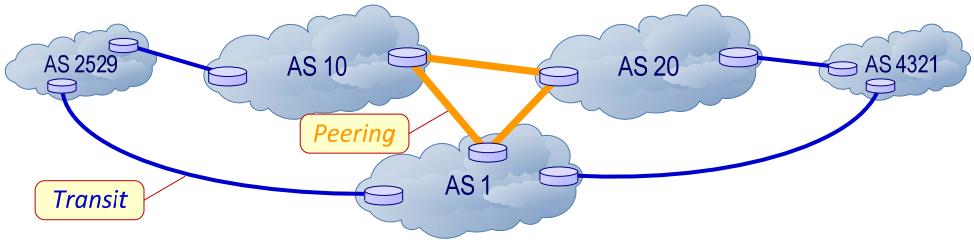


Figure 38: Responsibility when ASes are Closely Connected

In normal running, some proportion of AS2529's traffic may go via AS10, and the rest via AS1. AS10 will then forward that traffic to AS20 or AS1, who in turn will forward it on to AS4321. (AS1 is paid for the traffic it passes to AS4321, so AS1 will not pass any of it to AS20. Similarly AS20 will not pass traffic to AS1.) Similarly traffic from AS4321 to AS2529 will go first to AS20 and/or AS1, and hence via AS10 or directly to AS2529. (At each step, the decision on which path to use is made by the forwarding AS's routers – the receiving AS has no say in the matter.)

All the transit providers are paid by AS2529 or AS4321 to carry traffic in either direction – AS1 is in the happy position of being paid by both. So there are obligations on all the transit providers to look after the traffic between these ASes. These obligations mean that it is in the transit providers' mutual interests to ensure that the peering connections between them are effective.

[31] *Using n(log n) the larger network still sees ~10% improvement, but the smaller one sees a more modest ~35-fold improvement – still very different.*

[32] *Again, using n log n: 2,000:1,000 gives 1.6- and 3.5-fold improvements, while 2,000:1,500 gives 1.9- and 2.6-fold. So, as might be expected, n log n gives more moderate results.*

However, the SLAs which each transit provider offers do not formally cover traffic beyond their own network, so once traffic reaches the peering links between AS1, AS10 and AS20 not one of them takes formal responsibility for it. All things being equal, it would be best for AS2529 and AS4321 to use the route via AS1 when sending each other packets. That route has the shorter AS Path.

To see what happens when ASes are a little further removed, let us consider AS64500, which buys transit from AS4321 and AS1234, thus:

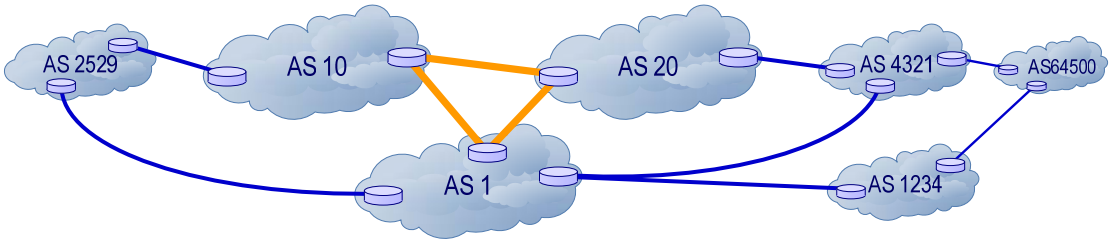


Figure 39: Responsibility when ASes are Less Closely Connected

Now, AS4321 and AS1234 have obligations to AS64500, and in turn AS20 and AS1 have obligations to them – the transit customer/provider relationship is itself transitive. Of course AS20 has no direct obligation to AS64500, so when buying transit from AS4321, AS64500 must (somehow) ensure that AS4321 is making proper provision for its traffic. The more ASes a given path crosses, the less direct interest the ASes in the middle part of the path have in the traffic – shorter AS Paths do not necessarily translate to shorter physical paths or more effective paths, but absent any other information, it makes sense to prefer shorter AS Paths, where there is a choice. In this way all ASes in all paths to and from AS2529 have an obligation, directly or indirectly, to either the sender or the receiver. So the pattern of transit and peering connections and their supporting commercial arrangements keeps the Internet running.

However, note again the limitations of the SLAs offered by transit providers. Even where the transit provider AS4321 has an SLA from AS20 at least to the edges of AS20's network, AS4321 does not extend its SLA that far. This is not entirely unreasonable. AS4321 also has an SLA from AS1, which may be different in some respects, and it is not really possible to predict

So what happens in the event of a failure? Suppose the transit connection between AS4321 and AS20 fails:

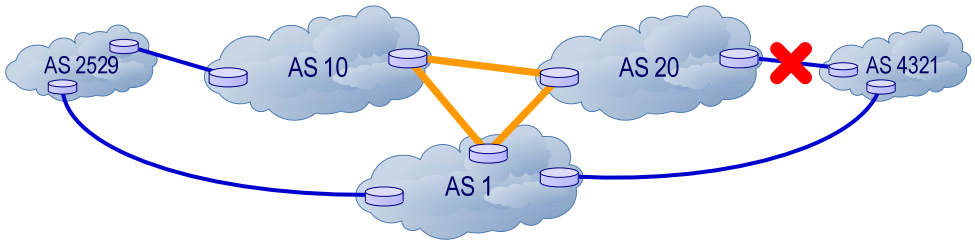


Figure 40: Responsibility in the Event of Failure

Now all the traffic that was using that connection must use the remaining transit connection to AS1. The first issue is whether AS4321 has made sure that it has sufficient capacity to and on that transit connection. Assuming it does, the next issue is whether AS1 has sufficient capacity to absorb the extra traffic, wherever it may go across AS1.

AS1 will be managing its network to meet the usual demands, with some margin to absorb the unusual. AS1 has no way of telling what extra capacity will be required if something goes wrong in its

customers' networks or its customers' customers' networks. So AS1 will use rules of thumb and previous experience to estimate how much spare capacity to maintain, with a reasonable expectation of not being caught out. It is likely that the entirety of AS4321's traffic will be relatively small compared to AS1's. It is likely that AS4321's traffic patterns are not hugely different from other customers' traffic patterns. So a relatively small percentage of spare capacity across AS1 may well absorb the extra traffic, and the Internet will do its job and cope.

The web of transit and peering arrangements forms a system of commercial and economic incentives to develop and maintain the underlying system of interconnections, and is a vital layer of the Interconnection Ecosystem. However, the formal SLAs that go with the commercial arrangements do not extend to the interconnection system. It is not clear what happens if large amounts of traffic are displaced from their usual paths between ASes, for whatever reason. Nobody is paying the larger transit providers to maintain enough spare capacity to compensate for large scale failures in other ASes or their interconnections. In fact, nobody can really predict what amount of spare capacity a large transit provider should have against such an eventuality.

The mechanisms that keep the Internet running on a day-to-day basis, coping with the usual round of unexpected failures, do not necessarily prepare it for large scale failure. Nor do the transit providers offer any formal undertaking that extends beyond their borders.

3.8 Mapping the Ecosystem

In order to assess the resilience of the Interconnection Ecosystem it would be useful to be able to map it and measure it [**92**]. But each layer of the Ecosystem presents its own problems:

a. the physical infrastructure. To understand the system of interconnections we need to be able to map the clusters of sites, the networks within those clusters and the networks between them. But some of this information is considered commercially confidential, and much of it is not published because to do so is thought to give too much information to anyone who might wish to damage the infrastructure. For many purposes, however, it is sufficient to be able to identify where bits of infrastructure are close to each other, without needing to know exactly where they are.

b. the peering and transit connections. For each peering and transit connection the map should show where ASes are connected, the capacity of each connection, and whether the connections are peering or transit connections (or something more exotic). Multiple, separate connections between ASes are important for resilience – any attempt to assess resilience without knowing about them is incomplete. But attempting to map the connections between ASes by looking at route collector data or using traceroute is problematic, as described in Section 3.1.7 above. Information stored in Internet Routing Registries (IRRs) is inherently incomplete[33]. It may be possible to form a partial view on whether ASes are interconnected, but this all tells us little about how the ASes are interconnected (in how many locations and at what capacity). It may be possible to deduce whether a given connection is for peering or transit, but that is not

[33] *Information in IRRs is secondary – the primary information is how each router in an AS is configured. There is no general requirement to publish routing information in an IRR, nor is there any general need for it to be up to date or accurate.* ***[240]***

guaranteed. The locations and capacities of interconnections may be deemed commercially confidential.

c. the distribution of traffic flows. Mapping how things are connected is useful, but what really matters is where traffic is actually flowing, and in what volume. How those traffic volumes might respond to the loss of some part of the system is of no small interest. Very little is known about traffic flows, and data about them will be commercially sensitive [**93**] [**94**].

d. the commercial and operational arrangements which keep those connections running. The response of the system to large scale failure may depend on the effectiveness of these, so knowing something about them would be useful. The very existence of some transit and peering arrangements is deemed commercially sensitive, the form of those arrangements more so.

Then, of course, there is the sheer scale of the Internet and the number of connections between ASes is an obvious problem. We have to ask whether it is practical to consider any approach to resilience that depends on accurate or complete mapping. If not, then what approach can be usefully taken?

3.8.1 On Topology

There is a great deal of interest in the topology of the interconnection system. A lot of academic effort has gone into attempting to capture accurate views of the topology of connections between ASes, and to draw inferences from those. Modelling connections between ASes as a graph, and then applying graph theory techniques to those models has also been popular.

As discussed in Section 3.1.7 above, the best information available on connections between ASes is incomplete. See: [**95**] which bemoans the "*very weak theoretical foundation for Internet topology modelling*"; [**96**] which observes that BGP-derived AS maps get a reasonable view of the major transit providers, but miss a large number of peering relationships between non-Tier 1 ASes; [**97**] finds a lot more peering relationships than other studies, a lot of them at IXPs. For an excellent analysis of the weaknesses of the graph theoretical approach see [**98**].

As well as trying to discover which ASes are connected to each other, some of the topology work also attempts to infer type of relationship between the ASes: transit or peering. This gives some insight into the commercial layer as well as the network routing layer. [**99**] discusses this.

From a resilience perspective, however, these AS topologies are not enough. First, because they tend to miss peering relationships, they underestimate the diversity of (predominantly) local connections between ASes. Second, these topologies completely ignore the fact that some ASes connect to each other in more than one place, which are particularly important for the connections between the larger ASes. More useful are what is known as 'Router-Level Topologies', which seek to find ASes' border routers and the connections between them. Attempts to do this are described in [**100**], [**101**] and [**102**].

When we talk about mapping the ecosystem we mean a greater level of detail than router-level topologies. To assess resilience the physical location of the border routers and the links between them is required. So called 'PoP-Level Topologies' are router-level topologies where there routers are identified as being located in a particular site. An analysis of resilience may then be done in which complete sites, and all links in and out of the site, fail together, which is not as precise as a full physical map, but might be sufficient.

All of this activity, using combinations of different data and different probing techniques to try to discover the relationships between ASes, is necessary because ASes do not publish the data, because it is commercially sensitive.

3.9 The Problem of Value

Unlike telephone calls, exchanges of data across the Internet are not accompanied by an exchange of money. Packets travelling across the Internet have no marginal value to the networks that carry them. It is true that within an ISP's network the operator may give some packets greater priority, and charge for the service – differentiated services are valuable to the ISP world, and it is possible to extend such services between ISPs by implementing special forms of interconnection. However, on the open Internet it is essentially impossible to ascribe a value to each packet or hope to collect revenue on a per-packet basis.

Packets do have value at the end points either to the sender or the receiver or both. So an ISP can charge its own customers for the service of providing access to the Internet and for carrying traffic to and from anywhere in the Internet.

ISPs do not charge end users for traffic, they charge for capacity. This approach side-steps various issues: first, the issue of which packets the end user feels they should pay for; second, the fact that the end user does not entirely control the volume of traffic sent or received; third, that there is no guarantee that all packets will be delivered; and fourth that accounting for every packet would create significant work.

The economics of an ISP network are based on the aggregation of traffic from many end users, who do not, generally, use their entire capacity all of the time. An ISP offering, say, 10Mbit/sec ADSL service to 100,000 users does not expect to deal with 1,000Gbit/sec of traffic – the ISP might work on the basis of 2% of that, so arrange for their network to cope with 20Gbit/sec peak traffic. The mechanics and economics of access networks are outside the scope of this study, but it is obvious that if third parties encourage an ISP's end users to use more of their capacity, more of the time, then those end users may be dissatisfied and may demand that the ISP spends more on their network.

The costs of constructing and running an ISP network are nearly all fixed costs, irrespective of how much traffic the network then carries. So the cost of the network is linked to the peak demand. Transit pricing is linked to peak traffic in each month. Some ISPs will apply so-called 'fair use' restrictions on their end users' traffic in order to limit the peak demand, and manage their costs.

The current debate over 'Network Neutrality' raises the question of what, exactly, the end user is buying when they pay for Internet Access. As discussed in [**103**], as soon as we start to consider what giving some traffic preference means, we start to realise that we have taken for granted what normal service levels are. We may have assumed that normal service means:

a. equal service for all destinations and all types of traffic. All destinations are not equal, so more realistically we must be content with there being no discrimination between destinations or types of traffic. (The 'traditional' view that all traffic was 'best efforts' traffic corresponds to this [**104**].)

b. that a *'N'*Mbit/sec service will work at that speed all day, every day. ISP networks are not designed to do this [**105**]. – if they were, they would be impossibly expensive and mostly empty.

The problem is that whatever the limitations are on service, ISPs tend not to shout about them and users may well not ask. Comcast adopted a 250GB/month limit for its residential customers [**106**] in 2008, stating that "*the median monthly data usage by our residential customers is approximately 2 - 3 GB*". But this was not universally popular [**107**] [**108**] [**109**], despite being a rather higher limit than other providers.

Some ISPs have chosen to implement restrictions on some sorts of traffic. It is the discrimination against different types of traffic which is the essence of the network neutrality issue. In order to discriminate, an ISP must be able to identify which packets carry what sort of traffic [**110**], which is not entirely straightforward (the contents of a packet may be encrypted, for example), leading to a possible 'arms race' between the ISP and its customers [**111**]. P2P traffic is a favourite target [**112**] for several reasons: first, heavy users of P2P file sharing can be using the full capacity of their connection most of the time; second, P2P traffic is reported to be 50% or more of total traffic; third, the feeling that some proportion of P2P traffic is not entirely legitimate. Comcast had run into difficulty with some of its customers over its traffic management of P2P [**113**]; a class action against them has been settled [**114**] at a cost of $16M.

The EC recently conducted a public consultation on "Open internet and net neutrality" [**115**], which reported on 9 November 2010 [**116**] and found instances of throttling and blocking of P2P and VoIP, noting that:

> *"However, it appears that many of these issues were resolved voluntarily, without any formal proceedings, although some such practices still remain".*

The shift towards the CDNs delivering large volumes of traffic direct to the ISP that serves the end users puts some content providers in direct contact with those ISPs. That may give those content providers both a technical and a commercial advantage [**117**]. Where the ISP negotiates a paid peering arrangement, the advantage may be stronger. The ISP's end customers may feel that they fees they pay for service entitle them to equal access to anywhere in the Internet, but if some content providers are buying improved access to them, whose interests are being served? End users who access the preferred content are better served, at no extra cost to themselves. End users who do not access that preferred content may want to be reassured that their service is not being adversely affected, which will require expectations to be set and met. The report on the EC consultation touches on this, looking to the future:

> *"A number of respondents pointed to managed services, such as internet protocol television (IPTV), as an area that could present difficulties. For example, some content providers voiced concerns that network operators could favour certain services over others, to the detriment of competition and innovation."*

Many businesses now use the Internet, extracting value from the ability to reach their customers, and exchange ever increasing amounts of data with those customers. The ISPs which transport that data are a key part of that value chain, but have no way of tapping into any of that value, other than the market price for transport – which is tending towards zero. With the CDNs increasingly bypassing the transit providers, the question of who extracts value from the network is being posed in new terms [**118**]. The report on the EC consultation touches on transit and peering, stating:

> *"There is general agreement that the commercial arrangements that currently govern the provision of internet access (question 10), such as peering arrangements and paid transit, have worked well until now. However, opinion is divided on future approaches. A number of respondents cite inefficiencies in the two-sided market and advocate a new business model for the internet that takes account of advances in broadband technology and enables innovation in the area of managed services. In*

contrast, content providers are concerned that a change in market structure that leads to their being charged additionally for network access would invest operators with too much power and would represent, according to a few respondents, a 'tax on innovation'. Consumer organisations also state their concerns about the market power of large operators. BEREC agrees that the current arrangements are adequate, but notes that market developments need to be monitored to ensure that regulatory interventions can take place in the future if the need arises."

Finally, ISPs carry traffic on the open Internet on a 'best-efforts' basis. This is jargon for: "there is no guarantee that any given packet will reach its intended destination". One way of looking at this is that every packet is treated as having the same value – "value unknown". In normal running, which is most of the time, the Internet delivers packets very reliably. In abnormal times parts of the network may be congested, and packets will be lost, and the service offered by the network will be degraded, but not necessarily fatally. This is an essential property of the Internet. It is not obvious how a network the size and diversity of the Internet could be made to offer hard guarantees of packet delivery. It is, however, obvious that harder guarantees come at ever increasing cost. When considering the resilience of the system, these considerations should temper expectations.

3.9.1 P2P Traffic

The ipoque Internet Study 2008/2009 [**119**] analyses Internet traffic in eight regions of the world. Their figures for the proportions of total traffic in late 2008, early 2009, are:

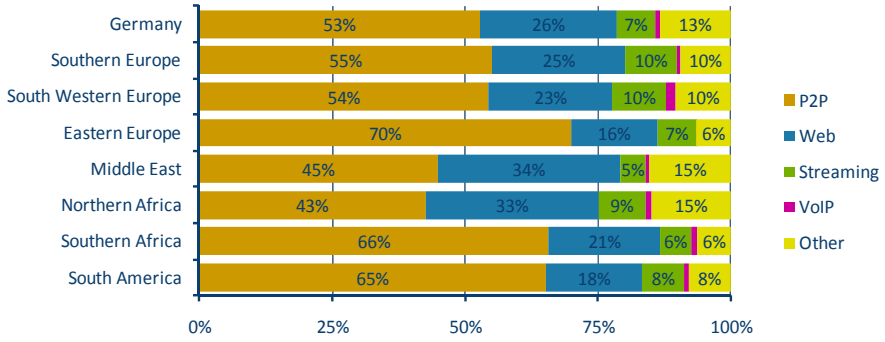


Figure 41: Proportions of Traffic Types, 2008/2009 – Source: ipoque

which shows that in Europe a little over 50% of all traffic in their sample was P2P. Cisco [**14**] reported 37% of all traffic as P2P in 2009. On the other hand [**120**] gives figures taken from packet-level monitoring of 20,000 residential DSL lines at a major European ISP, where P2P traffic is 14%-25%. Where there is some consensus, however, is that P2P traffic is falling as a proportion of the total, and web (HTTP) traffic is increasing, and that increase is mostly video traffic. Between 2007/2008 and 2008/2009, ipoque show P2P falling by ~24% (from 69% to 53%), and Web traffic increasing by ~80% (from 14% to 26%), in Germany.

This illustrates a number of issues. First, the available data is fragmentary and not entirely consistent. Second, the Internet is a big place and usage varies from place to place and across different types of user.

3.10 Regulation

Regulation is the final layer of the Internet interconnect ecosystem. By and large there is no regulation of Internet interconnections or of Internet operators. The Internet is held up as an

example of what a free market can achieve when unfettered by regulation. Internet insiders generally believe that any external regulation would be bad for the Internet.

Up to the early 1990s the Internet was a Government funded or subsidised initiative. The deregulation of telecommunications markets played a part in the privatisation of the Internet during the 1990s. Carriers that emerged from deregulation became key players in the Internet, and by 1997/1998 there were the 'big three', Worldcom (née UUNet and ANS), MCI and Sprint, who were together deemed to be the majority of the 'Internet Backbone' – although there were at least three times as many other significant networks: PSINet, GTE, et al. The speculative boom that drove the industry for the last half of the 1990s burst in the first quarter of 2000.

It is clear to everybody that the consolidation of market power into a small number of hands would be a bad thing, and regulators have acted to prevent that. To avoid the merger of Worldcom and MCI being vetoed by the competition authorities, MCI sold its Internet infrastructure and wholesale customer base (to Cable and Wireless). In late 1999 a merger between MCI Worldcom and Sprint was initiated, but abandoned in mid-2000 as a result of scrutiny of the merger by regulators and following the bursting of the 'dot-com bubble' in early 2000.

At present there are perhaps a dozen major global transit providers, and several dozen large regional/semi-global ones. So, largely without regulation, a reasonably diverse and fiercely competitive market has developed – via a speculative boom and spectacular bust.

The EC Directive on Interconnection in Telecommunications is designed to ensure that new entrants to the telecommunications market cannot be frozen out by existing businesses either refusing to interconnect or doing so at uneconomic rates. The directive provides a right to peer at market cost. However, because anyone can buy transit from a range of competitive providers, this right has not been relevant to Internet interconnection [**121**].

While there is a competitive market in transit service, the business of peering is clearly an area where market power is key. Smaller networks would reduce their costs if larger networks agreed to peer with them, and more peering could improve diversity of interconnections and contribute to greater resilience of the system. A regulator might see all that as beneficial. On the other hand, the larger networks would lose some of the revenue which supports the infrastructure that carries the Internet around the world. So the larger networks (apparently) live in fear of regulation that might require more open peering policies.

The 'Subject to Criteria' peering policy is a response to the fear of regulation. The theory is that if the largest networks are seen to decide peering requests in an apparently arbitrary manner, then they will invite calls to the courts and/or the regulators to overturn claimed predatory or discriminatory behaviour. This is a particular and justified fear in the context of what is known as de-peering, where one party decides that the other no longer qualifies as a peer, and the peering arrangement is terminated. De-peering has an impact on the former peer's costs and their standing in the industry.

So, 'Subject to Criteria' peering policies are intended to show that the operator is behaving in an open and even-handed manner when deciding whether to start or end a peering arrangement. In any case, the spectre of regulation can be invoked when a large network expresses its deep regret that, despite the excellence of the petitioning peer, it cannot afford to make an exception to its rules, however much it might wish to in this case.

Another example where regulation and the threat of regulation might have a chilling effect on peering is the case of large (say regional) ISPs and smaller ISPs from outside the region. In this case,

both can benefit from a peering arrangement, and since the larger has no prospect of selling transit it has nothing to lose. So a large European ISP may happily peer with a small ISP from Latin America, particularly if the Latin American ISP appears at one of the larger European IXPs, and is therefore covering the cost of carrying the traffic! This improves diversity of connections and hence contributes to resilience – the only losers are the global transit providers who would otherwise carry the traffic (though one of their number may provide the circuit used). However, it is clear that the large European ISP is discriminating against other European ISPs of similar scale to the Latin American peer. If competition lawyers had reason to be concerned about that, it could eliminate a generally beneficial form of peering arrangement.

In the mid 1990s the Internet was largely centred on the USA. For almost everybody else, connecting to the Internet meant first buying a circuit to the USA, and second buying transit in the USA – or buying transit from somebody locally who had already done that. Many outside the USA were upset that they were not only carrying the large cost of international circuits, but also paying for the privilege of exchanging the traffic. – see [**122**] for a view from 2000. The cost of connecting to the Internet outside of the USA was significantly greater than within it. This is in contrast to the long standing, regulated arrangements for international telephony, in which costs and revenues were shared between source and destination. Some argued then that the Internet should be governed by similar arrangements, but without success. But, the rapid fall in the cost of international circuits and of transit, and hence the ready availability of low cost transit in most parts of the world, has largely killed the issue except for developing countries [**123**]. The ITU[34] Study Group 3 has been looking into this [**124**], and came up with Recommendation D.50 [**125**] in Oct-2000, to which they added an appendix [**126**] in June-2004 and submitted it to the first Internet Governance Forum (IGF[35]) meeting in Athens in October-2006. Recommendation D.50 was updated again [**127**] in October-2008. For a moderately up to date view see [**128**]. Whether anything will come of this, and whether international regulation would have made, or may make, any difference is open to debate.

In [**129**] the authors consider the events of 9/11, and among other things note that "*Deregulation falsely raised expectations of users of the ability to have resiliency in services by using different carriers when in fact many carriers share the same core network, conduit or co-location facilities.*" This captures two issues: first, that even when regulators deregulate, and create competition and lower prices, there are unintended consequences; second, that shared physical infrastructure is a problem. The regulators concentrated on the market issues, but did not appreciate the significance of lower layers of the system.

Thus far, the Internet has thrived with little regulation. Internet insiders tend to believe that it has thrived because there is little regulation.

3.11 Summary of the Ecosystem

In this section we have described the components of the Internet Interconnection Ecosystem, covering the essential features of each one and how they form layers of systems which together

[34] *The ITU has no real rôle in the Internet scheme of things, but feels that this is an accident of history that should quickly be repaired. See also the ITU "IP Policy Manual" [**228**].*

[35] *http://www.intgovforum.org/*

comprise the Ecosystem. For each component we have touched on the issues which affect the resilience of the Ecosystem. Those issues will be covered further in later sections.

The layers of the Internet interconnection ecosystem are:

1. the physical infrastructure – clusters of sites, with fibre infrastructure within and between the sites, and the fibre networks that connect those clusters. Although the clusters make good cost sense, they may reduce resilience as many apparently independent ASes may share some physical infrastructure, often without being aware of it.
2. the peering and transit connections – the 'network layer', implemented over the physical infrastructure. The different sorts and scales of network are interconnected in a variety of ways, and exchange routes and traffic. That builds up to allow every part of the Internet to reach every other part, and send packets back and forth. The network layer has two distinct parts:
    a. routing (reachability): the 'BGP mesh' and its ability to distribute routing information;
    b. traffic: the distribution of traffic flows – the dynamic view of the system.

    The purpose of the system is to carry traffic so the volume and distribution of traffic are key (though largely hidden), and the behaviour of traffic in the event of failure is the essence of the ecosystem's resilience. However, the underlying BGP mechanisms are limited:

    a. they make it hard to tell where traffic will actually flow;
    b. they offer very limited means to control where traffic flows once it has left an AS and, particularly, on its way to an AS;
    c. they only establish how a given destination may be reached – there is no information about the capacity or quality of routes.
    d. they propagate route changes relatively slowly and it can take a while to converge after a change (minutes and tens of minutes), which degrades service for many applications while the system settles down, and may break real-time applications.
3. the operational arrangements which keep the connections running. The automatic mechanisms at the network layer only do so much; on top of those there are operational systems for capacity management, network monitoring, repair and so on. The system depends on them to ensure sufficient capacity is available and that traffic is properly looked after. The stability of traffic flows allows for capacity management to be based on recent history and rules of thumb (based on experience) for suitable levels of spare capacity.
4. the commercial agreements which govern the connections. Every connection in the interconnection system is bilateral and so are the commercial agreements and contracts which govern them. Each AS acts on its own, in its own commercial, best interests.
5. the economic imperatives. The web of commercial arrangements between ASes is part of a larger economic system, which governs costs and the ability to set prices. A resilient Internet must emerge by the 'invisible hand' as an equilibrium from the self-interested behaviour of tens of thousands of participating ASes who act strategically.
6. regulation. The Internet is largely unregulated, and so far, so good. Regulation to improve resilience would likely be resisted by the industry unless it were well-thought-through and clearly beneficial.

4 On Resilience

In the last few years, there has been a surge of interest and research into resilience as a property of large and complex systems. This is not limited to 'computer' systems but extends across ecology and even finance, drawing ideas from fields as diverse as systems biology and thermodynamics.

As different disciplines use the words slightly differently, we will now explain what we mean by reliability, robustness and resilience. There is a huge literature on reliability where engineers study the failure rates of components, the prevalence of bugs in software, and the effects of wear, maintenance etc; this is aimed at designing machines or systems with a known low rate of failure in predictable operating conditions [**1**]. Robustness relates to designing systems to withstand overloads, environmental stresses and other insults, for example by specifying equipment to be significantly stronger than is needed for normal operations. In traditional engineering, resilience was the ability of a material to absorb energy under stress and release it later. In modern systems thinking, it also means the opposite of 'brittleness' but now refers to the ability of a system or organisation to adapt itself to recover from a serious failure, or more generally to its ability to survive in the face of threats, including the prevention or mitigation of unsafe, hazardous or detrimental conditions that threaten its existence [**130**]. In the longer term, it can also mean evolvability: the ability of a system to adapt gradually as its environment changes – an idea borrowed from systems biology [**3**] [**4**]. One aspect of the recent surge in popularity of resilience as a goal of systems engineering is the growing realisation that recovering from terrorist attacks is generally cheaper than preventing them, and that bureaucratic risk aversion carries real costs: so organisations can gain competitive advantage by preparing to recover from many classes of contingency rather than seeking to reduce the risk of their occurrence to zero [**131**]. In addition, the growing concern about terrorism should shift the balance between anticipatory risk management and resilience in favour of resilience, and terrorists set out to defeat anticipation [**7**].

The concepts of reliability, robustness and resilience have some overlap, and in the context of this report, an interesting overlap is how an AS adapts to a route failure with new route. If this is a low-level automatic process, performed by router software using the BGP protocol, it might also be considered to be robustness: after all, classical robustness also includes the use of redundant components, and a route adaptation might be thought of as the Internet equivalent of a multi-engine passenger aircraft comfortably surviving the failure of a single engine. However, we take the view that the essence of resilience is adaptability. We will therefore use 'resilience' to refer both to failure recovery at the micro level, as when ASes recover from a cable cut or the failure of a router so quickly that users perceive a connection failure of perhaps a few seconds (if they notice anything at all); through coping with a mid-size incident, as when ASes provided extra routes in the hours immediately after the 9/11 terrorist attacks by running fibres across collocation centres; to disaster recovery at the strategic level, where we might plan for the next San Francisco earthquake or for a malware compromise of thousands of routers. In each case the desired outcome is that the system should continues to provide service in the event of some part of it failing, with service degrading gracefully if the failure is large.

There are thus two edge cases of resilience:

1. the ability of the system to cope with small local events such as machine failures and reconfigure itself essentially automatically and over a time scale of seconds to minutes. This enables the Internet to cope with day-to-day events with little or no effect on service – it is reliable. This is what most network engineers think of as resilience.

2. the ability of a system to cope with and recover from a major event, such as a large natural disaster or a capable attack, on a time scale of hours to days or even longer. This type of resilience includes, first, the ability of the system to continue to offer some service in the immediate aftermath, and second, the ability to repair and rebuild thereafter. The key words here are 'adapt' and 'recover'. This 'disaster recovery' is what civil authorities tend to think of as resilience.

This study is interested in the resilience of the ecosystem in the face of events which have medium to high impact and which have a correspondingly medium to low probability. It is thus biased toward the second of these cases.

The resilience research community provides a number of clarifications and insights. For example, Hollnagel argued that high-quality resilience requires four essential system attributes [**132**]:

1. the ability to respond to various disturbances, including regular and irregular threats;
2. the ability to flexibly monitor what is going on, including the system's own performance;
3. the ability to anticipate disruptions, pressures and their consequences;
4. finally, the ability to learn from experience.

At its heart, resilience is about enhancing people's adaptive capacity so that they can counter unanticipated threats. Spare resources matter, particularly in the form of broad resource networks; so do diversity, shared understanding, deep social capital, good technical communication, deference to expertise and an ability to imagine what might go wrong [**133**]. The Internet community is relatively rich in these assets; nonetheless, resilience is still something that can be purposefully developed and managed.

Robustness also matters. Where resilience is to do with adapting to the impact of events, robustness is to do with reducing their impact in the first place. Security is to a large extent an aspect of robustness, though some aspects of security contribute to adaptability and recovery as well. From outside the system, the result of robustness and resilience are often the same, in that events have less effect on service. Inside the system the distinction may be significant, because the means to achieve the improvement are quite different.

For more on resilience see [**134**] [**1**] [**130**] [**133**]. [**135**] looks at the resilience of the submarine cable systems that underpin the 'Global Internet'.

4.1 Incidents – Resilience and Response to Events

The function of the Internet interconnection system is to support various services by transporting data across the Internet. So we judge the degree of resilience of the system by how well those services continue to run.

We may consider an 'incident' as starting with some 'event', which has some 'impact' on the system. The impact may be assessed both by how services are affected (the user view) and how the system is affected (the system view). The severity of the impact will depend on the 'strength' of the system. The system will absorb the impact and attempt to mitigate the effect on services – the 'immediate response'. Where the immediate response is not enough to mitigate the effect, there is a 'recovery' phase, in which efforts are made to recover full service. The final phase is 'repair and/or replacement', in which all is made good and the system is 'restored' – which marks the end of the incident.

So the behaviour of a resilient system during an incident may be visualised as follows:

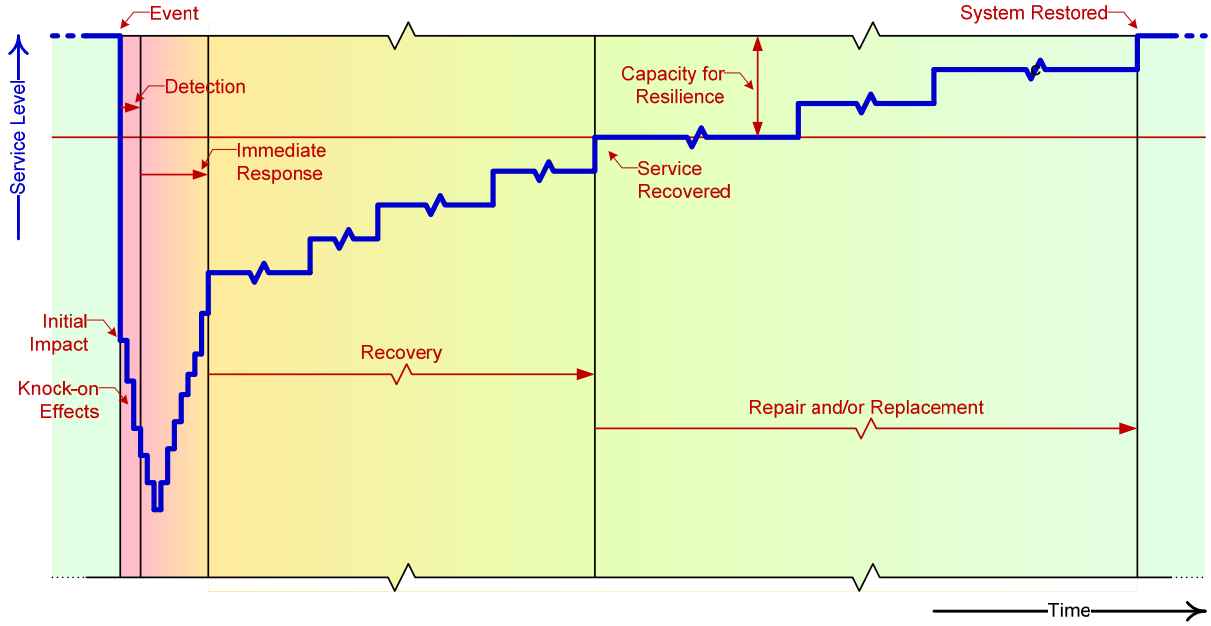


Figure 42: An Incident: Phases of Resilient Response

The general shape is more important than the exact relative scales. Indeed, the recovery and repair and/or replacement phases are likely to be much longer than the other phases – as indicated by the broken time line – and some repair and/or replacement may go on during the recovery phase.

The ideal response for a resilient system is for the impact of an event to be detected in a timely manner, and for the immediate response to be the timely switch-over to spare (redundant) parts of the system – all achieved so that users of the services provided by the system are not (unduly) inconvenienced. For routine events one would expect to see something approaching the ideal response.

To summarise, an incident consists of:

1. an event –which may be an external event such as earthquake, an equipment failure, and so on;
2. the impact or effect the event has on the system – how the event translates into system terms, the severity of which depends on the ability of the system to resist the event;
3. detection phase – until the system notices the impact of some event it cannot take steps to deal with it, and in the meantime service may be affected;
4. immediate (automatic) response – this is the designed-in response of the system;
5. secondary, possibly cascading, impact – secondary effects. Of particular concern here are problems in one part of the system which trigger problems in other parts, cascading across the entire system or problems which affect other systems on which this system depends;
6. recovery phase (longer term response) – where the emphasis is on improving levels of service by whatever means, including temporary arrangements, to recover full service but not necessarily normal levels of resilience;
7. repair and/or replacement phase – where the emphasis is on restoring the system to its normal state, with normal levels resilience.

8. restoration – the end of the process.

These elements of an incident are discussed further in the following sections.

4.1.1 Events

An incident starts with an event – something bad happens. Different sorts of events can be identified, along with examples for the interconnection system:

a. external events – such as the cutting of a fibre cable; a geomagnetic storm leading to an interruption to electricity supply; a flu pandemic causing many network staff not to work; earthquake; denial of service attack; etc.

b. failures – such as equipment or circuit failure. More serious are 'common-mode failures', in which a lot of equipment fails at the same time for the same reason, which has the potential for wide-spread impact.

c. overload – such as sudden increases in traffic from a sporting event; rapid and repeated changes of routes can overload the BGP mesh's ability to maintain a consistent set of routes across the Internet.

d. corruption –such as a misconfigured router injecting invalid routes into the BGP mesh.

e. design fault – software and hardware can suffer 'bugs' leading to common-mode failure which can extend across the entire Internet. Potentially more dangerous are design faults in the specifications themselves leading to failures or vulnerabilities across equipment from different suppliers.

f. internal events – including common mode failure, corruption and design faults, but also deliberate attempts to disrupt the system by attacking from the inside, for example by injecting invalid routes into the BGP mesh, or overloading it or otherwise.

Among the difficulties with assessing resilience is identifying a reasonable set of events to consider.

4.1.2 Impact or Effect on the System – Robustness

An event has an impact on some part or parts of the system. The severity of the impact depends on the robustness of the system – its ability to resist the event in question. The impact of an event may include various kinds of damage or failure:

a. total loss – part of the system no longer works at all;

b. partial loss – part of the system is partially working, but working more slowly or at reduced capacity, or less efficiently, etc.;

c. intermittent loss – part of the system is sometimes working, sometimes partially working, sometimes not working at all or any combination of those;

d. misbehaviour – part of the system is partially or possibly intermittently working, but some or all of what it is doing is not what is required/expected.

An event may have some or all of these kinds of effects on different parts of the system. It may also have an immediate impact on one part of the system, which causes knock-on effects in other parts. In the worst case, an initially minor event can become a major disaster if knock-on effects are not contained.

In general, however, different sorts and scales of event will have different effects and scales of effect. The number of different cases can grow quickly. It may be possible to classify events according to the impact they have. Different classes of events with different causes may have broadly the same impact – reducing the number of cases to consider.

4.1.3 Detection

Before a system can respond to an event it must detect the impact. If that takes a long time, then service will be affected for at least that long, and the risk of knock-on effects is greater. Detection may be complicated by the sorts of impact the event has had. For the Internet interconnection system we may see:

a. total loss. Where routers fail, this may be detected by other equipment connected to it as circuit failure. Some circuits may detect a break in a few tens of milliseconds, while others may take seconds, and yet others have no mechanism for detecting a break at all. BGP will itself detect the loss of a connection, or the loss of the router at the far end, within a minute or two.

b. partial loss. The partial loss of some part of the system does not lead to the loss of any routes it supports, but does reduce the effectiveness of those routes. The Internet interconnection system cannot detect partial loss.

c. intermittent loss. The intermittent loss of some part of the system is very bad news, because it leads to routes being lost, then restored, then lost again and so on. BGP routes 'flapping' in this way disrupt the system because changes propagate around the system relatively slowly, and reach different parts of the system at different times – so if routes flap quickly the system may never converge to a consistent state[36].

d. misbehaviour. Misbehaviour of some part of the system is a potential nightmare, depending on the nature of the misbehaviour, not least because it may be hard for the automatic systems to detect, either correctly or at all, that something is wrong.

Until the system detects something is wrong, it will continue as if nothing were wrong. In the case of the interconnection system that means forwarding packets towards parts of the system that can no longer (reliably) carry them – so service is lost or disrupted. Increasing the speed of detection of problems is one possible way of improving resilience, but it is counterproductive to increase the sensitivity of problem detection to the point that it falsely detects problems that do not exist. As with any other signal detection mechanism, the key is the Receiver Operating Characteristic, the trade-off between false alarms and missed alarms.

[36] *BGP has a 'Route Flap Damping' [**233**] mechanism to try to detect and absorb this effect. If it does detect flapping routes it effectively forces an intermittent loss to look like a total loss, which is a less bad effect ! The downside is that Route Flap Damping can slow down some day-to-day route changes [**243**], because they can appear to be a route flap (a false positive in the route flap detection logic). Current best practice [**235**] [**234**] is to not use Route Flap Damping.*

4.1.4 Immediate (Automatic) Response

Once the system has detected a problem, its immediate response is to:

a. adjust. The system will attempt to absorb the impact and adjust to offer as high a level of service as possible. Where resources are lost, for whatever reason, the system must be able to switch services to use remaining, previously spare resources – assuming there are any.

b. prioritise. If it can, the system may give priority to important services, so that they are less affected by any loss of resources.

For the Internet interconnection system the key resources are routes and capacity. Many subsystems in the network and physical layers will adjust to absorb the impact. At the network layer BGP will distribute new routes to replace any ones which have been lost. BGP will not find new capacity to replace any that has been lost – that is beyond it. All traffic on the 'open Internet' is lowest common denominator, 'best-efforts' traffic – the interconnection system cannot distinguish one service from another, so cannot prioritise.

4.1.5 Secondary and Possibly Cascading Impact

Cascade failures may be caused by common-mode failures of parts of the system. An event which causes some part of the system to misbehave may trigger further failures. Given a hidden flaw in BGP, a novel event, however trivial in itself, could knock over the entire Internet. Cascade failures may also be triggered by the immediate response amplifying the problem or creating another. An event which creates overload in one part of the system may create overload in another part, and so on. Route flaps are an example of this.

Cascade failures are not unknown to the Internet interconnection system; network engineers have inadvertently discovered a number of ways to misconfigure routers and disrupt large parts of the Internet.

A problem with the interconnection system might trigger a problem with electricity supply, where that system depends on the Internet, and the failure of electrical power might make it harder to recover the Internet.

4.1.6 Recovery (Longer Term Response)

Once the impact of an event has been absorbed by the system, and such immediate adjustments as are possible have been made, there are two, possibly overlapping, longer term responses: 'recovery', which is covered here, and 'repair and/or replacement', which is covered next.

Recovery is the process of improving degraded services and recovering interrupted services, to the point where, from the system users' perspective, full service is restored. If the system automatically recovers to offer full service, then the recovery phase is over before it has begun. If not, then some services will be degraded and some may have stopped altogether, so the priority is to recover from this.

For large scale events, where, more or less by definition, the immediate response of the system will not recover full service, the effectiveness of the recovery phase is key to the system's resilience. After the automatic recovery phase, it will be necessary to establish how well the system is working, and hence what further recovery is required. In a major event this will have its own challenges. To recover services the system may make further adjustments to how its remaining resources are used,

or add new (possibly temporary) resources, or temporarily repair (possibly partially) damaged or failed resources. During recovery, it may be possible to prioritise some services, temporarily, in order to recover them more quickly.

For the interconnection system, the immediate response is a matter for the network and physical layers, while the recovery phase is a matter for the operational layer – all the ASes' separate NOCs working to recover service. An important part of that will be dealing with any congestion.

4.1.7 Repair and/or Replacement

Once full service has been recovered, repair and/or replacement is the final phase – making good. What distinguishes 'recovery' from 'repair and/or replacement' is that recovery is to do with the service the system provides, while repair and/or replacement is to do with the system itself. Repair and/or replacement restores the system back to its normal state, replacing temporary repairs, and bringing any spare resources back to full working order and so to restore full resilience. Until all parts of the system are repaired or replaced, the system may be offering adequate service, but at reduced levels of resilience, so will be more vulnerable to any further untoward events.

4.1.8 Restoration

Restoration marks the end of the incident where full service with normal levels of resilience has been restored.

4.2 Assessing Resilience

Increasing resilience adds cost – a resilient a system must have spare resources. So in assessing resilience, the key question is: "is the system resilient enough?". In a commercial world, this depends on the customers: if the system is (or is perceived to be) too likely to fail, customers may go elsewhere or expect to pay less. If it is too resilient, it will be expensive and customers may go elsewhere to save money. There are (at least) two further problems. The first is imperfect information: customers cannot predict what failures are most likely, let alone how well the system will cope – how resilient it is. The second is externalities: the provider may not face the social cost of the failure of its own systems. An electricity utility, for example, may be penalised for lost customer minutes but not for the broader social costs of a supply interruption – so such a utility might not be prepared to pay for the socially optimal level of resilience in its Internet service.

In the general model for resilient behaviour, we may assess the resilience to a specific event by:

1. considering some possible event;
2. using a map of the system to estimate how the event translates into an impact on the system;
3. using a model of the system to assess how the system would respond, and in particular:
    a. how service levels are likely to be affected;
    b. how long they are likely to be affected – taking into account any staged recovery of service levels;
4. integrating all the aspects of the resilient response – the severity of the effect on services, the duration of service disruption, the effectiveness of any staged recovery, the time taken to restore the system, etc. – to provide a measure for how resilient the system is, for the given event.

This is known as 'Event Tree Analysis'. To assess the resilience of the system in general, we would have to consider all likely events, and combine the resilience for each event weighted by its probability. That is too complex to do exhaustively, apart from the problems of obtaining a good map of the system and a good model for its resilient response.

We will follow best practice elsewhere in the utility world by examining likely failure scenarios. The methodologies used for assessing resilience tend to be either top-down or bottom-up. In the first (Fault Tree Analysis) one starts off from undesired outcomes, while in the second (Failure Modes and Effects Analysis, or FMEA) the starting point is the failure or subversion of specific components. In this particular case we favour the latter approach, and consider the effects of various kinds of non-performance by various technical and other components of the Internet interconnection ecosystem. (For an example of this sort of analysis, see "Assessing Resilience in the U.S. National Energy Infrastructure [**136**].)

We define critical components to be those whose failure or loss can degrade the ecosystem or cause it to fail. Of particular interest is any 'Single Point of Failure': network engineers generally work to eliminate these, particularly at the physical level. So they will design networks so that the failure or loss of a single link, or a single router, or any other single element, does not stop the network working.

In assessing the resilience of critical components and of the system as a whole we identify the following considerations:

1. spare capacity – redundancy
2. diversity
3. independence
4. separacy – physical separation
5. vulnerabilities and single points of failure
6. best practice
7. supplier management and selection
8. preparation – disaster planning

These are discussed in more detail in the sections that follow.

4.2.1 Spare Capacity – Redundancy

Spare capacity is key to resilience. When some parts of the system are affected by some event, other parts must be able to take up the slack. In particular, a resilient network requires spare capacity across its links, so it can continue to carry the same traffic even when some links no longer work. But 'capacity' means more than bandwidth: every component must have some spare capacity to be used when others of its kind are no longer working.

There are two forms of spare capacity:

a. dedicated spare capacity. For example, we often find that:
   - a network site has redundant power and air-conditioning systems;

- within its network an AS has redundant equipment and circuits so that routine failures are barely noticeable
- a large IXP will arrange as far as possible to offer non-stop service, for example by having a backup site in case its primary site is lost; though it is up to its customers to arrange sufficient redundant connections to make most use of it.

This is often referred to as 'n+m redundancy'; for example, '1+1 redundancy', where one component is required and there is one redundant component, or '6+2', where 6 are required and there are 2 redundant. Redundancy provides high levels of resilience for a system or sub-system. However, a system or sub-system which is 1+1 redundant, has twice as much capacity as is required most of the time.

b. general spare capacity. This is where system resources have some capacity, but in normal running only part of that capacity is used. The spare capacity may be used to make up for lost capacity – assuming the load can be transferred effectively. Following a failure, BGP distributes alternative routes to maintain reachability, and traffic is redistributed to use the remaining capacity on those routes.

   This is generally more efficient than, say 1+1 redundancy. For network links, for example, this is more efficient because in normal running all the capacity of the links is available for use. Further, if one link in a network fails, the traffic it carries may spread across several other links, so it is not necessary to have as much spare capacity as 1+1 redundancy requires. So, 1+1 redundancy is generally reserved for critical links. However, the greater efficiency of reliance on general spare capacity comes at a cost: in the event of a failure, the resilience of the network depends on how quickly traffic can be redistributed and whether enough spare capacity exists in the links to which traffic is redistributed.

Redundancy tends to overlap with separacy and diversity (see below): redundant circuits are most effective if they are physically separate from primary circuits, while using diverse suppliers' equipment helps protect against design faults or attacks affecting everything at once.

4.2.2 Diversity

An AS that buys transit from two or more transit suppliers is guarding against the possibility of a failure of a single supplier, and the same holds if it connects to more than one IXP. An AS that peers with many other ASes is spreading its traffic across more connections, so that the failure of any one has less impact.

An AS can also have separate, redundant connections to diverse transit suppliers, IXPs and peers, which further improves resilience. In particular, where an AS has two (or more) connections to a transit supplier, for example, if one connection fails, the routes available to the AS and to the outside world do not change, reducing the impact of the failure both on the AS and the rest of the system.

Equipment from different suppliers is less likely to fail at the same time or in the same way due to design faults or software bugs (unlike [**137**]). Failure is not entirely excluded, though: different vendors typically offer different implementations of the same Internet protocols, and experiments with multiple teams developing software to the same specification reveal that while quality improvements are achieved, they are not always as great as might be expected, because of common mode errors – different teams can make the same mistakes when implementing a system to the same specification [**138**]. In addition, there can be errors in the specification itself, which in this context translates to the vulnerabilities of BGP discussed in this report (and any further vulnerabilities that

come to light over time). Diversity of equipment also reduces the risk of cascade failure because it reduces the risk of common-mode failures.

For diversity to be effective there must also be spare capacity. There is no point having two transit providers unless the connections, the transit customer's network, and the transit providers' networks, all have sufficient capacity to cope if something fails.

4.2.3 Independence

A loosely coupled system of independent components is more resilient than a tightly coupled system of components that depend critically on each other – though the later is likely to be more efficient. For example, one of the advantages of peering connections is that they carry an independent, self-contained amount of traffic. Unlike a transit connection, an AS's peering connections are not affected by the failure of other connections. Furthermore, a peering connection will carry just a fraction of an AS's total traffic, so the more peering connections an AS has, the more it is dividing down its total traffic. When a single peering connection fails, its relatively modest amount of traffic will spill over to the AS's transit connections.

However, a lot of peering connections happen at IXPs, which can undermine the peering connections' independence, unless there is redundancy at the IXP and in all (or most) ASes connections to it.

The interconnection system consists of many independent networks, coupled by their connections to each other and the BGP mesh. The problems with BGP are perhaps where those networks are too tightly coupled.

4.2.4 Separacy – Physical Separation

The clusters of sites which are home to the Internet interconnect system tend to concentrate infrastructure in relatively small areas. Similarly, the networks of fibre that connect sites within a cluster and between clusters, tend to concentrate a lot of traffic into relatively small numbers of common runs of fibre.

A number of incidents have demonstrated that different providers run circuits through the same trench or rent bandwidth on the same fibre: so a single cable cut destroys supposedly redundant or diverse circuits. It can actually be quite difficult to ensure that two circuits have no shared point of failure, given that service provision can involve multiple levels of subcontracting, outsourcing and virtualisation.

The term 'separacy' describes the property that two circuits have no common point of failure: namely that they are carried on separate fibres that are physically separate: so they have no common or neighbouring runs, and do not share any critical supporting infrastructure. The term is used generally to describe the physical separation of pieces of infrastructure – undersea cable systems, collocation sites, electricity supply cables, etc. – and hence the reduced risk that an event will damage all the pieces at once.

Redundancy and diversity are good attributes, but they can be negated by a lack of separacy, and that is a key issue in the resilience of the interconnection system.

4.2.5 Vulnerabilities and Single Points of Failure

A system may have components which are particularly fragile, so more likely to be affected or more likely to be severely affected by many sorts of events. It may have components which if affected have

a particularly severe impact on service, such as single points of failure. These are the system's vulnerabilities.

The first approach to a vulnerability may be to protect that part of the system from likely events – to strengthen the system or make it more secure or less likely to fail. Adding redundant or other spare capacity eliminates single points of failure in equipment and circuits. Diversity eliminates single points of failure in equipment software; increased testing may reduce the risk of such failures. Separacy eliminates single points of failure in cable runs and other physical infrastructure, while other measures may reduce the risk of damage in the first place.

There can be other less obvious vulnerabilities. An unexpected and perhaps instructive example comes from the blockading of UK petrol distribution centres by protesting lorry drivers and farmers in September 2000. The government had made contingency plans to allocate fuel rations to doctors and nurses, so as to keep hospitals open; however teachers did not get fuel rations, so schools closed, so nurses had to stay at home to look after their children. When studying the resilience of a critical component (such as a large transit provider, or an incumbent phone company), or when looking for common vulnerabilities shared by large numbers of small ASes, it is a mistake to draw the scope of the exercise too narrowly.

4.2.6 Best Practice

Best practice often encapsulates collective wisdom on vulnerabilities and how to deal with them. Best practice also encompasses operational procedures and techniques that are a vital part of keeping the interconnection system running, and responding to events when they occur. It is important to the resilience of the Internet interconnect that best practice be able to evolve so as to incorporate lessons that are learned from new types of failure worldwide, rather than becoming an exercise in box-ticking compliance. At present much of the 'wisdom' of the Internet interconnect is informal knowledge passed on between technical experts as war stories over coffee or beer; hopefully this report will help formalise and disseminate the wisdom a little better. But more is needed.

4.2.7 Supplier Management and Selection

An AS may improve the reliability of its network by careful choice of equipment, equipment suppliers, transit provider(s), IXP(s) and other critical services. This has systemic as well as local effects: the more that transit customers press their providers for resilient service, the more effort the providers will make and the more resilient the system as a whole will be.

4.2.8 Preparation – Disaster Planning

Major events often have unprecedented or at least unusual effects, so an AS may be faced with novel problems during the recovery phase. This means having operational procedures to deal with major events, where those procedures must cover how to deal with the unexpected – which means having communication and decision making systems ready to deal with incomplete knowledge and the ability to improvise and adapt to changing and possibly chaotic conditions.

For each network it is particularly important that key people in the organisation and its critical suppliers know how to contact each other. Simple things like communicating with field repair teams can suddenly become a problem if the mobile phone system has failed. So contact must be possible not only by email and mobile phone, but also by fixed line phone, and it must even be possible to

make contact at their domestic addresses - and all of these numbers need to be available on paper or on local machines. Storage 'in the cloud' or on remote systems is of little use if the Internet is ‘down’.

5 Resilience and the Interconnect Ecosystem

The preceding sections have covered resilience and the Interconnect Ecosystem separately and in general terms. Now we pull the two topics together. This section proceeds as follows:

- In Section 5.1 we describe the probable course of the immediate response and recovery of the Interconnect Ecosystem from a major event.
- The types of events and the scale of these events for which resilience may be required are considered in Section 5.2.
- Section 5.3 considers the impact of those types of event on the Interconnection Ecosystem.
- Seven key potential vulnerabilities of the Internet Interconnection Ecosystem are listed in Section 5.4.
- The issues to consider when undertaking disaster planning are considered in Section 5.5.
- In Section 5.6 we consider a number of well known past incidents and the lessons that may be learned from them.
- Section 5.7 considers resilience issues in general and as they affect the main parts of the system: Client ASes; IXPs; Large Transit Providers, and Content Delivery Networks.
- Section 5.8 addresses the problems of managing BGP and the BGP mesh and the effect that those have on resilience.
- Possible systemic failures are considered next in Section 5.9.
- And lastly, in Section 5.10, we touch on the fact that parts of the system which are more remote may be less resilient, as a natural consequence of the market.

5.1 Interconnection Ecosystem Response to a Major Event

The two main phases of the response to an event to consider are the immediate response and the recovery phase that follows it.

The outcome of the immediate response depends first on the resilience of each of the affected ASes and IXPs. In a major event we assume that there will be some loss of routes, so, second, it depends on BGP distributing alternative routes across the interconnection system [**139**]. Once the BGP mesh has converged, traffic will again find its way across the Internet, following the new routes. It might take the BGP mesh tens of minutes to converge, assuming that the event does not exceed its capacity to do so. While the BGP mesh is converging some traffic will be lost as it tries to follow routes that used to work, but which no longer do. Time-sensitive services may be badly affected – VoIP calls, for example, are likely to drop.

Once the BGP mesh has converged, the issue will be congestion. It could be that there is sufficient spare capacity across the interconnection system, that the alternative routes that BGP provides have sufficient capacity. Let us assume that, by definition, in a major incident that is not the case. Most Internet applications are not time or bandwidth critical. TCP (the protocol used by most Internet applications) contains a mechanism to throttle back the application, so that it demands less network capacity when congestion is detected. Since all applications that detect congestion reduce the rate at which they send data, the level of congestion falls and the available capacity is shared.

Where BGP looks after restoring routes, TCP looks after adapting to the capacity on the replacement routes. This is an important property of the interconnection system, and how it is designed to work. The 'best efforts' Internet does not guarantee performance, and if it did it would be at a substantially higher cost. ISPs may offer service levels which indicate that most of the time customers can expect reasonably reliable delivery of packets to and from anywhere on the Internet. But any SLA given will reflect the fact that the actual position is vague. Further, the ISP may offer little in compensation if the SLA is breached, and as quality of service is hard to measure, claims could be hard to prove.

The assumption that a major event will create congestion is supported by some of the incidents discussed in 5.6 below. However, it is believed that in Europe the major transit providers connect to each other in many places, so the loss of all connections following a failure of electrical power in, say Frankfurt, might have only a small effect on the interconnection system in Europe. Whether this is indeed true, there is no way of knowing beforehand.

Following the immediate response is the recovery phase: now the Network Operations Centres across all the affected ASes and IXPs will set about trying to recover services first to acceptable and then to full service. One of the strengths of the Internet is that each AS and IXP operates independently and each one seeks to recover as much as possible, as quickly as possible, including:

a. reorganising traffic internally, as far as possible, to relieve the strain on congested parts of their own network – internal traffic-engineering;

b. attempting to reorganise traffic entering or leaving its network, either to relieve internal congestion or to avoid congestion in others' networks – interdomain traffic-engineering;

c. shedding some traffic to reduce congestion, by disconnecting or reducing capacity to some customers and/or limiting certain types of traffic;

d. making temporary repairs to as much of its disabled or damaged equipment and circuits as possible –jury rigging and patching things up as quickly as possible;

e. adding extra circuits either internally or to other ASes to create temporary extra capacity.

Depending on the scale of the disaster, this activity could take days or even weeks. Acting independently, all the affected ASes will naturally do some things that are in the interests of the system as a whole (for example maximising the use of their resources and finding ways to introduce temporary capacity). Other activity might be selfish (for example chasing each other's tails trying to find uncongested routes or limiting traffic which happens to not be as important to the AS in question). To add to the problem, one would expect a sharp increase in demand for up to date news, to communicate with friends and family, to upload and download video footage of the disaster, etc.

Assuming the recovery phase is a long one, so that congestion will persist for a long time, the ability to prioritise some traffic, or the ability to curtail traffic that is deemed unimportant, could protect key services. The difficulties here may simply be an argument for not putting key services on the open Internet.

There are two issues with an AS trying to avoid congestion by attempting some interdomain traffic engineering. First, it is hampered by the limited facilities in BGP for traffic engineering. Second, it is hampered because the AS cannot tell where (or if) there is spare capacity that it can use. Any progress would be by trial and error, by diverting some traffic and seeing if the result was better or worse. Depending on how wide-spread the loss of capacity is, if many ASes hunt around for spare capacity in this way there could be a storm of BGP route changes across the BGP mesh, and the result

could be that congestion is chased around the system, making things worse rather than better. Effective means to engineer traffic in a crisis may be desirable, but are not available.

For an AS, the practical approach to congestion beyond its borders (assuming it can establish that its customers are being affected) is either to keep ringing its transit provider(s) until they recover service in their networks, or seek alternative, temporary transit arrangements. This may take time, but the AS will not be breaching its SLAs, though it may nevertheless have a lot of angry customers.

While congestion persists, applications such as VoIP and streaming video which are time and bandwidth critical, will suffer badly. It may thus be of concern that such applications, and particularly video, make up an ever larger proportion of Internet traffic. Policy issues follow: would it be reasonable for ISPs, or regulators, facing severe congestion following a regional failure, to seek to turn off YouTube and BBC iPlayer? Or would they just hope that most people would give up trying to watch these services as they became very slow and intermittent?

Beyond the recovery phase is the repair and/or replacement phase. Following a major event, completely restoring the system could take weeks or months. Providers who performed well may gain more customers; providers who did not may lose them. We assume, however, that the shape of the restored system would be much what it had been before the event – though a major failure, lasting many weeks in, say, Amsterdam, where there is a large cluster of Internet infrastructure, could reduce confidence and cause operators to move elsewhere.

5.2 Scale and Types of Event

In this study we are interested in medium to high impact events whose probability of occurrence is medium to low, with the potential to affect the Internet interconnection ecosystem. This ecosystem has shown itself to be robust in the face of major local disasters such as the 9/11 terrorist attacks, Hurricane Katrina, the Indian Ocean and Japanese tsunamis, the Haiti earthquake and the Buncefield fire. Based on this experience, we are optimistic about other possible disasters whose prospect may arouse considerable apprehension locally, such as the flooding of London Docklands or the next San Francisco earthquake. The ASes, IXPs and other infrastructure providers in areas that are known to be at risk already take extensive precautions, providing capacity, diversity and disaster planning to cope. Furthermore, even if a disaster should overwhelm them for a period, the likely effect on the rest of the Internet is limited.

In this report, we are interested in events on a scale or of a type that the system is perhaps not prepared for, and that will have an effect on the Internet globally or at least across a significant part of Europe. As a precedent for the horizon scanning we present here, the reader may consult the NERC report on the high-impact low-frequency (HILF) risks to the North American bulk power distribution system [**140**]. In fact, the three main risks considered there apply also to the Internet, and are the first three scenarios we discuss now together with a fourth Internet specific risk.

We consider the following four classes of potential adverse event:

1. Regional failure of other critical infrastructure on which the Internet depends.

   An example would be the failure of the bulk electricity power supply across part of Europe, and NREC discussed how such a failure could occur as the effect of a geomagnetic disturbance (GMD): intense solar activity, particularly large solar flares and associated coronal mass ejections, could do widespread damage to high-voltage transmission networks by damaging transformers and switchgear that will take time to replace, particularly if too many units are

damaged. Similar effects might follow the high-altitude burst of a nuclear weapon, leading to an electromagnetic pulse (EMP) which can cause similar damage.

A failure of the bulk power supply system over a region of Europe or North America for a period of weeks to months would disable enough ASes to have a significant effect on the Internet. Of course, such an event would also have exceptionally severe consequences for other sectors of the economy including food distribution, transport and other utilities (water, sewage, gas, fuel pipelines and telephony).

Experience of regional outages following natural disasters suggests that while an economy can weather an electricity outage of two to three days, after two weeks almost all economic activity comes to a standstill. Indeed, as the Internet comes to pervade every aspect of human life, its interconnection ecosystem will become as critical to economic activity and human wellbeing as the bulk power distribution system.

2. Disruption of the human infrastructure on which the Internet depends.

   The world has just experienced a mild pandemic of the 2009 A/H1N1 or 'Swine Flu' virus. Much more severe events have occurred in the past (such as the 1918 pandemic) and are likely to occur again in the future. Here the operational risk is not to the hardware or software of the Internet, but to the people who run it; ASes might suffer severe workforce reduction due to illness, fear of contracting illness, family issues (such as school closings) or travel restrictions imposed by governments in a bid to contain a pandemic. AS staff may be more able than typical employees to work from home, assuming they can still get online; but there are further complexities in that pandemic disease might interrupt the electricity supply or other critical services on which as AS depends. Although the sickness period for individuals might be only a week, a pandemic could come in waves lasting six to eight weeks; and staff who lose family or friends could be affected for a significant time period.

3. Coordinated attack.

   Perhaps the most significant novel threat facing the Internet is a well-planned cyber attack, or perhaps an attack involving both cyber and physical components, aimed at disrupting the interconnection ecosystem. This might be an overt act of war, or an attack by a rogue state (sometimes referred to by network engineers as 'Elbonia'), or an attack by a sub-state group who might be conventional terrorists, environmental activists or even an individual malware writer.

   Such an attack might involve propagating corrupt routing information, perhaps following a software compromise of a common make of router. A variant would be to rapidly announce and then withdraw large numbers of invalid routes, from multiple locations for maximum impact – so as to disrupt, and if possible to crash, routers in the core of the interconnection system.

   An attack that is more frequently talked about, but seems less likely to succeed, is a 'Distributed Denial of Service' (DDoS) attack. These have been used against online casinos to extort money. However, components of the core Internet infrastructure, such as DNS Root Name Servers, are designed to withstand DDoS attack and recent attacks have failed to stop them working. (Good data on these attacks is hard to come by.) It is conceivable that a very large-scale DDoS attack (for example, following a compromise of tens of millions of PCs) might have an effect.

4. Design faults.

   Routers consist of hardware and software, both of which may have bugs, and they are built according to device and protocol specifications that may also contain errors and oversights. Experience shows that software faults in BGP implementations can have a very widespread effect, but it is short lived. For example, the Internet has experienced software failures where it turns out that some routers will not tolerate unusual forms of route announcement, while others will. The less tolerant routers either fail completely, or drop the BGP session, while the more tolerant ones happily propagate the troublesome announcements across the BGP mesh[37]. The design of BGP itself has been tested in the live network for years and has evolved and we hope that the impact of newly discovered flaws will continue to diminish over time. However there is no guarantee of this, and in particular the ability of the BGP mesh to cope with an unprecedented pattern or volume of route changes is unknown.

5.3 Impact on Internet Interconnection

The likely impact on the Internet Interconnection system for the various types of event identified above is as follows.

1. An external event of regional scale, such as the loss of electricity supply to a country or perhaps even a major city in Europe, would disable part of the system.

   The impact would include loss of capacity and loss of routes. Where connections between ASes are damaged or disabled, the routes which use those connections will be lost; if an IXP is damaged or disabled, a lot of connections between ASes will be affected at the same time. Loss of routes will propagate across the BGP mesh and previously unused routes will be activated. The traffic which spills over onto the remaining routes may exceed their capacity. If enough routes are lost it may no longer be possible to reach all parts of the Internet from all other parts.

   From a customer perspective, the impacts might be congestion (leading to failure of time or bandwidth critical applications) and lost traffic due to route instability: the BGP mesh will propagate route changes, and then the changes caused by the initial changes, and then any further changes those cause and so on until things settle down and the system has converged. During this period of instability some traffic will be misrouted and lost.

   From the perspective of the Internet interconnection system as a whole, the overall impact depends on how the large transit providers and other large ASes respond – and in particular whether their networks suffer congestion. The impact of the redistribution of traffic is key. While ASes can be expected to plan for some degree of resilience for their own network and their own customers' traffic, it is not clear what, if any, provision is made for extra demand caused by problems in other ASes or at IXPs. It is common for an AS to manage capacity on the basis of history plus some safety margin, so the typical transit provider is not managing capacity to cope with a large shift in routes. The system cannot have infinite capacity, but that begs the question of how much spare capacity it should have, and who should decide, and how.

[37] *In the light of this there are proposals to standardise the behaviour of routers to be more forgiving of certain types of invalid announcements – the latest draft* ***[155]*** *is dated 28-Sep-2010.*

2. An external event such as a flu pandemic that causes widespread shortages of skilled labour is likely to have a more diffuse impact.

   It might be hoped that if the effect were to cause (say) 80% of all staff to stay at home then although small ASes with only a few staff might be unable to continue operations, larger ASes with dozens or even hundred of staff would have at least some engineers available and so would be able to continue operations. The organisation of the Internet, with a power-law distribution of AS size, makes it very resistant to random failure of ASes although it is vulnerable to failure of a small number of large ASes. So there are grounds for cautious optimism that the impact of a pandemic on the Internet might be less than on the bulk power distribution system. However, this then reverts to case 1 above: even if the large ASes can still staff their control rooms, if the electricity supply fails because linemen stay at home for fear of contracting a deadly strain of flu, the Internet would rapidly follow.

3. The effects of a coordinated attack are hard to predict because of the many options available to an attacker.

   Disruption of the BGP fabric appears to be a likely target, although it may be combined with overload (whether as a knock-on effect or a separate component of an external attack). The overload might be general – the result of normal traffic trying to fit in restricted capacity, or of abnormal volumes of traffic generated by public anxiety about the attack; or it might be targeted, as for example if many millions of infected PCs simultaneously tried to attack some part of the infrastructure.

   It is not impossible that a cyber attack might be blended with a physical attack, such as cable cuts or (in the event of war) kinetic attacks on IXP hosting centres or on supporting infrastructure such as electricity substations.

4. In the case of a design fault, the impact of the event triggered by a design fault is hard to gauge, but it is of interest here only if it affects a significant part of the Internet.

   Software faults in BGP can effectively disable large numbers of routers across the entire Internet, the impact of which is similar to a large number of simultaneous external events hitting individual routers at the same time. The impact of internal attempts to disrupt the system is potentially severe and widespread, because it uses the power of the system against itself; a combination of the two might arise if an attacker discovered a vulnerability and exploited it to take over many routers (though such an attack would properly fall under 3 above).

5.4 Vulnerabilities

The four scenarios discussed above could exploit a number of shortcomings in the interconnection system including:

a. the concentration of equipment and connections in clusters of sites, often at or near IXPs and other important facilities. This leads to dependence of those clusters on stable electricity supply, shared physical infrastructure which undermines resilience measures without anyone realising it, particularly with respect to communications;

b. the variability of resilience across ASes and of the connections between them – in particular the major transit providers;

c. the possible spillover of unknowable volumes of traffic onto routes with unknown capacity, and the difficulty of managing traffic generally – as BGP is ignorant of the available capacity of the routes it selects and using it to direct traffic away from congested routes is problematic;

d. the inability to prioritise some traffic, for example VoIP;

e. the openness of BGP to corruption;

f. the possibility of common software faults in BGP; and

g. the possibility of the BGP mesh failing to converge.

Having identified these vulnerabilities, the question is what to do about them. In October 2006 the Internet Architecture Board (IAB) held a "Workshop on Routing and Addressing", whose primary goal was to look at the problems that the large backbone operators have with the scalability of the Internet routing system. The report from that was published as RFC4984 [**141**]in September 2007. Tackling these issues remains a work-in-progress.

5.5 Disaster Planning

Let us assume that occasional disasters will damage or disable equipment and circuits causing a major upheaval of routes and a large-scale reduction in capacity. The system's response to such an event is described in 5.1 above. If such scenarios are rare – say, a geomagnetic storm once every fifty years, which knocks out the electricity supply to a quarter of the EU – then perhaps the emphasis should be on how rapidly the system can recover. It may be prohibitively expensive to harden the Internet against such events (especially if the bulk power distribution system is not simultaneously hardened), but some general preparedness can speed recovery.

In planning to improve the ability to recover quickly after a disaster, the following should be considered:

- what are acceptable levels of service in the event of a disaster of this magnitude?
- would coordination between ASes improve or accelerate some parts of the recovery process, and if so how to achieve it?
- how do ASes communicate with each other, their suppliers, their customers, their field staff, and so on during an incident where ordinary means of communication may be lost?
- would pre-prepared schemes for setting up temporary connections speed things up?
- would mechanisms to share resources between ASes help (if say some networks were hit harder than others)?
- could service for important traffic could be recovered more quickly (or unimportant traffic throttled back)?
- would the civil authorities set priorities?

This is the province of those responsible for local critical infrastructure, but this sort of preparation would also be of benefit for incidents short of a disaster.

5.6 Well Known Incidents

In this section we describe a few well-known incidents, partly because they illustrate many of the issues that have been covered in this review, but also to illustrate the quality (or lack of it) of the information about important events and some of the misconceptions that engenders.

5.6.1 A Compendium of Route Leaks and Hijacks

On a fairly regular basis some AS in the Internet manages either to leak routes or to hijack some. A route leak is generally the result of the misconfiguration of a router. A route hijack can be the result of misconfiguration or of some other fault, it may also be deliberate. The following sections describe the mechanics of route leaks and route hijacks, and their effects on the interconnection system.

5.6.1.1 Simple Route Leak

Consider some fraction of the Internet as shown:

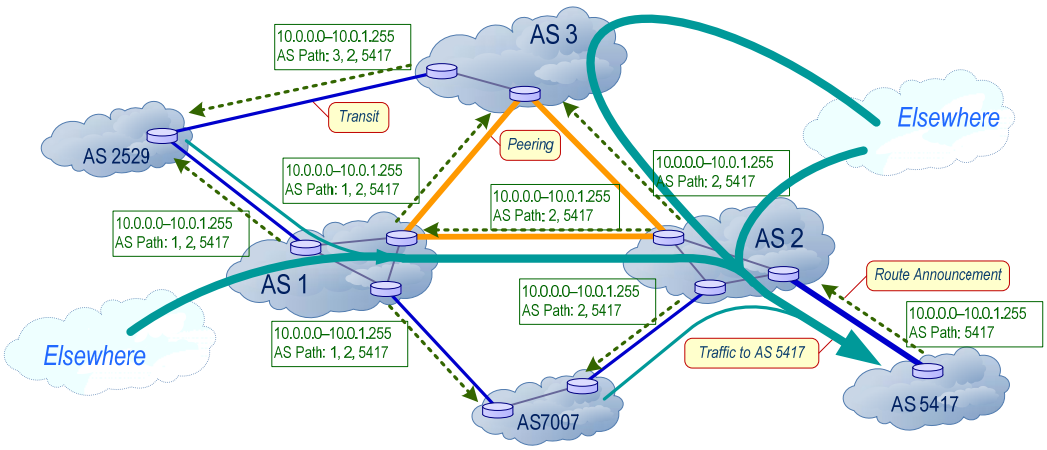


Figure 43: Before a Route Leak

in which AS5417, bottom right, as the origin of the address block 10.0.0.0–10.0.1.255. The other ASes and routers choose and announce routes for this address block, and traffic to AS5417 flows as shown. Note that: AS2529 has two routes with equal length AS Paths, and chooses to use the AS1 path; AS3 chooses the route with the shorter AS Path, as does AS7007. The *Elsewhere* clouds represent other parts of the Internet connected to AS1, AS2 and AS3, who are large transit providers.

AS7007 is a transit customer of both AS1 and AS2, and will normally only announce its own and its customers' routes to them, so it would not normally announce a route to any of AS5417's address blocks. *Router-Q* (in AS7007) has a full global routing table, so for the BGP connection between AS7007 and AS1, *Router-Q* is configured not to announce most of what it knows. If when making a configuration change to *Router-Q* a mistake is made, and it announces everything it knows to AS1, then that will include the route to 10.0.0.0–10.0.1.255, via AS2 – since that is AS7007's preferred route. The effect would be as shown:

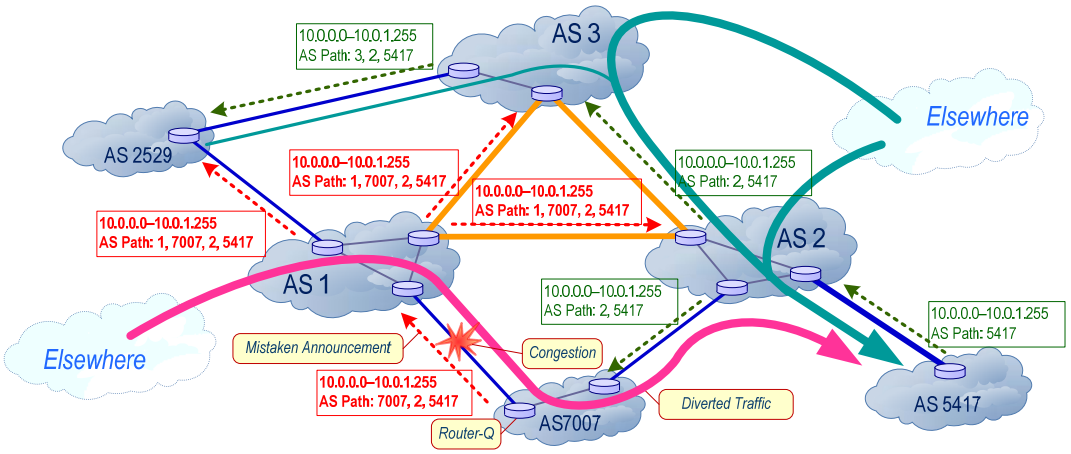


Figure 44: After a Route Leak

Here we see the mistaken announcement from AS7007 to AS1 in which it offers to carry traffic to 10.0.0.0-10.0.1.255. Now, AS1 will prefer to use routes that come from a customer – the more traffic sent to a customer, the more money the transit provider makes. So, AS1 will announce the new, but mistaken route to its customers and peers. We see that AS2529, which was using the route via AS1, now uses the AS3 routes, because the AS Path is shorter – it could have stayed with the AS1 route if, say, AS1 were less expensive. AS3 ignores the mistaken route, because it is choosing between peers on the basis of the AS Path length. AS2 ignores the route because it has a customer route to AS5417, so prefers that. The upshot is that some traffic going to AS5417 is now going via AS7007. Since AS7007 is announcing every route it knows to AS1, and AS1 will prefer them all, a great deal of traffic may be attracted to the connection between AS1 and AS7007, and the link will be completely overloaded, so traffic is effectively being sucked into a black hole. A simple mis-configuration of a router can have a wide-spread effect.

Note that there is nothing invalid about the announcements that AS7007 is making, and that AS1 is accepting. The routes that AS7007 is announcing are valid ways of reaching the addresses they are for. The only problem is that AS7007 doesn't have the capacity to carry the traffic! The RPKI system, discussed in Section 3.1.12, would not help in this case.

Note also that the effect of a route leak like this depends on the policies applied by the routers that hear the announcements, and on whether the routes are selected or not. It is hard to predict what the effect would be if one set out deliberately to collect traffic in this way to inspect or interfere with some traffic. This mechanism is also only collecting packets to the affected addresses, so may or may not collect both halves of any conversation.

When the mistake is detected, AS7007 can correct its router configuration, or AS1 can turn off the connection temporarily, or other ASes can implement filters to ignore what appear to be mistaken routes with AS7007 in the AS Path.

As discussed in Section 5.8 below, there are ways in which this kind of mistake can be mitigated:

a. 'maximum prefix' feature: in a route leak the AS usually announces a very large number of routes compared to the number they usually announce. In this case, if AS1 sets a reasonable 'maximum prefix' limit, then the mistake at AS7007's end would almost immediately lead to the router at the AS1 end closing the connection (dropping the BGP session) and the mistake would be contained. One would expect a well run transit provider to apply a 'maximum prefix' limit as a matter of course.

b. egress filtering in AS7007. The idea here is that AS7007 configures a permanent set of filters on its connections only allow the address blocks AS7007 should announce to be announced. This is a belt and braces approach, to avoid embarrassment if a mistake of the sort discussed here from having any effect.

c. ingress filtering in AS1. The idea here is that AS1 should configure its end to only accept routes for address blocks it knows that AS7007 should announce.

The practicalities of egress and ingress filtering are discussed in Section 5.8.3.

5.6.1.2 Route Hijack

Where a route leak lets out valid routes by mistake, a route hijack is the announcement of invalid routes by mistake or otherwise. A route hijack could be the announcement of valid routes deliberately to collect the traffic, but that is not necessarily very effective, and we have covered it already. A route hijack may also be the use of some unallocated address space, but since that does not disrupt the interconnection system, we will not discuss it here.

If AS7007 wanted to attract traffic to AS5417, it could announce routes as if AS5417's addresses belonged to it, as shown:

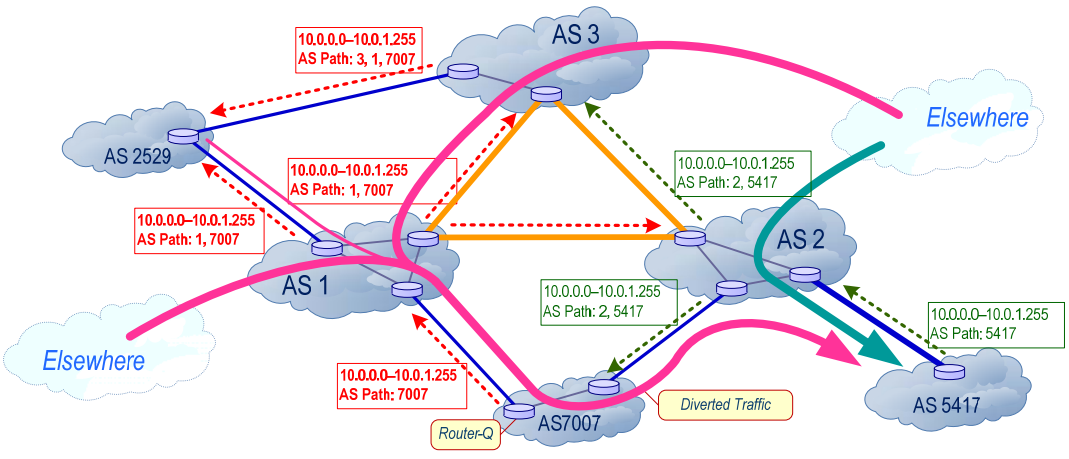


Figure 45: Route Hijack – Partial

The main difference from the route leak is that the AS Path in the announced routes is shorter, which may persuade more ASes to use it. In this case we assume that given two routes with the same AS Path length, that AS3 will prefer a path via AS1 – it has to choose one or the other.

In the scenario above, packets that diverted to AS7007 can still make their way to AS5417, because AS7007 is not announcing the hijack route to AS2. This is a rather special case, which here depends on the hijacker and the victim sharing a transit provider and the hijacker arranging the announcements to make use of that fact. If the hijacker wishes packets to still reach the true destination, then they

must take care to not completely hijack the addresses, though that will mean that not all packets will be diverted to AS7007.

If AS7007 announces the hijacked addresses to every AS, then we see:

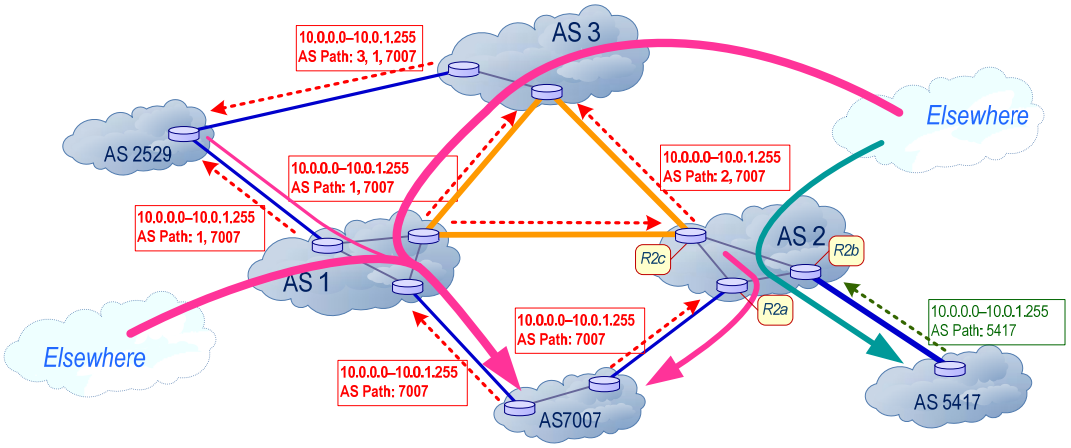


Figure 46: Route Hijack – Complete

in which more packets are diverted to AS7007, but no packets can make their way out again, unless AS7007 makes some very special arrangements indeed, for which it would probably need the assistance of another AS. What is going on here depends on how AS2 has configured its routers. In this scenario it is assumed that router *R2a* will prefer the route from AS7007, because that is the directly connected route. Similarly, *R2b* prefers the route from AS5417. *R2c* will hear one route from *R2a* and another from *R2b* and chooses the route via *R2a*, because in this instance that is the shorter route.

If AS7007 and AS5417 do not share a transit provider, then the situation is a little different, because they would then be greater separation and more opportunity for more ASes to make different routing decisions.

The form of hijack shown in Figure 46 will disrupt routing, to an extent which depends on the routing decisions made by all the ASes that receive the invalid announcements. This form of hijacking can be used to divert some proportion of the traffic to a given destination. It is more difficult to also arrange for packets to eventually reach their final destination. It has been noted that BGP offers limited means to affect the routing of traffic beyond an AS – this is an example of that!

The routes being announced by AS7007 are invalid; AS7007 is not the true origin of the address block 10.0.0.0-10.0.1.255. This sort of hijacking can be mitigated by ingress filtering and would be picked up by RPKI. However, AS7007 could lie as follows:

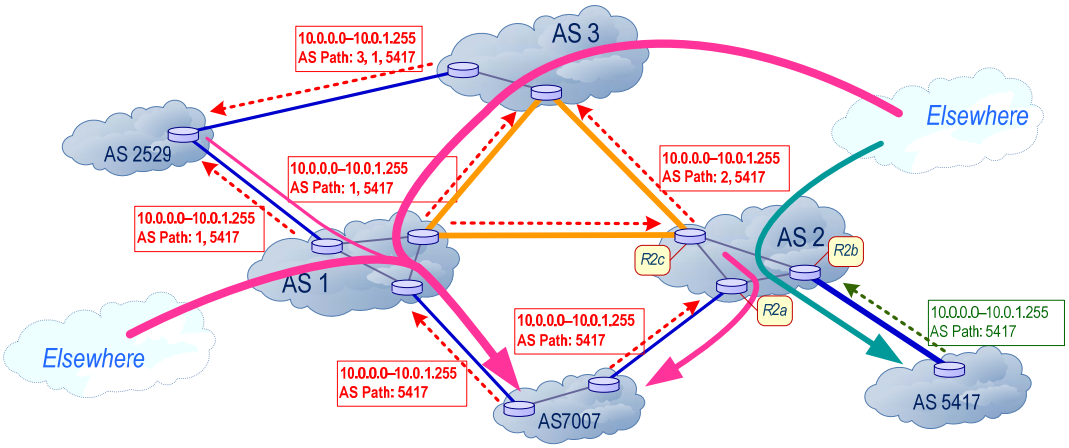


Figure 47: Route Hijack – Lie Direct

Which would not be picked up by RPKI, but really ought not to get past AS1 or AS2, because they can see that the AS Path is clearly invalid. A more plausible lie would be:

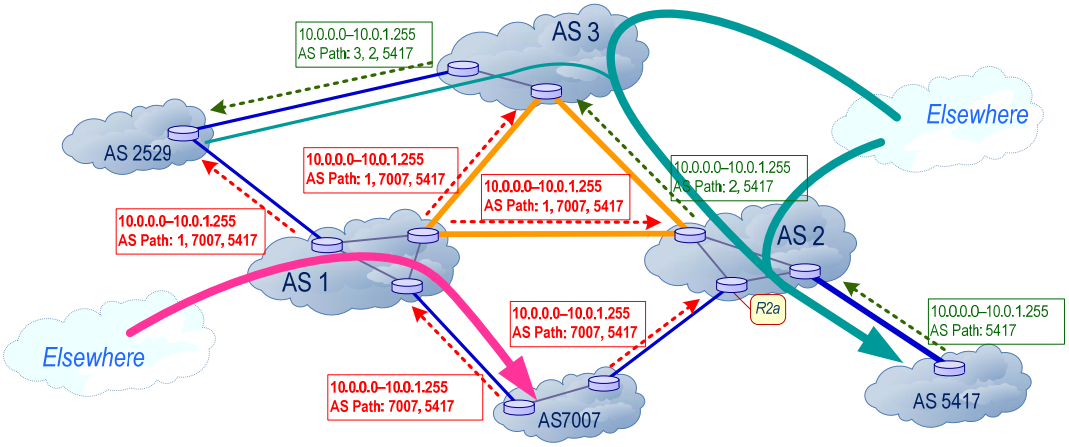


Figure 48: Route Hijack – Lie Indirect

Now Router *R2a* has a route with a shorter AS Path than the route from *AS7007*, so AS2 ignores the hijack attempt, as does *AS3*. In this case AS2529 prefers the route via *AS3*, so it too ignores the attempted route hijack. This would get past RPKI, because the invalid routes appear to originate in the right place. However, the effect of the hijack is reduced because the AS Path has been lengthened to get past RPKI – so RPKI is having some beneficial effect.

5.6.1.3 Route Hijack with 'More-Specific' Routes

A 'more-specific' route is universally preferred to a 'less-specific' route. This is an unconditional requirement of BGP. So, the most effective way to hijack routes is to announce 'more-specific' routes for the ones to be hijacked.

If AS7007 is determined to hijack 10.0.0.0–10.0.1.255, then it can announce routes for 10.0.0.0–10.0.0.255 and 10.0.1.0–10.0.1.255 , which are the two halves of the address block to be hijacked, and that is shown:

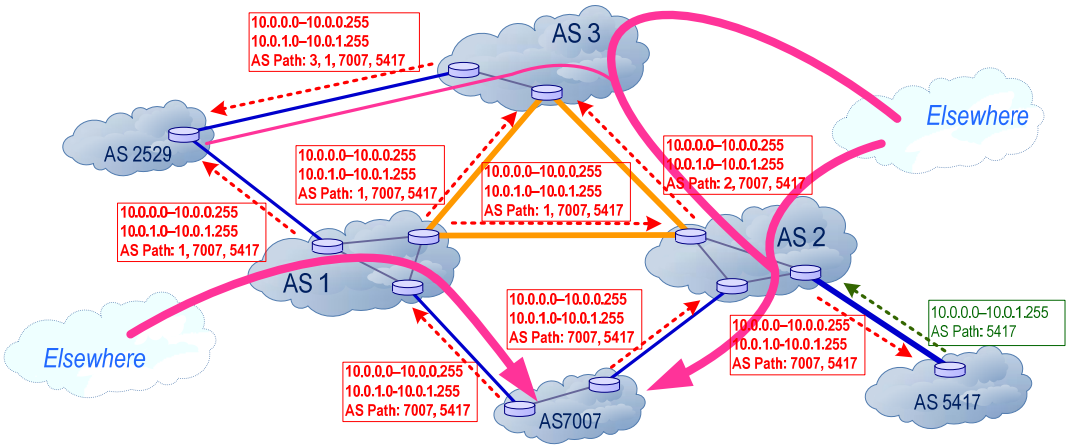


Figure 49: Route Hijack – with 'More-Specifics'

which shows the two more-specific routes announced by AS7007, and every other AS choosing and using those routes in preference to the real, less-specific, route announced by AS5417. To add insult to injury these more-specific routes will be announced to AS5417 itself.

Using more-specifics is the most effective way of hijacking routes. However, because the entire interconnection system will prefer the more-specific routes, it is hard to get packets to AS5417, so this hijack mechanism is not well suited to eavesdropping.

This form of hijacking is detected by RPKI, unless AS5417 happens to have registered the two more-specifics as well as the whole address block.

5.6.2 AS9121 Route Leak — December 2004

This is an example of a route leak. On the 24th December 2004 AS9121 announced over 100,000 invalid routes – almost a complete Global Routing Table, at the time. In [**142**] there is an analysis of this incident. It appears that one of AS9121's transit providers did not have a maximum-prefix limit configured for AS9121, so the contagion spread further than it should have done. The transit AS announced all of the invalid routes to its peers, who did have maximum-prefix limits configured, so many sessions between the transit AS and its peers were terminated, which, as a side effect, affected perfectly valid routes. The incident had a serious impact on the Internet, disrupting routing and traffic, for about an hour and a half, though there were two secondary incidents of 10 and 35 minutes [**143**]. See also [**144**] for an analysis a year on.

5.6.3 AS7007 Route Hijack — April 1997

This is perhaps the most famous route hijack. On 25th April 1997 AS7007 announced a very large number of invalid routes to their transit provider, attracting traffic for most of the Internet into a

'black hole'. Most of the invalid routes were 'more specific' than the standard ones[38]. In addition, the AS Path had been stripped back to just one AS number, 7007. The unusual features of this incident made the routes being announced particularly attractive, which accounts for the severity of the impact on the Internet. It is not clear how these routes were generated.

AS7007 quickly detected its mistake, and stopped announcing the invalid routes within an hour, and generally the incident was over in about two hours.

But the invalid routes persisted elsewhere, and some effect was felt for up to 7 or 8 hours in total. It appears that the sheer number of routes that AS7007 had announced caused many routers to crash because they were simply overwhelmed. When the crashed routers restarted they were sent the same routes all over again, causing them to crash again. This flapping may have exacerbated the problem.

Interestingly, although this is a famous incident, little seems to have been understood at the time about why invalid routes persisted for so long after the root cause was dealt with – or if it was understood, it does not appear to have been documented. If this was a real example of BGP failing to converge, interesting data appear to have been lost. AS7007's Director Network Services wrote this explanation [**145**], see also [**146**], and a later analysis [**147**].

The result of this incident was an increased interest in route filtering in general, particularly filtering by transit providers of the routes announced to them by their customers. Work on S-BGP started in 1997.

5.6.4 YouTube Hijack — Feb 2008

This is an example of a route hijack, using more-specific routes. On Sunday, 24th February 2008, Pakistan Telecom (AS17557) started an unauthorised announcement of the prefix 208.65.153.0/24. One of Pakistan Telecom's upstream providers, PCCW Global (AS3491) forwarded this announcement to the rest of the Internet, which resulted in the hijacking of YouTube traffic on a global scale. RIPE reported on this event [**148**], see also [**149**], [**150**], and [**151**].

Apart from being a good example of how effective a more-specific route hijack can be, this incident also had a political cause. The Pakistan Government instructed ISPs [**152**] to block a particular 'offensive' YouTube URL, and gave three IP addresses. It is not known why AS17557 announced the more specific to their transit provider.

The incident lasted about 2 hours and 15 minutes. For all the attention this incident received, its effect was limited in scope and time. This illustrates the capability of the operational layer of the ecosystem to respond to events.

All the indications are that Pakistan Government intended local ISPs to block particular content locally. The fact that this censorship 'leaked' illustrates the difficulties with BGP when attempting to influence routing and traffic. Nevertheless, this raises the very large topic of how local requirements interact with each other, and how that interacts with the limitations of the infrastructure.

[38] *For completeness: it appears that they announced the first 'classful' address block for every 'classless' one – for example for 172.16.0.0/14, 172.16.0.0/16 would have been announced. Other accounts suggest that they deaggregated every route into its component /24's, creating a very large number of more specifics.*

5.6.5 RIPE Unexpected Attribute – August 2010

This is an example of bugs in BGP implementations causing problems across the BGP mesh. On 27th August 2010, the RIPE NCC's Routing Information Service (RIS) announced some BGP routes which included an unusual, but entirely legal 'attribute'. Unfortunately, there was a bug in some Cisco routers which caused them to mangle the unusual attribute, in such a way that when they passed on the route, the router it was passed to would see an invalid attribute, and drop the BGP session. The effect was to disrupt the BGP mesh for perhaps 40 minutes. See [**153**] and [**154**].

This clearly illustrates the potential for latent bugs in BGP implementations to affect the system, though in this case a limited number of routers were affected. It also illustrates how a problem can propagate across the mesh and how BGP's handling of invalid attributes[39] can amplify the effect of a failure. However, this also illustrates the capacity of the operational layer to detect and deal with problems.

BGP includes a mechanism for the creation of new attributes without requiring ever router in the world to be immediately upgraded to understand those attributes – when a BGP router sees an 'optional, transitive attribute' which it does not understand, it is required to pass it on unchanged. When one BGP router talks to another, the routes it sends out are some mixture of the routes it has learned from all the other BGP routers it is connected to. Now, if a router does not understand new attribute 'X', it will pass it on even if the contents of that attribute are, in fact, invalid. So, amongst perfectly good routes there may be some bad ones, and parts of the system will innocently propagate them. Further, because the response to one bad route is to discard all routes learned in a BGP conversation, the effect is greater than it should be. There are recent proposals to handle errors in 'optional transitive attributes' to mitigate the effect, see [**155**].

5.6.6 Alexandria Cable Cuts – January and December 2008

This is an example of the effect that cutting some cable systems has, where the number of cable systems is limited. On the 30th January 2008 two major cables were damaged somewhere off Alexandria in Egypt: Sea-Me-We 4, FLAG FEA. Repairs to these systems took some 10 days, during which service in various parts of the Middle East, India and beyond were severely affected. Later that year, on 19th December there were repeat cuts to Sea-Me-We 4, FLAG FEA and a partial cut of Sea-Me-We 3. See [**156**] [**157**] [**158**] [**159**] [**160**] [**161**].

Compared to the BGP related incidents above, these simple failures of basic infrastructure had long lasting effects. The fact that two or three apparently separate cable systems were affected at the same time clearly illustrates the importance of redundancy, diversity and separacy!

The available analysis illustrates the difficulty of assessing the impact of such events. There is some information about loss of routes, but no good information on how traffic was affected (other than that large numbers of people were clearly inconvenienced). Further, the reasons for the cable cuts are not generally known – reports that the cables were damaged by ships' anchors are contradicted

[39] *When BGP detects something invalid in the information it receives, the standard response is to drop the connection with the router sending that invalid information. The effect of this is that all the routes the two routers have learned from each other are lost in the process. So, a small fault can be amplified. One of the difficulties is that, having received something which is invalid, BGP has no reliable way of determining what parts of the rest of the information are valid.*

by other reports. All the usual suspects in that part of the world are thought by some to have set out to disrupt communications. For all that these incidents could tell us about how the system performs when something very unusual happens, there seems to be a shortage of good information.

5.6.7 Taiwan Earthquake

On 27th December 2006 an earthquake cut seven out of nine cable systems serving Taiwan. Disruption continued for two to three weeks, see [**162**] [**163**].

The cause of this incident is clear: an earthquake of magnitude ~7 with its epicentre in the Luzon strait, some 20km south of Taiwan. The geography of the area means that this is a popular path for undersea cables, and seven out of nine cables in the area were damaged, affecting connections to and from Taiwan, Japan, Hong Kong, China, Korea, and Singapore. The strait is 250km wide, so these cables need not be that close, however as the map in [**164**] shows, several start and/or end in the same places, and the longer a cable is the more it costs and the greater the latency through it.

In [**165**].the authors describe the experience of research and education networks in the region. In the immediate aftermath of the earthquake BGP managed to maintain some reachability, but with the loss of so much capacity, there was severe congestion. Over the following days, operators worked to establish what capacity remained where, and adjusted routing to move traffic around. Interestingly, the while the operators of the research and education networks did what they could with their own routing, they found that the cable operators were moving traffic around under their feet.

5.6.8 Brazil Power Cuts – November 2009

On 11th November 2009 southern Brazil suffered an electrical power cut which affected some 60m people from Rio de Janeiro, through Sao Paulo to Paraguay [**166**]. The cut lasted about 6 hours. Apart from the obvious effects [**167**], of interest is the rumour that this was caused by 'hackers'.

On 26th/27th September 2007 there was a power cut in the state of Espirito Santo (north of Rio de Janeiro) affecting about 3m people. On 6th November 2009 the US CBS network broadcast an edition of their "60 Minutes" news programme "Cyber War: Sabotaging the System" [**168**], in which it was claimed – amongst other things – that "prominent intelligence sources confirmed that there were a series of cyber attacks in Brazil", causing black-outs, including the one in Espirito Santo. For some people, the much larger event a few days after the broadcast was all the confirmation they needed.

The Brasilian federal agency responsible for electrical power generation and distribution (Agência Nacional de Energia Elétrica – ANEEL) found that the Espirito Santo incident of 2007 was caused by a build-up of soot on insulators on pylons, caused by fires in the region [**169**], which is what Furnas Centrais Elétricas, the company responsible, said on the 29th September 2007 [**170**]. On the 27th January 2009, ANEEL fined Furnas R$5.54m (€1.86m at the time) for failing to maintain the transmission systems properly.

The November 2009 blackout was caused by a series of short circuits in and near a substation, which caused various protection systems to trip. These short circuits are attributed to extreme conditions – gales and heavy rain – see [**171**] and [**172**]. There was some early speculation that the cause was lightning strike, which became controversial a weather expert at Brazil's National Institute for Space Research said that satellite imagery proved that the nearest lightning strikes were at least 10 kilometres away. In the days that followed the blackout a hacker managed to gain (unauthorised) access to the National Electricity System Operator's (ONS) corporate network, but ONS were quick to

point out that their operational systems were not connected to that network, or indeed any other part of the open Internet.

So, discounting the possibility of an elaborate cover-up, it appears that 'hackers' were not to blame [**173**]. However, the 'hackers' story is sadly compelling, and in the absence of good information, such stories can persist.

5.6.9 China Telecom – April 2010

On 8th April 2010 China Telecom (AS4134) leaked routes amounting to about 15% of total address space[40], apparently for 18 minutes. What is interesting about this is not just the leak itself, but the hysteria and paranoia it seems to have created [**174**] [**175**] [**176**] [**177**] [**178**], and [**179**].

The source of the route leak was IDC Beijing China Telecom (AS23724) who appear to be customers of China Telecom (AS4134) and CNIX-AP China Networks Inter-Exchange (AS4847). AS4134 is well connected around the world, and it was AS4134 who passed on the invalid routes to its peers and transit providers[41], which is why the incident had the impact it did. This suggests that AS4134 did not filter the announcements from its customer, which is not unusual. Given that about 37,000 invalid routes were announced, it also suggests that they did not have (an effective) maximum prefix limit set, which is sloppy. In April 2010 AS23724 appears to have been announcing some 30 routes, though that jumped to about 55 in the latter half of that month, and has grown to about 125 since – so a maximum prefix limit of, say, 1,000 would be reasonable. However, if AS23724 did the usual thing, which is to announce all routes – about 330,000 at the time – then something in AS4134 managed to avoid announcing the majority of them.

Reports of the incident talk of traffic being "redirected through" China Telecom [**178**]. Most route leaks of this sort result in traffic being sucked into a black hole, never to be seen again. If traffic was diverted to AS4134 and came back out again to reach its destination, that would be remarkable – apart from anything else, once an AS has accepted the invalid routes from its customer, one would expect it to send any traffic it has for those destinations to that customer; getting the traffic to remerge from the AS is tricky, as is ensuring that the traffic does not get sucked back. It may well be that some reports are less well informed than others – but it would be nice to know what really happened.

The timescale is of some interest. That the incident lasted just 18 minutes suggests that network monitoring in AS4134 detected the problem – perhaps when their connection(s) to AS23724 filled up with all the traffic diverted to it. The importance of the operational layer cannot be underestimated.

[40] *which some reports give as 37,000 routes, or 11% of all routes – the amount of address space covered by a route varies.*

[41] *which currently includes: Global Crossing (AS3549), TINET (AS3257), Verizon (AS701 née UUNet), Sprint (AS1239), Cogent (AS174), Telia (AS1299), NTT (2914), Telecom Italia(AS6762).*

5.6.10 World Trade Centre – 9/11 – September 2001

This is one of the best documented incidents – see "The Internet Under Crisis Conditions" [**5**], in which the authors noted:

> ""The decentralised architecture of the Internet – although widely characterized as one of the Internet's strengths – further confounds the difficulty of collecting comprehensive data about how the Internet is performing."
>
> "It is therefore unsurprising that no definitive analyses exist on the impact of September 11 on the Internet, though a few conflicting anecdotal reports about its performance that day – such as several presentations at NANOG indicating relatively little effect and press accounts suggesting that the impact was severe – have appeared."

This report will not attempt to summarise [**5**], but leave it as recommended reading.

5.6.11 Cogent 'De-Peering' – October 2005, March 2008 and October 2008

Cogent was 'de-peered' by Level(3) in October 2005 [**180**], by Telia in March 2008 [**181**] and by Sprint in October 2008 [**182**]. These partitionings of the Internet lasted two or three days each.

In a 'de-peering' incident, one of the two parties to a peering relationship turns off all peering connections with the other. This is usually because one party no longer feels that the arrangement is (sufficiently) equitable, and the other party is not prepared to change the arrangement (see 3.6.1 above).

These disputes between Tier 1 providers affect the providers' customers who have no other connection to the Internet – often referred to as 'single-homed' customers. So, if 'n' customers who only reach the Internet via Level(3), and 'm' only reach the Internet via Cogent, those 'n' customers would be cut off from the 'm' customers, and vice versa.

The scale of these incidents can be measured in terms of the number of routes or addresses affected, but neither 'n' nor 'm' are going to be much more than 1% of all routes, and rather less than that in terms of addresses. Further, what really matters is how much traffic usually flows between those groups of customers – about which even less is known.

Compared to BGP related incidents, two or three days is a long time. Compared to an undersea cable cut, it is not a long time.

Apart from illustrating the difficulty of assessing the impact of events on the system, these occasional incidents show that the system can fail at the commercial level as well as other levels.

5.7 Resilience Issues

In our simplest possible view of the interconnection system, shown opposite, we have the 'core' of the system, the major transit providers, the IXPs and the CDNs, and a sea of 'client' ASes around them. The client ASes buy transit from the transit providers, connect to each other at IXPs and connect to CDNs either at IXPs, or directly, or via transit providers. With the exception of any direct peering connections between themselves, all the client ASes' traffic goes to and from the core[42]. The proportion of total traffic delivered by each of the three parts of the core is not known, but we speculate that it may be very approximately a third each, though some of the CDN traffic will be via IXPs, particularly, and via the major transit providers.

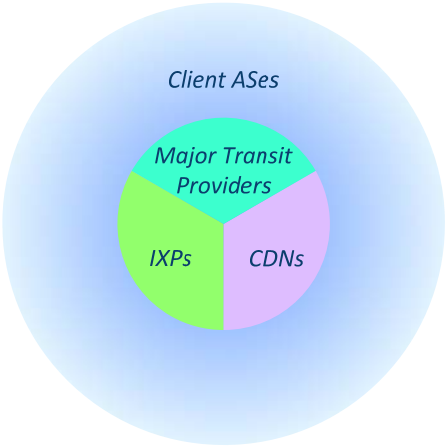


Figure 50: Simplified View of the Interconnection System

In this section we consider some general resilience issues and some issues specific to each of these parts of the interconnection system.

5.7.1 General Issues

Diversity and separacy are standard measures to improve resilience. Diversity without separacy is, obviously, less effective. The difficulty when provisioning and managing circuits is how to ensure that separacy is maintained. The business is multi-layered: a circuit may be provided by one operator, who buys a wavelength from another, who in turn buys a fibre pair from another, who may actually own the cable. Then there may be several cables owned by different people in close proximity to each other. Further, once a circuit has been provisioned, it may later be rerouted. Operators do not like to give guarantees to customers that then constrain their own operations.

There are a limited number of long distance cables. In some parts of the world there are very few cables, while in others multiple cables converge at key points.

Resilience costs. No utility likes to maintain extra capacity to cover for the possible failure of its competitors, and in more established markets the solution to this problem is regulation. For example, in electricity markets, it is common for regulators to add a tax to market prices so as to provide for enough 'spinning reserve' and grid resilience to maintain something close to five-nines electricity supply. The resilience of the interconnection system as a whole is no different, in principle, but as the business is global and there is no global regulator, there is no mechanism to share the cost of resilience across the ASes, and no reason to suppose that each AS will spontaneously pay extra costs for the benefit of the whole.

The system expects the self-interest of ASes to drive resilience, and for the market to find a suitable resilience level. But the market basis is very weak. First, capacity planning is done on a historical

[42] *Among the client ASes there are some who provide transit to other client ASes. Those smaller transit providers will connect to the major transit providers. They may also directly peer with fellow smaller transit providers, and any traffic on those peering connections does not go via the core either.*

basis, with little consideration of major incidents for which there is no precedent. Second, it is hard to measure quality of service at the best of times, and even harder to determine what it is likely to be when things go wrong. Third, there is no support on the open Internet for differential quality of service, so no mechanism to protect important services in the event of serious congestion. Thus the incentive to take on extra cost against an unquantified, hypothetical need is weak.

5.7.2 Resilience Issues for Client ASes

By Client ASes we mean the great majority of Internet networks – the small and medium size ISPs, the medium size and large enterprises – whose connection to the Internet is provided largely by transit providers, augmented, perhaps, by local peering. Client ASes have little direct impact on the resilience of the interconnection system.

The resilience of a client AS's connections to and from the rest of the Internet is ultimately in the hands of their transit providers. So a client AS managers, the key tasks in obtaining resilient service are:

a. selection and management of their transit suppliers. As noted elsewhere, at least two transit providers would be recommended. And some transit providers are better than others: the client should choose carefully and monitor its providers' performance.

b. management of the transit connections and their capacity. These are essential parts of the client AS's network, so will be managed as such. Managers must expect one transit connection to fail from time to time – so be ready for traffic to spill over to the other(s).

c. management of any peering connections and their capacity. The client AS must also make provision for peering connections to fail from time to time – and arrange for adequate spare transit capacity to cope.

When things go wrong in the interconnection system the client ASes can really only wait for things to be put right, and monitor the performance of their transit providers throughout. If a transit provider underperforms, then the client can always change. This gives a useful market signal in normal times. However, in the event of a regional or global incident, it can be expected that many clients of many transit providers will suffer dislocation and congestion. This prospect undermines any incentive that transit providers might have to differentiate themselves as more resilient by buying extra capacity.

5.7.3 Resilience Issues for IXPs

An IXP is, by its very nature, a potential single point of failure. How far the IXP should go to mitigate this depends on the IXP's users who pay for it. The IXP is, in fact, an extension of their networks.

The users of a small IXP may treat any traffic exchanged there as a bonus (saving money on transit) rather than a necessity. Those users may decide that it is not worth spending a lot of money to make the IXP more resilient, because even if it does fail occasionally their transit providers will cope. Provided any failure of the IXP does not exceed 36 hours in a month, the extra transit traffic will not cost extra, under the usual 95th percentile charging scheme.

The users of a very large IXP may well feel differently. They may exchange a significant proportion of their traffic at the IXP and it would be a serious matter if that traffic had to spill over to transit providers (or, possibly, another IXP). The large IXPs, therefore, arrange for the IXP to be fully 1+1 redundant, to minimise the probability of total failure.

An IXP hosted in a single site is vulnerable to the site failing (loss of power or cooling) or being damaged or destroyed. So the very largest IXPs are spread across more than one site, connected by a redundant network. In fact, a well run IXP will be as resilient as it is possible to be, and may have two internal networks using different suppliers' equipment as well as two sites and two fibres connecting them.

Yet, from the point of view of the interconnection system, the overall effectiveness of an IXP, even a highly resilient one, still depends on how each AS organises itself:

a. each AS must ensure it has sufficient capacity to and from the IXP to handle the traffic (some IXPs even have rules about this, to maintain their reputation);

b. a prudent AS will have redundant connections to the IXP, and some alternative, just in case it does fail;

c. where it is connected at multiple IXPs, it must buy sufficient capacity should one fail and traffic spill over to others.

Also, from the point of view of the resilience of the interconnection system, the significance of a given IXP depends on how much traffic it carries, relative to the alternatives available to the IXP's users – which is a measure of the impact on the system if the IXP stops working. And locally an IXP may have greater significance than that. It may carry a large proportion of a country's internal traffic, and for this reason be regarded as a key resource by the local Critical National Infrastructure organisation.

5.7.4 Resilience Issues for Large Transit Providers

The large transit providers and the connections between them are an essential part of the Internet interconnection system, and have a direct impact on the resilience of the whole system. The level of resilience within each large provider's network depends on the cost of that resilience and the need to gain and retain customers at a price they are willing to pay.

From the point of view of the interconnection system, it is possible for a large-scale event to cause traffic to spill over from one transit provider on to others. A key issue here is that the big players have neither the incentive, nor the information, to prepare properly for a spill over of traffic from their competitors' networks.

5.7.5 Resilience Issues for Content Delivery Networks

These networks make money by keeping copies of their content in sites from which they can connect as cheaply and effectively as possible to as many 'eyeball' networks as possible. Connecting to IXPs is a perfect way to do this, but for large eyeball networks a direct peering connection may be justified. The CDNs do this in large part to maintain the quality of their offering. Being close to the eyeball networks improves the responsiveness of the service, and allows for high capacity connections at minimum cost. The peering connection also obviates the need for either party to pay for transit, so there is a cost saving, too. CDNs deliver a significant proportion of total Internet traffic. One third party CDN claims to deliver 20% of all Internet traffic. So they are a significant part of the interconnection system.

The CDNs contribute to overall resilience by all this distribution and peering activity. They in turn must look after their connections to IXPs and direct peers. While they largely bypass the transit providers, they still use transit providers to reach ASes they cannot peer with, and as a backup for their other connections. The hard question is whether there will be enough capacity available from

the transit providers when it is needed. For example, the simultaneous failure of a number of European IXPs, whether as a result of a utility failure or an attack, might lead CDNs to dump large quantities of traffic into the transit providers' networks.

5.8 Managing BGP and Interconnections

The workings of BGP and the BGP mesh are at the heart of the interconnection system. We have seen that there are serious problems with BGP itself [**183**] [**184**], and occasionally with how it is used. Here we look at what may be done in the management of individual interconnections in order to mitigate these issues. This would come under the heading of Best Practice.

We will cover four examples

1. route filtering –accepting only known valid routes, or rejecting obvious rubbish.
2. route monitoring – watching out for invalid routes
3. maximum prefix filtering.
4. source address filtering – discarding packets with 'spoofed' source addresses.
5. deaggregation – rejecting deaggregated routes.

They all illustrate a general point. Although individual ASes can take operational steps to improve the system as a whole, the public spirited AS incurs a direct cost, but only enjoys a small and indirect benefit. It is the classic public-goods problem, and real benefit may only be felt if large numbers of ASes make the effort; although perhaps if appropriate procedures were adopted by the major transit providers, this would measurably improve the system as a whole.

At a lower level, if two ASes have a single connection between them, the failure of that connection will change the routes available to each AS, which may in turn affect the routes those ASes announce to other ASes, and so on. So the failure of a single connection can have a wide-spread effect, which can be exacerbated by delays in BGP [**185**]. Where ASes maintain two, separate connections, the effect of a single failure is mitigated (provided, of course, the remaining connection has sufficient capacity). An approach to reducing the effect of the failure of BGP connections is for each AS to establish 'Failover Matrices' for its BGP connections, see [**186**].

5.8.1 BGP Route Filtering

BGP allows invalid routing information to be injected into the BGP mesh, and will automatically propagate that across the entire Internet: BGP routers simply accept whatever they are told. So if one BGP router makes a mistake (or lies), all the BGP routers it speaks to will accept the invalid information, and promptly pass it on (in good faith).

BGP routers believe what they are told because there is no built-in mechanism to validate the routes they are given. More secure versions of BGP are covered in Section 3.1.12, above. It is very difficult to make changes to BGP, simply because of the scale of the system and the problems of making changes to what is a permanently live system. On balance, it is unlikely that a complex system of trust and verification on top of BGP is going to be adopted in the near future. Apart from the fear of breaking things, there is concern that the overhead of performing all the trust and verification operations would adversely affect the BGP mesh's ability to converge in a timely fashion.

Route filtering is an operational way to compensate for this deficiency in BGP, by either discarding apparently invalid routes or only using apparently valid ones. Route filtering and the issues with it are discussed in Section 3.1.11, above.

For a route to be valid it must:

1. be for a block of IP addresses that the origin AS is entitled to use. Unfortunately, there is no authoritative source for this basic information. The RPKI initiative (see Section 3.1.12, above) is addressing this fundamental issue.
2. have a path which is genuine in that each AS in the path will forward packets to the destination. If this is not the case, then the route is counterfeit. Unfortunately there is no practical way of knowing this, and a general solution requires a more secure form of BGP.

The Internet Routing Registries (IRR) were intended to provide information that could be used to verify routes. An AS can maintain an up to date, publicly accessible record of their routes in an IRR. Unfortunately, there is no direct link between the IRR information and an AS's actual routes, which are defined by what the AS does, not what it says it does. It makes no difference to the AS if its IRR information is out of date, incomplete or a work of fiction. Publishing information in an IRR is voluntary. The information that should be published in an IRR may be deemed to be secret, so some ASes publish nothing and some may publish incomplete information.

Nevertheless, the IRRs are better than nothing, and are used as the basis for some route filtering. Where that happens it creates an incentive on the ASes concerned to maintain accurate, complete and up to date IRR information (at least for the parts that the route filtering uses).

Various forms of route filtering which have been proposed are partial solutions, intended to be reasonably practical:

a. 'Bogon' and 'Martian' Filtering.

   A 'Bogon' is a piece of address space that has not been allocated. A 'Martian' is any address that has a known special use – the designated private address spaces, for example. Any route for such address space is clearly invalid, and whoever announces it is either up to no good or simply deranged.

   Filtering out Bogons and Martians is quite straightforward, and is of some benefit to the AS that does the filtering, because it shields their customers from potential harm. It is a small step in the right direction. However, it is essential to keep Bogon Filters up to date. Pieces of address space are regularly allocated to satisfy the demand for new addresses. An out-of-date Bogon Filter can filter out a valid address block, which will not impress customers, no matter how worthy Bogon Filters are in principle.

   As a practical matter, Bogon Filters are based on the allocations made at the top level, that is, when IANA allocates blocks to the RIRs. It takes some time for an RIR to allocate all the addresses in one of those blocks to ASes, so there is a period in which an address is not recognised as a Bogon, but is not really valid either.

b. egress filtering – where an AS checks the routes it announces to others.

   If an AS checks the routes it is announcing to its transit providers and peers, it reduces the possibility of sending out invalid routes by mistake.

If all ASes did this, it would be a good thing. If only a few ASes do it, it does not make much difference. It is not possible to tell whether an AS is doing this, or whether it is doing it effectively. Moreover, an AS must ensure that its filters are kept fully up to date and consistent across all of its routers – which is a lot more difficult than it should be.

Any slip up with egress filters would mean that some customer's routes would not be properly announced. So egress filters can be much more trouble than they are worth – which is not difficult, given that they are of no direct benefit to the AS in any case.

c. ingress filtering – where an AS checks the routes it receives from others.

   As discussed in Section 3.1.11 above, it is not currently possible to do this comprehensively. A partial approach is for every transit provider to check its customers' own routes, on the basis that it is more or less practical to establish what those routes should be. Accepting the estimate that 80% of ASes are stub ASes, this approach would check the vast majority of ASes' routes as they enter the BGP mesh. If every transit provider did this, all routes would be checked as they enter the BGP mesh. However, since each transit provider is only checking its customers' own routes, there is nothing preventing it accepting invalid routes that appear to come from its customers' customers.

d. checking the first AS number in the AS Path

   The first AS number in the AS Path should be the AS the route was learned from. If it is not, something is amiss (or something special is going on). This does not prevent invalid routes from entering the system, but it does prevent them from entering the system anonymously.

What is really needed is an effective way for the Tier 1 and Tier 2 providers to verify the routes they receive from their customers and from each other. Then the 'core' of the interconnection system would be less susceptible to disruption. Partial solutions are not effective, so the RPKI work is an essential first step. When a definitive source of what each AS is entitled to announce is available, it should be possible to detect invalid routes, at least those generated by accident. By doing most of the work outside the BGP routers, the extra overhead affects them as little as possible. However, a more complete system is required if deliberate attempts to disrupt the system are to be forestalled.

5.8.2 BGP 'Maximum Prefix' Feature

In a peering connection the ASes announce their own and their customers' routes to each other, and there are generally a limited number of these. A common mistake made when configuring a peering, or a transit, connection is for one peer, or the transit customer, to announce every route it has, i.e. a complete global routing table. The effect of this may be to attract a great deal of traffic, which neither the peering connection nor the peer is ready to handle. A well-known example of this is the AS7007 incident of 25th April 1997, described in 5.6.3 above.

BGP implementations have evolved over the years, in the light of experience. Following the AS7007 incident, most BGP implementations were enhanced to allow a limit to be placed on the number of routes a peer or customer may announce – this generally known as the 'maximum-prefix' feature[43]. The number of routes that a customer or peer will announce is small, certainly compared to the

[43] *In the jargon a block of Internet addresses is referred by its 'network prefix' or just 'prefix'.*

global routing table. So it is possible to set a limit on the number of routes that will be accepted, which that will detect this sort of mistake, and close down the peering connection before the contagion can spread. Note that this does not verify the routes themselves.

The great thing about the maximum-prefix feature is its simplicity and ease of use. There is no real need to be precise about how many routes are expected from a given peer or customer; setting a limit ten times higher than the expected number will still easily and quickly detect an attempt to announce a full routing table.

There is no good reason not to use the maximum-prefix feature, so, now when a peer or customer does inadvertently make this mistake, the effect is generally contained.

The AS9121 incident, described in 5.6.2 above, is interesting because most of AS9121's transit providers did have a 'maximum prefix' limit set, but one did not, so the incident was not contained as it should have been. However, what is more interesting is that the transit provider that did not have a maximum prefix limit, announced all of the leaked routes to its peers. That triggered their maximum prefix limits, so the transit provider lost a number of peering connections as a side effect of the route leak. Further, those peering connections would have been carrying traffic for routes which had nothing to do with AS9121, so some traffic suffered collateral damage.

5.8.3 BGP Route Monitoring

It would be preferable if what BGP distributes could be verified and invalid routes filtered out. What is clear, however, is that this will be difficult and expensive to achieve.

Invalid route announcements are relatively infrequent. When they happen NOCs leap into action and they are dealt with quickly. Perhaps the market is, in fact, making the right (or the efficient) choice in not implementing more secure BGP or BGP practices. There are systems such as Cyclops [**187**] that monitor what routes are being distributed, and signal suspicious announcements. The scheme suggested in [**62**] is designed to monitor for unusual announcements. Rather than spend a lot of time and money trying to perfect BGP, it may be more cost effective to deal with the occasional problem, and perhaps speed up detection and response rates.

5.8.4 Source Address Filtering

Source address filtering, also known as "Network Ingress Filtering', deals with invalid addresses being used in IP data packets. As discussed in Section 3.1.13 above, it is possible for the source address in an IP packet to be invalid, and such packets are almost invariably up to no good. See [**188**] for a fuller discussion of the state of IP spoofing defence, and also [**189**].

If all ASes checked packets coming from their users and direct customers, and rejected any that do not have a valid source address, then spoofed source addresses would be a thing of the past. This would not eliminate DoS or DDoS attacks but at least the source(s) of the attack packets could be traced back to their real origin. The RFC2827/BCP38 document [**6**] recommends 'Network Ingress Filtering' to prevent IP packets with invalid source addresses from entering the network, or being passed from one network to another.

Since an AS allocates the addresses that its users and direct customers use, it knows what is or is not valid. Also, it can configure the routers that its users and direct customers first connect to, and so trap invalid source addresses both as early as possible, and with a fair degree of precision. In the best of all possible worlds, an AS would not only filter out packets with invalid source addresses, it

would also investigate where they came from – it is possible that the user's or customer's machine has been compromised without them realising it.

A second type of source address filtering is 'Reverse Path Filtering'. When an AS announces routes across a BGP connection it is saying that it is happy to receive packets for the addresses in those routes. The source address of an outbound packet should be one that packets may be sent back to, so the source address should match an address in a route previously announced by the sender. Unfortunately, this is not a cast iron rule, nor do all routers support the facility, nor is it without some cost.

Much like filtering BGP announcements, it is reasonably practical for an AS to apply Network Ingress Filtering to its users and direct customers, but there is no effective way to implement source address filtering elsewhere. Also as with route filtering, source address filtering would be of general benefit to the whole system but has no immediate benefit to any AS that does such filtering, despite the cost in time and effort.

5.8.5 Rejecting Deaggregated Routes

Every route in the Internet has to be known to every BGP router in the Internet (in general terms). For every route which an AS announces, there is a small but finite load on the BGP mesh to distribute the route, and on every BGP router to process and store it. It is in everybody's interests to minimise the load on the BGP mesh – the less work that BGP has to do, the less likely it is to become overloaded.

As discussed in Section 3.1.9 above, deaggregation is used by some ASes to manage their traffic to their own advantage. Unfortunately this costs processing and memory resources in every single BGP router in the Internet. The regular 'CIDR Report' [**11**] suggest that the global routing table is about 50% bigger than the absolute minimum (though it is not entirely clear that the absolute minimum is achievable). In [**190**] the authors examine current levels and trends in deaggregation, and conclude that the problem remains, but is not becoming (proportionally) any worse.

There are efforts to "name and shame" ASes who deaggregate. This appears to have little effect on those ASes (but may have some deterrent effect on ASes who might feel tempted to). It would be possible for ASes to ignore deaggregated routes (given a definitive source of what are valid routes). If every AS did this, it would certainly be a disincentive, and the global routing table would shrink. But, an AS will worry that it might ignore a valid route which just happens to appear deaggregated. Further, individual action would probably have little effect, except perhaps to upset the AS's users and customers, who could find themselves cut off from some part of the Internet and could not care less about deaggregation and its cost implications for the system as a whole. So yet again there is no benefit from individual action and no mechanism for collective action.

5.9 Systemic Failure

Systemic failures are the most alarming threats, because they can damage or disable large parts of the system at the same time.

One form of systemic failure is 'common-mode failure', where many components of the system fail together in response to the same event. Design faults are a major cause of common-mode failure, as illustrated by events in which some BGP implementations have failed when they receive announcements of a particular form. To reduce the impact of such a failure some systems use a

range of equipment, each designed separately. But there are not many suppliers of the types of router that are used in the interconnection system.

Another cause of common-mode failure is an error in a specification – if the specification of some type of component turns out to be faulty, or incomplete, all components of that type can fail together. In this context, the universal use of BGP is a concern.

A cascade failure occurs when some other event affects part of the system in such a way that nearby parts are affected, and that affects further parts, and so on. Common mode failures may be a mechanism for the spread of a cascade failure. Cascades can also be caused by overload having a knock-on effect that creates more overload and so on.

5.10 Local vs Global

The Internet interconnection system is a global system. When looking at its resilience it is natural to concentrate on the bulk of the system – the 80% that requires 20% of the effort to run. But we have to keep reminding ourselves that 20% of the Internet is still a huge amount of network.

The Internet has no notion of Universal Service. So, unlike many communications systems, parts of the network remote from the main centres may find themselves paying to bridge the distance between themselves and the main centres. From a resilience perspective that may mean that parts of the interconnection system are poorly connected to the rest – with limited capacity and redundancy.

This has a number of interesting consequences. For example, Denial of Service attacks directed at a target in a poorly connected part of the system can have a much wider impact than a similar attack directed elsewhere, simply because of the size of the attack relative to the local capacity. For all victims of Denial of Service attacks, it would be better, and the system would be more resilient, if there were better mechanisms for stopping the attack traffic further away from the victim.

Although it is a global system, as has been noted elsewhere, a majority of traffic is local – anything from 60% to 90%. This suggests that efforts to improve the local resilience of the interconnection infrastructure will benefit the system for a large part of its use. It further suggests that from a resilience perspective it may be worth considering the global system as a number of local systems interconnected.

6 The Wider Issues

So far, we have looked at largely technical, operational issues and contractual issues. In this section we consider some of the wider issues, as follows:

- The Internet has a distinct culture which is examined in Section 6.1. The culture rejects regulation even though there are identifiable 'market failures' for which regulation might perhaps be desirable.
- All ASes have incentives to look after day-to-day traffic flows and routine events, as described in Section 6.2, but they have fewer incentives to provide resilience against unusual events, and none to consider the resilience of the system as a whole.
- Service Level Agreements exist and ISP customers have the ability to choose other ISPs if the service is poor, but there is little actual performance measurement available to buyers and in Section 6.3. we suggest that this might result in a "market for lemons".
- Transit Pricing at marginal cost is not economic as marginal cost is close to zero and there are significant fixed costs. Companies are pulling out of the transit provision market and some that remain are making significant losses. Section 6.4 points to this aspect of Transit provision as a possible vulnerability of the Internet Interconnection Ecosystem.
- Section 6.5. looks at the economics of peering, particularly at IXPs, and whether falling transit costs are a significant disincentive to peer.
- Section 6.6 suggests that with applications moving from the Desktop to the Cloud, consumers may be underestimating the risk of reliance on a best efforts system.
- Section 6.7 considers potential market failures and whether new incentives are required underpin resilience of the system.
- Section 6.8 gives some examples of past Government interventions which have failed.

6.1 Cultural Issues

The Internet has a distinct culture. The Internet succeeded where other contemporary network initiatives failed – notably the Open Systems Interconnection (OSI) initiative. Where OSI was large and bureaucratic, the Internet was small and ad hoc. Where OSI worked on meticulous specification, the Internet favoured 'rough consensus and running code'.

The Internet sees itself as a triumph of the free-market, where the invisible hand guided the explosive growth of a service that has provided a great public good, delivering ever-increasing power at ever decreasing cost. Unlike the telephone and telecommunications systems that came before, the Internet was not regulated, and so was not held back by bureaucracy. Coupled with the free-market ideals are notions of individual freedom; rather than the diplomats' view of nation speaking peace unto nation, the Internet is seen to give individuals the free and open ability to communicate with anyone at any time, transcending the nation state, and liberating them from government control.

Within the 'Internet Community' there is scepticism of Government. Government is perceived as:

- clueless – not understanding how the Internet really works, and trying to look at it in old-fashioned telecommunications or broadcast terms;

- wishing to assert some control – where the point of the Internet is that it is free and open, transcending national boundaries;
- wishing to interfere with traffic – whether to ban cryptography as a means of facilitating surveillance, or to ban peer-to-peer traffic in response to lobbying from the music industry;
- using Critical National Infrastructure and counter-terror concerns as excuses to interfere in general.

There are numerous examples of Governments who do want to restrict their people, and some who (partially) succeed. So, for the Internet Community, no regulation is good regulation; the Internet is where it is today without being regulated – indeed, it is where it is today because it was not regulated – and changing that would wreck things.

However, there are obvious cases where the market fails:

a. the absence of a mechanism to ensure that ASes, especially large ones, provide the socially optimal level of spare capacity for resilience;

b. the failure of most ASes to deploy technical measures that would increase resilience, such as route filtering, more secure forms of BGP, source address filtering, and address deaggregation;

c. the lack of a system-wide mechanism to deal with Distributed Denial of Service (DDoS) attacks and of a system-wide drive against bot-nets. An ISP can often detect when a customer's machine has been compromised, and could isolate the machine from the rest of the world; but it is quite a lot of work and mostly benefits other ASes and their customers;

   Once a DDoS attack is detected, it would be best if the attack traffic were identified as close to its sources as possible. That way there would be less impact on the 'innocent bystanders' close to the target of the attack – who are affected by congestion. And if DDoS attacks could be stopped before the traffic is concentrated onto the victim, then the Internet might be a safer place.

d. levels of preparation for IPv6 and the transition to IPv6. The last blocks of IPv4 space were allocated by IANA on 3rd February 2011, and the RIRs may run out of IPv4 addresses by June 2011. The end-game for IPv4 addresses is discussed further in [**191**] and [**192**].

   The world does not seem to be fully ready for the long heralded exhaustion of IPv4. In [**193**] it is reported that, at the end of December 2010, some 6% of all web sites in the EU had IPv6 addresses. In [**194**], RIPE reported (June 2010) that across all their members, some 60% or more have no IPv6 capability at all. In [**195**] it was estimated that in early 2010 "*some **5%** of Internet's end systems are capable of supporting end-to-end IPv6.*"

   In [**196**] Geoff Huston notes that the expectation was for IPv6 to replace IPv4, and discusses whether the complete lack of progress is a market failure.

Nevertheless, as noted in [**197**]:

> *"The Internet that we know today arose in a delicate balance with the competitive market forces that tie service providers, technology developers, and content providers together in a global, voluntary agreement to maintain these practices and standards. This agreement has been maintained out of an implicit belief that cooperation to keep the Internet functioning as an open, interconnected, and non-discriminatory platform serves the interests of the parties individually as well as collectively."*

Perhaps the answer is that the current Internet is the worst possible system, apart from all the others?

6.2 Structure of Incentives

Despite these market failures, the existing incentives in the Internet interconnection ecosystem do actually achieve a lot. End users pay their ISP to provide a connection to the Internet, and to send data to and receive data from any other part of the Internet. End users who are not satisfied with their service will generally have a number of ISPs they can change to. In turn, ISPs who are dissatisfied with their transit providers will generally have a number of others they can change to. Competition should ensure that end users expectations for price and performance are met.

It is worth noting that choice for end users has not arisen spontaneously. In some cases it has been made possible by deregulation – removing regulation that protected incumbent operators – and in other cases it has been made possible by new regulation – forcing incumbent operators to unbundle the local loop, for example.

We have seen that the mechanics of transit arranges that end users, indirectly, compensate the large transit providers for connecting to all parts of the Internet and transporting data all over the world. Again, competition amongst transit providers should ensure that their direct customers' expectations for price and performance are met. And where there is a chain of transit providers, a 'market signal' travels along the chain. Thus:

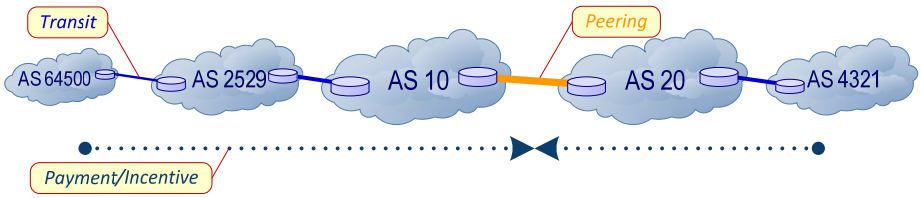


Figure 51: Payments/Incentives up to a Peering Connection

where each path between locations on the Internet has a chain of transit providers and transit arrangements from each end meeting either at a peering connection, or within a common transit provider:

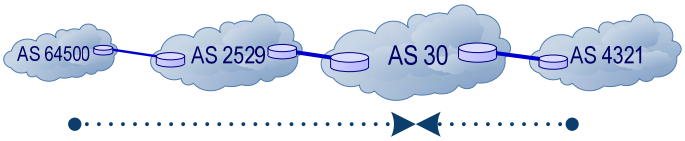


Figure 52: Payments/Incentives up to a Common Transit Provider

so that there is a trail of payment and incentives covering every possible path in the Internet.

So all ASes have an incentive to arrange for the day-to-day flows of traffic to be looked after, and that the routine events are absorbed without unacceptable degradation of service or unacceptably long periods of unacceptable service. And, of course, all ASes try to do this at minimum cost.

But the chains of incentives do have limits. From AS64500's perspective, it has a direct contract with AS2529, and knows who to call about a problem. It has to accept that its traffic may be a small part of AS2529's business (they may even be competing for the same end users). When things go wrong, who knows what AS2529's immediate priorities will be? Further, AS64500 has no direct relationship with AS30 at all, or any of AS2529's other transit providers. The chain of incentives becomes diffuse –

AS64500 is depending on AS2529's relationships with its transit providers, and how they will treat AS2529's traffic when times are hard.

There is also an unspoken assumption that the two chains of incentives that meet in the middle are roughly equal. There is no way of telling whether that is the case, but there is no particular advantage to AS30 in spending a lot of money on resilience if all its peers and transit providers have rather lower standards.

Also, as noted in Section 3.4.8 above and further in Section 6.3 below, formal SLAs stop at the edge of each provider's network. For the apparent chains of payments/incentives to be truly effective they would require back-to-back formal agreements across the entire system. That is not the case now, and is not likely to happen in the near future.

The incentive that appears missing is the incentive to provide for major but infrequent events: in particular, where a significant amount of traffic which is usually carried by some transit provider(s) spills over onto other transit provider(s). It would be good for the system as a whole if there was provision for this, but it would be an extra expense for the transit providers, for which they are not compensated. In other words, the overall resilience of the interconnection system is an externality. And the resilience to rare events is also an externality.

The discussion above assumes that a path will only ever have one peering connection in it, and that will indeed be the case. Suppose there were a set of connections such as:

Figure 53: No Transit via Peering Connections

where AS30 is peering with AS10 and AS20. In this case there would not be a path from AS2529 through AS30 to AS4321 (or vice versa). AS30 will learn a route to AS2529 via AS10 from AS10 – because AS10 provides transit to AS2529, and in a peering connection each AS announces all its customers' routes to the peer. But AS30 will not announce the route it has to AS2529 via AS10 to AS20 – because in a peering connection each AS does not announce routes learned from other peers – this is the 'no valley' rule in operation. So, there is no path from AS4321 to AS2529 via AS30, and similarly no path from AS2529 to AS4321. If there were to be paths through AS30, then it would be carrying the traffic gratis, which clearly makes no sense at all. So AS10 and AS20 must make other arrangements to complete a path between AS2529 and AS4321.

6.3 SLAs and the Market for Lemons[44]

The system of incentives assumes that end users can choose ISPs on the basis of their performance, and that ISPs can then choose transit providers similarly on the basis of performance, and if resilience matters to a user than that performance should include resilience.

[44] *"The Market for Lemons: Quality Uncertainty and the Market Mechanism" [**222**] by George Akerlof discusses what happens in a market in which the seller knows more about the goods than the buyer. The market in second hand cars is given as an example, in which there are some good second hand cars and some bad ones, the 'lemons'. If the buyer cannot distinguish a good car from a lemon, then there is no incentive on one seller to go to any effort ensuring the quality of their goods, since the buyer is likely to buy a cheaper alternative. As a result, lemons come to dominate the market.*

In fact it is hard to measure the performance of an ISP. It is even harder to measure the resilience of an ISP and its connections to the rest of the Internet. Similarly, it is hard for an ISP to measure the performance and resilience of transit providers.

The classic way of dealing with the problem of this type of failure('a market for lemons') is to introduce a warranty, or some other signal that one seller's goods are indeed of higher quality than another's. For ISPs the Service Level Agreement (SLA) is a form of warranty. A good SLA for transit may cover:

a. Availability. This measures the percentage of some period (usually a month) for which the transit customer can reach the router at the transit provider end of the connection, and whether that router is announcing routes. (This is equivalent to measuring whether the BGP session is active.)

b. Latency. This measures the round trip time between designated points in the transit provider's network. The transit provider will, presumably, measure this on a regular basis. The form of the guarantee may be that some maximum will never be exceeded, or that the 95th or other percentile will not exceed some value, or perhaps the average over a month will not exceed some value.

c. Packet Loss. Packet loss is an indirect measure of congestion. Where there is no congestion, one expects no packet loss. Where there is congestion, some packets will be lost. To measure packet loss a transit provider will regularly send test packets across their network, and see what percentage do not make the round trip. The test can be combined with the Latency test. Again the guarantee might be for a given maximum loss, a percentile or an average.

d. Jitter. Jitter is a close friend of latency, and measures how variable the latency is. This test can also be combined with the latency test. Jitter may be expressed as the latency plus or minus some maximum time, or maximum percentage, or again some percentile or an average.

These measures tell the customer something about the performance of the transit provider, between points of the transit provider's choosing, within the transit provider's network. However they say nothing at all about the actual service of transporting packets to and from the rest of the Internet. Indeed, the SLA will specifically exclude anything that happens beyond the borders of the transit provider's network.

Other limitations of SLAs to look for include:

a. under what conditions the guarantees may be voided – for example during routine maintenance or if service is disrupted by something which is not the provider's fault. Such clauses effectively remove any incentive for a transit provider to plan for the large-scale low-probability outages that are the focus of this report;

b. a high availability guarantee may exclude failures of the connection, and possibly the router interface – so the guarantee may look good, but it excludes parts that are likely to fail;

c. if a percentile guarantee is given, remember that 5% of a month is 36 hours, 2.5% 18 hours and so on. The transit provider may be giving themselves considerable latitude;

d. if an average guarantee is given, consider how bad things have to be and for how long to reach the given average;

e. what actual financial penalties the provider offers, and whether those are automatic or have to be claimed.

It is not surprising that Internet SLAs take this somewhat empty form. So much of what happens is outside the transit provider's control. Obviously, everything outside its network is beyond its control. Changes in its customers' traffic volumes and patterns are outside its control.

An SLA is, essentially, a bet. The provider bets the customer the value of the financial penalties that the service will meet the stated service levels – under normal conditions. In the event of an electricity failure, a flu pandemic, a software bug or a cyber-attack, all bets are off...

For day-to-day performance, an ISP may run regular tests to some selection of destinations to see how its transit providers' service varies over time. It might compare the performance of its providers using this data. The ISP may form a view about the effectiveness of its current providers, and take appropriate action. But it has little way to judge how good a replacement provider might be, though a month's trial is as good a way as any.

In terms of resilience, however, there is essentially no information available. Serious events are rare, and knowing whether a given transit provider was well prepared first requires knowledge of whether the provider was affected at all. If the customer is unable to tell how good the product is, there is no incentive on the supplier to provide a good product – and there is a market for lemons. SLAs may appear to mitigate this, but in reality do not.

The problem this poses, but we cannot answer, is what can be done about that? More information would certainly help – information about day-to-day performance, and about resilience. What is less clear is how such information might be collected and made available. For more on the issues to be addressed by SLAs for consumers see [**198**].

6.4 Transit Pricing – Zero Marginal Cost

The price of transit is tending towards zero, driven not just by improvements in technology and economies of scale, but by the inexorable logic of a zero marginal cost of supply. This is not unlike a number of other markets where costs are dominated by fixed costs and where average cost declines up to capacity, such as phone service, trains, cinemas and indeed information goods and services in general.

Transit has fallen in price continuously over the last fifteen years. Prices in the US have led the way, and as infrastructure has extended and expanded, prices in other parts of the world have followed suit. DrPeering[45] has been tracking the commodity U.S. Internet Transit pricing since 1998 [**199**], and his figures are shown using two scales, below:

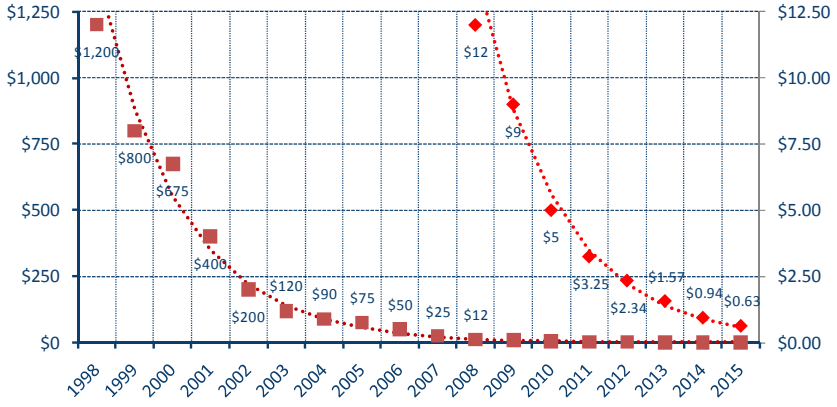


Figure 54: US Commodity Internet Transit Pricing 1998 to 2015, (US Dollars, per Mbit/sec per month) – Source DrPeering

The points shown as diamonds, to the right of the chart, are the same as the points below them, shown as squares, but on the right hand scale. The right hand scale is 1/100th of the left hand scale. The curve which roughly fits this data is a fall of 37% compound, year on year – as the chart shows, in 10 years the price drops to 1%. On the subject of this data DrPeering says:

> *The method for data collection is informal, mostly based on discussions at Internet Operations Forums from 1998 to 2010. It is important to note a few things about these informal surveys:*
>
> 1. *Data point acquisition has been difficult since transit agreements are often protected under NDA.*
> 2. *Transit pricing collected has always had a very large range as the 2006 survey results show.*
>
>   *Historically transit pricing were dependent on the level of commitment. We are assuming a low or zero level commit with these recent pricing data.*
> 3. *This survey data has been reviewed hundreds of times over a dozen years now. A seemingly equal number of people indicated that the pricing data was biased on the high side as said the pricing was biased on the low side.*

Note the "*low or zero level commit*". This means that the $5 price for 2010 is likely to be on the high side for any volume of traffic. Current pricing in the EU is under €3 at a 1Gbit/sec commitment level (< $4) and down to €1 ($1.35) for commitments in the 10s of Gbit/sec. Elsewhere in this review $3 per Mbit/sec per month is used as an indicative current figure. (In [**200**] the author suggests $15 in the US for 1Gbit/sec commitment in 2005, and $1.50 for 10Gbit/sec commitment in 2010.)

[45] *http://drpeering.net/*

Telegeography also provide figures (at a price) and occasionally publish extracts from their data. The following is taken from [**201**], and shows their "Median GigE Transit Price" in four cities:

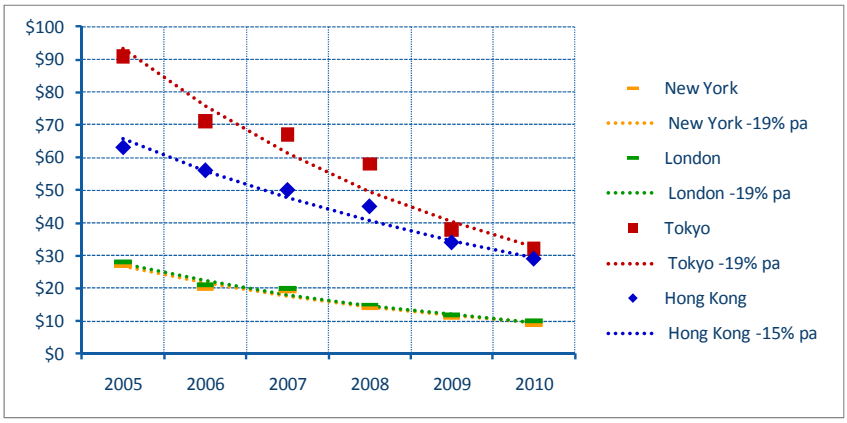


Figure 55: Global Transit Prices 2005 to 2010 – Source Telegeography

These show that there is now almost no difference between London and New York. They also show the same percentage drop per annum fits London, New York and Tokyo, but not Hong Kong, though the difference is not huge. In [**90**] the effect of increasing competition in Eastern Europe during 2009 is described.

The Telegeography figures and the DrPeering figures do not match up. Telegeography say that "*median GigE prices* [as shown in Figure 55] *offer a useful benchmark over time, prices for higher volumes of capacity can be much lower*". The DrPeering figures are for "*low or zero*" commitment, so why they suggest $5 in the US for 2010, but Telegeography suggest $10 is a puzzle. It could be that the Telegeography median is on the high side. For 2005 DrPeering suggests $75 not $28, which is more consistent with DrPeering being for no commitment but Telegeography being for 1Gbit[46] commitment level.

Nobody, however, doubts that the price has fallen continuously and steeply. And, as DrPeering notes:

> *Every year, everyone believed that the Internet Transit pricing drops could not possibly continue. And yet the gravitational forces pulling transit prices downward appears to be a natural law.*
>
> *"No one can make money at $___ /Mbps" and "the pricing has to level off now" they said. Yet every year the pricing dropped again. As you will see in the data* [above], *it is not a good time to be in the Internet Transit market as a supplier."*

There are obvious reasons for prices to fall:

- constantly improving technologies;
- economies of scale: the cost of (say) doubling the capacity of an existing network is far less than building that network in the first place;
- efficiency gains in a maturing industry;

[46] *One of the authors believes the $28 dollar price to be plausible for ~500Mbit/sec commitment level in 2005, and $4 or less for 2010.*

- when the dot-com bubble burst it left a lot of equipment and fibre valued at a fraction of its purchase price.

The triumph of the Internet is seen as a triumph of the free market, coupled to the triumph of ever more capable technology at ever-lower cost per Mbit/sec of traffic. The falling price of transit is apparently explained by the falling cost of equipment and fibre, helped out by bondholders and stock holders who saw hundreds of billions of dollars vanish in the dotcom crash.

However, in a competitive market, the issue is not how much it actually costs to provide a new customer with some amount of bandwidth, but how much more will it cost to do that – the marginal cost. Because a network grows in discrete increments, it will at any given moment have unused capacity. So, at any given moment a network can add a new customer without any increase in its costs – everything the new customer pays is profit. When there are several competing networks, two or more may be in this position at any given time, so competition drives price down towards the marginal cost of zero.

The market in the underlying network infrastructure has a similar structure. At any given moment a network provider has spare capacity, and the marginal cost of using it is either zero or small.

In a competitive market in which goods have zero marginal cost, the price will tend to zero.

This is an unsettling and counter-intuitive problem. It seems to make no sense that transit providers will drive prices down below levels which they need to charge to recoup the cost of their capital investment. But that is the logic of the market, all the available evidence supports the conclusion, and in fact it is a pervasive problem with the information goods and services industries. Similar problems have affected other parts of the telecommunications industry, where phone companies resort to techniques such as confusion pricing to maintain revenues; and to the software industry which is dominated by monopolies enforced by technical lock-in and network externalities.

In Appendix II we look at a number of major transit providers to see what may be learned by looking at their accounts. The numbers suggest that the telecommunications business is tough, but transit is generally a relatively small part of the major providers' business, so it is hard to discern the effect of falling transit prices. From the brief analysis we can, however, see:

- how many of the current major transit providers have either been through Chapter 11, or have acquired components of their business from Chapter 11.
- losses have been heavy in the past, though have generally reduced. Among the mainly Internet businesses:
    - Level 3 lost $3.5 Billion in 2005 to 2009 (average $707 Million per annum), despite several acquisitions reduced losses to just $618 Million in 2009 (on revenues of $3.7 Billion).
    - Global Crossing lost $1.4 Billion in 2005 to 2009 (average $282 Million per annum) – losses in 2009 were $141 Million (on revenues of $2.5 Billion).
    - Savvis lost $175 Million in 2005 to 2009 (average $35 Million per annum) – losses in 2009 were $21 Million (on revenues of $607 Million).
    - Cogent lost $453 Million in 2001 to 2009 (average $50 Million per annum) – losses in 2009 were $3.8 Million (on revenues of $236 Million).

- Abovenet has moved from losses of $36 Million in 2004 (on revenues of $189 Million) steadily to profits of $95 Million in 2009 (on revenues of $360 Million). Abovenet came out of Chapter 11 in Sep-2004, and has focussed on enterprise rather than wholesale customers since then.

Level 3, Global Crossing, Savvis and Cogent are ranked 1st, 2nd, 5th and 12th by Renesys.

- there are a wide range of sizes of supplier: at one end, Tinet with revenues of $53 Million in 2009, some part of which is transit, to Level 3 and Global Crossing whose wholesale or carrier arms had revenues of $2 Billion, each, in 2009.

  Cisco's projections [**14**] suggest total Internet traffic of 15,205 Peta[47]Bytes/month for 2010, which divides down to 70,000 Gbits/sec[48] peak traffic– the Euro-IX 2010 annual report [**16**] reports 4,400 Gbits/sec peak traffic across all European IXPs, so the numbers seem reasonable. If 50%[49] of that is carried by transit providers, and if Level 3 carries 20% of all transit, then its share would be 7,000 Gbits/sec, which at an average of $3 per Mbit/sec per month is ~$250 Million per annum. The CAIDA data [**82**] shows Level 3 as having approximately 2,500 directly connected ASes, so Level 3's average capacity per customer would be ~3 Gbits/sec , which does not seem unreasonable and suggests that our rough estimate of transit revenues is not orders of magnitude out.

  If all of Tinet's $53 Million revenue were transit (which we do not believe is the case), then at $3 per Mbit/sec per month it might be carrying ~4.2% of all transit. At $2 per Mbit/sec per month, it would be carrying ~6%.

  Finally, 35,000 Gbits/sec of transit at $3 Mbit/sec per month makes the total transit market worth $1.3 Billion per annum; at $2: $840 Million. For comparison, one transatlantic cable system, Flag-Atlantic, cost ~$1.1 Billion to build to its initial capacity of 160Gbit/sec in the early 2000s [**202**].

The following quotes from 2009 accounts reinforce the picture of a difficult market:

- Level 3: *"The Company continued to experience price compression in the high-speed IP market in 2009 and expects that pricing for its high-speed IP services will continue to decline in 2010."*.
- Global Crossing: *"Revenue attrition generally results from market dynamics and not customer dissatisfaction. Pricing for our VPN and managed services products has continued to decline at a relatively modest rate over the last few quarters, while pricing for specific data products such as high-speed transit and capacity services (specifically internet access arrangements used by content delivery and broadband service providers) has continued to decline at a greater rate."*
- Cogent: *"We believe two of the most important trends in our industry are the continued long-term growth in Internet traffic and a decline in Internet access prices within carrier neutral data centers . As Internet traffic continues to grow and prices per unit of traffic continue to decline, we believe our*

[47] *Giga: 10^9; Tera: 10^{12}; Peta: 10^{15}; Exa 10^{18}; Zetta: 10^{21}.*

[48] *This is on the basis that traffic varies over the day in a sinusoidal pattern, with the minimum traffic being ⅓ of the peak, so the average is ⅔ of the peak; so 15,205 PetaBytes is ~(15,205 × 8 × 10^{15}) / (10^9 × 30 × 24 × 60 × 60 × ⅔) Gbits/sec.*

[49] *Which may be on the high side, given the increase in CDN traffic.*

ability to load our network and gain market share from less efficient network operators will continue to expand. However, continued erosion in Internet access prices will likely have a negative impact on the rate at which we can increase our revenues and our profitability."

- Abovenet: *"The Internet connectivity business is intensely competitive and includes many providers such as AT&T, Verizon, Level 3 and Cogent. As a result of this competition, while Internet traffic has continued to grow at a substantial rate over the past five years, pricing has generally declined, which has negatively affected revenue growth."*

In [**203**], in Nov-2008, it is observed that:

"... business models are in trouble because of price erosion driven by vicious competition. Level 3 and Cogent are routinely blamed for the sharp decline in prices (to levels below $3 per megabit per second in gigabit and above speeds).

Note the $3 price in 2008, which makes the pricing for 2010, given above, look a little optimistic.

The Renesys rankings of the largest providers, published occasionally in their blog [**85**] [**204**] [**205**] make interesting reading. The following shows their rankings for two and a half years to the end of Jun-2010:

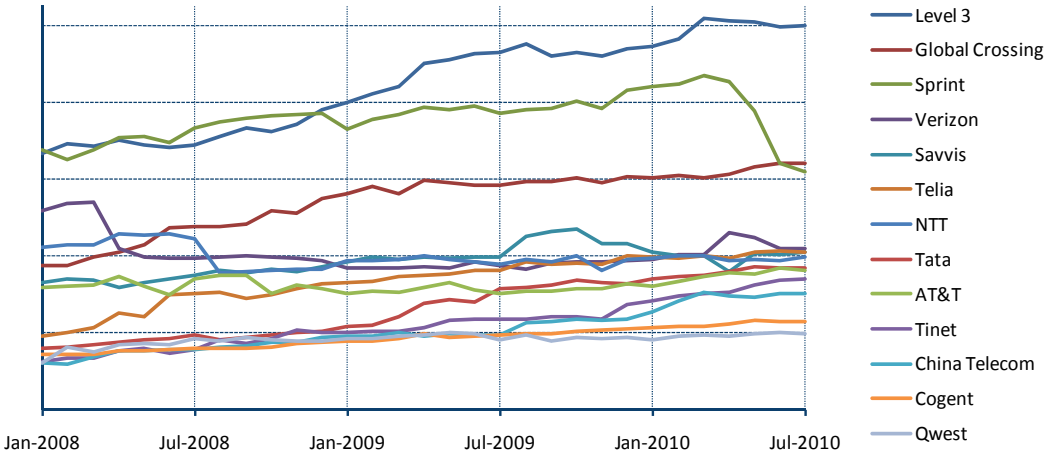


Figure 56: Renesys Top 13 Providers Jan-2008 to Jun-2010 – Source Renesys

The table opposite shows the state at the end of Jun-2010, and compares that with the state at the start of 2008. How networks have moved in the rankings is shown in the second column – so Level 3 has moved up +1, Global Crossing up +3, but Sprint has slipped -2 to third place.

1	+1	Level 3
2	+3	Global Crossing
3	-2	Sprint
4	-1	Verizon
5	+1	Savvis
6	+2	Telia
7	-3	NTT
8	+1	Tata
9	-2	AT&T
10	+1	Tinet
11	+2	China Telecom
12	-2	Cogent
13	-1	Qwest

Table 2: Renesys Rankings End Jun-2010

The larger telecommunications companies have slipped back. The recent fall at Sprint looks dramatic, though it is not known what significance may be put on the relative scale of the ratings. In their 2009 accounts [**206**] Sprint note:

"Some competitors are targeting the high−end data market and are offering deeply discounted rates in exchange for high−volume traffic as they attempt to utilize excess capacity in their networks."

Whether what we see here is Sprint no longer being prepared to match competitors' prices, we cannot say. At the end of Oct-2010 [**200**] reports that Sprint regained second

place, but the feeling remains that some providers have "*refocused their sales activities on enterprise-managed services where pricing and margins proved more stable.*".

Renesys's own conclusion is:

> *"Internet transit is an extremely tough business [**90**], one with ever falling profit margins. With lower Internet penetration, fewer competitors and higher margins, the Middle East and Asia have provided somewhat of a refuge for providers who can operate effectively in these geographies. As a result, you can expect many traditional US-centric carriers, such as AT&T, Sprint and Verizon, to either grow very slowly or decline, while those with strong global diversity, such as Level 3, Global Crossing, Tinet and Tata, should continue expand proportionally to the markets they serve. And if older, less nimble players "leave the field", such departures might just relieve some of the extreme pricing pressure found in the industry today, allowing the rest of us to continue to enjoy all that great Internet "content", but at slightly higher (and more sustainable) pricing levels."*

For the very large networks the only thing worse than selling transit is not selling transit. If a large network can persuade some large customer to buy transit, then that is a slice of the large network's traffic for which it is being paid. The alternative is to source that traffic from a peer or, worse, a transit provider, which not only costs money, but also puts some money in a competitor's pocket. Selling transit is the cheaper option. Related to this is the notion of 'on-net' customer traffic, which is traffic flowing from one customer to another customer, both of whom are paying the transit provider – unlike the case where traffic flows between a customer and a peer or, worse, a transit provider. In Table 1 (on page 91) we see that the largest networks have 80% and more of all IP addresses in their 'customer cones', whether that means that 80% of their traffic is on-net, we simply cannot tell. The decision to provide transit at the market rate becomes a strategic one, more related to the impact on the provider's costs and less related to transit as a business. Further, where a transit provider can use network capacity for services with better margins than commodity transit, we may expect them to do so, perhaps reducing the capacity used for, or available to, transit.

For the ordinary ISP, the falling cost of transit offsets to some extent the ever-increasing demand for bandwidth, though that cost is an ever-reducing part of their costs. To get an idea of the scale of this, consider providing customers with an 8Mbit/sec service, at a 20:1 contention ratio[50], where 50%[51] of traffic is sourced from a transit provider at $3[52] per Mbit/sec per month; the monthly cost of transit for each customer is $0.60. Ten years ago, when transit cost 100 times as much, things were different, in particular access rates were 16 times less and contention ratios a bit higher. The falling cost of transit also has a knock-on effect on peering, which is discussed in Section 6.5 below.

[50] *Contention ratios are contentious. A ratio of 20:1 used to be for the higher quality services, 50:1 was common for domestic DSL service. But the demands of video reduce the contention ratio that an ISP can get away with. However, in this context, more video traffic will tend to reduce transit traffic, since that is more likely to be delivered by a CDN across a peering connection – in some cases a paid peering connection in which the ISP is paid.*

[51] *This is probably on the high side for an ISP which peers effectively at a local IXP and connects to the major CDNs.*

[52] *This is also probably on the high side for any significant volume.*

The combined impact of the CDNs, the increase in video traffic (delivered increasingly by CDNs) and the relative fall in P2P (file sharing) traffic should not be underestimated. It can be seen in the projections given by the Cisco Visual Networking Index (VNI) [**14**]:

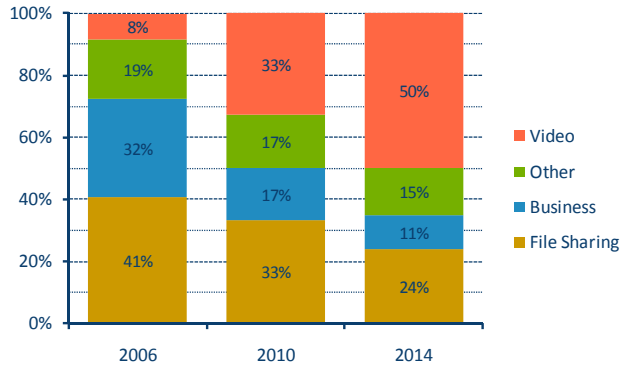


Figure 57: Projected Share of Traffic Types – Source: Cisco VNI

which shows the percentage of total Internet backbone traffic which video and other broad categories represent. If we assume that the video traffic more or less maps to the CDN traffic, then the impact of the CDNs is clear[53]. The effect of the rise of video content is discussed in [**207**].

Taking the total traffic in 2010 as '3', the following shows the growth in traffic as well as the change in its relative proportions:

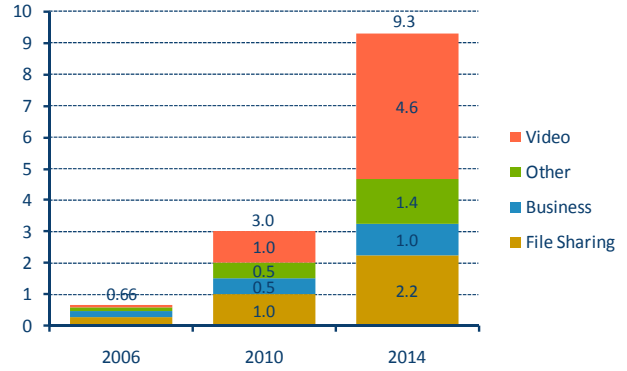


Figure 58: Projected Traffic Volumes, Relative to 2006 – Source: Cisco VNI

which shows overall traffic growing at 34% per annum, compound, between 2010 and 2014; but video traffic growing at ~45% and all other traffic at ~20%. Note that while P2P is less popular, it is still growing.

If we assume that video traffic is predominantly delivered by the CDNs, this means that transit traffic may be expected to grow ~20% per annum, compound. If transit prices fall by 37% per annum (as

[53] *There is a caveat here: in [**13**] it is observed that Akamai, who claim to deliver 20%* "of the world's Internet traffic" *[**224**], or* between 15-30% "of all Web traffic" *[**225**], bypass the interconnection system altogether for some proportion of their traffic, by locating servers within some (the larger) ISPs. Unfortunately, it is not clear whether the Cisco figures for 'Internet' traffic, which they define as* "all IP traffic that crosses the Internet backbone" *– which is what matters to the transit providers – takes this into account. The figures suggest that Cisco are treating all Akamai delivered traffic as 'Internet' traffic, and that is what we are assuming.*

suggested by DrPeering) then transit revenues fall by ~24% per annum. If transit prices fall by a more modest 19% (as suggested by Telegeography) then revenues fall by ~3% per annum. The transit providers are being squeezed by each other and by the CDNs – delivering more traffic each year, while total revenues shrink. Level 3 has moved into the content delivery market, and other providers have arrangements with independent CDNs.

Following the dotcom boom and the bursting of the bubble in 2000, networks acquired large amounts of fibre and equipment at a huge discount – either via Chapter 11 or from receivers. This helped create many of today's large networks. So, where networks now need to expand, they usually have the fibre they need, but must buy new equipment to light new wavelengths and carry more traffic. If their pricing has been based on the hugely discounted cost of the fibre and equipment acquired after the bubble, the pricing may not properly account for the capital cost, today. On the other hand, the cost of equipment has been falling sharply with technological advance and increased competition. So the networks' average costs are also falling steadily.

Encouraging and promoting competition has been the goal of deregulation in the telecommunications industry, and others. The result has been better services at better prices for users. However, nothing in this brief survey of the state of the transit business suggests that the market is healthy. It may be that the provision of transit can be viewed by providers as a cost of doing business. For every ISP there is a strong incentive to reduce costs and maximise network utilisation – reducing the spare capacity that is key to resilience.

The data this is based on is patchy and occasionally contradictory. The sense that "things cannot go on like this" has been strong for years. The Internet has survived numerous bankruptcies, some huge, some simply very big. Perhaps it does not matter if transit does not pay for itself, though it is a key part of the interconnection system. Perhaps we do not understand the economics, and it does pay for itself. Like many other things about the interconnection ecosystem, we do not appear to know enough.

6.5 Peering and IXPs

For all but the large transit providers, peering is a way for an AS to get a good connection to other ASes locally and reduce its bill for transit. A good proportion of Internet traffic is exchanged locally, so an ISP might source and sink 30%-60% of its traffic at a good IXP. The CDNs have good reasons to connect to IXPs, so with the proportion of traffic that an AS may source at an IXP is increasing. The Euro-IX 2010 annual report [**16**] reports that between 2009 and 2010 traffic at all European IXPs increased by 63%, where the increase in all traffic thought to be 35%-45%.

For an AS, the cost of connecting to an IXP is shared across all the peering connections made there, so once connected there is a financial incentive to peer with as many other ASes as possible, which encourages diversity of interconnection and is generally good for resilience. However, to be economic, the cost of peering must be comparable to the cost of sourcing the same traffic by transit; [**208**] discusses the business case for peering.

A pair of 10GE ports (10Gbits/sec capacity, each) at a large IXP might cost $0.60 per month per Mbit/sec (assuming total peak traffic of 7Gbits/sec – which is the amount that fits comfortably into just one connection, so this is assuming a full 1+1 redundant connection). Compared to ~$3 per month per Mbit/sec for transit the decision is simple – though not as simple as two years ago, when transit cost more but the IXP did not. A pair of 1GE ports, however, weighs in at more like $1.50 per

month per Mbit/sec (assuming 700Mbits/sec total peak traffic). So still a saving, but not as exciting a saving – though at this lower level of traffic, transit may cost a little more.

Peering connections are preferable where possible. So these figures suggest that transit prices can fall further before peering at an IXP will start to look expensive. Moreover, taking the $0.60 per Mbit/sec cost at an IXP, even if transit falls to $0.10 per Mbit/sec, the extra cost of peering would be $3,500 per month for 7Gbits/sec, or $0.025 per customer (at 20:1), so with these small numbers the advantages of peering could well outweigh the overprice. And, of course, IXP costs fall with improving technology.

The IXP costs we have looked at are just the cost for the IXP itself, they do not include the costs of the connection(s) to the IXP, the router port(s) at the AS end, the operational costs of making and looking after the peering connections and so on. When peering at a local IXP those are likely to be relatively modest (and mostly one-time costs), but in any case similar to the cost of connecting to a local transit provider. So for a local IXP the analysis above stands.

Connections to more distant IXPs are another matter. The falling cost of circuits between clusters of sites has made such connections more cost effective, and enabled some ASes to diversify their connectivity, improving their resilience and the resilience of the system as a whole. However, the cost of the circuits is a significant part of the cost of traffic exchanged at the distant IXP, so for these beneficial connections to continue, the circuit costs will need to stay below transit costs, at least to the point that the absolute cost becomes so small that the benefits outweigh any overprice.

DrPeering in [**199**] takes the view that "*With Internet Transit Pricing dropping so fast, peering using Public Peering ports will become very difficult to justify financially*", which suggests that peering, at least at IXPs, is under pressure. The picture is complicated, but given the value of local exchange of local traffic, and of the diversity of connections supported by IXPs, it would be good to understand this better.

In any event, since a large proportion of total traffic is local traffic, supporting resilient local interconnection may well be a good step toward supporting the resilience of the interconnection system.

6.6 Misunderstanding the Risk

The Internet does not offer full guarantees or minimum service levels. At any time its performance depends on demand, and TCP pushes back on applications when congestion occurs. To put it another way, when we use the Internet there is some risk that we will be disappointed – either because we cannot reach somewhere, or because the service to somewhere is slower than we expect. That risk is reflected in the limited nature of any SLA offered.

The Internet works well most of the time. It is inexpensive. In fact, compared to all previous networks, the Internet is extremely inexpensive. So moving applications from existing systems to the Internet saves money. For example, 'Cloud Computing' exploits economies of scale to reduce the cost of common applications, and to allow them to be used from anywhere – all driven by universal and cheap Internet access.

There are many questions here:

1. are the risks properly explained?
2. are they properly appreciated?

3. do the cost savings blind users to the risks?
4. what about the social costs?

How resilient do we actually expect services that depend on the Internet to be? If constant, high-quality access to 'the Cloud' is accepted as essential to how we run our lives and our businesses, have we fallen into a trap of false expectations? Do we, say, expect the service to be available 99.5% of the time – so that there are no more than ~4 hours unavailability per month? If not, what do we expect, and is that what the interconnection system can provide? What are the costs to the economy and to our way of life of losing Internet service for more than (say) three days? If those costs are as catastrophic as the costs of losing electricity supply for a comparable period, then how should these social costs be reflected in the incentives facing firms who provide various aspects of Internet service? At present, as we have discussed, their SLAs exclude all effective liability for extended or systemic failures.

6.7 Introducing New Incentives

In this review of the Internet interconnection system and its resilience, the following appear to be key:

a. resilience is an externality for the key parts of the system;
b. customers cannot tell whether their suppliers contribute to the resilience of the overall system, or not;
c. in fact nobody knows how resilient the system may be, so there is no market signal for better resilience.
d. inexorable price erosion is reducing the numbers of large transit providers, and squeezing revenues for the remaining ones.

If we look on the question of resilience as a 'safety' issue, then there are parallels in other industries – where customers do not have the information to judge how safe a given product is. There are further parallels with other forms of network where resilience is both important and to some extent an externality, such as electricity supply. Regulatory input is often considered necessary in such industries. For example, in the electricity supply business, regulators constrain markets by adding taxes to free market prices that are then used to pay for reserve generation capacity and network redundancy. In the airline industry safety standards must be met, and all incidents and accidents are thoroughly investigated by an independent body.

6.8 Government Intervention

Government intervention has a patchy record. From time to time governments try to encourage development of infrastructure where the market appears to need a push, for example:

- regular enthusiasms for improving the breadth and depth of Internet penetration in Africa, using aid/subsidy from the developed world;
- subsidising local IXPs, such as a proposal in the Netherlands to build satellite sites, connected to the AMS-IX, in depressed areas, and the launch of ScotiX to serve the Scottish Internet, and act as a magnet for investment and development in Internet businesses in Scotland (it attracted very little traffic and has vanished without trace);

- local loop unbundling – though that is far removed from the interconnection system.

Some Governments would like the Internet to be more like the telephone system. China has been pushing within ITU for settlement based peering to be required; that is to say, they want a scheme in which the cost of traffic to and from China is met in part by the other end. In the past the Australians have expressed a similar view: that it was not fair that they had to pay for long circuits to the USA, and that US carriers should cover part of that cost. But with the falling cost of transit and long-haul circuits, they are no longer concerned. Many speculate that Chinese enthusiasm for settlement based peering in fact stems from a wish to see more logging and hence more political control[54].

Some people argue that as the Internet is global and the large transit providers are multinational businesses, it would be difficult for any government to intervene to influence, say, levels of spare capacity in the large transit providers. This argument is no longer made in respect of multinational banks, and presumably if a global system-wide failure of the Internet were sufficiently severe to cause economic disruption, governments would try to influence large ASes domiciled in their jurisdictions so as to minimise the probability of a recurrence. This might involve direct regulation, or less direct means such as using the public sector's purchasing power to favour service providers who met minimum standards for capacity, diversity and disaster planning.

[54] *http://news.bbc.co.uk/1/hi/8417680.stm*

7 Is there Cause for Concern?

In general the Internet works wonderfully well. Every year it connects more people, delivers more traffic, provides more services, creates new opportunities, and enables greater efficiencies. And, every year costs fall.

So far the problems have been minor. The occasional natural disaster disrupts service locally, and echoes may be felt around the world. Occasionally a system fault causes a system wide hiccup, but those are generally cleared in hours. We would be better off without bot-nets and the like, but nobody would turn off the Internet to achieve that. So the big question is whether there is sufficient cause for concern to do anything more.

In this section we consider:

- in Section 7.1 we ask whether we have realistic expectation for the resilience of the system;
- Section 7.2 reviews the issue of how little good information we have about the system, and what that means for our ability to assess its resilience;
- in Section 7.3 we look at essential role of the major transit providers in the resilience of the system;
- Section 7.4 touches on the lack of any continuous or rigorous monitoring of the system;
- finally, in Section 7.5 we consider the perception and reality of risk to the system.

7.1 Realistic Expectations

We expect the system to be as resilient as possible, given what we wish to pay for it and given what we consider to be ordinary events and what we consider extraordinary.

We already discussed four possible disaster scenarios that could lead to global disruption of service;

1. a regional failure of technical infrastructure on which the Internet depends, such as electric power;
2. problems with the skilled labour force on which it depends, such as a flu pandemic;
3. a coordinated attack, most probably on the routing infrastructure, though that is not the only technical possibility;
4. a technical failure resulting (for example) from software bugs in routers or other critical infrastructure.

We believe that these scenarios are realistic. They are all low-probability but the probability of none of them is zero. The impact is less easy to predict. If half the capacity of the Internet were disabled, things would no doubt be difficult for a while. The problem of mapping realistic events to realistic impacts is addressed below.

The next question is: what degree of resilience is it realistic to expect? It is not reasonable, for example, to expect civilian networks such as the Internet to remain up for an extended period of time in the absence of electric power – but how much diesel should switching centres have for their backup generators? Three hours, three days, or three weeks? It may make sense for a military network centre to be provisioned for months, yet it's difficult to justify keeping an IXP running for

three weeks if almost all of the ASes that peer there would cease operations within three hours of a power cut. So, what do we expect and under what conditions do we expect it?

7.2 The System is Opaque

Attempts to translate scenarios into their impact on the system and on services run up against the problem that the system is opaque. Its design, the daily problems it copes with routinely, and the occasional larger events that are dealt with, all strongly support the belief that the Internet is resilient. Some may believe the Internet is so capacious and diverse that no realistic future event could have a materially greater impact than past events.

However, as our economies, or standard of living and even our survival come to depend on the Internet, there comes a point beyond which faith is not enough. How can we test and verify its resilience, given that:

a. the Internet continues to grow and develop rapidly;

b. we must consider events on a scale not encountered to date;

c. its criticality is growing steadily– for example, as applications move from the desktop to the cloud.

When considering how resilience might be verified, we run into a number of problems:

a. the system is extremely big. Modelling it as a core or 'virtual backbone' and its clients reduces the complexity; the virtual backbone is physically a number of clusters of sites, and the fibre networks within and between them. But even so, a map of the Internet– if there was one – would be extremely big and extremely complicated, and constantly changing.

b. mapping the connections between ASes is very hard, and we cannot do this from the outside. It is not enough to know the logical connections; we would need to know the number of physical connections, their separacy and their capacity. We might start with large transit providers, and some others – but this information is considered commercially sensitive.

c. as we do not have good maps of the physical infrastructure, it is hard to construct scenarios and analyse, say, the effect of a flood in London Docklands or an earthquake in San Francisco. It also hampers efforts to improve resilience because it is hard to ensure separacy. In fact, such maps as there are are confidential because of 'security' [**209**], but it is not clear whether the benefit derived from not assisting attackers is greater than the harm done by making it difficult to manage separacy.

d. as we do not have maps of traffic flows and volumes, it is hard to concentrate attention on the parts of the system that matter, to assess spare capacity, or to predict what traffic will be diverted to which route in the event of a regional disaster.

In short, the system is opaque. But perhaps it is unrealistic to expect to be able to assess the resilience of this huge system at any level of detail.

7.3 Resilience at the Transit Provider Level

Perhaps it is sufficient to focus on the core of the system – on the transit providers, the major CDNs and the IXPs. If this core keeps working then most other failures can be fixed in time, and hopefully

in not very much time, as most ASes have an arrangement with at least one transit provider as a backup.

Perhaps the contract between transit customer and transit provider, and the competition between providers, can provide a sufficiently strong incentive for each provider to ensure that its service is adequately resilient. The very limited information available to customers, and the very limited SLAs offered by transit providers, suggests that the main incentive is not losing customers to competitors – which means being no worse than the competition, most of the time.

It could be that in any realistic scenario:

a. levels of resilience/redundancy within networks mean that relatively little traffic will actually spill over between ASes;

b. that the spare capacity that the networks maintain, for their own purposes, is sufficient to cope with any realistic level of spill over.

In which case, the resilience of the system as a whole is maintained as a side effect of the resilience measures taken by each AS independently. But this is hard to verify. The resilience of the system as a whole depends on the individual ASes (particularly the large transit providers) having the necessary spare capacity to cope when traffic spills over following some event. This resilience suffers from strong externalities: individual providers do not face the full social costs of failure of the system so may have insufficient incentive to provide enough redundancy and resilience. There are also strong asymmetric-information effects in that the typical transit customer has no way of assessing its providers' resilience. Furthermore, the zero marginal cost of transit has led to an inexorable fall in transit price and an apparent reduction in the number of transit providers, leading some to question whether the system, in its current form, is sustainable.

7.4 Lack of Monitoring

The asymmetric-information problem at least might be mitigated if the levels of resilience offered by the different transit providers could be monitored. However nobody monitors even the performance of the Internet interconnection ecosystem, let alone its resilience! Although many third parties have run projects to collect data on different aspects of the Internet, there is no consistent and sustained effort.

When significant incidents occur, interested parties may try to find out what actually happened, and what the impact was. The operators involved may assist or, indeed, mislead efforts to analyse the event. Quite quickly the incident moves into the folklore, and everybody moves on.

Compare this with, say, the airlines. Incidents large and small are investigated, analysed and lessons learned and disseminated. Perhaps there is something to learn here.

7.5 Perception and Reality of Risk

The Internet is a success story. Within the industry there is a belief that the risk of a major incident having a major impact is mitigated by the Internet's size and diversity, and this view is reinforced by the system's ability to cope so far. There is, at least, a reasonable understanding of the limitations of the open Internet, and the risk that sometimes, some things will not work well, or not work at all.

Outside the industry these risks are not well understood – possibly because they are not well explained, or possibly because they are hard to understand. Firms are coming to depend for all sorts

of critical services on a network that is designed explicitly to give no guarantees at all. Offices move their word-processing and spreadsheets into the Cloud; supermarkets rely on the Internet for just-in-time stock replenishment; hospitals and electricity utilities put more and more of their communications over networks which, although they may be VPNs, still depend on the Internet. And assumptions about backup channels of communication are becoming wrong: more and more phone services, for example, are carried at least part of the way over an IP network. And because they do not understand the risks, customers are not applying suitable pressure on their suppliers. As a result, more of our economic activity is coming to depend on the Internet than is prudent given its likely level of resilience against low-probability high-impact events.

PART III – the Report on the Consultation

Introduction to the Report on the Consultation

The consultation for the "Inter-X" study on the "Resilience of the Internet Interconnection Ecosystem" comprised a questionnaire with 15 questions – some with more than one part. The questionnaire was sent out to a broad range of stakeholders, and 36 responses were received, from a range of respondents:

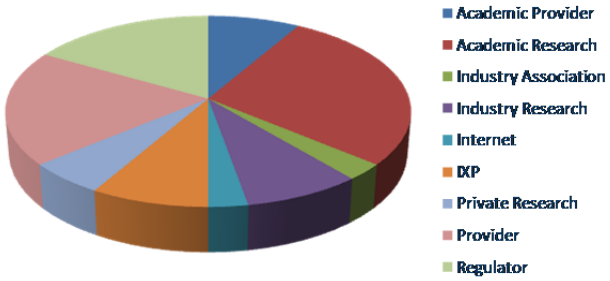


Figure 59: Breakdown of Respondents

The study covers a large and complex issue. The Internet interconnection system has many layers, which interact with each other particularly strongly when we look at its resilience properties. Resilience itself is a difficult concept to pin down. A number of people responded saying that they were either not confident, or did not feel competent enough, to respond, but that they would very much like to see the results of the study.

This report on the consultation is divided into three sections:

1. general themes or points identified. The responses to the questions were quite varied. In the analysis of the responses some general themes emerged, in many cases from responses to more than one question. Those general themes are presented in this section, supported by some selected quotes which are particularly apposite;
2. the questionnaire and the responses. This presents the questions and a summary of the responses to each one. The responses were solicited on the basis that the published results would be a summary of the actual responses, and that no attribution would be made or be readily deducible;
3. the 'Introduction to the Study' that accompanied the questionnaire. This was intended to set the context for the questions.

The questionnaire was quite open-ended, so the analysis is necessarily qualitative.

Respondents

Our thanks go to all those who gave their time freely to help with this study and responded to the questionnaire:

Olivier Bonaventure	Professor	UCLouvain, Belgium
Scott Bradner	University Technology Security Officer, Office of the CIO	Harvard University
Bob Briscoe	Chief Researcher	Networks Research Centre, BT Group plc
kc claffey	Principal Investigator	CAIDA
Andrew Cormack	Chief Regulatory Adviser	JANET(UK)
Jon Crowcroft	Marconi Professor of Communications Systems	Computer Lab, Cambridge University
John Curran	CEO	ARIN
Dai Davies	General Manager	Dante
Nicolas Desmons	Chargé de Mission	ARCEP, France
Amogh Dhamdhere	Post-Doctoral Researcher	CAIDA
Giuseppe Di Battista	Professor of Computer Science	Roma Tre University
Nico Fischbach	Director, Network Architecture	Colt
Mark Fitzpatrick	Engineer	Federal Office of Communications, OFCOM, Switzerland
David Hutchison	Professor of Computing	Lancaster University
Malcolm Hutty	Head of Public Affairs	LINX
Christian Jacquenet	Director of the Strategic Program Office	France Telecom Group
Balachander Krishnamurthy	Researcher	AT&T Labs Research
Craig Labovitz	Chief Scientist	Arbor Networks
Ulrich Latzenhofer		Rundfunk und Telekom Regulierungs, Austria
Simon Leinen	Network Engineer	SWITCH
Tony Leung	Global Internet and Network Convergence Manager	REACH
Kurt Erik Lindqvist	CEO	Netnod
Neil Long	Researcher and Founder	Team Cymru Research NFP
Patricia Longstaff	David Levidow Professor of Communication Law and Policy James Martin Senior Visiting Fellow, Oxford Martin School Visiting Scholar	Syracuse University Trinity College, Oxford
Paolo Lucente	Architect/Designer	KPN International
Bill Manning		USC/ISI
Maurizio Pizzonia	Assistant Professor, Computer	Roma Tre University

	Science	
Andrew Powell	Manager of Advice Delivery to the Communications Sector	UK Centre for the Protection of National Infrastructure
Edwin Punt	Product Manager	KPN International
Bruno Quoitin	Assistant Professor	University of Mons
Anders Rafting	Expert Adviser	Swedish Post and Telecom Agency
Jennifer Rexford	Professor	Department of Computer Science, Princeton University
Stefan Stefansson	Network Security Specialist	Post and Telecom Administration in Iceland
David Sutton	Director	tacit.tel (Telecommunications and Critical Infrastructure Technologies) Limited.
Guy Tal	Director of Strategic Relations	Limelight Networks
Rob Thomas	CEO and Founder	Team Cymru Research NFP
Nigel Titley	Head of Peering and Transit Strategy	Easynet/Sky
Andreas Wildberger	Generalsekretär	Internet Service Providers Austria (ISPA)

8 General Themes or Points

This section presents seven general themes or points which came out of the consultation.

1. Complexity and Lack of Data: observations on the complexity of the system and the lack of good data about the system – which is exacerbated by the tendency to treat information about interconnection as a commercial secret.
2. Resilience Issues: relating either to resilience in general and to resilience of the interconnection system in particular.
3. Physical Layer: issues to do with the physical infrastructure that supports Internet interconnections, from the sites and their dependency on electricity supply, through the fibre cables that run within and between sites, up to the equipment that does the routing and forwarding.
4. Network Layer: mostly observations on the limitations of BGP.
5. Operational Layer: comments on operational aspects of dealing with a crisis or disaster.
6. Contract and Economic Layers: observations on the incentives for resilience, broader economic and contractual issues (in particular SLAs) and the general 'tragedy of the commons'.
7. Regulatory Layer: the desirability, or otherwise, of regulatory interference,

The themes labelled **[C:xx]** are where there was some general agreement by the respondents. Given the variety of the responses it is not possible to identify full consensus, but these themes at least reflect some common views. The points labelled **[Q:xx]** are quotes from individual respondents which were particularly interesting or telling.

8.1 Complexity and Lack of Data

[C:1] Scale and complexity

The sheer scale and complexity of the system are such that it is hard to understand it, let alone to assess how well it might work in a major crisis. The complexity comes not only from the large number of different networks and different interconnections, but also from the number of distinct layers, each with its own properties, and the relationships and dependencies between those layers.

"The Internet interconnect ecosystem is complex with many layered dependencies, which is what also makes it uniquely robust, and at the same time make it hard to oversee and grasp."

"One barrier to certainty in the eco-system's resilience is that the enormous system complexity makes it extremely difficult to understand fully, and so impossible to predict outcomes or to measure the overall system resilience."

"...it is very difficult to assess exactly where traffic will flow in the event of failures."

"The ecosystem not being a single unified domain [means that it] *is difficult to verify or assess. ...*[also] *BGP achieves scalability by hiding non-optimal information."*

"...resilience at the scale of the whole ecosystem is harder to achieve..."

"Everything beyond your directly connected peers is mostly out of your control. The whole ecosystem is so dynamic in nature that you can't model it or adapt your resiliency to it..."

"All of the Internet protocols and services are key components, from the physical to the application layer, because each is a potential point of vulnerability in the face of faults or challenges to the ecosystem."

[C:2] Lack of information and trade secrecy

A common theme was the shortage of information about who interconnects with whom, how they interconnect (number of connections, capacity of each, routes exchanged, etc.) and what traffic is exchanged. It was noted that this reflects the difficulty of observing this information, and the secrecy surrounding interconnection. For example:

"Publicly available data about the Internet's AS-level topology is known to be incomplete..."

"...no one shares data with each other and all peering details are considered trade secrets..."

"How does one measure resiliency when so many of the agreements and topologies are protected by NDAs?"

"Access to infrastructure data / statistics is extremely challenging. As a result, most research focuses on end-systems and small components. There continues to be a need for broader access to data and systemic studies of broader inter-dependent infrastructure."

"Unless networks see an incentive to share their connectivity (and no such incentive currently exists), the research community will have to resort to ad-hoc methods to obtain more connectivity information, e.g., targeted traceroutes over IXPs to determine peering links, discovering AS links through traceroutes, etc."

"Transparency is impossible given the competitive nature of the business."

"Other non-technical approaches -- such as requirements by governments to provide topology and shared-risk information -- would help, though there would be significant resistance to this kind of government involvement." [in the context of improving resilience by identifying shared infrastructure.]

[C:3] Difficulty of estimating how the system will respond to events

For a number of reasons, including both the complexity of the system and the lack of information at every level, it is difficult to estimate how the system will respond, and hence difficult to assess its resilience.

"... it is very difficult to assess exactly where traffic will flow in the event of failures."

"Hidden failures (i.e. untested backup paths, traffic load capacity, etc.) are also significant dangers."

"...a WDM system failure might lose a wavelength over which one operator runs IP but over that IP network perhaps several other operators are running Layer 2 VPN or PWE3 (pseudo-wire) services. Identifying this type of cascading failure and layered dependencies is virtually impossible."

"[Because] *you don't know how your neighbours will send you traffic or handle the traffic you send to them, it is even harder to predict where that traffic will go when a failure occurs."*

"...impossible to predict outcomes or to measure the overall system resilience."

[C:4] Value of studying previous incidents

It is well known that the Internet suffers incidents at various scales all the time – these may be seen as random experiments that test its resilience, and identify weaknesses.

"A first solution is to look at the past and study all the failures that happen. This implies that enough data is available about the failures of ISPs. For this, I'd suggest to evaluate the possibility of forcing network operators to publicly release information about their failures that affect a given fraction of their customers. This is used in the US by the FCC, at least for the telephone service, and gives good results in the long term because it penalises the operators that have regular failures. The US also used a similar trick to force database owners to release information about database breaches. This has forced database owners to take breaches more seriously than in the past."

"For example, the Pakistan/YouTube incident resulted in many millions (tens of millions) of people experiencing a routing failure with the result that an Internet resource was unavailable for them to a couple of hours."

"Some information is available from observations of incidents… [which] *may indicate areas where tabletop exercises are useful."*

"Post-incident investigation and discovery…" [to assess resilience.]

[Q:1] The need for data

"…be useful to gather on a regular basis qualitative and quantitative data allowing the better understanding of the state of the Internet resilience…"

[Q:2] The essential role of information about the system

"Resilience requires a trusted source of information about how the system is working – the good and the bad. If specific incidents are not made public for security purposes the aggregate data and specifics without identifiers are critical."

[Q:3] The Internet is only one of a number of interdependent networks

"Look at inter-dependencies of multiple network types (energy, transport, etc)"

8.2 Resilience Issues

[C:5] Problem of definition and measurement of resilience

Compounding the problem of what, exactly, is meant by resilience, particularly given the nature of the Internet, there is the problem of measuring performance in general and resilience in particular.

"Assess the resilience of the ecosystem as a whole is much, much harder, not only because participating entities do not necessarily give access to their component-specific resilience information (hence introducing opaque domains that are likely to jeopardize the monitoring of the overall resilience). But also because there is a lack of reliable techniques that can measure the ecosystem's resilience globally."

"The Internet is 'best-efforts', so when my site goes down for repair, does that mean the Internet is down?"

"In general, the design philosophies of end-to-end (i.e. simple backbone and smart edge nodes) and loose coupling have made Internet generally more resilient than other telecommunication networks."

"A universally accepted measure of diversity might be useful for customers to use to assess how interconnected a provider is."

"…there is no formal discipline yet for analyzing for a given situation whether the additional peering is adding value, or even how to define 'fitness' or 'resilience' of a network…"

"…it is very hard to say what it actually means when one buys 'The Internet' - connectivity to what? round trip time to what?"

[C:6] Diversity, complexity and cost

There was no doubt amongst the respondents that diversity is a Good Thing. There was also general agreement that diversity at the interconnection level can be undermined by connections sharing lower level infrastructure – often unknowingly – and that this is a pervasive problem, see **[C:7]** below.

However, some doubts were raised about the extra complexity and cost that comes with increased diversity of interconnection, which are a disincentive to adding more interconnections and reduce the benefit of more connections:

"Yes, [more and more diverse interconnection] *would improve resilience as long as it doesn't create a tightly coupled system or a very complex system."*

"However, an excessive degree of interconnection introduces unnecessary complexity into the network which would ultimately lower the resilience of the eco-system."

"True diversity increases costs and requires motivation."

"But too much diversity can also lead to new types of failures (e.g. delayed convergence / BGP path exploration, state explosion, etc.)"

"More and diverse interconnects translates in higher recurring costs for operators"

"…each new connection adds cost as well as benefit to the ecosystem, there is no formal discipline yet for analyzing for a given situation whether the additional peering is adding value…"

"Increasing diversity of interconnections is good up to a point … at which this becomes untenable from a commercial perspective…"

"Yes, but these cost more of course; this is a balance of cost versus assurance which each operator must assess."

[Q:4] Europe's particularly rich infrastructure

"Europe is probably the most interconnected region in the world today with many transit and private peering interconnects, as well as a functioning and well built out co-location market as well as the highest number of Internet Exchanges in the world with over 120."

[Q:5] Efficiency and resilience are not fully compatible

"Loose coupling so that an infected or damaged part of the system can be severed and worked around. In case of big system problems the system should break down to lowest possible scale with parts that operate independently. Pretty standard – but not efficient."

[Q:6] Much traffic is local

"I disagree that most traffic on the internet does not pass directly between the source and destination networks. I'd also stress that most internet traffic stays local (or at least within language boundaries)."

[Q:7] Not clear that enough spare capacity exists

"...it is unclear that in the presence of an outage of a route that the surviving routes will have enough capacity to host the extra traffic."

8.3 Physical Layer

[C:7] Problem of common, and in places, limited infrastructure

This came up in a number of contexts. There are some heavy concentrations of sites and fibre connections, particularly near IXPs; some parts of the world depend on a small number of undersea fibres; and even where firms think they've bought diverse connections, they may suddenly find that these all pass through the same duct or even over the same fibre, as their suppliers adapt and reconfigure their networks. Examples:

"Common mode failures are more common and affect such things as undersea cables and land based fibre routes where fibre paths are unwittingly routed through common ducts or cable chambers."

"...a difficult problem since many shared risks are only apparent after a failure." [In the context of attempting to assess the resilience of the system by simulation, the problem being how to identify shared infrastructure before a failure.]

"Lack of diversity of physical plant in some key locations..."

"...many ISPs have routers and links located in the same place (e.g., co-lo hotels, fiber running through the same tunnels) that lead to correlated failures, where the dependencies may be hard to realize in advance."

"Diversity can be a canard." [Because diversity of interconnection may be undermined by a lack of diversity of lower level infrastructure.]

"There remains the question over the reality of physical diversity of international infrastructure, at least in certain parts of Europe."

"The vulnerability of underlying physical network, such as sea cables which are easy to cut apart and the need for redundancy in that regard..."

[C:8] Dependence on electrical power and other infrastructure

The dependence of the system on electrical power, on collocation sites with their cooling systems, and fibre and transmission systems was a common theme. The dependence on the telephone system in an emergency was also noted. The possibility of a mutual dependence between electricity networks, telephone networks and the Internet was noted as a particular danger.

Examples:

"...the physical infrastructure and all its dependencies such as cooling, power, a secure diesel supply for back up generators etc."

"…ensuring working power and cooling, even in times of stress on national infrastructure is of utmost importance…"

"…the risk of large-scale cascade failures due to interdependence with the power grid…"

[C:9] Monoculture

The lack of diversity of equipment, software, protocols, etc. was noted by many respondents as a weakness, particularly in the context of cascade or common-mode failure, but also generally as a weakness.

"…true technology diversity…" [...as a key contributor to resilience.]

"…the number of vendors and diversity of equipment and software among operators as well as Internet Exchange Points is relatively low. This has been demonstrated a few times where severe bugs have had to be globally fixed…"

"… technology monoculture…" [...detracting from resilience.]

"Routing code monoculture (systemic bug)" [...detracting from resilience.]

"The most apparent highly-concentrated commonality exists in network routing equipment…"

"The practice of employing separate makes of network router and transmission systems reduces the impact of common mode failures, …"

"…two/three vendors cover almost all the market of ISP-level routers."

[C:10] Vulnerability and testing of equipment and protocols

A number of respondents, noting the dependence of the system on a relatively small number of protocols, implementations and equipment, suggested that more (third party) testing to some (higher) standards would be beneficial – particularly testing for vulnerabilities.

"Robustness testing of all routing and switching protocols to a common standard…" [...would contribute to assessing resilience, and reduce vulnerability.]

"A greater understanding of the security vulnerabilities of network equipment is required…"

[Q:8] An approach to avoiding accidental damage to cables

"The Swedish telecommunication regulator have devised a plan whereby planned digging work can be reported and compared to existing cable plants, without disclosing the actual location of the cables."

[Q:9] Diversity and outsourcing

"Outsourcing can trump diversity."

8.4 Network Layer

[C:11] Invalid announcements and securing BGP

Unsurprisingly, the ability of BGP to rapidly propagate invalid route announcements was mentioned often. Invalid announcements can be the result of human error, or they can be deliberately created to disrupt the system or for dishonest gain.

"The Internet routing architecture lacks the most primitive SECURITY mechanisms."

"The highest risk is a corruption in the global routing table is propagated. This has occurred in practice, and is easy to do."

Existing means to mitigate this problem, to the extent currently possible, by filtering announcements etc. were recognised as Best Common Practice... but that was seen as not as commonly followed as it could be – see **[C:21]** below.

"Filtering of upstream providers announcement might improve protection against erroneous advertisements..."

"Best common practices for filtering BGP updates – either to prevent bogus update messages or prevent excessive message loads – directly improve the reliability of the ISP to its customers."

"ISPs should implement filtering guidelines."

"Best-common practices for filtering BGP update messages can help..."

But the feasibility of doing extensive route filtering, particularly for the largest networks, is questioned, see **[Q:12]** below. The long term prospect of more secure forms of BGP as solutions for this were mentioned, but the lack of urgency to deploy any more secure form of BGP is also noted – see **[C:21]** below.

"...it might be good to identify economic incentives to deploy security mechanisms more widely."

[C:12] BGP's behaviour under stress

Some respondents noted that BGP under stress may, temporarily, be part of the problem during recovery from some event.

"BGP converges relatively slowly, which also disrupts end-to-end communication."

"Certainly we already see cross-layer cascade failures, such as (i) slow BGP convergence causing application-layer timeouts that lead to many application-level retries and (ii) layer-two failures leading to excessive rerouting at layer three that causes missed routing-protocol timeouts at layer three. BGP could have cascading failures due to (say) route leaks that exhaust the router memory, leading to session resets and path exploration that affect neighbouring domains." [in the context of cascade failure.]

"In the last decade, the use of BGP path flap dampening has also proven to make faults worse..."

[Q:10] IPv6 and future shock

"Mixed IPv4/IPv6 operations might also generate quite some chaos in the coming years."

[Q:11] Connectivity is not everything... but that is all that BGP understands

"Lack of any performance culture is a major barrier. The Internet has always been about providing connectivity not performance. As a technology it is inherently resilient. The performance that can be achieved is something of a lottery."

[Q:12] Connectivity is one thing, capacity is another

"Fail over of routes may lead to connectivity but with much reduced capacities and higher latencies."

[Q:13] Limitations of the tools to hand

"Most interconnections between large companies have no restrictions on route announcements at all. Both sides rely on the other to filter their customers, which they aggregate and announce to their peers. This isn't the best way to interconnect with each other, but the alternative is to have to update filters every time anyone adds or loses a customer, or even adds or withdraws a route. This is clearly not manageable or even possible on today's Internet architectures."

[Q:14] RPKI a Good Thing

"Deployment of Resource PKI (RPKI) should be encouraged to provide for end-to-end authentication of routing announcements."

[Q:15] More secure routing may not be more resilient

"Filtering should be used with care: It should protect against catastrophes due to misconfiguration or malice, but still leave leeway so that 'unusual' routing announcements can be accepted in crisis situations. This is easier where BGP is configured manually (but that also makes errors more likely). Where this is done automatically, it would be nice to have escape mechanisms that would allow setting up extraordinary filters."

[Q:16] BGP is fundamentally limited

"BGP was not designed for resilience, and it is difficult, if not impossible, to configure BGP to deliver the resilience, dependability, survivability, and performability that we expect from the future Internet. If a clean-slate redesign of BGP is not feasible, then significant changes to the current BGP protocol and architecture must be made."

8.5 Operational Layer

[C:13] Technicians and the Internet Ethos

Looking after the interconnection system are the technicians, who work day and night to keep it running. It is felt that technicians have a stronger sense of 'the good of the Internet' than the management of the organisations they work for. Thus:

"… what makes the Internet run (BGP, DNS, etc) and what keeps it running (coffee and people)"

"…readiness within the ISP community to help each other out in times of emergency, e.g. by improvising short-term backup transit arrangements…"

"rapid/unencumbered communications between parties on a technical level." [Best practice for dealing with a crisis.]

"… cooperation may be hampered by management staff (if they find out about it) …" [In the context of technicians in different networks cooperating in a crisis.]

[C:14] Pre-arranged mutual aid in a crisis

To improve the resilience of the system it is suggested that networks that peer with each other, and generally only exchange their own traffic over the peering connection, could enter into 'mutual aid agreements', so that in a crisis they could offer transit to each other. By sharing what remained of each others' interconnections, both networks might maintain better connectivity.

"BGP routing policies restrict the set of paths that can be used, meaning that two hosts may be unable to communicate even though the Internet topology remains connected."

"…the most important component is to enable the interconnection of networks. I.e even if they are not interconnected at all times, that they can be if need arise is important. This has been showed on several occasions…"

"Forming an agreement with a peering partner(s) for ad-hoc 'back up' transit should catastrophic failure occurs in each other's network or failure on their upstream….a bit like subsea cable restorations."

One respondent suggested:

"Governments or industrial consortia could subsidize…"

[C:15] Crisis management is generally ad-hoc

When there are issues the communication between networks tends to be informal and ad-hoc. This is related to the cult of the technician-hero, see **[C:13]**.

"There are some ways that companies can communicate with each other in a crisis that are unofficial, but those channels are somewhat dependant on the personnel at the companies."

"Existing communication channels and the social fabric woven by ISP communities provide an excellent basis for handling critical situations in an informal and effective manner."

"…this is currently very ad-hoc and could stand some forums to develop practices…"

"Personal contact and knowledge between ISP technicians/supervisory personnel…" [In the context of what practice or systems exist for crisis management.]

"Not sure any more formal crisis management could be built given the diversity and span of the network, or it might contribute to slowing down the fix vs fixing it quicker."

"In case where multiple operators, IXPs and others must cooperate, there is no formal procedure to the best of my knowledge."

"Currently, operators tend to use ad-hoc solutions such as mailing lists (NANOG, internet-incidents etc.) to report issues with connectivity, performance and security."

[C:16] Value of Exercises ("War Games")

A number of respondents recommended the use of exercises or "war games", in which various event/disaster scenarios are worked through. During such exercises much may be learned about actual readiness for such events, and about otherwise hidden dependencies.

"[disaster exercise that involved mobile network operators, ISP, power companies, etc.] *has been a very valuable exercise, not only to test the alternative paths, but also to test communications between the participants and their ability to communicate and co-operate. A similar exercise in other countries as well as perhaps on a European scale might be a good way to assess and verify co-operation and robustness of the ecosystem."*

"CIIP preparedness and national and large scale exercises on network incidents will bring forward weaknesses."

"…national and large scale exercises and preparedness…" [as recommended best practice.]

"…periodic large scale tests need to be done much like joint military exercises and emergency drills."

"…place an X over a site and see what happens. There are a number of locations if you do this at, would cripple internet capacity…" [as a means to assess resilience..]

[Q:17] Dependence on phone system for communication between operators

"Mailing lists and phone calls continue to be main mechanism for crisis management. Other mechanisms (e.g. noc dba phones) never really gained traction."

[Q:18] Limits to preparedness

"Most companies barely adequately plan for a single failure, let along cascading failures."

[Q:19] Prioritisation of traffic in an emergency

"ISPs could develop disaster recovery plans to deal with the worse failures, e.g. by prioritising some traffic over other."

"Prioritisation of traffic categories could enable more critical traffic to flow but such decisions tend to be commercial and public sectors which may conflict."

8.6 Contract and Economic Layers

[C:17] Keeping customers is the key motivation – reputation

It was generally agreed that the main incentive on each network was its need to provide its customers with acceptable service, and maintain its reputation as a reliable provider. So:

"For their own networks, the impacts on brand, reputation and commercial disadvantage must be foremost in the minds of network operators as regards their own networks."

This incentive was certainly held to contribute to the resilience of individual networks. Whether it contributes to the resilience of the overall ecosystem was seldom addressed, but where it was there was doubt (see also **[C:21]** below), for example:

"Whether that translates into a more resilient ecosystem is up for debate…"

[C:18] Customers' ability or wish to influence resilience

As noted in **[C:17]** above, the behaviour of individual networks is strongly linked to what customers demand. However, a number of respondents noted that customers tend not to treat resilience as a priority, or tend to choose providers mostly on the basis of price.

"…resilience doesn't seem to be a big selling point for many internet customers…"

"…the market lacks ways to competitively evaluate/measure SLAs."

"… always using the low-cost bidder…"

[C:19] SLAs stop at the edge of the provider's network

In the context of whether SLAs do or might contribute to the ecosystem as a whole, a number of respondents pointed out that SLAs never (or at least seldom) cover anything beyond the edges of the network giving the SLA:

"…there is no guarantee that crosses the own border…"

"Suppliers in my experience never give SLA guarantees related to interconnectiveness. Without very strongly binding peering agreements I find it difficult to believe that any provider will give a general ecosystem guarantee."

"SLA contributes to resilience of the supplier; but any supplier drops responsibility of what happens in the ecosystem past its boundaries. Which makes sense given the fact an operator can not guarantee something which it does not have control over."

"It is unusual for end users to be offered a serious service level agreement. This is not surprising, since so many elements of service are outside of a service provider's control…"

The fact that SLAs stop at the edge of the provider's network appears natural and inevitable.

[C:20] Doubtful value of SLAs

The value of SLAs in general was doubted by a number of respondents. For special services, notably 'on-net' services such as VPNs, some SLAs were thought to have some value. But generally:

"[an SLA] *is usually more favourable towards the supplier's side…"*

"Terms in SLAs in general don't really encourage even the resilience of the supplier."

"SLAs are meaningless"

"Today's SLAs are quite coarse-grain… There are legitimate reasons why the SLAs are not more precise – specifically the difficulty of any one ISP to control end-to-end performance through many ISPs, and the difficulty of efficient and accurate measurement to verify the SLAs."

"currently nothing useful in the way of end user SLAs in most cases…"

"I do not think SLAs play a critical role / incentive structure today. Mainly because the market lacks ways to competitively evaluate / measure SLAs."

"[SLAs] *only refer to the portion of the network that is under the control of the ISP. Hence, their contribution to the resiliency of the Ecosystem is questionable."*

[C:21] "Tragedy of the Commons"

There were a number of observations that networks operating in their immediate self interest do not always operate in the best interests of the overall ecosystem – some using the notion of the tragedy of the commons. This includes:

- bemoaning the patchy or incomplete adherence to best common practice, notably in connection with BGP filtering.
- suggesting that to get networks to adopt more secure forms of BGP may require specific, external incentives.
- observing the failure to deal with issues that affect all Internet users, for which a collective response is required.

- calling for more and better testing of equipment and protocols, particularly in the context of dealing with unusual (or invalid) data or unusual load.
- Tier 1 providers and their occasional 'de-peering' disputes.

Examples:

"*Inconsistent application of best practice…*" [A factor detracting from resilience.]

"*lackadaisical attitude about longstanding problems like virus, spam, phishing*" [Also a factor detracting from resilience.]

"*Tragedy of the commons situation … individual motivations dominate decision making.*"

"*Very few incentives … to ensure that the ecosystem as a whole is resilient.*"

"*There are no incentives for IXPs to consider resilience of the ecosystem as a whole. The same consideration applies to network operators.*"

"*Not many incentives exist when they have to compete on price – making them get rid of redundancy and loose coupling (e.g., several suppliers) in order to reduce costs by increasing efficiency.*"

"*For the broader Internet, market forces alone do not provide incentives for resilience, security (i.e. botnet / SPAM), or other 'tragedy of the commons' sorts of problems.*"

"*…many operators will follow the herd and maximise their own network resilience while providing a little more than the bare minimum of resilience to the remainder of the ecosystem.*"

"*Ultimately, the incentives are for individual networks to provide the best possible availability to their own customers.*"

"*Governments could provide financial incentives to upgrade to a more secure variant of BGP…*"

"*Most companies barely adequately plan for a single failure, let along cascading failures.*"

"*We badly need economic incentives for providers to implement BCPs and basic policy around BGP.*"

"*If all 'major' ISPs would follow BCPs the number of incidents would be even lower than today.*"

8.7 Regulatory Layer

[C:22] Regulation? No thank you.

Regulation is mentioned once (in Question 10), but is referred to in answers to nine other questions. The consensus is that regulation is at best unnecessary:

- things work just fine without regulation, so none is required;
- consumers at all levels of the Internet have plenty of choice, keeping suppliers at all levels honest;
- where it exists, regulation is a barrier to interconnect; conversely, the unencumbered ability to interconnect in any way is key to the willingness to interconnect;

- regulation would be counter-productive. For example: if interconnection were required to be more formal, there would be less of it – which would be worse than any perceived weakness in informal interconnection;
- regulators do not understand the Internet, and any attempt to apply 'old world' telephony style regulation would be destructive;
- the Internet develops at a pace that out runs the ability of any regulation to be relevant or useful – even if it was effective or not actually destructive;
- the market will respond to change far more quickly than any regulator could, and continue to operate efficiently – where last year's regulations might prevent that;
- regulation would not work. For example: if interconnection were mandated, experience shows that it can be made so difficult or so ineffective as to be useless.

Where respondents acknowledged that regulation could be considered, it was generally as a last resort or subject to proof, or very strong proof, that it was essential or unavoidable.

Examples:

"...other regions of the world, where regulation is a real barrier to interconnects and an evolution of the interconnect ecosystem..."

"...regulatory intervention therefore carries a high burden of proving that externalities leading to market failure outweigh all these factors..."

"NO regulation required the eco system works fine."

"Regulation should be regarded as a last resort..."

"Regulatory intervention cannot be justified unless both scale and impact are high..." [In response to Question 3, "What can be done to assess and verify the resilience...".]

"Regulations are unlikely to succeed..."

"...far from clear that regulators would have enough understanding of Internet operations to add anything but confusion to the mix (particularly if they try to impose current regulations designed for circuit-based telephony on the packet based Internet)."

"How can regulation keep up in such a fast-moving space?" [In the context of the shift of traffic to Content and Content Delivery Networks, over the last two or three years.]

"you seem to be approaching the issue with the assumption that additional regulations will make the Internet more reliable – I'm far from convinced that is the case"

[C:23] Need for statutory disclosure regime

Given the need for information about the interconnection system, some respondents thought that perhaps this is an area where regulation is required, though it would be resisted!

"...would have to be persuaded to ... share freely high-level information about the shape and size of its network."

"...we need statutory data requirements just like we have for other critical infrastructure."

"...requirement by governments to provide topology and shared-risk information would help, though there would be significant resistance..."

[Q:20] Incentive needed for deployment of more secure BGP

"Governments could provide financial incentives to upgrade to a more secure variant of BGP, to get the ball rolling by 'buying' a critical mass of ISPs who have deployed."

[Q:21] Role for trusted third party in a crisis

"A large failure would cause traffic spikes that could be difficult to handle by current BGP. If such a failure happens, I guess that human operators will have to fine tune their BGP configurations to reroute traffic. In this case, visibility about the routes and the congestion in remote networks could be useful to avoid pushing traffic to overloaded links. Operators are reluctant to share this information, but in case of emergency it might be possible to ask them to give information to the regulator (in a format to be defined before hand) and allow the regulator to share this information during the crisis."

9 The Questionnaire and Summary of Responses

This section gives the questions from the questionnaire, along with the subheadings used, and a summary of the response to each question.

The respondents were told that the response to the questionnaire would be summarised and published with the report, but that responses would be neither attributed nor attributable.

The Ecosystem, Risks and Resilience

1. What are the key components of the Internet interconnection ecosystem?

The Introduction to the Study (annexed below) gave an outline for the ecosystem which is the subject of the study. This question was intended to be read in that context, with an emphasis on the key components. It was also intended to allow respondents to add anything which had been missed in the Introduction.

This question was deliberately open-ended... but this did leave a number of respondents overwhelmed. A number of respondents noted the discussion in the Introduction and agreed with it.

The respondents generally noted the key roles of:

- Internet Exchange Points;
- BGP – the protocol, its implementation and use;
- underlying fibre and other transmission infrastructure – including undersea systems – notably provided by third parties;
- equipment and equipment vendors;
- interconnection policies;
- data/collocation centres, power and so on – also notably provided by third parties;
- the major Transit Providers – echoing one of the points in the Introduction;
- the Content Delivery Networks, and their growing significance;

Less generally, the following were noted:

- the Domain Name System (strictly speaking, DNS is not really part of the interconnection ecosystem. It is, however, a key service which is accessed across the interconnection system.);
- the influence of governance – IETF, ICANN, the Regional Internet Registries, ...
- regulation, or the absence of it, and the independence of ISPs and IXPs;
- a shared ethos amongst network operators;
- the people who make the networks and their interconnections and keep them running;
- the importance of 'true technology diversity';
- the system's complexity and the many dependencies between the layers, making it hard to grasp let alone to oversee;

- the particularly dense connectivity in Europe which has 120-odd IXPs.

2. *What are the key factors that contribute to, or barriers that detract from, the resilience of key components of the ecosystem and of the ecosystem as a whole?*

Having identified the key components, the respondents were invited to discuss resilience. Their responses here very varied, but the following were mentioned by a number of them.

- Connections are concentrated in relatively small areas, sharing collocation sites, ducting, fibre or other infrastructure – quite possibly unknowingly. See **[C:7]** and **[C:8]** above. IXPs were noted as a particular example, although it was also noted that some maintain high standards of resilience. In some instances the sharing of facilities is inevitable because of the presence of a monopoly supplier, so neutral collocation sites contribute to resilience.
- BGP will propagate invalid route advertisements, whether those are created by accident or deliberately to damage the system or for dishonest gain; this insecurity detracts from resilience. See **[C:11]** above. The 'inconsistent' application of Best Common Practice to reduce these effects was mentioned in this context, as was failure to deploy more secure versions of BGP. See **[C:21]** above.
- BGP is also fragile in that invalid or unusual data can disrupt implementations, it is vulnerable to operator error, and the number of equipment suppliers is relatively small. See **[C:11]**, **[C:10]** and **[C:9]** above.
- Yet the operation of BGP and the ability to reroute is a key attribute of the Internet interconnection ecosystem, and contributes to its resilience – however insecure or vulnerable it is.
- The diversity and multiplicity of interconnections contribute to resilience in many ways. The existence of multiple providers at every level, and the flexibility of association between networks (in the absence of regulation) contributes to more interconnection; so do the trend for content and content Delivery networks to peer openly in multiple locations and multiple, diverse connections between the Tier 1 networks. However, diversity and multiplicity add cost and complexity, which limits the extent to which networks will implement them.

The following were mentioned by two or three respondents:

- multi-homing clearly improves the resilience of the system;
- peering disputes at Tier 1 can lead to 'de-peering' incidents, in which those who connect only to one of the networks in dispute will be cut off from those who connect only to the other. See **[C:21]** above.
- commercial issues may limit or prevent interconnect. See **[C:4]** above.
- high standards in the large transit providers contribute to resilience.

Other individual points included:

- the small number of Tier 1 providers in some parts of the world;
- the lack of effort made by ISPs to improve the security of the system;
- TCP's response to congestion, which contributes to resilience;

- the 'best efforts' character of the Internet, which assumes that temporary loss or degradation of service should be acceptable;
- the lack of testing of protocols and implementations;
- readiness and ability to improvise in an emergency;
- the impossibility of predicting the outcome of failures or measuring actual resilience;
- the difficulty of telling whether there is spare capacity in the system for traffic to fail-over between ISPs – noting the multiple fibre cuts in the Mediterranean in early 2008;
- restrictive routing policies that can limit the utility of some interconnections (e.g. peering connections), particularly in an emergency;
- slow convergence of the BGP mesh;
- back-up paths that are not tested until they are needed;
- whether the importance of resilience to each network translates to ecosystem resilience;
- lack of data about almost all parts of the system – some of which is treated as trade secret – which limits the ability to assess how well it works or evaluate ways to improve it;
- the replication of data and servers across the world in the CDNs;
- the continuous activity across the Internet community, discussing issues and seeking solutions;
- that the CDNs by drawing traffic away from the general Internet, and towards more closed networks and systems, may be reducing resilience: by separating parts of the Internet from each other, by concentrating a lot of traffic in a few hands and by undermining the transit providers.

3. *What can be done to assess and verify the resilience of key components of the ecosystem and of the ecosystem as a whole?*

This is clearly a very large question, and addresses a central concern of the study. Perhaps not surprisingly the respondents were not able to solve the problem. But there was some consensus that the problem is hard and that the information required to tackle it is not currently available. The following points were made.

- Some respondents pointed to the problem of defining and measuring resilience – always assuming there was an accepted definition for the ecosystem which is to be assessed.
- Assuming that is possible, is it possible to actually monitor the resilience of the system? See **[C:5]** above. This is a hard problem as it requires a model of the system, at economic, commercial and technical levels. At the technical level, we need maps of interconnections, capacity, traffic, routes etc., and the mapping of interconnections needs to take into account the problem of shared infrastructure, including collocation sites and their dependence on power etc. See **[C:7]** and **[C:8]** above. There was consensus that such a model would be extremely difficult to construct, as networks do not publish the required data, and we cannot collect it from the outside. Some respondents suggested that to obtain the data would require some regulatory intervention, which would be resisted. It was noted, however, that other Critical Infrastructure is subject to this kind of disclosure requirement. See **[C:1]** and **[C:23]** above.

- A number of respondents suggested studying incidents to learn not only how to avoid them, or mitigate their effect, but also to help build a model of the system using hard data. See **[C:4]** above.
- The value of "War Games" was noted, as a means to discover likely problems (for example, shared infrastructure) and testing preparedness. See **[C:16]** above.

One respondent pointed out that the system does appear to be reliable:

"…real world - notice that the Internet is VERY reliable, it is very rare that there are significant outages -- even when there are attacks or failures ISPs recover quickly…".

Other individual points included:

- it is unlikely that the entire system could collapse;
- scenario testing is a good thing, particularly if based on plausible/relevant emergency situations;
- surveys of physical diversity are too rare;
- it is very difficult to assess where traffic will go in the event of failures;
- a recommendation for a system to monitor the interconnection system, so that its status and performance can be continuously assessed;
- a recommendation for a mechanism to audit the interconnection system, to encourage improved security and resilience.

4. *What is the risk of cascade failures bringing down a large part of the Internet? What cascade or common mode failures are likely, and what can be done to reduce the risk or the impact, or both?*

Given that it is impossible to imagine an external event that could possibly affect a large part of the Internet, the most likely cause of widespread problems is some internal, systemic problem. There was some agreement as to the most likely risks.

- BGP issues were, unsurprisingly, at the top of the list as a likely risk of cascade failure. Its ability to cope under stress is a particular concern (see **[C:12]** above) and a more secure version of BGP would mitigate some of these risks (see **[C:11]** above). Related risks included the relatively small number of equipment vendors, the problem of cross-layer dependencies and the possibility of many interconnections failing at once, with possible knock on effects, thanks to the dependency on underlying fibre, other transmission facilities, collocation sites, power etc.
- A few respondents cited congestion as a potential cascade failure – in which traffic diverted away from some failure overloads the networks to which it is diverted, overwhelming them so that yet more traffic is diverted. Such knock-on congestion could affect BGP's ability to keep BGP sessions running, which would add route instability to the mix, possibly exacerbating the congestion. Large scale DoS attack was suggested by a few respondents as the seed for such a cascading congestion failure.
- To mitigate the risks from software and protocol issues, some respondents thought there should be more and open testing of equipment and protocols. See **[C:10]** above.

There was some scepticism. A number of respondents felt that there is a low risk of bringing down large part of Internet, even from the inside. The experience of incidents in which some software problem has spread across the Internet is that only some equipment was affected, and/or operators have responded quickly to squelch the problem.

Other individual points included:

- the danger of mutual cascade failure – e.g. due to mutual dependence of power networks and the Internet;
- cascade failure are most likely to be caused by something not predicted;
- common mode failures (e.g. common fibre) are more common/likely than cascade failure, though cascade failure is not unknown;
- if part of one transit provider's network fails, traffic could/would be diverted to other transit providers, who may not have enough capacity, causing congestion and further diversion of traffic;
- the need to be able to monitor the behaviour of the system to detect and diagnose problems;
- now the CDNs are replicating and distributing data locally, the system is less dependent on global transport;
- the serious deployment of IPv6 is likely to cause interesting problems.

Incentives, Agreements and Economics

5. What incentives exist for operators of networks, IXPs etc to ensure both their own resilience and that of the ecosystem as a whole? How might those incentives be strengthened and/or incentives be created?

One of the motivations for the study is that it is not clear whether there are sufficient incentives to support the resilience of the ecosystem as a whole. On two points there was agreement among a number of respondents:

1. Networks are motivated by simple commercial pressures to offer reliable service for their customers, on their own network. See **[C:17]** above. They want to keep customers and to gain new ones, in a market where the customer has a number of alternative suppliers. Reputation and/or Service Level Agreements play a part in this.
2. There are no incentives for networks to consider the resilience of the ecosystem as a whole. See **[C:21]** above. Extra incentives may be required to encourage the deployment of more secure BGP. There was one dissenting voice, who felt that SLAs *"ensure resilience of the ecosystem"*. But see below for the mixed views on SLAs.

Other individual points included:

- increased transparency could reveal which networks do a good job, and which do not;
- one respondent suggested "naming and shaming" poor networks;
- one respondent suggested that independent audit and consumer bodies could have a rôle in this;

- it would be helpful to have better systems to identify where congestion occurs – e.g. IETF ConEx;
- educated customers are the best incentive;
- price competition requires each network to increase its efficiency, which implies more tightly coupled networks, which are less resilient,
- regulator intervention could only be justified if "*the customer was unduly incapable of switching…due to market failure*";
- the large Content and Content Delivery Networks own an increasing share of the infrastructure and traffic;
- resilience at the scale of the ecosystem requires new inter-AS technical and contractual arrangements;
- new services – e.g. cloud computing – are likely to increase the demand for reliability.

6. *Would more and more diverse interconnections improve the resilience of the ecosystem? If so, how might that be encouraged or supported?*

Diversity is generally associated with resilience. This question was looking for any indication that the ecosystem is less resilient than it might be. There was general agreement that more diverse interconnections would be a Good Thing (particularly for singly connected networks), but this doesn't always happen.

- Extra interconnections increase cost and complexity – so for each extra interconnection the network operator must decide whether any extra resilience is worth it.
- It is hard to be sure that each extra interconnection truly is diverse! Respondents cited examples such a number of ASes at one IXP connecting to a distant IXP for extra diversity, but using the same infrastructure; connecting both privately and through an IXP, but doing so at the same site as the IXP; problems where the physical infrastructure has limited resilience, notably some undersea cable systems; and questions on whether it is better to spend money on increasing diversity, or on hardening the existing infrastructure.
- It would be valuable if peering connections allowed for mutual transit in an emergency. See **[C:14]** above.
- Extra connectivity may be driven by cost reduction rather than resilience – notably peering to reduce transit volumes.
- More connections may slow down BGP convergence (by introducing more routes to be considered).
- Interconnection may be limited by commercial considerations. In other types of network the regulator can require interconnection, perhaps at some defined cost – which might increase diversity of connection.
- Interconnection is driven by business and commercial considerations, so there is no need to do anything to encourage connections.

Yes, the last two points are contradictory!

Some respondents wished first to have a means to measure the improvement provided by a given extra interconnection – so that the value of one more interconnection could be properly assessed. Some were unconvinced that this was a problem.

7. *What terms in end-user Service Level Agreements contribute to the resilience of the supplier and of the ecosystem as whole? How might such agreements contribute more?*

If the market is to sustain the resilience of the ecosystem then the contractual commitments to customers must have a role. This question asked about end-user agreements (the next one is about inter-ISP agreements). Generally respondents were less than convinced of the usefulness of SLAs in general, and even less convinced of their effect on the ecosystem as a whole. One or two felt that SLAs do contribute at least to the resilience of the provider's own network. A few suggested that better SLAs would have more to contribute. The common points raised were:

- availability is a key SLA metric.
- SLAs almost never go beyond the provider's network, so however good or bad they are, they do not cover the interconnection ecosystem (see **[C:19]** above). Respondents noted that because everything beyond the edge of the provider's network is outside its control, so it cannot offer guarantees, and so it is hard to see how SLAs could contribute much to the resilience of the ecosystem.
- One respondent noted that binding peering agreements might form the basis for wider SLAs, while another noted that end-to-end SLAs would be required if the Internet were to be better than 'best efforts', but that this remains a research topic.
- it is unusual for end-user SLAs to be useful. Respondents noted different aspects of this:
  - SLA metrics tend to be simple, limited and measured by the provider within their network, almost as if they are designed not to fail. Network availability, for example, may be measured at the provider's router even if no traffic can leave that router, and other metrics may only cover performance between defined points in the provider's network;
  - more comprehensive metrics and the means to measure them are lacking;
  - as a result, SLA metrics tend not to cover what the customer thinks is important, their particular pattern of use;
  - SLAs tend to favour the provider, so that if claims are made, it is cheaper for the provider to pay up than to improve resilience;
  - the market lacks the ability to evaluate and compare competing providers' SLAs, so they neither provide a means to differentiate service nor a means to exercise informed choice;
  - for domestic users SLAs are limited and are particularly hard for individual users to enforce in any meaningful way, from the perspective of resilience.
- Most succinctly, one respondent said:

  "SLAs are meaningless."
- Conversely it was noted that:
  - more precise/effective/enforceable SLAs could improve resilience, but at a cost;

- to be effective SLAs should cover from the customer to the customer's chosen destinations and back (end-to-end);
- if customers demanded better SLAs that might drive improvement;
- new services – e.g. cloud computing – may call for new and better SLAs;
- two respondents thought that SLAs contribute to the resilience of the provider;
- if providers were required to publish performance and failure information (at least to their own customers) that might be an incentive to improve resilience (at least in their own networks).

Other individual points raised included:

- if SLAs described what is not covered – notably anything beyond the edges of the provider's network – customers could better assess their risks, and perhaps consider multi-homing;
- if SLAs were required to guarantee greater resilience that would make service more expensive, and discourage multi-homing – which would be bad, because multi-homing is more beneficial than an incremental improvement in the resilience of one network;
- mandating SLA terms might force some customers to pay for levels of service they do not need, requiring them to subsidise higher levels of service for others;
- independent audit and consumer bodies could help customers demand better SLAs.

8. *What parts of the formal or informal agreements that govern inter-ISP connections contribute to the resilience of the ISPs and of the ecosystem as a whole? How might such agreements contribute more (e.g. more formality and/or transparency)?*

This question was perhaps too broad. There are different types of connections between ISPs. A transit agreement is likely to be more formal, because money is involved, while a peering agreement between two small ISPs is different from that between two Tier-1 networks; and so on. Yet on two points there was some agreement.

1. If the connectivity and performance of networks were more transparent, it would help networks decide to whom it would be best to connect, and help them manage their networks once connected. There was no doubt that peering arrangements are opaque and treated by some as secret. The problems caused by Tier 1 de-peering incidents were mentioned in this context.
2. Peering in multiple, diverse locations is a good thing – and for the larger networks is definitely best practice. Some respondents note that this might be improved if agreements:
   - included requirements for diverse connections;
   - covered the provision and maintenance of adequate capacity;
   - included sharing of routing and performance (QoS) information, to improve each network's ability to manage its network;
   - covered the provision of transit in the event of an emergency.

A couple of respondents did feel that greater transparency, formality and rigour would be beneficial. However, there were also definite reservations about such an approach:

- most peering arrangements are informal;
- the existence of an interconnection is (much) more important than anything else. So it would be a mistake to create barriers to interconnection by attempting to require formal agreements or particular standards, which would increase the cost of interconnection. In practice, IXPs do not attempt to mandate whether or how users interconnect;
- where arrangements are formal, they reflect the business and commercial needs of the parties, so are necessarily private arrangements. Any requirement for transparency would be fiercely resisted (and probably not be very useful);
- attempting to standardise in this area could simply reduce things to the lowest common denominator – assuming that it is clear who should set such a standard;
- if some peering arrangements are more formal than others, and have particular quality requirements, that may clash with 'net neutrality'.

Topics of broader scepticism included:

- whether the inter-ISP arrangements have any bearing on the resilience of the ecosystem, or whether it is even a consideration;
- whether there is any evidence to suggest there is any problem in this area;
- whether intervention in this area would be beneficial, as imposing a requirement to peer would, effectively, force larger ISPs to subsidise smaller ones.

Some operational issues were considered important:

- cooperation when dealing with DDoS, or invalid route announcements, etc.;
- filtering to remove invalid announcements;
- willingness to act on reports from other operators, who are not paying customers;
- good change management between the parties.

9. Are costs falling in line with the continuing fall in end-user and wholesale prices and the continuing increase in demand? If not, how can the market be sustained in the long term?

The main thrust of this question was whether the costs of providing transit are falling in line with prices paid by end-users (most of whom are consuming ever more bandwidth) and prices paid by wholesale transit customers. The large transit providers are key components of the ecosystem, so the health of this market is of immediate concern. The desire to avoid 'leading the witness' may have obscured this, and some respondents did not know what to make of the question, while others answered it from their perspective. But on the intended subject of the question, the health of the transit market, the following points were made:

- wholesale prices continue to fall. One respondent detects some signs of this bottoming out, while another noted that the trend has been consistent for a decade, so there is unlikely to be a short term sustainability problem;

- the transit market may be sustained by cross-subsidy from other parts of each provider's business. For example: transit in highly competitive markets in Europe and North America may be subsidised by business in less competitive markets; or, ordinary transit may be subsidised by other added value services;
- there was some doubt, particularly among the small number of transit providers in the consultation, about the long term health of the market;
- consolidation in the wholesale market is thought likely and some providers are expected simply to move away from transit provision;
- the possibility of market failure was considered by one or two, who touched on the possibility of the need for regulation;
- conversely, one or two respondents had faith that the market will provide.
- the costs of the technologies that underlie the cost of providing transit were reported as continuing to fall at 20-30% pa. But outside Europe and North America, capacity is still relatively expensive.
- one of the service provider respondents referred to some ISPs still living on cheap assets from the dot-com boom, which will sooner or later be exhausted.

From the point of view of the costs of providing end-user services, compared to their price:

- costs are falling more slowly than revenues – end-users demand more for the same or lower price;
- running out of cheap (dot-com boom) assets will be a shock for some ISPs;
- there is a mismatch between (low) flat rate end-user pricing and providers' actual (peak traffic related) costs.

Finally, the lack of good data in this area was noted as limiting analysis!

Good Practice, Policies and Management

10. What best practice, policies or regulations apply to Internet interconnection? What additional best practice, policies or regulations could enhance the resilience of interconnections and of the ecosystem as a whole?

This question was a fishing expedition. There was some agreement on:

- the advisability of following IETF official Best Common Practice and other Best Practice generally known to the industry (via NANOG, other operator groups, the RIRs, IXPs, Euro-IX, Team Cymru, national CERTs, etc. etc.);
- that best practice for interconnections includes multiple and diverse connections, with adequate capacity, a competent NOC which responds to peers as well as customers, and cooperation between networks to manage DDoS;
- generally Internet interconnection is not regulated.

Beyond that there was some agreement that regulation would be unwelcome for a number of reasons:

- no regulation is required;
- it has been OK so far... any change would need to be fully justified;
- agreements and the IXP model are fairly mature;
- new policy must not hinder competition or create barriers to entry for new operators;
- regulation would be unlikely to succeed given political and cultural diversity;
- regulators do not understand enough to do anything but harm – especially if they try to apply old telecom models;
- regulation could not keep up;
- interconnection is business case driven and improves costs – so we must not create barriers to it;
- regulation to require peering (from experience) may be defeated by technical means to make peering connections ineffective.

For good measure, the following arguments were also made:

- there is no ecosystem as a whole... just different companies each with their own agenda;
- best practice implies similar actors and it is not clear that this is true.
- it is not clear whether additional policies would help.

However, there were also the following suggestions:

- statutory ISP licensing;
- the need for protocol robustness standards;
- interconnected ISPs should arrange for mutual back-up connectivity in a crisis;
- government could offer incentives for secure BGP – to buy a critical mass;
- there should be mandatory reporting of incidents to the relevant CERT.

11. What best practice, procedures or systems exist for crisis management – particularly where multiple network operators, IXPs and others must cooperate?

The objective of this question was to gather some information about current practice, and the common responses were as follows:

- Current procedures and systems depend largely on personal contacts between ISP/IXP technicians – so are largely ad hoc, but they are not accidental: these contacts are cultivated. The technicians' role is absolutely crucial, as rapid and unencumbered technical communication is essential; more formal channels usually do not involve the right people; more formality would probably slow things down; and while, in a crisis, operations staff will work with colleagues in other organisations 'for the good of the Internet', that may be hampered by management.

- One respondent argued that more transparency is required in crisis management, and the current informal systems cannot last.
- Working with the local CIIP (Critical Information Infrastructure Protection) body was mentioned by several respondents, but fewer than noted the 'traditional' ad hoc inter-provider channels.
- Exercises ('war games') for crisis management were recommended as good opportunities to learn and to establish best practice. So far these have been on a national basis, but perhaps should be attempted on a Europe-wide basis.

Other, individual, points made included:

- IXPs often play strong coordinating role, and that should be encouraged;
- things were better before competition;
- higher tiers are better organised than lower ones;
- an official forum is required, because otherwise anti-trust considerations and the like make it hard for operators to cooperate;
- mailing lists and phones are still the preferred mechanisms for communication between operators... other mechanisms, such as INOC-DBA phones have not caught on.

12. Do you share information about threats and vulnerabilities with other network operators, IXPs or others? If so, what information do you share and how do you do that?

The objective here was to gather some information about current practice, and common responses included:

- share and gather information at established fora, informally – that might be NANOG, RIPE, local IXP, Euro-IX, etc.;
- share and gather information using National CERT and/or CIIP;
- expect to receive notifications from vendors;
- participate in one or more of the closed communities – nsp-sec, ops-trust, etc. (though one respondent noted even in these closed communities, not much is shared about vulnerabilities prior to publication);
- a couple of respondents share only with partners and customers;
- a few respondents did not think information is shared between operators. (The responses from operators were mixed.)

Other, individual, points raised included:

- the desirability of large scale tests and emergency drills;
- physical diversity tends to be treated as secret information, which makes it hard to ensure separacy;
- coordination among CERTs is useful/important;
- the ENISA NEISAS project for sharing information in secure manner should contribute here;

- in the usual, informal forums the information can be superficial – useful detail is typically treated as confidential.

13. Do you report security and integrity incidents that affect service? If so, how and to whom? How effective is this mechanism?

This question was intended to gather some information about current behaviour. The response was a little mixed, some appeared to answer 'yes we do report' while others appeared to answer 'yes we should report'. The impression is that in Europe the answer is generally 'do' and elsewhere 'should'. Very few respondents addressed the 'how effective' part of the question.

Among the respondents that 'do':

- reports are made to the national CERT and/or CIIP;
- reports are made to the usual fora – essentially information sharing, as per Question 12, above;
- reports are made, variously, to: affected customers (only), members or funding agencies, affected peers (only), or to some closed group.

Among the respondents who felt they 'should':

- a standard definition of 'security and integrity' incident and a standard reporting format should be established;
- respondents said that reporting would need some central coordination – a clearing house, or perhaps some official body to collect and act on reports, or perhaps the Internet should be regulated in the same way as other CII, with more organised and centralised reporting.

On 'effectiveness':

- a few respondents suggested that publication might 'encourage' improvement – whether publication to customers or broader publication, while another rejected 'name and shame' and others commented on the tendency to gloss over issues in the interests of corporate image and reputation;
- one respondent noted that the dissemination of information through national CERTs can be useful;
- and one doubted that reporting would really contribute to resilience.

Other points included:

- *"resilience requires a trusted source of information about how the system is working"*;
- if the reporting mechanism hides specifics for security reasons, aggregated and anonymised data must still be published;
- how to report and publish in a useful way which is not also useful to attackers?
- if reports go outside the 'technical community', will the detail required by the technical people be filtered out?
- reporting can be a distraction from problem resolution;

- it is appropriate to involve the police in the reporting process – if only to help them to set priorities;
- EU Telecom Framework and reporting will help.

14. How can BGP be configured and routing policy be organised to improve resilience? What mechanisms exist, or should exist, to manage traffic across multiple networks – as might be required in a large scale incident?

On reflection, this would have been better as two separate questions. On the question of BGP configuration and policy there was general agreement that multi-homing and diverse connections were good, and on two further things: the need for route filtering and BGP security.

1. Improved filtering, of routes and packets, would help – so a further vote for implementation of Best Practice. However, it was also noted that:
    - economic incentives for the implementation of Best Practice are "badly needed";
    - reducing the number of routes (by re-aggregating) would also help, but that this is probably a "utopian goal";
    - Route Flap Damping has been proven to make things worse – not using it is now Best Practice;
    - for connections between large transit providers, it is currently neither feasible nor practical to apply route filtering – large transit providers must depend on their peers and customers filtering their routes;
    - filtering may restrict the ability to implement unusual routing in the event of a crisis;
    - filtering using the available IRR information is seldom attempted.
2. Some securer form of BGP is needed (!):
    - RPKI was cited as a good start – which should be encouraged;
    - economic incentives for the development and deployment are very weak;
    - data to help evaluate the likely effectiveness of any more secure BGP scheme, or the likely deployment issues, is lacking;
    - if a clean-slate redesign is not possible, then significant changes to existing BGP must be made.

Other points included:

- the desirability of improving crisis management systems and skills;
- the desirability of keeping some spare router capacity, CPU and memory, for use in an emergency;
- that political decisions implemented at routing level – e.g. Pakistan/YouTube, Iran/gmail, China/Google –muddy the question of what is a routing failure;
- the lack of accurate published routing policy;
- the desirability of being able to blackhole DDoS attack traffic;

- the need for inter-domain traffic engineering to improve resilience – and the need for more QoS routing support in BGP;
- that some of the theoretical problems with BGP are a current research topic;
- that even now subtle problems with BGP are being encountered.

On the question of managing traffic across multiple networks the response was spotty, but included:

- traffic management across multiple domains is not feasible;
- there is no known mechanism for managing traffic across multiple networks. Possibly data from Routing Registries could be used to help implement something, but it is probably best to involve the major operators, who have the scale to do something here, noting that they are, of course, competitors;
- the operators are best placed to manage traffic and should be left to it;
- *"We question the value of an organisation such as ENISA planning to manage the traffic across networks, even in (perhaps especially in) a crisis situation."*;
- *"you seem to be approaching the issue with the assumption that additional regulations will make the Internet more reliable – I'm far from convinced that is the case"*.

Finally

15. Please tell us if you feel we have missed out any important questions or subject areas which should be addressed when looking at the resilience of the Internet interconnection ecosystem.

We got a variety of distinct responses to this question.

- ENISA should consider the interdependence of multiple network types – for example the mutual dependence of the Internet and electricity supply networks – and the vulnerability of the underlying physical network... eg undersea cables.
- The study should have *"more focus on the regulation part as many people fear it's coming…"* – meaning that more attention should have been given to the arguments for and (especially) against regulatory intervention.
- The European Commission should pay heed to the tendency to oligopoly in: networks, equipment supply, CDN, operating systems, etc.
- The study and questionnaire should have considered privacy issues.
- They should have considered the effects of misbehaving applications and protocols.
- They should have considered the effects of organised crime.
- They were high-level and extremely generic. In particular, they should have considered the rapidly-growing role of social networks, or other application areas.
- One respondent disagreed with the premise that most traffic on the Internet does not pass directly between source and destination – citing the tendency of traffic to be local and to be exchanged locally (within language boundaries and e.g. at local IXPs), and the increasing amounts of traffic delivered locally by CDNs.

- Because the Internet is collection of individual networks, the resilience of the ecosystem as a whole cannot be separate from the resilience of the parts.
- Is there a baseline against which to measure improvements in resilience?
- It would be useful to gather qualitative and quantitative data, regularly, to better understand and monitor the state of Internet resilience... and be ready to act if required.
- Fail-over routes may be short of capacity – which could be mitigated by prioritising critical traffic, but there may well be conflict over what is critical.
- The spread of cloud providers and CDNs affects resilience, because they are replicating data across the network.
- The debate over net neutrality should not be ignored – it will have a bearing on the resilience of the ecosystem.
- *"…top-down regulation with regard to network management— including resilience — is unnecessary".*
- We could look further into the notion of a "business layer" – dynamic support for inter-domain traffic engineering.
- The study does not consider future work on improving the ecosystem – including research. It also does not consider whether or how to migrate from current (deficient) Internet architecture to clean-slate solutions (such as ResiliNets and ResumeNet, or programs such as FIRE and FIND).

Introduction to the Study *(sent with the Questionnaire)*

This document is a brief introduction to the study "Resilience of the Internet Interconnection Ecosystem", which is work sponsored by the "European Network and Information Security Agency – ENISA.

About ENISA

Communication networks and information systems have become an essential factor in economic and social development. Computing and networking are now becoming ubiquitous utilities in the same way as electricity or water supply. The security of communication networks and information systems, in particular their availability, is therefore of increasing concern to society as they deliver services critical to the well-being of European citizens.

The "European Network and Information Security Agency" (ENISA) was established on 10 March 2004. Its purpose is to ensure a high and effective level of network and information security within the Community and to develop a culture of network and information security (NIS), for the benefit of the citizens, consumers, enterprises, and public sector organisations within the European Union (EU), and thereby contributing to the smooth functioning of the Internal Market.

Objectives of the Agency

The Agency's objectives are as follows:

- to enhance the capability of the Community, EU Member States and, as a consequence, the business community to prevent, to address, and to respond to network and information security problems.
- to assist and advise the Commission and EU Member States on issues related to network and information security falling within its competencies as set out in the Regulation[55].
- building on national and Community efforts, to develop a high level of expertise.
- to use this expertise to stimulate broad cooperation between actors from the public and private sectors.

Additional Information

Further information about ENISA can be obtained from its website: www.enisa.europa.eu.

The Subject of the Study

The Internet connects a large number of independent networks, which cooperate to ensure that each network's users can reach every other network's users – directly or, much of the time, indirectly.

The resilience of the Internet as a whole depends on each network – from its end users to its interconnections with other networks – being resilient That is under the control of each network, individually and independently.

Each direct connection between two networks is a bilateral and generally private arrangement. Each direct connection is under the shared control of the two networks.

Most traffic, however, does not pass directly between networks, but crosses one or more other networks between source and destination. These indirect connections are underpinned by a system of incentives and bilateral agreements (formal and informal).

[55] *Regulation (EC) No 460/2004 of the European Parliament and of the Council of 10 March 2004 establishing the European Network and Information Security Agency.*

See: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32004R0460:EN:HTML

The system of direct and indirect connections between networks, and the incentives and agreements that underpin those are, together, the Internet Interconnection Ecosystem.

The resilience of the Internet as a whole depends on the resilience of that interconnection ecosystem, which is beyond the control of any network.

That is the subject of this study.

In the following the Internet Interconnection Ecosystem will generally be referred to simply as the ecosystem.

The Motivation for the Study

Internet technologies are designed to cope with change whatever the cause.

Individual networks are designed to cope automatically with some anticipated level of change. Where a change exceeds the capability of the automatic systems to adjust, the network operations centre will step in. Network operators constantly monitor and maintain their networks, and their day-to-day work is dealing with changes and events.

The interconnection ecosystem is not the entire Internet, but it is an important and central part of it. There is no design or management of the ecosystem. An invisible hand causes tens of thousands of individual networks to work coherently to provide the Internet.

It is believed that the ecosystem is resilient. Experience suggests that it is.

But, there is no way to verify either that it is, or that it will remain so.

Given the importance of the ecosystem, that is a serious deficiency.

That is the motivation for this study.

The Scope of the Study

This study is interested in the resilience of the ecosystem, looking at:

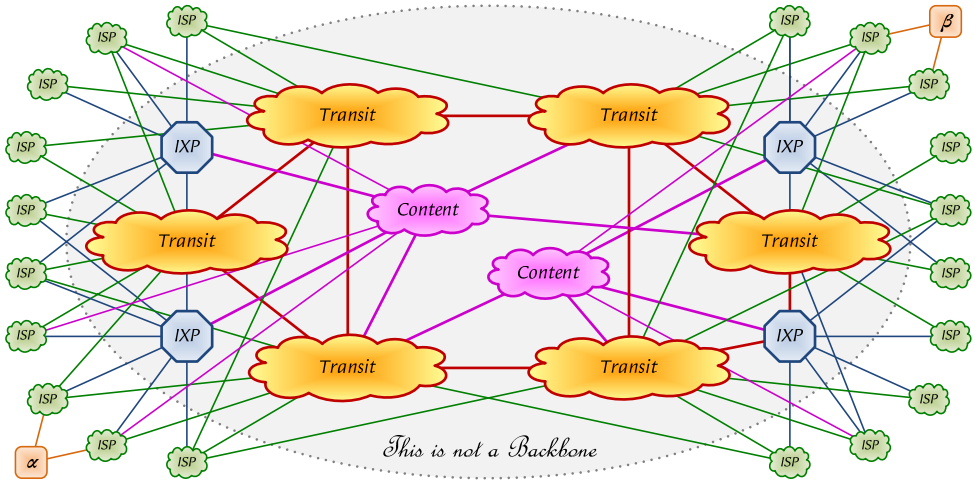
- its response to events with medium to high impact, and which have a medium to low probability.
- how that resilience may be assured and improved.
- what may influence that resilience in the long term.

particularly from a European perspective but, as with anything to do with the Internet, the context is clearly global.

Note that this excludes the day-to-day running of the ecosystem and individual networks. It also excludes the resilience of end-user connections to their ISPs.

A Working Model of the Interconnection Ecosystem

The ecosystem is very complicated, or at least very large. Some model is required, and the following diagram provides a simple one:



where the ecosystem comprises:

1. a relatively small number of (large) transit providers, generally connected to each other (but not universally), either directly or at an IXP. Those transit providers each serve a number of ISPs and content providers/delivery networks.
2. a number of Internet Exchange Points (IXPs), at which transit providers, content providers/delivery networks and ISPs connect to each other.
3. a small number of content providers/delivery networks, connecting to transit providers, IXPs and ISPs.
4. a much larger number of Internet Service Providers (ISPs), connected in various ways to transit providers, to IXPs and to content providers/delivery networks.
5. a still larger number of customers of the ISPs – represented on the diagram by just two of their number: networks α and β. These are relatively sophisticated, multi-homed customers. There is a yet larger number of single homed end users.

The system is, of course, much richer and more varied than the diagram suggests. The model is intended to be simple enough to be tractable, without being too simple to be useful.

Consider for a moment the exchange of traffic between α and β: the resilience of this interconnection depends on all the networks and connections between them, and the interlocking incentives and agreements that keep those willing and able to carry the traffic.

The term backbone suggests some centrally managed infrastructure on which the rest of the system depends. There is no Internet Backbone in that sense. However, as the diagram suggests, key parts of the ecosystem may be treated as a "de facto backbone". The resilience of the ecosystem is then the resilience of this "backbone" and the connections with that.

In general this study is not concerned with the resilience of individual networks. However, where a network is a component of the de facto backbone, its resilience directly affects the resilience of the ecosystem.

The Approach to Resilience

A system is resilient if it continues to offer acceptable service despite damage to, or failure of, parts of the system. This begs a lot of questions about what represents acceptable service for any given degree of damage or failure. Almost any analysis of resilience can quickly become complex as the number of possibilities to consider grows rapidly.

This is a large system so the problem of its resilience is only tractable if a broad approach is taken. It is important to remember that the ecosystem is the sum of a large number of independently managed networks – so considering its resilience is not quite like considering the resilience of even a very large individual network.

Having identified the major components of the ecosystem, it may be possible to identify key factors which affect its resilience. For example:

a. physical diversity – which is not easy for an individual network to achieve and maintain, and much, much harder for a system of independently managed networks.
b. spare capacity –which may be designed into an individual network, but may or may not exist, in the right places, in the system of interconnections between networks.
c. management systems – which in this context means the ability to organise the recovery of the system as a whole, not just individual operator's networks.
d. systemic problems – which, by definition, may be capable of simultaneously affecting large parts of this large system.
e. cascade failure – which could also simultaneously affect large parts of the system.

The other side of the coin is some consideration of what sorts of events might have medium to high impact, and whether, given the nature of the ecosystem, it is possible to identify some general effects on the system. By linking real possibilities to (hopefully) a relatively small number of general effects, it may be possible to draw real conclusions without having to consider an impossibly large number of cases.

Mapping the Ecosystem

To assess the resilience of the ecosystem first it must be possible to see it.

Assuming that a practical model for the ecosystem and its resilience can be devised, little further progress can be made while the system is almost completely hidden.

The model suggests that the IXPs and a relatively small number of networks make up a de facto backbone. Mapping that ecosystem is, at least, a much smaller task than mapping the entire Internet. However, it is still more or less impossible to discover:

a. how the components are connected to each other,
b. whether that mesh is diverse,
c. if there is suitable spare capacity,
d. how well it would respond to a major event,

...or anything else.

The connections and relationships with the clients of a de facto backbone are also part of the ecosystem. For any analysis of this to be tractable it will be necessary to divide the clients into a (hopefully) small number of classes.

The Wider Issues

The Internet works because a system of incentives causes the networks in it to behave coherently, without any central coordination. Internet insiders not only believe that this is the best way to organise such a large system, they believe it is the only way.

However, there appears to be little incentive for a network to spend time and money making the ecosystem as a whole resilient or more resilient. Each individual network gets no direct benefit from such expenditure. Further, it is not clear that the resilience of the ecosystem arises naturally out of the resilience of the individual networks or their individual interconnections, at least when considering large scale events.

With falling prices and increasing demand, the resilience of the ecosystem may not be a priority – even if an operators knew how to help achieve that resilience. Does continual cost reduction reduce the resilience of the components of the ecosystem?

Customers have no way of assessing whether the suppliers they are choosing between are better or worse than each other in terms of resilience, even if they were sophisticated enough to care. There does not appear to be any consumer choice or price signal to promote resilience – even though that appears to be in the end-users' interests.

Customers are motivated by price, and applications are moved to the Internet to reduce costs. Much of the time the Internet works. It is not clear to anyone what risk they are accepting in return for the cost saving.

As transit prices continue to fall, apparently inexorably towards zero, how will that affect the incentives, and what will that do for the resilience of the ecosystem? Indeed, is the ecosystem sustainable in the long term?

The speculative boom created a lot of infrastructure which became very cheap when the bubble burst. If that is still a factor, what happens when new infrastructure has to be paid for at real prices?

The Objectives of the Study

The resilience of the Internet Interconnection Ecosystem is central to the resilience of the Internet and the services delivered over the Internet.

The importance of the Internet means that its resilience cannot simply be assumed, or be accepted as an article of faith.

The objectives of the study are to:

survey the current state-of-the-art:

- what the ecosystem is
- what may affect its resilience
- whether or how that resilience may be assessed (or even verified)

where possible, make recommendations for:

- further work
- action to improve resilience
- other research

PART IV – Annexes

Bibliography

[1] DJ Smith, *Reliability, Maintainability and Risk*, 7th ed.: Elsevier, 2005.

[2] D.D. Woods, N. Leveson E. Hollnagel, *Resilience Engineering: Concepts and Precepts*.: Ashgate Publishing, 2006.

[3] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, pp. 1662-1664, 2002.

[4] A. Wagner, "Robustness and Evolvability: A Paradox Resolved," *ProcBiolSci*, vol. 275, pp. 91-100, Jan 2008.

[5] Craig Partridge et al., *The Internet under Crisis Conditions: Learning from September 11*. Washington: The National Acadamies Press, 2002.

[6] P. Ferguson and D. Senie. (2000, May) RFC2827/BCP38: Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing. [Online]. http://www.ietf.org/rfc/rfc2827.txt

[7] M. De Bruijne, "Systems that Should Have Failed: Critical Infrastructure Protection in an Institutionally Fragmented Environment," *Journal of Contingencies and Crisis Management*, vol. 15, no. 1, pp. 1-29, Mar 2007.

[8] Andrew Odlyzko. (2010, Jan) Collective hallucinations and inefficient markets: the British railway mania of the 1840s. [Online]. http://www.dtc.umn.edu/~odlyzko/doc/hallucinations.pdf

[9] P. Faratin et al., "The Growing Complexity of Internet Interconnection," *Communications & Strategies, No. 72, p. 51, 4th Quarter 2008* , vol. 72, p. 51, 2008.

[10] Amogh Dhamdhere and Constantine Dovrolis, "Ten Years in the Evolution of the Internet Ecosystem," in *IMC'08*, Vouliagmeni, Greece, 2008.

[11] Geoff Huston et al. CIDR Report. [Online]. http://www.cidr-report.org/

[12] Andrew Odlyzko. Minnesota Internet Traffic Studies (MINTS). [Online]. http://www.dtc.umn.edu/mints/

[13] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian, "Internet Inter-Domain Traffic," in *SIGCOMM'10*, New Delhi, 2010.

[14] Cisco Systems. (2010, Jun) Cisco Visual Networking Index: Forecast and Methodology, 2009-2014. [Online]. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[15] Andrew Odlyzko. (2000, Nov) Internet Growth: Myth and Reality, Use and Abuse. [Online]. http://www.dtc.umn.edu/~odlyzko/doc/internet.growth.myth.pdf

[16] Serge Radovcic. (2010, Oct) European Internet Exchange Association 2010 Report on European IXPs. [Online]. https://www.euro-ix.net/resources/reports/euro-ix_report_2010.pdf

[17] Y. Rekhter, T. Li, and S. Hares. (2006, Jan) RFC 4271: A Border Gateway Protocol 4 (BGP-4). [Online]. http://www.ietf.org/rfc/rfc4271

[18] Matthew Caesar and Jennifer Rexford, "BGP Routing Policies in ISP Networks," *IEEE Network*, November/December 2005.

[19] Nick Feamster, Jared Winick, and Jennifer Rexford, "A Model of BGP Routing for Network Engineering," in *SIGMETRICS/Performance'04*, New York, 2004, pp. 331-342.

[20] Sharad Agarwal, Chen-Nee Chuah, Supratik Bhattacharyya, and Christophe Diot, "The Impact of BGP Dynamics on Intra-Domain Traffic," in *SIGMETRICS/Performance'04*, New York, 2004, pp. 319-330.

[21] F. Bruce Shephard, Gordon Wilfong Timothy G. Griffin, "The Stable Paths Problem and Interdomain Routing," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, Apr 2002.

[22] Timothy G. Griffin and Brian J. Premore, "An Experimental Analysis of BGP Convergence Time," in *ICNP'01*, Riverside, California, 2001 .

[23] Xiaoliang Zhao, Beichuan Zhang, Dan Massey, Lixia Zhang Mohit Lad, "Analysis of BGP Update Surge during Slammer Worm Attack," in *IWDC'03*, 2003.

[24] Xiaoliang Zhao, Dan Pei, Randy Bush, Daniel Massey, Allison Mankin, S. Felix Wu, Lixia Zhang Lan Wang, "Observation and Analysis of BGP Behavior under Stress," in *IMW'02*, Marseille, France, 2002.

[25] Peidong Zhu, Xicheng Lu, Bernhard Plattner Wenping Deng, "On Evaluating BGP Routing Stress Attack," *Journal of Communications*, vol. 5, no. 1, Jan 2010.

[26] George Kesidis Glenn Carl, "Large-Scale Testing of the Internet's Border Gateway Protocol (BGP) via Topological Scale-Down," *ACM Transactions onModeling and Computer Simulation*, vol. 18, no. 3, Jul 2008.

[27] Feng Wang, Jian Qiu, Lixin Gao, and Jia Wang, "On Understanding Transient Interdomain Routing Failures," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 17, no. 3, pp. 740-751, Jun 2009.

[28] Feng Wang, Zhouqing Morely Mao, Jia Wang, Lixin Gao, and Randy Bush, "A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance," in *SIGCOMM'06*, Pisa, 2006.

[29] Z. Morley Mao, Jia Wang Ying Zhang, "A Framework for Measuring and Predicting the Impact of Routing Changes," in *INFOCOM'07*, 2007.

[30] Abba Ahuja, Abhijit Bose, Farnam Jahanian Craig Lavovitz, "Delayed Internet Routing Convergence," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, Jun 2001.

[31] Nate Kushman, Srikanth Kandula, and Dina Katabi, ""Can You Hear Me Now? It Must Be BGP," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 2, pp. 75-84, Apr 2007.

[32] Gianluca Iannaccone, Christophe Diot Catherine Boutremans, "Impact of link failures on VoIP performance," in *NOSSDAV '02*, 2002.

[33] Randy Bush, Olaf Maennel, Matthew Roughan, and Steve Uhlig, "Internet Optometry: Assessing the Broken Glasses in Internet Reachability," in *IMC'09*, Chicago, 2009.

[34] Ricardo Oliveira, Lixia Zhang Ying-Ju Chi, "Cyclops: The AS-level Connectivity Observatory," *ACM SIGCOMM Computer Communications Review*, vol. 38, no. 5, pp. 7-16, Oct 2008.

[35] Randy Bush, Timothy G. Griffin, Matthew Roughan Z. Morley Mao, "BGP Beacons," in *IMC'03*, Miami Beach, Florida, 2003.

[36] Yuval Shavitt and Eran Shir, "DIMES: Let the Internet Measure Itself," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 5, Oct 2005.

[37] Brice Augustin, Balachander Krishnamurthy, and Walter Willinger, "IXPs: Mapped ?," in *IMC'09*, Chicago, 2009, pp. 336-349.

[38] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang, "The (In)Completeness of the Observed Internet AS-level Structure," *IEEE/ACM Transactions on Networking*, vol. 18, no. 1, pp. 109-112, Feb 2010.

[39] Fan Chung, kc claffy, Marina Formenkov, Alessandro Vespignani, Walter Willinger Dmitri Krioukov, "The Workshop on Internet Topology (WIT) Report," *ACM SIGCOMM Computer Communications Review*, vol. 37, no. 1, Jan 2007.

[40] Jay Borkenhagen, Jennifer Rexford Nick Feamster, "Guidelines for Interdomain Traffic Engineering," *ACM SIGCOMM Computer Communications Review*, vol. 33, no. 5, Oct 2003.

[41] Olivier Bonaventure, Vincent Magnin, Chris Rapier, Lica Deri Steve Uhlig, "Implications of the Topological Properties of Internet Traffic on Traffic Engineering," in *SAC'04*, Nicosia, Cyprus, 2004.

[42] Cristel Pelsser, Louis Swinnen, Olivier Bonaventure, Steve Uhlig Bruno Quoitin, "Interdomain Traffic Engineering with BGP," *IEEE Communications Magazine*, vol. 41, no. 5, May 2003.

[43] Aditya Akella, Almir Mutapcic Gureesh Shrimali, "Cooperative Inter-Domain Traffic Engineering Using Nash Bargaining and Decomposition," in *INFOCOM'07*, 2007.

[44] Aman Shaikh, Tim Griffin, Jennifer Rexford Renata Teixeira, "Dynamics of hot-potato routing in IP networks," in *SIGMETRICS '04*, New York, New York, 2004.

[45] S. Murphy. (2006, Jan) RFC4272: BGP Security Vulnerabilities Analysis S. Murphy. [Online]. http://www.ietf.org/rfc/rfc4272.txt

[46] Kevin Butler, Toni R. Farley, Patrick McDaniel, and Jennifer Rexford, "A Survey of BGP Security Issues and Solutions," *Proceedings of the IEEE*, vol. 98, no. 1, pp. 100-122, Jan 2010.

[47] Sparta, Inc. (2006, Sep) Secure Protocols for the Routing Infrastructure (SPRI) Initiative: A Road Map (First Draft). [Online]. http://www.cyber.st.dhs.gov/docs/spriRoadmap.pdf

[48] Stephen Kent, Charles Lynn, and Karen Seo, "Secure Border Gateway Protocol (S-BGP)," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 4, Apr 2000.

[49] Stephen T. Kent, "Securing the Border Gateway Protocol," *The Internet Protocol Journal*, vol. 6, no. 3, Sep 2003.

[50] Russ White, "Securing BGP Through Secure Origin BGP," *The Internet Protocol Journal*, vol. 6, no. 3, Sep 2003.

[51] James Ng (Ed). (2004, Apr) Extensions to BGP to Support Secure Origin BGP (soBGP). [Online]. http://tools.ietf.org/pdf/draft-ng-sobgp-bgp-extensions-02.pdf

[52] S. Turner M. Lepinski. (2011, Mar) An Overview of BGPSEC. [Online]. http://tools.ietf.org/pdf/draft-lepinski-bgpsec-overview-00.pdf

[53] S. Kent M. Lepinski. (2011, Feb) An Infrastructure to Support Secure Internet Routing. [Online]. http://tools.ietf.org/pdf/draft-ietf-sidr-arch-12.pdf

[54] R. Loomans, G. Michaelson G. Huston. (2011, Feb) A Profile for Resource Certificate Repository Structure. [Online]. http://tools.ietf.org/pdf/draft-ietf-sidr-repos-struct-07.pdf

[55] S. Kent. (2011, Feb) Threat Model for BGP Path Security. [Online]. http://tools.ietf.org/pdf/draft-kent-bgpsec-threats-01.pdf

[56] Meiyuan Zhao, Sean Smith, and David M. Nicol, "The Performance Impact of BGP Security," *IEEE Network*, Nov 2005.

[57] R. Bush. (2011, Feb) BGPsec Operational Considerations. [Online]. http://tools.ietf.org/pdf/draft-ymbk-bgpsec-ops-00.pdf

[58] Geoffrey Goodell et al., "Working Around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing," in *NDSS Symposium 2003, Internet Society (ISOC)*, 2003.

[59] Yih-Chun Hu, Adrian Perrig, and Marvin Sirbu, "SPV: Secure Path Vector Routing for Securing BGP," in *SIGCOMM'04*, Portland, Oregon, 2004.

[60] Barath Raghavan, Saurabh Panjwani, and Anton Mityagin, "Analysis of the SPV Secure Routing Protocol: Weaknesses and Lessons," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 2, Apr 2007.

[61] Tao Wan, Evangelos Kranakis, and P.C. van Oorschot, "Pretty Secure BGP (psBGP)," in *NDSS Symposium 2005, Internet Society (ISOC)*, 2005.

[62] Josh Karlin, Stephanie Forrest, and Jennifer Rexford, "Autonomous Security for Autonomous Systems," *Computer Networks 52 (2008) 2908–2923*, vol. 52, pp. 2908–2923, Jun 2008.

[63] Sharon Goldberg, Michael Schapira, Peter Hummon, and Jennifer Rexford, "How secure are secure interdomain routing protocols," in *SIGCOMM '10*, New Delhi, 2010.

[64] Debabrata Dash, Adrian Perrig, Hui Zhang Haowen Chan, "Modeling Adoptability of Secure BGP Protocols," in *SIGCOMM'06*, Pisa, Italy, 2006, pp. 279-290.

[65] Fouad Tobagi, Christophe Diot Chuck Fraleigh, "Provisioning IP Backbone Networks to Support Latency Sensitive Traffic," in *INFOCOM'03*, 2003.

[66] Yann Labit, Jordi Domingo-Pascual, Philippe Owezarski Rene Serral-Gracia, "Towards an Efficient Service Level Agreement Assessment," in *INFOCOM'09*, 2009.

[67] Constantine Dovrolis Ravi S. Prasad, "Measuring the Congestion Responsiveness of Internet Traffic," in *PAM*, 2007.

[68] David Clark, William Lehr Steven Bauer, "The Evolution of Internet Congestion," in *TPRC*, Arlington, Virginia, 2009.

[69] Rüdiger Martin, Joachim Charzinski Michael Menth, "Capacity Overprovisioning for Networks with Resilience," in *SIGCOMM'06*, Pisa, Italy, 2006, pp. 87-98.

[70] Albert Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford Anja Feldmann, "Deriving Traffic Demands for Operational IP Networks: Methodology and Experience," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, Jun 2001.

[71] Guillaume Dewaele, Kensuke Fukuda, Patrice Abry, Kenjiro Cho Pierre Borgnat, "Seven Years and One Day: Sketching the Evolution of Internet Traffic," in *INFOCOM'09*, 2009.

[72] kc claffy, Marina Fomenkov, Dhiman Barman, Michalis Faloutsos, KiYoung Lee Hyunchul Kim, "Internet Traffic Classification Demystified: Myths, Caveats and the Best Practices," in *CoNEXT'08*, 2008.

[73] Zhenhai Duan, Zhi-Li Zhang, Jaideep Chandrashekar Kuai Xu, "On Properties of Internet Exchange Points and Their Impact on AS Topology and Relationship," in *Networking 2004*, 2004.

[74] Geoff Huston, "Interconnection, Peering and Settlements - Part I," *Internet Protocol Journal*, vol. 2, no. 1, Mar 1999.

[75] Geoff Huston, "Interconnection, Peering and Settlements - Part II," *Internet Protocol Journal*, vol. 2, no. 2, Jub 1999.

[76] Chris Owens Lyman Chapin. (2005) Interconnection and Peering among Internet Service Providers: A Historical Perspective. [Online]. http://www.interisle.net/sub/ISP%20Interconnection.pdf

[77] Nicolas Economides, "The Economics of the Internet Backbone," in *Handbook of Telecommunications Economics, Vol 2*, S Majumdar et al, Ed.: Elsevier, 2005, ch. 9.

[78] Paul Hurley, Andreas Kind, Marc Ph. Stoecklin Xenofontas Dimitropoulos, "On the 95-Precentile Billing Method," in *PAM*, 2009.

[79] Bruce Maggs, Srinivasan Seshan, Anees Shaikh, Ramesh Sitaraman Aditya Akella, "A Measurement-Based Analysis of Multihoming," in *SIGCOMM'03*, Karlsruhe, Germany, 2003.

[80] Bill Norton. (2003, Nov) The Evolution of the US Internet Peering Ecosystem. [Online]. http://www.nanog.org/meetings/nanog31/presentations/norton.pdf

[81] Bill Norton. (2010, Aug) Internet Service Providers and Peering v3.0. [Online]. http://drpeering.net/white-papers/Internet-Service-Providers-And-Peering.html

[82] CAIDA. (2010, Nov) AS Rank. [Online]. http://as-rank.caida.org/

[83] CAIDA. (2010, Nov) Introduction to Relationship-based AS Ranking. [Online]. http://www.caida.org/research/topology/rank_as/

[84] Renesys. (2007, Nov) Rankings, Damned Rankings and Statistics. [Online]. http://www.renesys.com/tech/presentations/pdf/menog2.pdf

[85] Renesys Blog, Earl Zmijewski. (2010, Jul) What happened to Sprint? [Online]. http://www.renesys.com/blog/2010/07/what-happened-to-sprint.shtml

[86] Bill Norton. (2010) The Art of Peering: The Peering Playbook. [Online]. http://drpeering.net/white-papers/Art-Of-Peering-The-Peering-Playbook.html

[87] Bill Norton. (2010) Peering Policy. [Online]. http://drpeering.net/white-papers/Peering-Policies/Peering-Policy.html

[88] Bill Norton. (2010) A Study of 28 Peering Policies. [Online]. http://drpeering.net/white-papers/Peering-Policies/A-Study-of-28-Peering-Policies.html

[89] Greg Goth, "New Internet Economics Might Not Make it to the Edge," *IEEE Internet Computing*, Feb 2010.

[90] Renesys Blog, Bob Fletcher. (2009, Nov) IP Backbone: Hard sell, not so much. [Online]. http://www.renesys.com/blog/2009/11/ip-backbone-hard-sell-not-so-m.shtml

[91] Bill Norton. (2005, Oct) The Folly of Peering Ratios (as a Peering Candidate Discriminator). [Online]. http://drpeering.net/white-papers/The-Folly-Of-Peering-Ratios.html

[92] Marina Fomenkov, Ethan KatzBassett, Robert Beverly, Beverly A. Cox, Matthew Luckie kc claffy, "The Workshop on Active Internet Measurements (AIMS) Report," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 5, Oct 2009.

[93] David Clark, William Lehr Steven Bauer, "Broadband Microfoundations: the Need for Traffic Data," in *Beyond Broadband Access*, Washington, DC, 2009.

[94] kc claffy Scott Bradner, "The (un)Economic Internet?," *IEEE Internet Computing*, Jun 2007.

[95] Hyunseok Chang and Walter Willinger, "Difficulties Measuring the Internet's AS-Level Ecosystem," in *Information Sciences & Systems 2006*, 2006.

[96] Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger, "Towards Capturing Representative AS-level Internet Topologies," *Computer Networks*, vol. 44, pp. 737-755, 2004.

[97] Yihua He, Georgos Siganos, Michalis Faloutsos, and Srikanth Krishnamurthy, "A Systematic Framework for Unearthing the Missing Links: Measurements and Impact," in *NSDI'07*, 2007, pp. 187-200.

[98] Lun Li, David Alderson, Walter Willinger, and John Doyle, "A First Principles Approach to Understanding the Internet's Router-Level Topology," in *SIGCOMM'04*, Portland, 2004.

[99] Xenofontas Dimitropoulos et al., "AS Relationships: Inference and Validation," *ACM SIGCOMM Computer Communications Review*, vol. 37, no. 1, Jan 2007.

[100] Lun Li, Walter Willinger, John C. Doyle David Alderson, "Understanding Internet Topology: Principles, Models, and Validation," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, Dec 2005.

[101] Brian Eriksson, Gautam Dasarathy, Paul Barford, and Robert Nowak, "Toward the Practical Use of Network Tomography for Internet Topology Discovery," in *INFOCOM 2010*, San Diego, 2010.

[102] Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson, "Measuring ISP Topologies With Rocketfuel," *IEEE/ACM TRANSACTIONS ON NETWORKING*, vol. 12, no. 1, Feb 2004.

[103] Dean Bubley Chris Barraclough et al. (2010, Sep) Net Neutrality 2.0: Don't Block the Pipe, Lubricate the Market. [Online]. http://www.telco2research.com/downloads/20100915

[104] Scott Jordan, "Implications of Internet Architecture on Net Neutrality," *ACM Transactions on Internet Technology*, vol. 9, no. 2, May 2009.

[105] Greg Goth, "Traffic Management Becoming High Priority Problem," *IEEE Internet Computing*, Dec 2008.

[106] Comcast. (2008, Aug) Announcement Regarding An Amendment to Our Acceptable Use Policy. [Online]. http://www.comcast.net/terms/network/amendment/

[107] Sam Diaz. (2008, Aug) Comcast's Web limits: Saving bandwidth or stifling innovation? [Online]. http://www.zdnet.com/blog/btl/comcasts-web-limits-saving-bandwidth-or-stifling-innovation/9838

[108] Scott J. Berry. (2008, Sep) Comcast Limits User Downloads: Wrong Solution. [Online]. http://seekingalpha.com/article/93860-comcast-limits-user-downloads-wrong-solution

[109] Carl Weinschenk. (2008, Sep) Comcast Bandwidth Limits Anger Observers. [Online]. http://www.itbusinessedge.com/cm/blogs/weinschenk/comcast-bandwidth-limits-anger-observers/?cs=13286

[110] Jon M. Peha, "The Benefits and Risks of Mandating Network Neutrality, and the Quest for a Balanced Policy," in *TPRC*, Arlington, Virginia, 2006.

[111] Sharon Gillett, Marvin A. Sirbu, Jon M. Peha William Lehr, "Scenarios for Network Neutrality Arms Race," in *TPRC*, Arlington, Virginia, 2006.

[112] P2P ON! (2010) List of Internet Service Providers That Throttle P2P Traffic. [Online]. http://www.p2pon.com/guides/list-of-internet-service-providers-that-throttle-p2p-traffic/

[113] Daniel J. Weitzner, "Net Neutrality. Seriously this Time," *IEEE Internet Computing*, Jun 2008.

[114] Grant Gross. (2010, Jun) Judge Approves Comcast Traffic Throttling Settlement. [Online]. http://www.pcworld.com/businesscenter/article/200790/judge_approves_comcast_traffic_throttling_settlement.html

[115] EC DG Information Society and Media. (2010, Jun) Open internet and net neutrality. [Online]. http://ec.europa.eu/information_society/policy/ecomm/library/public_consult/net_neutrality/index_en.htm

[116] EC DG Information Society and Media. (2010, Nov) Report on the public consultation on 'The open internet and net neutrality in Europe'. [Online]. http://ec.europa.eu/information_society/policy/ecomm/doc/library/public_consult/net_neutrality/report.pdf

[117] Martin Arlitt, Zongpeng Li, Anirban Mahanti Phillipa Gill, "The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse ?," in *PAM'08*, 2008.

[118] Constantine Dovrolis Amogh Dhamdhere, "Can ISPs be Profitable Without Violating "Network Neutrality" ?," in *NetEcon'08*, 2008.

[119] Hendrik Schulze and Klaus Mochalski. (2009, Feb) Internet Study 2008/2009. [Online]. http://www.ipoque.com/userfiles/file/ipoque-Internet-Study-08-09.pdf

[120] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman, "On Dominant Characteristics of Residential Broadband Internet Traffic," in *IMC'09*, Chicago, 2009, pp. 90-102.

[121] Margit A. Vanberg, "Competition and Cooperation in Internet Backbone Services," in *Telecommunication Markets, Drivers and Impediments*, Brigitte Preissl, Justus Haucap, and Peter Curwen, Eds.: Physica Verlag Heidelberg, 2009.

[122] Jane van Beelen and John Rolland. (2000, Oct) The International Internet Interconnection Issue. [Online]. http://www.isoc.org/oti/articles/1000/vanbeelen.html

[123] ITU Study Group 3. (2006, Jul) Cost of international internet connectivity (IIC) too high says ITU group. [Online]. http://www.itu.int/ITU-T/newslog/Cost+Of+International+Internet+Connectivity+IIC+Too+High+Says+ITU+Group.aspx

[124] ITU Study Group 3. (2008, Oct) International Internet Connectivity. [Online]. http://www.itu.int/ITU-T/studygroups/com03/iic/index.html

[125] ITU Study Group 3. (2000, Oct) Recommendation D.50: International Internet Connection (10/00). [Online]. http://www.itu.int/rec/T-REC-D.50-200010-S/en

[126] ITU Study Group 3. (2004, Jun) Recommendation D.50: International Internet Connection (06/04). [Online]. http://www.itu.int/rec/T-REC-D.50-200406-S!Amd1/en

[127] ITU Study Group 3. (2008, Oct) Recommendation D.50: International Internet Connection (10/08). [Online]. http://www.itu.int/rec/T-REC-D.50-200810-I/en

[128] Eric Lie. (2007, Feb) International Internet Interconnection. [Online]. http://www.itu.int/ITU-D/treg/Events/Seminars/GSR/GSR07/discussion_papers/Eric_lie_international_interconnection.pdf

[129] James Alleman and Jonathan Liebenau, "Network Resilience and its Regulatory Inhibitors," in *Global Economy and Digital Society*, Erik Bohlin et al., Eds.: Elsevier Science Ltd, 2004, pp. 379-394.

[130] E. Hollnagel, D.D. Woods, and N Leveson, *Reslience Engineering: Concepts and Precepts*.: Ashgate Publishing, 2006.

[131] Yossi Sheffi, *The Resilient Enterprise: Overcoming Vulnerability for Competitive Enterprise*.: MIT Press, 2005.

[132] E Hollnagel, CP Nemeth, and S Dekker, *Resilience Engineering Perspectives, vol 1: Remaining Sensitive to the Possibility of Failure*.: Ashgate Publishing Ltd., 2008.

[133] E Hollnagel, CP Nemeth, and S Dekker, *Resilience Engineering Perspectives, vol 2: Preparation and Restoration*.: Ashgate Publishing Ltd., 2009.

[134] Patricia Longstaff. (2005, Nov) Security, Resilience, and Communication in Unpredictable Environments Such as Terrorism, Natural Disasters and Complex Technology. [Online]. http://pirp.harvard.edu/pubs_pdf/longsta/longsta-p05-3.pdf

[135] Roshanak Nilchiani, Ali Mostashari Mayada Omer, "Measuring the Resilience of the Global Internet Infrastructure System," in *SysCon'09*, Vancouver, Canada, 2009.

[136] Roshanak Nilchiani Jason Hoffman, "Assessing Resilience in the US National Energy Infrastructure," Centre for Complex Adaptive Sociotechnical Systems, 2008.

[137] Marguerite Reardon. (2002, Apr) WorldCom's IP Outages: Whodunnit? [Online]. http://www.lightreading.com/document.asp?doc_id=14501

[138] J. C. Knight and N. G. Leveson, "An experimental evaluation of the assumption of independence in multiversion programming.," *IEEE Trans. Softw. Eng.*, vol. 12, no. 1, pp. 96-109, Jan 1986.

[139] Ying Zhang, Z. Morley Mao, Kang G. Shin Jian Wu, "Internet Routing Resilience to Failures: Analysis and Implications," in *CoNEXT'07*, New York, New York, 2007.

[140] North American Electric Reliability Corporation (NERC). (2010, Nov) Critical Infrastructure Protection: High-Impact, Low-Frequency Risks. [Online]. http://www.nerc.com/page.php?cid=6%7C69%7C327

[141] L. Zhang, K. Fall (Eds) D. Meyer. (2007, Sep) RFC4984: Report from the IAB Workshop on Routing and Addresssing. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc4984.txt.pdf

[142] Alin C. Popescu, Brian J. Premore, and Todd Underwood. (2005, May) The Anatomy of a Leak: AS9121. [Online]. http://www.renesys.com/tech/presentations/pdf/renesys-nanog34.pdf

[143] Larry J. Blunk. (2005, Mar) New BGP analysis tools and a look at the AS9121 Incident. [Online]. http://iepg.org/march2005/bgptools+as9121.pdf

[144] Todd Underwood. (2005, Dec) Internet-Wide Catastrophe—Last Year. [Online]. Internet-Wide Catastrophe—Last Year

[145] Vincent J. Bono. (1997, Apr) 7007 Explanation and Apology. [Online]. http://www.merit.edu/mail.archives/nanog/1997-04/msg00444.html

[146] CNet. (1997, Apr) Router glitch cuts Net access. [Online]. http://news.cnet.com/2100-1033-279235.html

[147] Adrian Chadd. (2006, Aug) Murphy's Law Strikes Again: AS7007. [Online]. http://lists.ucc.gu.uwa.edu.au/pipermail/lore/2006-August/000040.html

[148] RIPE NCC. (2008, Feb) YouTube Hijacking: A RIPE NCC RIS case study. [Online]. http://www.ripe.net/news/study-youtube-hijacking.html

[149] Declan McCullagh. (2008, Feb) How Pakistan knocked YouTube offline (and how to make sure it never happens again). [Online]. http://news.cnet.com/8301-10784_3-9878655-7.html

[150] Martin A. Brown. (2008, Feb) Pakistan hijacks YouTube. [Online]. http://www.renesys.com/blog/2008/02/pakistan_hijacks_youtube_1.shtml

[151] Danny McPherson. (2008, Feb) Internet Routing Insecurity:Pakistan Nukes YouTube? [Online]. http://asert.arbornetworks.com/2008/02/internet-routing-insecuritypakistan-nukes-youtube/

[152] Pakistan Telecommunication Authority. (2008, Feb) Blocking of Offensive Website. [Online]. http://www.renesys.com/blog/pakistan_blocking_order.pdf

[153] Erik Romijn. (2010, Aug) RIPE NCC and Duke University BGP Experiment. [Online]. http://labs.ripe.net/Members/erik/ripe-ncc-and-duke-university-bgp-experiment

[154] Erik Romijn. (2010, Aug) Re: Did your BGP crash today? [Online]. http://www.merit.edu/mail.archives/nanog/msg11505.html

[155] J. Scudder and E. Chen. (2010, Sep) Error Handling for Optional Transitive BGP Attributes, draft-ietf-idr-optional-transitive-03.txt. [Online]. http://tools.ietf.org/pdf/draft-ietf-idr-optional-transitive-03.pdf

[156] Malcolm Fried and Lars Klemming. (2008, Dec) Severed Cables in Mediterranean Disrupt Communication. [Online]. http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aBa0lTN.dcoQ

[157] Renesys Blog, Alin Popescu. (2008, Dec) Deja Vu All Over Again: Cables Cut in the Mediterranean. [Online]. http://www.renesys.com/blog/2008/12/deja-vu-all-over-again-cables.shtml

[158] Renesys Blog, Earl Zmijewski. (2008, Jan) Mediterranean Cable Break - Part II. [Online]. http://www.renesys.com/blog/2008/01/mediterranean-cable-break-part-1.shtml

[159] Renesys Blog, Earl Zmijewski. (2008, Feb) Mediterranean Cable Break - Part III. [Online]. http://www.renesys.com/blog/2008/02/mediterranean-cable-break-part.shtml

[160] Renesys Blog, Earl Zmijewski. (2008, Feb) Mediterranean Cable Break - Part IV. [Online]. http://www.renesys.com/blog/2008/02/mediterranean-cable-break-part-3.shtml

[161] Telegeography. (2008, Feb) Four international cable breaks in a week. [Online]. http://www.telegeography.com/cu/article.php?article_id=21567

[162] Alin Popescu, Todd Underwood, and Earl Zmijewski. (2007, Feb) The Taiwan Earthquakes and the Internet Routing Table. [Online]. http://renesys.com/tech/presentations/pdf/nanog39.pdf

[163] Renesys Blog, Todd Underwood. (2007, Jan) The Shape of Disaster on the Net. [Online]. http://www.renesys.com/blog/2007/01/the_shape_of_disaster_on_the_n.shtml

[164] Telegeography. (2007, Jan) Earthquake Highlights Asian Dependency on Submarine Cables. [Online]. http://www.telegeography.com/wordpress/index.html%3Fp=45.html

[165] Youngseok Lee, Ryo Sakiyama, Koji Okamura Yasuichi Kitamura, "Experience with Restoration of Asia Pacific Network Failures from Taiwan Earthquake," *IEICE Trans. Commun.*, vol. Vol E90-B, no. No 11, Nov 2007.

[166] The Guardian. (2009, Nov) Brazilian power cut leaves 60 million in the dark. [Online]. http://www.guardian.co.uk/world/2009/nov/11/brazil-power-cut-rio-madonna

[167] James Cowie Renesys Blog. (2009, Nov) Lights Out in Rio. [Online]. http://www.renesys.com/blog/2009/11/lights-out-in-rio.shtml

[168] CBS "60 Minutes". (2009, Nov) Cyber War: Sabotaging the System. [Online]. http://www.cbsnews.com/stories/2009/11/06/60minutes/main5555565.shtml

[169] Agência Nacional de Energia Elétrica (ANEEL) do Brasil. (2009, Jan) Recurso Administrativo interposto pela empresa Furnas Centrais Elétricas S.A. - FURNAS, em face do Auto de Infração – AI nº 036/2008-SFE. [Online]. http://www.aneel.gov.br/cedoc/adsp2009278_1.pdf

[170] ambientebrasil. (2007, Sep) Furnas diz que apagão no RJ e ES foi causado por fuligem de queimadas. [Online]. http://noticias.ambientebrasil.com.br/clipping/2007/09/29/33797-furnas-diz-que-apagao-no-rj-e-es-foi-causado-por-fuligem-de-queimadas.html

[171] Jorge Miguel Ordacgi Filho, "Brazilian Blackout 2009," *PAC World* , Mar 2010.

[172] Operador Nacional do Sistema Elétrico. (2009, Dec) Análise da Perturbação do dia 10/11/2009 às 22:13 Envolvendo o Desligamento dos Três Circuitos da LT 765 kV Itaberá-Ivaiporã. [Online]. http://www.mme.gov.br/mme/galerias/arquivos/conselhos_comite/CMSE/2010/Anexo_2_-_Relatxrio_de_Anxlise_da_Perturbaxo_-_RAP_xONS-RE-3-252-2009x.pdf

[173] Juan Pablo Conti, "'Hackers not to blame' for Brazil blackout," *Engineering and Technology Magazine*, vol. 4, no. 21, Dec 2009.

[174] TheRegister, John Leyden. (2010, Apr) China routing snafu briefly mangles interweb - Cockup, not conspiracy. [Online]. http://www.theregister.co.uk/2010/04/09/china_bgp_interweb_snafu/

[175] Renesys Blog, James Cowrie. (2010, Apr) How To Build A Cybernuke. [Online]. http://www.renesys.com/blog/2010/04/how-to-build-a-cybernuke.shtml

[176] National Defense Blog, Stew Magnuson. (2010, Nov) Cyber Experts Have Proof That China Has Hijacked U.S.-Based Internet Traffic: UPDATED. [Online]. http://www.nationaldefensemagazine.org/blog/Lists/Posts/Post.aspx?ID=249

[177] MacAfee Blog, Dmitri Alperovitch. (2010, Nov) April Route Hijack: Sifting through the confusion. [Online]. http://blogs.mcafee.com/mcafee-labs/april-route-hijack-sifting-through-the-confusion-2

[178] McAfee Blog, Dmitri Alperovitch. (2010, Nov) U.S.-Based Internet Traffic Redirected to China. [Online]. http://blogs.mcafee.com/mcafee-labs/u-s-based-internet-traffic-redirected-to-china

[179] U.S.-China Economic and Security Review Commission. (2010, Nov) 2010 Annual Report to Congress. [Online]. http://www.uscc.gov/annual_report/2010/Chapter5_Section_2(page236).pdf

[180] TheRegister. (2005, Oct) Level 3 depeers Cogent. [Online]. http://www.theregister.co.uk/2005/10/06/level3_cogent/

[181] Iljitsch van Beijnum. (2008, Mar) Playing chicken: ISP depeering a high-school lovers' quarrel. [Online]. http://arstechnica.com/old/content/2008/03/isps-disconnect-from-each-other-in-high-stakes-chicken-game.ars

[182] TheRegister. (2008, Oct) Sprint accused of 'partitioning internet'. [Online]. http://www.theregister.co.uk/2008/10/31/congent_sprint_spat/

[183] Xavier Masip-Bruin, Oliver Bonaventure Marcelo Yannuzzi, "Open Issues in Interdomain Routing: A Survey," *IEEE Network*, Nov 2005.

[184] Amund Kvalein, Constantine Dovrolis Ahmed Elmokashfi, "On the Scalability of BGP: The Roles of Topology Growth and Update Rate-Limiting," in *CoNEXT'08*, 2008.

[185] Malleswari Saranu, Joel M. Gottlieb, Dan Pei Lan Wang, "Understanding BGP Session Failures in a Large ISP," in *INFOCOMM'07*, 2007.

[186] Thomas Telkamp. (2009, Sep) Peering Planning Cooperation: Failover Matrices. [Online]. http://www.uknof.org.uk/uknof14/Telkamp-Failover.pdf

[187] Cyclops Project. (2010) Welcome to Cyclops. [Online]. http://cyclops.cs.ucla.edu/

[188] Jun Li Toby Ehrenkranz, "On the State of IP Spoofing Defence," *ACM Transactions on Internet Technology, Vol 9, No 2, May 2009*, vol. 9, no. 2, May 2009.

[189] Arthur Berger, Young Hyun, k claffy Robert Beverly, "Understanding the Efficacy of Deployed Internet Source Address Validation Filtering," in *IMC'09*, Chicago, Illinois, 2009, pp. 356-369.

[190] Wolfgang Mühlbauer, Steve Uhlig, Randy Bush, Pierre Francois, Olaf Maennel Luca Cittadini, "Evolution of Internet Address Space Deaggregation: Myths and Reality," *IEEE Journal on Selected Areas in Communications,* , vol. 28, no. 8, Oct 2010.

[191] Geoff Huston. (2010, Oct) The ISP Column: When? [Online]. http://www.potaroo.net/ispcol/2010-10/when.pdf

[192] Geoff Huston. (2010, Sep) The ISP Column: A Rough Guide to Address Exhaustion. [Online]. http://www.potaroo.net/ispcol/2010-09/exhaustguide.pdf

[193] IPv6 Deployment Monitoring. [Online]. http://www.ipv6monitoring.eu/

[194] Emile Aben. (2010, Jun) IPv6 Ripeness - the Sequel. [Online]. http://labs.ripe.net/Members/emileaben/content-ipv6-ripeness-sequel

[195] Geoff Huston. (2010, Apr) Measuring More IPv6. [Online]. http://www.potaroo.net/ispcol/2010-04/ipv6-measure.pdf

[196] Geoff Huston. (2009, Sep) The ISP Column: Is the Transition to IPv6 a "Market Failure?" [Online]. http://www.potaroo.net/ispcol/2009-09/v6trans.pdf

[197] Daniel J. Weitzner, "Net Neutrality. Seriously this Time," *IEEE Internet Computing*, pp. 86-89, May-Jun 2008.

[198] Lee W. McKnight William Lehr, "Show me the money: contracts and agents in service level agreement markets," *INFO*, vol. 4, no. 1, Jan 2002.

[199] Bill Norton. (2010, Aug) Internet Transit Prices - Historical and Projected. [Online]. http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php

[200] Renesys Blog, Bob Fletcher. (2010, Oct) Internet Transit Sales: 2005-10. [Online]. http://www.renesys.com/blog/2010/10/internet-transit-sales-2005-10.shtml

[201] Telegeography. (2010, Nov) IP transit prices continue their downward trend. [Online]. http://www.telegeography.com/cu/article.php?article_id=35206

[202] Flag Telecom Holdings Limited. (2000, Mar) 10-K for year ending 31-Dec-1999. [Online]. http://www.sec.gov/Archives/edgar/data/1102752/0000912057-00-015174.txt

[203] Renesys Blog, Todd Underwood. (2008, Nov) Will Work For Bandwidth. [Online]. http://www.renesys.com/blog/2008/11/will-work-for-bandwidth.shtml

[204] Renesys Blog, Earl Zmijewski. (2009, Dec) A Baker's Dozen in 2009. [Online]. http://www.renesys.com/blog/2009/12/a-bakers-dozen-in-2009.shtml

[205] Renesys Blog, Earl Zmijewski. (2008, Dec) Rising to the Top: A Baker's Dozen. [Online]. http://www.renesys.com/blog/2008/12/winners-and-losers-for-2008.shtml

[206] Sprint Nextel Corporation. (2010, Feb) Annual 10-K Filing for 2009. [Online]. http://investors.sprint.com/phoenix.zhtml?c=127149&p=irol-sec

[207] Bill Norton. (2008, Feb) Video Internet: The Next Wave of Massive Disruption to the U.S. Peering Ecosystem (v1.7). [Online]. http://drpeering.net/white-papers/Video-Internet-The-Next-Wave-Of-Massive-Disruption-To-The-U.S.-Peering-Ecosystem.html

[208] Bill Norton. (2010, Sep) A Business Case for Peering in 2010. [Online]. http://drpeering.net/white-papers/A-Business-Case-For-Peering.php

[209] Laura Blumenfeld. (2003, Jul) Washington Post: Dissertation Could Be Security Threat. [Online]. http://www.washingtonpost.com/ac2/wp-dyn/A23689-2003Jul7

[210] SAVVIS, Inc. (2010, May) Savvis 10-K Filing 2009. [Online]. http://www.savvis.net/en-US/Pages/Investors.aspx

[211] Cogent Communications Group. (2009, Dec) 10-K Filing. [Online]. http://www.cogentco.com/Reports/10k_Report.pdf

[212] AboveNet, Inc. (2010, Mar) AboveNet 10-K Filing 2009. [Online]. http://phx.corporate-ir.net/phoenix.zhtml?c=147513&p=irol-sec

[213] Neutral Tandem. (2010, Oct) Acquisition of Tinet. [Online]. http://www.neutraltandem.com/investorRelations/acquisition-details.htm

[214] NTT Communications. (2010, May) NTT Com Announces Financial Results for Fiscal Year Ended March 31, 2010. [Online]. http://www.ntt.com/aboutus_e/news/data/20100514.html

[215] China Telecom. (2010, Oct) http://www.chinatelecom-h.com/eng/ir/reports.php. [Online]. http://www.chinatelecom-h.com/eng/ir/reports.php

[216] Internet Engineering Steering Group. (1993, Sep) RFC 1517: Applicability Statement for the Implementation of Classless Inter-Domain Routing (CIDR). [Online]. http://www.rfc-editor.org/rfc/rfc1517.txt

[217] Y. Rekhter and T. Li. (1993, Sep) RFC 1518: An Architecture for IP Address Allocation with CIDR. [Online]. http://www.rfc-editor.org/rfc/rfc1518.txt

[218] V. Fuller, T. Li, J. Yu, and K. Varadhan. (1993, Sep) RFC 1519: Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy. [Online]. http://www.rfc-editor.org/rfc/rfc1519.txt

[219] V. Fuller and T. Li. (2006, Aug) RFC 4632: Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan. [Online]. http://www.rfc-editor.org/rfc/rfc4632.txt

[220] B. Briscoe, A. Odlyzko, and B. Tilly, "Metcalfe's law is wrong - communications networks increase in value as they add members-but by how much?," *IEEE Spectrum*, vol. 43, no. 7, pp. 26-31, July 2006.

[221] IPv4 Address Report. [Online]. http://www.potaroo.net/tools/ipv4/

[222] George A. Akerlof, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488-500, Aug 1970.

[223] Feng Qian et al., "TCP Revisited: A Fresh Look at TCP in the Wild," in *IMC'09*, Chicago, 2009.

[224] Akamai. (2001, Oct) Visualizing the Internet. [Online]. http://www.akamai.com/html/technology/visualizing_akamai.html

[225] Akamai. (2010, Oct) Facts & Figures. [Online]. http://www.akamai.com/html/about/facts_figures.html#6

[226] Level 3 Communications, Inc. (2009, Dec) Level 3 10-K Filling for 2009. [Online]. http://lvlt.client.shareholder.com/secfiling.cfm?filingID=1047469-10-1553

[227] Rajiv C. Shah and Jay P. Kesan, "The Privatization of the Internet's Backbone Network," *Journal of Broadcasting & Electronic Media, 51: 1, 93 — 109*, vol. 51, no. 1, pp. 93-109, 2007.

[228] ITU. (2005, Jul) IP Policy Manual. [Online]. http://www.itu.int/ITU-T/special-projects/ip-policy/

[229] Sprint Nextel Inc. (2010, Feb) Sprint 10-K Filing for 2009. [Online]. http://investors.sprint.com/phoenix.zhtml?c=127149&p=irol-sec

[230] Barry M. Leiner et al. (2010) A Brief History of the Internet. [Online]. http://www.isoc.org/internet/history/brief.shtml

[231] Katie Hafner and Matthew Lyon, *Where Wizards Stay Up Late, The Origins of the Internet.*: Simon & Schuster Ltd, 1996.

[232] Y. Rekhter and C. Topolcic. (1993, Sep) RFC 1520: Exchanging Routing Information Across Provider Boundaries in the CIDR Environment. [Online]. http://www.rfc-editor.org/rfc/rfc1520.txt

[233] C. Villamizar, R. Chandra, and R. Govindan. (1998, Nov) BGP Route Flap Damping. [Online]. http://www.ietf.org/rfc/rfc2439.txt

[234] NIST: Rick Kuhn, Kotikalapudi Sriram, Doug Montgomery. (2007, Jul) Border Gateway Protocol Security - Recommendations of the National Institute of Standards and Technology. [Online]. http://csrc.nist.gov/publications/nistpubs/800-54/SP800-54.pdf

[235] RIPE Routing Working Group: Philip Smith, Christian Panigl. (2006, May) Recommendations on Route-flap Damping. [Online]. http://www.ripe.net/ripe-378.txt

[236] UK Government, The Cabinet Office. (2010, Oct) Fact Sheet 18: Cyber Security. [Online]. http://download.cabinetoffice.gov.uk/sdsr/factsheet18-cyber-security.pdf

[237] Sun Newspaper (UK), Deputy Political Editor, Graeme Wilson. (2010, Oct) Fight Cyber War Before Planes Fall Out of Sky. [Online]. http://www.thesun.co.uk/sol/homepage/news/3186185/Security-chiefs-warn-Britain-must-protect-itself-against-cyber-warfare-amid-government-cuts.html

[238] Mathieu Feuillet, Alexandre Proutiere Thomas Bonald, "Is the "Law of the Jungle Sustainable for the Internet ?," in *INFOCOM'09*, 2009.

[239] Constantine Dovrolis, Ellen W. Zegura Ruomei Gao, "Avoiding Oscillations due to Intelligent Route Control Systems," in *INFOCOM'06*, 2006.

[240] Michalis Faloutsos Georges Siganos, "Analysing BGP Policies: Methodology and Tool," in *INFOCOM'04*, 2004.

[241] Tom Vest, Eliot Lear William Lehr, "Running on Empty: the Challenge of Managing Internet Addresses," in *TPRC*, Arlington, Virginia, 2008.

[242] Kartik Gopalan, Michael R. Hines, Aman Shaikh, Jacobus E. van der Merwe Nick Duffield, "Measurement Informed Route Selection," in *PAM*, 2007.

[243] Ramesh Govindan, George Varghese, Randy H. Katz Zhuoqing Morley Mao, "Route Flap Damping Exacerbates Internet Routing Convergence," , Pittsburg, Pennsylvania, 2002.

[244] V. Paxson J. Mahdavi. (1999, Sep) RFC2678: IPPM Metric for Measuring Connectivity. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc2678.txt.pdf

[245] G. Almes, J. Mahdavi, M. Mathis V. Paxson. (1998, May) RFC2330: Framework for IP Performance Metrics (IPPM). [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc2330.txt.pdf

[246] S. Kalidindi, M. Zekauskas G. Almes. (1999, Sep) RFC2679: A One Way Delay Metric for IPPM. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc2679.txt.pdf

[247] S. Kalidindi, M. Zekauskas G. Almes. (1999, Sep) RFC2680: A One Way Packet Losss Metric for IPPM. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc2680.txt.pdf

[248] S. Kalidindi, M. Zekauskas G. Almes. (1999, Sep) RFC2681: A Round Trip Delay Metric for IPPM. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc2681.txt.pdf

[249] R. Ravikanth R. Koodli. (2002, Aug) RFC3357: One-way Loss Pattern Sample Metrics. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc3357.txt.pdf

[250] P. Chimento C. Demichelis. (2002, Nov) RFC3393: IP Packet Delay Variation Metric for IPPM. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc3393.txt.pdf

[251] J. Ishac P. Chiment. (2008, Feb) RFC5136: Defining Network Capacity. [Online]. http://www.rfc-editor.org/rfc/pdfrfc/rfc5136.txt.pdf

Appendix I – Trivial Internet Global Routing Table

Figure 13, on page 49, shows a trivial internet with four ASes with three routers each. Each AS is home to a single address block. Every router has its version of the 'global routing table', which contains at least one route for each of the four address blocks. This table shows that table for all twelve routers.

Router: R1a			Router R1b			Router R1c		
Address Block	**AS Path**	**via**	**Address Block**	**AS Path**	**via**	**Address Block**	**AS Path**	**via**
10.0.1.0-10.0.1.255	*Local*	–	10.0.1.0-10.0.1.255	*Local*	–	10.0.1.0-10.0.1.255	*Local*	–
10.0.2.0-10.0.2.255	2	R2a	10.0.2.0-10.0.2.255	2	R1a	10.0.2.0-10.0.2.255 or	2 4, 2	R1a R4c
10.0.3.0-10.0.3.255	3	R1b	10.0.3.0-10.0.3.255	3	R3b	10.0.3.0-10.0.3.255 or	3 4, 3	R1b R4c
10.0.4.0-10.0.4.255 or	4 2, 4	R1c R2a	10.0.4.0-10.0.4.255 or	4 3, 4	R1c R3b	10.0.4.0-10.0.4.255	4	R4c

Router: R2a			Router R2b			Router R2c		
Address Block	**AS Path**	**via**	**Address Block**	**AS Path**	**via**	**Address Block**	**AS Path**	**via**
10.0.1.0-10.0.1.255	1	R1a	10.0.1.0-10.0.1.255 or	1 4, 1	R2a R4b	10.0.1.0-10.0.1.255	1	R2a
10.0.2.0-10.0.2.255	*Local*	–	10.0.2.0-10.0.2.255	*Local*	–	10.0.2.0-10.0.2.255	*Local*	–
10.0.3.0-10.0.3.255 or	1, 3 4, 3	R1a R2b	10.0.3.0-10.0.3.255 or	4, 3 1, 3	R4b R2a	10.0.3.0-10.0.3.255 or	1, 3 4, 3	R2a R2b
10.0.4.0-10.0.4.255 or	4 1, 4	R2b R1a	10.0.4.0-10.0.4.255	4	R4b	10.0.4.0-10.0.4.255	4	R2b

Router: R3a			Router R3b			Router R3c		
Address Block	**AS Path**	**via**	**Address Block**	**AS Path**	**via**	**Address Block**	**AS Path**	**via**
10.0.1.0-10.0.1.255 or	1 4, 1	R3b R4a	10.0.1.0-10.0.1.255	1	R1b	10.0.1.0-10.0.1.255	1	R3b
10.0.2.0-10.0.2.255 or	4, 2 1, 2	R4a R3b	10.0.2.0-10.0.2.255 or	1, 2 4, 2	R1b R3a	10.0.2.0-10.0.2.255 or	4, 2 1, 2	R3a R3b
10.0.3.0-10.0.3.255	*Local*	–	10.0.3.0-10.0.3.255	*Local*	–	10.0.3.0-10.0.3.255	*Local*	–
10.0.4.0-10.0.4.255	4	R4a	10.0.4.0-10.0.4.255 or	4 1, 4	R3a R1b	10.0.4.0-10.0.4.255	4	R3a

Router: R4a			Router R4b			Router R4c		
Address Block	**AS Path**	**via**	**Address Block**	**AS Path**	**via**	**Address Block**	**AS Path**	**via**
10.0.1.0-10.0.1.255 or	1 3, 1	R4c R3a	10.0.1.0-10.0.1.255 or	1 2, 1	R4c R2b	10.0.1.0-10.0.1.255	1	R1c
10.0.2.0-10.0.2.255	2	R4b	10.0.2.0-10.0.2.255	2	R2b	10.0.2.0-10.0.2.255 or	2 1, 2	R4b R1c
10.0.3.0-10.0.3.255	3	R3a	10.0.3.0-10.0.3.255	3	R4a	10.0.3.0-10.0.3.255 or	3 1, 3	R4a R1c
10.0.4.0-10.0.4.255	*Local*	–	10.0.4.0- 10.0.4.255r	*Local*	–	10.0.4.0-10.0.4.255	*Local*	–

Table 3 'Global Routing Tables' for Figure 13

Appendix II – Major Transit Provider Financials

This appendix is a brief analysis of the publically available data on the financial state of sixteen major transit providers, mostly Tier 1.

Table 4 shows the change in revenues from the previous year, for the segments which appear to cover the provision of transit. The table also shows the revenue for the segment for 2009, and the percentage of total revenue excluding any Wireless Telephone revenue that could be identified[56]. Also shown is which networks are believed to be Tier 1 and their ranking according to Renesys, as of July-2010 [**85**].

Changes in Segment Revenues 2007 to 2009							2009	
Type	T1	Company	*Segment*	2007	2008	2009	Revenue	% Total
Internet	1	Level 3	*Wholesale Markets*	*Note 1*	6.5%	-8.7%	$ 1,987.0	53.8%
	2	Global Crossing	*Enterprise, Carrier, …*	*Note 1*	21.0%	-0.6%	$ 2,159.0	85.1%
	5	Savvis	*Network Services*	-5.8%	-5.6%	-8.7%	$ 267.1	30.5%
	12	Cogent		24.5%	16.1%	9.4%	$ 235.8	100.0%
	-	Abovenet		7.1%	26.1%	12.6%	$ 360.1	100.0%
	10	Tinet		*Note 2*	30.8%	16.0%	$ 52.7	100.0%
US: ILEC	9	AT&T	*Data*	31.4%	5.3%	5.4%	$ 26,723.0	40.7%
	4	Verizon	*Global Wholesale*	*Note 1*	-3.6%	-7.0%	$ 9,637.0	20.0%
	13	Qwest	*Strategic*	4.3%	2.2%	-13.4%	$ 1,222.0	9.9%
			Legacy	-4.6%	-6.8%	-13.1%	$ 1,621.0	
US: Other	3	Sprint Nextel	*Internet*	37.8%	36.4%	6.8%	$ 2,293.0	40.7%
			Data	-15.5%	-20.7%	-31.0%	$ 662.0	
	-	XO	*Broadband*	0.8%	3.4%	3.0%	$ 798.3	52.5%
International	8	NTT	*IP Services*	*Note 2*	*Note 2*	3.0%	$ 4,364.5	33.8%
			Data	*Note 2*	*Note 2*	-9.8%	$ 1,437.2	
	6	TeliaSonera	*Broadband Services*	8.8%	1.0%	-3.4%	$ 6,492.1	39.8%
	7	Tata	*Enterprise & Carrier*	9.4%	-11.2%	18.3%	$ 298.1	41.1%
	11	China Telecom	*Internet*	34.1%	28.0%	26.6%	$ 7,777.8	24.6%
	-	Colt	*Wholesale Data*	7.2%	5.9%	5.4%	$ 150.9	12.8%
Note 1	*Change of segments makes comparison impossible.*						Millions of Dollars	
Note 2	*Data not located.*							

Table 4: Segment Revenues 2007 to 2009

Over the period we mostly see either increasing declines in revenue, or reducing increases. This is particularly marked in 2009, though 2009 was not a good year generally.

The companies are grouped very approximately, as follows:

- Internet: these are companies whose main business is providing Internet services, including transit. Results for these companies are more closely related to the health of the interconnection system.
- US: ILEC ('Incumbent Local Exchange Carrier'): these are companies for whom the provision of transit is a small part of their business.

[56] *This affects Verizon 56% of whose total revenues are Wireless, AT&T 45% and Sprint Nextel 83%.*

- US: Other: these are large US carriers offering a range of services, including telephone services, but who are not ILECs.
- International: significant non-US companies.

The rest of this appendix provides more detail for each of these providers.

Hard data about the costs associated with transit and peering are not available, nor are the revenues from transit. The available financial statements provide some information about revenues from business segments and the costs allocated to the business segments. Profits from a segment are less reliable because the costs of a segment are allocated costs. Revenue numbers are more reliable as they are directly related to the segment. Unfortunately the segments defined by the companies do not include "transit segments" and transit revenues may be included in more than one segment. Any conclusions drawn from this financial information must therefore be tentative.

II.1 Internet Companies

II.1.1 Level 3 Communications

Level 3 is seen to be the market leader in global transit. According to Renesys, [**85**] [**204**] [**205**], as of July 2010 Level 3 were the top global provider, 40% ahead of roughly equal second Sprint and Global Crossing – Renesys does not publish their scores, except to their customers, but the basis of the scoring is described in [**84**].

Level 3 has made more than 20 acquisitions since 1998, including: nearly all of Genuity[57] (2003), Wiltel Communications (Dec-2005), Progress Telecom (Mar-2006), ICG Communications (May-2006), Looking Glass Networks (Aug-2006), Broadwing Corporation (Jan-2007), Savvis's CDN (Jan-2007) and Servecast Limited (Jul-2007).

[57] *which was a spin off from GTE and included BBN – so Level 3 own AS 1.*

Level 3 splits its revenues into four Groups: Wholesale Markets; Business Markets; Content Markets and European Markets. The Wholesale Markets Group is the group that includes IP transit provision[58].

Level 3 Communications Inc					
Core Communications Services Revenue in Millions of Dollars					
	2005	**2006**	**2007**	**2008**	**2009**
Wholesale Markets Group			$ 2,045.0	$ 2,177.0	$ 1,987.0
Total	$ 1,645.0	$ 3,311.0	$ 4,199.0	$ 4,226.0	$ 3,695.0
Change in Wholesale Markets Group				6.5%	-8.7%
Change in Total		101.3%	26.8%	0.6%	-12.6%
Loss from continuing operations	$ (658.7)	$ (798.0)	$ (1,142.0)	$ (318.0)	$ (618.0)
Sources: 10-K Filings for 2009, 2008, 2007, 2006, and 2005					

Table 5: Level 3 Communications Inc., Core Communications Services Revenue 2005 to 2009

It is not possible to compare 2006 or 2005 Wholesale Markets Group revenues with those from 2007 onward as the company re-defined its groups in 2007. The losses total $3.5 Billion. The increase in revenues in 2006 and 2007 are from growth and acquisitions[59].

In their 10-K filing for 2009 [**210**] the management state (in Item 7):

> *"The Company believes that one of the largest sources of future incremental demand for the Company's Core Communications Services will be from customers that are seeking to distribute their feature rich content or video over the Internet. Revenue growth in this area is dependent on the continued increase in usage by both enterprises and consumers and the pricing environment. An increase in the reliability and security of information transmitted over the Internet and declines in the cost to transmit data have resulted in increased utilization of e-commerce or web based services by businesses. Although the pricing for data services is currently stable, the IP market is generally characterized by price compression and high unit growth rates depending upon the type of service. The Company continued to experience price compression in the high-speed IP market in 2009 and expects that pricing for its high-speed IP services will continue to decline in 2010."*

[58] *"The Wholesale Markets Group targets customers that include the largest national and global service providers, including carriers, cable companies, wireless companies, voice service providers, systems integrators and the federal government. These customers typically integrate Level 3 services into their own products and services to offer to their end user customers." Level 3 10-K filing for 2009 [**226**].*

[59] *"The 84% increase in Core Communications Services revenue for 2007 compared to 2006 is due to growth in the Company's revenue from existing services, as well as revenue from the Progress Telecom, ICG Communications, TelCove, Looking Glass, Broadwing, DN Business and Servecast acquisitions." Level 3 10-K filing for 2009 [**226**].*

II.1.2 Global Crossing Ltd.

Global Crossing was founded in 1997, entered Chapter 11 in Jan-2002 and emerged, restructured in Dec-2003.

The segment of Global Crossing Ltd. that includes transit provision is "carrier data" which is part of the "Enterprise, carrier data and indirect sales channel".

Global Crossing Ltd.					
Revenues in Millions of Dollars					
	2005	**2006**	**2007**	**2008**	**2009**
Enterprise, Carrier Data and Indirect Sales Channel	-	-	$ 1,794.0	$ 2,171.0	$ 2,159.0
Total Revenues	$ 1,968.0	$ 1,871.0	$ 2,265.0	$ 2,599.0	$ 2,536.0
Change in Enterprise, etc…	-	-	-	21.0%	-0.6%
Change in Total Revenues	-	-4.9%	20.8%	14.9%	-2.4%
Loss from continuing operations	$ (363.0)	$ (324.0)	$ (306.0)	$ (277.0)	$ (141.0)
Sources: 10-K Filings for 2009, 2008 and 2007					

Table 6: Global Crossing Ltd., Revenues 2005 to 2009

2007 saw the acquisition of Impsat Fiber Networks, Inc. which partially accounts for the increase in Total Revenues in 2007. The acquisition also led the company to redefine its business segments, so it is not possible to compare 2006 or 2005 Enterprise, carrier data and indirect sales channel revenues with those from 2007 onward.

In their 10-K filing for 2009 [**210**] the management state :

> *"We expect overall price erosion in our industry to continue at varying rates based on our service portfolio and reflective of marketplace demand and competition relative to existing capabilities and availability."*

and:

> *"Revenue attrition generally results from market dynamics and not customer dissatisfaction. Pricing for our VPN and managed services products has continued to decline at a relatively modest rate over the last few quarters, while pricing for specific data products such as high-speed transit and capacity services (specifically internet access arrangements used by content delivery and broadband service providers) has continued to decline at a greater rate."*

II.1.3 Savvis Inc.

Savvis Inc. provides Hosting, Co-location and Network Services. Savvis acquired C&W USA (via Chapter 11) in 2004. C&W USA had acquired MCI's Tier 1 Internet backbone in 1998 (when MCI merged with Worldcom the regulators would not allow the MCI and UUNet networks to be conjoined), and Exodus in 2001.

Savvis Inc					
Revenues in Millions of Dollars					
	2005	**2006**	**2007**	**2008**	**2009**
Colocation & Managed Hosting	-	$ 389.3	$ 474.6	$ 564.5	$ 607.3
Network Services	-	$ 328.9	$ 309.9	$ 292.5	$ 267.1
Change Colocation & Managed Hosting	-	-	21.9%	18.9%	7.6%
Change in Network Services Revenues	-	-	-5.8%	-5.6%	-8.7%
Income (Loss) from operations	$ (69.0)	$ (44.0)	$ (18.6)	$ (22.0)	$ (21.0)
Sources: 10-K Filings for 2009, 2008, 2007 and 2006					

Table 7: Savvis Inc., Revenues 2005 to 2009

The 2007 filing notes: "*The significant changes in 2007 reflect the impact of gains on sale of certain data center assets of $180.5 million in June 2007 and CDN assets of $125.2 million in January 2007 and the impact of the loss on debt extinguishment of $45.1 million in June 2007 related to our subordinated notes.*"

In their 10-K filing for 2009 [**210**] the management state (in Item 7):

> *"However, we have seen decreases from non−core, below−market margin customers, including certain of those in the internet content business, and certain of our network products, which we expect to continue to be under pressure throughout 2010."*

II.1.4 Cogent Communications Group.

Cogent started in 1999 and has acquired 13 other networks, including a number of the Internet pioneer companies: PSINet, NetRail and Aleron (originally AGIS/Net99), Allied Riser, OnSite Access, Fiber City, Fiber Network Solutions, Applied Theory, LambdaNet France and Spain, Carrier1, Unlimited Fiber Optics, Global Access and NTT/Verio.

Cogent Communications Group									
Service Revenues in Millions of Dollars									
	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
	$ 3.0	$ 51.9	$ 59.4	$ 91.3	$ 135.2	$ 149.1	$ 185.7	$ 215.5	$ 235.8
Change		1,620.1%	14.5%	53.6%	48.1%	10.2%	24.5%	16.1%	9.4%
Operating Income (Loss) before taxes in Millions of Dollars									
	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
	$ (61.1)	$ (62.3)	$ (81.2)	$ (84.1)	$ (62.1)	$ (46.6)	$ (29.9)	$ (22.2)	$ (3.8)
Sources: 10-K Filings, 2009 and 2006									

Table 8: Cogent Communications Group, Revenues & Income 2001 to 2009

Cogent is yet to make a profit, and is less than one tenth of the size of Level 3 in revenue terms. Revenue growth is apparently slowing, but so are the annual losses.

In their 2009 10-K filing [**211**], the management state (in Item 7):

> *"We believe two of the most important trends in our industry are the continued long-term growth in Internet traffic and a decline in Internet access prices within carrier neutral data centers[60]. As Internet traffic continues to grow and prices per unit of traffic continue to decline, we believe our ability to load our network and gain market share from less efficient network operators will continue to expand. However, continued erosion in Internet access prices will likely have a negative impact on the rate at which we can increase our revenues and our profitability."*

II.1.5 Abovenet Inc.

The current Abovenet was Metromedia Fiber Network (MFN), who acquired AboveNet Communications in Sep-1999. The parent company took the name AboveNet in Sep-2003 when it emerged from bankruptcy, and shifted the focus of the business to high−bandwidth solutions, primarily to enterprise customers (away from wholesale business).

AboveNet Inc.						
Revenues in Millions of Dollars						
	2004	**2005**	**2006**	**2007**	**2008**	**2009**
	$ 189.3	$ 219.7	$ 236.7	$ 253.6	$ 319.9	$ 360.1
Change		16.1%	7.7%	7.1%	26.1%	12.6%
Operating Income (Loss) before taxes in Millions of Dollars						
	2004	**2005**	**2006**	**2007**	**2008**	**2009**
	$ (35.4)	$ (12.2)	$ (3.5)	$ 3.2	$ 55.1	$ 94.9
Sources: 10-K Filings for 2009 and 2007						

Table 9: AboveNet Inc. Revenues and Income 2004 to 2009

Abovenet have become profitable in recent years, perhaps because of its focus on enterprise customers rather than wholesale customers.

In their 10-K filing for 2009 [**212**] the management state (in Item 1):

> *"The telecom industry is intensely competitive and has undergone significant consolidation over the past few years. Although there are multiple reasons for this consolidation, among the most prominent is the need to rationalize capacity created as a result of the telecommunications investment boom which occurred in the late 1990s. With respect to our larger competitors, Verizon and AT&T (formerly SBC) have accounted for most of the consolidation through their purchases of MCI and AT&T, respectively. In the mid−market, Level 3 was responsible for a significant portion of the consolidation by acquiring a large number of facilities−based telecommunications providers. At the same time, regulatory rulings have reduced the obligations of the ILECs to provide portions of their networks, referred to as unbundled network elements ("UNEs"), at historical cost prices making it more difficult for non−facilities−based operators to continue to provide services by utilizing UNEs from the ILECs."*

and:

> *"The Internet connectivity business is intensely competitive and includes many providers such as AT&T, Verizon, Level 3 and Cogent. As a result of this competition, while Internet traffic has continued to grow at a substantial rate over the past five years, pricing has generally declined, which has negatively affected revenue growth."*

[60] *By which they mean transit prices.*

II.1.6 Tinet

Tiscali International Network, the carrier arm of Tiscali Group, was acquired by BS Private Equity SpA in May-2009 (in an MBO worth €45.4 Million) and became Tinet. Tinet offered IP Transit and Ethernet services. Tinet was acquired in Sep-2010 by Neutral Tandem (for €74.5 Million) [**213**].

Tinet			
Revenue in Millions of Dollars	2007	2008	2009
Revenue	$ 34.8	$ 45.5	$ 52.7
Change in Revenue		30.8%	16.0%
Source: Neutral Tandem Aquisition Presentation, Oct-2010			

Table 10: Tinet Revenues 2007 to 2009

It is not known what proportion of these revenues is transit. It is claimed to be the 10th largest transit provider by volume in Oct-2010 – according to the Acquisition Presentation given by Neutral Tandem, which gives Renesys as the source.

II.2 US: ILEC

II.2.1 AT&T.

AT&T was acquired by SBC (Southwestern Bell Corporation) in Nov-2005, and SBC promptly adopted the name AT&T. AT&T acquired BellSouth at the end of 2006. AT&T splits its business into segments. The "Wireline"[61] segment includes: Voice, Data[62] and Other. "Wireline Data" includes IP transit revenues.

AT&T Inc					
Wireline Revenues in Millions of Dollars	2005	2006	2007	2008	2009
Data	$ 10,734.0	$ 18,317.0	$ 24,075.0	$ 25,353.0	$ 26,723.0
Change in Data Revenues		70.6%	31.4%	5.3%	5.4%
Total Wireline	$ 39,505.0	$ 57,473.0	$ 71,583.0	$ 69,855.0	$ 65,670.0
Change in Total Wireline		45.5%	24.6%	-2.4%	-6.0%
Source: 10-K Filings for 2010, 2008, 2007 and 2005					

Table 11: AT&T Inc., Revenues 2005 to 2009

The jumps in revenue in 2006 and 2007 are clearly associated with the acquisitions at the end of 2005 and the end of 2006. The increases in Data revenues in 2008 and 2009 are modest, but these do not reflect just Transit provision.

[61] *"The Wireline segment uses our regional, national and global network to provide consumer and business customers with landline voice and data communications services, AT&T U-verseSM TV, high-speed broadband and voice services (U-verse) and managed networking to business customers. Additionally, we offer satellite television services through our agency arrangements." - AT&T's 10-K Filing for 2009.*

[62] *Data is defined as: "Data includes traditional products, such as switched and dedicated transport, Internet access and network integration, and data equipment sales, and U–verse services. Additionally, data products include high–speed connections such as private lines, packet, dedicated Internet and enterprise networking services, as well as products such as DSL/broadband, dial–up Internet access and Wi–Fi (local radio frequency commonly known as wireless fidelity)." - AT&T's 10-K Filing for 2009.*

II.2.2 Verizon Communications Inc.

Verizon was formed when Bell Atlantic merged with GTE in Jun-2000. (Genuity, formerly BBN Planet, had been acquired by GTE in 1997 but was spun off prior to the merger with Bell Atlantic.) Verizon acquired MCI in Jan-2006, so Verizon Business is now the home of the one-time UUNet network.

Verizon splits revenues between Wireline and Wireless. Wireline Business covers sub-segments: Mass Markets; Global Enterprise[63]; Global Wholesale and Other. Global Wholesale covers IP transit.

Verizon Communications Inc.			
Wireline Businesses - Consolidated Revenues			
	2007	**2008**	**2009**
Global Wholesale	$ 10,750.0	$ 10,360.0	$ 9,637.0
Change in Global Wholesale		-3.6%	-7.0%
Total Wireline	$ 51,136.0	$ 50,222.0	$ 48,089.0
Change in Total Wireline		-1.8%	-4.2%
			Source: 10-K Filing, 2009

Table 12: Verizon Communications Inc., Revenues 2007 to 2009

A change in reporting categories in 2008 means that it is not possible to identify changes in revenues consistently earlier than 2007.

II.2.3 Qwest.

Qwest started life in 1996 running fibre cables beside railway tracks, for others and for themselves. In Jun-2000 they merged with US West (one of the Baby Bells). They were part owners of the ill-fated KPNQwest which crashed spectacularly in Jul-2002.

Qwest reports revenues for Wholesale Markets[64], segmented into: Strategic Services[65] and Legacy Services[66]. Transit revenues are likely to be in Strategic Services, but it is not entirely obvious from

[63] *Global Wholesale is described as follows in the 2009 10-K filing: "Global Wholesale revenues are primarily earned from long distance and other carriers who use our facilities to provide services to their customers. Switched access revenues are generated from fixed and usage-based charges paid by carriers for access to our local network, interexchange wholesale traffic sold in the U.S., as well as internationally destined traffic that originates in the U.S. Special access revenues are generated from carriers that buy dedicated local exchange capacity to support their private networks. Wholesale services also include local wholesale revenues from unbundled network elements and interconnection revenues from competitive local exchange carriers and wireless carriers. A portion of Global Wholesale revenues are generated by a few large telecommunication companies, many of whom compete directly with us."*

[64] "Wholesale Markets*: Our wholesale markets customers are other telecommunications carriers and resellers that purchase our products and services in large quantities to sell to their customers or that purchase our access services that allow them to connect their customers and their networks to our network." Qwest 2009 10K filing.*

[65] *"Nearly all of the strategic services revenue we generate from wholesale markets customers is from private line services. Our wholesale customers use our private line services to connect their customers and their networks to our network. We also provide private line services to wireless service providers that use our fiber-optic services to support their next generation wireless networks." Qwest 2009 10K filing.*

[66] *"Our wholesale markets legacy services include long-distance, access, local and traditional WAN services. Local services include primarily unbundled network elements, or UNEs, which allow our wholesale customers to use our network or a combination of our network and their own networks to provide voice and data services to their customers. Our local services also include network transport, billing services and access to our network by other telecommunications*

the descriptions. Qwest's Wholesale Markets Revenues declined each year from 2006 to 2009 at an increasing rate.

Qwest				
Wholesale Markets Revenues in Millions of Dollars				
	2006	**2007**	**2008**	**2009**
Strategic services	$ 1,323.0	$ 1,380.0	$ 1,411.0	$ 1,222.0
Legacy services	$ 2,354.0	$ 2,129.0	$ 1,860.0	$ 1,621.0
Total Wholesale Markets	$ 3,677.0	$ 3,509.0	$ 3,271.0	$ 2,843.0
Change in Strategic Services		4.3%	2.2%	-13.4%
Change in Legacy Services		-9.6%	-12.6%	-12.8%
Change in Total Wholesale Markets		-4.6%	-6.8%	-13.1%
Sources: 10-K Filings for 2009, 2008 and 2007				

Table 13: Qwest Communications International Inc., Revenues 2006 to 2009

Although they are a Tier 1 transit provider, their business appears to be focused on their local area end-user services.

providers and wireless carriers. These services allow other telecommunications companies to provide telecommunications services that originate or terminate on our network. Long-distance services include domestic and international long-distance services.

Access services include fees that we charge to other telecommunications providers to connect their customers and their networks to our network so that they can provide long-distance, transport, data, wireless and Internet services." Qwest 2009 10-K filing.

II.3 US: Other

II.3.1 Sprint Nextel Corp

Sprint is one of the original commercial Internet 'backbones'.

Sprint Nextel Corporation splits its operations between Wireline and Wireless and is predominantly Wireless.[67]. Wireline Operating Revenues are shown for four categories: Voice; Data; Internet[68] and Other.

Sprint Nextel Corp					
Wireline "Operating Revenues" in Millions of Dollars					
	2005	**2006**	**2007**	**2008**	**2009**
Data	$ 1,620.0	$ 1,432.0	$ 1,210.0	$ 959.0	$ 662.0
Internet	$ 829.0	$ 1,143.0	$ 1,575.0	$ 2,148.0	$ 2,293.0
Total	$ 6,818.0	$ 6,560.0	$ 6,463.0	$ 6,332.0	$ 5,629.0
Change in Data		-11.6%	-15.5%	-20.7%	-31.0%
Change in Internet		37.9%	37.8%	36.4%	6.8%
Change in Data and Internet		5.1%	8.2%	11.6%	-4.9%
Change in Total		-3.8%	-1.5%	-2.0%	-11.1%
Source: 10-K Filings for 2009, 2008, 2007 and 2006					

Table 14: Sprint Nextel Corp., Revenues 2005 to 2009

Total Wireline "Operating Revenues" fell each year between 2005 and 2009, driven by the fall in "traditional" Data revenues, which is partly offset by subscribers moving to IP-based services.

In their 10-K filing for 2009 [**206**] the management state (in Item 1):

> *"Some competitors are targeting the high-end data market and are offering deeply discounted rates in exchange for high-volume traffic as they attempt to utilize excess capacity in their networks."*

[67] *Wireline represented a small portion of the company's revenues in 2009 with Wireless $27.8; Wireline $5.6 Billion.*

[68] *Internet Revenues are described as follows in the 10-K filing for the period to 31-Jan-2009 [**229**]: "Internet revenues reflect sales of IP-based data services, including MPLS. Internet revenues increased 38% in 2007 as compared to 2006 and increased 38% in 2006 as compared to 2005. The increases were due to higher IP revenues as business customers increasingly migrate to MPLS services, as well as revenue growth in our cable VoIP business, which experienced an 80% increase in 2007 as compared to 2006 and a 127% increase in 2006 as compared to 2005."*

II.3.2 XO Holdings Inc.

XO is the result of the merger in Jun-2000 of Nextlink Communications and Concentric Network. XO emerged from Chapter 11 in Jan-2003. It acquired Allegiance Telecom in Jun-2004, also through Chapter 11. XO is a Competitive Local Exchange Carrier (CLEC) as well as an ISP.

XO Holdings Inc.								
Revenues in Millions of Dollars								
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Total	$ 1,259.0	$ 1,110.0	$ 1,300.0	$ 1,437.0	$ 1,416.0	$ 1,428.0	$ 1,477.0	$ 1,521.0
Change		-11.8%	17.1%	10.5%	-1.5%	0.8%	3.4%	3.0%
Broadband						$ 530.3	$ 670.6	$ 798.3
Data & IP				$ 384.4	$ 426.6	$ 527.1		
Data Services	$ 472.2	$ 392.7	$ 414.7	$ 432.4				
Operating Income (Loss) before taxes in Millions of Dollars								
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
	$ (1,208.8)	$ (111.8)	$ (370.0)	$ (128.9)	$ (113.7)	$ (110.1)	$ (84.8)	$ (47.2)
Sources: 10-K filings for 2009, 2007, 2005 and 2003								

Table 15: XO Revenue and Income 2002 to 2009

The segmentation of the accounts has evolved over the years. The "Broadband", "Data &IP" and "Data Services" segments probably cover the provision of transit, however exactly how they relate to each other is not known. These figures also include a lot of other IP services, including end-user broadband connections.

II.4 International

II.4.1 NTT Communications

NTT Com is a subsidiary of NTT (Nippon Telegraph and Telephone Corporation). Data was located for 2009 and 2010 [**214**].

NTT Communications			
Business Results (Non-Consolidated Operating Revenues) in Millions of Dollars at 83.6 Yen to the US Dollar	**Mar-2009**	**Mar-2010**	**Change**
Voice Transmission Services (excl. IP services)	$ 4,950.1	$ 4,532.8	-8.4%
IP Services	$ 4,235.4	$ 4,364.5	3.0%
Open Computer Network Services	$ 1,878.9	$ 1,951.2	3.8%
IP-Virtual Private Network Services	$ 932.5	$ 934.1	0.2%
Wide-Area Ethernet services	$ 666.2	$ 689.9	3.6%
Other	$ 757.8	$ 789.2	4.1%
Data Communications (excl. IP Services)	$ 1,593.5	$ 1,437.2	-9.8%
Leased circuit services	$ 1,148.0	$ 1,059.5	-7.7%
Other	$ 445.5	$ 377.8	-15.2%
Solution Services	$ 2,357.4	$ 2,231.3	-5.3%
Others	$ 346.8	$ 344.0	-0.8%
Total operating revenues	$ 13,483.1	$ 12,909.9	-4.3%
Source: Financial Results for Fiscal Year Ended March 31, 2010			

Table 16: NTT Communications Revenues 2009 and 2010.

It's not clear where NTT Global IP Transit revenues fit. It could be among the "IP Services: Other", or it could be part of "Data Communications: Other", it depends on whether the "IP Services" category includes all IP activity, or just the end-user service.

II.4.2 TeliaSonera

TeliaSonera International Carrier runs the Telia IP transit network, and its revenues are counted under "Broadband Services"[69].

TeliaSonera				
Revenue in Millions of US Dollars at 6.69 SEK to the Dollar	**2006**	**2007**	**2008**	**2009**
Broadband Services Revenues	$ 6,110.6	$ 6,648.4	$ 6,717.9	$ 6,492.1
Change in Broadband Revenues		8.8%	1.0%	-3.4%
Source: 2009 Report and Accounts				

Table 17: TeliaSonera Broadband Services Revenues 2006 to 2009

[69] *To quote the 2009 Accounts, "Business area Broadband Services provides mass-market services for connecting homes and offices. Services include broadband over copper, fiber and cable, IPTV, voice over internet, home communications services, IP-VPN/Business internet, leased lines and traditional telephony. The business area operates the group common core network, including the data network of the international carrier business, and comprises operations in Sweden, Finland, Norway, Denmark, Lithuania, Latvia (49 percent), Estonia and international carrier operations."*

II.4.3 Tata Communications Limited

Tata Communications are the current owners of the one-time Tyco Global Network and of Teleglobe.

Tata Communications Limited					
Revenues from Telecommunications and Other Services in Millions of Dollars at $0.02254 Dollars to the Rupee	2006	2007	2008	2009	2010
Enterprise and Carrier Data	$ 284.4	$ 311.2	$ 276.4	$ 326.9	$ 298.1
Total	$ 852.2	$ 890.7	$ 740.1	$ 845.1	$ 725.3
Change in Enterprise and Carrier Data		9.4%	-11.2%	18.3%	-8.8%
Change in Total		4.5%	-16.9%	14.2%	-14.2%
Source: http://www.nseindia.com/marketinfo/companyinfo/eod/corp_res.jsp?symbol=TATACOMM					

Table 18: Tata Communications Limited Revenues 2006 to 2010

No significant conclusions can be drawn from this data although Tata are experiencing declining revenues in their Data and Network Services businesses.

II.4.4 China Telecom

China Telecom Corporation Limited provides basic telecommunications services such as wireline telecommunications services and mobile telecommunications services, and value-added telecommunications services such as Internet access services and information services in the PRC.

China Telecom Corporation Limited								
Revenues in Millions of Dollars at $0.1508 to the CNY	2002	2003	2004	2005	2006	2007	2008	2009
Total	$ 16,525.5	$ 22,858.7	$ 24,315.5	$ 25,537.0	$ 26,488.1	$ 27,282.4	$ 28,134.1	$ 31,579.2
Change in Total	-	38.3%	6.4%	5.0%	3.7%	3.0%	3.1%	12.2%
Internet	$ 741.2	$ 1,509.4	$ 2,128.1	$ 2,694.1	$ 3,578.3	$ 4,798.9	$ 6,142.8	$ 7,777.8
Change in Internet	-	103.6%	41.0%	26.6%	32.8%	34.1%	28.0%	26.6%
Sources: 2009, 2008, 2007, 2006, 2005, 2004 and 2003 Annual Accounts								

Table 19: China Telecom Corporation Limited Revenues 2002 to 2009

The Internet segment covers "*amounts charged to customers for the provision of Internet access services*" [**215**], which may include the provision of transit.

II.4.5 Colt Telecom Group S.A.

Colt started in 1992 as an alternative carrier in the City of London (hence the name). They now offer a mix of business and carrier services across Europe.

COLT Telecom Group SA								
Revenues in Millions of Dollars at €0.7295 to the Dollar	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Wholesale Data	-	-	-	*Note 1*	$ 126.1	$ 135.2	$ 143.2	$ 150.9
Change in Wholesale Data						7.2%	5.9%	5.4%
Total	$ 1,096.0	$ 1,244.4	$ 1,300.2	$ 1,328.9	$ 1,313.8	$ 1,225.3	$ 1,222.2	$ 1,183.6
Change in Total		13.5%	4.5%	2.2%	-1.1%	-6.7%	-0.3%	-3.2%
Operating Income (Loss) before taxes in Millions of Dollars								
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
	$ (203.1)	$ (84.8)	$ (68.6)	$ (57.7)	$ 18.0	$ 40.3	$ 55.7	$ 63.0
Note 1	*Prior to Jun-2006 the group was COLT Telecom Group plc. Segment reporting changed for the 2006 accounts.*							
	Figures from the 2002 to 2004 accounts converted to Euros at €1.4626 to the Pound used in the 2005 accounts.							
	Sources: 2009, 2008, 2007, 2006, 2005 and 2004 Annual Accounts							

Table 20: Colt Telecom Group S.A. 2002 to 2010

Colt are a Tier 2 transit provider.