1
Introduction

The Internet has irrevocably invaded many aspects of daily life. What was once an obscure scientific research tool has blossomed into a communications platform used by hundreds of millions of people. Telecom providers use the Internet to carry critical voice communications. Banking institutions use it to provide access to account services and market trading. Airline tickets, hotel reservations, and car rentals can all be booked with a click of the mouse. Whole industries have sprung into existence with business models that depend entirely on Internet infrastructure to reach their customers. More users than ever depend on the Internet to connect with family and colleagues using email, instant messaging, Voice over IP, photo and video sharing services, and online journals. Children born in the last ten years have grown up in a time when the Internet has always been available.

This point of view is popular among Internet users, but it does not necessarily reflect the experience of all, or even most, of the rest of the world. According to the ITU*, more than half of the users on the Internet are concentrated in the G8 countries (Canada, France, Germany, Italy, Japan, Russia, the UK, and the US). In 2004, less than 3% of Africans used the Internet, compared with an average of one 50% of the inhabitants of the G8 countries. The entire African continent accounts for about 13% of the total world population, yet in 2004 it had fewer Internet users than France alone.

Fortunately, in places where the Internet has not yet penetrated, it is all but certain to happen soon. There is a global push to bridge the so-called digital divide by bringing modern telecommunications to the developing world. State and private investment in public infrastructure, in the form of fibre optic backbones, wireless networks, and satellite connectivity are bringing the Internet to the most remote locations at a pace that is accelerating over time. People all

* Source: http://www.itu.int/ITU-D/ict/statistics/ict/

over the globe are beginning to realise that in order to effectively participate in the global marketplace, they need access to the global communications network.

But superhighways aren't built overnight. As with any major undertaking to build infrastructure, extending fast network connections to all of the ends of the earth takes time. Technologies such as VSAT make it possible to install an Internet connection just about anywhere, particularly in the absence of existing wired infrastructure. While this does extend the footprint of the Internet to otherwise unreachable places, the capacity of the connection provided is far from infinite. The cost of these connections is also quite high for many organisations. This often leads to the practice of stretching an insufficient network connection to serve many users simultaneously.

Bandwidth, throughput, latency, and speed

There are a few technical words used to describe how fast an Internet connection may go. Users often find these terms confusing, so it's best to be clear about their definitions from the beginning.

- ***Bandwidth*** refers to a measure of frequency ranges, typically used for digital communications. The "band" part of broadband is short for bandwidth, meaning that the device uses a relatively wide range of frequencies. In recent years, the term bandwidth has been popularly used to refer to the ***capacity*** of a digital communications line, typically measured in some number of bits per second. In its popular usage, you might read that a T1 provides a theoretical maximum "bandwidth" of 1.544 Mbps.

  While some purists insist that we should speak of capacity when talking about data transfer speeds and bandwidth when talking about frequency ranges, the popular usage of the term "bandwidth" has been reinforced by years of product marketing and misleading documentation. There simply is no going back now. Therefore, we will use the terms bandwidth and capacity interchangeably in this book.

- ***Throughput*** describes the actual amount of information flowing through a connection, disregarding protocol overhead. Like bandwidth, it is expressed in some number of bits per second. While a T1 may provide 1.544 Mbps between the endpoints, the protocol spoken on the physical line reduces the effective throughput to about 1.3 Mbps. When you factor in the additional overhead of Internet protocols, the available throughput is even less. When you measure the actual usage of a connection or perform a "speed test" on a line, you are measuring throughput.

- ***Latency*** refers to the amount of time it takes for a packet to travel from one point on a network to another. A closely related concept is ***Round Trip Time*** (***RTT***), which is the amount of time it takes for a packet to be acknowledged from the remote end of a connection. Latency is measured as some amount of time, usually in milliseconds. The latency of Ethernet is about 0.3 ms. A T1 connection has a latency of 2 to 5 ms, while a VSAT connection requires at least 500 ms before an acknowledgment can be received, due to the speed of light and the large distances involved. Some factors that contribute to latency are network congestion, overutilised servers, and the distance between the two points.

- ***Speed*** is an ambiguous term that refers to some combination of these other terms. An Internet connection may "feel slow" when using an interactive service (such as Voice over IP or gaming) on a line with high latency, even if there is sufficient bandwidth. Users will also complain when transferring large files on a connection with insufficient capacity, even if the latency is very low.

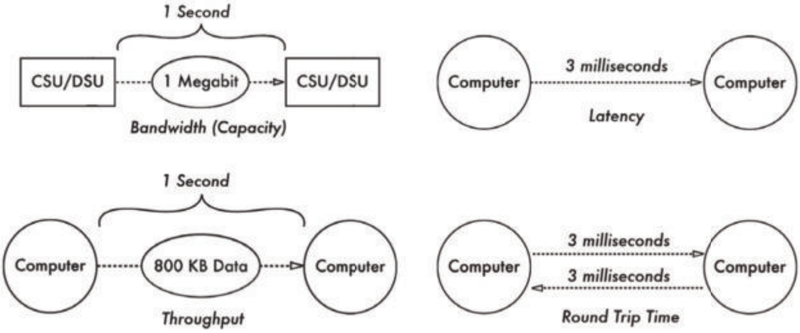

Figure 1.1: Bandwidth, Capacity, Throughput, Latency, and Round Trip Time.

The goal of this book is to show you how to optimise your Internet connection so that it provides the greatest possible throughput and lowest possible latency. By eliminating wasted bandwidth, the cost of operating your network connection will be reduced, and the usability of the network will be improved.

Not enough to go around

What actually causes a slow Internet connection? Obviously, the capacity of a given connection is finite, so if too many people request information at once, then someone will have to wait. In an ideal world, organisations would simply order more bandwidth to accommodate the increased traffic. But as we all know, Internet access costs money, and most organisations do not have infinite budgets.

It is an interesting fact of online life that users tend to consume more bandwidth over time. It is very rare to find a user who, once they have had access to a broadband connection, is satisfied with going back to a low speed dialup line. As users are exposed to Internet services earlier in life and in a variety of venues (for example at home, at work, at University, or at a cyber-cafe), they be-come accustomed to using it in a certain way. They are increasingly unlikely to know or care about the bandwidth required to listen to Internet radio, or to download the latest video game, or to watch funny movies on a video sharing service. They "just want it to work," and may complain when the Internet "is slow." Users often have no idea that they can single-handedly bring an organisation's Internet connection to a halt by running a simple file sharing program on their computer.

User education is obviously critical to every stage of implementing a plan to manage your bandwidth. While users can be forced to adhere to certain behaviour patterns, it is always far easier to implement a plan with their voluntary compliance. But how does such a plan come into being? If you simply order people to change their behaviour, little is likely to change. If you install technical hurdles to try to force them to change, they will simply find a way around the obstacles.

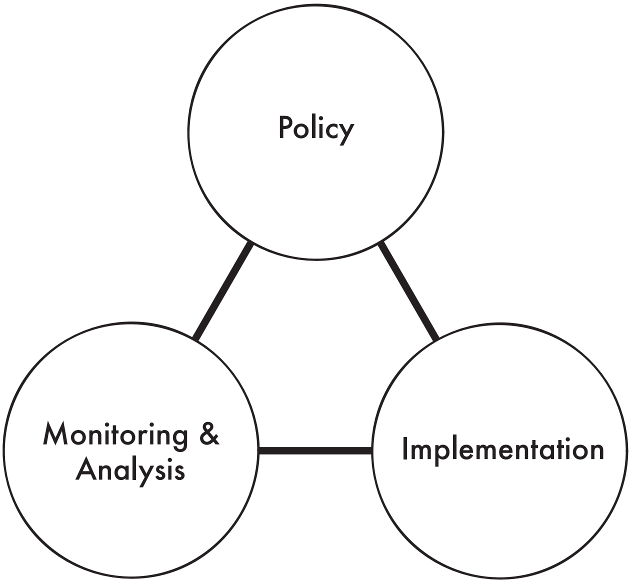

Figure 1.2: Policy, Monitoring & Analysis, and Implementation are three critical (and interdependent) components of bandwidth management.

In order to effectively manage a network connection of any size, you will need to take a multifaceted approach that includes effective ***network monitoring***, a sensible ***policy*** that defines acceptable behaviour, and a solid ***implementation*** that enforces these rules. Each component is important for effective bandwidth management in any network that consists of more than a few users. This book includes chapters devoted to each of these three major areas.

A **policy** is a statement of opinions, intentions, actions and procedures that guide the overall use of the network. An **acceptable use policy** is a subset of this, setting out in technical detail what uses of the network are believed by the network operators to be acceptable, and what they intend to do to anyone who uses it in a manner that they consider unacceptable. It should be a written document that defines acceptable forms of network access, as well as guidelines for how network problems are dealt with, definitions of abuse, and other operational details. The policy also typically includes definitions of legal constraints for network users (such as the exchange of copyrighted material, requesting inappropriate materials, etc.). Having a policy makes it much easier to enforce certain types of network behaviour, as you will be able to hold people to a set of agreed rules.

Network monitoring is the ongoing process of collecting information about various aspects of your network operations. By carefully analysing this data, you can identify faults, find cases of waste and unauthorised access, and spot trends that may indicate future problems.

Implementation is the step of implementing traffic shaping, filtering, caching, and other technologies within your network to help bring actual usage in line with policy. The actions you need to take are indicated by the data collected through monitoring and analysis, and are constrained by the network policy. Many people expect to begin the task of bandwidth management by starting with this step. But without good monitoring techniques, you are effectively blind to the problem. Without a policy, your users will not understand what you are doing or why, and will complain or subvert your actions instead of helping you to achieve your goal.

Don't underestimate the value of personally interacting with your network users, even at a very large institution. At Carnegie Mellon University (page **248**), social interactions made a far greater impact on bandwidth consumption than did technical constraints. But at an organisation as large as CMU, personal attention could only have had this effect by operating within a well-defined policy, with the support of a good network implementation and watched by careful network monitoring.

Where to begin

Effective bandwidth management can only happen by applying a combination of technical computer skills, effective network monitoring, and a sensible policy that is understood by all users. If your organisation has a small network, one person may need to work on all of these areas. Larger organisations will likely require a team of people to effectively manage busy networks, with each person specialising in a particular area.

This book is designed to be used as both a guide and a reference to anyone who needs to tackle this difficult problem. While you may read it cover-to-cover, each chapter is designed to stand on its own and address a particular aspect of bandwidth management. If you don't know where to begin, these guidelines should help you find a good starting place.

Do you need to fix your network immediately?

- Is something wrong with your computers or Internet access?
- Do the problems get in the way of people getting legitimate work done?
- Is your job at risk if you don't do something now?

If you answered **yes** to any of these questions, go to the **Troubleshooting** chapter (page **159**). When you've solved the immediate problem, continue with the steps below.

Do you know what's happening on your network?

- Do you monitor your network?
- Do you know what your bandwidth usage is, on average?
- Do you know who is using your bandwidth?
- Do you know how your bandwidth is being used? How much bandwidth is used for email, as compared to web traffic and peer-to-peer applications?
- Do you know about network outages before your users complain?
- Are you certain that your network only being used for appropriate services, and has not been compromised by a malicious user?

If you answered **no** to any of these questions, take a look at the **Monitoring & Analysis** chapter on page **25**. When you have a clear idea of what's happening on your network, continue with the steps below.

Do you want to change how users behave on your network?

- Is inappropriate user behaviour (e.g. peer-to-peer file sharing or excessive downloads) causing problems on your network?
- Do you need to create a written policy on network usage?
- Do you need to update an existing policy?
- Are your users largely unaware of what the network policy is, and why it is important?
- Do you need to guarantee the availability of certain services on your network?

If you answered **yes** to any of these questions, you will want to start with the **Policy** chapter (page **9**). When you have established a policy, please continue with the steps below.

Are you using basic optimisation techniques?

- Do you operate your network without a site-wide web cache?
- Do responses to DNS requests seem sluggish?
- Are spam and viruses wasting a significant amount of your bandwidth?
- Do your users make extensive use web mail services, such as Hotmail or Yahoo! Mail?

If you answered **yes** to any of these questions, you should start with the **Implementation** chapter on page **101**. Please be aware that technical solutions, while important, are unlikely to help unless you already have a well-defined and well-known network usage policy, and have already implemented good network monitoring.

Do you need to enforce further technical constraints on the network?

- Do you need to reduce the bandwidth used by certain services?
- Do you need to guarantee bandwidth for certain services (such as email) at the expense of others (such as web browsing)?
- Do you need to block some kinds of traffic entirely?
- Are some users able to monopolise the available bandwidth, effectively blocking access for all other users?
- Does your network usage exceed the available capacity of a single line, requiring you to make use of multiple Internet connections?

If you answered **yes** to any of these questions, you will want to start with the **Performance Tuning** chapter on page **177**. These steps should only be taken after basic optimisation methods have been implemented.

Do you need to convince someone else of the importance of bandwidth management?

Go to the **Case Studies** chapter (page **235**) to see examples of how bandwidth management is used in real organisations.

Do you want to know how to reduce your personal bandwidth use?

See the **General Good Practices** section on page **105**.