Efficient Network Routeing

Richard Mortier and Ian Pratt

{richard.mortier,ian.pratt}@cl.cam.ac.uk

Systems Research Group,
University of Cambridge Computer Laboratory,
New Museums Site,
Pembroke Street,
Cambridge, UK. CB2 3QG.

Tel: +44 1223 334600.
Fax: +44 1223 334678.

October 2000

Abstract

We believe that the control currently provided by routeing in the Internet is becoming demonstrably insufficient for the demands of network operators and users. Current routeing protocols provide little or no support for automation of network management, leaving such things as construction and implementation of SLAs and the pricing of resources entirely to the network manager. The research proposed in this document intends to attack the problem of efficient IP network management via the application of pricing to network resource allocation. Our goal is to devise schemes for the automatic negotiation, selection, and monitoring of SLAs within high-level policy descriptions provided by the network manager. We then aim to investigate how these algorithms can be implemented within the framework of existing IP routeing protocols such as BGP and OSPF.

Introduction

We begin by describing the current structure of the Internet in more detail, including a brief description of extant routeing protocols. We continue by describing existing and developing problems, and then listing the aims of the proposed research. A description of our view of the solution space follows, with some background on our suitability to carry out this research. We conclude with a proposed project plan, including a programme of work, plans for project management, and funding and intellectual property arrangements.

The structure of the network

Currently the Internet functions as an ad-hoc collection of IP (INTERNET PROTOCOL) networks, structured as a collection of ASs (AUTONOMOUS SYSTEMS). ISPs (INTERNET SERVICE PROVIDERS) operate small groups of ASs, peering with one another to provide end-to-end connectivity for users. ASs can loosely be divided into two types: *transit* and *stub*. Stub ASs form the edges of the network, where users connect to the network, and where content providers and server farms are situated. Transit ASs carry traffic between other ASs, providing the network's essential connectivity. IP traffic is accepted into a provider's stub AS, routed through that AS, and then off-loaded into another AS. This may either be the destination stub AS, or a transit AS; in either case the traffic is routed, respectively to its destination host, or to a further AS where the process repeats. By this means traffic reaches its destination. Peering arrangements between ASs are defined by SLAs (SERVICE LEVEL AGREEMENTS), legal documents which specify such details as the amount of traffic to be carried, the cost of the agreement, the routes that will be advertised to other carriers, and the protocols which will be used to advertise these routes.

The routeing protocols

There are two forms of routeing in the Internet, *intra*-AS routeing, and *inter*-AS routeing. The former deals with routeing traffic *within* an AS and is typically carried out by the OSPF (OPEN SHORTEST PATH FIRST) protocol. The latter deals with routeing traffic *between* ASs and is typically carried out by the BGP (BORDER GATEWAY PROTOCOL) protocol. In both cases routeing is based solely on the destination IP address contained within each

packet; although IP supports a per-packet 'type-of-service' designation, this is only available as a hint to the routeing protocol, and gives no guarantee of differentiated service.

Existing and developing problems

There are two fundamental problems which we intend to address with this research, both of which pertain to the general unmanageability of the Internet from the point of view of the network operator. The first is the difficulty with which different values can be associated with different traffic, and the second is the difficulty with which SLAs can be specified and maintained.

The value of a bit

It seems clear that different users associate different values with their traffic, dependent on a variety of parameters, such as time of day, state of the network, and the application generating the traffic. There are currently very few ways within the Internet framework to express these different valuations to the network, in order that the traffic may be appropriately treated. Although the IETF (INTERNET ENGINEERING TASK FORCE) has attempted to address this problem through the Integrated Services [1], and more recently the Differentiated Services [2, 3] communities, neither sets of proposals are entirely satisfactory.

The Integrated Services proposals suffer from the fact that they require the deployment of a completely new signalling protocol throughout the Internet before improvement in service can be seen. Furthermore, they address the provisioning of quality of service from the point of view of providing various types of guaranteed bandwidth for individual flows; it is widely held that such an approach is not sufficiently scalable in a network the size of the Internet.

The Differentiated Services proposals address the problem of scalability by defining PHBs (PER-HOP BEHAVIOURS) for IP packets, allowing a given AS to identify aggregates of traffic as requiring certain transport properties. For example, there are two PHBs currently defined in addition to the default *best effort* service: *assured forwarding* [4], and *expedited forwarding* [5]. These attempt to give guarantees on drop probability and bandwidth respectively.

Whilst these proposals address the mechanisms for requesting differentiated

service from the network, little progress has so far been made on methods to express policy to the network. The COPS (COMMON OPEN POLICY SERVICE) proposals [6, 7] attempt to do so for Integrated Services, but there has been little equivalent effort for Differentiated Services. In summary, current protocols do not allow for the route traffic will take to be specified in terms of the quality of service the traffic will receive.

Management of service level agreements

SLAs are bi-lateral agreements between network operators concerning the treatment one operator's traffic will receive when carried by the other operator, and *vice versa*. Since they must be specified and managed by intervention of the operator, they have a high associated management cost. This acts as a disincentive to operators entering into many SLAs at one time, and to modifying the SLAs into which they have entered. This disincentive is at odds with the more usual behaviour of a network, that its utility increases with its connectivity. Furthermore, it leads to inefficiencies in the distribution of traffic through the network, since peering arrangements cannot change at timescales short enough to diminish localised congestion.

There are other motivations for the pricing of SLAs, in addition to the associated management costs. Traffic will flow between the operators' networks, increasing the potential for congestion. The transitory nature of congestion means that the harm caused to an operator's network by transiting traffic for another operator is related, not only to the volume or volume-distance product of traffic carried, but to other factors such as the type of traffic, or the times at which traffic is carried. Additionally, settlement should be based on more than simple 'sender pays' since the utility of the traffic carried relates to more than simple reception. For example, shortening the response times experienced by interactive web traffic might be considered to be more valuable to users than that doing so for bulk data transfer traffic. Similarly, the modem banks through which users connect to the network do not source a significant amount of traffic (generating mostly the HTTP GET requests), but do cause a large amount of traffic to be generated and sunk (the data forming the responses to those requests). Within the responses, different data can also have different values, most obviously to users, but also to network operators. Traffic sourced from popular web sites might be considered more valuable than other traffic since carrying it can make the operator a more desirable peering partner.

Multi-protocol label switching

MPLS (MULTI-PROTOCOL LABEL SWITCHING) [8] is a technology enabling operators to fix the paths that packets belonging to particular sets of flows take across their network. This allows them greater scope for traffic management within their own networks, enabling traffic to be spread across multiple paths, with potentially different qualities of service. Although a useful tool, MPLS does not directly solve the problem of selecting routes for paths and monitoring their quality of service.

Work to allow MPLS paths to cross multiple ASs is still in its infancy, but could in principle allow a carrier to build paths above the 'default' BGP routed network, giving it greater control over how traffic is routed to and from other ASs. Deployment of such a scheme would clearly require renegotiation of many of the SLAs between ISPs. Consequently, a mechanism for more dynamic negotiation of SLAs is required before such a scheme could achieve its full potential.

Aims of the research

The research we propose intends to address the two principle problems identified above: the difficulty of expressing policy for differing treatment of traffic, and the difficulties associated with the specification and management of peering arrangements. We believe that the addressing these problems will manifest in a variety of ways. Apart from allowing automated generation of SLAs and automating the associated settlement process, it will allow users to better specify the quality of service they wish their traffic to receive.

More radically, it should encourage richer peering between ISPs, as the disincentive to rich peering currently caused by the high associated management costs will no longer exist. In addition, we aim to allow the *operators* to express the quality of service they wish to provide to customers, both for traffic they transmit *and for traffic they receive*. The end result should be a more efficient and robust network, with greater controllability. A likely side effect is that the increased 'openness' about pricing and delivered quality of service will encourage healthy competition between carriers.

The goal for the operator will be to allow themselves (and possibly their customers) to influence the service experienced by traffic carried. This includes the desire to influence the route taken by traffic as it traverses the network,

and has an implicit requirement for feedback on the quality of service traffic actually receives. This will allow operators to select their operating point more easily, and allow users to choose an operator appropriate to their desires.

The solution space

We now describe the so-called 'solution space,' the methods by which we intend to attack the problems detailed above. A key requirement of this work is the ability to monitor the quality of service delivered by network paths. We envisage using a number of different sources of data about *local* network conditions, such as information available through the IETF standards and from commercial products such as Cisco NetFlow, and more recent IETF developments such as packet marking. It will then be necessary to infer path characteristics by combining sets of local data.

Current suggestions for using packet marking for traffic engineering have either been highly conservative such as the IETF's ECN (EXPLICIT CONGESTION NOTIFICATION) proposals [9], or have concentrated on controlling congestion by generating a price based on marked packets received to influence user behaviour [10, 11, 12]. These approaches do not directly address the problem of management of traffic within the network. We propose that given the information provided by packet marking, operators can better manage the traffic within their network via intra-AS routeing.

We believe that addressing the problems described is most appropriately done at these timescales. Dealing with aggregates of traffic is simpler from an operational standpoint, as one has more time, and often more information, than when dealing with individual packets. Additionally, dealing with the problems at the borders of ASs allows for much easier realization of generated solutions. A significant problem with a number of other proposed pricing mechanisms is that they operate in the end-system, leading to the well-known problems of deployment of new end-system protocols.

Intra-AS routeing

From the point of intra-AS routeing, areas of interest include the manner in which the price is generated since it must take account, on a per-link basis, of at least prevailing congestion conditions and the available bandwidth, and

possibly other parameters such as link latency and loss likelihood. Furthermore, the properties of extensions required by extant routeing protocols to allow prices to be distributed so that they may affect routeing within the AS are of interest.

Inter-AS routeing

Areas of interest for inter-AS routeing include the methods used to calculate the settlement price for traffic and the way in which this might be connected to the current state of the intra-AS routeing protocol, to the way in which ASs can be made aware of the treatment traffic they originated or transited subsequently received in the Internet. Since ASs on the back-path for a route will also see how the requesting AS has traffic routed to it, an incremental, highly distributed version of such an option might be viable.

Work programme

We intend this to be a one year project in the first instance, with the option to extend it for a further year or more on the agreement of all parties after evaluation of the first year's progress.

Due to the nature of this research, we do not believe a detailed calendar of work is appropriate. In general, we intend to further develop the thesis outlined above through a mixture of analysis, modelling, and simulation. Our goal is to devise pricing schemes and routeing algorithms that promote desirable network behaviour. In the first instance such schemes may rely on 'global knowledge' about network state. We then intend to investigate approximations to these algorithms that can be implemented in a distributed fashion in a manner similar to conventional routing protocols.

It should be noted performing simulation of this nature is relatively difficult as suitable simulation environments do not currently exist. Furthermore, the stimuli necessary to drive such a simulation in a realistic manner are not well known. We intend to learn more about these factors through discussion with a number of ISPs; a key part of this work is its relevance to the commercial state of the network, and it is important that we build on an accurate and up-to-date picture of the network which can be used to drive the simulation.

Once simulation has enabled suitable algorithms to be developed, we would

hope to be able to produce prototype implementations by extending existing routeing protocols. However, this goal may not be achievable within the first year. Producing 'production quality' implementations would require further resources.

Dissemination and intellectual property

We expect the principle results of this work to be submitted as one or more papers to high-quality academic conferences or journals. In addition, the project will develop certain software, including the routeing simulator, and possibly prototype implementation of extended routeing protocols. It is our intention that all code developed would eventually be released under an Open Source license. Patent protection may be sought for routeing algorithms developed by the project.

Sponsors of the project would be given early access to results through regular meetings and on-going communication with the investigators. A formal presentation of the results of the work will be made to the sponsors at the end of the first year. The sponsors support would be prominently noted in any published material, including web pages. The sponsor would be granted a non-exclusive royalty-free license to any intellectual property developed directly by the project. However, it is the nature of the inter-AS routeing work we propose that its utility will depend upon its widespread implementation and deployment. This would be dependent on it becoming an IETF standard, which would probably require that the intellectual property be made freely available.

Management and resources

The Principle Investigator (Ian Pratt) will be responsible for management and direction of the project, and will supervise the work of the full-time Research Associate (Richard Mortier).

Research context

This research would be carried out in the context of the University of Cambridge Computer Laboratory's extensive experience in management and control of resources in computer systems. Specifically, previous work in the Systems Research Group has addressed the problems of mechanism for resource management in operating systems [13], and given such mechanisms, the need for policy specification [14]. Similarly, the Systems Research Group has considerable experience in resource control and management for networks from ATM [15, 16] to the Internet [17] and also MPLS [**?**].

Personnel

Richard Mortier received his B.A. in Mathematics in 1996 and his Diploma in Computer Science in 1997, both from Cambridge University. He is currently studying for a Ph.D. in the area of network resource allocation and control through pricing, with the Systems Research Group at Cambridge University Computer Laboratory. He is interested in many areas of systems, especially resource allocation and control in computer networks, operating systems and distributed systems.

Ian Pratt is a currently a member of faculty at the University of Cambridge Computer Laboratory. He holds a Ph.D. in Computer Science and was elected a Fellow of King's College, Cambridge in 1996. As a member of the Lab's Systems Research Group for over five years, he has worked on number of influential projects, including the Fairisle ATM LAN, the Desk Area Network Workstation, and the Nemesis operating system. His research interests cover a broad range of Systems topics, including computer architecture, networking, and operating system design.

References

[1] J. Wroclawski. The Use of RSVP with IETF Integrated Services. RFC 2210, IETF, September 1997.

[2] K. Nichols, S. Blake, F. Baker, and D. Black. Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers. RFC 2474, IETF, December 1998.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Service. RFC 2475, IETF, December 1998.

[4] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. Assured Forwarding PHB Group. RFC 2597, IETF, June 1999.

[5] V. Jacobson, K. Nichols, and K. Poduri. An Expedited Forwarding PHB. RFC 2598, IETF, June 1999.

[6] R. Yavatkar, D. Pendarakis, and R. Guerin. A Framework for Policy-based Admission Control. RFC 2753, IETF, January 2000.

[7] J. Boyle, R. Cohen, D. Durham, S. Herzog, R. Rajan, and A. Sastry. The COPS (Common Open Policy Service) Protocol. RFC 2748, IETF, January 2000.

[8] E. Rosen, A. Viswanathan, and R. Callon. Multiprotocol label switching architecture. Internet Draft, July 2000. Expires January 2001; available as `http://www.ietf.org/internet-drafts/draft-ietf-mpls-arch-07.txt`.

[9] S. Floyd. TCP and explicit congestion notification. *Computer Communication Review*, 24(5):10–23, October 1994.

[10] R.J. Gibbens and F.P. Kelly. Resource pricing and the evolution of congestion control. *Automatica*, 35:1969–1985, 1999.

[11] F.P. Kelly, P.B. Key, and S. Zachary. Distributed admission control. *IEEE Journal on Selected Areas in Communications*, 18(12):2617–2628, 2000.

[12] P. Key, D. McAuley, P. Barham, and K. Laevens. Congestion pricing for congestion avoidance. Technical Report MSR-TR-99-15, Microsoft Research, February 1999. `http://www.research.microsoft.com/research/network/disgame.htm`.

[13] I. Leslie, D. McAuley, R. Black, T. Roscoe, P. Barham, D. Evers, R. Fairbairns, and E. Hyden. The design and implementation of an operating system to support distributed multimedia applications. *IEEE Journal on Selected Areas in Communications*, 14(7):1280–1297, September 1996.

[14] N. Stratford and R. Mortier. An economic approach to adaptive resource management. In *Proceedings of the Seventh Workshop on Hot Topics in Operating Systems (HotOS-VII)*, 1999.

[15] S. Rooney, J.E. van der Merwe, S.A. Crosby, and I.M. Leslie. The Tempest: A framework for safe, resource-assured programmable networks. *IEEE Communications Magazine*, 36(10):42–53, October 1998.

[16] J.E. van der Merwe, S. Rooney, I. Leslie, and S. Crosby. The Tempest — a practical framework for network programmability. *IEEE Network Magazine*, 12(3):20–28, May 1998.

[17] R. Mortier, I. Pratt, C. Clark, and S. Crosby. Implicit admission control. *IEEE Journal on Selected Areas in Communications*, 18(12):2629–2639, December 2000.