Position paper: Dynamic Circuit Networks

Ian Pratt, 14 April 1998

University of Cambridge Computer Laboratory,
Pembroke Street, Cambridge CB2 3QG, U.K.
Ian.Pratt@cl.cam.ac.uk

1 Introduction

Will wide-area network bandwidth ever be cheap and plentiful? If it were, it would surely enable a plethora of new applications, and allow many of the current ones to function much better, bringing an end to the World Wide Wait.

New network access technologies such as ADSL and cable-TV modems promise to make significantly more bandwidth available for the final-hop into the home (2-8Mb/s as opposed to the 56Kb/s supplied by conventional analogue modems). However, the Internet Service Providers (ISPs) are struggling to expand the backbone Internet infrastructure to meet the demand caused by the growth in number of users. Its unlikely that a user with a 4Mb/s ADSL connection would actually receive much more bandwidth than a conventional modem user when accessing sites external to the local ISP.

In part, this is caused by the high percentage of Internet traffic that is 'long-haul'; people routinely access data from servers all over the world, not just those in their locality. Caching can help, but hit rates are not particularly high, certainly not enough to make up for the 100x increase in demand that would be created were everyone to get ADSL. The backbone infrastructure needs to drastically improve, with only minimal cost to the end user if growth in the user community is to continue.

2 Up-rating the infrastructure

A large part of the cost of building a world-wide network is in installing the fibre-optic cables that provide the physical media. Relatively recent advances in photonics such as Wave Division Multiplexing (WDM) now mean that many existing fibres can potentially be used to carry several orders of magnitude more data, with only limited additional cost.

However, this is not the full story. Access to large amounts of relatively cheap physical bandwidth will not necessarily result in cheap, plentiful bandwidth for the Internet user: The current Internet does a lot of 'work' per bit transported. Packets typically get forwarded to their destination through 15 or more IP routers. Each router must delve inside the header of the packet and perform a routing lookup to determine which output port the packet should be sent on. Making per-packet forwarding and queueing decisions requires significant hardware, which can make it both difficult and expensive to build really fast IP routers.

For this reason, many groups are investigating techniques to 'switch' rather than route some IP traffic. High-bandwidth 'flows' of data taking the same path through some group of

switches in the network are identified, and assigned a 'flow-Id'. The upstream switch spots packets for which a flow-Id has been assigned, and forwards the packet prefixed with a flow-Id tag. These tags are recognised by the downstream switches, and are used to index a table of known flows, enabling packets to be forwarded without the complexity of a full routing lookup.

3 Hybrid networks

IP switching techniques may help to reduce the complexity of IP routers, but perhaps a better technique would be to avoid any per-packet based routing or switching in the *core* of the network at all.

This could be achieved with a core that consisted of a large, reconfigurable circuit switched network, built mainly from all-optical switches operating in a combination of Wave Division and Time Division switching. Similar to the conventional phone network, bandwidth would be allocated in a large number of fixed size chunks (e.g. units of 100Mb/s). Circuits can be established across the network, enabling data inserted at one end to be transported to the other with minimal effort from the network. There would be no per-packet queueing in the core of the network at all – important since all-optical variable-length buffering is hard.

IP routeing would be only be used towards the logical periphery of the network. Routers would try to dispatch packets onto circuits that terminate 'near' the packets' destination, from where the journey will be completed using conventional routeing.

The routers will also be monitoring traffic flow, and making decisions about opening, closing or re-assigning circuits as flow patterns change over time. For example, if a queue is building for packets destined for dispatch to a particular circuit, the router may choose to negotiate the setting-up of another parallel circuit to the same destination to form a 'bundle' (it is expected than bundles consisting of up to 10-30 circuits will be fairly commonplace). Alternatively, the router may choose to create a circuit to a different, more specific destination, thereby capturing some of the load dealt with by the other circuit.

Sometimes, a router may decide it currently has no circuit that terminates sufficiently 'near' to the packet's destination, and it will choose to route the packet to one of its neighbours (also at the periphery) which does. In due course, a direct circuit may be established if traffic levels warrant.

Where multiple circuits are collected together into a bundle, individual packets will not be 'striped' across the circuits, but scheduled for transmission down a single circuit. This would be done so as to avoid issues of skew between circuits, and because there is little gain from striping in the wide-area: propagation delay is likely to dominate serialization delay.

This combination of IP routeing and circuit switching has a number of attractions:

The circuit switches should be simpler to construct than either Packet Transfer Mode (PTM) or even Asynchronous Transfer Mode (ATM) switches, and can be built to handle substantially higher bandwidths. They enable a very richly interconnected mesh of circuits to be established between network nodes, with huge aggregate bandwidth. IP routeing is only required towards the periphery of the network, where the total bandwidth through a single router is relatively

small, hence making their construction relatively straightforward.

By using a layer of IP routers at the periphery, many of the benefits of statistical multiplexing are retained. Packets sent from different (but relatively local) sources heading in the same general direction will be sent along the same circuit (or bundle).

Although this might not achieve quite the same link utilization as a pure PTM network, the waste should be acceptable. Under utilized circuits going to a similar destination on a busy link can be merged to free up bandwidth. Furthermore, due to the elastic nature of much traffic, provided the bandwidth bottleneck is not in the access network we will typically see flow's bandwidth expand to fill any empty allocation in a circuit bundle. Hence, ensuring bandwidth is shared 'fairly' among flows on other circuits is likely to be much more of an issue than maximising utilization (Besides, bandwidth is cheap, so we shouldn't care too much about high utilization).

Another benefit of the statistical multiplexing is that the behaviour of traffic flows should be relatively stable, thus the rate at which circuits will need to be established and modified will hopefully be relatively slow, perhaps only tens per second (this is important, since making good decisions here is likely to be computationally intensive).

A further benefit of this hybrid network is that QoS provision should be easier, since the core of the network is effectively passive (no queuing), and hence can't interfere with any QoS guarantees. Establishing a QoS guarantee (e.g. using RSVP) for some real-time traffic is likely to be easier when the data passes through only a handful of IP routers rather than 15+ experienced by many current Internet connections.

Many current Internet links already operate over a circuit switched network, using circuits leased from the Public Telephone Network Operators. The key difference between this and what is being proposed is that currently the circuit switched network is typically used to provide static point-to-point links between routers rather than a richly connected dynamically reconfigurable network that can be used to *replace* many routers.

4 Problems

Even if this paper's assumptions about the availability of cheap physical bandwidth and low-cost, high-performance optical circuit switches are proved correct, there are other hurdles to implementing hybrid networks.

The IP routeing tables of the peripheral network routers will be rather large, but perhaps no worse than today's core routers. Generating the routing tables will be altogether more complicated, and is likely to require considerable knowledge about the global state of the network. Perhaps this can be computed and disseminated in a distributed manner?

Other problems exist, such as the need for ISPs to co-operate and peer using circuits rather than IP routing (as inter-ISP border routers are precisely the kinds of monster router we're trying to avoid). This will probably mean that each ISP will have to 'advertise' their internal link structures to each other, which many ISPs currently regard as commercially sensitive information. The circuit set-up must take account of peering policies (e.g. as currently specified using BGP).

5 Thoughts about switch design

Due to the wide availability of highly accurate atomic clocks, it may be possible to run the optical core entirely asynchronously: Different regions of the network may be using different clock sources but the frequencies will be very accurate. Hence, the input to each switch only needs to compensate for phase, which will vary over time and can be tracked (e.g. due to fibre temperature variations).

Rather than using multiplexing at the unit of an octet, the optical switch design is likely to be much easier if a single bit is used.

As network speeds increase it would be nice if it were possible to increase the size of a single circuit, e.g. by trying to send two bits in the time where one was previously sent. However, its highly unlikely that any of the older-generation switch designs would be sufficiently transparent for this.

Ian Pratt (ian.pratt@cl.cam.ac.uk) 14 April 1998