IP over Photons: How not to waste the waist of the Hourglass

Jon Crowcroft,
UCL,
jon@cs.ucl.ac.uk

15 April 1999

Abstract

IP is often referred to as the waist of the hourglass, because of the simple model the Internet provides packet level communication compared to other approaches (e.g. FR, X.25 or B-ISDN), but runs over a complex plethora of links and switched networks, and is used by a firmament of application and transport protocols.

The ever increasing demand for capacity is pushing us past the 40Gbps barrier, and towards the terabit link.

Dense mode WDM offers 128 wavelengths at 10s Gigabytes per wavelength. However re-tuning takes geological epochs compared to IP packet switching, or flow switching, or even QoS routing time frames. The problem resembles that of IP over ATM or other virtual circuits, but this may be misleading - what is needed is a novel approach to a network architecture, that avoids the same design flaws and pitfalls we have been forced into in IP over ATM (including MPLS).

This paper will briefly describe why these approaches are misled, and attempt to outline an approach based on network edge packet sorting and scheduling, which may allow us to use pure optical core networks without sacrificing the flexibility of what we might call legacy electron-datagram services.

1 Introduction

Exponential growth in end system customers, as well as linear growth in access link speed is leading to near hyper-exponential demand for trunk link performance and switching/routing speeds. However, the Internet has traffic patterns that exhibit poor locality of reference, and therefore requires continued support for arbitrary routes without classical telecom network hierarchy. This may be alleviated by distance or volume or time related charging, but it seems a shame to kill the goose that has laid the golden egg.

This position paper is about approaches to continue the anarchic organisation of the Internet by enabling direct IP routing over photonic switching.

2 Goals

If we are to forward IP at optical speeds, we must consider the packet rate of $1Tbps = 1Gpkt/Sec$ which implies switching times of 1 ns per packet, per interface.

If we are not to incur excessive delay in forwarding, we must not handoff to electrical (even parallel) logic, which is "becoming" unacceptably slow.[1].

However, we cannot incur excessive cost in optical router. We can't build very clever cheap optical logic yet. It would appear that we need to do all this by end of 2001, and cannot wait for optical device people, so what is to be done?

2.1 Problem Description

We need to get a good understanding of the design space.

- What are properties of optical transmission?
- What are properties of IP (backbone) traffic?

[1] While Moore's law provides exponential speedup, in practice the half life is too slow: in general, as we add functionality, we defeat some of the speed up mechanisms such as RISC, multi-level I+D Caches, and super-scalar pipelining architectures

- What are future requirements for IP backbone traffic?

What are we not discussing:

- Gigabit (or more) to the desktop - in any case, it seems a non requirement (even CAVEs don't need more that few 100Mbps).
- Multiple queues (diff or int serv) in an optical router!

2.2 Competitive Analysis

Other approaches include:

- Parallel electrical routing (Pluris)
- Optical Logic (need to support enough of a device to implement a RISC processor with a cycle time fast enough to do 20 instructions per packet - I.e. 20 trillion instructions a second - some way off :-)
- Increasing the number of hops - this seems like a bad idea - increases delays and routing complexity (time to converge after change etc) and increases cost of network,
- Getting rid of IP and go for circuit switching is definitely a bad idea - a Tbps is a lot, but not when split between 100,000 video conferences, which are adaptive and potentially statistically multiplex-able too. In any case, circuits are an anathema to multicast and mobile applications.

3 Internetworking Technology

We have IP over X where X is currently one of:

- LAN - traditionally shared medium.
- Serial (PPP) - typically dial up or "leased line".
- NBMA (ATM, FR, etc) - Non-Broadcast Multiple Access.
- RF - Radio, can be one of:
  - CATV 802.14, possibly with ATM
  - HiperLAN
  - Satellite.

Is optical switching a good match to any of these?

- Typically, we multiplex the optical layer with different wavelengths and have wavelength "routers".
- There are about as many wavelengths per fiber as backbone IP routers currently have neighbours.
- Wavelength switching times are SLOW (millions of packet times).

IP flows are not currently long lived, demand is growing for long lived flows (RTP/RTSP streamed data, Web cache server based traffic, persistent HTTP, Internet Telephony etc). The spatial locality of reference should improve with these temporal locality trends too. We need to manage mapping IP packets to wavelengths This is a very similar task to matching IP packets to VCs on NBMA networks (and VCs on FR and ATM are scarce resources too, so 1-1 is not sensible) Main difference is that each packet can be shredded to ATM cells, whereas we want to burst multiple packets onto the same wavelength before going to the next one.

4 Propose: Edge Sorting, and Labelling

We propose we use MPLS and use LDP (or other) to distribute mapping between wavelength and destination set. We assume the "edge" of the optically routed region is multi-gigabit router - then the delay (and jitter) incurred in sorting packets to queues there, before forwarding onto a wavelength will not be significant. This also allows experimentation with different optical route+labelling algorithms.

We are going to try building this in a UK Project called LEARNET (with partners in BT, Cambridge, UCL and Essex and others) Not sure if it makes a lot of sense, since if anyone does build pure optical computing, this will be obsolete, but its seems like there is stopgap requirement for this approach at least.

There is a need for a GSMP-like protocol to interface to the optical router gear (not too hard

though). We also need to look at switching times and flow duration in more detail, and think about making system adaptive. We also need quite a lot of money (its about 1M US$ per optical router). This project is starting sometime in 1999.

5 Research Questions

Some interesting research questions include:

- Can we at least monitor traffic enough to tell what timers we can put into the edge sorting routines?
- Could we (should we) think about metering and billing?
- What about faults (e.g. if optical path fails, can we re-route in time to save a queue full in an edge router)?

6 Acknowledgements

Steve Deering first came up with (a version of) the title. Thanks to Tony Li at Juniper and Phil Lane in UCL EE for discussion of the ideas here.

References

[1] Multi-Protocol Label Switching is a working group of the IETF, whose web page is at `http://www.ietf.org/html.charters/mpls-charter.html`

[2] The figures on optical equipment costs and performances are back-of-the-envelope.