Wide Area Networks :

Backbone Infrastructure

Ian Pratt
University of Cambridge
Computer Laboratory

Outline

- Demands for backbone bandwidth
- Fibre technology
  - DWDM
- Long-haul link design
- Backbone network technology
  - IP Router Design
  - The near future : reducing layering
  - Longer term : all-optical networks

Internet Backbone growth

- ~125 million Internet hosts, ~350 million users
  - Host/user growth rate at 40-80% p.a.
  - Metcalfe's Law: "the utility of a network is proportional to the number of users squared"
- Access bandwidth increasing at 25%p.a.
  - Set to jump with DSL & Cable Modem
- High percentage of long-haul traffic
  - Unlike phone service where call freq. ? 1/?distance
  - Web caches & Content Distribution Nets may help
- Huge future requirements for backbone b/w

Optical Fibre

- Multi-mode fibre : 62.5/125?m
  - Typically used at 850nm
  - Requires less precision hence cheaper : LANs
  - Fibre ribbons
- Single-mode fibre : 8-10/125?m
  - Better dispersion properties
    - Normally best at 1310nm, can be shifted
    - 1310nm typically used in Metropolitan area
  - Minimum attenuation at 1550nm
    - NZDSF at 1550nm used for long-haul
- Fibers joined by "splicing"

Transceiver Technology

- Currently at 100Gb/s for a single channel
  - 2.5 and 10 Gb/s in common use (OC-48, OC-92)
  - Use TDM to subdivide channel
  - Improving at ~70%p.a.
- Wavelength Division Multiplexing
  - Use multiple 'colours' (?'s) simultaneously
  - 1310 & 1550nm – fused fibre couplers for de/mux
  - 4 channel 20nm spacing around 1310nm
    - Proposed for 10Gb/s Ethernet
  - So-called "Coarse WDM"

Dense WDM (DWDM)

- 100's or even 1000's of ? possible
  - e.g. 100x10Gb/s at 50GHz spacing
- need very precise and stable lasers
  - Temperature controlled, external modulator
  - wavelength tuneable lasers desirable
- gratings to demux and add/drop
  - Photo receivers are generally wide-band
- Fibre cap. currently increasing at ~180% p.a. !

Optical Amplifiers

- Erbium Doped Fibre Amplifiers (EDFA)
  - few m's of Erbium doped fibre & pump laser
  - wide bandwidth (100nm), relatively flat gain
  - 1550 'C' band, 1585 'L' band, also 'S' band
- Raman amplification
  - counter-propagating pump laser
  - Improve S/N on long-haul links
- Amplification introduces noise
  - Need 3R's eventually: reshape, retime, retransmit

Long-haul links

- E.g. as installed by "Level (3) Inc.":
  - NZDSF fibre (1550nm)
  - 32x10Gb/s = 320Gb/s per fibre
  - 12 ducts, 96 cables/duct, 64 fibres/cable
  - 100km spans between optical amplification
    - Renting sites for equipment is expensive
    - 8 channel add/drop at each site, O/E terminated
  - 600km between signal regeneration
    - Expensive transceiver equipment
- US backbone capacity up 8000% in 5 years!
  - Level 3, Williams, Frontier, Qwest, GTE, IXC, Sprint, MCI, AT&T,...

SONET/SDH

- SONET US standard, SDH European
  - OC-3 / STM-1 155Mbp/s
  - OC-12 / STM-4 622Mbp/s
  - OC-48 / STM-16 2.4Gbp/s
  - OC-192 / STM-64 10Gbp/s
- Can use as a point-to-point link
- Enables circuits to be mux'ed, added, dropped
- Often used as bi-directional TDM rings with ADMs
  - 50ms *protection* switch-over to other ring
    - "wastes" bandwidth, particularly for meshes
    - SONET/SDH switches under development
  - Perceived as expensive, provisioning relatively slow

IP Routers

- Need big, fast routers
  - Particularly at POPs for interconnecting ISPs
    - Densely connected mesh of high speed links
    - Often need features too : filtering, accounting etc.
- Rapidly becoming a bottleneck
  - Best today: sixteen OC-192 ports
- Fortunately, routeing is parallelize-able
  - Have beaten Moore's Law 70% vs. 60% p.a.
  - Recent DWDM advances running at 180%...

Router Evolution

- First generation
  - Workstation with multiple line cards connected via a bus
  - Software address lookup and header rewrite
  - Buffering in main memory
- Second generation
  - Forwarding cache & header rewrite on line card
  - Peer to peer transfers between line cards
    - Buffer memory on line cards to decouple bus scheduling

Router Evolution

- Shared bus became a bottleneck
- Third generation
  - Space-division switched back plane
    - pt2pt connections between fabric and line cards
  - All buffering on line cards
  - Full forwarding table
  - CPU card only used for control plane
    - Routeing table calculation
- Fourth generation
  - Optical links between line cards and switch fabric

IP Address Lookup

- Longest prefix match lookup
  - (find most specific route)
  - Map to output port number
- Currently, about 120k routes and growing
  - Need full table in core
  - 99.5% of prefixes = 24 bits (50% are 24 bits)
- Packet rates high on high speed links
  - 40 byte packet every 32ns on OC-192 10Gb/s

Hardware address lookup

- Binary trie
  - Iterative tree descent until leaf node reached
  - Compact representation, but
  - Lots of memory accesses in common case
- 24-8 direct lookup trie
  - 2^{24} entry lookup table (16.8MB) with 2^{nd} level table for the infrequent longer prefixes
  - Vast majority of entries will be duplicates, but
  - Only $20 of DRAM
  - Normally one lookup per memory access

Packet Buffer Requirements

- Routers typically have 1x b/w delay product of buffering per port
  - e.g. for OC-768 : 250ms x 40Gb/s =1.25GB/port
- Need DRAM for density, but random access to slow
  - currently around 50ns and improving at only 7% p.a.
  - 40 byte packet every 8ns at OC-768
- Use small SRAM at head and tail of a DRAM FIFO to batch packets and make use of DRAM's fast sequential access modes to the same DRAM row

Switch fabric design

- Ideal fabric would allow every input port to send to the same output port simultaneously
  - So-called output buffered switch
  - Implementation infeasible / unnecessary
- Input-buffered switches used in practice
  - Simple design suffers from head-of-line blocking
    - Limit of 58% of max throughput for random traffic
    - May be able to run fabric at greater than line speed

Switch Fabric Design

- Use "virtual output queues" on input ports
  - Scheduler to try and maximise fabric utilization
    - Choose links on request graph such as to maximise the number of output ports in use in each slot time
    - Bipartite match
  - Maximum Weight Matching now realisable
    - Previously used an approximation
- In future, parallel packet switching with load balancing looks promising

IP over ATM over SONET

- Uses SONET to provide point-to-point links between ATM switches
- Hang ATM switches off SONET ADMs
  - VC/VPs used to build a densely connected mesh
  - flexible traffic shaping/policing to provision paths
  - Can provide *restoration* capability ~100ms
- Hang IP routers off ATM switches
  - Routers see dense mesh of pt-to-pt links
  - Reduces # of high-performance routers required
    - Don't carry "through traffic"
  - IP capable of relatively slow restoration
  - MPLS to better exploit underlying ATM in the future

Near future: IP over SONET

- "Packet over SONET" (PoS)
- Build traffic shaping into routers/tag switches
- tag-switching to make routing more efficient
  - CDIR routing tricky, especially if packet classification for QoS required
  - Virtual circuit identifier pre-pended to packets
    - "soft-state" only
- Route at the edges, tag switch in the core
- Use MPLS to fix paths for flows
  - provision alternate paths
  - provide QoS etc.

Near future: IP over "not SONET"

- CISCO "Dynamic Packet Transport"
  - Replace SONET higher layers with something more amenable to packet transfer mode
  - still uses SONET physical layer (allows tunnelling)
  - Ring based architecture
    - Rapid self-healing through ring wrapping
      - Don't over commit critical traffic!
    - Flow-through and Local TX FIFOs in each station
    - Spatial Reuse Protocol (SRP) is bandwidth efficient
    - Uses 802.3 (Ethernet) 48 bit station addresses
    - Rudimentary QoS with two priority classes
      - Watermarks on FIFOs with back-pressure to other stations

All Optical Networks

- Really fast routers and ATM switches difficult and expensive
  - Variable buffering tricky
  - Optical-electrical-optical (OEO) conversion expensive
  - "only" on the semiconductor performance curve...
- Exploit DWDM : "transparent optical networks"
  - Use DWDM to build a *network* rather than a fat pipe
  - Use ?'s like ATM Virtual Paths

Optical Components

- ? Add-Drop Multiplexers (ADMs)
  - Fibre Bragg Gratings – in common use
  - Tuneable lasers - available
  - Tuneable filters – getting there
- Optical Cross Connects (OXCs)
  - Beam steering devices
    - holographic devices – typically very lossy
    - micro-mirrors
- ? converters – some promising technologies

All Optical Networks

- What functionality can we do all-optically?
  - IP routeing
    - Looks very hard
  - Packet switching (MPLS like)
    - Variable length packets may be tricky, as is header lookup
  - Cell switching
    - Buffering slightly easier, but still variable length
  - TDM
    - Fixed length buffering, out-of-band switch configuration
    - Looks do-able
    - Good enough for carrying traffic aggregates in core?