SUNETS TEKNISKA

TREF

REFERENSGRUPP

local area network next generation

table of contents

a short summary – in swedish

Under höstterminen 2000 formade SUNET:s tekniska referensgrupp en s k LAN-grupp med uppgift att fundera på – och föreslå – hur de lokala näten på högskolor och universitet ska byggas i framtiden för att kunna dra nytta av ett framtida GigaSunet.

I LAN-gruppen ingår:

- Jan Engvald, Lunds universitet
- Bengt Gördén, KTH
- Christer Holgersson, Umeå universitet
- Magnus Höglund, Högskolan Dalarna
- Börje Josefsson, Luleå tekniska universitet
- Johan Sandfeldt, Karolinska institutet

Resultatet av deras arbete presenteras i denna rapport. LAN-gruppens rapporter finns även tillgängliga på nätet, **http://tref.sunet.se/lang/**

LAN-gruppen har i sin rapport i stor utsträckning utgått från de intryck gruppen fick under en studieresa till ett flertal leverantörer under hösten 2000.

De tekniska lösningar som presenteras i rapporten – för stora och för mindre högskolor – är i dagsläget relativt kostsamma.

Förhoppningar finns dock att den nya utrustning som behövs för att förverkliga LAN-gruppens förslag, redan under sommaren 2002 ska vara tillgänglig till betydligt förmånligare priser än vad dagens utrustning är.

Det är med andra ord klokt att vänta med att förverkliga de förslag som LAN-gruppen förordar. Dock bör högskolor och universitet redan nu börja installera infrastruktur för nästa generations lokala nät.

Det gör man lämpligast genom att först läsa och tänka igenom föreliggande rapport noggrant.

SUNET:s tekniska referensgrupp
och LAN-gruppen, nedtecknat i mars 2001

report from vendor visits 2000

During November 2000 the SUNET technical reference group visited computer network software and hardware vendors in the USA. The vendors we planned to visit were:

- Computer Associates, makers of the Unicenter NMS
- Aprisma, makers of the Spectrum NMS
- Cisco Systems
- Extreme Networks
- Intel, canceled, due to misunderstandings between Intel Sweden and US offices
- Stanford University
- Nortel Networks

The motivation for the visits was to get firsthand information about which way these vendors thought networking technology was headed, and to hear what new software and hardware these vendors were planning on releasing the next two years.

This report tries to summarize the information received, which is not covered by NDA agreements.

Summary

There was only one really interesting new idea for how to build campus networks presented, see the "Future" section below.

10 gigabit Ethernet is supported by everyone, and is seen as the big technology for building future LAN/MAN and even WAN networks. There are two main versions of 10 GE, one that leverages SDH technology to make a manageable 10 GE WAN solution, and one meant for direct connections over a single fiber/wavelength in a LAN environment.

The industry sees the wireless market exploding. There will be between 4 and 100 times as many wireless devices as wired the prediction goes. Speeds for wireless will lag orders of magnitude behind wired though, so wireless will be a complement to, not a replacement for, wired networks.

Regarding network management/monitoring progress seems to be slow, a few improvements are to be expected, but at the cost of much more complex network management systems. The best bet still seems to be Spectrum.

Infrastructure

Cat5 twisted pair cabling is enough to support up to gigabit speeds at 100 meters. The tolerances are low though, so you should test any cable that you plan to run gigabit Ethernet on. Cat5E has had such testing done, so use Cat5E for new installations (same cost as Cat5). There seems to be no need to put in Cat6 or higher cabling.

Single-mode fiber is the only viable alternative for more than very short-range gigabit Ethernet, even more so for the upcoming 10 gigabit Ethernet standard. Deploy single-mode fiber only on any long-range connections, and even on short distances single-mode should be predominant.

If you are planning on fiber to the desktop you should install single-mode to be future-proof speed-wise, but there is a case for using multimode if all you want is to increase the distance between desktop and wiring-closet, not speed. The equipment cost is still 3-4 times copper though.

The cost for single-mode lasers will drop significantly starting late next year (2001).

The current generation of single-mode fiber will handle up to 40 Gbps per wavelength. To go faster than that per wavelength new fibers might have to be deployed. This might possibly affect the predicted life-length of nationwide fiber deployments.

Power-over-ethernet (via for example BIAS-T connectors or special switches with built-in power) for powering devices such as IP telephones and wireless base stations is gaining momentum. Gives 5W today, aiming for 15W. See IEEE 802.3af (DTE power via MDI).

There will be GBICs for specific colors in a DWM network, and inexpensive very short-range lasers for connecting to external multiplexors. Range around 300 meters with one-tenth of the price of traditional short or medium range lasers.

The choice of connectors for future fiber cards does not seem to be obvious. We did not get a definite answer on that. SC, LC and MT-RJ connectors were all mentioned.

10 gigabit Ethernet, IEEE 802.3ae

There is overwhelming support for 10 gigabit Ethernet. 10 GE will be widely deployed in LAN/MAN and even WAN networks. 10 GE will be fiber only, full duplex only.

There are two main versions of 10 GE. One using a LAN PHY that is meant for simple fiber connections with no repeaters, and hence has no means to monitor the underlying infrastructure, and one WAN PHY that can monitor each hop of a

10 GE link over a diverse underlying infrastructure, for example using wavelengths, repeaters or traditional SDH system hops (OC-192c / VC-4-64c).

The cost for 10 GE will be 3-4 times the cost of 1 GE, and 1/5 to 1/10 the cost of the corresponding SDH equipment (due to the need for very accurate clocking equipment for SDH).

The standard for 10 gigabit Ethernet is meant to be finalized in March 2002, and so far there is an 80% likelihood that it will meet that date. Despite that fact, everyone we talked with believed that after March 2001 there will be mostly editorial changes, so pre-standard equipment will be available in the summer of 2001.

Distances:

• 850 nm 50µ multimode	65 meters
• 1310 nm 62.5µ WWDM (4 channels) multimode, expensive	300 meters
• 1310 nm 9µ singlemode	10 km
•1550 nm 9µ singlemode	40 km

Channel multiplexing, IEEE 802.3ad, is part of the standard, but jumbo-frames are not (though there are vendor specific extensions). 10 GEchannel will probably be used more for resiliency than for raw bandwidth.

Resilient packet rings, IEEE 802.17

Current examples of such technology are DPT and DTM.

No vendor, except Cisco, said they had any plans to make DPT equipment, and they did not see DPT as a good technical solution for LAN environments (but might be usable for a MAN). Cisco will have DPT cards for the future GSRs, and plan to make DPT cards for the Catalyst 6500 series in the future (but not in the immediate future).

Cisco will release a router with 24 ports fast Ethernet (copper or fiber) and 2 ports for connecting to an 2.4 Gbps SRP (DPT) ring priced like a 7200 router (for the copper model) soon, there will also be a another model with 4 GE ports and 12 fast Ethernet in addition to the SRP interfaces.

Wireless

There was a general consensus that wireless will be big. The predictions ran from 4 to 100 times as many wireless devices as wired, but the wireless networks will lag orders of magnitude behind in speed, and will thus be a complement to, and not a replacement for, wired networks.

Equipment for IEEE 802.11b will become cheap. Expect $120 for a PC-card at the beginning of next year, and $50 at the end of the year. There is a problem with Bluetooth interference with 802.11b networks, which will grow as more Bluetooth devices are deployed.

Check for the "WiFi" (wireless fidelity) mark on wireless equipment. WiFi marked devices have been proven to work together with all other WiFi marked devices.

There is work an 22 Mbps, 54 Mbps and ~100 Mbps wireless. The 22 Mbps stuff will be available for around $200 per PC-card during 2001. The 54 Mbps equipment will be available during the second half of 2001, and the 100 Mbps equipment during 2002, but the higher speed equipment will be prohibitively expensive initially.

One interesting research was to make ultra-wide-band (UWB) wireless that uses very low power (less than the background noise) per wavelength. In this way you could have license-free 1 gigabit per second wireless with a 10 meter range. This could be used for example to have one such basestation per cubicle at work and all equipment there wireless.

Currently the Lucent wireless equipment seems to be the best, used in for example the Apple AirPort base-stations (which are cheap ;-)). These devices also have management/monitoring support from major vendors.

There are concerns regarding security on 802.11b networks. Not all vendors implement WEP (wire-equivalent privacy), and the implementations do not interoperate well. See also work on IEEE 802.1x. Will probably not be standardized until 2002.

IPv6

There is no drive for going to IPv6 in existing networks, however, the mobile phone 3rd generation networks will be based on IPv6, and countries in the third world that have no existing networks might (mostly) deploy IPv6 directly.

Will probably be IPv6 core networks with IPv4 used in the networks connecting the end users for the foreseeable future.

Cisco will include IPv6 support in the standard IOS release from this summer. IPv6 will require much more memory in routers.

Multicast

All the visited vendors have equipment that handles layer 2 multicast at wire speed. Beware that IGMPv3 will require software changes in intelligent switches, and of course to the IP stacks on the hosts.

The network management vendors had no plans on making tools for layer three multicast fault detection and monitoring within the next two years.

MSDP/MBGP is widely supported, and seen as the only viable cross-domain multicast interconnect today.

Noteworthy is that as multicast becomes more widely used, and the backbone networks increase in capacity faster than the edge devices, one solution is to always send all multicast everywhere in the backbone and just have the edge devices filter out the groups wanted. The overhead for keeping state for lots of constantly changing multicast groups is higher than for just installing fatter pipes.

Management

The network management/monitoring people seem to be moving very slowly towards better tools for intelligent root cause analysis, i.e. only telling the user that for example a certain trunk has failed instead of reporting every single network/device/host that became unreachable because of this.

The focus is moving towards monitoring the services running on the hosts, and of course the traditional layer 2 network monitoring. We felt that there is much missing between these two as regards to layer 3 monitoring, especially regarding routing and multicast.

What the vendors really wanted to sell was systems for doing software upgrades of hosts, keeping inventories etc. etc. This is in itself interesting, but not what we wanted to hear now. There was also talk about single-sign-on systems, and user administration.

Aprisma will support making automatic overlays of VLAN, OSPF, and multicast topologies on top of the physical network map in the February 2001 release. They also had the good sense to separate the network discovery process from the actual live management. You could now use the network discovery to cut and paste into your live manager, or to make automatic checks if the network has changed for example.

Computer Associates had some interesting work on using neural network agents (neugents) to correlate the vast amounts of data a NMS system collects and making predictions for future failures (as the network changes rapidly it is very hard to make static rules for error prediction, the NMS system has to "learn" what causes errors). For example they had a router neugent that could tell if something was wrong with the routing in a device, although not what.

There was also an agreement that as the complexity of the networks grows, more and more of the monitoring will have to be done in the wiring closets, either on special add-on cards to the networking equipment, or by putting special "manager" computers in each wiring closet that only report data to the central manager when needed.

Management of IP telephony is still in its infancy. The IP telephony vendors do not provide enough information to make intelligent monitoring.

In the US it is common to require that the service provider places agents in their network that can be used to verify SLA agreements.

LANs

There is talk that VLANs are a bad thing as they separate your physical and logical network. The "having to draw the map twice" problem. The recommendation is that you limit your VLAN use, and try to make sure that your logical and physical networks match.

There is a move towards partitioning networks purely geographically, for example all users in one wiring-closet on one network, one network per wiring-closet.

Most vendors see more and more intelligence moving towards the user to be able to keep up with the networking speed. The most extreme bid being one routed port per user. See the solution described in the "Future" section below for the other extreme.

Regarding monitoring of VLAN trunks (e.g. to see which VLAN on the trunk is sending the most traffic) Extreme Networks were the only ones who would have support for this in the near future. They will support individual counters for up to 96 VLANs on a trunk.

Extreme were also working on replacing spanning-tree, which takes ages to switch over to an alternative path, with a statically configured alternative path, this would give sub-second failovers. This makes a lot of sense as it is easy to build your network so that you know what the best alternative path is. There is also work ongoing on speeding up SPT convergence. See IEEE 802.1w.

There is a concern that most vendors did not have large enough buffers on their equipment to handle the future SUNET case where we have cross-Atlantic gigabit connections that will not see any bottleneck before the last switch hop (closest to the end user/consumer). This means that the last (first from our direction) switch should be able to handle at least 200 ms, preferably 350 ms, buffering of data per port. For gigabit speed this corresponds to 25-50 MB of buffering. Now most vendors seem to have 3-4 MB per 24 ports on shared buffer devices. Crossbar architectures like the Cisco 6500 only has a 64 kB buffer for 100 Mbps ports and a 512 kB buffer for Gb ports.

None of the vendors had any plans on making this upgradeable for customers who want larger buffers.

WANs

10 gigabit Ethernet using WAN PHYs are seen as a major contender for making cheaper WANs, and especially MANs.

Wavelength multiplexing will be used heavily, and the cost for low-density multiplexing (say 16 channels) will drop. There will be GBICs for different wavelengths, removing the need for wavelength conversion in the external multiplexors.

The bandwidth over a single fiber doubles every nine months. You can now run 6.4 Tbps on a single fiber for 4000 km with only optical repeaters (we didn't say it was cheap... :-)).

Using wavelength multiplexing the increase in networking speed is now higher than Moore's law, meaning that you can always build networks that are faster than the end equipment can handle. It is now more effective to build really fat networks that can handle all traffic than to do any sort of QoS in the core. The future core networks will be pure optical as electronics cannot keep up with the bandwidth explosion on the Internet.

There is work on making sub 100 msec route convergence a reality. This is needed for heavy IP telephony deployment.

Misc.

Some points, comments and links that we got, and that haven't been mentioned elsewhere in this report:

- High density line cards are much harder to make for fiber than for copper, mainly due to power consumption and heat. A guess is that the port density for fiber cards will stay below 2/3 of the copper versions.

- Work with Van Jacobsen shows that the TCP/IP feedback model of today is almost optimal.

- White paper on Gigabit Ethernet IEEE 802.3ab and 802.3z. (http://www.gigabit-ethernet.org/technology/whitepapers/gige_97/technology.html)

- Ethernet glossary (http://www.techfest.com/networking/lan/etherneta.htm)

- Optimized Engineering Technical Compendium (http://www.optimized.com/COMPENDI/index.html)

- Lightreading.com - about optical networks. (http://www.lightreading.com/)

Quotes

”Keep it simple” — see the gigabit network design presentation from Cisco (http://kokbok.sunet.se/lang/ha-design.pdf)

”Bandwidth is cheaper than complex logic”- regarding QoS in the core.

”It is simpler to go faster” — another complex logic comment.

”Bandwidth grows faster than Moore's law” – expect fast future networks.

”The future core network will be all optical with low-speed terrabit routers at the edges” – low speed terabit... ;-)

Future

The only really interesting new idea for how to build a campus network cannot be described in detail as it was covered by an NDA agreement. What follows is a high-level description.

Premises:

- In a campus environment where most of the traffic is going to/through some central equipment (the inverse 80/20 rule) it is quite stupid to put more and more intelligent and costly equipment into the wiring closet for the purpose of off-loading the central equipment as most of the traffic will have to be handled by the central equipment anyway.

- The campus has enough fibre that, using cheap low-density wavelength multiplexing, it is no problem to make a really fat campus core network.

- Low-density wavelength multiplexing equipment, and especially 10 gigabit ethernet equipment will be quite cheap.

- The next generation high-end layer three switches/routers will handle terrabit speeds.

The interesting solution proposed is then to build really cheap, stupid, low-end devices to put in the wiring closets that just multiplex all ports onto trunks based on 10 GE technology, which are then aggregated, per house for example, onto a wavelength multiplexed backbone and fed to a redundant central high-end switch/router that does all the work.

This creates a setup where you only have one (redundant) central switch/router that has every port on campus directly connected (via these multiplexors).

The rest of the network is just stupid, cheap, low-end devices.

Benefits:

- Cheap equipment where there is a lot of it, i.e in the wiring closets.
- Everything redundant except the edge devices with twisted pair ports. These have redundant uplinks, but not redundant electronics.
- All intelligence/management/monitoring done in the (redundant) core. This makes things like VLANs (you can put any port anywhere on the campus into the same LAN as any other port as they are all connected to the same switch), multicast (handled on the backplane of the central switch), management (only one piece of equipment to configure), monitoring (only one piece of equipment to monitor) etc etc easy.
- The monitoring includes detailed error detection to be able to see which of the multiplexors out on the campus have failed if such is the case.
- The central router/switch can be put in an optimal location without regard to network limitations. This also includes all servers etc that now can be placed anywhere in the network without network bottlenecks.
- A true non-blocking campus network. The design focuses on no overbooking at all to make the non-core devices really simple. Tests show that for example 10 times overbooking in the core works flawlessly.
- Yes, it will be an expensive central switch/router, but balance this against only stupid, low-cost edge devices, and that you probably would have to have a really expensive central piece of equipment anyway to handle the 80% traffic, and then consider the benefits described above.

Is this realistic?

Yes, we believe it is. The example we were shown used no equipment that is not available today or will be available soon. This includes cheap low-density wavelength multiplexors, 10 gigabit ethernet equipment, and terrabit routers. In fact it is the wait for cheap 10 GE equipment that was seen as the major reason for not introducing this earlier. They now expect to have a beta system up and running next summer, and for a system that handles around 15.000 end users connected at 100 Mbps to be generally available sometime around the summer of 2002.

Recommendations

The technology described in the "Future" section above is seen as a strong future path for large campus networks, and for that reason the technical reference group recommends that, if possible, you wait for more information on this (possibly until 2002) before making any major, costly upgrades to large campus networks.

For smaller campuses anything goes. Gigabit and 10 gigabit Ethernet seem to be the most cost effective solutions. Just make sure you get equipment that handle jumbo frames and buffering to leverage the future multi-gigabit SUNET connections.

For more discussions and current design examples, see the next generation campus LAN design report.

Disclaimer

SUNET, the individual members of the technical reference group or their organisations are not to be held responsible for the use of the information in this document. The information and configuration guidelines must be tailored for the individual organisation!

local area network next generation

This is a report written by the SUNET TRef LAN group (LANG). The purpose of the report is to give recommendations for how to build a next generation campus LAN that matches the next generation SUNET/NORDUnet backbone giving at least 2.5 Gbps to every campus.

This report should be read with the information in the "travel report, vendor visits 2000" fresh in mind.

General idea, modular design

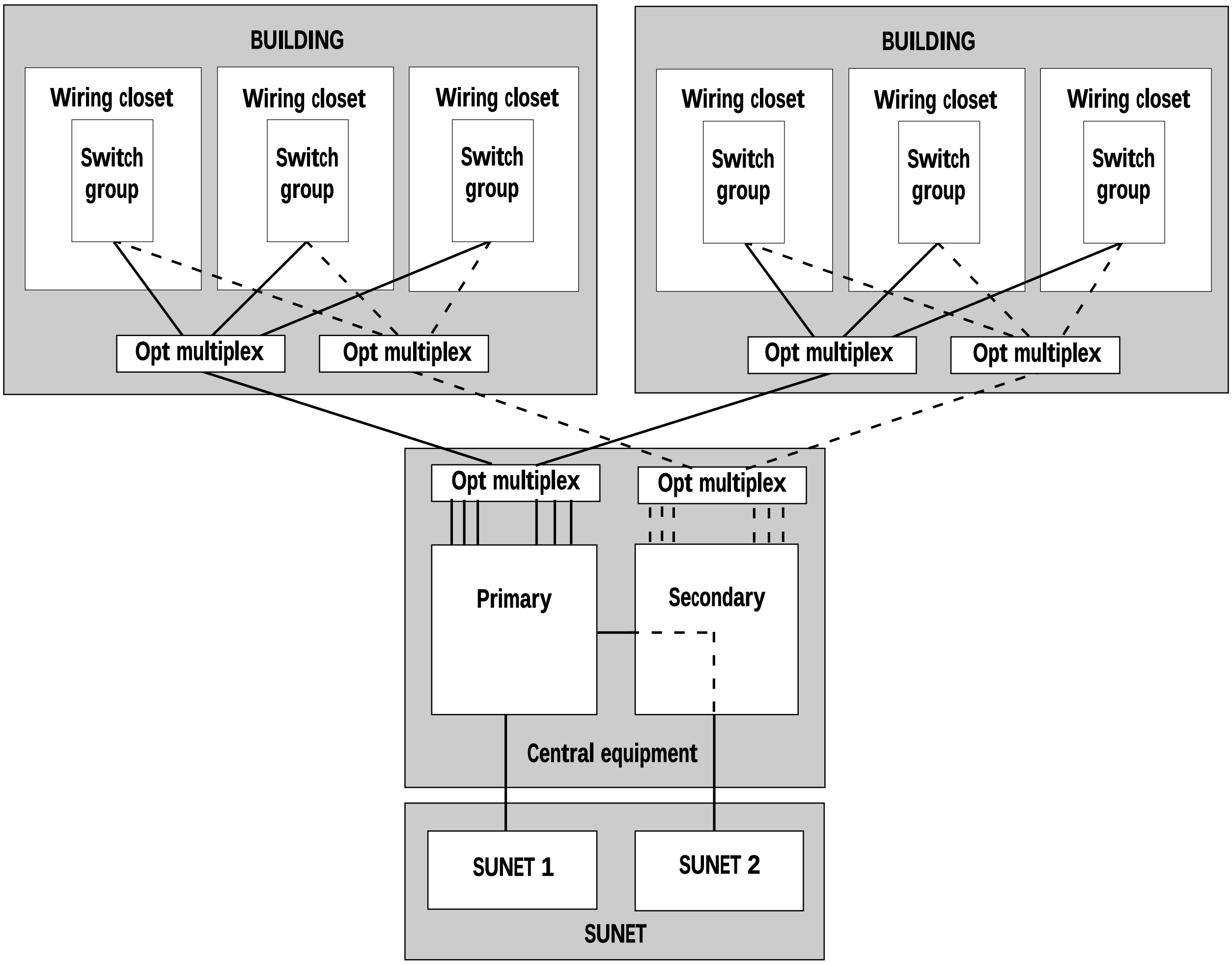
The general design is the same both for the current next generation campus LAN proposal, and for the LAN described in the "Future" section of the travel report - the proposed infrastructure works in both cases. By upgrading the building blocks this infrastructure will last several generations of campus LAN upgrades.

In short – one (two in the redundant case) pair of single-mode fibre from every wiring closet to the central equipment, using wavelength multiplexing between buildings if not enough fibre is available.

In the current proposal the building blocks interconnect with at least gigabit capacity. Gigabit ethernet, preferrably with jumbo frame support (9 kB MTU), is seen as the most cost-effective technology to achieve this, with 10 gigabit ethernet around the corner.

Most users connect with full duplex fast ethernet, this is deemed enough for the majority of users. Power-users and servers connect with gigabit ethernet. One gigabit user per 20 normal users is deemed a good design parameter. The wiring closet uplinks are also gigabit ethernet. One gigabit uplink per 200 users (10 power users) is deemed reasonable, with capacity growth by using gigabit etherchannel.

The equipment should be sized to handle full gigabit speed between the wiring closets and to the servers, with the SUNET/NORDUnet connection being at least 2.5 Gbps now, and maybe 10 Gbps



within the lifelength of the equipment. The edge switches should preferrably handle at least 200 ms buffering of gigabit traffic, roughly 25 MB shared packet memory, to handle a cross-atlantic gigabit connection, or a few within Sweden.

Multiplex

Ideally the multiplex component should not be used.

If you have (or can deploy) enough fibre to connect the central equipment to each wiring closet directly with one (two in the redundant case) fibre pair – do so!

These connections are assumed to be full duplex gigabit ethernet, and will normally need to be single-mode fibre unless the distances are short. You will definitely need single-mode fibre to every wiring closet for future higher speeds, so start deploying now.

If you do not have enough fibre to connect every wiring closet directly to the central equipment we recommend using WWDM (4 channel) or future cheap 16 channel DWDM (CWDM?) equipment to multiplex the connections to each building over 4 or 16 times less physical fibre. In this case you could also get away with using multi-mode fibre within the building, distances permitting.

Some of the products have back-haul protection switching, meaning that you could connect a multiplex with two fibers taking different paths to the central equipment. If one of the fibers is cut traffic will still flow uninterrupted. This gives a reasonable degree of protection in the not fully redundant case as it is the fibers between buildings that are most likely to get cut.

Another alternative (discouraged as it will introduce an extra switch hop with associated latency/jitter, packet loss and buffering issues) is to use a switch with one 10 gigabit ethernet connection towards the central equipment, and multiple gigabit ethernets towards the wiring closets.

Note that the multiplex equipment should handle jumbo frames (9 kB MTU) and VLAN trunking.

Example multiplex equipment

- Canoga-Perkins UCS/6004/L602 WWDM, 4 gigabit ethernet per fibre pair, up to 8 gigabit ethernet over 2 fibre pairs in a single chassis.
- When WWDM GBICs (1310, 1480, 1543, 1557 nm) are released this can be simplified to a standalone Canoga-Perkins 6004-1012-4001 for 4 gigabit ethernet over one fibre pair as no wavelength conversions will have to be done.
- Finisar Opticity 3000, 4-8 channels, GBIC based.
- Cisco Metro 1500 MAN DWDM, 8-32 channels
- Redfernnetworks GigaWave TMX-16, 16-64 channels
- Nortel Optera Metro
- Cienna MultiWave Metro
- Products by Sycamore, Vitesse, Lumenon et al.

Future cheap 16 channel DWDM (CWDM?) equipment is expected to be in the same price-range as a 16 port gigabit ethernet switch. We eagerly await GBICs for specific wavelengths.

Anyone who knows about cheap WWDM products or wavelength specific GBICs - please contact us!

Switch group

At the edges are VLAN capable switches with one (two in the redundant case) full duplex gigabit ethernet VLAN trunk connection, preferrably with jumbo frame (9 KB MTU) support, towards the central equipment. Using IEEE 802.1Q trunking is recommended.

The switch group might be a cluster of fixed-configuration switches (with at least one gigabit per second interconnect) or a larger modular chassis. The important thing is that the chosen switches have as large port buffers as possible to handle a large bandwidth*delay product. Shared memory architectures are normally better than having a fixed buffer per port as not all ports will be used simultaneously at the edge. A 25 MB shared memory buffer for gigabit speeds corresponds to rougly 200 ms, which is reasonable.

Avoid daisy-chaining switches to keep the number of switch hops as low as possible. Each switch hop increases delay and most importantly jitter and packet loss, and of course also lowers the overall MTBF.

It is beneficial if the switches respect and can set IP priority (ToS, DiffServ), or at least IEEE 802.1p CoS, as we see a need for prioritizing for example voice over IP and network based video end-to-end. Two to four different queues is deemed enough.

Most ports should be VLAN capable full duplex fast ethernet. Any port on campus can be connected to any other. This simplifies deployment and moves, and gives better utilization of switch-ports, leading to better economy. If you do not want to deploy campus-wide VLANs (for example if you prefer just one subnet per wiring closet) there is no problem doing that either, and you still have the possibility to add ports elsewere.

There should be at least one full duplex gigabit ethernet (1000Base-T or GBIC) port per 20 normal users for power-user or server connections. One gigabit uplink per 200 fast ethernet (10 gigabit ethernet) connections is deemed reasonable, with capacity growth by using gigabit etherchannel.

The uplinks will normally be 1000Base-LX/LH single- or possibly multi-mode fibre (see the multiplex section).

Note that if you deploy auto-sense 10/100 switches you should configure the ports statically whenever possible to avoid the all too common and hard to detect duplex mismatch problem. The building block interconnects at least should always be statically configured.

Do not use the spanning tree protocol, use etherchannel links and layer three redundancy instead.

Example edge switch equipment

- Cisco Catalyst 4006 modular 192 ports 10/100,
12 ports 1000Base-T, 2 ports GBIC
24 Gbps, 18 Mpps, 24 MB shared buffer
NO ToS/DiffServ, 802.1p with 2 queues
NO jumbo frame support.

- Extreme Networks Alpine 3808 modular
160 ports 10/100, 8 ports 1000Base-T, 4 ports
GBIC 64 Gbps, 48 Mpps, 16 MB shared buffer
ToS/DiffServ and 802.1p with 8 queues Jumbo
frame support.

Note that only the Catalyst 4006 really passes muster with 24 MB shared buffer, the lack of jumbo frames will hopefully be corrected in a later software release. The Alpine 3808 has better QoS support though.

Other switches like the Cisco Catalyst 35xx/29xx XL series and the Extreme Networks Summit48i have only 4 MB buffer, which is way too little for a gigabit edge switch. A little better for the Cisco Catalyst 2980G/2948G at 8 MB buffer, but this is still too little.

Central equipment, primary only

In the "primary only" case all wiring closets connect with full duplex gigabit ethernet VLAN trunks, preferrably with jumbo frame (9 KB MTU) support, to a central high-capacity gigabit ethernet switch/router. The central equipment should allow both VLAN (layer 2) and routing (layer 3) switching at full wire speed, and should preferrably handle future 10 gigabit ethernet modules also.

To get the needed performance you will need layer three hardware switching in the central equipment. You want hardware support for multicast too. This almost certainly implies that internal routing must be done by an integrated gigabit switch/router.

In the primary only case the central equipment should have as much redundancy as possible, in particular dual switch engines where you can upgrade one at a time, and that can do as seamless service takeover as possible, without having to reboot etc (minimum downtime). Of course you should also have dual power supplies, all modules should be hot-swapable etc.

You should also have redundant routing in the central equipment, either by having redundant router modules similar to the switch engines (upgradeable one at a time, fast takeover) or by using VRRP (HSRP, ESRP...) between two completely separate router modules. The first alternative means less complex configuration, the second allows for load balancing between the routers.

Whether to let the same routers handle internal routing and peering with SUNET is a choice you must make.

Using the same routers means less equipment, and as both internal and peering routers must handle routing at multiple gigabit speeds (SUNET will be at least 2.5 Gbps now, going towards 10 Gbps) this makes economic sense as such routers are fairly expensive. The router feature sets found on integrated switch/routers do not always include BGP4+ and in particular MBGP/MSDP support though, and for configuration simplicity splitting internal routing and external peering might also be beneficial.

Central equipment, primary and secondary

In the "primary and secondary" case all wiring closets connect with full duplex gigabit ethernet VLAN trunks, preferrably with jumbo frame (9 KB MTU) support, to a central primary high-capacity gigabit ethernet switch/router, and also via separate fibres to a secondary high-capacity switch/router. These can either be configured to handle half the traffic each (for example odd/even VLANs) or the secondary might only handle traffic in case the primary fails.

On possible alternative in the latter case might be to have a secondary switch/router with lower capacity, possibly only fast ethernet connections to the wiring closets (discouraged), as it will only be used when the primary equipment fails. Note that you will still need gigabit capacity in this box for the backup SUNET connection.

As this case assumes box redundancy, the boxes themselves need not have high internal redundancy – a single switch engine in each, and one router to each chassis using VRRP (HSRP, ESRP...) between the boxes for routing redundancy.

The SUNET peering in this case has the backup connection to/via the secondary box.

All comments for the "primary only" case are valid here too. The equipment should allow both VLAN (layer 2) and routing (layer 3) switching, hardware support is needed for layer 3 and multicast etc etc.

Example central equipment

- Cisco Catalyst 6000 series, see the example section.
- Extreme Black Diamond 6800 series, see the example section.

In the central equipment jumbo frame support is deemed necessary.

SUNET

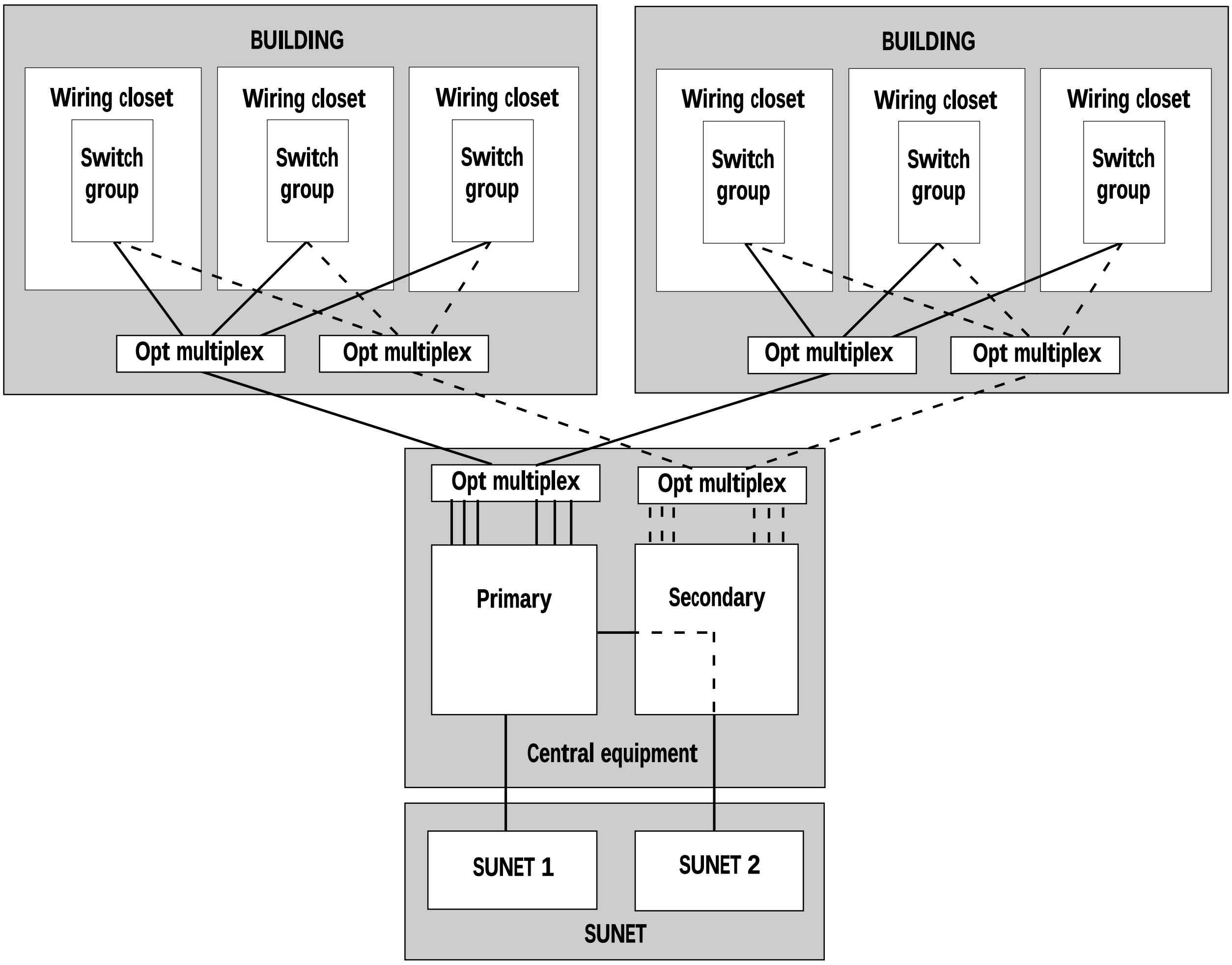
The future SUNET hand-off will most likely be with jumbo-frame (9 kB MTU) gigabit ethernet, maybe two per router, to handle a 2.5 Gbps connection to a 10 Gbps backbone. The hand-off will be via layer 3 (routing) load-balancing over the two gigabit ethernet connections from each of the SUNET routers in case a 2 gigabit ethernet hand-off is chosen. Normally all traffic will use the SUNET1 connection, SUNET2 will only be used in case of failures.
(NOTE: Nothing decided)

Example networks

Note that the following design examples are based on theory only, they have not been tested in reality. The Cisco equipment we are fairly confident about, the others need a more thourough evaluation.

The fine print is – the information and configuration guidelines must be tailored for the individual organisation, and SUNET TRef will not be held responsible for how you use this information.

For a really huge campus the idea is to use several of the "large campus" building blocks fully meshed via a 10 gigabit ethernet layer 2 switch with jumbo frame support (capacity growth by using 10 gigabit etherchannel).



next generation campus LAN design

The general design is the same both for the current next generation campus LAN proposal, and for the LAN described in the ”Future” section of the main report – the proposed infrastructure works in both cases.

By upgrading the building blocks this infrastructure will last several generations of campus LAN upgrades.

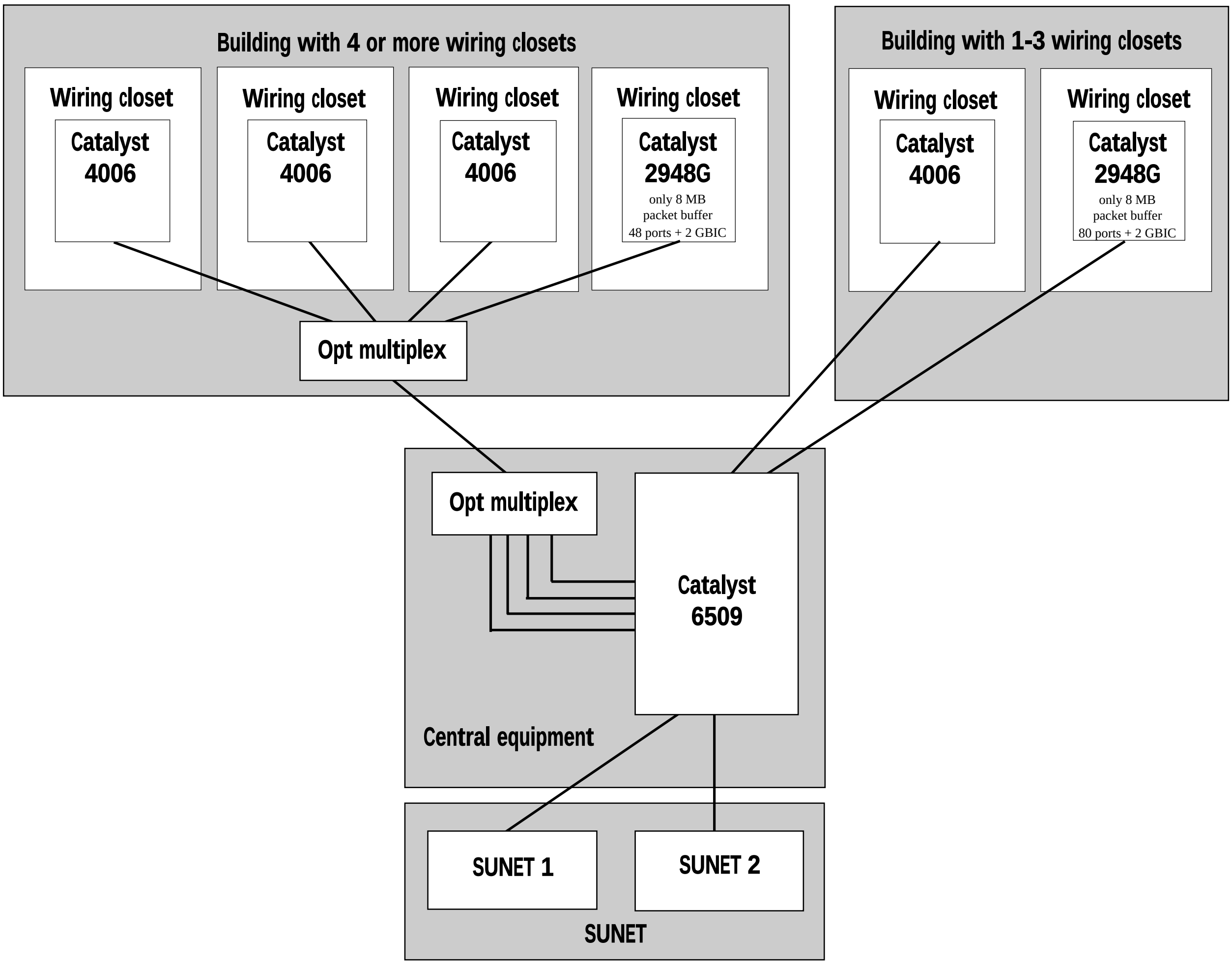
In short - one (two in the redundant case) pair of single mode fibre from every wiring closet to the central equipment, using wavelength multiplexing between buildings if not enough fibre is available.

In the current proposal the building blocks interconnect with at least gigabit capacity. Gigabit ethernet, preferrably with jumbo frame support (9kB MTU), is seen as the most cost-effective technology to achieve this, with 10 gigabit ethernet around the corner.

Most users connect with full duplex fast ethernet, this is deemed to be enough for the majority of users.

Power-users and servers connect with gigabit ethernet. One gigabit user per 20 normal users is deemed a good design parameter. The wiring closet uplinks are also gigabit ethernet. One gigabit uplink per 200 users (10 power users) is deemed reasonable, with capacity growth by using gigabit etherchannel.

The equipment should be sized to handle full gigabit speed between the wiring closets and to the servers, with the SUNET/NORDUnet connection being at least 2.5 Gbps now, and maybe 10 Gbps within the lifelength of the equipment. The edge switches should preferrably handle at least 200 ms buffering of gigabit traffic, roughly 25 MB shared packet memory, to handle a cross-atlantic gigabit connection, or a few within Sweden.



example network mainly Cisco equipment large campus

This is an example network for large campuses up to around 80 wiring closets, 15.000 connections, assuming 4 pairs of single mode fiber to each building.

Which Multiplex to use is still unclear. The Cisco Metro 1500 MAN is too expensive.

Used above is the less expensive Canoga-Perkins UCS/6004/L600 WWDM (4 channels, up to 8 channels in a single chassis) solution. Alternatives are eagerly awaited. Cisco will have GBICs for different wavelengths, meaning that the Multiplex can be reduced to only a standalone optical splitter (the 6004 module connecting the two chassis above in a standalone box).

Catalyst 4006 with Supervisor II	1 WS-C4006-S2
3 power supplies (needs 2).	1 WS-X4008/3
• Supervisor II (with 2 GBIC)	(do not use these GBIC)
• 48 port 10/100 RJ45 (or -21)	1-4 WS-X4148-RJ
• 48 port 10/100 RJ45 (or -21)	(or WS-X4148-RJ21)
• 48 port 10/100 RJ45 (or -21)	+
• 48 port 10/100 RJ45 (or -21)	+
• 12 x 1000Base-TX + 2 x GBIC	1 WS-X4412-2GB-T
• 2 GBIC 1000Base-LX/LH	2 WS-G5486

This gives 48-192 ports 10/100, 4 full 1000Base-TX ports (12 oversubscribed) and 2 usable GBIC ports in a chassis with redundant power and hot-swapable modules (but otherwise no redundancy) in the wiring closet.

Switch capacity is 24 Gbps (up to 64 Gbps), 18 Mpps. 24 MB shared memory buffer (200 ms at gigabit speed). Two QoS queues, only 802.1p CoS support, NOT IP ToS. NO gigabit jumbo frame support (limiting gigabit reach). Efficient multicasting utilizing the shared memory.

The GBIC ports on the Supervisor cannot be used if switch acceleration is used (enabled by ”set port disable 1/1-2” and ”set switchacceleration enable 1”).

The 1000Base-TX module is oversubscribed, grouped ports 1-4, 5-8, 9-12. The GBIC ports are linespeed and one is used as the uplink port with a 1000Base-LX/LH GBIC that can be used both over multimode (max 550 meters) and singlemode fibre (max 10 km). Alternatively a 1000Base-SX multimode only GBIC reaching 220 meters.

The C2948G and C2980G are basically fixed configuration C4003, and thus only have 8 MB packet memory, which is limiting. The C29xx/35xx XL only have 4 MB packet memory, which is too little for a gigabit edge switch.

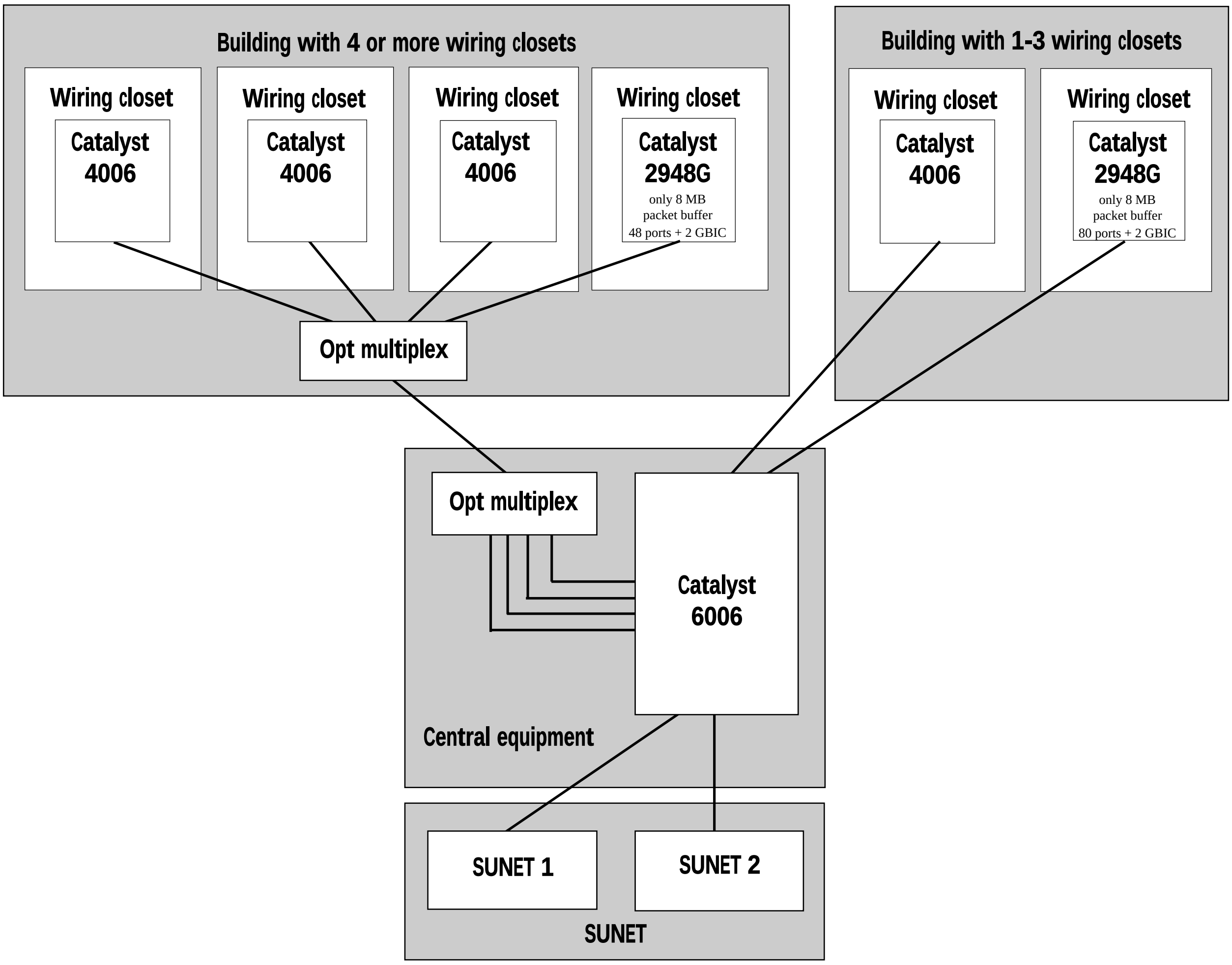
Catalyst 6509 chassis with 2500W power and redundant 2500W power supply	1 WS-C6509-2500AC 1 WS-CAC-2500W/2
• Supervisor II including 2 GBICs with PFC II (policy feature card 2) and MSFC II (multilayer switch feature card 2) with 512 MB memory, 24 MB flash	1 WS-X6K-S2-MSFC2 2 MEM-MSFC2-512MB 2 MEM-C6K-FLC24M
• Redundant Supervisor exactly as above.	1 WS-X6K-S2-MSFC2/2
• SFM (switch fabric module), in slot 6.	1 WS-C6500-SFM
• 1-5 GBIC 16 port modules, fabric enabled	1-5 WS-X6516-GBIC
• future 10 gigabit DPT (or ethernet) module.	1 WS-???

This gives 16-80 GBIC wiring closet connections, and 4 Supervisor GBICs for the SUNET connections, in a chassis with everything hot-swapable and redundant; power supplies, switching engines (supervisors), routers (MSFCs), and backplanes (256 Gbps SFM backed up by 32 Gbps backplane). 30 Mpps both layer 2 and 3.

All gigabit modules have jumbo frame support (9kB MTU) and handle CoS and ToS with 2 or 3 queues per port. Gigabit etherchannels from ports on separate modules for increased redundancy. NO server connections on this box as not enough buffering to handle long-delay gigabit connections (64 kB for 10/100 ports gives 0.5 ms, 512 kB for Gb ports gives 4 ms). Gigabit connected servers feasible, but better to connect these to gigabit ports on separate C4006 unless jumbo frame support is needed.

Note that DFC (distributed forwarding card) daugher cards are NOT used. Using these you would need two SFM for redundancy, and furthermore they are only supported in the integrated Supervisor IOS software, which lacks a lot of desirable features, like stateful failover, hitless software upgrades, jumbo frames, 802.1Q tunneling etc.

Spanning tree turned off on all ports. Dual MSFC modules for layer three redundancy. Use gigabit etherchannel to get layer two redundancy and higher uplink capacity to critical edge switches.



example network mainly Cisco equipment small campus

This is an example network for small campuses up to around 30 wiring closets, 5.000 connections, assuming 4 pairs of single mode fiber to each building.

Which Multiplex to use is still unclear. The Cisco Metro 1500 MAN is too expensive.

Used above is the less expensive Canoga-Perkins UCS/6004/L600 WWDM (4 channels, up to 8 channels in a single chassis) solution. Alternatives are eagerly awaited. Cisco will have GBICs for different wavelengths, meaning that the Multiplex can be reduced to only a standalone optical splitter (the 6004 module connecting the two chassis above in a standalone box).

Catalyst 4006 with Supervisor II	1 WS-C4006-S2
3 power supplies (needs 2).	1 WS-X4008/3
• Supervisor II (with 2 GBIC)	(do not use these GBIC)
• 48 port 10/100 RJ45 (or -21)	1-4 WS-X4148-RJ
• 48 port 10/100 RJ45 (or -21)	(or WS-X4148-RJ21)
• 48 port 10/100 RJ45 (or -21)	+
• 48 port 10/100 RJ45 (or -21)	+
• 12 x 1000Base-TX + 2 x GBIC	1 WS-X4412-2GB-T
2 GBIC 1000Base-LX/LH	2 WS-G5486

This gives 48-192 ports 10/100, 4 full 1000Base-TX ports (12 oversubscribed) and 2 usable GBIC ports in a chassis with redundant power and hot-swapable modules (but otherwise no redundancy) in the wiring closet.

Switch capacity is 24 Gbps (up to 64 Gbps), 18 Mpps. 24 MB shared memory buffer (200 ms at gigabit speed). Two QoS queues, only 802.1p CoS support, NOT IP ToS. NO gigabit jumbo frame support (limiting gigabit reach). Efficient multicasting utilizing the shared memory.

The GBIC ports on the Supervisor cannot be used if switch acceleration is used (enabled by ”set port disable 1/1-2” and ”set switchacceleration enable 1”).

The 1000Base-TX module is oversubscribed, grouped ports 1-4, 5-8, 9-12. The GBIC ports are linespeed and one is used as the uplink port with a 1000Base-LX/LH GBIC that can be used both over multimode (max 550 meters) and singlemode fibre (max 10 km). Alternatively a 1000Base-SX multimode only GBIC reaching 220 meters.

The C2948G and C2980G are basically fixed configuration C4003, and thus only have 8 MB packet memory, which is limiting. The C29xx/ C35xx XL only have 4 MB packet memory, which is too little for a gigabit edge switch.

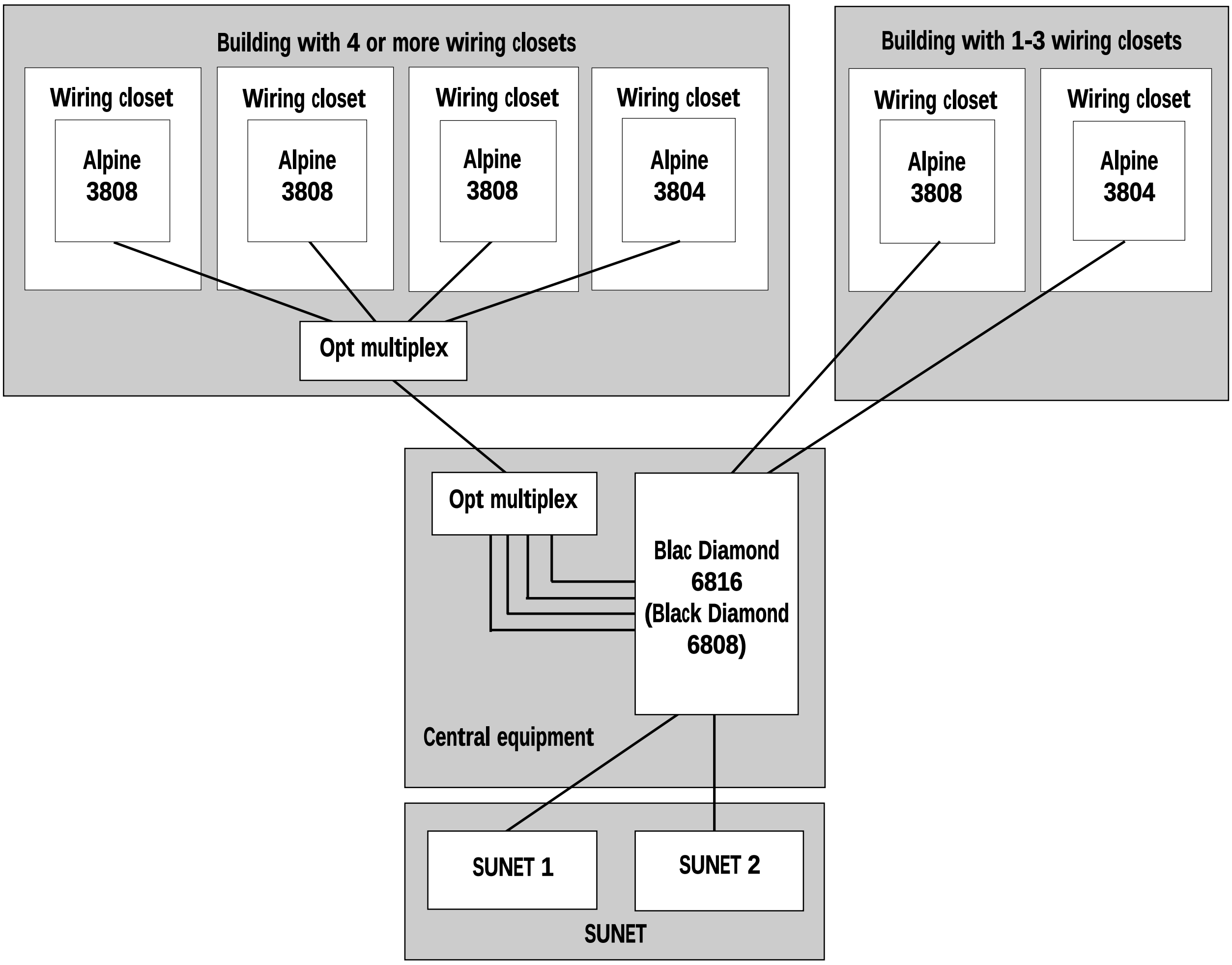
Catalyst 6006 chassis with1300W power and redundant 1300W power supply	1 WS-C6006-1300AC 1 WS-CAC-1300W/2
• Supervisor IA including 2 GBICs with PFC (policy feature card) and MSFC II (multilayer switch feature card 2) with 512 MB memory, 24 MB flash	1 WS-X6K-S1A-MSFC2 2 MEM-MSFC2-512MB 2 MEM-C6K-FLC24M
• Redundant Supervisor exactly as above.	1 WS-X6K-S1A-MSFC2/2
• 1-4 GBIC 8 port modules.	1-4 WS-X6408A-GBIC

This gives 8-32 GBIC wiring closet connections, and 4 Supervisor GBICs for the SUNET connections, in a chassis with almost everything hot-swapable and redundant; power supplies, switching engines (supervisors), routers (MSFCs), but NOT backplane. Switch capacity is 32 Gbps, 15 Mpps both layer 2 and 3.

All gigabit modules have jumbo frame support (9kB MTU) and handle CoS and ToS with 2 or 3 queues per port. Gigabit etherchannels from ports on separate modules for increased redundancy. NO server connections on this box as not enough buffering to handle long-delay gigabit connection (64 kB for 10/100 ports gives 0.5 ms, 512 kB for Gb ports gives 4 ms). Gigabit connected servers feasible, but better to connect these to gigabit ports on separate C4006 unless jumbo frame support is needed.

Note that if possible buying redundant Super-visor II (WS-X6K-S2-MSFC2) instead gives signi-ficant performance improvements. Do not use the integrated Supervisor IOS software, it lacks a lot of desirable features, like stateful failover, hitless soft-ware upgrades, jumbo frames, 802.1Q tunneling etc.

Spanning tree turned off on all ports. Dual MSFC modules for layer three redundancy. Use gigabit etherchannel to get layer two redundancy



example network mainly Extreme equipment

This is an example network for large (and small) campuses up to around 120 (60) wiring closets, 19.000 (9.000) connections, assuming 4 pairs of single mode fiber to each building.

Which Multiplex to use is still unclear.

Used above is the Canoga-Perkins UCS/ 6004/L600 WWDM (4 channels, up to 8 channels in a single chassis) solution. Alternatives are eagerly awaited.

The edge switches can be run with only basic layer 3 support, while the Black Diamond runs the full layer 3 license.

Compared to the Cisco alternative this gives much better QoS; extensive classification and 8 queues per port to the edge. All switches have jumbo frame support. There will also be per VLAN counters (up to 96) for VLAN trunks making monitoring much easier. The Black Diamond has better buffering than the Catalyst 6000, but the Alpine has less than the Catalyst 4006.

On the downside MBGP/MSDP is not supported, meaning you will have to have a separate multicast peering router. There is also no NetFlow accounting, no generic tunnels, no NTP server capability, no automatic verify reverse path, and you have to reboot to fail over between supervisors (causing at least 3 minutes downtime). Most of this will be corrected by the end of Q3 2001 says Extreme. The handling of ”secondary” addresses is also different, each secondary net becomes a separate VLAN.

There is a concern regarding the amount of packet memory for the Alpine switches, and manual configuration to increase buffering for specific ports, but all in all they seem a good alternative to the Cisco 4006 if you want/need the extra features.

A Black Diamond switch centrally seems to be a viable alternative to a Cisco Catalyst 6000 if you can wait for the Q3 2001 release, but this must of course be tested when that release is available.

The Extreme switches all have the same basic architecture.

The Alpine 3808 is 64 Gbps, 48 Mpps, 16 MB shared memory [limiting] switch, and can be configured with 160 ports 10/100, 8 ports 1000-TX, and 4 ports GBIC. Redundant power and hot-swapable modules, otherwise no redundancy. Expensive to use as layer 2 switch only, but very good QoS/classification support.

The Alpine 3804 is 32 Gbps, 24 Mpps, 16 MB shared memory [limiting] switch, and can be configured with 64 ports 10/100, 4 ports 1000-TX, and 4 ports GBIC. Redundant power and hot-swapable modules, otherwise no redundancy. Expensive to use as layer 2 switch only, but very good QoS/classification support.

The Summit 48i is a fixed configuration 17.5 Gbps, 10.1 Mpps, 4 MB shared memory [which is too little] switch with 48 ports 10/100 and 2 ports GBIC. Do not use a switch with this little packet memory at the edge.

The Black Diamond 6816 is a 256 Gbps, 192 Mpps, 16 MB shared memory per supervisor switch, and can be configured with up to 128 ports GBIC. The BD is fully redundant with 4 load-balancing switch modules and power supplies. Everything hot-swapable.

The Black Diamond 6808 is a 128 Gbps, 96 Mpps, 16 MB shared memory per supervisor switch, and can be configured with up to 64 ports GBIC. The BD is fully redundant with 2 load-balancing switch modules and power supplies. Everything hot-swapable.

Issues impacting Gigabit Networks:

Why don’t most users experience high data rates?

Phil Dykstra
Chief Scientist
WareOnEarth Communications, Inc.
phil@wareonearth.com

The United States has several high speed nationwide networks that support Research, Engineering, and Education. These networks should support data rates in excess of 100 Mbps, with many OC3 (155 Mbps), OC12 (622 Mbps), and even OC48 (2.4 Gbps) links. Routine end-to-end data rates approaching 100 Mbps is even a goal of the Next Generation Internet program. Yet most users today see perhaps one tenth of that goal. Why is this, and what should we do to improve the situation?

Recent measurements on the Defense Research and Engineering Network (DREN), vBNS, and Abilene networks have painted a rather grim picture of typical end-to-end performance. There appear to be many obstacles to high data rate flows. We briefly discuss some of them below, along with network design concepts that have become increasingly important as data rates have increased. Several of these concepts come directly from the estimate of TCP throughput:
bps < min(rwin/rtt, MSS/(rtt*sqrt(loss)))

Window Size Matters

Most of our end systems still default to offering ~16KB TCP receive windows (rwin), or even 8KB. These values are fine for high-speed local area networks, and low-speed wide area networks, but they severely limit throughput on high-speed wide area networks. For example, over a coast-to-coast path (rtt = 40 msec), TCP could not exceed about 3.3 Mbps, even if it was running over a gigabit per second path.

The answer is not as simple as setting a large default rwin. Too large of a default window can run your system out of memory, since every connection will use it. A large window can also be bad for local area performance, and bad for some interactive sessions or applications that don't require high data rates. What is needed are tuned applications - ones that use large windows when and where appropriate - and/or adaptive TCP stacks that adjust rwin based on actual use. Web100 is an example of one project that aims to provide an adaptive TCP for the masses. We could do more today to improve typical user performance by improving end system software than we could by installing more high speed links.

Latency Matters

”High speed” networks are really high capacity networks. The "speed" with which data moves down a T1 line, or an OC48 line, are both dictated by the speed of light in the media.

Architectural changes in our high performance networks in the past few years have often resulted in increased delay or latency between pairs of sites. Examples include the relatively small number of Network Access Points (NAPs) where networks interconnect, the concentration of sites behind Gigapops, and the reliance on a fairly small number of very high capacity trunks.

On low capacity networks, latency cause by propagation delays wasn't a very critical factor. Today however we see numerous cases where the performance an application sees is directly impacted by the geographic path length of the network. Everything else being equal, TCP can go twice as fast if the path length (latency) is cut in half. Yet our routers today usually choose paths based on minimizing the number of hops, and following the highest capacity path, even if that means routing across the country and back. Many applications will do better over a low latency OC3 path than over a high latency OC48 path, yet we have no way to ask for such a path from the network.

On a single high performance network today, measured latencies are typically ~1.5x - 3x that expected from the speed of light in fiber. This is mostly due to taking longer than line-of-site paths. Between different networks (via NAPs) latency is usually much worse. Some extra distance is required, based on the availability of fiber routes and interconnects, but much more attention should be given to minimizing latency as we design our network topologies and routing.

MTU Matters

Packet size can have a major impact on throughput. The dynamics of TCP are such that, for a given latency and loss rate, there is a maximum packet per second rate than can be achieved. To increase throughput, you have to increase the packet size (or reduce latency or loss, which is something the end systems can't control).

Today the world is rapidly heading to where 1500 bytes is the largest supported end-to-end packet size. This is because of the dominance of ethernet technology, and the use of 1500 bytes even at gigabit data rates. Such small packets are a major obstacle to high performance TCP flows. At one gigabit per second, this equates to over 83000 packets per second, or only 12 microseconds per packet. There is no reason to require such small packets at gigabit data rates.

In the short term, the author hopes that the 9KB "jumbo frame" proposal for gigabit ethernet becomes widespread. In the longer term, we should build high speed networks that can support much larger packet sizes. The backbone links and NAPs are particularly important, because if they restrict MTU, the end systems are helpless. It is hard to overstate the importance of this issue.

Loss Matters

In the old days we though 10% packet loss was acceptable. After all, TCP does error recovery, and 90% isn't bad, right? Today, many service level agreements (SLAs) target a loss of 1% or less (often averaged over 24 hours). For gigabit data rates however, loss has to be extraordinarily low!

For example, to achieve a gigabit per second with TCP on a coast-to-coast path (rtt = 40 msec), with 1500 byte packets, the loss rate can not exceed 8.5x10^-8! If the loss rate was even 0.1% (far better than most SLAs), TCP would be limited to just over 9 Mbps. [Note that large packet sizes help. If packets were n times larger, the same throughput could be achieved with n^2 times as much packet loss.]

Buffering Matters

Gigabit networks thus need to be nearly lossless. We believe that one of the reasons that such low loss isn't being observed is because most of today's routers and switches have insufficient buffering for such high bandwidth-delay products. Few of the high performance paths we have studied show stable queueing regions. Also, the concept of depending on loss to indicate congestion to TCP may not apply very well at extreme bandwidth-delay products.

Bugs Are Everywhere (especially duplex ones)

Recent measurements over numerous high performance paths have turned up a wealth of bad behavior, much of which is still unexplained. Slow forwarders, insufficient buffering, strange rate shaping behavior, duplex problems, packet reordering, and low level link or hardware errors are all playing a part. Very few paths can sustain the packet per second data rates that you would expect from the underlying hardware and links.

At least one problem deserves special mention. The failure of ethernet auto-negotiation, and the resulting duplex problems, are perhaps the single biggest performance bug on the internet today. The results of this bug only show up under load which makes them difficult to notice. Low rate pings show almost no loss, but high data rate loads result in dramatic loss. Our tests, and similar reports from others, are indicating that this bug has reached epidemic proportions.

Why Debugging Is Hard

Network test platforms and programs are usually only available at the edges of the network. When end-to-end tests are performed, they span many different devices and links. The result is a messy convolution of all of the behaviors along the path. When bad behavior is observed, it is sometimes nearly impossible to figure out where in the path the problem lies.

Performance problem debugging would be vastly easier if routers provided some kind of high performance testing service. Routers are designed to forward packets well, but are usually very bad at answering traffic directed to them. This means that tests can't be directed at a router in order to debug a path problem hop-by-hop. If you target a router in the middle of the path, you get such poor performance that other problems you are looking for are usually masked. The participation of routers in something like the proposed IP Measurement Protocol (IPMP), and/or a high speed echo service, would greatly aid in debugging.

Security Isn't Helping

The ever increasing security threat, and level of abuse on the internet, has led to numerous measures that decrease performance and make performance measurement and debugging more difficult. ICMP is often blocked making ping and/or traceroute impossible. Deliberate rate limits are sometimes imposed on ICMP or other traffic as a measure to defend against denial of service attacks. Sometimes only a limited number of TCP and UDP port numbers are left unblocked, which can prohibit measurement applications that use other ports. And an increasing number of Firewall and Network Address Translation (NAT) boxes are in the path, creating a loss of end-to-end transparency.

The performance impact of all of these measures has not been well studied. What exactly is the slowdown of different routers given certain kinds of filter lists? How fast do various firewall and NAT devices forward packets under different traffic situations? Can you bypass these security mechanisms for authenticated applications? The use of ICMP for measurements should probably be phased out, but acceptable alternatives to ICMP need to be created.

Measurements Are Easy, Analysis Is Hard

We are doing well today at collecting basic measurements. Projects like AMP, Surveyor, PingER, and RIPE, are gathering a wealth of delay, loss, and route information. What we aren't very good at yet is learning things from all of that data. Major progress could be made from detailed automated analysis of the data: detection of anomalies, correlation of events, high level abstraction of causes. There are several projects working in this direction, but we are just beginning.

ABOUT THE AUTHOR: Phil Dykstra is the Chief Scientist of WareOnEarth Communications, Inc., and heads the San Diego office which is focused on the measurement and analysis of high performance networks. Prior to WareOnEarth, he was the head of Advanced Development in the Army Research Laboratory High Performance Computing Division. With over 20 years of Internet and HPC experience, he fostered US Federal Networking for many years, and until recently, co-chaired the Joint Engineering Team (JET) which coordinates Next Generation Internet (NGI) and Internet2 engineering and plans. He has taught Computer Science courses at Johns Hopkins University

Gigabit Ethernet Jumbo Frames – and why you should care...

Phil Dykstra
Chief Scientist
WareOnEarth Communications, Inc.
phil@wareonearth.com

Whether or not Gigabit Ethernet (and beyond) should support frame sizes (i.e. packets) larger than 1500 bytes has been a topic of great debate. With the explosive growth of Gigabit ethernet, the impact of this decision is critically important and will affect Internet performance for years to come.

Most of the debate about jumbo frames has focused on local area network performance and the impact that frame size has on host processing requirements, interface cards, memory, etc. But what is less well known, and of critical concern for high performance computing, is the impact that frame size has on wide area network performance.

This document discusses why you should care, and about the largely ignored but important impact that frame size has on the wide area performance of TCP.

How jumbo is a jumbo frame anyway?

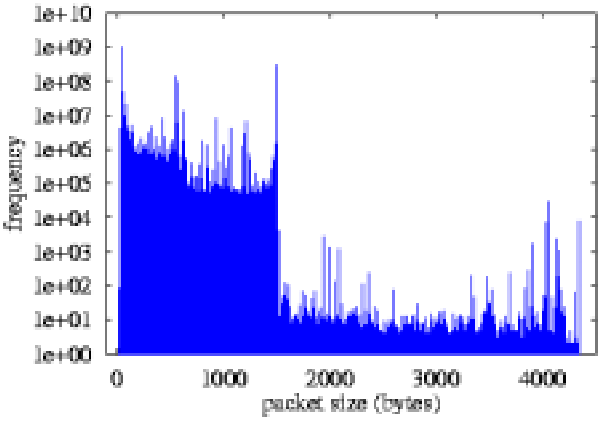
Ethernet has used 1500 byte frame sizes since it was created (around 1980). To maintain backward compatibility, 100 Mbps ethernet used the same size, and today ”standard” gigabit ethernet is also using 1500 byte frames. This is so a packet to/from any combination of 10/100/1000 Mbps ethernet devices can be handled without any layer two fragmentation or reassembly.

”Jumbo frames” extends ethernet to 9000 bytes. Why 9000? First because ethernet uses a 32 bit CRC that loses its effectiveness above about 12000 bytes. And secondly, 9000 was large enough to carry an 8 KB application datagram (e.g. NFS) plus packet header overhead. Is 9000 bytes enough? It's a lot better than 1500, but for pure performance reasons there is little reason to stop there. At 64 KB we reach the limit of an IPv4 datagram, while IPv6 allows for packets up to 4 GB in size. For ethernet however, the 32 bit CRC limit is hard to change, so don't expect to see ethernet frame sizes above 9000 bytes anytime soon.

How can jumbo frames and 1500 byte frames coexist?

Two basic approaches exist:

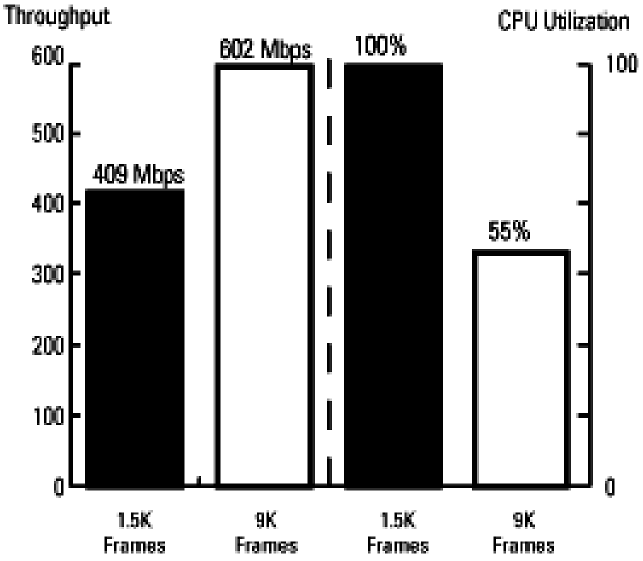
- On a port by port basis, where everything ”downstream” from a given port is known to support jumbo frames.
- Using 802.1q Virtual LANs, where jumbo frame and non-jumbo frame devices are segregated to different VLANs.



What frame sizes are actually being used?

The above graph is from a study[1] of traffic on the InternetMCI backbone in 1998. It shows the distribution of packet sizes flowing over a particular

Extended Ethernet Frames vs. Standard Ethernet Frames*



* Using Gigabit Ethernet. Throughput on tests was limited to SBus capacity. TCP tests used dual 300 Mhz Sun servers running Solaris 2.5.1

backbone OC3 link. There is clearly a wall at 1500 bytes (the ethernet limit), but there is also traffic up to the 4000 byte FDDI MTU. But here is a more surprising fact: while the number of packets larger than 1500 bytes appears small, more than 50% of the bytes were carried by such packets because of their larger size. Also, the above traffic was limited by FDDI interfaces (thus the 4000 byte limit). Many high performance flows have been achieved over ATM WAN's offering 9180 byte MTU paths.

Local performance issues

Smaller frames usually mean more CPU interrupts and more processing overhead for a given data transfer size. Often the per-packet processing overhead sets the limit of TCP performance in the LAN environment. The above graph, from a white paper[2] by Alteon is an often cited study showing an example where jumbo frames provided 50% more throughput with 50% less CPU load than 1500 byte frames.

Such local overhead can be reduced by improved system design, offloading work to the NIC interface cards, etc. But however you feel about these often debated local performance issues, it is the WAN that we are most concerned about here.

WAN TCP performance issues

The performance of TCP over wide area networks (the Internet) has been extensively studied and modeled. One landmark paper by Matt Mathis et al.[3] explains how TCP throughput has an upper bound based on the following parameters:

Throughput <= ~0.9 * MSS / (rtt * sqrt(packet_loss))

So maximum TCP throughput is directly proportional to the Maximum Segment Size (MSS, which is MTU minus TCP/IP headers). All other things being equal, you can double

your throughput by doubling the packet size! This relationship seems to have escaped most of the arguments surrounding jumbo frames. [Packet_loss may also increase with

MSS size, but does so at a sub-linear rate, and in any case has an inverse square effect on throughput, i.e. MSS size still dominates throughput.]

In the local area network or campus environment, rtt and packet loss are both usually small enough that factors other than the above equation set your performance limit (e.g. raw available link bandwidths, packet forwarding speeds, host CPU limitations, etc.). In the WAN however, rtt and packet loss are often rather large and something that the end systems can not control. Thus their only hope for improved performance in the wide area is to use larger packet sizes.

Let's take an example: New York to Los Angeles. Round Trip Time (rtt) is about 40 msec, and let's say packet loss is 0.1% (0.001). With an MTU of 1500 bytes (MSS of 1460), TCP throughput will have an upper bound of about 8.3 Mbps! And no, that is not a window size limitation, but rather one based on TCP's ability to detect and recover from congestion (loss). With 9000 byte frames, TCP throughput could reach about 51 Mbps.

Or let's look at that example in terms of packet loss rates. Same round trip time, but let's say we want to achieve a throughput of 500 Mbps (half a "gigabit"). To do that with 9000 byte frames, we would need a packet loss rate of no more than 1x10^-5. With 1500 byte frames, the required packet loss rate is down to 2.8x10^-7! While the jumbo frame is only 6 times larger, it allows us the same throughput in the face of 36 times more packet loss.

But aren't jumbo frames bad for multimedia?

For applications that are sensitive to burst drops, delay jitter, etc., it can be argued that large frames are a bad idea.

No application has to use large frames however, so the question is really whether other application's large frames will negatively impact your application's small ones. This is primarily an issue of slot time, i.e. how much will a large packet delay (or quantize) the time(s) available to transmit the small packets.

A 9000 byte GigE packet takes the same amount of time to transmit as a 900 byte fast ethernet packet or a 90 byte 10 Mbps ethernet packet.

So jumbo frames on gigabit ethernet at worse add less delay variation than 1500 byte frames do on slower ethernets. And no one is suggesting that slower ethernets use 9000 byte frames.

As for queueing delay concerns, that could happen whether packets are large or small. If delivery QoS is required, than the routers need to implement some kind of priority or expedited forwarding, regardless of the packet sizes. Tiny frames (including 53 byte ATM cells) may be helpful when multiplexing lower bit rate streams, but they become increasingly ridiculous on gigabit and beyond links.

Does GigE have a place in a NAP?

Not if it reduces the available MTU!
Network Access Points (NAPs) are at the very "core" of the internet. They are where multiple wide area networks come together. A great deal of internet paths traverse at least one NAP. If NAPs put a limitation on MTU, then all WANs, LANs, and end systems that traverse that NAP are subject to that limitation. There is nothing the end systems could do to lift the performance limit imposed by the NAP's MTU.

Because of their critically important place in the internet, NAPs should be doing everything they can to remove performance bottlenecks. They should be among the most permissive nodes in the network as far as the parameter space they make available to network applications.

The economic and bandwidth arguments for GigE NAPs however are compelling. Several NAPs today are based on switched FDDI (100 Mbps, 4 KB MTU) and are running out of steam. An upgrade to OC3 ATM (155 Mbps, 9 KB MTU) is hard to justify since it only provides a 50% increase in bandwidth. And trying to install a switch that could support 50+ ports of OC12 ATM is prohibitively expensive!

A 64 port GigE switch however can be had for about $100k and delivers 50% more bandwidth per port at about 1/3 the cost of OC12 ATM. The problem however is 1500 byte frames, but GigE with jumbo frames would permit full FDDI MTU's and only slightly reduce a full Classical IP over ATM MTU (9180 bytes).

A recent example comes from the Pacific Northwest Gigapop in Seattle which is based on a collection of Foundry gigabit ethernet switches. At Supercomputing '99, Microsoft and NCSA demonstrated HDTV over TCP at over 1.2 Gbps from Redmond to Portland. In order to achieve that performance they used 9000 byte packets and thus had to bypass the switches at the NAP! Let's hope that in the future NAPs don't place 1500 byte packet limitations on applications.

What about GigE on the campus?

The Gartner Group predicts that 95% of all large-enterprise LAN backbones will be based on high-speed ethernet technology by 2002. Cost, bandwidth, compatibility, and easy administration are all driving this.

So the technology will be there, but it shouldn't come at the cost of our future wide area performance.

If you want high performance, the best network design advice that I can give for a campus, regardless of the networking technology being used is this: Every host in the campus should have a path between it and the wide area network that,

1. does not reduce the link bandwidth, and
2. does not reduce the MTU.

So if, e.g. a host has an OC3 ATM interface running Classical IP, there should be an end to end path between it and the WAN of at least OC3 speed and that supports at least a 9000 byte MTU; every FDDI host should have at least a 100 Mbps 4000 byte MTU path, etc. Wherever you have installed jumbo frame GigE, you should have a jumbo frame path to (and through) the Internet.

Summary:
Gigabit Ethernet needs Jumbo Frames

If you intend to leave the local area network at high speed, the dynamics of TCP will require you to use large frame sizes. Without them, the packet loss rate over a high bandwidth-delay product path would have to be extraordinarily low. Core internet infrastructure, from campus backbones to Network Access Points (NAPs), should be particularly careful not to limit the permitted MTU to 1500 bytes. In the long run there is no reason to stop at 9000 byte frames, but given the current ethernet CRC limitation it is a good evolutionary step for gigabit data rates.

References

[1] the nature of the beast: recent traffic measurements from an Internet backbone
[2] Extended Frame Sized for Next Generation Ethernets - a white paper by Alteon
[3] The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm - the performance equation used above
[4] Jumbo Frames? Yes!

ABOUT THE AUTHOR: Phil Dykstra is the Chief Scientist of WareOnEarth Communications, Inc., and heads the San Diego office which is focused on the measurement and analysis of high performance networks. Prior to WareOnEarth, he was the head of Advanced Development in the Army Research Laboratory High Performance Computing Division. With over 20 years of Internet and HPC experience, he fostered US Federal Networking for many years, and until recently, co-chaired the Joint Engineering Team (JET) which coordinates Next Generation Internet (NGI) and Internet2 engineering and plans. He has taught Computer Science courses at Johns Hopkins University and the University of Delaware.