The following paper was originally presented at the
Ninth System Administration Conference (LISA '95)
Monterey, California, September 18-22, 1995

From Twisting Country Lanes to MultiLane Ethernet SuperHighways

Stuart McRobert
Department of Computing, Imperial College, London

For more information about USENIX Association contact:

1. Phone: 510 528-8649
2. FAX: 510 548-5738
3. Email: office@usenix.org
4. WWW URL: http://www.usenix.org

From Twisting Country Lanes to MultiLane Ethernet SuperHighways

Stuart McRobert – Department of Computing, Imperial College, London

ABSTRACT

This paper describes a slightly different approach to solving network capacity problems between workstations and servers by significantly increasing the number of conventional Ethernet interfaces on each server from just a few to typically a dozen or more. So rather than installing a single faster network backbone (e.g., FDDI, ATM, Fast Ethernet, etc.) to carry all the traffic to and from the servers, coupled with some form of step down hubs to connect to the local workstation Ethernets, our approach bypasses the backbone completely and brings many local Ethernets directly to each of the servers (typically Sun Sparc Station 10s or 20s). This technique has worked very well for our size of operation with several file and CPU servers, 50+ workstations and around 100 X-terminals, with still room for some further expansion too.

Over the past year this approach has been very successful in our main teaching laboratories, significantly reducing network congestion and providing many more well connected networks to support both existing and additional workstations and X-terminals, yet with fewer clients per network, so easing local network contention problems. This, coupled with enhancements to the workstations and servers themselves, has yielded significant performance improvements all round and made for much happier and contented users.

Early Days – Twisting Country Lanes

A long time ago a colleague and I very carefully installed a pair of VAX 750s as the first hosts on our new Ethernet – a thick yellow heavy coaxial cable that ever so gently snaked its way around under the computer room floor – such care with a networking cable was probably never shown again! But users soon discovered how easy and convenient a rich set of new remote access commands were to use, e.g., *rcp*, *rlogin*, and *rsh*, and just how amazingly fast they could now transfer data between hosts. Meanwhile local file transfers successfully moved from the *uucp* tty port based era (cf. countryside foot paths) to this new amazingly quick single Ethernet (cf. a quiet single country lane). Incidentally *uucp* soon fought back for a while by offering queued user file transfers over Ethernet, which still appealed to some users.

Demand for Ethernet connectivity from research groups quickly became virtually unstoppable, almost like the modern day rush to get onto the *Internet* – everybody wanted to be connected. Fortunately for us there was just one moderating factor – cost.

Meanwhile the capacity of the network was at that stage never considered to be an issue, after all Ethernet had 10 Mbps bandwidth compared with only a few 19.2 Kbps tty circuits used before – capacity was almost considered to be infinite. However, as the single thick Ethernet cable began to spread, snaking its way out of the computer room and up the building, concerns were soon raised about its vulnerability to both physical and electrical damage. These were soon laid to rest with the installation of network repeaters on each floor, but fortunately this never became a real problem (Figure 1). The network was also extended via a bridge across campus, complete with our original and officially registered Class B IP network number, although fun and games

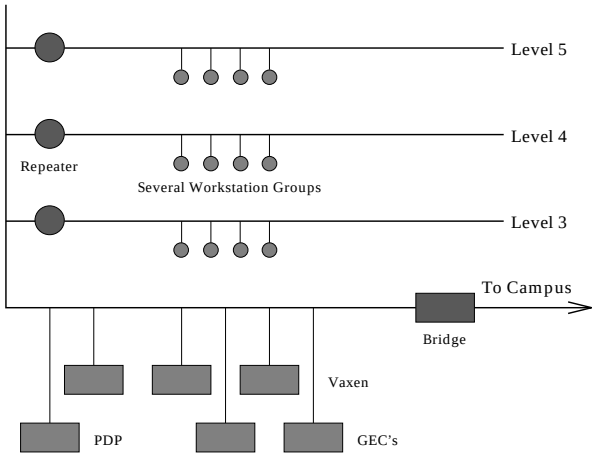

Figure 1: Early Departmental Ethernet (circa mid-80s)

were played with the netmask, eventually settling on a rather interesting 7/9 bit split which although politically acceptable caused no end of grief with various pieces of software. The bridge was a wise investment in terms of providing a surprising degree of protection and isolation from some rather strange and otherwise campus wide networking disasters. It

was however rather slow at forwarding packets (an issue we will return to later), but for now this wasn't much of a problem or concern since most host interfaces were also rather slow too.

With the Department's research groups successfully networked, our attention now turned to advancing teaching facilities and central services. After great debate and a lengthy search, a new powerful twin CPU Gould Powernode 9082 was purchased to act as our new central server. This system was great at handling I/O and since it was also very much more powerful than any of our other computers (both then or for the next few years) it soon took over nearly all central services. However, it did become a classic single point of failure, something that strongly influenced our later drive towards a far more distributed and fault tolerant or at least fault limiting approach. Ten 4 MB Sun 3/50 disk less student workstations and a small Sun 3/160 file server chiefly for Yellow Pages (YP, now NIS) and Network Disk support (ND was needed for booting workstations, root and swap areas in those days) were purchased for the undergraduate teaching laboratory.

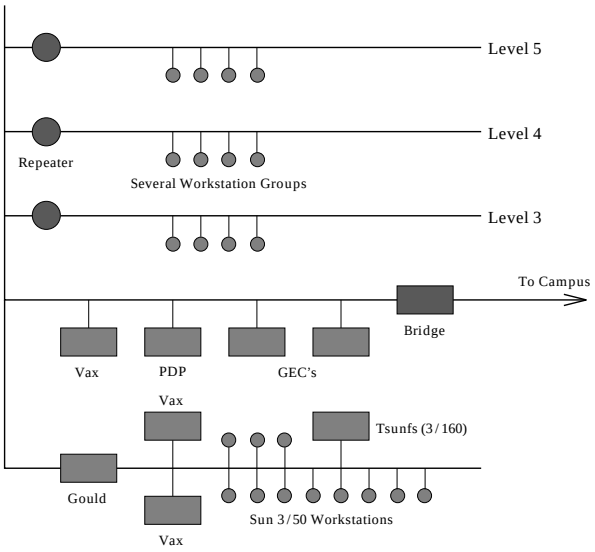

Figure 2: Departmental network including teaching (circa 1987)

However, for the first time concerns were raised about Ethernet's capacity to handle large traffic levels typically generated by many disk less workstations, especially due to all the additional paging and swap traffic produced because of a lack of enough workstation memory. So it was wisely decided to introduce a new teaching network rather than risk overloading the existing network any further. Probably the most interesting and significant choice at that time was the decision to connect this new network via a second Ethernet interface on the new central server, so creating the first of many multi-homed hosts (Figure 2). The Gould was superb at handling I/O and could easily and efficiently handle the extra traffic and still make good use of having access to twice the network bandwidth, in fact it later gained a third Ethernet interface and still coped well.

Over the next few years the number of disk less teaching workstations more than doubled with many additional Sun and HP workstations, along with several multi-purpose severs often with twin Ethernet interfaces, for both CPU and file serving work (Figure 3).

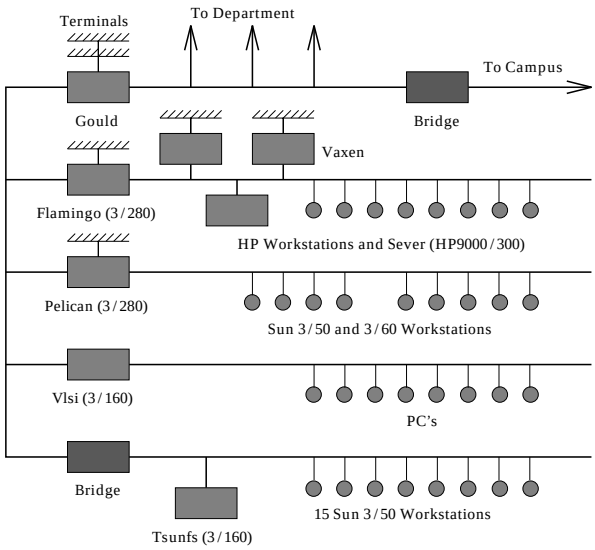

Figure 3: Teaching network (circa 1989)

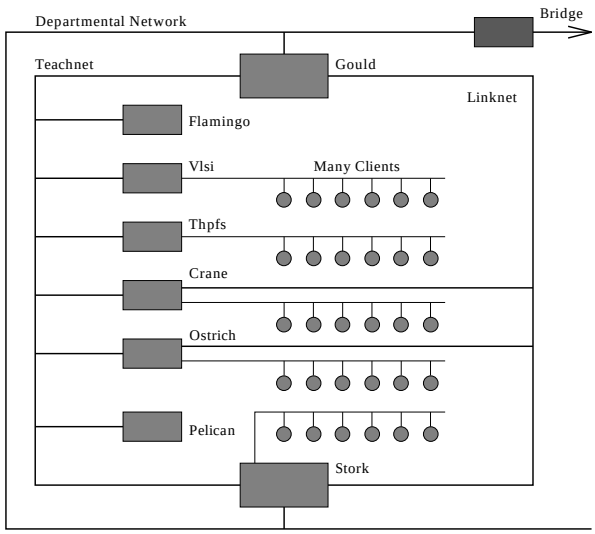

Figure 4: Teaching network (circa 1990)

Such growth continued especially as the use of glass ttys dwindled and graphical workstations proved highly successful, but traffic levels on the teaching networks rose at an alarming rate, coupled with high network collision levels during the ever lengthening peak periods. A full discussion of the problems faced at that time is outside the scope of this paper,

but can be found in [SunUG'91]. However it is worth noting the key changes in network topography carried out at this time to better cope with the ever rising traffic levels (Figure 4).

A new server (see *Stork* in Figure 4) with four network interfaces was introduced and along with *Crane* and *Ostrich* were the first Sparc servers purely dedicated to serving, i.e., they supported no user logins, and were locally known as Network Support Nodes or NSNs for short. They provided the workstation users with a much better response since their CPUs were never tied up with user jobs and had fairly good network connectivity to the other servers. For now they also had a speed advantage over the earlier generation of workstations, something that wouldn't last for long. Note that the bridge used earlier to help ease network congestion has been removed, since it was actually found to cause more of a network bottleneck than a help, since it was unable to forward packets at anything like network speeds (of course bridges today generally can and easily do achieve such performance).

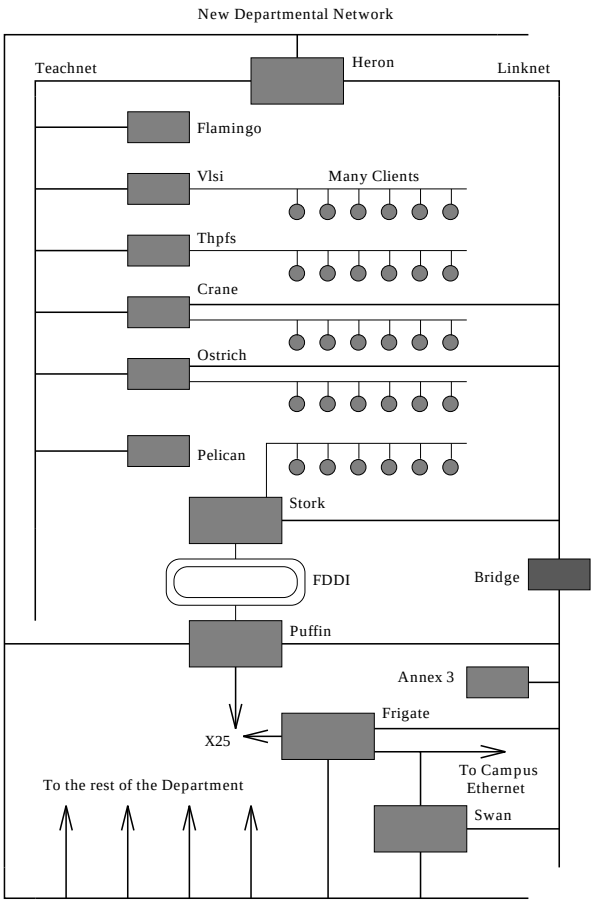

Figure 5: Departmental network overview (cica 1991)

Finally in 1991 the *Gould* came to the end of its life, chiefly due to reliability problems, but the main lesson learnt from it was to avoid at all cost designing any part of a system with a possibly devastating single point of failure, since not only would failures be a problem, but also routine work like software or hardware upgrades too. The *Gould* was replaced by several distributed systems, three Sun Sparc Station 2s (*swan*, *heron* and *frigate*) took over most of its work and the old Sun 3/160 file server was upgraded to a Sparc 4/360 and renamed *Puffin*. An experimental FDDI network ran between *Puffin* and *Stork*, but mainly due to hardware costs it never took off as a possible new departmental network backbone. Figure 5 shows the network overview from that time with a second departmental Ethernet installed to help ease some of the backbone congestion problems seen at that time.

In summary, by this stage multi-homed dedicated servers had been shown to be a good idea, especially where the hardware was capable of easily sustaining the I/O rates required. Single points of devastating failure needed to be avoided wherever possible, hence the distributed approach was much better for most services (e.g., spreading home directory and replicated /usr type file systems, mail, external communications, adequate network routing with alternative routes, etc.). However it has to be recognized that there are additional system management overheads in terms of keeping everything consistent across multiple servers and platforms, but it is possible and tools do exist to help (e.g., *rdist* and *track*).

Meanwhile everyone was buying cars, sorry workstations, bigger faster workstations with higher performance Ethernet interfaces. As a result more and more locations were being networked, the backbone networks were becoming congested carrying an ever increasing amount of traffic, and network cables just seemed to mushroom everywhere. A good few miles of thick and thin Ethernet cable typically ran from the computer room and up the building risers, even filling them to capacity in places, and then off along the corridors to various rooms. Of course physical cable navigation was just as skilled as map reading (where there were maps) and identification signs just as rare and accurate as old road signs at remote country road junctions, e.g., two roads/cables going in different directions to the same place! (and quite correctly labeled when installed). Things change, just as roads get bypassed so do network cables, and just to make things worse there are all those thick Ethernet drop cables too, and the one you have to trace always seems to go for miles crossing several others in its path until you take the wrong turn and follow the wrong one – really just like *twisting country lanes*.

The Problem – Growing Pains

From now on let us mainly concentrate on the teaching side of the departmental network, since it is far more interesting! Having established the idea of

dedicated Network Support Nodes (NSNs) to look after groups of workstations, whilst the NSNs were themselves all well connected to both teaching backbones for good network access, e.g., to all home directories and central services like mail and news, now was the time to expand this successful idea even further (Summer 1991).

Two additional Sun Sparc Station 2 file servers (SS2s) allowed us to significantly improve student NFS file serving by spreading student home directories over three instead of one file server (*Heron*, *Toucan* and *Lorikeet*), all transmitting data to and from clients via the existing two teaching backbones (teach and link net, Figure 6). *Heron* also acted as a second route to the main departmental network. In addition, the two new file servers were also directly connected to a mixture of nine Sun mono ELC and ten color IPX workstations, all with local 207 MB

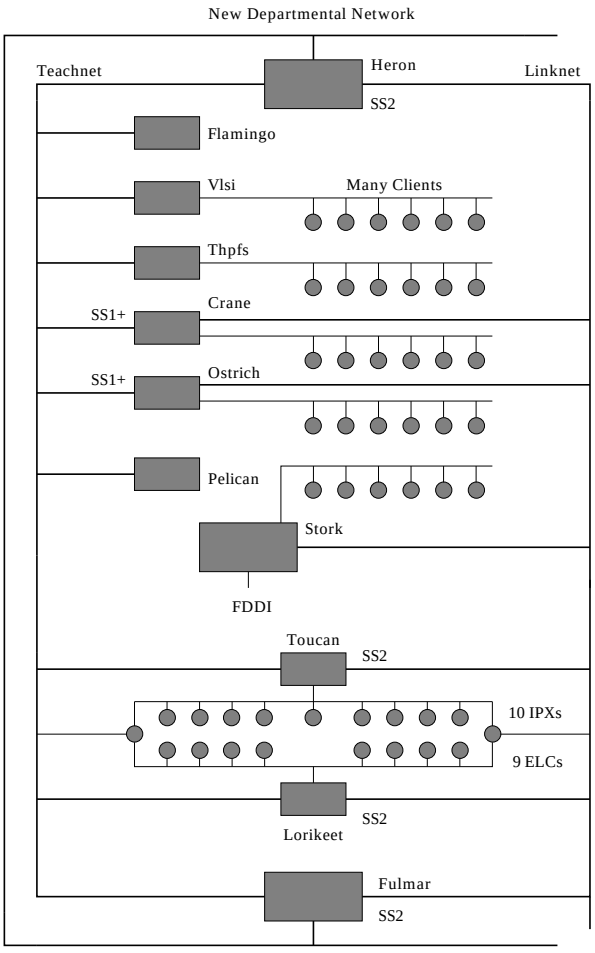

Figure 6: Teaching network Autumn 1991

disks. Just like the first NSNs and their disk less clients, these SS2s locally provided everything their workstations might need and couldn't be stored on the workstation's somewhat limited capacity local disk. However, for student home directory requests, one third would be available locally whilst the remainder would have to come from one of the other two file servers, which were never more than one network hop away from the client workstation. As backup routers, a pair of IPXs also had additional network connectivity to allow the service to be reconfigured should the need ever arise. All in all this solution worked well.

Over the next couple of years another ten IPXs were added with bigger local disks and more memory, but with no additional networking capacity the network soon started to show signs of strain. High levels of collision rates returned, and overall the system was approaching its design capacity. Meanwhile the number of Sun 3s now being used as X-terminals also steadily increased, adding to both network and workstation CPU burdens. Further expansion in the form of three additional Sun Sparc Station 10s (SS10s), two as central CPU servers (*Finch* and *Motmot*) to help improve X-terminal response and one as an upgrade for file server *Heron*, soon took the network at times to near breaking point. Also by then many of the workstations and servers were well under configured for the teaching load now being imposed on them, and so the quality of service degraded, especially at peak times. Not surprisingly users and support staff were increasingly less than happy with the system, especially with the obviously overloaded networks, but not all fully understood why.

Now was the time to study the problem and find a cost effective solution, since further expansion was called for and clearly the existing networking structure could no longer cope.

MultiLane Ethernet SuperHighways

But Why Ethernet?

Early on in the design stage of this project it was recognized that the only viable solution to delivering networking to the desktop was to remain, for now, with Ethernet technology. Quite simply many of the older workstations and X-terminals couldn't accept anything else, whereas for those newer ones with expansion slots available, the costs involved in equipping whole teaching labs with faster interface cards (be it FDDI or its copper based equivalent CDDI, or even Fast Ethernet) was prohibitive and also of questionable benefit considering the overall power of the systems involved. However, any new physical network wiring to the desktop is now installed and fully tested to 100 Mbps specifications, i.e., UTP category 5, making much of the cabling system ready for faster networking whenever it does arrive, be it either of the Fast Ethernet standards, CDDI or even ATM.

Furthermore, there was no perceived need nor support for general workstation networking faster than 10 Mbps, we just needed to get the existing technology working well. Another big plus for continuing with Ethernet was the assimilation of

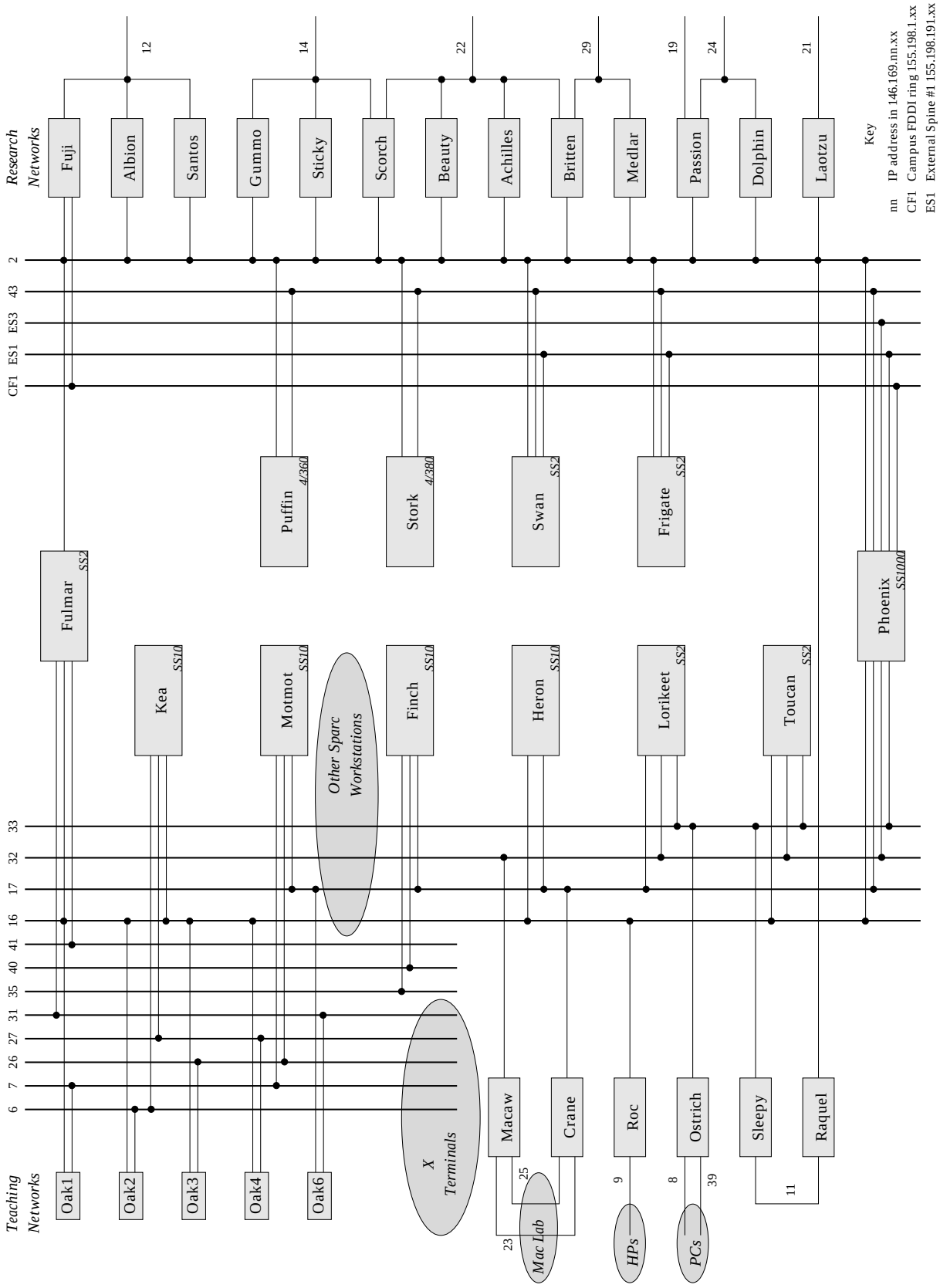

Figure 7: Departmental Network, prior to any changes, Summer 1994

several years practical experience – a considerable asset, especially in terms of rapid problem resolution, traffic capacity planning, and overall understanding – that feel good factor. On the other hand, the extensive deployment to every desktop of some new networking technology, even if previously used in a backbone environment, and however good, carried with it a much higher risk of the unknown – we preferred to minimize such risks.

The Problem Revisited

Having accepted that we would stay with Ethernet to the desktop, the next big issue was how to connect the file servers to their client workstations.

Previously the network had been organized in such a way that no server was ever more than one network hop away from any client. Initially this seemed to be an acceptable compromise between the number of direct network connections (chiefly limited by the number of suitable host server expansion slots), and backbone connections (bandwidth). It was certainly a vast improvement over earlier network topographies.

Originally there was just one coaxial teaching network backbone, extended (by those that knew better) to also support some workstations elsewhere, cost effective perhaps, but when it broke remotely one day, the whole of teaching stopped. Of course one might say that nowadays with UTP wiring and the hubs used now this wouldn't be a problem – but hubs do fail, UTP cables get damaged, the wrong interface gets connected to the wrong hub – things can and do go wrong. So this dual backbone approach remains with us to this day – belt and braces perhaps, but the extra resilience has proved itself time and again to be very well worth having.

In many respects the no more than one hop approach was actually quite good since it also allowed us to successfully implement the idea of distributing class file serving, spreading each teaching class over multiple file servers rather than confining them to a single specific host. Although one might now consider such an approach to have rather obvious advantages and to be the only cost effective scalable approach, resistance stemmed from two areas. The first was financial, where a single new server had been funded for a specific class, and the second was reliability. Not so long ago computer hardware wasn't nearly as reliable as it is today, and so it was felt only fair that should a fault occur the whole class should suffer equally, otherwise some students might have an unfair advantage when it came to marking over others. Fortunately we were able to happily resolve both issues and reap the obvious advantages with few long term difficulties.

However, the one hop approach doesn't scale well with an increasing number of users or parallel teaching (serving two classes at once instead of just one), since as the number of servers and workstations increased, fewer and fewer users found themselves sitting in front of a workstation with direct access to their home directory file server. Furthermore, the number and power of workstations always tended to increase faster than the power of the server(s) assigned to support them. So as the servers became even busier, the latency through them rose significantly, such that even a hop count of one, was one hop too many.

What was generally happening at a user's workstation was that it would try to route a NFS request over the local Ethernet via a busy locally attached file server, which would eventually route it out over one of the two busy busy backbones to the desired file server, which would then reply, or at least try to.

Meanwhile, back at the workstation, things would be going rather slowly, retransmission of UDP packets would be sent out which would again have to be handled by the servers, so increasing network traffic and collisions along with server load. The poor users simply received a worse response and they tended to load balance the system, hopefully coming back later. This wasn't good since full use of facilities wasn't possible nor could demand be adequately satisfied.

So far it would appear that all our problems were chiefly network related, but in fact the workstations themselves were a major contributor to the problem since they were under configured for the tasks now being performed. The most glaring problem was inadequate local disk space and physical memory, resulting in increased network traffic to and from the servers since frequently required pages were flushed rather than remaining locally cached. Workstation configurations would also need to be improved.

There was also a requirement to expand the number of workstations being supported and improve the performance of both the CPU and file servers. Better access to the CPU servers from a large number of X-terminals (based on old Sun 3s) also required urgent attention.

Alternative Solutions

The most obvious solution to improving network performance would be to install a significantly faster backbone, say FDDI/CDDI, or Fast Ethernet or just maybe early ATM. Staying with Ethernet speeds but using an Ethernet switch was also considered, as was the need for file server independent routing, e.g., an additional direct connection of each workstation subnet to a network hub. We also needed to improve the ratio of the number of workstations per Ethernet, i.e., have less workstations per network, which along with the installation of new workstations would require significantly more Ethernet subnets be connected with good network access to the servers.

All this was possible, but so far most solutions also required an expensive network hub, and that required money that was difficult to find. Although the entry price didn't seem too high, by the time one had included all the necessary interfaces for the number of Ethernet networks desired to adequately support all the workstations, they all looked very expensive solutions indeed.

Overall it was preferred to spend funds on workstations and servers, along with more disk space, memory, etc. rather than on an albeit a very high performance network hub, yet still find a network solution to allow good use of the facilities purchased.

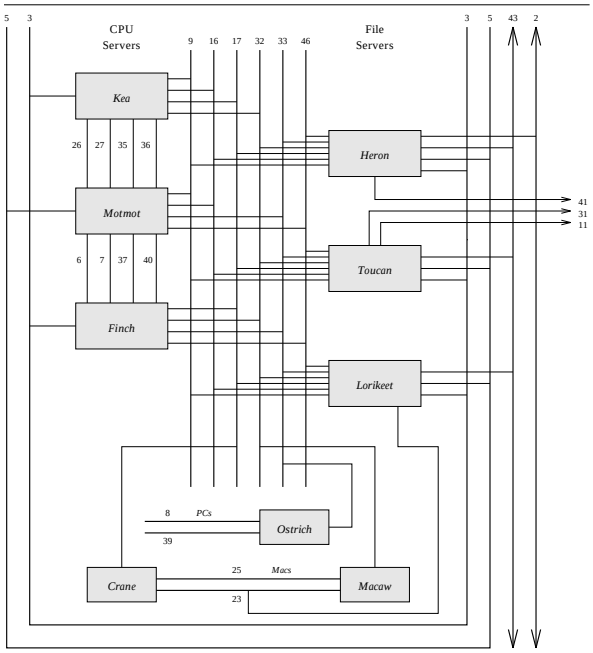

Figure 8: Teaching network Spring 1995

An Ethernet SuperHighway

One promising solution arose from the the idea of significantly increasing the number of workstation networks and directly connecting each network to every file server. This was now possible by upgrading the file servers from old Sun SS2s to SS10s, which were much faster and had four instead of three Sbus slots, and utilizing Sun quad buffered Ethernet Sbus cards which provide four UTP Ethernet interfaces per Sbus slot. Typically each file server would now have two or three quad Ethernet cards plus possibly a combined fast SCSI and buffered Ethernet Sbus card, giving the server two fast SCSI buses (for disk and tape drives) and up to 14 UTP Ethernet interfaces. A dozen of these networks could then be used for workstation subnets since two were still retained for backbone connectivity. Currently each file server only has two quad cards installed, giving a total of around 10 UTP connections, and the remaining Sbus slot remains free for future expansion.

Having such a large number of workstation networks available allowed for a significant reduction in the number workstations per network, down to around eight, dramatically improving the available bandwidth per workstation and helping to reduce the previous excessive levels of network collisions.

The number and power of student workstations also increased dramatically. The main teaching laboratory now supports 18 new 48 MB Sun SS5 color workstations with ½ GB local disks and eight new 64 MB HP 712/80 color workstations with 1 GB drives, along with two 64 MB SS20-SX multi media workstations. In addition ten of the newer IPXs were upgraded to 52 MB of memory and nine ELCs to 24 MB, whilst the remaining ten older IPXs were redeployed for research student use elsewhere.

The direct network access idea has also been extended to supporting X-terminals by placing their networks between pairs of powerful 128 MB Sun Sparc CPU servers *Kea*, *Motmot*, and *Finch* (two SS20s and one SS10), also using quad cards to increase network connectivity and provide direct access to all the file servers and many of the workstations too. Two quad cards are typically installed in each CPU server and eight dedicated X-terminal networks have been constructed. This solves the two old problems of having too many X-terminals per network (now down to an average of 16 mono or 8 color per network), and poor network access to both workstations and dedicated CPU servers.

Twisted Networks

Quite obviously with this amount of networking there is a corresponding large volume of network wiring. Fortunately this is now all UTP which is much easier to handle and much quicker to reconfigure and install, especially in bulk as structured wiring. The workstation end is conventional enough UTP wiring not to merit comment, except that purely on cost grounds small UTP-to-Thin converters are used to wire benches of old Sun 3s where conversion to UTP couldn't be cost justified.

The server end is much more interesting. Typically servers are installed on shelves inside 19 racks, with locally constructed SCSI disk trays on either side. Since there is a very high demand for UTP connections, bulk UTP wiring is run under the computer room floor from the central hub area to each server rack, where it is terminated at a 110-block. From there manufactured UTP patch cables are cut in two and the cut end punched down on the 110-block, providing a fully compliant (and tested) Category 5 cabling system that terminates in a standard RJ-45 which can then be plugged into the server. The other end is just as easy, but instead of a server there are banks of either SNMP managed hubs for the key networks, or cheaper unmanaged ones for less critical uses, e.g., X-terminal networks.

The whole installation was completed over several months, a couple of them very busy involving quite a complex set of phased network changes to allow new networks and services to be gradually phased in whilst others were smoothly removed to be redeployed later. About the only key software aspect worth serious note is that in order to make sensible use of such a highway, lane discipline is very important. DNS needs to return the local IP address of the name requested from the point of view of the local workstation asking, otherwise needless IP routing can and will take place. Also some non-UNIX software can't handle the concept of a host having a dozen or more IP interfaces, shame, but we generally created an alias.

The Results

Quite simply it worked phenomenally well, first time, no problems, good old Ethernet! In fact very few people actually realized what had been done, and apart from a few students who studied the host tables and didn't believe them at first, there have been very few comments. There hasn't been one complaint about network response attributable to the local teaching networks, and it was a cost effective solution delivered on time and within budget.

The Future

The original design has room for further expansion both in terms of supporting more workstations (file serving) and X-terminals (CPU serving). Plans are currently underway for a new student project laboratory which will hopefully integrate well with the existing facilities. Fast Ethernet could also be a very interesting hot topic, especially since the two competing standards have done a lot to bring this technology quickly to market as a working deliverable product. Currently hub prices continue to fall and interface cards are readily available on many platforms, and it will happily run over our existing networking infrastructure, so just plug-and-play. Meanwhile ATM slowly moves through various committees, maybe one day.

Conclusions

This Ethernet SuperHighway approach quickly provided an expandable, cost effective, highly integrated, fast, low congestion and latency, direct (zero hop) connection for each and every workstation to all the teaching file servers. It also directly connects X-terminals between powerful CPU servers, which themselves have multiple direct connections to all the file servers as well. In addition the design is also reasonably network fault tolerant and damage limiting in terms of what becomes unavailable should any single component fail.

It has worked very well and is indeed a very simple solution. Best of all it has many happy users, and room for further expansion to hopefully keep them that way. All in all it has been one of those great behind the scenes successes.

References

[SunUG'91] Stuart McRobert, Divide and Conquer, *README*, Sun User Group, Vol. 6, No. 3, Fall 1991; also in the Sun User Group Conference Proceedings, Atlanta, GA, June 1991.

W. Richard Stevens, TCP/IP Illustrated, Volumes 1 and 2, 1994, 1995, Addison Wesley.

Author Information

Stuart McRobert received his BSc and College prize in Physics at Imperial College London in 1982, where he is now Head of Systems and Chair of Netman (the local Network Management team) in the Department of Computing there. His work has moved from the support of individual systems of the PDP/VAX era, through several local networking firsts (Ethernet, FDDI, UTP wiring) along with the introduction and management of client/server computing and overseeing its subsequent growth into the highly distributed multiprocessor systems of today. He has also been involved in the installation of large parallel systems including a Fujitsu AP1000, and along with a colleague, in their spare time, manages *Sun SITE Northern Europe*, one of the larger and rapidly expanding archives on the Internet, which will be extensively involved in next years *Internet 1996 World Exposition.*

He can be reached by post at the Department of Computing, Imperial College, 180 Queen's Gate, London, UK, SW7 2BZ, or preferably via email to sm@doc.ic.ac.uk .