Internet Architecture Workshop: Future of the Internet System Architecture and TCP/IP Protocols[1,2]

David L. Mills, INARC Chair; Paul Schragger and Michael Davis, Editors

University of Delaware
Newark, Delaware
1-2 June 1989

The Internet Architecture Task Force (INARC) is a unit of the Internet Activities Board (IAB), which oversees the research, development and standards of the Internet community and TCP/IP protocol suite. This is a report on the INARC workshop held at the University of Delaware on 1-2 June 1989 at which the future of the Internet and its protocols, as well as their relevance to other protocol communities, was explored and debated.

Introduction

This is a report on a workshop convened to discuss and debate the future of the Internet architecture and TCP/IP protocols. At the direction of the Internet Activities Board (IAB), the Internet Architecture Task Force (INARC) tracks significant happenings, research activities and evolutionary forces shaping the Internet architecture and its protocols. This workshop concentrated on the architectural and protocol issues that the Internet must face in the years to come. The lessons learned should serve as guidance for research planning, policy formulation and system engineering for the Internet community of the future.

The first day of the two-day workshop included several invited or volunteered formal sessions which exemplify current technology and policy issues. Due to the large number of volunteered informal presentations, the first day concluded with a number of fifteen-minute mini-sessions. The second day included four panel discussions in which panelists presented some aspect of the topic and other panelists and members of the audience discussed it at length.

In this report, which is condensed from the original 45-page edited transcript, the words are the editors', unless noted to the contrary. While an effort was made to balance the emphasis on all presentations, some of them were more provocative than others and generated much heat and some light. In the concise summaries to follow the editors tried to capture the full intent of the speakers; but, sometimes, using words that never actually came out of their mouths.

Day 1: 1 June 1989

SESSION 1

Navigation Aids for the Future Internet

David Mills, INARC Chair, U. Delaware

Presenter's summary: In order to set the tone for this workshop we need to mark regions on the map which are only beginning to be explored by Internet navigators. This presentation raises several issues designed to provoke interest, stimulate discussion and foster debate in the remaining sessions and panels. The following list of objectives are presented as guidelines for debate.

1. Conventional wisdom cites visualization, remote sensing and national filestores as drivers for huge and fast. The machines which require such speeds are the supercomputers and earth stations of today and the workstations and space stations of the future. Are the Internet architecture and protocols suitable for use on very high-speed networks operating in the 1000-Mbps range and up? If the network-level or transport-level protocols are not usable directly, can they be modified or new ones developed to operate effectively at these speeds?

2. We occasionally see cases of Internet routing bobbles, meltdowns and black holes, even with only 700 nets and uncoordinated backdoor paths which invite sinister routing loops. Are the Internet addressing and routing algorithms adequate for very large networks with millions of subscribers? If not, is it possible to extend the addressing scope and/or develop new routing paradigms without starting over from scratch?

[1] Sponsored by: Defense Advanced Research Projects Agency contract number N00140-87-C-8901 and by National Science Foundation grant number NCR-89-13623.

[2] Reprinted from: Mills, D.L., P. Schragger and M. Davis. Internet Architecture Workshop: Future of the Internet System Architecture and TCP/IP Protocols. *ACM Computer Communication Review 20, 1* (January 1990), 6-17.

3. Can the Internet model of stateless networks and stateful hosts be evolved to include sophisticated algorithms for flow management, congestion control and effective use of multiple, prioritized paths? Can this be done without abandoning the estimated 60,000 hosts and 700 networks now gatewayed in the system? Should we evolve to more stateful designs in order to embrace new principles of flow and policy management?

4. The routing technology of the future will be coupled less to available hardware and routing algorithms than to issues of resource utilization, cost recovery and administrated access. Can the existing Internet of about 300 routing domains be evolved to support the policy and engineering mechanisms for many thousands of domains including education, research, commercial and government interests? Can this be done with existing decentralized management styles and funding sources? If not, what changes are needed and how can they be supported, given practical limits on infrastructure funding?

SESSION 2

The Next Generation of Internetworking

Guru Parulkar, Washington U.

Editor's note: This session consisted of a presentation and discussion of the paper "The Next Generation of Internetworking," which appears elsewhere in this issue.

SESSION 3

Fair Queueing for Unfair Gateways

Scott Shenker, Xerox PARC

Editor's note: This session consisted of a preview and discussion of the paper "Analysis and Simulation of a Fair Queueing Algorithm," by Alan Demers, Srinivason Keshav and Scott Shenker from Xerox PARC. The full text of the presentation is given in the SIGCOM 89 Proceedings. The following abstract is reprinted with permission.

"Gateway queueing algorithms have a role in controlling congestion in datagram networks. A fair queueing algorithm, based on an earlier suggestion by Nagle, was proposed. Analysis and simulations were used to compare this algorithm to other congestion control schemes. They found that fair queueing provides several important advantages over the usual first-come-first-serve queueing algorithms: fair allocation of bandwidth, lower delay for sources using less than their full share of bandwidth, and protection from ill-behaved sources.

SESSION 4

Service Requirements from the Science Community

Peter Shames, NASA Space Telescope Science Institute

Presenter's summary: The national and international space research communities have significant data management and network concerns for the near term future because of present and planned launches of major new science data instrumentation. The networking needs of the space science community are driven by requirements for the integration of information from many different discipline sources, for visualization mechanisms that can operate on large data cubes, and by data transport, storage, display, and computational resources to handle the data volumes. In addition, the proposed environment which will support collaboration on cross-disciplinary, multi-agency, and multinational science projects requires ubiquitous service between all member sites.

However, scientists are faced with consideration of a major tradeoff between funding for science or funding for information support infrastructure. The science community must seriously examine cost/benefit tradeoffs when dealing with network development and use. If new services are too costly then they will be done without or alternatives will be used. While new distributed systems offer potential benefits in convenience and can enable fundamentally new kinds of science, these systems may not be achieved if both the network infrastructure and the science information system infrastructure are not adequately funded.

The space station and instrument platforms are specific examples of projects that require massive amounts of bandwidth. The data volumes from some instruments will be in the range of 100-300 Mbps. In addition to these high data flow rates, many of the proposed projects will require remote control of instruments in space from laboratory and university sites. Consideration must be given to both the need for secure and safe remote control paths and protocols, and for reliable, low latency, communications to distributed sites.

The many terabytes of data collected from these instruments and space platforms will be stored and maintained in online archives. These archives will be accessible to all collaborating individuals, who may be at any of several remote locations. So, in addition to access to the data, other policy issues must be resolved (such as privacy and routing) in terms of the multi-agency, multinational, and public-private carrier environment that is being nurtured.

Other areas of scientific network needs include access to supercomputer sites, control and monitoring of terrestrial experiments (such as accelerators and telescopes), and access to other unique resources that may become available online (libraries, information servers, laboratories). In addition there is the desire to have multimedia text, graphics, image, voice, and video services to enhance the exchange of ideas and knowledge among colleagues.

The major system issue that must be faced by the community is that this network infrastructure must be long lived, large scaled and support a heterogeneous, distributed environment. The complexity and longevity of the systems requires planning for growth and evolutionary change. These systems will be based on a variety of vendor platforms and be widely dispersed. There is a need for the current community to accept and develop "de-facto" standards in a number of key areas (e.g., interface protocols, data and media formats for interchange), and also plan for the anticipated use of the ISO/OSI standards.

The areas that require technological developments can be placed in three categories: standards, databases, and workstations. In the standards arena we need international cooperation between all protocol-specifying and policy-making bodies. Some of these coordinations are now occurring at the level of ISO and ANSI committees, some are occurring in space-related venues such as the Consultative Committee for Space Data Systems (CCSDS), formed of the appropriate agencies of key space-faring nations.

In summary, there are a number of identified concerns and challenges that must be met before the space science community will have the network environment that they desire. Since networks compete with science for dollars, the advantages of networks must be demonstrated with realized systems accessible to a broad spectrum of users. Everyone understands the need for network mail, yet few understand the opportunities for distributed systems, largely because few (no) public ones exist.

Common distributed system interfaces and protocol standards are extremely important for overall system success. As new services such as archives, science databases, remote-controlled instruments are brought online there is an enormous opportunity for protocol chaos to ensue. Point solutions to such interfaces are relatively easy to achieve, but the general inter-operability that is desired requires the planned development of robust interface protocols.

Access to science resources by a distributed community require nearly ubiquitous access to networks, regardless of the actual agency or group that is funding the research. However, present agency policies relating to network access and use can adversely affect functionality of existing installations in national and international arenas. Research efforts in policy-based routing can help alleviate some of these problems, when routers implementing such policies can be deployed.

The major planned science activities require an advanced national (and international) network infrastructure if these systems are to be realized. National and international coordinating bodies such as the Federal Research Internet Coordinating Committee (FRICC) and the Consultative Committee on Intercontinental Research Networks (CCIRN) are providing a venue where agreements on cooperative use of networks can be, and are, worked out.

Questions and Comments

Q: Tell us about the FRICC.

A: The FRICC is the Federal Research Internet Coordinating Committee. Its membership consists of a number of national agencies that have major networks: NSF, NASA, DOE, USGS, and HHS. The FRICC provides a venue that sensitizes these agencies to the needs and opportunities of network usage. One of the activities of the FRICC is the Research Internet Backbone (RIB). This is to be an operational prototype of a high speed network that is intended to eventually run at gigabit speeds.

Q: What is the justification of gigabits, wouldn't your needs be satisfied by megabit circuits? Is there a need for gigabits other than visualization?

A: Besides aggregation of all the users, the Jupiter mission is a prime example of the need for large bandwidth network to support international cooperation. There will be multiple megabyte files coming from the mission that requires global collaboration on the data. A material science group wants to follow 1000 frames per second, high definition video at rates to 50 Gps from space. The instruments exist, but a network that can move that data does not exist.

Q. Has compression been considered?

A. Yes it has been considered and the estimates are that the data compression would reduce the bandwidth 2 or 3 times the raw rate at best. The reason for this is that lossless compression techniques must be used in order to preserve all of the information in the original data.

SESSION 5A

The Once and Future Internetwork Survey

Paul Tsuchiya, MITRE

Editor's summary: This is an informal survey and discussion of how the group foresees the future of the Internet and how the nodes and nets will be connected together. Out of the many confusing, individual views, six alternative views of the future Internet emerged:

1. Only common carriers connect all sites. Even in a corporation there are no private wide-area nets.
2. Common carriers (only) connect possibly global private nets to other private nets.
3. Same as (2), but now direct connections between private networks are allowed.
4. Same as (3), but now the private networks can carry transit traffic between other private networks.
5. Common carriers exist but the majority of the network traffic is provided by resource sharing on private nets. This is the view today.
6. Common carriers find no worthwhile return in data networking and become entertainment distribution systems.

SESSION 5B

Open Routing Work in Progress

Marianne Lepp, BBN CC

Editor's summary: The open routing working group (ORWG), which is a unit of the Internet Engineering Task Force (IETF), has constructed a model of the Internet as it may appear in five years. The model involves an architecture for the design of an inter-domain, policy-based routing (PBR) protocol that works with a full mesh topology of autonomous regions (AR). The protocol must work in today's Internet and grow gracefully as the Internet does. This session presented a concise summary of the model and generated a spirited discussion with many participants and questions asked.

Questions and Comments

Q. What about the granularity for forming policy in PBR? Is policy based on users, hosts, networks, ARs?

A: In this model we are using David Clark's view of policy as a generic term used by ARs to provide a uniform administration. There could be a finer granularity that appears within the AR.

Q: Will you exclusively disallow policies that are at the user level.

A: No, the problem that we are addressing is the routing problem not the policy implementation. The ORWG world view is between the common-carrier-only view and the private-net-only view. There will be a number of common carriers, and backbone networks (in the 10's), along with a larger number of ARs (in the 1000's). Most ARs will be stubs off of the backbones. They will generate and receive traffic but will provide either no or limited transit service based on bilateral agreements. The implicit question is if this is our view it has certain design implications, in the sense that we are attempting to optimize for the reasonably simple system. The organization is hierarchical with a fairly ubiquitous set of backbones providing the bulk of the services. But, we have to allow semi-automatic ways to short-cut through neighbors' backyards. We do not want to explicitly introduce restrictions that a region must use a predetermined path. If the protocol is designed correctly it will be able to handle more interconnects than our predicted connection model.

The driving requirements that face the routing protocol is that there will be as many as 10,000 routing entities. The millions of end user addresses will not be visible except locally. There will be a general topology and a group of complicated policies, but not all policies will be supported. This system will be large scale and heterogeneous, involving different boxes, protocols, addresses, and network technologies. Between each of these groups there will be only limited cooperation to provide resources, transit, policy implementation, etc. Security must be addressed in the exchange of protocol information, the installation of policies and the installation of paths. There is a need for privacy of local data, and authentication of outside data. We must remain cognizant of performance issues such as CPU speeds, link bandwidths, and protocol robustness. And finally, all this must be supported and yet remain conceptually simple.

Q: How did the 10,000 entities come about? Could you define the term "routing entities?"

A: The 10,000's number is mostly our intuition developed by talking to knowledgeable people and groups in the Internet community. Entities are what routers must deal with.

Q: How does database size depend on type of service, etc? Is it linear, exponential? There could be major difficulties if it grows too quickly.

A: This is work in progress the database has not been defined. There will be inherent data hiding, so the database size can remain manageable.

The architecture assumptions provide for a number of conceptual elements: A routing agent computes routes and maintains the topology database. A policy agent validates, controls and maintains policies. A forwarding agent provides data forwarding functions. A user agent negotiates type of service, authentication and private policy elements. A data collection agent monitors link and entity status.

One of the features of this model is that data reduction is accomplished through clustering. We use source routing and route caching. There is a route setup that permits the packets to pass through the system without lengthy source route addresses in all packets. The protocol is dependent on link states. Also security issues will be implemented from the beginning of design.

SESSION 6

Policy Issues and Other Subjects

Danny Cohen, USC Information Sciences Institute

Presenter's essay: There are three aspects of policy-based routing (PBR) that we are addressing today: why, what, and how. We have heard quite a bit of how, but very little about why and what. I invite anyone to display a document that tells us why we want PBR.

My understanding is that the reason PBR makes so much sense is that we want to guarantee a certain amount of service to our own people. It is not that we want to be nice to others outside. There are two ways of thinking about using PBR: 1) we may wish to select which networks they use and 2) networks may wish to select which users they serve. They are very different problems.

The following are proposed guidelines for the PBR for the national network testbed (NNT). Any PBR scheme for the NNT should: 1) consider the possibility of improved performance for privileged users, 2) not depend on cooperation or changes by others, 3) not deny service or degrade their performance for others who do not change their systems to comply with the PBR, 4) operate independent of (and possibly uncoordinated with) the PBR implementations of the other agencies, 5) not be expensive (in performance and dollars), and not have any performance penalty during OPEN periods, 6) be non-trivial (not necessarily impossible) to forge and 7) include inter-agency backup, not just limited load-sharing.

What must be changed to implement the PBR? Will we have to change hosts , agents, gateways, protocols and other things not foreseen? When we make plans to implement PBR we should try to change as elements as possible. Therefore we should only address schemes that do not depend on too much change.

At this point let's ask some pointed issue questions. Can we identify a "good-user" explicitly or implicitly? Will there be some IP option for good-user, or will they appear on some good-user list? Is the system passive or active? Does it just keep track of users and send the bill or does it act like a policeman and prevent non-privileged usage? If you cannot accomplish perfect allocation, do we err in the direction of giving service to non-privileged users or in the direction of denying a privileged user. Will access to routes be like tunnels or bridges? Tunnels are hidden and you have to know where to look, while bridges are visible landmarks. Note that there is no typing of users in the exterior gateway protocol (EGP). There is no mechanism to provide for information that is only good for any subset of users. Therefore, since the connectivity under PBR is not uniform and the connectivity under EGP is uniform, source routing looks like a reasonable tool to use.

On the Internet Architecture

In the good old days we knew ARPA ran the network, today we don't know who runs what. This is an application of Conway's Law: "Organizations tend to build systems that reflect their own organizational charts." What we have is lots of organization but no chart, no relations, and we occasionally meet to decide to cooperate in some way. Another observation is that the research community is not always leading the industry. It is as if we follow the saying: "Here go my people. I better run after them because I am their leader." Sometimes under the auspices of research we are examining what is happening already, and not leading the direction.

In the beginning the ARPANET was created. Like many successes it had lots of parents. The ARPANET was made of three components: "stupid" lines, IMPs, and hosts. This provided the purest packet switching. There was absolutely no state information stored, no advance information and no connections. Connections was considered the dirty "C"-word. This system espoused equality; all lines were equal; all IMPs were equal; and all hosts were equal. An interesting observation was that there were no hostless IMPs by choice, not by design. The most beautiful thing about this network is that if there was a possible connection, it was always found by the IMPs. A network may be large/small or fast/slow but it was always a network. A net was a net was a net.

Then came the Internet (notice the capital I). The Internet is a collection of networks connected by split-personality hosts called gateways. The gateways are made of a collection of half gateways, one for each connected network. The overall Internet architecture is similar to the old ARPANET, where gateways replace IMPs and networks replace lines. IMPs are interconnected through lines and gateways are interconnected through networks. The original Internet addressing took the form of flat IP address that were used in the ARPANET, except the IP-addresses were now divided into [Net-ID][the-rest]. Net-ID's were totally unstructured, flat-spaced, and with a fixed length, just like the old 6-bit IMP-addresses of the original ARPANET. The later use of address classes was a packing trick, not structuring.

Then subnetting was invented. Subnetting is structuring but is limited to one level. It was not permitted to subnet a subnet. A structured address is like a street in the US, the addresses reflect position. To go up in street-number you go one way on the street, to go down you go the other way. Unstructured addressing is like many old streets in London, where the number of a building tells you in what order they were built, and not where they are located. You cannot compute which way to go to find a given street-number, because you have no idea where a building is without talking to someone with local knowledge.

Flat addressing requires everyone to know about everyone, and to make routing decisions on the same granularity regardless of distance. Consider my work address:

Danny Cohen
USC Information Sciences Institute
Eleventh Floor
4676 Admiralty Way
Marina del Rey
California 90292
USA

It has seven layers, like all good layered products should. Does routing decisions made in Kyoto need to consider all seven levels? Obviously not. Kyoto should use either one or two of the above addressing layers, without having to "understand" and consider the rest of the layers. What is the right number of levels in an addressing hierarchy, 7 or ? A side effect of what we have done to the Internet is that existing connection is not always found, and rich connectivity is not always an asset.

Gateways often serve as routers (IMPs) of networks, eliminating the need for yet another local net protocol. Networks are often interconnected by long lines. Now a single line may be a subnet or even a whole network. Networks are very heterogeneous with respect to performance size and number of users. Networks can have long-haul lines, regional networks, campus networks, departmental networks, home networks, ad infinitum. they form a hierarchy but not a tree. The world is not a mesh of equal nets. Routing now goes "up, away, then down". Our routing granularity should depend on distance, like mail sorting. Addressing hierarchy is not routing hierarchy, they are separate things. Addressing is used for routing, but you do not have to follow the route if you can do better.

What should a network address reflect? Should a single corporate entity (e.g., DoD and IBM) have a single address space no matter where the end-point entities exist geographically? That is not how their telephone numbers are assigned. Addressing should reflect location, not ownership. I recommend that hosts should be only on local networks. One of the things that caused the addressing difficulties that we have today is that the hosts all existed directly on the ARPANET when it started, and therefore had Net-10 addresses, not reflecting their positions We ran into all the problems because it was a given axiom: Routing had to deal with a flat address space. It would be much better if hosts were only on local networks. This would be a great step toward addresses that reflect position.

About the "C"-word

We have connections all over, but we do not like to talk about them. On private networks, TCP creates a connection between hosts. On public networks X.25 does it. Since there is a confusion as to what is a private network and a public network, I provide this explanation: "Private networks are the ones paid for by the public, and public networks (in the US) are owned privately." What percentage of the traffic is "C"-oriented. I expect it to be fairy high. Why don't we allow our gateways to share in the information about C-things?

In order to discuss the C-things rationally, we should consider these two questions: Is the processing of subsequent packets in the same connection less then that of the first packet? How much work is needed to establish and tear down a connection? The answer to the first question is yes. The answer to the second will take some more work. The break-even point is when the time to establish a connection is equivalent to the processing time that is common to all connectionless packets. This relation tells us when it makes sense to use a connection scheme in the gateways.

What is in the connection that is common to all its packets? The routing is the same for all the packets. So is policy. The decision about access and privileges is the same. Billing, load sharing, security, etc., are all things that can be accomplished once for an entire connection lifetime. Therefore, supporting connections is something that is important for networks to do.

The Bad News

The bad news was mentioned before, the networks are not going to be subsidized forever and ever. Both the Disney World model and the restaurant model of payment are equivalent. Do you pay a-la-carte, all-you-can-eat, or don't care because the parents pay? The scheme of payment has been observed to modify behavior. More people overeat in the second scheme than under the first. Will we continue to live in the beautiful world of "parents pay". Other people pay for our network usage, some of us believe that it is written in the Bill of Rights the guarantee of free packets. How long will they pay for our communications? (Also while doing it why don't they pick up our food, travel, phone and other bills?)

When would be a good time to introduce billing and accounting? Most of us continue to say later. Consider that approximately 15 percent of ATT's operating expenses was billing and accounting. This is not an insignificant amount. We can no longer consider this issue as an afterthought. Maybe something should have been done very early to support the eventual requirements of billing.

In the “parents pay” world, we have another interesting effect. Every group of scientists needs a communication network of their own. This need depends on the fact that someone else is willing to pay. We don’t insist that we have special airlines for scientists only. Having our own special airline is feasible, but expensive, and we are not willing to spend our research funds for it and don’t mind sharing the airplanes with other people. However, if the parent pays, many scientists insist on their own networks, and have no interest in any economy of scale arguments. If we are forced to go to a pay-your-own-way scheme, then we can expect to see several side effects. The first side effect is that users may become friends of the networks, instead of foes. The biggest enemy of the ARPANET was its success, the fact is that it has a growing user community. Now there are too many users to get the great service we had when hardly anyone knew we were there. The phone companies do not have such an attitude towards their users. Outsiders will be welcomed to join our networks. They pay for their usage.

We have a mindset that you pay for what you get. This is only half the equation. The other half states that you get what you pay for. Payment is a terrific alternative to PBR. Every goal that PBR makes can be reached with payment. For example: if the DoD put in a resource to help only DoD people, they could charge $101 per packet to use that resource, refunding $100 per packet only to the DoD people. If too many outsiders continue to use that resource, DoD could increase the price to reduce their usage or to create enough revenue for additional resources.

In conclusion where we want to get is to provide gigabit service for many megausers. There are two different paths. The common carriers are first dealing with gigausers then increasing speeds. They already serve about 500 megausers. We are trying to go to gigabits with very few users and believe we can then scale up the system for megausers. Who will get to the desired point first? I believe that billing is required. We cannot throw a party for 500 million people and expect the parents to pay for it forever. Such an operation must be based on solid economics, and must obey its basics rules (e.g., economy of scale).

The most sensible way to proceed is for us to do research, then have the CCITT choose and adopt a system. The carriers will implement it. Then we will use it “as is”. Even if it means changing our addressing scheme. There may be much more to gain from being part of this economic system than to stay apart from it.

MINI-SESSIONS

Policy Based Routers and a New Model

Susan Hares, MERIT

Editor’s summary: The rich interconnectivity within the Internet causes routing problems today. However, the presenter believes the problem is not the high degree of interconnection, but the routing protocols and models upon which these protocols are based. Rich interconnectivity can provide redundancy which can help packets moving even through periods of outages.

Our model of interdomain routing needs to change. The model of autonomous confederations and autonomous systems (RFC-975) no longer fits the reality of many regional networks. The ISO models of administrative domain and routing domains better fit the current internet’s routing structure.

With the first NSFNET backbone, NSF assumed that the Internet would be used as a production network for research traffic. We cannot stop these networks for a month and install all new routing protocols. The Internet will need to evolve its changes to networking protocols while still continuing to serve its users. This reality colors how plans are made to change routing protocols.

NSFNET-centric views on IP limitations

Bilal Chinoy, MERIT

Editor’s summary: The goal is to insure that the NSFNET is reliable and efficient through the regulating and tuning of traffic patterns. In the near future, MCI will provide Digital Reconfiguration Service (DRS) to make it possible to tailor the circuit topology to the users (NSFNET) needs. This provides the ability to dictate to the common carrier what the user requires and is willing to pay for. To make reconfiguration worthwhile, rationals must be developed to constrain costs and bandwidth. Flow models and architecture limitations must be considered when developing these rationals. The reconfiguration is done by feeding traffic measurements to a matrix and a resulting optimal topology is produced to give to the DRS. The reconfiguration is done in units of T1 lines at a cost of $100 per reconfiguration. In the beginning, the reconfigurations will be fed in manually and gradually converting the system over to automatic setup.

Pros and Cons of Stateful Gateways

Charles Eldridge, SPARTA

Presentor’s summary: This presentation briefly noted how the evolution of the current internet architecture has explicitly involved stateless gateways. Both the possibility of arbitrarily unreliable networks and the expense of machine cycles and memory have motivated use of stateless gateways. However, the stateless gateways themselves have become sources of non-reliability between end communication points, due to buffer and processing shortages.

Stateful gateways should be considered for new stages in the internet evolution, because they can provide more intelligent management of internet traffic in the forms of Fair Queueing and other flow-management techniques. It is also possible that stateful gateways could provide more efficient use of network bandwidth by providing reliability services along the communication paths. Arguments for this feature would be based upon Deming's quality control arguments. Deming argues that quality control should be practiced at individual process levels, rather than only at the collective process output stage. For evaluating the utility of this policy in networks, we should ask whether the throughput of a segmented, lossy connection is better with ARQ error control practiced on each segment, or only at the connection endpoints under specific assumptions regarding link speed and error rates.

Giving gateways "state" would have its costs and impacts. For example, a single gateway would have many sources sending traffic through it. Significant processing would be required to track the intensity of the sources over time, as would be required for the "govern flow of traffic" role. In order to perform the "provide reliability" role, the streams between hosts would need some kind of sequencing. However, IP data streams have no features (protocol header information) with which to define sequence numbers. The cost of giving the host-pair data streams sequence information might be prohibitive, because it would necessitate either changes to IP or definition of a new IP derivative. The very wide scope of the internet means that any given gateway would need to maintain and process a very large number of state variables. For each gateway, the potential number of states would increase with the square of the number of interacting end-systems.

Investigation of the Domain Name System

Chuck Cranor, U. Delaware

Editor's summary: A report was presented on the domain name system, which is designed to provide a consistent name space to be used for referring to resources (i.e., host name to IP address mapping). The data are cached locally to speed up the mapping process. The data collected from the name server *named* showed an average cache hit ratio of only about 60 percent and that took an average of 18 hours to stabilize. Also, the data showed the cache entries had an average time to live of only 12 hours. The report recommends implementing incremental updates of local data. Also, it recommends that the times to live should be larger, since the stabilizing time is larger than the time to live. Finally, the report recommends future protocols to keep more information around about search paths, take into consideration the frequency of reference and implement error messages to flush the cache of old data.

A Policy-Based Model and Paradigm

Zaw-Sing Su, SRI International

Editor's summary: This session raised the question: do we have a single administrative paradigm for policy based routing, or are we looking at a multiple sourced system? There was some discussion, but no definitive conclusions.

First Packet Routing in Gigabit Networks

Debbie Deutsch, BBN STC

Editor's summary: When routing packets at gigabit speeds, the call-setup time becomes appreciable with respect to transmission time. Along in this setup time is the time required to consult name servers for the destination address. The presenter proposes that, since you must consult the name server first, then route the first packet by sending it to the name server. In the current scheme the distance traveled by the first packet is the sum of the distances to all name servers polled plus the source/destination distance. In the proposed scheme, the distance is reduced to the sum of the distances between source/name server, name server/name server, and finally name server/destination which is effectively the source/destination distance! This scheme can be improved upon by using authentication and policy servers in the same way. The trip the packet takes back from the destination is flexible in that it is totally stateless and the routes are written on the packets and stored in the nodes as the packet flies across the network. With these parameters, the designer can pick a back routing method that goes with the choice of routing architecture.

How Slow is a Gigabit?

Craig Partridge, BBN STC

Editor's note: This session consisted of a presentation and discussion of the paper "How Slow is One Gigabit per Second?" which appears elsewhere in this issue.

Day 2: 3 June 1989

PANEL 1

International Standards

Steve Goldstein (Chair), Susan Hares, Philip Prindeville, Debbie Deutsch

Editor's summary: During the discussion two major broad architectural issues emerged: (1) procedures and mechanisms for management, control and/or coordination of increasing numbers of interconnected international (as well as national networks), and (2) development, deployment and management of protocol application gateways, especially mail, file transfer and virtual terminal between the Internet and ISO/OSI networks in Europe.

The Canadian National Research Net (56 Kbps evolving to T1; multi-protocol), NORDUNET (Scandinavia), EASINET (France/CERN), Bellvue (Baden-Wuremburg, FRG), JANET (UK) and PACCOM (Hawaii, Australia, New Zealand, Japan) are all connecting to the Internet backbones. The NSFNET approach is to assign autonomous system (AS) numbers to each and run EGP at the NSFNET entry nodes. Each AS has a registered responsible person who can be contacted to coordinate and resolve improper behavior. (Presently, Canadian provincial nets peer with NSFNET mid-level nets.

RARE WG1 (X.400 NHS) has asked the Internet to submit a mapping of Internet domain namespace onto X.400 or name schema. The IETF OSI working group has established an OSI mailing list (send requests to ietf-osi-or-request@cs.wisc.edu). This is needed to facilitate RFC-987 SMPT-X.400 mail exchange. X.500 Directory Services offers a general extensible schema on which white pages services could be built for the Internet. Other possibilities are Larry Peterson's Profile and the DEC Network Architecture Name Service (DNANS). X.500 needs extensions (e.g. methods/rules for replication, access control) to be useful as a DNS for the Internet. In any event much homework is needed for naming entities in the US for both X.400 (registering) and X.500 (advertising bound names).

Questions and Comments

Steve Goldstein provided an overview of the international networking arena. The European community has a organization similar to the US FRICC called RARE. RARE is a group of academic research network people and different companies in Europe. The Coordinating Committee for Intercontinental Research Networking (CCIRN) is made up of RARE, FRICC, and Canadian groups. This group plans to expand its membership to Australia and Japan. Several things came up in regards to standards at the RARE April meeting, especially including X.400. RARE WG1 is working on this service. X.500 can be used to serve names for X.400, since RFC-987 specifies a mapping.

PANEL 2

Internet Policy Management

Scott Brim (Chair), Danny Cohen,
Susan Hares, Mike Little

Chair's Summary: Given a policy-based routing (PBR) architecture, what will it behave like when it is actually used in the real world? Any architecture must be tested for practical manageability. In our panel we wanted to examine current concepts of policy-based routing with regard to a few key questions (which we do not consider to be the only areas of concern, by any means):

1) Local autonomy versus global harmony: What balance will be sought between central and distributed control? While we must allow some local independence, how will we constrain unexpected global effects of local change? How will we assure composability of administrative decisions by the autonomous regions (AR)? How will accounting systems be coordinated between ARs?

2) Problem diagnosis and repair: Since a single policy change may dramatically change the observed Internet for individual users, perhaps for individual user processes, and since policy changes can occur for obscure reasons at unexpected times, how can network troubleshooting and diagnosis be done?

3) Abstraction versus the need to control one's own destiny: How will we balance the need for abstraction of information (in order to keep the Internet from collapsing from the load of its own internal information) with the need for a richness of information at end ARs in order for them to have confidence in their PBR decisions?

These are large topics and, since we opened the discussions up to the audience a great deal, we really only worked on the first one. Viewpoints on how the balance between local autonomy and centralized control should be achieved in a functioning Internet covered a wide spectrum, from a belief in the need to respect the rights of the individual ARs, no matter how small, to a belief that as soon as people are billed for Internet usage the solutions to such problems will become obvious. We also discussed just how much policy activity is necessary for a network to be considered managed, and in fact how much policy activity a network should have.

We managed to reach very few conclusions (reaching conclusions was not our goal), except that having multiple models for how much autonomy the ARs have will lead to great difficulty in actually running the Internet, so we need to agree on one model very soon. Any architecture containing PBR should minimize its dependence on such policy or the cost of having it would be greater than not having it. A good way to ease dependence on PBR appears to be by taking into account the eventual need to bill for network use, since some apparent PBR requirements then cease to exist. The only other solid conclusion was that, if only in self-defense, we need to create a much better program for educating administrators of networks newly attaching to the Internet.

Questions and Comments

Q. One of the major issues facing the policy based management design is determining the use of a single policy model or a multi-modal model. The ISO is looking to create a hierarchy with a single central control. NSFNET thinks they are a single administrative model with the regionals as subordinate. The global policies are hard to police. An important question to ask is: “Can policy changes made locally affect other autonomous systems?”

A. NSFNET is really a community of effort. The regionals all help make it go. It is a hierarchical structure with autonomous domains. The policies allow a set of local autonomy with global restrictions.

Q. As an example of some difficulty that can arise from differing regional policies, a small university wants to hook to two regionals. Whose policy do they implement?

A. Regionals with different polices shouldn’t be bridged if the “host” network will obey only some of one regional’s policies and some of the other’s. That sort of bridging creates only a half-baked souffle. If they wish to connect to two regionals, then they should obey all the policies of both the regionals that they connect to.

Q. Is it possible to create a management policy model which provides for each regional a definitive expectation of service? For example, if several regionals are operating in a particular area, must I buy a line to each one of them if I require ubiquitous routing? Should there be a canon that stipulates the services required (like ubiquitous routing) in order to qualify as a regional?

A. [There was no agreement on a consensus answer. Ed.]

Q. Are the PBR examples a realistic model for the future? What we have now is the parents (the funding agencies) are saying that in eight years the children (regionals) must pay for themselves. We must move to a world where true networking costs are identified, understood, and eventually paid. One should expect that the way we use networks will change significantly then. For example, many things will be programmed to reduce cost. Night-time delivery of mail and large databases will occur because they are the low cost times.

A1. The money that researchers spend is gotten from elsewhere. All that will happen is that we will have to ask Uncle for more overhead funds.

A2. Those numbers do not fully represent the true cost of a self-supporting structure. If the real prices are that low, why do users complain about performance rather than just buy more connectivity? Policy coupled with true cost will provide the motivation to become network knowledgeable. Things will change when we have to pay. We will tend to choose the cheapest way of doing anything.

A3. When the money is gotten from elsewhere, if won’t force a fundamental change, just force the getting of it in different ways.

A4. Some of the services that were discussed are public and should be subsidized, therefore a telephone - pay for service - analogy is not rich enough. Perhaps instead we should consider a tax model with progressive taxation and exemptions to implement policy.

A5. All we still need an access model. How do we allocate shares of network requirements in a policy to allocate priorities? Also, how does a service organization recover costs of providing service to its users and outside entities?

Steve Goldstein commented that the FRICC has a proposal, which is to be distributed soon, for a gigabit network project that is expected to transition to the commercial providers at the end of the project. Communication cost can then be included in budget proposals

PANEL 3

Stretching the Internet Protocols

Bob Braden (Chair), Jacob Rekhter, Craig Partridge,
Gurudatta Parulkar, Phil Karn

Editor’s summary: This session was a discussion of the lessons learned from designing the NSFNET backbone routers. The backbone routing is autonomous-system (AS) based as defined in RFC-904. However, network-based routing is used at the exit points from the backbone. Because of this the backbone address is created using a combination of the AS number and the network number.

Two protocols are used to provide the routes - the Interior Gateway Protocol (IGP), and the Exterior Gateway Protocol (EGP). IGP deals with routes inside the backbone and uses a subset of the ANSI IS-IS protocol. EGP is used to communicate routes between the backbone and the regionals. while the EGP model and protocol were not changed, the backbone routers interpret the information differently. Two areas were affected by the changes: 1) verification of incoming routing information to prevent black holes. 2) controlled distribution of outgoing routing information.

By using both of these protocols, NSFNET obtains normal routing, fallback routing and “hot-backup” routing. Stable routing provides the users with predictable performance when using the backbone. Fallback routing from AS to AS provides alternate paths when the primary paths fail. Hot-backup routing occurs within a single AS when primary or secondary EGP peers fail.

The NSFNET backbone can support a limited subset of policy-based routing (PBR). This is based on distribution of routing information between backbone and regional networks. This cannot support any arbitrary policy, because enforcement occurs at the administrative domain level. The existing software allows the engineering of large WAN's, allows arbitrary mesh topology, and allows network policy control without impact on performance.

Questions and Comments

Q: How long can you continue to stretch, in terms of number of nets, nodes, ASs, etc.

A: In terms of IGP, this can be stretched for a very large number. An ANSI paper suggests about 10,000 systems. The new draft of the ANSI IS-IS protocol will address performance issues and partial updates. In terms of EGP, when the work started the feeling was that EGP could not be used. But we have stretched it a little, and could probably stretch it a little bit more. The problems with EGP is that it requires complete updates, and an arbitrary mesh topology requires more intelligent table usage. BGP was created from what we learned from using EGP in the backbone.

Q: NSFNET has tampered with AS model. It was originally modeled as a system of gateways, you are routing essentially on a system of gateways when the network itself does not belong to an autonomous system. If we had a single network attached to two AS's that would be legal, but your model will not handle it.

A: This question of "what do we mean by an Autonomous System?", has been discussed before. And I agree we have broken the model.

Craig Partridge

There is a class of users that travel with their PC'S and expect to be connected to the Internet wherever they travel. Phone access seems to be the minimum requirement. IP address assignment is preferred by system administrators. However knowledgeable users have been known to provide promiscuous gatewaying. For instance, the MIT PC gateway assigns an IP address at connect time. I have a vision of people flying across the country while reading their electronic mail on the plane. People will want to run their home environments wherever they go, how will this impact our vision of future network addressing.

PANEL 4

Internet Scaling to Gigabits and Megahosts

Sandy Fraser (Chair), Lixia Zhang, Gary Delp and Roy Perry

Editor's note: Synopses of the panel presentations and discussions are given below.

Lixia Zhang

A simulation was described of a network with hosts running TCP and slow-start. The simulations were driven by data monitored on a trunk in a backbone network. The simulations revealed pronounced oscillations in queue lengths. These oscillations were apparently due to synchronization of the slow-start/back-off congestion control mechanisms. Observations were made on the network without slow-start. In the latter cases, substantial numbers of packets were lost due to congestion and to redundant retransmissions.

It was observed that slow-start significantly improved network efficiency, but conjectured that, in order to avoid oscillations and further improve efficiency, it would be necessary to keep more state information.

Sandy Fraser

There is a strong economy of scale in transmission. As evidence of this we see the rapid deployment of long haul, fiber optic systems spanning the country. Since fiber optic systems were introduced in 1980, transmission speed has increased from 45 Mbps to 1.7 Gbps, while costs per Gbps per mile has rapidly declined.

Many users share high speed transmission lines through the use of multiplexers which today form a hierarchy. Twenty-four 64-Kbps digital circuits are multiplexed onto a single 1.5 Mbps DS1 line. Twenty-eight DS1 circuits are multiplexed onto a 45 Mbps DS3, and thirty-six DS3 circuits are combined to form a 1.7 Gbps FTG circuit.

There is a certain amount of information hiding which makes this hierarchy practical to construct and maintain, i.e., a single FTG transmission line handles more than 24,000 voice circuits, but the multiplexed equipment only has to know about the 36 DS3 circuits from which the 1.7 Gbps information rate is derived.

Perhaps there should be a packet-multiplexed hierarchy analogous to this synchronous digital hierarchy, so that high capacity transmission lines that move packets for thousands of people do so without having to pay attention to every individual conversation. It is not clear what would be the best form of aggregation. Jacob Rekhter described routing for Autonomous Systems which hides the details of destination addresses. Another possibility is to use encapsulation so that many packets are handled as a single packet ion the high speed long haul circuits.

The equipment needed to multiplex packet communication at a given transmission rate is usually more complex than synchronous multiplexing equipment for the same data rate. Thus it seems likely that the fastest available packet multiplexer will never be as fast as the fastest available synchronous multiplexer.

These thoughts suggest a packet multiplexing hierarchy which merges at the top end into the synchronous digital hierarchy. Hierarchical organization is also important for large networks of switching machines. The telephone network employs small switches on customer premises (PBX), exchange switches which serve a small town, and backbone network switches that carry large quantities of long distance communications.

Very likely there should be a similar hierarchy of switching machines (routers) for a national data network. Local area networks would be analogous to PBX's, and there would be at least two varieties of packet switches to be used in regional and backbone networks respectively. The exchange networks must handle many types of access circuits, and must interface with thousands of independent circuits. The backbone network need not have as many independent attachments, but must handle a very large volume of traffic.

When we talk of the need for high speed data communications, we must distinguish between high order to obtain volume, as in backbone, and high speed in order to serve supercomputers in local area networks. Switches for use in different contexts probably demand a range of technologies. The components of a switching machine with small fan-out might be interconnected by a single backplane bus. A high capacity switch will probably use a space division network, such as a binary routing network.

Propagation delays and transmission speeds for various transmission media were reviewed, as well as the number of instructions available for packet processing at these speeds. The conclusions were that adequate CPU cycles will be available for basic packet routing procedures, given hardware translation, while remote hosts will seem increasingly distant as instruction times become smaller relative to propagation delay.

Roy Perry

The emerging plans for SMDS that are being laid by US West, other RBOC's and Bellcore were described. In the near term these plans assume that an 802.6 MAN will be used as the vehicle for switched service. Initial data rates for access at up to 45 Mbps will be supported with an interface adapter to encapsulate IP datagrams. Gateways will interconnect MANs to form a wide area network.

While SMDS does not assume BISDN as a prerequisite, it will be supported when it comes, along with 802.6 and ATM. SMDS speeds are expected to be from 1.5 - 45 Mbps to customers. The backbone of SMDS can be any speed with packet sizes up to 8K. However, there is still a need to do research on how to integrate all this onto one network.

List of Attendees

Babu Bangaru, US WEST
Bob Beach, Ultra Network Tech
Steve Bellovin, ATT Labs
Ernst Biersack, Bellcore
Erv Blythe, Virginia Tech
Bob Braden, USC ISI
Scott Brim, Cornell U
Dan Brown, Textronix
Graham Campbell, BNR
Bilal Chinoy, MERIT
Maj. Glen Carter, DCA
Danny Cohen, USC ISI
Mike Collins, LLNL
Chuck Cranor, U Delaware
Pete Crumpacker, MITRE
Mike Davis, U Delaware
Gary Delp, IBM Yorktown
Debbie Deutsch, BBN STC
Charles Eldridge, SPARTA
Dave Feldmeier, Bellcore
Richard Fox, Tandem
Jose Garcia-luna, SRI International
Douglas Franke, ATT
A.G. (Sandy) Fraser, ATT
Steve Goldstein, NSF
Jim Griffioen, Purdue U
Phil Gross, NRI
Susan Hares, MERIT
Steve Holmgren, CMC
Michael Hrybyk, BITNIC
Mike Hui, BNR
Craig Hunt, NIST
Prasad Jayanti, U Delaware
Thomas Joseph, Olivetti Research
Phil Karn, Bellcore
Charley Kline, U Illinois
Wai Sum Lai, ATT Bell Labs
Marianne Lepp, BBN CC
Stefan Levie, U Delaware
Mike Little, SAIC
Andrea Lobo, U Delaware
Bryan Lyles, Xerox PARC
David Mills, U Delaware
Russ Mundy, Trusted Info Sys
Thomas Narten, Purdue U
Carolyn Nguyen, ATT Bell Labs
Yaoshuang Qu, George Mason U
MichaelPadlipsky, UNISYS
Phillip Park, BBN STC
Craig Partridge, BBN STC
Guru Parulkar, Washington U
Drew Daniel Perkins, CMU
Roy Perry, US WEST
Philip Prindeville, Wellfleet
S. Ramanathan, U Delaware
Aswath Rao, ATT Bell Labs
Allen Rehert, ATT Bell Labs
Jacob Rekhter, IBM Yorktown
Joel Replogle, NCSA
Ira Richer, DARPA
Kenneth Rodemann, Purdue U
Paul Schragger, U Delaware
Doretta Schrock, Textronix
Tim Seaver, MCNC
Adarsh Sethi, U Delaware
Nachum Shacham, SRI International
Scott Shenker, Xerox PARC
Peter Shames, NASA Space Telescope
Judy Smith, HP
Frank Solensky, RACAL-InterLAN
Zaw-Sing Su, SRI International
Michael Ting, NIST
Ron Toth, ATT Bell Labs
Claudio Topolcic, BBN
David Tsao, COMEX
Howard Tsai, ATT Bell Labs
Paul F. Tsuchiya, MITRE
Ben Verghese, HP
Mike Wenzel, HP
Charlie Wickham, DEC
Joe Wieclawek, JPL
Bill Williams, BNR
Lixia Zhang, MIT LCS