TUBA: Replacing IP with CLNP

Running Internet protocols on top of the Connectionless Network Protocol can solve the Internet's addressing and routing problems.

Dave Katz and Peter S. Ford

The Internet has evolved considerably from its original role of supporting research in packet-switched computer networking. In the last decade, the Internet has become the computer network infrastructure for the global research and education community. During the 1990s, the Internet continues to evolve to meet the networking requirements of both the public and private sectors. The growth in networks and hosts connected to the Internet has revealed problems in addressing and routing the Internet Protocol (IP)[21]. We propose using the Connectionless Network Protocol (CLNP) [7] supported by the associated OSI routing protocols, as a replacement for IP. The basis of the proposal is to run the Internet transport protocols, the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP), on top of CLNP in an approach known as TCP and UDP with bigger addresses (TUBA) [1].

This paper details the fundamentals of CLNP and the OSI connectionless routing architecture, the operation of the IP suite with CLNP replacing IP, the support of Internet applications operating on top of TUBA, and a transition plan to a TUBA Internet.

Problems with Internet Growth

Growth

Availability and use of the IP suite has resulted in many research networks using TCP/IP, e.g., the NASA Science Internet, the Department of Energy's Energy Sciences Network, the European IP Backbone (EBONE), and Japan's Widely Integrated Distributed Environment (WIDE). U.S. and international research networks subsequently have interconnected, resulting in today's Internet. Demand for computer networking outside the research and education community contributes to the continued growth of the Internet.

The growth in the number of individual networks connected to the Internet is illustrated in Fig. 1. The interconnection of organizations outside of the research and education community has opened up a large market for Internet services as seen in the growth rate since 1990. This graph only encompasses those networks that are part of the public Internet; there is a greater number of IP networks allocated that are not connected to the global public Internet.

Limitations of IP Addressing

IP addresses are fixed-length, four-octet values, addressing up to 4 billion hosts. IP source and destination addresses are located in fixed locations in an IP header, so simply changing the length of IP addresses would require changing every IP protocol engine. The designers of the Internet suite of protocols did not anticipate the application of IP in the environment of the large numbers of hosts and networks that the advent of microprocessor technology makes possible:

"TCP addressing is intimately bound up in routing issues, since a HOST or GATEWAY must choose a suitable destination HOST or GATEWAY for an outgoing internetwork packet. Let us postulate the following address format for the TCP address. The choice of network identification (eight bits) allows up to 256 distinct networks. This size seems sufficient for the foreseeable future. Similarly, the TCP identifier field permits up to 65,536 distinct TCPs to be addressed, which seems more than sufficient for any given network." [3]

Local area networks (LANs) have led to greater use of computer networks and a reinvestigation of the IP address structure. The four-octet address space is currently partitioned into the following three unicast address classes (see Fig. 2):

Class A — one octet of network id, three octets of host id;
126 networks, 16 million hosts/network.

Class B — two octets of network id, two octets of host id;
16,000 networks, 65,000 hosts/network.

Class C — three octets of network id, one octet of host id;
2 million networks, 250 hosts/network.

General Scaling Issues

Scalability of a protocol involves more than just being able to address large numbers of hosts. As the

DAVE KATZ is a software engineer at Cisco Systems, Inc.

PETER S. FORD is a member of the technical staff at Los Alamos National Laboratory.

0890-8044/92/$03.00 ©IEEE

Internet grows, routing databases must grow more slowly than linearly with respect to the number of addressable hosts.

To accomplish this, network addresses should (to some measure) reflect network topology — the greater the congruence between the address hierarchy and the network topology, the greater the ability to use topological abstraction to reduce the size of routing information. This implies the need for multiple levels of hierarchy, and sufficiently flexible address structure and assignment to support them.

Network addresses must be carefully assigned to achieve significant reductions in routing data base size. Wholesale addressing changes may be necessary as site connectivity is altered to avoid addressing entropy. Routing protocols need to allow exceptions to the addressing hierarchy to be supported at a reasonable cost, to ease addressing transitions and provide for the attachment of networks to multiple service providers.

Scaling Problems of IP

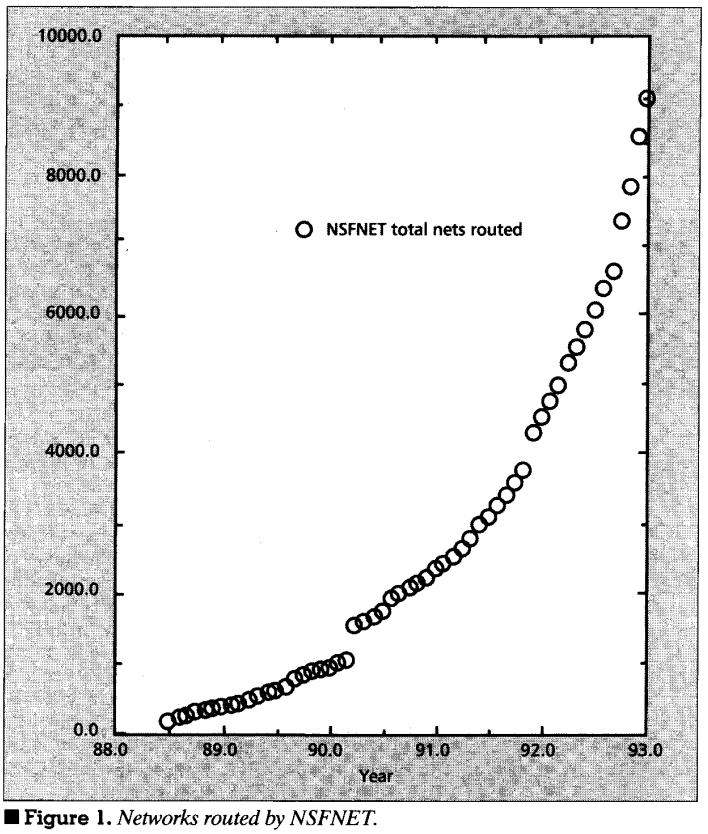
Internet routing treats IP network numbers as a set of flat identifiers, with each network requiring a routing table entry. As a result, Internet routing becomes less tractable with respect to processor and memory requirements as the number of IP network addresses increases. Demand for Class B addresses will result in exhaustion of the B space before 1995. Then, Class C addresses will be assigned to new Internet sites, and sites with more than 254 hosts will need to use multiple network addresses. The current load of 10,000 networks already taxes the routing infrastructure of the Internet, and the current system will not scale to 2 million Class C networks.

A solution to the scaling problems of the IP routing system is to hierarchically assign IP network addresses by allocating blocks of addresses to sites from a block of addresses assigned to their network service provider. The routing system then routes blocks of addresses instead of routing individual networks. Sites served by a single provider lie within a block, so routes to all of those sites may be represented by the single routing entry for the provider, allowing for significant abstraction in the routing system. The application of these techniques to IP is known as Classless Interdomain Routing (CIDR) [5], since the class structure of IP addresses is no longer used to identify elements in the routing system. Although CIDR has the potential to extend the useful life of IP, it does not address the fundamental limitations of 32-bit addresses.

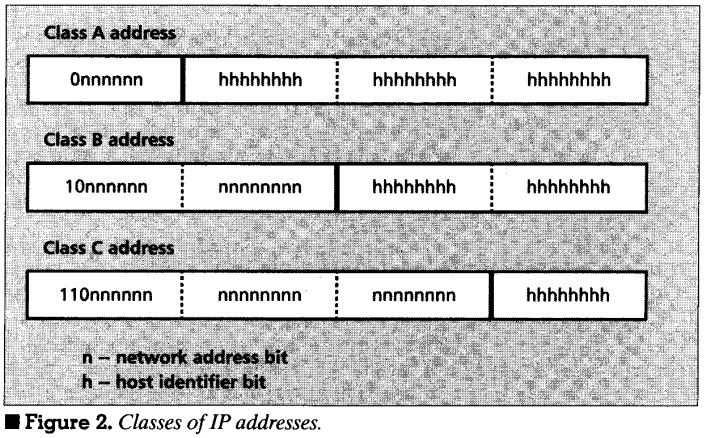
Beyond 32-Bit Addresses

CIDR addresses the immediate routing limitations of IP. As the Internet becomes a global public internet, it must be able to address all possible computer systems that wish to communicate. One can imagine each telephone termination evolving into a computer network, and that over time there may be more computers (networks?) than there are people. Thus, the global internet must be able to support millions, or perhaps billions of networks, connecting several billion hosts.

CLNP retains the overall architecture of IP, but provides a much larger and more flexible address space. In conjunction with the OSI routing protocols, scaling to the size of a ubiquitous global internetwork can be achieved. By using the existing Internet transport and application protocols, it is possible to continue to use the broad base of existing Internet applications on top of a CLNP infrastructure.



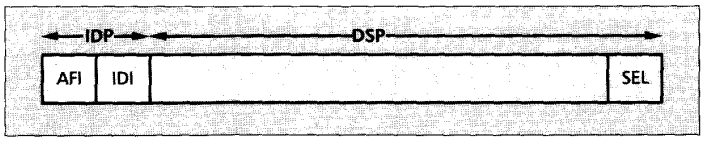
■ **Figure 1.** *Networks routed by NSFNET.*



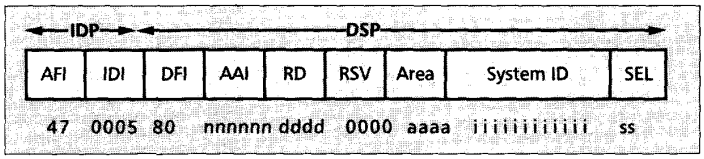
■ **Figure 2.** *Classes of IP addresses.*

Motivation for TUBA

The TUBA approach is motivated primarily for pragmatic reasons. CLNP and its associated protocols are mature technologies, for which both router and host implementations already exist in off-the-shelf products from vendors. Furthermore, CLNP service in the Internet already is being deployed by many service providers. TUBA provides a way



■ **Figure 3.** *OSI NSAP.*



■ **Figure 4.** *GOSIP NSAP address structure.*

to leverage this investment in development, deployment, and training to address the growth problems of the Internet.

TUBA is explicitly not revolutionary, but only evolutionary. Although some may believe that radical new technologies are necessary to provide for data communication needs in the 1990s and beyond, it does not seem feasible to abandon the current internetwork paradigm, given the time required to design, implement, debug, and deploy protocols that are currently in the research phase, if reliable and ubiquitous connectivity is to be maintained. Of course, research and development efforts into new approaches to data networking should be pursued, while preserving the integrity of the global networking infrastructure.

Addressing

Network layer addresses for TUBA are standard OSI network service access point (NSAP) addresses [6]. These are defined to be variable length of up to 20 octets, although it is worth noting that CLNP itself does not impose this address length limitation (CLNP uses an octet to represent NSAP lengths and could support addresses approaching 100 octets in length, but this should not be necessary).

The format of an NSAP address is shown in Fig. 3. The authority and format identifier (AFI) is an addressing plan identifier used to discriminate between addressing and numbering plans from internationally recognized organizations. ISO and CCITT assign AFI codepoints to organizations willing and able to administer addressing plans. The AFI and initial domain identifier (IDI), taken together as the initial domain part (IDP), describe the administrative authority for this part of the address space. The administrative authority may be a country, a multinational entity, or some other body.

Following the IDP is the domain specific part (DSP), which is formatted according to the authority described in the IDP. The DSP typically is a combination of administrative and topological information. The last octet of the DSP is the NSAP selector. This field is used to select among multiple transport layer entities in a system, serving much the same function as a protocol ID field in other protocols.

One example of a full NSAP address format is the one defined by U.S. GOSIP (Fig. 4) [16]. Operating under IDP 47/0005 (United States government), the DSP contains a DSP format identifier (effectively a version number), an administrative authority identifier (identifying the next level administrative authority), a routing domain identifier, a reserved field (for future expansion), an area ID, a system ID, and an NSAP selector.

OSI NSAP addresses were designed to be large enough to contain embedded addresses from other numbering plans, such as X.121 and E.164 addresses. Although primarily intended as a means of delegating address administration, these may prove to be useful as a way of determining subnetwork address bindings if ubiquitous level-2 services (e.g., ATM) become available.

On the surface, the ability to support multiple address formats may seem needlessly complex and difficult to implement. In fact, however, the routing protocols do not require a single, fixed, address structure to operate efficiently. Similarly, hosts do not need to know anything about the structure of addresses, viewing them as opaque strings. Furthermore, the ability to define new address structures allows for flexibility in the face of future technologies and deployment requirements. This extensibility removes the need to define a single, fixed, global addressing plan.

One aspect about OSI addressing that sets it apart from IP addressing is that network addresses generally are assigned to systems, rather than to interfaces. Routers and multihomed hosts typically need only one address. Furthermore, there is no conceptual equivalent to an IP subnet address (effectively the address of the subnetwork itself, e.g., a piece of Ethernet cable). Identifying hosts instead of interfaces simplifies configuration, and can add robustness — if an interface on an IP system fails, packets destined to that interface's IP address may be undeliverable, even when the system is reachable through another interface. This is not an issue if addresses are not bound to interfaces.

CLNP

The heart of the TUBA proposal is the OSI CLNP [7]. Semantically it is similar to IP; in fact, it originally was derived from IP. It is a datagram protocol, carrying full source and destination addresses in every packet. Independence from frame-size constraints imposed by subnetwork media is provided by a fragmentation/reassembly mechanism.

Optional features include source routing and route recording, as well as multiple types of service. Error diagnostics are provided by error report packets. Echo request and reply packets provide low-level diagnostic capability.

Work is progressing at this time on extensions to CLNP, including network layer multicast capabilities, arbitrary packet coloring for policy routing purposes, and more flexible type of service selection.

Routing

The OSI network layer has a full complement of routing protocols. TUBA uses these routing protocols without modification.

The OSI routing framework [8] describes a global routing environment divided into routing domains. A routing domain is a set of hosts, routers, and

subnetworks operated according to a single policy model (most likely by a single administrative authority). A routing domain is expected to be very tightly coupled in terms of its routing, with a high degree of trust between member routers. Routing domains should be able to contain many thousands of hosts and still operate efficiently.

Routing between domains, or interdomain routing, is controlled by policy issues, rather than by purely topological issues. Interdomain routing is expected to be much more loosely coupled, with only limited trust between routing domains. Since it is responsible for establishing ubiquitous connectivity, interdomain routing must be able to scale to arbitrarily large internetworks.

Routers and hosts are modeled with very different capabilities. The philosophical approach used is that hosts should have virtually no routing intelligence whatsoever, allowing them to concentrate on running applications, and avoiding architectural limitations that could cause future problems.

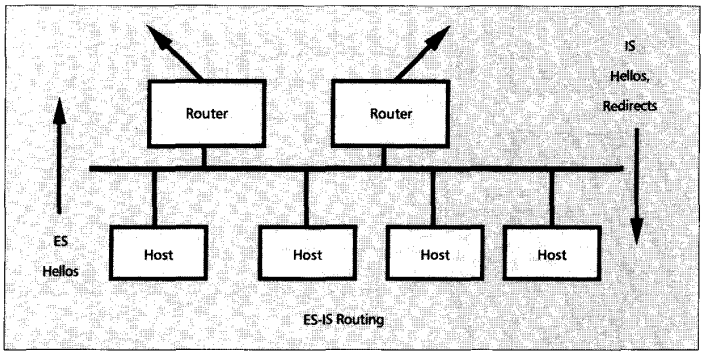
Routing Between Hosts and Routers

Routing between hosts (end systems) and routers (intermediate systems) is accomplished using the end system to intermediate system (ES-IS) protocol [9]. The ES-IS protocol has three major functions: the announcement of reachability between hosts and routers (using hello packets), the population of host routing caches (using redirect packets), and the configuration of host NSAP addresses (using assign address packets).

Reachability Maintenance — Hosts and routers periodically send one another ES-IS hello packets that contain their network addresses. Two multicast subnetwork addresses are used, one with the semantic "all routers," and the other with the semantic "all hosts." Hosts typically do not listen to the all routers multicast address; this has the desirable property that hosts need not incur the overhead of processing the more numerous ES hello packets. The use of multicast, rather than broadcast, also means that other machines on the subnetwork that are not participating in ES-IS will not have to receive (and ignore) these packets.

The mapping between subnetwork addresses and network addresses is noted from the subnetwork-specific envelope in which the hello packet is carried, rather than being carried within the data portion of the packet itself. Each hello packet contains a holding time, which tells the packet's receiver the length of time for which the information is valid.

The result of this exchange is that each host knows the identity (and subnetwork address) of all routers on the subnetwork, and all routers conversely know the identity of all hosts. This information is aged, and will be deleted if not refreshed periodically by subsequent hello packets. The longevity of this information, and thus the frequency of its advertisement, is configurable. Furthermore, routers can set the value of the holding time that the hosts put in their hellos, allowing the entire subnetwork to be tuned from the routers. This provides a means of controlling the balance between overhead and rapid convergence.



■ **Figure 5.** *ES-IS routing.*

Note that, in most cases, the frequency of advertisement by hosts can be set quite low, to reduce bandwidth utilization on shared media with many hosts present. The case of a router commencing operations on a subnetwork can be optimized by having the hosts send unicast hello packets to the router when they first hear its hello packet.

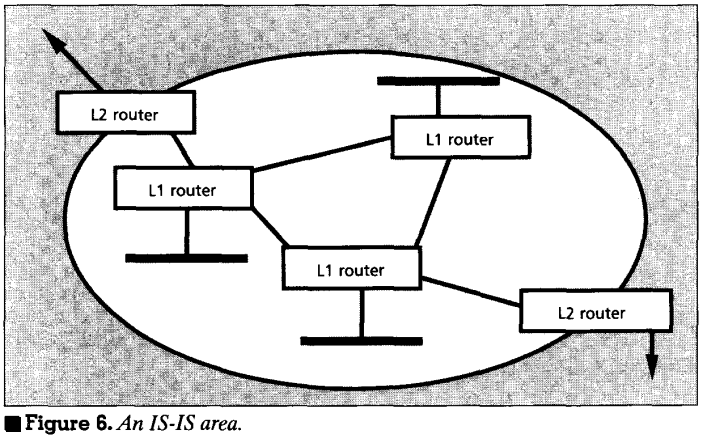
Populating Host Routing Caches — When a host wishes to emit a CLNP packet, it examines its routing cache to see whether it has any information about the destination NSAP address. If so, the packet is forwarded to the next hop subnetwork address contained in the cache.

If the host does not have any cached information about the destination address, the host simply forwards the packet to any router on the subnetwork (because of the adjacency list built from incoming IS hellos). The router can be chosen by any local algorithm. If the router chosen by the host is not the best path to the destination (either because the destination is better reached through another router, or because the destination is on the same subnetwork as the source), the router sends an ES-IS redirect packet back to the host, containing the appropriate next hop and a holding time. The host then inserts the information contained in the redirect packet into its route cache (Fig. 5).

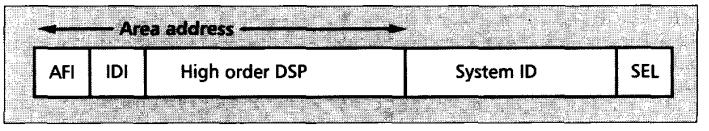
Once a route is cached, the cache entry can be refreshed by noting the arrival of reverse traffic with network and subnetwork addresses that match the cache entry, thus reducing the number of redirect packets generated.

If a host wishes to communicate with another host when no routers are present (known because of the lack of router adjacencies), it simply multicasts the CLNP packet to the "all hosts" subnetwork address. The target host will then send a unicast ES hello back to the originating host, allowing subsequent communication to be unicast rather than multicast.

Two additional optimizations are possible when a router generates a redirect packet. First, the router may choose to specify an equivalence class of destinations for which the redirect applies by returning a mask in the redirect packet. Often this is possible for destinations located outside the local subnetwork. Second, if an equivalence class of addresses containing the target destination is known to have subnetwork addresses embedded within the corresponding network addresses, the router may relay this fact to the host. This is typically useful only when the destinations share a common sub-



■ **Figure 6.** *An IS-IS area.*



■ **Figure 7.** *IS-IS NSAP address structure.*

network with the source.

Dynamic Address Assignment — Network addresses cannot be preconfigured in the way that subnetwork addresses typically are (using information stored in ROM) because they are topologically significant — the location of a particular host in the global network cannot be predetermined at the factory.

The ability for hosts to learn their own network addresses without direct human intervention has a number of useful properties. The most obvious is that it reduces the complexity of managing a large number of hosts — the less information that needs to be set up on a host-by-host basis, the greater the probability that it will be done properly (and with less administrative overhead as well). In addition, it is useful to be able to change the network addresses of all hosts in a network, something that is highly onerous with existing networks.

ES-IS provides a mechanism for hosts to determine their network addresses dynamically. If a host wishes to find out its NSAP address, it multicasts a request address packet to the "all routers" subnetwork address. One or more routers (or perhaps an address assignment entity that resides on the subnetwork) may then respond directly to the host (via a unicast) with an assign address packet, which contains a network address to use as well as a holding time during which the address is valid. The host then chooses among the possible multiple responses through local means.

An entity sending an address assignment may not assume that the host has chosen that particular address; that binding occurs later when the host sends an ES hello. The assignor may not re-use the address until the holding time expires, whether or not it ever hears an ES hello.

The method with which the assignor determines the network address is deliberately left open in the standard. One obvious method is to build the address out of the host's subnetwork address (e.g., an IEEE 802 MAC address), combined with a prefix. This approach requires minimal manual configuration; in fact, a brand new system could be uncrated, plugged into the network, and be able to communicate without any manual configuration. Some may not view this ability as a virtue; another approach would be to keep very tight reins on the assignment of network addresses, only assigning them to those systems whose subnetwork addresses have been preconfigured in a server.

Intradomain Routing

Routing within a routing domain is handled by the intermediate system-intermediate system (IS-IS) protocol [ISO 10589]. IS-IS is a member of the link state protocols class.

IS-IS views a routing domain as a connected set of areas, each of which is a connected set of routers and hosts (see Fig. 6).

NSAP addresses are minimally constrained by IS-IS to having a fixed-length system ID adjacent to the NSEL. IS-IS expects NSAP addresses to have the format depicted in Fig. 7.

The System ID must be unique within the area. The IS-IS protocol allows system IDs of lengths one to eight octets (inclusive), but existing implementations currently fix the size at six octets. This allows the use of an IEEE 802 MAC address as a system ID, if desired.

Each area may have multiple area addresses, allowing for phased address changes for an entire domain, as well as for the splitting and combining of areas. Area addresses need not share any common prefix.

IS-IS has a three-tiered view of routing:

- Intra-area (Level 1).
- Inter-area (Level 2).
- Exterior.

Within an area, addressing and routing is flat. This requires flooding routes for each host within the area, but allows a host to be moved anywhere within the area without having to change its address. To reduce routing overhead only the system ID portion of the address is distributed. No routing information about individual hosts ever leaves an area. Conversely, no routing information about other areas, or about destinations outside the routing domain, ever enters the area. A router internal to an area (Level 1) knows only whether the destination is inside the area (the address matches one of the area addresses) or outside the area; if it is outside the area, the packet is forwarded toward the nearest inter-area router.

Level 2 (inter-area) routers operate over a flat space of area addresses within the domain (Fig. 8). Information about all area addresses within the routing domain is flooded to all Level 2 routers, as well as any exterior (interdomain) routing information imported into the domain. If a destination is within the routing domain, a Level 2 router will forward packets to the nearest entry point into the destination area, whereupon the packet will be forwarded according to Level 1 routing.

Routes to destinations outside the routing domain are represented as address prefixes with four-bit granularity. Packets destined outside the domain are forwarded along the route with the longest prefix that matches the destination. If there are multiple paths to the destination, the best path can be selected either by virtue of the nearest exit

point from the domain, or by an external metric (which is likely to reflect a ranking of the quality of interdomain paths).

Within a routing domain, the Level 2 and each of the Level 1 topologies are expected to be connected. If a Level 1 area becomes partitioned, however, the partition may be healed at Level 2 so long as each component of the partitioned area still has Level 2 connectivity. Partitions at Level 2 cannot be healed completely with either Level 1 routing or interdomain routing.

Area boundaries are modeled as occurring on links (rather than within a router). This implies that a single router cannot be used to connect two areas, a supposition that is borne out in existing implementations. However, it is feasible to implement a single router that can route between two areas, should such a capability prove necessary. Due to the flexible nature of area addressing, it has not yet appeared worthwhile to do so.

Interdomain Routing

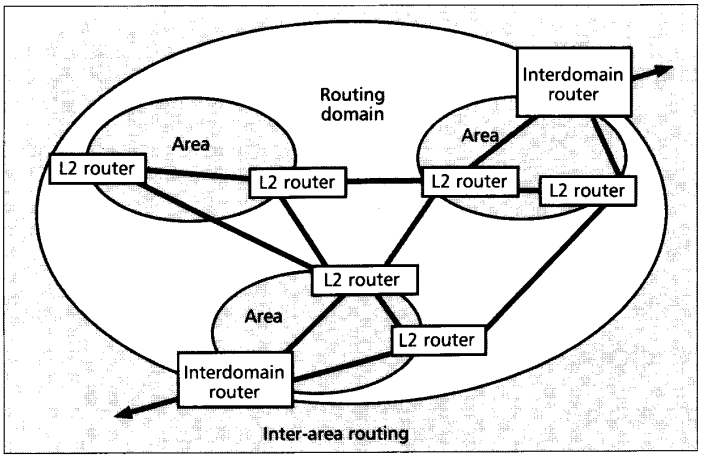
The final component in the routing mechanism is the interdomain routing protocol (IDRP) (Fig. 9) [11]. This is the newest piece of work in the OSI routing architecture, and is the only one that is not a full international standard (although it is expected to become one during 1993).

IDRP is based on the IP border gateway protocol [13], and has been generalized to be a protocol-independent interdomain routing protocol. It is likely that protocols other than CLNP will be routed using IDRP in the future. Work already has begun on specifying the use of IDRP for the interdomain routing of IP, and interest has been expressed in using IDRP for other routable protocols such as Novell's IPX and Appletalk. No changes to the IDRP protocol itself are necessary for use with other protocols.

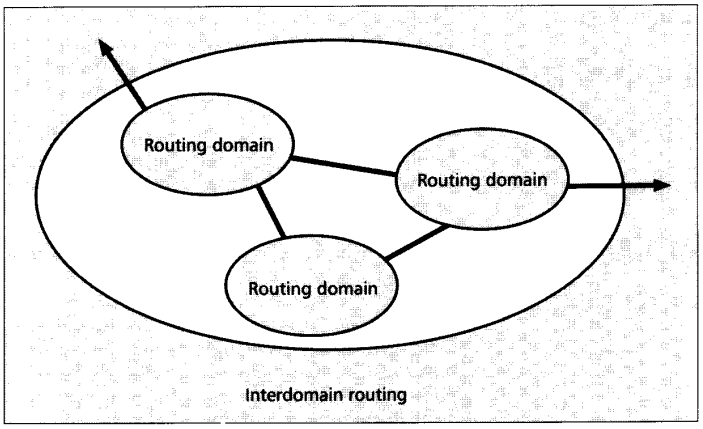
IDRP calculates routes according to policy constraints, rather than by the shortest path. This allows the interdomain topology physically to be an arbitrary mesh, while still providing a means for ensuring that calculated routes meet transit policy constraints. Policy information is not carried in IDRP. Instead, each routing domain is responsible for implementing its routing policies by first selecting among potential routes to a destination, and then by limiting the further distribution of the chosen route.

Each route carried within IDRP contains two components — a list of reachable destinations, and a set of path attributes. The most important path attribute is the RD path, a representation of the path of routing domains through which a route passes. This provides automatic loop suppression, since a routing domain cannot select an offered route that already traverses itself. Additionally, the RD path supplies data on which policy determination can be made. When the routing domain further distributes the selected route, it adds itself to the RD path.

Scalability to extremely large internetworks is provided through two mechanisms. The first is the representation of destinations as address prefixes with bitwise granularity. This mechanism allows many individual routes to be aggregated into a single prefix, effectively providing numerous levels of routing and vastly reducing the volume of routing data exchanged. When routes are aggregated, their RD paths are merged, ensuring protection against loops.



■ **Figure 8.** *Inter-area routing.*



■ **Figure 9.** *Interdomain routing with confederations.*

Simply allowing aggregation of reachable destinations is not sufficient to allow scaling, since the size of the RD path will continue to grow. Indeed, a default route (a zero-length prefix) will carry an RD path containing the set union of all routing domains in the internetwork. This problem is solved by the routing domain confederation (RDC). An RDC is a connected set of routing domains that are grouped together as a single topological entity. An RDC is essentially a SuperDomain, in the sense that, when viewed from the outside, an RDC is indistinguishable from a single routing domain. This eliminates all the detail of the inner workings of the RDC from the RD path, as the collection of member RDs is replaced with the identifier of the RDC. RDCs may be disjoint, nested (forming confederations of confederations), or may overlap.

Member routing domains of an RDC need not have coordinated routing policies beyond what is normally necessary to provide connectivity (i.e., the intersection of the policies should not be null if there is to be any transit traffic). This allows RDCs to be formed without a significant level of administrative and configuration over-

Using CLNP in place of IP requires that all the services of IP visible from the transport layer be mapped to CLNP, a fairly straightforward task.

head. The RDC mechanism automatically imposes the restriction that routes from one RDC member to another cannot exit the RDC (necessary to maintain the loop-free property of IDRP).

The multiprotocol capabilities of IDRP are provided by carrying destination address information and associated next hop network addresses as opaque data, qualified by protocol ID. Information for multiple protocols can be carried within a single instance of IDRP, as long as the routing domain boundaries of the various protocols are congruent. If this restriction cannot be satisfied, separate instances of IDRP can be used for each protocol.

Scaling Properties of NSAP Addressed Networks

Tractable scaling is only possible if routing complexity increases much more slowly than linearly concerning the number of destinations. This type of growth can be achieved if routing information can be separated hierarchically, with bounded complexity at each level of the hierarchy.

In the OSI routing architecture, the theoretical number of levels is bounded only by the number of addressing bits that map to real topological elements. In practical terms, however, the number of levels of routing is determined by the way that addresses are assigned, and by the actual topology of the Internet.

There is debate within the Internet community as to the best approach for address allocation, with some favoring a very geographical bias to address assignment (similar to the telephone numbering plan in North America), and some promoting embedding carrier identification into addresses. By definition, the latter approach has the advantage that addressing and topology are tightly coupled, providing very high efficiency and complexity reduction. The former approach decouples individual destinations from their carriers, potentially allowing easy migration between carriers, but places significant requirements on the relationships between carriers and their topological interconnection.

The two approaches have similar scaling properties. OSI routing can support either approach, or even both simultaneously. OSI routing and addressing do not impose a particular address format, other than the position of the system ID field, since all routing above the intra-area level is done based on address prefixes.

For purposes of illustration, we will assume the use of GOSIP-format addresses in a carrier-based addressing scheme. It is the case in the current operational Internet that routers in the central part of the infrastructure handle about 10,000 individual routes. This can be viewed as a feasible lower bound for route table complexity at any given level in the routing hierarchy.

At the low-order end of the address, six octets of system ID are used for intra-area routing. GOSIP then allocates two octets to define the area within a routing domain. Routers at this level are likely to be less capable than infrastructure routers, but should be able to easily handle 10^5 destinations per routing domain.

GOSIP next specifies a two-octet routing domain number within carrier. This could be treated as a flat space of 65,000 domains, or may be subdivided into multiple levels. Either way, this provides another order of 10^4 allowing about 10^9 destinations within a domain.

The next two octets currently are declared to be reserved. Additional levels of hierarchy could be inserted at this point, should it become necessary, providing routing over 10^4 superdomains, but for the moment this will not be considered.

The next three octets (the administrative authority identifier) identify the carrier. If we assume 10^4 carriers, we reach 10^{13} hosts within the GOSIP address space, which should be sufficient for the foreseeable future.

The GOSIP space is identified by the high order octets as being administered by the United States government. Other parts of the NSAP address space are administered by other nations, allowing another level of information reduction if necessary.

Mapping IP Functionality to CLNP

Using CLNP in place of IP requires that all the services of IP visible from the transport layer be mapped to CLNP. Since CLNP was originally derived from IP, this is a fairly straightforward task. Most functions of IP map directly to matching functions in CLNP [18]. A few aspects deserve further discussion.

Protocol Identification

IP carries a one-octet protocol identifier, which specifies the protocol being carried as payload in the data portion of the packet. CLNP does not have a separate protocol ID field, but instead treats this functionality as part of addressing, using a one-octet NSAP selector. For use with TUBA, the NSEL field in the destination NSAP address is set to the value normally carried in the IP protocol ID field (e.g., 6 = TCP, 17 = UDP). If OSI transport layer protocols are to be used with the same network address as TUBA, NSEL values must be used that do not conflict with existing IP transport layer protocols. A protocol ID value for the OSI connection-oriented transport protocol already has been assigned; another value can easily be assigned for the OSI connectionless transport protocol. Note that since NSEL values are viewed in OSI as being of local significance, changing existing NSEL usage to conform to these restrictions is consistent with the intended OSI usage.

Error Notification

Error conditions in IP, such as unreachable destinations, expiration of packet lifetime, etc., are reported using the Internet control message protocol (ICMP). CLNP has a separate packet type for error reporting. Reference 18 contains the mapping between ICMP message types and CLNP error report types.

TCP and UDP Pseudoheaders

The TCP and UDP protocols compute checksums using not only information contained in the transport layer segment, but also the source and destination IP addresses, the IP protocol ID, and total length of the transport layer segment. These transport protocols do not have separate connection identifiers and use network layer addresses in

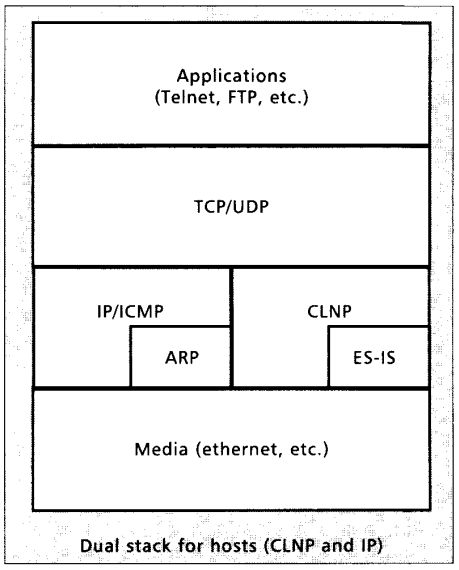


Figure 10. *Dual stacked host.*

part to demultiplex incoming packets. The checksum over the pseudoheader protects this information as part of transport layer semantics.

TUBA preserves these semantics by including the source and destination NSAP addresses, including their lengths and selector values (in which the transport protocol is identified), in the pseudoheader checksum.

Network Management

Network management in the TUBA environment is done using the simple network management protocol (SNMP)[2]. SNMP management information bases exist for CLNP and its associated routing protocols [24], and SNMP Version 2 contains explicit ASN.1 definitions for OSI addresses. SNMP can operate over either IP or CLNP in this environment. In addition, standard network-layer diagnostic tools such as ping [25] and traceroute are widely implemented.

Transition to TUBA

TUBA transition is a phased plan consisting of:

- Infrastructure networks (e.g., internet service providers) turn on CLNP capability in their routers.
- Root domain name servers deploy TUBA name services.
- TUBA host software becomes available.
- Local routing domains at sites turn on CLNP routing and register TUBA zones in their local domain name servers.
- Hosts running TUBA software register in their local domain name servers.

Transition Overview

TUBA defines a method for hosts to run Internet transport protocols over a CLNP infrastructure. The transition plan for TUBA specifies that Internet hosts go through two stages: the initial phase where a host only uses IP, and a second phase where a host is TUBA capable and TCP and UDP operate over both IP and CLNP (see Fig. 10). TUBA-capable hosts are dual-stacked. IP-only hosts use IP to communicate with all other hosts. TUBA-capable hosts use IP to talk to IP-only hosts, but they use CLNP when talking to other TUBA-capable hosts. Over time IP-only hosts will upgrade to TUBA-capable software and new systems will come equipped with TUBA-capable software. Hosts using IP strictly for local communication will not need to be upgraded.

The transition to TUBA has a strong dependency on continued operation of the IP infrastructure, since TUBA is a dual network layer strategy. The regional, national, and intercontinental IP networks form this IP infrastructure, making it a simple matter for a site to attach and gain worldwide IP connectivity. Initially, the TUBA transition augments the capabilities of the current Internet infrastructure to route and forward CLNP simultaneously with IP. Many infrastructure networks already carry CLNP (e.g., NSFNET, Alternet, ESnet, NSI, etc.). Most commercial routers already have CLNP routing and forwarding "out of the box," so most infrastructure networks are capable of managing CLNP traffic.[1] Several countries have CLNP-based initiatives and trials, and CLNP routing and addressing plans exist to provide guidance for Internet providers [4].

The TUBA transition requires each host to have both IP and CLNP addresses. Systems currently attached to the IP Internet will obtain NSAPs as they add TUBA software and attach to the CLNP Internet. Over time, most hosts will be TUBA-capable and use CLNP for network layer services. If the TUBA transition is vigorously pursued, the use of IP on the Internet could become vestigial before the exhaustion of IP addresses.

IP Address Space Exhaustion

If TUBA transition is not vigorously pursued, there may be a significant amount of IP traffic on the Internet at the time IP addresses become scarce. Once the IP address space runs out, hosts that are only capable of using IP will not be able to communicate with all Internet hosts, since the IP address space will be incapable of uniquely identifying each host in the Internet.

IP address exhaustion would force the existence of islands of IP hosts that cannot communicate with each other using the current Internet infrastructure. Before IP address space exhaustion, a block of IP addresses will be designated, such that addresses assigned from this block will not be globally unique, and these network addresses will not be routed across the Internet infrastructure. By administrative control, these addresses can be guaranteed to be unique within each local IP routing domain, so that hosts with these addresses can use IP to communicate with IP-only hosts within their own domain. New hosts will use CLNP to communicate globally, yet continue to use IP for communicating with IP-only hosts within their own IP routing domain. Old hosts may still use IP; however, they will reach an increasingly small subset of the Internet, since most systems will be using CLNP. It is important to remember that as the Internet grows, most systems will be new and TUBA-capable when purchased.

Host Name Resolution

Routing and forwarding of data traffic is not the only critical piece of Internet infrastructure. The domain name system [15] provides the operational name

■ ■ ■ ■ ■

The transition to TUBA has a strong dependency on continued operation of the IP infrastructure, since TUBA is a dual network layer strategy.

[1] *The current demand for CLNP services is not sufficient at present for some networks to turn on and manage CLNP.*

Document Availability

The TUBA mailing list is archived and available by anonymous ftp from merit.edu in the pub/tuba-archive directory. Subscription to the tuba list is available upon receipt of Internet mail to tuba-request@lanl.gov.

The full set of OSI connectionless network layer protocol specifications is available in electronic form via anonymous FTP. They can be found at merit.edu, in the pub/iso directory:

clnp.ps
esis.ps
isis.ps
idrp.ps

Request for comments (RFC) documents are available by anonymous ftp from nic.ddn.mil. Several of the documents referenced are internet-drafts that will become RFCs. They are available in the nic.ddn.mil anonymous ftp site.

space upon which Internet services depend. Internet domain names will be used with TUBA, so the DNS will need to be augmented to map between domain names and CLNP addresses, as well as IP addresses. Operationally, the root name servers will need to be upgraded to respond to DNS requests by TUBA-capable hosts, to respond with IP and CLNP addresses, and to delegate TUBA-based zones. Systems that are responsible for serving DNS zones will subsequently convert to TUBA-capable implementations of the DNS when hosts within that zone wish to become known to the TUBA Internet. Since the nature of TUBA capability is defined on a host-by-host basis, any host announcing that it is TUBA capable is announcing it can use CLNP for all services expected of that host.

Other Transition Issues

There are several Internet protocols and services that assume IP as the sole network layer. The most common assumption made is that IP addresses are used to identify hosts. The file transfer protocol (FTP) identifies the host IP address and port number to be used when opening a data connection. This is a common practice in protocols and systems that need to pass network bindings as data, such as remote procedure call protocols, authentication protocols, and protocols that involve third parties or proxies. These protocols and systems can be re-engineered to eliminate their dependence on a single protocol by passing the name of the system instead of a network layer address. Another option would be to extend these protocols to pass network layer information (e.g., addresses and port numbers) and an identifier for the network layer protocol.

It is necessary to engineer applications to simultaneously offer services over both IP and CLNP stacks. Hosts can run two separate applications offering the same service, one over IP and the other over CLNP. Alternatively, a host could provide an application programming interface (API) allowing a single version of the application to access both CLNP and IP network services through the same interface. The latter system architecture is preferable, since it reduces the number of separate instances of essentially identical software.

CLNP-Incapable Hosts

Some hosts never will use CLNP, either because they cannot, or because they choose not to convert to TUBA. Many of these hosts do not need to communicate with the entire Internet, and will only need to communicate with systems within some limited scope (e.g., their local ethernet or routing domain). These hosts are not considered by the TUBA plan, since TUBA expressly addresses the problem of globally interconnected sites. As discussed earlier, hosts with globally unique IP addresses still will be able to communicate with one another.

There may be cases where IP-only hosts without globally unique IP addresses require global Internet connectivity, but will be unable to use CLNP. It is possible to translate between IP datagrams and CLNP datagrams, provided there is a simple mapping between the IP address and CLNP NSAP. A network layer translating gateway (NLTG) could convert TCP and UDP packets over IP to TCP and UDP packets over CLNP. NLTGs are prosthetics for hosts that cannot be converted to TUBA, and are not considered to be in the mainstream of TUBA transition planning.

Current Status of TUBA

The bulk of implementation work for TUBA requires changing host software. There are trial implementations on five separate hardware platforms including personal computers, workstations, and routers. Interoperability testing has successfully demonstrated interoperation of Telnet, TFTP, and Finger over TUBA.

Conclusion

Eventually it will be necessary to provide the means for addressing more hosts than is possible with the current version of the IP. OSI NSAP addresses can address very large Internets, and with TUBA it is possible to use the applications and protocols that Internet users currently enjoy, running over CLNP in place of IP. TUBA is a pragmatic solution to IP's lack of adequate address space, taking advantage of the current investment in Internet applications and protocols, as well as the development, testing, and deployment of CLNP. We envision a two-step transition to a TUBA Internet. Initially, IP transit networks add CLNP services, a step that already is in progress by many transit networks. Subsequently, a longer phase where Internet services transition from using only IP to delivery of the same services using both IP and CLNP.

Acknowledgments

The authors would like to thank Dave Piscitello (Bellcore), Yakov Rekhter (T.J. Watson Research Center, IBM Corp.), Sue Hares (Merit/NSFNET), Mark Knopper (Merit/NSFNET), Richard Colella (NIST), and the anonymous reviewers for comments that significantly improved this paper.

Peter Ford acknowledges support from the United States Department of Energy's Office of Energy Research (Contract No. KC0702), Los Alamos National Laboratory (Contract No. W-7405-ENG-36) and the National Science Foundation.

References

[1] R. Callon, "TCP and UDP with Bigger Addresses (TUBA), A Simple Proposal for Internet Addressing and Routing," RFC 1347, June 1992.
[2] J. D. Case *et al.*, "Simple Network Management Protocol (SNMP)," RFC 1157, May 1990.
[3] V. G. Cerf and R. E. Kahn, "A Protocol for Packet Network Intercommunication," *IEEE Trans. Commun.*, vol. COM-22, no. 5, May 1974.
[4] R. Colella and R. Callon, "Guidelines for OSI NSAP Allocation in the Internet," RFC 1237, July 1991.

[5] V. Fuller *et al.*, "Supernetting: An Address Assignment and Aggregation Strategy," RFC 1338, June 1992.
[6] "Information Processing Systems — Data Communications — Network Service Definition Addendum 2: Network Layer Addressing," ISO 8348/Addendum 2, 1988.
[7] "Protocol for Providing the Connectionless-mode Network Service," ISO 8473, 1988.
[8] "OSI Routing Framework," ISO TR 9575, 1989.
[9] "End System to Intermediate System Routing Exchange Protocol for Use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)," ISO 9542, 1988.
[10] "Intermediate System to Intermediate System Intra-domain Routing Information Exchange Protocol for Use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589, 1992.
[11] "Protocol for Exchange of Inter-domain Routing Information Among Intermediate Systems to Support Forwarding of ISO 8473 PDUs," ISO/IEC DIS 10747, 1992.
[12] L. Kleinrock and K. Farouk, "Hierarchical Routing for Large Networks," *Computer Networks 1* (North-Holland Publishing Company, 1977).
[13] K. Lougheed, ed. and Y. Rekhter, "A Border Gateway Protocol 3 (BGP-3)," RFC 1267, Oct. 1991.
[14] B. Manning, "DNS NSAP RRs," RFC 1348, July 1992.
[15] P. V. Mockapetris, "Domain Names — Implementation and Specification," RFC 1035, Nov. 1987.
[16] NIST, "U.S. Government Open Systems Interconnection Profile (GOSIP) Version 2.0," Oct. 1990.
[17] D. Oran, "OSI IS-IS Intra-domain Routing Protocol," RFC 1141, Feb. 1990.
[18] D. Piscitello, "Use of ISO CLNP in TUBA Environments," Internet Draft draft-ietf-tuba-clnp-02.txt, Jan. 1993.
[19] J. B. Postel, "User Datagram Protocol," RFC 768, Aug. 1980.
[20] J. B. Postel, "Internet Protocol," RFC 791, Sept. 1981.
[21] J. B. Postel, "Transmission Control Protocol," RFC 793, Sept. 1981.
[22] Y. Rekhter, ed. and T. Li, "A Border Gateway Protocol 4 (BGP-4)," Internet Draft, Dec. 1992.
[23] Y. Rekhter and T. Li, "An Architecture for IP Address Allocation with CIDR," Internet Draft, Feb. 1993.
[24] G. Satz, "CLNS MIB for Use with Connectionless Network Protocol (ISO 8473) and End System to Intermediate System (ISO 9542)," RFC 1238, June 1991.
[25] C. Wittbrodt, ed., "An Echo Function for ISO 8473," Internet Draft draft-ietf-noop-echo-00.txt, Feb. 1993.

Biographies

PETER S. FORD received a B.G.S. from the University of Michigan. He is a member of the technical staff at Los Alamos National Laboratory, Los Alamos, New Mexico, where he works on computer networking and high performance computing. He is currently working with the National Science Foundation on the evolution of the NSFNET.

DAVE KATZ has been a software engineer at cisco Systems in Menlo Park, California since 1992. Previously he worked on the NSFNET project at Merit, Inc. in Ann Arbor, Michigan. He has been active in standards activities since 1986, involved with both TCP/IP (IETF) and OSI (ANSI X3S3.3), concentrating on routing issues.