A Quarterly Technical Publication for Internet and Intranet Professionals

In This Issue

You can download IPJ
back issues and find
subscription information at:
www.cisco.com/ipj

F R O M   T H E   E D I T O R

If you are reading the printed version of this journal you will notice a subtle change in the paper. This issue is printed on an uncoated stock, specifically Exact® Offset Opaque White 60#, a recycled paper made by Wausau Paper Corporation. This paper is slightly thinner, and thus lighter, than the paper we have been using. It is also less reflective and easier to write notes on. We invite your feedback on this paper as we experiment with various solutions to reduce our carbon footprint. As always, send your comments to: **ipj@cisco.com**

This journal has a long history of covering existing and emerging technologies that form part of the underlying infrastructure for both the global Internet and private enterprise networks. Recent articles have focused on wireless systems such as WiMAX, and we have other articles on wireless technologies in the pipeline. This time, however, we look at *optical networking*, specifically *Generalized Multiprotocol Label Switching* (GMPLS) as a technology for next-generation internets. The article is by Francesco Palmieri.

The topic of IP Version 4 address exhaustion has been discussed in several articles in this journal, and is currently being heavily debated in the *Regional Internet Registries* (RIRs). As we approach the inevitable date when the IPv4 address pool "runs out," we are returning to this topic with several articles. The first of these articles is included in this issue. Geoff Huston sets the stage by reviewing some of the history and answering the basic question of "why" we find ourselves at a point in history where the IPv4 addresses will run out before we have deployed any significant amount of IPv6 systems. In future issues we will follow Geoff's introduction with several other perspectives on this situation.

Once again, let me remind you to visit our Website at **http://www.cisco.com/ipj**, where you can renew and update your subscription, download back issues, and find additional resources such as our online forum at **http://ipjforum.org**

—*Ole J. Jacobsen, Editor and Publisher*
ole@cisco.com

GMPLS Control Plane Services in the Next-Generation Optical Internet

by Francesco Palmieri, Federico II University of Napoli, Italy

One of the major concerns in the Internet-based information society today is the tremendous demand for more and more bandwidth. Optical communication technology has the potential for meeting the emerging needs of obtaining information at much faster yet more reliable rates because of its potentially limitless capabilities—huge bandwidth (nearly 50 terabits per second[1]), low signal distortion, low power requirement, and low cost. The challenge is to turn the promise of optical networking into reality to meet our Internet communication demands for the next decade. With the deployment of *Dense Wavelength Division Multiplexing* (DWDM) technology, a new and very crucial milestone is being reached in network evolution. The speed and capacity of such wavelength switched networks—with hundreds of channels per fiber strand—seem to be more then adequate to satisfy the medium to long term connectivity demands. In this scenario, carriers need powerful, commercially viable and scalable devices and control plane technologies that can dynamically manage traffic demands and balance the network load on the various fiber links, wavelengths, and switching nodes so that none of these components is over- or underused.

This process of adaptively mapping traffic flows onto the physical topology of a network and allocating resources to these flows—usually referred to as *traffic engineering*—is one of the most difficult tasks facing Internet backbone providers today. *Generalized Multiprotocol Label Switching* (GMPLS) is the most promising technology. GMPLS will play a critical role in future IP pure optical networks by providing the necessary bridges between the IP and optical layers to deliver effective traffic-engineering features and allow for interoperable and scalable parallel growth in the IP and photonic dimension. The GMPLS control plane technology, when fully available in next-generation optical switching devices, will support all the needed traffic-engineering functions and enable a variety of protection and restoration capabilities, while simplifying the integration of new photonic switches and existing label switching routers.

Wavelength Division Multiplexing

Traditional *Electronic Time-Division Multiplexed* (ETDM) networks use an electrical signal form to switch traffic along routes and restore signal strength. These networks do not fully exploit the bandwidth available on optical fibers because only a single frequency (wavelength or *lambda*) of light is used on each fiber to transmit data signals that can be modulated at a maximum bit rate of the order of 40 Gbps. The high bandwidth of optical fibers can be better used through WDM technology by which distinct data signals may share an optical fiber, provided they are transmitted on carriers having different wavelengths[2].

In more detail, the optical transmission spectrum is divided into numerous nonoverlapping wavelengths, with each wavelength supporting a single communication channel. Each channel, which can be viewed as a *light path,* is transmitted at a different wavelength (or frequency). Multiple wavelengths are multiplexed into a single optical fiber and multiple light-path data is transmitted as shown in Figure 1.

Figure 1: WDM Functional Model
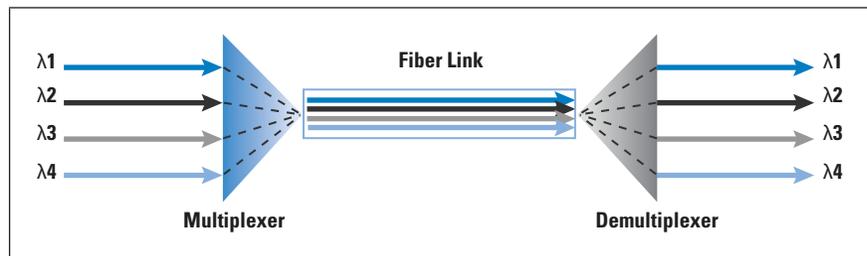
Dense WDM (DWDM), an evolution of WDM referring essentially to the closer spacing of channels, is the current favorite multiplexing technology for long-haul communications in modern optical networks. Hence, all the major carriers today devote significant effort to developing and applying DWDM technology in their business.

All-optical networks employing the concept of WDM and wavelength routing are thought to be the transport networks for the future[3]. In such networks, two adjacent nodes are connected by one or multiple fibers, each carrying multiple wavelengths or channels. Each node consists of a dynamically configurable optical switch that supports fiber switching and wavelength switching; that is, the data on a specified input fiber and wavelength can be switched to a specified output fiber on the same wavelength[4]. In order to transfer data between source–destination node pairs, a light path needs to be established by allocating the same wavelength throughout the route of the transmitted data. Benefiting from the development of all-optical amplifiers, light paths can span more than one fiber link and remain entirely optical from end to end. It has been demonstrated that the introduction of wavelength-routing networks not only offers the advantages of higher transmission capacity and routing node throughput, but also satisfies the growing demand for protocol transparency and simplified operation and management[3] [5].

Optical Transport Backbones
The modern Internet transport infrastructure can be physically seen as a very complex mesh of variously interconnected optical or traditional ETDM subnetworks, where each subnetwork consists of several heterogeneous routing and switching devices built by the same or different vendor and operating according to the same control plane protocols and policies. With these very different types of devices, all the forwarding decisions will be based on a combination of packet or cell, timeslot, wavelengths, or physical ports, depending on the position (edge or core) and role (intermediate or termination or gateway node) of the switching devices in the network layout.

In particular, WDM-switched optical subnetworks are typically used as backbone infrastructures to interconnect a large number of different IP as well as other packet networks such as SDH, ATM, and Frame Relay.

New optical devices such as DWDM multiplexers, *Add/Drop Multiplexers* (ADM), and *Optical Cross-Connects* (OXC) are making possible an intelligent all-optical core where packets are routed through the network without leaving the optical domain. The optical network and the surrounding IP networks are independent of each other, and an edge IP router interacts with its ingress switching node only over a well-defined *User-Network Interface* (UNI). Clearly, the optical network is responsible for setting up light paths between the edge IP routers. A light path can be either switched or permanent. Switched light paths are established in real time using proper signaling procedures, and they may last for a short or a long period of time. Permanent light paths are set up administratively by subscription, and they typically last for a very long time. An edge IP router requests a switched light path from its ingress optical switching device using a proper signaling protocol over the UNI. See Figure 2.

*Figure 2: The Optical Transport Infrastructure*
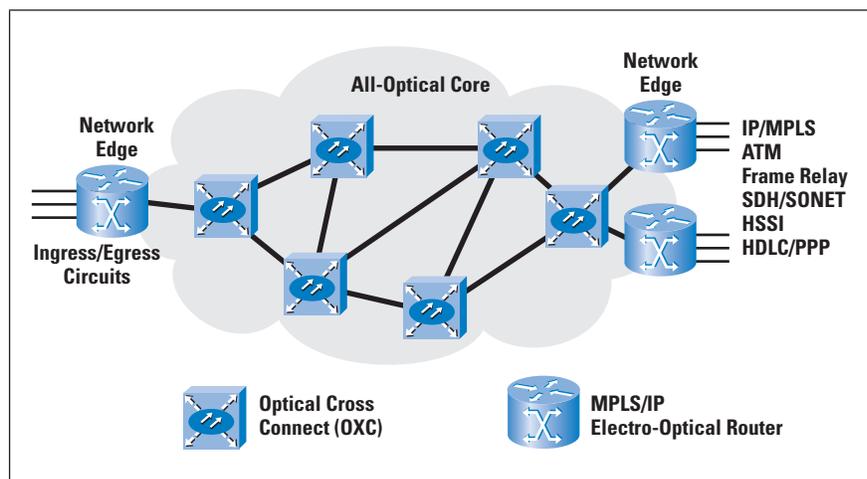
The key concept to guarantee desirable speeds and correct functional behavior in these networks is to maintain the signal in pure optical form, thereby avoiding the prohibitive overhead of conversion to and from electrical form. Such a network would be "optical transparent" in the sense that it would be able to transport client signals with any format and with a wide range of bit rates (at least from about 10 Mbps to more than 10 Gbps). In particular, transparent OXCs, used to selectively switch wavelengths between their input and output ports, are likely to emerge as the preferred option for switching multigigabit or even terabit data streams, because any slow electronic per-packet processing is avoided.

Transparent Optical Switching Nodes

Transparent OXC systems are expected to be the cornerstone of the photonic layer, offering carriers more dynamic and flexible options in building network topologies with enhanced performance and scalability. The development of large and flexible transparent OXCs, now enabled by a new generation of optical components such as optical amplifiers, tunable lasers, and wavelength filters, is still a significant challenge[1]. Their architecture makes use of optical switching fabrics, wavelength multiplexers and demultiplexers, and transparent wavelength converters, which eliminate the need for optoelectronic transponders. A simple and linear architectural model for an optical transparent OXC is shown in Figure 3.

Figure 3: OXC Architectural Model
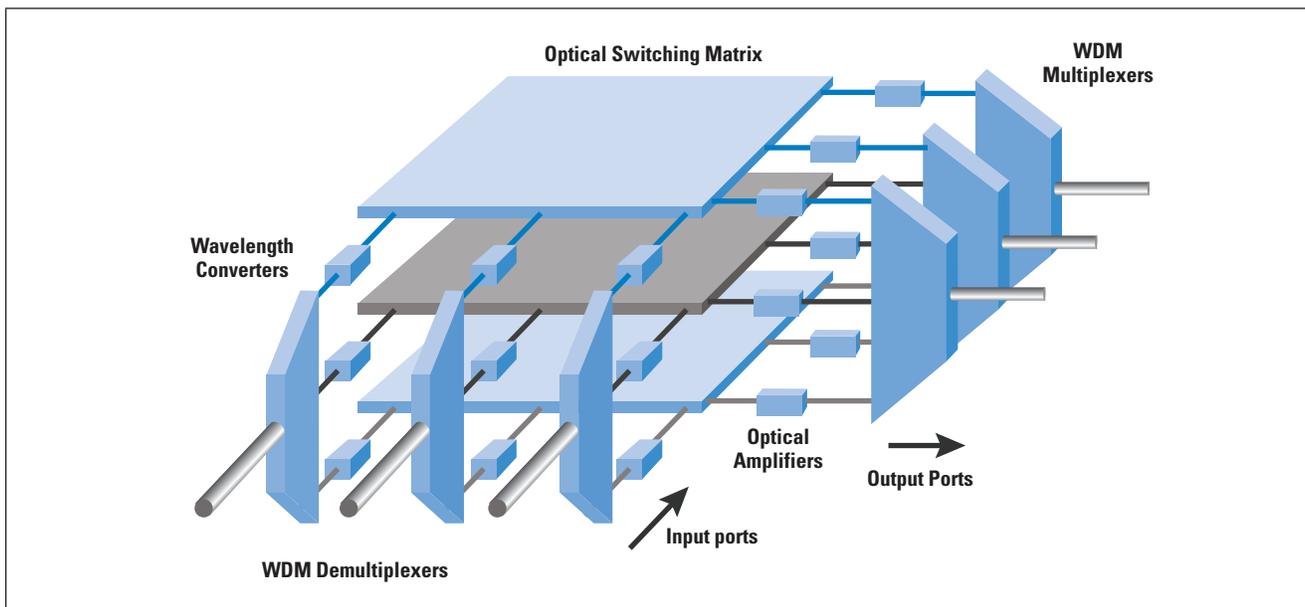
Here, the WDM demultiplexers separate incoming grouped wavelengths from input ports into individual lambdas. A sufficiently large low-loss connectivity and compact-design, all-optical switching fabric can be realized by using the reflection of light and *Micro-Electromechanical Systems* (MEMS) technology, now widely available on the market. This multilayer switching fabric driven by a micro-machined electrical actuator redirects, according to the control plane instructions, each wavelength into appropriate output ports passing through optical amplifiers, typically *Erbium-Doped Fiber Amplifiers* (EDFAs), which boost the signal power in line without the need for any optoelectronic conversion to cope with the effects of light dispersion and attenuation on long distances. The WDM multiplexer then groups the wavelengths from the above multiple layers of cross-connects. Furthermore, the wavelength that arrives into an OXC can be directly passed to the optical switching fabric, to be switched to the appropriate output fiber or previously converted, based on the control plane instructions, to another particular wavelength with the use of a tunable wavelength converter (without being transformed to electricity) if the former output wavelength is not available.

This architecture is transparent; that is, the optical signal does not need to be transformed to electricity at all, implying that this architecture can support any protocol and any data rate. Hence, possible upgrades in the wavelength transport capacity can be accommodated at no extra cost. Furthermore, this architecture decreases the cost because it involves the use of fewer devices than the other architectures. In addition, transparent wavelength conversion eliminates constraints on conversions. In this way the real switching capacity of the OXC is increased, leading to cost reduction. First-generation OXCs require manual configuration. Clearly, an automatic switching capability allowing optical nodes to dynamically modify the network topology based on changing traffic demand is highly desirable.

Automatically Switched Optical Networks

For automatically switched networks, where network nodes may directly initiate or terminate new connections or perform wavelength-level switching in the network, sophisticated and flexible control functions are needed.

The *control plane* supports connection management by clients and also provides protection and restoration services. The control plane of an optical network is also responsible for tracking the network topology and for notifying the state of the network resources. Two families of protocols achieve this task:

- *Routing protocols* are specifically responsible for the reliable advertisement of the optical network topology and the available bandwidth resources within and between network domains. In particular, some areas are relevant within this context: the bundling of links with equivalent or logically bundled characteristics, the definition of the routing areas in an optical domain, the rich specifications of an optical link resource as opposed to a typical advertisement of the up or down interface of IP networks, and the advertisement of the shared risk group (optical fibers flowing in the same cable or duct) to which an optical connection belongs.

- *Signaling protocols* are responsible for provisioning, maintaining, and deleting connections. Optical networks are characterized by connection-oriented paradigms that require a resource reservation protocol. State-of-the-art control plane technologies operating on traditional IP-based networks focus on soft-state protocols that require periodic refresh throughout the participating nodes. In optical networks, where the data plane is separated from the control plane, a possible solution is also to adopt a hard state reservation protocol without periodic refresh to limit the effect caused on the data plane by a failure in the control plane. Furthermore, redundant, generalized label binding is encouraged to reserve protection paths in the mesh network.

Data transport is the most obvious task and the main purpose of an optical network *data plane*. It provides uni- or bidirectional information transport (transmission and switching) between users, detects faults, and monitors signal quality. More specifically, the data plane performs, under the directions of the control plane, data routing to the appropriate ports; channel adds and drops to external, older networks (using the edge interfaces); and label or lambda swapping through an array of WDM demultiplexers, wavelength converters, OXCs, optical amplifiers, and multiplexers.

An important concern that must be addressed in designing an optical network is the cross effect of the failure of a data or control plane. Failures of the data plane are usually addressed by the control plane itself by rerouting the disrupted flows at the appropriate level. The control plane must then advertise quickly the new network state to the neighboring nodes to avoid the presence of stale information in the link databases. A failure of the IP-based control plane usually significantly affects the data plane.

Traffic Engineering in Optical Networks
Traffic engineering should be viewed as assistance to the routing and switching infrastructure that provides additional information used in forwarding traffic along alternate paths across the network, trying to optimize service delivery throughout the network by improving its balanced usage and avoiding congestion caused by uneven traffic distribution. Traffic engineering is required in the modern Internet mainly because the current dynamic routing protocols always use the shortest paths to forward traffic. This practice, obviously, conserves network resources, but it causes some of them to be overused while the other resources remain underused. Furthermore, the routing protocols mentioned earlier never account for specific traffic flow requirements such as bandwidth and *Quality of Service* (QoS) needs. Practitioners in the field often assert that traffic engineering essentially signifies the ability to place traffic where the capacity exists to accommodate it—whereas network engineering denotes the ability to install capacity where the traffic exists.

When a traffic-engineering application implements the right set of features, it should provide precise control over the placement of traffic flows within a routing and switching domain, gaining better network use and realizing a more manageable network. A traffic-engineering solution suitable for transparent optical networks always consists of numerous basic functional components; for example:

- *Traffic monitoring, analysis, and aggregation*—This function collects traffic statistics from the network elements; for example, the OXCs. Then the statistics are analyzed or aggregated to prepare for the traffic engineering and network reconfiguration related to decision making.

- *Bandwidth demand projection*—Bandwidth demand projection estimates the bandwidth requirements in the near future based on past and present measurements and the characteristics of the traffic arrival processes. The bandwidth projections are used for subsequent allocation.

- *Reconfiguration trigger*—This variable consists of a set of policies that decide when a network-level reconfiguration is performed. This decision is based on traffic measurements, bandwidth predictions, and operational areas; for example, to suppress the influence of transitional factors and reserve adequate time for the network to converge.

- *Topology design*—Topology design provides a network topology based on the traffic measurements and predictions. Conceptually this process can be considered as optimizing a graph (that is, OXC connected by light paths at the WDM layer) for specific objectives (for example, maximizing throughput), subject to certain constraints (for example, nodal degree or interface capacity), for a given load matrix (that is, traffic load applied to the network.) This area is, in general, a NP-hard problem. Because reconfiguration is regularly triggered by continually changing traffic patterns, an optimized solution may not be stable. It may be more practical to develop heuristics that place more emphasis on factors such as fast convergence, and less on ongoing traffic, rather than on optimality.

- *Topology migration*—Topology migration consists of algorithms to coordinate the network migration from an old topology to a new one. Because WDM reconfiguration deals with large-capacity channels, changing allocation of channel resources in this coarse granularity significantly affects a large number of end-user flows. Traffic flows have to adapt to the light-path changes at and after each migration step. These effects can potentially spread over the routing pattern of the network, in turn possibly affecting more user flows.

Traditionally, all provisioning and engineering in optical networks has required manual planning and configuration, resulting in setup times of days or even weeks and a marked reluctance among network managers to de-provision resources in case doing so would affect other services. In the last few years, during which control protocols have been deployed to dynamically provide traffic engineering and provisioning or management assistance in optical networks, the control protocols have been proprietary and have greatly suffered from interoperability problems. Consequently, a new standardized control plane framework, supporting evolutionary traffic-engineering features, is needed for automatically switched optical transport networks to foster the expedited development and deployment of a new class of versatile optical switches that specifically address the optical transport needs of the Internet.

The important remaining challenge to be addressed in developing a dynamically reconfigurable optical network is that of controlling the optical resources, especially under distributed control where the network elements exchange information among themselves in a standardized multivendor environment. Performance and reliability requirements make this challenge of paramount importance to photonic networks. Beyond eliminating proprietary "islands of deployment," this common control plane enables independent innovation curves within each product class, and faster service deployment with end-to-end provisioning using a single set of semantics.

The GMPLS Paradigm

GMPLS, the emerging paradigm for the design of control planes for OXCs, aims to address and solve all the challenges mentioned previously, trying to automatically and dynamically configure any kind of network element. It was proposed shortly after *Multiprotocol Label Switching* (MPLS) to extend its packet control plane to encompass time division (for example, for SONET/SDH), wavelength (for optical lambdas) and spatial switching (for example, for incoming port or fiber to outgoing port or fiber). Nongeneralized MPLS overlays a packet-switched IP network to facilitate traffic engineering and allow resources to be reserved and routes predetermined. It provides virtual links or tunnels through the network to connect nodes that lie at the edge of the network. For packets injected into the ingress of an established tunnel, normal IP routing procedures are suspended; instead the packets are label-switched so that they automatically follow the tunnel to its egress.

With the success of MPLS in packet-switched IP networks, optical network providers have accelerated a process to generalize the applicability of MPLS to cover all-optical networks as well. The premise of GMPLS is that the idea of a label can be generalized to be anything that is sufficient to identify a traffic flow. For example, in an optical fiber whose bandwidth is divided into wavelengths, the whole of one wavelength could be allocated to a requested flow. The *Label Switch Routers* (LSRs) at either end of the fiber simply have to agree on which frequency to use. From a control plane perspective, an LSR bases its functions on a table that maintains relations between incoming label or port and outgoing label or port. It should be noted that in the case of the OXC, the table that maintains the relations is not a software entity but it is implemented in a more straightforward way, for example, by appropriately configuring the micro-mirrors of the optical switching fabric.

There are several constraints in reusing the GMPLS control plane. These constraints arise from the fact that LSRs and OXCs use different data technologies. More specifically, LSRs manipulate packets that bear an explicit label, and OXCs manipulate wavelengths that bear the label implicitly; that is, the label value is implicit in the fact that the data is being transported within the agreed frequency band.

Furthermore, because the analogy of a label in the OXC is a wavelength or an optical channel, there are no equivalent concepts of label merging nor label push and pop operations in the optical domain, and label swapping can be realized through wavelength conversion. The transparency and multiprotocol properties of such a control plane approach would allow an OXC to route optical channel trails carrying various types of digital payloads (including IP, ATM, SDH, etc.) coherently and uniformly.

GMPLS Control Plane Functions and Services

GMPLS focuses mainly on the control plane services that perform connection management for the data plane (the actual forwarding logic) for both packet-switched interfaces and non-packet-switched interfaces. The GMPLS control plane essentially facilitates four basic functions:

- *Routing control*—Provides the routing capability, traffic engineering, and topology discovery

- *Resource discovery*—A mechanism to keep track of the system resource availability such as bandwidth, multiplexing capability, and ports

- *Connection management*—Provides end-to-end service provisioning for different services, including connection creation, modification, status query, and deletion

- *Connection restoration*—Implements an additional level of protection to the networks by establishing for each connection one or more presignaled backup paths and enabling very fast switching in case of failure between them.

The fundamental service offered by the GMPLS control plane is dynamic end-to-end connection provisioning. The operators need only to specify the connection parameters and send them to the ingress node. The network control plane then determines the optical paths across the network according to the parameters that the user provides and signals the corresponding nodes to establish the connection. The whole procedure can be done within seconds instead of hours. The other important service is bandwidth on demand, which extends the ease of provisioning even further by allowing the client devices that connect to the optical network to request the connection setup in real time as needed. In order to establish a connection that will be used to transfer data between a source–destination node pair, a light path needs to be established by allocating, in presence of the so-called *continuity constraint,* the same wavelength throughout the route of the transmitted data or selecting the proper wavelength conversion-capable nodes across the path. In fact, if the wavelength continuity constraint is not fully enforced, some wavelength conversion-capable nodes can be placed in the network to reduce the overall blocking probability in case of wavelength resource exhaustion on some nodes. Light paths can span more than one fiber link and remain entirely optical from end to end.

However, according to the mandatory clash constraint, two light paths traversing the same fiber link cannot share the same wavelength on that link. That is, each wavelength on a given fiber is not a sharable resource between light paths.

In general, if there are multiple feasible wavelengths (lambdas) between a source node and a destination node, then a Wavelength Assignment algorithm is required to select a wavelength for a given light path. The wavelength selection can be performed either after an optical route has been determined (in the so-called *decoupled approach*), or in parallel with finding a route. In the latter case, we refer to the coupled approach, in which the entire job is accomplished by a single *Routing and Wavelength Assignment* (RWA) algorithm. When light paths are established and taken down dynamically, routing and wavelength assignment decisions must be made as connection requests arrive to the network. It is possible that, for a given connection request, there may be insufficient network resources to set up a light path, in which case the connection request is blocked. The connection may also be blocked if there is no common wavelength available on all the links along the chosen route. Thus, the objective in the dynamic situation is to choose a route and a wavelength that maximizes the probability of setting up a given connection, while at the same time attempting to minimize the blocking for future connections.

In addition, because the quality of an optical signal degrades as it travels through several optical components and fiber segments, the deployment of "long-distance" light paths may require signal regeneration at strategic locations in a nationwide or global WDM network. As a result, the algorithms performing routing and wavelength assignment, virtual-topology embedding, wavelength conversion, etc. must also be mindful of the locations of the sparse signal regenerators in the network. Such regenerators, which are placed at select locations in the network, "clean up" the optical WDM signal either entirely in the optical domain or through an optoelectronic conversion followed by an electro-optic conversion. Thus the signal from the source travels through the network as far as possible before its quality drops below a certain threshold, thereby requiring it to be regenerated at an intermediate node. The same signal could be regenerated several times in the network before it reaches the destination.

Furthermore, in current multilayer transport networks the bandwidth demanded by traffic typically is orders of magnitude lower than the capacity of lambda links, and the number of available wavelengths per fiber is limited and costly. Hence, it is not worth assigning exclusive end-to-end light paths to these demands, so a better sub-lambda granularity is required. Thus, to increase the throughput of a network with a limited number of lambdas per fiber, *traffic grooming* is required in certain nodes, typically those on the network edge.

The GMPLS control plane ensures traffic-grooming capability on edge nodes by operating on a two-layer model; that is, an underlying pure optical wavelength routed network and an "optoelectronic" time-division multiplexed layer built over it. In the wavelength routed layer, operating exclusively at lambda granularity, when a transparent light path connects two physically adjacent or distant nodes, these nodes will seem adjacent for the upper layer. The upper layer can perform multiplexing of different traffic streams into a single wavelength-based light path through simultaneous time and space switching. Similarly it can demultiplex different traffic streams of a single lambda path. It can also perform remultiplexing: some of the demands demultiplexed can be again multiplexed into some other wavelength paths and handled together along it. This is due to the "generalized" and hence multilayer nature of the GMPLS control plane.

The electronic layer is clearly required for multiplexing packets coming from different ports. This upper electronic layer can be a classical or "next-generation" technology, such as IP/MPLS, but it can also be based on any other networking technology (that is SDH/SONET, ATM, Ethernet, etc.). However, the technology of the upper layer must be unique for all traffic streams that have to be demultiplexed and then multiplexed again, because the network cannot directly multiplex, for example, ATM cells with Ethernet frames.

Another service that gives greatest flexibility to users in handling their own virtual network topologies on the transport core is the *Optical Virtual Private Network* (OVPN), which allows users to have full network resource control of a defined partition of the carrier optical network. Although users have full network resource control of that portion of the network, the OVPN is just a logical network partition and the end users still do not have access and visibility to the carrier's networks. This service can save the carrier's operation resources by allowing end users to perform circuit provisioning and setup procedures.

GMPLS Interfaces

GMPLS encompasses control plane signaling for multiple interface types. The diversity of controlling not only switched packets and cells but also TDM network traffic and optical network components makes GMPLS flexible enough to position itself in the direct migration path from electronic to all-optical network switching. The five main interface types supported by GMPLS follow:

- *Packet Switching Capable* (PSC)—These interfaces recognize packet boundaries and can forward packets based on the IP header or a standard MPLS "shim" header.

- *Layer 2 Switch-Capable* (L2SC)—These interfaces recognize frame and cell headers and can forward data based on the content of the frame or cell header (for example, an ATM LSR that forwards data based on its *Virtual Path Identifier/Virtual Circuit Identifier* (VPI/VCI) value, or Ethernet bridges that forward the data based on the MAC header).

- *Time-Division Multiplexing-Capable* (TDMC)—These interfaces forward the data based on the time slot in a repeating cycle (for example, SDH cross-connect or ADM, interfaces implementing the Digital Wrapper G.709, and *Plesichronous Digital Hierarchy* [PDH] interfaces).

- *Lambda Switch-Capable* (LSC)—These interfaces are for wavelength-based MPLS control of optical devices and wavelength switching devices, such as *optical ADMs* (OADMs) and OXCs, operating at the granularity of the single wavelength or group of wavelengths (waveband). These interfaces forward the optical signal from an incoming optical wavelength to an outgoing optical wavelength. Traffic is forwarded based upon wavelength or waveband.

- *Fiber-Switch-Capable* (FSC)—These interfaces forward the signal from one or more incoming fibers to one or more outgoing fibers for spatial control of interface selection, automated patch panels, and physical fiber switching systems. Traffic is forwarded based on port, fiber, or interface.

These supported interfaces are hierarchal in structure and controlled simultaneously by GMPLS.

Generalized Label

GMPLS defines several new forms of label—the *generalized label* objects. These objects include the generalized label request, the generalized label, the explicit label control, and the protection flag. The generalized label can be used to represent timeslots, wavelengths, wavebands, or space-division multiplexed positions.

With plain MPLS labels embedded in the cell or packet structure for in-band control plane signaling, with the different kinds of interfaces supported by GMPLS it is impossible to embed label-specific information, in terms of fiber port or wavelength switching, into the traffic packet structure. Consequentially, new "virtual" labels have been added to the MPLS label structure. These virtual labels comprise specific indicators that represent wavelengths, fiber bundles, or fiber ports and are distributed to GMPLS nodes through out-of-band GMPLS signaling. GMPLS out-of-band signaling causes a control-channel separation problem.

With MPLS, the control information is found in the label, which is directly attached to the data payload. However, when you send the control information out of band, the label is separated from the data that it is attempting to control. GMPLS provides a means for identifying explicit data channels. Having the ability to identify data channels allows the control message to be associated with a particular data flow, whether it is a wavelength, fiber, or fiber bundle.

Generalized Label-Switched Paths

The handling of *label-switched paths* (LSPs) under GMPLS differs from that of MPLS. MPLS does not provide for bidirectional LSPs. Each direction LSP has to be established in turn. Under GMPLS, the LSP can be established bidirectionally. The traffic-engineering requirements for the bidirectional LSP are the same in both directions, and it is established for both directions through only one signaling message, allowing for reductions in latency-related setup time. In the optical environment, OXC translates label assignments into corresponding wavelength assignments and sets up *generalized LSPs* (G-LSPs) using their local control interfaces to the other switching devices. Subsequent to G-LSP setup, no explicit label or lambda lookup or processing operations are performed by the OXC nodes.

GMPLS supports traffic engineering by allowing the node at the network ingress to specify the route that a G-LSP will take by using explicit light-path routing. An explicit route is specified by the ingress as a sequence of hops and wavelengths that must be used to reach the egress, which is different from the hop-by-hop routing that is usually associated with PSC networks.

Figure 4: G-LSPs Ensuring Traffic Engineering
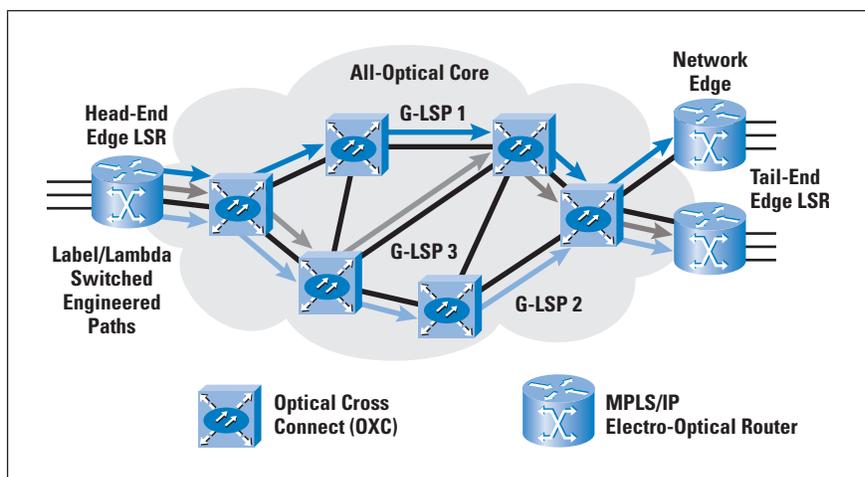
GMPLS also maintains the capability already available with MPLS to nest G-LSPs. Nested G-LSPs make possible the building of a forwarding hierarchy. At the top of this hierarchy are nodes that have FSC interfaces, followed by nodes that have LSC interfaces, followed by nodes that have TDMC interfaces, and followed by nodes with PSC interfaces. Nesting of G-LSPs between interface types increases flexibility in service definition and makes it possible for service providers operating a GMPLS network to deliver both bundled and unbundled services.

Because the deployment of DWDM equipment makes feasible the creation a large number of individual connections between two adjacent nodes, another very useful feature of bundling is the ability to simultaneously handle multiple adjacent links. Link bundling treats the traffic of these links as a single link.

In order for the adjacent links to be bundled, they must be on the same GMPLS segment, they must be of the same type, and they must have the same traffic-engineering requirements. These requirements reduce the amount of link advertisements that need to be maintained throughout the network, thereby increasing the control plane scalability. Just as in MPLS label stacking, GMPLS labels only contain information about a single level of hierarchy. The difference for GMPLS is that this hierarchy can be fiber-, wavelength-, timeslot-, packet- or cell-based.

For instance, if a connection is desired from one PSC interface to another PSC interface, and the traffic traverses physically separate fibers, a unique LSP has to be established for each level in turn. First, the FSC LSP, then the LSC LSP, then the TDMC LSP, and finally the PSC LSP have to be established through GMPLS signaling.

Signaling and Routing Protocols

In order to set up a light path, a signaling protocol is also required to exchange control information among nodes, to distribute labels, and to reserve resources along the path. In our case, the signaling protocol is closely integrated with the routing and wavelength assignment protocols. Suitable GMPLS signaling protocols for the GMPLS control plane include *Resource Reservation Protocol* (RSVP) and *Constraint-Based Label Distribution Protocol* (CR-LDP). Any of the objects that are defined within the GMPLS specification can be carried within the message of either of these signaling protocols that are responsible for all the connection management actions such as setup, modify, or remove the G-LSPs. Clearly, support for provisioning and restoration of end-to-end optical trails within a photonic network consisting of heterogeneous networking elements imposes new requirements for these signaling protocols. Specifically, optical trails require small setup latency (especially for restoration purposes), support for bidirectional trails, rapid failure detection and notification, and fast intelligent trail restoration.

Both RSVP and CR-LDP can be used to reserve a single wavelength for a light path if the wavelength is known in advance. These protocols can also be modified to incorporate wavelength selection functions into the reservation process[7]. In RSVP, signaling takes place between the source and destination nodes. The signaling messages may contain information such as QoS requirements for the carried traffic and label requests for assigning labels at intermediate nodes that reserve the appropriate resources for the path. CR-LDP uses TCP sessions between nodes in order to provide a hop-by-hop reliable distribution of control messages, indicating the route and the required traffic parameters for the route. Each intermediate node reserves the required resources, allocates a label, and sets up its forwarding table before backward signaling to the previous node.

To correctly perform resource reservation, allocation, and topology discovery on the available optical link resources, each node needs to maintain a representation of the state of each link in the network. The link state includes the total number of active channels, the number of allocated channels, and the number of channels reserved for light-path restoration. Additional parameters can be associated with allocated channels; for example, some light paths can be preemptable or have associated hold priorities. When the local inventory is constructed, the node engages in a routing protocol to distribute and maintain the topology and resource information. Standard IP routing protocols, such as *Open Shortest Path First* (OSPF) or *Intermediate System-to-Intermediate System* (IS-IS) with GMPLS Traffic Engineering extensions, can be used to reliably propagate the information.

The extensions to OSPF and IS-IS add additional information about links and nodes into the link-state database. Such information includes the type of LSPs that can be established across a given link (for example, packet forwarding, SONET/SDH trails, wavelengths, or fibers), as well as the current unused bandwidth, the maximum size of G-LSP that can be established, and the administrative groups supported. This information allows the node computing the explicit route for an LSP to do so more intelligently. Furthermore, any switching node cooperating in the GMPLS control plane will maintain a per-interface or per-fiber *Wavelength Forwarding Information Base* (WFIB) because lambdas and channels (labels) are specific to a particular interface or fiber, and the same lambda or channel (label) could be used concurrently on multiple interfaces or fibers.

Link Management Protocol

GMPLS also uses the *Link Management Protocol* (LMP) to communicate proper cross-connect information between the network elements. LMP runs between adjacent systems for link provisioning and fault isolation. It can be used for any type of network element, particularly in natively photonic switches. LMP automatically generates and maintains associations between links and labels for use in label swapping[6]. Automating the labeling process simplifies management and avoids the errors associated with manual label assignment. LMP provides control-channel management, link-connectivity verification, link-property correlation, and fault isolation. Control-channel management establishes and maintains connectivity between adjacent nodes using a keepalive protocol. Link verification verifies the physical connectivity between nodes, thereby detecting loss of connections and misrouting of cable connections. Fault isolation pinpoints failures in both electronic and optical links without regard to the data format traversing the link.

In order for these link bundles to be handled accordingly, GMPLS needed a method to manage the links between adjacent nodes. LMP was developed to address several link-specific problems that surfaced when generalizing the MPLS protocol across different interface types. The main responsibilities of the LMP follow:

- *Control-Channel Management*—Establishment of a control channel is critical to GMPLS signaling. The maintenance of the control channel between adjacent nodes must be able to exchange information related to LSP establishment.

- *Link-Property Correlation*—When link bundling occurs, GMPLS requires a way to verify that all traffic-engineering requirements are similar between links of adjacent nodes. Link-property correlation performs the verification and the aggregation of such links.

- *Link-Connectivity Verification*—This feature is used by GMPLS to verify the connectivity between data links when the control channel is separate from each data link.

- *Fault Management*—Fault management helps the network isolate faults down to the individual link.

Although LMP assumes the messages are IP encoded, it does not dictate the actual transport mechanism used for the control channel. However, the control channel must terminate on the same two nodes that the bearer channels span. Therefore, this protocol can be implemented on any OXC, regardless of the internal switching fabric. A requirement for LMP is that each link has an associated bidirectional control channel and that free bearer channels must be opaque (that is, able to be terminated); however, when a bearer channel is allocated, it may become transparent. Note that this requirement is trivial for optical cross-connects with electronic switching planes, but is an added restriction for photonic switches.

Conclusion

Innovations in the field of optical components will take advantage of the introduction of all-optical networking in all areas of information transport and will offer system designers the opportunity to create new solutions that will allow smooth evolution of all telecommunication networks. A new class of versatile IP-addressable optical switching devices is emerging, operating according to a common GMPLS-based control plane to support full-featured traffic engineering in modern optical transparent infrastructures.

The main advantage of this approach is that it is based on already existing and widely deployed protocols while simplifying network management and engineering tasks that can be performed in a unified way in both the data and the optical domains. Furthermore, it offers a function framework that can accommodate future expectations concerning the way networks will work and the way services will be provided to clients. Thus we envision a horizontal network, harmonized by a common GMPLS-based control plane, where all network elements work as peers to dynamically establish optical paths through the network.

This new photonic internetwork will make it possible to provision high bandwidth in tenths of seconds, and enable new revenue-generating services and dramatic cost savings for service providers.

In the same way that digital communication technologies changed the twentieth century into the "electronic century," the optical technologies discussed in this article will make the next century "the photonic century." All winning strategies must rely on such GMPLS-based photonic infrastructures—an environment in which innovations work at the speed of light.

For Further Reading

[1] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics,* John Wiley & Sons Inc., 1991.

[2] P. Raghavan and E. Upfal, "Efficient Routing in All-Optical Networks," *Proceedings of ACM STOC'94,* 1994.

[3] B. Mukherjee, *Optical Communication Networks,* McGraw-Hill, 1997.

[4] A. Mokhtar and M. Azizoglu, "Adaptive Wavelength Routing in All-Optical Networks," *IEEE/ACM Transactions on Networking,* vol. 6, pp. 197–206, April 1998.

[5] E. Karasan and S. Banerjee, "Performance of WDM Transport Networks," *IEEE Journal on Selectected Areas in Communications,* vol. 16, pp. 1081–1096, September 1998.

[6] A. Banerjee, J. Drake, J. Lang, B. Turner, K. Kompella, and Y. Rekhter, "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements," *IEEE Communications Magazine,* January 2001.

[7] A. Banerjee, J. Drake, J. Lang, B. Turner, D. O. Awduche, L. Berger, K. Kompella, and Y. Rekhter, "Generalized Multiprotocol Label Switching: An Overview of Signalling Enhancements and Recovery Techniques," *IEEE Communications Magazine,* July 2001.

FRANCESCO PALMIERI holds two Computer Science degrees from Salerno University, Italy. Since 1997, he has led the network management and operation centre of the Federico II University, in Napoli, Italy. He has been closely involved with the development of the Internet in Italy in the last few years, particularly within the academic and research sector, and is actually a member of the Technical Scientific Committee and of the Computer Emergency Response Team of the Italian NREN GARR. He worked for several international companies on a variety of networking-related projects concerned with nationwide communication systems, network management, transport protocols, and IP networking. He is an active researcher in the fields of high-performance, evolutionary networking, and network security. He regularly publishes in leading technical journals and conferences and gives invited talks and keynote speeches. E-Mail: `Francesco.Palmieri@unina.it`

The Changing Foundation of the Internet: Confronting IPv4 Address Exhaustion

by Geoff Huston, APNIC

Throughout its relatively brief history, the Internet has continually challenged our preconceptions about networking and communications architectures. For example, the concepts that the network itself has no role in management of its own resources, and that resource allocation is the result of interaction between competing end-to-end data flows, were certainly novel innovations, and for many they have been very confrontational. The approach of designing a network that is unaware of services and service provisioning and is not attuned to any particular service whatsoever—leaving the role of service support to end-to-end overlays—was again a radical concept in network design. The Internet has never represented the conservative option for this industry, and has managed to define a path that continues to present significant challenges.

From such a perspective it should not be surprising that the next phase of the Internet story—that of the transition of the underlying version of the IP protocol from IPv4 to IPv6—refuses to follow the intended script. Where we are now, in late 2008, with IPv4 unallocated address pool exhaustion looming within the next 18 to 36 months, and IPv6 still largely not deployed in the public Internet, is a situation that was entirely uncontemplated and, even in hindsight, entirely surprising.

The topic examined here is *why* this situation has arisen, and in examining this question we analyze the options available to the Internet to resolve the problem of IPv4 address exhaustion. We examine the timing of the IPv4 address exhaustion and the nature of the intended transition to IPv6. We consider the shortfalls in the implementation of this transition, and identify their underlying causes. And finally, we consider the options available at this stage and identify some likely consequences of such options.

When?

This question was first asked on the TCP/IP list in November 1988, and the responses included foreshadowing a new version of IP with longer addresses and undertaking an exercise to reclaim unused addresses[1]. The exercise of measuring the rate of consumption of IPv4 addresses has been undertaken many times in the past two decades, with estimates of exhaustion ranging from the late 1990s to beyond 2030. One of the earliest exercises in predicting IPv4 address exhaustion was undertaken by Frank Solensky and presented at IETF 18 in August 1990. His findings are reproduced in Figure 1.

At that time the concern was primarily the rate of consumption of Class B network addresses (or of /16 prefixes from the address block `128.0.0.0/2`, to use current terminology). Only 16,384 such Class B network addresses were within the class-based IPv4 address plan, and the rate of consumption was such that the Class B networks would be fully consumed within 4 years, or by 1994. The prediction was strongly influenced by a significant number of international research networks connecting to the Internet in the late 1980s, with the rapid influx of new connections to the Internet creating a surge in demand for Class B networks.

Figure 1: Report on IPv4 Address Depletion[2]
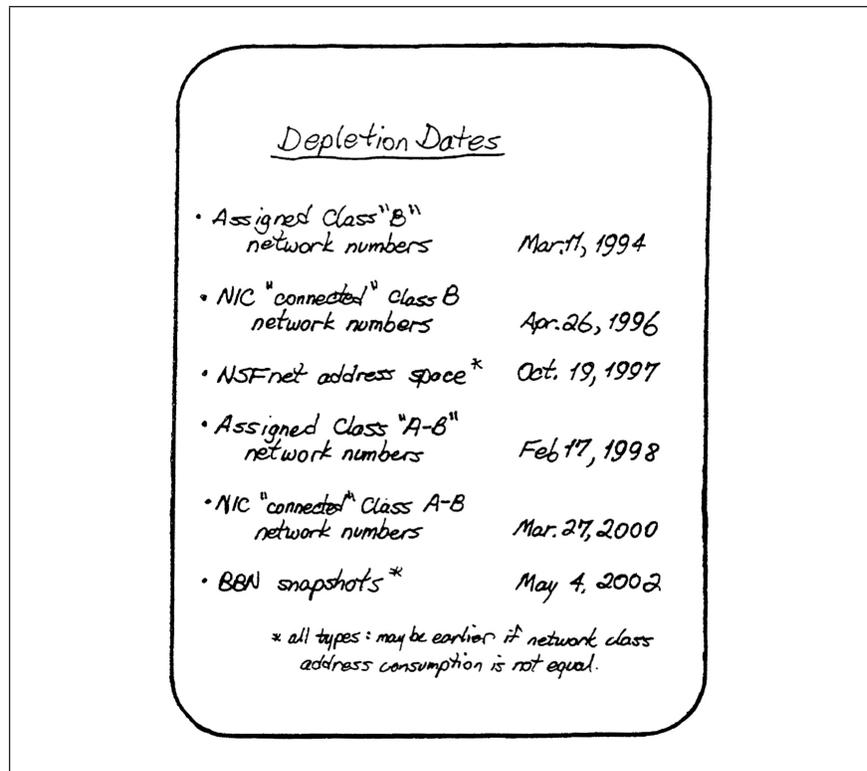
Successive predictions were made in the context of the *Internet Engineering Task Force* (IETF) in the *Address Lifetime Expectancy* (ALE) Working Group, where the predictive model was refined from an exponential growth model to a logistical saturation function, attempting to predict the level at which all address demands would be met.

The predictive technique described here is broadly similar, using a statistical fit of historical data concerning address consumption into a mathematical model, and then using this model to predict future address consumption rates and thereby predict the exhaustion date of the address pool.

The predictive technique models the IP address distribution framework. Within this framework the pool of unallocated /8 address blocks is distributed by the *Internet Assigned Numbers Authority* (IANA) to the five *Regional Internet Registries* (RIRs). (A "/8 address block" refers to a block of addresses where the first 8 bits of the address values are constant. In IPv4 a /8 address block corresponds to 16,777,216 individual addresses.) Within the framework of the prevailing address distribution policies, each RIR can request a further address allocation from IANA when the remaining RIR-managed unallocated address pool falls below a level required to meet the next 9 months of allocation activity. The amount allocated is the number of /8 address blocks required to augment the RIR's local address pool to meet the anticipated needs of the regional registry for the next 18 months. However, in practice, the RIRs currently request a maximum of 2 /8 address blocks in any single transaction, and do so when the RIR-managed address pool falls below a threshold of the equivalent of 2 /8 address blocks.

As of August 2008 some 39 /8 address blocks are left in IANA's unallocated address pool. A predictive exercise has been undertaken using a statistical modeling of historical address consumption rates, using data gathered from the RIRs' records of address allocations and the time series of the total span of address space announced in the Internet interdomain default-free routing table as basic inputs to the model. The predictive technique is based on a least-squares best fit of a linear function applied to the first-order differential of a smoothed copy of the address consumption data series, as applied to the most recent 1,000 days' data.

The linear function, which is a best fit to the first-order differential of the data series, is integrated to provide a quadratic time-series function to match the original data series. The projection model is further modified by analyzing the day-of-year variations from the smoothed data model, averaged across the past 3 years, and applying this daily variation to the projection data to account for the level of seasonal variations in the total address consumption rate that has been observed in the historical data. The anticipated rate of consumption of addresses from this central pool of unallocated IPv4 addresses is expected to be about 15 /8s in 2009, and slightly more in 2010.

RIR behaviors are modeled using the current RIR operational practices and associated address policies, which are used to predict the times when each RIR will be allocated a further 2 /8s from IANA. This RIR consumption model, in turn, allows the IANA address pool to be modeled.

This anticipated rate of increasing address consumption will see the remaining unallocated addresses that are held by IANA reach the point of exhaustion in February 2011. The most active RIRs are anticipated to exhaust their locally managed unallocated address pools in the months following the time of IANA exhaustion.

The assumptions behind this form of prediction follow:

• The current policy framework relating to the distribution of addresses will continue to apply without any further alteration through to complete exhaustion of the unallocated address pool.

• The demand curves will remain consistent, meaning that there will be no forms of disruption to demand, such as a panic rush on the remaining addresses or some introduced externality that affects total address demand.

• The level of return of addresses to the unallocated address pool will not vary significantly from existing levels of address return.

Although the statistical model is based on a complete data set of address allocations and a detailed hourly snapshot of the address span advertised in the Internet routing table, a considerable level of uncertainty is still associated with this prediction.

First, the behavior of the *Internet Service Provider* (ISP) industry and the other entities that are the direct recipients of RIR address allocations and assignments are not ignorant of the impending exhaustion condition, and there is some level of expectation of some form of last-minute rush or panic on the part of such address applicants when exhaustion of this address pool is imminent. The predictive model described here does not include such a last-minute acceleration of demand.

The second factor is the skewed distribution of addresses in this model. From 1 January 2007 until 20 July 2008, 10,402 allocation or assignments transactions were recorded in the RIRs' daily statistics files. These transactions accounted for a total of 324,022,704 individual IPv4 addresses, or the equivalent of 19.3 /8s. Precisely one-half of this address space was allocated or assigned in just 107 such transactions.

In other words, some 1 percent of the recipients of address space in the past 18 months have received some 50 percent of all the allocated address space. The reason why this distribution is relevant here is that this predictive exercise assumes that although individual actions are hard to predict with any certainty, the aggregate outcome of many individuals' actions assumes a much greater level of predictability.

This observation about aggregate behavior does not apply in this situation, however, and the predictive exercise is very sensitive to the individual actions of a very small number of recipients of address space because of this skewed distribution of allocations. Any change in the motivations of these larger-sized actors that results in an acceleration of demand for IPv4 will significantly affect the predictions of the longevity of the remaining unallocated IPv4 address pool.

The third factor is that this model assumes that the policy framework remains unaltered, and that all unallocated addresses are allocated or assigned under the current policy framework, rather than under a policy regime that is substantially different from today's framework. The related assumption here is that the cost of obtaining and holding addresses remains unchanged, and that the perceptions of future scarcity of addresses do not affect the policy framework of address distribution of the remaining unallocated IPv4 addresses.

Given this potential for variation within this set of assumptions, a more accurate summary of the current expectations of address consumption would be that the exhaustion of the IANA unallocated IPv4 address pool will occur sometime between July 2009 and July 2011, and that the first RIR will exhaust all its usable address space within 3 to 12 months from that date, or between October 2009 and July 2012.[3]

What Next?
Apart from the exact date of exhaustion that is predicted by this modeling exercise, none of the information relating to exhaustion of the unallocated IPv4 address pool should be viewed as particularly novel information. The IETF *Routing and Addressing* (ROAD) study of 1991 recognized that the IPv4 address space was always going to be completely consumed at some point in the future of the Internet[4].

Such predictions of the potential for exhaustion of the IPv4 address space were the primary motivation for the adoption of *Classless Inter-Domain Routing* (CIDR) in the *Border Gateway Protocol* (BGP), and the corresponding revision of the address allocation policies to craft a more exact match between planned network size and the allocated address block. These predictions also motivated the protracted design exercise of what was to become the IPv6 protocol across the 1990s within the IETF. The prospect of address scarcity engendered a conservative attitude to address management that, in turn, was a contributory factor in accelerating the widespread use of *Network Address Translation* (NAT)[5] in the Internet during the past decade. By any reasonable metric this industry has had ample time to study this problem, ample time to devise various strategies, and ample time to make plans and execute them.

And this reality has been true for the adoption of classless address allocations, the adoption of CIDR in BGP, and the extremely widespread use of NAT. But all of these measures were short-term, whereas the longer-term measure, that of the transition to IPv6, was what was intended to come after IPv4. But IPv6 has not been the subject of widespread adoption so far, while the time of anticipated exhaustion of IPv4 has been drawing closer. Given almost two decades of advance warning of IPv4 address exhaustion, and a decade since the first stable implementations of IPv6 were released, we could reasonably expect that this industry—and each actor within this industry—is aware of the problem and the need for a stable and scalable long-term solution as represented by IPv6. We could reasonably anticipate that the industry has already planned the actions it will take with respect to IPv6 transition, and is aware of the triggers that will invoke such actions, and approximately when they will occur.

However, such an expectation appears to be ill-founded when considering the broad extent of the actors in this industry, and there is little in the way of a common commitment as to what will happen after IPv4 address exhaustion, nor even any coherent view of plans that industry actors are making in this area.

This lack of planning makes the exercise of predicting the actions within this industry following address exhaustion somewhat challenging, so instead of immediately describing future scenarios, it may be useful to first describe the original plan for the response of the Internet to IPv4 address exhaustion.

What Was Intended?

The original plan, devised in the early 1990s by the IETF to address the IPv4 address shortfall, was the adoption of CIDR as a short-term measure to slow down the consumption of IPv4 addresses by reducing the inefficiency of the address plan, and the longer-term plan of the specification of a new version of the Internet Protocol that would allow for adoption well before the IPv4 address pool was exhausted.

The industry also adopted the use of NAT as an additional measure to increase the efficiency of address use, although the IETF did not strongly support this protocol. For many years the IETF did not undertake the standardization of NAT behaviors, presumably because NAT was not consistent with the IETF's advocacy of end-to-end coherence of the Internet at the IP level of the protocol stack.

Over the 1990s the IETF undertook the exercise of the specification of a successor IP protocol to Version 4, and the IETF's view of the longer-term response was refined to be advocacy of the adoption of the IPv6 protocol and the use of this protocol as the replacement for IPv4 across all parts of the network.

In terms of what has happened in the past 15 years, the adoption of CIDR was extremely effective, and most parts of the network were transitioned to use CIDR within 2 years, with the transition declared to be complete by the IETF in June 1996. And, as noted already, NAT has been adopted across many, if not most, parts of the network. The most common point of deployment of NAT has not been at an internal point of demarcation between provider networks, but at the administrative boundary between the local customer network and the ISP, so that the common configuration of *Customer Premises Equipment* (CPE) includes NAT functions. Customers effectively own and operate NAT devices as a commonplace aspect of today's deployed Internet.

CIDR and NAT have been around for more than a decade now, and the address consumption rates have been held at very conservative levels in that period, particularly so when considering that the bulk of the population of the Internet was added well after the advent of CIDR and NAT.

The longer-term measure—the transition to IPv6—has not proved to be as effective in terms of adoption in the Internet.

There was never going to be a "flag-day" transition where, in a single day, simultaneously across all parts of every network the IP protocol changed to using IPv6 instead of IPv4. The Internet is too decentralized, too large, too disparate, and too critical for such actions to be orchestrated, let alone completed with any chance of success. A flag day, or any such form of coordinated switchover, was never a realistic option for the Internet.

If there was no possibility of a single, coordinated switchover to IPv6, the problem is that there was never going to be an effective piecemeal switchover either. In other words, there was never going to be a switchover where host by host, and network by network, IPv6 is substituted for IPv4 on a piecemeal and essentially uncoordinated basis. The problem here is that IPv6 is not "backward-compatible" with IPv4. When a host uses IPv6 exclusively, then that host has no direct connectivity to any part of the IPv4 network. If an IPv6-only host is connected to an IPv4-only network, then the host is effectively isolated. This situation does not bode well for a piecemeal switchover, where individual components of the network are switched over from IPv4 to IPv6 on a piecemeal basis. Each host that switches over to IPv6 essentially disconnects itself form the IPv4 Internet at that point.

Given this inability to support backward compatibility, what was planned for the transition to IPv6 was a "dual-stack" transition. Rather than switching over from IPv4 to IPv6 in one operation on both hosts and networks, a two-step process has been proposed: first switching from IPv4 only to a "dual-stack" mode of operation that supports both IPv4 and IPv6 simultaneously, and second—and at a much later date—switching from dual-stack IPv4 and IPv6 to IPv6 only.

During the transition more and more hosts are configured with dual stack. The idea is that dual-stack hosts prefer to use IPv6 to communicate with other dual-stack hosts, and revert to use IPv4 only when an IPv6-based end-to-end conversation is not possible. As more and more of the Internet converts to dual stack, it is anticipated that use of IPv4 will decline, until support for IPv4 is no longer necessary. In this dual-stack transition scenario, no single flag day is required and the dual-stack deployment can be undertaken in a piecemeal fashion. There is no requirement to coordinate hosts with networks, and as dual-stack capability is supported in networks the attached dual-stack hosts can use IPv6. This scenario still makes some optimistic assumptions, particularly relating to the achievement of universal deployment of dual stack, at which point no IPv4 functions are used, and support for IPv4 can be terminated. Knowing when this point is reached is unclear, of course, but in principle there is no particular timetable for the duration of the dual-stack phase of operation.

There are always variations, and in this case it is not necessarily that each host must operate in dual-stack mode for such a transition. A variant of the NAT approach can perform a rudimentary form of protocol translation, where a *Protocol-Translating NAT* (or NAT-PT[6]) essentially transforms an incoming IPv4 packet to an outgoing IPv6 packet, and conversely, using algorithmic binding patterns to map between IPv4 and IPv6 addresses. Although this process relieves the IPv6-only host of some additional complexity of operation at the expense of some added complexity in *Domain Name System* (DNS) transformations and service fragility, the essential property still remains that in order to speak to an IPv4-only remote host, the combination of the local IPv6 host and the NAT-PT have to generate an equivalent IPv4 packet. In this case the complexity of the dual stack is now replaced by complexity in a shared state across the IPv6 host and the NAT-PT unit. Of course this solution does not necessarily operate correctly in the context of all potential application interactions, and concerns with the integrity of operation of NAT-PT devices are significant, a factor that motivated the IETF to deprecate the existing NAT-PT specification[7]. On the other hand, the lack of any practical alternatives has led the IETF to subsequently reopen this work, and once again look at specifying the standard behavior of such devices[8].

The detailed progress of a dual-stack transition is somewhat uncertain, because it involves the individual judgment of many actors as to when it may be appropriate to discontinue all support for IPv4 and rely solely on IPv6 for all connectivity requirements. However, one factor is constant in this envisaged transition scenario, and whether it is dual stack in hosts or dual stack through NAT-PT, or various combinations thereof, the requirement that there are sufficient IPv4 addresses to span the addressing needs of the entire Internet across the complete duration of the dual-stack transition process is consistent.

Under this dual-stack regime every new host on the Internet is envisaged to need access to both IPv6 and IPv4 addresses in order to converse with any other host using IPv6 or IPv4. Of course this approach works as long as there is a continuing supply of IPv4 addresses, implying that the envisioned timing of the transition was meant to have been completed by the time that IPv4 address exhaustion happens.

If this transition were to commence in earnest at the present time, in late 2008, and take an optimistic 5 years to complete, then at the current address consumption rate we will require a further 90 to 100 /8 address blocks to span this 5-year period. A more conservative estimate of a 10-year transition will require a further 200 to 250 /8 address blocks, or the entire IPv4 address space again, assuming that we will use IPv4 addresses in the future in precisely the same manner as we have used them in the past and with precisely the same level of usage efficiency as we have managed to date.

Clearly, waiting for the time of IPv4 unallocated address pool exhaustion to act as the signal to industry to commence the deployment of IPv6 in a dual-stack transition framework is a totally flawed implementation of the original dual-stack transition plan.

Either the entire process of dual-stack transition will need to be undertaken across a far faster time span than has been envisaged, or the manner of use of IPv4 addresses, and, in particular their usage efficiency in the context of dual-stack transition support, will need to differ markedly from the current manner of address use. Numerous forms of response may be required, posing some challenging questions because there is no agreed precise picture of what markedly different and significantly more efficient form of address use is required here. To paraphrase the situation, it is clear that we need to do "something" differently, and do so as a matter of some urgency, but we have no clear agreement on what that something is that we should be doing differently. This situation obviously is not an optimal one.

What was intended as a transition mechanism for IPv6 is still the only feasible approach that we are aware of, but the forthcoming exhaustion of the unallocated IPv4 address pool now calls for novel forms of use of IPv4 addresses within this transitional framework, and these novel forms may well entail the deployment of various forms of address translation technologies that we have not yet defined, let alone standardized. The transition may also call for scaling capabilities from the interdomain routing system that also head into unknown areas of technology and deployment feasibility.

Why?

At this point it may be useful to consider how and why this situation has arisen.

If the industry needed an abundant supply of IPv4 addresses to underpin the entire duration of the dual-stack transition to IPv6, then why didn't the industry follow the lead of the IETF and commence this transition while there was still an abundant supply of IPv4 addresses on hand? If network operators, service providers, equipment vendors, component suppliers, application developers, and every other part of the Internet supply chain were aware of the need to commence a transition to IPv6 well before effective exhaustion of the remaining pool of IPv4 addresses, then why didn't the industry make a move earlier? Why was the only clear signal for a change in Internet operation to commence a dual-stack transition to IPv6 one that has been activated too late to be useful for the industry to act on efficiently?

One possible reason may lie in a perception of the technical immaturity of IPv6 as compared to IPv4. It is certainly the case that many network operators in the Internet are highly risk-adverse and tend to operate their networks in a mainstream path of technologies rather than constantly using leading-edge advance releases of hardware and software solutions. Does IPv6 represent some form of unacceptable technical risk of failure that has prevented its adoption? This reasoning does not appear to be valid in terms of either observed testing or observation of perceptions about the technical capability of IPv6. The IPv6 protocol is functionally complete and internally consistent, and it can be used in almost all contexts where IPv4 is used today. IPv6 works as a platform for all forms of transport protocols, and is fully functional as an internetwork layer protocol that is functionally equivalent to IPv4. IPv6 NAT exists, *Dynamic Host Configuration Protocol Version 6* (DHCPv6) provides dynamic host configuration for IPv6 notes, and the DNS can be completely equipped with IPv6 resource records and operate using IPv6 transport for queries and responses.

Perhaps the only notable difference between the two protocols is the ability to perform host scans in IPv6, where probe packets are sent to successive addresses. In IPv6 the address density is extremely low because the low-order 64-bit interface address of each host is more or less unique, and within a single network the various interface addresses are not clustered sequentially in the number space. The only known use of address probing to date has been in various forms of hostile attack tools, so the lack of such a capability in IPv6 is generally seen as a feature rather than an impediment. IPv6 deployment has been undertaken in a small scale for many years, and although the size of the deployed IPv6 base remains small, the level of experience gained with the technology functions has been significant. It is possible to draw the conclusion that IPv6 is technically capable and this capability has been broadly tested in almost every scenario except that of universal use across the Internet.

It also does not appear that the reason was a lack of information or awareness of IPv6. The efforts to promote IPv6 adoption have been under way in earnest for almost a decade now. All regions and many of the larger economies have instigated programs to promote the adoption of IPv6 and have provided information to local industry actors of the need to commence a dual-stack transition to IPv6 as soon as possible. In many cases these promotional programs have enjoyed broad support from both public and industry funding sources. The coverage of these promotional efforts has been widespread in industry press reports. Indeed, perhaps the only criticism of this effort is possibly too much promotion, with a possible result that the effectiveness of the message has been diluted through constant repetition.

A more likely area to examine in terms of possible reasons why industry has not engaged in dual-stack transition deployment is that of the business landscape of the Internet. The Internet can be viewed as a product of the wave of progressive deregulation in the telecommunications sector in the 1980s and early 1990s. New players in the deregulated industry searching for a competitive edge to unseat the dominant position of the traditional incumbents found the Internet as their competitive lever. The result was perhaps unexpected, because it was not one that replaced one vertically integrated operator with a collection of similarly structured operators whose primary means of competition was in terms of price efficiency across an otherwise undifferentiated service market, as we saw in the mobile telephony industry. In the case of the Internet, the result was not one that attempted to impose convergence on this industry, but one that stressed divergence at all levels, accompanied by branching role specialization at every level in the protocol stack and at every point in the supply chain process. In the framework of the Internet, consumers are exposed to all parts of the supply process, and do not rely on an integrator to package and supply a single, all-embracing solution. Consumers make independent purchases of their platform technology, their software, their applications, their access provider, and their means of advertising their own capabilities to provide goods and services to others, all as independent decisions, all as a result of this direct exposure to the consumer of every element in the supply chain.

What we have today is an industry structure that is highly diverse, broadly distributed, strongly competitive, and intensely focused on meeting specific customer needs in a price-sensitive market, operating on a quarter-by-quarter basis. Bundling and vertical integration of services has been placed under intense competitive pressure, and each part of the network has been exposed to specialized competition in its right. For consumers this situation has generated significant benefits. For the same benchmark price of around US$15 to US$30 per month, or its effective equivalent in purchasing power of a local currency, today's Internet user enjoys multimegabit-per-second access to a richly populated world of goods and services.

The price of this industry restructure has been a certain loss of breadth and depth of the supply side of the market. If consumers do not value a service, or even a particular element of a service, then there is no benefit in incurring marginal additional cost in providing the service. In other words, if the need for a service is not immediate, then it is not provided. For all service providers right through the supply side the focus is on current customer needs, and this focus on current needs, as distinct from continued support of old products or anticipatory support of possible new products, excludes all other considerations.

Why is this change in the form of communications industry operation an important factor in the adoption of IPv6? The relevant question in this context is that of placing IPv6 deployment and dual-stack transition into a viable business model. IPv6 was never intended to be a technology visible to the end user. It offers no additional functions to the end user, nor any direct cost savings to the customer or the supplier. Current customers of ISPs do not need IPv6 today, and neither current nor future customers are aware that they may need it tomorrow. For end users of Internet services, e-mail is e-mail and Web-based delivery of services is just the Web. Nothing will change that perspective in an IPv6 world, so in that respect customers do not have a particular requirement for IPv6, as opposed to a generic requirement for IP access, and will not value such an IPv6-based access service today in addition to an existing IPv4 service. For an existing customer IPv6 and dual stack simply offer no visible value. So if the existing customer base places no value on the deployment of IPv6 and dual stack, then the industry has little incentive to commit to the expenditure to provide it.

Any IPv6 deployment across an existing network is essentially an unfunded expenditure exercise that erodes the revenue margins of the existing IPv4-based product. And as long as sufficient IPv4 address space remains to cover the immediate future needs, looking at this situation on the basis of a quarter-by-quarter business cycle, then the decision to commit to additional expenditure and lower product margins to meet the needs of future customers using IPv6 and dual-stack deployments is a decision that can comfortably be deferred for another quarter. This business structure of today's Internet appears to represent the major reason why the industry has been incapable of making moves on dual-stack transition within a reasonable timeframe as it relates to the timeframe of IPv4 address pool exhaustion.

What of the strident calls for IPv6 deployment? Surely there is substance to the arguments to deploy IPv6 as a contingency plan for the established service providers in the face of impending IPv4 address exhaustion, and if that is the case, why have service providers discounted the value of such contingency motivations? The problem to date is that IPv4 address exhaustion is now not a novel message, and, so far, NAT usage has neutralized the urgency of the message.

The NAT protocol is well-understood, it appears to work reliably, applications work with it, and it has influenced the application environment to such an extent that now no popular application can be fielded unless is can operate across this protocol. For conventional client-server applications, NAT represents no particular problem. For peer-to-peer–based applications, the rendezvous problem with NAT has been addressed through application gateways and rendezvous servers. Even the variability of NAT behavior is not a service provider liability, and it is left to applications to load additional functions to detect specific NAT behavior and make appropriate adjustments to the behavior of the application.

The conventional industry understanding to date is that NAT can work acceptably well within the application and service environment. In addition, NAT usage for an ISP represents an externalized cost, because it is essentially funded and operated by the customer and not the ISP. The service provider's perspective is that considering that this protocol has been so effective in externalizing the costs of IPv4 address scarcity from the ISP for the past 5 years, surely it will continue to be effective for the next quarter. To date the costs of IPv4 address scarcity have been passed to the customer in the form of NAT-equipped CPE devices and to the application in the form of higher complexity in certain forms of application rendezvous. ISPs have not had to absorb these costs into their own costs of operation. From this perspective, IPv6 does not offer any marginal benefits to ISPs. For an ISP today, NATs are purchased and operated by customers as part of their CPE equipment. To say that IPv6 will eliminate NATs and reduce the complexities and vulnerabilities in the NAT service model is not directly relevant to the ISP.

The more general observation is that, for the service provider industry currently, IPv6 has all the negative properties of revenue margin erosion with no immediate positive benefits. This observation lies at the heart of why the service provider industry has been so resistant to the call for widespread deployment of IPv6 services to date.

It appears that the current situation is not the outcome of a lack of information about IPv6, nor a lack of information about the forthcoming exhaustion of the IPv4 unallocated address pool. Nor is it the outcome of concerns over technical shortfalls or uncertainties in IPv6, because there is no evidence of any such technical shortcomings in IPv6 that prevent its deployment in any meaningful fashion. A more likely explanation for the current situation is an inability of a highly competitive deregulated industry to be in a position to factor longer-term requirements into short-term business logistics.

What Next?

Now we consider some questions relating to IPv4 address exhaustion. Will the exhaustion of the current framework that supplies IP addresses to service providers cause all further demand for addresses to cease at that point?

Or will exhaustion increase the demand for addresses in response to various forms of panic and hoarding behaviors in addition to continued demand from growth?

The size and value of the installed base of the Internet using IPv4 is now very much larger than the size and value of incremental growth of the network. In address terms the routed Internet currently (as of 14 August 2008) spans 1,893,725,831 IPv4 addresses, or the equivalent of 112.2 /8 address blocks. Some 12 months ago the routed Internet spanned 1,741,837,080 IPv4 addresses, or the equivalent of 103.8 /8 address blocks, representing a net annual growth of 10 percent in terms of advertised address space.

These facts lead to the observation that, even in the hypothetical scenario where all further growth of the Internet is forced to use IPv6 exclusively while the installed base still uses IPv4, it is highly unlikely that the core value of the Internet will shift away from its predominate IPv4 installed base in the short term.

Moving away from the hypothetical scenario, the implication is that the relative size and value of new Internet deployments will be such that these new deployments may not have sufficient critical mass by virtue of their volume and value as to be in a position to force the installed base to underwrite the incremental cost to deploy IPv6 and convert the existing network assets to dual-stack operation in this timeframe. The corollary of this observation is that new Internet network deployments will need to communicate with a significantly larger and valuable IPv4-only network, at least initially. The fact that IPv6 is not backward-compatible with IPv4 further implies that hosts in these new deployments will need to cause IPv4 packets with public addresses in their packet headers to be sent and received, either by direct deployment of dual stack or by proxies in the form of protocol-translating NATs. In either case the new network will require some form of access to public IPv4 addresses. In other words, after exhaustion of the unallocated address pools, new network deployments will continue to need to use IPv4 addresses.

From this observation it appears highly likely that the demand for IPv4 addresses will continue at rates comparable to current rates across the IPv4 unallocated address pool and after it is exhausted. The exhaustion of the current framework of supply of IPv4 addresses will not trigger an abrupt cessation of demand for IPv4 addresses, and this event will not cause the deployment of IPv6-only networks, at least in the short term of the initial years following IPv4 address pool exhaustion. It is therefore possible to indicate that immediately following this exhaustion event there will be a continuing market need for IPv4 addresses for deployment in new networks.

Although a conventional view is that this market need is likely to occur in a scenario of dual-stacked environments, where the hosts are configured with both IPv4 and IPv6, and the networks are configured to also support the host operation of both protocols, it is also conceivable to envisage the use of deployments where hosts are configured in an IPv6-only mode and network equipment undertakes a protocol-translating NAT function. In either case the common observation is that we apparently will have a continuing need for IPv4 addresses well after the event of IPv4 unallocated pool exhaustion, and IPv6 alone is no longer a sufficient response to this problem.

How?

If demand continues, then what is the source of supply in an environment where the current supply channel, namely the unallocated pool of addresses, is exhausted? The options for the supply of such IPv4 addresses are limited.

In the case of established network operators, some IPv4 addresses may be recovered through the more intensive use of NAT in existing networks. A typical scenario of current deployment for ISPs involves the use of private address space in the customer's network and NAT performed at the interface between the customer network and the service provider infrastructure (the CPE). One option for increasing the IPv4 address usage efficiency could involve the use of a second level of NAT within the service provider's network, or the so-called "carrier-grade" NAT option[9]. This option has some attraction in terms of increasing the port density use of public IPv4 addresses, by effectively sharing the port address space of the public IPv4 address across multiple CPE NAT devices, allowing the same number of public IPv4 addresses to be used across a larger number of end-customer networks.

The potential drawback of this approach is that of added complexity in NAT behavior for applications, given that an application may have to traverse multiple NATs, and the behavior of the compound NAT scenario becomes in effect the behavior of the most conservative of the NATs in the path in terms of binding times and access. Another potential drawback is that some applications have started to use multiple simultaneous transport sessions in order to improve the performance of the download of multipart objects. For single-level CPE NATs with more than 60,000 ports to be used for the customer network, this application behavior had little effect, but the presence of a carrier NAT servicing a large number of CPE NATs may well restrict the number of available ports per connection, in turn affecting the utility of various forms of applications that operate in this highly parallel mode. Allowing for a peak simultaneous demand level of 500 ports per customer provides a potential use factor of some 100 customers per IP address.

Given a large enough common address pool, this factor may be further improved by statistical multiplexing by a factor of 2 or 3, allowing for between 200 and 300 customers per NAT address. Of course such approximations are very coarse, and the engineering requirement to achieve such a high level of NAT usage would be significant. Variations on this engineering approach are possible in terms of the internal engineering of the ISP network and the control interface between the CPE NATs and the ISP equipment, but the maximal ratio of 200 to 300 customers per public IP address appears to be a reasonable upper bound without unduly affecting application behaviors.

Another option is based on the observation that, of the currently allocated addresses, some 42 percent of them, or the equivalent of some 49 /8 address blocks, are not advertised in the interdomain routing table, and are presumed to be either used in purely private contexts, or currently unused. This pool of addresses could also be used as a supply stream for future address requirements, and although it may be overly optimistic to assume that the entirety of this unadvertised address space could be used in the public Internet, it is possible to speculate that a significant amount of this address pool could be used in such a manner, given the appropriate incentives. Speculating even further, if this address pool were used in the context of intensive carrier-grade NATs with an achieved average deployment level of, say, 10 customers per address, an address pool of 40 /8s would be capable of sustaining some 7 billion customer attachments.

Of course, no such recovery option exists for new entrants, and in the absence of any other supply option, this situation will act as an effective barrier to entry into the ISP market. In cases where the barriers to entry effectively shut out new entrants, there is a strong trend for the incumbents to form cartels or monopolies and extract monopoly rentals from their clients. However, it is unlikely that the lack of supply will be absolute, and a more likely scenario is that addresses will change hands in exchange for money. Or, in other words, it is likely that such a situation will encourage the emergence of markets in addresses. Existing holders of addresses have the option to monetize all or part of their held assets, and new entrants, and others, have the option to bid against each other for the right to use these addresses. In such an open market, the most efficient usage application would tend to be able to offer the highest bid, in an environment dominated by scarcity tending to provide strong incentives for deployment scenarios that offer high levels of address usage efficiency.

It would therefore appear that options are available to this industry to increase the usage efficiency of deployed address space, and thereby generate pools of available addresses for new network deployments. However, the motive for so doing will probably not be phrased in terms of altruism or alignment to some perception of the common good. Such motives sit uncomfortably within the commercial world of the deregulated communications sector.

Nor will it be phrased in terms of regulatory impositions. It will take many years to halt and reverse the ponderous process of public policy and its expression in terms of regulatory measures, and the "common-good" objective here transcends the borders of regulatory regimes. This consideration tends to leave this argument with one remaining mechanism that will motivate the industry to significantly increase the address usage efficiency: monetizing addresses and exposing the costs of scarcity of addresses to the address users. The corollary of this approach is the use of markets to perform the address distribution function, creating a natural pricing function based on levels of address supply and demand.

References

[1] TCP/IP Mailing List, Message Thread: "Running out of Internet Addresses," November 1988.
`http://www-mice.cs.ucl.ac.uk/multimedia/misc/tcp_ip/8813.mm.www/index.html#121`

[2] F. Solenksy, "Internet Growth," Steering Group Report, p. 61, Proceedings of the 18th IETF Meeting, August 1990.
`http://www.ietf.org/proceedings/prior29/IETF18.pdf`

[3] G. Huston, "The IPv4 Internet Report," August 2008,
`http://ipv4.potaroo.net`

[4] P. Gross and P. Almquist, "IESG Deliberations on Routing and Addressing," RFC 1380, November 1992.

[5] K. Egevang and P. Francis, "The IP Network Address Translator (NAT)," RFC 1631, May 1994.

[6] G. Tsirtsis and P. Srisuresh, "Network Address Translation – Protocol Translation (NAT-PT)," RFC 2766, February 2000.

[7] C. Aoun and E. Davies, "Reasons to Move the Network Address Translator – Protocol Translator (NAT-PT) to Historic Status," RFC 4966, July 2007.

[8] M. Bagnulo, P. Matthews, and I. van Beijnum, "NAT64/DNS64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers," Internet Draft, work in progress, `draft-bagnulo-behave-nat64-00.txt`, June 2008.

[9] T. Nishitani and S. Miyakawa, "Carrier Grade Network Address Translator (NAT) Behavioral Requirements for Unicast UDP, TCP and ICMP," Internet Draft, work in progress, `draft-nishitani-cgn-00.txt`, July 2008.

[10] Olaf Maennel, Randy Bush, Luca Cittadini, Steven M. Bellovin, "A Better Approach than Carrier-Grade-NAT," `http://rip.psg.com/~randy/080820.alt-to-cgn.pdf`

[11] William Lehr, Tom Vest, Eliot Lear, "Running on Empty: The Challenge of Managing Internet Addresses," to be presented at the 36th Research Conference on Communication, Information and Internet Policy (TPRC), on 27 September 2008. `http://eyeconomics.com/backstage/References_files/Lehr-Vest-Lear-TPRC2008-080915.pdf`

[12] Hain, Tony, "A Pragmatic Report on IPv4 Address Space Consumption," *The Internet Protocol Journal,* Volume 8, No. 3, September 2005

[13] `http://icann.org/en/announcements/proposal-ipv4-report-29nov07.htm`
(See also "Fragments" on page 46.)

GEOFF HUSTON is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He graduated from the Australian National University with a B.Sc. and M.Sc. in Computer Science. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of numerous Internet-related books, and was a member of the Internet Architecture Board from 1999 until 2005; he served on the Board of Trustees of the Internet Society from 1992 to 2001.
E-mail: `gih@apnic.net`

Letters to the Editor

I sincerely congratulate you for Geoff Huston's excellent article in *The Internet Protocol Journal,* June 2008, on the "Decade of Internet Evolution." The article shows an amazing insight into the Internet as it has recently evolved and deserves as wide an audience as possible.

The only comment I could make is that though Huston hints about separating the IP address from the host name, he does not explicitly mention the *Host Identity Protocol* (HIP)[1]. Previous issues of the Journal have this omission as well.

Note: As we struggle in the IETF and everywhere else in the industry with NAT traversal, mobility, and multihoming, we see countless approaches for each application layer protocol separately. HIP seems to fulfill the promise of solving these problems comprehensively.

Thanks for the privilege to continue reading the Journal; keep such papers coming.

—Henry Sinnreich, Adobe Systems, Inc.
hsinnrei@adobe.com

[1] R. Moskowitz, P. Nikander, P. Jokela, Ed., and T. Henderson, "Host Identity Protocol," RFC 5201, April 2008. See also: **http://www.ietf.org/html.charters/hip-charter.html**

The author responds:

Thank you for your generous comments.

At some point I was toying (dangerously!) with writing an article that attempted to predict the next 10 years, looking at what appears to be important today and what that could mean in the future. There is no doubt that the tight binding of identity and location is one of the assumptions that has made the Internet both simple and effective for the past decade. But where we sit today, in a world dominated by scale, mobility, a dense mesh of interconnectivity, highly capable end devices, dense middleware, and a panoply of specialized requirements, we need to look forward to methods that allow separation of identity and location. Now this separation could be at the level of the Internet Protocol itself, as in HIP or *Site Multihoming by IPv6 Intermediation* (SHIM6); or at the level of the transport session, as exemplified at present by the *Stream Control Transmission Protocol* (SCTP); or even at the application level, where the various offerings related to *Voice over IP* (VoIP) and *Peer-to-Peer* (P2P) have been working at the level of multiparty application rendezvous and application identity that sit on top of an adaptive platform of dynamic discovery of the characteristics of the underlying transport subsystem.

Each approach appears to offer some significant leverage in scaling the network in diverse ways, while at the same time presenting us with some fascinating insights into possible architectures that could address our needs in the next decade. No doubt the next 10 years will present us with some quite novel challenges with the imminent exhaustion of the unallocated IPv4 address pool and the associated observation that the schedule for the update of IPv6 has proceeded so slowly that we will be forced to be remarkably inventive with IPv4. HIP may well be a central part of such invention, but, more generally, I have no doubt that we will examine more generally how we can devise refinements to the networking model that preserve useful notions of identity across a rather fluid sea of shared location tokens.

Regards,

—*Geoff Huston, APNIC*
`gih@apnic.net`

Ten Years of IPJ

We received many congratulatory messages in response to our June 2008 Anniversary Issue. The following are some quotes from our readers:

"Compliments and congratulations for the tenth anniversary of this great Journal. It is great because it is making us realize the synergy between what has been and what is to come."

—*John Okewole, Lagos, Nigeria*

"This week I received the June 2008 issue of IPJ. I have been a subscriber for several years and it has been a great pleasure to find great contents in IPJ, such as the current issue that brings reviews on Internet evolution. I would like to send my congratulations to the IPJ team for 10 years of publication and my best wishes for future success."

—*Frederico Fari, Belo Horizonte, Brazil*

"I think that IPJ is a great journal. I hope you will not be forced to give up the paper edition because is a beautiful one (and it allows me to read during the evening hours when all computers and children in the house are shut down :–)"

—*Andrea Montefusco, Rome, Italy*

Book Reviews

Code and Other Laws of Cyberspace, by Lawrence Lessig, Basic Books, 1999, ISBN 0-465-03913-8. `http://code-is-law.org/`

Code 2.0, by Lawrence Lessig, Basic Books, 2006, ISBN-10: 0-465-03914-6, ISBN 13: 978-0-465-03914-2. `http://codev2.cc/`

First published in 1999, then Harvard Law School Professor Lawrence Lessig's cautionary tale about the inescapable influence of certain material features of the built Internet has since become a foundational "Internet studies" text in universities and laws schools around the world. Lessig, who now occupies an endowed chair at Stanford Law School, makes a series of troubling observations about the Internet, his chosen sector of focus since setting aside his mid-1990s work on legal and institutional development in post-Soviet societies.

Lessig's key findings from that previous work are that rules matter—especially the sort of rules embodied in "constitutions" and other foundational institutions; that rules are artifacts of contingent human intent and design; and that rules can be changed. Being a "classical liberal" on the model of John Stuart Mill, Lessig advocates the sort of rules that afford maximum liberty for individuals against a triumvirate of coercive influences, including not only governments but also market power and oppressive social mores.

Now however, a fourth challenge to personal liberty has been exposed by the advent of the Internet—or rather, of *cyberspace,* which Lessig describes as the lived experience of participants in the rich application space that has been built atop the Internet. This new constraining factor is "architecture," which Lessig defines as "the built environment," or "the way the world is," that is, the cumulative result of all of the contingent historical events and decisions that have shaped the material circumstances confronting Internet users (or *cyberspace denizens*) today. *Code* is Lessig's term for the instruction sets (that is, programs, applications, etc.) that are the building blocks of the architecture of cyberspace; it is the stuff that emerges from the decision making of a relatively few (the *code writers*), which accretes over time into the less-malleable architecture that shapes the everyday choices and possibilities of everyone else whom the Internet or cyberspace touches.

New Code Means New Power(s)

According to Lessig, the code that defines cyberspace—which he calls "West Coast Code"—demands particular attention, both because of its omnipresence and because of how it differs from the other, more familiar factors that can impinge on individual liberty.

Like the canons of law (also known as "East Coast Code"), code is basically a collection of rules written with human goals and objectives in mind. However, in its effects code more closely resembles the laws of nature, because it requires neither the awareness nor the consent of its subjects in order to be effective. Although this claim sounds suspiciously like a variant, or perhaps an illustration of Arthur C. Clarke's *Third Law of Prediction* (which states that any sufficiently advanced technology will be indistinguishable from the supernatural), there is purpose behind Lessig's observation. The self-enforcing character of code is doubly problematic in the case of cyberspace, he suggests, because unlike the law, code affords no appeal, no recourse, and no formal, institutional review and interpretation of the kind that lawyers and judges exercise in legal matters. Without such expert oversight, code might come to be used as a tool to subvert individual liberties or public values, for either commercial or political gain, without anyone's being the wiser. In fact, he implies, the lack of transparency of code almost invites such abuses.

At this point some might be tempted to dismiss Lessig's program as just "sour grapes" from a high-profile industry spokesman sensing this erosion of the traditional prominence and centrality of his profession in a new code-centric world. Lessig believes passionately in the exercise of law and judicial review as master tools for keeping other important forces—government power, market power, and social norms—broadly aligned with "important public values." He extols the relationships among the rule of law, democracy, and politics, the latter of which invests law with legitimacy to raise or lower the cost of particular individual actions (for example, by taxing, criminalizing, valorizing, or subsidizing them) to encourage conformity with publicly chosen goals and values. He observes that "architecture is a kind of law" and that "code codifies values, and yet, oddly, most people speak as if code were just a question of engineering." It takes no great leap of imagination to conclude that code too should be subject to the same kind of legal and judicial oversight that keeps the rest of society running smoothly. Eliminating any doubt, Lessig asserts that:

> Technology is plastic. It can be remade to do things differently. We should expect—and demand—that it can be made to reflect any set of values that we think important. The burden should be on the technologists to show us why that demand can't be met.

However, such a dismissal would indeed be too easy, for Lessig also expresses misgivings about the professionalization and segregation of "constitutional thinking" within the legal sector. "Constitutional thought has been the domain of lawyers and judges for too long," Lessig writes, and as a result everyone else has grown less comfortable—and also less competent—in engaging in fruitful conversation about fundamental, "constitutional" values.

And yet Lessig suggests that this skill has also atrophied within the legal community, as more and more jurists have embraced an "originalist" interpretive philosophy that holds that the U.S. Constitution provides no guidance for how to resolve conflicts between old values—what Lessig calls *latent ambiguities*—or how to address wholly novel concerns raised by technologies such as the Internet. Originalists (Lessig mentions U.S. Supreme Court Justice Antonin Scalia) assert that in such cases the only recourse is the political and legislative processes—where, one assumes, limited experience with both technology and constitutional debate make the prospects for success even dimmer. Lessig writes that "We (legal scholars) have been trapped by a mode of reasoning that pretends that all the important questions have already been answered," but that "the constitutional discourse of our present Congress is far below the level at which it must be to address the questions about constitutional values that will be raised by cyberspace."

Diagnosis from a Distance

Lessig is without question eminently qualified to make such observations about his home-turf legal and political spheres. However, it is less clear that his blanket charge of deliberative incompetence is equally valid across the full range of Internet and cyberspace stakeholders. Neither is it clear that the architecture of cyberspace is as uniquely problematic as he suggests, compared to the architecture of other, more familiar domains. Finally, Lessig's own admittedly limited technical expertise may lead him to misapprehend the boundary between cyberspace and the Internet, and to underestimate the radicalness of his proposed cyberspace fix.

Taking these ideas in reverse order, Lessig's conception of the structural and functional distinction between the Internet and cyberspace merits closer scrutiny. As explained later, Lessig advocates profound technical changes to bring the functions of code under the rule of law (or laws, because Lessig wishes to accommodate subsidiary jurisdictions as well as sovereign differences in law). However, he envisions this intervention affecting only the "code" domain, not the "Internet's core protocols":

> When I speak about regulating the code, I'm not talking about changing these core TCP/IP protocols...In my view these components of the network are fixed. If you required them to be different, you'd break the Internet. Thus rather than imagining the government changing the core, the question I want to consider is how the government might either (1) complement the core with technology that adds regulability, or (2) regulate applications that connect to the core.

Lessig's specific ideas for achieving this function while preserving the core are not fully detailed in this context until *Code 2.0* (2006), which Lessig describes as an update rather than a full rewrite, albeit one with new relevance to match a "radically different time." The central idea involves the introduction of an "identity layer" that permits authoritative in-band querying and signaling of the jurisdiction(s) to which every would-be Internet user is subject. The deployment of this system would be accompanied by the development of a comprehensive distributed database of Internet usage restrictions mandated by every legally recognized jurisdiction around the world. Together, these components would operate as a kind of "domain interdiction system" that would automatically black-hole all Internet resource queries that are legally impermissible to individuals based on their jurisdiction(s) of origin, regardless of their actual location.

This proposal is clearly vulnerable to criticism of many kinds—technical, ethical, practical, etc.—and to be fair Lessig anticipates and preemptively responds to several of the most obvious ones. Space limitations preclude any review of those arguments here, but it is impossible to resist a few short observations. First, it is not clear why Lessig imagines that his proposed system would be anything less than a fundamental intervention in the core function and protocols of the Internet. Today several different high-profile technical developments that could plausibly be described as changing TCP/IP are under way, but they (hopefully) will not break the Internet. At the same time, TCP/IP is not the only technology that is essential to the Internet "core." The system that Lessig advocates is clearly inspired by the *Domain Name System* (DNS), it would of necessity be similarly global and ubiquitous in scope and scale, and it would likely function by selectively blocking some DNS responses based on the initiator's identity. Although some once regarded the DNS as a mere application (for example, shortly after it was invented), few today would categorize it as anything other than a core protocol. Also, given the degree to which any implementation of the proposed identity system would preempt many "normative" features that are associated with the Internet core (for example, the principles behind the *end-to-end* arguments), it is unclear what would remain "unbroken" therein that might still warrant any special consideration or separate treatment. We can only hope that Lessig's optimism on this question is justified, because looming developments in certain wireless standards as well as in the management of IP addressing may provide for more concrete—and less revisable—answers in the very near future.

Objects in View May Be Closer Than They Appear

Then there is the question of how much code really makes the architecture of cyberspace different from the architecture of other domains. Many of Lessig's claims on this point date back to the first version of the book, when Internet exceptionalism was still new enough for deflationary counterarguments to seem provocative.

Although the revolutionary potential of the Internet continues to inspire many (this reviewer included), the past decade of booms, busts, compromises, and indictments have done much to temper that faith. It is not that Lessig's concerns about the opaque nature of cyberspace architecture, about the substantial influence that code writers and network owners command, and about the vulnerability of the whole system to a crisis-induced authoritarian turn aren't reasonably well-founded. But they are equally apropos to most other important spheres of life. The phrase "possession is nine-tenths of the law" has multiple meanings, and was coined many decades before the Internet was invented. The inexplicability of many current "real-world" legislative and judicial outcomes without recourse to some cynical theory of unacknowledged interests and unobservable influence certainly raises many questions about the architecture of the space beyond cyberspace. And Lessig's warnings about national security fears precipitating a sudden loss of freedoms (taken from Jonathan Zittrain's *Z-Theory*) now seem prophetic—albeit less for the Internet than for the earliest and largest host society of the Internet. One might observe that Lessig is guilty of his own kind of exceptionalism—one that, ironically, may obscure the degree to which constitutional challenges in the real and virtual worlds are more or less the same. In fact, Lessig's subsequent shift of priorities from code to intellectual property law recently ended with a return to his original home turf of law and politics—perhaps in belated recognition that sometimes, even when you have a good story, East Coast Code is still the only durable recourse.

Finally, there is the question of constitutional acumen. This question is the critical one for Lessig (he uses some form of the term *constitution* more than 250 times in the main text), because for him the term evokes nothing less than "an architecture... a way of life that structures and constrains social and legal power, to the end of protecting fundamental values." In this sense, he adds, constitutions are built rather than found. Moreover, they have been built in different (albeit sometimes overlapping) places by different institutions and societies, many with quite different conceptions of which fundamental values to uphold. From whence will the architecture of values of cyberspace emerge? Who will be its authors? Lessig never quite gives a final answer, even for his own home jurisdiction, but he does help to winnow out several likely suspects. As noted previously, he invests little faith in the current U.S. legislative branch. He also has reservations about many members of his own legal profession, although the need to preserve backward compatibility with the primary U.S. Constitution and to reconcile newly revealed "latent ambiguities" therein obviously recommends some legal training at the very least. Government and industry represent the most likely perpetrators of liberty-undermining code, Lessig claims, so he looks for no help from those quarters.

In the end Lessig provides some oblique advice for judges (abandon formalism), hackers (open source), and voters (educate yourself, and don't give up hope), but ultimately concludes with a call for more lawyerly deliberation: if only our leaders could act more like lawyers, telling stories that persuade "not by hiding the truth or exciting the emotion, but by using reason," and our fellow citizens could act like juries, resisting the fleeting passions of the mob and making decisions based on the facts alone, then perhaps we could overcome the architectural challenges of both cyberspace and physical space.

Story Boards and Internet Constitutions

Notwithstanding its solipsistic aspects, advice like that discussed in the last section is hard to find fault with. Professor Lessig is unquestionably a person of good conscience, and has a long, distinguished, and very well-documented record of putting this advice into practice in a wide range of good causes, including many that are wholly unrelated to code or cyberspace. However, one could argue (perhaps with equal solipsism) that many of the behaviors and virtues that he commends are now regularly on display in the mailing lists, message boards, and other deliberative records of the Regional Internet Registries, the IETF, and the IAB—in particular in discussions on the form that IPv4 and IPv6 address-allocation policies should take, in the design of future routing systems that balance scalability with the freedom to choose between competing providers, and in the reconciliation of traditional policies and their beneficiaries with the changing realities of Internet resource stewardship. Closer scrutiny of these records reveals that successful consensus policies are almost invariably borne of good, well-reasoned stories, the vast majority of which are offered by individuals who are affiliated neither with government agencies nor with any of the largest and most powerful ISPs. Many of the storytellers are old hands, but new voices regularly emerge and command attention based on nothing more than the strength of their reasoning. Participating in these discussions, one can *occasionally* experience the same feeling that inspires Lessig in the courtroom, where "some, for the first time in their lives, see power constrained by reason. Not by votes, not by wealth, not by who someone knows—but by an argument that persuades."

That this "architectural" work has gone largely unrecognized to date in law schools, university humanities and social science departments, and even in some civil society-oriented Internet governance fora is not entirely unexpected, because the context and terminology of those discussions is invariably technical, even if many participants recognize that the underlying principles are essentially "constitutional" in nature. No doubt a more complete conversation between code writers and constitutionalists is inevitable over time, and with luck more cross-fertilization will lead to better protocols, better policies, and better architecture.

However, this rapprochement is unlikely to be initiated by technologists seeking to take up the study and application of legal principles. Lessig, whose own intellectual project builds substantially on the antiformalist, "legal realist" school of thought, should understand this reality better than most. In the crudest of forms, legal realism holds that "the Law is whatever lawyers happen to say it is." Stated as neither a boast nor a claim of entitlement but rather as a practical observation of the challenges that lawyers face in applying ambiguous old laws to incommensurable new circumstances, this maxim nevertheless clearly conveys a sense of both the great responsibility and the great power that lawyers command. Perhaps it is time that Mr. Lessig and his counterparts consider the possibility that a similar school of thought may inform (consciously or unconsciously) the perspectives of network builders and code writers. Being of no less good conscience, perhaps code writers and other "cyberspace realists" are merely waiting for the moment when the Law and lawyers come calling with a good story, under the banner of reason rather than power. So long as the story now unfolding continues to make sense and satisfy the ever-expanding audience, we needn't fear either.

Code may not be *that* particular story, but it's an excellent read, and an important contribution to a dialogue that must be engaged.

—*Tom Vest*
`tvest@eyeconomics.com`

Read Any Good Books Lately?

Then why not share your thoughts with the readers of IPJ? We accept reviews of new titles, as well as some of the "networking classics." In some cases, we may be able to get a publisher to send you a book for review if you don't have access to it. Contact us at `ipj@cisco.com` for more information.

Fragments

Global Policy Proposal for Remaining IPv4 Address Space

Global Internet Number Resource Policies are defined by the *Address Supporting Organization* (ASO) MoU[1]—between the *Internet Corporation for Assigned Names and Numbers* (ICANN) and the *Number Resource Organization* (NRO)—as "Internet number resource policies that have the agreement of all RIRs according to their policy development processes and ICANN, and require specific actions or outcomes on the part of the *Internet Assigned Numbers Authority* (IANA) or any other external ICANN-related body in order to be implemented." Attachment A of this MoU describes the *Development Process of Global Internet Number Resource Policies,* including the adoption by every *Regional Internet Registry* (RIR) of a global policy to be forwarded to the ICANN Board by the ASO, as well as its ratification by the ICANN Board. In this context, the ICANN Board adopted its own Procedures[2] for the Review of Internet Number Resource Policies Forwarded by the ASO for Ratification.

Among other features, these Procedures state that the Board will decide, as and when appropriate, that ICANN staff should follow the development of a particular global policy, undertaking an "early awareness" tracking of proposals in the addressing community. To this end, staff should issue background reports periodically, forwarded to the Board, to all ICANN Supporting Organizations and Advisory Committees and posted at the ICANN Web site.

At its meeting on 20 November 2007, the Board resolved to request tracking of the development of a global policy proposal for allocation of remaining IPv4 address space, under discussion in the Regional Internet Registries. The status overview presented below is compiled in response to this request and will be further updated as developments proceed, for information to ICANN entities and the wider community. This is the fifth issue of the tracking of this policy.

Originally, two slightly different global policy proposals were introduced for allocation of the remaining IPv4 address space:

- A version (1) "Global Policy for the Allocation of the Remaining IPv4 Address Space," first presented at LACNIC X in May 2007

- A version (2) "End Policy for IANA IPv4 allocations to RIRs," first presented at APNIC 24 in September 2007

Both featured the same approach, distribution of an equal number N of /8 IPv4 address blocks to each RIR when the IANA free pool would reach the threshold value of $5 \times N,$ but differed in the proposed value of $N,$ notably 2 or 1, respectively. The proposals were discussed in parallel in the RIRs and regarded essentially as one proposal, with a view to converging on a value for $N.$ In February 2008, agreement was reached for a unified proposal (3).

The current proposal is thus:

• Version (3) "Global Policy for the Allocation of the Remaining IPv4 Address Space," first presented at APNIC 25 in February 2008.

The proposal was introduced at the subsequent meetings of all other RIRs. It has now been adopted in ARIN, AfriNIC, LACNIC and RIPE, and is in final call in APNIC. If adopted by all the RIRs, the proposal will subsequently be handled by the NRO Executive Council and the ASO Advisory Council according to their procedures before being submitted to the ICANN Board for ratification. A table[3] can be found on the ICANN Website that indicates the status within each RIR for the current proposal. Hyperlinks are included for easy access.

It should be noted that other policy proposals have been put forward and are being discussed regarding IPv4 address space exhaustion, although only those mentioned above have been scoped as global policy proposals in the sense of the ASO MoU, that is, focusing on address allocation from IANA to the RIRs, and recognized by the ASO AC as global policy proposals in that meaning.

[1] `http://aso.icann.org/docs/aso-mou2004.html`

[2] `http://icann.org/en/general/review-procedures-pgp.html`

[3] `http://www.icann.org/en/announcements/proposal-ipv4-report-29nov07.htm`

Upcoming Events

The *Internet Engineering Task Force* (IETF) will meet in Minneapolis, Minnesota, November 16 – 21, 2008. In 2009, IETF meetings are scheduled for San Francisco, California (March 22 – 27), Stockholm, Sweden (July 26 – 31) and Hiroshima, Japan (November 8 – 13). For more information see `http://www.ietf.org/`

The *North American Network Operators' Group* (NANOG) will meet in Los Angeles, California, October 12 – 14. Immediately following the NANOG meeting, the *American Registry for Internet Numbers* (ARIN) will meet in the same location, October 15 – 17. See `http://nanog.org` and `http://arin.net`

The *Internet Corporation for Assigned Names and Numbers* (ICANN) will meet in Cairo, Egypt, November 2 – 7, 2008. For more information see: `http://icann.org`

The *Asia Pacific Regional Internet Conference on Operational Technologies* (APRICOT) will be held in Manila, Philippines, February 18 – 27, 2009. See: `http://www.apricot2009.net/`

The Internet Protocol Journal, Cisco Systems
170 West Tasman Drive
San Jose, CA 95134-1706
USA

ADDRESS SERVICE REQUESTED

PRSRT STD
U.S. Postage
PAID
PERMIT No. 5187
SAN JOSE, CA