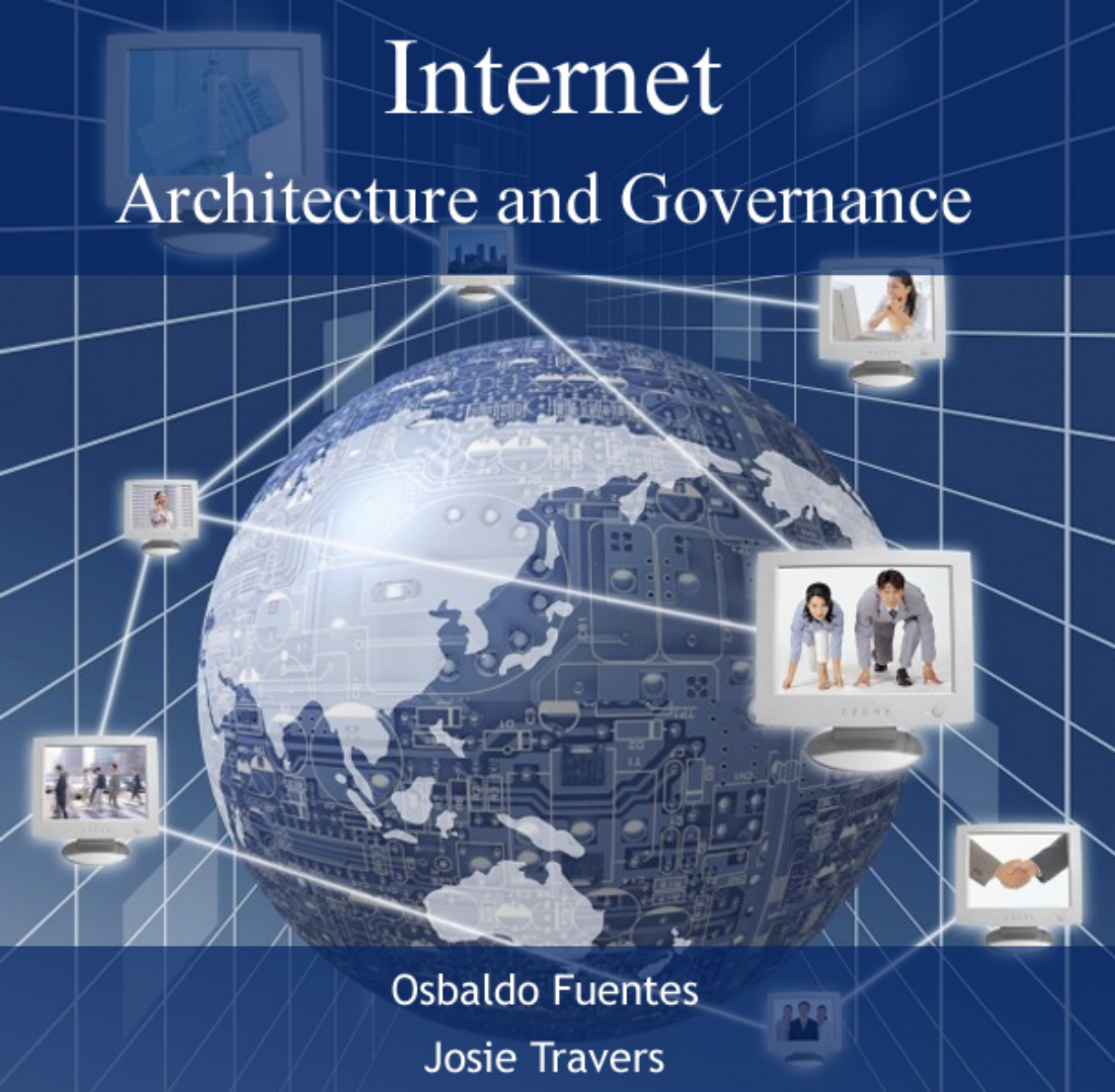
Internet

Architecture and Governance



Osbaldo Fuentes

Josie Travers

First Edition, 2012

ISBN 978-81-323-1282-6

Published by:
College Publishing House
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Border Gateway Protocol

The **Border Gateway Protocol** (**BGP**) is the protocol backing the core routing decisions on the Internet. It maintains a table of IP networks or 'prefixes' which designate network reachability among autonomous systems (AS). It is described as a path vector protocol. BGP does not use traditional Interior Gateway Protocol (**IGP**) metrics, but makes routing decisions based on path, network policies and/or rulesets. For this reason, it is more appropriately termed a reachability protocol rather than routing protocol.

BGP was created to replace the Exterior Gateway Protocol (**EGP**) routing protocol to allow fully decentralized routing in order to allow the removal of the NSFNet Internet backbone network. This allowed the Internet to become a truly decentralized system. Since 1994, version four of the BGP has been in use on the Internet. All previous versions are now obsolete. The major enhancement in version 4 was support of Classless Inter-Domain Routing and use of route aggregation to decrease the size of routing tables. Since January 2006, version 4 is codified in RFC 4271, which went through more than 20 drafts based on the earlier RFC 1771 version 4. RFC 4271 version corrected a number of errors, clarified ambiguities and brought the RFC much closer to industry practices.

Most Internet users do not use BGP directly. Since most Internet service providers must use BGP to establish routing between one another (especially if they are multihomed), it is one of the most important protocols of the Internet. Compare this with Signaling System 7 (SS7), which is the inter-provider core call setup protocol on the PSTN. Very large private IP networks use BGP internally. An example would be the joining of a number of large Open Shortest Path First (OSPF) networks where OSPF by itself would not scale to size. Another reason to use BGP is multihoming a network for better redundancy either to multiple access points of a single ISP (RFC 1998) or to multiple ISPs.

Operation

BGP neighbors, peers are established by manual configuration between routers to create a TCP session on port 179. A BGP speaker will periodically send 19-byte keep-alive messages to maintain the connection (every 60 seconds by default). Among routing protocols, BGP is unique in using TCP as its transport protocol.

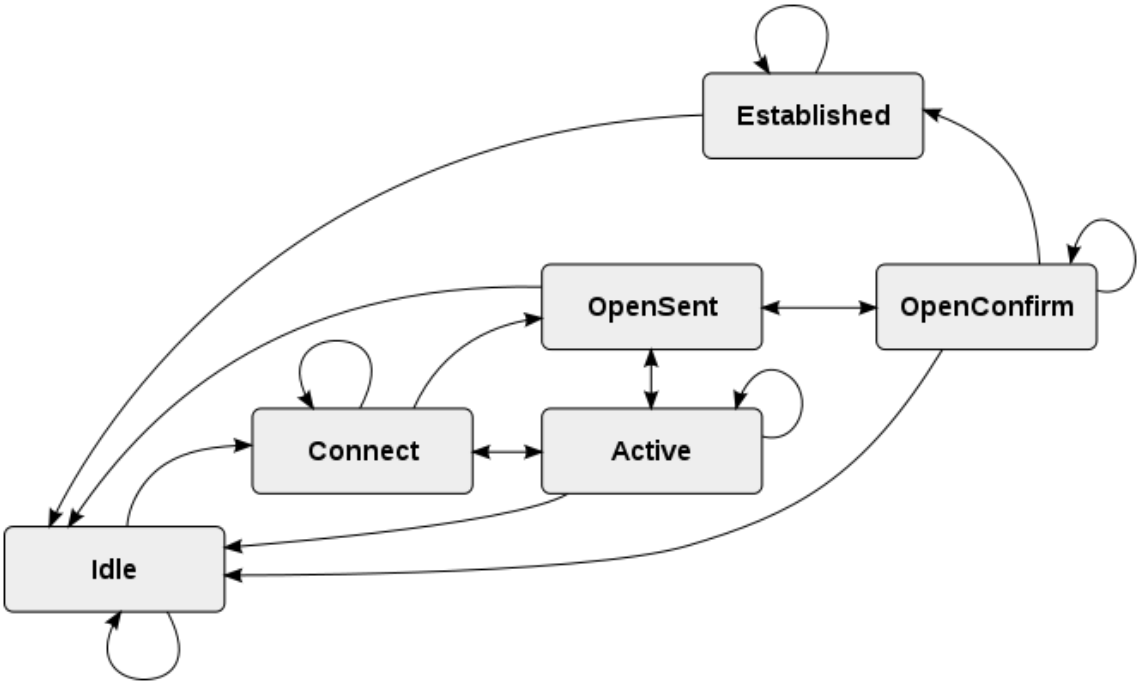
When BGP is running inside an autonomous system (AS), it is referred to as *Internal BGP* (*IBGP* or *Interior Border Gateway Protocol*). When it runs between autonomous systems, it is called *External BGP* (*EBGP* or *Exterior Border Gateway Protocol*). Routers on the boundary of one AS exchanging information with another AS are called border or edge routers. In the Cisco operating system, IBGP routes have an administrative distance of 200, which is less preferred than either external BGP or any interior routing protocol. Other router implementations also prefer EBGP to IGPs, and IGPs to IBGP.

It uses 20 bytes per header.

Extensions Negotiation

During the OPEN BGP speakers can negotiate optional capabilities of the session, including multiprotocol extensions and various recovery modes. If the multiprotocol extensions to BGP are negotiated at the time of creation, the BGP speaker can prefix the Network Layer Reachability Information (NLRI) it advertises with an address family prefix. These families include the IPv4 (default), IPv6, IPv4/IPv6 Virtual Private Networks and multicast BGP. Increasingly, BGP is used as a generalized signaling protocol to carry information about routes that may not be part of the global Internet, such as VPNs.

Finite State Machine



BGP state machine

In order to make decisions in its operations with other BGP peers, a BGP peer uses a simple finite state machine (FSM) that consists of six states: Idle; Connect; Active; OpenSent; OpenConfirm; and Established. For each peer-to-peer session, a BGP implementation maintains a state variable that tracks which of these six states the session is in. The BGP protocol defines the messages that each peer should exchange in order to change the session from one state to another. The first state is the “Idle” state. In the “Idle” state, BGP initializes all resources, refuses all inbound BGP connection attempts and initiates a TCP connection to the peer. The second state is “Connect”. In the “Connect” state, the router waits for the TCP connection to complete and transitions to the "OpenSent" state if successful. If unsuccessful, it resets the ConnectRetry timer and transitions to the "Active" state upon expiration. In the "Active" state, the router resets the ConnectRetry timer to zero and returns to the "Connect" state. In the "OpenSent" state, the router sends an Open message and waits for one in return. Keepalive messages are exchanged and, upon successful receipt, the router is placed into the “Established” state. In the “Established” state, the router can send/receive: Keepalive; Update; and Notification messages to/from its peer.

- **Idle State**:
  - Refuse all incoming BGP connections
  - Start event triggers the initialization of resources for the BGP process.
  - Initiates a TCP connection with its configured BGP peer.
  - Listens for a TCP connection from its peer.
  - Changes its state to Connect.
  - If an error occurs at any state of the FSM process, the BGP session is terminated immediately and returned to the Idle state. Some of the reasons why a router does not progress from the Idle state are:
    - TCP port 179 is not open.
    - A random TCP port over 1023 is not open.
    - Peer address configured incorrectly on either router.
    - AS number configured incorrectly on either router.
- **Connect State**:
  - Waits for successful TCP negotiation with peer.
  - BGP does not spend much time in this state if the TCP session has been successfully established.
  - Sends Open message to peer and changes state to OpenSent.
  - If an error occurs, BGP moves to the Active state. Some reasons for the error are:
    - TCP port 179 is not open.
    - A random TCP port over 1023 is not open.
    - Peer address configured incorrectly on either router.
    - AS number configured incorrectly on either router.
- **Active State**:
  - If the router was unable to establish a successful TCP session, then it ends up in the Active state.
  - BGP FSM will try to restart another TCP session with the peer and, if successful, then it will send an Open message to the peer.

    - If it is unsuccessful again, the FSM is reset to the Idle state.
    - Repeated failures may result in a router cycling between the Idle and Active states. Some of the reasons for this include:
        - TCP port 179 is not open.
        - A random TCP port over 1023 is not open.
        - BGP configuration error.
        - Network congestion.
        - Flapping network interface.
- **OpenSent State**:
    - BGP FSM listens for an Open message from its peer.
    - Once the message has been received, the router checks the validity of the Open message.
    - If there is an error it is because one of the fields in the Open message doesn't match between the peers, e.g. BGP version mismatch, MD5 password mismatch, the peering router expects a different My AS. The router will then send a Notification message to the peer indicating why the error occurred.
    - If there is no error, a Keepalive message is sent, various timers are set and the state is changed to OpenConfirm.
- **OpenConfirm State**:
    - The peer is listening for a Keepalive message from its peer.
    - If a Keepalive message is received and no timer has expired before reception of the Keepalive, BGP transitions to the Established state.
    - If a timer expires before a Keepalive message is received, or if an error condition occurs, the router transitions back to the Idle state.
- **Established State**:
    - In this state, the peers send Update messages to exchange information about each route being advertised to the BGP peer.
    - If there is any error in the Update message then a Notification message is sent to the peer, and BGP transitions back to the Idle state.
    - If a timer expires before a Keepalive message is received, or if an error condition occurs, the router transitions back to the Idle state.

Basic BGP updates

Once a BGP session is running, the BGP speakers exchange UPDATE messages about destinations to which the speaker offers connectivity. In the protocol, the basic CIDR route description is called Network Layer Reachability Information (NLRI). NLRI includes the expected destination prefix, prefix length, path of autonomous systems to the destination and next hop in **attributes**, which can carry a wide range of additional information that affects the acceptance policy of the receiving router. BGP speakers incrementally announce new NLRI to which they offer reachability, but also announce *withdrawals* of prefixes to which the speaker no longer offers connectivity.

BGP router connectivity and learning routes

In the simplest arrangement all routers within a single AS and participating in BGP routing must be configured in a full mesh: each router must be configured as peer to every other router. This causes scaling problems, since the number of required connections grows quadratically with the number of routers involved. To alleviate the problem, BGP implements two options: route reflectors (RFC 4456) and confederations (RFC 5065). The following discussion of basic UPDATE processing assumes a full IBGP mesh.

Basic update processing

A given BGP router may accept NLRI in UPDATEs from multiple neighbors and advertise NLRI to the same, or a different set, of neighbors. Conceptually, BGP maintains its own "master" routing table, called the *Loc-RIB* (Local Routing Information Base), separate from the main routing table of the router. For each neighbor, the BGP process maintains a conceptual *Adj-RIB-In* (Adjacent Routing Information Base, Incoming) containing the NLRI received from the neighbor, and a conceptual *Adj-RIB-Out* (Outgoing) for NLRI to be sent to the neighbor.

Conceptual, in the preceding paragraph, means that the physical storage and structure of these various tables are decided by the implementer of the BGP code. Their structure is not visible to other BGP routers, although they usually can be interrogated with management commands on the local router. It is quite common, for example, to store the two Adj-RIBs and the Loc-RIB together in the same data structure, with additional information attached to the RIB entries. The additional information tells the BGP process such things as whether individual entries belong in the Adj-RIBs for specific neighbors, whether the per-neighbor route selection process made received policies eligible for the Loc-RIB, and whether Loc-RIB entries are eligible to be submitted to the local router's routing table management process.

By *eligible to be submitted*, BGP will submit the routes that it considers best to the main routing table process. Depending on the implementation of that process, the BGP route is not necessarily selected. For example, a directly connected prefix, learned from the router's own hardware, is usually most preferred. As long as that directly connected route's interface is active, the BGP route to the destination will not be put into the routing table. Once the interface goes down, and there are no more preferred routes, the Loc-RIB route would be installed in the main routing table. Until recently, it was a common mistake to say *BGP carries policies*. BGP actually carried the information with which rules inside BGP-speaking routers could make policy decisions. Some of the information carried that is explicitly intended to be used in policy decisions are communities and multi-exit discriminators (MED).

Route selection

The BGP standard specifies a number of decision factors, more than are used by any other common routing process, for selecting NLRI (Network Layer Reachability Information) to go into the Loc-RIB (Routing Information Base). The first decision point for evaluating NLRI is that its next-hop attribute must be reachable (or resolvable). Another way of saying the next-hop must be reachable is that there must be an active route, already in the main routing table of the router, to the prefix in which the next-hop address is located.

Next, for each neighbor, the BGP process applies various standard and implementation-dependent criteria to decide which routes conceptually should go into the Adj-RIB-In. The neighbor could send several possible routes to a destination, but the first level of preference is at the neighbor level. Only one route to each destination will be installed in the conceptual Adj-RIB-In. This process will also delete, from the Adj-RIB-In, any routes that are withdrawn by the neighbor.

Whenever a conceptual Adj-RIB-In changes, the main BGP process decides if any of the neighbor's new routes are preferred to routes already in the Loc-RIB. If so, it replaces them. If a given route is withdrawn by a neighbor, and there is no other route to that destination, the route is removed from the Loc-RIB, and no longer sent, by BGP, to the main routing table manager. If the router does not have a route to that destination from any non-BGP source, the withdrawn route will be removed from the main routing table.

Per-neighbor decisions

After verifying that the next hop is reachable, if the route comes from an internal (i.e. IBGP) peer, the first rule to apply according to the standard is to examine the LOCAL_PREF attribute. If there are several IBGP routes from the neighbor, the one with the highest LOCAL_PREF is selected unless there are several routes with the same LOCAL_PREF. In the latter case the route selection process moves to the next tie breaker. While LOCAL_PREF is the first rule in the standard, once reachability of the NEXT_HOP is verified, Cisco and several other vendors first consider a decision factor called WEIGHT which is local to the router (i.e. not transmitted by BGP). The route with the highest WEIGHT is preferred.

LOCAL_PREF, WEIGHT, and other criteria can be manipulated by local configuration and software capabilities. Such manipulation is outside the scope of the standard but is commonly used. For example the COMMUNITY attribute (see below) is not directly used by the BGP selection process. The BGP neighbor process however can have a rule to set LOCAL_PREFERENCE or another factor based on a manually programmed rule to set the attribute if the COMMUNITY value matches some pattern matching criterion. If the route was learned from an external peer the per-neighbor BGP process computes a LOCAL_PREFERENCE value from local policy rules and then compares the LOCAL_PREFERENCE of all routes from the neighbor.

At the per-neighbor level - ignoring implementation-specific policy modifiers - the order of tie breaking rules is:

1. Prefer the route with the shortest AS_PATH. An AS_PATH is the set of AS numbers that must be traversed to reach the advertised destination. AS1-AS2-AS3 is shorter than AS4-AS5-AS6-AS7.
2. Prefer routes with the lowest value of their ORIGIN attribute.
3. Prefer routes with the lowest MULTI_EXIT_DISC (multi-exit discriminator or MED) value.

Before the most recent edition of the BGP standard, if an UPDATE had no MULTI_EXIT_DISC value, several implementations created a MED with the least possible value. The current standard however specifies that missing MEDs are to be treated as the highest possible value. Since the current rule may cause different behavior than the vendor interpretations, BGP implementations that used the nonstandard default value have a configuration feature that allows the old or standard rule to be selected.

Decision factors at the Loc-RIB level

Once candidate routes are received from neighbors, the Loc-RIB software applies additional tie-breakers to routes to the same destination.

1. If at least one route was learned from an external neighbor (i.e., the route was learned from EBGP), drop all routes learned from IBGP.
2. Prefer the route with the lowest interior cost to the NEXT_HOP, according to the main Routing Table. If two neighbors advertised the same route, but one neighbor is reachable via a low-bitrate link and the other by a high-bitrate link, and the interior routing protocol calculates lowest cost based on highest bitrate, the route through the high-bitrate link would be preferred and other routes dropped.

If there is more than one route still tied at this point, several BGP implementations offer a configurable option to load-share among the routes, accepting all (or all up to some number).

1. Prefer the route learned from the BGP speaker with the numerically lowest BGP identifier
2. Prefer the route learned from the BGP speaker with the lowest peer IP address

Communities

BGP communities are attribute tags that can be applied to incoming or outgoing prefixes to achieve some common goal (RFC 1997). While it is common to say that BGP allows an administrator to set policies on how prefixes are handled by ISPs, this is generally not possible, strictly speaking. For instance, BGP natively has no concept to allow one AS to tell another AS to restrict advertisement of a prefix to only North American peering customers. Instead, an ISP generally publishes a list of well-known or proprietary

communities with a description for each one, which essentially becomes an agreement of how prefixes are to be treated. Examples of common communities include local preference adjustments, geographic or peer type restrictions, DoS avoidance (black holing), and AS prepending options. An ISP might state that any routes received from customers with community XXX:500 will be advertised to all peers (default) while community XXX:501 will restrict advertisement to North America. The customer simply adjusts their configuration to include the correct community(ies) for each route, and the ISP is responsible for controlling who the prefix is advertised to. The end user has no technical ability to enforce correct actions being taken by the ISP, though problems in this area are generally rare and accidental.

It is a common tactic for end customers to use BGP communities (usually ASN:70,80,90,100) to control the local preference the ISP assigns to advertised routes instead of using MED (the effect is similar). It should also be noted that the community attribute is transitive, but communities applied by the customer very rarely become propagated outside the next-hop AS.

Extended communities

The BGP Extended Community Attribute was added in 2006 in order to extend the range of such attributes and to provide a community attribute structuring by means of a type field. The extended format consists of one or two octets for the type field followed by seven or six octets for the respective community attribute content. The definition of this Extended Community Attribute is documented in RFC 4360. The IANA administers the registry for BGP Extended Communities Types. The Extended Communities Attribute itself is a transitive optional BGP attribute. However, a bit in the type field within the attribute decides whether the encoded extended community is of a transitive or non-transitive nature. The IANA registry therefore provides different number ranges for the attribute types. Due to the extended attribute range, its usage can be manifold. RFC 4360 exemplarly defines the "Two-Octet AS Specific Extended Community", the "IPv4 Address Specific Extended Community", the "Opaque Extended Community", the "Route Target Community" and the "Route Origin Community". A number of BGP QoS drafts also use this Extended Community Attribute structure for inter-domain QoS signalling.

Uses of multi-exit discriminators

MEDs, defined in the main BGP standard, were originally intended to show to another neighbor AS the advertising AS's preference as to which of several links are preferred for inbound traffic. Another application of MEDs is to advertise the value, typically based on delay, of multiple AS that have presence at an IXP, that they impose to send traffic to some destination.

BGP problems and mitigation

Internal BGP scalability

An autonomous system with internal BGP (IBGP) must have all of its IBGP peers connect to each other in a full mesh (where everyone speaks to everyone directly). This full-mesh configuration requires that each router maintain a session to every other router. In large networks, this number of sessions may degrade performance of routers, due either to a lack of memory, or too much CPU process requirements.

Route reflectors and confederations both reduce the number of IBGP peers to each router and thus reduce processing overhead. Route reflectors are a pure performance-enhancing technique, while confederations also can be used to implement more fine-grained policy.

Route reflectors reduce the number of connections required in an AS. A single router (or two for redundancy) can be made a route reflector: other routers in the AS need only be configured as peers to them.

Confederations are sets of autonomous systems. In common practice, only one of the confederation AS numbers is seen by the Internet as a whole. Confederations are used in very large networks where a large AS can be configured to encompass smaller more manageable internal ASs.

Confederations can be used in conjunction with route reflectors. Both confederations and route reflectors can be subject to persistent oscillation, unless specific design rules, affecting both BGP and the interior routing protocol, are followed.

However, these alternatives can introduce problems of their own, including the following:

- route oscillation,
- sub-optimal routing,
- increase of BGP convergence time

Additionally, route reflectors and BGP confederations were not designed to ease BGP router configuration. Nevertheless, these are common tools for experienced BGP network architects. These tools may be combined, for example, as a hierarchy of route reflectors.

Instability

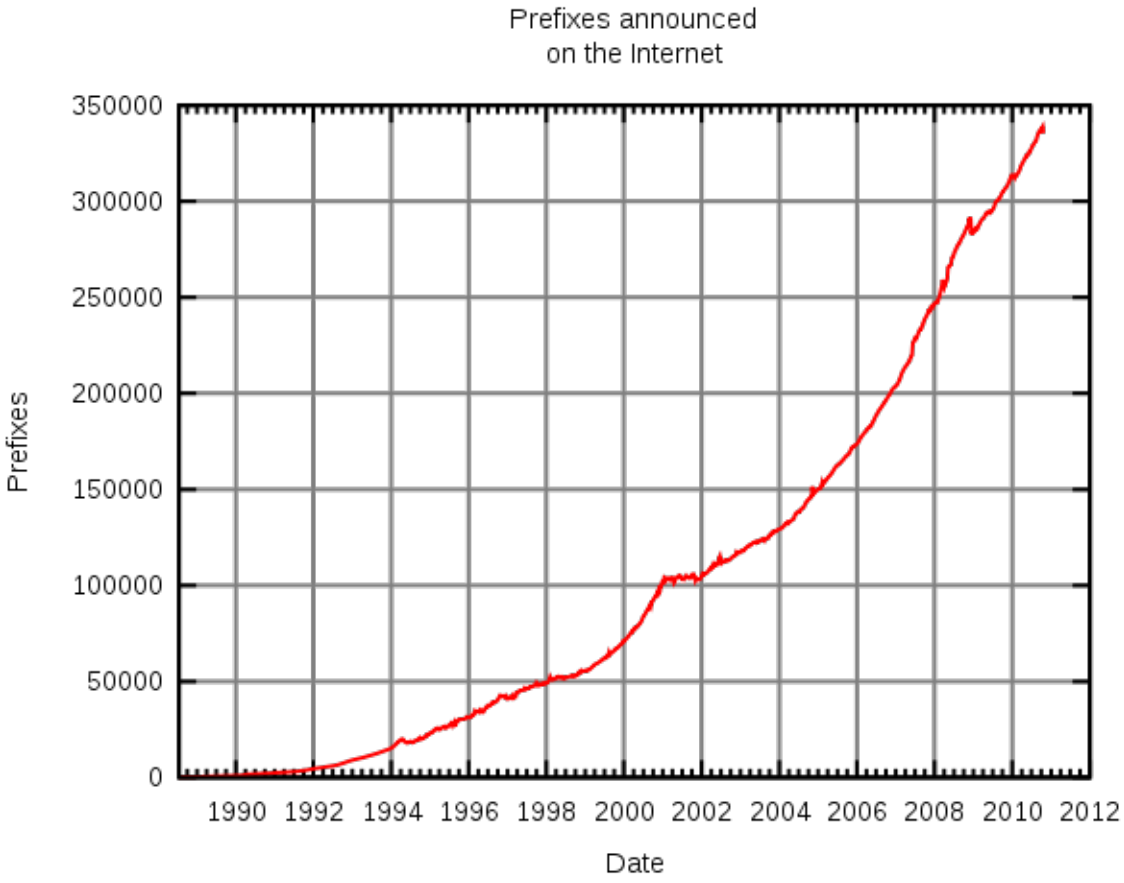
The routing tables managed by a BGP implementation are adjusted continually to reflect actual changes in the network, such as links breaking and being restored or routers going down and coming back up. In the network as a whole it is normal for these changes to happen almost continuously, but for any particular router or link changes are supposed to be relatively infrequent. If a router is misconfigured or mismanaged then it may get into a rapid cycle between down and up states. This pattern of repeated withdrawal and

reannouncement, known as route flapping, can cause excessive activity in all the other routers that know about the broken link, as the same route is continuously injected and withdrawn from the routing tables. The BGP design is such that delivery of traffic may not function while routes are being updated. On the Internet, a BGP routing change may cause outages for several minutes.

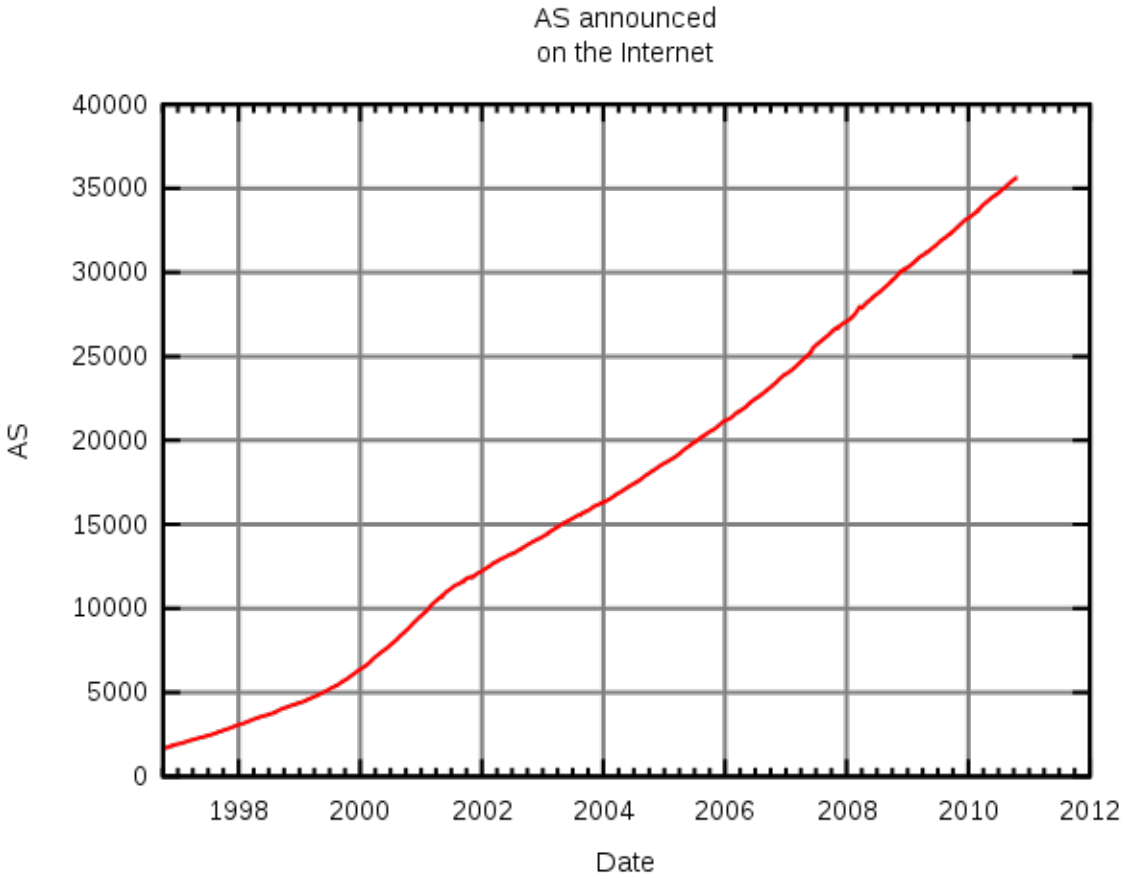
A feature known as *route flap damping* (RFC 2439) is built into many BGP implementations in an attempt to mitigate the effects of route flapping. Without damping the excessive activity can cause a heavy processing load on routers, which may in turn delay updates on other routes, and so affect overall routing stability. With damping, a route's flapping is exponentially decayed. At the first instance when a route becomes unavailable and quickly reappears, damping does not take effect, so as to maintain the normal fail-over times of BGP. At the second occurrence, BGP shuns that prefix for a certain length of time; subsequent occurrences are timed out exponentially. After the abnormalities have ceased and a suitable length of time has passed for the offending route, prefixes can be reinstated and its slate wiped clean. Damping can also mitigate denial of service attacks; damping timings are highly customizable.

However, subsequent research has shown that flap damping can actually lengthen convergence times in some cases, and can cause interruptions in connectivity even when links are not flapping. Moreover, as backbone links and router processors have become faster, some network architects have suggested that flap damping may not be as important as it used to be, since changes to the routing table can be absorbed much faster by routers. This has led the RIPE Route Working Group to write that *"with the current implementations of BGP flap damping, the application of flap damping in ISP networks is NOT recommended. ... If flap damping is implemented, the ISP operating that network will cause side-effects to their customers and the Internet users of their customers' content and services ... . These side-effects would quite likely be worse than the impact caused by simply not running flap damping at all."* Improving stability without the problems of flap damping is the subject of current research.

Routing table growth



BGP table growth on the Internet



Number of AS on the Internet

One of the largest problems faced by BGP, and indeed the Internet infrastructure as a whole, is the growth of the Internet routing table. If the global routing table grows to the point where some older, less capable, routers cannot cope with the memory requirements or the CPU load of maintaining the table, these routers will cease to be effective gateways between the parts of the Internet they connect. In addition, and perhaps even more importantly, larger routing tables take longer to stabilize (see above) after a major connectivity change, leaving network service unreliable, or even unavailable, in the interim.

Until late 2001, the global routing table was growing exponentially, threatening an eventual widespread breakdown of connectivity. In an attempt to prevent this, ISPs cooperated in keeping the global routing table as small as possible, by using Classless Inter-Domain Routing (CIDR) and route aggregation. While this slowed the growth of the routing table to a linear process for several years, with the expanded demand for multihoming by end user networks the growth was once again exponential by the middle of 2004. As of April 2010, the routing table has in excess of 310,000 entries.

Route summarization is often used to improve aggregation of the BGP global routing table, thereby reducing the necessary table size in routers of an AS. Consider AS1 has been allocated the big address space of 172.16.0.0/16, this would be counted as one route in the table, but due to customer requirement or traffic engineering purposes, AS1 wants to announce smaller, more specific routes of 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18. The prefix 172.16.192.0/18 does not have any hosts so AS1 does not announce a specific route 172.16.192.0/18. This all counts as AS1 announcing four routes.

AS2 will see the 4 routes from AS1 (172.16.0.0/16, 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18) and it is up to the routing policy of AS2 to decide whether or not to take a copy of the four routes or, as 172.16.0.0/16 overlaps all the other specific routes, to just store the summary, 172.16.0.0/16.

If AS2 wants to send data to prefix 172.16.192.0/18, it will be sent to the routers of AS1 on route 172.16.0.0/16. At AS1's router, it will either be dropped or a destination unreachable ICMP message will be sent back, depending on the configuration of AS1's routers.

If AS1 later decides to drop the route 172.16.0.0/16, leaving 172.16.0.0/18, 172.16.64.0/18 and 172.16.128.0/18, AS1 will drop the number of routes it announces to three. AS2 will see the three routes, and depending on the routing policy of AS2, it will store a copy of the three routes, or aggregate the prefix's 172.16.0.0/18 and 172.16.64.0/18 to 172.16.0.0/17, thereby reducing the number of routes AS2 stores to only two: 172.16.0.0/17 and 172.16.128.0/18.

If AS2 wants to send data to prefix 172.16.192.0/18, it will be dropped or a destination unreachable ICMP message will be sent back at the routers of AS2 (not AS1 as before), because 172.16.192.0/18 would not be in the routing table.

Load-balancing problem

Another factor causing this growth of the routing table is the need for load balancing of multi-homed networks. It is not a trivial task to balance the inbound traffic to a multi-homed network across its multiple inbound paths, due to limitation of the BGP route selection process. For a multi-homed network, if it announces the same network blocks across all of its BGP peers, the result may be that one or several of its inbound links become congested while the other links remain under-utilized, because external networks all picked that set of congested paths as optimal. Like most other routing protocols, the BGP protocol does not detect congestion.

To work around this problem, BGP administrators of that multihomed network may divide a large continuous IP address block into smaller blocks, and tweak the route announcement to make different blocks look optimal on different paths, so that external networks will choose a different path to reach different blocks of that multi-homed network. Such cases will increase the number of routes as seen on the global BGP table.

Requirements of a router for use of BGP for Internet and backbone-of-backbones purposes

Routers, especially small ones intended for Small Office/Home Office (SOHO) use, may not include BGP software. Some SOHO routers simply are not capable of running BGP using BGP routing tables of any size. Other commercial routers may need a specific software executable image that contains BGP, or a license that enables it. Open source packages that run BGP include GateD, GNU Zebra, Quagga, OpenBGPD, BIRD, XORP and Vyatta. Devices marketed as Layer 3 switches are less likely to support BGP than devices marketed as routers, but high-end Layer 3 Switches usually can run BGP.

Products marketed as switches may or may not have a size limitation on BGP tables, such as 20,000 routes, far smaller than a full Internet table plus internal routes. These devices, however, may be perfectly reasonable and useful when used for BGP routing of some smaller part of the network, such as a confederation-AS representing one of several smaller enterprises that are linked, by a BGP backbone of backbones, or a small enterprise that announces routes to an ISP but only accepts a default route and perhaps a small number of aggregated routes.

A BGP router used only for a network with a single point of entry to the Internet may have a much smaller routing table size (and hence RAM and CPU requirement) than a multihomed network. Even simple multihoming can have modest routing table size. The actual amount of memory required in a BGP router depends on the amount of BGP information exchanged with other BGP speakers, and the way in which the particular router stores BGP information. The router may have to keep more than one copy of a route, so it can manage different policies for route advertising and acceptance to a specific neighboring AS. The term view is often used for these different policy relationships on a running router.

If one router implementation takes more memory per route than another implementation, this may be a legitimate design choice, trading processing speed against memory. A full BGP table as of April 2010 is in excess of 310,000 prefixes. Large ISPs may add another 50% for internal and customer routes. Again depending on implementation, separate tables may be kept for each view of a different peer AS.

Free and open source implementations of BGP

- Bird Internet routing daemon, a GPL routing package for Unix-like systems.
- GNU Zebra, a GPL routing suite supporting BGP4.
- OpenBGPD, a BSD licensed implementation by the OpenBSD team.
- Quagga, a fork of GNU Zebra for Unix-like systems.
- Vyatta, a commercial open-source router/firewall/VPN - network operating system.
- XORP, the eXtensible Open Router Platform, a BSD licensed suite of routing protocols.

BGP simulators

- BGPviz, a Flash application that presents a graphical visualization of BGP routes and updates for any real AS on the Internet
- SSFnet, SSFnet network simulator includes a BGP implementation developed by BJ Premore
- C-BGP, a BGP simulator able to perform large-scale simulation trying to model the ASes of the Internet or modelling ASes as large as Tier-1.
- BGP++, a patch integrating GNU Zebra software on ns-2 and GTNetS network simulators
- ns-BGP, a BGP extension for ns-2 simulator based on the SSFnet implementation
- NetViews, a Java application that monitors and visualizes BGP activity in real time.

Test equipment

Systems for testing BGP conformance, load or stress performance come from vendors such as:

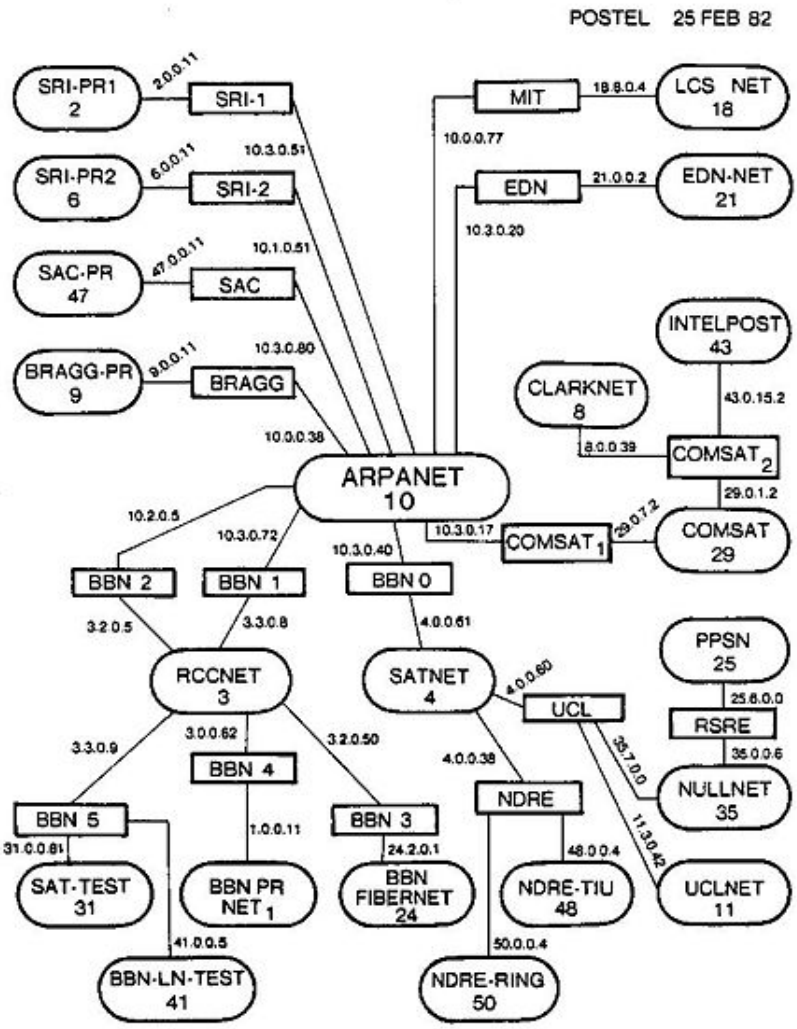
- Ixia Communications
- Spirent Communications
- Agilent Technologies

Chapter 2

Classful Network

A **classful network** is a network addressing architecture used in the Internet from 1981 until the introduction of Classless Inter-Domain Routing in 1993. The method divides the address space for Internet Protocol Version 4 (IPv4) into five address classes. Each class, coded in the first four bits of the address, defines either a different network size, i.e. number of hosts for unicast addresses (classes A, B, C), or a multicast network (class D). The fifth class (E) address range is reserved for future or experimental purposes.

Since its discontinuation, remnants of classful network concepts remain in practice only in limited scope in the default configuration parameters of some network software and hardware components (e.g., default subnet mask), but the terms are often still heard in general discussions of network structure among network administrators.



Map of the prototype Internet in 1982, showing 8-bit-numbered networks (ovals) only, interconnected by routers (rectangles).

Background

Originally, a 32-bit IPv4 address was logically subdivided into the *network number* field, the most-significant 8 bits of an address, which specified the particular network a host was attached to, and the local address, also called *rest field* (the rest of the address), which uniquely identifies a host connected to that network. This format was sufficient at a time when only a few large networks existed, such as the ARPANET which was assigned the network number 10, and before the wide proliferation of local area networks (LANs). As a consequence of this architecture, the address space supported only a low number (254) of independent networks, and it became clear very early on that this would not be enough.

Introduction of address classes

Expansion of the network had to ensure compatibility with the existing address space and the Internet Protocol (IP) packet structure, and avoid the renumbering of the existing networks. The solution was to expand the definition of the network number field to include more bits, allowing more networks to be designated, each potentially having fewer hosts. All existing network numbers at the time were smaller than 64, they only used the 6 least-significant bits of the network number field. Thus it was possible to use the most-significant bits of an address to introduce a set of address classes, while preserving the existing network numbers in the first of these classes.

The new addressing architecture was introduced by RFC 791 in 1981 as a part of the specification of the Internet Protocol. It divided the address space into primarily three address formats, henceforth called address *classes*, and left a fourth range reserved to be defined later.

The first class, designated as *Class A*, contained all addresses in which the most significant bit is zero. The network number for this class is given by the next 7 bits, therefore accommodating 128 networks in total, including the zero network, and including the existing IP networks already allocated. A *Class B* network was a network in which all addresses had the two most-significant bits set to 1 and 0. For these networks, the network address was given by the next 14 bits of the address, thus leaving 16 bits for numbering host on the network for a total of 65536 addresses per network. *Class C* was defined with the 3 high-order bits set to 1, 1, and 0, and designating the next 21 bits to number the networks, leaving each network with 256 local addresses.

The leading bit sequence *111* designated an "*escape to extended addressing mode*", which was later subdivided in to Class D (*1110*) for multicast addressing, while leaving as reserved for future use the *1111* block designated as Class E.

This addressing scheme is illustrated in the following table:

Class	Leading bits	Size of *network number* bit field	Size of *rest* bit field	Number of networks	Addresses per network	Start address	End address
Class A	0	8	24	128 (2^7)	16,777,216 (2^{24})	0.0.0.0	127.255.255.255
Class B	10	16	16	16,384 (2^{14})	65,536 (2^{16})	128.0.0.0	191.255.255.255
Class C	110	24	8	2,097,152 (2^{21})	256 (2^8)	192.0.0.0	223.255.255.255
Class D (multicast)	1110	not defined	not defined	not defined	not defined	224.0.0.0	239.255.255.255

Class E (reserved)	1111	not defined	not defined	not defined	not defined	240.0.0.0	255.255.255.255

The number of addresses usable for addressing specific hosts in each network is always $2^N - 2$ (where N is the number of rest field bits, and the subtraction of 2 adjusts for the use of the all-bits-zero host portion for network address and the all-bits-one host portion as a broadcast address. Thus, for a Class C address with 8 bits available in the host field, the number of hosts is 254.

Today, IP addresses are associated with a subnet mask. This was not required in a classful network because the mask was implicitly derived from the IP address itself. Any network device would inspect the first few bits of the IP address to determine the class of the address.

Bit-wise representation

In the following table:

- *n* indicates a binary slot used for network ID.
- *H* indicates a binary slot used for host ID.
- *X* indicates a binary slot (without specified purpose)

```
Class A
  0.  0.  0.  0 = 00000000.00000000.00000000.00000000
127.255.255.255 = 01111111.11111111.11111111.11111111
                  0nnnnnnn.HHHHHHHH.HHHHHHHH.HHHHHHHH

Class B
128.  0.  0.  0 = 10000000.00000000.00000000.00000000
191.255.255.255 = 10111111.11111111.11111111.11111111
                  10nnnnnn.nnnnnnnn.HHHHHHHH.HHHHHHHH

Class C
192.  0.  0.  0 = 11000000.00000000.00000000.00000000
223.255.255.255 = 11011111.11111111.11111111.11111111
                  110nnnnn.nnnnnnnn.nnnnnnnn.HHHHHHHH

Class D
224.  0.  0.  0 = 11100000.00000000.00000000.00000000
239.255.255.255 = 11101111.11111111.11111111.11111111
                  1110XXXX.XXXXXXXX.XXXXXXXX.XXXXXXXX

Class E
240.  0.  0.  0 = 11110000.00000000.00000000.00000000
255.255.255.255 = 11111111.11111111.11111111.11111111
                  1111XXXX.XXXXXXXX.XXXXXXXX.XXXXXXXX
```

The replacement of classes

The first architecture change extended the addressing capability in the Internet, but did not prevent IP address shortage. The principal problem was that many sites needed larger address blocks than a Class C network provided, and therefore they received a Class B block, which was in most cases much larger than required. In the rapid growth of the Internet, the pool of unassigned Class B addresses (2^{14}, or about 16,000) was rapidly being depleted. Classful networking was replaced by Classless Inter-Domain Routing (CIDR), starting in 1993 with the specification of RFC 1518 and RFC 1519, to attempt to solve this problem.

Early allocations of IP addresses by the Internet Assigned Numbers Authority (IANA) were in some cases not made efficiently, which contributed to the problem. However, the commonly held notion that some American organizations unfairly or unnecessarily received Class A networks is wrong; most such allocations date to the period before the introduction of address classes, when the only address blocks available were what later became known as Class A networks.

Chapter 3

Classless Inter-Domain Routing

Classless Inter-Domain Routing (**CIDR**) is a methodology of allocating IP addresses and routing Internet Protocol packets. It was introduced in 1993 to replace the prior addressing architecture of classful network design in the Internet with the goal to slow the growth of routing tables on routers across the Internet, and to help slow the rapid exhaustion of IPv4 addresses.

IP addresses are described as consisting of two groups of bits in the address: the most significant part is the *network address* which identifies a whole network or subnet and the least significant portion is the *host identifier*, which specifies a particular host interface on that network. This division is used as the basis of traffic routing between IP networks and for address allocation policies. Classful network design for IPv4 sized the network address as one or more 8-bit groups, resulting in the blocks of Class A, B, or C addresses. Classless Inter-Domain Routing allocates address space to Internet service providers and end users on any address bit boundary, instead of on 8-bit segments. In IPv6, however, the interface identifier has a fixed size of 64 bits by convention, and smaller subnets are never allocated to end users.

CIDR notation is a syntax of specifying IP addresses and their associated routing prefix. It appends to the address a slash character and the decimal number of leading bits of the routing prefix, e.g., 192.168.0.0/16 for IPv4, and 2001:db8::/32 for IPv6.

Background

During the first decade of the modern Internet after the invention of the Domain Name System (DNS) it became apparent that the devised system based on the classful network scheme of allocating the IP address space and the routing of IP packets was not scalable.

To alleviate the shortcomings, the Internet Engineering Task Force published in 1993 a new set of standards, RFC 1518 and RFC 1519, to define a new concept of allocation of IP address blocks and new methods of routing IPv4 packets. A new version of the specification was published as RFC 4632 in 2006.

An IP address is interpreted as composed of two parts: a network-identifying prefix followed by a host identifier within that network. In the previous classfull network architecture, IP address allocations were based on dividing the 32 bits into 8-bit segments

called octets. An address was considered to be the combination of an 8, 16, or 24-bit network prefix along with a 24, 16, or 8-bit individual or "node" address. Thus, the smallest allocation and routing block contained only 256 addresses—too small for most enterprises, and the next larger block contained 65,536 addresses—too large to be used efficiently by even large organizations. This led to inefficiencies in address use as well as routing because the large number of allocated small (class-C) networks with individual route announcements, being geographically dispersed with little opportunity for route aggregation, created heavy demand on routing equipment.

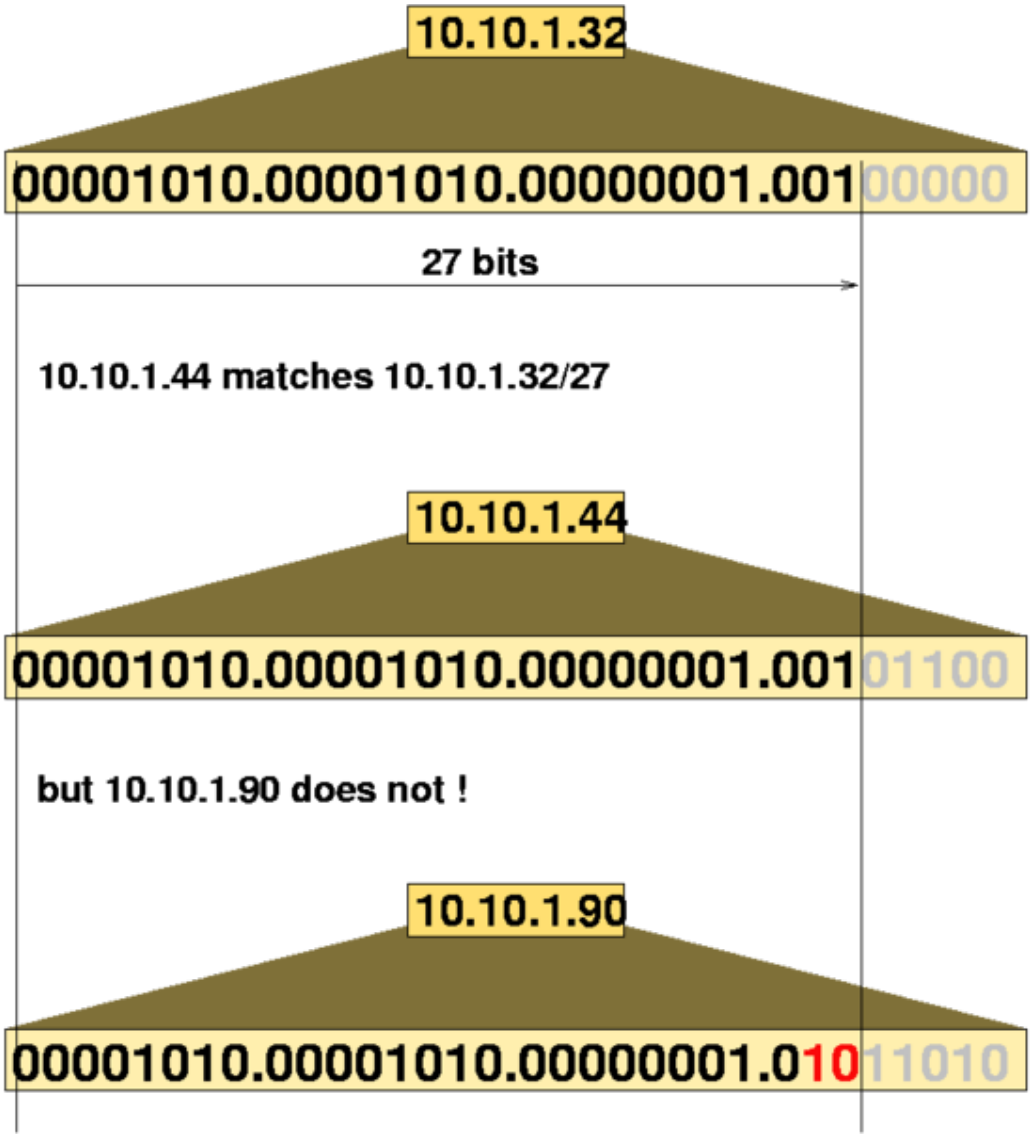
As the experimental TCP/IP network expanded into the Internet during the 1980s, the need for more flexible addressing schemes became increasingly apparent. This led to the successive development of subnetting and CIDR. Because the old class distinctions are ignored, the new system was called *classless routing*. It is supported by modern routing protocols, such as RIP-2, EIGRP, IS-IS and OSPF. This led to the original system being called, by back-formation, classful routing.

Classless Inter-Domain Routing is based on *variable-length subnet masking* (VLSM), which allows a network to be divided into different-sized subnets. CIDR avoids wasting IP addresses. Variable-length subnet masks are mentioned in RFC 950.

CIDR encompasses:

- the VLSM technique with effective qualities of specifying arbitrary-length prefixes;
- the CIDR notation in which an address or routing prefix is written with a suffix indicating the number of bits in the address, such as 192.168.0.0/16;
- the administrative process of allocating address blocks to organizations based on their actual and short-term projected needs;
- the aggregation of multiple contiguous prefixes into supernets, and, wherever possible in the Internet, advertising aggregates, thus reducing the number of entries in the global routing table.

CIDR blocks



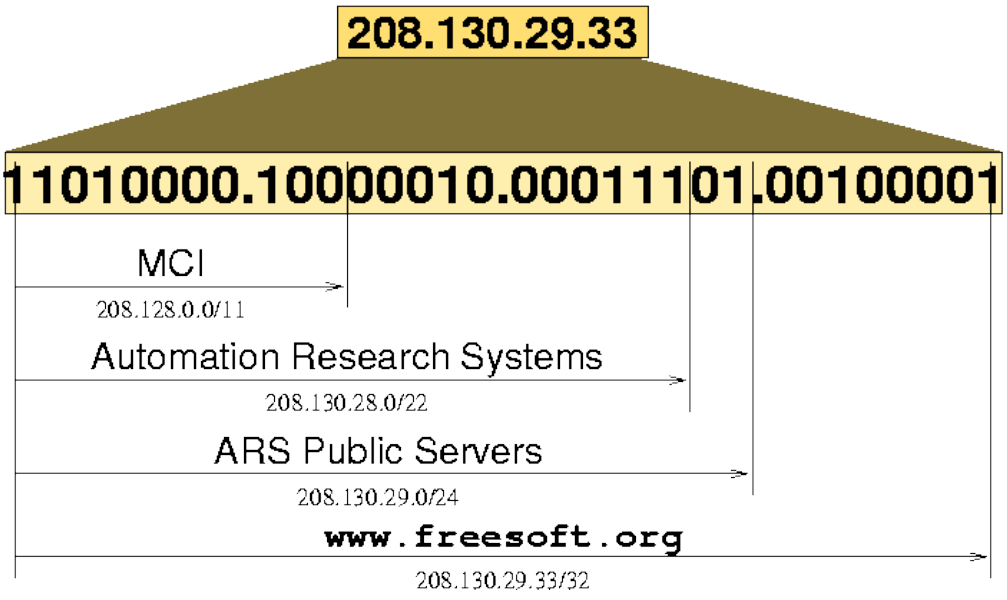
CIDR is principally a bitwise, prefix-based standard for the interpretation of IP addresses. It facilitates routing by allowing blocks of addresses to be grouped into single routing table entries. These groups, commonly called *CIDR blocks*, share an initial sequence of bits in the binary representation of their IP addresses. IPv4 CIDR blocks are identified using a syntax similar to that of IPv4 addresses: a four-part dotted-decimal address, followed by a slash, then a number from 0 to 32: *A.B.C.D/N*. The dotted decimal portion is interpreted, like an IPv4 address, as a 32-bit binary number that has been broken into four octets. The number following the slash is the prefix length, the number of shared initial bits, counting from the most significant bit of the address. When emphasizing only the size of a network, terms like */20* are used, which is a CIDR block with an unspecified 20-bit prefix.

An IP address is part of a CIDR block, and is said to match the CIDR prefix, if the initial N bits of the address and the CIDR prefix are the same. Thus, understanding CIDR requires that IP address be visualized in binary. Since the length of an IPv4 address has 32 bits, an N-bit CIDR prefix leaves 32-N bits unmatched, meaning that 2^{32-N} IPv4 addresses match a given N-bit CIDR prefix. *Shorter* CIDR prefixes match more addresses, while *longer* CIDR prefixes match fewer. An address can match multiple CIDR prefixes of different lengths.

CIDR is also used with IPv6 addresses and the syntax semantic is identical. A prefix length can range from 0 to 128, due to the larger number of bits in the address, however, by convention a subnet on broadcast MAC layer networks always has 64-bit host identifiers. Larger prefixes are rarely used even on point-to-point links.

Assignment of CIDR blocks

The Internet Assigned Numbers Authority (IANA) issues to Regional Internet Registries (RIRs) large, short-prefix (typically /8) CIDR blocks. For example, 62.0.0.0/8, with over sixteen million addresses, is administered by RIPE NCC, the European RIR. The RIRs, each responsible for a single, large, geographic area (such as Europe or North America), then subdivide these allocations into smaller blocks and issue them to local Internet registries. This subdividing process can be repeated several times at different levels of delegation. End user networks receive subnets sized according to the size of their network and projected short term need. Networks served by a single ISP are encouraged by IETF recommendations to obtain IP address space directly from their ISP. Networks served by multiple ISPs, on the other hand, may often obtain independent CIDR blocks directly from the appropriate RIR.



For example, in the late 1990s, the IP address 208.130.29.33 (since reassigned) was used by `www.freesoft.org`. An analysis of this address identified three CIDR prefixes.

208.128.0.0/11, a large CIDR block containing over 2 million addresses, had been assigned by ARIN (the North American RIR) to MCI. Automation Research Systems, a Virginia VAR, leased an Internet connection from MCI and was assigned the 208.130.28.0/22 block, capable of addressing just over 1000 devices. ARS used a /24 block for its publicly accessible servers, of which 208.130.29.33 was one.

All of these CIDR prefixes would be used at different locations in the network. Outside of MCI's network, the 208.128.0.0/11 prefix would be used to direct to MCI traffic bound not only for 208.130.29.33, but also for any of the roughly two million IP addresses with the same initial 11 bits. Within MCI's network, 208.130.28.0/22 would become visible, directing traffic to the leased line serving ARS. Only within the ARS corporate network would the 208.130.29.0/24 prefix have been used.

Subnet masks

A subnet mask is a bitmask that encodes the prefix length in quad-dotted notation: 32 bits, starting with a number of 1 bits equal to the prefix length, ending with 0 bits, and encoded in four-part dotted-decimal format. A subnet mask encodes the same information as a prefix length, but predates the advent of CIDR. However, in CIDR notation, the prefix bits are always contiguous, whereas subnet masks may specify non-contiguous bits. However, this has no practical advantage for increasing efficiency.

Prefix aggregation

CIDR provides the possibility of fine-grained *routing prefix aggregation*, also known as *supernetting* or *route summarization*. For example, sixteen contiguous /24 networks can be aggregated and advertised to a larger network as a single /20 route, if the first 20 bits of their network addresses match. Two aligned contiguous /20s may then be aggregated to a /19, and so forth. This allows a significant reduction in the number of routes that have to be advertised.

IPv4 CIDR

IP/CIDR	Δ to last IP addr	Mask	Hosts (*)	Class	Notes
a.b.c.d/**32**	+0.0.0.0	255.255.255.255	1	1/256 C	
a.b.c.d/**31**	+0.0.0.1	255.255.255.254	2	1/128 C	d = 0 ... (2n) ... 254
a.b.c.d/**30**	+0.0.0.3	255.255.255.252	4	1/64 C	d = 0 ... (4n) ... 252
a.b.c.d/**29**	+0.0.0.7	255.255.255.248	8	1/32 C	d = 0 ... (8n) ... 248
a.b.c.d/**28**	+0.0.0.15	255.255.255.240	16	1/16 C	d = 0 ... (16n) ... 240
a.b.c.d/**27**	+0.0.0.31	255.255.255.224	32	1/8 C	d = 0 ... (32n) ... 224

a.b.c.d/**26**	+0.0.0.63	255.255.255.192	64	1/4 C	d = 0, 64, 128, 192
a.b.c.d/**25**	+0.0.0.127	255.255.255.128	128	1/2 C	d = 0, 128
a.b.c.0/**24**	+0.0.0.255	255.255.255.000	256	1 C	
a.b.c.0/**23**	+0.0.1.255	255.255.254.000	512	2 C	c = 0 ... (2n) ... 254
a.b.c.0/**22**	+0.0.3.255	255.255.252.000	1,024	4 C	c = 0 ... (4n) ... 252
a.b.c.0/**21**	+0.0.7.255	255.255.248.000	2,048	8 C	c = 0 ... (8n) ... 248
a.b.c.0/**20**	+0.0.15.255	255.255.240.000	4,096	16 C	c = 0 ... (16n) ... 240
a.b.c.0/**19**	+0.0.31.255	255.255.224.000	8,192	32 C	c = 0 ... (32n) ... 224
a.b.c.0/**18**	+0.0.63.255	255.255.192.000	16,384	64 C	c = 0, 64, 128, 192
a.b.c.0/**17**	+0.0.127.255	255.255.128.000	32,768	128 C	c = 0, 128
a.b.0.0/**16**	+0.0.255.255	255.255.000.000	65,536	256 C = 1 B	
a.b.0.0/**15**	+0.1.255.255	255.254.000.000	131,072	2 B	b = 0 ... (2n) ... 254
a.b.0.0/**14**	+0.3.255.255	255.252.000.000	262,144	4 B	b = 0 ... (4n) ... 252
a.b.0.0/**13**	+0.7.255.255	255.248.000.000	524,288	8 B	b = 0 ... (8n) ... 248
a.b.0.0/**12**	+0.15.255.255	255.240.000.000	1,048,576	16 B	b = 0 ... (16n) ... 240
a.b.0.0/**11**	+0.31.255.255	255.224.000.000	2,097,152	32 B	b = 0 ... (32n) ... 224
a.b.0.0/**10**	+0.63.255.255	255.192.000.000	4,194,304	64 B	b = 0, 64, 128, 192
a.b.0.0/**9**	+0.127.255.255	255.128.000.000	8,388,608	128 B	b = 0, 128
a.0.0.0/**8**	+0.255.255.255	255.000.000.000	16,777,216	256 B = 1 A	
a.0.0.0/**7**	+1.255.255.255	254.000.000.000	33,554,432	2 A	a = 0 ... (2n) ... 254
a.0.0.0/**6**	+3.255.255.255	252.000.000.000	67,108,864	4 A	a = 0 ... (4n) ... 252
a.0.0.0/**5**	+7.255.255.255	248.000.000.000	134,217,728	8 A	a = 0 ... (8n) ... 248
a.0.0.0/**4**	+15.255.255.255	240.000.000.000	268,435,456	16 A	a = 0 ... (16n) ...

						240
a.0.0.0/**3**	+31.255.255.255	224.000.000.000	536,870,912	32	A	a = 0 ... (32n) ... 224
a.0.0.0/**2**	+63.255.255.255	192.000.000.000	1,073,741,824	64	A	a = 0, 64, 128, 192
a.0.0.0/**1**	+127.255.255.255	128.000.000.000	2,147,483,648	128	A	a = 0, 128
0.0.0.0/**0**	+255.255.255.255	000.000.000.000	4,294,967,296	256	A	

(*) For routed subnets bigger than /31 or /32, two reserved addresses need to be subtracted from the number of available host addresses: the largest address, which is used as the broadcast address, and the smallest address, which is used to identify the network itself. It is also common for the IP gateway for that subnet to use an address, meaning that you would subtract three from the number of hosts that can be used on the subnet.

Chapter 4

Differentiated Services

Differentiated Services or **DiffServ** is a computer networking architecture that specifies a simple, scalable and coarse-grained mechanism for classifying, managing network traffic and providing Quality of Service (**QoS**) guarantees on modern IP networks. DiffServ can, for example, be used to provide low-latency to critical network traffic such as voice or video while providing simple best-effort traffic guarantees to non-critical services such as web traffic or file transfers.

DiffServ uses the 6-bit **Differentiated Services Code Point** (**DSCP**) field in the header of IP packets for packet classification purposes. DSCP replaces the outdated IP precedence, a 3-bit field in the Type of Service byte of the IP header originally used to classify and prioritize types of traffic.

Background

Since modern data networks carry many different types of services, including voice, video, streaming music, web pages and email, many of the proposed QoS mechanisms that allowed these services to co-exist were both complex and failed to scale to meet the demands of the public Internet. In December 1998, the IETF published RFC 2474 (An Architecture for Differentiated Services), which replaced the ToS field with the DiffServ field. In the DiffServ field, a range of eight values (class selector) is used for backward compatibility with IP precedence. Today, DiffServ has largely supplanted other Layer 3 QoS mechanisms (such as IntServ) as the primary protocol routers use to provide different levels of service.

Traffic management mechanisms

DiffServ is a *coarse-grained*, **class-based** mechanism for traffic management. In contrast, IntServ is a *fine-grained*, **flow-based** mechanism.

DiffServ operates on the principle of *traffic classification*, where each data packet is placed into a limited number of traffic classes, rather than differentiating network traffic based on the requirements of an individual flow. Each router on the network is configured to differentiate traffic based on its class. Each traffic class can be managed differently, ensuring preferential treatment for higher-priority traffic on the network.

The DiffServ model does not incorporate premade judgements of what types of traffic should be given priority treatment; that is left up to the network operator. DiffServ simply provides a framework to allow classification and differentiated treatment. DiffServ does recommend a standardized set of traffic classes (discussed below) to make interoperability between different networks and different vendors' equipment simpler.

DiffServ relies on a mechanism to *classify* and *mark* packets as belonging to a specific class. DiffServ-aware routers implement *Per-Hop Behaviors* (PHBs), which define the packet forwarding properties associated with a class of traffic. Different PHBs may be defined to offer, for example, low-loss, low-latency forwarding properties or best-effort forwarding properties. All the traffic flowing through a router that belongs to the same class is referred to as a *Behavior Aggregate* (BA).

DiffServ domain

A group of routers that implement common, administratively defined DiffServ policies are referred to as a *DiffServ Domain*.

Classification and marking

Network traffic entering a DiffServ domain is subjected to classification and conditioning. Traffic may be classified by many different parameters, such as source address, destination address or traffic type and assigned to a specific traffic class. Traffic classifiers may honor any DiffServ markings in received packets or may elect to ignore or override those markings. Because network operators want tight control over volumes and type of traffic in a given class, it is very rare that the network honors markings at the ingress to the DiffServ domain. Traffic in each class may be further conditioned by subjecting the traffic to rate limiters, traffic policers or shapers.

Per-hop behavior

The Per-Hop Behavior (PHB) is determined by the differentiated services (DS) field of the IPv4 header or IPv6 header. The DS field was formerly used as the type of Service field. The DS field consists of a 6bit differentiated services code point (DSCP) RFC 2474. Explicit Congestion Notification occupies the least-significant 2 bits.

In theory, a network could have up to 64 (i.e. 2^6) different traffic classes using different markings in the DSCP. The DiffServ RFCs recommend, but do not require, certain encodings. This gives a network operator great flexibility in defining traffic classes. In practice, however, most networks use the following commonly-defined Per-Hop Behaviors:

- Default PHB—which is typically best-effort traffic
- *Expedited Forwarding* (EF) PHB—dedicated to low-loss, low-latency traffic
- *Assured Forwarding* (AF) PHB— which gives assurance of delivery under conditions

- *Class Selector* PHBs—which are defined to maintain backward compatibility with the IP Precedence field.

Default PHB

A default PHB is the only required behavior. Essentially, any traffic that does not meet the requirements of any of the other defined classes is placed in the default PHB. Typically, the default PHB has best-effort forwarding characteristics. The recommended DSCP for the default PHB is '000000' (in binary).

Expedited Forwarding (EF) PHB

The IETF defines Expedited Forwarding behavior in RFC 3246. The EF PHB has the characteristics of low delay, low loss and low jitter. These characteristics are suitable for voice, video and other realtime services. EF traffic is often given strict priority queuing above all other traffic classes. Because an overload of EF traffic will cause queuing delays and affect the jitter and delay tolerances within the class, EF traffic is often strictly controlled through admission control, policing and other mechanisms. Typical networks will limit EF traffic to no more than 30%—and often much less—of the capacity of a link. The recommended DSCP for expedited forwarding is 101110_B, or $2E_H$.

Assured Forwarding (AF) PHB group

The IETF defines the Assured Forwarding behavior in RFC 2597 and RFC 3260. Assured forwarding allows the operator to provide assurance of delivery as long as the traffic does not exceed some subscribed rate. Traffic that exceeds the subscription rate faces a higher probability of being dropped if congestion occurs.

The AF behavior group defines four separate AF classes. Within each class, packets are given a drop precedence (high, medium or low). The combination of classes and drop precedence yields twelve separate DSCP encodings from AF11 through AF43 (see table)

Assured Forwarding (AF) Behavior Group

	Class 1	**Class 2**	**Class 3**	**Class 4**
Low Drop	AF11 (DSCP 10)	AF21 (DSCP 18)	AF31 (DSCP 26)	AF41 (DSCP 34)
Med Drop	AF12 (DSCP 12)	AF22 (DSCP 20)	AF32 (DSCP 28)	AF42 (DSCP 36)
High Drop	AF13 (DSCP 14)	AF23 (DSCP 22)	AF33 (DSCP 30)	AF43 (DSCP 38)

Some measure of priority and proportional fairness is defined between traffic in different classes. Should congestion occur *between* classes, the traffic in the higher class is given priority. Rather than using strict priority queueing, more balanced queue servicing algorithms such as fair queueing or weighted fair queuing are likely to be used. If congestion occurs *within* a class, the packets with the higher drop precedence are discarded first. To prevent issues associated with tail drop, the random early detection

(RED) or weighted random early detection (WRED) algorithms are often used to drop packets.

Usually, traffic policing is required to encode drop precedence. Typically, all traffic assigned to a class is initially given a low drop precedence. As the traffic rate exceeds subscription thresholds, the policer will increase the drop precedence of packets that exceed the threshold.

Class selector PHB

Prior to DiffServ, IP networks could use the *Precedence* field in the Type of Service (TOS) byte of the IP header to mark priority traffic. The TOS byte and IP precedence was not widely used. The IETF agreed to reuse the TOS byte as the DS field for DiffServ networks. In order to maintain backward compatibility with network devices that still use the Precedence field, DiffServ defines the Class Selector PHB.

The Class Selector codepoints are of the form 'xxx000'. The first three bits are the IP precedence bits. Each IP precedence value can be mapped into a DiffServ class. If a packet is received from a non-DiffServ aware router that used IP precedence markings, the DiffServ router can still understand the encoding as a Class Selector codepoint.

Advantages of DiffServ

One advantage of DiffServ is that all the policing and classifying is done at the boundaries between DiffServ clouds. This means that in the core of the Internet, routers can get on with doing the job of routing, and not care about the complexities of collecting payment or enforcing agreements. That is, DiffServ requires no advance setup, no reservation, and no time-consuming end-to-end negotiation for each flow, as with integrated services. This leads DS to be relatively easy to implement.

IP differs from several legacy protocols such as SDH, PDH and ATM that have end to end service assurance. IP does not enforce its service level end-to-end. IP is not connection oriented and there is no end-to-end signaling in the network in order to let every device in the path know about a session and then set up requested priority or decline the session. Only packet marking takes place with preferred QoS or service description (DiffServ), no service level enforcement. This is extremely scalable since there is no need for common end to end methodology on how proper service levels are achieved. Quality enforcement can be implemented hop by hop and be adapted to underlying technology and challenges.

Disadvantages of DiffServ

End-to-end and peering problems

One disadvantage is that the details of how individual routers deal with the type of service field is somewhat arbitrary, and it is difficult to predict end-to-end behaviour.

This is complicated further if a packet crosses two or more DiffServ clouds before reaching its destination.

From a commercial viewpoint, this is a major flaw, as it means that it is impossible to sell different classes of end-to-end connectivity to end users, as one provider's Gold packet may be another's Bronze. Internet operators could fix this, by enforcing standardised policies across networks, but are not keen on adding new levels of complexity to their already complex peering agreements. One of the reasons for this is set out below.

Diffserv operation only works if the boundary hosts honour the policy agreed upon. However, this assumption is naive. A host can always tag its own traffic with a higher precedence, even though the traffic doesn't qualify to be handled with that importance. This in fact has already been exploited: Microsoft Windows 2000 always tags its traffic with IP precedence 5, making the traffic classing useless. On the other hand, the network is usually quite within its rights to traffic shape and otherwise ration the amount of network traffic ingress with any particular precedence, and so where this is enforced, overall network traffic flow provided to a host could be reduced by such a tactic.

DiffServ or any other IP based QoS marking does not ensure quality of the service or a specified service level (SLA). By marking the packets the sender wants the packets to be treated as a specific service, but it can only hope that this happens. It is up to all the service providers and their routers in the path to ensure that their policies will take care of the packets in an appropriate fashion.

DiffServ vs. more capacity

Some people believe that the problem addressed by DiffServ should not exist, and instead the capacity of Internet links should be chosen large enough to prevent packet loss altogether.

The logic is as follows. Since DiffServ is simply a mechanism for deciding to deliver or route at the expense of others in a situation where there is not enough network capacity, consider that when DiffServ is working by dropping packets selectively, traffic on the link in question must already be very close to saturation. Any further increase in traffic will result in Bronze services being taken out altogether. This will happen on a regular basis if the *average* traffic on a link is near the limit at which DiffServ becomes needed.

For a few years after the tech wreck of 2001, there was a glut of fibre capacity in most parts of the telecoms market, with it being far easier and cheaper to add more capacity than to employ elaborate DiffServ policies as a way of increasing customer satisfaction. This is what is generally done in the core of the Internet, which is generally fast and dumb with "fat pipes" connecting its routers.

However, this logic is flawed in many respects:

First, the problem of Bronze traffic being starved can be avoided if the network is provisioned to provide a minimum Bronze bandwidth, by limiting the maximum amount of higher priority traffic admitted.

Simple over-provisioning is an inefficient solution, since Internet traffic is highly bursty. If the network is dimensioned to carry all traffic at such times, then it will cost an order of magnitude more than a network dimensioned to carry typical traffic, with traffic management used to prevent collapse during such peaks.

It is not even possible to dimension for "peak load". In particular, when sending a large file, the TCP protocol continues to request more bandwidth as the loss rate decreases, and so it is simply not possible to dimension links to avoid end-to-end loss altogether: increasing the capacity of one link eventually causes loss to occur on a different link.

Finally, with wireless links such as EV-DO, where the air-interface bandwidth is several orders of magnitude less than the backhaul, QoS is being used to efficiently deliver VoIP packets where it would not otherwise be achievable.

When discussing DiffServ vs. More Capacity it is important to look beyond the black and white scenarios with either too little capacity on one side and enough capacity on the other side. In the real life and in the real networks it is more differentiated than that.

Let us first look at the dark side with too low capacity. In a situation with too low capacity something has to suffer. How do we as users experience a situation where packets are lost due to congestion in the network? What type of services and applications are more likely to decrease quality in such way that they are useless? Is it voice or video, or is it e-mail? I guess I can wait a few minutes extra for the e-mail, but I can't view a video with packet loss, and I cannot understand voice if it is chopped in to bits and pieces. So what do we do if we don't have enough capacity? Use codecs that ramp up the bit rate and fill the networks with redundant information and thereby fill up the already congested network? Or should we look at the services and prioritize the packets that need to be delivered in order to provide a service that is usable for the end user?

In the gray zone we have the almost ultimate network with fat pipes and where such terminology as congestion and queuing are unheard of. Is it possible to build such a network all over the world with fat pipes in to every connected host on the planet or even in the universe? I don't think so. What happens if some of the fat pipes fail and everything is routed via a bottleneck? Should the "bottleneck" be wider than the bottle to remedy this challenge? Should we use DiffServ to ensure that the end users or systems get sufficient service quality?

So at last the ultimate networks with fat pipes where congestion never occurs. Such network cannot exist. Even with little or no traffic on the fat pipes we will have small congestions on the interfaces. Let's imagine a router with the three interfaces A, B, and C. If two packets arrive at interface A and B at the same time that are destined for a network reachable via interface C then we might have a queuing issue. First come first

served with FIFO buffering is easy and might seem “fair”. If these two packets arrives at almost the same time or even at the perfect same time it might not be fair to serve first come first. Propagation delay is an issue if we want to keep jitter and latency down for certain applications. We simply cannot afford to let some services struggle with jitter and latency and it is fair that interactive voice and video is prioritized. What if we start using Internet and IP for medical applications etc? Is it fair that my BitTorrent have the same priority than a video session for a surgery specialist that is supervising a remote operation? My BitTorrent can handle jitter.

To set the focus on “Diffserv vs. More Capacity” vs. “DiffServ and More Capacity” is the first challenge.

Effects of dropped packets

Dropping packets wastes the resources that have already been expended in carrying these packets so far through the network. In many cases, this traffic will be re-transmitted, causing further bandwidth consumption at the congestion point and elsewhere in the network. To minimize this waste, packets must be discarded as close to the edge of the network as possible, while Diffserv is often implemented throughout a network (edge and core).

Thus, dropping packets amounts to betting that congestion will have resolved by the time the packets are re-sent, or that (if the dropped packets are TCP datagrams) TCP will throttle back transmission rates at the sources to reduce congestion in the network. The TCP congestion avoidance algorithms are subject to a phenomenon called TCP global synchronization unless special approaches (such as Random early detection) are taken when dropping TCP packets. In Global Synchronization, all TCP streams tend to build up their transmission rates together, reach the peak throughput of the network, and all crash together to a lower rate as packets are dropped, only to repeat the process.

Delays caused by re-scheduling packets due to Diffserv can cause packets to drop by the IPsec anti-replay mechanism.

DiffServ as rationing

Hence, DiffServ is for most ISPs mainly a way of rationing customer network utilisation to allow greater overbooking of their capacity. A good example of this is the use of DiffServ tools to suppress or control peer-to-peer traffic, because of its ability to saturate customer links indefinitely, disrupting the ISP's business model which relies on 1%-10% link utilization for most online customers.

Bandwidth broker

RFC 2638 from IETF defines the entity of the Bandwidth Broker in the framework of DiffServ. According to RFC 2638, a Bandwidth Broker is an agent that has some knowledge of an organization's priorities and policies and allocates bandwidth with

respect to those policies. In order to achieve an end-to-end allocation of resources across separate domains, the Bandwidth Broker managing a domain will have to communicate with its adjacent peers, which allows end-to-end services to be constructed out of purely bilateral agreements. Bandwidth Brokers can be configured with organizational policies, keep track of the current allocation of marked traffic, and interpret new requests to mark traffic in light of the policies and current allocation. Bandwidth Brokers only need to establish relationships of limited trust with their peers in adjacent domains, unlike schemes that require the setting of flow specifications in routers throughout an end-to-end path. In practical technical terms, the Bandwidth Broker architecture makes it possible to keep state on an administrative domain basis, rather than at every router in the same way as the DiffServ architecture makes it possible to confine per flow state to just the leaf routers.

- Manages each cloud's resources (Bandwidth Broker)
- Packets are "coloured" to indicate forwarding "behavior"
- Focus on aggregates and NOT on individual flows
- Policing at network periphery to get services
- Used together with Multiprotocol Label Switching (MPLS) and Traffic Engineering (TE)
- "Aggregated" QoS guarantees only!
- Poor on the guarantees for end-to-end applications

DiffServ RFCs

- RFC 2474—Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers
- RFC 2475—An Architecture for Differentiated Services
- RFC 2597—Assured Forwarding PHB Group
- RFC 3140—Per Hop Behavior Identification Codes (Obsoletes RFC 2836)
- RFC 3246—An Expedited Forwarding PHB (Obsoletes RFC 2598)
- RFC 3260—New Terminology and Clarifications for Diffserv
- RFC 4594—Configuration Guidelines for DiffServ Service Classes

Chapter 5

End-to-end Principle and Forwarding Plane

End-to-end principle

The **end-to-end principle** is one of the central design principles of the Internet and is implemented in the design of the underlying methods and protocols in the Internet Protocol Suite. It is also used in other distributed systems. The principle states that, whenever possible, communications protocol operations should be defined to occur at the end-points of a communications system, or as close as possible to the resource being controlled.

According to the end-to-end principle, protocol features are only justified in the lower layers of a system if they are a performance optimization, hence, Transmission Control Protocol (TCP) retransmission for reliability is still justified, but efforts to improve TCP reliability should stop after peak performance has been reached.

History

The concept and research of end-to-end connectivity and network intelligence at the end-nodes reaches back to packet-switching networks in the 1970s, cf. CYCLADES. A 1981 presentation entitled *End-to-end arguments in system design* by Jerome H. Saltzer, David P. Reed, and David D. Clark, argued that reliable systems tend to require end-to-end processing to operate correctly, in addition to any processing in the intermediate system. They pointed out that most features in the lowest level of a communications system have costs for all higher-layer clients, even if those clients do not need the features, and are redundant if the clients have to reimplement the features on an end-to-end basis.

This leads to the model of a *dumb, minimal network* with smart terminals, a completely different model from the previous paradigm of the smart network with dumb terminals. However, the End-to-end principle was always meant to be a pragmatic engineering philosophy for network system design that merely prefers putting intelligence towards the end points. It does not forbid intelligence in the network itself if it makes more practical sense to put certain intelligence in the network rather than the end-points. David D. Clark

along with Marjory S. Blumenthal wrote in 2001 in *Rethinking the design of the Internet: The end to end arguments vs. the brave new world*:

from the beginning, the end to end arguments revolved around requirements that could be implemented correctly at the end-points; if implementation inside the network is the only way to accomplish the requirement, then an end to end argument isn't appropriate in the first place.

Indeed, as noted in RFC 1958 edited by Brian Carpenter in June 1996, entitled "Architectural Principles of the Internet," "[i]n searching for Internet architectural principles, we must remember that technical change is continuous in the information technology industry. The Internet reflects this. . . .In this environment, some architectural principles inevitably change. Principles that seemed inviolable a few years ago are deprecated tomorrow. The principle of constant change is perhaps the only principle of the Internet that should survive indefinitely." This is particularly true with respect to the so-called "end-to-end" principle.

As noted by Bob Kahn, co-inventor of the Internet Protocol:

The original Internet involved three individual networks, namely the ARPANET, the Packet Radio network and the Packet Satellite network, all three of which had been developed with DARPA support. One early consideration that was rejected was to change each of these networks to be able to interpret and route internet packets so that there would be no need for external devices to route the traffic. However, this would have required major changes to all three networks and would have required synchronized changes in all three to accommodate protocol evolutions. Instead, it was decided to create what were called "gateways," the forerunner of today's routers, to handle the IP protocol-based networks. Reliable packet communication was handled by a combination of factors, but, ultimately, the TCP protocol provided an end-to-end means of reassembly of packet fragments, error checking and acknowledgment back to the source. The resulting fact that no changes were needed in the individual networks was interpreted by some as implying that the Internet design assumed only dumb networks with all the smarts being at the boundaries. Nothing could have been further from the truth. The initial choice of using gateways/routers was purely pragmatic and should imply nothing about how the Internet might operate in the future.

In 1995, the Federal Networking Council adopted a resolution defining the Internet as a "global information system" that is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons; is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and provides, uses or makes accessible, either publicly or privately, high level services layered on this communications and related infrastructure .

In comments submitted by Patrice Lyons to the United Nations Working Group on Internet Governance (November 4, 2004), entitled "The End-End Principle and the

Definition of Internet," on behalf of Bob Kahn's non profit research organization, Corporation for National Research Initiatives (CNRI), it was noted that:

To argue today that the only stateful elements that may be active in the Internet environment should be located at the edges of the Internet is to ignore the evolution of software and other technologies to provide a host of services throughout the Internet. The layering approach has many advantages and should be retained along with more integrated system architectures; the approach was a practical way of overlaying the Internet architecture over existing networks when it was difficult to coordinate the modification of these networks, if indeed such modifications could have been agreed upon and implemented. For some newer applications, maintaining state information within the network may now be desirable for efficiency if not overall performance effectiveness. In addition, current research efforts may need to draw upon innovative methods to increase security of communications, develop new forms of structuring data, create and deploy dynamic metadata repositories, or real-time authentication of the information itself.

Specifically, CNRI proposed that, in the third element of the FNC definition of Internet, after the phrase "high level services layered on," it is advisable to add the following words: "or integrated with," and observed that this point is "directly relevant to the ongoing discussions about the so-called 'end-to-end' principle that is often viewed as essential to an understanding of the Internet." Further, while the end-to-end principle may have been relevant in the environment where the Internet originated, it has not been critical for a number of years going back "at least to the early work on mobile programs, distributed searching, and certain aspects of collaborative computing."

Examples

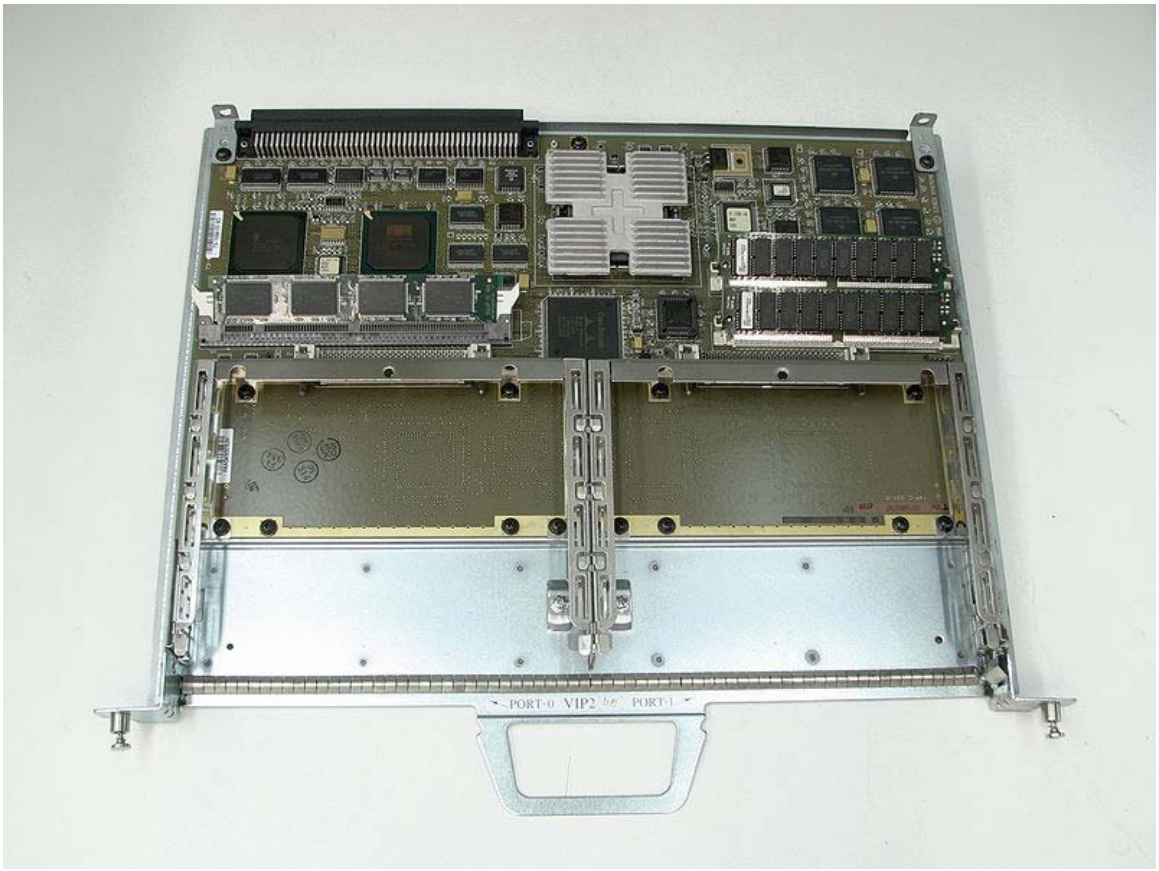
In the Internet Protocol Suite, the Internet Protocol is a simple ("dumb"), stateless protocol that moves datagrams across the network, and TCP is a smart transport protocol providing error detection, retransmission, congestion control, and flow control end-to-end. The network itself (the routers) needs only to support the simple, lightweight IP; the endpoints run the heavier TCP on top of it when needed.

A second canonical example is that of file transfer. Every reliable file transfer protocol and file transfer program should contain a checksum, which is validated only after everything has been successfully stored on disk. Disk errors, router errors, and file transfer software errors make an end-to-end checksum necessary. Therefore, there is a limit to how secure TCP checksum should be, because it has to be reimplemented for any robust end-to-end application to be secure.

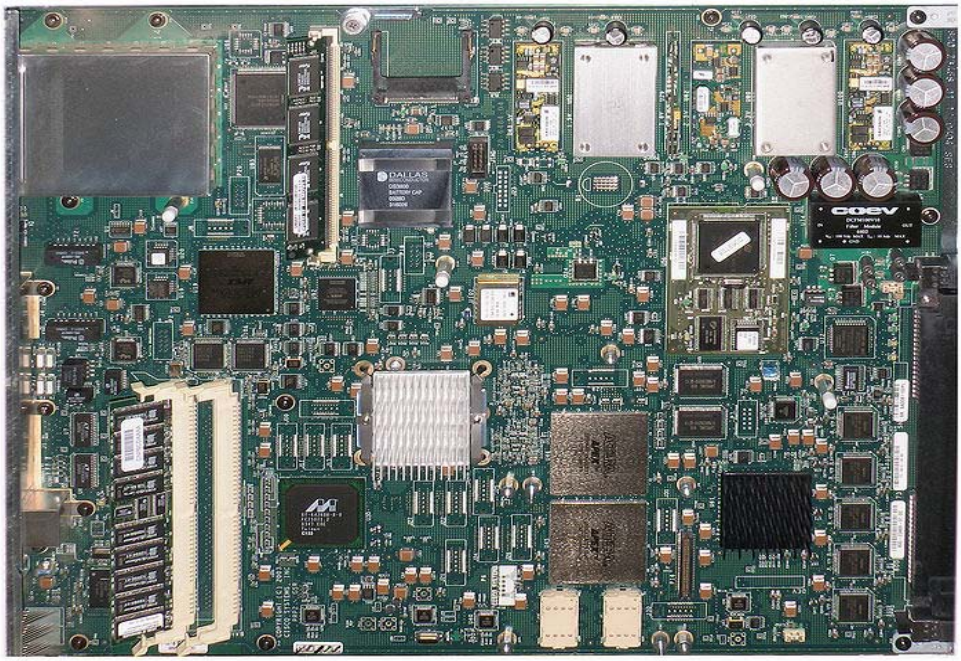
A third example (not from the original paper) is the EtherType field of Ethernet. An Ethernet frame does not attempt to provide interpretation for the 16 bits of type in an original Ethernet packet. To add special interpretation to some of these bits would reduce the total number of Ethertypes, hurting the scalability of higher layer protocols, i.e. all higher layer protocols would pay a price for the benefit of just a few. Attempts to add

elaborate interpretation (e.g. IEEE 802 SSAP/DSAP) have generally been ignored by most network designs, which follow the end-to-end principle.

Forwarding plane



Cisco VIP 2-40, from an older generation of routers



Performance Route Processor, from the high-end Cisco 12000 series

In routing, the **forwarding plane**, sometimes called the **data plane**, defines the part of the router architecture that decides what to do with packets arriving on an inbound interface. Most commonly, it refers to a table in which the router looks up the destination address of the incoming packet and retrieves the information necessary to determine the path from the receiving element, through the internal **forwarding fabric** of the router, and to the proper outgoing interface(s). The IP Multimedia Subsystem architecture uses the term **transport plane** to describe a function roughly equivalent to the routing control plane.

The table also might specify that the packet is discarded. In some cases, the router will return an ICMP "destination unreachable" or other appropriate code. Some security policies, however, dictate that the router should be programmed to drop the packet silently. By dropping filtered packets silently, a potential attacker does not become aware of a target that is being protected.

The incoming forwarding element will also decrement the time-to-live (TTL) field of the packet, and, if the new value is zero, discard the packet. While the Internet Protocol (IP) specification indicates that an Internet Control Message Protocol (ICMP) "TTL exceeded" message should be sent to the originator of the packet (i.e., the node with the source address in the packet), routers may be programmed to drop the packet silently.

Depending on the specific router implementation, the table in which the destination address is looked up could be the routing table (also known as the routing information

base, RIB), or a separate forwarding information base (FIB) that is populated (i.e., loaded) by the routing control plane, but used by the forwarding plane to look up packets, at very high speed, and decide how to handle them. Before or after examining the destination, other tables may be consulted to make decisions to drop the packet based on other characteristics, such as the source address, the IP protocol identifier field, or Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) port number.

Forwarding plane functions run in the forwarding element. High-performance routers often have multiple distributed forwarding elements, so that the router increases performance with parallel processing.

The outgoing interface will encapsulate the packet in the appropriate data link protocol. Depending on the router software and its configuration, functions, usually implemented at the outgoing interface, may set various packet fields, such as the DSCP field used by differentiated services.

In general, the passage from the input interface directly to an output interface, through the fabric with minimum modification at the output interface, is called the *fast path* of the router. If the packet needs significant processing, such as segmentation or encryption, it may go onto a slower path, which is sometimes called the *services plane* of the router. Service planes can make forwarding or processing decisions based on higher-layer information, such as a Web URL contained in the packet payload.

Issues in router forwarding performance

Vendors design router products for specific markets. Design of routers intended for home use, perhaps supporting several PCs and VoIP telephony, is driven by keeping the cost as low as possible. In such a router, there is no separate forwarding fabric, and there is only one active forwarding path: into the main processor and out of the main processor.

Routers for more demanding applications accept greater cost and complexity to get higher throughput in their forwarding planes.

Several design factors affect router forwarding performance:

- Data link layer processing and extracting the packet
- Decoding the packet header
- Looking up the destination address in the packet header
- Analyzing other fields in the packet
- Sending the packet through the "fabric" interconnecting the ingress and egress interfaces
- Processing and data link encapsulation at the egress interface

Routers may have one or more processors. In a uniprocessor design, these performance parameters are affected not just by the processor speed, but by competition for the processor. Higher-performance routers invariably have multiple processing elements,

which may be general-purpose processor chips or specialized application-specific integrated circuits (ASIC).

Very high performance products have multiple processing elements on each interface card. In such designs, the main processor does not participate in forwarding, but only in control plane and management processing.

Benchmarking performance

In the Internet Engineering Task Force, two working groups in the Operations & Maintenance Area deal with aspects of performance. The Interprovider Performance Measurement (IPPM) group focuses, as its name would suggest, on operational measurement of services. Performance measurements on single routers, or narrowly defined systems of routers, are the province of the Benchmarking Working Group (BMWG).

RFC 2544 is the key BMWG document. A classic RFC 2544 benchmark uses half the router's (i.e., the device under test (DUT)) ports for input of a defined load, and measures the time at which the outputs appear at the output ports.

Forwarding information base design

Originally, all destinations were looked up in the RIB. Perhaps the first step in speeding routers was to have a separate RIB and FIB in main memory, with the FIB, typically with fewer entries than the RIB, being organized for fast destination lookup. In contrast, the RIB was optimized for efficient updating by routing protocols.

Early uniprocessing routers usually organized the FIB as a hash table, while the RIB might be a linked list. Depending on the implementation, the FIB might have fewer entries than the RIB, or the same number.

When routers started to have separate forwarding processors, these processors usually had far less memory than the main processor, such that the forwarding processor could hold only the most frequently used routes. On the early Cisco AGS+ and 7000, for example, the forwarding processor cache could hold approximately 1000 route entries. In an enterprise, this would often work quite well, because there were fewer than 1000 server or other popular destination subnets. Such a cache, however, was far too small for general Internet routing. Different router designs behaved in different ways when a destination was not in the cache.

Cache miss issues

A **cache miss** condition might result in the packet being sent back to the main processor, to be looked up in a **slow path** that had access to the full routing table. Depending on the router design, a cache miss might cause an update to the fast hardware cache or the fast cache in main memory. In some designs, it was most efficient to invalidate the fast cache

for a cache miss, send the packet that caused the cache miss through the main processor, and then repopulate the cache with a new table that included the destination that caused the miss. This approach is similar to an operating system with virtual memory, which keeps the most recently used information in physical memory.

As memory costs went down and performance needs went up, FIBs emerged that had the same number of route entries as in the RIB, but arranged for fast lookup rather than fast update. Whenever a RIB entry changed, the router changed the corresponding FIB entry.

FIB design alternatives

High-performance FIBs achieve their speed with implementation-specific combinations of specialized algorithms and hardware.

Software

Various search algorithms have been used for FIB lookup. While well-known general-purpose data structures were first used, such as hash tables, specialized algorithms, optimized for IP addresses, emerged. They include:

- Binary tree
- Radix tree
- Four-way trie
- Patricia tree

Since 2006, multicore CPU is changing the design of Fast Path thanks to innovative processors such as the Cavium's Octeon, RMI's XLS/XLR/XLP, Freescale's QorIQ or even using Intel's multicore processors. Now, it is based on dedicated data path software into the cores of those CPUs. The coding rules of software have changed. The system has to be made of a dedicated packet engine stack on each core, which cannot be a regular OS stack; each instance is spread into the cluster of cores in order to be the Fast Path. Into this Fast Path, since it runs in a highly parallel system, in order to achieve the highest performance of a data plane, only lock free algorithms are allowed.

Hardware

Various forms of fast RAM and, eventually, basic content addressable memory (CAM) were used to speed lookup. CAM, while useful in layer 2 switches that needed to look up a relatively small number of fixed-length MAC addresses, had limited utility with IP addresses having variable-length routing prefixes. Ternary CAM (CAM), while expensive, lends itself to variable-length prefix lookups.

One of the challenges of forwarder lookup design is to minimize the amount of specialized memory needed, and, increasingly, to minimize the power consumed by memory.

Distributed forwarding

A next step in speeding routers was to have a specialized forwarding processor separate from the main processor. There was still a single path, but forwarding no longer had to compete with control in a single processor. The fast routing processor typically had a small FIB, with hardware memory (e.g., static random access memory (SRAM)) faster and more expensive than the FIB in main memory. Main memory was generally dynamic random access memory (DRAM).

Early distributed forwarding

Next, routers began to have multiple forwarding elements, that communicated through a high-speed **shared bus** or through a **shared memory**. Cisco used shared busses until they saturated, while Juniper preferred shared memory.

Each forwarding element had its own FIB. See, for example, the Versatile Interface Processor on the Cisco 7500

Eventually, the shared resource became a bottleneck, with the limit of shared bus speed being roughly 2 million packets per second (Mpps). Crossbar fabrics broke through this bottleneck.

Shared paths become bottlenecks

As forwarding bandwidth increased, even with the elimination of cache miss overhead, the shared paths limited throughput. While a router might have 16 forwarding engines, if there was a single bus, only one packet transfer at a time was possible. There were some special cases where a forwarding engine might find that the output interface was one of the logical or physical interfaces present on the forwarder card, such that the packet flow was totally inside the forwarder. It was often easier, however, even in this special case, to send the packet out the bus and receive it from the bus.

While some designs experimented with multiple shared buses, the eventual approach was to adapt the crossbar switch model from telephone switches, in which every forwarding engine had a hardware path to every other forwarding engine. With a small number of forwarding engines, crossbar forwarding fabrics are practical and efficient for high-performance routing. There are multistage designs for crossbar systems, such as Clos networks.

Chapter 6

IPv4 Address Exhaustion and Locator/Identifier Separation Protocol

IPv4 address exhaustion

IPv4 address exhaustion is the decreasing supply of unallocated Internet Protocol Version 4 (IPv4) addresses available at the Internet Assigned Numbers Authority (IANA) and the regional Internet registries (RIRs) for assignment to end users and local Internet registries, such as Internet service providers. IPv4 provides for approximately 4 billion addresses, and with the current allocation granularity of /8 blocks, each approximately 16.8 million addresses, IANA's primary address pool exhaustion is estimated to be reached by early February 2011. As of November 30, 2010, only seven of the 255 allocation blocks remain available, or less than 3% of the IPv4 address space.

The depletion of the IPv4 allocation pool has been a concern since the late 1980s when the Internet started to experience dramatic growth. The Internet Engineering Task Force (IETF) created the Routing and Addressing Group (ROAD) in November 1991 to respond to the scalability problem caused by the classful network allocation system in place at the time.

The anticipated shortage has been the driving factor in creating and adopting several new technologies, including classful networks in the 1980s, Classless Inter-Domain Routing (CIDR) methods in 1993, network address translation (NAT) and a new version of the Internet Protocol, IPv6, in 1998.

The transition of the Internet to IPv6 is the only practical and readily available long-term solution to IPv4 address exhaustion. Although the predicted IPv4 address exhaustion was approaching its final stages, most providers of Internet services and software vendors were just beginning IPv6 deployment in 2008.

IP addressing

Every host on an IP network, such as a computer or networked printer, is assigned an IP address that is used to communicate with other hosts on the same network or globally. Internet Protocol version 4 provides 2^{32} (approximately 4.3 billion) addresses. However,

large blocks of IPv4 addresses are reserved for special uses and are unavailable for public allocation.

The IPv4 addressing structure provides an insufficient number of publicly routable addresses to provide a distinct address to every Internet device or service. This problem has been mitigated for some time by changes in the address allocation and routing infrastructure of the Internet. Classful networking and particularly Classless Inter-Domain Routing delayed the exhaustion of addresses substantially.

In addition, network address translation permitted large Internet service providers to allocate only one public IP address to each of their customers, by masquerading the customer network behind this address with specially configured customer-premise Internet routers.

Address depletion

While the primary reason for IPv4 address exhaustion is insufficient design capacity of the original Internet infrastructure, several additional driving factors have aggravated the shortcomings. Each of them increased the demand on the limited supply of addresses, often in ways unanticipated by the original designers of the network.

Mobile devices
: As IPv4 increasingly became the *de facto* standard for networked digital communication, the cost of embedding substantial computing power into handheld devices dropped. Mobile phones have become viable Internet hosts. New specifications of 4G devices require IPv6 addressing.

Always-on connections
: Throughout the 1990s, the predominant mode of consumer Internet access was telephone modem dialup. The rapid growth of the dialup networks increased address consumption rates, although it was common that the modem pools, and as a result, the pool of assigned IP addresses, were shared amongst a larger customer base. By 2007, however, broadband Internet access had begun to exceed 50% penetration in many markets. Broadband connections are always active, as the gateway devices (routers, broadband modems) are rarely turned off, so that the address uptake by Internet service providers continued at an accelerating pace.

Internet demographics
: There are hundreds of millions of households in the developed world. In 1990, only a fraction of these had Internet connectivity. Just 15 years later, almost half of them had persistent broadband connections.

Inefficient address use
: Organizations that obtained IP addresses in the 1980s were often allocated far more addresses than they actually required, because the initial allocation method was inadequate to reflect reasonable usage. For example, large companies or universities were assigned class A address blocks with over 16 million IPv4 addresses each, because the next smaller allocation unit, a class B block with 65536 addresses, was too small for their intended deployments.

Many organizations continue to utilize public IP addresses for devices not accessible outside their local network. From a global address allocation viewpoint, this is inefficient in many cases, but scenarios exist where this is preferred in the organizational network implementation strategies.

Due to inefficiencies caused by subnetting, it is difficult to use all addresses in a block. The host-density ratio, as defined in RFC 3194, is a metric for utilization of IP address blocks, that is used in allocation policies.

Virtualization

With advances in hardware performance and processor features of server systems and the advent of sophisticated hardware abstraction layers it became possible to host many instantiations of an operating system on a single computer. Each of these systems may require a public IP address.

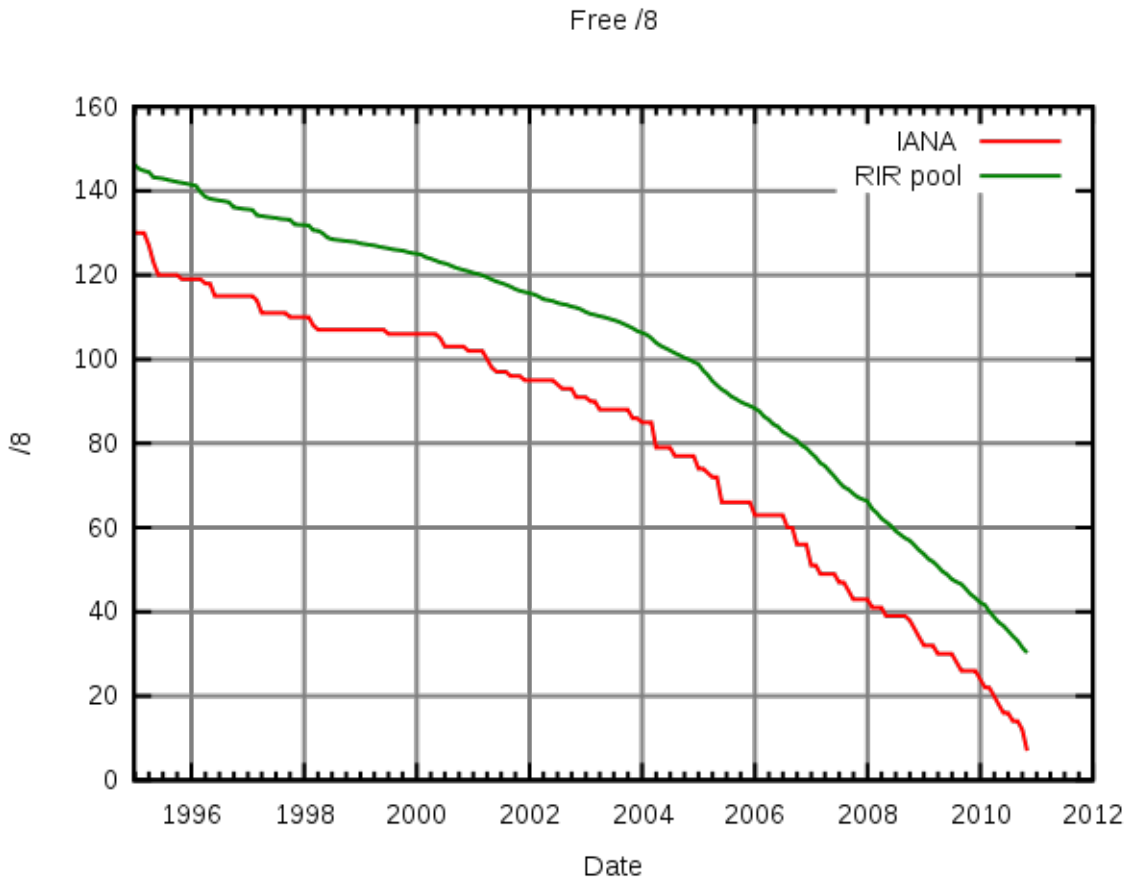
Mitigation efforts

Some methods of mitigation of IPv4 address exhaustion have been

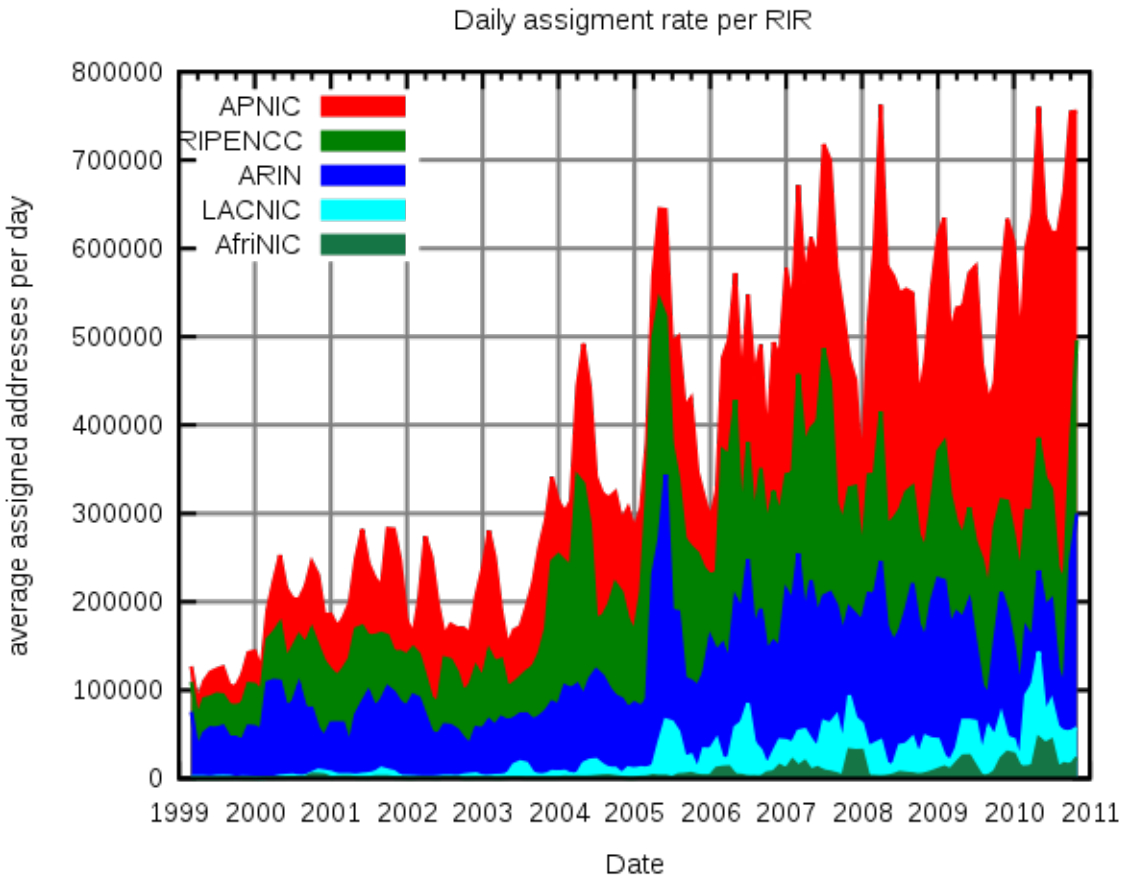
- Network address translation
- Use of private network addressing
- Name-based virtual hosting of web sites
- Tighter control by regional Internet registries on the allocation of addresses to local Internet registries
- Network renumbering and subnetting to reclaim large blocks of address space allocated in the early days of the Internet

Several organizations have returned large blocks of IP addresses. Notably, Stanford University relinquished their Class A IP block in 2000, making 16 million IP addresses available.

Exhaustion date



IPv4 adresses exhaustion since 1995



IPv4 addresses allocation rate per RIR

Estimates of the time of complete IPv4 address exhaustion varied widely in the early 2000s. In 2003, Paul Wilson (director of APNIC) stated that, based on then-current rates of deployment, the available space would last for one or two decades. In September 2005, a report by Cisco Systems suggested that the pool of available addresses would deplete in as little as 4 to 5 years.

The most prominent analysis of the exhaustion progress is published by Geoff Huston, Chief Scientist at APNIC. As of November 2010, his daily *IPv4 Address Report* predicts the exhaustion date of the unallocated IANA pool for February 2011. Huston's predictions are derived from past and current allocation trends and policies by IANA and the RIRs. Two other regularly updated sites, those of Tony Hain, and Stephan Lagerholm, were in approximate agreement with this estimate as of August 2010.

As of November 2010, approximately 3% or 7 of the 255 IANA assignment blocks (/8 CIDR blocks) remain unallocated.

Exhaustion will first occur at IANA, then at APNIC, and then at the other RIRs. Only specific organizations that requested addresses prior to the introduction of CIDR possibly

have significant unused address space remaining. After the IANA pool exhaustion, the studies indicate that each RIR may be able to supply from their last assigned addresses for another 8 months, when at least one of the RIRs is expected to be depleted. In the last year before exhaustion, IPv4 allocations are accelerating, resulting in exhaustion trending to earlier dates. By early 2012, new devices and services are expected to appear on the Internet that are only reachable by IPv6. These will only be accessible from the IPv4 Internet if older hosts that cannot implement IPv6 utilize special translator gateway services.

The time remaining until the first RIR exhaustion is a short time for the entire industry to transition to IPv6. This situation is aggravated by the likelihood that until exhaustion there will be no significant consumer demand for IPv6. David Conrad, the general manager of IANA, acknowledges: "*I suspect we are actually beyond a reasonable time frame where there won't be some disruption. Now it's more a question of how much.*" Geoff Huston claims the transition to IPv6 should have started much earlier, such that by the exhaustion date it would be *completed*, with all devices IPv6-capable, and IPv4 being phased out.

Notable exhaustion advisories

- On May 21, 2007, the American Registry for Internet Numbers (ARIN), the North American RIR, advised the Internet community that due to the expected exhaustion in 2010 "migration to IPv6 numbering resources is necessary for any applications which require ongoing availability from ARIN of contiguous IP numbering resources". "Applications" includes general connectivity between devices on the Internet, as some devices only have an IPv6 address allocated.
- On June 20, 2007, the Latin American and Caribbean Internet Addresses Registry (LACNIC), advised "preparing its regional networks for IPv6" by January 1, 2011, for the exhaustion of IPv4 addresses "in three years time".
- On June 26, 2007, the Asia-Pacific Network Information Centre (APNIC), the RIR for the Pacific and Asia, endorsed a statement by the Japan Network Information Center (JPNIC) that to continue the expansion and development of the Internet a move towards an IPv6-based Internet is advised. This with an eye on the expected exhaustion around 2010 which will create a great restriction on the Internet.
- On April 15, 2009, the American Registry for Internet Numbers (ARIN), the North American RIR, sent a letter to all CEO/Executives of companies who have IPv4 addresses allocated informing them that ARIN expects the IPv4 space will be depleted within the next two years.
- On 25 August 2009 ARIN announced a joint series event in the Caribbean region to push for the implementation of IPv6. ARIN reported at this time that less than 10.9% of IPv4 address space is remaining.
- Tony Hain of networking equipment manufacturer Cisco Systems predicts the exhaustion date of the unallocated IANA pool to be early in 2011 (updated monthly). His predictions use the same data source as Geoff Huston's, but the

trends are generated from different subsets, and account for the different distribution rules for the "last 5".

Reclamation of unused IPv4 space

Before and during the time when classful network design was still used as allocation model, large blocks of IP addresses were allocated to some organizations. The Internet Assigned Numbers Authority (IANA) could potentially reclaim these ranges and reissue the addresses in smaller blocks. ARIN, the North American Internet registry, has a transfer policy, such that addresses can get returned to ARIN, with the purpose to be reassigned to a specific recipient. However, it can be expensive in terms of cost and time to renumber a large network, so these organizations will likely object, with legal conflicts possible. However, even if all of these were reclaimed, it would only result in postponing the date of address exhaustion.

Similarly, IP address blocks have been allocated to entities that no longer exist or never used them. No strict accounting of IP address allocations has been undertaken, and it would take quite a bit of effort to track down which addresses really are unused, as many are only in use on intranets.

Some address space that was previously reserved by IANA has been added to the available pool. There have been proposals to use the class E network addresses, but many computer and router operating systems and firmware do not allow the use of these addresses. For this reason, the proposals have sought not to designate the class E space for public assignment, but instead propose to permit private use for networks that require more address space than is currently available through RFC 1918.

ISP-wide network address translation

When Internet service providers (ISPs) implement network address translation within their network, rather than at the demarcation to customer networks, they may allocate private addresses to customers and need only one global scope address for a potentially large group of customers. However, many customers must use the gateway for traffic to the Internet.

This has been successfully implemented in some countries like Russia, where many broadband providers now use Carrier Grade NAT, and offer publicly routable IP address at an additional cost. Similarly, Research In Motion (RIM), the maker of BlackBerry devices, currently routes all Blackberry data to central network operating centers for encryption and decryption purposes; this has the side effect of reducing the number public IP addresses necessary assigned.

However, ISP-wide NAT is not scalable, and limited to the number of ports available (approximately 65000) in the Transport Layer protocols. In addition, network address translation is not suitable for all applications.

Markets in IP addresses

The creation of markets to buy and sell IPv4 addresses has been proposed many times as an efficient means of allocation. The primary benefit of an address market would be that IPv4 addresses would continue to be available, although the market price of addresses would be expected to rise over time. These schemes have major drawbacks that have prevented their implementation:

- The creation of a market in IPv4 addresses would only delay the practical exhaustion of the IPv4 address space for a relatively short time, since the public Internet is still growing. This implies that absolute exhaustion of the IPv4 space would follow within at most a couple of years after the exhaustion of addresses for new allocations.
- The concept of legal "ownership" of IP addresses as property is explicitly denied by ARIN and RIPE policy documents and by the ARIN Registration Services Agreement. It is not even clear in which country's legal system the lawsuits would be resolved.
- The administration of such a scheme is outside the experience of the current regional address registries.
- Ad-hoc trading in addresses would lead to fragmented patterns of allocation that would vastly expand the global routing table, resulting in severe routing problems for many network operators which still use older routers with limited forwarding information base memory or low-powered routing processors. This large cost placed on everyone who uses the Internet by those that buy/sell IP addresses is a negative economic externality that any market would need to correct for.
- Trading in IP blocks that are large enough to prevent fragmentation problems would reduce the number of potentially tradeable units to a few million at most.
- The cost of changing from one set of IP addresses to another is very high, reducing the market liquidity. Organizations that can potentially reorganize their usage of IP addresses to free them up so that they can be sold will demand a high price and, once bought, will not be resold without a large profit. The cost of renumbering an organization's IP address space each time is comparable to the cost of switching to IPv6 once.

Long-term solution

The deployment of IPv6 is the only viable solution to the IPv4 address shortage. IPv6 is endorsed and implemented by all Internet technical standards bodies and network equipment vendors. It encompassed many design improvements, including the replacement of the 32-bit IPv4 address format, which allows 4.3 billion possible addresses, with a 128-bit address for a theoretical capacity of 3.4×10^{38} addresses. IPv6 has been in active production deployment since June 2006, when organized worldwide efforts of testing and evaluation ceased (6bone).

Locator/Identifier Separation Protocol

LISP is a "map-and-encapsulate" protocol which is currently developed by the Internet Engineering Task Force LISP Working Group. The basic idea behind the separation is that the Internet architecture combines two functions, routing locators (where you are attached to the network) and identifiers (who you are) in one number space: the IP address. LISP supports the separation of the IPv4 and IPv6 address space following a network-based map-and-encapsulate scheme (RFC 1955). In LISP, both identifiers and locators can be IP addresses or arbitrary elements like a set of GPS coordinates or a MAC address.

The LISP Logo

Historical origin

The Internet Architecture Board's October 2006 Routing and Addressing Workshop renewed interest in the design of a scalable routing and addressing architecture for the Internet. Key issues driving this renewed interest include concerns about the scalability of the routing system and the impending exhaustion of IPv4 address space. Since the IAB workshop, several proposals have emerged that attempted to address the concerns expressed both at the workshop. All of these proposals are based on a common concept: the separation of Locator and Identifier in the numbering of Internet devices, often termed the "Loc/ID split".

Current Internet Protocol Architecture

The current addressing architecture used by the Internet Protocol uses IP addresses for two separate functions:

- as an end-point addressing identifier to uniquely identify a network interface within its local network addressing context
- as a locator for routing purposes, to identify where a network interface is located within a larger routing context

LISP

Advantages

There are several advantages to decoupling Location and Identifier, and to LISP specifically.

- Improved routing scalability
- BGP-free multihoming in active-active configuration
- Address family traversal: IPv4 over IPv4, IPv4 over IPv6, IPv6 over IPv6, IPv6 over IPv4
- Inbound traffic engineering
- Mobility
- Simple deployability
- No host changes are needed

Terminology

- **Routing Locator (RLOC)**: A RLOC is an IPv4 or IPv6 address of an egress tunnel router (ETR). A RLOC is the output of a EID-to-RLOC mapping lookup.
- **Endpoint ID (EID)**: An EID is an IPv4 or IPv6 address used in the source and destination address fields of the first (most inner) LISP header of a packet.
- **Egress Tunnel Router (ETR)**: An ETR is a router that accepts an IP packet where the destination address in the "outer" IP header is one of its own RLOCs. ETR functionality does not have to be limited to a router device. A server host can be the endpoint of a LISP tunnel as well.
- **Ingress Tunnel Router (ITR)**: An ITR receives IP packets from site end-systems on one side and sends LISP-encapsulated IP packets toward the Internet on the other side.
- **Proxy ETR (PETR)**: A PETR is used for inter-networking between LISP and Non-LISP sites, a PETR acts like an ETR but does so on behalf of LISP sites which send packets to destinations at non-LISP sites.
- **Proxy ITR (PITR)**: A PITR is used for inter-networking between Non-LISP and LISP sites, a PITR acts like an ITR but does so on behalf of non-LISP sites which send packets to destinations at LISP sites.

- **xTR**: A xTR is a reference to an ITR or ETR when direction of data flow is not part of the context description. xTR refers to the router that is the tunnel endpoint.

The LISP mapping system

In the Locator/Identifier Separation Protocol the network elements (routers) are responsible for looking up the mapping between end-point-indentifiers (EID) and route locators (RLOC) and this process is invisible to the Internet end-hosts. The mappings are stored in a distributed database called the mapping system, which responds to the lookup queries. The LISP beta network uses a BGP-based mapping system called LISP ALternative Topology (LISP+ALT). The protocol design makes it easy to plug in a new mapping system, if a different design proves to have additional benefits. Some proposals have already emerged and have been compared.

Implementations

- Cisco has released public IOS and NX-OS images which support LISP.
- A team of researchers from the Université catholique de Louvain and T-Labs have written a FreeBSD implementation called OpenLISP.

LISP beta network

A testbed has been developed to gain real-life experience with LISP. Participants include Google, Facebook, NTT, Level3, InTouch N.V. and the Internet Systems Consortium. As of October 2010 around 50 companies from 13 countries are involved.

Other approaches

Several proposals for separating the two functions and allowing the Internet to scale better have been proposed, for instance GSE/8+8 as network based solution and SHIM6, HIP and ILNP as host based solutions.

Chapter 7

Mbone and Multicast

Mbone

Mbone (short for "multicast backbone") was an experimental backbone for IP multicast traffic across the Internet developed in the early 1990s. It required specialized hardware and software. Since most Internet routers have IP multicast disabled due to concerns of bandwidth tracking and billing, the Mbone evolved to connect multicast-capable networks over the existing Internet infrastructure. The commercialization of multicast routers is difficult because there are no efficient access control capabilities to the multicast trees (multicast routers and their protocols), and because Internet service providers have difficulty computing charges for multicast traffic.

A November 1994 Rolling Stones concert at the Cotton Bowl in Dallas with 50,000 fans was the "first major cyberspace multicast concert." Mick Jagger opened the concert by saying, "I wanna say a special welcome to everyone that's, uh, climbed into the Internet tonight and, uh, has got into the M-bone. And I hope it doesn't all collapse."

By 1995 there were M-bone links in Russia, as well as at the McMurdo Sound research station in Antarctica.

Mbone is currently of practical use for shared communication such as videoconferences or shared collaborative workspaces. It is not generally connected to Internet service providers, but often to universities and research institutions. Some other projects and network testbeds, such as Internet2's Abilene Network, have made Mbone obsolete.

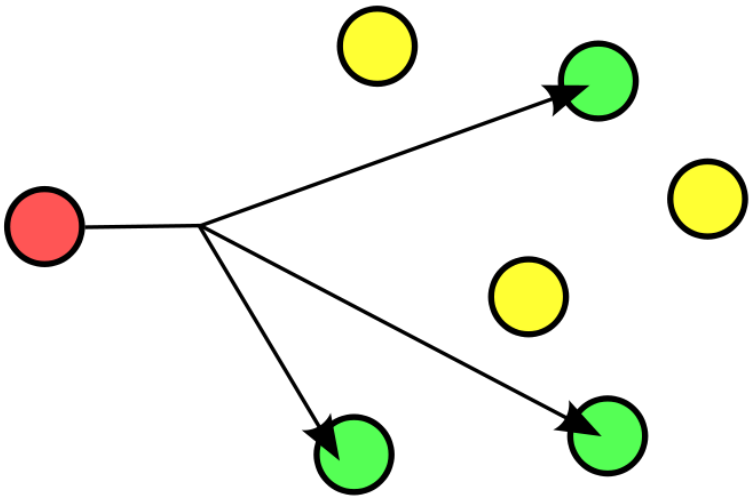
A recent application with support over Mbone was Virtual Room Videoconferencing System (VRVS).

Details

- Mbone is a virtual network built on top of the Internet, invented by Van Jacobson, Steve Deering and Stephen Casner in 1992. The purpose of Mbone is to minimize the amount of data required for multipoint audio/video-conferencing.
- Mbone is free; it uses a network of mrouters that can support IP multicast, and it enables access to real-time interactive multimedia on the Internet.

- Many older routers do not support IP multicast. To cope with this tunnels must be set up on both ends; also known as a tunneling protocol: multicast packets are encapsulated in unicast packets and sent through a tunnel.
- Mbone uses a small subset of the class D IP address space (224.0.0.0–239.255.255.255) assigned for multicast traffic. Mbone uses 224.2.0.0 for multimedia conferencing.
- Characteristics:
  - topology: combination of mesh and star networks
  - IP addresses: 224.2.0.0; routing schemes: DVMRP, MOSPF
  - session registration: IGMP
  - traffic requirement: audio 32-64 kbit/s, video 120 kbit/s
- Mbone tools:
  - Videoconferencing: vic -t ttl destination-host/port (supports: NV, H.261, CellB, MPEG, mJPEG)
  - Audioconferencing: vat -t ttl destination-host/port (supports: LPC, PCMU, DVI4, GSM)
  - Whiteboard: wb destination-host/port/ttl
  - Session Directory: sdr

Multicast



In computer networking, **multicast** is the delivery of a message or information to a group of destination computers simultaneously in a single transmission from the source creating copies automatically in other network elements, such as routers, only when the topology of the network requires it.

Multicast is most commonly implemented in IP multicast, which is often employed in Internet Protocol (IP) applications of streaming media and Internet television. In IP multicast the implementation of the multicast concept occurs at the IP routing level, where routers create optimal distribution paths for datagrams sent to a multicast destination address.

At the Data Link Layer, *multicast* describes one-to-many distribution such as Ethernet multicast addressing, Asynchronous Transfer Mode (ATM) point-to-multipoint virtual circuits (P2MP) or Infiniband multicast.

IP multicast

IP multicast is a technique for one-to-many communication over an IP infrastructure in a network. It scales to a larger receiver population by not requiring prior knowledge of who or how many receivers there are. Multicast uses network infrastructure efficiently by requiring the source to send a packet only once, even if it needs to be delivered to a large number of receivers. The nodes in the network take care of replicating the packet to reach multiple receivers only when necessary.

The most common transport layer protocol to use multicast addressing is User Datagram Protocol (UDP). By its nature, UDP is not reliable—messages may be lost or delivered out of order. Reliable multicast protocols such as Pragmatic General Multicast (PGM) have been developed to add loss detection and retransmission on top of IP multicast.

IP multicast is widely deployed in enterprises, commercial stock exchanges, and multimedia content delivery networks. A common enterprise use of IP multicast is for IPTV applications such as distance learning and televised company meetings.

Other multicast technologies

As of 2006, most effort at scaling multicast up to large networks have concentrated on the simpler case of single-source multicast, which seems to be more computationally tractable.

Still, the large state requirements in routers make applications using a large number of trees unworkable using IP multicast. Take presence information as an example where each person needs to keep at least one tree of its subscribers, if not several. No mechanism has yet been demonstrated that would allow the IP multicast model to scale to millions of senders and millions of multicast groups and, thus, it is not yet possible to make fully-general multicast applications practical. For these reasons, and also reasons of economics, IP multicast is not, in general, used in the commercial Internet backbone. The increasing availability of WiFi Access Points that support multicast IP is facilitating the emergence of WiCast WiFi Multicast that allows the binding of data to geographical locations.

Explicit Multi-Unicast (XCAST) is an alternate multicast strategy to IP multicast that provides reception addresses of all destinations with each packet. As such, since the IP packet size is limited in general, XCAST cannot be used for multicast groups of large number of destinations. The XCAST model generally assumes that the stations participating in the communication are known ahead of time, so that distribution trees can be generated and resources allocated by network elements in advance of actual data traffic.

Other multicast technologies not based on IP multicast are more widely used. Notably the Internet Relay Chat (IRC), which is more pragmatic and scales better for large numbers of small groups. IRC implements a single spanning tree across its overlay network for all conference groups. This leads to suboptimal routing for some of these groups however. Additionally, IRC keeps a large amount of distributed states that limit growth of an IRC network, leading to fractioning into several non-interconnected networks. The lesser known PSYC technology uses custom multicast strategies per conference. Also some peer-to-peer technologies employ the multicast concept when distributing content to multiple recipients.

Commercial deployment

Starting in 2005, the BBC has begun encouraging UK-based Internet service providers to adopt multicast-addressable services in their networks by providing BBC Radio at higher quality than is available via their unicast-addressed services. This has also been supported by a variety of commercial radio networks, including GCAP, EMAP, and Virgin Radio.

The German public-service broadcasters ARD and ZDF and the Franco-German network Arte offer their TV program multicasted on several networks. Austrian Internet service provider Telekom Austria offers its Digital Subscriber Line (DSL) customers a TV set-top box that uses multicast addressing in receiving TV and radio broadcasts. In Germany, T-Home, a brand of Deutsche Telekom, offers a similar service.

TV multicasting

Digital television technology increased the total available bandwidth for each broadcast channel to permit high-definition (HD) picture and audio quality. This additional bandwidth is used by many television broadcasters to deliver muliple channels of programming within the main HD channel. USA Today reported in 2004 that 213 of 1700 broadcast stations in the US (less than 13%) are using this type of transmission.

Chapter 8

Peering

Peering is a voluntary interconnection of administratively separate Internet networks for the purpose of exchanging traffic between the customers of each network. The pure definition of peering is settlement-free or "sender keeps all," meaning that neither party pays the other for the exchanged traffic; instead, each derives revenue from its own customers. Marketing and commercial pressures have led to the word peering routinely being used when there is some settlement involved, even though that is not the accurate technical use of the word. The phrase "settlement-free peering" is sometimes used to reflect this reality and unambiguously describe the pure cost-free peering situation.

Peering requires physical interconnection of the networks, an exchange of routing information through the Border Gateway Protocol (BGP) routing protocol and is often accompanied by peering agreements of varying formality, from "handshake" to thick contracts.

How peering works

The Internet is a collection of separate and distinct networks, each one operating under a common framework of globally unique IP addressing and global BGP routing.

The relationships between these networks are generally described by one of the following three categories:

- Transit (or *pay*) – You pay money (or *settlement*) to another network for Internet access (or *transit*).
- Peer (or *swap*) – Two networks exchange traffic between each other's customers freely, and for mutual benefit.
- Customer (or *sell*) – Another network pays you money to provide them with Internet access.

Furthermore, in order for a network to reach any specific other network on the Internet, it must either:

- Sell *transit* (or Internet access) service to that network (making them a 'customer'),

- Peer directly with that network, or with a network who sells transit service to that network, or
- Pay another network for transit service, where that other network must in turn also sell, peer, or pay for access.

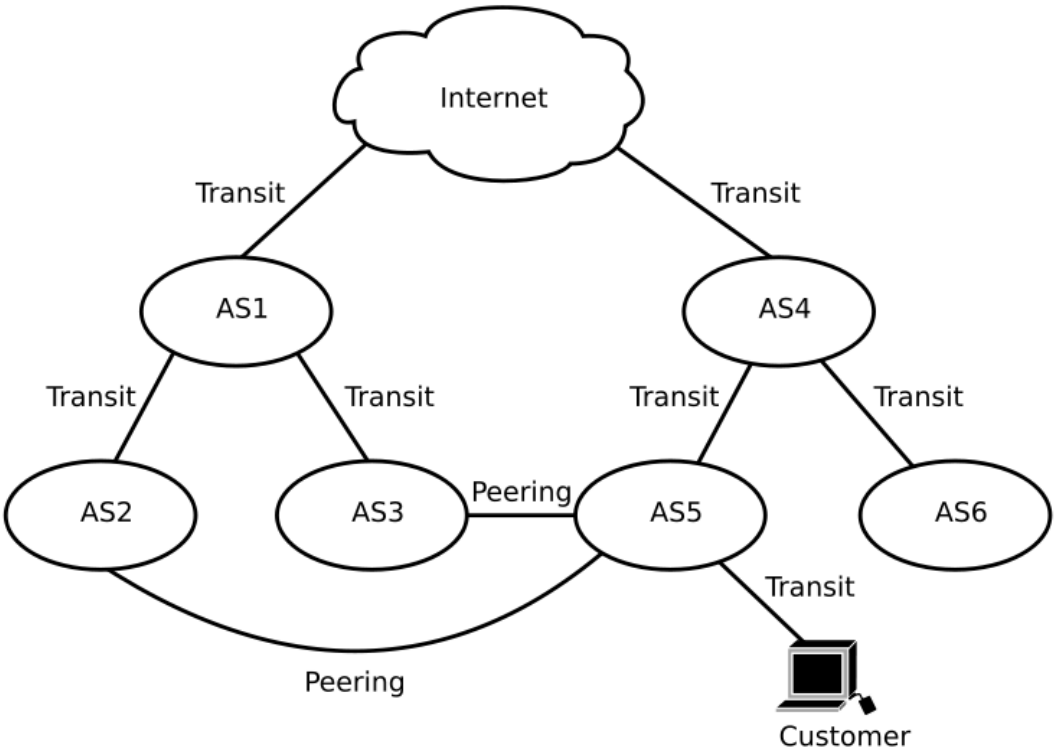
The Internet is based on the principle of *global reachability* (sometimes called *end-to-end reachability*), which means that any Internet user can reach any other Internet user as though they were on the same network. Therefore, any Internet connected network must by definition either pay another network for transit, or peer with every other network who also does not purchase transit.

Motivations for peering

Peering involves two networks coming together to exchange traffic with each other freely, and for mutual benefit. This 'mutual benefit' is most often the motivation behind peering, which is often described solely by "reduced costs for transit services". Other less tangible motivations can include:

- Increased redundancy (by reducing dependence on one or more transit providers).
- Increased capacity for extremely large amounts of traffic (distributing traffic across many networks).
- Increased routing control over your traffic.
- Improved performance (attempting to bypass potential bottlenecks with a "direct" path).
- Improved perception of your network (being able to claim a "higher tier").
- Ease of requesting for emergency aid (from friendly peers).

Physical interconnections for peering



Scheme of interconnection and peering of autonomous systems

The physical interconnections used for peering are categorized into two types:

- Public peering – Interconnection utilizing a multi-party shared switch fabric such as an Ethernet switch.
- Private peering – Interconnection utilizing a point-to-point link between two parties.

Public peering

Public peering is accomplished across a Layer 2 access technology, generally called a *shared fabric*. At these locations, multiple carriers interconnect with one or more other carriers across a single physical port. Historically, public peering locations were known as network access points (NAPs); today they are most often called exchange points or *Internet exchanges* ("IXP" or "IX"). Many of the largest exchange points in the world can have hundreds of participants, and some span multiple buildings and colocation facilities across a city.

Since public peering allows networks interested in peering to interconnect with many other networks through a single port, it is often considered to offer "less capacity" than private peering, but to a larger number of networks. Many smaller networks, or networks

who are just beginning to peer, find that public peering exchange points provide an excellent way to meet and interconnect with other networks who may be open to peering with them. Some larger networks utilize public peering as a way to aggregate a large number of "smaller peers", or as a location for conducting low-cost "trial peering" without the expense of provisioning private peering on a temporary basis, while other larger networks are not willing to participate at public exchanges at all.

A few exchange points, particularly in the United States, are operated by commercial carrier-neutral third parties. These operators typically go to great lengths to promote communication and encourage new peering, and will often arrange social events for these purposes.

Private peering

Private peering is the direct interconnection between only two networks, across a Layer 1 or 2 medium that offers dedicated capacity that is not shared by any other parties. Early in the history of the Internet, many private peers occurred across 'telco' provisioned SONET circuits between individual carrier-owned facilities. Today, most private interconnections occur at carrier hotels or carrier neutral colocation facilities, where a direct crossconnect can be provisioned between participants within the same building, usually for a much lower cost than telco circuits.

Most of the traffic on the Internet, especially traffic between the largest networks, occurs via private peering. However, because of the resources required to provision each private peer, many networks are unwilling to provide private peering to "small" networks, or to "new" networks who have not yet proven that they will provide a mutual benefit.

Peering agreements/contracts

Throughout the history of the Internet, there have been a spectrum of kinds of agreements between peers, ranging from handshake deals to peering contracts which may be required by one or both sides. Such a contract sets forth the details of how traffic is to be exchanged, along with a list of expected activities which may be necessary to maintain the peering relationship, a list of activities which may be considered abusive and result in termination of the relationship, and details concerning how the relationship can be terminated. Detailed contracts of this type are typically used between the largest ISPs, and the ones operating in the most heavily-regulated economies, accounting for about 1-2% of peering relationships overall.

History of peering

The first Internet exchange point was the Commercial Internet Exchange (CIX), formed by Alternet/UUNET (now Verizon Business), PSI, and CERFNET to exchange traffic without regard for whether the traffic complied with the acceptable use policy (AUP) of the NSFNet or ANS' interconnection policy. The CIX infrastructure consisted of a single router, managed by PSI, and was initially located in Santa Clara, California. Paying CIX

members were allowed to attach to the router directly or via leased lines. After some time, the router was also attached to the Pacific Bell SMDS cloud. The router was later moved to the Palo Alto Internet Exchange, or PAIX, which was developed and operated by Digital Equipment Corporation (DEC).

Another early exchange point was Metropolitan Area Ethernet, or MAE, in Tysons Corner, Virginia. When the United States government decided to de-fund the NSFNET backbone, Internet exchange points were needed to replace its function, and initial governmental funding was used to aid the MAE and bootstrap three other exchanges, which they dubbed NAPs, or "Network Access Points," in accordance with the terminology of the National Information Infrastructure document. All four are now defunct or no longer functioning as Internet exchange points:

- MAE-East - Located in Tysons Corner, VA, and later relocated to Ashburn, Virginia
- Chicago NAP - Operated by Ameritech and located in Chicago, Illinois
- New York NAP - Operated by Sprint and located in Pennsauken, New Jersey
- San Francisco NAP - Operated by PacBell and located in the Bay Area

As the Internet grew, and traffic levels increased, these NAPs became a network bottleneck. Most of the early NAPs utilized FDDI technology, which provided only 100 Mbit/s of capacity to each participant. Some of these exchanges upgraded to ATM technology, which provided OC-3 (155 Mbit/s) and OC-12 (622 Mbit/s) of capacity.

Other prospective exchange point operators moved directly into offering Ethernet technology, such as gigabit Ethernet (1000 Mbit/s), which quickly became the predominant choice for Internet exchange points due to the reduced cost and increased capacity offered. Today, almost all significant exchange points operate solely over Ethernet, and most of the largest exchange points offer ten gigabit Ethernet (10,000 Mbit/s) service.

During the dot-com boom, many exchange point and carrier neutral colocation providers had plans to build as many as 50 locations to promote carrier interconnection in the United States alone. Essentially all of these plans were abandoned following the dot-com bust, and today it is considered both economically and technically infeasible to support this level of interconnection among even the largest of networks.

Depeering

By definition, peering is the voluntary and free exchange of traffic between two networks, for mutual benefit. If one or both networks believes that there is no longer a mutual benefit, they may decide to cease peering: this is known as **depeering**. Some of the reasons why one network may wish to depeer another include:

- A desire that the other network pay settlement, either in exchange for continued peering or for transit services.

- A belief that the other network is "profiting unduly" from the settlement free interconnection.
- Concern over *traffic ratios*, which related to the fair sharing of cost for the interconnection.
- A desire to peer with the upstream transit provider of the peered network.
- Abuse of the interconnection by the other party, such as *pointing default* or utilizing the peer for transit.
- Instability of the peered network, repeated routing leaks, lack of response to network abuse issues, etc.
- The inability or unwillingness of the peered network to provision additional capacity for peering.
- The belief that the peered network is unduly peering with your customers.
- Various external political factors (including personal conflicts between individuals at each network).

In some situations, networks who are being depeered have been known to attempt to fight to keep the peering by intentionally breaking the connectivity between the two networks when the peer is removed, either through a deliberate act or an act of omission. The goal is to force the depeering network to have so many customer complaints that they are willing to restore peering. Examples of this include forcing traffic via a path that does not have enough capacity to handle the load, or intentionally blocking alternate routes to or from the other network. Some very notable examples of these situations have included:

- BBN Planet *vs* Exodus Communications
- PSINet *vs* Cable & Wireless

- AOL Transit Data Network (ATDN) *vs* Cogent Communications
- Teleglobe *vs* Cogent Communications
- France Telecom *vs* Cogent Communications
- France Telecom (Wanadoo) *vs* Proxad (Free)
- Level 3 Communications *vs* XO Communications
- Level 3 Communications *vs* Cogent Communications
- Telecom/Telefonica/Impsat/Prima *vs* CABASE (Argentina)
- Cogent Communications *vs* TeliaSonera
- Sprint-Nextel *vs* Cogent Communications

Modern peering

Peering locations

The modern Internet operates with significantly more peering locations than at any time in the past, resulting in improved performance and better routing for the majority of the traffic on the Internet. However, in the interests of reducing costs and improving efficiency, most networks have attempted to standardize on relatively few locations within these individual regions where they will be able to quickly and efficiently interconnect with their peering partners.

The primary locations for peering within the United States are generally considered to be:

- San Francisco Bay Region (San Jose CA, Palo Alto CA, Santa Clara CA, San Francisco CA)
- Washington DC / Northern Virginia Region (Washington, DC, Ashburn VA, Reston VA, Vienna VA)
- New York City Region (New York NY, Newark NJ)
- Chicago Region (Chicago IL)
- Los Angeles Region (Los Angeles, CA)
- Dallas Region (Dallas, TX, Plano, TX, Richardson, TX)
- Miami, FL
- Seattle, WA

For international traffic, the most important locations for peering are generally considered to be

Europe;

- Amsterdam, Netherlands
- London, United Kingdom
- Frankfurt, Germany

Asia;

- Tokyo, Japan
- Hong Kong, China
- Seoul, South Korea
- Singapore

Exchange points

The largest individual exchange points in the world are AMS-IX in Amsterdam, followed by DE-CIX in Frankfurt Germany and LINX in London. The next largest exchange point is generally considered to be JPNAP in Tokyo, Japan. The United States, with a historically larger focus on private peering and commercial public peering, has a much smaller amount of traffic on public peers compared to other regions which operate non-profit exchange points. The combined exchange points in multiple cities operated by Equinix are generally considered to be the largest and most important, followed by the PAIX facilities which are operated by Switch and Data. Other important but smaller exchange points include LIPEX and LONAP in London UK, NYIIX in New York, and NAP of the Americas in Miami, Florida.

URLs to some public traffic statistics of exchange points include:

- AMS-IX
- DE-CIX

- LINX
- MSK-IX
- TORIX
- NYIIX
- LAIIX
- TOP-IX
- Netnod

Peering and BGP

A great deal of the complexity in the BGP routing protocol exists to aid the enforcement and fine-tuning of peering and transit agreements. BGP allows operators to define a policy that determines where traffic is routed. Three things commonly used to determine routing are local-preference, multi exit discriminators (MEDs) and AS-Path. Local-preference is used internally within a network to differentiate classes of networks. For example, a particular network will have a higher preference set on internal and customer advertisements. Settlement free peering is then configured to be preferred over paid IP transit.

Networks that speak BGP to each other can engage in multi exit discriminator exchange with each other, although most do not. When networks interconnect in several locations, MEDs can be used to reference that network's interior gateway protocol cost. This results in both networks sharing the burden of transporting each others traffic on their own network (or *cold potato*). *Hot-potato* or nearest-exit routing, which is typically the normal behavior on the Internet, is where traffic destined to another network is delivered to the closest interconnection point.

Law and policy

Internet interconnection is not regulated in the same way that public telephone network interconnection is regulated. Nevertheless, Internet interconnection has been the subject of several areas of federal policy. Perhaps the most dramatic example of this is the attempted MCI Worldcom/Sprint merger. In this case, the Department of Justice signaled that it would move to block the merger specifically because of the impact of the merger on the Internet backbone market. In 2001, the Federal Communications Commission's advisory committee, the Network Reliability and Interoperability Council recommended that Internet backbones publish their peering policies, something that they had been hesitant to do beforehand. The FCC has also reviewed competition in the backbone market in its Section 706 proceedings which review whether advanced telecommunications are being provided to all Americans in a reasonable and timely manner.

Finally, Internet interconnection has become an issue in the international arena under something known as the International Charging Arrangements for Internet Services (ICAIS). In the ICAIS debate, countries underserved by Internet backbones have complained that it is unfair that they must pay the full cost of connecting to an Internet

exchange point in a different country, frequently the United States. These advocates argue that Internet interconnection should work like international telephone interconnection, with each party paying half of the cost. Those who argue against ICAIS point out that much of the problem would be solved by building local exchange points. A significant amount of the traffic, it is argued, that is brought to the US and exchanged then leaves the US, using US exchange points as switching offices but not terminating in the US. In some worst-case scenarios, traffic from one side of a street is brought to all the way to Miami, exchanged, and then returned to another side of the street. Countries with liberalized telecommunications and open markets, where competition between backbone providers occurs, tend to oppose ICAIS.

Chapter 9

Introduction to Internet Governance

Policies and mechanisms for **Internet governance** have been topics of debate between many different Internet stakeholders, some of whom have very different opinions for how and indeed whether the Internet should facilitate free communication of ideas and information.

Definition

The definition of Internet governance has been contested by differing groups across political and ideological lines. One of the main debates concerns the authority and participation of certain actors, such as national governments, corporate entities and civil society, to play a role in the Internet's governance.

A Working group established after a United Nations-initiated World Summit on the Information Society (WSIS) proposed the following definition of Internet governance as part of its June 2005 report:

Internet governance is the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet.

Law professor Yochai Benkler developed a conceptualization of Internet governance by the idea of three "layers" of governance: the "physical infrastructure" layer through which information travels; the "code" or "logical" layer that controls the infrastructure; and the "content" layer, which contains the information that signals through the network.

History

To understand how the Internet is managed today, it is necessary to know some of the main events of Internet governance.

The original ARPANET, one of the components which evolved eventually into the Internet, connected four Universities: University of California Los Angeles, University of California Santa Barbara, Stanford Research Institute and Utah University. The IMPs,

interface minicomputers, were built during 1969 by Bolt, Beranek and Newman in accord with a proposal by the US Department of Defense Advanced Research Projects Agency, which funded the system as an experiment. By 1973 it connected many more systems and included satellite links to Hawaii and Scandinavia, and a further link from Norway to London. ARPANET continued to grow in size, becoming more a utility than a research project. For this reason during 1975 it was transferred to the US Defense Communications Agency.

During the development of ARPANET, a numbered series of Request for Comments (RFCs) memos documented technical decisions and methods of working as they evolved. The standards of today's Internet are still documented by RFCs, produced through the very process which evolved on ARPANET.

Outside of the USA the dominant technology was X.25. The International Packet Switched Service, created during 1978, used X.25 and extended to Europe, Australia, Hong Kong, Canada, and the USA. It allowed individual users and companies to connect to a variety of mainframe systems, including Compuserve. Between 1979 and 1984, a system known as Unix to Unix Copy Program grew to connect 940 hosts, using methods like X.25 links, ARPANET connections, and leased lines. Usenet News, a distributed discussion system, was a major use of UUCP.

The Internet protocol suite, developed between 1973 and 1977 with funding from ARPA, was intended to hide the differences between different underlying networks and allow many different applications to be used over the same network.

RFC 801 describes how the US Department of Defense organized the replacement of ARPANET's Network Control Program by the new Internet Protocol during January 1983. During the same year, the military systems were removed to a distinct MILNET, and the Domain Name System was invented to manage the names and addresses of computers on the "ARPA Internet". The familiar top-level domains .gov, .mil, .edu, .org, .net, .com, and .int, and the two-letter country code top-level domains were deployed during 1984.

Between 1984 and 1986 the US National Science Foundation created the NSFNET backbone, using TCP/IP, to connect their supercomputing facilities. The combined network became generally known as the Internet.

By the end of 1989 Australia, Germany, Israel, Italy, Japan, Mexico, the Netherlands, New Zealand, and the United Kingdom had connected to the Internet, which now contained over 160,000 hosts.

During 1990, ARPANET formally terminated, and during 1991 the NSF ended its restrictions on commercial use of its part of the Internet. Commercial network providers began to interconnect, extending the Internet.

Today almost all Internet infrastructure is provided and owned by the private sector. Traffic is exchanged between these networks, at major interconnect points, in accordance with established Internet standards and commercial agreements.

Actors

During 1979 the Internet Configuration Control Board was founded by DARPA to oversee the network's development. During 1984 it was renamed the Internet Advisory Board (IAB), and during 1986 it became the Internet Activities Board.

The Internet Engineering Task Force (IETF) was formed during 1986 by the US Government to develop and promote Internet standards. It consisted initially of researchers, but by the end of the year participation was available to anyone, and its business was performed largely by email.

From the early days of the network until his death during 1998, Jon Postel oversaw address allocation and other Internet protocol numbering and assignments in his capacity as Director of the Computer Networks Division at the Information Sciences Institute of the University of Southern California, under a contract from the Dept. of Defense. This function eventually became known as the Internet Assigned Numbers Authority (IANA), and as it expanded to include management of the global Domain Name System (DNS) root servers, a small organization grew. Postel also served as RFC Editor.

Allocation of IP addresses was delegated to four Regional Internet Registries (RIRs):

- American Registry for Internet Numbers (ARIN) for North America
- Réseaux IP Européens - Network Coordination Centre (RIPE NCC) for Europe, the Middle East, and Central Asia
- Asia-Pacific Network Information Centre (APNIC) for Asia and the Pacific region
- Latin American and Caribbean Internet Addresses Registry (LACNIC) for Latin America and the Caribbean region

In 2004 a new RIR, AfriNIC, was created to manage allocations for Africa.

After Jon Postel's death during 1998, the IANA became part of the Internet Corporation for Assigned Names and Numbers (ICANN), a newly created Californian non-profit corporation, initiated during September 1998 by the US Government and awarded a contract by the US Department of Commerce. Initially two board members were elected by the Internet community at large, though this was changed by the rest of the board during 2002 in a little- attended public meeting in Accra, in Ghana.

During 1992 the Internet Society (ISOC) was founded, with a mission to *"assure the open development, evolution and use of the Internet for the benefit of all people throughout the world"*. Its members include individuals (anyone may join) as well as corporations, organizations, governments, and universities. The IAB was renamed the Internet

Architecture Board, and became part of ISOC. The Internet Engineering Task Force also became part of the ISOC. The IETF is overseen currently by the Internet Engineering Steering Group (IESG), and longer term research is carried on by the Internet Research Task Force and overseen by the Internet Research Steering Group.

During 2002, a restructuring of the Internet Society gave more control to its corporate members.

At the first World Summit on the Information Society (WSIS) in Geneva 2003 the topic of Internet governance was discussed. ICANN's status as a private corporation under contract to the U.S. government created controversy among other governments, especially Brazil, China, South Africa and some Arab states. Since no general agreement existed even on the definition of what comprised Internet governance, United Nations Secretary General Kofi Annan initiated a Working Group on Internet Governance (WGIG) to clarify the issues and report before the second part of the World Summit on the Information Society in Tunis 2005. After much controversial debate, during which the US delegation refused to consider surrendering the US control of the Root Zone file, participants agreed on a compromise to allow for wider international debate on the policy principles. They agreed to establish an Internet Governance Forum, to be convened by United Nations Secretary General before the end of the second quarter of the year 2006. The Greek government volunteered to host the first such meeting.

Controversy

The position of the US Department of Commerce as the controller of the Internet gradually attracted criticism from those who felt that control should be more international. A hands-off philosophy by the US Dept. of Commerce helped limit this criticism, but this was undermined in 2005 when the Bush administration intervened to help kill the .xxx top level domain proposal.

When the IANA functions were given to a new US non-profit Corporation called ICANN, controversy increased. ICANN's decision-making process was criticised by some observers as being secretive and unaccountable. When the directors' posts which had previously been elected by the "at-large" community of Internet users were abolished, some feared the worst. ICANN stated that they were merely streamlining decision-making processes, and developing a structure suitable for the modern Internet.

Other topics of controversy included the creation and control of generic top-level domains (.com, .org, and possible new ones, such as .biz or .xxx), the control of country-code domains, recent proposals for a large increase in ICANN's budget and responsibilities, and a proposed "domain tax" to pay for the increase.

There were also suggestions that individual governments should have more control, or that the International Telecommunication Union or the United Nations should have a function in Internet governance.

Chapter 10

Alternative DNS Root and Domain Name Registry

Alternative DNS root

The Internet uses the Domain Name System (DNS) to associate the names of computers with their numeric IP addresses and with other information. The top level of the domain name hierarchy, the DNS root, contains the top-level domains that appear as the suffixes of all Internet domain names. The official DNS root is administered by the Internet Corporation for Assigned Names and Numbers (ICANN). In addition, several organizations operate **alternative DNS roots** (often referred to as **alt roots**). These alternative domain name systems operate their own root nameservers and administer their own specific name spaces consisting of custom top-level domains.

The Internet Architecture Board has spoken out strongly against alternate roots in RFC 2826.

Description

The DNS root zone consists of pointers to the authoritative domain name servers for all TLDs (top-level domains). The root zone is hosted on a collection of root servers operated by several organizations around the world that all use a specific, approved list of domains that is managed by ICANN.

Alternative roots typically include pointers to all of the TLD servers for domains delegated by ICANN, as well as name servers for other, custom top-level domains that are not sanctioned by ICANN. Some alternate roots are operated by the organizations that manage these alternative TLDs.

Alternative DNS roots may be characterized as three groups: those run for idealistic or ideological reasons, those run as profit-making enterprises, and those run internally by an organization for its own use.

While technically trivial to set up, the maintenance of a reliable root server network is a serious undertaking. In order for the system to be effective, multiple servers must be run continuously without interruption in geographically diverse locations.

During the dot-com boom, some alternate root providers believed that there were substantial profits to be made from providing alternative top-level domains.

Only a small portion of Internet service providers actually use any of the domains served by alternate root operators, generally supporting only ICANN-sanctioned root servers. This has led to the commercial failure of several alternative DNS root providers.

A `BIZ` TLD created by Pacific Root was in operation before ICANN approved the official `BIZ` domain, operated by Neulevel. For some time after the creation of the official domain, several alternate roots continued to resolve `BIZ` domains to Pacific Root's servers rather than Neulevel's. Therefore, some domain names existed in different roots and pointed to different IP addresses. The possibility of such conflicts, and their potential for destabilizing the Internet, is the main source of controversy surrounding alternate roots. Many of the alternate roots try to coordinate with each other, but many do not, and no conflict resolution processes exist between them.

List of alternative roots and their domains

This section lists the known alternate DNS roots, and for each root, lists the TLDs carried in addition to the ICANN-sanctioned gTLDs and ccTLDs.

Active public root zones

Public-Root

- Public-Root resolves multiple kinds of TLDs globally. It is created to offer an alternative, open DNS infrastructure with its own 13 root servers around the world.
- Administrated by INAIC
- Open for registration of new TLDs through an approved registrar, such as GQNET

OpenNIC

Public Access Website:

- `bbs` — aimed toward (Telnet style) bulletin board system servers, and affiliated/related/owned Websites.
- `dyn` — Approved by the OpenNIC Community, and will be introduced in mid-2008. Used to resolve dynamic DNS.
- `free` — non-commercial use of the Internet
- `fur` — Furry and Furry Fandom related sites

- `geek` — anything geeky
- `glue` — Sites related to infrastructure
- `indy` — Independent news and media
- `ing` — fun TLD. Further details to be confirmed
- `null` — miscellaneous non-commercial individual sites
- `oss` — Open source software
- `parody` — Parodies
- `eco` — Intended for the use in socially responsible investing (SRI) and ecological cooperatives, wholly owned subsidiaries, and other organisations that exist to promote or support the said co-operative.

New.net

Website:

- `agent`
- `art`
- `auction`
- `chat`
- `shop`
- `free`
- `golf`
- `llc`
- `llp`
- `love`
- `ltd`
- `school`
- `scifi`
- `soc`
- `video`
- `travel` — conflicts with ICANN-sanctioned TLD `travel`
- `tech`
- `kids`
- `church`
- `game`
- `mp3`
- `med`
- `mail`
- `xxx` — conflicts with TLD `xxx` which is in review by ICANN as of 2010
- `club`
- `inc`
- `law`
- `family`
- `sport`

UnifiedRoot

Website:

- UnifiedRoot enables all existing TLDs and allows new TLDs to be registered at a cost of €50,000 each (plus annual maintenance fees of €12,500).

UnifiedRoot offers a downloadable tool to modify the name server configuration on Windows. UnifiedRoot have also made agreements with ISPs and telcos to enable access to the provided TLDs. UnifiedRoot supports internationalized domain names (IDN) for top level domains (TLDs).

MobileTLD

Website:

- MobileTLD claims to resolve domains for mobile devices.

Open RSC

One of the notable challengers to ICANN's control of the DNS namespace was *Open RSC*, a group which grew out of private discussions and morphed into a public mailing list which grew large enough the group decided to submit an application to the US government to run the DNS.

Bylaws and articles of incorporation were posted outlining ORSC's position following extensive public discussion regarding the manner in which DNS was being run.

ICANN chairwoman Esther Dyson acknowledged adopting features such as membership from ORSC in her response to the US Department of Commerce.

ORSC publishes a root zone containing additional top level domains not found in the official root zone.

Website:

- `per` — personal pages
- `etc` — anything
- `web` — for the web
- `shop` — online shops
- `pickle` — just a general funny name
- `sco` — for Scottish culture
- `mail` - a tld for email - to reduce spam and clearly identify email servers.

Inactive public root zones

AlterNIC

AlterNIC ceased operation in 1997.

- `exp` —
- `llc` —
- `lnx` —
- `ltd` —
- `med` —
- `nic` —
- `noc` —
- `porn` —
- `xxx` —

eDNS

eDNS stopped in 1998.

- `biz` — General business use
- `corp` — For use by corporations
- `fam` — For and about Family
- `k12` — For and about Kids
- `npo` — Non-profit organizations
- `per` — Personal Domain Name services
- `web` — Web-based sites (ie: web pages)

Iperdome

Iperdome stopped in 1999.

- `per` — Personal Domain Name services
- later the TLDs changed to:
  - `biz` — General business use
  - `corp` — For use by corporations
  - `gay` — For and about the Gay Community
  - `k12` — For and about Kids
  - `npo` — Non-profit organizations
  - `pol` — Related to Poland and Polish organizations
  - `web` — Web-based sites (ie: web pages)

Open Root Server Network (ORSN)

(Shutdown 31.12.2008 00:00 UTC) Website:

- Used to be a mirror of the ICANN root.

Active private root zones

A number of organizations have alternative top-level domains configured on their internal DNS infrastructures, accessible only from within the enterprise. For instance, the National Security Agency operates the `nsa` domain; many NSA internal email addresses are of the form `username@r21.r.nsa`, mirroring the NSA organizational group structure.

Domain name registry

A **domain name registry** is a database of all domain names registered in a top-level domain. A registry operator, also called a **network information center** (NIC), is the part of the Domain Name System (DNS) of the Internet that keeps the database of domain names, and generates the zone files which convert domain names to IP addresses. Each NIC is an organisation that manages the registration of Domain names within the top-level domains for which it is responsible, controls the policies of domain name allocation, and technically operates its top-level domain. It is potentially distinct from a domain name registrar.

Domain names are managed under a hierarchy headed by the Internet Assigned Numbers Authority (IANA), which manages the top of the DNS tree by administrating the data in the root nameservers.

IANA also operates the .int registry for intergovernmental organisations, the .arpa zone for protocol administration purposes, and other critical zones such as root-servers.net.

IANA delegates all other domain name authority to other domain name registries such as VeriSign.

Country code top-level domains (ccTLD) are delegated by IANA to national registries such as DENIC in Germany and Nominet in the United Kingdom.

Operation

Some name registries are government departments (e.g., the registry for Sri Lanka *nic.lk*). Some are co-operatives of Internet service providers (such as DENIC) or not-for profit

companies (such as Nominet UK). Others operate as commercial organizations, such as the US registry (*nic.us*).

The allocated and assigned domain names are made available by registries by use of the WHOIS system and via their Domain name servers.

Some registries sell the names directly (like SWITCH in Switzerland) and others rely on separate entities to sell them. For example, names in the .com TLD are in some sense sold "wholesale" at a regulated price by VeriSign, and individual domain name registrar sell names "retail" to businesses and consumers.

Policies

Allocation policies

Generally, domain name registries operate a first-come-first-served system of allocation but may reject the allocation of specific domains on the basis of political, religious, historical, legal or cultural reasons.

For example, in the United States, between 1996 and 1998, InterNIC automatically rejected domain name applications based on a list of perceived obscenities.

Registries may also control matters of interest to their local communities: for example, the German, Japanese and Polish registries have introduced internationalized domain names to allow use of local non-ASCII characters.

Dispute policies

Domains which are registered with ICANN registrars, generally have to use the Uniform Domain-Name Dispute-Resolution Policy (UDRP), however, Germany's DENIC requires people to use the German civil courts, and Nominet UK deals with Intellectual Property and other disputes through its own dispute resolution service.

Prices of registration

Prices of domain registrations are set by each registry.

Third-level domains

Domain name registries may also impose a system of third-level domains on users. DENIC, the registry for Germany (.de), does not impose third level domains. AFNIC, the registry for France (.fr), has some third level domains, but not all registrants have to use them, and Nominet UK, the registry for the United Kingdom (.uk), requires all names to have a third level domain (e.g. `.co.uk or .org.uk`).

Many ccTLDs have moved from compulsory third or fourth-level domain to the availability of registrations of second level domains. Among them are .us (April 2002), .mx (May 2009), and .co (March 2010).

Domain Sub-Registration

Registrants of second-level domains sometimes act as a registry by offering sub-registrations to their registration. For example, registrations to `.fami.ly` are offered by the registrant of `fami.ly` and not by GPTC, the registry for Libya (.ly).

Chapter 11

Internet Governance Forum

Internet Governance Forum, Rio de Janeiro 2007

The **Internet Governance Forum** (**IGF**) is a multi-stakeholder forum for policy dialogue on issues of Internet governance. The establishment of the IGF was formally announced by the United Nations Secretary-General in July 2006 and it was first convened in October / November 2006.

Structure and Function

The formation of the Internet Governance Forum was first recommended in the report of the Working Group on Internet Governance following a series of open consultations. This report was one of the inputs to the second phase of the World Summit on the Information Society in Tunis in 2005, which formally called for the creation of the IGF and set out its mandate.

Following an open consultation meeting called in February 2006, the UN Secretary-General established an Advisory Group, the MAG, and a Secretariat as the main institutional bodies of the IGF.

These organizational divisions should not be considered concrete since the organizational structures will continue to be adjusted and to be changed until they fit into the needs of the members.

Multistakeholder Advisory Group - MAG

The Advisory Group, now referred to as the MAG (Multistakeholder Advisory Group), was set up by the former Secretary General of the United Nations, Mr Kofi Annan on May 17, 2006. The MAG was originally made up of 46 Members from international governments, the commercial private sector and public civil society, including academic and technical communities, and was chaired by Nitin Desai- the Secretary-General's Special Adviser for the World Summit on the Information Society. All stakeholders participate as equals. The purpose for which the MAG was set up was to assist the Secretary General in convening the Internet Governance Forum. On August 20, 2007, the mandate of the MAG was renewed with a new structure of 47 members, and a Co-Chairmanship by Nitin Desai, and Brazilian diplomat Hadil da Rocha Vianna. The mandate of the MAG was further extended on April 30, 2008 with a renewed one third of its members within each stakeholder group and Nitin Desai serving as the sole Chairman. The MAG meet three times each year - in February, May and September. All three meetings take place in Geneva at the Palais des Nations and they are preceded by open consultations meeting.

The details on MAG's operating principles and selection criteria are contained in the summary report of its February meeting available at this link.

On August 22, 2008, the United Nations Office in Geneva renewed the membership of MAG to prepare for the Internet Governance Forum Meeting in Hyderabad, India. There were a total of 50 members, among them 17 new appointed members, which represents 1/3 of its membership. Nitin Desai continues to be the Chairman for the Advisory Group. (Source: UN Department of Public Information, United Nations Office in Genev. Accessed online at:

- Actual List of Members
- MAG Meetings

Secretariat

The Secretariat, based in the United Nations Office in Geneva, assists and coordinates the work of the MAG, Multistakeholder Advisory Group. The Secretariat is headed by Markus Kummer with the designation of Executive Coordinator and Chengetai Masango is Programme and Technology Manager. The Secretariat also hosts fellowships. Markus Kummer has also been involved with the WGIG as its Executive Coordinator of the Secretariat.

History and Development of the Internet Governance Forum

WSIS Follow Ups

The IGF is considered an important development of the World Summit on Information Technology (WSIS). This important outcome was reaffirmed by paragraphs 37 and 38 of the Tunis 2005 Commitment. Paragraph 37 states that *"...goals can be accomplished through the involvement, cooperation and partnership of governments and other stakeholders, i.e. the private sector, civil society and international organizations, and that international cooperation and solidarity at all levels are indispensable if the fruits of the Information Society are to benefit all."* Corollary to this commitment, paragraph 38 states, too, that all efforts from here on *"should not stop with the conclusion of the Summit…emergence of the global Information Society to which we all contribute provides increasing opportunities for all our peoples and for an inclusive global community…we must harness these opportunities today and support their further development and progress."*

The Tunis Summit of 2005 made significant headway when the mandate of the IGF was formulated. In paragraph 72 of the Tunis Agenda, the UN Secretary-General was asked to convene a meeting with regards to the new multi-stakeholder forum, otherwise known as the IGF. In this mandate, different stakeholders are encouraged to strengthen engagement, particularly those from developing countries. In paragraph 72(h), the mandate focused on capacity-building for developing countries and the drawing out of local resources. This particular effort, for instance, has been reinforced through *Diplo Foundation's* Internet Governance Capacity Building Programme (IGCBP) that allowed participants from different regions to benefit from valuable resources with the help of regional experts in IG.

The involvement of different stakeholders in the policy framework of the IGF is a re-affirmation of commitment as per paragraph 39 of the Tunis Commitment. In this particular context, there is a deep resolve to *"...develop and implement an effective and sustainable response to the challenges and opportunities of building a truly global Information Society that benefits all our peoples."* During the OECD Civil Society-Organized Labour Forum held last June 16, 2008, in Seoul, Korea, Ambassador David A. Gross of the US Department of State talked about the transformation of the Internet in the social lives of people. He believed that this transformation made an impact in the free flow of information that politically drives challenges. Ambassador Gross commented on the 2005 WSIS because of the powerful language used on paragraph 4 of the Tunis agenda that reiterated on openness.

Formation of the IGF

A multi-stakeholder's approach was reiterated in the coordination of international activities for the IGF. This adaptation was set from paragraphs 29 to 35 of the Tunis agenda. These stakeholders were defined as coming from governments, the private technical and economic sector, civil society, intergovernmental organizations, and

international organizations. In paragraph 32, the UN Secretary-General was commended for his efforts in establishing the Working Group on Internet Governance (WGIG).

The suggested need of an organization like the IGF was first pointed out in the WGIG Report. After reaching a clear consensus among its members the WGIG proposed in paragraph 40 of the Report that:

"(t)he WGIG identified a vacuum within the context of existing structures, since there is no global multi-stakeholder forum to address Internet-related public policy issues. It came to the conclusion that there would be merit in creating such a space for dialogue among all stakeholders. This space could address these issues, as well as emerging issues, that are cross-cutting and multidimensional and that either affect more than one institution, are not dealt with by any institution or are not addressed in a coordinated manner".

The IGF was one of four proposals made in the report.

The idea of the Forum was also proposed by Argentina, as stated in its proposal made during the last Prepcom 3 in Tunis:

"(t)In order to strengthen the global multistakeholder interaction and cooperation on public policy issues and developmental aspects relating to Internet governance we propose a forum. This forum should not replace existing mechanisms or institutions but should build on the existing structures on Internet governance, should contribute to the sustainability, stability and robustness of the Internet by addressing appropriately public policy issues that are not otherwise being adequately addressed excluding any involvement in the day to day operation of the Internet. It should be constituted as a neutral, non-duplicative and non-binding process to facilitate the exchange of information and best practices and to identify issues and make known its findings, to enhance awareness and build consensus and engagement. Recognizing the rapid development of technology and institutions, we propose that the forum mechanism periodically be reviewed to determine the need for its continuation."

The convening of the IGF was announced on 18 July 2006, with the inaugural meeting of the Forum being held in Athens, Greece from 30 October to 2 November 2006.

Consultations

There were two rounds of consultations with regards to the convening of the first IGF:

16 – 17 of February 2006 – The first round of consultations was held in Geneva. The transcripts of the two-day consultations are available in the IGF site.

19 May 2006 – The second round of consultations was open to all stakeholders and was coordinated for the preparations of the inaugural IGF meeting. The meeting chairman

was *Nitin Desai* who is the United Nations Secretary-General's Special Adviser for Internet Governance.

The Second Meeting of the IGF

Consultations held in Geneva last May 23, 2007 were open to all stakeholders. This consultation was part of a cluster of related events of the WSIS that took place last 15-25 of May 2007. An advisory group was also facilitated for the IGF meeting in Rio de Janeiro, Brazil. The IGF open Consultations held last 3 September 2007 was held in Geneva.

For further information, a summary of the IGF consultations and meetings can be found below:

Date	Event
16–18 November 2005	Second Phase of the WSIS in Tunis
16 – 17 February 2006	First Round of Consultations
2 March 2006	Establishment of the IGF Secretariat
19 May 2006	Second Round of Consultations
22 – 23 May 2006	Establishment and First Meeting of the IGF Advisory Group
18 July 2006	Convening of the IGF
7 – 8 September 2006	Second Meeting of the IGF Advisory Group
30 October – 2 November 2006	Inaugural Meeting of the IGF in Athens
12–15 November 2007	Second Meeting of the IGF in Rio de Janeiro, Brazil

13 May 2008	Open Consultations
14–15 May 2008	Meeting of the IGF Multistakeholder Advisory Group(**MAG**)
3–6 December 2008	Third meeting of the IGF in Hyderabad, India
15–18 November 2009	Fourth Meeting of the IGF in Sharm El Shiekh, Egypt
14–17 September 2010	Fifth Meeting of the IGF in Vilnius, Lithuania

The government of Egypt offered to host the 2009 IGF meeting, while the governments of Lithuania and Azerbaijan made a bid for the 2010 meeting.

Mandate and Outcome

The mandate of the IGF is principally that of a discussion forum for facilitating dialogue between the participants. The IGF may *"identify emerging issues, bring them to the attention of the relevant bodies and the general public, and, where appropriate, make recommendations,"* but does not have any direct decision-making authority.

Activities at the IGF

The following are the activities that take place during the IGF: Workshops, Best Practice Forums, Open Forums and meetings of the Dynamic Coalitions.

The main themes of IGF are: openness, security, diversity and access. A new theme was introduced in IGF Brazil: critical Internet resources being one of the most debatable topics in the IG field at the moment.

Dynamic Coalitions

The most tangible result of the first IGF in Athens is the establishment of a number of so-called *Dynamic Coalitions*. These coalitions are relatively informal, issue-specific groups consisting of stakeholders that are interested in the particular issue.

Most coalitions allow participation of anyone interested in contributing. Thus, these groups gather not only academics and representatives of governments, but also members of the civil society interested in participating on the debates and engaged in the coalition's works.

So far, the following Dynamic Coalitions were brought to the attention of the IGF Secretariat:

- The StopSpamAlliance
- Dynamic Coalition on Privacy
- The IGF Dynamic Coalition on Open Standards (IGF DCOS)
- The Dynamic Coalition on Access and Connectivity for Remote, Rural and Dispersed Communities
- Dynamic Coalition on the Internet Bill of Rights
- Dynamic Coalition for Linguistic Diversity
- A2K@IGF Dynamic Coalition
- Freedom of Expression and Freedom of the Media on the Internet (FOEonline)
- Online Collaboration Dynamic Coalition
- Gender and Internet Governance (GIG)
- Framework of Principles for the Internet
- Dynamic Coalition on Child Online Safety
- Dynamic Coalition on "Accessibility and Disability"
- Dynamic Coalition for Online Education

Active Dynamic Coalitions

Workshops

In 2007, IGF hosted a number of workshops which attracted great interest with the public. In particular, the theme of child protection was one of the topics that increased the engagement of the participants in the events.

For 2008 the IGF page stipulates that workshops can be proposed on the draft main session headings:

```
  * Universalization of the Internet - How to reach the next billion
(Expanding the Internet)
  * Low cost sustainable access
  * Multilingualization
  * Implications for development policy
  * Managing the Internet (Using the Internet)
  * Critical Internet resources
  * Arrangements for Internet governance
  * Global cooperation for Internet security and stability
  * Taking stock and the way forward
  * Emerging issues
```

The following workshops have been proposed as of 15 May 2008, according to the Workshop page . These proposals will be reviewed and an attempt will be made to merge propositions into a manageable number of workshops.

Number Proposed	Workshop Theme
15	Access
9	Diversity
15	Openness
21	Security
13	Critical Internet Resources
9	Development
6	Capacity Building
17	Other

I IGF Athens 2006

The host webpage brings interesting information about the evolution of the first IGF.

II IGF Rio 2007

There were 84 events happening in parallel to the main sessions, organized under the 5 main themes: (i) critical Internet resources; (ii) access; (iii) diversity; (iv) openness and (v) security. There were 36 workshops, 23 best practices forums, 11 dynamic coalitions meetings, 8 open forums and 6 events covering other issues (like the Giganet Symposium)

The host webpage keeps video and audio records from main sessions and some parallel events such as workshops, best practices and open forums, as well as the tool for translation into Arabic.

Regarding the participation by region, around 35% of the attendees came from the Latin America and Caribbean of which 29% were from the host country (Brazil).

There are also some interesting statistics such as:

Region	Participation
Latin America and Caribbean	35%
Western Europe	20%
North America	13%
Asia	13%
Africa	10%
Eastern Europe	7%
Oceania	2%

Main sessions

The main sessions were developed according to the five themes chosen for this year: Critical Internet Resources, Access, Diversity, Openness and Security.

Please see below the summary of the main sessions:

Opening Ceremony/Opening Session

The multistakeholder approach was highlighted by many speakers and panelists during the Opening Session, including the message from the UN Secretary-General Ban Ki-Moon, which was read by the UN Under-Secretary-General for Economic and Social Affairs, M. Sha Zukang.

M. Ban Ki-Moon assured that it is not a UN goal to take over Internet Governance but the UN will offer an opportunity to bring people together, with the same interest, in a global reach.

M. Sha Zukang concludes that the IGF was a unique experience because *"it brings together people who normally do not meet under the same roof."*

"Development" was a key discussion during the IGF Rio Meeting. It will still be an important aspect for discussion, together with the issue of bridging the digital divide - a key element of discussion for the IGF Hyderabad and reflects the theme of the IGF Hyderabad which is "Internet for All."

The nature and prospective of the IGF were also discussed, as the Chairman properly summarizes:

"Several participants underlined that the IGF was not only a space for dialogue, but also a medium that should encourage fundamental change at the local level to empower communities, build capacity and skills enable the Internet's expansion, thereby contributing to economic and social development."

Critical Internet Resources

This is a new session that was introduced during the IGF Rio Meeting. Basically, it covered some issues pertaining to the infrastructure of the Internet. ICANN discussions were not missed, as well as the role of governments in shaping policies.

Access

The issue of "access" is more on how to get the billion of users around the world to go online in the next years to come. Such initiatives to this cause are reminiscent of pilot projects in Africa wherein laptops were given to children under an open source software agreement.

Diversity

The issue of "diversity" calls for multilingualism in the Net. Such promotion on multilingualism would increase users whose main language is not English. In order to open the Net to a diverse population, international domain names (IDN) were added to facilitate the language needs of other users.

Openness

The strong support on closed software has not been favorable to some people. This is because there were long-lasting agreements between governments and large software companies. Such actions were considered critical, as it binds different entities to proprietary or closed source technologies. Many believed that the shift from closed to open software can only happen with the full-scale participation of both the private and public sectors. As such, many people fear the turning of the Internet into a "private" network if there is much insistence on the use of closed technologies.

Talks on open standards, open architecture and open software are clear indicators of what the issue on openness is all about.

Read this literature entitled "Free Culture" by Lawrence Lessig to know more on "Openness on the Internet."

Security Issues

The question of Internet Security is one of the most important debate in the Information Society. Internet is becoming an important communication and business tool, as such, that the question of security comes as a cross-cutting issue to be addressed in all its dimension. As indicated by Michael Harrop, Rapporteur SG 17 Q4, Communications Security Project in 2006, *"without effective security, all systems and processes that rely on electronic communications are at risk and, as a consequence, large numbers of resources are now devoted to countering threats, protecting systems and recovering from successful attacks."* The Rio de Janeiro Meeting mentioned that *"...achieving the Internet's full potential to support commercial and social relationships required an environment that promoted and ensured users' trust and confidence and provided a stable and secure platform."*

Cyber-security, in this case, focused heavily on child protection, particularly on child pornography. Participants gathered were called to seek ways to harmonize legislative agendas to counter-act such crimes. This was a call of legislation between countries that can work together in order to enforce laws that would protect children. As such that some laws are not applicable online, this call also promoted formulation of legislation that would be applicable in the online or virtual world.

Internet Security has been mentioned in the Substantive Agenda of the Rio de Janeiro Meeting. It was also present in the Agenda of the Athens Meeting. Even before the Athens Meeting, Internet Security was mentioned at the Tunis WSIS under "Building Confidence and Security in the Use of ICT's." At the coming Hyderabad Meeting in December 2008, two panels will again discuss questions related to Internet Security. This gives an idea on how important the question has been in each of the IGF meetings so far.

Internet Security issues can be folded under the following:

- secure telecommunication which deals most with the security of telecommunication infrastructure
- cyber-security as Internet users deal with it in their daily operations and use of the Internet
- identity theft
- children pornography
- hacking and other virus and cyber threats (scams, spams, etc.)
- cyber-terrorism

Internet Security on the Athens Agenda

The International Telecommunication Union (ITU) is at the forefront of contributors to the field of Telecommunication Security. At the Athens meeting, ITU mentioned the major contributions made in this domain by International organizations. ITU took the necessary steps to set up a number of initiatives that were presented at the Athens meeting. It presented a telecommunication security guideline and set up the road map towards Internet Security. The question of security of telecommunication was somehow dominant at the Athens meeting.

ITU mentioned the difficulty of experts in the field. One of the difficulties mentioned was related to the question of standardisation - as many international organizations were developing domain initiatives at the same time.

As a follow-up activity of the WSIS Conference, a number of ITU study groups have been assigned tasks related to Internet Security. At the Athens Meeting, findings of these study groups were presented to address diverse Internet security questions such as:

- Telecommunication management
- Protection against electromagnetic environment effects
- Outside Plant and related indoor installations
- Security, languages and telecommunication software
- Mobile Telecommunications Networks

Internet Security on the Rio de Janeiro Agenda

At the Rio de Janeiro meeting, a whole session was dedicated to the question of Internet security, emphasizing the importance of this question nowadays, as well as the threats, that users are facing more and more in their daily operations over the Internet. Internet Security questions put on the agenda at Rio were related to:

- cybercrime
- cyber-terrorism
- protection of individuals and automatic processing of personal data
- action against trafficking in human beings
- protection of children against sexual exploitation and sexual abuse

The Rio de Janeiro meeting called for international cooperation and coordinated action to counter cybercrime because of its trans-national dimension. Recommendations were forwarded towards the direction of responsibility of governments in order to raise awareness among Internet users and in the direction of ICANN because of the responsibility it has for the Domain Name System. It is required of ICANN since it accepts responsibility for controlling illegal online content for the protection of children from Internet pornography.

Taking Stock and the Way Forward

Emerging Issues

This session aims to identify key issues in Internet Governance that should be addressed in the Forum. The first obstacle was to filter some themes, as there is a variety of interests to be held in such a generic target. There were four themes proposed:

(i) **demand and supply side initiatives** (by Robert Pepper). He brought into debate the economic concept of demand and supply applied to Internet Governance. On the demand side, there were interesting proposals, such as the need for educating through capacity-building Internet users, the ability of people controlling their web ID (part of educating the usage in Internet), local content in local languages (enforcing local community) and improving public policies (but not over regulating, such as prohibiting or limiting access to VoIP, which can suppress the demand). On the supply side, there were the common concern of extending Internet users/access, but also considering *"the opportunities created by the release of spectrum through the switch to digital broadcasting were highlighted. Some speakers suggested that such spectrum could be used to support new broadband networks and support new investment and innovative services, while others held the view that this would not be a sustainable solution."*

(ii) **social, cultural and political issues of Web 2.0** (by Andrew Keen);

(iii) **access** (particularly in Africa, by Nii Quaynor) and

(iv) **innovation, research and development** (by Robert Kahn).

Another challenge was to discuss emerging issues in a global forum with different perspectives, for example, developed and developing countries realities; democratic and non-democratic political regimes; and etc.

III IGF Hyderabad 2008

The third meeting of the IGF was held in Hyderabad, India. The over-all theme for the meeting was "Internet for All." The chairman's summary can be accessed via the official IGF website.

In terms of attendance, there were 1280 participants from 94 countries. The actual breakdown of participants by region can be found here.

Renewal of the Multistakeholder Advisory Group (MAG)

Stakeholders from different sectors - government, civil society, private, academe and technical communities - were invited to submit proposals/nominations for new MAG members. The mandate behind the rotation of its members are based on recommendations

from different sectors. The official IGF website carries the list of updated MAG members.

Remote Participation in the IGF 2008

The Remote Participation Working Group (RPWG) has been working closely with the IGF Secretariat for allowing remote participants across the globe to interact in the meeting. There were 522 remote participants from around the world who joined the main sessions and workshops.

The entire meeting in Hyderabad was webcast in real-time using high quality video, audio streaming and live chat.

Remote Hubs were also introduced with remote moderators leading the discussions in their region. Most of the hubs were able to discuss pertinent local and domestic Internet Governance issues. The Remote Hubs were found in Buenos Aires, Argentina, Belgrade, Serbia, São Paulo (Brazil), Pune (India), Lahore (Pakistan), Bogotà (Colombia), Barcelona and Madrid (Spain).

The platform used for remote participation in Hyderabad was DimDim.

IV IGF Sharm El Sheikh 2009

Egypt hosted the fourth IGF meeting in Sharm el Sheikh from 15–18 November 2009 in Sharm El Sheikh. “Internet Governance – Creating Opportunities for all” is the overall title of the meeting. It marks the beginning of a new multi-stakeholder process.

The main sessions on the agenda points are Managing Critical Internet Resources; Security, Openness and Privacy; Access; Diversity; Internet governance in the Light of WSIS Principles; Taking Stock and the Way forward – on the Desirability of the Continuation of the Forum; and Emerging Issues - Impact of Social Networks.

One key focus of IGF 2009 is encouraging youth participation towards Internet Governance issues.

Remote Participation in the IGF 2009

Following the success of remote participation in the IGF Hyderabad, the Remote Participation Working Group (RPWG) has come up with improved guidelines on intervention for the training of remote moderators. Webex was also used as the platform for this year's remote participation. There are 11 registered remote hubs for this year's meeting and the complete list can be found in the official IGF website.

Chapter 12

InterNIC and Internet Watch Foundation

InterNIC

The **Internet Network Information Center**, known as **InterNIC**, was the Internet governing body primarily responsible for domain name and IP address allocations until September 18, 1998 when this role was assumed by the Internet Corporation for Assigned Names and Numbers (ICANN). It was accessed through the domain name **internic.net**, with email, FTP and World Wide Web services run by Network Solutions, Inc and AT&T.

Term

InterNIC is a registered service mark of the U.S. Department of Commerce. The use of the term is licensed to the Internet Corporation for Assigned Names and Numbers (ICANN).

History

The first central authority to coordinate the operation of the network was the Network Information Center (NIC) at the Stanford Research Institute (SRI) in Menlo Park, California. In 1972, management of network resources was transferred to the newly created Internet Assigned Numbers Authority (IANA). Jon Postel fulfilled the role of manager of IANA, in addition to his role as the RFC Editor, until his death in 1998.

On the ARPANET, hosts were given names to be used in place of numeric addresses and a HOSTS.TXT file was distributed by SRI International and manually installed on each host on the network to provide a mapping between these names and their corresponding network address. As the network grew, this became increasingly cumbersome. A technical solution came in the form of the Domain Name System, created by Paul Mockapetris. The Defense Data Network Network Information Center (DDN-NIC) at SRI handled all registration services, including the top-level domains `mil`, `gov`, `edu`, `org`, `net`, `com` and `us`. DDN-NIC also performed root nameserver administration and Internet number assignments under a United States Department of Defense contract. In 1991, the Defense Information Systems Agency (DISA) awarded the administration and maintenance of DDN-NIC, which had been up until this point under the management of

SRI for many years, to Government Systems, Inc. which subcontracted it to the small private-sector firm Network Solutions, Inc.

Up to this time, most of the growth of the Internet was in the non-military sector. Therefore, it was decided that the Department of Defense would no longer fund registration services outside of the `mil` domain. In 1993, the National Science Foundation of the United States, after a competitive bidding process in 1992, created the Internet Network Information Center, known as *InterNIC*, to manage the allocations of addresses and awarded the contract to three organizations. Registration services were to be provided by Network Solutions, directory and database services were to be run by AT&T, and information services by General Atomics. Later, General Atomics was disqualified from the contract after a review found their services not conforming to the standards of its contract. General Atomics' InterNIC functions were assumed by AT&T. AT&T discontinued InterNIC services after their contract expired.

In 1998 both IANA and InterNIC were reorganized under the control of ICANN, a California non-profit corporation contracted by the US Department of Commerce to manage a number of Internet-related tasks. The role of operating the DNS system was privatized, and opened up to competition, while the central management of name allocations would be awarded on a contract tender basis.

Domain name restrictions

Via internic.net, domain names were distributed through an automated system. Beginning in 1996, Network Solutions began restricting the distribution of domain names containing a number of words on a "restricted list" through an automated filter. The filter is known to have rejected domain names containing the "least agreeable words in the English language" Applicants whose domain names were rejected would receive a form email containing the notice: "Network Solutions has a right founded in the First Amendment to the U.S. Constitution to refuse to register, and thereby publish, on the Internet registry of domain names words that it deems to be inappropriate."

This filter came under heavy scrutiny, as legitimate domain names such as "shitakemushrooms.com" would be rejected, but the domain name "shit.com" was active, as it had been registered before 1996. Network Solutions eventually allowed domain names containing the words on a case-by-case basis, after manually reviewing the names for obscene intent. This profanity filter was never enforced by the government and its use was not continued by ICANN when it took over governance of the distribution of domain names to the public.

Internet Watch Foundation

Internet Watch Foundation

Type	Registered charity
Founded	1996
Employees	14 (2007)

The **Internet Watch Foundation (IWF)** is a non-governmental charitable body based in the United Kingdom. It offers an online service for the public and IT professionals to report content on the Internet that it considers to be "potentially illegal". As part of its function, the IWF produces a blacklist of Internet sites and content that it deems to be in contravention/potentially in contravention to UK laws. Since 2010, blocking Internet users from accessing the content on this list is mandatory for all UK based ISPs that want to be eligible for contracts with government agencies and other public bodies.

The IWF operates in informal partnership with the police, government, public and Internet service providers. Originally formed to police suspected child pornography online, the IWF's remit was later expanded to cover racist and criminally obscene material.

The IWF is an incorporated charity, limited by guarantee, and largely funded by voluntary contributions from UK communications service providers, including ISPs, mobile phone operators, Internet trade associations, search engines, hardware manufacturers, and software providers. It also receives funding from the Association for Payment Clearing Services and the European Union.

The IWF is governed by a Board of Trustees which consists of an independent chair, six non-industry representatives, and three industry representatives. The Board monitors and reviews IWF's remit, strategy, policy and budget to enable the IWF to achieve its objectives. The IWF operates from offices in Oakington, near Cambridge.

History

Background

During 1996 the Metropolitan Police told the Internet Service Providers Association (ISPA) that the content carried by some of the newsgroups made available by them was illegal, that they considered the Internet Service Providers (ISPs) involved to be publishers of that material, and that they were therefore breaking the law. In August 1996, Chief Inspector Stephen French, of the Metropolitan Police Clubs & Vice Unit, sent an open letter to the ISPA, requesting that they ban access to a list of 132 newsgroups, many of which were deemed to contain pornographic images or explicit text.

This list is not exhaustive and we are looking to you to monitor your newsgroups identifying and taking necessary action against those others found to contain such material. As you will be aware the publication of obscene articles is an offence. This list is only the starting point and we hope, with the co-operation and assistance of the industry and your trade organisations, to be moving quickly towards the eradication of this type of newsgroup from the Internet ... We are very anxious that all service providers should be taking positive action now, whether or not they are members of a trade association. We trust that with your co-operation and self regulation it will not be necessary for us to move to an enforcement policy.
—Chief Inspector Stephen French, quoted in *Web Control*

The list was arranged so that the first section consisted of unambiguously titled paedophile newsgroups, then continued with other kinds of groups which the police wanted to restrict access to, including *alt.binaries.pictures.erotica.cheerleaders* and *alt.binaries.pictures.erotic.centerfolds*.

Although this action had taken place without any prior debate in Parliament or elsewhere, the police, who appeared to be doing their best to create and not simply to enforce the law, were not acting entirely on their own initiative. Alan Travis, Home Affairs editor of the newspaper *The Guardian*, explained in his book "Bound and Gagged" that Ian Taylor, the Conservative Science and Industry Minister at the time, had underlined an explicit threat to ISPs that if they did not stop carrying the newsgroups in question, the police would act against any company that provided their users with "pornographic or violent material". Taylor went on to make it clear that there would be calls for legislation to regulate all aspects of the Internet unless service providers were seen to wholeheartedly "responsible self-regulation".

Demon Internet regarded the police request as "unacceptable censorship"; however, its attitude annoyed ISPA chairman Shez Hamill, who said:

We are being portrayed as a bunch of porn merchants. This is an image we need to change. Many of our members have already acted to take away the worst of the Internet.

But Demon have taken every opportunity to stand alone in this regard. They do not like the concept of our organisation.
—*Observer*, 25 August 1996

Following this, a tabloid-style exposé of ISP Demon Internet appeared in the *Observer* newspaper, which alleged that Clive Feather (a director of Demon) "provides paedophiles with access to thousands of photographs of children being sexually abused".

During the summer and autumn of 1996 the UK police made it known that they were planning to raid an ISP with the aim of launching a test case regarding the publication of obscene material over the Internet. The direct result of the campaign of threats and pressure was the establishment of the Internet Watch Foundation (initially known as the Safety Net Foundation) in September 1996.

Foundation of IWF

Facilitated by the Department of Trade & Industry (DTI), discussions were held between certain ISPs, the Metropolitan Police, the Home Office, and a body called the "Safety Net Foundation" (formed by the Dawe Charitable Trust). This resulted in the "R3 Safety Net Agreement", where "R3" referred to the triple approach of rating, reporting, and responsibility. In September 1996, this agreement was made between the ISPA, LINX, and the Safety Net Foundation, which was subsequently renamed the Internet Watch Foundation. The agreement set requirements for associated ISPs regarding identifiability and traceability of Internet users; ISPs had to cooperate with the IWF to identify providers of illegal content and facilitate easier traceability.

Demon Internet was a driving force behind the IWF's creation, and one of its employees, Clive Feather, became the IWF's first chair of the Funding Board and solicitor Mark Stephens the First Chair of the IWF's Policy Board. The Policy Board developed codes, guidance, operational oversight and a hotline for reporting content.

The Funding Board, made up of industry representatives and Chair of Policy Board, provided the wherewithall for the IWF's day to day activities as set down and required by the Policy Board.

After 3 years of operation, the IWF was reviewed for the DTI and the Home Office by consultants KPMG and Denton Hall. Their report was delivered in October 1999 and resulted in a number of changes being made to the role and structure of the organisation, and it was relaunched in early 2000, endorsed by the government and the DTI, which played a "facilitating role in its creation", according to a DTI spokesman.

At the time, Patricia Hewitt, then Minister for E-Commerce, said: "The Internet Watch Foundation plays a vital role in combating criminal material on the Net." To counter accusations that the IWF was biased in favour of the ISPs, a new independent chairman was appointed, Roger Darlington, former head of research at the Communication Workers Union.

The website

The IWF's website offers a web-based government-endorsed method for reporting suspect online content and remains the only such operation in the United Kingdom. It acts as a Relevant Authority in accordance with the Memorandum of Understanding concerning Section 46 of the Sexual Offences Act 2003 (meaning that its analysts will not be prosecuted for looking at illegal content in the course of their duties). Reports can be submitted anonymously. The IWF aims to minimise the availability of potentially illegal Internet content, specifically:

- Indecent images of under-18s hosted anywhere in the world;
- criminally obscene content hosted in the UK, or anywhere in the world if uploaded by a British citizen (under the Obscene Publications Acts);
- *incitement to racial hatred content hosted in the UK*

However, almost the whole of the IWF site is concerned with suspected child pornography with little mention of the rest of their remit (racial hatred and criminally obscene material). Images judged by the IWF to be child pornography are blocked, whilst other possibly illegal content is reported to the police for further action.

The Government claimed that they would also be handling images of adult "extreme pornography" which are now illegal for UK citizens to possess as of 26 January 2009. The IWF now includes "extreme pornography" as an example under "criminally obscene content", meaning that they will report material hosted in the UK, or uploaded by a British citizen, but has stated that it has no plans to block any such material, or handle sites hosted outside on the UK.

The IWF states that it works in partnership with UK Government departments such as the Home Office and the Department for Business, Enterprise and Regulatory Reform to influence initiatives and programmes developed to combat online abuse.

They are funded by the European Union and the online industry. This includes Internet service providers, mobile operators and manufacturers, content service providers, telecommunications and filtering companies, search providers and the financial sector as well as blue-chip and other organisations who support the IWF for corporate social responsibility reasons.

Through their "Hotline" reporting system, the organisation helps ISPs to combat abuse of their services through a "notice and take down" service by alerting them to any potentially illegal content within their remit on their systems and simultaneously invites the police to investigate the publisher.

The IWF has connections with the Virtual Global Taskforce, the Serious Organised Crime Agency and the Child Exploitation and Online Protection Centre.

Management

Peter Robbins OBE, QPM is IWF Chief Executive

Sarah Robertson is IWF Director of Communications

Fred Langford is IWF Director of Technology and Content

Cross-border aspects

Previously, the IWF passed on notifications of suspected child pornography hosted on non-UK servers to the UK National Criminal Intelligence Service which in turn forwards it to Interpol or the relevant foreign police authority. It now works with the Serious Organised Crime Agency instead. The IWF does not, however, pass on notifications of other types of illegal content hosted outside the UK.

Blacklist

The IWF compiles and maintains a blacklist, mainly of what it considers child pornography URLs, from which 95% of commercial Internet customers in the UK are filtered. A staff of four police-trained analysts are responsible for this work, and the director of the service has claimed that the analysts are capable of adding an average of 65-80 new URLs to the list each week, and act on reports received from the public rather than pursuing investigative research.

Between 2004 and 2006, BT Group introduced its Cleanfeed technology which was then used by 80% of internet service providers. BT spokesman Jon Carter described Cleanfeed's function as "to block access to illegal Web sites that are listed by the Internet Watch Foundation", and described it as essentially a server hosting a filter that checked requested URLs for Web sites on the IWF list, and returning an error message of "Web site not found" for positive matches.

In 2006, Home Office minister Alan Campbell pledged that all ISPs would block access to child abuse websites by the end of 2007. By the middle of 2006 the government reported that 90% of domestic broadband connections were either currently blocking or had plans to by the end of the year. The target for 100% coverage was set for the end of 2007, however in the middle of 2008 it stood at 95%. In February 2009, the Government said that it is looking at ways to cover the final 5%. In an interview in March 2009, a Home Office spokesperson mistakenly thought that the IWF deleted illegal content, and didn't look at the content they rate.

Although the IWF's blacklist causes content to be censored even if the content has not been found to be illegal by a court of law, IWF Director of Communications Sarah Robertson claimed, on 8 December 2008, that the IWF is opposed to the censorship of legal content.

In March 2009 a Home Office spokesperson said that ISPs were being pressured to sign up to the IWF's blacklist in order to block child pornography websites and said that there was no alternative to using the IWF's blacklist. One of the ISPs which refused to subscribe to the blacklist, Zen Internet, has said that it has "concerns over its effectiveness".

As of 2009, the blacklist was said to contain about 450 URLs. A 2009 study by researcher Richard Clayton at the University of Cambridge found that about a quarter of them were on (otherwise) legitimate free file hosting services, among them RapidShare, Megaupload, SendSpace and Zshare. According to the *Times*, the list contained "between 500 and 800 websites" as of March 2010, and was updated two times per day.

It appears, around July/August 2010, Megaupload, and Megavideo were added to the blacklist again. Access to these sites via some exchange routers used by O2 broadband is restricted. A few members appear to be blocked with no way of appealing this decision.

Incidents

Sex stories

On 26 July 2007, UK tabloid newspaper *The Daily Star* reported that it had discovered an online text story about British pop group Girls Aloud that it described as "a chilling story detailing each singer's gory death in scenes that could be straight out of a horror movie", characterizing its author as "a vile internet psycho" and "a cyber-sicko". The news story said that *The Daily Star* had reported the content of the hosting website, "Kristen Archives" (a subsite of the ASSTR archive), to the IWF, and that the IWF had traced the site to the US. It also claimed that Interpol had been notified to help track down the site's operators and the writer of the story. An IWF spokesperson was reported as saying that since the site was hosted in the US, it fell outside the organization's remit, but that they were aware of the site. The spokesperson added that the site also contained "child abuse fantasy stories" and that they had passed on details of it to the British police.

Although the story, entitled "Girls (Scream) Aloud", had been published on a US website, British police carried out the investigation because the alleged author was identified as living in the UK. Although he had submitted the story under a pseudonym, he included an email address which was reportedly traced. Officers from Scotland Yard's Obscene Publications Unit decided to take action over the story after consulting the Crown Prosecution Service (CPS), and on 25 September 2008 it was announced that the author, Darryn Walker, was to be prosecuted for the online publication of material that the police and the CPS believed was obscene. It was the first such prosecution for written material in nearly two decades, and was expected to have a significant impact on the future regulation of the Internet in the UK.

Walker appeared in court on 22 October 2008 to face charges of "publishing an obscene article contrary to Section 2(1) of the Obscene Publications Act 1959". He was granted unconditional bail, and his case was set for trial on 16 March 2009. However, at a

directions hearing in January, the defendant made it known that given the seriousness of the case he would be represented by a QC (Queen's Counsel), following which the Crown Prosecution Service gave notice of its intention to similarly employ a QC, and the trial date was put back to 29 June 2009, where the defendant was found not guilty, and cleared of all charges of obscenity.

Wayback Machine

On 14 January 2009 some UK users reported that all of the 85 billion pages of the Internet Archive (Wayback Machine) had been blocked, in spite of the fact that the IWF's policy is to try to only censor the exact webpage in question and not the whole domain. According to IWF chief executive Peter Robbins this happened due to a "technical hitch". Because the Internet Archive's web site contained URLs on the IWF's blacklist, requests sent there from the ISP Demon Internet carried a particular header, which clashed with the Internet Archive's internal mechanism to convert web links when serving archived versions of web pages. The actual blocked URL which had caused the incident never became publicly known.

Criticism

Charity status

In February 2009 a Yorkshire-based software developer lodged a formal complaint regarding the IWF status as a charity with the Charity Commission, in which he pointed out that "regulating the worst of the internet" was "not really a charitable purpose", and that the IWF existed mainly to serve the interests of ISPs subscribing to it rather than the public. An IWF spokesperson said that the IWF had attained charitable status in 2004 "in order to subject itself to more robust governance requirements and the higher levels of scrutiny and accountability which charity law, alongside company law, brings with it". The IWF is listed by fakecharities.org, "a directory of those so-called charities that receive substantial funding from either the UK or EU governments". It has also been termed a quango by critics, implying poor management and lack of accountability.

False positives

Following the IWF's blacklisting of the article, the organisation's operating habits came under scrutiny. J.R. Raphael of PC World stated that the incident had raised serious free-speech issues, and that it was alarming that one non-governmental organisation was ultimately acting as the "morality police" for about 95% of UK's Internet users. Frank Fisher of *The Guardian* criticized the IWF for secretiveness and lack of legal authority, among other things, and noted that the blacklist could contain anything and that the visitor of a blocked address may not know if their browsing is being censored.

Forced adoption

The government believes that a self-regulatory system is the best solution, and the Metropolitan Police also believe that working with ISPs, rather than trying to force them via legislation, is the way forward. The IWF has a blacklist of URLs which is available to ISPs, but ISPs are not forced to subscribe to it. However, ISPs may feel inclined or even forced to join (and contribute) to the IWF's activities as a failure to do so may harm their reputation as responsible providers. Subscribing to the IWF may also be seen as a marketing tool by ISPs.

Legality

As a "self-appointed, self-regulated internet watchdog, which views user-submitted content and compiles a list of websites that it deems to contain illegal images" there have been questions raised regarding the legality of their viewing content that would normally constitute a criminal offense.

Secrecy

The IWF has been criticized for blacklisting legal content and for not telling websites that they are being blocked and also for not making their blocked website list public.

Technical issues

In addition to introducing performance problems the blacklisting of sites may be concealed by generic HTTP 404 "file not found" errors rather than a more appropriate HTTP 403 "forbidden" message; it should be noted, however, that the exact method of censorship is completely reliant on the implementing ISP; BT, for example, return 404 pages, whereas Demon return an honest message stating that, and why, the page is censored.

List of IWF filter servers on each internet provider network

By doing a traceroute to a particular website you can see the path it takes to see if it does go through the internet service providers invisible IWF filter. The following list will show the server to look for in the traceroute and be able to determine whether the website is blacklisted.

Chapter 13

Legal Status of Internet Pornography

Due to the international nature of the Internet, Internet pornography carries with it special issues with regard to the law. There is no one set of laws that apply to the distribution, purchase, or possession of Internet pornography. Only the laws of one's home nation apply with regard to distributing or possessing Internet pornography. This means that, for example, even if a pornographer is legally distributing pornography, the person receiving it may not be legally doing so due to local laws.

Areas of legal concern within many countries

Some areas of legal concern regarding adult pornography are:

- Prohibiting certain or all types of pornography that are illegal within a government's jurisdiction. For countries that do not prohibit all pornography, this might include pornography featuring violence or bestiality, for example.
- Preventing those under the legal age (for most this means a minor under 18 or 21) from accessing pornographic content.
- Enforcing laws designed to ensure that performers in pornography are of legal age.

In jurisdictions that heavily restrict access or outright ban pornography, various attempts have been made to prevent access to pornographic content. The mandating of Internet filters to try preventing access to porn sites has been used in some nations such as China and Saudi Arabia. Banning porn sites within a nation's jurisdiction does not necessarily prevent access to that site, as it may simply relocate to a hosting server within another country that does not prohibit the content it offers.

Many nations that allow at least some types of pornography attempt to ensure that those under their legal age for accessing porn (often 18 or 21) cannot easily access it. Various measures have been tried but with varying success. Within the United States, most websites have taken voluntary steps to ensure that visitors to their sites are not underage. Many Web sites provide a warning upon entry, warning minors and those not interested in viewing porn not to view the site, and requiring one to affirm that one is at least 18 and wishing to view pornographic content. Such warning pages have little effect in preventing access by minors to porn, as any minor interested in viewing the site can simply click on the "I am an adult over 18" button without having to prove his or her age.

Thus, such warnings are generally not used by themselves but with other techniques. Commercial porn sites generally do not restrict access to any pornographic content until a membership has been purchased using a credit card, as most have explicit 'free trial' content as a major part of their sales strategy. So-called age verification services have also sprung up that offer access to any Web site that participates in their program without additional charge. The users need only verify their age with the verification service, which then issues a username and password that can access all sites that use its services. Most age verification sites charge either a monthly or yearly fee to those wanting access to participating sites.

Within nations that allow at least some types of pornography, models are often required to be at least a specific age (18 is most common). Various nations have various rules as to how a site must ensure that all porn models featured on it are of age such as strict record-keeping laws.

Child pornography and the Internet

According to the United States organization The National Center for Missing and Exploited Children (NCMEC) and other international sources, child pornography is a multi-billion dollar industry and among the fastest growing criminal segments on the Internet. According to the NCMEC, approximately one fifth of all Internet pornography is child pornography.

Child pornography is illegal in most countries with coordinated enforcement by Interpol and policing institutions of various governments, including among others the United States Department of Justice. Even so, the UK based NSPCC said that worldwide an estimated 2% of websites still had not been removed a year after being identified. Recent investigations include Operation Cathedral that resulted in multi-national arrests and 7 convictions as well as uncovering 750,000 images with 1,200 unique identifiable faces being distributed over the web; Operation Amethyst which occurred in the Republic of Ireland; Operation Auxin; Operation Avalanche; Operation Ore based in the United Kingdom; Operation Pin; Operation Predator; the 2004 Ukrainian child pornography raids and the 2008 US child pornography raid. New technology that aids those who produce this material include inexpensive digital cameras and Internet distribution has made it easier than ever before to produce and distribute child pornography. The producers of child pornography try to avoid prosecution by distributing their material across national borders, though this issue is increasingly being addressed with regular arrests of suspects from a number of countries occurring over the last few years.

The legal status of simulated or "virtual" child pornography varies around the world; for example, it is legal in the United States, it is illegal in the European Union, and in Australia its legal status is unclear and so far untested in the courts. Child pornography may be simulated by the use of computers or adults made to look like children.

In 2008, it was discovered that the United States will post fake hyperlinks claiming to be child pornography and then raiding, arresting, and prosecuting anyone who was found

using the IP address that visited them, even someone whose computer was an open wifi. In 2008, a man in Middlesbrough was found guilty of downloading "child pornography" when he downloaded computer generated cartoons.

Internet pornography laws in various countries

United States

With the exception of child pornography, the legal status of accessing Internet pornography is still somewhat unsettled, though many individual states have indicated that the creation and distribution of adult films and photography are legally listed as prostitution within them.

The legality of pornography at the federal level has been traditionally determined by the Miller test, which dictates that community standards are to be used in determining whether a piece of material is obscene. Thus, if a local community determines a pornographic work to meet its standard for obscenity then it could be banned. This means that a pornographic magazine that might be legal in California could be illegal in Alabama. This standard poses a problem when it comes to the Internet because restricting the communities some pornographic material is available in is much more difficult over the Internet. It has been argued that if the Miller test were applied to the Internet then, in effect, the community standards for the most conservative community would become the standard for all U.S.-based Web sites. The courts are currently examining this issue.

The first attempt to regulate pornography on the Internet was the federal Communications Decency Act of 1996, which prohibited the "knowing" transmission of "indecent" messages to minors and the publication of materials which depict, in a manner "patently offensive as measured by contemporary community standards, sexual or excretory activities or organs", unless those materials were protected from access by minors, for example by the use of credit card systems. Immediately challenged by a group of organizations spearheaded by the ACLU, both of these provisions were struck down by the U.S. Supreme Court in Reno v. American Civil Liberties Union (1997). The "indecent transmission" and "patently offensive display" provisions were ruled to limit the freedom of speech guarantee of the First Amendment.

A second attempt was made with the narrower Child Online Protection Act (COPA) of 1998, which forced all *commercial* distributors of "material harmful to minors" to protect their sites from access by minors. "Material harmful to minors" was defined as materials that by "contemporary community standards" are judged to appeal to the "prurient interest" and that show sexual acts or nudity (including female breasts). Several states have since passed similar laws. An injunction blocking the federal government from enforcing COPA was obtained in 1998. In 1999, the 3rd Circuit Court of Appeals upheld the injunction and struck down the law, ruling that it was too broad in using "community standards" as part of the definition of harmful materials. In May 2002, the Supreme Court reviewed this ruling, found the lower court's given reason insufficient and returned the case to the circuit court. In March 2003, the 3rd Circuit Court again struck down the law

as unconstitutional, this time arguing that it would hinder protected speech among adults. The administration appealed; in June 2004 the Supreme Court upheld the injunction against the law, ruling that it was most likely unconstitutional but that a lower court should determine whether newer technical developments could have an impact on this question. On March 22, 2007, COPA was found to violate the First and Fifth Amendments of the United States Constitution and was struck down.

Another act intended to protect children from access to Internet pornography was the Children's Internet Protection Act (CIPA) of 2000. It required that public libraries, as a condition of receiving federal subsidies for Internet connectivity, employ filtering software to prevent patrons from using Internet terminals to view images of obscenity and child pornography, and to prevent children from viewing images "harmful to minors", a phrase typically used for otherwise legal pornography. The act allowed librarians to disable the filtering software for adult patrons with "bona-fide research or other lawful purposes". The act was challenged by the American Library Association on First Amendment grounds, and enforcement of the act was blocked by a lower court. In June 2003, the Supreme Court reversed and ruled that the act was constitutional and could go into effect.

The production of sexually explicit materials is regulated under 18 U.S.C. 2257, requiring "original" producers to retain records showing that all performers were over the age of 18 at the time of the production for inspection by the Attorney General. The 18 U.S.C. 2257 disclaimer is common on Internet sites distributing pornography, but the Department of Justice has rarely if ever enforced the provision. Although the law had been on the books for over 10 years, the Justice Department never actually inspected anyone. It was not until pressure from Congress, and right-wing religious groups spurred the Administration of George W. Bush and Attorney General Gonzales to begin inspections of larger commercial porn companies primarily in the Los Angeles area. Despite fearing mass inspections, harassment and prosecution, the Justice department inspected less than two dozen companies (out of several thousand operating) and no prosecutions resulted from any of the inspections. The inspections were conducted by retired FBI agents, and according to porn executives agents were always courteous and professional, and agents suggested changes or modifications to the companies record keeping process. Agents generally arrived with a list of films which they wanted to inspect the records for, most likely to avoid potential 4th amendment conflicts on issues of probable cause. Once Attorney General Alberto Gonzales departed the Justice Department, the inspections ended.

On July 1, 2005, new regulations took effect requiring among other things, "secondary" producers to retain the same records. This has been seen both as a prelude to increased inspection of records by the Department of Justice, and also as a potential assault on the Internet pornography industry by increasing the burden of compliance for distributors.

On Oct. 24, 2007 the Sixth Circuit court of appeals in Ohio, issued a judgment against the 2257 law, ruling it as unconstitutional according to the first amendment, however the Sixth Circuit subsequently reheard the case *en banc* and issued an opinion on February

20, 2009, upholding the constitutionality of the record-keeping requirements, albeit with some dissents. The Sixth Circuit en banc decision was appealed to the US Supreme Court where on Monday October 5, 2009, the US Supreme Court denied certiorari without comment not addressing the Sixth Circuit decision that 18 USC 2257 is not constitutionally "vague and overbroad" and able to be enforced.

New York sentenced ISP, BuffNET, after they plead guilty to fourth-degree criminal facilitation for not stopping child pornography after being asked to remove it.

United Kingdom

The sale or distribution of hardcore pornography through any channel was prohibited until the rules were relaxed in 2002, however the rules are still quite strict . The possession of pornographic images for private use has never been an offence in the UK. This means that UK citizens have always been able to access content on sites overseas without breaking any laws, except for child pornography.

Adult pornography that falls under the Government's classification of "extreme pornography" is illegal to possess as of January 26, 2009, carrying a three year prison sentence. This was proposed by the Government after the murder of Jane Longhurst, claiming that such material was viewed by murderer Graham Coutts. Critics of the law point out that the law will criminalise images of legal acts between consenting adults and have criticised the lack of evidence of a link between viewing such material, and violent crime. The perils behind the law are debated in the 2010 documentary Hanging Perverts.

Internet service providers started the Internet Watch Foundation in 1996 to watch for pornographic content that is in violation of British law and report it to the police. The web filter Cleanfeed is used by the largest ISP BT Group to block sites on the IWF's list which includes sites that are "criminally obscene" as well as child pornography.. The government ordered all ISPs to have a cleanfeed system by the end of 2007.

Australia

Internet pornography in Australia is subject to a multifaceted regulatory framework. Criminal legislation is in force at the Commonwealth, state and territory levels targeting those involved in the production, dissemination and consumption of illegal internet pornography (including online child abuse pornography and online pornography featuring adults portrayed as children).

It is illegal for internet content providers within Australia to 'broadcast' internet pornography classified as MA15+ to R18+ unless such internet pornography is subject to an age verification system or internet pornography which may be classified as X18+ to RC content that is not subject to an ACMA infringement notice through exceptions.

Under an internet filter, proposed by Sen. Stephen Conroy, internet pornography hosted outside Australia classified by the ACMA under the Classification Board legislation will

be blocked if such internet pornography is deemed by the AMCA to be refused classification (RC), or 'potentially' refused classification. Refused classification (RC) does include real child abuse internet pornography and bestiality internet pornography, however it may also include content discussing or illustrating examples of internet pornography (including both, illegal internet pornography and internet pornography featuring adults portrayed as children) which may limit discussion and debate to authorised statutory persons only, rather than open and free public debate.

Criminal legislation is complemented by a further tier of regulation which provides a range of administrative remedies designed to deal with the availability of inappropriate content by removing it from the internet or by blocking access to it.

Online content scheme

Since January 1999, internet pornography considered offensive or illegal has been subject to a statutory scheme administered by Australia's media regulator, the Australian Communications and Media Authority (ACMA).

Established under Schedule 5 to the *Broadcasting Services Act 1992*, the online content scheme evolved from a tradition of Australian content regulation in broadcasting and other entertainment media. This tradition embodies the principle that – while adults should be free to see, hear and read what they want – children should be protected from material that may be unsuitable for (or harmful to) them, and everyone should be protected from material that is highly offensive.

The online content scheme seeks to achieve these objectives by a number of means such as complaint investigation processes, government and industry collaboration, and community awareness and empowerment. While administration of the scheme is the responsibility of ACMA, the principle of 'co-regulation' underpinning the scheme reflects parliament's intention that government, industry and the community each plays a role in managing internet safety issues in Australia.

Investigations into internet pornography

A central feature of the online content scheme is the complaints mechanism that allows members of the Australian public to submit complaints to ACMA about offensive and illegal internet content.

Offensive and illegal internet content will be 'prohibited' under the scheme if it meets certain classification thresholds, irrespective of where the content is hosted. If prohibited content is hosted in Australia, ACMA will direct the internet content host to remove the content from its service. If prohibited content is not hosted in Australia, ACMA will notify the content to the suppliers of accredited filters in accordance with the Internet Industry Association's internet content code of practice so that access to that content is blocked for users of those filters.

In addition, sufficiently serious internet content (for example, illegal material such as child pornography) will be referred by ACMA under specialized agreements to the appropriate law enforcement agency, or, where appropriate, to a fellow member of the Internet Hotline Providers' Association (INHOPE).

Between January 2000 and June 2006, ACMA received over 5,000 complaints from the public about offensive and illegal internet content hosted in Australia and overseas, resulting in the removal or blocking of almost 4,000 individual items of online content. Approximately 60% of such content was also referred to law enforcement agencies on the basis that it related to material classifiable as 'RC' (see below).

Classification of internet pornography

Internet pornography will be 'prohibited' by ACMA if certain classification thresholds are met. These thresholds form part of the National Classification Scheme (which also applies to other forms of media such as publications, films and video games) and are agreed by the Attorneys-General of the Commonwealth, States and Territories.

The thresholds are articulated in a National Classification Code and in Guidelines. The Classification Board (part of the Attorney-General's Department) is Australia's official classification body. In the course of investigating potentially prohibited internet content, ACMA may seek a formal classification decision from the Classification Board, or it may make its own assessment of the content against the National Classification Code and in Guidelines.

In summary, the following categories of internet content are prohibited: • Content classifiable as 'RC' ('refused classification'). Such content includes, for example, illegal material (such as child sexual abuse material) and other highly offensive material (such as bestiality). • Content classifiable as 'X18+'. Such content includes material containing real depictions of actual sexual activity. • Content hosted in Australia which is classified 'R18+' and not subject to a restricted access system which complies with criteria determined by ACMA. Content classified R18+ is not considered suitable for minors. Such content includes, for example, material containing implied (or simulated) sexual activity.

Internet pornography will be prohibited if it falls within the 'RC' or 'X18+' classifications or, for content hosted in Australia that is not restricted by an adult verification procedure, if it falls within the 'R18+' classification.

Indonesia

The legal situation in Indonesia tightened sharply in 2008 with the passing of the Bill against Pornography and Pornoaction. Law books of Indonesia KUHP (Kitab Undang-Undang Hukum Perdata) article number 282 says that "it is forbidden to spread pornographic content". But there have been Indonesian pornographic pay sites with Indonesian nude models that exploit legal loopholes.

Hong Kong

Pursuant to the Control of Obscene and Indecent Articles Ordinance (Cap 390), it is an offence to publish an obscene article. Publication covers distribution, circulation, selling, hiring, giving, or lending the obscene article. Distribution by email would fall within the definition of distribution, as would the placing of an obscene article on a web site. It should also be noted that distribution does not require any element of financial gain to be present. The definition of article includes "anything consisting of or containing material to be read or looked at or both read and looked at, any sound recording, and any film, video-tape, disc or other record of a picture or pictures." The article will be considered obscene if, by reason by its obscenity, "it is not suitable to be published by any person." Obscenity includes "violence, depravity and repulsiveness". The penalty for this offence is up to three years imprisonment and a fine of up to HK$1,000,000.

Related cases:

- On January 27, 2008, The Hong Kong Police Force arrested suspects who were accused of uploading pornographic images after a multi-billion entertainment company filed a complaint about these photos available on the internet having been fabricated and might charge the offender for defamation.

Moreover, the Prevention of Child Pornography Ordinance, Cap.579, was enacted to deal with the problems associated with child pornography in Hong Kong. Under Section 3, dealing in any of the following manners with child pornography, such as “prints, makes, produces, reproduces, copies, imports or exports”; “publishes” or “has in his possession” is an offence. A child is a person under the age of 16. “Child pornography” means a photograph, film, computer-generated image or other visual depiction that is a pornographic depiction of a child. “Pornographic depiction” means a visual depiction that depicts a person as being engaged in explicit sexual conduct, whether or not the person is in fact engaged in such conduct; or a visual depiction that depicts, in a sexual manner or context, the genitals or anal region of a person or the breast of a female person.

Singapore

The Media Development Authority, a government-run agency in Singapore, blocks a "symbolic" number of websites containing "mass impact objectionable" material, including Playboy, YouPorn, and Sex.com. In addition, the Ministry of Education, Singapore blocks access to pornographic websites.

Chapter 14

Internet Assigned Numbers Authority

The **Internet Assigned Numbers Authority** (**IANA**) is the entity that oversees global IP address allocation, autonomous system number allocation, root zone management in the Domain Name System (DNS), media types, and other Internet Protocol-related symbols and numbers. IANA is operated by the Internet Corporation for Assigned Names and Numbers, also known as ICANN.

Prior to the establishment of ICANN for this purpose, IANA was administered primarily by Jon Postel at the Information Sciences Institute of the University of Southern California, under a contract USC/ISI had with the United States Department of Defense, until ICANN was created to assume the responsibility under a United States Department of Commerce contract.

Responsibilities

IANA is broadly responsible for the allocation of globally-unique names and numbers that are used in Internet protocols that are published as RFC documents. These documents describe methods, behaviors, research, or innovations applicable to the working of the Internet and Internet-connected systems. IANA also maintains a close liaison with the Internet Engineering Task Force (IETF) and RFC Editorial team in fulfilling this function.

In the case of the two major Internet namespaces, namely IP addresses and domain names, extra administrative policy and delegation to subordinate administrations is required because of the multi-layered distributed use of these resources.

IP addresses

IANA delegates allocations of IP address blocks to regional Internet registries (RIRs). Each RIR allocates addresses for a different area of the world. Collectively the RIRs have created the Number Resource Organization formed as a body to represent their collective interests and ensure that policy statements are coordinated globally.

The RIRs divide their allocated address pools into smaller blocks and delegate them in their respective operating regions to Internet service providers and other organizations. Since the introduction of the CIDR system, IANA typically allocates address space in the size of /8 prefix blocks for IPv4 and /12 prefix blocks from the 2000::/3 IPv6 block to requesting regional registries as needed.

Domain names

IANA administers the data in the root nameservers, which form the top of the hierarchical DNS tree. This task involves liaising with top-level domain operators, the root nameserver operators, and ICANN's policy making apparatus.

ICANN also operates the .int registry for international treaty organizations, the .arpa zone for Internet infrastructure purposes, including reverse DNS service, and other critical zones such as root-servers.org.

Protocol parameters

IANA administers many parameters of IETF protocols. Examples include the names of Uniform Resource Identifier (URI) schemes and character encodings recommended for use on the Internet. This task is undertaken under the oversight of the Internet Architecture Board, and the agreement governing the work is published in RFC 2860.

Oversight

IANA is managed by the Internet Corporation for Assigned Names and Numbers (ICANN) under contract to the United States Department of Commerce (DOC). The Department of Commerce also provides an ongoing oversight function, whereby it verifies additions and changes made in the root to ensure IANA complies with its policies.

On January 28, 2003 the Department of Commerce, via the Acquisition and Grants Office of the National Oceanic and Atmospheric Administration, issued a notice of intent to grant ICANN the IANA contract for three more years. It invited alternative offerors to submit in writing a detailed response on how they could meet the requirements themselves. Such responses were to be received no later than 10 days following publication of the invitation and the decision on whether to open the "tender" to competition was to remain solely within the discretion of the government.

In August 2006, the U.S. Department of Commerce extended its IANA contract with ICANN by an additional five years, subject to annual renewals.

Since ICANN is managing a worldwide resource, but being controlled by U.S. interests, a number of proposals have been brought forward to decouple the IANA function from ICANN. However, some believe that it would be impractical to change the current control structure without risking fracturing the Internet.

History

IANA was established informally as a reference to various technical functions for the ARPANET, that the Information Sciences Institute performed for the Defense Advanced Research Project Agency (DARPA) of the United States Department of Defense.

On March 26, 1972, Vint Cerf and Jon Postel called for establishing a socket number catalog in RFC 322. Network administrators were asked to submit a note or place a phone call, "*describing the function and socket numbers of network service programs at each HOST*". This catalog was subsequently published as RFC 433 in December 1972. In it Postel first proposed official assignments of port numbers to network services and suggested a dedicated administrative function, which he called a *czar*, to maintain a registry.

The first reference to the name "IANA" in the RFC series is in RFC 1060, published in 1990, but the function, and the term, was well established long before that; RFC 1174 says that "Throughout its entire history, the Internet system has employed a central Internet Assigned Numbers Authority (IANA)...", and RFC 1060 lists a long series of earlier editions of itself, starting with RFC 349.

In 1996 the "DNS Wars" began as the FNAC ordered the NSF to instruct its contractor, Network Solutions who ran the Internic project, to begin charging for com/net/org domain names. There was widespread dissatisfaction with this concentration of power (and money) in one company and people looked to IANA for a solution. Postel wrote up a draft on the creation of new top level domains.

USC/ISI would not back Postel in the legal sense and IANA, which was a part time "task" had no legal personality - it could not sign contracts - and there was some resentment in the community at paying IANA large sums of money to add one or two lines to the legacy root zone. Jon was trying to institutionalize IANA.

Postel was threatened by Ira Magaziner with the statement "You'll never work on the Internet again" after he split the root zone, assuming authority for the entire domain name system in an attempt to repatriate the root to IANA; Jon had plans to add hundreds of new tlds, a plan he had advocated for a while. This would let him do it, however it lasted less than a day.

Jon Postel managed the IANA function from its inception until his death in October 1998. Postel had been given defacto authority to perform the IANA function, as he had always done it in his position at the Information Sciences Institute, under its Department of Defense contract. After his death, Joyce Reynolds, who had worked with him at IANA for many years, managed the transition of the IANA function to ICANN.

Starting in 1988, IANA was funded by the U.S. government under a contract between the Defense Advanced Research Projects Agency and Information Sciences Institute (ISI). This contract expired in April 1997, but was extended to preserve IANA's function.

- On December 24, 1998, USC entered into a transition agreement with the Internet Corporation for Assigned Names and Numbers ICANN, transferring the IANA function to ICANN, effective January 1, 1999, thus making IANA an operating unit of ICANN.
- On February 8, 2000, the Department of Commerce entered into an agreement with ICANN to perform the IANA functions.
- In June 1999, at its Oslo meeting, IETF signed an agreement with ICANN concerning the tasks that IANA would perform for the IETF; this is published as RFC 2860.
- In November 2003, Doug Barton was appointed IANA manager.
- In 2005, David Conrad was appointed as IANA manager.
- in 2010, Elise Gerich was appointed as IANA manager.