Network Working Group	R. Whittle
Internet-Draft	First Principles
Intended status: Experimental	March 06, 2010
Expires: September 7, 2010	

TOC

Glossary of some Ivip and scalable routing terms
draft-whittle-ivip-glossary-01.txt

Abstract

This glossary is intended to help with the understanding of terms used in the Ivip core-edge separation architecture and of some non-Ivip terms which are pertinent to scalable routing. These are not "official" definitions of terms as used in scalable routing, but I hope they will help newcomers to the field. Please suggest corrections, additions and improvements.

Status of this Memo

Copyright Notice

described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

Table of Contents

---

1. Introduction

TOC

Please see the Ivip-arch ID [I-D.whittle-ivip-arch] (Whittle, R., “Ivip (Internet Vastly Improved Plumbing) Architecture,” January 2010.) and other IDs mentioned there for a detailed description of Ivip.
Significant developments regarding Ivip are at http://www.firstpr.com.au/id/ivip/ along with links to the IRTF Routing Research Group wiki, mailing list etc. I assume anyone with an interest in scalable routing is keeping up with the RRG mailing list discussions.
For a discussion of the meaning of Core-Edge Elimination (CEE) and Core-Edge Separation (CES), the history of these important terms and some debates about their true meaning, or value, please see my February 2010 RRG message: "CES & CEE: GLI-Split; GSE, Six/One Router; 2008 sep./elim. paper (v3)" http://www.ietf.org/mail-archive/web/rrg/current/msg06110.html and any later versions or discussion which follows.

---

2. Glossary

TOC

TOC

2.1. BR - Border Router

Border Router of an ISP - where the ISP network connects to the routers of other networks. See also PE and CE.

2.2. CE - Customer Edge router

TOC

Customer Edge router. A router in an end-user network which connects to one or more ISP networks. See also BR and PE.

2.3. Core-Edge Elimination (CEE)

TOC

This class of scalable routing architectures implements the "Locator / Identifier Separation" naming model, which is different from that used by IPv4 and IPv6 today. (LISP - the "Locator / Identifier Separation Protocol" - is badly named, since it does not do this and is an example of the other kind of architecture: Core-Edge Separation.) CEE is a scalable routing architecture in which hosts in end-user networks gain one or more "Locator" AKA "physical" addresses from each upstream ISP. These addresses can be scalably supplied to many end-user networks, since they are part of larger ISP prefixes. End-user networks do not retain these Locator addresses when they choose another ISP.
Host applications use a separate system (separate namespace) of "logical" AKA "edge" or "Identifier" addresses. The host's stack (or perhaps the application) determines how to create addresses for packets which the routing system will use to get the packet to the correct destination network via one of its ISPs, as determined by which "Locator" address the stack affixes to the packet.
The "Identifiers" (which are not regarded as "addresses") are retained by the hosts of the end-user network no matter which ISPs they use. There are no "core" or "edge" addresses - just two separate systems: one to identify hosts and the other to use as a routing locator to get the packet from one network to another. So "Elimination" refers to the elimination of the need for two different classes of address space - there being no need for special "edge" prefixes. All such systems involve changes to existing host stacks and perhaps applications. They generally attempt to be backwards compatible with IPv6. They are not practical for IPv4, for several reasons of which one is that a multihomed EUN (End User Network) consumes at least two or perhaps more times the address space it provides for its hosts.
Some architectures allow ISP routers to alter locator addresses to control packet flows. Generally, the hosts have to do more work than at present since there are no ITRs or ETRs or the like in the network. The network remains simple, compared to the additional elements added to

create a Core-Edge Separation architecture. (However some CEE architectures do involve significant new functionality in routers.) There is usually at least one additional global mapping lookup system, or an extension to DNS to support mapping lookups, such as using an Identifier to find that host's valid Locator or Locators. HIP and ILNP are examples of Core-Edge Elimination architectures. See also the start of the Architectural Choices section in Ivip-arch.

2.4. Core-Edge Separation (CES)

TOC

A scalable routing architecture in which hosts in end-user networks use a subset of the global unicast address space which are called "edge" (AKA "EID" or "SPI") addresses. The remainder of this space retains its current properties and is known as "core" (AKA "RLOC" or "conventional") space. End-user networks retain their edge address space no matter which one or more ISPs they use for Internet access.
A system of ITRs, ETRs and a mapping system transports packets addressed to "edge" addresses across the DFZ by tunneling from the ITR to the ETR address. Only a small number of large (short) prefixes need to be advertised in the DFZ to cover very large numbers of these "edge" prefixes (AKA, in Ivip, micronets of SPI space), so the impact on the DFZ is very small. In Ivip, these DFZ-advertised covering prefixes are known as MABs (Mapped Address Blocks).
This edge space can be sliced into many small pieces for very large number of end-user networks. The "edge" addresses are separated out from the "core" addresses, but remain part of the same namespace. Only ITRs treat packets differently according to whether the destination address is "edge" or "core". Hosts on both kinds of address communicate normally and the host requires no new protocols or knowledge of whether an address is "core" or "edge". IRON-RANGER LISP, APT, Ivip, TRRP and TIDR are all CES architectures. See also the start of the Architectural Choices section in Ivip-arch.

2.5. BGP - Border Gateway Protocol

TOC

Border Gateway Protocol. A protocol by which routers communicate in order that each can develop an optimal, or at least a good, set of best-path rules for its FIB, to handle packets matching all the prefixes the router handles. BGP is used in the interdomain routing system, which is also loosely referred to as the DFZ.

TOC

2.6. COTS

Commercial Off The Shelf server - no specific brand. For instance a rack-mount server running GNU/Linux, BSD, or any other operating system - usually remote controlled, and so without display or keyboard. High performance COTS servers today typically have multicore CPUs from Intel or AMD, gigabytes of RAM and one or more hard drives.

2.7. DITR - Default ITR in the DFZ

TOC

Default ITR in the DFZ. Previously known as an OITRD (Open ITR in the DFZ) and before that, erroneously, as an "Anycast ITR in the core/DFZ". The LISP equivalent is the PTR (Proxy Tunnel Router). DITRs advertise MABs (Mapped Address Blocks) and so attract packets addressed to SPI space which were sent by hosts in networks which have no ITRs.
DITRs (or PTRs) are essential for ensuring that networks adopting SPI (EID) space get all the packets which are sent to them, with full support for portability, multihoming and TE. In principle, a DITR could advertise every MAB in the Ivip system. In practice, there are likely to be multiple independent sets of DITRs, with each set having at least one DITR in a DITR-site, with these sites typically being widely distributed around the world to reduce the total path length between sending host the ETR used by the destination network. Since DITRs will generally be run by, or for, the MABOC (MAB Operating Companies) who lease SPI space to thousands of EUNs (end-user networks), the one or more DITRs at any one site will typically only advertise the MABs of the MABOCs this site is serving. See also "DSOC".

2.8. DFZ - Default-Free Zone

TOC

The large subset of the interdomain routing system which consists of routers which have more than one "upstream" link - meaning there is more than one path to "the rest of the Internet". If the router is a BR of an ISP or a PI-using end-user network which connects to the DFZ, then it will have one or more other links which take packets to this local network. If the router has no "local network" then it is a transit router in the DFZ and is operated by a transit provider.
A router at the border of an ISP or PI-using end-user network which has a single upstream link (probably to an ISP network) can have the interface for upstream link as the "default path" in its FIB and RIB. Routers with two or more links to the rest of the Net can't have such a default route, and so are considered to be in the "default-free" part of the interdomain routing system. DFZ routers need to have a route in their FIB and RIB for every prefix (route) which is advertised in the

interdomain routing system. (Often "DFZ" is used to refer to the interdomain routing system.) Since there are 300k or more such prefixes, this means the router needs to have a fast route processor (main CPU) to run its RIB and BGP sessions with neighbours.
Each DFZ router also needs a high capacity (and typically very expensive) FIB to figure out, for each incoming packet, which of the 300k+ prefixes best matches the packet's source address. DFZ routers are regarded as being multihomed. A "single-homed" router has a single upstream link. Its RIB and FIB have much fewer demands placed upon them, since they contain routes for the local network, accessible by one or more interfaces, and then a "default" rule, which catches all packets not yet matched, which causes the FIB to forward those packets to the single upstream link. "Single-homed" routers don't need their RIB or FIB to consider all the 300k prefixes which are advertised in the DFZ - just the ones this router advertises.
DFZ routers are very expensive and there are an unknown number of them - maybe 100,000 or so of them. They are run mainly by ISPs (who sell connectivity to end-user networks) and transit providers (who sell connectivity to ISPs and other transit providers. See the August 2007 RRG thread "Routers in DFZ - reliable figures from iPlane".
DFZ routers may also be operated by larger PI-using end-user networks, such as those of universities, which are multihomed to two or more upstream ISPs, and which choose to send out packets on the link with the optimal path to the destination, rather than just nominating one link as the "default".
A router which is inside a large network and is operating as a Route Reflector may also be considered part of the DFZ, if it needs to carry all DFZ routes in its RIB.

2.9. DFZ Control Plane

TOC

Broadly speaking, the system of all DFZ routers and their route processors communicating with each other using BGP messages so that each one can determine the optimal (or at least "good enough") best path for packets which are addressed to every prefix (route) which is advertised in the interdomain routing system. The entire global system behaves as a system - although its exact behaviour is not necessarily well understood. Geoff Huston's site http://bgp.potaroo.net is an excellent source of information on the BGP control plane. Please also see his 2010-3-01 message to the RRG (msg06152) which contains his latest analysis of the DFZ control plane's burdens.
The "control plane" is separate from the "data plane" - which actually handles traffic packets. The "control plane" includes the RIBs of all the DFZ routers. It is an essential goal of scalable routing to contain the growing load on the DFZ's control plane while providing portability, multihoming and TE for far more end-user networks than

currently have these things. (Reducing the load on the DFZ data plane is not possible in terms of the number of packets, but anything which reduces the load by limiting or reducing the number of prefixes in DFZ routers' FIBs, while allowing many more multihoming end-user networks, would also be achieving a vital goal of scalable routing.)

2.10. DSOC - DITR-site Operating Company

TOC

A MABOC which runs its own DITRs is typically runs them at multiple (perhaps 5 to 30 or more) DITR-sites. In this case, the MABOC is its own DSOC. A MABOC or some other company which is a DSOC may also run the DITR(s) and QSA(s) (Authoritative Query Servers) at that site to handle the MABs of multiple MABOCs.
In principle, a single DITR-site could handle all MABs in the Ivip system, but in general it is assumed that there will be multiple DSOCs and that each will have multiple, ideally numerous, DITR sites - each of which handles a subset of all the MABs. In addition to running these DITR routers/servers and QSA servers, the DSOC needs to do at least two other things.
Firstly, they need to get real time mapping changes from each MABOC's system which collects mapping change commands from EUNs (or their appointees) and reliably, securely and rapidly fan this out to all the QSAs in their DITR-sites. The one or more QSAs at each site are used by the DITRs and for answering mapping queries from the QSRs (Resolving Query Servers) of typically nearby ISPs and EUNs which have ITRs and QSRs.
Secondly, they need to collect traffic statistics on the usage of the DITRs in a manner that they can charge the MABOCs for this work, and so the MABOCs can charge their individual EUNs who use the space in each MAB.

2.11. EAF - ETR Address Forwarding

TOC

ETR Address Forwarding. The MHF (Modified Header Forwarding) technique for IPv4 - as an alternative to encapsulation. See:
[I-D.whittle-ivip-etr-addr-forw] (Whittle, R., “Ivip4 ETR Address Forwarding,” January 2010.)

TOC

2.12. EUN - End-User Network

A network which is not used for selling Internet connectivity - although the term "end-user network" does apply to a network such as that of a hosting company, which leases the capacity of its servers to its customers. "Internet connectivity" in this sense means connecting a user's network or mobile device to the Internet, which is what ISPs do. Most of the end-user networks referred to in scalable routing are those which want or need portability, multihoming and inbound traffic engineering (TE). However, this is just a subset of end-user networks. Most end-user networks, such as those of home and SOHO users, are fine without portability, multihoming or inbound TE. With TTR (Translating Tunnel Router) mobility, each mobile node (MN) is regarded as a separate EUN, though it may have only a single IPv4 address and isn't really a "network" since it is a single host. An MN could also have multiple IPv4 addresses, multiple prefixes of IPv6 addresses etc. The network of a passenger jet or cruise ship is physically mobile and is also regarded as a mobile EUN, even though in practice this network may charge its customers for Internet connectivity to their PCs etc.

2.13. FEC - Forwarding Equivalence Class

TOC

Within a router, FEC can be thought of as a number of some kind which the FIB chooses for each incoming packet. A simple type of FEC is which interface to forward the packet from. However, routers may maintain multiple queues for packets going out a single interface, so as to give priority to different types of packet. Each such queue would be identified by a different FEC.

2.14. FIB - Forwarding Information Base

TOC

Forwarding Information Base. This refers either to the body of data in a router which directly controls how traffic packets are processed, and/or to the hardware and software which performs this plus the data which controls them. Earlier routers had a single FIB, with multiple input/output interfaces. Many modern, larger, high-speed routers integrate an FIB into each interface to handle the packets arriving on that interface alone - or have multiple FIBs each dedicated to one or more interfaces.
The FIB has arguably the most demanding task of any part of the router - though the interconnect between the interfaces/FIBs is has a daunting task too.

2.15. IPTM

TOC

Ivip's "ITR Probes Tunnel MTU" arrangement for handling the PMTUD problems inherent in encapsulated tunnels between the ITR and ETR. This will be the subject of a future ID. For now please refer to: http://www.firstpr.com.au/ip/ivip/pmtud-frag/ .

2.16. ITFH - ITR Function in sending Host

TOC

ITR Function in sending Host. An ITR which is implemented purely by software which is added to a host, and which only processes the packets that host sends. The host can be on a conventional "core" address or on an SPI ("edge") address. It cannot be behind NAT. Generally, ITFH should not be implemented on hosts with slow or unreliable links, such as any host relying on a 3G or similar wireless link.

2.17. ITR - Ingress Tunnel Router

TOC

An existing router, or a function within a server or existing host, which accepts packets addressed to an SPI address and which alters the packet in some way. The altered packet is forwarded so the DFZ routing system (plus any internal routers of ISPs and end-user networks which are on path) will transport ("tunnel") the modified packet to an ETR, which reverses the modifications and forwards the packet to the destination network.
ITRs need to look up some mapping for each packet - and they need to get the mapping quickly when a packet arrives which they have no cached mapping for. The ITR then caches the mapping for some time, so it can handle packets addressed to this address (or any other address in the micronet which contained the first packet's destination address) without requesting mapping again.
ITRs in the DFZ are called DITRs. An ITR function in a sending host is called an ITFH. ITRs in other Core-Edge Separation schemes always use encapsulation to tunnel the packet to the ETR. Ivip ITRs will be able to use MHF (Modified Header Forwarding) instead of encapsulation.

2.18. Ivip

TOC

Internet vastly improved plumbing. The origins of the acronym and guidance on capitalization are in section 3.

2.19. MHF - Modified Header Forwarding

TOC

Modified Header Forwarding. An method for ITR tunneling traffic packets to an ETR - as an alternative to encapsulation. The IP header is modified, so all routers between the ITR and ETR must be upgraded to handle the new format. For IPv4: ETR Address Forwarding (EAF) and for IPv6: Prefix Label Forwarding (PLF).

2.20. MAB - Mapped Address Block

TOC

A Mapped Address Block is a DFZ-advertised prefix containing SPI space - typically the UABs (User Address Blocks) and their constituent micronets for many end-user networks.
This is an Ivip term with no direct equivalent in LISP, although LISP too has the same concept. (In the RRG list, Dino Farinacci has used the term "coarse prefix" to refer to the LISP equivalent of Ivip's MABs.)
MABs are advertised by ITRs in the local routing system and by DITRs in the DFZ. Typically an ordinary ITR (an ITR inside an EUN or an ISP's network - all ITRs except DITRs) will advertise all the MABs of the Ivip system while DITRs will only typically advertise a subset of MABs.
ITFH functions in sending hosts don't "advertise" MABs, but they intercept all outgoing packets with destination addresses which match any MAB.

2.21. MABOC - MAB Operating Company

TOC

Mapped Address Block Operating Company. An organization, here assumed to be a company (otherwise MABOO...) which controls the address space within a MAB.
Most MABOCs will use this space to lease SPI space to large numbers of EUNs, each of which is therefore a customer of the MABOC. A MABOC my have one or many MABs. The space leased to a given EUN is a User Address Block (UAB). Within each UAB, each EUN dynamically assigns the space into one or more micronets, each with its own mapping. However, if the UAB may be just a single IPv4 address or IPv6 /64, then it can only be used as a single micronet.
An EUN may also have a complete MAB, in which case it is the MABOC of this MAB. Perhaps it leases SPI space to other EUNs, or perhaps not. A network which leasing SPI space to others is perhaps no longer strictly speaking an EUN, but it is not providing actual Internet connectivity - so we do not regard it as an ISP. A MABOC may or may not be an ISP.

MABOCs are also responsible for handling the mapping changes for all micronets in their MABs and for running DITRs to advertise these MABs and so handle packets addressed to any micronet in the MAB which are sent from hosts in networks without ITRs. MABOCs are also responsible for providing multiple QSAs (Authoritative Query Servers) which are typically numerous and widely dispersed so that ISPs and other networks with ITRs and QSRs (Resolving Query Servers) can quickly and reliably obtain mapping for any micronet in this MABOC's MABs. Therefore, a MABOC may run DITRs and QSAs at its own DITR-sites, or it may contract another company to perform these functions - a DSOC.

2.22. Mapping

TOC

In a CES architecture such as Ivip, mapping is information which tells an ITR which ETR to tunnel a packet to, when the destination address matches one of the MAB prefixes - when the packet's destination address is in the SPI "edge" subset of the global unicast address range. There is only mapping for SPI addresses.
In Ivip, a range of contiguous addresses covered by one mapping is a "micronet" and the mapping consists purely of a single ETR address. In LISP and other CES architectures, the mapping of an EID prefix (~AKA "micronet") typically consists of multiple ETR addresses with various priorities and weightings so the ITR (or Default Mapper, in APT) can choose one for the purposes of load balancing TE and/or multihoming service continuity.

2.23. Mapping Distribution System

TOC

CES architectures need a method by which ITRs can quickly find the mapping which applies to a particular "edge" (AKA SPI or EID) address which is the destination address of a traffic packet the ETR is handling . The device which ultimately controls this mapping could be anywhere in the world - and there could be very large numbers of such devices, scattered all over the Net. The Mapping Distribution System is how the ITRs get the mapping - as quickly and reliably as possible.
Full-push mapping distribution involves all mapping being pushed to all ITRs, so the ITR already has the mapping. (e.g. LISP-NERD.)
Full-pull involves a global system by which the ITR's request is directed to the one authoritative query server (or one of just a few such servers) which is the authoritative source of the mapping - and which sends back the map reply to the ITR. (e.g. LISP-CONS, LISP-ALT and TRRP.)
A third approach is to push all mapping to "local" full database query servers, such as in each ISP. ITRs request mapping from these. (e.g.

APT and Ivip before DRTM [I-D.whittle-ivip-drtm] (Whittle, R., “DRTM - Distributed Real Time Mapping for Ivip and LISP,” March 2010.).)
From late February 2010, Ivip uses DRTM - in which DSOCs and MABOCs push the full mapping database, with real-time updates, to QSAs (Authoritative Query Servers) at multiple DITR-sites. These QSAs are not "local" as were the QSDs in the previous approach to Ivip, but are typically "nearby" enough to the querying QSRs (Resolving Query Servers) to provide mapping replies reliably and within a few tens of milliseconds. DRTM is therefore "push" to the QSA, "pull" in the sense of the ITR requesting, via a QSR, from the QSA, the mapping - and also "push" from the QSA to the ITRs for any Cache Updates changes affecting micronets whose mapping was sent to the ITRs within the current caching time.

2.24. Micronet

TOC

A micronet is a contiguous range of SPI address space which is mapped to a single ETR address. A UAB (User Address Block) contains one or more micronets. The units of splitting SPI space are IPv4 addresses and IPv6 /64s. Micronets and UABs are integer numbers of these units - so they are not restricted to being binary boundary prefixes. The equivalent in LISP is an "EID" prefix.

2.25. Mobility

TOC

Mobility in TCP/IP networks refers not directly to a host being physically mobile and connecting to different networks. Nor does it necessarily imply the device has wireless interfaces to those networks. It generally refers to the ability of a host to maintain its communication sessions while it is changing its physical point of connection.
Some mobility systems meet these requirements by giving the host the same IP address no matter where it physically connects to a particular access network. A global approach to mobility would enable session continuity when the host connects to any network at all, and so may have completely different IP addresses from time-to time. One approach is to use special IP protocol stack capabilities so applications are not affected by changes in physical address. Another is to keep the current host stack and (with some additional software and usually the involvement of some devices in the network, such as ITRs and TTRs - Translating Tunnel Routers) give the host a single IP address no matter how or where it is connected. Such a system is the TTR Mobility architecture. [TTR Mobility] (Whittle, R. and S. Russert, “TTR Mobility

Extensions for Core-Edge Separation Solutions to the Internets Routing Scaling Problem," August 2008.)

2.26. MN - Mobile Node

TOC

Mobile Node. Synonymous with "mobile host".

2.27. MTU - Maximum Transmission Unit

TOC

Maximum Transmission Unit. The maximum length of a packet, measured in bytes, which a particular interface of a router, or the data link it drives, can handle. See also PMTU and PMTUD.

2.28. Multihoming

TOC

The ability of an end-user network (as large as a corporation network, or as small as a home network or handheld wireless device) to maintain all its communication sessions, and the identity of all its hosts, when the connection it is using via one ISP fails, and is replaced quickly by that of another ISP.
One way of doing this is to ensure the hosts never see any changes - that is, the hosts always retain their own IP addresses. This is achieved with the currently only approach to multihoming - having the EUN advertise its own PI prefix in the DFZ. CES architectures also maintain the host's IP address, but enable multihoming to be done in a scalable way - without each EUN's address space being separately advertised in the DFZ.
Another approach, as used by CEE architectures is to have the host IP stack manage the host identity (which suitably written application programs use to set up and maintain communications with other hosts rather than using IP addresses) in a stable way so that applications are unaware of the ISP link changes, while operating from either a physical (Locator) address obtained from the first ISP or that obtained from the second. CEE use the Locator / Identifier Separation naming model for this purpose.

TOC

2.29. Outer header

When a packet AA is encapsulated, another one or more headers is prepended to it. The outer header is the IP header of the new packet BB which contains just the original packet AA (Ivip), or (LISP) a UDP header and a LISP header, which is followed by the AA packet. The destination address of the outer header will be recognised by all routers and the packet will be forwarded towards that address - which in the case of ITR encapsulation, will be an ETR which can decapsulate the packet an forward it to the destination network.

2.30. PA - Provider Aggregatable

TOC

Provider Aggregatable - address space, prefix or IP address. ("Provider Assigned" is also in common use, but "Provider Aggregatable" seems to be more appropriate. See Brian Carpenter's message to the RRG on 2010-02-28.) Global unicast address space which is used by an end-user network (EUN) and comes from an ISP's prefix.
Typically the prefix it comes from is a large (short) prefix which is therefore not a problem in terms of scalable routing due to there not being too many of such prefixes in the DFZ. The same large (short) prefix which the ISP advertised may be used for its own internal purposes and for many other EUNs. The provider (the ISP) is said to be able to aggregate many such PA prefixes into a single prefix of its own which it advertises in the DFZ. ISPs typically do so via multiple uptream links to other ISPs or transit providers - so this single large (short) prefix is multihomed via these links.
PA prefixes are good for scalable routing, but bad for any end-user network which wants portability, since they only get these particular addresses with a particular ISP. Likewise, unless special techniques are used (CEE), an end-user network can't achieve multihoming (with session continuity during an outage of one ISP or the link to that ISP) with PA space - or inbound TE. See also PI space.

2.31. PE - Provider Edge router

TOC

Provider Edge router. A router in an ISP network which connects to one or more end-user networks. See also BR and CE.

TOC

2.32. PI

Provider Independent address space, prefix or IP address. Global unicast address space which is used by an end-user network and which the network retains no matter which ISPs it uses for connecting to the Net and which is advertised as a separate prefix in the DFZ. SPI (Scalable PI) space is also Provider Independent, but each EUN's SPI space is not separately advertised in the DFZ.
PI space is good for the end-user network, since it is portable and can be used for multihoming and TE, with full session continuity in the event of failure, by having its two or more ISPs advertise the prefix in the DFZ - or to have one advertise it and the other advertise it if the link to the first ISP fails. This use of PI prefixes is bad for routing scalability, since each such PI prefix and any changes to its advertisement is an additional burden on all DFZ routers and on the DFZ control plane in general. PI space is also too costly to obtain and advertise in the DFZ for many EUNs. A further problem, at least with IPv4, is the convention of not propagating prefixes longer than /24 between ASes (Autonomous Systems) in the DFZ - so every prefix of PI space uses at least 256 IP addresses. See also PA and SPI.

2.33. PLF - Prefix Label Forwarding

TOC

Prefix Label Forwarding. The MHF (Modified Header Forwarding) technique for IPv6 - as an alternative to encapsulation. See: http://www.firstpr.com.au/ip/ivip/PLF-for-IPv6/ .

2.34. PMTU - Path Maximum Transmission Unit

TOC

Path MTU. The MTU (maximum packet length, in bytes) not of a single interface but of a path from one device to another - and so of all the devices, input and output interfaces and data links in that path. In a core-edge separation architecture the PMTU of most interest is that between the ITR and the ETR, since encapsulation disrupts the RFC 1191 or RFC 1981 PMTU Discovery process which normally operates with all routers between the sending and destination hosts.

2.35. PMTUD - Path MTU Discovery

TOC

Path MTU Discovery. RFC 1191 (IPv4) and RFC 1981 (IPv6) PMTUD is a protocol by which the sending host can try sending different length

packets (which must be unfragmentable in IPv4: DF=0 - in IPv6, all packets are unfragmentable) to a destination host and being able to choose the longest packet length which will fit in the PMTU, by using ICMP PTB (Packet Too Big) messages from any router where the packet will not fit within the next-hop MTU. RFC 1191 and RFC 1981 are universally used, but there are some problems. See the 2010-02-03 RRG thread "Fred's IPv4 PMTUD research: RFC1191 support frequently broken". There is also a more complex and recent PMTUD technique - RFC 4821 - which has not been adopted to any significant degree. RFC 4821 does not rely on PTBs, but involves stack packetization layers such as TCP and the packetization layers of applications discovering the PMTU to a host by end-to-end means, and then sharing that PMTU information with other such layers.

2.36. Portability

TOC

"Portable address space" means the ability of an end-user network to retain its address space when using different ISPs. This may involve the network having a single link to one ISP - or multiple links, and so being multihomed. Being free to change ISPs is important for competition and flexibility. While there have been proposals, especially for IPv6, to make it easy to change host and network addresses so as to make it easy to change to a new ISP's PI address space, this has never been accepted as providing the convenience, low cost and reliability of actual portable address space.
"Ease of choosing ISPs" has been one way of stating a major goal of scalable routing, and some people have stated this in terms of assuming the end-user network can't keep its own address space: ease of renumbering when changing ISPs". However, the only practical way the needs of end-user networks can be met when choosing another ISP is to retain the current address space - which means this address space must be "portable". The impossibility of making renumbering acceptably free of cost and risk is in part due to EUNs having the addresses of their hosts, or of their entire network, in configuration files, DNS zone files, ACLs (Access Control Lists) etc. in places outside their direct control. Having to get these changed, correctly, at exactly the right time, is impossible or at least so impractical as to preclude renumbering from being a viable approach to "ease of choosing ISPs". Therefore, "portability" of a network's address space is the only way of enabling them to choose a new ISP without unreasonable cost or risk of disruption.

TOC

2.37. PTB - Packet Too Big ICMP message

ICMP Packet Too Big message. Part of RFC 1191 and RFC 1981 Path MTU Discovery (PMTUD). Ordinarily sent to the sending host by a router which determined that the packet was too long for either the data link or next device in the next-hop.
The PTB includes initial parts of the original packet, and an MTU number which the sending host can use as a maximum packet length, so that future packets will not breach this limit. When ITRs use encapsulation to tunnel packets to an ETR, the routers between the ITR and ETR are unable to generate a valid PTB to the sending host (unless they were specially modified, in some way). So the ITR has to take care of whatever MTU limits exist between it and the ETR, and generate PTBs to the sending host in order to ensure its packets are not longer than the PMTU (Path MTU) of the path to the ITR.

2.38. Query Server

TOC

In Ivip, a "query server" is a server which responds to queries about mapping. See: QSA, QSC, QSD (obsolete) and QSR.
LISP and APT do not use the term "query server", but I use it to describe whatever it is in these architectures which responds to the ITR's request for mapping. In LISP-ALT, the query servers are either ETRs or MSes (Map Servers) - and are distributed all over the Net. So LISP-ALT is a global query server system. APT's query servers are local to the ISP and are called Default Mappers.

2.39. QSA - Authoritative Query Server

TOC

QSAs are authoritative mapping query servers for the one or more MABs they serve.
In principle, a single QSA could be authoritative for all MABs, in which case it would resemble the now obsolete "QSD". With DRTM [I-D.whittle-ivip-drtm] (Whittle, R., “DRTM - Distributed Real Time Mapping for Ivip and LISP,” March 2010.), there is no need for any QSA to carry mapping for all MABs, so it is assumed that each QSA carries the full mapping database for one or more MABs, but not for all of them.
QSAs require a secure, robust, real-time feed of mapping updates for all the MABs they serve. QSAs are typically located at DITR-sites. Since the DITRs at DITR-sites are likely to be busy, they need a close-by QSA - such as a server in the same rack - so as not to have to rely on sending packets to a QSA at some other site. A QSA could be implemented in the same device as the DITR, but it is generally assumed

it would be implemented in a separate COTS server. The term "QSA" applies both to QSAs which serve the one or more local DITRs, and to those which handle queries from QSRs (Resolving Query Servers) in (typically) nearby ISP networks and EUNs.
A QSA at a DITR-site is "full-database" for all the MABs this DITR site serves. However, the one DITR-site could have multiple QSAs accepting queries from QSRs, and it would be possible for some of those QSAs to handle one subset of the DITR-site's MABs and other QSAs to handle other subsets. In principle, it would also be possible for a QSA to be located somewhere other than at a DITR site. For instance if the MABOC for a set of MABs was happy with having 10 DITR sites, but wanted to establish a larger number of QSAs to spread the query load better, or to be closer to some QSRs than is possible with the DITR-sites, then it could run QSAs in other places.
QSAs need a reliable, robust, feed of mapping updates for all the MABs they serve. How they get this is not at present defined in Ivip, since it needs to be achieved solely within the networks of a DSOC and however many MABOCs whose MABs the DSOC handles. The limited scaling and interoperability challenges of doing this are assumed to be solvable, but in the future it would be good to have a protocol by which MABOCs could send updates to DSOCs, and which DSOCs could use for their internal real-time fanning out of this information to their DITR-sites.

2.40. QSC - Caching Query Server

TOC

Caching query servers responds to map requests from ITRs or other QSCs. Each QSC sends map requests to one or more upstream local QSCs and/or QSRs. Each QSC also receives Cache Update messages from whatever device it queries and then passes on Cache Update messages to the one or more queriers which were sent mapping for the micronet concerned, during the current caching time.
QSCs are generally always in EUNs or ISP networks. They are not usually used in DITR-sites, since a DITR-site should have its own QSA and it probably makes sense for the one or more DITRs at the site to query the QSA directly. However, QSCs could also be used in a DITR-site, to reduce the load on the QSA. Since QSCs cache the map reply information they receive, they will sometimes - perhaps frequently - be able to answer map requests from their queriers from their cached mapping, so eliminating the need to query whatever query server they would normally query. Likewise, if multiple queriers (ITRs or other QSCs) have recently (in the current caching time) been sent mapping for a given micronet and the QSC receives a Cache Update for that micronet, it will send out multiple Cache Update messages to those queriers - so saving its upstream QSC or QSR from having to send more than one Cache Update.

2.41. QSD (obsolete term)

TOC

Prior to the introduction of DRTM [I-D.whittle-ivip-drtm] (Whittle, R., "DRTM - Distributed Real Time Mapping for Ivip and LISP," March 2010.) in late February 2010, the QSD - Full database query server - played a central part in Ivip.
Each ISP or EUN with its own ITRs was to run within its network one or ideally two or three QSDs, each of which received the full feed of mapping updates for all MABs. An EUN with ITRs could also use the QSDs of its one or more ISPs. In March 2010, Ivip no longer needs QSDs. There role is taken by QSRs - Resolving Query Servers. QSRs are caching query servers, but it will remain an option for a QSR to be sent full feeds of mapping updates for one or more MABs. If such a QSR received full mapping feeds for all MABs, then it would be functioning identically to a pre-February 2010 QSD. A QSD never needs to ask any other device for mapping information, whereas to the extent that a QSR is caching, it always has to ask QSAs for mapping information. Ivip should be able to scale to the greatest levels required with purely caching QSRs.

2.42. QSR - Resolving Query Server

TOC

Resolving query server. With DRTM [I-D.whittle-ivip-drtm] (Whittle, R., "DRTM - Distributed Real Time Mapping for Ivip and LISP," March 2010.), this takes the role previously performed by the now obsolete "QSD".
An ISP which has its own ITRs, and or which has customer networks with ITRs needs at least one, and ideally two or three, QSRs in its network. EUNs with ITRs can also install their own QSRs, or they may be able to use the QSRs of their one or more ISPs instead.
QSRs answer mapping queries from devices internal to the network in which they are located - ITRs or QSCs. They can also send Cache Updates to these queriers. While an ITR may be configured to send queries to one or a few upstream query servers - QSCs or QSRs, always in its own network, or in an ISP network used by its own network - and while the optional, intermediate, caching QSCs do the same, a QSR does not query any server in its own network. It only queries authoritative QSAs, which are typically located "nearby" (within a few thousand km). Since QSAs are typically only authoritative for a subset of MABs, each QSA needs to automatically discover two or three ideally "nearby" QSAs for each of the MABs in the Ivip system.
Since MABs will generally be run by a smaller number of MABOCs, and since the MABOCs will directly or indirectly run a still smaller number of DITR-sites, where the QSAs are located, for each set of DITR-sites, the QSA will typically find one or ideally two or three, "nearby" QSAs at these sites. This discovery is done automatically and on a

continuing basis via a DNS-based system as described in [I-D.whittle-ivip-drtm] (Whittle, R., “DRTM - Distributed Real Time Mapping for Ivip and LISP,” March 2010.).

2.43. Replicator (obsolete term)

TOC

Replicators" were a central part of Ivip's mapping distribution system until late February 2010, when they were made unnecessary by DRTM [I-D.whittle-ivip-drtm] (Whittle, R., “DRTM - Distributed Real Time Mapping for Ivip and LISP,” March 2010.). A Replicator was a COTS server within the Ivip fast-push mapping distribution system and is still described in [I-D.whittle-ivip-fpr] (Whittle, R., “Fast Payload Replication mapping distribution for Ivip,” March 2010.). A Replicator receives two or more streams of update packets from upstream devices, such as other Replicators as part of a fully or partially meshed flooding system for rapidly and robustly propagating real-time changes to mapping to full database query servers (QSDs - also now obsolete). As noted in the QSD section above, these are no longer needed with DTRM, but remain an option which could be used either within the networks of DSOCs or as part of pushing real-time mapping feeds for one or more MABs to QSRs, which, instead of being purely caching, are made to be full-database for one or more MABS.

2.44. RIB - Routing Information Base

TOC

Within a router, the RIB is the body of data - as maintained by software which controls the route processor (administrative CPU of the router) - by which the router decides how it will handle traffic packets.
When the router is running BGP (as all DFZ routers do) the RIB is not just a product of messages received, but also controls the BGP messages which will be sent to neighbours. The RIB is used to generate data which is written into the FIB so the FIB classifies, processes and forwards packets in the manner specified by the RIB. Many routers, in addition to running BGP, also run other routing protocols - and the RIB contains routes generated by those systems too.

2.45. SPI - Scalable Provider Independent

TOC

Scalable Provider Independent address space. The Ivip term for the new "edge" subset of the global unicast space which is suitable for end-

user networks, providing portability, multihoming and inbound TE in a manner which is "scalable" - does not overly burden the DFZ control plane.
The LISP equivalent is "EID".
Global unicast space which is not SPI is known as "conventional" or "core" space - or in LISP, as "RLOC" - space.

2.46. TE - Traffic Engineering

TOC

Most references to TE in the scalable routing field refer to inbound TE - steering incoming traffic streams between two or more ISPs and their data links.
Both inbound and outbound TE is typically practised to balance traffic volumes over multiple links to make best use of each link's capacity. Other reasons for preferring one link over another for particular subsets of the total traffic include one link being more reliable, lower latency or lower cost. Also, it may be desired for various policy reasons to avoid some traffic traversing one link, which would cause it to pass through some ISP or country jurisdiction which was not desired.

2.47. TTR Mobility architecture

TOC

A Translating Tunnel Router behaves like an ETR to the core-edge separation scheme and communicates with the Mobile Node (MN) by a two-way tunnel initiated by the MN. The TTR is ideally topologically close to the MN - no more than 1000km or so distant. The MN tunnels to one or more TTRs. TTRs are commercially operated (by TTROCs) and are ideally numerous and well connected.
The MN's outgoing packets from its SPI address are sent out to the TTR which forwards them to the destination - since the access network the MN is connected to will probably not forward packets with such a source address. See: [TTR Mobility] (Whittle, R. and S. Russert, “TTR Mobility Extensions for Core-Edge Separation Solutions to the Internets Routing Scaling Problem,” August 2008.).

2.48. TTROC - TTR Operating Company

TOC

An organization, assumed to be a company, which operates a complete (typically global) system of TTRs. The entire TTR system of a TTROC operates as a single system and instructs how MNs choose which TTRs to

tunnel to. The MN user is therefore a customer of the TTROC, since they pay for access to the TTROCs network of TTRs.
The MN user may provide their own SPI address space - such as a single IPv4 micronet of one IPv4 address - for use by their MN with the TTROC's system. Alternatively, the TTROC may supply this micronet - in which case it is either a MABOC or is obtaining the micronet from a MABOC. In both cases, the TTROC controls the mapping of the micronet as long as the MN is using its TTR network.
Multiple TTROCs can compete. If there was a standardised tunneling and management protocol for all MNs to use with all TTRs, then a single piece of software in MNs could be used for all TTROC systems. Since there is considerable scope for innovation, service differentiation etc. in the TTR Mobility field, it may be more likely that TTROCs will develop their own specialised software for the major types of MN, and distribute this to their MN customers. Theoretically a single MN could operate with the TTR systems of multiple TTROCs at the same time, but each system would provide it with a separate micronet of SPI space.

2.49. UAB - User Address Block

TOC

A contiguous range of SPI address controlled by a single end-user network. May be used as a single micronet or split into multiple micronets. A MAB typically contains many UABs. ITRs, QSCs and QSRs and QSAs don't work with UABs - they only work with micronets. As with micronets, UABs are of integer length, with any starting point within the MAB - and the units are IPv4 addresses or IPv6 /64 prefixes.
Each micronet must be fully contained within a UAB - and each UAB must be fully contained within a MAB.

2.50. WAG ...

TOC

Wild Assed Guess. Technique employed where some kind of figure is required, but the constraints on the realistic range for the figure are unknown or difficult to use precisely.
Useful for discussing order-of-magnitude questions concerning future Internet developments, due to our current inability to obtain data about the future. Similar to "Stab in the Dark", but used for serious technical discussions and made with full awareness of its speculative nature.

TOC

3. The Ivip acronym

The "vip" in "Ivip" comes from the 1961 Doris Day, Rock Hudson and Tony Randall romp "Lover Come Back". Advertising executive Jerry Webster (Rock Hudson) finds himself in trouble - from which he believes he can extract himself by convincing a dancer (Edie Adams) that he will introduce her to Hollywood by making her the star of a promotional campaign for a hot new product. She is keen and keeps asking him what the product is. Casting his eyes around the room, he sees a newspaper with a headline about a VIP. "Vip!" he exclaims - and spends the rest of the movie trying to figure out what this great new product will be.
Capitalization of the four characters is user selectable but defaults to "Ivip". Lower-case 'i' is not recommended since "iVIP" might be mistaken for an abrasive bath and sink cleanser from Apple Inc. (A low cost product for those unable to afford a Macintosh computer or i**** product - the mere possession of which instantly renders the owner's whole dwelling spic-and-span.)
The capital 'I' raises a potential problem with sans-serif fonts such as Helvetica, since it is indistinguishable from lower-case "L". This has bedevilled the 3GGP term "Iub" (capital 'i') which seems to be more widely known outside the organisation as "lub" (lower-case 'L').
Ivip predates and has absolutely no connection with the UK "IVIP" iPhone application.

---

4. History of Ivip's mapping system

TOC

DRTM (Distributed Real Time Mapping) was first described on the RRG list on 2010-02-26, but it took about two weeks to update the IDs accordingly. If you have not read any Ivip material before this, and if you are not concerned about critiques of Ivip made according to the pre-DRTM version of Ivip, and if you are not interested in Replicators, Lost Payload Servers and the like, then there's no need to read this section.
The terms "Plan A" etc. are purely to help describe how the design of Ivip's mapping system has progressed - these terms are not used in the IDs themselves.

Plan-A
2007-07-15: Original system with a tree-like structure of Replicators - with the top-level being "Launch Servers" with a fancy protocol between them.

ivip-arch-00/01/02	} All
ivip-db-fast-push-00/01	} obsolete.

2010-01-13: Same system, but all-new ivip-arch and revised ivip-db-fast-push.

ivip-arch-03	Completely rewritten.
ivip-db-fast-push-02	Better documentation of the original Launch Server system.

Plan-B
2010-01-18: "Launch servers" replaced by Level 0 Replicators which are fully meshed and have a flooding arrangement which is simpler, faster and more robust.

ivip-db-fast-push-03	Significant simplifications and new material to give an overview of Plan-B.
ivip-fpr-00	All new ID with goals and non-goals, better description of Replicators and the best Plan-B documentation.

Plan-C
2010-02-07: Ivip's (short-lived, and not fully documented) distributed mapping distribution system which also used Replicators, but not in a single global system. Described in RRG message msg05975.html

This keeps the Replicator concept, but has no central inverted tree structure of Replicators. Instead, one or more ISPs (or large end-user networks) make their own small tree (or non-tree-structured mesh) of Replicators, and get feeds of mapping changes for the MABs of all MABOCs from the one or typically more than one mapping coordination companies (now known as DSOCs) or the MABOCs themselves - whoever runs the nearest one or two DITR-Sites for each MABOC. So there is no central inverted tree of Replicators - just smaller trees or meshes or even single QSDs getting feeds from MABOC-run DITR-Site sources of

mapping generally not too far away.

In Plan-A and Plan-B, the MABOCs were either RUAS (Root Update Authorisation Server) companies, or contracted RUAS companies to handle the mapping of the micronets in their MABs. The RUAS companies collectively ran a decentralised but still unified inverted tree-like structure of Replicators to fan out mapping changes in real-time all over the world to ISPs' full database QSDs.

In Plan-C, there is no global inverted tree of Replicators and the MABOCs invest more and reach out to ISPs from their widely distributed DITR-Sites. ISPs don't absolutely need ITRs and QSDs (and therefore mapping feeds and probably Replicators) but they will probably want them after a while (assuming some of their customers are using SPI space) since having their own ITRs will reduce traffic going out to a DITR and returning to these customers' ETRs.

Missing Payload Servers are also needed so the ISP's QSDs can get mapping which is somehow missing from the two or more upstream Replicators - due to temporary outages affecting the two or more feeds.

Plan-D

2010-02-24: DRTM - Distributed Real Time Mapping - no need for Replicators, Missing Payload Servers or QSDs.

ISPs (or EUNs) which want to run their own ITRs can still use the Plan-C approach of having their own full-database QSDs, with full feeds, Replicators, Missing Payload servers etc. However this is entirely optional and as far as I know, is not required for Ivip to scale well to the largest numbers of micronets and EUNs imaginable.

The main plan is for ISPs (and end-user networks) which want ITRs to use new query servers at these nearby MABOC-operated (directly or indirectly) sites where the DITRs are. These QSAs (referred to in the RRG message as "DITR-Site-QSD query servers") are "full database" for the subset of MABs each such DITR-site handles. The ISP's ITRs query these via a QSR - which is like a caching QSC query server but which knows, for each MAB, the addresses of two or more of these typically "nearby" QSAs authoritative,

full-database, query servers for each MAB.

Therefore, the ITRs in an ISP or an EUN are relying on full-database query servers are no longer strictly "local" - as they were in Plans A, B and C. They are (typically) "nearby". This means that they are normally "close" or "close enough" that delay times and query/response packet losses are insignificant. So this is fully distributed, but is not a "global" query server system like LISP-ALT: with queries and responses frequently traversing the Earth - with consequent delays, losses and scaling problems.

Figure 1: History of Ivip's mapping system, to early March 2010.

---

5. Security Considerations

TOC

None.

---

6. IANA Considerations

TOC

None.

---

7. Informative References

TOC

[I-D.whittle-ivip-arch]	Whittle, R., “Ivip (Internet Vastly Improved Plumbing) Architecture,” draft-whittle-ivip-arch-03 (work in progress), January 2010 (TXT).
[I-D.whittle-ivip-drtm]	Whittle, R., “DRTM - Distributed Real Time Mapping for Ivip and LISP,” draft-whittle-ivip-drtm-01 (work in progress), March 2010 (TXT).
[I-D.whittle-ivip-etr-addr-forw]	Whittle, R., “Ivip4 ETR Address Forwarding,” draft-whittle-ivip-etr-addr-forw-00 (work in progress), January 2010 (TXT).
[I-D.whittle-ivip-fpr]	Whittle, R., “Fast Payload Replication mapping distribution for Ivip,” draft-whittle-ivip-fpr-01 (work in progress), March 2010 (TXT).
[TTR Mobility]	Whittle, R. and S. Russert, “TTR Mobility Extensions for Core-Edge Separation Solutions to the Internets Routing Scaling Problem,” August 2008.

Author's Address

TOC

	Robin Whittle
	First Principles
Email:	rw@firstpr.com.au
URI:	http://www.firstpr.com.au/ip/ivip/