Network Working Group
Internet-Draft
Obsoletes: RFC6179 (if approved)
Intended status: Informational
Expires: September 19, 2013

F. Templin, Ed.
Boeing Research & Technology

March 18, 2013

The Interior Routing Overlay Network (IRON)
draft-templin-ironbis-13.txt

Abstract

Since large-scale Internetworks such as the public Internet must continue to support escalating growth due to increasing demand, it is clear that Autonomous Systems (ASes) must avoid injecting excessive de-aggregated prefixes into the interdomain routing system and instead mitigate de-aggregation internally. This document describes an Interior Routing Overlay Network (IRON) architecture that supports sustainable growth within AS-interior routing domains while requiring no changes to end systems and no changes to the exterior routing system. In addition to routing scaling, IRON further addresses other important issues including mobility management, mobile networks, multihoming, traffic engineering, NAT traversal and security. While business considerations are an important determining factor for widespread adoption, they are out of scope for this document.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 19, 2013.

Table of Contents

1. Introduction 4
2. Differences With RFC6179 5
3. Terminology 6
4. The Interior Routing Overlay Network 7
4.1. IRON Client 9
4.2. IRON Serving Router 10
4.3. IRON Relay Router 11
5. IRON Organizational Principles 12
6. IRON Control Plane Operation 14
6.1. IRON Client Operation 14
6.2. IRON Server Operation 14
6.3. IRON Relay Operation 15
7. IRON Forwarding Plane Operation 15
7.1. IRON Client Operation 15
7.2. IRON Server Operation 16
7.3. IRON Relay Operation 17
8. IRON Reference Operating Scenarios 18
8.1. Both Hosts within Same IRON Instance 18
8.1.1. EUNs Served by Same Server 18
8.1.2. EUNs Served by Different Servers 20
8.1.3. Client-to-Client Tunneling 22
8.2. Mixed IRON and Non-IRON Hosts 23
8.2.1. From IRON Host A to Non-IRON Host B 23
8.2.2. From Non-IRON Host B to IRON Host A 25
8.3. Hosts within Different IRON Instances 26
9. Mobility, Multiple Interfaces, Multihoming, and Traffic Engineering 26
9.1. Mobility Management and Mobile Networks 27
9.2. Multiple Interfaces and Multihoming 27
9.3. Traffic Engineering 28
10. Renumbering Considerations 28
11. NAT Traversal Considerations 28
12. Multicast Considerations 29
13. Nested EUN Considerations 29
13.1. Host A Sends Packets to Host Z 31

13.2. Host Z Sends Packets to Host A 31
14. Implications for the Internet 32
15. Additional Considerations 33
16. Related Initiatives 33
17. IANA Considerations 34
18. Security Considerations 34
19. Acknowledgements 35
20. References 36
20.1. Normative References 36
20.2. Informative References 36
Appendix A. IRON Operation over Internetworks with Different Address Families 39
Appendix B. Scaling Considerations 40
Author's Address 41

1. Introduction

Growth in the number of prefix entries instantiated in the Internet routing system has led to concerns regarding unsustainable routing scaling [RFC4984][RADIR] [I-D.narten-radir-problem-statement]. Operational practices such as de-aggregation and the increased use of multihoming with Provider-Independent (PI) addressing are resulting in more and more prefixes being injected into the Internet routing system. Furthermore, depletion of the public IPv4 address space has raised concerns for both increased de-aggregation and an impending address space run-out scenario. At the same time, the IPv6 routing system is beginning to see growth [BGPMON] which must be managed in order to avoid the same routing scaling issues the IPv4 Internet now faces. Since the Internet must continue to scale to accommodate increasing demand, it is clear that new methodologies and operational practices for managing Autonomous System (AS) interior routing systems are needed in order to avoid excessive routing scaling due to de-aggregation.

These same issues apply also to Internetworks other than the public Internet, including critical infrastructure networks such as corporate enterprise networks, civil aviation networks, emergency response networks, power grid networks, medical care networks, etc. The architectural principles presented in this document therefore apply equally to any such Internetwork.

Several related works have investigated routing scaling issues. Virtual Aggregation (VA) [GROW-VA] and Aggregation in Increasing Scopes (AIS) [EVOLUTION] are global routing proposals that introduce routing overlays with Virtual Prefixes (VPs) to reduce the number of entries required in each router's Forwarding Information Base (FIB) and Routing Information Base (RIB). Routing and Addressing in Networks with Global Enterprise Recursion (RANGER) [RFC5720] examines recursive arrangements of enterprise networks that can apply to a very broad set of use-case scenarios [RFC6139]. IRON specifically adopts the RANGER Non-Broadcast, Multiple Access (NBMA) tunnel virtual-interface model, and uses Virtual Enterprise Traversal (VET) [INTAREA-VET] the Subnetwork Adaptation and Encapsulation Layer (SEAL) [INTAREA-SEAL] and Asymmetric Extended Route Optimization [RFC6706] as its functional building blocks.

This document introduces an Interior Routing Overlay Network (IRON) architecture with goals of supporting scalable routing and addressing while requiring no changes to the Internetwork's interdomain routing system [RFC4271]. IRON observes the Internet Protocol standards [RFC0791][RFC2460], while other network-layer protocols that can be encapsulated within IP packets (e.g., OSI/CLNP [RFC0994], etc.) are also within scope.

IRON borrows concepts from VA and AIS, and further borrows concepts from the Internet Vastly Improved Plumbing (Ivip) [IVIP-ARCH] architecture proposal along with its associated Translating Tunnel Router (TTR) mobility extensions [TTRMOB]. Indeed, the TTR model to a great degree inspired the IRON mobility architecture design discussed in this document. The Network Address Translator (NAT) traversal techniques adapted for IRON were inspired by the Simple Address Mapping for Premises Legacy Equipment (SAMPLE) proposal [SAMPLE] [I-D.carpenter-softwire-sample] and by Teredo [RFC4380].

IRON is specifically adapted for Virtual Service Provider (VSP) overlay networks that connect to the Internetwork as an AS and service Aggregated Prefixes (APs) from which more-specific Client Prefixes (CPs) are delegated. IRON is motivated by a growing end user demand for mobility management, mobile networks, multihoming, traffic engineering, NAT traversal and security while using stable addressing to minimize dependence on network renumbering [RFC4192][RFC5887]. IRON VSP overlay network instances use the existing IPv4 and/or IPv6 Internetwork as virtual NBMA links for tunneling inner network layer packets within outer network layer headers (see Section 4). Each IRON instance requires deployment of a small number of relays and servers in the Internetwork, as well as client devices that connect End User Networks (EUNs). No modifications to hosts, and no modifications to existing routers, are required. The following sections discuss details of the IRON architecture.

2. Differences With RFC6179

An earlier version of IRON was published as RFC6179. This version clarifies that IRON operates at the intradomain level within an AS, and is therefore not intended as an interdomain solution. IRON is therefore complimentary with the approaches documented in interdomain solutions such as the Identifier / Locator Network Protocol (ILNP) [RFC6740] and the Locator I/D Split Protocol (LISP) [RFC6830]. This version of IRON further introduces significant improvements in security and route optimization, as well as a direct client-to-client route optimization capability not found in RFC6179.

Some terminology has been changed for greater clarification, including Virtual Service Provier (VSP), Aggregated Prefix (AP) and Client Prefix (CP). This document further introduces Asymmetric Extended Route Optimization (AERO) [RFC6706] as the primary route discovery mechanism. The document finally adds a new section on renumbering considerations and adds enhanced security considerations.

3. Terminology

This document makes use of the following terms:

Aggregated Prefix (AP):
a short network-layer prefix (e.g., an IPv4 /16, an IPv6 /20, an OSI Network Service Access Protocol (NSAP) prefix, etc.) that is owned and managed by a Virtual Service Provider (VSP).

Client Prefix (CP):
a more-specific network-layer prefix (e.g., an IPv4 /28, an IPv6 /56, etc.) derived from an AP and delegated to a client end user network.

Client Prefix Address (CPA):
a network-layer address belonging to a CP and assigned to an interface in an End User Network (EUN).

End User Network (EUN):
an edge network that connects an end user's devices (e.g., computers, routers, printers, etc.) to the Internetwork. IRON EUNs are mobile networks, and can change their ISP attachments without having to renumber.

Interior Routing Overlay Network (IRON):
an AS-interior overlay network instance that appears as a virtual enterprise network, and connects to the Internetwork the same as for any AS.

IRON Client Router/Host ("Client"):
a customer device that logically connects EUNs to an IRON instance via an NBMA tunnel virtual interface. The device is normally a router, but may instead be a host if the "EUN" is a singleton end system.

IRON Serving Router ("Server"):
a VSP's IRON instance router that provides forwarding and mapping services for Clients.

IRON Relay Router ("Relay"):
a VSP's router that acts as a relay between the IRON instance and the Internetwork.

IRON Agent (IA):
generically refers to any of an IRON Client/Server/Relay.

IRON Instance:
a set of IRON Agents deployed by a VSP to service EUNs through automatic tunneling over the Internetwork.

Internetwork Service Provider (ISP):
a service provider that connects an IA to the Internetwork. In other words, an ISP is responsible for providing IAs with data link services for basic connectivity.

Locator:
an IP address assigned to the interface of a router or end system connected to a public or private network over which tunnels are formed. Locators taken from public IP prefixes are routable on a global basis, while locators taken from private IP prefixes [RFC1918] are made public via Network Address Translation (NAT).

Routing and Addressing in Networks with Global Enterprise Recursion (RANGER):
an architectural examination of virtual overlay networks applied to enterprise network scenarios, with implications for a wider variety of use cases.

Subnetwork Encapsulation and Adaptation Layer (SEAL):
an encapsulation sublayer that provides extended identification fields and control messages to ensure deterministic network-layer feedback.

Virtual Enterprise Traversal (VET):
a method for discovering border routers and forming dynamic tunnel neighbor relationships over enterprise networks (or sites) with varying properties.

Asymmetric Extended Route Optimization (AERO):
a means for a destination IA to securely inform a source IA of a more direct path.

Virtual Service Provider (VSP):
a company that owns and manages a set of APs from which it delegates CPs to EUNs.

VSP Overlay Network:
the same as defined above for IRON Instance.

4. The Interior Routing Overlay Network

The Interior Routing Overlay Network (IRON) operates at the AS level and provides a number of important services to End User Networks

(EUNs) that are not well supported in the current architecture, including routing scaling, mobility management, mobile networks, multihoming, traffic engineering and NAT traversal. This is accomplisehd through the establishment of IRON instances as overlays configured over the underlying Internetwork.

Each IRON instance consists of IRON Agents (IAs) that automatically tunnel the packets of end-to-end communication sessions within encapsulating headers used for Internetwork routing. IAs use the Virtual Enterprise Traversal (VET) [INTAREA-VET] virtual NBMA link model in conjunction with the Subnetwork Encapsulation and Adaptation Layer (SEAL) [INTAREA-SEAL] to encapsulate inner network-layer packets within outer network layer headers, as shown in Figure 1.

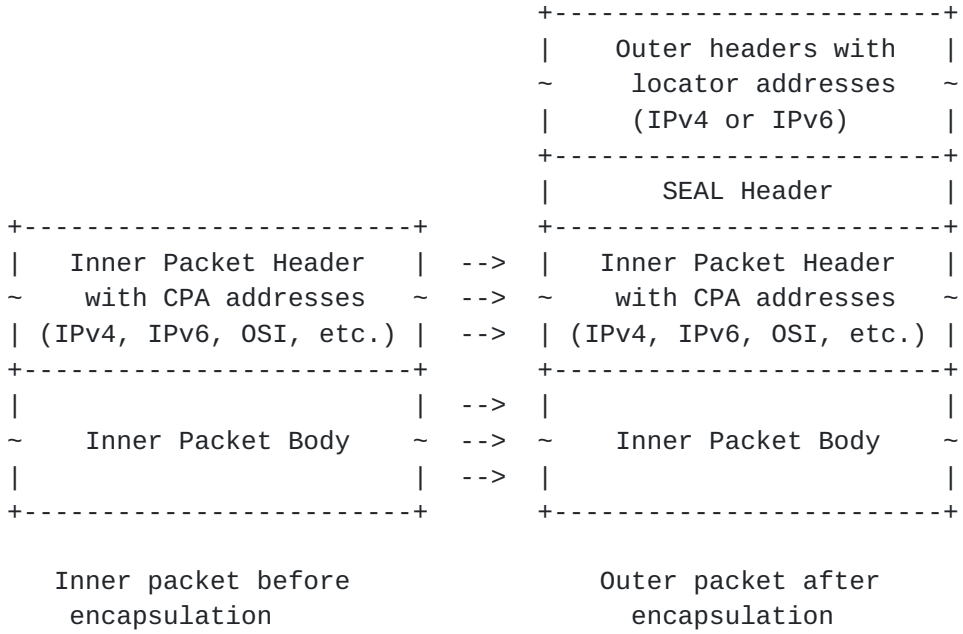


Figure 1: Encapsulation of Inner Packets within Outer IP Headers

VET specifies automatic tunneling and tunnel neighbor coordination mechanisms, where IAs appear as neighbors on an NBMA tunnel virtual link. SEAL specifies the format and usage of the SEAL encapsulating header. Additionally, Asymmetric Extended Route Optimization (AERO) [RFC6706] specifies the method for route optimization to reduce routing path stretch. Together, these documents specify a set of control messages used to deterministically exchange and authenticate neighbor discovery messages, route redirections, indications of Path Maximum Transmission Unit (PMTU) limitations, destination unreachables, etc.

Each IRON instance comprises a set of IAs distributed throughout the Internetwork to provide routing services for a set of Aggregated

Prefixes (APs). (The APs may be owned either by the VSP, or by an enterprise network customer that hires the VSP to manage its APs.) VSPs delegate sub-prefixes from APs, which they provide to end users as Client Prefixes (CPs). In turn, end users assign CPs to Client IAs which connect their End User Networks (EUNs) to the VSP IRON instance.

VSPs may have no affiliation with the ISP networks from which end users obtain their basic Internetwork connectivity. In that case, the VSP can service its end users without the need to coordinate its activities with ISPs or other VSPs. Further details on VSP business considerations are out of scope for this document.

IRON requires no changes to end systems or to existing routers. Instead, IAs are deployed either as new platforms or as modifications to existing platforms. IAs may be deployed incrementally without disturbing the existing Internetwork routing system, and act as waypoints (or "cairns") for navigating VSP overly networks. The functional roles for IAs are described in the following sections.

4.1. IRON Client

An IRON Client (or, simply, "Client") is a router that logically connects EUNs to the VSP's IRON instance via tunnels, as shown in Figure 2. Clients obtain CPs from their VSPs and use them to number subnets and interfaces within the EUNs.

Each Client connects to one or more Servers in the IRON instance which serve as default routers. The Servers in turn consider this class of Clients as "dependent" Clients. Clients also dynamically discover destination-specific Servers through the receipt of redirection messages. These destination-specific Servers in turn consider this class of Clients as "visiting" Clients.

A Client can be deployed on the same physical platform that also connects EUNs to the end user's ISPs, but it may also be deployed as a separate router within the EUN. (This model applies even if the EUN connects to the ISP via a Network Address Translator (NAT) -- see Section 7.7). Finally, a Client may also be a simple end system that connects a singleton EUN and exhibits the outward appearance of a host.

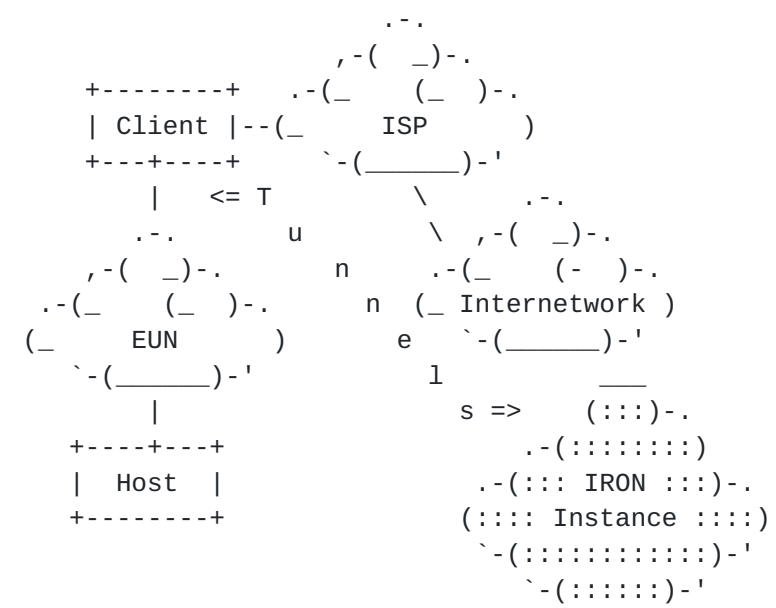


Figure 2: IRON Client Connecting EUN to IRON Instance

4.2. IRON Serving Router

An IRON serving router (or, simply, "Server") is a VSP's router that provides forwarding and mapping services within the IRON instance for the CPs that have been delegated to end user Clients. In typical deployments, a VSP will deploy many Servers for the IRON instance in a globally distributed fashion (e.g., as depicted in Figure 3) around the Internetwork so that Clients can discover those that are nearby.

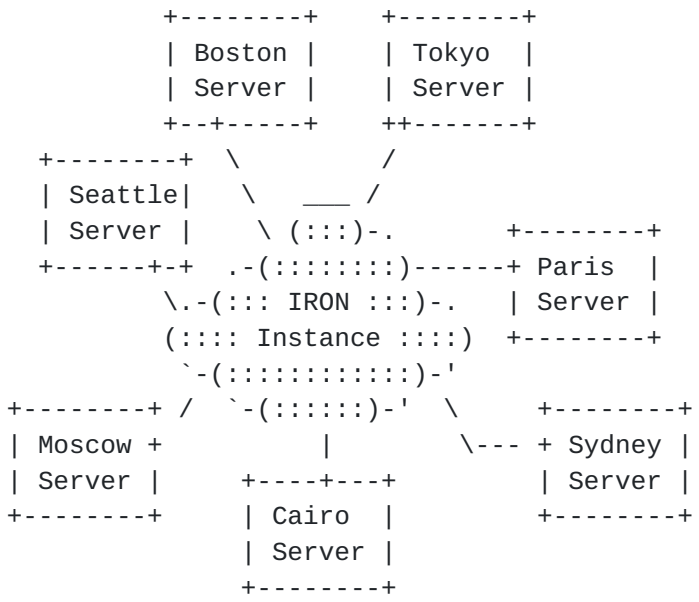


Figure 3: IRON Server Global Distribution Example

Each Server acts as a tunnel-endpoint router. The Server forms bidirectional tunnel neighbor relationships with each of its dependent Clients, and can also serve as the unidirectional tunnel neighbor egress for dynamically discovered visiting Clients. (The Server can also form bidirectional tunnel neighbor relationships with visiting Clients, e.g., if a symmetric security association is necessary.) Each Server also forms bidirectional tunnel neighbor relationships with a set of Relays that can forward packets from the IRON instance out to the native Internetwork and vice versa, as discussed in the next section.

4.3. IRON Relay Router

An IRON Relay Router (or, simply, "Relay") is a router that connects the VSP's IRON instance to the Internetwork as an AS. The Relay therefore also serves as an Autonomous System Border Router (ASBR) that is owned and managed by the VSP.

Each VSP configures one or more Relays that advertise the VSP's APs into the IPv4 and/or IPv6 Internetwork routing systems. Each Relay associates with the VSP's IRON instance Servers, e.g., via tunnel virtual links over the IRON instance, via a physical interconnect such as an Ethernet cable, etc. The Relay role is depicted in Figure 4.

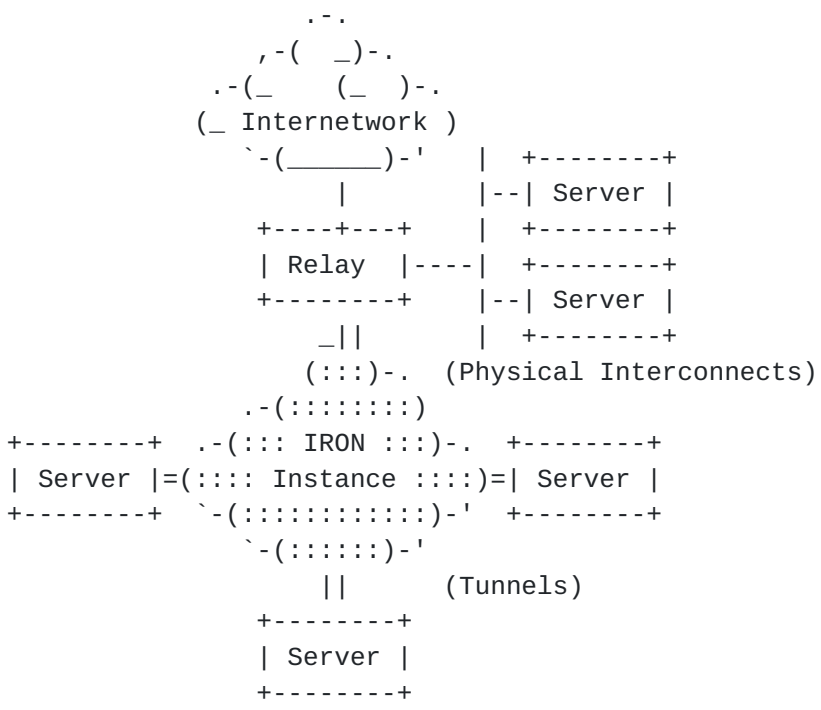
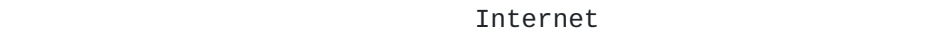


Figure 4: IRON Relay Router Connecting IRON Instance to Native



5. IRON Organizational Principles

Each IRON instance represents a distinct "patch" on the underlying Internetwork "quilt", where the patches are stitched together by standard routing. When a new IRON instance is deployed, it becomes yet another patch on the quilt and coordinates its internal routing system independently of all other patches.

Each IRON instance connects to the Internetwork as an AS in the interdomain routing system using a public Border Gateway Protocol (BGP) Autonomous System Number (ASN). The IRON instance maintains a set of Relays that serve as ASBRs as well as a set of Servers that provide routing and addressing services to Clients. Figure 5 depicts the logical arrangement of Relays, Servers, and Clients in an IRON instance.

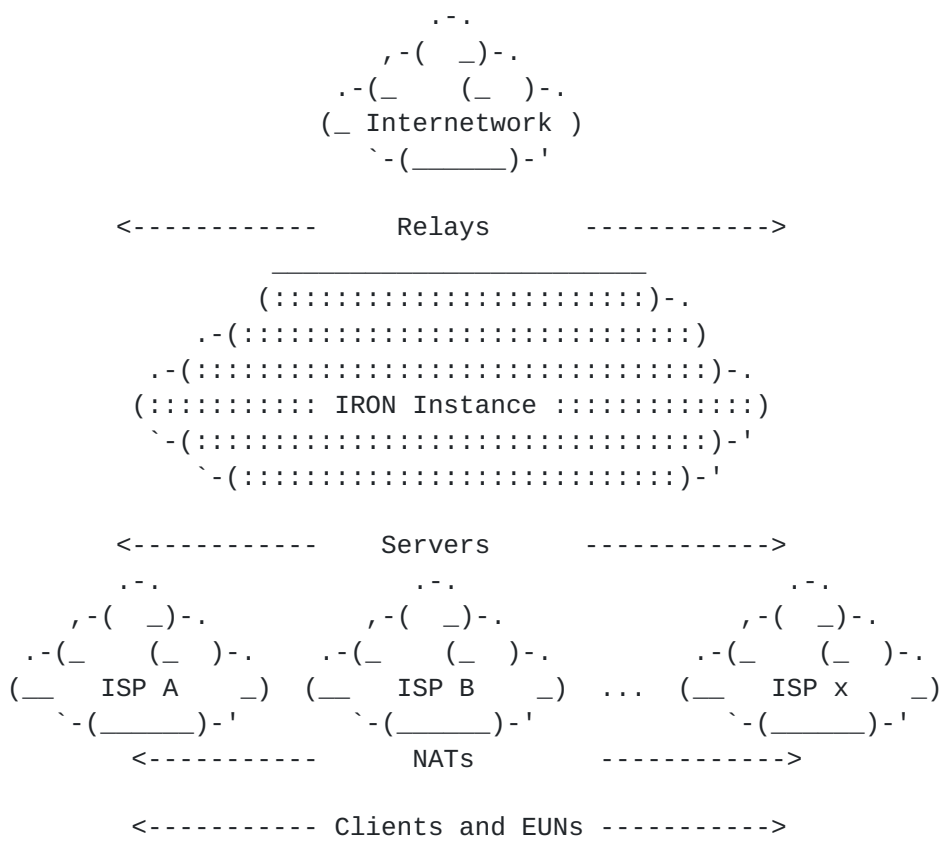


Figure 5: IRON Organization

Each Relay connects the IRON instance directly to the underlying IPv4 and/or IPv6 Internetworks via external BGP (eBGP) peerings with

neighboring ASes. It also advertises the IPv4 APs managed by the VSP into the IPv4 Internetwork routing system and advertises the IPv6 APs managed by the VSP into the IPv6 Internetwork routing system. Relays will therefore receive packets with CPA destination addresses sent by end systems in the Internetwork and forward them to a Server that connects the Client to which the corresponding CP has been delegated. Finally, the IRON instance Relays maintain synchronization by running interior BGP (iBGP) between themselves the same as for ordinary ASBRs.

In a simple VSP overlay network arrangement, each Server can be configured as an ASBR for a stub AS using a private ASN [RFC1930] to peer with each IRON instance Relay the same as for an ordinary eBGP neighbor. (The Server and Relay functions can instead be deployed together on the same physical platform as a unified gateway.) Each Server maintains a working set of dependent Clients for which it caches CP-to-Client mappings in its forwarding table. Each Server also, in turn, propagates the list of CPs in its working set to its neighboring Relays via eBGP. Therefore, each Server only needs to track the CPs for its current working set of dependent Clients, while each Relay will maintain a full CP-to-Server forwarding table that represents reachability information for all CPs in the IRON instance.

Each Client obtains its basic Internetwork connectivity from ISPs, and connects to Servers to attach its EUNs to the IRON instance. Each EUN can further connect to the IRON instance via multiple Clients as long as the Clients coordinate with one another, e.g., to mitigate EUN partitions. Clients may additionaly use private addresses behind one or several layers of NATs. Each Client initially discovers a list of nearby Servers then forms a bidirectional tunnel neighbor relationship with one or more Servers through an initial exchange followed by periodic keepalives.

After a Client connects to Servers, it forwards initial outbound packets from its EUNs by tunneling them to a Server, which may, in turn, forward them to a nearby Relay within the IRON instance. The Client may subsequently receive redirection messages informing it of a more direct route through a different IA within the IRON instance that serves the final destination EUN.

IRON can also be used to support APs of network-layer address families that cannot be routed natively in the underlying Internetwork (e.g., OSI/CLNP over the public Internet, IPv6 over IPv4-only Internetworks, IPv4 over IPv6-only Internetworks, etc.). Further details for the support of IRON APs of one address family over Internetworks based on different address families are discussed in Appendix A.

6. IRON Control Plane Operation

Each IRON instance supports routing through the control plane startup and runtime dynamic routing operation of IAs. The following sub-sections discuss control plane considerations for initializing and maintaining the IRON instance routing system.

6.1. IRON Client Operation

Each Client obtains one or more CPs in a secured exchange with the VSP as part of the initial end user registration. Upon startup, the Client discovers a list of nearby VSP Servers via, e.g., a location broker, a well known website, a static map, etc.

After the Client obtains a list of nearby Servers, it initiates short transactions to connect to one or more Servers, e.g., via secured TCP connections. During the transaction, each Server provides the Client with a CP and a symmetric secret key that the Client will use to sign and authenticate messages. The Client in turn provides the Server with a set of link identifiers ("LINK_ID"s) that represent the Client's ISP connections. Finally, the Client provides a "willingness" indication as to whether or not it will accept direct Client-to-Client communications without involving the Server as an intermediary. The protocol details of the connection transaction are specific to the VSP, and hence out of scope for this document.

After the Client connects to Servers, it configures default routes that list the Servers as next hops on the tunnel virtual interface. The Client may subsequently discover more-specific routes through receipt of redirection messages.

6.2. IRON Server Operation

In a simple VSP overlay network arrangement, each IRON Server is provisioned with the locators for Relays within the IRON instance. The Server is further configured as an ASBR for a stub AS and uses eBGP with a private ASN to peer with each Relay.

Upon startup, the Server uses eBGP to announce the list of CPs it is currently serving to the overlay network Relays. The Server then actively listens for Clients that register their CPs as part of the connection establishment procedure described in Section 6.1. When a new Client connects, the Server uses eBGP to announce the new CP routes to its neighboring Relays; when an existing Client disconnects, the Server withdraws its CP announcements. This process can often be accommodated through standard eBGP router configurations, e.g., on routers that can announce and withdraw prefixes based on kernel route additions and deletions.

6.3. IRON Relay Operation

Each IRON Relay is provisioned with the list of APs that it will serve, as well as the locators for Servers within the IRON instance. The Relay is also provisioned with eBGP peerings with neighboring ASes in the Internetwork -- the same as for any ASBR.

In a simple VSP overlay network arrangement, each Relay connects to each Server via IRON instance-internal eBGP peerings for the purpose of discovering CP-to-Server mappings, and connects to all other Relays using iBGP either in a full mesh or using route reflectors. (The Relay only uses iBGP to announce those prefixes it has learned from AS peerings external to the IRON instance, however, since all Relays will already discover all CPs in the IRON instance via their eBGP peerings with Servers.) The Relay then engages in eBGP routing exchanges with peer ASes in the IPv4 and/or IPv6 Internetworks the same as for any ASBR.

After this initial synchronization procedure, the Relay advertises the APs to its eBGP peers in the Internetwork. In particular, the Relay advertises the IPv6 APs into the IPv6 interdomain routing system and advertises the IPv4 APs into the IPv4 interdomain routing system, but it does not advertise the full list of the IRON overlay's CPs to any of its eBGP peers. The Relay further advertises "default" via eBGP to its associated Servers, then engages in ordinary packet-forwarding operations.

7. IRON Forwarding Plane Operation

Following control plane initialization, IAs engage in the cooperative process of receiving and forwarding packets. IAs forward encapsulated packets over the IRON instance using the mechanisms of VET [INTAREA-VET], SEAL [INTAREA-SEAL] and AERO [RFC6706], while Relays additionally forward packets to and from the native IPv6 and/or IPv4 Internetworks. IAs also use VET, SEAL and AERO control messages to coordinate with other IAs, including the process of sending and receiving redirection messages, error messages, etc. Each IA operates as specified in the following sub-sections.

7.1. IRON Client Operation

After connecting to Servers as specified in Section 6.1, the Client registers its active ISP connections with each of its connected Servers. Thereafter, the Client sends periodic beacons (e.g., cryptographically signed SEAL Control Message Protocol (SCMP) Router Solicitation (SRS) messages) to the Server via each ISP connection to maintain tunnel neighbor address mapping state. The beacons should

be sent at no more than 60 second intervals (subject to a small random delay) so that state in NATs on the path as well as on the Server itself is refreshed regularly. Although the Client may connect via multiple ISPs (each represented by a different LINK_ID), the CP itself is used to represent the bidirectional Client-to-Server tunnel neighbor association. The CP therefore names this "bundle" of ISP connections.

If the Client ceases to receive acknowledgements from a Server via a specific ISP connection, it marks the Server as unreachable from that ISP. (The Client should also inform the Server of this outage via one of its working ISP connections.) If the Client ceases to receive acknowledgements from the Server via multiple ISP connections, it disconnects from the failing Server and connects to a different nearby Server. The act of disconnecting from old servers and connecting to new servers will soon propagate the appropriate routing information among the IRON instance's Relays.

When an end system in an EUN sends a flow of packets to a correspondent in a different network, the packets are forwarded through the EUN via normal routing until they reach the Client, which then tunnels the initial packets to one of its connected Servers as its default router. In particular, the Client encapsulates each packet in outer headers with its locator as the source address and the locator of the Server as the destination address.

The Client uses the mechanisms specified in VET and SEAL to encapsulate each packet to be forwarded, and uses the redirection procedures described in AERO to coordinate route optimization. The Client further accepts control messages from its Servers, including neighbor coordination exchanges, indications of PMTU limitations, redirections and other control messages. When the Client is redirected to a foreign Server that serves a destination CP, it forms a unidirectional tunnel neighbor association with the foreign Server as the new next hop toward the CP. (The visiting Client can also form a bidirectional tunnel neighbor association with the foreign Server, e.g., if a symmetric security association is necessary.)

Note that Client-to-Client tunneling is also enabled when the foreign Client has indicated its willingness to accept Client-to-Client communications. In that case, the foreign Server can allow the final destination Client to return the redirection message, which removes the foreign Server from the fowarding path.

7.2. IRON Server Operation

After the Server associates with nearby Relays, it accepts Client connections and authenticates the SRS messages it receives from its

already-connected Clients. The Server discards any SRS messages that failed authentication, and responds to authentic SRS messages by returning signed SCMP Router Advertisement (SRA) messages.

When the Server receives a SEAL-encapsulated data packet from one of its dependent Clients, it uses normal longest-prefix-match rules to locate a forwarding table entry that matches the packet's inner destination address. The Server then re-encapsulates the packet (i.e., it removes the outer header and replaces it with a new outer header), sets the outer destination address to the locator address of the next hop and forwards the packet to the next hop.

When the Server receives a SEAL-encapsulated data packet from a visiting Client, it accepts the packet only if the packet's signature is correct; otherwise, it silently drops the packet. The Server then locates a forwarding table entry that matches the packet's inner destination address. If the destination does not correspond to one of the Server's dependent Clients, the Server silently drops the packet. Otherwise, the Server re-encapsulates the packet and forwards it to the correct dependent Client. If the Client is in the process of disconnecting (e.g., due to mobility), the Server also returns a redirection message listing a NULL next hop to inform the visiting Client that the dependent Client has moved.

When the Server receives a SEAL-encapsulated data packet from a Relay, it again locates a forwarding table entry that matches the packet's inner destination. If the destination does not correspond to one of the Server's dependent Clients, the Server drops the packet and sends a destination unreachable message. Otherwise, the Server re-encapsulates the packet and forwards it to the correct dependent Client.

7.3. IRON Relay Operation

After each Relay has synchronized its APs (see Section 6.3) it advertises them in the IPv4 and/or IPv6 interdomain routing systems. These APs will be represented as ordinary routing information in the interdomain routing system, and any packets originating from the IPv4 or IPv6 Internetwork destined to an address covered by one of the APs will be forwarded to one of the VSP's Relays.

When a Relay receives a packet from the Internetwork destined to a CPA covered by one of its APs, it behaves as an ordinary IP router. Specifically, the Relay looks in its forwarding table to discover a locator of a Server that serves the CP covering the destination address. The Relay then simply forwards the packet to the Server, e.g., via SEAL encapsulation over a tunnel virtual link, via a physical interconnect, etc.

When a Relay receives a packet from a Server destined to a CPA serviced by a different Server, the Relay forwards the packet toward the correct Server while also sending a "predirect" indication as the initial leg in the AERO redirection procedure. When the target IA returns a redirection message, the Relay proxies the message by re-encapsulating it and forwarding it to the previous hop.

8. IRON Reference Operating Scenarios

The following sections discuss the IRON reference operating scenarios.

8.1. Both Hosts within Same IRON Instance

When both hosts are within EUNs served by the same IRON instance, it is sufficient to consider the scenario in a unidirectional fashion, i.e., by tracing packet flows only in the forward direction from source host to destination host. The reverse direction can be considered separately and incurs the same considerations as for the forward direction. The simplest case occurs when the EUNs that service the source and destination hosts are connected to the same server, while the general case occurs when the EUNs are connected to different Servers. The two cases are discussed in the following sections.

8.1.1. EUNs Served by Same Server

In this scenario, the packet flow from the source host is forwarded through the EUN to the source's IRON Client. The Client then tunnels the packets to the Server, which simply re-encapsulates and forwards the tunneled packets to the destination's Client. The destination's Client then removes the packets from the tunnel and forwards them over the EUN to the destination. Figure 6 depicts the sustained flow of packets from Host A to Host B within EUNs serviced by the same Server via a "hairpinned" route:

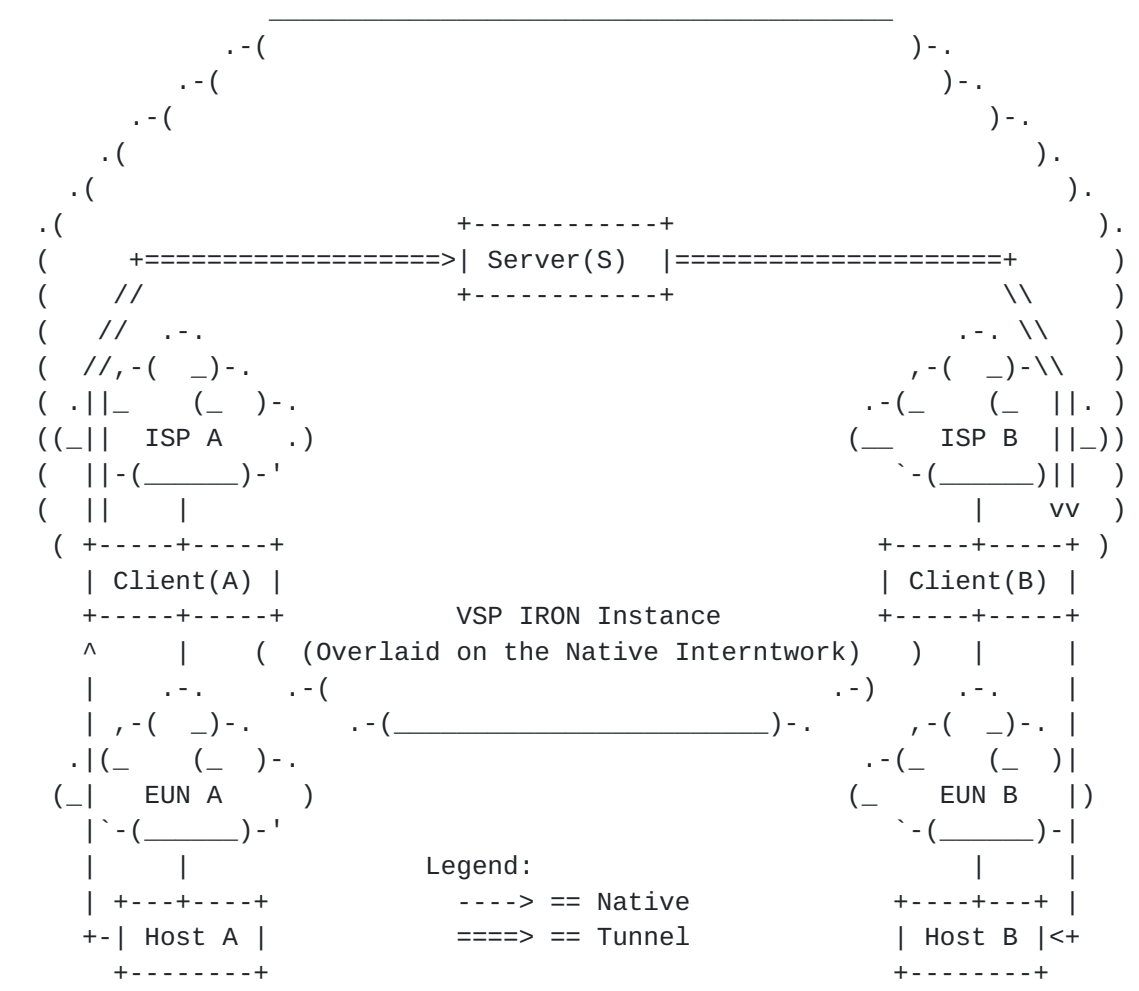


Figure 6: Sustained Packet Flow via Hairpinned Route

With reference to Figure 6, Host A sends packets destined to Host B via its network interface connected to EUN A. Routing within EUN A will direct the packets to Client(A) as a default router for the EUN, which then encapsulates them in outer IP/*/SEAL headers with its locator address as the outer source address, the locator address of Server(S) as the outer destination address, and the identifying information associated with its tunnel neighbor state as the identity. Client(A) then simply forwards the encapsulated packets into the ISP network connection that provided its locator. The ISP will forward the encapsulated packets into the Internetwork without filtering since the (outer) source address is topologically correct. Once the packets have been forwarded into the Internetwork, routing will direct them to Server(S).

Server(S) will receive the encapsulated packets from Client(A) then check its forwarding table to discover an entry that covers destination address B with Client(B) as the next hop. Server(S) then

re-encapsulates the packets in a new outer header that uses the source address, destination address, and identification parameters associated with the tunnel neighbor state for Client(B). Server(S) then forwards these re-encapsulated packets into the Internetwork, where routing will direct them to Client(B). Client(B) will, in turn, decapsulate the packets and forward the inner packets to Host B via EUN B.

8.1.2. EUNs Served by Different Servers

In this scenario, the initial packets of a flow produced by a source host within an EUN connected to the IRON instance by a Client must flow through both the Server of the source host and a nearby Relay, but route optimization can eliminate these elements from the path for subsequent packets in the flow. Figure 7 shows the flow of initial packets from Host A to Host B within EUNs of the same IRON instance:

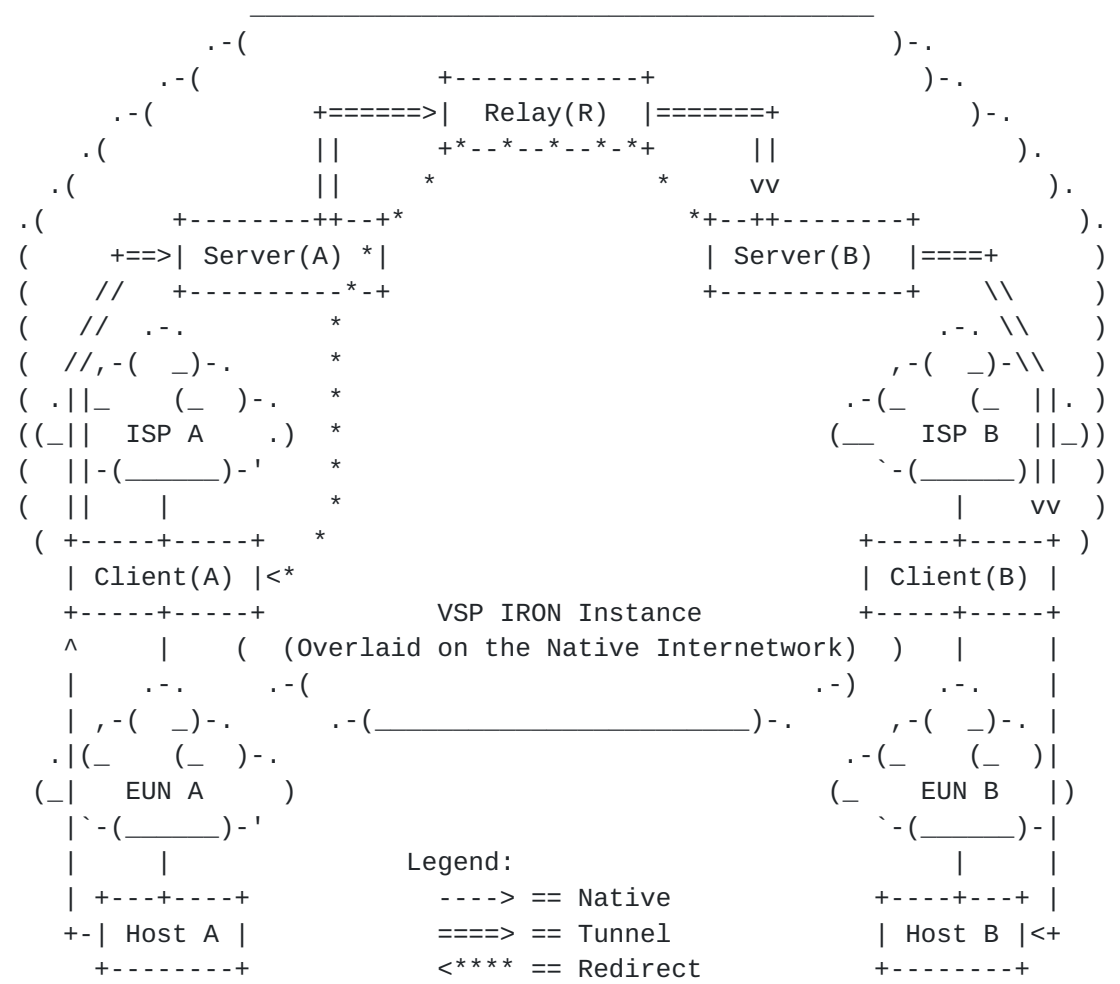


Figure 7: Initial Packet Flow Before Redirects

With reference to Figure 7, Host A sends packets destined to Host B via its network interface connected to EUN A. Routing within EUN A will direct the packets to Client(A) as a default router for the EUN, which then encapsulates them in outer IP/*/SEAL headers that use the source address, destination address, and identification parameters associated with the tunnel neighbor state for Server(A). Client(A) then forwards the encapsulated packets into the ISP network connection that provided its locator, which will forward the encapsulated packets into the Internetwork where routing will direct them to Server(A).

Server(A) receives the encapsulated packets from Client(A) and consults its forwarding table to determine that the most-specific matching route is via Relay(R) as the next hop. Server(A) then re-encapsulates the packets in outer headers that use the source address, destination address, and identification parameters associated with Relay (R), and forwards them into the Internetwork where routing will direct them to Relay(R). (Note that the Server could instead forward the packets directly to the Relay without encapsulation when the Relay is directly connected, e.g., via a physical interconnect.)

Relay(R) receives the forwarded packets from Server(A) then checks its forwarding table to discover a CP entry that covers inner destination address B with Server(B) as the next hop. Relay(R) then sends a "predirect" indication forward to Server(B) to inform the server that a redirection message must be returned. Relay(R) finally re-encapsulates the packets in outer headers that use the source address, destination address, and identification parameters associated with Server(B), then forwards them into the Internetwork where routing will direct them to Server(B). (Note again that the Relay could instead forward the packets directly to the Server, e.g., via a physical interconnect.)

Server(B) receives the "predirect" and forwarded packets from Relay(R), then checks its forwarding table to discover a CP entry that covers destination address B with Client(B) as the next hop. Server(B) returns a redirection message to Relay(R), which proxies the message back to Server(A), which then proxies the message back to Client(A).

Server(B) then re-encapsulates the packets in outer headers that use the source address, destination address, and identification parameters associated with Client(B), then forwards them into the Internetwork where routing will direct them to Client(B). Client(B) will, in turn, decapsulate the packets and forward the inner packets to Host B via EUN B.

After the initial flow of packets, Client(A) will have received one or more redirection messages listing Server(B) as a better next hop, and will establish unidirectional tunnel neighbor state listing Server(B) as the next hop toward the CP that covers Host B. Client(A) thereafter forwards its encapsulated packets directly to the locator address of Server(B) without involving either Server(A) or Relay(B), as shown in Figure 8.

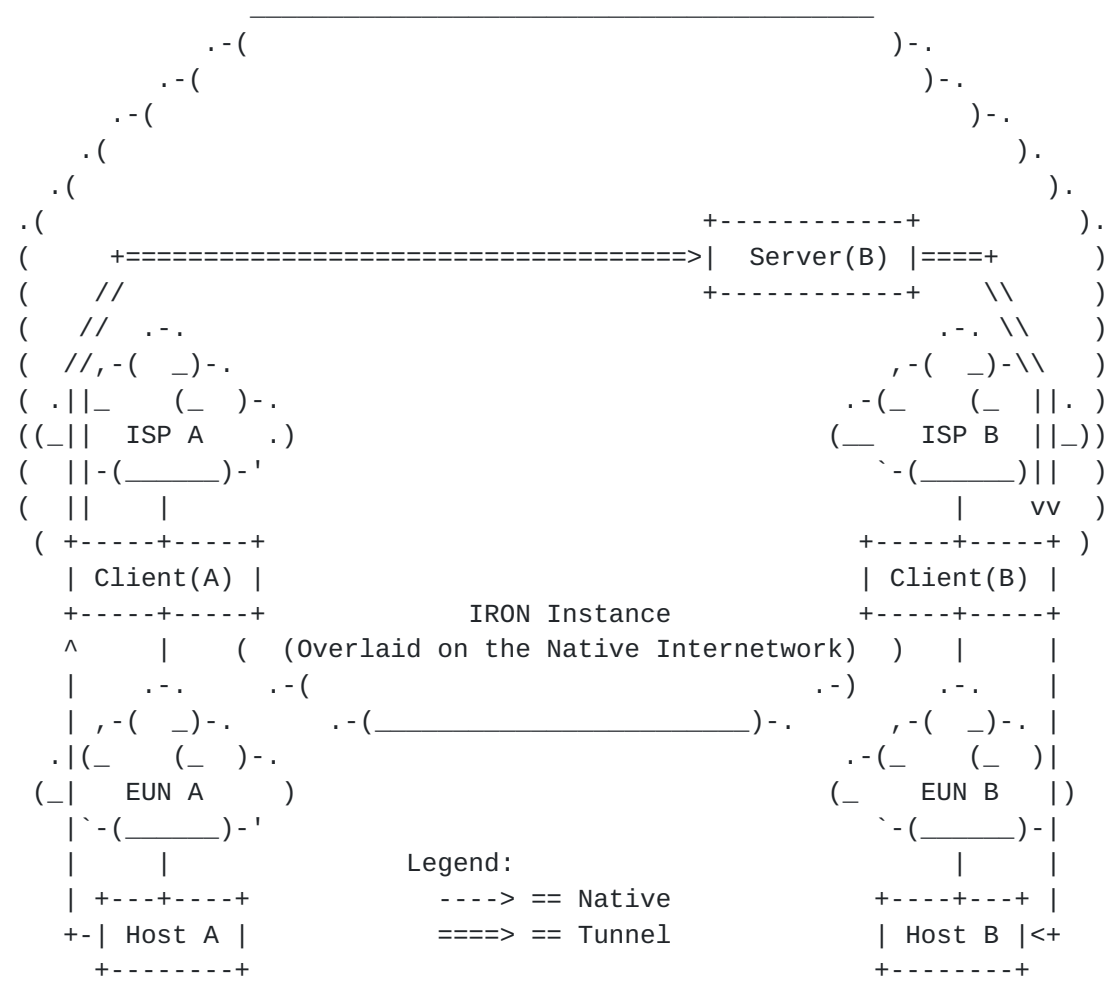


Figure 8: Sustained Packet Flow After Redirects

8.1.3. Client-to-Client Tunneling

In the scenarios shown in Sections 8.1.1 and 8.1.2, if the foreign Client has indicated its willingness to accept Client-to-Client communications, then the foreign Server can allow the foreign Client to return the redirection message, i.e., by passing the "predirect" message on to the Client. In that case, the two Clients become peers in either a unidirectional or bidirectional tunnel neighbor relationship as shown in Figure 9:

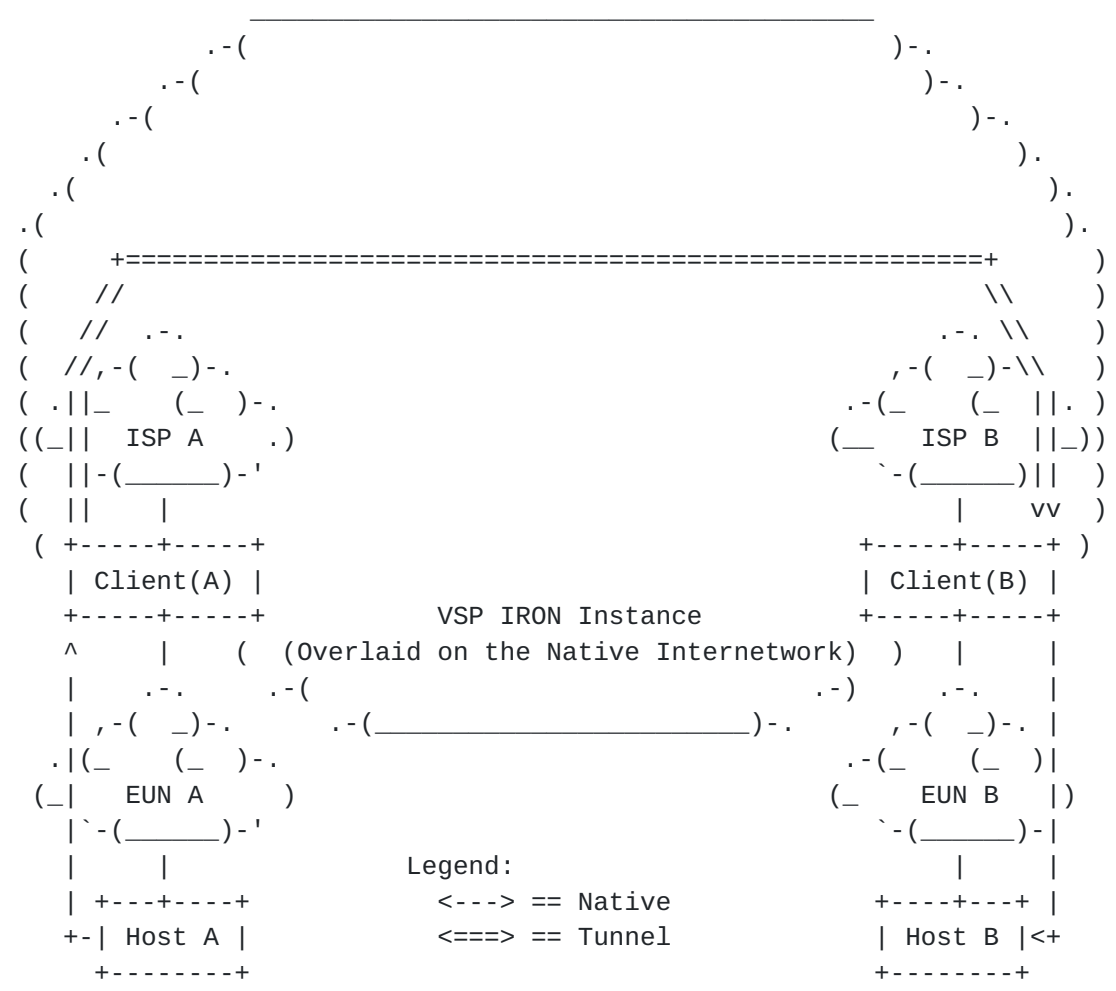


Figure 9: Client-to-Client Tunneling

8.2. Mixed IRON and Non-IRON Hosts

The cases in which one host is within an IRON EUN and the other is in a non-IRON EUN (i.e., one that connects to the native Internetwork instead of the IRON) are described in the following sub-sections.

8.2.1. From IRON Host A to Non-IRON Host B

Figure 10 depicts the IRON reference operating scenario for packets flowing from Host A in an IRON EUN to Host B in a non-IRON EUN.

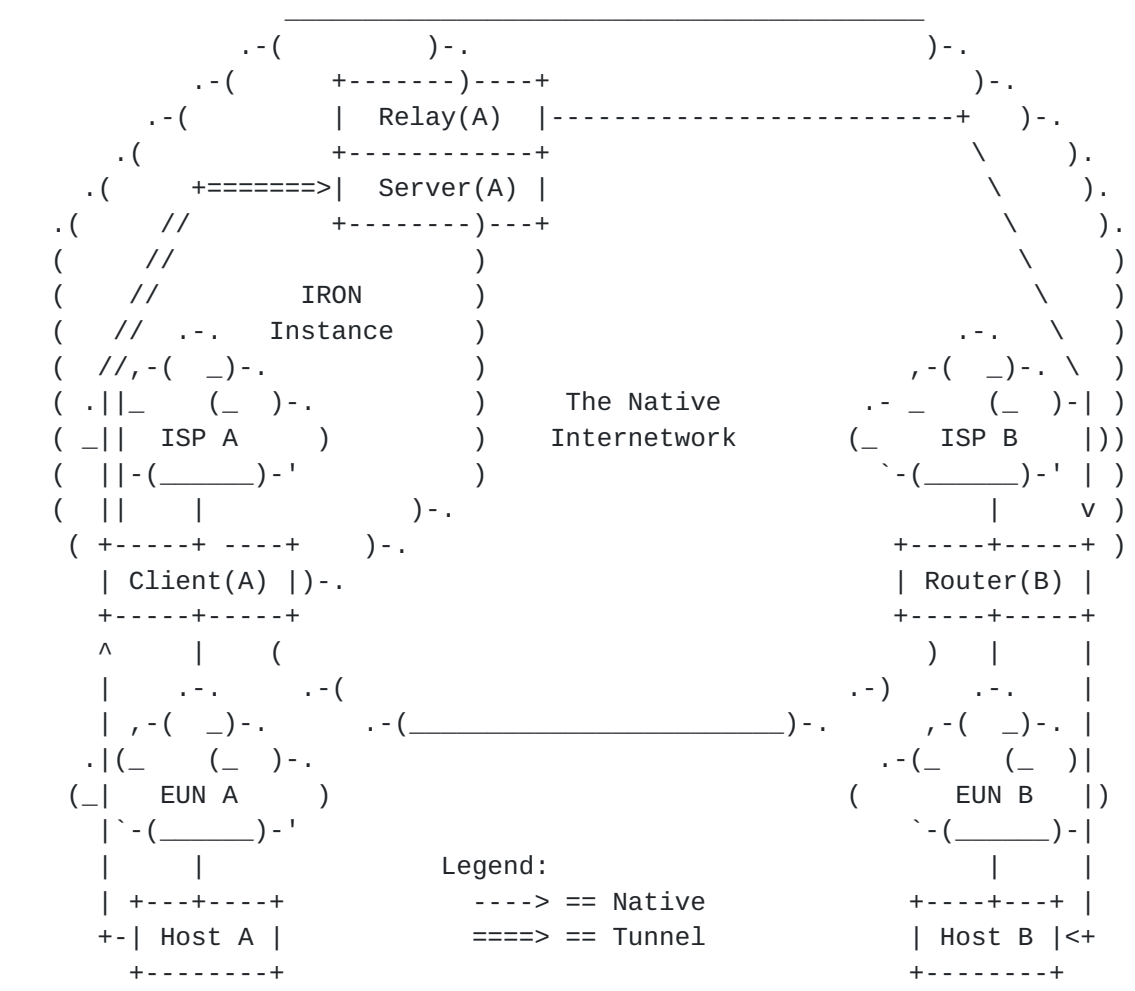


Figure 10: From IRON Host A to Non-IRON Host B

In this scenario, Host A sends packets destined to Host B via its network interface connected to IRON EUN A. Routing within EUN A will direct the packets to Client(A) as a default router for the EUN, which then encapsulates them and forwards them into the Internetwork routing system where they will be directed to Server(A).

Server(A) receives the encapsulated packets from Client(A) then forwards them to Relay(A), which simply forwards the unencapsulated packets into the Internetwork. Once the packets are released into the Internetwork, routing will direct them to the final destination B. (Note that for simplicity Server(A) and Relay(A) are depicted in Figure 10 as two concatenated "half-routers", and the forwarding between the two halves is via encapsulation, via a physical interconnect, via a shared memory operation when the two halves are within the same physical platform, etc.)

8.2.2. From Non-IRON Host B to IRON Host A

Figure 11 depicts the IRON reference operating scenario for packets flowing from Host B in an Non-IRON EUN to Host A in an IRON EUN.

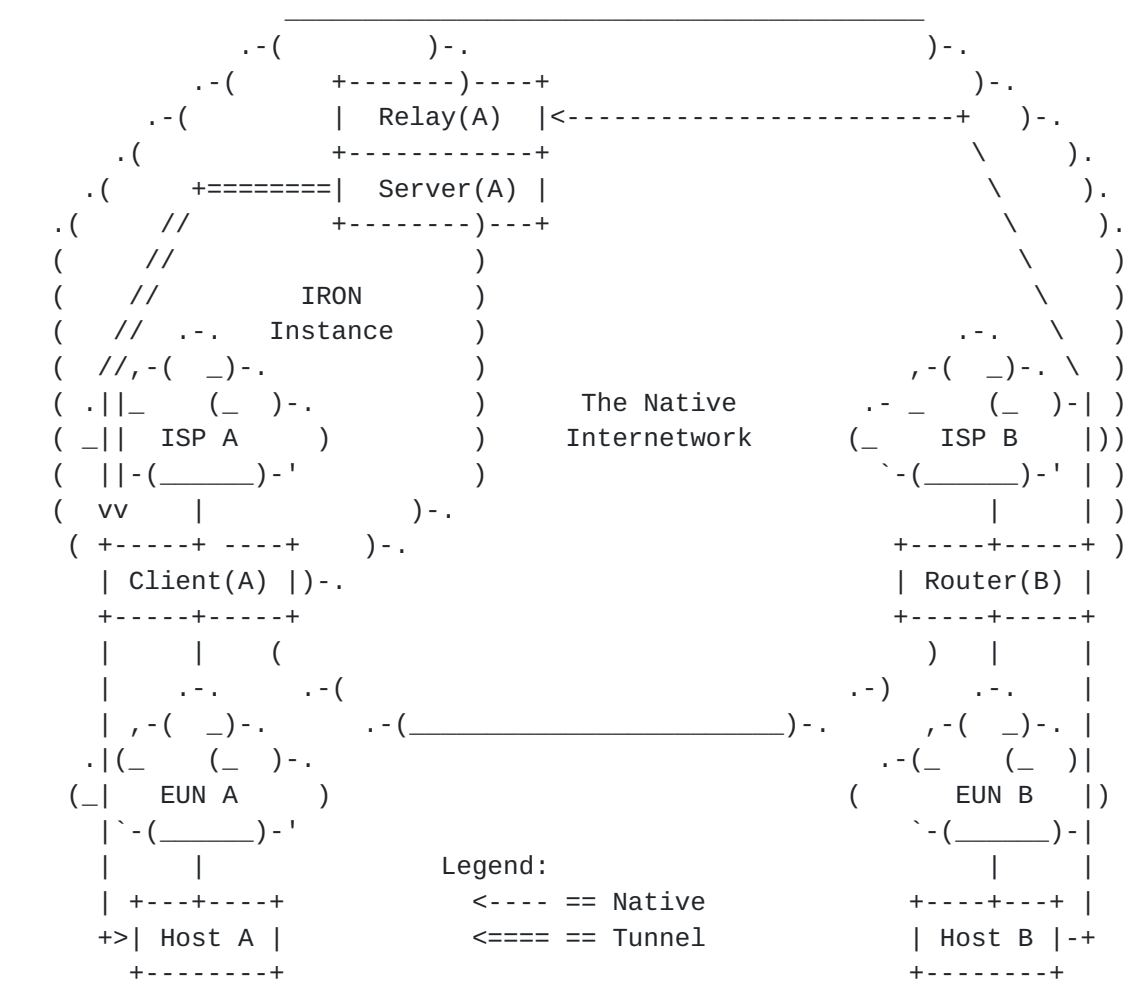


Figure 11: From Non-IRON Host B to IRON Host A

In this scenario, Host B sends packets destined to Host A via its network interface connected to non-IRON EUN B. Interdomain routing will direct the packets to Relay(A), which then forwards them to Server(A).

Server(A) will then check its forwarding table to discover an entry that covers destination address A with Client(A) as the next hop. Server(A) then (re-)encapsulates the packets and forwards them into the Internetwork, where routing will direct them to Client(A). Client(A) will, in turn, decapsulate the packets and forward the inner packets to Host A via its network interface connected to IRON EUN A.

8.3. Hosts within Different IRON Instances

Figure 12 depicts the IRON reference operating scenario for packets flowing between Host A in an IRON instance A and Host B in a different IRON instance B. In that case, forwarding between hosts A and B always involves the Servers and Relays of both IRON instances, i.e., the scenario is no different than if one of the hosts was serviced by an IRON EUN and the other was serviced by a non-IRON EUN.

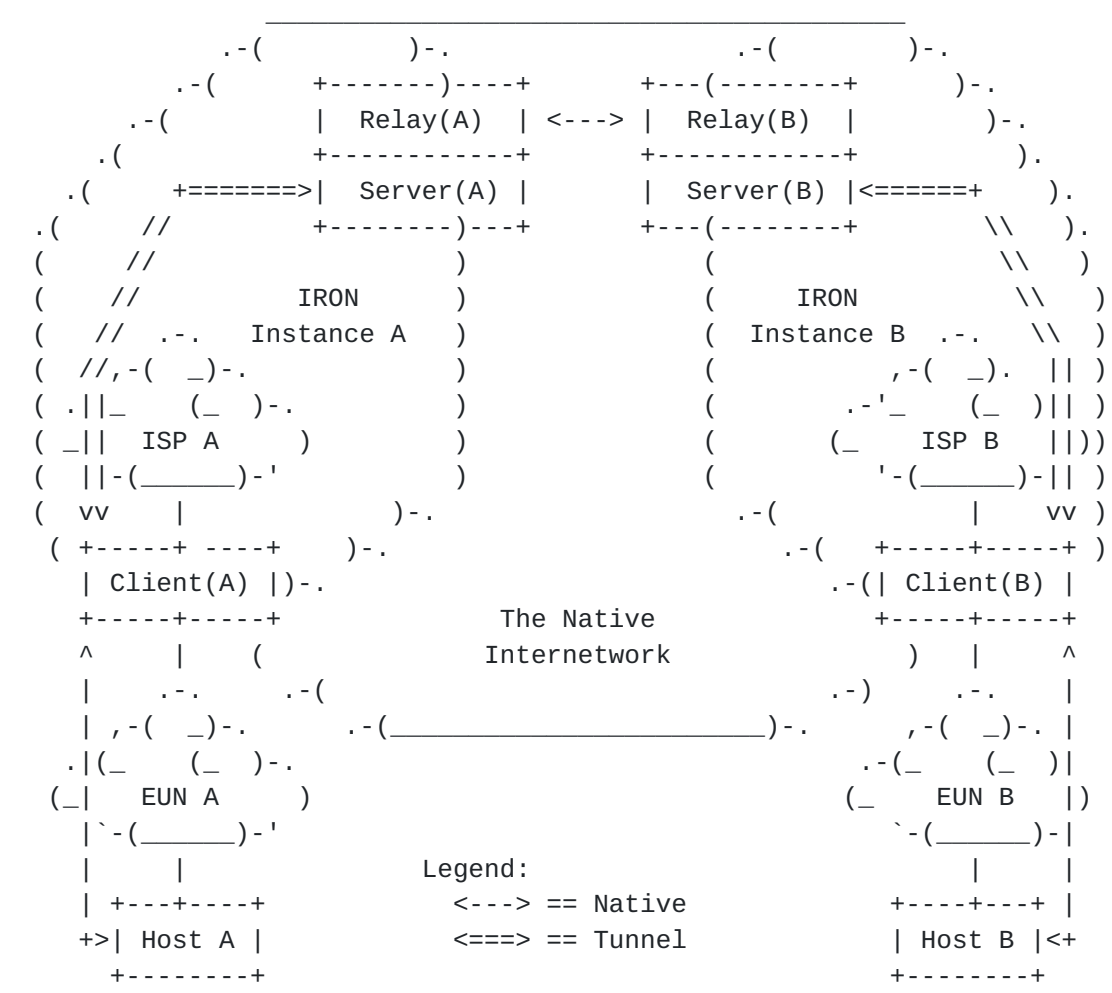


Figure 12: Hosts within Different IRON Instances

9. Mobility, Multiple Interfaces, Multihoming, and Traffic Engineering

While IRON Servers and Relays are typically arranged as fixed infrastructure, Clients may need to move between different network points of attachment, connect to multiple ISPs, or explicitly manage their traffic flows. The following sections discuss mobility, multihoming, and traffic engineering considerations for IRON Clients.

9.1. Mobility Management and Mobile Networks

When a Client changes its network point of attachment (e.g., due to a mobility event), it configures one or more new locators. If the Client has not moved far away from its previous network point of attachment, it simply informs its connected Server and any Client neighbors of any locator changes. This operation is performance sensitive and should be conducted immediately to avoid packet loss. This aspect of mobility can be classified as a "localized mobility event".

If the Client has moved far away from its previous network point of attachment, however, it re-issues the Server discovery procedure described in Section 6.3. If the Client's current Server is no longer close by, the Client may wish to move to a new Server in order to reduce routing stretch. This operation is not performance critical, and therefore can be conducted over a matter of minutes/ seconds instead of milliseconds/microseconds. This aspect of mobility can be classified as a "global mobility event".

To move to a new Server, the Client first engages in the CP registration process with the new Server, as described in Section 6.3. The Client then informs its former Server that it has departed; again, via a VSP-specific secured reliable transport connection. The former Server will then withdraw its CP advertisements from the IRON instance routing system and retain the (stale) forwarding table entries until their lifetime expires. In the interim, the former Server continues to deliver packets to the Client's last-known locator addresses for the short term while informing any unidirectional tunnel neighbors that the Client has moved.

Note that the Client may be either a mobile host or a mobile router. In the case of a mobile router, the Client's EUN becomes a mobile network, and can continue to use the Client's CPs without renumbering even as it moves between different network attachment points.

9.2. Multiple Interfaces and Multihoming

A Client may register multiple ISP connections with each Server such that multiple interfaces are naturally supported. This feature results in the Client "harnessing" its multiple ISP connections into a "bundle" that is represented as a single entity at the network layer, and therefore allows for ISP independence at the link-layer.

A Client may further register with multiple Servers for fault tolerance and reduced routing stretch. In that case, the Client should register its full bundle of ISP connections with each of its Servers unless it has a reason for carefully coordinating its

individual ISP-to-Server mappings.

Client registration with multiple Servers results in "pseudo-multihoming", in which the multiple homes are within the same VSP IRON instance and hence share fate with the health of the IRON instance itself.

9.3. Traffic Engineering

A Client can dynamically adjust its ISP-to-Server mappings in order to influence inbound traffic flows. It can also change between Servers when multiple Servers are available, but should strive for stability in its Server selection in order to limit VSP network routing churn.

A Client can select outgoing ISPs, e.g., based on current Quality-of-Service (QoS) considerations such as minimizing delay or variance.

10. Renumbering Considerations

As new link-layer technologies and/or service models emerge, end users will be motivated to select their basic Internetwork connectivity solutions through healthy competition between ISPs. If an end user's network-layer addresses are tied to a specific ISP, however, they may be forced to undergo a painstaking renumbering even if they wish to change to a different ISP [RFC4192][RFC5887].

When an end user Client obtains CPs from a VSP, it can change between ISPs seamlessly and without need to renumber the CPs. IRON therefore provides ISP independence at the link layer. If the end user is later compelled to change to a different VSP, however, it would be obliged to abandon its CPs and obtain new ones from the new VSP. In that case, the Client would again be required to engage in a painstaking renumbering event.

In order to avoid any future renumbering headaches, a Client that is part of a cooperative collective (e.g., a large enterprise network) could join together with the collective to obtain a suitably large PI prefix then and hire a VSP to manage the prefix on behalf of the collective. If the collective later decides to switch to a new VSP, it simply revokes its PI prefix registration with the old VSP and activates its registration with the new VSP.

11. NAT Traversal Considerations

The Internet today consists of a global public IPv4 routing and

addressing system with non-IRON EUNs that use either public or private IPv4 addressing. The latter class of EUNs connect to the public Internet via Network Address Translators (NATs). When an IRON Client is located behind a NAT, it selects Servers using the same procedures as for Clients with public addresses and can then send SRS messages to Servers in order to get SRA messages in return. The only requirement is that the Client must configure its encapsulation format to use a transport protocol that supports NAT traversal, e.g., UDP, TCP, etc.

Since the Server maintains state about its dependent Clients, it can discover locator information for each Client by examining the transport port number and IP address in the outer headers of the Client's encapsulated packets. When there is a NAT in the path, the transport port number and IP address in each encapsulated packet will correspond to state in the NAT box and might not correspond to the actual values assigned to the Client. The Server can then encapsulate packets destined to hosts in the Client's EUN within outer headers that use this IP address and transport port number. The NAT box will receive the packets, translate the values in the outer headers, then forward the packets to the Client. In this sense, the Server's "locator" for the Client consists of the concatenation of the IP address and transport port number.

In order to keep NAT and Server connection state alive, the Client sends periodic beacons to the server, e.g., by sending an SRS message to elicit an SRA message from the Server. IRON does not otherwise introduce any new complications for NAT traversal or for applications embedding address referrals in their payload.

12. Multicast Considerations

IRON Servers and Relays are topologically positioned to provide Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD) proxying for their Clients [RFC4605]. Further multicast considerations for IRON (e.g., interactions with multicast routing protocols, traffic scaling, etc.) are out of scope and will be discussed in a future document.

13. Nested EUN Considerations

Each Client configures a locator that may be taken from an ordinary non-CPA address assigned by an ISP or from a CPA address taken from a CP assigned to another Client. In that case, the Client is said to be "nested" within the EUN of another Client, and recursive nestings of multiple layers of encapsulations may be necessary.

For example, in the network scenario depicted in Figure 13, Client(A) configures a locator CPA(B) taken from the CP assigned to EUN(B). Client(B) in turn configures a locator CPA(C) taken from the CP assigned to EUN(C). Finally, Client(C) configures a locator ISP(D) taken from a non-CPA address delegated by an ordinary ISP(D).

Using this example, the "nested-IRON" case must be examined in which a Host A, which configures the address CPA(A) within EUN(A), exchanges packets with Host Z located elsewhere in a different IRON instance EUN(Z).

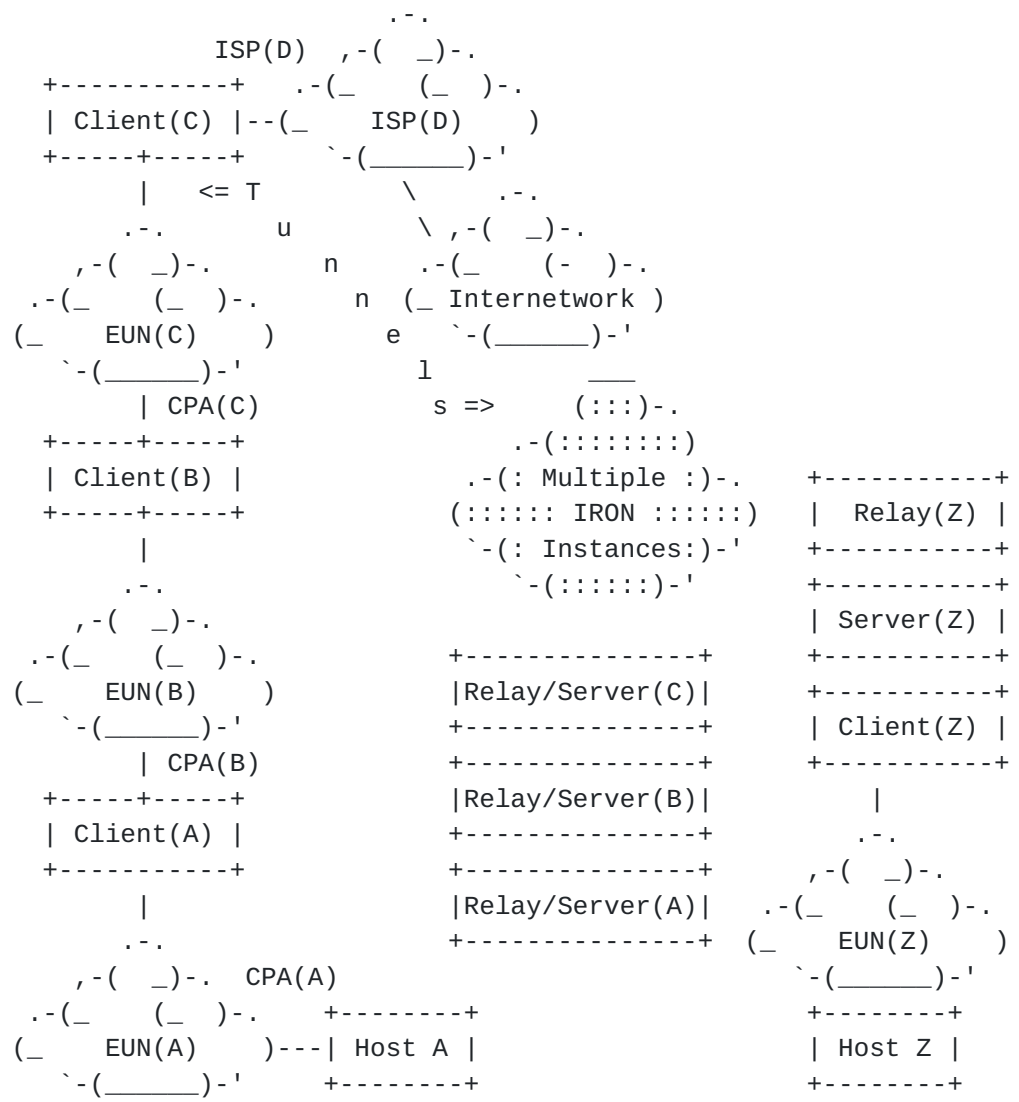


Figure 13: Nested EUN Example

The two cases of Host A sending packets to Host Z, and Host Z sending packets to Host A, must be considered separately, as described below.

13.1. Host A Sends Packets to Host Z

Host A first forwards a packet with source address CPA(A) and destination address Z into EUN(A). Routing within EUN(A) will direct the packet to Client(A), which encapsulates it in an outer header with CPA(B) as the outer source address and Server(A) as the outer destination address then forwards the once-encapsulated packet into EUN(B).

Routing within EUN(B) will direct the packet to Client(B), which encapsulates it in an outer header with CPA(C) as the outer source address and Server(B) as the outer destination address then forwards the twice-encapsulated packet into EUN(C). Routing within EUN(C) will direct the packet to Client(C), which encapsulates it in an outer header with ISP(D) as the outer source address and Server(C) as the outer destination address. Client(C) then sends this triple-encapsulated packet into the ISP(D) network, where it will be routed via the Internetwork to Server(C).

When Server(C) receives the triple-encapsulated packet, it forwards it to Relay(C) which removes the outer layer of encapsulation and forwards the resulting twice-encapsulated packet into the Internetwork to Server(B). Next, Server(B) forwards the packet to Relay(B) which removes the outer layer of encapsulation and forwards the resulting once-encapsulated packet into the Internetwork to Server(A). Next, Server(A) forwards the packet to Relay(A), which decapsulates it and forwards the resulting inner packet via the Internetwork to Relay(Z). Relay(Z), in turn, forwards the packet to Server(Z), which encapsulates and forwards the packet to Client(Z), which decapsulates it and forwards the inner packet to Host Z.

13.2. Host Z Sends Packets to Host A

When Host Z sends a packet to Host A, forwarding in EUN(Z) will direct it to Client(Z), which encapsulates and forwards the packet to Server(Z). Server(Z) will forward the packet to Relay(Z), which will then decapsulate and forward the inner packet into the Internetwork. Interdomain will convey the packet to Relay(A) as the next-hop towards CPA(A), which then forwards it to Server(A).

Server (A) encapsulates the packet and forwards it to Relay(B) as the next-hop towards CPA(B) (i.e., the locator for CPA(A)). Relay(B) then forwards the packet to Server(B), which encapsulates it a second time and forwards it to Relay(C) as the next-hop towards CPA(C) (i.e., the locator for CPA(B)). Relay(C) then forwards the packet to Server(C), which encapsulates it a third time and forwards it to Client(C).

Client(C) then decapsulates the packet and forwards the resulting twice-encapsulated packet via EUN(C) to Client(B). Client(B) in turn decapsulates the packet and forwards the resulting once-encapsulated packet via EUN(B) to Client(A). Client(A) finally decapsulates and forwards the inner packet to Host A.

14. Implications for the Internet

For IRON instances configured over the public Internet as the underlying Internetwork, the IRON system requires a VSP deployment of new routers/servers throughout the Internet to maintain well-balanced virtual overlay networks. These routers/servers can be deployed incrementally without disruption to existing Internet infrastructure as long as they are appropriately managed to provide acceptable service levels to end users.

End-to-end traffic that traverses an IRON instance may experience delay variance between the initial packets and subsequent packets of a flow. This is due to the IRON system allowing a longer path stretch for initial packets followed by timely route optimizations to utilize better next hop routers/servers for subsequent packets.

IRON instances work seamlessly with existing and emerging services within the native Internet. In particular, end users serviced by an IRON instance will receive the same service enjoyed by end users serviced by non-IRON service providers. Internet services already deployed within the native Internet also need not make any changes to accommodate IRON end users.

The IRON system operates between IAs within the Internet and EUNs. Within these networks, the underlying paths traversed by the virtual overlay networks may comprise links that accommodate varying MTUs. While the IRON system imposes an additional per-packet overhead that may cause the size of packets to become slightly larger than the underlying path can accommodate, IAs have a method for naturally detecting and tuning out instances of path MTU underruns. In some cases, these MTU underruns may need to be reported back to the original hosts; however, the system will also allow for MTUs much larger than those typically available in current Internet paths to be discovered and utilized as more links with larger MTUs are deployed.

Finally, and perhaps most importantly, the IRON system provides in-built mobility management, mobile networks, multihoming and traffic engineering capabilities that allow end user devices and networks to move about freely while both imparting minimal oscillations in the routing system and maintaining generally shortest-path routes. This mobility management is afforded through the very nature of the IRON

service model, and therefore requires no adjunct mechanisms. The mobility management and multihoming capabilities are further supported by forward-path reachability detection that provides "hints of forward progress" in the same spirit as for IPv6 Neighbor Discovery (ND).

15. Additional Considerations

Considerations for the scalability of interdomain routing due to multihoming, traffic engineering, and provider-independent addressing are discussed in [RADIR] [I-D.narten-radir-problem-statement]. Other scaling considerations specific to IRON are discussed in Appendix B.

Route optimization considerations for mobile networks are found in [RFC5522].

In order to ensure acceptable end user service levels, the VSP should conduct a capacity analysis and distribute sufficient Relays and Servers for the IRON instance globally throughout the Internet. As for common practices in the Internet today, such capacity analysis can be conducted in parallel with actual deployment of the service.

16. Related Initiatives

IRON builds upon the concepts of the RANGER architecture [RFC5720] , and therefore inherits the same set of related initiatives. The Internet Research Task Force (IRTF) Routing Research Group (RRG) mentions IRON in its recommendation for a routing architecture [RFC6115].

Virtual Aggregation (VA) [GROW-VA] and Aggregation in Increasing Scopes (AIS) [EVOLUTION] provide the basis for the Virtual Prefix concepts.

Internet Vastly Improved Plumbing (Ivip) [IVIP-ARCH] has contributed valuable insights, including the use of real-time mapping. The use of Servers as mobility anchor points is directly influenced by Ivip's associated TTR mobility extensions [TTRMOB].

[RO-CR][I-D.bernardos-mext-nemo-ro-cr] discusses a route optimization approach using a Correspondent Router (CR) model. The IRON Server construct is similar to the CR concept described in this work; however, the manner in which Clients coordinate with Servers is different and based on the NBMA virtual link model [RFC5214].

Numerous publications have proposed NAT traversal techniques. The

NAT traversal techniques adapted for IRON were inspired by the Simple Address Mapping for Premises Legacy Equipment (SAMPLE) proposal [SAMPLE][I-D.carpenter-softwire-sample].

The IRON Client-Server relationship is managed in essentially the same way as for the Tunnel Broker model [RFC3053]. Numerous existing provider networks that provide service similar to tunnel broker (e.g., Hurricane Electric, SixXS, freenet6, etc.) provide existence proofs that IRON-like overlay network services can be deployed and managed on a global basis [BROKER].

IRON is further related to the Identifier-Locator Network Protocol (ILNP) [RFC6740] and Locator / ID Split Protocol (LISP) [RFC6830] proposals which address routing scaling aspects at the interdomain level. IRON is therefore complimentary to these approaches.

17. IANA Considerations

There are no IANA considerations for this document.

18. Security Considerations

Security considerations that apply to tunneling in general are discussed in [RFC6169]. Additional considerations that apply also to IRON are discussed in RANGER [RFC5720][RFC6139] , VET [INTAREA-VET] and SEAL [INTAREA-SEAL].

The IRON system further depends on mutual authentication of IRON Clients to Servers and Servers to Relays. As for all Internet communications, the IRON system also depends on Relays acting with integrity and not injecting false advertisements into the interdomain routing system (e.g., to mount traffic siphoning attacks).

IRON Agents must perform message origin authentication on the packets they accept from correspondents. IAs must therefore include a signature on each packet that the destination can use to verify that the IA is authorized to use the source address.

IRON Servers must ensure that any changes in a Client's locator addresses are communicated only through an authenticated exchange that is not subject to replay. For this reason, Clients periodically send digitally-signed SRS messages to the Server. If the Client's locator address stays the same, the Server can accept the SRS message without verifying the signature. If the Client's locator address changes, the Server must verify the SRS message's signature before accepting the message. Once the message has been authenticated, the

Server updates the Client's locator address to the new address.

Each IRON instance requires a means for assuring the integrity of the interior routing system so that all Relays and Servers in the overlay have a consistent view of CP<->Server bindings. Also, Denial-of-Service (DoS) attacks on IRON Relays and Servers can occur when packets with spoofed source addresses arrive at high data rates. However, this issue is no different than for any border router in the public Internet today.

Middleboxes can interfere with tunneled packets within an IRON instance in various ways. For example, a middlebox may alter a packet's contents, change a packet's locator addresses, inject spurious packets, replay old packets, etc. These issues are no different than for middlebox interactions with ordinary Internet communications. If man-in-the-middle attacks are a matter for concern in certain deployments, however, IRON Agents can use IPsec [RFC4301] or TLS/SSL [RFC5246] to protect the authenticity, integrity and (if necessary) privacy of their tunneled packets.

19. Acknowledgements

The ideas behind this work have benefited greatly from discussions with colleagues; some of which appear on the RRG and other IRTF/IETF mailing lists. Robin Whittle and Steve Russert co-authored the TTR mobility architecture, which strongly influenced IRON. Eric Fleischman pointed out the opportunity to leverage anycast for discovering topologically close Servers. Thomas Henderson recommended a quantitative analysis of scaling properties.

The following individuals provided essential review input: Jari Arkko, Mohamed Boucadair, Stewart Bryant, John Buford, Ralph Droms, Wesley Eddy, Adrian Farrel, Dae Young Kim, and Robin Whittle.

Discussions with colleagues following the publication of RFC6179 have provided useful insights that have resulted in significant improvements to this, the Second Edition of IRON.

This document received substantial review input from the IESG and IETF area directorates in the February 2013 timeframe. IESG members and IETF area directorate representatives who contributed helpful comments and suggestions are gratefully acknowledged.

20. References

20.1. Normative References

[INTAREA-SEAL]
Templin, F., Ed., "The Subnetwork Encapsulation and Adaptation Layer (SEAL)", Work in Progress, February 2011.

[INTAREA-VET]
Templin, F., Ed., "Virtual Enterprise Traversal (VET)", Work in Progress, January 2011.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC6706] Templin, F., "Asymmetric Extended Route Optimization (AERO)", RFC 6706, August 2012.

20.2. Informative References

[BGPMON] net, B., "BGPmon.net - Monitoring Your Prefixes, http://bgpmon.net/stat.php", June 2010.

[BROKER] Wikipedia, W., "List of IPv6 Tunnel Brokers, http://en.wikipedia.org/wiki/List_of_IPv6_tunnel_brokers", August 2011.

[EVOLUTION]
Zhang, B., Zhang, L., and L. Wang, "Evolution Towards Global Routing Scalability", Work in Progress, October 2009.

[GROW-VA] Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "FIB Suppression with Virtual Aggregation", Work in Progress, February 2011.

[I-D.bernardos-mext-nemo-ro-cr]
Bernardos, C., Calderon, M., and I. Soto, "Correspondent Router based Route Optimisation for NEMO (CRON)", draft-bernardos-mext-nemo-ro-cr-00 (work in progress), July 2008.

[I-D.carpenter-softwire-sample]
Carpenter, B. and S. Jiang, "Legacy NAT Traversal for IPv6: Simple Address Mapping for Premises Legacy Equipment (SAMPLE)", draft-carpenter-softwire-sample-00 (work in progress), June 2010.

[I-D.narten-radir-problem-statement]
Narten, T., "On the Scalability of Internet Routing", draft-narten-radir-problem-statement-05 (work in progress), February 2010.

[IVIP-ARCH]
Whittle, R., "Ivip (Internet Vastly Improved Plumbing) Architecture", Work in Progress, March 2010.

[RADIR] Narten, T., "On the Scalability of Internet Routing", Work in Progress, February 2010.

[RFC0994] International Organization for Standardization (ISO) and American National Standards Institute (ANSI), "Final text of DIS 8473, Protocol for Providing the Connectionless-mode Network Service", RFC 994, March 1986.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC1930] Hawkinson, J. and T. Bates, "Guidelines for creation, selection, and registration of an Autonomous System (AS)", BCP 6, RFC 1930, March 1996.

[RFC3053] Durand, A., Fasano, P., Guardini, I., and D. Lento, "IPv6 Tunnel Broker", RFC 3053, January 2001.

[RFC4192] Baker, F., Lear, E., and R. Droms, "Procedures for Renumbering an IPv6 Network without a Flag Day", RFC 4192, September 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4380] Huitema, C., "Teredo: Tunneling IPv6 over UDP through Network Address Translations (NATs)", RFC 4380, February 2006.

[RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

[RFC4984] Meyer, D., Zhang, L., and K. Fall, "Report from the IAB

Workshop on Routing and Addressing", RFC 4984, September 2007.

[RFC5214] Templin, F., Gleeson, T., and D. Thaler, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214, March 2008.

[RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.

[RFC5522] Eddy, W., Ivancic, W., and T. Davis, "Network Mobility Route Optimization Requirements for Operational Use in Aeronautics and Space Exploration Mobile Networks", RFC 5522, October 2009.

[RFC5720] Templin, F., "Routing and Addressing in Networks with Global Enterprise Recursion (RANGER)", RFC 5720, February 2010.

[RFC5743] Falk, A., "Definition of an Internet Research Task Force (IRTF) Document Stream", RFC 5743, December 2009.

[RFC5887] Carpenter, B., Atkinson, R., and H. Flinck, "Renumbering Still Needs Work", RFC 5887, May 2010.

[RFC6115] Li, T., "Recommendation for a Routing Architecture", RFC 6115, February 2011.

[RFC6139] Russert, S., Fleischman, E., and F. Templin, "Routing and Addressing in Networks with Global Enterprise Recursion (RANGER) Scenarios", RFC 6139, February 2011.

[RFC6169] Krishnan, S., Thaler, D., and J. Hoagland, "Security Concerns with IP Tunneling", RFC 6169, April 2011.

[RFC6740] Atkinson,, RJ., "Identifier-Locator Network Protocol (ILNP) Architectural Description", RFC 6740, November 2012.

[RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, January 2013.

[RO-CR] Bernardos, C., Calderon, M., and I. Soto, "Correspondent Router based Route Optimisation for NEMO (CRON)", Work in Progress, July 2008.

[SAMPLE] Carpenter, B. and S. Jiang, "Legacy NAT Traversal for

IPv6: Simple Address Mapping for Premises Legacy Equipment (SAMPLE)", Work in Progress, June 2010.

[TTRMOB] Whittle, R. and S. Russert, "TTR Mobility Extensions for Core-Edge Separation Solutions to the Internet's Routing Scaling Problem, http://www.firstpr.com.au/ip/ivip/TTR-Mobility.pdf", August 2008.

Appendix A. IRON Operation over Internetworks with Different Address Families

The IRON architecture leverages the routing system by providing generally shortest-path routing for packets with CPA addresses from APs that match the address family of the underlying Internetwork. When the APs are of an address family that is not routable within the underlying Internetwork, however, (e.g., when OSI/NSAP [RFC0994] APs are used over an IPv4 Internetwork) a global Master AP mapping database (MAP) is required. The MAP allows the Relays of the local IRON instance to map APs belonging to other IRON instances to addresses taken from companion prefixes of address families that are routable within the Internetwork. For example, an IPv6 AP (e.g., 2001:DB8::/32) could be paired with one or more companion IPv4 prefixes (e.g., 192.0.2.0/24) so that encapsulated IPv6 packets can be forwarded over IPv4-only Internetworks. (In the limiting case, the companion prefixes could themselves be singleton addresses, e.g., 192.0.2.1/32).

The MAP is maintained by a globally managed authority, e.g. in the same manner as the Internet Assigned Numbers Authority (IANA) currently maintains the master list of all top-level IPv4 and IPv6 delegations. The MAP can be replicated across multiple servers for load balancing using common Internetworking server hierarchies, e.g., the DNS caching resolvers, ftp mirror servers, etc.

Upon startup, each Relay advertises IPv4 companion prefixes (e.g., 192.0.2.0/24) into the IPv4 Internetwork routing system and/or IPv6 companion prefixes (e.g., 2001:DB8::/64) into the IPv6 Internetwork routing system for the IRON instance that it serves. The Relay then selects singleton host numbers within the IPv4 companion prefixes (e.g., 192.0.2.1) and/or IPv6 companion prefixes (e.g., as 2001:DB8::0), and assigns the resulting addresses to its Internetwork interfaces. (When singleton companion prefixes are used (e.g., 192.0.2.1/32), the Relay does not advertise a the companion prefixes but instead simply assigns them to its Internetwork interfaces and allows standard Internet routing to direct packets to the interfaces.)

The Relay then discovers the APs for other IRON instances by reading the MAP, either a priori or on-demand of data packets addressed to other AP destinations. The Relay reads the MAP from a nearby MAP server and periodically checks the server for deltas since the database was last read. The Relay can then forward packets toward CPAs belonging to other IRON instances by encapsulating them in an outer header of the companion prefix address family and using the Relay anycast address as the outer destination address.

Possible encapsulations in this model include IPv6-in-IPv4, IPv4-in-IPv6, OSI/CLNP-in-IPv6, OSI/CLNP-in-IPv4, etc. Details of how the DNS can be used as a MAP are given in Section 5.4 of VET [INTAREA-VET].

Appendix B. Scaling Considerations

Scaling aspects of the IRON architecture have strong implications for its applicability in practical deployments. Scaling must be considered along multiple vectors, including interdomain core routing scaling, scaling to accommodate large numbers of EUNs, traffic scaling, state requirements, etc.

In terms of routing scaling, each VSP will advertise one or more APs into the interdomain routing system from which CPs are delegated to end users. Routing scaling will therefore be minimized when each AP covers many CPs. For example, the IPv6 prefix 2001:DB8::/32 contains 2^24 ::/56 CP prefixes for assignment to EUNs; therefore, the VSP could accommodate 2^32 ::/56 CPs with only 2^8 ::/32 APs advertised in the interdomain routing core. (When even longer CP prefixes are used, e.g., /64s assigned to individual handsets in a cellular provider network, many more EUNs can be represented within only a single AP.)

In terms of traffic scaling for Relays, each Relay represents an ASBR of a "shell" enterprise network that simply directs arriving traffic packets with CPA destination addresses towards Servers that service the corresponding Clients. Moreover, the Relay sheds traffic destined to CPAs through redirection, which removes it from the path for the majority of traffic packets between Clients within the same IRON instance. On the other hand, each Relay must handle all traffic packets forwarded between the CPs it manages and the rest of the Internetwork. The scaling concerns for this latter class of traffic are no different than for ASBR routers that connect large enterprise networks to the Internet. In terms of traffic scaling for Servers, each Server services a set of CPs. The Server services all traffic packets destined to its own CPs but only services the initial packets of flows initiated from its own CPs and destined to other CPs. Therefore, traffic scaling for CPA-addressed traffic is an asymmetric

consideration and is proportional to the number of CPs each Server serves. When possible, the Server can also be removed from the path in order to allow direct Client-to-Client communications as described in Section 8.1.3. In that case, the Server's burden in handling data packets is greatly reduced.

In terms of state requirements for Relays, each Relay maintains a list of Servers in the IRON instance as well as forwarding table entries for the CPs that each Server handles. This Relay state is therefore dominated by the total number of CPs handled by the Relay's associated Servers. Keeping in mind that current day core router technologies are only capable of handling fast-path FIB cache sizes of O(1M) entries, a large-scale deployment may require that the total CP database for the VSP overlay be spread between the FIBs of a mesh of Relays rather than fully-resident in the FIB of each Relay. In that case, the techniques of Virtual Aggregation (VA) may be useful in bridging together the mesh of Relays. Alternatively, each Relay could elect to keep some or all CP prefixes out of the FIB and maintain them only in a slow-path forwarding table. In that case, considerably more CP entries could be kept in each Relay at the cost of incurring slow-path processing for the initial packets of a flow.

In terms of state requirements for Servers, each Server maintains state only for the CPs it serves, and not for the CPs handled by other Servers in the IRON instance. Finally, neither Relays nor Servers need keep state for final destinations of outbound traffic.

Clients source and sink all traffic packets originating from or destined to the CP. Therefore, traffic scaling considerations for Clients are the same as for any site border router. Clients also retain tunnel neighbor state for final destinations of outbound traffic flows. This can be managed as soft state, since stale entries purged from the cache will be refreshed when new traffic packets are sent.

Author's Address

Fred L. Templin (editor)
Boeing Research & Technology
P.O. Box 3707 MC 7L-49
Seattle, WA 98124
USA

EMail: fltemplin@acm.org