The Internet Routing Overlay Network (IRON)

Abstract

   Since the Internet must continue to support escalating growth due to
   increasing demand, it is clear that current routing architectures and
   operational practices must be updated. This document proposes an
   Internet Routing Overlay Network (IRON) that supports sustainable
   growth while requiring no changes to end systems and no changes to
   the existing routing system. IRON further addresses other important
   issues including routing scaling, mobility management, multihoming,
   traffic engineering and NAT traversal. While business considerations
   are an important determining factor for widespread adoption, they are
   out of scope for this document. This document is a product of the
   IRTF Routing Research Group.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

   Growth in the number of entries instantiated in the Internet routing
   system has led to concerns regarding unsustainable routing scaling
   [RADIR]. Operational practices such as the increased use of
   multihoming with Provider-Independent (PI) addressing are resulting
   in more and more fine-grained prefixes being injected into the
   routing system from more and more end user networks. Furthermore,
   depletion of the public IPv4 address space has raised concerns for
   both increased address space fragmentation (leading to yet further
   routing table entries) 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 routing methodologies and operational practices are
   needed.

   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] and the Subnetwork Adaptation and Encapsulation Layer
   (SEAL) [INTAREA-SEAL] as its functional building blocks.

   This document proposes an Internet Routing Overlay Network (IRON)
   with goals of supporting sustainable growth while requiring no
   changes to the existing routing system. 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].

   IRON supports scalable addressing without changing the current BGP
   [RFC4271] routing system. IRON observes the Internet Protocol
   standards [RFC0791][RFC2460]. Other network-layer protocols that can
   be encapsulated within IP packets (e.g., OSI/CLNP (Connectionless
   Network Protocol) [RFC1070], etc.) are also within scope.

   The IRON is a global routing system comprising virtual overlay
   networks managed by Virtual Prefix Companies (VPCs) that own and
   manage Virtual Prefixes (VPs) from which End User Network (EUN)
   prefixes (EPs) are delegated to customer sites. The IRON is
   motivated by a growing customer demand for multihoming, mobility
   management, and traffic engineering while using stable addressing to
   minimize dependence on network renumbering [RFC4192][RFC5887]. The
   IRON uses the existing IPv4 and IPv6 global Internet routing systems
   as virtual NBMA links for tunneling inner network protocol packets
   within outer IPv4 or IPv6 headers (see Section 3). The IRON requires
   deployment of a small number of new BGP core routers and supporting
   servers, as well as IRON-aware routers/servers in customer EUNs. No
   modifications to hosts, and no modifications to most routers, are
   required.

   Note: This document is offered in compliance with Internet Research
   Task Force (IRTF) document stream procedures [RFC5743]; it is not an
   IETF product and is not a standard. The views in this document were
   considered controversial by the IRTF Routing Research Group (RRG),
   but the RG reached a consensus that the document should still be
   published. The document will undergo a period of review within the
   RRG and through selected expert reviewers prior to publication. The
   following sections discuss details of the IRON architecture.

2. Terminology

   This document makes use of the following terms:

   End User Network (EUN):
      an edge network that connects an organization's devices (e.g.,
      computers, routers, printers, etc.) to the Internet.

   End User Network Prefix (EP):
      a more specific inner network-layer prefix derived from a Virtual
      Prefix (VP) (e.g., an IPv4 /28, an IPv6 /56, etc.) and delegated
      to an EUN by a Virtual Prefix Company (VPC).

   End User Network Prefix Address (EPA):
      a network-layer address belonging to an EP and assigned to the
      interface of an end system in an EUN.

Forwarding Information Base (FIB):
   a data structure containing network prefixes to next-hop mappings;
   usually maintained in a router's fast-path processing lookup
   tables.

Internet Routing Overlay Network (IRON):
   a composite virtual overlay network that comprises the union of
   all VPC overlay networks configured over a common Internetwork.
   The IRON supports routing through encapsulation of inner packets
   with EPA addresses within outer headers that use locator
   addresses.

IRON Client Router/Host ("Client"):
   a customer's router or host that logically connects the customer's
   EUNs and their associated EPs to the IRON via an NBMA tunnel
   virtual interface.

IRON Serving Router ("Server"):
   a VPC's overlay network router that provides forwarding and
   mapping services for the EPs owned by customer Clients.

IRON Relay Router ("Relay"):
   a VPC's overlay network router that acts as a relay between the
   IRON and the native Internet.

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

Internet Service Provider (ISP):
   a service provider that connects customer EUNs to the underlying
   Internetwork. In other words, an ISP is responsible for providing
   basic Internet connectivity for customer EUNs.

Locator
   an IP address assigned to the interface of a router or end system
   within a public or private network. Locators taken from public IP
   prefixes are routable on a global basis, while locators taken from
   private IP prefixes 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 packet
    identification and a Control Message Protocol 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.

Virtual Prefix (VP):
    a prefix block (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 Prefix Company (VPC).

Virtual Prefix Company (VPC):
    a company that owns and manages a set of VPs from which it
    delegates EPs to EUNs.

VPC Overlay Network
    a specialized set of routers deployed by a VPC to service customer
    EUNs through a virtual overlay network configured over an
    underlying Internetwork (e.g., the global Internet).

3. The Internet Routing Overlay Network

The Internet Routing Overlay Network (IRON) is a system of virtual
overlay networks configured over a common Internetwork. While the
principles presented in this document are discussed within the
context of the public global Internet, they can also be applied to
any autonomous Internetwork. The rest of this document therefore
refers to the terms "Internet" and "Internetwork" interchangeably
except in cases where specific distinctions must be made.

The IRON consists of IRON Agents (IAs) that automatically tunnel the
packets of end-to-end communication sessions within encapsulating
headers used for Internet 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 headers, as shown in Figure 1.

```
                                  +------------------------+
                                  |  Outer headers with    |
                                  ~    locator addresses    ~
                                  |    (IPv4 or IPv6)       |
                                  +------------------------+
                                  |      SEAL Header        |
   +------------------------+     +------------------------+
   |  Inner Packet Header   | --> |  Inner Packet Header   |
   ~   with EP addresses    ~ --> ~   with EP addresses    ~
   | (IPv4, IPv6, OSI, etc.)| --> | (IPv4, IPv6, OSI, etc.)|
   +------------------------+     +------------------------+
   |                        | --> |                        |
   ~   Inner Packet Body    ~ --> ~   Inner Packet Body    ~
   |                        | --> |                        |
   +------------------------+     +------------------------+

        Inner packet before            Outer packet after
          encapsulation                  encapsulation
```
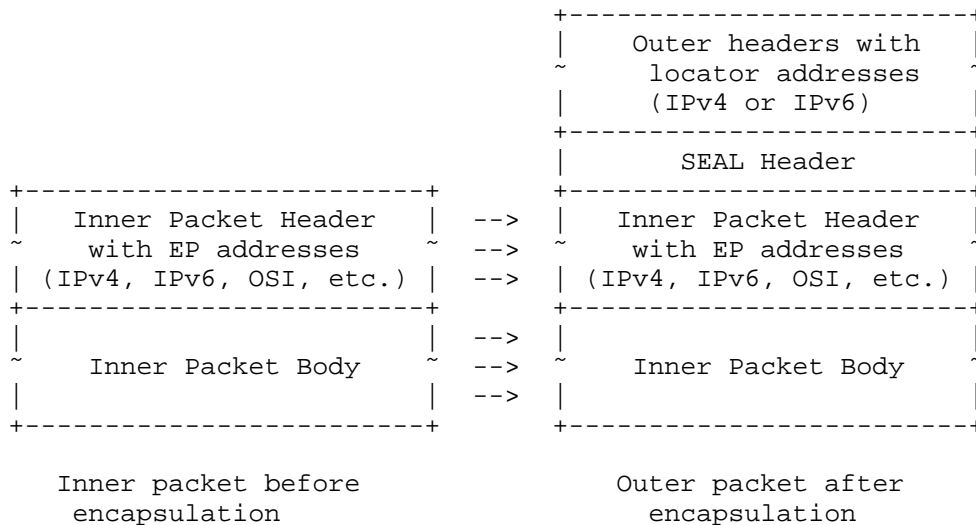
          Figure 1: Encapsulation of Inner Packets within Outer IP Headers

   VET specifies the automatic tunneling mechanisms used for
   encapsulation, while SEAL specifies the format and usage of the SEAL
   header as well as a set of control messages. Most notably, IAs use
   the SEAL Control Message Protocol (SCMP) to deterministically
   exchange and authenticate control messages such as route
   redirections, indications of Path Maximum Transmission Unit (PMTU)
   limitations, destination unreachables, etc. IAs appear as neighbors
   on an NBMA virtual link, and form bidirectional and/or unidirectional
   tunnel-neighbor relationships.

   The IRON is the union of all virtual overlay networks that are
   configured over a common underlying Internet and are owned and
   managed by Virtual Prefix Companies (VPCs). Each such virtual
   overlay network comprises a set of IAs distributed throughout the
   Internet to serve highly aggregated Virtual Prefixes (VPs). VPCs
   delegate sub-prefixes from their VPs, which they lease to customers
   as End User Network Prefixes (EPs). In turn, the customers assign
   the EPs to their customer edge IAs, which connect their End User
   Networks (EUNs) to the IRON.

   VPCs may have no affiliation with the ISP networks from which
   customers obtain their basic Internet connectivity. Therefore, a
   customer could procure its summary network services either through a
   common broker or through separate entities. In that case, the VPC
   can open for business and begin serving its customers immediately

   without the need to coordinate its activities with ISPs or other
   VPCs. Further details on business considerations are out of scope
   for this document.

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

3.1. IRON Client

   An IRON client (or, simply, "Client") is a customer's router or host
   that logically connects the customer's EUNs and their associated EPs
   to the IRON via tunnels, as shown in Figure 2. Client routers obtain
   EPs from VPCs and use them to number subnets and interfaces within
   their EUNs. A Client can be deployed on the same physical platform
   that also connects the customer's EUNs to its ISPs, but it may also
   be a separate router or even a standalone server system located
   within the EUN. (This model applies even if the EUN connects to the
   ISP via a Network Address Translator (NAT) -- see Section 6.7).
   Finally, a Client may also be a simple end system that connects a
   singleton EPA and exhibits the outward appearance of a host.

```
                             .-.
                          ,-(   _)-.
        +--------+      .-(_     (_  )-.
        | Client |--(_       ISP      )
        +---+----+       `-(_____)-'
            |   <= T          \       .-.
           .-.        u        \ ,-(   _)-.
         ,-(   _)-.      n       .-(_     (-  )-.
       .-(_     (_  )-.      n  (_    Internet   )
      (_      EUN      )      e   `-(_____)-
        `-(_____)-'          l            ___
            |                    s =>      (:::)-.
        +----+---+                       .-(:::::::::)
        | Host   |                     .-(:::::::::::::)-.
        +--------+                    (::::: The IRON ::::)
                                       `-(:::::::::::::)-'
                                          `-(:::::::)-'
```
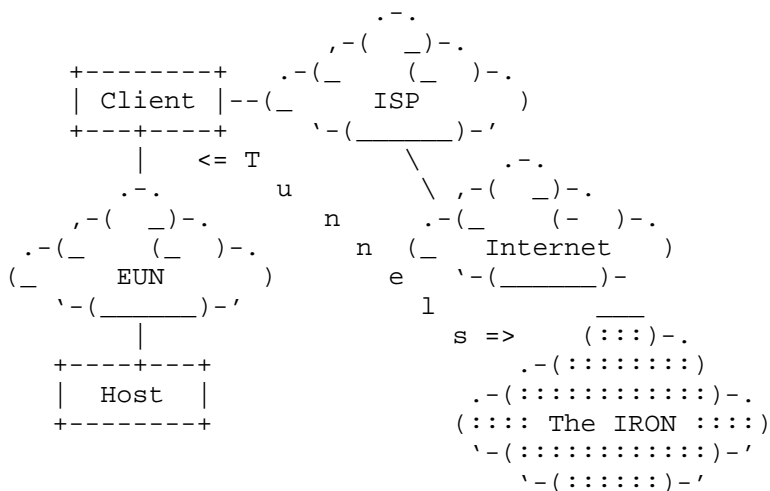
        Figure 2: IRON Client Router Connecting EUN to the IRON

3.2. IRON Serving Router

   An IRON serving router (or, simply, "Server") is a VPC's overlay
   network router that provides forwarding and mapping services for the
   EPs owned by customer Client routers. In typical deployments, a VPC
   will deploy many Servers around the IRON in a globally distributed
   fashion (e.g., as depicted in Figure 3) so that Clients can discover
   those that are nearby.

```
              +--------+    +--------+
              | Boston |    | Tokyo  |
              | Server |    | Server |
              +--+-----+    ++-------+
    +--------+   \         /
    | Seattle|    \  ___  /
    | Server |     \ (:::)-.        +--------+
    +------+-+  .-(:::::::::)------+ Paris  |
            \.-(:::::::::::::)-.   | Server |
            (::::: The IRON :::)   +--------+
             '-(:::::::::::::)-'
    +--------+ /  '-(:::::::)-'  \    +--------+
    | Moscow +        |          \--- + Sydney |
    | Server |    +----+---+         | Server |
    +--------+    | Cairo  |         +--------+
                  | Server |
                  +--------+
```
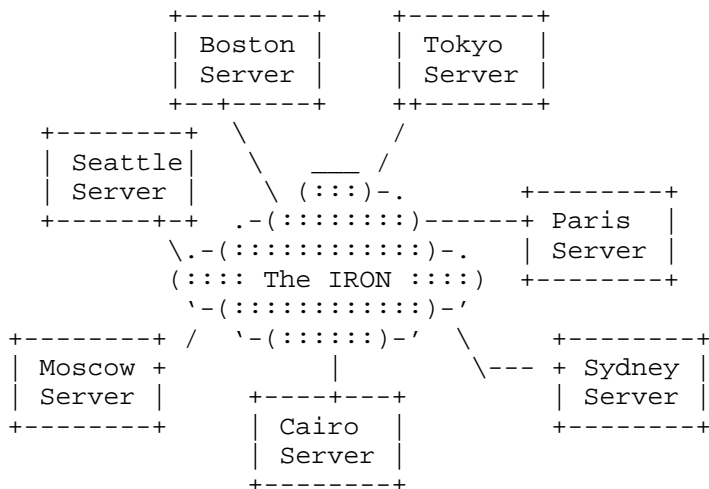
        Figure 3: IRON Serving Router Global Distribution Example

   Each Server acts as a tunnel-endpoint router that forms a
   bidirectional tunnel-neighbor relationship with each of its Client
   customers. Each Server also associates with a set of Relays that can
   forward packets from the IRON out to the native Internet and vice
   versa, as discussed in the next section.

3.3. IRON Relay Router

   An IRON Relay Router (or, simply, "Relay") is a VPC's overlay network
   router that acts as a relay between the IRON and the native Internet.
   Therefore, it also serves as an Autonomous System Border Router
   (ASBR) that is owned and managed by the VPC.

   Each VPC configures one or more Relays that advertise the company's
   VPs into the IPv4 and IPv6 global Internet BGP routing systems. Each
   Relay associates with all of the VPC's overlay network Servers, e.g.,
   via tunnels over the IRON, via a direct interconnect such as an
   Ethernet cable, etc. The Relay role (as well as its relationship
   with overlay network Servers) is depicted in Figure 4.

```
                      .-.
                    ,-(   _)-.
                  .-(_        (_  )-.
                 (_    Internet    )
                  `-(_____)-'   |  +--------+
                      |          |--| Server |
                  +----+---+     |  +--------+
                  | Relay  |----|  +--------+
                  +--------+     |--| Server |
                     _||         |  +--------+
                   (:::)-.   (Ethernet)
                 .-(::::::::)
   +--------+  .-(:::::::::::::)-.  +--------+
   | Server |=(::::: The IRON ::::)=| Server |
   +--------+  `-(:::::::::::::)-'  +--------+
                 `-(:::::::)-'
                     ||       (Tunnels)
                 +--------+
                 | Server |
                 +--------+
```
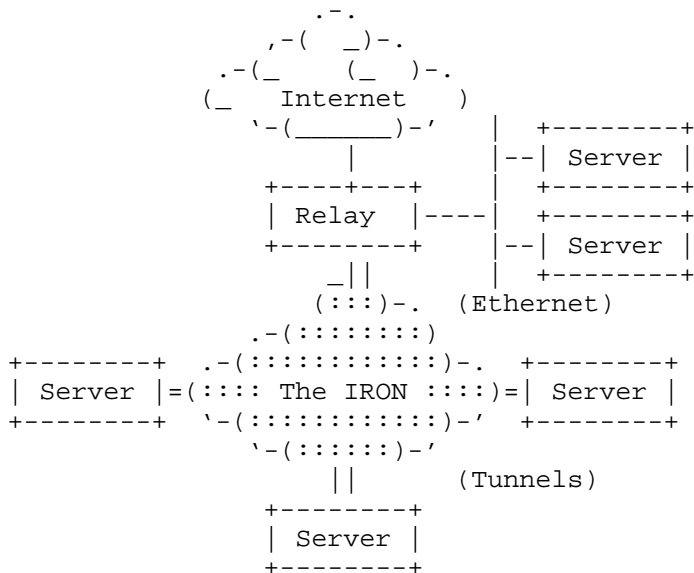
         Figure 4: IRON Relay Router Connecting IRON to Native Internet

4. IRON Organizational Principles

   The IRON consists of the union of all VPC overlay networks configured
   over a common Internetwork (e.g., the public Internet). Each such
   overlay network represents a distinct "patch" on the Internet
   "quilt", where the patches are stitched together by tunnels over the
   links, routers, bridges, etc. that connect the underlying
   Internetwork. When a new VPC overlay network is deployed, it becomes
   yet another patch on the quilt. The IRON is therefore a composite
   overlay network consisting of multiple individual patches, where each
   patch coordinates its activities independently of all others (with
   the exception that the Servers of each patch must be aware of all VPs
   in the IRON). In order to ensure mutual cooperation between all VPC
   overlay networks, sufficient address space portions of the inner
   network-layer protocol (e.g., IPv4, IPv6, etc.) should be set aside
   and designated as VP space.

   Each VPC overlay network in the IRON maintains a set of Relays and
   Servers that provide services to their Client customers. In order to
   ensure adequate customer service levels, the VPC should conduct a
   traffic scaling analysis and distribute sufficient Relays and Servers
   for the overlay network globally throughout the Internet. Figure 5
   depicts the logical arrangement of Relays, Servers, and Clients in an
   IRON virtual overlay network.

```
                            .-.
                         ,-(   _)-.
                       .-(_       (_  )-.
                      (__   Internet   _)
                        '-(_____)-'

              <------------    Relays      ------------>
                    _____
                    (:::::::::::::::::::::::)-.
                  .-(:::::::::::::::::::::::::::)
                .-(:::::::::::::::::::::::::::::::)-.
               (:::::::::::    The IRON  :::::::::::::)
                '-(:::::::::::::::::::::::::::::::)-'
                  '-(:::::::::::::::::::::::::)-'

              <------------    Servers     ------------>
             .-.               .-.                      .-.
           ,-(   _)-.        ,-(   _)-.               ,-(   _)-.
         .-(_       (_  )-.  .-(_       (_  )-.     .-(_       (_  )-.
        (__    ISP A    _)  (__    ISP B    _)  ... (__    ISP x    _)
          '-(_____)-'        '-(_____)-'           '-(_____)-'
             <----------     NATs       ------------>


              <---------- Clients and EUNs ---------->
```
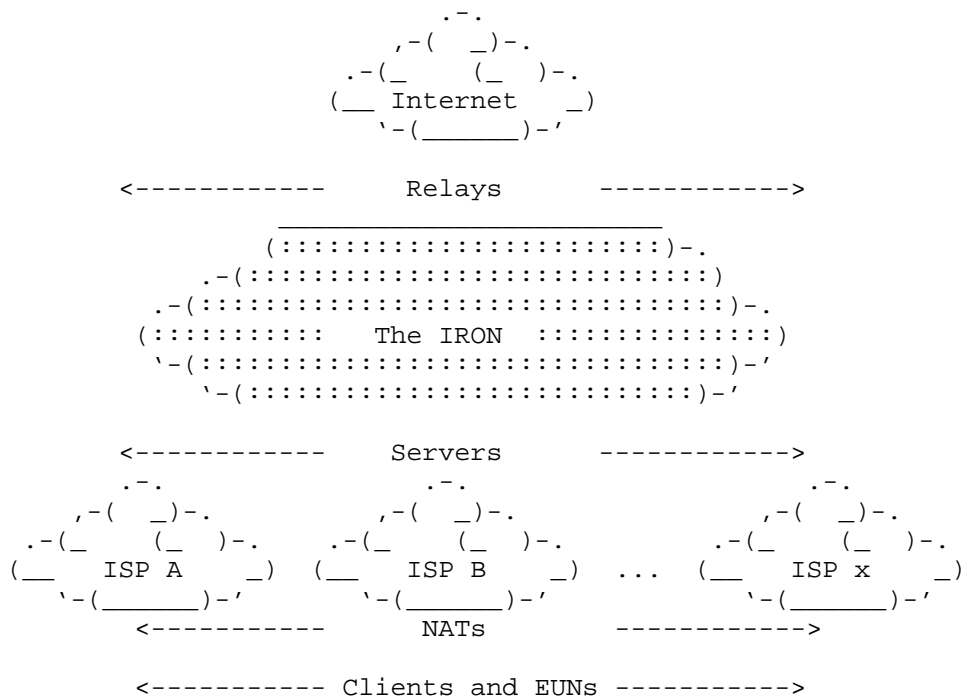
             Figure 5: Virtual Overlay Network Organization

Each Relay in the VPC overlay network connects the overlay directly
to the underlying IPv4 and IPv6 Internets. It also advertises the
VPC overlay network's IPv4 VPs into the IPv4 BGP routing system and
advertises the overlay network's IPv6 VPs into the IPv6 BGP routing
system. Relays will therefore receive packets with EPA destination
addresses sent by end systems in the Internet and direct them toward
EPA-addressed end systems connected to the VPC overlay network.

Each VPC overlay network also manages a set of Servers that connect
their Clients and associated EUNs to the IRON and to the IPv6 and
IPv4 Internets via their associations with Relays. IRON Servers
therefore need not be BGP routers themselves; they can be simple
commodity hardware platforms. Moreover, the Server and Relay
functions can be deployed together on the same physical platform as a
unified gateway, or they may be deployed on separate platforms (e.g.,
for load balancing purposes).

Each Server maintains a working set of Clients for which it caches
EP-to-Client mappings in its Forwarding Information Base (FIB). Each
Server also, in turn, propagates the list of EPs in its working set
to each of the Relays in the VPC overlay network via a dynamic

   routing protocol (e.g., an overlay network internal BGP instance that
   carries only the EP-to-Server mappings and does not interact with the
   external BGP routing system). Therefore, each Server only needs to
   track the EPs for its current working set of Clients, while each
   Relay will maintain a full EP-to-Server mapping table that represents
   reachability information for all EPs in the VPC overlay network.

   Customers establish Clients that obtain their basic Internet
   connectivity from ISPs and connect to Servers to attach their EUNs to
   the IRON. Each EUN can connect to the IRON via one or multiple
   Clients as long as the Clients coordinate with one another, e.g., to
   mitigate EUN partitions. Unlike Relays and Servers, Clients may use
   private addresses behind one or several layers of NATs. Each Client
   initially discovers a list of nearby Servers through an anycast
   discovery process (described below). It then selects one of these
   nearby Servers and forms a bidirectional tunnel-neighbor relationship
   with the server through an initial exchange followed by periodic
   keepalives.

   After the Client selects a Server, it forwards initial outbound
   packets from its EUNs by tunneling them to the Server, which, in
   turn, forwards them to the nearest Relay within the IRON that serves
   the final destination. The Client will subsequently receive redirect
   messages informing it of a more direct route through a Server that
   serves the final destination EUN.

   The IRON can also be used to support VPs 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 VPs of one address family
   over Internetworks based on other address families are discussed in
   Appendix A.

5. IRON Initialization

   IRON initialization entails the startup actions of IAs within the VPC
   overlay network and customer EUNs. The following sub-sections
   discuss these startup procedures.

5.1. IRON Relay Router Initialization

   Before its first operational use, each Relay in a VPC overlay network
   is provisioned with the list of VPs that it will serve as well as the
   locators for all Servers that belong to the same overlay network.
   The Relay is also provisioned with external BGP interconnections --
   the same as for any BGP router.

Upon startup, the Relay engages in BGP routing exchanges with its
peers in the IPv4 and IPv6 Internets the same as for any BGP router.
It then connects to all of the Servers in the overlay network (e.g.,
via a TCP connection over a bidirectional tunnel, via an Internal BGP
(IBGP) route reflector, etc.) for the purpose of discovering EP-to-
Server mappings. After the Relay has fully populated its EP-to-
Server mapping information database, it is said to be "synchronized"
with regard to its VPs.

After this initial synchronization procedure, the Relay then
advertises the overlay network's VPs externally. In particular, the
Relay advertises the IPv6 VPs into the IPv6 BGP routing system and
advertises the IPv4 VPs into the IPv4 BGP routing system. The Relay
additionally advertises an IPv4 /24 companion prefix (e.g.,
192.0.2.0/24) into the IPv4 routing system and an IPv6 ::/64
companion prefix (e.g., 2001:DB8::/64) into the IPv6 routing system
(note that these may also be sub-prefixes taken from a VP). The
Relay then configures the host number '1' in the IPv4 companion
prefix (e.g., as 192.0.2.1) and the interface identifier '0' in the
IPv6 companion prefix (e.g., as 2001:DB8::0), and it assigns the
resulting addresses as subnet-router anycast addresses
[RFC3068][RFC2526] for the VPC overlay network. (See Appendix A for
more information on the discovery and use of companion prefixes.)
The Relay then engages in ordinary packet-forwarding operations.

5.2. IRON Serving Router Initialization

Before its first operational use, each Server in a VPC overlay
network is provisioned with the locators for all Relays that
aggregate the overlay network's VPs. In order to support route
optimization, the Server must also be provisioned with the list of
all VPs in the IRON (i.e., not just the VPs of its own overlay
network) so that it can discern EPA and non-EPA addresses.
(Therefore, the Server could be greatly simplified if the list of VPs
could be covered within a small number of very short prefixes, e.g.,
one or a few IPv6 ::/20's). The Server must also discover the VP
companion prefix relationships discussed in Section 5.1, e.g., via a
global database such as discussed in Appendix A.

Upon startup, each Server must connect to all of the Relays within
its overlay network (e.g., via a TCP connection, via an IBGP route
reflector, etc.) for the purpose of reporting its EP-to-Server
mappings. The Server then actively listens for Client customers that
register their EP prefixes as part of establishing a bidirectional
tunnel-neighbor relationship. When a new Client registers its EP
prefixes, the Server announces the new EP additions to all Relays;
when an existing Client unregisters its EP prefixes, the Server
withdraws its announcements.

5.3. IRON Client Initialization

   Before its first operational use, each Client must obtain one or more
   EPs from its VPC as well as the companion prefixes associated with
   the VPC overlay network (see Section 5.1). The Client must also
   obtain a certificate and a public/private key pair from the VPC that
   it can later use to prove ownership of its EPs. This implies that
   each VPC must run its own public key infrastructure to be used only
   for the purpose of verifying its customers' claimed right to use an
   EP. Hence, the VPC need not coordinate its public key infrastructure
   with any other organization.

   Upon startup, the Client sends an SCMP Router Solicitation (SRS)
   message to the VPC overlay network subnet-router anycast address to
   discover the nearest Relay. The Relay will return an SCMP Router
   Advertisement (SRA) message that lists the locator addresses of one
   or more nearby Servers. (This list is analogous to the Intra-Site
   Automatic Tunnel Addressing Protocol (ISATAP) Potential Router List
   (PRL) [RFC5214].)

   After the Client receives an SRA message from the nearby Relay
   listing the locator addresses of nearby Servers, it initiates a short
   transaction with one of the Servers carried by a reliable transport
   protocol such as TCP in order to establish a bidirectional tunnel-
   neighbor relationship. The protocol details of the transaction are
   specific to the VPC, and hence out of scope for this document.

   Note that it is essential that the Client select one and only one
   Server. This is to allow the VPC overlay network mapping system to
   have one and only one active EP-to-Server mapping at any point in
   time, which shares fate with the Server itself. If this Server
   fails, the Client can select a new one that will automatically update
   the VPC overlay network mapping system with a new EP-to-Server
   mapping.

6. IRON Operation

   Following the IRON initialization detailed in Section 5, IAs engage
   in the steady-state process of receiving and forwarding packets. All
   IAs forward encapsulated packets over the IRON using the mechanisms
   of VET [INTAREA-VET] and SEAL [INTAREA-SEAL], while Relays (and in
   some cases Servers) additionally forward packets to and from the
   native IPv6 and IPv4 Internets. IAs also use SCMP to coordinate with
   other IAs, including the process of sending and receiving redirect
   messages, error messages, etc. (Note however that an IA must not
   send an SCMP message in response to an SCMP error message.) Each IA
   operates as specified in the following sub-sections.

6.1. IRON Client Operation

   After selecting its Server as specified in Section 5.3, the Client
   should register each of its ISP connections with the Server for
   multihoming purposes. To do so, it sends periodic beacons (e.g., SRS
   messages) to its Server via each of its ISPs to establish additional
   tunnel-neighbor state. This implies that a single tunnel-neighbor
   identifier (i.e., a "nonce") is used to represent the set of all ISP
   paths between the Client and the Server. Therefore, the nonce names
   this "bundle" of ISP paths.

   If the Client ceases to receive acknowledgements from its Server via
   a specific ISP connection, it marks the Server as unreachable from
   that address and therefore over that ISP connection. (The Client
   should also inform its Server of this outage via one of its working
   ISP connections.) If the Client ceases to receive acknowledgements
   from its Server via multiple ISP connections, it marks the Server as
   unusable and quickly attempts to register with a new Server. The act
   of registering with a new Server will automatically purge the stale
   mapping state associated with the old Server, since dynamic routing
   will propagate the new client/server relationship to the VPC overlay
   network Relay Routers.

   When an end system in an EUN sends a flow of packets to a
   correspondent, the packets are forwarded through the EUN via normal
   routing until they reach the Client, which then tunnels the initial
   packets to its Server as the next hop. In particular, the Client
   encapsulates each packet in an outer header with its locator as the
   source address and the locator of its Server as the destination
   address. Note that after sending the initial packets of a flow, the
   Client may receive important SCMP messages, such as indications of
   PMTU limitations, redirects that point to a better next hop, etc.

   The Client uses the mechanisms specified in VET and SEAL to
   encapsulate each forwarded packet. The Client further uses the SCMP
   protocol to coordinate with Servers, including accepting redirects
   and other SCMP messages. When the Client receives an SCMP message,
   it checks the nonce field of the encapsulated packet-in-error to
   verify that the message corresponds to the tunnel-neighbor state for
   its Server and accepts the message if the nonce matches. (Note
   however that the outer source and destination addresses of the
   packet-in-error may be different than those in the original packet
   due to possible Server and/or Relay address rewritings.)

6.2. IRON Serving Router Operation

   After the Server is initialized, it responds to SRSs from Clients by
   sending SRAs. When the Server receives a SEAL-encapsulated packet
   from one of its Client tunnel neighbors, it examines the inner
   destination address. If the inner destination address is not an EPA,
   the Server decapsulates the packet and forwards it unencapsulated
   into the Internet if it is able to do so without loss due to ingress
   filtering. Otherwise, the Server re-encapsulates the packet (i.e.,
   it removes the outer header and replaces it with a new outer header
   of the same address family) and sets the outer destination address to
   the locator address of a Relay within its VPC overlay network. It
   then forwards the re-encapsulated packet to the Relay, which will, in
   turn, decapsulate it and forward it into the Internet.

   If the inner destination address is an EPA, however, the Server
   rewrites the outer source address to one of its own locator addresses
   and rewrites the outer destination address to the subnet-router
   anycast address taken from the companion prefix associated with the
   inner destination address (where the companion prefix of the same
   address family as the outer IP protocol is used). The Server then
   forwards the revised encapsulated packet into the Internet via a
   default or more specific route, where it will be directed to the
   closest Relay within the destination VPC overlay network. After
   sending the packet, the Server may then receive an SCMP error or
   redirect message from a Relay/Server within the destination VPC
   overlay network. In that case, the Server verifies that the nonce in
   the message matches the Client that sent the original inner packet
   and discards the message if the nonce does not match. Otherwise, the
   Server re-encapsulates the SCMP message in a new outer header that
   uses the source address, destination address, and nonce parameters
   associated with the Client's tunnel-neighbor state; it then forwards
   the message to the Client. This arrangement is necessary to allow
   SCMP messages to flow through any NATs on the path.

   When a Server ('A') receives a SEAL-encapsulated packet from a Relay
   or from the Internet, if the inner destination address matches an EP
   in its FIB, 'A' re-encapsulates the packet in a new outer header and
   forwards it to a Client ('B'), which, in turn, decapsulates the
   packet and forwards it to the correct end system in the EUN.
   However, if 'B' has left notice with 'A' that it has moved to a new
   Server ('C'), 'A' will instead forward the packet to 'C' and also
   send an SCMP redirect message back to the source of the packet. In
   this way, 'B' can leave behind forwarding information when changing
   between Servers 'A' and 'C' (e.g., due to mobility events) without
   exposing packets to loss.

6.3. IRON Relay Router Operation

   After each Relay has synchronized its VPs (see Section 5.1) it
   advertises the full set of the company's VPs and companion prefixes
   into the IPv4 and IPv6 Internet BGP routing systems. These prefixes
   will be represented as ordinary routing information in the BGP, and
   any packets originating from the IPv4 or IPv6 Internet destined to an
   address covered by one of the prefixes will be forwarded to one of
   the VPC overlay network's Relays.

   When a Relay receives a packet from the Internet destined to an EPA
   covered by one of its VPs, it behaves as an ordinary IP router. In
   particular, the Relay looks in its FIB to discover a locator of the
   Server that serves the EP covering the destination address. The
   Relay then simply encapsulates the packet with its own locator as the
   outer source address and the locator of the Server as the outer
   destination address and forwards the packet to the Server.

   When a Relay receives a packet from the Internet destined to one of
   its subnet-router anycast addresses, it discards the packet if it is
   not SEAL encapsulated. If the packet is an SCMP SRS message, the
   Relay instead sends an SRA message back to the source listing the
   locator addresses of nearby Servers then discards the message. The
   Relay otherwise discards all other SCMP messages.

   If the packet is an ordinary SEAL packet (i.e., one that encapsulates
   an inner packet), the Relay sends an SCMP redirect message of the
   same address family back to the source with the locator of the Server
   that serves the EPA destination in the inner packet as the redirected
   target. The source and destination addresses of the SCMP redirect
   message use the outer destination and source addresses of the
   original packet, respectively. After sending the redirect message,
   the Relay then rewrites the outer destination address of the SEAL-
   encapsulated packet to the locator of the Server and forwards the
   revised packet to the Server. Note that in this arrangement, any
   errors that occur on the path between the Relay and the Server will
   be delivered to the original source but with a different destination
   address due to this Relay address rewriting.

6.4. IRON Reference Operating Scenarios

   The IRON supports communications when one or both hosts are located
   within EP-addressed EUNs, regardless of whether the EPs are
   provisioned by the same VPC or by different VPCs. When both hosts
   are within IRON EUNs, route redirections that eliminate unnecessary
   Servers and Relays from the path are possible. When only one host is
   within an IRON EUN, however, route optimization cannot be used. The
   following sections discuss the two scenarios.

6.4.1. Both Hosts within IRON EUNs

   When both hosts are within IRON EUNs, 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.

   In this scenario, the initial packets of a flow produced by a source
   host within an EUN connected to the IRON by a Client must flow
   through both the Server of the source host and a Relay of the
   destination host, but route optimization can eliminate these elements
   from the path for subsequent packets in the flow. Figure 6 shows the
   flow of initial packets from host A to host B within two IRON EUNs
   (the same scenario applies whether the two EUNs are within the same
   VPC overlay network or different overlay networks).

```
                        _____
               .-(                      .-.                       )-.
            .-(                      ,-(  _)-.                        )-.
         .-(            +========+(_       (_  +=====+                  )-.
       .(              ||     (_||  Internet ||_) ||                      ).
      .(               ||       ||-(_____)-||  vv                        ).
    .(       +--------++--+     ||           ||  +-----------+              ).
   (      +==>| Server(A)  |    vv           ||  | Server(B) |====+          )
   (     //   +---------|\-+   +--++----++--+    +-----------+    \\         )
   (    //  .-.         |  \   | Relay(B)  |                .-.   \\         )
   (   //,-(  _)-.      |   \  +-v---------+             ,-(  _)-\\          )
   ( .||_    (_   )-.   |    \____|                    .-(_     (_  ||. )    )
   ( _||  ISP A    .)   |                             (__   ISP B  ||_))
   (  ||-(_____)-'     |  (redirect)                   '-(_____)||  )      )
   (  ||    |           |                                   |   vv   )
    ( +-----+-----+     |                               +-----+-----+ )
      | Client(A) | <--+                                | Client(B) |
      +-----+-----+           The IRON                  +-----+-----+
            |     (   (Overlaid on the Native Internet)    )   |
          .-.       .-(                              .-)       .-.
        ,-(  _)-.       .-(_____)-.       ,-(  _)-.
      .-(_    (_  )-.                              .-(_    (_   )-.
     (_   IRON EUN A  )                           (_   IRON EUN B  )
        '-(_____)-'                                 '-(_____)-'
            |                                             |
       +---+----+                                    +---+----+
       | Host A |                                    | Host B |
       +--------+                                    +--------+
```
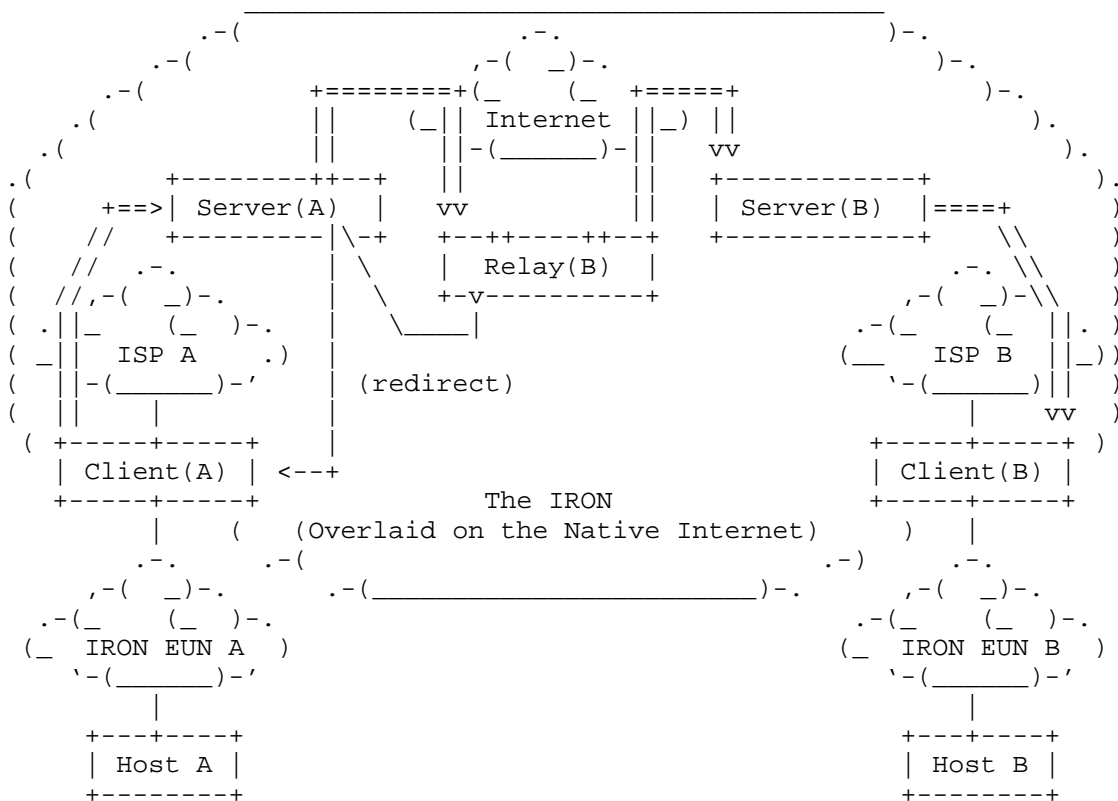
              Figure 6: Initial Packet Flow before Redirects

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 uses VET and SEAL to encapsulate them in outer headers
with its locator address as the outer source address and the locator
address of Server(A) as the outer destination address. Client(A)
then simply forwards the encapsulated packets into its ISP network
connection that provided its locator. The ISP will forward the
encapsulated packets into the Internet without filtering since the
(outer) source address is topologically correct. Once the packets
have been forwarded into the Internet, routing will direct them to
Server(A).

Server(A) receives the encapsulated packets from Client(A) then
rewrites the outer source address to one of its own locator addresses
and rewrites the outer destination address to the subnet-router
anycast address of the appropriate address family associated with the
inner destination address. Server(A) then forwards the revised
encapsulated packets into the Internet, where routing will direct
them to Relay(B), which services the VPC overlay network associated
with host B.

Relay(B) will intercept the encapsulated packets from Server(A) then
check its FIB to discover an entry that covers inner destination
address B with Server(B) as the next hop. Relay(B) then returns SCMP
redirect messages to Server(A) (*), rewrites the outer destination
address of the encapsulated packets to the locator address of
Server(B), and forwards these revised packets to Server(B).

Server(B) will receive the encapsulated packets from Relay(B) then
check its FIB to discover an entry that covers destination address B
with Client(B) as the next hop. Server(B) then re-encapsulates the
packets in a new outer header that uses the source address,
destination address, and nonce parameters associated with the tunnel-
neighbor state for Client(B). Server(B) then forwards these re-
encapsulated packets into the Internet, 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.

(*) Note that after the initial flow of packets, Server(A) will have
received one or more SCMP redirect messages from Relay(B) listing
Server(B) as a better next hop. Server(A) will, in turn, forward the
redirects to Client(A), which will establish unidirectional tunnel-
neighbor state and thereafter forward its encapsulated packets
directly to the locator address of Server(B) without involving either
Server(A) or Relay(B), as shown in Figure 7.

```
                 _____
             .-(                    .-.                     )-.
           .-(                    ,-(  _)-.                     )-.
          .-( +=============> .-(_     (_  )-.======+             )-.
        .(     //                (__ Internet _)     ||             ).
       .(     //                  '-(_____)-'       vv             ).
      .(      //                                +-----------+         ).
      (      //                                 | Server(B) |====+      )
      (      //                                 +-----------+    \\     )
      (     //   .-.                                        .-. \\      )
      (    //,-(  _)-.                                   ,-(  _)-\\     )
      ( .||_     (_  )-.                               .-(_     (_ ||. )
      ( _||  ISP A    .)                              (__    ISP B ||_))
      (  ||-(_____)-'                                   '-(_____)||   )
      (  ||     |                                             |    vv   )
      ( +-----+-----+              The IRON              +-----+-----+ )
        | Client(A) |    (Overlaid on the native Internet)  | Client(B) |
        +-----+-----+                                      +-----+-----+
              |      (                                  )    |
            .-.       .-(                        .-)       .-.
         ,-(  _)-.         .-(_____)-.    ,-(  _)-.
       .-(_     (_  )-.                                     .-(_     (_  )-.
      (_   IRON EUN A  )                                   (_   IRON EUN B  )
        '-(_____)-'                                         '-(_____)-'
              |                                                   |
        +---+----+                                          +---+----+
        | Host A |                                          | Host B |
        +--------+                                          +--------+
```
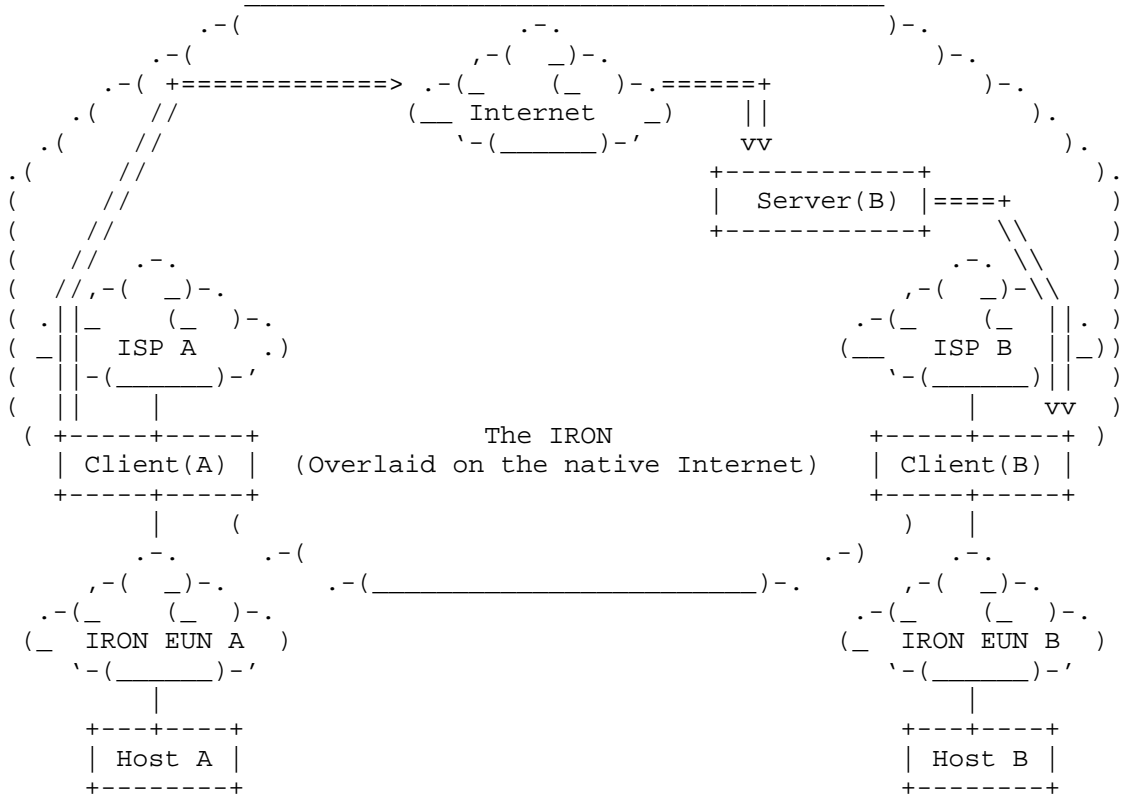
               Figure 7: Sustained Packet Flow after Redirects

6.4.2. Mixed IRON and Non-IRON Hosts

   When one host is within an IRON EUN and the other is in a non-IRON
   EUN (i.e., one that connects to the native Internet instead of the
   IRON), the IA elements involved depend on the packet-flow directions.
   The cases are described in the following sub-sections.

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

   Figure 8 depicts the IRON reference operating scenario for packets
   flowing from host A in an IRON EUN to host B in a non-IRON EUN.

```
                  _____
              .-(            )-.                            )-.
            .-(      +--------)----+                           )-.
          .-(        | Relay(A)  |-------------+                  )-.
         .(          +-----------+              \                   ).
        .(       +======>| Server(A) |           \                    ).
       .(      //        +--------)---+           \                     ).
      (      //                    )               \                      )
      (     //        The IRON     )                \                     )
      (    //   .-.                )                 \        .-.          )
      (   //,-(  _)-.              )                  \  ,-(  _)-.         )
      ( .||_     (_  )-.           ) The Native Internet  .-|_    (_  )-. )
      ( _||  ISP A     )           )                   (_  |   ISP B     ))
      (  ||-(_____)-'             )                      |-(_____)-'   )
      (  ||    |                 )-.                       v     |         )
       ( +-----+ ----+      )-.                     +-----+-----+        )
         | Client(A) |)-.                           |  Router B  |
         +-----+-----+                              +-----+-----+
             |  (                                        )  |
           .-.    .-(_____)-.    .-.
          ,-(  _)-.                                      ,-(  _)-.
        .-(_     (_  )-.                               .-(_     (_  )-.
       (_   IRON EUN A  )                             (_non-IRON EUN B)
         '-(_____)-'                                   '-(_____)-'
            |                                              |
        +---+----+                                     +---+----+
        | Host A |                                     | Host B |
        +--------+                                     +--------+
```
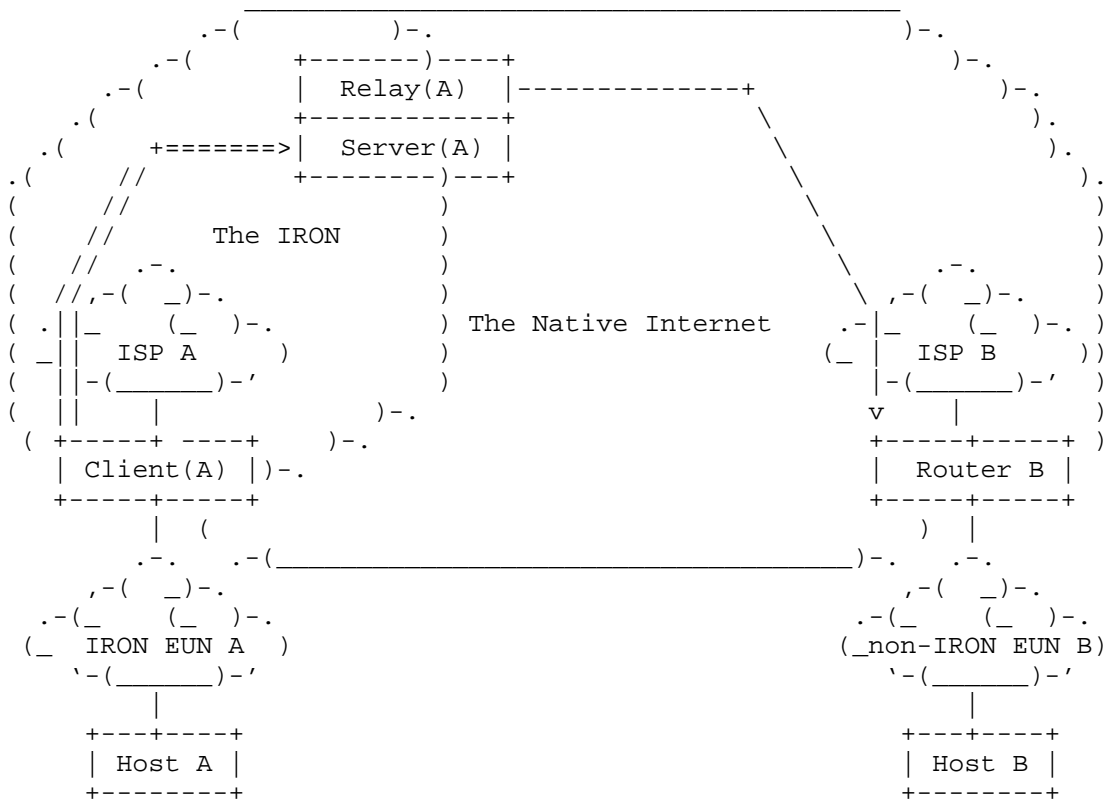
              Figure 8: From IRON Host A to Non-IRON Host B

   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 uses VET and SEAL to encapsulate them in outer headers
   with its locator address as the outer source address and the locator
   address of Server(A) as the outer destination address. The ISP will
   pass the packets without filtering since the (outer) source address
   is topologically correct. Once the packets have been released into
   the native Internet, routing will direct them to Server(A).

   Server(A) receives the encapsulated packets from Client(A) then re-
   encapsulates and forwards them to Relay(A), which simply decapsulates
   them and forwards the unencapsulated packets into the Internet. Once
   the packets are released into the Internet, routing will direct them
   to the final destination B. (Note that Server(A) and Relay(A) are

depicted in Figure 8 as two halves of a unified gateway. In that
case, the "forwarding" between Server(A) and Relay(A) is a zero-
instruction imaginary operation within the gateway.)

This scenario always involves a Server and Relay owned by the VPC
that provides service to IRON EUN A. Therefore, it imparts a cost
that would need to be borne by either the VPC or its customers.

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

Figure 9 depicts the IRON reference operating scenario for packets
flowing from host B in an Non-IRON EUN to host A in an IRON EUN.

```
                             _____
             .-(          )-.                               )-.
           .-(       +--------)----+                           )-.
         .-(         |  Relay(A)  |<-------------+              )-.
        .(           +------------+               \              ).
       .(       +=======|  Server(A) |             \              ).
      .(        //      +--------)---+              \              ).
     (         //                )                   \              )
     (        //       The IRON   )                   \             )
     (       //   .-.             )                    \    .-.      )
     (      //,-(   _)-.          )                     \ ,-(   _)-. )
     ( .||_     (_   )-.          ) The Native Internet  .-|_     (_   )-. )
     ( _||  ISP A      )          )                     (_ |  ISP B     ))
     (  ||-(_____)-'             )                      |-(_____)-'   )
     (  vv     |              )-.                         |    |         )
      ( +-----+ ----+     )-.                         +-----+-----+ )
        | Client(A) |)-.                              | Router B  |
        +-----+-----+                                 +-----+-----+
             |  (                                          ) |
           .-.    .-(_____)-.  .-.
          ,-(  _)-.                                  ,-(  _)-.
        .-(_     (_  )-.                           .-(_     (_  )-.
       (_   IRON EUN A  )                         (_non-IRON EUN B)
         '-(_____)-'                               '-(_____)-'
             |                                           |
        +---+----+                                  +---+----+
        | Host A |                                  | Host B |
        +--------+                                  +--------+
```
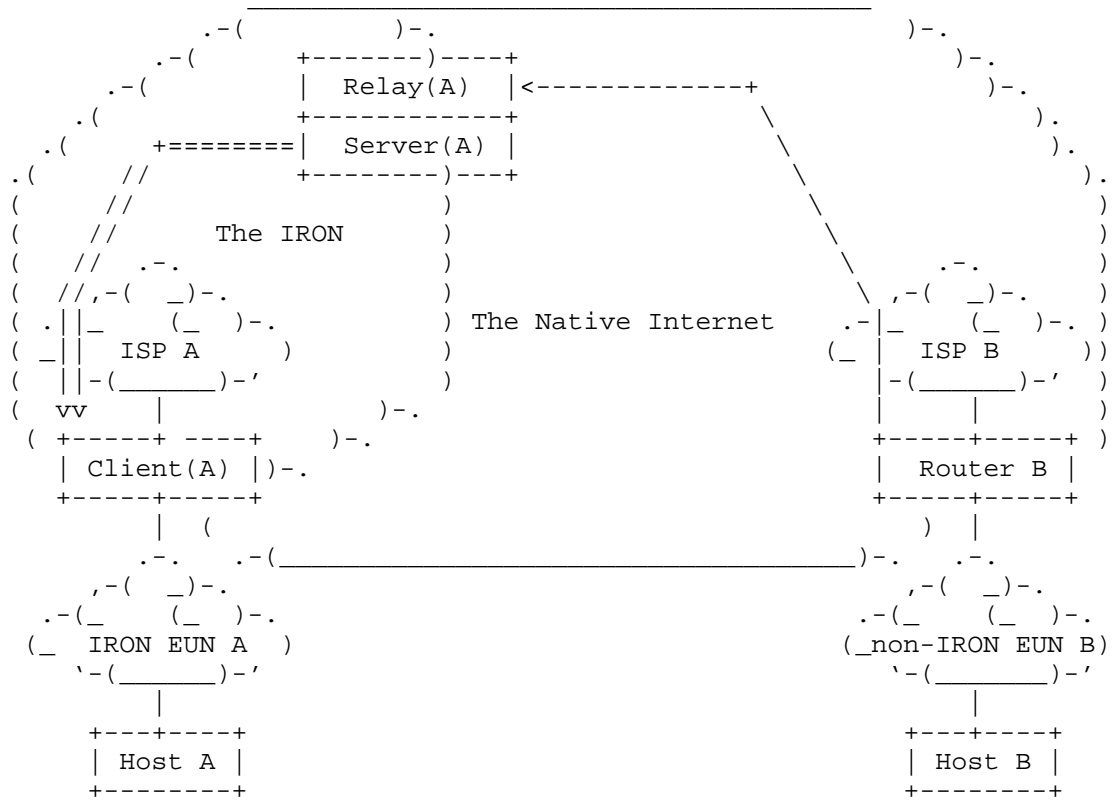
                Figure 9: From Non-IRON Host B to IRON Host A

In this scenario, host B sends packets destined to host A via its
network interface connected to non-IRON EUN B. Routing will direct
the packets to Relay(A), which then forwards them to Server(A) using
encapsulation if necessary.

Server(A) will then check its FIB to discover an entry that covers
destination address A with Client(A) as the next hop. Server(A) then
(re-)encapsulates the packets in an outer header that uses the source
address, destination address, and nonce parameters associated with
the tunnel-neighbor state for Client(A). Next, Server(A) forwards
these (re-)encapsulated packets into the Internet, 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.

This scenario always involves a Server and Relay owned by the VPC
that provides service to IRON EUN A. Therefore, it imparts a cost
that would need to be borne by either the VPC or its customers.

6.5. Mobility, Multihoming, and Traffic Engineering Considerations

   While IRON Servers and Relays can be considered 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 client
   routers.

6.5.1. Mobility Management

   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 Server of any locator additions or
   deletions. This operation is performance sensitive and should be
   conducted immediately to avoid packet loss.

   If the Client has moved far away from its previous network point of
   attachment, however, it re-issues the anycast discovery procedure
   described in Section 6.1 to discover whether its candidate set of
   Servers has changed. If the Client's current Server is also included
   in the new list received from the VPC, this provides indication that
   the Client has not moved far enough to warrant changing to a new
   Server. Otherwise, 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 seconds/
   minutes instead of milliseconds/microseconds.

   To move to a new Server, the Client first engages in the EP
   registration process with the new Server, as described in Section
   5.3. The Client then informs its former Server that it has moved by

providing it with the locator address of the new Server; again, via a
VPC-specific reliable transaction. The former Server will then
garbage-collect the stale FIB entries when their lifetime expires.

This will allow the former Server to redirect existing correspondents
to the new Server so that no packets are lost.

6.5.2. Multihoming

A Client may register multiple locators with its Server. It can
assign metrics with its registrations to inform the Server of
preferred locators, and it can select outgoing locators according to
its local preferences. Therefore, multihoming is naturally
supported.

6.5.3. Inbound Traffic Engineering

A Client can dynamically adjust the priorities of its prefix
registrations with its Server 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 VPC network routing churn.

6.5.4. Outbound Traffic Engineering

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

6.6. Renumbering Considerations

As new link-layer technologies and/or service models emerge,
customers will be motivated to select their service providers through
healthy competition between ISPs. If a customer's EUN addresses are
tied to a specific ISP, however, the customer may be forced to
undergo a painstaking EUN renumbering process if it wishes to change
to a different ISP [RFC4192][RFC5887].

When a customer obtains EP prefixes from a VPC, it can change between
ISPs seamlessly and without need to renumber. If the VPC itself
applies unreasonable costing structures for use of the EPs, however,
the customer may be compelled to seek a different VPC and would again
be required to confront a renumbering scenario. The IRON approach to
renumbering avoidance therefore depends on VPCs conducting ethical
business practices and offering reasonable rates.

6.7. 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 a
   Client is located behind a NAT, it selects Servers using the same
   procedures as for Clients with public addresses, e.g., it can send
   SRS messages to Servers in order to get SRA messages in return. The
   only requirement is that the Client must configure its SEAL
   encapsulation to use a transport protocol that supports NAT
   traversal, namely UDP.

   Since the Server maintains state about its Client customers, it can
   discover locator information for each Client by examining the UDP
   port number and IP address in the outer headers of the Client's
   encapsulated packets. When there is a NAT in the path, the UDP 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 UDP 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 UDP port
   number.

   IRON does not introduce any new issues to complications raised for
   NAT traversal or for applications embedding address referrals in
   their payload.

6.8. 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.) will be discussed in a
   separate document.

6.9. Nested EUN Considerations

   Each Client configures a locator that may be taken from an ordinary
   non-EPA address assigned by an ISP or from an EPA address taken from
   an EP 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 10, Client(A)
configures a locator EPA(B) taken from the EP assigned to EUN(B).
Client(B) in turn configures a locator EPA(C) taken from the EP
assigned to EUN(C). Finally, Client(C) configures a locator ISP(D)
taken from a non-EPA 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 EPA(A) within EUN(A), exchanges
packets with host Z located elsewhere in the Internet.

```
                        .-.
            ISP(D)   ,-(   _)-.
    +----------+    .-(_      (_   )-.
    | Client(C) |--(_      ISP(D)     )
    +-----+-----+    `-(_____)-'
          |    <= T      \        .-.
        .-.       u       \ ,-(   _)-.
      ,-(   _)-.       n     .-(_      (-   )-.
    .-(_      (_  )-.     n  (_     Internet    )
   (_     EUN(C)    )      e   `-(_____)-'
     `-(_____)-'         l        ___
          | EPA(C)         s =>     (:::)-.
    +-----+-----+                 .-(:::::::::)
    | Client(B) |              .-(:::::::::::::)-.  +----------+
    +-----+-----+             (::::: The IRON :::::) | Relay(Z) |
          |                    `-(:::::::::::::)-'  +----------+
        .-.                      `-(:::::::)-'       +----------+
      ,-(   _)-.                                     | Server(Z) |
    .-(_      (_  )-.          +----------+          +----------+
   (_     EUN(B)    )          | Server(C) |           +----------+
     `-(_____)-'            +----------+            | Client(Z) |
          | EPA(B)             +----------+           +----------+
    +-----+-----+              | Server(B) |            +--------+
    | Client(A) |             +----------+             | Host Z |
    +----------+               +----------+            +--------+
          |                    | Server(A) |
        .-.                    +----------+
      ,-(   _)-.  EPA(A)
    .-(_      (_  )-.    +--------+
   (_     EUN(A)    )---| Host A |
     `-(_____)-'      +--------+
```
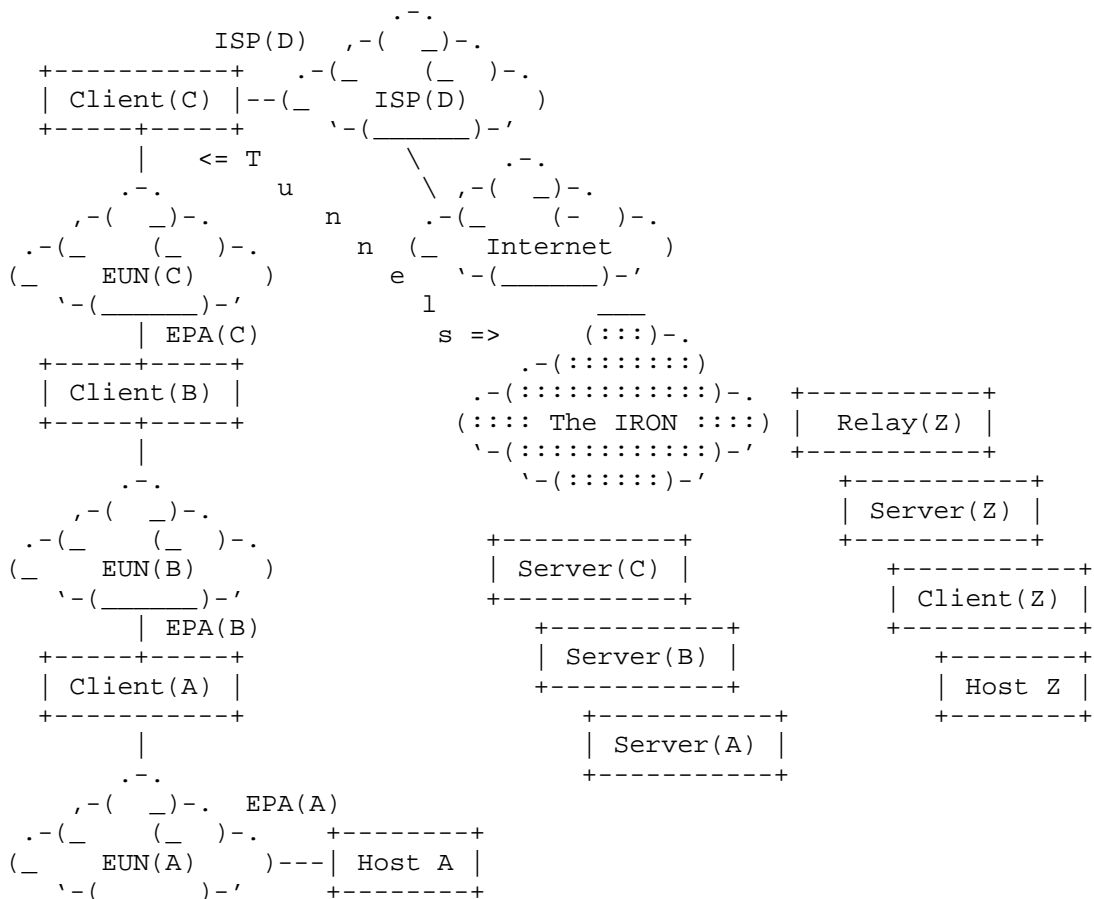
                   Figure 10: Nested EUN Example

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.

6.9.1. Host A Sends Packets to Host Z

   Host A first forwards a packet with source address EPA(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 EPA(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 EPA(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
   into the Internet to Server(C).

   When Server(C) receives the triple-encapsulated packet, it removes
   the outer layer of encapsulation and forwards the resulting twice-
   encapsulated packet into the Internet to Server(B). Next, Server(B)
   removes the outer layer of encapsulation and forwards the resulting
   once-encapsulated packet into the Internet to Server(A). Next,
   Server(A) checks the address type of the inner address 'Z'. If Z is
   a non-EPA address, Server(A) simply decapsulates the packet and
   forwards it into the Internet. Otherwise, Server(A) rewrites the
   outer source and destination addresses of the once-encapsulated
   packet and forwards it to Relay(Z). Relay(Z), in turn, rewrites the
   outer destination address of the packet to the locator for Server(Z),
   then forwards the packet and sends a redirect to Server(A) (which
   forwards the redirect to Client(A)). Server(Z) then re-encapsulates
   the packet and forwards it to Client(Z), which decapsulates it and
   forwards the inner packet to host Z. Subsequent packets from
   Client(A) will then use Server(Z) as the next hop toward host Z,
   which eliminates Server(A) and Relay(Z) from the path.

6.9.2. Host Z Sends Packets to Host A

   Whether or not host Z configures an EPA address, its packets destined
   to host A will eventually reach Server(A). Server(A) will have a
   mapping that lists Client(A) as the next hop toward EPA(A).
   Server(A) will then encapsulate the packet with EPA(B) as the outer
   destination address and forward the packet into the Internet.
   Internet routing will convey this once-encapsulated packet to
   Server(B), which will have a mapping that lists Client(B) as the next
   hop toward EPA(B). Server(B) will then encapsulate the packet with
   EPA(C) as the outer destination address and forward the packet into
   the Internet. Internet routing will then convey this twice-
   encapsulated packet to Server(C), which will have a mapping that

lists Client(C) as the next hop toward EPA(C). Server(C) will then
encapsulate the packet with ISP(D) as the outer destination address
and forward the packet into the Internet. Internet routing will then
convey this triple-encapsulated packet to Client(C).

When the triple-encapsulated packet arrives at Client(C), it strips
the outer layer of encapsulation and forwards the twice-encapsulated
packet to EPA(C), which is the locator address of Client(B). When
Client(B) receives the twice-encapsulated packet, it strips the outer
layer of encapsulation and forwards the once-encapsulated packet to
EPA(B), which is the locator address of Client(A). When Client(A)
receives the once-encapsulated packet, it strips the outer layer of
encapsulation and forwards the unencapsulated packet to EPA(A), which
is the host address of host A.

7. Implications for the Internet

   The IRON architecture envisions a hybrid routing/mapping system that
   benefits from both the shortest-path routing afforded by pure dynamic
   routing systems and the routing-scaling suppression afforded by pure
   mapping systems. Therefore, IRON targets the elusive "sweet spot"
   that pure routing and pure mapping systems alone cannot satisfy.

   The IRON system requires a deployment of new routers/servers
   throughout the Internet and/or provider networks to maintain well-
   balanced virtual overlay networks. These routers/servers can be
   deployed incrementally without disruption to existing Internet
   infrastructure and appropriately managed to provide acceptable
   service levels to customers.

   End-to-end traffic that traverses an IRON virtual overlay network 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 virtual overlay networks also work seamlessly with existing and
   emerging services within the native Internet. In particular,
   customers serviced by IRON virtual overlay networks will receive the
   same service enjoyed by customers serviced by non-IRON service
   providers. Internet services already deployed within the native
   Internet also need not make any changes to accommodate IRON virtual
   overlay network customers.

   The IRON system operates between routers within provider networks and
   end user networks. 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, IRON
routers have a method for naturally detecting and tuning out all
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 an
in-built mobility management and multihoming capability that allows
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 customer/provider relationship,
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).

8. Additional Considerations

   Considerations for the scalability of Internet Routing due to
   multihoming, traffic engineering, and provider-independent addressing
   are discussed in [RADIR]. Other scaling considerations specific to
   IRON are discussed in Appendix B.

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

9. Related Initiatives

   IRON builds upon the concepts of the RANGER architecture [RFC5720]
   [RFC6139], 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] 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 customer
EUNs coordinate with Servers is different and based on the
redirection model associated with NBMA links.

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].

10. Security Considerations

Security considerations that apply to tunneling in general are
discussed in [V6OPS-TUN-SEC]. 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. This is accomplished
through initial authentication exchanges followed by tunnel-neighbor
nonces that can be used to detect off-path attacks. As for all
Internet communications, the IRON system also depends on Relays
acting with integrity and not injecting false advertisements into the
BGP (e.g., to mount traffic siphoning attacks).

Each VPC overlay network 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 Client<->Server bindings. Finally,
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.

11. 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.

12. References

12.1. Normative References

   [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.

12.2. Informative References

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

   [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.

   [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.

   [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.

   [RFC1070] Hagens, R., Hall, N., and M. Rose, "Use of the Internet as
              a subnetwork for experimentation with the OSI network
              layer", RFC 1070, February 1989.

   [RFC2526] Johnson, D. and S. Deering, "Reserved IPv6 Subnet Anycast
              Addresses", RFC 2526, March 1999.

   [RFC3068] Huitema, C., "An Anycast Prefix for 6to4 Relay Routers",
              RFC 3068, June 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.

   [RFC4548] Gray, E., Rutemiller, J., and G. Swallow, "Internet Code
              Point (ICP) Assignments for NSAP Addresses", RFC 4548,
              May 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.

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

   [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.

   [V6OPS-TUN-SEC]
               Krishnan, S., Thaler, D., and J. Hoagland, "Security
               Concerns With IP Tunneling", Work in Progress,
               October 2010.

Appendix A. IRON VPs over Internetworks with Different Address Families

   The IRON architecture leverages the routing system by providing
   generally shortest-path routing for packets with EPA addresses from
   VPs that match the address family of the underlying Internetwork.
   When the VPs are of an address family that is not routable within the
   underlying Internetwork, however, (e.g., when OSI/NSAP [RFC4548] VPs
   are used within an IPv4 Internetwork) a global mapping database is
   required to allow Servers to map VPs to companion prefixes taken from
   address families that are routable within the Internetwork. For
   example, an IPv6 VP (e.g., 2001:DB8::/32) could be paired with a
   companion IPv4 prefix (e.g., 192.0.2.0/24) so that encapsulated IPv6
   packets can be forwarded over IPv4-only Internetworks.

   Every VP in the IRON must therefore be represented in a globally
   distributed Master VP database (MVPd) that maintains VP-to-companion
   prefix mappings for all VPs in the IRON. The MVPd is maintained by a
   globally managed assigned numbers authority 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 database
   can be replicated across multiple servers for load balancing, much in
   the same way that FTP mirror sites are used to manage software
   distributions.

   Upon startup, each Server discovers the full set of VPs for the IRON
   by reading the MVPd. The Server reads the MVPd from a nearby server
   and periodically checks the server for deltas since the database was
   last read. After reading the MVPd, the Server has a full list of VP-
   to-companion prefix mappings.

   The Server can then forward packets toward EPAs covered by a VP by
   encapsulating them in an outer header of the VP's companion prefix
   address family and using any address taken from the companion prefix
   as the outer destination address. The companion prefix therefore
   serves as an anycast prefix.

   Possible encapsulations in this model include IPv6-in-IPv4, IPv4-in-
   IPv6, OSI/CLNP-in-IPv6, OSI/CLNP-in-IPv4, etc.

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 customer EUNs,
   traffic scaling, state requirements, etc.

   In terms of routing scaling, each VPC will advertise one or more VPs
   into the global Internet routing system from which EPs are delegated
   to customer EUNs. Routing scaling will therefore be minimized when
   each VP covers many EPs. For example, the IPv6 prefix 2001:DB8::/32
   contains 2^{24} ::/56 EP prefixes for assignment to EUNs; therefore,
   the IRON could accommodate 2^{32} ::/56 EPs with only 2^{8} ::/32 VPs
   advertised in the interdomain routing core. (When even longer EP
   prefixes are used, e.g., /64s assigned to individual handsets in a
   cellular provider network, considerable numbers of EUNs can be
   represented within only a single VP.) Each VP also has an associated
   anycast companion prefix; hence, there will be one anycast prefix
   advertised into the global routing system for each VP.

   In terms of traffic scaling for Relays, each Relay represents an ASBR
   of a "shell" enterprise network that simply directs arriving traffic
   packets with EPA destination addresses towards Servers that service
   customer EUNs. Moreover, the Relay sheds traffic destined to EPAs
   through redirection, which removes it from the path for the vast
   majority of traffic packets. On the other hand, each Relay must
   handle all traffic packets forwarded between its customer EUNs and
   the non-IRON Internet. 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 the VPC overlay network's
   customer EUNs. The Server services all traffic packets destined to
   its EUNs but only services the initial packets of flows initiated
   from the EUNs and destined to EPAs. Therefore, traffic scaling for
   EPA-addressed traffic is an asymmetric consideration and is
   proportional to the number of EUNs each Server serves.

   In terms of state requirements for Relays, each Relay maintains a
   list of all Servers in the VPC overlay network as well as FIB entries
   for all customer EUNs that each Server serves. This state is
   therefore dominated by the number of EUNs in the VPC overlay network.
   Sizing the Relay to accommodate state information for all EUNs is
   therefore required during VPC overlay network planning. In terms of
   state requirements for Servers, each Server maintains tunnel-neighbor
   state for each of the customer EUNs it serves, but it need not keep

state for all EUNs in the VPC overlay network. 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 customer EUN. Therefore, traffic scaling
considerations for Clients are the same as for any site border
router. Clients also retain state for the Servers 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