Curtis Villamizar
ANS
Tony Li
Juniper
February 22, 1999

IS-IS Optimized Multipath (ISIS-OMP)

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

IS–IS may form multiple equal cost paths between points. This is true of any link state protocol. In the absence of any explicit support to take advantage of this, a path may be chosen arbitrarily. Techniques have been utilized to divide traffic somewhat evenly among the available paths. These techniques have been referred to as Equal Cost Multipath (ECMP). An unequal division of traffic among the available paths is generally preferable. Routers generally have no knowledge of traffic loading on distant links and therefore have no basis to optimize the allocation of traffic.

Optimized Mulitpath is a extension to IS–IS, utilizing additional Type/Length/Value (TLV) tuples to distribute loading information. An algorithm to adjust forwarding, gradually enough to insure stability yet provide reasonably fast adjustment when needed, is provided in the related document OSPF–OMP [6]. The IS–IS encapsulation and minor differences in the dynamics of ISIS–OMP relative to OSPF–OMP are described here.

1 Overview

IS–IS is OSI link state routing protocol [2]. OSPF is a link state routing protocol defined within the IETF process [4]. IS–IS can also be used as an IP interior gateway protocol (IGP) much as OSPF is used [1].

Networks are often heavily loaded. Topologies often evolve to include multiple paths. Multiple paths may be initially designed to provide redundancy but also result from incremental addition of circuits to accommodate traffic growth. The redundant paths provide a potential to distribute traffic loading and reduce congestion. Optimized Mulitpath (OMP) provides a means for a link state protocol such as IS–IS or OSPF to make better use of this potential to distribute loading. Please refer to [6] for a more complete introduction to OMP.

2 Differences Between ISIS–OMP and OSPF–OMP

Both IS–IS and OSPF are link state protocols. The technique of equal cost multipath (ECMP) routing is formally defined in OSPF [4] but has also been applied to IS–IS implementations. Because these routing protocols are quite similar, application of OMP techniques is very similar.

Differences between OSPF and IS–IS that impact OMP are:

1. The encapsulation of IS–IS link state information is different so the encapsulation of link loading information will be different.
2. IS–IS link state information is flooded in no more than 256 LSPDU fragments which must be reflooded in their entirety when any TLV in the LSPDU fragment changes.
3. IS–IS Level 2 routing differs from OSPF inter–area routing.
4. IS–IS handling of external routes in Level 2 differs from OSPF handling of external routes.

2.1 Encapsulation

IS–IS packs link state information in Type/Length/Value tuples. The type and length are one byte each. The length is a unsigned byte providing the length in bytes of the Value portion of the tuple. The "extended IS reachability TLV" is defined by Li as type 22 [3]. Subtypes of this TLV include the following, however if there are any differences in a descendant to [3], consider that list to be authoritative.

```
IPv4 interface address           subtype 6, length 4
IPv4 neighbor address            subtype 8, length 4
Maximum (out) link bandwidth     subtype 9, length 4
Current (out) bandwidth usage    subtype 12, length 4
Maximum inbound bandwidth usage  subtype 15, length 4
Outbound fractional packet loss  subtype 17, length 4
```

```
Inbound fractional packet loss    subtype 18, length 4
Maximum inbound link bandwidth    subtype 19, length 4
```

The value part these TLV tuples are described below:

IPv4 interface address, subtype 6, length 4 This is the IPv4 address of the near end of the interface.

IPv4 neighbor address, subtype 8, length 4 This is the IPv4 address of the far end of a point to point interface.

Maximum (out) link bandwidth, subtype 9, length 4 This is an IEEE 754 format 32 bit floating point number representing the bandwidth of the interface in bytes per second. See subtype 19.

Current (out) bandwidth usage, subtype 12, length 4 This is an IEEE 754 format 32 bit floating point number representing the outbound link utilization of the interface in bytes per second.

Maximum inbound bandwidth usage, subtype 15, length 4 This is an IEEE 754 format 32 bit floating point number representing the inbound link utilization of the interface in bytes per second.

Outbound fractional packet loss, subtype 17, length 4 This is an IEEE 754 format 32 bit floating point number representing the ratio of lost outbound packets (generally lost due to queueing drop) to total outbound packets.

Inbound fractional packet loss, subtype 18, length 4 This is an IEEE 754 format 32 bit floating point number representing the ratio of lost inbound packets to total inbound packets.

Maximum inbound link bandwidth, subtype 19, length 4 This is an IEEE 754 format 32 bit floating point number representing the inbound bandwidth of the interface in bytes per second. If this subtype is not present the link is assumed to be symmetric and the value given in a type 9 subtype TLV is used as both the inbound and outbound bandwidth of the interface.

2.2 Flooding

OSPF link state information is flooded in individual Link State Attributes that may be independently flooded or packed for efficiency. IS–IS considers the entire link state database to be somewhat of an atomic entity expressed as the Link State Protocol Data Unit (LSPDU). The LSPDU can be flooded in up to 256 LSPDU fragments. When a fragment is reflooded with a change, all link information in that fragment must be reflooded. Link loading information which otherwise might not be flooded for some time will be updated. This is expected to have minimal impact on the dynamics of the OMP algorithm.

An algorithm for filtering raw counter values and determining when to flood is provided in [6] stated quite concisely in the appendices. The IS–IS algorithm differs only in that link loading may be flooded at times simply because the link resides in an LSPDU fragment that must be reflooded due to a loading change in another link.

2.3 Relaxing Best Path Criteria

The IS–IS best path metric criteria can be relaxed just as the best path criteria is relaxed in OSPF–OMP. The same restrictions apply. In no circumstance should traffic toward an exit point with equal external cost or toward an internal destination be sent to an neighboring router where the path from that router has the same internal cost or greater internal cost. This restriction is sufficient to avoid the possibility of routing loops.

This restriction can be stated more precisely. Consider the following example. The best path from Router A to destination D has IS–IS path metric M_{AD}. The best path from Router B to destination D has IS–IS path metric M_{BD}. The link metric from Router A to Router B is M_{AB}. The path A to B to D is not an equal cost path if:

$$M_{AB} + M_{BD} > M_{AD}$$

Router A can relax the best path metric criteria is the following holds true:

$$M_{AD} > M_{BD}$$

2.4 Level 2 Routing

An implied assumption of the IS–IS protocol specification is that the Level 2 area has greater capacity and is less congested than the Level 1 only areas. For this reason, Level 2 routers consider the external metric before the internal metric and Level 1 routers consider the shortest path to any Level 2 router. In practice the Level 2 area may not be less congested than Level 1 areas and these restrictions are relaxed.

If consistency with the IS–IS model were maintain, path loading information can be passed from Level 1 into Level 2, but not from Level 2 into Level 1. This information would supplement the "IP Internal Reachability Information" (type 128) TLV tuple. Because the type 128 tuple contains fixed sized entries, a new tuple would have to be defined whose key is the IP address and subnet mask. The tuple defined in Section 2.1 would be used for this purpose.

The Level 2 routers may balance load across destinations with equal Level 1 metric but may not route toward a higher Level 1 metric. To do so would cause a routing loop if any router in the path decided to prefer the lower Level 1 metric as is now specified by IS–IS.

2.5 Handling External Routes

In IS–IS, as specified only Level 2 routers may carry external routing information. Level 1 routers are expected to route only to destinations in their own area or to the nearest Level 2 router. The Level 2 routers have complete information and can route the packet toward the correct area or to a router advertising the external route. In practice this is rarely an issue and implementations allow this aspect of the specification to be violated.

Generally ISIS is not used to carry exterior routing information. Instead, a separate exterior gateway protocol is used, specifically BGP–4 [5]. When BGP–4 is used to carry external routes within an AS (this is referred to as Internal BGP or IBGP), IS–IS is used strictly as an Interior Gateway Protocol (IGP). IBGP learned routes contain a NEXT_HOP attribute which contains an address that is reachable via the IGP but often not directly connected. This is commonly referred to as a recursive route lookup. When the IS–IS learned IGP route is used, the load balancing used to reach the IGP destination specified by the BGP NEXT_HOP is applied to the traffic routed toward that intermediate destination.

Some relaxation of routing criteria is possible to improve load balancing without creating route loops. For IS–IS external routes the external metric is considered before the internal metrics. This rule cannot be relaxed. With BGP–4, the BGP LOCAL_PREF and other BGP attributes are considered first. This rule cannot be relaxed. Given otherwise equal external routes, the IGP cost to reach one or more equally preferred exit point is considered. In this case the IGP cost is the IS–IS path metric. The best IS–IS path criteria can be relaxed as described in Section 2.3.

The path loading tuple defined in Section 2.4 could be used to provide exterior loading information if exterior routes are injected into IS–IS, a strongly discouraged practice. Carrying path loading information in an exterior routing protocol such as BGP–4 is outside the scope of this document.

Acknowledgments

Numerous individual have provided valuable comments regarding this work. Dave Ward made a very substantial contribution by pointing out that the best path criteria could be relaxed. Dave Ward, John Scudder, and Daniel Awduche have provided particularly valuable review and comments.

References

[1] R.W. Callon. Use of osi is-is for routing in tcp/ip and dual environments. Technical Report RFC 1195, Internet Engineering Task Force, 1990. ftp://ftp.isi.edu/in-notes/rfc1195.txt.

[2] ISO/IEC. Iso/iec 10589 - information processing systems - telecommunications and information exchange between systems - intermediate system to intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode network service (iso 8473). Technical report, International Organization for Standardization, 1992. ftp://merit.edu/pub/iso/iso10589.ps.gz.

[3] Tony Li and Henk Smit. Is-is extensions for traffic engineering. Internet Draft (Work in Progress) draft-ietf-isis-traffic-00, Internet Engineering Task Force, 2 1999. ftp://ftp.isi.edu/internet-drafts/draft-ietf-isis-traffic-00.txt.

[4] J. Moy. Ospf version 2. Technical Report RFC 2328, Internet Engineering Task Force, 1998. ftp://ftp.isi.edu/in-notes/rfc2328.txt.

[5] Y. Rekhter and T. Li. A border gateway protocol 4 (bgp-4). Technical Report RFC 1771, Internet Engineering Task Force, 1995. ftp://ftp.isi.edu/in-notes/rfc1771.txt.

[6] Curtis Villamizar. Ospf optimized multipath (ospf-omp). Internet Draft (Work in Progress) draft-ietf-ospf-omp-01, Internet Engineering Task Force, 10 1998. ftp://ftp.isi.edu/internet-drafts/draft-ietf-ospf-omp-01.txt.

Security Considerations

In deployments which use a strong IS–IS authentication method, and require signatures on LSP fragments from the originating router, no leveraging of a partial compromise beyond a localized disruption of service is possible. In deployments which use a strong IS–IS authentication method, but do not require signatures on LSP fragments from the originating router, compromise of a single router can be leveraged to cause significant disruption of service with or without the use of these TLV tuples, but disruption of service cannot be achieved without such a compromise. In deployments which use a weak IS–IS authentication method, significant disruption of service can be caused through forged protocol interactions with or without the use of these TLV tuples.

Author's Addresses


```
Curtis Villamizar            Tony Li
ANS                          Juniper Networks
<curtis@ans.net>             <tli@juniper.net>
```


A Configuration Options

Many of the capabilities must be configurable. The ability to enable and disable capability subsets is needed. Many parameters used by the algorithm should also be configurable. Please refer to [6] for parameter settings. ISIS–OMP differs from OSPF–OMP in the capability subsets.

A.1 Capability Subsets

There should at least be the ability to enabled and disabled the following.

```
default description of capability
  ON   Flooding any loading information
  ON   Flooding loading information for specific links
  -    Relaxing best path criteria
  -    Adjusting traffic shares (default to even split)
  OFF  Flooding loading information into IS-IS Level 2
  OFF  Flooding loading information for specific IS-IS Level 1 routes
  OFF  Flooding loading information for external routes
  OFF  Flooding loading information for specific external routes
  OFF  Considering loading information for IS-IS Level 2
```

```
OFF  Considering loading information for specific IS-IS Level 1 routes
OFF  Considering loading information for external routes
OFF  Considering loading information for specific external routes
```

Flooding and considering Level 1 routes loading information into Level 2 and flooding external route loading information should be disabled by default. Flooding loading information should be enabled by default. Relaxing best path criteria and adjusting traffic shares could be enabled or disabled by default, depending on vendor preference.