Chair
IEEE US TAG
to
ISO/IEC JTC1
SC22/WG13
But I Have Recovered
Where it Started

Entering the Rabbit Hole (an example)
IIJ is Building a Second Medium Scale Data Center (MSDC) in Shiroi/Chiba Capacity of 6k Racks
OSPF OK to 500 Nodes

IS-IS good to 1,000

Limited Because They Repeatedly Flood Everything
Your Clos on IS-IS or OSPF
BGP Scales Because
It Signals
Only Changes

So BGP has become common in MSDCs
But What is the Decision Process?
Consult the Professor

Edsger W Dijkstra
1930–2002

Shortest Path First
BGP-SPF

The Path Calculation of IS-IS

With the Update Rate of BGP
BGP Only Signals Changes Does Not Repeatedly Flood Link State
Neighbor Distribution Route Reflection Outbound Policy

SPF

Inbound Policy Link State

AS-Path Length EGP vs IGP Arrival Order Non-deterministic MED IGP metric Tie Break

BGP-SPF

Removed!
But ... How Does BGP-SPF Learn Link State?
Link Discovery and Liveness

What do we really need?
Application
Presentation
Session
Transport
Network
Data Link
Physical

We Are Here
Trying to Discover
Two Kinds of Standards

Union - the accumulation of all the features anybody wanted

Intersection - only those things everybody absolutely had to have

Either Tony Hoare or Klaus Wirth - I can not find the quote <blush>
IETF asks the ITU

Q: So you add features until the "NO"s stop?
A: We don’t like to think of it that way
Must Haves

- Discover Nodes and Links
- Discover Link Encapsulations:
  - IPv4, IPv6, MPLS4/6
- Maintain Liveness
- Northbound API to BGP-SPF
Security
Security?

• Datacenter Ops seem not to think of security at this layer (or any!)

• Do we want to add Authentication and maybe Integrity?

• One of the things which are likely to drive the size over 1,500
Non-Features

• Routing Data, BGP-SPF does that

• Access to IGP Databases, This is discovery and liveness, not routing

• Just want the Link

• Transport, not our job
Desiderata

- Discovery & Liveness for BGP-SPF
- Simple but usable in Massively Scalable networks of >10,000 nodes
- May be useful for other applications
- Simple
- Extensible (e.g. authentication, cost)
- Simple
- No IPR
Why Simple?

We are here to produce easily understood, implementable, and securable standards, not build résumés.
Why Simple?

A high goal of software engineering is to remove the need for features. It's a vital part of designing for simplicity, even invisibility. -- Rob Pike
LLDP

• IEEE Protocol
• A Little Noisy
• Beacons, not KeepAlives
• Viable but potential Show-Stopper
IS-IS Discovery

- Flooding & Noisy
- Complex enough that BGP-LS was invented so normals could get the link state database
- IS-IS not commonly implemented on MSDC devices, so would need to profile and develop
Edge Control Protocol

- It is a transport & Controlled by IEEE
- A Reliable layer two transport, on top of LLC
- Has flow control, reliable, non-reorder, ... transport
- used for EVP and PD/CSP
- Reinventing TCP over 802.1
BGP Neighbor Autodiscovery

- IETF Unrealistic & Incomplete Protocol
- Very new
- Needs the peering address to get the peering address
- AS Based, can not use other idents
- Not really discovery at all, configuration
- No liveness
Link State Over Ether

- Custom made for the job
- Very bare bones, brutally simple
- Only does discovery and liveness
- New, therefore risky, still in flux
- But so is BGP-SPF
- No measurement or monitoring tools
<table>
<thead>
<tr>
<th></th>
<th>LLDP</th>
<th>IS-IS</th>
<th>ECP</th>
<th>BNA</th>
<th>LSOE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Who Owns</strong></td>
<td>IEEE</td>
<td>IETF</td>
<td>IEEE</td>
<td>IETF</td>
<td>IETF</td>
</tr>
<tr>
<td><strong>Maturity</strong></td>
<td>Mature</td>
<td>Mature</td>
<td>Recent</td>
<td>New</td>
<td>New</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Somewhat</td>
<td>Noisy</td>
<td>Rather</td>
<td>Somewhat</td>
<td>Almost too Simple</td>
</tr>
<tr>
<td><strong>Discovery</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Configure</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Liveness</strong></td>
<td>Beacons</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
So a New Protocol
East West Protocol
A PDU (Protocol Data Unit) is an application layer message.

- It may occupy multiple Datagrams.
- Datagrams are one per Ethernet Frame.
**Datagram**

A Datagram is one Ethernet Frame

A Datagram has Number, Length, and Checksum

The L flag is set on the last datagram of an application layer PDU

This Transport Layer assembles PDUs

---

2018.11.10 IETF/IEEE Discovery

Creative Commons: Attribution-NonCommercial-ShareAlike 2.0
### Every Datagram a TLV

<table>
<thead>
<tr>
<th>Type</th>
<th>PDU Length</th>
<th>Value ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - HELLO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - OPEN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - KEEPALIVE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 - ACK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 - IPv4 Announcement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 - IPv6 Announcement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 - MPLS IPv4 Announcement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 - MPLS IPv6 Announcement</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sessions are Pretty Clear

OPEN and Encapsulation

PDUs are ACKed
## Inter-Link Ether Protocol

<table>
<thead>
<tr>
<th>HELLO</th>
<th>Mandatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPEN</td>
<td>MACs, IDs</td>
</tr>
<tr>
<td>OPEN</td>
<td>Mandatory</td>
</tr>
<tr>
<td>Interface IPv4 Addresses</td>
<td>Interface IPv4 Addresses</td>
</tr>
<tr>
<td>Interface IPv6 Addresses</td>
<td>Interface IPv4 Addresses</td>
</tr>
<tr>
<td>Interface MPLSv4 Labels</td>
<td>MPLS IPv4 Addresses</td>
</tr>
<tr>
<td>Interface MPLSv6 Labels</td>
<td>MPLS IPv6 Addresses</td>
</tr>
<tr>
<td>Layer 2 KeepAlives</td>
<td>Layer 2 Liveness</td>
</tr>
</tbody>
</table>
• HELLO is Multicast, à la LLDP
• Each device learns the other's MAC from its HELLO whining. All devices on a wire/interface know each others MACs and learn each other’s IDs
• Respond with OPEN
• A multi-point topology is a set of point-to-point links
• Config to be link-only or piercing switches
Local/Remote IDs

Might be

- an ASN with high order bits zero
- a classic RouterID with high order bits zero
- a catenation of the two
- a 80-bit ISO System-ID
- or any other identifier unique to a single device in the current routing space
Attributes

A node may have zero or more user-defined attributes, e.g. spine, leaf, backbone, route reflector, arabica, ...

Nodes exchange their attributes only in the OPEN message
Authentication Data

• Specific to the Operational Environment

• Might be Certificate derived from Op’s CA

• Failure to authenticate is a failure to start the LSOE association, and HELLOs MUST BE restarted.
ACK

The Receiver ACKs with a Type=3

PDU Type is the Type of the PDU being ACKed

Might add PDU Number being ACKed

If the Sender does not receive an ACK in one second, they retransmit. Operator configured failure count.
Once We Know Each Other’s MACs

Layer Two KeepAlives May be Started
This is in addition to L3 BFD etc.

We assume that one or more Encapsulation addresses will be used to ping, BFD, or whatever the operator configures
We Know MAC/Ether Link State of This Device & Neighbor

And Node IDs (often ASNs)

Now Announce Encapsulations of the Link Interfaces
## Encapsulation PDU Header

<table>
<thead>
<tr>
<th>Type</th>
<th>PDU Length</th>
<th>Count</th>
</tr>
</thead>
</table>

---

<table>
<thead>
<tr>
<th>...</th>
<th>Encapsulation List...</th>
<th></th>
</tr>
</thead>
</table>

---

`0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1`
IPv4 Encapsulations

An Encapsulation message describes zero or more addresses of the encapsulation type.

An Encapsulation message of Type T replaces all previous encapsulations of Type T.
An Interface may have multiple Encapsulations
For each Encapsulation there might be multiple Addresses
One Address per Encapsulation SHOULD be marked as Primary
An Address may be marked as a loopback
IPv6 Encapsulations

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Type = 5       |    PDU Length    |   Count    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       ...       | PrimLoop Flags  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           |           |
|           |           |
|           |           |
|           |           |
|          IPv6 Address        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               |   PrefixLen       |   more ...   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
MPLS IPv4 Encapsulations

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type = 6 | PDU Length | Count |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ... | PrimLoop Flags| MPLS Label List ... |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ... | IPv4 Address |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| ... | PrefixLen | PrimLoop Flags|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| MPLS Label List ... |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| IPv4 Address |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Prefix Len | more ... |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```
## MPLS Label List

<table>
<thead>
<tr>
<th>Label Count</th>
<th>Label</th>
<th>Exp</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

+---------------------------------+
| Label Count | Label | Exp | S |
+---------------------------------+

+---------------------------------+
| Label | Exp | S | more ... |
+---------------------------------+
Use Multiple MPLS Label Encapsulations to Allow One Label to be Associated with Multiple AFI/SAFIIs and/or Multiple IP Addresses
MPLS IPv6 Encapsulations

+----------------------------------+
| Type = 7                         |
+----------------------------------+
| PDU Length                       |
+----------------------------------+
| Count                            |
+----------------------------------+
| ...                             |
| PrimLoop Flags                  |
| MPLS Label List ...             |
+----------------------------------+
| ...                             |
+----------------------------------+
| IPv6 Address                     |
+----------------------------------+
| +                                |
| +                                |
| +                                |
| +                                |
| +                                |
| IPv6 Address                     |
+----------------------------------+
| +                                |
| +                                |
| +                                |
| +                                |
| +                                |
| IPv6 Address                     |
+----------------------------------+
| +                                |
| +                                |
| +                                |
| +                                |
| +                                |
| Prefix Len                       |
| +                                |
| +                                |
| +                                |
| +                                |
| +                                |
| Prefix Len                       |
| +                                |
| +                                |
| +                                |
| +                                |
| +                                |
| more ...                         |
+----------------------------------+

Creative Commons: Attribution-NonCommercial-ShareAlike 2.0
We're Looking at Security
Are you the Droid I was talking to Earlier?
Thinking of Security

- OPEN has public key plus ...
- Signed with private key, proves possession
- All PDUs signed with 512-bit suffix
- KEEPALIVE could get a sequence number to reduce replay attack window
- Maybe a later Proof of Possession Challenge / Response PDU pair
Layer-3 IP/Label Liveness Should Also be Tested

One or more Discovered AFI/SAFI Addresses Are Used to Ping, BFD, ... to Assure Layer-3 Liveness
We now know all links, IDs, Encapsulation Types, and Addresses of this Device

Now Present an API to Topology and Dijkstra Layers
BGP-LS (RFC 7752) an extension to BGP to distribute the network's link-state (LS) topology
Open Questions
Should HELLO go Through an intermediate Layer Two Switch
Are HELLO and KEEPALIVE Redundant?
Should PDUs be Numbered so ACKs are not Ambiguous
Should the Version Number be Fail on MisMatch?
A Python3 Implementation is in Progress