Link-State Routing for Compute Nodes within a Cluster

Colette Caskie, New Mexico State University
Mentors: Lowell Wofford, Susan Coulter, Paul Peltz

Goal: Implement an open source dynamic routing protocol to increase reliability and maintainability while reducing resource overhead for high performance computing clusters.

Current Configuration Overview

IO Nodes

- Gateways for compute nodes used to access high speed parallel storage file systems that contain crucial data required for scientific research.
- Failure of an IO node hinders data transmission leading to possible non-recoverable data loss and job failure.

Dead Gateway Detection

- Custom script developed and maintained by LANL to determine the current health state of IO nodes.
- ICMP (ping) is utilized to identify broken links
- Reconfigures routing information within each of the compute nodes if a failure exists.

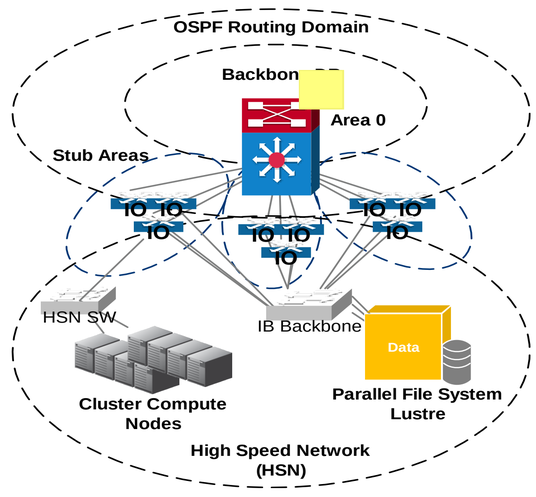

Figure 1: Standing OSPF Configuration

Open Shortest Path First (OSPF)

- Interior gateway routing protocol utilizing the link-state routing algorithm.
- Utilizes a tree topology requiring all areas to be attached to the backbone area 0.
- Routers assign a cost to each network segment/link, lower integers identify a link with higher preference.
- A Route is calculated all at once by each router using Dijkstra's Shortest Path First (SPF) algorithm and added to the shortest-path tree.
- OSPF Hello packet (equivalent to a half ICMP) is utilized to keep neighbor adjacencies
- Network traffic is minimized by only updating the routing information thru a process referred to as "reliable flooding".

Stretch Cluster

Stretch Cluster: Stub vs Totally Stub Areas

```
                Summary Link States (Area 0.0.0.0)

Link ID         ADV Router      Age  Seq#       CkSum  Route
172.16.0.0      10.15.6.1       194  0x80000031 0xe65f 172.16.0.0/16
192.168.0.0     10.15.6.1       194  0x80000032 0xb8df 192.168.0.0/16
```

```
                Summary Link States (Area 0.0.0.6 [Stub])

Link ID         ADV Router      Age  Seq#       CkSum  Route
0.0.0.0         10.15.6.1       194  0x80000002 0x8eae 0.0.0.0/0
10.15.1.0       10.15.6.1       194  0x80000001 0x0d04 10.15.1.0/24
10.15.6.0       10.15.6.1       104  0x80000002 0x6fa5 10.15.6.0/24
172.16.0.0      10.15.6.1       194  0x80000002 0x6314 172.16.0.0/16
```

```
                Summary Link States (Area 0.0.0.7 [Stub])

Link ID         ADV Router      Age  Seq#       CkSum  Route
0.0.0.0         10.15.6.1       204  0x80000001 0x90ad 0.0.0.0/0

                AS External Link States

Link ID         ADV Router      Age  Seq#       CkSum  Route
0.0.0.0         10.15.1.254     1356 0x8000002f 0xda8f E2 0.0.0.0/0 [0x0]
```

Figure 2: Type 3 LSA comparison between areas

- OSPF Totally Stub areas prevent communication between clusters.
- Totally stub areas reduce inter-area network traffic.

Designs Explored

Efforts to minimize production configuration changes displayed in *figure 1,* involved planning for an extra layer of separation for the compute nodes.

Compute Node Area Separation

- Virtual Links between areas not connected to the backbone and area 0
  Limitations: Virtual links are not supported by Arista hardware & Quagga
- IS-IS (intermediate system to intermediate system) is a link-state routing protocol similar to OSPF supports areas in the form of levels.
  Limitations: Redistribution and IP over IB not supported

OSPF Routing Domain Re-design

- Inner Router (IR): Designated Router (DR) Compute Nodes
- Area Border Router (ABR): IO Nodes
- Area X totally stub – prevents communication between clusters
- Non-Broadcast network for compute nodes - establishes only adjacency with IO nodes
- Broadcast configuration for IO nodes – establishes adjacencies with all directly connected routers.
- Default-Routes defined by OSPF (allows external domain communication)

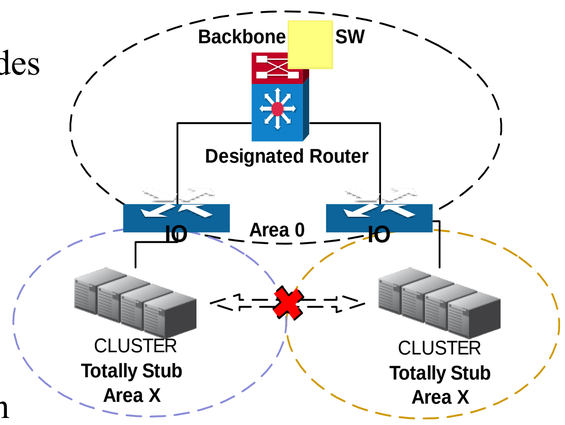

Figure 3: NEW OSPF Routing Domain

Results & Future Tests

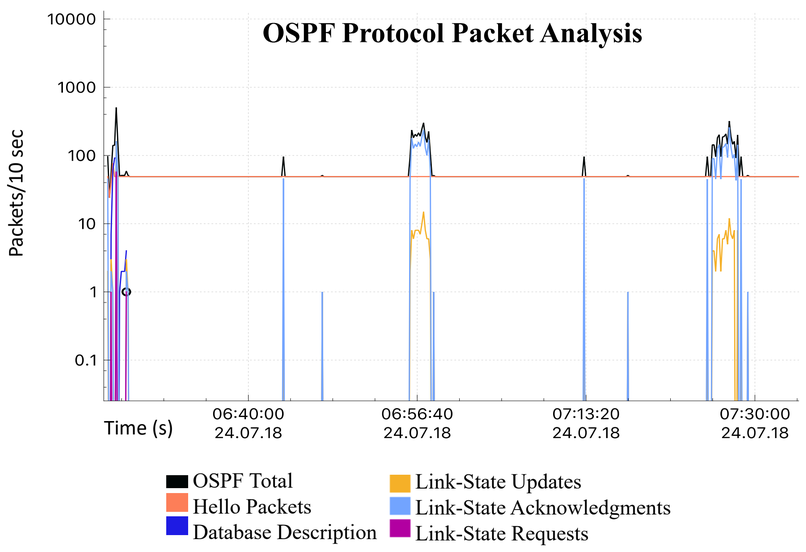

Figure 4: TCP packet capture on Stretch IO node

Results & Observations:

- Compute nodes only form adjacencies with IO Nodes
- OSPF Total packet count is highest when service is started or restarted: Database description packets account for this spike
- Packet count is minimal with packet no larger than ~2,000 bits
- "Reliable Flooding" occurs every ~30 min

Future Tests:

- Kit Cluster: 4-8 IO Nodes ~30 Compute Nodes
- Moon Cluster: ~50 IO Nodes ~1500 Compute Nodes

DGD overhead comparison to OSPF overhead
OSPF protocol scalability - Maximum Packet size ~ 65,000 Bytes

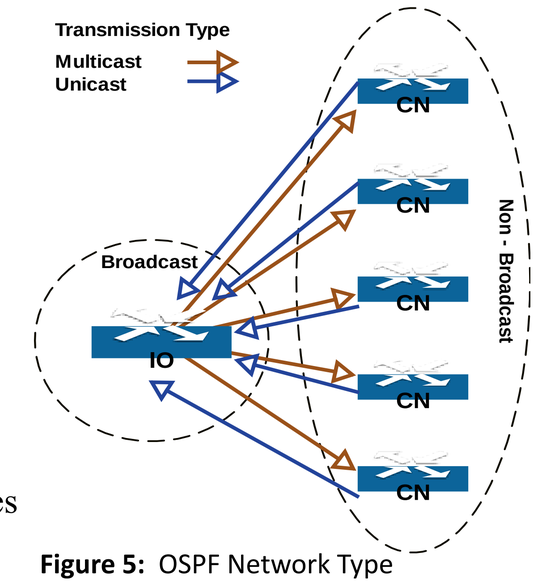

Figure 5: OSPF Network Type

Acknowledgements

LANL HPC Networking Team, Lowell Wofford, Paul Peltz, Susan Coulter, Dave Morton, Hunter Easterday

References

[1] Moy, J. T. (1998). *OSPF: Anatomy of an Internet routing protocol*. Reading, MA: Addison-Wesley.
[2] OSPF Version 2. (1998, April). Retrieved May 29, 2018, from https://tools.ietf.org/html/rfc2328
[3] Quagga. (1999). Retrieved May 25, 2018, from https://www.nongnu.org/quagga/docs/docs-multi/

LA-UR-18-27112