Layer 3 Network Configuration Best Practices for z/OS with OSPF

Mike Fox

Table of Contents

Executive Summary

z/OS® Communications Server provides TCP/IP and RDMA network capability on z/OS. It includes multiple networking applications, including the OMPROUTE routing daemon which implements both RIP and OSPF for both IPv4 and IPv6. These capabilities provide significant flexibility for designing network topology around z/OS. There are several best practices that enable maximum efficiency and flexibility in exploiting z/OS's networking capabilities, and those best practices evolve as the hardware technology evolves.

This paper discusses the evolution of networking hardware available for z/OS and how the evolution of that hardware, specifically the evolution of OSA Express connectivity from copper to fiber, has resulted in evolution of best practices for configuration of mainframe data centers.

This paper is provided for data center designers and implementers who have an interest in the performance and flexibility of data center layer 3 networks. It is assumed that readers already have a basic background in TCP/IP protocols and the related z/OS implementation of those protocols.

.

History

The topic of layer 3 network configuration best practices was addressed in detail in a 2001 White Paper, “OSPF Design and Interoperability Recommendations for Catalyst 6500 and OSA-Express Environments”, jointly published by IBM and Cisco. This White Paper can be downloaded from http://www.ibm.com/support/docview.wss?uid=swg27005474&aid=1 and was focused on designing networks that optimized the capabilities of OSPF, OSA Express features and Cisco switches that were available at the time for quick recovery from network outages.

Technology that featured prominently in the recommendations of that time was copper-based OSA Express features and Cisco Switches supporting a function called MSFC Autostate. Copper OSA Express features were not able to reliably detect hardware failures at the switch ports or the wires attaching them to switches (for example, un-plugging a wire or physical failure of the switch port).

Because of this inability of copper OSA Express features to reliably detect these failures, the 2001 White Paper recommended relying on a switch function known as MSFC Autostate to detect and recover from hardware failures that the OSA could not reliably detect. However to perform this detection quickly, MSFC Autostate required there be only one host per LAN or VLAN segment. This is because MSFC Autostate works by detecting that the switch is the only active port on the LAN or VLAN segment, so if there were more than one host on the segment this function would not detect the outage of one of them (note: because this paper discusses networks at a layer 3 level, for purposes of this paper, the terms LAN and VLAN are used interchangeably).

This recommendation resulted in “point to point” LAN segments, so called because each LAN segment contains only a single host and a single switch/router. An implication of this type of configuration is that any traffic that travels between hosts must go through a router. An example of a point to point LAN configuration is shown in Figure 11.

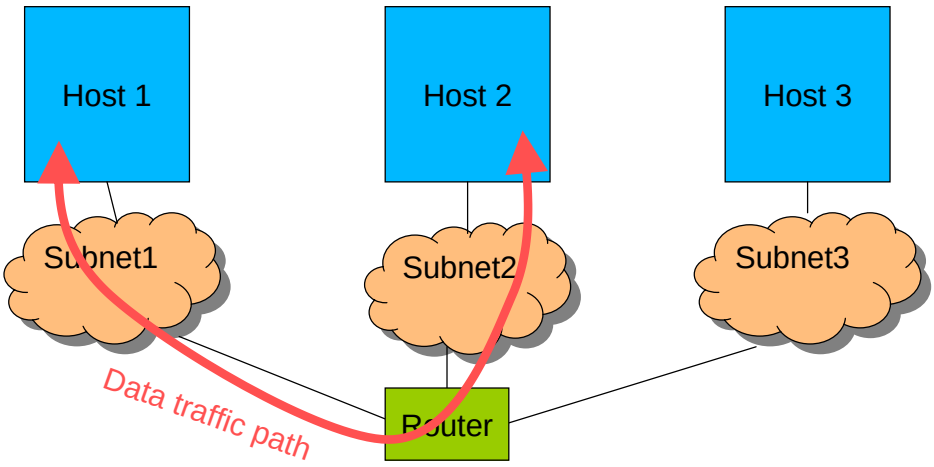

Figure 1: Point to point network configuration

Note that if OSPF is running on the LAN segment, outages that are not detected at lower levels are usually detected by the OSPF Hello protocol. However, this detection is not immediate because it relies on expiration of the configured Dead Router Interval, which on z/OS defaults to 40 seconds. There are also some configurations that preclude Dead Router detection; these will be discussed later in this paper.

These point to point LAN networks have some disadvantages. The main disadvantage is that all network traffic between hosts must go through a router, adding latency and the cost of the routers. This configuration also inhibits use of RDMA technology to communicate between hosts, because RDMA traffic is not routable and requires a direct network connection.

If there is a business reason to require all traffic between hosts to go through a router, for example a requirement for all traffic to be examined by a filter or a firewall, then this type of configuration is unavoidable. If the reason for this type of configuration is the recommendation made in 2001 based on the technology available at that time, this type of configuration may no longer be necessary.

Current Technology

Modern OSA Express features support fiber connectivity technology. One advantage of fiber-based OSA technology is that, unlike copper-based technology, it can reliably detect hardware failures in the optics, cabling and switch port. These failures are referred to as “loss of light” and because of fiber OSA’s capability to detect these failures, it is no longer necessary to rely on MSFC Autostate to detect hardware failures in the wire or the switch port.

When fiber OSA detects a loss of light, it immediately notifies the TCP/IP stack which immediately takes action. If OSPF is being used, the host immediately notifies its neighbors of the topology change and every host in the OSPF area is notified through the OSPF link state update protocol.

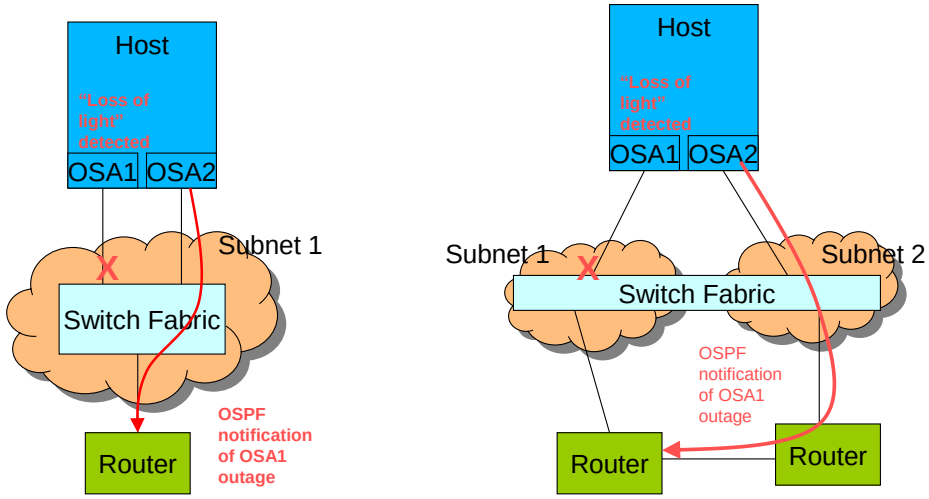

Figure 2: OSPF notification of OSA outage

Figure 22 illustrates loss of light detection combined with OSPF link state updates, providing quick notification and recovery from loss of light type events. In scenario A, both OSAs are on the same subnet and the surviving OSA immediately notifies the attached router of the failure of the other. Per the OSPF architecture, the routers immediately recalculate their routing tables to route around this failure.

In scenario B, the OSA that fails is the only OSA on its subnet so the host cannot notify the attached router of the failure. However the host notifies each of its neighbors of this failure over all of its remaining interfaces, and then each of the neighbors cascades this notification to their neighbors over their interfaces, and so on until all OSPF routers in the area have been notified. Each OSPF router immediately recalculates its routing table to route around the failure as soon as it receives the notification.

Note that because of the immediacy of OSPF link state flooding, the failure is detected and routed around in real time, without waiting for expiration of dead router intervals. Therefore, MSFC Autostate is no longer needed for quick recovery from these outages, so the need for point to point LAN segments is obviated.

Soft Failures

In a soft failure, the hardware continues to function but the switch port becomes unresponsive. An example of a soft failure would be an abend or a hang of the switch process that manages the port attached to the OSA. Because in a soft failure the hardware continues to function, there is no loss of light to be detected by the OSA Express feature so this failure must be detected by other means.

The primary means for detecting a soft failure is the expiration of the dead router interval as configured to OSPF. An OSPF host or router sends keepalive signals called "Hello" packets to all if its neighbors. If the dead router interval elapses without receipt of a Hello packet from a neighboring host or router, the neighbor is declared to be down and OSPF initiates recovery:

Dead router detection is slower than loss of light detection because instead of affirmatively detecting a hardware failure, it relies on the expiration of a keepalive timer. The value of this timer is configurable and on z/OS defaults to 40 seconds. This value can be shortened but doing so would also require shortening the Hello intervals. This is because the dead router interval is recommended to be four times the Hello interval, so that 1-2 dropped packets don't cause an outage, especially since Hello packets are delivered using multicast, which is unreliable. Decreasing these intervals increases responsiveness but at the cost of increased network traffic and CPU usage by the OSPF routing process.

MSFC Autostate and the resulting point to point networks may at first glance appear to be a quicker solution for detecting soft failures. However since soft failures are caused by software problems on the very switches and routers that would be expected to detect the outage, this strategy cannot be considered reliable.

Figure 3 illustrates OSPF detection of a soft switch failure. In this case the switch port to OSA1 on Subnet1 has hung, so the switch is no longer responding. This means OSPF Hello packets from the router will stop reaching OSA1. After the expiration of the configured dead router interval on the host, the host will declare the router adjacent to OSA1 to be unreachable and will propagate that information to the network.

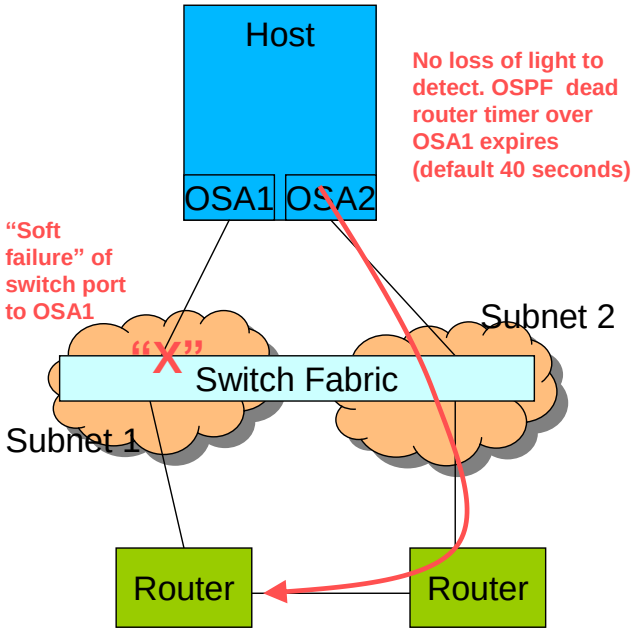

Figure 3: OSPF detection of soft switch failure

Soft failures that are undetectable by OSPF

Care must be taken when designing data center networks to ensure that the dead router interval can detect all soft failures. If soft failure detection using dead router intervals is a requirement, then no host should have two or more OSAs on the same subnet. This is because when there are two or more OSAs attached to the same subnet, OSPF designates one of them as Primary and all the others as Secondary.

An OSPF host uses the Primary interface for all OSPF protocol messages sent and received on the attached subnet. No OSPF protocol messages flow on Secondary interfaces, including Hello packets. This means that dead router detection is not available on secondary OSPF interfaces, as shown in

Figure 44. It is important to note that the OSPF Primary/Secondary designation has no effect on network data traffic, only on OSPF protocol messages. If configured OSPF costs are the same, OSPF treats all interfaces equally when computing routes for data traffic.

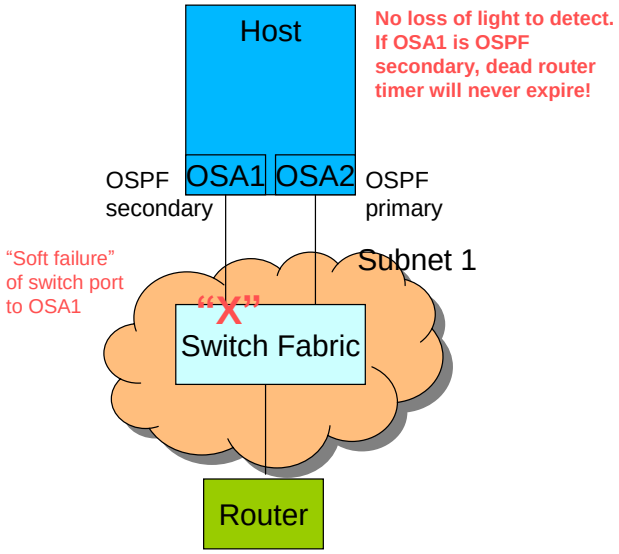

Figure 4: Soft failure on OSPF secondary interface

Recommendation

To avoid the soft failure over OSPF Secondary problem while still providing direct z/OS to z/OS IP connectivity with redundancy, a configuration along the lines of Figure 5 Configuration to avoid OSPF secondary is recommended. In this configuration, no host has more than one OSA on the same subnet therefore there are no OSPF Secondary interfaces. This means that all interfaces are monitored by the OSPF Hello protocol. However the hosts are still connected using shared subnets so traffic can flow between them without traversing routers. This configuration has additional advantages.

- It eliminates latency and complexity introduced by router hops
- It enables the use of RDMA (SMC-R) between the z/OS hosts.
- If VIPAROUTE is in use, direct connections enable direct forwarding of traffic.
- Jumbo frames can be enabled on the shared subnets, which eliminates fragmentation issues for VIPAROUTE and generally provides better IP performance for z/OS to z/OS communication

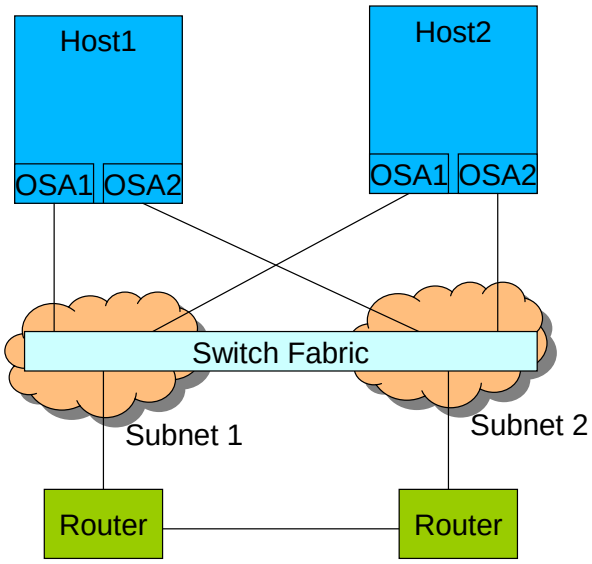

Figure 5: Configuration to avoid OSPF secondary

About the Author:

Mike Fox is a Senior Software Design and Strategy Architect in IBM Software Group's Enterprise Network Solutions team, focusing on routing and connectivity. He has over 25 years of experience as a developer and architect on IBM System z® and its predecessors. Mike can be reached at mjfox@us.ibm.com

IBM Systems and Technology Group
Route 100
Somers, New York 10589
U.S.A.
Produced in the United States of America,
06/2014

ZSW03262-USEN-00