VIOS 101: Network

July 2015 | by Jaqui Lynch

In Part 1, I wrote an introduction to virtualization and then discussed CPU virtualization with PowerVM. In Part 2, we addressed memory technologies. In this article, we'll be looking at network options and then in Part 4, we'll review I/O.

Network Options

As with nonvirtualized LPARs, dedicated network adapters are always an option, even for LPARs using micropartitions (shared processor pool LPARs). The key point to remember is that LPM (live partition mobility) requires that all adapters be virtualized at the time of the move, so no dedicated adapters can be in use at that time. With respect to virtualized network options, you need to understand some specific terminologies, specifically Link Aggregation (standard and 8023AD/LACP), Virtual Ethernet and Shared Ethernet Adapters.

Link Aggregation

EtherChannel and IEEE 802.3ad Link Aggregation are network port aggregation technologies that allow several Ethernet adapters to be aggregated together to form a single pseudo Ethernet device.

For example, the physical adapters ent0 and ent1 can be aggregated to a pseudo adapter ent3; interface en3 would then be configured with an IP address. The system treats the aggregated adapters as one adapter. All adapters in the EtherChannel or Link Aggregation are given the same hardware (MAC) address, so they are treated by remote systems as if they were one adapter.

The main benefit of EtherChannel and IEEE 802.3ad Link Aggregation is that they have the network bandwidth of all of their adapters in a single network presence and, if an adapter fails, the packets are automatically sent on the next available adapter without disruption to existing user connections. The adapter is automatically returned to service on the EtherChannel or Link Aggregation when it recovers. Thus, link aggregation provides both additional bandwidth and redundancy.

For redundancy, I typically have an aggregate on VIO1 that is connected to one network switch and an aggregate on VIO2 that is connected to a separate network switch. That way, there is redundancy within the VIO and across switches. All the physical ports in the aggregation group must reside on the same switch except in the case of a switch stack, where they can reside on different switches on the stack. Thus, having each VIO on a different switch helps provide redundancy. There are now network switches that can handle aggregates across switches but, for our purposes, we will assume those are not being used. Lastly, the network switch will need to be set to either standard or 802.3ad (LACP) aggregation. I typically use LACP but this should be discussed with the network team to see what they prefer/support.

NIB – Network Interface Backup

NIB is a type of EtherChannel that is used for high-availability only. NIB allows an aggregated adapter to have a backup. If all adapters that compose the aggregation fail, then communication is switched to the backup adapter until any adapter in the main channel recovers

In the NIB mode of operation, you have an adapter in the main channel and a backup adapter. While NIB by itself doesn't provide better bandwidth than the physical adapter, it can be used to work around switch failures. Usually port aggregation requires all adapters to be connected to the same switch, which makes the switch the single point of failure. By using NIB, and by connecting the primary and backup adapters to different switches, communication won't be lost by the failure of a single switch.

To help detect loss of network reachability (in addition to detecting failures in the adapter and its connection to the switch), NIB allows specifying an address to be pinged. If the given address cannot be reached after a given number of attempts (both specified when NIB is defined), then the current active adapter is considered down, resulting in the backup adapter taking over communication.

Virtual Ethernet

Virtual Ethernet has been around since AIX 5.3 and requires the use of an HMC or IVM (Integrated Virtualization Manager). IVM only supports the use of a single VIO server, which reduces redundancy. Virtual Ethernet allows LPARs to communicate with each other without having to assign physical hardware to the LPARs. The LPARs communicate via the Hypervisor over Virtual Ethernet channels. It also allows for VLANs and other security mechanisms.

Virtual Ethernet adapters are connected to an IEEE 802.1q (VLAN)-style Virtual Ethernet switch, which allows LPARs to share a common logical network. The system transmits packets by copying the packet directly from the memory of the sender LPAR to the receive buffers of the receiver LPAR without any intermediate buffering of the packet.

Each Virtual Ethernet adapter can be used to access up to 20 networks—the port VLAN ID and up to 19 additional VLAN IDs. The HMC generates a locally administered Ethernet MAC address for the virtual Ethernet adapters so that these addresses don't conflict with physical Ethernet adapter MAC addresses.

Virtual Ethernets are a great way to set up a private, secure, fast network between LPARs on the same server. Virtual Ethernet requires some CPU and memory to transfer network traffic. Performance and overhead (CPU and memory) are affected by entitlement as well as by the MTU size. If the data is never going external to the server then it's worth looking at either 9000 or 65394 as potential MTU sizes for the Virtual Ethernet as this can significantly reduce CPU overhead while increasing network bandwidth across the Virtual Ethernet. Details on this are provided in Alexander Paul's presentation at http://bit.ly/1eLHFT6.

Shared Ethernet Adapter

Shared Ethernet adapters (SEAs) are designed to bridge real adapters to Virtual Ethernet adapters. Rather than having a dedicated Ethernet adapter in every LPAR, SEA involves having the actual adapter (or a group of adapters aggregated together) assigned to the VIO server. The LPARs talk using Virtual Ethernet to the VIO server and any traffic that needs to go outside the box is sent out via the Shared Ethernet adapter in the VIO server. For redundancy, normally two VIO servers are acting as SEAs and it's common to set one up as primary with the other running as the SEA failover VIO server. SEA is used to reduce the number of Ethernet cards on a system and to allow for failover and redundancy.

SEAs can be set up in failover mode or load sharing mode. With failover, the Ethernet adapters on the backup SEA sit idle until a planned or unplanned failure on the primary VIO's SEA. With load sharing, you can set some VLANs to use the SEA on VIOS1, and other VLANs to use VIOS2, and still have both SEAs ready to take over should the other VIOS fail. Another flavor is SEA failover with NIB.

SEA Failover. SEA failover is the technology most people are familiar with and is designed to work with two VIO servers, which requires an HMC. An aggregate is created on each VIO server using the mkvdev command. That aggregate is then turned into an SEA where one VIO is set up with a trunk priority of 1 and the other with a trunk priority of 2. It's important to get this right to avoid spanning tree loops. Prior to PowerVM 2.2.3, a VLAN ID had to be reserved for a control channel that is used by the two VIO servers to send commands as well as keep-alive messages. PowerVM 2.2.3 introduced simplified SEA failover where a control channel doesn't need to be defined. If it's not defined, then the SEA will default to using VLAN 4095. It's important that this VLAN is not defined anywhere else.

With SEA failover, one VIO is primary and has the full bandwidth of all the adapter ports in the SEA on that primary VIO. If the primary VIO fails or loses its connectivity, then the secondary VIO becomes primary and network connectivity continues. If each VIO is on a separate switch, then you also get redundancy across switches. Assuming you have two adapters in the aggregate on each VIO (total of four) then you can use the bandwidth of two adapters.

Depending on the actual setup, a failover for SEA failover can take up to 30 seconds. This depends on network switch and spanning tree setup.

SEA with NIB. SEA with NIB is an SEA that's defined with one of the adapters set up as a backup adapter. That allows the NIB adapter to be on a separate switch to the primary aggregate. This means that if the primary aggregate or switch fails, then the NIB adapter on the same VIO gets activated. Additionally, there is also failover to the secondary VIO. As already mentioned, NIB uses ping to monitor the active aggregate so it creates a lot of ping traffic. It also reduces available bandwidth. Comparing it to the previous example: Assuming you have two adapters in the aggregate on each VIO (total of four) then you can use the bandwidth of only one adapter, as the second adapter on each VIO is reserved for NIB. This is a very expensive solution when using 10Gb/s adapters.

SEA Failover with load balancing/sharing. Load sharing was introduced with PowerVM v2.2.1. In this case, we define two SEAs and the SEAs negotiate the set of VLAN IDs that they are responsible for bridging. Once negotiation is complete, each SEA bridges the assigned trunk adapters and the associated VLANs. Thus, both SEAs are in use, allowing you to take advantage of all of the adapters in both VIO servers. If a failure occurs, the remaining active SEA bridges all trunk adapters and the associated VLANs. To define this, each SEA needs to have the ha_mode=sharing attribute rather than the ha_mode=auto:

One major difference is the available bandwidth. Comparing the three options where we have two adapters on each VIO (four adapters total) we see:

Type	**Bandwidth**
SEA Failover	2 adapters
SEA Failover with NIB	1 adapter
SEA Failover with load sharing	4 adapters

SEA Failover Versus NIB

SEA failover has some advantages over NIB. In particular, it's simpler as SEA failover is implemented on the VIO server and the client configuration is simple. NIB is more complex, especially at the client LPAR where a second Virtual Ethernet adapter on a different VLAN is required for the NIB adapter.

Additionally, SEA failover has the added support of IEEE 802.1Q VLAN tagging and it simplifies NIM installation because only a single virtual Ethernet device is required on the client partition. Finally, SEA failover is significantly less "pingy" than NIB. With SEA failover, only the SEAs send out periodic ping requests for checking the network availability. With NIB, every client partition will send out ping requests, resulting in more network traffic.

Summary

As you can see, you have multiple ways to virtualize the network for LPARs on your servers. Networking can be as complex or as simple as you want to make it—it's highly recommended that the design be as simple as possible and that the network team be part of the design so you can best determine where to take advantage of larger MTU sizes to improve performance. Additionally, techniques such as flow control, large receive and large send can assist with performance. More information on these is provided in the network performance session at the movies site in the references (see box).

RESOURCES VIDEO SOLUTIONS EDITION BLOGS WEBINARS SUBSCRIBE ABOUT US

Systems

Connect With Us:

Magazine Archives

AIX LINUX ON POWER MAINFRAME POWER

Search

IBM i ADMINISTRATOR DEVELOPER TRENDS TIPS & TECHNIQUES CASE STUDIES STORAGE PRODUCT NEWS ENDPGM

References

< Return to main article

Print Email

PowerVM Enhancements in 2013

http://publib-b.boulder.ibm.com/abstracts/sg248198.html?Open

PowerVM Virtualization Introduction and Configuration

http://publib-b.boulder.ibm.com/abstracts/sg247940.html?Open

PowerVM Virtualization Managing and Monitoring

http://publib-b.boulder.ibm.com/abstracts/sg247590.html?Open

PowerVM Best Practices

http://publib-b.boulder.ibm.com/abstracts/sg248062.html?Open

SEA Techdoc

http://www-01.ibm.com/support/docview.wss?uid=isg3T7000527

Movies site for replays of performance and other sessions

http://www.circle4.com/movies

< Return to main article

READ THE CURRENT ISSUE: DIGITAL | ONLINE | eNEWSLETTER

IBM i | AIX | LINUX ON POWER | MAINFRAME | POWER |

Connect With Us:

Homepage About Us Contact Us Subscriptions Editorial Calendar

Advertise With Us Reprints Privacy Policy Terms of Service Sitemap