IBM PowerVM Technical Webinars

Session 7

23rd Nov 2011

Virtualisation and the world of 10Gbit Ethernet

Gareth Coates

gaz@uk.ibm.com

IBM Power Systems,

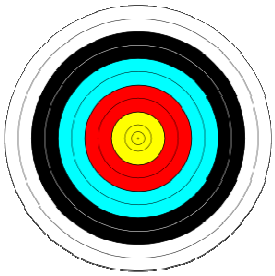
Advanced Technical Support (EMEA)

PowerVM

IBM Systems Director

Aim of this session

- Be aware and be prepared
  - Don't make false assumptions
  - You'll need to do some configuration
- The good news
  - What can you get?
  - What are realistic expectations?
  - Some experimental results
- More info
  - Other sources of information



Be aware and be prepared

- There are a number of things which people may overlook
- It is easy to make false assumptions
- It probably will "just work straight out of the box"
- But you can do so much better

Don't Assume 1 - The same boost as last time

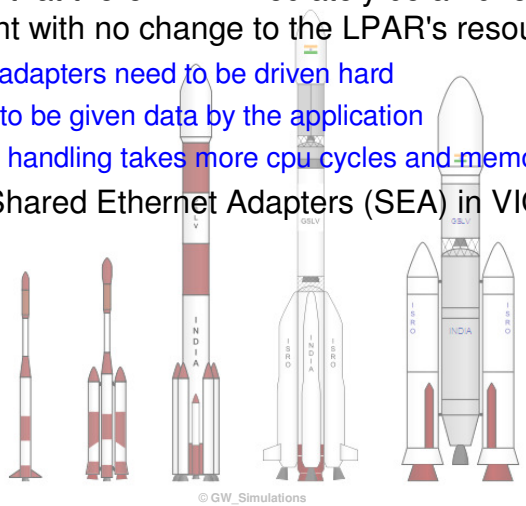
- It may be that you migrated:
  - 2Mbit to 10Mbit
    - saw approximately a 5 fold increase in performance
  - 10Mbit to 100Mbit
    - saw approximately a 10 fold increase in performance
  - 100Mbit to 1Gbit
    - saw approximately a 10 fold increase in performance
- Don't assume that when you go from:
  - 1Gbit to 10Gbit
    - that you will *automatically* see another 10 fold improvement.
  - you are trying to push a lot more through the system
    - hardware and software

Don't Assume 2 - Your old Ethernet cables will work

- Your structured cabling may be good enough to carry 1Gbit, but will it support 10Gbit?
- See "**A Primer on Power Systems 10 Gb Ethernet**"
  - http://www.ibmsystemsmag.com/aix/administrator/networks/A-Primer-on-Power-Systems-10-Gb-Ethernet/
- If you are using Copper:
  - Category 6A cables are needed for up to distances of 100 metre.
  - Category 6 cables can be used for shorter distances
- For Fibre:
  - it is not quite so simple
  - check rather than assume

Don't Assume 3 - This won't use any more CPU or memory

- Don't assume that if you simply replace a 1Gbit card with a 10Gbit card that there will immediately be a 10 fold improvement with no change to the LPAR's resources.
  - These fast adapters need to be driven hard
  - They need to be given data by the application
  - All the data handling takes more cpu cycles and memory
- Applies to Shared Ethernet Adapters (SEA) in VIOS as well.



Don't Assume 4 - Everything can be Max-ed out at the same time

- It not safe to assume that all the ports of a switch can go at full speed all of the time.
- It is not necessarily the case that a switch can handle all of its ports going at full speed.
- Check the specifications.
- The size of the buffers can be important too.
- Same goes for multiple port adapters - don't expect all four ports to Max-out at the same time as the on-board CPU can be a limit - especially with small packets
- Remember all four ports share the PCI slot which has a particular bandwidth

Don't Assume 5 - A faster network fixes everything

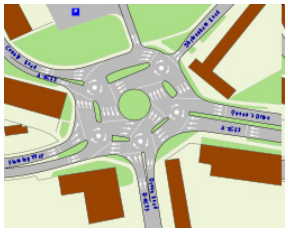
- A hyperthetical company has some POWER6 based servers.
- They are consolidating to POWER7 and are considering adopting 10Gbit technology.
  - A particular LPAR runs a production database and is using 36 rPerfs.
  - They want to have about 2X the database performance, so plan an LPAR with 72 rPerfs on the POWER7 server, and expect a 10X performance increase on the LAN.
- They have doubled the rPerfs
  - OK for the database,
  - but expecting the network device drivers and TCP/IP stack to be able to do **10** times as much work

Don't Assume 5 - A faster network fixes everything

- To be able to see a 10 fold increase in bandwidth, the data needs to be available.
- It is quite possible that applications running on smaller POWER7 LPARs, and their attached storage; will not be able to generate that sort of data flow.
- If performance is not as expected
  - The bottleneck is not necessarily the network
  - It could be a single threaded process or internal application locking on logical resources or internal buffering of the data arriving.

You'll need to do some configuration

- Now that we know some of the bad assumptions to avoid…



- **It can work really well**
  - It is possible to achieve respectable bandwidth when using 10GBit Ethernet on IBM System Power equipment.
  - With some careful consideration, planning, configuration and tuning, it is possible to achieve **very** good results.
- **so here are our first set of hot tips ...**

Top Tip 1 - Flow Control

- Turn on Flow Control - everywhere.
- With 10 Gbit Ethernet it is very useful to turn on flow control to stop the need for retransmission.
  - It is very easy at high bandwidth to completely fill buffers on switches and adapters so that transmitted packets are dropped.
  - This leads to time-outs and retransmits. Together, these lead to delays, wasted bandwidth and compute cycles and uses more energy.
- If flow control is enabled, packets flow most efficiently
- This is MUCH more important with 10Gbit than before
- There are multiple places to turn on flow control - see following tips....

- Imagine a railway system with no signals ⇨ chaos

Top Tip 2 - Flow Control on the network switches

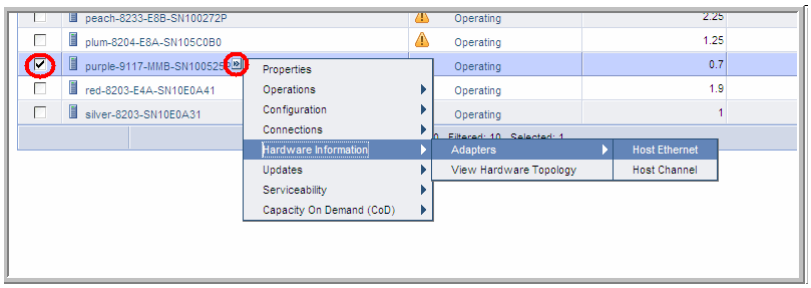
- Turn on flow control on the network switch
- The method depends upon the manufacturer of the switch
- Don't assume it is on
- Don't even assume that the switch port is configured to connect at 10Gbit

Top Tip 3 - Flow Control on AIX networks

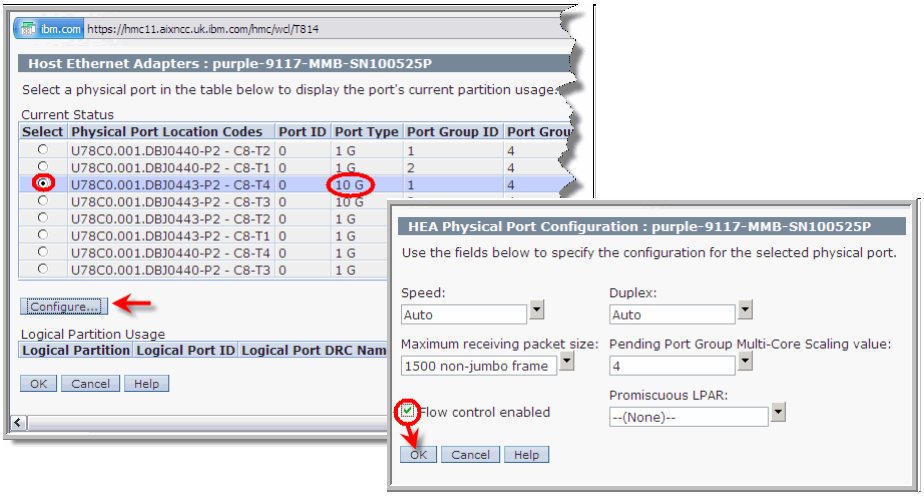
- Turn on flow control In the Operating System
- For AIX, you can use
  - `chdev -l ent# -a flow_ctrl=yes`
  - replace # with the number of the adapter
- to turn on flow control

Top Tip 4 - Flow Control on the Host Ethernet Adapter (HEA/IVE)

- If the adapter is part of a Host Ethernet Adapter(HEA), you need to turn on flow control there too.
- You can do this via the HMC:
  - Select the Managed System-> Hardware Information -> Adapters-> Host Ethernet



Top Tip 4 - Flow Control on the Host Ethernet Adapter (HEA/IVE)



- Select the appropriate adapter and click "**Configure**"

Top Tip 5 - If possible map the HEA directly, don't use a SEA

- You *can* use a 10 Gbit Host Ethernet adapter as part of a Shared Ethernet Adapter (SEA).
  - It is completely possible and supported
- **Ideally, don't do it, though.**
- Much better to map a logical HEA port directly to each of the client LPARs,
  - Configured via the HMC
  - Much better performance with less tuning.
  - For LPARs which need a lot of bandwidth, you can dedicate the adapter to the particular LPAR.
- This by itself would mean that Live Partition Mobility would not be possible.
  - It is possible to dedicate a network adapter to an LPAR as desired and use it as part of an Ether-channel with an SEA as a backup device.
  - LPM is possible with the use of a shell script to temporarily remove the HEA.
  - There is some smitty support for this type of thing
    - http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=%2Fcom.ibm.aix.baseadmn%2Fdoc%2Fbaseadmndita%2Flpm_overview.htm
    - http://publib.boulder.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.baseadmn/doc/baseadmndita/lpm_running.htm

Top Tip 6 - Set the network option largesend

- Turn on the network option largesend (and large_receive)
  - It allows TCP to build a message up to 64 KB long and send it in one call down the stack
  - That would take 44 calls at 1500 bytes
- When using a 10Gbit adapter in AIX, turn on largesend using the ifconfig or chdev command on the VIOS SEA.
- AIX Example:
  - `ifconfig en0 largesend`
  - `chdev -l en0 -a largesend=on`
- VIOS Example:
  - `chdev -dev ent2 -attr largesend=1`
  - `chdev -dev ent2 -attr large_receive=yes`

Top Tip 7 - Large packets with MTU

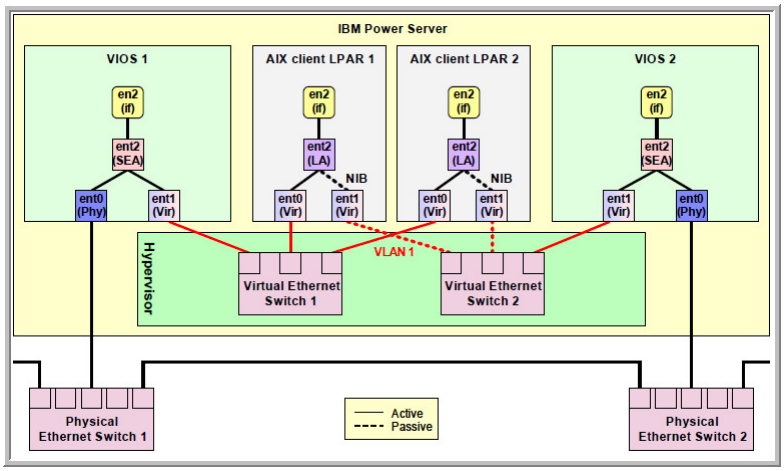
- We have found that increasing MTU to 64K for virtual adapters can have a massive benefit when using 10Gbit physical adapters in Power servers.
- This can drastically reduce the number of transactions (one per packet) on the CPUs and adapters.
  For quickness: smitty chif
- This uses chdev as in this example:
  - `chdev -s en0 -a mtu=65535`
- The actual number you can use can depend on the device you are using, so if the above does not work try 65536 (the full 64K), 65394 (64K minus overhead), 65390 (64K minus VLAN overhead) or at least try 9000.
- The packets will automatically be divided into the physical MTU before leaving the hardware

Top Tip 8 - Multiple virtual switches in POWER6 and POWER7 machines

- Most people haven't heard of multiple virtual switches but they are worth considering, so don't worry if you are uncertain :-)
- The "IBM PowerVM Virtualization Introduction and Configuration" Redbook, talks about them in section 2.10.3.
- At first it describes the concept of one virtual switch and then goes on to talk about multiple virtual switches, which was introduced in POWER6 based servers, multiple virtual switches offer performance and resilience benefits.
- Also see the few notes in Infocenter
  - Configuring Details Virtual Ethernet Switch
  - http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hby/configdetailvirtethernetswitch.htm
- Here is an excellent Whitepaper:
  - Using Virtual Switches in PowerVM to Drive Maximum Value of 10 Gb Ethernet (format PDF 1.5 MB)
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101752

Top Tip 8 - Multiple virtual switches in POWER6 and POWER7 machines

- They are configured using the HMC



Top Tip 9 - Set the network option rfc1323

- Enable the rfc1323 Option
  - Enables TCP enhancements as specified by RFC 1323
  - TCP Extensions for High Performance.
- AIX example
  - `no -p -o rfc1323=1`

Top Tip 10 - do NOT use ftp as your benchmark tool

- ftp has some significant single threadedness
- ftp is not efficiently written for 21st century hardware
- netperf is a much better bet

- If you want to break the sound barrier – don't use a glider!

The good news - What can you get?

- We carried out some tests using a p795
  - 2 Nodes
    - 64 cpus, POWER7, 4GHz
    - 256GB RAM
- One server was used to both send and receive data.
  - sending LPARs and VIO servers were on vswitch1
  - receiving VIO server and LPARs on vswitch2.
- Two physically separate servers would have been better
- Placement of cards in single IO drawer
  - in slots P1-C4 C7 C9 C10 and P2- C4 C7 C9 C10
  - this not optimal
  - C1,C4,C2,C5 would have been better.
- One of the four 10Gbit Ethernet cards was a FCoEE adapter rather than a 10Gbit Ethernet adapter
- Testing was performed using the commonly available “netperf” network performance test program

The good news - Inter LPAR tests

- All LPAR virtual networks are configured with
  - largesend and an mtu size of 65390
- For a single process on a single VNET, LPAR to LPAR
  - maximum observed bandwidth was 1700MB/sec – 13.6Gb/sec
  - average observed bandwidth being 1010MB/sec – 8Gb/sec
- The best aggregate performance was 72.5Gb/sec
  - 4LPAR to 4LPAR 8 processes in each LPAR over 4 VNET

The good news - LPAR to LPAR via VIO Servers

- Single process in 1 LPAR to a single process in another LPAR via VIO servers and 10Gbit Ethernet
  - averages 450MB/sec (max seen 720MB/sec)
- Using a single SEA over single Etherchannel with 4x 10Gbit Ethernet adapters
  - (mode =802.3ad,src_dst_port,jumbo_frames,largesend,large_receive,flow_control)
  - 1LPAR to 1Lpar 1 process in each LPAR over 1 VNet 694MB/sec
  - 4LPAR to 4LPAR 32 process in each LPAR over 4 VNet 2400MB/sec
- Using 2x SEA over Dual Etherchannel with 2x 20Gb Ethernet adapters in each EtherChannel
  - 2LPAR to 2LPAR 2 process in each LPAR over 2 VNet 1036MB/sec
  - 4LPAR to 4Lpar 8 process in each LPAR over 4 VNet 2650MB/sec

The good news - LPAR to LPAR via VIO Servers

- Using 4x SEA each over 1x 10Gbit Ethernet
  - 4LPAR to 4LPAR 4 process in each LPAR over 4 VNET 1850 MB/sec
  - 4LPAR to 4LPAR 32 process in each LPAR over 8 VNET2500 MB/sec

The good news - What are realistic expectations?

- An SEA returns the maximum aggregate bandwidth when two Virtual network adapters are configured to it rather than one. It is not possible to achieve line speed of 10Gbit Ethernet through an SEA with a single attached Virtual Network – it is possible with two virtual network adapters.
- An individual process communicating to a remote process will see an average of 450MB/sec bandwidth through VIO servers using 10Gbit Ethernet
- Affinity and locality of thread/data may impact performance of an individual thread
  - Tip: use the latest firmware in your server

The good news - What are realistic expectations?

- For an LPAR to achieve the maximum throughput over virtual networks the LPAR must communicate with 2 (or more) virtual networks and use more than one process on each virtual network.
- For maximum aggregate bandwidth use 2 or 4 VIO servers, each with 2x 10Gbit Ethernet adapters configured in an Etherchannel. If possible, the attached SEA should have two virtual networks.

Reminder

- Using the correct configuration options is essential
  - In each LPAR
    - enable largesend
    - use the maximum available mtu size (64k)
  - Ensure each 10Gbit Ethernet adapter has
    - largesend, flow_control, large_receive and jumbo_frames enabled
  - Etherchannel adapters should use
    - 8023ad mode, src_dst_port hash mode and have jumbo frames enabled
  - The SEA adapter itself needs to have
    - largesend, large_receive and jumbo_frames enabled
- You can only use the available bandwidth if your application can drive that amount of data

More info

- The AIXpert Blog
  - http://tinyurl.com/AIXpert
  - look for "10Gbit"
  - will be updated as we find out more
- See "**A Primer on Power Systems 10 Gb Ethernet**"
  - http://www.ibmsystemsmag.com/aix/administrator/networks/A-Primer-on-Power-Systems-10-Gb-Ethernet/
- See "**Controlling the Flow**"
  - http://www.ibmsystemsmag.com/aix/administrator/networks/Controlling-the-Flow
- IBM PowerVM Virtualization Introduction and Configuration
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg247940.pdf

Summary

- If you:
  - plan carefully
  - configure your system appropriately
  - have realistic expectations
- Then you can
  - achieve great performance using 10Gbit Ethernet on Power Systems

W5: Pillars of Star Formation © NASA

Thank you for attending this session!