Technical Tivoli Storage Talk

February 10, 2009

Preparing for the next TSM Version Release

Brandon W. Moore

brmoore@us.ibm.com

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

Introduction

- 17 Years in Information Technology
  - 4 Years in Academia
    - **First webmaster of Florida Agricultural and Mechanical University**
    - **First webmaster of FAMU-FSU College of Engineering**
  - 13 Years in Corporate America
    - **Engineering Systems Engineer, EDS and General Motors**
    - **Subcontractor, IBM Global Services**
    - **Systems Analyst, Cingular Wireless**
    - **System Engineer (SE), IBM Software Group**
      - **Tivoli Storage**
      - **West Region, Southwest Business Unit**
    - Tivoli Technical Storage Leader, GTS SO SWG Sales

Introduction (cont.)

- This technical Storage Talk will focus on how to help your TSM customers prepare their current TSM environment for an upgrade to TSM 6.
- We will look at proven practices around the TSM database, including the importance of IOPS (Input/Output Per Second) from your disk subsystem.
- We will then cover suggestions for hardware selection, LAN and SAN for TSM servers.

The challenge…

- What is the most important piece of the transition to the next version of TSM is…

The TSM Database upgrade

The challenge (cont.)…

- Understanding the physical layout of the TSM database will be key to your ability to quickly migrate and improve your overall performance
- We now have to consider things that are considered best practices in the database world within TSM world
- This presentation is designed to help you consider these ideas

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

Background

- ***Disk I/O Primer***
- ***Sequential vs. Random***
- ***Caching vs. No caching***
- ***RAID***

Background:
Disk I/O Primer

- A single physical disk is capable of 100-200 IOPs
  - Random I/O is typically lower IOPs
  - Sequential I/O is typically higher IOPs
- The TSM database writes 4k chunks in batches as large 64 chunks at a time
- The TSM database during backup reads in 256k chunks

Background:
Sequential vs. Random

- Sequential IO
  - TSM recovery logs
    - Write Heavy
    - Read Heavy
  - TSM Database
    - Write Light
    - Read Light, except during Backup

Background:
Sequential vs. Random

- Random IO
  - TSM recovery logs
    - Write Light
    - Read Light
  - TSM Database
    - Write Heavy
    - Read Heavy

Background:
Caching vs. No caching

- Caching
  - TSM database caches IOPs
    - TSM bufpool
    - TSM bufpool size: 1 gigabyte or greater
  - RAID Subsystems also perform caching

- No Caching
  - SVC/Virtualized storage
  - Non SVC/Virtualized storage

Background:
RAID

- ***Strips and stripes forever***
- ***RAID 0, RAID 1, RAID 5 and RAID 10***
- ***“RAID TSM”***

Strips and Stripes forever

- The idea is that a certain amount of data, typically some multiple of the file system's or database's defined block size, is written in a single operation to a single member of the array.

- The data written to the individual array member is called the data strip.

Strips and Stripes forever

- The amount of data (in bytes) written to a strip is referred to as the *stripe depth* (yes, stripe depth not strip depth).

- By aligning the stripe depth with the block size of the host file system, data I/O operations can be performed consistently and quickly.

RAID 0, RAID 1, RAID 5 and RAID 10

- Using RAID help increase the number of IOPs
  - RAID 0: Striping, no redundancy
    - Best Read performance
    - Best Write performance
  - RAID 1: No striping, redundancy
    - Better Read performance
    - Basic Write performance
  - RAID 5: Striping and fault tolerance
    - Best Read Performance
    - Worst write performance
  - RAID 10: Striping and redundancy
    - Best Read performance
    - Best Write performance

“RAID TSM”

“RAID TSM” or TSM mirroring

- If TSM is managing the mirror and it detects corrupt data during a write, it will not write the corrupt data to the second copy.
- TSM can then use the good copy to fix the corrupt mirror.

“RAID TSM”

- TSM also mirrors at transaction level, and hardware at IO level.
- Hardware will always mirror every IO, but TSM will only mirror complete transactions. This also protects the mirror from corruption.

“RAID TSM”

- RAID TSM performs RAID 1 with application level intelligence
  - Load balancing on read write operation
  - Parallel or sequential writes
  - Redundancy
  - Mirror across subsystems
- Allow the operating system or RAID controller to provide RAID 0 protection

“RAID TSM”

- Do not use RAID 1 at the operating system or RAID controller level
  - Potential for data corruption
  - Cheaper and better for TSM to do it.

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

Operating Systems

- ***AIX***
- ***Windows***
- ***Linux***
- ***Others***

Operating Systems

- AIX, Linux and other *nix
  - Filesystems: Yes
  - Raw Logical Volumes: No, except during DR
  - Caution: Logical Volume Managers
  - Decrease the amount of physical RAM used

- Windows
  - Filesystems: Yes
  - Raw Logical Volumes: No, except during DR
  - Logical Volumes Managers: No

NOTE:
Do not fill the filesystems past 95% capacity with the TSM databases volumes. Additional space is needed for bad block allocation and other overhead

Operating Systems

- AIX supports RAID0, RAID1 and RAID10.
  - RAID0 is not really a good idea, as if a logical volume was spread over five physical volumes, then the logical volume is five times more likely to be affected by a disk crash. If one disk crashes, the whole file is lost.
  - RAID1 is straight disk mirroring with two stripes and requires twice as much disk.
  - RAID10 combines striping and mirroring, and also uses twice as much disk.
- If AIX is mirroring raw logical volumes it is possible for it to overwrite some TSM control information, as they both write to the same user area on a disk. The impact would be that TSM would be unable to vary volumes online.

Operating Systems

- AIX supports RAID0, RAID1 and RAID10.
  - RAID0 is not really a good idea, as if a logical volume was spread over five physical volumes, then the logical volume is five times more likely to be affected by a disk crash. If one disk crashes, the whole file is lost.
  - RAID1 is straight disk mirroring with two stripes and requires twice as much disk.
  - RAID10 combines striping and mirroring, and also uses twice as much disk.
- If AIX is mirroring raw logical volumes it is possible for it to overwrite some TSM control information, as they both write to the same user area on a disk. The impact would be that TSM would be unable to vary volumes online.

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

Subsystems

- ***FC***
- ***Ethernet Network Attached***
- ***Virtualized***

Fibre Channel

- Disk HBA ports to Tape HBA ports ratio should be 2:1
- Physical disk to HBA ratios should be maintained
- The database should be mirrored on internal and external disk
  - TSM writes to the database volumes in the order they are created
  - Remember to remove the default TSM database volume(s) from your initial setup of TSM
- Align the RAID strip/stripe parameters, with filesystem blocks/chunks and TSM database blocksize

Fibre Channel

- Which type of physical disk for TSM database
  - For low or mid tier disk subsystems this means that multiple disk heads can be seeking, reading, and writing simultaneously.
  - High tier subsystems perform most of their I/O in cache, so this is less of an issue but you have cost trade-offs. Writing to cache for the TSM database is fine, but writing to disk subsystem cache for diskpool volume is not per se

Fibre Channel

- Do not use RAID 1 FC disk
  - Most disk subsystems support RAID1 mirroring, which is expensive as it needs twice as much disk, and will not detect logical errors in the data.
  - All data is mirrored even if it is corrupt.
  - **Always** use RAID TSM

Ethernet Network Attached

- Do not to use Ethernet Network Attached storage for TSM
- Use internal, fibre channel or virtualized before you use Ethernet network attached storage

Virtualized

- Virtualized storage like SVC can help TSM immensely
  - Striped vDisk
  - Advanced copy features
- Virtualized storage like SVC can hurt the whole environment when TSM is used

SOLUTION: Use cache disabled vDisks for TSM

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

TSM

- ***How the TSM Database writes***
- ***TSM Database***
- ***TSM Recovery log***

How the TSM Database writes

- Writes transactions in 4k chunks to the recovery log in sequential order
- At a certain point, determined by your logmode, these transactions are written to the TSM database files in random order

TSM Database

- Lots of smaller volumes
  - The TSM Server process can be more efficient by assigning multiple threads
  - Minimizes your exposure the certain types of failures
  - TSM will schedule one concurrent operation for each database disk it makes sense to allocate a lot of small disks, rather than a few large ones.

TSM Database

- TSM RAID is enhanced
  - One thread per database volume allows TSM to respond quicker to IOP requests
  - “TSM RAID” will retrieve the information from the fastest volume

TSM Database

- This sounds obvious, but the mirrors need to be on different disks.
- It is also possible to mirror to three ways as well as two ways. With three-way mirroring you get three copies of the data.
- Use three way mirrors, if at all possible

TSM Database

- Database backup and Expire Inventory are both CPU intensive processes that can be used to indicate server performance problems in general.
- The only sensible answer to 'how big should a TSM database be?' is to let you database grow until these following processes start to become an issue.
  - Expire Inventory should finish before your daily backups start
  - Database Backups should take no longer than an hour
  - Database Restores should take an hour and half no more

TSM Database

- How do you verify if your database is running well?
  - Expiration

    select activity, cast((end_time) as date) as "Date", (examined/cast ((end_time-start_time) seconds as decimal(24,2))*3600) "Objects Examined Up/Hr" from summary where activity='EXPIRATION'

  - Database backups

    select activity, ((bytes/1048576)/cast ((end_time-start_time) seconds as decimal(18,13))*3600) "MB/Hr" from summary where activity='FULL_DBBACKUP' and days(end_time) - days(start_time)=0

- Determine database size using the database page size

  select dec((dec(page_size)*dec(used_pages))/1073741824,10,2) as "GB" from db

TSM Database

- The actual size will depend on how fast your hosting server is, how good your disks are and what level of service you need to provide.
- Proper disk layout and configuration at the hardware level can improve this behavior
- 85% capacity of two formatted 15k RPM, 146.8 gigabyte physical disk is my upper limit for TSM database size as of Version 5.X

TSM Database

- ESTIMATE DBREORGSTATS
  - Conduct on a restored database copy
  - Use TSM database snapshots
  - This is a good way to verify how a new disk layout for TSM database will perform

TSM Database

- For older releases of TSM use the QUERY DB to see if you need to defrag your TSM DB.
- How do we do this with the older TSM instances

TSM Database

- For older TSM server releases(GASP!!!)
  - The following formula can be used to see how much space could be reclaimed by an unload/reload.

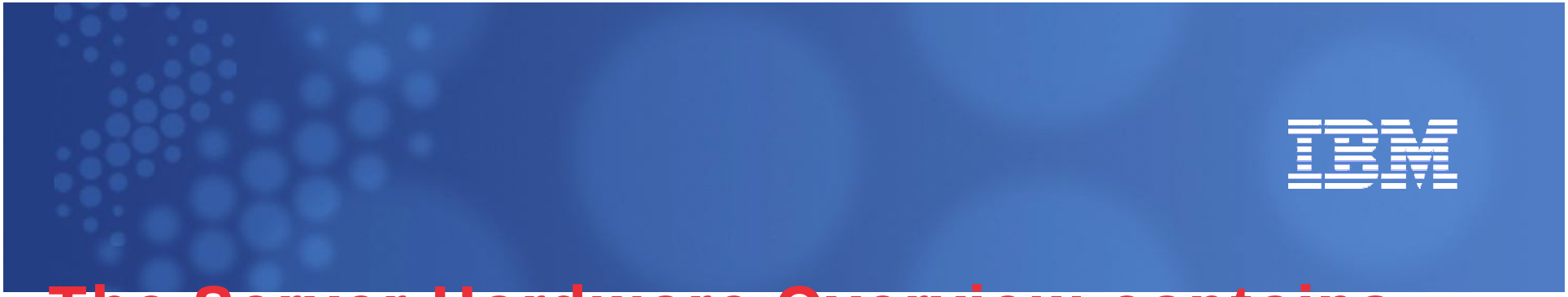
  - SELECT CAST((100 - (CAST(MAX_REDUCTION_MB AS FLOAT) * 256) / (CAST(USABLE_PAGES AS FLOAT) - CAST(USED_PAGES AS FLOAT)) * 100) AS DECIMAL(4,2)) AS PERCENT_FRAG FROM DB

- A high PERCENT_FRAG value can indicate problems. If you think your database needs a defrag, then if possible, take a copy and try that first. That will give you an indication of how much time is needed for the load.

- Use expiration and database backup scripts to monitor database performance using Operational Reporting

Agenda

- ***Introduction***
- ***The challenge…***
- ***Background***
- ***Operating Systems***
- ***Subsystems***
- ***TSM***
- ***Server Hardware Overview***
- ***TSM Server Reference Platforms***
- ***Questions***

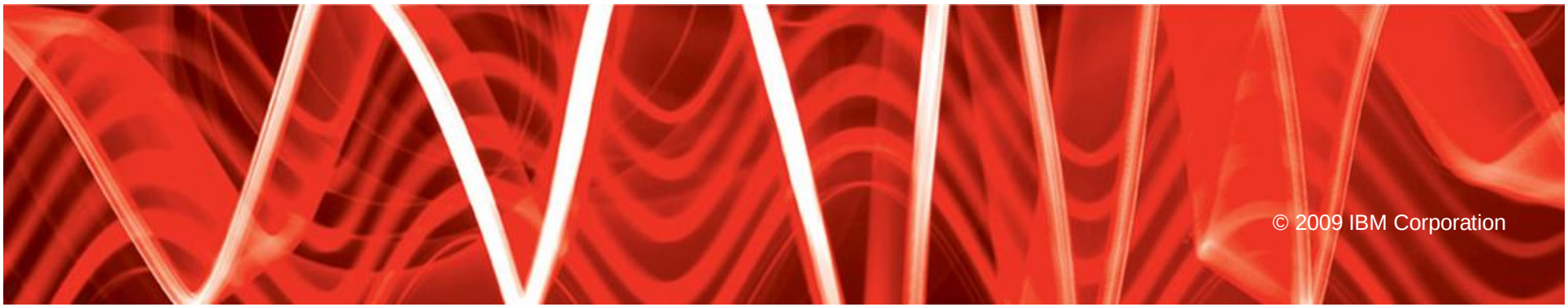


The Server Hardware Overview contains excerpts from

A Practical Approach to Hardware Selection for TSM Servers and Clients

Presented at
Software University 2006 - Session 2678

by Brandon W. Moore



Abstract

One aspect of backup and recovery (BAR) is moving mass amounts of data within a certain amount of time. A key to developing successful Tivoli Storage Manager solutions is an understanding I/O, PCI Subsystems, Disk and Tape Storage Hardware: essentially the plumbing of the backup server and backup clients. This presentation explores real world examples as a case study and utilizes technical specifications from organizations like PCI-SIG and SNIA to develop a practical approach to sizing a Tivoli Storage Manager server and client hardware.

All with one goal in mind:
To meet the customer's stated Recovery Time Objectives in a DR situation.

Why should this subject be important to you?

- Server Hardware is the most important piece of our solution
- Proper Server Hardware ensures the success of the proposed solution
- Learning practical methods of hardware selection simplifies solution development
- Enabling S.O.S.W.O.S with our server hardware counterparts

Input/Output (I/O) Subsystem consideration

What I/O subsystem technology matters.

What is PCI?

An interconnection system between a microprocessor and attached devices in which expansion slot s are spaced closely for high speed operation.

Why does PCI matter?

Input/Output (I/O) Subsystem consideration

Who enables I/O Subsystem technology?

What is the PCI-SIG?

Formed in 1992, the PCI-SIG is the industry organization chartered to develop and manage the PCI standard.

The PCI-SIG is chartered to:

- Maintain the forward compatibility of all PCI revisions or addenda
- Contribute to both the establishment of PCI as an industry-wide standard and to the technical longevity of the PCI architecture
- Maintain the PCI specification as a simple, easy-to-implement, stable technology that supports the spirit of its design

Input/Output (I/O) Subsystem consideration

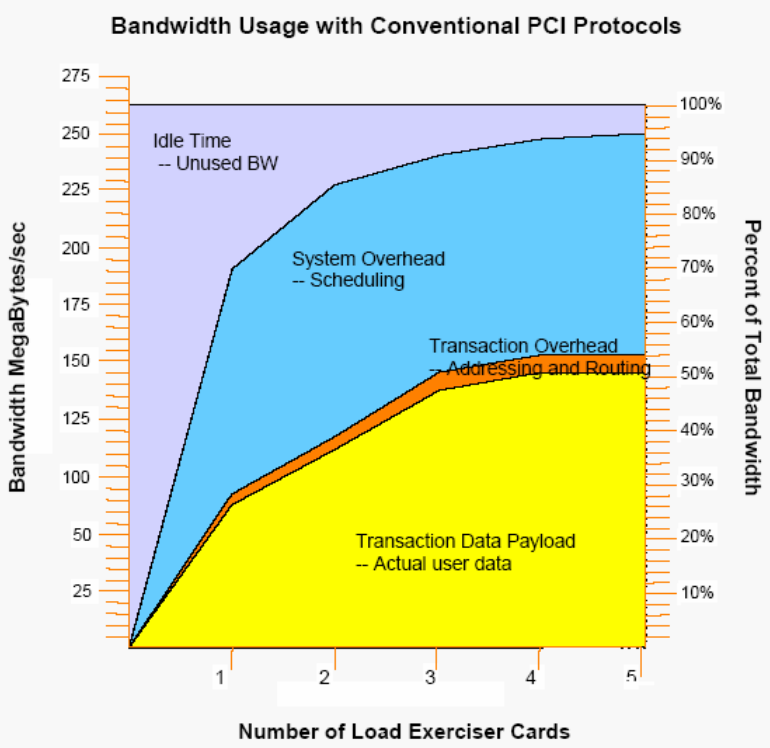
Current and near-term I/O subsystem future technology

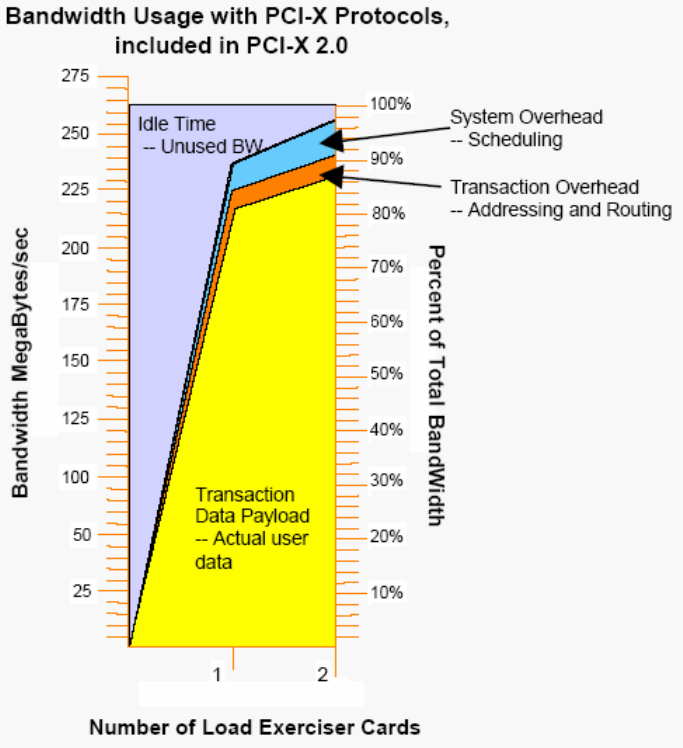
- Current Standards
  - PCI-X 2.0 64-bit, 66-533
  - PCI Express, also known as PCIe, formerly known as 3GIO
- Current Trends
  - PCI-X is still an entrenched incumbent due to backwards compatible interconnect
  - PCIe is a newly emerging form factor aimed initially at Intel platforms

Input/Output (I/O) Subsystem consideration

Current and near-term I/O subsystem future technology

- PCI and PCI-X has undergone several code changes to improve overall efficiency of data transfer.





The PCI-X protocol is more efficient than traditional PCI.

Input/Output (I/O) Subsystem consideration

Current and near-term I/O subsystem future technology

- What is PCI Express (PCIe)?
  - Serial technology providing scalable performance.
  - High bandwidth—Initially, 5-80 gigabits per second (Gbps) peak theoretical bandwidth, depending on the implementation.
  - Point-to-point link dedicated to each device, instead of the PCI shared bus.
  - Advanced features—Quality of service (QoS) via isochronous channels for guaranteed bandwidth delivery when required, advanced power management, and native hot plug/hot swap support.

Input/Output (I/O) Subsystem consideration

Current and near-term I/O subsystem future technology

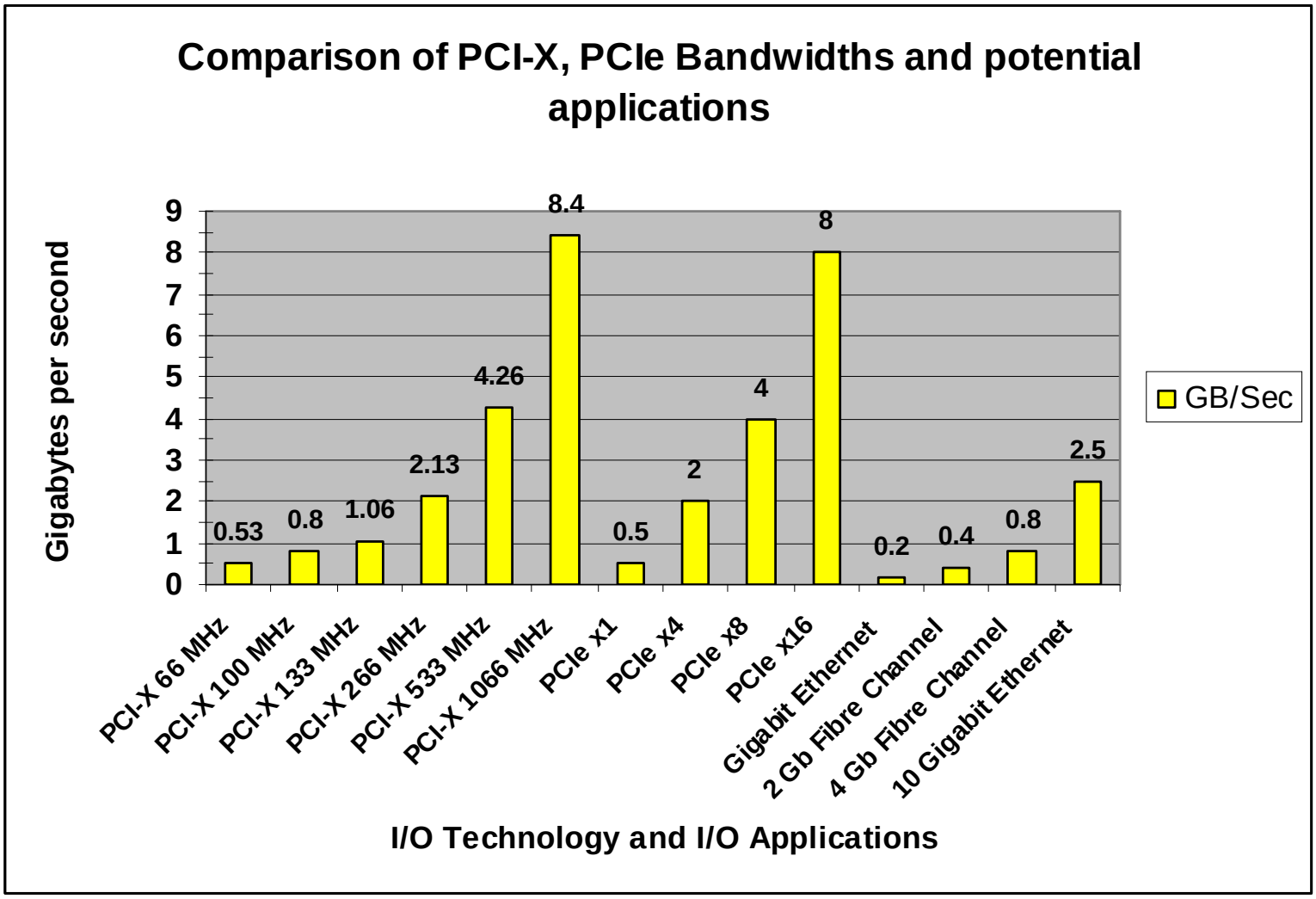
Comparison between PCI-e and PCI-X bandwidths

- PCIe
  - x1 = 0.5 GB/s
  - x4 = 2.0 GB/s
  - x8 = 4.0 GB/s
  - x16 = 8.0 GB/s
- Serial Bus technology

- PCI-X
  - PCI-X 133 MHz = 1 GB/s
  - PCI-X 266 MHz = 2.1 GB/s
  - PCI-X 533 MHz = 4.2 GB/s
  - PCI-X 1066 MHz = 8.4 GB/s*
- Parallel Bus Technology

Input/Output (I/O) Subsystem consideration

Current and near-term I/O subsystem future technology



Input/Output (I/O) Subsystem consideration

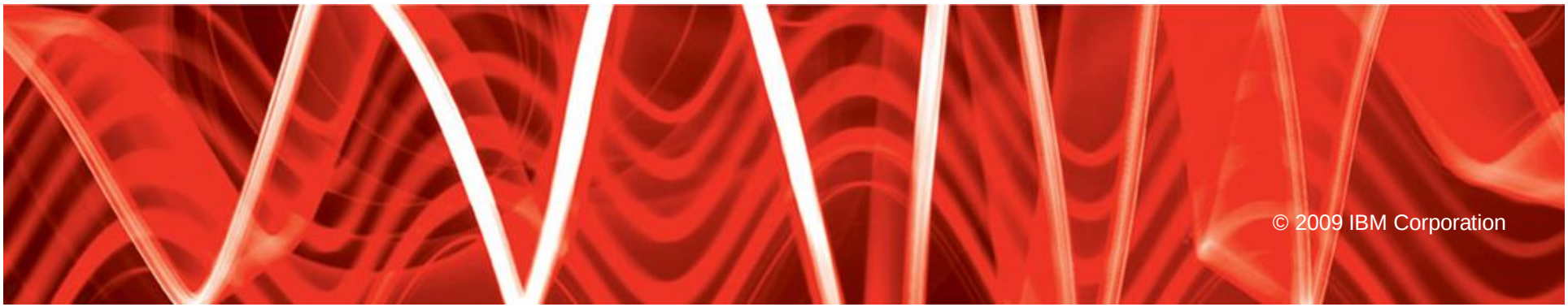
Current and near-term I/O subsystem future technology

Current design standards for servers/workstations using PCIe

- x16 lanes are reserved for primarily graphics support to replace AGP(1X-8X) Graphics connections
- x1 is utilized for general purpose I/O and be satisfactory for a single gigabit Ethernet card.
- x4 is used for a single or dual 2 or 4 Gb FC HBA connection or dual-port gigabit Ethernet card

TSM Server Reference Platforms

POWER and System x



Let's simplify this whole process

- I have taken out the guesswork TSM servers properly
- Etherchannel is needed to increase throughput to your TSM Servers
- If the customer does not support etherchannel on their LAN networks, then you have other things to consider
- Each of the these platforms have enough NIC bandwidth on board without using a PCIe or PCI-X slot
- Connections for disk should be to the PCI-X HBAs, where applicable

Summary

Best price per performance for TSM server hardware

- AIX
  - Approximately 1000 gigabytes per hour via the LAN (4-port gigabit etherchannel)
  - Quad Core 4.2 GHz POWER 520
  - The POWER 550 and POWER 520 main CECs are identical
  - QC4.2POWER520 is less than QC3.5POWER550
- Windows
  - Approximately 500 gigabytes per hour via the LAN (2-port gigabit etherchannel)
  - Two Dual Core 3.5 GHz System x3650
  - Linux or Windows
- Other vendors
  - Those servers will be available by request
  - The reference server hardware list will be refreshed twice a year

eConfig List Pricing

Use these list price configuration files to impress, influence and motivate your hardware sales brethren. They are meant to be a guide for them when you need to sell your TSM solution. For now send me an email to request the eConfigs.

POWER 520 Configurations

4-core 4.2 GHz POWER520

4-core 4.2 GHz)WER520 w/ 4Gb HI

System x3650 Configurations

Two Dual-Core 3.5 GHz Systemx3650

Two Dual-Core 3.5 z Systemx3650 w/ 4

POWER 520 Configurations

Special Consideration Hardware Platforms
will be covered in a later presentation and/or discussions

- VMware Consolidated Backup
- SAP
- PeopleSoft
- Clustered Databases
- Data Warehouse
- Large Fileservers and Shares
- System x3850 versus x3950
- POWER 550 versus 570 versus 595

Thanks and References

- Steve Lipton, Stephanie Woods, Andrew LaMothe
- Doffie Benjamin, Daniel Thompson, Michael Barton, Geoff Cecil, Charlie Fitzgerald, Colleen Gobbi
- Charles Fisher, Daniel Crouse, George Blackwood, Larry Coyne, Toby Marek
- Larry Mills, Ken Scott, Jerry Brown, Araceli Loya, Todd Virnoche, Scott Sullivan, Ken Keefer and Susan Bowlin

- Herman Dierks, IBM AIX TCP/IP Performance

- PCI-SIG, http://www.pci-sig.com
- IBM Power™ Systems Family Quick Reference Guide, POY03001USEN.pdf
- Planning and Installing the IBM eServer X3 Architecture Servers, SG24-6797-00
- xSeries Performance: http://perform.raleigh.ibm.com:8090/

Trugarez
Breton

Merci
French

Спасибо
Russian

Gracias
Spanish

شكرا
Arabic

감사합니다
Korean

תודה רבה
Hebrew

धन्यवाद
Hindi

谢谢你
Chinese

Tack så mycket
Swedish

Obrigado
Brazilian Portuguese

Thank You
English

Tak
Danish

Grazie
Italian

Dankon
Esperanto

Dank u
Dutch

Danke
German

நன்றி
Tamil

ありがとうございました
Japanese

ขอบคุณ
Thai

go raibh maith agat
Gaelic

Dekujeme Vam
Czech

The following terms are trademarks or registered trademarks of the IBM Corporation in either the United States, other countries or both.

- AIX
- AIX 5L
- BladeCenter
- Chipkill
- DB2
- DB2 Universal Database
- DFSMSdss
- DFSMShsm
- DFSMSrmm
- Domino
- e-business logo
- Enterprise Storage Server
- ESCON
- eServer
- FICON
- FlashCopy
- GDPS
- Geographically Dispersed Parallel Sysplex
- HiperSockets
- i5/OS
- IBM
- IBM eServer
- IBM logo
- iSeries
- Lotus
- ON (button device)
- On demand business
- OnForever
- OpenPower
- OS/390
- OS/400
- Parallel Sysplex
- POWER
- POWER5
- Predictive Failure Analysis
- pSeries
- S/390
- Seascape
- ServerProven
- System z9
- System p5
- Tivoli
- TotalStorage
- TotalStorage Proven
- TPF
- Virtualization Engine
- X-Architecture
- xSeries
- z/OS
- z/VM
- zSeries

Linear Tape-Open, LTO, LTO Logo, Ultrium logo, Ultrium 2 Logo and Ultrium 3 logo are trademarks in the United States and other countries of Certance, Hewlett-Packard, and IBM.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

Intel, Intel Inside (logos), MMX and Pentium are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.