Linux for zSeries Performance Update

Klaus Bergmann

02/24/2003, Session # 2559

Trademarks

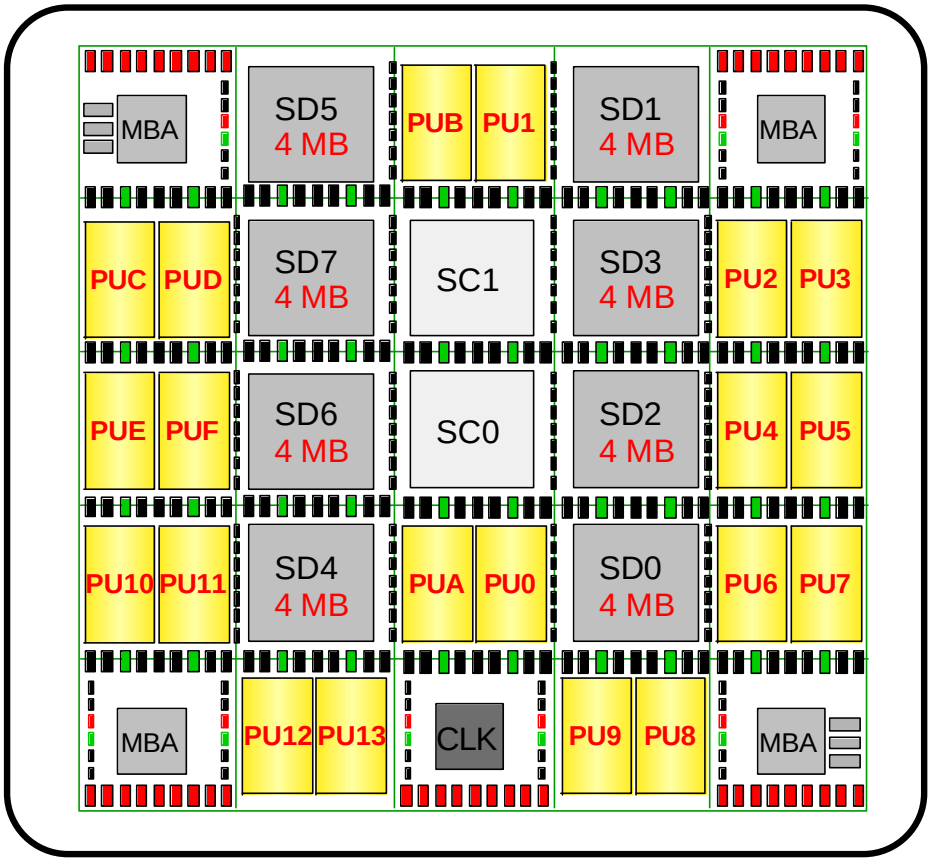
The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.
Enterprise Storage Server
ESCON*
FICON
FICON Express
HiperSockets
IBM*
IBM logo*
IBM eServer
Netfinity*
S/390*
VM/ESA*
WebSphere*
z/VM
zSeries
* Registered trademarks of IBM Corporation
The following are trademarks or registered trademarks of other companies.
Intel is a trademark of the Intel Corporation in the United States and other countries.
Java and all Java-related trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc., in the United States and other countries.
Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.
Linux is a registered trademark of Linus Torvalds.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
Penguin (Tux) compliments of Larry Ewing.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* All other products may be trademarks or registered trademarks of their respective companies.
I

Agenda

- Hardware
- Scalability
- Networking
- Disk I/O
  - FCP / SCSI
  - ESS
  - LVM

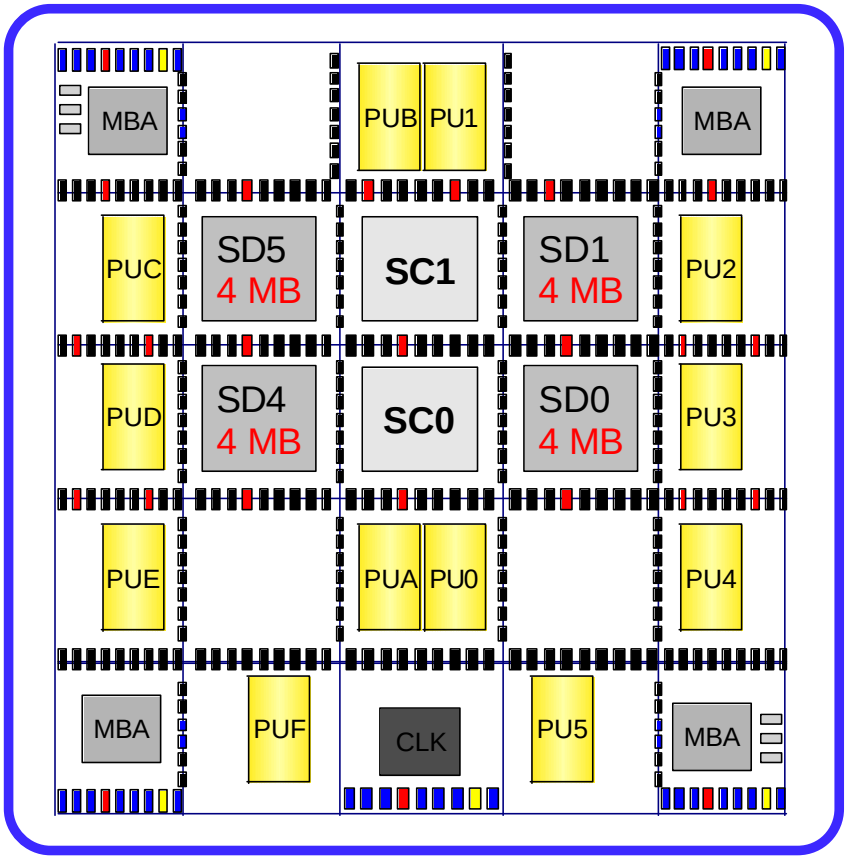
All measurement results are
based on IBM internal benchmarks in a controlled environment and results may vary.

20-PU Module 110-116, 1C1-1C9 210-216, 2C1-2C9



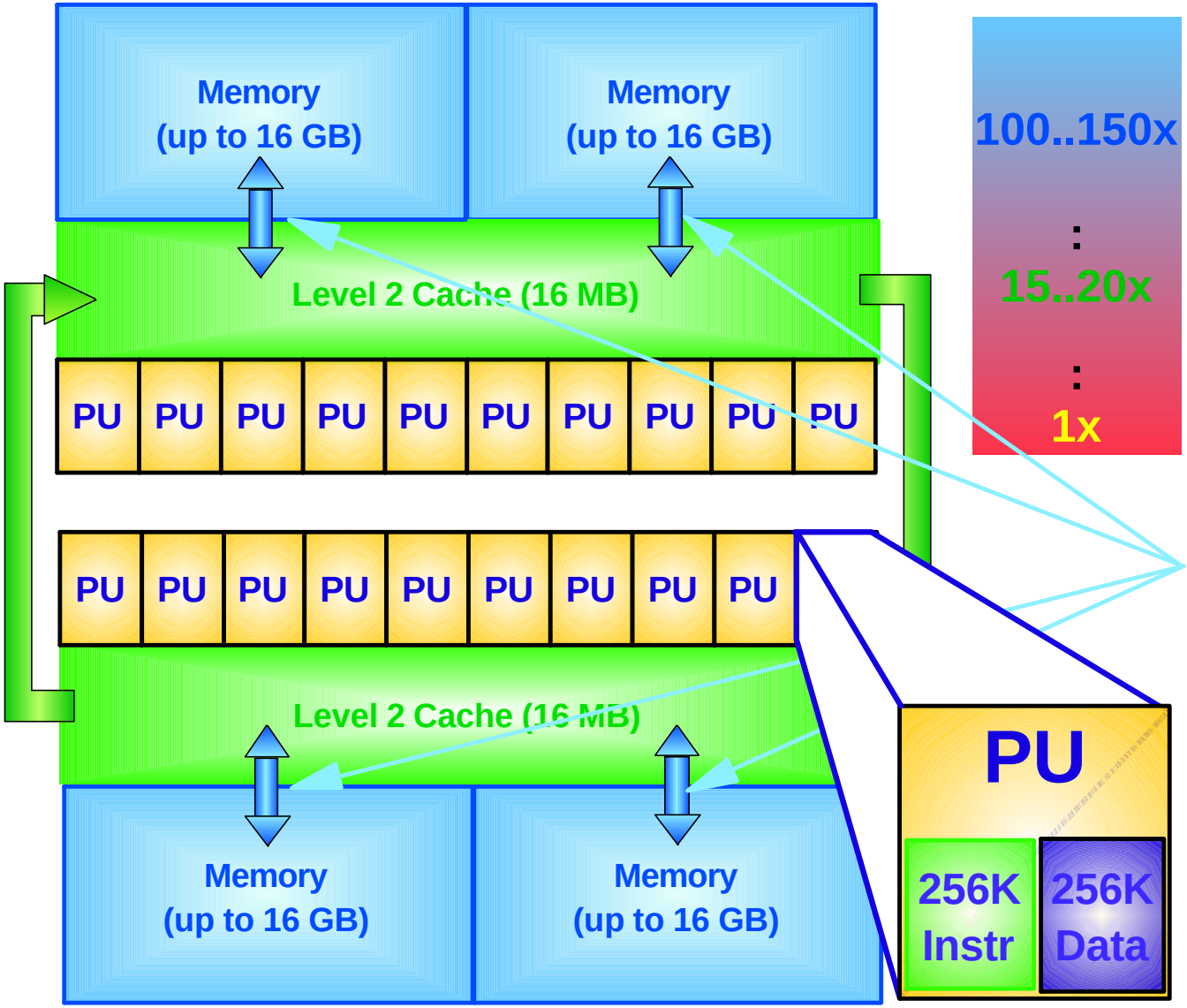
- 20 PU Chips @ 1.3 / 1.09 ns
- 3 SAP's, 1 spare
- up to 16 CP's
- up to 8 ICF's/IFL's

12-PU Module 101..109, 100 (CF)



- 12 PU Chips @ 1.3 ns
- 2 SAP's, 1 spare
- up to 9 CP's
- up to 8(9) ICF's/IFL's

z900 Systemstructure:
Optimized for maximum internal bandwidth



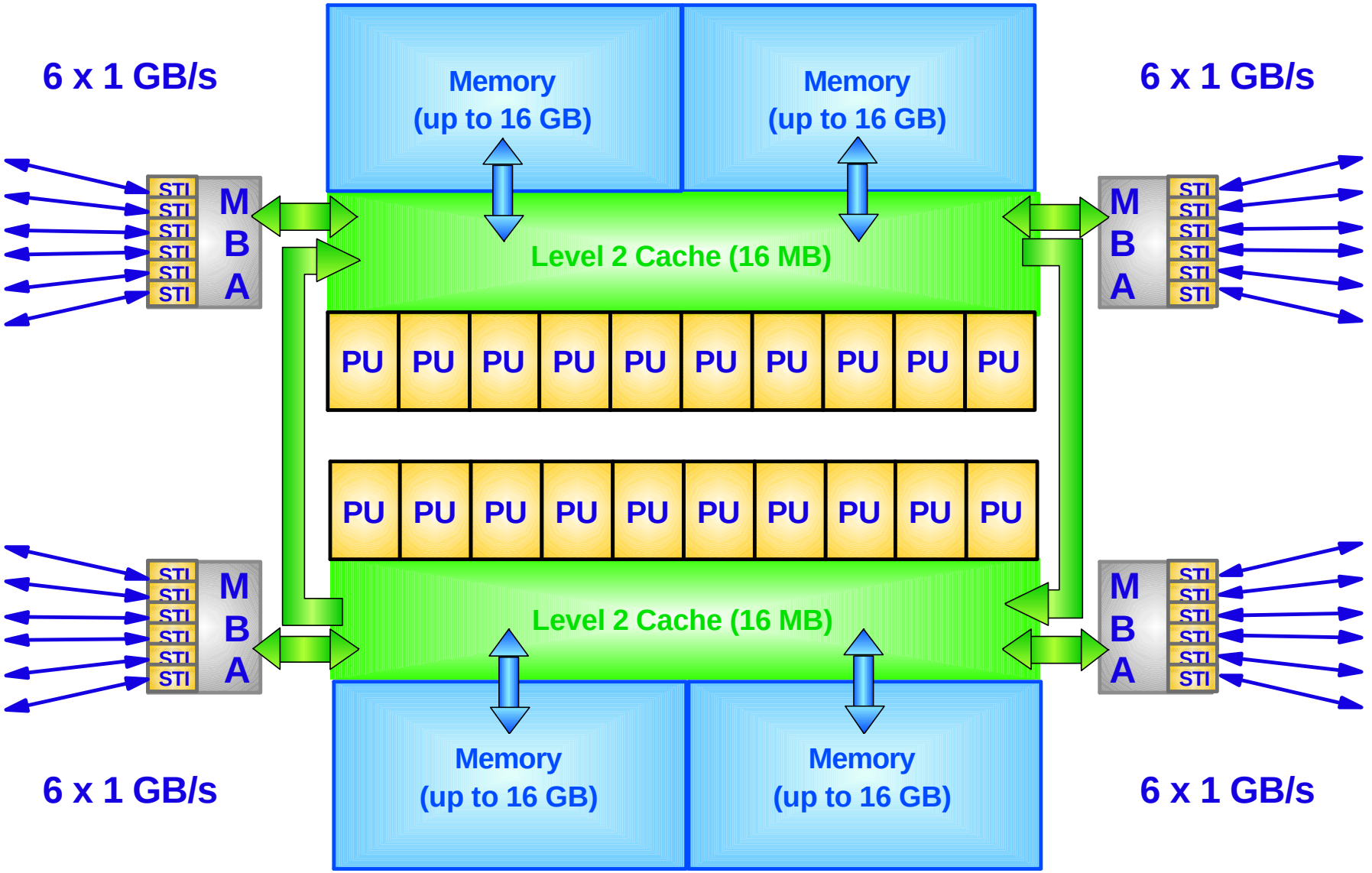
The problem

- Memory access does not scale with CPU-cycle

Solutions

- High bandwidths (=> throughput):
  4 x 16 Byte @ 2.18 ns
  ==> 28 GB/sec
- Cache Hierarchies (latencies)
  - Level 1 Caches on CPU-Chip
  - Level 2 Cache 'shared by 10'

z900 Systemstructure:
Optimized for maximum external bandwidth



New functions since 04/2002 :

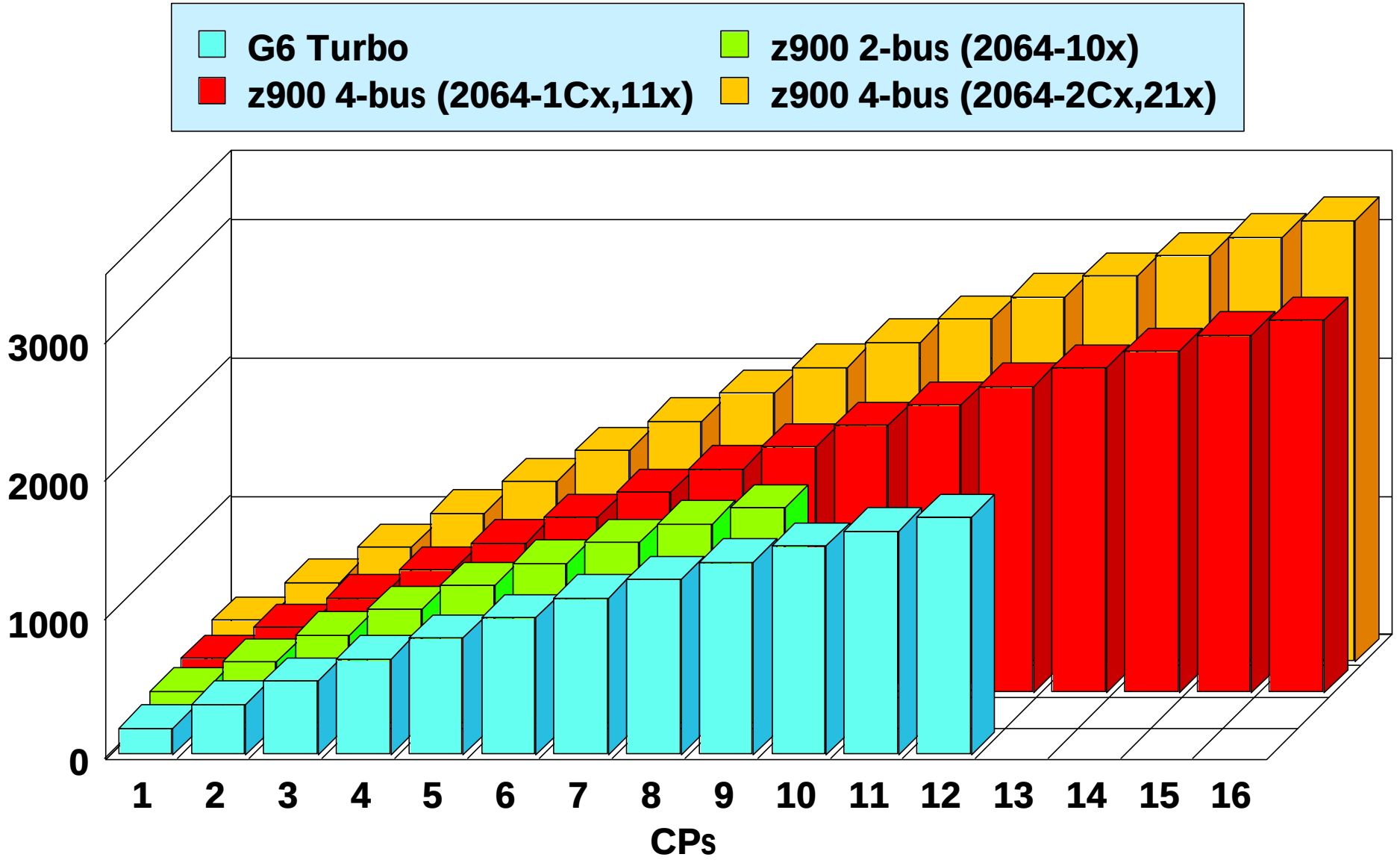
Additional z900-functionalities

- Support for FCP and SCSI on FICON Feature in Linux environments
  - LA since 6/2002, GA 1Q 2003 (with new distributions)
- CIU: Customer Initiated Upgrade
  - Web-Interface for processor or memory upgrade
  - LIC/microcode download and upgrade via RSF
- OSA-Express(QDIO):
  - IPv6, VLAN, SNMP,...

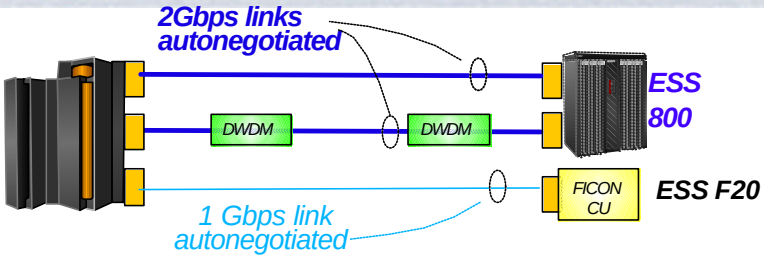
z900-'Turbo' @ 1.09 nsec

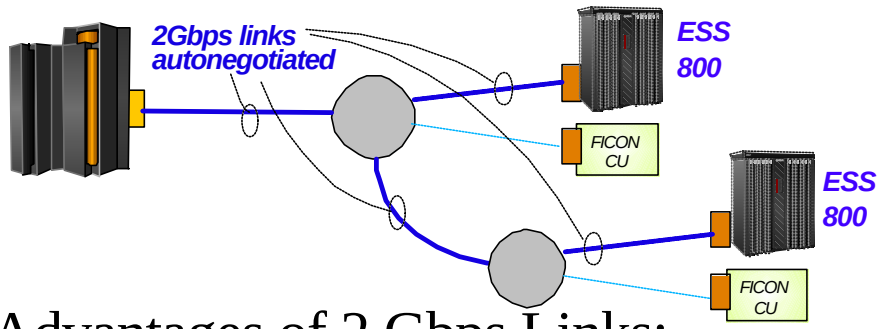
- 16 additional models
  2C1..2C9,210..216
- ..20..% faster systems

Relative Performance z900 vs G6 (2064-1C1 = 250)



FICON Express Connectivity Options - 2 Gbps Links





- Direct-attachment
  - IBM Enterprise Storage Server 800
- DWDM and optical amplifiers
  - Cisco ONS 15540 ESP (LX, SX) and optical amplifiers (LX, SX)
  - Nortel Optera Metro 5200 and 5300E* and optical amplifiers*
  - IBM 2029 Fiber Saver*
- FICON Switched and Cascaded Connectivity
  - McDATA ED-6064 (2032-064)
  - INRANGE FC/9000-001/-128 (2042)

- Advantages of 2 Gbps Links:
  - Higher throughput, improved performance
  - Extended opportunities for channel-consolidation
- Prerequisite: FICON Express cards (FC 2319, 2320)
- Native FICON, FICON CTC, FICON Cascaded Directors, Fibre Channel
- Link speeds negotiated between server and device
  - transparent for application and user

Our Hardware for Measurements

2064-216 (z900)

- 1.09ns (917MHz)
- 2 * 16 MB L2 Cache (shared)
- 64 GB
- LPAR
- ESCON
- FICON
- HiperSockets
- OSA Express GbE

2105-F20 (Shark)

- 384 MB NVS
- 16 GB Cache
- 128 * 36 GB disks
- 10.000 RPM
- FCP (1 Gbps)
- FICON (1 Gbps)

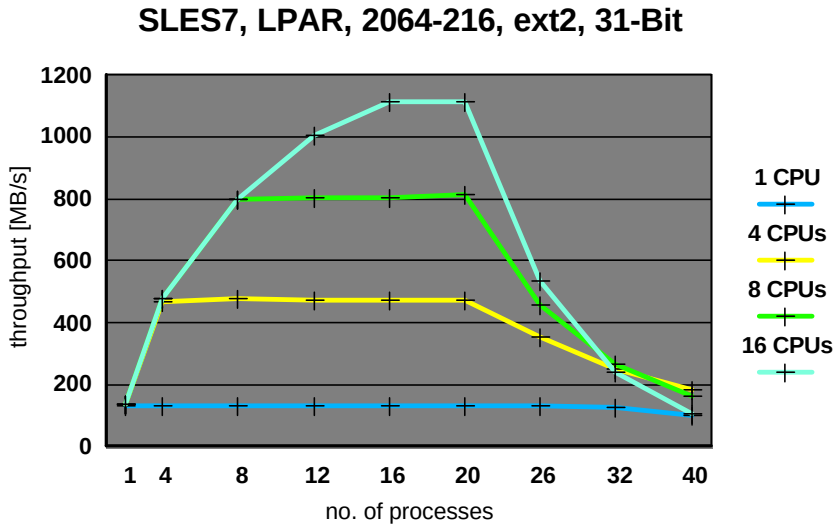
8687-3RX (8-way X440)

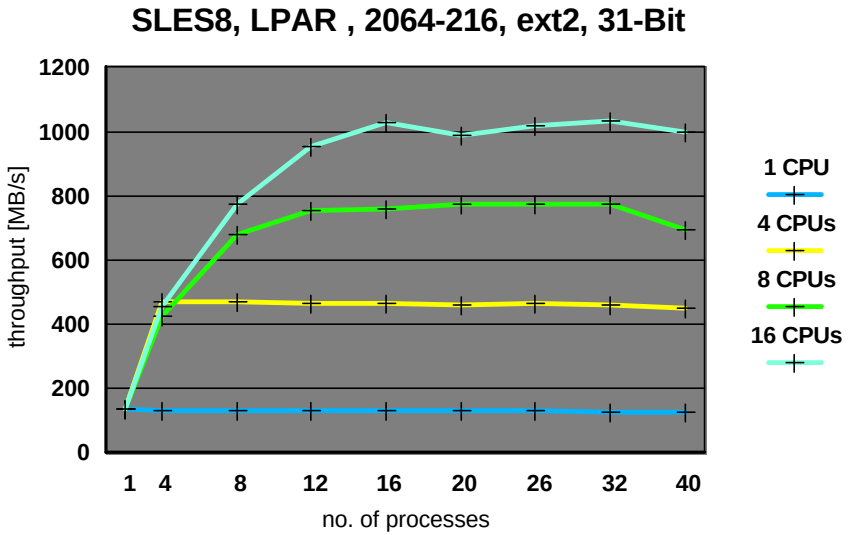
- 8-way Intel Pentium 3 Xeon
- 1.6 GHz
- 8 * 512K L2 Cache (private)
- hyperthreading
- summit chipset

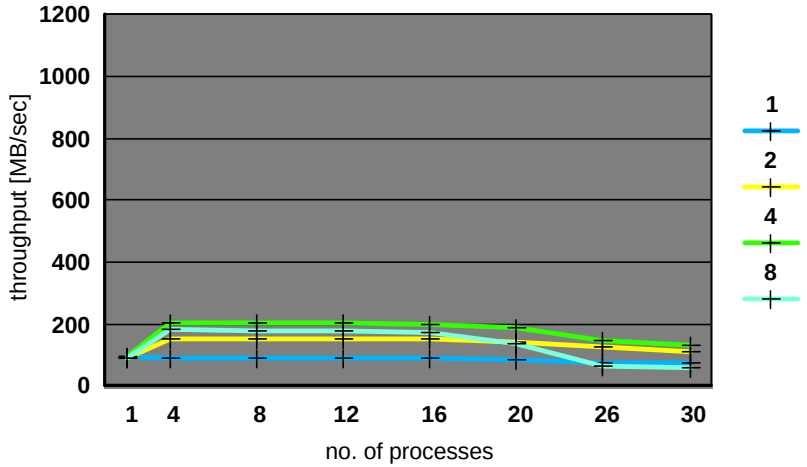
SuSE SLES7 versus SuSE SLES8

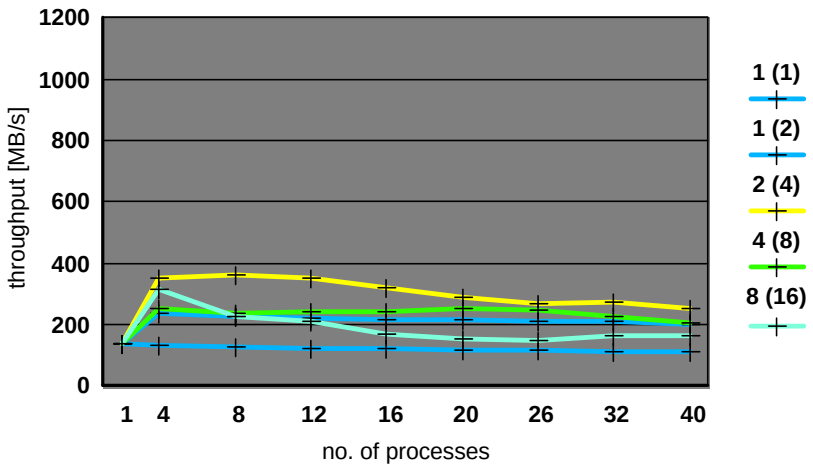
- From Kernel version 2.4.7 / 2.4.17 to version 2.4.19
- From glibc version 2.2.4-31 to version 2.2.5-84
- From gcc version 2.95.3 to version 3.2-31
- Huge number of United Linux patches
  1.3 MLOC (including x,p,i changes)
  New Linux scheduler
  Async I/O
  ...

Dbench File I/O - Scalability





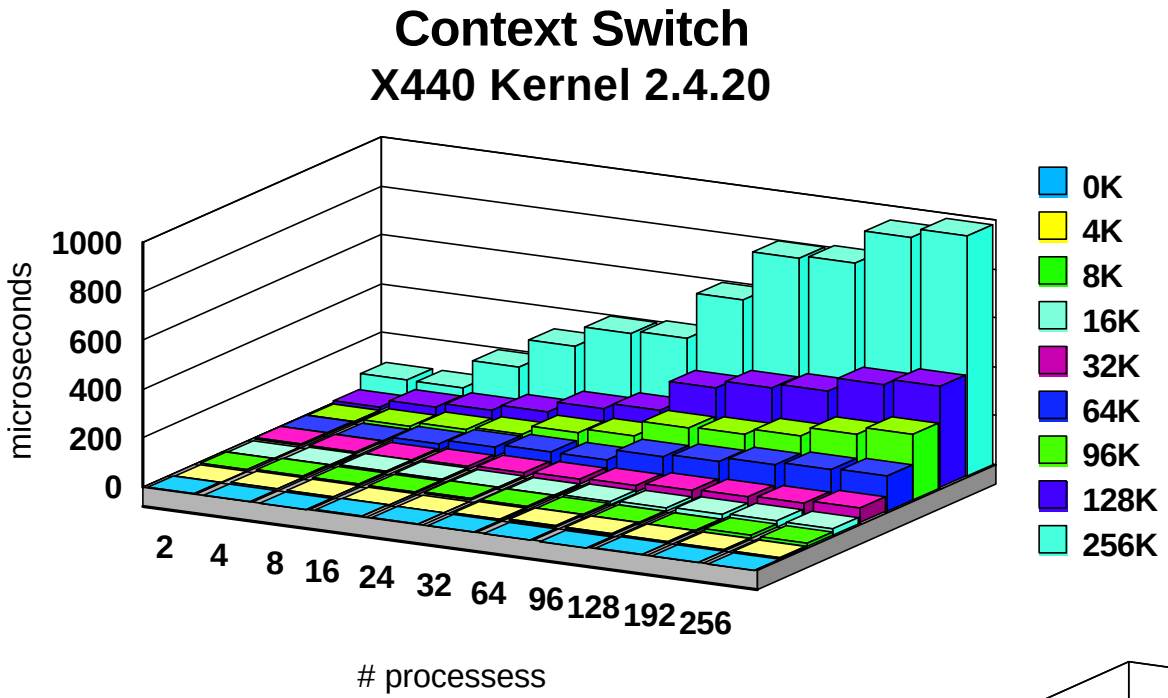


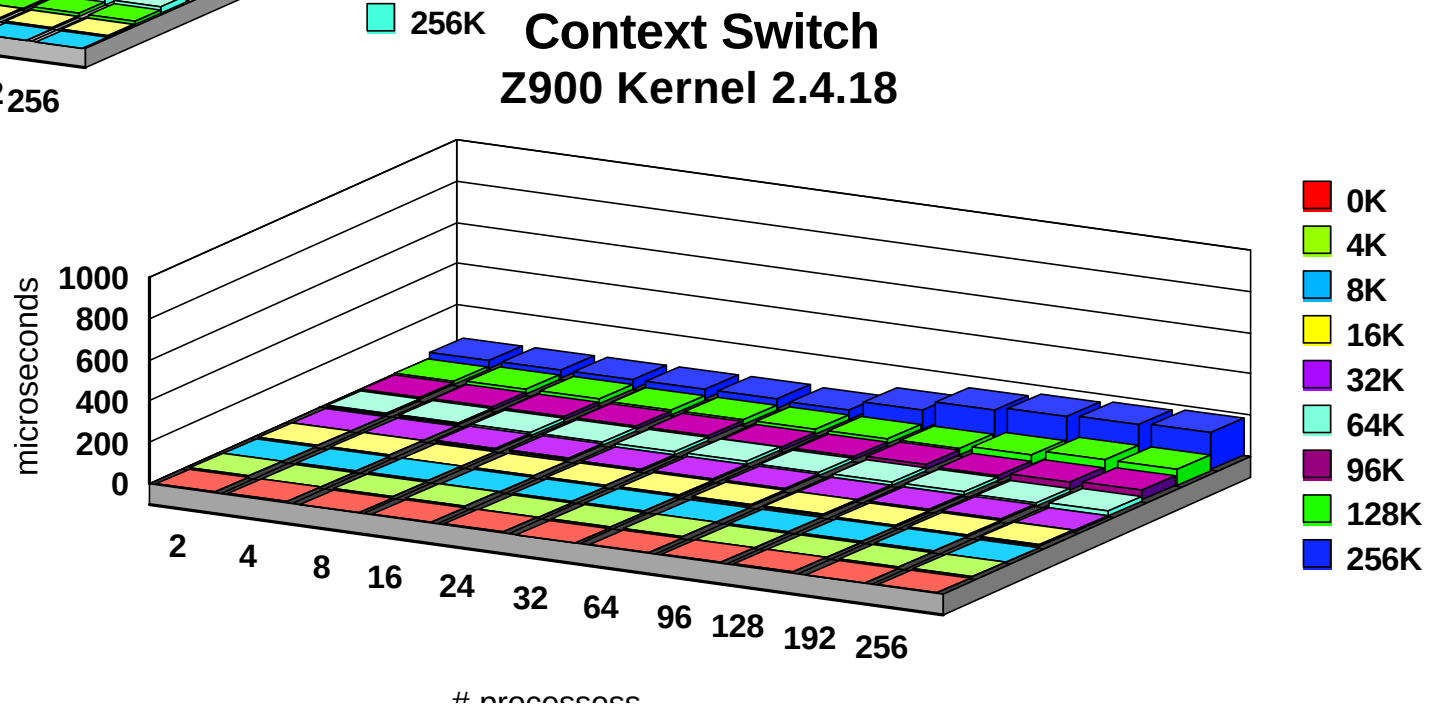


- 8-way Intel Pentium III, 700 MHz

- 8-way Intel Pentium III Xeon, 1.6 GHz
- hyperthreading, summit chipset
- values in () are Linux maxcpus parameter

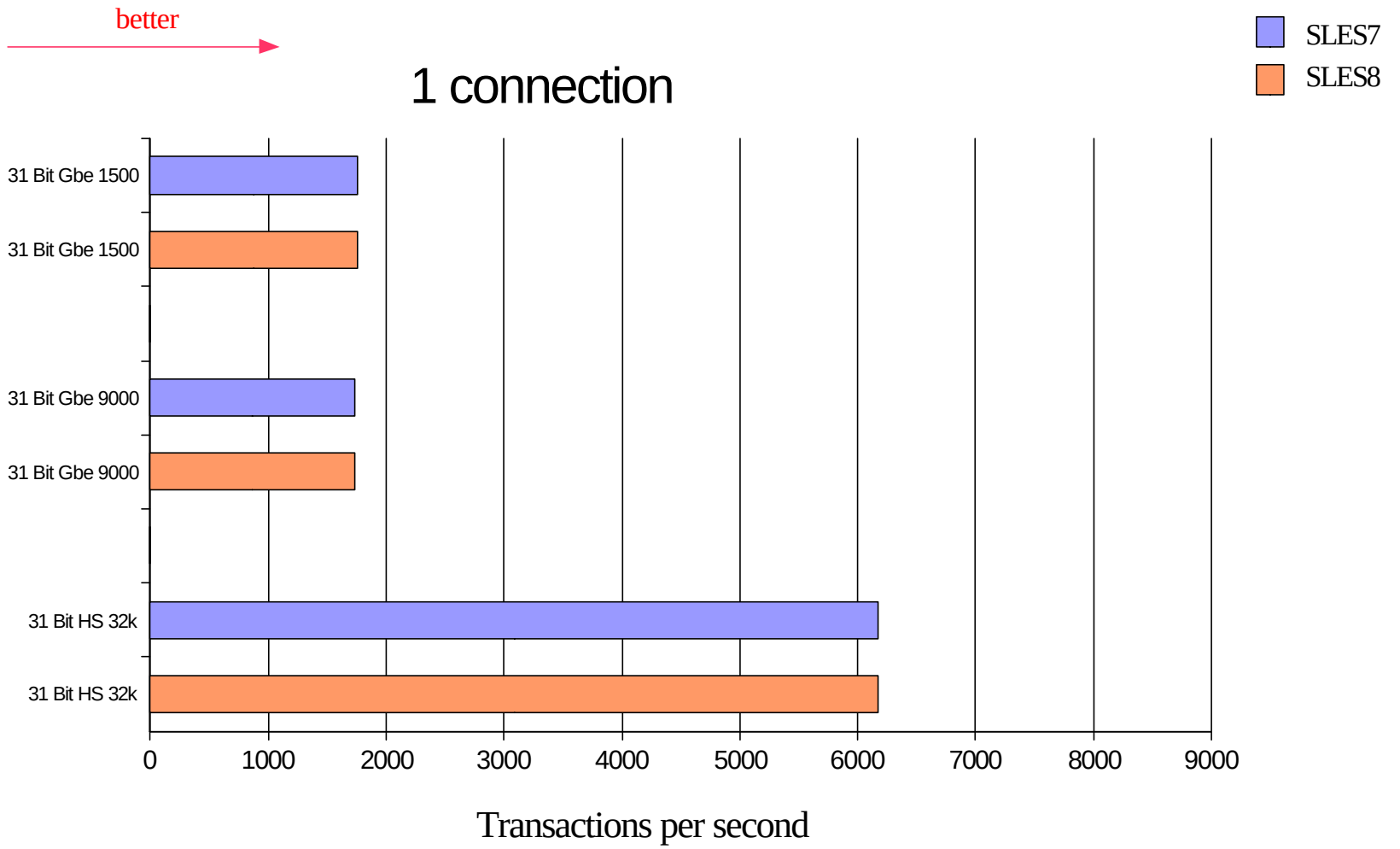
Context Switching



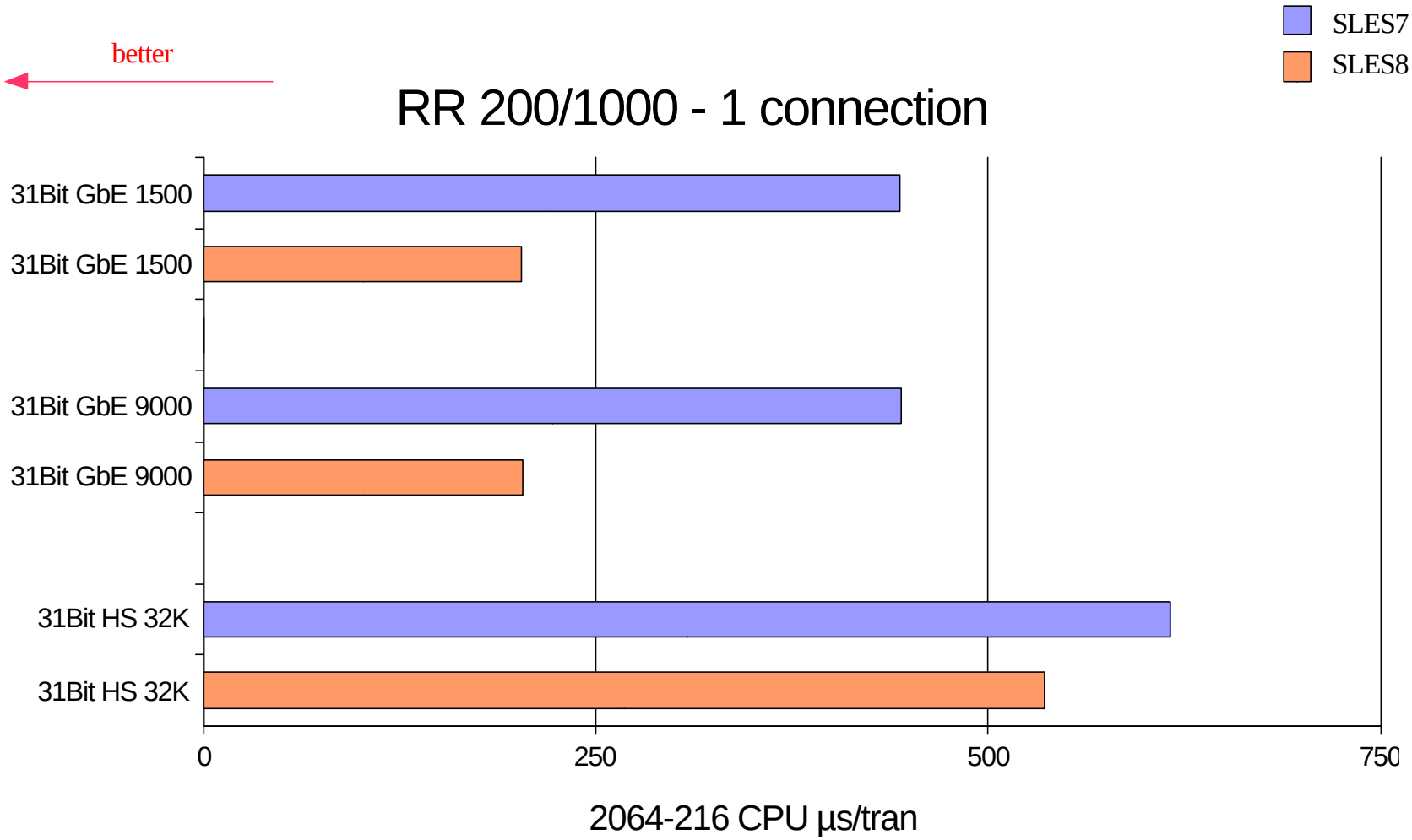


Netmarks Results

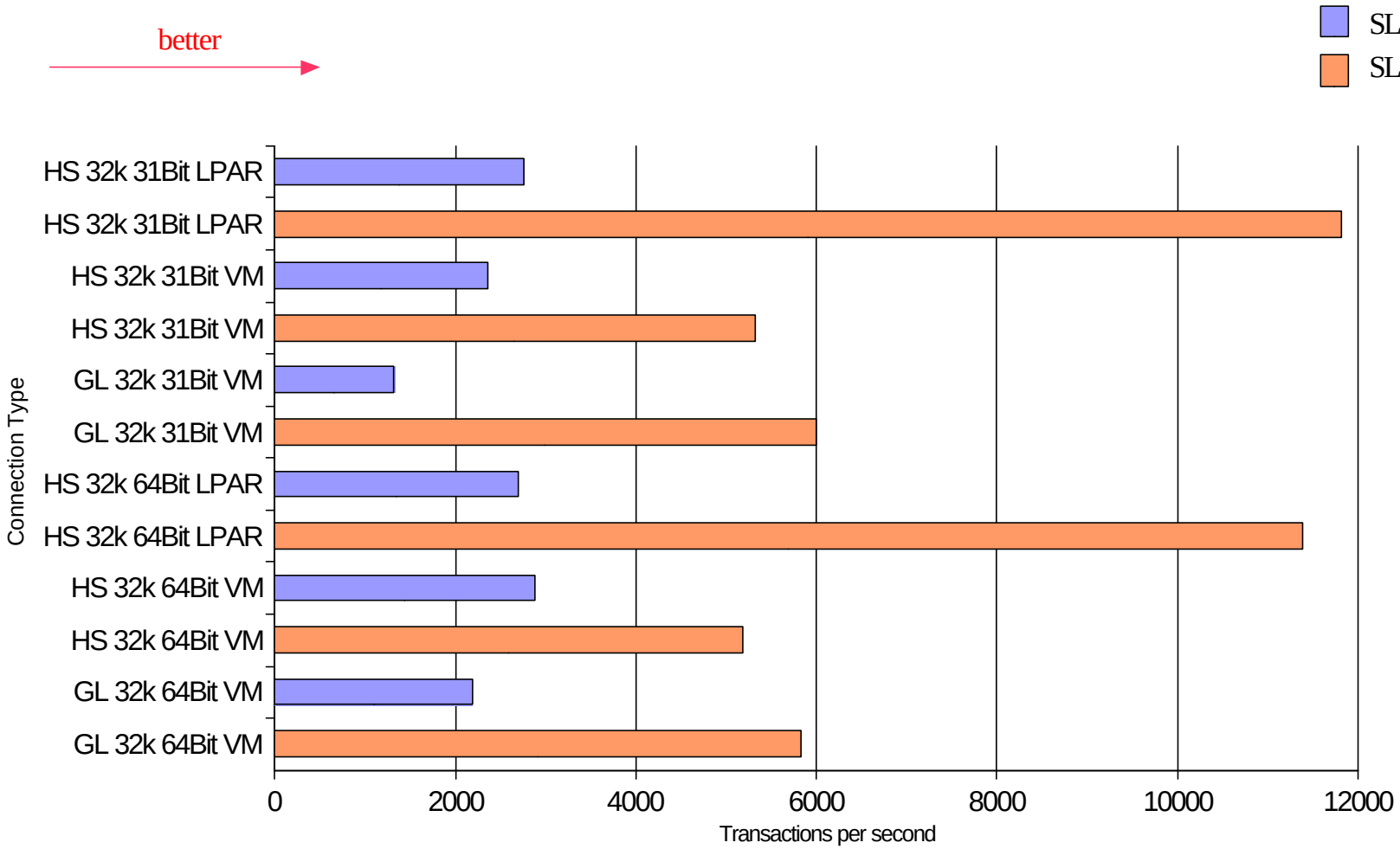
Transaction workload - RR 200/1000



Network cost VM - transaction workload



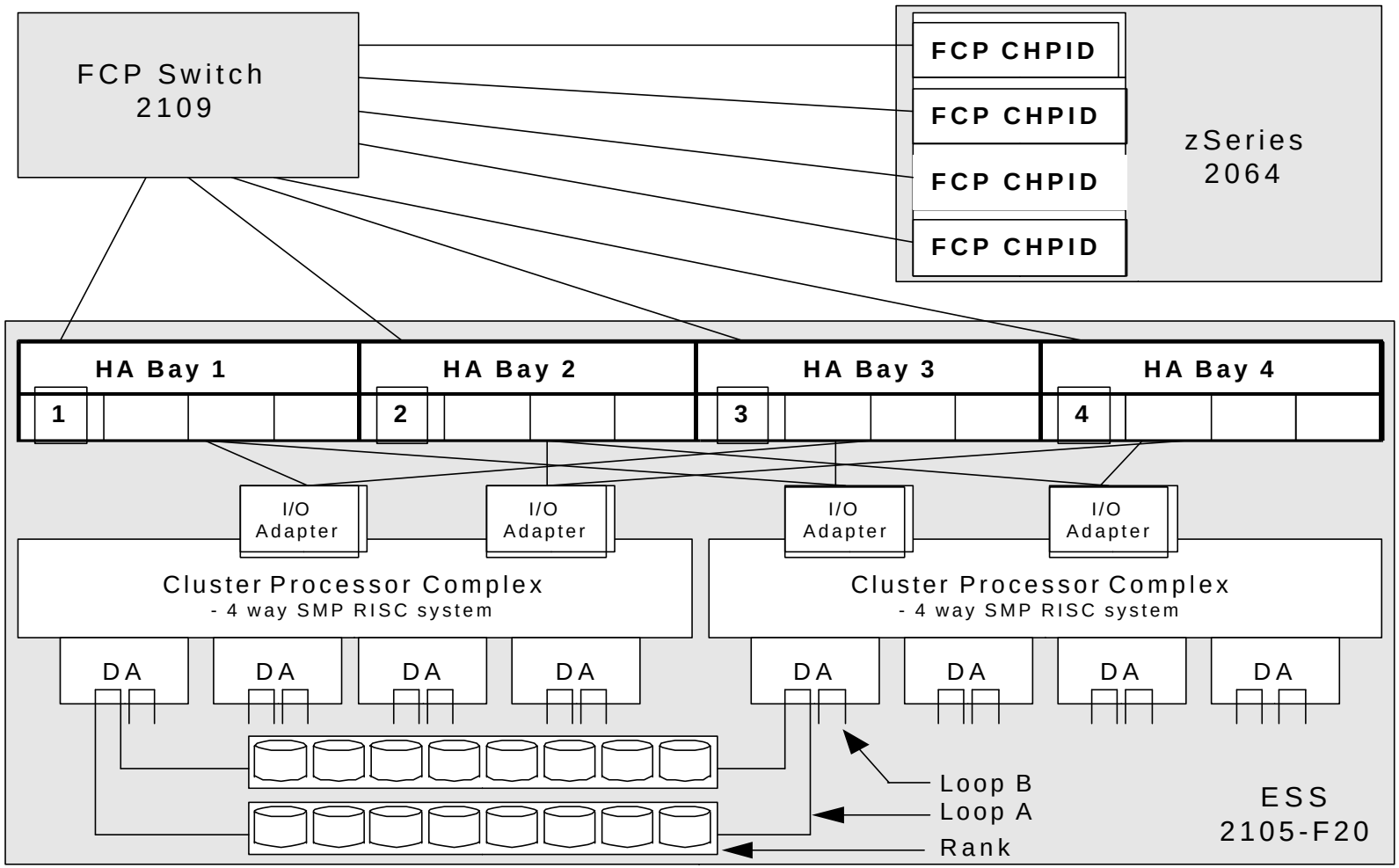
Connect Request Response workload (CRR 64/8k-10 connections)



Disk I/O

- Don't treat ESS as a black box, understand its structure
- The default is close to worst case:
  You ask for 16 disks and your SysAdmin gives you addresses 5100-510F
- What's wrong with that?

ESS Architecture



CHPIDs

Host Adapters
16 HA in 4 bays

Device Adapter Pairs
supporting 2 loops

Disk ranks, each rank is one RAID-5 array

Scaling Tests

Scenario	CHPIDs	Host Adapt.	Ranks	Disks	bottleneck
single disk	1	1	1	1	1 HA
single rank	1	1	1	8	1 HA
single HA	1	1	4	8	1 HA
single CHPID	1	4	4	8	1 CHPID
two CHPIDs	2	4	4	8	4 HA
Max. Avail.	4	4	4	8	4 HA

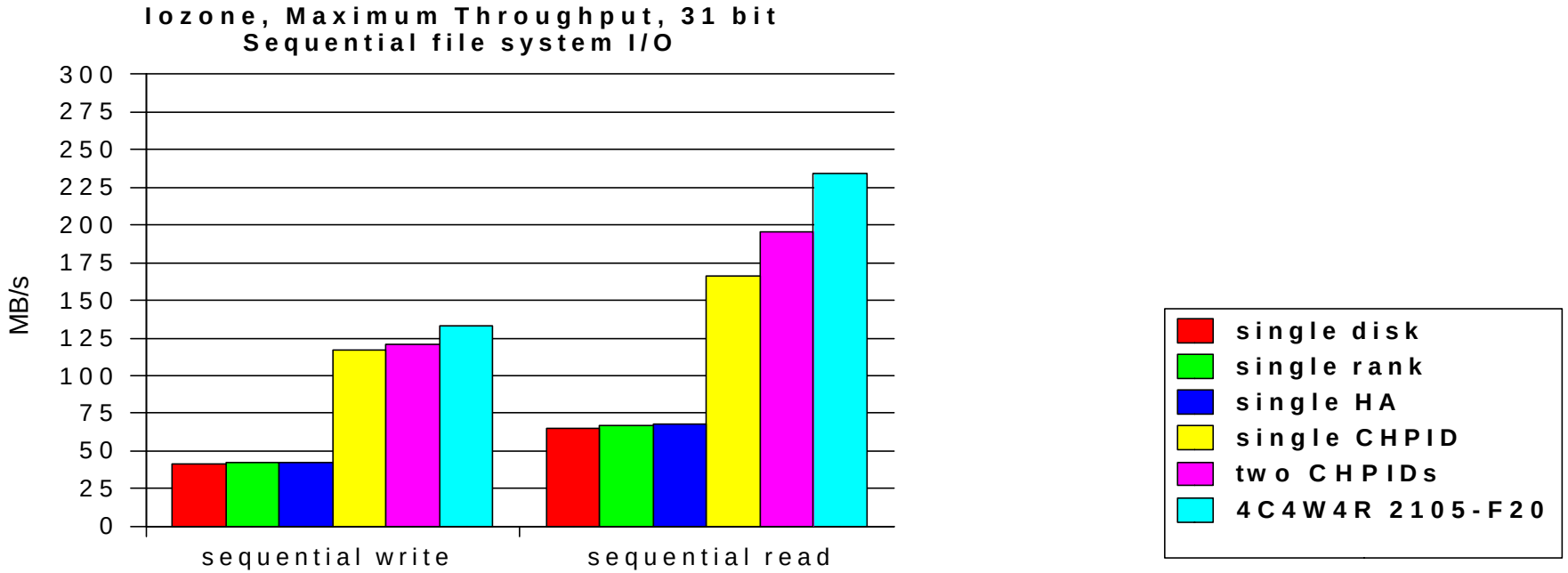
Benchmark used for measuring: Iozone (http://www.iozone.org)

- multi process sequential file system I/O, scaling 1-8/16 processes
- each process writes and reads a 350 MB file on a separate disk
- System: LPAR, 4 CPUs, 128 MB main memory, linux 2.4.17

Hardware Setup

- 2064-216, 917 Mhz, 256MB LPAR
- 4 FICON Express channels used for FCP (IOCDS: type FCP)
- 6 FICON Express Channels used for FICON (IOCDS: type FC)
- 2109-F16 FCP switch
- ESS 2105-F20:
  - 16GB cache, 4 FCP host adapters, 6 FICON host adapters
  - 4 device adapter pairs
  - only A-loops contain disks (36.4 GB, 10,000 RPM):
    - 4 ranks for FB (fixed block) disks used for FCP
    - 8 ranks ECKD disks used for FICON measurements

Results for sequential file system I/O



- 1 HA limits to 40MB/s write and 65 MB/s read, regardless of the number of ranks
- 4 HA are limiting to 125 MB/s write and 240 MB/s read,but 4 CHPIDs are required to make use of
- 31 bit and 64 bit difference is small
- it is expected that the values further increase using more ranks, HA, CHPIDs

Locating disks

- use the tool 'ESS Specialist' to generate a html report of all disks
  (Storage Allocation -> Tabular View -> print table)

- for ECKD disk it looks like:

Volume	Location	LSS	Volume Type	Size	Storage Type	Host Port	Host Nicknames
SSID: 0x5400	LCU: 0x000 Device ID: 0x00	0X5400	3390-3	002.87 GB	ESCON/ FICON	Device Adapter Pair 1 Cluster 1, Loop A Array 2, Vol 000	n/a

logical control unit address (IOCDS: CUADD)
unit address (IOCDS: UNITADD)
ECKD disk
rank information

- for a FCP disk it looks like

Volume	Location	LSS	Volume Type	Size	Storage Type	Host Port	Host Nicknames
400-17403	Device Adapter Pair 3 Cluster 1, Loop A Array 4, Vol 000	0x14	Open System	010.0 GB	RAID Array	Fibre Channel ID 00, LUN 5400	FR39A, FR39B, FR39C, FR39D

rank information
FCP disk
16 bit LUN *0x1000000000000=64 bit FCP LUN

FCP disks are not defined in IOCDS!

Sample Setup

- CHPIDs: C1, C2, C3, C4
- Host Adapters: HA1, HA2, HA3, HA4
- device numbers / LUNs:
  - rank1: 5000, 5001, 5002, 5003, ...
  - rank2: 5400, 5401, 5402, 5403, ...
  - rank3: 5800, 5801, 5802, 5803, ...
  - rank4: 5C00, 5C01, 5C02, 5C03, ...
- For FCP use paths:
  - C1 -> HA1 -> rank1
  - C2 -> HA2 -> rank2
  - C3 -> HA3 -> rank3
  - C4 -> HA4 -> rank4
- FICON: Define a path from each CHPID to each LCU

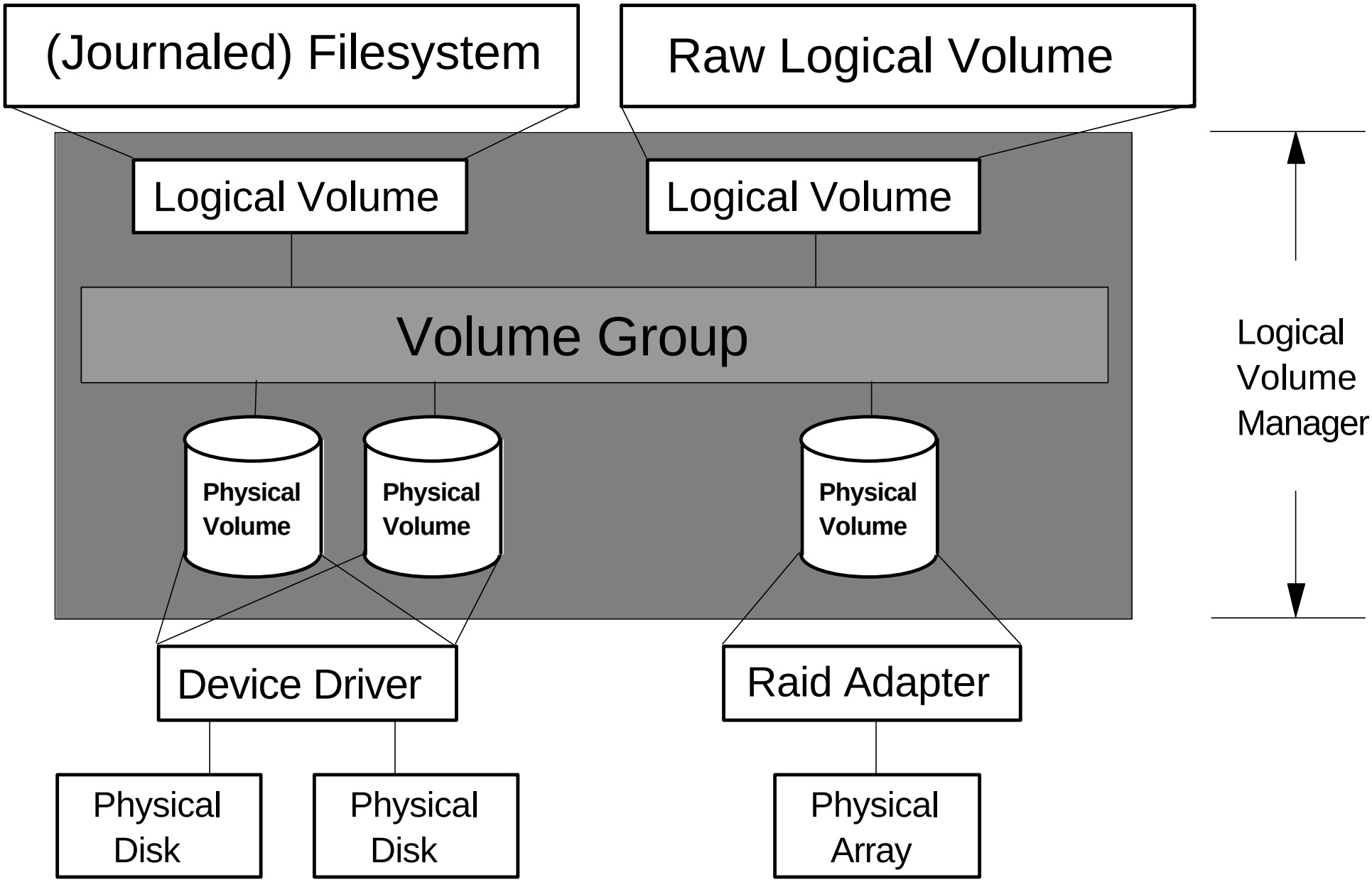
Sample Setup

- LVM:
  - add the disks to one VG in the following order:
    5000, 5400, 5800, 5C00, 5001, 5401, 5801, 5C01, ...
  - Use striping (32 KB or 64 KB)!
- For Linux VM guests:
  - guest1: 5000, 5400, 5800, 5C00, ...
  - guest2: 5001, 5401, 5801, 5C01, ...
  - guest3: 5003, 5403, 5803, 5C03, ...
  - guest4: 5004, 5404, 5804, 5C04, ...

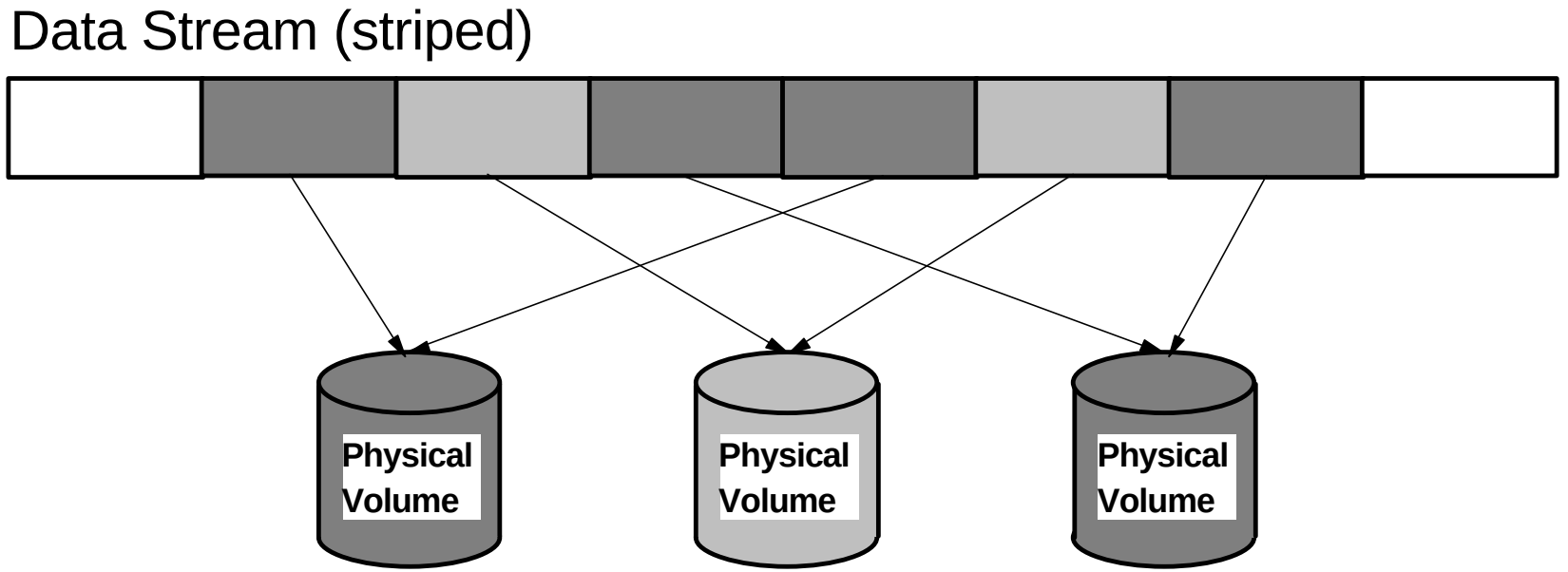
LVM: Logical Volume Manager

- What is LVM?
  - LVM builds an abstraction layer which hides hardware details
  - Data is stored on a virtual disk managed by LVM
  - LVM is contained in SuSE SLES7 and SLES8
  - http://www.sistina.com/products_lvm.htm
- Benefits of LVM
  - file and file system sizes independent of physical hardware size
  - allows to manage multiple devices as one entity (scalability!)
  - offers much higher performance when configured appropriately

LVM: operating system view



LVM: improving performance

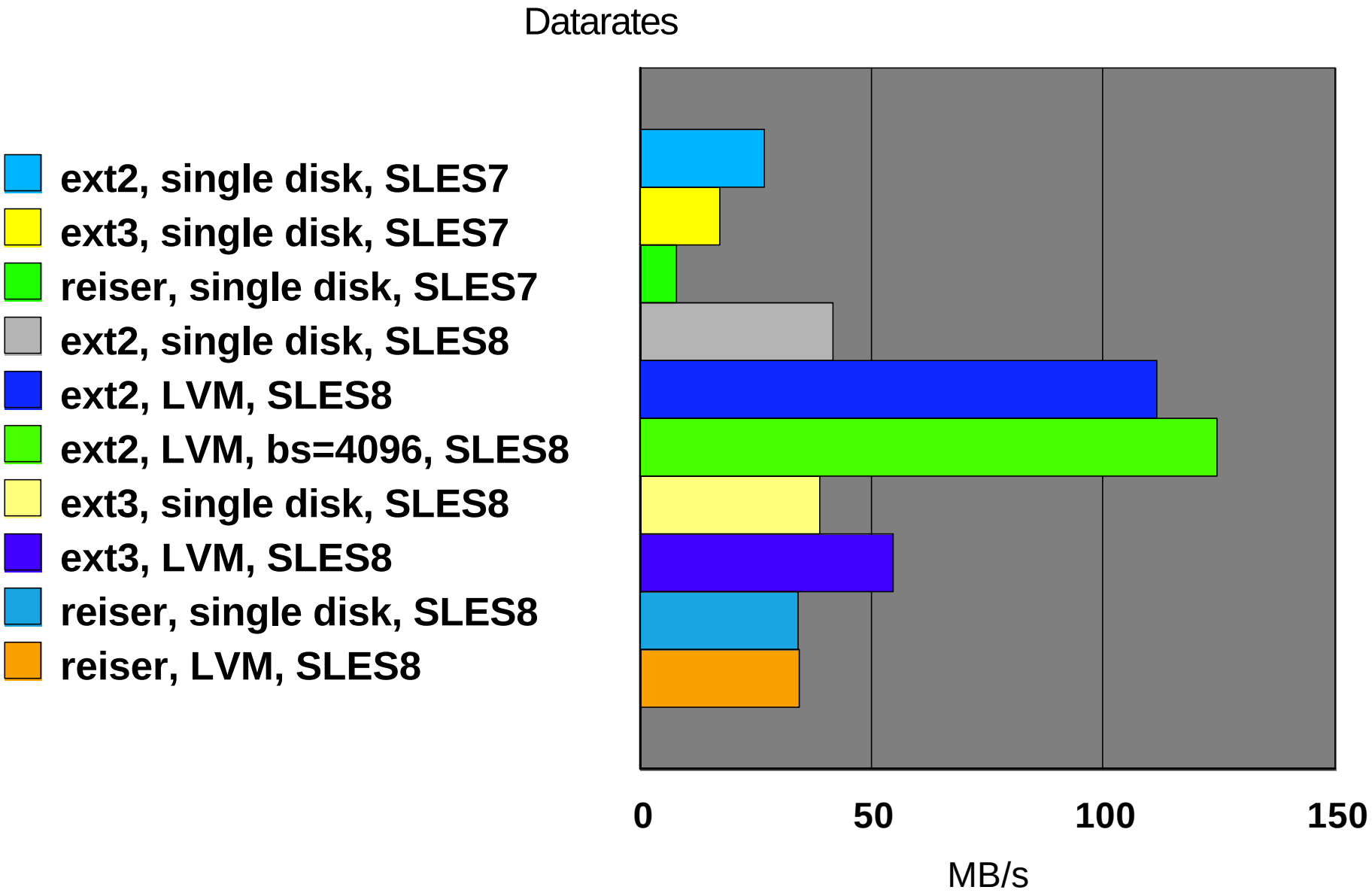


- With LVM ***and*** striping parallelism is achieved.
- http://publib-b.boulder.ibm.com/Redbooks.nsf/RedbookAbstracts/tips0128.html?open

A simple experiment (1)

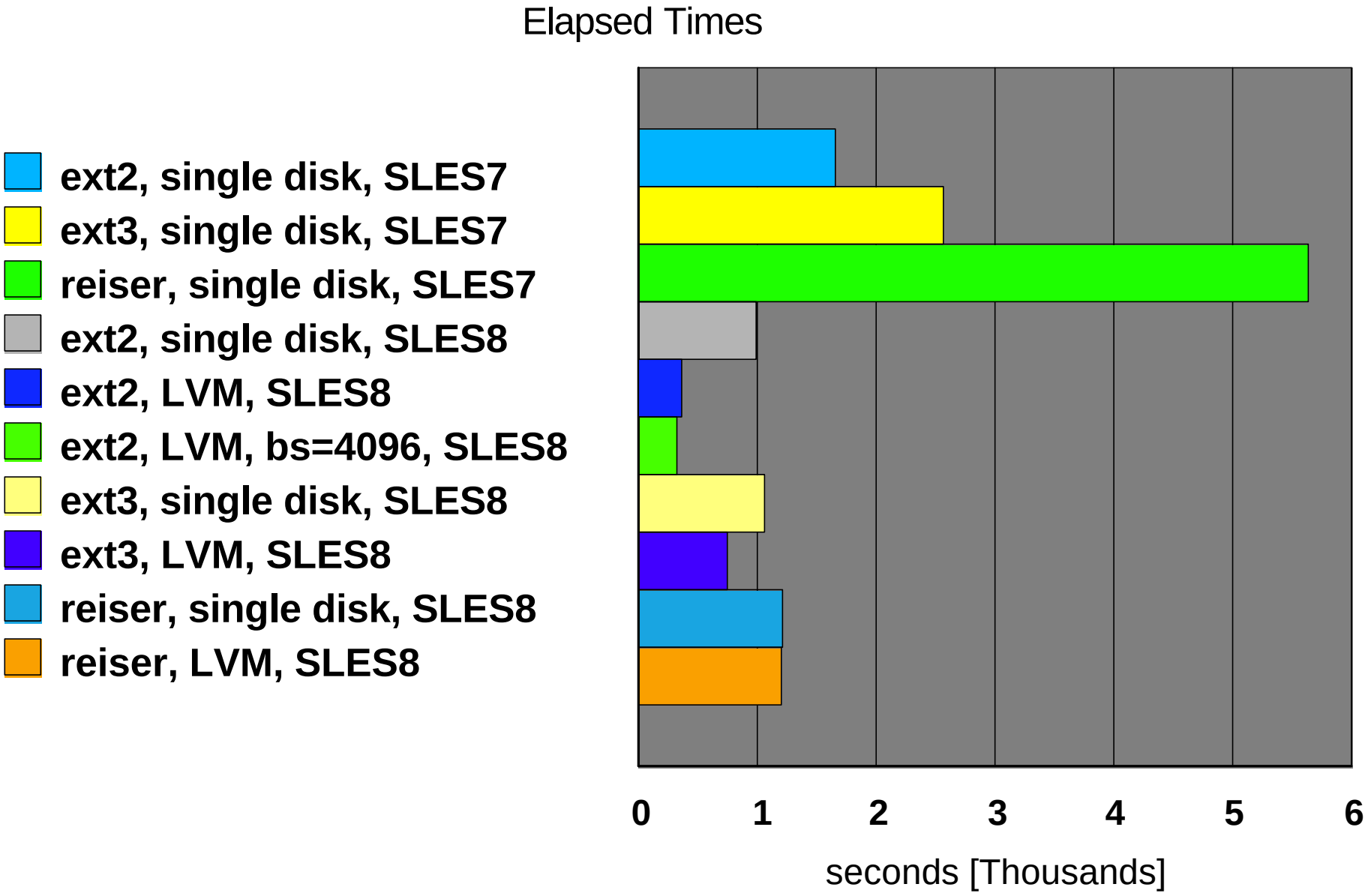
System setup: SuSE SLES8, VM guest, 1 CPU, 256MB RAM
Single disk setup: 50GB SCSI disk
LVM setup: 4 CHPIDs, 4 WWPNs, 4 ranks, 16 disks,
16 stripes, 32KB stripe size, 62GB total disk size
FS block size: 4096 bytes
Write block size: 1024 bytes (4096 bytes)
Write command: dd if=/dev/zero of=/mnt/dummy.file bs=1024 count=43352000

A simple experiment (2)



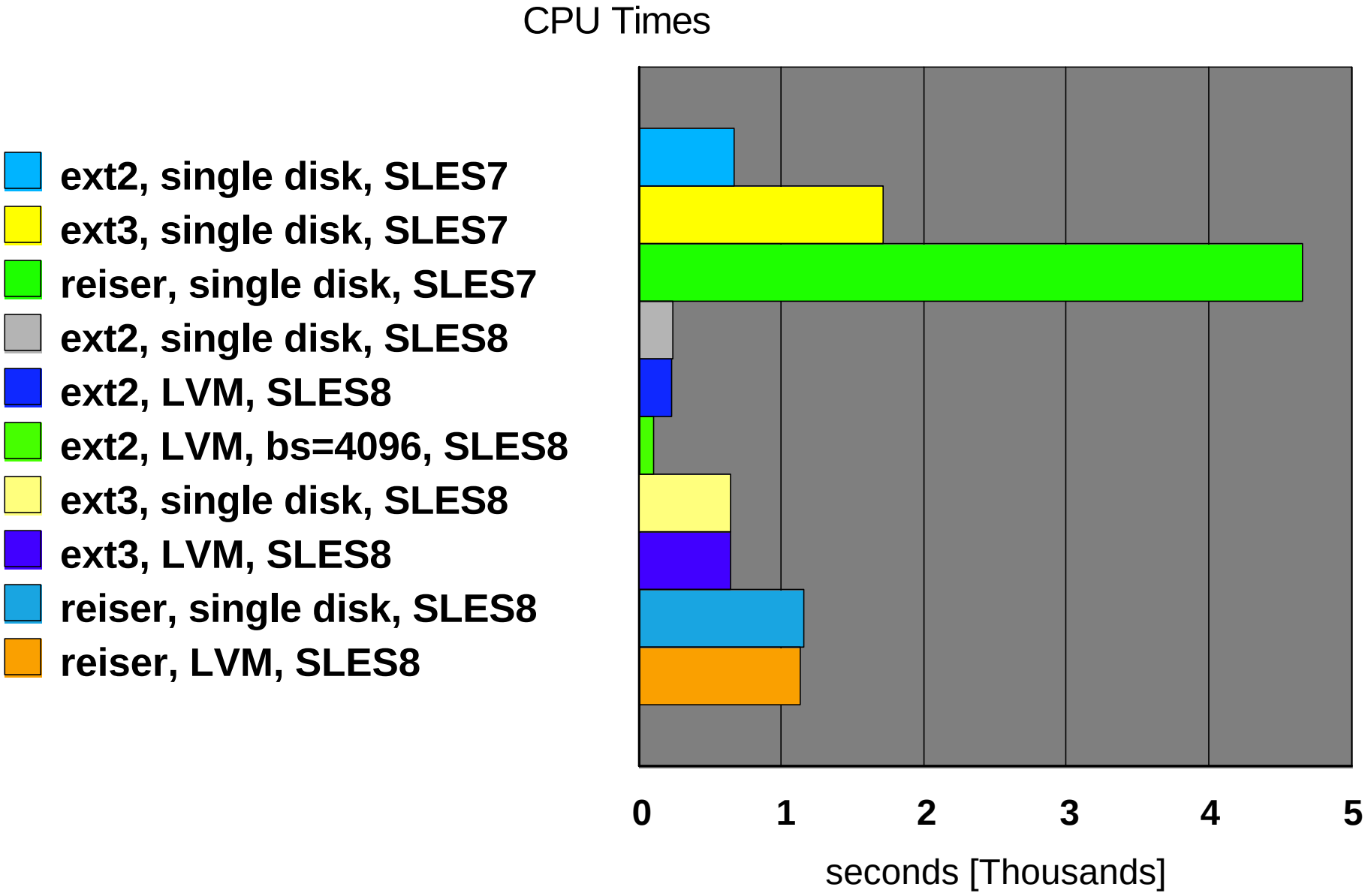
dd if=/dev/zero of=/mnt/dummy.file bs=1024 count=43352000

A simple experiment (3)



dd if=/dev/zero of=/mnt/dummy.file bs=1024 count=43352000

A simple experiment (4)



dd if=/dev/zero of=/mnt/dummy.file bs=1024 count=43352000

Visit us

Linux for zSeries Performance Website:

http://www10.software.ibm.com/developerworks/opensource/linux390/whatsnew.shtml

Linux-VM Performance Website:

http://www.vm.ibm.com/perf/tips/linuxper.html

Questions

Thank you!