Linux on zSeries Performance Update

Klaus Bergmann
IBM Boeblingen, Germany

Trademarks

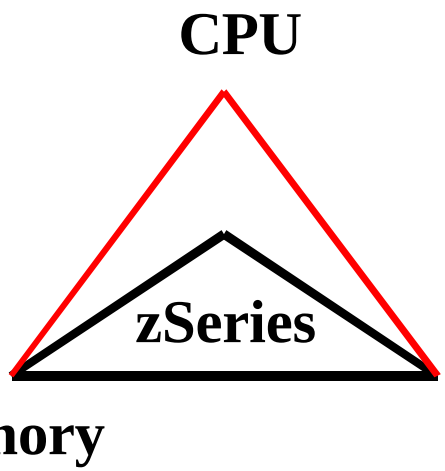
The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.
Enterprise Storage Server
ESCON*
FICON
FICON Express
HiperSockets
IBM*
IBM logo*
IBM eServer
Netfinity*
S/390*
VM/ESA*
WebSphere*
z/VM
zSeries
* Registered trademarks of IBM Corporation
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.

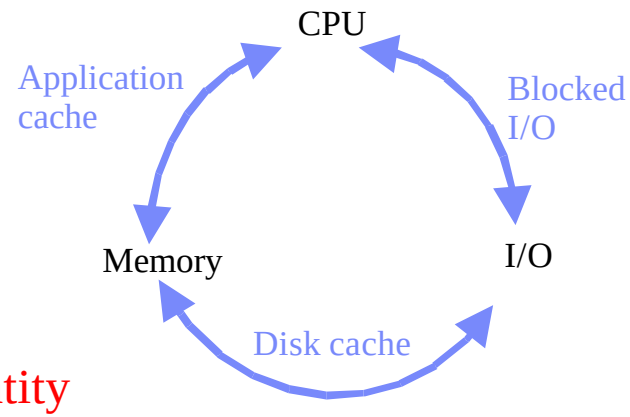
Agenda

- **Relative System Capacity**
- zSeries Hardware
- Scalability
- Java
- Disk I/O
  - Parallel Access Volume (PAV)
  - ESS Architecture

Relative System Capacity

- A system provides different types of resources
- Capacity for each resource type may be different
- The ideal machine provides enough capacity of each type
- Don't forget additional Resources (Network, Skilled staff, Money, availability of software, reliability, time ...)

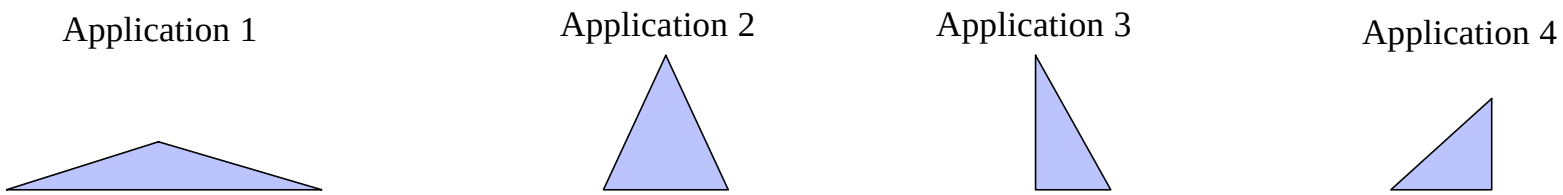




The ideal platform requires a mix of resources in right quantity

Resource Profiles

- Each application has its specific requirements
  - CPU intensive
  - I/O intensive
  - Memory
- Applications can often be tuned to change the resource profile
  - Exchange one resource for the other
  - Requires knowledge about available resources
- Some platforms can be extended better than others
  - Not every platform runs every application well
  - It's not easy to determine the resource profile of an appl.



Agenda

- Relative System Capacity
- **zSeries Hardware**
- Scalability
- Java
- Disk I/O
  - Parallel Access Volume (PAV)
  - ESS Architecture

The Evolution of zSeries Mainframes since 2000

Continous extension of 2000/2001 z900-Base

- Functionality
- Capacity spectrum

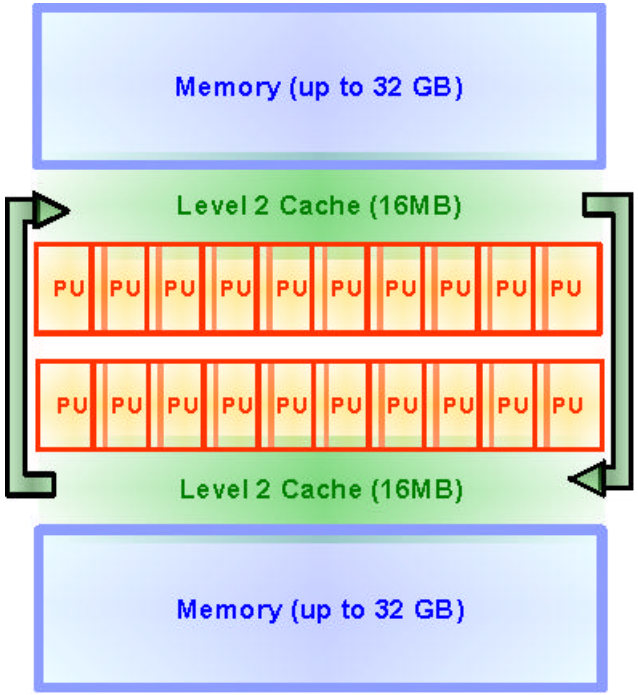
12/00 10/01 02/02 04..08/02 10..12/02

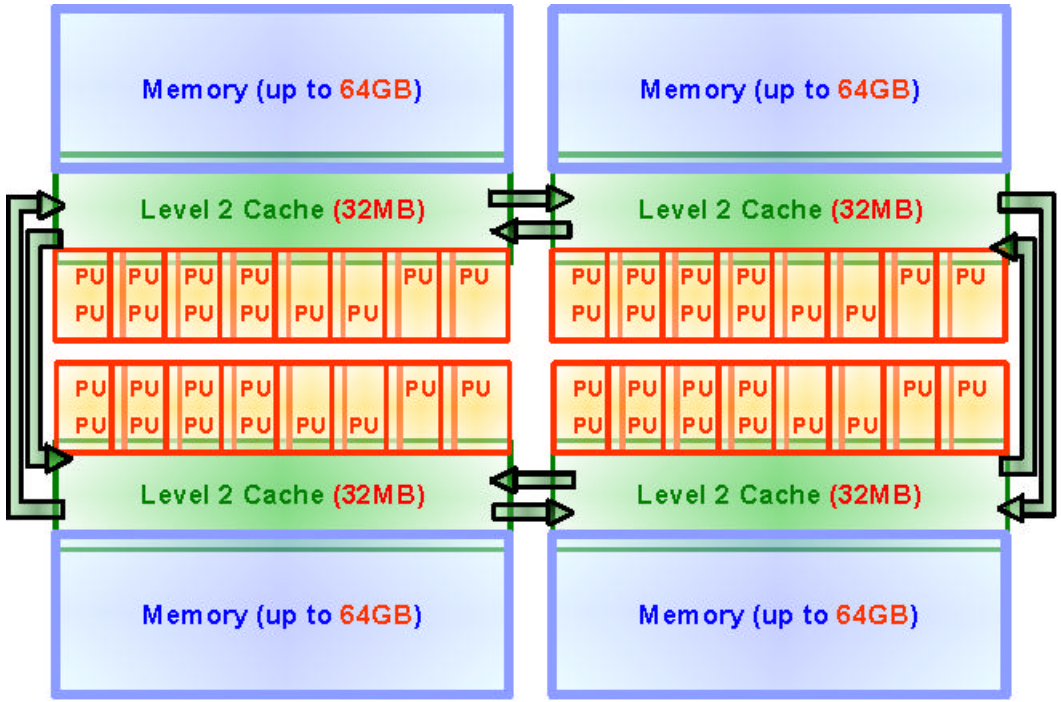
New in 2003: z990 GA1/2

New in 2004: z990 GA3, z890

06/03 10/03 04/04

zSeries 2003: Extended Multi-Book(Node)-Structures:

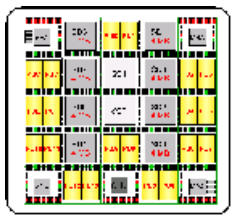


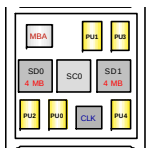


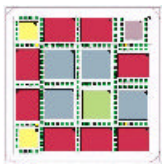
From z900/z800 ...

... to modular z990 systems with up to 3-fold capacity

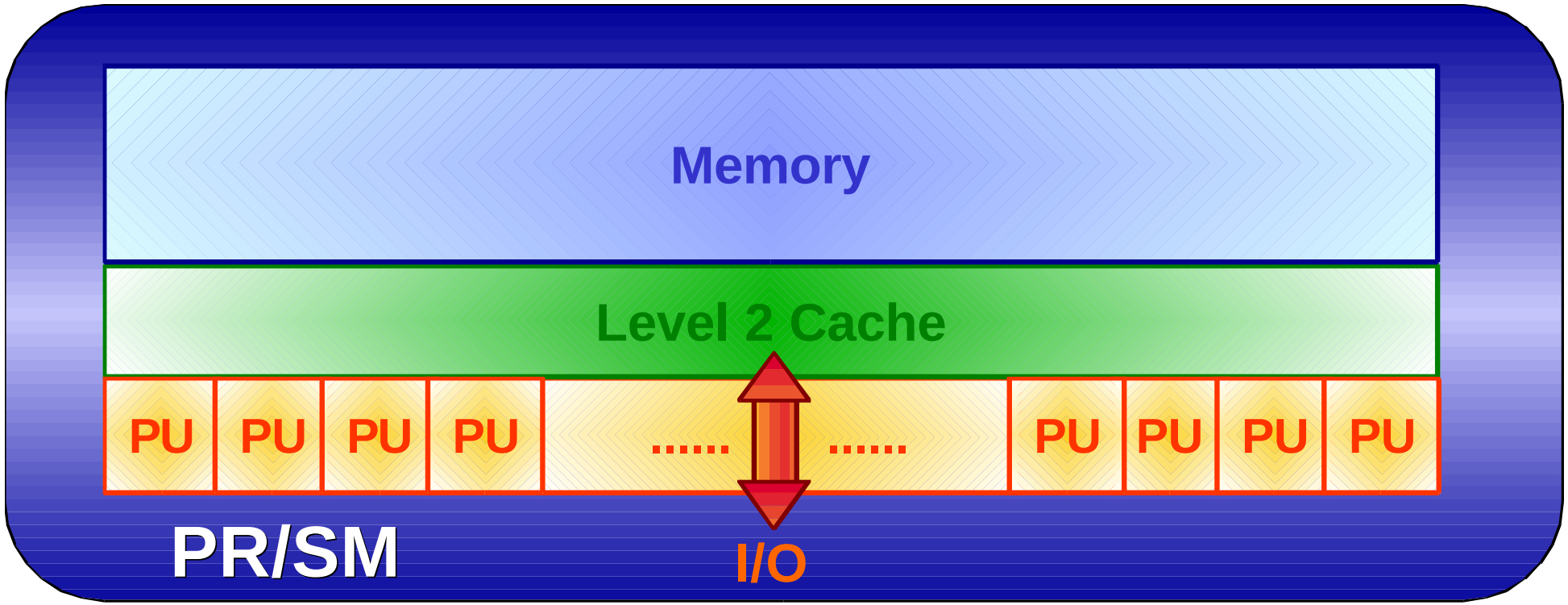
- 0.83 nsec CPU-Cycle (1.2 GHz)
- Superscalar design
- ..50..60.. % more UP-Performance vs 2C1







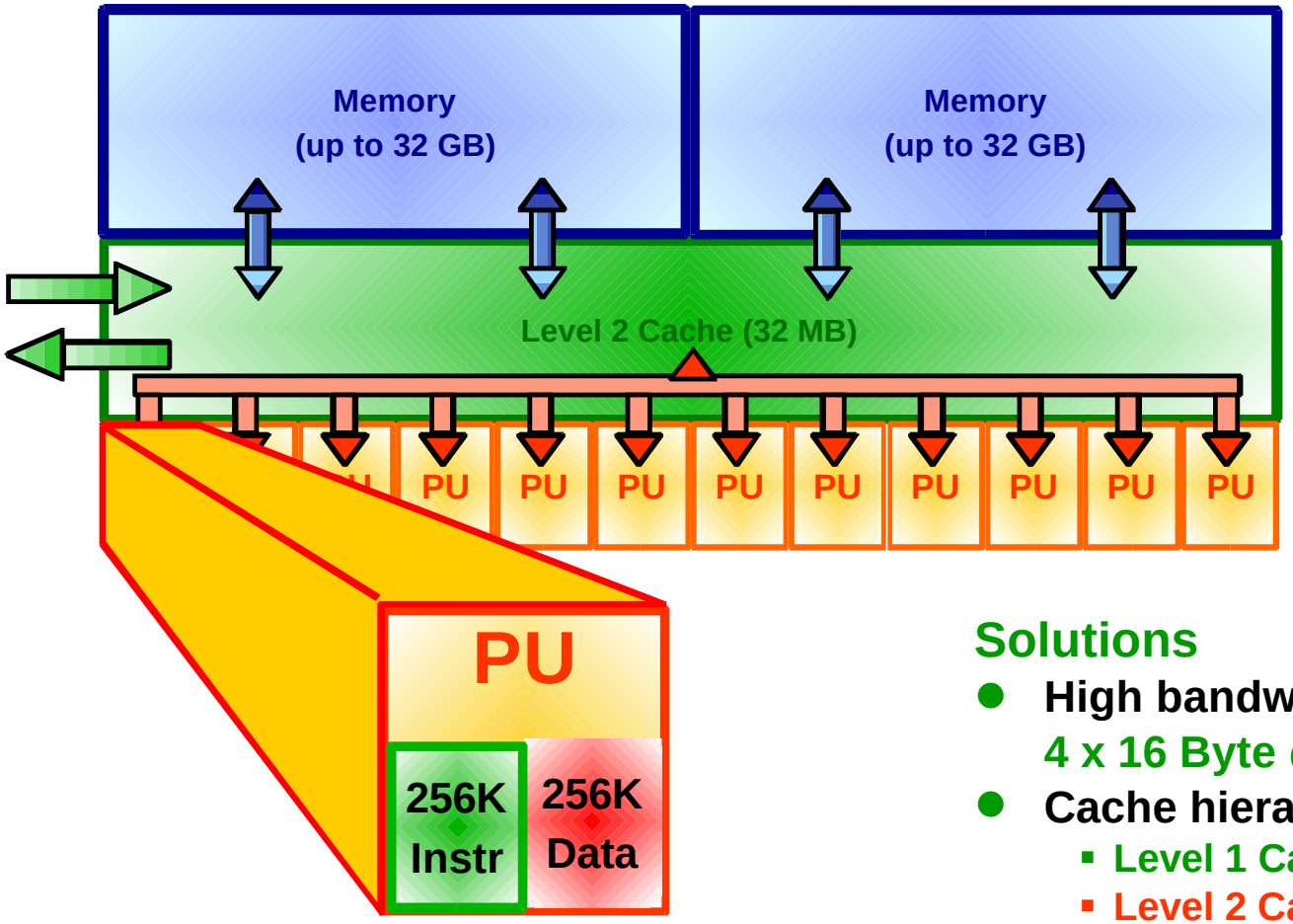
zSeries 2003: Multi-Book (-Node) Structures (Logical View)

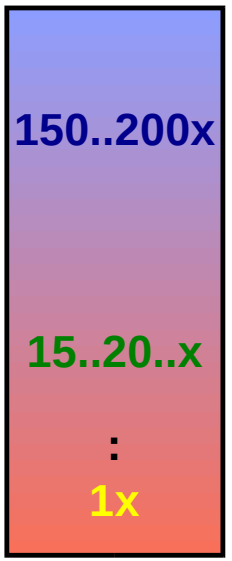


- LPAR mode only
- Single Pool of physical resources (CPU's, Memory, I/O) in modular implementation (1/2/3/4 Nodes/'books')
- Multiple Channel Subsystems: Up to 4 (z890: 2) x 256 CHPIDs)
- Exploitation by virtual servers: Up to 30 LPARs ...100+... (VM)

zSeries System Structure: Optimized for maximum internal bandwidth

z990





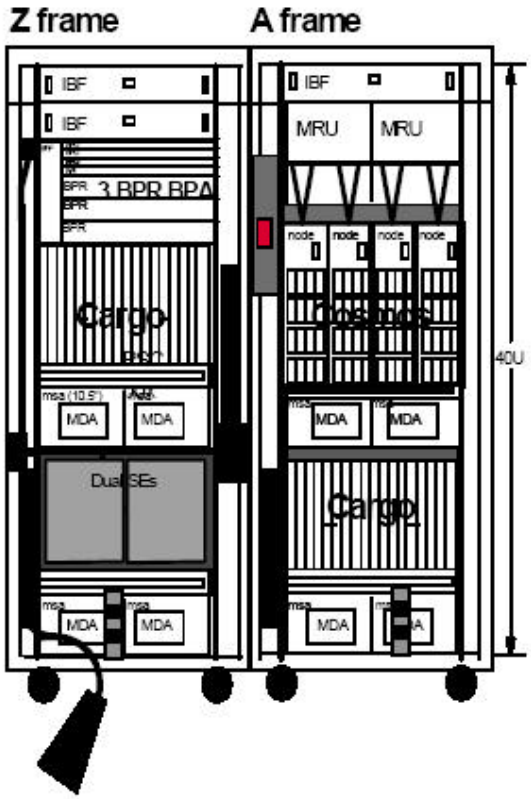
The Problem:

- Memory access does not scale with CPU-frequency

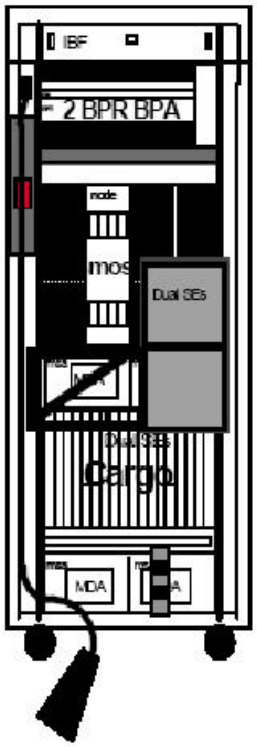
Solutions

- High bandwiths: Throughput !
  4 x 16 Byte @ 2 ns : Elastic Interface
- Cache hierarchies: Latency !
  - Level 1 Caches on CPU-Chip
  - Level 2 Cache 'shared by 12'
  - Connected to other books by double ring

2004: From z990 to z890



z990: A08/B16/C24/D32



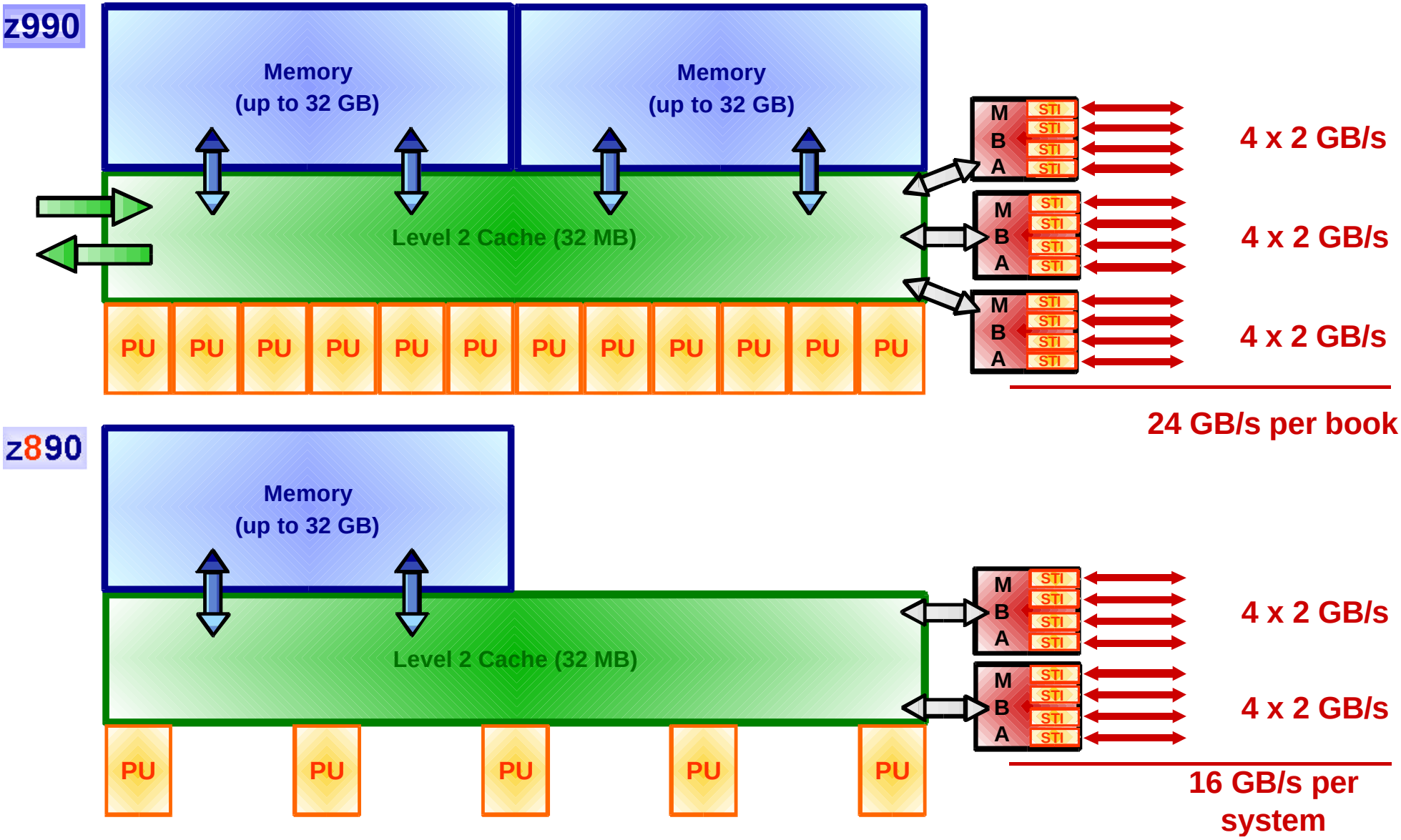
z890: A04

- Machine Type: 2086
- Model A04
- Single Frame
- No MRU, aircooled @1.0 nsec
- Up to 4 CPs, ICFs, IFLs, zAAPs…
- Functionality = z990
- Highly granular capacity range
  - 28 versions from 26 thru 1365 MIPS (“Mixed LSPR”)
  - 7 capacity families
    - 1- to -4-way each
  - concurrent upgrade any to any

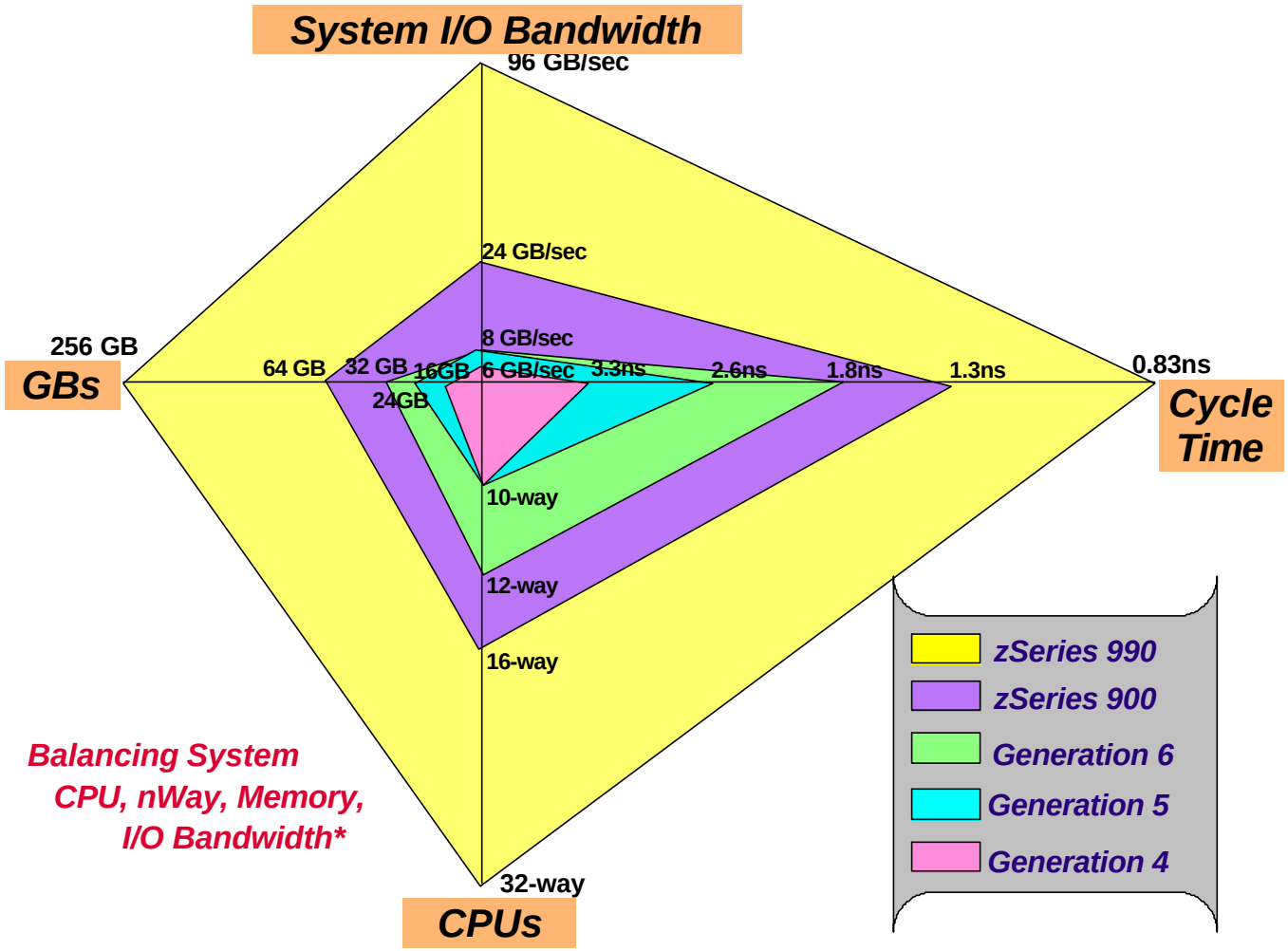
Not just a z800 FO: full blown z990 functions at SMB TCO !

Staying current will pay off !

zSeries System Structure: Optimized for maximum external bandwidth



IBM S390 and zSeries Servers – Balanced Scaling



* External I/O or STI bandwidth only (Internal Coupling Channels and HiperSockets not included)
zSeries MCM internal bandwidth is 500 GB/s. Memory bandwidth not included (not a system constraint)

Performance results

Our Hardware for Measurements

2064-216 (z900)

1.09ns (917MHz)
2 * 16 MB L2 Cache (shared)
64 GB
FICON
HiperSockets
OSA Express GbE
z/VM 4.3

2105-F20 (Shark)

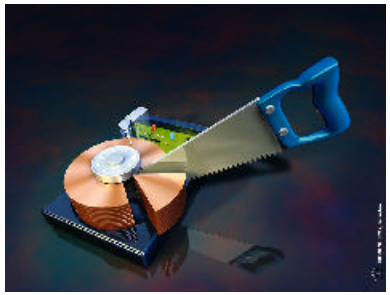
384 MB NVS
16 GB Cache
128 * 36 GB disks
10.000 RPM
FCP (2 Gbps)
FICON (1 Gbps)

2084-B16 (z990)

0.83ns (1.2 GHz)
2 Books each with 8 CPUs
64 GB
FICON
HiperSockets
OSA Express GbE
z/VM 4.4

8687-3RX (8-way X440)

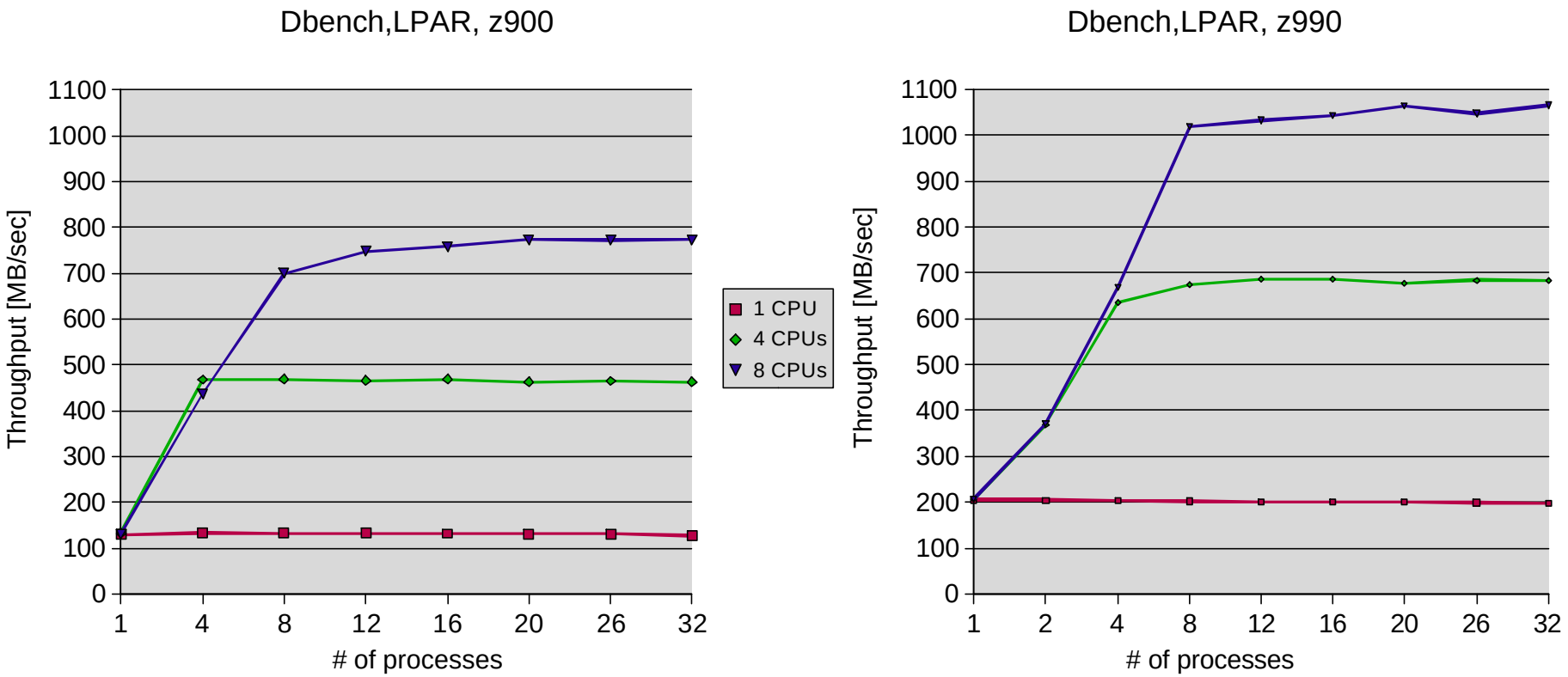
8-way Intel Pentium III Xeon
1.6 Ghz
8*512K L2 Cache (private)
hyperthreading
summit chipset



Agenda

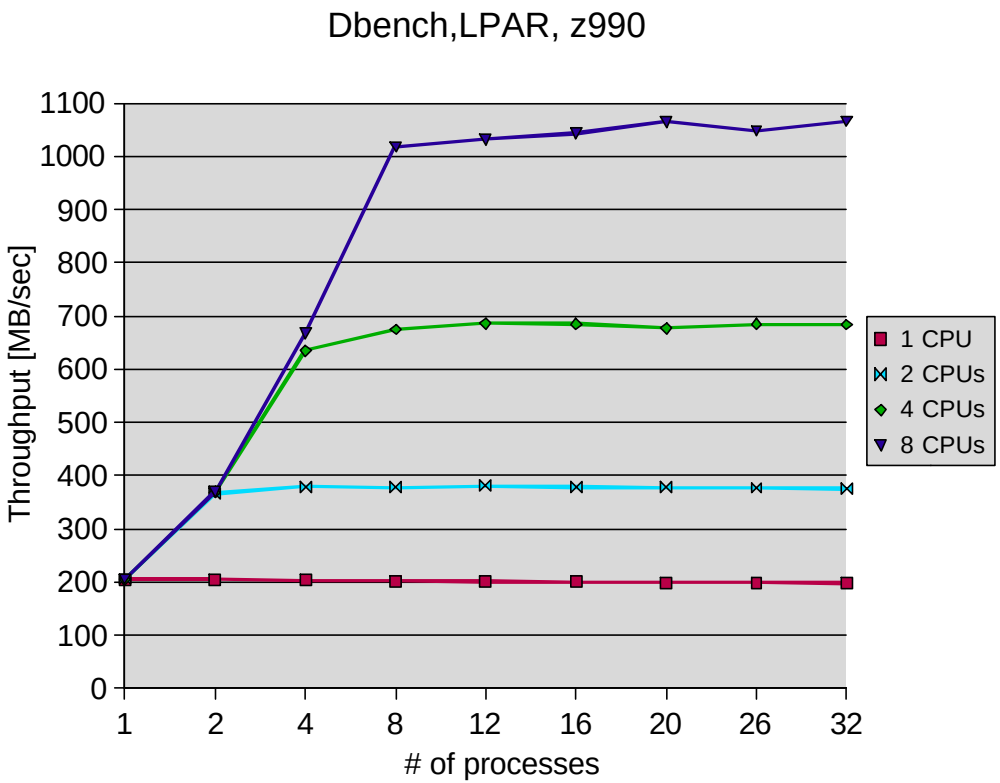
- Relative System Capacity
- zSeries Hardware
- **Scalability**
- Java
- Disk I/O
  - Parallel Access Volume (PAV)
  - ESS Architecture

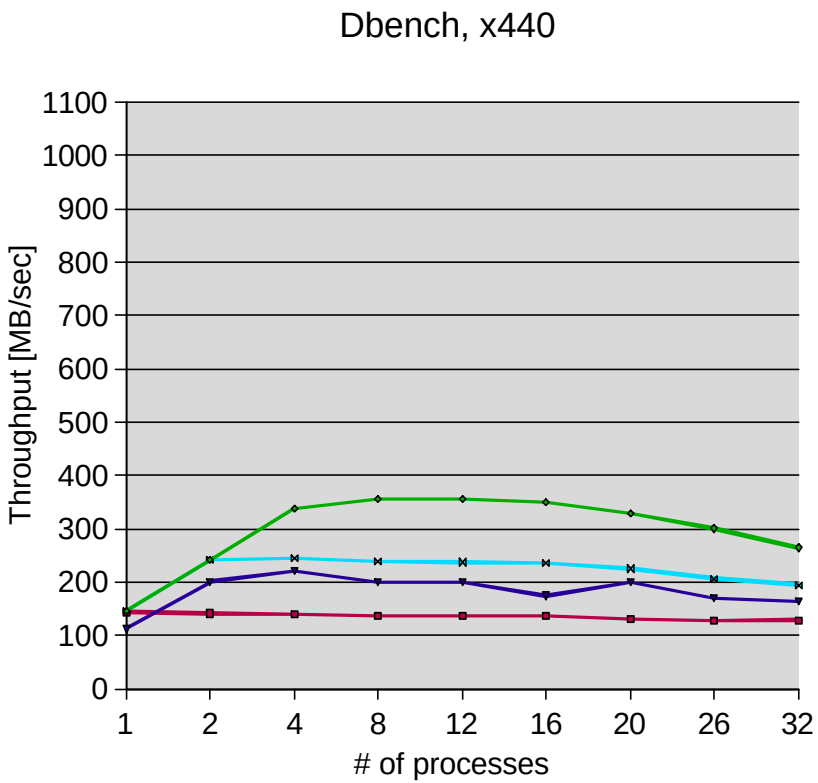
Scalability - z900 vs z990, ext2, 31 Bit



- z990 takes advantage of higher memory bandwidth

Scalability - z990 vs Intel, ext2, 31/32Bit

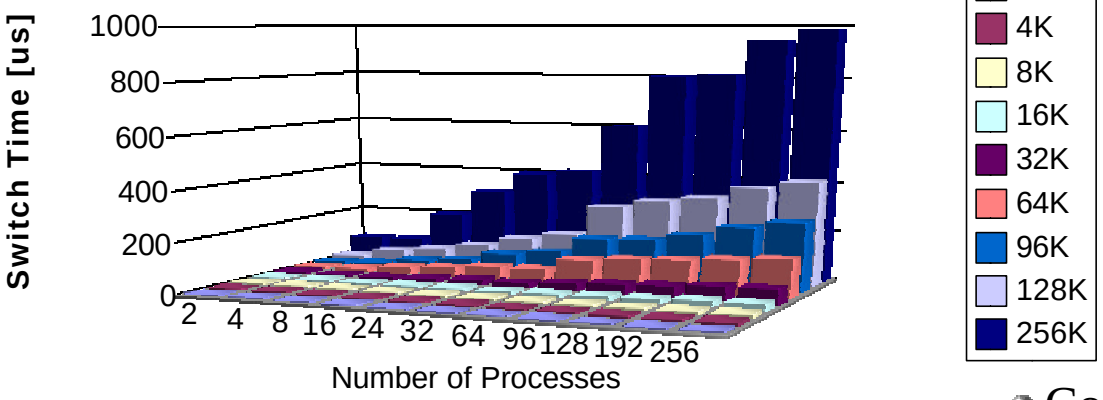




- z990 shows good scaling behavior
- x440 shows best throughput with 4 CPU, strong throughput degradation with more than 4 CPUs

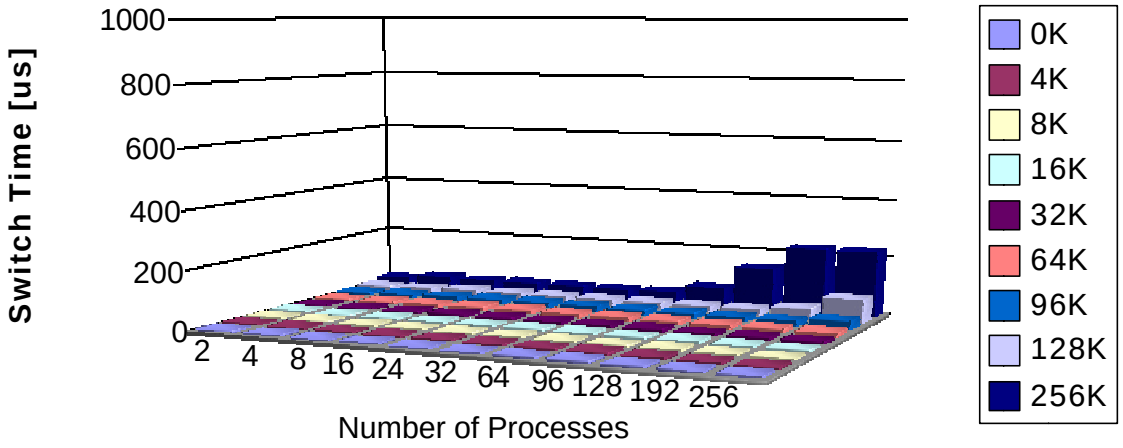
Kernel – Context Switches

x440, SLES-8

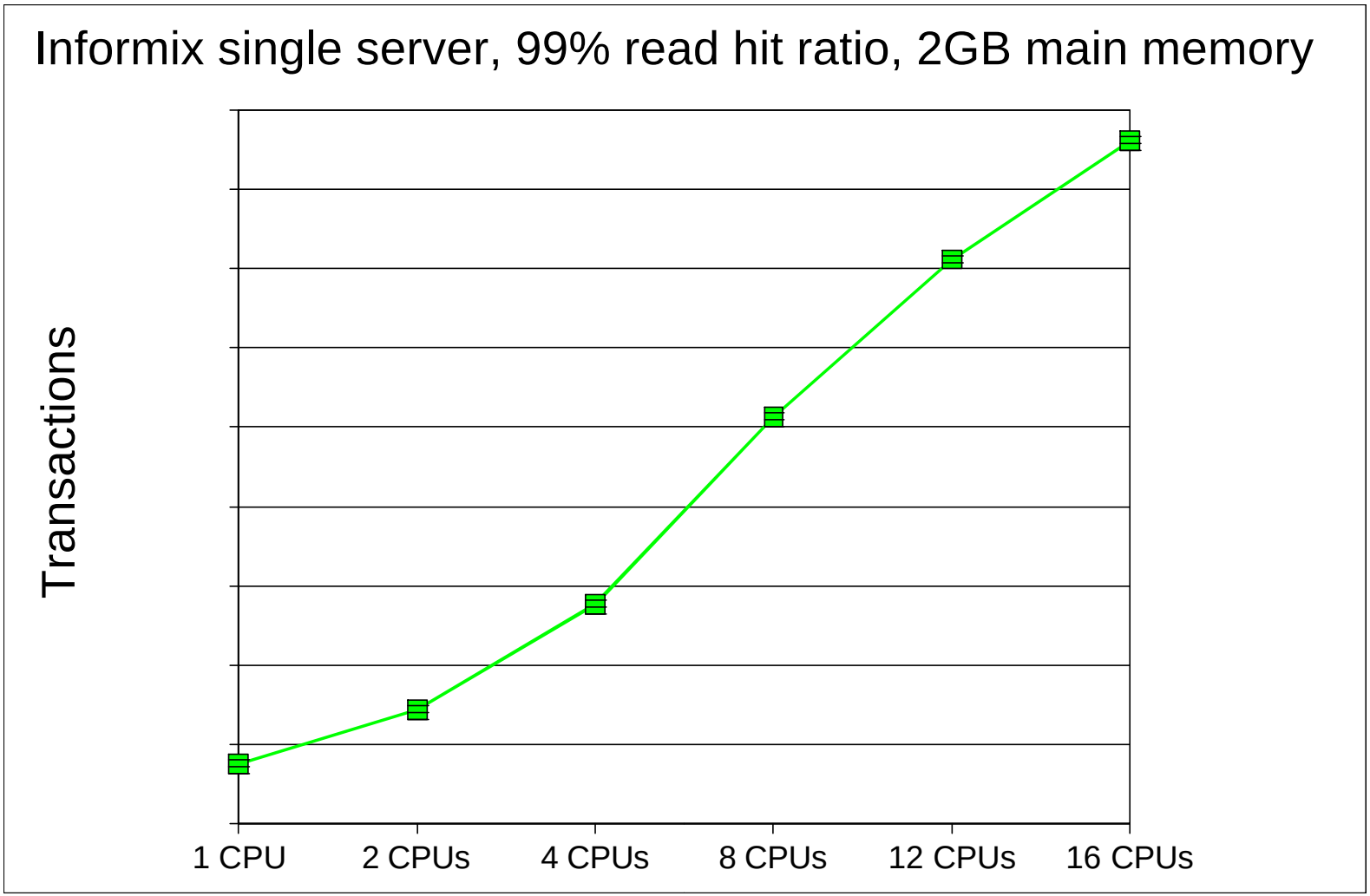


- Context Switches much faster on zSeries because of large shared caches

z990, LPAR, SLES-8, 31-bit



OLTP Workload Informix - CPU Scaling



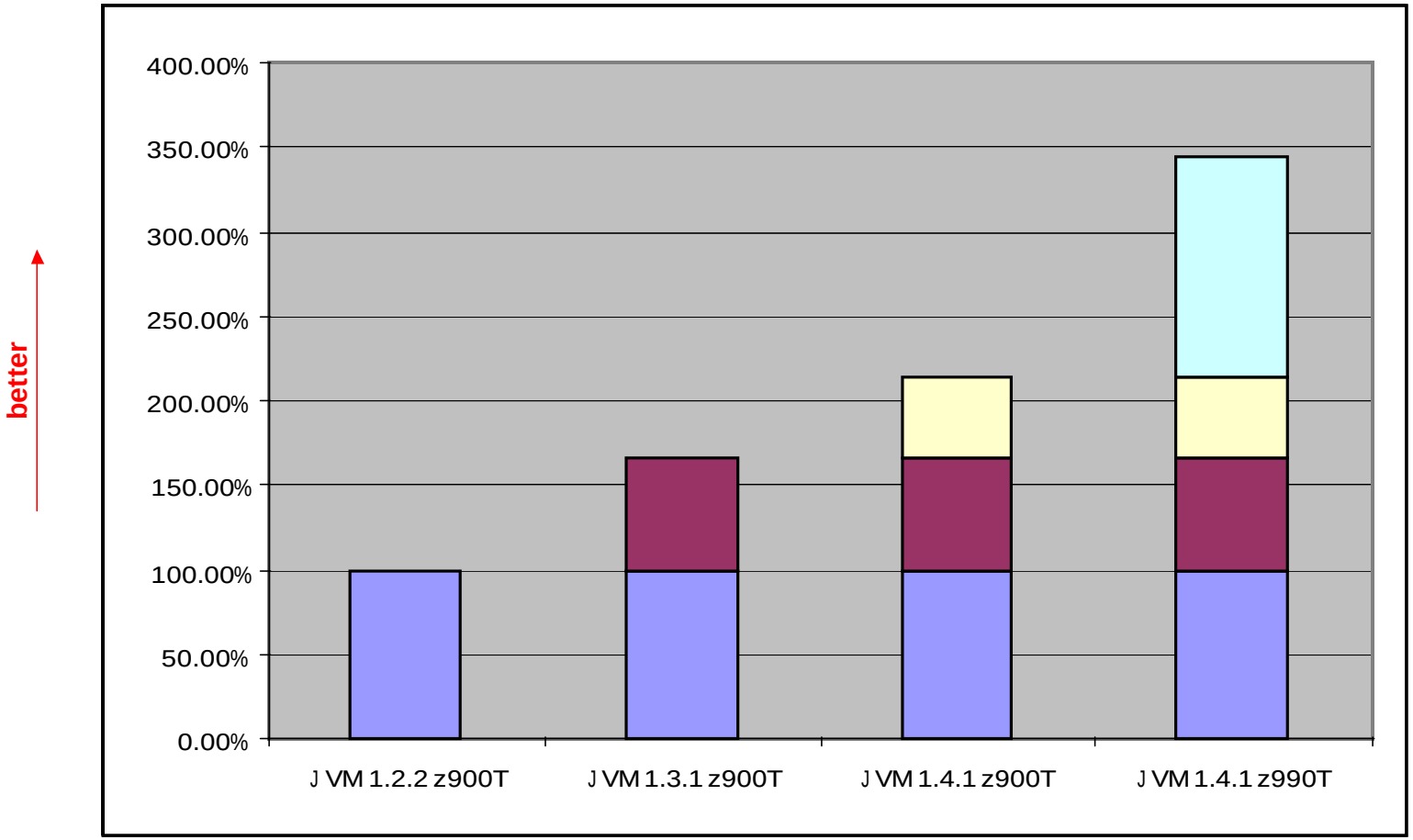
Agenda

- Relative System Capacity
- zSeries Hardware
- Scalability
- **Java**
- Disk I/O
  - Parallel Access Volume (PAV)
  - ESS Architecture

Java

- Java Virtual Machine improved
- zSeries Just in Time Compiler improved
- 2001: JVM 1.2.2, Websphere 3.x
- 2002: JVM 1.3.1, Websphere 4.x, 5.0
- 2003: JVM 1.4.1, Websphere 5.1

Java

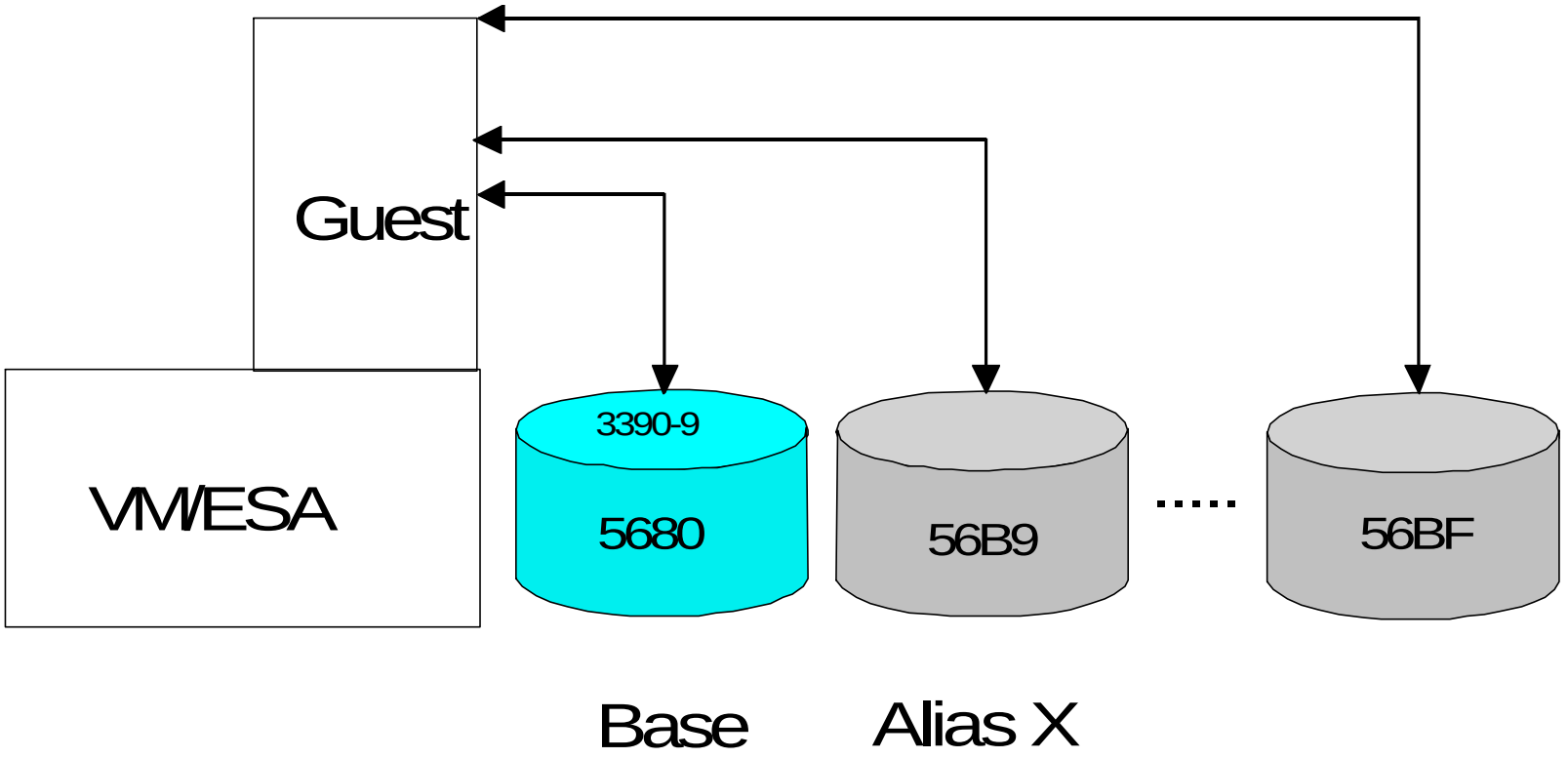

400.00%
350.00%
300.00%
250.00%
200.00%
150.00%
100.00%
50.00%
0.00%
JVM 1.2.2 z900T
JVM 1.3.1 z900T
JVM 1.4.1 z900T
JVM 1.4.1 z990T

better

Agenda

- Relative System Capacity
- zSeries Hardware
- Scalability
- Java
- **Disk I/O**
  - **Parallel Access Volume (PAV)**
  - ESS Architecture

Parallel Access Volume (PAV)



Linux cannot enable PAV on the ESS but can use it under VM

Base and Aliases (PAV Cont.)

- IOCDS changes

```
IODEVICE ADDRESS=(5680,024),UNITADD=00,CUNUMBR=(5680),   *
        STADET=Y,UNIT=3390B
IODEVICE ADDRESS=(5698,040),UNITADD=18,CUNUMBR=(5680),   *
        STADET=Y,UNIT=3390A
```

- ATTACH Base and Aliases to the guest
- QUERY PAV shows base and alias addresses

cat /proc/dasd/devices

```
5794(ECKD) at ( 94:  0) is dasda      : active at blocksize: 4096, 1803060 blocks, 7043 MB
5593(ECKD) at ( 94:  4) is dasdb      : active at blocksize: 4096, 601020 blocks, 2347 MB
5680(ECKD) at ( 94:  8) is dasdc      : active at blocksize: 4096, 1803060 blocks, 7043 MB
56bf(ECKD) at ( 94: 12) is dasdd      : active at blocksize: 4096, 1803060 blocks, 7043 MB
```

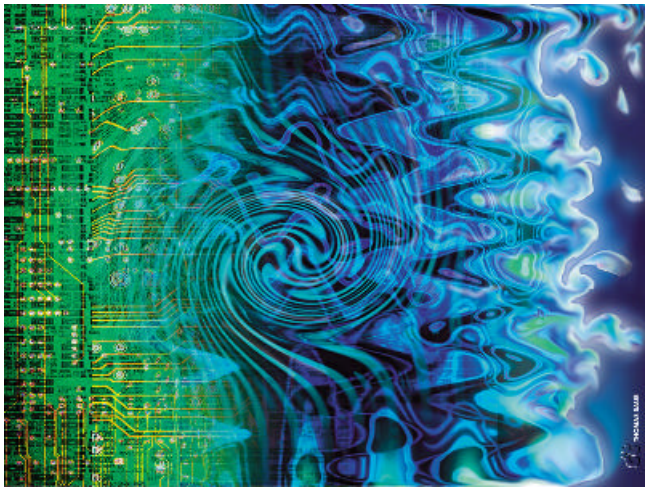
cat /proc/subchannels | egrep "5680|56BF"

```
5680    0030  3390/0C   3990/E9  yes     FC  FC  FF  C6C7C8CA CBC90000
56BF    0031  3390/0C   3990/E9  yes     FC  FC  FF  C6C7C8CA CBC90000
```

This works only with z/VM

LVM commands (PAV Cont.)

- vgscan: create configuration data
  - scans all discs for volume groups
- pvcreate /dev/dasdc1
  - has to be done for each physical volume
- vgcreate vg_kb /dev/dasdc1
  - creates the volume group vg_kb
- vgdisplay



vgdisplay

vgdisplay -v vg_kb

```
--- Volume group ---
VG Name                vg_kb
VG Access              read/write
VG Status              available/resizable
VG #                   0
MAX LV                 256
Cur LV                 0
Open LV                0
MAX LV Size            255.99 GB
Max PV                 256
Cur PV                 1
Act PV                 1
VG Size                6.87 GB
PE Size                4 MB
Total PE               1759
Alloc PE / Size        0 / 0
Free  PE / Size        1759 / 6.87 GB
VG UUID                3nwJYn-SxW1-gKym-OvZs-TYIf-CrHP-inO5Yp

--- No logical volumes defined in "vg_kb" ---
```

More LVM commands

lvcreate --name lv_kb --extents 1759 vg_kb

cat /proc/lvm/global

```
LVM module LVM version 1.0.5(mp-v6)(15/07/2002)
Total:  1 VG  1 PV  1 LV (0 Lvs open)
Global: 32300 bytes malloced   IOP version: 10   3:18:35 active
VG:  vg_kb  [1 PV, 1 LV/0 open]  PE Size: 4096 KB
  Usage [KB/PE]: 7204864 /1759 total  7204864 /1759 used  0 /0 free
  PV:  [AA] dasdc1                  7204864 /1759     7204864 /1759
         0 /0
       +-- dasdd1
    LV:  [AWDL  ] lv_kb                        7204864 /1759     close
```

lvscan

```
lvscan -- ACTIVE            "/dev/vg_kb/lv_kb" [6.87 GB]
lvscan -- 1 logical volumes with 6.87 GB total in 1 volume group
lvscan -- 1 active logical volumes
```

Enable Paths

pvpath-change or query path attributes of a physical multipathed volume

pvpath -qa

```
Physical volume /dev/dasdc1 of vg_kb has 2 paths:
       Device  Weight Failed Pending State
 #  0:  94:9        0      0       0 enabled
 #  1:  94:13       0      0       0 disabled
```

The second path can be enabled:

pvpath -p1 -ey /dev/dasdc1

```
vg_kb: setting state of path #1 of PV#1 to enabled
```

pvpath -qa

```
Physical volume /dev/dasdc1 of vg_kb has 2 paths:
       Device  Weight Failed Pending State
 #  0:  94:9        0      0       0 enabled
 #  1:  94:13       0      0       0 enabled
```

Now LVM is ready to use both paths to the volume

Results

iozone sequential write/read 1 disk

Base / Alias	Write MB/s	Read MB/s
1 / 0	15	26
1 / 1	26	52
1 / 2	28	78
1 / 3	28	100
1 / 4	30	120
1 / 5	30	134
1 / 6	30	141
1 / 7	30	147

Agenda

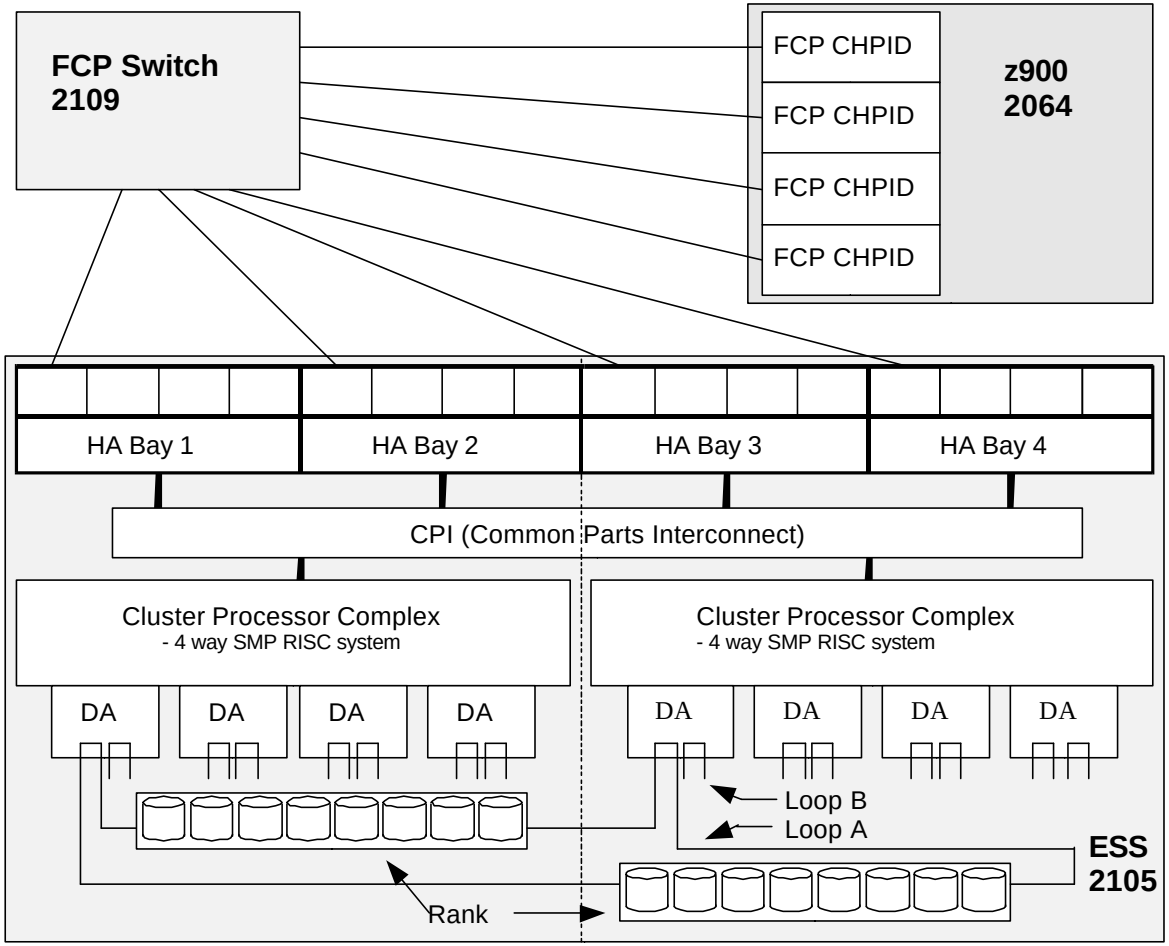
- Relative System Capacity
- zSeries Hardware
- Scalability
- Java
- **Disk I/O**
  - Parallel Access Volume (PAV)
  - **ESS Architecture**

ESS – 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

Let's have a deeper look to the elements of the scenario:



- **CHPIDs**
- **Host Adapter (HA) supporting FCP (FCP port)**
  -16 Host Adapters, organized in 4 bays, 4 ports each
- **Device Adapter Pairs (DA)**
  - each one supports two loops
- **Disks are organized in ranks**
  - each rank (8 physical disks) implements one RAID 5 array (with logical disks)

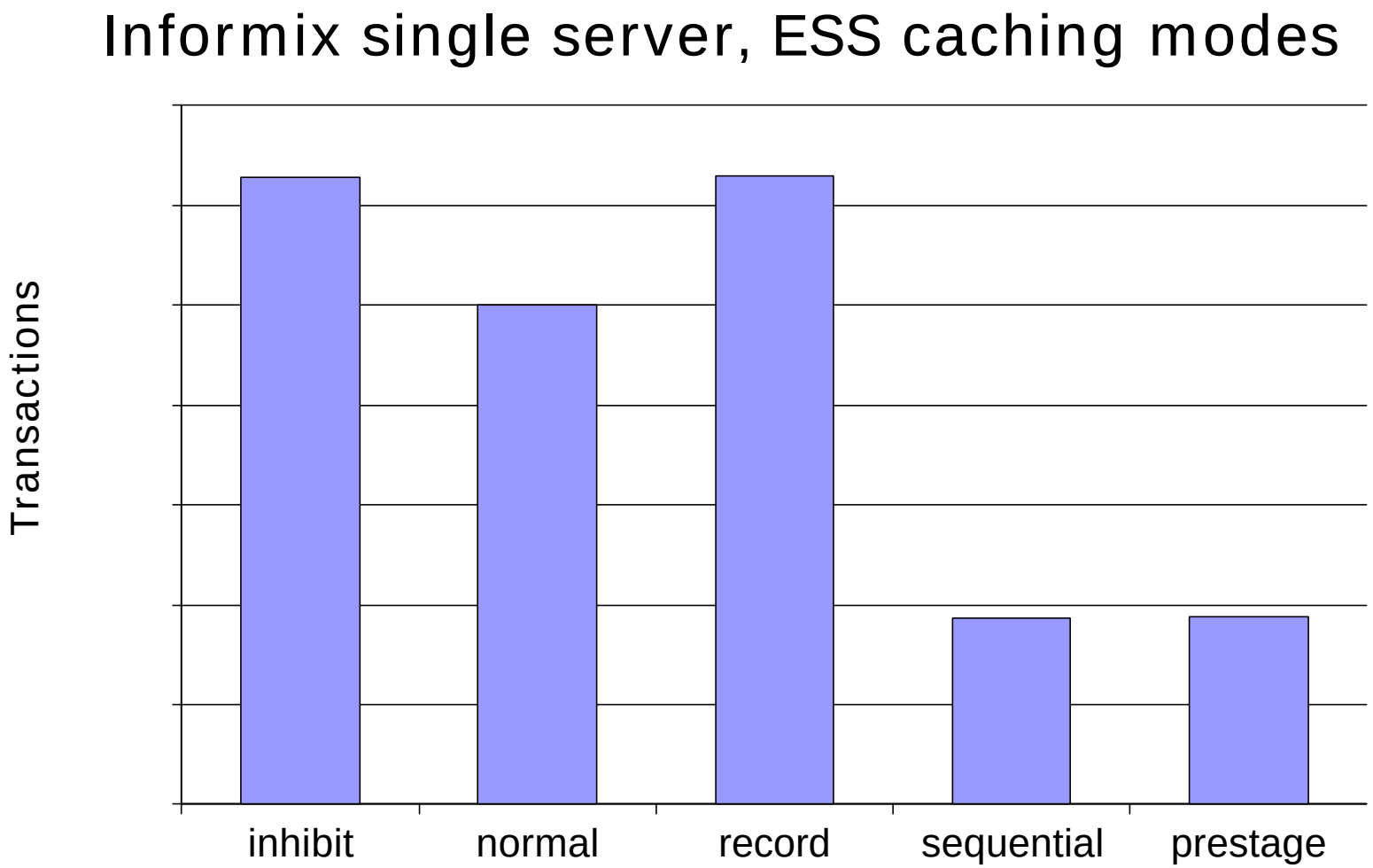
General Rules

- this makes it **slow**:
  - when all disks are from one rank and accessed via the same path
- this makes it **fast**:
  - use many host adapters
  - spread the host adapters used across all host adapter bays
  - use as much CHPIDs as possible and
    access each disk through all CHPIDs, if possible (FICON, LVM1-mp)
  - spread the disks used over all ranks equally
- this applies to FCP and FICON

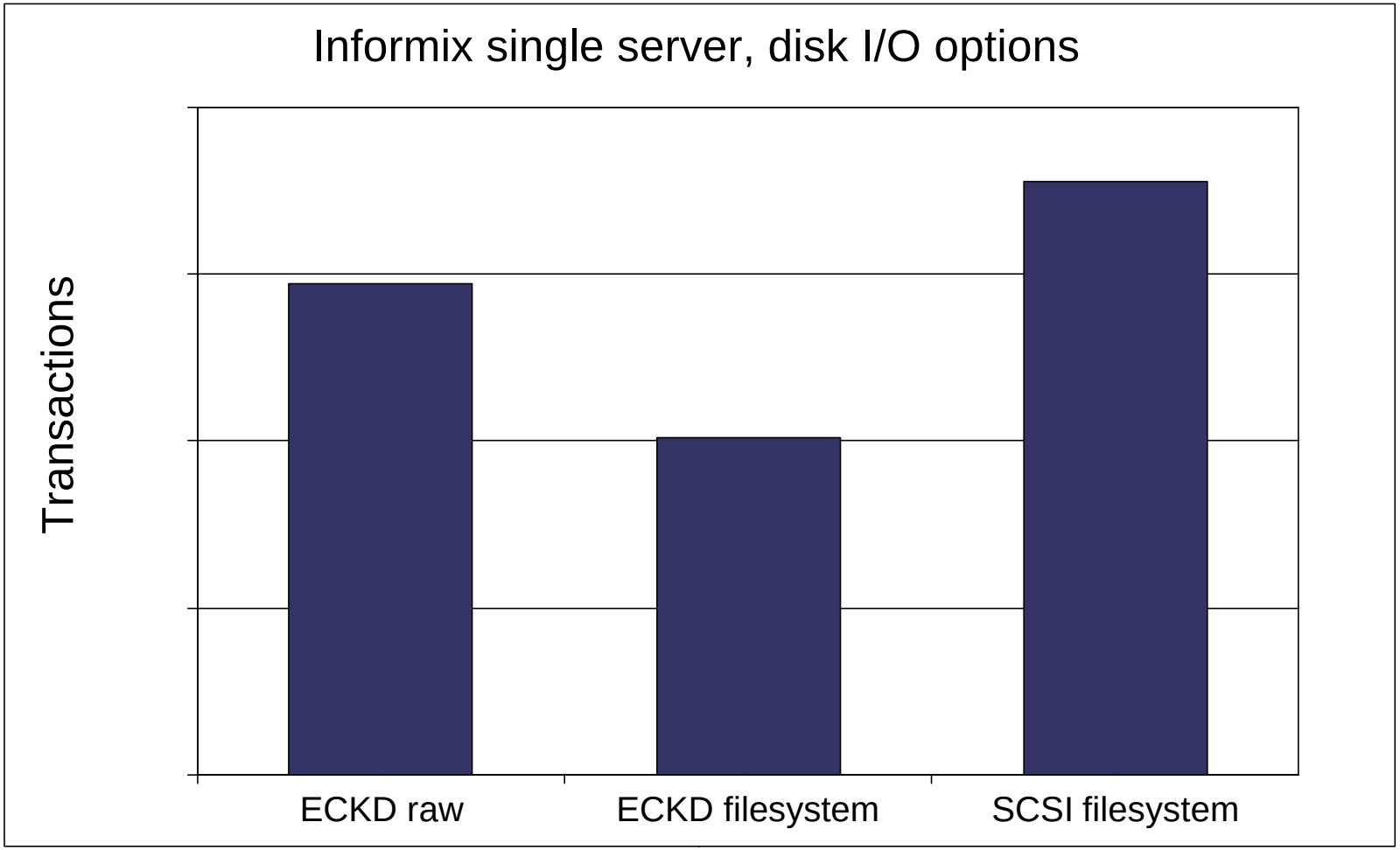
ESS Caching Modes

- Caching Modes: inhibit, normal, record, sequential, prestage
- Description: System/390 Command Reference 2105 Models, SC26-7298
- Default with SLES8, RHEL: “sequential”
- Our Experience with database workloads: switch to “normal”
- Patch available on SUSE's maintenance website
  (Kernel Version 2.4.21-107)
- See our “Linux on zSeries Tuning Hints and Tips” website for details

OLTP Workload Informix – I/O Options

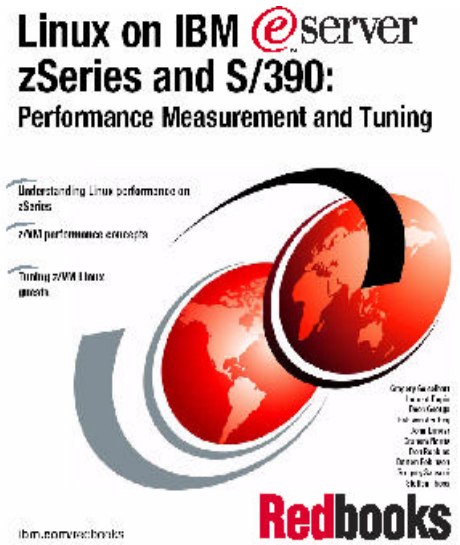


OLTP Workload Informix – I/O Options



Visit us !

- Linux on zSeries Tuning Hints and Tips
  - http://www10.ibm.com/developerworks/oss/linux390/perf/index.shtm
- Linux-VM Performance Website:
  - http://www.vm.ibm.com/perf/tips/linuxper.html
- Performance Redbook:
  - SG24-6926-00



Questions