Linux for zSeries Performance Measuring and Tuning

Session 9304

Jens Osterkamp (Jens.Osterkamp@de.ibm.com)

SHARE, February 22-27, 2004 | Longbeach, CA

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.

Performance Measuring and Tuning

- Resource profiles
- Linux and z/VM
- Tools
- Storage
- Networking
- Future enhancements

Tracking down performance problems

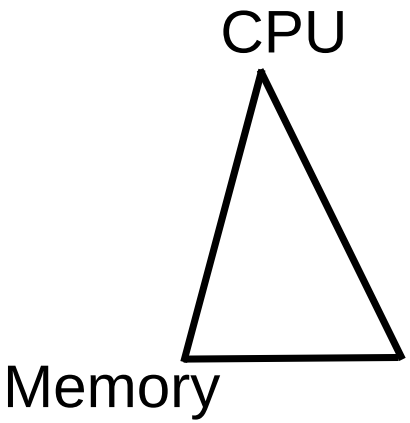
Questions to be answered

- What do I expect ?
- Do I have numbers for comparison ?
- Do I really have a problem ?
- Where do I suspect the problem ?
- What data do I want to collect ?
- Which tools can I use ?
- What do the numbers tell me ?
- What measures evolve from the numbers ?

resource profiles

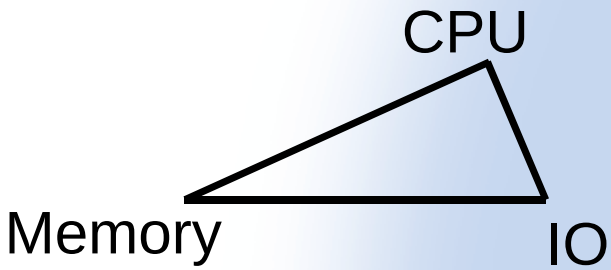
- Know your applications resource profile
- Know your systems resource profile

What your application needs



Does it really match ?

What your system provides



know your setup...

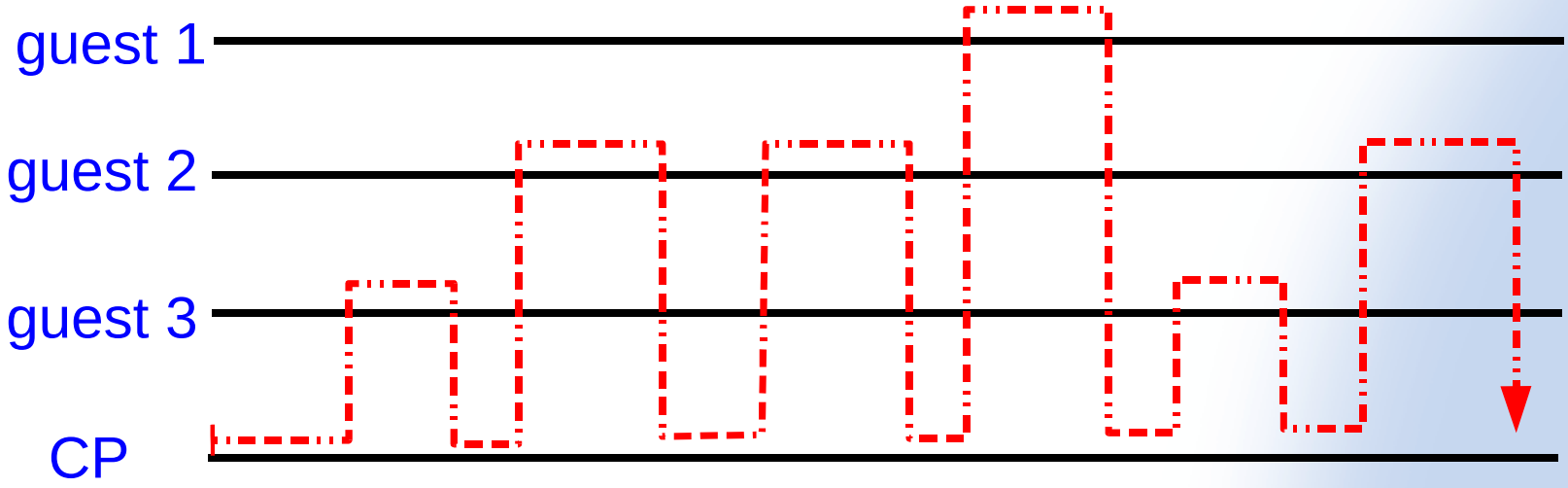
1. know your applications
2. know your environment (hardware, software, network)
3. know your resource limits (memory, cpu, IO, shares)
4. use tools to identify or ask your administrator
5. know your Linux

Linux and z/VM

Linux and z/VM : resource sharing

If your Linux is a VM guest have in mind :

- Linux tools do only see the share they got in VM
  - CPU might only be a part of a physical CPU
- use VM tools to monitor performance for your guests



sharing of one physical CPU

on-demand timer patch

- Linux image gets external timer interrupt (tick) every 10 ms
- several idle Linux images running in z/VM cause significant load
- solution : use larger timer intervals on idle Linux guests
- timer behaviour may be changed during run time by writing a '0' or a '1' to /proc/sys/kernel/hz_timer

but :

- older versions of the on-demand timer patch could lead to a misaccounting of Linux CPU load for certain workloads (high volume network traffic)

Linux and z/VM : VM monitor data

- In case you want to report a performance problem with z/Linux running as a VM guest, you may be asked to collect monitor data :
- data set with information about CPU load, storage, IO and guest activity
- prepare your VM system to collect monitor data. For a detailed description on how to do that, see

  http://www.vm.ibm.com/perf/tips/collect.html
- be sure to enable collection for all domains
- collect the data
- send the **compressed** file to us (eMail or ftp)

Linux and z/VM : monitor data collection

- logon to userid

```
cp monitor start
monwrite MONDCSS *monitor DISK mon dat a
monwstop
cp monitor stop
```

- send result file mon dat a to user FCONX for analysis

Tools

Performance tools Overview

A wide range of tools is available :

- System Tools (top, vmstat)
- rmfpms
- sysstat Package (collect a large set of Linux system data)
- VM Tools (FCON, Monitor...) discussed in later sessions
- OSA SNMP (capture OSA Express card data)
- Kernel Profiler (set of facilities for profiling the Linux kernel)
- Lockmeter (capture spin lock activity)

system tools : vmstat

- most important Linux system data at a glance
- low performance impact
- no setup necessary

- example :

```
[wolf@wolf wolf]$ vmstat 3
   procs                  memory    swap          io     system
  cpu
 r  b  w   swpd   free   buff  cache  si  so    bi    bo   in    cs  us  sy
  id
 0  0  0      0 663248  24204 196816   0   0    66    32  582   663   4   1
  95
 0  0  0      0 663248  24220 196816   0   0     0    56  555   493   2   1
  97
 0  0  0      0 663248  24220 196816   0   0     0     0  548   482   1   0
  99
 0  0  0      0 663248  24228 196816   0   0     0     4  549   493   1   1
  98
```

system tools : top

- even more system data at a glance
- data on a per-process basis
- high performance impact
- no setup necessary

- example :

```
[wolf@wolf wolf]$ top
      1:40pm  up  3:17, 11 users,  load average: 0.00, 0.02, 0.03
      134 processes: 127 sleeping, 6 running, 1 zombie, 0 stopped
      CPU states:  1.9% user,  1.7% system,  0.0% nice, 96.2% idle
      Mem:  1030464K av,  675028K used,  355436K free,       0K shrd,  156256K buff
      Swap: 1663160K av,       0K used, 1663160K free                 296068K
  cached

      PID USER     PRI  NI  SIZE  RSS SHARE STAT %CPU %MEM   TIME COMMAND
      1746 wolf      15   0 54772  14M 11600 R     1.1  1.4   0:00 kdeinit
      6147 wolf      15   0  1080 1080   840 R     0.9  0.1   0:00 top
```

rmfpms

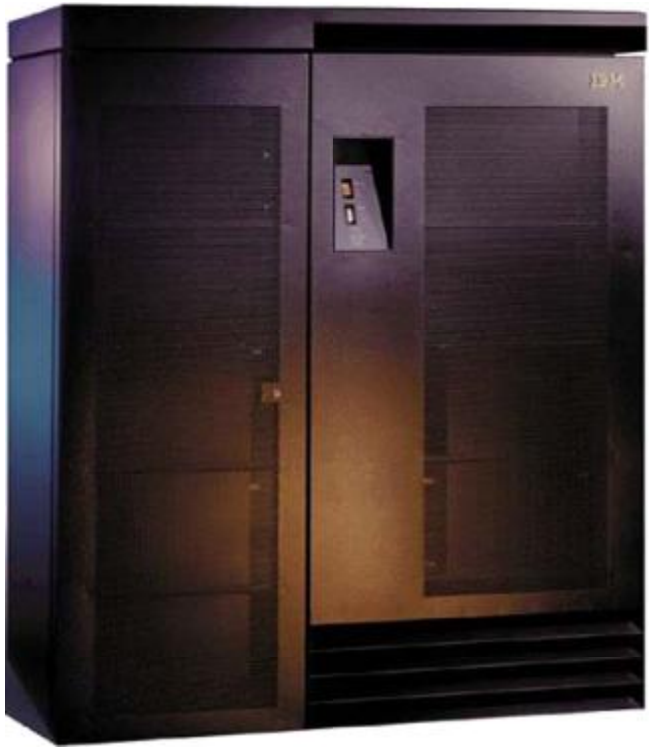
- long term data gathering
- xml over http interface
- independent from z/os; with z/os, you can also have an ldap interface to linux performance data
- modular architecture
- low performance impact
- see *http://www.ibm.com/servers/eserver/zseries/zos/rmf/rmfhtmls/pmweb/pmlin.htm for more info*

sysstat Package

- collection of linux tools to collect system data
- available as open source package at
  http://perso.wanadoo.fr/sebastien.godard/
- on Linux for zSeries only recompile needed
- latest stable version is 5.0.1
- contains multiple components :

sadc	Data gatherer stores data in binary file
sar	reporting tool reads binary file and converts it to readable output
mpstat	Processor utilization
iostat	IO utilization

Storage



DASD statistics

- contained in kernel since SUSE SLES8
- statistics collected by dasd driver
- can be easily switched on/off in proc filesystem

```
echo set on > /proc/dasd/statistics
echo set off > /proc/dasd/statistics
```

- setting off and back on resets all counters

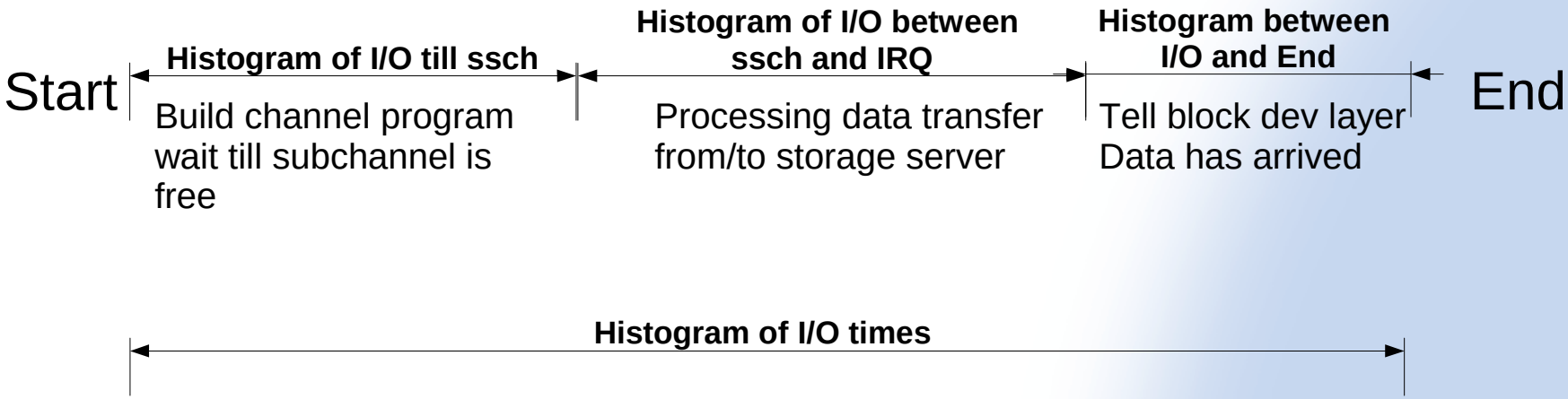
DASD statistics : example

/proc/dasd/statistics – Example

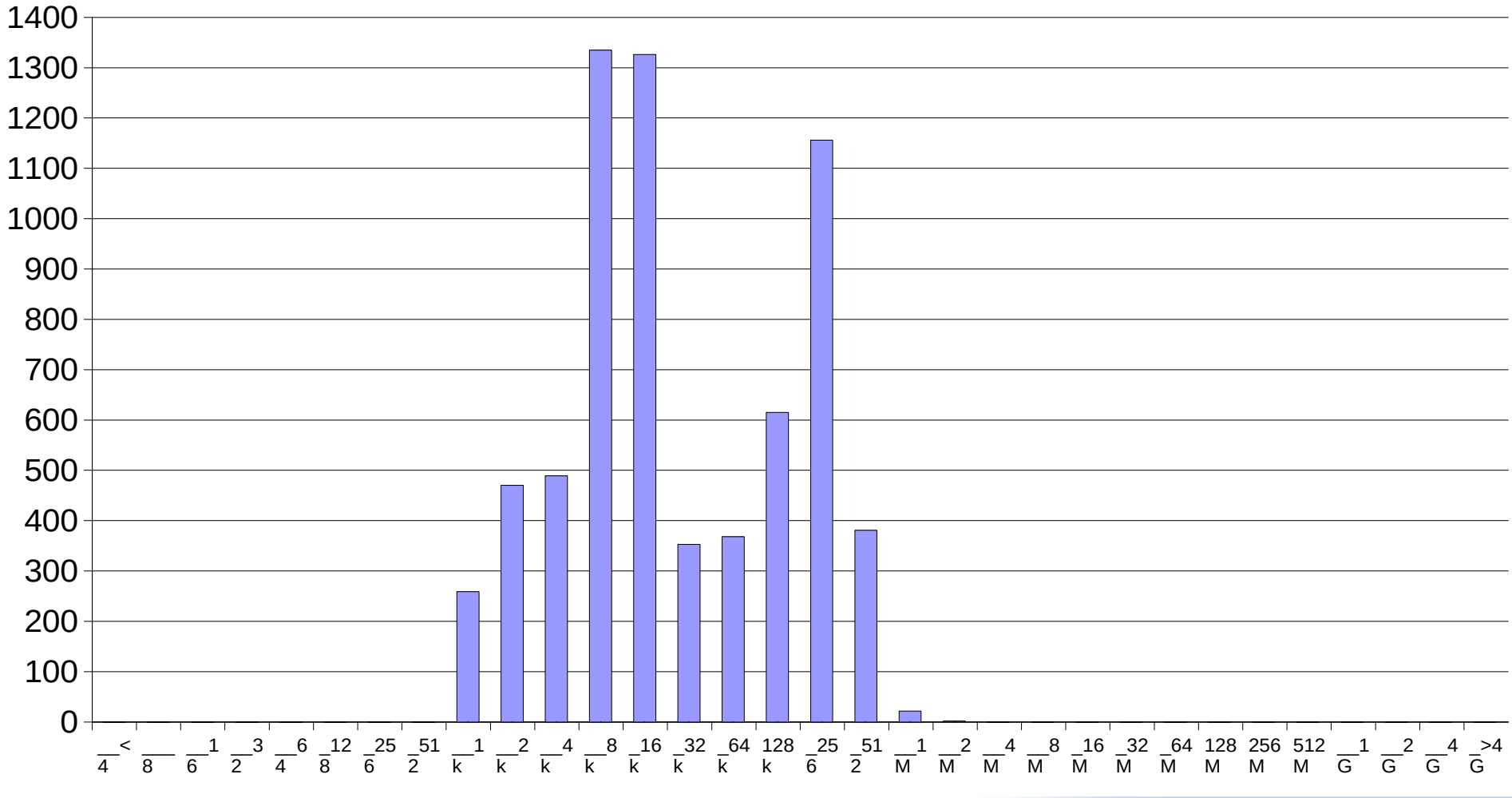
```
root@g73vm1:~# cat /proc/dasd/statistics
56881 dasd I/O requests
with 5270816 sectors(512B each)
   __<4     ___8     __16     __32     __64     _128     _256     _512     __1k     __2k     __4k     __8k     _16k     _32k     _64k     128k
   _256     _512     __1M     __2M     __4M     __8M     _16M     _32M     _64M     128M     256M     512M     __1G     __2G     __4G     _>4G
Histogram of sizes (512B secs)
      0        0     1039     4799     8102    36557     4475      292      195     1422        0        0        0        0        0        0
      0        0        0        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O times (microseconds)
      0        0        0        0        0        0        0        0        2        8      109     3244    25570    17480     7666     1248
   1390      153       11        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O times per sector
      0        0        0        0      176     4141    24084    15639     9506     2513      601      173       41        7        0        0
      0        0        0        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O time till ssch
      5        1        2        0        0        0        0        0        2        4      301    11527    25339    12278     5156     1759
    383      118        6        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O time between ssch and irq
      0        0        0        0        0        0        0        0     2584    23896    18720     5307     2325     2725     1217       62
     23       21        1        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O time between ssch and irq per sector
      0        0        0    21722    26243     3939     2184     1798      774      159       47       12        3        0        0        0
      0        0        0        0        0        0        0        0        0        0        0        0        0        0        0        0
Histogram of I/O time between irq and end
      7        0    43393    11341      457      179     1494        3        3        1        2        1        0        0        0        0
      0        0        0        0        0        0        0        0        0        0        0        0        0        0        0        0
# of req in chanq at enqueuing (1..32)
      8        3        4        5    56861        0        0        0        0        0        0        0        0        0        0        0
      0        0        0        0        0        0        0        0        0        0        0        0        0        0        0        0
```

DASD statistics : the details

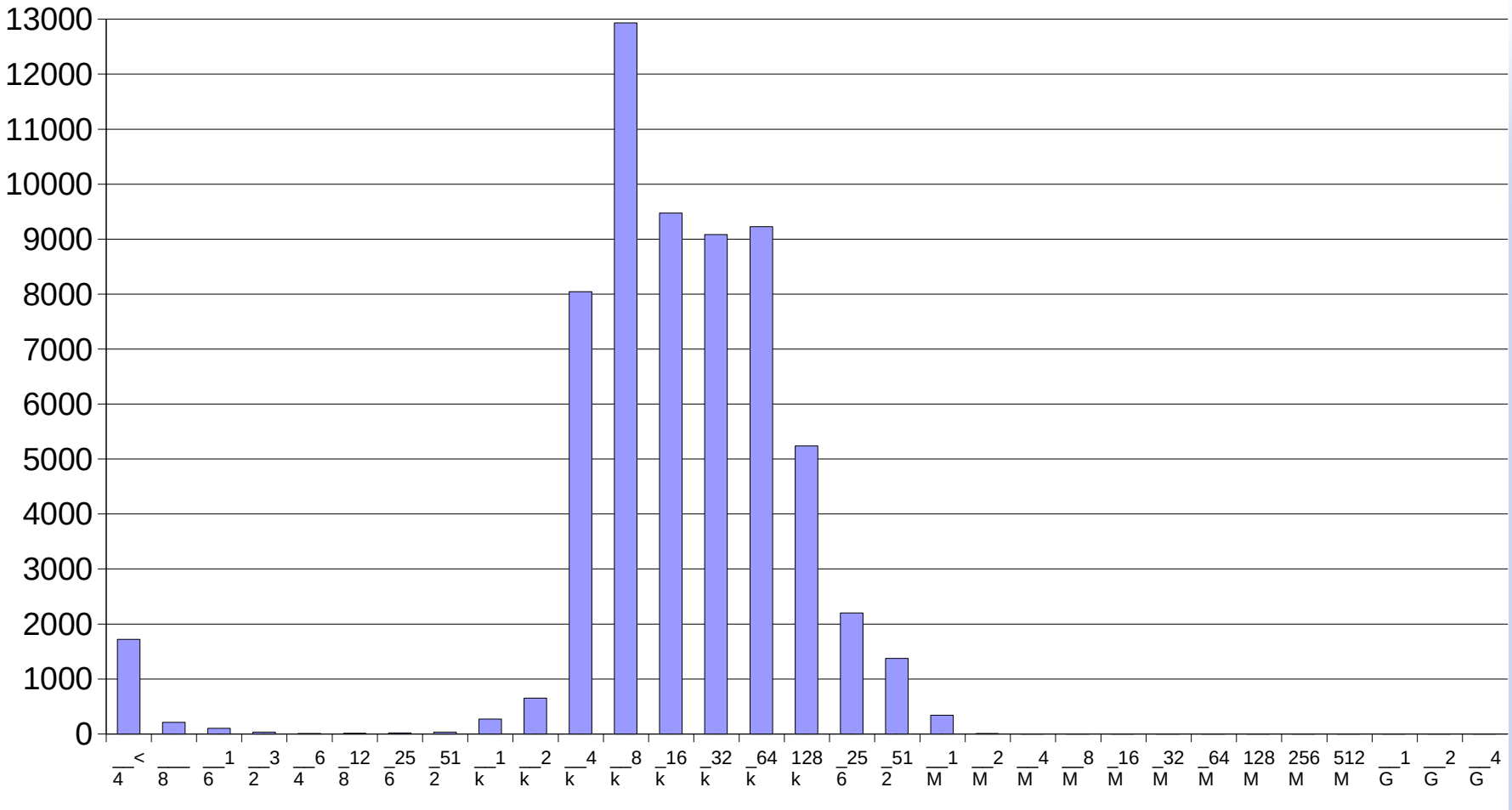
- collects statistics (mostly processing times) of IO operations
- each line represents a histogram of times for a certain operation
- operations split up into the following :



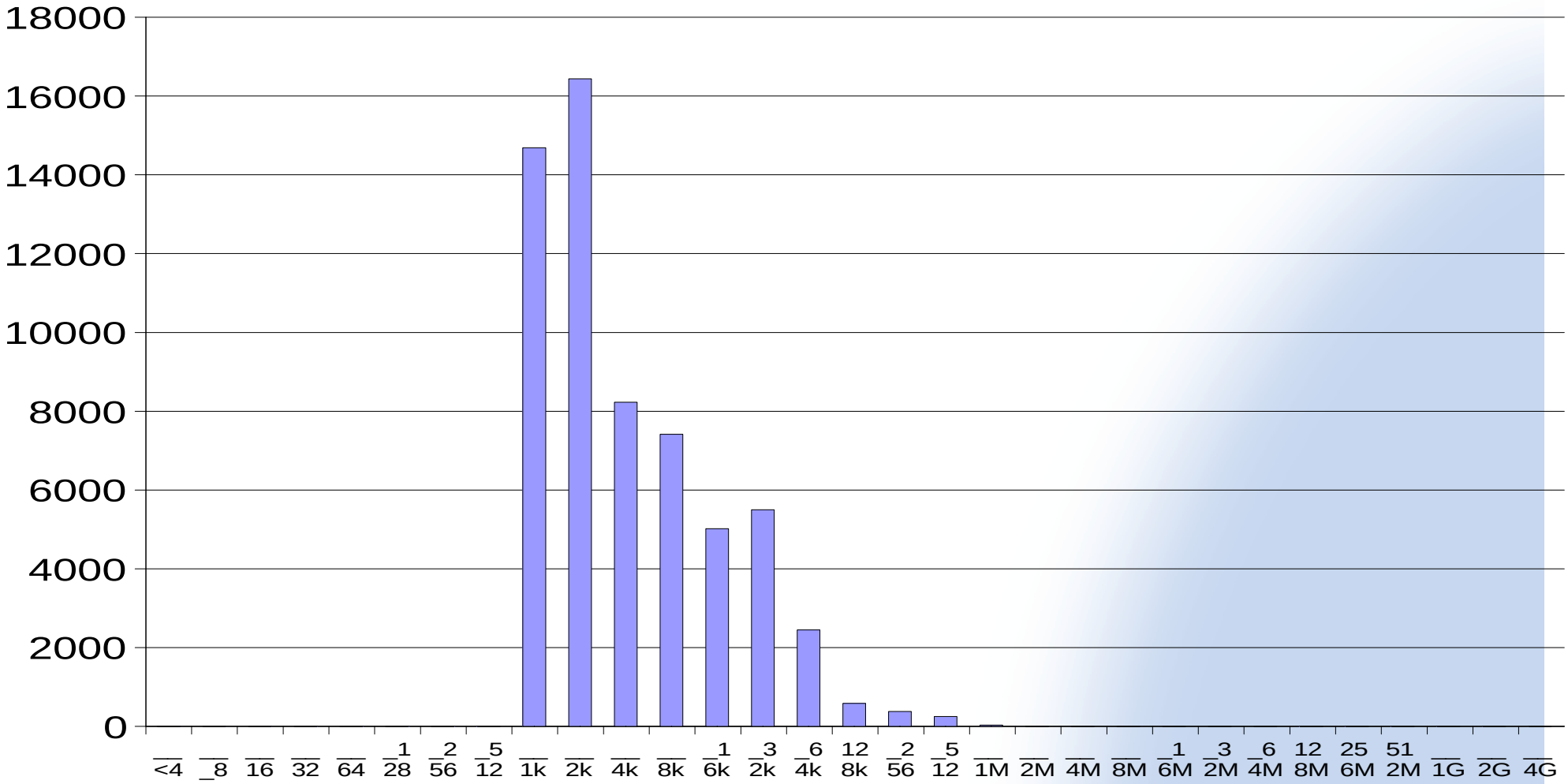
ext2, 8 Processes
Histogram of I/O times (microseconds)



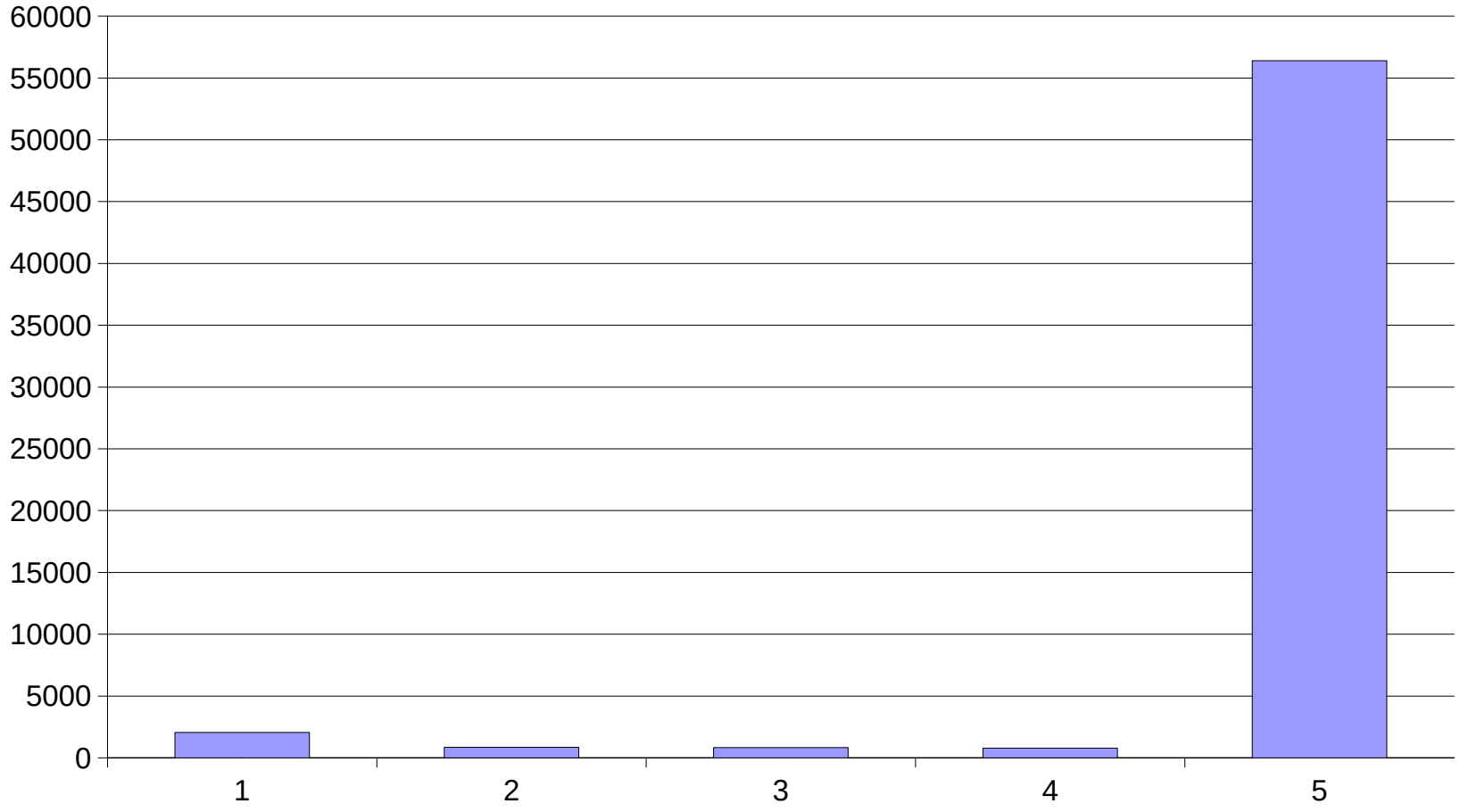
ext3, 16 Processes
Histogram of I/O time before SSCH (IOSQ)



Ext3, 16 Processes
Histogram of I/O time between SSCH and IRQ



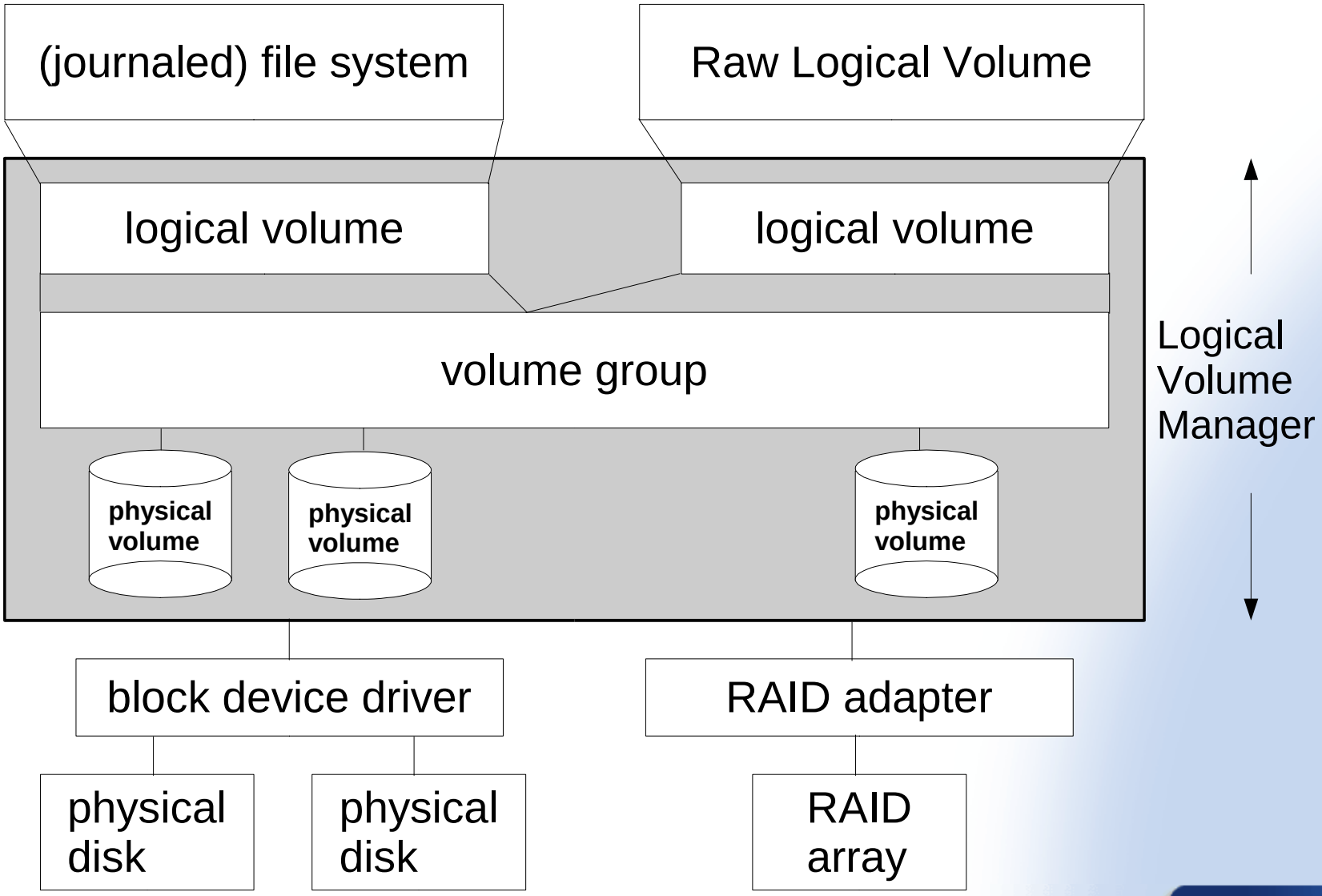
Ext3, 16 Processes
number of requests in subchannel-queue at enqueuing



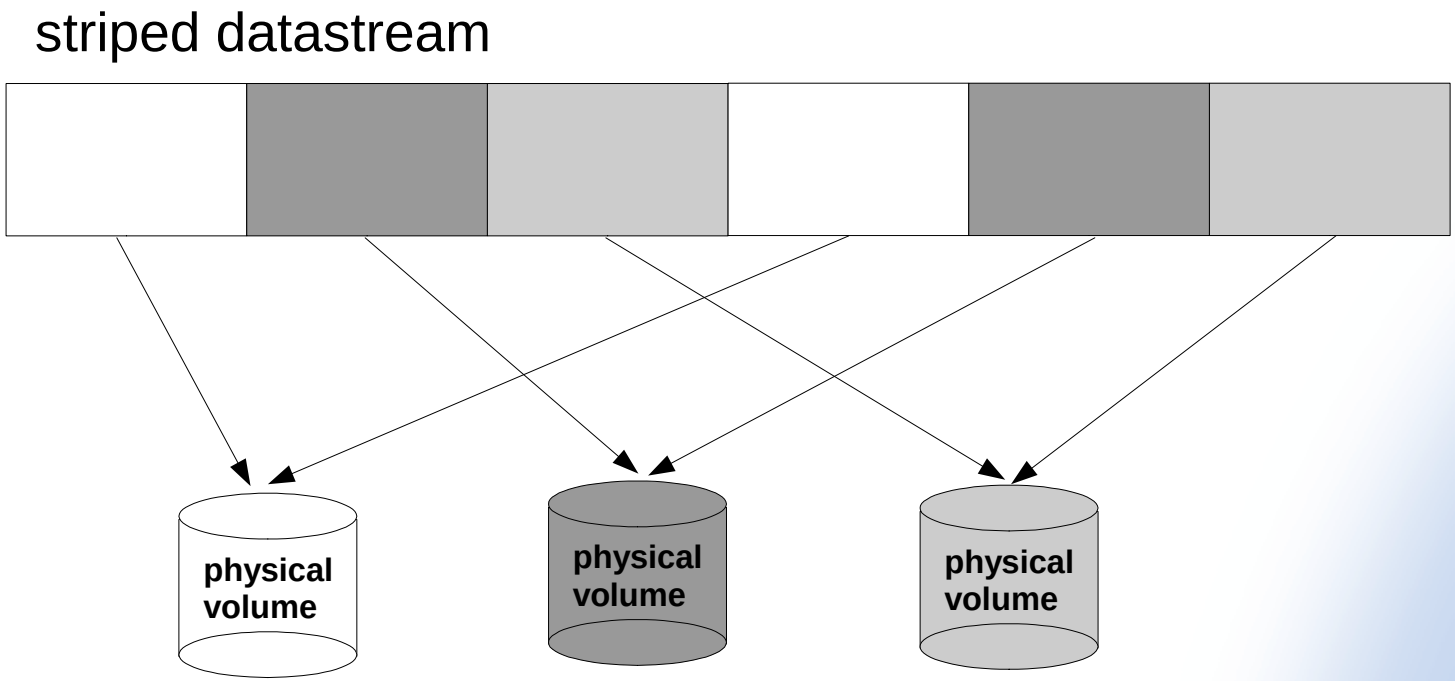
SISTINA Logical Volume Manager (LVM)

- Linux software raid with raid levels 0,1, 4 and 5
- excellent performance
- excellent flexibility (resizing, adding/removing disks)
- available in SLES7, SLES8, and RedHat RHEL 3
- on zSeries, support multipath and PAV (under z/VM)
- http://www.sistina.com/products_lvm.htm

LVM system structure



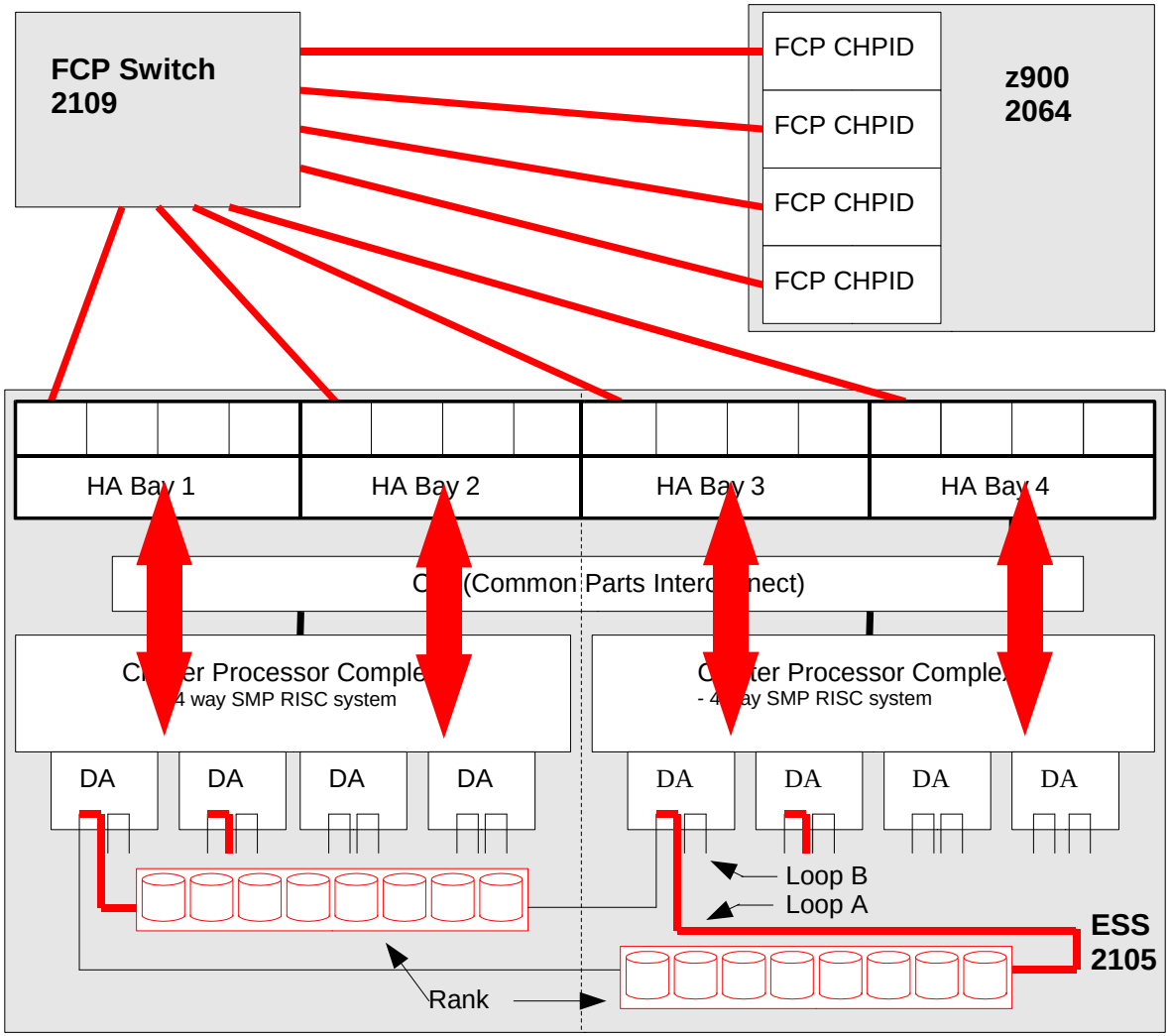
Improving disk performance with LVM



- With LVM and striping parallelism is achieved

ESS Architecture

Scenario: four CHPIDs



- **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)

ESS setup rules

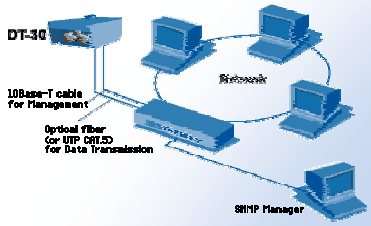
- spread your accesses over as much chpids as possible
- use as much host adapter bays as possible
- spread disks equally over as much ranks as possible

⇒ maximize parallel access to disks

Networking

OSA SNMP

- provides means to readout a lot of useful information from an OSA Express Card
- distributed as part of s390-tools
- grab the latest MIB from www.ibm.com/servers/resourcelink (needs resourcelink sign-in)
- open Library→Open Systems Adapter (OSA) Library→OSA-Express Direct SNMP MIB module



OSA SNMP : example

```
root@g73vm8:~# cat /proc/qeth
devnos (hex) CHPID     device     cardtype port chksum prio-q'ing rtr fsz C cnt
------------- --- --------- ------------- -- -- ---------- --- --- - ---
F118/F119/F11A x6A       eth0        OSD_100  0     no always_q_2  no 64k   128
FA0C/FA0D/FA0E x7A       eth1       OSD_1000  0     no always_q_2  no 64k   128
F006/F007/F008 x7C       eth2       OSD_1000  0     no always_q_2  no 64k   128
8209/820A/820B xFD       hsi5   HiperSockets  0     no always_q_2  no 40k   128
7000/7001/7002 x03      hsi12 GuestLAN Hiper  0     no always_q_2  no 40k   128
9000/9001/9002 x05      eth16  GuestLAN QDIO  0     no always_q_2  no 64k   128

root@g73vm8:~# /usr/bin/snmpwalk -Os localhost private ibmOSAExpChannelNumber
ibmOSAExpChannelNumber.6 =  Hex: 00 7C
ibmOSAExpChannelNumber.7 =  Hex: 00 7A
ibmOSAExpChannelNumber.8 =  Hex: 00 6A
```

OSA SNMP : example

```
root@g73vm8:~# /usr/bin/snmpwalk -Os localhost private ibmOSAExpChannelPCIBusUtil1Min
ibmOSAExpChannelPCIBusUtil1Min.6 = 7
ibmOSAExpChannelPCIBusUtil1Min.7 = 7
ibmOSAExpChannelPCIBusUtil1Min.8 = 12


root@g73vm8:~# /usr/bin/snmpwalk -Os localhost private ibmOSAExpChannelProcUtil1Min
ibmOSAExpChannelProcUtil1Min.6 = 2
ibmOSAExpChannelProcUtil1Min.7 = 1
ibmOSAExpChannelProcUtil1Min.8 = 2
```

OSA SNMP : example

```
root@g73vm8:~# /usr/bin/snmpwalk -Os localhost private ibmOSAExpChannelProcUtil1Min
ibmOSAExpChannelProcUtil1Min.6 = 2
ibmOSAExpChannelProcUtil1Min.7 = 1
ibmOSAExpChannelProcUtil1Min.8 = 20

root@g73vm8:~# /usr/bin/snmpwalk -Os localhost private ibmOSAExpChannelPCIBusUtil1Min
ibmOSAExpChannelPCIBusUtil1Min.6 = 7
ibmOSAExpChannelPCIBusUtil1Min.7 = 7
ibmOSAExpChannelPCIBusUtil1Min.8 = 21
```

Application performance

Adjust Java Performance

- Check if Just In Timer Compiler is enabled
  JIT is not enabled if compat library is not available
  or Environment variable java_compiler is set to none

```
java -version
```

```
java version "1.4.0"
Java(TM) 2 Runtime Environment, Standard Edition (build 1.4.0)
Classic VM (build 1.4.0, J2RE 1.4.0 IBM build cxia32140-20020917a (JIT enabled: jitc))
```

- Check if garbage collector adjustments could improve your performance

```
java -Xgcpolicy:optthruput <java class>
java -Xgcpolicy:optavgpause <java class>
```

- To monitor garbage collector (show statistics)

```
java -verbosegc <java class>
```

New features

- Channel measurement blocks
- z/VM monitor stream stage
  - Linux guest exports performance data into “APPLDATA monitor records”
  - performance data may be collected or display by z/VM performance monitoring tools
- virtual CPU timers

web resources

- Linux on zSeries Performance website

  http://www.ibm.com/developerworks/oss/linux390/perf_hints_tips.shtml
- z/VM Performance website

  http://www.vm.ibm.com/perf
- Linux on zSeries Performance redbook

  http://www.redbooks.ibm.com/redbooks/pdfs/sg246926.pdf

Questions ?