Problem Determination with Linux on System z

Dr. Holger Smolinski
IBM Germany Research & Development GmbH

2010-08-04
9279

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market. Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply. All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions. This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the Performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda – Part I

- Introduction
- Problem Description
- Troubleshooting First aid-kit
- System
    - dbginfo script, sos report
    - system z debug feature
    - sadc/sar
    - vmstat
- Disk
    - iostat
    - DASD/SCSI statistics
- Network
    - netstat
- Processes
    - top, ps
- Linux Documentation by IBM

Introductory Remarks

- Problem analysis looks straight forward on the charts but it might have taken weeks to get it done.

  A problem does not necessarily show up on the place of origin

- The more information is available, the sooner the problem can be solved, because gathering and submitting additional information again and again usually introduces delays.
- This presentation can only introduce some tools and how the tools can be used, comprehensive documentation on their capabilities is to be found in the documentation of the corresponding tool.
- Do not forget to update your systems

Describe the problem

- **Get as much information as possible about the circumstances:**
  - What is the problem ?
  - When did it happen ? - date and time, important to dig into logs
  - Where did it happen ? - one or more systems, production or test environment ?
  - Is this a first time occurrence ?
  - If occurred before:
    - how frequently does it occur ?
  - is there any pattern ?
  - Was anything changed recently ?
  - Is the problem reproducible ?
- **Write down as much information as possible about the problem !**

Describe the environment

- Machine Setup
  - Machine type (z10, z9, z990 ...)
  - Storage Server (ESS800, DS8000, other vendors models)
  - Storage attachment (FICON, ESCON, FCP, how many channels)
  - Network (OSA (type, mode), Hipersocket)

  ...
- Infrastructure setup
  - Clients
  - Other Computer Systems
  - Network topologies
  - Disk configuration
- Middleware setup
  - Databases, web servers, SAP, TSM, ...including version information

Trouble-Shooting First Aid kit

- Install packages required for debugging
  - s390-tools/s390-utils
    - dbginfo.sh
  - sysstat
    - sadc/sar
    - iostat
  - procps
    - vmstat, top, ps
  - net-tools
    - netstat
  - dump tools crash / lcrash
    - lcrash (lkcdutils) available with SLES9 and SLES10
    - crash available on SLES11
    - crash in all RHEL distributions

Trouble-Shooting First Aid kit (cont'd)

- Collect dbginfo.sh output
  - Proactively in healthy system
  - When problems occur – then compare with healthy system
- Collect system data
  - Always archive syslog (/var/log/messages)
  - Start sadc (System Activity Data Collection) service when appropriate
  - Collect z/VM MONWRITE Data if running under z/VM when appropriate

Trouble-Shooting First Aid kit (cont'd)

- When System hangs
  - Take a dump
    - Include System.map, Kerntypes (if available) and vmlinux file
  - See “Using the dump tools” book on http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26ddt02.pdf
- Enable extended tracing in /sys/kernel/debug/s390dbf for subsystem

Trouble-Shooting First Aid kit (cont'd)

- Attach comprehensive documentation to problem report:
  - Output file of dbginfo.sh, any (performance) reports or logs
  - z/VM MONWRITE data
    - Binary format, make sure, record size settings are correct.
    - For details see http://www.vm.ibm.com/perf/tips/collect.html
  - When opening a PMR upload documentation to directory associated to your PMR at
    - ftp://ecurep.ibm.com/, or
    - ftp://testcase.boulder.ibm.com/

  See Instructions: http://www.ibm.com/de/support/ecurep/other.html
- When opening a Bugzilla (bug tracker web application) at Distribution partner attach documentation to Bugzilla
- Think of global support structures

dbginfo script

- dbginfo.sh is a script to collect various system related files, for debugging purposes. It generates a tar-archive which can be attached to PMRs / Bugzilla entries
- part of the s390-tools package in SUSE and recent Red Hat distributions
  - **dbginfo.sh gets continuously improved by service and development**

  Can be downloaded at the developerWorks website directly

  http://www.ibm.com/developerworks/linux/linux390/s390-tools.html
- It is similar to the RedHat tool sosreport

```
root@larsson:~>  dbginfo.sh
Create target directory /tmp/DBGINFO-2009-04-15-22-06-
20-t6345057
Change to target directory /tmp/DBGINFO-2009-04-15-22-
06-20-t6345057
[...]
```

dbginfo script (cont'd)

- dbginfo.sh captures the following information:
  - /proc/[version, cpu, meminfo, slabinfo, modules, partitions, devices ...]
  - System z specific device driver information: /proc/s390dbf (RHEL 4 only) or /sys/kernel/debug/s390dbf
  - Kernel messages /var/log/messages
  - Reads configuration files in directory /etc/

    [ccwgroup.conf, modules.conf, fstab]
  - Uses several commands: ps, dmesg
  - Query setup scripts
    - lscss, lsdasd, lsqeth, lszfcp, lstape
  - And much more

dbginfo script (cont'd)

- dbginfo.sh captures the following information, when your system runs as guest under z/VM:
  - Release and service Level: `q cplevel`
  - Network setup: `q [lan, nic, vswitch, v osa]`
  - Storage setup: `q [set, v dasd, v fcp, q pav ...]`
  - Configuration/memory setup: `q [stor, v stor, xstore, cpus...]`
- In order to run the script properly, ensure that it is run as root user.
- When the system runs as z/VM guest, ensure that the guest has the appropriate privilege class authorities to issue the commands

sosreport

sosreport generates a compressed tarball of debugging information for the system it is run on that can be sent to technical support that will give them a more complete view of the overall system status.

```
root@larsson:~>  sosreport
sosreport (version 1.7)
[...]
This process may take a while to complete.
No changes will be made to your system.

Press ENTER to continue, or CTRL-C to quit.

Please enter your first initial and last name [h42lp27]: ABC
Please enter the case number that you are generating this report for:
DEF

Creating compressed archive...

Your sosreport has been generated and saved in:
  /tmp/sosreport-ABC-427338-6e8879.tar.bz2
[...]
```

(supportconfig from SLES similar)

System z debug feature

- System z specific driver tracing environment
- Uses wraparound memory buffers
- Available in live system and in system dumps
- Debug filesystem must be mounted (except RHEL 4) :
  - `mount -t debugfs /sys/debug /sys/kernel/debug`
- Views: hex_ascii, sprintf, flush and pages
- Trace levels between 0 <-> 6 (lowest-highest) default: 2
- set/change trace level via `'echo 2 >level'`
- Flush s390dbf: `'echo - >flush'`
- Increase buffer size: `'echo 10 >pages'`

```
==> /sys/kernel/debug/s390dbf/qeth_trace/level <==
==> /sys/kernel/debug/s390dbf/qeth_trace/hex_ascii <==
01132180673:456679 0 - 00 788606ba  4e 4f 4d 4d 20 20 20 38 | NOMM   8
01132180673:456810 0 - 00 788606ba  4e 4f 4d 4d 20 20 20 38 | NOMM   8
01132180673:456936 0 - 00 788606ba  4e 4f 4d 4d 20 20 20 38 | NOMM   8
```

SADC/SAR

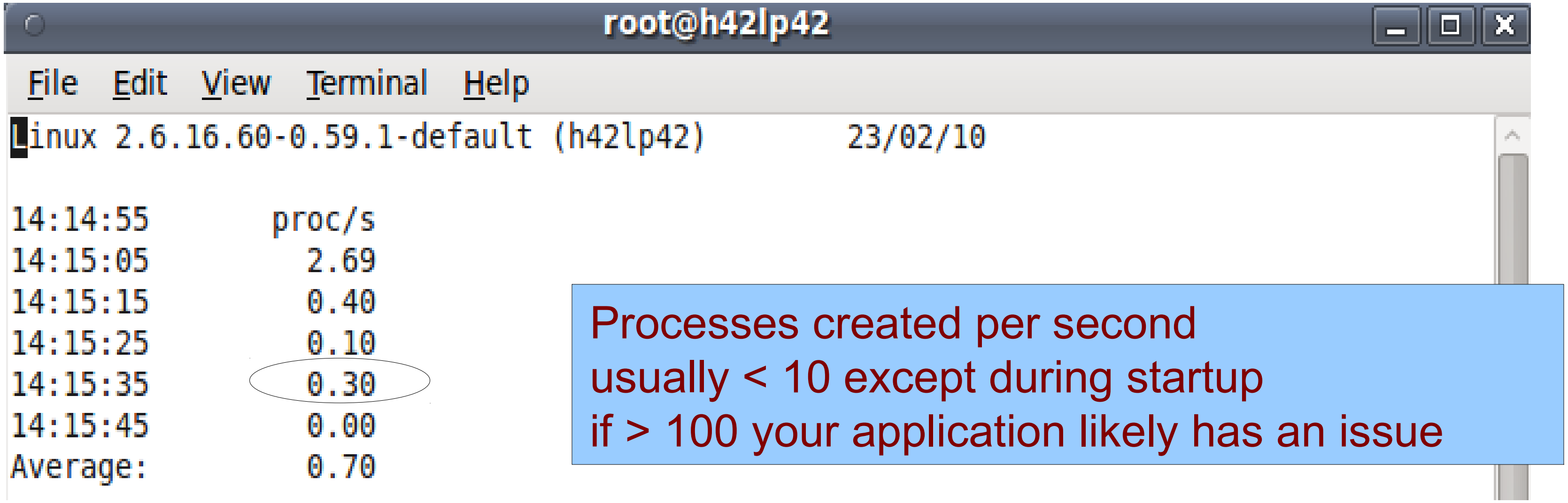
- Capture Linux performance data with sadc/sar
  - CPU utilization
  - Disk I/O overview and on device level
  - Network I/O and errors on device level
  - Memory usage/Swapping
  - … and much more
  - Reports statistics data over time and creates average values for each item
- SADC example (for more see man sadc)
  - **S**ystem **A**ctivity **D**ata **C**ollector (sadc) --> data gatherer
  - /usr/lib64/sa/sadc [options] [interval [count]] **[binary outfile]**
  - /usr/lib64/sa/sadc 10 20 sadc_outfile

SADC/SAR (cont'd)

- /usr/lib64/sa/sadc **-d** 10 sadc_outfile
- -d option: statistics for disk
- Should be started as a service during system start

* SAR example (for more see man sar)
  - **S**ystem **A**ctivity **R**eport (sar) command --> reporting tool
  - sar **-A**
    - -A option: reports all the collected statistics
    - sar -A -f sadc_outfile >sar_outfile
- Please include the binary sadc data and sar -A output when submitting SADC/SAR information to IBM support

Processes created

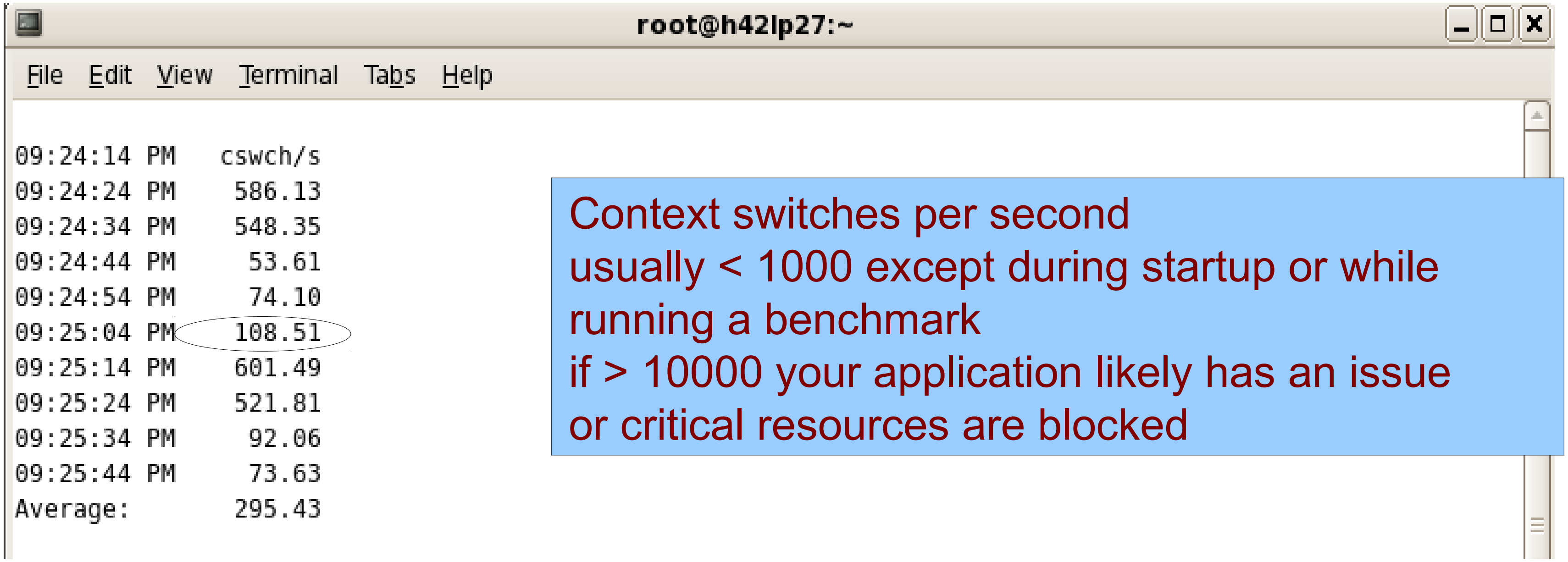


```
Linux 2.6.16.60-0.59.1-default (h42lp42)        23/02/10

14:14:55        proc/s
14:15:05          2.69
14:15:15          0.40
14:15:25          0.10
14:15:35          0.30
14:15:45          0.00
Average:          0.70
```

Processes created per second
usually < 10 except during startup
if > 100 your application likely has an issue

Context Switch Rate



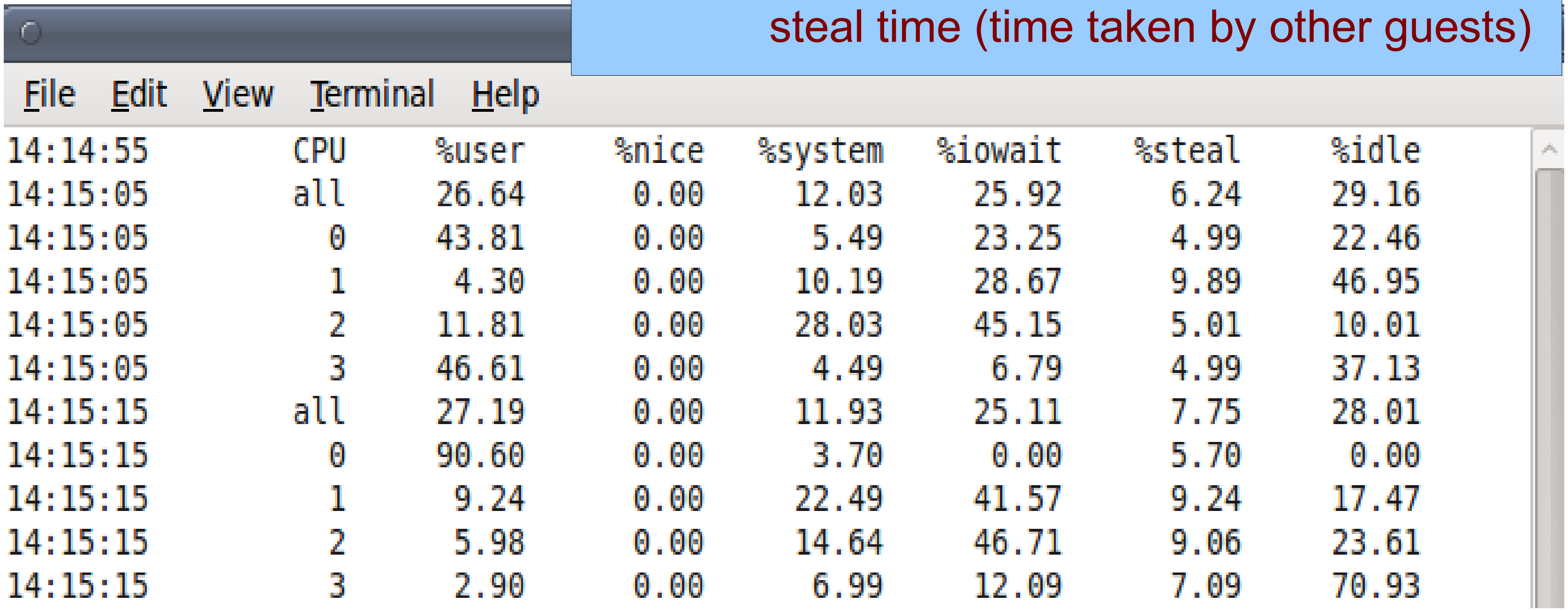
```
09:24:14 PM    cswch/s
09:24:24 PM     586.13
09:24:34 PM     548.35
09:24:44 PM      53.61
09:24:54 PM      74.10
09:25:04 PM     108.51
09:25:14 PM     601.49
09:25:24 PM     521.81
09:25:34 PM      92.06
09:25:44 PM      73.63
Average:        295.43
```

Context switches per second
usually < 1000 except during startup or while running a benchmark
if > 10000 your application likely has an issue or critical resources are blocked

CPU utilization

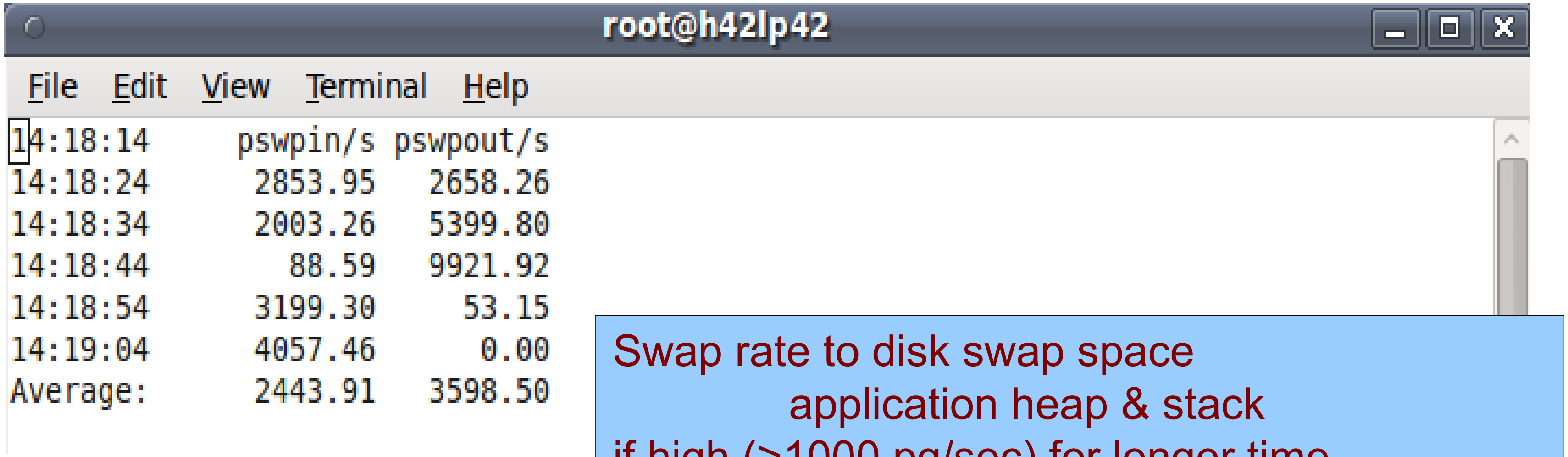
Per CPU values:
watch out for

- system time (kernel time)
- iowait time (slow I/O subsystem)
- steal time (time taken by other guests)



```
14:14:55       CPU     %user     %nice   %system   %iowait    %steal     %idle
14:15:05       all     26.64      0.00     12.03     25.92      6.24     29.16
14:15:05         0     43.81      0.00      5.49     23.25      4.99     22.46
14:15:05         1      4.30      0.00     10.19     28.67      9.89     46.95
14:15:05         2     11.81      0.00     28.03     45.15      5.01     10.01
14:15:05         3     46.61      0.00      4.49      6.79      4.99     37.13
14:15:15       all     27.19      0.00     11.93     25.11      7.75     28.01
14:15:15         0     90.60      0.00      3.70      0.00      5.70      0.00
14:15:15         1      9.24      0.00     22.49     41.57      9.24     17.47
14:15:15         2      5.98      0.00     14.64     46.71      9.06     23.61
14:15:15         3      2.90      0.00      6.99     12.09      7.09     70.93
```

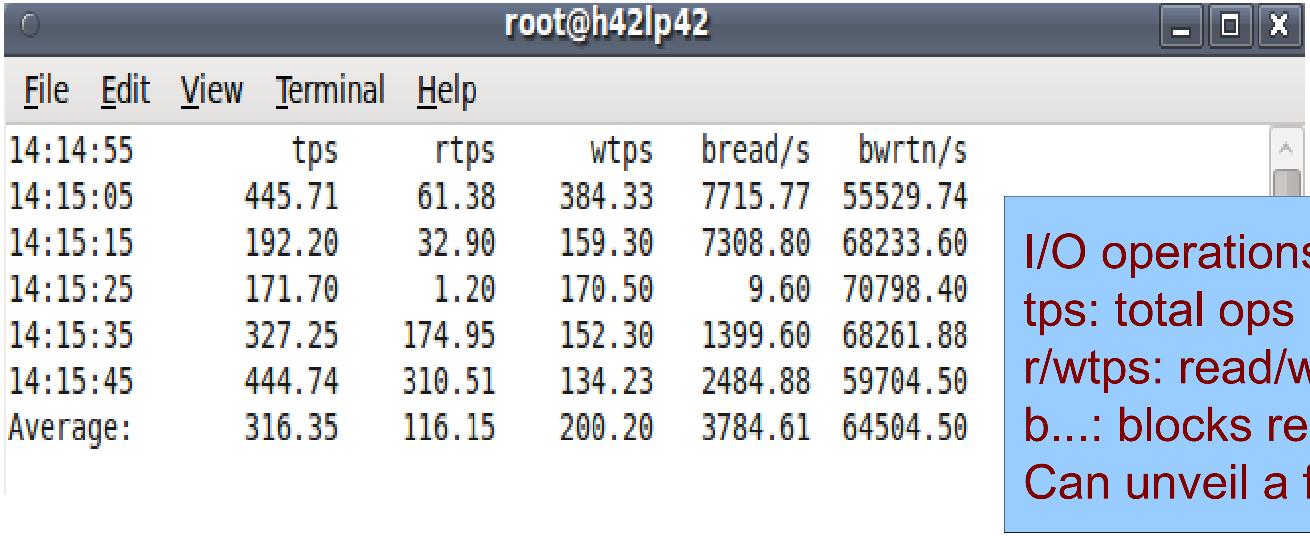
Swap rate



```
14:18:14      pswpin/s pswpout/s
14:18:24       2853.95   2658.26
14:18:34       2003.26   5399.80
14:18:44         88.59   9921.92
14:18:54       3199.30     53.15
14:19:04       4057.46      0.00
Average:       2443.91   3598.50
```

Swap rate to disk swap space
application heap & stack
if high (>1000 pg/sec) for longer time
you are likely short on memory
or your application has a memory leak

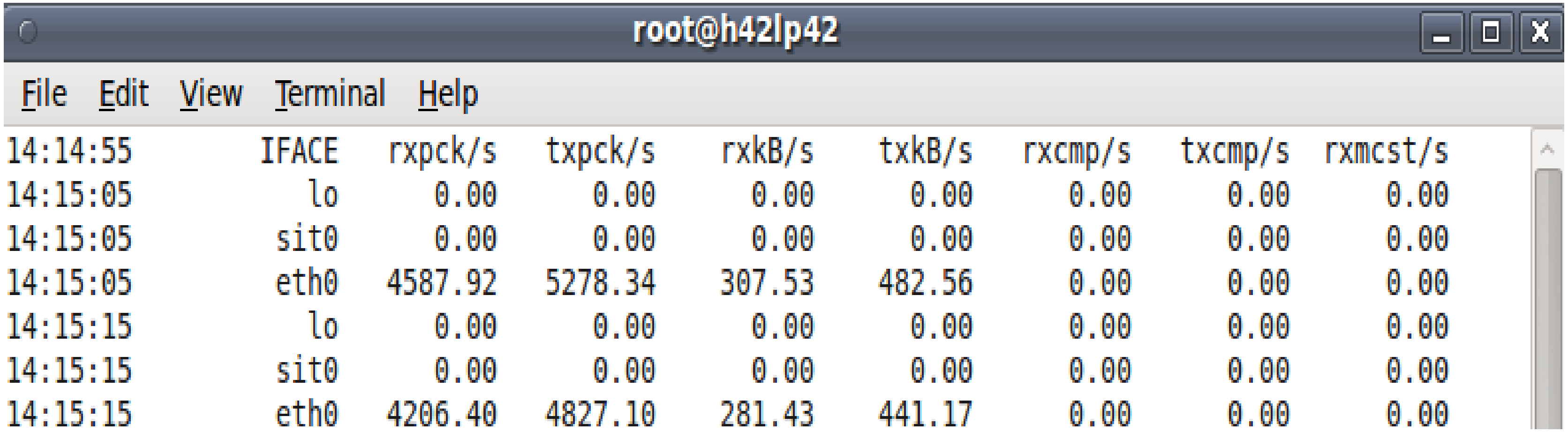
I/O rates



```
14:14:55          tps      rtps      wtps   bread/s   bwrtn/s
14:15:05       445.71     61.38    384.33   7715.77  55529.74
14:15:15       192.20     32.90    159.30   7308.80  68233.60
14:15:25       171.70      1.20    170.50      9.60  70798.40
14:15:35       327.25    174.95    152.30   1399.60  68261.88
14:15:45       444.74    310.51    134.23   2484.88  59704.50
Average:       316.35    116.15    200.20   3784.61  64504.50
```

I/O operations per second
tps: total ops
r/wtps: read/write operations
b...: blocks read/written
Can unveil a fabric problem...

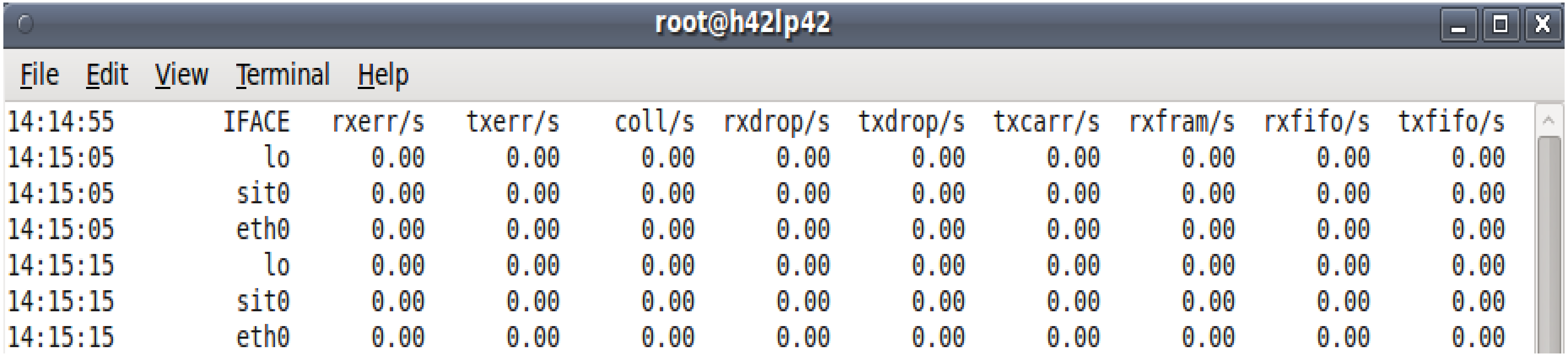
Networking data (1)



```
14:14:55        IFACE   rxpck/s   txpck/s    rxkB/s    txkB/s   rxcmp/s   txcmp/s  rxmcst/s
14:15:05           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:05         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:05         eth0   4587.92   5278.34    307.53    482.56      0.00      0.00      0.00
14:15:15           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:15         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:15         eth0   4206.40   4827.10    281.43    441.17      0.00      0.00      0.00
```

- Rates of successful transmits/receives
  - Per interface
  - Packets and bytes

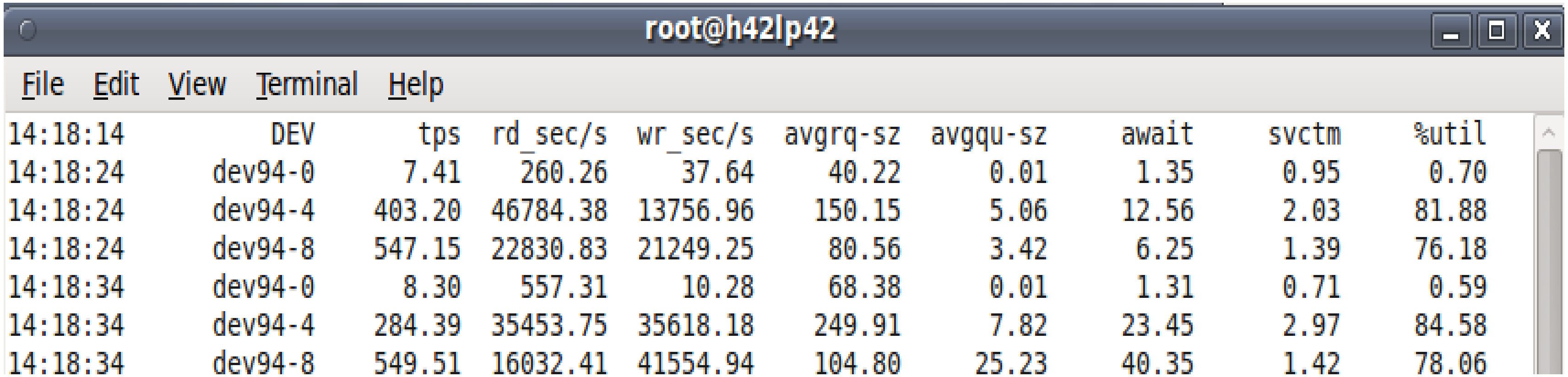
Networking data (2)



```
14:14:55        IFACE   rxerr/s   txerr/s    coll/s  rxdrop/s  txdrop/s  txcarr/s  rxfram/s  rxfifo/s  txfifo/s
14:15:05           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:05         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:05         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:15           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:15         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
14:15:15         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
```

- Rates of unsuccessful transmits/receives
  - Per interface
  - rx/tx Errors
  - Dropped packets
    - Inbound: potential memory shortage

Disk I/O rates

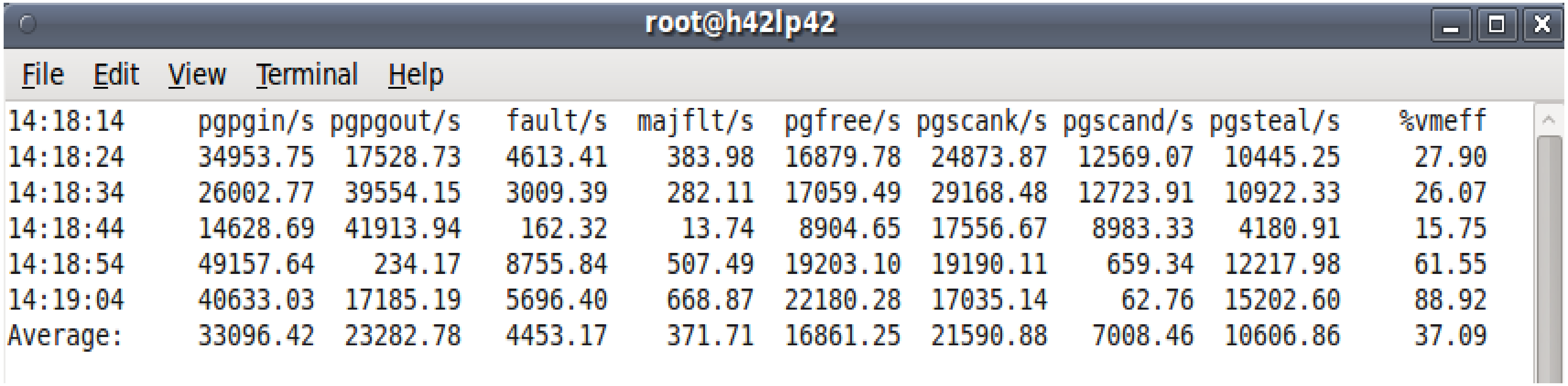


```
14:18:14        DEV      tps  rd_sec/s  wr_sec/s  avgrq-sz  avgqu-sz     await     svctm     %util
14:18:24    dev94-0     7.41    260.26     37.64     40.22      0.01      1.35      0.95      0.70
14:18:24    dev94-4   403.20  46784.38  13756.96    150.15      5.06     12.56      2.03     81.88
14:18:24    dev94-8   547.15  22830.83  21249.25     80.56      3.42      6.25      1.39     76.18
14:18:34    dev94-0     8.30    557.31     10.28     68.38      0.01      1.31      0.71      0.59
14:18:34    dev94-4   284.39  35453.75  35618.18    249.91      7.82     23.45      2.97     84.58
14:18:34    dev94-8   549.51  16032.41  41554.94    104.80     25.23     40.35      1.42     78.06
```

read/write operations
- per I/O device
- tps: transactions
- rd/wr_secs: sectors

is your I/O balanced?
Maybe you should stripe your LVs

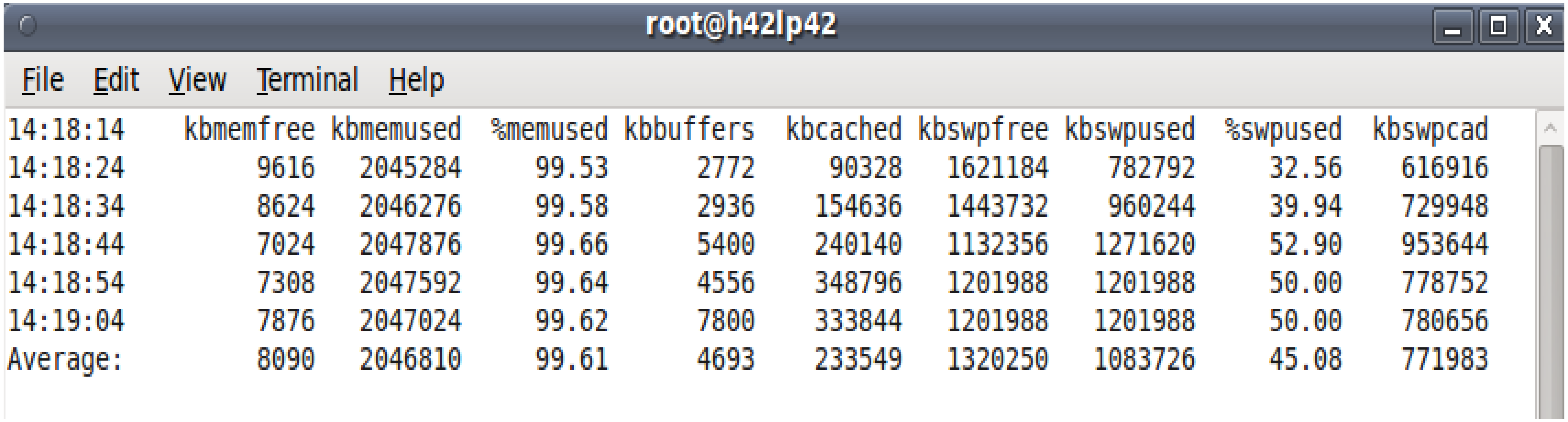
Disk I/O paging statistics



```
14:18:14      pgpgin/s pgpgout/s   fault/s  majflt/s  pgfree/s pgscank/s pgscand/s pgsteal/s    %vmeff
14:18:24      34953.75  17528.73   4613.41    383.98  16879.78  24873.87  12569.07  10445.25     27.90
14:18:34      26002.77  39554.15   3009.39    282.11  17059.49  29168.48  12723.91  10922.33     26.07
14:18:44      14628.69  41913.94    162.32     13.74   8904.65  17556.67   8983.33   4180.91     15.75
14:18:54      49157.64    234.17   8755.84    507.49  19203.10  19190.11    659.34  12217.98     61.55
14:19:04      40633.03  17185.19   5696.40    668.87  22180.28  17035.14     62.76  15202.60     88.92
Average:      33096.42  23282.78   4453.17    371.71  16861.25  21590.88   7008.46  10606.86     37.09
```

Watch for major page faults, if high,
short on available memory
I/O overhead - consumes a lot of CPU time

Memory statistics



```
14:18:14     kbmemfree kbmemused  %memused kbbuffers  kbcached kbswpfree kbswpused  %swpused  kbswpcad
14:18:24          9616   2045284     99.53      2772     90328   1621184    782792     32.56    616916
14:18:34          8624   2046276     99.58      2936    154636   1443732    960244     39.94    729948
14:18:44          7024   2047876     99.66      5400    240140   1132356   1271620     52.90    953644
14:18:54          7308   2047592     99.64      4556    348796   1201988   1201988     50.00    778752
14:19:04          7876   2047024     99.62      7800    333844   1201988   1201988     50.00    780656
Average:          8090   2046810     99.61      4693    233549   1320250   1083726     45.08    771983
```

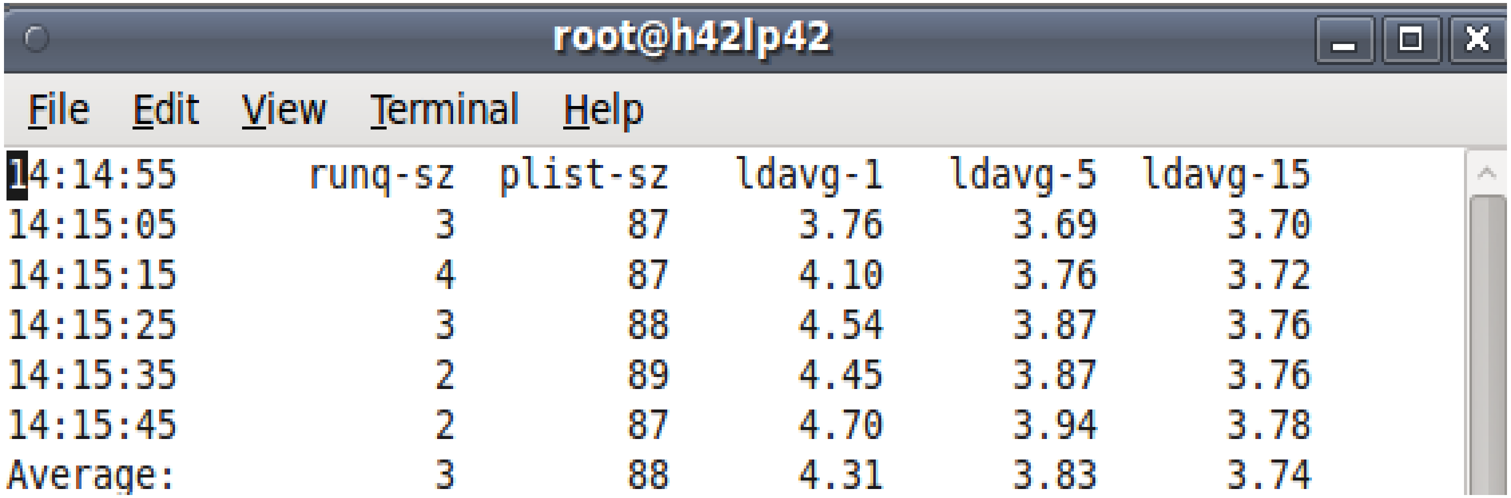
Watch

%memused and kbmemfree: short on available memory

kbswapfree: if not swapped but short on memory

the problem is not heap & stack but I/O buffers

System Load



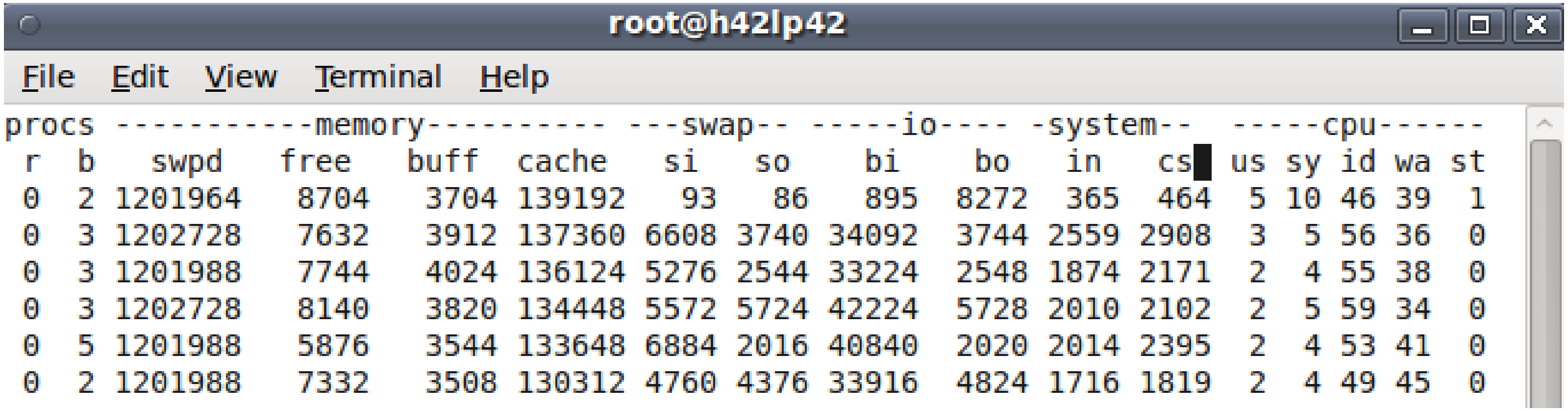
```
14:14:55      runq-sz  plist-sz   ldavg-1   ldavg-5  ldavg-15
14:15:05            3        87      3.76      3.69      3.70
14:15:15            4        87      4.10      3.76      3.72
14:15:25            3        88      4.54      3.87      3.76
14:15:35            2        89      4.45      3.87      3.76
14:15:45            2        87      4.70      3.94      3.78
Average:            3        88      4.31      3.83      3.74
```

Watch runqueue size snapshots runq-sz
Many (>5) processes on runqueue are critical
Blocked by shortage on available CPUs
Being bound in IOWAIT state
Load average is runqueue length average in 1/5/15 minutes

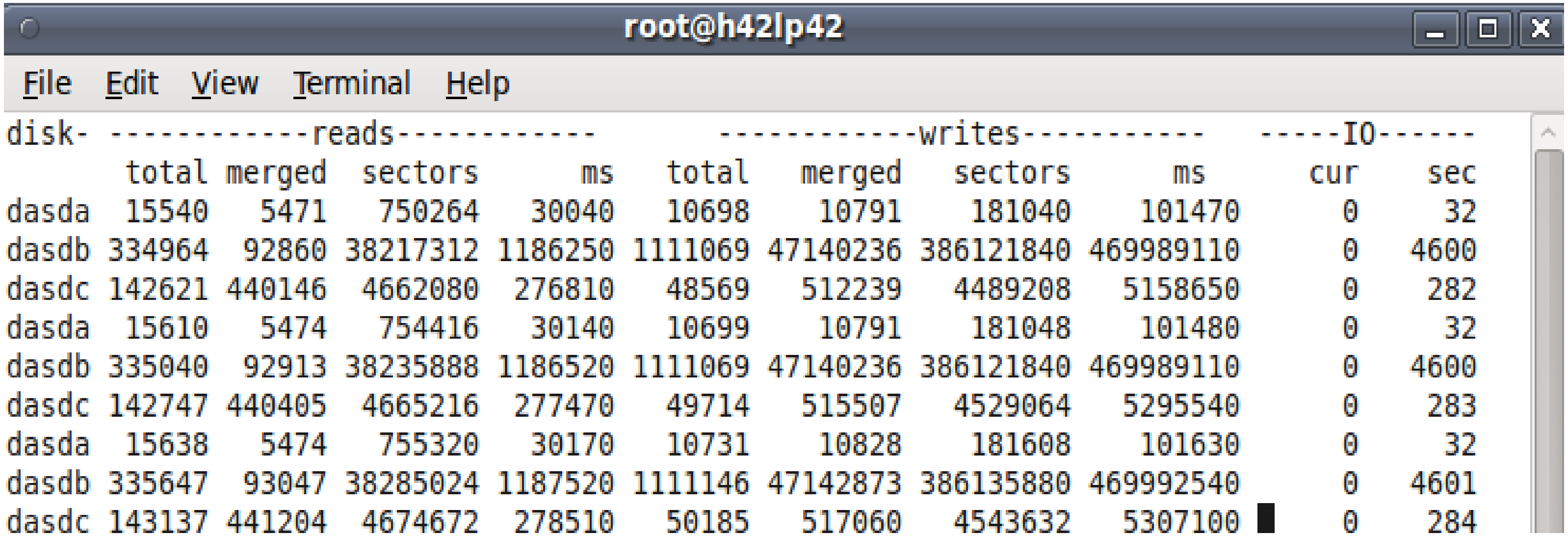
vmstat

- vmstat reports information about
  - Data per time interval
  - CPU utilization
  - Disk I/O
  - Memory usage/Swapping
- vmstat example (for more see man vmstat)
  - vmstat **[delay [count]]**
  - vmstat 10 5
  - vmstat **-d**
  - -d option: statistics for disks

vmstat (cont'd)



```
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  2 1201964   8704   3704 139192   93   86   895  8272  365  464  5 10 46 39  1
 0  3 1202728   7632   3912 137360 6608 3740 34092  3744 2559 2908  3  5 56 36  0
 0  3 1201988   7744   4024 136124 5276 2544 33224  2548 1874 2171  2  4 55 38  0
 0  3 1202728   8140   3820 134448 5572 5724 42224  5728 2010 2102  2  5 59 34  0
 0  5 1201988   5876   3544 133648 6884 2016 40840  2020 2014 2395  2  4 53 41  0
 0  2 1201988   7332   3508 130312 4760 4376 33916  4824 1716 1819  2  4 49 45  0
```



```
disk- ------------reads------------ ------------writes----------- -----IO------
       total merged sectors      ms  total  merged   sectors      ms    cur    sec
dasda  15540   5471  750264   30040  10698   10791    181040  101470      0     32
dasdb 334964  92860 38217312 1186250 1111069 47140236 386121840 469989110      0   4600
dasdc 142621 440146 4662080  276810  48569  512239   4489208  5158650      0    282
dasda  15610   5474  754416   30140  10699   10791    181048  101480      0     32
dasdb 335040  92913 38235888 1186520 1111069 47140236 386121840 469989110      0   4600
dasdc 142747 440405 4665216  277470  49714  515507   4529064  5295540      0    283
dasda  15638   5474  755320   30170  10731   10828    181608  101630      0     32
dasdb 335647  93047 38285024 1187520 1111146 47142873 386135880 469992540      0   4601
dasdc 143137 441204 4674672  278510  50185  517060   4543632  5307100      0    284
```

iostat

- iostat shows
  - Device queue information
  - Service times
- IOSTAT example (for more see man iostat)
  - **iostat** command --> I/O utilization
  - iostat [options] [interval [count]]
  - iostat ALL -kx **-->** Analyse cpu and io related performance data
  - iostat -c **-->** Analyse only cpu related performance data
  - iostat -dkx **-->** Analyse io related performance data for all disks

iostat (cont'd)

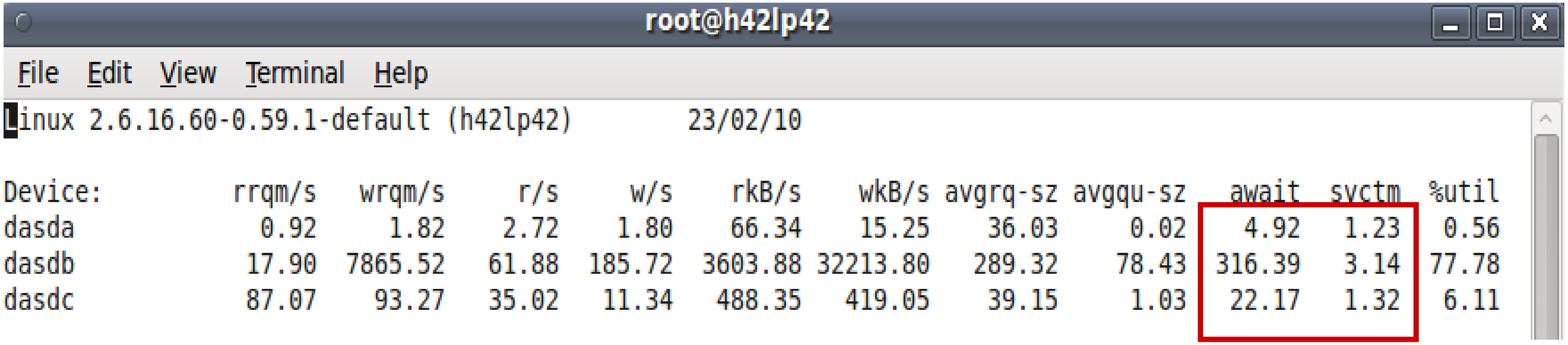
- iostat shows averaged performance data per device
  - Sample *iostat -dkx* output:

– Especially watch queue size and await/svctm

avgqu-sz: average length of queue, how many i/o requests are not dispatched

await (in millisec.): average time for i/o requests issued to the device to be serviced (total time of an i/o, incl. Time on queue).

svctm (in millisec.): average service time for i/o requests that were issued to the device.



```
Linux 2.6.16.60-0.59.1-default (h42lp42)        23/02/10

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
dasda             0.92     1.82    2.72    1.80    66.34    15.25    36.03     0.02    4.92   1.23   0.56
dasdb            17.90  7865.52   61.88  185.72  3603.88 32213.80   289.32    78.43  316.39   3.14  77.78
dasdc            87.07    93.27   35.02   11.34   488.35   419.05    39.15     1.03   22.17   1.32   6.11
```

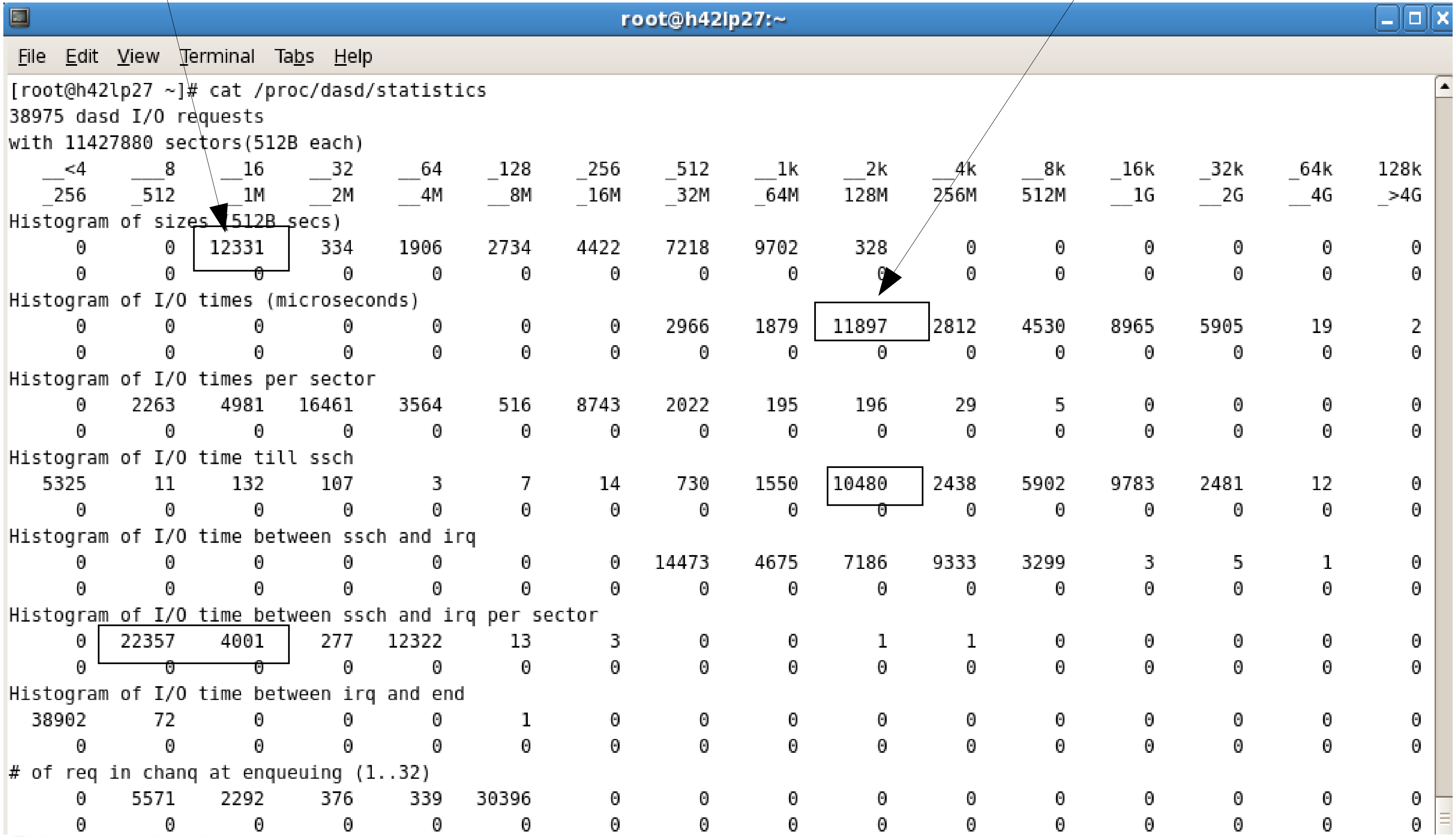
DASD statistics

- DASD statistics records (mostly processing time) of I/O operations of a specific period as statistic data
- Capture DASD statistics data
  - Activate via '`echo set on > /proc/dasd/statistics`'
  - Summarized histogram information available in /proc/dasd/statistics
  - '`cat /proc/dasd/statistics`'
  - Deactivate via '`echo set off > /proc/dasd/statistics`'
  - tunedasd -P /dev/dasda --> for individual DASD

DASD statistics (cont'd)

4 kb <= request size <= 8 kb

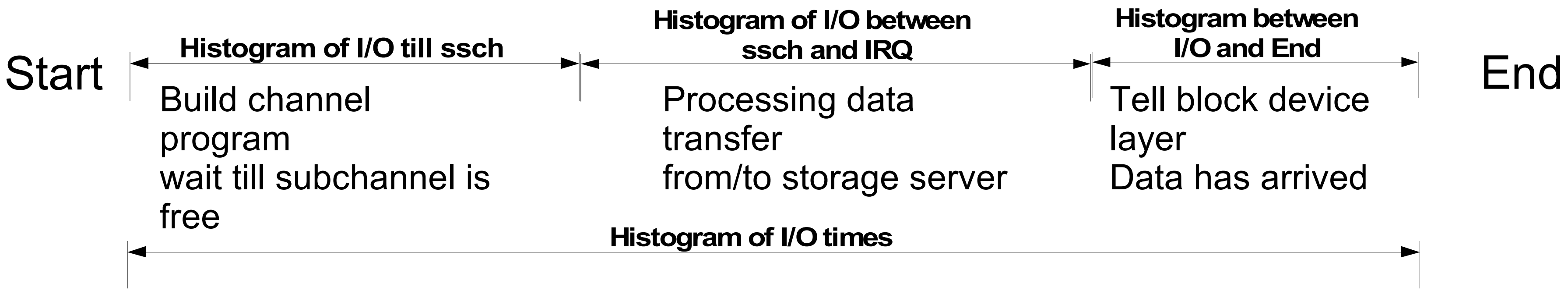
1 ms <= response time <= 2 ms



```
[root@h42lp27 ~]# cat /proc/dasd/statistics
38975 dasd I/O requests
with 11427880 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   12331     334    1906    2734    4422    7218    9702     328       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    2966    1879   11897    2812    4530    8965    5905      19       2
      0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Histogram of I/O times per sector
      0    2263    4981   16461    3564     516    8743    2022     195     196      29       5       0       0       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
   5325      11     132     107       3       7      14     730    1550   10480    2438    5902    9783    2481      12       0
      0       0       0       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   14473    4675    7186    9333    3299       3       5       1       0
      0       0       0       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   22357    4001     277   12322      13       3       0       0       1       1       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 time between irq and end
  38902      72       0       0       0       1       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
# of req in chanq at enqueuing (1..32)
      0    5571    2292     376     339   30396       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 (cont'd)

- DASD statistics decomposition
  - Each line represents a histogram of times for a certain operation
  - Operations split up into the following :



Start

Histogram of I/O till ssch

Build channel program
wait till subchannel is free

Histogram of I/O between ssch and IRQ

Processing data transfer
from/to storage server

Histogram between I/O and End

Tell block device layer
Data has arrived

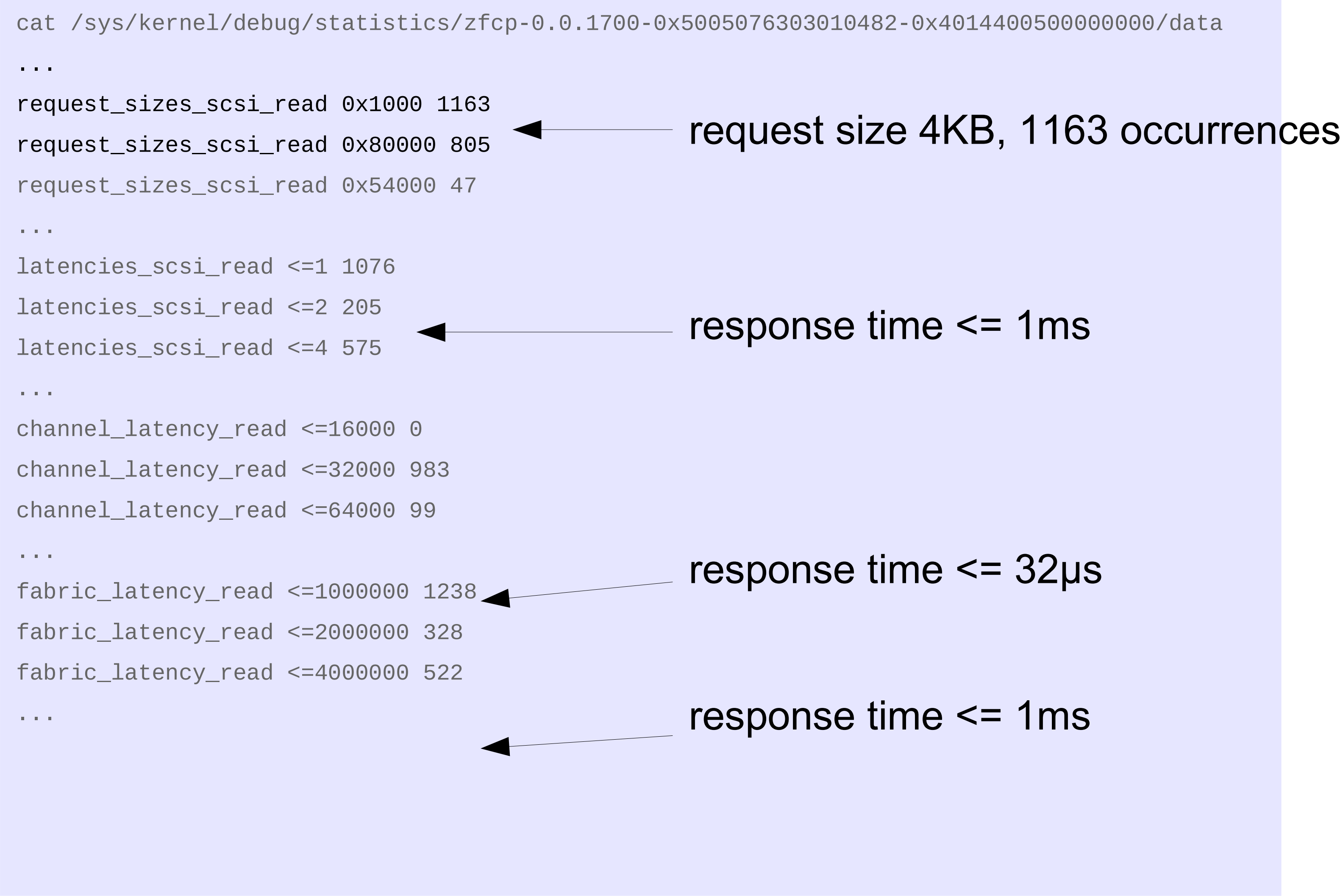
End

Histogram of I/O times

SCSI statistics (SLES9 and SLES10 only)

- Detailed latency information
- Collects statistics of I/O operations on FCP devices on request base, separate for read/write
- CONFIG_STATISTICS=y must be set in the kernel config file
- If debugfs is mounted at /sys/kernel/debug/, all the statistics data collected can be found at /sys/kernel/debug/statistics/ as
  - zfcp-<device-bus-id> for an adapter and
  - zfcp-<device-bus-id>-<WWPN>-<LUN> for a LUN.
- Each subdirectory contains two files, a data and a definition file.
- Activate data gathering via `'echo on=1 >definition'`
- Deactivate via `'echo on=0 >definition'`
- Reset collected data to 0 via `'echo data=reset >definition'`

SCSI statistics (SLES9 and SLES10 only) (cont'd)



```
cat /sys/kernel/debug/statistics/zfcp-0.0.1700-0x5005076303010482-0x4014400500000000/data
...
request_sizes_scsi_read 0x1000 1163
request_sizes_scsi_read 0x80000 805
request_sizes_scsi_read 0x54000 47
...
latencies_scsi_read <=1 1076
latencies_scsi_read <=2 205
latencies_scsi_read <=4 575
...
channel_latency_read <=16000 0
channel_latency_read <=32000 983
channel_latency_read <=64000 99
...
fabric_latency_read <=1000000 1238
fabric_latency_read <=2000000 328
fabric_latency_read <=4000000 522
...
```

request size 4KB, 1163 occurrences

response time <= 1ms

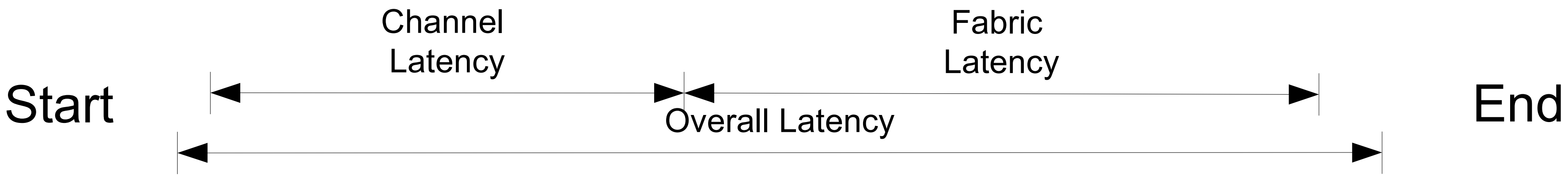
response time <= 32µs

response time <= 1ms

SCSI statistics (SLES9 and SLES10 only) (cont'd)

- The channel latency roughly corresponds to the time a request spent in the channel. (µsec)
- The fabric latency is the time a request spent outside the system z machine. This includes latencies caused by the SAN and the SCSI device (storage server). (µsec)
- The passthrough latency is the delay caused by QDIO (the FCP transport between Linux device driver and FCP channel adapter) and, if applicable, a hypervisor which makes FCP subchannels available to a hosted Linux system. The passthrough latency can be estimated as

passthrough latency = overall latency – (channel latency + fabric latency).

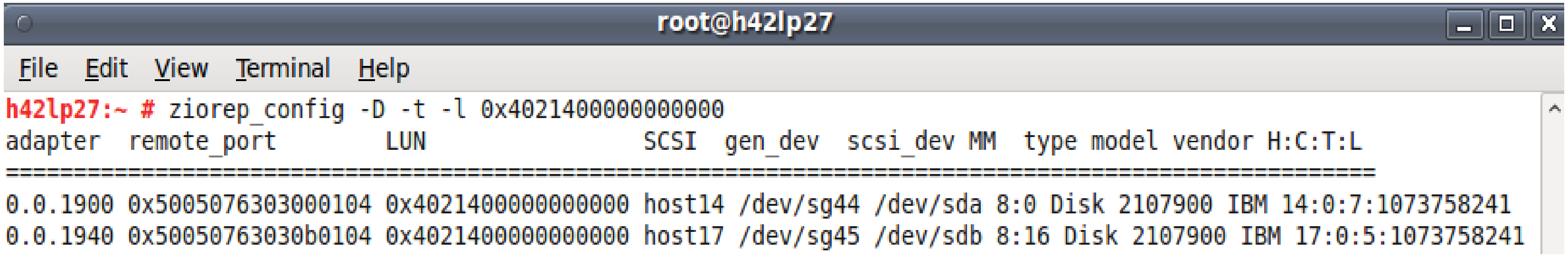


SCSI statistics (SLES11 only)

- Analyse FCP performance with ziomon and ziorep tools
- Capture FCP relevant performance data with the monitor ziomon
  - FCP I/O configuration,
  - I/O workload
  - utilization of FCP resources
- ziomon example (for more see man ziomon)
  - ziomon -i <interval> -d <duration> -l <size limit of output file> -o <output file> <device node> [<device node>]
  - ziomon -i 20 -d 5 -l 50M -o trace_data /dev/sda /dev/sdb
  - ziomon can be stopped with CTRL-C before time period runs out
  - needs Vmalloc space for each device node and CPU

SCSI statistics (SLES11 only) (cont'd)

- ziomon creates 2 output files
  - <output file>.cfg holds various configuration data from the system
  - <output file>.log holds the raw data samples taken during the data collection phase in a binary format
- Use the ziorep tools to analyse the reports created by ziomon
- ziorep_config
  - generates a report on the multipath, SCSI and FCP I/O configuration
  - ziorep_config example (for more see man ziorep_config)
  - ziorep_config -D -t -l 0x4021400000000000



```
h42lp27:~ # ziorep_config -D -t -l 0x4021400000000000
adapter  remote_port        LUN                SCSI  gen_dev  scsi_dev MM  type model vendor H:C:T:L
=====================================================================================================
0.0.1900 0x5005076303000104 0x4021400000000000 host14 /dev/sg44 /dev/sda 8:0 Disk 2107900 IBM 14:0:7:1073758241
0.0.1940 0x50050763030b0104 0x4021400000000000 host17 /dev/sg45 /dev/sdb 8:16 Disk 2107900 IBM 17:0:5:1073758241
```

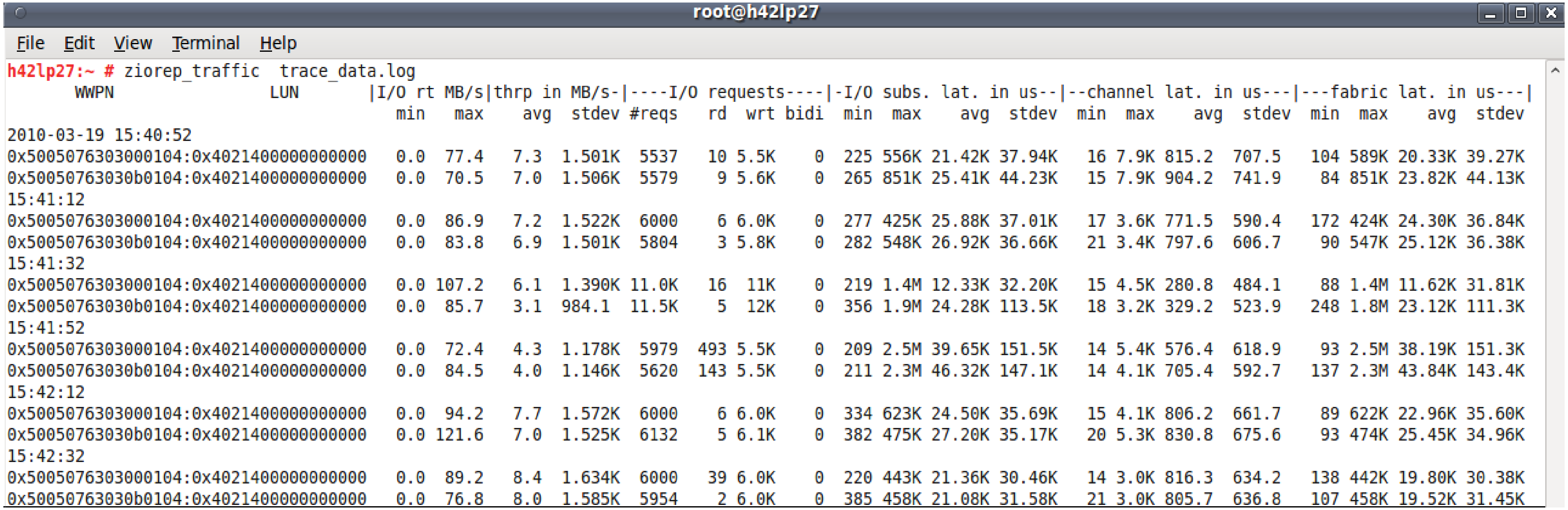
SCSI statistics (SLES11 only) (cont'd)

- ziorep_utilization
  - provides a central detailed analysis of adapters' utilizations, errors, and queue fill levels
  - ziorep_utilization example (for more see man ziorep_utilization)
  - ziorep_utilization <output file>.log

```
CHP|adapter in %-|--bus in %---|--cpu in %---|
 ID min max    avg min max    avg min max    avg
2010-03-19 15:40:52
 58    0    1    0.0    4   12    9.0    0    1    0.0
 5a    0    3    0.0    3   15    9.0    0    1    0.0
.........
CHP Bus-ID  |qdio util.%|queu|fail|-thp in MB/s-|I/O reqs-|
 ID            max    avg full   erc     rd     wrt    rd  wrt
2010-03-19 15:40:52
 58/0.0.1900  98.4   1.9    2    0   4.8    7.3    10 5.5K
 5a/0.0.1940  99.2   2.3    0    0   0.9    7.0     9 5.6K
```

SCSI statistics (SLES11 only) (cont'd)

- ziorep_traffic
  - provides a central detailed analysis of systems I/O traffic through FCP adapters
  - ziorep_traffic example (for more see man ziorep_traffic)
  - ziorep_traffic <output file>.log

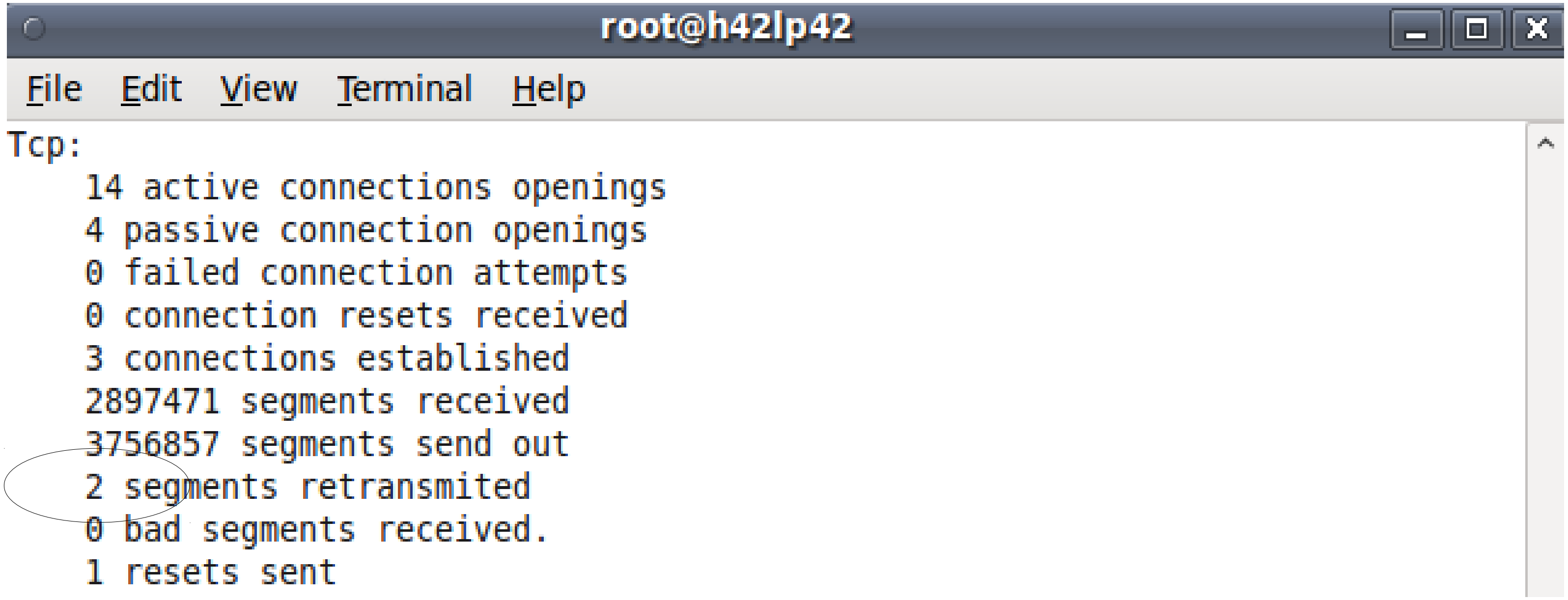


```
h42lp27:~ # ziorep_traffic  trace_data.log
       WWPN                LUN       |I/O rt MB/s|thrp in MB/s-|----I/O requests----|-I/O subs. lat. in us--|--channel lat. in us---|---fabric lat. in us---|
                                        min   max    avg  stdev #reqs   rd  wrt bidi  min  max    avg  stdev  min  max    avg  stdev  min  max    avg  stdev
2010-03-19 15:40:52
0x5005076303000104:0x4021400000000000   0.0  77.4   7.3  1.501K  5537   10 5.5K    0  225 556K 21.42K 37.94K   16 7.9K 815.2  707.5   104 589K 20.33K 39.27K
0x50050763030b0104:0x4021400000000000   0.0  70.5   7.0  1.506K  5579    9 5.6K    0  265 851K 25.41K 44.23K   15 7.9K 904.2  741.9    84 851K 23.82K 44.13K
15:41:12
0x5005076303000104:0x4021400000000000   0.0  86.9   7.2  1.522K  6000    6 6.0K    0  277 425K 25.88K 37.01K   17 3.6K 771.5  590.4   172 424K 24.30K 36.84K
0x50050763030b0104:0x4021400000000000   0.0  83.8   6.9  1.501K  5804    3 5.8K    0  282 548K 26.92K 36.66K   21 3.4K 797.6  606.7    90 547K 25.12K 36.38K
15:41:32
0x5005076303000104:0x4021400000000000   0.0 107.2   6.1  1.390K 11.0K   16  11K    0  219 1.4M 12.33K 32.20K   15 4.5K 280.8  484.1    88 1.4M 11.62K 31.81K
0x50050763030b0104:0x4021400000000000   0.0  85.7   3.1  984.1  11.5K    5  12K    0  356 1.9M 24.28K 113.5K   18 3.2K 329.2  523.9   248 1.8M 23.12K 111.3K
15:41:52
0x5005076303000104:0x4021400000000000   0.0  72.4   4.3  1.178K  5979  493 5.5K    0  209 2.5M 39.65K 151.5K   14 5.4K 576.4  618.9    93 2.5M 38.19K 151.3K
0x50050763030b0104:0x4021400000000000   0.0  84.5   4.0  1.146K  5620  143 5.5K    0  211 2.3M 46.32K 147.1K   14 4.1K 705.4  592.7   137 2.3M 43.84K 143.4K
15:42:12
0x5005076303000104:0x4021400000000000   0.0  94.2   7.7  1.572K  6000    6 6.0K    0  334 623K 24.50K 35.69K   15 4.1K 806.2  661.7    89 622K 22.96K 35.60K
0x50050763030b0104:0x4021400000000000   0.0 121.6   7.0  1.525K  6132    5 6.1K    0  382 475K 27.20K 35.17K   20 5.3K 830.8  675.6    93 474K 25.45K 34.96K
15:42:32
0x5005076303000104:0x4021400000000000   0.0  89.2   8.4  1.634K  6000   39 6.0K    0  220 443K 21.36K 30.46K   14 3.0K 816.3  634.2   138 442K 19.80K 30.38K
0x50050763030b0104:0x4021400000000000   0.0  76.8   8.0  1.585K  5954    2 6.0K    0  385 458K 21.08K 31.58K   21 3.0K 805.7  636.8   107 458K 19.52K 31.45K
```

netstat

- netstat shows
  - Summary information to each protocol
  - Amount of incoming and outgoing packages
  - Various error states, for example TCP segments retransmitted!
- NETSTAT example (for more see man netstat)
  - **netstat** command
  - netstat **-s**
  - „-s“ option displays summary statistics for each protocol

netstat (cont'd)

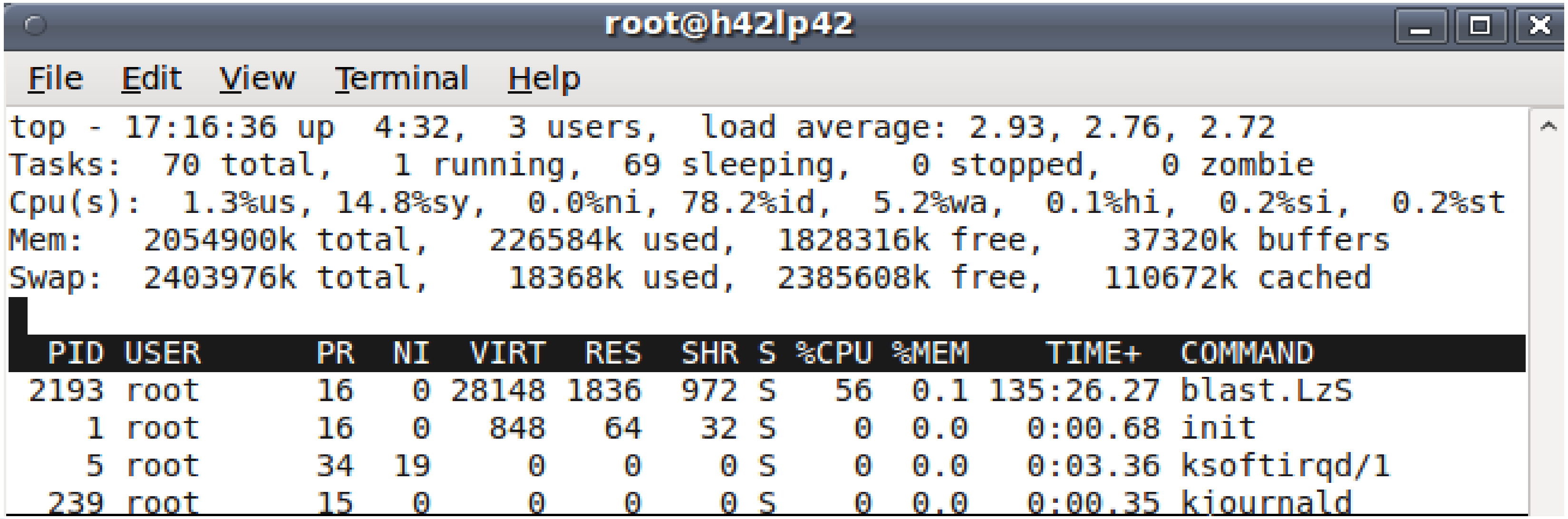


```
Tcp:
    14 active connections openings
    4 passive connection openings
    0 failed connection attempts
    0 connection resets received
    3 connections established
    2897471 segments received
    3756857 segments send out
    2 segments retransmited
    0 bad segments received.
    1 resets sent
```

Watch segments retransmitted
When the system is not able to receive, then the sender shows retransmits

top program

- The top program shows resource usage on process thread level
- `top example (for more see man top)`
  - top [options] -d [delay] -n [iterations] -p [pid, [pid]]
  - top -d 1
  - top -b -d 1 -n 180 >top.log 2>&1 & => batch mode, 3 minutes

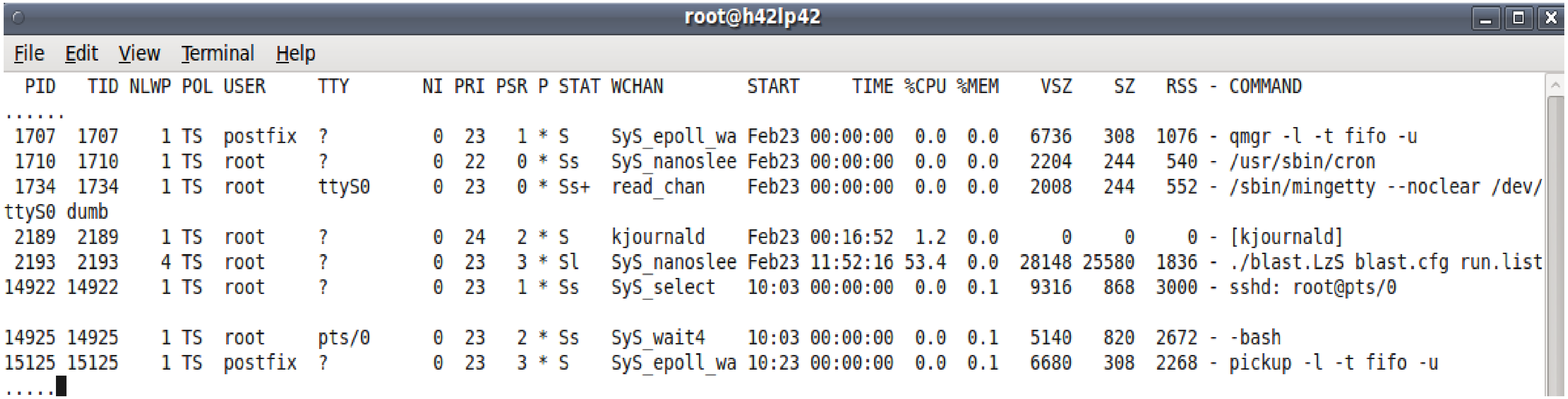


```
top - 17:16:36 up  4:32,  3 users,  load average: 2.93, 2.76, 2.72
Tasks:  70 total,   1 running,  69 sleeping,   0 stopped,   0 zombie
Cpu(s):  1.3%us, 14.8%sy,  0.0%ni, 78.2%id,  5.2%wa,  0.1%hi,  0.2%si,  0.2%st
Mem:   2054900k total,   226584k used,  1828316k free,    37320k buffers
Swap:  2403976k total,    18368k used,  2385608k free,   110672k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
 2193 root      16   0 28148 1836  972 S   56  0.1 135:26.27 blast.LzS
    1 root      16   0   848   64   32 S    0  0.0   0:00.68 init
    5 root      34  19     0    0    0 S    0  0.0   0:03.36 ksoftirqd/1
  239 root      15   0     0    0    0 S    0  0.0   0:00.35 kjournald
```

ps command

- The ps command reports a snapshot of the current processes
- `ps example (for more see man ps)`
  - to see every process with a user-defined format
  - ps -eo pid,tid,nlwp,policy,user,tname,ni,pri,psr,sgi_p,stat,wchan:12,start_time,time, pcpu,pmem,vsize,size, rss,share,command



```
  PID   TID NLWP POL USER     TTY        NI PRI PSR P STAT WCHAN        START     TIME %CPU %MEM    VSZ    SZ   RSS - COMMAND
......
 1707  1707    1 TS  postfix  ?           0  23   1 * S    SyS_epoll_wa Feb23 00:00:00  0.0  0.0   6736   308  1076 - qmgr -l -t fifo -u
 1710  1710    1 TS  root     ?           0  22   0 * Ss   SyS_nanoslee Feb23 00:00:00  0.0  0.0   2204   244   540 - /usr/sbin/cron
 1734  1734    1 TS  root     ttyS0       0  23   0 * Ss+  read_chan    Feb23 00:00:00  0.0  0.0   2008   244   552 - /sbin/mingetty --noclear /dev/
ttyS0 dumb
 2189  2189    1 TS  root     ?           0  24   2 * S    kjournald    Feb23 00:16:52  1.2  0.0      0     0     0 - [kjournald]
 2193  2193    4 TS  root     ?           0  23   3 * Sl   SyS_nanoslee Feb23 11:52:16 53.4  0.0  28148 25580  1836 - ./blast.LzS blast.cfg run.list
14922 14922    1 TS  root     ?           0  23   1 * Ss   SyS_select   10:03 00:00:00  0.0  0.1   9316   868  3000 - sshd: root@pts/0

14925 14925    1 TS  root     pts/0       0  23   2 * Ss   SyS_wait4    10:03 00:00:00  0.0  0.1   5140   820  2672 - -bash
15125 15125    1 TS  postfix  ?           0  23   3 * S    SyS_epoll_wa 10:23 00:00:00  0.0  0.1   6680   308  2268 - pickup -l -t fifo -u
.....
```

Agenda – Part II

- Remarks about customer incidents
- Customer reported incidents
  - Disk I/O bottlenecks
  - FCP disk configuration issues
  - Long response time
  - Guest spontaneously reboots
  - Kernel Panic: Low Address Protection
  - IPL of LPAR takes hours
  - Unable to mount file system after LVM changes
  - High CPU consumption in VM but not in Linux
  - Bonding throughput not matching expectations
  - Service time bigger than average wait time
  - More customer problems: in a nutshell

Introductory Remarks

- The incidents reported here are real customer incidents
  - Red Hat Enterprise Linux, and Novell Linux Enterprise Server distributions
  - Linux running in LPAR and z/VM of different versions
- While problem analysis looks rather straight forward on the charts, it might have taken weeks to get it done.
- The more information is available, the sooner the problem can be solved, because gathering and submitting additional information again and again usually introduces delays.
  - See First Aid Kit at the beginning of this presentation.
- This presentation focuses on how the tools have been used, comprehensive documentation on their capabilities is in the docs of the corresponding tool.

Performance: 'disk I/O bottlenecks'

- Configuration:
  - Customer has distributed I/O workload to multiple volumes using VM minidisk and LVM striping
  - This problem also applies to non-LVM and non minidisk configurations
- Problem Description:
  - Multi-disk I/O performance is worse than expected by projecting single disk benchmark to more complex solution.
- Tools used for problem determination:
  - dbginfo.sh
  - Linux for System z Debug Feature
  - Linux SADC/SAR, IOSTAT and DASD statistics
  - z/VM monitor data
  - Storage Controller DASD statistics

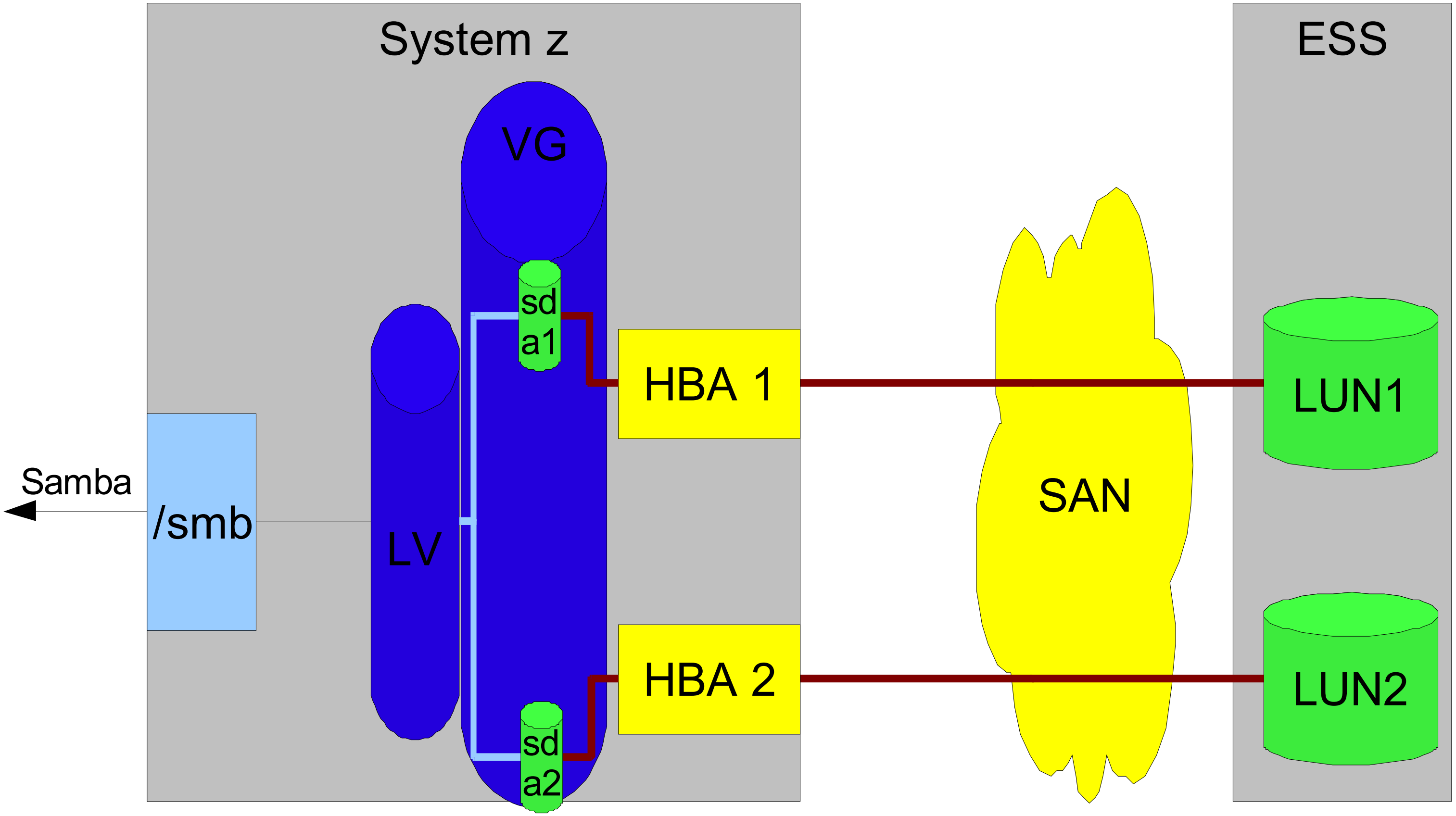
Performance: 'disk I/O bottlenecks' (cont'd)

- Problem Origin:
  - bottleneck other than the device – e.g.:
    - z/VM minidisks are associated to same physical disk
    - SAN bandwidth not sufficient
    - Storage controller HBA bandwidth not sufficient
    - Multiple disks used are in the same rank of storage controller
- Solution:
  - Check your disk configuration and configure for best performance
  - Make sure, minidisks used in parallel are not on the same physical disk
  - Distribution of I/O workload (striped LVs, PAV or HyperPAV)
  - For optimal disk performance configurations read and take into account http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_optimizedisk.html

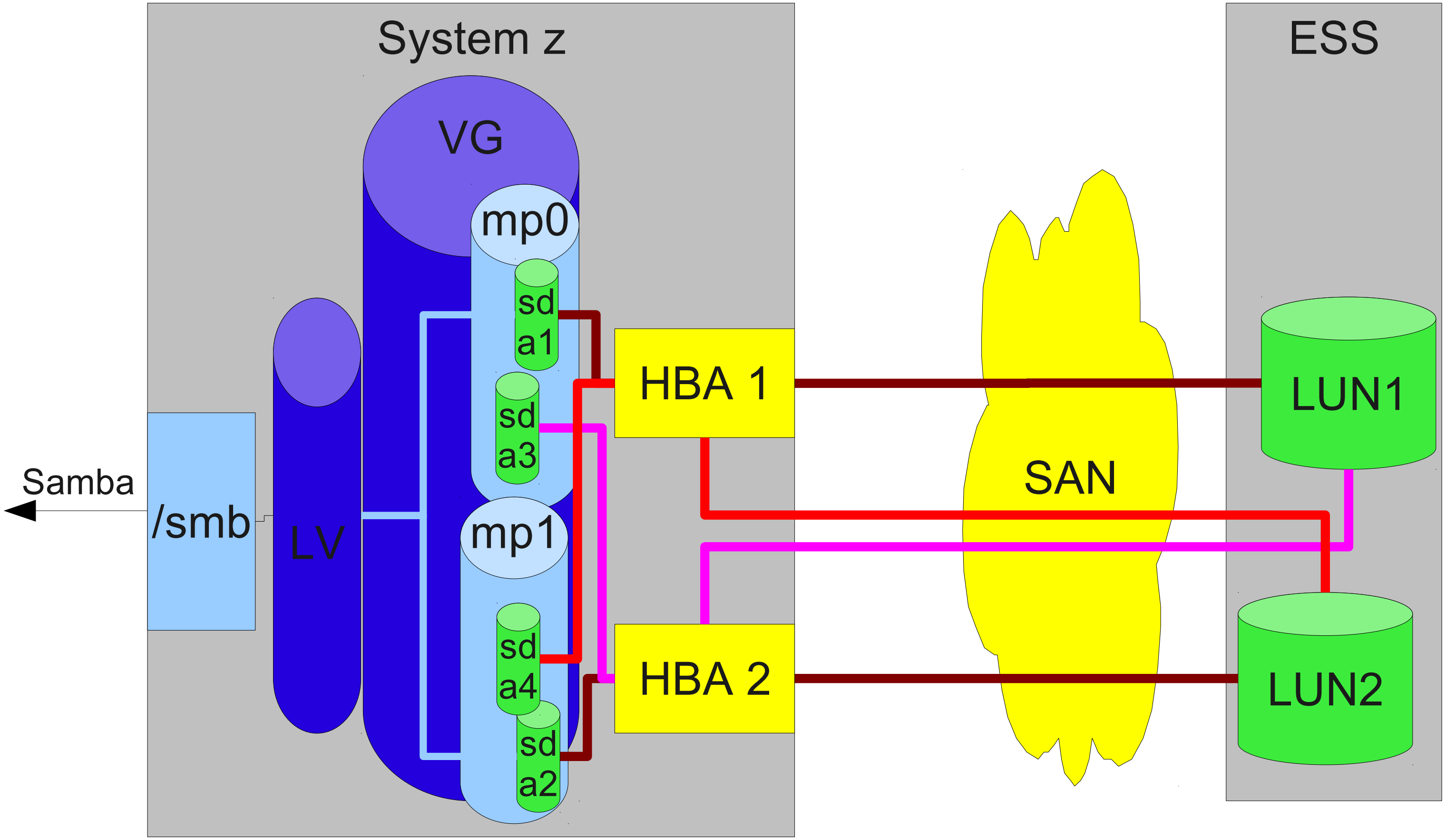
FCP disk: ‘multipath configuration’

- Configuration:
  - Customer is running Samba server on Linux with FCP attached disk managed by Linux LVM.
  - This problem also applies to any configuration with FCP attached disk storage
- Problem Description:
  - Accessing *some files* through samba causes the system to hang while accessing other files works fine
  - Local access to the same file cause a hanging shell as well
    - Indicates: this is not a network problem!
- Tools used for problem determination:
  - dbginfo.sh
- Problem Indicators:
  - Intermittent outages of disk connectivity

FCP disk: ‘multipath configuration’ (cont'd)



FCP disk: ‘multipath configuration’ (cont'd)



Performance: Long response time

- Configuration:
  - Oracle RAC server or other databases on guest under z/VM
- Problem Description:
  - Access to database did not meet customer's expectations
- Tools used for problem determination:
  - dbginfo.sh
  - Linux SADC/SAR
  - z/VM monitor data

Performance: Long response time (cont'd)

- Problem Origin:
  - Insufficient CPU resources for z/VM guest or LPAR – e.g.:
    - Undersized z/VM guest after migration from non z-platform
    - Additional workload without changing physical resources
      - *On the very same guest*
      - *Additional guests or more workload on other guests*
    - Inappropriate CPU shares in z/VM and/or LPAR hypervisor level
- Solution:
  - Reduce CPU overcommitment
    - Offload workload from overloaded z/VM (guest) or LPAR
    - Assign appropriate priorities to guests by setting SHARE
    - Resize the CPU resource need based on the current workload and for further workload extensions
    - Get additional CPU (IFL) resources

Availability: Guest spontaneously reboots

- Configuration:
  - Oracle RAC server or other HA solution under z/VM
- Problem Description:
  - Occasionally guests spontaneously reboot without any notification or console message
- Tools used for problem determination:
  - cp instruction trace of (re)IPL code
  - Crash dump taken after trace was hit

Availability: Guest Spontaneously reboots (cont'd)

- Problem Origin:
  - HA component erroneously detected a system hang
    - hangcheck_timer module did not receive timer IRQ
    - z/VM 'time bomb' switch
    - TSA monitor
- z/VM cannot guarantee 'real-time' behavior if overloaded
  - Longest 'hang' observed: 37 seconds(!)
- Solution:
  - Offload HA workload from overloaded z/VM
    - e.g. use separate z/VM
    - Or: run large Oracle RAC guests in LPAR

Kernel panic: Low address protection

- Configuration:
  - z10 only
  - High work load
  - The more likely the more multithreaded applications are running
- Problem Description:
  - Concurrent access to pages to be removed from the page table
- Tools used for problem determination:
  - crash/lcrash
- Problem Origin:
  - Race condition in memory management
- Solution:
  - Upgrade to latest kernels – fix to be integrated in all supported distributions

Performance: IPL of LPAR takes hours

- Configuration:
  - Customer is running in LPAR with many (>10k) subchannels
- Problem Description:
  - IPL takes hours,
  - network interfaces and file systems are not activated during IPL
- Tools used for problem determination:
  - dbginfo.sh (lscss)
- Problem Origin:
  - Unused subchannels delay IPL
- Solution:
  - Use cio_ignore to restrict system to used subchannels

Unable to mount file system after LVM changes

- Configuration:
  - Linux HA cluster with two nodes
  - Accessing same dasds which are exported via ocfs2
- Problem Description:
  - Added one node to cluster, brought Logical Volume online
  - Unable to mount the filesystem from any node after that
- Tools used for problem determination:
  - dbginfo.sh
- Problem Origin:
  - LVM metadata was overwritten when adding 3rd node
- Solution:
  - Extract meta data from running node and write to disk again

High CPU consumption in VM but not in Linux

- Configuration:
  - SLES10 SP2 system with Tivoli Monitoring
  - No other workload, relatively idle
- Problem Description:
  - Seeing 6% IFL usage in VM
  - Seeing 2% CPU usage in Linux
- Tools used for problem determination:
  - dbginfo.sh, top
- Problem Origin:
  - Bug in Linux Kernel prevented VM from putting it on to the idle run queue
- Solution:
  - Apply service, fixed since 2.6.16.60-0.34

Bonding throughput not matching expectations

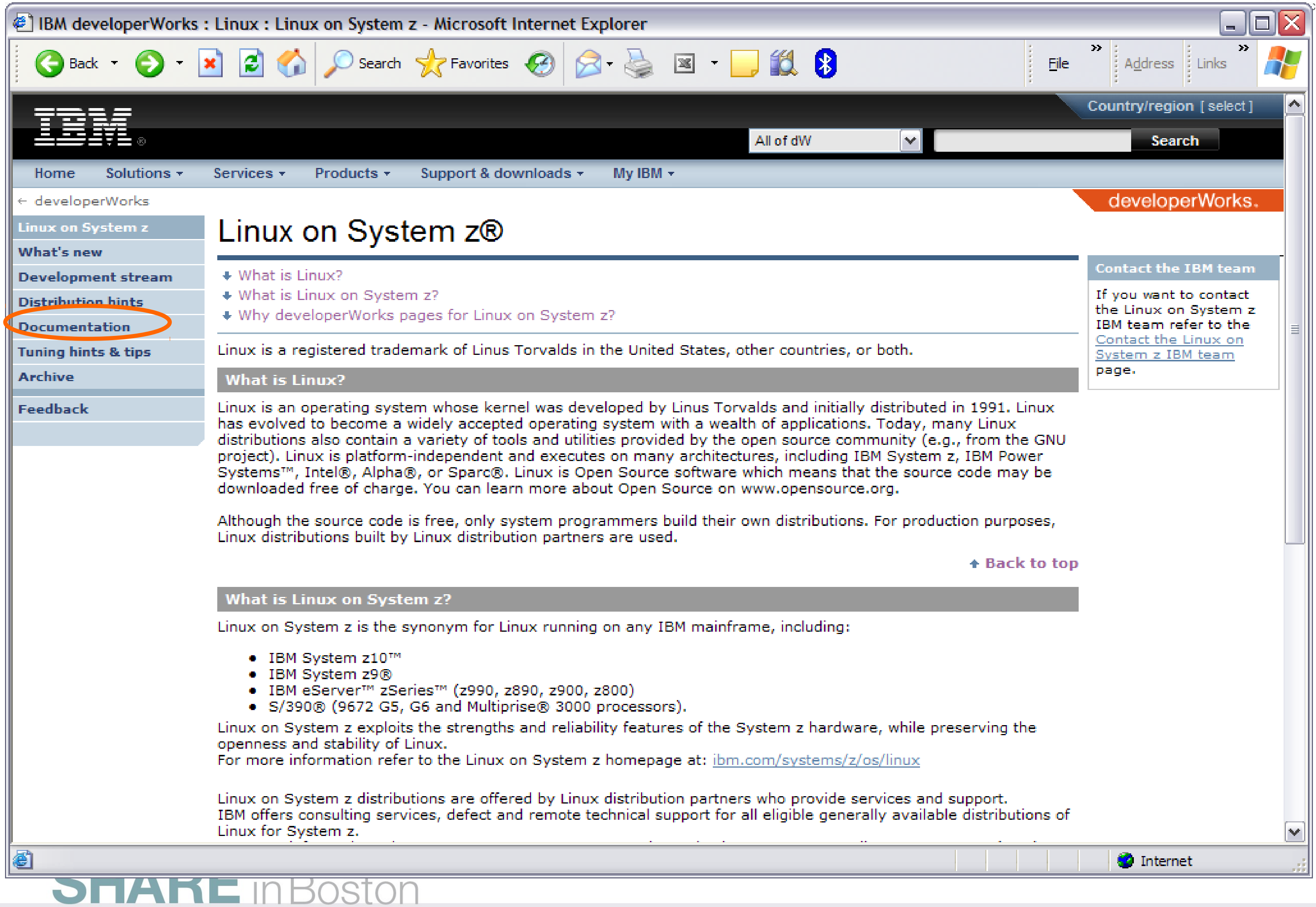
- Configuration:
  - SLES10 system, connected via OSA card and using bonding driver
- Problem Description:
  - Bonding only working with 100mbps
  - FTP also slow
- Tools used for problem determination:
  - dbginfo.sh, netperf
- Problem Origin:
  - ethtool cannot determine line speed correctly because qeth does not report it
- Solution:
  - Ignore the 100mbps message – upgrade to SLES11

Service time bigger than average wait time

- Configuration:
  - SLES9 system, SCSI storage
- Problem Description:
  - Service time (scvtm) sometimes higher than average wait time (await)
- Tools used for problem determination:
  - dbginfo.sh, scsi statistics
- Problem Origin:
  - with very low utilisation the times might be wrong because of interval boundaries
- Solution:
  - Look at the complete picture:
    - Include scsi statistics
    - Do not focus on one line from iostat/syssat

Questions?

developerWorks – entry page for documentation



Linux on System z®

- What is Linux?
- What is Linux on System z?
- Why developerWorks pages for Linux on System z?

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

What is Linux?

Linux is an operating system whose kernel was developed by Linus Torvalds and initially distributed in 1991. Linux has evolved to become a widely accepted operating system with a wealth of applications. Today, many Linux distributions also contain a variety of tools and utilities provided by the open source community (e.g., from the GNU project). Linux is platform-independent and executes on many architectures, including IBM System z, IBM Power Systems™, Intel®, Alpha®, or Sparc®. Linux is Open Source software which means that the source code may be downloaded free of charge. You can learn more about Open Source on www.opensource.org.

Although the source code is free, only system programmers build their own distributions. For production purposes, Linux distributions built by Linux distribution partners are used.

Back to top

What is Linux on System z?

Linux on System z is the synonym for Linux running on any IBM mainframe, including:

- IBM System z10™
- IBM System z9®
- IBM eServer™ zSeries™ (z990, z890, z900, z800)
- S/390® (9672 G5, G6 and Multiprise® 3000 processors).

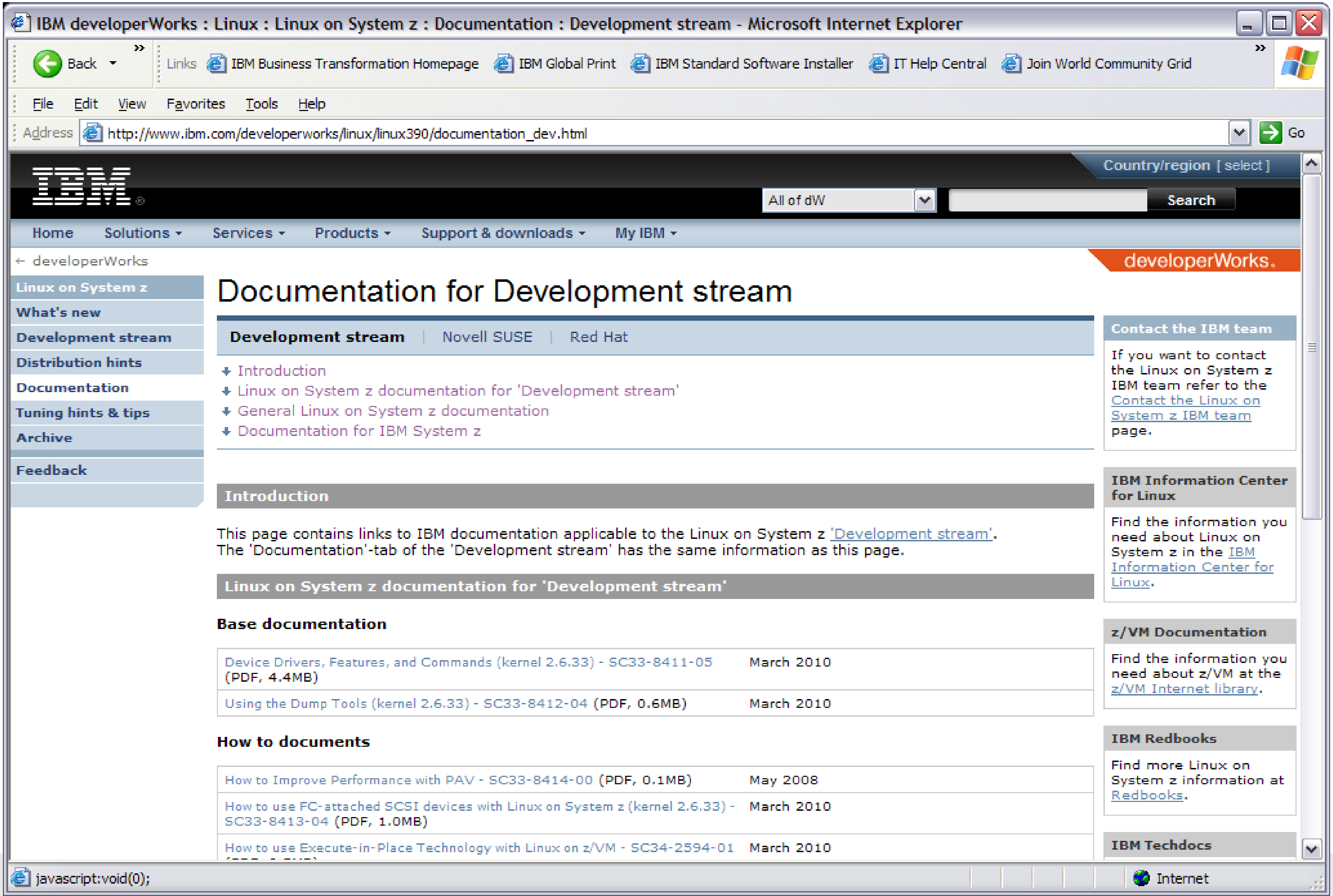
Linux on System z exploits the strengths and reliability features of the System z hardware, while preserving the openness and stability of Linux.
For more information refer to the Linux on System z homepage at: ibm.com/systems/z/os/linux

Linux on System z distributions are offered by Linux distribution partners who provide services and support.
IBM offers consulting services, defect and remote technical support for all eligible generally available distributions of Linux for System z.

Contact the IBM team

If you want to contact the Linux on System z IBM team refer to the Contact the Linux on System z IBM team page.

Development stream – Novell SUSE – Red Hat documentation



IBM developerWorks : Linux : Linux on System z : Documentation : Development stream - Microsoft Internet Explorer

http://www.ibm.com/developerworks/linux/linux390/documentation_dev.html

← developerWorks

- Linux on System z
- What's new
- Development stream
- Distribution hints
- Documentation
- Tuning hints & tips
- Archive
- Feedback

Documentation for Development stream

Development stream | Novell SUSE | Red Hat

- Introduction
- Linux on System z documentation for 'Development stream'
- General Linux on System z documentation
- Documentation for IBM System z

Introduction

This page contains links to IBM documentation applicable to the Linux on System z 'Development stream'. The 'Documentation'-tab of the 'Development stream' has the same information as this page.

Linux on System z documentation for 'Development stream'

Base documentation

Device Drivers, Features, and Commands (kernel 2.6.33) - SC33-8411-05 (PDF, 4.4MB)	March 2010
Using the Dump Tools (kernel 2.6.33) - SC33-8412-04 (PDF, 0.6MB)	March 2010

How to documents

How to Improve Performance with PAV - SC33-8414-00 (PDF, 0.1MB)	May 2008
How to use FC-attached SCSI devices with Linux on System z (kernel 2.6.33) - SC33-8413-04 (PDF, 1.0MB)	March 2010
How to use Execute-in-Place Technology with Linux on z/VM - SC34-2594-01	March 2010

Contact the IBM team

If you want to contact the Linux on System z IBM team refer to the Contact the Linux on System z IBM team page.

IBM Information Center for Linux

Find the information you need about Linux on System z in the IBM Information Center for Linux.

z/VM Documentation

Find the information you need about z/VM at the z/VM Internet library.

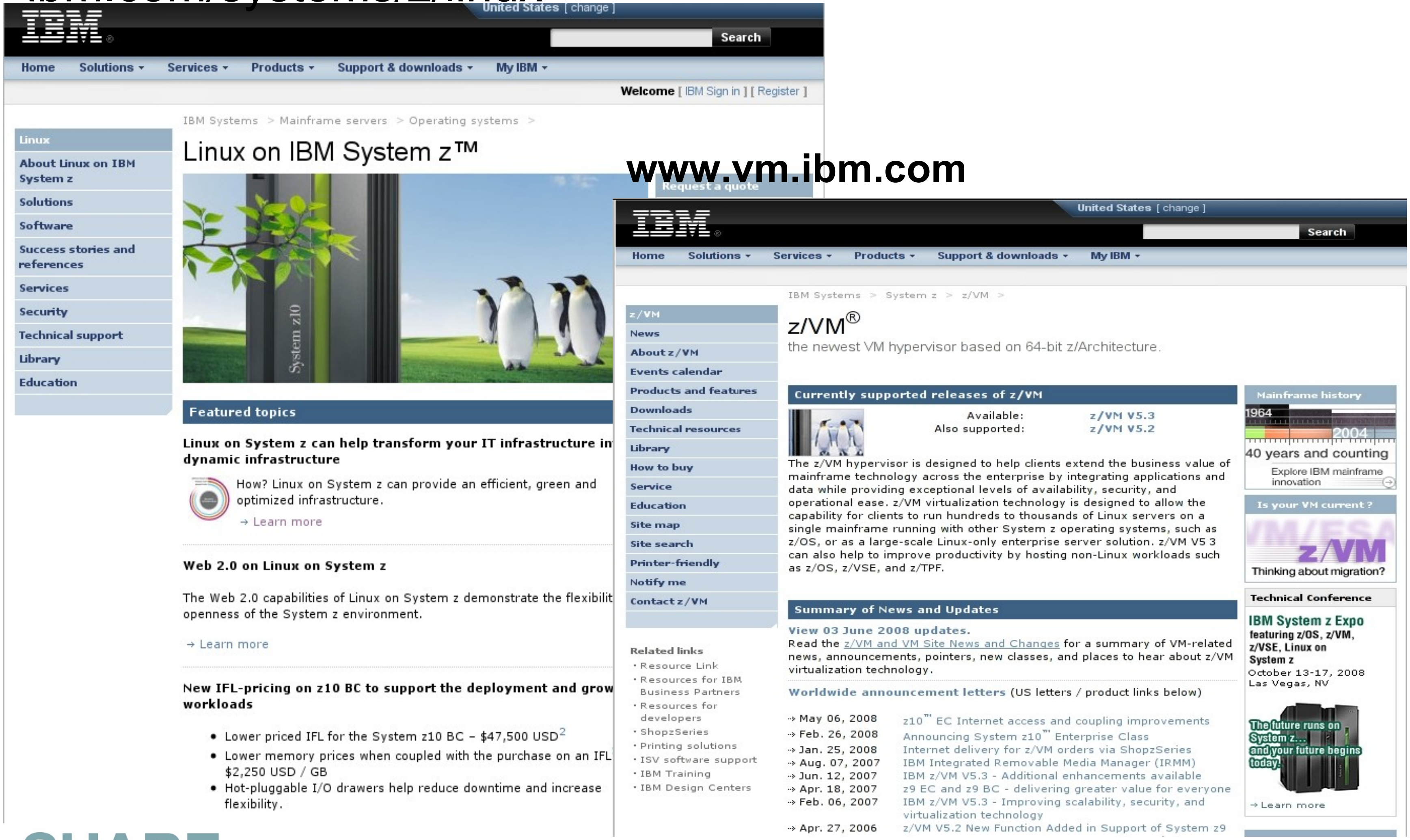
IBM Redbooks

Find more Linux on System z information at Redbooks.

IBM Techdocs

More information

ibm.com/systems/z/linux



United States [change]

Search

Home | Solutions | Services | Products | Support & downloads | My IBM

Welcome [IBM Sign in] [Register]

IBM Systems > Mainframe servers > Operating systems >

Linux
- About Linux on IBM System z
- Solutions
- Software
- Success stories and references
- Services
- Security
- Technical support
- Library
- Education

Linux on IBM System z™

Request a quote

Featured topics

Linux on System z can help transform your IT infrastructure in dynamic infrastructure

How? Linux on System z can provide an efficient, green and optimized infrastructure.

→ Learn more

Web 2.0 on Linux on System z

The Web 2.0 capabilities of Linux on System z demonstrate the flexibilit openness of the System z environment.

→ Learn more

New IFL-pricing on z10 BC to support the deployment and grow workloads

- Lower priced IFL for the System z10 BC – $47,500 USD[2]
- Lower memory prices when coupled with the purchase on an IFL $2,250 USD / GB
- Hot-pluggable I/O drawers help reduce downtime and increase flexibility.

www.vm.ibm.com

United States [change]

Search

Home | Solutions | Services | Products | Support & downloads | My IBM

IBM Systems > System z > z/VM >

z/VM
- News
- About z/VM
- Events calendar
- Products and features
- Downloads
- Technical resources
- Library
- How to buy
- Service
- Education
- Site map
- Site search
- Printer-friendly
- Notify me
- Contact z/VM

Related links
- Resource Link
- Resources for IBM Business Partners
- Resources for developers
- ShopzSeries
- Printing solutions
- ISV software support
- IBM Training
- IBM Design Centers

z/VM®

the newest VM hypervisor based on 64-bit z/Architecture.

Currently supported releases of z/VM

Available:	z/VM V5.3
Also supported:	z/VM V5.2

The z/VM hypervisor is designed to help clients extend the business value of mainframe technology across the enterprise by integrating applications and data while providing exceptional levels of availability, security, and operational ease. z/VM virtualization technology is designed to allow the capability for clients to run hundreds to thousands of Linux servers on a single mainframe running with other System z operating systems, such as z/OS, or as a large-scale Linux-only enterprise server solution. z/VM V5 3 can also help to improve productivity by hosting non-Linux workloads such as z/OS, z/VSE, and z/TPF.

Summary of News and Updates

View 03 June 2008 updates.
Read the z/VM and VM Site News and Changes for a summary of VM-related news, announcements, pointers, new classes, and places to hear about z/VM virtualization technology.

Worldwide announcement letters (US letters / product links below)

→ May 06, 2008	z10™ EC Internet access and coupling improvements
→ Feb. 26, 2008	Announcing System z10™ Enterprise Class
→ Jan. 25, 2008	Internet delivery for z/VM orders via ShopzSeries
→ Aug. 07, 2007	IBM Integrated Removable Media Manager (IRMM)
→ Jun. 12, 2007	IBM z/VM V5.3 - Additional enhancements available
→ Apr. 18, 2007	z9 EC and z9 BC - delivering greater value for everyone
→ Feb. 06, 2007	IBM z/VM V5.3 - Improving scalability, security, and virtualization technology
→ Apr. 27, 2006	z/VM V5.2 New Function Added in Support of System z9

Mainframe history

40 years and counting

Explore IBM mainframe innovation

Is your VM current ?

Thinking about migration?

Technical Conference

IBM System z Expo featuring z/OS, z/VM, z/VSE, Linux on System z
October 13-17, 2008
Las Vegas, NV

The future runs on System z... and your future begins today.

→ Learn more

Problem Determination with Linux on System z

Dr. Holger Smolinski
IBM Germany Research & Development GmbH

2010-08-04
9279

Appendix

Corrupted Data: When paging starts, programs dump core!

- Configuration:
  - Customer has configured CDL formatted DASDs as swapspace
- Problem Description:
  - When swapping starts, programs arbitrarily die or dump core
- Tools used for problem determination:
  - dbginfo.sh
- Problem Origin:
  - Customer has configured full disk /dev/dasda as swapspace instead of partition. First blocks of CDL are padded with 0x5e when read, since block length <4k.
- Solution:
  - Configure partition /dev/dasda1 as swapspace
  - Or use LDL formatted devices

NFS: NFS write to z/OS server is slow

- Configuration:
  - Customer is configuring Linux guests with NFS mount to VSAM/PSD datasets on z/OS NFS server
- Problem Description:
  - NFS write of large file takes hours
- Problem Indicator:
  - NFS server writes VSAM datasets
  - Sync mount is faster
- Workaround:
  - Switch to HFS/zFS
  - Use Sync-NFS mount
- Solution:
  - Some relief given by patched Red Hat 5.2 kernel

Performance: ‘disk cache bits settings’

- Configuration:
  - This customer was running database workloads on FICON attached storage
  - The problem applies to any Linux distribution and any runtime environment (z/VM and LPAR)
  - The problem also applies to other workloads with inhomogeneous I/O workload profile (sequential and random access)
- Problem Description:
  - Transaction database performance is within expectation
  - Warm-up basically consisting of database index scans, takes longer than expected.

Performance: ‘disk cache bits settings’ (cont'd)

- Tools used for problem determination:
  - Linux **SADC/SAR** and **IOSTAT**
  - Linux **DASD statistics**
  - **Storage Controller DASD statistics**
  - Scripted testcase
- Problem Indicators:
  - Random Access I/O rates and throughtput are as expected
  - Sequential IO throughput shows variable behaviour
    - always lower than expected
    - As expected for small files, lower than expected for large files
  - Test case showed even stronger performance degradation, when storage controller cache size was exceeded

Performance: ‘disk cache bits settings’ (cont'd)

- Problem Origin:
  - Storage controller cache is utilized inefficiently
    - Sequential data not prestaged
    - Used data not discarded from cache
- Solution:
  - Configure volumes for sequential I/O different from ones for random I/O
  - And use the tunedasd tool to set appropriate cache-setting bits in CCWs for each device. See http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_cachemode.html

Function: no login prompt on integrated ASCII console in HMC

- Configuration:
  - Customer is running in LPAR using integrated ASCII console
- Problem Description:
  - Integrated ASCII console is not enabled as a login terminal
- Problem Origin:
  - Integrated ASCII console must be registered properly
- Solution:
  - Add `'console=ttyS1 conmode=sclp'` to parmline
  - Add console to `/etc/securetty`
  - `Change getty statement in /etc/inittab to:`

```
1:2345:respawn:/sbin/mingetty --noclear /dev/console dumb
```

Networking: 'tcpdump fails'

- Configuration:
  - Customer is trying to sniff the network using tcpdump
- Problem Description (Various problems):
  - tcpdump does not interpret contents of packets or frames
  - tcpdump does not see network traffic for other guests on GuestLAN/HiperSockets network
- Problem Indicators:
  - OSA card is running in Layer 3 mode
  - HiperSocket/Guest LAN do not support promiscuous mode
- Solution:
  - Use the layer-2 mode of your OSA card to add Link Level header
  - Use the tcpdump-wrap.pl script to add fake LL-headers to frames
  - Use the fake-ll feature of the qeth device driver
  - Wait for Linux distribution containing support for promiscuous mode

Networking: 'dhcp fails'

- Configuration:
  - Customer is configuring Linux guests with dhcp and using VLAN
- Problem Description (Various problems):
  - Dhcp configuration does not work on VLAN because
    - Dhcp user space tools do not support VLAN packets
- Problem Indicators:
  - When VLAN is off, dhcp configuration works fine.
- Workaround:
  - Apply service to Linux to hide VLAN information from dhcp tools
    - Ask Distributor/IBM for appropriate kernel levels
- Solution:
  - Request VLAN aware dhcp tools from your distributor

Performance:
'aio (POSIX asynchronous I/O) not used'

- Configuration:
  - Customer is running DB2 on Linux
- Problem Description:
  - Bad write performance is observed, while read performance is okay
- Tools used for problem determination:
  - DB/2 internal tracing
- Problem Origin:
  - libaio is not installed on the system
- Solution:
  - Install libaio package on the system to allow DB2 using it.

Memory:
‘higher order allocation failure’

- Configuration:
  - Customer is running CICS transaction gateway in 31 bit emulation mode
- Problem Description:
  - After several days of uptime, the system runs out of memory
- Tools used for problem determination:
  - Dbginfo.sh
- Problem Indicators:
  - Syslog contains messages about failing 4th-order allocations
    - Caused by compat_ipc calls in 31bit emulation, which request 4th-order memory chunks
- Problem Origin:
  - Compat_ipc code makes order-4 memory allocations
- Solution:
  - Switch to 31 bit system to avoid compat_ipc
  - Upgrade to SLES10
  - Request a fix from distributor or IBM

System stalls: 'PFAULT loop'

- Configuration:
  - Customer is running 35 Linux guests (SLES 8) in z/VM with significant memory overcommit ratio.
- Problem Description:
  - After a couple of days of uptime, the systems hang.
- Tools used for problem determination:
  - System dump
- Problem Origin:
  - CPU loop in the pfault handler caused by
    - Linux acquiring a lock in pfault handler although not needed
- Solution:
  - Request a fix for Linux from SUSE and/or IBM

System stalls: 'reboot hangs'

- Configuration:
  - Customer is running Linux and issuing 'reboot'-command to re-IPL
- Problem Description:
  - 'reboot' shuts down the system but hangs.
- Tools used for problem determination:
  - System dump
- Problem Indicators:
  - 'reboot' hangs, but LOAD-IPL works file
- Problem Origin:
  - Root cause: CHPIDs are not reset properly during 'reboot'
- Solution:
  - Apply Service to Linux, ask SUSE/IBM for appropriate kernel level.

Cryptography:
'HW not used for AES-256'

- Configuration:
  - Customer wants to use Crypto card acceleraton for AES-encryption
- Problem Description:
  - HW acceleration is not used – system falls back to SW implementation
- Tools used for problem determination:
  - SADC/SAR
- Problem Indicators:
  - CPU load higher than expected for AES-256 encryption
- Problem Origin:
  - System z Hardware does not support AES-256 for acceleration.
- Solution:
  - Switch to AES 128 to deploy HW acceleration
  - Expect IBM provided Whitepapers on how to use cryptography appropriately

Cryptography: 'glibc error in openssl'

- Configuration:
  - Customer is performing openssl speed test to check whether crypto HW functions are used in SLES10
- Problem Description:
  - Openssl speed test fails with an error in glibc: “glibc detected openssl: free(): invalid next size (normal)”
- Solution:
  - Upgrade Linux to SLES10 SP1 or above

Storage:
'zipl fails in EAL4 environment'

- Configuration:
  - Customer installs an EAL4 compliant environment with ReiserFS
- Problem Description:
  - Zipl refuses to write boot records due to an ioctl blocked by the auditing SW
- Problem Indicators:
  - Zipl on ext3-FS works well
- Solution:
  - Use ext3-FS at least for /boot

Storage:
'non-persistent tape device nodes'

- Configuration:
  - Customer uses many FCP attached tapes
- Problem Description:
  - Device nodes for tape drives are named differently after reboot
- Solution:
  - Create UDEV-rule to establish persistent naming
  - Wait for IBMtape device driver to support persistent naming

Storage: 'tape device unaccessible'

- Configuration:
  - Customer has FCP attached tape
- Problem Description:
  - Device becomes unaccessible
- Problem Indicators:
  - ELS messages in syslog, or
  - Device can be enabled manually, but using hwup-script it fails
- Solution:
  - Apply service to get fixed version of hwup scripts
  - Apply service to Linux and µCode and disable QIOASSIST if appropriate
    - See: http://www.vm.ibm.com/perf/aip.html for required levels.
  - If tape devices remain reserved by SCSI 3rd party reserve use the ibmtape_util tool from the IBMTape device driver package to break the reservation

Storage: 'QIOASSIST'

- Configuration:
  - Customer is running SLES10 or RHEL 5 under z/VM with QIOASSIST enabled
- Problem Description:
  - System hangs
- Problem Indicators:
  - System stops operation because all tasks are in I/O wait state
  - System runs out of memory, because I/O stalls
  - When switching QIOASIST OFF, the problems vanish
- Solution:
  - **Apply service to Linux, z/VM and System z µCode**
    - See: http://www.vm.ibm.com/perf/aip.html for required levels.

Memory: '31bit address space exhausted'

- Configuration:
  - Customer is migrating database contents to different host in a 31bit system.
- Problem Description:
  - Database reports system caused out-of-memory condition: 'SQL1225N The request failed because an operating system process, thread, or swap space limit was reached.' indicating that a sycall returned -1 and set errno to ENOMEM
- Tools used for problem determination:
  - DB/2 internal tracing
- Problem Origin:
  - System out of resources due to 31bit kernel address space
- Solution:
  - Try to reduce memory footprint of workload (nr of threads, buffer sizes...)
  - Run migration in 31bit compatibility environment of 64 bit system

Storage: 'DASD unaccessible'

- Configuration:
  - Customer is running SLES9 with LVM configuration
- Problem Description:
  - DASDs become not accessible after boot
- Problem Indicators:
  - Intermitting errors due to race between LVM and device recognition
- Solution:
  - Apply service to Linux
  - Race fixed, due to which partition detection couldn’t complete, because LVM had devices already in use.

Networking: 'firewall cuts TCP connections'

- Configuration:
  - Customer is running eRMM in a firewalled environment
- Problem Description:
  - After certain period of inactivity eRMM server loses connectivity to clients
- Problem Indicators:
  - Disconnect occurs after fixed period of inactivity
  - Period counter appears to be reset when activity occurs
- Solution:
  - Tune TCP_KEEPALIVE timeout to be shorter than firewall setting, which cuts inactive connections

Networking: 'Channel Bonding'

- Configuration:
  - Customer is trying to configure channel bonding on SLES 10 system
- Problem Description (Various problems):
  - Interfaces refuse to get enslaved
  - Failover/failback does not work
  - Kernel Panic when issuing 'ifenslave -d' command
- Solution:
  - Apply Service to Linux, System z HW and z/VM
    - ask SUSE/IBM for appropriate kernel and µCode levels.