Problem Determination for Linux on System z

Mario Held
IBM Research & Development, Germany

August 26, 2009
Session Number 9279

Trademarks

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

DB2*
DB2 Connect
DB2 Universal Database
e-business logo
GDPS*
Geographically Dispersed Parallel Sysplex
HyperSwap
IBM*
IBM eServer
IBM logo*
Parallel Sysplex*
System z
Tivoli*
VM/ESA*
WebSphere*
z/OS*
z/VM*
zSeries*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Intel is a registered trademark of the Intel Corporation in the United States, other countries or both.

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

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries.

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

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* 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

- **In case a problem shows up**
  - **Describe problem**
  - **Describe the environment**
- Tools used to determine problems
  - Troubleshooting “first-aid kit”
  - sysstat package
  - Disk statistics
- Customer reported incidents
  - Networking: ‘TSM - breaking TCP connections'
  - SCSI disk: 'Multipath configuration'
  - More customer problems in a nutshell

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.
  - Some useful links are included

Describe the problem

Get as much information as possible about the circumstances:

- What is the problem ?
- When did it appear ? - date and time, important to dig into logs
- Where did it appear ? - 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 by will ?
- **Write down as much as possible information 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 if relevant

Agenda

- In case a problem shows up
  - Describe problem
  - Describe the environment
- **Tools used to determine problems**
  - **Troubleshooting “first-aid kit”**
  - **sysstat package**
  - **Disk statistics**
- Customer reported incidents
  - Networking: ‘TSM - breaking TCP connections'
  - SCSI disk: 'Multipath configuration'
  - More customer problems in a nutshell

Trouble-shooting “first-aid kit”

- Install packages required for debugging
  - s390-tools / s390utils
  - sysstat
  - Dump tools crash / lkcdutils
    - Lkcdutils / lcrash available with SLES9 and SLES10
    - crash and lcrash available on SLES11
    - crash in all RHEL distributions
- Collect system data
  - Always archive syslog (/var/log/messages)
  - Start sadc (System Activity Data Collection) service when appropriate
  - Collect z/VM Monitor Data if running under z/VM when appropriate
  - Enable disk statistics if needed
- Collect dbginfo.sh output
  - Pro-actively in healthy system
  - When problems occur – then compare with healthy system

Trouble-shooting “first-aid kit” (cont'd)

- dbginfo.sh where to get ?
  - 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
- dbginfo.sh captures the following information:
  - /proc/[version, cpu, meminfo, slabinfo, modules, partitions, devices ...]
  - System z specific device driver information: /proc/s390dbf
  - Kernel messages /var/log/messages
  - Reads configuration files in directory /etc/

    [ccwgroup.conf, chandev.conf, modules.conf, fstab]
  - Uses several commands: ps, dmesg
  - DASD setup
    - /proc/dasd/devices or lsdasd (part of s390-tools package)
    - LVM lvdisplay, vgdisplay
  - And much more

Trouble-shooting “first-aid kit” (cont'd)

- Network:
  - Draw a picture of you network setup if possible
  - Run lsqeth (part of s390-tools package)

```
h3730002:~ # lsqeth
Device name                     : eth2
-------------------------------------------------
        card_type               : OSD_10GIG
        cdev0                   : 0.0.4104
        cdev1                   : 0.0.4105
        cdev2                   : 0.0.4103
        chpid                   : 82
        online                  : 1
        portname                : OSAPORT
        portno                  : 0
        route4                  : no
        route6                  : no
        checksumming            : hw checksumming
        state                   : SOFTSETUP
        priority_queueing       : always queue 2
        fake_ll                 : 0
        fake_broadcast          : 0
        buffer_count            : 128
        add_hhlen               : 0
        layer2                  : 0
        large_send              : no
```

Trouble-shooting “first-aid kit” (cont'd)

- z/VM:
  - Release and service Level: `q cplevel`
  - Network setup: `q [lan, nic, vswitch, v osa]`
  - General/DASD: `q [set, v dasd ...]`
  - Issue above commands in 3270 console or use `vmcp` or `hcp` in Linux

```
h3730002:~ # modprobe vmcp
h3730002:~ # vmcp 'q cplevel'
z/VM Version 5 Release 4.0, service level 0801 (64-bit)
Generated at 01/07/09 09:48:41 CST
IPL at 08/24/09 08:25:42 CST
h3730002:~ #

h3730002:~ # vmcp 'q v stor'
STORAGE = 2047M
```

Trouble-shooting “first-aid kit” (cont'd)

- When System hangs
  - Take a dump
    - Include the System.map and (if available) Kerntypes file from /boot
    - Refer to the “Using the Dump Tools” book on:

      http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26ddt02.pdf

- If a function does not work as expected
  - Enable extended tracing in /proc/s390dbf or /sys/s390dbf for a subsystem

- In case of a performance problem
  - Enable sadc (System Activity Data Collection) service
  - Collect z/VM Monitor Data if running under z/VM
  - Enable disk statistics if appropriate

Trouble-shooting “first-aid kit” (cont'd)

- Attach comprehensive documentation to problem report:
  - Output file of dbginfo.sh script (/tmp/DBGINFO-<date>.tgz)
  - z/VM monitor 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://testcase.boulder.ibm.com/** or
  - **ftp://ecurep.ibm.com/**
  - See Instructions: **http://www.ibm.com/de/support/ecurep/other.html**
- When opening a Bugzilla (bug tracker web application) at a distribution partner attach documentation directly

Use and configure sysstat:

- Capture Linux performance data with sysstat package
  - Part of the distribution (but might be not preinstalled)
  - **S**ystem **A**ctivity **D**ata **C**ollector (sadc)
  - **S**ystem **A**ctivity **R**eport (sar) command
  - iostat command
- sadc example
  - /usr/lib/sa/sadc <interval> <count> <binary outfile>
  - **`/usr/lib/sa/sadc 10 30 sadc_outfile`**
  - Should be started as a service during system start
- sar example
  - sar -A -f <binary outfile> analyze data from current sadc data collection
- iostat example
  - iostat -dkx --> analyze io related performance data for all disks
- Please include the binary sadc data and sar -A output when submitting sadc information to IBM support

CPU utilization

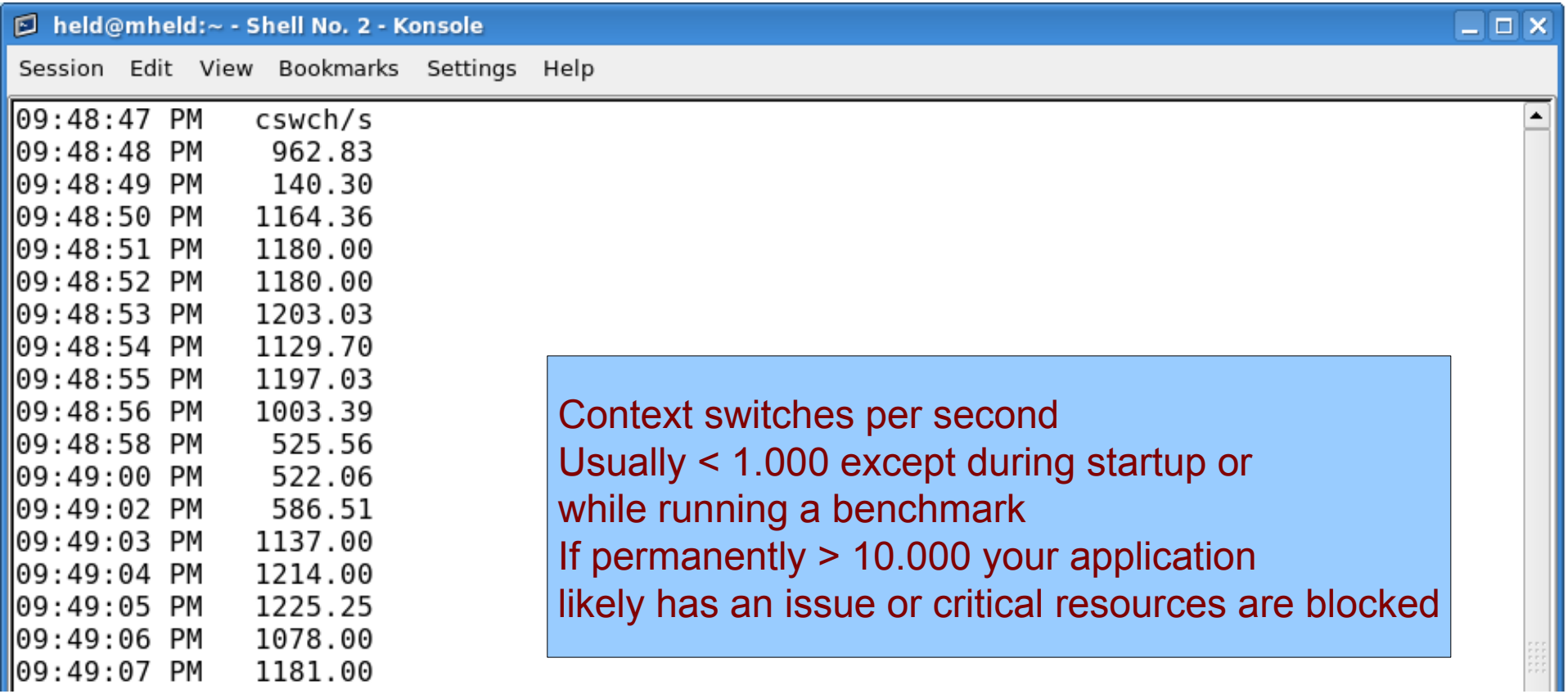
held@mheld:~ - Shell No. 2 - Konsole

Session Edit View Bookmarks Settings Help

```
09:48:47 PM     CPU     %user     %nice   %system   %iowait    %steal     %idle
09:48:55 PM     all     22.75      0.00     30.74      0.00      0.20     46.31
09:48:55 PM       0     42.57      0.00     57.43      0.00      0.00      0.00
09:48:55 PM       1     43.00      0.00     57.00      0.00      0.00      0.00
09:48:55 PM       2     42.42      0.00     57.58      0.00      0.00      0.00
09:48:55 PM       3      0.00      0.00      0.00      0.00      0.00    100.00
09:48:55 PM       4     43.43      0.00     56.57      0.00      0.00      0.00
09:48:55 PM       5      0.00      0.00      0.00      0.00      0.00    100.00
09:48:55 PM       6      0.00      0.00      0.00      0.00      0.00      0.00
09:48:55 PM       7      0.00      0.00      0.00      0.00      0.00      0.00
09:48:55 PM       8      0.00      0.00      0.00      0.00      0.00      0.00
09:48:55 PM       9      0.00      0.00      0.00      0.00      0.00      0.00
09:48:55 PM      10     42.42      0.00     57.58      0.00      0.00      0.00
09:48:55 PM      11     43.00      0.00     57.00      0.00      0.00      0.00
09:48:55 PM      12     42.57      0.00     56.44      0.00      0.00      0.99
09:48:55 PM      13      0.00      0.00      0.00      0.00      0.00    100.00
09:48:55 PM      14                                              0.00     99.57
09:48:55 PM      15                                              2.97      0.00
09:48:56 PM     all                                              0.13     73.35
09:48:56 PM       0                                              0.00     90.68
09:48:56 PM       1                                              0.00     94.74
09:48:56 PM       2                                              1.98     93.07
09:48:56 PM       3                                              2.00      1.00
09:48:56 PM       4                                              0.00     16.00
```

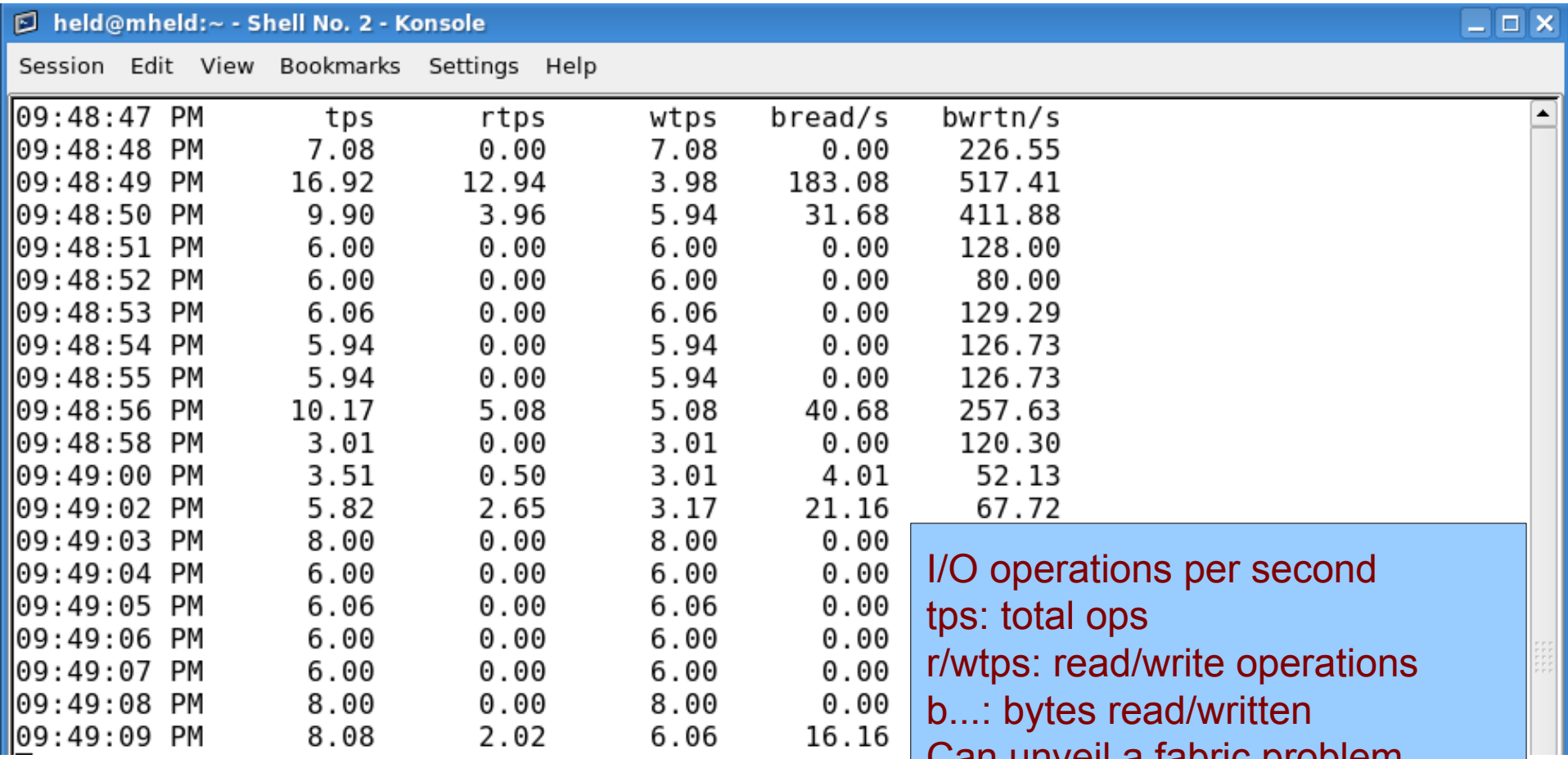
Per CPU values:
watch out for
- system time (kernel time)
- iowait time (slow I/O subsystem)
- steal time (time taken by other guests)

Context Switch Rate



Context switches per second
Usually < 1.000 except during startup or while running a benchmark
If permanently > 10.000 your application likely has an issue or critical resources are blocked

I/O rates


held@mheld:~ - Shell No. 2 - Konsole
Session Edit View Bookmarks Settings Help
09:48:47 PM tps rtps wtps bread/s bwrtn/s
09:48:48 PM 7.08 0.00 7.08 0.00 226.55
09:48:49 PM 16.92 12.94 3.98 183.08 517.41
09:48:50 PM 9.90 3.96 5.94 31.68 411.88
09:48:51 PM 6.00 0.00 6.00 0.00 128.00
09:48:52 PM 6.00 0.00 6.00 0.00 80.00
09:48:53 PM 6.06 0.00 6.06 0.00 129.29
09:48:54 PM 5.94 0.00 5.94 0.00 126.73
09:48:55 PM 5.94 0.00 5.94 0.00 126.73
09:48:56 PM 10.17 5.08 5.08 40.68 257.63
09:48:58 PM 3.01 0.00 3.01 0.00 120.30
09:49:00 PM 3.51 0.50 3.01 4.01 52.13
09:49:02 PM 5.82 2.65 3.17 21.16 67.72
09:49:03 PM 8.00 0.00 8.00 0.00
09:49:04 PM 6.00 0.00 6.00 0.00
09:49:05 PM 6.06 0.00 6.06 0.00
09:49:06 PM 6.00 0.00 6.00 0.00
09:49:07 PM 6.00 0.00 6.00 0.00
09:49:08 PM 8.00 0.00 8.00 0.00
09:49:09 PM 8.08 2.02 6.06 16.16
I/O operations per second
tps: total ops
r/wtps: read/write operations
b...: bytes read/written
Can unveil a fabric problem...

Memory statistics

held@mheld:~ - Shell No. 2 - Konsole

Session Edit View Bookmarks Settings Help

```
09:48:47 PM kbmemfree kbmemused  %memused kbbuffers  kbcached kbswpfree kbswpused  %swpused  kbswpcad
09:48:48 PM   1732996    321468     15.65    151480    107048   7212136         0      0.00         0
09:48:49 PM   1547888    506576     24.66    154028    280232   7212136         0      0.00         0
09:48:50 PM   1543956    510508     24.85    157016    278316   7212136         0      0.00         0
09:48:51 PM   1542496    511968     24.92    159108    282744   7212136         0      0.00         0
09:48:52 PM   1542568    511896     24.92    160076    280068   7212136         0      0.00         0
09:48:53 PM   1534512    519952     25.31    161300    286668   7212136         0      0.00         0
09:48:54 PM   1538080    516384     25.13    162128    281824   7212136         0      0.00         0

~~~~~~~~~~

09:52:28 PM   1353904    700560     34.10    342792    280172   7212136         0      0.00         0
09:52:29 PM   1531736    522728     25.44    342824    107812   7212136         0      0.00         0
Average:      1443313    611151     29.75    259045    276074   7212136         0      0.00         0
```

Watch

- %memused and kbmemfree: if short on available memory
- kbswapfree: if not swapped but short on memory
  - the problem is not heap & stack but I/O buffers

Swap rate

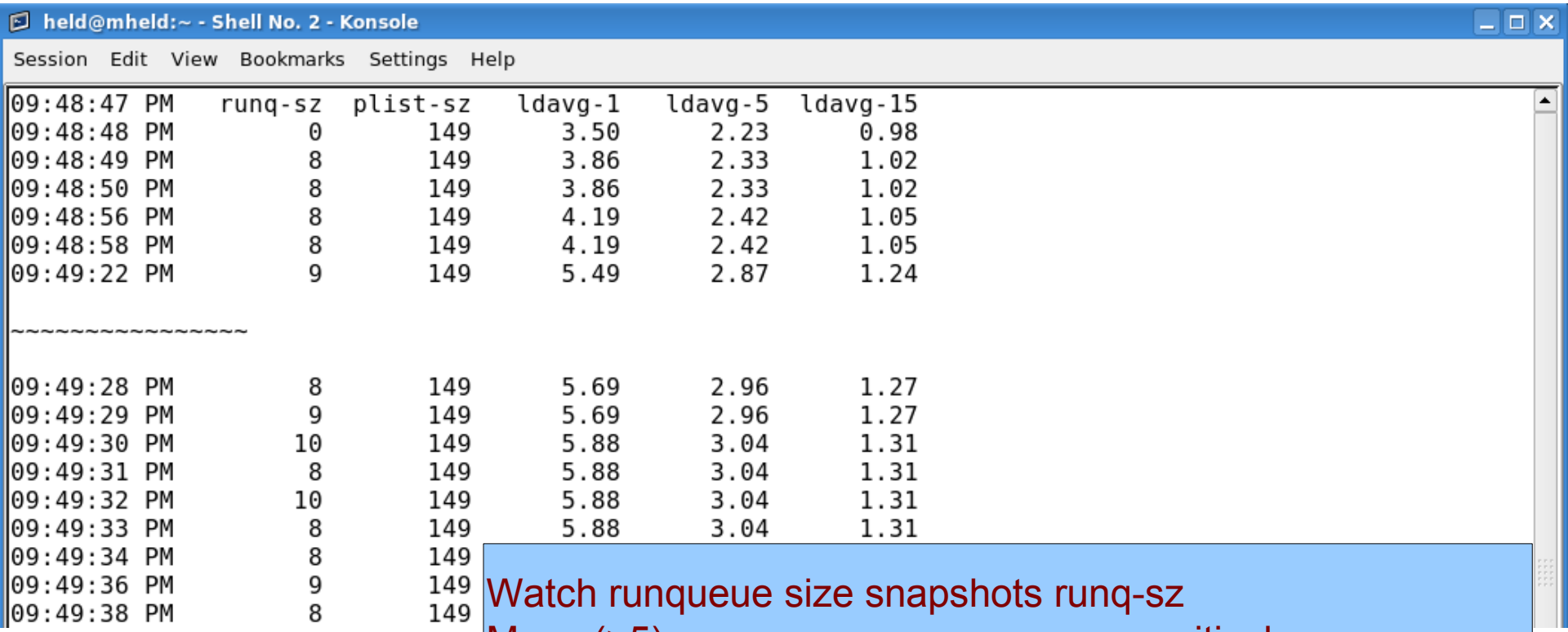
held@mheld:~ - Shell No. 2 - Konsole

Session Edit View Bookmarks Settings Help

```
09:48:47 PM  pswpin/s pswpout/s
09:48:48 PM      0.00      0.00
09:48:49 PM      0.00      0.00
09:48:50 PM      0.00      0.00
09:48:51 PM      0.00      0.00
09:48:52 PM      0.00      0.00
09:48:53 PM      0.00      0.00
09:48:54 PM      0.00      0.00
09:48:55 PM      0.00      0.00
09:48:56 PM      0.00      0.00
09:48:58 PM      0.00      0.00
```

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

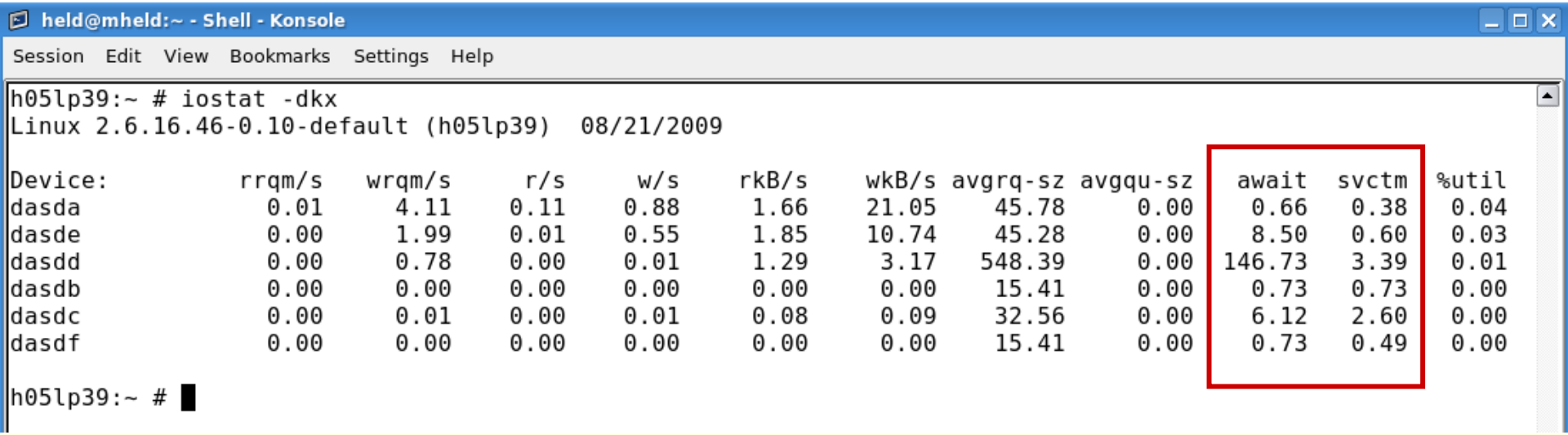
System Load



- 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

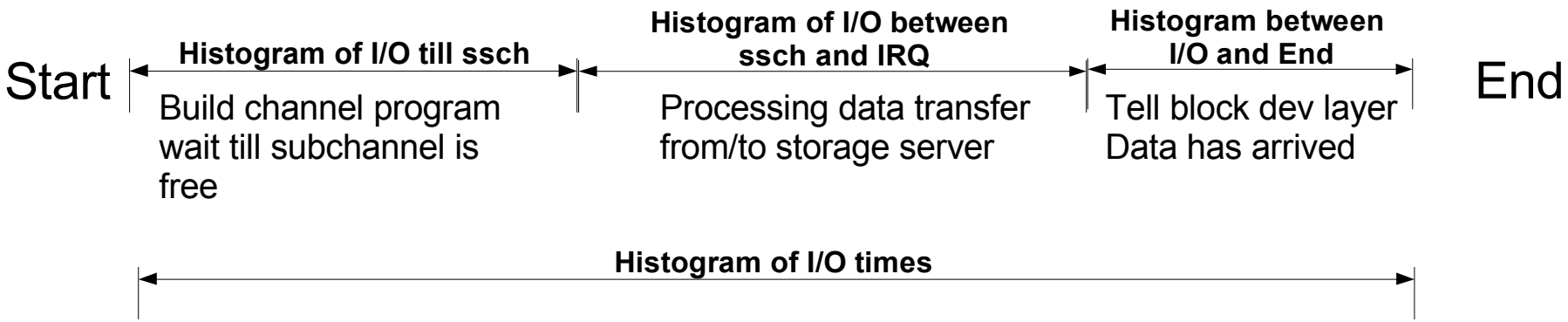
iostat

- iostat: shows averaged performance data per device
  - More detailed decomposition than achieved with sadc
  - Especially watch queue size and await / svctm
    - await (in millisec.): average time for i/o requests issued to the device to be serviced (incl. Time on queue).
    - svctm (in millisec.): average service time for i/o requests that were issued to the device.



Linux DASD statistics

- Collects statistics of DASD I/O operations
  - Histogramm of request sizes
  - Histogramm of processing times
  - Number of requests already chained in channel queue
- Each line represents a histogram of times for a certain operation
- Processing times split up into the following :

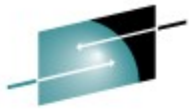


http://www.ibm.com/developerworks/linux/linux390/perf/tuning_how_tools_dasd.html

DASD statistics (cont‘d)

- Linux can collect performance stats on DASD activity as seen by Linux(!)
- Summarized histogram information available in /proc/dasd/statistics
- Turn on with
  `echo on > /proc/dasd/statistics`
- Turn off with
  `echo off > /proc/dasd/statistics`
- To reset: turn off and then on again
- Can be read for the whole system by
  `cat /proc/dasd/statistics`
- Can be read for individual DASDs by
  `tunedasd -P /dev/dasda`

Linux DASD statistics (cont'd)



Seattle SHARE

```
thoss-11:20:27~/temp#cat statistics
36092283 dasd I/O requests
with -1725707784 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 1008619  655629 3360987 2579503 1098338  215814   86155   18022       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  204086  551833  376809  487413  760823 1020219  948881 1447413 1752571
1036560  274399  123980   36916    1162       0       0       0       0       0       0       0       0       0       0       0
Histogram of I/O times per sector
      0    1244  106729  462435  645039  687343  673292 1073946 1697563 1921045 1212557  429291   82078   23062    5681    1409
    345       6       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Histogram of I/O time till ssch
4202149   97492  144602   41229    6349    6189   13122   30505   70775  112524  199203  337873  494914  624231  892960  961439
 513787  173339   80344   19694     343       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  234574 1417573  730299  784908  841778 1158314 1008186 1291285 1148930
 315034   70795   21271     113       6       0       0       0       0       0       0       0       0       0       0       0
Histogram of I/O time between ssch and irq per sector
      0    7572  253750 1291491  863359  967642 1057080 1452901 1692525 1082657  319214   29180    5252     421      22       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
3538030 1224909 2667755  970430  369618  185642   43442   14481    6120    1779     427     202      81      66      39      39
      4       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
# of req in chanq at enqueuing (1..32)
4487074 1970046  987103  687097  891750       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
thoss-11:20:30~/temp#
```

SCSI statistics

- Since SLES9 SP3 (plus maintenance) SCSI statistics can be collected
- The parameter **`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/`**
- The names of these subdirectories consist of
  - **`zfcp-<device-bus-id>`** for an adapter and
  - **`zfcp-<device-bus-id>-<WWPN>-<LUN>`** for a LUN.

SCSI statistics (cont'd)

- Each subdirectory contains two files, a data and a definition file.
- Using

  ```
  echo on=1 > definition
  ```

  the data gathering can be switched on for each device,
- With

  ```
  echo on=0 > definition
  ```

  the gathering is switched off again. It defaults to data gathering being turned off.
- The command

  ```
  echo data=reset > definition
  ```

  enables you to reset the collected data to 0.

SCSI statistics example

```
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
request_sizes_scsi_read 0x2d000 44
request_sizes_scsi_read 0x2a000 26
request_sizes_scsi_read 0x57000 25
request_sizes_scsi_read 0x1e000 25
request_sizes_scsi_read 0x63000 24
request_sizes_scsi_read 0x6f000 19
request_sizes_scsi_read 0x12000 19
...
latencies_scsi_read <=1 1076
latencies_scsi_read <=2 205
latencies_scsi_read <=4 575
latencies_scsi_read <=8 368
latencies_scsi_read <=16 0
...
channel_latency_read <=16000 0
channel_latency_read <=32000 983
channel_latency_read <=64000 99
channel_latency_read <=128000 115
channel_latency_read <=256000 753
channel_latency_read <=512000 106
channel_latency_read <=1024000 141
channel_latency_read <=2048000 27
channel_latency_read <=4096000 0
...
fabric_latency_read <=1000000 1238
fabric_latency_read <=2000000 328
fabric_latency_read <=4000000 522
fabric_latency_read <=8000000 136
more ...
```

Agenda

- In case a problem shows up
  - Describe problem
  - Describe the environment
- Tools used to determine problems
  - Troubleshooting “first-aid kit”
  - sysstat package
  - Disk statistics
- **Customer reported incidents**
  - **Networking: ‘TSM - breaking TCP connections'**
  - **SCSI disk: 'Multipath configuration'**
  - **More customer problems in a nutshell**

Networking:
‘TSM - breaking TCP connections'

- **Problem reporting - advanced:**
  - **Describe your problem:**

“Our backup clients lost connection to the TSM server for several minutes during the overnight backup. Therefore the clients are not able to finish their backups. The problem appears only during our overnight backups.”

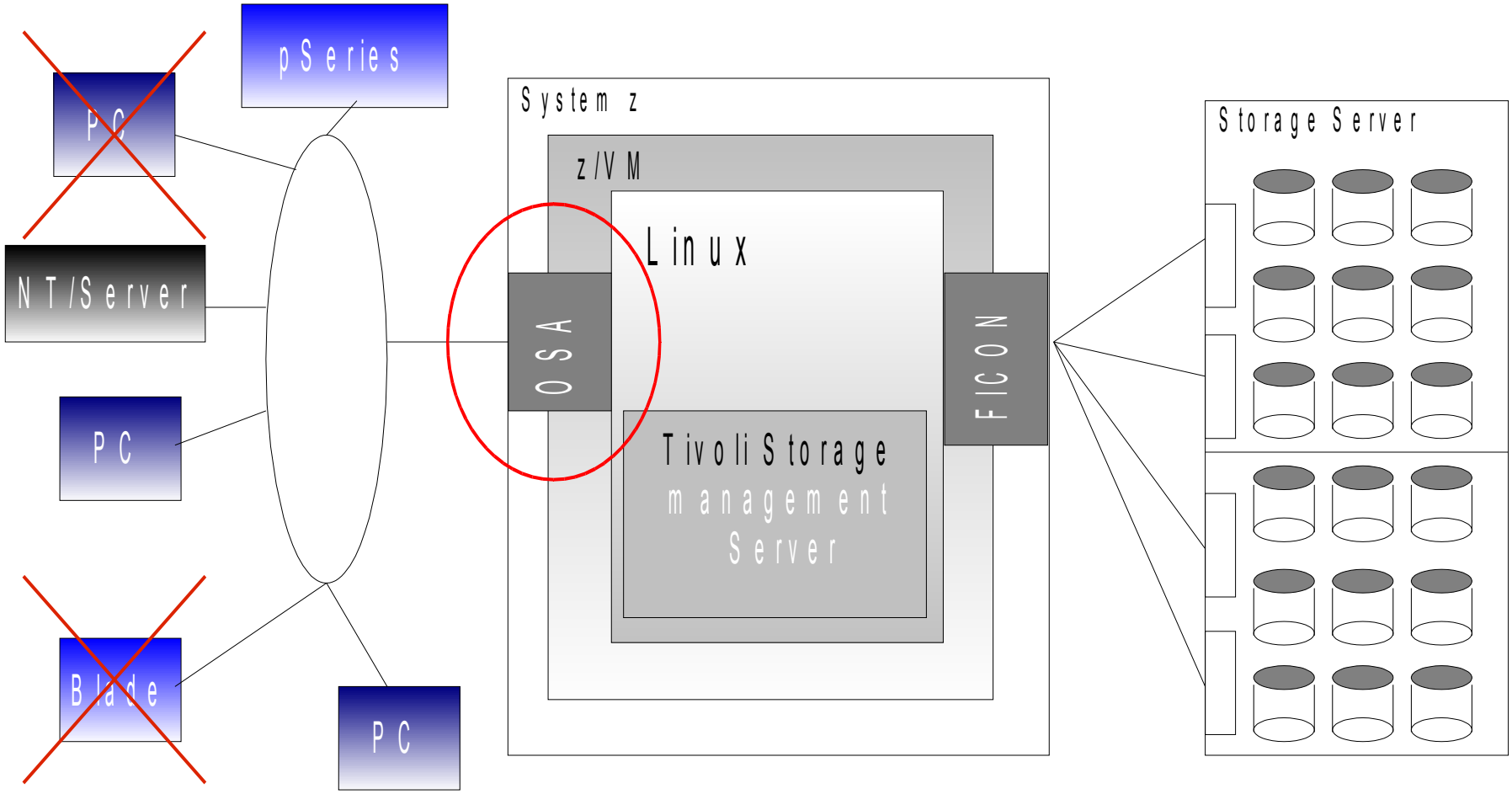
  - **Describe your setup:**

“We are running a TSM Server 5.1 under SLES8 SP3. The Linux runs under z/VM 5.1. The Disk attachment is XYZ storage server connected with 8 FICON channels to the z9 box. A picture about our networking structure is attached”

  - **What recently changed ?**

“It happens every night since we moved our TSM environment from a z990 to a z9 system. We also migrated our disk attachment from a ESS800 to a XYZ system. The operating system and software levels did not change.”

Networking:
‘TSM - breaking TCP connections'



Networking:
'TSM - breaking TCP connections'

- dbginfo.sh collects /var/log/messages at the time of the outages

Seattle SHARE

```
Jan 17 22:40:55 zlinp03 last message repeated 6 times
Jan 17 22:40:55 zlinp03 kernel: NET: 3 messages suppressed.
Jan 17 22:40:55 zlinp03 kernel:  qeth: no memory for packet from eth0
Jan 17 22:40:55 zlinp03 kernel: __alloc_pages: 0-order allocation failed (gfp=0x20/0)
Jan 17 22:40:55 zlinp03 kernel:  qeth: no memory for packet from eth0
Jan 17 22:40:55 zlinp03 kernel: __alloc_pages: 0-order allocation failed (gfp=0x20/0)
Jan 17 22:40:55 zlinp03 kernel:  qeth: no memory for packet from eth0
Jan 17 22:40:55 zlinp03 kernel: __alloc_pages: 0-order allocation failed (gfp=0x20/0)
Jan 17 22:40:55 zlinp03 kernel:  qeth: no memory for packet from eth0
Jan 17 22:40:55 zlinp03 kernel: __alloc_pages: 0-order allocation failed (gfp=0x20/0)
:
```

- And also the contents of Debug Feature for Linux on System z

- **==> /proc/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

Networking: ‘TSM - breaking TCP connections'

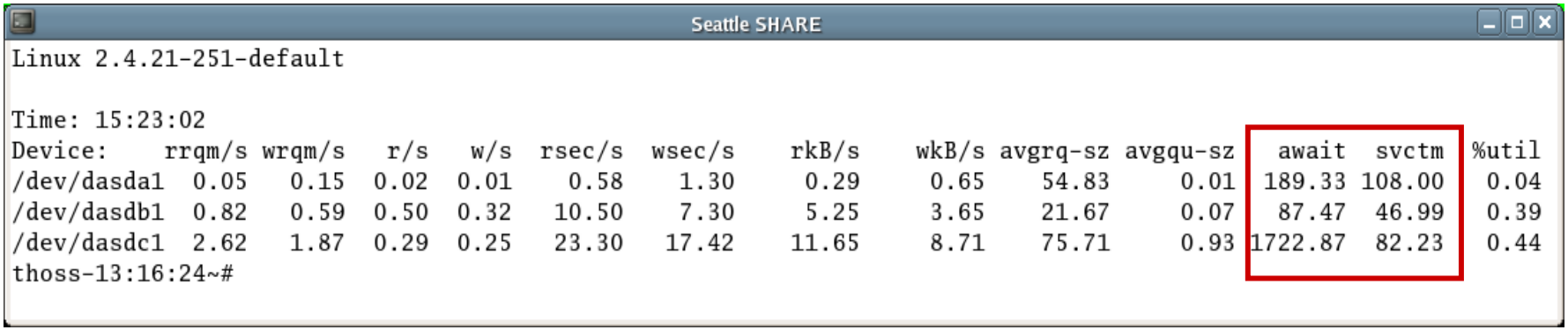
- sadc data collection shows system low on memory at the time of the outages
- Pages in question are not pageable

Seattle SHARE

```
Linux 2.4.21-251-default

23:00:00          CPU     %user     %nice   %system     %idle
23:01:01          all     13.09      0.02     27.33     59.57
23:02:00          all     10.96      0.00     23.20     65.84

23:00:00     pgpgin/s pgpgout/s  activepg  inadtypg  inaclnpg  inatarpg
23:01:01      2738.79  36069.55      8324         0         0         0
23:02:00      2949.09  32550.58      8374         0         0         0

23:00:00          tps      rtps      wtps   bread/s   bwrtn/s
23:01:01       524.22    264.40    259.82   4091.32  14252.31
23:02:00       425.83    274.72    151.11   4435.16   9932.33

23:00:00    kbmemfree kbmemused  %memused kbmemshrd kbbuffers  kbcached kbswpfree kbswpused  %swpused
23:01:01         2724   1029972     99.74         0     27376    537260   2457068         48      0.00
23:02:00         2344   1030352     99.77         0     27400    541240   2457068         48      0.00

23:00:00        IFACE   rxpck/s   txpck/s   rxbyt/s   txbyt/s
23:01:01         eth1 817548.06 1776428.44 66012742.46  37864.67
23:01:01         eth0  25412.79   6994.23 37754460.48 821214.90

thoss-14:14:29~/win/data/vortrag/seattle/data#
```

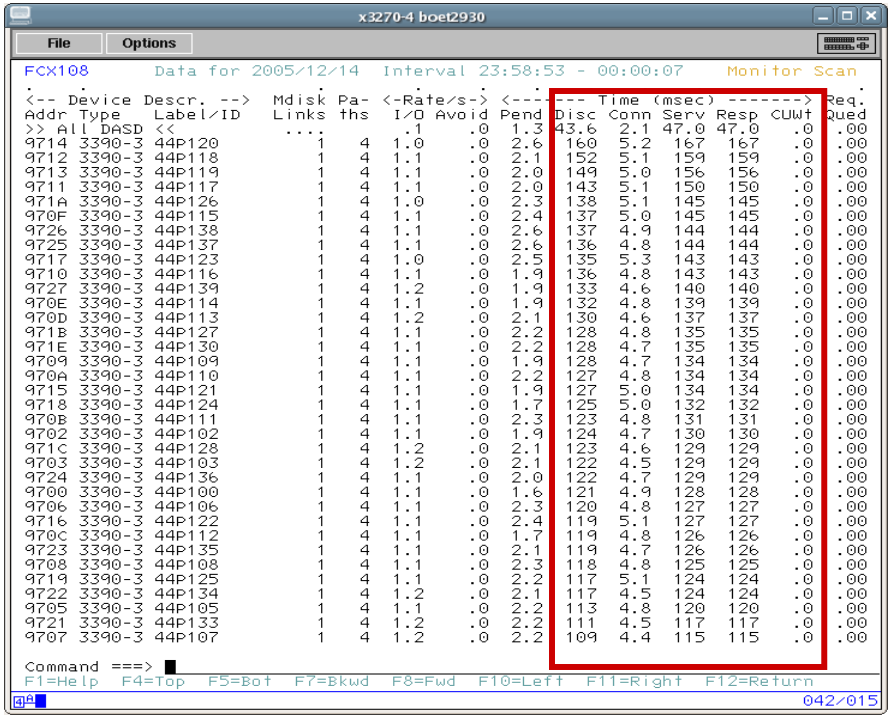
Networking:
‘TSM - breaking TCP connections'

- iostat shows long response times for disk I/O requests on certain devices
  - Good values would be between 8-15ms
    - await: The average time for I/O requests issued to the device to be served.
    - svctm: The average service time for I/O requests that were issued to the device.
  - Keep in mind if you run virtualization like z/VM: this is the Linux view



Networking:
‘TSM - breaking TCP connections'

- Used tool is PerfTK (FCX108)
- z/VM Monitor data shows high service times in disconnected state while FICON channel utilization is rather low
- If you run on z/VM collect information using z/VM tools
- Try to match the information of Linux and z/VM tools



- If you run on z/VM collect information using z/VM tools
- Try to match the information of Linux and z/VM tools

Networking:
‘TSM - breaking TCP connections'

- **Tools used for problem determination:**
  - dbginfo.sh
    - /var/log/messages
    - Linux for System z Debug Feature
  - Linux SADC / SAR and IOSTAT
  - Storage Controller DASD statistics
- **Problem Indicators:**
  - Network connections break, because buffers for inbound packets cannot be allocated due to insufficient memory
  - Disk I/O shows high service time on the storage controller
  - z/VM monitor data show long disconnect times while FICON channels still have capacity.
  - Disks with poor performance are configured as non-full-pack z/VM minidisks
  - Storage Controller statistics data shows large number of cache misses for write operations

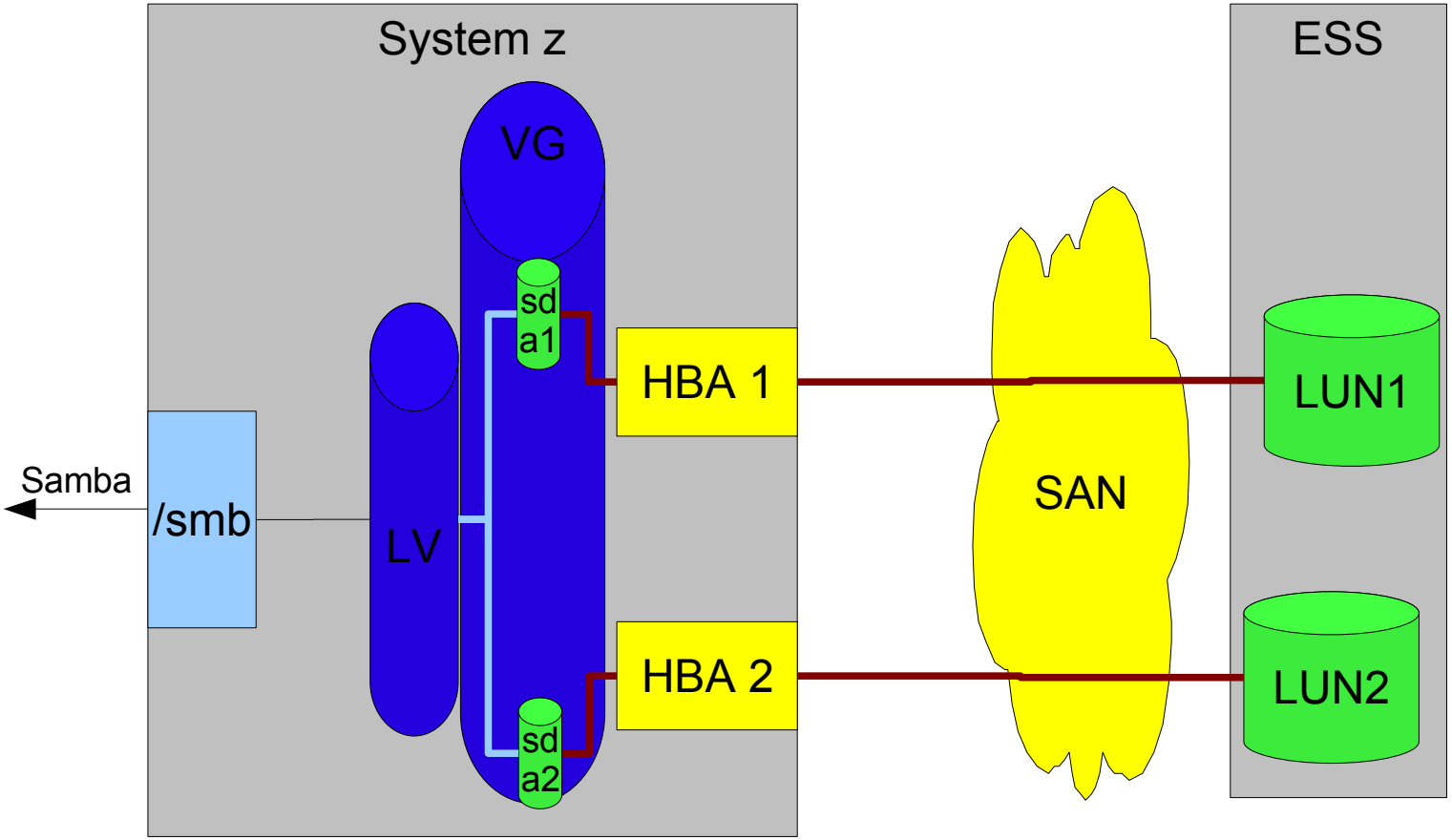
Networking:
‘TSM - breaking TCP connections'

- **Problem origin:**
  - Disk Storage Controller (this one was provided by an independent storage vendor) treated write requests to non-full-pack z/VM minidisks as cache miss and performed a write through operation instead of fast write to NVS cache.
- **Solution / Circumvention:**
  - Use fullpack minidisk or dedicated disk as storage pool

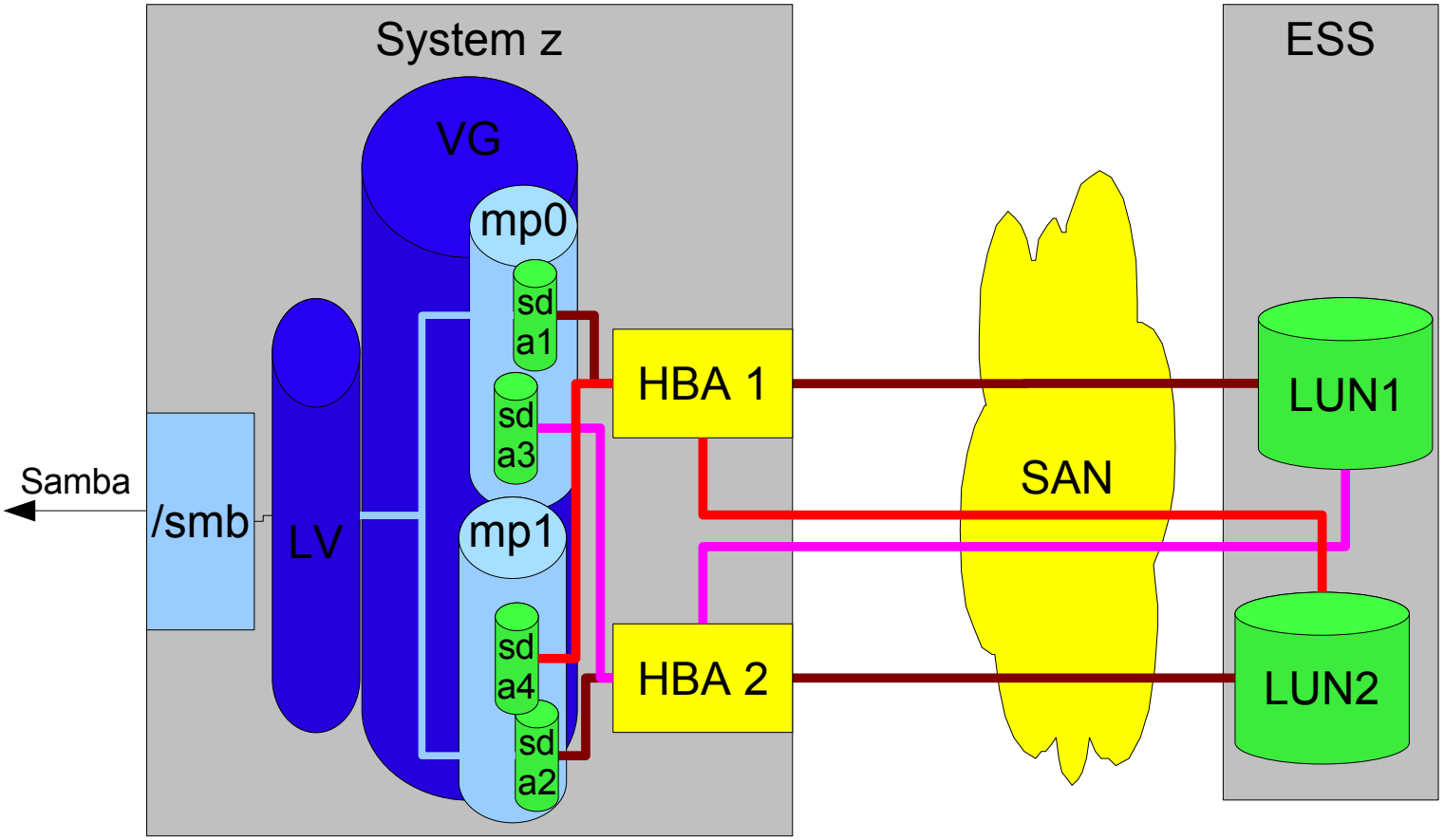
SCSI disk: ‘Multipath configuration’

- **Configuration:**
  - Customer is running Samba server (Samba = file and printer sharing e.g. with Windows clients) 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!

SCSI disk:
‘Multipath configuration’



SCSI disk:
‘Multipath configuration’



SCSI disk: ‘Multipath configuration’

- **Tools used for problem determination:**
  - dbginfo.sh
- **Problem Indicators:**
  - Intermittent outages of disk connectivity
- **Solution:**
  - Configure multipathing correctly:
    - Establish independent paths to each volume
    - Group the paths using the device-mapper-multipath package
    - Base LVM configuration on top of multipath devices instead of sd<#>
  - For a more detailed description how to use FCP attached storage appropriately with Linux on System z see http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26cts02.pdf

More customer problems in a nutshell

Networking:
'Firewall cuts TCP connections'

- **Configuration:**
  - Customer is running Enterprise Removable Media Mgr (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: '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

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.

Cryptography: 'HW not used for AES-256'

- **Configuration:**
  - Customer wants to use Crypto card acceleration 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
  - Use SLES11 on a System z10

http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100810

Links

- Linux on System z project at IBM DeveloperWorks:
  http://www.ibm.com/developerworks/linux/linux390/
- Redbook “Problem Determination for Linux on System z”

  http://www.redbooks.ibm.com/abstracts/sg247599.html

  Table of contents

  Chapter 1. Problem determination methodology

  Chapter 2. Problem determination tools for z/VM

  Chapter 3. Problem determination tools for Linux on System z

  Chapter 4. Network problem determination

  Chapter 5. Performance problem determination

  Chapter 6. Storage problems

  Chapter 7. Eligibility lists

  Chapter 8. Hardware-related problems

  Chapter 9. Installation and setup problems

  Chapter 10. Booting problems

  Chapter 11. Case study: slow responding Web site

  Publish Date 25 August 2008

Questions ?

- Ask right now!
- Submit it by email to
  - Mario Held: mario.held@de.ibm.com
  - Linux S390 mail account: linux390@de.ibm.com
  - Please refer to this presentation

BACKUP


SHARE in Denver
August 23-28, 2009 | Colorado Convention Center | Denver, Colorado
2009 CONFERENCE THEMES: Total Enterprise Virtualization
& Service Orientation: The Foundation for IT Modernization
into your world