2009 System z Expo

October 5 – 9, 2009 – Orlando, FL

Problem determination with Linux on System z

zLG05

Ursula Braun (ursula.braun@de.ibm.com)

IBM Germany Research and Development

Linux on System z Development

Authorized IBM Training

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

- Troubleshooting First aid-kit
- Tools: Sadc / sar , iostat
- Customer reported incidents 2006 – 2009
  - TSM - Network connectivity breaks
  - Guest spontaneously reboots
  - FCP disk configuration issues
  - Disk I/O bottlenecks
  - More customer problems in a nutshell (for reference)
- Service Offerings for pain relief

First Aid Kit

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

Trouble-Shooting First Aid kit

- Install packages required for debugging
  - s390-tools/s390-utils
  - Sysstat (sadc/sar, iostat)
  - Dump tools: lkcdutils, lcrash, crash
- 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
  - Enable /proc/dasd/statistics (see Device Drivers book)

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 (see backup chart)
    - Include System.map and (if available) Kerntypes file from /boot
    - Include vmlinux.gz containing debugging info
  - See “Using the dump tools” book on http://www.ibm.com/developerworks/linux/linux390/development_documentation.html
- In case of a performance problem
  - Enable sadc (System Activity Data Collection) service
  - Collect z/VM MONWRITE Data if running under z/VM
  - Enable DASD statistics:
    See /proc/dasd/statistics on how to enable

Trouble-Shooting First Aid kit (cont'd)

- Attach comprehensive documentation to problem report:
  - Output file of dbginfo.sh
  - 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 asociated to your PMR at
    - ftp://ecurep.mainz.ibm.com/, or
    - ftp://testcase.boulder.ibm.com/
- When opening a Bugzilla at Distribution partner attach documentation to Bugzilla

SADC / SAR
Iostat

Use and configure SADC/SAR and iostat:

- Capture Linux performance data with sysstat package
  - **S**ystem **A**ctivity **D**ata **C**ollector (sadc) --> data gatherer
  - **S**ystem **A**ctivity **R**eport (sar) command --> reporting tool
  - **iostat** command --> I/O utilization
- SADC example (for more see man sadc)
  - /usr/lib64/sa/sadc [options] [interval [count]] **[binary outfile]**
  - /usr/lib64/sa/sadc 5 10 sadc_outfile
  - /usr/lib64/sa/sadc -d 10 >/tmp/sadc_outfile
  - statistics for disk: option **`-d`** resp. **`-s DISK`**
  - Should be started as a service during system start

Use and configure SADC/SAR and iostat: (cont'd)

- SAR example (for more see man sar)
  - sar -A **-->** Analyse data from current sadc data collection
  - sar -A -f sadc_outfile >sar_outfile
- Please include the binary sadc data and sar -A output when submitting SADC information to IBM support
- IOSTAT example (for more see man iostat)
  - 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

sadc/sar - 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                                                      99.57
09:48:55 PM      15                                                       0.00
09:48:56 PM     all                                                      73.35
09:48:56 PM       0                                                      90.68
09:48:56 PM       1                                                      94.74
09:48:56 PM       2                                                      93.07
09:48:56 PM       3                                                       1.00
09:48:56 PM       4                                                      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)

sadc/sar - Processes created

root@t6345020:/var/log/sa

File Edit View Terminal Tabs Help

```
Linux 2.6.9-42.UI310.3.EL (t6345020)    2007-09-03

02:30:01        proc/s
02:40:01          0.19
02:50:01          0.00
03:00:01          0.05
03:10:01          0.06
03:20:01          0.00
03:30:01          0.01
03:40:02          0.06
03:50:01          0.02
04:10:01          0.39
04:20:01          0.00
04:30:01          0.01
04:40:01          0.01
04:50:01          0.00
05:00:01          0.01
05:10:01          0.01
05:20:01          0.00
05:30:01          0.01
05:40:01          0.00
05:50:01          0.01
06:00:01          0.01
:
```

Processes created per second

- usually < 10 except during startup
- if > 100 your application likely has an issue

sadc/sar - Context Switch Rate

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

Session Edit View Bookmarks Settings Help

```
09:48:47 PM   cswch/s
09:48:48 PM    962.83
09:48:49 PM    140.30
09:48:50 PM   1164.36
09:48:51 PM   1180.00
09:48:52 PM   1180.00
09:48:53 PM   1203.03
09:48:54 PM   1129.70
09:48:55 PM   1197.03
09:48:56 PM   1003.39
09:48:58 PM    525.56
09:49:00 PM    522.06
09:49:02 PM    586.51
09:49:03 PM   1137.00
09:49:04 PM   1214.00
09:49:05 PM   1225.25
09:49:06 PM   1078.00
09:49:07 PM   1181.00
```

Context switches per second

- usually < 1000 except during startup
- if > 10000 your application likely has an issue or critical resources are blocked

sadc/sar - Swap rate

root@t6345020:/var/log/sa

File Edit View Terminal Tabs Help

```
02:30:01     pswpin/s pswpout/s
02:40:01         0.00      0.00
02:50:01         0.00      0.00
03:00:01         0.00      0.00
03:10:01         0.00      0.00
03:20:01         0.00      0.00
03:30:01         0.00      0.00
03:40:02         0.00      0.00
03:50:01         0.00      0.00
04:10:01         0.00      0.00
04:20:01         0.00      0.00
04:30:01         0.00      0.00
04:40:01         0.00      0.00
04:50:01         0.00      0.00
05:00:01         0.00      0.00
05:10:01         0.00      0.00
05:20:01         0.00      0.00
05:30:01         0.00      0.00
05:40:01         0.00      0.00
05:50:01         0.00      0.00
06:00:01         0.00      0.00
06:10:01         0.00      0.00
06:20:01         0.00      0.00
:
```

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

sadc/sar - 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    368.00
09:49:04 PM      6.00      0.00      6.00      0.00    144.00
09:49:05 PM      6.06      0.00      6.06      0.00     80.81
09:49:06 PM      6.00      0.00      6.00      0.
09:49:07 PM      6.00      0.00      6.00      0.
09:49:08 PM      8.00      0.00      8.00      0.
09:49:09 PM      8.08      2.02      6.06     16.
```

I/O operations per second
- tps: total ops
- r/wtps: read/write operations
- b...: bytes read/written

Can unveil a fabric problem...

sadc/sar - I/O rates

```
root@t6345020:/var/log/sa
File  Edit  View  Terminal  Tabs  Help
02:30:01          DEV       tps  rd_sec/s  wr_sec/s
02:40:01       dev1-0      0.00      0.00      0.00
02:40:01       dev1-1      0.00      0.00      0.00
02:40:01       dev1-2      0.00      0.00      0.00
02:40:01       dev1-3      0.00      0.00      0.00
02:40:01       dev1-4      0.00      0.00      0.00
02:40:01       dev1-5      0.00      0.00      0.00
02:40:01       dev1-6      0.00      0.00      0.00
02:40:01       dev1-7      0.00      0.00      0.00
02:40:01       dev1-8      0.00      0.00      0.00
02:40:01       dev1-9      0.00      0.00      0.00
02:40:01      dev1-10      0.00      0.00      0.00
02:40:01      dev1-11      0.00      0.00      0.00
02:40:01      dev1-12      0.00      0.00      0.00
02:40:01      dev1-13      0.00      0.00      0.00
02:40:01      dev1-14      0.00      0.00      0.00
02:40:01      dev1-15      0.00      0.00      0.00
02:40:01      dev94-0      0.23      0.19      3.98
02:40:01      dev94-4      0.00      0.00      0.00
02:40:01      dev94-8      0.00      0.00      0.00
02:40:01     dev94-12      0.00      0.00      0.00
02:40:01       dev9-0      0.00      0.00      0.00
02:50:01       dev1-0      0.00      0.00      0.00
:
```

- read/write operations
  - Per I/O device
  - tps: transactions
  - rd/wr_secs: sectors
- Is your I/O balanced?
  - Maybe you should stripe your LVs!

sadc/sar - Networking data (1)

root@t6345020:/var/log/sa

File Edit View Terminal Tabs Help

```
02:30:01        IFACE   rxpck/s   txpck/s   rxbyt/s   txbyt/s   rxcmp/s   txcmp/s  rxmcst/s
02:40:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:40:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:40:01         eth0      0.27      0.25     19.32     63.72      0.00      0.00      0.00
02:50:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:50:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:50:01         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01         eth0      0.08      0.08      5.59     21.17      0.00      0.00      0.00
03:10:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:10:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:10:01         eth0      0.37      0.32     26.13     94.94      0.00      0.00      0.00
03:20:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:20:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00
:
```

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

sadc/sar - Networking data (2)

root@t6345020:/var/log/sa

File Edit View Terminal Tabs Help

```
02:30:01        IFACE   rxerr/s   txerr/s    coll/s  rxdrop/s  txdrop/s  txcarr/s  rxfram/s  rxfifo/s  txfifo/s
02:40:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:40:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:40:01         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:50:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:50:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
02:50:01         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:00:01         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:10:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:10:01         sit0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:10:01         eth0      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:20:01           lo      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00
03:20:01         sit0      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

sadc/sar - 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: short on available memory
- kbswapfree: if not swapped but short on memory
  - the problem is not heap & stack but I/O buffers

sadc/sar - System Load

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

Session Edit View Bookmarks Settings Help

```
09:48:47 PM   runq-sz  plist-sz   ldavg-1   ldavg-5  ldavg-15
09:48:48 PM        0       149      3.50      2.23      0.98
09:48:49 PM        8       149      3.86      2.33      1.02
09:48:50 PM        8       149      3.86      2.33      1.02
09:48:56 PM        8       149      4.19      2.42      1.05
09:48:58 PM        8       149      4.19      2.42      1.05
09:49:22 PM        9       149      5.49      2.87      1.24

~~~~~~~~~~~~~~~~

09:49:28 PM        8       149      5.69      2.96      1.27
09:49:29 PM        9       149      5.69      2.96      1.27
09:49:30 PM       10       149      5.88      3.04      1.31
09:49:31 PM        8       149      5.88      3.04      1.31
09:49:32 PM       10       149      5.88      3.04      1.31
09:49:33 PM        8       149      5.88      3.04      1.31
09:49:34 PM        8       149      5.88      3.04      1.31
09:49:36 PM        9       149
09:49:38 PM        8       149
```

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

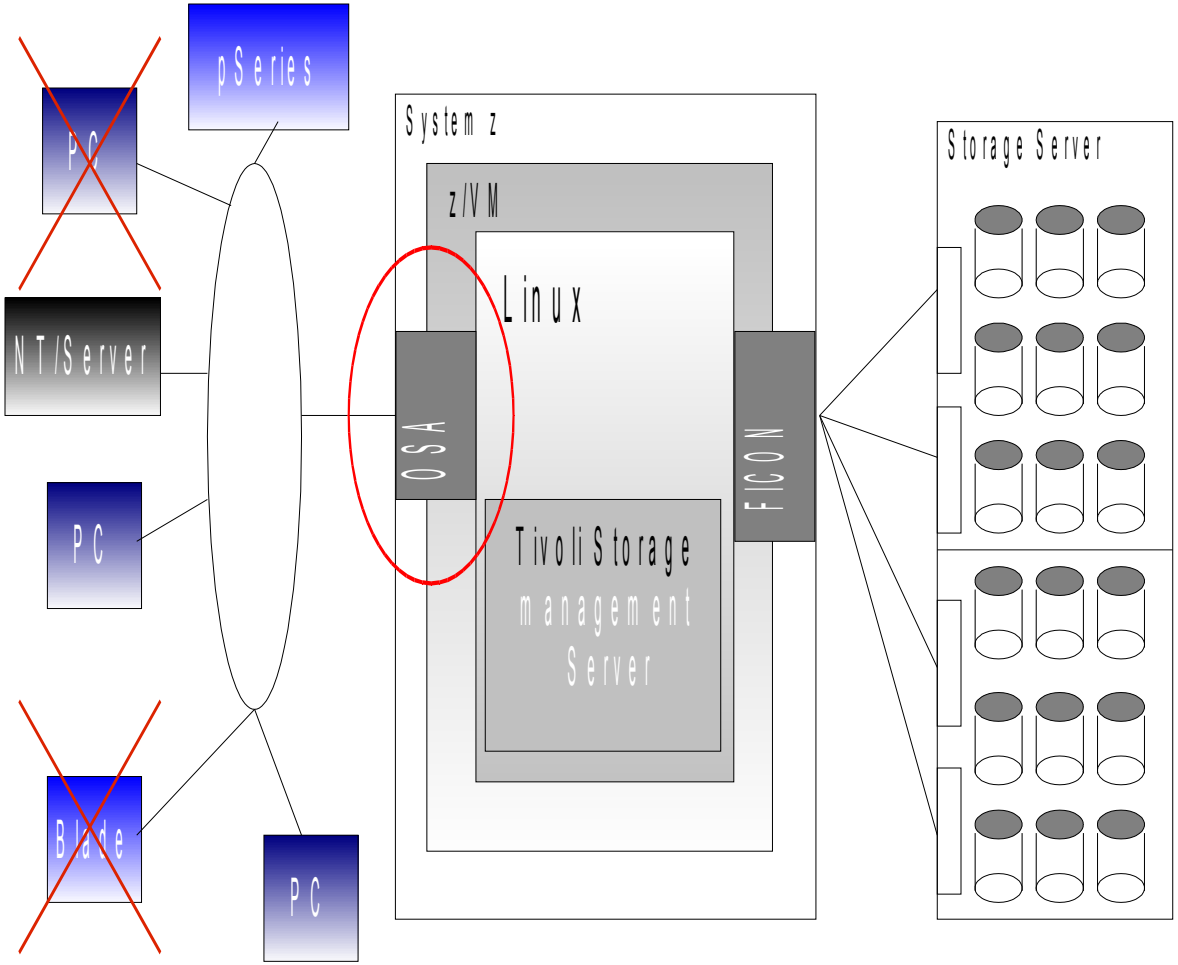
Customer Incidents

Introductory Remarks

- The incidents reported here are real customer incidents
  - Out of years 2006 - 2009
  - Red Hat Enterprise Linux, and Novell Linux Enterprise Server distributions
  - Linux running in LPAR and z/VM of different versions
- While problem analysis look 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.

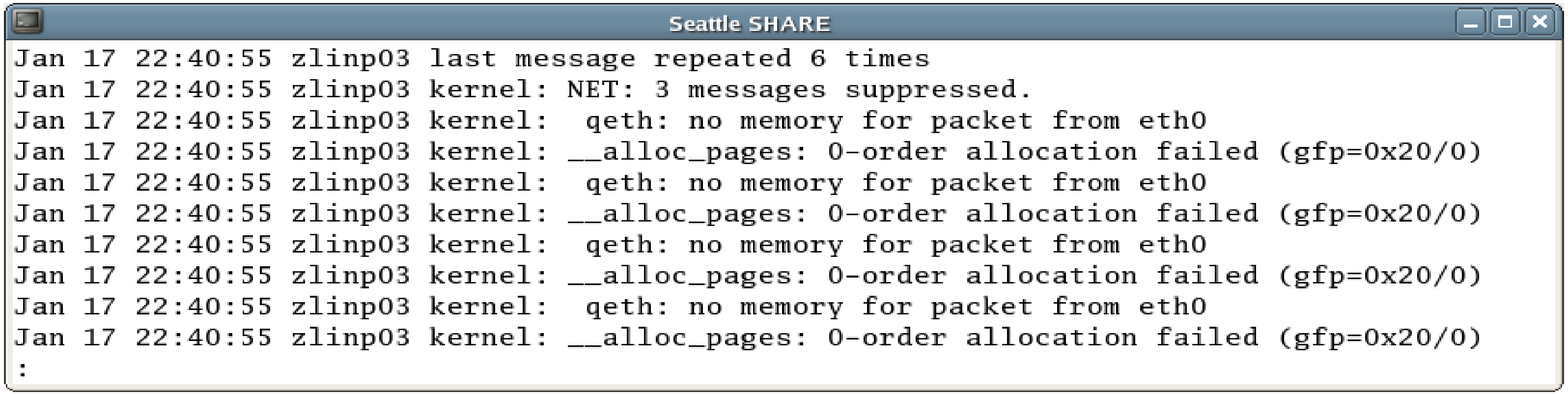
Networking: 'TSM - breaking TCP connections'

- Configuration:
  - Customer is running TSM backup over LAN with storage pool on minidisks provided by vendor supplied storage controller
- Problem Description:
  - During overnight backup runs the TSM clients report backup failure due to TCP/IP disconnect



Networking:
‘TSM - breaking TCP connections'

- dbginfo.sh collects /var/log/messages
  - Look at the time of the outages
  - Here messages show directly, why inbound network packets get lost



```
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)
:
```

Networking: 'TSM - breaking TCP connections'

- dbginfo.sh also collects 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
```

```
/usr/src/linux/drivers/s390/qeth.c
} else {          skb=qeth_get_skb(length);
                  if (!skb) goto nomem;}
nomem:
          sprintf(dbf_text,"NOMM%4x",card->irq0);
          QETH_DBF_TEXT0(0,trace,dbf_text);
```

Networking: 'TSM - breaking TCP connections'

- SADC data collection shows system low on memory at the time of the outages

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 averaged performance data per device
  - More detailled decomposition than with sadc
  - Watch queue size and await/svctm
- iostat shows long response times for disk I/O requests on certain devices
  - Good values would be between 8-15ms

Seattle SHARE

```
Linux 2.4.21-251-default

Time: 15:23:02
Device:    rrqm/s wrqm/s   r/s   w/s  rsec/s  wsec/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
/dev/dasda1  0.05   0.15  0.02  0.01    0.58    1.30     0.29     0.65    54.83     0.01  189.33 108.00   0.04
/dev/dasdb1  0.82   0.59  0.50  0.32   10.50    7.30     5.25     3.65    21.67     0.07   87.47  46.99   0.39
/dev/dasdc1  2.62   1.87  0.29  0.25   23.30   17.42    11.65     8.71    75.71     0.93 1722.87  82.23   0.44
thoss-13:16:24~#
```

Networking: 'TSM - breaking TCP connections'

- z/VM Monitor data shows high service times in disconnected state while FICON channel utilization is rather low
- Try to match information of Linux and z/VM tools

x3270-4 boet2930

File Options

FCX108 Data for 2005/12/14 Interval 23:58:53 - 00:00:07 Monitor Scan

<-- Device Descr. --> Addr	Type	Label/ID	Mdisk Links	Pa- ths	<-Rate/s-> I/O	Avo	d	<------- Time (msec) ---> Pend	Disc	Conn	Serv	Resp	CUWt	Req. Qued
>> All DASD <<			....		.1		.0	1.3	43.6	2.1	47.0	47.0	.0	.00
9714	3390-3	44P120	1	4	1.0		.0	2.6	160	5.2	167	167	.0	.00
9712	3390-3	44P118	1	4	1.1		.0	2.1	152	5.1	159	159	.0	.00
9713	3390-3	44P119	1	4	1.1		.0	2.0	149	5.0	156	156	.0	.00
9711	3390-3	44P117	1	4	1.1		.0	2.0	143	5.1	150	150	.0	.00
971A	3390-3	44P126	1	4	1.0		.0	2.3	138	5.1	145	145	.0	.00
970F	3390-3	44P115	1	4	1.1		.0	2.4	137	5.0	145	145	.0	.00
9726	3390-3	44P138	1	4	1.1		.0	2.6	137	4.9	144	144	.0	.00
9725	3390-3	44P137	1	4	1.1		.0	2.6	136	4.8	144	144	.0	.00
9717	3390-3	44P123	1	4	1.0		.0	2.5	135	5.3	143	143	.0	.00
9710	3390-3	44P116	1	4	1.1		.0	1.9	136	4.8	143	143	.0	.00
9727	3390-3	44P139	1	4	1.2		.0	1.9	133	4.6	140	140	.0	.00
970E	3390-3	44P114	1	4	1.1		.0	1.9	132	4.8	139	139	.0	.00
970D	3390-3	44P113	1	4	1.2		.0	2.1	130	4.6	137	137	.0	.00
971B	3390-3	44P127	1	4	1.1		.0	2.2	128	4.8	135	135	.0	.00
971E	3390-3	44P130	1	4	1.1		.0	2.2	128	4.7	135	135	.0	.00
9709	3390-3	44P109	1	4	1.1		.0	1.9	128	4.7	134	134	.0	.00
970A	3390-3	44P110	1	4	1.1		.0	2.2	127	4.8	134	134	.0	.00
9715	3390-3	44P121	1	4	1.1		.0	1.9	127	5.0	134	134	.0	.00
9718	3390-3	44P124	1	4	1.1		.0	1.7	125	5.0	132	132	.0	.00
970B	3390-3	44P111	1	4	1.1		.0	2.3	123	4.8	131	131	.0	.00
9702	3390-3	44P102	1	4	1.1		.0	1.9	124	4.7	130	130	.0	.00
971C	3390-3	44P128	1	4	1.2		.0	2.1	123	4.6	129	129	.0	.00
9703	3390-3	44P103	1	4	1.2		.0	2.1	122	4.5	129	129	.0	.00
9724	3390-3	44P136	1	4	1.1		.0	2.0	122	4.7	129	129	.0	.00
9700	3390-3	44P100	1	4	1.1		.0	1.6	121	4.9	128	128	.0	.00
9706	3390-3	44P106	1	4	1.1		.0	2.3	120	4.8	127	127	.0	.00
9716	3390-3	44P122	1	4	1.1		.0	2.4	119	5.1	127	127	.0	.00
970C	3390-3	44P112	1	4	1.1		.0	1.7	119	4.8	126	126	.0	.00
9723	3390-3	44P135	1	4	1.1		.0	2.1	119	4.7	126	126	.0	.00
9708	3390-3	44P108	1	4	1.1		.0	2.3	118	4.8	125	125	.0	.00
9719	3390-3	44P125	1	4	1.1		.0	2.2	117	5.1	124	124	.0	.00
9722	3390-3	44P134	1	4	1.2		.0	2.1	117	4.5	124	124	.0	.00
9705	3390-3	44P105	1	4	1.1		.0	2.2	113	4.8	120	120	.0	.00
9721	3390-3	44P133	1	4	1.2		.0	2.2	111	4.5	117	117	.0	.00
9707	3390-3	44P107	1	4	1.2		.0	2.2	109	4.4	115	115	.0	.00

Command ===>

F1=Help F4=Top F5=Bot F7=Bkwd F8=Fwd F10=Left F11=Right F12=Return

042/015

Networking: 'TSM - breaking TCP connections'

- Tools used for problem determination:
  - dbginfo.sh
  - Linux for System z Debug Feature
  - Linux SADC/SAR and IOSTAT
  - Linux DASD statistics
  - Storage Controller DASD statistics

Networking: 'TSM - breaking TCP connections'

- 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
  - Observed here, but not relevant: Paging space almost unused, because all memory is used for TSM I/O buffers, which are not pageable.

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:
  - Use fullpack minidisk or dedicated disk as storage pool
  - For optimal disk configuration see http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_optimizedisk.html

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

- 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

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:
  - I/O performance is worse than expected by projecting single disk benchmark to more complex solution

Performance: 'disk I/O bottlenecks'

- Tools used for problem determination:
  - dbginfo.sh
  - Linux for System z Debug Feature
  - Linux SADC/SAR and IOSTAT
  - Linux DASD statistics
  - z/VM monitor data
  - Storage Controller DASD statistics
- Problem Indicators:
  - Multi-disk performance is worse than projected single-disk performance.

Performance: 'disk I/O bottlenecks'

- 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 (e.g. for swapspace!)
    - For optimal disk performance configurations read and take into account

      http://www.ibm.com/developerworks/linux/linuxs390/perf/tuning_rec_dasd_optimizedisk.html

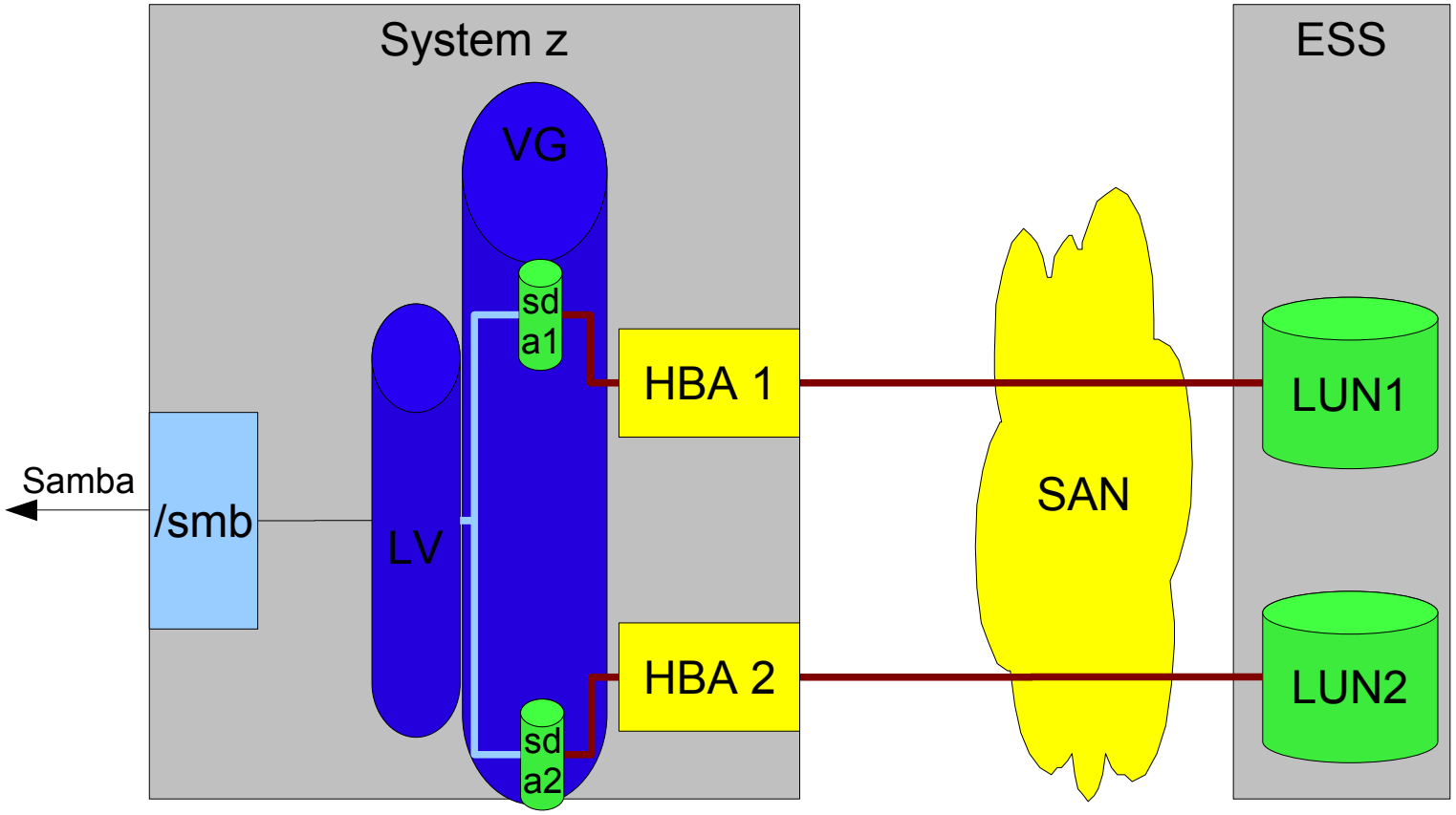
FCP disk: ‘multipath configuration’

- 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!

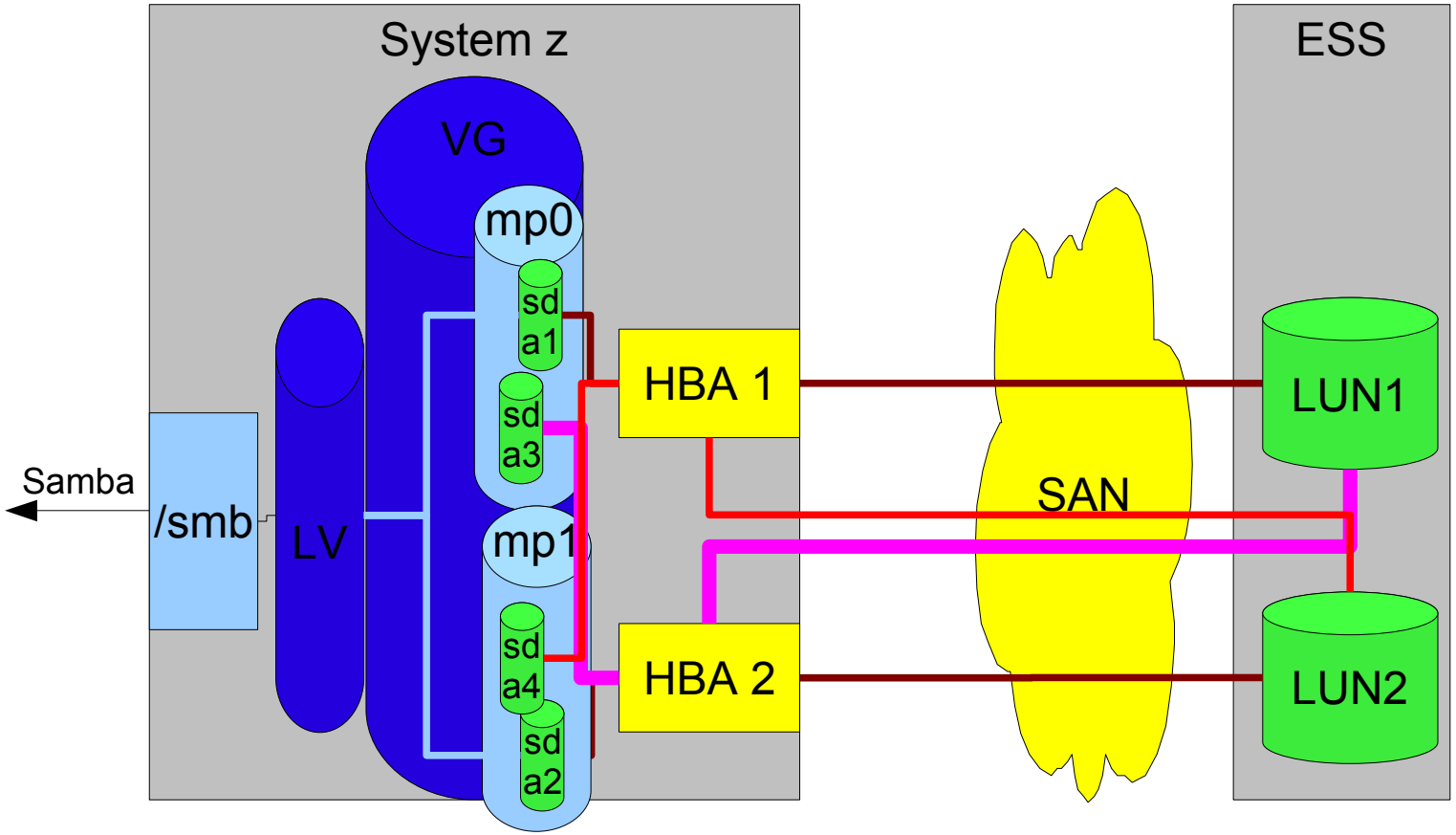
FCP disk: ‘multipath configuration’

- Tools used for problem determination:
  - dbginfo.sh
- Problem Indicators:
  - Intermittent outages of disk connectivity

FCP disk: ‘multipath configuration’



FCP disk: ‘multipath configuration’



FCP disk: ‘multipath configuration’

- Solutions
  - Configure multipathing correctly:
    - Establish independent paths to each volume
    - Group the paths using the device-mapper-multipath package
    - Base LVM configuration on top of mpath 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

Service Offerings For Problem Avoidance

Under construction:
Lab based service offerings for our customers

- System Quality Assurance (“Health Check”)
  - Proactive check of system configuration
  - Risk assessment
  - Recommendations to optimize configuration
- Performance and Capacity Monitoring
  - Assess and plan system capacity and performance
  - Proactively tune/augment system
- Your ideas for more other offerings are welcome!
  - Michael Daubman/Poughkeepsie/IBM
  - Holger Smolinski/Germany/IBM

Your feedback and questions:

- Raise it right now!
- Write it on the feedback sheets!
- Submit it by email to
  - Ursula Braun (ursula.braun@de.ibm.com)
  - Sven Schuetz (sven@de.ibm.com)
  - linux390@de.ibm.com
  - Please refer to this presentation

Backup:
More customer problems
In a nutshell

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.
- Solution:
  - Configure partition /dev/dasda1 as swapspace

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:
  - lscss
- Problem Origin:
  - Unused subchannels delay IPL
  - Hotplug event processing takes very long
- Solution:
  - Use cio_ignore to restrict system to used subchannels

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
```

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

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

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: '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.

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.

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’

- 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 throughput 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'

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

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.

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

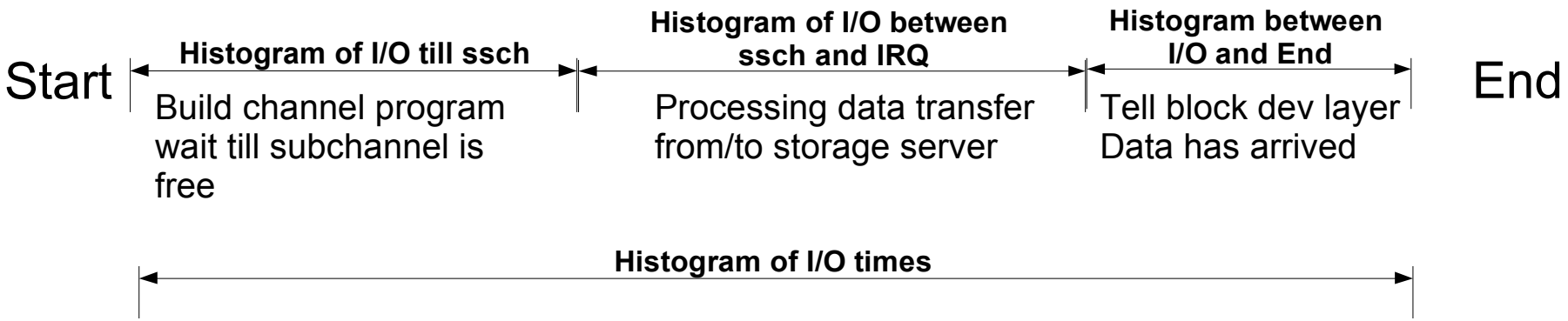
NFS: NFS write to Z/OS server is slow

- 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 Indicators:
  - 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

Backup: Miscellaneous

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

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#
```

DASD statistics (cont'd)

- DASD statistics decomposition
  - Summarized histogram information available in /proc/dasd/statistics
  - Also accessible per device via `BIODASDPRRD` and `BIODASDPRRST` ioctls

```
typedef struct dasd_profile_info_t {
        unsigned int dasd_io_reqs;       /* number of requests processed at all */
        unsigned int dasd_io_sects;      /* number of sectors processed at all */
        unsigned int dasd_io_secs[32];   /* histogram of request's sizes */
        unsigned int dasd_io_times[32];  /* histogram of requests's times */
        unsigned int dasd_io_timps[32];  /* histogram of requests's times per
sector */
        unsigned int dasd_io_time1[32];  /* histogram of time from build to start
*/
        unsigned int dasd_io_time2[32];  /* histogram of time from start to irq */
        unsigned int dasd_io_time2ps[32]; /* histogram of time from start to irq */
        unsigned int dasd_io_time3[32];  /* histogram of time from irq to end */
        unsigned int dasd_io_nr_req[32]; /* histogram of # of requests in chanq */
} dasd_profile_info_t;
```

Storage Controller Cache Statistics

- Available on selected distributions:
- ioctl BIODASDPSRD, returning:

```
typedef struct dasd_rssd_perf_stats_t {
        unsigned char  invalid:1;
        unsigned char  format:3;
        unsigned char  data_format:4;
        unsigned char  unit_address;
        unsigned short device_status;
        unsigned int   nr_read_normal;
        unsigned int   nr_read_normal_hits;
        unsigned int   nr_write_normal;
        unsigned int   nr_write_fast_normal_hits;
        unsigned int   nr_read_seq;
        unsigned int   nr_read_seq_hits;
        unsigned int   nr_write_seq;
        unsigned int   nr_write_fast_seq_hits;
        unsigned int   nr_read_cache;
        unsigned int   nr_read_cache_hits;
        unsigned int   nr_write_cache;
        unsigned int   nr_write_fast_cache_hits;
        unsigned int   nr_inhibit_cache;
        unsigned int   nr_bybass_cache;
        unsigned int   nr_seq_dasd_to_cache;
        unsigned int   nr_dasd_to_cache;
        unsigned int   nr_cache_to_dasd;
        unsigned int   nr_delayed_fast_write;
        unsigned int   nr_normal_fast_write;
        unsigned int   nr_seq_fast_write;
        unsigned int   nr_cache_miss;
        unsigned char  status2;
        unsigned int   nr_quick_write_promotes;
        unsigned char  reserved;
        unsigned short ssid;
        unsigned char  reseved2[96];
} __attribute__((packed)) dasd_rssd_perf_stats_t;
```

- Shows details about storage controller cache utilization
  - Nr or R/W requests and corrsponding cache hits
- Available through storage controller interface (Controller HMC) or Linux ECKD device driver as an ioctl.

Dump Tools Summary

Tool	Stand alone tools			VMDUMP
	DASD	Tape	SCSI	
Environment	VM&LPAR		LPAR	VM
Preparation	Zipl -d /dev/<dump_dev>		Mkdir /dumps/mydumps zipl -D /dev/sda1 ...	---
Creation	Stop CPU & Store status ipl <dump_dev_CUU>			Vmdump
Dump medium	ECKD or FBA	Tape cartridges	LINUX file system on a SCSI disk	VM reader
Copy to filesystem	Zgetdump /dev/<dump_dev> > dump_file		---	Dumpload ftp ... vmconvert ...
Viewing	Lcrash or crash			

See “Using the dump tools” book on
http://www-128.ibm.com/developerworks/linux/linux390/index.html

Links

- Linux on System z project at IBM DeveloperWorks:
  http://www.ibm.com/developerworks/linux/linux390/
- HW and SW level requirements for QIOASSIST:

  http://www.vm.ibm.com/perf/aip.html
- Fixed I/O buffers with z/VM 5.1:

  http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_fixed_io_buffers.html
- Optimize disk configuration for performance:
  http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_optimizedisk.html
- DASD cache bit tuning:

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