



Problem Determination with Linux on System z



Steffen Thoss
IBM

Session 9279

Agenda

- Troubleshooting First aid-kit
- Customer reported incidents
 - Storage Controller caching strategies
 - TSM - Network connectivity breaks
 - Disk I/O bottlenecks
 - FCP disk configuration issues
 - z/VM 5.1 problem with memory shortage < 2GB
- More customer problems: in a nutshell
- Ideas to give relief

Introductory Remarks



- The incidents reported here are real customer incidents
 - Out of years 2006 and 2007
 - 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
- This presentation focuses on how the tools have been used, comprehensive documentation on their capabilities is in the docs of the corresponding tool.

Trouble-Shooting First Aid kit



- Install packages required for debugging
 - s390-tools
 - sysstat
 - Lkcdutils (SUSE), crash (RedHat)
- 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 Monitor Data if running under z/VM when appropriate
 - Enable /proc/dasd/statistics (see Device Drivers book)

Trouble-Shooting First Aid kit (cont'd)

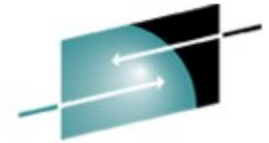
- When System hangs
 - Take a dump
<http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26cdt02.pdf>
- In case of a performance problem
 - Enable sadc (System Activity Data Collection) service
 - Collect z/VM Monitor Data if running under z/VM
 - Enable DASD statistics:
See `/proc/dasd/statistics` on how to enable
- Function does not work as expected
 - Enable extended tracing in `/proc/s390dbf` or `/sys/s390dbf` for subsystem

Trouble-Shooting First Aid kit (cont'd)



- Attach comprehensive documentation to problem report:
 - Output file of dbginfo.sh
 - 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://ecurep.mainz.ibm.com/>
 - <ftp://testcase.boulder.ibm.com/>
 - See Instructions: <http://www-05.ibm.com/de/support/ecurep/other.html>
- When opening a Bugzilla at Distribution partner attach documentation to Bugzilla

- Customer reported incidents



SHARE

Technology • Connections • Results



MORITZ WELER, tamedien

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

Use and configure SADC/SAR and iostat:

- Capture Linux performance data with **sysstat** package
 - Part of the Distro (but might be not pre-installed)
 - **System Activity Data Collector** (sadc)
 - **System Activity Report** (sar) command
 - **iostat** command
- SADC example (for more see man sadc)
 - `/usr/lib/sa/sadc <interval> <count> <binary outfile>`
 - `/usr/lib/sa/sadc 5 10 sadc_outfile`
 - Should be started as a service during system start
- SAR example (for more see man sar)
 - `sar -A -->` Analyse data from current sadc data collection
- IOSTAT example (for more see man iostat)
 - `iostat -dkx -->` Analyse io related performance data for all disks
- **Please include the binary sadc data and sar -A output when submitting SADC information to IBM support**

SADC/SAR Demo

```
thoss@takeda:~  
File Edit View Terminal Tabs Help  
Linux 2.6.16.54-0.2.5-default (t2930035) 02/10/2008  
  
10:41:27 AM      proc/s  
10:41:29 AM      0.00  
10:41:31 AM      0.00  
10:41:33 AM      0.00  
10:41:35 AM      0.00  
10:41:37 AM      0.00  
10:41:39 AM      0.00  
10:41:41 AM      0.00  
10:41:43 AM      0.00  
10:41:45 AM      0.00  
Average:         0.00  
  
10:41:27 AM      cswch/s  
10:41:29 AM      37.50  
10:41:31 AM      15.76  
10:41:33 AM      12.00  
10:41:35 AM      11.50  
10:41:37 AM      15.00  
10:41:39 AM      12.94  
10:41:41 AM      14.57  
10:41:43 AM      10.00  
10:41:45 AM      16.67  
Average:         15.54  
  
10:41:27 AM      CPU      %user    %nice    %system  %iowait  %steal   %idle  
10:41:29 AM      all      0.00     0.00     0.14     0.00     0.00     99.86  
10:41:29 AM      0        0.00     0.00     0.00     0.00     0.00     100.00  
10:41:29 AM      1        0.00     0.00     0.00     0.00     0.00     100.00  
10:41:29 AM      2        0.00     0.00     0.00     0.00     0.00     100.00  
10:41:29 AM      3        0.00     0.00     0.00     0.00     0.00     100.00  
lines 1-32
```

lostat

- `lostat`: shows averaged performance data per device
 - More detailed decomposition than achieved with `sadc`
 - Especially watch queue size and `await/svctm`

```
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   kB/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~#
```

Linux DASD statistics



SHARE

Technology • Connections • Results

```

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 chang at enqueueing (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 chang */  
} 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    reseed2[96];
} __attribute__((packed)) dasd_rssd_perf_stats_t;

```

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

Performance: 'disk cache bits settings'

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

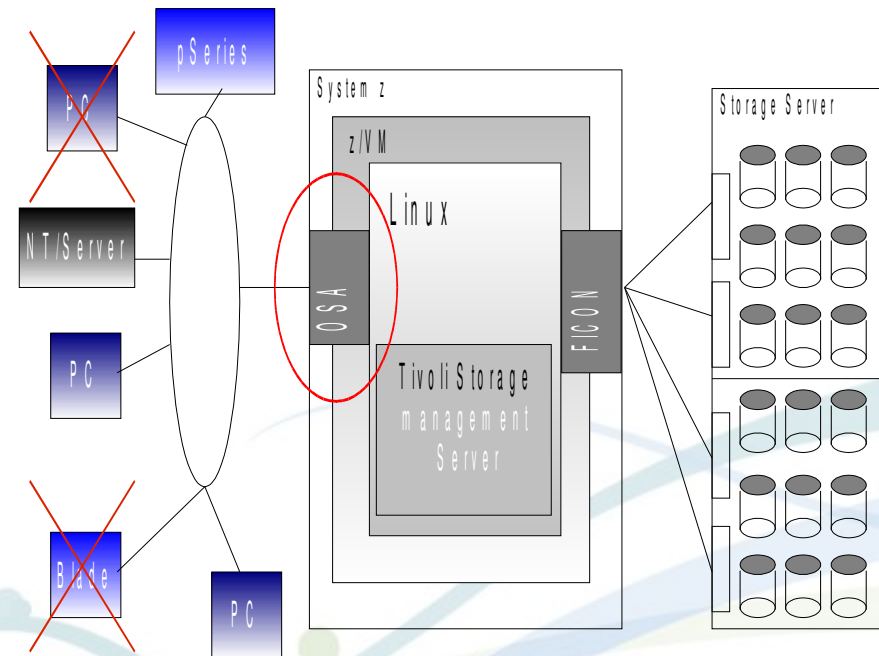
Networking: 'TSM - breaking TCP connections'

- **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'

- **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'

- 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)
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

```
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      ppggin/s  ppggout/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'

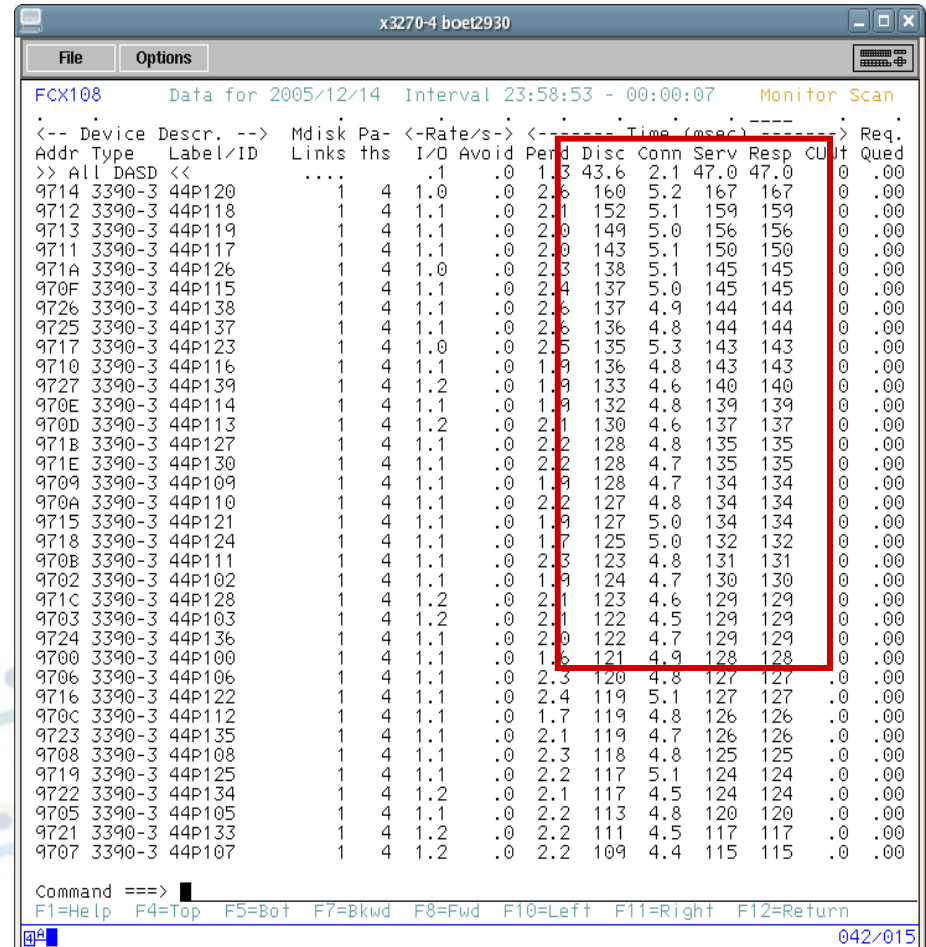
- iostat shows long response times for disk I/O requests on certain devices
 - Good values would be between 8-15ms
 - **Await:** The average time for I/O requests issued to the device to be served.
 - **Svctm:** The average service time for I/O requests that were issued to the device.

```
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   kB/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



The screenshot shows a z/VM Monitor window titled 'x3270-4 boet2930'. The window displays data for 'FCX108' on '2005/12/14' with an interval of '23:58:53 - 00:00:07'. The data is presented in a table format with columns for Device, Descr., Mdisk, Pa, Rate/s, Time (ms), and Req. A red box highlights the 'Time (ms)' column, which shows high values (e.g., 43.6, 160, 152, 149, 143, 138, 137, 136, 135, 136, 143, 143, 133, 132, 130, 127, 128, 128, 127, 125, 124, 124, 123, 122, 122, 122, 120, 119, 119, 118, 118, 117, 117, 113, 113, 111, 109) indicating high service times in the disconnected state.

Device	Descr.	Mdisk	Pa	Rate/s	Time (ms)	Req.
>> All DASD <<						
9714	3390-3 44P120	1	4	1.0	43.6	47.0
9712	3390-3 44P118	1	4	1.1	160	167
9713	3390-3 44P119	1	4	1.1	152	159
9711	3390-3 44P117	1	4	1.1	149	156
971A	3390-3 44P126	1	4	1.0	143	150
970F	3390-3 44P115	1	4	1.1	138	145
9726	3390-3 44P138	1	4	1.1	137	145
9725	3390-3 44P137	1	4	1.1	136	144
9717	3390-3 44P123	1	4	1.0	135	143
9710	3390-3 44P116	1	4	1.1	136	143
9727	3390-3 44P139	1	4	1.2	133	140
970E	3390-3 44P114	1	4	1.1	132	139
970D	3390-3 44P113	1	4	1.2	130	137
971B	3390-3 44P127	1	4	1.1	128	135
971E	3390-3 44P130	1	4	1.1	128	135
9709	3390-3 44P109	1	4	1.1	128	134
970A	3390-3 44P110	1	4	1.1	127	134
9715	3390-3 44P121	1	4	1.1	127	134
9718	3390-3 44P124	1	4	1.1	125	132
970B	3390-3 44P111	1	4	1.1	123	131
9702	3390-3 44P102	1	4	1.1	124	130
971C	3390-3 44P128	1	4	1.2	123	129
9703	3390-3 44P103	1	4	1.2	122	129
9724	3390-3 44P136	1	4	1.1	122	129
9700	3390-3 44P100	1	4	1.1	121	128
9706	3390-3 44P106	1	4	1.1	120	127
9716	3390-3 44P122	1	4	1.1	119	127
970C	3390-3 44P112	1	4	1.1	119	126
9723	3390-3 44P135	1	4	1.1	119	126
9708	3390-3 44P108	1	4	1.1	118	125
9719	3390-3 44P125	1	4	1.1	117	124
9722	3390-3 44P134	1	4	1.2	117	124
9705	3390-3 44P105	1	4	1.1	113	120
9721	3390-3 44P133	1	4	1.2	111	117
9707	3390-3 44P107	1	4	1.2	109	115

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

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 swap space!)
 - For optimal disk performance configurations read and take into account http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_optimizedisk.html

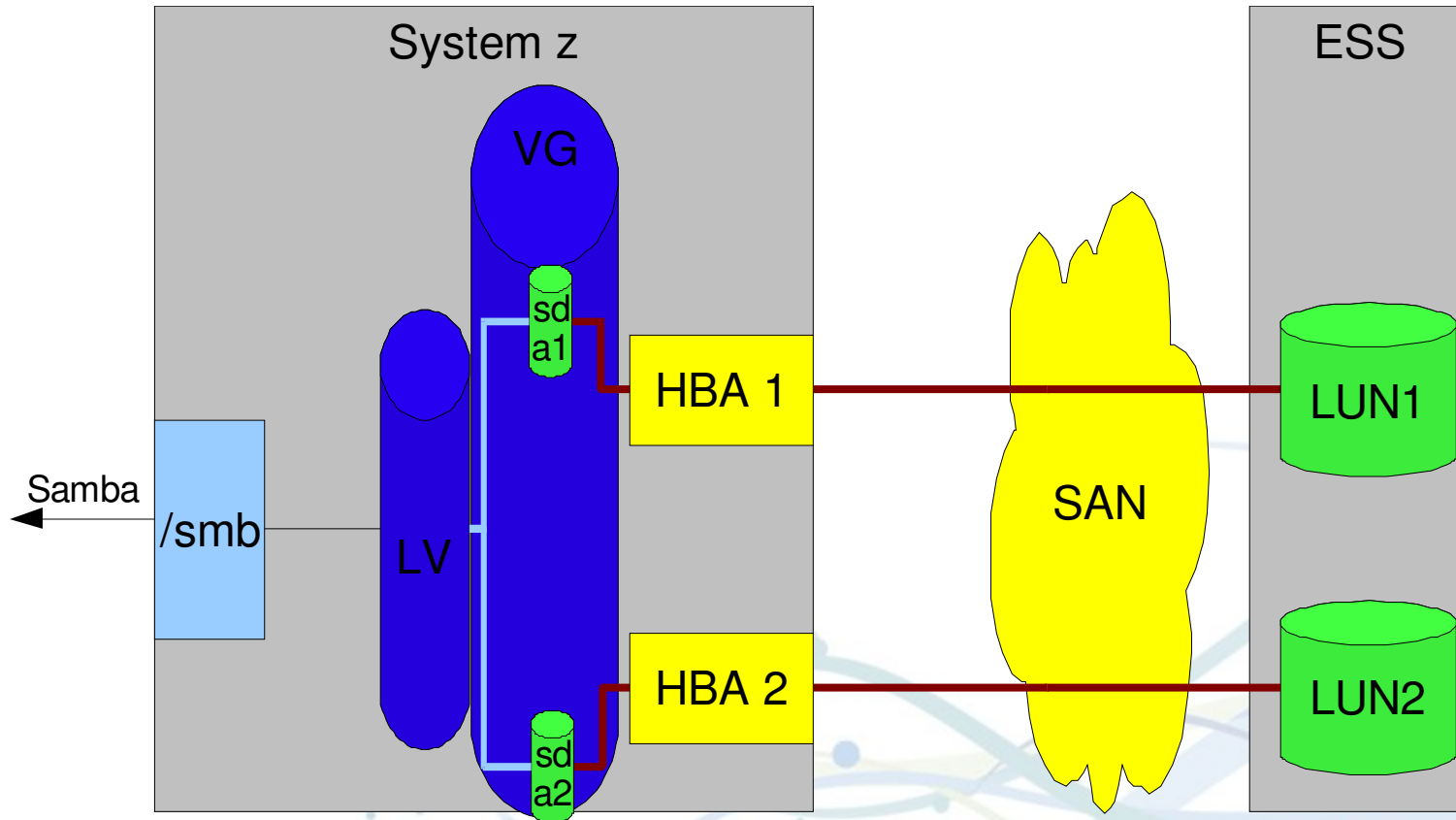
FCP disk: 'multipath configuration'

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

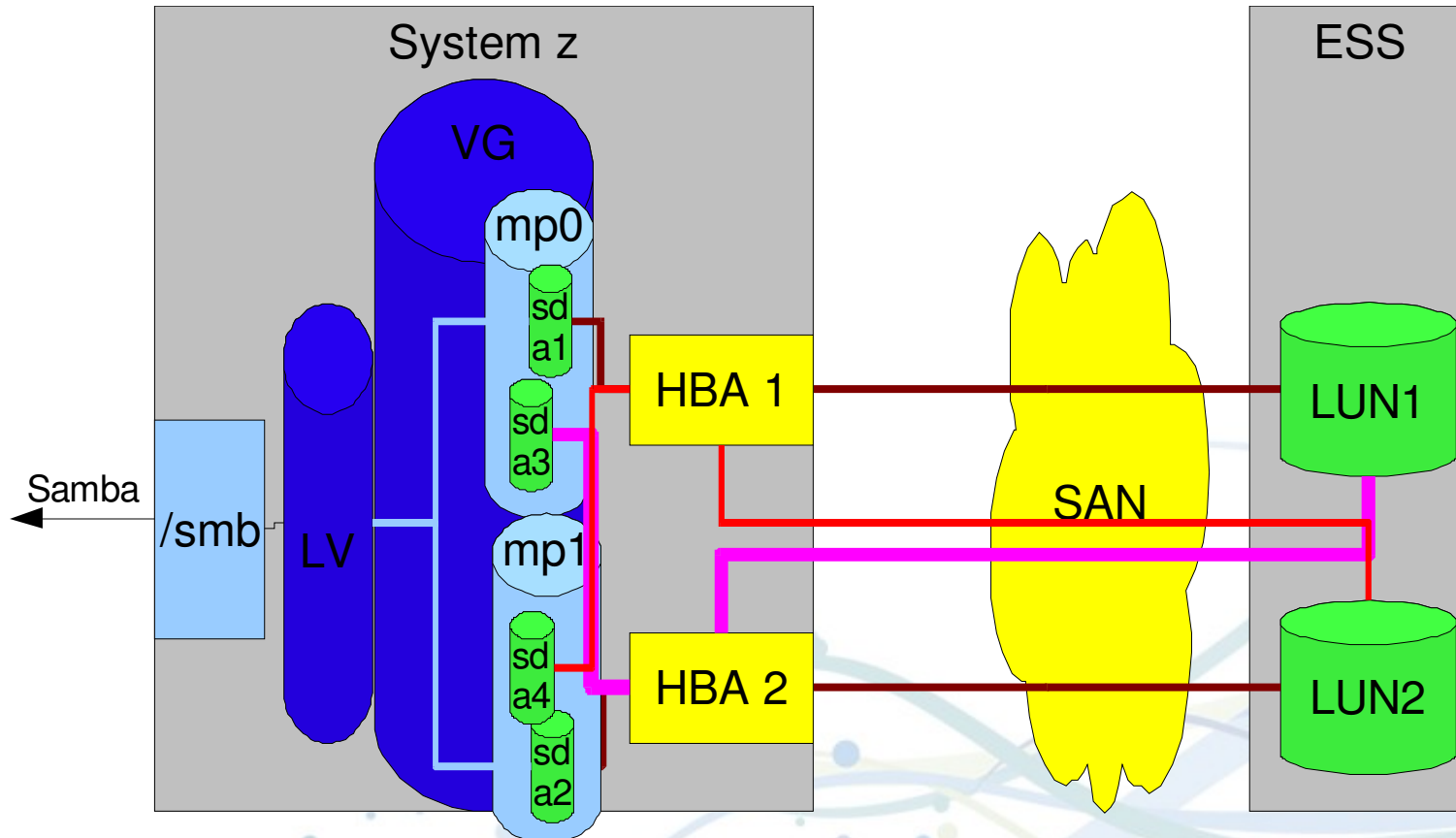
FCP disk: 'multipath configuration'

- **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>

Performance:

'z/VM 5.1- 2GB problem'

- **Configuration:**

- This Customer was running Informix database on a SLES 9 guest hosted by z/VM 5.1
- Other applications and distributions are also affected.

- **Problem Description:**

- Performance scales well up to a certain point, at which the system responsiveness immediately drops very close to zero.

Performance:

'z/VM 5.1- 2GB problem'

- **Tools used for problem determination:**
 - System Observation by users
 - **z/VM MONITOR** data
- **Problem Indicators:**
 - **System Observation:** System throughput drops to almost zero
 - **z/VM MONITOR data:** XSTOR highly utilized (> 90 %)
 - **z/VM MONITOR data:** High Page Migration rate from <2GB to XSTOR (> 300 pg/sec)

Performance: 'z/VM 5.1- 2GB problem'

```
x3270-4 boet2930
File Options
FCX124 Performance Screen Selection (FL520 BASE ) Monitor Scan

General System Data      I/O Data      History Data (by Time)
1. CPU load and trans.   11. Channel load  31. Graphics selection
2. Storage utilization  12. Control units 32. History data files*
3. Reserved             13. I/O device load* 33. Benchmark displays*
4. Priv. operations     14. CP owned disks* 34. Correlation coeff.
5. System counters     15. Cache extend. func.* 35. System summary*
6. CP IUCV services    16. DASD I/O assist 36. Auxiliary storage
7. SPPOOL file display* 17. DASD seek distance* 37. CP communications*
8. LPAR data           18. I/O prior. queueing* 38. DASD load
9. Shared segments     19. I/O configuration 39. Minidisk cache*
A. Shared data spaces  1A. I/O config. changes 3A. Storage mgmt. data*
B. Virt. disks in stor. 21. User resource usage* 3B. Proc. load & config*
C. Transact. statistics 22. User paging load* 3C. Logical part. load
D. Monitor data        23. User wait states* 3D. Response time (all)*
E. Monitor settings   24. User response time* 3E. RSK data menu*
F. System settings    25. Resources/transact.* 3F. Scheduler queues
G. System configuration 26. User communication* 3G. Scheduler data
H. VM Resource Manager 27. Multitasking users* 3H. SFS/BFS logs menu*
I. Exceptions         28. User configuration* 3I. System log
K. User defined data* 29. Linux systems*    3K. TCP/IP data menu*
                    3L. User communication
                    3M. User wait states

Pointers to related or more detailed performance data
can be found on displays marked with an asterisk (*).

Command ==> 2
F1=Help F4=CP F5=Bot F7=Bkwd F8=Fwd F12=Return
042/016
```

Performance: 'z/VM 5.1- 2GB problem'

```
x3270-4 boet2930
File Options
FCX103 Data for 2005/12/14 Interval 22:28:53 - 22:29:53 Monitor Scan

Main storage utilization:
Total real storage 19'456MB
Total available 19'456MB
Offline storage frames 0kB
SYSGEN storage size 19'456MB
CP resident nucleus 9'556kB
Shared storage 3'316kB
FREE storage pages 24'532kB
FREE stor. subpools 7'500kB
Subpool stor. utilization 94%
Total DPA size 1'962MB
Locked pages 79'576kB
Trace table 700kB
Pageable 1'884MB
Storage utilization 32%
Tasks waiting for a frame 0
Tasks waiting for a page 0/s

V=R area:
Size defined 0kB
FREE storage 0kB
V=R recovery area in use ...%
V=R user .....

Paging / spooling activity:
Page moves <2GB for trans. 9/s
Fast path page-in rate 15/s
Long path page-in rate 0/s
Long path page-out rate 19/s
Page read rate 0/s
Page write rate 0/s
Page read blocking factor ...
Page write blocking factor ...
Migrate-out blocking factor ...
Paging SSCH rate 0/s
SPOOL read rate 0/s
SPOOL write rate 0/s

XSTORE utilization:
Total available 6'144MB
Att. to virt. machines 0kB
Size of CP partition 6'144MB
CP XSTORE utilization 14%
Low threshold for migr. 1'200kB
XSTORE allocation rate 19/s
Average age of XSTORE blks 10177s
Average age at migration ...s

MDCACHE utilization:
Min. size in XSTORE 0kB
Max. size in XSTORE 0kB
Ideal size in XSTORE 0kB
Act. size in XSTORE 0kB
Bias for XSTORE 1.00
Min. size in main stor. 0kB
Max. size in main stor. 1'024MB
Ideal size in main stor. 1'024MB
Act. size in main stor. 1'022MB
Bias for main stor. 1.00
MDCACHE limit / user 18'724kB / 5
Users with MDCACHE inserts 5
MDISK cache read rate 15/s
MDISK cache write rate ...../s
MDISK cache read hit rate 1/s
MDISK cache read hit ratio 9%

VDISKS:
System limit (blocks) 245760

Command ===> █
F1=Help F4=Top F5=Bot F7=Bkwd F8=Fwd F12=Return
042/015
```

Be alerted when
page moves > 700/sec and
XSTOR utilization > 90%

Performance:

'z/VM 5.1- 2GB problem'

- **Problem origin:**

- Certain I/O algorithms up to z/VM 5.1 need data buffers below 2GB line
 - Data buffers need to be moved 'below the line' and remain there until I/O is over
 - Not only guest I/O is affected but also z/VM paging etc...

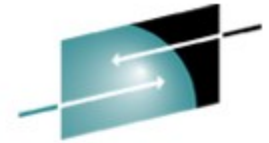
- **Workaround might help:**

- Configure sufficient XSTOR to your system
 - Rule of Thumb: Divide configured storage as 75/25 (Central Storage/XSTOR)
 - XSTOR is still needed for 64bit systems, since z/VM algorithms are tuned for XSTOR usage

- **Solution:**

- Upgrade to z/VM 5.2
 - I/O algorithms have been adapted to use any storage address for data buffers
- Or: Apply service to SLES 9 and enable 'fixed I/O buffers'
 - See: http://www.ibm.com/developerworks/linux/linux390/perf/tuning_rec_fixed_io_buffers.html

More customer problems: In a nutshell



SHARE
Technology • Connections • Results



tsamedien

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 syscall 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 accelerator 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 μ Code 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.

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 μ Code 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

Ideas for relief

- Provide more webpages like shown on the 'Links'-chart
- Create regular Linux on System z Newsletter
 - Provide information about current hot topics
 - Hints & tips about system configuration
- Establish Linux & VM Health Check Offerings for the GEOs
 - Proactive check of system configuration
 - Risk assessment
 - Recommendations to optimize configuration
- Your ideas are welcome!

Your feedback and questions:

- Raise it right now!
- Write it on the feedback sheets!
- Submit it by email to
 - Steffen Thoss (thoss@de.ibm.com)
 - Holger Smolinski (smolinski@de.ibm.com)
 - linux390@de.ibm.com
 - Please refer to this presentation

- **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

Dump Tools Summary



SHARE

Technology • Connections • Results

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>

SHARE Orlando 9279, Steffen Thoss