

# Linux on System z What to do when there is a problem



Session L17  
IBM System z9 and zSeries Expo Orlando 2006  
Oct. 9-13

**Ursula Braun ([braunu@de.ibm.com](mailto:braunu@de.ibm.com))**  
IBM Development Lab, Boeblingen, Germany

# Trademarks

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

Enterprise Storage Server

ESCON\*

FICON

FICON Express

HiperSockets

IBM\*

IBM logo\*

IBM eServer

Netfinity\*

S/390\*

VM/ESA\*

WebSphere\*

z/VM

zSeries

\* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Intel is a trademark of the Intel Corporation in the United States and other countries.

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

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.

Linux is a registered trademark of Linus Torvalds.

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

Penguin (Tux) compliments of Larry Ewing.

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

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

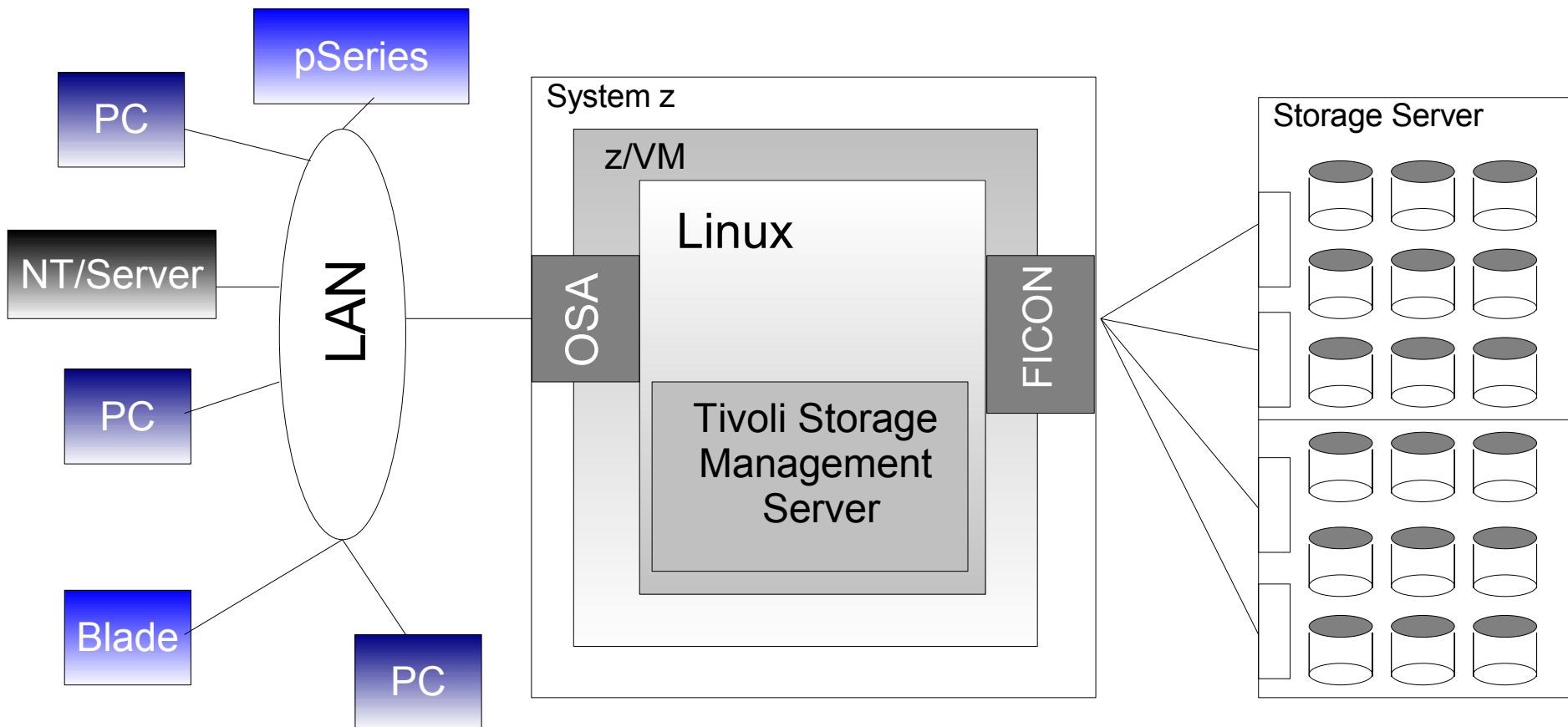
All other products may be trademarks or registered trademarks of their respective companies.



## Agenda

- A customer scenario – The customer problem !
- Collect Information about Problem and Setup
  - ◆ Problem description
  - ◆ Hardware setup/Infrastructure
  - ◆ Linux & z/VM setup information
- Handle a Linux Crash
  - ◆ Linux dump (z/VM and LPAR)
  - ◆ Crash & Lcrash
- Debug a performance problem
  - ◆ Linux messages & s390dbf
  - ◆ Sysstat
  - ◆ Perfsvm and raw monitor data
  - ◆ DASD statistics and tunedasd

# A customer scenario



Task: backup clients every night on TSM Server



## The customer problem

- The customer:

“Our backup clients lost connection to the TSM server for several minutes during the overnight backup. Therefore the clients are not able to finish their backups. Please help us !!”

- You:

What do you think ? Where is the problem !!



## A problem appears

**Get as much information as possible about the circumstances:**

- What is the problem ?
- When did it appear (Date, Time) ?
- Where did it appear (on which systems) ?
- How frequently does it appear ?
- Have I changed something ?
- Has somebody else changed something ?
- Is it a problem ?
- Is it reproducible ?

**→ Write down as much information as possible about the problem !**

# Describe your problem and setup

- Describe your problem:
  - “Our backup clients lost connection to the TSM server for several minutes during the overnight backup. Therefore the clients are not able to finish their backups. The problem appears only during our overnight backups. Another problem we observed is a system crash during enabling the OSA layer 2 feature”
- Describe your setup:
  - “We are running a TSM Server 5.1 under SLES8 SP3. The Linux runs under z/VM 5.1. The Disk attachment is an IBM DS8000 storage server connected with 8 FICON channels to the z9 box. A picture about our networking structure is attached”
- What changed ?
  - “It happens every night since we have moved our TSM environment from a z990 to a z9 system. We also have migrated our disk attachment from an ESS800 to a DS8000 system. The Operating system and Software levels have not changed.”



# Collect information about HW setup

- Machine Setup
  - ◆ Machine type (z9 s38, z990 ...)
  - ◆ Storage Server (ESS800, DS8000)
  - ◆ Storage attachment (FICON, ESCON, FCP, how many channels)
  - ◆ Network (OSA (type, mode), Hipersocket, zVM GuestLAN)
  - ◆ ...
- Infrastructure setup
  - ◆ Clients
  - ◆ Other Computer Systems
  - ◆ Network topologies
  - ◆ ...



# Collect technical setup information

- Linux:
  - ◆ Run script **dbginfo.sh** (part of the **s390-tools** package in SUSE)
  - ◆ RedHat does not include it yet
  - ◆ s390-tools package can be downloaded at our developer works page
  - ◆ **dbginfo.sh** captures the following information:
    - ★ `/proc/[version,cpuinfo,meminfo,slabinfo,modules,partitions,devices ...]`
    - ★ System z specific device driver information: `/proc/s390dbf`
    - ★ Kernel messages: `/var/log/messages`
    - ★ Config files `/etc/[ccwgroup.conf,chandev.conf,modules.conf,fstab]`
    - ★ Several commands: `ps`, `vgdisplay`, `dmesg`
    - ★ Will be enhanced in the future
  - ◆ DASD setup
    - ★ `/proc/dasd/devices` or `lsdasd` (part of s390-tools package)
    - ★ In case of LVM: `lvdisplay`, `vgdisplay`



## Collect technical setup information

- Linux:
  - ◆ Draw a picture of your network setup if possible
  - ◆ Run `lsqeth` (part of s390-tools package)

```
t2930036:~ # lsqeth
```

```
Device name           : eth0
```

```
-----
```

```
card_type             : OSD_1000
```

```
cdev0                 : 0.0.f5f0
```

```
cdev1                 : 0.0.f5f1
```

```
cdev2                 : 0.0.f5f2
```

```
chpid                 : 76
```

```
online                : 1
```

```
portname              : OSAPORT
```

```
portno                : 0
```

```
route4                : no
```

```
route6                : no
```

```
layer2                : 0
```

# Collect technical setup information

- 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

```
t2930036:~ # modprobe vmcp
t2930036:~ # vmcp 'q cplevel'
z/VM Version 5 Release 2.0, service level 0501 (64-bit)
Generated at 01/18/06 06:48:57 CET
IPL at 01/18/06 07:22:09 CET

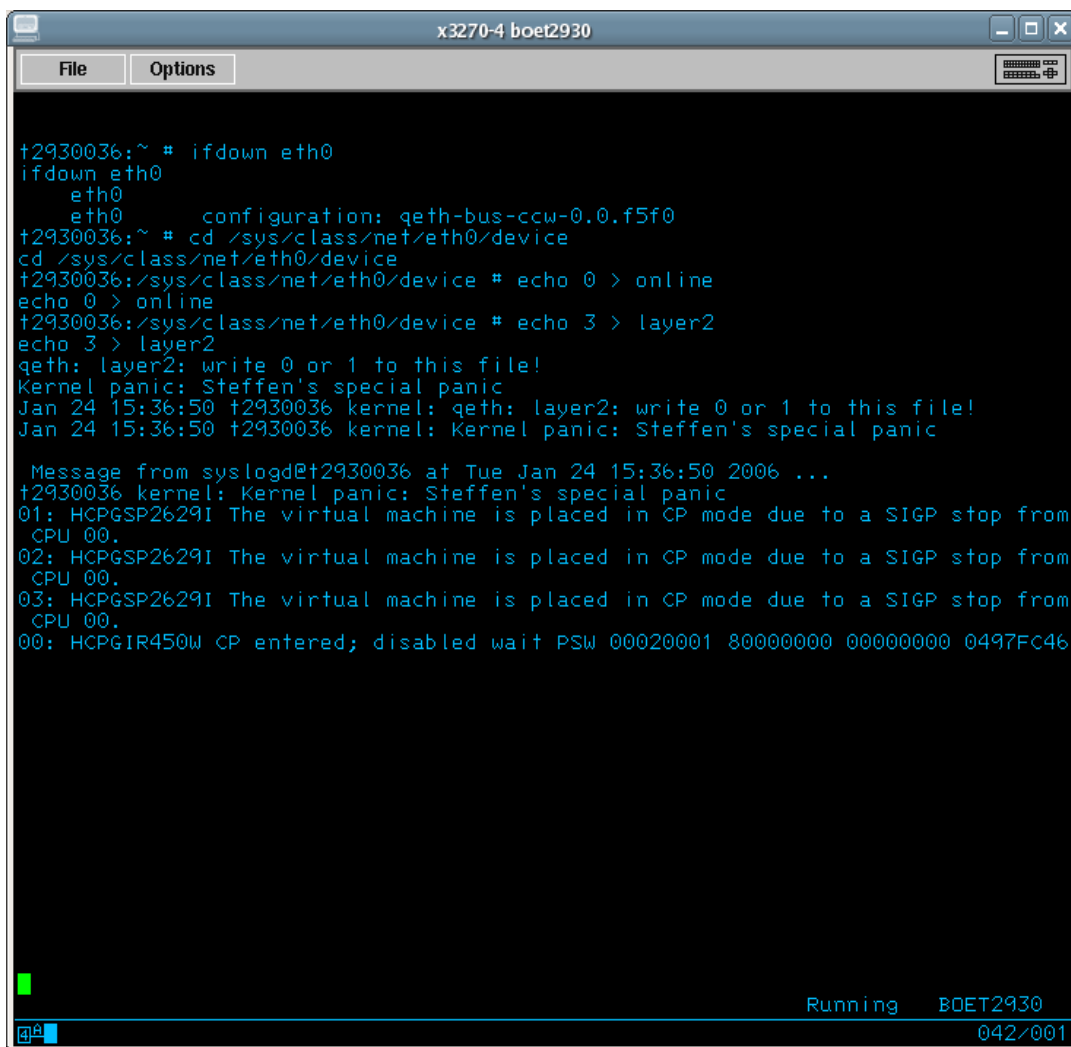
t2930036:~ # hcp 'q cplevel'
z/VM Version 5 Release 2.0, service level 0501 (64-bit)
Generated at 01/18/06 06:48:57 CET
IPL at 01/18/06 07:22:09 CET
```

## What's next ?

Customer actually reported 2 problems:

- Linux crash during enabling OSA layer 2
  - Linux crash dump necessary
- TSM backup to slow
  - Analysis of delivered setup data
  - Performance analysis necessary

# Dump



```
x3270-4 boet2930
File Options
t2930036:~ # ifdown eth0
ifdown eth0
  eth0
  eth0      configuration: qeth-bus-ccw-0.0.f5f0
t2930036:~ # cd /sys/class/net/eth0/device
cd /sys/class/net/eth0/device
t2930036:/sys/class/net/eth0/device # echo 0 > online
echo 0 > online
t2930036:/sys/class/net/eth0/device # echo 3 > layer2
echo 3 > layer2
qeth: layer2: write 0 or 1 to this file!
Kernel panic: Steffen's special panic
Jan 24 15:36:50 t2930036 kernel: qeth: layer2: write 0 or 1 to this file!
Jan 24 15:36:50 t2930036 kernel: Kernel panic: Steffen's special panic

Message from syslogd@t2930036 at Tue Jan 24 15:36:50 2006 ...
t2930036 kernel: Kernel panic: Steffen's special panic
01: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP stop from
CPU 00.
02: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP stop from
CPU 00.
03: HCPGSP2629I The virtual machine is placed in CP mode due to a SIGP stop from
CPU 00.
00: HCPG1R450W CP entered; disabled wait PSW 00020001 80000000 00000000 0497FC46

Running  BOET2930
042/001
```

## Necessary packages:

- s390-tools
- Lkcd (SUSE)
- Crash (RedHat)
  - Starting with RHEL 4 U3

## Dump with z/VM vmdump command

- First method to come up with a dump for a zVM Linux
- Create dump  
`#cp vmdump`
- System stops only for dump process !
- IPL CMS:  
`#cp i cms`
- Receive dump from reader into CMS dump file:  
`dumpload`

## Dump with z/VM ... (cont.)

- Upload dump from CMS disk to Linux guest
  - ◆ Put to Linux (z/VM ftp client: bin, locsite fix 80)
  - ◆ Get from z/VM (Linux ftp client: bin, quote site fix 80)
- Convert dump created with `vmdump` to Linux dump format with Linux command `vmconvert`

```
t2930036:~/dump # vmconvert -f dump_vmdump_format -o
dump_linux_format

vmdump information:

architecture: 64 bit

date.....: Fri Jan 13 16:24:37 2006

storage.....: 1024 MB

cpus.....: 2

1024 of 1024 |
#####| 100%

'dump_linux_format' has been written successfully.
```

## Dump on dump device

- Second method to come up with a dump for a zVM Linux
- Prepare dump device in Linux, if possible on 64Bit environment:  
`zipl -d /dev/<dasd>`
- **After Linux crash issue these commands on 3270 console:**  
`#cp cpu all stop`  
`#cp store status`  
`#cp i <dasd_cuu>`
- Wait until dump is saved on device: disabled wait appears
- Attach dump device to a Linux guest with dump tools installed
- Store dump to Linux filesystem from dump device:  
`zgetdump /dev/<dasd> > dump_file`

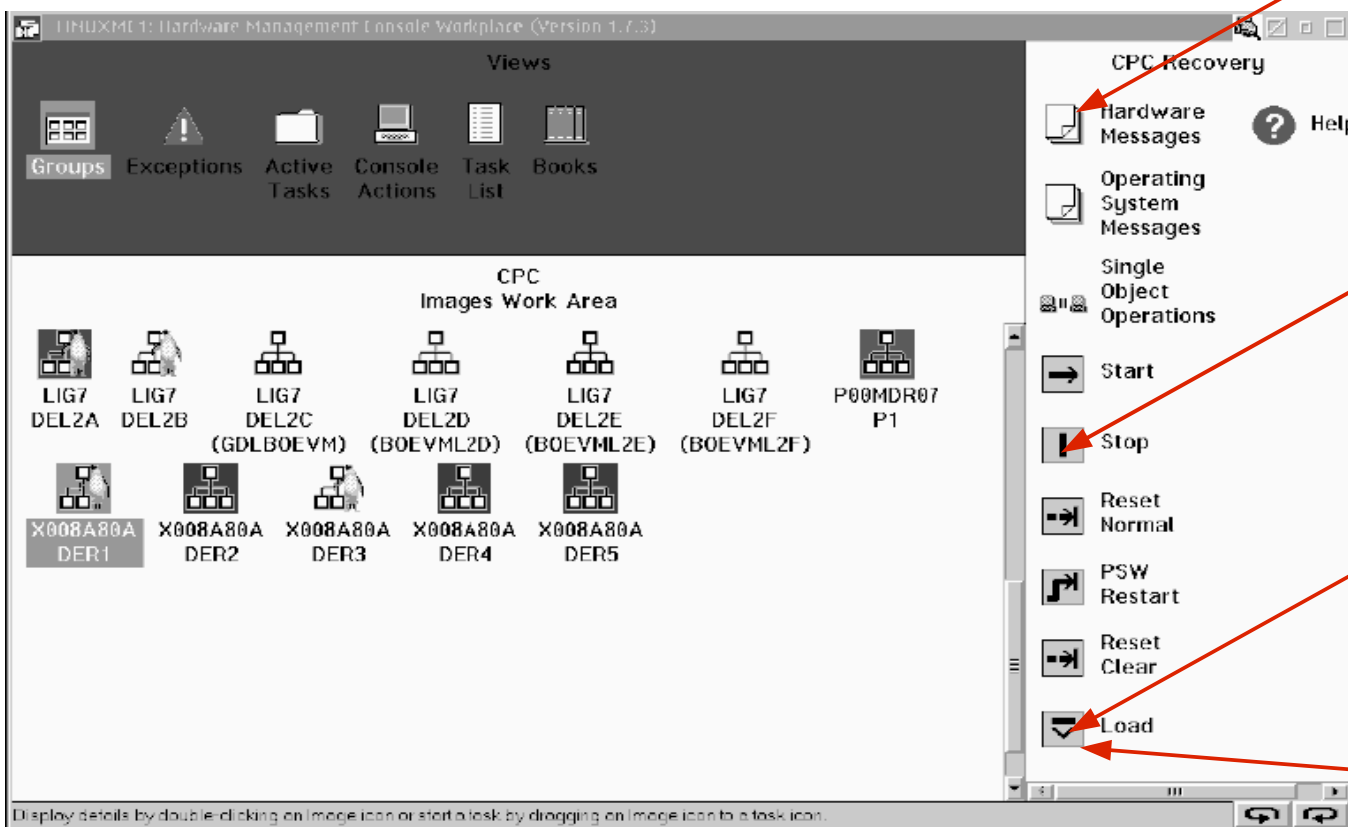


# Dump with LPAR

- Prepare dump device in Linux, if possible on 64Bit environment:

```
zipl -d /dev/<dasd>
```

3.) Check for disabled wait



1.) Stop CPU(s) (after LPAR selection !)

2.) IPL from dump Device (load type "normal", store status selected)

4.) IPL Linux again

## Store Dump in Linux

- Attach dump device to Linux guest
- Store dump to Linux fs from dump device:

```
t2930035:~ # zgetdump /dev/dasdb1 > dump_file
Dump device: /dev/dasdb1
>>> Dump header information <<<
Dump created on: Wed Feb 22 14:43:44 2006
Magic number: 0xa8190173618f23fd
Version number: 2
Header size: 4096
Page size: 4096
Physical memory: 134217728
Number of pages: 32768
cpu id: 0xff00012320948000
System Arch: s390x (ESAME)
Build Arch: s390x (ESAME)
>>> End of Dump header <<<
Reading dump content .....
Dump End Marker found: this dump is valid.
```



## Load dump with lcrash or crash

- Check if dump is loadable with **lcrash** (SUSE) or **crash** (RHEL4 U3):

```
lcrash /boot/System.map-2.6.5-7.244-s390x
dump_linux_format
      /boot/Kerntypes-2.6.5-7.244-s390x
```

```
crash -S /boot/System.map-2.6.5-7.244-s390x
      /boot/vmlinux
      dump_linux_format
```

- Vmlinux: Kernel compiled with debug information.

## Send dump to Linux support

- Send the following files to IBM support team:
  - ◆ Dump in Linux format
  - ◆ System.map
    - ★ Contains kernel addresses of all functions and data structures used in the kernel
  - ◆ Kerntypes (only for SUSE)
    - ★ Contain information about all variable and their type used in the kernel
  - ◆ Vmlinux (only for RedHat)
    - ★ Kernel with debug information (only RedHat)
- Keep dump in case of transfer errors to IBM service !

## More information about Dump tools

- lcrash  
<http://lkcd.sourceforge.net/>
- crash  
[http://people.redhat.com/anderson/crash\\_whitepaper/](http://people.redhat.com/anderson/crash_whitepaper/)
- Lcrash has build in help feature:

```
help s390dbf
```

```
COMMAND: s390dbf [-w outfile] [-v] [debug_log] [debug_log view]
```

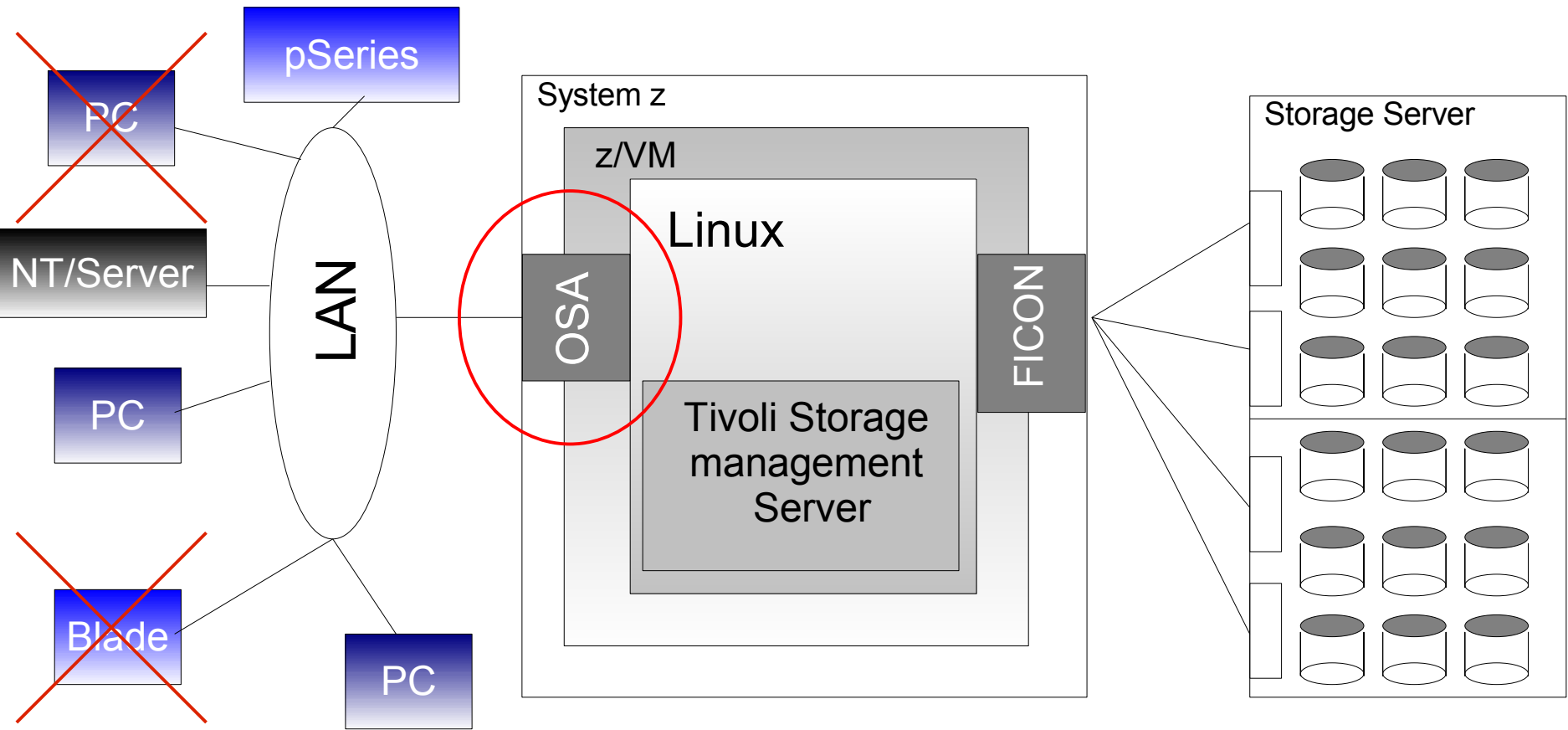
```
    Display Debug logs:
```

```
    + If called without parameters, all active debug logs are listed.
```

```
    ...
```



# TSM backup too slow



## Check Kernel messages

- Check `/var/log/messages` or `runtimeout` in `dbginfo.sh` output

```
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)
:
```

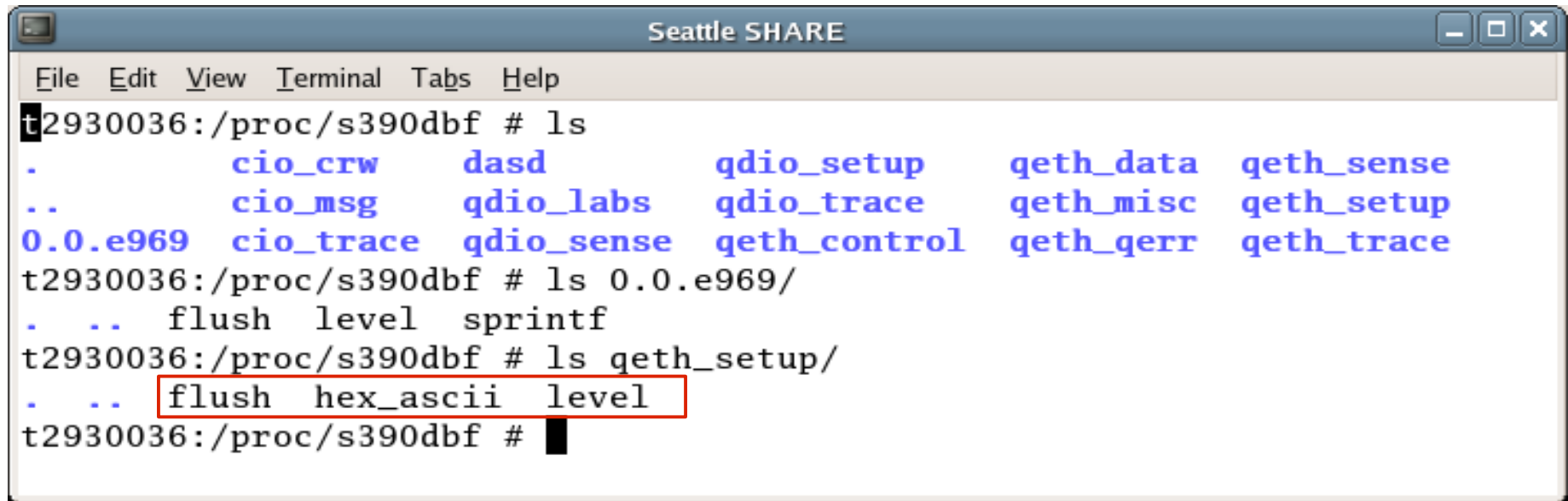
Qeth device driver does not get a 4k kernel memory page for incoming data packages (`gfp=0x20 = GFP_ATOMIC`)

- ➔ Packets dropped
- ➔ Network hangs
- ➔ Check network device driver debug messages



## Device driver debug messages

- Located in `/proc/s390dbf` or `/sys/kernel/debug/s390dbf`, can be collected with `dbginfo.sh`
- Contains detailed info from device drivers (`sprintf` or `hex_ascii`) depending on the trace level (`level`)
- Can be reset with `flush`



```
Seattle SHARE
File Edit View Terminal Tabs Help
t2930036:/proc/s390dbf # ls
.      cio_crw   dasd      qdio_setup  qeth_data  qeth_sense
..     cio_msg   qdio_labs qdio_trace  qeth_misc  qeth_setup
0.0.e969 cio_trace qdio_sense qeth_control qeth_qerr  qeth_trace
t2930036:/proc/s390dbf # ls 0.0.e969/
.  .. flush level sprintf
t2930036:/proc/s390dbf # ls qeth_setup/
.  .. flush hex_ascii level
t2930036:/proc/s390dbf #
```

## Device driver debug messages

- Increase trace level of s390dbf (e.g. qeth,qeth device driver)

- Before problem appears

```
for i in /proc/s390dbf/q*/level;
do
    echo 6 > $i;
done
```

- Capture data from s390dbf (e.g qeth,qdio device driver)

- Shortly after problem appears because s390dbf is a ringbuffer

```
for i in /proc/s390dbf/q*/*;
do
    echo $i >>outfile;
    cat $i>>outfile;
done
```

- Send outfile to IBM

# S390dbf output

```

Seattle SHARE
File Edit View Terminal Tabs Help
t2930036:/proc/s390dbf/qeth_setup # cat level
6
t2930036:/proc/s390dbf/qeth_setup # head -n20 hex_ascii
00 01138787589:133777 2 - 00 000000002097b024 70 72 6f 62 65 64 65 76 | probedev
00 01138787589:133778 2 - 00 000000002097b082 30 2e 30 2e 66 35 66 30 | 0.0.f5f0
00 01138787589:133778 2 - 00 000000002097b054 61 6c 6c 6f 63 63 72 64 | allocrd
00 01138787589:133779 2 - 00 000000002097b098 00 00 00 00 1d a8 d0 00 | .....
00 01138787589:133780 2 - 00 0000000020969562 73 65 74 75 70 63 68 00 | setupch.
00 01138787589:133786 2 - 00 0000000020969562 73 65 74 75 70 63 68 00 | setupch.
00 01138787589:133794 2 - 00 000000002097b0c8 64 65 74 63 64 74 79 70 | detcdtyp
00 01138787589:133794 2 - 00 000000002097b504 63 68 6b 5f 31 39 32 30 | chk_1920
00 01138787589:133796 2 - 00 000000002097b4d4 72 63 3a 30 00 00 00 00 | rc:0
00 01138787589:133796 2 - 00 000000002097b3ae 73 65 74 75 70 63 72 64 | setupcrd
00 01138787589:133796 2 - 00 000000002097b37e 00 00 00 00 1d a8 d0 00 | .....
00 01138787589:196013 2 - 03 000000002097962a 73 65 74 6f 6e 6c 69 6e | setonlin
00 01138787589:196013 2 - 03 00000000209795fa 00 00 00 00 1d a8 d0 00 | .....
00 01138787589:196114 2 - 03 000000002097173a 68 72 64 73 65 74 75 70 | hrdsetup
00 01138787589:196494 2 - 03 00000000209720aa 67 65 74 75 6e 69 74 00 | getunit.
00 01138787589:196642 2 - 03 000000002096c674 69 64 78 61 63 74 63 68 | idxactch
00 01138787589:197535 2 - 00 0000000020966808 69 64 78 72 64 63 62 00 | idxrdcb.
00 01138787589:197567 2 - 03 000000002096c7dc 69 64 78 61 6e 73 77 72 | idxanswr
00 01138787589:198245 2 - 00 0000000020966808 69 64 78 72 64 63 62 00 | idxrdcb.
00 01138787589:198272 2 - 03 000000002096c674 69 64 78 61 63 74 63 68 | idxactch
t2930036:/proc/s390dbf/qeth_setup # █

```

## S390dbf from customer

```
==> /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);
```

## Next steps

- Kernel messages show 0-order allocation problems
- Network device driver debug data show error conditions about insufficient memory for incoming packets, therefore network device driver cannot store received packets
- ➔ Find out what causes the memory problem
- ➔ Get Linux system data

## Linux performance data - sysstat

- Capture Linux performance data with sysstat package
- Consists of **S**ystem **A**ctivity **D**ata **C**ollector (sadc) and **S**ystem **A**ctivity **R**eport (sar) command
- Sadc example (for more see [man sadc](#))  

```
/usr/lib/sa/sadc <interval> <count> <binary outfile>  
/usr/lib/sa/sadc 5 10 sadc_outfile
```
- Sar example (for more see [man sar](#))  

```
sar -A -f sadc_outfile      --> Analyse files from sadc  
sar -B <interval> <count>  --> realtime Paging rate
```
- Please send the binary `sadc` and `sar -A` output to IBM service team

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  1
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   2
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#

```

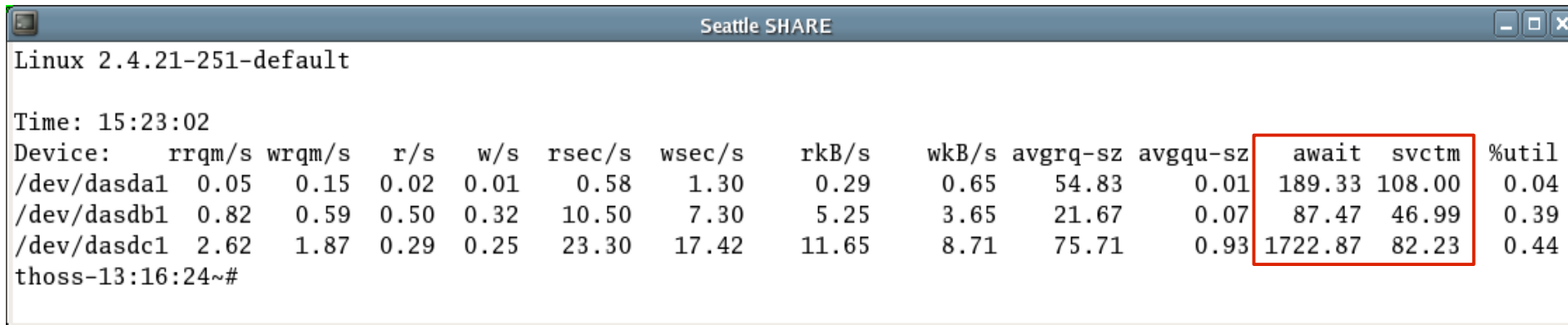
1. Memory paging:      pgpout means pages dropped  
                           pgpin  means pages load into memory

2. Block I/O:      tps      - transfers per second  
                       rtps      - read tps  
                       wtps      - write tps  
                       bread     - block read per second (x 4KByte)  
                       bwrtn    - block writes per second (x 4KByte)

## Linux performance data – iostat

- Input/Output statistics for devices and partitions.

```
iostat -dkx
```



```
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~#
```

- Await: average time (ms) for I/O req. in queue + time to service them
- Svct: average time (ms) for I/O req. issued to device (time spend outside of Linux)





## Next steps

- Sysstat data shows high memory and Disk activity
- iostat shows extreme high service time with up to 100 ms for servicing an I/O request. Good values would be between 8-15ms
- ➔ Analyze z/VM performance data
- ➔ See if z/VM reports the same bad service times
- ➔ First Problem Analysis:
  - ➔ System run out of memory because incoming data cannot written fast enough to Disk attachment, Reason maybe long service times
  - ➔ Network device driver does not get memory to store incoming packets
  - ➔ Network hangs
  - ➔ TSM backup slow down

## z/VM Monitor data – capture

<http://www.vm.ibm.com/perf/tips/collect.html>

- 5 Steps to get raw monitor data
  - ◆ Create a monitor DCSS (Discontinuous saved segment)
  - ◆ Setup userid to issue monwrite command
  - ◆ Starts and configure monitor
  - ◆ Start monwrite
  - ◆ Start your test
  - ◆ Stop monwrite and save data
- Transfer Monitor Data in binary mode with blocksize of 4096
- Collect HIST LOG data: `fcontrol moncoll perflog on`
- Histlog data are stored on your A disk
- Transfer Histlog Data in binary mode with blocksize of 1468

Note: Linux offers monitor stream support for z/VM

==> capture Linux data for z/VM Monitor

## z/VM performance data – read

- Load z/VM raw Monitor data with Performance toolkit
  - ◆ Logon to perfsvm user
  - ◆ Attach disk with raw monitor data
  - ◆ Stop real time monitor: `fcontrol moncoll dcsc off`
  - ◆ Read raw monitor data into perfkit  
`monscan disk <fn> <ft> <fm> [ from <hh:mm> ]`
- Start “Realtime” Monitor
  - ◆ Logon to perfsvm user
  - ◆ Start Performance toolkit if not done automatically: `perkit`
  - ◆ Start Monitor: `mon`
- Same interface for “Real time Monitor” and Raw Monitor data analysis

Note: Linux offers z/VM \*MONITOR record reader device driver

==> read z/VM performance data from Linux

File

Options

FCX124

Performance Screen Selection (FL520 BASE )

Monitor Scan

## General System Data

1. CPU load and trans.
2. Storage utilization
3. Reserved
4. Priv. operations
5. System counters
6. CP IUCV services
7. SPOOL file display\*
8. LPAR data
9. Shared segments
- A. Shared data spaces
- B. Virt. disks in stor.
- C. Transact. statistics
- D. Monitor data
- E. Monitor settings
- F. System settings
- G. System configuration
- H. VM Resource Manager
- I. Exceptions
- K. User defined data\*

## I/O Data

11. Channel load
12. Control units
13. I/O device load\*
14. CP owned disks\*
15. Cache extend. func.\*
16. DASD I/O assist
17. DASD seek distance\*
18. I/O prior. queueing\*
19. I/O configuration
- 1A. I/O config. changes

## User Data

21. User resource usage\*
22. User paging load\*
23. User wait states\*
24. User response time\*
25. Resources/transact.\*
26. User communication\*
27. Multitasking users\*
28. User configuration\*
29. Linux systems\*

## History Data (by Time)

31. Graphics selection
32. History data files\*
33. Benchmark displays\*
34. Correlation coeff.
35. System summary\*
36. Auxiliary storage
37. CP communications\*
38. DASD load
39. Minidisk cache\*
- 3A. Storage mgmt. data\*
- 3B. Proc. load & config\*
- 3C. Logical part. load
- 3D. Response time (all)\*
- 3E. RSK data menu\*
- 3F. Scheduler queues
- 3G. Scheduler data
- 3H. SFS/BFS logs menu\*
- 3I. System log
- 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 (\*).

## Performance Toolkit Book SC24-6136:

<http://www-03.ibm.com/servers/eserver/zseries/zos/bkserv/zvmpdf/zvm52.html>

Command ==&gt; 2

F1=Help F4=Top F5=Bot F7=Bkwd F8=Fwd F12=Return



File Options

FCX100

Data for 2005/12/14 Interval 21:55:53 - 21:56:53

Monitor Scan

```

CPU Load
PROC  %CPU  %CP  %EMU  %WT  %SYS  %SP  %SIC  %LOGLD  Vector Facility  Status or
P00   5     1     4     95   0     0     95     5     %VTOT %VEMU  REST  ded. User
P01   5     0     5     95   0     0     95     5     0     0     .0   .....

Total SSCH/RSCH      173/s      Page rate      .0/s      Priv. instruct.  431/s
Virtual I/O rate     24/s      XSTORE paging  2.3/s     Diagnose instr.  11/s
Total rel. SHARE     3900      Tot. abs SHARE  0%

Queue Statistics:      Q0      Q1      Q2      Q3      User Status:
VMDBKs in queue      16      0      2      0      # of logged on users      19
VMDBKs loading        0      0      0      0      # of dialed users          0
Eligible VMDBKs      0      0      0      0      # of active users          14
El. VMDBKs loading   0      0      0      0      # of in-queue users        18
Tot. WS (pages)     1350k   0     95303  0      % in-Q users in PGWAIT     0
Expansion factor     53.97   6.746  53.97  323.8  % in-Q users in IOWAIT     1
85% elapsed time     53.97   6.746  53.97  323.8  % elig. (resource wait)    0

Transactions          Q-Disp  trivial  non-trv  User Extremes:
Average users         4.4     .0       4.4     Max. CPU %      ZPWASF      3.2
Trans. per sec.       .8      .1       .6     Max. VECT %     .....
Av. time (sec)       5.494   .265    8.134   Max. IO/sec     ZPWASF      7.4
UP trans. time        .000    .283
MP trans. time        .265    8.380
System ITR (trans. per sec. tot. CPU)  15.3
Emul. ITR (trans. per sec. emul. CPU)  .0
Max. RESPG          LINUX03  257977
Max. MDCIO          .....
Max. XSTORE         ZLINP02  60227

```

Command ==&gt; nextsamp 22:30

F1=Help F4=Top F5=Bot F7=Bkwd F8=Fwd F12=Return

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

XSTORE even with 64 Bit z/VM necessary !  
 Rule: 25% of Main storage should be XSTOR  
 (e.g. 4 GB=3GB central + 1 GB XSTOR)  
<http://www.vm.ibm.com/perf/tips/storconf.html>

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
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
User limit (blocks)	81920

Be alerted when >700/s; see 2 GB problem descr.:  
<http://www.vm.ibm.com/perf/tips/2gstorag.html>

Device Descr.		Mdisk Pa-	Rate/s		Time (msec)					Req.			
Addr	Type	Label/ID	Links	ths	I/O	Avoid	Pend	Disc	Conn	Serv	Resp	CUWt	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	44P117	1	4	1.1	.0	2.6	137	4.9	144	144	.0	.00
9725													.00
9717													.00
9710													.00
9727													.00
970E													.00
970D													.00
971B													.00
971E													.00
9709													.00
970A													.00
9715													.00
9718													.00
970B													.00
9702													.00
971C													.00
9703													.00
9724													.00
9700													.00
9706													.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

**Resp:** Service time plus the time during which an I/O req. was waiting to be started

**Disc:** Average time that the device remained disc. from channel while executing I/O request  
(High values = overloaded path and/or long SEEK)

**Serv:** Sum of Function Pending (Pend), Connected (Conn) and Disconnect (Disc) time

**Good response time values are between 5-12 ms**

## DASD statistics with Linux

- Check I/O times from a Linux point of view to see if they match with z/VM data
- Linux DASD statistics – shows I/O statistics for the whole system
- **enable:** `echo set on > /proc/dasd/statistics`
- **disable:** `echo set off > /proc/dasd/statistics`
- **reset:** turn off and on again



# /proc/dasd/statistics

```

thoss-11:20:27~/temp#cat statistics
36092283 dasd I/O requests
with -1725707784 sectors(512B each)
  <u>_<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)
  <u>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
<u>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
  <u>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
<u>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 per device

- Tunedasd is part of s390-tools package and shows DASD statistics per device  
echo set on > /proc/dasd/statistics  
tunedasd -P <device>
- Output format same as `cat /proc/dasd/statistics`
- Tunedasd can also change storage server caching behavior  
get caching behavior:  
tunedasd -g /dev/dasda1  
set caching behavior:  
tunedasd -c sequential /dev/dasda1

## Final problem analysis

- z/VM Monitor data and Linux I/O statistics show very high Disc I/O response times
- Channel load utilization is around 20%
- Problem seems to be in the disk attachment !
- How does this fit with our network problem ?
  - ➔ TSM tries to write backup data from the connected clients to TSM storage pools
  - ➔ Data stored in the main Linux memory and flagged as “dirty” because it must be written to the TSM storage. Therefore Linux will not swap this data !
  - ➔ Linux runs out of memory because write I/O is too slow
  - ➔ Network driver cannot get memory to store incoming packets
  - ➔ Network hangs

## Next steps

- Speak with your Hardware people about errors in the storage attachment (Storage server, Cable, Switches ...)
- Speak with your Storage server admin about how DASD devices arranged in your Storage server:

[http://www-128.ibm.com/developerworks/linux/linux390/perf/tuning\\_rec\\_dasd\\_optimizedisk.html](http://www-128.ibm.com/developerworks/linux/linux390/perf/tuning_rec_dasd_optimizedisk.html)

- Do some I/O benchmark measurements to reproduce the problem outside the production system. Open Source benchmark IOZone would be a good tool for our problem:

<http://www.iozone.org/>

## Conclusion

1. Describe your problem as detailed as possible.
2. Save as much information as possible about your environment.
3. Store a dump in case of a Linux crash.
4. Store performance data (sysstat, Monitor data) in case of performance problems.
5. Check all system components even if they do not appear within the problem area in the beginning

## Links

- Check out Linux Development Developer Works Page for
  - ◆ **Documentation:**  
[http://www-128.ibm.com/developerworks/linux/linux390/april2004\\_documentation.html](http://www-128.ibm.com/developerworks/linux/linux390/april2004_documentation.html)
    - ★ Device Driver, Features and Commands book
    - ★ Using the Dump Tools
    - ★ ...
  - ◆ **Tuning Hints&Tips**  
<http://www-128.ibm.com/developerworks/linux/linux390/perf/index.html>
  - ◆ **Useful addons**  
[http://www-128.ibm.com/developerworks/linux/linux390/useful\\_add-ons.html](http://www-128.ibm.com/developerworks/linux/linux390/useful_add-ons.html)
  - ◆ **Latest s390-tools package if not included in the Distro:**  
<http://www-128.ibm.com/developerworks/linux/linux390/s390-tools-1.5.3.html>

# Thank you !

