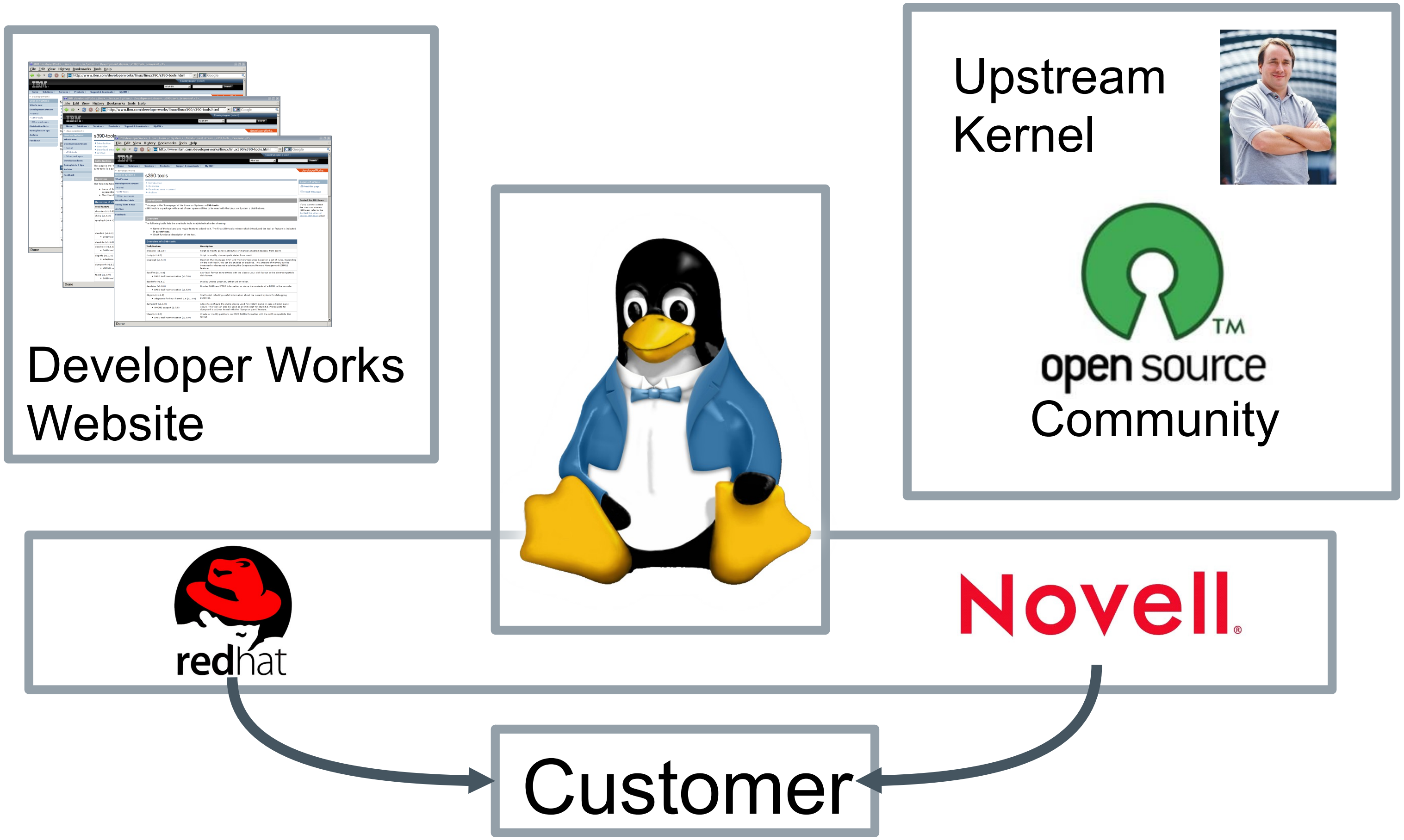
Hans-Joachim Picht, Linux on System z Initiatives & Linux Champion EMEA <hans@de.ibm.com>

Running Linux on System z as a z/VM Guest: Useful Things to Know

Thursday, August 11, 2011: 4:30 PM-5:30 PM, Dolphin, Southern Hemisphere I-II
Session 09463

IBM Linux on System z Development

IBM Linux on System z Development contributes in the following areas: Kernel, s390-tools, Open Source Tools (e.g. eclipse, ooprofile), GCC, GLIBC, Binutils



Developer Works Website

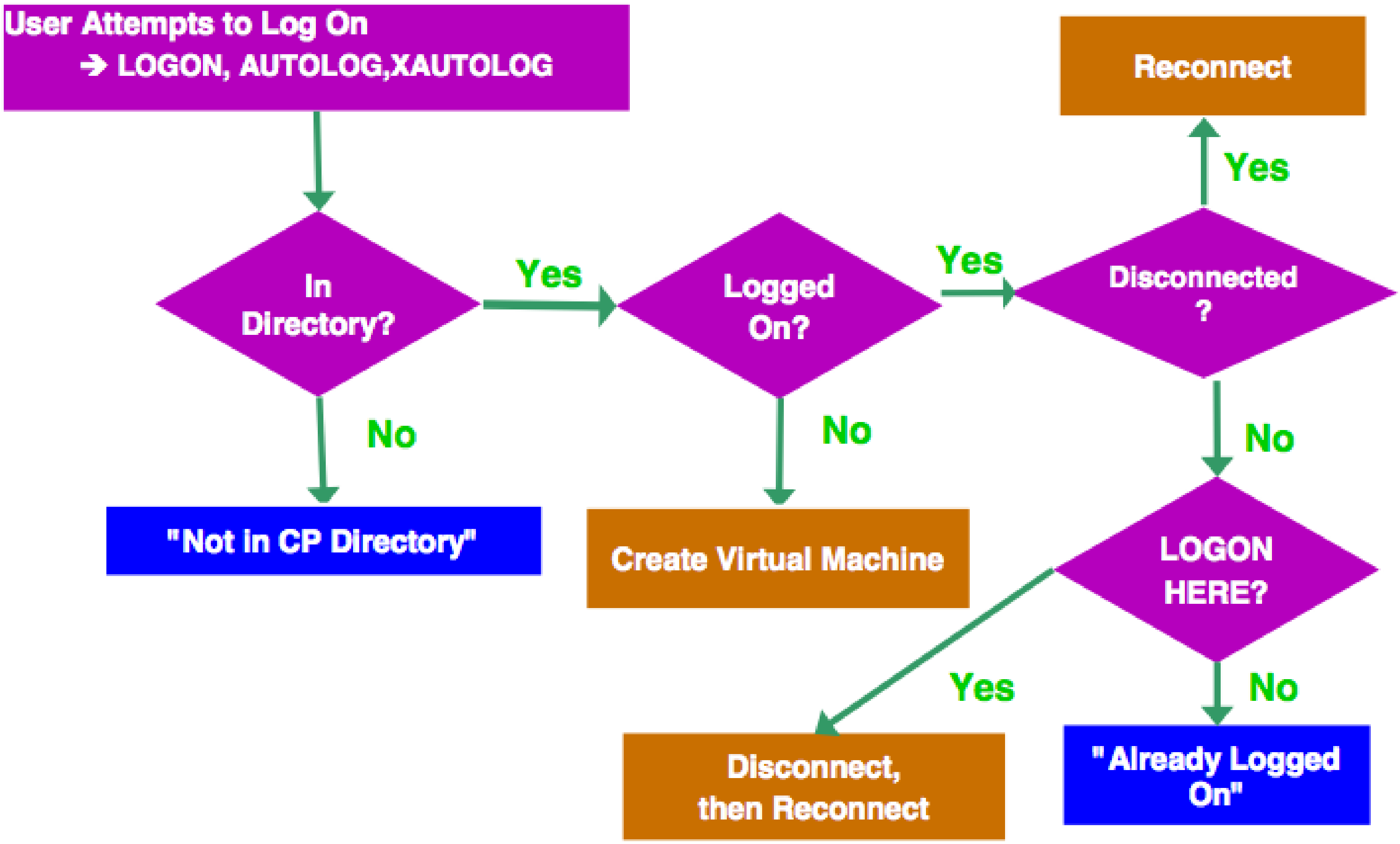
Upstream Kernel

Community

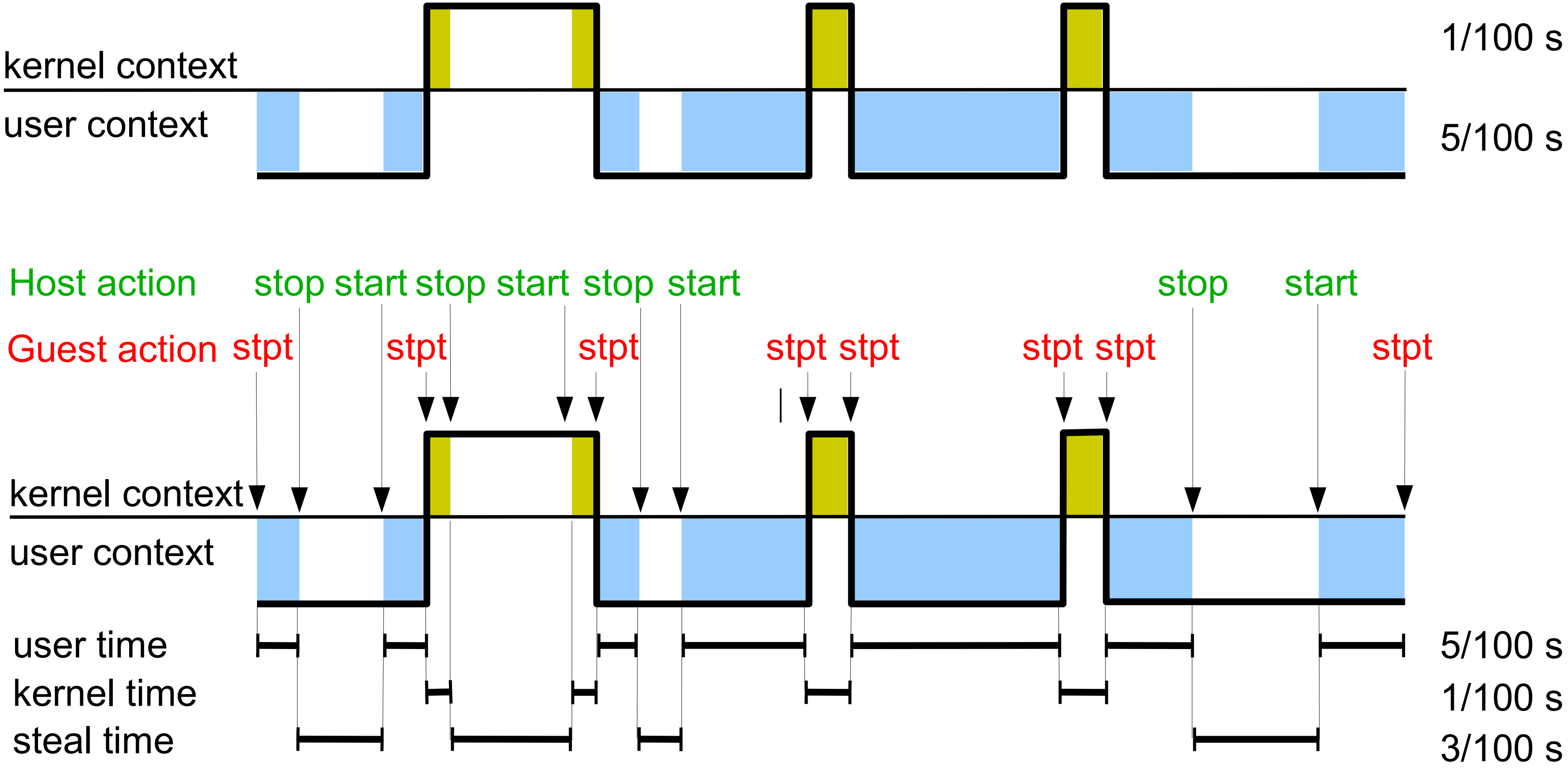
Customer

....the code you use is the result of the efforts of an anonymous army of blue penguins involved in developing, testing, documenting, ...

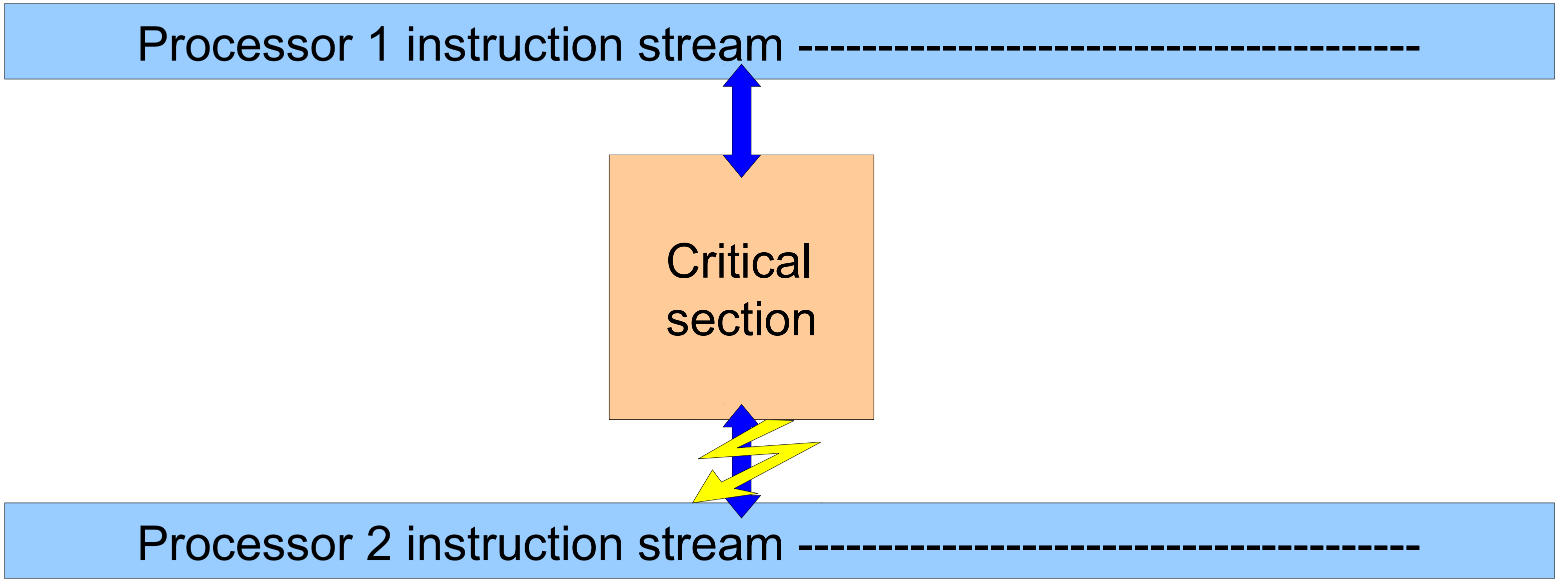
Logging on to z/VM (creating a virtual machine)



Timer based CPU accounting & virtual CPUs



z/VM – Linux Locking



Processor 1 instruction stream ------------------------------------

Critical section

Processor 2 instruction stream ------------------------------------

- Linux kernel uses spin locks for exclusive use of kernel resources
- Traditional implementation uses busy waiting (“spinning”) if lock cannot b acquired
  - Bad idea with virtual CPU's

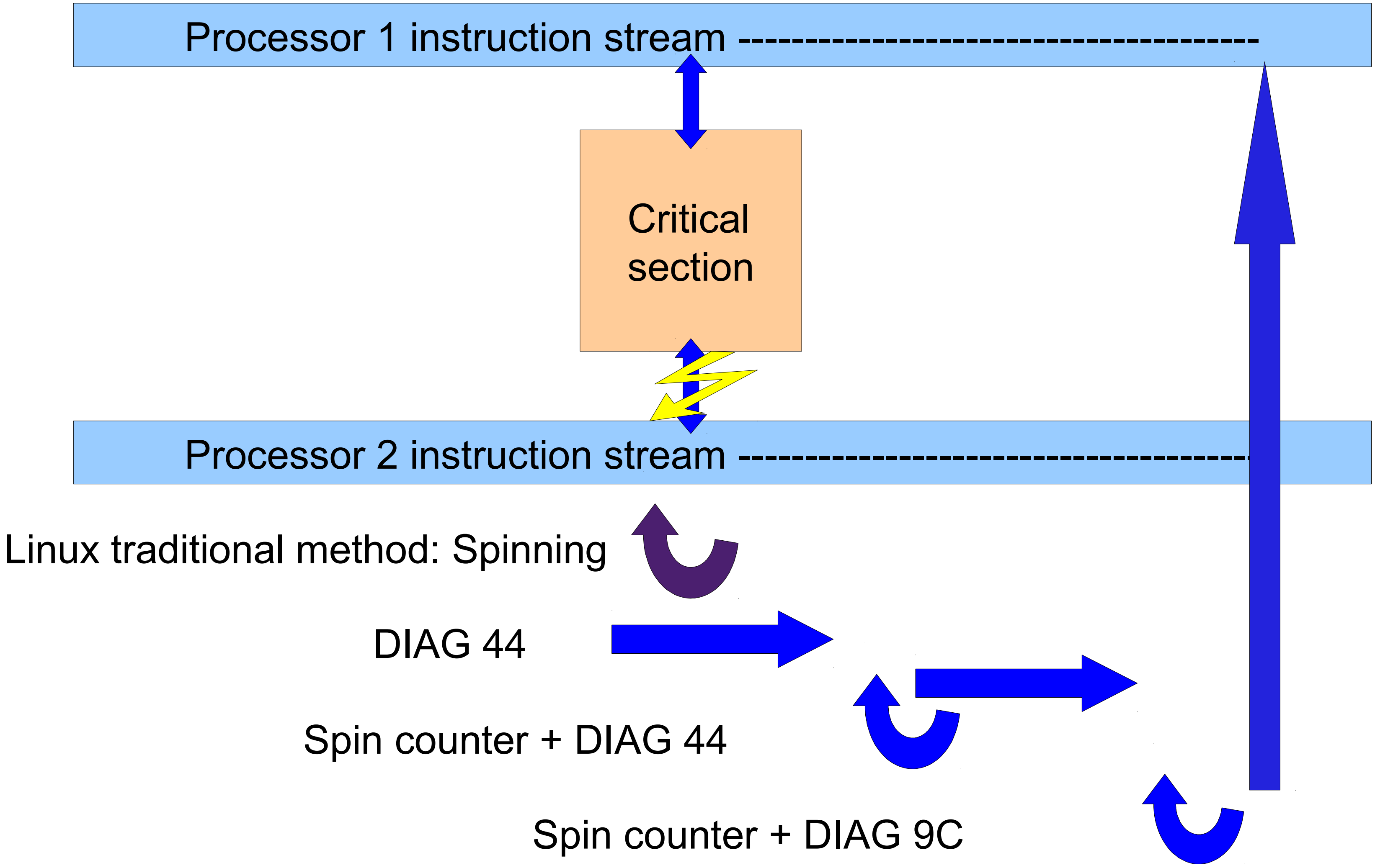
z/VM – Linux Locking Improvements

- DIAG 44: to the LPAR hypervisor or to z/VM to give the processor back instead of looping on the lock, to allow other more useful work to be done
  spin_retry counter in Linux to avoid excessive use of diagnose instructions

```
[root@h4245005 ~]# cat /proc/sys/kernel/spin_retry
1000
```

- DIAG 9C: remember CPU that has the lock and tell z/VM who should get the processor
  - Available since RHEL5 U1 and SLES10 SP1

z/VM – Linux Locking with diag



Processor 1 instruction stream ---------------------------------------

Critical section

Processor 2 instruction stream ---------------------------------------

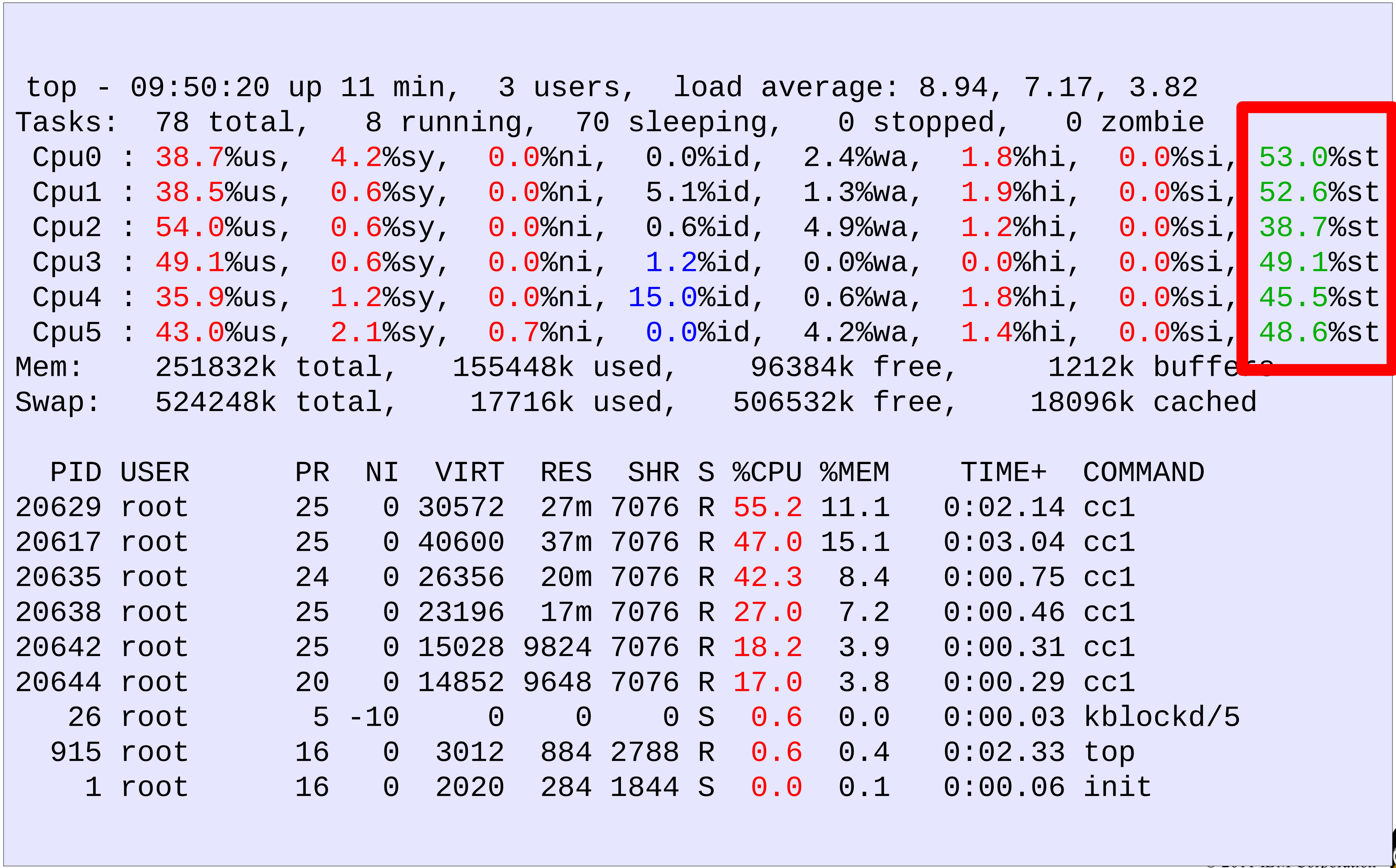
Linux traditional method: Spinning

DIAG 44

Spin counter + DIAG 44

Spin counter + DIAG 9C

top



```
 top - 09:50:20 up 11 min,  3 users,  load average: 8.94, 7.17, 3.82
Tasks:  78 total,   8 running,  70 sleeping,   0 stopped,   0 zombie
 Cpu0 : 38.7%us,  4.2%sy,  0.0%ni,  0.0%id,  2.4%wa,  1.8%hi,  0.0%si, 53.0%st
 Cpu1 : 38.5%us,  0.6%sy,  0.0%ni,  5.1%id,  1.3%wa,  1.9%hi,  0.0%si, 52.6%st
 Cpu2 : 54.0%us,  0.6%sy,  0.0%ni,  0.6%id,  4.9%wa,  1.2%hi,  0.0%si, 38.7%st
 Cpu3 : 49.1%us,  0.6%sy,  0.0%ni,  1.2%id,  0.0%wa,  0.0%hi,  0.0%si, 49.1%st
 Cpu4 : 35.9%us,  1.2%sy,  0.0%ni, 15.0%id,  0.6%wa,  1.8%hi,  0.0%si, 45.5%st
 Cpu5 : 43.0%us,  2.1%sy,  0.7%ni,  0.0%id,  4.2%wa,  1.4%hi,  0.0%si, 48.6%st
Mem:    251832k total,   155448k used,    96384k free,     1212k buffers
Swap:   524248k total,    17716k used,   506532k free,    18096k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
20629 root      25   0 30572  27m 7076 R 55.2 11.1   0:02.14 cc1
20617 root      25   0 40600  37m 7076 R 47.0 15.1   0:03.04 cc1
20635 root      24   0 26356  20m 7076 R 42.3  8.4   0:00.75 cc1
20638 root      25   0 23196  17m 7076 R 27.0  7.2   0:00.46 cc1
20642 root      25   0 15028 9824 7076 R 18.2  3.9   0:00.31 cc1
20644 root      20   0 14852 9648 7076 R 17.0  3.8   0:00.29 cc1
   26 root       5 -10     0    0    0 S  0.6  0.0   0:00.03 kblockd/5
  915 root      16   0  3012  884 2788 R  0.6  0.4   0:02.33 top
    1 root      16   0  2020  284 1844 S  0.0  0.1   0:00.06 init
```

z/VM – Linux Idle Guests

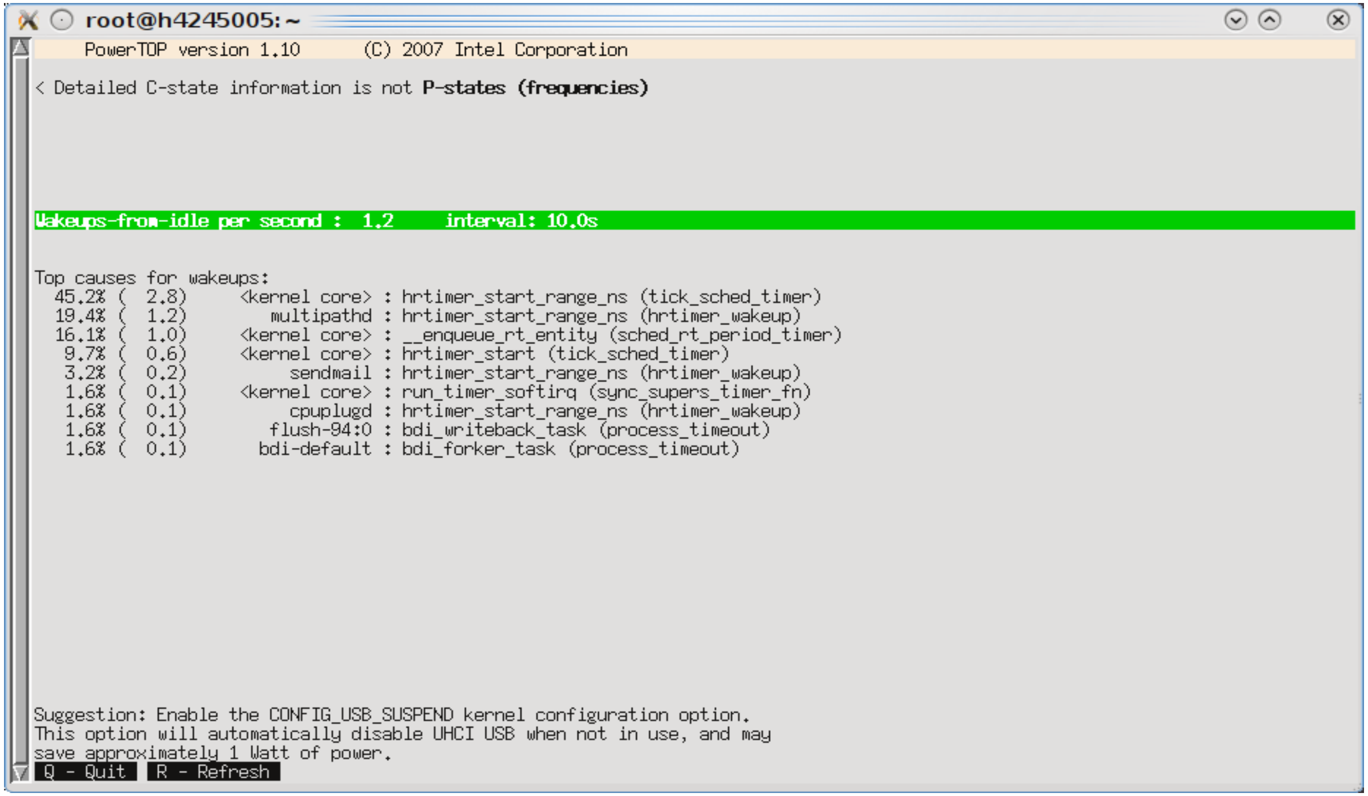
- Many applications use timeouts to check for work
- Modern distributions come with many background processes (daemons) that wake up regularly
- Older kernels had a regular timer interrupt of up to 1000 times per second
  - Unnecessary CPU load in z/VM that also influences z/VM scheduler decisions

- Linux now uses Dynamic ticks for event-based wakeup

```
[root@h4245005 ~]# cat /proc/interrupts
           CPU0       CPU1       CPU2       CPU3       CPU4       CPU5

EXT:     269258      27381     100415      46332      19236     181209
I/O:       7990       7330       6246       7852       7067       6011
```

z/VM Linux Idle Guests



```
     PowerTOP version 1.10      (C) 2007 Intel Corporation

< Detailed C-state information is not P-states (frequencies)



Wakeups-from-idle per second :  1.2     interval: 10.0s


Top causes for wakeups:
  45.2% (  2.8)     <kernel core> : hrtimer_start_range_ns (tick_sched_timer)
  19.4% (  1.2)        multipathd : hrtimer_start_range_ns (hrtimer_wakeup)
  16.1% (  1.0)     <kernel core> : __enqueue_rt_entity (sched_rt_period_timer)
   9.7% (  0.6)     <kernel core> : hrtimer_start (tick_sched_timer)
   3.2% (  0.2)          sendmail : hrtimer_start_range_ns (hrtimer_wakeup)
   1.6% (  0.1)     <kernel core> : run_timer_softirq (sync_supers_timer_fn)
   1.6% (  0.1)          cpuplugd : hrtimer_start_range_ns (hrtimer_wakeup)
   1.6% (  0.1)        flush-94:0 : bdi_writeback_task (process_timeout)
   1.6% (  0.1)       bdi-default : bdi_forker_task (process_timeout)



Suggestion: Enable the CONFIG_USB_SUSPEND kernel configuration option.
This option will automatically disable UHCI USB when not in use, and may
save approximately 1 Watt of power.
 Q - Quit   R - Refresh
```

Recommendation:

Turn off any unneeded services and daemons

Watch out for applications with bad wakeup behaviour

How can you read & write files on a CMS disk with Linux?

About the CMS user space file system (fuse) support

6.1

- Allows to mount a z/VM minidisk to a Linux mount point
- z/VM minidisk needs to be in the enhanced disk format (EDF)
- The cmsfs fuse file system transparently integrates the files on the minidisk into the Linux VFS, no special command required

```
root@larsson:~> cmsfs-fuse /dev/dasde /mnt/cms
root@larsson:~> ls -la /mnt/cms/PROFILE.EXEC
-r--r----- 1 root root 3360 Jun 26 2009
/mnt/fuse/PROFILE.EXEC
```

- By default no conversion is performed
  - Mount with '-t' to get automatic EBCDIC to ASCII conversion

```
root@larsson:~> cmsfs-fuse -t /dev/dasde /mnt/cms
```

- Write support is also available
  - use “vi” to edit PROFILE.EXEC anyone ?
- Use fusermount to unmount the file system again

```
root@larsson:~> fusermount -u /mnt/cms
```

vmcp

Using the z/VM CP interface device driver (vmcp), you can send control program (CP) commands to the VM hypervisor and display VM’s response.

```
root@larsson:~> modprobe vmcp
root@larsson:~> vmcp q v cpus
CPU 02  ID  FF20012320978000 CP           CPUAFF ON
CPU 00  ID  FF00012320978000 (BASE) CP    CPUAFF ON
CPU 01  ID  FF10012320978000 CP           CPUAFF ON
root@larsson:~> vmcp q priv
Privilege classes for user HANS
        Currently: GU
        Directory: GU
The privilege classes are not locked against changes.
root@larsson:~> vmcp def store 32G
HCPDST094E Storage size (32G) exceeds directory maximum (5G)
Error: non-zero CP response for command 'DEF STORE 32G': #94
```

Be careful, when executing disruptive commands!

lsmem - Show online status information about memory blocks

11.1

The lsmem command lists the ranges of available memory with their online status.

6.1

- The listed memory blocks correspond to the memory block representation in sysfs.
- The command also shows the memory block size, the device size, and the amount of memory in online and offline state.

The output of this command, shows ranges of adjacent memory blocks with similar attributes.

```
root@larsson:~> lsmem
Address range                        Size (MB) State Removable Device
=========================================================================
0x0000000000000000-0x000000000fffffff 256       online    no    0
0x0000000010000000-0x000000002fffffff 512       online    yes   1-2
0x0000000030000000-0x000000003fffffff 256       online    no    3
0x0000000040000000-0x000000006fffffff 768       online    yes   4-6
0x0000000070000000-0x00000000ffffffff 2304      offline    -    7-15
Memory device size : 256 MB
Memory block size : 256 MB
Total online memory : 1792 MB
Total offline memory: 2304 MB
```

Configuring standby memory

To see how much central and expanded storage (memory) are installed and allocated to a system use the QUERY STORAGE and QUERY XSTOR commands. For example:

```
==> q stor
STORAGE = 16G CONFIGURED = 16G INC = 256M STANDBY = 0
RESERVED = 0
```

Modify the directory entry by adding a COMMAND statement. This will give the virtual machine an additional 768 MB of standby memory:

```
USER LINUX01 LNX4VM 256M 2G G
INCLUDE LNXDFLT
COMMAND DEFINE STORAGE 256M STANDBY 768M
OPTION APPLMON
MDISK 100 3390 3339 3338 UM63A9 MR LNX4VM LNX4VM LNX4VM
MDISK 101 3390 6677 3338 UM63A9 MR LNX4VM LNX4VM LNX4VM
```

chmem – Setting memory online or offline

11.1

6.1

- **The chmem command sets a particular size or range of memory online or offline**
  - Setting memory online might fail if the hypervisor does not have enough memory left
    - For example, because memory was overcommitted
  - Setting memory offline might fail if Linux cannot free the memory
  - If only part of the requested memory could be set online or offline, a message tells you how much memory was set online or offline instead of the requested amount

- **To request 1024 MB of memory to be set online, issue:**

```
root@larsson:~# chmem --enable 1024
```

- **To request the memory range starting with 0x00000000e4000000 and ending with 0x00000000f3ffffff to be set offline, issue:**

```
root@larsson:~# chmem --disable 0x00000000e4000000-0x00000000f3ffffff
```

- **This command requests 1024 MB of memory to be set online.**

```
root@larsson:~# chmem --disable 512
```

How can the terminal server using IUCV help you?

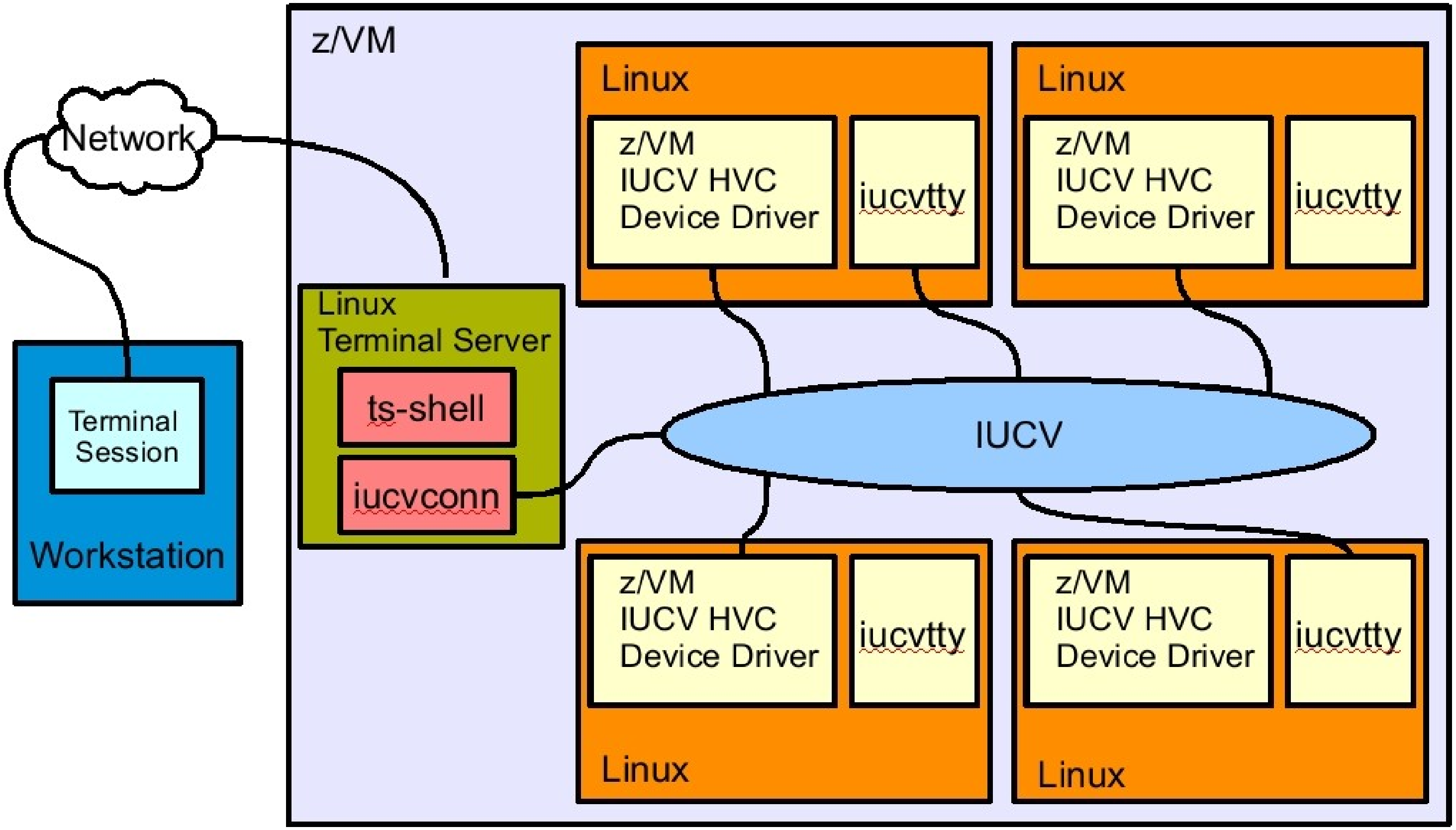
- **Full-screen terminal access to Linux instances on the same z/VM**
- **Access to Linux instances that are not connected to an Internet Protocol (IP) network**
- **Use cases**
  - Provide an alternative terminal access to 3270 and 3215 line-mode terminals
  - Increase availability by providing emergency access if the network for a system fails
  - Centralize access to systems by providing a terminal server environment
  - Heighten security by separating user networks from administrator networks or by isolating sensitive Linux instances from public IP networks

IUCV Terminals

11

5.4

- Full-screen terminal access to Linux guest operating systems on the same z/VM
- Access Linux instances with no external network because IUCV is independent from TCP/IP



IUCV terminal applications – examples

11

5.4

Using the iucvconn program: To access the first z/VM IUCV HVC terminal on the Linux instance in z/VM guest LNXSYS02

```
root@larsson:~> iucvconn LNXSYS02 lnxhvc0
```

To create a transcript of the terminal session to the Linux instance in z/VM guest LNXSYS99

```
root@larsson:~> iucvconn -s ~/transcripts/lnxsys99 LNXSYS99
lnxhvc0
```

Using the iucvtty program: To allow remote logins using the terminal identifier „lnxterm“

```
root@larsson:~> iucvtty lnxterm
```

To access the „lnxterm“ terminal on the Linux instance in z/VM guest LNXSYS01

```
root@larsson:~>  iucvconn LNXSYS01 lnxterm
```

To use /sbin/sulogin instead of /bin/login for terminal “suterm”

```
root@larsson:~>  iucvtty suterm -- /sbin/sulogin
```

How can you enable a terminal server for iucvconn?

- **Authorizing the z/VM guest virtual machine for IUCV**
  - Adding an IUCV user directory statement, for example, `IUCV ANY`
  - The z/VM user directory for a terminal server might look like:

```
USER LNXTS    XSECRETX 768M 1G G
* General statements
 IPL 0150
 MACH ESA 8
* IUCV authorization
 IUCV ANY
 OPTION MAXCONN 128
* Generic device statements
 CONSOLE 0009 3215 T
 SPOOL 000C 2540 READER *
*    ...
```

Multi Volume Dump

How to prepare a set of ECKD DASD devices for a multi-volume dump? bit systems only).

10.3

5.4

- We use two DASDs in this example:

```
root@larsson:~> dasdfmt -f /dev/dasdc -b 4096
root@larsson:~> dasdfmt -f /dev/dasdd -b 4096
```

- Create the partitions with fdasd. The sum of the partition sizes must be sufficiently large (the memory size + 10 MB):

```
root@larsson:~> fdasd /dev/dasdc
root@larsson:~> fdasd /dev/dasdd
```

- Create a file called sample_dump_conf containing the device nodes (e.g. /dev/dasda1) of the two partitions, separated by one or more line feed characters
- Prepare the volumes using the zipl command.

```
root@larsson:~>  zipl -M sample_dump_conf
[...]
```

How to obtain a dump

To obtain a dump with the multi-volume DASD dump tool, perform the following steps:

- Stop all CPUs, Store status on the IPL CPU.
- IPL the dump tool using one of the prepared volumes, either 4711 or 4712.
- After the dump tool is IPLed, you'll see a messages that indicates the progress of the dump. Then you can IPL Linux again

```
==> cp cpu all stop
==> cp cpu 0 store status
==> cp ipl 4711
```

- Copying a multi-volume dump to a file
- Use zgetdump command without any option to copy the dump parts to a file:

```
root@larsson:~>  zgetdump /dev/dasdc > mv_dump_file
```

How to obtain information about a multi volume dumps

Display information on the involved volumes:

```
root@larsson:~>  zgetdump -d /dev/dasdc
'/dev/dasdc' is part of Version 1 multi-volume dump,which is
spread along the following DASD volumes:
0.0.4711 (online, valid)
0.0.4712 (online, valid)
[...]
```

Display information about the dump itself:

```
root@larsson:~>  zgetdump -i /dev/dasdc
Dump device: /dev/dasdc
>>>  Dump header information  <<<
Dump created on: Thu Feb  25 15:12:41 2010
[...]
Multi-volume dump: Disk 1 (of 2)
Reading dump contents from
0.0.4711.................................
Dump ended on:   Thu Feb  25 15:12:52 2010
Dump End Marker found: this dump is valid.
```

Handling large dumps

Compress the dump and split it into parts of 1 GB

```
root@larsson:~>  zgetdump /dev/dasdc1 | gzip | split -b 1G
```

Several compressed files such as xaa, xab, xac, ... are created

Create md5 sums of the compressed files

```
root@larsson:~>  md5sum xa* > dump.md5
```

Upload all parts together with the md5 information.

Verification of the parts for a receiver

```
root@larsson:~>  md5sum -c dump.md5
xaa: OK
[....]
```

Merge the parts and uncompress the dump

```
root@larsson:~>  cat xa* | gunzip -c > dump
```

Transferring dumps

Transferring single volume dumps with ssh

```
root@larsson:~>  zgetdump /dev/dasdc1 | ssh user@host "cat >
dump_file_on_target_host"
```

Transferring multi-volume dumps with ssh

```
root@larsson:~>  zgetdump /dev/dasdc | ssh user@host "cat >
multi_volume_dump_file_on_target_host"
```

Transferring a dump with ftp.

Establish an ftp session with the target host, login and set the transfer mode to Binary. Send the dump to the host

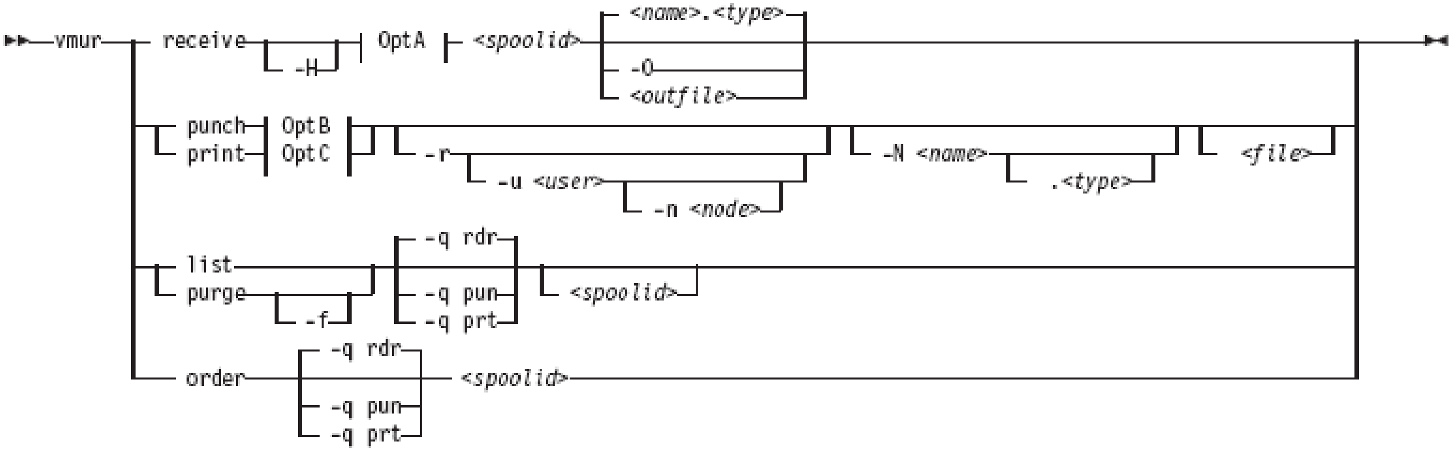
```
root@larsson:~>  ftp> put |"zgetdump /dev/dasdc1"
<dump_file_on_target_host>
```

vmur – Working with z/VM unit record devices

10.1

5.2

- **The vmur command provides functions required to work with z/VM spool file queues**
  - *Receive*: Read data from the z/VM reader file queue
  - *Punch* or *print*: Write data to the z/VM punch or printer file queue and transfer it to another user's virtual reader, optionally, on a remote z/VM node
  - *List*: Display detailed information about one or all files on the specified spool file queue
  - *Purge*: Remove one or all files on the specified spool file queue
  - *Order*: Position a file at the top of the specified spool file queue



vmur – Working with z/VM unit record devices

Logging and reading console output of Linux guest operating systems

- Begin console spooling with:

```
root@larsson:~#  vmcp sp cons start
```

- Produce output to z/VM console (for example, with CP TRACE)

- Close the console spool file and transfer it to the reader queue, find the spool ID behind the FILE keyword in the corresponding CP message:

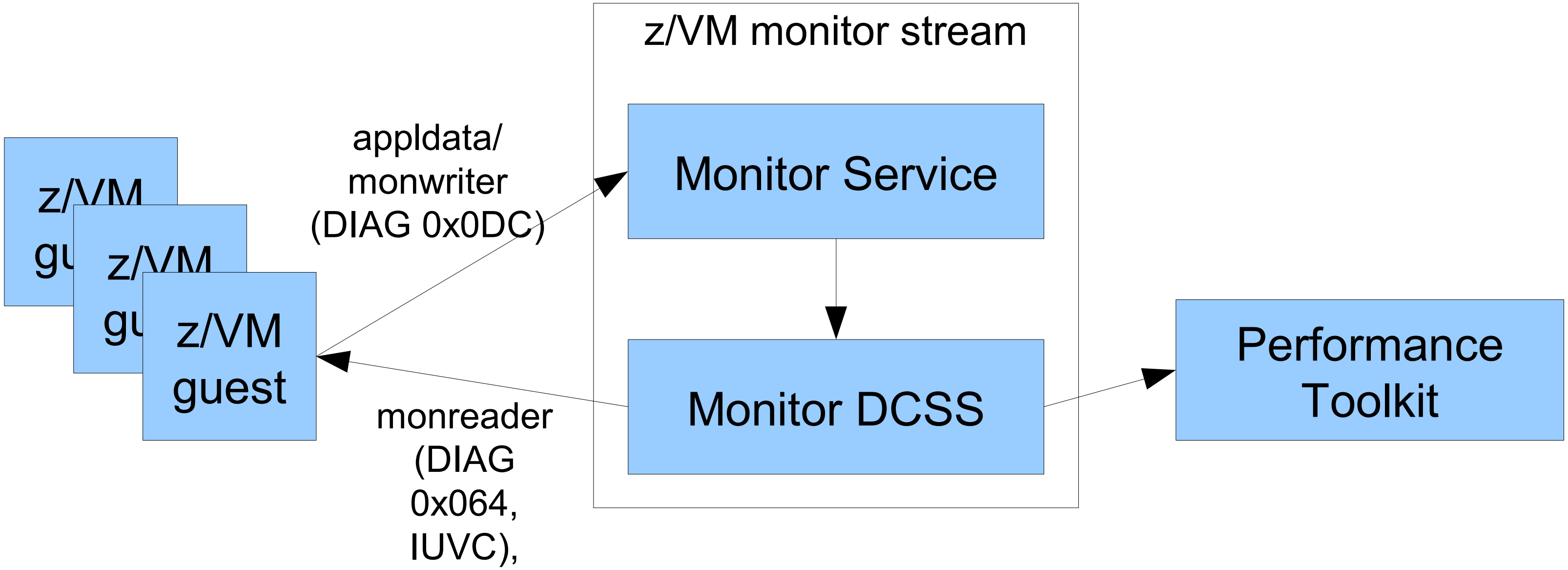
```
root@larsson:~#  vmcp sp cons clo \* rdr
RDR FILE 0398 SENT FROM T6360025 CON WAS 0398 RECS 1872 CPY
001 T NOHOLD NOKEEP
```

- Receive the console spool file and save it on the Linux file system in the current directory:

```
root@larsson:~#  chccwdev -e 000c
root@larsson:~#  vmur re -t 398 linux_cons
```

z/VM Monitor Service Infrastructure

- **Provides monitor data through the monitor stream**
  - z/VM monitor service collects data in a shared memory segment (DCSS)
  - Producer: a range of facilities, e.g. Linux through appldata / monwriter
  - Consumer: Performance Toolkit, or Linux application through monreader



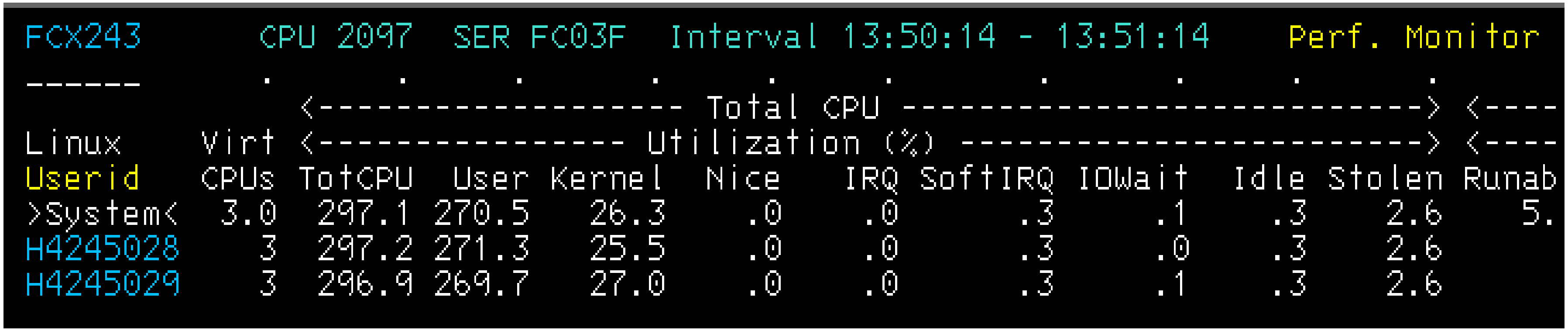
appldata - Linux monitoring modules

- Kernel modules which gather information from the Linux kernel
- appldata_os
  - CPU utilization, processes
- appldata_mem
  - memory, paging, cache
- appldata_net_sum
  - packets, bytes, errors
- Usage:

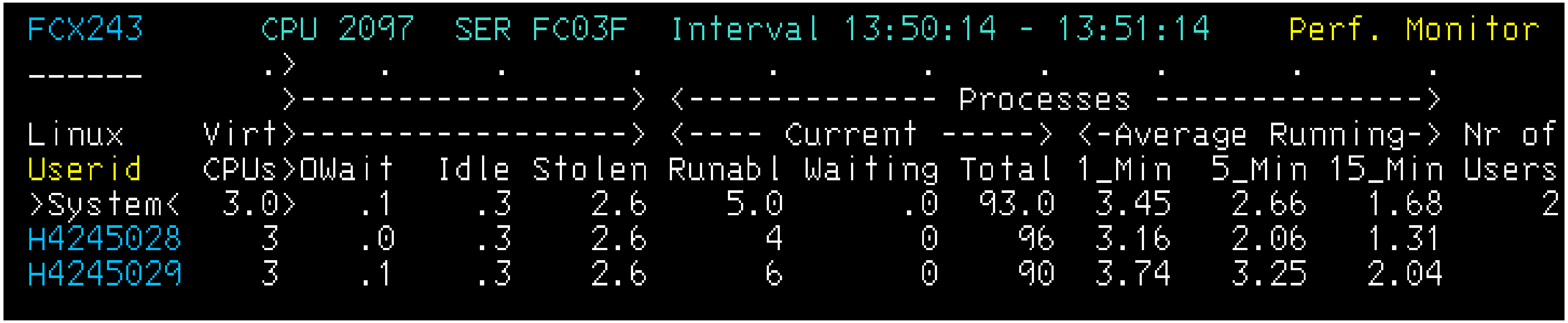
```
root@larsson:~# modprobe appldata_os
root@larsson:~# modprobe appldata_os
root@larsson:~# echo 1 >  /proc/sys/appldata/os
root@larsson:~# modprobe appldata_mem
root@larsson:~# echo 1 >  /proc/sys/appldata/mem
root@larsson:~# modprobe appldata_net_sum
root@larsson:~# echo 1 >  /proc/sys/appldata/net_sum
```

appldata – os monitor

Linux monitoring data collected by appldata_os as processed and displayed by z/VM Performance Toolkit:



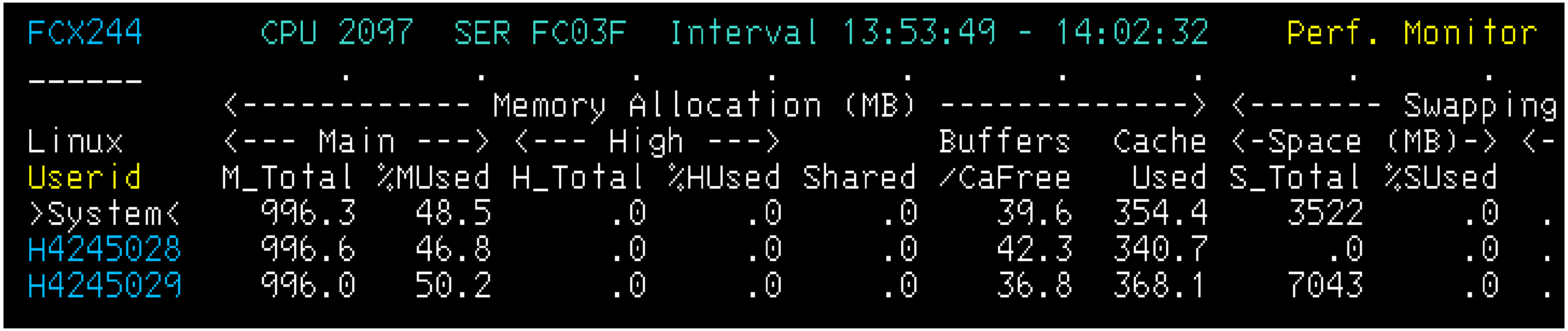
```
FCX243      CPU 2097   SER FC03F   Interval 13:50:14 - 13:51:14    Perf. Monitor
------          .      .     .      .     .     .      .      .     .      .
                 <------------------- Total CPU ---------------------------> <----
Linux     Virt  <--------------- Utilization (%) -------------------------> <----
Userid    CPUs  TotCPU  User Kernel  Nice    IRQ SoftIRQ IOWait  Idle Stolen Runab
>System<   3.0   297.1 270.5   26.3    .0     .0      .3     .1    .3    2.6    5.
H4245028     3   297.2 271.3   25.5    .0     .0      .3     .0    .3    2.6
H4245029     3   296.9 269.7   27.0    .0     .0      .3     .1    .3    2.6
```



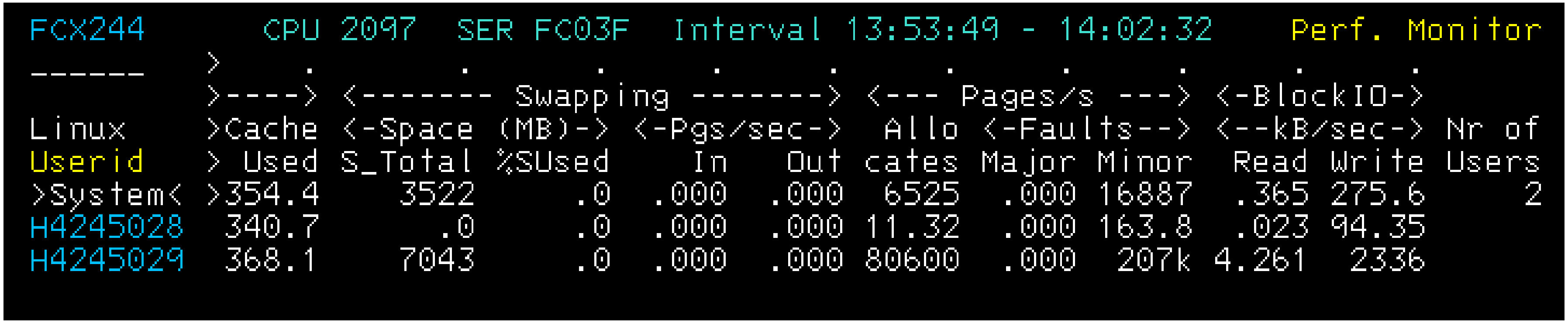
```
FCX243      CPU 2097   SER FC03F   Interval 13:50:14 - 13:51:14    Perf. Monitor
------         .>     .     .      .      .      .      .     .      .     .
                >-----------------> <------------ Processes ------------->
Linux     Virt>-----------------> <---- Current -----> <-Average Running-> Nr of
Userid    CPUs>OWait   Idle Stolen Runabl Waiting Total 1_Min  5_Min 15_Min Users
>System<   3.0>    .1    .3    2.6    5.0      .0  93.0  3.45   2.66   1.68     2
H4245028     3     .0    .3    2.6      4       0    96  3.16   2.06   1.31
H4245029     3     .1    .3    2.6      6       0    90  3.74   3.25   2.04
```

appldata – memory monitor

Linux monitoring data collected by appldata_mem as processed and displayed by z/VM Performance Toolkit:



```
FCX244      CPU 2097   SER FC03F   Interval 13:53:49 - 14:02:32    Perf. Monitor
______          .      .      .      .      .      .      .      .      .
          <------------ Memory Allocation (MB) -------------> <------- Swapping
Linux     <--- Main ---> <--- High --->       Buffers  Cache <-Space (MB)-> <-
Userid    M_Total %MUsed H_Total %HUsed Shared /CaFree   Used S_Total %SUsed
>System<    996.3   48.5      .0     .0     .0    39.6  354.4    3522     .0  .
H4245028    996.6   46.8      .0     .0     .0    42.3  340.7      .0     .0  .
H4245029    996.0   50.2      .0     .0     .0    36.8  368.1    7043     .0  .
```



```
FCX244      CPU 2097   SER FC03F   Interval 13:53:49 - 14:02:32    Perf. Monitor
______    >     .      .      .      .      .      .      .      .      .
          >----> <------- Swapping -------> <--- Pages/s ---> <-BlockIO->
Linux     >Cache <-Space (MB)-> <-Pgs/sec->  Allo <-Faults--> <--kB/sec-> Nr of
Userid    > Used S_Total %SUsed    In   Out cates Major Minor  Read Write Users
>System<  >354.4    3522     .0  .000  .000  6525  .000 16887  .365 275.6     2
H4245028   340.7      .0     .0  .000  .000 11.32  .000 163.8  .023 94.35
H4245029   368.1    7043     .0  .000  .000 80600  .000  207k 4.261  2336
```

Monitoring with hypfs

Virtual Linux file system

- Uses diagnose calls to gather guest data from hypervisor
- Works with LPAR hypervisor or z/VM
- resources controlled by hypervisor, i.e. physical CPUs
- resources provided to guest systems, i.e virtual CPUs

Preconditions

- LPAR: enable “Global performance data control” checkbox in HMC activation profile of the guest where hypfs is mounted
- z/VM: privilege class B required for the guest where hypfs is mounted

mounting hypfs:

```
root@larsson:~# mount -t s390_hypfs /sys/hypervisor/s390/
```

hypfs is populated with initial data when being mounted and hypfs data is only updated on request:

```
root@larsson:~# echo 1 > /sys/hypervisor/s390/update
```

Hypfs structure under z/VM

```
/sys/hypervisor/s390
|-- update
|-- cpus
|   `-- count
|-- hyp
|   `-- type
`-- systems
    |-- <guest-name>
    |   |-- onlinetime_us
    |   |-- cpus
    |   |   |-- capped
    |   |   |-- count
    |   |   |-- cputime_us
    |   |   |-- dedicated
    |   |   |-- weight_cur
    |   |   |-- weight_max
    |   |   `-- weight_min
    |   |-- mem
    |   |   |-- max_KiB
    |   |   |-- min_KiB
    |   |   |-- share_KiB
    |   |   `-- used_KiB
    |   `-- samples
    |       |-- cpu_delay
    |       |-- cpu_using
    |       |-- idle
    |       |-- mem_delay
    |       |-- other
    |       `-- total
    |
```

- systems/onlinetime_us: time since guest activation
- systems/cpus:
  - capped: 0=off, 1=soft, 2=hard
  - count: number of virtual CPUs
  - cputime_us: actual use time
  - dedicated: 0=no, 1=yes
  - weight_cur, weight_min, weight_max: current, minimum and maximum share of guest (1-10000; 0=ABSOLUTE SHARE)
- systems/mem:
  - max_KiB: memory limit granted to guest
  - min_KiB: minimum memory requirement of guest
  - share_KiB: suggested guest memory size estimated by z/VM
  - used_KiB: current memory footprint of guest
- systems/samples:
  - cpu_delay: guest waiting for CPU
  - cpu_using: guest doing work
  - idle: guest being idle
  - mem_delay: guest waiting for memory to be paged in
  - other: other samples
  - total: total samples

Hypfs example under z/VM

```
[root@h4245005 ~]# find /sys/hypervisor/s390/systems/H4245005/
  -type f | while read f; do echo -n "$f: "; cat $f; done
/sys/hypervisor/s390/systems/H4245005/samples/total: 500061
/sys/hypervisor/s390/systems/H4245005/samples/other: 30152
/sys/hypervisor/s390/systems/H4245005/samples/idle: 469694
/sys/hypervisor/s390/systems/H4245005/samples/mem_delay: 0
/sys/hypervisor/s390/systems/H4245005/samples/cpu_delay: 43
/sys/hypervisor/s390/systems/H4245005/samples/cpu_using: 172
/sys/hypervisor/s390/systems/H4245005/mem/share_KiB: 319004
/sys/hypervisor/s390/systems/H4245005/mem/used_KiB: 319004
/sys/hypervisor/s390/systems/H4245005/mem/max_KiB: 1048576
/sys/hypervisor/s390/systems/H4245005/mem/min_KiB: 0
/sys/hypervisor/s390/systems/H4245005/cpus/weight_cur: 100
/sys/hypervisor/s390/systems/H4245005/cpus/weight_max: 10000
/sys/hypervisor/s390/systems/H4245005/cpus/weight_min: 6
/sys/hypervisor/s390/systems/H4245005/cpus/count: 6
/sys/hypervisor/s390/systems/H4245005/cpus/dedicated: 0
/sys/hypervisor/s390/systems/H4245005/cpus/capped: 0
/sys/hypervisor/s390/systems/H4245005/cpus/cputime_us: 203792603
/sys/hypervisor/s390/systems/H4245005/onlinetime_us: 166806841739
```

hyptop - Display hypervisor performance data

6.1

The hyptop command provides a dynamic real-time view of a hypervisor environment on System z.

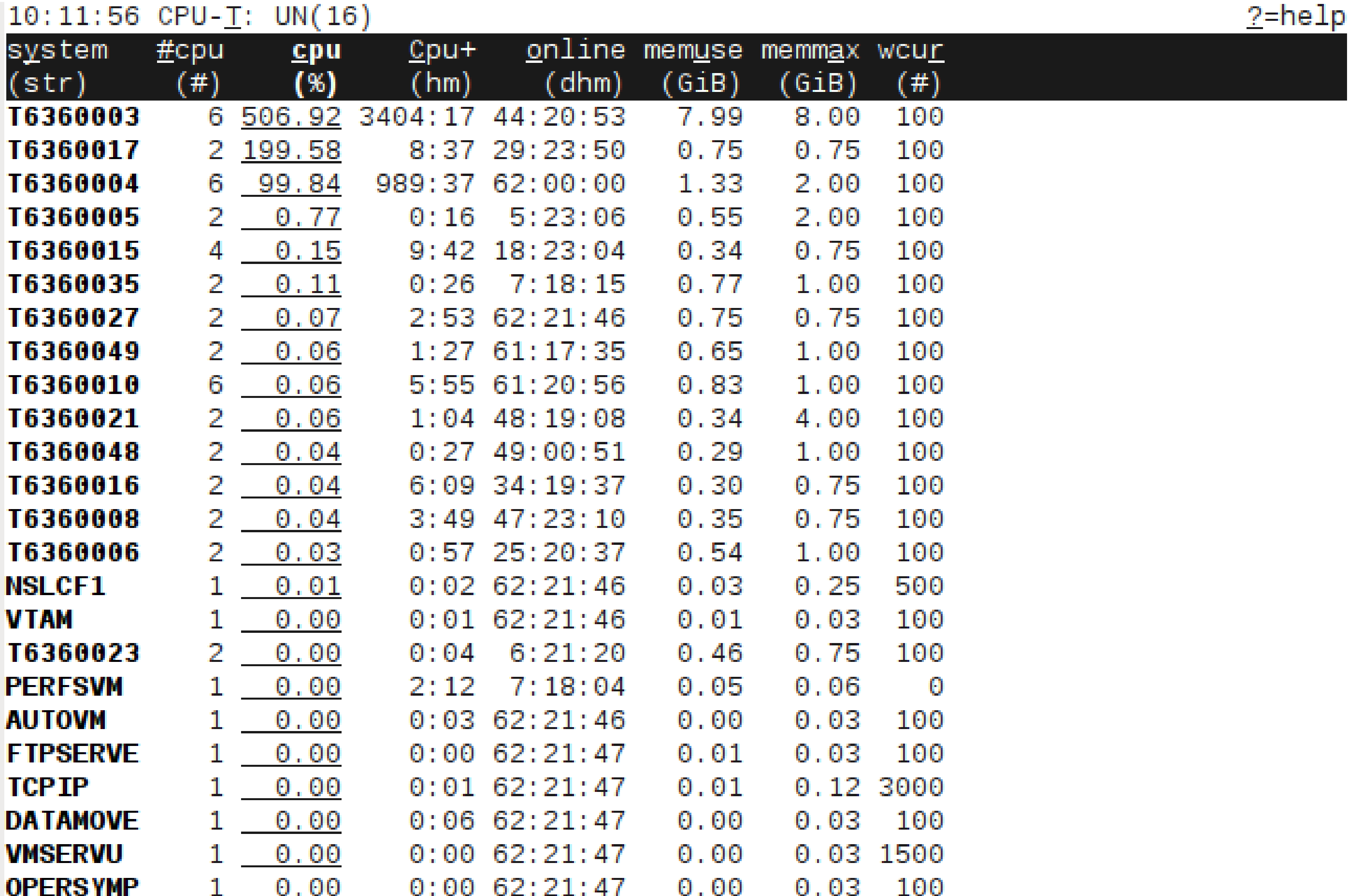
- It works with both the z/VM and the LPAR PR/SM hypervisor.
- Depending on the available data it shows, for example, CPU and memory information about running LPARs or z/VM guest operating systems.

The following things are required to run hyptop:

- The debugfs file system must be mounted.
- The hyptop user must have read permission for the required debugfs files:
  - z/VM: <debugfs mount point>/s390_hypfs/diag_2fc
  - LPAR: <debugfs mount point>/s390_hypfs/diag_204
- To monitor all LPARs or z/VM guest operating systems of the hypervisor, your system must have additional permissions:
  - For z/VM: The guest must be privilege class B.
  - For LPAR: On the HMC or SE security menu of the LPAR activation profile, select the Global performance data control checkbox.

hyptop – Displaying hypervisor performance data

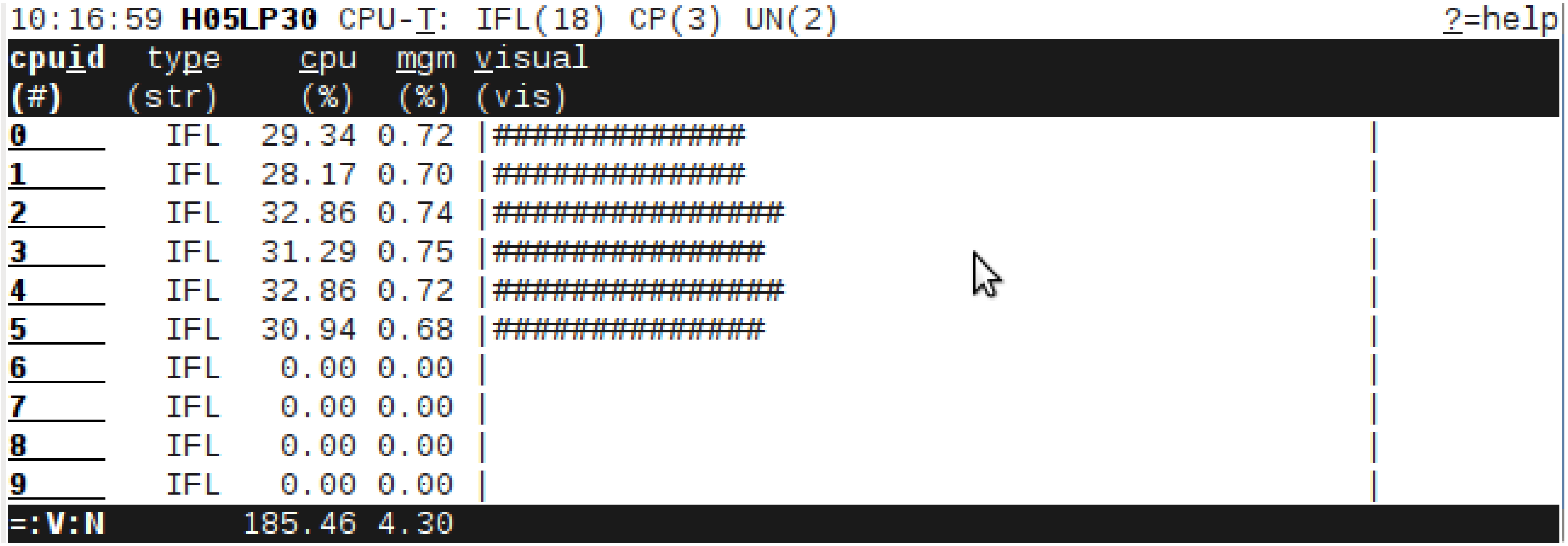
Displaying performance data for the z/VM hypervisor



```
10:11:56 CPU-T: UN(16)                                                    ?=help
system   #cpu    cpu    Cpu+   online memuse memmax wcur
(str)     (#)    (%)    (hm)    (dhm)  (GiB)  (GiB)  (#)
T6360003    6 506.92 3404:17 44:20:53    7.99    8.00  100
T6360017    2 199.58    8:37 29:23:50    0.75    0.75  100
T6360004    6  99.84  989:37 62:00:00    1.33    2.00  100
T6360005    2   0.77    0:16  5:23:06    0.55    2.00  100
T6360015    4   0.15    9:42 18:23:04    0.34    0.75  100
T6360035    2   0.11    0:26  7:18:15    0.77    1.00  100
T6360027    2   0.07    2:53 62:21:46    0.75    0.75  100
T6360049    2   0.06    1:27 61:17:35    0.65    1.00  100
T6360010    6   0.06    5:55 61:20:56    0.83    1.00  100
T6360021    2   0.06    1:04 48:19:08    0.34    4.00  100
T6360048    2   0.04    0:27 49:00:51    0.29    1.00  100
T6360016    2   0.04    6:09 34:19:37    0.30    0.75  100
T6360008    2   0.04    3:49 47:23:10    0.35    0.75  100
T6360006    2   0.03    0:57 25:20:37    0.54    1.00  100
NSLCF1      1   0.01    0:02 62:21:46    0.03    0.25  500
VTAM        1   0.00    0:01 62:21:46    0.01    0.03  100
T6360023    2   0.00    0:04  6:21:20    0.46    0.75  100
PERFSVM     1   0.00    2:12  7:18:04    0.05    0.06    0
AUTOVM      1   0.00    0:03 62:21:46    0.00    0.03  100
FTPSERVE    1   0.00    0:00 62:21:47    0.01    0.03  100
TCPIP       1   0.00    0:01 62:21:47    0.01    0.12 3000
DATAMOVE    1   0.00    0:06 62:21:47    0.00    0.03  100
VMSERVU     1   0.00    0:00 62:21:47    0.00    0.03 1500
OPERSYMP    1   0.00    0:00 62:21:47    0.00    0.03  100
```

hyptop – Displaying hypervisor performance data

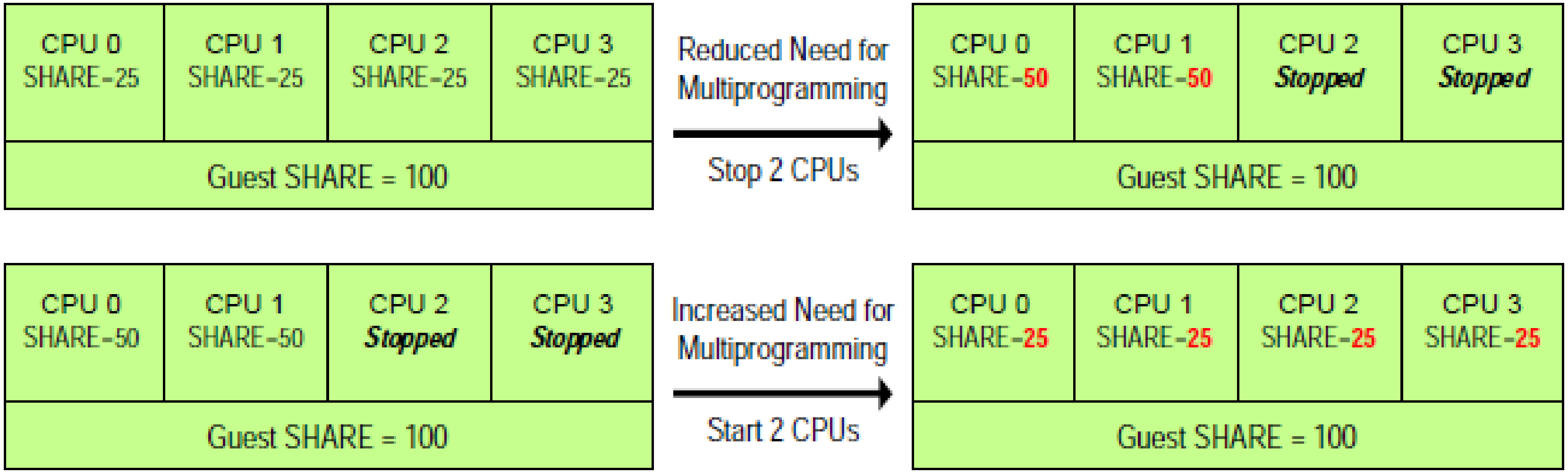
Displaying performance data for a single LPAR



```
10:16:59 H05LP30 CPU-T: IFL(18) CP(3) UN(2)                               ?=help
cpuid  type    cpu  mgm visual
(#)   (str)    (%)  (%) (vis)
0       IFL  29.34 0.72 |#############                                |
1       IFL  28.17 0.70 |#############                                |
2       IFL  32.86 0.74 |###############                              |
3       IFL  31.29 0.75 |##############                               |
4       IFL  32.86 0.72 |###############                              |
5       IFL  30.94 0.68 |##############                               |
6       IFL   0.00 0.00 |                                             |
7       IFL   0.00 0.00 |                                             |
8       IFL   0.00 0.00 |                                             |
9       IFL   0.00 0.00 |                                             |
=:V:N       185.46 4.30
```

z/VM 5.4: Virtual CPU SHARE Redistribution

- Allows z/VM guests to expand or contract the number of virtual processors it uses without affecting the overall CPU capacity it is allowed to consume
  - Guests can dynamically optimize their multiprogramming capacity based on workload demand
  - Starting and stopping virtual CPUs does not affect the total amount of CPU capacity the guest is authorized to use
  - Linux CPU hotplug daemon starts and stops virtual CPUs based on Linux load average value
- Helps enhance the overall efficiency of a Linux-on-z/VM environment
- Previously, stopped virtual processors were given a portion of the guest share.



CPU 0 SHARE=25	CPU 1 SHARE=25	CPU 2 SHARE=25	CPU 3 SHARE=25
Guest SHARE = 100			

Reduced Need for Multiprogramming → Stop 2 CPUs

CPU 0 SHARE=50	CPU 1 SHARE=50	CPU 2 *Stopped*	CPU 3 *Stopped*
Guest SHARE = 100			

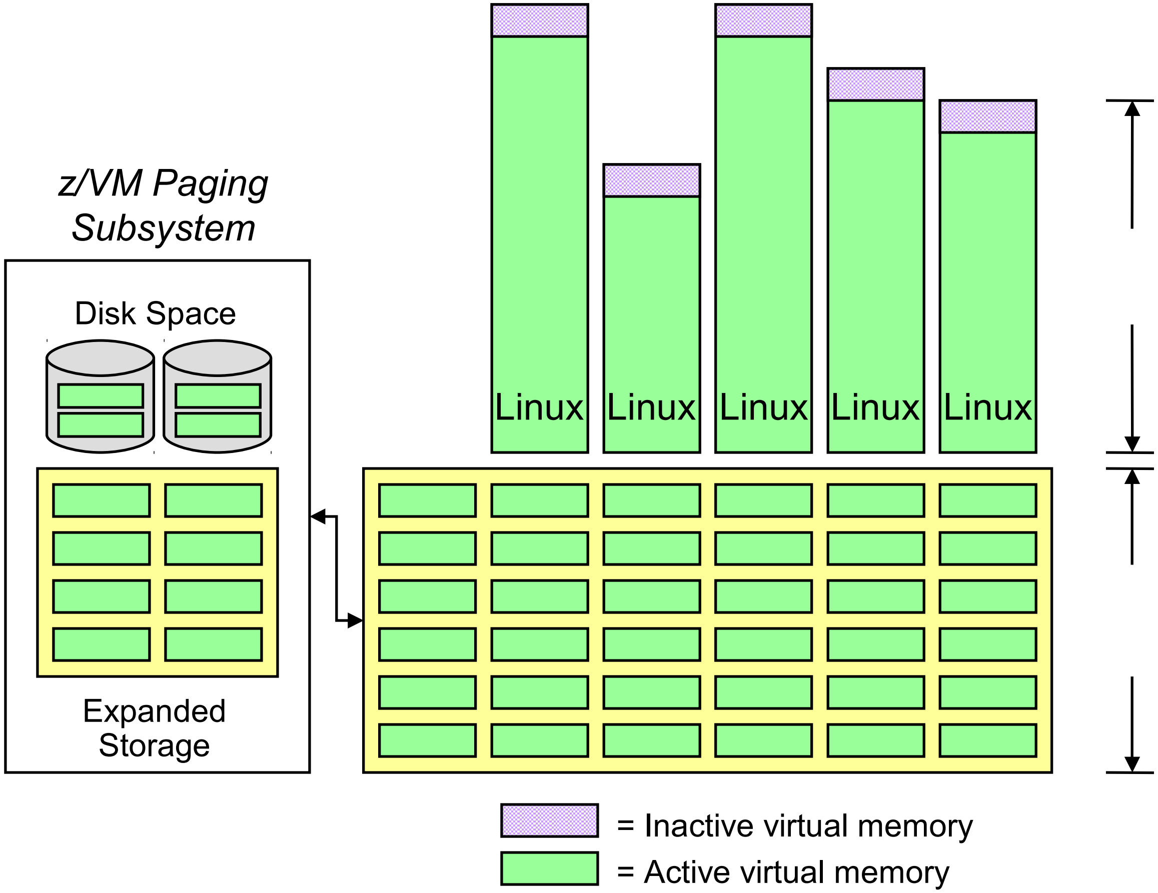
CPU 0 SHARE=50	CPU 1 SHARE=50	CPU 2 *Stopped*	CPU 3 *Stopped*
Guest SHARE = 100			

Increased Need for Multiprogramming → Start 2 CPUs

CPU 0 SHARE=25	CPU 1 SHARE=25	CPU 2 SHARE=25	CPU 3 SHARE=25
Guest SHARE = 100			

Cooperative Memory Management (CMM)

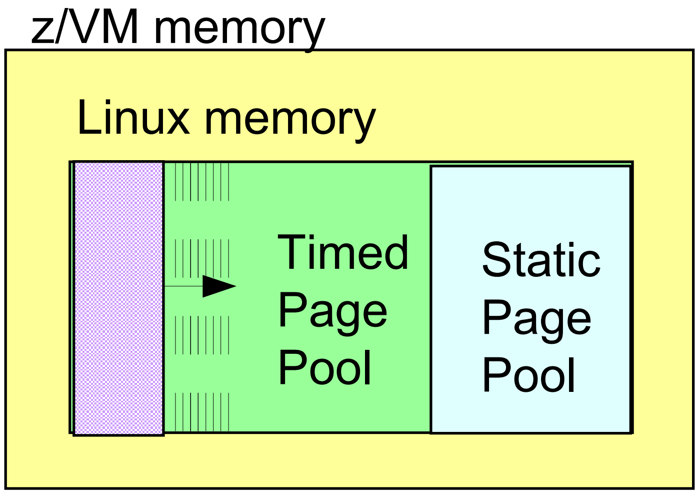
- The z/VM hypervisor maps guest virtual memory into the real memory storage of the System z machine.
- If there aren't enough real memory frames to contain all the active guests' virtual memory pages, some pages are moved to expanded storage
- When expanded storage is full, the pages of the guest are stored on the paging disk space (dasd)
- Inactive virtual memory pages inside the Linux guest must be recovered for use by other guest systems

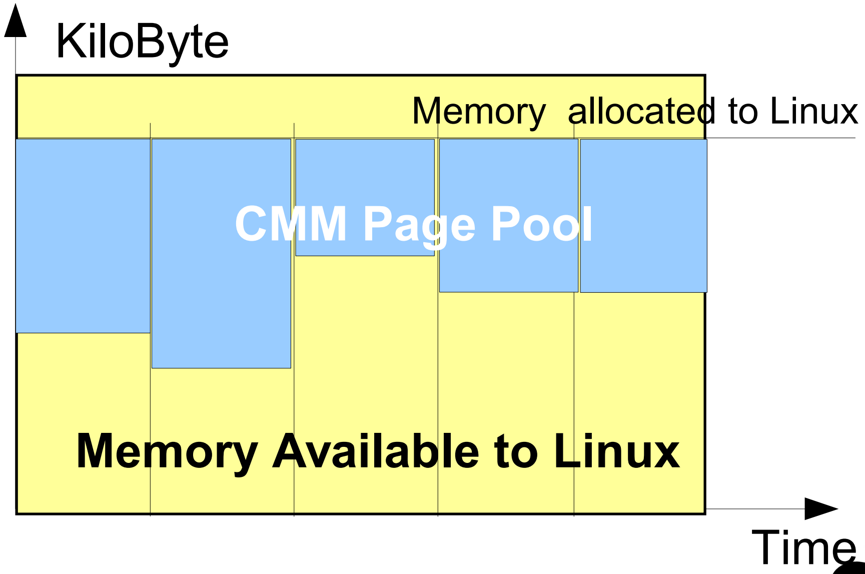


Learn more at: http://ibm.com/servers/eserver/zseries/zvm/sysman/vmrm/vmrmcmm.html

CMM: Linux Implementation

- To reduce Linux guest memory size CMM allocates pages to page pools (“balloon”) that make the pages unusable to Linux.
- Currently two such page pools exist for a Linux guest
  - **Timed page pool:** Pages are released from this pool at a speed set in the release rate. According to guest activity and overall memory usage on z/VM, a resource manager adds pages at intervals. If no pages are added and the release rate is not zero, the pool will empty
  - **Static page pool:** The page pool is controlled by a resource manager that changes the pool size at intervals according to guest activity as well as overall memory usage on z/VM





Cooperative Memory Management - Linux

- Linux uses a IUCV special message interface for z/VM interaction (CMMSHRINK/CMMRELEASE/CMMREUSE)
- Linux support is available since SLES9 SP3 and RHEL4 U7

```
root@larsson:~# modprobe cmm
root@larsson:~# echo 100 > /proc/sys/vm/cmm_timed_pages
root@larsson:~# echo 10 1 > /proc/sys/vm/cmm_timeout
root@larsson:~# cat /proc/sys/vm/cmm_timed_pages
60
root@larsson:~# echo 100 > /proc/sys/vm/cmm_pages
```

- `/proc/sys/vm/cmm_pages`
  - Read to query number of pages permanently reserved
  - Write to set new target (will be achieved over time)
- `/proc/sys/vm/cmm_timed_pages`
  - Read to query number of pages temporarily reserved
  - Write increment to add to target
- `/proc/sys/vm/cmm_timeout`
  - Holds pair of N pages / X seconds (read/write)
  - Every time X seconds have passed, release N temporary pages

cpuplugd: Example Configuration

```
UPDATE="60"

CPU_MIN="2"
CPU_MAX="10"

HOTPLUG = "(loadavg > onumcpus +0.75) & (idle < 10.0)"
HOTUNPLUG = "(loadavg < onumcpus -0.25) | (idle > 50)"


CMM_MIN="0"
CMM_MAX="8192"
CMM_INC="256"

MEMPLUG = "swaprate > freemem+10 & freemem+10 < apcr"
MEMUNPLUG = "swaprate > freemem + 10000"
```

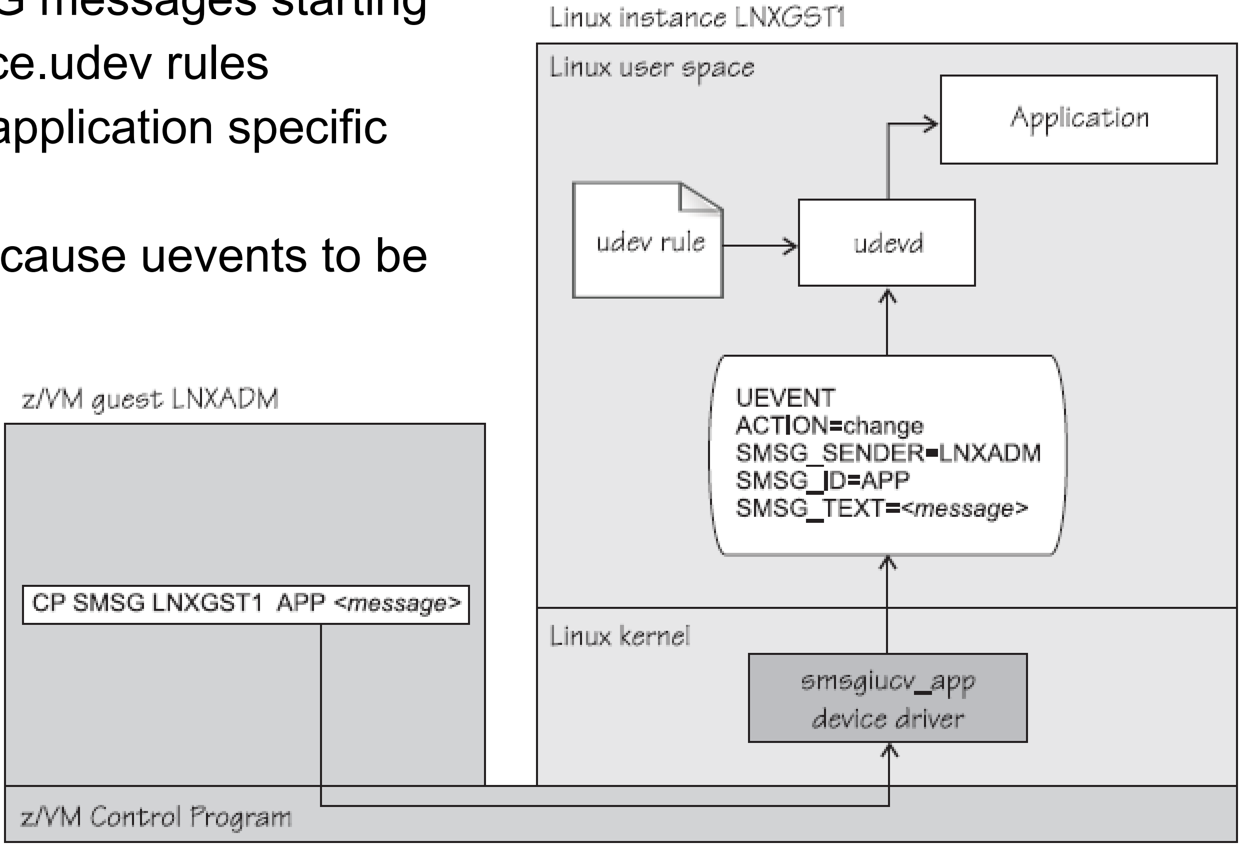
Deliver z/VM CP special messages as uevent

6.1

Allows to forward SMSG messages starting with “APP” to user space.udev rules can be used to trigger application specific actions

The special messages cause uevents to be generated.

See “Writing udev rules for handling CP special messages” on page 229 in the Device Driver Book for information about handling the uevents.



```
UEVENT
ACTION=change
SMSG_SENDER=LNXADM
SMSG_ID=APP
SMSG_TEXT=<message>
```

zipl integration of device mapper devices

6.1

- zipl provides a helper script, ***zipl_helper.device-mapper***, that detects the required information and provides it to zipl for you.
- To use the helper script run zipl as usual, specifying the parameters for the kernel image, parameter file, initial RAM disk, and target.
- Assuming an example device for which the location of the kernel image is /boot/image-5, the location of an initial RAM disk as /boot/initrd-5, a kernel parameter file /boot/parmf-5, and which writes the required boot loader code to /boot and is a device mapper device, the command then becomes:

```
root@larsson:~# zipl -i /boot/image-5 -r /boot/initrd-5 -p
/boot/parmf-5 -t /boot
```

The corresponding configuration file section becomes:

```
[boot5]
image=/boot/image-5
ramdisk=/boot/initrd-5
paramfile=/boot/parmf-5
target=/boot
```

cio_ignore

When a Linux on System z instance boots, it senses and analyses all available devices. You can use the cio_ignore kernel parameter to specify a list of devices that are to be ignored.

The following applies to ignored devices:

- Ignored devices are not sensed and analyzed. The device cannot be used unless it has been analyzed.
- Ignored devices are not represented in sysfs.
- Ignored devices do not occupy storage in the kernel.
- The subchannel to which an ignored device is attached is treated as if no device were attached.
- cio_ignore might hide essential devices such as the console under z/VM. The console is typically device number 0.0.0009.
-

This example specifies that all devices in the range 0.0.b100 through 0.0.b1ff, and the device 0.0.a100 are to be ignored.

```
cio_ignore=0.0.b100-0.0.b1ff,0.0.a100
```

cio_ignore (cont.)

Display ignored devices:

```
root@larsson:~>  cat /proc/cio_ignore
0.0.0000-0.0.78ff
0.0.f503-0.0.ffff
```

Free a individual device from the ignore list

```
root@larsson:~>  echo free 0.0.4711 >/proc/cio_ignore
```

Free all devices from the ignore list

```
root@larsson:~>  echo free all >/proc/cio_ignore
```

Use cio_ignore tool to manage the I/O device exclusion list

```
root@larsson:~>  cio_ignore -l
Ignored devices:
=================
0.0.0000-0.0.0008
0.0.000a-0.0.6365
[...]
```

cio_ignore (cont'd)

Use the -L option to display the devices which are accessible

```
root@larsson:~>  cio_ignore -L
Accessible devices:
===================
0.0.0009
0.0.6366
0.0.f5f0-0.0.f5f2
```

Use the -r option to remove devices from the exclusion list

```
root@larsson:~>  cio_ignore -r 6366
```

The the -R option is used to free all devices

Use the -a option to add devices to the exclusion list

```
root@larsson:~>  cio_ignore -a 4000-5fff
```

Use the -k option to create the kernel parameter list string

```
root@larsson:~>  cio_ignore -k
cio_ignore=all,!0009,!6366,!f5f0-f5f2
```

More Information

Live Virtual Classes for z/VM and Linux

http://www.vm.ibm.com/education/lvc/

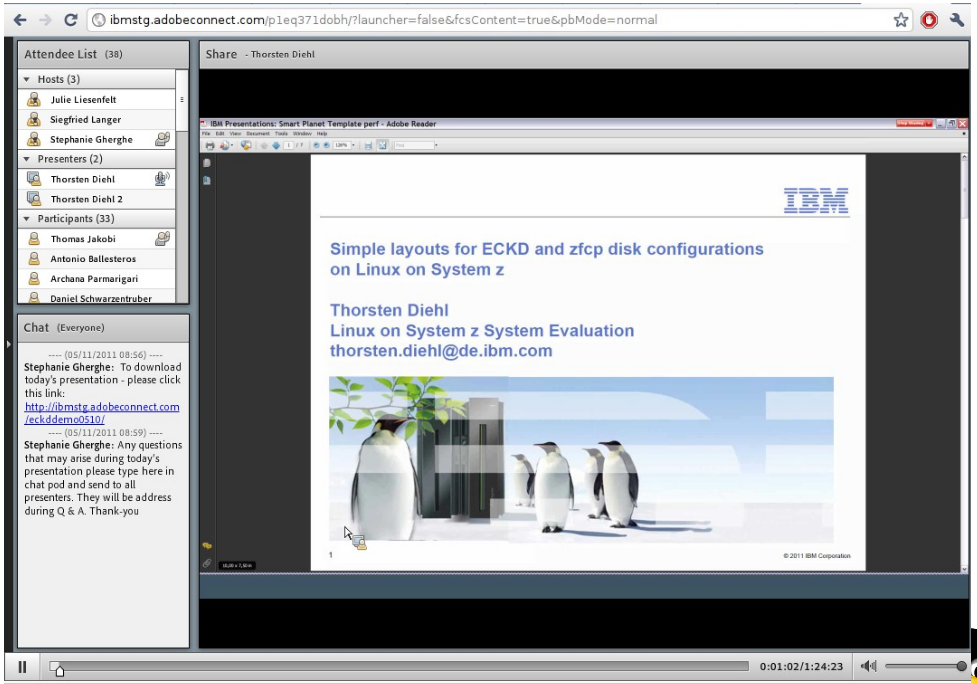
IBM offers education on a variety of z/VM, Linux on System z and z/VSE topics in the form of 'Live Virtual Classes' (LVC) available on the Internet <u>for Customers, Business Partners and IBMers</u>

The day of the LVC broadcast, you can see the charts and listen to the speaker 'live'. In addition, you are able (and are encouraged) to ask questions of the speaker during a Q&A session following the prepared presentation.

* The day following each LVC, we post the the charts in PDF format.

* Shortly thereafter we provide a replay where you can read the charts, hear the recording and the Q's and A's in MP3 Format

*. You are welcome to read the charts or listen to the replay without registration when you can't participate 'live' or even if you wish to hear it all again.



ibmstg.adobeconnect.com/p1eq371dobh/?launcher=false&fcsContent=true&pbMode=normal

Attendee List (38)

Hosts (3)
Julie Liesenfelt
Siegfried Langer
Stephanie Gherghe

Presenters (2)
Thorsten Diehl
Thorsten Diehl 2

Participants (33)
Thomas Jakobi
Antonio Ballesteros
Archana Parmarigari
Daniel Schwarzentruber

Chat (Everyone)

---- (05/11/2011 08:56) ----
Stephanie Gherghe: To download today's presentation - please click this link: http://ibmstg.adobeconnect.com/eckddemo0510/

---- (05/11/2011 08:59) ----
Stephanie Gherghe: Any questions that may arise during today's presentation please type here in chat pod and send to all presenters. They will be address during Q & A. Thank-you

Share - Thorsten Diehl

IBM Presentations: Smart Planet Template perf - Adobe Reader

Simple layouts for ECKD and zfcp disk configurations on Linux on System z

Thorsten Diehl
Linux on System z System Evaluation
thorsten.diehl@de.ibm.com

1 © 2011 IBM Corporation

LVC 2011

January 26, 2011

- **Best Practices for WebSphere Application Server on System z Linux**
  An introduction to setting up an infrastructure that will allow WebSphere applications to run efficiently on Linux for System z.
  Speaker: Steve Wehr

February 16 & 17 (3 sessions – U.S. am + pm, Asia & Europe)

- **Lessons learned from putting Linux on System z in Production**
  This session will give you a candid insight on how customers around the world dealt with these topics.
  Recommendations of “best practices” will be included.
  Speaker: Hans-Joachim Picht

March 16 & 17 (3 sessions – U.S. am + pm, Asia & Europe)

- **Linux on System z RHEL 6 Performance Report**
  This presentation covers the overall status of RHEL6 from a System z performance focus.
  Speaker: Christian Ehrhardt

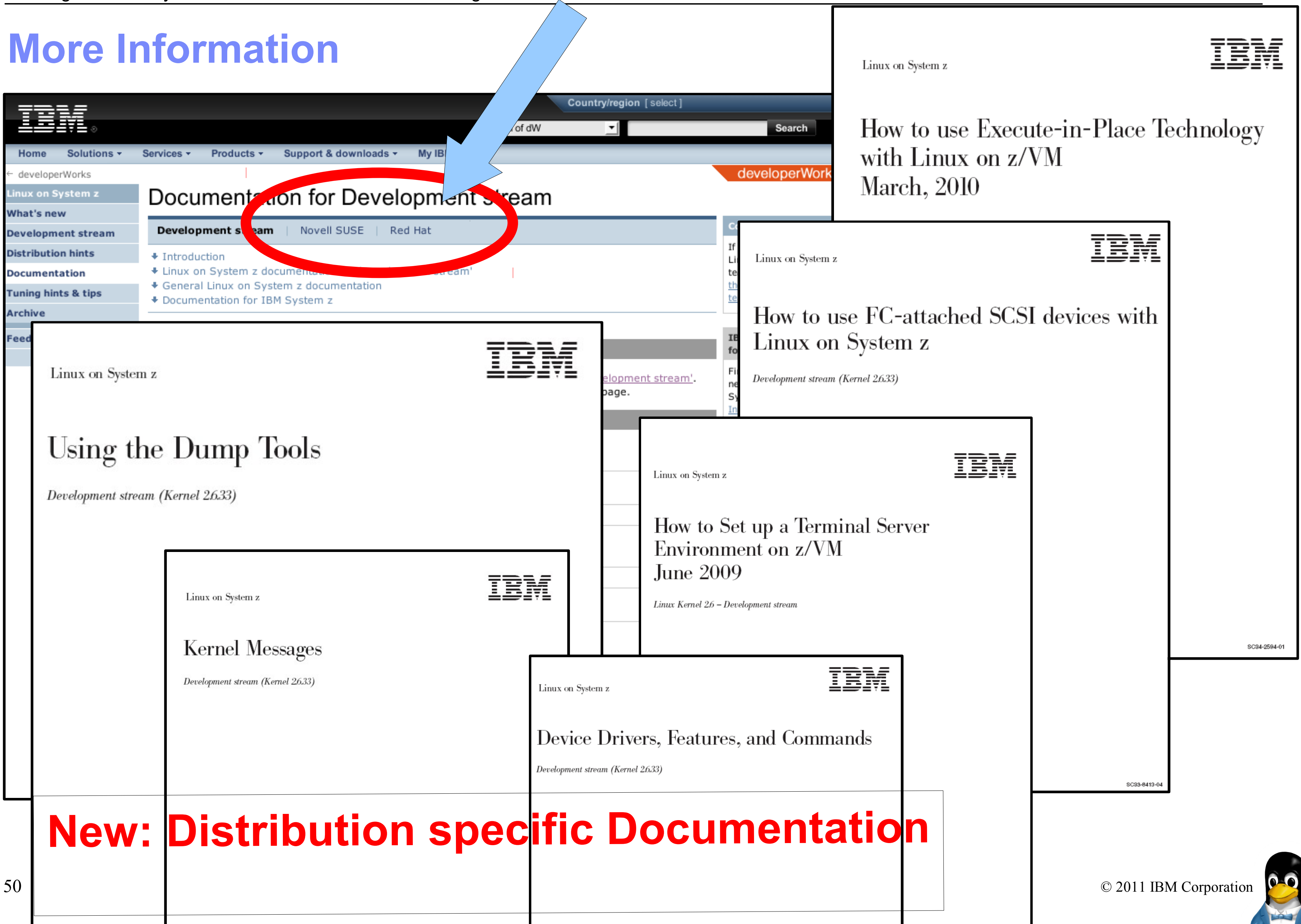
April 6 & 7 (2 session – U.S. pm, Asia & Europe)

- **Problem Reporting and Analysis Linux on System z - How to survive a Linux critical situation**
  You encounter a problem with Linux on System z and you don't know what to do. This webcast will introduce you to a trouble shooting "First Aid Kit" for Linux on System z.
  Speaker: Sven Schuetz

May 10 & 11

- **Live Demo: Setup of simple and multipathed disk I/O configurations of ECKD and zfcp Volumes on Linux on System z**
  During this "Live Demo" you will see how ECKD DASD is added to a running SLES 11 Service Pack 1 on System z and how you can exploit HyperPAV to improve DASD performance. In a second part watch zfcp volumes being added to a Linux single path and multipath with LVM.
  Speaker: Thorsten Diehl

More Information



Documentation for Development stream

Development stream | Novell SUSE | Red Hat

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

Linux on System z

How to use Execute-in-Place Technology with Linux on z/VM
March, 2010

SC34-2594-01

Linux on System z

How to use FC-attached SCSI devices with Linux on System z

Development stream (Kernel 2.6.33)

SC33-8413-04

Linux on System z

Using the Dump Tools

Development stream (Kernel 2.6.33)

Linux on System z

How to Set up a Terminal Server Environment on z/VM
June 2009

Linux Kernel 2.6 – Development stream

Linux on System z

Kernel Messages

Development stream (Kernel 2.6.33)

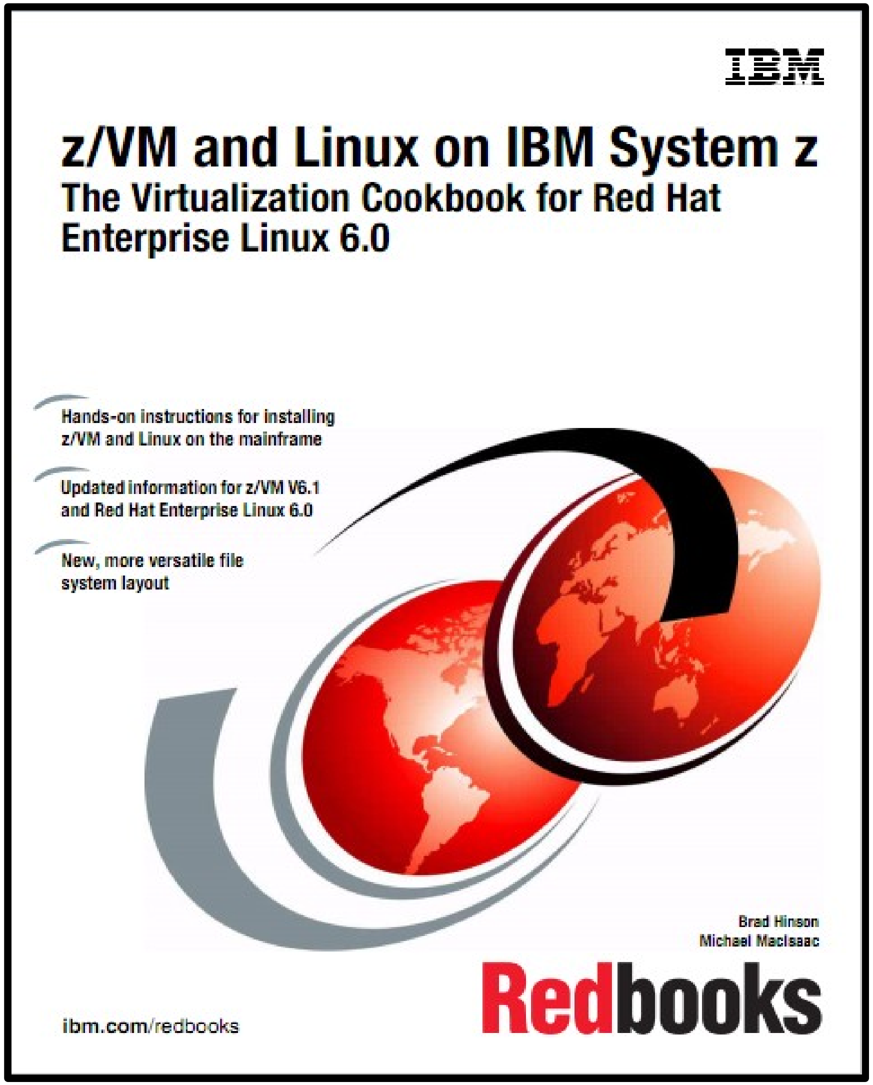
Linux on System z

Device Drivers, Features, and Commands

Development stream (Kernel 2.6.33)

New: Distribution specific Documentation

More Information



IBM

z/VM and Linux on IBM System z

The Virtualization Cookbook for Red Hat Enterprise Linux 6.0

Hands-on instructions for installing z/VM and Linux on the mainframe

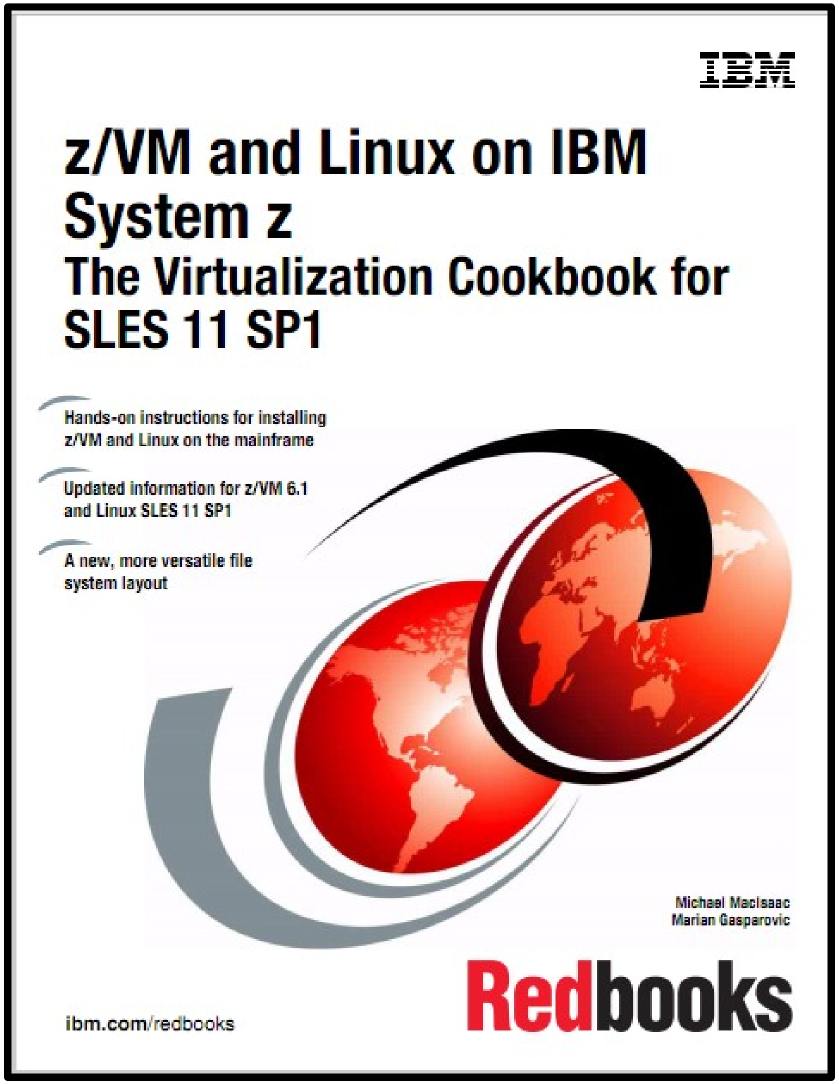
Updated information for z/VM V6.1 and Red Hat Enterprise Linux 6.0

New, more versatile file system layout

Brad Hinson
Michael MacIsaac

ibm.com/redbooks

Redbooks



IBM

z/VM and Linux on IBM System z

The Virtualization Cookbook for SLES 11 SP1

Hands-on instructions for installing z/VM and Linux on the mainframe

Updated information for z/VM 6.1 and Linux SLES 11 SP1

A new, more versatile file system layout

Michael MacIsaac
Marian Gasparovic

ibm.com/redbooks

Redbooks

Your Linux on System z Requirements?

Are you missing a certain feature, functionality or tool? **We'd love to hear from you!**

We will evaluate each request and (hopefully) develop the additional functionality you need.

Send your input to hans@de.ibm.com

Questions?

Hans-Joachim Picht

Linux Technology Center

IBM Deutschland Research
& Development GmbH
Schönaicher Strasse 220
71032 Böblingen, Germany

Mobile +49 (0)175 - 1629201
hans@de.ibm.com

Trademarks & Disclaimer

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DB2, e-business logo, ESCON, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, System Storage, System z9, VM/ESA, VSE/ESA, WebSphere, xSeries, z/OS, zSeries, z/VM.

The following are trademarks or registered trademarks of other companies

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries. LINUX is a registered trademark of Linux Torvalds in the United States and other countries. UNIX is a registered trademark of The Open Group in the United States and other countries. Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation. SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC. Intel is a registered trademark of Intel Corporation. * 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. References in this document to IBM products or services do not imply that IBM intends to make them available in every country. Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use. The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice. Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.