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

Introduction to Linux on System z

2010 Blooming Basics T3 for z/VM Linux on IBM System z
June 22-23-24, 2010, 2:00 - 5:00 PM EDT (NY), T3 Conference Calls

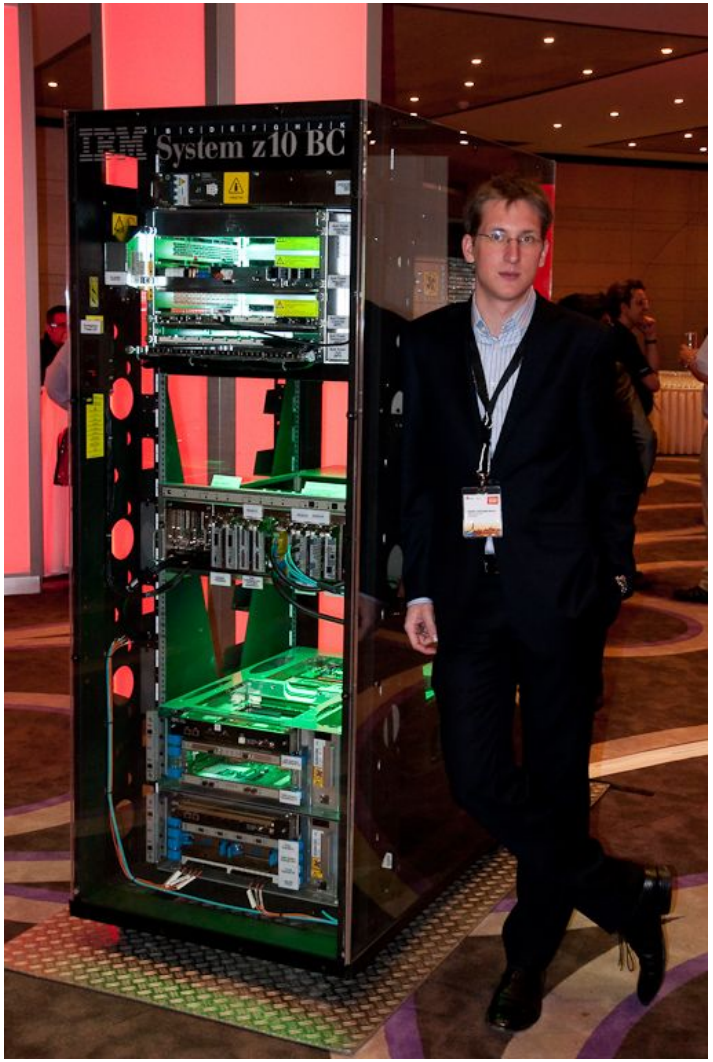
Facts on Linux

- Last year, **75%** of the Linux code was developed by **programmers working for corporations**.
- **$7.37 billion:** projected cost to produce the 283 million lines of code which are contained in Linux Distribution **in a commercial environment.**
- IDC forecasts show that **Linux server revenue will grow by 85.5%** between 2008 and 2012 **in the non-x86 server space** equalling a four year compound annual growth rate of 16.7%.
- **Linux is Linux**, but …features, properties and quality differ dependent on your platform

Source: Intelligence Slideshow: 40 Fast Facts on Linux http://www.baselinemag.com/c/a/Intelligence/40-Fast-Facts-on-Linux-727574/
http://www.internetnews.com/dev-news/article.php/3659961
http://public.dhe.ibm.com/software/au/downloads/IBM_zLinux_DAG_FINAL.pdf

What you see depends on the perception...or your background

<<This is a mainframe, and can lately also be used to run the Linux Operating System>>



<<This is a high end server using Linux & Virtualization Technology for massive Server Consolidation (IBM Enterprise Linux Server)>>

What is Linux on System z

Linux on System z exploits the strengths and reliability features of the System z hardware, while preserving the openness of Linux.

- Linux on System z is the synonym for Linux running on any IBM mainframe, including:
  - IBM System z10™,
  - IBM System z9®
  - IBM eServer™ zSeries™ (z990, z890, z900, z800)
  - S/390® (9672 G5, G6 and Multiprise® 3000 processors).
- It is a native System z operating environment
  - Pure Linux, an ASCII environment
  - Exploits IBM S/390 hardware, including IEEE floating point
- Not a replacement for other IBM System z operating systems
- Can be run under z/VM or in its own LPAR
- There is not a unique version of Linux or other operating system
  - Developed by the Open Source Community
  - Distributed and Supported by SuSE/Novell, Red Hat, and Others

What System z Hardware Brings to Linux

- The most reliable hardware platform available
  - MTF measured in decades
  - RAS features built into hardware
- Scalability
  - Both Physical and Logical
  - Non-disruptive capacity upgrade on demand
- Designed to support mixed work loads
  - Complete work load isolation
  - High speed inter-server connectivity
  - High Internal Bandwidth, sophisticated cache nest
  - Virtualization
- Hipersockets
  - Virtual network between LPARs, at memory speed
- Integrated Facility for Linux (IFL)
  - Lower priced hardware
  - Protects against software costs

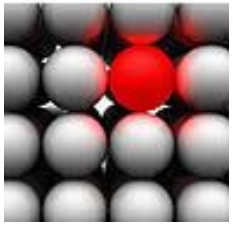
Linux on System z Value Proposition

A Linux on System z infrastructure provides significant IT cost savings opportunities:

- **People Cost** – increase the productivity of the IT staff
- **Software Costs** – reduce software license fees by consolidating a large number of virtual machines per System z core
- **Energy Costs** – dramatic reduction in power usage
- **Facilities Costs** – reduce floor space with dense server consolidation on System z
- **Networking Costs** – consolidate networks inside the box too







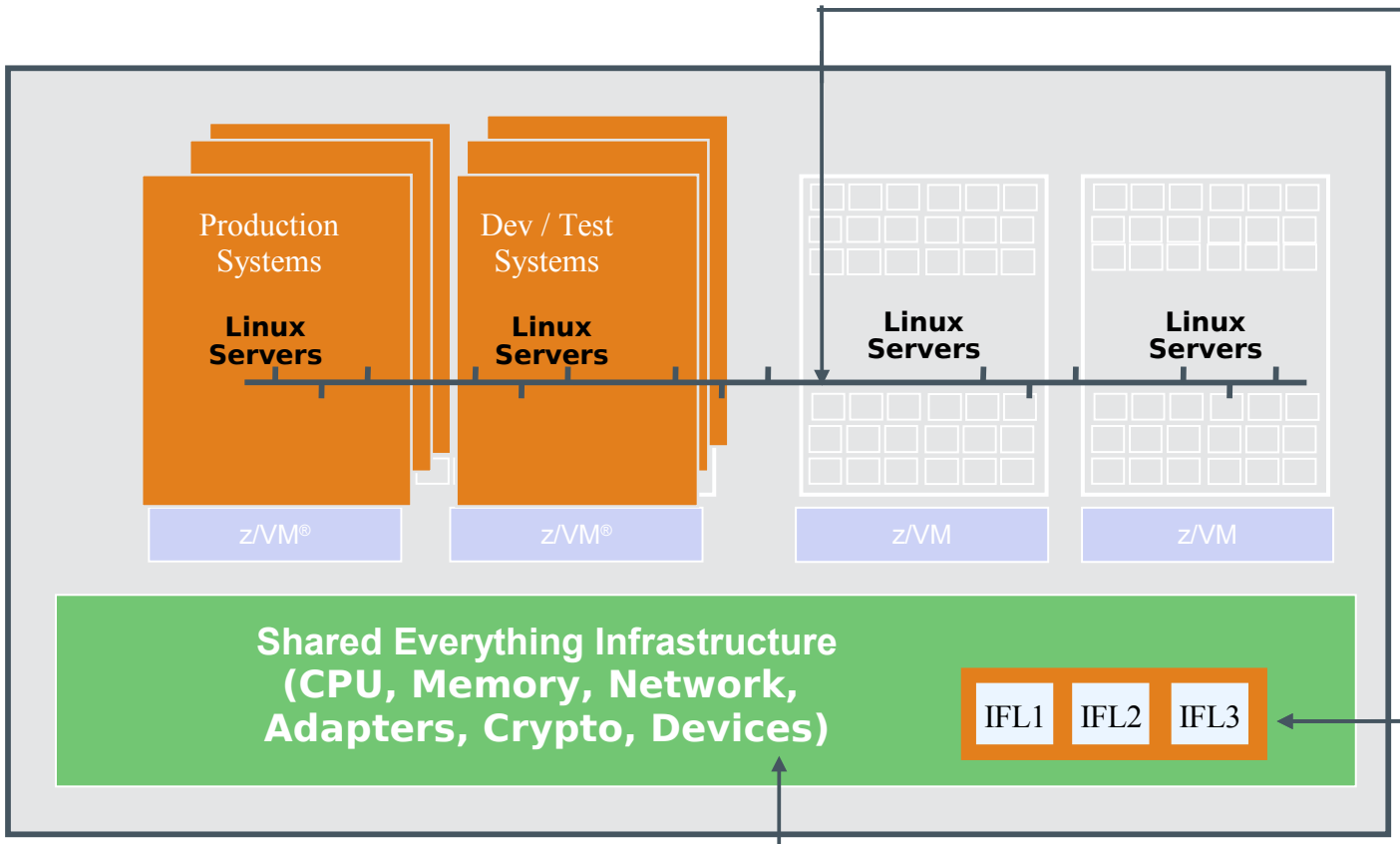

Linux vs. Mainframe Terminology

- **Linux**
  - System administrator
  - Network management
  - Boot
  - 4-processor machine
  - Main memory
  - Disk
  - Scheduler
  - NIC
- **Mainframe**
  - System programmer
  - Systems management
  - IPL
  - 4-way
  - Main storage
  - DASD
  - Dispatcher
  - OSA

While I worked as a Linux consultant for 8 years people said: "You don't look like someone doing Linux" (no long hair and geeky look)

After I joined IBM to work in the mainframe world people said: "You don't look like someone doing mainframe" (some years ago half of today's mainframe pros were at least 50 years old and nearing retirement)

The Linux on System z Server High Level Architecture



"Inside the box"virtual networking

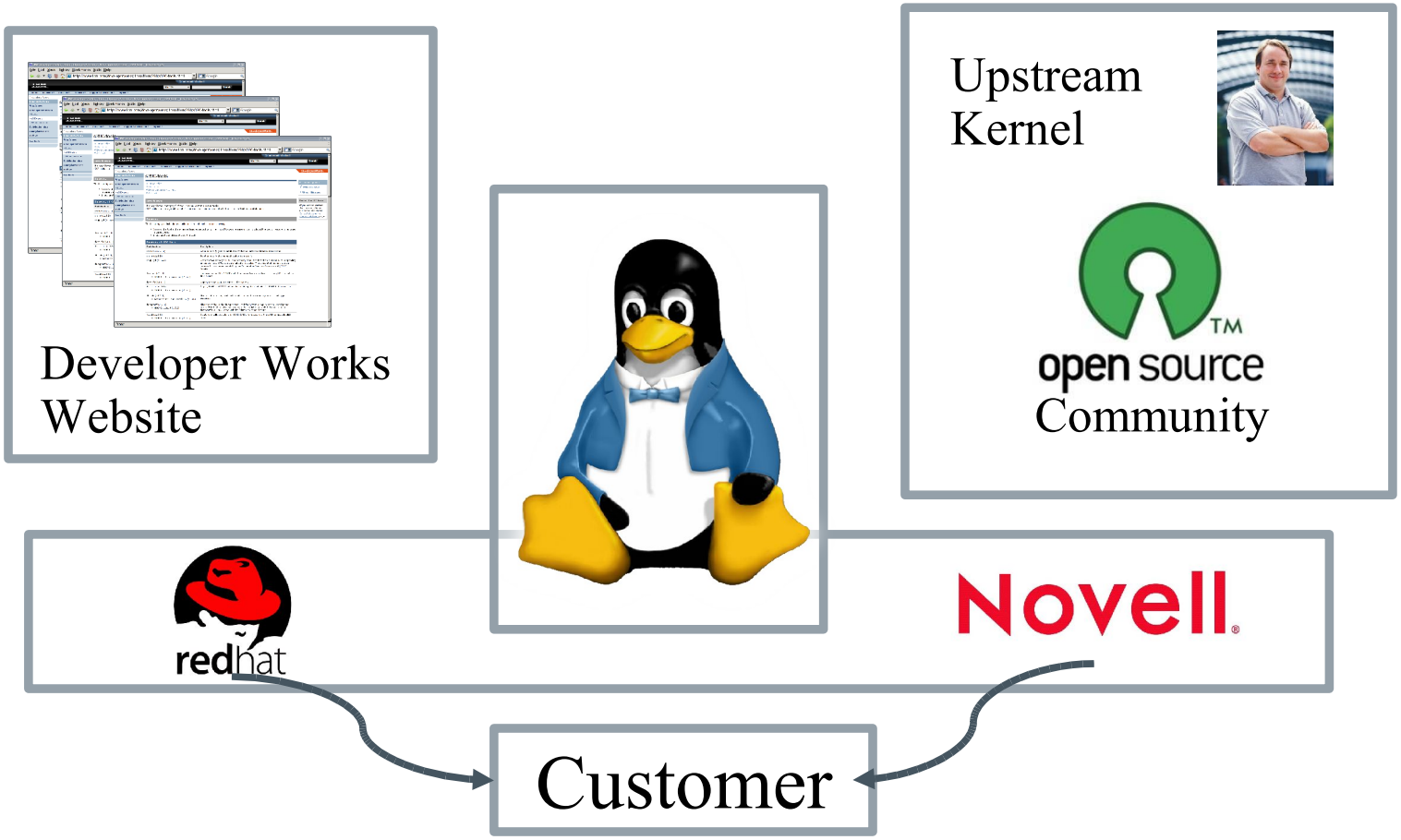
... most Linux ISV software fees are priced on real CPU capacity...

...a potential source of cost savings given z/VM's ability to overcommit CPU capacity

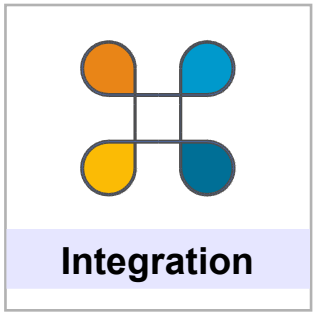
Non-disruptively scale your environment by adding hardware assets that can be shared with *every* virtual server

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



Linux on System z Development Focus



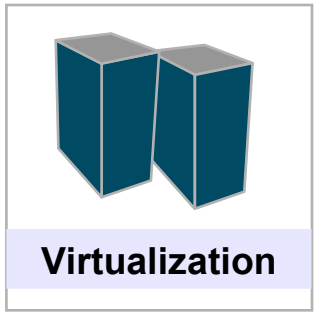
Integration

Application Serving

- z/OS integration

Data Hub

- Database Consolidation



Virtualization

Virtualization &

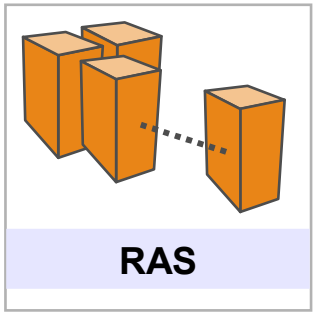
Virtualization Management

- Ease of Use
- Serviceability
- Hosting capacity

Security

Security

- Certifications
- Data security & privacy

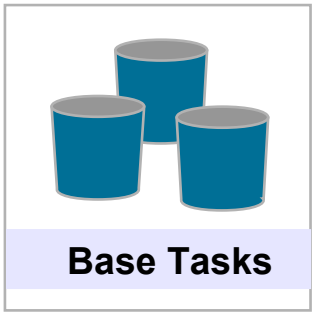


RAS

Continuous Availability

& Data Replication

- RAS
- Differentiation for mission critical workloads



Base Tasks

Customer Requirements

- Address customer observed deficiencies

Competitiveness

- Close competitive gaps
- Differentiation / innovation that matters

Hardware Support

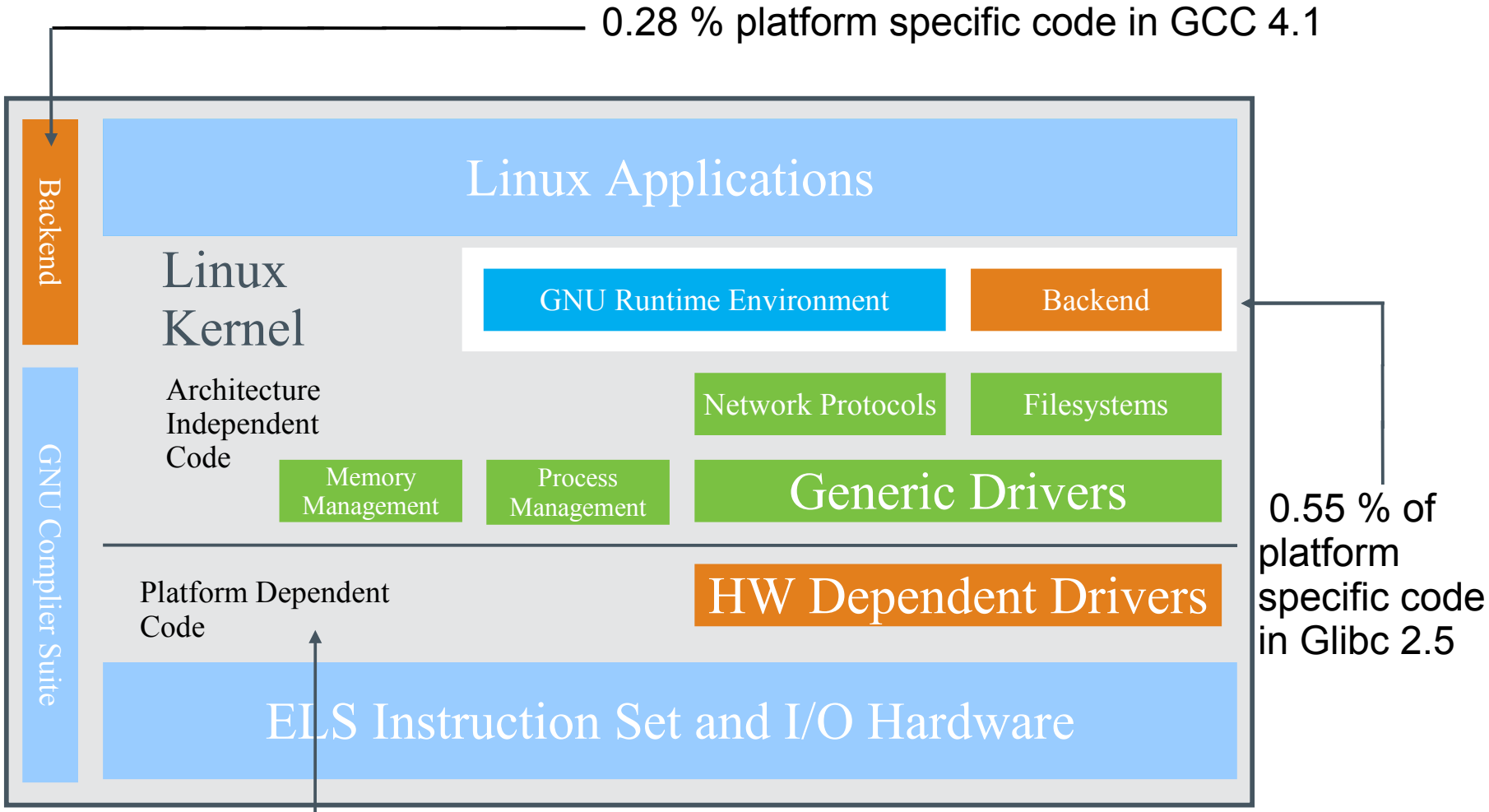
- Exploitation of new System z HW
- Storage exploitation

Linux

- Maintainership & code currency

Structure of Linux on System z

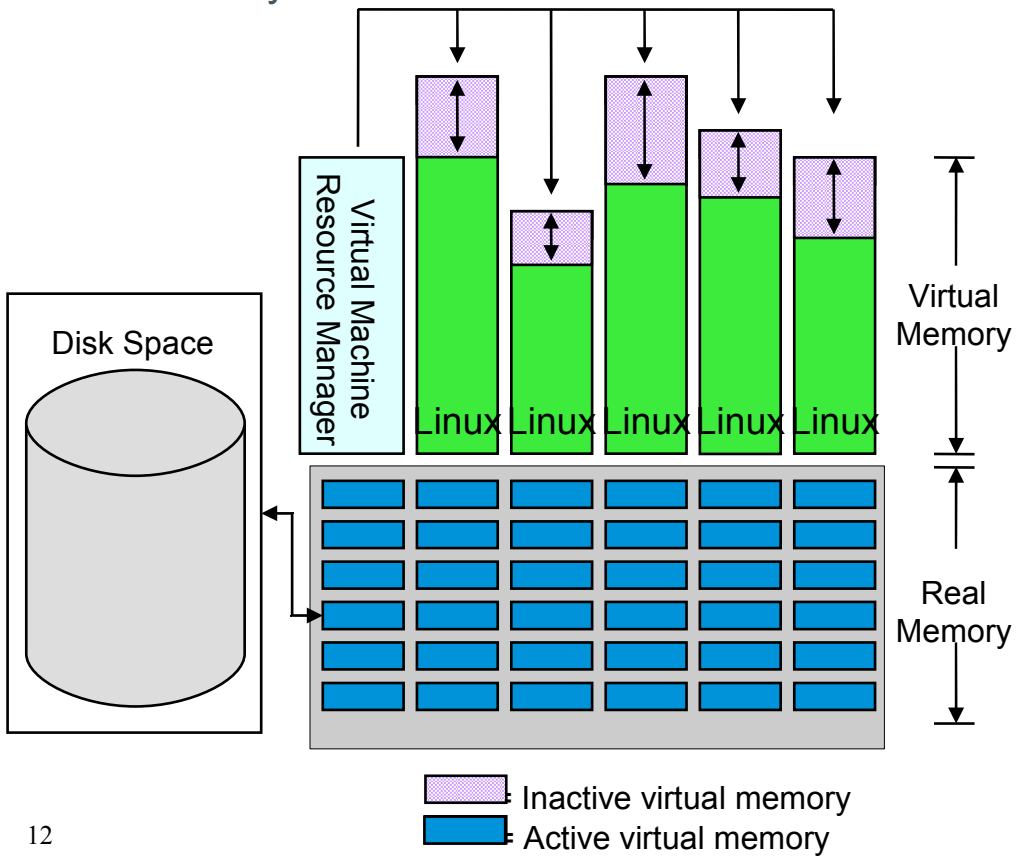
Many Linux software packages did not require any code change to run on Linux on System z

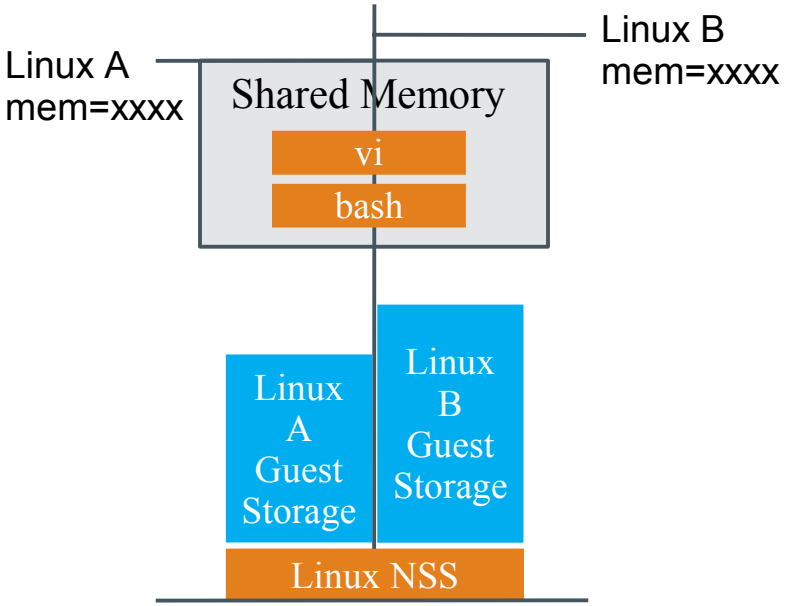


Advanced Virtualization Capabilities

A fundamental strength of ELS running the z/VM hypervisor is its ability to over-commit system resources: "**Do more with less**"

- This can translate into cost savings for hardware *and* software
- Consider a Linux environment with a 25-to-1 overcommitment of CPU capacity
- Memory Overcommitment 1.5-3:1 for Production, 4+:1 for Non Production Systems

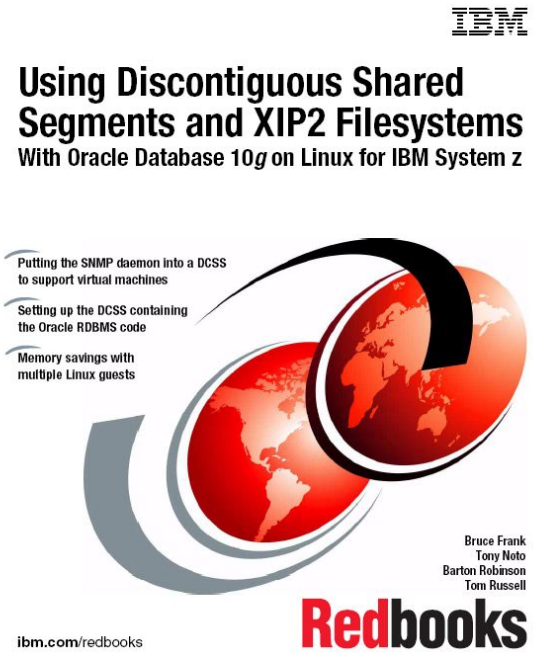


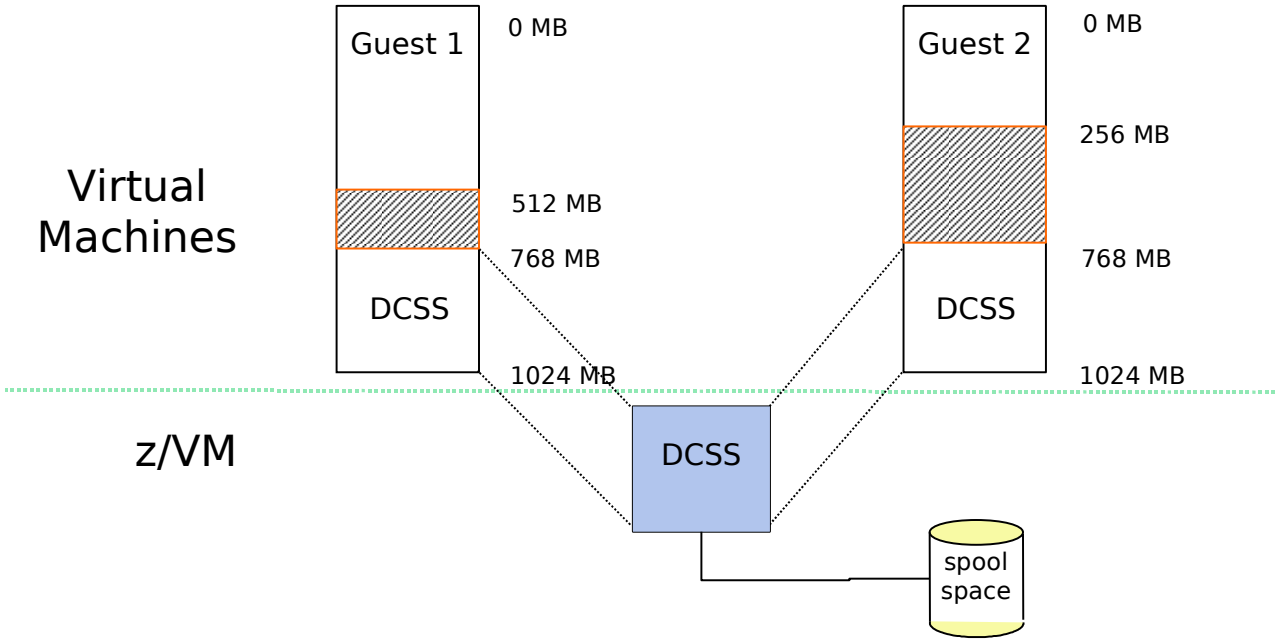


z/VM included advanced memory concepts to save real memory: Cooperative Memory Management & Shared Memory Technologies (NSS, XIP)

Enhancing scalability of Linux on z/VM using execute-in-place technology

- Applications are being executed directly from where they are permanently stored
- Was invented for embedded systems that do not have disk drives
- Applications can be run directly in flash or ROM memory
- Reduces memory consumption





Where's SYS1.PARMLIB Kept?

(This chart is borrowed from Mark Post)

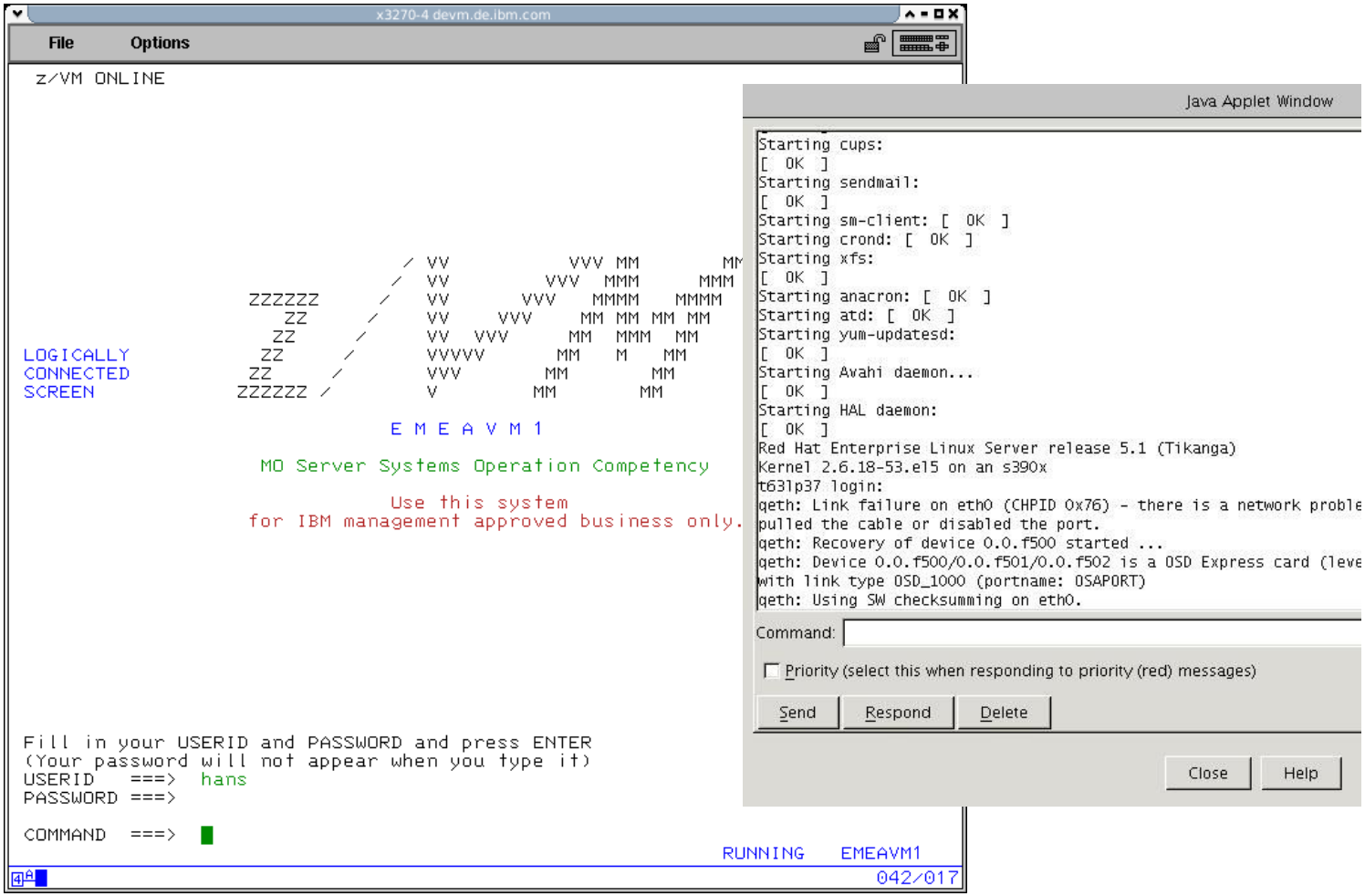
- Just about everything you need is kept under /etc (at some level of hierarchy)
- Individual text files (or groups of them), since no concept of a PDS in Linux.
- Some fairly important ones:
  - /etc/passwd
  - /etc/group
  - /etc/shadow
  - /etc/inittab
  - /etc/fstab
  - ....
- Do I Really Have to Know All This Stuff?
  - No, but shouldn't you?
  - Would you let a junior systems programmer or system operator loose on SYS1.PARMLIB or SYSTEM CONFIG via a GUI?
  - If you really don't want to know what's going on or have a large virtual farm used the graphical tool from Red Hat & Novell

How to discover that you are logged into a Linux System on a different Hardware Architecture?

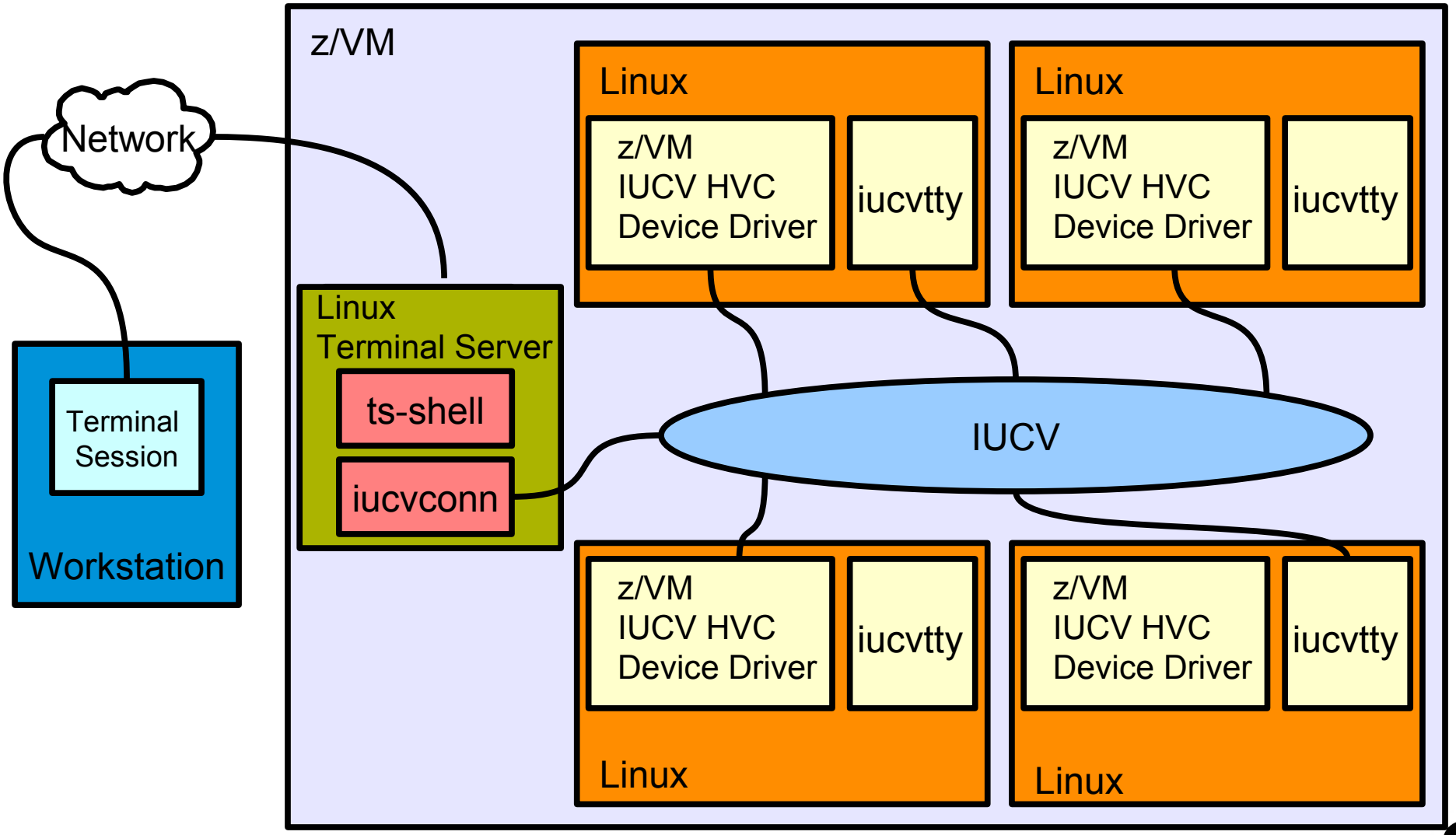
```
hans@tuxmaker:~> cat /proc/cpuinfo
vendor_id       : IBM/S390
# processors    : 8
bogomips per cpu: 3761.76
features    : esan3 zarch stfle msa ldisp eimm dfp edat
processor 0: version = 00,  identification = 28C03F,  machine = 2097
processor 1: version = 00,  identification = 28C03F,  machine = 2097
processor 2: version = 00,  identification = 28C03F,  machine = 2097
processor 3: version = 00,  identification = 28C03F,  machine = 2097
processor 4: version = 00,  identification = 28C03F,  machine = 2097
processor 5: version = 00,  identification = 28C03F,  machine = 2097
processor 6: version = 00,  identification = 28C03F,  machine = 2097
processor 7: version = 00,  identification = 28C03F,  machine = 2097
hans@tuxmaker:~> uname -a
Linux tuxmaker 2.6.16.60-0.42.10-default #1 SMP Tue Apr 27 05:11:27
UTC 2010 s390x s390x s390x GNU/Linux
```


Using a Terminal / Console on System z

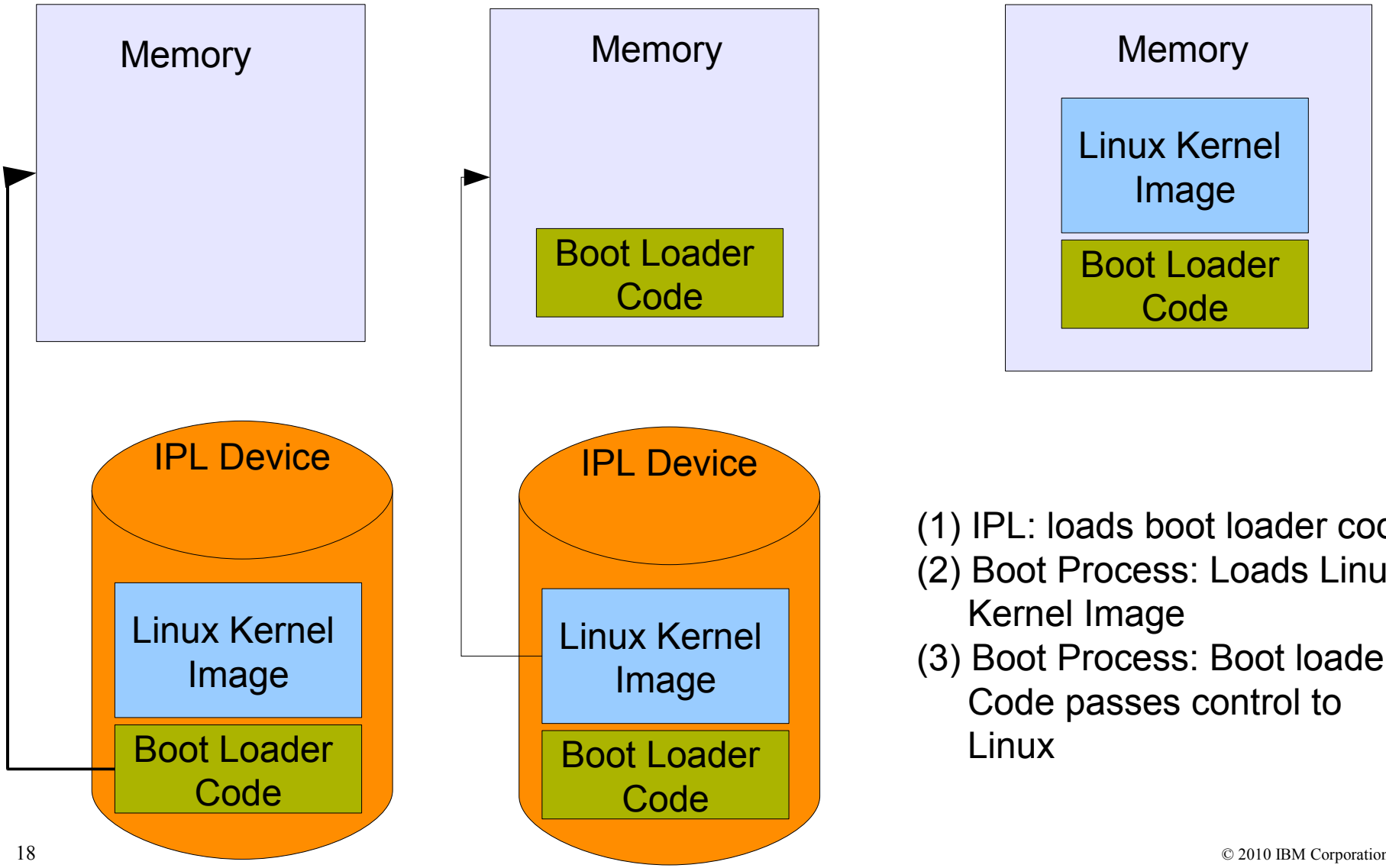
- Line-mode terminal for booting, then SSH connection
  - In case of network/setup problems line-mode terminal may be required
  - ed may safe your system...



IUCV terminal environment



Linux IPL (Initial Program Load)



(1) IPL: loads boot loader code
(2) Boot Process: Loads Linux Kernel Image
(3) Boot Process: Boot loader Code passes control to Linux

The Bootmanager - /etc/zipl.conf

```
[defaultboot]
defaultmenu=menu
[2.6.25]
        image=/boot/vmlinuz-2.6.25
        ramdisk=/boot/initrd-2.6.25.img
        target=/boot/
        parameters="root=/dev/disk/by-path/ccw-0.0.beef TERM=dumb"
[...]
:menu
target = "/boot"
1 = "2.6.25"
2 = "2.6.27"
default = 2
prompt = 1
timeout = 10
```

How device drivers are accessed by Linux

- The Linux kernel represents the character and block devices it knows as a pair of numbers <major>:<minor>.
- Some major numbers are reserved for particular device drivers, others are dynamically assigned to a device driver when Linux boots or the driver is loaded.
- For example, major number 94 is always the major number for DASD devices while the device driver for channel-attached tape devices has no fixed major number.

```
hans@larsson:~> ls -la /dev/dasd*
brw-r----- 1 root disk 94,  0 Sep 14 22:52 /dev/dasda
brw-r----- 1 root disk 94,  1 Sep 14 22:52 /dev/dasda1
brw-r----- 1 root disk 94,  2 Sep 14 22:52 /dev/dasda2
[...]
```

Querying information about the current DASD Setup & Activating a new Disk

Printing a list of active DASD devices (The same information can also be obtained from the file /proc/dasd/devices)

```
hans@larsson:~> lsdasd
Bus-ID     Status      Name      Device  Type  BlkSz  Size      Blocks
==============================================================================
0.0.ec24   active      dasda     94:0    ECKD  4096   7043MB    1803060
```

Activating a new Disk

```
root@larsson:~>  modprobe dasd_mod dasd=ec27
root@larsson:~>  modprobe dasd_eckd_mod
root@larsson:~>  chccwdev -e ec27
Setting device 0.0.ec27 online
Done
root@larsson:~>  dmesg|tail|grep dasd
dasd(eckd): 0.0.ec27: 3390/0C(CU:3990/01) Cyl:10017 Head:15 Sec:224
dasd(eckd): 0.0.ec27: volume analysis returned unformatted disk
```

The device needs to be activated, after the driver Is loaded. Otherwise it can't be used

DASD low level format:

```
root@larsson:~> dasdfmt -d cdl -b 4096 -f /dev/dasdb -p
Drive Geometry: 10017 Cylinders * 15 Heads =  150255 Tracks

I am going to format the device /dev/dasdb in the following way:
   Device number of device : 0xec27
   Labelling device        : yes
   Disk label              : VOL1
   Disk identifier         : 0XEC27
   Extent start (trk no)   : 0
   Extent end (trk no)     : 150254
   Compatible Disk Layout  : yes
   Blocksize               : 4096

--->> ATTENTION! <<---
All data of that device will be lost.
Type "yes" to continue, no will leave the disk untouched: yes
Formatting the device. This may take a while (get yourself a coffee).

cyl   385 of  3339 |#####----------------------------------------|
11%
```


DASD: Partitioning

Compared to other architectures, Linux on System z makes use of its own partitioning tool for DASD devices.
The common Linux tool fdisk can not be used in this environment! Nevertheless the handling is Similar. **The system is limited to 3 partitions per disk when using DASD**

```
root@larsson:~> fdasd /dev/dasdb
reading volume label ..: VOL1
reading vtoc ..........: ok

Command action
   m   print this menu
   p   print the partition table
   n   add a new partition
   d   delete a partition
   v   change volume serial
   t   change partition type
   r   re-create VTOC and delete all partitions
   u   re-create VTOC re-using existing partition sizes
   s   show mapping (partition number - data set name)
   q   quit without saving changes
   w   write table to disk and exit

Command (m for help):
```


Filesystem

Now we have a new device partition (e.g. /dev/dasdb1) which can be used as any other Linux Disk

```
root@larsson:~> mke2fs -j /dev/dasdb1
mke2fs 1.41.4 (27-Jan-2009)
[...]
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 28 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

For a static setup the driver has to be loaded using the bus id during system startup

```
root@larsson:~> echo “options dasd_mod dasd=ec27” >>
/etc/modprobe.conf
```


Miscellaneous

In case of debugging performance problems:

```
root@larsson:~> echo “set on” > /proc/dasd/statistics
root@larsson:~> cat /proc/dasd/statistics
```

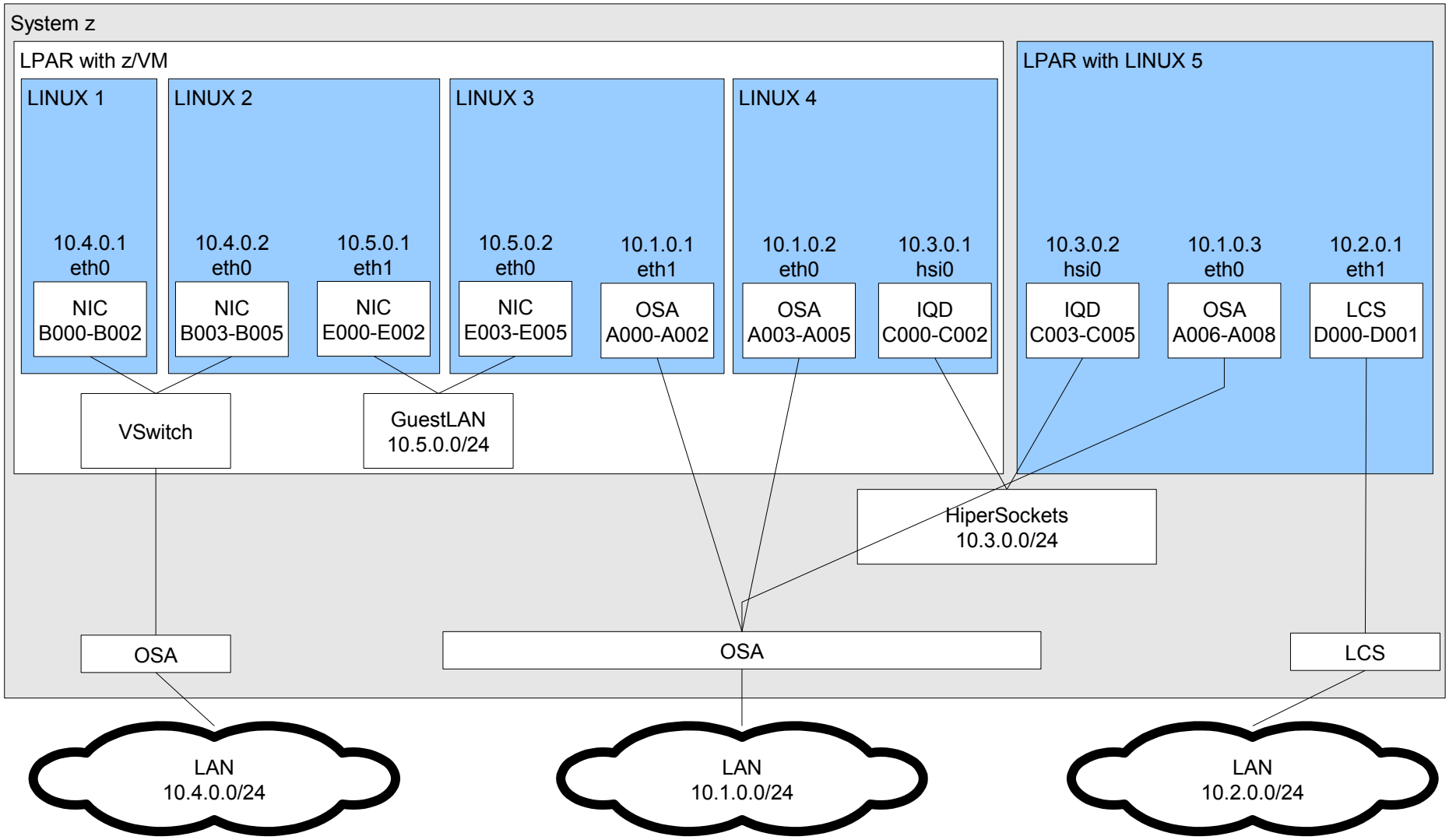
I highly recommend the use of a technology like Logical Volume Manager: Why?

- You can combine several hard disks or partitions
- You can enlarge a logical volume when free space is exhausted
- You can add hard disks to the volume group in a running system
- You can add logical volumes in a running system
- You can use several hard disks with improved performance in the RAID 0 (striping) mode
- You can add up to 256 logical volumes
- The Snapshot feature enables consistent backups

Benefits for Linux on System z

Minidisks on z/VM cannot span more than one physical DASD volume.
Without a volume management system like LVM the size of a file system is limited to the size of a DASD volume.

Networking Example



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
```


dbginfo.sh

dbginfo.sh is a script to collect various system related files, for debugging purposes.

- It generates a tar-archive which can be attached to PMRs / Bugzilla entries
- It is similar to the RedHat tools sosreport / Novell supportutils

```
root@larsson:~>  dbginfo.sh
Create target directory /tmp/DBGINFO-2010-04-25-22-06-20-
t6345057
Change to target directory /tmp/DBGINFO-2010-04-25-22-06-
20-t6345057
Get procfs entries
Saving runtime information into runtime.out
Get file list of /sys
Get entries of /sys
[...]
```

Please use the data from this tool is you open a Bugzilla (Novell/ RedHat) or a PMR !!!

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
```

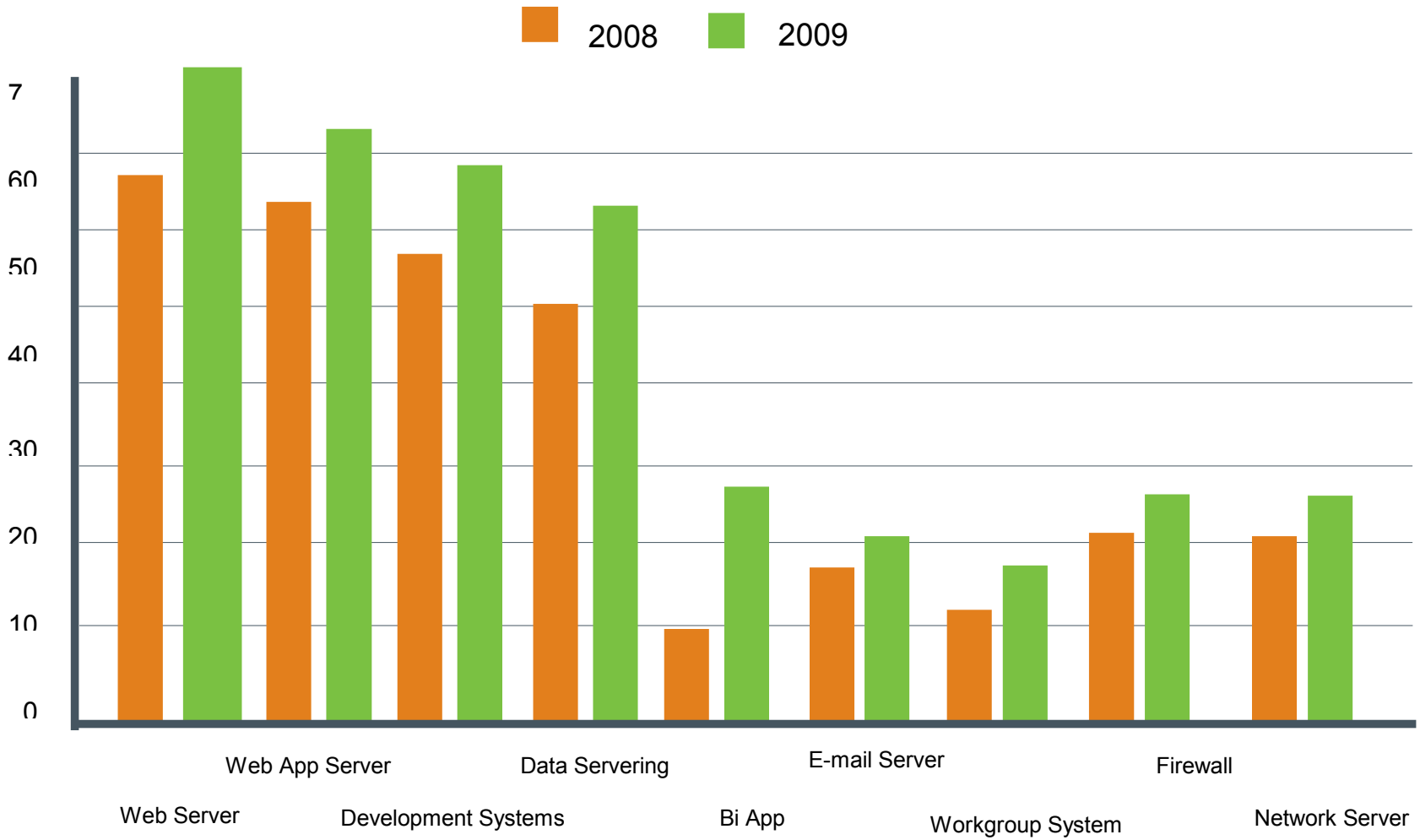
Be careful, when executing disruptive commands!

Workload share on utilized IFLs
Primary applications in the past

60%	Application serving for z/OS e.g. WebSphere, SAP, CICS TG, DB2 Connect
30%	Data serving e.g. Oracle DB, DB2 UDB
5%	Workplace serving e.g. Domino, Scalix, other e-mail
5%	Infrastructure serving e.g. Apache, Samba, NFS, etc.
<1%	Linux application development/deployment

Notes: extrapolation based on analyzing 1/3 of inventory, excludes all IBM

What are Clients Consolidating to Linux on System z?



A “typical” customer: ACME Inc.

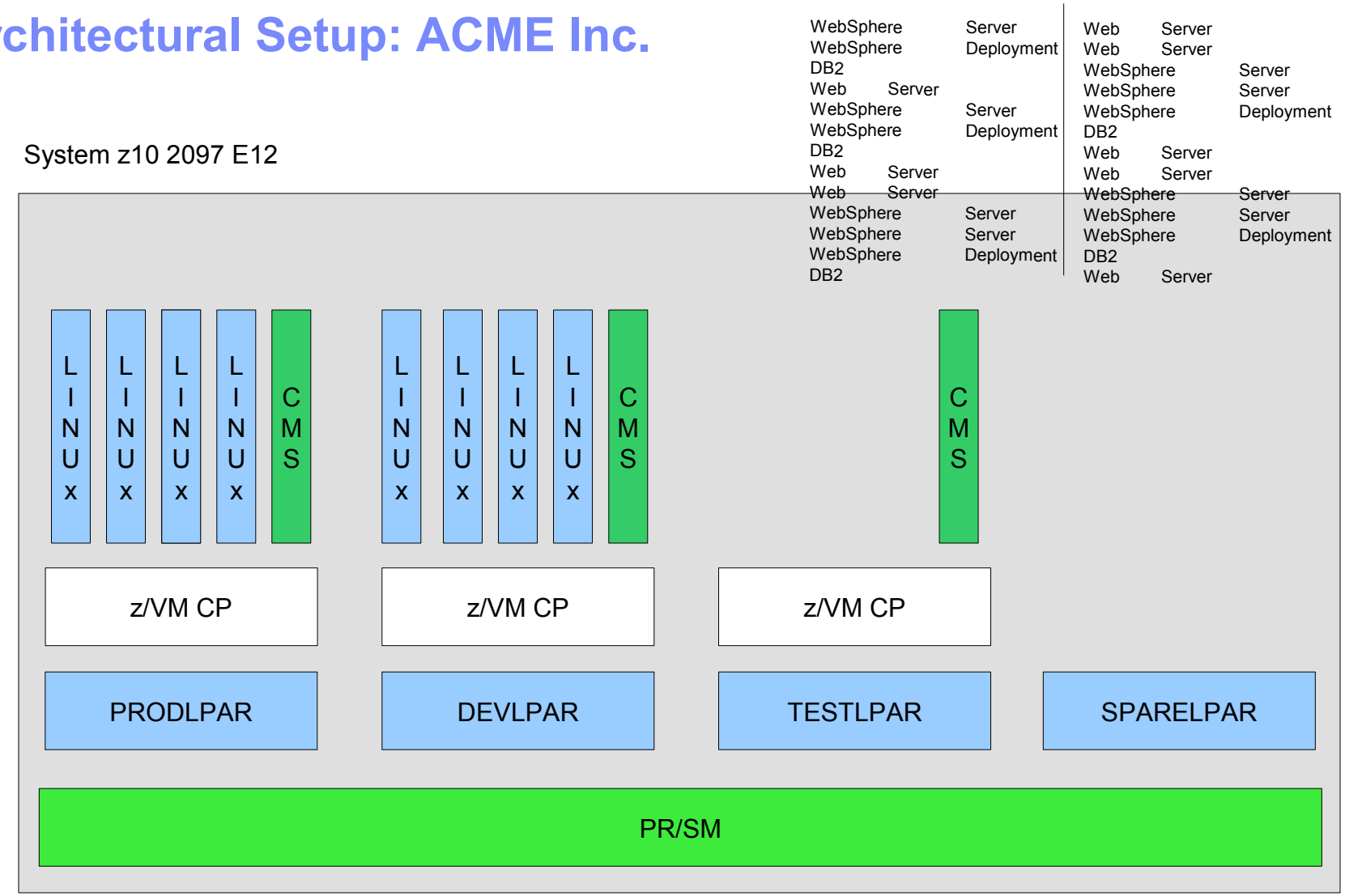
- During the second half of 2008 ACME Inc. purchased an IBM System z mainframe to act as a server consolidation platform.
- Hardware (excerpt)
  - IBM System z10 Enterprise Class
    - Model: 2097-E12
    - 96GB memory
    - 3 Integrated Facility for Linux (IFL) CPU's
  - IBM System Storage DS6800 Disk
    - Model: 1750 522
    - Parallel Access Volume (PAV) licence.
- In z10 has been configured with 4 LPARs: Production, Development, Software and one reserved for future use.
- The system is going to be used as a server consolidation platform.
  - Multiple WebSphere servers running on Intel machines will be consolidated to
  - Linux servers running as virtualized guests hosted by the z/VM operating system.
  - Each LPAR will run a z/VM 5.4 operating system.

Linux & z/VM environment at ACME Inc.

- Excerpt from the customer documentation:
- In order to build a new Linux guest virtual machine a number of steps need to be completed.
  - Identify DASD required for the new Linux guest
  - Edit the CP User Directory file – adding definitions for the new guest
  - Complete RACF changes for the new guest
  - Grant access to the relevant VSWITCH
  - Edit the Kickstart files
  - Run the Kickstart process to install the Linux system

- In addition to the base **z/VM** operating system ACME Inc. has licensed a number of z/VM products.
  - Performance Toolkit
  - Omegamon XE on z/VM and Linux
  - Operations Manager for z/VM
  - Backup/Restore Manager for z/VM
  - Tape Manager for z/VM
  - Hardware Configuration
  - Definition (HCD)/Hardware Configuration Manager/(HCM)
  - RACF
  - DFSMS/VM

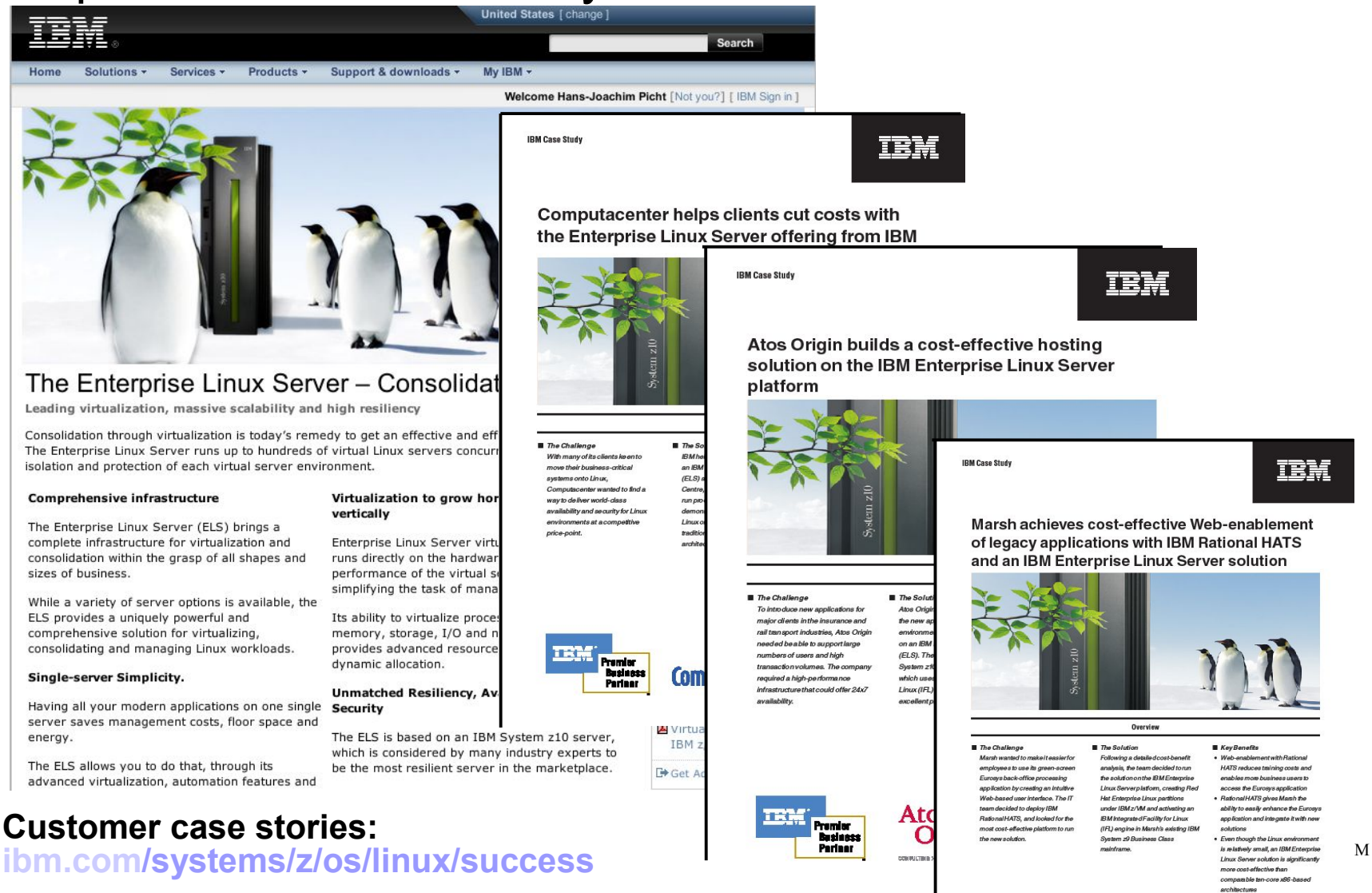
Architectural Setup: ACME Inc.



For the initial implementation each LPAR has been given access to 3 shared IFL's i.e. no dedicated IPL's have been configured
IFL Weight: 70% Production, 10 % Development, 10 % Test, 10 % Spare LPAR

More Information

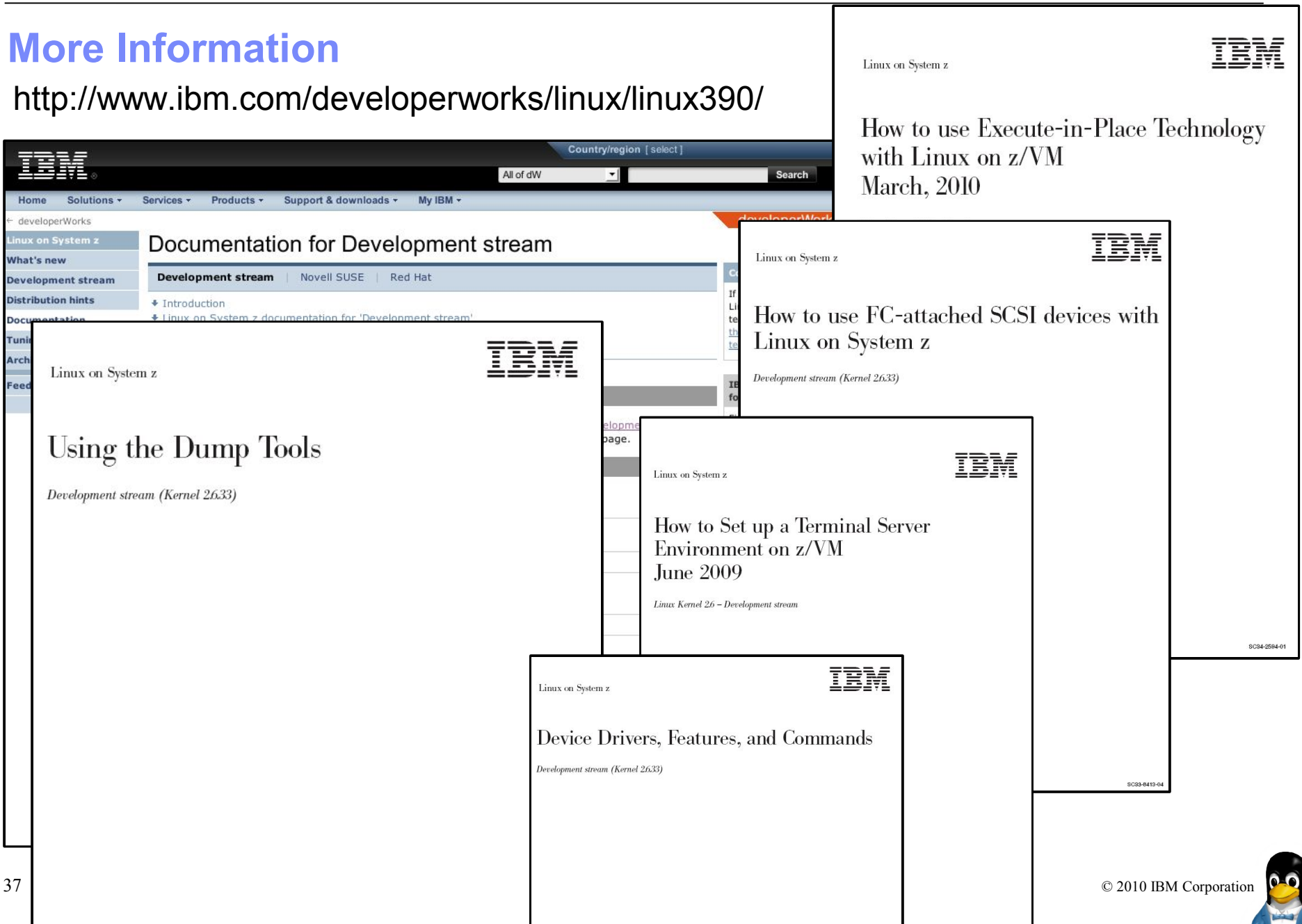
http://www-03.ibm.com/systems/z/os/linux/els.html



Customer case stories:
ibm.com/systems/z/os/linux/success

More Information

http://www.ibm.com/developerworks/linux/linux390/



Questions?

Hans-Joachim Picht

Linux on System z Initiatives

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

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

How to explain the benefits of running Linux on System z in 2:39?

http://www.youtube.com/watch?v=0i7kBnhN3Lg

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

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:
IBM, the IBM logo, BladeCenter, Calibrated Vectored Cooling, ClusterProven, Cool Blue, POWER, PowerExecutive, Predictive Failure Analysis, ServerProven, System p, System Storage, System x , System z, WebSphere, DB2 and Tivoli are trademarks of IBM Corporation in the United States and/or other countries. For a list of additional IBM trademarks, please see http://ibm.com/legal/copytrade.shtml.

The following are trademarks or registered trademarks of other companies: Java and all Java based trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries or both Microsoft, Windows,Windows NT and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both. Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.UNIX is a registered trademark of The Open Group in the United States and other countries or Both. Linux is a trademark of Linus Torvalds in the United States, other countries, or both. Cell Broadband Engine is a trademark of Sony Computer Entertainment Inc. InfiniBand is a trademark of the InfiniBand Trade Association.Other company, product, or service names may be trademarks or service marks of others.
NOTES: Linux penguin image courtesy of Larry Ewing (lewing@isc.tamu.edu) and The GIMP
Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development- level systems. There is no guarantee these measurements will be the same on generally-available systems. Users of this document should verify the applicable data for their specific environment. IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
Information is provided “AS IS” without warranty of any kind. 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