IBM Linux and Technology Center

# What's New in Linux on System z

Martin Schwidefsky
IBM Lab Böblingen, Germany
August 2nd, 2010 – Session 6923

# 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 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 are suggested US list prices and are subject to change without notice. Starting price may not include a hard drive, operating system or other features. Contact your IBM representative or Business Partner for the most current pricing in your geography. 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

# 10 YEARS of Enterprise Linux on System z®

# A simple idea: 10 years ago

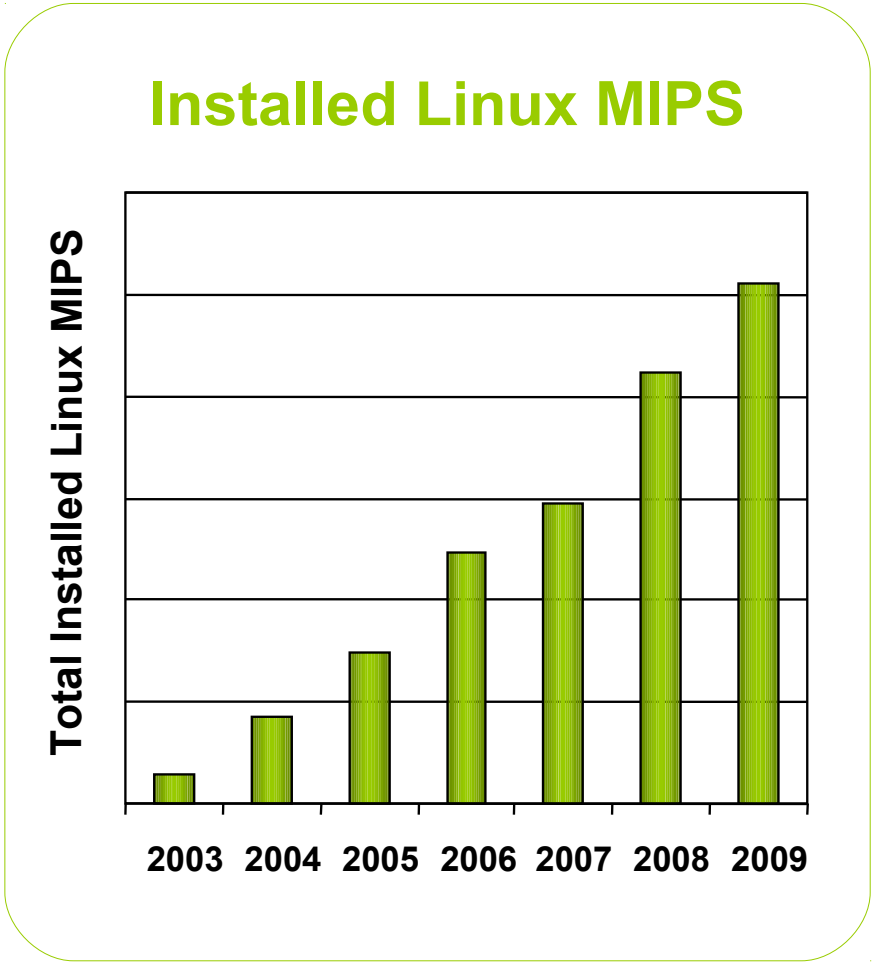
- **Why not try to run Linux on a mainframe?**
- **Increased solutions through Linux application portfolio**
- **Large number of highly skilled programmers familiar with**
- **Integrated business solutions**
  - Data richness from IBM eServer™ zSeries®
  - Web capability of Linux applications
- **Industrial strength environment**
  - Flexibility and openness of Linux
  - Qualities of server of zSeries and S/390®
- **Unique ability to easily consolidate large number of servers**

# A simple question

- **Will Linux be accepted by clients on the world's most secure server platform?**
  - Free Software Foundation UNIX®-like operating system
  - GNU - GNU is Not UNIX
  - GPL - GNU General Public License
- **December 1999: IBM published a collection of patches / additions to the Linux 2.2.13 kernel for System/390® to start a market evaluation**

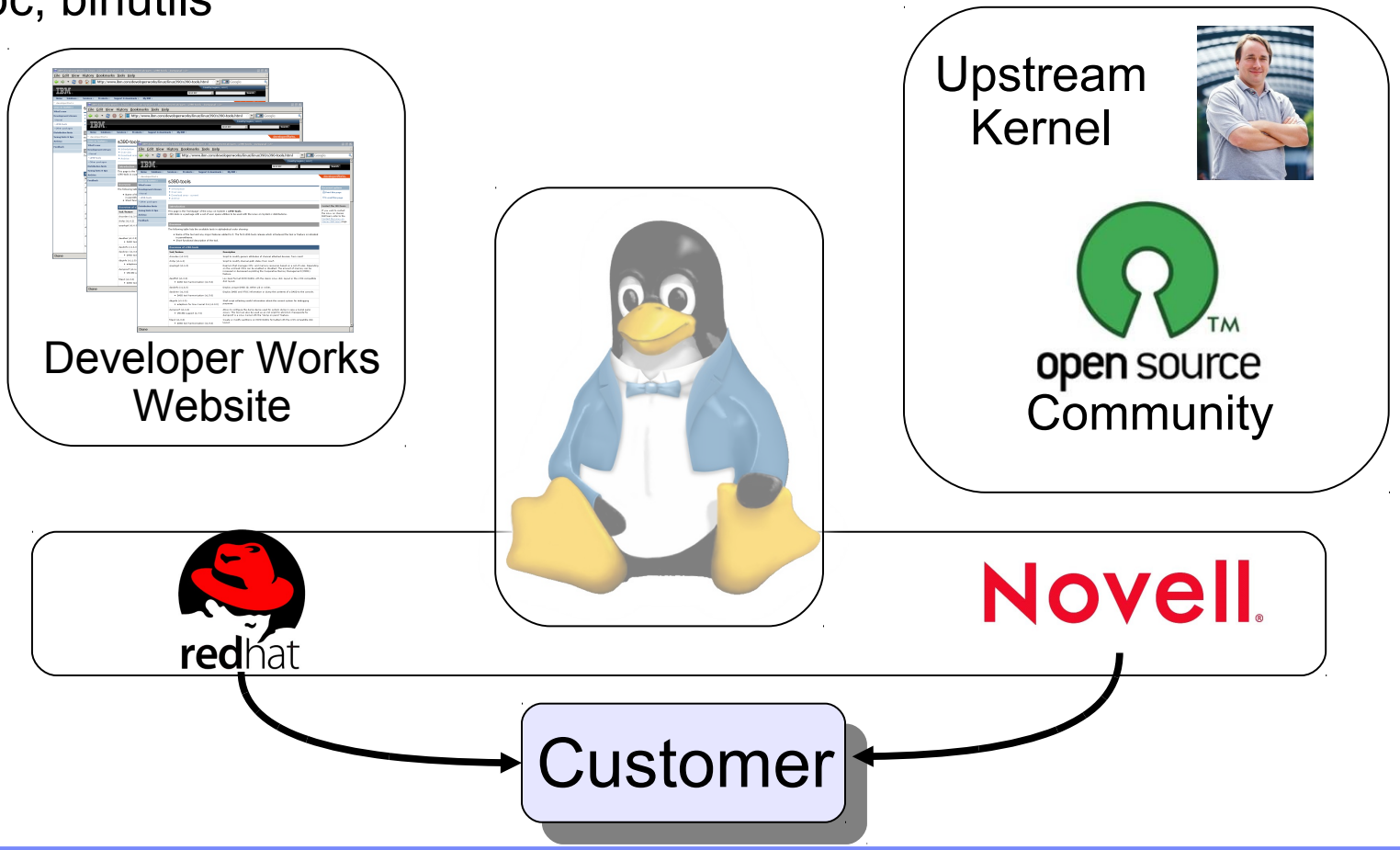
# Client adoption continues to drive Linux success

- **The momentum continues:**
  - Shipped IFL engine volumes increased 35% from YE07 to YE09
  - Shipped IFL MIPS increased 65% from YE07 to YE09
- **Linux is 16% of the System z customer install base (MIPS)**
- **70% of the top 100 System z clients are running Linux on the mainframe**
- **More than 3,100 applications are available for Linux on Sytem z**



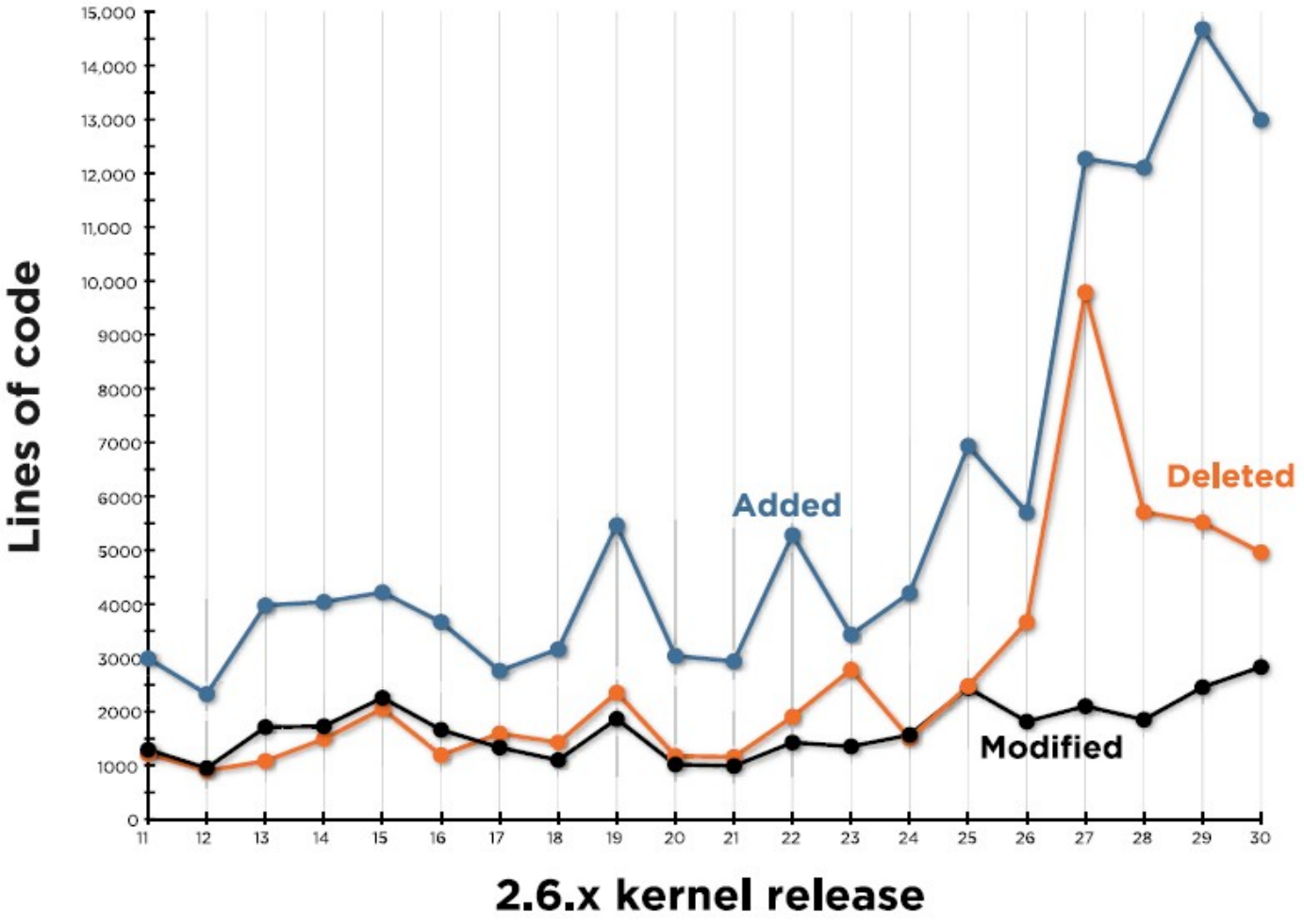
# The IBM Linux development process

- IBM Linux on System z development contribues in the following areas: Kernel, s390-tools, open source tools (e.g. eclipse, ooprofile), gcc, glibc, binutils



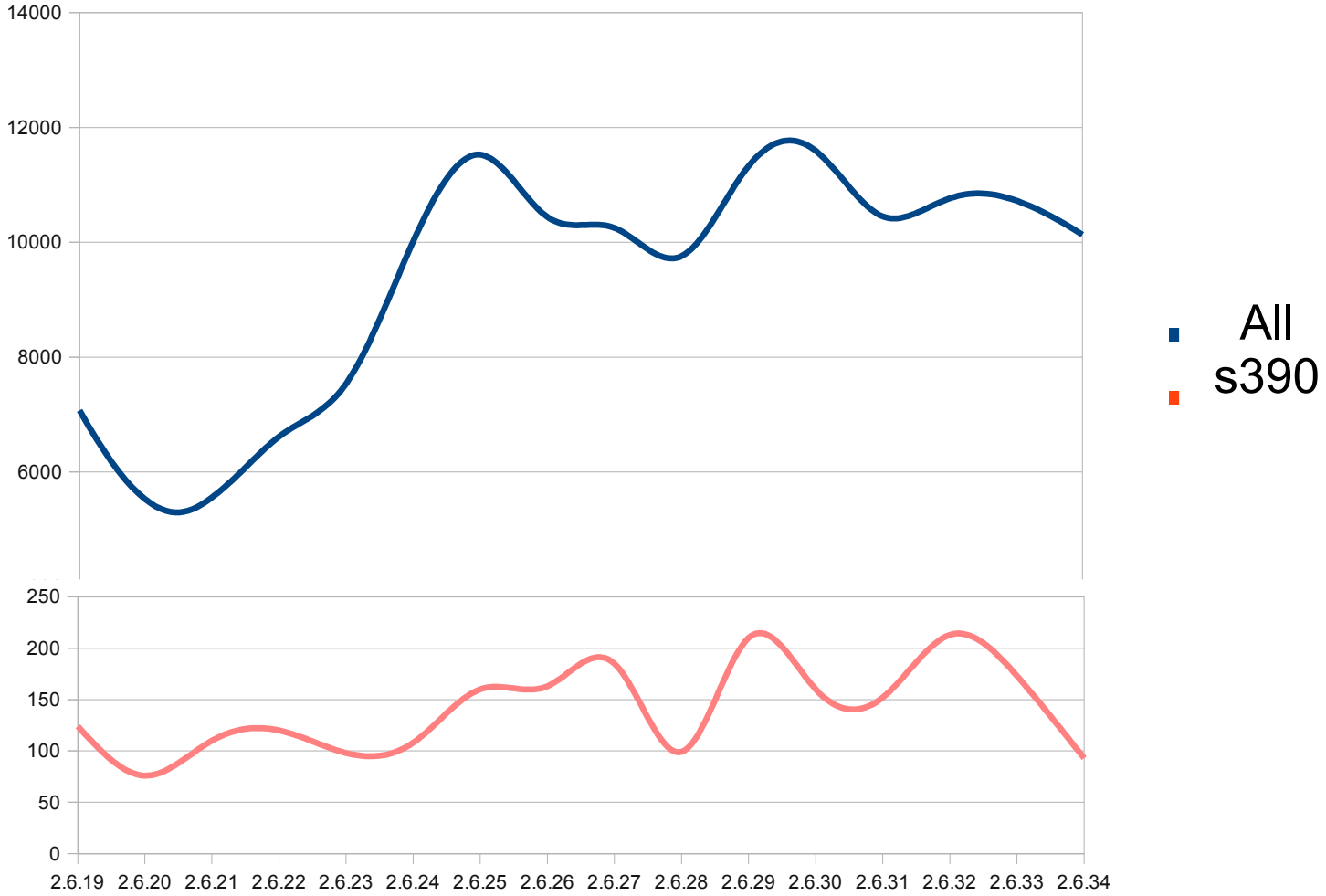
# Linux kernel development: rate of change

Average: 6,422 lines added, 3,285 lines removed, and 1,687 lines changed every day for the past 4 1/2 years.



# Linux kernel development: System z contributions

- Changesets per release



# Linux on System z distributions (Kernel 2.6 based)

- **SUSE Linux Enterprise Server 9 (GA 08/2004)**
  - Kernel 2.6.5, GCC 3.3.3, Service Pack 4 (GA 12/2007)
- **SUSE Linux Enterprise Server 10 (GA 07/2006)**
  - Kernel 2.6.16, GCC 4.1.0, Service Pack 3 (GA 09/2009)
- **SUSE Linux Enterprise Server 11 (GA 03/2009)**
  - Kernel 2.6.27, GCC 4.3.3, Service Pack 1 (GA 06/2010), Kernel 2.6.32
- **Red Hat Enterprise Linux AS 4 (GA 02/2005)**
  - Kernel 2.6.9, GCC 3.4.3, Update 8 (GA 05/2009)
- **Red Hat Enterprise Linux AS 5 (GA 03/2007)**
  - Kernel 2.6.18, GCC 4.1.0, Update 5 (GA 03/2010)
- **Others**
  - Debian, Slackware,
  - Support may be available by some third party

# Supported Linux Distributions

| Distribution | zEnterprise 196 | System z10 | System z9 | zSeries |
|---|---|---|---|---|
| RHEL 5 | ✔ | ✔ | ✔ | ✔ |
| RHEL 4 (1) | — | ✔ | ✔ | ✔ |
| RHEL 3 (1) | — | — | * | ✔ |
| SLES 11 | ✔ | ✔ | ✔ | ✘ |
| SLES 10 | ✔ | ✔ | ✔ | ✔ |
| SLES 9 (1) | — | ✔ | ✔ | ✔ |

(1) Also available as 31-bit distributions.

✔ Indicates that the distribution (version) has been tested by IBM on the hardware platform, will run on the system, and is an IBM supported environment. Updates or service packs applied to the distribution are also supported.

✘ Indicates that the distribution is not supported by IBM.

— Indicates that the distribution has not been tested by IBM.

* Supported on customer request (RPQ).

# Kernel news – Common code

**Linux version 2.6.31 (2009-08-03)**

- Performance counters
- Ftrace function tracer extensions
- Per partition blktrace

**Linux version 2.6.32 (2009-08-03)**

- Per-backing-device based writeback (pdflush replaced by flush-<major>)
- Kernel Samepage Merging (memory deduplication)
- CFQ I/O scheduler low latency mode
- S+core architecture support

**Linux version 2.6.33 (2010-02-24)**

- DRDB (Distributed Replicated Block Device)
- TCP Cookie Transactions for DNSSEC protocol
- Swappable KSM pages
- Compcache: memory compressed swapping

# Kernel news – Common code

**Linux version 2.6.34 (2010-05-16)**

- Ceph distributed network file system
- LogFS flash memory file system
- Asynchronous suspend / resume
- 'Perf' performance analysis improvements, cross architecture support

**Linux version 2.6.35-rc5 (2010-07-12)**

- BKL removal work (incomplete)
- XFS delayed logging, allows to accumulate multiple transaction
- Support for multiple multicast route tables
- Support for Layer 2 Tunneling Protocol L2TP Version 3

# System z kernel features – z/VM

- **Extra kernel parameter for SCSI IPL (kernel 2.6.32)**
  - Modify the SCSI loader to append append extra parameters specified with the z/VM VMPARM option to the kernel command line.
- **Deliver z/VM CP special messages as uevent (kernel 2.6.34)**
  - Allows to forward SMSG messages starting with “APP” to user space.
  - udev rules can be used to trigger application specific actions
- **Automatic detection of read-only devices (2.6.34)**
  - Improve usability by automatically detection of read-only dasd devices with diagnose 210
- **CMSFS user space file system support (s390-tools 1.9.0 for the read-only cmsfs support)**
  - Implement a FUSE file system that allows to read from and write to CMSFS minidisks.
  - Writing is difficult, the record based CMSFS does not fit well into the byte steam oriented Linux VFS

# CMSFS user space file system support

- **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**

```
# cmsfs-fuse /dev/dasde /mnt/cms
# ls -la /mnt/fuse/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

```
# cmsfs-fuse -t /dev/dasde /mnt/cms
```

- **Write support is work in progress, almost completed**
  - use “vi” to edit PROFILE.EXEC anyone ?
- **Use fusermount to unmount the file system again**

```
# fusermount -u /mnt/cms
```

# System z kernel features – Storage

- **FCP adjustable queue depth (kernel 2.6.31)**
  - Customizable queue depth for SCSI commands in zfcp
- **Resume reordered devices (kernel 2.6.34)**
  - Allow resume of a guest with different subchannels for individual devices
  - Allow suspend of a system with devices in the disconnected state
- **Unit check handling (kernel 2.6.35)**
  - Improve handling of unit checks for internal I/O started by the common-I/O layer
  - After a unit check certain setup steps need to be repeated, e.g. for PAV
- **Store I/O status and initiate logging (SIOSL) (> kernel 2.6.35)**
  - Enhance debug capability for FCP attached devices
  - Enables operating system to detect unusual conditions on a device of channel path

# System z kernel features – Storage

- **E2E data consistency checking (> kernel 2.6.35)**
  - Use checksum of SCSI data payload to check end-to-end consistentcy
  - Needs common code changes to file systems
- **Automatic LUN scanning (> kernel 2.6.35)**
  - Scan and attach accessible LUNs automatically
  - Available only for a NPIV FCP attachment
- **Automatic menu support in zipl (> s390-tools 1.9.0)**
  - Zipl option that will create a boot menu for all eligible non-menu sections in the zipl configuration file
- **reIPL from device-mapper devices (> s390-tools 1.9.0)**
  - The automatic re-IPL function only works with a physical device
  - Enhance the zipl support for device-mapper devices to provide the name of the physical device if the zipl target is located on a logical device

# System z kernel features – Networking

- **OSA QDIO Data Connection Isolation (kernel 2.6.33)**
  - Isolate data traffic from Linux on System z guests sharing an OSA card
  - Communication between guests needs to go over via external entity
- **HiperSockets Network Traffic Analyser (kernel 2.6.34)**
  - Trace HiperSockets network traffic for problem isolation and resolution.
  - Supported for layer 2 and layer 3
- **Offload outbound checksumming (kernel 2.6.35)**
  - Move calculation of checksum for non-TSO packets from the driver to the OSA network card
- **Toleration of optimized latency mode (kernel 2.6.35)**
  - OSA devices in optimized latency mode can only serve a small number of stacks / users. Print a helpful error message if the user limit is reached.
  - Linux does not exploit the optimized latency mode

# System z kernel features – Networking

- **NAPI support for QDIO and QETH (> kernel 2.6.35)**
  - Convert QETH to the NAPI interface, the “new” Linux networking API
  - NAPI allows for transparent GRO (generic receive offload)
- **QETH debugging per single card (> kernel 2.6.35)**
  - Split some of the global QETH debug areas into separate per-device areas
  - Simplifies debugging for complex multi-homed configurations
- **Configuration tool for System z network devices (s390-tools 1.8.4)**
  - Provide a shell script to ease configuration of System z network devices

# znetconf network device configuration tool

- **Allows to list, add, remove & configure System z network devices**
- **For example: list all potential network devices:**

```
# znetconf -u
Device Ids                 Type    Card Type  CHPID Drv.
------------------------------------------------------------
0.0.f500,0.0.f501,0.0.f502 1731/01 OSA (QDIO) 00    qeth
0.0.f503,0.0.f504,0.0.f505 1731/01 OSA (QDIO) 01    qeth
```

- **Configure device 0.0.f503**

```
znetconf -a 0.0.f503
```

- **Configure device 0.0.f503 in layer2 mode and portname “myport”**

```
znetconf -a 0.0.f503 -o layer2=1 -o portname=myport
```

- **Remove network device 0.0.f503**

```
znetconf -r 0.0.f503
```

# System z kernel features – Usability / RAS

- **Suspend / resume support (kernel 2.6.31)**
  - Add the ability to stop a running Linux system and resume operations later on. The image is stored on the swap device and does not use any system resource while suspended.
  - Only suspend to disk is implemented, suspend to RAM is not supported.
- **Add Call Home data on halt and panic if running in LPAR (kernel 2.6.32)**
  - Report system failures (kernel panic) via the service element to the IBM service organization.
  - Improves service for customers with a corresponding service contract.
  - Opt-in, by default this features is deactivated.
- **Trigger SCSI dump on remove container (snipl 2.1.9)**
  - Extend the snIPL tool to force a dump to a SCSI device on a remove z/VM guest of LPAR

# Suspend / resume support

- **Ability to stop a running Linux on System z instance and later continue operations**
- **Memory image is stored on the swap device specified with a kernel parameter: `resume=/dev/dasd<x>`**
- **Lower the swap device priority for the resume partition**

```
# cat /etc/fstab
...
/dev/dasdb1 swap swap pri=-1 0 0
/dev/dasdc1 swap swap pri=-2 0 0
```

- **Suspend operation is started with a simple echo:**

```
# echo disk > /sys/power/state
```

- **Resume is done automatically on next IPL**
- **Use signal quiesce to automatically suspend a guest**

```
ca::ctrlaltdel:/bin/sh -c "/bin/echo disk > /sys/power/state || \
                           /sbin/shutdown -t3 -h now"
```

# System z kernel features – Usability / RAS

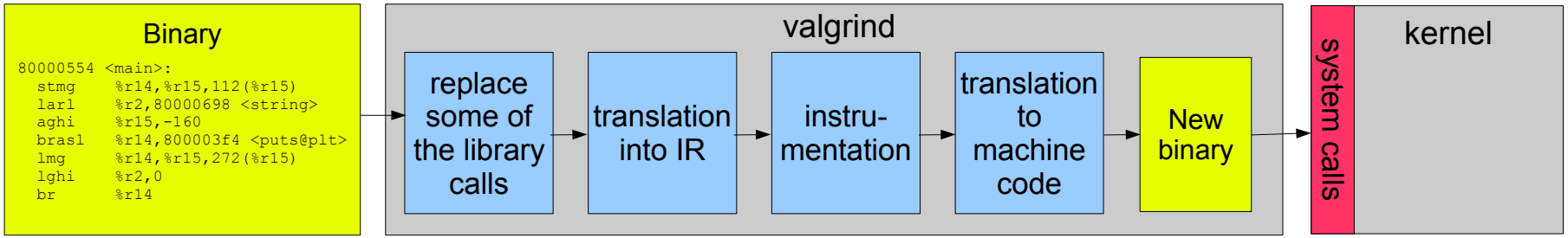
- **Dump on panic – prevent reipl loop (s390-tools 1.8.4)**
  - Delay arming of automatic reipl after dump
  - Avoids dump loops where the restarted system crashes immediately
- **Add support for makedumpfile tool (kernel 2.6.34, s390-tools 1.9.0)**
  - Convert Linux dumps to the ELF file format
  - Use the makedumpfile tool to remove user data from the dump
  - Multi-volume tape dump will be removed
- **Breaking event address for user space (kernel 2.6.35)**
  - Store the breaking-event-address for user space programs
  - Valuable aid in the analysis of wild branches
- **Precise process accounting (> kernel 2.6.36)**
  - Extend the taskstats interface to provide better process accounting values
  - Quality goal is a resolution of 10ths of microseconds

# System z toolchain

- **64 bit register in 31 bit compat mode**
  - Make use of 64 bit registers in 31 bit application running in z/Architecture mode.
  - Allows to use instruction operating on 64 bits, e.g. 64 bit multiplication
  - Needs kernel support for asynchronous signals
- **Oprofile hardware customer mode sampling**
  - Provide CPU measurement data to applications for performance tuning
  - Based on hardware counters and samples built into the CPU
  - Use oprofile to communicate the information to user space programs
- **Valgrind System z support**
  - Valgrind is a generic framework for creating dynamic analysis tools
  - Valgrind is in essence a virtual machine using just-in-time (JIT) compilation techniques
  - Valgrind can be used for memory debugging, memory leak detection, and profiling (e.g. cachegrind)

# Valgrind System z support

- valgrind –tool=memcheck [--leak-check=full] [--track-origins] <program>
  - Detects if your program accesses memory it shouldn't
  - Detects dangerous uses of uninitialized values on a per-bit basis
  - Detects leaked memory, double frees and mismatched frees
- valgrind –tool=cachegrind
  - Profile cache usage, simulates instruction and data cache of the cpu
  - Identifies the number of cache misses
- valgrind –tool=massif
  - Profile heap usage, takes regular snapshots of program's heap
  - Produces a graph showing heap usage over time



# System z kernel features - Misc

- **Crypto Express 3 (kernel 2.6.33)**
  - Toleration support for Crypto Express 3 in Accelerator and Coprocessor mode
- **Kernel image compression (kernel 2.6.34)**
  - The kernel image size can be reduced by using one of three compression algorithms: gzip, bzip2 or lzma.
- **KULI (2009-06-24)**
  - kuli is experimental userspace sample to demonstrate that KVM can be used to run virtual machines on Linux on System z.
  - This experimental proof of concept is unsupported and should not be used for any production purposes.
- **Oprofile**
  - Starting with version 0.9.4, oprofile suppors sampling of Java byte code applications for Linux on System z.
- **Eclipse 3.3**
  - Starting with Eclipse 3.3, Linux on System z is officially supported.

# s390-tools package: what is it?

- **s390-tools is a package with a set of user space utilities to be used with the Linux on System z distributions.**
  - It is **the** essential tool chain for Linux on System z
  - It contains everything from the boot loader to dump related tools for a system crash analysis .
- **Version 1.8.3 was released on 2009-09-23**
- **This software package is contained in all major (and IBM supported) enterprise Linux distributions which support s390**
  - RedHat Enterprise Linux 4
  - RedHat Enterprise Linux 5
  - SuSE Linux Enterprise Server 10
  - SuSE Linux Enterprise Server 11
- **Website: http://www.ibm.com/developerworks/linux/linux390/s390-tools.html**
- **Feedback: linux390@de.ibm.com**

# s390-tools package: the content

**CHANGE**
- chccwdev
- chchp
- chreipl
- chshut
- chcrypt
- chmem

**DASD**
- dasdfmt
- dasdinfo
- dasdview
- fdasd
- tunedasd

**DUMP & DEBUG**
- dbginfo
- dumpconf
- zfcpdump
- zfcpdbf
- zgetdump
- scsi_logging_level

**DISPLAY**
- lscss
- lschp
- lsdasd
- lsluns
- lsqeth
- lsreipl
- lsshut
- lstape
- lszcrypt
- lszfcp
- lsmem

**MONITOR**
- mon_fsstatd
- mon_procd
- ziomon

**z/VM**
- vmconvert
- vmcp
- vmur
- cms-fuse

**NETWORK**
- ip_watcher
- osasnmpd
- qetharp
- qethconf

**MISC**
- cpuplugd
- iucvconn
- iucvtty
- ts-shell
- ttyrun

**TAPE**
- tape390_display
- tape390_crypt

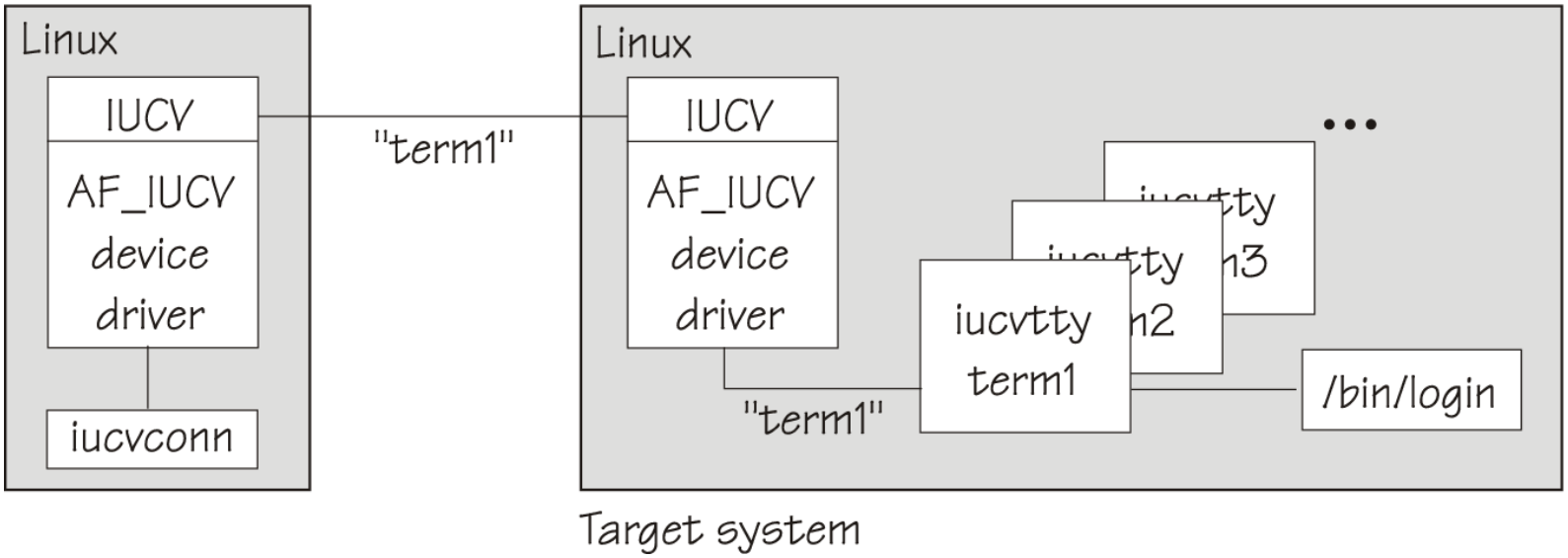
**BOOT**
- zipl

# s390-tools package: version 1.8.x

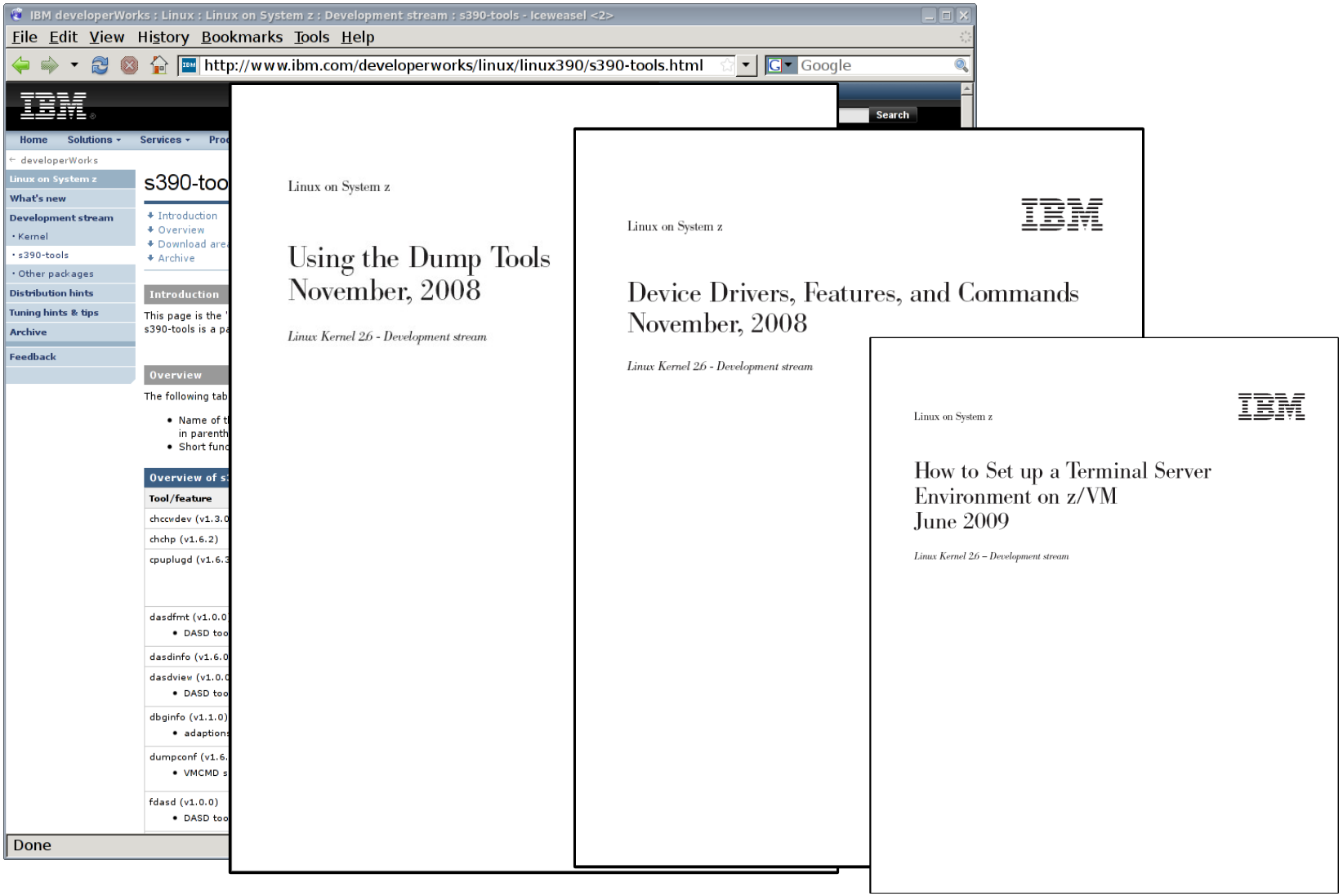
- **Version 1.8.1 (2009-05-09)**
  - *iucvterm*: z/VM iucv terminal app
  - A set of applications to provide terminal access via IUCV. The terminal access does not require an active TCP/IP connection.
  - DASD related tools: Add large volume support for ECKD DASDs
  - The DASD device driver with “Large Volume Support” allows to access move than 65520 cyclinders.
- **Version 1.8.2 (2009-09-18)**
  - cio_ignore: query and modify the contents of the CIO layer blacklist
  - znetconf: list and configure network devices for System z specific network devices
- **Version 1.8.3 (2009-09-23)**
  - zipl: add support for device mapper devices
  - zipl now allows installation of and booting from a boot record on logical devices, i.e. devices managed by device mapper (or similar packages), e.g. multipath devices.
- **Version 1.8.4 (2010-03-12)**
  - readahead: udev rule to increase default maximum readahead
- **Version 1.9.0 (2010-05-28)**
  - cmsfs-fuse: access CMS disks
  - lsmem/chmem: scripts to display / change memory setup
  - ttyrun: tool to safely start getty

# z/VM IUCV hypervisor console (HVC)

- **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**



# More information



# The IBM zEnterprise System – A new dimension in computing


- **The IBM zEnterprise 196 (z196) was announced 2010-07-22**
- **Naturally it is supported by Linux on System z**
- **The upstream kernel supports the following new features of the z196:**
  - Third subchannel set
  - Up to 32 HiperSockets
  - Support for the new OSA CHPID types OSX and OSM has been accepted into kernel 2.6.35, appropriate s390-tools will be published shortly
  - More features to follow

Get all the important hardware details at:
*http://www.ibm.com/common/ssi/rep_ca/0/897/ENUS110-170/ENUS110-170.PDF*

# Questions?

***Martin Schwidefsky***

*Linux on System z Development*

*Schönaicher Strasse 220*
*71032 Böblingen, Germany*

*Phone +49 (0)7031-16-2247*
*schwidefsky@de.ibm.com*