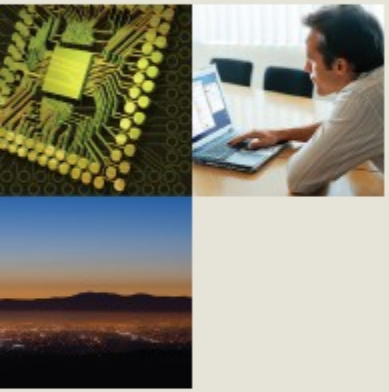


What's New in Linux on System z

Martin Schwidefsky
IBM Lab Boeblingen, Germany

August 12th, 2008
Session 9262

Agenda

- Linux on System z Distributions
- Linux Common Code news
- What's New in System z
  - Kernel
  - GCC
  - s390-tools
  - Kuli
  - Sysload

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 2 (GA 05/2008)
- **Red Hat Enterprise Linux AS 4 (GA 02/2005)**
  - Kernel 2.6.9, GCC 3.4.3, Update 6 (GA 11/2007)
- **Red Hat Enterprise Linux AS 5 (GA 03/2007)**
  - Kernel 2.6.18, GCC 4.1.0, Update 2 (GA 05/2008)
- **Others**
  - Debian, Slackware, ...
  - Support may be available by some third party

Supported Linux Distributions

Hardware Platform and Operating System Software Compatibility - 64-bit enviroment			
Release	**zSeries**	**System z9**	**System z10**
SLES 9	✓	✓	✓
SLES 10	✓	✓	✓
RHEL 3	✓	**	--
RHEL 4	✓	✓	✓
RHEL 5	✓	✓	✓

Hardware Platform and Operating System Software Compatibility - 31-bit enviroment			
Release	**zSeries**	**System z9**	**System z10**
SLES 9	✓	✓	✓
SLES 10	*	*	*
RHEL 3	✓	**	--
RHEL 4	✓	✓	✓
RHEL 5	*	*	*

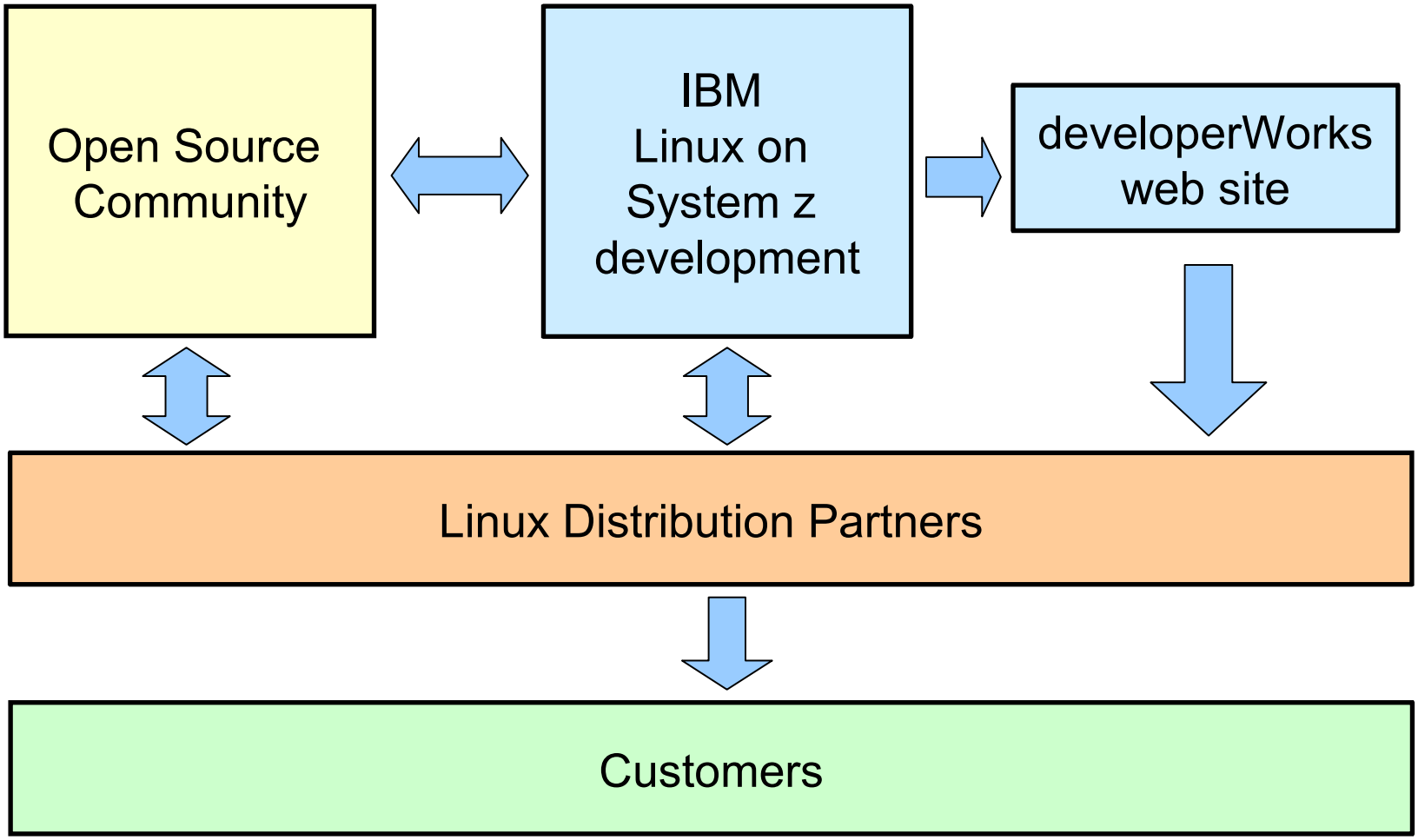
✓ indicates that the operating system release has been tested by IBM in the environment, will run on the system, and is an IBM supported environment. Updates or service packs applied to the release/version are also supported. Newer releases/versions are not supported unless they are listed here.

-- Indicates that the operating system release has not been tested by IBM.

* Distribution does not run in 31 bit mode. Note that 31-bit applications can be run on 64-bit Linux with 31-bit emulation layer.

** Provided on customer request for existing zSeries workloads only. No System z9 feature exploitation.

Linux on System z development process



Kernel news – Common code

- **Linux version 2.6.21 (2007-04-25)**
  - KVM updates
  - Dynticks and Clockevents
- **Linux version 2.6.22 (2007-07-08)**
  - SLUB in kernel memory allocator
  - Signal/timer events through file descriptors
  - Process footprint measurement facility
- **Linux version 2.6.23 (2007-10-09)**
  - Completely Fair Scheduler (CFS)
  - On-demand read-ahead (readahead trashing /3)
  - Variable argument length (no more “arg list too long”)
  - Movable Memory Zone

Kernel news – Common code

- **Linux version 2.6.24 (2008-01-24)**
  - CFS improvements: performance, fair group scheduling, guest time
  - Anti-fragmentation patches
  - Per-device dirty memory thresholds
  - Task Control Groups
- **Linux version 2.6.25 (2008-04-16)**
  - Real Time Group scheduling
  - SMACK, simplified mandatory access control
  - Latencytop
  - BRK and PIE executable randomization
- **Linux version 2.6.26 (2008-07-13)**
  - KVM ported to IA64, PPC and s390 (alias System z)
  - Kgdb kernel debugger
  - Read-only bind mounts (started with 2.6.24)

Current Linux Kernel Development

SHARE
Technology · Connections · Results

Most active 2.6.26 employers

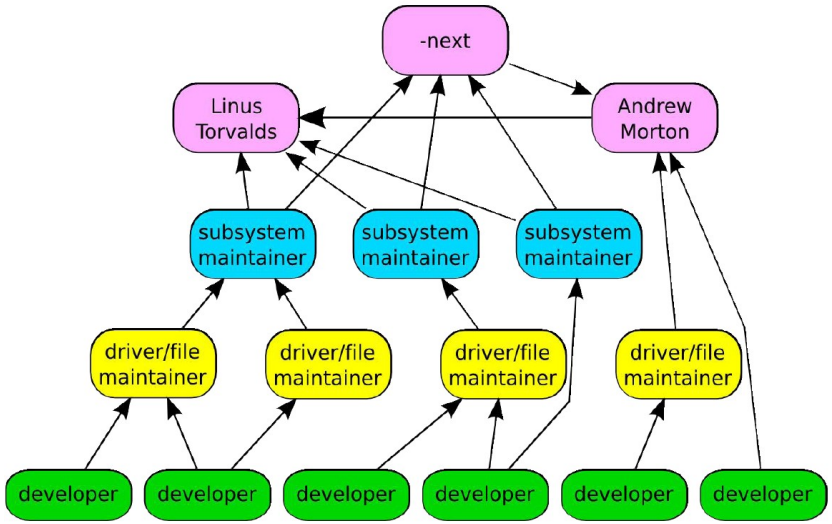
By changesets			By lines changed		
(None)	2085	20.6%	(None)	111703	15.7%
Red Hat	1130	11.2%	IBM	73601	10.3%
(Unknown)	906	8.9%	Red Hat	56331	7.9%
IBM	609	6.0%	Intel	50297	7.1%
Novell	597	5.9%	(Unknown)	44699	6.3%
Intel	469	4.6%	Vyatta	41835	5.9%
Parallels	312	3.1%	Novell	33745	4.7%
SGI	211	2.1%	Movial	28632	4.0%
Movial	180	1.8%	Hauppauge	20234	2.8%
Oracle	142	1.4%	Analog Devices	18363	2.6%
Analog Devices	134	1.3%	(Consultant)	16397	2.3%
HP	124	1.2%	Solarflare	15585	2.2%
MontaVista	122	1.2%	Freescale	15090	2.1%
(Consultant)	116	1.1%	MontaVista	14013	2.0%
Freescale	109	1.1%	QLogic	13327	1.9%
QLogic	97	1.0%	SGI	10351	1.5%
Fujitsu	95	0.9%	Marvell	7881	1.1%
Google	94	0.9%	Wind River	7770	1.1%
(Academia)	89	0.9%	Oracle	7680	1.1%
Marvell	88	0.9%	Pengutronix	7334	1.0%

Source:http://lwn.net/Articles/288233/

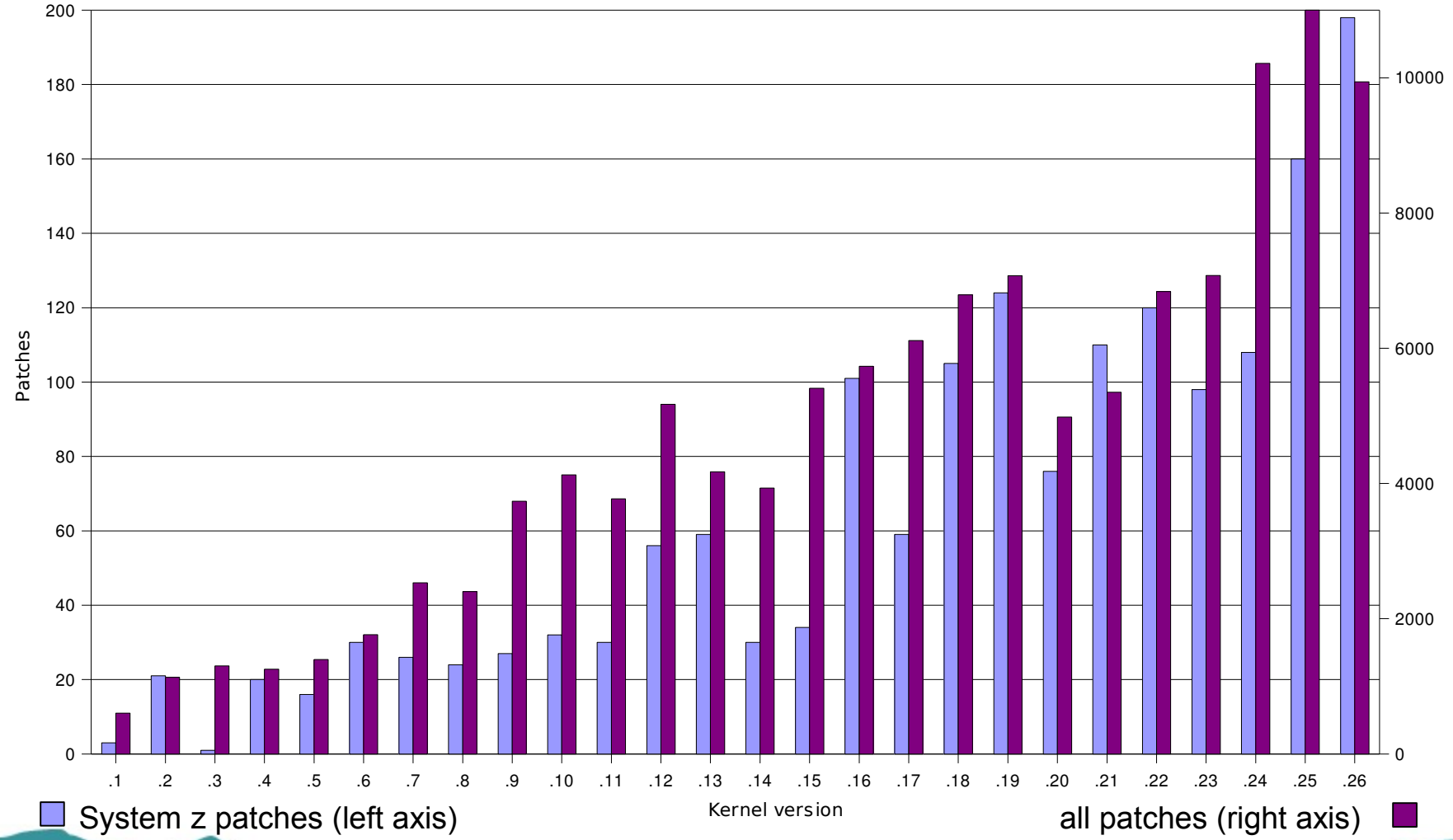
4.300 lines added
1.800 lines removed
1.500 lines modified

per day 2007-2008

Source: Greg KH



Linux kernel – System z contributions



System z kernel features – z10 support

- **CPU node affinity (kernel 2.6.25)**
  - With this feature the kernel uses CPU topology information as supplied by the IBM System z10. This information is used by the scheduler to build scheduling domains and should increase overall performance on SMP machines.
  - This support is available only on IBM System z10, when running Linux on System z in an LPAR.

10.2

- **Vertical CPU management (kernel 2.6.25)**
  - With this feature it is possible to switch between horizontal and vertical CPU polarization via a sysfs attribute.
  - If vertical CPU polarization is active then the hypervisor will dispatch certain CPUs for a longer time than other CPUs for maximum performance.
  - There are three different types of vertical CPUs: high, medium and low. "Low" CPUs get hardly any real CPU time, while "high" CPUs get a full real CPU; "medium" CPUs get something in between.
  - By default the old horizontal CPU polarization is active.
  - This support is available only on z10, running Linux on System z in an LPAR.

System z kernel features – z10 support

- **Large page support - with large page emulation on older hardware (kernel 2.6.25)**
  - This adds hugetlbfs support on System z, using both hardware large page support if available (IBM System z10), and software large page emulation (with shared hugetlbfs pagetables) on older hardware.
  - Exploitation of the IBM System z10 hardware large page support is only available when running Linux on System z in an LPAR.

- **STSI change for capacity provisioning (kernel 2.6.25)**
  - Make the permanent and temporary capacity information as provided by the STSI instruction of the IBM System z10 available to user space via `/proc/sysinfo`.
  - Using this support when running Linux on System z on IBM System z10 as a VM guest requires z/VM 5.3

System z kernel features – z10 support

- **Support for hardware accelerated crypto**
  - Add support for the new hardware accelerated crypto algorithms.
  - The new algorithms are SHA-512 (including SHA-384) and AES-192, AES-256.
  - This support is available only on IBM System z10, running Linux on System z in an LPAR or as a VM guest.
  - The new algorithms have been added to the in-kernel crypto API with kernel version 2.6.25.
  - The new algorithms have been added to the user space library libica 1.3.9 which is part of openCryptoki, see http://opencryptoki.sourceforge.net/.

10.2


5.2

- **System z HiperSockets layer-2 support (kernel 2.6.25)**
  - HiperSockets are enhanced to support layer-2 functionality.
  - The existing OSA layer-2 support is utilized to enable HiperSockets layer-2. This includes IPv6 support for HiperSocket layer-2. Connecting layer-2 and layer-3 hosts is not supported by the System z firmware.
  - This support is available only on z10, running Linux on System z in an LPAR or as a VM guest (z/VM 5.2 or later).

10.2


5.2

System z kernel features – Channel subs.

10.1

5.2

- **Dynamic CHPID reconfiguration via SCLP (kernel 2.6.22)**
  - Allows to react on hardware changes and to configure the path for the LPAR from within a running Linux system
- **I/O configuration support (> kernel 2.6.26)**
  - Adds the infrastructure to allow Linux system to change the I/O configuration of a System z system.
  - Operations are addition, removal and reconfiguration/reassignment of I/O channels, control units and subchannels.
  - This support is available only when running Linux on System z in an LPAR.
- **Modularization of qdio and thin interrupts (> kernel 2.6.26)**
  - Make the thin interrupt layer independent from qdio and improve the code layering in the qdio module. This splits thin interrupts, memory queues and the initialization of the subchannels into seperate, independent units.

System z kernel features – z/VM

- **z/VM Unit-record device driver (kernel 2.6.23)**
  - This is a character device to access the virtual (punched card) reader that is available under VM and allows an ipl from the reader.

- **Linux support for z/VM DIAG 2FC (eWLM) (kernel 2.6.21)**
  - Expand the Hypfs functionality to expose Linux image velocity data that z/VM exposes through a DIAG 2FC call. EWLM will use the hypfs data for velocity management.

- **KERNEL NSS (kernel 2.6.21)**
  - Allows to save a kernel image in a shared memory area called Named Saved Segment (NSS) and IPL from it.

- **Extra kernel parameter via VMPARM (> kernel 2.6.26)**
  - Modify the IPL records to append extra parameters specified wit the z/VM VMPARM option to the kernel command line.

System z kernel features – z/VM

- **AF_IUCV Protocol Support (kernel 2.6.21)**
  - Enables IUCV communication via the BSD socket interface between Linux VM guest or between a Linux guest and CMS

10.2


5.2

- **TTY terminal server over IUCV (> kernel 2.6.27)**
  - Provide central access to the Linux console for the different guests of a z/VM.
  - The terminal server connects to the different guests over IUCV.
  - The IUCV based console is ASCII based.
  - Fullscreen applications like *vi* are usable on the console.

- **Support for enhanced z/VM DASD IUDs (> kernel 2.6.26)**
  - When z/VM provides two virtual devices (minidisks) that reside on the same real device, both will receive the configuration data from the real device and thus get the same uid. To fix this problem, z/VM provides an additional configuration data record that allows to distinguish between minidisks.
  - z/VM APAR VM64273 needs be installed to enable enhanced DASD IUDs.

System z kernel features – DASD

- **DASD HyperPAV support (kernel 2.6.25)**
  - Parallel access volumes (PAV) is a storage server feature, that allows to start multiple channel programs on the same DASD in parallel. It defines alias devices which can be used as alternative paths to the same disk.
  - HyperPAV is activated automatically when the necessary prerequisites are there: (DS8000 with HyperPAV LI and z/VM 5.3, when running Linux on System z as a VM guest)
  - See Document: "How to Improve Performance with PAV"
- **SIM: system information messages (kernel 2.6.25)**
  - With this feature the system reports system information messages (SIM) to the user. The System Reference Code (SRC), which is part of the SIM, is reported to the user and allows to look up the reason of the SIM online in the documentation of the storage server.
- **4G FICON Express support (test only)**
  - Ensure that the new 4G FICON links work with the existing DASD driver.

System z kernel features – FCP

SHARE
Technology · Connections · Results

- **FCP performance data collection:I/O statistics (2.6.25)**
  - The FCP adapter statistics (available since IBM System z9) provide a variety of information about the virtual adapter (subchannel). In order to collect this information the zfcp device driver is extended on one side to query the adapter and on the other side summarize certain values which can then be fetched on demand. This information is made available via files (attributes) in the sysfs filesystem.

SUSE A NOVELL BUSINESS 10.2

- **FCP performance data collection: adapter statistics (2.6.26)**
  - The zFCP adapter collects a number of statistics about the virtual adapter. This information is fetched by the driver and is exported to user space via sysfs.
  - This support is available only on IBM System z9 or later.
- **FCP qdio rate improvements (test only)**
  - The ops/second rate of the zFCP adapter has been increased significantly (x2).
  - This support is available only on IBM System z9 GA3 or later.
- **4G FICON Express support (test only)**
  - Ensure that the new 4G FICON links work with the existing zFCP driver.

System z kernel features – FCP

- **FCP automated port discovery (kernel 2.6.25)**
  - Scan the connected fiber channel SAN and automatically activate all available and accessible target ports. This requires a proper SAN setup with zoning.
- **FCP LUN discovery tool (user space)**
  - A command line tool to display the available LUNs for a specified remote-port.
  - A replacement for the functionality provided by the san-discovery tool based on the zFCP HBA-API.
- **FCP enhanced trace facility (kernel 2.6.21)**
  - The new zfcp error recovery trace code allows to understand all operations related to zfcp error recovery.
  - The trace output can be found in the debug feature /sys/kernel/debug/s390dbf

System z kernel features - Networking

- **Support two OSA ports per CHPID - Four-port exploitation (kernel 2.6.25)**
  - Exploit next OSA adapter generation which offers two ports within one CHPID The additional port number 1 can be specified with the qeth sysfs-attribute "portno".
  - This support is available only for OSA-Express3 GbE SX and LX on z10, running Linux on System z in an LPAR or as a VM guest (PTF for z/VM APAR VM64277 required).

10.2


5.2

- **QETH componentization (kernel 2.6.25)**
  - The qeth driver module is split into a core module and layer2-/layer3-specific modules. The default operation mode for OSA-devices is changed to layer2; for HiperSockets devices the layer3 default-mode is kept.
  - For layer3 mode devices the existence of (possibly faked) ethernet-headers is guaranteed to enable smooth integration of qeth devices into Linux.

- **SKB scatter-gather support for large incoming messages**
  - This avoids allocating big chunks of consecutive memory and should increase networking throughput in some situations for large incoming packets.

10.2


5.2

System z kernel features - Crypto

- **Support for large random numbers (kernel 2.6.25)**
  - Allow user space applications to access large amounts of truly random data. The random data source is the built-in hardware random number generator on the CEX2C cards.
- **Generic algorithm fallback (kernel 2.6.25)**
  - Use software implementation of the in-kernel crypto library for key lengths not supported by hardware
  - Without the fallback support it is not possible to use in-kernel crypto with a key length that is not supported by the hardware module.

System z kernel features - Usability

SHARE
Technology · Connections · Results

- **Standby CPU activation/deactivation (kernel 2.6.25)**
  - With this feature it is possible to make use of standby CPUs for instruction execution.
  - A CPU can be in one of the states "configured", "standby", or "reserved". Before a CPU can be used for instruction execution it must be in "configured" state. Previously, the kernel was limited to operate only with "configured" CPUs. With this feature it is possible to change the state of "standby" CPUs to "configured" state and vice versa via a sysfs attribute.
  - This support is available only on IBM System z10, when running Linux on System z in an LPAR.
- **Shutdown Actions Interface (kernel 2.6.25)**
  - The new shutdown actions interface allows to specify for each shutdown trigger (halt, power off, reboot, panic) one of the five available shutdown actions (stop, ipl, reipl, dump, vmcmd).
  - A sysfs interface under `/sys/firmware` is provided for that purpose.
  - Possible use cases are e.g. to specify that a vmdump should be automatically triggered in case of a kernel panic or the z/VM logoff command should be executed on halt.

System z kernel features - Usability

- **Dynamic memory add / remove (> kernel 2.6.26)**
  - Use the SCLP interface to attach and detach storage elements to the image.
  - Provide the platform support for Linux memory add / remove interface.
- **Struct page elimination (kernel 2.6.26)**
  - Remove the need to allocate a “struct page” structure for pages of a DCSS.
  - No more “mem=” to include the memory areas of the DCCS segments in the memory map.

System z kernel features – Misc

- **ETR Support (kernel 2.6.21)**
  - Support for clock synchronization to an external time reference (ETR)
  - This support is available only when running Linux on System z in an LPAR.

10.2

- **STP Support (> kernel 2.6.26)**
  - Support for clock synchronization using the server time protocol (STP)
  - This support is available only when running Linux on System z in an LPAR.
- **Support for Processor Degradation (kernel 2.6.22)**
  - Generates uevents for all CPUs if the CPU-capability changes (e.g. because the CPUs are overheating).

10.2

- **Cleanup SCSI dumper for upstream integration (kernel 2.6.23)**

Upcoming Kernel Message Documentation

- We plan to document all System z related kernel messages
- A man page can be generated for every message
- Distributors generate man pages for their distributions

```
xpram.1(9)                                                    xpram.1(9)

Message
       xpram.1: %d is not a valid number of XPRAM devices

Severity
       Error

Parameters
       @1: number of partitions

Description
       The  number  of  XPRAM partitions specified for the 'devs' module parameter or with
       the 'xpram.parts' kernel parameter must be an integer in the range  1  to  32.  The
       XPRAM  device  driver created a maximum of 32 partitions that are probably not con-
       figured as intended.

User action
       If the XPRAM device driver has been complied as a separate module, unload the  mod-
       ule  and  load  it  again with a correct value for the into the kernel, correct the
       'xpram.parts' parameter in the kernel parameter line and restart Linux.

LINUX                          Linux Messages                 xpram.1(9)
```

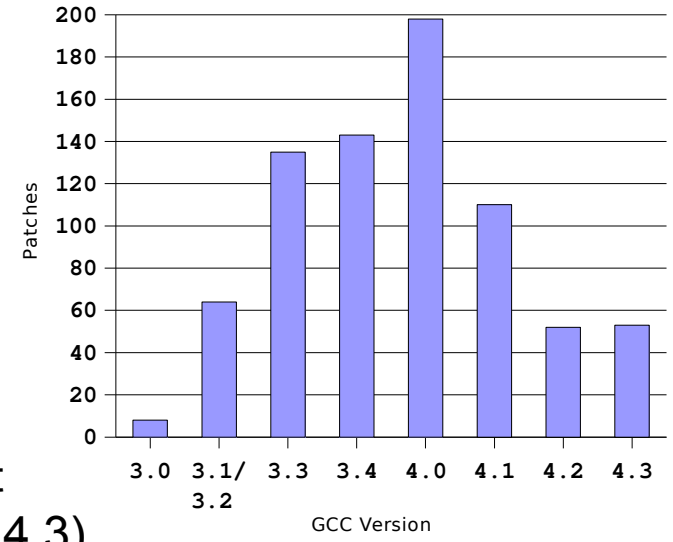
Linux Kernel Directions

- Diversity: now 25 architectures
- Bigger servers (large SGI machines, Mainframes, ...)
- Embedded systems, real-time (Cell-phones, PDAs)
- Appliances (network router, digital video recorder)
- Virtualization (KVM, paravirt, XEN), stronger than ever
- Linus opened the merge window 24 hours after the 2.6.26 release. As of 31. Juli 7885 changesets have been merged.

- **Linux is Linux, but**
  - Features, properties and quality differ dependent on your platform

GNU Compiler

- **General optimizer improvements**
  - New data flow analyzer framework (GCC 4.3)
- **System z machine support**
  - System z10 processor support (> GCC 4.3)
    - Exploit instruction new to z10
    - Selected via -march=z10 / -mtune=z10
  - Decimal floating point support (GCC 4.3)
    - For newer machines with hardware DFP support
  - 64 bit registers for 31 bit applications (> GCC 4.3)
    - Work in progress



- **System z compiler performance**
  - Overall enhancement > 10% on z9 with industry-standard integer benchmark
    - 8% comparing GCC 3.4 and GCC 4.1
    - 5.9% comparing GCC 4.1 and GCC 4.2
    - 0.5% comparing GCC 4.2 and GCC 4.3

s390-tools

- s390-tools is a package with a set of user space utilities to be used with Linux on System z. Latest changes:
  - ***cpuplugd*** - s390-tools 1.6.3
    A daemon that manages CPU and memory based on a set of rules

  - ***New option -x/--extended-uid for dasdinfo*** – s390-tools 1.7.0
    With the PTF for APAR VM64273 installed, z/VM provides a unique identifier that allows to distinguish between virtual disks which are defined on the same real device. This identifier will be part of the uid. To allow for an easier upgrade, the original -u/--uid option will print the uid without this token and the -x/--extended-uid will return the full uid.
  - ***zipl IPL-retry on IFCC*** – s390-tools 1.7.0
    This feature causes the hardware to retry a CCW IPL operation on an alternate channel-path if an interface-control check is detected during execution of the CCW IPL operation.

s390-tools (cont)

- More change in s390-tools version 1.7.0
  - ***Replacement of the kernel parameter string in the zipl menu***
    zipl's boot menu for DASD devices has been changed to allow replacing the complete kernel parameter string with user input when the first character of user input is an equals sign ('=').
  - ***Add VMCMD support to dumpconf***
    The dumpconf init script now exploits the new "shutdown actions interface" introduced with upstream linux kernel 2.6.25. Up to five VM commands can be specified to be triggered in case of a kernel panic.
  - ***mon_fsstatd: Remove init script and sysconfig file***
    mon_fsstatd init script and sysconfig file are replaced by mon_statd, which controls both monitor daemons mon_fsstatd and mon_procd.
  - ***lszfcp: Source code cleanup (No external interfaces changed.)***
- ***Website: http://www.ibm.com/developerworks/linux/linux390/s390-tools.html***

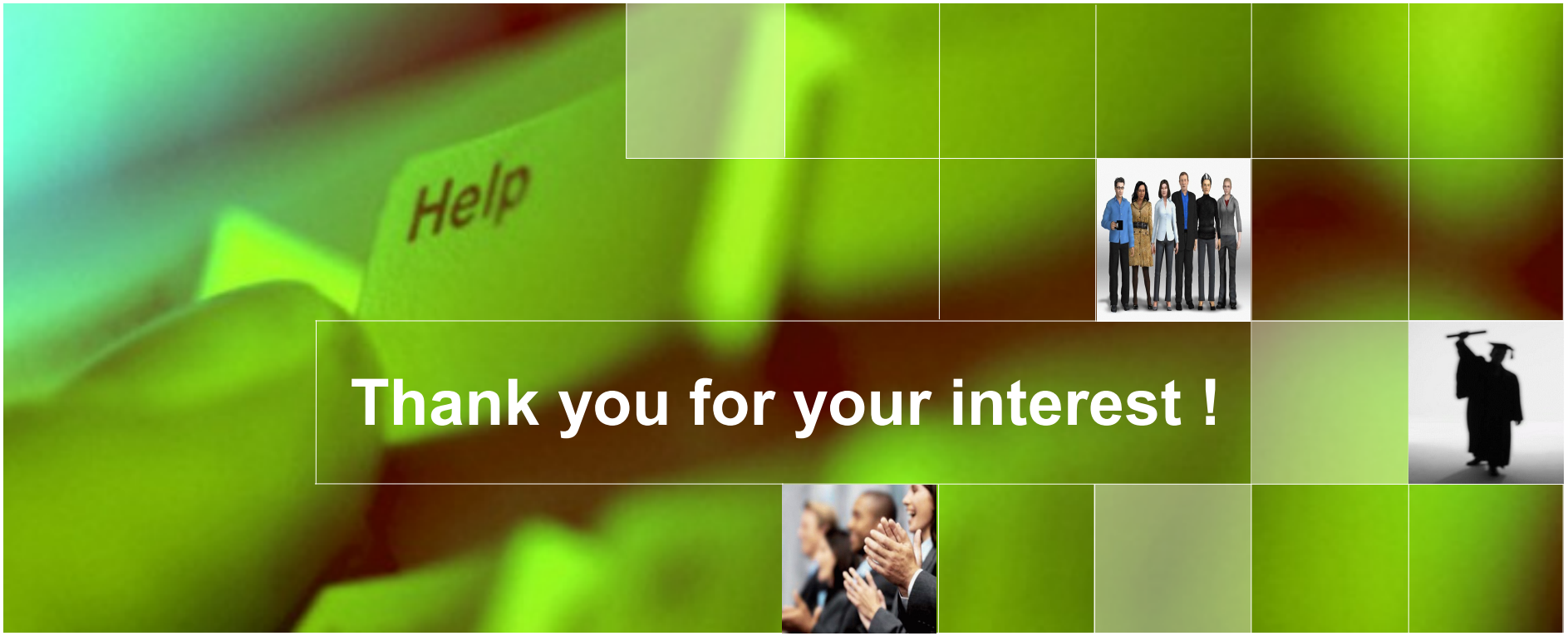
Other packages

- Experimental (unsupported) userspace sample kuli 1.0.0 demonstrating KVM usage (2008-07-04)
  - kuli" is an experimental (unsupported) 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.
- System Loader (sysload) 1.0.0 for Linux on System z.(2008-05-16)
  - System Loader uses a minimal RAM disk based Linux system to run a boot loader application and kexec to boot the final Linux system.
  - zipl from s390-tools is required as a first stage boot loader.
  - Benefits of the System Loader sysload:
    - no need to re-initialize after kernel update
    - flexible boot menu
    - network-based boot via ftp, http, ssh
    - can be used as a generic rescue system

Useful Web links

- www.ibm.com/developerworks/linux/linux390
  - www.ibm.com/developerworks/linux/linux390/whatsnew
  - www.ibm.com/developerworks/linux/linux390/development_recommended
    - www.ibm.com/developerworks/linux/linux390/kernel
    - www.ibm.com/developerworks/linux/linux390/s390-tools
    - www.ibm.com/developerworks/linux/linux390/other_packages
  - www.ibm.com/developerworks/linux/linux390/distribution_hints
  - www.ibm.com/developerworks/linux/linux390/perf/tuning_papers
- publib.boulder.ibm.com/infocenter/systems/index.jsp?topic=/linuxinformation/concep

SHARE
Technology · Connections · Results



Thank you for your interest !

Disclaimer

IBM®, DB2®, MVS/ESA, AIX®, S/390®, AS/400®, OS/390®, OS/400®, iSeries, pSeries, xSeries, zSeries, z/OS, AFP, Intelligent Miner, WebSphere®, Netfinity®, Tivoli®, Informix und Informix® Dynamic ServerTM, IBM, BladeCenter, and POWER and others are trademarks of the IBM Corporation in US and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.
The information and materials are provided on an ”as is” basis and are subject to change.