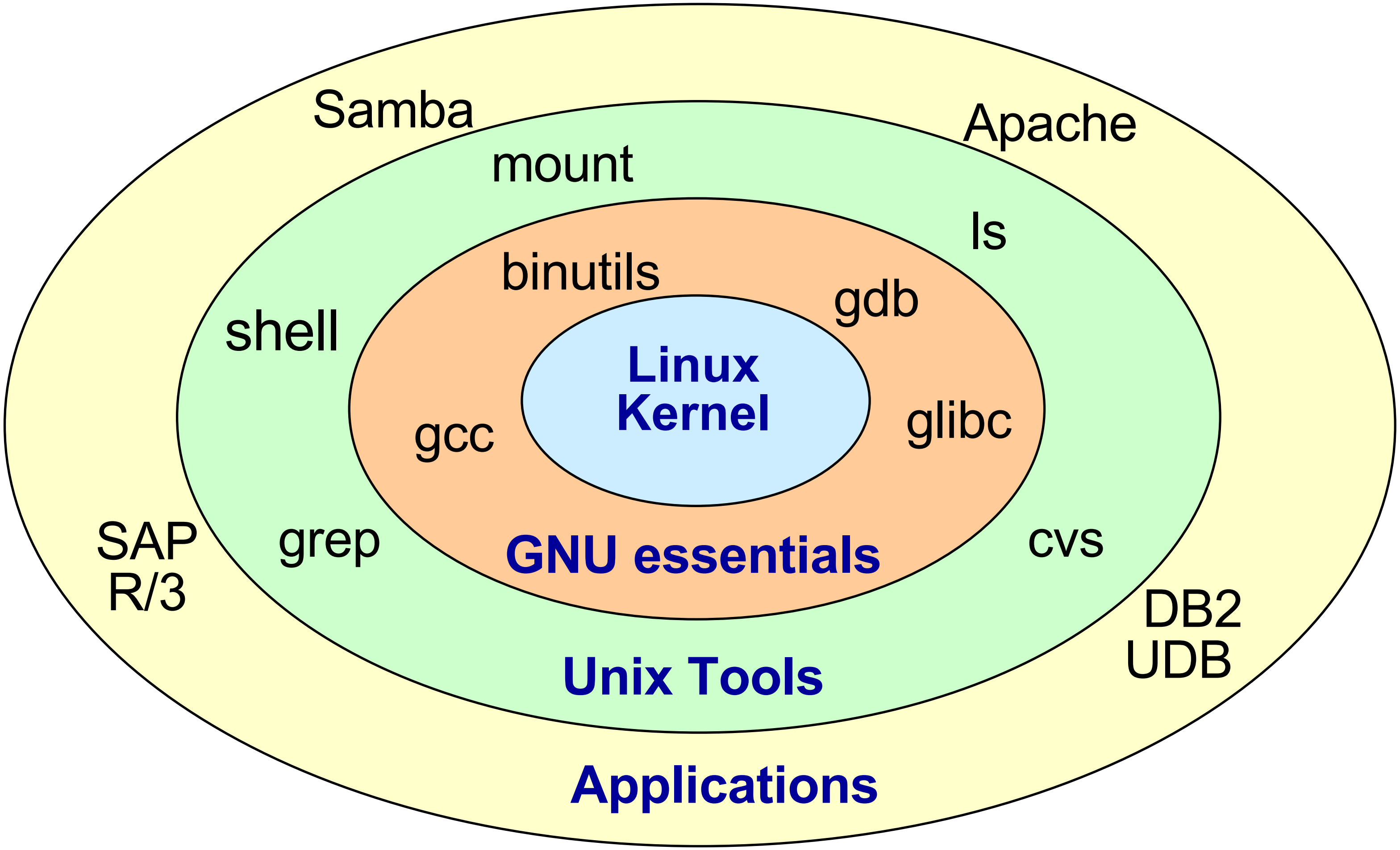
What's New
for Linux on System z

Michael Holzheu (holzheu@de.ibm.com)
Linux on System z Development
IBM Lab Boeblingen, Germany

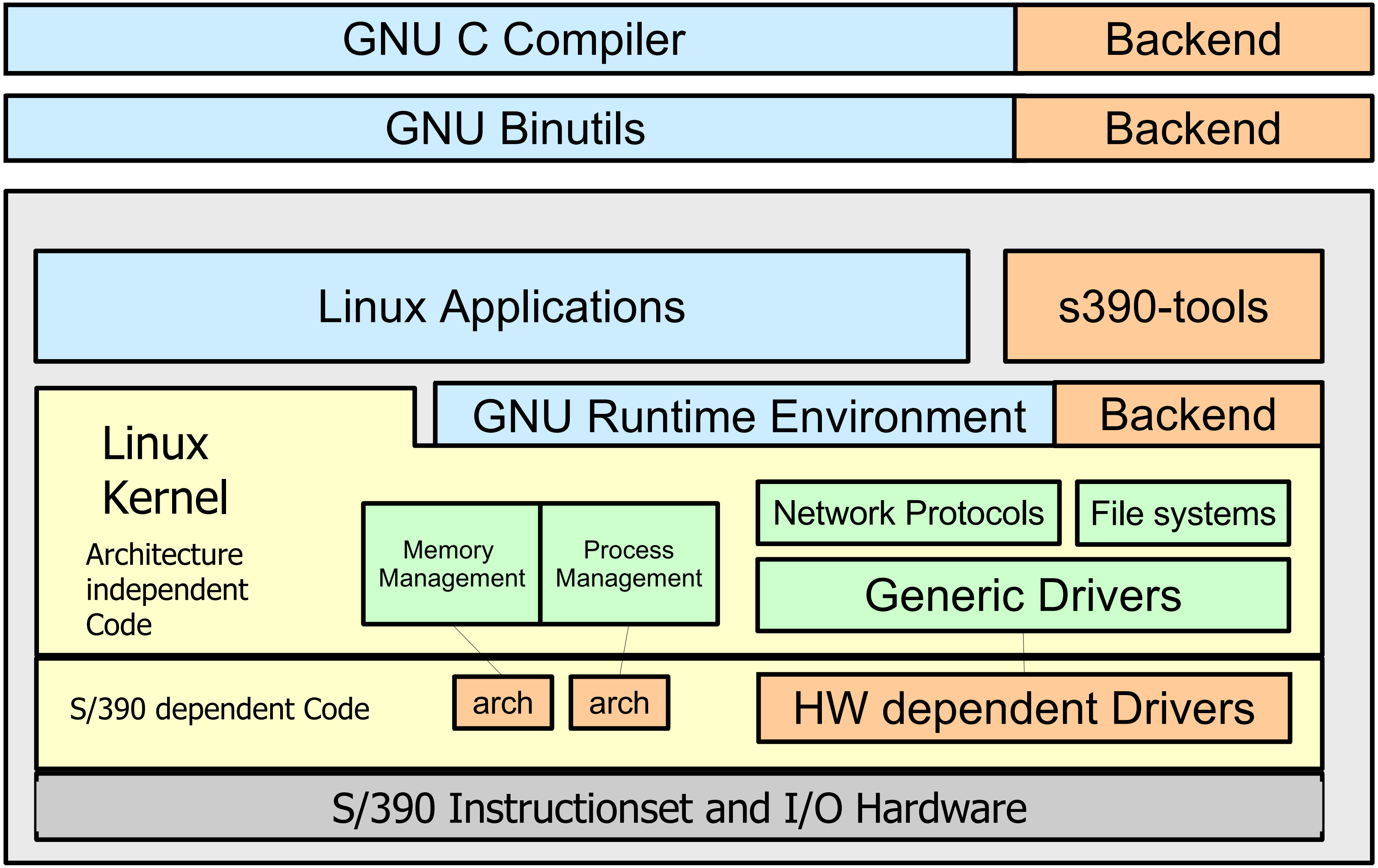
Agenda

- Linux on System z Overview
- Compiler News
- Linux Kernel News
- Linux on System z News

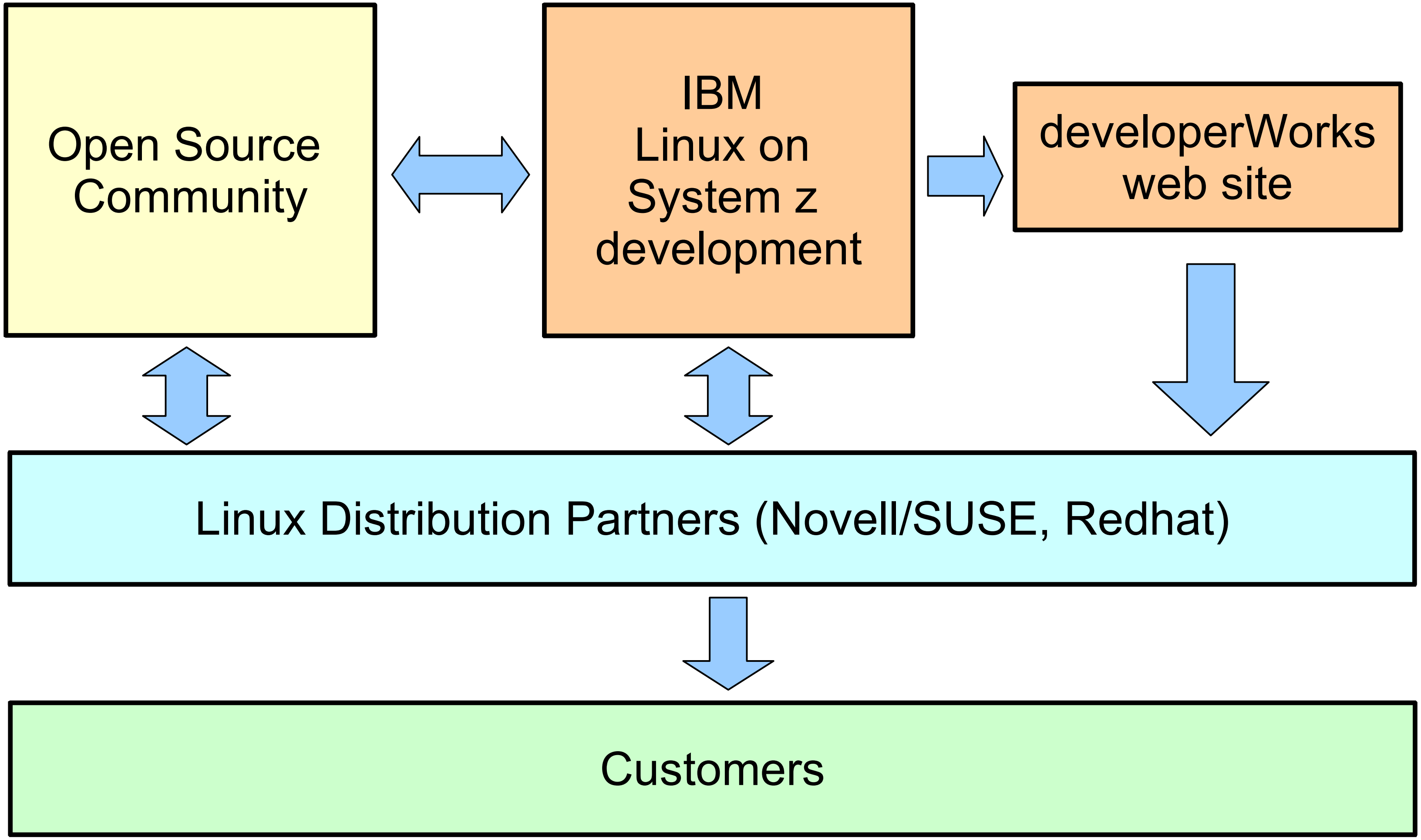
Linux system components



Linux on System z system structure



Linux on System z development process



Open Source Community

IBM Linux on System z development

developerWorks web site

Linux Distribution Partners (Novell/SUSE, Redhat)

Customers

Linux on System z distributions (Kernel 2.6 based)

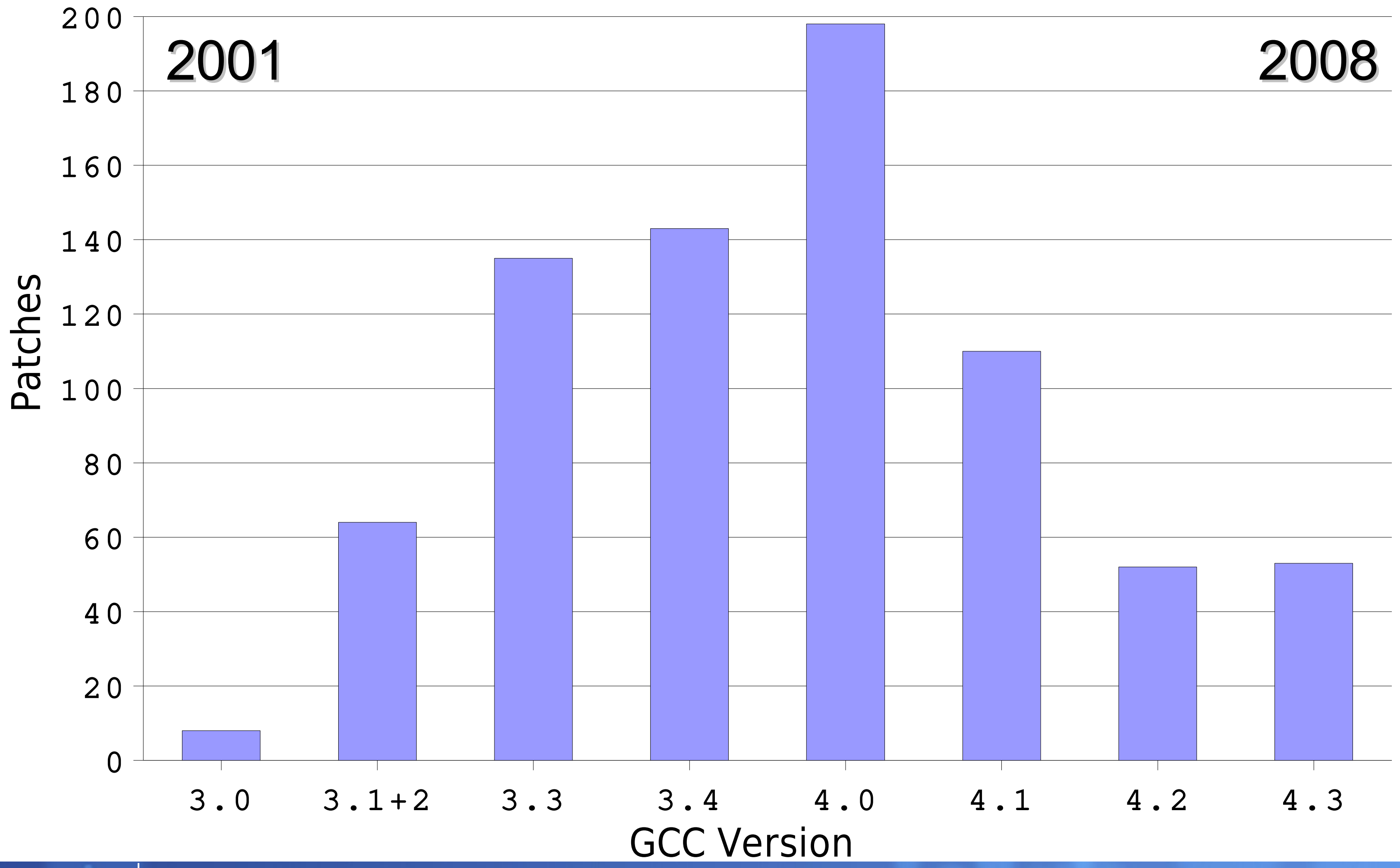
- **SUSE Linux Enterprise Server 9 (GA 08/2004)**
  - Kernel 2.6.5, GCC 3.3.3
  - Service Pack 4 (GA 12/2007)
- **SUSE Linux Enterprise Server 10 (GA 07/2006)**
  - Kernel 2.6.16, GCC 4.1.0
  - Service Pack 1 (GA 06/2007)
- **Red Hat Enterprise Linux AS 4 (GA 02/2005)**
  - Kernel 2.6.9, GCC 3.4.3
  - Update 6 (GA 11/2007)
- **Red Hat Enterprise Linux AS 5 (GA 03/2007)**
  - Kernel 2.6.18, GCC 4.1.0
  - Update 1 (GA 11/2007)
- **Others**
  - Debian, Slackware, ...
  - Support may be available by some third party

Compiler

Open Source development process GCC

- **Centralized development model**
  - Master repository hosted by the Free Software Foundation
    - Read access to the general public, write access to maintainers
    - All copyright owned by / transferred to the FSF
  - Global maintainers (ca. 12), Subsystem maintainers (ca. 130)
- **Release process**
  - New major release every 8-12 months
  - “Dot releases” every 2 months containing regression fixes only
- **System z integration**
  - Back-end maintainers from Böblingen:
    Ulrich Weigand, Hartmut Penner, Andreas Krebbel

GNU Compiler Collection - System z contributions



Compiler News – System z machine support

- **System z10 processor support (> GCC 4.3)**
  - Exploit instruction new to z10
  - Selected via `-march=z10` / `-mtune=z10`
- **System z9 109 processor support (GCC 4.1)**
  - Exploit instructions provided by the *extended immediate facility*
  - Selected via `-march=z9-109` / `-mtune=z9-109`
- **Support for 128-bit IEEE "long double" data type (GCC 4.1)**
  - Provide extended range of floating point exponent and mantissa
  - Selected via `-mlong-double-128`
- **Decimal floating point support (GCC 4.3)**
  - For newer machines with hardware DFP support
  - Selected via `-march=z9-ec, -mhard-dfp/-mnohard-dfp`
  - Software support for older machines without hardware DFP support

Compiler News - System z features

- **Kernel stack overflow avoidance/detection (GCC 4.0)**
  - Compile / Run-time time detection
  - Stack frame size reduction

- **GCC support for the z/TPF OS (GCC 4.0/4.1)**
  - z/TPF uses Linux / GCC as cross-build environment

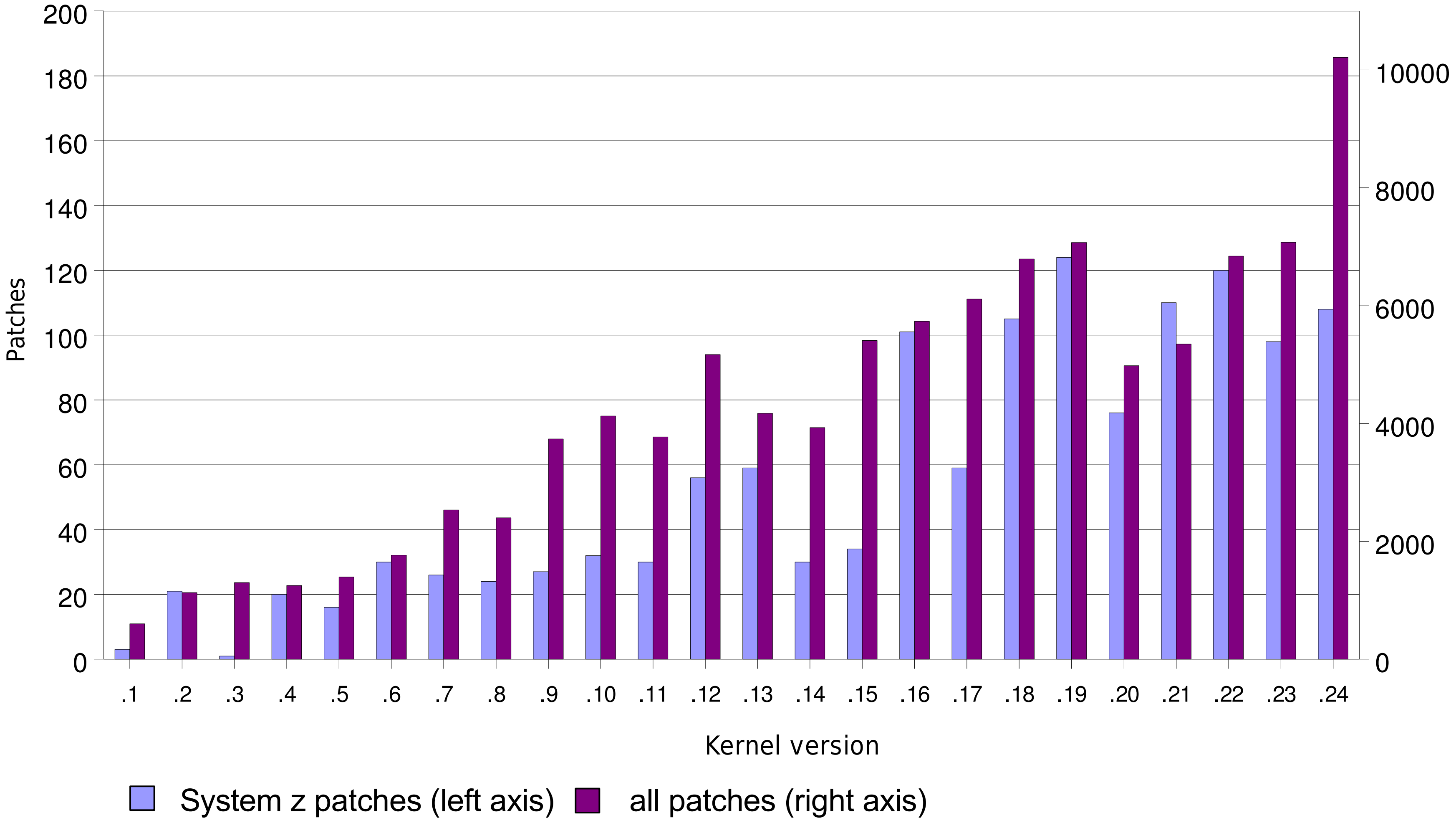
- **Performance enhancements on z9 (compiled code)**
  - Industry-standard integer performance benchmark
  - 8% comparing GCC 3.4 and GCC 4.1
  - 5.9% comparing GCC 4.1 and GCC 4.2
  - 0.5% comparing GCC 4.2 and GCC 4.3

Kernel

Open Source development process - Linux Kernel

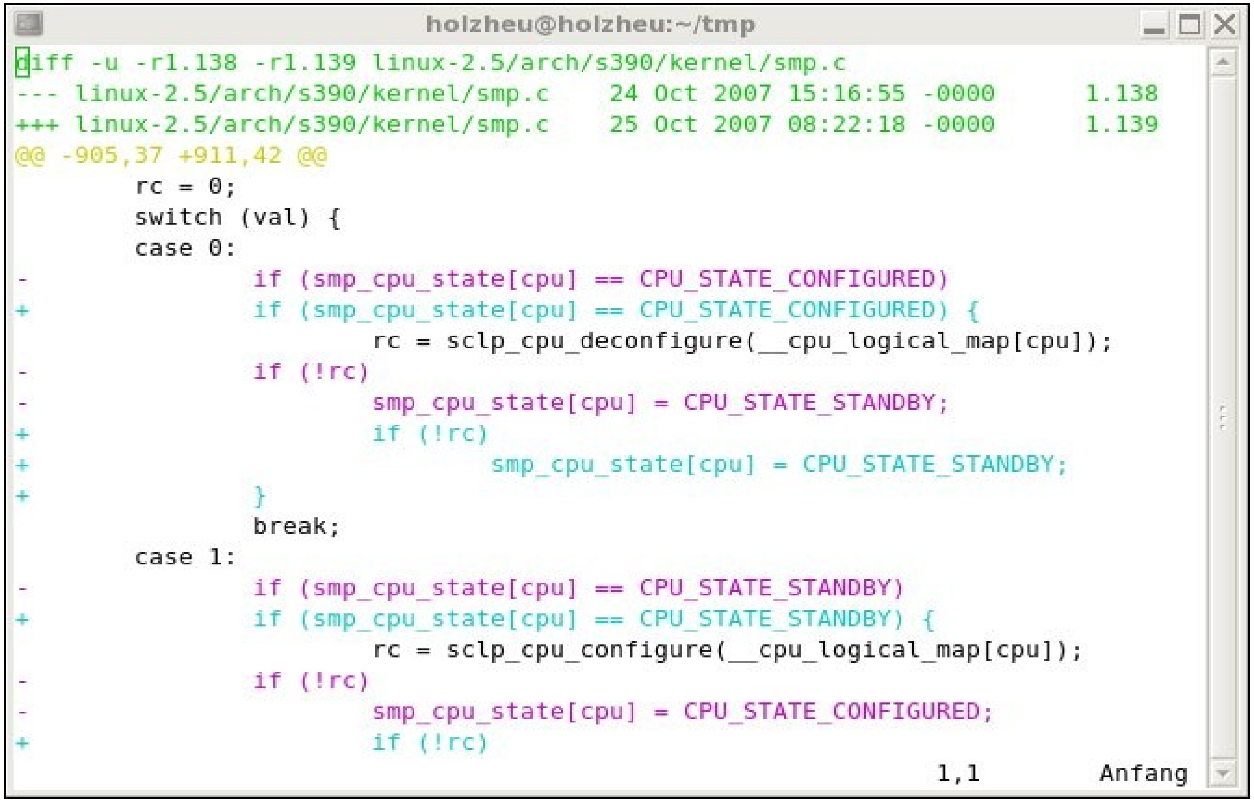
- **Distributed development model**
  - Source code control tool: git
  - 'Master' repository maintained by Linus Torvalds
  - 'Experimental' repository maintained by Andrew Morton
  - Flow of code tracked via “Signed-Off” and “Acked-By” statements
- **Release process**
  - New 2.6.x version released every 2-3 months by Linus
  - First two weeks to merge new features, leading to first -rc
  - Sequence of multiple release candidates to stabilize
- **System z integration**
  - Platform subsystem maintainer in Böblingen:
    - Martin Schwidefsky, Heiko Carstens
  - Repository for System z features hosted on non-IBM site
    - Staging area for IBM and third-party System z patches

Linux kernel – System z contributions



Kernel Patches (Development from 2.6.23 to 2.6.24)

- Patch Size: 1.697.585 lines
- Change Log: 154.444 lines



holzheu@holzheu:~/tmp

```
diff -u -r1.138 -r1.139 linux-2.5/arch/s390/kernel/smp.c
--- linux-2.5/arch/s390/kernel/smp.c    24 Oct 2007 15:16:55 -0000      1.138
+++ linux-2.5/arch/s390/kernel/smp.c    25 Oct 2007 08:22:18 -0000      1.139
@@ -905,37 +911,42 @@
        rc = 0;
        switch (val) {
        case 0:
-               if (smp_cpu_state[cpu] == CPU_STATE_CONFIGURED)
+               if (smp_cpu_state[cpu] == CPU_STATE_CONFIGURED) {
                        rc = sclp_cpu_deconfigure(__cpu_logical_map[cpu]);
-               if (!rc)
-                       smp_cpu_state[cpu] = CPU_STATE_STANDBY;
+                       if (!rc)
+                               smp_cpu_state[cpu] = CPU_STATE_STANDBY;
+               }
                break;
        case 1:
-               if (smp_cpu_state[cpu] == CPU_STATE_STANDBY)
+               if (smp_cpu_state[cpu] == CPU_STATE_STANDBY) {
                        rc = sclp_cpu_configure(__cpu_logical_map[cpu]);
-               if (!rc)
-                       smp_cpu_state[cpu] = CPU_STATE_CONFIGURED;
+                       if (!rc)
                                                              1,1        Anfang
```

How to get new features into distributions

- **Upstream feature (ideal case)**
  - Develop feature against mainline kernel, accepted in kernel version 2.6.x
  - Distribution release based on 2.6.x or later will usually include feature
- **Backport of upstream feature (usually acceptable)**
  - Code already accepted in some kernel version 2.6.x
  - Develop back-port against previous kernel release, provide on developerWorks and/or to distributor
  - Distribution release/update based on earlier kernel may add the feature as additional patch
- **Feature not upstream (difficult)**
  - Code provided only on developerWorks and/or to distributor, not yet accepted in any upstream kernel
  - Distributors are generally reluctant to add such features as additional patches due to maintenance concerns

Kernel news - Linux version 2.6.20 - 2.6.24

- **Virtualization:**
  - Kernel Virtual Machine (KVM)
  - Lguest and Xen
- **New Functions:**
  - New Filesystems: GFS2, Ext4, ecryptfs
  - Read-only bind mounts
- **Performance:**
  - High resolution timers
  - Better kernel memory allocator (SLUB)

Kernel news - Linux version 2.6.20 - 2.6.24

- **Performance (cont.):**
  - Completely Fair Scheduler (CFS)
  - On-demand read-ahead
  - Anti-memory-fragmentation
  - Per-device dirty memory thresholds
- **Measurement:**
  - Process footprint measurement facility
  - I/O Accounting for processes

Kernel directions

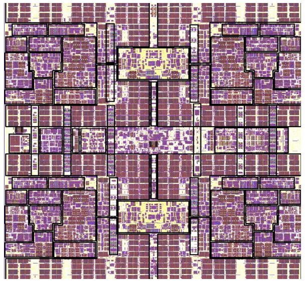
- **Diversity: now 24 architectures**
- **Bigger servers (Mainframes, large SGI machines, ...)**
- **Embedded systems, real-time (Cell-phones, PDAs)**
- **Appliances (network router, digital video recorder)**
- **Virtualization (KVM, XEN, etc.), stronger than ever**

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

Kernel (System z)

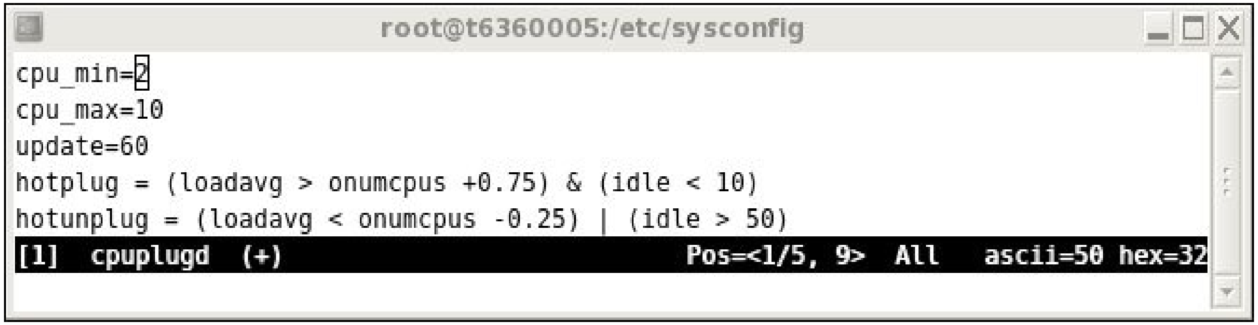
System z kernel features - CPU

- **CPU node affinity**
  - Exploits LPAR hardware interface to query cpu topology
  - Gives hints to Linux scheduler
  - /sys/devices/system/cpu/cpuX/topology/core_siblings
- **STSI changes for capacity provisioning**
  - Permanent and temporary capacity
  - /proc/sysinfo
- **Standby CPU activation / deactivation**
  - echo 0 > /sys/devices/system/cpu/cpuX/configure
  - echo 0 > /sys/devices/system/cpu/cpuX/online



System z kernel features - CPU

- **Dynamic CPU hotplug daemon**
  - Automatically configures guest/LPAR resources
    - Sets CPUs online and offline
    - Adjusts Linux memory footprint (z/VM - CMM1)
  - Rules defined in /etc/sysconfig/cpuplugd:



```
cpu_min=2
cpu_max=10
update=60
hotplug = (loadavg > onumcpus +0.75) & (idle < 10)
hotunplug = (loadavg < onumcpus -0.25) | (idle > 50)
[1]  cpuplugd  (+)                    Pos=<1/5, 9>  All   ascii=50 hex=32
```

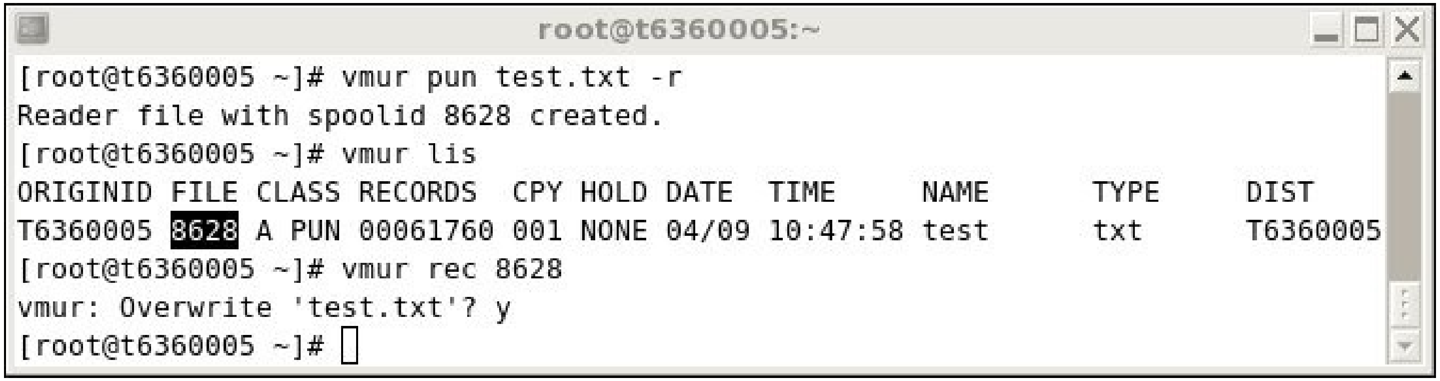
- **Support for processor degradation**
  - Kernel message + Userspace Events

System z kernel features - Crypto

- **New hardware support – System z10 processor**
  - New crypto instructions for:
    - AES 192/256
    - SHA 384/512
  - “Libica” userspace support → Faster ssh!
  - Kernel support → Faster IPSEC!
- **Generic algorithm fallback**
  - Use software implementation for key lengths not supported by hardware
- **Crypto driver**
  - Support for long random numbers
    - Character device driver: /dev/hwrng
  - Capability for dynamic crypto device add

System z kernel features - z/VM

- **Linux process data in monitor APPLDATA**
  - Write Linux process specific data to monitor stream
- **Unit record device support**
  - Vmur character device driver and vmur user space tool
  - Support for z/VM punch, printer and reader



```
[root@t6360005 ~]# vmur pun test.txt -r
Reader file with spoolid 8628 created.
[root@t6360005 ~]# vmur lis
ORIGINID FILE CLASS RECORDS  CPY HOLD DATE  TIME     NAME      TYPE     DIST
T6360005 8628 A PUN 00061760 001 NONE 04/09 10:47:58 test      txt      T6360005
[root@t6360005 ~]# vmur rec 8628
vmur: Overwrite 'test.txt'? y
[root@t6360005 ~]# 
```

- **IUCV access to z/VM services (user space netcat / nc6)**

System z kernel features – Networking / SCSI

- **QETH network driver**
  - HiperSockets MAC layer routing
  - QETH componentization
- **Support for skb scatter-gather**
  - Increases performance for inbound traffic
- **FCP performance**
  - FCP performance data collection - I/O statistics + adapter statistics
    - New sysfs attributes for throughput, latency etc.
  - 4G FICON Express support for FCP

System z kernel features - Usability and RAS

- **IPL**
  - IPL through IFCC / multipath IPL
  - Shutdown actions interface
- **System dump**
  - Cleanup SCSI dumper for upstream integration
- **DASD sense data reporting**
  - SIM/MIM handling for ECKD DASD
- **Dynamic CHPID reconfiguration via SCLP**
  - User space tools to modify and list chpids:
    - chchp
    - lschp

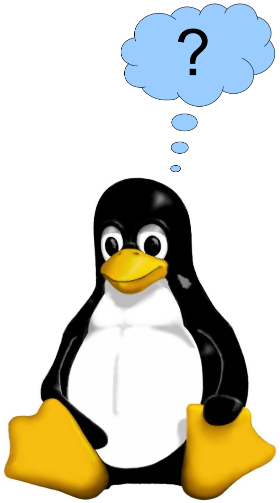
System z kernel features - Performance

- **Kernel Large Page Support**
  - System z10 hardware support (1MB pages instead of 4KB)
  - Benefits:
    - Reduce page table memory consumption (2KB per MB / process)
    - Speed up address translation
  - Software emulation for older machines
  - SysV shared memory (shmget / SHM_HUGETLB)
  - hugetlbfs filesystem (mmap)
  - Java option for using large pages (-Xlp)
- **DASD Hyper PAV (Parallel Access Volume)**
  - Base & Alias Devices
  - Start multiple channel programs on a single DASD in parallel
  - Alias devices are not assigned to specific base devices
  - User space multipath setup is now obsolete

Outlook

- **New hardware exploitation**
- **Enhanced Linux – z/VM synergy**
- **Basic support for KVM virtualization**
- **Keep current with open source**

Questions?



Trademarks

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, DBE, e-business logo, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/30, VM/ESA, VSE/ESA, Websphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

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