What’s New for Linux on System z?

Volker Sameske (sameske@de.ibm.com)
Linux on System z Development
IBM Lab Boeblingen, Germany

System z Technical Conference, Munich
April 16-20, 2007
Session L01
Agenda

- Linux on System z Overview
- Development Process
  - Linux Kernel
  - Compiler gcc
- Distributor Support
- Linux Kernel News
- What’s new on System z
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 3 (GA 12/2005)
- SUSE Linux Enterprise Server 10 (GA 07/2006)
  - Kernel 2.6.16, GCC 4.1.0
- Red Hat Enterprise Linux AS 4 (GA 02/2005)
  - Kernel 2.6.9, GCC 3.4.3
  - Update 4 (GA 07/2006)
- Red Hat Enterprise Linux AS 5 (GA 03/2007)
  - Kernel 2.6.18, GCC 4.1.0

- Others
  - Debian, Slackware, ...
  - Support may be available by some third party
Linux system components

- Linux Kernel
- GNU essentials
- Unix Tools
- Applications
  - SAP R/3
  - grep
- applications
  - gdb
- glibc
  - cvs
- DB2 UDB
  - Samba
- Apache
  - mount
- ls
  - binutils
- Linux Kernel
Linux on System z system structure

- GNU C Compiler
- GNU Binutils
- Linux Applications
  - Linux Kernel
    - Architecture independent Code
    - S/390 dependent Code
  - GNU Runtime Environment
    - Network Protocols
    - File systems
  - Generic Drivers
    - Memory Management
    - Process Management
    - arch
  - HW dependent Drivers
- S/390 Instructionset and I/O Hardware
Linux on System z development process

- Open Source Community
- IBM Linux on System z development
- developerWorks web site
- Linux Distribution Partners
- Customers
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
  - Secondary repositories maintained by subsystem maintainers and others
  - 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 release candidate
  - Sequence of multiple release candidates to stabilize

- **System z integration**
  - Platform subsystem maintainer: Martin Schwiddefsky
  - **New**: git repository for System z features hosted on non-IBM site
    - Staging area for IBM and third-party System z patches
    - Experimental System z features
Linux kernel – System z contributions
Open Source development process: GCC

- **Centralized development model**
  - Source code control tool: subversion
  - 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
  - GCC Steering Committee oversees the project
  - SC delegates design/development to maintainers
    - Global maintainers (ca. 12), Subsystem maintainers (ca. 130)

- **Release process**
  - New major release every 8-12 months
  - Development stages: Major changes, minor changes, bugs, regressions
  - “Dot releases” every 2 months containing regression fixes only

- **System z integration**
  - Platform back-end maintainers: Ulrich Weigand, Hartmut Penner
  - Generally all System z features merged upstream
GNU Compiler Collection – System z contributions

- GCC version Patches

- GCC version
  - 3.0
  - 3.1/3.2
  - 3.3
  - 3.4
  - 4.0
  - 4.1
  - 4.2

- Patches
  - 0
  - 20
  - 40
  - 60
  - 80
  - 100
  - 120
  - 140
  - 160
  - 180
  - 200
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
Object-code only kernel modules

- **Issues**
  - OCO modules need to be re-built with every kernel change
  - Distributors reluctant to include OCO modules

- **Currently, we have no OCO module**
  - lcs: open source since 2002-03-04, upstream in 2.4.x
  - z90crypt: open source since 2002-07-31, upstream in 2.4.x
  - qdio: open source since 2002-09-13, upstream in 2.4.x
  - qeth: open source since 2003-06-30, upstream in 2.4.x
  - tape_3590: open source since 2006-03-28, upstream in 2.6.17

- **Future strategy: No more OCO modules!**
Kernel news – Linux version 2.6.17 (2006-06-17)

- Niagara multicore cpu support (Sparc64)
- Wlan improvements (Softmac layer, Broadcom 43xx driver)
- Splice/tee/vms splice system calls for faster process communication
- New scheduler domain for multi core cpus with shared cache
- Block queue I/O tracing (blktrace)
- Lightweight robust futexes
- Generic RTC interface
- ...
Kernel news – Linux version 2.6.18 (2006-09-19)

- Lightweight user space priority inheritance
- Lockdep – a kernel lock validator
- New power saving policy for multi core system
- Swapless page migration, per-zone VM counters
- New default I/O scheduler: CFQ
- Generic core time subsystem
- Devfs removal
- vDSO randomization
- ...
- [tons of architecture and driver updates]
Kernel news – Linux version 2.6.19 (2006-11-29)

- New file systems: GFS2, Ext4, ecryptfs
- Parallel ATA subsystem
- AVR32 Architectures (32 bit embedded RISC processor)
- RCU enhancements (sleepable RCU)
- Configurable block layer
- Vectored AIO support
- Namespaces for IPC and UTS
- ...
- [tons of architecture and driver updates]
Kernel news – Linux version 2.6.20 (2007-02-04)

- Kernel Virtual Machine (KVM)
- Relocatable kernel images (i386)
- Asynchronous SCSI scanning
- Multithreaded USB probing
- I/O Accounting
- Relative atime support
- Bus event notifications
- ...
- [tons of architecture and driver updates]
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)

- **Linux is Linux, but**
  - Features, properties and quality differ dependent on your platform
System z kernel features – hardware support

- **Channel tape enhancements**
  - 3592 Control unit recognition (*in 2.6.17, 4Q06*)
  - 3592 crypto tape support (>2.6.20, 1Q07)

- **Kernel**
  - External time reference (ETR) support (>2.6.20, 1Q07)
System z kernel features – Virtualization

- **Reduction of virtualization overhead**
  - QDIO pass-through stage 2 (in 2.6.16, DW 1Q06)
  - Collaborative memory management stage 2 (under discussion, DW 4Q06)
  - z/VM DIAG250 I/O support for 64-bit (in 2.6.14, DW 1Q06)
  - Directed yield of spinlocks with diagnose 0x9C (in 2.6.19, DW 4Q06)

- **Usability enhancements**
  - Guest LAN sniffer support (in 2.6.15, DW 1Q06)

- **Memory savings**
  - Virtual mem_map array (in 2.6.20, no DW)
System z kernel features – Virtualization z/VM

- **z/VM APPLDATA enhancements**
  - Application support (2.6.19, DW 1Q07)
  - CPU Hotplug enhancements (2.6.18, DW 4Q06)
  - Add steal time information to record layout (2.6.18 DW 4Q06)
  - Linux guest file size in monitor APPLDATA (>2.6.20 DW 1Q07)

- **z/VM integration**
  - Kernel NSS support (>2.6.20, DW 1Q07)
  - AF_IUCV support (>2.6.20, DW 1Q07)
System z kernel features – Operational Simplification

- **Runtime configuration**
  - Switch for qeth and qdio performance statistics (in 2.6.20, DW 1Q07)
  - Switch for DASD error logging (in 2.6.20, DW 1Q07)

- **FCP enhancements**
  - snIPL SCSI load (DW 1Q07)
  - Program directed IPL support/no XML in system dumper (>2.6.20, DW 1Q07)
  - FCP performance data collection (>2.6.20, DW 1Q07)
System z kernel features – RAS

○ Kernel
  • Reboot with alternate parameters from FCP (2.6.20, DW 1Q07)
  • IPL/dump on panic (2.6.20, DW 1Q07)

○ Common I/O Layer
  • Improved handling of dynamic subchannel mapping (2.6.20, DW 1Q07)

○ DASD
  • Fast fail support (in 2.6.16, DW 1Q06)
  • Enhanced error reporting (in 2.6.17, no DW)

○ FCP
  • Best effort SAN notifications (in 2.6.16, DW 1Q06)
System z kernel features – Performance

○ Scalability enhancements
  • Multiple Subchannel Set support (in 2.6.16, DW 1Q06)
  • Linux PAV support for LPAR (in 2.6.18, no DW)
  • HyperSwap support DASD & Common I/O layer (2.6.18, DW 4Q06)

○ Hardware/kernel performance data collection
  • Channel path measurement data (in 2.6.17, no DW)
  • Access to LPAR performance data (in 2.6.18, no DW)

○ User and kernel space code profiling
  • Oprofile in-kernel call graph support (in 2.6.16, DW 1Q06)
System z kernel features – Security

- Intrusion prevention
  - Enhanced Linux system layout or NX support (>2.6.20, DW 1Q07)

- New hardware support – z9 processor
  - Support user-space AES+SHA+PRNG crypto CP Assists
  - Support in-kernel AES+SHA crypto CP Assists (in 2.6.16, DW 1Q06)
  - PRNG in kernel pseudo random numbers (>2.6.20, DW 1Q07)

- Crypto device driver
  - Secure Key cryptography (in 2.6.19, no DW)
Compiler – Common features

- General optimizer improvements
  - SSA-based common optimization infrastructure (GCC 4.0)
  - Inter-procedural optimization infrastructure (GCC 4.1)

- Languages and language features
  - Fortran 95 front end (GCC 4.0)
  - Decimal Floating Point support (GCC 4.2)

- Other improvements
  - Stack Protector feature (GCC 4.1)
  - Builtins for atomic operations (GCC 4.1)
Compiler – System z features

- 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 quad “long double” data type (GCC 4.1)
  - Provide extended range of floating point exponent and mantissa
  - Selected via `-mlong-double-128`

- Kernel stack overflow avoidance/detection (GCC 4.0)
  - Compile time detection: `-mwarn-framesize / -mwarn-dynamicstack`
  - Run-time detection: `-mstack-size / -mstack-guard`
  - Stack frame size reduction: `-mpacked-stack`

- GCC support for the z/TPF OS (GCC 4.0/4.1)
  - z/TPF uses Linux / GCC as cross-build environment
  - New target `s390x-ibm-tpf`
Compiler – System z performance

- Compiler back-end improvements
  - Improved condition code handling (GCC 4.0)
  - Improved function prologue/epilogue scheduling (GCC 4.0)
  - Improved use of memory-to-memory instructions (GCC 4.0)
  - Added sibling call support (GCC 4.0)
  - Enhanced use of string instructions (SRST, MVST, ...) (GCC 4.1)
  - More precise register tracking (r13, r6, ...) (GCC 4.1)
  - Use LOAD ZERO (GCC 4.1)
  - ICM/STCM, BRCT, vararg enhancements (GCC 4.1)

- Overall performance enhancement 8%
  - Industry-standard integer performance benchmark
  - Comparing GCC 3.4 and GCC 4.1 on System z
Outlook

- New hardware exploitation
- Enhanced Linux – z/VM synergy
- Enhanced integration with z/OS
- Keep current with open source
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/390, 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
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.