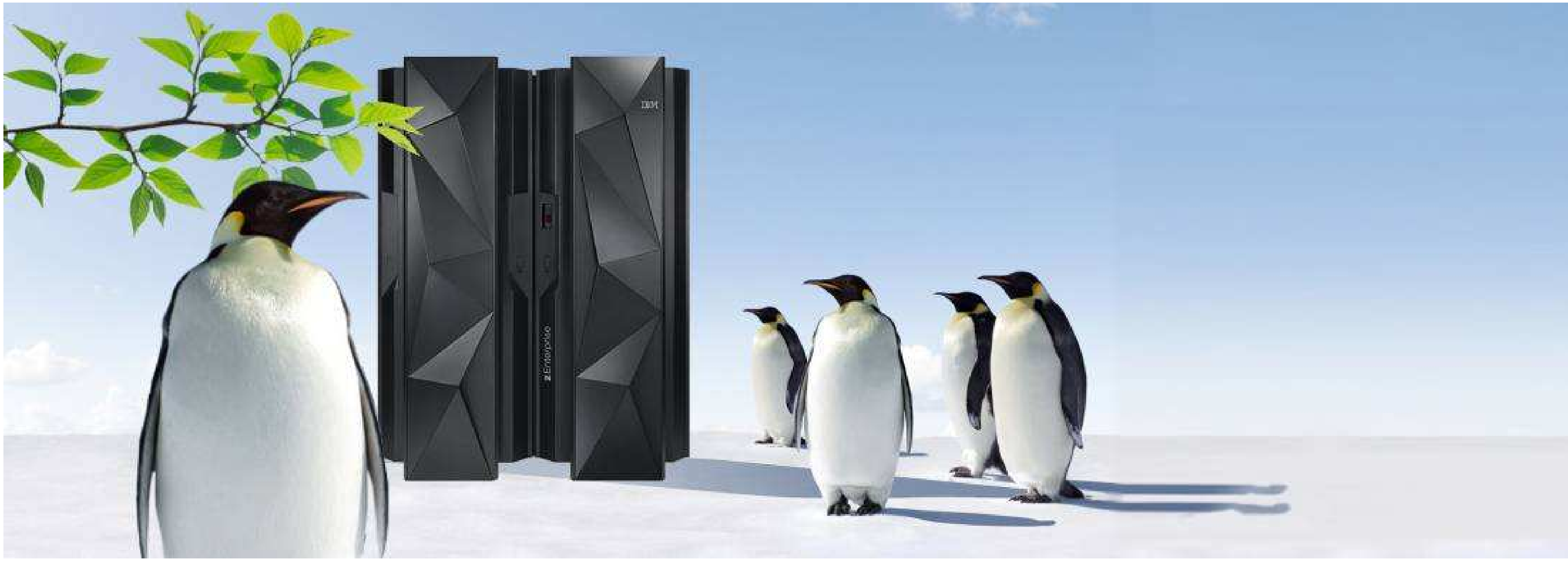
Dr. Stefan Reimbold
IBM Research & Development

Linux on System z - What’s New ?



Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at Copyright and trademark information at www.ibm.com/legal/copytrade.shtml.

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.

Agenda

- Linux Development
- Distributions
- System z Code News
- Tool-Chain

Linux Trivia

- Kernel 1.0.0 176,250 lines of code
- Kernel 3.3 15,000,000 lines of code in 2012
- 3/4 is driver code
- 3 Billion USD estimated development costs
- 30 CPU architectures with many machine architectures
- 476 of the Top500 systems running Linux (performance 97.4%)
  - and growing
- 1.91% of desktop clients (browser stats)

source: http://en.wikipedia.org/wiki/Linux_kernel
http://www.top500.org
www.w3counter.com

IBM Integration with Linux Community

- Since 1999
- One of the leading contributors
- > 600 full-time developers in Linux and Open Source

Linux Kernel & Subsystem Development

- Kernel Base
- Security
- Systems Mgmt
- Virtualization
- Filesystems
- and more . . .

Expanding the OpenSource Ecosystem

- Apache
- Eclipse
- Firefox
- OpenOffice
- and more . . .

Promoting Open Standards & Community Collaboration

- The Linux Foundation
- Linux Standards Base
- Common Criteria Certification
- and more . . .

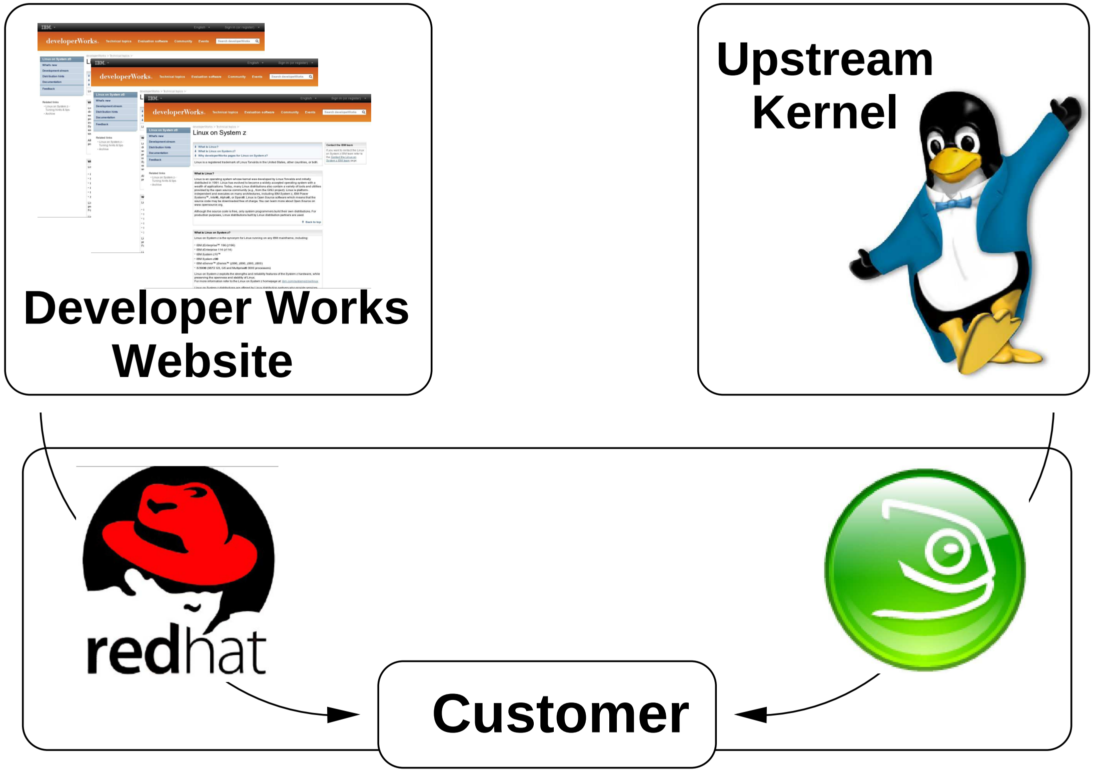
Foster and Protect the Ecosystem

- Software Freedom Law Center
- Free Software Foundation (FSF)
- and more . . .

IBM Linux Development Process

IBM Linux on System z development contributes in the following areas

- kernel
- s390-tools
- Open source tools (e.g. eclipse)
- gcc and glibc
- binutils



Developer Works Website

Upstream Kernel

Customer

Distributions

- SUSE Linux Enterprise Server
  - SLES 10 Service Pack 4 (GA 05/2011) end of regular life cycle
  - SLES 11 (GA 03/2009) kernel 2.6.32 gcc 4.3.3
    - Service Pack 3 (GA 07/2013) kernel 3.0.93
- Red Hat Enterprise Linux
  - RHEL 4 Update 9 (GA 02/2011) end of regular life cycle
  - RHEL 5 Update 9 (GA 01/2013)
  - RHEL 6 (GA 11/2010) kernel 2.6.32 gcc 4.4.7
    - Update 4 (GA 02/2013)
- Others
  - Debian
  - Slackware

Supported Linux Distributions

	zEnterprise EC12 & BC12	zEnterprise z196 & z114	System z10	System z9	zSeries
RHEL 6	✓*	✓	✓	✓	X
RHEL 5	✓*	✓	✓	✓	✓
RHEL 4	X	✓*	✓	✓	✓
SLES 11	✓*	✓	✓	✓	X
SLES 10	✓*	✓	✓	✓	✓
SLES 9	X	✓*	✓	✓	✓

* specific release level recommended or required, some new functions may not be available

see http://www-03.ibm.com/systems/z/os/linux/resources/testedplatforms.html

System z Linux Features - Core

- Enable spinning mutex

  redhat 6.3 11.2

  - Make use of new common code for adaptive mutexes
  - Add new architecture primitive arch_mutex_cpu_relay to exploit sigp sense running to avoid mutex lock retries if hypervisor has not scheduled the CPU holding the mutex

- Jump label support (3.0)

  11.2

  - Branch optimization for conditions that are rarely toggled e.g. tracepoints

- Two stage dumper - kdump support

  redhat 6.3 11.3

  - Uses Preloaded crash kernel
  - Either panic triggered or stand-alone
  - Can reduce dump size
  - Can't dump z/VM Named Saved System (NSS)

System z Linux Features - Core

- Allow to compare dump system with boot system — redhat 6.4 | 11.3
  - z/VM 6.2 allows relocation of guests to other z/VM host systems
  - Provide log of live-guest-relocations in runtime system and dump system for debugging
- Physical memory > 4 TB (kernel 3.3) — 11.3
- libhugetlbfs support — 11.3
  - Enables the transparent use of large pages in C/C++ programs
  - Provide large pages of anonymous data
- Transparent huge page support (kernel 3.7) — 11.2
  - Improve performance in memory intensive applications
  - Reduce number of TLB entries and Page Faults
  - Waste more memory when using

System z Linux Features - Core

- System z hardware counters (kernel 3.4)
  - Counters for running in LPAR
    - basic counter set
    - problem-state counter set
    - crypto-activity
    - counter set,
    - extended counter set with System z10
    - System zEC12 counter (kernel 3.7)
- Compile & disassemble support for zEC12 (kernel 3.8)
  - Add new instructions to the kernel disassembler and allow compiling with -march=zEC12

System z Linux Features - I/O

- End-To-End data consistency checking — redhat 6.4 | 11.2
- Support for hardware data router — redhat 6.4 | 11.3
  - FCP on FICON Express8S
  - Improve performance by reducing path length for data
- Extended DASD statistics — redhat 6.3 | 11.3
  - Add detailed per-device debugging of DASD I/Os via debugfs
  - Useful to analyze problems in particular for PAV and HPF
- Store I/O and initiate logging - SIOSL — redhat 6.1 | 11.2
  - Enhance debug capability for FCP attached devices
  - Enables operating system to detect unusual conditions on a FCP channel

System z Linux Features - I/O

- Safe offline interface for DASD devices (kernel 3.8)
  - Gracefully complete all outstanding I/O requests before a DASD is set offline
- DASD enhancements (kernel 3.11)
  - Add ’timeout’ attribute
  - Implement block timeout handling
  - Number of retries configurable
- Native PCI feature cards (kernel 3.8)
  - Support for native PCIe adapters visible to the operating system

PCI Express

- Native PCIe feature cards introduced on zEC12 and zBC12
  - 10GbE RoCE Express, network card for SMC-R
  - zEDC Express, data compression/decompression card
- Native PCIe adapter concept
  - Plugged into an PCIe I/O drawer
  - Managed by an internal firmware processor (IFP)
  - Device driver for the PCIe function is located in the operating system
- Uses standard Linux PCI support and drivers with some constraints
  - Only MSIX, no port I/O, memory mapped I/O by use of PCI load/store instructions
  - Provides ability to assign individual functions of an adapter to an LPAR
  - Converted System z architecture code to use generic hardirqs
  - Only selected PCIe adapters are known to the IFP and surfaced to the OS

10GbE RoCE Express

- Native PCIe networking card
  - 10 Gigabit remote direct memory access (RDMA) capable network card
  - Uses Infiniband RDMA over Converged Ethernet (RoCE) specification
  - Up to 16 10GbE RoCE Express adapters per machine
  - Reduced latency and lower CPU overhead
  - Supports point-to-point connections and switch connection with an enterprise-class 10 GbE switch
- Software support
  - z/OS V2R1 with PTFs supports SMC-R with RoCE
  - z/VM support planned
  - Linux support in principle available but not available in any distribution yet

zEDC Express

- Native PCIe data compression / decompression card
  - Up to 8 adapters can be installed into a single machine
  - With large blocks, it can compress data at more than 1 GB per second
  - Implements compression as defined by RFC1951 (DEFLATE)
  - Comparable to `gzip -1`
- Software support
  - z/OS V2R1, V1R13 and V1R12 with PTFs
  - The zlib open source library is a C implementation commonly used to provide compression and decompression services

System z Linux Features - Network

- Improved QDIO performance statistics (2.6.33) — 11.2
  - Converts global statistics to per-device statistics and adds adds new counter for the input queue full condition

- QDIO outbound scan algorithm (2.6.38) — 11.2
  - Improve scheduling of QDIO tasklets
  - OSA, HiperSockets and zfcp need different thresholds

- Offload outbound checksumming (2.6.35) — redhat 6.1 — 11.2
  - Move calculation of checksum for non-TSO packets from the driver to the OSA network card

- IPv6 support for the qetharp tool — redhat 6.3 — 11.2
  - Extend the qetharp tool to provide IPv6 information in case of a layer 3 setup
  - Required for communication with z/OS via HiperSockets using IPv6

System z Linux Features - Network

- Support Virtual Ethernet Port Aggregator (VEPA) mode 11.3
  - Send all packages to networking switch to enable external routing
  - Reduce CPU overhead in virtual machine
  - Ensure isolation mode never falls back to non-isolated
  - Check switch supports required configuration modes
- Toleration of optimized latency mode (2.6.35) 11.2
  - 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
- QETH debugging per single card (2.6.36) 11.2
  - Split some of the global QETH debug areas into separate per-device areas
  - Simplifies debugging for complex multi-homed configurations

System z Linux Features - Network

- Change default standard blkt settings for OSA Express — 11.3
- Add OSA concurrent hardware trap — redhat 6.3 — 11.2
  - For better problem determination the qeth driver requests a hardware trace when the device driver or the hardware detect an error
  - Allows correlation between OSA and Linux traces
- AF_IUCV HiperSockets transport (kernel 3.2) — redhat 6.2 — 11.2
  - Use HiperSockets completion queues to control traffic
- Muliple paths with netiucv between z/VM guests (kernel 3.3)
  - Performance improvement with parallel IUCV paths
- Query OSA address table (kernel 3.4)
  - Diagnostic option by gettting a table of physical and logical device information

System z Linux Features - Crypto

- 4096 bit RSA fast path (kernel 2.6.38) — 11.2
  - Make use of 4096 bit RSA acceleration available with Crypto Express3 GA2 cards
- CPACF exploitation of z196 — redhat 6.2 — 11.2
  - Add support for new crypto modes
    - Cipher feedback mode (CFB)
    - Output feedback mode (OFB)
    - Counter mode (CTR)
    - Galois counter mode (GCM)
    - XEX based Tweaked Code Book with Cipher Text Stealing (XTS),
    - Cipher based message authentication mode (CMAC)
    - Counter with cipher block chaining message authentication (CCM)

System z Linux Features - Crypto

- libica APIs for supported crypto modes — redhat 6.2 — 11.2
  - Programmatic way to query for supported crypto ciphers, modes and key sizes
  - Information wether cryptographic features are implemented in hardware or software
- CPACF Support — redhat 6.4 — 11.3
- Crypto Express4S Support — redhat 6.4 — 11.3
- Support the SHA-256 in the opencryptoki CCA token — 11.3

System z Linux Features - Tools

- Fuzzy live dump — 11.3
  - Dump live system without stopping
  - Possibly some data structures are inconsisent
    - But still useful in most cases
- Extend lscpu and add new chcpu tool — redhat 6.4 | 11.3
  - Display CPU topology and CPU state
  - chcpu can change rescan, change state and dispatching mode of CPUs
- SCSI device management tool (s390-tools 1.14.0) — 11.3
  - Tool analog to chccwdev to enable or disable SCSI LUNs addressed by HBA/target port/LUN
- CMSFS user space filesystem support — redhat 6.1 | 11.2

System z Linux Features - Compiler

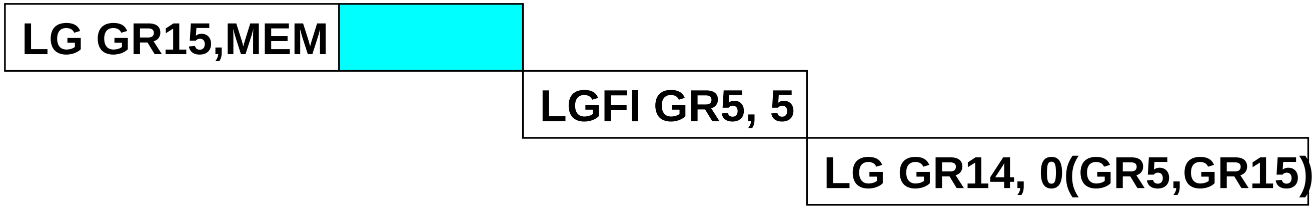
redhat 6.1 | 11.2

- z196 exploitation
  - gcc 4.6
  - Use new instructions -march=z196
  - Use -mtune=z196 to use out-of-order execution
  - Performance improvements with new instructions - needs recompile
  - Use -mtune=z196 to use out-of-order execution

Out of Order Execution

- Change order of instructions that have no dependencies
  - Use wait time to execute other instructions
- Improves instructions with long latencies, like memory access
- Faster Millicode execution

In Order Execution

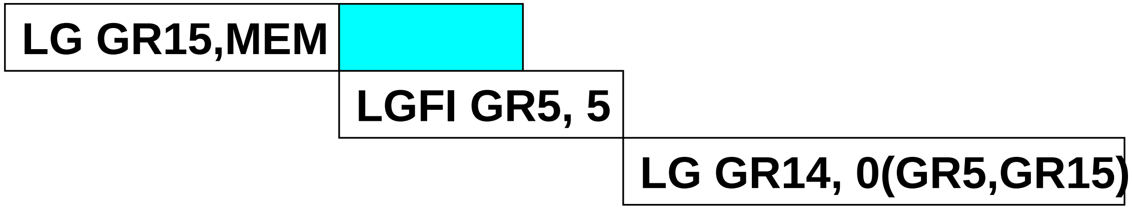


LG GR15,MEM

LGFI GR5, 5

LG GR14, 0(GR5,GR15)

Out of Order Execution



LG GR15,MEM

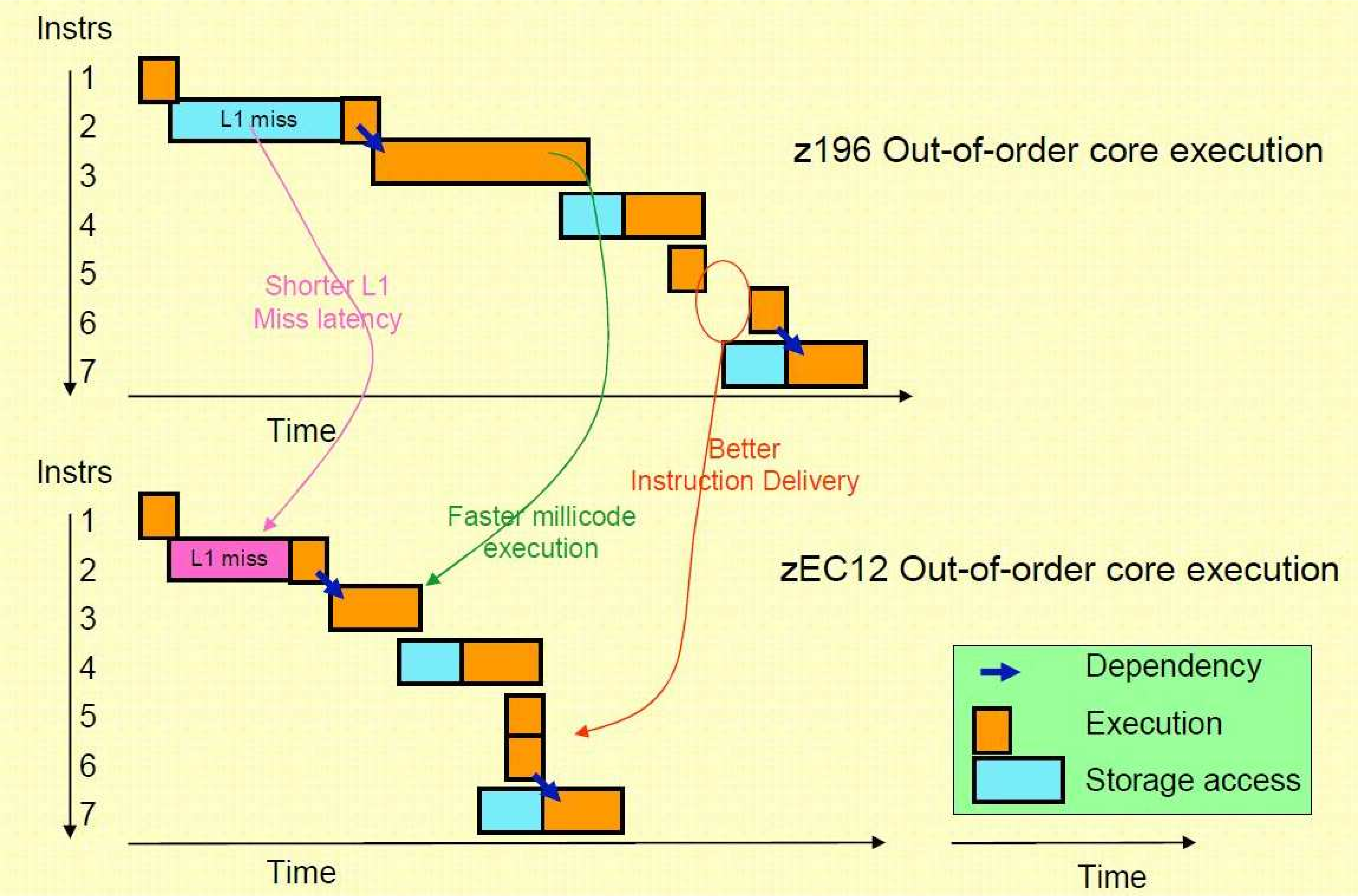
LGFI GR5, 5

LG GR14, 0(GR5,GR15)

Out of Order Execution

● Instruction Fetch	● Instruction Fetch
	● Dispatch to Instruction Queue
● Wait for operands	● Wait for operands
● Dispatch to functional unit	● Dispatch to functional unit
● Execute instruction	● Execute instruction
	● Queue Results
● Write back results to register file	● Write back results to register file

Out of Order Execution



System z Linux Features - zEC12 support

- Flash Express
  - Internal Solid State Disk
  - Up to 4 pairs of cards with max 6.4 TB
  - Concurrent update (kernel 3.8)

redhat 6.4 | 11.3

- Crypto Express4S
  - Indicates capabilities through bit field
- Compiler (gcc 4.8)
  - New instructions
  - Optimization for instruction pipeline

redhat 6.4 | 11.3

- Runtime instrumentation support

redhat 6.4 | 11.3

System z Linux Features - zEC12 support

redhat 6.4

11.3

- Transactional Execution Facility
  - Also known as hardware transactional memory
  - CPU features that allows to execute a group of instructions atomically
  - Optimistic execution, if a transaction conflicts a rollback to a saved state is done

Transactional Execution

- Typical pattern
  1. Lock
  2. Short operation
  3. Unlock

```
spin_lock(&list_lock, 0, 1);
list_add(new, &list_head);
spin_unlock(&list_lock, 1, 0);
```

- Use case
  - Speculative execution
  - Avoid locks for code segments
  - Kernel support required for control register setup
- Transaction abort is expensive

Transactional Execution

```
spin_lock(&list_lock, 0, 1);
list_add(new, &list_head);
spin_unlock(&list_lock, 1, 0);
```

Traditional Code

```
# spin_lock
larl %r3,list_lock
lhi %r1,1
lock: lhi %r0,0
cs %r0,%r1,0(%r3)
ltr %r0,%r0
jne lock
# list_add
larl %r4,list_head
lg %r5,0(%r4)
stg %r4,0(%r2)
stg %r5,8(%r2)
stg %r2,0(%r5)
stg %r2,8(%r4)
# spin_unlock
cs %r1,%r0,0(%r3)
br %r14 br %r14
```

Transaction Execution Code

```
# begin transaction
tbeginc 0,0





# list_add
larl %r4,list_head
lg %r5,0(%r4)
stg %r4,0(%r2)
stg %r5,8(%r2)
stg %r2,0(%r5)
stg %r2,8(%r4)
# end transaction
tend
br %r14
```

s390-tools

- A package with a set of user space utilities to be used with the Linux on System z distributions.
- THE essential tool chain for Linux on System z
- Contains everything from the boot loader to dump related tools for a system crash analysis .
- Contained in all major (and IBM supported) Enterprise Linux distributions which support s390
- RedHat Enterprise Linux
- SUSE Linux Enterprise Server
- Website: http://www.ibm.com/developerworks/linux/linux390/s390-tools.html
- Feedback: linux390@de.ibm.com

s390-tools

chccwdev
chchp
chreipl
chshut
chcrypt
chmem
CHANGE

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

dasdfmt
dasdinfo
dasdstat
dasdview
fdasd
tunedasd
DASD

mon_fsstatd
mon_procd
ziomon
hyptop
MONITOR

ip_watcher
osasnmpd
qetharp
qethconf
NETWORK

tape390_display
tape390_crypt
TAPE

dbginfo
dumpconf
zfcpdump
zfcpdbf
zgetdump
scsi_logging_level
DEBUG

vmconvert
vmcp
vmur
cms-fuse
z/VM

cpuplugd
iucvconn
iucvtty
ts-shell
ttyrun
MISC

zipl
BOOT

s390-tools

- Dump on panic - prevent reIPL loop (1.8.4)
  - Delay arming of automatic reIPL after dump
  - Avoids dump loops where the restarted system crashes immediately
- Automatic menu support in zipl (1.11.0)
  - zipl option to create a boot menu for all eligible non-menu sections in zipl.conf
- re-IPL from device-mapper devices (1.12.0)
  - Automatic reIPL 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
- Configuration tool for System z network devices (1.8.4)
  - Provide a shell script to ease configuration of System z network devices

s390-tools

- Safe offline feature for DASD devices (1.21.0)
- Add Flash Express support to lscss (1.20.0)
- Live Dump support for zgetdump (1.19.0)
  - Use /dev/mem as source dump
  - creation of live dumps in all supported target formats
- Queury OSA address table with qethqoat (1.18.0)
  - Display physical and logical device information
- Support for stand-alone kdump (1.18.0)
- Support for AF_IUCV Completion Queue (1.17.0)
  - New hsuid attribute for lsqeth

Common Kernel News

- btrfs
  - Reduce CPU contention while waiting for delayed extent operations (3.9)
  - Reduce lock contention on extent buffer locks (3.9)
  - Smaller, more space-efficient extent tree (3.10)
  - Offline data deduplication support in btrfs (3.12)
- ext4
  - Add punching hole support for non-extent-mapped files (3.9)
- NFS
  - Parallel NFS (pNFS)
  - NFS Server Side Copy (SSC)

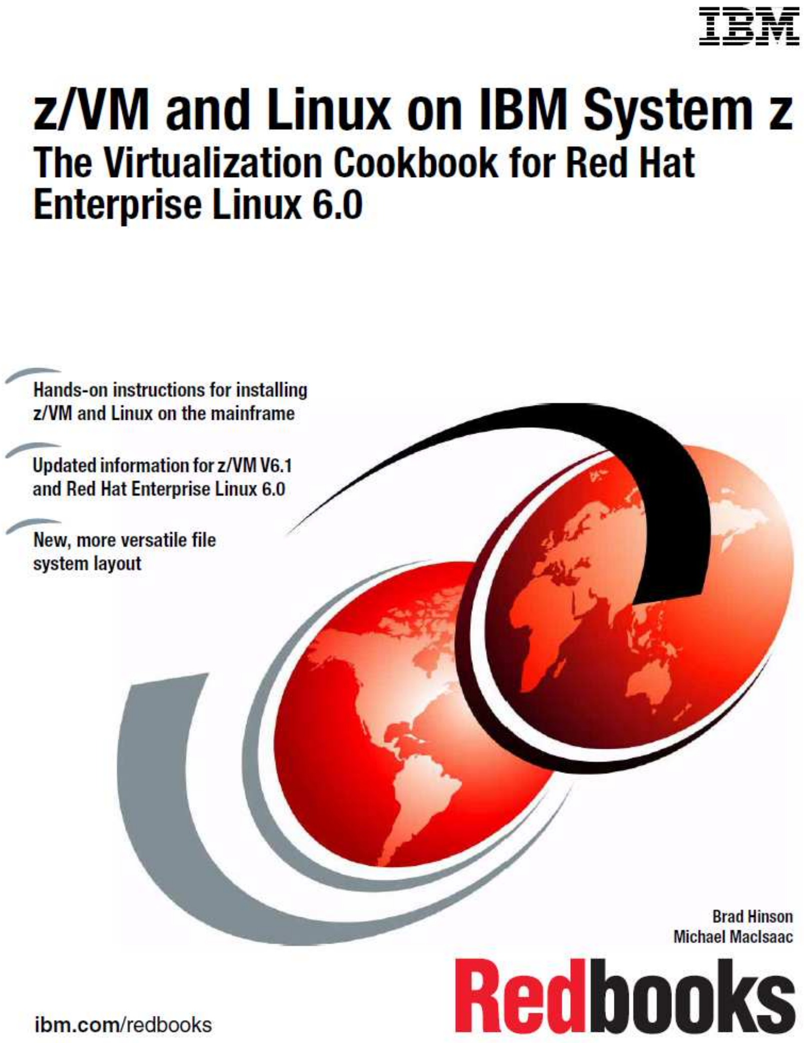
Common Kernel News

- Locking
  - Implement writer lock-stealing for better scalability (3.9)
  - Add support for wound/wait style locks (3.10)
  - Mutex locking scalability improvements (3.10)
  - Improved locking performance for virtualized guests (3.12)
  - New lockref locking scheme, VFS locking improvements (3.12)
  - Improved tty layer locking (3.12)
  - IPC locking improvements (3.12)
- Multiprocessor and Virtualization
  - Add a tuning knob to allow changing SCHED_RR timeslice (3.9)
  - Implement NUMA affinity for unbound workqueues (3.10)
  - Timerless multitasking (3.10)

Common Kernel News

- TCP optimization: Tail loss probe (3.10)
- Better Out-Of-Memory handling (3.12)
- Device mapper target dm-cache allows to use SSD as cache for spinning disk (3.9)

RedBooks



IBM

z/VM and Linux on IBM System z
The Virtualization Cookbook for Red Hat Enterprise Linux 6.0

Hands-on instructions for installing z/VM and Linux on the mainframe

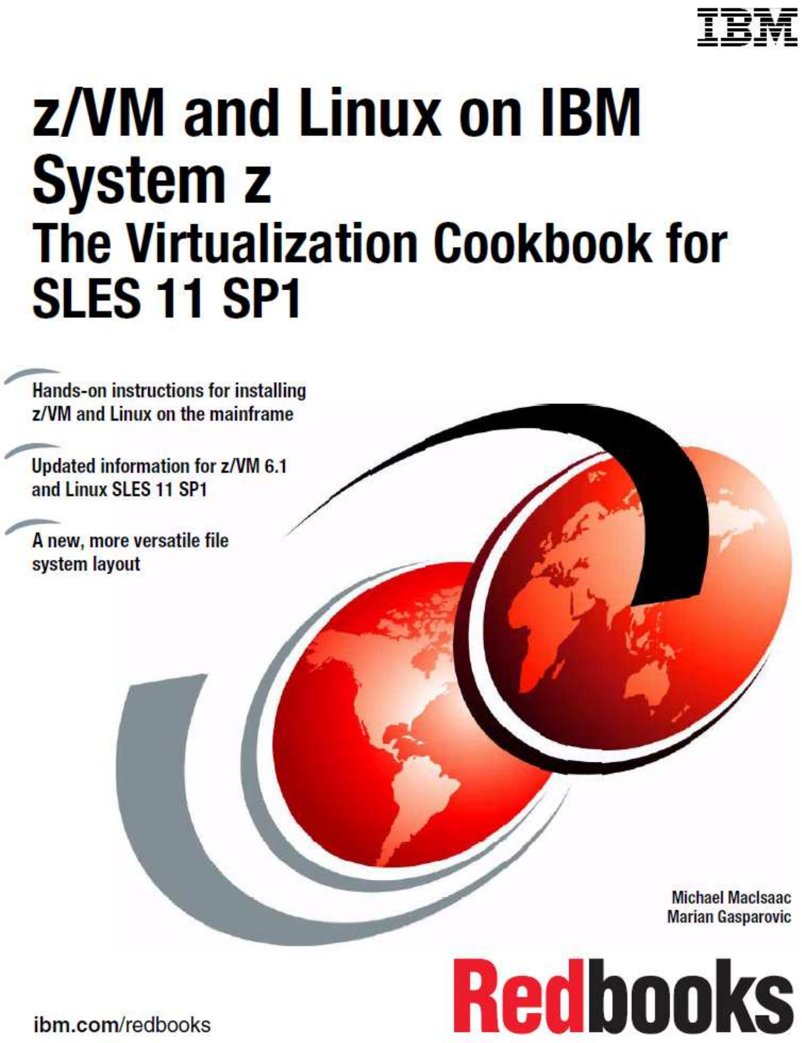
Updated information for z/VM V6.1 and Red Hat Enterprise Linux 6.0

New, more versatile file system layout

Brad Hinson
Michael MacIsaac

ibm.com/redbooks

Redbooks



IBM

z/VM and Linux on IBM System z
The Virtualization Cookbook for SLES 11 SP1

Hands-on instructions for installing z/VM and Linux on the mainframe

Updated information for z/VM 6.1 and Linux SLES 11 SP1

A new, more versatile file system layout

Michael MacIsaac
Marian Gasparovic

ibm.com/redbooks

Redbooks

Links

- developerWorks

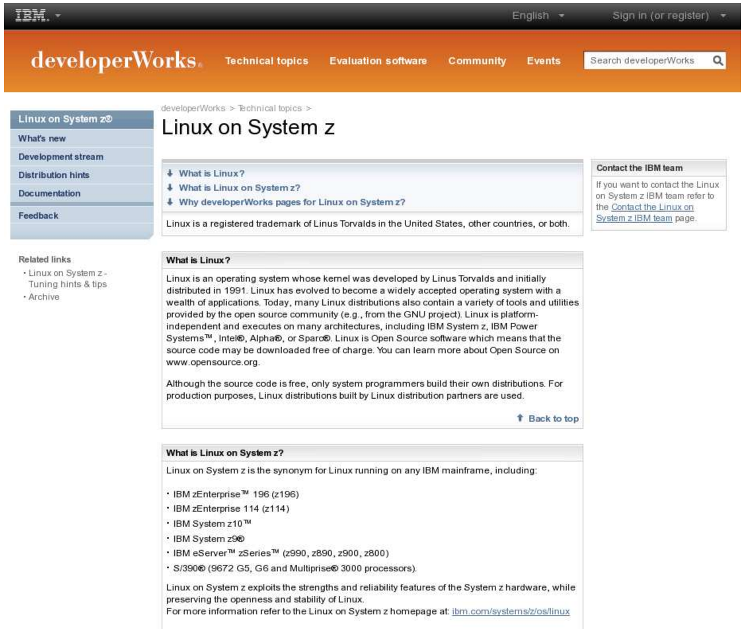
  `http://www.ibm.com/developerworks/linux/linux390`

- Resources for Linux on System z

  `http://www-03.ibm.com/systems/z/os/linux/resources/index.html`

- IBM Redbooks

  `http://www.redbooks.ibm.com`



Linux on System z

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

Contact the IBM team

If you want to contact the Linux on System z IBM team refer to the Contact the Linux on System z IBM team page.

What is Linux?

Linux is an operating system whose kernel was developed by Linus Torvalds and initially distributed in 1991. Linux has evolved to become a widely accepted operating system with a wealth of applications. Today, many Linux distributions also contain a variety of tools and utilities provided by the open source community (e.g., from the GNU project). Linux is platform-independent and executes on many architectures, including IBM System z, IBM Power Systems™, Intel®, Alpha®, or Sparc®. Linux is Open Source software which means that the source code may be downloaded free of charge. You can learn more about Open Source on www.opensource.org.

Although the source code is free, only system programmers build their own distributions. For production purposes, Linux distributions built by Linux distribution partners are used.

What is Linux on System z?

Linux on System z is the synonym for Linux running on any IBM mainframe, including:

- IBM zEnterprise™ 196 (z196)
- IBM zEnterprise 114 (z114)
- IBM System z10™
- IBM System z9®
- IBM eServer™ zSeries™ (z990, z890, z900, z800)
- S/390® (9672 G5, G6 and Multiprise® 3000 processors).

Linux on System z exploits the strengths and reliability features of the System z hardware, while preserving the openness and stability of Linux.
For more information refer to the Linux on System z homepage at: ibm.com/systems/z/os/linux

Thank You !

- Martin Schwidefsky
- Einar Lueck

Questions ?

Dr. Stefan Reimbold
Diplom-Physiker

Linux on System z Service

Schoenaicher Strasse 220
D-71032 Boeblingen
Mail: Postfach 1380
D-71003 Boeblingen

Phone +49-7031-16-2368
Stefan.Reimbold@de.ibm.com