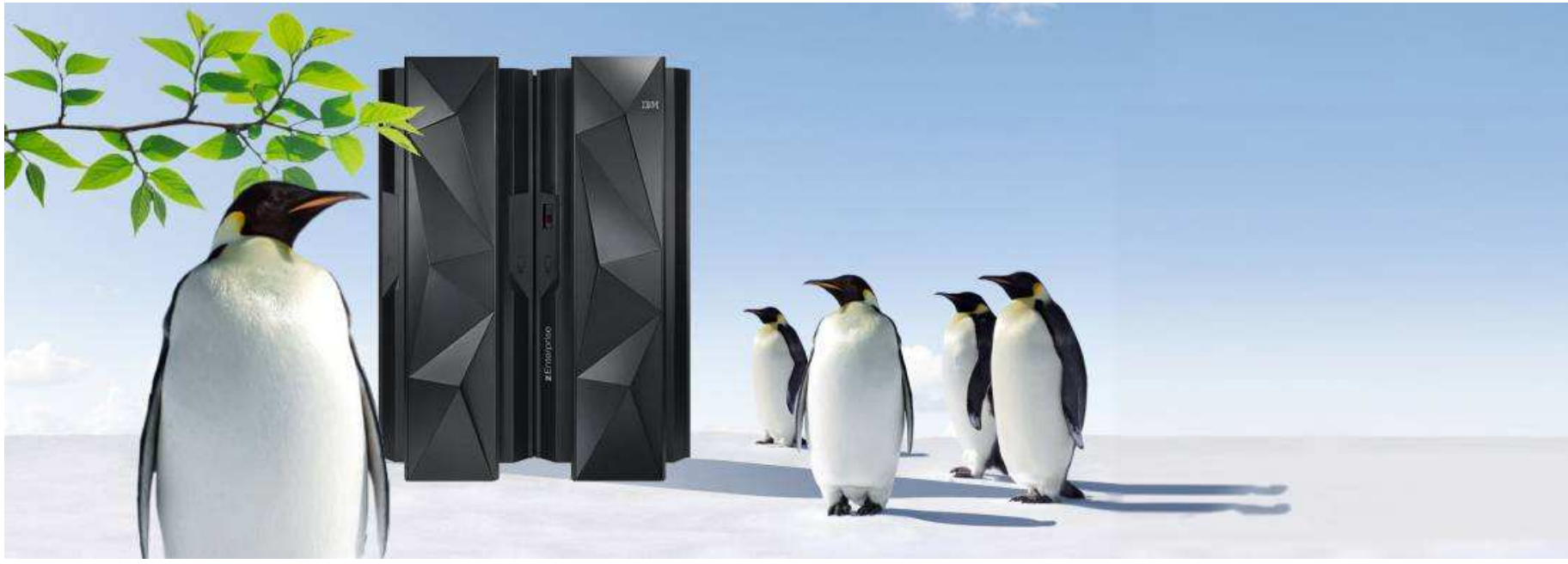
Dr. Stefan Reimbold
IBM Research & Development

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

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

Expanding the OpenSource Ecosystem

- Apache
- Eclipse
- Firefox
- OpenOffice

Promoting Open Standards & Community Collaboration

- The Linux Foundation
- Linux Standards Base
- Common Criteria Certificati-on

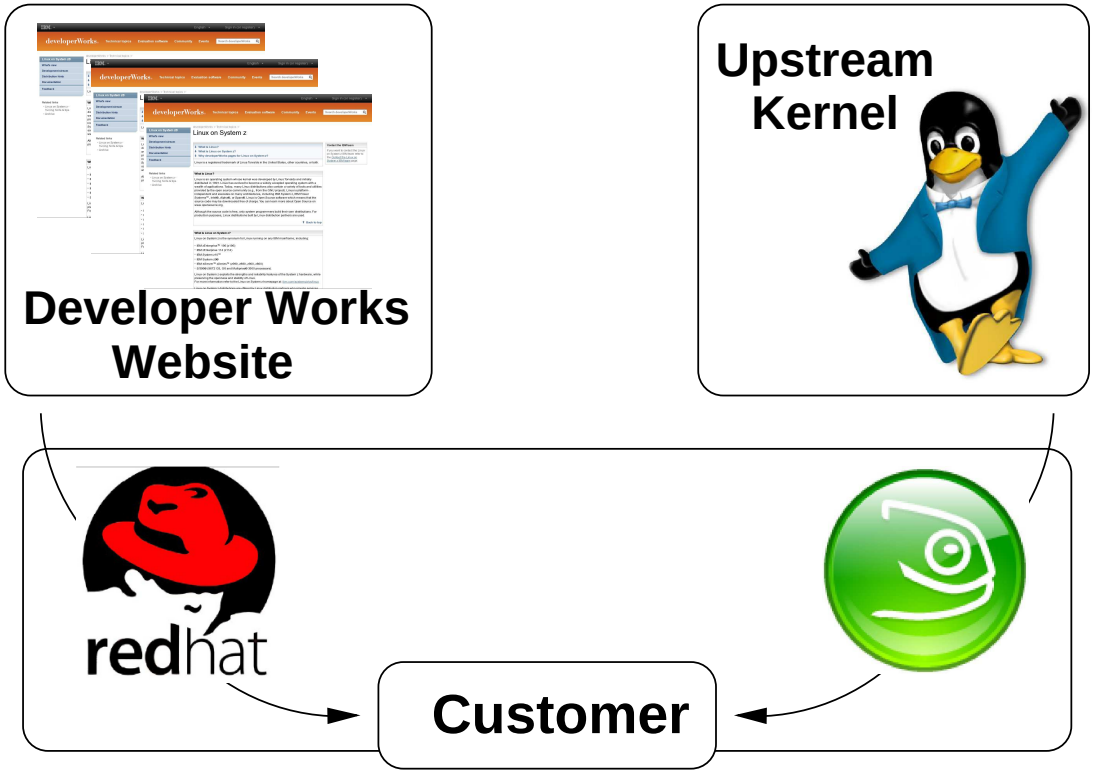
Foster and Protect the Ecosystem

- Software Freedom Law Center
- Free Softwa-re Foundation (FSF)

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



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 10 (GA 10/2013)
  - RHEL 6 (GA 11/2010) kernel 2.6.32 gcc 4.4.7
    - Update 5 (GA 11/2013)
- 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

SUSE SLES 12

- systemd
- Architecture level set for IBM System z196 and newer
- Default file system btrfs
- Linux Containers (since SLES 11)

Red Hat RHEL 7

- systemd
- journald
- Architecture level set for IBM System z196 and newer
- Default file system xfs
- Linux Containers
- MariaDB replaces MySQL

systemd

```
root> service sshd status
openssh-daemon (pid  3045) is running...
```

```
root> systemctl status sshd
sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled)
   Active: active (running) since Mon 2014-03-31 16:53:38 CEST; 3 days ago
  Process: 800 ExecStartPre=/usr/sbin/sshd-keygen (code=exited, status=0/SUCCESS)
 Main PID: 808 (sshd)
   CGroup: /system.slice/sshd.service
           ??808 /usr/sbin/sshd -D

Apr 03 11:27:11 p2330008 sshd[39308]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39318]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39337]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39362]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39382]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39391]: Accepted publickey for root from 9.15...68
Apr 03 11:27:12 p2330008 sshd[39401]: Accepted publickey for root from 9.15...68
Apr 03 11:27:13 p2330008 sshd[39420]: Accepted publickey for root from 9.15...68
Apr 03 11:27:13 p2330008 sshd[39429]: Accepted publickey for root from 9.15...68
Apr 04 11:32:23 p2330008 sshd[40968]: Accepted password for root from 9.164...h2
Hint: Some lines were ellipsized, use -l to show in full.
```

systemd

```
root> service sshd status
Redirecting to /bin/systemctl status  sshd.service
sshd.service - OpenSSH server daemon
...
```

```
root> chkconfig --list
cpi             0:off   1:on    2:on    3:on    4:on    5:on    6:off
cpuplugd        0:off   1:off   2:off   3:off   4:off   5:off   6:off
dumpconf        0:on    1:on    2:on    3:on    4:on    5:on    6:on
iprdump         0:off   1:off   2:on    3:on    4:on    5:on    6:off
iprinit         0:off   1:off   2:on    3:on    4:on    5:on    6:off
iprupdate       0:off   1:off   2:on    3:on    4:on    5:on    6:off
mon_statd       0:off   1:off   2:on    3:on    4:on    5:on    6:off
netconsole      0:off   1:off   2:off   3:off   4:off   5:off   6:off
network         0:off   1:off   2:on    3:on    4:on    5:on    6:off
rhnsd           0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

systemd

```
root> systemctl list-unit-files
UNIT FILE                                   STATE
proc-sys-fs-binfmt_misc.automount           static
dev-hugepages.mount                         static
dev-mqueue.mount                            static
proc-sys-fs-binfmt_misc.mount               static
sys-fs-fuse-connections.mount               static
sys-kernel-config.mount                     static
sys-kernel-debug.mount                      static
tmp.mount                                   disabled
...
initrd-switch-root.service                  static
initrd-udevadm-cleanup-db.service           static
iscsi.service                               enabled
iscsid.service                              disabled
iscsiuio.service                            disabled
kdump.service                               enabled
kmod-static-nodes.service                   static
...
```

systemd

OLD

```
root> chkconfig  httpd on
```

```
root> chkconfig  httpd off
```

NEW

```
root> systemctl enable httpd.service
```

```
root> systemctl disable httpd.service
```

systemd

OLD

```
root> init 3
```

NEW

```
root> systemctl isolate runlevel3.target
```

OR

```
root> systemctl isolate multi-user.target
```

change to runlevel 5

```
root> systemctl isolate runlevel5.target
```

OR

```
root> systemctl isolate graphical.target
```

systemd

Setting default runlevel 3

```
root> systemctl enable multi-user.target --force
```

journald

OLD

```
root> tail -f /var/log/messages
```

NEW

```
root> journalctl -f
```

OLD

```
root> dmesg
```

NEW

```
root> journalctl -k
```

System z Linux Features - Core

- Enable spinning mutex  
  - 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 (kernel 3.0) 
  - Branch optimization for conditions that are rarely toggled e.g. tracepoints
- Two stage dumper - kdump support  
  - 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
  - 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)

- libhugetlbfs support
  - Enables the transparent use of large pages in C/C++ programs
  - Provide large pages of anonymous data

- Transparent huge page support
  - 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

  - 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

  - 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
  

- Support for hardware data router
  

  - FCP on FICON Express8S
  - Improve performance by reducing path length for data
- Extended DASD statistics
  

  - Add detailed per-device debugging of DASD I/Os via debugfs
  - Useful to analyze problems in particular for PAV and HPF

System z Linux Features - I/O

- Safe offline interface for DASD devices
  - 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

System z Linux Features - Network

- IPv6 support for the qetharp tool

  - 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

- Support Virtual Ethernet Port Aggregator mode VEPA

  - 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

- Support for HiperSockets bridgeport (kernel 3.14)

System z Linux Features - Network

- QETH debugging per single card (2.6.36) 

  - Split some of the global QETH debug areas into separate per-device areas
  - Simplifies debugging for complex multi-homed configurations
- Query OSA address table (kernel 3.4)
  - Diagnostic option by gettting a table of physical and logical device information
- Change default standard blkt settings for OSA Express 

- Muliple paths with netiucv between z/VM guests (kernel 3.3)
  - Performance improvement with parallel IUCV paths

System z Linux Features - Network

- 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
- Add OSA concurrent hardware trap 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

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 — 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 — RHEL 6.2, SLES 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 — RHEL 6.4, SLES 11.3
- Crypto Express4S Support — RHEL 6.4, SLES 11.3
- Support the SHA-256 in the opencryptoki CCA token — SLES 11.3
- Support for EP11 coprocessor cards — SLES 12.x

System z Linux Features - Tools

- Fuzzy live dump
  - Dump live system without stopping
  - Possibly some data structures are inconsisent
    - But still useful in most cases

- Extend lscpu and add new chcpu tool
  - 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)
  - Tool analog to chccwdev to enable or disable SCSI LUNs addressed by HBA/target port/LUN

System z Linux Features - Compiler

- 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

System z Linux Features - zEC12 support

- 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
```

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)

- Crypto Express4S
  - Indicates capabilities through bit field

- Compiler (gcc 4.8)
  - New instructions
  - Optimization for instruction pipeline
- Runtime instrumentation support

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

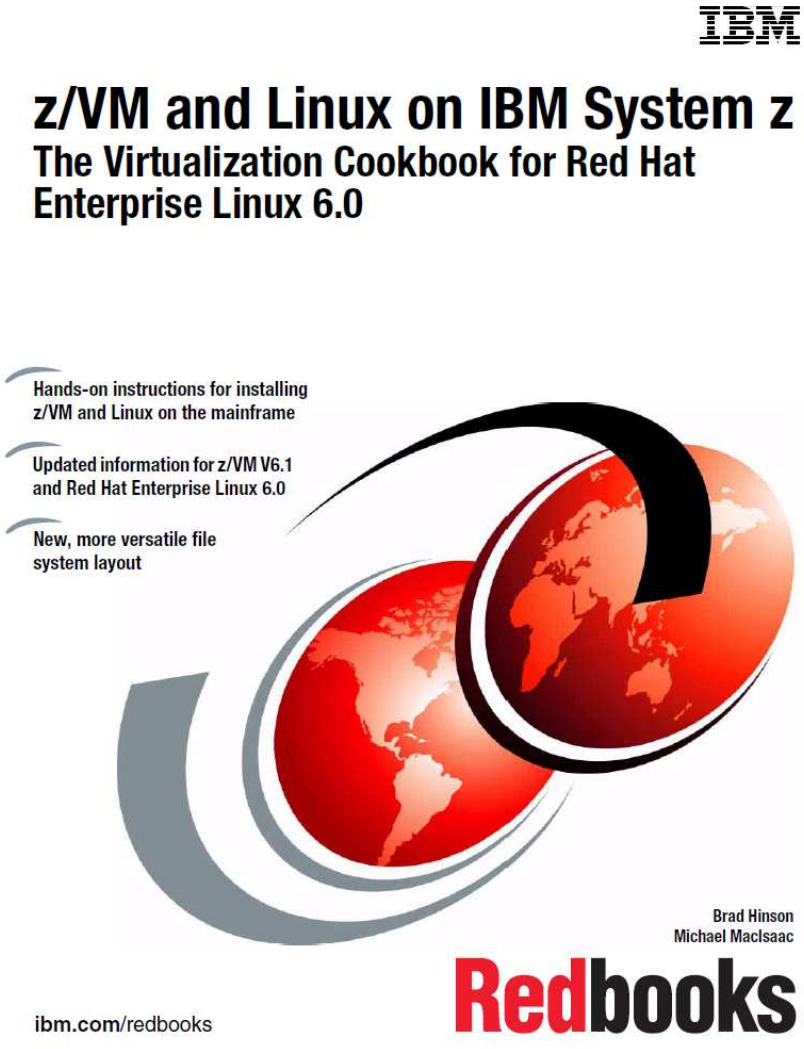
s390-tools

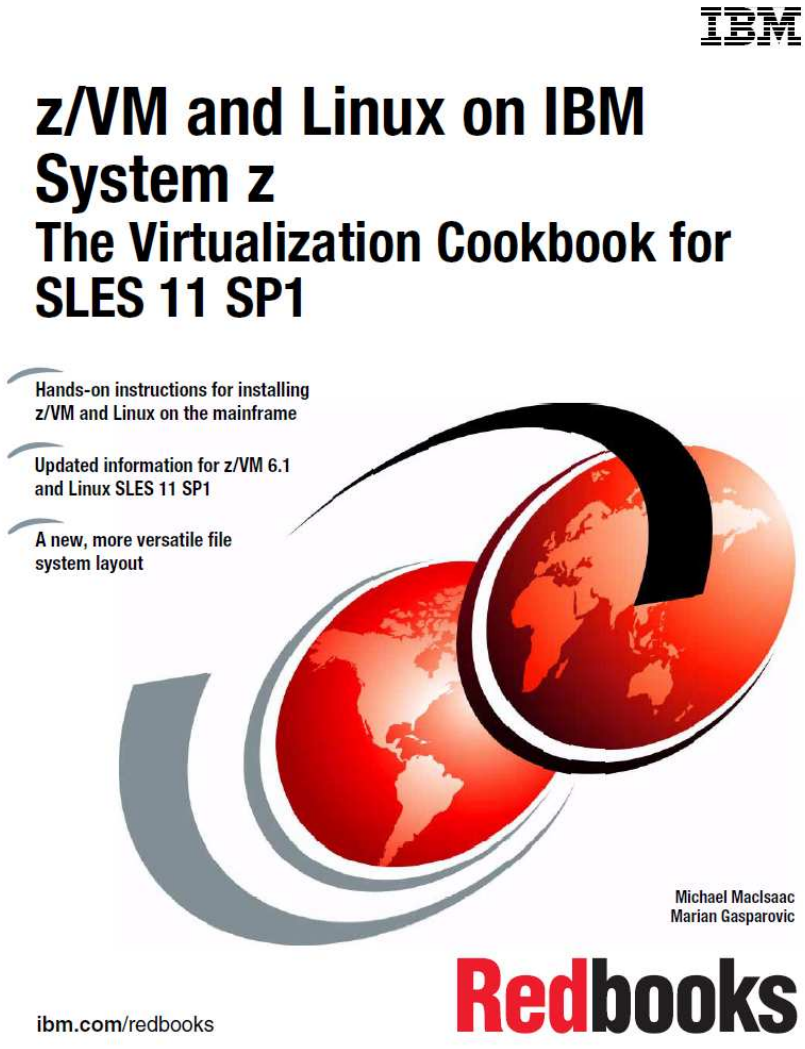
- zdsfs - mount z/OS DASD as Linux file system (1.24.0)
- dasdfmt - better performance(1.23.0)
- Safe offline feature for DASD devices (1.21.0)
- Add Flash Express support to lscss (1.20.0)
- Live Dump support for zgetdump (1.19.0)
  - Use /dev/mem as source dump
  - creation of live dumps in all supported 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

Kernel News

- Better Out-Of-Memory handling (3.12)
- Deadline scheduling class for better real-time scheduling (3.14)
- Kernel address space randomization (3.14)
- nftables, the successor of iptables 3.13)
- Parallel NFS - pNFS (3.11)
- TCP optimization: Tail loss probe (3.10)
  - Reduce latency of short transactions
  - Use fast recovery instead of waiting for retransmission timeout

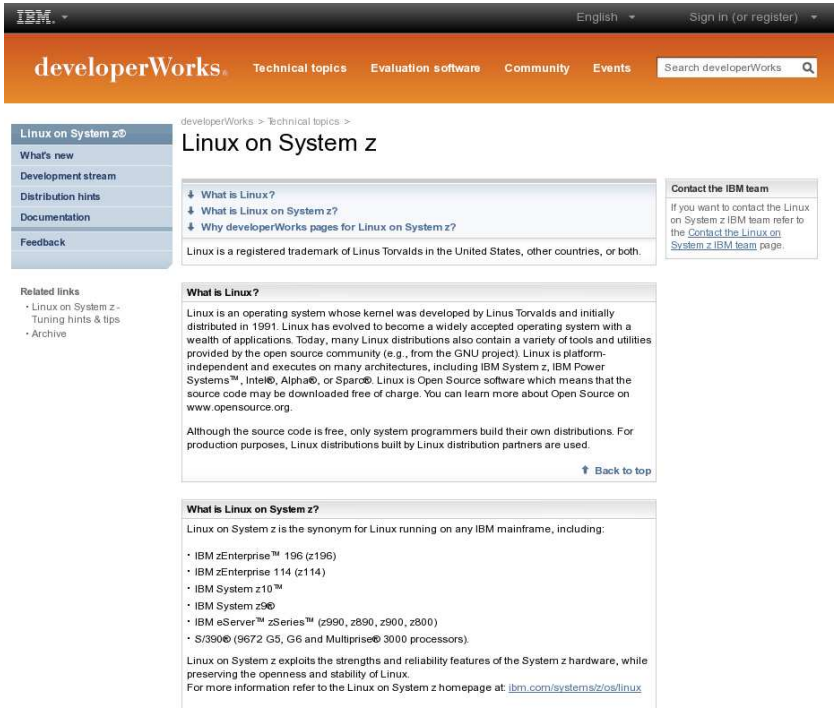
RedBooks





Links

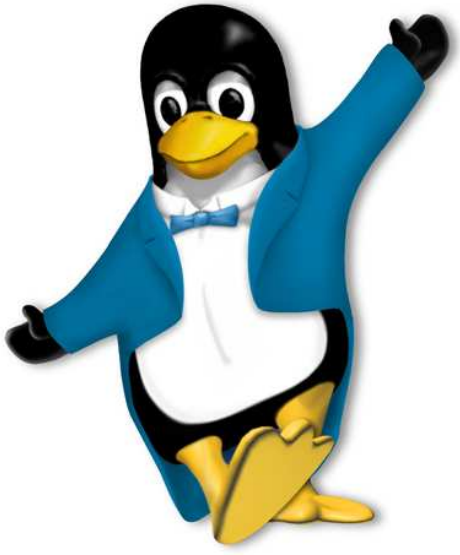
- developerWorks
  `http://www.ibm.com/developerworks/linux/linux390`
- Resources for Linux on System z
  `http://www-03.ibm.com/systems/z/os/linux/resources/index.html`
- Linux Kernel Newbies - Kernel News
  `http://www.kernelnewbies.org`
- IBM Redbooks
  `http://www.redbooks.ibm.com`



Thank You !

- Martin Schwidefsky
- Einar Lueck

Questions ?



Dr. Stefan Reimbold
Diplom-Physiker

Linux on System

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

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