2012
IBM System z Technical University

Enabling the infrastructure for smarter computing

What's New ?
Linux on System z

zLG02

Dr. Stefan Reimbold

© 2012 IBM Corporation

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

© 2012 IBM Corporation

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
- 28 CPU architectures with many machine architectures
- 462 of the Top500 systems running Linux (performance 94.2%)
- 1.73% of desktop clients (browser stats)

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

© 2012 IBM Corporation

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

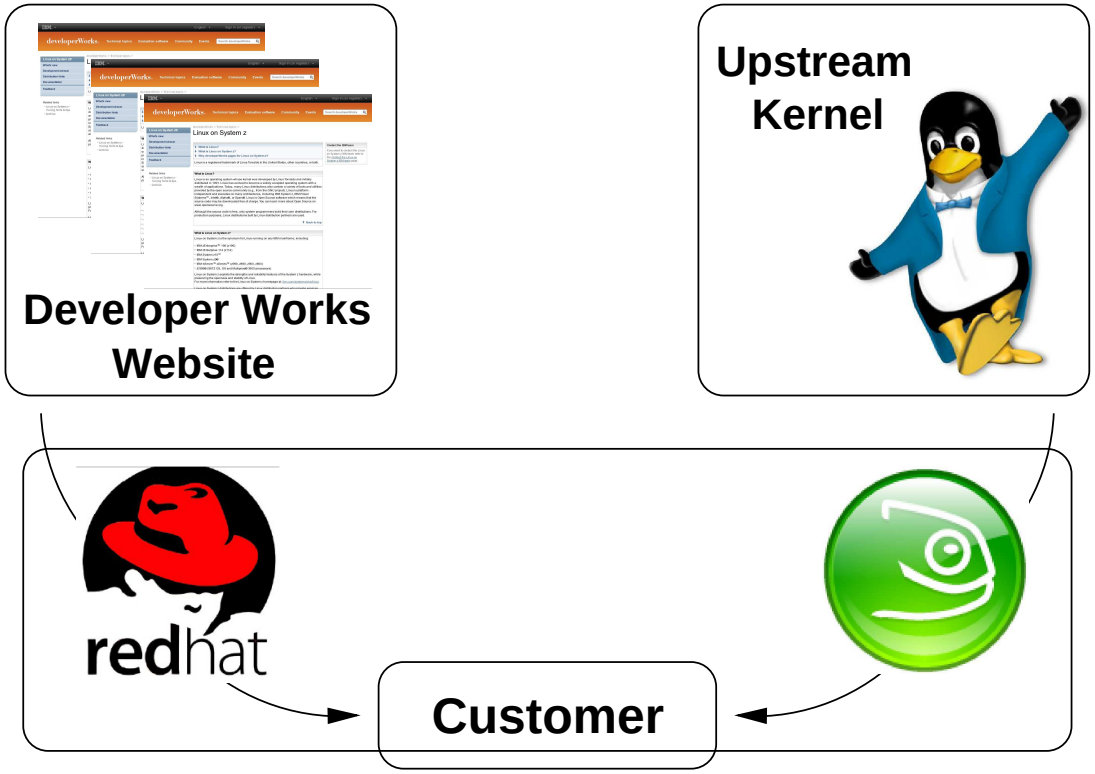
- Software Freedom Law Center
- Free Software Foundation (FSF)
- and more . . .

© 2012 IBM Corporation

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 9 Service Pack 4 (GA 12/2007) end of regular life cycle
  - SLES 10 Service Pack 4 (GA 05/2011)
  - SLES 11 kernel 2.6.32 gcc 4.3.3
    - Service Pack 1 (GA 06/2010) kernel 2.6.32 gcc 4.3.4
    - Service Pack 2 (GA 02/2012) kernel 3.0.13
- Red Hat Enterprise Linux AS
  - RHEL 4 Update 9 (GA 02/2011) end of regular life cycle
  - RHEL 5 Update 8 (GA 02/2012)
  - RHEL 6 (GA 11/2010) kernel 2.6.32 gcc 4.4.0
    - Update 3 (GA 06/2012)
- Others
  - Debian
  - Slackware

© 2012 IBM Corporation

Supported Linux Distributions

	zEnterprise EC12	zEnterprise z114 and z196	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

© 2012 IBM Corporation

System z Linux Features - Core

- breaking event address for user space programs (2.6.35)
  - remember last break in sequential flow of instructions
  - valuable aid in analysis of wild branches
- z196 enhanced node affinity support (2.6.37)
  - allows Linux Scheduler to optimize decisions on z196 topology
- enable spinning mutex (2.6.28)
  - 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
- address space randomization (2.6.38)
  - enable flexible mmap layout for 64 bit to randomize start address for runtime stack and mmap area

© 2012 IBM Corporation

System z Linux Features - I/O

- unit check handling (2.6.35)
  - improve handling of unit checks for internal I/O started by common-I/O layer
  - after a unit check certain setup steps need to be repeated, e.g. for PAV
- dynamic PAV toleration (2.6.35)
  - tolerate dynamic Parallel Access Volume changes for base PAV
  - system management tools can reassign PAV alias device to different base devices
- tunable default grace period for missing interrupts in DASD (2.6.36)
  - provide a user interface to specify the timeout for missing interrupts for standard I/O operations on DASD

© 2012 IBM Corporation

System z Linux Features - I/O

- query DASD reservation status (2.6.37)
  - new DASD ioctl to read the ’Sense Path Group ID’ data
  - allows to determine the reservation status of a DASD in relation to the current system
- multi-track extension for HPF (2.6.38)
  - allows to read from and write to multiple tracks with a single CCW
- access to raw ECKD data from Linux (2.6.38)
  - allows to access ECKD disks in raw mode
  - use ’dd’ command to copy the disk level content of an ECKD disk to a Linux file and vice versa
  - storage array needs to support read-track and write-full-track command

© 2012 IBM Corporation

System z Linux Features - I/O

- store I/O and initiate logging - SIOSL (2.6.36)
  - enhance debug capability for FCP attached devices
  - enables operating system to detect unusual conditions on a FCP channel
- add NPIV information to symbolic port name (2.6.39)
  - add the device bus-ID and the network node to the symbolic port name if the NPIV mode is active
- SAN utilities (2.6.36)
  - two new utilities: zfcp_ping and zfcp_show
  - useful to discover a storage area network

© 2012 IBM Corporation

System z Linux Features - Network

- improved QDIO performance statistics (2.6.33)
  - 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)
  - improve scheduling of QDIO tasklets
  - OSA, HiperSockets and zfcp need different thresholds
- offload outbound checksumming (2.6.35)
  - move calculation of checksum for non-TSO packets from the driver to the OSA network card
- OSX/OSM CHPIDs for hybrid data network (2.6.35)
  - OSA cards for zBX Blade Center Extension will have a new CHPID type
  - allows communication between zBX and Linux on System z

© 2012 IBM Corporation

System z Linux Features - Network

- toleration of optimized latency mode (2.6.35)
  - 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
- NAPI support for QDIO and QETH (2.6.36)
  - convert QETH to the NAPI interface, the ’new’ Linux networking API
  - NAPI allows for transparent GRO (generic receive offload)
- 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

© 2012 IBM Corporation

System z Linux Features - Network

- support for assisted VLAN null tagging (2.6.37)
  - z/OS may sent null-tagged frames to Linux
  - close a gap between OSA and Linux to process null tagged frames correctly
- new default qeth configuration values (2.6.39)
  - receive checksum offload
  - generic receive offload
  - number of inbound buffers

© 2012 IBM Corporation

System z Linux Features - Network

- IPv6 support for the qetharp tool (2.6.38)
  - 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
- add OSA concurrent hardware trap (3.0)
  - 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

© 2012 IBM Corporation

System z Linux Features - Tools

- performance indicator bytes (2.6.37)
  - display capacity adjustment indicator introduced with z196 via /proc/sysinfo
- add support for makedumpfile tool (2.6.34)
  - convert Linux dumps to ELF file format
  - use makedumpfile tool to remove user data from dump
  - multi-volume dump will be removed
- get CPC name (2.6.39)
  - useful to identify a particular hardware system in a cluster
  - CPC name and HMC network name are provided

© 2012 IBM Corporation

CMSFS user space file system support

- allows to mount a z/VM minidisk to a Linux mount point
- z/VM minidisk needs to be in the enhanced disk format (EDF)
- cmsfs fuse file system transparently integrates the files on the minidisk into the Linux VFS, no special command required

```
# cmsfs-fuse /dev/dasde /mnt/cms
# ls -la /mnt/fuse/PROFILE.EXEC
-r--r----- 1 root root 3360 Jun 26 2009 /mnt/cms/PROFILE.EXEC
```

- by default no conversion is performed
  - mount with `-t` to get automatic EBCDIC to ASCII conversion

```
# cmsfs-fuse -t /dev/dasde /mnt/cms
```

© 2012 IBM Corporation

CMSFS user space file system support

- write support is work in progress - almost completed
- use fusermount to unmount the file system again

```
# fusermount -u /mnt/cms
```

- RHEL 6.1 and SLES 11 SP2

© 2012 IBM Corporation

Two stage dumper / kdump support

- use a Linux kernel to create a system dump
  - use a preloaded crashkernel to run in case of a system failure
  - can be triggered either as panic action or by the stand-alone dumper, integrated into the shutdown actions framework
- Pro
  - enhanced dump support that is able to reduce dump size, shared disk space, dump to network, dump to a file-system etc.
  - makedumpfile tool can be used to filter the memory of the crashed system
- Con
  - kdump is not as reliable as the stand-alone dump tools
  - kdump cannot dump a z/VM named saved system (NSS)
  - for systems running in LPAR kdump consumes memory
- kernel 3.2 - s390-tools-1.17.0

© 2012 IBM Corporation

Two stage dumper / kdump support

- add a crashkernel to the kernel command line

```
crashkernel=<size>@<offset>
```

- boot your system and check the reservation

```
# cat /proc/iomem
00000000-3fffffff : System RAM
00000000-005f1143 : Kernel code
005f1144-00966497 : Kernel data
00b66000-014c4e9f : Kernel bss
40000000-47ffffff : Crash kernel
48000000-7fffffff : System RAM
```

- load the kdump kernel with kexec

```
# kexec -p kdump.image initrd kdump.initrd --command-line="dasd=1234 root=/dev/ram0"
```

- manually trigger for kdump under z/VM

```
#cp system restart
```

© 2012 IBM Corporation

Changes Kernel 3.2

- Btrfs
  - faster scrubbing
  - automatic backup of tree roots
  - detailed corruption messages
  - manual inspection of metadata
- ext4
  - support 1 MB block size
- I/O-less dirty throttling - reduce filesystem writeback from page reclaim
- Network
  - TCP Proportional Rate Reduction
- New architecture
  - Hexagon

© 2012 IBM Corporation

Changes Kernel 3.3

- Btrfs
  - restriping between different RAID levels
  - improved balancing
  - improved debugging tools
- Open vSwitch
- teaming
  - Better bonding of network interfaces
- Network
  - Per-cgroup TCP buffer limits
  - Network priority control group
- Better ext4 online resizing
- New architecture
  - TI C6X

© 2012 IBM Corporation

Changes Kernel 3.4

- Btrfs updates
  - repair and data recovery tools
  - metadata blocks bigger than 4KB
  - performance improvements
  - better error handling
- remove *resize* mount option for ext4
  - no longer useful in the age of online resize2fs
- new X32 ABI - 64-bit mode with 32-bit pointers
- Virtualization
  - KVM - several changes including 1 s390 change
  - Hyper-V - several changes
  - Xen - ACPI change and netconsole support
  - virtio-pc - S3 support
  - rpmsg - remote processor message bus

© 2012 IBM Corporation

Changes Kernel 3.5

- Network
  - TCP connection repair
  - relocate a network connection to another host
  - TCP Early Retransmit
- Btrfs
  - I/O failure statistics
  - latency improvements
- task children info in /proc/<pid>/task/<tid>/children
  - useful for process checkpointing or relocation

© 2012 IBM Corporation

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

© 2012 IBM Corporation

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

© 2012 IBM Corporation

s390-tools

CHANGE
- chccwdev
- chchp
- chreipl
- chshut
- chcrypt
- chmem

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

DASD
- dasdfmt
- dasdinfo
- dasdstat
- dasdview
- fdasd
- tunedasd

MONITOR
- mon_fsstatd
- mon_procd
- ziomon
- hyptop

NETWORK
- ip_watcher
- osasnmpd
- qetharp
- qethconf

TAPE
- tape390_display
- tape390_crypt

DEBUG
- dbginfo
- dumpconf
- zfcpdump
- zfcpdbf
- zgetdump
- scsi_logging_level

z/VM
- vmconvert
- vmcp
- vmur
- cms-fuse

MISC
- cpuplugd
- iucvconn
- iucvtty
- ts-shell
- ttyrun

BOOT
- zipl

© 2012 IBM Corporation

LNXHC - Linux Health Checker

- command line tool for Linux.
- to identify potential problems before they impact your system performance, availability or cause outages.
- collect and compare the active Linux settings and system status with the values provided by health-check authors or defined by the customer
- produces detailed messages, which describe potential problems and the suggests solutions
- Linux Health Checker runs on any Linux platform which meets the software requirements
- can be easily extended by writing new health check plug-ins
- The Linux Health Checker is an open source project sponsored by IBM. It is released under the Eclipse Public License v1.0. http://lnxhc.sourceforge.net

© 2012 IBM Corporation

SAN Utilities

- 2 new utilities
  - zfcp_show
  - zfcp_ping
- useful to discover a storage area network
- kernel 2.6.36 - lib-zfcp-hbaapi 2.1

© 2012 IBM Corporation

zfcp_show

- Query Fibre Channel nameserver about ports available for my system

```
# zfcp_show -n
Local Port List:
        0x500507630313c562 / 0x656000 [N_Port] proto = SCSI-FCP FICON
        0x50050764012241e4 / 0x656100 [N_Port] proto = SCSI-FCP
        0x5005076401221b97 / 0x656400 [N_Port] proto = SCSI-FCP
```

- Query SAN topology, requires FC management server access

```
# zfcp_show
Interconnect Element Name 0x100000051e4f7c00
Interconnect Element Domain ID 005
Interconnect Element Type Switch
Interconnect Element Ports 256
       ICE Port 000 Online
              Attached Port [WWPN/ID] 0x50050763030b0562 / 0x650000 [N_Port]
       ICE Port 001 Online
              Attached Port [WWPN/ID] 0x50050764012241e5 / 0x650100 [N_Port]
       ICE Port 002 Online
              Attached Port [WWPN/ID] 0x5005076303008562 / 0x650200 [N_Port]
       ICE Port 003 Offline
```

© 2012 IBM Corporation

zfcp_ping

- Check if remote port responds (requires FC management service access)

```
# zfcp_ping 0x5005076303104562
Sending PNG from BUS_ID=0.0.3c00 speed=8 GBit/s
       echo received from WWPN (0x5005076303104562) tok=0 time=1.905 ms
       echo received from WWPN (0x5005076303104562) tok=1 time=2.447 ms
       echo received from WWPN (0x5005076303104562) tok=2 time=2.394 ms
---------- ping statistics -----------
min/avg/max = 1.905/2.249/2.447 ms
----------------------------------------
```

- zfcp_show and zfcp_ping are part of the zfcp-hbaapi 2.1 package
  http://www.ibm.com/developerworks/linux/linux390/zfcp-hbaapi-2.1.html

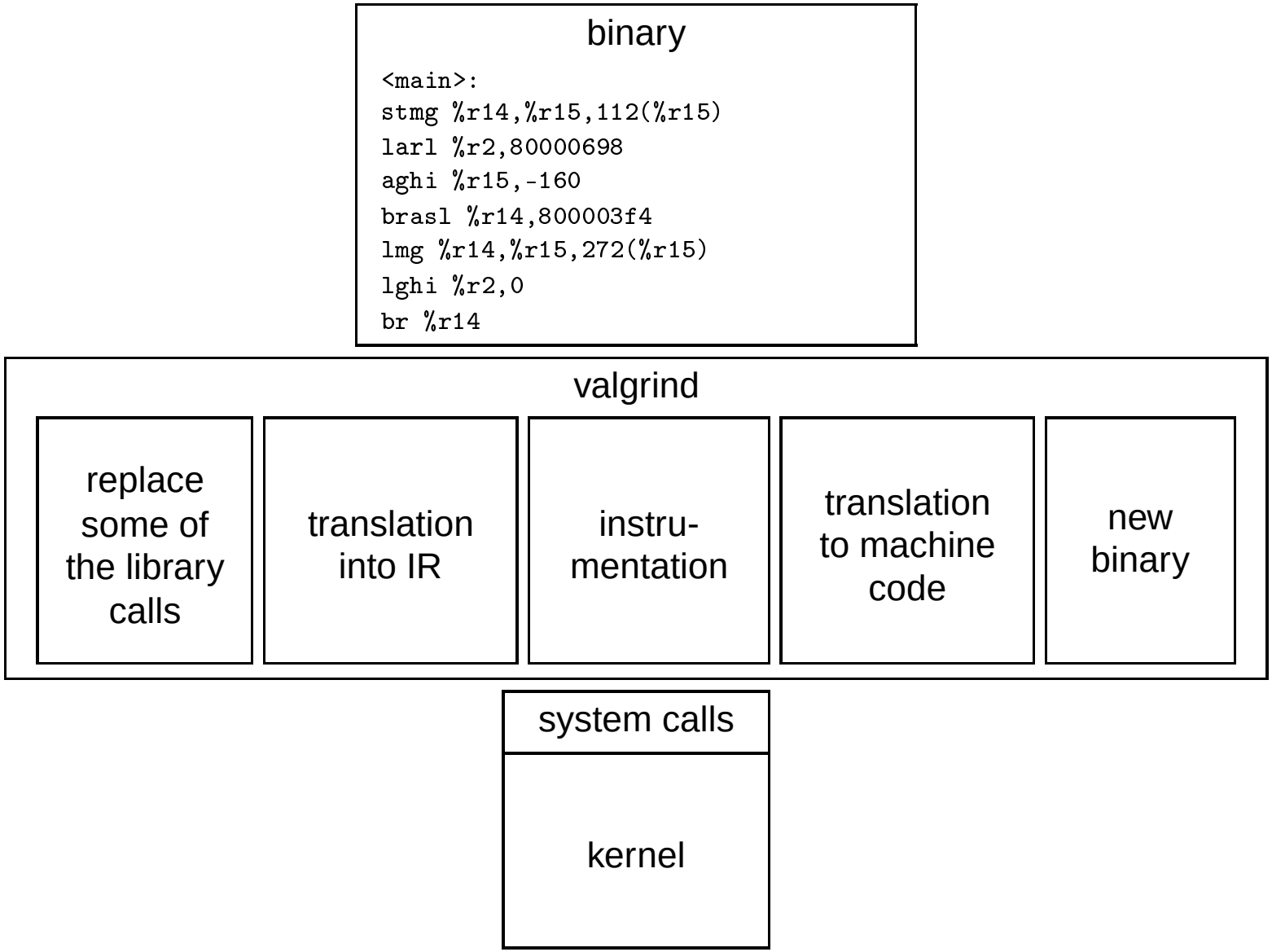
© 2012 IBM Corporation

valgrind System z Support

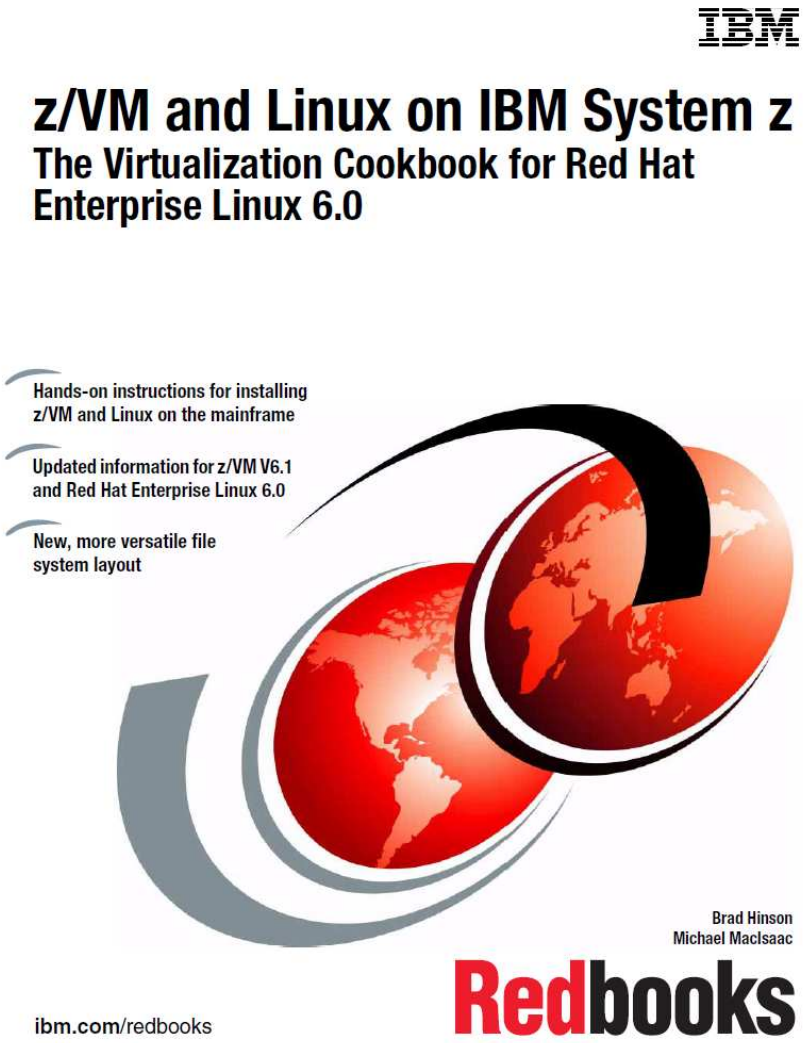
- valgrind -tool=memcheck [–leak-check=full] [–track-origins] <program>
  - detects if your program accesses memory it shouldn't
  - detects dangerous uses of uninitialized values on a per-bit basis
  - detects leaked memory, double frees and mismatched frees
- valgrind -tool=cachegrind
  - profile cache usage, simulates instruction and data cache of the cpu
  - identifies the number of cache misses
  - needs cache line size, Extract Cache Attributes (ECAG) instruction introduced with z10
- valgrind -tool=massif
  - profile heap usage, takes regular snapshots of program's heap
  - produces a graph showing heap usage over time

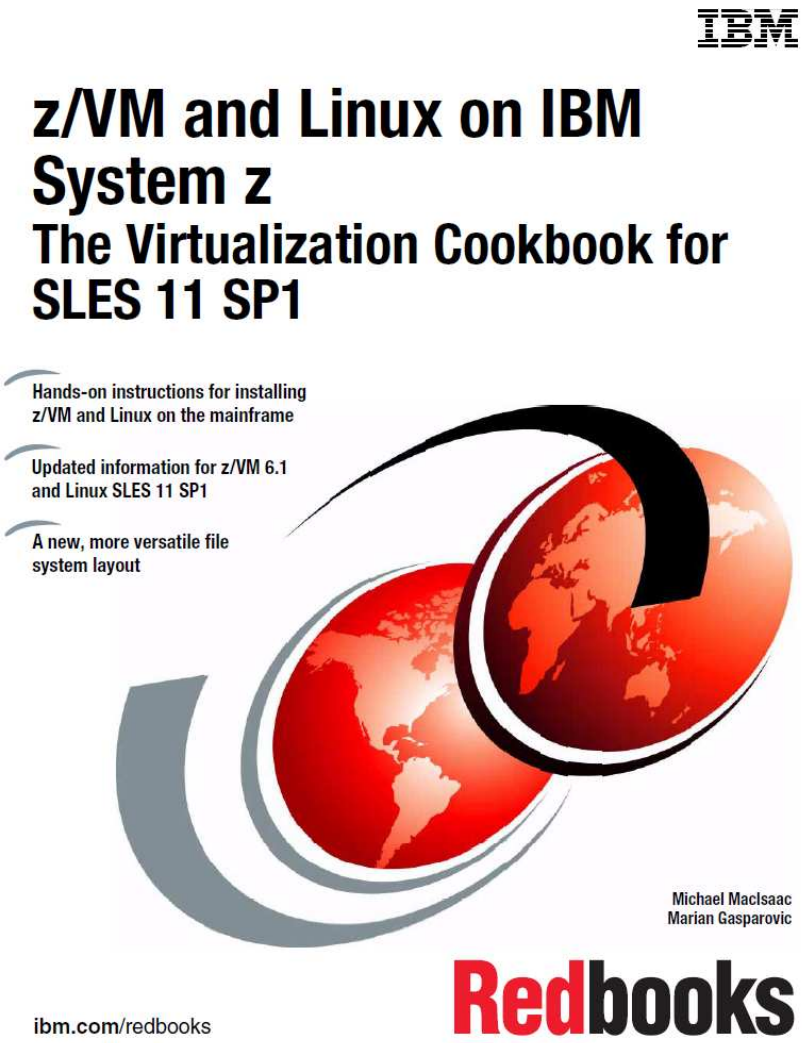
© 2012 IBM Corporation

valgrind System z Support



RedBooks

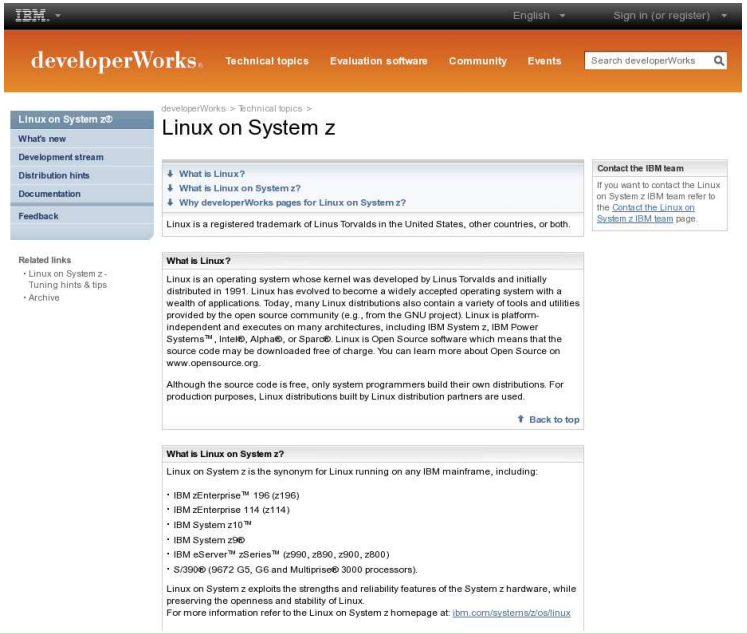




© 2012 IBM Corporation

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`



Thank You !

- Martin Schwidefsky

© 2012 IBM Corporation

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

© 2012 IBM Corporation