

Linux on System z Performance Update - Part 1 z10, CPU and Memory

Mario Held
IBM Research & Development, Germany

August 28, 2009
Session Number 2190



maxi
ove agility save tim
enges colleagues

Trademarks



The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

DB2*	System z
DB2 Connect	Tivoli*
DB2 Universal Database	VM/ESA*
e-business logo	WebSphere*
GDPS*	z/OS*
Geographically Dispersed Parallel Sysplex	z/VM*
HyperSwap	zSeries*
IBM*	
IBM eServer	
IBM logo*	
Parallel Sysplex*	

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Intel is a registered trademark of the Intel Corporation in the United States, other countries or both.

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

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.

* 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.

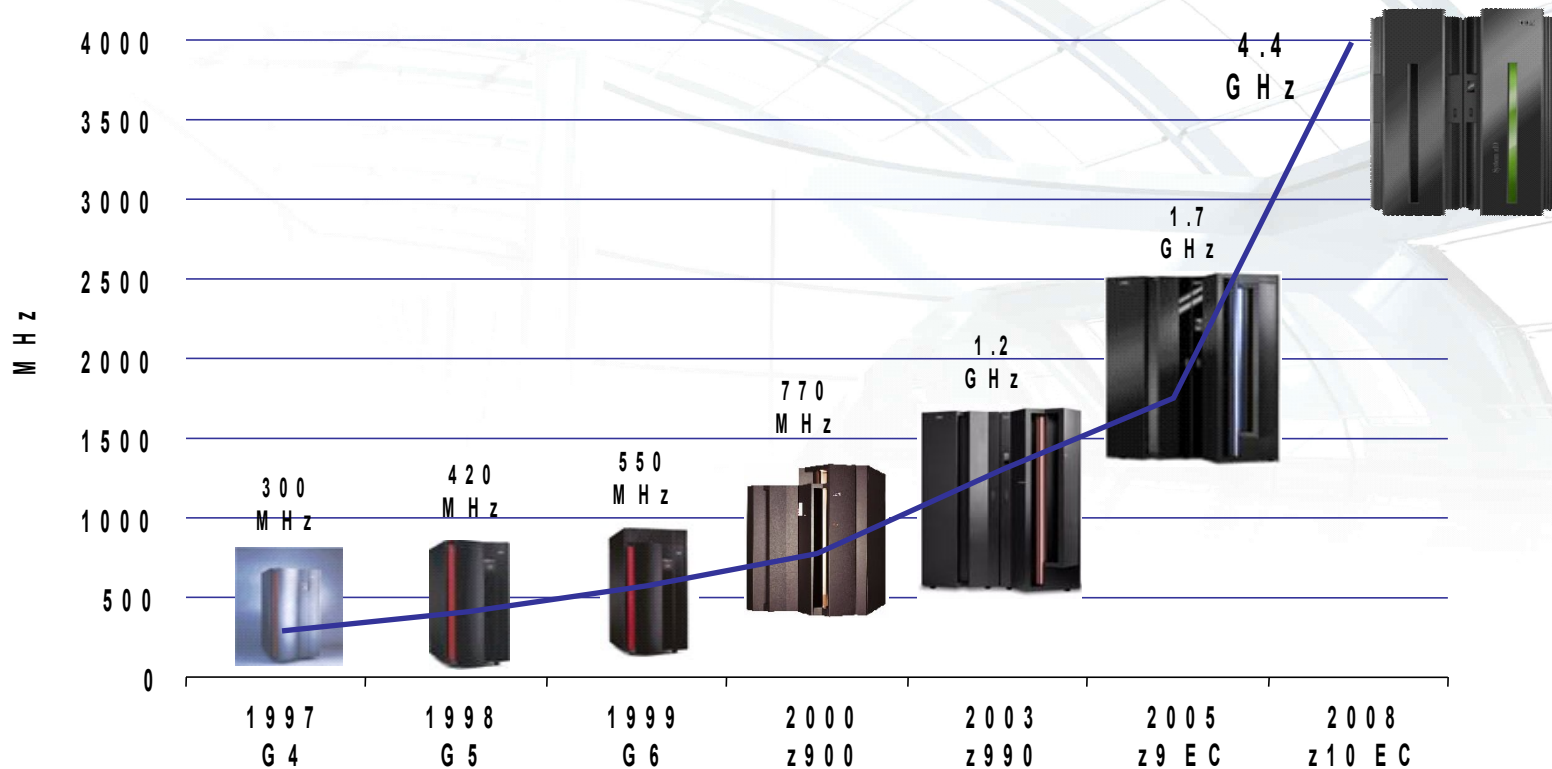
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.



Agenda

- **System z10**
- GCC compiler
- Java
- CPU hotplug
- Oprofile

IBM z10 EC - the CMOS Mainframe Heritage

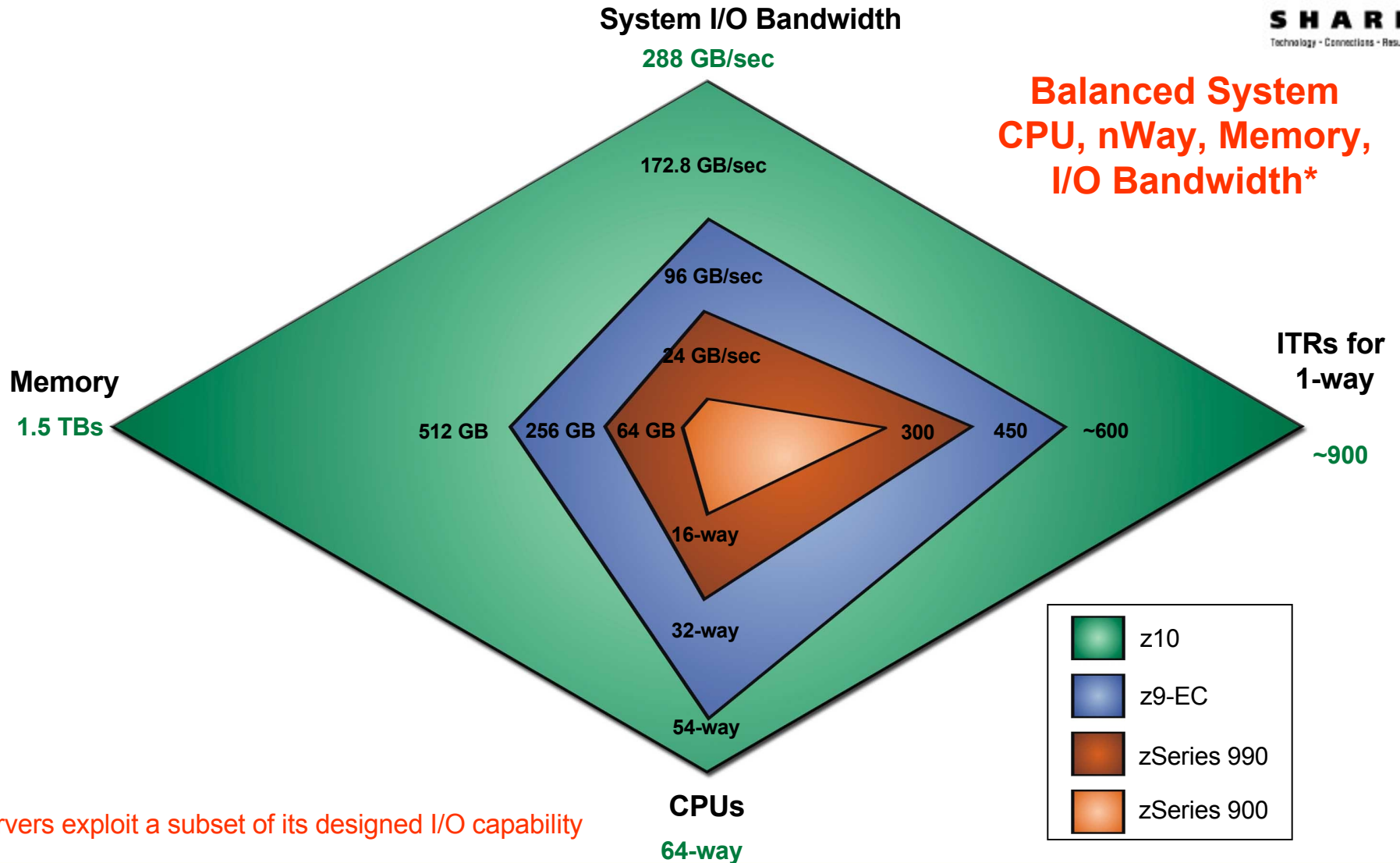


- G 4 - 1st full-custom CMOS S/390®
- G 5 - IEEE-standard BFP; branch target prediction
- G 6 - Cu BEOL

- IBM eServer zSeries 900 (z900) - Full 64-bit z/Architecture®
- IBM eServer zSeries 990 (z990) - Superscalar CISC pipeline
- z9 EC - System level scaling

- z10 EC - Architectural extensions

IBM System z – system design comparison



*Servers exploit a subset of its designed I/O capability

LSPR Mixed Workload for System z10 EC



z10 EC to z9 EC

Ratios

Mixed workload, multi-image with HiperDispatch active on z10 EC!

Uni-processor

1.62

16-way z10 EC to 16-way z9 EC

1.49

32-way z10 EC to 32-way z9 EC

1.49

56-way z10 EC to 54-way z9 EC

1.54

64-way z10 EC to 54-way z9 EC

1.70

File server benchmark description

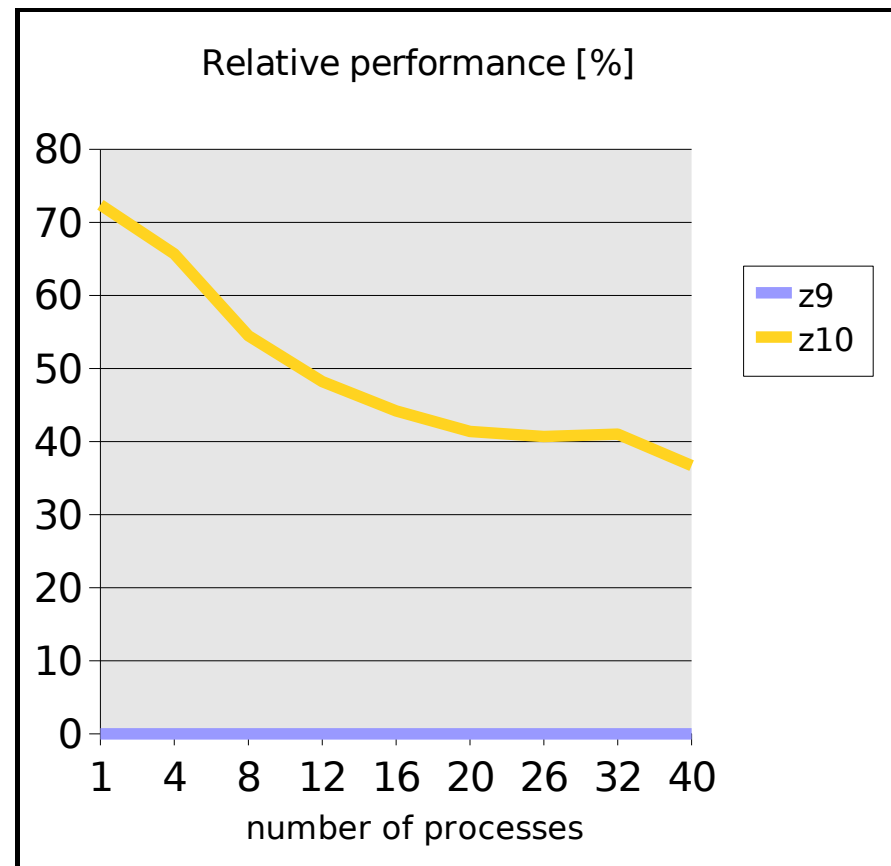
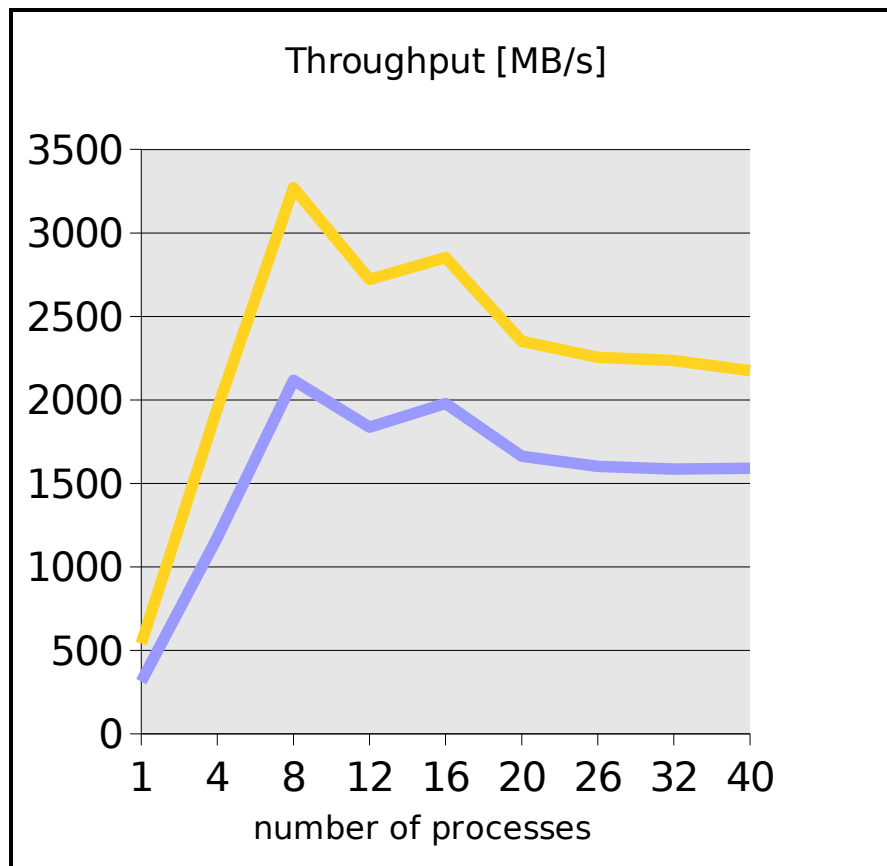


- dbench 3
 - Emulation of Netbench benchmark, rates windows file servers
 - Mainly memory operations
 - Mixed file operations workload for each process: create, write, read, append, delete
 - 8 CPUs and 1, 4, 8, 12, 16, 20, 26, 32, 40 processes
 - 2 GB memory

z10 Performance: dbench 3



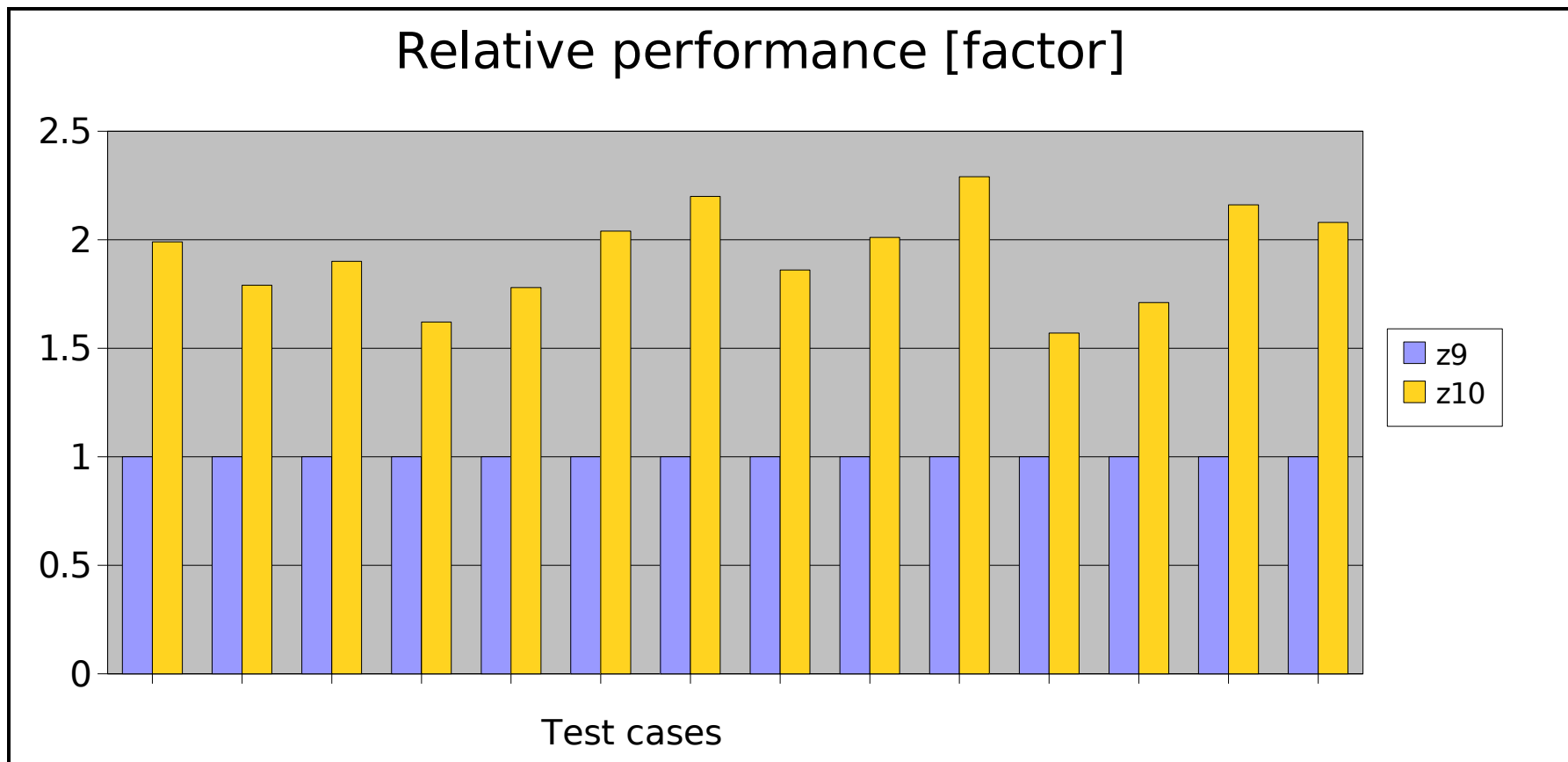
- Improvement z10 versus z9:
 - Measured with 8 CPUs: average improvement is 50%





z10 performance: CPU intensive workloads

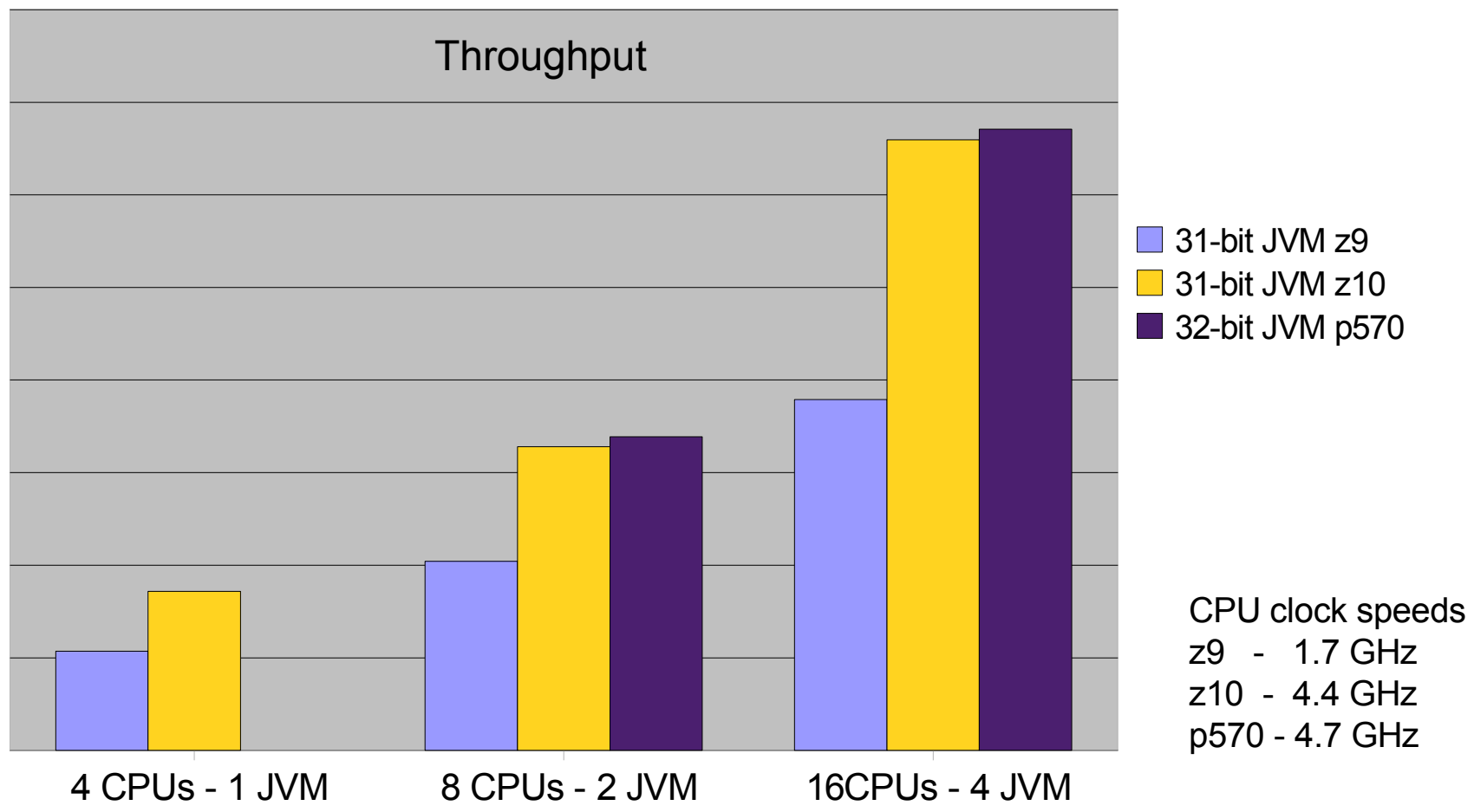
- Overall improvement with z10 versus z9: 1.9x
- gcc-4.3 compiler using -march=z9-109 or -march=z10 option





z10 Performance: Java workload

- System z versus System p, IBM J2RE 1.6.0

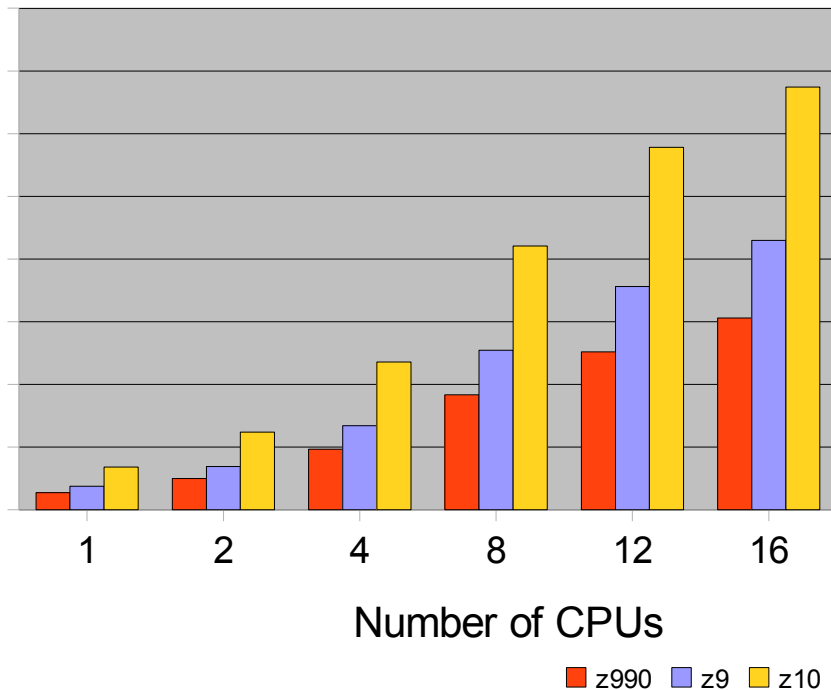


z10 with Informix IDS 11 OLTP workload

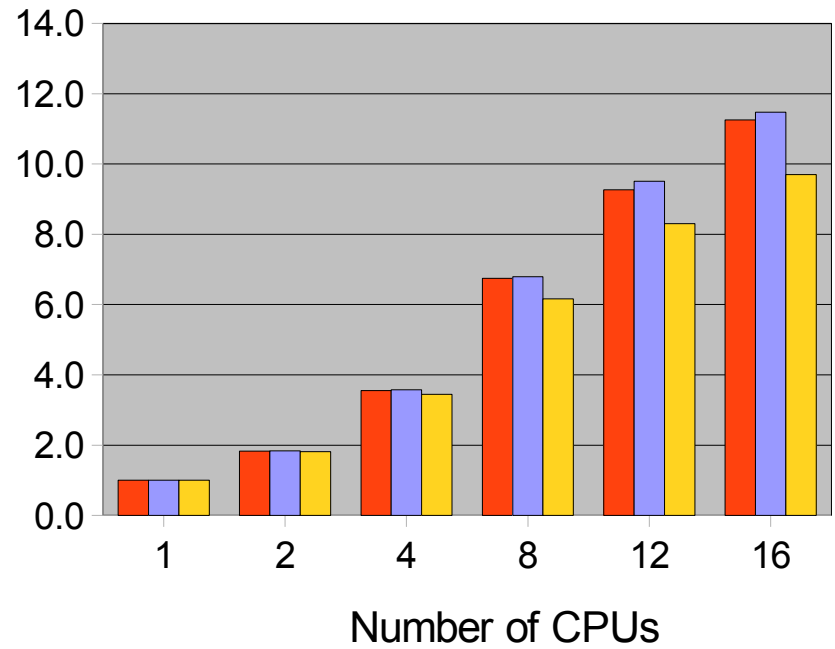


- Throughput improvements
 - z9 to z10: 65% - 82%
 - A number of z10 CPUs can do the same work as the double number of z9 CPUs

Transactions



Scaling factor



z10 performance summary

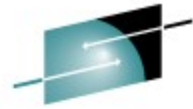


- System z evolution continues
- Sharp performance boost from z9 to z10
- Balanced System
- LSPR expectations met
- Excellent on compute intensive and Java workloads



Agenda

- System z10
- **GCC compiler**
- Java
- CPU hotplug
- Oprofile



GCC compiler evolution

Development makes new GCC features available for System z

Development exploits new System z hardware features

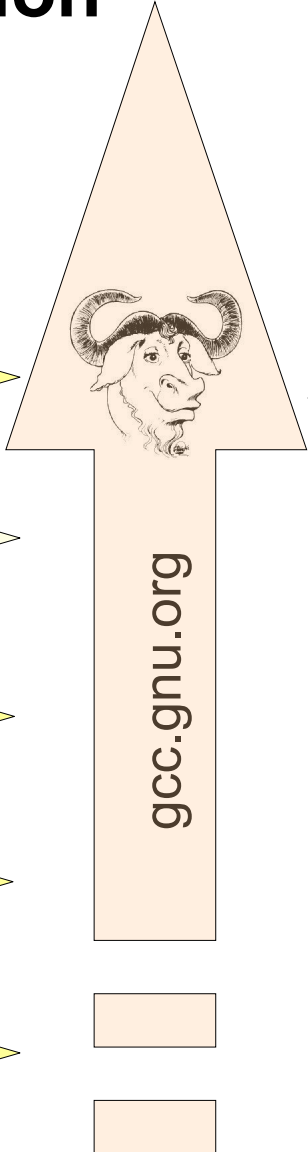
GCC-4.3 patches

GCC-4.2 + software DFP

GCC-4.1 patches

GCC-3.4 patches

GCC-3.3 patches



z10 support hardware DFP

z9-ec | bc support

z9-109 support

z990 support

GCC versions supported on System z



GCC version	Used in SUSE distribution	Used in Red Hat distribution
GCC-3.3	SLES9	
GCC-3.4		RHEL4
GCC-4.0		
GCC-4.1	SLES10	RHEL5
GCC-4.2		
GCC-4.3	SLES11	
GCC-4.4		RHEL6 ?
GCC-4.5		

The Novell logo consists of the word 'Novell.' in red text on a black rectangular background.



Optimizing C/C++ code

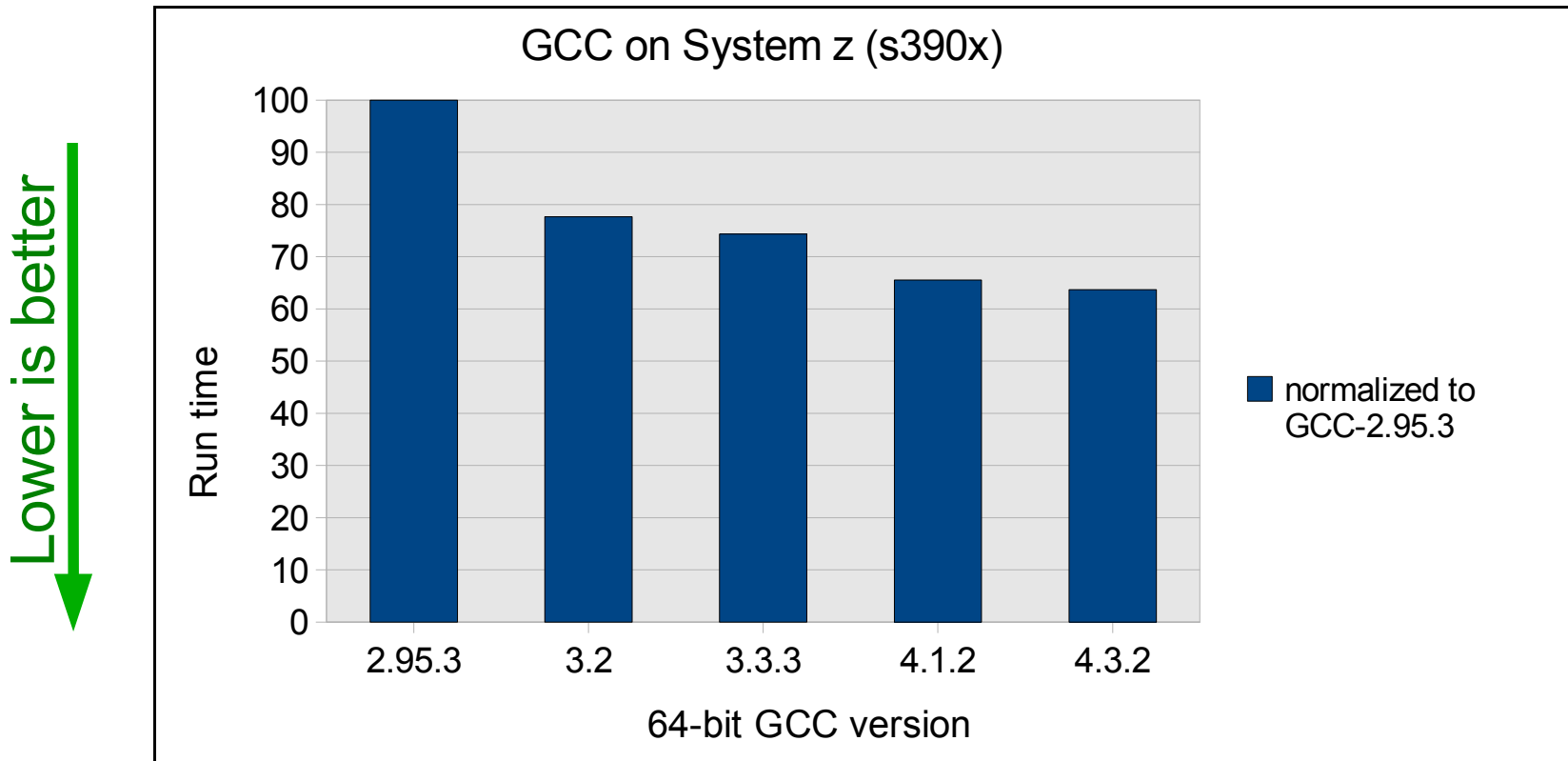


- Produce optimized code
 - Options -O3 or -O2 (often found in delivered Makefiles) are a good starting points
 - Optimize GCC instruction scheduling with the performance critical target machine in mind
 - -mtune=values <z9-109 from gcc-4.1 and up> <z10 with SLES11 gcc-4.3>
 - If you know the target machine exploit improved machine instruction set
 - -march=values <z9-109 from gcc-4.1 and up> <z10 with SLES11 gcc-4.3>
 - -march is only upward compatible
- Fine Tuning: additional general options on a file by file basis
 - Use of inline assembler for performance critical functions may have advantages
 - -funroll-loops --param max-unrolled-insns=100 has advantages
 - -ffast-math speeds up calculations (if not exact implementation of IEEE or ISO rules/specifications for math functions is needed)
 - Don't use debugging options in the final executable

GCC performance evolution on System z



- Run time of industry standard benchmark applications with newer GCC versions is much shorter



DFP - support added in GCC



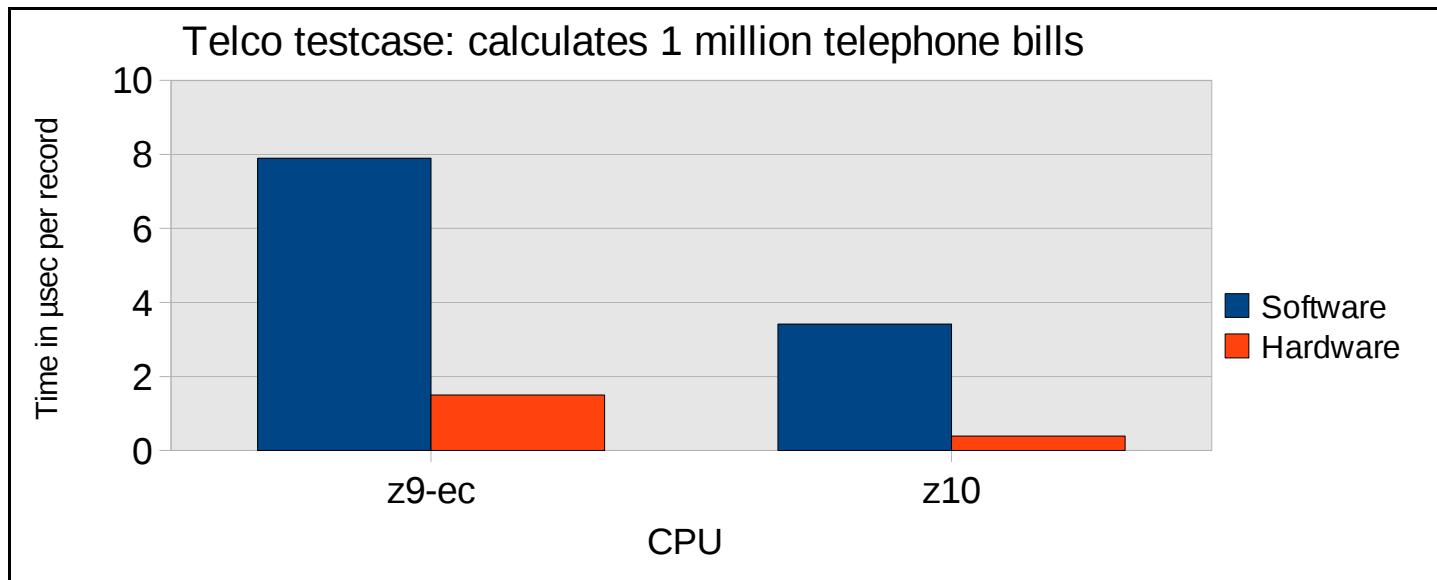
- Front end support (C, C++, Fortran, Java):
 - Support for the 3 new data types: `_Decimal{32|64|128}`
 - Support for DFP constants written with DF suffix
- Middle end support:
 - Complete DFP arithmetic layer for constant folding
 - Support for integer or IEEE floating point conversion routines
- GCC - versions
 - Software DFP support in GCC-4.2 added
 - Hardware DFP support in GCC-4.3 added (usable with z9-ec, z10)
- For the first time GCC-4.3.2 as available in SLES11 offers DFP in a supported environment on Linux on System z
 - The usage of DFP arithmetics in applications requires the explicit use of DFP data types
 - If GCC is used with `-march=z9-ec` or `-march=z10` the HW DFP support is used by default

DFP - decimal floating point performance



- Telco testcase models a telephone company's billing system
 - Billing of one million telephone calls including tax using DFP arithmetics
- Big advantage if DFP hardware support is exploited
 - z9-ec DFP hardware support in millicode
 - z10 DFP hardware support by real hardware -> much faster

Lower is better





Agenda

- System z10
- GCC compiler
- **Java**
- CPU hotplug
- Oprofile

Java on servers: Workload



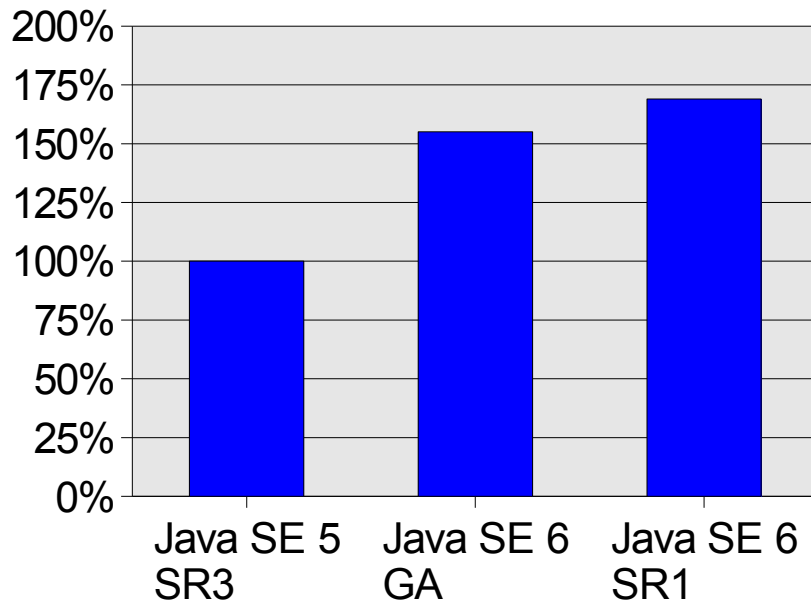
- Evaluates server side Java
 - Emulates 3-tier system
 - Random input from user
 - Middle tier business logic implemented in Java
 - No explicit database --> emulated by Java objects
- Stressed components
 - Java
 - Virtual Machine (VM)
 - Just-In-Time compiler (JIT)
 - Garbage Collection (GC)
 - Linux operating system
 - Threads
 - CPUs
 - Caches and Memory

Java on servers: Performance Improvements

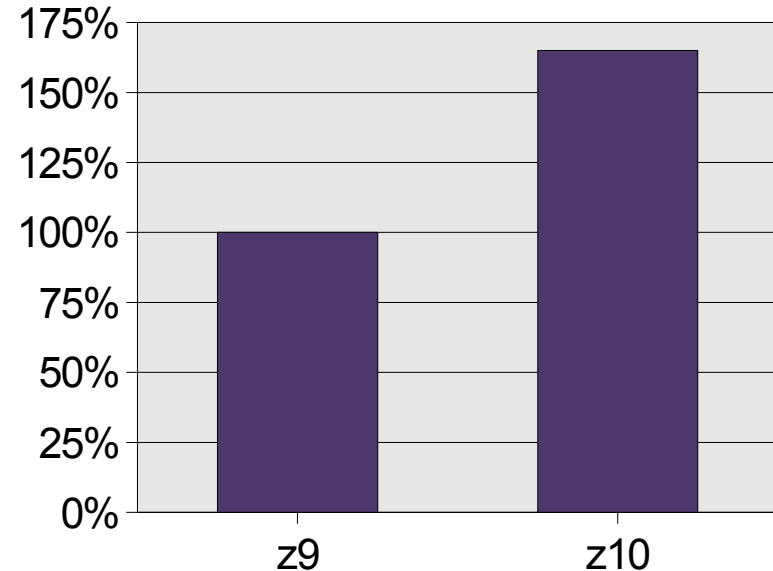


- Better virtual machines (VMs) and just-in-time (JIT) compilers
- Better garbage collection (GC) technologies
- Improvements through new hardware

History of Java versions



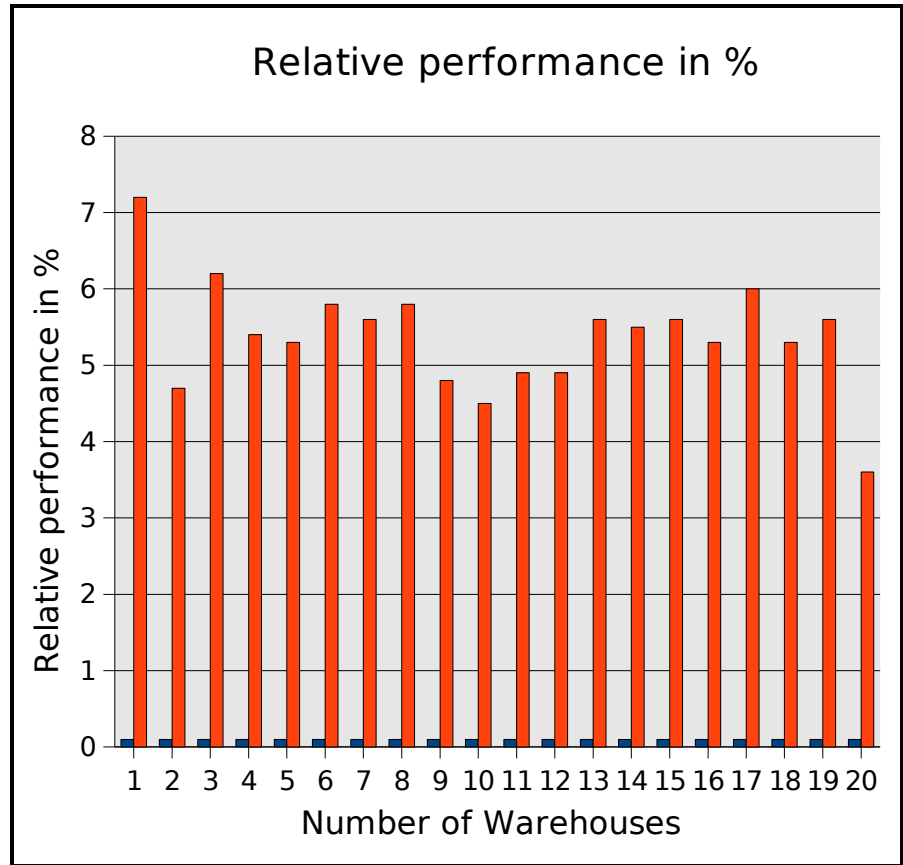
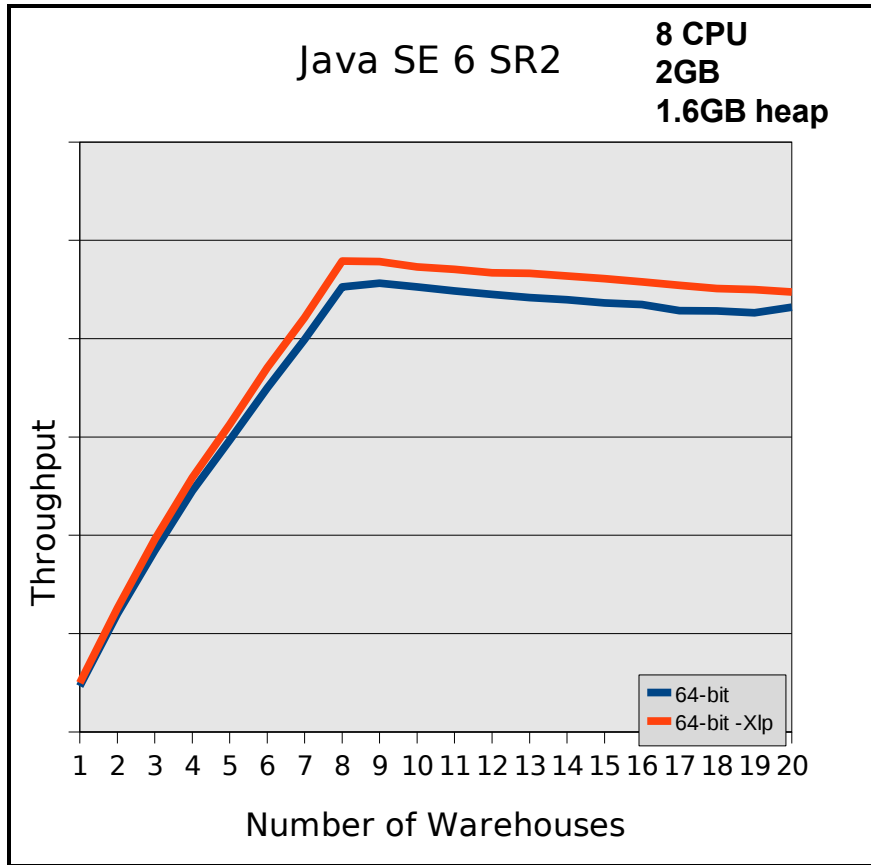
System z with Java SE 6 GA

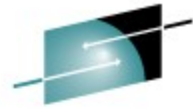




Java: Large page support (z10 feature)

- use of -Xlp improves throughput
- large page size was 2 MB (default for SLES10, RHEL5)

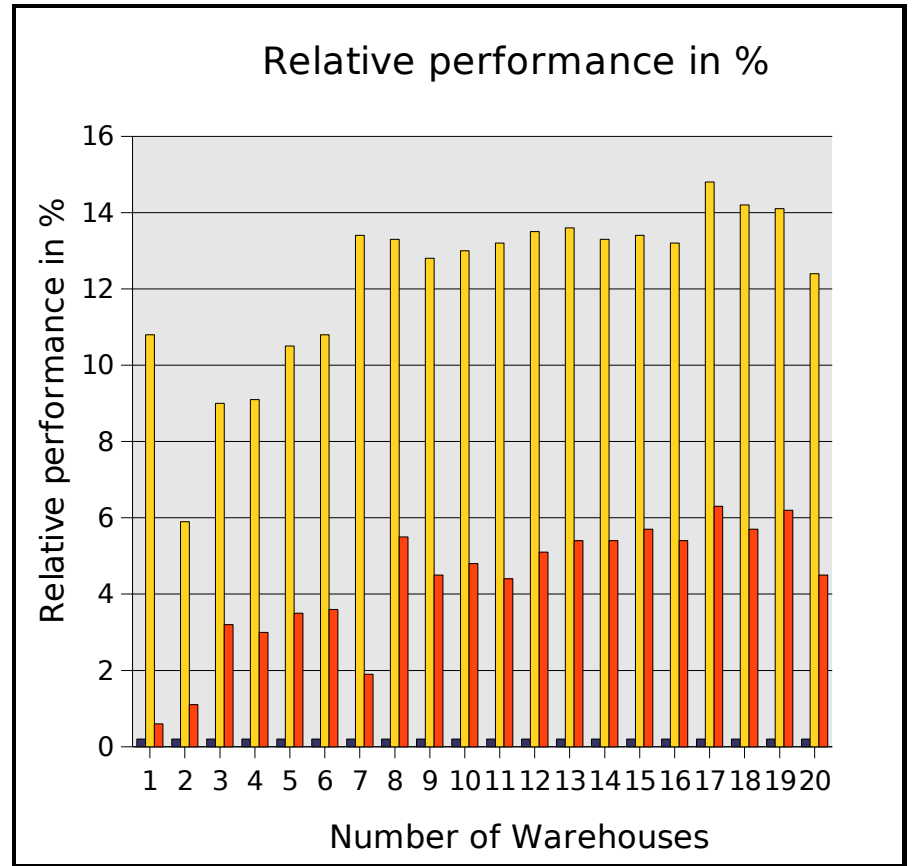
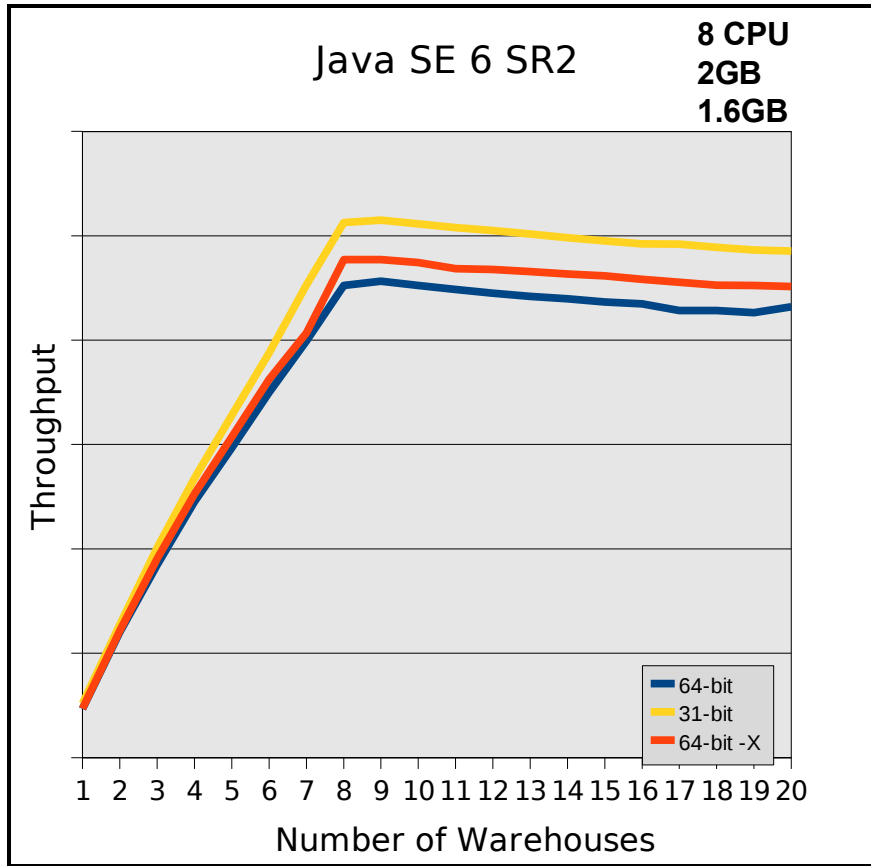




SHARE
Technology • Connections • Results

Java on servers: 31-bit vs. 64-bit

- Use of -Xcompressedrefs provides relief for 64-bit (new with Java SE 6 SR2)



Java on servers: Heap size



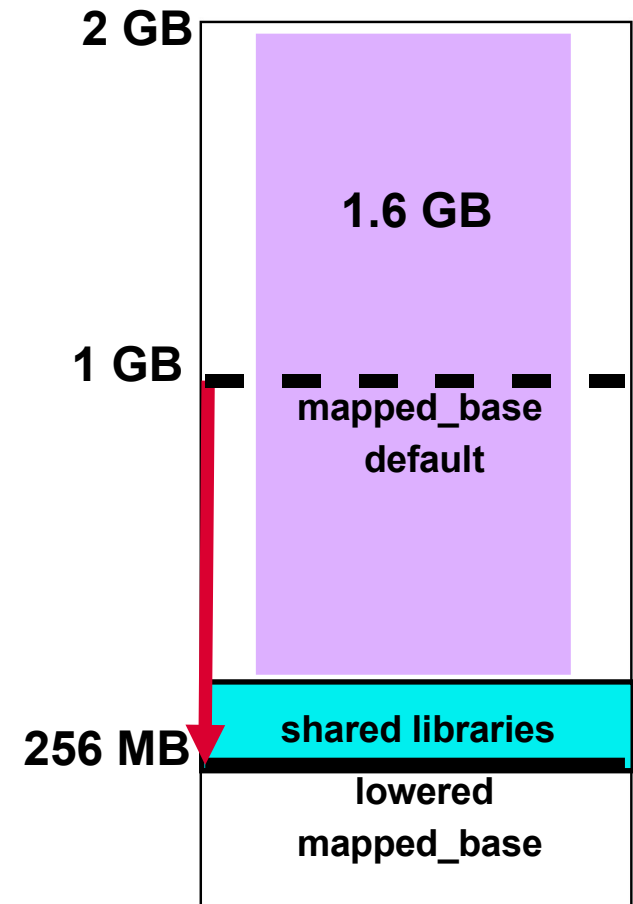
- Heap size needs to be sized adequately
 - Maximum heap size \leq available memory
 - avoids paging in Linux and z/VM
 - Heap too small: frequent garbage collection and OutOfMemoryErrors
 - Heap too big: infrequent garbage collection; Linux starts swapping
 - 31-bit Java kits: larger heap sizes up to 1.6 GB (modify memory layout)
 - also true for 31-bit Java kits in a 64-bit Linux environment
- Useful Java interpreter parameters for fine tuning – workload dependent
 - Setting a fixed heap size: -Xms (initial), -Xmx (maximum), when initial==maximum
 - Monitor garbage collection (GC): -verbose:gc
 - -Xlp tries to allocate large pages for the heap
 - prereq: Linux kernel needs to be setup for large pages (vm.nr_hugepages)
 - Control GC behavior: -Xgcpolicy:[optthruput, optavgpause, gencon]
 - 64-bit: smaller size of heap objects: -Xcompressedrefs

Java: larger heaps for 31-bit Java kits (1)



- Modify Linux memory layout
 - Reorder mapped base for shared libraries
- 31-bit emulation mode for Novell SLES9, SLES10
- HOWTO:
 - PID is the process ID of the process you want to change the layout (usually the bash shell)
 - \$\$ gives the current shell PID, cat /proc/self/maps works as well
 - Display memory map of any PID by
 - cat /proc/<PID>/maps
 - Check the mapped base value by
 - cat /proc/<PID>/mapped_base
 - Lower the value to e.g. 256 MB by
 - echo 268435456 >/proc/<PID>/mapped_base

==> Now retry to allocate a larger heap size



Java: larger heaps for 31-bit Java kits (2)



- Modify Linux memory layout
 - RHEL includes flex-mmap patch; turn off Linux prelinking
- Applies to RHEL4, RHEL5 distributions (31-bit emulation mode)
- HOWTO:
 - Show state of flex-mmap patch
 - `cat /proc/sys/vm/legacy_va_layout`
 - 0 means flex-mmap is enabled; 1 means old memory layout
 - Enable flex-mmap if disabled
 - `echo 0 > /proc/sys/vm/legacy_va_layout`
 - Disable Linux prelinking
 - in `/etc/sysconfig/prelink` set `PRELINKING=no`
 - Apply setting by running the daily cron prelink job immediately
 - `# /etc/cron.daily/prelink <ENTER>`

==> Now retry to allocate a larger heap size

Java: Summary & Hints



- Try to use the **latest Java version**
 - Up to 60% release to release improvements
 - Up to 15% with newer service releases (SR) for a release
 - Middleware applications often bring their own Java Kit
- Make sure that you've got **JIT enabled**
 - Command 'java -version' says "JIT enabled/disabled"
- Lots of java interpreter **-X... parameters** for fine tuning
 - To get an idea type 'java -X'
- Provide an **optimal heap size** to your application
- Don't use the java interpreter in batch mode
 - Don't call x-times 'java Myprog'
 - Instead try to put the loop logic into your Java application

Agenda



- System z10
- GCC compiler
- Java
- **CPU hotplug**
- Oprofile

CPU hotplug function



- Changes the number of used processors on the fly, depending on the current overall utilization and load
- Is available with SLES10 SP2 and SLES11
- Expectation:
 - **Increases the performance of single threaded applications within a z/VM or LPAR environment with multiple CPUs**
- Enables or disables CPUs based on a set of rules
- Is enabled in the kernel configuration by setting

Base setup --->

--- Processor type and features ---

64 bit kernel (CONFIG_64BIT)

Symmetric multi-processing support (CONFIG_SMP)

└ Support for hot-pluggable CPUs (CONFIG_HOTPLUG_CPU)

CPU hotplug parameters



- The control information is stored at `/etc/sysconfig/cpuplugd`
- Minimum number of CPUs is set with `cpu_min=<number>`
- Maximum number of CPUs is set with `cpu_max=<number>`
- The update interval is set with `update=<value in seconds>`
- Consider the effect of kernel “cpu” parameters:
 - `maxcpus=<n>` sets the number of processors which will be active after system boot
 - `possible_cpus=<n>` is the upper limit for hotpluggable CPUs
 - If `possible_cpus` is not specified but `maxcpus` is, then `maxcpus` is the upper limit for hot-pluggable CPUs

CPU hotplug rules



- The default rule for increasing the number of CPUs is `HOTPLUG="(loadavg > onumcpus + 0.75) & (idle < 10.0)"`
 - An additional CPU is enabled, if the loadaverage is greater than the number of active (online) CPUs plus 0.75 and the current idle percentage is less than 10 percent.
- The default rule for decreasing the number of CPUs is `HOTUNPLUG="(loadavg < onumcpus - 0.25) | (idle > 50)"`
 - A CPU is disabled, either if the current load is below the number of active CPUs minus 0.25 or if the idle percentage is greater than 50%.
- The formulas for these rules can be modified. See “Device Drivers, Features and Commands” for valid expressions.
- Note:
 - `loadavg` is a value that changes slowly
 - `idle` changes fast
 - Increments and decrements of active CPUs are done in steps of 1 every time when the rules are checked.

CPU hotplug test workload



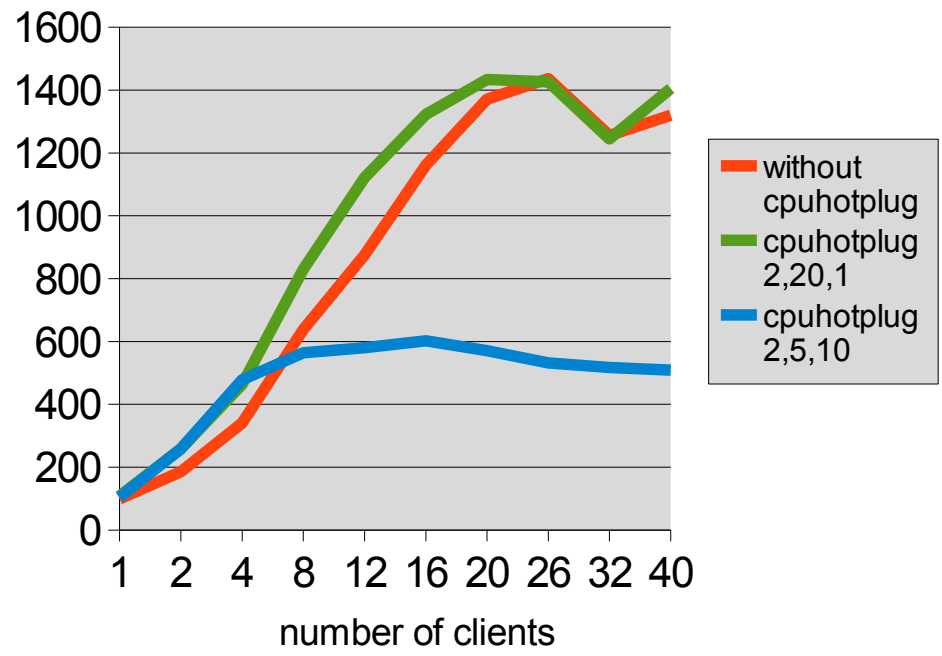
- dbench 3
 - Emulation of Netbench benchmark, rates windows file servers
 - Mainly memory operations
 - Mixed file operations workload for each process: create, write, read, append, delete
 - Scaling with 1,2,4,8,16 CPUs and 1,4,8,12,16,20,26,32 and 40 clients
 - 2 GB memory
- Modification to the standard code:
 - Purpose: Need more interaction between clients
 - Create two processes per client and communicate with POSIX message queues
 - First process:
 - *Read the I/O commands from the control file*
 - *Pass this information to the second process*
 - Second process:
 - *Performs the execution of this command*
 - *Reports the end of the operation back to the first process*



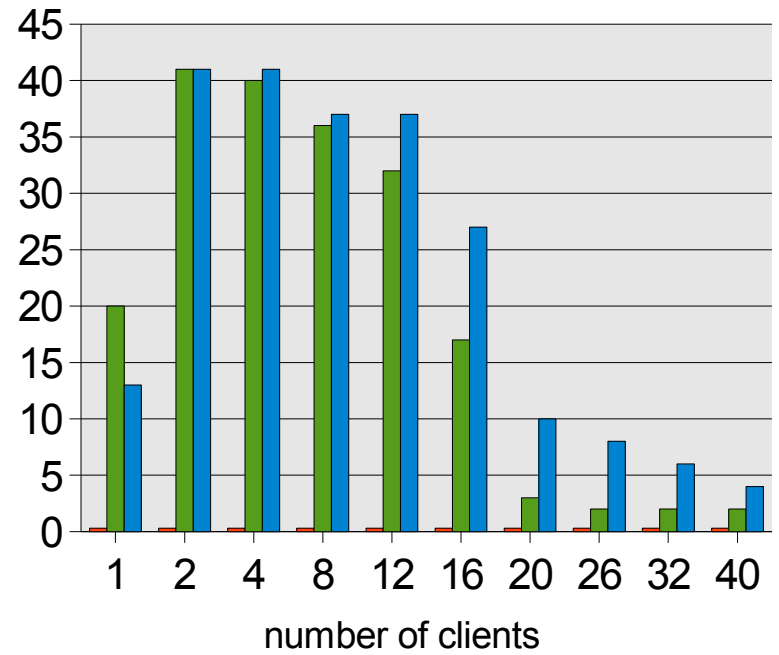
CPU hotplug performance results

- Improvements in case where the default (high) number of CPUs is not needed
- Up to 40% more throughput, up to 40% CPU cost savings

Throughput by dbench [MB/s]



Relative CPU consumption savings based on the test run without cpu hotplug [%]



CPU hotplug summary



- This feature improves the performance by
 - sizing the correct amount of processors for a Linux system depending on its current load
 - avoiding the Linux scheduler queue balancing in partial load situations
- Set the minimum and maximum number of CPUs to values which apply to the real workload:
 - Setting `cpu_min` to 2 may be too high
 - `cpu_max` should be set so that it really covers the peaks
- Linux guests under z/VM: use z/VM 5.4
 - Guarantees that stopped processors are no longer included in virtual processor prioritization calculations
 - Ensures share redistribution

Agenda



- System z10
- GCC compiler
- Java
- CPU hotplug
- **Oprofile**

Oprofile – the Open Source sampling tool



- Oprofile offers profiling of all running code on Linux systems, providing a variety of statistics
 - By default, kernel mode and user mode information is gathered for configurable events
- System z hardware currently does not have support for hardware performance counters, instead timer interrupt is used
 - Enable the hz_timer(!)
- The timer is set to whatever the jiffy rate is and is not user-settable
- Novell / SUSE: OProfile is on the SDK CDs
- Also available with RHEL4 and RHEL5
- More info at:

<http://oprofile.sourceforge.net/docs/>

<http://www.redhat.com/docs/manuals/enterprise/RHEL-4-Manual/sysadmin-guide/ch-oprofile.html>

Oprofile – short HowTo



```
sysctl -w kernel.hz_timer=1
```

```
gunzip /boot/vmlinux-2.6.16.46-0.4-default.gz
```

specify the kernel level of `uname -r`

```
opcontrol --vmlinux=/boot/vmlinux-2.6.16.46-0.4-  
default
```

```
▶ opcontrol --start
```

<DO THE TEST>

```
opcontrol --shutdown
```

```
opreport
```

any next test to run? If yes

```
opcontrol --reset
```



opreport

```
>opreport
CPU: CPU with timer interrupt, speed 0 MHz (estimated)
Profiling through timer interrupt
```

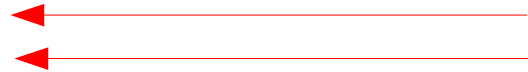
samples	TIMER:0	%	
140642	94.0617		vmlinux-2.6.16.46-0.4-default ← Kernel
3071	2.0539		libc-2.4.so ← glibc
1925	1.2874		dbench ← application
1922	1.2854		ext3 ← file system
1442	0.9644		jbd ← journaling
349	0.2334		dasd_mod ← dasd driver
152	0.1017		apparmor ← security
6	0.0040		oprofiled
5	0.0033		bash
5	0.0033		ld-2.4.so
1	6.7e-04		dasd_eckd_mod
1	6.7e-04		oprofile

...

Opreport -- long-filenames | -l



```
>opreport -l
warning: /apparmor could not be found.
warning: /dasd_eckd_mod could not be found.
warning: /dasd_mod could not be found.
warning: /ext3 could not be found.
warning: /jbd could not be found.
warning: /oprofile could not be found.
CPU: CPU with timer interrupt, speed 0 MHz (estimated)
Profiling through timer interrupt
samples %      app name                symbol name
130852  87.5141 vmlinux-2.6.16.46-0.4-default cpu_idle
1922    1.2854  ext3                    (no symbols)
1442    0.9644  jbd                     (no symbols)
734     0.4909  vmlinux-2.6.16.46-0.4-default memcopy
662     0.4427  libc-2.4.so            strchr
619     0.4140  dbench                 next_token
567     0.3792  vmlinux-2.6.16.46-0.4-default do_gettimeofday
536     0.3585  vmlinux-2.6.16.46-0.4-default __link_path_walk
525     0.3511  vmlinux-2.6.16.46-0.4-default copy_to_user_std
435     0.2909  libc-2.4.so            strstr
413     0.2762  dbench                 child_run
349     0.2334  dasd_mod               (no symbols)
347     0.2321  vmlinux-2.6.16.46-0.4-default _spin_lock
328     0.2194  vmlinux-2.6.16.46-0.4-default sysc_do_svc
285     0.1906  dbench                 all_string_sub
283     0.1893  vmlinux-2.6.16.46-0.4-default __d_lookup
251     0.1679  vmlinux-2.6.16.46-0.4-default __find_get_block
231     0.1545  libc-2.4.so            __strtoul_l_internal
216     0.1445  dbench                 vsnprintf
209     0.1398  vmlinux-2.6.16.46-0.4-default filldir64
205     0.1371  vmlinux-2.6.16.46-0.4-default memset
196     0.1311  vmlinux-2.6.16.46-0.4-default _atomic_dec_and_lock
166     0.1110  vmlinux-2.6.16.46-0.4-default strchr
155     0.1037  libc-2.4.so            memmove
152     0.1017  apparmor               (no symbols)
148     0.0990  libc-2.4.so            readdir
147     0.0983  vmlinux-2.6.16.46-0.4-default __brelse
146     0.0976  vmlinux-2.6.16.46-0.4-default generic_file_buffered_write
```



almost idle
unresolved symbols

opreport -l --image-path -p [paths]



```
>opreport -l --image-path=/lib/modules/2.6.16.46-0.4-default/kernel/fs/ext3/,/lib/modules/2.6.16.46-0.4-  
default/kernel/fs/jbd/,/lib/modules/2.6.16.46-0.4-default/kernel/drivers/s390/block/,/lib/modules/2.6.16.46-0.4-  
default/kernel/security/apparmor/,/lib/modules/2.6.16.46-0.4-default/kernel/arch/s390/oprofile  
CPU: CPU with timer interrupt, speed 0 MHz (estimated)
```

Profiling through timer interrupt

samples	%	image name	app name	symbol name
130852	87.5141	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	cpu_idle
734	0.4909	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	memcpy
662	0.4427	libc-2.4.so	libc-2.4.so	strchr
619	0.4140	dbench	dbench	next_token
567	0.3792	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	do_gettimeofday
536	0.3585	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	__link_path_walk
525	0.3511	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	copy_to_user_std
435	0.2909	libc-2.4.so	libc-2.4.so	strstr
413	0.2762	dbench	dbench	child_run
361	0.2414	ext3.ko	ext3	ext3_get_block_handle
347	0.2321	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	_spin_lock
328	0.2194	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	sysc_do_svc
285	0.1906	dbench	dbench	all_string_sub
283	0.1893	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	__d_lookup
251	0.1679	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	__find_get_block
231	0.1545	libc-2.4.so	libc-2.4.so	___strtol_l_internal
226	0.1511	ext3.ko	ext3	ext3_try_to_allocate
223	0.1491	dasd_mod.ko	dasd_mod	dasd_smallocc_request
216	0.1445	dbench	dbench	vsnprintf
209	0.1398	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	filldir64
205	0.1371	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	memset
196	0.1311	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	_atomic_dec_and_lock
188	0.1257	ext3.ko	ext3	ext3_new_inode
166	0.1110	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	strchr
157	0.1050	jbd.ko	jbd	journal_init_dev
155	0.1037	libc-2.4.so	libc-2.4.so	memmove
148	0.0990	libc-2.4.so	libc-2.4.so	readdir
147	0.0983	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	__brelse
146	0.0976	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	generic_file_buffered_write
144	0.0963	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	generic_permission
140	0.0936	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	__getblk
140	0.0936	vmlinux-2.6.16.46-0.4-default	vmlinux-2.6.16.46-0.4-default	kmem_cache_free

Visit us !



- Linux on System z: Tuning Hints & Tips
<http://www.ibm.com/developerworks/linux/linux390/perf/>
- Linux-VM Performance Website:
<http://www.vm.ibm.com/perf/tips/linuxper.html>
- IBM Redbooks
<http://www.redbooks.ibm.com/>
- IBM Techdocs
<http://www.ibm.com/support/techdocs/atmastr.nsf/Web/Techdocs>

Questions

