http://linuxmain.blogspot.com/

Dr. Eberhard Pasch (epasch@de.ibm.com)

IBM

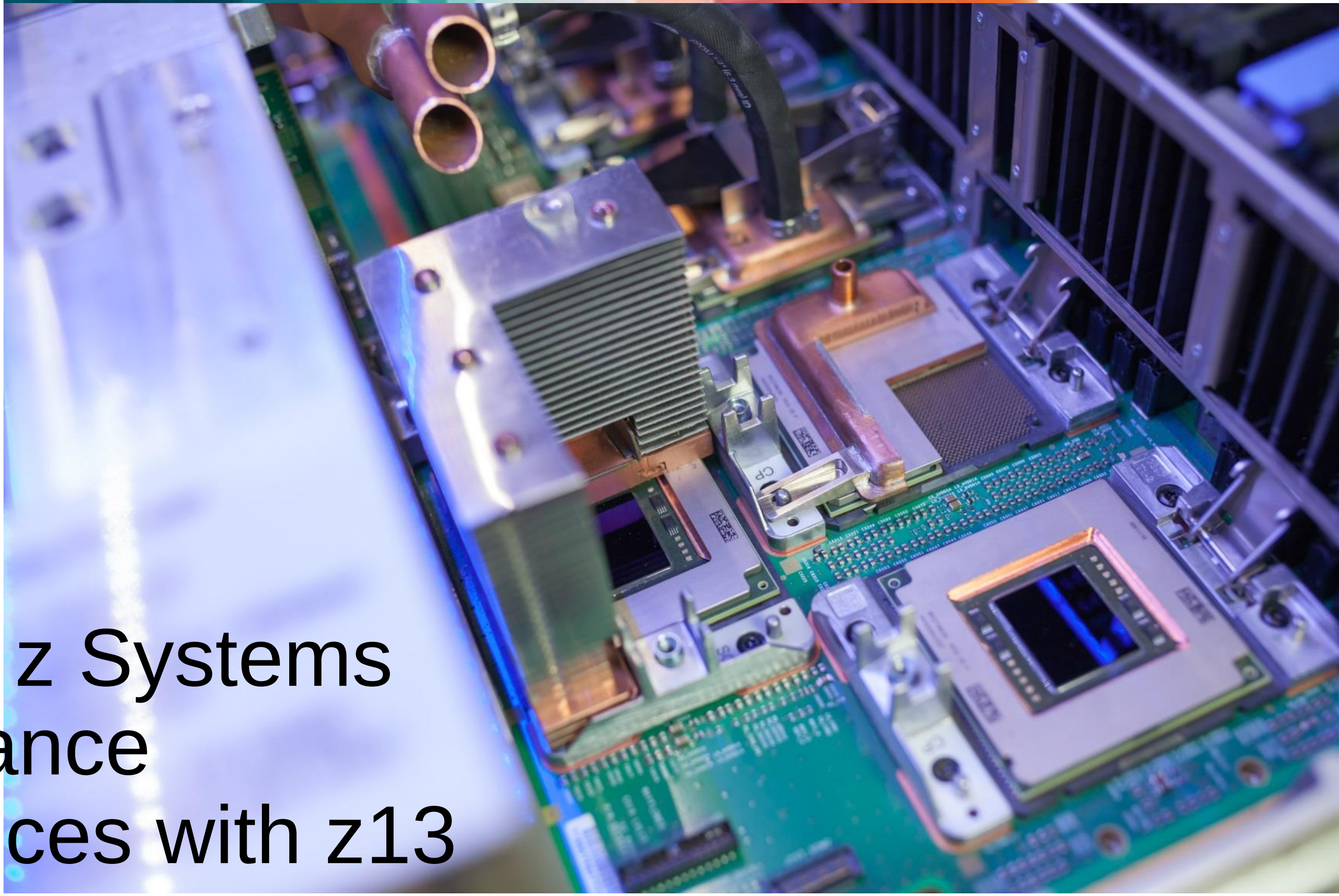
Experiences, People and Ideas Converge to Power Business Outcomes

Celebrating 60 Years of SHARE

March 1-6
Sheraton Seattle
Seattle, WA

SHARE in Seattle 2015

Linux on z Systems Performance Experiences with z13



Trademarks

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

BlueMix
BigInsights
Cognos*
DB2*
DB2 Connect
Domino*
DS8000*
ECKD
FICON*
FileNet*
FlashSystem
GDPS*
GPFS
IBM*
Ibm.com
IBM (logo)*
IMS
Informix*
InfoSphere
Maximo*
MQSeries*
Performance Toolkit for VM
POWER*
Quickr*
Rational*
Sametime*
Smarter Cities*
Smarter Analytics
SPSS*
Storwize*
System Storage*
Tivoli*
WebSphere*
XIV*
z13
zEnterprise*
z/OS*
z Systems
z/VSE*
z/VM*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

* Other product and service names might be trademarks of IBM or other 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. **All z13 numbers have been measured on pre GA hardware with pre GA software.**

Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here. All z13 numbers have been measured on pre GA hardware.

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.

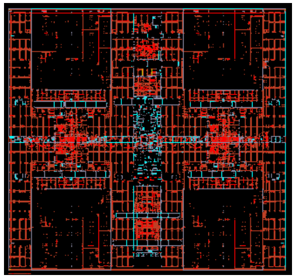
This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

- **z13 structure and characteristics**
- base performance
- SMT2
- recommendations and outlook

z Systems - Processor Roadmap

z10
2/2008



Workload Consolidation and Integration Engine for CPU Intensive Workloads

Decimal FP

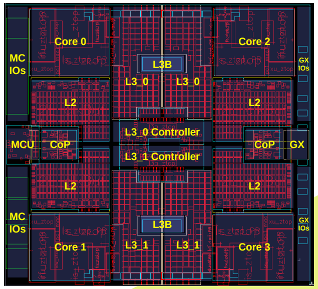
Infiniband

64-CP Image

Large Pages

Shared Memory

z196
9/2010



Top Tier Single Thread Performance,System Capacity

Accelerator Integration

Out of Order Execution

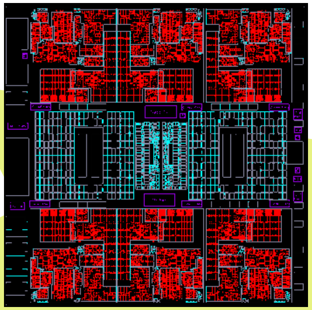
Water Cooling

PCIe I/O Fabric

RAIM

Enhanced Energy Management

zEC12
8/2012



Leadership Single Thread, Enhanced Throughput

Improved out-of-order

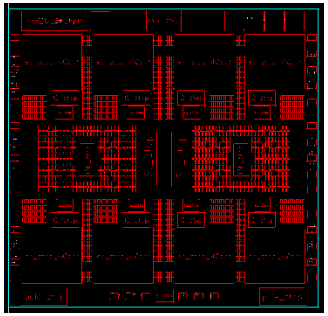
Transactional Memory

Dynamic Optimization

2 GB page support

Step Function in System Capacity

z13
1/2015



Leadership System Capacity and Performance

Modularity & Scalability

Dynamic SMT

Supports two instruction threads

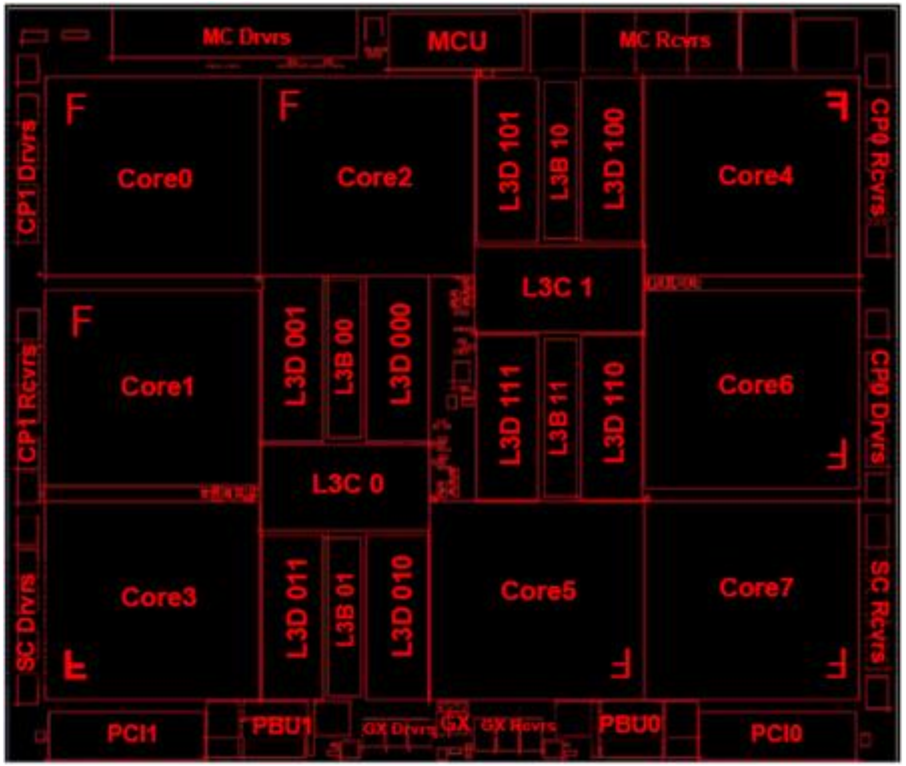
SIMD

PCIe attached accelerators

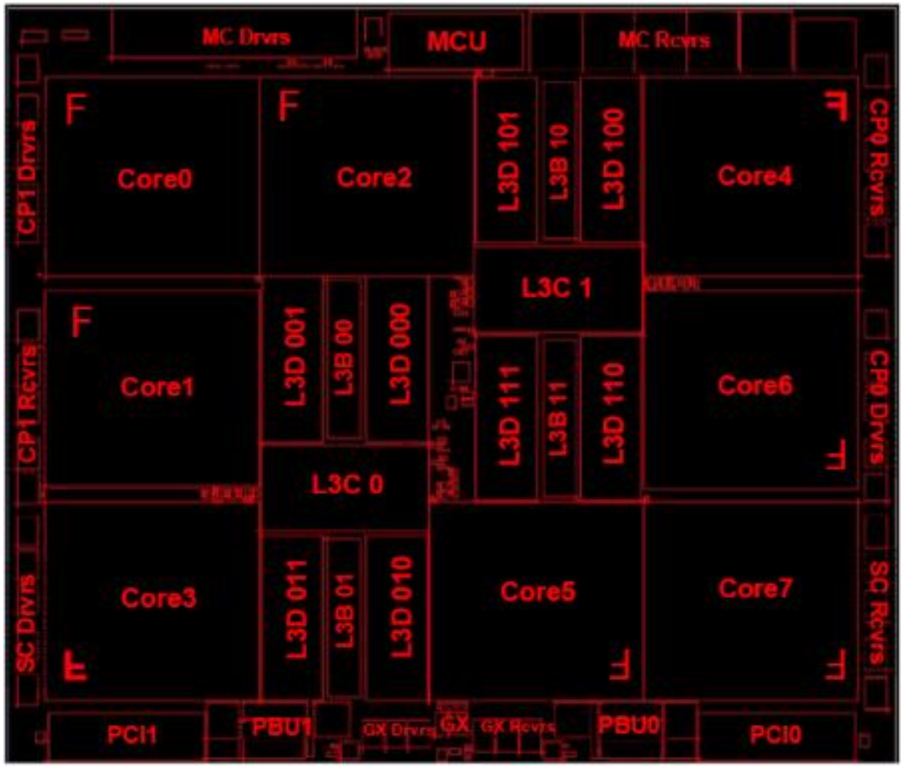
Business Analytics Optimized

z13 Processor Overview

- 2X Instruction pipe width
  - – Improves IPC for all modes
  - – Symmetry simplifies dispatch/issue rules
  - – Required for effective SMT
- Added FXU and BFU execution units
  - – 4 FXUs
  - – 2 BFUs, DFUs
  - – 2 new SIMD units
- SIMD unit plus additional registers
- Pipe depth re-optimized for power/performance
  - – Product frequency reduced
  - – Processor performance increased
- SMT support
  - – Wide, symmetric pipeline
  - – Full architected state per thread
  - – SMT-adjusted CPU usage metering



z13 8-Core Processor Unit (PU) Chip Detail

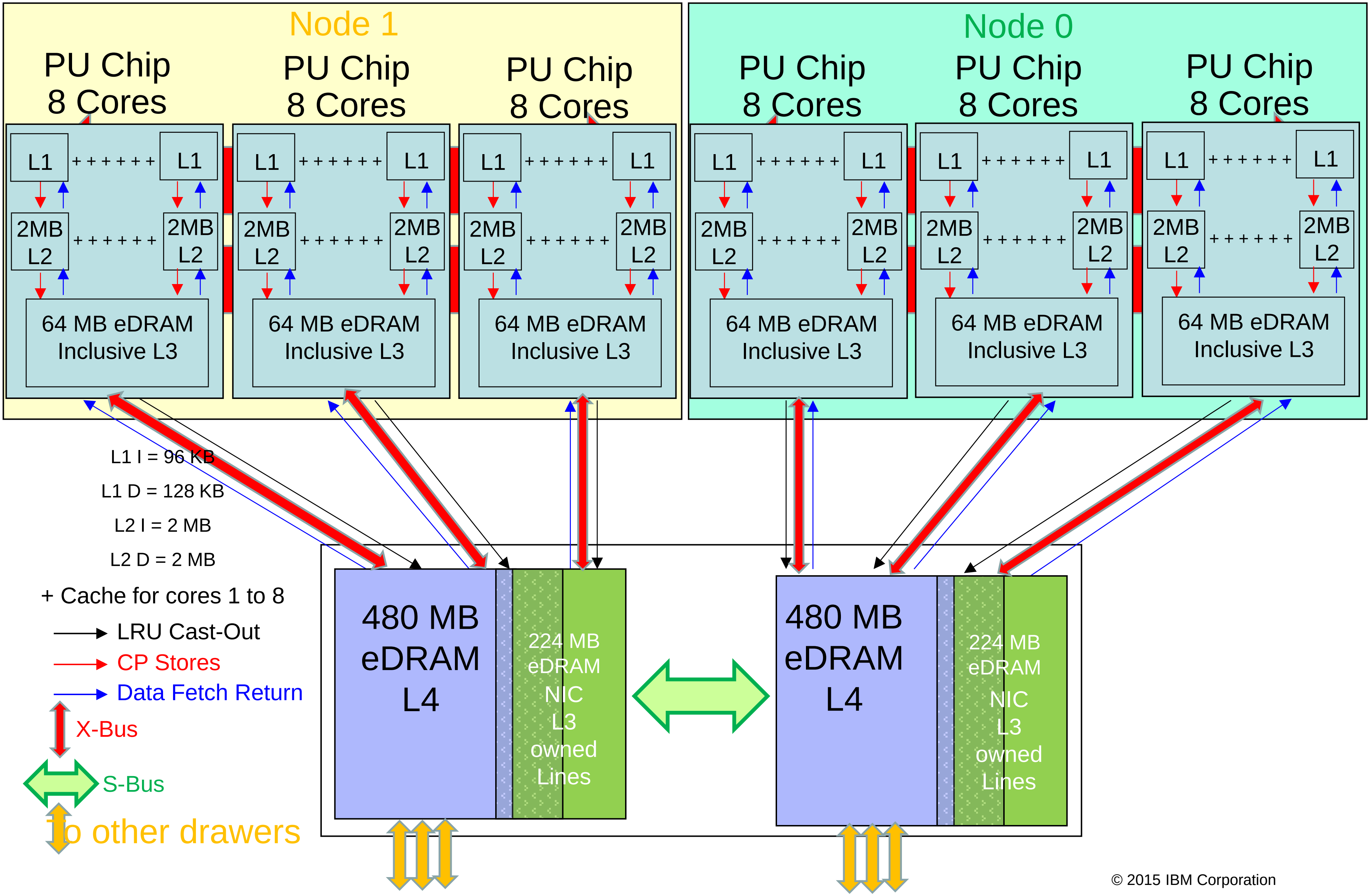


- 14S0 22nm SOI Technology
  - 17 layers of metal
  - 3.99 Billion Transistors
  - 13.7 miles of copper wire

- Chip Area
  - 678.8 mm^2
  - 28.4 x 23.9 mm
  - 17,773 power pins
  - 1,603 signal I/Os

- Up to eight active cores (PUs) per chip
  - 5.0 GHz (v5.5 GHz zEC12)
  - L1 cache/ core
    - 96 KB I-cache
    - 128 KB D-cache
  - L2 cache/ core
    - 2M+2M Byte eDRAM split private L2 cache
- Single Instruction/Multiple Data (SIMD)
- Single thread or 2-way simultaneous multithreading (SMT) operation
- Improved instruction execution bandwidth:
  - Greatly improved branch prediction and instruction fetch to support SMT
  - Instruction decode, dispatch, complete increased to 6 instructions per cycle
  - Issue up to 10 instructions per cycle
  - Integer and floating point execution units
- On chip 64 MB eDRAM L3 Cache
  - Shared by all cores
- I/O buses
  - One InfiniBand I/O bus
  - Two PCIe I/O buses
- Memory Controller (MCU)
  - Interface to controller on memory DIMMs
  - Supports RAIM design

z13 CPC Drawer Cache Hierarchy



Node 1

PU Chip
8 Cores

PU Chip
8 Cores

PU Chip
8 Cores

L1 + + + + + + L1

2MB L2 + + + + + + 2MB L2

64 MB eDRAM
Inclusive L3

Node 0

PU Chip
8 Cores

PU Chip
8 Cores

PU Chip
8 Cores

L1 + + + + + + L1

2MB L2 + + + + + + 2MB L2

64 MB eDRAM
Inclusive L3

L1 I = 96 KB

L1 D = 128 KB

L2 I = 2 MB

L2 D = 2 MB

+ Cache for cores 1 to 8

LRU Cast-Out

CP Stores

Data Fetch Return

X-Bus

S-Bus

To other drawers

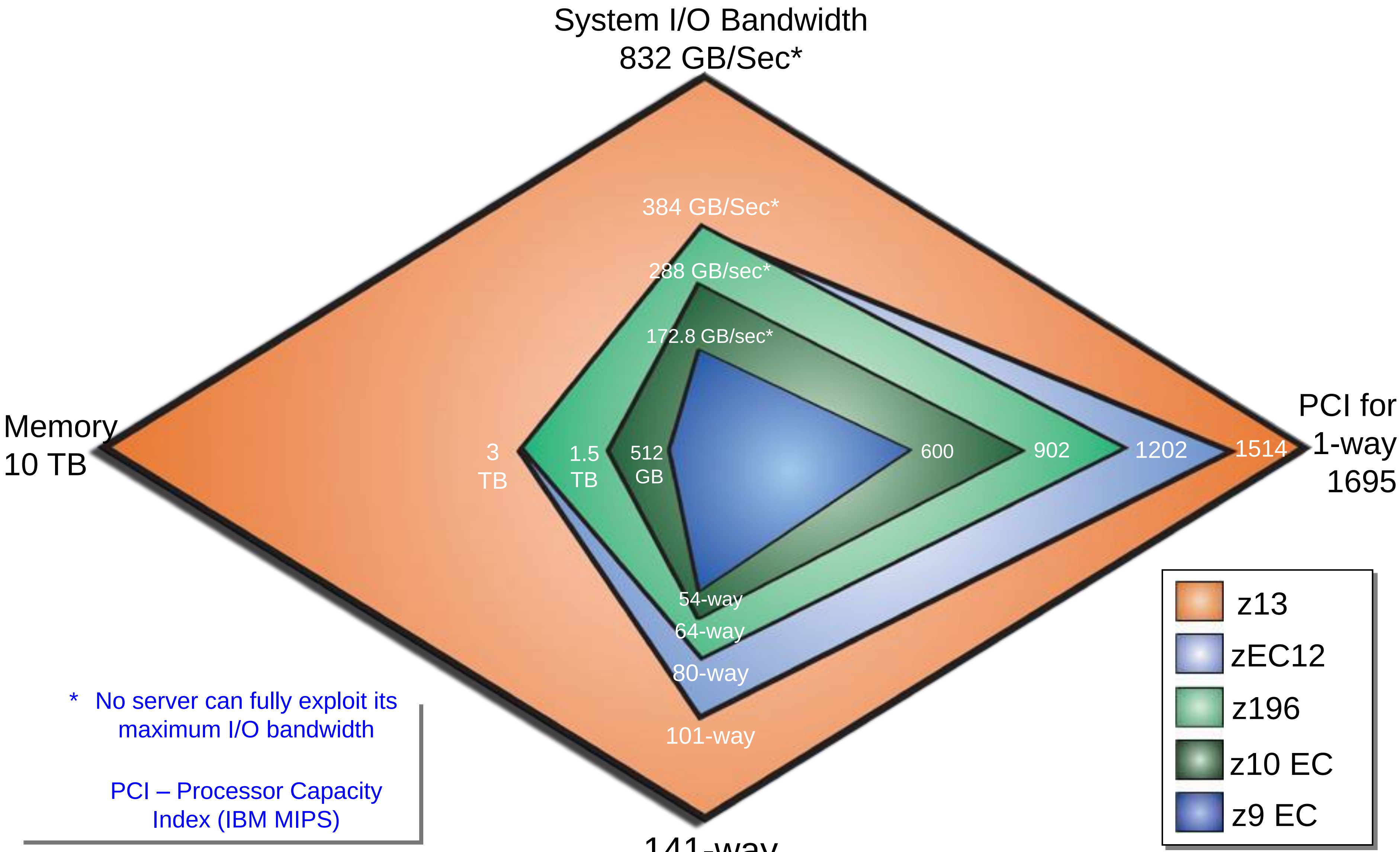
480 MB
eDRAM
L4

224 MB
eDRAM
NIC
L3
owned
Lines

480 MB
eDRAM
L4

224 MB
eDRAM
NIC
L3
owned
Lines

IBM z13: Advanced system design



* No server can fully exploit its maximum I/O bandwidth

PCI – Processor Capacity Index (IBM MIPS)

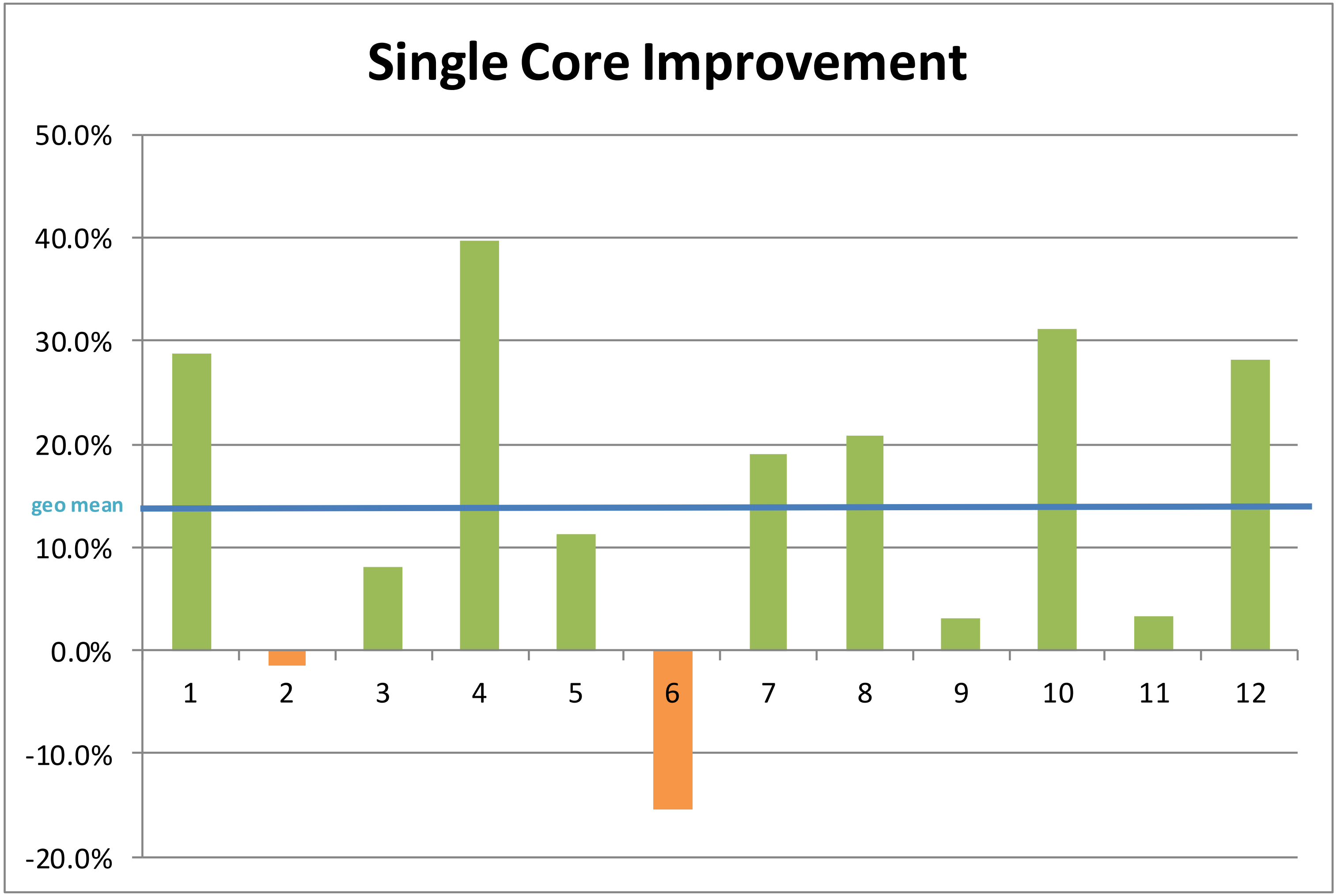
Large Memory Value – Potential Performance Gains

- 2.5 TB per drawer for a total of 10 TB available, special pricing!
- Enables more caching for classical databases
  - larger Oracle SGAs
- Helps with storage pressure under z/VM
- Enables In-Memory Databases
  - Dramatic reduction in response time by avoiding I/O wait
  - DB2 BLU / Oracle 12c
- Enables in memory analytics
- Java heaps can be increased
  - for older Java versions be sure to use –Xcompressedrefs

Agenda

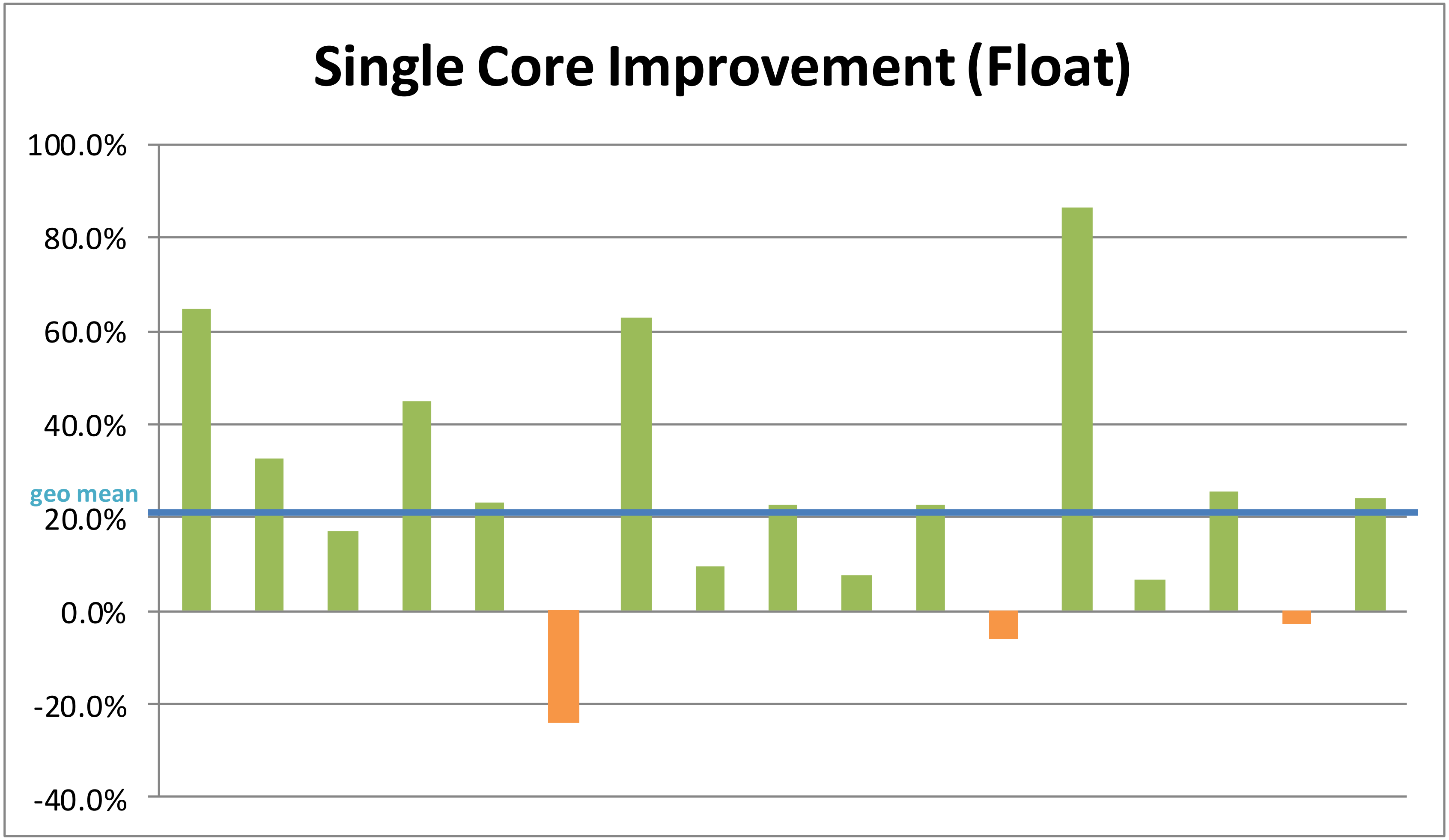
- z13 structure an characteristics
- **base performance**
- SMT2
- recommendations and outlook

Single Core improvement for existing C/C++ workloads



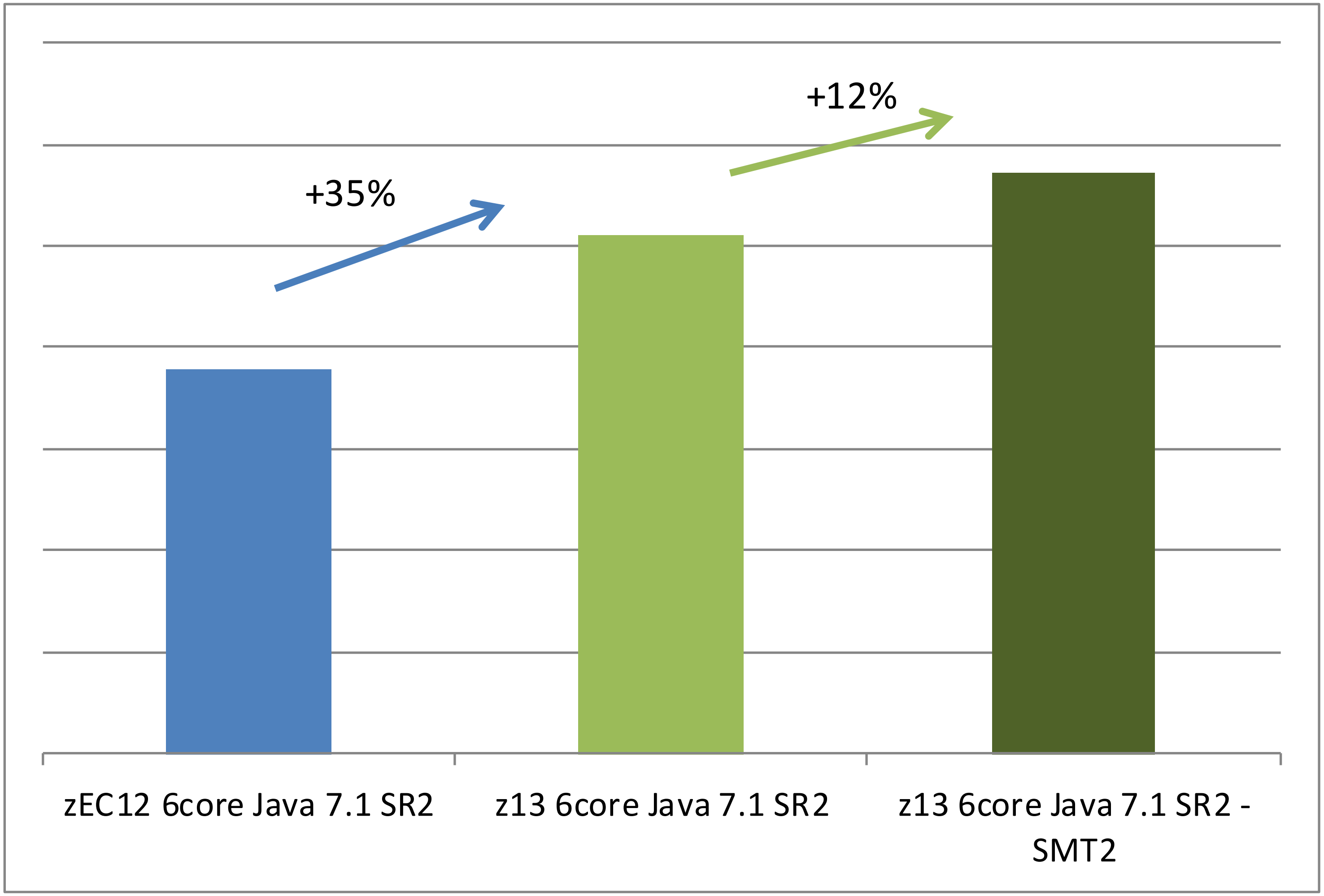
standard compiler RHEL7.0, -O3, -march=z196

Single vore improvement for exiting workloads (floating point intense)

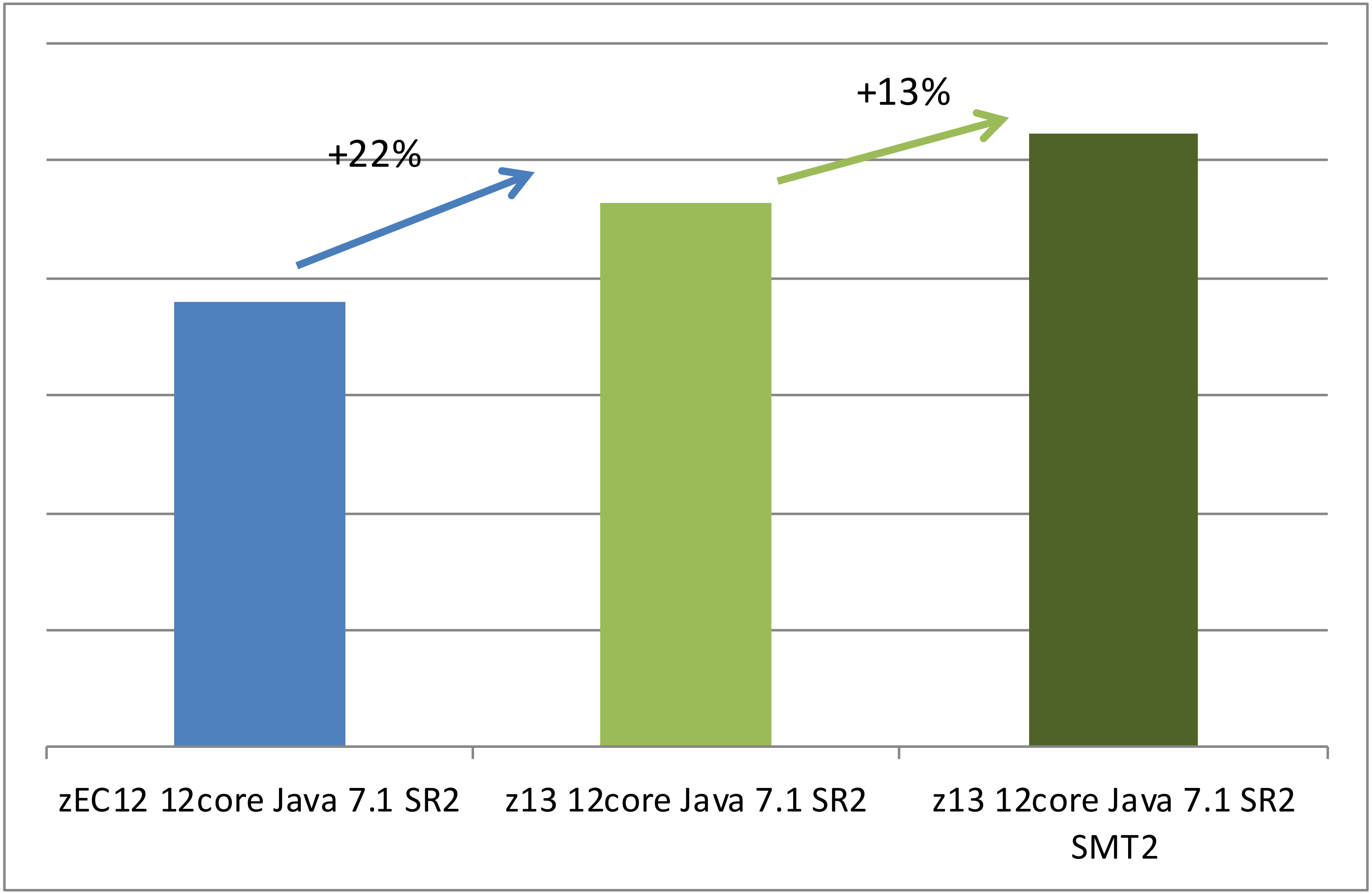


standard compiler RHEL7.0, -O3, -march=z196, the two FXUs do make a difference!

A perfect fit for the new z13 chip – all on one chip



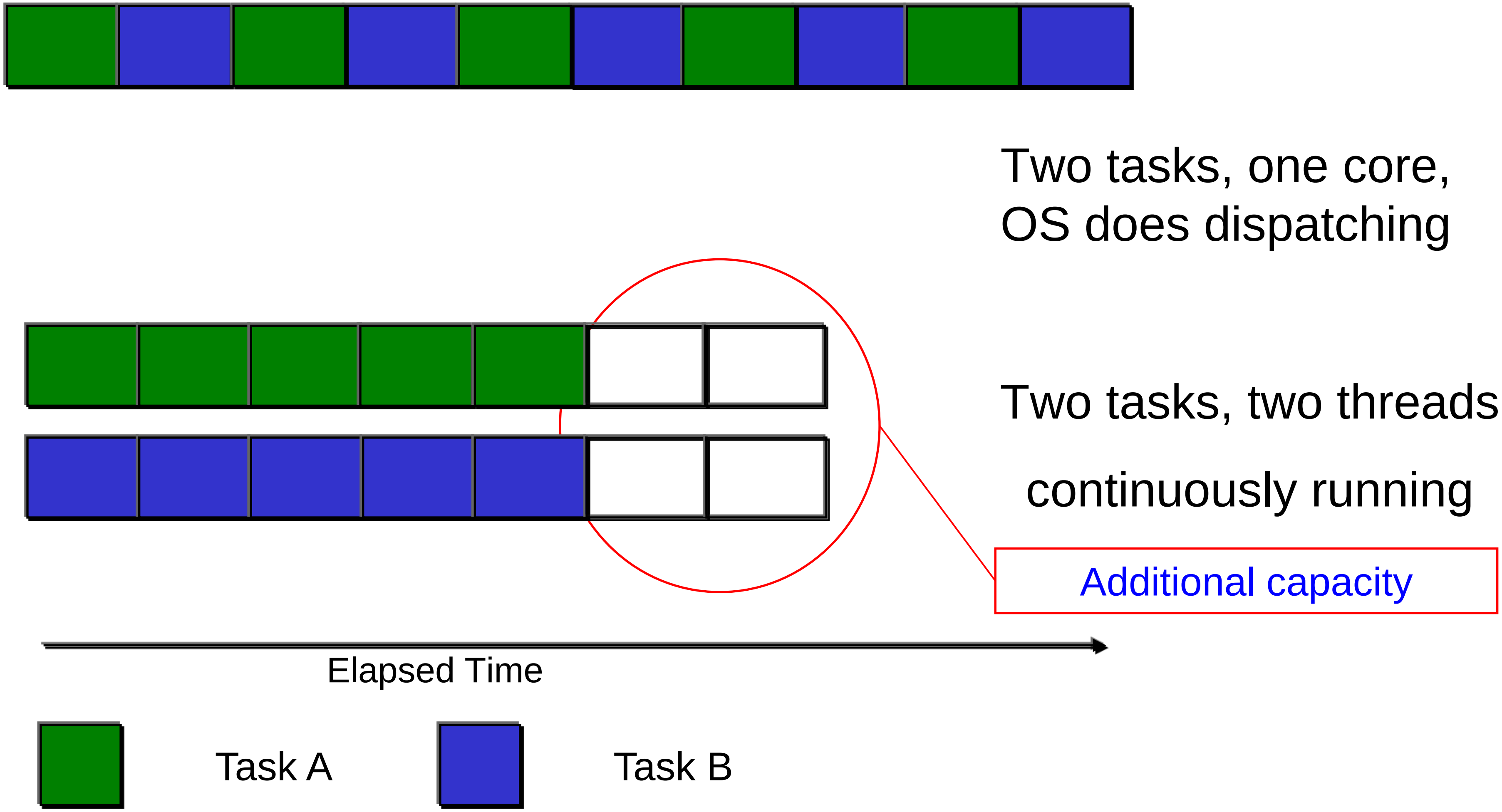
12 cores – 2 chips – one node



Agenda

- z13 structure an characteristics
- base performance
- **SMT2**
- preparation and planning
- recommendations and outlook

Simultaneous Multithreading Value Example



Two tasks, one core, OS does dispatching

Two tasks, two threads continuously running

Additional capacity

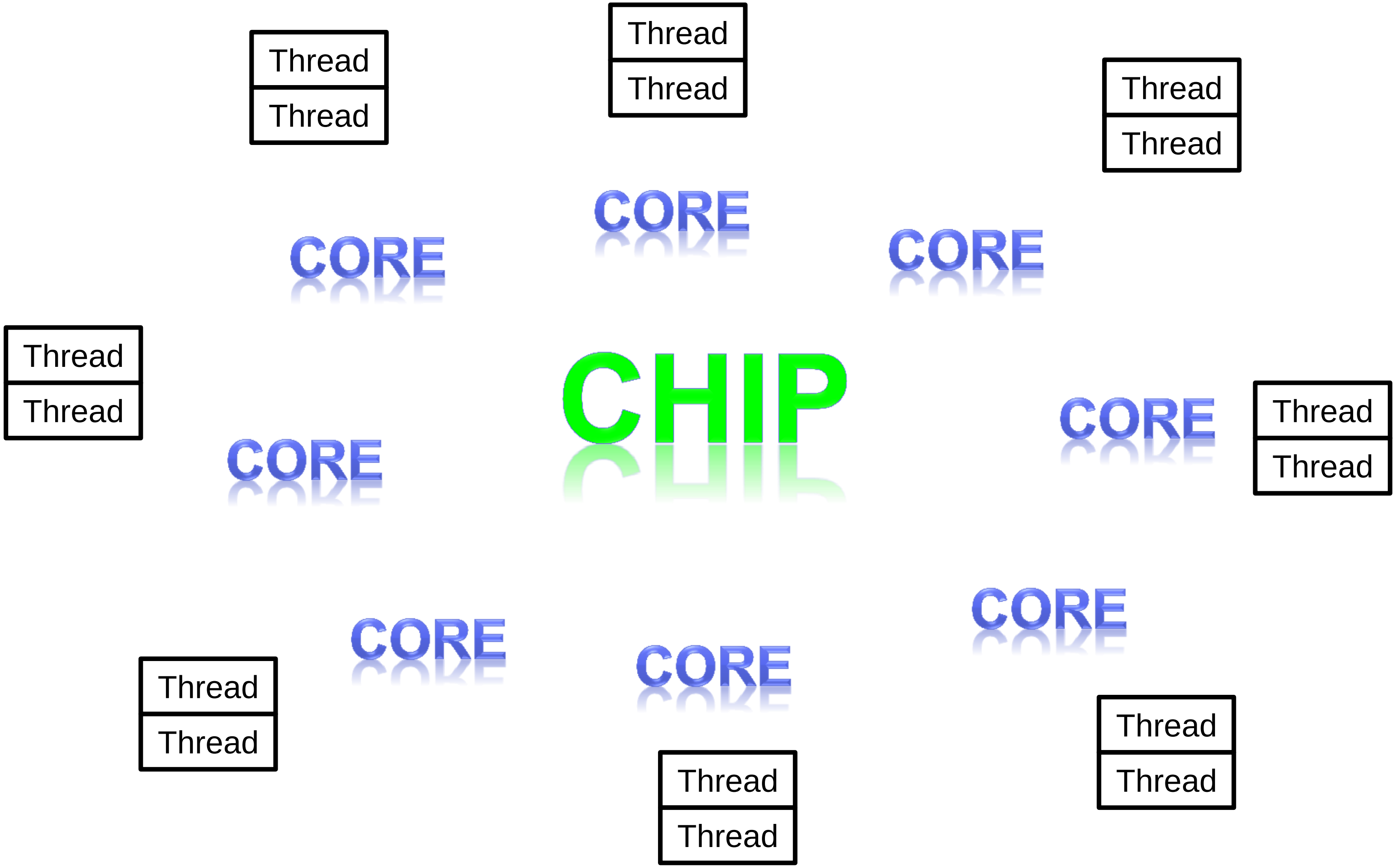
Elapsed Time

🟩 Task A

🟦 Task B

(assumes SMT2 efficiency of 1.4)

Name is Sound and Smoke (Goethe, Faust I)



Thread
Thread

Thread
Thread

Thread
Thread

CORE

CORE

CORE

Thread
Thread

CHIP

CORE

Thread
Thread

CORE

CORE

CORE

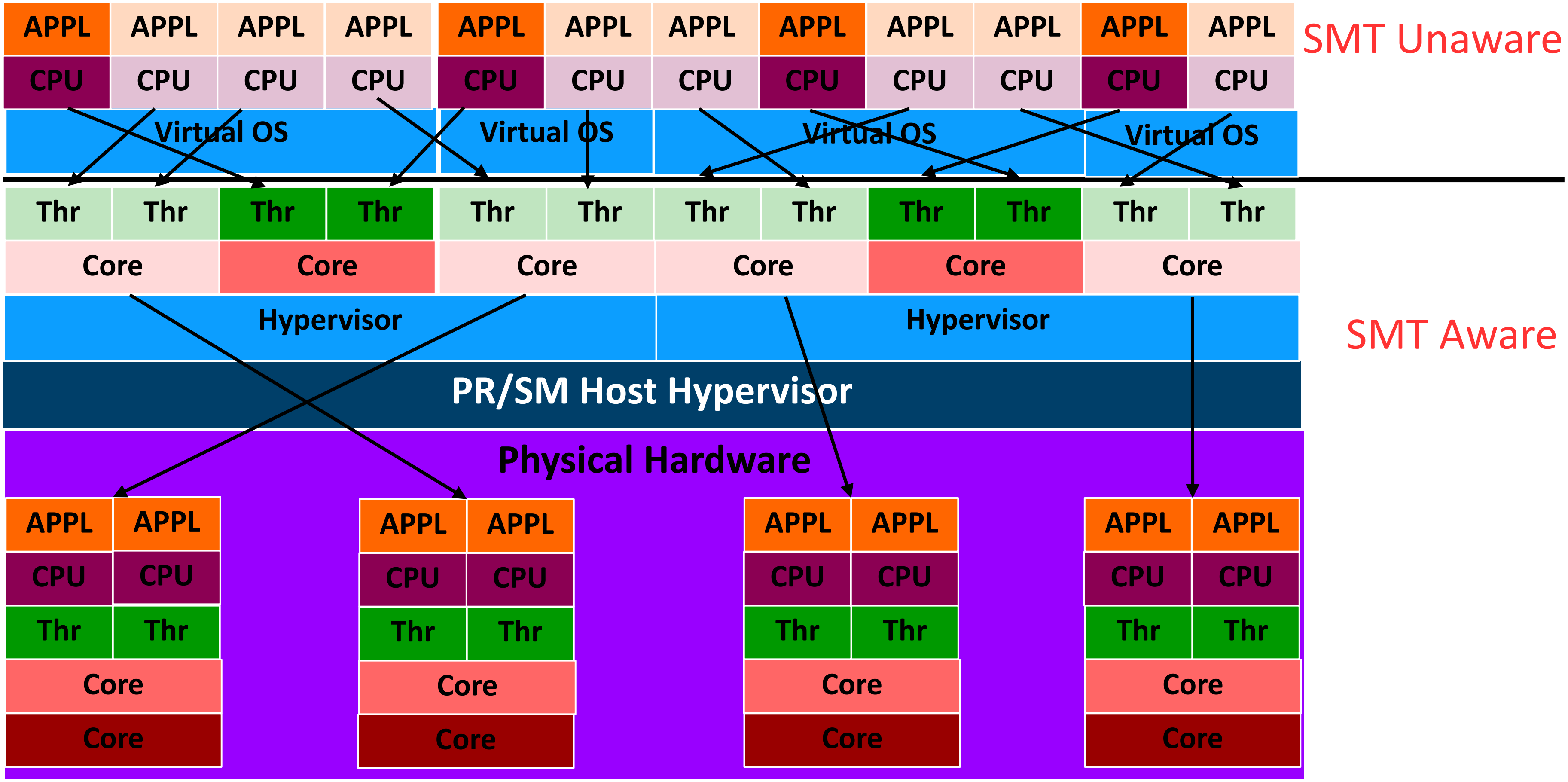
CORE

Thread
Thread

Thread
Thread

Thread
Thread

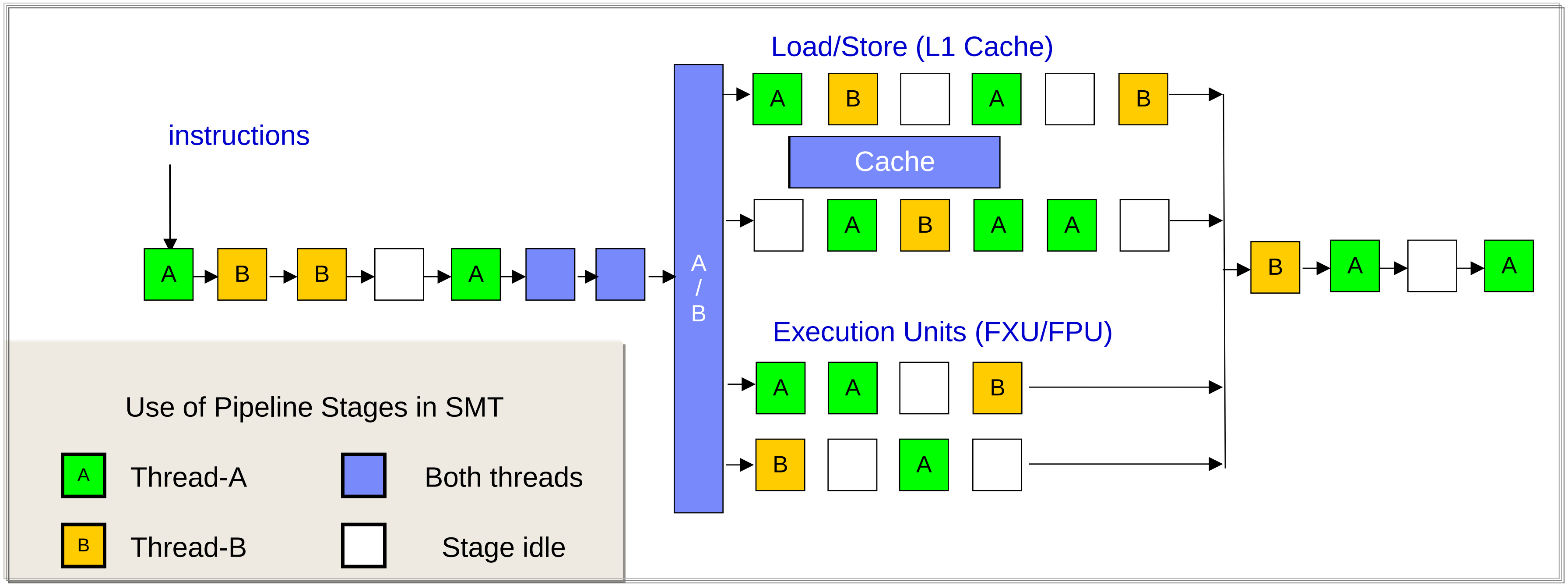
SMT2 – Hypervisor picture



- PR/SM supports SMT for SMT aware hypervisor like z/VM via core dispatching
  - z/VM controls and manages whole core (all threads)
  - SMT transparent to virtual OSes and applications

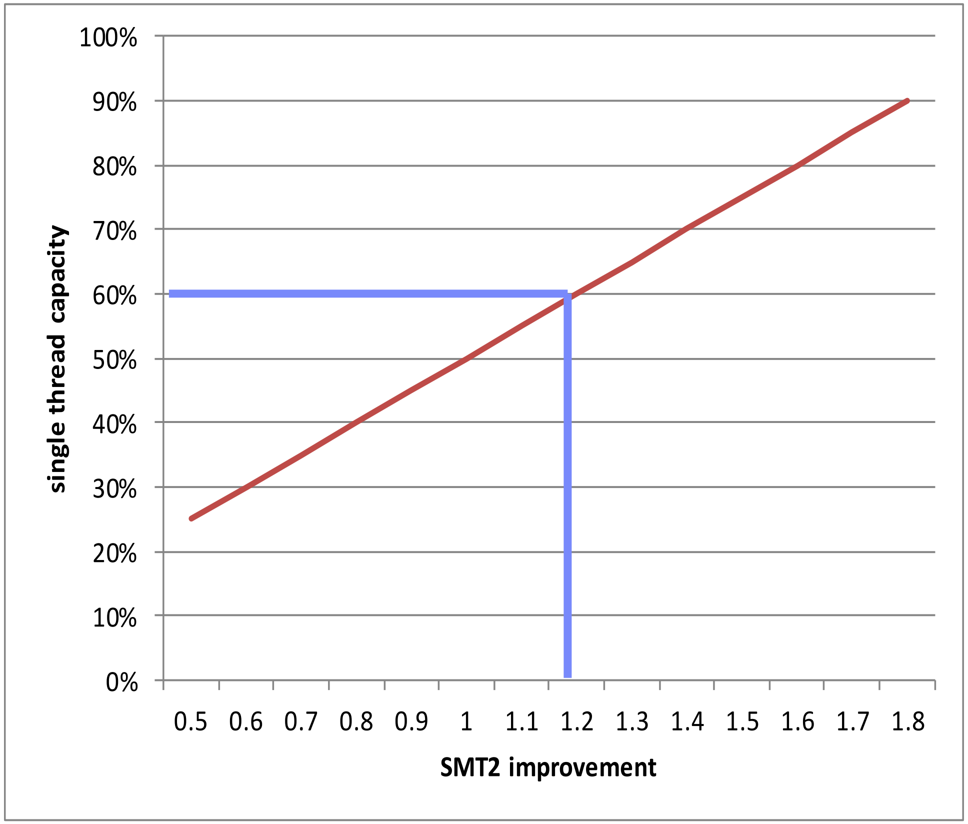
Simultaneous Multithreading – The Technology

- **Simultaneous Multithreading (SMT) technology**
  - Multiple programs (software threads) run on the same processor core
  - More efficient use of the core hardware
- **Active threads share core resources**
  - In space: data and instruction caches, TLBs, branch history tables, etc.
  - In time: pipeline slots, execution units, address translator, etc.
- **Typically Increases overall throughput per core when SMT is active**
  - Amount that increase, varies widely with workload
  - Each thread runs more slowly than on a single-thread core



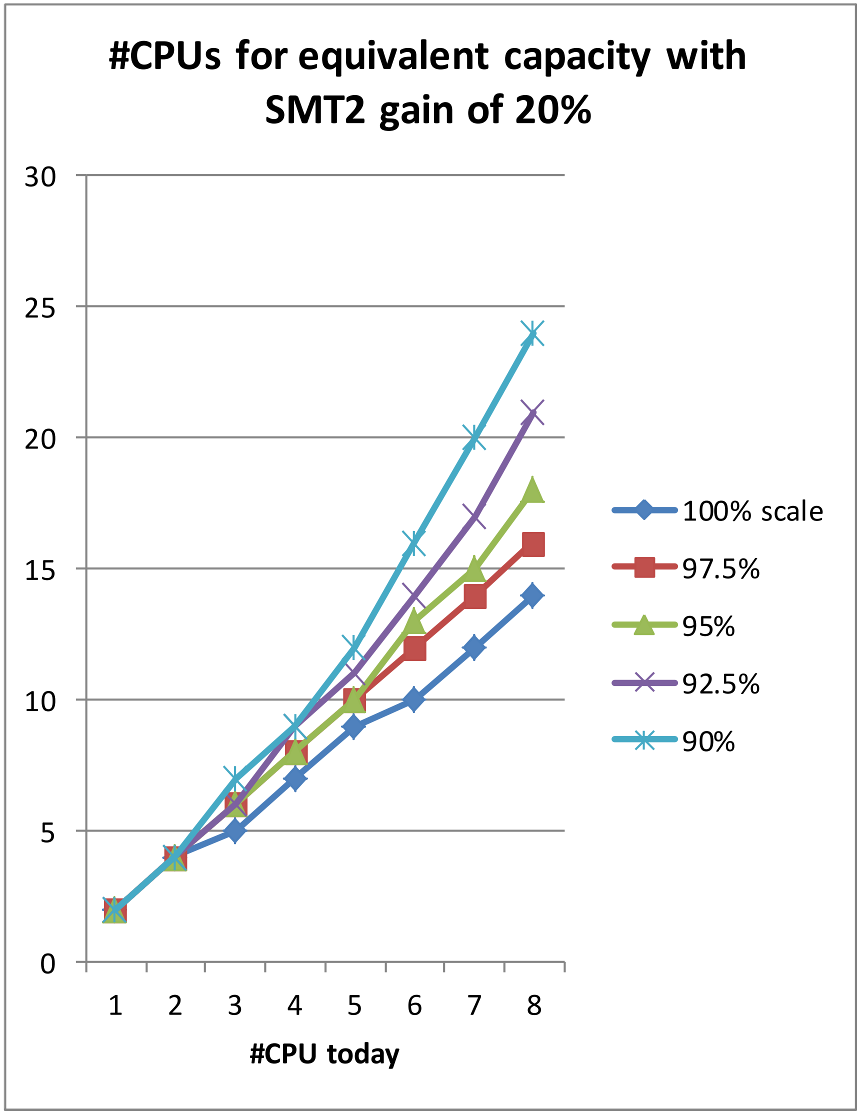
SMT2 implies that each logical CPU is slower

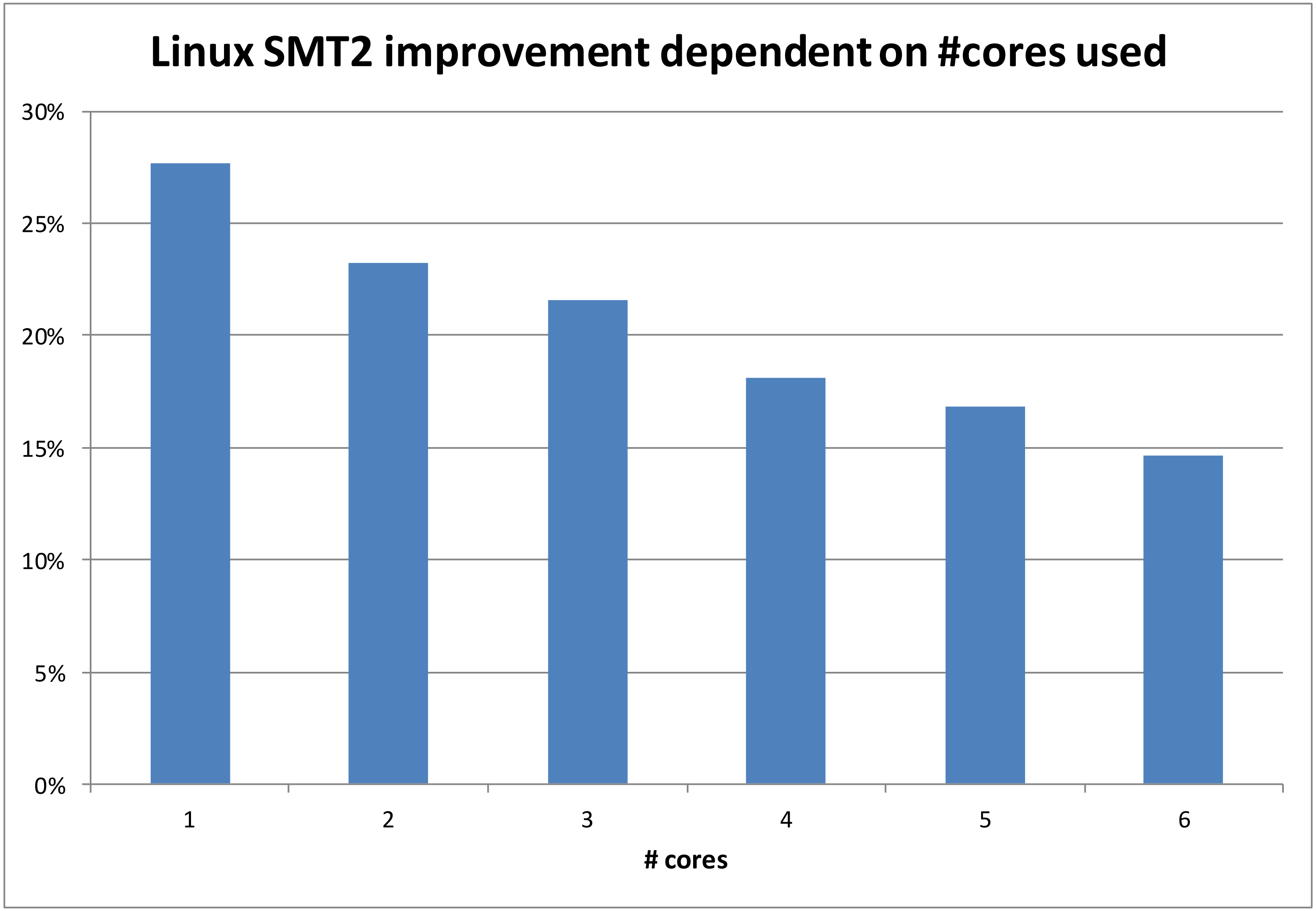
- Evaluate your workload
- single thread speed dependency?
- logwriter process?
- I/O processing?



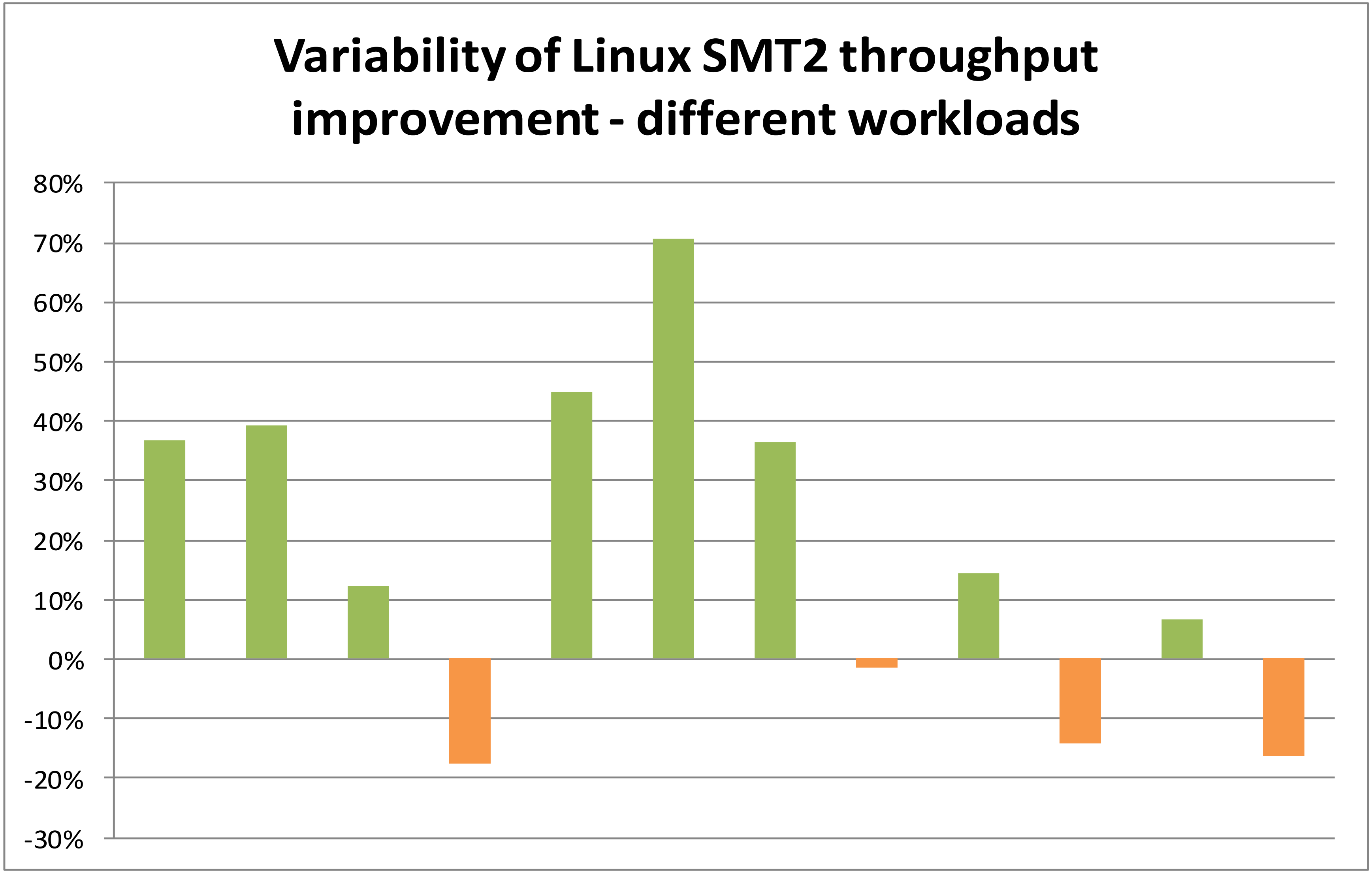
SMT2 requires more logical CPUs

- reduced capacity → more CPUs required
- more CPUs → SMP n-way effect
- workload scalability is really important
- revisit your #virtual CPU sizing
- measure before you deploy





This is one example. There is a lot of variability wrt workload benefit



measure your workload!

ETR / ITR – be careful with calculations

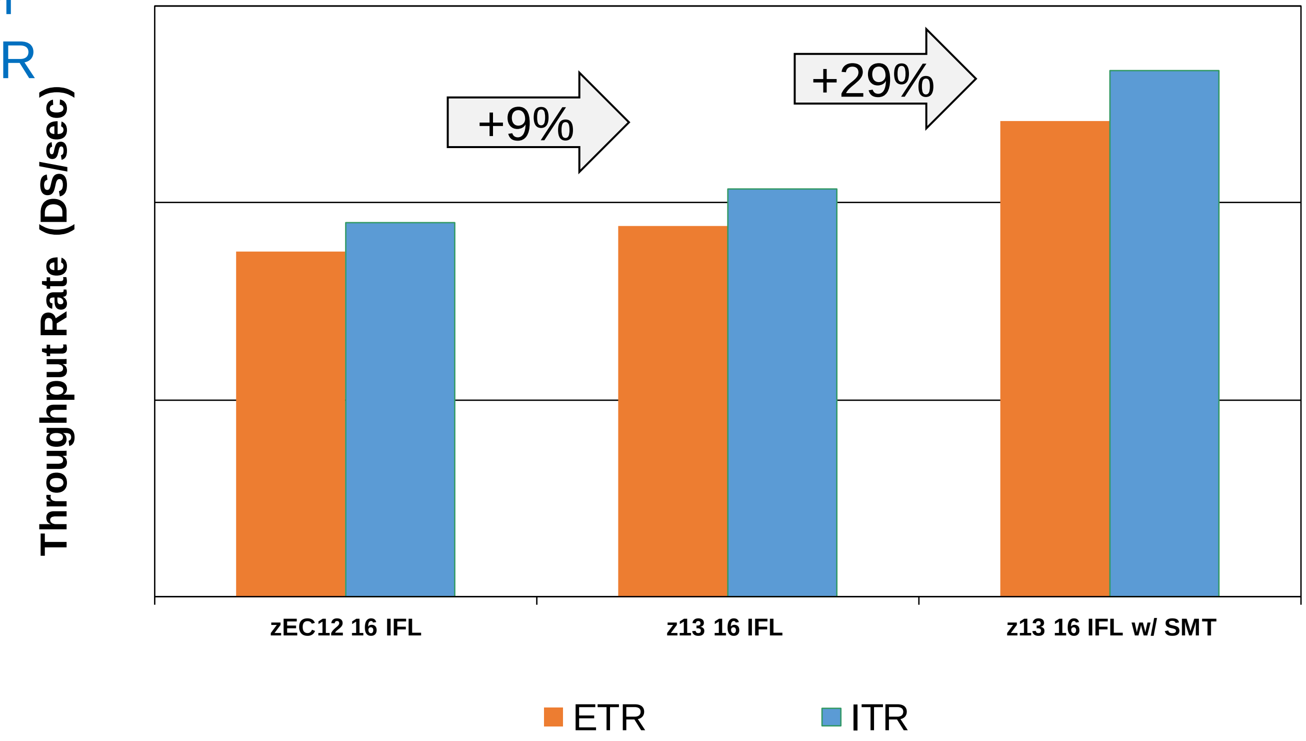
- ETR (External Throughput Rate) / ITR (Internal Throughput Rate)

- $ETR = \frac{\#Transactions}{Elapsed\ time} = ITR\ * CPU\ \ utilization$

- Example with SMT2 enabled
  - 50% of the logical CPUs utilized
  - scheduler puts them on different cores
  - throughput is roughly equivalent to what you get with SMT1, so ~5/6 of total capcity
  - normal calculation ITR = 2*ETR
    - assumes a SMT gain factor of 100%
    - wrong result

SMT real world example

- –SAP Workload
- –2CPs, 2 zIIPs, DB server
- –16 IFL App Server
- –SMT2 with z/VM
- –overall +41% ITR



Agenda

- z13 structure an characteristics
- base performance
- SMT2
- **recommendations and outlook**

Recommendations for enabling SMT

- **z/VM**
  - create a new LPAR with a z/VM that has SMT enabled
  - move one workload (type) / guest at a time
    - remember to increase the # of virtual CPUs
    - check the memory!
    - measure throughput, CPU utilization and response time **before** and **afte**r the movement, keep your monitor record!
  - workloads not showing enough benefit should be run on the z/VM with SMT disabled

- **LPAR**
  - test on separate LPAR with SMT2 turned on
    - you can do this directly on your test LPAR
  - virtual CPUs will automatically double
  - check memory
  - measure throughput, CPU utilization and response time **before** and **afte**r the movement
  - depending on the outcome turn on SMT in the production LPAR

Recommendation OSA5 – layer 3 –TSO

- Intermediate recommendation: disable TCP Segmentation Offload (TSO) if
  - MTU 1492/MTU 1500 is used
  - 10 Gbit Ethernet cards are used
  - the bandwidth is needed
  - you are on RHEL X.X or SLES12

Linux performance fixes coming

- During testing several problems got indentified
- patches are in test
- expect release later this year into the service stream of RHEL6.x, RHEL7.x, SLES11.x, SLES12.x

Linux outlook

- Exploitation coming
  - SMT2
  - SIMD

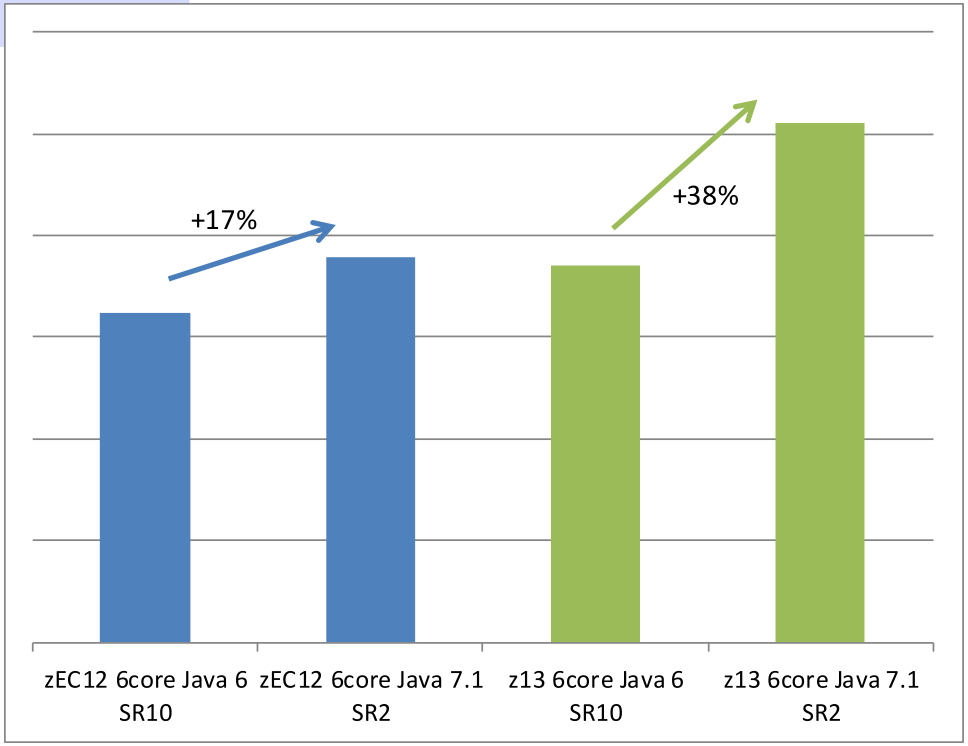
- RHEL 7.x, SLES12.x

- Further compiler and glibc optimizations / exploitations in development
  - hopefully more at one of the next SHARE conferences

- See SHARE session from Matin Schwidefsky **16450: What's New in Linux on System z?**

Toleration is required to ensure that existing JVMs in the field can exploit z9, z10, z196, zEC12 optimizations

Java Release	SR or FP	Aavailability
Java6	SR16 FP3	Jan 2015
Java7	SR8 FP3	Jan 2015
Java7.1	SR2 FP10	Jan 2015



Ensure that you update all the middleware that comes with an embedded Java version

Dr. Eberhard Pasch
epasch@de.ibm.com

Linux on System z – Tuning hints and tips:
http://www.ibm.com/developerworks/linux/linux390/perf/index.html

Mainframe Linux blog: http://linuxmain.blogspot.com