IBM System z Forum

IBM zEnterprise EC12 overview

Simon Williams – zSoftware Client Architect

27th November 2012

Trademark

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

AIX*	DS8000*	Lotus*	System x*	z10 EC
BladeCenter*	FICON*	Power*	System z*	zEnterprise
CICS*	GDPS*	Rational*	System z10*	z/OS*
Cognos*	HyperSwap	Smarter Cities	Tivoli*	z/VM*
DB2*	IMS	Smarter Planet	WebSphere*	z/VSE*

* 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. 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.

- Genealogy
- zEnterprise EC12 technical overview
  - Processor
  - Architecture Extensions
  - Capacity and Performance
- Flash Express
- zAWARE
- Q & A

IBM System z (and zSeries) Generations

N-5

z900
- Announced 10/2000
- 770 MHz
- Up to 16 assignable cores
- CP, IFL, ICF
- Up to 64 GB Memory

z800
- Announced 2/2002
- 625 MHz
- Up to 4 assignable cores
- CP, IFL, ICF
- Up to 32 GB Memory

N-4

z990
- Announced 5/2003
- 1.2 GHz
- Up to 32 assignable cores
- CP, IFL, ICF, zAAP
- Up to 256 GB Memory

z890
- Announced 4/2004
- 1.0 GHz
- Up to 4 assignable cores
- CP, IFL, ICF, zAAP
- Up to 32 GB Memory

N-3

z9 Enterprise Class
- Announced 7/2005
- 1.7 GHz
- Up to 54 assignable cores
- CP, IFL, ICF, zAAP, zIIP
- Up to 512 GB Memory

z9 Business Class
- Announced 4/2006
- 1.4 GHz
- Up to 7 assignable cores
- CP, IFL, ICF, zAAP, zIIP
- Up to 64 GB Memory

N-2

z10 Enterprise Class
- Announced 2/2008
- 4.4 GHz
- Up to 64 assignable cores
- CP, IFL, ICF, zAAP, zIIP
- Up to 1.5 TB Memory

z10 Business Class
- Announced 10/2008
- 3.5 GHz
- Up to 10 cfg cores (5 CP)
- CP, IFL, ICF, zAAP, zIIP
- Up to 248 GB Memory

N-1

zEnterprise 196
- Announced 7/22/2010
- 5.2 GHz
- Up to 80 assignable cores
- CP, IFL, ICF, zAAP, zIIP
- Up to 3 TB Memory

zEnterprise 114
- Announced 7/12/2011
- 3.8 GHz
- Up to 10 cfg cores (5 CP)
- CP, IFL, ICF, zAAP, zIIP
- Up to 256 GB Memory

zEnterprise EC12 technical overview

zEC12 Overview

- Machine Type
  - 2827
- 5 Models
  - H20, H43, H66, H89 and HA1
- Processor Units (PUs)
  - 27 (30 for HA1) PU cores per book
  - Up to 16 SAPs per system, standard
  - 2 spares designated per system
  - Dependant on the H/W model - up to 20, 43, 66,89, 101 PU cores available for characterization
    - Central Processors (CPs), Internal Coupling Facility (ICFs), Integrated Facility for Linux (IFLs), System z Application Assist Processors (zAAPs), System z Integrated Information Processor (zIIP), optional - additional System Assist Processors (SAPs)
  - Sub-capacity available for up to 20 CPs
    - 3 sub-capacity points
- Memory
  - RAIM Memory design
  - System Minimum of 32 GB
  - Up to 768 GB per book
  - Up to 3 TB for System and up to 1 TB per LPAR
    - 32 GB Fixed HSA, standard
    - 32/64/96/112/128/240/256 GB increments
  - Flash Express
- I/O
  - 6 GBps I/O Interconnects – carry forward only
  - Up to 48 PCIe interconnects per System @ 8 GBps each
  - Up to 4 Logical Channel Subsystems (LCSSs)
    - Up to 3 Sub-channel sets per LCSS
- STP - optional (No ETR)

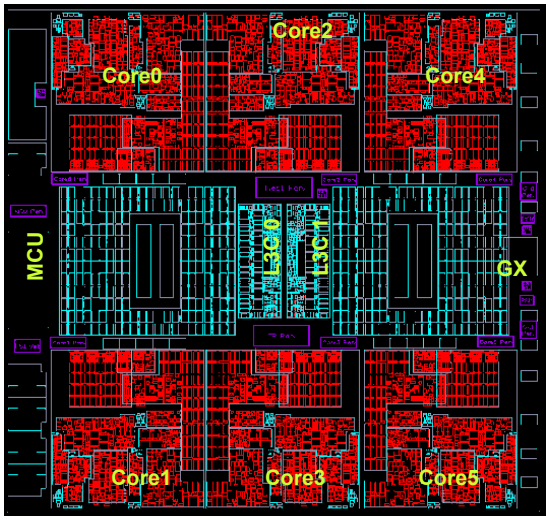
zBX Overview

- Machine Type/Model 2458-003
- Racks – Up to 4 (B, C, D and E)
  - 42U Enterprise, (36u height reduction option)
  - 4 maximum, 2 chassis/rack
  - 2-4 power line cords/rack
  - Non-acoustic doors as standard
  - Optional Acoustic Doors
  - Optional Rear Door Heat Exchanger (conditioned water required)
- Chassis – Up to 2 per rack
  - 9U BladeCenter
  - Redundant Power, cooling and management modules
  - Network Modules
  - I/O Modules
- Blades (Maximum 112 single width blades in 4 racks)
  - Customer supplied POWER7 Blades (0 to 112)
  - Customer supplied IBM System x Blades (0 to **56**)
  - DataPower XI50z, M/T 2462-4BX (0 to 28 – double width)
- Management Firmware
  - Unified Resource Manager
- Top of Rack (TOR) Switches - 4
  - 1000BASE-T intranode management network (INMN)
  - 10 GbE intraensemble data network (IEDN)
  - GbE IEDN for customer network
- Network and I/O Modules in the BladeCenter
  - 1000BASE-T and 10 GbE modules
  - 8 Gb Fibre Channel (FC) connected to customer supplied disks

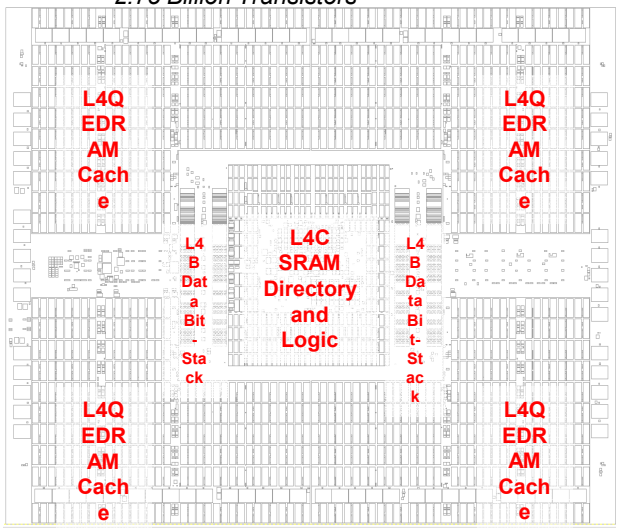
zEC12 Processor

zEC12 Processor Design

- Built on solid foundation of z196
  - Leverage IBM 32nm SOI technology with eDRAM (embedded Dynamic RAM – allows combination of dense DRAM caches with high-speed logic)
- Improved high-frequency (5.5 GHz) and O-o-O Execution (2nd gen)
  - Improved grouping of instructions means the instruction pipeline streamlined for smoother flow
  - Better branch prediction – added BTB2 (Branch Target Buffer)
  - Faster engine for fixed-point division
  - Millicode performance improvements
- 23 new instructions
- Cache hierarchy leadership extended
  - New structure for 2nd-level private cache
    - Separate optimizations for instructions and data
    - Reduced access latency for most L1 misses
  - 3rd-level on-chip shared cache doubled to 48MB
  - 4th-level book-shared cache doubled to 384MB
  - Focus on keeping data closer to the PU
- More processors in the same package as z196
  - 6 processor cores per CP chip
  - Crypto/compression co-processor per core (no longer shared between two cores)
  - Same power consumption as z196

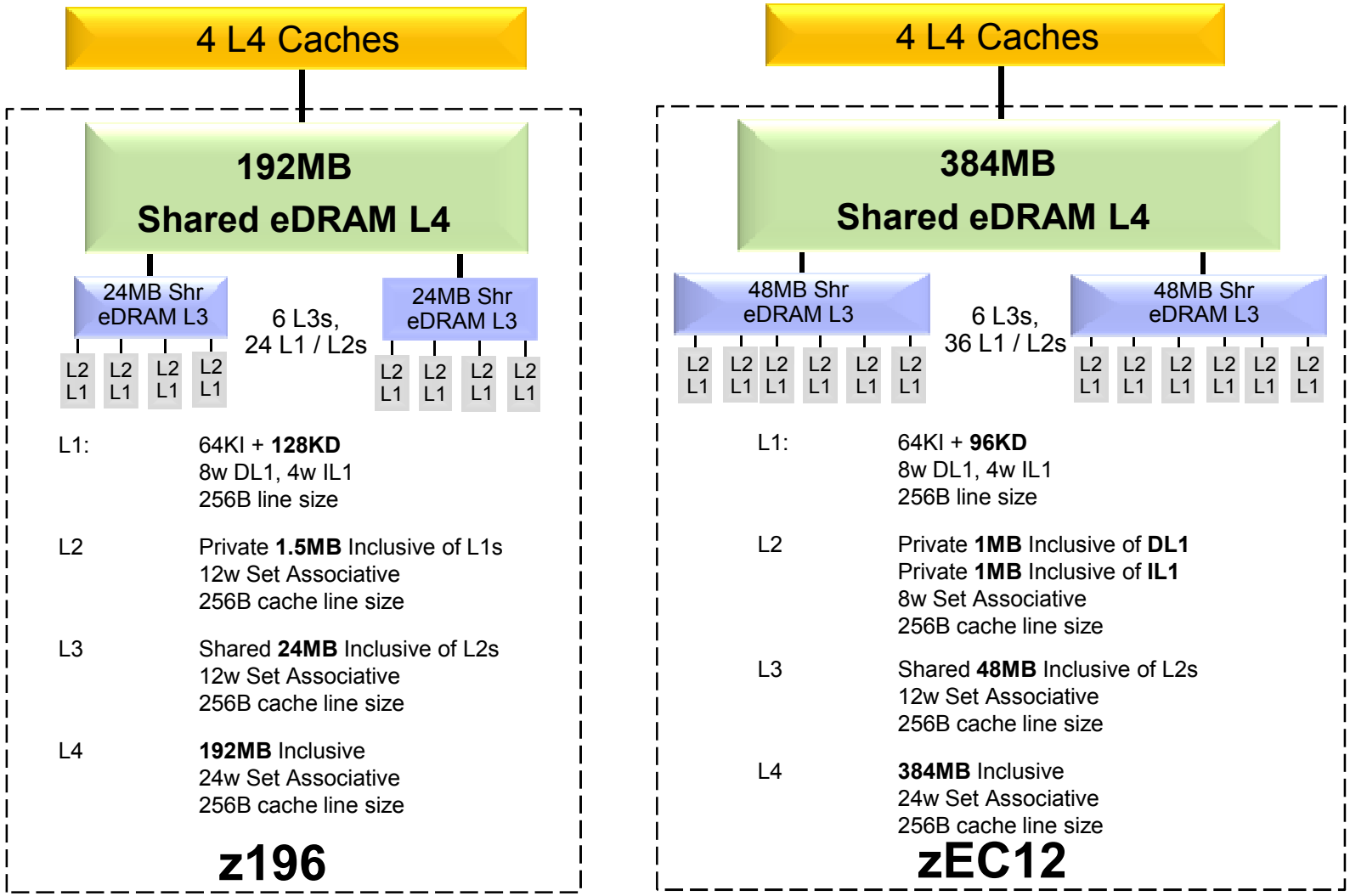


zEC12 PU Chip: 6 cores, 598 mm² chip
2.75 Billion Transistors



zEC12 SC Chip: 192MB cache, 526 mm² chip
3.3 Billion Transistors

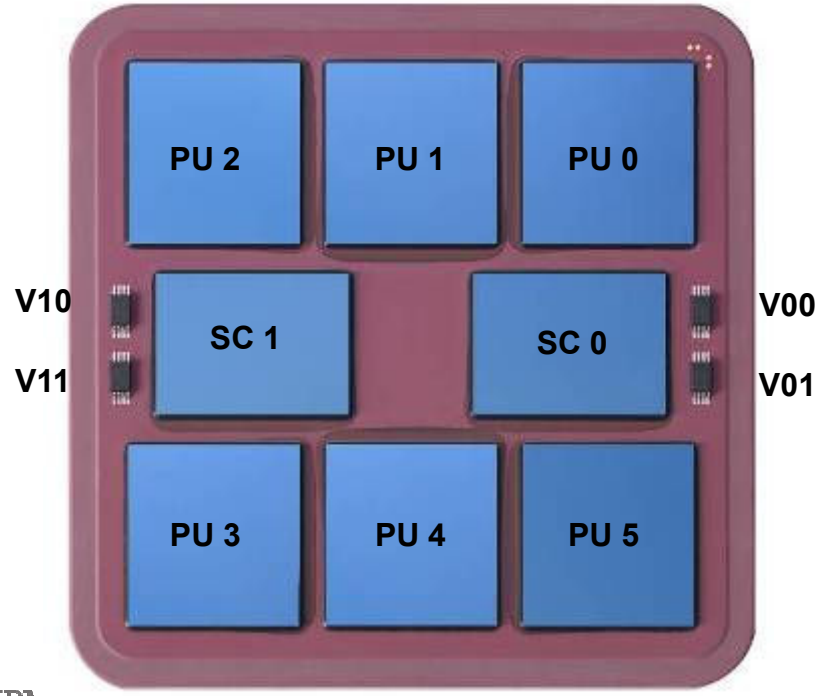
System z Cache Topology – z196 vs. zEC12 Comparison



All together, a zEC12 system containing four books—each with six microprocessor chips and two hub chips—has an aggregate of 1.5 GB of L4 caches and more than 1 GB of L3 caches

zEC12 Multi-Chip Module (MCM) Packaging

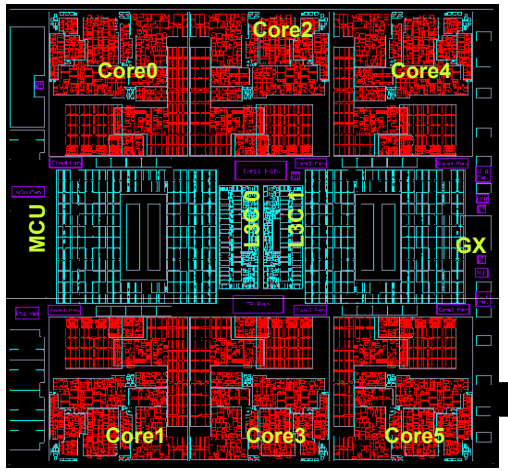
- 96mm x 96mm MCM
  - 102 Glass Ceramic layers
  - 8 chip sites
- 7356 LGA connections
  - 27 and 30 way MCMs
  - Maximum power used by MCM is 1800W



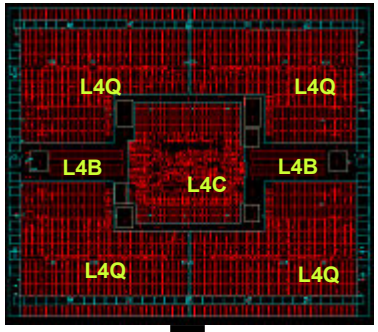
- CMOS 13s chip Technology
  - PU, SC, S chips, 32nm
  - *6 PU chips/MCM – Each up to 6 active cores*
    - 23.7 mm x 25.2 mm
    - 2.75 billion transistors/PU chip
    - L1 cache/PU core
      - 64 KB I-cache
      - 96 KB D-cache
    - L2 cache/PU core
      - 1 MB I-cache
      - 1 MB D-cache
    - L3 cache shared by 6 PUs per chip
      - 48 MB
    - 5.5 GHz
  - *2 Storage Control (SC) chip*
    - 26.72 mm x 19.67 mm
    - 3.3 billion transistors/SC chip
    - L4 Cache 192 MB per SC chip (384 MB/Book)
    - L4 access to/from other MCMs
  - 4 SEEPROM (S) chips – 1024k each
    - 2 x active and 2 x redundant
    - Product data for MCM, chips and other engineering information
  - Clock Functions – distributed across PU and SC chips
    - Master Time-of-Day (TOD) function is on the SC

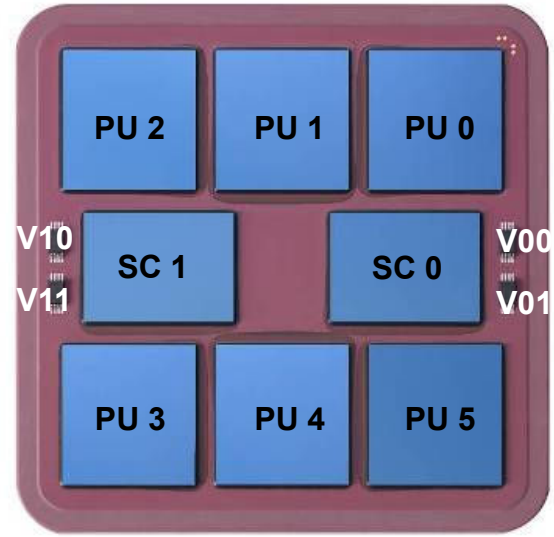
zEC12 PU chip, SC chip and MCM

zEC12 Hexa-core PU chip



SC chip





MCM

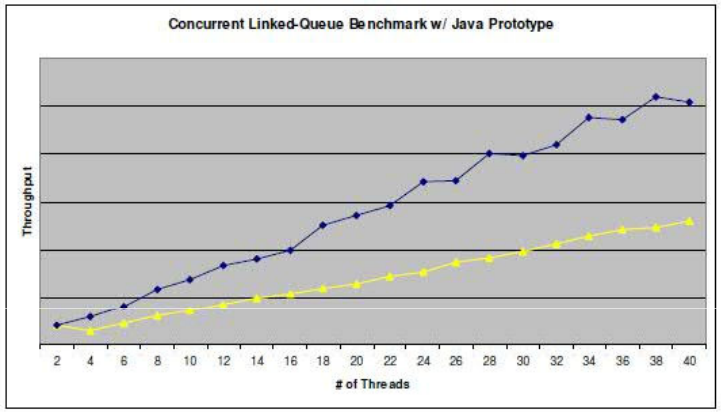
BOOK Side View

Front View Fanouts

zEC12 Architecture Extensions

Transactional Execution – hardware/software synergy

- Transactional execution (a.k.a. Hardware Transactional Memory)
  - Software-defined sequence treated by hardware as atomic “transaction”
    - “All or nothing” execution
    - New ‘wrapper’ instructions
  - Implemented in core hardware
- Monitor storage locations accessed
    - Buffer updates until transaction completes
    - Auto-retry for “constrained” transactions
  - Enables significantly more efficient software
    - Highly-parallelized applications
    - Speculative code generation
    - Lock elision (avoidance of pessimistic locking)
  - Staged software exploitation plan
    - Initial support in Java (Java7SR3) - no source code changes required
    - XL C/C++ compiler for z/OS V1R13) – other languages to follow
    - Leverage for high-n-way scaling enhancements

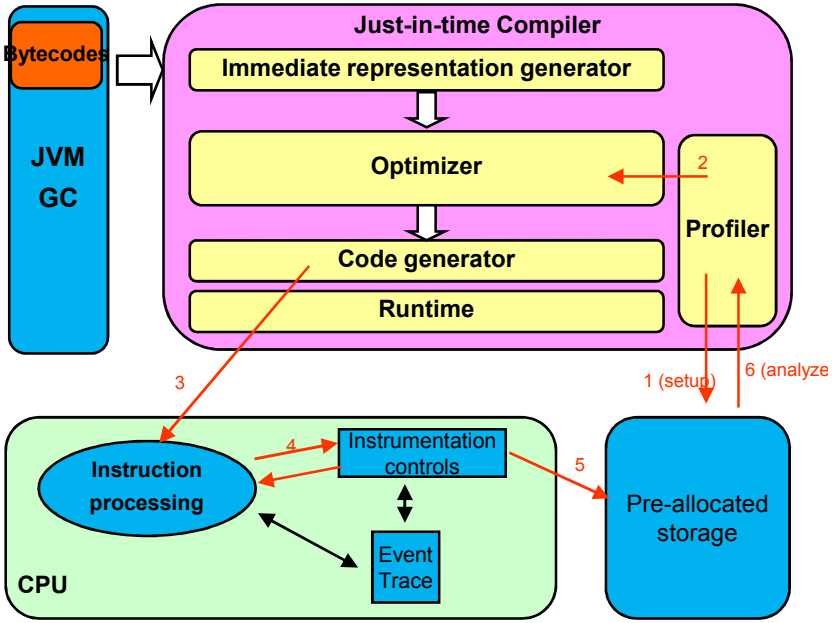


1st general purpose processor to support hardware transactional memory.

Performance gains of up to 45% seen in initial testing

Run-time Instrumentation

- A new hardware facility for managed runtimes
  - Tailored towards Java Runtime Environment (JRE) / Just-in-Time (JIT) compiler
  - Low cost (overhead) profiling
- Allows dynamic optimization on code generation as it is being executed:
  - Enhances JRE decision-making by providing real-time feedback on the execution



- Key features
  - A collection buffer capturing a run-time trace till an instruction sample point, providing
    - "how we got here information", e.g. branch history
  - Meta-data collected for "what happened" information with the sample instruction
    - Cache miss
    - Branch prediction/resolution
  - 3 modes of sampling; by
    - cycle count, instruction count, or explicit indication
    - Sample reports include hard-to-get information

      Event traces, e.g. taken branch trace

      "costly" events of interest, e.g. cache miss information
  - Not the same as current CPU Measurement Facility (CPUMF)
    - Both can run concurrently
    - With the benefit of CPUMF keeping tabs on how RI is actually running and affecting things.

1st microprocessor to support run-time instrumentation

zEC12 Architecture Extensions

- **2 GB page frame support**
  - Increased efficiency for DB2 buffer pools, Java heaps and other large structures
  - Better performance by decreasing the number of TLB misses that an exploiter application incurs
  - Less time spent converting virtual addresses into physical addresses
  - Less real storage used to maintain DAT structures

- **Software directives to improve hardware performance**
  - Data usage intent improves cache management
    new *Next Instruction Access Intent (NIAI) Instruction*
  - Branch pre-load improves branch prediction effectiveness
    new *Branch Prediction Preload (BPP)* and *Branch Prediction Relative Preload (BPRP) instructions*

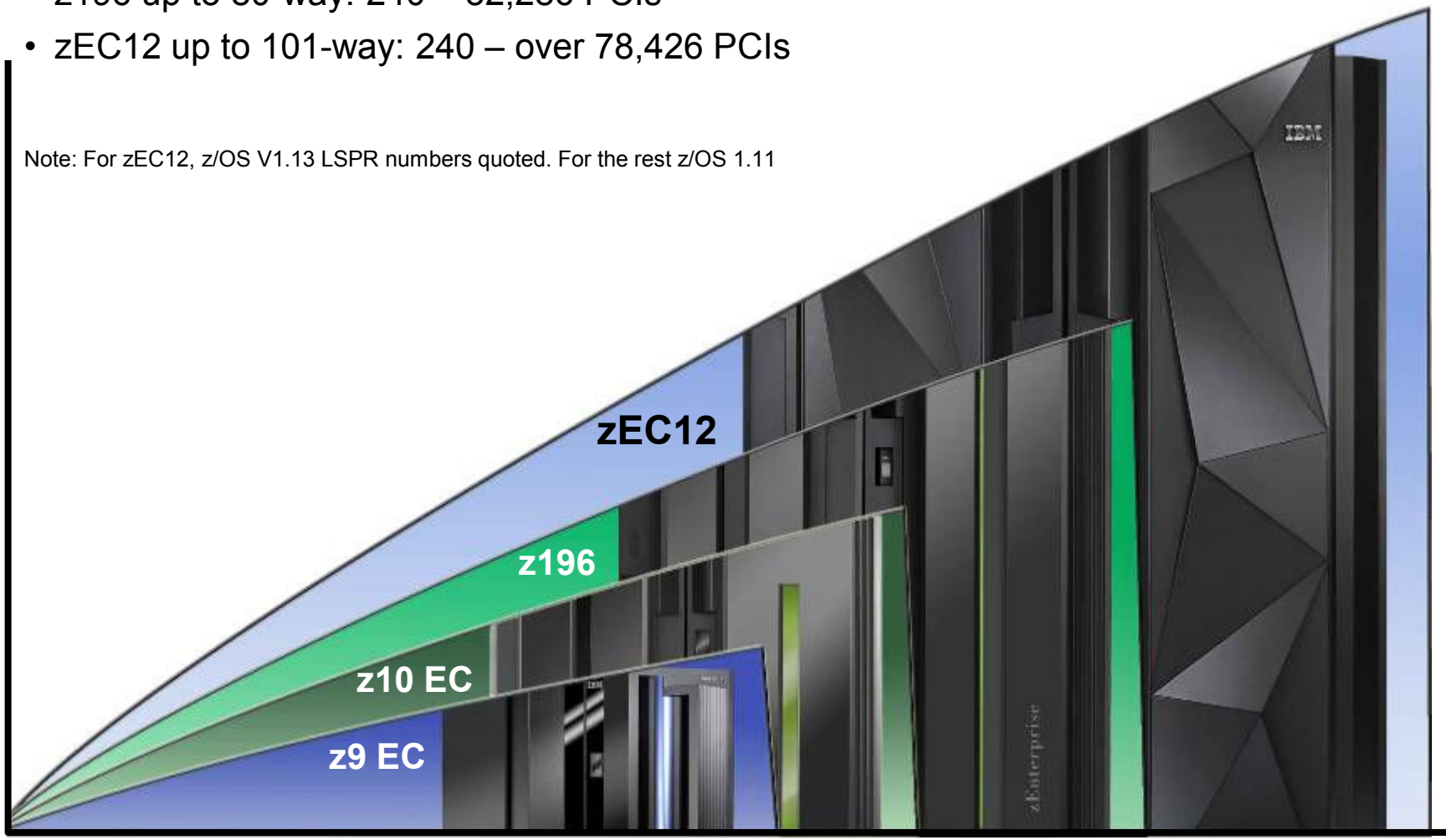
- **Decimal format conversions**
  - Enable broader exploitation of Decimal Floating Point facility by COBOL programs (exploited by the latest COBOL compiler)

zEC12 Capacity and Performance

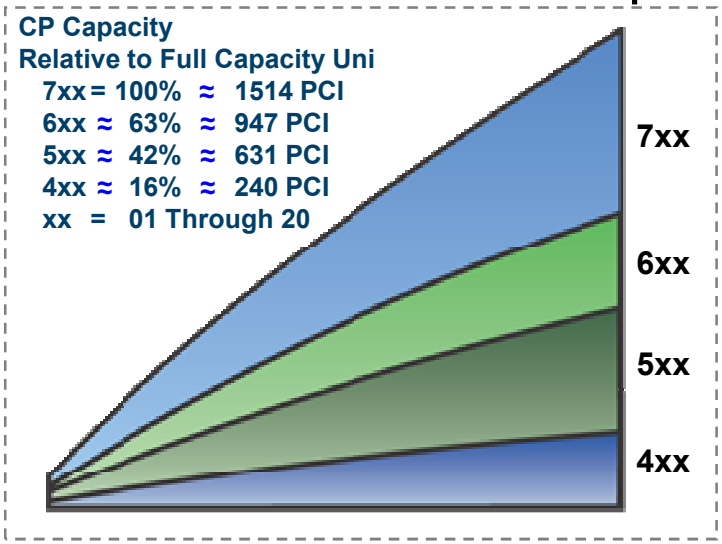
zEC12 Vs z196 Vs z10 EC Vs z9 EC capacity comparison

- z9 EC up to 54-way: 193 – 18,505 PCIs
- z10 EC up to 64-way: 214 – 31,826 PCIs
- z196 up to 80-way: 240 – 52,286 PCIs
- zEC12 up to 101-way: 240 – over 78,426 PCIs

Note: For zEC12, z/OS V1.13 LSPR numbers quoted. For the rest z/OS 1.11

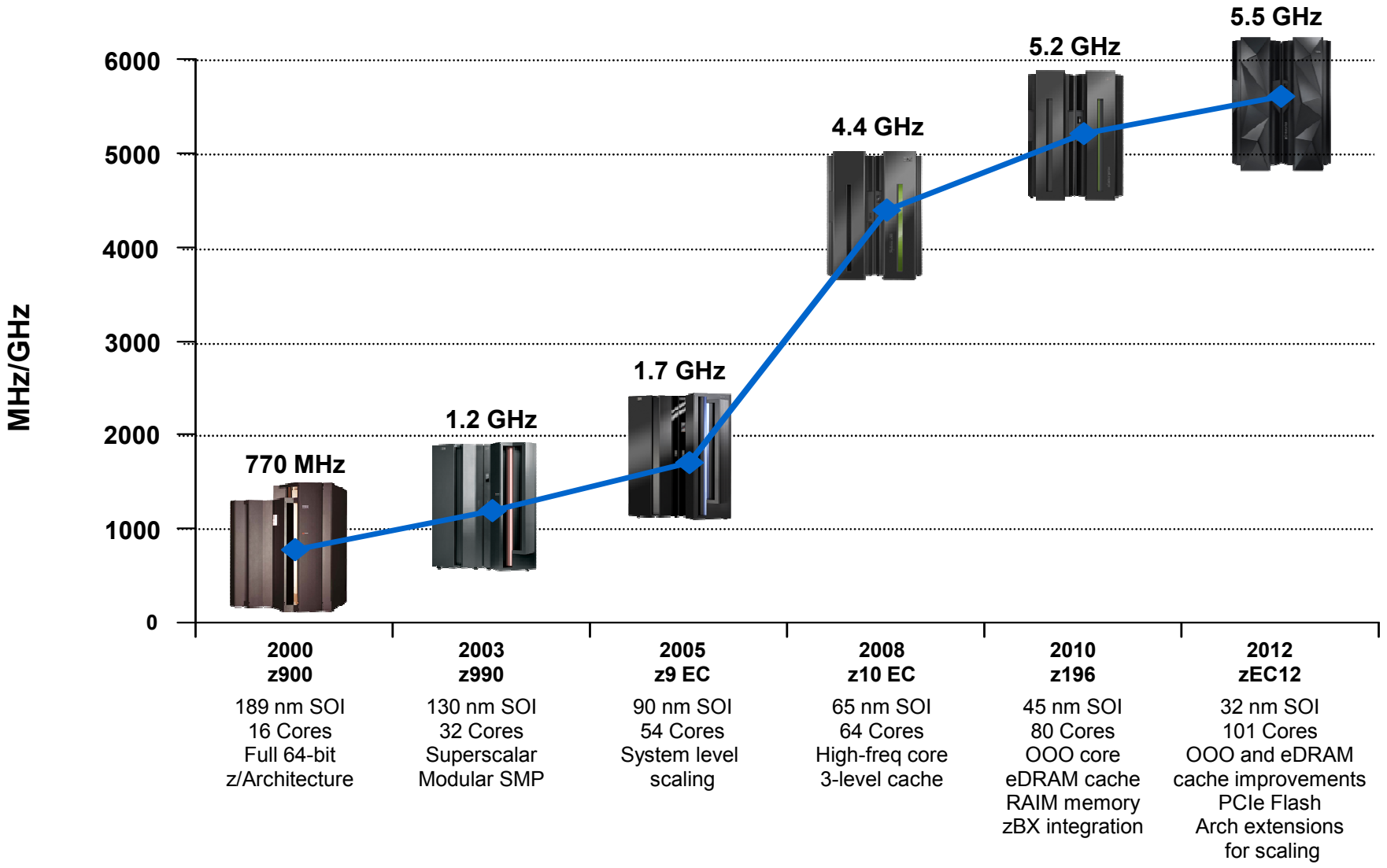


zEC12 Full and Sub-Capacity CP Offerings



- Subcapacity CPs, up to 20, may be ordered on ANY zEC12 model. If 21 or more CPs are ordered all must be full 7xx capacity
- All CPs on a zEC12 CPC must be the same capacity
- All specialty engines run at full capacity. The one for one entitlement to purchase one zAAP and one zIIP for each CP purchased is the same for CPs of any capacity.
- Only 20 CPs can have granular capacity but other PU cores may be characterized as full capacity specialty engines
- For no CPs, the capacity setting is 400
- PVU for zEC12 is 120

Frequency evolution



5.7% improvement in clock speed from z196 to zEC12…but 25% improvement in engine capacity

IBM zEnterprise EC12: An optimized system

Latest Java Runtime performance numbers…

- Java 7 SR3, a no charge upgrade, shipped 8th November 2012
- ***60% improvement*** for Java Multi-Threaded Benchmark vs z196 (z/OS and Linux on System z)
- ***~40% improvement*** over z196 running the CPU Intensive benchmark
- WAS8.5 running DayTrader2.0 EJB ***improves throughput 66%*** vs z196 and WAS7.0
- WAS8.5 Liberty
  - TradeLite ***83% improved*** vs. WAS8.5 on z196 (Servlet and JSP throughput)
  - Up to ***5x start-up time reduction*** (<5s), ***reduced RAM requirements*** (up to 81%)***, improved zAAP offload to 95%+***
- IMS Java Message Processing region performance up to **32% improved** vs z196
- **~30% improvement** for ILOG WebSphere Operational Decision Management on zEC12 vs z196

Software Performance Improvements

- **New hardware functions optimized for software performance**
  - Up to **30%** improvement in **IMS** throughput due to faster CPU, cache and compliers
  - Up to **31%** improvement to **PL/I** based CPU intensive applications
  - Up to **30%** improvement in throughput for **DB2** for z/OS operational analytics[2]
  - More than **30%** improvement in throughput for **SAP** workloads[1]
  - Up to **27%** improvement in throughput for CPU intensive integer & float **C/C++** applications[1]

[1] Based on preliminary internal measurements and projections
[2] As measured by the IBM 9700 Solution Integration Center. The measured operational BI workload consists of 56 concurrent users executing a fixed set of 160,860 Cognos reports . Compared DB2 v10 workload running on IBM's z196 w/10 processors to an zEC12 w/10 processors

Flash Express

Flash Express – What is it?

- Physically comprised of internal storage on NAND Flash SSDs
- Used to deliver a new tier of memory: Storage Class Memory (SCM)
- Supported on z/OS V1.13 plus web deliverable
- IBM is working with the Linux Distribution partners to include support in future Linux on System z distribution releases

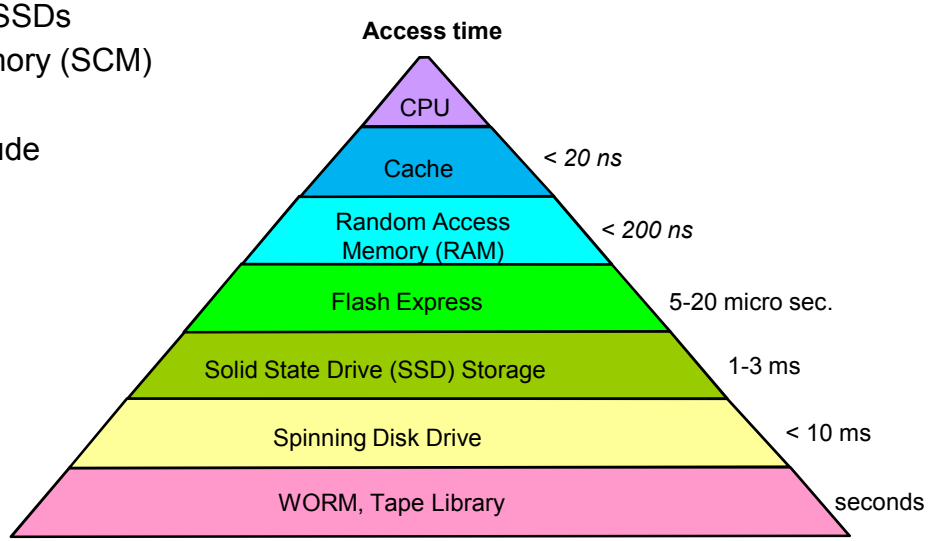
- Uses PCIe I/O drawer
- A Card Pair is Treated as one FRU.

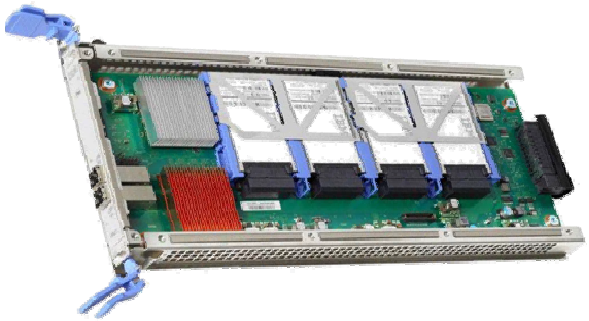
- Sized to accommodate *all LPAR paging*
  - A **card pair** provides **1.4 TB** usable storage
  - Maximum 4 card pairs (4 X1.4=5.6 TB)

- Immediately usable
  - No capacity planning needed
  - No intelligent data placement needed
  - Full virtualization across partitions
  - No HCD/IOCP requirements

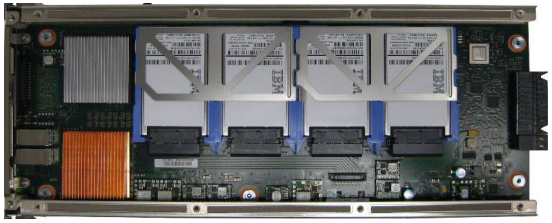
- Robust design
  - Delivered as a **RAID10** mirrored pair
  - Designed for long life
  - Designed for concurrent firmware upgrade

- Secured
  - Flash Express adapter is protected with 128-bit AES encryption.
  - Key Management provided based on a Smart Card





z/OS FLASH Use Cases



1) Paging

•z/OS paging subsystem will work with mix of internal Flash and External Disk
•Self Tuning based on measured performance
•Improved Paging Performance, Simplified Configuration

2) Dumping

•Is expected to yield substantial improvements in SVC dump data capture time (reduced times)
•Reduce Stand Alone Dump duration (Read time for paged out data)

3) Pageable Large Page exploiters by 14th December 2012 (NB: Pageable Large Pages are only exploited by Flash Express)

•z/OS V1.13 *Language Environment*

•*IMS 12 Common Queue Server*

•*DB2 10 **

•*Java SDK601 SR4, and Java SDK7 SR3* and by extension exploiters such as
- *CICS Transaction Server 5.1*
- WAS Liberty Profile v8.5
- Traditional WAS 8.0.0x and Traditional WAS 8.5.5 (future) **

*DB2 date to be determined. Support for V10 with APARs is planned

Flash Express Performance benefits

- The **WAS Day Trader** 64-bit showed **8%** performance improvement using Flash. The test used Java 7 SR3 with JIT code cache & Java Heap leveraging Flash and PLP.

- **SVC Dump** is faster, so higher availability for workloads directly involved as well as other impacted workloads
  - Transaction steady state reached in **14 seconds** with Flash Express vs **73 seconds** DASD (80% reduction)

- Using Flash Express for **workload transition**, peak throughput reached **23% faster** than DASD
  - Paging to Flash reduced **response by 90% and increased throughput by 37%** in first 45 seconds

- **DB2** Initial results: Pageable Large Pages for DB2 helps DB2 achieve up to a **3% transaction throughput improvement** from CPU savings. The savings are due to reduced buffer pool management.

z/OS Java SDK 7:16-Way Performance Shows up to 60% Improvement

Aggregate 60% improvement from zEC12 and Java7SR3

- zEC12 offers a ~45% improvement over z196 running the Java Multi-Threaded Benchmark
- Java7SR3 offers an additional ~13% improvement (-Xaggressive + Flash Express pageable 1Meg large pages)

IBM zAware

Systems are more complex and more integrated than ever

- Errors can occur anywhere in a complex system
- Difficult to detect, difficult to diagnose, symptoms / problems can manifest hours or days later
- Problems can grow, cascade, snowball
- Volume of data is unmanageable – need information and insight.
- Systematic ‘soft failures’ (a.k.a. “sick but not dead”) much harder to detect – anomalies can build up over time

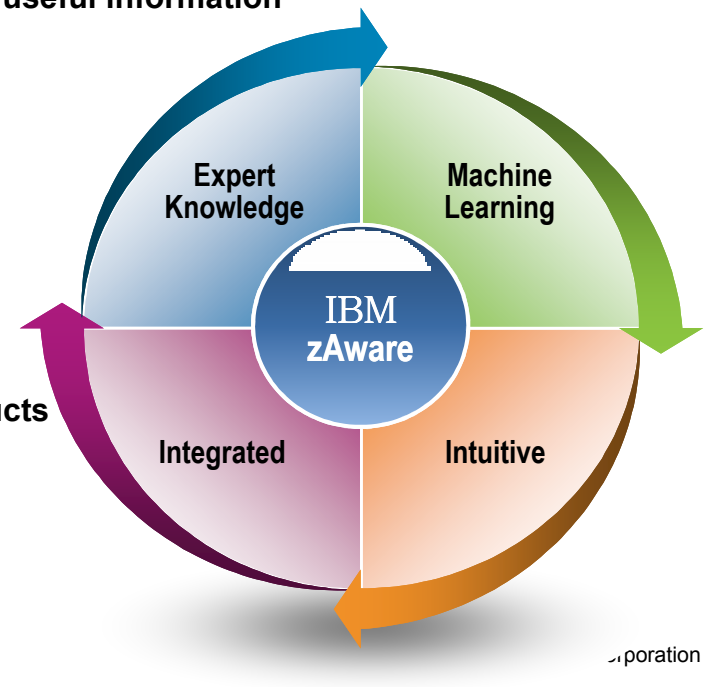
IBM zAware - *IBM System z Advanced Workload Analysis Reporter*

- Cutting edge **pattern recognition analytics** looks at the health of a z/OS system
  - **Perform machine learning, pattern recognition, and statistical analysis on streaming messages** to look for unexpected patterns to give faster, more pinpointed recognition of problems

- A 'watch dog' to detect unusual behavior of z/OS images in near real time – enabling you to act on system issues sooner - pushes z/OS high availability even beyond what it is today
  - Diagnose problems/ critical events/ outages while they are occurring in real time
  - Helps heighten awareness of small problems so they can be corrected quickly
  - Determine the cause of problems so the operation team can establish procedures to prevent a reoccurrence

- It can **analyse huge amounts of OPERLOG messages and turn it into useful information**
  - Works "out of the box" with little customization
  - A single browser based view
  - **Out of band – minimal effect on existing workloads**
  - Complementary to existing tools

- **Can monitor across a parallel sysplex (and multiple sysplexes)**
- Detects anomalies monitoring systems miss:
  - Messages may be suppressed or rare
  - Messages may indicate a trend
- **XML Output consumable through published API, can drive ISV products**

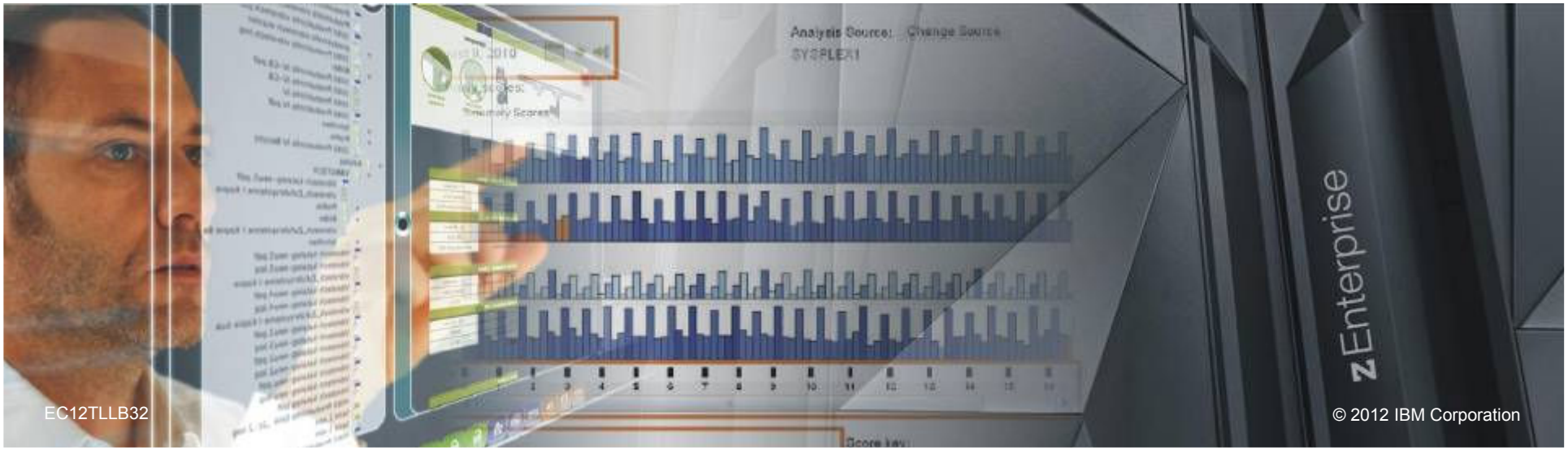


Critical Questions Answered by IBM zAware

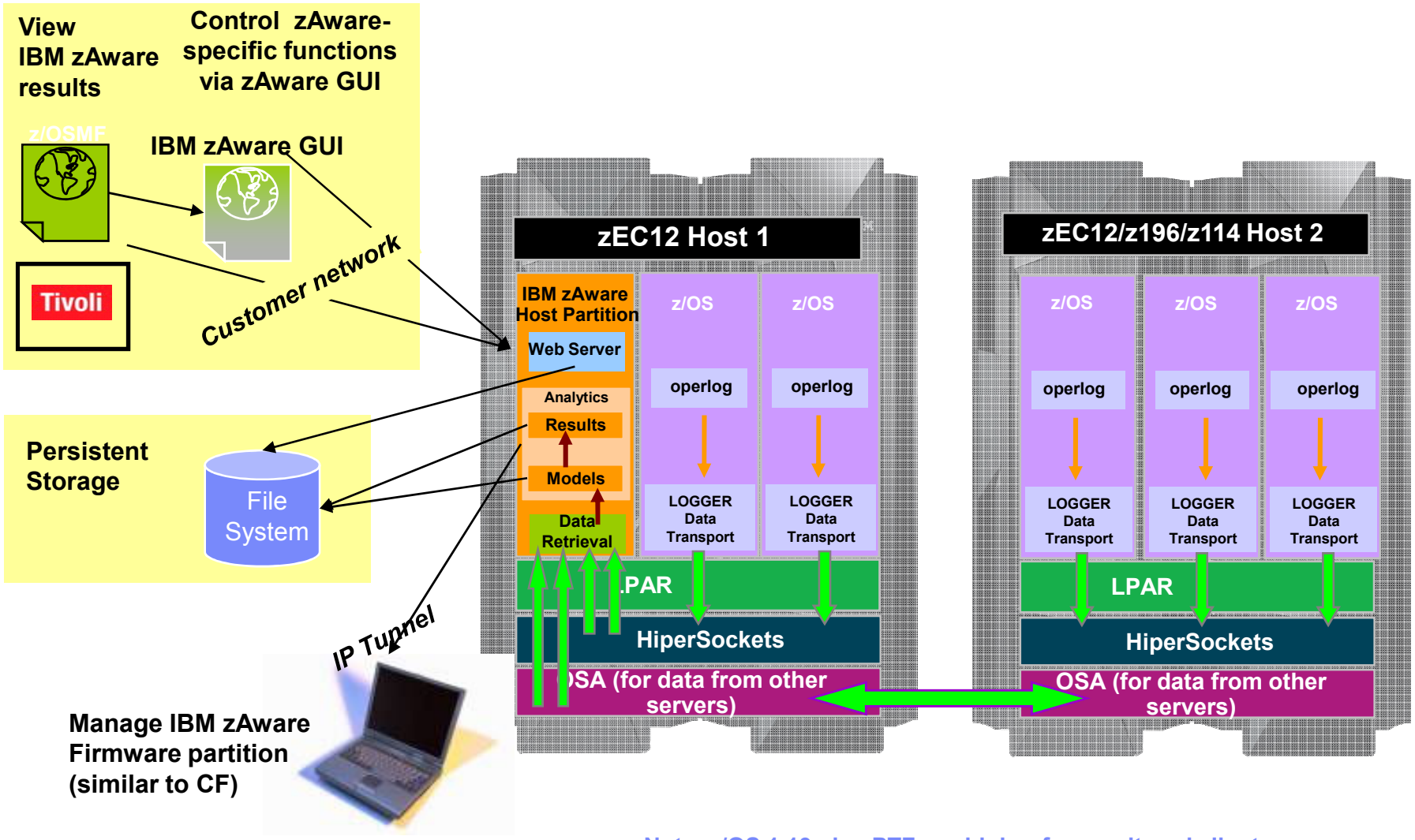
- When a problem occurs
  - **Which z/OS LPAR is behaving abnormally?**
    - Lots of unique messages
    - High score generated by unusual message ids or an unusual patterns of message ids
  - **When did the z/OS LPAR start to behave abnormally?**
  - For a selected 10 minute interval either the current 10 minute interval or intervals in the past
    - What message ids are unusual?
    - Are messages issued in context within expected messages pattern?
    - Is a component emitting unusual message ids?
    - How often did the message id occur?
    - Within the 10 minute interval when did the message id start occur?
  - **Did the z/OS LPAR produce the same messages for the corresponding interval in the past?**
    - Yesterday, last week, last month, …

- **After a change has been made**
  - Are new unusual messages being issued during periods immediately following changes like
    - New software levels (operating system, middleware , applications)?
    - Updated system settings / system configurations?
  - Are more messages issued than expected?

- **When looking for the cause of a random, intermittent problem**
  - Are new unusual messages being issued during periods before the problem is reported or during the periods when the problem is being reported?
  - Are more messages issued then expected?
  - Are messages issued out of context ?



A closer look inside IBM zAware



Note: z/OS 1.13 plus PTFs or higher for monitored client

IBM zAware itself uses an LPAR. This will reduce the number of LPARs available for customer use

Thanks for listening…

Questions?