System z EC12 (deep dive)
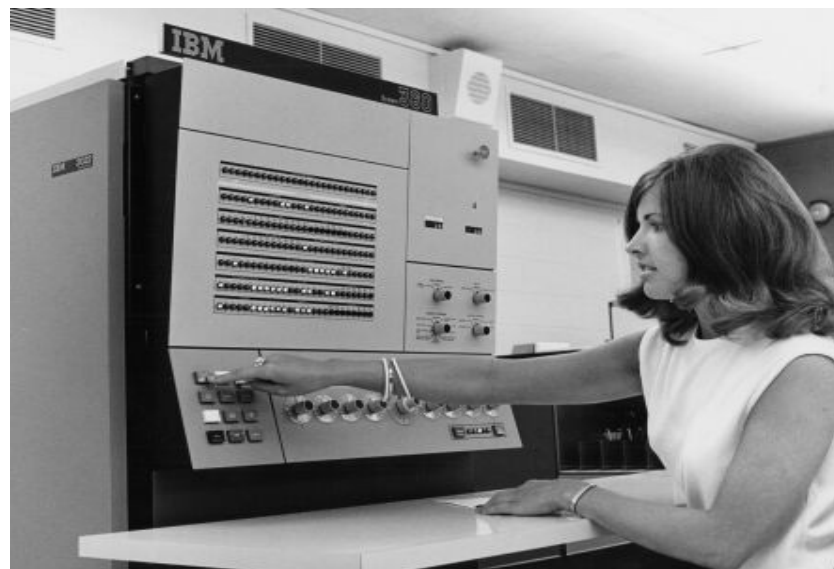
Carsten Otte
System z Firmware Developer
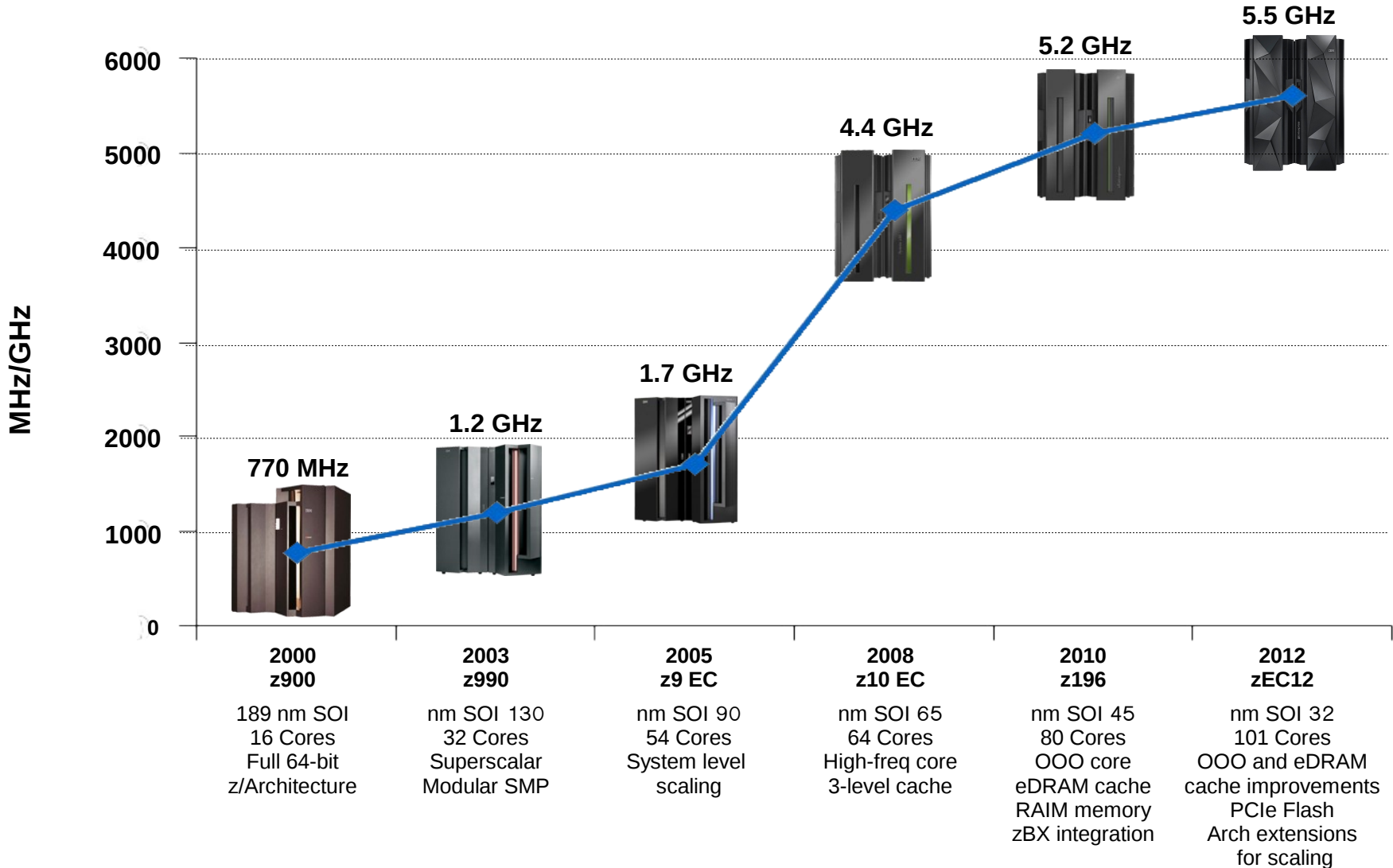
What is System z? – A Server for large Enterprises!

- **A System z server is what businesses use to host the largest commercial databases, transaction servers, and applications that require a greater degree of security and availability than is commonly found on smaller-scale machines.**
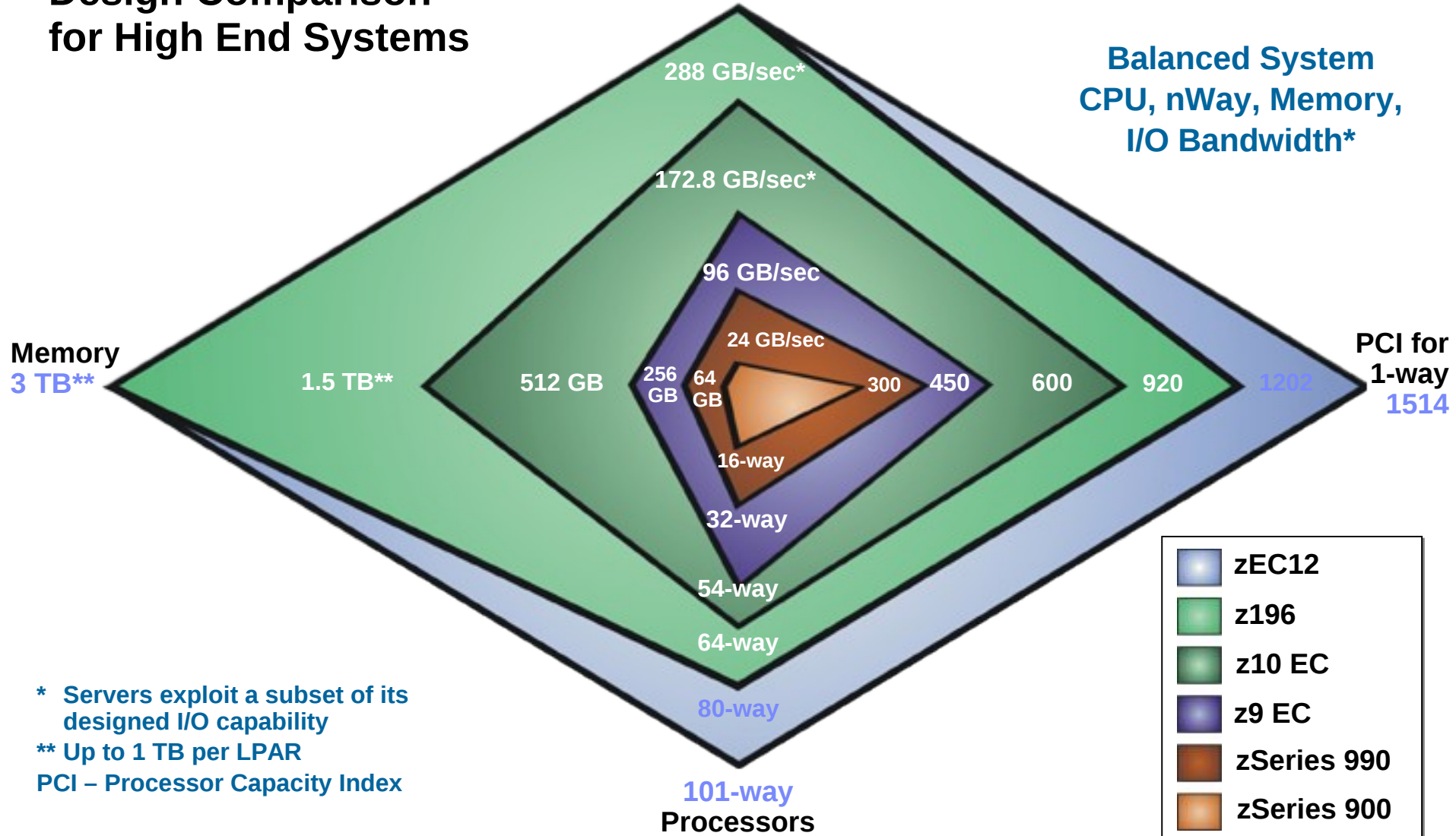
zEC12 Continues the CMOS Mainframe Heritage Begun in 1994

MHz/GHz

6000
5000
4000
3000
2000
1000
0

770 MHz
1.2 GHz
1.7 GHz
4.4 GHz
5.2 GHz
5.5 GHz

2000 z900	2003 z990	2005 z9 EC	2008 z10 EC	2010 z196	2012 zEC12
189 nm SOI 16 Cores Full 64-bit z/Architecture	nm SOI 130 32 Cores Superscalar Modular SMP	nm SOI 90 54 Cores System level scaling	nm SOI 65 64 Cores High-freq core 3-level cache	nm SOI 45 80 Cores OOO core eDRAM cache RAIM memory zBX integration	nm SOI 32 101 Cores OOO and eDRAM cache improvements PCIe Flash Arch extensions for scaling

3

IBM System z:
Design Comparison for High End Systems

System I/O Bandwidth
384 GB/Sec*

288 GB/sec*

172.8 GB/sec*

96 GB/sec

24 GB/sec

**Balanced System
CPU, nWay, Memory,
I/O Bandwidth***

**Memory
3 TB****

1.5 TB** | **512 GB** | **256 GB** | **64 GB** | **300** | **450** | **600** | **920** | **1202**

**PCI for
1-way
1514**

16-way

32-way

54-way

64-way

80-way

101-way
Processors

* **Servers exploit a subset of its designed I/O capability**
** **Up to 1 TB per LPAR**
PCI – Processor Capacity Index

Legend:
- zEC12
- z196
- z10 EC
- z9 EC
- zSeries 990
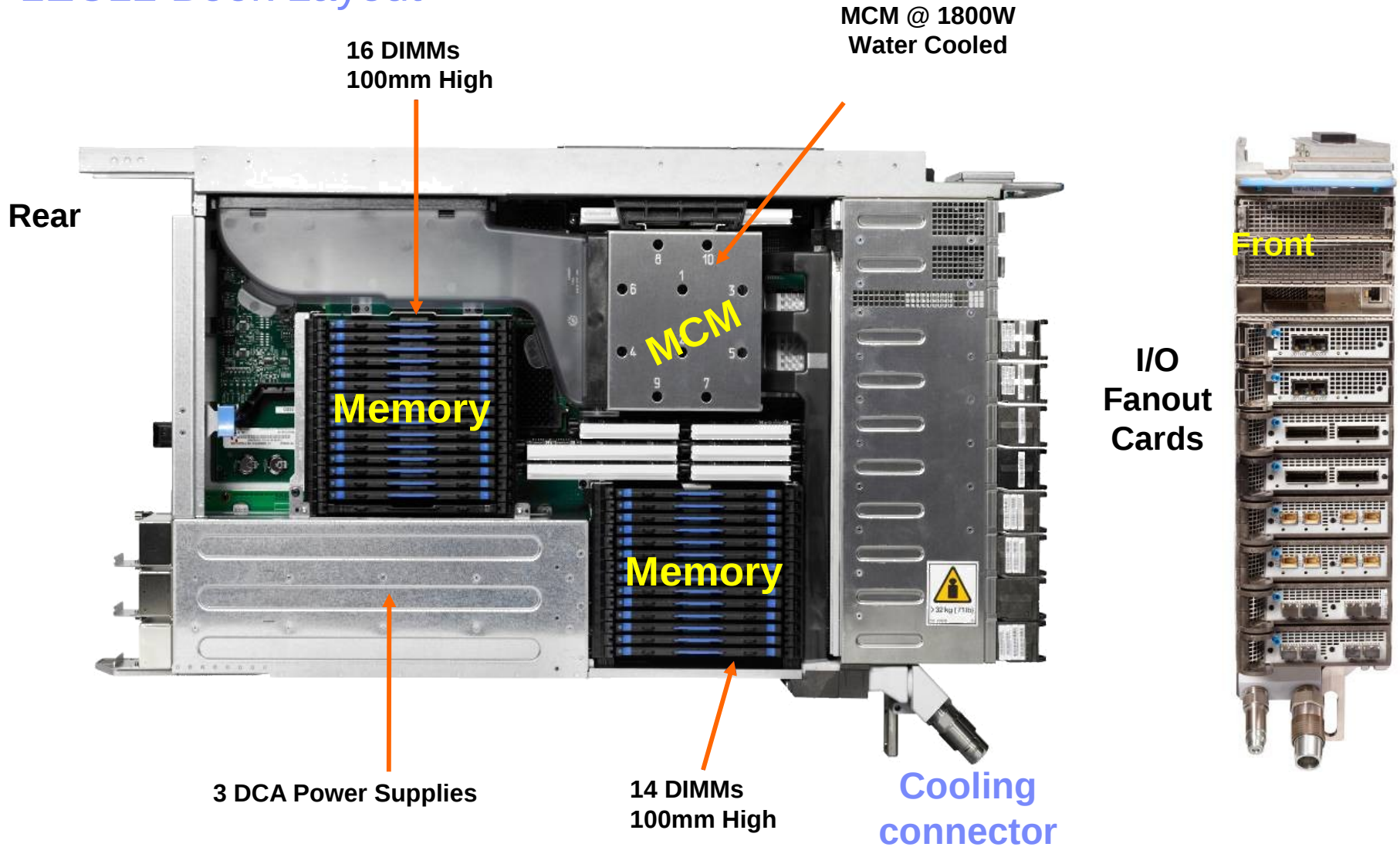- zSeries 900

4

zEC12 – Overall Attributes Highlights (compared to z196)

- **50% more cores in a CP chip**
  - **Up to 5.7% faster core running frequency**
  - **Up to 25% capacity improvement over z196 uni-processor**
- **Bigger caches and shorter latency**
  - **Total L2 per core is 33% bigger**
  - **Total on-chip shared L3 is 100% bigger**
  - **Unique private L2 designed to reduce L1 miss latency by up to 45%**
- **3rd Generation High Frequency, 2nd Generation Out of Order Design**
  - **Numerous pipeline improvements based on z10 and z196 designs**
  - **# of instructions in flight is increased by 25%**
- **New 2nd level Branch Prediction Table for enterprise scale program footprint**
  - **3.5x more branches**
- **Dedicated Co-processor per core with improved performance and additional capability**
  - **New hardware support for Unicode UTF8<>UTF16 bulk conversions**
- **Multiple innovative architectural extensions for software exploitation**

5

zEC12 New Build Radiator-based Air cooled – Under the covers (Model H89 and HA1) Front view

Overhead Power Cables (option)

Internal Batteries (option)

Power Supplies

2 x Support Elements

PCIe I/O drawers (Maximum 5 for zEC12)

Processor Books with Flexible Support Processors (FSPs), PCIe and HCA I/O fanouts

PCIe I/O interconnect cables and Ethernet cables FSP cage controller cards

Radiator with N+1 pumps, blowers and motors

Overhead I/O feature is a co-req for overhead power option

Optional FICON LX Fiber Quick Connect (FQC) not shown

6

© 2013 IBM Corporation

zEC12 Book Layout

16 DIMMs 100mm High

MCM @ 1800W Water Cooled

Rear

Front

MCM

Memory

I/O Fanout Cards

Memory

3 DCA Power Supplies

**14 DIMMs 100mm High**

Cooling connector

Note: Unlike the z196, zEC12 Books are the same for the Radiator based Air and Water cooled Systems

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

    - L2 cache/PU core

    - L3 cache shared by 6 PUs per chip

    - 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
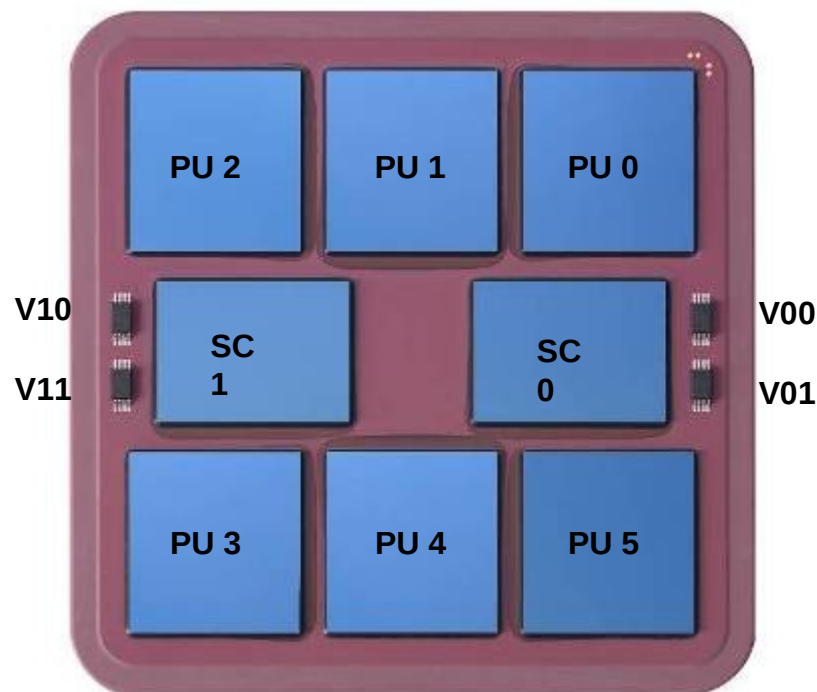
zEC12 Processor Design

- Built on solid foundation of z196
  - Leverage IBM 32nm SOI technology with eDRAM
- Enhanced high-frequency out-of-order core
  - Instruction pipeline streamlined for smoother flow
  - 2nd-level BTB expands branch prediction coverage
  - Faster engine for fixed-point division
  - Millicode performance improvements
- 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
- More processors in the same package as z196
  - 6 processor cores per CP chip
  - Crypto/compression co-processor per core
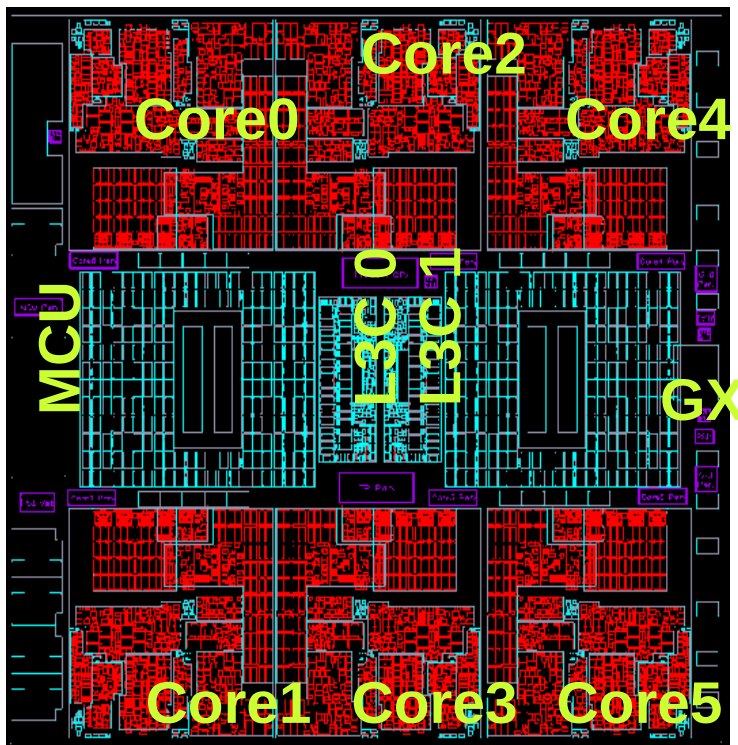 - Same power consumption as z196
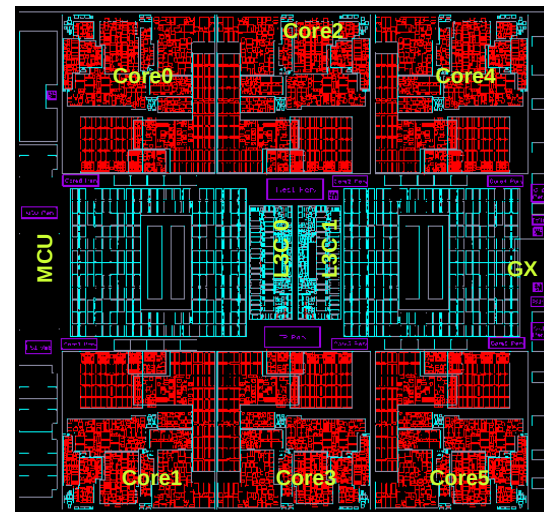
zEC12 PU Chip: 6 cores, 598 mm^2 chip
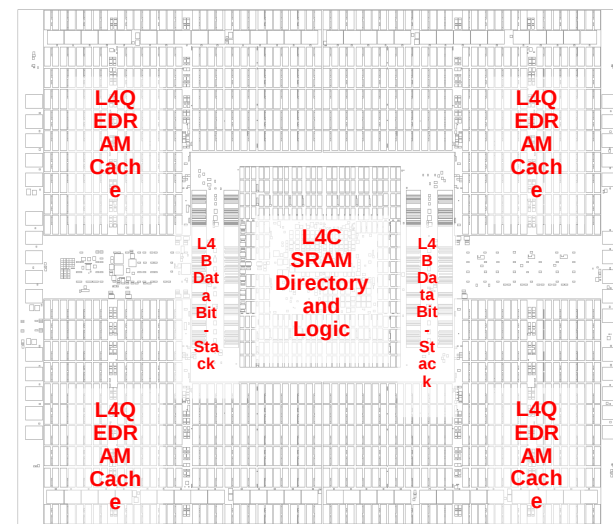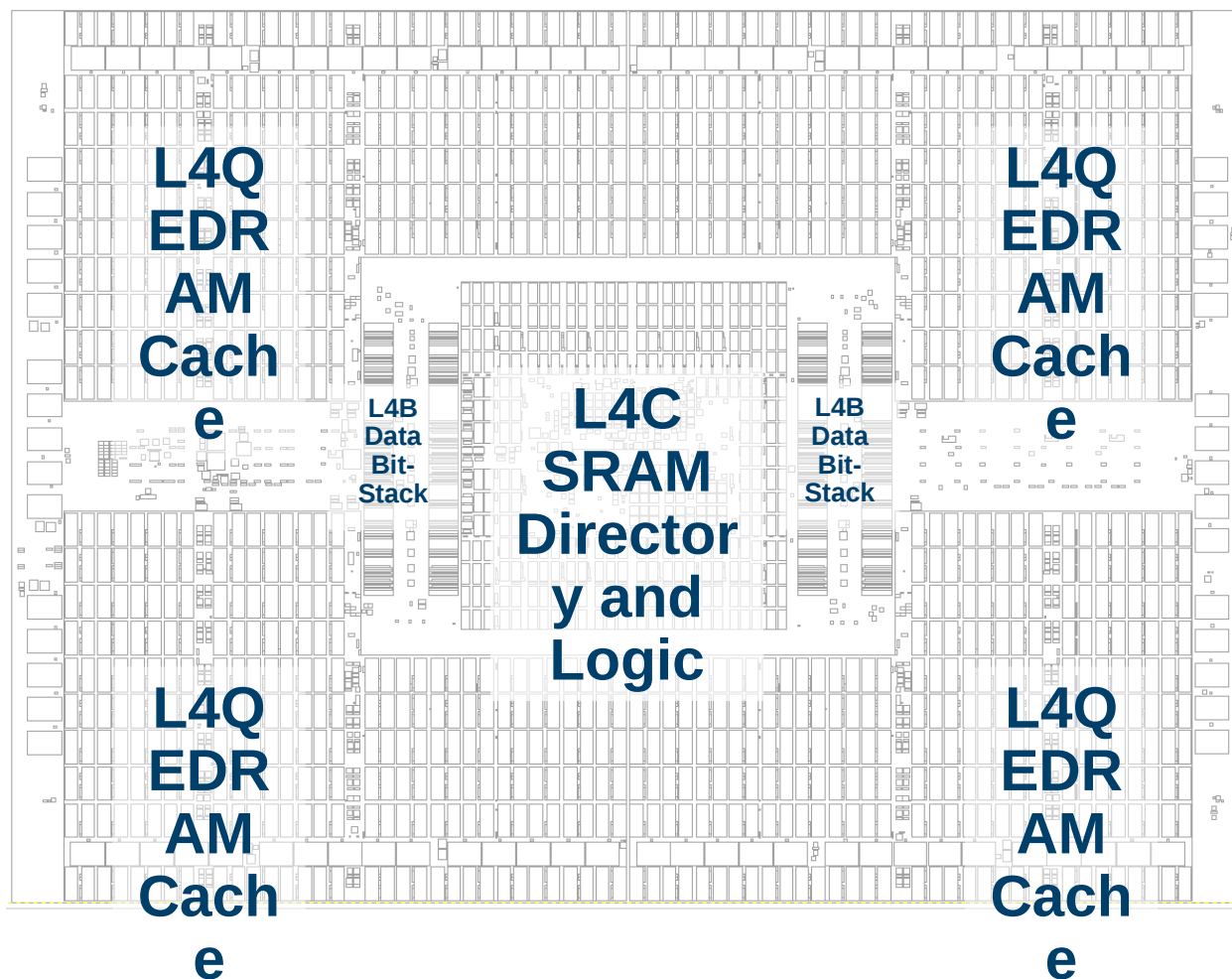

zEC12 SC Chip: 192MB cache, 526 mm^2 hip

9

zEC12 Hexa Core PU Chip Details

- **13S 32nm SOI Technology**
  - **15 layers of metal**
  - **7.68 km wire**
- **2.75 Billion Transistors**

- **Chip Area**
  - **597 mm^2**
  - **23.7mm x 25.2mm**
  - **10000+ Power pins**
  - **1071 signal I/Os**

- **Up to Six active cores per chip**
  - **5.5 GHz**
  - **L1 cache/ core**
    - **64 KB I-cache**
    - **96 KB D-cache**
  - **L2 cache/ core**
    - **1M+1M Byte hybrid split private L2 cache**
- **Dedicated Co-processors (COP) per core**
  - **Crypto & compression accelerators**
  - **Includes 16KB cache**
- **On chip 48 MB eDRAM L3 Cache**
  - **Shared by all six cores**
- **Interface to SC chip / L4 cache**
  - **44 GB/sec to each of 2 SCs (5.5 GHz)**
- **I/O Bus Controller (GX)**
  - **Interface to Host Channel Adapter (HCA)**
- **Memory Controller (MC)**
  - **Interface to controller on memory DIMMs**
  - **Supports RAIM design**

10

zEC12 SC Chip Detail

- 13S 32nm SOI Technology
  - 15 layers of metal
- Chip Area –
  - 526mm^2
  - 26.72mm x 19.67mm
  - 7311 Power C4's
  - 1819 signal C4's
- 3.3 Billion Transistors
  - 2.1 Billion eDRAM transistors
- eDRAM Shared L4 Cache
  - 192 MB per SC chip
  - 384 MB per Book
- 6 CP chip interfaces
- 3 Fabric interfaces
- 1 clock domain
- 4 unique chip voltage supplies

11

IBM zEnterprise EC12: An optimized system

Semiconductor Technology	Microprocessor Design	Systems Design	Virtualization & Operating Systems	Compilers & Java Virtual Machine	Optimized Middleware

zEC12 Architecture Extensions

- **Transactional Execution (a/k/a Transactional Memory)**
  - **Software-defined sequence treated by hardware as atomic "transaction"**
  - **Enables significantly more efficient software**
    - **Highly-parallelized applications**
    - **Speculative code generation**
    - **Lock elision**
  - **Designed for exploitation by Java; longer-term opportunity for DB2, z/OS, others**

- **Runtime instrumentation**
  - **Real-time information to software on dynamic program characteristics**
  - **Enables increased optimization in JVM/JIT recompilations**
  - **Additional exploitation opportunities in the works**

- **2 GB page frames**
  - **Increased efficiency for DB2 buffer pools, Java heap, other large structures**

- **Software directives to improve hardware performance**
  - **Data usage intent improves cache management**
  - **Branch pre-load improves branch prediction effectiveness**
  - **Block prefetch moves data closer to processor earlier, reducing access latency**

- **Decimal format conversions**
  - **Enable broader exploitation of Decimal Floating Point facility by COBOL programs**

zEC12 Processor Unit allocation/usage

Model	Books/ PUs	CPs	IFLs uIFLs	zAAPs	zIIPs	ICFs	Std SAPs	Optional SAPs	Std. Spares	Rsvd. PUs
H20	1/27	0-20	0-20 0-19	0-10	0-10	0-20	4	0-4	2	1
H43	2/54	0-43	0-43 0-42	0-21	0-21	0-43	8	0-8	2	1
H66	3/81	0-66	0-66 0-65	0-33	0-33	0-66	12	0-12	2	1
H89	4/108	0-89	0-89 0-88	0-44	0-44	0-89	16	0-16	2	1
HA1	4/120	0-101	0-101 0-100	0-50	0-50	0-101	16	0-16	2	1

- **zEC12 Models H20 to H89 use books with 27 core MCMs. The Model HA1 has 4 books with 30 core MCMs**
  - **Each MCM uses PU chips with a combination of 4, 5 and 6 active cores**
- **The maximum number of logical ICFs or logical CPs supported in a CF LPAR is 16**
- **The Reserved PU is not available for customer purchase**
- **Concurrent Book Add is available to upgrade from model H20 to model H89**

Notes: 1. At least one CP, IFL, or ICF must be purchased in every machine
2. One zAAP and one zIIP may be purchased for each CP purchased even if CP capacity is "banked".
3. "uIFL" stands for Unassigned IFL

IBM zEnterprise EC12

IBM zEnterprise EC12 (2827)

- Announced 08/12 – Server w/ up to 101 PU cores
- 5 models – Up to 101-way
- Granular Offerings for up to 20 CPs
- PU (Engine) Characterization
  - CP, SAP, IFL, ICF, zAAP, zIIP
- On Demand Capabilities
  - CoD, CIU, CBU, On/Off CoD, CPE, FoD
- Memory – up to 3 TB for Server and up to 1 TB per LPAR
  - 32 GB Fixed HSA
- Channels
  - PCIe bus
  - Four LCSSs
  - 3 Subchannel Sets
  - FICON Express8 and 8S
  - zHPF
  - OSA 10 GbE, GbE, 1000BASE-T
  - InfiniBand Coupling Links
  - Flash Express
- Configurable Crypto Express4S
- Parallel Sysplex clustering
- HiperSockets – up to 32
- Up to 60 logical partitions
- Enhanced Availability
- IBM zAware
- Unified Resource Manager
- Operating Systems
  - z/OS, z/VM, z/VSE, z/TPF, Linux on System z

IBM zEnterprise Blade Extension (2458)

- First Announced 7/10
- Model 003 for zEC12 – 08/12
- zBX Racks with:
  - BladeCenter Chassis
  - N + 1 components
  - Blades
  - Top of Rack Switches
  - 8 Gb FC Switches
  - Power Units
  - Advance Management Modules

- Up to 112 Blades
  - POWER7 Blades
  - IBM System x Blades
  - IBM WebSphere DataPower Integration Appliance XI50 for zEnterprise (M/T 2462-4BX)
- Operating Systems
  - AIX 5.3 and higher
  - Linux for Select IBM x Blades
  - Microsoft Windows for x Blades
- Hypervisors
  - PowerVM Enterprise Edition
  - Integrated Hypervisor for System x

Introducing the newest members of the zEnterprise System family
The zEnterprise EC12 and zEnterprise BladeCenter Extension Model 003

IBM zEnterprise EC12 (zEC12)

- zEC12 has the industry's fastest superscalar chip with each core at 5.5 GHz
- New innovation to drive availability with IBM zAware and Flash Express
- Optimized for the corporate data serving environment
- Hardware functions boost software performance for Java™, PL/I, DB2®

IBM zEnterprise Unified Resource Manager and zEnterprise BladeCenter® Extension (zBX) Mod 003

- Supports the new zEC12 platform
- Hosts PS701 and HX5 blades
- Provides workload-awareness resource optimization
- Enhancements to System Director support zBX
- System z will continue to expand hybrid computing

Plus more flexibility and function by connecting to IDAA

- IBM DB2 Analytics Accelerator (IDAA) allows deployment of business analytics on the same platform as operational applications
- Analytics and OLTP can be run as the same workload

16

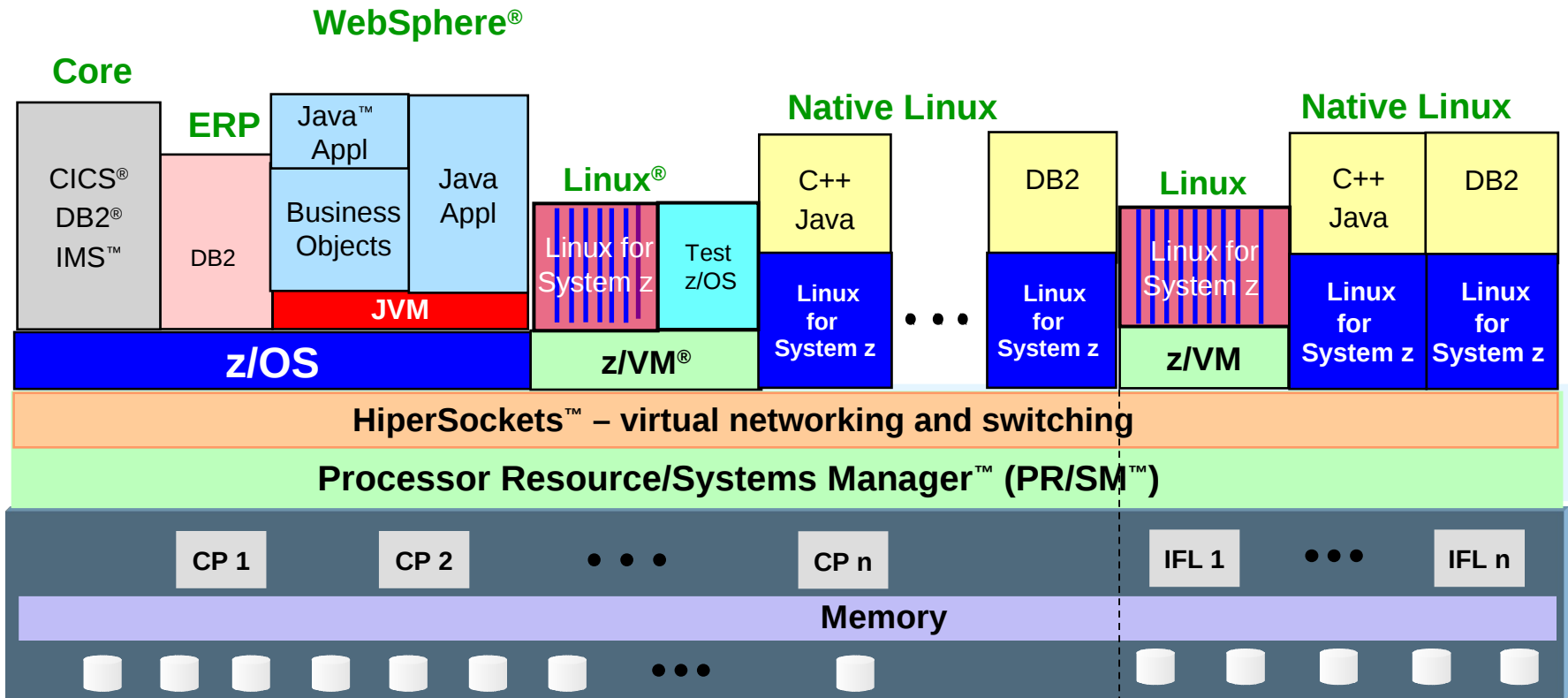# Unified Resource Manager optimizes system resources

- Bringing mainframe governance for System z resources

- Single view of virtualized resources across platforms

- Integrated network for better security, control and faster time to value

- Management of resources as defined by your business goals and objectives
  - IBM intends to deliver workload-aware optimization for System x blades (Statement of Direction)*

- Programmable interfaces (APIs) to connect with system management tools for total management capabilities

- New dynamic storage capabilities for System x blades

> ***Unified Resource Manager** transforms the way resources are managed and deployed*

* All statements regarding IBM future direction and intent are subject to change or withdrawal without notice and represents goals and objectives only.

System z – The Ultimate Virtualization Resource

WebSphere®

Core

| CICS® DB2® IMS™ | ERP DB2 | Java™ Appl / Business Objects | Java Appl | Linux® Linux for System z | Test z/OS | Native Linux C++ Java | DB2 | Linux Linux for System z | Native Linux C++ Java | DB2 |

JVM

| z/OS | z/VM® | Linux for System z | • • • | Linux for System z | z/VM | Linux for System z | Linux for System z |

HiperSockets™ – virtual networking and switching

Processor Resource/Systems Manager™ (PR/SM™)

| CP 1 | CP 2 | • • • | CP n | | IFL 1 | • • • | IFL n |

Memory

- Massive, robust consolidation platform; virtualization is built in, not added on
- Up to 60 logical partitions on PR/SM; 100's to 1000's of virtual servers on z/VM
- Virtual networking for memory-speed communication, as well as virtual layer 2 and layer 3 networks supported by z/VM
- Most sophisticated and complete hypervisor function available
- Intelligent and autonomic management of diverse workloads and system resources based on business policies and workload performance objectives

Flash Express – What is it?

FLASH Express

- **Physically comprised of internal storage on Flash SSDs**
- **Used to deliver a new tier of memory - storage class memory (SC**
- **Supported on z/OS V1.13 plus web deliverable (PTF)**
- **Uses PCIe I/O drawer**

- **Sized to accommodate *all LPAR paging***
  - **Each <u>card pair</u> provides 1.6 TB usable storage (3.2 TB total)**
  - **Maximum 4 card pairs   (4 X1.6 = 6.4 TB)**

- **Immediately usable**
  - **No capacity planning needed**
  - **No intelligent data placement needed**

- **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.**
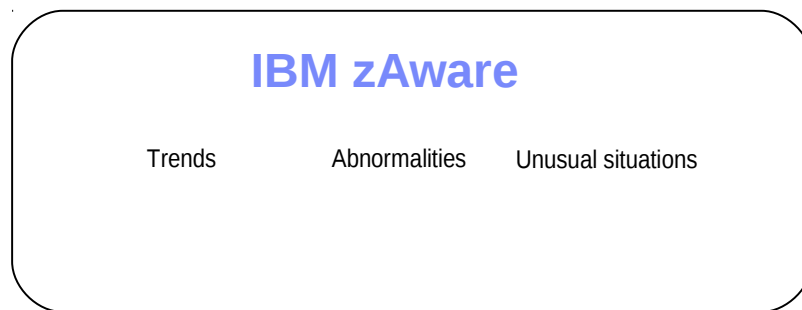 - **Key Management provided based on a Smart Card**

One Flash Express Card

CPU nanos

Flash - micros

Disk - ms
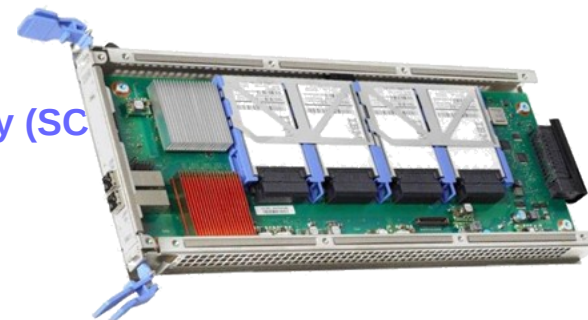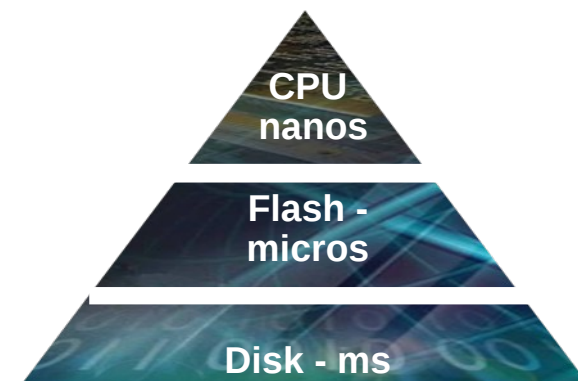
IBM zAware delivers smarter message monitoring capabilities

- **Out-of-band high speed analytics application delivered as an integrated firmware stack**
- **Analyzes system messages to provide a near real-time view of your system**
- **Uses self learning to recognize message patterns of your environment**
- **User-friendly web interface**
- **Easy drill down techniques help identify problematic messages and unusual patterns to help speed up diagnostic time**
- **Provides information, in XML format, that can feed other processes or tools**
  - **Tivoli Integrated Service Management intends to work with IBM zAware to provide alert and event notifications***

IBM zAware

Trends   Abnormalities   Unusual situations

zEC12 IBM zAware host

z/OS   z/OS

IBM zAware Host Partition

IBM zAware GUI

IBM zAware monitored client

z/OS   z/OS   z/OS

***All statements regarding IBM future direction and intent are subject to change or withdrawal without notice and represents goals and objectives only.**

Synergy with zEC12 Operating Systems

z/OS

- **Java exploitation of *Transactional Execution* for *increased parallelism and scalability***
- **Enhanced security support for *digital signatures***
- **Faster problem determination with *IBM zAware for improved availability***
- **Improve availability and performance with *Flash Express***
- ***2 GB page* support**
- **Simpler Specialty Engine (*zIIP*) exploitation**
- **z/OS v1.13 exploitation of new hardware**
- **z/OS health checks for SAN for new channel path selection**
- **Plus *over 4,100* applications enabled on z/OS**

Linux on System z

- ***Improved consolidation ratio* through new capacity performance**
- **Improved I/O performance using *High Performance FICON (zHPF)***
- ***Application and Linux optimization* enabled by full exploitation of zArchitecture extensions**
- **Optimized system setup via *Linux health checker***
- ***FCP end-to-end data integrity checking* for applications and storage subsystems**
- **Plus *over 3,000* applications on System z**

z/TPF

- **Support for *86 CPUs***
- ***Hardware exploitation* for performance improvements**

z/VM

- *z/VM Compatibility support*
- **Guest exploitation *support for new OSA and encryption technology***
- ***Simplified data exchange* of virtual Linux servers using z/VM software**
- **Improved I/O performance using *High Performance FICON (zHPF)* for guest exploitation**

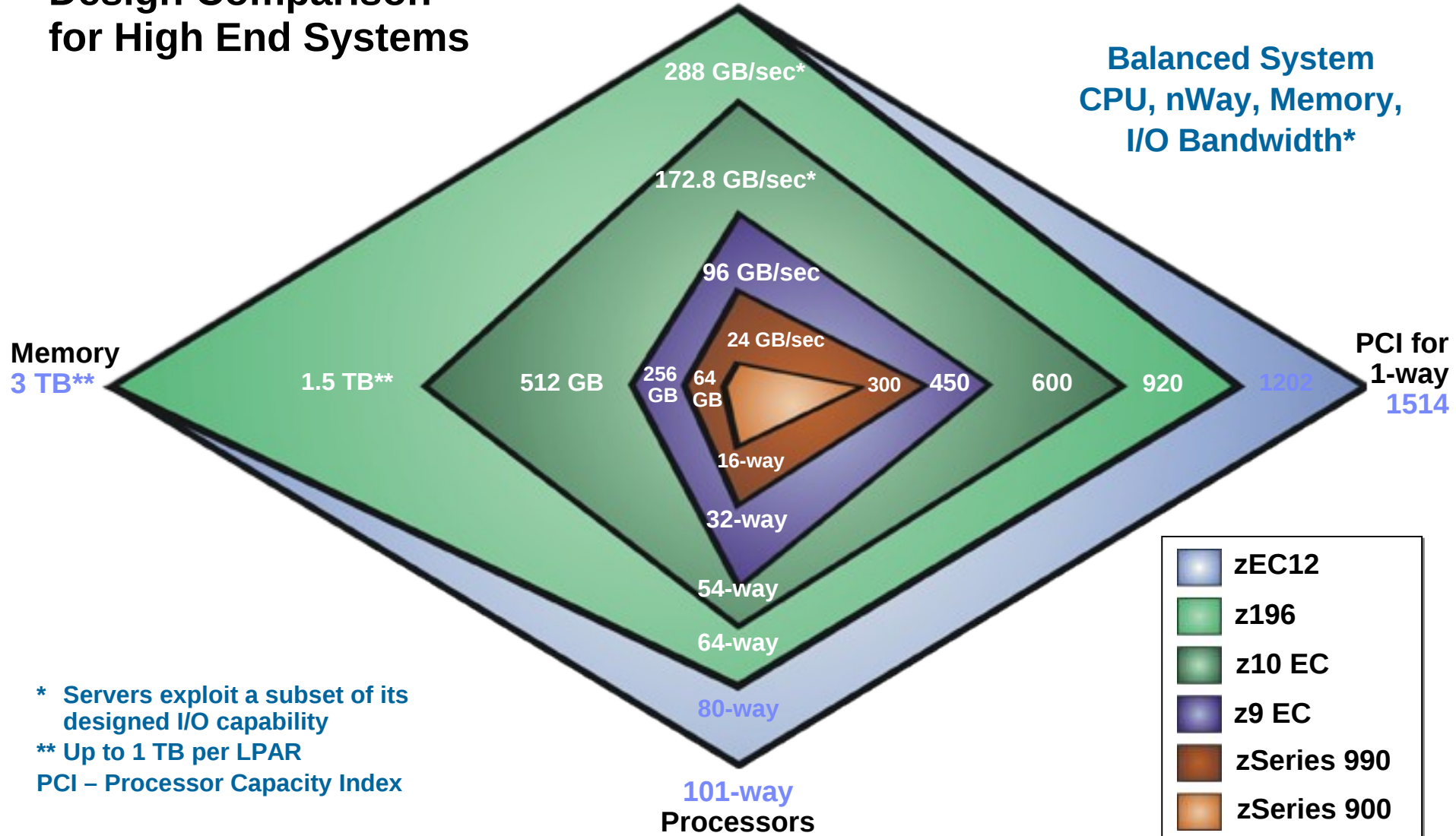## z/VSE

- ***64-bit addressing* with z/VSE V5.1**
- **Strong *interoperability with Linux* on System z**
- **New CICS functionality (*CICS Explorer*)**

- ***AND* with blades on the zBX there are even more options with applications on *AIX, Linux on System x or Microsoft Windows***

IBM System z:
Design Comparison
for High End Systems

System I/O Bandwidth
384 GB/Sec*

Balanced System
CPU, nWay, Memory,
I/O Bandwidth*

288 GB/sec*

172.8 GB/sec*

96 GB/sec

24 GB/sec

Memory
3 TB**

PCI for
1-way
1514

1.5 TB**   512 GB   256 GB   64 GB   300   450   600   920   1202

16-way

32-way

54-way

64-way

80-way

* Servers exploit a subset of its
  designed I/O capability
** Up to 1 TB per LPAR
PCI – Processor Capacity Index

101-way
Processors

zEC12
z196
z10 EC
z9 EC
zSeries 990
zSeries 900

22

© 2012 IBM Corporation

IBM