IBM's System z Forum

Make the most of your mainframe.

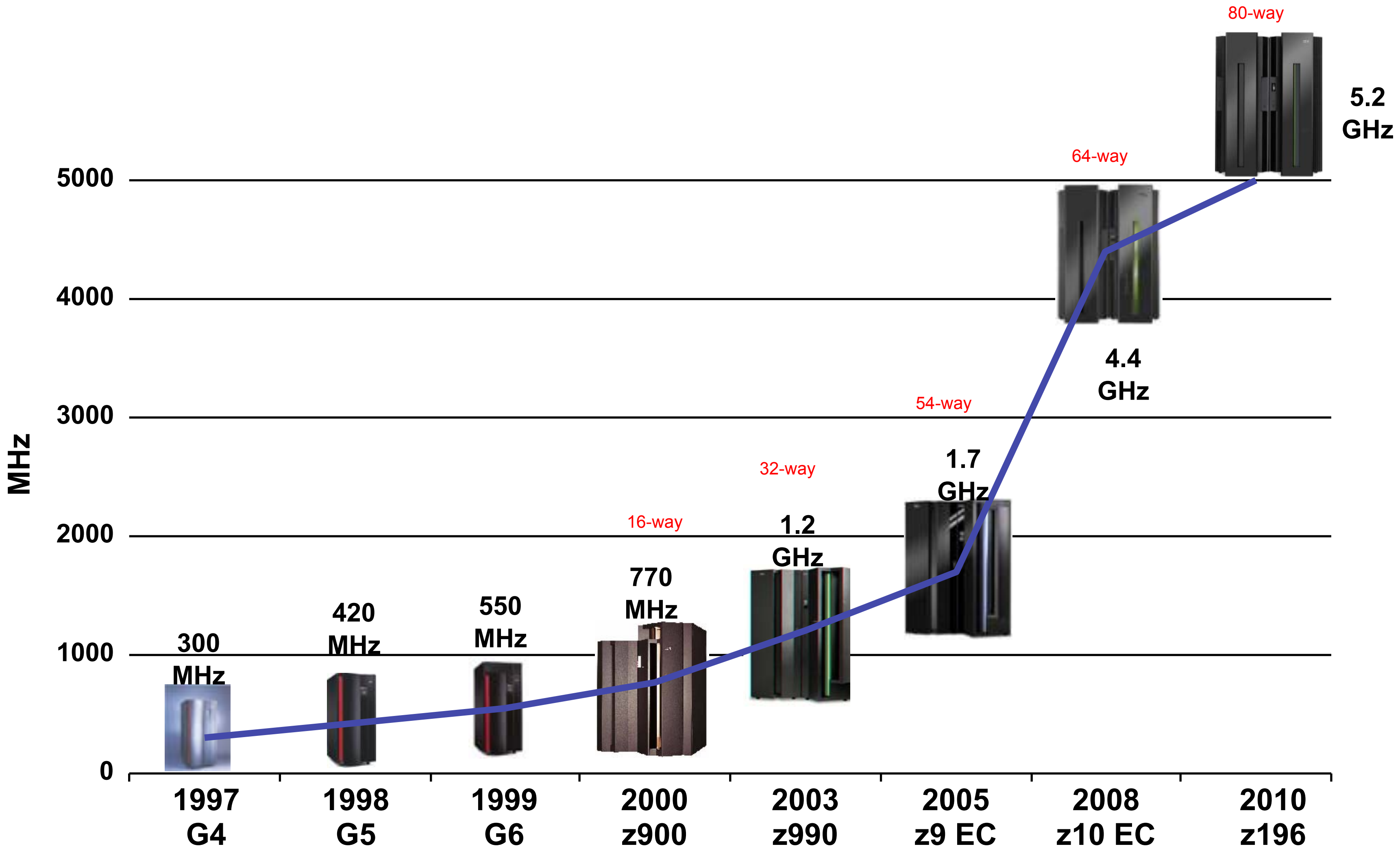
IBM System z: A Peek Under the Bonnet

Steve Talbot-Walsh
Senior System zSW Client Architect (zCA), zEvangelist and zChampion

Topics

- Review of recent System z mainframes
- Processor Hardware Overview
- New Instruction Set Architecture for zEnterprise
- System z Compilers and JAVA
- Out of Order (OOO) Execution
- Redundant Array of Independent Memory (RAIM)
- Security
- zBX: a System of Systems
- Energy Efficiency

IBM z196 Continues the CMOS Mainframe Heritage

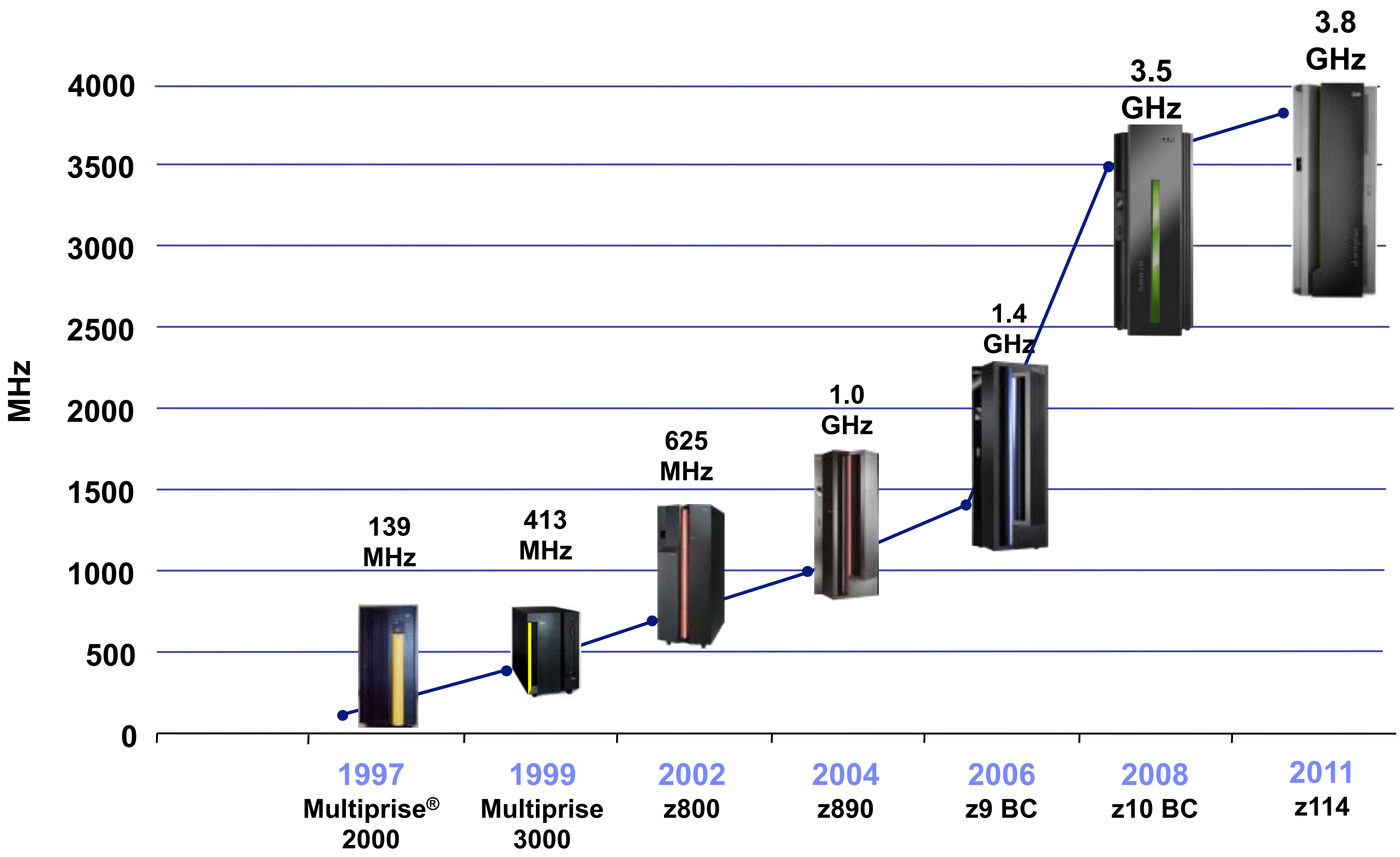


- **G4** – 1st full-custom CMOS S/390®
- **G5** – IEEE-standard BFP; branch target prediction
- **G6** – Copper Technology (Cu BEOL)

- **z900** – Full 64-bit z/Architecture®
- **z990** – Superscalar CISC pipeline
- **z9 EC** – System level scaling

- **z10 EC** – Architectural extensions
- **z196** – Out of order, improved superscalar, new architecture

IBM z114 Continues the CMOS Mainframe Heritage

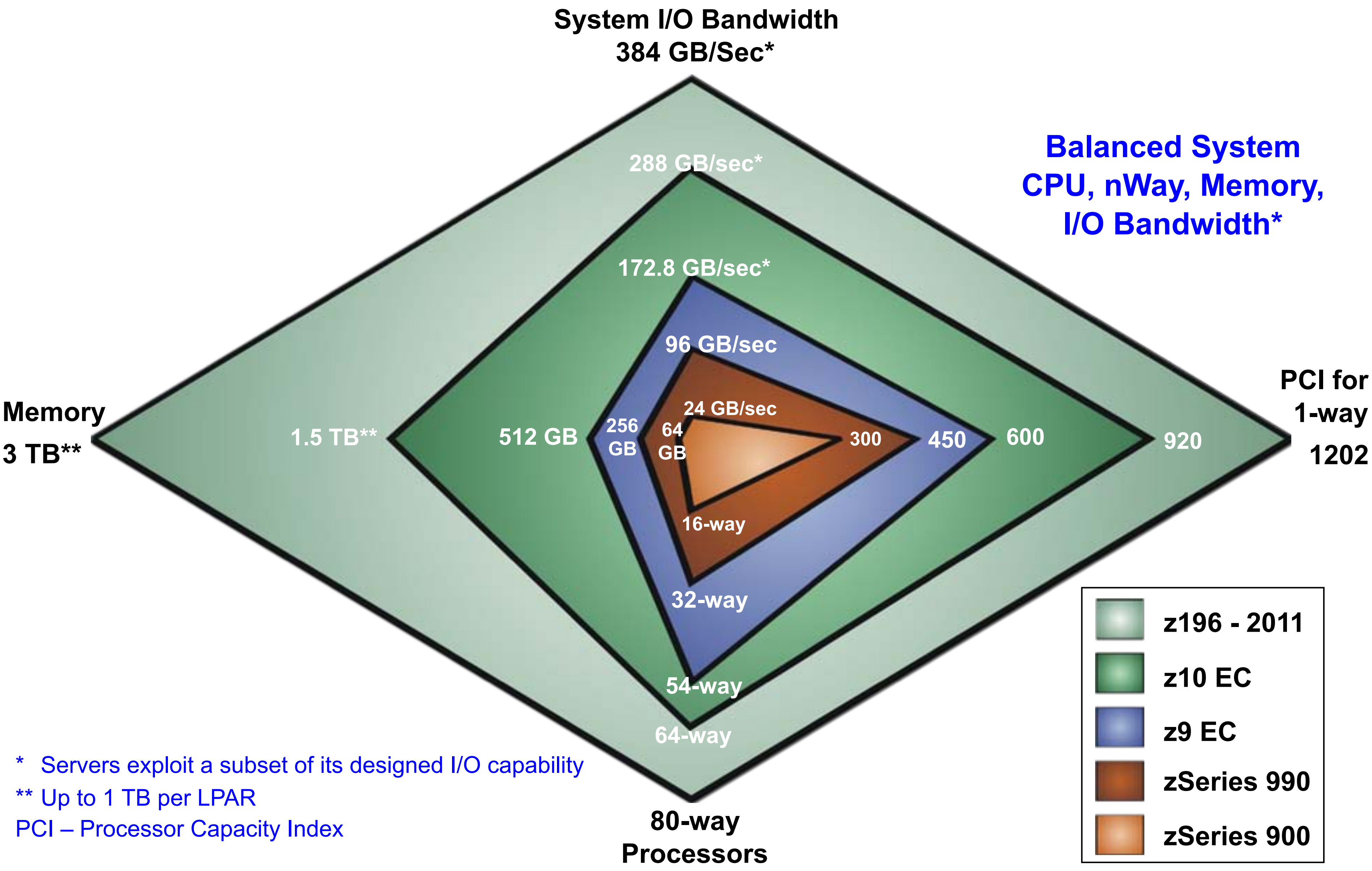


- Multiprise 2000 – 1st full-custom Mid-range CMOS S/390
- Multiprise 3000 – Internal disk, IFL introduced on midrange

- z800 - Full 64-bit z/Architecture®
- z890 - Superscalar CISC pipeline
- z9 BC - System level scaling

- z10 BC - Architectural extensions
  - Higher frequency CPU
- z114 – Additional Architectural extensions and new cache structure

System z Design Comparison for High-End Systems



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

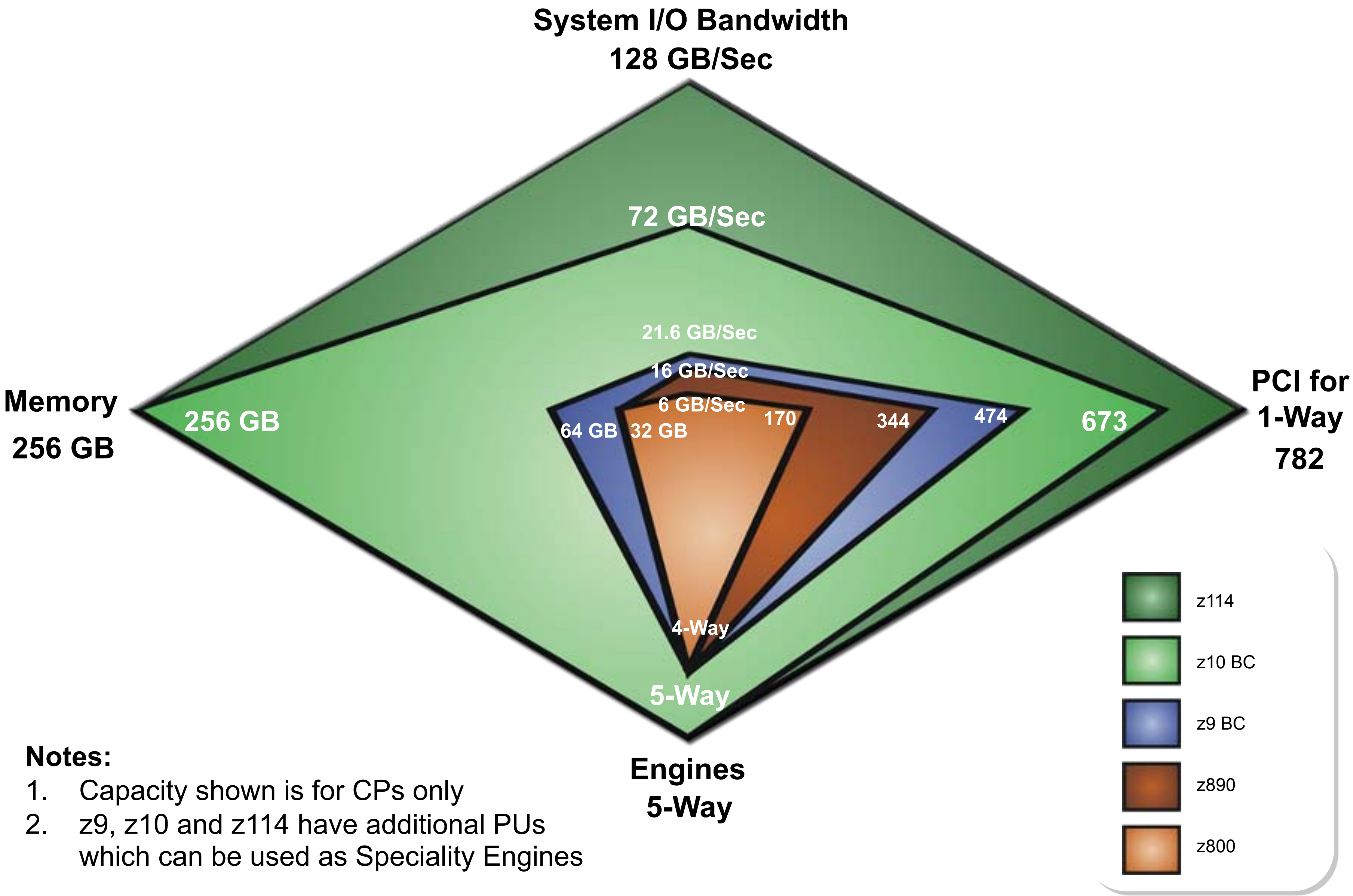
🟩	z196 - 2011
🟩	z10 EC
🟦	z9 EC
🟫	zSeries 990
🟧	zSeries 900

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

** Up to 1 TB per LPAR

PCI – Processor Capacity Index

System z Design Comparison for Entry-Level Systems

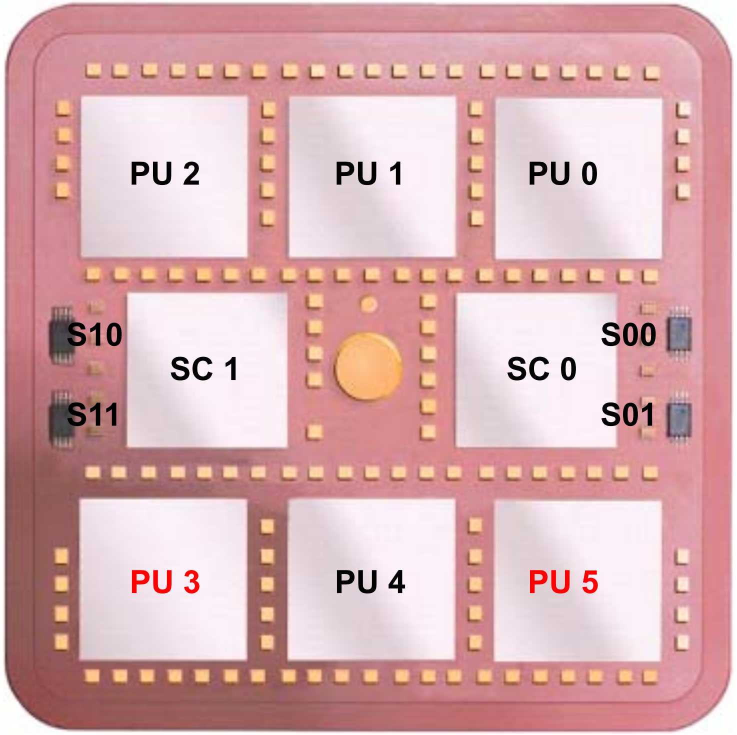


Notes:

1. Capacity shown is for CPs only
2. z9, z10 and z114 have additional PUs which can be used as Speciality Engines

z196 Multi-Chip Module (MCM) Packaging

- **96mm x 96mm MCM**
  - 103 Glass Ceramic layers
  - 8 chip sites
  - 7356 LGA connections
  - 20 and 24 way MCMs
  - Maximum power used by MCM is 1800W

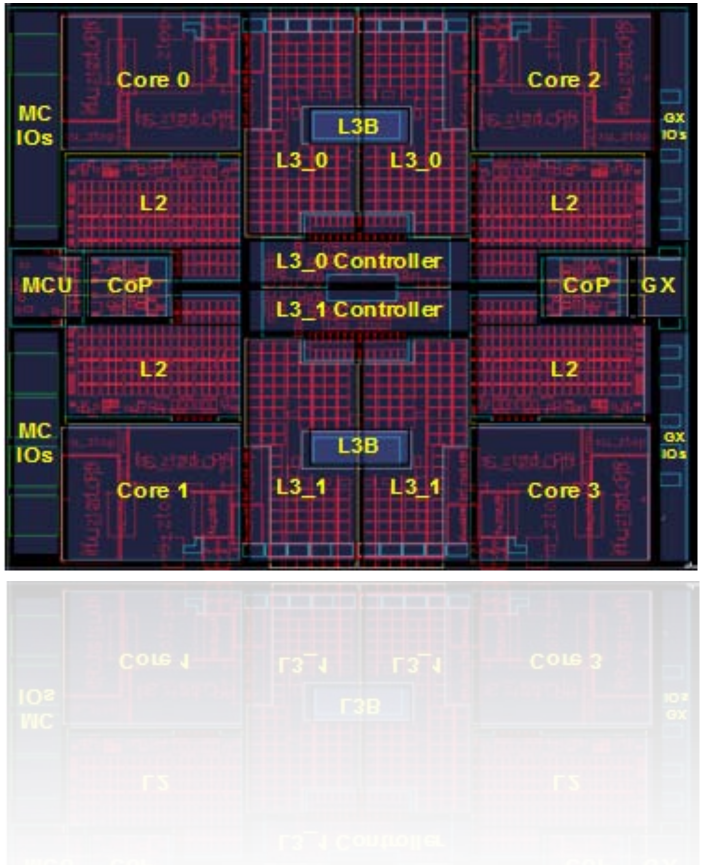


- **CMOS 12s chip Technology**
  - PU, SC, S chips, 45 nm
  - 6 PU chips/MCM – Each up to 4 cores
    - One memory control (MC) per PU chip
    - 23.498 mm x 21.797 mm
    - 1.4 billion transistors/PU chip
    - L1 cache/PU core
      - 64 KB I-cache
      - 128 KB D-cache
    - L2 cache/PU core
      - 1.5 MB
    - L3 cache shared by 4 PUs per chip

      24 MB
    - 5.2 GHz
  - 2 Storage Control (SC) chip
    - 24.427 mm x 19.604 mm
    - 1.5 billion transistors/SC chip
    - L4 Cache 96 MB per SC chip (192 MB/Book)
    - L4 access to/from other MCMs
  - 4 SEEPROM (S) chips
    - 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

z196 – IBM Leadership Technology At the Core

- New 5.2 GHz Quad Core Processor Chip boosts hardware price/performance
  - 100+ new instructions – improvements for CPU intensive, Java™, and C++ applications
  - Over twice as much on-chip cache as System z10 to help optimize data serving environment
  - Out-of-order execution sequence gives significant performance boost for compute intensive applications
  - Significant improvement for floating point workloads
- Performance improvement for systems with large number of cores – improves MP ratio
- Data compression and cryptographic processors right on the chip



z196 CPU Core

- Each core is a superscalar, out of order processor:
  - Cycle time is 5.2 GHz (which means a 0.19 ns cycle time)
  - Six RISC-like execution units
    - 2 fixed-point (integer), 2 load/store, 1 binary floating point, 1 decimal floating point
  - Up to three instructions decoded per cycle (vs. 2 in z10)
  - Up to five instructions/operations executed per cycle (vs. 2 in z10)
    - Execution can occur out of (program) order
    - Memory address generation and memory access can occur out of (program) order
    - Special circuitry to make execution and memory access appear to be in order to s/w
    - 211 complex instructions cracked into multiple internal operations
    - 246 of the most complex z/Architecture instructions are implemented via millicode
  - Each core has 3 private caches
    - 64KB 1st level cache for instructions, 128KB 1st level cache for data
    - 1.5MB L2 cache containing both instructions and data

- The same physical processor can be used for all the following CPU types:
  - Normal client CPU's (general processors)
  - Specialty Engines: zIIP (DB2), zAAP (JAVA), IFL (Linux)
  - Coupling Facilities
  - SAPs – I/O and service processors
  - Spare CPU's – used for Dynamic Processor Sparing in the event of a failing processor

Extensive Use of Hardware Speculation

- Based on the principles of speculative parallelism for MP systems – basically, doing work speculatively, the result of which may not be needed
- zArchitecture places many strict constraints on how the CPU has to appear to behave:
  - Example – Strict storage ordering rules (see zArchitecture POPS Chapter 5)
  - Good for software developers – significantly easier and more robust MP programming than other Instruction Set Architectures (ISA's)
  - Bad for the IBM CPU design team as its difficult to achieve good performance
- CPU has to make use of speculative processing techniques:
  - Assume things will go well, and have mechanisms to detect and back-off if they do not
  - In CPU design … "It's OK to cheat so long as you don't get caught".
  - Under the covers, the CPU violates the storage ordering rules in POPS, but has extensive/ complex logic to detect if software might observe it violating those rules. If it detects possible observation, it needs to redo the operation precisely following POPS rules.
  - Result is software only can observe the CPU following all the rules in POPS

zEnterprise New Instruction Set Architecture (ISA)

- Re-compiled code/apps get further performance gains through 100+ new instructions
- High-Word Facility (30 new instructions)
  - Independent addressing to high word of 64-bit GPRs
  - Effectively provides compiler/ software with 16 additional 32-bit registers (GPRCR)
- Interlocked-Access Facility (12 new instructions)
  - Interlocked (atomic) load, value update and store operation in a single instruction
  - Immediate exploitation by Java
- Load/Store-on-Condition Facility (6 new instructions)
  - Load or store conditionally executed based on condition code
  - Dramatic improvement in certain code with highly unpredictable branches
- Distinct-Operands Facility (22 new instructions)
  - Independent specification of result register (different than either source register)
  - Reduces register value copying
- Population-Count Facility (1 new instruction)
  - Hardware implementation of bit counting ~5x faster than prior software implementations
- Integer to/from Floating point convertions (39 new instructions)

System z Compilers and JAVA

- Compilers are the invisible bridge between user applications and the underlying systems and infrastructure that run your business
- Convert source code to machine executable instructions
- Impacts application performance, programmer productivity, and return on investment (ROI)
- Designed to unleash the full power of IBM System z processors
  - System z9, System z10, zEnterprise 114/196 & future z/Architectures
- Designed to support IBM Middleware
  - CICS, DB2, IMS
- Strong investment for strategic compilers on System z
  - Enterprise COBOL for z/OS, Enterprise PL/I for z/OS and z/OS XL C/C++
- Highly skilled development and research teams
  - 350 engineers in development, test and service roles
  - Close ties with IBM Research teams (Watson, Tokyo, China, and Haifa)
  - IBM Processor design teams
  - IBM Middleware development teams

z/OS XL C/C++ v1.12 New!

- Fully exploits the zEnterprise z114/z196 processor
  - Support for new Instructions
- Performs aggressive optimizations to C/C++ programs
  - Loop optimizations, whole program optimization, profile-directed feedback
  - Offers up to 60% performance improvement on zEnterprise 196 server over System z10
- Enables straightforward porting of C/C++ applications to z/OS
  - Supports industry C and C+ language standards and extensions
  - Added new C++0x language features
- Supports 31-bit and 64-bit application development
- “Metal C” option simplifies system programs development on System z
  - Supports CICS application development
  - Allows users to develop freestanding C programs
    - Obtain system services by calling assembler services directly
    - Works with HLASM
    - Supports MVS system linkage conventions

Enterprise PL/I for z/OS v4.1

New!

- Exploits zEnterprise 196 processor
  - Support new zEnterprise Instructions
  - Offers up to 27% performance improvement on zEnterprise 196 server over System z10
  - Leverage the same optimization technology as z/OS XL C/C++
- Supports integration of PL/I applications with web-based business processes
  - Introduced capability to validate an XML document against a schema while it is being processed by the PL/I application
  - Support offloading of XML parsing to zAAP specialty processors
- Improved support for Debug Tool
  - Option to reduce generated code size
  - Improved support for automonitor
- Improved support for SQL preprocessor and enforcement of coding rules

Enterprise COBOL for z/OS v4.2 (GA Sept 2009)

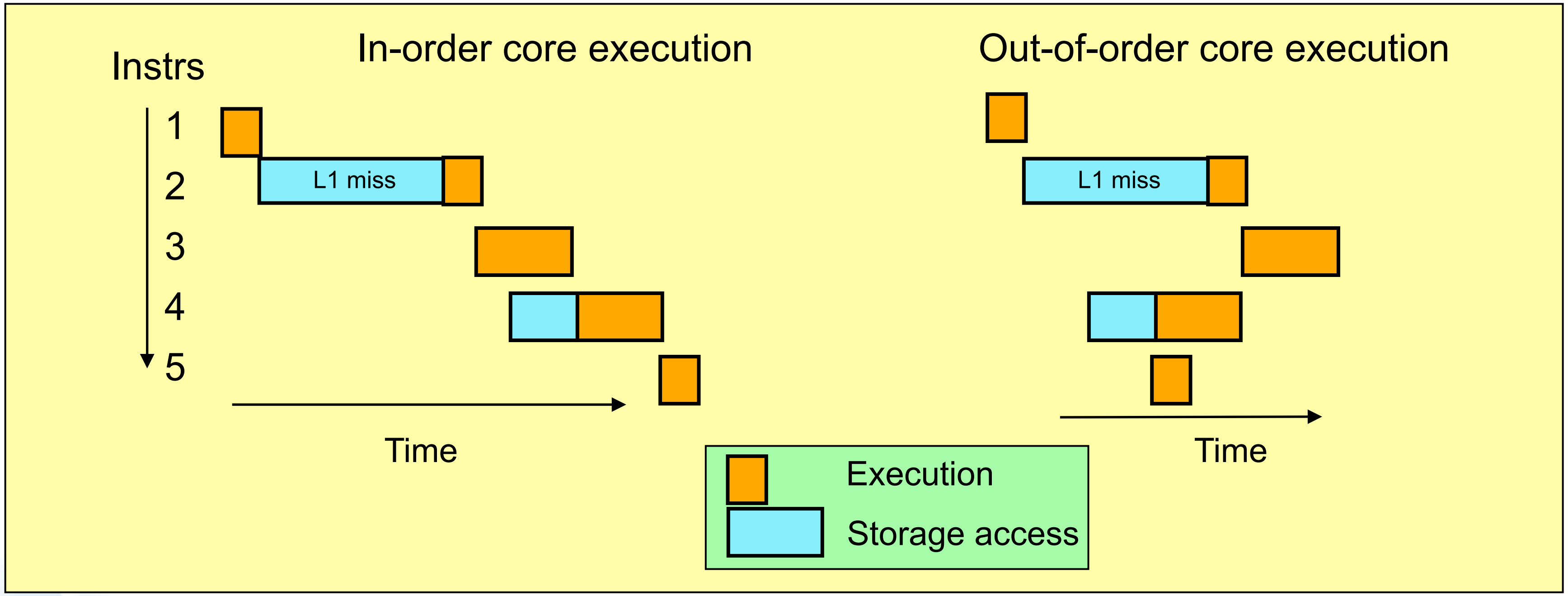
- Validated on zEnterprise 196 server with IBM's latest middleware.

JAVA

- z196 and JAVA v6.0.1 engineered together (GA Mar 2011)
- Utilizes the IBM J9 v2.6 VM
- Leveraging 70+ new zEnterprise instructions
- Performance improvements of circa 65% with Linux on System z
- Average of 2.1x improvement to multi-threaded workloads
- Average of 1.93x improvement to CPU intensive workloads
- H/W optimization technology for JAVA
  - Reducing pressure on instruction cache and data cache
  - New architectural facilities designed for scalability and concurrency
  - General optimizer and codegen improvements
- Tighter integration with System z facilities
- Current JAVA for z/OS releases (GA Aug 2011)
  - IBM 31-Bit SDK for z/OS, JAVA Technology Edition v7.0.0 (5655-W43)
  - IBM 64-Bit SDK for z/OS, JAVA Technology Edition v7.0.0 (5655-W44)

z196 Out of Order (OOO) Execution

- OOO yields significant performance benefit for compute intensive apps through
  - –Re-ordering instruction execution
    - Later (younger) instructions can execute ahead of an older stalled instruction
  - –Re-ordering storage accesses and parallel storage accesses
- OOO maintains good performance growth for traditional apps



z196 Out of Order Detail

- Out of order yields significant performance benefit through
  - Re-ordering instruction execution
    - Instructions stall in a pipeline because they are waiting for results from a previous instruction or the execution resource they require is busy
    - In an in-order core, this stalled instruction stalls all later instructions in the code stream
    - In an out-of-order core, later instructions are allowed to execute ahead of the stalled instruction
  - Re-ordering storage accesses
    - Instructions which access storage can stall because they are waiting on results needed to compute storage address
    - In an in-order core, later instructions are stalled
    - In an out-of-order core, later storage-accessing instructions which can compute their storage address are allowed to execute
  - Hiding storage access latency
    - Many instructions access data from storage
    - Storage accesses can miss the L1 and require 10 to 500 additional cycles to retrieve the storage data
    - In an in-order core, later instructions in the code stream are stalled
    - In an out-of-order core, later instructions which are not dependent on this storage data are allowed to execute

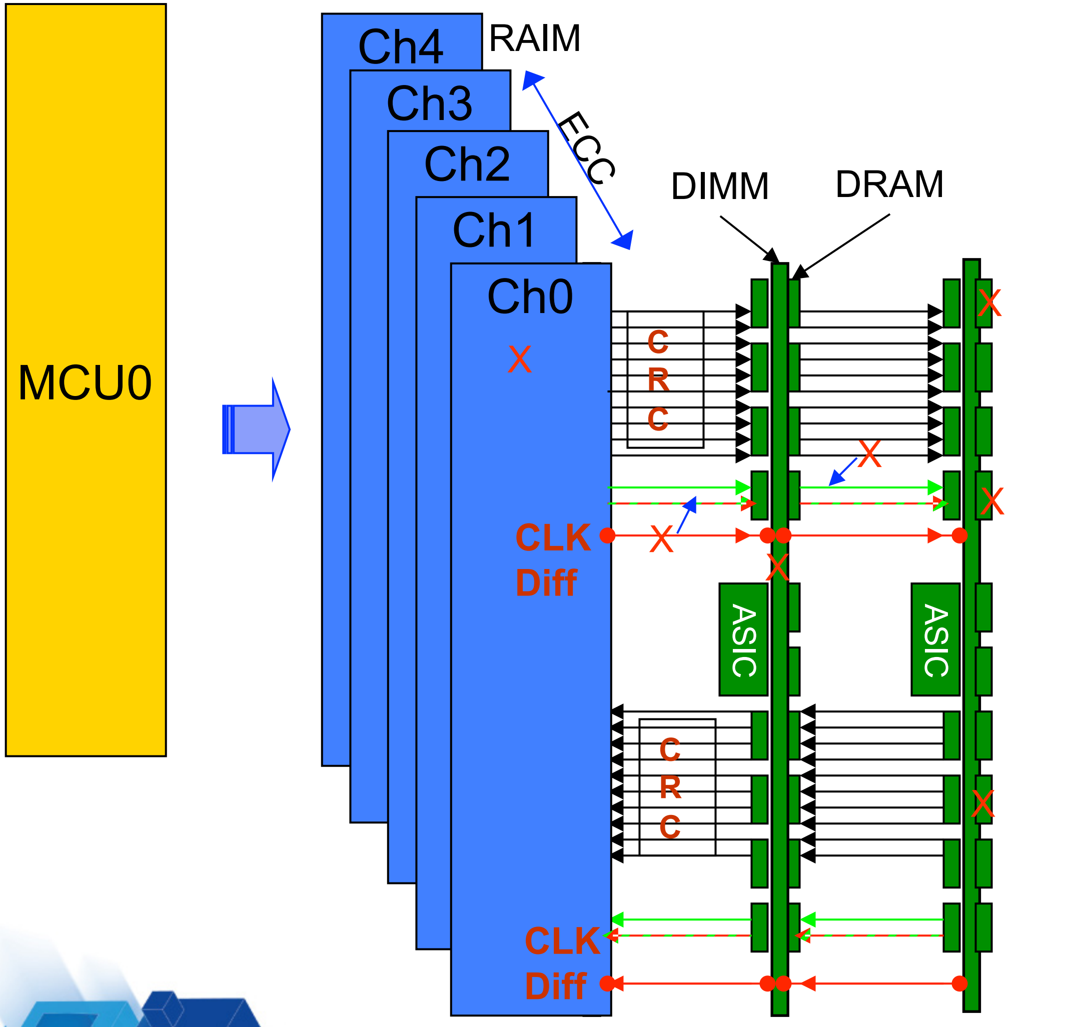
z196 Redundant Array of Independent Memory (RAIM)

- System z10 EC memory design:
  - Four Memory Controllers (MCUs) organized in two pairs, each MCU with four channels
  - DIMM technology is Nova x4, 16 to 48 DIMMs per book, plugged in groups of 8
  - 8 DIMMs (4 or 8 GB) per feature – 32 or 64 GB physical memory per feature
    Equals 32 or 64 GB for HSA and customer purchase per feature
  - 64 to 384 GB physical memory per book = 64 to 384 GB for use (HSA and customer)

- z196 memory design:
  - Three MCUs, each with five channels. The fifth channel in each z196 MCU is required to implement memory as a Redundant Array of Independent Memory (RAIM). This technology adds significant error detection and correction capabilities. Bit, lane, DRAM, DIMM, socket, and complete memory channel failures can be detected and corrected, including many types of multiple failures.
  - DIMM technology is SuperNova x81, 10 to 30 DIMMs per book, plugged in groups of 5
    5 DIMMs (4, 16 or 32 GB) per feature – 20, 80 or 160 GB physical RAIM per feature
    Equals 16, 64 or 128 GB for use per feature. RAIM takes 20%. (There is no non-RAIM option.)
  - 40 to 960 GB RAIM memory per book = 32 to 768 GB of memory for use
    (Minimum RAIM for the M15 is 60 GB = 48 GB = 16 GB HSA plus 32 GB customer memory)

- For both z196 and z10
  - The Hardware System Area (HSA) is 16 GB fixed, outside customer memory

z196 RAIM Memory Controller Overview



2- Deep Cascade
Using Quad High DIMMs

Layers of Memory Recovery

ECC
- Powerful 90B/64B Reed Solomon code

DRAM Failure
- Marking technology; no half sparing needed
- 2 DRAM can be marked
- Call for replacement on third DRAM

Lane Failure
- CRC with Retry
- Data – lane sparing
- CLK – RAIM with lane sparing

DIMM Failure (discrete components, VTT Reg.)
- CRC with Retry
- Data – lane sparing
- CLK – RAIM with lane sparing

DIMM Controller ASIC Failure
- RAIM Recovery

Channel Failure
- RAIM Recovery

Enhancing System z world-class security and resiliency

- Cryptographic enhancements on zEnterprise
  - Cryptography is in the “DNA” of System z hardware with Processor and Coprocessor based encryption capabilities
    - Processor Clear Key for bulk encryption – key material visible in storage
    - System z exclusive Protected Key CPACF helps to protect sensitive keys from inadvertent disclosure -- not visible to application or OS
  - Crypto Express3 enhanced to support key ANSI and ISO standards for the banking, finance and payment card industry.
  - NEW Enhanced display of cryptographic cards and simplified card configuration and management capabilities via the Trusted Key Entry workstation (TKE).
  - NEW Simplified master key management with ICSF enhancements providing a single point of administration within an z/OS Sysplex.
  - NEW Continued support for the next generation of public key technologies , ECC support is ideal for constrained environments such as mobile devices.
  - Crypto Express3 Coprocessor FIPS 140-2 Level 4 hardware evaluation.
- PR/SM™ designed and certified for EAL-5 certification
- RACF for z/OS v1.12 certified for EAL-5 (Feb 2012) NEW
- Policy driven flexibility to add capacity to real or virtual processors.
- High Availability, Backup and Disaster Recovery solutions
  - Leverage z114 and z196 as part of the new GDPS®/active-active continuous availability solution

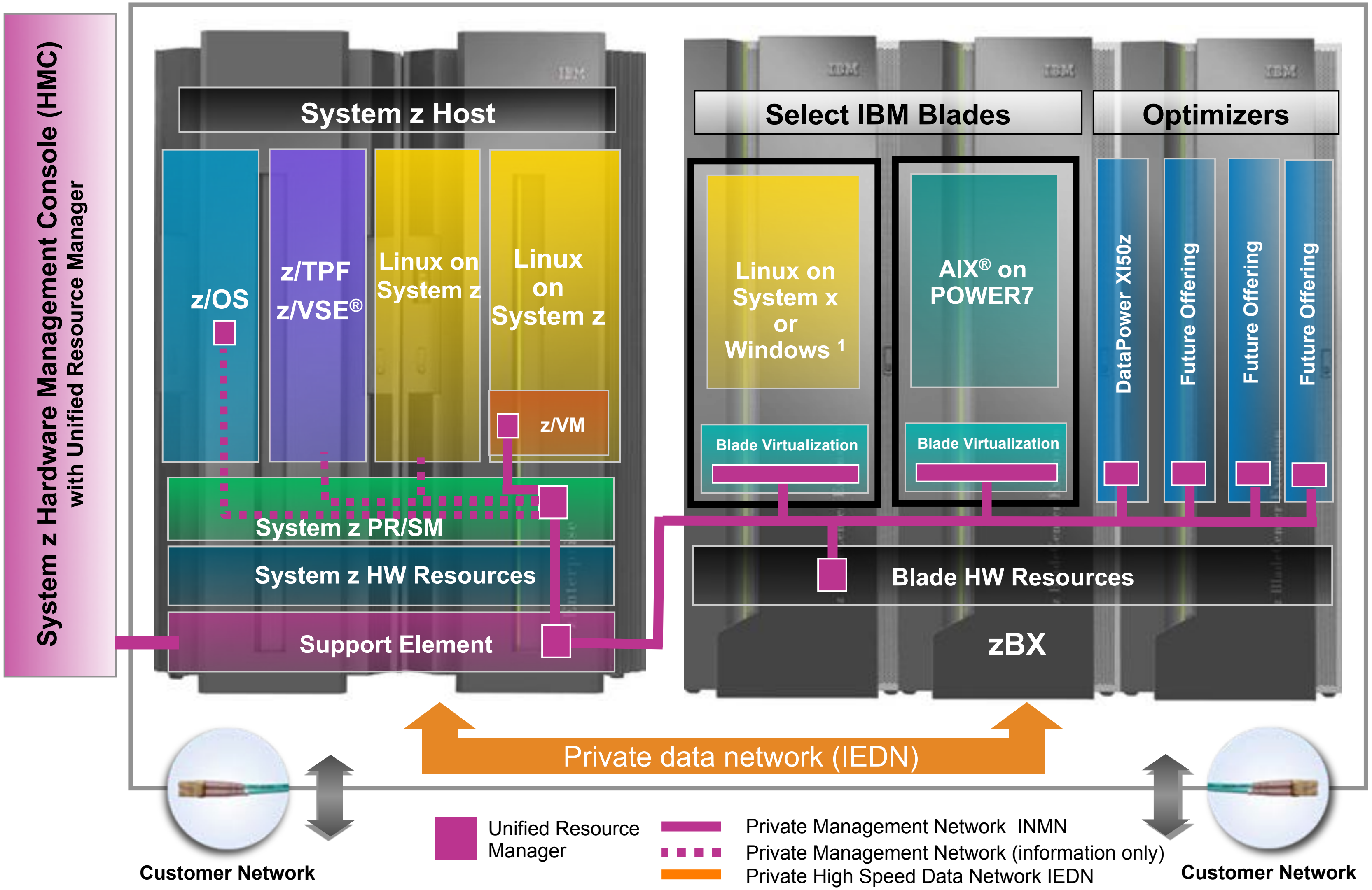
In July 2010, the IBM zEnterprise system introduced the first hybrid computing technology enabling clients to:

- Optimize the deployment of workloads by utilizing the best fit technology and operating environment
- Deploy enterprise private clouds that are ready for mission critical applications
- Establish a common management infrastructure for both mainframe and distributed-systems
- Take actionable insight based upon real time analytics

Central Processing Complex (CPC)
IBM zEnterprise™ Unified Resource Manager
IBM zEnterprise BladeCenter® Extension (zBX)

Putting zEnterprise System to the Task

Use the smarter solution to improve your application design



- ■ Unified Resource Manager
- ━ Private Management Network INMN
- ┅ Private Management Network (information only)
- ━ Private High Speed Data Network IEDN

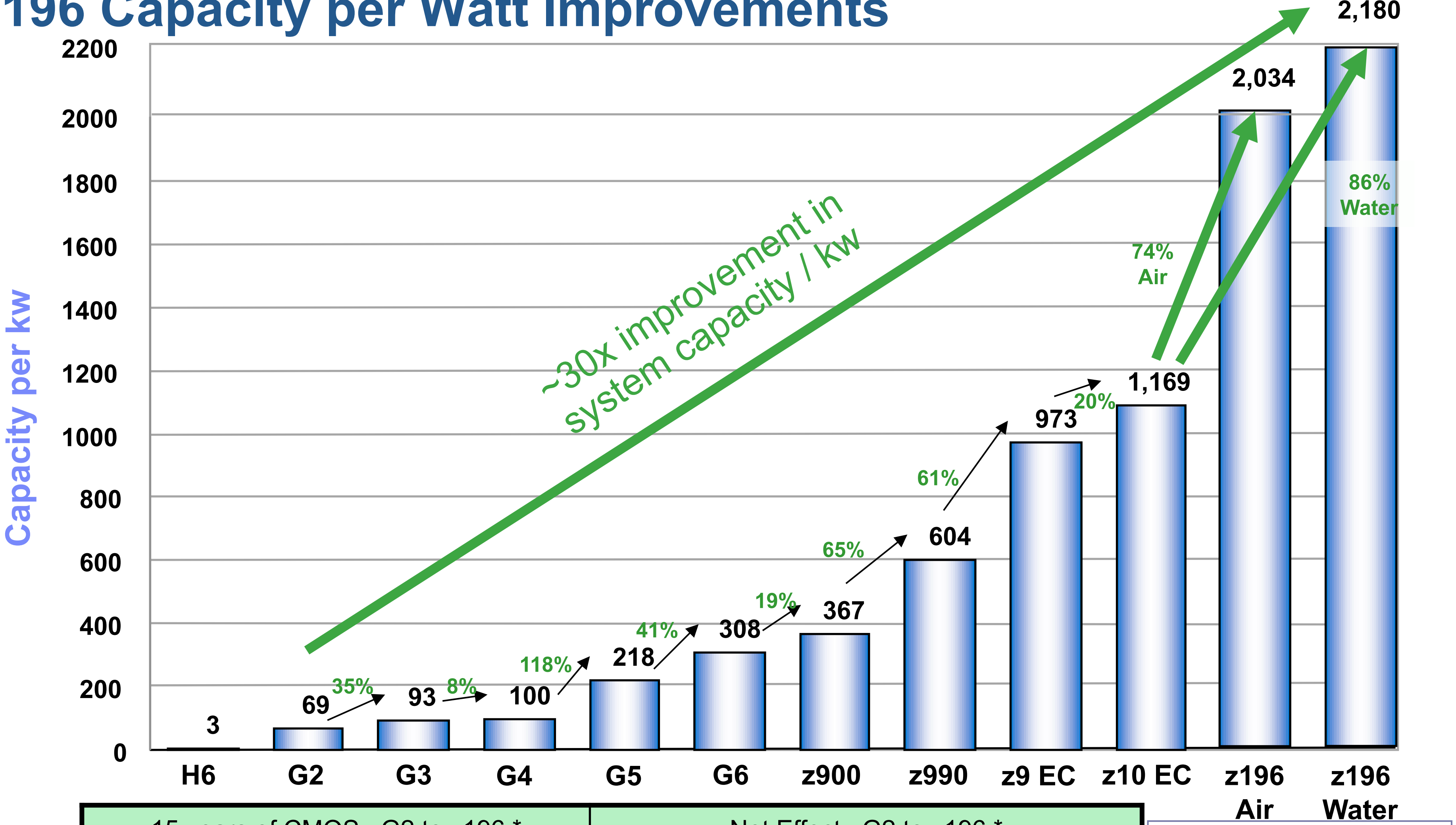
[1] All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represents goals and objectives only.

z196 – Helping to Control Energy Consumption in the Data Center

- Better control of energy usage and improved efficiency in your data center
- New water cooled option allows for energy savings without compromising performance
  - Maximum capacity server has improved power efficiency of 60% compared to the System z10 and a 70% improvement with water cooled option
- Savings achieved on input power with optional High Voltage DC by removing the need for an additional DC to AC inversion step in the data center
- Improve flexibility with overhead cabling option while helping to increase air flow in a raised floor environment
- z196 is same footprint as the System z10 EC[1]

[1] With the exception of water cooling and overhead cabling

z196 Capacity per Watt Improvements



15 years of CMOS: G2 to z196 *		Net Effect: G2 to z196 *	
Power Increase:	17% **per year**	**Performance increased by:**	~300x
Performance increase:	46% **per year**	Performance / kWatt increased by:	~30x
Power density increase:	13% **per year**	**Performance / sq ft increased by:**	~190x

Note: Capacity/kWatt assumes hot room, max plugged I/O power, max memory power and all engines turned on. Real world max capacity system is about 3/4 of this.

Summary

- There is a whole lot of hardware/firmware complexity under the covers for …
  - Performance
  - Reliability
  - Integrity/Security

  …. “But we worry about the details so you don’t have to”

- zEnterprise Instruction Set Architecture (ISA) continues to evolve
  - Close collaboration with software to optimize performance and functionality
- zBX opens up a new “hybrid computing” dimension to System z
  - Will likely continue this trend with more accelerator functions
- Energy efficiency will continue to improve

References

- IBM zEnterprise System Technical Introduction (SG24-7832-01)
- IBM zEnterprise 196 Technical Guide (SG24-7833-01)
- IBM zEnterprise 114 Technical Guide (SG24-7954-00)
- IBM zArchitecture Principles of Operation (POPS) (SA22-7832-08)

Thank You!

Email: stwalsh@au1.ibm.com