IBM System z: A Peek Under the Hood

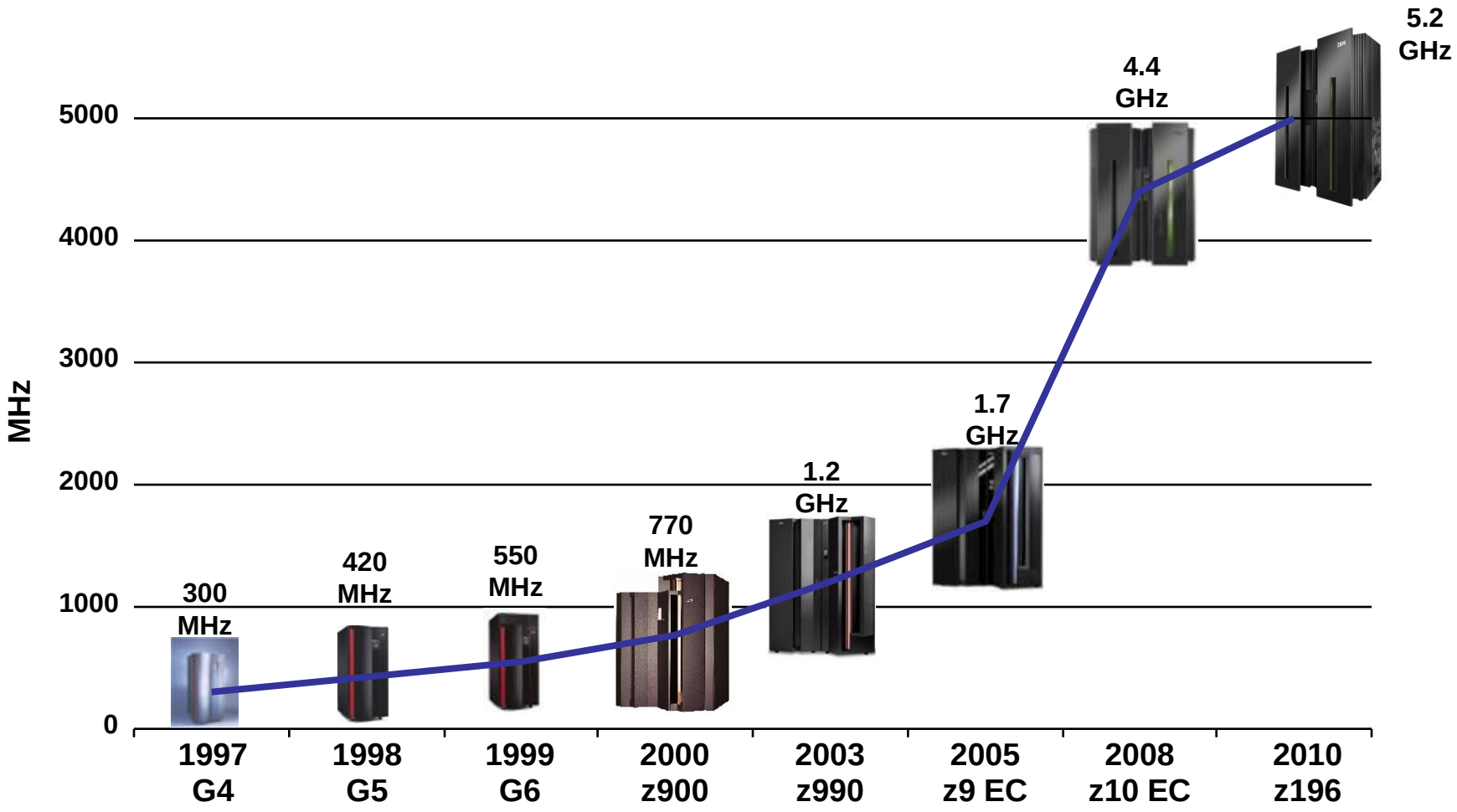
Tim Slegel
IBM Distinguished Engineer
System z Processor Development

June 13, 2011

Topics

- **Review of recent System z mainframes**
- **Processor hardware overview**
- **Millicode and Virtualization**
- **Cache/memory subsystem**
- **New Instruction Set Architecture for z196**
- **zBX: A system of systems**
- **Energy efficiency**

IBM zEnterprise 196 Continues the CMOS Mainframe Heritage



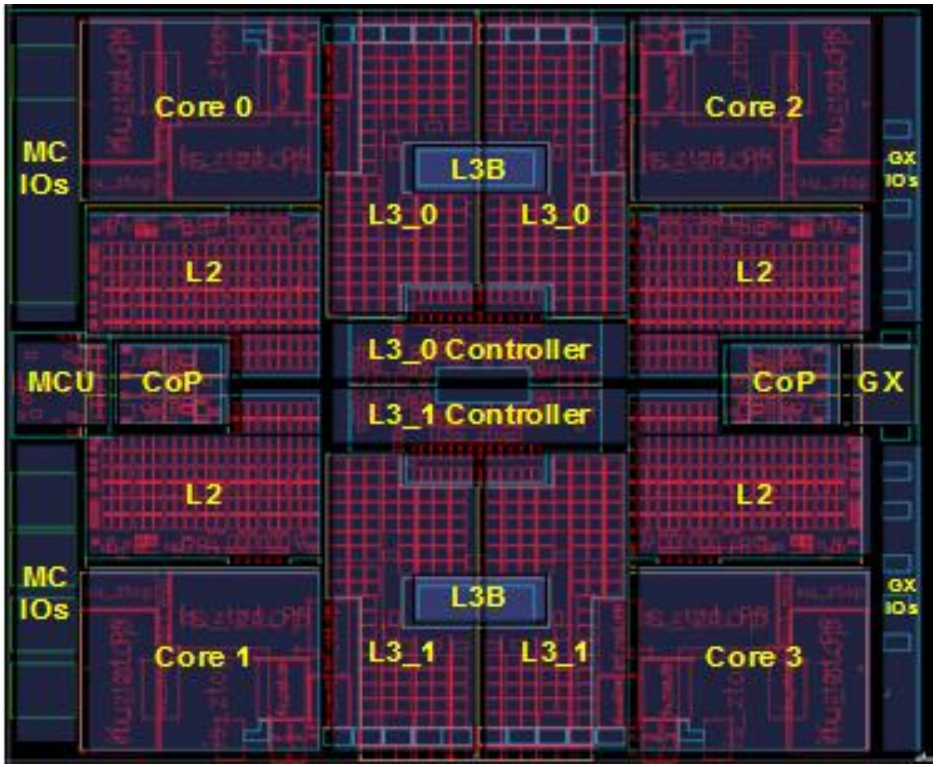
- **G4** – 1st full-custom CMOS S/390®
- **G5** – IEEE-standard BFP; branch target prediction
- **G6** – Copper Technology (Cu BEOL)

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

- **z10 EC** – Architectural extensions
- **zEnterprise** – Additional Architectural extensions

System Hardware, Firmware, and Software

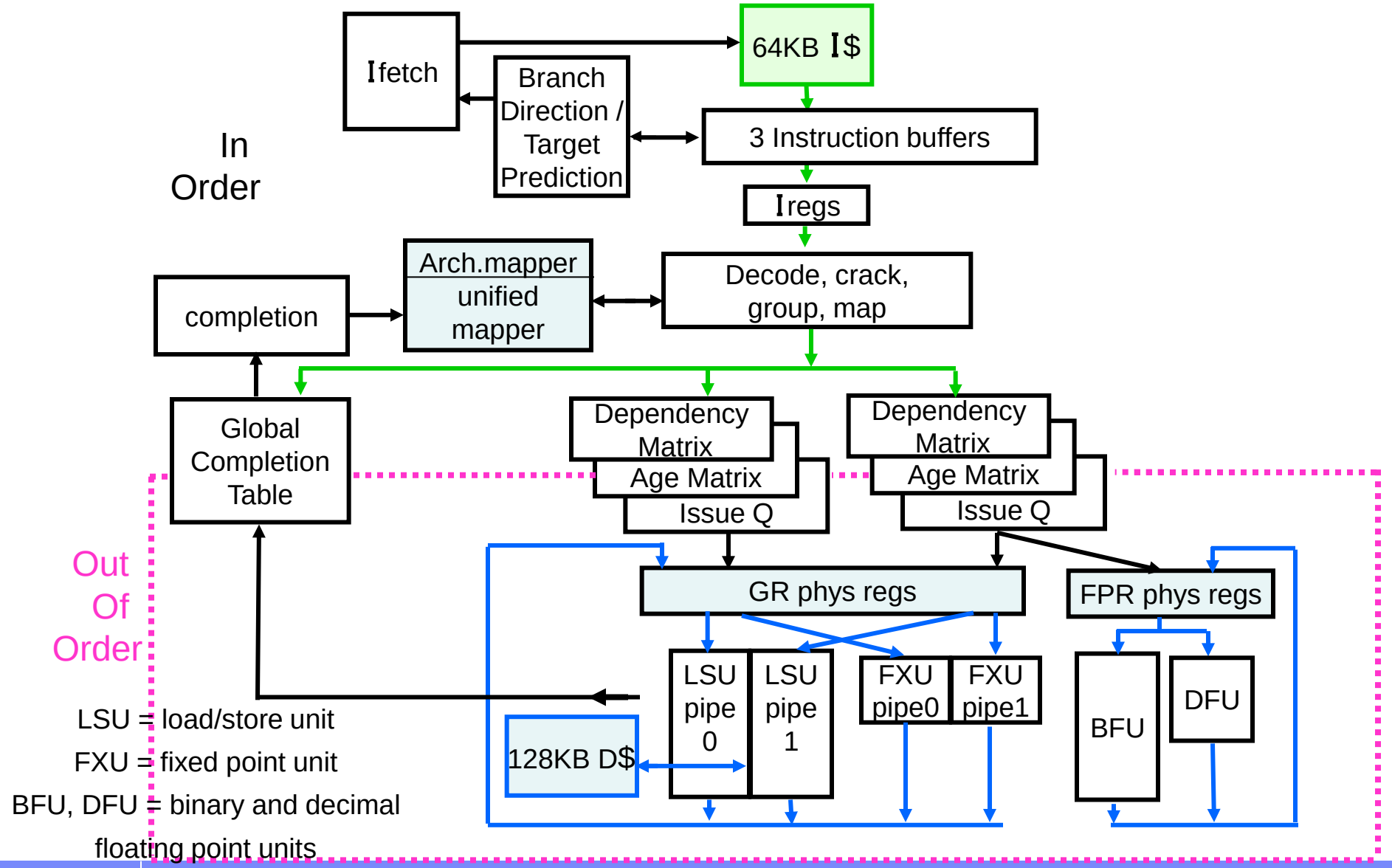
Quad Core zEnterprise 196 Processor Chip



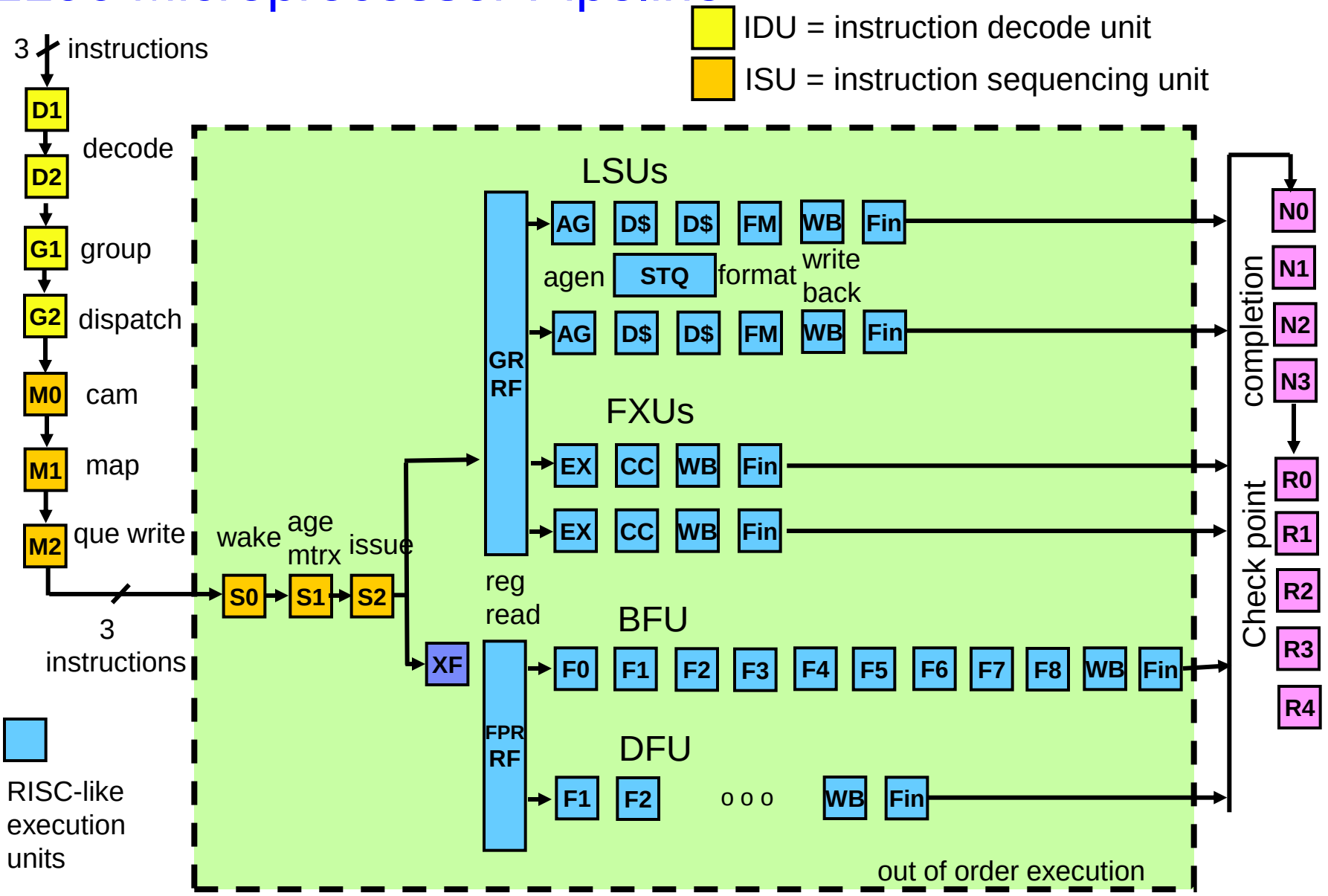
- **45nm SOI Technology**
  - 13 layers of metal
  - 3.5 km wire
- **Chip Area – 512.3mm²**
  - 23.5mm x 21.8mm
  - 8093 Power C4's
  - 1134 signal C4's
- **1.4 Billion Transistors**

- **Up to Four active cores per chip**
  - 5.2 GHz system operation – fastest processor in the world
  - L1 cache/ core
    - 64 KB I-cache
    - 128 KB D-cache
  - 1.5 MB private L2 cache/ core
- **Two Co-processors (COP)**
  - **Crypto & compression accelerators**
  - Includes 16KB cache
  - Shared by two cores
- **24MB eDRAM L3 Cache**
  - Shared by all four cores
- **Interface to SC chip / L4 cache**
  - 40+ GB/sec to each of 2 SCs
- **I/O Bus Controller (GX)**
  - Interface to Host Channel Adapter (HCA)
- **Memory Controller (MC)**
  - Interface to controller on memory DIMMs
  - Supports RAIM design

z196 Microprocessor Core



z196 Microprocessor Pipeline



z196 CPU core

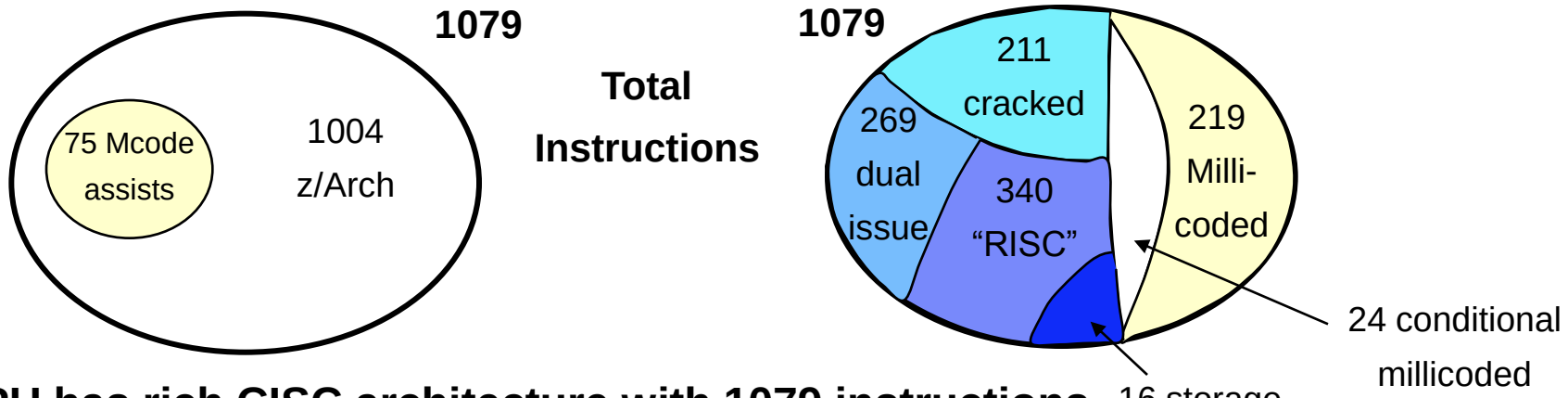
- **Each core is a superscalar, out of order processor:**
  - **Cycle time is 5.2 GHz**
  - **Six RISC-like execution units**
    - **2 fixed point (integer), 2 load/store, 1 binary floating point, 1 decimal floating point**
  - **Up to three instructions decoded per cycle (vs. 2 in z10)**
  - **Up to five instructions/operations executed per cycle (vs. 2 in z10)**
    - **Execution can occur out of (program) order**
    - **Memory address generation and memory accesses can occur out of (program) order**
    - **Special circuitry to make execution and memory accesses appear in order to software**
  - **Each core has 3 private caches**
    - **64KB 1st level cache for instructions, 128KB 1st level cache of data**
    - **1.5MB L2 cache containing both instructions and data**

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

Extensive use of hardware speculation

- **z/Architecture places many strict constraints on how the CPU has to <u>appear</u> to be behave**
  - Example – Strict storage ordering rules (see POPS chapter 5)
  - Good for software – significantly easier and more robust MP programming than other ISAs
  - Bad for the CPU design team – 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 as 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 rules

Instruction Set Architecture (ISA)



- **CPU has rich CISC architecture with 1079 instructions**
  - 75 assists usable by millicode (vertical microcode) only
- **Most complex instructions are executed by millicode**
  - Another 24 instructions are conditionally executed by millicode
- **Medium complexity instructions cracked at decode into 2 or more μops**
- **Most RX instructions cracked at issue → dual issued**
  - RX have one storage operand and one register operand
- **Some storage-storage ops executed by LSU sequencer**
- **Remaining z instructions are RISC-like and map to single μop**

Millicode

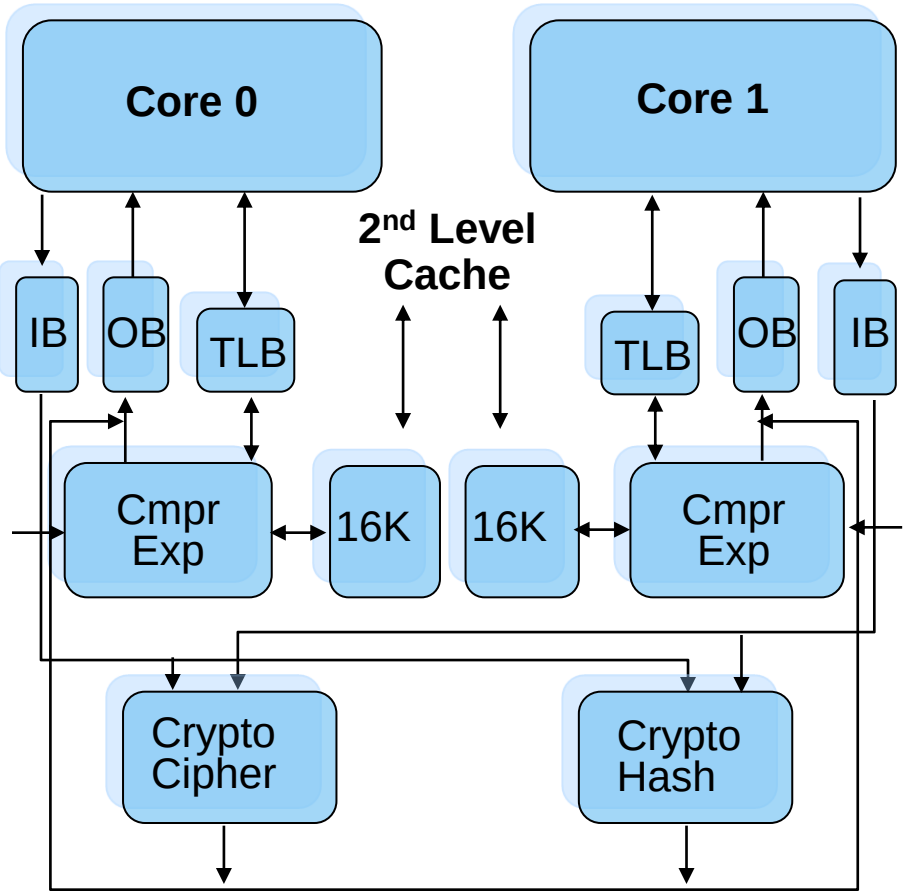
- **Our name for the vertical microcode that executes on the processor**
- **Runs in a special mode on the normal processor pipeline – no specialized microcode engine**
- **Written in assembler (with optimizers and semantic correctness tools)**
- **Most z/Architecture instructions are available for use in millicode routines**
- **Resides in HSA and is cached in the I-cache. Storage operands can be in the D-cache.**
- **Separate set of millicode General Purpose Registers**
- **Special millicode assist instructions**
  - Move data to/from micro-architected control registers and facilities
  - Performance enhancing instructions (over the years, some of these have been transferred to POPS and are usable by normal software)
  - Pipeline controls
  - Ability to move data anywhere in storage – between LPAR partitions or to/from HSA
  - CoP access for crypto and compression
  - Perform System Operations (page mover engine, multi-CPU operations such as broadcast TLB purges, I/O operations, service functions, etc.)
- **Interestingly, for some millicode instructions full pipeline interlocks are not maintained in hardware**
  - E.g., read after write of a special register may not yield the updated value
  - Improves performance and simplifies hardware complexity but makes it more difficult to write millicode

Hardware/Millicode Support for Virtualization

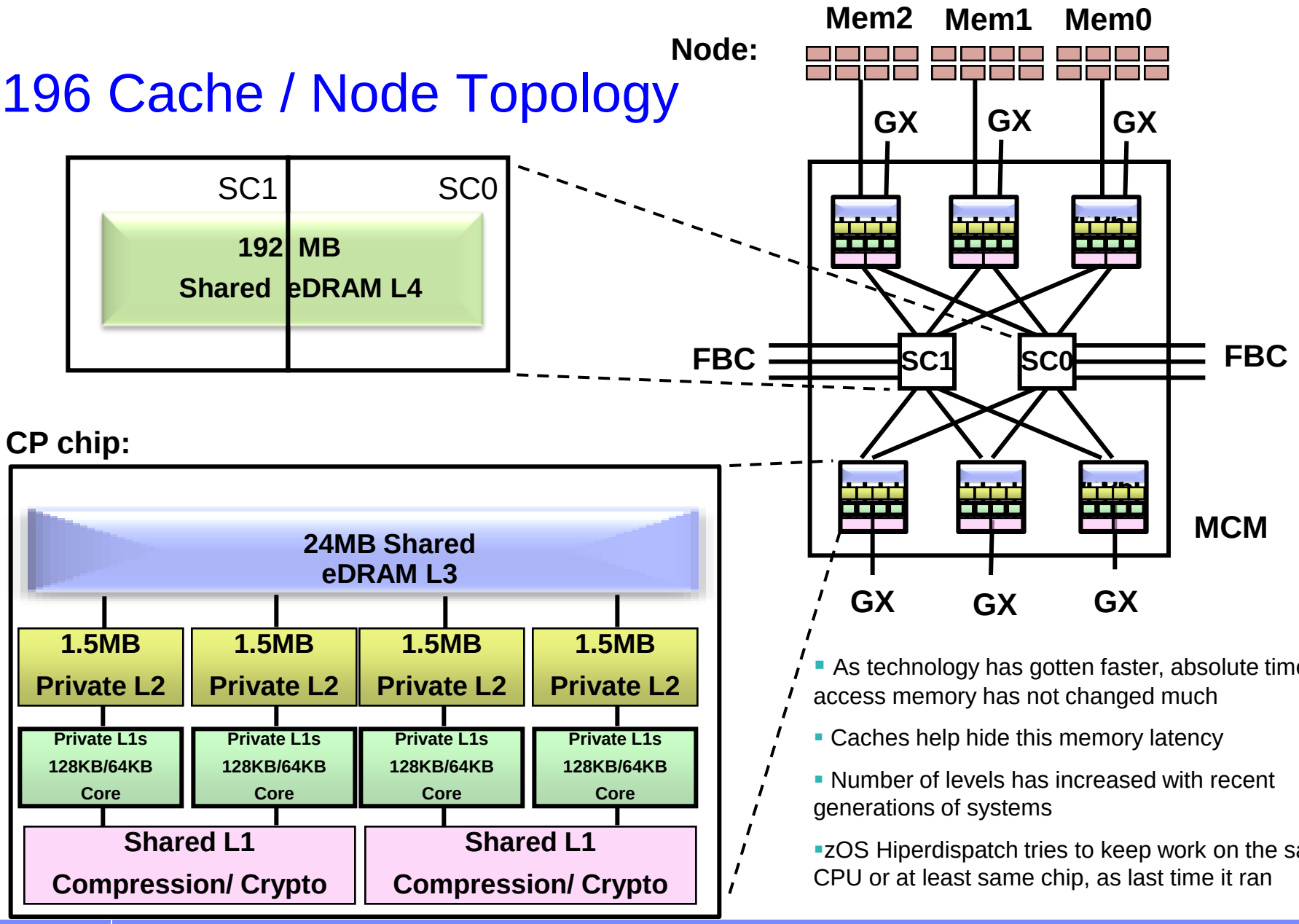
- **Full logical virtualization via the START INTERPRETIVE EXECUTION (SIE) instruction**
- **Nested SIE supports two level guests:**
  - LPAR Hypervisor (firmware) runs natively. First level guests are normal OSes (e.g., zOS, zLinux, zVM). Up to 60 1st level partitions.
  - If zVM is running as a first level guest, then it supports hundreds (or thousands) of second level guests (e.g., zLinux)
- **Separate hardware Host/Guest-1/Guest-2 facilities:**
  - z/Architecture control registers
  - Timing Facility (including interrupt controls)
- **All important SIE State Description controls are buffered into hardware control registers during SIE-entry/exit, which is performed by millicode**
- **Hardware detects most SIE Intercept and Intervention conditions**
- **Full hardware support for SIE address translation:**
  - RRF supports zone relocation (and zone based I/O interrupts)
  - Multi-level pageable guest support (up to 56 table fetches required for a single 2nd level guest ART/DAT translation)
  - MCDS handling of ARs
  - TLB2 holds multiple SIE guests entries simultaneously
  - Appropriate TLB purging on all CPUs for IPTE/IDTE operations with filtering

z196 Compression and Cryptography Accelerator

- **Data compression engine**
  - **Static dictionary compression and expansion**
  - **Dictionary size up to 64KB (8K entries)**
    - **Local 16KB cache per core for dictionary data**
- **CP Assist for Cryptographic Function (CPACF)**
  - **Enhancements for new NIST standard**
  - **Complemented prior ECB and CBC symmetric cipher modes with XTS, OFB, CTR, CFB, CMAC and CCM**
  - **New primitives (128b Galois Field multiply) for GCM**
- **Accelerator unit shared by 2 cores**
  - **Independent compression engines**
  - **Shared cryptography engines**

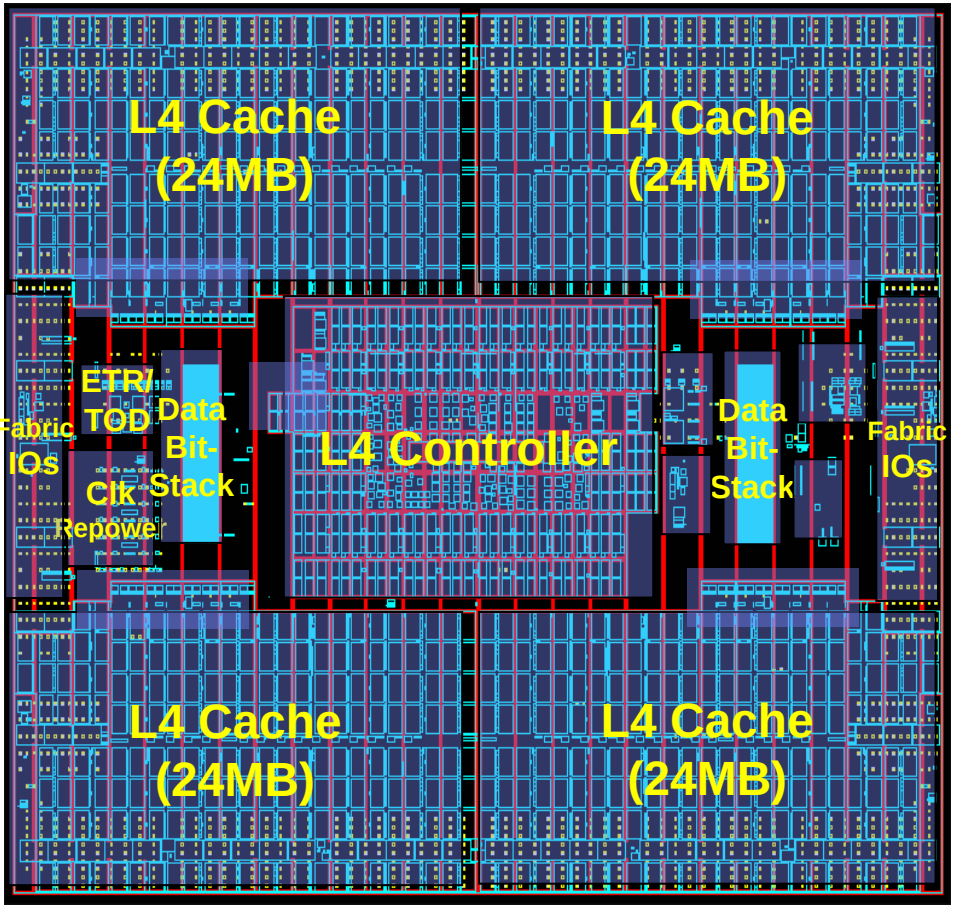


z196 Cache / Node Topology



- As technology has gotten faster, absolute time to access memory has not changed much
- Caches help hide this memory latency
- Number of levels has increased with recent generations of systems
- zOS Hiperdispatch tries to keep work on the same CPU or at least same chip, as last time it ran

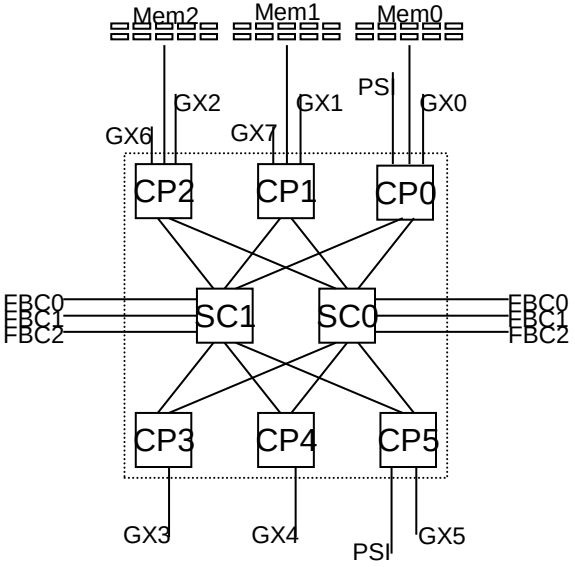
Hub / Shared Cache Chip for z196

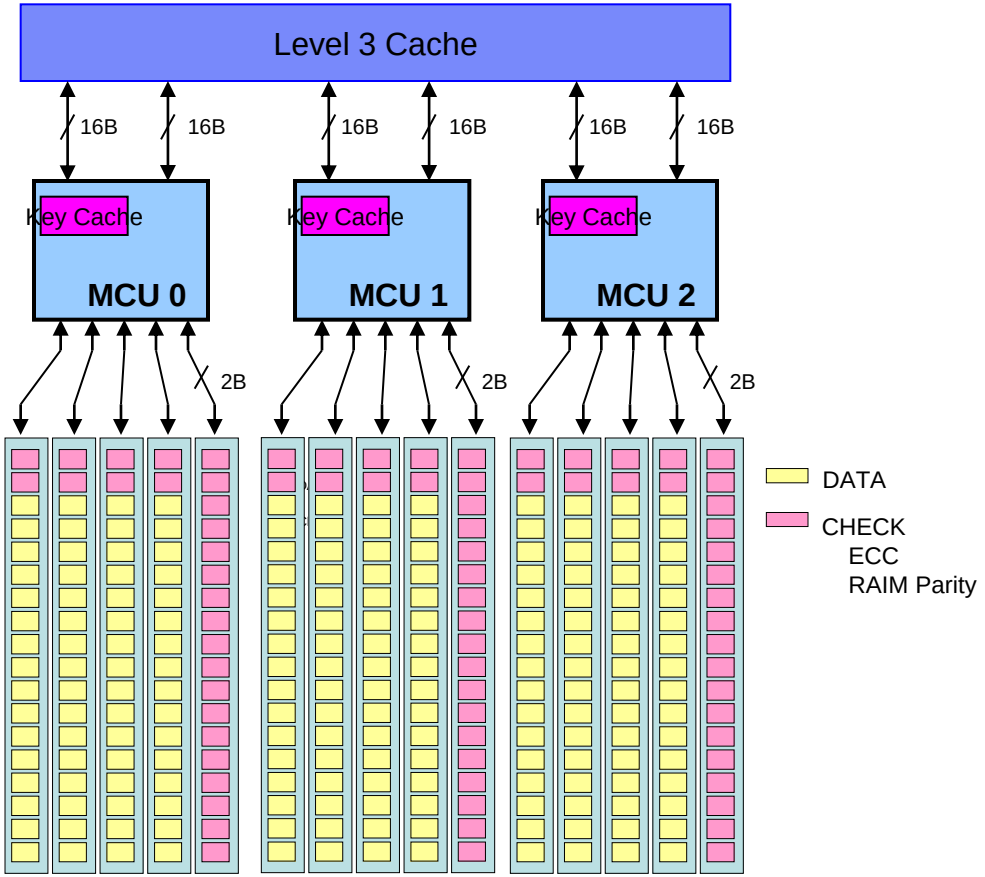


- **eDRAM Shared L4 Cache**
  - 96 MB per SC chip
  - 192 MB per Node
- **6 CP chip interfaces**
  - 40+ GB/sec each
- **3 Fabric interfaces**
  - 40+ GB/sec each
- **45nm SOI Technology**
  - 13 layers of metal
- **Chip Area – 478.8mm^2**
  - 24.4mm x 19.6mm
  - 7100 Power C4's
  - 1819 signal C4's
- **1.5 Billion Transistors**
  - 1 Billion cells for eDRAM

z196 RAIM Memory Structure

- **Redundant Array of Independent Memory**
  - 5 channel memory controller
  - DIMM bus CRC error retry
  - Industry leading reliability
- **Up to 3TB Memory capacity**
  - 3 MCUs per MCM
  - 2-deep DIMM cascade





Reliability/Availability Features

- **Near 100% hardware error detection for logic faults – far higher than other platforms**
- **Multi-level error recovery capability:**
  - On-the-fly error correction of array errors. Automatically deletes failing sections of arrays for solid errors.
  - Within the processor, all instructions are checkpointed in fault-hardened registers/arrays. If a hardware error is detected, processor retry allows for the re-execution of the failed instruction. Effective for soft-errors.
  - In the event of a hard-error where retry is unsuccessful, Dynamic Processor Sparing moves the entire micro-architected state to a spare processor. Happens transparently to software and even the OS.

z196 New Instruction Set Architecture

- **High-word extension**
  - General register high-word independently usable for loads, stores, arithmetic, logical operations
  - Gives software up to 32 word-sized registers
- **Conditional load, store, register copy**
  - Based on condition code
  - Used to eliminate unpredictable branches
- **New atomic instructions**
  - Load and arithmetic (ADD, AND, XOR, OR)
  - (Old) storage location value loaded into GR with arithmetic/logical result updating storage
- **Load Pair Disjoint**
  - Load from two different storage locations into two GRs. Condition code in indicates whether fetches were atomic
- **Distinct-Operands Facility (22 new instructions)**
  - Independent specification of result register (different than either source register) which reduces value copying
- **Population-Count Facility**
- **Virtual Architecture Level**
  - Allows the zVM Live Guest Relocation Facility to make a z196 behave architecturally like a z10 system
  - Facilitates moving work transparently between z196 and z10 systems for backup and capacity reasons
- **Non-quiescing SSKE**
  - Significant performance improvement for systems with large number CPUs
- **Integer to/from Floating point converts (39 new instructions)**
- **New truncate and OR inexactness Binary Floating Point rounding mode**
- **New Decimal Floating Point quantum exception**
- **Eliminates need for test data group for every operation**
- **Other minor architectural enhancements**

IBM zEnterprise System – Best-in-class systems and software technologies

A "System of Systems" that unifies IT for predictable service delivery

IBM zEnterprise 196 (z196)

- Optimized to host large-scale database, transaction, and mission-critical applications
- The most efficient platform for large-scale Linux consolidation
- Capable of massive scale-up
- New easy-to-use z/OS V1.12

zEnterprise Unified Resource Manager

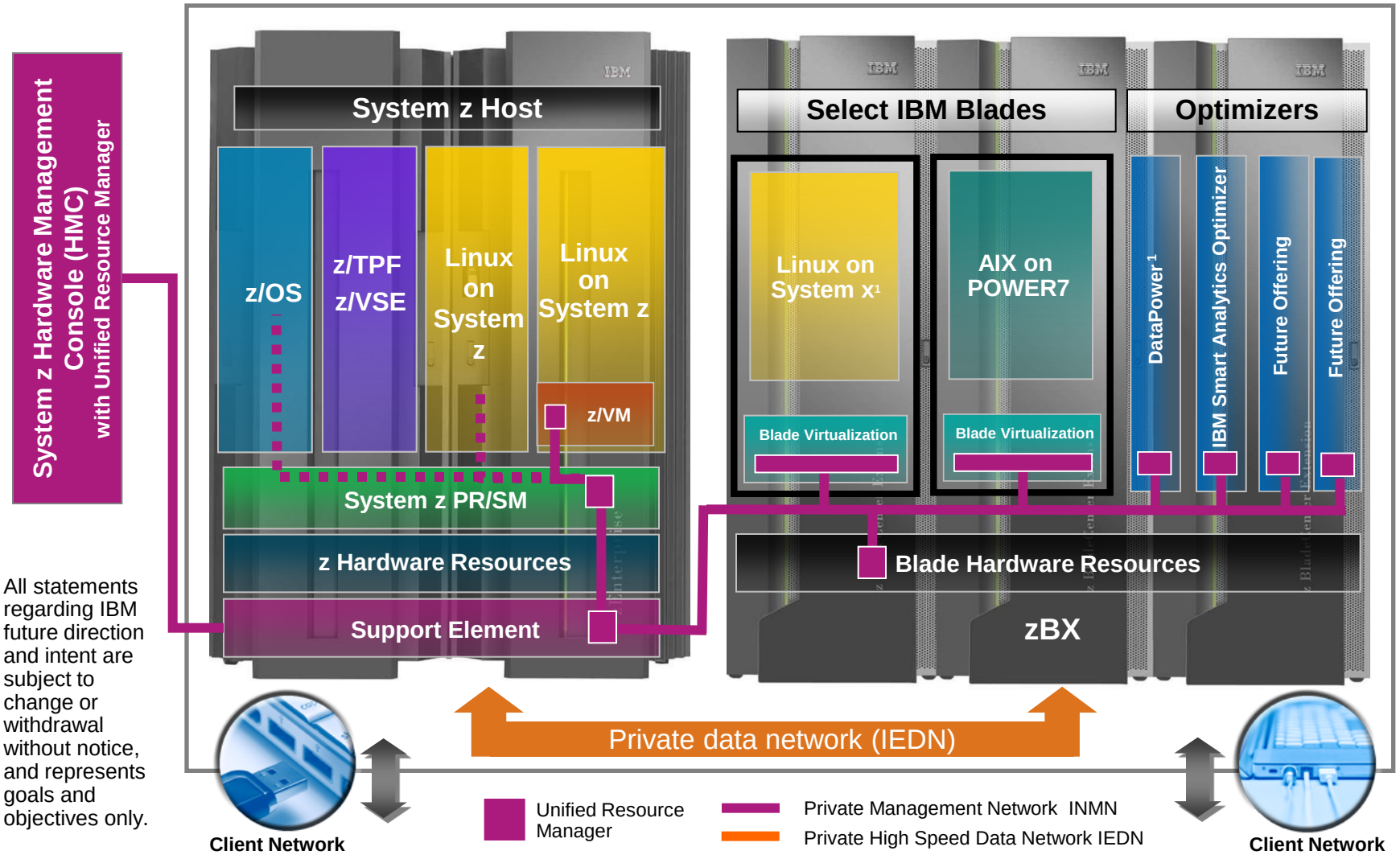
- Unifies management of resources, extending IBM System z qualities of service end-to-end across workloads
- Provides platform, hardware and workload management

zEnterprise BladeCenter Extension (zBX)

- Select IBM POWER7® and IBM x86* blades for tens of thousands of AIX, Linux and Windows applications
- High-performance optimizers and appliances to accelerate time to insight and reduce cost
- Dedicated high-performance private network

Putting zEnterprise System to the task

Use the smarter solution to improve your application design



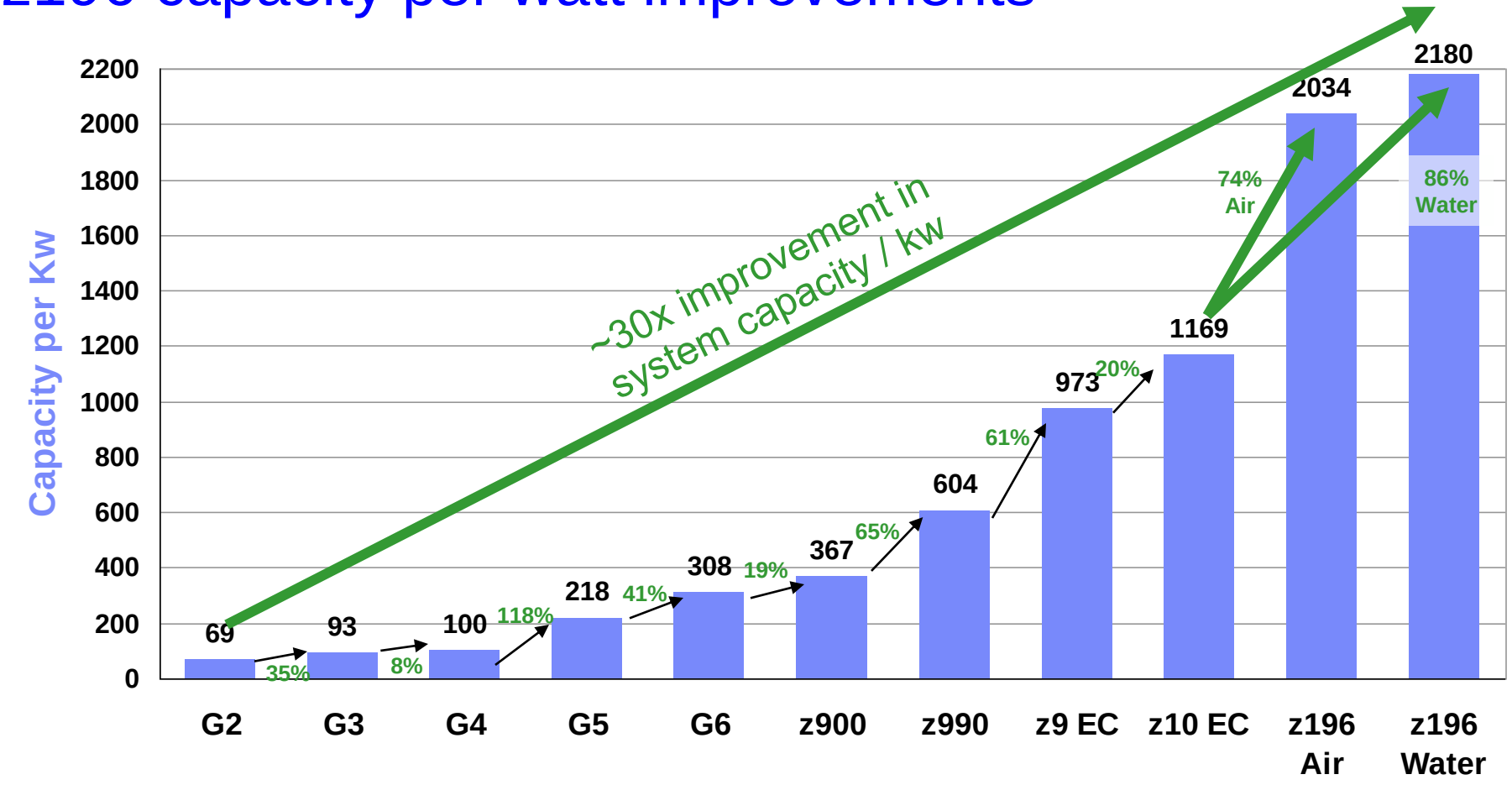
1. All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represents goals and objectives only.

z196 – Helping to control energy consumption in the data center

- **Better control of energy usage and improved efficiency in the 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 lot of hardware/firmware complexity under the covers for:**
  - Performance
  - Reliability
  - “But, we worry about the details, so you don’t have to.”
- **Instruction Set Architecture continues to evolve**
  - Close collaboration with software to optimize performance and functionality
- **zBX opens up a new dimension in System z**
  - Will likely continue this trend with more accelerator functions
- **Energy efficiency will continue to improve**

Thank you!

- **Feel free to contact me offline with questions on IBM System z performance, functionality, etc.**
- **e-mail: slegel@us.ibm.com**