IBM Z / LinuxONE Processor Optimization Primer

v5 (February 2026) – updates vs v4 in blue

David Hutton, STSM, Master Inventor

IBM Z Performance and Design

Poughkeepsie Engineering

hutton@us.ibm.com

Trademarks

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

CICS	FICON	IBM	IMS	PR/SM	z10, z10 EC, z10 BC
Cloud Paks	GDPS	ibm.com	LinuxONE	Spectrum Scale	z196, z114
Db2	HiperSockets	IBM logo	Parallel Sysplex	System Storage	zEC12, zBC12
DFSMS	Hyper Swap	IBM Sterling Connect:Direct	Power	z/OS, z/VM, z/VSE	z13, z13s
					z14, z14 ZR1
					z15, z15 T01, z15 T02
					z16, z16 A01, z16 A02, Telum
					z17, z17 ME1, Telum II

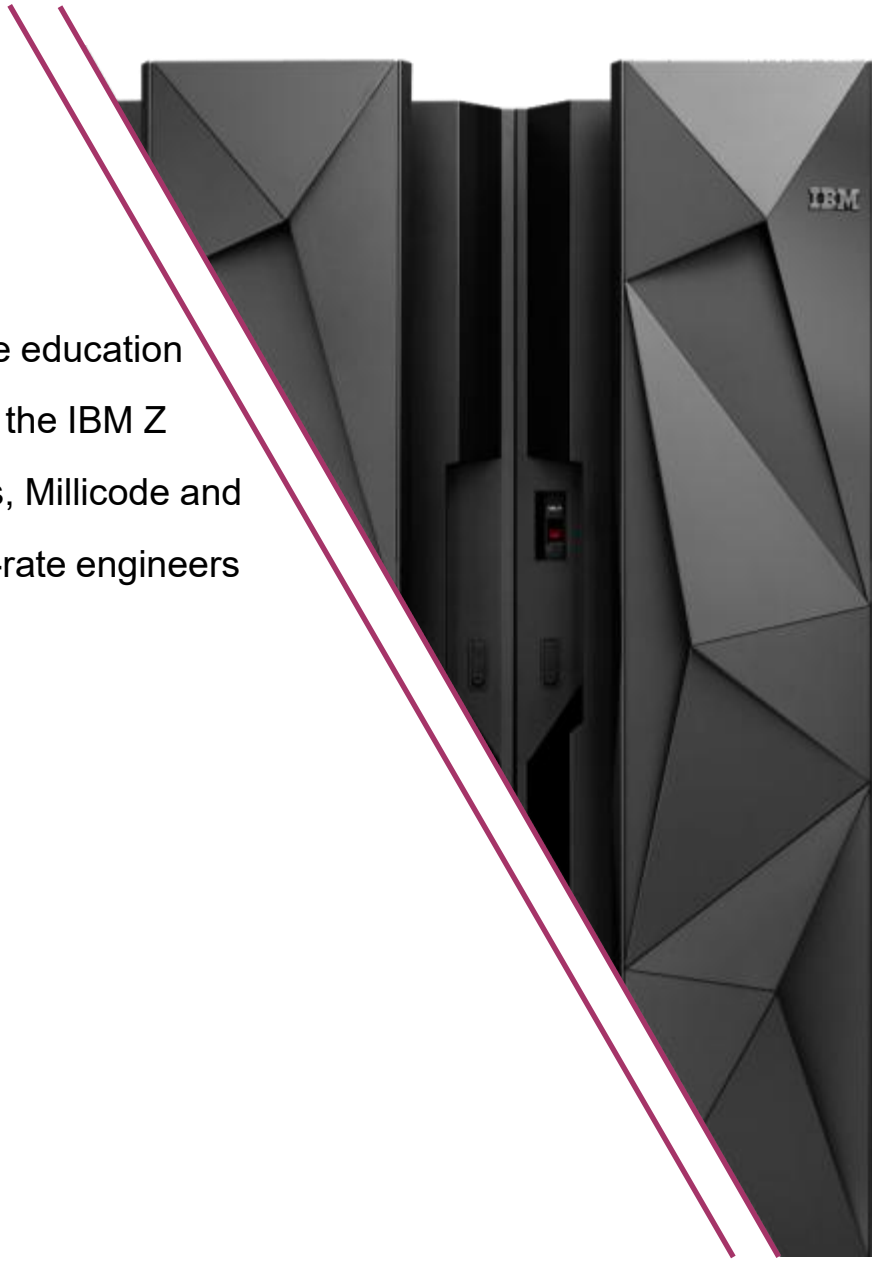
The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Red Hat®, JBoss®, OpenShift®, Fedora®, Hibernate®, Ansible®, CloudForms®, RHCA®, RHCE®, RHCSA®, Ceph®, and Gluster® are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.

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.
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.

Acknowledgments

While this document in its various forms has had multiple owners, it is an externalized compilation of many in-house education and design documents and represents contributions from the IBM Z Microprocessor and Nest Design and Performance teams, Millicode and Firmware development, and Architecture – too many first-rate engineers worldwide to name here.

Thank you!!



Documentation Objectives

This document provides an overview of the processor subsystems of IBM Z / LinuxONE systems, with focus on the core microarchitectures from z196 to z17

> Version 5 (v5) includes minor corrections, clarifications, additional information, and z17 design updates from z16

As optimal performance comes from a balance of hardware and software, this document gives architectural (IBM Z Principles Of Operation) and microarchitectural (z196..z17-specific) insights with information and potential methods to optimize for code performance and is intended for IBM Z binary or assembly code developers.

This document fosters a deep technical exchange with non-IBM development teams within the open source and z/OS assembler communities to encourage performance optimization for applications running on IBM Z / LinuxONE system processors

> *“Tell us what you would like to see added in future releases!”*

This document is not intended to be a comprehensive write-up and should not replace any formal architecture documents

Note: When a machine short-name is used with a + sign, like “z196+”, it means the statement applies to z196 and machines after.

z/Architecture and Implementation

- z/Architecture[1] is a 64-bit architecture that is supported by IBM Z / LinuxONE microprocessors
  - A Complex Instruction Set Computer (CISC) architecture, including highly capable (and thus complex) instructions
  - Big-Endian (BE) architecture (vs. Little-Endian) where bytes of a multi-byte operand data element are stored with the most significant byte (MSB) at the lower storage address
- z/Architecture grows compatibly upon each generation, and includes many innovative features
  - Typical load/store/register-register/register-storage instructions, including logical and arithmetic functions
  - Branch instructions supporting absolute and relative offsets, and subroutine linkages
  - Storage-storage instructions, e.g., “MOVE characters (MVC)” (for copying characters), including decimal arithmetic
  - Hexadecimal, binary, and decimal (both IEEE 754-2008 standard) floating-point operations
  - Vector (SIMD) operations on z13+, including fixed-point, floating-point, and character string operations; decimal operations added on z14+
  - Atomic operations including COMPARE AND SWAP, LOAD AND ADD, and OR (immediate) instructions
  - *Hardware transactional memory, through the Transactional Execution Facility (since zEC12), including the definition of a constrained transaction that can be retried by the hardware – support being sunset starting with z17*
  - Two-way Simultaneously Multi-Threading (SMT-2) support (since z13)
  - Further z14+ updates beginning on page 11
- Highly complex instructions are implemented through a special firmware layer – millicode[2]
  - Millicode is a form of vertical microcode that is pre-optimized for each processor generation
  - An instruction that is implemented in millicode is executed by the hardware similar to a built-in subroutine call that transparently returns back to the program when the millicode routine ends
  - A millicode instruction routine consists a subset of the existing instructions in the z/Architecture, with access to its own pool of internal registers in addition to program registers and specialized hardware instructions
  - Some complex routines might involve operating along with a private co-processor or special hardware that is only accessible by millicode

Microprocessor CPU State

- Under z/Architecture, the architected states of a software thread running on a microprocessor core, referred to as **CPU**, involves the following highlighted components*
  - Program Status Word (PSW)
    - Instruction Address (aka Program Counter, including where the next instruction address is for execution)
    - Condition Code (2 bits, set depending on results of some previously executed instructions)
    - Addressing Mode (2 bits, indicating 24-bit, 31-bit or 64-bit addresses)
    - DAT Mode (when 1, indicates that implicit dynamic addressing translation is needed to access storage)
    - Address Space Control (controls translation modes: Access Register specified, Primary, Secondary, or Home)
    - Key (4-bit access key that is used to access storage when key-controlled protection applies)
    - Wait state (waiting, no instructions are processed)
    - Problem State (when 1, indicates problem state, not supervisor state; privileged instructions cannot be executed, and semi-privileged instructions can be executed only if certain authority tests are met)
    - Masks (control various kinds of interrupt enablement)
  - Registers
    - Access Registers (ARs): 16 total, 32 bits each, used mainly during access-register based translation mode
    - General Registers (GRs): aka general-purpose registers, 16 total, 64 bits each, with high and low 32-bit-word independent operations for address arithmetic, general arithmetic, and logical operations
    - Floating-Point Registers (FPRs): 16 total, used by all floating-point instructions regardless of formats; a register can contain either a short (32-bit) or a long (64-bit) floating-point operand; while a pair can be used for extended (128-bit) operands
    - Vector Registers (VRs): available since z13, 32 total, 128 bits each, when present, FPRs overlay the VRs
    - Floating-Point-Control Register: 32-bit, contains mask, flag and rounding mode bits, and a data exception code
    - Control Registers (CRs): 16 total, bit positions in the registers are assigned to defined architectural facilities in the system

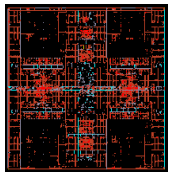
**For more information, see z/Architecture Principles of Operations[1] (POPs)*

Highlights of the Recent Microprocessor Cores

- The z10 processor[3,4] (65nm technology) started the recent ultra-high frequency pipeline design in Z processors
- z196[5,6] (45nm) introduces the first generation out of order pipeline design
  - Runs at 5.2 GHz on the EC class machines
  - Introduces high-word architecture with operations on upper 32 bits of general registers (GRs)
  - Adds more nondestructive arithmetic instructions
  - Adds conditional load and store instructions, for reducing potential branch wrong penalties
- zEC12[7] (32nm) improves upon the first generation out of order design
  - Runs at 5.5 GHz on the EC class machines
  - Introduces level-2 branch prediction structure[8]
  - Introduces a set of split level-2 (L2) caches, providing low-latency large capacity instruction and operand data caching per processor core
  - Integrates tightly L2 data cache lookup into level-1 (L1) data cache design, further improves L2 data cache access latency
  - *Supports Hardware Transactional Memory[9] (Execution) and Run-Time Instrumentation facilities – support being sunset starting with z17*
- z13[10] (22nm) improves further on top of the zEC12 design
  - Runs at a slightly lower maximum frequency of 5 GHz; with a much wider pipeline (2x) to handle more instructions per cycle for a net increase in overall instruction execution rate
  - Integrates L2 instruction cache lookup into L1 instruction cache design to improve L2 instruction cache access latency
  - Supports simultaneous multi-threading (SMT) for 2 threads
  - Introduces Single-Instruction-Multiple-Data (SIMD) instructions for vector operations[1]
- z14[14] (14nm) and z15[16] (14nm) improve upon z13
  - Runs at a slightly higher maximum frequency of 5.2 GHz
  - Improves performance with innovative core and cache subsystem enhancements
  - Provides new system level features by supporting new architecture
  - Further updates highlighted in page 12 and 13
- z16[18,19,20] (7nm) and z17 (5nm) improve upon z15
  - z16 continues at 5.2GHz while z17 runs at 5.5GHz and ties zEC12 for the industry's fastest-clocked processor to date
  - Adds an on-chip AIU (artificial intelligence unit) HW accelerator that supports 8x8x2 matrix operation for common tensor transformations
  - Adds dynamic container-based branch prediction to improve utilization of branch history real estate which yields better predictions
  - Physical Level 2 caches bidirectionally ring-interconnected at the chip level while the chips are point-to-point interconnected at the drawer level to virtualize chip-level L3 and drawer-level L4 caches – ***another industry first!***
  - z16 processor marketed as "Telum", z17 processor marketed as "Telum II"
  - z17 redesigned and relocated its principal IO-card ASIC onto the CP chip as the Data Processing Unit or DPU

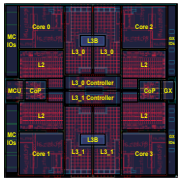
IBM Z Processor Historic Roadmap

65 nm
z10
2/2008



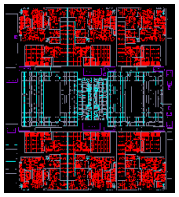
- Workload Consolidation and Integration Engine for CPU Intensive Workloads
- Decimal FP
- InfiniBand
- 64-CP Image
- Large Pages
- Shared Memory

45 nm
z196
9/2010



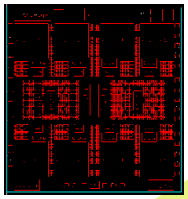
- Top Tier Single Thread Performance, System Capacity
- Accelerator Integration
- Out of Order Execution
- Water Cooling
- PCIe I/O Fabric
- RAIM
- Enhanced Energy Management

32 nm
zEC12
9/2012



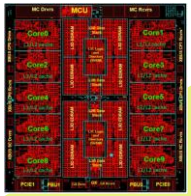
- Leadership Single Thread, Enhanced Throughput
- Improved out-of-order
- Transactional Memory
- Dynamic Optimization
- 2 GB page support
- Step Function in System Capacity

22 nm
z13
3/2015



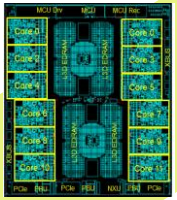
- Leadership System Capacity and Performance
- Modularity & Scalability
- Dynamic SMT
- Supports two instruction threads
- SIMD
- PCIe attached accelerators
- Business Analytics Optimized

14 nm
z14
9/2017



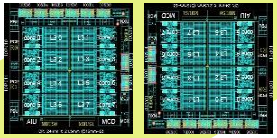
- Pervasive encryption
- Low latency I/O for acceleration of transaction processing for DB2 on z/OS
- Pause-less garbage collection for enterprise scale JAVA applications
- New SIMD instructions
- Optimized pipeline and enhanced SMT
- Virtual Flash Memory

14 nm
z15
9/2019



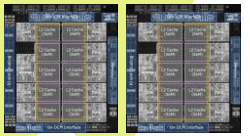
- Focus on power efficiency and new on-chip architectures
- Improved and enlarged caches
- Optimized Out-of-Order architecture
- IBM Integrated Accelerator for zEDC (On-chip compression support (DEFLATE)) and Z Sort
- Elliptic Curve Cryptography Accelerator

7 nm
z16
5/2022



- **Industry-first nest design that virtualizes chip-level 3 and drawer-level 4 caches from physical core-attached 32MB level 2 caches**
- Artificial Intelligence Unit (AIU) on-chip accelerator
- Dynamic container-based branch prediction
- Sysplex latency and scalability
- Quantum-safe and homomorphic cryptography

5 nm
z17
6/2025



- **Data Processing Unit (DPU) which redesigns and relocates IO-card ASIC to CP Chip and facilitates double-wide IO ports**
- **2nd gen Industry-first nest design that virtualizes chip-level 3 and drawer-level 4 caches from physical core-attached 36MB level 2 caches**
- 2nd gen Artificial Intelligence Unit (AIU) on-chip accelerator
- Branch prediction enhanced with additional PHT and capacity accelerator

System Cache Structure

- An IBM Z system consists of multiple computing nodes that are connected through the global fabric interface. Each system node includes a number of processor (CP) chips: 6 in z196 and zEC12, 3 in z13, 6 in z14, 4 in z15 and 8 (4 DCMs) in z16 and z17
  - In z10, z196, and zEC12, the system consists of up to four nodes, with each node fully interconnected to every other node through the level-4 (L4) caches. The L4 caches are managed in the System Controller (SC) chips.
  - In z13, the system consists of up to eight nodes, packaged as one pair of nodes per drawer
    - The nodes on each drawer are connected to each other through the L4 caches
    - Each node is connected to the corresponding node on each other drawer through the L4 caches
    - The three CP chips in each node are connected to each other through the shared on-chip level-3 (L3) caches
  - z14 consists of up to 4 nodes while z15 consists of up to 5 nodes, with further details provided in page 10 & 11
  - z16 and z17 bond a pair of CP chips into a dual chip module, or DCM, with 4 DCMs comprising a node and 4 total nodes in a system
- Each processor (CP) chip includes a number of processor cores
  - There are 4 cores in a z196 CP chip, 6 in zEC12, 8 in z13, 10 in z14, 12 in z15 and 8 cores-per-chip x 2 chips-per-DCM=16 in z16 and z17
  - Each core includes both local level-1 (L1) instruction and operand data caches, and a local level-2 (L2) cache
  - In zEC12 through z15, a pair of L2 caches supports instruction and operand data separately, and each L2 cache is connected to the on-chip (shared) L3 cache
  - In z16 and z17, the L2 cache is “unified” (i.e., no longer separate instruction and operand data caches), and each L2 cache is bidirectional-ring coupled to others on its chip to virtualize a L3 cache
- Prior to z16, all caches are managed “inclusively” such that contents in lower-level caches are contained (or tracked) in the higher-level caches using cache management algorithms derived from the MOESI (modified, owned, exclusive, shared, invalid) protocol with additional innovative cache states and features
  - Cache lines are managed in different states (simplistic view):
    - “exclusive” (at most 1 core can own the line to store or update at any time)
    - “shared” or "read-only" (can be read by 1 or more cores at any time)
    - “unowned” (where no core currently owns the cache line)
  - When a cache line is shared and a processor wants to store (update) one of the elements, a cache coherency delay is required to invalidate all existing read-only lines in other caches so this processor can be the exclusive owner
  - Similarly, this exclusive line will need to be invalidated before another processor can read or write to it
  - In z13, the L4 maintains a non-data inclusive coherency (NIC) directory to keep track of cache-line states in the L3 without having to save a copy of the actual cache-line data.
  - Starting with z16, data are brought directly into the local L2 non-inclusively and eventually age out to the shared L3 and L4
    - The L2 is still inclusive of the L1 on z16+, however the shared vL3 and vL4 caches are no longer inclusive of the L2 and L1

Near-Core Cache Operations

- The L1 cache on z196+ and L2 (private) cache on z196..z15 are store-through, i.e., each storage update is forwarded immediately to the shared L3 cache after the instruction performing the update completes;
  - L3 and L4 (shared) caches on z196+ are store-in, i.e., storage updates are kept in the cache until the cache entry is replaced by a new cache line or evicted to move to another L3 or L4 cache
  - On z16 and z17 the L2 (private) cache is also store-in
- The cache-line size (for all caches) being managed across the cache subsystem is currently 256 bytes
  - Although the cache-line size remains stable across recent machines, it should not be relied upon
  - However, it is unlikely that the cache-line size will grow beyond 256 bytes
  - EXTRACT CPU ATTRIBUTE (ECAG) instruction should be used to obtain information about the cache subsystem, e.g., cache sizes and cache-line sizes for each cache level
- The z/Architecture and the processor design supports self-modifying code
  - However, supporting self-modifying code can be costly due to movement of cache lines between the instruction and operand data caches (L1 and L2). More details are provided in “Optimization – Insn/Data Placement”
  - Due to out of order and deep pipelining, self-modifying code becomes even more expensive to use and is not advised
  - Even if there is no intention to update the program code, false sharing of program code and writeable operand data in the same cache line will suffer similar penalties; more details are provided in “Optimization – Shared Data”
- The L1 implements a “store-allocate” design where it must obtain the exclusive ownership before it can store into a cache line
  - The storing instruction will stall in the pipeline until the correct cache state is obtained
  - It is important not to share writeable operand data elements in the same cache line for independent multiprocessor operations
- The associativity of a cache (as specified in subsequent pages) reflects how many compartments are available for a particular cache line to be stored in
  - For an 8-way associative cache, a cache line (based on its line address) can be saved in one of 8 compartments

High-level updates about z14

- z14[14] improves upon the z13 SMT design with focus on special functions
  - Runs at a faster maximum frequency of 5.2 GHz; with a similar pipeline to z13
  - Maintains the tight integration of each L2 cache lookup with the corresponding L1 cache
  - Integrates the level-1 Translation-Lookaside-Buffer (TLB1) function into the L1 directory for both instruction and data cache accesses
  - Operates TLB2 lookup in parallel with L2 directory lookup pipeline to drastically reduce TLB miss penalties
    - TLB2 enhancements: 2x CRSTE (combined region segment table entry) and 1.25x PTE (page table entry) growth
  - Branch prediction improvements; 33% BTB1-and-2 growth, new perceptron predictor and simple call-return stack
  - Further improves handling of simultaneous multi-threading (SMT) for 2 threads, focusing on maximizing execution overlap within the pipeline, and parallelizing TLB and cache misses
    - To this end, there are now four HW-implemented translation engines on z14 vs one picocoded engine on z13
  - Doubles FP32/64 (single/double precision floating point) SIMD throughput
  - Adds amazing performance and functionality in the Co-Processors (COP) for compression and cryptography
- z14 introduces many notable z/Architecture features, including, but not limited to:
  - Guarded Storage Facility to enable pause-less garbage collection for Java
  - New compression modes to improve compression ratio and to provide order-preserving compression
  - New encryption modes including SHA3, AES-GCM
  - True (hardware) Random Number Generation support
  - New SIMD instructions, e.g., Binary-Coded Decimal (BCD) arithmetic, single & quad precision floating-point, long-multiply
- z14 system structure and cache topology
  - The processor subsystem consists of up to 4 nodes, with 1 node per drawer
    - Each node is connected to each other node through the SC chips
    - Each node consists of 2 clusters, with 3 CP chips per cluster
  - Each processor (CP) chip includes 1 L3 cache, which is shared by 10 processor cores through the L2/L1 caches similar to z13 design

High-level updates about z15

- z15[16] improves upon the z14 overall design with focus on power efficiency and performance improvements
  - Runs at the same maximum frequency of 5.2 GHz; with a similar core pipeline as z14 (or z13)
  - Utilizes the same 14nm technology, with bigger L2 to L4 caches on CP and SC chips, and with 2 more processor cores on each CP chip
  - New “TAGE” or TAgged GEometric history length-based PHT or pattern history table branch predictor design
    - can learn multiple patterns of varying lengths and outcomes for the same branches –
    - https://www.irisa.fr/caps/people/seznec/JILP-COTTAGE.pdf
  - Redesigns hexadecimal and binary floating-point pipelines with a one cycle shorter pipeline, 2x FP32 (single precision floating point) throughput and a new divide engine providing improvements in SIMD performance
  - Further streamlines store processing pipeline and fetch/store conflict handling in the L1 data cache design
  - Increases out-of-order window size vs z14, number of outstanding L1 cache misses, and TLB2 size for supporting 2G pages
- z15 introduces many notable z/Architecture features, including, but not limited to:
  - On-chip hardware accelerator that performs DEFLATE compliant compression and decompression, just like the Integrated Accelerator for zEnterprise Data Compression (zEDC), thus replacing the need of the zEDC Express adapter
  - Hardware-based Secure Execution, a security technology protecting workload from external and internal threats
  - A sort accelerator inside each core to perform a loser tree merge sort on data of any size
  - A modulo arithmetic (MA) unit inside each core to accelerate elliptic curve cryptography (ECC) implementing the modulo arithmetic operations behind it
  - New SIMD instructions to operate decimal data with built-in validation, and to provide faster string search
- z15 system structure and cache topology
  - The processor subsystem consists of up to 5 nodes (1 more than z14), with 1 node per drawer
    - Each node is connected to each other node through the L4 caches (SC chips)
    - Each node consists of 2 clusters, with 2 CP chips per cluster
  - Each processor (CP) chip includes 1 L3 cache, which is shared by 12 processor cores through the L2/L1 caches similar to the z14 design
  - Enhancements in L3/L4 MESI cache protocols to prevent L4 capacity evictions from intersecting with actively used L3 cache content

High-level updates about z16

- z16[18,19,20] (7nm) improves upon z15
  - Continues at 5.2GHz
  - Adds an on-chip AIU (artificial intelligence unit) HW accelerator that supports matrix operation (up to 8x8x2 per cycle and up to 32Kx32K per invocation) for common tensor transformations and mathematic operations
  - **Switch to industry standard 7nm technology drove dense SRAM based innovations**
    - Redesigned branch prediction logic

      Adds dynamic container or size-based branch history organization to improve utilization of real estate which yields better predictions

      Improves efficiency by not reindexing the lookup structure when continuing branch identification down an in-progress 128B half-cache line

      Now skips over lines with no known branches
    - Optimized support for new nest-based 32MB physical Level 2 caches paired 1:1 with cores

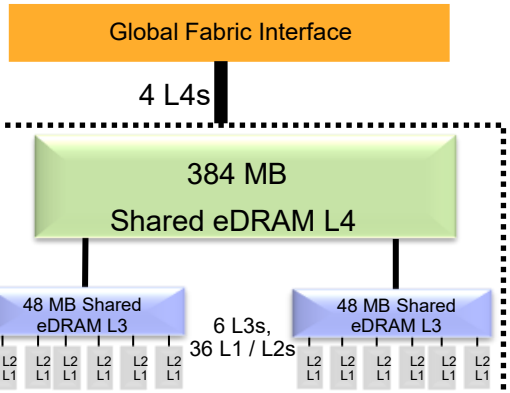
      The 8 L2s on a chip are accessed in a bi-directional ring and facilitate a virtual chip-level L3 cache

      The 8 chips on a drawer are point-to-point interconnected to virtualize a drawer-level L4 cache
- z16 system structure and cache topology
  - The processor subsystem consists of up to 4 nodes with 1 node (virtual L4) per drawer
    - Each node is connected to each other node through the direct CP chip connections (no SC chip!)
    - Each node consists of 4 clusters or dual chip modules or DCMs, that by definition contain 2 CP chips each
    - Each processor (CP) chip includes 1 virtual L3 cache, which is shared by 8 processor cores through the L2/L1 caches logically similar to previous traditional designs
- In addition to AIU and Vector-Packed-Decimal-Enhance Facility 2, z16 also adds privileged operations:
  - Cryptography Counter Set Support: Query Processor Activity Counter Information
  - Breaking Event Address Recording for debuggers that encounter “wild branches“
  - Reset DAT Protection to streamline storage management consistency or “system quiesce“ operations

High-level updates about z17

- z17 (5nm) improves upon z16[21,22,23,24,25]
  - Processor clocking frequency increased to 5.5GHz (tied with zEC12) up from z16's 5.2GHz
  - System-level focus on performance improvement for Artificial Intelligence (AI) workloads
    - Second generation on-chip AIU (artificial intelligence unit) HW accelerator; official performance claims:
      - IBM z17 demonstrates up to 48% reduction in latency for single threaded inference operations using IBM Integrated Accelerator for AI versus a similarly configured IBM z16[(reference forthcoming)].
      - On IBM z17, by allowing routing of inference requests to any idle IBM Integrated Accelerators for AI within the same drawer, the IBM Integrated Accelerator for AI can increase inference throughput by up to 7.5x as compared to IBM z16[24].
    - NNPA enhancements (see AIU slide later in this document)
    - z17 redesigned its principal IO-card ASIC as the Data Processing Unit ("DPU") and relocated onto the CP chip
      - Facilitates double-wide IO ports which save on frame (and power and floorspace) utilization that may be repurposed to AIU IO-expansion (Spyre) cards
  - Second generation SRAM-redesigned nest with larger shared L3 and L4 caches virtualized from larger physical L2 caches and with improved control logic
    - Overall topology similar to z16: 4 drawers of eight 8-core chips though each chip now has 10 L2s (up from 8)
  - New General Instructions
    - Added GPR based Count Leading Zeros and Count Trailing Zeros instruction
    - Added Parallel Bit Deposit and Extract instructions
  - Vector Enhancements
    - Adding 64 and 128bit multiply instructions
    - Added Vector Divide and Vector Remainder
    - Added 128bit support to most other vector arithmetic instructions
  - CPACF Enhancements
    - Adding new functions to perform XTS-mode AES more efficiently on small disk blocks
    - Adding functions to be able to perform an HMAC function (incl. "last lock" and support for new sizes)
    - Adding controls to improve the performance of short block SHA3/SHAKE (new algorithm) operations
  - *Secure Execution Enhancements*
    - *2G pages in secure (and non-secure) guests, 1M host pages, PIC 3D interpretation, Retrievable secrets*

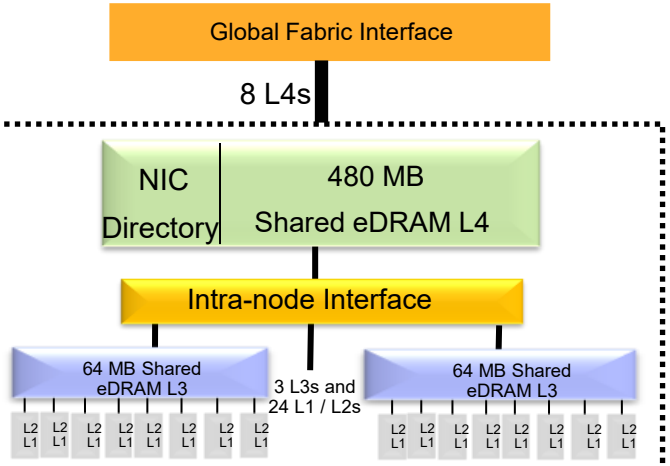
Cache Hierarchy and sizes (zEC12 and z13)



Store-through Parity protected	L1:	64 KB I-L1 + 96 KB D-L1 6-way Set Associative D-L1 4-way Set Associative I-L1 256B cache-line size
Store-through Parity protected	L2:	1 MB L1+* Inclusive of D-L1 1 MB L2 Inclusive of L1, L1+* 8-way* Set Associative 256B cache-line size
Store-in ECC protected	L3:	48 MB Inclusive of L1, L1+, L2 12-way Set Associative 256B cache-line size
Store-in ECC protected	L4:	384 MB Inclusive of L1, L1+, L2, L3 24-way Set Associative 256B cache-line size

zEC12 (book)

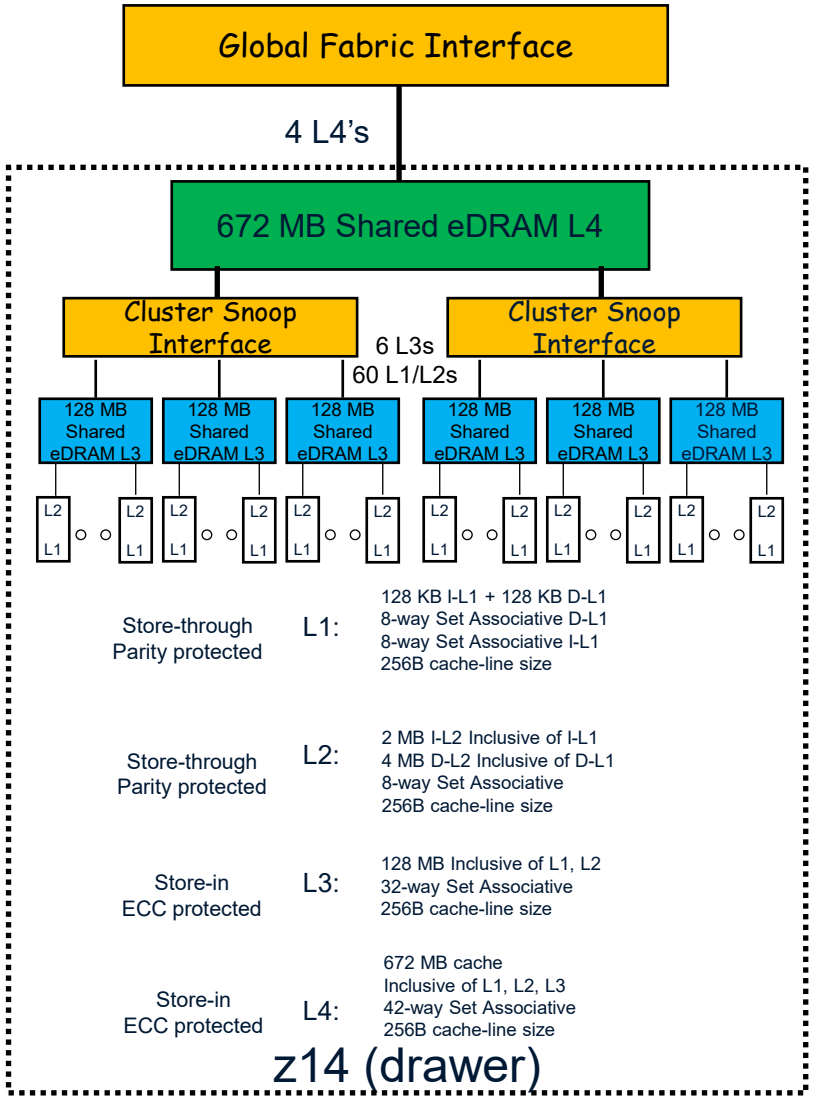
* The L1+/L2 design in zEC12 is too complicated for this document. One can treat it as two L2s, each 1MB and 8-way set associative

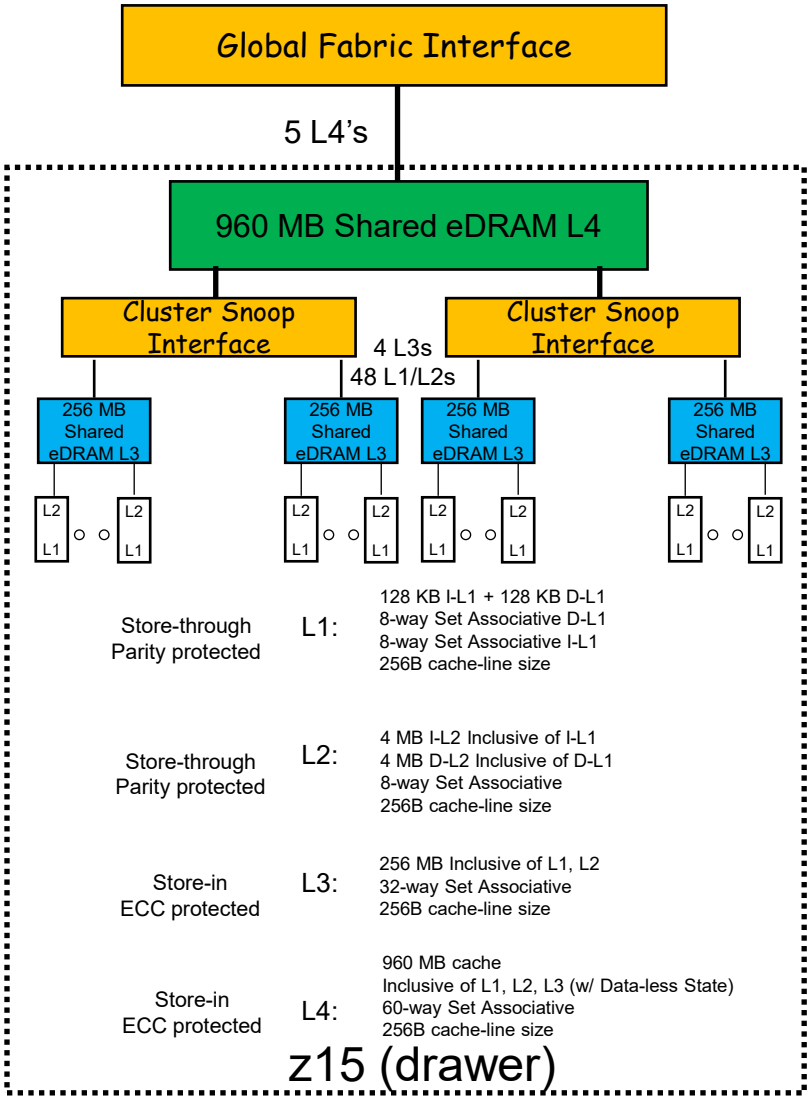


Store-through Parity protected	L1:	96 KB I-L1 + 128 KB D-L1 8-way Set Associative D-L1 6-way Set Associative I-L1 256B cache-line size
Store-through Parity protected	L2:	2 MB I-L2 Inclusive of I-L1 2 MB D-L2 Inclusive of D-L1 8-way Set Associative 256B cache-line size
Store-in ECC protected	L3:	64 MB Inclusive of L1, L2 16-way Set Associative 256B cache-line size
Store-in ECC protected	L4:	480 MB cache + 224 MB Non-data Inclusive Coherency (NIC) Directory Inclusive of L1, L2, L3 30-way Set Associative L4 14-way Set Associative NIC 256B cache-line size

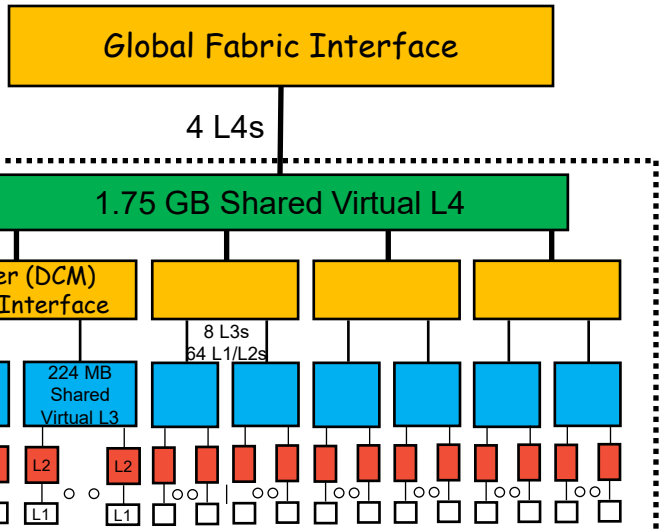
z13 (half drawer)

Cache Hierarchy and sizes (z14 and z15)



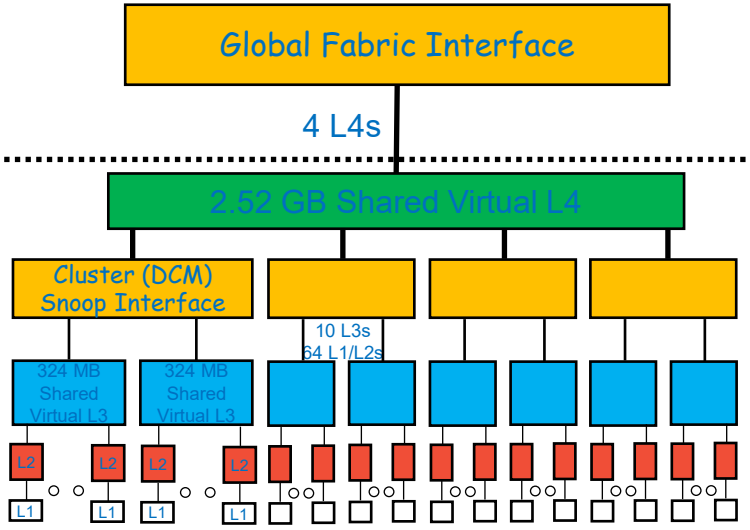


Cache Hierarchy and sizes (z16 and z17)



Store-through Parity protected	L1:	128 KB I-L1 + 128 KB D-L1 8-way Set Associative D-L1 8-way Set Associative I-L1 256B cache-line size
Store-in ECC+RAIM protected	L2:	32 MB unified I+D L2 Inclusive of I-and-D L1s 16-way Set Associative 256B cache-line size **SRAM**
Store-in ECC+RAIM protected	Virtual L3:	224 MB (7 L2s x 32 MB) **Not** inclusive of L1, L2 112w Set Associative (7 L2s x 16-way assoc) 256B cache-line size
Store-in ECC+RAIM protected	Virtual L4:	1.75 GB (7 chips x 8 L2s x 32 MB) **Not** inclusive of L1, L2, L3 896w Set Associative (7 chips x 8 L2s x 16-way) 256B cache-line size

z16 (drawer)



Store-through Parity protected	L1:	128 KB I-L1 + 128 KB D-L1 8-way Set Associative D-L1 8-way Set Associative I-L1 256B cache-line size
Store-in ECC+RAIM protected	L2:	36 MB unified I+D L2 Inclusive of I-and-D L1s 16-way Set Associative 256B cache-line size **SRAM**
Store-in ECC+RAIM protected	Virtual L3:	324 MB (9 L2s x 36 MB) **Not** inclusive of L1, L2 144w Set Associative (9 L2s x 16-way assoc) 256B cache-line size
Store-in ECC+RAIM protected	Virtual L4:	2.52 GB (7 chips x 10 L2s x 36 MB) **Not** inclusive of L1, L2, L3 1120w Set Associative (7 chips x 10 L2s x 16-way) 256B cache-line size

z17 (drawer)

High-Level understanding of the microprocessor core

- The z microprocessor cores can be simplified into a number of functional units (which are further described in some published papers):
  - Branch prediction unit
    - Two-level structure of branch histories; advanced design predicts both targets and directions
  - Instruction caching and fetching unit
    - Based on branch prediction information, delivers instructions in a seamless fashion
  - Instruction decoding and issuing unit
    - Decodes instructions in groups; issues micro-operations out-of-order to the execution units
  - Fixed-Point Execution unit
    - Executes most of the fixed-point operations, and (since z13) fixed-point divides
  - Vector & Floating-Point Unit
    - Handles floating-point arithmetic operations, complicated fixed-point and decimal operations, and (since z13) vector operations
  - Modulo Arithmetic Unit
    - An elliptic curve cryptographic accelerator that is an extension to the Vector & Floating-Point unit (since z15); operates through millicode routines
  - Load/Store (or Operand Data-caching) unit
    - Accesses operand data for both fetch (load) or store (update) operations
  - Co-processor unit
    - Supports data compression, cryptographic functions, UTF translations (since zEC12), sort acceleration (since z15); operates through millicode routines
  - Second-Level Translation and Cache unit
    - Maintains the private second-level translation-lookaside-buffer (TLB2) and cache (L2)
- We will give a high-level overview of the microprocessor design features
  - For more information, see articles that are listed in the reference section near the end

Branch Prediction Unit

- Branch prediction in Z processors is performed *asynchronously* to instruction processing
  - The branch prediction logic can find/locate/predict future occurrences of branch-type instructions (including calls and returns) and their corresponding directions (taken or not-taken) and targets (where to go next) on its own, without requiring or waiting for the downstream pipeline to decode or detect a branch instruction
  - The branch prediction logic tries its best to predict the program path much further ahead than where the instruction fetching unit is currently delivering instructions and thus should be way ahead of where the execution engines are executing

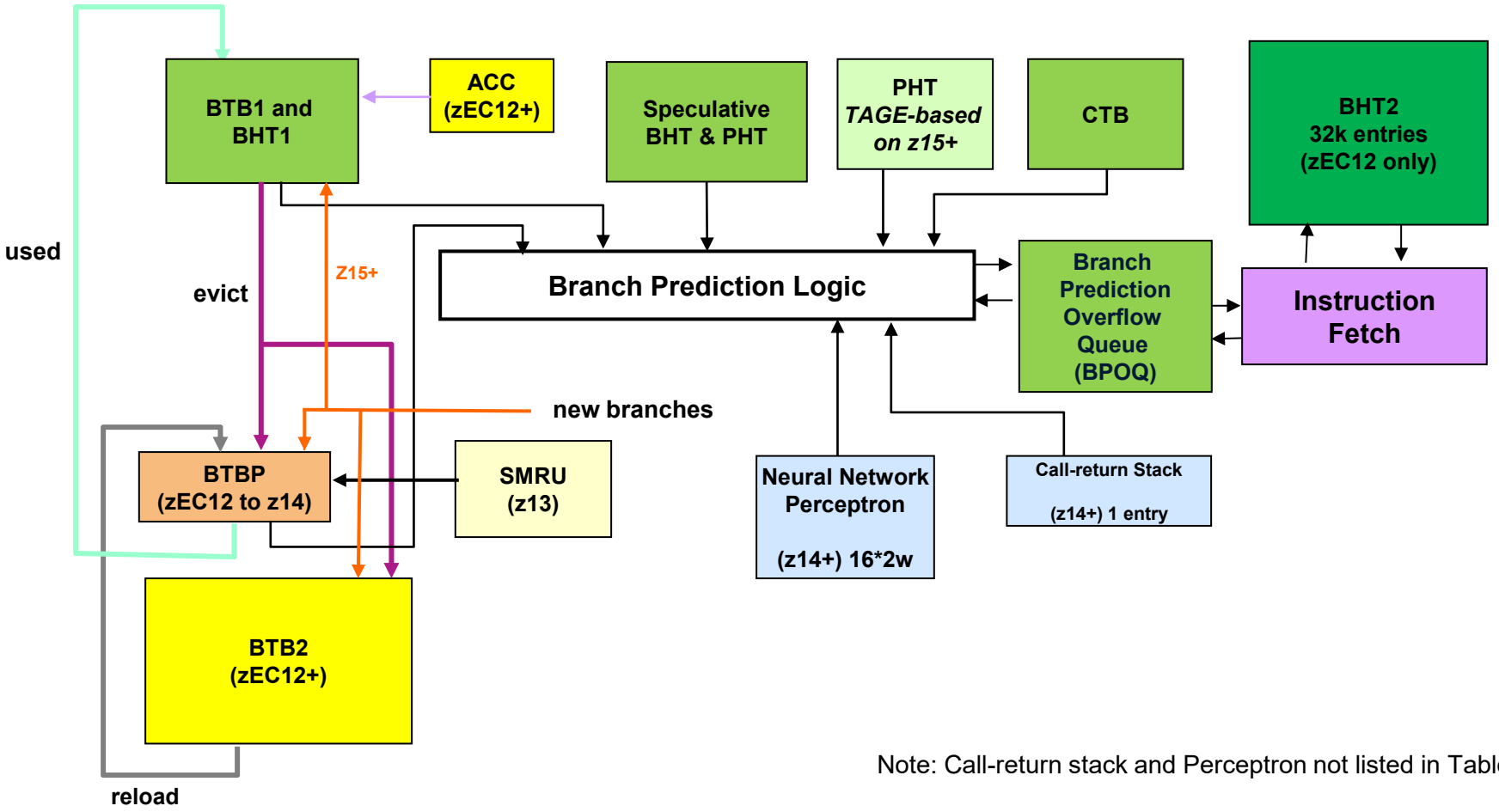
- The branch prediction logic employs many advanced algorithms and structures for predicting branching behaviors in program code as seen in the following “Branch Prediction Structure” figure, including:
  - First-level branch target buffer (BTB1) and branch (direction) history table (BHT1)
  - Second-level target and history buffers (BTB2 and BHT2) (introduced in zEC12) with a pre-buffer (BTBP) used as a transient buffer to filter out unnecessary histories. In z15 the BTB1/BHT1 is doubled in size and the BTBP structure is removed.
    - Note: the BHT2 is used separately in z196/zEC12 and is fully integrated into the BTB2 in z13+
  - Since zEC12, accelerators for improving prediction throughput (ACC) by “predicting the prediction” can make a taken prediction every other cycle for a limited subset of branches
  - Pattern-based direction and target predictors (PHT and CTB) anticipate how the program will progress based on a branch history pattern that represents “how the program got here”, e.g., for predicting an ending of a branch-on-count loop, or a subroutine return that has multiple callers. The PHT is improved to a design based on tagged geometric length predictor (TAGE) as of z15[17] as is the CTB starting on z16.
  - Starting in z14, both a neural-network-based perceptron engine for enhanced direction prediction and a simple call-return stack for target prediction are introduced for additional accuracy
    - *Only branch targets of > 512 byte-blocks away will be considered as a potential call-return pair*

Branch Prediction Unit

- The branch prediction logic communicates its prediction results to the instruction fetching logic through an overflow queue (BPOQ) such that the branch prediction pipeline will not be directly coupled to and be potentially throttled by instruction fetch operations (e.g., during an I-cache miss).

- z16+ branch prediction is inclusive of all z15 structures and is substantially redesigned for SRAM
  - Container-based organization optimizes real estate by conserving bits for short-and-medium distance branches rather than consuming full entries designed for long-distance branches
  - Improves efficiency by not reindexing the lookup structure when continuing branch identification down an in-progress 128B line
  - Now skips over lines with no known branches
  - z17 adds a (slower) “PHT2” alongside the existing (faster) “PHT” which is now known as the “PHT1”, and is now collectively 2048x2x2x2
    - That is 2048 rows x 2 tables (PHT1 and PHT2) indexed with different history lengths x 2 parent pairs x 2 sets.
  - z17 adds a 4k x 2 “capacity accelerator” that operates similarly to the existing “latency accelerator”.

Branch Prediction Structure



Note: Call-return stack and Perceptron not listed in Table 1.

Table 1: Branch Prediction Resources

Label	Structure Name	Description	z13	z14	z15	z16	z17
BTBP	Br Tgt Pre-buffer	0.5th level branch instruction address and target predictor. Look-up in parallel to BTB1, upon usage, transfer to BTB1	128 x 6	128 x 6	NA	NA	NA
BTB1	L1 Br Tgt Buffer	1st level branch instruction address and target predictor	1024 x 6	2048 x 4	2048 x 8	512 x 4 x 4..6 (8K..12K)	512 x 4 x 4..6 (8K..12K)
BHT1	L1 Br History Table	1st level direction predictor (2-bit) : weakly, strongly taken, or not-taken	1024 x 6	2048 x 4	2048 x 8	512 x 4 x 4..6 (8K..12K)	512 x 4 x 4..6 (8K..12K)
BTB2	L2 Br Tgt Buffer	2nd level branch instruction address and target history buffer	16K x 6	32K x 4	32K x 4	128K..256K	128K..256K
BHT2	L2 Br History Buffer	2nd level direction 1-bit predictor for branches not predicted ahead of time	NA	NA	NA	128K..256K	128K..256K
ACC	Col Pred (z13..z15) Latency Accelerator (z16+) Capacity Accelerator (z17)	Accelerate BTB1 throughput in finding the “next” branch	1024 (search-based)	1024 (stream-based)	1024 x 8	512 x 2	512 x 2 (latency acc) 4K x 2 (capacity acc)
SBHT/ PHT	Speculative BHT & PHT	Speculative direction prediction with transient updates at (out-of-order) resolution time prior to actual completion	8 + 8	8 + 8	8 + 8	NA	NA
PHT	Pattern History Table	Pattern-based tagged direction prediction “TAGE” or TAgged GEometric history length”-based on z15+	1024 x 6	2048 x 4	512 x 8 x 2 (short & long tables)	2048 x 2 x 2 (short & long tables)	2048 x 2 x 2 x 2 (short & long tables now dubbed “PHT1” plus a 2nd-level capacity-increasing “PHT2” set alongside it)
CTB	Changing Target Buffer	Pattern-based target prediction predicts branches with multiple targets, typically subroutine returns and branch tables	2048	2048	2048	1024 x 2 (short & long tables)	1024 x 2 (short & long tables)
SMRU	Super MRU table (z13+)	Protect certain branches from normal LRU out to make the BTBP more effective	128	128	NA (no BTBP)	NA (no BTBP)	NA (no BTBP)
CRS (+ RAT)	Call-Return Stack (plus Return Address Table on z16+)	simple call-return stack for target prediction	NA	1	1	16	16
PCP	Perceptron	neural-network-based perceptron engine for enhanced direction prediction	NA	16x2	16x2	32x3	32x3

Instruction Delivery

- Since z/Architecture instructions are of variable lengths (2, 4 or 6 bytes), an instruction can start at any halfword (integral 2-byte) granularity
- Instruction fetching logic fetches “chunks” of storage-aligned instruction data from the instruction cache, starting at a disruption point, e.g., after a taken branch (including subroutine calls and returns) or after a pipeline flush
  - Up to 2 16-byte chunks for z196 and zEC12 and up to 4 8-byte chunks for z13 and after
- These “chunks” of instruction data are then written into an instruction buffer (as a “clump”) where instructions are extracted (or parsed) into individual z-instructions in program order
- The instruction decode logic then figures out high-level characteristics of the instructions and which/how the execution engines will handle them
  - Is it a storage access? A fixed-point instruction? Which execution units will be involved?
  - Is it a branch-type instruction? If yes, did the branch prediction logic predict that? If not, notify the branch prediction logic (to restart its search) and then proceed based on predefined static prediction rules (e.g., branch-on-conditions are default to be not-taken, while branch-on-counts are defaulted to be taken)
  - Is it going to be implemented in millicode? If yes, did the branch prediction logic predict that? If not, reset the front-end to start at the corresponding millicode routine entry instruction
  - For a complex instruction, does it need to be “cracked” or “expanded” into simpler internal instructions, called micro-operations (µops)? For example, a LOAD MULTIPLE instruction will be expanded into multiple “load” µops that fetch from storage and write individual general registers (GRs)
- Instructions (and µops) are then bundled to form an instruction group (for pipeline management efficiency), and dispatched (written) into the instruction issue queue

Instruction Cracking, Expansion and Dual-Issuing

IBM Z Architecture is (arguably super-) complex instruction set computing (“CISC”), yet our high frequency pipelines innately support only reduced instruction set computing (“RISC”)

- Hence complex IBM Z instructions must be:
  - **cracked** into RISC micro-operations or “µops” comprising definitely-at-most a single group or triplet, OR
  - **expanded** into µops comprising one or possibly more groups or triplets, AND/OR
  - **dual-issued**, where a single instruction or µop is issued to two execution units, typically the LSU and an arithmetic unit

There are multiple ways we can go about this CISC -> RISC conversion

- Always crack/expand (due to inherent multiple operations needed), e.g.

```
– BRANCH ON COUNT (BCTR)     ------------> add register with immediate value of -1
–                                |                              |------>  scratch condition code
–                                |                                                |
–                                |-----> branch evaluation <----------------------|
```

- Length based cracking/expansion (multiple operations based on length), e.g.

```
– 8-byte MOVE characters (MVC)  ------------------>  load into scratch register
–                                  |------------>  store from scratch register

– 16-byte LOAD MULTIPLE (LM) ---------->  load into register 1
–                             |------>  load into register 2(displacement adjusted at dispatch)
–                             |------>  load into register 3(displacement adjusted at dispatch)
–                             |------>  load into register 4(displacement adjusted at dispatch)
```

- Typical **register-storage** (“RX”) instructions and µops are dual-issued like the “ADD” in the example below
  - ADD: Register1 <= Register1 + memory((Base register) + (Index register) + Displacement)

```
– Register-storage ADD (A) ---------------> load from storage into target register
–                           |                  .. Some cache access cycles later
–                           |-----------> add R1 with target register
```

  - The instruction is **not** considered to be cracked because it is tracked as 1 instruction by using 1 issue queue entry (and 1 global completion table entry), though it is issued to both the LSU and a non-LSU execution unit - hence it's 'dual issued'

Instruction Grouping

- Instructions (and µops) are dispatched (or written) in-order into the out-of-order issue queue as a group. They are then tracked in the global completion table (GCT) until every instruction in the group finishes its processing. When all instructions in a group finish processing, the group is completed and retired
- As instructions (and µops) are grouped, they are subject to various grouping rules, which prevent certain instructions (and µops) from being grouped with others
- During a dispatch cycle, z196 and zEC12 support one group of up to 3 instructions (i.e., “a triplet”), while z13+ allows two groups of up to 3 instructions (or two triplets)
- Some basic rules of grouping
  - Simple instructions, including most “register-register” (“RR”) and “register-storage” (“RX”) type instructions, can be grouped
  - Branch instructions, if second in the group, or if predicted taken, will be the last instruction in the group
    - Best group size if taken branches are the third in a group
  - µops that are cracked/expanded from the same instruction will usually be grouped with others from the same crack/expansion
    - But not with other instructions (or µops) in z196, zEC12
    - Certain zInsns that crack/expand into only 2 µops may be grouped with one other simple instruction on z13+
  - Storage-storage instructions are usually grouped alone, except for the µops that they may be expanded into
  - Other instructions that are alone in a group:
    - Register-pair writers, e.g., DIVIDE (D, DR, DL, DLR), MULTIPLY (M, MR)
    - Non-branch condition code readers, e.g., ADD LOGICAL WITH CARRY (ALC*), SUBTRACT LOGICAL WITH BORROW (SLB*)
    - Explicit floating-point control register readers or writers
    - Instructions with multiple storage operands
    - EXECUTE or EXECUTE RELATIVE instruction or its target
  - Since z13, max group size will be 2 if any µop has more than 3 register sources (including Access Register usage in AR mode)

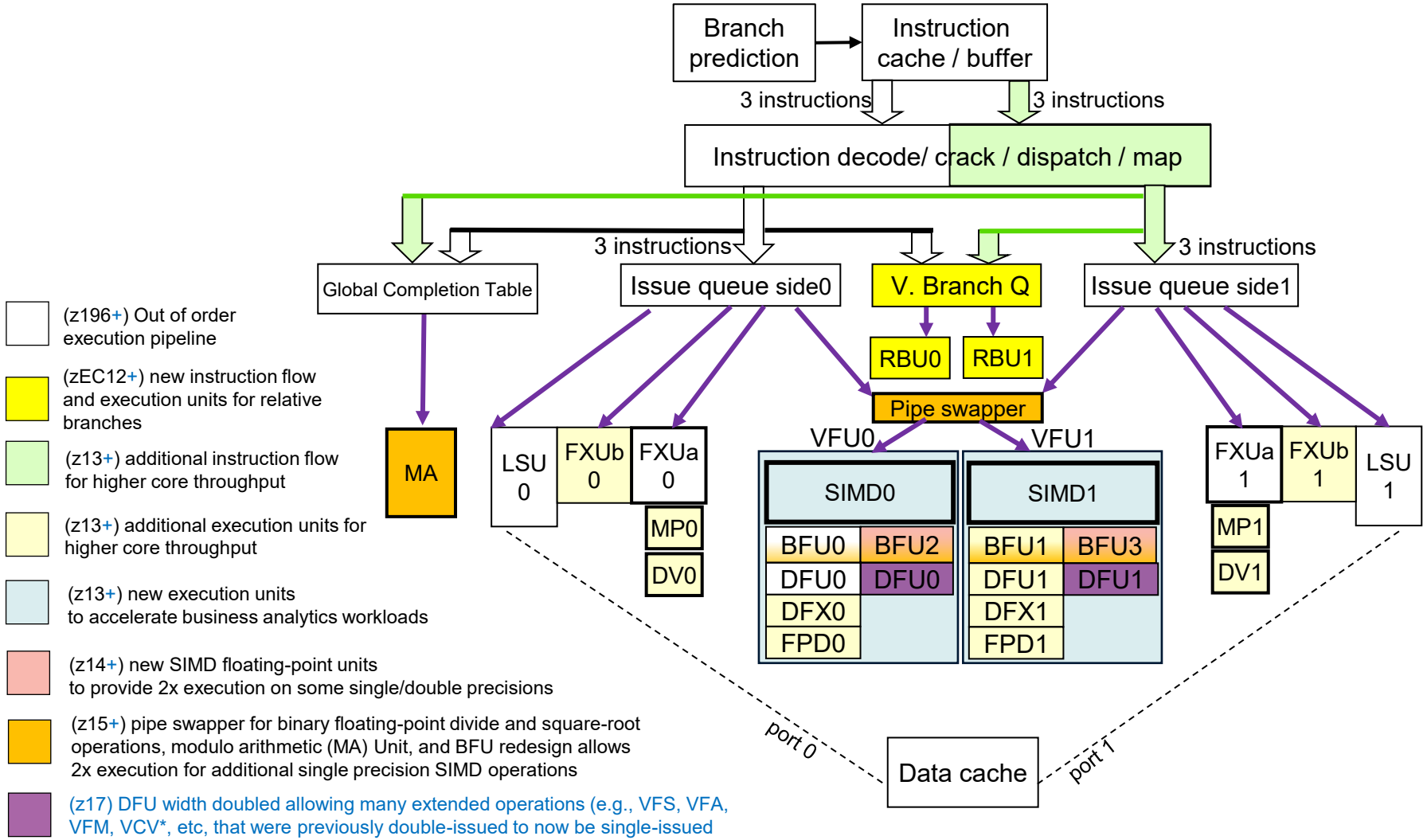
Instruction Dispatching

- As instructions are dispatched, the source and target architected registers are renamed into a virtual pool of physical registers and are tracked accordingly
  - The amount of rename tracking resources (how many in-flight mappings can be tracked) and physical registers available are key factors of the effectiveness of an out-of-order design
  - In z196 and zEC12, the mapping tracker (the mapper) consists of 1 bucket of 48 mappings
    - GRs: 1 mapping per each 32-bit register write, 1 mapping per each full 64-bit register write
    - FPRs: 1 mapping per each 32-bit register write, 1 mapping per each full 64-bit register write
    - ARs: 1 mapping per each 32-bit register write
  - In z13 and z14, the mapping tracker consists of 2 buckets of 64 mappings each = 128 total mappings
    - GRs: 1 mapping per each 32-bit register write, the GR #'s LSB decides which bucket to use; a 64-bit register write will require 2 mappings, one from each bucket
    - FPRs: 1 mapping per each write, the FPR #'s 2nd LSB decides which bucket to use
    - ARs: 1 mapping per each write, the AR #'s LSB decides which bucket to use
  - Since z13, multiple writes to the same register in the same group does not require separate trackers
  - Since z15, the mapping tracker consists of 4 buckets of 60 mappings each = 240 total mappings, where each bucket is defined by the second and third least significant bits of the register number for all register types. 64-bit write only require one entry.
- Instructions in a group are dispatched into one of the two issue queues (side 0 and side 1).
  - The total size of issue queue directly relates to the overall out-of-order window and thus affects performance
  - In z196 and EC12, only one instruction group can be written into one of the two queue sides in any cycle; in an alternating fashion
  - Since z13, two groups can be written in any cycle with one group into each side; with the older group on side 0
- The issue queue includes a dedicated "virtual branch queue" since zEC12, 1 per side, that handles relative branch instructions whose targets are less than 64 Kilobytes away
  - These branches will alternate to the different sides of the virtual branch queue independently of the other instructions in the group

Instruction Issue and Execution

- After instructions are dispatched into the issue queues, the issue queues will issue the oldest (and ready) instruction from each issue port to the corresponding execution engine
- Each issue-queue side is connected to a number of specific processing engines, using z15 as an example in Fig. 4,
  - There are 5 issue ports (per side; 10 total per core); each to a different engine, including
    - A relative branch unit (**RBU**) handles relative branches
    - A GR writing fixed-point unit (**FXUa**) handles most of the fixed-point arithmetic and logical operations, and also includes a multiply engine (**MP**) and a divide engine (**DV**) (both being non-blocking)
    - A non-GR writing fixed-point unit (**FXUb**) handles other fixed-point operations that do not write any GR results
    - A load/store unit (**LSU**) port, with accesses to the operand data-cache, handles memory accesses
    - A vector & floating-point unit (**VFU**) handles complicated operations
  - Inside each of the **VFU**, there are multiple engines that execute different functions in parallel to each other
    - **BFU** that handles both hexadecimal and binary (IEEE standard) floating-point arithmetic operations, and vector floating-point operations
    - **DFU** that handles decimal (IEEE standard) and quad-precision floating-point arithmetic operations; and since z14, BCD vector convert, multiply, and divide operations
    - **SIMD** that further consists of multiple subunits: PM engine that performs vector permute functions; XS engine that performs fixed-point arithmetic and logical functions; XM engine that performs several multiply functions and ST engine that performs string-related functions
    - **DFX** that handles decimal (BCD) fixed-point arithmetic operations, and since z14, simple BCD vector operations
    - **FPD** that handles divide and square root operations for both binary and hexadecimal floating-point arithmetic
  - Typical pipeline delays through each of the execution engines are shown in Fig. 5
- Generational differences are shown as colored boxes in Fig. 4
- Starting with z15, a modulo arithmetic (MA) unit can execute micro-operations sent from the global completion table (instead of the issue queues) to accelerate elliptic curve cryptography (ECC) as directed by millicode.

z196+ high-level instruction & execution flow



z14+ Execution Engine Pipelines

Only 1 of 2 issue sides shown

- **Typical pipeline depths and bypass capabilities shown**
- **Some instructions may take longer to execute or bypass results**
- **Access registers not shown**

ACC – GR access
WB – GR write back

V-ACC – FPR/VR access
VWB – FPR/VR write back

CC – condition code calculation

BYP – data bypass network cycle

FPD, DFU – functions, e.g., divide, square-root, may take multiple passes through the pipeline

G2F – GR to VR/FPR moves
F2G – VR/FPR to GR moves

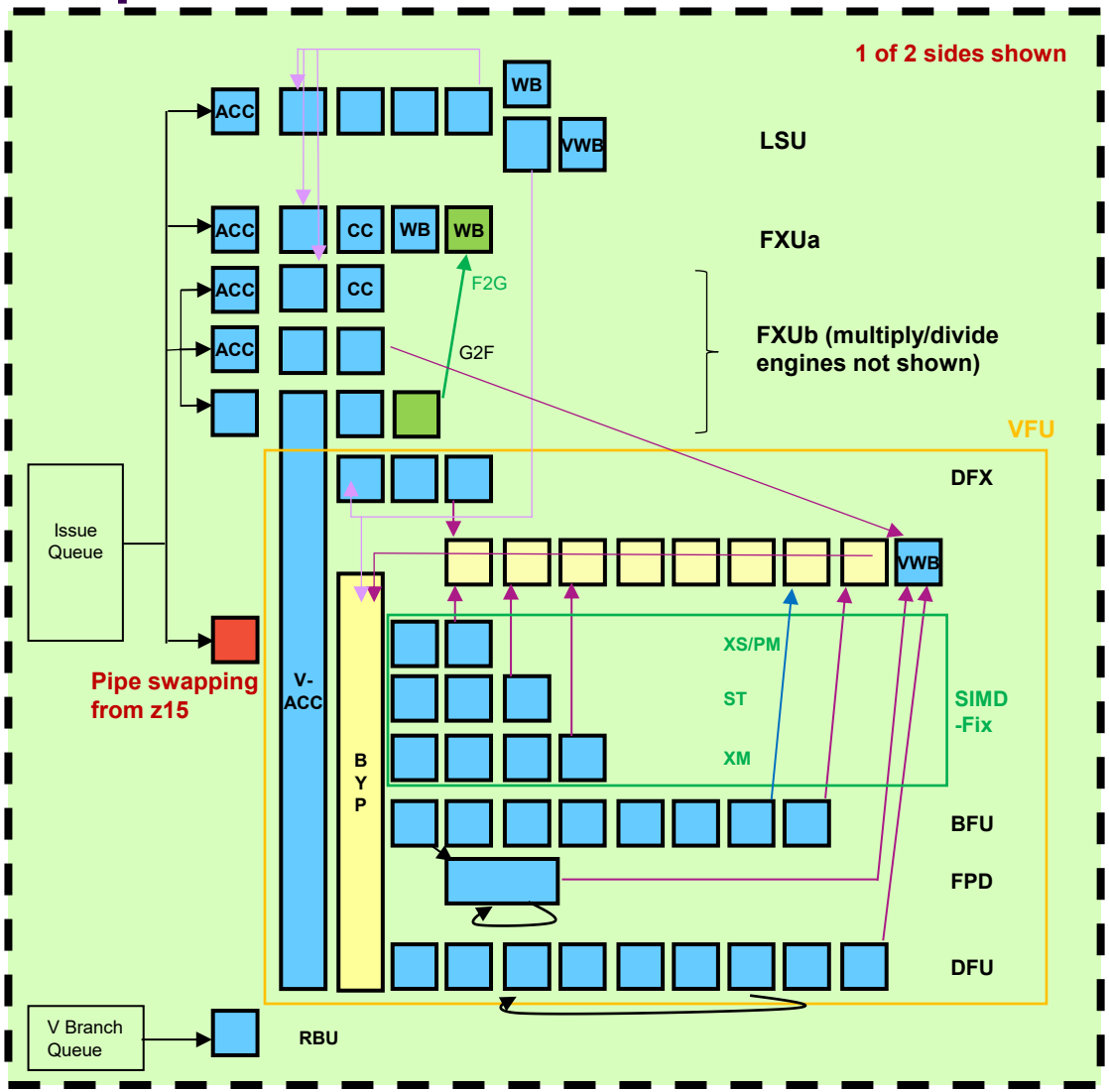


Table 2: Out of order resources

	z196	zEC12	z13	z14	z15 and z16	z17
GR	80 (16 permanently reserved for millicode)	80 (up to 16 reserved for millicode) + 16 immediate value entries	120 (up to 16 reserved per thread while in millicode) + 8 immediate value entries	Same as z13	Same as z13	Same as z13
FPR / VR(z13+)	48	64	127 (up to 8 reserved per thread while in millicode) + a zero-value entry	Same as z13	Same as z13	159 (up to 8 reserved per thread while in millicode) + a zero-value entry
AR (access register)	56 (16 permanently reserved for millicode)	56 (16 permanently reserved for millicode)	96 (up to 8 reserved for each thread while in millicode)	Same as z13	Same as z13	Same as z13
Issue Queue (z196+) plus Relative Branch Queue (zEC12+)	ISQ: 20 x 2 sides	ISQ: 20 x 2 sides + RBQ: 12 x 2 sides	ISQ: 30 x 2 sides + RBQ: 14 x 2 sides	ISQ: 30 x 2 sides+ RBQ: 16 x 2 sides	ISQ: 36 x 2 sides + RBQ: 16 x 2 sides	Same as z15
Global Completion Table	24 x 3 instructions (complete up to 3 instructions/cycle)	30 x 3 instructions (complete up to 3 instructions/cycle)	24 x 2 x 3 instructions (complete up to 6 instructions / cycle)	Same as z13	30 x 2 x 3 instructions (complete up to 6 instructions / cycle)	Same as z15
Unified Mapping Trackers	48	48	64 + 64	Same as z13	4 x 60 (new design)	Same as z15

The Load / Store Unit

- The Load / Store unit (LSU) handles the operand data accesses with its L1 data-cache and, prior to z16, its tightly coupled L2 data-cache
- The L1 data cache has 2 access ports, and each port can support an operand access of data elements of up to 8 bytes** a cycle
  - There is no performance penalty on alignment except for when the element crosses a cache line boundary
  - Prior to z15, vector elements of more than 8 bytes are accessed in two successive cycles
  - Starting with z15, vector loads using doubleword or **quadword alignment hints can be accessed in one cycle
- In addition to the prefetching of cache misses as part of the natural behavior of the out-of-order pipeline
  - LSU supports software prefetching through PREFETCH DATA type instructions
  - LSU also includes a stride-prefetching engine that prefetches +1, +2 cache-line strides, when a consistent stride is detected between cache miss address patterns at the **same** instruction address across loop iterations
- To minimize pipeline bubbles typically caused by “store-load” dependencies through storage, LSU provides a sophisticated bypass network to bypass pending storage updates that are not yet available in the L1 cache into dependent loads as if the operand data was in L1 (subject to certain limitations). But in general,
  - Data should be bypass-able even if bytes are required from different storing instructions for a load request
  - Data should be bypass-able if the store data is ready a small number of cycles before the dependent load request
  - Multiple mechanisms are used to predict dependencies (based on prior pipeline processing history) between load and store instructions, and will stall load instructions just long enough to enable “perfectly” timed bypasses
  - If a store operation is performed after its dependent load (due to out-of-order operations), a flush occurs
  - If a store operation is performed before its dependent load and the data is not bypass-able (due to timing or hardware limitations), the load is rejected and retried
  - More discussion follows – search for “store forwarding”

On-chip Core Co-Processor

- On-chip core co-processors (COPs) are available to enable hardware acceleration of data compression, cryptography, and, on zEC12 and after, Unicode conversions
  - Each COP is private to each core since zEC12, but is shared by two cores in z10 and z196
- The co-processor also handles COMPRESSION CALL (CMPSC) instruction that compresses data and cryptographic functions (under the CPACF facility, next page) that support latest NIST standards
  - In addition, Unicode UTF8<>UTF16 conversions are supported in zEC12; and since z13, all Unicode conversions (UTF 8<>16<>32) are supported
- Starting with z15, the co-processor provides support for SORT LISTS (SORTL) instruction that can turn up to 128 lists of unsorted input data into one or more lists of sorted output data. It also provides a means to merge multiple lists of sorted-input data into a single list of sorted-output data.
- Co-processors are driven through commands of millicode, as it emulates the corresponding complex z instruction
  - Millicode interprets the instruction, tests storage areas, and sets up the co-processor
  - Millicode fetches the source operand
  - Millicode writes source operand data into the co-processor to be processed
  - Millicode sets up result storage areas for co-processor to use - often it is the actual target areas
  - Coprocessor works on the instruction with the provided source data and generates output data
    - For CMPSC, the coprocessor will also fetch dictionary tables accordingly
  - Co-processor writes into the pre-set result storage areas
    - In some cases, millicode will transfer the Co-processor results to the target areas
  - Millicode analyzes status information from the co-processor and repeats work if needed
  - Millicode ends when the instruction (or a unit-of-operation) is completed
- In SMT mode (since z13), the co-processor handles one thread at a time. If the second thread requires the COP, it waits until the first thread finishes an appropriate unit-of-operation or the whole instruction

CPACF - CP Assist for Cryptographic Functions

- Also known as the Message-Security Assist (MSA) instructions
- Runs synchronously as part of the program on the processor
- Provides a set of symmetric cryptographic and hash functions for:
  - Data privacy and confidentiality
  - Data integrity
  - Random Number generation
  - Message Authentication
- Enhances the encryption/decryption performance of clear-key operations for
  - SSL/TLS transactions
  - Virtual Private Network (VPN)-encrypted data transfers
  - Data storing applications
- Since z15, (not shown in table to the right), an extension 9 is provided to support for elliptic curve cryptographic authentication of messages, the generation of elliptic curve keys, and scalar multiplication.
  - new instruction COMPUTE DIGITAL SIGNATURE AUTHENTICATION (KDSA) supports the ECDSA and EdDSA algorithms using curves P-256, P-384,P-521, Ed25519, and Ed448
  - compliant with the Digital Signature Standard (DSS), National Institute of Standards and Technology (NIST) July 2013
  - Existing PERFORM CRYPTOGRAPHIC COMPUTATION instructions (PCC and PCKMO) are also modified

Supported Algorithms	Clear Key	Protected Key
DES, T-DES	Y	Y
AES128	Y	Y
AES192	Y	Y
AES256	Y	Y
AES-GCM(z14)	Y	Y
GHASH	Y	N/A
SHA-1	Y	N/A
SHA-256	Y	N/A
SHA-384	Y	N/A
SHA-512	Y	N/A
SHA-3 224 (z14)	Y	N/A
SHA-3 256 (z14)	Y	N/A
SHA-3 384 (z14)	Y	N/A
SHA-3 512 (z14)	Y	N/A
PRNG	Y	N/A
DRNG	Y	N/A
TRNG (z14)	Y	N/A

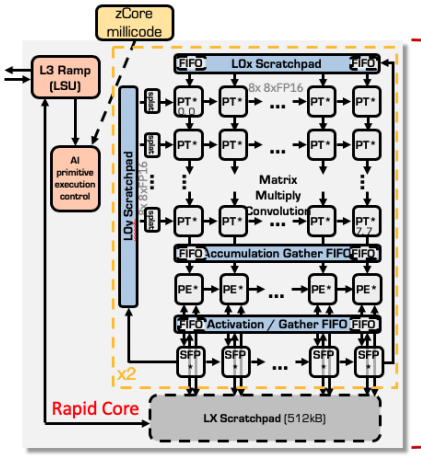
On-chip Integrated Deflate Accelerator

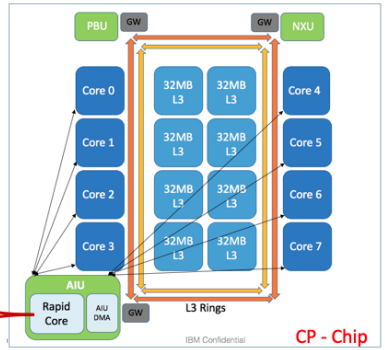
- On z15, with the Integrated Accelerator for zEnterprise Data Compression, the industry standard compression offered through the zEnterprise Data Compression (zEDC) Express PCIe I/O card is now integrated at the processor chip level
  - Data compressed with the zEDC Express adapter can be read and decompressed with the new Integrated Accelerator for zEDC on z15 and vice versa
- A single compression acceleration engine, called the Nest Accelerator Unit (NXU), is provided on each CP chip
  - The NXU is directly connected to the shared on-chip L3 cache and operate in tandem w/ the processor core (thread) that is currently running the new DEFLATE CONVERSION CALL (DFLTCC) instruction
- DFLTCC is interpreted by the millicode and executed in NXU
  - Millicode manages the locking mechanism which enables a single core to use NXU at a given time
  - Millicode divides bigger data into 256KB chunks for timesharing and combines the result
  - Millicode tests source data area, history area, target data area and creates a data structure with all the control information
  - Millicode enqueues the source and target address to the accelerator and sends a start operation signal
  - NXU fetches the control block, history data, source data and compress/decompress the data and writes the result to target data area
  - NXU also provides the status and writes back the resulted control block
  - Millicode analyses the status and control block and repeats/continue/finishes the instruction, report the result to software

On-chip Integrated Artificial Intelligence Accelerator[18,19,20,21,22,25]

z16 adds Neural-Network-Processing Assist Facility

- Uses CP-chip level AI hardware accelerator unit (“AIU”) to assist neural network processing for:
  - Training : train a model with input data
  - Inferencing : input data to a trained model to infer/predict an outcome
- Introduces Deep Learning data Format (DLF) that represents a 16-bit signed floating-point number in a proprietary format with a range and precision tailored toward neural-network processing that may be DMA-streamed into the AIU’s “rapid core” 8x8x2-per-cycle HW accelerator
- New DLF to/from BFP short/tiny conversion vector-unit executed instructions with operands in VRs
- **NNPA instruction**
  - Multiple Function Codes provide various mathematical operations
    - performed on a "tensor" : an array of elements with each element having DLF format
  - Millicode does the initialization and points the AIU to the tensors and lets it rip!
  - z17 Enhancements:
    - HW accelerated tensor formatting
    - New HW functions to support foundation model primitives and a wider variety of ML models
    - Support for quantized matrix multiplication (INT8 x DLF16 + DLF16 and INT8 x INT8 + DLFT16)
    - Transparent NNPA load balancing across all 8 AIU in a drawer
    - Support for tensor transformation to/from the AIU tensor format (data padding and format)





Function group	AI Function Name
Tensor manipulation	NNPA_TRANSFORM
	NNPA_RESHAPE
	NNPA_CONCAT_1
	NNPA_REDUCE_1
Elementwise ops	NNPA_EL_ADD
	NNPA_EL_SUB
	NNPA_EL_MUL
	NNPA_EL_DIV
	NNPA_EL_MIN
	NNPA_EL_MAX
	NNPA_EL_SQRT
	NNPA_EL_INVSQRT
Activation ops	NNPA_LOG
	NNPA_EXP
	NNPA_RELU w/ clipping and leaky
	NNPA_TANH
	NNPA_SIGMOID
	NNPA_GELU
Norm ops	NNPA_SOFTMAX w/ masking
	NNPA_BATCHNORM
	NNPA_L2NORM
	NNPA_LAYERNORM
Pooling	NNPA_AVGPOOL2D
	NNPA_MAXPOOL2D
Systolic ops	NNPA_CONVOLUTION_ACT
	NNPA_MATMUL_OP w/ transpose, w/ INT8-quantize
	NNPA_MATMUL_OP_BCAST w/ transpose, w/ INT8-quantize
RNN	NNPA_LSTMACT
	NNPA_GRUACT
Utility	NNPA_QAF

Availability color coding:
z16 and z17
z17 only

Instructions of Interest

- We will discuss some of the instructions in z/Architecture and their handling that might be of general interest:
  - Simple instructions, including descriptions of some interesting ones
  - Special Storage-to-Storage instructions
  - MOVE LONG instructions
  - High Word instructions
  - Conditional instructions
  - EXECUTE instructions
  - BRANCH PREDICTION PRELOAD instructions
  - DATA PREFETCH instructions
  - NEXT INSTRUCTION ACCESS INTENT instruction
  - Atomic and locking instructions
- And a few architecture features:
  - *Hardware Transactional Execution – support being sunset starting with z17*
  - Guarded Storage Handling
  - Secure Execution
  - Vector (SIMD) instructions
  - BCD Vector Instructions
  - Neural Network Processing Assist (NNPA)
- And some storage usage model highlights

Simple Instructions

- Simple instructions
  - Fixed-point results are bypassed without delay into the next dependent fixed-point instruction if the instructions are in the **same side** of the issue queue; otherwise, there will be at least a one-cycle delay
  - An instruction with a storage access operand will need to wait for 4 cycles if the operand is a hit in L1 data cache
  - An operand written by a store instruction to a storage address followed by a load instruction of the same address will require at least 2 to 4 cycles to be bypassed as L1 cache hit data
  - Floating-point instructions are generally pipelined but can be of different latencies. The design forwards dependent data as soon as it is available
  - Non-floating-point vector (SIMD) instructions (since z13) have shorter latencies than floating-point ones
    - SIMD results are also bypassed when available
- Destructive and non-destructive instructions
  - Many z/Architecture instructions specify just two operands, with one operand doubling as both a source and a target
    - These instructions are shorter (in length) and occupy less space in storage
    - If a register-based operand that will be overwritten is still required after an instruction execution, software must first make a copy of the register before the overwriting instruction
  - Many non-destructive instructions were introduced since z196, such that the register copy operations can be avoided
- Load and Store Reversed instructions
  - To facilitate conversion between big-endian (BE) and little-endian (LE) formats, a few instructions (LOAD REVERSED and STORE REVERSED) are provided to reverse the byte ordering of a data element to/from memory
  - Both load and store operations are supported
  - 2, 4, and 8-byte operands are supported
  - MOVE INVERSE is also available for more than 8-bytes storage to storage data swap
    - Millicode implements this instruction by doing a byte-by-byte copy
  - Starting with z15, vector enhancements facility 2 includes new instructions (VECTOR LOAD/STORE BYTE/ELEMENTS REVERSED ELEMENT<S> <AND ZERO/REPLICATE>) to load or store elements or arrays of elements in the little-endian format through the vector registers

Special Storage-to-Storage Instructions

- z/Architecture includes a set of storage-storage instructions in which the data size is specified in the instruction as the length field
  - Mostly defined to be left-to-right and byte-at-a-time operations
  - Special hardware is being used to speed up certain common cases
- MOVE Characters (MVC)
  - If <= 16 bytes, it is cracked into separate load and store µops
  - If > 16 bytes, it is handled by sequencing logic inside the LSU
  - If the destination address is 1 byte higher than the source address (and they overlap), it is special cased into hardware as a 1-byte storage-padding function (with faster handling)
  - If the destination address is 8 bytes higher than the source address (and they overlap), it is special cased into hardware as an 8-byte storage-padding function (with faster handling)
  - For other kinds of address overlap, it will be forced into millicode to be handled a byte at a time
  - Since z10, special case detection is done at decode time, not detected during address generation, and thus requires the instructions to have B1=B2

– COMPARE LOGICAL Characters (CLC)
  - If <= 8 bytes, it is cracked into separate load and compare µops
  - If > 8 bytes, it is handled by the sequencing logic inside the LSU

- EXCLUSIVE OR Characters (XC)
  - If <= 8 bytes, it is cracked into separate load and “or-and-store” µops
  - If > 8 bytes and if base register values and displacement values are equal, i.e., an exact overlap on addresses, it is special cased into hardware as a storage clearing function (with faster handling)
    - Since z10, this special case detection is done at decode time, not detected during address generation, and thus requires the instructions to have B1=B2 and D1=D2
  - If > 8 bytes and no exact overlap on addresses is detected, it is handled by sequencing logic inside the LSU
  - For other kinds of address overlap, it will be forced into millicode to be handled a byte at a time
  - AND Characters (NC) and OR Characters (OC) instructions are implemented similarly, without the special clearing function

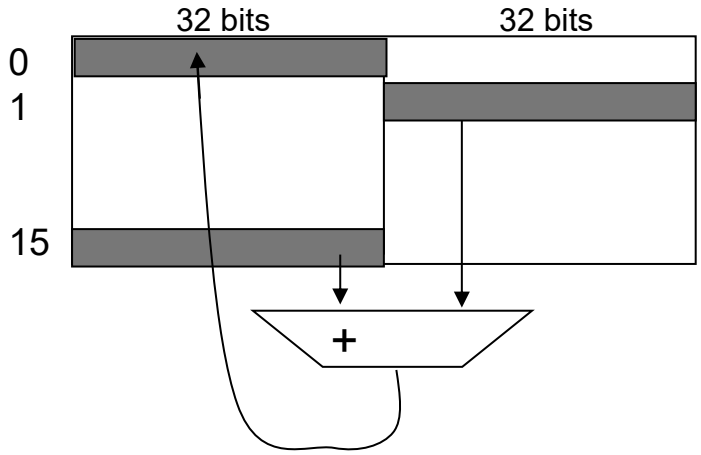
MOVE LONG Instructions (MVCL*)

- MOVE LONG instructions can copy a large amount of data from one storage location to another
- A special architected functional variant can also be used to pad storage
- These instructions are implemented in millicode
- A special engine is built per CP chip for aligned copying or padding functions at a page granularity
  - The page-aligned copying or padding is done “near memory”, instead of through caches, if
    - Not executed inside a transaction
    - Padding character specified is neither X’B1’ nor X’B8’
    - A preceding NIAI instruction does not indicate that the storage data will be used subsequently
    - The operands must not have an access exception
    - Length >= 4K bytes
    - For moves, source and destination addresses are both 4K-byte aligned
    - For padding, destination address is 4K-byte aligned
  - Otherwise, the move process will operate through the caches (L1, L2…)
  - Note that the evaluation is revised every unit-of-op
    - For padding, even if starting address is not aligned, millicode pads in cache to the first 4K-byte boundary, then uses “near memory” pad engine for the next aligned 4K-byte pages until the remaining length is less than 4K bytes. After that, padding is done in cache again
- Near-Memory engine usage is best when the amount of data involved is large and the target memory is not to be immediately consumed in subsequent processes
  - Since the special engine is shared within a CP chip, contention among processors is possible
  - Such contention is handled transparently by millicode, and additional delay may be observed

** Further discussion about MVCL vs. MVCLE in “Frequently Asked Questions (7)”*

High Word Instructions

- Provided since z196
  - Intended to provide register-constraint relief for compilers
- High words of GRs are made independently accessible from the low words of GRs
- Software can use up to 32 word-based GRs, 16 doubleword-based GRs, or combination of word and doubleword GRs
- For register dependencies, including address-generation interlocks, the high-word GRs are treated separately from the low-word GRs
- Various types of operations are supported
  - Add, subtract, compare, rotate, load, store, branch-on-count



Conditional Instructions

- In many applications (for instance, sorting algorithms), conditional-branch outcomes are highly data dependent and thus highly unpredictable
  - A mispredicted branch can result in a pipeline flush, and may incur many cycles of branch correction penalty
- A limited set of conditional load/store instructions are provided (z196+) where the execution is predicated on the condition code
  - Highly **unpredictable** branches can be replaced with conditional instructions
- In the example, the old code shows a COMPARE register instruction (CR) followed by a BRANCH ON CONDITION instruction (BRNE for BC), and a LOAD instruction (L) that may or may not be executed depending on the outcome of the branch
- The new code sequence replaces the branch and load instructions with a LOAD ON CONDITION (LOC) instruction
  - It is cracked into a load from storage and a conditional select µop
  - The conditional select µop uses the condition code to select between the original register value and the new value from storage
  - This sequence now avoids potential branch wrong flushes

NOTE: Access exceptions may be reported whether the storage content is effectively accessed or not

```
Old Code
        CR   R1, R3
        BRNE skip
        L    R4, (address X)
skip    AR   R4, R3
        ..



New Code
        CR   R1, R3
        LOC  R4,(address X), b'1000'
        AR   R4, R3
        ..


*Pseudo-code for illustration only
```

EXECUTE Instructions

- "Execute" instruction is commonly used* with targets being storage-related instructions (e.g., MVC, CLC mentioned before) where the length field (specifying the number of bytes) can be substituted with the contents of a general register (GR) without actually modifying the instruction in memory (and without explicit branch to or from the 'target' instruction)
- "Execute" is handled by the processor like a branch, by
  - Jumping to the target of the execute instruction as a branch target, and fetching it
  - Decoding and executing the target instruction (modifying as needed)
  - Immediately returning back to the subsequent instruction after the "execute" (except when the target is a taken branch itself)

– This "implied" branch handling is supported by the branch prediction logic to reduce the overall processing delay

- Certain pipeline delay is required between the reading of the GR and the "modification" of the target instruction
  - The delay is reduced since z13 for a selected group of instructions: MVC, CLC, and TRANSLATE AND TEST (TRT)
- When the operand length is mostly random during run-time, the alternative of using a branch table is not preferred due to its potential inaccuracy in branch prediction

```
Example where MVC's length depends
on compare of R1 and R3:

        LHI  R4, x'1'
        LHI  R5, x'2'
        CR   R1, R3
        LOCR R4, R5, b'1000'
        EX   R4, move
        ..

move    MVC  0(length, R13), 0(R14)

*Pseudo-code for illustration only
```

*other **tricky** EXECUTE usages are not discussed here; e.g., in modifying register ranges, lengths of operand 1 or operand 2, and branch masks*

BRANCH PREDICTION PRELOAD Instructions

- BRANCH PREDICTION PRELOAD (BPP) and BRANCH PREDICTION RELATIVE PRELOAD (BPRP) instructions introduced with zEC12 specify the location of a future to-be-taken branch and the target address of that branch
- By providing such directives to the hardware's branch prediction logic, the limitation of the hardware branch table's capacity may be overcome
  - The processor may now predict the presence of branches without having seen them before or if their history was displaced
  - The directives will not override or modify an existing hardware history entry's target address
- As described earlier, the branch prediction logic should always search ahead 'asynchronously' of where in the program instructions are currently being decoded and executed
  - Just like requesting a stop on a bus, the request needs to be activated BEFORE the bus passes the desired stop; to be effective, the preload instruction needs to be executed **before** the prediction logic may search pass the branch address to be effective
  - The preload instructions are thus best used when the program's run-time behavior involves a lot of somewhat cold modules; such that (taken) branches are likely not being predicted and the instructions are likely not in the cache; such that the preload instructions can have a good chance of being executed AHEAD of the search logic
  - The actual usage of the preload instruction is therefore most effective when in conjunction with profile-directed feedback (PDF), or in a JIT environment where the run-time characteristic can be extracted and analyzed
- The more (taken) branches in-between and the further away in sequential memory address, the more likely a preload will succeed
  - At a minimum, the target branch should be more than 1 (taken) branches and 256 sequential bytes away
- The relative form of preload instruction, BPRP, should be used if possible as it can be activated earlier in the pipeline, providing a better chance of being effective
- The preload mechanism may also perform an instruction cache touch (and thus a potential prefetch) on the branch target
  - Do not use for purely instruction cache prefetches, as that will pollute the branch prediction history structure

PREFETCH DATA Instructions

- Starting with z10, PREFETCH DATA (PFD) and PREFETCH DATA RELATIVE LONG (PFDRL) instructions were introduced to enable program code a way to manipulate the local data cache
  - It is architecturally a no-op
- The provided prefetch function allows code to potentially acquire a cache line (into L1) in a correct cache state (for read-only or for write) ahead of the actual load/store instructions that will access the data
  - Note: prefetching a cache line that is contested among multiple processors is usually a bad idea
- These prefetch instructions not only allow operand data prefetching, but they also provide a way to release a local cache line's ownership (also known as untouch)
  - The untouch function is to allow software code to proactively release (or invalidate) its ownership (from the processor that it is running on) of a specified cache line; as such it can be used when done using a shared data structure
  - Such that, when a different processor accesses this same cache line some time later, the shared cache (L3/L4) will not need to spend time in removing the line from this "last-owning" processor before granting ownership to the "newly-requesting" processor
- These directives should be used carefully, and some experimentation may be required to yield desired performance effect
  - Prefetch function can be redundant with given hardware capabilities
    - The out-of-order pipeline inherently performs "baseline" prefetching
    - The stride-prefetch engine also prefetches cache lines based on fetching patterns and miss history
    - The L4 cache does limited prefetching from memory based on certain miss criteria
    - Prefetch can hurt if the cache line is contested with other processors
  - Untouch function can be tricky to use
    - If it is a highly contested cache line, demote operation might hurt (by adding more related operations to the system)
    - If the cache line is cold, it might not matter
    - In general, the demote variant (code 6) is preferred to the full untouch variant (code 7) since it usually incurs less overhead; as it can be used when done updating a shared data structure
- *NOTE: as stated earlier, while IBM Z's 256-byte cache lines aren't likely to change anytime soon, rather than hardcoding to that or any other cache-related assumptions, use EXTRACT CPU ATTRIBUTE (ECAG) to minimize the chance of observing adverse effects on different hardware models*

NEXT INSTRUCTION ACCESS INTENT (NIAI) Instruction

- The NIAI instruction was introduced in zEC12 for program code to provide some hints to the cache system on the intention of the next immediate instruction's operand accesses, so the hardware can adjust its related handling
  - The instruction behaves like a "prefix" instruction but architecturally it is a separate (no-op) instruction
  - Hints will also be passed into instructions that are implemented in millicode
- The cache subsystem provides heuristic to maintain cache ownership among multiple processors
  - Upon a cache miss from a "current" processor core for a "fetch-only" (non-storing) instruction, the cache subsystem may return an exclusive state if the cache line was previously updated by a "previous" processor
  - This design anticipates that this "current" processor will likely follow suit of the "previous" processor and store to the cache line after this fetch-only miss, saving coherency delays (otherwise seen when changing from a shared state to an exclusive state)
  - In the case where the heuristic is not working perfectly, e.g., when there are multiple "readers" on a cache line, the NIAI instruction (code 1 - "write") can be used by a "writer" process to indicate subsequent store intention upon an initial fetch
- The NIAI instruction can also be used to indicate "truly read-only" usage of a cache line.
  - Given the "reader and writer" processes described above, a NIAI (code 2 – "read") can be used to specify the read-only intention of the consumer (or reader) process's accesses to a cache line; thus, preventing the line from potentially migrated to the reading processor as exclusive (write) ownership
  - The hint can now help reduce the coherency penalty on the next round when the producer (or writer) process is writing into the cache line again
- Cache lines are usually managed from most recently used (MRU) to least recently used (LRU) in the cache, so lines that have not been used recently are evicted first when new cache lines are installed
  - This scheme generally works well, but is suboptimal in cases where the process is operating on streaming data where data is only accessed once and then becomes uninteresting
  - In these streaming cases, it is desirable to label such data as LRU so that it's not retained at the expense of other data that will be used again
  - The NIAI instruction (code 3 – "use once") can be used to indicate streaming data accesses such that the local cache will keep those data in compartments that will be evicted sooner

Atomic and Locking Instructions

- z/Architecture provides a set of instructions that can be used for atomic operations
  - e.g., TEST AND SET (TS), COMPARE AND SWAP (CS)
  - They check a value in storage (fetch) and then conditionally update the storage value (store) such that the fetch and the store are observed to be “atomic”, meaning to an outside observer the two actions appear to have occurred simultaneously
- The Interlocked-Access Facility instructions were added on z196
  - Load and “arithmetic” instructions for unconditional updates of storage values
    - (Old) storage location value loaded into GR
    - Arithmetic or logical operation (*add, and, xor and or*) result overwrites value at storage location
    - Best for unconditionally updating global information, like a counter or a flag
  - Interlocked storage updates with an immediate operand are also supported
    - Supported operations include *ADD (LOGICAL) IMMEDIATE*, *AND IMMEDIATE,* XOR IMMEDIATE and OR IMMEDIATE
  - LOAD PAIR DISJOINT (LPD, LPDG)
    - Load from two different storage locations into GR N, N+1
    - Condition code indicates whether the fetches were atomic
- Hint: For software locks, if the lock is likely concurrently used by multiple processors (i.e., often contested), the following sequence should be considered
  - It is more desirable to test the lock value before using atomic instruction (e.g., CS) to set the lock

```
        LHI   R2, 1          ; value to set lock
LOOP    LT    R1, lock       ; load from memory and test value; always test first
        BCR   14,0           ; serialization to architecturally guarantee getting new value
        JNZ   LOOP           ; repeat if non-zero
        CS    R1, R2, lock   ; set lock if lock was empty
        JNE   LOOP           ; retry if lock became set
*Pseudo-code for illustration only
*See additional discussion in “Frequently Asked Questions (2)”
```

Hardware Transactional Memory

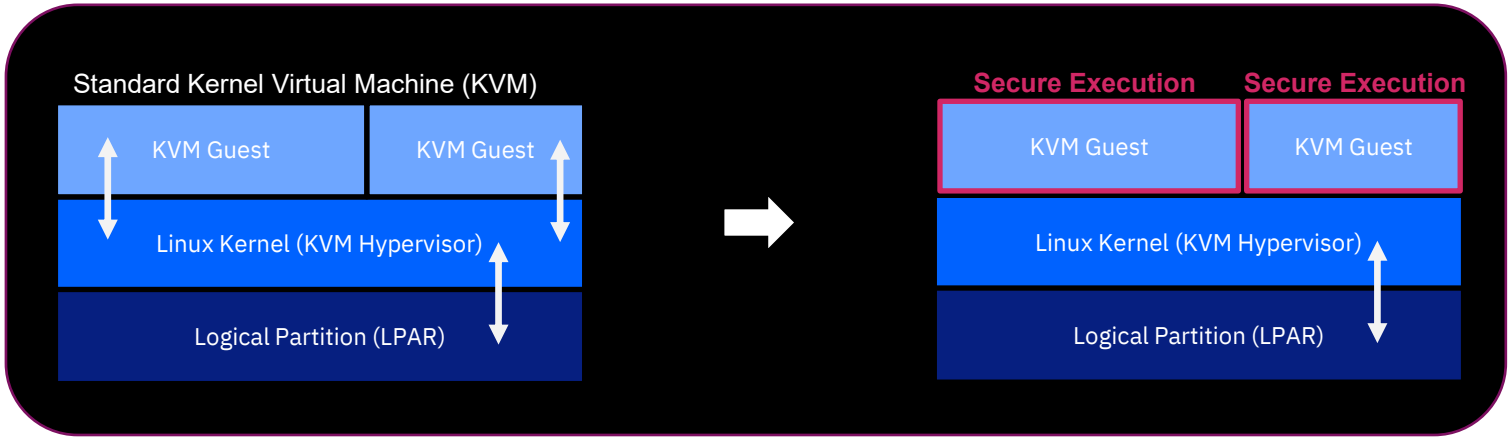
- Statement of direction (with z16 announce):
  - **"Removal of support of the transactional execution and constrained transactional execution facility:** In a future IBM Z hardware system family, the transactional execution and constrained transactional execution facility will no longer be supported. Users of the facility on current servers should always check the facility indications before use."
  - ***More specifically, the facility is still fully supported on z16, however z17 ~~is expected to~~ supports only constrained transactions while the follow-on to that is not expected to support any transactions***
- TX removal-related instruction support on z17:
  - Adding extensions to Perform Locked Operation (PLO) to provide functionality with simpler HW
    - Including up to 4 way Compare and Swap (CS) on octwords
- *Beginning in zEC12, z/Architecture supports hardware transactional (memory) execution through the transaction execution facility, occasionally referred to as TX*
  - *A group of instructions can be observed to be performed with atomicity, or not done at all (aborted)*
  - *Non-transactional stores are allowed within a transaction*
  - *A form of constrained transaction (transaction with restrictions) is also supported where the hardware will automatically retry an aborted/failed transaction until the transaction is successful*
  - *Optional detail debug data can be provided*
- *Transaction usage is not advisable if the contention of used storage is already high*
  - *Likely end up wasting CPU cycles if the transaction keeps aborting due to real-time cross-CPUs memory access contentions*
  - *Aborts are expensive (>200 cycles) and worse if abort debug information is requested*
- *Hint: compute complex results outside of a transaction, then use transaction with only a small number of instructions to check data, and then store the results away*
- *Access (fetch) footprint* is limited by L2 cache associativity and size*
  - *Up to 1 Mbyte in zEC12, 2 Mbyte in z13, 4 Mbyte in z14..z15*
- *Update (store) footprint* is limited by L2 cache associativity and size of an internal store buffer*
  - *The buffer design can support up to 64 blocks of 128-byte (storage-aligned) data changed within a transaction*
  - *The L1 data cache is updated as each store instruction completes within a transaction, but L2 cache update from the buffer is deferred until transaction completes*
- *Note: Access footprint may be counted for fetches done through mispredicted branches. Footprint limitations are shared by the 2 threads when SMT2 is enabled such that effective footprint may be smaller than when one thread is running*

Guarded Storage Handling

- Beginning in z14, z/Architecture provides the guarded-storage facility for programming languages to more efficiently implement storage-reclamation techniques commonly referred to as garbage collection (GC)
  - Allows application/user threads to continue running concurrently during phase of GC compaction
  - Accomplished by providing hardware-assisted read barriers for guarded storage (GS) involved in a compaction event
- Prior to compaction, software defines a "range of guarded-storage segments" by specifying the GS starting/ending address and segment size to cover the entire region containing all guarded segments. A bit mask vector is set to define which segments are protected.
  - Done through the LGSC (LOAD GUARDED STORAGE CONTROLS) and STGSC (STORE GUARDED STORAGE CONTROLS) instructions
- When guarded storage is set up and enabled
  - When the second operand of the LGG (LOAD GUARDED) or LLGFSG (LOAD LOGICAL AND SHIFT GUARDED) instruction designates a GS segment as specified by software, a guarded-storage event is recognized, and control is passed to a guarded-storage event handler
  - Otherwise, the respective instructions perform their defined load operations
  - All other instructions that access a segment of guarded storage are unaffected by the facility
    - Only the new LGG and LLGFSG instructions can potentially generate a guarded-storage event
- It is expected that the problem-state guarded-storage event handler would
  - First save program GRs, move offending data to a non-guarded area and fix up any pointer addresses if necessary
  - Then after restoring program GRs, it will branch back to the interrupted program address previously saved during the GS event and continue normal program operations
- NOTE: Some refer to this feature as 'pause-less garbage collection', where 'pause-less' should be understood to mean 'less-pausing' and *not* 'pause-free'

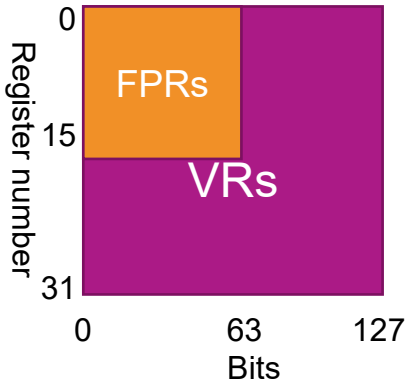
Secure Execution

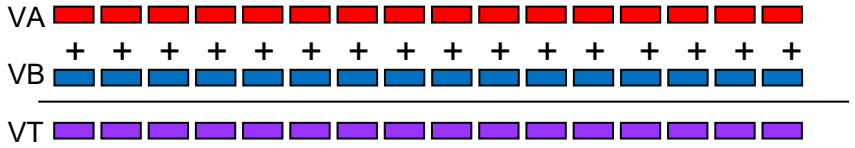
- Beginning with z15, a hardware-based security technology, called Secure Execution (SE) architecture, is introduced
- Secure Execution
  - Provides a virtual machine that is fully isolated and protected from the hypervisor with encryption keys that only the IBM z hardware and firmware have access to
  - Enables hosted workloads to process unencrypted memory securely without exposing it to the host or any other workloads in the same environment
- As an example, KVM guests enabled in secure execution mode will have their memory and execution states protected
  - Secure memory can only be accessed in secure mode
  - Instructions are only executed from the secure memory of the guest in secure mode
  - A new trusted firmware layer called the Ultravisor sits between the hardware and hypervisor to implement high-level security requirements, e.g., encrypting memory blocks before export (paging), and decrypting them on import
  - Although there will be some modest overhead passing controls through the Ultravisor, application performance should perform similarly while in SE mode vs. not in SE mode



Single-Instruction-Multiple-Data (SIMD)

- SIMD instructions, sometimes also referred to as vector instructions, were introduced in z13
  - See Eric Schwarz's Journal article[11] for additional color
  - More instructions are added on z14 and onward, but only a selected set will be discussed here**
- To support these instructions, new vector registers (VRs) are architected
  - 32 x 128-bit architected registers are defined per thread
  - FPRs overlay VRs as follows:
    - FPRs 0-15 == Bits 0:63 of SIMD registers 0-15
    - Update to FPR <x> alters **entire** SIMD register <x>
    - Whenever an instruction writes to a floating-point register, bits 64-127 of the corresponding vector register are unpredictable
- Each SIMD instruction provides fixed-sized vectors ranging from one to sixteen elements
  - Some instructions operate only on a subset of elements
- The use of vector-compares and vector-select operations can help avoid unpredictable branch penalties similar to the simple conditional instructions described earlier





Most instructions have a non-destructive operand encoding (T=A+B vs. A=A+B)

*** Use latest z/Architecture for full reference*

Table 3: Types of SIMD instructions*

Integer

16 x 8b, 8 x 16b, 4 x 32b, 2 x 64b, 1 x 128b

- 8b to 128b add, subtract
- 128b add/subtract with carry
- 8b to 64b minimum, maximum, average, absolute, compare
- 8b to 16b multiply, multiply/add
  32b x 32b multiply/add
- Logical operations, shifts
- Carry-less multiply (8b to 64b), Checksum (32b)
- Memory accesses efficient with 8-Byte alignment; minor penalties for byte alignment
- Gather / Scatter by Step; Permute; Replicate
- Pack/Unpack
- (z14) Pop-count (8b to 64b)
- (z14) Large integer multiplication for matrix operations that are used in cryptography

String

- Find 8b, 16b, 32b, equal or not equal with zero character end
- Range compare
- Find any equal
- Isolate String
- Load to block boundary - load/store with length (to avoid access exceptions)
- (z15) (sub)String search

Decimal (z14)

- Register-based versions of storage-based arithmetic/convert instructions
- Multiply-shift, shift-divide

Floating-point

2 x 64b

- Binary Floating-Point operations with double precision
- 2 BFUs with an effective increase in architected registers
- All IEEE trapping exceptions reported through VXC, and will not trigger interrupts
- Compare/Min/Max (per language usage) added in z14
- Two more precision binary floating-point operations are added in z14

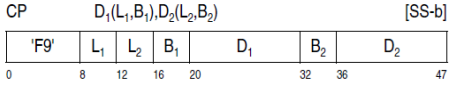
4 x 32b
1 x 128b

** More instructions are added on z14+, but only a selected set will be discussed here*

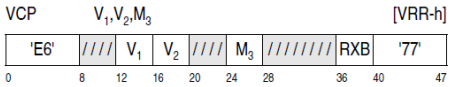
Binary Coded Decimal (BCD) Vector Register-based Insns

- In z/Architecture, decimal numbers may be represented in formats that are of variable length.
  - Each byte of a format consists of a pair of 4-bit codes; the 4-bit codes include decimal digit codes, sign codes, and a zone code
- New architecture is introduced in z14 to support BCD operations through vector registers, reducing memory accesses:
  - Load, store with length right-aligned
  - BCD arithmetic
- Special features that are targeted to COBOL (vs. existing BCD instructions)
  - Static & dynamic length specification
  - Elaborate signed / unsigned control
  - Condition code setting is optional
  - New variants of multiply and divide
- NOTE: Available with COBOL V6.2 and ARCH(12) compilation switch; NOTE: if ARCH() isn't specified the default is ARCH(7) for z9!
- Performance of conversion between decimal and binary formats is improved in z15
- More instructions are added with z15, but not all are discussed here

COMPARE DECIMAL



VECTOR COMPARE DECIMAL



	OLD	NEW (since z14)
Function	Classic, storage-based	VR-based
Add	AP	VAP
Subtract	SP	VSP
Zero & add	ZAP	VPSOP
Compare	CP	VCP
Test	TP	VTP
Shift & round	SRP	VSRP
Multiply/multiply-shift	MP	VMP, VMSP
Divide/remainder/shift-divide	DP	VDP, VRP, VSDP
BCD → Binary	CVB*	VCVB*
Binary → BCD	CVD*	VCVD*
Load immediate		VLIP
Load rightmost		VLRL(R)
Store rightmost		VSTRL(R)

Vector-Packed-Decimal-Enhance Facility 2 – *new to z16*

- Adds VR-based instructions to convert
  - Between hexadecimal floating point (HFP) and signed packed decimal (SPD) formats
  - Between SPD and zoned decimal (ZD) formats

input	operation	output
SPD number	scale, convert to HFP & round	S, L, or X HFP number
SPD number	scale, convert to HFP & “split” (“*split*” *is a COBOL special request)*	S HFP “top” & L HFP “bottom”
X HFP number	convert to SPD, scale & round	SPD number
ZD number	convert to SPD	SPD number
SPD number	shift & round	SPD number
SPD number	convert left 15 digits to ZD	high part of ZD number
SPD number	convert right 16 digits, sign to ZD	low part of ZD number

- Adds instruction to count leading zero digits of an SPD number

Uniprocessor Storage Consistency

- Uniprocessor view of storage consistency
  - General rules (important for full software compatibility):
    - Program must behave as if executed serially
    - Each instruction can use all results of previous instructions
  - Operand storage accesses must be observed to be done in program order
    - Store / fetch conflicts are recognized by **real*** address
    - Most operands are processed left to right
    - Fixed-point decimal operands are processed right to left
    - Storage-storage (SS) instructions are observed to operate in a byte-by-byte fashion
  - Instruction pre-fetches may be observed
    - Must still detect store updates / instruction fetch conflicts; where detection is on **logical*** address only
    - Instructions executed must reflect prior stores
    - Serialization can add further restrictions (see details in next 2 pages)

**Logical address*
- *What program specifies*
- *May be virtual or real, depending on DAT (dynamic address translation) mode specified in the program status word (PSW)*

**Real address*
- *Result of dynamic address translation process or the logical address when DAT mode is off in PSW*
- *Subject to prefixing*

Multiprocessor Storage Consistency

- Must be able to define consistent ordering of accesses
  - “as seen by this and other processors"
  - Some instruction operations are allowed to have ambiguous results (See the section “Storage-Operand Consistency” in the z/Architecture Principles of Operation for details)
- Operand fetches and stores must appear to occur in proper order
- All processors must obey uniprocessor rules
  - Although the processor is designed to do things out-of-order, the observed results must be consistent
  - The processor has states and checking in place, such that when the out-of-order accesses might be observed to be inconsistent, the pipeline will flush and retry the operations; possibly in a “safer” (slower) mode
- Operand accesses must be DW-consistent
  - No "score-boarding" should be observed
  - e.g., DW consistency is maintained for LOAD MULTIPLE (LM) when the loads are expanded into individual GR writing operations
- Instruction fetches are generally allowed in any sequence

```
CPU1                CPU2

Store  R1, AA       Store   R1, BB
Load   R2, AA       Load    R2, BB
Load   R3, BB       Load    R3, AA

As an example, if both final Load instructions get “old” (pre-store) values : Violation!
```

Serialization

- z/Architecture defines a set of situations in which additional restrictions are placed on the storage access sequence
- Defined as “A serialization operation consists in completing all conceptually previous storage accesses and related reference-bit and change-bit settings by the CPU, **as observed by other CPUs** and by the channel subsystem, before the conceptually subsequent storage accesses and related reference-bit and change-bit settings occur”
- Defined for specific points in instruction stream
  - Usually "before and after" specific opcodes
  - Includes Instruction fetches as well as operand accesses
  - Exception: Instruction fetch for the serializing instruction itself
- See additional discussion in “frequently asked questions (2)”

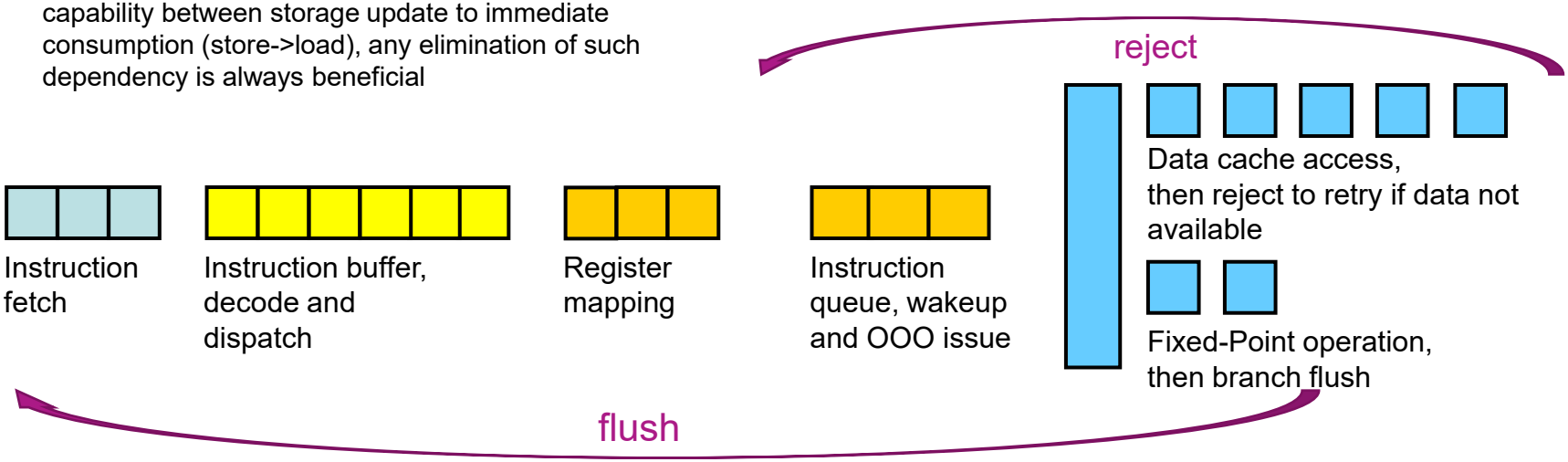
```
CPU 1                   CPU 2

MVI   A, X'00'          G   CLI   A, X'00'
BCR   14, 0                 BNE G

The BCR 14, 0 instruction executed by CPU 1 is a serializing instruction that forces the store
by CPU 1 at location A to become visible to other CPUs. However, per the architecture, CPU 2 may
loop indefinitely, or until the next interruption on CPU 2, because CPU 2 may already have
fetched from location A for every execution of the CLI instruction. The architecture requires
that a serializing instruction be placed in the CPU-2 loop to ensure that CPU 2 will again fetch
and refresh the data value from location A.
```

General Optimization Guidelines

- See “References” near the end of this document on topics you might be interested
- CPU Measurement Facilities (CPU MF) are provided with user-accessible hardware instrumentation data to understand performance characteristics
  - Documentation and education materials can be found on online with IBM’s supporting websites
  - Extracted data is presently made available in z/OS through SMF 113 records and in z/VM through MONWRITE data
  - Linux on Z now also surfaces CPU MF data: https://linux.mainframe.blog/cpumf/
- Some general recommendations will be provided next, including some that have been mentioned in previous pages
  - All descriptions provided are of general guidance only
  - It will not be practical to describe all intricate design details within the systems in this document
  - There may be counter-examples (usually rare occurrences) that will observe hardware behavior differently than described, or not adhere to optimization recommendations provided
  - Detailed instruction-by-instruction classifications and timings will not be provided in this document
- Z processors are designed for processing both cache-intensive and CPU-centric workloads and are optimized to handle code that was hand-written many years ago or was generated from the latest compilers, with that code running in applications, middleware or operating systems
  - However, code that was hand-written or generated using older versions of compilers many years ago could benefit from a scrub or refresh as there are very likely new instructions and optimizations that would further improve code performance
  - General rules that help produce good performance code for modern processor microarchitectures usually apply to z processors too
  - Microprocessor pipeline, branch prediction algorithm, cache subsystem structure, and their characteristics will likely change from generation to generation to obtain better general performance improvements and bigger system capacity
  - Code sequence can be tuned to get more performance by optimizing to a new processor pipeline, or by using new instructions or new architectures
  - Performance variations should be expected on highly optimized code that is tuned to a specific processor generation vs. another generation

Optimization - General Pipeline

- Recent z microprocessor cores have fairly deep instruction pipelines
  - Driven by high-frequency design (up to 5+ GHz since zEC12)
  - Latest z13+ pipeline: 20+ cycles from instruction fetch to instruction finish
- Pipeline hazards can be expensive
  - Branch wrong flush (for either direction or target) – 20+ cycles
  - Cache reject when cache data is not available, or when cache resources are not available – 12+ cycles
- Code optimization technique can help, for example (more guidelines provided in subsequent pages):
  1. Arrange frequent code in “fall through” paths
     Although the z processors improve upon branch prediction design every generation to get better accuracy and larger coverage, allowing the frequent code in a ”straight-line” sequence is always beneficial
  2. Pass values via registers rather than storage
     Although the z processors improve data bypassing capability between storage update to immediate consumption (store->load), any elimination of such dependency is always beneficial



Optimization - Branch Handling (1)

- Align frequently called functions to start at storage boundaries for efficient instruction fetching
  - at least at quadword (16-byte) boundary, but potentially even better if at octword (32-byte) or cache-line boundaries
- Rearrange code path around conditional branches such that the not-taken path (i.e., fall-through path) is the most frequent execution path
- Although the branch predictor attempts to predict every cycle, keeping loops to be at least 12 instructions on z13+ (6 instructions on z196 and zEC12) will allow branch prediction to catch up
  - If more instructions can be used, branch prediction will be able to stay ahead of instruction fetching
- Although z processors before z14 do not include a call-return predictor, z14 included a simple call-return predictor such that pairing up calls and returns, i.e., no nesting of calls, may facilitate the design to work more effectively
- Consider inlining subroutines if they are small and used often
- Unroll loops to parallelize dependency chains to maximize the advantage of parallel and out-of-order processing
- Use relative branches instead of non-relative (indirect branches) when possible
- There is usually an advantage to use a branch-on-count or a branch-on-index type instruction versus doing the operations as individual instructions, due to
  - Smaller instruction footprint and less hardware overhead
  - Branch-on-count type and branch-on-index-low-or-equal type instructions are predicted taken whenever the branch prediction logic is not able to predict its direction ahead of time
- Similarly, load-and-test or compare-and-branch type instructions will perform better than a pair of individual instructions
- Avoid hard-to-predict branches by using conditional instructions
  - Conditional instruction is usually slower than a correctly predicted branch + load/store instruction; thus "hard-to-predict" is an important criteria

Optimization - Branch Handling (2)

- The main branch prediction structure (BTB1) is managed on a fixed sized memory block basis, the total number of branches within a block that can be saved is therefore limited by the number of sets in the set-associative design
- The block size and set associativity is shown in table below:

System	Block Size	Number of Branches
z196	32 Byte	4
zEC12	32 Byte	4
z13	32 Byte	6
z14	32 Byte	4
z15	64 Byte	8
z16 and z17	128 Byte	up to 24

- For example, in z196, zEC12 and z14, when the BTB1 is designed as a 4-way set associative structure, it can keep prediction history of up to 4 branches per a 32-byte block of code in memory
- To account for other potential conflicts, it may be advantageous to limit the number of branches within a block to half the number of sets available, e.g., up to 2 branches per a 32-byte block for z196, zEC12 and z14
- Typical code sequence that has many branches within a block is due to the usage of branch tables. One can help by
  - Rearranging the branch table with intermixed infrequent entries
  - Padding the branch table entries with no-op instructions

Optimization - Instruction Selection (1)

- Register-storage format instruction is often more efficient than a 2-instruction sequence of “load” + “register-register” operations
- Use instruction variants that do not set condition code if available (and when the resulting condition code is not required)
- Use instructions of shorter instruction lengths if possible
- An instruction accessing storage by using a Base + Index + Displacement form (3-way) of address generation incurs no additional penalty vs. a Base + Displacement form (2-way) or a register-based form
  - Similarly, Base + Index + Displacement form for branch target calculation incurs no additional delays vs. a register form; e.g. BC vs. BCR
  - Precompute storage address only if you can use it for branch prediction preloading or operand data prefetching
  - However, “Load Address” type instructions will take an extra cycle through the FXU when both base and index registers are not using GR#0
- Understand rotate-then-*-selected-bits instructions, and see whether they can be used
  - The second-operand register is rotated left by a specified amount; then one of four operations (*and, xor, or, insert*) is performed using selected bits of the rotated value and the first-operand register
- Use compare-and-trap instructions where practical, in particular, for null-pointer checking
- Take advantage of the additional high-word GRs instead of performing register spill-and-fill through storage
  - Since z13, VRs might also be used
- Starting with z15, SELECT (and SELECT HIGH) instructions are provided to select one out of two input registers depending on the condition code, and place it into a third register
  - Can be used instead of using branch instructions
- Starting with z16, Neural Network Processing Assist or NNPA can employ the AIU (Artificial Intelligence Unit) accelerator to perform 2x8x8 matrix operations with direct memory access or DMA
  - Significantly more performant than manually aggregating discrete vector operations
  - Requires up front tensor conversion into IBM’s Deep Learning tensor Format (DLF)

Optimization - Instruction Selection (2)

- Regular register clearing instructions are fast-pathed in the pipeline, and their results do not use any physical register renames (since zEC12)
  - SUBTRACT or EXCLUSIVE OR register (SR/SGR, XR/XGR of the same register); which sets CC=0
  - LOAD HALFWORD IMMEDIATE (LHI, LGHI of immediate value 0..7), which leaves CC unchanged
  - LOAD ADDRESS (LA) where Base, Index, and Displacements are all zero's
  - Since z13, LOAD ZERO {long}, {extended} (LZDR, LZXR), INSERT/AND IMMEDIATE {high} {low} (IIHF, IILF, NIHF, NILF) with immediate value of 0, and VECTOR GENERATE BYTE MASK (VGBM) with immediate value of 0 (aka VZERO)
- Use the long-displacement variants, with a 20-bit signed displacement field, that provide a positive or negative displacement of up to 512K bytes if necessary
- A set of instructions (ends with RELATIVE or RELATIVE LONG) is provided to operate on data elements where the address of the memory operand is based on an offset of the program counter rather than an explicitly defined address location. The offset is defined by an immediate field of the instruction that is added (as a sign extended, halfword-aligned address) to the value of the program counter
  - Load, store, and various kinds of compares are provided
  - Such accesses are treated as data accesses (except for EXECUTE RELATIVE LONG); these data elements should not be placed anywhere near the same cache lines as the program instructions to avoid potential cache conflicts
- For operations on large amounts of memory, e.g., copying or padding storage, consider using instructions that can handle long operand lengths, e.g., MOVE characters (MVC), instead of doing individual loads or stores
- Complex instructions, e.g., COMPRESSION CALL (CMPSC), convert-UTF-UTF instructions, and cryptographic instructions, with help of the per-core co-processor, are usually faster than software routines especially for large datasets
- For serialization, a BCR 14, 0 (supported since z196) is better than BCR 15, 0 (which also requires checkpoint synchronization needed for software checkpoints that might incur additional delays)

Optimization - Instruction Selection (3)

- For storing clock value, use STOCK CLOCK EXTENDED (STCKE); if uniqueness is not required, use STORE CLOCK FAST (STCKF)
  - Design of z14..z15 further optimizes towards STCKE/STCKF, STCK usages might observe relative slow-down
- Use simple “interlocked-access” instructions, e.g., LOAD AND ADD (LAA), OR/AND/XOR immediate (OI, NI, XI), instead of conditional loops using compare-and-swap type instructions, for any **unconditional** atomic updates
  - OI, NI, XI (and their long displacement variants, OIY, NIY, XIY) were used in examples that did not interlock in earlier architecture; these instructions are now interlocking since z196
- Control instructions that change the PSW masks/modes/spaces will introduce some forms of in-order operations within the pipeline and their usage should be limited
- EXTRACT and STORE FLOATING POINT CONTROLS instructions will also introduce some forms of in-order operations and should be avoided if possible
- Translation/Key changing/purging control instructions may involve some forms of serialization and should be limited
  - For SET STORAGE KEY EXTENDED, usage of the non-quiescing (NQ) variant is encouraged
- With z15, the memory loading bandwidth from the L1 data cache to the vector SIMD engines has doubled
  - This extra bandwidth is enabled by specifying the Doubleword/Quadword **alignment hints** on vector load instructions
  - When a correct alignment is provided (and if the vector load doesn’t cross a cache line), it will execute the vector load in 1 cycle instead of 2 cycles in prior designs
  - However, if alignment is provided and incorrect, the vector load will cost a pipeline reject
- With z15, efficient instructions are provided, such as
  - non-destructive 32- and 64-bit binary operations (NAND, NOR, AND/OR WITH COMPLEMENT, NOT EXCLUSIVE OR)
  - POPCNT can count all the bits in a GR
  - MOVE RIGHT TO LEFT (MVCRL) or “backward MVC” is provided to copy storage in the opposite direction vs. MVC
    - Length (at most 256 bytes) in GR0
    - Can be used to “open a hole” in a storage array by moving an element and everything else further to the right
    - Unpredictable in case of destructive overlap

Optimization - Instruction Scheduling (1)

- Optimizing instruction grouping efficiency might yield better performance
  - Arrange code such that 3 instructions that can be grouped together to optimize dispatch bandwidth
  - Instruction clump formation (instruction storage alignment) affects how instructions are fetched from the instruction cache, and may affect grouping effectiveness
  - Branch instruction ends a group in z196; but since zEC12, it ends only if it is predicted taken or if second in the group

- Execution results can be bypassed without any additional latency to a dependent instruction if the sourcing and receiving instructions are on the FXUs (FXUa, but not FXUb in z13+) of the same side of the issue queue
  - This bypass can be arranged by placing the related instructions consecutively, and thus usually in the same group (and the same side)

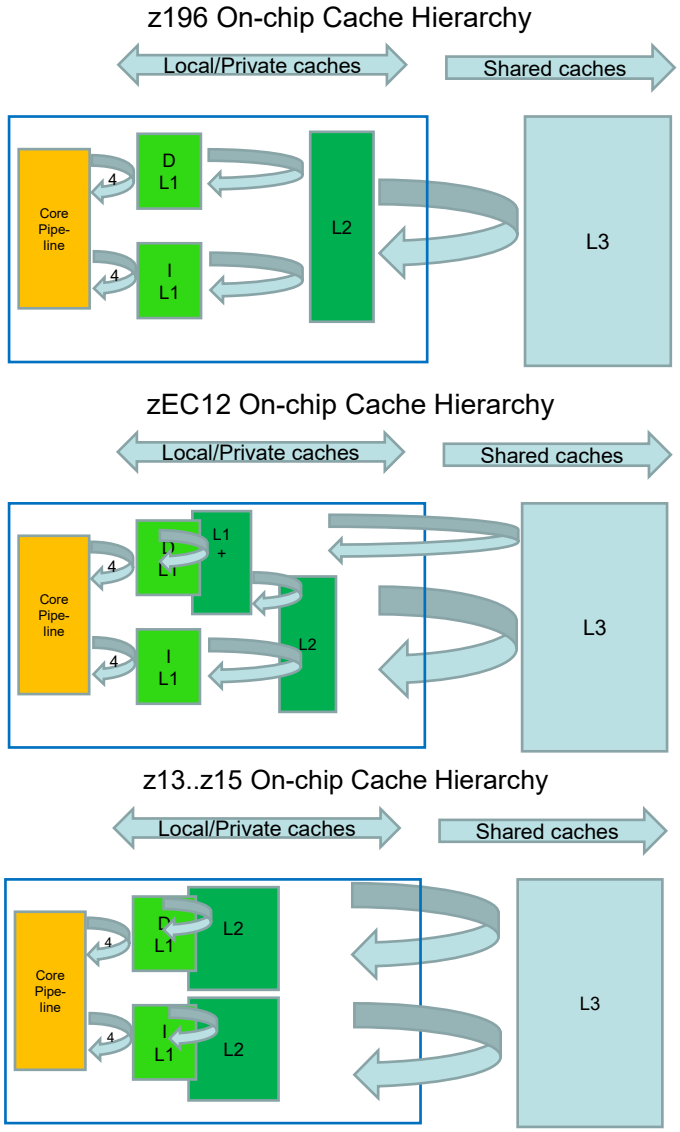
- Floating-point (FP) operations
  - Mixed mode FP (e.g., short->long, long->short, hex->bin, bin->hex) operations should be avoided; results are typically not bypassed, and can cost pipeline rejects or flushes
  - Since z13, the simpler mapper tracker design used for VRs (and FPRs) can lead to false dependencies in single precision FP operations; where possible, double precision FP operations should be used
  - Since z13, execution functions are evenly distributed (symmetric) among the 2 sides of the issue queue, scheduling that enables parallel processing among the 2 different sides can potentially achieve better performance
    - For reference, in z196 and zEC12, floating-point unit and fixed-point multiply engines are only provided on one side of the issue queue
    - Since z13, FP results bypassing capability are symmetric among FP operations from the two issue queue sides
    - The pipeline design can be very consistent in terms of instruction grouping when processing instructions in a loop, it is best to have long running or non-pipelined instructions, like floating-point SQUARE-ROOT/DIVIDE, distributed evenly between side0 and side1 of the dispatch groups by estimating the group boundaries such that they don’t end up in the same side

Optimization - Instruction Scheduling (2)

- Software directives like branch prediction preload and prefetch data instructions are designed to help hardware optimize performance
  - Hardware is implemented to optimize performance (based on patterns) by predicting the intended program behavior. Hints about behavior from the programmer can help.
  - These directives potentially modify behavior of heuristic / history-based hardware mechanisms
    - The net performance improvement might vary between hardware implementations and the hints can potentially be ignored
  - Use responsibly
    - As usage might have adverse effects by increasing overall code size, these directives are best used by applying insights based on run-time profiles such that “blind” insertions can be avoided
    - Experimentation is highly advised
  - These directives should be placed as far back from actual branches or storage accesses as possible to be effective

Optimization - Local Caches

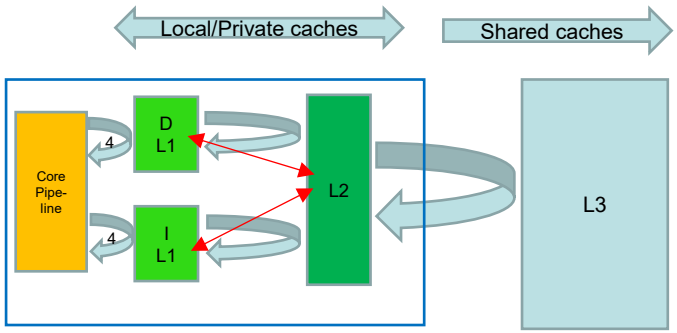
- Many traditional and current workloads running on IBM Z systems are data-centric, where a lot of time is spent in accessing storage and moving data around
  - Understanding the cache topology and optimizing data placement can be very important
- The pictures on the right depict the evolution of the per-core private L2 cache structure. Each generation improves the overall L2 cache size and latency (since z196):
  - By progressing from a unified (z196) to a hybrid (zEC12) to a fully split (z13..z15) design
  - The z13..z15's fully split L2s (vs. unified) for instructions (I-L2) and operands (D-L2) help keep data closer to the corresponding L1s
  - Contrast to the traditional serial lookup design, the integration of L2 directory lookup pipeline into L1 helps reduce access latency for L2 hit and L2 miss since the L2 status is now known much earlier
    - Integrated lookup approach is implemented in zEC12 for operands, then in z13..z15 for both operands and instructions
- The z processors' design allows large and fast L2 caches
  - L2 sizes are comparable to other competitors' L3s (MBs) by leveraging eDRAM technology in z13..z15
  - L2 access latency is also extremely aggressive, at around 10 cycles



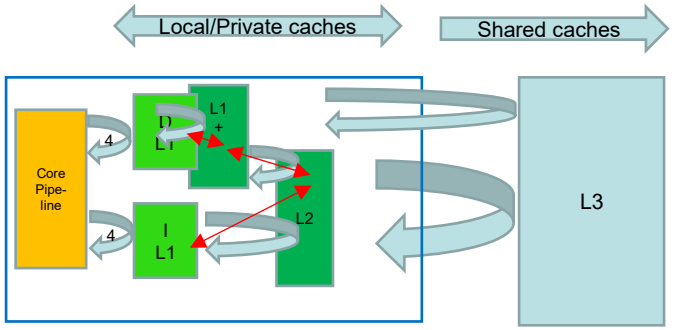
Optimization – Insn/Data Placement

- The on-chip L3 is shared by all cores on the CP chip
  - In z13..z15, it is also the common/sharing point for I-L2 and D-L2
- Split L1 caches design was (re-)introduced in z900 (2000)
  - Designs were optimized for well-behaved code
    - Increased cost of I and D cache contention if they are sharing contents in the same cache line in a conflicting way
  - With split L2 cache design in z13+, cache conflict is resolved through the L3
- What happens when Instructions & Operands share same cache line in a split L2 design (since z13)?
  - OK (maybe inefficient) if operands usage is also read-only /shared since instructions are read-only by design
  - PROBLEM (performance wise) if stores are expected to those operand locations, including those stores on potentially predictive code paths
    - Insn/data cache thrashing and pipeline flushes may be encountered, driving long delays
    - Identified as Store Into Instruction Stream (SIIS) inefficiency
- In general, the split L2 design is not a problem for
  - Re-entrant code or dynamic run-time code
  - Any LE-based compiler generated code
- However, there are problematic examples where conflicts can happen (and should be avoided):
  - True self-modifying code
  - Classic save area
  - Local save of return address
  - In-line macro parameters
  - Local working area right after code

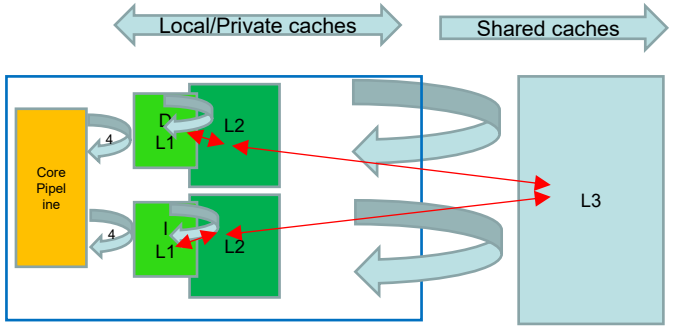
z196 On-chip Cache Hierarchy



zEC12 On-chip Cache Hierarchy



z13..z15 On-chip Cache Hierarchy



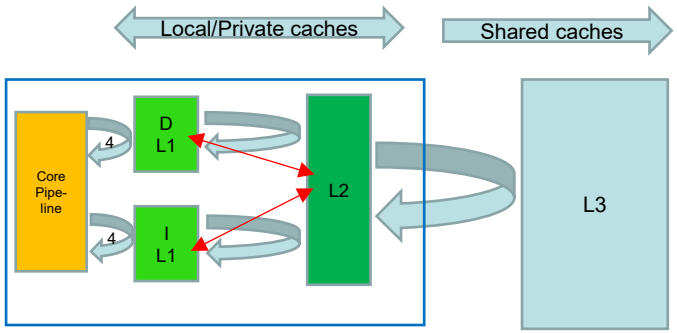
Optimization – Insn/Data Placement

- The cache-line size is 256 bytes throughout the cache hierarchy (L1..L4)
  - One line of separation between insns and writable data will ensure the “heavy hammer” of SIIS-related pipeline flushes will not occur
  - Four lines of separation are recommended and should be practically sufficient on current models to avoid insn and data prefetch thrashing
  - A full 4K page (16 lines) of separation should fully cover prefetch thrashing on future designs

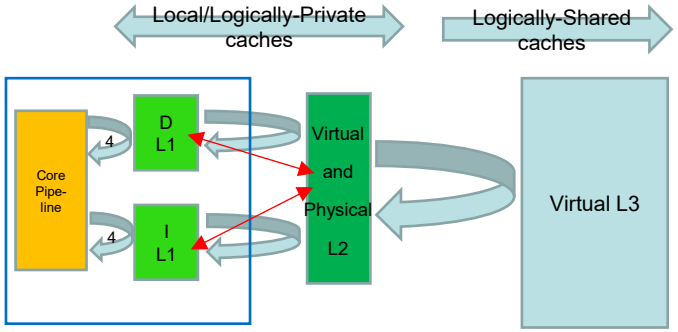
- z13..z15 (lower right) require L3 SIIS resolution and thus have a considerably higher SIIS penalty than z196, zEC12, z16 and z17
  - Regarding z16 and z17, note that for a single core, its nest-coupled, unified physical L2 serves as its logical L2 (middle right) which operates much like z196’s unified on-core L2 (upper right)
    - Despite being off-core, a z16 and z17 L2 is much closer to its core physically or “cycle-wise” than an off-core L3 on prior generations
    - **Ergo z16 and z17’s SIIS penalty is substantially similar to pre-z13 era machines like z196 and zEC12**

- While the SIIS penalty varies significantly by generation, make no mistake: **overhead is incurred on all generations and should be avoided whenever possible**

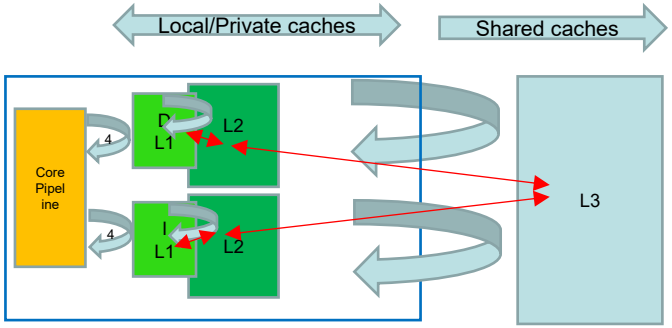
z196 On-chip Cache Hierarchy



z16 and z17 On-chip Cache Hierarchy



z13..z15 On-chip Cache Hierarchy



Optimization - Instruction Cache

- As described in previous pages, instructions (executable code) and operand data (working storage or stack storage) in the same cache lines, which can be costly due to moving cache lines between the separated (split) local caches (instruction/data L1/L2), should be avoided
  - Since both instruction and operand accesses can be predictive in nature, if the instruction and operand storage locations **can be located further apart**, the possibility of leading to unintended cache transfer delays can be reduced
  - The target operand of an EXECUTE-type instruction is treated as an instruction fetch (not data operand) and should be located as part of the instruction cache lines
  - Self-modifying code (or store-into-instruction-stream) is supported in hardware functionally, but in general, the sequence can become costly due to out-of-order pipelining and movement of cache lines
  - Pay attention to local (static) save areas and macro expansions with in-line storage parameters, especially in manually-assembled code, to avoid unintended sharing

- Instruction Cache optimization
  - Minimize the number of cache lines needed through the most frequent execution path
    - Separating out frequently and infrequently used code to different storage areas can improve both cache and translation-lookaside-buffer (TLB) efficiency
  - Software hints, e.g., prefetch data and branch prediction preload instructions, should not be added blindly
    - Unnecessary hints may increase instruction cache footprint and instruction processing delay
    - Branch prediction preload instruction also does instruction cache touch (as a way of prefetching)
  - Unrolling and in-lining should be done to improve potential processing parallelism, but should be targeted with a reasonable resulting loop size; i.e., aim for maximum processing with minimal loop size

Optimization – Shared Data

- Shared data structures among software threads / processes are common
  - Sharing is not necessarily bad and can be very useful to leverage the strongly consistent z/Architecture
- However, when updates are coming from multiple cores, the cache lines will bounce around among caches
  - Depending on locations of cores and states of the cache lines, increased access and possible intervention latencies can be troublesome
- In the case of **false sharing**
  - Where independent structures / elements are in same cache line
    - Potential performance problems can be avoided by separating independent structures into different cache lines
  - As a non-coding example, structures may reside within a non-local cache due to OS dispatch decisions (i.e., multi-processor affinity queues)
- In the case of **true sharing**
  - Where real-time sharing of data structure is required among multiple software threads / processes
  - Often involved with the usage of atomic updates or software locks
  - When such operations are involved at a higher n-Way (concurrent software threads), the more frequent and parallel accesses that are performed at any given time, the more likely the cache lines involved are contested in real time
  - These cases can lead to “hot-cache-line” situations, where care and optimization as described here will be needed to help minimize cache coherency delays of cache lines movement latencies
  - A typical coding example was described in “Atomic and Locking Instructions”
- A recommendation is to consider splitting single locks into multiple locks, when possible, in highly-contested MP situations

Cache hit location	HW generation	Nominal effective access latency (clocks or cycles)	Intervention overhead *if another core owns exclusive*
L1	z13..z17	4	-
L2	z13..z15	~10	-
	z16..z17	~19	-
L3 (on-chip)	z13	~36	~43
	z14	~45	~44
	z15	~45	~52
	z16..z17	~63	~27

L1..L3-on-chip cache latencies plus possible intervention overhead (z13..z17)

Optimization – Data Cache (for operands) (1)

- As explained in prior page, don't mix multiple distinct shared writeable data in the same cache line to avoid potential tug-of-war among multiple processors
  - Avoid putting multiple shared (and contested) locks in the same cache line
- Avoid using any storage element as a running variable that will get fetched and updated many times in close proximity
  - Consider using a general register (GR) instead
  - Similarly, avoid spill and fill through storage within a short number of instructions
- NIAI may be used to provide hints to the hardware about intentions of storage accesses to avoid delays from potential cache state changes
- Data Prefetch instructions with both prefetch and untouch functions are provided
  - For cache lines that are contested among many processors, it might not be desirable to prefetch the cache line ahead of time since it may add unnecessarily data movement in the system causing extra delays
- L1 cache access pipeline (from issue of data fetch to issue of dependent instruction) is currently 4 cycles
  - scheduling non-dependent operations in-between storage accesses and subsequent usages allows maximum parallel processing

Optimization – Data Cache (for operands) (2)

- The hardware has mechanisms to detect store-to-load dependencies and to provide substantial bypass capabilities (known as “store forwarding”) which improve with every generation, minimizing storage access dependencies can yield better performance
  - In general, simple store and load instructions are handled well while more complicated instructions or address overlaps may
    observe more pipeline rejects
  - On z13+, a generic temporary store data buffer is used to bypass or “forward” pending store data to a dependent load. However, the following cases are not bypassable-from:
    - XC of > 8 bytes with perfect or “exact” overlap (clears storage with zeros)
    - MVC of > 8 bytes with 1-byte or 8-bytes of destructive overlap (pads storage with one or eight characters)

      In the event of a loop where the final MVC 1-byte destructive overlap instance in the current iteration feeds the first instance in the next…

      If forwarding is desired in a clearing case, replace MVC-1-byte overlap with XC-exact overlap

      If forwarding is desired in a store-padding case, insert an MVI to feed the MVC-1-byte overlap
    - Due to out-of-order handling of stores, multiple close-by store updates to the same doubleword location with dependent fetches in-between (sometimes observed due to loops)

Store Forwarding Processor Design Comparison

zEC12's store queue (STQ) w/ store forwarding buffer (SFB) vs. z13+'s STQ plus 4-set-associative store forwarding cache (SFC)

Feature	zEC12	z13+
Can stitch together the data to be forwarded from multiple SFB/SFC entries and/or the L1 cache	Yes	Yes
Footprint	16 DWs	128 DWs
Bytes forwardable (potentially)	Final DW	All
SMT capable	No	Yes
Susceptible to younger databeats executing OOO and preventing bypassing to older loads by even older databeats	No	Yes
Can forward from XC-exact and MVC 1-or-8-byte overlap	Yes	No
4-set-associative design allows at most 4 in-order databeats to the same 56:60 before overwriting an older entry**	**No**	**Yes**

- z13+'s SF design is superior in most scenarios, with a notable exception being the final category
  - Moral of the story: *****don't do a flurry of stores to the same doubleword!*** ***Use registers!!***

IBM Z Accelerator Summary

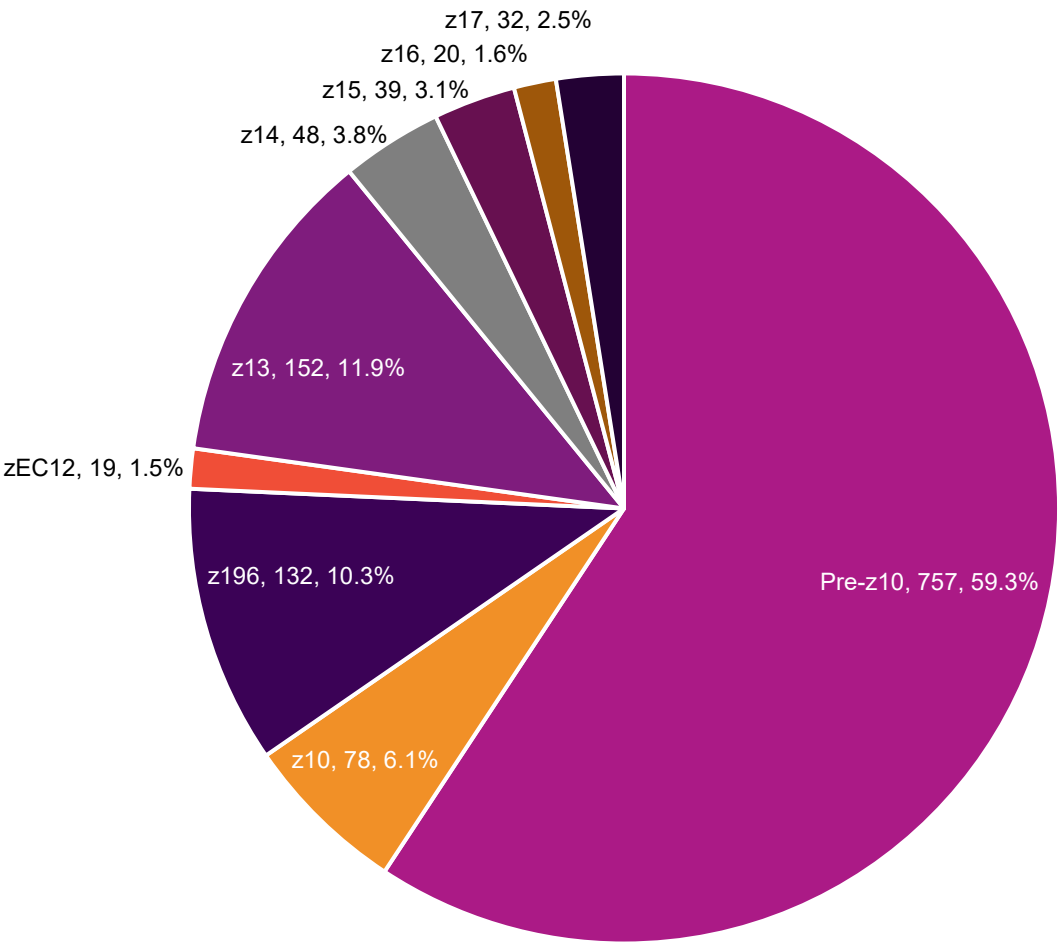
- Core-level**
  - z10
    - COP (co-processor) unit that accelerated compression and cryptography

      *The COP was shared by two cores** on z10 and z196 before going fully private on zEC12*

      The COP handled Unicode conversions starting on zEC12

      In SMT2 mode starting with z13 the COP handles one thread at a time, where the second waits until the COP insn or a unit of op of the COP insn has completed
  - z13
    - SIMD (single instruction multiple data) accelerator; new execution unit + VRs (vector registers) + instructions
  - z15
    - MA (modulo arithmetic) unit for elliptic curve cryptography
    - SORTL (sort/merge lists) accelerator within the COP
- Chip-level
  - Long, long ago
    - HAE “hardware accelerator engine” a.k.a. the “page mover” or “page-move/pad facility”
  - z15
    - NXU (nest accelerator unit) for G(UN)ZIP inflate/deflate operations that effectively moved zEDC on chip
  - z16
    - AIU (artificial intelligence unit) that facilitates dual 8x8 matrix operations common in NNP (neural network processing) algorithms
  - z17
    - DPU (data processing unit) that facilitates former PCIe IO management function

IBM Z Instructions per Generation, ~1277 total



HW/SW synergy is essential in this post-significant-frequency-growth era of computing

- *"Moore's Law is not dead yet, though we can see the checkered flag on the GPS" – Dave Hutton*
- **Keep current on the latest instruction set architecture (ISA)!!**
  - Real world hand-written assembler example – *from **storage-based** variable manipulation to purely **register-based**:*

```
ST    12,FLOATJSR              ; STORE SIGN BIT
NI    FLOATJSR,X'80'           ; RETAIN SIGN BIT
OI    FLOATJSR,X'4E'           ; EXPONENT
XC    FLOATJSR+1(3),FLOATJSR+1
LPR   12,12                    ; GENERAL REGISTER MADE POSITIVE
ST    12,FLOATJSR+4
AD    0,FLOATJSR               ; NORMALIZED FLOATING POINT NUMBER
```

→

```
RISBG 12,12,0,0,32     ; COPY SIGN BIT
NIHF  12,X'80000000'   ; RETAIN SIGN BIT
OIHF  12,X'4E000000'   ; EXPONENT
LPR   12,12            ; GENERAL REGISTER MADE POSITIVE
LDGR  12,12            ; GR12 -> FPR12
ADR   0,12             ; NORMALIZED FLOATING POINT NUMBER
```

  - **SIMD for packed decimal (COBOL) operations available since z14**
- The latest compilers tune their scheduling to the latest processor's specific microarchitecture
  - If decode and dispatch are 3-uops wide, the compiler will attempt to schedule 3-uop groups
  - If multiple instruction combination solutions exist to perform a specific action, the compiler will choose the optimal sequence for a given processor generation
  - Modern compilers separate instructions from data to avoid SIIS
- If you have handwritten assembler, the onus to remain current is on you, *as it has always been*
- ***The consequences of not remaining current are much higher now than they were in the past***

Frequently Asked Questions (1)

- Question:
  - What should be used to move or clear large blocks of data?
- Answer:
- There are several ways to move or clear a large block of storage provided in the z/Architecture
  1. One MVCL instruction
  2. Loops of MVCs to move data
  3. Loops of MVC <Len>,<Addr>+1,<Addr> or XC <Len>,<Addr>,<Addr> to pad/clear an area
- As discussed in “MOVE LONG instructions”,
  - MVCL is implemented through millicode routines
  - Millicode is a firmware layer in the form of vertical microcode
    - Incurs some overhead in startup, boundary/exception checking, and ending
  - MVCL function implemented using loops of MVCs or XCs
  - Millicode has access to special near-memory engines that can do page-aligned move and page-aligned padding
    - Can be faster than dragging cache lines through the cache hierarchy
    - However, the destination data will NOT be in the local cache
- As such, the answer is “it depends” as there is no one answer to all situations. There are many factors to consider
  - Will the target be needed in local cache soon?
    - Then moving/padding “locally” will be better by using MVCs or XCs
  - Is the source in local cache?
    - Then moving/padding “locally” may be better by using MVCs, or XCs
  - How much data is being processed?
    - The more data you are required to process, the more you may benefit from using MVCL due to special hardware engines used by millicode
  - Experimentation is, therefore, highly advised

Frequently Asked Questions (2)

- Question:
  - The "**Atomic and Locking Instructions**" chart recommends the sequence:

```
LOOP   LT  R1, lock ; load from memory and test value; always test first
       BCR  14,0    ; serialization to architecturally guarantee getting new value
       JNZ LOOP     ; repeat if non-zero
```

  …and the later "**Serialization**" chart shows a sequence

```
CPU 2
G   CLI  A, X'00'
    BNE  G
```

  with the comment: *"A serializing instruction must be in the CPU-2 loop to ensure that CPU 2 will again fetch and refresh data value from location A".*
  We have code that does not perform the serializing instruction and it works just fine. Can you explain how this works, and whether the serialization instruction is really needed?

- Answer:
  - From an architecture point of view, the serialization is needed to guarantee/ensure updated values will be used on both cases
  - On the other hand, *the serialization is not "needed" in practice* because all recent designs will always reacquire the cache line eventually. Such behavior is due to our hardware cache protocol and the presence of both architectural and micro-architectural interruptions. So, even without the serializing instruction, the two examples will not loop indefinitely.
  - However, *future hardware implementations may not work the same way.* Although the IBM designs ensure that an update to a location will eventually be seen by all processors without the proactive serialization, relying on the design behavior can be a potential performance issue depending on when the hardware refreshes itself.
  - On all recent processors, the serialization instruction is normally executed as a fast single-cycle instruction, so there is negligible performance impact to including it in the code. "Proactive" serialization should be added whenever necessary.

Frequently Asked Questions (3)

- Question:
  - I have stack pointer update code that I'm considering changing from L/AR/ST to LAA, but I see in POPs that LAA does a specific-operand-serialization function, presumably to support block-concurrent interlocked-update. However, since my stack cache lines are guaranteed to be unique per CP (unique 8K-byte blocks per CP with no overlap) I really don't need the interlocked-update. **Will the LAA unnecessarily cause cache messaging or the cache line to be pushed out to memory and thus make the single instruction slower than the L/AR/ST trio?**
- Short answer: **no**
- Longer answer:
  - Among all recent implementations, the enforcement of serialization (or atomic update) is done with minimal performance overhead
    - The handling for special-operand-serialization is no different than normal operand access ordering/coherency and should not lead to any additional overhead that can be observable
  - The only time serialization / atomic overhead can be observed is when the involved cache line has real-time contention among multiple processes (on multiple processors)
    - With high contention, the pipeline may run in a slow mode to adhere to the serialization/atomic requirement

      Since LAA is one storage-access instruction vs. L/AR/ST having 2 storage-access instructions, LAA will likely be faster
    - Without contention, LAA is expected to be a bit faster because it is 1 instruction (vs. 3) and the pipeline will handle it seamlessly.
  - Therefore, LAA is preferred to sequence of 3 instructions
- Additional note:
  - In general, the z processors are built to optimize single **"storage + arithmetic/logic"** operations through a "dual issue" design, such that instructions like "A", "S", "OI", "NI" are handled very efficiently
    - These single instructions should be used over multiple simple instructions

Frequently Asked Questions (4)

- Question:
  - What instruction is best to use in setting registers to zeros?
- Answer:
- For most cases (in real life code), the differences between SR/XR vs. LHI or SGR/XGR vs. LGHI will not be noticeable
- Obtaining the ultimate performance depends on neighboring code, specifically on 1) code address alignment, 2) condition code (CC) dependencies and usage, and 3) register dependencies and usage
- To expand on the rule of thumb:
  1. If one doesn't want to think too hard, use L(G)HI. L(G)HI is generally good as one does not need to worry about register or condition code dependencies. L(G)HI doesn't write the CC, which is a potential plus in terms of processor resource usage. The only downside is its instruction length being a little longer. If someone would want to optimize further, X(G)R or S(G)R can be used instead to relieve instruction fetching limitations which leads to #2
  2. If someone wants to think a little more, then the recommendation is to optimize for instruction path-length unless there is a dependency conflict. Thus,
     - If one is doing a 32/64-bit GR clear and doesn't care about the CC or would prefer the CC to be cleared, use X(G)R or S(G)R
     - If you're doing a 32/64-bit GR clear and cannot destroy the CC, use L(G)HI
     - If the GR to be cleared was used near-by (similar to CC concern), use L(G)HI to avoid register dependencies
  3. For extreme optimization, one will have to understand the processor pipeline a lot more in terms of instruction grouping, register and condition code buffer size, instruction alignment, etc.
     - All these considerations can get very complicated and convoluted. So, it is best to leave to the compiler
- Note that for writing zeros to FPRs, LZXR or LZDR can be used, while for VRs, VGBM w/ I2=0 can be used

Frequently Asked Questions (5)

- Question:
  - For operand data access, what hardware prefetching is available and how can you best use the PREFETCH DATA instruction as a form of software prefetching?
- Answer:
- Since z196, the z processor pipeline is out of order. As each generation improves upon the out of order window, the pipeline inherently overlaps storage accesses that are independent from each other; depending upon instruction fetching/dispatching, ”prefetching” is performed whenever out of order processing can ”get to it”
- Since z10, the z processor implements a hardware-based stride prefetch engine and its capability and accuracy has been improving on each subsequent system
  - It captures history of operand access addresses corresponding to a set of instruction address (PC) offsets
  - If it detects a striding pattern (+/- 1 to 2047) at a certain PC offset and determines the pattern has happened more than 2 times, it will start launching the next access in the form of a prefetch by predicting the next several striding accesses
  - For example, it will detect X, X+Y, X+2*Y at instruction address A (usually in a loop), and then launch hardware data prefetches at X+3*Y and X+4*Y; if it detects with more confidence, then it can prefetch up to X+5*Y (i.e., 3 strides ahead)
  - When the stride detected is small (less than a cache-line size), the HW prefetch engine can decide to be aggressive and prefetch the next cache line
- Since the HW prefetch engine is built to handle stride-based prefetching within loops, it is advisable not to insert additional line-based PFD instructions into the code since it might become redundant and potentially incur unnecessary overhead
- Software prefetching using line-based PFD instructions is best done far ahead of the actual load
  - An instruction distance of at least half of the out of order window, i.e., > 40 instructions for z13 onward, to see potential benefits

Frequently Asked Questions (6)

- Question:
  - During “Optimization – Insn/Data Placement”, Classic save area is mentioned, can you elaborate more?
  - I have some code which frequently references a fixed lookup table that's near the executable code. Is there any problem of having the same cache line read-only in both the I-cache and D-cache?
- Answer:
- The "classic save area" refers to subroutine linkage for s/360 assembler in which the caller provides an area into which the callee can save GPRs; it dates from the "classical" era before re-entrant programming, linkage stacks, etc., when this method was standard practice, and the save area often ended up being allocated right after the caller's code so it often ended up being in the same cache line as some of the code - hence the Store Into Instruction Stream (SIIS) inefficiency.
- No performance concern is expected with read-only copies of the same cache line in both the instruction and data caches. The SIIS inefficiency occurs when the processor detects the same line is in both the instruction and data caches **and** the data cache's copy is potentially to be updated (including any conditional paths not expected to be executed), at which point an expensive cache synchronization action is needed. So long as both copies of the line in the instruction and data caches remain identical, the synchronization action does not occur, and there should be no performance penalty.

Frequently Asked Questions (7)

- Question:
  - Can you discuss MVCL vs. MVCLE?
- Answer:
- As defined in z/Architecture, MVCL handles up to 24-bit lengths while MVCLE handles up to 32-bit lengths in 24- or 31-bit mode, and 64-bit lengths in 64-bit mode.
- Both instructions are implemented by millicode. The millicode implementation of MVCL and MVCLE are similar. Both use the special engine, when possible, as discussed earlier in “MOVE LONG Instructions”.
- MVCL is interruptible. Without any interrupts, MVCL can process through the full length in one pass.
- MVCLE, on the other hand, is not interruptible. In z/Architecture, “*The amount of processing that results in the setting of condition code 3 is determined by the CPU on the basis of improving system performance, and it may be a different amount each time the instruction is executed. The maximum amount is approximately 4K bytes of either operand*”.
- Therefore, MVCLE takes a mandatory break after every 4-Kbyte move as defined by the architecture, to allow for interrupt handling whether one is pending or not. A typical code sequence will include a condition code 3 check to loop back to the same MVCLE. This break thus includes the overhead of exiting and then re-entering millicode again. Meanwhile, MVCL will only take such a break if an interrupt is pending, and the checking for pending interrupt inside millicode is very minimal. If there are no interrupts pending on this CPU, MVCL proceeds to the next 4K move without exiting millicode.
- MVCL has the advantage of being much faster if the CPU isn't the target of lots of interrupts. If the CPU is being interrupted frequently, then MVCL is no faster than MVCLE.

Frequently Asked Questions (8)

- Question:
  - What is the pipeline penalty with address mode changing instructions?
- Answer:
- In z/Architecture, SET ADDRESS MODE instructions (SAM24, SAM31 and SAM64) can be used to change the addressing mode
- When any of these instructions doesn’t change the addressing mode, i.e., executing SAM31 when we're already in 31-bit mode, it will not incur any penalty. The pipeline will recognize that there is no change and not do anything.
- However, if any of these instructions changes the addressing mode, it will cause a front-end pipeline restart like a taken relative branch that is not predicted by the branch prediction logic.

Frequently Asked Questions (9)

- Question:
  - How do I debug or tune application/system performance? What counter information will be useful in relations to some of the topics discussed in this document?
- Answer:
- This document is not intended to provide information about how to debug performance issues or tuning applications
- Please refer to other IBM documentation and consult with our support / client care teams for deeper investigations
- However, let us know what information will help you. If we see a strong need for additional documentation, we are open to suggestions.

References

1. "z/Architecture: Principles of operation," Int. Bus. Mach. (IBM) Corp., Armonk, NY, USA, Order No. SA22-7832-14, June 2025. https://www.ibm.com/docs/en/module_1678991624569/pdf/SA22-7832-14.pdf
2. M. Farrell et al, "Millicode in an IBM zSeries processor," IBM J. Res. & Dev., vol. 48, no. 3/4, pp. 425–434, 2004.
3. C.F. Webb, ''IBM z10: The Next-Generation Mainframe Microprocessor,'' IEEE Micro, vol. 28, no. 2, 2008, pp. 19-29.
4. C Shum, "Design and microarchitecture of the IBM System z10 microprocessor," IBM J. Res.& Dev., vol. 53, no. 1, 2009, pp. 1.1-1.12.
5. Brian W. Curran et al, "The zEnterprise 196 System and Microprocessor," IEEE Micro, vol. 31, no. 2, 2011, pp. 26-40.
6. F. Busaba et al, "IBM zEnterprise 196 microprocessor and cache subsystem," IBM J. Res.& Dev., vol. 56, no. 1/2, pp. 1:1–1:12, Jan./Feb. 2012.
7. K. Shum et al, "IBM zEC12: The third-generation high-frequency mainframe microprocessor," IEEE Micro, vol. 33, no. 2, pp 38–47, Mar./Apr. 2013.
8. Bonanno et al, "Two Level Bulk Preload Branch Prediction", HPCA, 2013
9. C. Jacobi et al, "Transactional Memory Architecture and Implementation for IBM System z," IEEE/ACM Symposium on Microarchitecture (MICRO), 2012.
10. B. Curran et al, "The IBM z13 multithreaded microprocessor," IBM J. Res. & Dev., vol. 59, no. 4/5, pp. 1:1–1:13, 2015.
11. E. M. Schwarz et al, "The SIMD accelerator for business analytics on the IBM z13," IBM J. Res. & Dev., vol. 59, no. 4/5, pp. 2:1–2:16, 2015.
12. B. Prasky et al, "Software can Provide Information Directly to the System z Microprocessor", IBM Systems Magazine, May 2014
13. C. Walters et al, "The IBM z13 processor cache subsystem", IBM J. Res. & Dev., vol. 50, no. 4/5, pp. 3:1-3:14, 2015

References (Cont'd)

14. C. Jacobi et al, "The IBM z14 microprocessor chip set", IEEE Hot Chips 29, 2017
15. John R. Ehrman's book on IBM Z system assembly programming. (Web link as of July 2021, may change by Marist) https://idcp.marist.edu/documents/33945/44724/Assembler.V2.alntext+V2.00.pdf/
16. A. Saporito et al, "Design of the IBM z15 microprocessor", IBM J. Res. & Dev., vol. 64, no. 5/6, pp. 7:1-7:18, 2020
17. N. Adiga et al, "The IBM z15 High Frequency Mainframe Branch Predictor", ISCA, 2020
18. C. Jacobi et al, "Real-time AI for Enterprise Workloads: the IBM Telum Processor", IEEE Hot Chips 33, 2022
19. z16: https://hc33.hotchips.org/assets/program/conference/day1/HC2021.C1.3%20IBM%20Cristian%20Jacobi%20Final.pdf
20. C. Lichtenau et al, "AI Accelerator on IBM Telum Processor", ISCA '22, June 18–22, 2022, New York City, NY
21. z17: https://hc2024.hotchips.org/assets/program/conference/day1/04_HC2024.IBM.CBerry.final.pdf
22. C Berry et al, "The IBM Telum II Processor" https://ieeexplore.ieee.org/document/10980413
23. C Walters et al, "Enterprise Class Modular Cache Hierarchy" https://ieeexplore.ieee.org/document/10946804
24. Parziale et al, "AI on IBM Z: Applications and Examples" IBM Redpaper https://www.redbooks.ibm.com/redpieces/pdfs/redp5758.pdf
25. D. Berger et al, "Enterprise Class On-Chip Accelerator Integration" https://2026.hpca-conf.org/program/program-hpca-2026/

THANK YOU

Suggestions, questions, comments:

hutton@us.ibm.com

https://www.linkedin.com/in/david-hutton-a9ab2a1/

z/TPF

VSEn

APPENDIX

z196 through z17 Topologies

z196
L1 private 64k i + 128k d
L2 private 1.5 MB
L3 shared 24 MB per CP chip
L4 shared 92 MB per book
4 cores + 1 L3 per CP chip
6 CP chips + 1 L4 per **book**
4 **books** (STAR) per CEC

zEC12
L1 private 64k i + 96k d
L2 private 1 MB i + "L1+" 1 MB d
L3 shared 48 MB per CP chip
L4 shared 384 MB per **book**
6 cores + 1 L3 per CP chip
6 CP chips + 1 L4 per **book**
4 **books** (STAR) per CEC

z13
L1 private 96k i, 128k d
L2 private 2 MB i + 2 MB d
L3 shared 64 MB per chip
L4 shared 480 MB per **node**
8 cores + 1 L3 per CP chip
3 CP chips + 1 L4 per **node**
2 **nodes** per **drawer**
4 **drawers** (NUMA) per CEC

z14 / z15
L1 private 128k i, 128k d
L2 private 2 / 4 MB i, 4 MB d
L3 shared 128 / 256 MB per chip
L4 shared 672 / 960 MB per **drawer**
10 / 12 cores + 1 L3 per CP chip
3 / 2 CP chips per **cluster**
2 **clusters** + 1 L4 per **drawer**
4 / 5 **drawers** (numa) per CEC

z16 / z17
L1 private 128k i, 128k d
L2 private 32 MB / 36 MB unified
virtual victim L3 up to 7x32 = 224 MB per CP chip / 9 x 32 = 324 MB per CP chip
virtual victim L4 up to 8x32x7 = 1.75 GB per drawer / 10x36x7 = 2.52 GB per drawer
8 (core + 8 L3s / 10 L3s) + 0 DPU / 1 DPU per CP chip
2 CP chips / DCM
4 DCMs (64 engines) / drawer
4 drawers / CEC