The IBM z14 microprocessor chip set and architectural enhancements

Jonathan Bradbury
Senior Engineer
Lead ISA Architect, IBM Z
IBM Poughkeepsie, NY

Trademarks

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

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a more complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*BladeCenter®, CICS®, DataPower®, DB2®, e business(logo)®, ESCON, eServer, FICON®, IBM®, IBM (logo)®, IMS, MVS, OS/390®, POWER6®, POWER6+, POWER7®, Power Architecture®, PowerVM®, PureFlex, PureSystems, S/390®, ServerProven®, Sysplex Timer®, System p®, System p5, System x®, z Systems®, System z9®, System z10®, WebSphere®, X-Architecture®, z13™, z13s ™, **z14 ™**, z Systems™, z9®, z10, z/Architecture®, z/OS®, z/VM®, z/VSE®, zEnterprise®, zSeries®, IBM Z ®

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

*** All other products may be trademarks or registered trademarks of their respective companies.**

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured Sync 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 Sync 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.

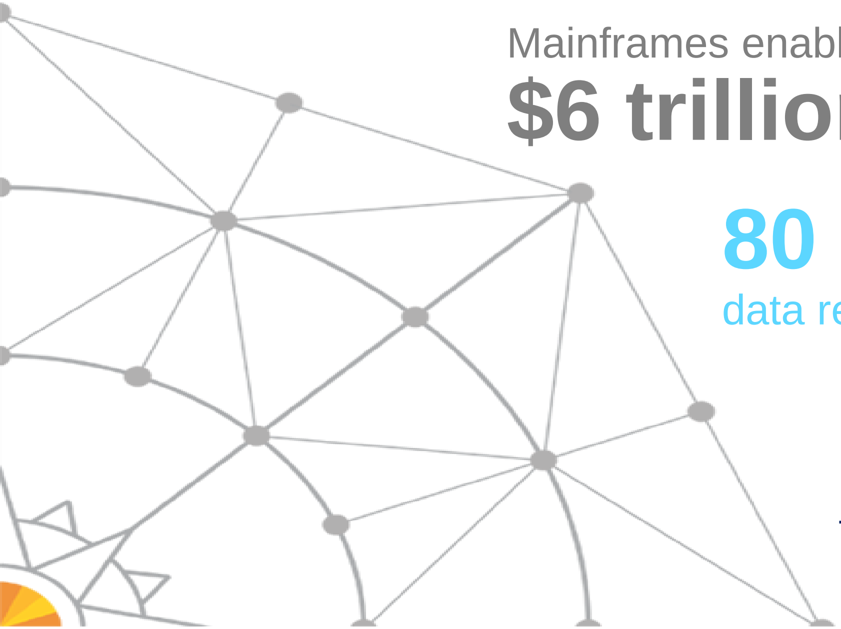
The mainframe is everywhere, making the world work better

Mainframes process
30 billion business transactions per day

Mainframes enable
$6 trillion in card payments annually

80 percent of the world's corporate data resides or originates on mainframes

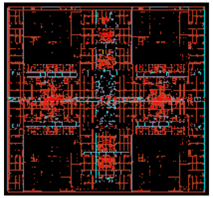
91 percent of CIOs said new customer-facing apps are accessing the mainframe



IBM Z – Processor Roadmap

65 nm

z10
2/2008



Workload Consolidation and Integration Engine for CPU Intensive Workloads

Decimal FP

Infiniband

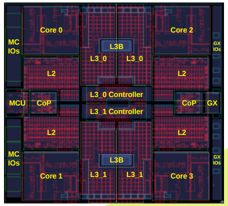
64-CP Image

Large Pages

Shared Memory

45 nm

z196
9/2010



Top Tier Single Thread Performance,System Capacity

Accelerator Integration

Out of Order Execution

Water Cooling

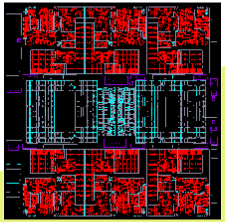
PCIe I/O Fabric

RAIM

Enhanced Energy Management

32 nm

zEC12
9/2012



Leadership Single Thread, Enhanced Throughput

Improved out-of-order

Transactional Memory

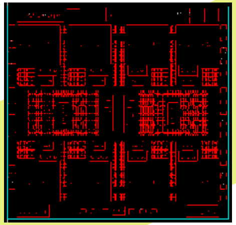
Dynamic Optimization

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

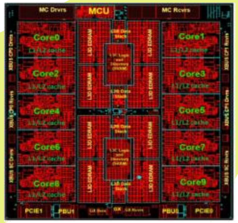
SIMD

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

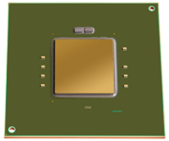
New SIMD instructions

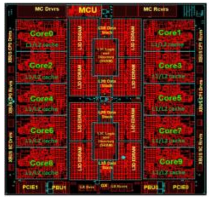
Optimized pipeline and enhanced SMT

Virtual Flash Memory

z14 Processor Chipset & Drawer Design

CP SCM

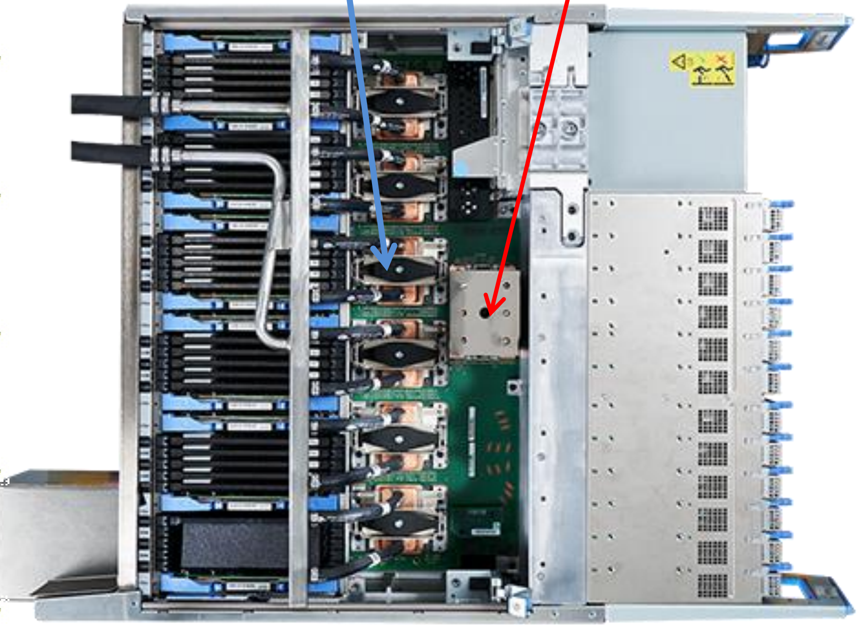




6x CP SCMs

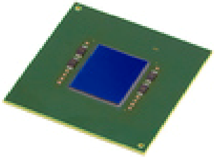
1x SC SCM
(Air Cooled)

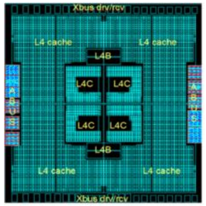
6x CP SCMs
under the cold-plates



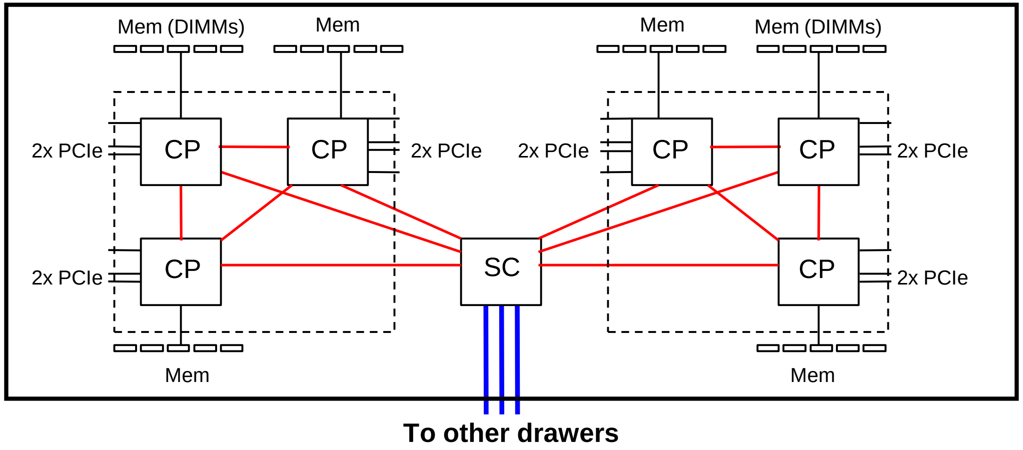
Capped
SC

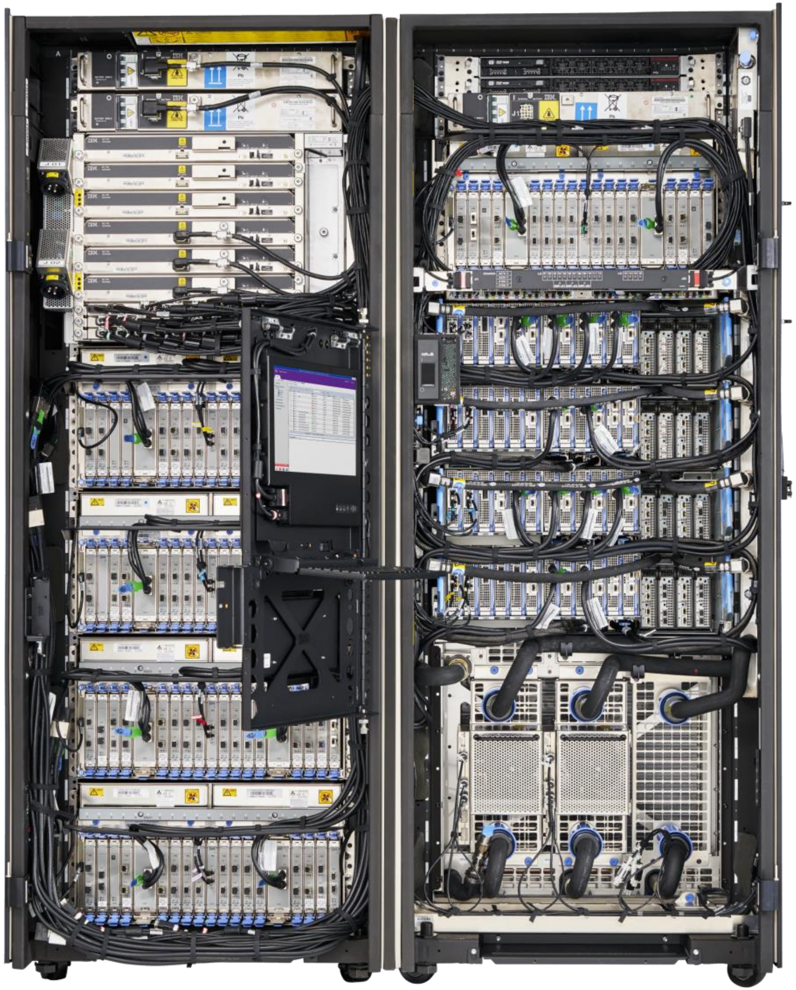
SC Chip





z14 On-Drawer and System Topology





CP chip, 696 sqmm, 14nm, 17 layers of metal
- 10 cores, each 2+4MB I+D L2 cache
- Shared 128MB L3 cache

SC chip, 696 sqmm, 14nm, 17 layers of metal
- System interconnect & coherency logic
- Shared 672MB L4 cache

Max System:
- 24 CP sockets in SMP interconnect
- 32TB RAIM-protected memory
- 40 PCI gen3x16 fanouts to IO-drawers
- 320 IO cards

z14 processor design summary

Micro-Architecture

- 10 cores per CP-chip
- 5.2GHz

- Cache Improvements:
  - 128KB I$ + 128KB D$
  - 2x larger L2 D$ (4MB)
  - 2x larger L3 Cache
  - symbol ECC

- New translation & TLB design
  - Logical-tagged L1 directory
  - Pipelined 2nd level TLB
  - Multiple translation engines

- Pipeline Optimizations
  - Improved instruction delivery
  - Faster branch wakeup
  - Improved store hazard avoidance
  - 2x double-precision FPU bandwidth
  - Optimized 2nd generation SMT2

- Better Branch Prediction
  - 33% Larger BTB1 & BTB2
  - New Perceptron & Simple Call/Return Predictor

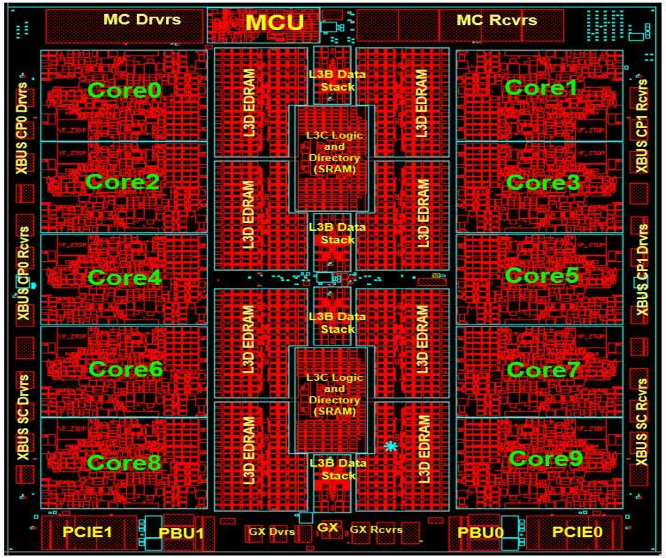
Architecture

- PauseLess Garbage Collection
- Vector Single & Quad precision
- Long-multiply support (RSA, ECC)
- Register-to-register BCD arithmetic

Accelerators

- Redesigned in-core crypto-accelerator
  - Improved performance
  - New functions (GCM, TRNG, SHA3)

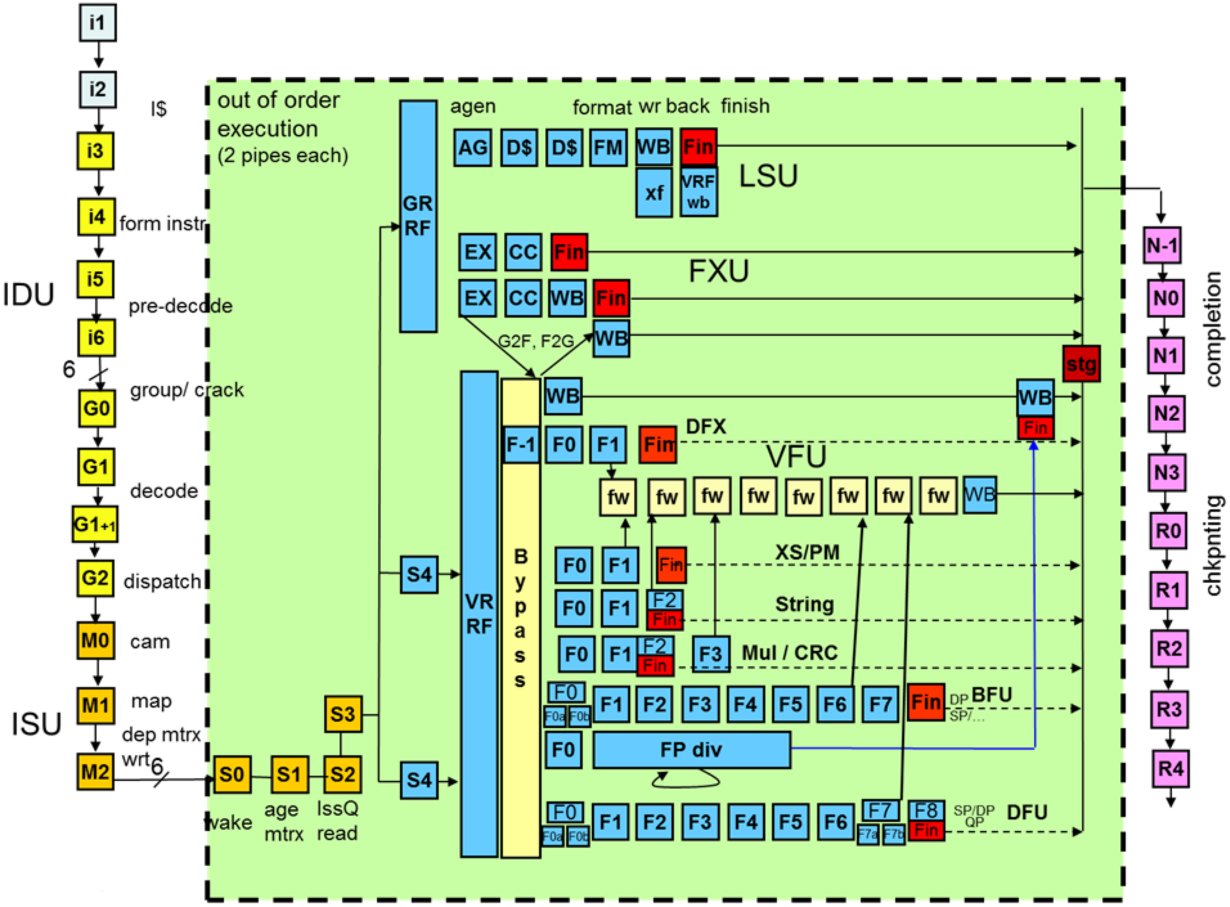
- Optimized in-core compression accelerator
  - Improved start/stop latency
  - Huffman encoding for better compression ratio
  - Order-preserving compression



z14 Pipeline

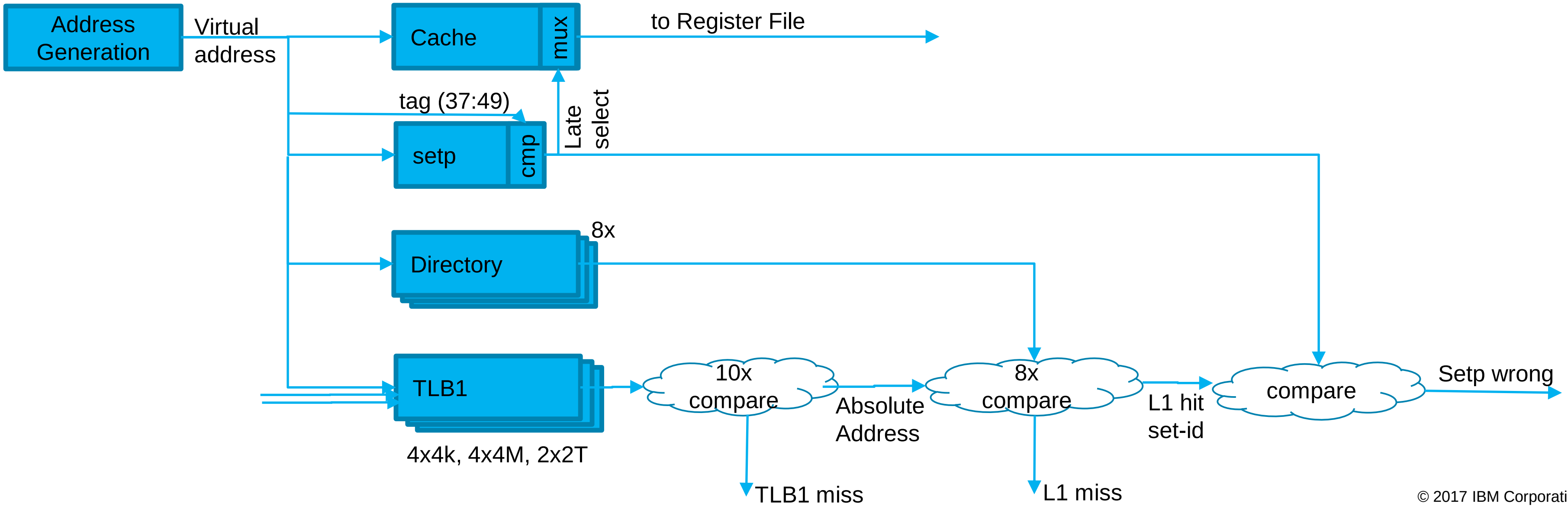
Deep high frequency pipeline

- Async branch prediction ahead of ifetch
- 32B/cycle ifetch
- 6 instruction / cycle parse & decode
- CISC instruction cracking
- Unified OOO issue queue
- 2 LSU, 4-cycle load-use
- 4 FXU, 2 SIMD/FP/BCD
- In-order completion & checkpoint



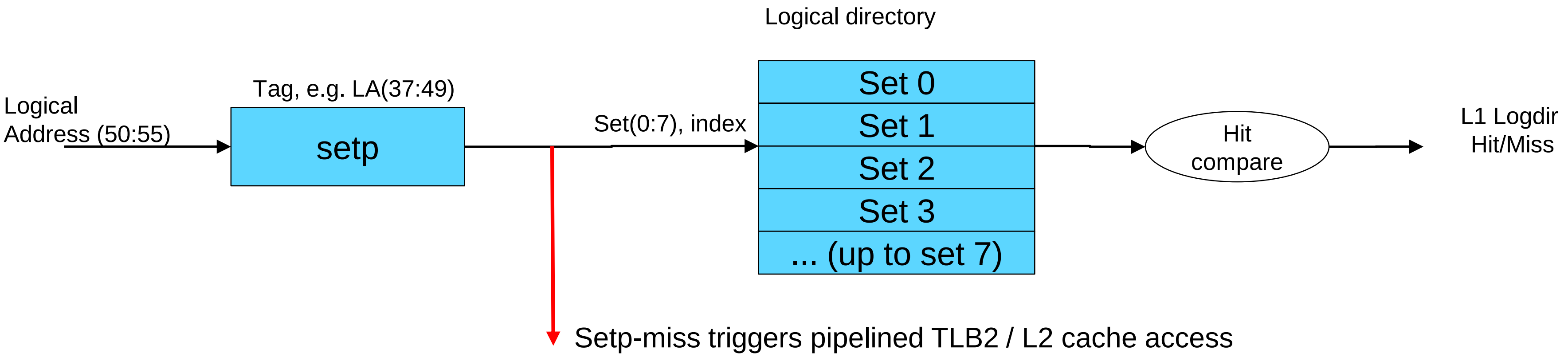
Traditional L1 cache directory

- Traditional cache design employs logical-indexed, absolute tagged directory
  - Use of partial compare set-predict array reduces latency of data return from L1 cache
  - TLB access and L1 directory access / compare happen in parallel with L1 cache read
  - Absolute-address based miscompare drives setp-correction and re-drive of the instruction
- Highly associative TLB and directory structures are area and power inefficient
  - Limited TLB1 size, in turn limits performance gain from growing L2 and L3 caches
  - Number of cache ports is limited



L1 logical tagged directory

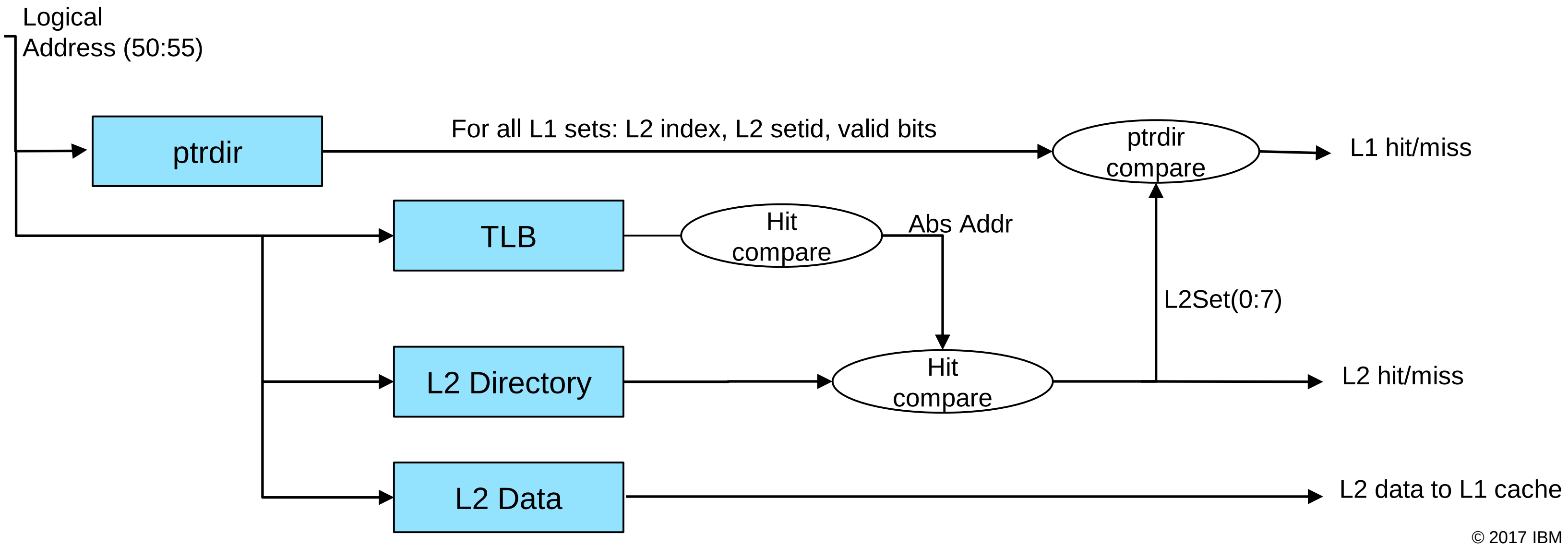
- I$ and D$ now use logically tagged directory
- Effectively combining TLB1 and cache directory into single structure
  - Directory entry is wider (address space ID + virtual address, versus absolute address only)
  - Set-predict is used to read a single directory entry for each Load/Store micro-op
    - 8-way directory is implemented as a single, 8x deep directory; set-predict output is used as row-addressed
  - Directory now part of local & global TLB purge operations (e.g. during OS paging)
- Significant area & power reduction for L1 hit



Setp-miss triggers pipelined TLB2 / L2 cache access

Integrated TLB2 & L2 cache pipeline

- Fixed-duration pipeline for TLB2 and L2 cache invoked on L1-setp miss
- L2 and TLB2 can scale to be very large
  - Power & area efficient because of single-ported design and only accessed on L1 miss
  - 2MB L2I and 4MB L2D on z14, 6k entries TLB2 for 4KB pages
  - 8 cycle L2-hit latency (1.5ns)
- L1 pointer directory keeps track of cache lines in L1 even when logdir invalidated
  - Used for reload of translation without reloading data into L1, as well as logical address synonyms

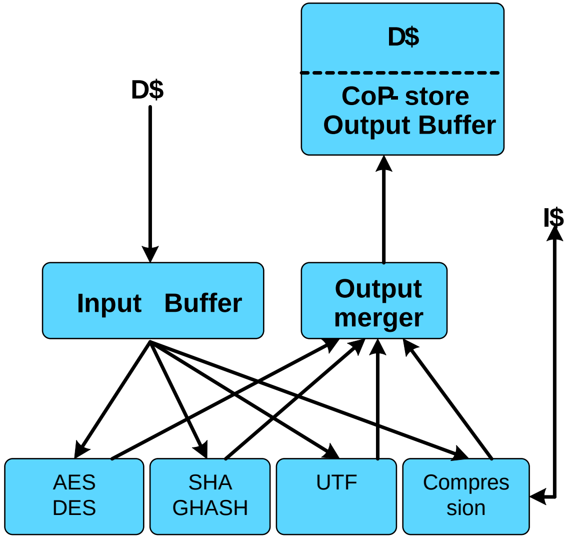


Crypto Accelerator

- IBM Z pervasive encryption reduces risk and auditing effort & cost
  - Pervasively encrypt data in flight and at rest with no application changes and no impact to SLAs
  - System-wide design optimized through silicon, firmware, OS, and middleware stack

- Redesigned crypto engines for 4-7x bandwidth vs z13
  - Pipeline and parallelize AES & Hash operations (GCM)
  - Execute 2 AES rounds in 3 cycles
  - Overlap multiple rounds where possible (ie. Non-CBC encryption)
  - Push limits of cycle time (lowVt)

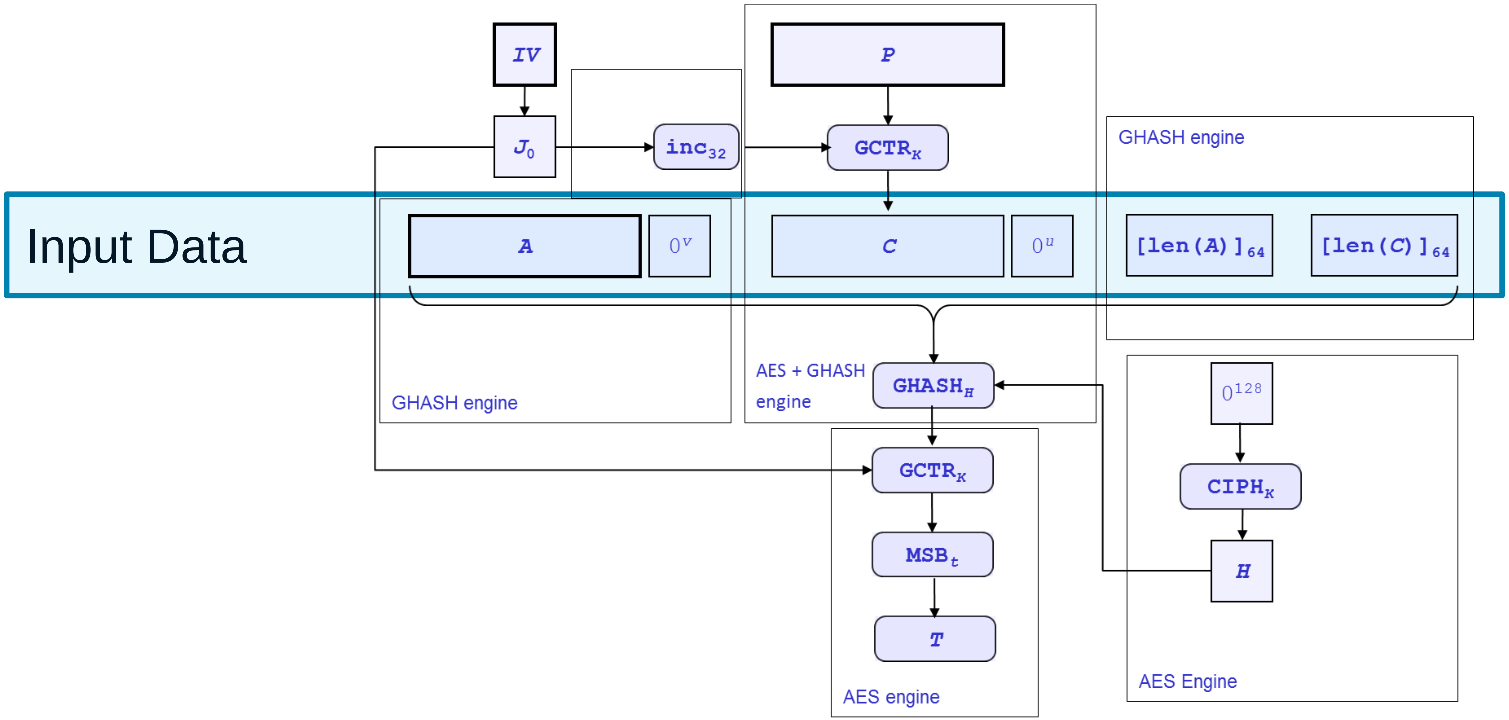
- Faster engines required redesign of interfaces to/from cache
  - New firmware instruction to copy up to 256B from D$ to Co-processor
  - Branch-avoidance to not slow down data delivery
  - Optimized prefetching for source & destination to not starve engine



- 13.2GB/sec per core in OpenSSL AES-256-XTS speed test with 4KB blocks

Galois Counter Mode (GCM)

- *AES-GCM* is frequently used mode to encrypt and authenticate messages
- z14 adds new instructions to directly implement AES-GCM algorithm
- Implemented as orchestration of AES and GHASH engines
- 12.5GB/sec AES-GCM-256 per core with 4KB blocks



Key protection

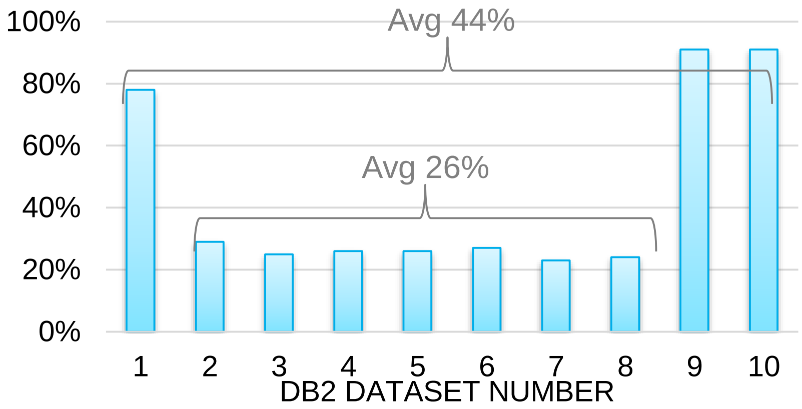
- Cryptographic encryption relies on the protection of the key
- Most processors support crypto accelerating instructions with the keys in user memory
  - When encrypting message from web browser to your bank a call to OpenSSL takes a clear text message and clear key to encrypt
  - Know as *clear key cryptography* in IBM Z
  - OS-admin, memory dump, core dump etc all pose risk of exposing key to adversaries

- CryptoExpress6S is a tamper-responding PCI crypto accelerator
- Holds a *master key* in physically protected memory on the card
- *Protected Key* cryptography wraps user-keys with master key
  - CPU crypto accelerator can interact with CryptoExpress to temporarily decrypt key to perform AES operation
  - Clear Key never exists in application or OS accessible memory
- *Secure Key* is another mode: all crypto operations directly performed on the card itself
  - Most secure environment but performance limited by PCI latency and bandwidth

Data Compression Accelerator

- IBM Z provides special instruction for dictionary-based data compression
  - Tailored towards short data (e.g. database rows) but employed broadly also for file & tape compression
  - Reduced storage cost, and improved performance (disk bandwidth, DB2 buffer pool efficiency, etc)
  - Implemented as firmware and co-processor specialized hardware
- z14 performance improvements in both start-up latency, and peak throughput
  - Optimized data load & store (same as crypto engine)
  - Optimized compression status return to firmware
  - Parallel search of dictionary for multiple symbols
- New architectural features to further improve data compression efficiency
  - Huffman Coding
  - Order Preserving Compression

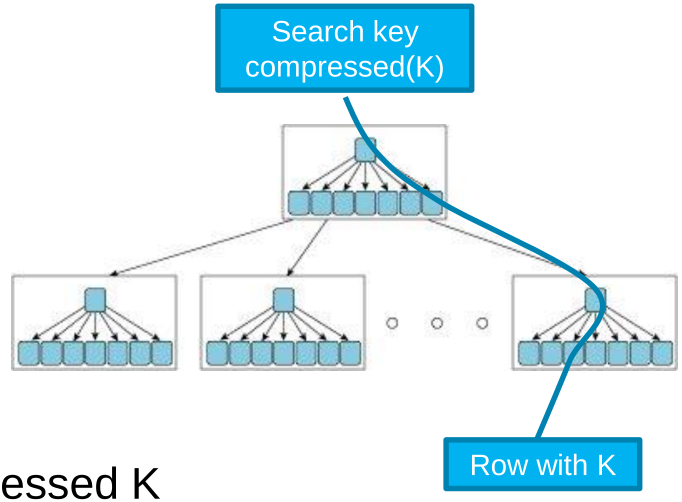
Disk reduction from Huffman Coding over Traditional Compression



Order-Preserving Compression

- Compression algorithm that ensures data can be compared <, =, >
  - Reduces compression efficiency slightly vs standard dictionary compression
  - Encodes symbols in a way that compressed data maintains ordering relation

    A < B   iff   compressed A < compressed B

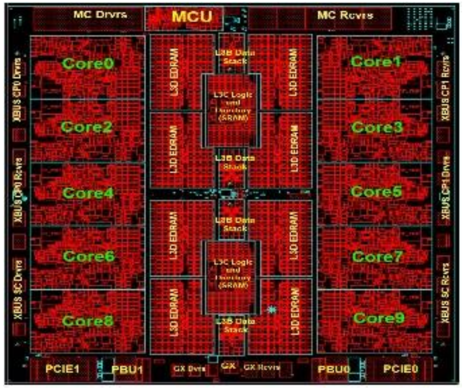


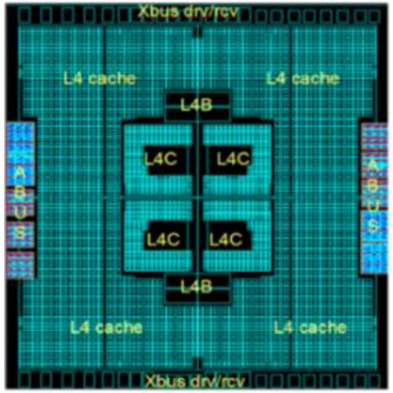
- Searching for key K is replaced with searching for compressed K
- Entries in DB2 index are stored in the new compressed format
  - Reduced disk space
  - Improved Buffer Pool efficiency
  - Improved Cache efficiency for Index searches

- Same technology for Sort Workfiles and other places where searchable data is stored

IBM z14 – designed for massive scale commercial workloads

- Processor Chip w/ L3 cache + System Control Chip w/ L4 cache
- 14nm SOI technology, 5.2GHz in water cooled enterprise server
  - CP: 6.1B transistors, 14 miles of wire
  - SC: 9.7B transistors, 13.5 miles of wire
- Up to 240 physical cores in 4-drawer shared-memory SMP
- 170 configurable customer CPUs, plus IO assist and firmware CPUs
- +35% capacity, +10% single thread / +25% SMT2 performance
- Micro-architectural and architectural enhancements for wide variety of workloads
  - BCD for Cobol, Garbage Collection for Java, Compression for Databases, …





z14 Selected New Instructions

- Vector Decimal Instructions
  - Used by COBOL and other software that operates on zoned or packed decimal data
  - Allows for operations to take place out of registers instead of storage
  - Lowers the latency for each instruction as well as avoids many hardware interlocks
  - New instructions:

New Instruction	Equivalent Existing Instruction	New Instruction	Equivalent Existing Instruction
VECTOR ADD DECIMAL	AP	VECTOR PACK ZONED	PACK
VECTOR COMPARE DECIMAL	CP	VECTOR PERFORM SIGN OPERATION DECIMAL	OI. NI, ZAP
VECTOR CONVERT TO BINARY	CVB/CVBG	VECTOR REMAINDER DECIMAL	DP
VECTOR CONVERT TO DECIMAL	CVD, CVDG	VECTOR SHIFT AND DIVIDE DECIMAL	
VECTOR DIVIDE DECIMAL	DP	VECTOR SHIFT AND ROUND DECIMAL	SRP
VECTOR LOAD IMMEDIATE DECIMAL		VECTOR SUBTRACT DECIMAL	SP
VECTOR MULTIPLY AND SHIFT DECIMAL		VECTOR TEST DECIMAL	TP
VECTOR MULTIPLY DECIMAL	MP	VECTOR UNPACK ZONED	UNPK

z14 Selected New Instructions

- General Instructions
  - Message Security Assists:
    - MSA6 – Enhances KIMD/KLMD to support the SHA-3 hash
    - MSA7 – Enhances PRNO (formerly PPNO) to support the generation on true random numbers
    - MSA8 – Adds KMA instruction and provides AES-GCM support in hardware
  - Miscellaneous General Instruction Extension 2
    - Adds new arithmetic instructions:
      - ADD HALFWORD (AGH)
      - MULTIPLY (MG, MGRK)
      - MULTIPLY HALFWORD (MGH)
      - MULTIPLY SINGLE (MSC, MSGC, MSGRCK, MSRKC)
      - SUBTRACT HALFWORD (SGH)
    - New Branch Instruction
      - BRANCH INDIRECT ON CONDITION

z14 Other Enhancements

- Multiple Epoch Facility
  - Will allow for the TOD clock to go past 2042
  - May still require software changes if using STCK/STCKF instead of STCKE
- Guarded Storage Facility
  - Used by runtime environments that use garbage collection (i.e. Java) to increase efficiency of the collection
- Configuration z/Architecture-architectural-mode facility
  - All LPARs now IPL directly into z/Architecture mode

THANK YOU

Jonathan Bradbury
jdbradbu@us.ibm.com

BACKUP

Glossary and Links

- CP – Central Processor Chip
- SC – System Control Chip
- SCM – Single Chip Module
- RAIM – Redundant Array of Independent Memory
  Meaney, P. J. et.al. "IBM zEnterprise redundant array of independent memory subsystem". *IBM Journal of Research and Development*, Vol 56, Issue 1.2, Jan-Feb 2012
- MC – memory controller
- LSU – Load Store Unit
- FXU – Fixed Point Unit
- Setp – Set Predictor (for associative caches)
- Logdir – logically tagged and indexed directory
- Crypto Express
  https://www-03.ibm.com/security/cryptocards/pciecc2/overview.shtml
- Details on Protected Key crypto:
  https://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100647