Smarter Systems for a
Smarter Planet

IBM zEC12 Processor Subsystem

The Foundation for a Highly Reliable, High Performance Mainframe SMP System

Robert Sonnelitter

System z Processor Development, Systems & Technology Group, IBM Corp.

rjsonnel@us.ibm.com

IBM zEC12 Processor Subsystem

- Historical Background
- Performance Characteristics
- CP/L3 Design Highlights
- SC/L4 Design Highlights
- System RAS Features

System z Shared Cache History

- 1990 – Fully shared second level cache
- 1995 – Cluster Shared L2
- 1997 – Distributed Cache Topology
- 1998 – Bi-Nodal Distributed Cache Design
- 2003 – Modular Nodal Design, Ring Topology
- 2008 – Three Level Cache Hierarchy, Fully Connected Topology
- 2010 – Four Level Cache Hierarchy, eDRAM Caches

1990	1993	1995	1994	1995	1996	1997	1998	1999	2000	2003	2005	2008	2010	2012
H2	**H5**	**H6**	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**z900**	**z990**	**z9**	**z10**	**z196**	**zEC12**
6w	6w	10w	6w	10w	10w	10w	10w	12w	20w	32w	54w	64w	80w	101w

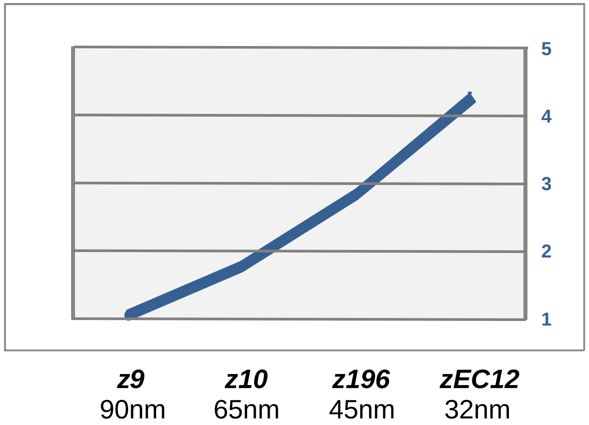
Customer Environment
& Workload Characteristics

- Highly virtualized workloads
  - Heavily shared system environment
  - Sustained high processor utilizations
  - Tasks dynamically dispatched across the system
- Large single image workloads
- High data sharing across processors
- Response time sensitive workloads
- Large memory footprint
- Extremely high system reliability
- zOS, zVM, zVSE, TPF, zLinux operating systems

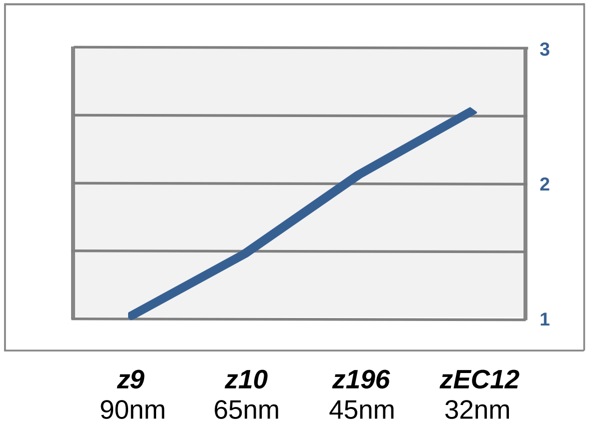
Performance Benchmarks

- Ensure Per-Thread and SMP Performance growth with increased system capacity
- Guaranteed customer performance targets with constant software
- Workloads
  - Large System Performance Reference (LSPR)
    HIDI / MIDI / LODI
    CB-L / WASDB / OLTP / etc.
  - Internal Custom Stressors
  - External Benchmarks
  - Only LSPR metrics are published

System Capacity



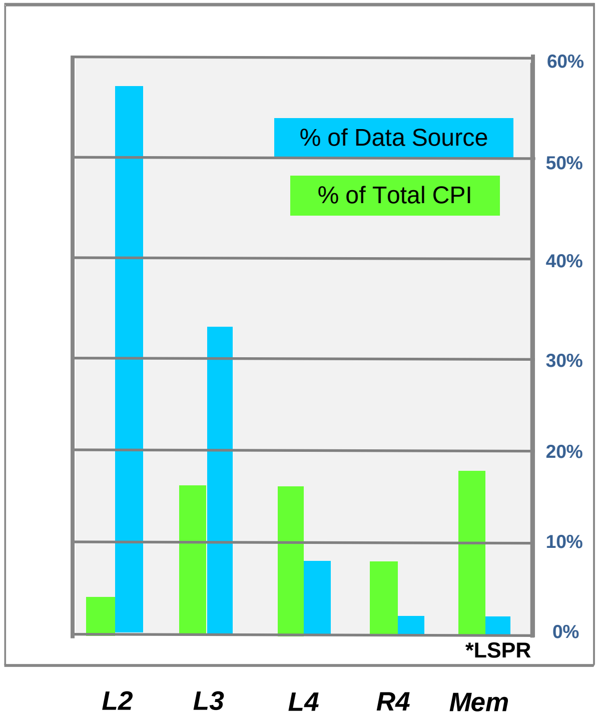
Per-Thread Performance



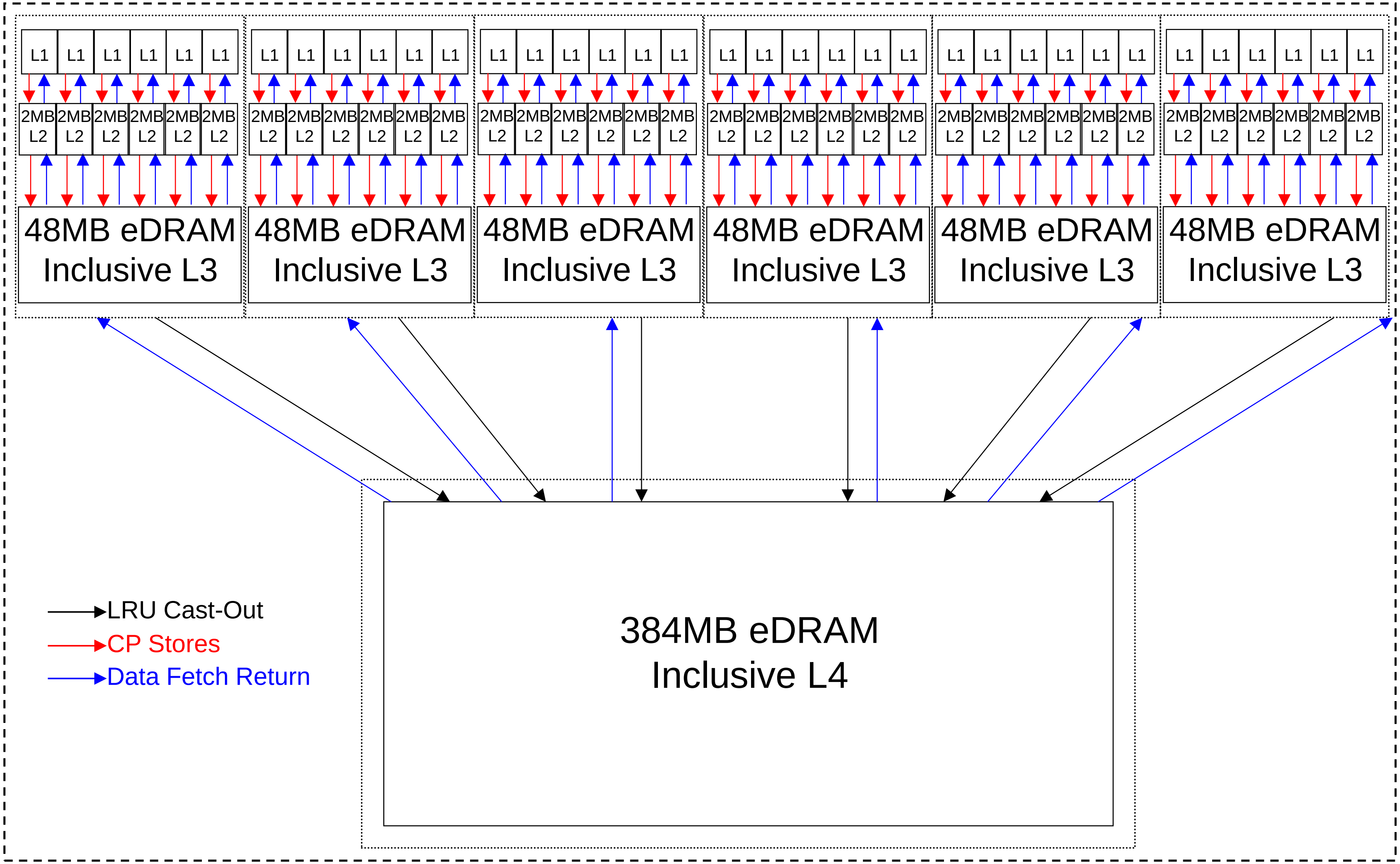
Design / Performance Intersection

- Private core cache
  - Low latency, high bandwidth access
  - Caching for performance critical data
- High capacity fully shared system caches
  - Fast shared latency, high bandwidth for smooth SMP scaling
  - Caching for cross-processor sharing & reentrant data use
- Tiered clustered multi-level cache structure
  - Localize processor and cache affinity
  - Ensure consistent SMP scaling within a book
  - Interconnect bandwidth matched for caching effects
- Distributed Switch Design
  - Ensure SMP flatness as system scales up
  - Balanced system effects

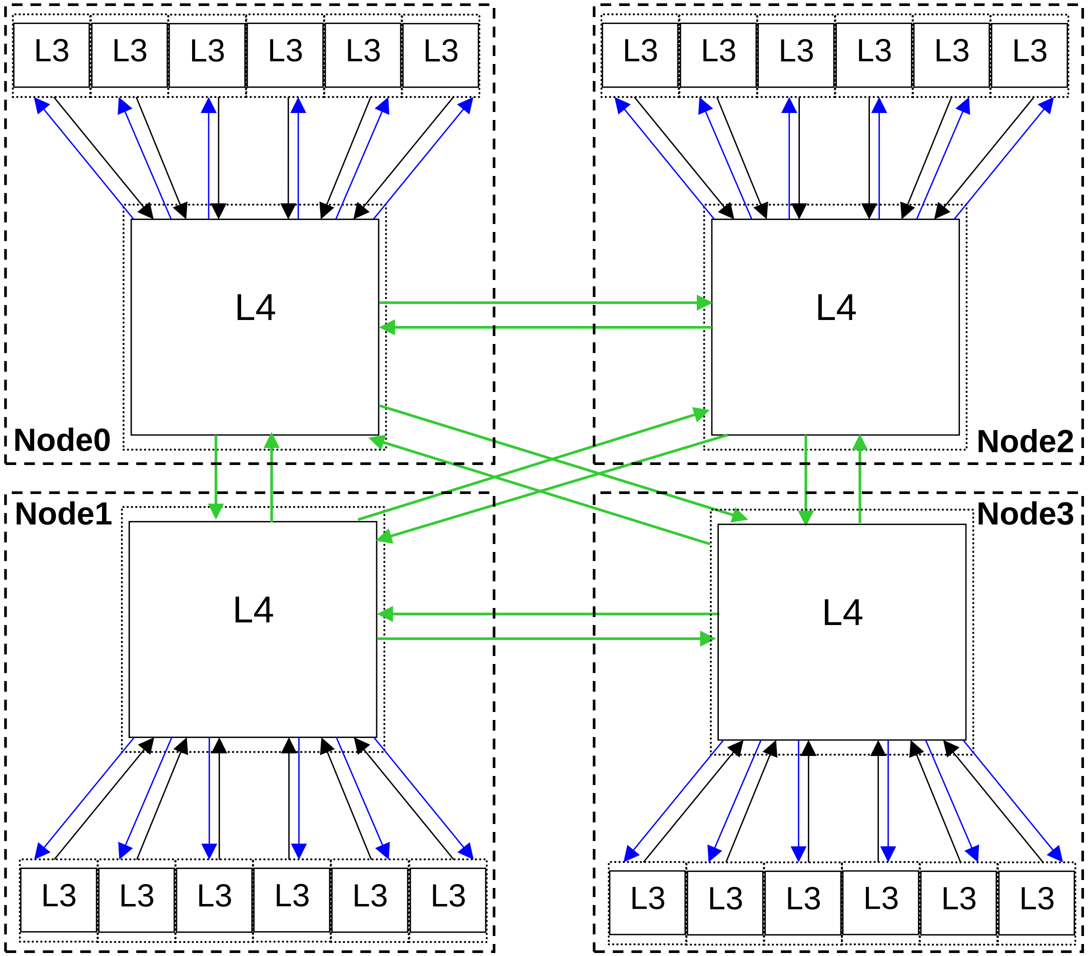
Performance Contributions



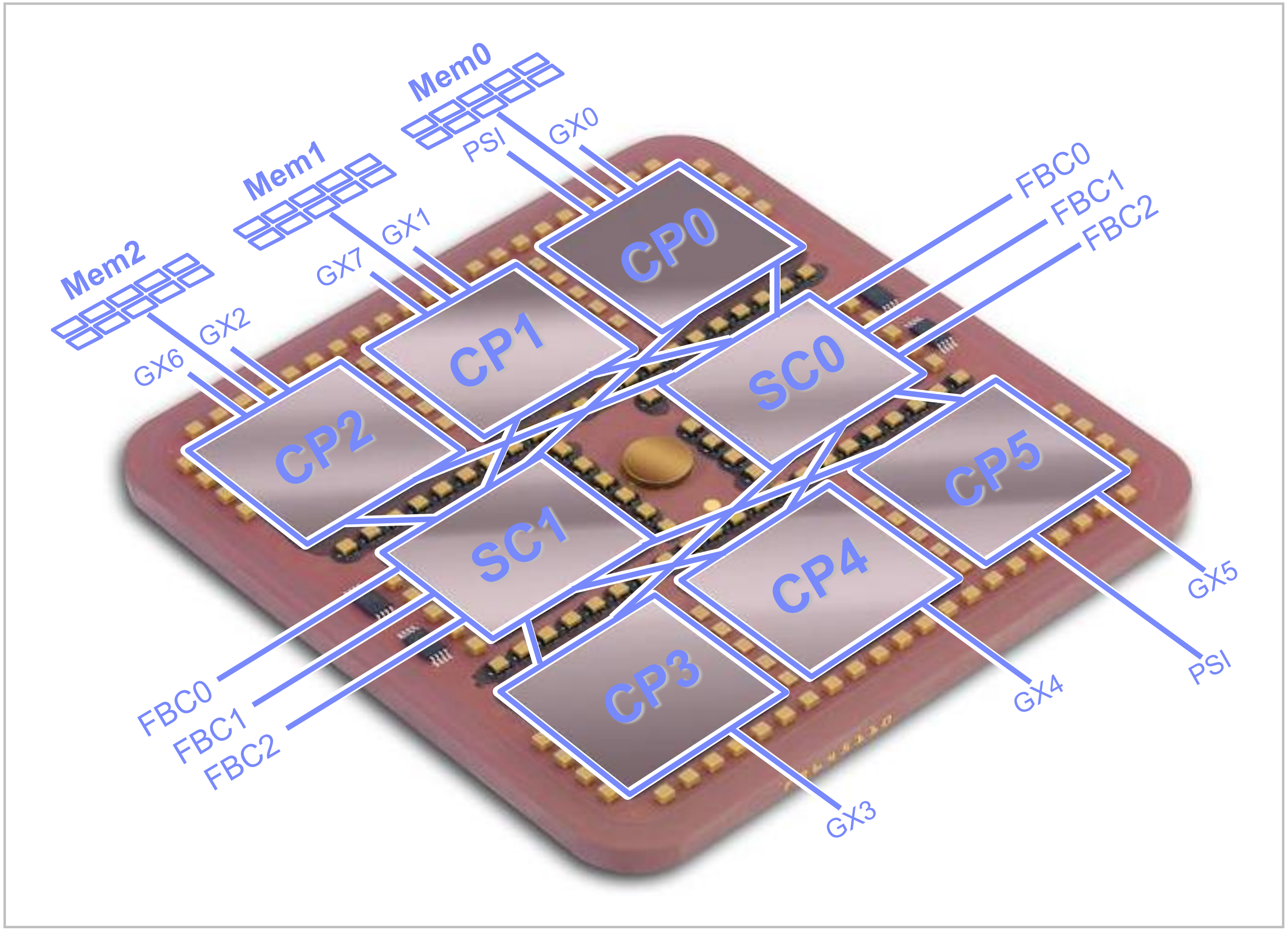
Logical Node Overview



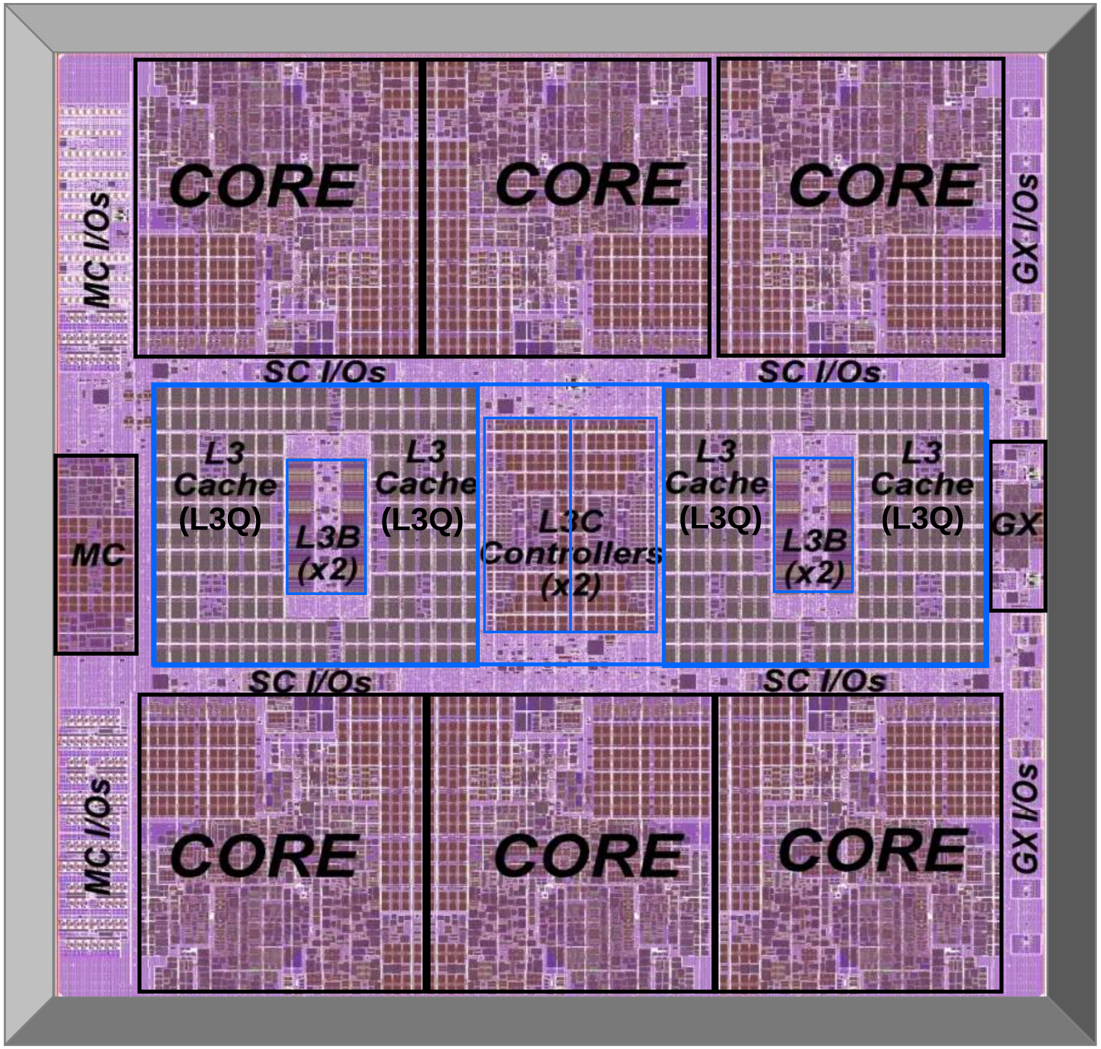
Logical System Overview



Node Multi-Chip Module (MCM)



CP (Central Processor) Chip Overview



- 6 Cores
- 48 MB Shared EDRAM L3
- 32 nm SOI Technology
- 5.5 GHz constant core frequency
- 4:1 L3 clock gear ratio
- 2.75 billion transistors
- 7.68 miles of wire
- 598 mm^2

L3 Cache Features

- Two independent slices based on low order address bit

- Each slice connects to six cores, memory and I/O controllers

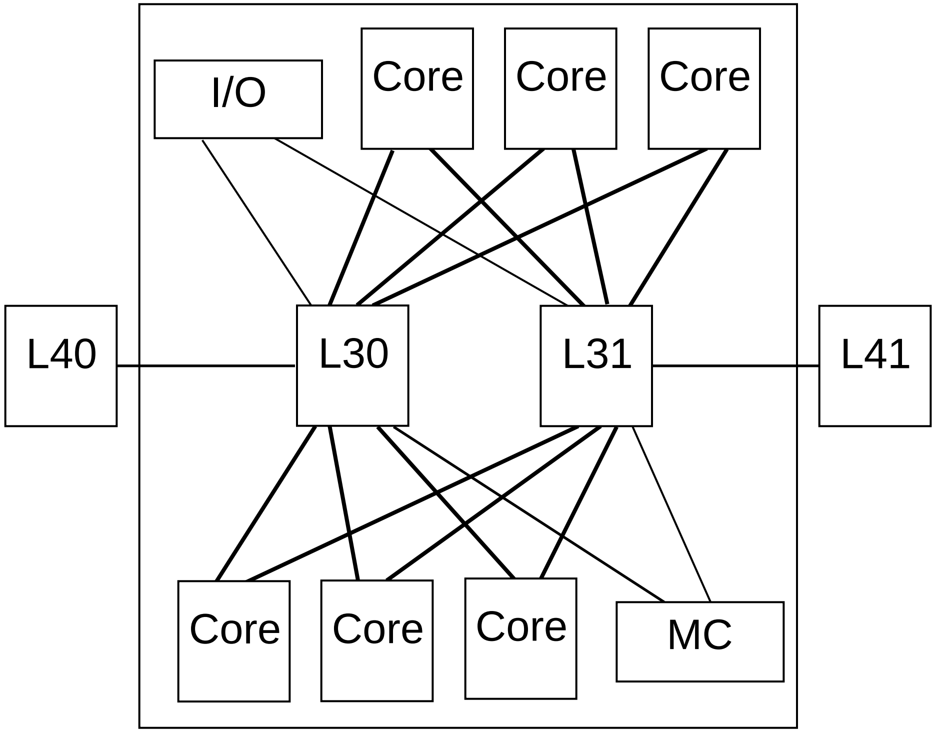
- 12 way set associative

- 16k congruence classes

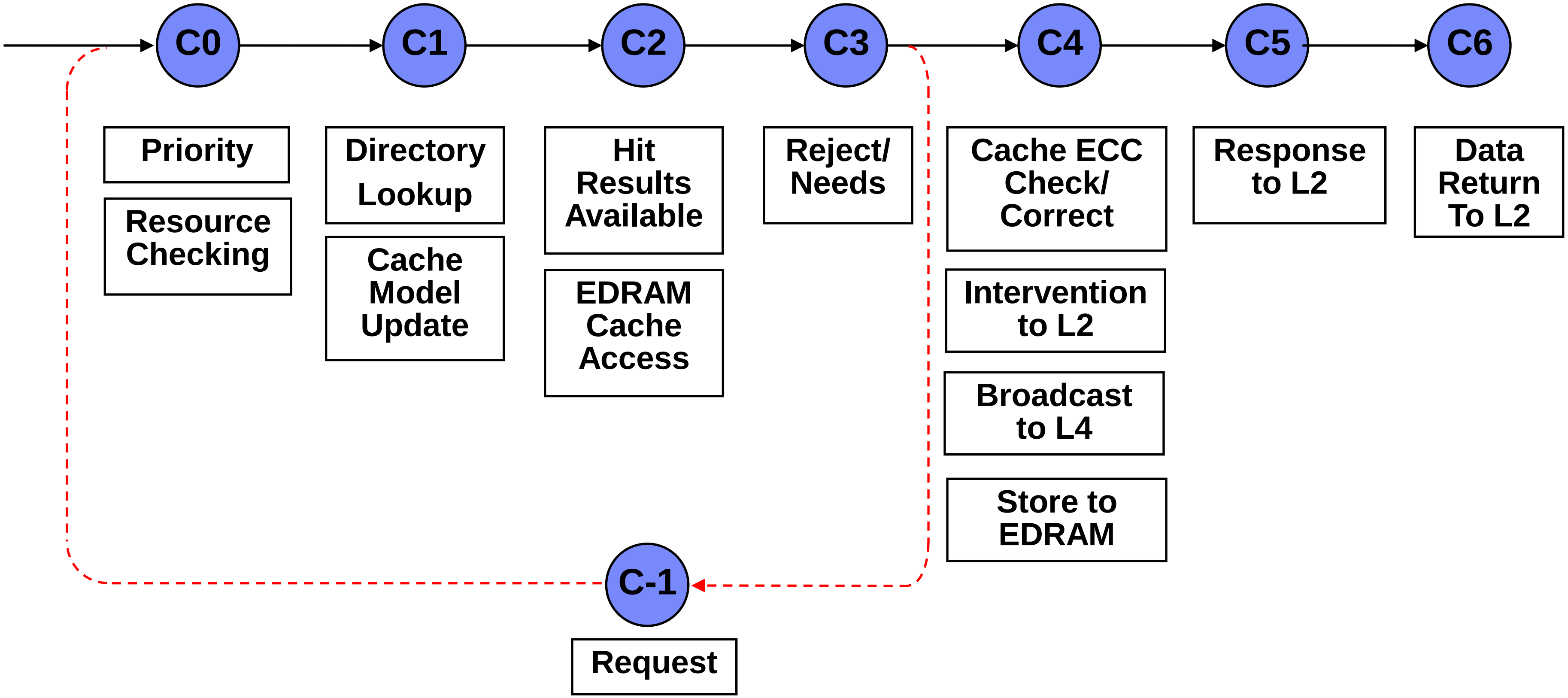
- Byte Merge Stations for DMA partial line operations

- HW Accelerators for page based operations

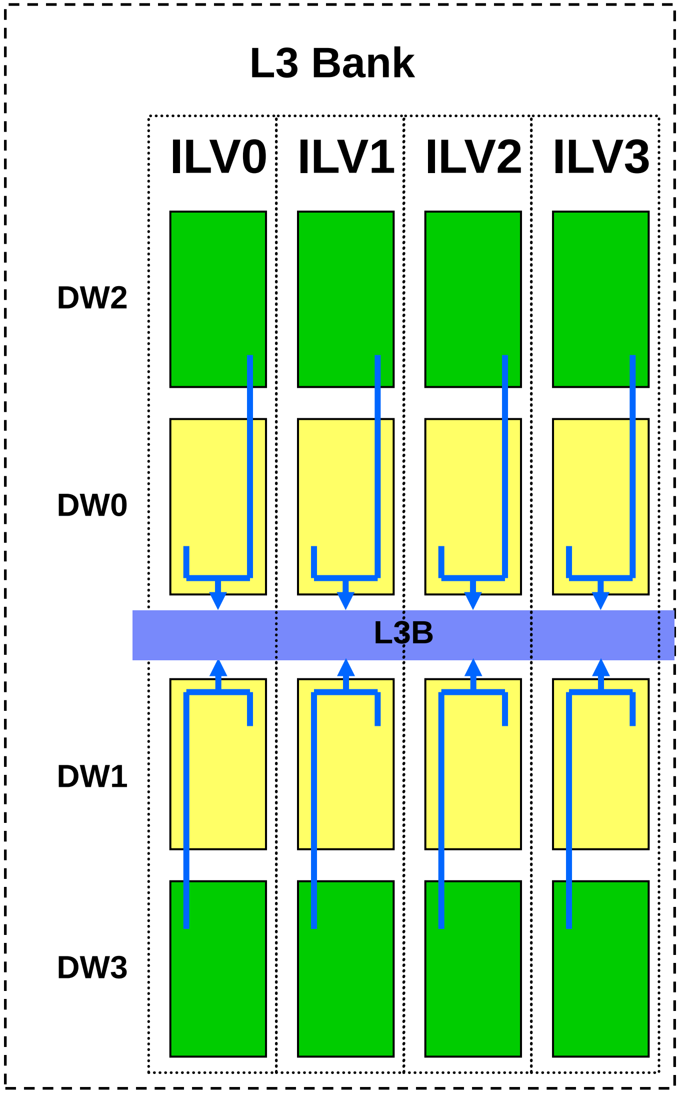
- Recent Store History for preemptive exclusive data access



L3 Pipeline


C0
C1
C2
C3
C4
C5
C6
Priority
Resource Checking
Directory Lookup
Cache Model Update
Hit Results Available
EDRAM Cache Access
Reject/ Needs
Cache ECC Check/ Correct
Intervention to L2
Broadcast to L4
Store to EDRAM
Response to L2
Data Return To L2
C-1
Request

L3 EDRAM Structure



- Two independent banks
- Four way interleave per bank
- Three cycle EDRAM busy time
- Match EDRAM busy and data bus busy time
- 256B cache line spread across four interleaves

L3 EDRAM Management

- EDRAM busy time vs write back cache
  - Stores from the cores represent a majority of operations processed by the L3
- Fetch vs Store Management
  - Fetches schedule EDRAM access for continuous data streaming
  - Flexible store scheduling to minimize resource conflicts

🟩	Fetch eDRAM busy
🟪	Store eDRAM busy
🟨	Fetch Start Blocked

Blocking Store Scheduler

store	C0	C1	C2	C3	C4	C5	C6		
fetch		Stall	Stall	Stall	Stall	Stall	C0	C1	C2
ILV0		🟨	🟨	🟨	🟨				
ILV1		🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟩
ILV2			🟨	🟨	🟪	🟪	🟪	🟩	🟩
ILV3		🟨	🟨	🟨	🟨				🟩

Flexible Store Scheduler

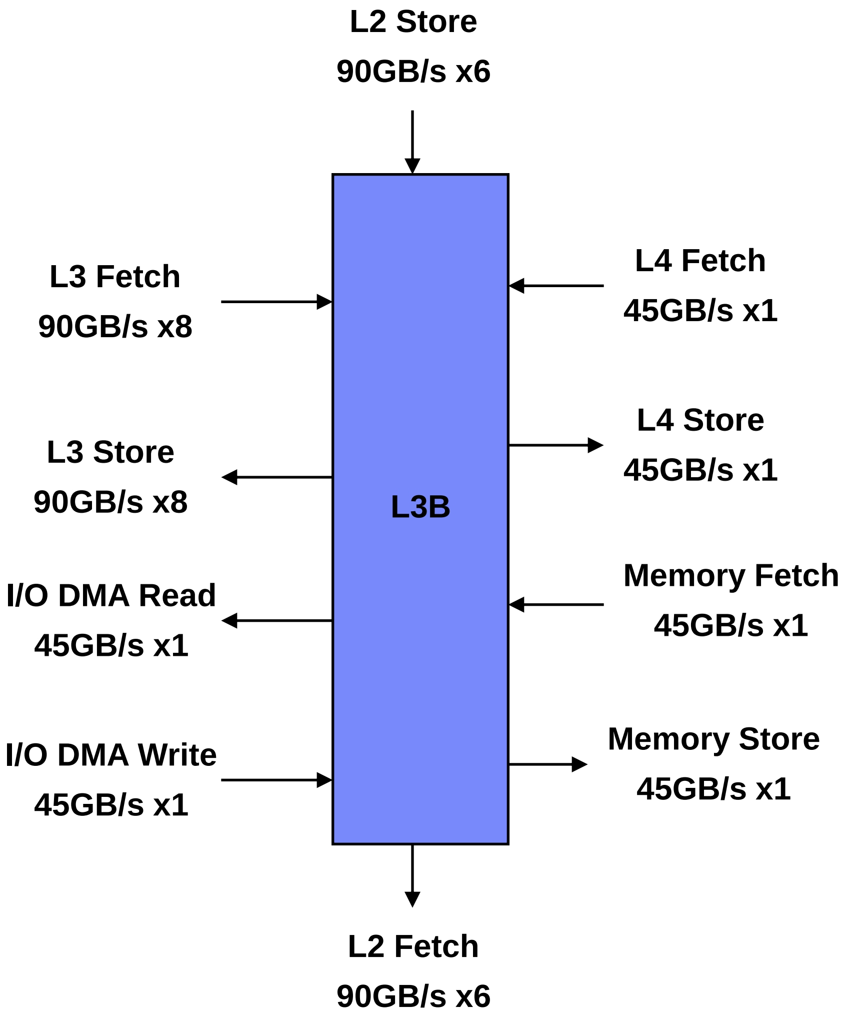
store	C0	C1	C2	C3	C4	C5	C6		
fetch		C0	C1	C2	C3	C4	C5	C6	C7
ILV0					🟩	🟩	🟩		
ILV1		🟩	🟩	🟩					
ILV2			🟩	🟩	🟩	🟪	🟪	🟪	
ILV3				🟩	🟩	🟩			

L3 Dataflow Challenges

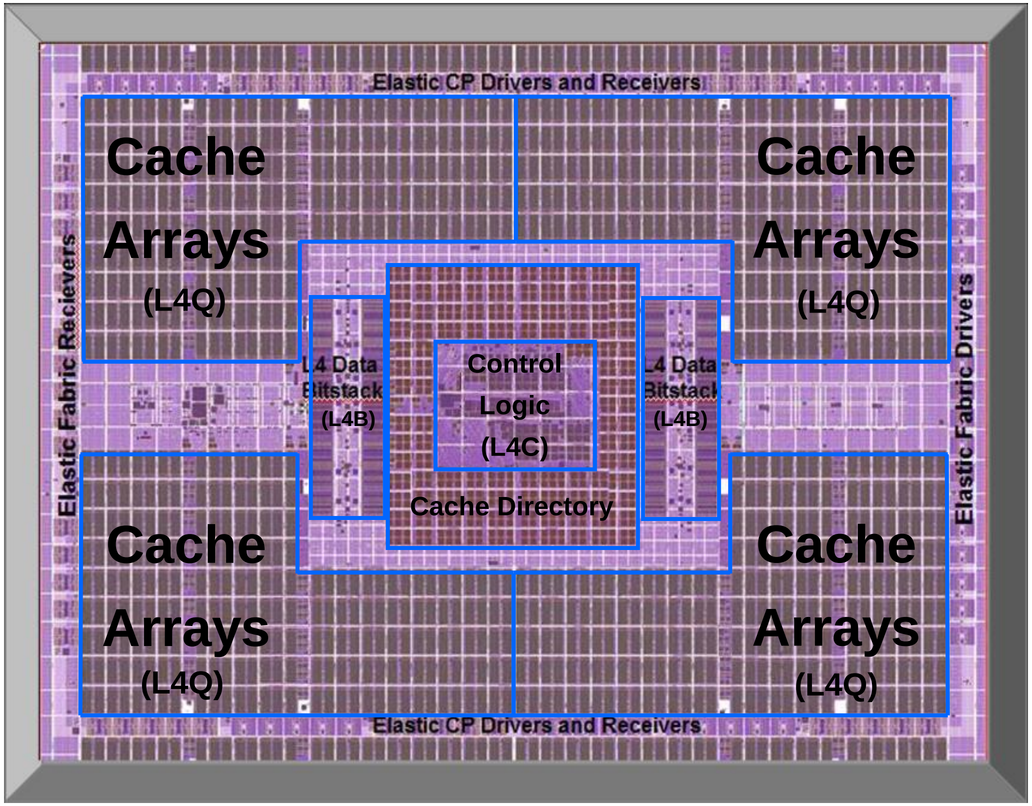
- Wiring and Reach
  - 384 EDRAM macros
  - ~100 cache line data buffers
  - Symmetric core latency

- Request Concurrency
  - 6 Fetch and 12 Store requests per core
  - 12 Fetch and 8 Store requests per IO port
  - 16 requests from L4

- Balanced Bandwidth



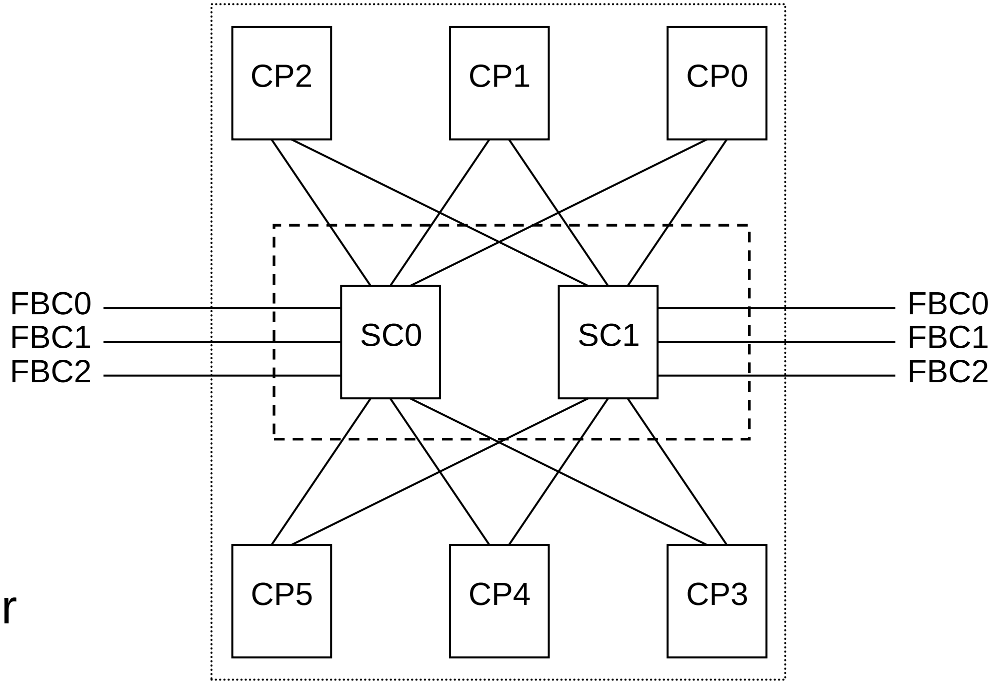
SC (System Controller) Chip Overview



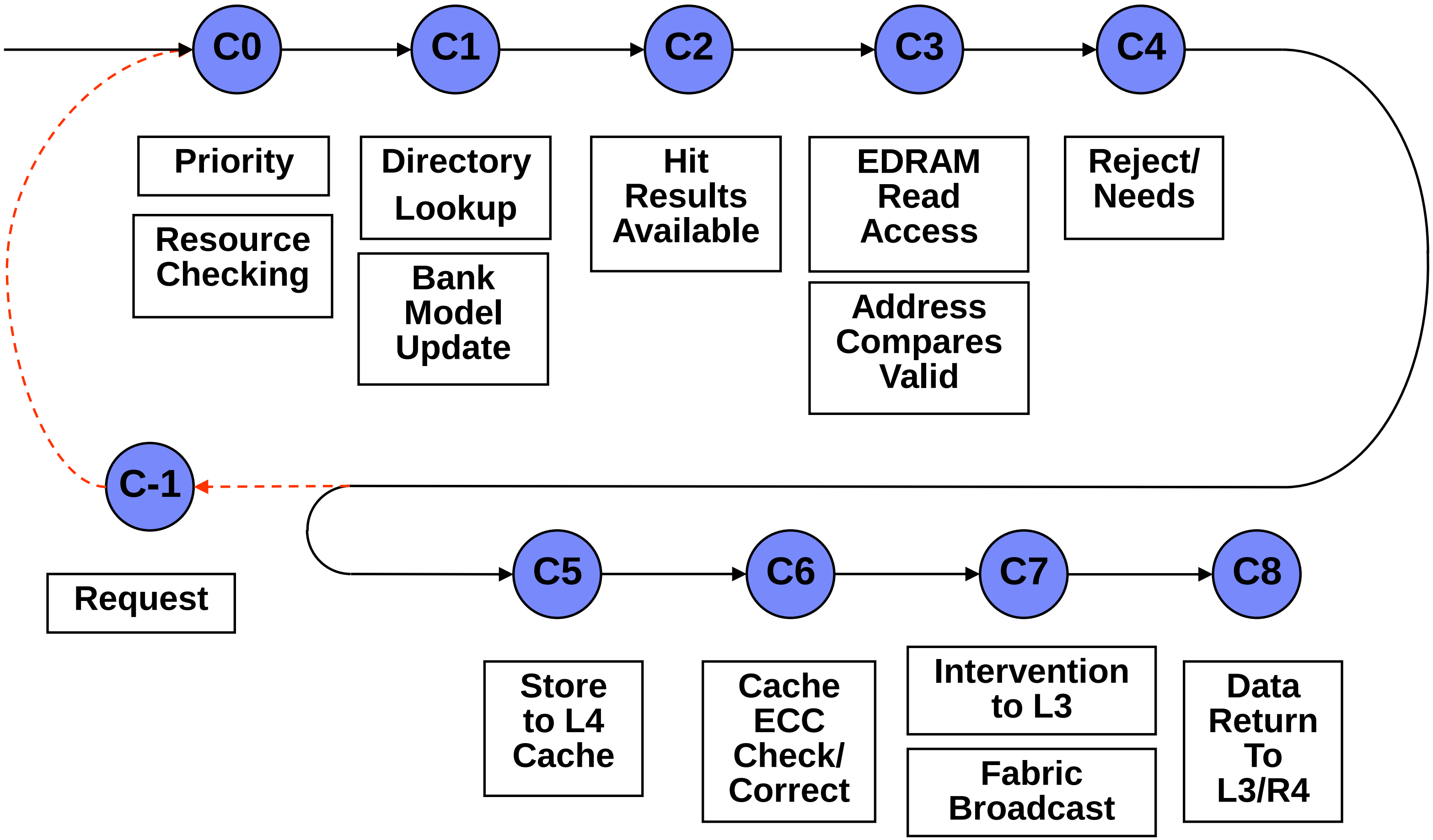
- 192 MB Shared EDRAM L4
- 6 CP Chip Interfaces
- 3 SMP Interfaces
- 3.3 billion transistors
- 32 nm SOI Technology
- 4:1 L4 clock gear ratio
- 526 mm^2

L4 Cache Features

- L4 spread across two SC chips
  - matches L3 address slicing
- Each SC chip connects to six CP chips and three SC on other nodes
- 24 way set associative
- 64k congruence classes
- L4 is the system coherency manager
- HW Pattern Based Prefetching



L4 Pipeline



C0
Priority
Resource Checking

C1
Directory Lookup
Bank Model Update

C2
Hit Results Available

C3
EDRAM Read Access
Address Compares Valid

C4
Reject/ Needs

C-1
Request

C5
Store to L4 Cache

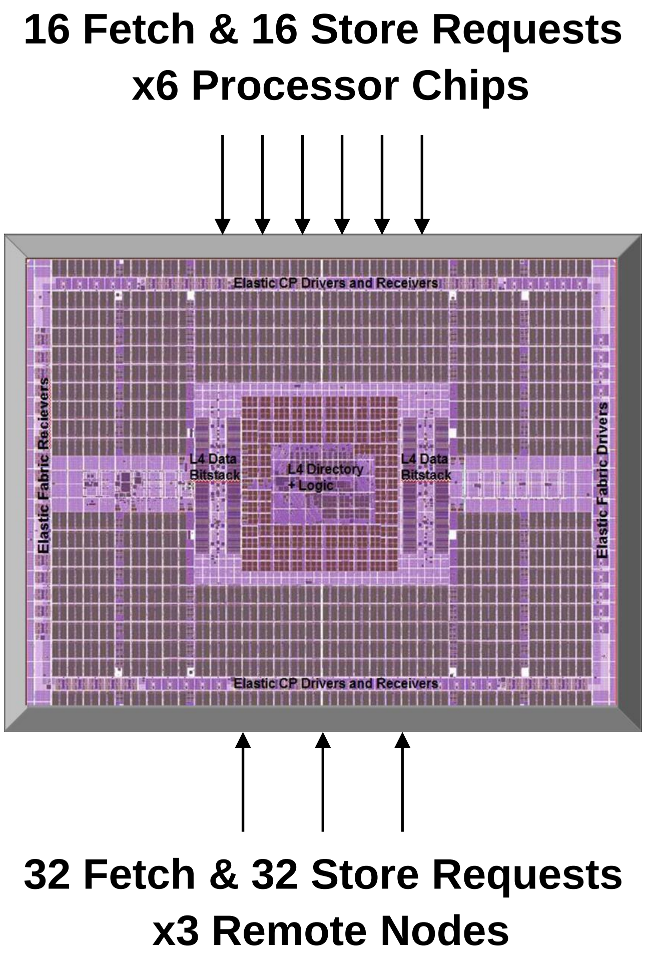
C6
Cache ECC Check/ Correct

C7
Intervention to L3
Fabric Broadcast

C8
Data Return To L3/R4

L4 Design Challenges

- Integration, Wireability, and Reach
  - 1024 eDRAM macros & management logic
  - 114 cache line data buffers
  - 230 address registers w/compare logic
- Request Concurrency
  - Support for 196 concurrent operations per chip
- Intelligent Request Scheduling
  - Operation Address Interlocks
  - Central Fairness/Ordering
- Bandwidth balancing
  - L3<->L4: 22GB/s per port, 132GB/s in/out
  - L4<->L4: 22GB/s per port,   66GB/s in/out
  - L4 Cache services >60% of L3 requests under storage hierarchy intense workloads



16 Fetch & 16 Store Requests x6 Processor Chips

32 Fetch & 32 Store Requests x3 Remote Nodes

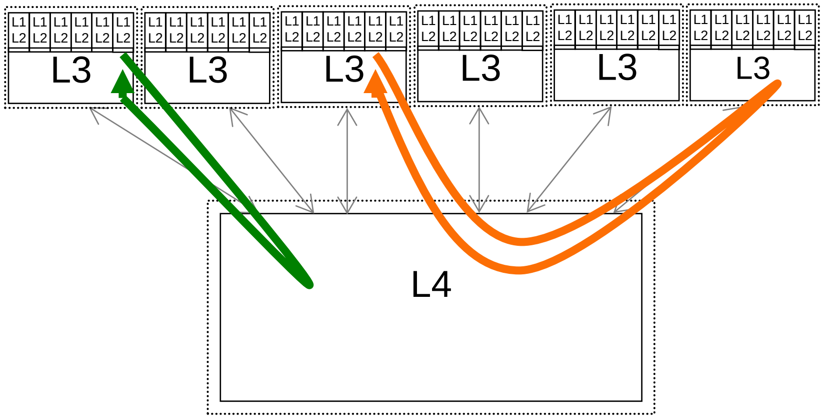
Intra-Node Cache Management

- MESI derived protocol
- L2, L3 and L4 are inclusive
- L1 and L2 are write through
- L3 and L4 are write back
- All L3 to L3 communication goes through L4
- L3 Miss Requests
  - L4 Hit Shared by one or more L3s

    Guaranteed intervention processing time by other L3s on shared lines

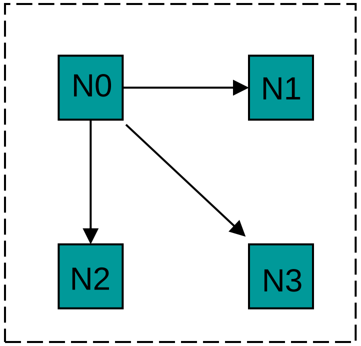
    Data sourced from L4
  - L4 Hit Exclusive or Modified to L3

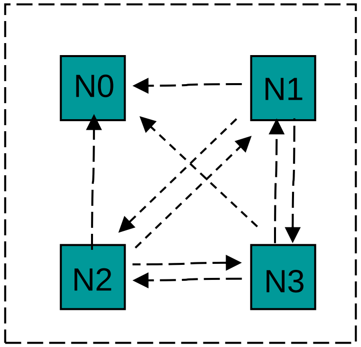
    Data sourced from other L3

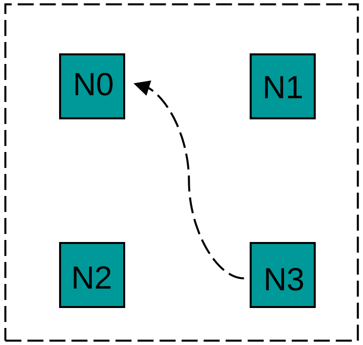


Inter-Node System Coherency Protocol

- Enhanced MOESI Protocol
  - Intervention Master (IM)
  - Memory Master (MM)
  - Multi-Copy (MC)
  - Exclusive
  - Invalid
  - Ghost
- Fabric Broadcast
  - Point to point communication
  - Local state information sent to remote nodes
- Partial Response Broadcast
  - Any to any, expediting system state information
  - Partial responses are ordered, no response identifier tags
- Data Return
  - Immediate return by IM node on hit clean or shared states
- Horizontal Persistence
  - Allows IM state to move to another node when evicted



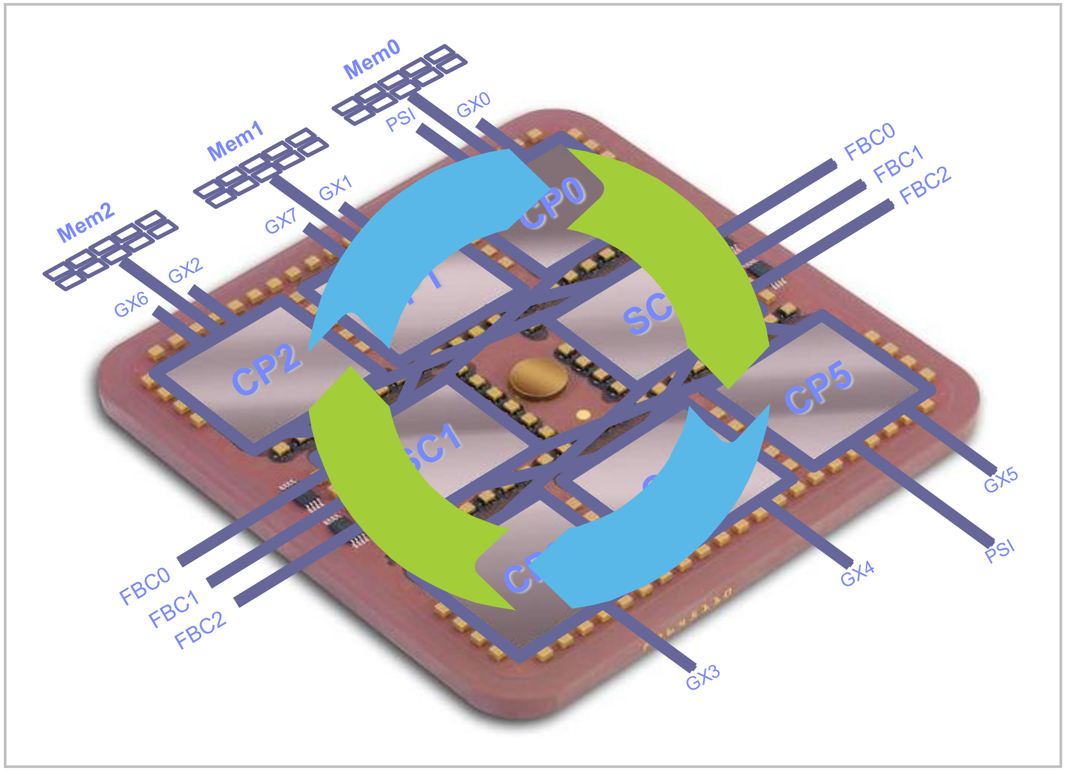




System RAS

- Pervasive coherent RAS handling through-out the hardware, firmware, and operating system

- System RAS Features
  - Bitline delete
  - Dynamic Array Masking
  - Cache Write Back Stepper
  - Memory RAIM
  - Write through L1 and L2
  - Alternate Processor Recovery
  - Concurrent maintenance
  - Dynamic Chip Interface Repair



Summary

The IBM zEC12

has a robust, multi-level shared cache hierarchy

that is designed to meet the needs of the enterprise class computing environment

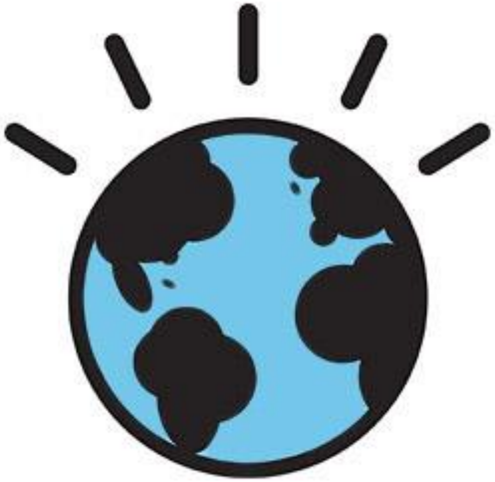
and represents a significant growth in system capacity and performance from its predecessor.

Acknowledgements

- Special thanks
  - – Craig R. Walters – Senior Engineer, System z Performance and Design
  - – Pak-kin Mak – Distinguished Engineer, System z Processor Subsystem Design

- Thanks to the entire System Z Hardware Design and Development teams

Thank You!



Trademarks

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

AIX*
BladeCenter*
CICS*
Cognos*
DataPower*
DB2*
FICON*
GDPS*
IMS
IBM*
IBM (logo)*
Parallel Sysplex*
POWER*
PR/SM
System z*
System z9*
System z10
WebSphere*
z/OS*
z/VM*
z/VSE*
zEnterprise*

* Registered trademarks of IBM Corporation

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 there from.
Java and all Java-based trademarks are trademarks of Oracle Corporation 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.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
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 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.