M7: Next Generation SPARC

Hotchips 26 – August 12, 2014

Stephen Phillips
Senior Director, SPARC Architecture
Oracle

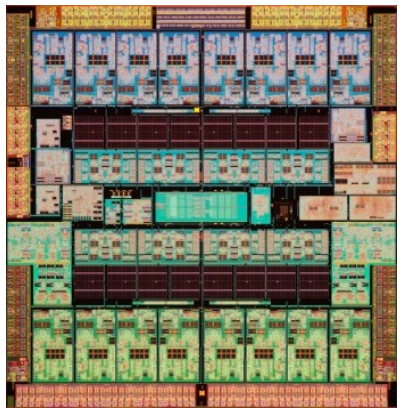
ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

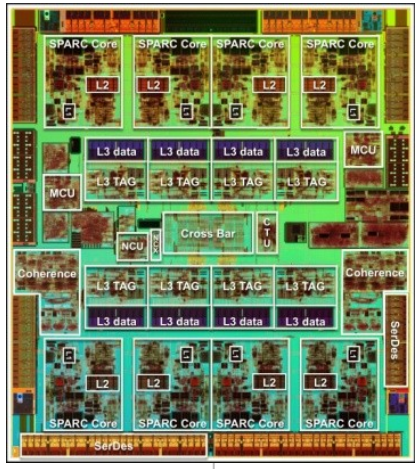
SPARC @ Oracle

5 Processors in 4 Years



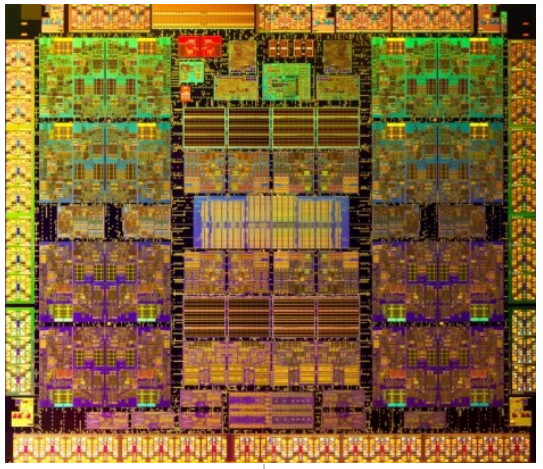
2010

SPARC T3
- 16 S2 cores
- 4MB L3$
- 40 nm technology
- 1.65 GHz



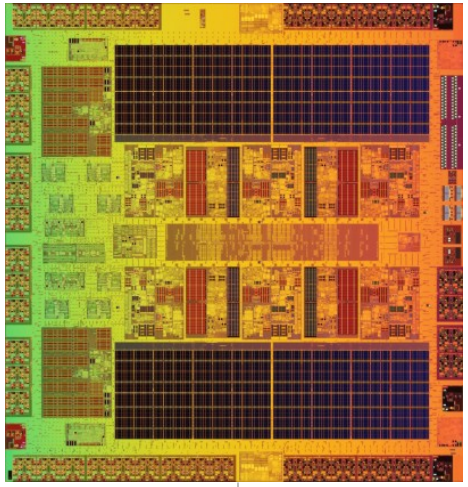
2011

SPARC T4
- 8 S3 Cores
- 4MB L3$
- 40nm Technology
- 3.0 GHz



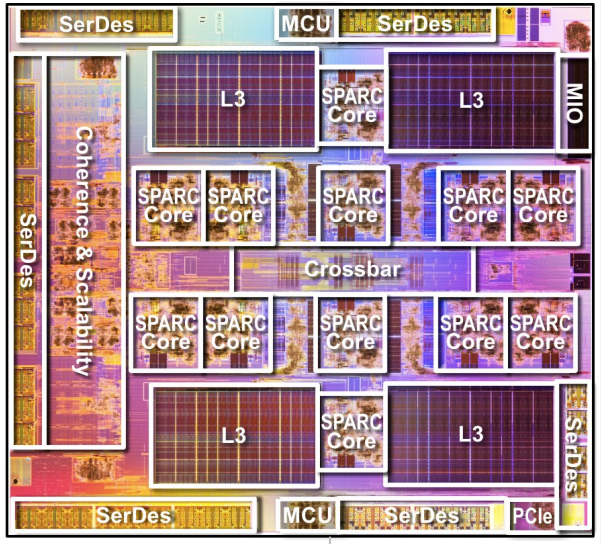
2012

SPARC T5
- 16 S3 Cores
- 8MB L3$
- 28nm Technology
- 3.6 GHz



2012

SPARC M5
- 6 S3 Cores
- 48MB L3 $
- 28nm Technology
- 3.6 GHz



2013

SPARC M6
- 12 S3 Cores
- 48MB L3$
- 28nm Technology
- 3.6 GHz

Oracle's SPARC Strategy

Extreme Performance

Best Performance and Price-Performance in the Industry

Deliver Services with Faster Speeds, and Lower the Costs for Existing Services

Computing Efficiency

Flexible Virtualization, Designed for Availability

Enable Higher Utilization of Capital Assets, and Reduce Risks to Availability of Data

Optimized for Oracle Software

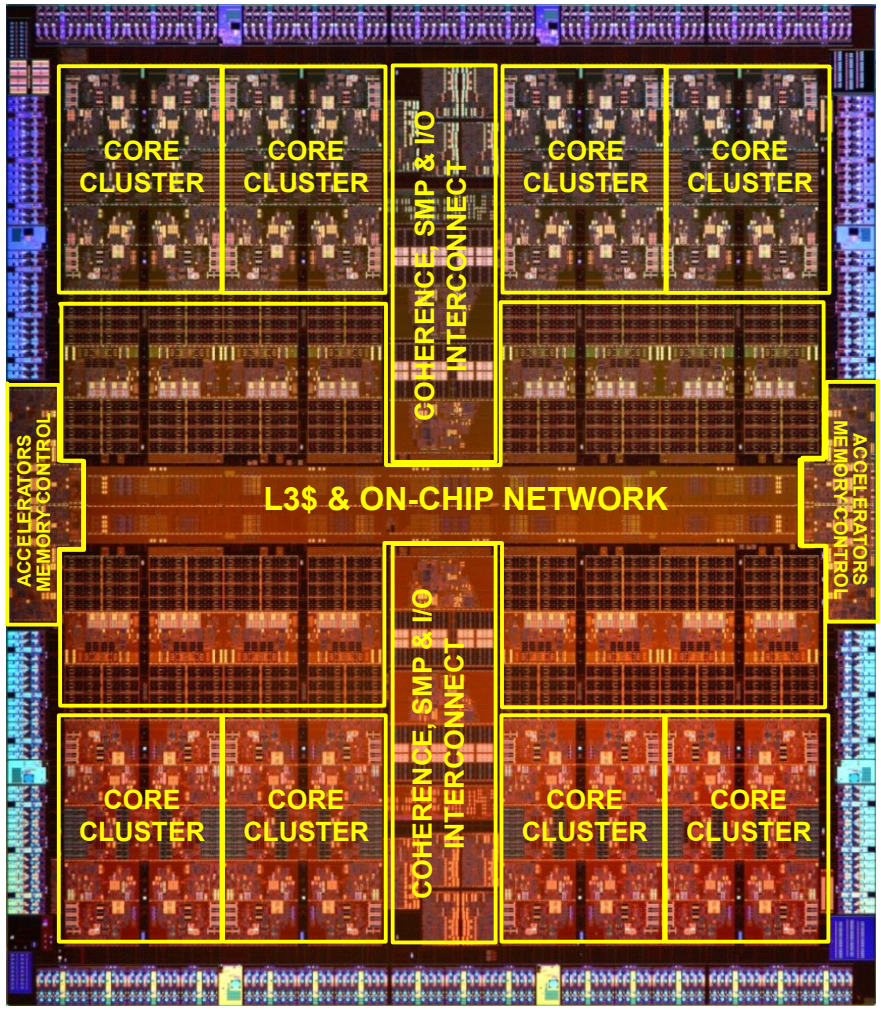
Hardware and Software Engineered, Tested, and Supported Together

Deploy Technology Faster, with Less Risk and Lower Costs

http://blog.oracle.com/bestperf

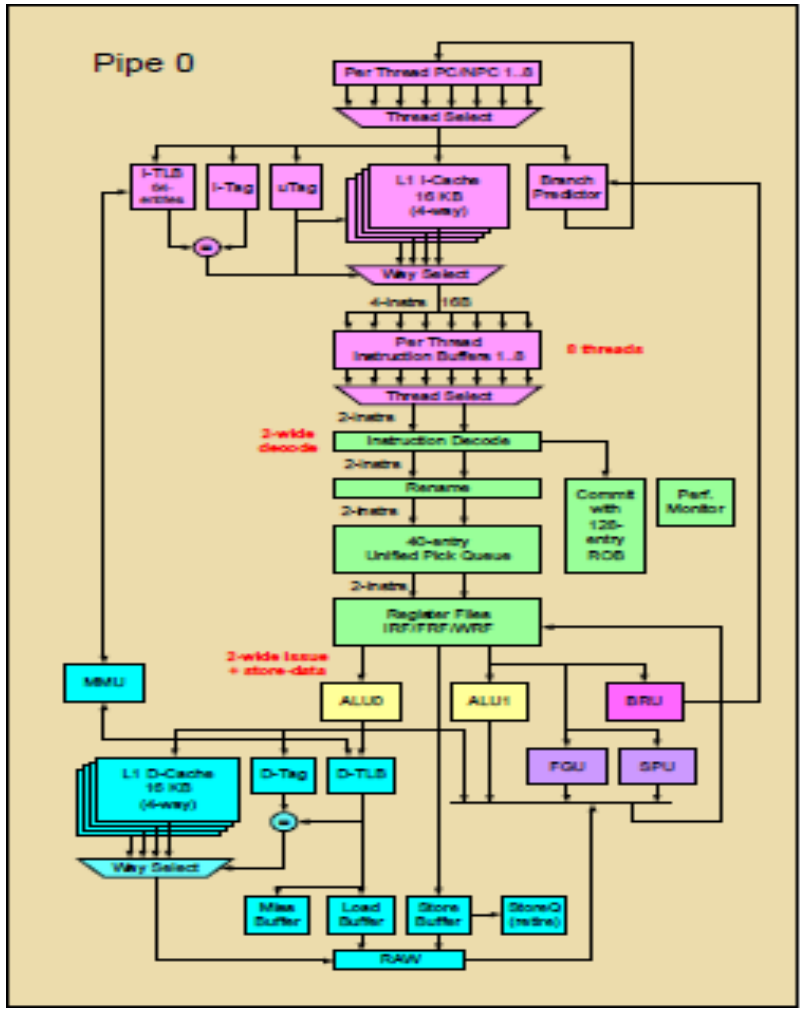
ORACLE

M7 Processor



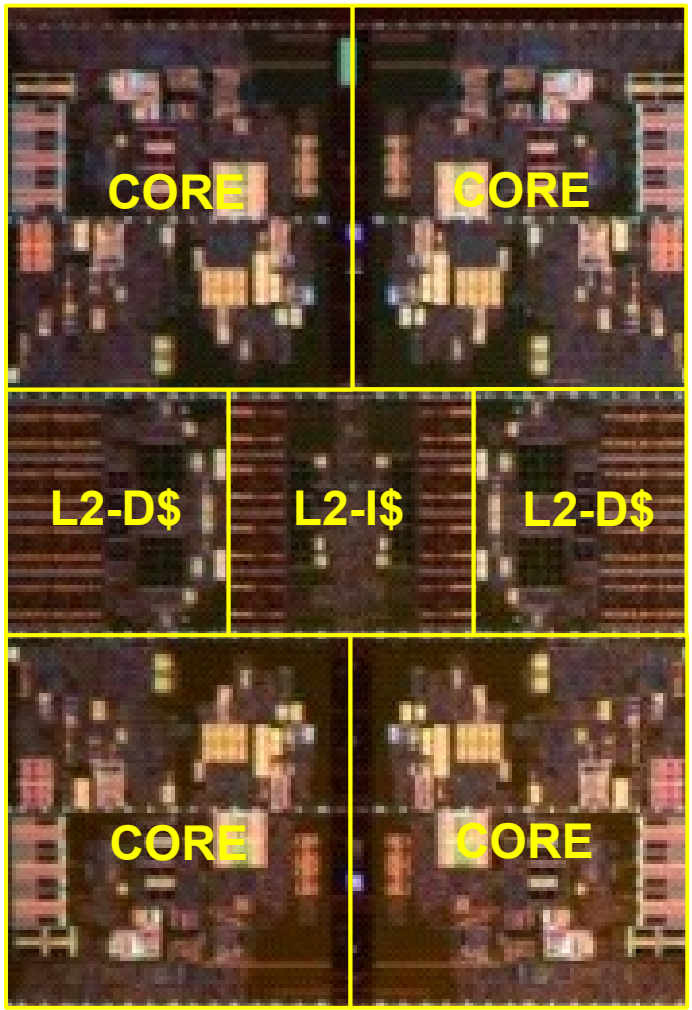
- 32 SPARC Cores
  - Fourth Generation CMT Core (S4)
  - Dynamically Threaded, 1 to 8 Threads Per Core
- New Cache Organizations
  - Shared Level 2 Data and Instruction Caches
  - 64MB Shared & Partitioned Level 3 Cache
- DDR4 DRAM
  - Up to 2TB Physical Memory per Processor
  - 2X-3X Memory Bandwidth over Prior Generations
- PCIe Gen3 Support
- Application Acceleration
  - Real-time Application Data Integrity
  - Concurrent Memory Migration and VA Masking
  - DB Query Offload Engines
- SMP Scalability from 1 to 32 Processors
- Coherent Memory Clusters
- Technology: 20nm, 13ML

M7 Core (S4)



- Dynamically Threaded, 1 to 8 Threads
- Increased Frequency at Same Pipeline Depths
- Dual-Issue, OOO Execution Core
  - 2 ALU's, 1 LSU, 1 FGU, 1 BRU, 1 SPU
  - 40 Entry Pick Queue
  - 64 Entry FA I-TLB, 128 Entry FA D-TLB
  - Cryptographic Performance Improvements
  - 54bit VA, 50bit RA/PA, 4X Addressing Increase
- Fine-grain Power Estimator
- Live Migration Performance Improvements
- Core Recovery
- Application Acceleration Support
  - Application Data Integrity
  - Virtual Address Masking
  - User-Level Synchronization Instructions

M7 Core Cluster

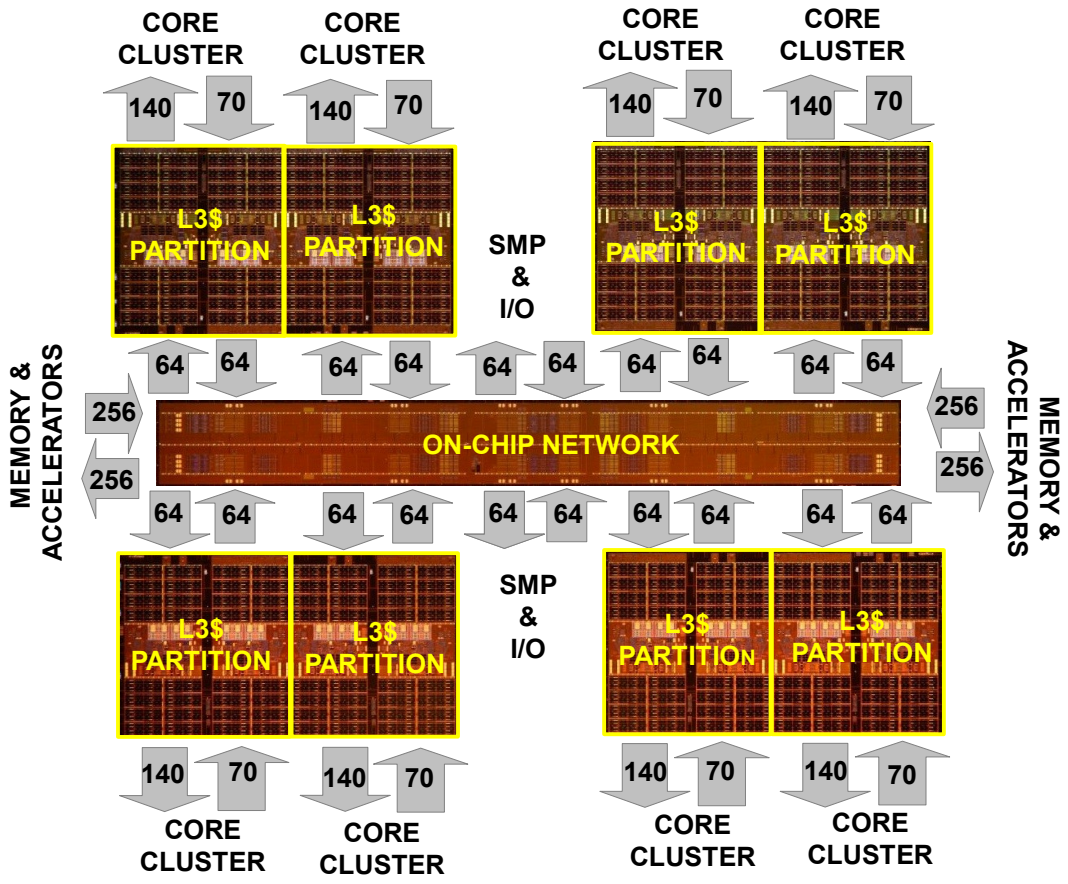


- 4 SPARC Cores Per Cluster
- New L2$ Architecture
  - 1.5X Larger at Same Cycle Count Latency as Prior Generations
  - 2X Bandwidth per Core, >1TB/s Per Core Cluster, >8TB/s per Chip
- 256KB L2-I$
  - Shared by All Cores in Core Cluster
  - 4-way SA, 64B Lines, >500GB/s Throughput
  - 4 Independent Core Interfaces @ >128GB/s Each
- 256KB Writeback L2-D$
  - Each L2-D$ Shared by 2 Cores (Core-pair)
  - 8-way SA, 64B Lines, >500GB/s Throughput per L2-D$
  - 2 Independent Core Interfaces @ >128GB/s per L2-D$
- Core-pair Dynamic Performance Optimizations
  - Doubles Core Execution Bandwidth at Prior Generation Thread Count to Maximize Per-Thread Performance
  - Utilize 8 Threads Per Core to Maximize Core Throughput

ORACLE

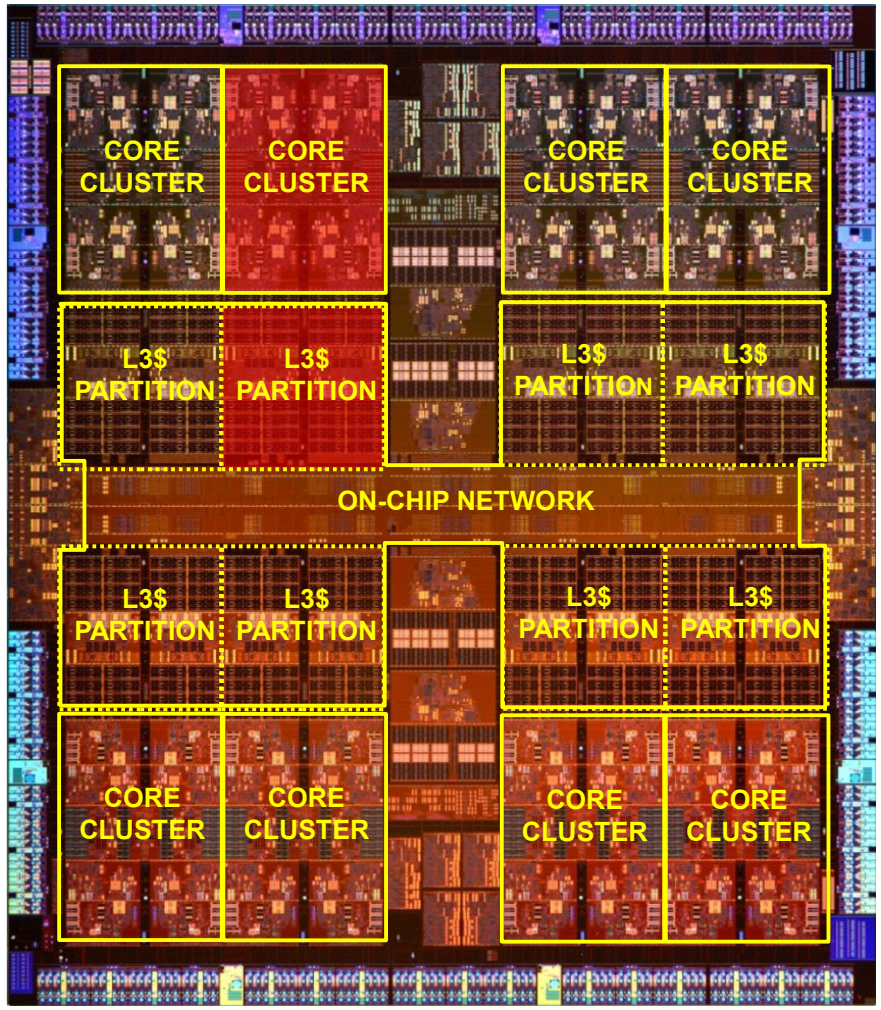
M7 Level 3 Cache and On-Chip Network

L3$ and OCN Data Bandwidth (GB/s)



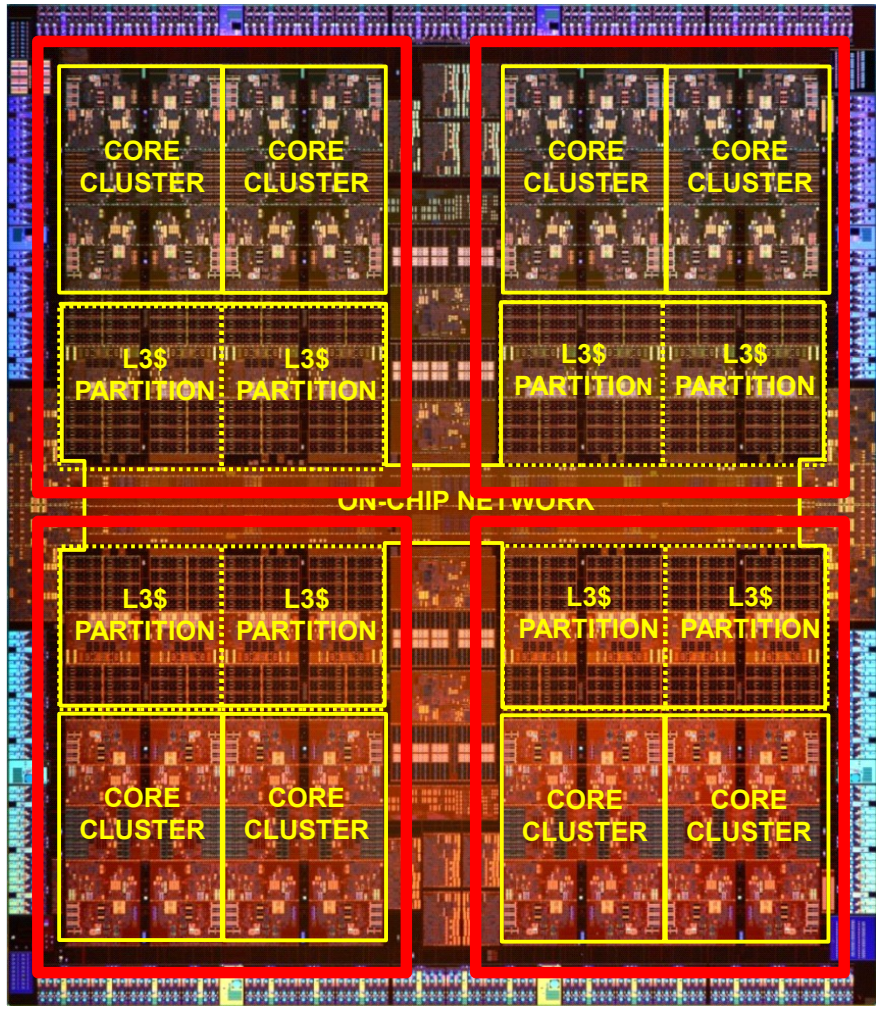
- 64MB Shared & Partitioned L3$
  - 8MB Local Partitions, Each 8-way SA
  - >25% Reduction in Local L3$ Cycle Latency Compared to Prior Generation (M6)
  - >1.6TB/s L3$ Bandwidth per Chip, 2.5x (T5) to 5X (M6) Over Previous Generations
  - Cache Lines May be Replicated, Migrated or Victimized Between L3$ Partitions
  - HW Accelerators May Directly Allocate into Target L3$ Partitions
- On-Chip Network (OCN)
  - Consists of Request (Rings), Response (Pt-Pt) and Data (Mesh) Interconnects
  - 0.5TB/s Bisection Data Bandwidth

M7 Process Group & Virtualization Affinity



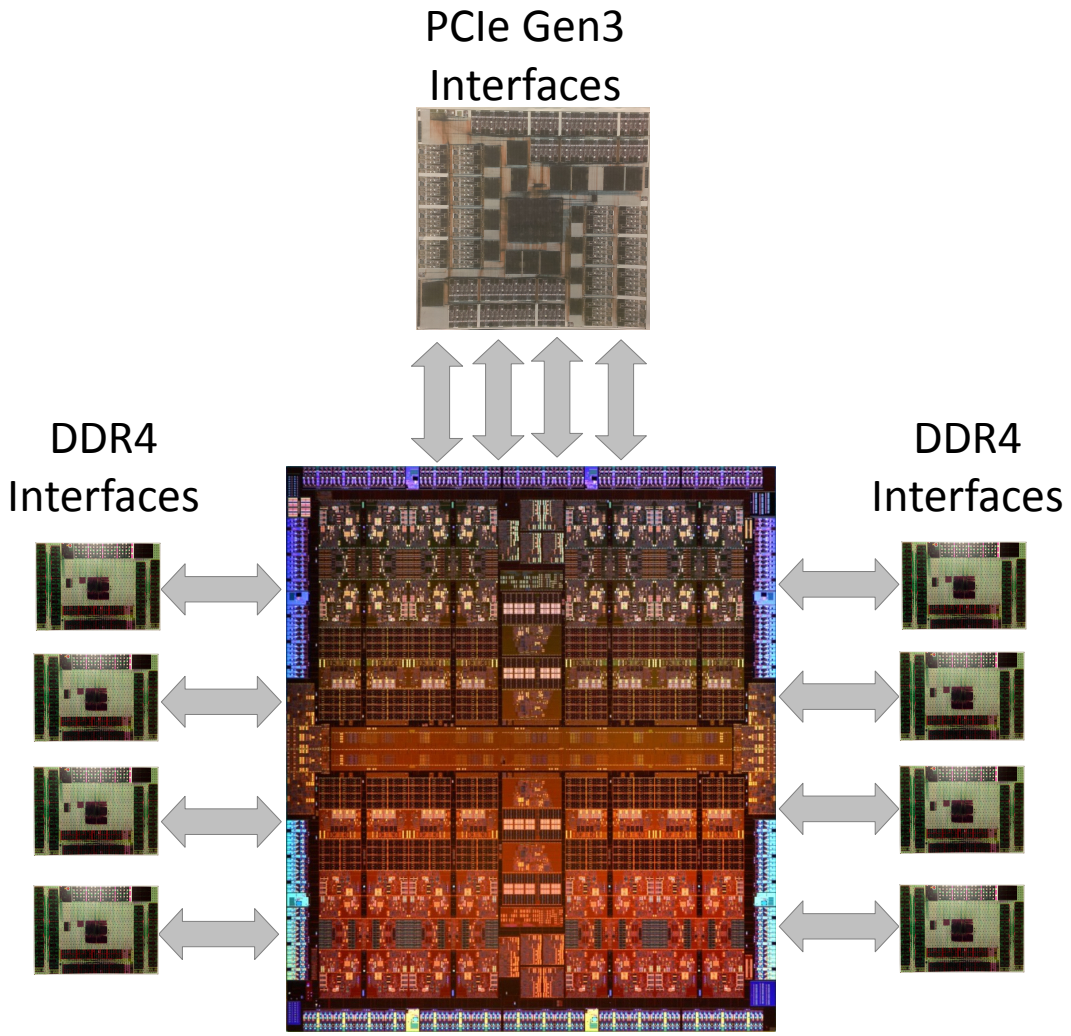
- Solaris Process Group Hierarchy
  - Thread Scheduling Level Corresponding to Core Cluster and Local L3$ Partition
  - Thread Load Balance Across L3$ Partitions
  - Thread Reschedule Affinity
  - Co-location of Threads Sharing Data
- Oracle Virtualization Manager (OVM)
  - Process Group Provisioning Applied to Logical Partitions
  - Minimize L3$ Thrashing Between Virtual Machine Instances, Improve QoS

M7 Fine-grain Power Management



- On-die Power Estimator Per Core
  - Generates Dynamic Power Estimates By Tracking Internal Core Activities
  - Estimates Updated at 250 Nanosecond Intervals
- On-die Power Controller
  - Estimates Total Power of Cores and Caches on a Quadrant Basis (2 Core Clusters + 2 L3$ Partitions)
  - Accurate to within a Few Percent of Measured Power
  - Individually Adjusts Voltage and/or Frequency within Each Quadrant Based on Software Defined Policies
- Performance @ Power Optimizations
  - Highly Responsive to Workload Temporal Dynamics
  - Can Account for Workload Non-uniformity Between Quadrants
- Quadrants May be Individually Power Gated

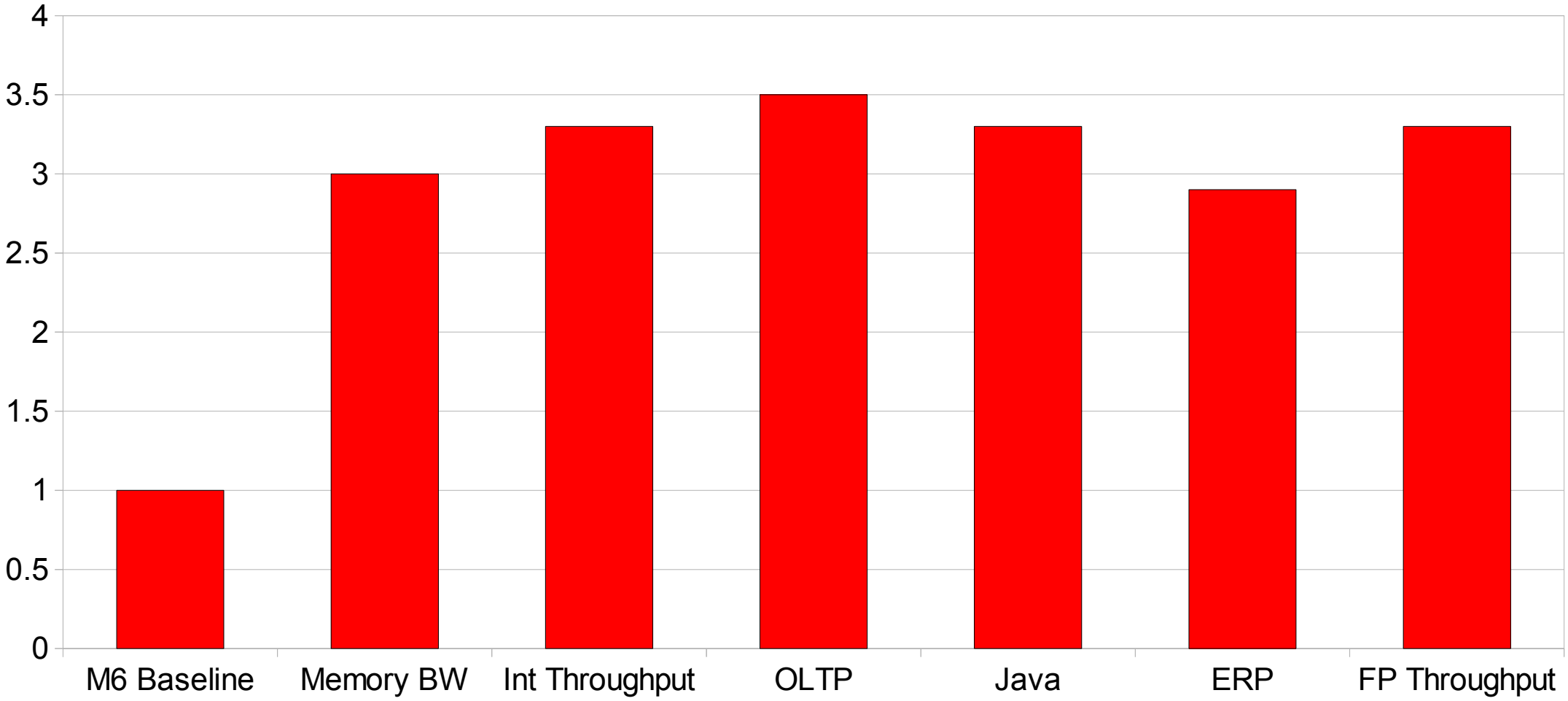
M7 Memory and I/O



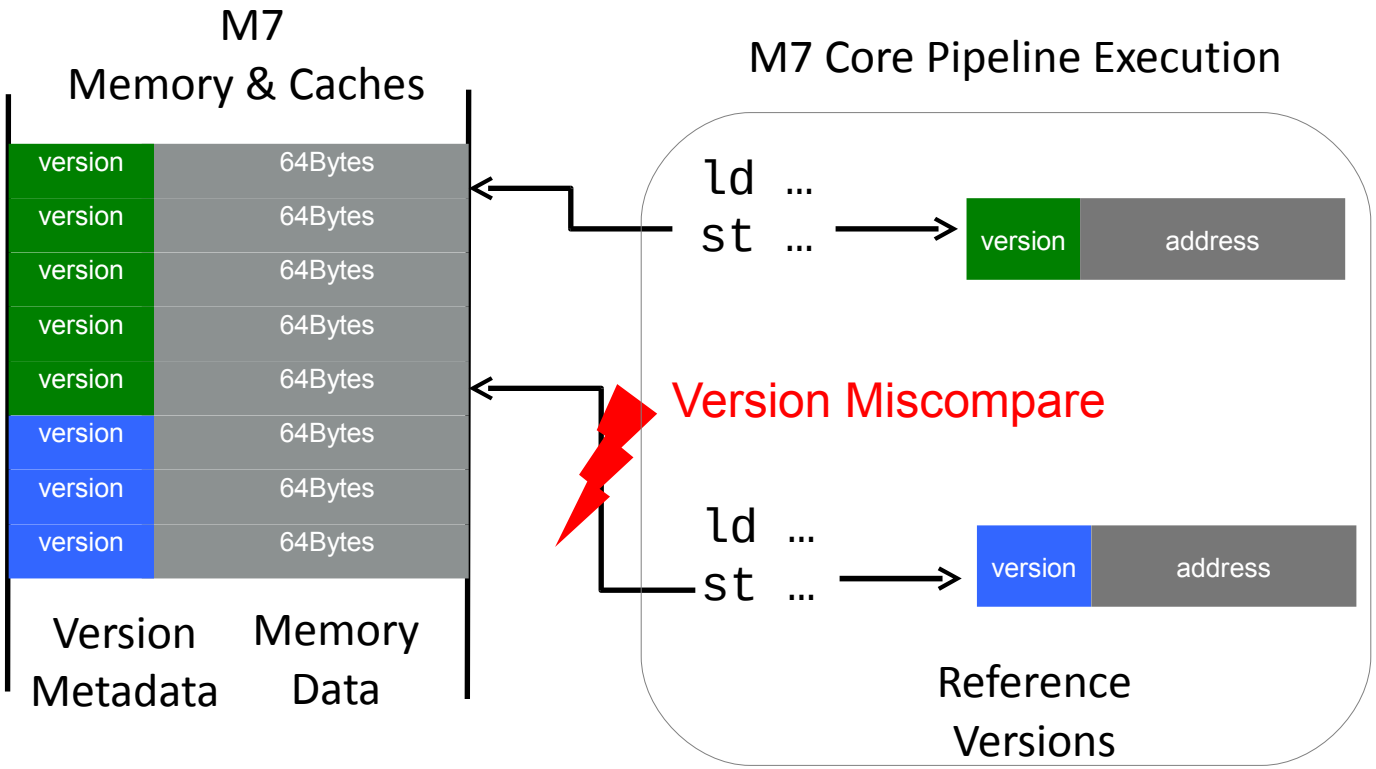
- 4 DDR4 Memory Controllers
  - 16 DDR4-2133/2400/2667 Channels
  - Very Large Memory, Up to 2TB per Processor
  - 160GB/s (DDR4-2133) Measured Memory Bandwidth (2X to 3X Previous Generations, T5 and M6)
  - DIMM Retirement Without System Stoppage
- Memory Links to Buffer Chips
  - 12.8Gbps/14.4Gbps/16Gbps Link Rates
  - Lane Failover with Full CRC Protection
- Speculative Memory Read
  - Reduces Local Memory Latency by Prefetching on Local L3$ Partition Miss
  - Dynamic per Request, Based on History (Data, Instruction) and Threshold Settings
- PCIe Gen3
  - 4 Internal Links Supporting >75GB/s
  - >2X Previous Generations, T5 and M6

ORACLE

M7 Processor Performance



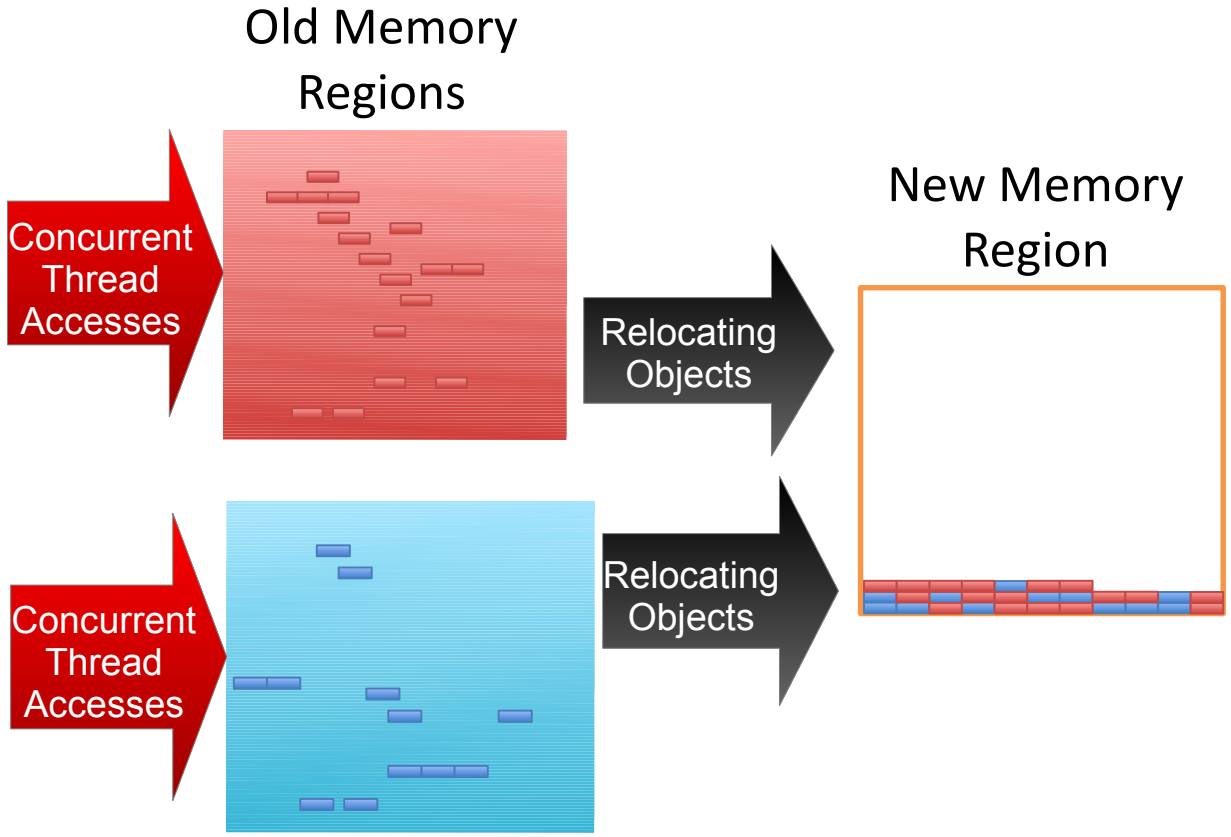
M7 Application Data Integrity



Safeguards Against Invalid/Stale References and Buffer Overruns for Solaris and DB Clients

- Real-time Data Integrity Checking in Test & Production Environments
  - Version Metadata Associated with 64Byte Aligned Memory Data
  - Metadata Stored in Memory, Maintained Throughout the Cache Hierarchy and All Interconnects
  - Memory Version Metadata Checked Against Reference Version by Core Load/Store Units
  - HW Implementation, Very Low Overhead
- Enables Applications to Inspect Faulting References, Diagnose and Take Appropriate Recovery Actions

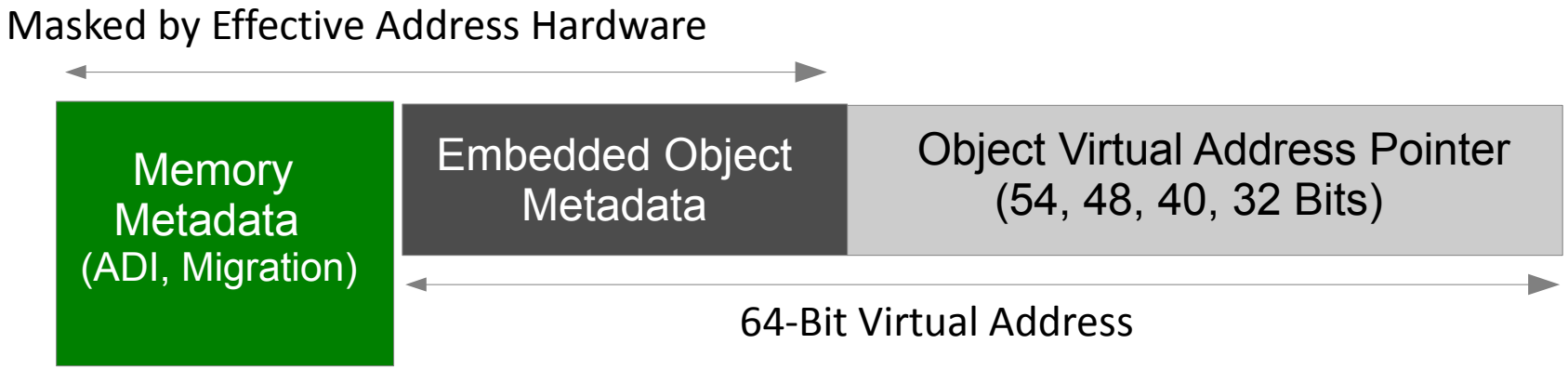
M7 Concurrent Fine-grain Memory Migration



Enables Concurrent and Continuous Operation

- Hardware Support For Fine Grain Access Control
  - Applicable to Fine-grain Objects and Large Memory Pages
  - Bypasses Operating System Page Protection Overheads
  - Scales with Threads and Memory Bandwidth
- Deterministic Memory Access Conflict Resolution
  - Memory Metadata of Relocating Objects Are Marked for Migration
  - User Level Trap on Detection of Memory Reference with Migrating Version

M7 Virtual Address (VA) Masking

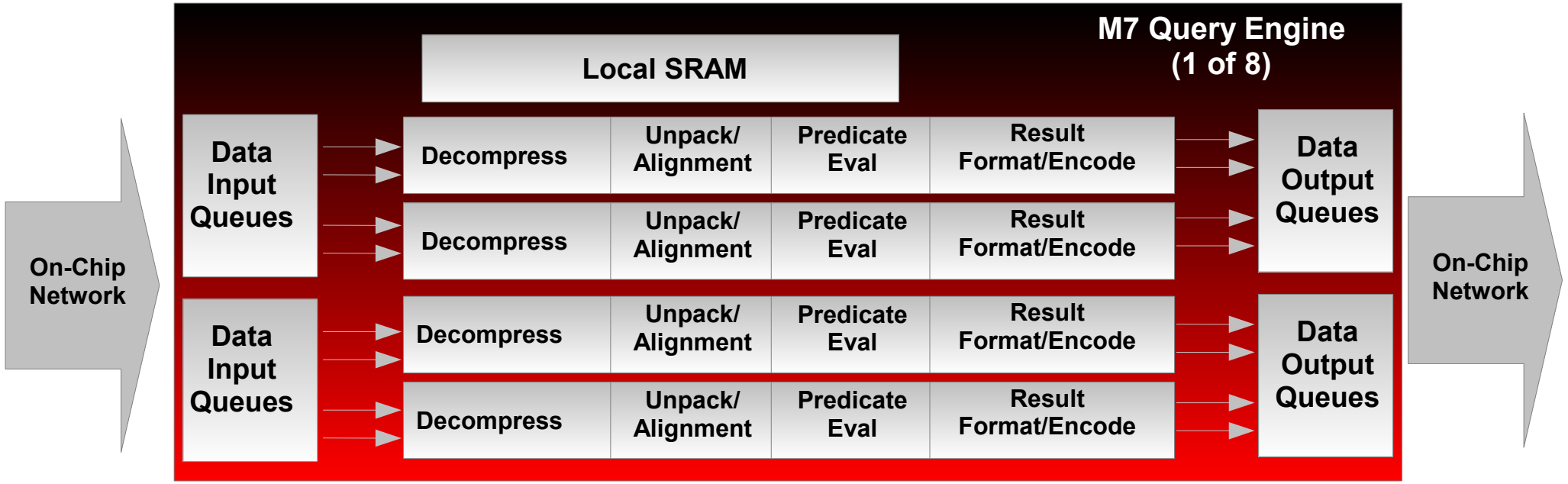


- Allow Programs to Embed Metadata in Upper Unused Bits of Virtual Address Pointers
  - Applications Using 64-bit Pointers Can Set Aside 8, 16, 24 or 32 Bits
  - Addressing Hardware Ignores Metadata
- Enables Managed Runtimes (e.g. JVM's) to Embed Metadata for Tracking Object Information
  - Caches Object State Table Information into Object Pointer (Pointer Coloring)
  - Eliminates De-reference and Memory Load from Critical Path

M7 Database In-Memory Query Accelerator

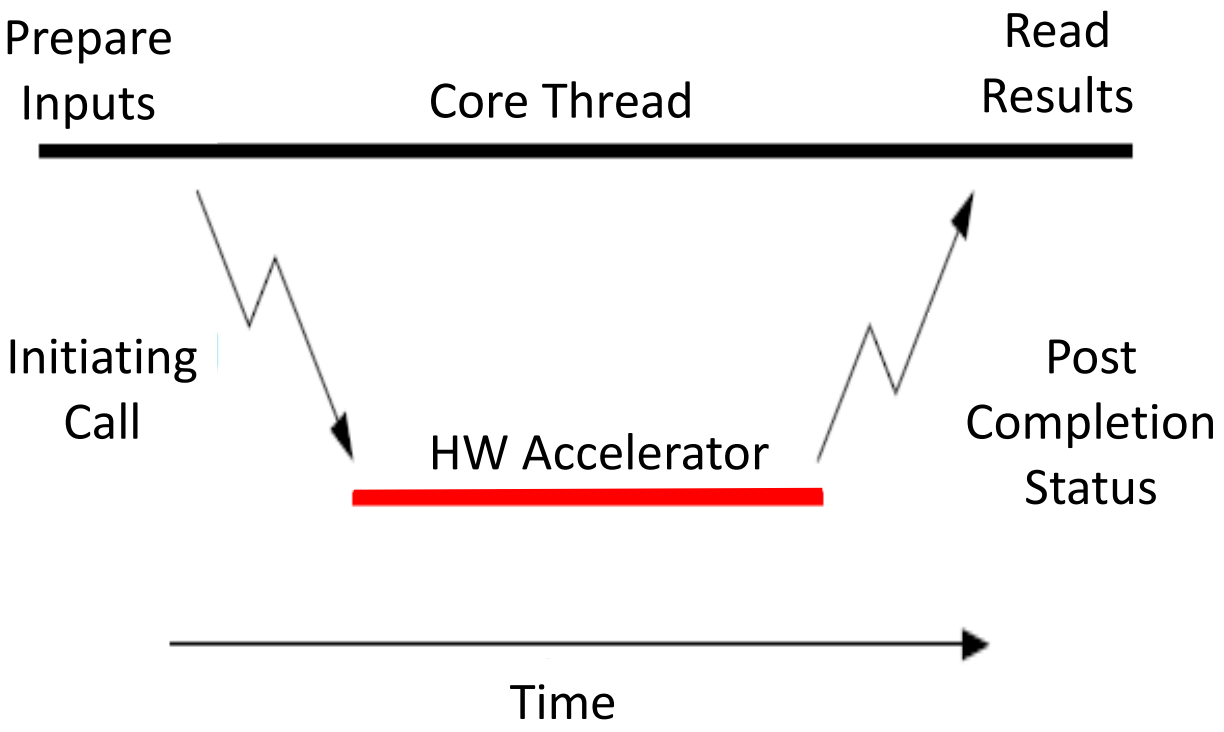
- Hardware Accelerator Optimized for Oracle Database In-Memory
  - Task Level Accelerator that Operates on In-Memory Columnar Vectors
  - Operates on Decompressed and Compressed Columnar Formats
- Query Engine Functions
  - In-Memory Format Conversions
  - Value and Range Comparisons
  - Set Membership Lookups
- Fused Decompression + Query Functions Further Reduce Task Overhead, Core Processing Cycles and Memory Bandwidth per Query

M7 Query Accelerator Engine



- Eight In-Silicon Offload Engines
- Cores/Threads Operate Synchronous or Asynchronous to Offload Engines
- User Level Synchronization Through Shared Memory
- High Performance at Low Power

M7 Accelerator Fine-grain Synchronization

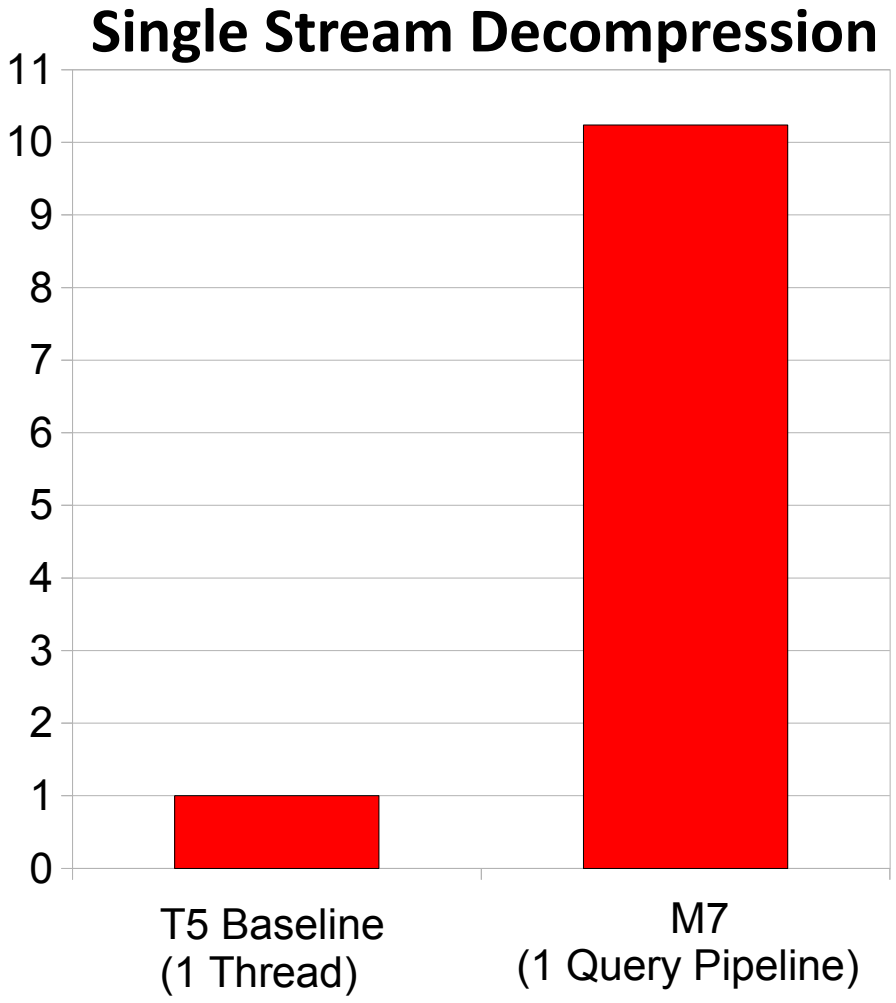


- Core Thread Initiates a Query Plan Task to Offload Engine
- User-Level LDMONITOR, MWAIT
  - Halts Hardware Thread for Specified Duration
  - Thread is Re-activated Once Duration Expires or Monitored Memory Location is Updated

Offload Engine Completion

- Results Written Back to Memory or Target L3$ Partition
- Completion Status Posted to Monitored Memory Location
- MWAIT Detection Hardware Resumes Core Thread Execution

M7 Query Offload Performance Example

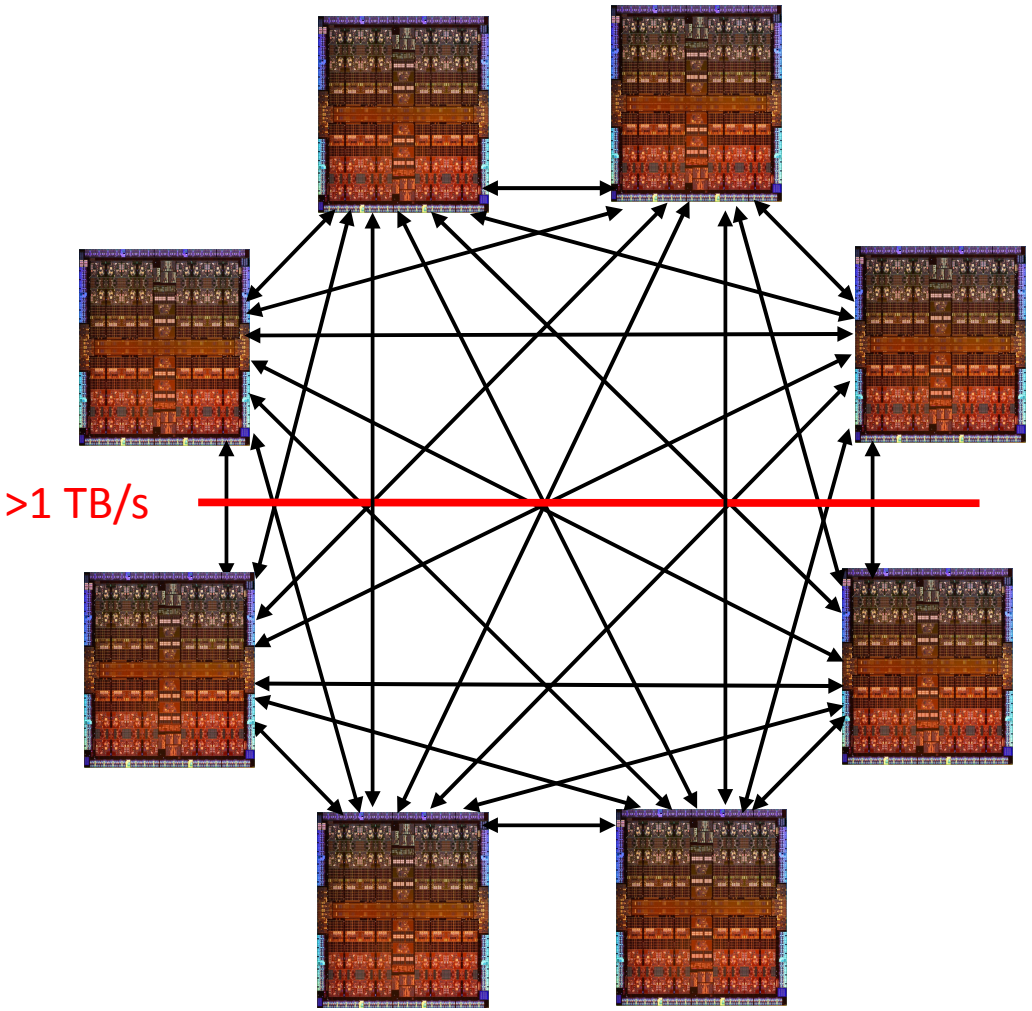


- SPARC T5, M7 & Oracle Database In-Memory
- Single Stream Decompression Performance
  - Decompression Stage of Query Acceleration
  - Unaligned Bit-Packed Columnar Formats
  - 1 of 32 Query Engine Pipelines
- M7 Hardware Accelerator Fuses Decompression Output with Filtering Operations
  - Further Accelerates the “WHERE” Clause in SQL Query
  - In-line Predicate Evaluation Preserves Memory Bandwidth
- Business Analytics Performed at System Memory Bandwidth

SPARC Accelerated Program

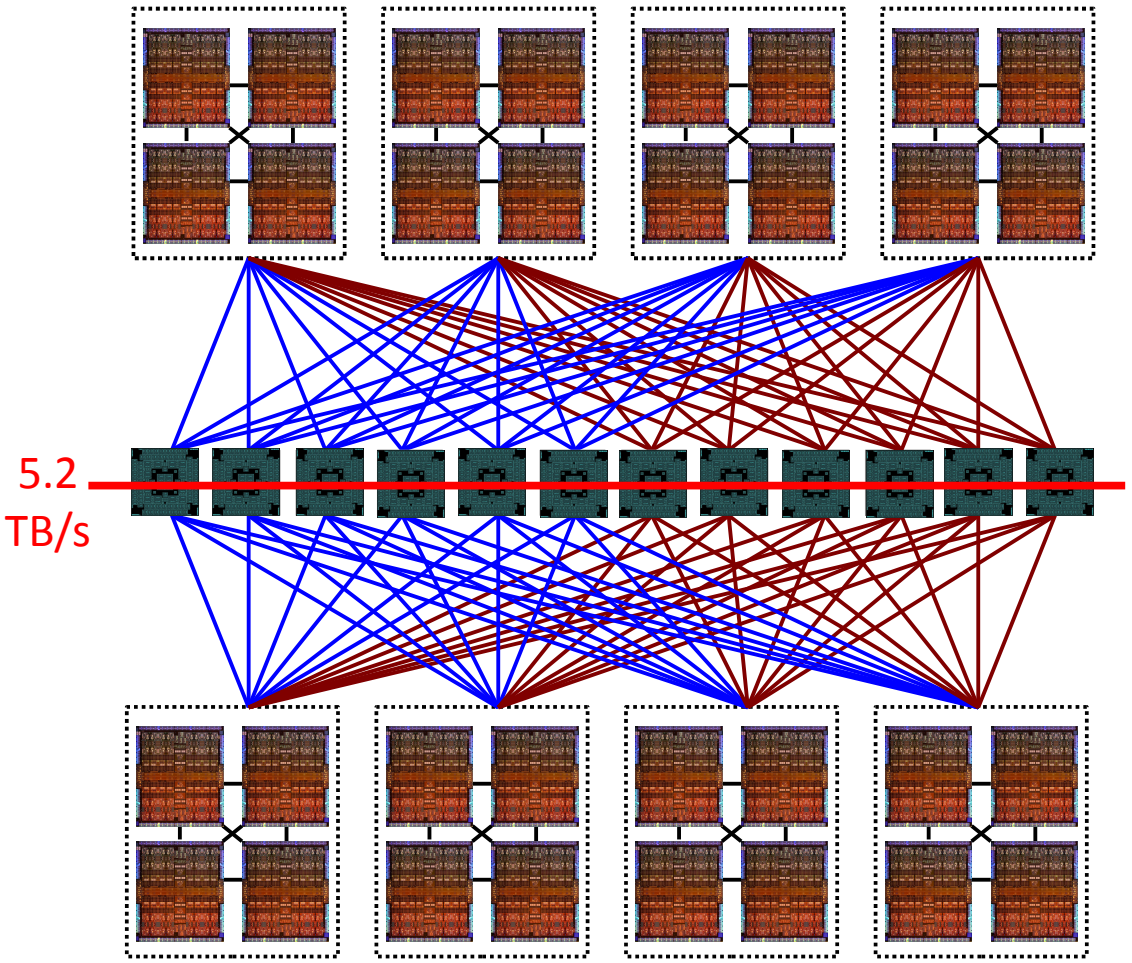
- Enables Third Party Software to Utilize SPARC Application Acceleration Features
- Application Data Integrity Support in Future Solaris and DB Memory Allocators, Compiler and Tools Chain
- Future Solaris Support for Memory Migration and Virtual Address (VA) Masking
- Query Engine Software Libraries

M7 SMP Scalability



- Fully Connected 8 Processor SMP
  - Up to 256 Cores, 2K Threads, 16TB Memory
  - >1TB/s Bisection Payload Bandwidth
- Fully Connected 2/4 Processor SMP Utilizing Link Trunking
- Directory-based Coherence
- 16Gbps to 18Gbps Link Rates
- Link Level Dynamic Congestion Avoidance
  - Alternate Path Data Routing
  - Based on Destination Queue Utilization
- Link Level RAS
  - Auto Frame Retry
  - Auto Link Retrain
  - Single Lane Failover

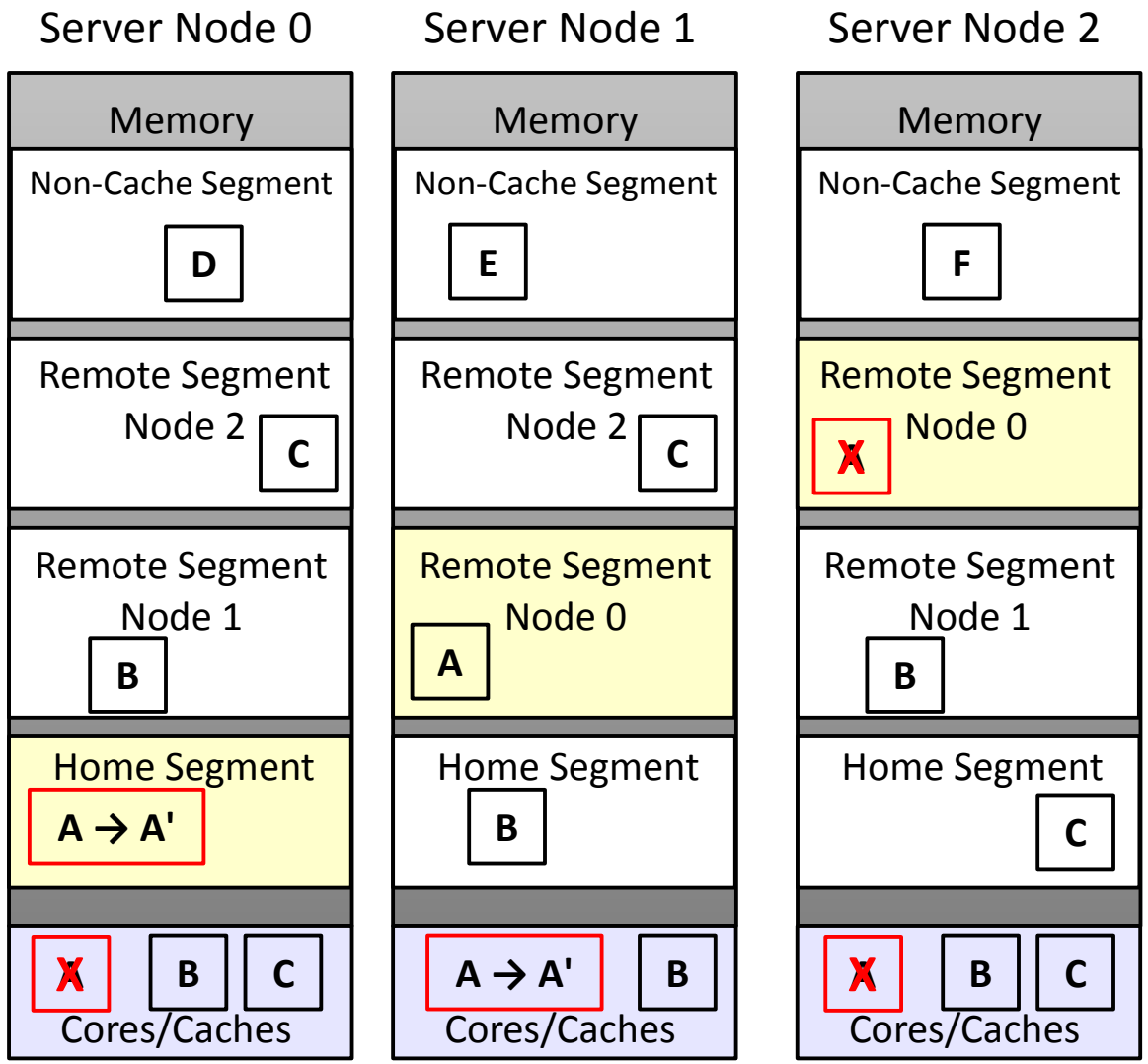
M7 SMP Scalability



- SMP Scalability to 32 Processors
  - Up to 1K Cores, 8K Threads
  - Up to 64TB Memory
- 64 Port Switch ASIC's
  - Divided into 2 Switch Groups of 6 Switch ASICs
  - 5.2TB/s Payload Bandwidth
  - 4X Bisection Payload Bandwidth over M6
- Physical Domains of 4 Processors Each
  - Fully Connected Local Topology
  - Dynamically Combine Processor Domains
- Fully Connected Switch Topology
  - 2 Links Connecting Every Processor and Switch
  - Coherence Directory Distributed Among Switches
  - Latency Reduction over M6 Generation
- SMP Interconnect RAS
  - Link Auto Retry & Retrain, Single Lane Failover
  - Link Level Multipathing
  - Operational with 5 of 6 ASIC's per Switch Group

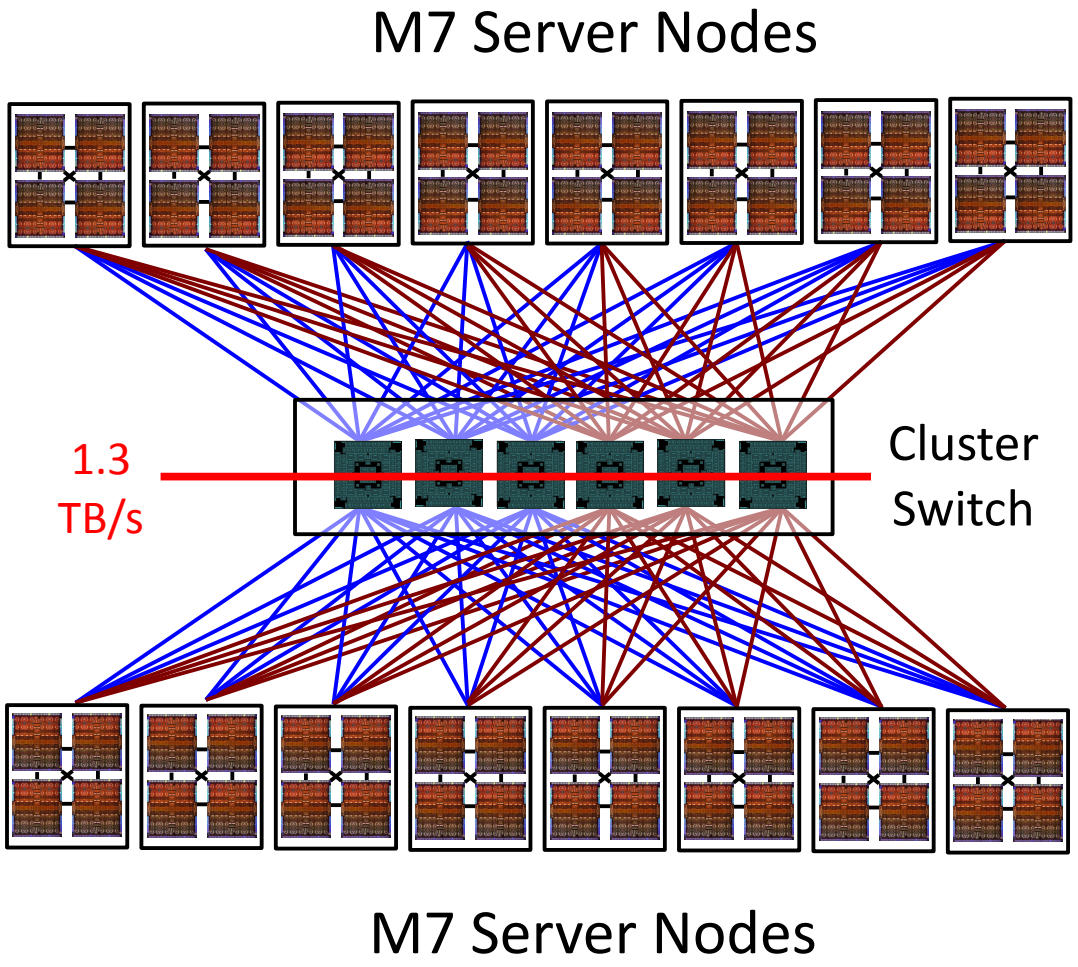
ORACLE

M7 Coherent Memory Clusters



- Highly Reliable and Secure Shared Memory Clustering Technology
  - Access Remote Node Memory Using Load/Store/Atomics, Relaxed Ordering
  - Integrated User Level Messaging and RDMA
- Cacheable Memory Segments
  - Remote Segments Cached in Local Node Memory & Caches at 64 Byte Granularity
  - Load-Hit Latency Accessing Remote Segments Same as Local Node Memory and Caches
  - Committed Stores to Remote Segments Update Home Node and Invalidate Remote Nodes
- Non-Cacheable Memory Segments
  - Always Access Remote Node Memory
- Cluster-wide Security
  - 64-bit Access Key Per Remote Request
  - Memory Version Checking Across Cluster

M7 Coherent Memory Clusters



- Up to 64 Processor Cluster
  - Combinations of 2P, 4P or 8P Server Nodes
  - Leverages M7 SMP HW and Interconnect
- Coherent Memory Cluster Protocol
  - Application Committed Stores Remain Consistent at the Home Node in Face of Requester Failure
  - 2-Party Dialogs for Failure Isolation
- 1.3TB/s Bisection Payload Bandwidth
- Self Redundant Cluster Switch
  - 64 Port Switch ASIC's
  - 6 Switching Paths Between Each Processor Pair, Divided in 2 Groups
  - Fault Tolerant Design Allowing Operation with a Single Switching Path per Group
  - Automatic Link and Switch Failover Without Involving Application Software

M7 Summary

Extreme Performance

Significant Increase in Processor Performance

Further Increase Core and Thread Performance

Increased Bandwidths Across Caches, Memory, Interconnects and I/O

Very Large Memory

Computing Efficiency

Increased Virtualization Density

Low Latency Application Migration

Flexible Logical and Physical Partitioning

Fine-grain Power Management

Optimized for Oracle Software

Improved Security and Reliability via Real-time Application Data Integrity

Concurrent Object Migration and Pointer Coloring

Database In-Memory Columnar Decompression, Query Offload and Coherent Memory Clusters

M7: Next Generation SPARC

Hotchips 26 – August 12, 2014

Stephen Phillips
Senior Director, SPARC Architecture
Oracle

ORACLE

Acronyms

- ADI: Application Data Integrity
- ALU: Arithmetic Logic Unit
- BRU: Branch Unit
- FA: Fully Associative
- FGU: Floating Point & Graphics Unit
- LSU: Load/Store Unit
- PA: Physical Address
- RA: Real Address
- SA: Set Associative
- SMP: Shared Memory Multiprocessor
- SPU: Stream Processing Unit
- TLB: Instruction or Data Translation Lookaside Buffer
- VA: Virtual Address