Oracle's Sonoma Processor: Advanced Low-cost SPARC Processor for Enterprise Workloads

HotChips 27 – Aug 24, 2015

Basant Vinaik

Senior Principal Engineer, CPU & I/O Verification

Rahoul Puri

Senior Architect, Networking & Low Latency I/O

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.

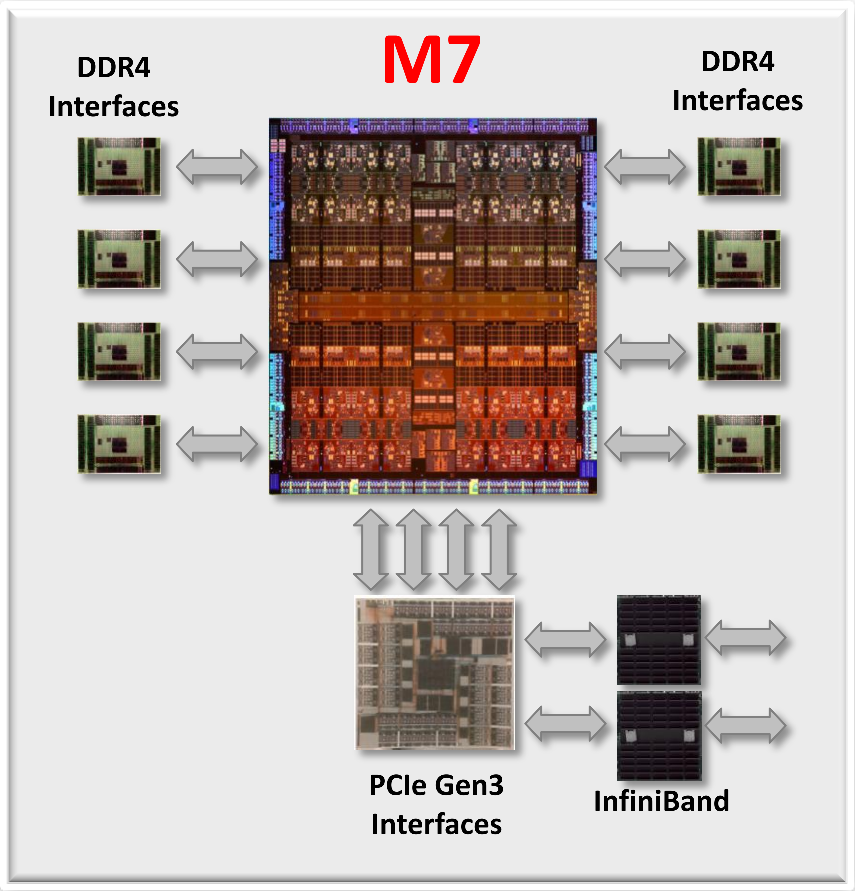
Oracle’s Sonoma Strategy

Extends SPARC portfolio to provide enterprise class performance and Software in Silicon features in significantly lower-cost form factors

Provides high level of system integration, excellent throughput, low memory latency, and high bandwidth IO interconnect

Delivers uncompromising price/performance for horizontal scale database, middleware, and cloud computing workloads

Fully Integrated to Lower Latency, Power, and Cost for Scale-Out

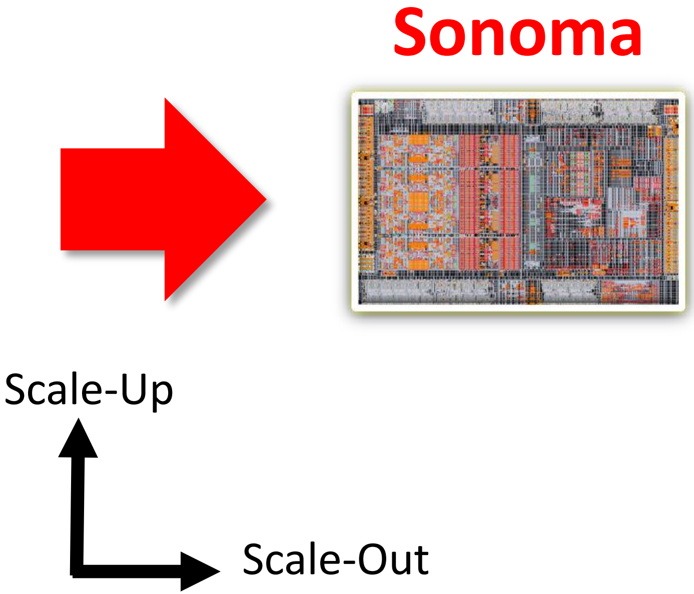

M7

DDR4 Interfaces

DDR4 Interfaces

PCIe Gen3 Interfaces

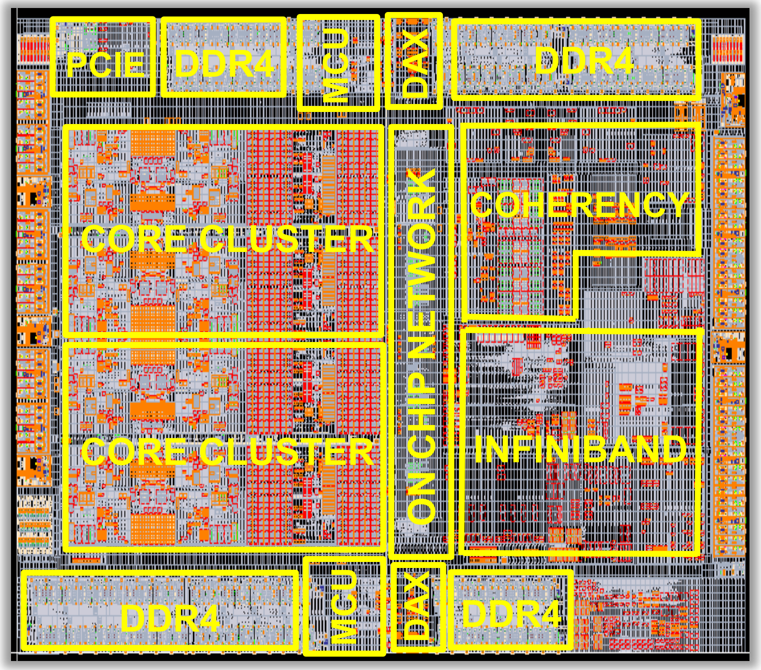
InfiniBand


Sonoma

Scale-Up

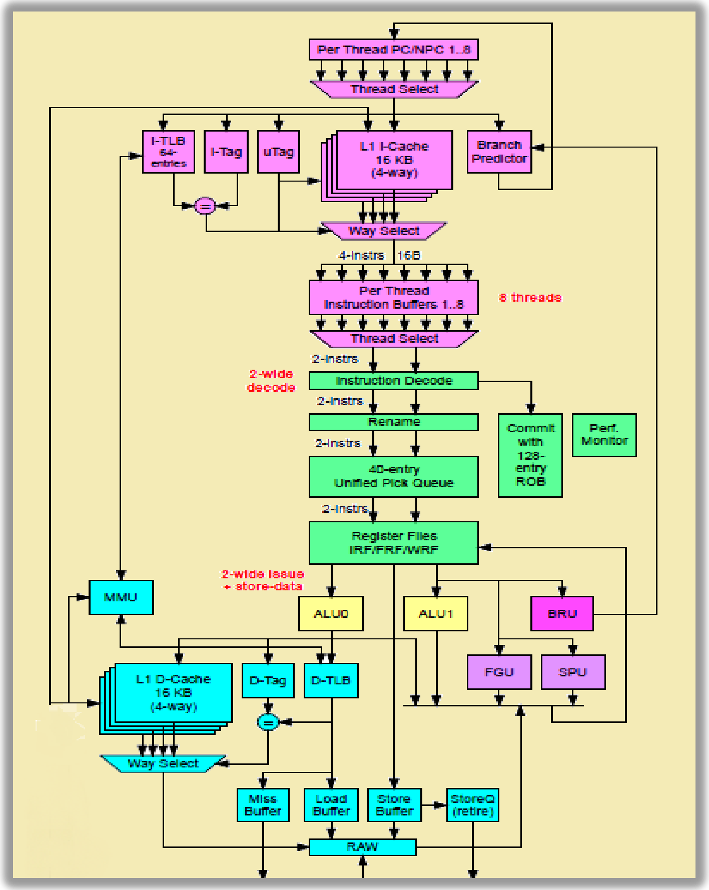
Scale-Out

Sonoma Processor



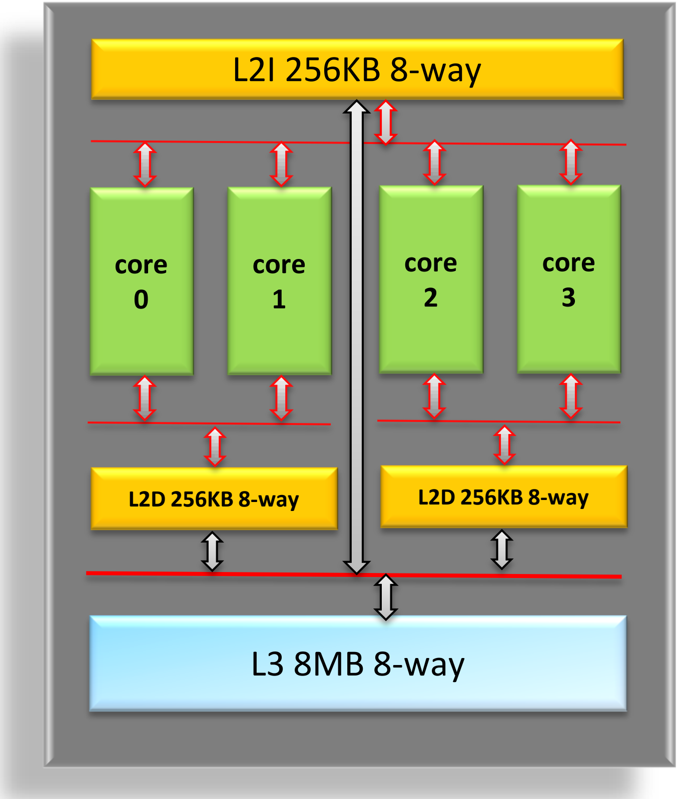
- 8 SPARC 4th generation cores
- Optimized cache organization
- Advanced Software in Silicon features
  - Real-time Application Data Integrity (ADI)
  - Concurrent Memory Migration and VA Masking
  - DB query offload engines
- Direct attached DDR4 memory
- Integrated PCIe Gen3
- Integrated InfiniBand HCA
- Scale-out IB interconnect
- Technology: 20nm, 13 Metal Layers

Enterprise Class Core with Crypto and Software in Silicon Features



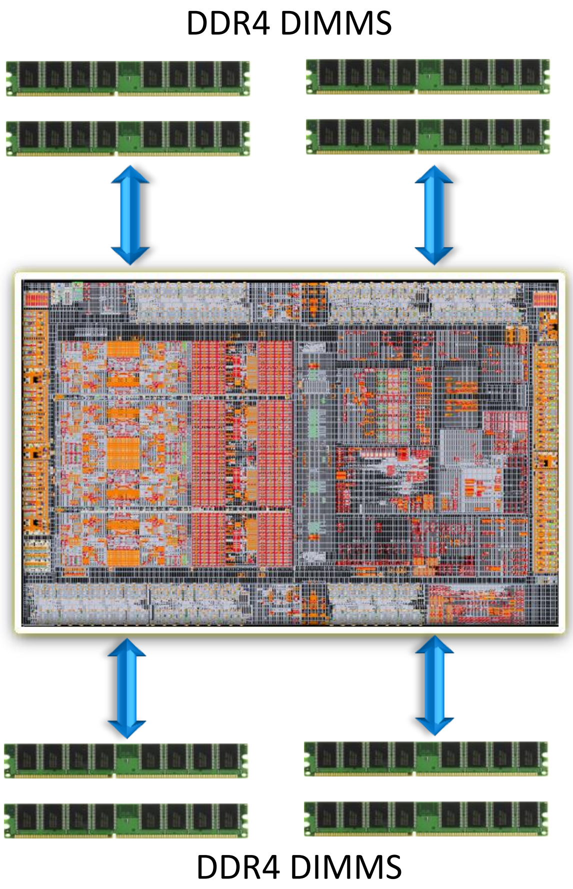
- Dynamically threaded, 1 to 8 Threads
- Dual-Issue, OOO execution core
  - 2 ALU, 1 LSU, 1 FGU, 1 BRU, 1 SPU
  - 40 entry Pick Queue
  - 64 entry FA I-TLB, 128 entry FA D-TLB
  - 54bit VA, 50bit RA/PA
- Integrated cryptographic unit
  - User level crypto instructions support:
    - AES, DES, 3DES, Camellia, CRC32c
    - MD5, RSA, DH, DSA, ECC
    - SHA-1, SHA-224, SHA-256, SHA-384, SHA-512
  - Provides security and transparent encryption across Oracle software stack
- Fine-grain power estimator to lower TDP
- Application acceleration with Software in Silicon features

Cache Hierarchy Optimized for Latency and Throughput



- Two core clusters with 4 cores/cluster
- Core private L1$
  - 16KB, 64B line, 4-way SA L1 I-$
  - 16KB, 32B line, write-through, 4-way SA L1 D-$
- Shared L2-I$
  - 8-way SA, 64B Lines, >500GB/s throughput
- Core pair shared writeback L2-D$
  - 8-way SA, 64B lines, >500GB/s throughput per L2-D$
- Shared & partitioned L3$
  - 8MB local partitions designed to reduce latency and improve performance
  - Cache lines can be replicated or victimized between L3$ partitions
  - HW accelerators, PCIe DMA, and IB DMA can directly allocate lines into targeted L3$ partition

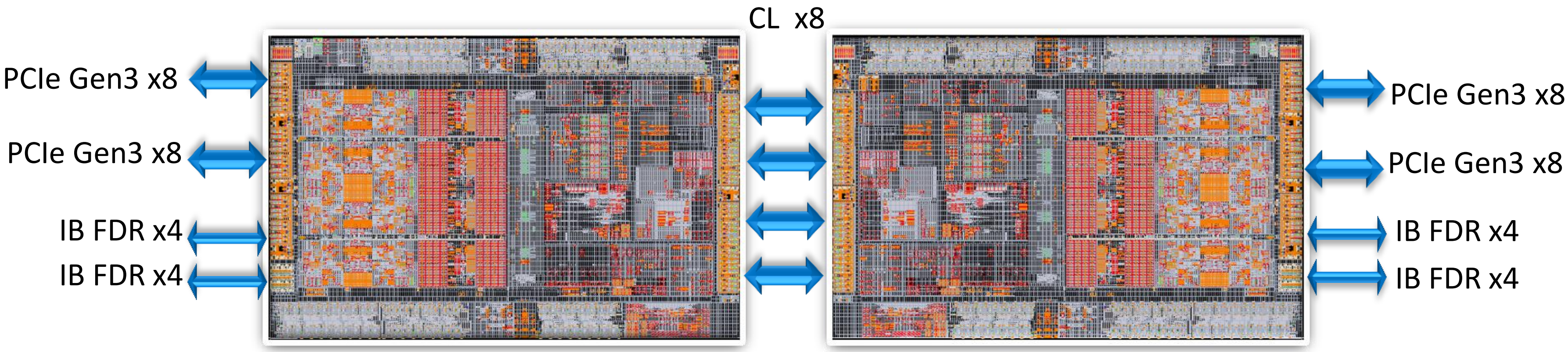
Direct Attached Low Latency Memory



- 2 DDR4 memory controllers
  - 4 direct attached DDR4-2133/2400 channels
  - Up to 2 DIMMS per channel
  - Up to 1TB memory per socket
  - 77GB/s peak memory bandwidth
  - Support for DIMM retirement

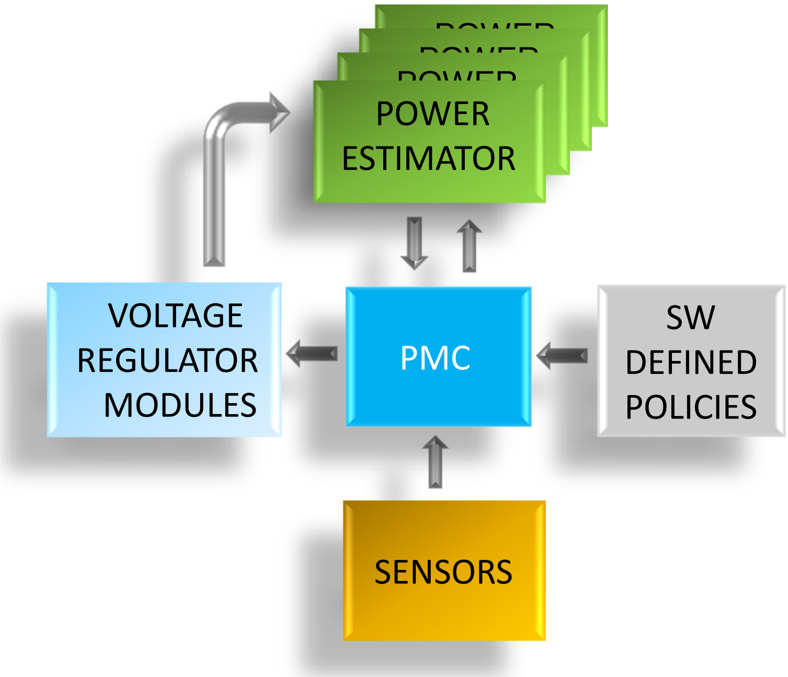
- Speculative memory read to reduce latency
  - Reduces local memory latency by pre-fetching data on local L3$ partition miss
  - Dynamic per request, based on history (data, instruction), and controlled by threshold settings

Connectivity Optimized for Scale-Out



- 2 InfiniBand links @ FDR (56Gbps)
  - Low latency scale-out networking interconnect for DB and clusters
  - 28 GB/s Bidirectional Bandwidth
- 2 PCIe links @ Gen3 (64Gbps)
  - 32 GB/s Bidirectional Bandwidth
- 4 Scale-Up Coherence links @ 16Gbps (128Gbps)
  - 128 GB/s bidirectional bandwidth
  - Auto frame retry, auto link retrain, and single lane failover

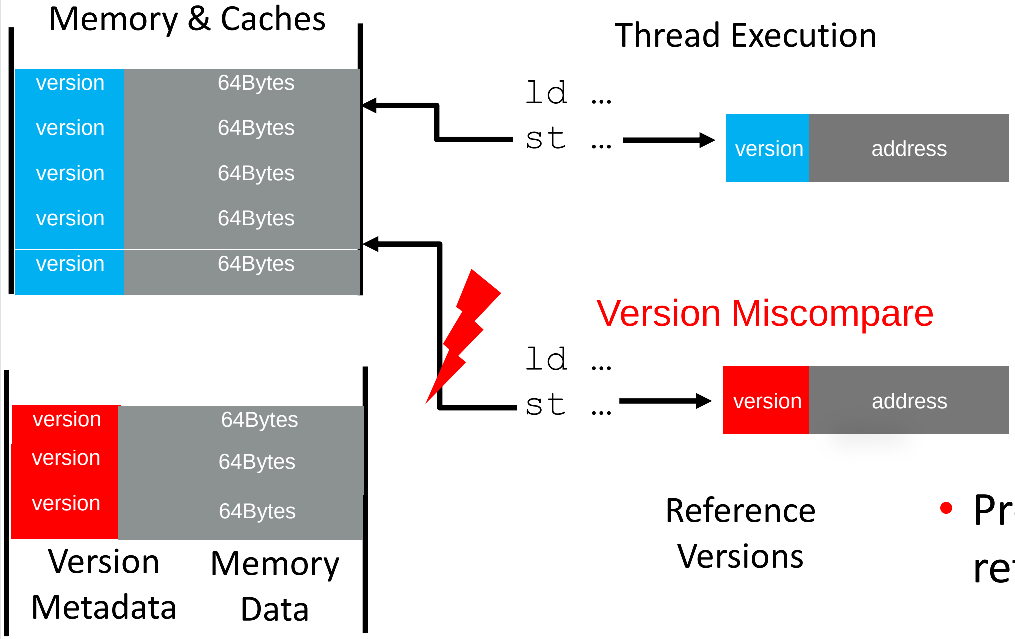
Fine-grain Power Management to Lower TDP



- On-die power estimator in each core and L3$
  - Tracks internal activity to estimate dynamic power
  - Self Governing Core (SGC) and L3 cache
  - Estimates updated at 250 nanosecond intervals
- On-die Power Management Controller (PMC)
  - Estimates total power of cores, caches, and SOC
  - Accurate to within a few percent of measured power
  - Dynamically adjusts voltage and/or frequency within core clusters based on software defined policies
- Power management policies
  - Power, current, temperature, and subsystem capping

➢ Lowers TDP and simplifies system design which in turn lowers cost

Real-time Application Security with ADI

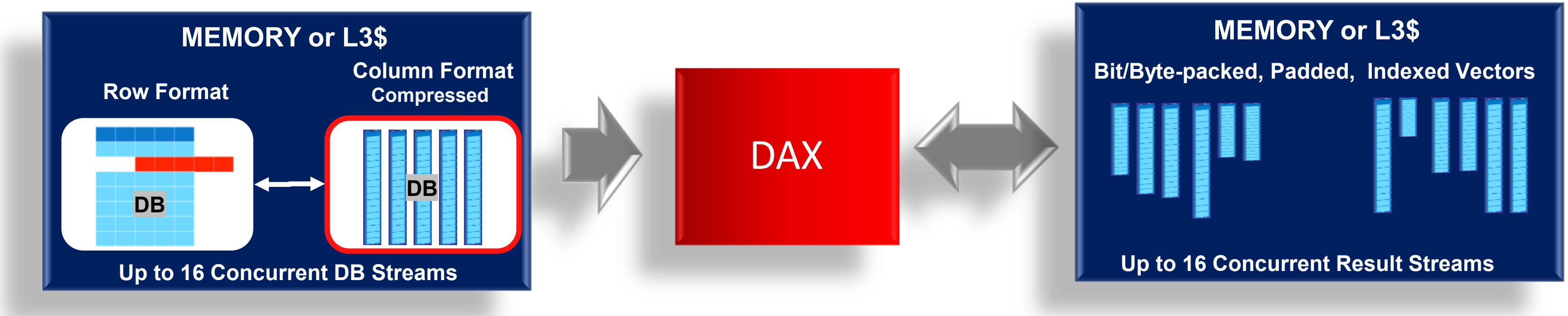


- Real-time Application Data Integrity (ADI)
  - Version metadata associated with 64Byte aligned memory data
  - Version stored in memory and maintained throughout the cache hierarchy and all interconnects
  - Memory Version checked against Reference Version by core Load/Store Units
  - Useful for both production and code development
- Protects against software Invalid/Stale references and buffer overruns

➢Real-time Application Security against malicious attacks like HeartBleed

➢ Secure cloud computing via ADI and Crypto

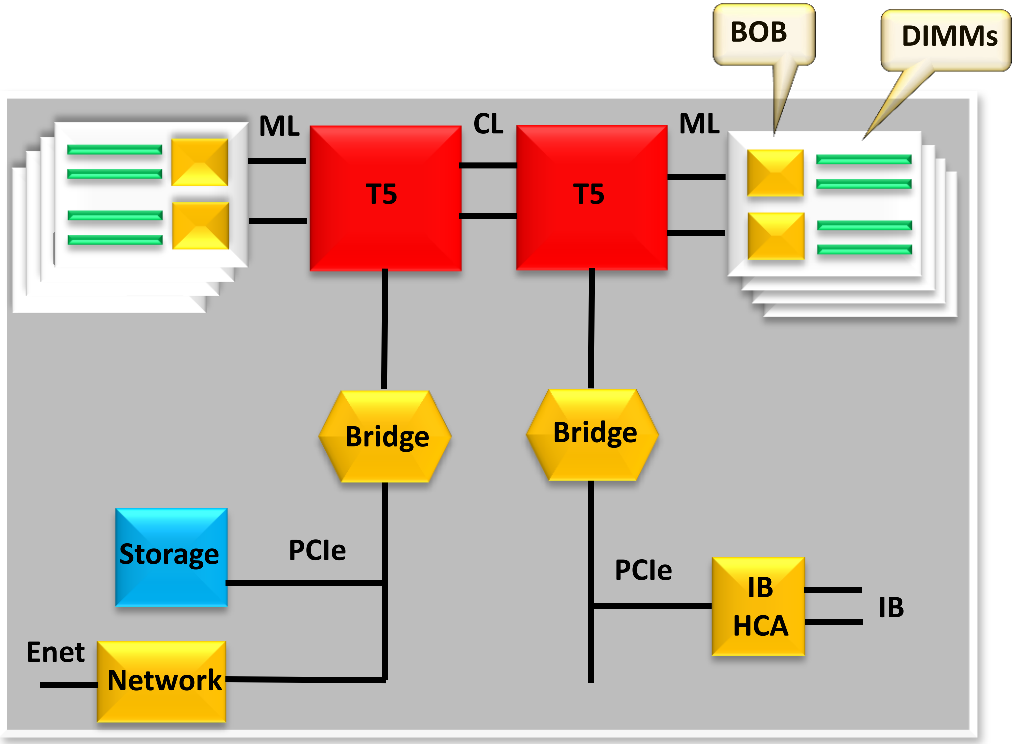
Database Accelerator (DAX) for Business Analytics

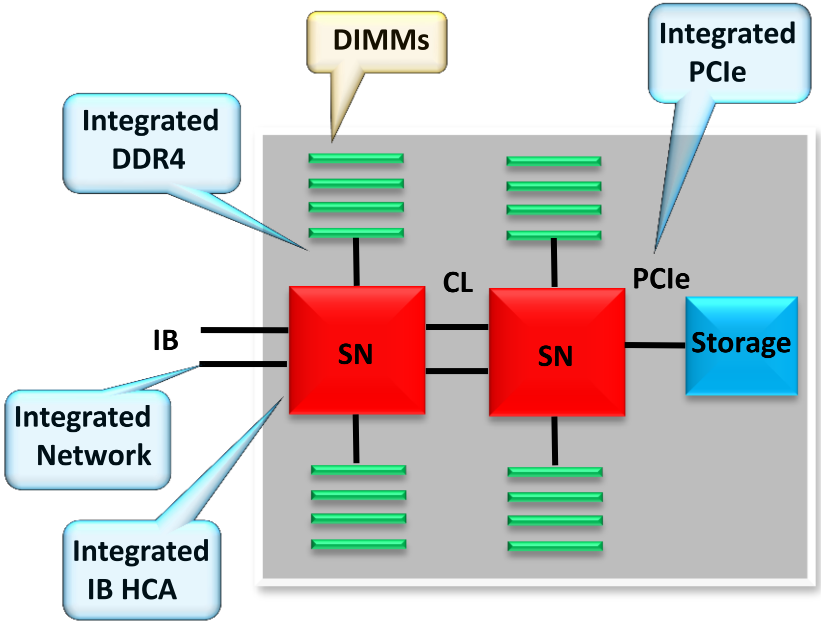

MEMORY or L3$
Row Format
Column Format Compressed
DB
DB
Up to 16 Concurrent DB Streams
DAX
MEMORY or L3$
Bit/Byte-packed, Padded, Indexed Vectors
Up to 16 Concurrent Result Streams

- 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
  - Applications submit work using Hypervisor API and synchronize using shared memory
- Query Engine Functions
  - In-Memory format conversions, value and range comparisons, and set membership lookups
- Inline decompression with query functions to improve performance

➢ Performs business analytics at system memory bandwidth

System Integration for Sonoma Compute Node

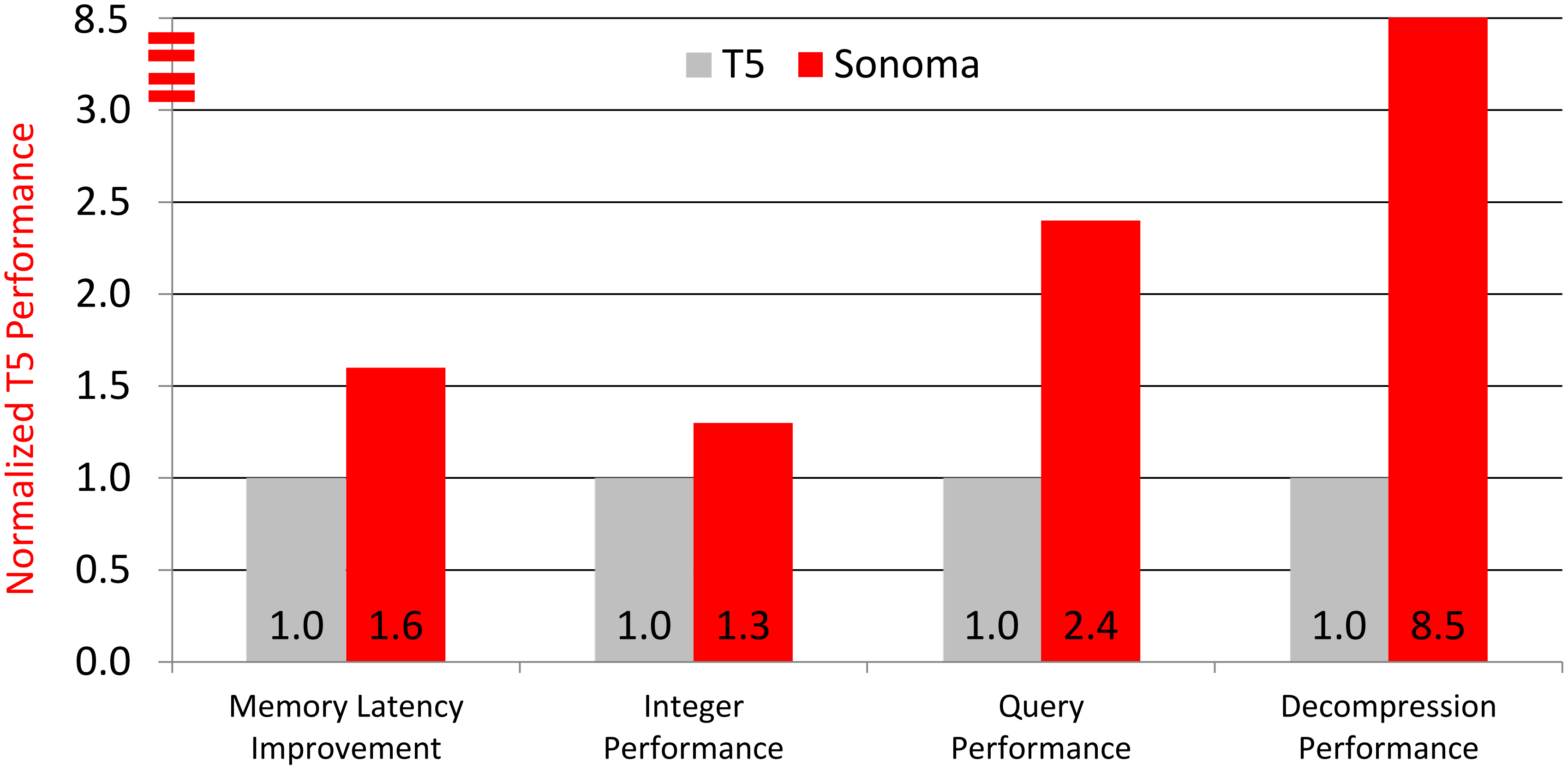




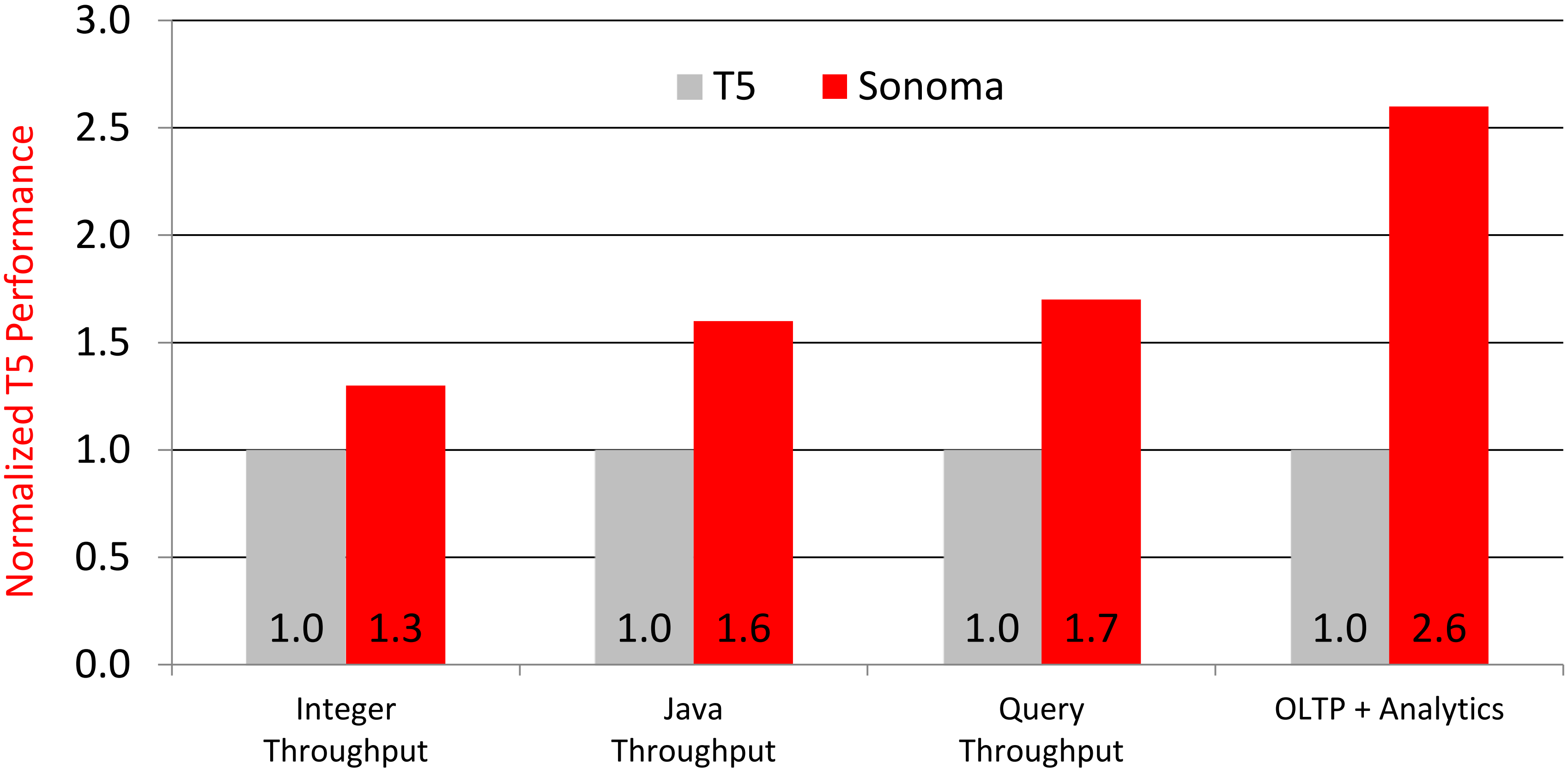
- More rack space
- More components and links
- Higher power
- Higher cost

- Less rack space
- Less components and Links
- Lower power
- Lower cost

T5 vs. SN: Single Thread Performance



T5 vs. SN: Per Core Performance



Sonoma Low Latency I/O Features

Delivers compelling and differentiated networking in lower cost systems

Serves as a highly scalable low latency backbone for enterprise, DB RAC cluster & cloud

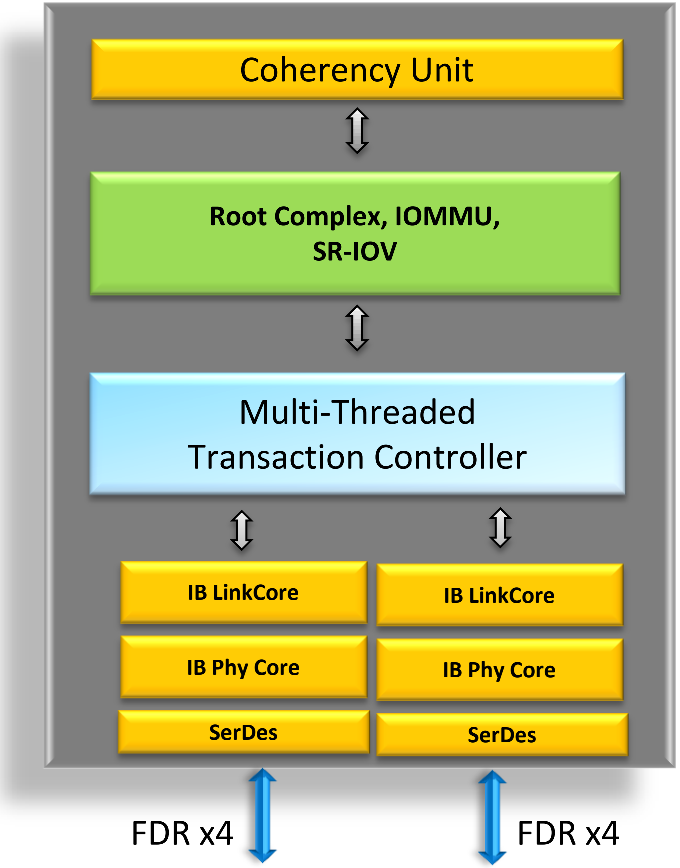
10x improvement in packet rate and predictable QoS for 100k+ processes

Reduces Memory Registration overhead for InfiniBand RDMA

Resource Scaling for large number of connections enables user-level IPC

Consolidates storage, networking, and IPC fabric

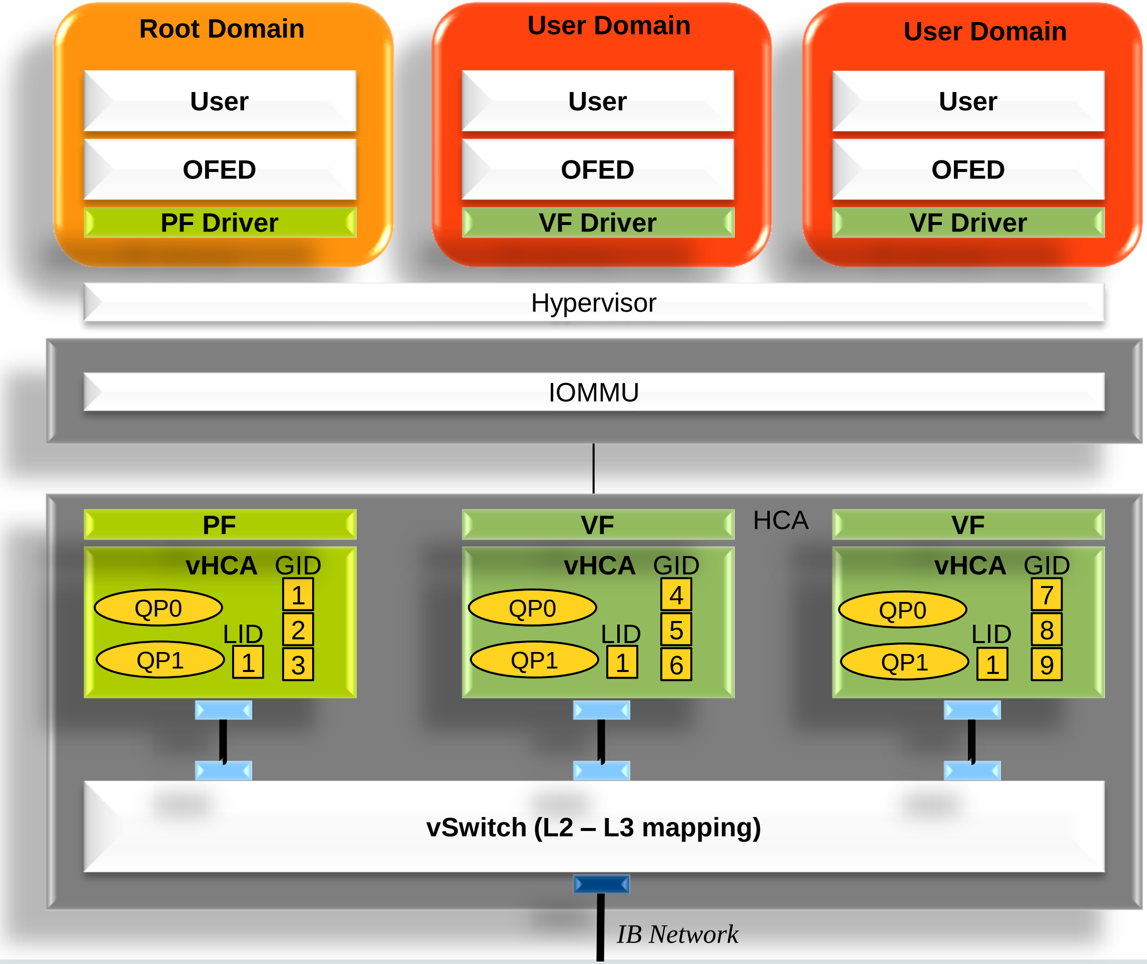
Sonoma Integrated InfiniBand HCA



- OFED Compliant IB HCA with 2 x4 FDR (56 Gbps) Ports
- SR-IOV EP with 32 Virtual Functions
- vHCA Virtualization with embedded vSwitch
- 16 M Queue Pairs per vHCA
- Line-Rate packet classification for Active-Active FDR
- Conditional RDMA
- Virtual Cut-Through messaging
- InfiniBand transport support (UD, UC, RC and XRC)
- HW assisted reliable multicast
- IP offloads
  - Checksum, LSO/TSO, RSS, Header/Split, Packet Classification
- IP Security features
  - ARP spoofing, VLAN, SMAC, vNIC enforcement

IB HCA Virtualization

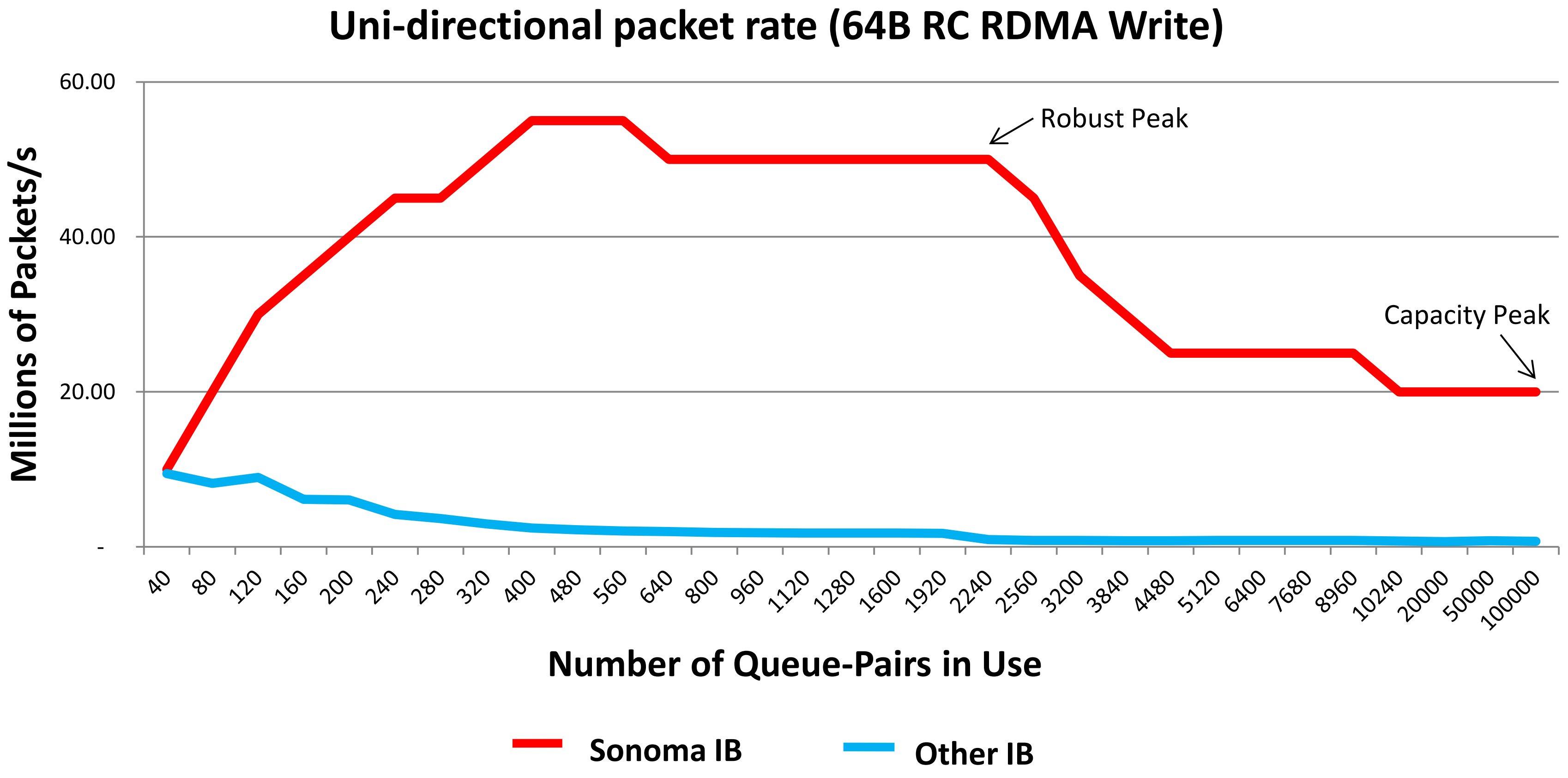
- Host observes
  - Each VF is complete vHCA
  - LID, GID Table, 16Million QPs
- Network observes
  - Multiple HCAs behind L2/L3 switch
- L2/L3 Switch
  - LID/GID mapping
  - VM-VM on same physical HCA through GID
- Advantages
  - Transparent virtualization
  - No re-configuration of L2 switch forwarding tables throughout fabric
  - Direct Device access



Differentiated Networking Optimized for Scale-Out

- Multi-threaded InfiniBand Transaction Controller
  - Scaling: 20M+ messages/sec rate sustained with 100K+ reliable connections
  - Optimized IB HW/SW interface provide direct access to HW resources for 16K simultaneous active processes
- Virtualization
  - vHCA virtualization with embedded vSwitch provide a full, private QP space for all virtual functions, and enable Live Migration of RDMA ULPs
  - Hardware enforced Security and Isolation for Guest Domains
  - Virtualized SMAs and GSAs
- MMU with Shared Page Tables support
  - Fast-Path Work Request based invalidations
- Multi network protocol support on a single interconnect
  - Optimized for Database Real Application Cluster (RAC) and storage applications
  - LAN, SAN, WAN, and IPC across single interconnect

Sonoma Queue-Pair Scaling and Packet Rate



Sonoma: The Perfect Choice for Scale-Out

Cost

High system integration: networking, memory, fabric

Mainstream volume process technology

Mainstream TDP

Hardware offloads

Convergence

Direct attached memory

Integrated PCIe

Integrated InfiniBand

Lower latency, higher bandwidth

Cloud

Real-time application security

Excellent throughput

Software in Silicon

Optimized for Oracle software

Acronyms

- ADI: Application Data Integrity
- ALU: Arithmetic Logic Unit
- BRU: Branch Unit
- DPC: DIMMs Per Channel
- EoIB: Ethernet Over InfiniBand
- FA: Fully Associative
- FGU: Floating Point & Graphics Unit
- FDR: Fourteen Data Rate (14Gbps)
- HCA: Host Channel Adapter
- IPoIB: Internet Protocol Over InfiniBand
- LSU: Load/Store Unit
- LDOM: Logical Domain
- OFED: Open Fabrics Enterprise Distribution
- PA: Physical Address
- PMC: Power Management Controller
- QOS: Quality Of Service
- RA: Real Address
- RC: Reliable Connection
- RDMA: Remote Direct Memory Access
- SA: Set Associative
- SR-IOV: Single Root IO Virtualization
- SMP: Shared Memory Multiprocessor
- SPU: Stream Processing Unit
- TLB: Instruction or Data Translation Lookaside Buffer
- UC: Unreliable Connection
- UD: Unreliable Datagram
- VA: Virtual Address
- XRC: Extended Reliable Connection

Hardware and Software

ORACLE®

Engineered to Work Together