Design and implementation of Sun's Niagara2 Processor

Umesh Nawathe
Senior Manager, Sun Microsystems
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Memory Bottleneck

Relative Performance

- CPU Frequency
- DRAM Speeds

2x Every 2 Years

Gap

2x Every 6 Years

Date (Year)

Comparing Modern CPU Design Techniques

- ILP Offers Limited Headroom
- TLP Provides Greater Performance Efficiency
For a single thread:
- Memory is THE bottleneck for improving performance.
- Commercial server workloads exhibit poor memory locality.
- Only a modest throughput speedup is possible by reducing compute time.
- Conventional single-thread processors optimized for ILP have low utilizations.

With many threads:
- It’s possible to find something to execute every cycle.
- Significant throughput speedups are possible.
- Processor utilization is much higher.
Microprocessor Power Evolution

Power Density (W/cm²) over Time

![Graph showing the evolution of power density over time for various microprocessors. The graph includes different markers and colors for Intel, Pentium, Itanium, Alpha, Sparc, PowerPC, and AMD x86 64. Points N1 and N2 are highlighted.]
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Niagara2's Key features

- 2nd generation CMT (Chip Multi-Threading) processor optimized for Space, Power, and Performance (SWaP).
- 8 Sparc Cores, 4MB shared L2 cache; Supports concurrent execution of 64 threads.
- >2x UltraSparc T1's throughput performance and performance/Watt.
- >10x improvement in Floating Point throughput performance.
- Integrates important SOC components on chip:
  > Two 10G Ethernet (XAUI) ports on chip.
  > Advanced Cryptographic support at wire speed.
- On-chip PCI-Express, Ethernet, and FBDIMM memory interfaces are SerDes based; pin BW > 1Tb/s.
Niagara2 Block Diagram

Key Point: System-on-a-Chip, CMT architecture => lower # of system components, reduced complexity/power => higher system reliability.
Sparc Core (SPC) Architecture Features

- Implementation of the 64-bit SPARC V9 instruction set.

- Each SPC has:
  - Supports concurrent execution of 8 threads.
  - 1 load/store, 2 Integer execution units.
  - 1 Floating point and Graphics unit.
  - 8-way, 16 KB I$; 32 Byte line size.
  - 4-way, 8 KB D$; 16 Byte line size.
  - 64-entry fully associative ITLB.
  - 128-entry fully associative DTLB.
  - MMU supports 8K, 64K, 4M, 256M page sizes; Hardware Tablewalk.
  - Advanced Cryptographic unit.

- Combined BW of 8 Cryptographic Units is sufficient for running the 10 Gb ethernet ports encrypted.
SPC Architecture Features (Cont'd.)

• 8-stage Integer Pipeline (Fetch, Cache, Pick, Decode, Execute, Memory, Bypass, Writeback).
  > 3-cycle load-use latency.

• 12-stage FP and Graphics Pipeline (Fetch, Cache, Pick, Decode, Execute, FX1, FX2, FX3, FX4, FX5, FB, FW).
  > 6-cycle latency for dependent FP operations.
  > Longer pipeline for Divide/Sqrt.

• Upto 4 instructions fetched per cycle in the 'Fetch' stage.

• Has 2 thread-groups (TGs); 'Pick' tries to find 2 instructions to execute every cycle – one per TG.
  > Can lead to hazards (e.g. Loads picked from both TGs).

• 'Decode' stage resolves hazards that 'Pick' cannot.
SPC Architecture Features (Cont'd.)

• Integer/Ld/St pipeline shown.

Key Point:
Different threads can occupy different pipeline stages in a given cycle with very few restrictions.

Thread Group 0
- IB0-3
- P0
- D2
- E0
- M3
- B1
- W2

Thread Group 1
- IB4-7
- P5
- D7
- E6
- M4
- B7
- W6

F = Fetch
C = Cache
P = Pick
D = Decode
E = Execute
M = Memory
B = Bypass
W = Writeback

Numbers 0 through 7 represent the thread identifier.
Niagara2 Die Micrograph

- 8 SPARC cores, 8 threads/core.
- 4 MB L2, 8 banks, 16-way set associative.
- 16 KB I$ per Core.
- 8 KB D$ per Core.
- FP, Graphics, Crypto, units per Core.
- 4 dual-channel FBDIMM memory controllers @ 4.8 Gb/s.
- X8 PCI-Express @ 2.5 Gb/s.
- Two 10G Ethernet ports @ 3.125 Gb/s.
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• **Physical Implementation**
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
### Physical Implementation Highlights

<table>
<thead>
<tr>
<th>Technology</th>
<th>65 nm CMOS (from Texas Instruments)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal Voltages</td>
<td>1.1 V (Core), 1.5V (Analog)</td>
</tr>
<tr>
<td># of Metal Layers</td>
<td>11</td>
</tr>
<tr>
<td>Transistor types</td>
<td>3 (SVT, HVT, LVT)</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.4 Ghz @ 1.1V</td>
</tr>
<tr>
<td>Power</td>
<td>84 W @ 1.1V</td>
</tr>
<tr>
<td>Die Size</td>
<td>342 mm^2</td>
</tr>
<tr>
<td>Transistor Count</td>
<td>503 Million</td>
</tr>
<tr>
<td>Package</td>
<td>Flip-Chip Glass Ceramic</td>
</tr>
<tr>
<td># of pins</td>
<td>1831 total; 711 Signal I/O</td>
</tr>
</tbody>
</table>
Level 2 Cache

- 4-MB shared L2 Cache:
  > 8 banks of 512 KB each.
  > 64 B line size; 16-way set associative.
  > Read 16 B per cycle per bank with 2-cycle latency.
  > Address hashing capability to distribute accesses across different sets.

- SEC DED ECC/parity protected.

- Data from different ways/words interleaved to improve SER.

- Tag arrays contain reverse-mapped directory:
  > Maintains L1 I$ and D$ coherency across 8 SPCs.
  > Store L2 Index/Way bits instead of all the tag bits.

- Memory cell NWELL power separated out as a test hook:
  > Helps identify weak memory bits susceptible to read-disturb fails due to PMOS NBTI effect.
  > Significantly improves DPPM/reliability.
Level2 Cache – Row Redundancy

- Redundancy implemented at 32-KB level.
- Spare rows for one array located in adjacent array.
- Adjacent array (which is normally not enabled) is enabled if 'incoming address' = 'defective row address'.
- Reduces X-decoder area by ~30%.
Crossbar

- Provides high-BW interface between 8 SPCs and 8 L2 cache banks/NCU.
- Consists of 2 blocks:
  > PCX (Processor to Cache/NCU transfer): 8-i/p, 9-o/p mux.
  > CPX (Cache/NCU to Processor transfer): 9-i/p, 8-o/p mux.
- PCX/CPX combined provide Rd/Wr BW of ~270 GB/s (Pin BW of ~400 GB/s).
- 4-stage pipeline: Request, Arbitration, Selection, Transmission.
- 2-deep queue for each source-destination pair to hold data transfer requests.
Key Point: Complex clocking; large # of clock domains; asynchronous domain crossings.
Clocking (Cont'd.)

- On-chip PLL generates Ratioed Synchronous Clocks (RSCs); Supported fractional divide ratios: 2 to 5.25 in 0.25 increments.

- Balanced use of H-Trees and Grids for RSCs to reduce power and meet clock-skew budgets.

- Periodic relationship of RSCs exploited to perform high BW skew-tolerant domain crossings.

- Clock Tree Synthesis used for Asynchronous Clocks; domain crossings handled using FIFOs and meta-stability hardened flip-flops.

- Cluster/L1 Headers support clock gating to save clock power.
Clocking (RSC domain crossings)

- FCLK = Fast-Clock
- SCLK = Slow-Clock
- Same 'Sync_en' signal for FCLK -> SCLK, and SCLK -> FCLK crossings.

**Key Point:** Equalizing setup and hold margins maximizes skew tolerance.
Niagara2's SerDes Interfaces

- All SerDes share a common micro-architecture.
- Level-shifters enable extensive circuit reuse across the three SerDes designs.
- Total raw pin BW in excess of 1Tb/s.
- Choice of FBDIMM (vs DDR2) memory architecture provides ~2x the memory BW at <0.5x the pin count.

<table>
<thead>
<tr>
<th></th>
<th>FBDIMM</th>
<th>PCI-Express</th>
<th>Ethernet-XAUI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Signalling Reference</strong></td>
<td>VSS</td>
<td>VDD</td>
<td>VDD</td>
</tr>
<tr>
<td><strong>Link-rate (Gb/s)</strong></td>
<td>4.8</td>
<td>2.5</td>
<td>3.125</td>
</tr>
<tr>
<td><strong># of North-bound (Rx) lanes</strong></td>
<td>14 * 8</td>
<td>8</td>
<td>4 * 2</td>
</tr>
<tr>
<td><strong># of South-bound (Tx) lanes</strong></td>
<td>10 * 8</td>
<td>8</td>
<td>4 * 2</td>
</tr>
<tr>
<td><strong>Bandwidth (Gb/s)</strong></td>
<td>921.6</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>
Niagara2's True Random Number Generator

- Consists of 3 entropy cells.
- Amplified n-well resistor thermal noise modulates VCO frequency; VCO o/p sampled by on-chip clock.
- LFSR accumulates entropy over a pre-set accumulation time.
  > Privileged software programs a timer with desired entropy accumulation time.
  > Timer blocks loads from LFSR before entropy accumulation time has elapsed.
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Niagara2's System on Chip Methodology

- Chip comprised of many subsystems with different design styles and methodologies:
  - Custom Memories & Analog Macros:
    - Full custom design and verification.
      - 40% compiled memories.
    - Schematic/manual layout based.
  - External IP:
    - SerDes full custom IP Macros.
  - Complex Clusters:
    - DP/Control/Memory Macro.
    - Higher speed designs.
  - ASIC designs:
    - PCI-Express and NIC functions.
  - CPU:
    - Integration of component abstracts.
    - Custom pre-routes and autoroute solution.
    - Proprietary RC analysis and buffer insertion methodology.

Key Point: Chip Design Methodologies had to comprehend blocks with different design styles and levels of abstraction.
Complex Design Flow

**Key Point:** Design Flow different for different design phases.

- Architectural pipeline reflected closely in the floorplanning of:
  - Memory Macros.
  - Control Regions.
  - Datapath Regions.

- Early Design Phase:
  - Fully integrated SUN toolset allows fast turnaround.
  - Less accurate, but fast - allows quick iterations to identify timing fixes involving RTL/floorplan changes.
  - Allows reaching route stability.

- Stable Design Phase:
  - More accurate, but not as fast, allows timing fixes involving logic and physical changes; Allows logic to freeze.

- Final Design Phase:
  - More accurate, but longer time to complete; More focus on physical closure then logic.

- Freeze and ECO Design Phases:
  - Allows preserving large portion of design from one iteration to next.
Design For Manufacturability (DFM)

- Single poly orientation (except I/O blocks).
- Larger-than-minimum design rules:
  - To minimize impact of poly/diffusion flaring.
  - Near stress-prone topologies to reduce chances of dislocations in Si-lattice.
  - Larger Metal overlap of via/contact where possible.
- Improved gate-CD control:
  - Dummy polys used for gate shielding.
  - Limited gate-poly pitches used; OPC algorithm tuned for them.
- OPC simulations of critical cell layouts to ensure sufficient manufacturing process margin.
- Extensive use of statistical simulations:
  - Reduces unnecessary design margin that could result from designing to FAB-supplied corner models that often are non-physical.
- Redundant vias placed without area increase.
- All custom ckts proven on testchips prior to 1st Si.
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Power

Niagara2 Worst Case Power = 84 W @ 1.1V, 1.4 GHz

- CMT approach used to optimize the design for performance/watt.
- Clock gating used at cluster and local clock-header level.
- 'GATE-BIAS' cells used to reduce leakage.
  > ~10 % increase in channel length gives ~40 % leakage reduction.
- Interconnect W/S combinations optimized for power-delay product to reduce interconnect power.
Power management

- Software can turn threads on/off.
- 'Power Throttling' mode controls instruction issue rates to manage power consumption.
- On-chip thermal diodes monitor die temperature.
  > Helps ensure reliable operation in case of cooling system failure.
- Memory Controllers enable DRAM power-down modes and/or control DRAM access rates to control memory power.

Effect of Throttling on Dynamic Power

<table>
<thead>
<tr>
<th>Degree of Throttling</th>
<th>None</th>
<th>Minimum</th>
<th>Medium</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of 'No Throttling' power</td>
<td>100</td>
<td>95</td>
<td>90</td>
<td>80</td>
</tr>
</tbody>
</table>

- High Workload
- Low Workload
Outline

• Why CMT Architecture ?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > Clocking Scheme
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Design for Testability

- Deterministic Test Mode (DTM) used to test core by eliminating uncertainty of asynchronous domain crossings.
- Dedicated 'Debug Port' observes on-chip signals.
- 32 scan chains cover >99 % flops; enable ATPG/Scan testing.
- All RAM/CAM arrays testable using MBIST and Macrotest.
  - Direct Memory Observe (DMO) using Macrotest enables fast bit-mapping required for array repair.
- Path Delay/Transition Test technique enables speed testing of targeted critical paths.
- SerDes designs incorporate loopback capabilities for testing.
- Architecture design enables use of <8 SPCs/L2 banks.
  - Shortened debug cycle by making partially functional die usable.
  - Will increase overall yield by enabling partial-core products.
Mission Mode vs DTM

Mission Mode Operation

Deterministic Test Mode Operation
Outline

• Why CMT Architecture?
• Key Features and Architecture Overview
• Physical Implementation
  > Key Statistics
  > On-chip L2 Caches
  > Crossbar
  > On-chip L2 Caches
  > SerDes interfaces
  > Cryptography Support
  > Physical Design Methodology
• Power and Power Management
• DFT Features
• Conclusions
Conclusions

• Sun's 2\textsuperscript{nd} generation 8-core, 64-thread, CMT SPARC processor optimized for Space, Power, and Performance (SWaP) integrates all major system functions on chip.

• Doubles the throughput and throughput/watt compared to UltraSparcT1.

• Provides an order of magnitude improvement in floating point throughput compared to UltraSparcT1.

• Enables secure applications with advanced cryptographic support at wire speed.

• Enables new generation of power-efficient, fully-secure datacenters.
Acknowledgements

• Jim Ballard, Mahmudul Hassan, Tim Johnson, Rob Mains, Paresh Patel, Alan Smith for helping put together the presentation.
• Niagara2 design team and other teams inside SUN for the development of Niagara2.
• Texas Instruments for manufacturing Niagara2.
Thank You!

Umesh Nawathe
Senior Manager,
Sun Microsystems