An 8-core, 64-thread, 64-bit, power efficient SPARC SoC (Niagara2)

Umesh Gajanan Nawathe, Mahmudul Hassan, Lynn Warriner, King Yen, Bharat Upputuri, David Greenhill, Ashok Kumar, Heechoul Park

Sun Microsystems Inc., Sunnyvale, CA
Outline

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

- 2\textsuperscript{nd} 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.
• 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.
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.
Physical Implementation Highlights

• Flat cluster composition allows better design optimization; custom clock insertion/routing to meet tight clock skew budgets.

• Static cell-based methodology for most design.

• Selective use of Low-VT gates to speed up critical paths.

• Extensive use of DFM:
  > Larger-than-minimum design rules.
  > Shielding gates using dummy polys.
  > OPC simulations of critical layouts.
  > Extensive use of statistical simulations.
  > All custom designs proven on testchips prior to 1st Si.

<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>
Level2 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 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.
Clocking

<table>
<thead>
<tr>
<th></th>
<th>REF</th>
<th>CMP</th>
<th>IO</th>
<th>IO2X</th>
<th>FSR.refclk</th>
<th>FSR.bitclk</th>
<th>FSR.byteclk</th>
<th>DR</th>
<th>PSR.refclk</th>
<th>PSR.bitclk</th>
<th>PSR.byteclk</th>
<th>PCI-Ex</th>
<th>ESR.refclk</th>
<th>ESR.bitclk</th>
<th>ESR.byteclk</th>
<th>MAC.1</th>
<th>MAC.2</th>
<th>MAC.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>133/167/200 MHz</td>
<td>1.4 GHz</td>
<td>350 MHz</td>
<td>700 MHz</td>
<td>133/167/200 MHz</td>
<td>1.6/2.0/2.4 GHz</td>
<td>267/333/400 MHz</td>
<td>267/333/400 MHz</td>
<td>100/125/250 MHz</td>
<td>1.25 GHz</td>
<td>250 MHz</td>
<td>250 MHz</td>
<td>156 MHz</td>
<td>1.56 GHz</td>
<td>312.5 MHz</td>
<td>312.5 MHz</td>
<td>156 MHz</td>
<td>125/25/2.5 MHz</td>
</tr>
</tbody>
</table>

- **REF**: Reference clock
- **CMP**: Central processing unit
- **IO**: Input/output
- **IO2X**: Input/output with twice the frequency
- **FSR**: Front-side reference
- **PSR**: Processor system reference
- **PCI-Ex**: Peripheral component interconnect express
- **ESR**: External system reference
- **MAC.1**, **MAC.2**, **MAC.3**: Memory access controller
- **NCU**, **DMU**, **PEU**, **PSR**, **SIU**, **NIU**, **RDP**, **RTX**, **TDS**: Various units for clock management

**Clocking frequencies**:
- **Mesochronous**: clock frequencies between 1.4 GHz and 350 MHz
- **Ratioed synchronous**: clock frequencies between 350 MHz and 700 MHz
- **Asynchronous**: clock frequencies above 700 MHz
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.
RSC domain crossings: Sync_en generation

- Example shows: \( \frac{F_{\text{FCLK}}}{F_{\text{SCLK}}} = 13/4 = 3.25 \)
- 'Sync_En' pulse identifies FCLK cycle for data transfers in both directions, i.e.
  > FCLK -> SCLK, and
  > SCLK -> FCLK.
- Desired FCLK cycle is the one whose rising edge is closest to the center of the SCLK cycle (yellow vertical lines in timing diagram).
RSC domain crossings

- Same 'Sync_en' signal used for FCLK -> SCLK and SCLK -> FCLK domain crossings.
- This methodology greatly reduces clock balancing requirements on all RSCs.
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>Signalling Reference</th>
<th>FBDIMM</th>
<th>PCI-Express</th>
<th>Ethernet-XAUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link-rate (Gb/s)</td>
<td>4.8</td>
<td>2.5</td>
<td>3.125</td>
</tr>
<tr>
<td># of North-bound (Rx) lanes</td>
<td>14 * 8</td>
<td>8</td>
<td>4 * 2</td>
</tr>
<tr>
<td># of South-bound (Tx) lanes</td>
<td>10 * 8</td>
<td>8</td>
<td>4 * 2</td>
</tr>
<tr>
<td>Bandwidth (Gb/s)</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.
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

% of 'No Throttling' power

Degree of Throttling

- None
- Minimum
- Medium
- Maximum

Effect of Throttling on Dynamic Power

- High Workload
- Low Workload

Table:

<table>
<thead>
<tr>
<th>Workload</th>
<th>Minimum</th>
<th>Medium</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>70%</td>
<td>75%</td>
<td>80%</td>
</tr>
<tr>
<td>None</td>
<td>75%</td>
<td>80%</td>
<td>85%</td>
</tr>
<tr>
<td>None</td>
<td>80%</td>
<td>85%</td>
<td>90%</td>
</tr>
<tr>
<td>None</td>
<td>85%</td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>None</td>
<td>90%</td>
<td>95%</td>
<td>100%</td>
</tr>
</tbody>
</table>
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
## F vs Vdd Shmoo

- **1st** Si very clean – booted Solaris in 5 days.
- Several parts from **1st** Si running in lab systems at **1.4 GHz**.

### Graph Details

1. **4 Ghz @ 1.1V, 95C**

### Table

<table>
<thead>
<tr>
<th>Frequency (MHz)</th>
<th>Delay (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>4.000</td>
</tr>
<tr>
<td>1939</td>
<td>4.125</td>
</tr>
<tr>
<td>1882</td>
<td>4.250</td>
</tr>
<tr>
<td>1829</td>
<td>4.375</td>
</tr>
<tr>
<td>1778</td>
<td>4.500</td>
</tr>
<tr>
<td>1730</td>
<td>4.625</td>
</tr>
<tr>
<td>1684</td>
<td>4.750</td>
</tr>
<tr>
<td>1641</td>
<td>4.875</td>
</tr>
<tr>
<td>1600</td>
<td>5.000</td>
</tr>
<tr>
<td>1561</td>
<td>5.125</td>
</tr>
<tr>
<td>1524</td>
<td>5.250</td>
</tr>
<tr>
<td>1488</td>
<td>5.375</td>
</tr>
<tr>
<td>1455</td>
<td>5.500</td>
</tr>
<tr>
<td>1422</td>
<td>5.625</td>
</tr>
<tr>
<td>1391</td>
<td>5.750</td>
</tr>
<tr>
<td>1362</td>
<td>5.875</td>
</tr>
<tr>
<td>1333</td>
<td>6.000</td>
</tr>
<tr>
<td>1306</td>
<td>6.125</td>
</tr>
<tr>
<td>1280</td>
<td>6.250</td>
</tr>
<tr>
<td>1255</td>
<td>6.375</td>
</tr>
<tr>
<td>1231</td>
<td>6.500</td>
</tr>
<tr>
<td>1208</td>
<td>6.625</td>
</tr>
<tr>
<td>1185</td>
<td>6.750</td>
</tr>
<tr>
<td>1164</td>
<td>6.875</td>
</tr>
<tr>
<td>1143</td>
<td>7.000</td>
</tr>
<tr>
<td>1123</td>
<td>7.125</td>
</tr>
<tr>
<td>1103</td>
<td>7.250</td>
</tr>
<tr>
<td>1085</td>
<td>7.375</td>
</tr>
<tr>
<td>1067</td>
<td>7.500</td>
</tr>
<tr>
<td>1049</td>
<td>7.625</td>
</tr>
<tr>
<td>1032</td>
<td>7.750</td>
</tr>
<tr>
<td>1016</td>
<td>7.875</td>
</tr>
<tr>
<td>1000</td>
<td>8.000</td>
</tr>
<tr>
<td>985</td>
<td>8.125</td>
</tr>
<tr>
<td>970</td>
<td>8.250</td>
</tr>
<tr>
<td>955</td>
<td>8.375</td>
</tr>
<tr>
<td>941</td>
<td>8.500</td>
</tr>
<tr>
<td>920</td>
<td>8.625</td>
</tr>
<tr>
<td>914</td>
<td>8.750</td>
</tr>
<tr>
<td>901</td>
<td>8.875</td>
</tr>
<tr>
<td>889</td>
<td>9.000</td>
</tr>
<tr>
<td>877</td>
<td>9.125</td>
</tr>
<tr>
<td>865</td>
<td>9.250</td>
</tr>
<tr>
<td>853</td>
<td>9.375</td>
</tr>
<tr>
<td>842</td>
<td>9.500</td>
</tr>
<tr>
<td>831</td>
<td>9.625</td>
</tr>
<tr>
<td>821</td>
<td>9.750</td>
</tr>
<tr>
<td>810</td>
<td>9.875</td>
</tr>
<tr>
<td>800</td>
<td>10.000</td>
</tr>
</tbody>
</table>
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

- Niagara2 design team and other teams inside SUN for the development of Niagara2.
- Texas Instruments for co-developing SerDes and manufacturing Niagara2.