

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

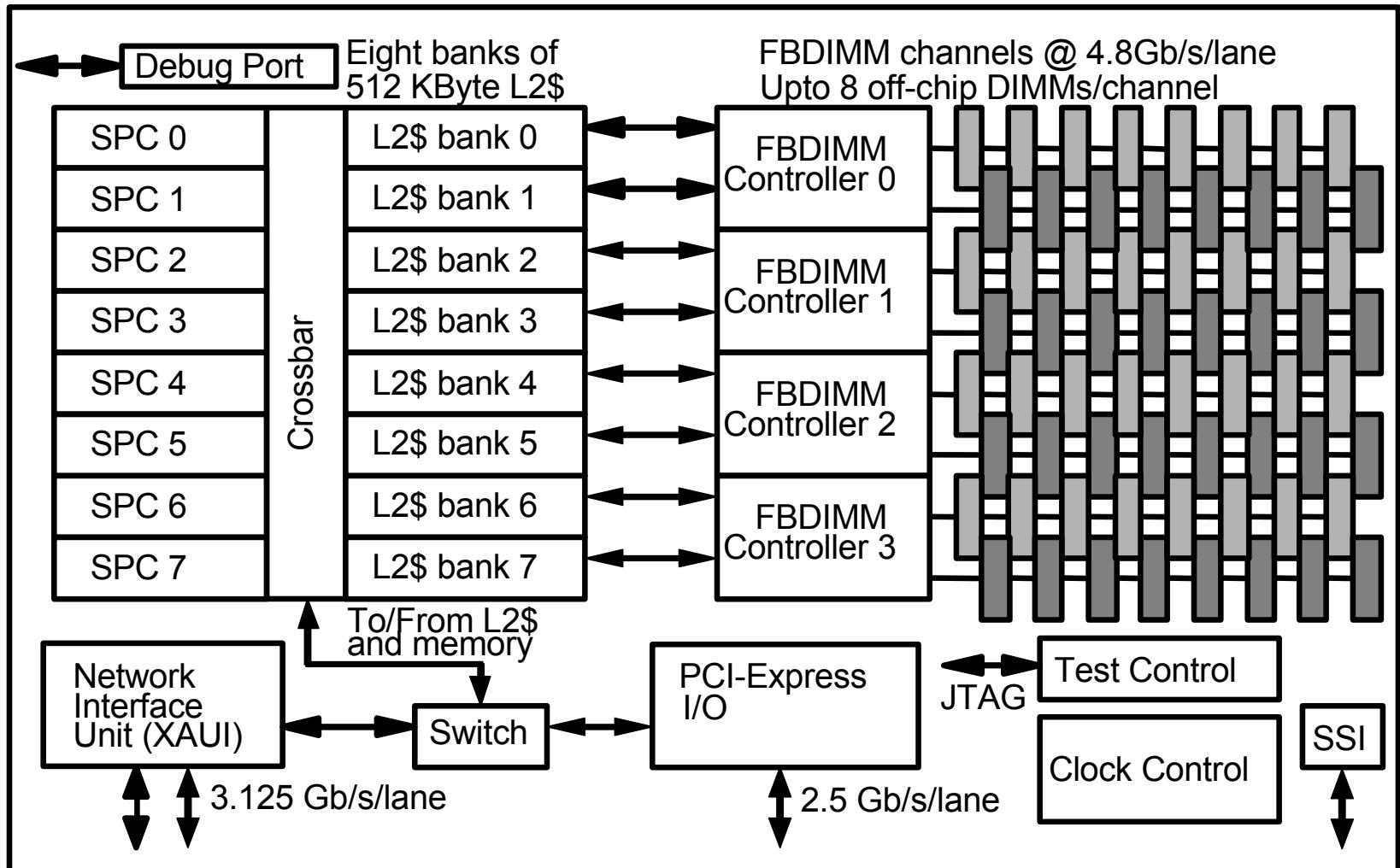
Outline

- 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

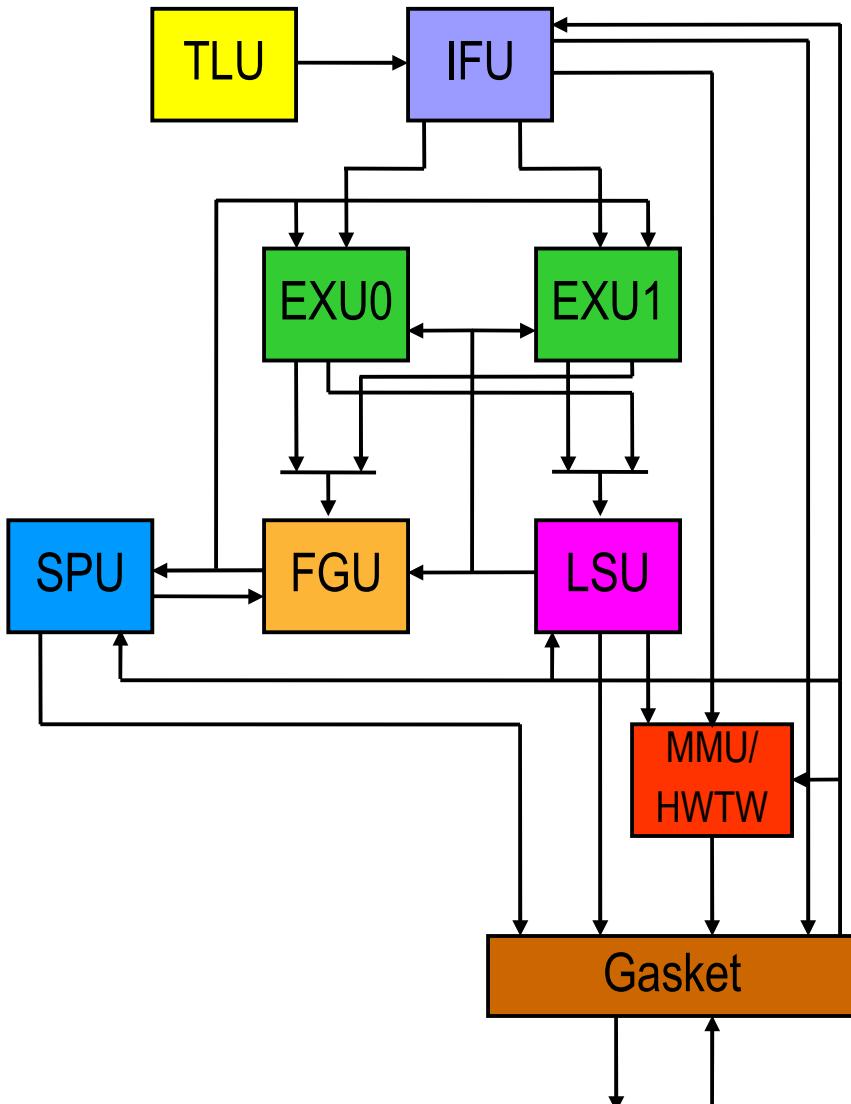
- 2nd generation CMT (Chip Multi-Threading) processor optimized for Space, Power, and Performance (SWaP).
- 8 Sparc Cores, 4MB shared L2 cache; Supports concurrent execution of 64 threads.
- >2x UltraSparc T1's throughput performance and performance/Watt.
- >10x improvement in Floating Point throughput 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.

Niagara2 Block Diagram



- System-on-a-Chip, CMT architecture => lower # of system components, reduced complexity, power => higher system reliability.

Sparc Core (SPC) Architecture Features



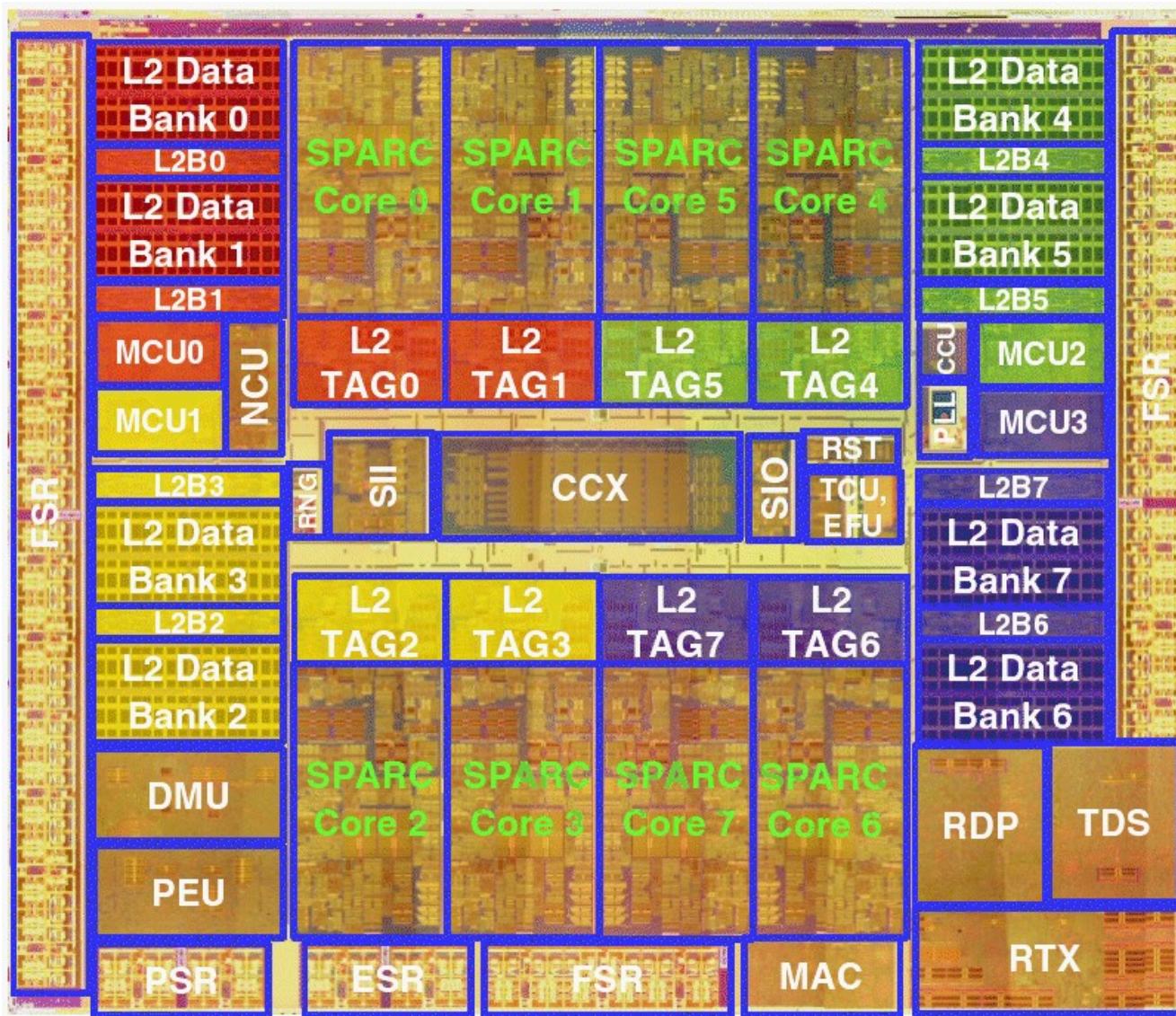
SPC Block Diagram

- 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.
- Up to 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

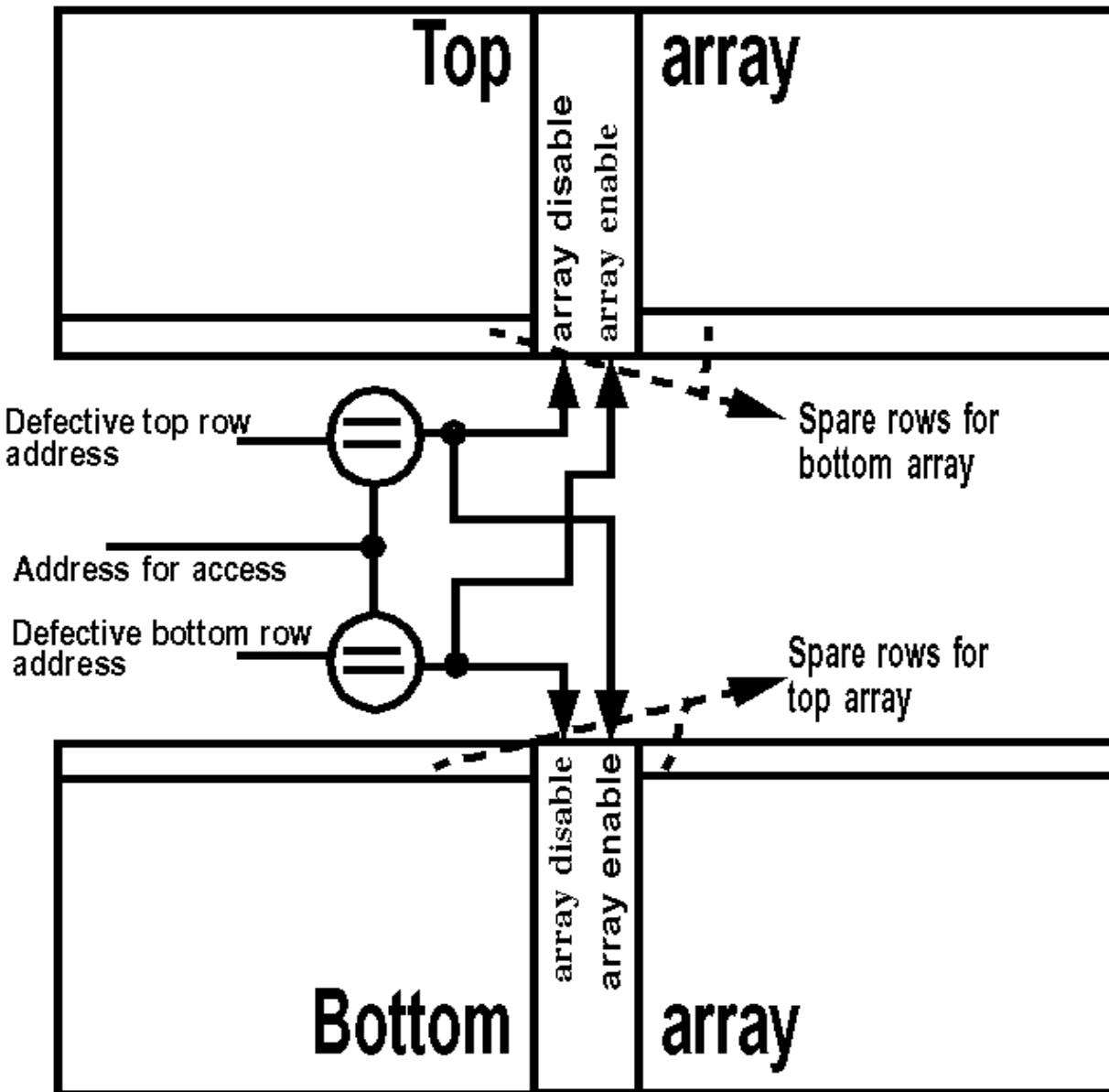
Technology	65 nm CMOS (from Texas Instruments)
Nominal Voltages	1.1 V (Core), 1.5V (Analog)
# of Metal Layers	11
Transistor types	3 (SVT, HVT, LVT)
Frequency	1.4 Ghz @ 1.1V
Power	84 W @ 1.1V
Die Size	342 mm ²
Transistor Count	503 Million
Package	Flip-Chip Glass Ceramic
# of pins	1831 total; 711 Signal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

- 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.

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 N WELL 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



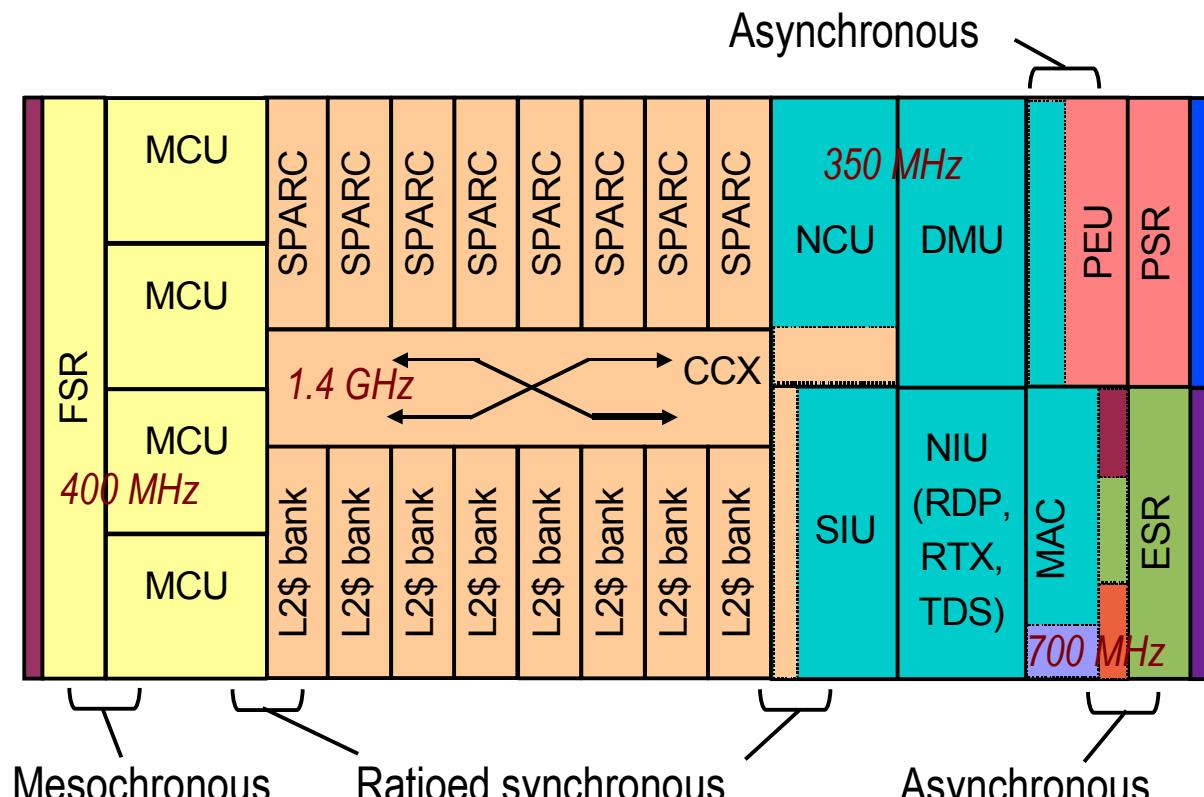
- Redundancy implemented at 32-KB level.
- Spare rows for one array located in adjacent array.
- Adjacent array (which is normally not enabled) is enabled if 'incoming address' = 'defective row address'.
- Reduces X-decoder area by ~30 %.

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.

Clocking

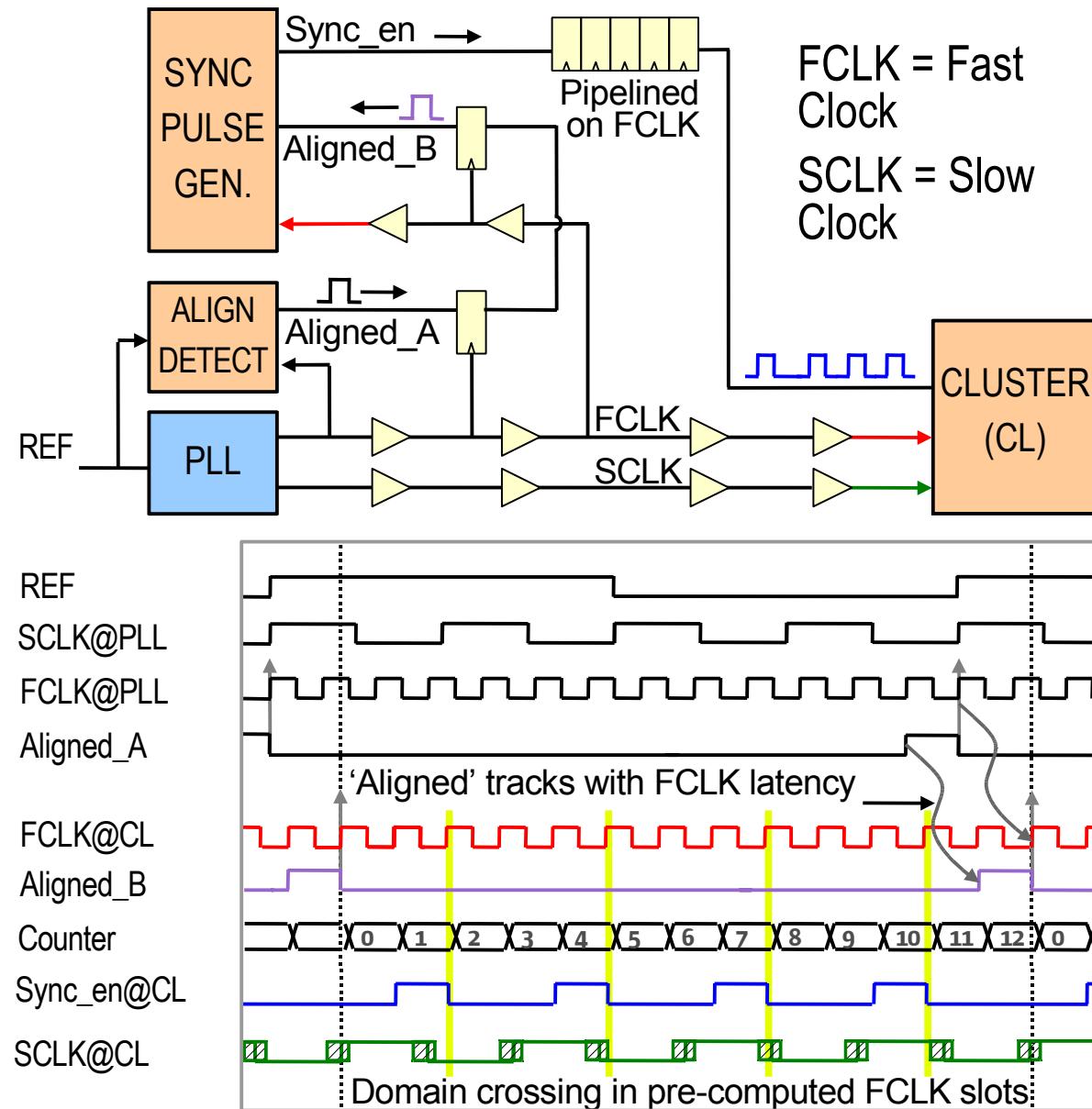


REF	133/167/200 MHz
CMP	1.4 GHz
IO	350 MHz
IO2X	700 MHz
FSR.refclk	133/167/200 MHz
FSR.bitclk	1.6/2.0/2.4 GHz
FSR.byteclk	267/333/400 MHz
DR	267/333/400 MHz
PSR.refclk	100/125/250 MHz
PSR.bitclk	1.25 GHz
PSR.byteclk	250 MHz
PCI-Ex	250 MHz
ESR.refclk	156 MHz
ESR.bitclk	1.56 GHz
ESR.byteclk	312.5 MHz
MAC.1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

Clocking (Cont'd.)

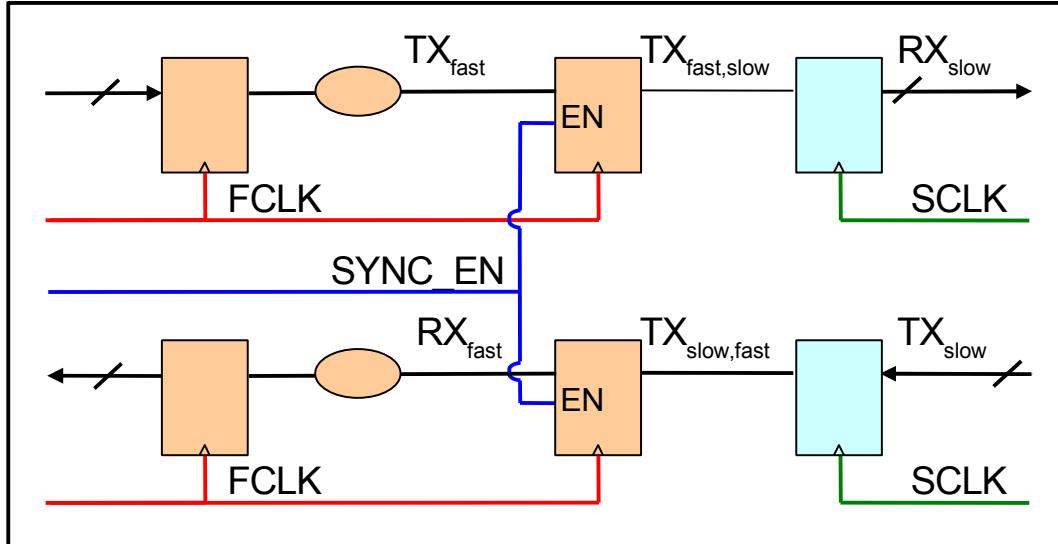
- On-chip PLL generates Ratioed Synchronous Clocks (RSCs); Supported fractional divide ratios: 2 to 5.25 in 0.25 increments.
- Balanced use of H-Trees and Grids for RSCs to reduce power and meet clock-skew budgets.
- Periodic relationship of RSCs exploited to perform high BW skew-tolerant domain crossings.
- 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

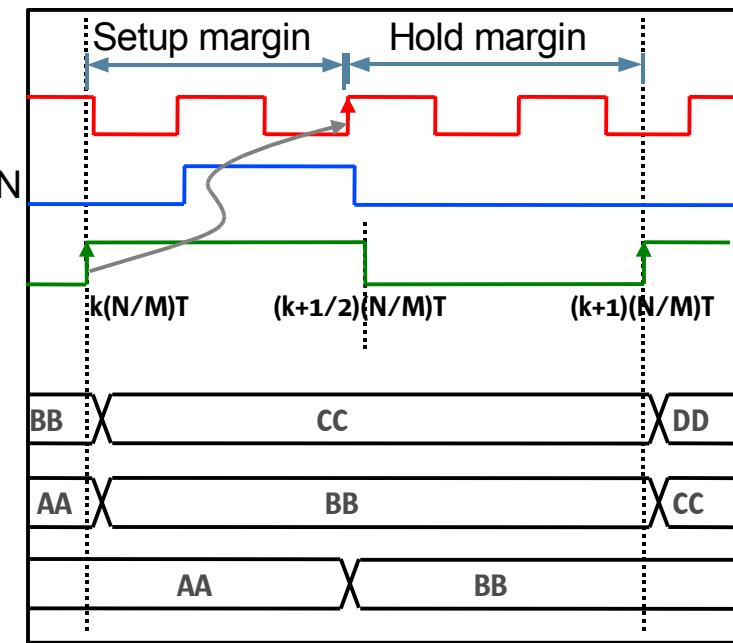
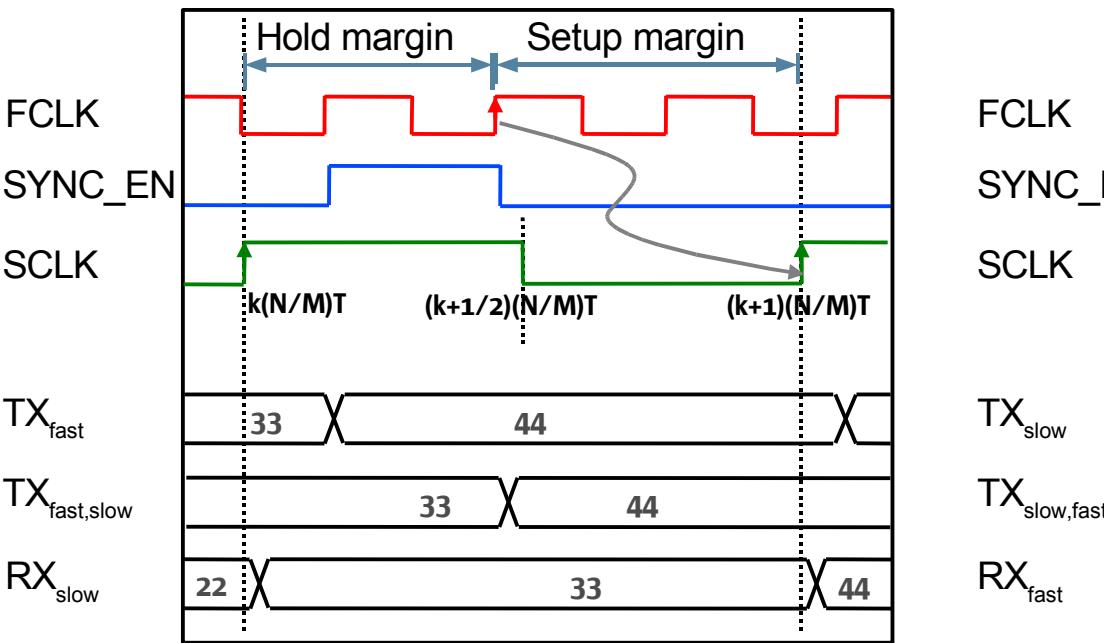


- Example shows: $\frac{F_{FCLK}}{F_{SCLK}} = 13/4 = 3.25$
- 'Sync_En' pulse identifies FCLK cycle for data transfers in both directions, i.e.
 - > FCLK \rightarrow SCLK, and
 - > SCLK \rightarrow FCLK.
- Desired FCLK cycle is the one whose rising edge is closest to the center of the SCLK cycle (yellow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

RSC domain crossings



- Same 'Sync_en' signal used for FCLK \rightarrow SCLK and SCLK \rightarrow FCLK domain crossings.
- This methodology greatly reduces clock balancing requirements on all RSCs.

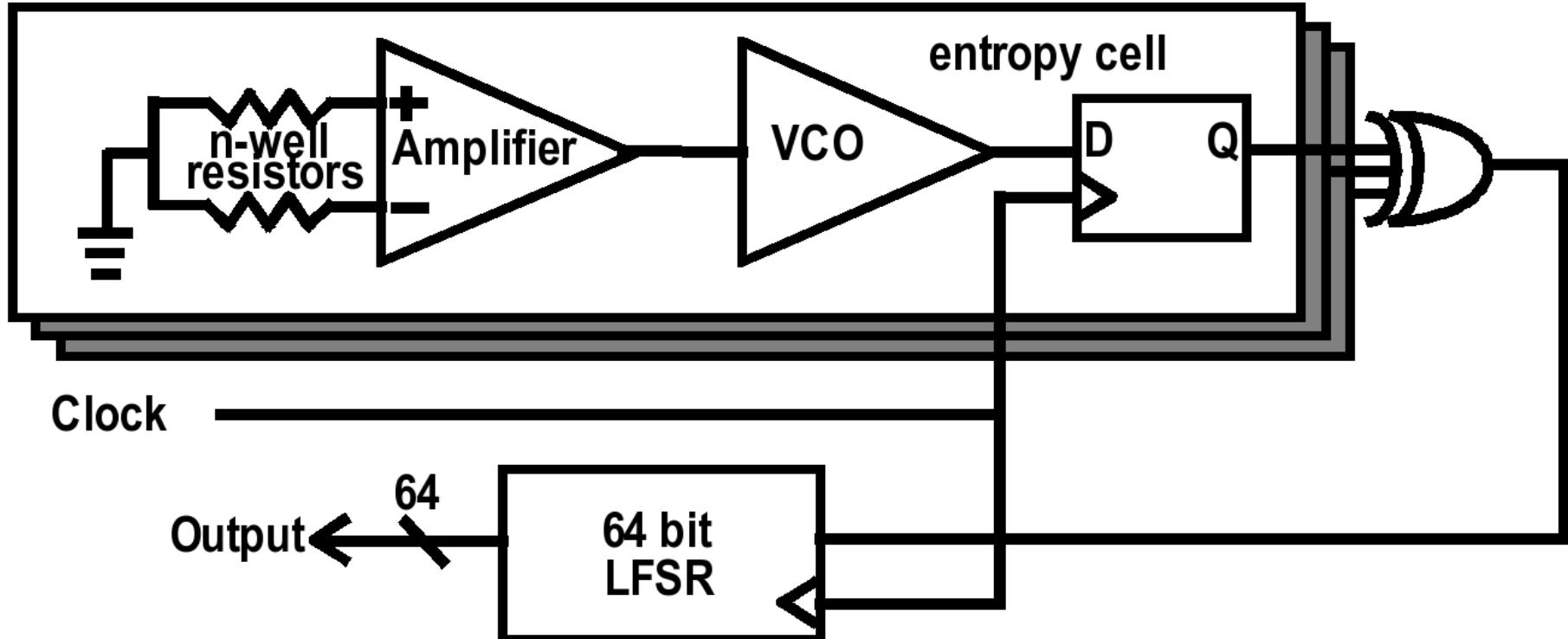


Niagara2's SerDes Interfaces

	FBDIMM	PCI-Express	Ethernet-XAUI
Signalling Reference	VSS	VDD	VDD
Link-rate (Gb/s)	4.8	2.5	3.125
# of North-bound (Rx) lanes	14 * 8	8	4 * 2
# of South-bound (Tx) lanes	10 * 8	8	4 * 2
Bandwidth (Gb/s)	921.6	40	50

- 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.

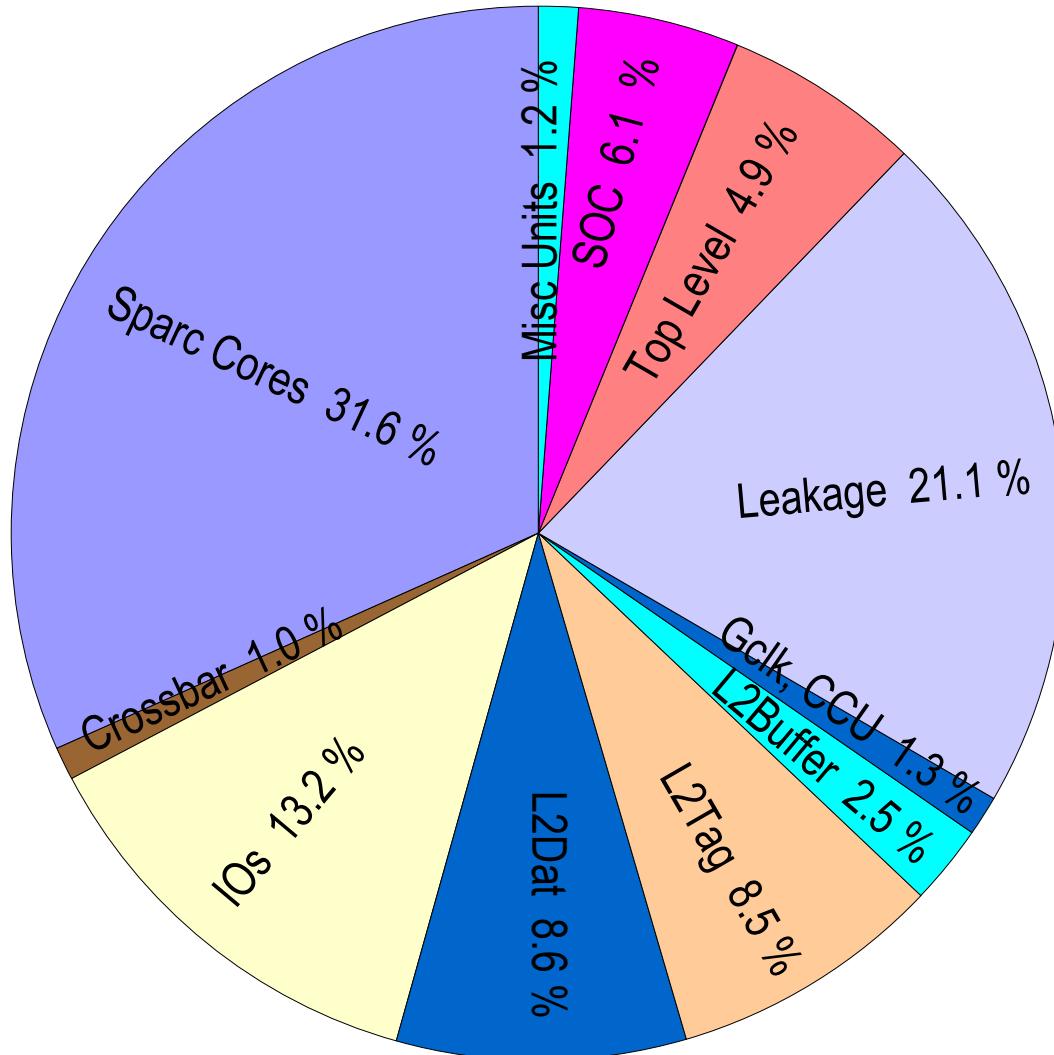
Niagara2's True Random Number Generator



- 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

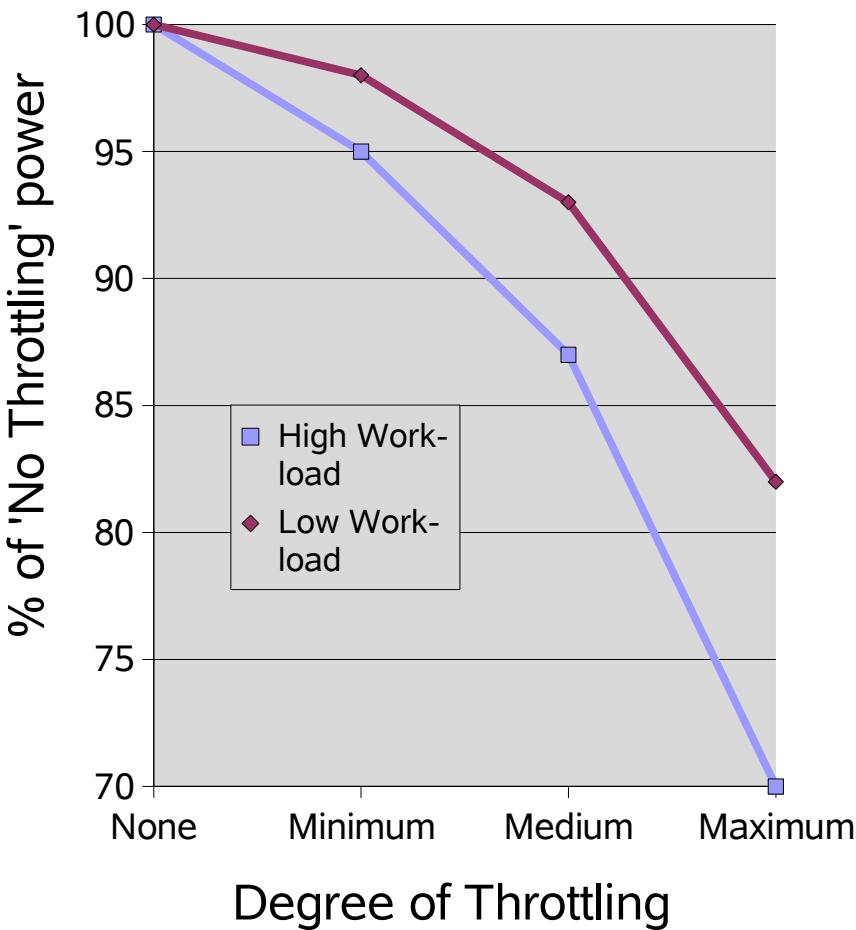
Niagara2 Worst Case Power =
84 W @ 1.1V, 1.4 GHz



- 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

Effect of Throttling on Dynamic Power



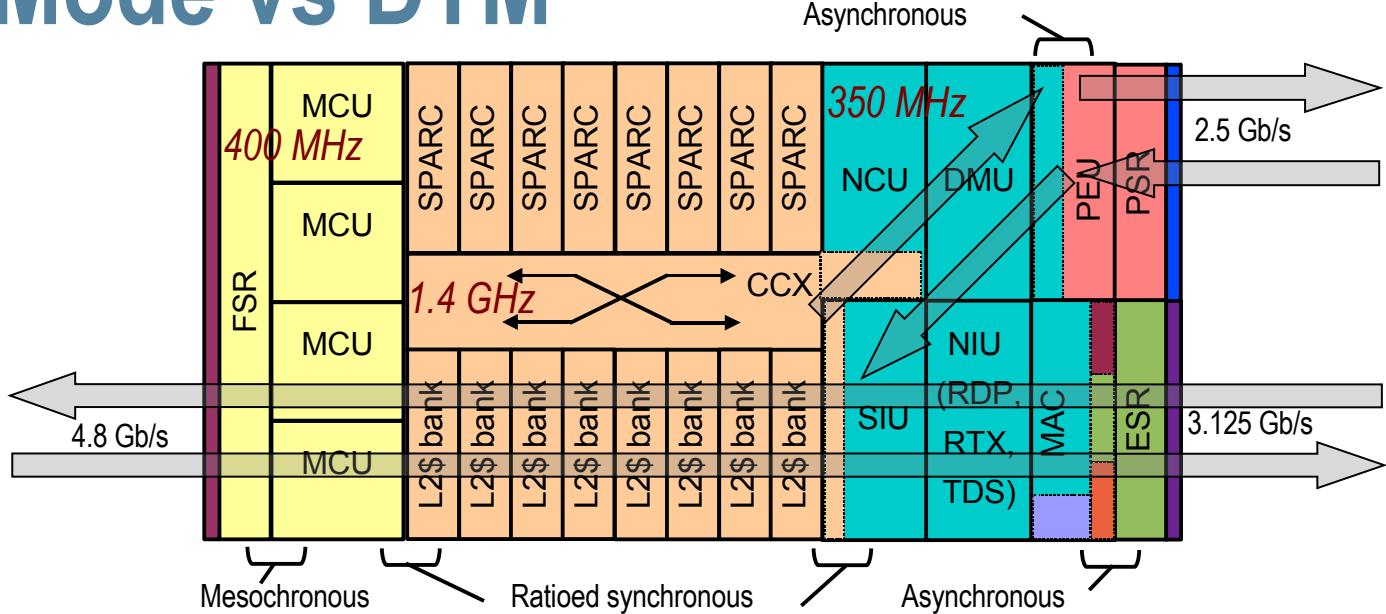
- Software can turn threads on/off.
- 'Power Throttling' mode controls instruction issue rates to manage power consumption.
- On-chip thermal diodes monitor die temperature.
 - Helps ensure reliable operation in case of cooling system failure.
- Memory Controllers enable DRAM power-down modes and/or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

Design for Testability

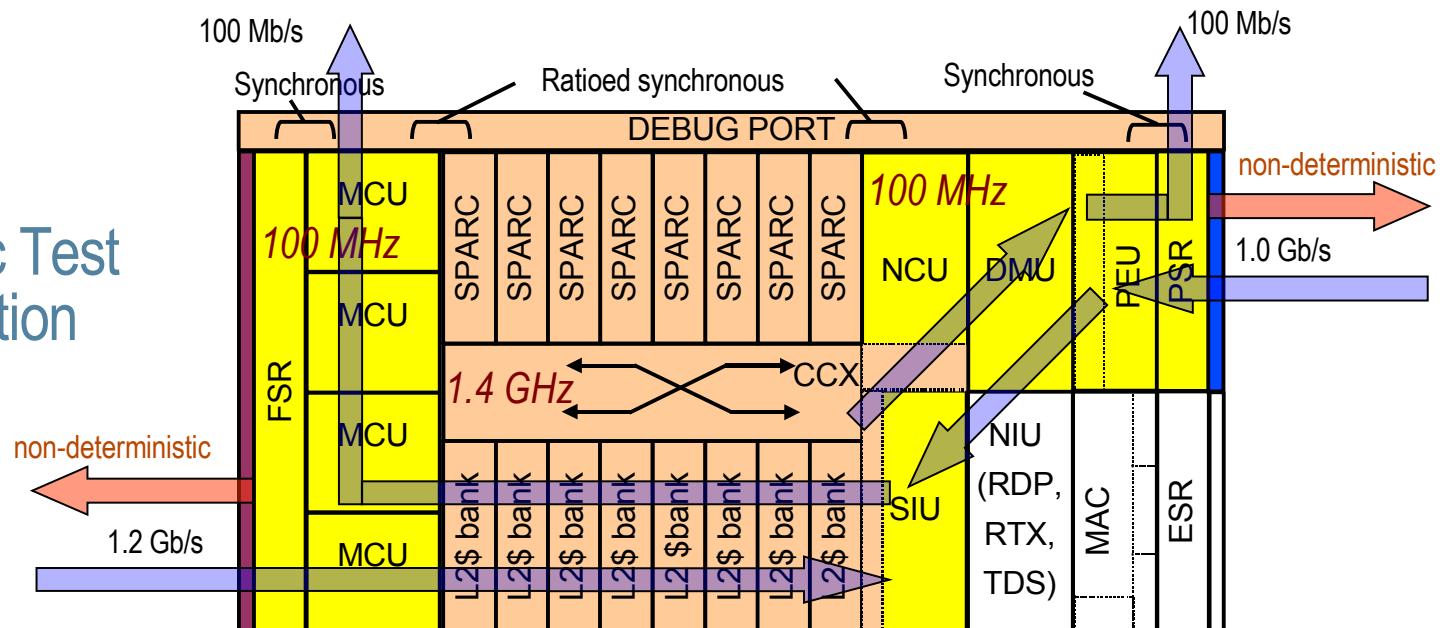
- Deterministic Test Mode (DTM) used to test core by eliminating uncertainty of asynchronous domain crossings.
- Dedicated 'Debug Port' observes on-chip signals.
- 32 scan chains cover >99 % flops; enable ATPG/Scan testing.
- All RAM/CAM arrays testable using MBIST and Macrotest.
 - Direct Memory Observe (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.

1.4 Ghz @
1.1V, 95C

2000MHz	4.000ns	
1939MHz	4.125ns	
1882MHz	4.250ns	
1829MHz	4.375ns	
1778MHz	4.500ns	
1730MHz	4.625ns	
1684MHz	4.750ns	
1641MHz	4.875ns	
1600MHz	5.000ns	
1561MHz	5.125ns	*****
1524MHz	5.250ns	*****
1488MHz	5.375ns	*****
1455MHz	5.500ns	*****
1422MHz	5.625ns	*****
1391MHz	5.750ns	*****
1362MHz	5.875ns	*****
1333MHz	6.000ns	*****
1306MHz	6.125ns	*****
1280MHz	6.250ns	*****
1255MHz	6.375ns	*****
1231MHz	6.500ns	*****
1208MHz	6.625ns	*****
1185MHz	6.750ns	*****
1164MHz	6.875ns	*****
1143MHz	7.000ns	*****
1123MHz	7.125ns	*****
1103MHz	7.250ns	*****
1085MHz	7.375ns	*****
1067MHz	7.500ns	*****
1049MHz	7.625ns	*****
1032MHz	7.750ns	*****
1016MHz	7.875ns	*****
1000MHz	8.000ns	*****
985MHz	8.125ns	*****
970MHz	8.250ns	*****
955MHz	8.375ns	*****
941MHz	8.500ns	*****
928MHz	8.625ns	*****
914MHz	8.750ns	*****
901MHz	8.875ns	*****
889MHz	9.000ns	*****
877MHz	9.125ns	*****
865MHz	9.250ns	*****
853MHz	9.375ns	*****
842MHz	9.500ns	*****
831MHz	9.625ns	*****
821MHz	9.750ns	*****
810MHz	9.875ns	*****
800MHz	10.000ns	*****

Conclusions

- Sun's 2nd 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.

