

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/224220004>

A 40nm 16-core 128-thread SPARC® SoC processor

Conference Paper in IEEE Journal of Solid-State Circuits · December 2010

DOI: 10.1109/ASSCC.2010.5716541 · Source: IEEE Xplore

CITATIONS

29

READS

1,164

6 authors, including:



Jinuk Luke Shin

24 PUBLICATIONS 611 CITATIONS

[SEE PROFILE](#)



Changku Hwang

The Ohio State University

34 PUBLICATIONS 596 CITATIONS

[SEE PROFILE](#)

All content following this page was uploaded by Jinuk Luke Shin on 07 June 2015.

The user has requested enhancement of the downloaded file.

A 40 nm 16-Core 128-Thread SPARC SoC Processor

Jinuk Luke Shin, Dawei Huang, *Member, IEEE*, Bruce Petrick, Changku Hwang, *Senior Member, IEEE*, Kenway W. Tam, *Member, IEEE*, Alan Smith, Ha Pham, Hongping Li, Timothy Johnson, Francis Schumacher, Ana Sonia Leon, *Senior Member, IEEE*, and Allan Strong

Abstract—This fourth generation UltraSPARC T3 SoC processor implements sixteen 8-threaded SPARC cores to double on-chip thread count and throughput performance over its previous generation. It enhances glueless scalability to enable up to 512 threads in a 4-way system. A 16-Bank 6 MB L2 Cache, a 512 GB/s hierarchical crossbar and a 312-lane SerDes I/O of 2.4 Tb/s support the bandwidth required by the large number of threads. This SoC processor integrates the memory controller, PCIE 2.0, 10 Gb Ethernet ports, and required cache coherency support in multi-chip configurations. Multiple clock and power domains are used to optimize performance and power for the SoC components. Extensive power management features, from architecture to circuit techniques, optimize both active and idle power. The 377 mm² die includes 1 billion transistors in a flip-chip ceramic package with 2117 pins. The chip is fabricated in TSMC's 40 nm high-performance process with 11 Cu metals and four transistor types.

Index Terms—Asynchronous crossing, body-bias, clocking, coherency, electromigration (EM), idle power, interconnect crossbar, L2 Cache, multi-core, multi-threaded, multiple power domain, power grid, power management, reliability, SerDes, System-On-Chip (SoC), throughput performance, UltraSPARC T3.

I. INTRODUCTION

COMMERCIAL workloads in datacenters, which traditionally exhibit large amounts of thread level parallelism (TLP) and high cache miss rates, continue to face increasing throughput, power and area challenges [1]. Since its introduction, the T-Series Family has effectively addressed these challenges by improving aggregate throughput performance per watt. Fig. 1 shows the evolution in number of threads for Oracle's T-series processors, doubling the throughput performance within the same power envelope. The 90 nm UltraSPARC T1 is the industry's first generation general purpose multi-threaded processor with 32 on-chip threads [2], [3]. The second generation T2, implemented in 65 nm, doubles on-chip thread count over its predecessor by introducing dual pipelines in the cores [4]. The T2Plus processor delivers the first glueless scalability in the T-series line providing 128 threads in a 2-way system configuration without the need for any additional bridging chips [5]. This next generation UltraSPARC T3 [6],

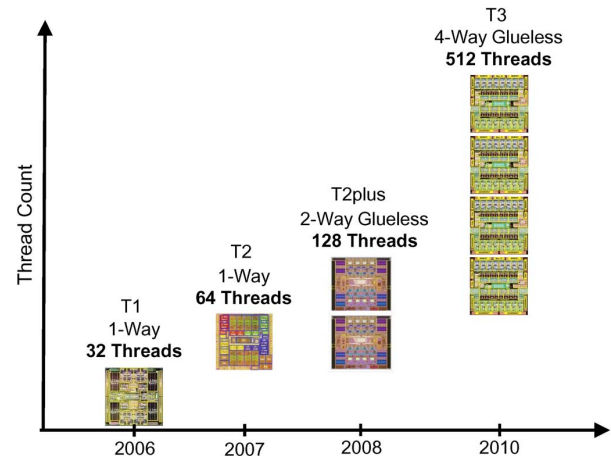


Fig. 1. T-Series doubles thread counts for each generation.

[7] integrates sixteen enhanced cores with 128 hardware threads on a single die. With expanded 4-way glueless scalability, it delivers an unprecedented total of 512 threads in a single 4-way system.

This paper highlights the design implementation of the UltraSPARC T3 processor. Section II gives the architecture overview and key physical statistics. Section III introduces the multiple clock domain features. Section IV describes the power domains and the power grid design methodology. Section V features power management techniques for both active and idle power. Section VI details the physical design methodology which addresses the challenges in advanced technology. Section VII shows the circuit design implementation for the SerDes, L2 cache and crossbar. Section VIII gives the overview of the DFT features. Section IX concludes this paper.

II. OVERVIEW

A. Architecture

The UltraSPARC T3 is the fourth generation scalable chip multi-threaded processor. The processor functional block diagram is shown in Fig. 2. Sixteen enhanced 64-bit SPARC cores (SPC) support the execution of eight fully independent hardware strands to provide 128 threads in a single die, delivering the highest single-chip thread count for a general purpose microprocessor. The core includes two integer execution pipelines, one shared floating-point execution pipeline, and one memory pipeline. The floating point unit is enhanced with a new 9-stage floating point pipeline with a fused multiply-add (FMAC) instruction and visual instruction set (VIS) 3.0 to double the performance for linear-algebra matrix and

Manuscript received April 15, 2010; revised July 18, 2010; accepted August 25, 2010. Date of publication October 25, 2010; date of current version December 27, 2010. This paper was approved by Guest Editor Tanay Karnik.

The authors are with Oracle, Santa Clara, CA 95054 USA (e-mail: jinuk.shin@oracle.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JSSC.2010.2080491

0018-9200/\$26.00 © 2010 IEEE

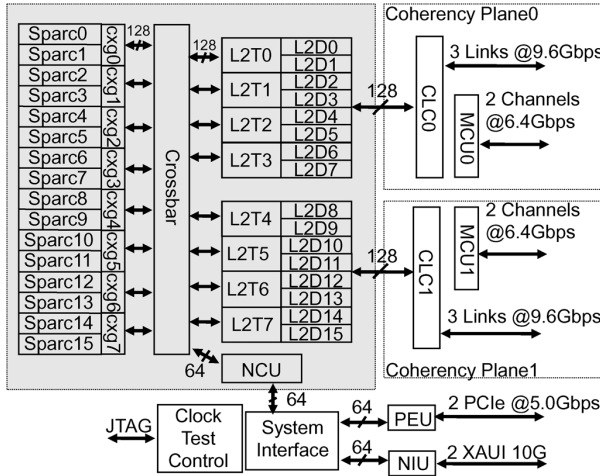


Fig. 2. Processor block diagram.

fast-fourier transforms (FFT) applications over the previous T2/T2Plus processors. The cryptographic stream processing unit includes new support for SHA-384/SHA-512 secure hashes algorithm (SHA), Kasumi Bulk Cipher and Galois field operations. The sixteen cores communicate to the 16-Bank unified 6 MB level 2 cache (L2C) through a high-bandwidth, non-blocking, crossbar switch (CCX). This 4-stage pipelined hierarchical crossbar delivers a bandwidth of 512 GB/s. A new gasket structure (CXG) is also introduced to manage the massive interconnect network between the sixteen cores and the crossbar. The CCX facilitates a synchronized delay control between any of the cores and any of the L2C bank, for partial core product binning and testing. Each L2C bank is 24-way set associative with a 64-byte line size. Its high banking and associativity levels provide adequate bandwidth for the sixteen SPC cores. The L2Cache micro-architecture is enhanced to share a common L2 controller (L2T) for every two L2Data banks. This reduces interconnect congestion to the crossbar and avoids duplication of logic and SRAM components.

The UltraSPARC T3 is a complete system-on-chip processor featuring: coherence control logic which empowers the connection of four T3's in a system without an external hub, two memory controllers to interface to a pair of two memory SerDes channels, two PCI-Express 2.0 controllers, and two 10 Gb Ethernet ports with XAUI interfaces. This high level of integration in the processor reduces the total component count in the system, therefore improving system power, reliability, availability and serviceability [8]. The coherence link cluster (CLC) handles coherent transactions from the corresponding L2C bank pairs, DMA transactions, and the coherence physical link data layer. There are six coherence links managed by two controllers. Each link is 14-bit wide and operates at 9.6 Gb/s in each direction. Eight L2C banks share a memory control unit (MCU). The MCU communicates to DDR3 DIMMs at a peak link rate of 6.4 Gb/s. In addition, the MCU implements various power management features such as dynamic clock enable (CKE) gating and memory-metering to reduce the DRAM power up to 2.7 times. The PCI-Express (PEU) unit supports

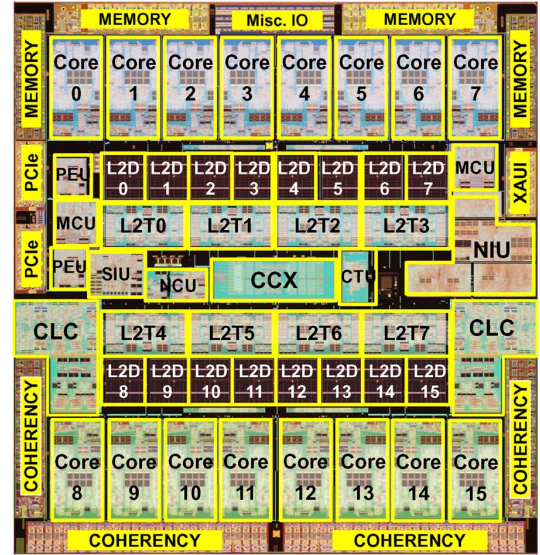


Fig. 3. Chip micrograph.

$\times 1$, $\times 2$, $\times 4$ and $\times 8$ configurations at a data rate of 5.0 Gb/s. The Network Interface Unit (NIU) links a pair of on-chip 10 Gb/s or 1 Gb/s Ethernet MACs to the CPU.

B. Key Physical Features

The processor micrograph (Fig. 3) shows the main functional blocks. The chip takes advantage of a uniform and symmetrical floorplan. The crossbar is placed in the center to evenly distribute data transfers from the eight cores at the top and the eight cores at the bottom to the eight L2 controller units. The non-cacheable-unit (NCU) sits next to the crossbar as it needs to deliver data to the system interface unit (SIU). The regular and symmetric distribution of the high frequency components such as the cores, L2 cache and crossbar produces a uniform temperature and a low gradient of only 14°C under a worst case work load. Each core includes a 16 KB Instruction cache and an 8 KB Data cache with a total area of 6.5 mm^2 . This corresponds to $\sim 25\%$ of the total processor area. The chip is fabricated in TSMC's 40 nm high performance CMOS process with four transistor types including standard Vth, low Vth, thick gate oxide device for the I/Os, and SRAM transistors. 11 layers of Cu metals, including ultra low-K dielectric layers, are used for signal and power routing. The aluminum redistribution layer (ALRDL) is only used to distribute power from the bumps. The 377 mm^2 die contains 1 billion transistors and 2.5 million flip-flops. The chip runs at 1.65–2.0 GHz consuming 120 W, which corresponds to less than 1 Watt per thread. There are 833 signals and 1284 power pins on the flip-chip ceramic LGA.

III. CLOCKING

This processor implements multiple clock domains to support the SoC components, each of them running at intrinsic clock frequencies for optimal performance and power. The main units, such as the SPARC cores, L2 cache and crossbar, operate at their

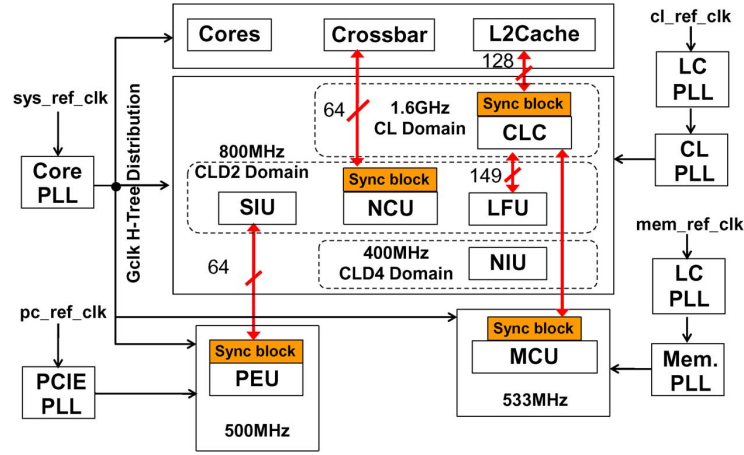


Fig. 4. Clocking overview with 6 major clock domains.

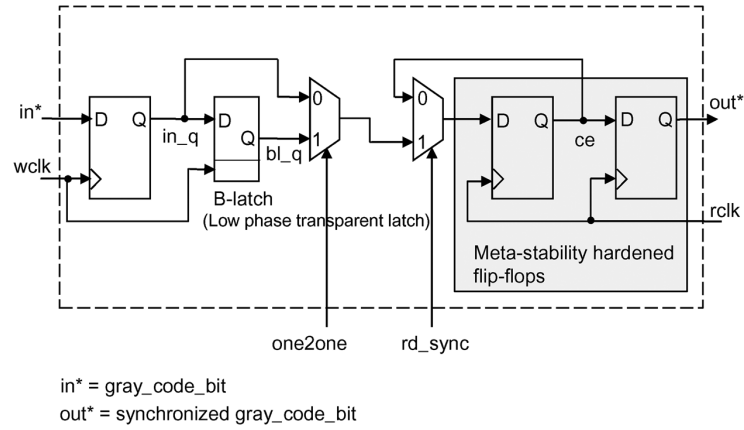


Fig. 5. Synchronization cell circuit diagram.

optimal frequency for performance, independently of the SoC clocks. The T2 processor [9] addressed this difference by using a common clock with synchronous enables to generate divided down ratios. However, this approach constrains the frequency options and increases implementation complexity. Instead, the T3 implements an asynchronous interface across the different clock domains, allowing each SoC cluster to operate at its optimal frequency. A special embedded test mode facilitates synchronous clock domain crossing for operation in the tester and the system.

A. Clock Generation and Distribution

Fig. 4 shows the six main synchronous and asynchronous clock domains. The core PLL generates the clock for the main high-speed clusters. This main clock *gclk* is distributed across the entire chip with a precise H-tree to minimize skew while reducing power. As the CLC, MCU and PEU clusters have a dedicated SerDes I/O, each SerDes generates a suitable clock from the internal PLLs. LC-based PLLs supply a clean reference for the coherency and memory PLLs. These three SoC PLLs clock the coherency, memory, and PEU units at 1.6 GHz, 533 MHz, and 500 MHz respectively. The 1.6 GHz coherence clock is further divided to 800 MHz and 400 MHz to feed the system inter-

face unit (SIU) and the NIU. Within each cluster, the clock control logic (CCL) receives the global clocks and drives the local clock grid for finer skew control. Each SoC cluster receives two input clocks, the *gclk* used in tester mode and the local clock derived from the SerDes used in normal functional mode. Each cluster CCL provides local synchronization pulses to allow synchronous domain crossing during testing mode.

B. Domain Crossing

All control and data communication between clusters is done synchronously at the chip level. This provides a more simplified top level implementation compared to the ratioed synchronous scheme in the previous generation. Asynchronous clock domain crossing is performed within each cluster. A “sync_cell” (Fig. 5) and asynchronous FIFOs with gray code pointers interface between the local cluster clock and a small domain operating at the frequency of the external crossing domain. Two clock grids in each cluster are driven by an individual CCL which controls the asynchronous or synchronous domain crossing mode selection. In the synchronous test mode, pulses from the CCL synchronize the domain crossing at each sync_cell. Each bit of the gray code includes a sync_cell to transfer the gray coded bit from the write clock domain to the read clock domain in both

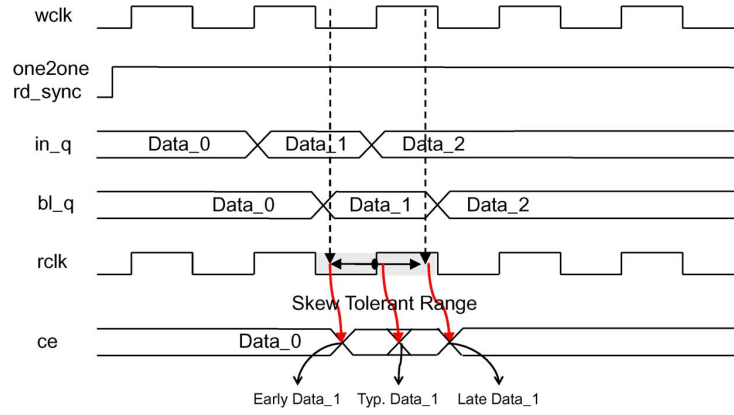


Fig. 6. Synchronous mode crossing timing diagram.

asynchronous and synchronous modes. In asynchronous mode, *rd_sync* is held high and *one2one* is held low. This configures the *sync_cell* into a write flip-flop followed by two flip-flops which function as a synchronizer. These two flip-flops are specially designed with robust meta-stability [10]. In synchronous mode, the signals *rd_sync* and *one2one* are driven by the CCL depending on the ratio between the write and read clocks, to ensure synchronous crossing of the gray code bit. Fig. 6 shows the timing diagram when the write and read clocks are at the same frequency. As the clock grids for the write and read clocks are not identical, there can be a skew between two clock grids. This skew impact is addressed by having a latch following the write flip-flop. Therefore, the two clock grids can be skewed up to half a cycle. The latch is also used for cases when the write clock is faster than the read clock. When the read clock is faster than the write clock, a synchronization signal, *rd_sync*, is pulsed to handle the different frequency ratios. Besides handling domain crossings, the CCL performs several other critical functions such as: handling reset sequences, generating divided down clocks for power management and supporting various DFT features such as clock stop, clock stretch and transition test. Down clocking during reset, enables a gradual ramp up and down of the clocks to minimize di/dt and also facilitates a more relaxed timing for the reset and DFT signals.

IV. POWER DOMAINS AND GRID DESIGN

A. Multiple Power Domains With Body-Bias Capability

The processor takes advantage of multiple power domains to provide performance and power optimization for different types of server applications. Fig. 7 summarizes the multiple supply domains in the chip and the corresponding operating ranges. *Vdd_core* powers most of the cores, L2cache, crossbar, and SoC units. *Vdd_memory* powers the L2Data arrays to support low-power *Vmin* operation while maintaining cell stability. Having a separate *Vdd_memory* and *VNWM* allows low-voltage weak-bit screening during wafer test. A clean 1.0 V *Vdd_analog* supply is filtered on the board with a ferrite bead and capacitors to power the sensitive SerDes analog components and their PLLs. The I/O circuits are driven by a 1.5 V

Power Supply	Function
<i>Vdd_core</i> (VDDC) 0.8V ~ 1.1V	All Logic
<i>Vdd_memory</i> 0.9V ~ 1.1V	L2D Memory Cells
<i>Vdd_analog</i> 1.0V	PLL/SerDes
VDDT/VDDR 1.5V	I/O
VNW (VDDC +/- 0.3V)	PMOS Bias
VNWM	L2D Memory Cells PMOS Bias
VSB (VSS +/- 0.3V)	NMOS Bias

Fig. 7. Power domains and operating ranges.

supply. This 1.5 V supply is also used to generate an internal regulated voltage for the core PLL and the analog components. The design implements body-bias capability for both PMOS (VNW) and NMOS (VSB) to enable electrical modulation of the transistors threshold voltage. Threshold sensitive analog components are placed in a deep nwell for isolation from substrate bias conditions.

Back-bias is used to increase yield by controlling leakage for high leakage parts. Initial silicon results show a potential upside of more than 15% recovered parts after applying back-bias while still meeting speed requirements (Fig. 8(a)). Compared to a per-part-Vdd control, this body-bias technique allows a two-dimensional leakage adjustment by independently controlling PMOS and NMOS depending on process variation. The proper bias level is programmed into an on-chip Efuse array during production test. The system reads the programmed bias levels to set the supplies at power up. As leakage power is the dominant component during burn-in test, this bias technique also allows a larger number of chips in the burn-in board, improving test throughput while maintaining the quality of the reliability test. PMOS back-bias further enables weak-bit screening for the different memory cell types, including multi-port register files and content addressable memories. As PMOS forward-bias can be used to improve frequency, it is also used to compensate

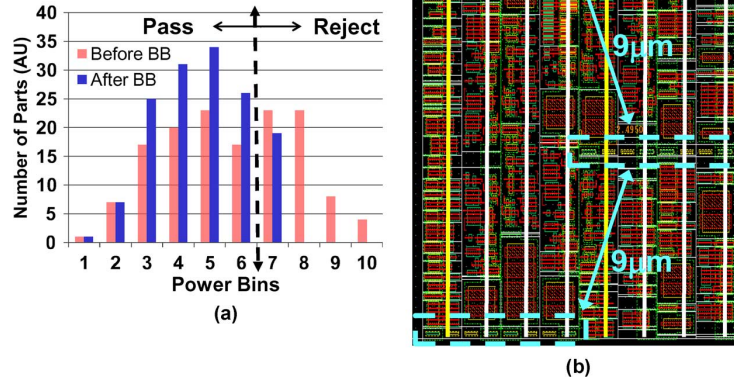


Fig. 8. (a) Part recovery after applying back-bias to reduce leakage. (b) Tap placement and M2 VNW/VSB grid.

for NBTI speed degradation, therefore reducing excessive speed guard-bands during manufacturing.

Body-bias implementation starts with a uniform distribution of taps. Tap cells in control and datapath units are placed as a row group of four every 18 μm . After the taps, the regular library cells are placed. Empty areas between tap rows are then filled with either de-coupling capacitors, which include both PMOS and NMOS transistors, or filler cells. This provides a context friendly environment and provides connectivity to the tap rows. Each library cell reserves one M2 track for VNW and VSB implementation. After tap and cell placement, this track is converted to either VNW or VSB every four cells (Fig. 8(b)). Unused tracks are then used for signal routes. All bias grids are routed between VDD or VSS supply lines for protection from any noise injection with minimal signal route overhead of only 1–2% for each metal layer. Thorough physical and electrical analysis of this bias grid is conducted to ensure the grid integrity. VNW and VSB are supplied from the same external Vdd_core regulator to minimize the component count and limit the bias supply overhead to the package and board routings. In addition, VNW and VSB sense lines are fed back to the external regulator along with VDD and VSS sense lines to maintain the desired bias levels with respect to VDD and VSS.

B. Power Grid Design

While technology scaling continues to increase transistor density, it also brings significant challenge on power grid design due to: higher current density, shrinking metal dimensions, increasing routing congestion and fixed bump pitch. It then becomes critical to design a proper power grid to address EM and IR requirements by construction. In addition, as the complete EMIR verification can only be performed towards the later part of the design cycle, it is necessary to develop an analysis strategy to predict the power grid robustness early in the design phase. This method needs to be as accurate as possible and guarantee a realistic worst case at all times without overdesigning. Several factors such as area utilization, gate types, activity and temporal gate switching are considered to estimate the realistic worst case current for the different metal

layer levels. After a thorough understanding of the currents behavior, it is observed that almost all the current flow from the bump to any distant cells is carried out by only few upper level metals, M11 to M8. On the other hand, the lower level metals conduct current only to a window enclosed by a particular metal (M_N) and its upper level metal (M_{N+1}). Thus, the worst case currents through a low metal layer can be analytically determined by using Kirchhoff's current law (KCL). Once the current consumed by all transistors within the window is estimated, the worst case current for the metal (M_N) can be found based on the number of metal (M_{N-1}) tracks within that window (Fig. 9). This current is used to define all lower level metal power grids while meeting EM requirements. For the upper level metals, simulations are performed to obtain the global currents. This is done based on an approximate model of the power grid between the bumps. With a single simulation, the ratio between the global current in the upper level metals to the corresponding local grid for the lower metals can be obtained. This ratio, which is around 9.2, is then used to estimate the global currents for the rest of the upper level metals without the need of further simulations.

IR drop analysis is conducted using a similar concept. At first, the segment resistance of the metal under evaluation (M_N) is calculated based on the geometry of the lower (M_{N-1}) and upper layer (M_{N+1}) enclosure as shown in Fig. 10. This resistance can be expressed as

$$R_{M_n} = r \frac{L_n}{W_n}$$

where

$$L_n = \frac{(\text{pitch} M_{n+1})}{(K_n + 1)}$$

$$K_n = \text{int} \left[\frac{(\text{pitch} M_{n+1})}{(\text{pitch} M_{n-1})} \right] + 1.$$

The IR drop can then be calculated depending on the potential gradient across the metal segment. For cases with a significant gradient, IR drop through that metal is calculated as

$$IR_{M_n} = R_{M_n} [K_n I_{Mh} - I_{(Mn-1)} K_n (K_n - 1) / 2].$$

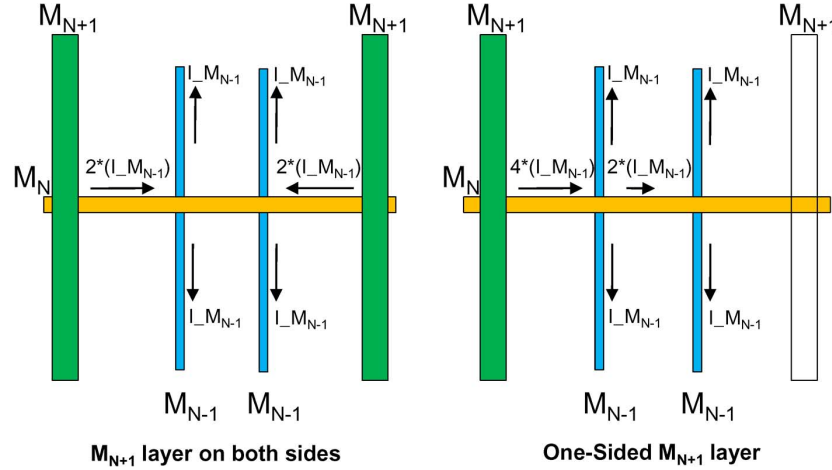


Fig. 9. Local-grid current calculation for analytical EM analysis.

When both sides of the upper level metal lines (M_{n+1}) are at identical potential, the current on layer M_n flows from both ends. Thus, the current at any end is half of the value of the case with a potential gradient. This current will only flow on half of the segment. The IR drop for each metal segment is then calculated as follows:

When K_n is even,

$$IR_{M_n} = R_{M_n} \frac{K_n}{2} \left[I_{M_n} - I_{(M_n-1)} \left(\frac{K_n}{2} - 1 \right) \right]$$

When K_n is odd,

$$IR_{M_n} = R_{M_n} \frac{(K_n + 1)}{2} \left[I_{M_n} - I_{(M_n-1)} \frac{(K_n - 1)}{2} \right].$$

This analytical technique enables an accurate and robust power grid design at the beginning of the project. The grid provides an optimal metal usage that balances the maximum number of signal tracks and a solid power grid to meet EMIR requirements. The EMIR verification performed at tape-out showed no design modifications required to the initial power grid.

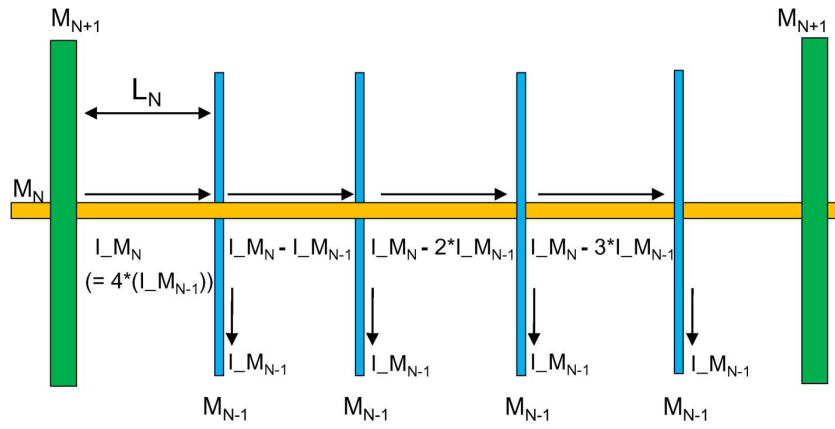
V. POWER MANAGEMENT

Even though the T-series processors utilize an inherent power-efficient architecture, doubling the thread count and integrating more system components increased the T3's power management challenge. Therefore, the design implements extensive power saving features across the full system stack, from the software to the micro-architecture, and down to the transistor level to maintain the same power envelope over its previous generation. In addition to the core, clock gating is now extended to the L2 cache and SoC components. SoC units achieve 60–90% clock gating to save both active and idle power. In an effort to reduce peak power by $\sim 75\%$, the L2 cache lines read and write operations are staged to 16 bytes as linefill instructions to the outside memory can be staged without any performance penalty. Since the T3 chip can support various multi-way system configurations, the coherence link

implements a gating function to enable only the necessary links. As the coherence links account for more than half of the total SerDes lanes and operate at 9.6 Gb/s, this feature provides about 20 W of savings in a 1-way configuration. From the physical design point of view, the design methodology strategy is to use a fully static CMOS circuit style with a low P/N ratio across the library and all custom designs. The massive interconnect network between the cores and the L2C banks is designed for an optimal power-delay product. As the flip-flop count is close to 2.5 million, special attention was paid to selecting the flip-flop type to balance speed and power trade-offs. A single-ended large-signal sensing scheme in all cache designs allows fine grain array and flip-flop gating, resulting in 10%–20% saving over the previous generation. In order to minimize the increased leakage, the use of gate bias (longer L) is more than doubled from the previous generation, while the use of low V_{th} transistors is avoided.

Fig. 11(a) shows the breakdown of components with highest power in a 2-way system. The sixteen SPARC cores consume 28% and the SerDes I/O consumes 22%. Even though the SerDes I/O consumes a high percentage of the total power, it does reduce the overall system power since no external hub chip is required. Increased leakage and total chip transistor width continue to make leakage a big portion of the chip power even with more aggressive leakage management features. The 16-Bank L2 cache system consumes 10% of the total power. The low clock frequency and extended clock gating makes the SoC power only 6% of the total chip power. The pie chart in Fig. 11(b) shows the breakdown of dynamic power excluding the SerDes links. Flip-flops are the dominant contributor to dynamic power consumption. SRAMs make the second largest contributor due to its dynamic operation nature. Interconnect power, which includes cluster and top level routing, represents 15% of the total dynamic power while the logic represents only 15%.

In addition to software controlled power throttling features, such as thread parking and instruction issue rates, fine grain core disable and down clocking features are implemented to manage idle power. The clock to each core is disabled to zero out the



Note: Assuming significant voltage gradient across M_N and M_{N-1}

Fig. 10. Analytical IR drop calculation for local metals.

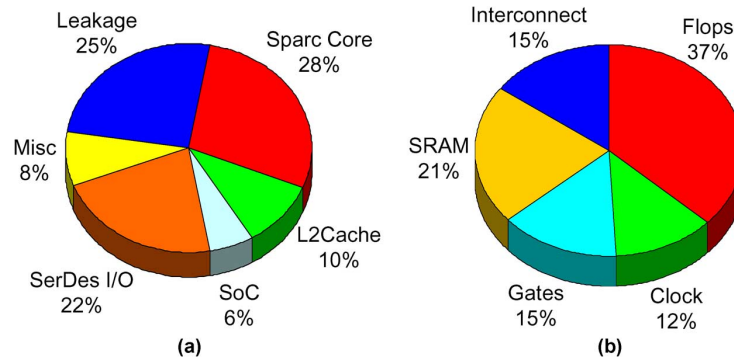


Fig. 11. (a) Two-way power breakdown. (b) Dynamic power component.

dynamic power of the fifteen cores. Only one thread in the remaining core will be active during idle mode to support the operating system. Down-clocking decrements the clock to one half of the operating frequency further reducing idle power. Frequency is reduced in steps of 1/8th to minimize Ldi/dt . Fig. 12 shows how power throttling reduces active power by $\sim 70\%$ including additional leakage savings obtained by the lower operating temperature from power throttling. During deep sleep mode, after core-disable and down-clocking are turned on, the $V_{dd-core}$ power supply can be lowered by the software power management.

VI. PHYSICAL DESIGN METHODOLOGY

Like in previous generations, the design utilizes a fully static cell-based design methodology to reduce implementation complexity. In order to balance performance and power, three different flip-flop types (sense-amplifier flip-flop, semi-dynamic flip-flop and standard master-slave flip-flop) are chosen based on functional and timing requirements. Since most of the design is implemented with cell-based automated flows, selective timing-driven wire classes and optimized repeater insertion techniques are used on critical signals to meet timing requirements. New design methodologies are developed to handle layout dependency effects in the advanced technology. Various

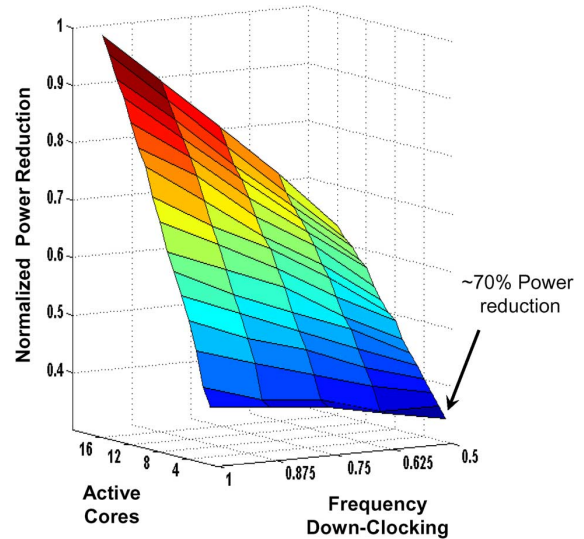


Fig. 12. Power throttling impact with core-disable and down-clocking.

techniques such as increasing poly spacing and adding dummy gates are used to improve performance on critical devices. Small

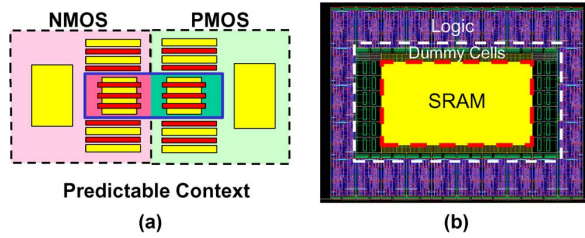


Fig. 13. (a) Library cell context. (b) Custom SRAM context.

devices are moved away from well boundaries to minimize device variations from proximity effects. Active and well joggings are avoided to reduce edge effects. Then, all library cells are characterized in a context that mimics the predicted cell environment in the actual design. Fig. 13(a) illustrates how well, diffusion and poly polygons are added around the library cell to emulate this context. Also, in order to avoid discontinuities between custom macros and composition blocks, proper filler cells are included around custom cell boundaries (Fig. 13(b)). Decoupling capacitors are specially designed with both P and N type transistors to provide a context friendly interface to the neighboring logic cells.

Increased process variability results in higher local mismatches due to V_t variation and lithography limitations. A traditional corner methodology is generally too pessimistic and often unable to catch circuit failures since all devices are skewed in the same direction. Historically, circuits were designed with 20% self-timed margins (STM), while flip-flops were characterized in a worst case hold time corner. Statistical methods are instead extensively deployed to analyze STM, custom memory cells, flip-flops, and all process sensitive circuits. Design of experiments reduces the required STM to less than 15%. The adopted methodology still guarantees the robustness of the high performance circuits for a 6-sigma yield and captures any failures across the full process spectrum. A new timing methodology based on statistical analysis is also developed to characterize the sensitive sense-amplifier flip-flop. A large number of statistical simulations are performed to first understand the timing induced failure mechanism and generate timing models with enough margins. Then, the failure rate is calculated based on this data and the margin is iteratively tuned until an acceptable chip yield is achieved. Fig. 14 shows how a too small design margin can lead to a catastrophic failure of the sense-amplifier flip-flop. A uniform distribution method is used to verify timing margins by equally sampling the full -3 sigma to 3 sigma window with emphasis to the tail of the distribution.

VII. CIRCUIT DESIGN

A. SerDes I/O

The main I/O interface is based on Serializer/Deserializer. This 312-lane high-speed SerDes delivers a maximum throughput of 2.4 Tb/s (~ 1.2 Tb/s per direction) in a 4-way system. The transmitter for the coherency and memory ports I/Os are implemented based on a current mode logic (CML) driver with 4-tap FIR. The pre-cursor provides compensation for the channel inter symbol interference (ISI) and a

zero-crossing lock point for the Muller-Muller (MM) clock data recovery (CDR). A small post-cursor tap 2 (5% of cursor) reduces the bandwidth requirement by the ADC to 3.5 GHz. The receiver features a digital MM-CDR and a 5-tap digital decision feedback equalizer (DFE) with feed-forward equalizer (FFE) based on a 13-level (3.7 bits) flash ADC [11]. Fig. 15 shows the receiver block diagram. At the receiver input, a 6 dB attenuator is integrated in the tapped termination block, doubling the dynamic range of the receiver. The ADC subsystem includes two track-and-hold switches and two interleaving half-rate ADCs. It converts the analog signal to a 4-bit sign magnitude digital code. The track-and-hold switch, not only reduces bandwidth requirement for the corresponding ADC, but also eliminates the signal related aperture timing (SRAT) error due to the variation of trans-conductance at the linear comparator. Inside each half-rate ADC (Fig. 16), a level shifter and a buffer are used to convert the signal common mode from ground-referenced to supply-referenced, and to drive the loading of the 13 ADC slicers. Each ADC slicer includes a linear comparator followed by a sampling latch, a sense amplifier and a RS flip-flop. A tunable clock delay path is used to match the delay of the clock and the analog paths, so that the digital sample-and-latch in the ADC slicer can happen at the end of the hold period. The reference voltages for the ADC slicers are generated by a resistor ladder with a 7-bit current DAC, which is controlled by an automatic gain control (AGC) loop. The 4-terminal differential linear comparator (Fig. 17) uses a common source for all four branches to improve the headroom for the bottom current source at 0.85 V. At the output of the linear comparator, a current steering technique is used to provide offset calibration with a 5-bit DAC.

Fig. 18 shows the high level schematic for the auto-zeroing scheme. Instead of using a pass-gate to short the inputs and reference levels, and of using a tri-state output buffer, an op-amp (OA1) is added to provide the virtual short between the data input and the corresponding reference levels. At system start up, the digital state machine steps through the calibration of Op-Amp1, the level shifter, and then each of the slices one by one. Once this calibration is done, the system starts the normal operation. Fig. 19 shows the measurement results and how differential nonlinearity (DNL) reduces by half with auto-zeroing. The receiver equalization is implemented in the digital domain and runs at $1/4$ of the bit rate to reduce latency. The digital FFE adds and subtracts one LSB from the current data sample based on the comparison of the threshold with a future data sample, therefore removing ISI at the pre-cursor location. In contrast to the pre-cursor tap used in the transmitter FIR, the use of FFE does not result in cursor level reduction, therefore benefitting the noise-limited link. The 5-tap digital DFE is implemented with a loop-unrolling technique. All DFE correction coefficients for the 32 combinations of the five history bits are stored in registers. The decision is then selected from the comparison results between the data samples and these registers based on the past decision bits. The DFE slice 0 and slice 1 run at clock phase 0 (Fig. 20) while EQ slice 2 and 3 run at clock phase 180. At pipeline stage 1, sixteen 2-to-1 multiplexers are used to select 16 out of 32 speculative results based on history bit 4. At Stage 2, a 16-to-1 multiplexer is used based on history bits 0 to 3. The op-

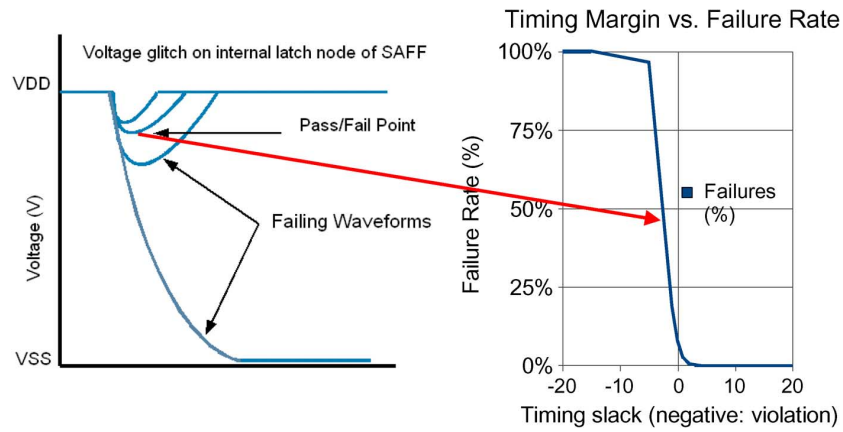


Fig. 14. Sense amplifier flip-flop failure rate sensitivity to timing margin.

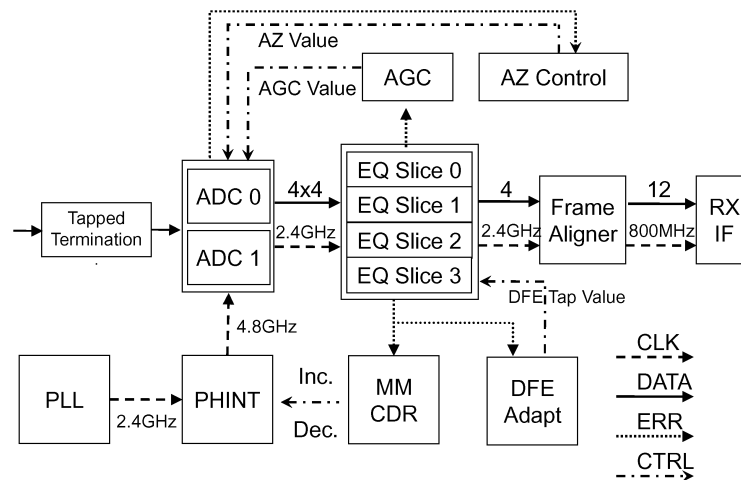


Fig. 15. SerDes receiver block diagram.

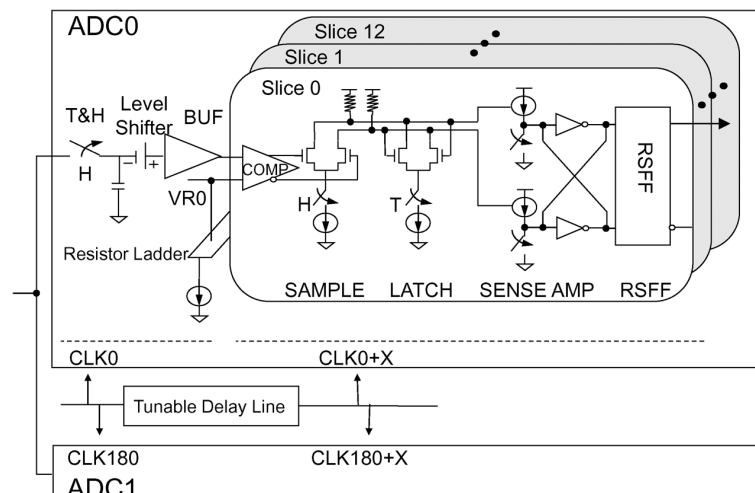


Fig. 16. Half-rate ADC architecture.

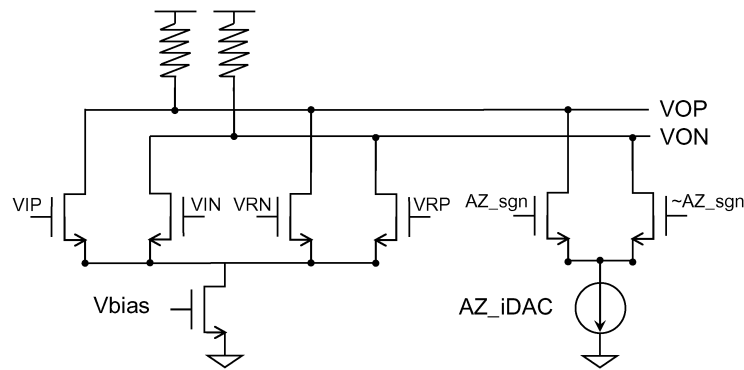


Fig. 17. Four-terminal linear comparator with auto-zeroing.

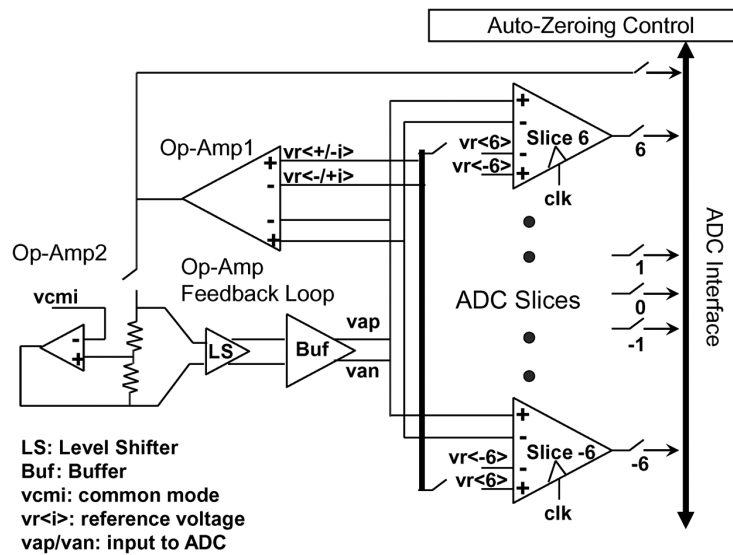


Fig. 18. Auto-zeroing scheme for 13-level ADC.

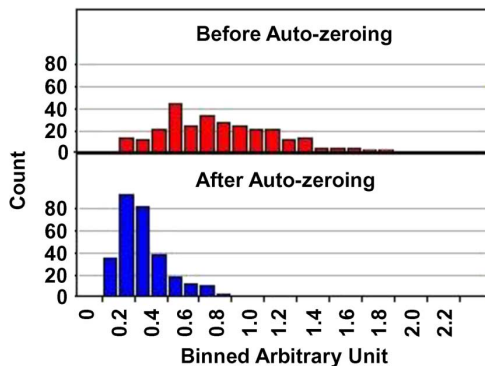


Fig. 19. Silicon measured histogram of differential nonlinearity (DNL).

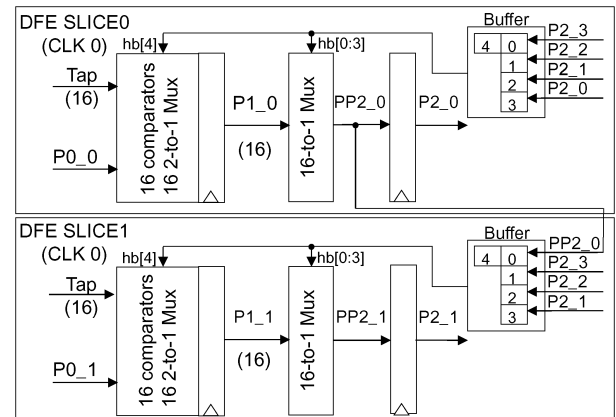


Fig. 20. Decision feedback equalizer (DFE) block diagram.

eration for the DFE slice 2 and 3 is similar. The DFE, CDR and the AGC adaptation run at 1/8 of the bit rate. The AGC loop adjusts the reference DAC code to lock the h0 (cursor height) of the signal after FFE to 4.5LSB. The DFE and CDR share the error signals generated by comparing the data sample after

DFE and h0 (4.5LSB). A central DFE adaptation unit is shared among seven lanes. This time sharing approach not only fulfills the DFE tracking requirement due to environmental changes,

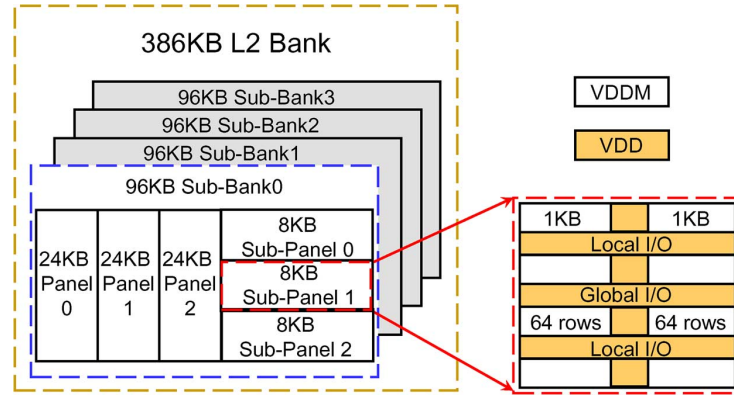


Fig. 21. 386 KB L2Data bank organization.

but also consumes significantly less power and area (70%) than a per-lane DFE adaptation scheme.

MM CDR is used instead of a bang-bang CDR to eliminate the need of comparators and quadrature clock trees usually associated with oversampling at the transition edge. This results in substantial power and area savings. It also reduces the input stage loading thus improving the overall receiver front-end bandwidth. The error function for this MM CDR implementation is $h(-1) = 0$. The error is computed based on samples from neighboring bits through pattern matching and grouping. In contrast to the use of an error function $h(-1) - h(1) = 0$ [12], this approach decouples the CDR adaptation loop from the adaptation of the first post-cursor tap in DFE. An advantage of using an all digital ADC approach is that all information can be accessed through the scan-chain. Therefore, the post processing of digital samples is performed at a relatively low cost. In diagnostics mode, the receiver is configured like a digital oscilloscope to obtain signal waveforms, such as pulse response, pseudo-random binary sequence (PRBS) patterns, and eye diagrams. The DFE, CDR and ADC gain data is also accessible. These DFE features are essential for the bring-up and test of hundreds of links.

B. L2 Cache

The 6 MB unified L2 cache consists of sixteen 24-way 386 KB banks to support the required bandwidth for the sixteen multi-threaded cores. Two L2Data banks share a common L2T and supports 16 B read and 4 B/8 B/16 B write operations. The L2T contains the L2 control logic, tag array and all cache status buffering registers. The L2Data is implemented using TSMC's $0.303 \mu\text{m}^2$ high performance bit cell with dual power rails to support the processor low voltage operation. Only column repair is implemented per each panel. Each L2Data bank contains Vdd_memory and VNWM bumps. Vdd_memory and Vdd_core grids are interleaved at all metal levels to balance power grid integrity and to maximize routing resources for top level signals. To simplify implementation complexity, Vdd_memory is used for cell arrays and the last two stages of the word line drivers. Each 386 KB bank is divided into four 96 KB sub-bank arrays (Fig. 21). Any of the sub-banks

can be accessed with back-to-back operations while access to the same sub-bank can only be requested every other cycle. Each sub-bank is implemented with a self-locking circuit to disregard the second consecutive request while servicing the first request to the same sub-bank. The 96 KB sub-bank is composed of four 24 KB panels. Each panel contains three 8 KB sub-panels which contain the 8-ways of the data set. Since way-select is used to multiplex the columns, one-hot detection is implemented to ensure only the required column turns on during each access.

Four 64-row arrays form an 8 KB sub-panel which implements a large signal sensing scheme (Fig. 22). The design is comprised of two cross-coupled inverters with two stacked NMOS and a pair of sensing PMOS for the top and bottom arrays. The separate sensing PMOS pair for top and bottom arrays reduces the leakage impact from the wide column select structure by half. The N1-N4 stacked NMOS are connected to the output of the NMOS column multiplexers. During read access, since one of the arrays is in pre-charge state, one pair of the NMOS is on while the other NMOS pair is waiting for signal to develop. Fig. 22 shows the simulated waveform of the sense latch operation. Once the wordline turns on, the bitline starts to droop to develop a signal. As the NMOS column multiplexer is electrically off until the bitline droops to below V_{th} , it decouples the capacitance of the sense latch from the bitline. This results in a faster signal developing on the bitline. Once the bitline droops to a sufficient level, the NMOS gate turns on and the BL signal is reflected to the sensing PMOS. A PMOS sensing device writes a new data to the sense latch. As DL feedbacks to the NMOS stack device (N2), it disables N2 therefore speeding up the writing of the data into the sense latch. This simple sense latch large-signal sensing scheme mitigates the impact of increased device mismatch in advanced technologies and eliminates the need of timing critical control signals. In addition shallow bitlines allows the use of smaller transistor sizes in the I/O circuitry. The NMOS only read and write column-select and sharing a single sense latch between each two arrays make the L2Data array area comparable to a conventional sense-amplifier based design. Since only one to four 2 KB arrays are active per sub-bank, 98% of the L2D is always disabled with fine-grain array gating.

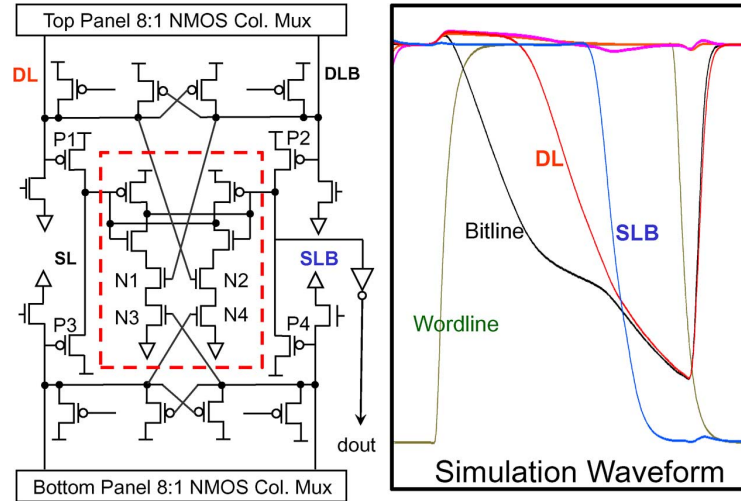


Fig. 22. L2Data sensing scheme and waveforms.

C. Crossbar

The crossbar delivers a maximum bandwidth of 512 GB/s. One of the top physical implementation challenges was how to handle the routing complexity from so many cores and caches with large sets of data buses. In addition, minimizing core to L2 latency is key to reduce L2 load-to-use performance penalty. To solve these challenges, we introduced a new structure called crossbar gasket (CXG), between the two cores and the crossbar (CCX). This approach creates a staged multiplexer next to the cores to reduce interconnect. In addition, a shared L2T with dual pipe-lines serves two L2Data banks to reduce routing from CCX to L2T. Sharing ports among cores and L2T further reduces routing congestion and timing penalty. Staging registers balance the latencies from the various core locations. There are two main datapaths in the crossbar: PCX handles the cores to the cache out-streaming data, while CPX processes in-streaming data from the cache to the cores. There are eight rows of each PCX and CPX datapaths and nine data columns (Fig. 23). The center of the crossbar is the arbitration logic which dispatches requests based on age for fair scheduling. Each datapath includes two queuing entries, which processes data in a 4-stage pipeline. In order to handle the high bandwidth and complex routing density, the design is implemented with a semi-custom approach. Each datapath slice is hand-built as a custom macro with pre-aligned pins and tracks. Then a rule-based automated flow is used to create predictable routes, thereby meeting performance targets and avoiding noise challenges by construction. This modular approach reduces both design and electrical verification complexity.

VIII. DESIGN FOR TESTABILITY

The design incorporates extensive DFT features, including logic and SRAM test using segmented scan chains, and SerDes internal and external loop-back tests [13]. Scan chains are segmented to enhance test efficiency. The scan chain compression helps to reduce scan pin count. Timing critical paths are evaluated with scan based transition test, path delay test, clock-stretch and clock-stop techniques. The design achieves 98%

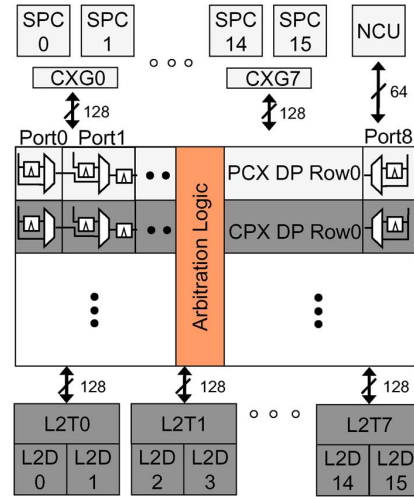


Fig. 23. Interconnect crossbar and gasket block diagram.

single stuck-at fault and 85% speed test coverage. All memories are tested through MBIST engines. In addition, single port cache arrays implement programmable MBIST to allow data pattern changes and direct observability for the repair bits. Small arrays include scan-based at-speed test to allow back-to-back read or write operations with a single load of the scan collar input flip-flops.

IX. CONCLUSION

We have introduced the fourth generation T-series processor in 40 nm technology. The UltraSPARC T3 doubles on-chip thread counts and includes support for 4-way glueless scalability to deliver up to 512 threads in a 4-way glueless system configuration. Therefore, the T-series line continues to double thread count per year. The high bandwidth required by the large number of threads is supported by, the 6 MB 16-Bank L2 cache, the hierarchical crossbar and the 2.4 Tb/s high-speed SerDes. This SoC processor implements multiple clock and power

domains as well as extensive power management techniques to optimize power and performance trade-offs. Special circuit techniques and a robust design methodology overcome the 40 nm silicon challenges.

ACKNOWLEDGMENT

The authors acknowledge the contributions from the entire UltraSPARC T3 CPU Development Team, and all those who supported the project at Oracle, TI and TSMC. Special thanks go to David Greenhill and Ray Heald for their technical advice and to Rodrigo Liang for his management support.

REFERENCES

- [1] P. Kongetira *et al.*, "A 32-way multithreaded SPARC processor," *IEEE Micro*, vol. 25, pp. 21–29, Mar. 2005.
- [2] A. S. Leon *et al.*, "A power-efficient high throughput 32-thread SPARC processor," in *2006 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech Papers*, Feb. 2006, pp. 98–99.
- [3] A. S. Leon *et al.*, "A power-efficient high throughput 32-thread SPARC processor," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 7–16, Jan. 2007.
- [4] U. G. Nawathe *et al.*, "Implementation of an 8-core, 64-thread, power-efficient SPARC server on a chip," in *2007 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech Papers*, Feb. 2007, pp. 108–109.
- [5] S. Phillips *et al.*, "Victoria Falls: Scaling highly-threaded processor cores," in *2007 Hot Chips 19 Symp.*, Stanford, CA, Aug. 2007.
- [6] S. Patel *et al.*, "Sun's next-generation multi-threaded processor—Rainbow Falls," in *2009 Hot Chips 21 Symp.*, Stanford, CA, Aug. 2009.
- [7] J. L. Shin *et al.*, "A 40 nm 16-core 128-thread CMT SPARC SoC processor," in *2010 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech Papers*, Feb. 2010.
- [8] A. S. Leon *et al.*, "The UltraSPARC T1 processor: CMT reliability," in *Proc. 2006 IEEE Custom Integrated Circuits Conf. (CICC)*, Sep. 2006, pp. 555–562.
- [9] U. G. Nawathe *et al.*, "Implementation of an 8-core, 64-thread, power-efficient SPARC server on a chip," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 6–20, Jan. 2008.
- [10] I. Jones *et al.*, "Synchroniser behaviour and analysis," in *Proc. 15th IEEE Int. Symp. Asynchronous Circuits and Systems*, May 2009, pp. 117–126.
- [11] M. Harwood *et al.*, "A 12.5 Gb/s SerDes in 65 nm CMOS using a baud rate ADC with digital receiver equalization and clock recovery," in *2007 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech Papers*, Feb. 2007, pp. 68–69.
- [12] V. Balan *et al.*, "A 4.8–6.4-Gb/s serial link for backplane applications using decision feedback equalization," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1957–1967, Sep. 2005.
- [13] R. Molyneux *et al.*, "Design for testability features of the SUN Microsystems Niagara2 CMP/CMT SPARC chip," in *Proc. Int. Test Conf.*, 2007, pp. 1–8.

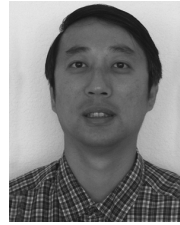


Jinuk Luke Shin received the M.S. degree in electrical engineering from the University of Texas, Austin, in 1995.

He joined Sun Microsystems Inc., Sunnyvale, CA, in 2000. He was a senior design manager for the Rainbow Falls (T3) processor, being responsible for SRAMs, analogs, power, technology development and silicon validation. Before that, he led memory designs for the first generation Niagara (T1) and the dual-core UltraSPARC processors. He is currently a senior design manager at Oracle Corp., Santa Clara, CA, focusing on clocking, power management, technology development and analog components for future generation microprocessors. His research interest is in high-speed and low-power circuit design techniques and methodology.

Prior to joining Sun, he was with Motorola, Austin, TX, from 1995 to 1997, where he was involved in the design of embedded flash memories for digital signal processors. From 1997 to 2000, he was with Hitachi Semiconductor America, San Jose, CA, where he developed the level-1 cache arrays for a SuperH processor. He is the author of 28 technical papers in design, device and test areas, and holds several issued and pending U.S. patents.

Mr. Shin is a member of Eta Kappa Nu and Tau Beta Pi.



Dawei Huang (M'00) received the B.S. degree in precision instruments from Tsinghua University, Beijing, China, in 1997 and the M.S. degree in electrical engineering from the University of California at San Diego in 1999.

He joined Sun Microsystems in 2000 and has worked on various high-speed optical and electrical system interconnect projects at Physical Science Center of Chief Technology Office. Since 2005, he has been engaged with the architect, design, verification and characterization of multiple generations of SerDes IO for UltraSPARC T-series CPU. He is currently a senior principal engineer at Oracle Incorporation's Mixed Signal Design group, where he leads the SerDes architecture development effort for next generation CPU. He has two granted patents and ten pending patent applications. His areas of interest include low-power high-speed SerDes IO architecture and short reach optical interconnect.



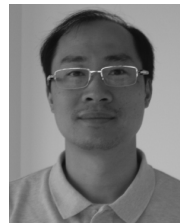
Bruce Petrick received the B.S. and M.S. degrees in electrical engineering from Montana State University, Bozeman, MT, in 1972 and 1974, respectively.

He joined Sun Microsystems in 1994 and worked as a design engineer on the UltraSPARC II. After that, he has worked on the UltraSPARC III, the dual-core UltraSPARC II, Niagara2 and most recently Rainbow Falls (T3). On Rainbow Falls, he was responsible for the design of the clocking structures and the interfaces between clusters which could be clocked in either an asynchronous or synchronous mode.



Changku Hwang (SM'97) received the B.S. degree in electronics engineering from Hanyang University, Seoul, Korea, and the M.S. and Ph.D. degrees in electrical engineering from the Ohio State University, Columbus, OH, in 1985, 1992, and 1996, respectively.

He was a post-doctoral researcher with Analog VLSI Lab of the Solid-State Microelectronics at the Ohio State University in 1997 and a visiting senior researcher in the Communication Systems Department of the Central Research Lab, Hitachi, Ltd., Tokyo, Japan, in 1998, focusing on design of low-voltage low-power analog and analog-digital mixed mode VLSI circuits and RF circuits for communication applications. Since then, he has been engaged with the design of high performance microprocessors with special emphasis on high frequency clock generation, power distribution, I/O circuits and special circuits. Currently, he is with Oracle where he is responsible for development of core analog circuits in various SPARC-based high performance low-power multi-thread microprocessors projects. He has published more than 36 technical papers.



Kenway W. Tam (S'89–M'94) received the B.S. and M.S. degrees in electrical engineering and computer science from the University of California at Berkeley in 1992 and 1994, respectively.

From 1994, he was a circuit designer with Sun Microsystems working on high-speed memory design, custom circuits, and design methodology for high-performance microprocessors, including the UltraSPARC II, III and IIe. In 2001 he joined Afara Websystems (later acquired by Sun) where he led the custom memory designs and methodologies on Niagara processor. He then managed all custom register files and CAMs, and co-managed custom design methodologies on Rainbow Falls processor. He is currently a senior SRAM design manager at Oracle microelectronics microprocessors. He holds six U.S. patents and is a coauthor of several technical publications.



Alan Smith received the B.Eng. degree in electronics from the University of Glasgow, Scotland, U.K., in 1990.

He joined Sun in 1997 and has worked on UltraSPARC III, IIIi, the dual-core UltraSPARC, Niagara, Niagara2 and most recently T3. He is currently a senior principal hardware engineer, focusing on physical design methodology and analysis. Prior to joining Sun, he worked at ST Microelectronics in Europe in DFT and methodology roles.



Francis Schumacher received the B.S.E.E. and M.S.E.E. degrees from Stanford University, Stanford, CA, in 1984 and 1986, respectively.

He has spent his career in the Santa Clara valley developing VLSI for server class computers. He has focused primarily on systems and circuit design, with an emphasis on IO and debug. He is currently the director for Oracle's mixed signal design group.



Ha Pham received the B.S. degree in electrical engineering and computer science from the University of California at Berkeley in 1996.

From 1996 to 1998, he was a component design engineer at Intel Corporation. Since 1998, he has been with Sun Microsystems, Santa Clara, CA, and then Oracle Corporation, Santa Clara, CA, as a Principle Engineer. He has been working on embedded SRAMs and other custom circuit designs on the UltraSPARC IIe, the dual-core UltraSPARC, the first generation Niagara (T1) and the Rainbow Falls (T3) processors. He is currently working on L1 Cache designs for the next generation SPARC microprocessor.



Ana Sonia Leon (S'83–M'88–SM'06) received the M.S.E.E. degree from the University of Southern California, Los Angeles, CA, in 1988.

From 1989 to 1996, she was a lead senior circuit designer with Motorola, Inc., Austin, TX, working on DSP processors. In 1996 she joined Chromatic Research, Sunnyvale, CA, as a senior circuit designer working on multimedia processors. Since 1998 she was with Sun Microsystems, Inc., Sunnyvale, CA, where she served as Director of Technology and led the physical design of several high-performance SPARC processors, including the UltraSPARC IIe, the dual-core UltraSPARC, and the first generation of CMT processors Niagara; being responsible for foundry interface, memories, digital and analog designs, clock, power, circuit techniques, design methodology, chip integration and silicon productization. She is currently a Senior Principal Engineer at Oracle Corp., Santa Clara, CA, focusing on the research and development of advanced design technologies for 20 nm and beyond. Her interests are in advanced circuit technologies to further extend the Moore's Law progression. She has authored 16 papers in microprocessor design and holds several U.S. patents in circuit design.

Ms. Leon serves on the ISSCC Technical Program Committee.



Hongping Li received the M.S.C.E. and B.S.E.E. degrees from HuaZhong University of Science and Technology, China, and the M.S.C.S. degree from the University of South Florida in 1991.

She joined Sun Microsystems, Sunnyvale, CA, in 2002 and has led physical design on several multi-core multi-thread microprocessors. She is currently the Director of physical design at Oracle for future generation processors. Prior to Oracle, she worked at IBM Microelectronics, Avanti, and Raza Foundries.



Allan Strong received the S.B.E.E. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, and the M.S.C.S. degree from Stanford University, Stanford, CA.

He spent 20 years at HP designing and managing VLSI chips for servers including processors, bus converters, memory controllers, cache controllers, and chipsets. In 2000 he joined inSilicon as VP of engineering and in 2001 he became VP of VLSI at Afara Websystems where he began developing what became the SPARC T1 processor. Afara was acquired in 2002 by Sun Microsystems, and as Senior Director he developed the T1, T2+ and T3 SPARC processors. Oracle acquired Sun in 2010, where he finished T3 productization.



Timothy Johnson received the M.Eng. degree from York University, U.K., in 1989.

His early contribution to microprocessor design was in the area of IEEE floating point RTL and physical design on the INMOS/SGS-Thomson T9000 Transputer. On moving to the United States, he joined Sun Microsystems working on the Sun's first integrated SPARC processor: UltraSparc Iii. While at SUN/Oracle, he has focused on the physical design of microprocessors including CPU/cluster place and routing, timing closure, design integrity

and product tape-out activities. Over the last 15 years, he has contributed to eight SPARC microprocessors. His latest role is as the Director of hardware engineering on 40 nm and 28 nm SPARC designs.