4.2 A 20nm 32-Core 64MB L3 Cache SPARC M7 Processor

Penny Li, Jinuk Luke Shin, Georgios Konstadinidis, Francis Schumacher, Venkat Krishnaswamy, Hoyeol Cho, Sudesna Dash, Robert Masleid, Chaoyang Zheng, Yuanjung David Lin, Paul Loewenstein, Heechoul Park, Vijay Srinivasan, Dawei Huang, Changku Hwang, Wenjay Hsu, Curtis McAllister

Oracle, Redwood Shores, CA

The SPARC M7 processor delivers more than 3× throughput performance improvement over its predecessor SPARC M6 for commercial applications. It introduces new design features, such as the S4 core, a 64MB L3 cache sub-system with application data integrity, a low-latency, high-throughput on-chip network (OCN), a database analytic accelerator (DAX), fine-grain adaptive power management and 1.5× higher SerDes I/O bandwidth for memory, coherency and system interfaces (Fig. 4.2.1) [1]. The enhancements in the S4 core over the S3 core [2] include a new L2 cache scheme, support for visual instruction set (VIS) extensions, virtual address masking and user-level synchronization instructions to provide continuous single-thread performance improvement for SPARC processors since SPARC T4. In addition, a hierarchical modular approach, called SPARC cache cluster (SCC), is used for the core-L2-L3 cache system. Within the SCC, all four cores share a single 256KB L2 instruction cache and each core pair has its own 256KB L2 data cache. The L2 caches are organized as 2-banks and 8-ways to deliver greater than 1TB/s bandwidth to the four cores. This L2 system delivers 2× more throughput for each core with 1.5× increase in size and the same latency as the previous generation L2 cache scheme. The L2 caches connect to an 8MB, 8-way set-associative partitioned L3 cache. Having a localized L3 cache within each SCC reduces L3 latency by 25%. The chip contains eight SCCs for a total of 32-cores with 256 threads and a 64MB L3 cache with 1.6TB/s bandwidth. In order to support the bandwidth and latency requirements from 256 threads and other system agents, the OCN architecture is implemented in place of a crossbar based network used in previous SPARC processors. Each SCC connects to the OCN, which in turn connects to four on-chip memory controllers (MCUs), coherency systems and eight database analytic accelerator (DAX) engines. The SPARC M7 introduces a customized DAX engine in an effort to optimize performance for Oracle databases. Eight DAX engines handle simple query predicates, decompression, message passing and interrupts across cluster nodes. This query accelerator provides up to 10× better performance for single stream decompression.

The 0.5TB/s data bandwidth OCN is logically made of a multi-stage data network, a request network with a 4-ring topology, and a point-to-point response network. The network stations are built with modular tiles with serving circuitries embedded in each tile. The data network utilizes long distance repeaters to convey data in 2 cycle hops between 6 micro-crossbar stations. Every station contains local queues from different sources with an arbiter to pick the forward station output and prevent stalls (Fig. 4.2.2). The request network uses a circular ring topology to distribute 72b multi-cast requests among all clients. The response network has 7 stations connected by 2-cycle-hop repeater chains. Dedicated wires per responder to each client allow numerous responses to arrive at the same station simultaneously. The bus organization and routing topologies are designed to balance routing density for area, and timing for performance. Both request and data networks use an opposite-direction routing topology to reduce the long wire coupling capacitance impact arising from timing window overlap. The OCN is able to reduce 35% of the coupling capacitance and 10% total RC delay without a routing channel increase.

The chip is implemented in TSMC's 20nm process utilizing four different threshold devices, a 1.5V I/O device, two different sizes of memory cells and 13 layers of Cu interconnect. Each SCC is a self-contained unit that includes its own clock generator, temperature sensors, power supply calibrator, global controller to handle clock initialization and SRAM/logic tests. The SCC communicates to the OCN through the voltage shift module (VSM) using source-synchronous interfaces to provide low latency between different voltage and frequency domains. The OCN runs at 4.0GHz to handle the bandwidth requirement. There are five additional SoC clock domains: coherence controllers operate at 2.4/1.2GHz, the DDR4 memory controllers at 1.33GHz, DAX engines at 2.0GHz and a system clock (Fig. 4.2.3). All SoC components are in the same supply domain. The third-generation asymmetric frequency lock loop (AFLL) in each SCC uses a regulated supply for the wire DCOs to reduce period compression during supply overshoot, furthering the benefits of prior adaptive clocking schemes [3]. In addition, having an AFLL within each SCC reduces the global clock tree latency and provides a faster feedback loop between the AFLL and logic paths to enable high-frequency supply noise compensation. The AFLL provides up to 2× guard-band reduction compared to the previous FLL (Fig. 4.2.4).

In order to limit the power envelope of the SPARC M7 processor, reducing power consumption and enhancing power management have been a priority since the inception of the architecture. As SRAMs and SerDes are the main contributors to the processor power, these custom circuits are optimized for power-speed trade-offs. The SRAMs use three different types of V_t devices to maximize leakage saving, while reducing dynamic power for high-activity nodes. The L3 data cache consumes a significant portion of chip leakage power. Designing the cache with higher V_t transistors and a stacked PMOS for the last wordline driver stage reduces its leakage by 89%. The dummy transistor in the strap area of wordline drivers is converted to a stacked PMOS with no impact to area and a speed degradation of only 3.8% of the wordline delay. As the NMOS of the wordline driver is a standby-on transistor, a lower V_t transistor of smaller width is used to achieve the same speed with reduced area and gate capacitance. Applying this "mixed V_t" approach to global and local wordline drivers offers 3% L3 power reduction. In the L3 cache, the read way hit (*rd_wayhit*) signal is too timing critical to be used to deactivate the read operation of its constituent sub-arrays. Therefore, a speculative way hit signal is introduced to achieve 16% dynamic power saving during L3 cache misses. This signal is used to disable the clock generation circuitry, sense amplifiers and all associated downstream logic (Fig. 4.2.5). The accuracy of the speculation is higher than 95% for most applications.

The SPARC M7 processor employs 280 SerDes lanes supporting up to 18Gb/s line rate and 1TB/s total bandwidth for memory, multi-socket coherency and IO interfaces. The common architecture is based on a 4-tap FIR transmitter and an Analog Decision-Feedback Equalization (ADFE) receiver with baud-rate (Mueller-Muller) CDR and odd/even data paths. Two different variations are used to reduce total power consumption: a short-reach (SR) 1-tap slicer ADFE employed for channels with 14dB or less loss, and a long-reach (LR) 10-tap summer-based ADFE for channels with up to 20dB of loss. Power for the SR design is <11mW/Gb/s per-lane (RX+TX), while the LR design is <14mW/Gb/s per-lane. Power savings is achieved with direct feedback and reduced error-slicer count without negatively impacting performance. With a large number of SerDes lanes, calibration and adaptation are essential to maintain yield and performance. Therefore, the SerDes design makes extensive use of feedback, power-up calibration and real-time adaptation circuits. Calibration is featured in multiple components: offset-cancelling auto-zeroing comparators; duty-cycle distortion adjustments in the TX output; PLL range and regulated voltage optimization for minimizing jitter. Automatic continuous run-time adaptation is achieved with PRBS scrambled data for DFE tap weights using adjustable resolution control, separate odd/even DFE taps and back-channel FIR adaptation on all taps. Single tap DFE in the SR is achieved in one cycle by dynamic generation of a reference voltage (Fig. 4.2.6). A reference generator was chosen over the traditional summing node in order to reduce power consumption and reduce capacitance on the data node. Further benefits include the ability to support large tap weight values, as well as allowing individual control of separate tap weights for data and error slicers (and odd/even instances) to compensate for manufacturing variations. Almost all adaptation and calibration controls are accessible via software, which was used to enable extensive characterization of margin, BER and even channel loss without adding power or area to the design.

Acknowledgements:
The authors acknowledge the contributions of the talented SPARC M7 team and TSMC for manufacturing.

References:
[1] Stephen Phillips, "M7: Next Generation SPARC," *Hot Chips*, 2014.
[2] J.L. Shin, *et al.*, "A 40nm 16-Core 128-Thread CMT SPARC SoC Processor," *ISSCC Dig. Tech. Papers*, pp. 98-99, Feb. 2010.
[3] Y. Yangong, *et al.* "A 28 nm Asymmetric Frequency Locked Loop," *Asian Solid-State Circuits Conf.*, 2014.
[4] V. Krishnaswamy, *et al.*, "Adaptive Power Management for SPARC M7 Processor," *ISSCC Dig. Tech. Papers*, Feb. 2015.

 978-1-4799-6224-2/15/$31.00

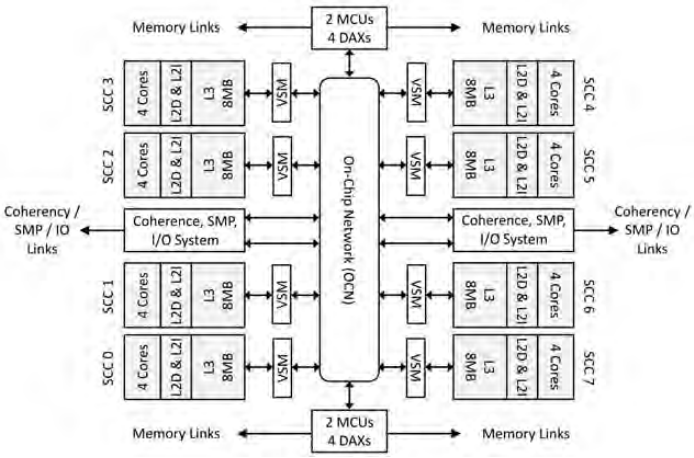

Figure 4.2.1: SPARC M7 processor functional block diagram.

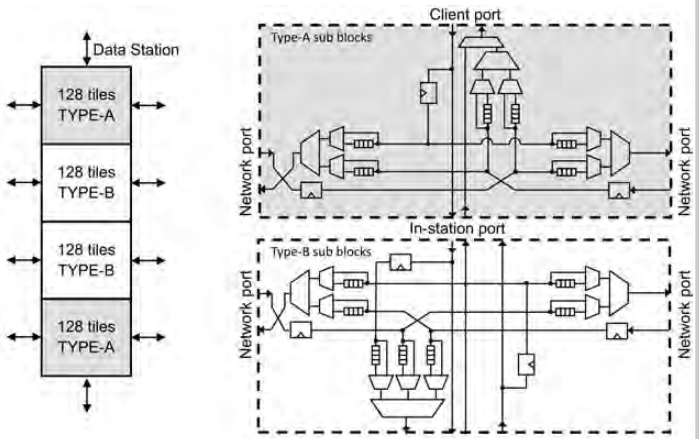

Figure 4.2.2: OCN data station diagram with two types of sub-blocks.

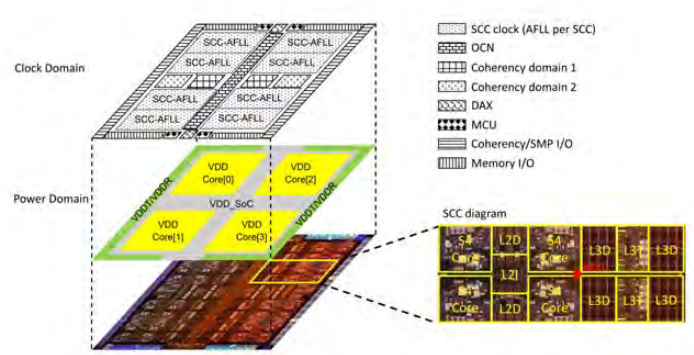

Figure 4.2.3: Clock and power domains with a detailed SCC view.

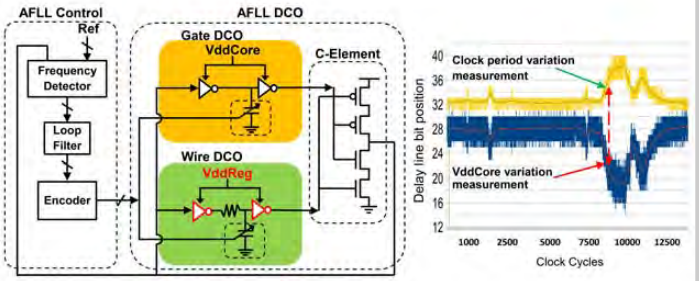

Figure 4.2.4: 3rd generation AFLL architecture. Clock period and supply variation measurements showing longer clock period when supply droops.

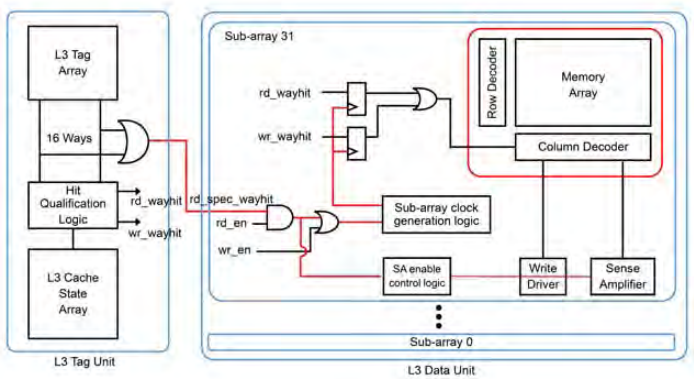

Figure 4.2.5: L3D speculative way selection and its usage for SRAM gating.

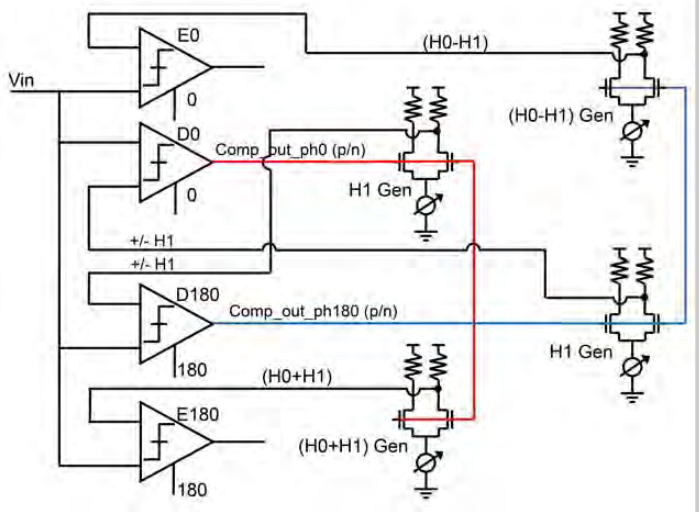

Figure 4.2.6: Single-tap DFE with dynamic reference switching.

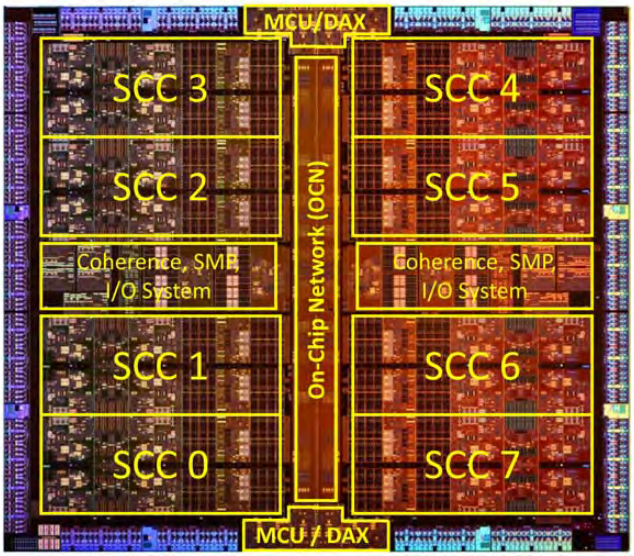

Figure 4.2.7: SPARC M7 die photograph.

978-1-4799-6224-2/15/$31.00