4.3 Fine-Grained Adaptive Power Management of the SPARC M7 Processor

Venkat Krishnaswamy, Jeffrey Brooks, Georgios Konstadinidis, Curtis McAllister, Ha Pham, Sebastian Turullols, Jinuk Luke Shin, Yifan YangGong, Haowei Zhang

Oracle, Redwood Shores, CA

The goal of the power management system of Oracle's SPARC M7 CPU [1] is to maximize the performance of commercial, cloud and big-data workloads subject to thermal and electrical constraints in a variety of system environments. Under thermal constraints, the goal is to maintain silicon die temperature within a target limit with a time constant of milliseconds to seconds, and under electrical constraints the current draw on a supply must be maintained below a threshold with a time constant of microseconds. The processor is designed to run a range of commercial workloads, potentially in virtualized environments, resulting in widely varying activity factors across the chip. Thus, the system must implement a fine granularity of control to achieve maximum performance across the entire range of workloads. These challenges are addressed with a number of advances over the previous generation processors [2] by implementing a low-latency on-die hardware power management system comprised of fast, accurate sensor designs (dynamic power meters and thermal sensors), a proportional feedback control system implementing thermal, current and chip power capping algorithms (Fig. 4.3.1), and actuation mechanisms with different latencies and structural granularities. The 32 cores and 64MB L3 cache are organized as 8 SPARC cache clusters (SCCs) each comprising 4 cores and an 8MB L3 Cache partition. Each SCC includes 6 dynamic power meters and 2 temperature sensors. The granularity of actuation for power management includes SCC-level clock cycle skipping, SCC level frequency scaling and voltage scaling for groups of 2 or 4 SCCs, as shown in Fig. 4.3.2. The power management system supports multiple DVFS power domain configurations.

The on-die digital dynamic power meter (DDPM) (Fig. 4.3.3) computes core average power using a rolling average of signal activity over a fixed number of core clock cycles enabling generation of a new power value every 16-1024 nanoseconds. It implements the function $P = k + \sum_i c[i]*x[i]$. This expression computes the dynamic power of a digital block as the sum of a fixed intercept, k, and the summation of rolling average signal transition activity ($x[i]$) over an interval, multiplied by an empirically determined coefficient ($c[i]$) for n signals. The accuracy of the power calculation depends on careful selection of signals, such that they correspond to the activity of structures that have high power consumption, while being minimally correlated within the averaging window, maximizing the possibility of positive coefficients. On the SPARC M7, there are two versions of the DDPM instantiated: one per S4 core that monitors the activity of 48 signals to estimate core, L1 and L2 cache power; the other monitors the activity of 8 signals to estimate L3 bank power. A linear regression-based fitting process is used to compute the coefficients using data gathered from netlist-level pre-silicon power analysis of over 15 million cycles of pre-silicon simulation. Fig. 4.3.4(a) shows the pre-silicon fit of the model versus power analysis at the core level, which predicts a fit with an R^2 of 0.9986 and a RMSE of less than 5%. Fig. 4.3.4(b) shows power computed by hardware vs. electrical measurements confirming that the difference between the DDPM and the measurement is less than 6% on a variety of workloads. The dimensions of the core DDPM in 20nm are 203×152μm^2 and it comprises 2231 flops organized into an arithmetic datapath and control unit.

The on-die temperature measurement is done using a bandgap-based sensing scheme. The bandgap circuit generates 2 voltages, V_{BE1} and V_{BE15}, from 2 thermal diodes with different current densities. The difference between V_{BE1} and V_{BE15} (ΔV_{BE}) is linearly dependent on temperature. The voltages are first converted to time by charging a capacitor with a constant current I_{REF} until the capacitor voltage equals the analog voltages V_{BE}. Time-to-digital conversion is done by counting the number of cycles of a fixed-frequency clock taken for charging the capacitor. The digital codes N_{BE1} and N_{BE15} are used to calculate the temperature (Fig. 4.3.5). Two temperature sensors are located symmetrically near hotspots between S4 cores within each SCC giving a total of 16 sensors per CPU. The area of each sensor is 300x50μm^2, and it achieves an accuracy of ±3°C. A new temperature sample can be generated at a rate of up to once every 20μs, which is well within the thermal time constant of the die.

Each SCC supports frequency scaling in steps of 133MHz using an asymmetric frequency-locked loop (AFLL [2]). Further, the frequency of the core clock to the SCC can be rapidly changed in steps of 133MHz using clock cycle skipping. Voltage scaling of groups of 2 or 4 SCCs is supported. A highly configurable on-die hardware-power-management controller (PMC) implements 3 proportional feedback control loops to cap the maximum SCC current, temperature, and full-chip power to software programmable limits. The PMC periodically samples dynamic power and temperature values and calculates leakage, as well as total power per SCC and for the chip, on the basis of which it may throttle or resume the frequency of any violating SCC. A novel feature of the PMC is the control of full-chip power to a given threshold, while supporting optional SCC priorities based on throttling using a weighted round-robin algorithm.

Current constraints can manifest as transients (μs), whereas thermal constraints are relatively steady state (ms to s). The PMC initially reacts to any capping request by skipping core clock cycles with extremely low latency (ns). The PMC hardware monitors whether the SCC cycle skipping state persists for a programmable dwell time, and initiates a change in DVFS operating point with a latency of several μs. In this way, the PMC can satisfy transient capping timing constraints, while steady-state responses are handled using more optimally using DVFS.

The clock generator used on the SPARC M7 CPU is the AFLL, an evolution of the previously reported AFLL [3] that reduces the *Ldi/dt* voltage guardband by 50% relative to a PLL by implementing an asymmetric voltage-to-frequency gain ($\partial V/\partial f$) that reacts more to voltage undershoots than to overshoots through the use of carefully sized wire and gate dominated DCOs whose outputs drive a Muller C element. In this implementation, a regulated supply is used to further reduce the frequency gain for voltage overshoots.

The PMC provides support for system-level power-supply-related RAS features at the DC-DC convertor level, as well as redundant AC grid-supply failures. The failure of a phase of a DC-DC convertor is detected and directly signaled to the processor in hardware leading to halving of current capping thresholds. Fig. 4.3.6 shows the steep power drop-off through cycle skipping occurring within 10μs (well within the convertor's over-current limit) followed by a gentler DVFS migration-enabled decline. The system also has the ability to detect power supply oversubscription that may occur upon loss of a redundant grid; in this case, it can throttle the processor through a dedicated hardware interface. The PMC supports all of the software power-management policies for idle mode power optimization described in [2]. In addition, there is support for VRM level power gating.

The power management system described in this paper enables more than 3× increase in power-constrained performance over the previous generation of SPARC server CPUs [2]. The low latency and high performance of the system is possible due to accurate, high-bandwidth sensors, fast on-die control and fine-grained actuation implemented using both clock cycle skipping and DVFS, as required by the time constants of system constraints.

Acknowledgements:
The authors would like to acknowledge the hard work and talent of the extended SPARC M7 development teams.

References:
[1] S. Phillips, "M7: Next Generation SPARC", *Hot Chips*, 2014.
[2] V. Krishnaswamy, *et al.*, "Bandwidth and Power Management of Glueless 8-Socket SPARC T5 System", *ISSCC Dig. Tech. Papers*, pp. 58-59, Feb. 2013.
[3] Y. YangGong, *et al.* "A 28 nm Asymmetric Frequency Locked Loop", *Asian Solid-State Circuits Conf.*, 2014.

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

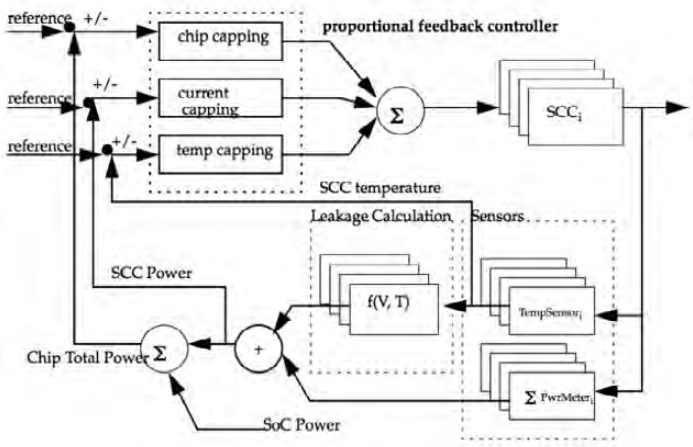


Figure 4.3.1: PMC proportional feedback control-loop architecture.

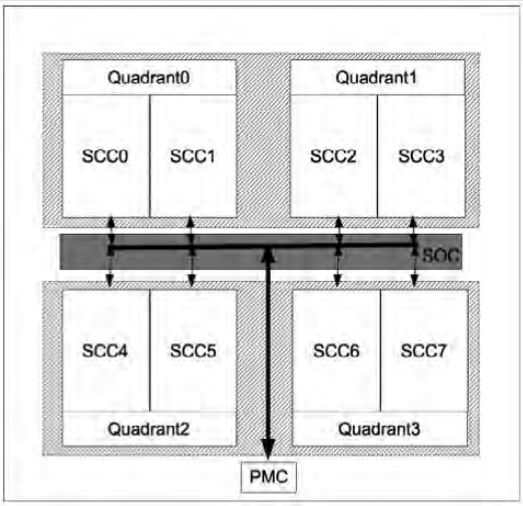


Figure 4.3.2: SPARC M7 showing 4 variable voltage SCC domains (quadrants) and one fixed voltage SOC power domain and transfer of sensor data and control information between PMC and SCCs.

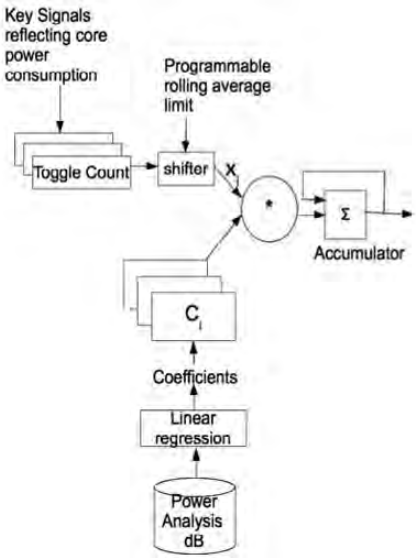
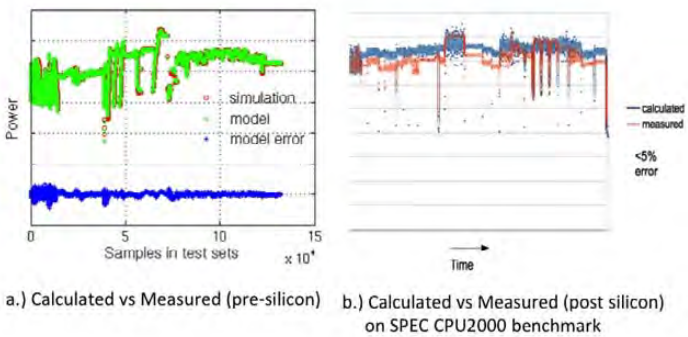


Figure 4.3.3: On-die digital dynamic power meter.



a.) Calculated vs Measured (pre-silicon)

b.) Calculated vs Measured (post silicon) on SPEC CPU2000 benchmark

Figure 4.3.4: On-die dynamic power meter accuracy.

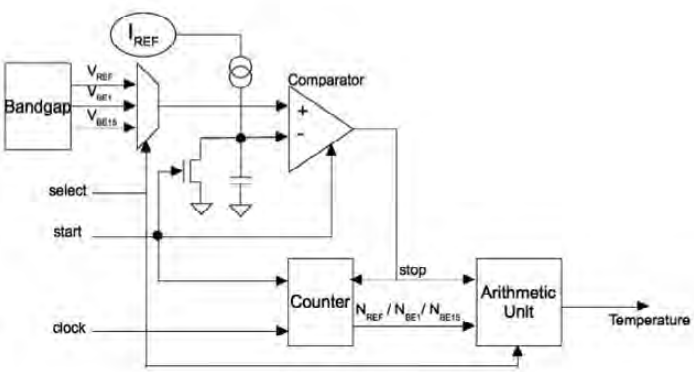


Figure 4.3.5: Temperature sensor block diagram.

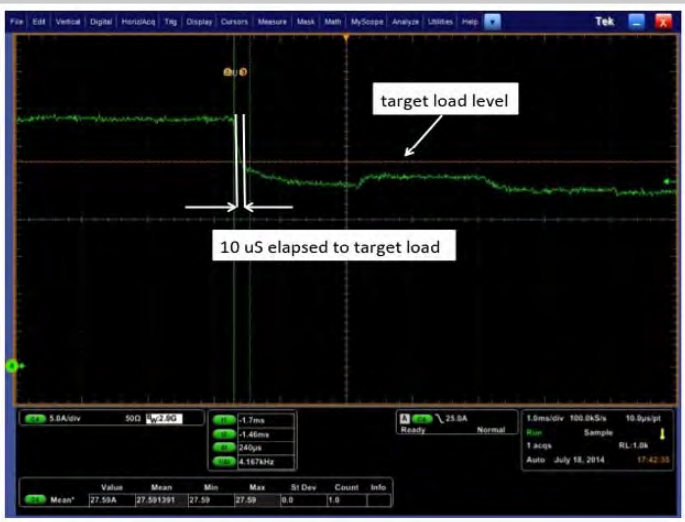


Figure 4.3.6: Rapid 10µs reaction to convertor phase loss with cycle skip followed by DVFS.

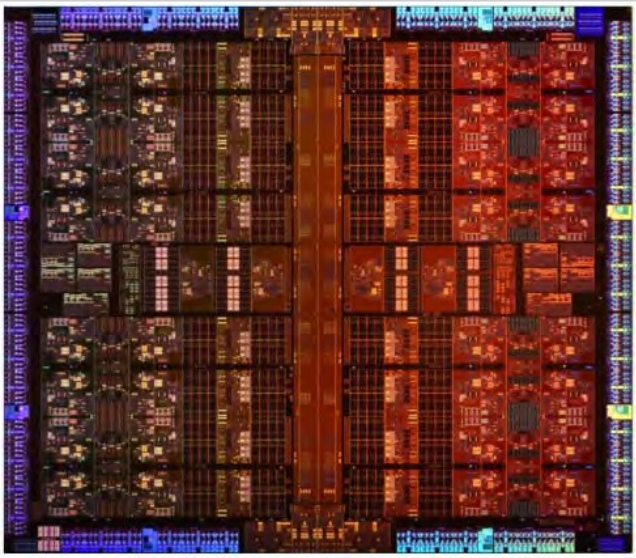


Figure 4.3.7: SPARC M7 die photograph.

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