4.4 Energy-Efficient Microserver Based on a 12-Core 1.8GHz 188K-CoreMark 28nm Bulk CMOS 64b SoC for Big-Data Applications with 159GB/s/L Memory Bandwidth System Density

Ronald Luijten[1], Dac Pham[2], Rolf Clauberg[1], Matteo Cossale[1], Huy N. Nguyen[2], Mihir Pandya[2]

[1]IBM Research, Rüschlikon, Switzerland,
[2]Freescale Semiconductor, Austin, TX

MicroServers integrate an entire server motherboard into a single Server-on-a-Chip (SoC), excluding DRAM, NOR-boot and power conversion circuits. This technology has evolved to 64b processing able to run server-class operating systems (OS), and the newest SoCs also target cloud computing and business SW [1]. The DOME µServer [2] packages the T4240 SoC in a dense node.

Figure 4.4.7(a) shows the 239mm^2 T4240 die with ~1.7B transistors including 111M bits of SRAM and 6M flipflops, implemented in 28nm bulk technology with 10 levels of copper interconnect. It has 3 clusters each with 4 e6500 cores. This core is 64b Power Architecture™ compliant and is fully dual-threaded with simultaneous multi-threading. It operates up to 1.8GHz and achieves 8.7CoreMark/MHz in dual-thread mode. Each core shares one floating-point and one Altivec SIMD unit. CoreNet is a low-latency coherent fabric connecting cores, memory and IO, operating at platform frequency and providing greater than 50GB/s bandwidth (BW). CoreNet eliminates bus contention and provides concurrent multicore connectivity capable of scaling up to 32 cores [3]. Each core furthermore implements state-retention power-gating (SRPG) registers and memories to enable quick wake-ups from power-down modes and the power management unit in each cluster can independently and automatically power down each core and the Altivec SIMD unit within each core.

Each cluster has 7 reduced power states including 2 power-gating states. While in the power-gating states, the always-on V_{DD} domain supports SRPG on all registers and memories, and the gated V_{DD} domain can zero the supply voltage for the majority of logic gates. During operation, the SoC power-management subsystem can initiate any combination of cores to enter or exit the gated mode while the rest of logic is running. This can cause a sudden change in current, thereby affecting the integrity of the power supply and the gate-oxide reliability due to voltage overshoot. The trickle and header switches, shown in Fig. 4.4.1, are staged into multiple chains to manage the rush-in current and maintain power grid stability for continuous correct operation. High V_t switches are used because of R_{on} and leakage requirements. The design has 3 trickle chains and 5 header chains with a total of 93K switches occupying ~5% of core area. It enables the gated V_{DD} domain to reach 99.8% of the nominal 1.0V within 200 core cycles, while achieving less than 25mV droop in RLC response, well within the design margin, for the worst-case wake-up condition. The implementation has been fully validated and functionally demonstrated on first-pass silicon at frequencies over 1.8GHz as shown in Fig. 4.4.2.

In early 2014, we started building a 139mm×55mm×7.6mm compute node board (Fig. 4.4.7(b)) employing the T4240 SoC for the DOME µServer project [2]. This SoC meets our µServer definition, 64b architecture and server class OS requirements. Our goal is to build a '*data center in a box*' demonstrator using commodity components and industry standards at high density, high energy efficiency and low cost. Our node board also contains a Cypress PSOC5™ for management, booting and thermal-monitoring functions, a Volterra VT1175MA / VT1198SA power converter for core T4240 power and three channels of DDR3 72b DRAM totaling 48GB. The system connector carries four 10GbE, two SATA, USB and the 1.35V DRAM, 3.3V and 1.8V power supplies. We achieve high system density through the use of indirect hot-water cooling, where the cooling infrastructure is also used to transport electrical power [2]. We remove the T4240 package heat spreader (lid, Fig. 4.4.7(b)) and replace it with our own copper plate, which additionally supplies 12VDC power as shown in Fig. 4.4.3. Our heat spreader uses 2mm OF R240 medium-hard Cu for the thick layer acting as the main thermal path and 0.2mm GOULD copper JTC (Grade 1) for the other 2 layers used for power contact and electromagnetic shielding. These layers are laminated together in a standard PCB fabrication process using the Arlon 92ML-type 106 0.1mm dielectric, a prepreg with enhanced thermal conductivity. We have built a thermal test vehicle (Fig. 4.4.4) and performed thermal characterization of our µServer cooling structure, using a Si test chip only containing heating resistors and die-temperature-measurement diodes. We measured a thermal package resistance of 1.11K/W and a thermal heat output of 36W can be removed with Tj=85°C with 45°C water. We also characterized the use of heat pipes embedded in our heat spreader to further reduce the thermal resistance [4]. First results show that the thermal resistance improves to better than 0.85K/W, and a heat output of 47W can be removed at Tj=85°C, or 36W can be removed with Tj=75°C with 45°C water (see Fig. 4.4.3).

We ran performance benchmarks on a T4240RDB-PB, a 1U rack unit with a single T4240 SoC running at 1.667GHz core clock and 1.867GT/s DDR3 DRAM. On this platform, we have successfully tested Linux Fedora 20 OS (F20), Specbench and CPMD. On a 4-way T4240RDB cluster, we also tested the latest IBM DB2 DPF code (v10.5FP3a). Compilation was not needed for DB2, underlining the suitability of the T4240 SoC for Linux-on-Power. The 4 RDBs, each with a local SATA root file system, are connected in a star topology with 1 GbE to an Ethernet switch, and as a functional test we successfully demonstrated the DB2 *Workload Multiuser Driver* with 1100 clients. Fig. 4.4.5 compares the integer ratios of the *Specbench* and *CoreMark* benchmarks against an Intel® Xeon® E3-1230Lv3 processor running at a comparable core clock rate of 1.8GHz. The analysis is performed under Fedora-20/19 (T4240/E3 respectively) using the standard gcc compiler. Benchmark test conditions are shown in Fig. 4.4.5. The T4240 shows 3× lower single thread performance, but 41% more aggregate performance than the E3-1230Lv3.

We target Big-Data applications with our T4240-based µServer system. Our planned 2U rack unit shown in Fig. 4.4.6 contains 128 compute nodes with 48GB DRAM each, yielding an aggregate 3072 HW threads and 6TB of DRAM. A compute node runs at 1.8GHz core as well as DRAM clock rate, yielding a total of 43.2GB/s peak memory BW per node. The volume of our rack unit (427×86.5×940mm^3) will be 34.7L. For Big-Data applications, the single-thread performance is of secondary importance, and node memory capacity and BW are more relevant. We introduce a new metric: *memory BW per liter* for aggregate DRAM bandwidth between compute nodes and main memory within a server enclosure. This metric is not sufficient on its own and must be considered together with memory capacity (6TB/rack unit) and memory per compute thread (2GB). Our 2U rack unit achieves 159GB/s/L (128 nodes at 43.2GB/s in 34.7L). The IBM Power System S822L (2 P8 CPUs), a server with currently the highest memory BW per socket, achieves 13.9GB/s/L (384GB/s in 427×86.5×747mm^3). Note that for our µServer the number is peak BW, whereas for the P8 system it is sustained BW.

At 1.8GHz and Tj=85°C, the T4240 SoC consumes 26W (Fig. 4.4.2) and at full SpecBench load our compute node is 41% faster than the Xeon E3-1230Lv3. Our compute node is expected to consume 36W, which is 70% of the Xeon compute node board. Our 28nm CMOS SoC-based µServer node has more than 2× better energy efficiency than the 22nm FinFET Xeon-based node, showing that system-level design choices can outweigh the advantage of a newer CMOS process. This result validates our µServer approach.

The authors acknowledge all contributions from IBM and Freescale on the system and the design of this SoC, and the Dutch government grant for DOME.

References:

[1] A. Yeung, *et al.*, "A 3GHz 64b ARM v8 Processor in 40nm Bulk CMOS Technology," *ISSCC Dig. Tech. Papers*, pp. 110-111, Feb. 2014.
[2] R. Luijten, *et al.*, "Dual Function Heat-Spreading and Performance of the IBM / Astron DOME 64-bit µServer Demonstrator," *IEEE International Conf. on IC Design and Tech.*, 2014.
[3] D. Pham, *et al.*, "Embedded Multicore Systems: Design Challenges and Opportunities," *Multiprocessor System-on-Chip: Hardware Design and Tool Integration*, Springer, pp. 197-222, 2011.
[4] S. Zimmermann, *et al.*, "Aquasar: A Hot Water Cooled Data Center with Direct Energy Reuse," *Energy*, vol. 43, no.1, pp. 237-245, 2012.

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

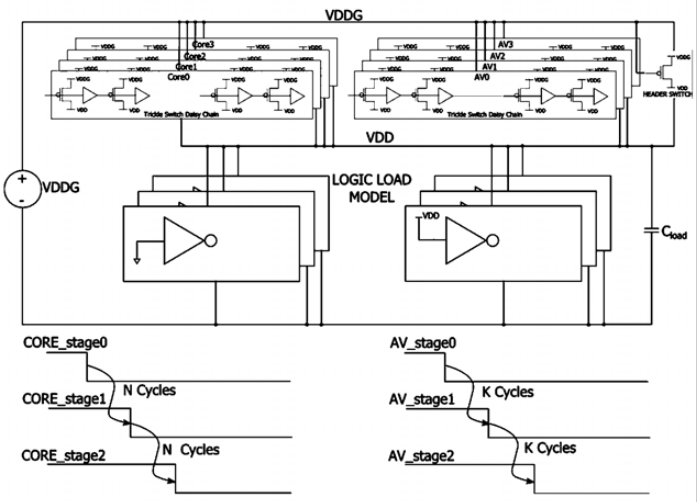


Figure 4.4.1: Power switch configuration.

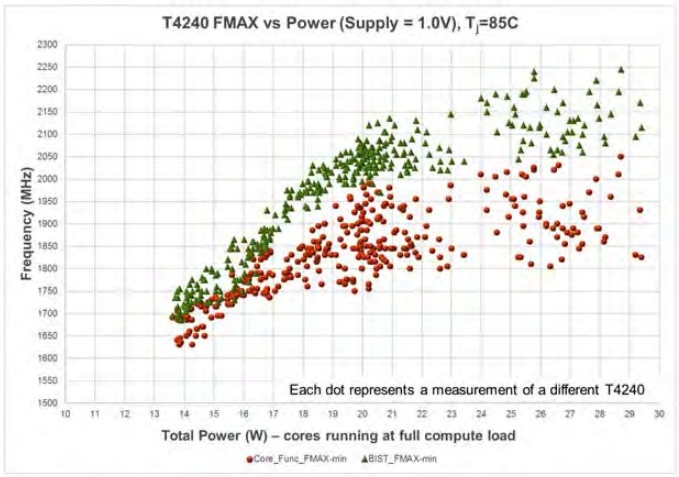
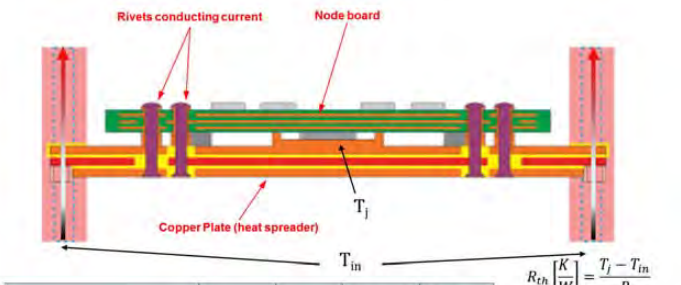


Figure 4.4.2: T4240 frequency vs. power hardware characterization.



Cooling variant	Inlet water temp [C]	Junction temp T_j [C]	Measured thermal resistance [K/W]	Maximum cooling capacity [W]
OF R240 Cu, no heat pipe	45	85	1.11	36
OF R240d Cu with heat pipe	45	85	0.85	47
OF R240d Cu with heat pipe	45	75	0.85	36

Figure 4.4.3: Compute node power and cooling and measurement results.

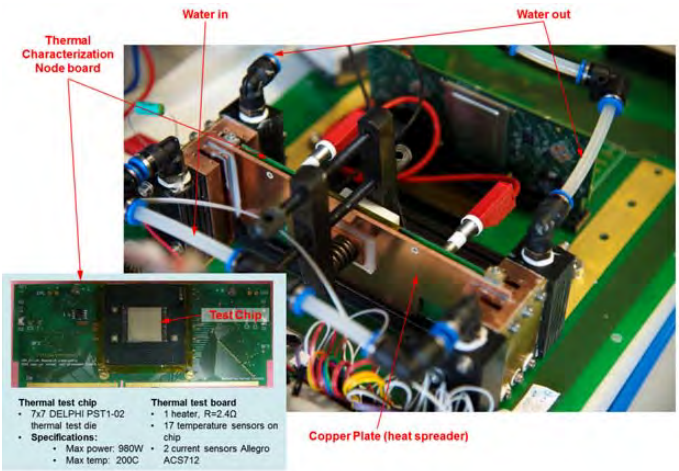


Figure 4.4.4: Thermal test vehicle lab setup.

Processor	Freescale **T4240** 12 cores; 24 threads, 28nm Bulk	Intel Xeon **E3-1230L v3** 4 cores; 8 threads, 22nm FinFET
Benchmark Test System and conditions	System: **Freescale T4240RDB-PB** **1.67 GHz core clock,** **1.87 GT/s 12GB DRAM in 3 channels** Fedora 20, Kernel 3.12.19 GCC 4.7.2 gcc compiler options: -O3 -mcpu=powerpc64	System: **Supermicro X10SAE** **1.8 GHz core clock; Turbo disabled** **1.666 GT/s 8 GB DRAM in 2 channels** Fedora 19, Kernel 3.13.9 GCC 4.8.2 gcc compiler options: -O3 -march=native -mtune=native
CINT-base – **1** thread	6.86	20.7
CINT-base – **all** threads	109.34 (24 threads)	77.6 (8 threads)
Coremark – **all** threads	188K (24 threads)	65K (8 threads)

Figure 4.4.5: T4240 performance comparison with Intel 1.8GHz Xeon.

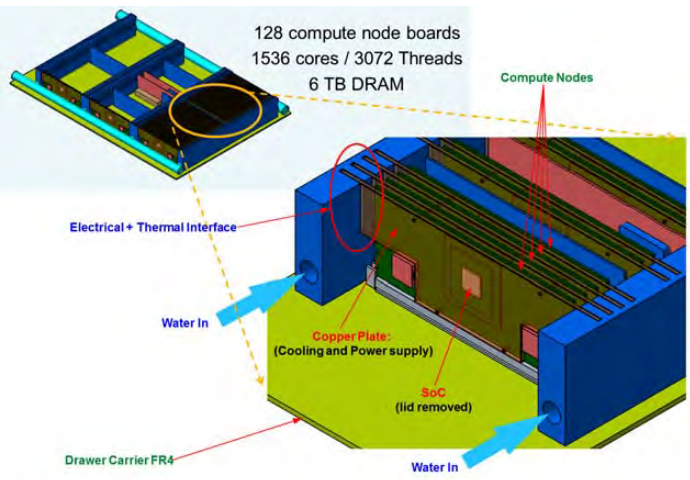


Figure 4.4.6: Rack unit, cooling and power.

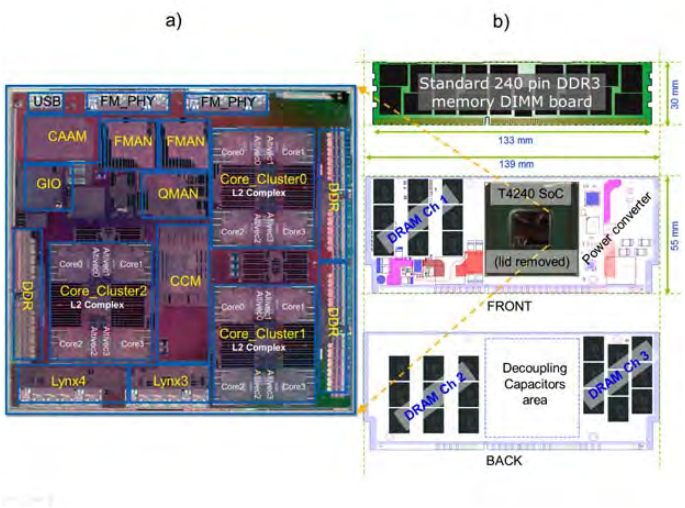


Figure 4.4.7: a) T4240 die pPhoto and b) Dome µServer compute node form factor.

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