How Good Are Low-Power 64-bit SoCs for Server-Class Workloads?

Reza Azimi
School of Engineering
Brown University
Providence, RI 02912
Email: reza_azimi@brown.edu

Xin Zhan
School of Engineering
Brown University
Providence, RI 02912
Email: xin_zhan@brown.edu

Sherief Reda
School of Engineering
Brown University
Providence, RI 02912
Email: sherief_reda@brown.edu

Abstract—**Emerging system-on-a-chip (SoC)-based microservers promise higher energy efficiency by drastically reducing power consumption albeit at the expense of loss in performance. In this paper we thoroughly evaluate the performance and energy efficiency of two 64-bit eight-core ARM and x86 SoCs on a number of parallel scale-out benchmarks and high-performance computing benchmarks. We characterize the workloads on these servers and elaborate the impact of the SoC architecture, memory hierarchy, and system design on the performance and energy efficiency outcomes. We also contrast the results against those of standard x86 servers.**

I. Introduction

Given the pressing needs for server energy efficiency, an emerging trend advocates using processor or System-on-a-Chip (SoC) architectures that mimic those of mobile systems, because mobile systems were highly designed and optimized for strict power limitations [1]. While the individual processor core of mobile systems offers much less performance than their traditional server counterparts, it offers orders of magnitude reduction in power consumption, leading to an overall improved energy efficiency. Since an individual mobile core is weaker in performance than a server core, it is imperative that servers designed with mobile cores end up using a larger number of cores to deliver the same performance as traditional servers. Thus, scale-out applications with high degree of parallelism are expected to leverage mobile-based parallel architectures to deliver improved energy efficiency.

Our goal is to better understand the strengths and weaknesses of emerging SoC-based microservers for server-class applications. We consider two 8-core 64-bit microservers based on x86 and ARM SoCs. We evaluate the microservers by stressing various architectural features of the processors with selected workloads. We show that the microservers have different strengths for various features. The better branch predictor of the ARM server leads to improved performance for workloads dominated by control flow for instruction fetching, but the improved cache organization of the Atom processor leads to better performance for workloads dominated by data memory accesses. For high-performance computing (HPC) workloads with abundant instruction-level parallelism, the two microservers are dominated by traditional servers.

II. Infrastructure

We consider two microservers, one fitted with an Atom C2750 SoC and the other fitted with a XGene-1 SoC. Each SoC features 8 cores that can run to a maximum of 2.40 GHz. Both processors have the same L1 data and instruction cache sizes, but have considerable differences in L2 and L3. The Atom C2750 SoC has large L2 caches but no L3 cache, while the XGene-1 SoC has smaller L2 caches and a large shared L3 cache [3]. The biggest difference between the two SoCs is their manufacturing technology. The XGene-1 SoC is manufactured with planar 40 nm transistor technology, while the Atom C2750 SoC is manufactured using 22 nm FinFet technology, which enables the Atom SoC to have ideally 3.6× the number of transistor per unit area and to have lower power consumption. A server with Xeon E5-2630 V3 processor running at 2.4 GHz is used as a reference to represent the performance of standard servers. All servers have 16 GB of DRAM in two channels with almost equivalent speed. We fully instrumented all servers to measure performance counters, network traffic statistics, and total power consumption.

III. Methodology and Results

To evaluate the performance of each SoC, we consider a few representative benchmarks that stress various architectural features of the processors.

− **Branch prediction performance.** We use the memcached application to stress the branch predictors given its control-dominant characteristics [2]. Table I gives the 95^{th} percentile latency for 150K request per seconds (rps). The results show that the ARM microserver is able to deliver improved latency compared to the Atom SoC by an average of 46%. To analyze the reason for this improved performance, the branch misprediction rate is given in Table II. The table shows that the Atom microserver suffers from large branch mispredictions that impact its branch fetching capabilities, which we believe is the reason for its worse performance.

Conclusion #1. Despite there is still a large performance gap between microservers and traditional datacenter servers for memcached, the microservers can offer acceptable performance for lower request rate (rps $<$ 100K) with good energy efficiency: 28% power saving by XGene-1 and 53%

	memcached rps:150K		FT: 8 threads			CG: 8 threads			SPECpower_ssj 100% utilization		
servers	response time (ms)	power (W)	normalized runtime	power (W)	normalized energy	normalized runtime	power (W)	normalized energy	ssj per second	power (W)	ssj per Watt
XGene-1	8.62	57	2.65×	80	1.96×	4.53×	74	3.22×	81481	66	1776
Atom C2750	16.11	37	2.62×	41	1.00×	1.76×	43	0.73×	107735	41	3634
Xeon E5-2630	2.20	79	1×	108	1×	1×	104	1×	295380	107	4066

Table I. Performance metrics of each workload at highest utilization.

	memcached rps:150K			FT: 8 threads			CG: 8 threads			SPECpower_ssj 100% utilization		
servers	IPC	branch miss rate	L2 miss per K instr	IPC	branch miss rate	L2 miss per K instr	IPC	branch miss rate	L2 miss per K instr	IPC	branch miss rate	L2 miss per K instr
XGene-1	0.265	1.5%	19	0.43	0.4%	37	0.12	6.6%	114	0.312	2.5%	21
Atom C2750	0.258	9.5%	12	0.77	2.9%	6	0.43	2.3%	24	0.518	4.5%	16
Xeon E5-2630	0.616	3.3%	24	2.04	0.2%	19	0.99	0.9%	42	0.866	1.1%	22

Table II. Performance analysis based on performance counter results.

power saving by Atom C2750 compared to the evaluated Xeon server. Given that servers are usually over provisioned in datacenters, using microservers is a promising way to achieve energy efficiency.

– **Arithmetic performance.** We select the CPU-intensive benchmark FT from the NAS Parallel benchmark (NPB) suite to stress the arithmetic units, especially that FT has both good instruction-level parallelism (ILP) and good thread-level parallelism. Normalized to Xeon E5-2630, the runtime of FT on each server is reported in Table I. The runtime results show large deterioration, 2.6×, in performance in comparison to the standard Xeon server. The IPC results reported in Table II demonstrate that the microservers do not have architectural resources to execute multiple instructions in parallel as good as traditional high performance computing servers. However, in term of energy efficiency, as shown in Table I, the normalized energy of Atom server is equal to the Xeon server.

Conclusion #2. Microservers seem to be dominated by traditional servers for HPC workloads with abundant ILP because they offer less performance at the same energy.

– **Memory subsystem performance.** We use CG from NPB benchmarks to stress the memory subsystem of the servers. It is configured with 8 threads to fully utilize the processor. Table I gives the normalized runtime of CG for our two microservers and the reference Xeon server. It shows that the Atom system delivers superior performance compared to the XGene-1. We believe this performance is attributed to the better memory organization of the Atom SoC. To understand the memory behavior of the microserver, we report the number of L2 cache misses per kilo-instruction of each server in Table II. The Atom processor has much lower misses compared to XGene-1 because of its larger distributed L2 caches.

Conclusion #3. For memory-bound benchmarks, the larger L2 caches of the Atom processor is better a choice than the smaller L2 caches and larger, high-latency L3 cache of the XGene-1 processor.

– **Energy Efficiency & Proportionality.** To evaluate power efficiency and energy proportionality, we use the SPECpower_ssj benchmark, which is a scalable, multi-threaded benchmark that represents server-side Java business applications. Table I gives the total number of ssj operations per second at 100% utilization for the two microservers and the reference Xeon server. It shows that the Atom system delivers improved performance than XGene-1 by about 32%. When comparing the energy efficiency, the Atom microserver delivers 2.1× more ssj operations per Watt than the XGene-1 processor. However, the Atom C2750 processor is fabricated using 22 nm technology, while XGene-1 is fabricated using 40 nm technology.

We observe that the base power of the motherboards constitute a significant portion of the total power as the idle power is about 45 W for the XGene-1 microserver, while it is about 26 W for the Atom microserver. Overall, both microservers offer a limited dynamic power range with non-ideal energy proportionality because the power consumption of the remaining components of the microservers are relatively high and non scalable.

Finally we compare some of the latest standard server results reported at spec.org against the SoC-based microservers. We consider a reported server with two Xeon E5-2699 processors with 11,284 ssj/W. If we compare the price, the ssj performance and the power of Atom C2750 microserver and the aforementioned server, we find that the two Xeon E5-2699 processors cost the same as 51 C2750 SoCs. Thus, the total scale-out performance of 51 Atom SoCs is equal to 5,494,485 ssj operations with a power of $20\times51 = 1020$ W. Thus, the two Xeons offer 11,284 ssj/W and the 51 C2750 SoCs offer 5,386 ssj/W. This calculation is a best case scenario for C2750-based microservers because it does not take into account the additional power consumption of the microserver boards.

Conclusion #4. High-end traditional servers deliver improved throughput per Watt since they pack on a single server board two or more many-core processors, which amortize the power of other components on the motherboard.

Acknowledgment: This project is supported by NSF grant 1305148.

References

[1] E. Blem, l. Menon, and K. Sankaralingarn, “Power struggles: Revisiting the RISC vs. CISC debate on contemporary ARM and x86 architectures,” in *ISCA*, 2013, pp. 1–12.

[2] K. Lim, *et al.*, “Thin servers with smart pipes: designing SoC accelerators for memcached,” in *ISCA*, 2013, pp. 36–47.

[3] A. Yeung *et al.*, “A 3GHz 64b ARM v8 processor in 40nm bulk CMOS technology,” in *ISSCC*, 2014, pp. 110–112.