POSTER: NUMA-aware Power Management for Chip Multiprocessors

Changmin Ahn, Camilo. A. Celis Guzman, and Bernhard Egger
Department of Computer Science and Engineering
Seoul National University, Seoul, Korea
{*changmin, camilo, bernhard*}*@csap.snu.ac.kr*

***Abstract*—Traditional approaches for cache-coherent shared-memory architectures running symmetric multiprocessing (SMP) operating systems are not adequate for future many-core chips where power management presents one of the most important challenges. In this work, we present a power management framework for many-core systems that does not require coherent shared memory and supports multiple-voltage/multiple-frequency (MVMF) architectures. A hierarchical NUMA-aware power management technique combines dynamic voltage and frequency scaling (DVFS) with workload migration. The conflicting goals of grouping workloads with similar utilization patterns and placing workloads as close as possible to their data are considered by a greedy placement algorithm. Implemented in software and evaluated on existing hardware, the proposed technique achieves a 30 and 8 percent improvement in performance-per-watt compared to DVFS-only and NUMA-unaware power management.**

***Keywords*-Power management, CMP, MVMF, NUMA**

I. Introduction

The past decade has brought a shift from single- or dual-core processors to chip multiprocessors (CMPs) integrating from a few tens up to a thousand cores into one processor die [1]–[3]. Chip-level power and thermal constraints are now one of the primary design constraints and performance limiters. Higher power consumption not only leads to increased energy cost but also causes higher die temperatures that adversely affect chip reliability and lifetime.

Processors support dynamic voltage and frequency scaling (DVFS) to allow at-runtime adjustments of the supply voltage and frequency in order to reduce power consumption. On CMPs, the required logic for individually controlling the voltage for each core is becoming too costly; instead, cores are physically clustered into voltage and frequency domains leading to so-called *multiple-voltage/multiple-frequency* (MVMF) designs where all cores within a domain run at the same voltage or frequency.

Managing energy consumption effectively on MVMF architectures remains a challenge. First, instead of individual cores, DVFS has to be applied to groups of cores. This requires grouping of workloads with similar performance requirements into the same group for maximal efficiency and has been address by previous work [4]. Second, the varying distance of the cores to the memory controller(s) of CMPs with tens of cores causes varying memory access latencies. To achieve good performance, workloads with a high memory affinity should be located close to the accessed memory controllers. Third, more and more recent CMPs drop support for coherent global shared memory [1]–[3].

In this work, we propose a scalable hierarchical power management technique for modern MVMF CMPs. The technique considers both the NUMA properties of the chip and the restrictions imposed by the MVMF design. The potentially conflicting goals of NUMA affinity and grouping similar workloads are resolved by a greedy core allocation algorithm. The technique can be applied to monolithic kernels running on a cache-coherent SMP processor as well as non-coherent memory architectures running distributed kernels. A working implementation is evaluated on real hardware [1] with real-world workloads [5]. We compare the technique to a DVFS-only approach [6] and prior work that only considered the MVMF limitations [7]. On average, we achieve a 55, 30, and 8% higher performance-per-watt efficiency over no power management, DVFS-only and a NUMA-unaware approach at no performance degradation.

II. Cooperative Power Management

The goal of the proposed power management technique is to achieve the best possible power efficiency measured as power-per-watt (PPW) without sacrificing performance. To apply DVFS effectively, workloads with similar performance requirements should be placed in the same voltage/frequency domains. At the same time, memory-bound workloads experience a significant performance loss in dependence on the location on the chip; we measured up to 33% of performance loss at a fixed frequency caused by NUMA characteristics on our hardware. Both goals require relocation of workloads on the chip, i.e., assigning workloads to cores that best match the goal of best possible power efficiency.

The properties of workloads are profiled at runtime by querying the cores' performance counters. In order to remain scalable, performance monitoring, frequency scaling, and voltage regulation are performed in a distributed manner. Performance is monitored at each core, one core on each frequency domain sets the frequency, and one core on each voltage domain is responsible to regulate the voltage. These core, frequency, and voltage controllers form a hierarchical structure resembling that of the MVMF architecture. Information about the workloads running on the individual cores is sent up in the hierarchy to the top-level controller, the chip

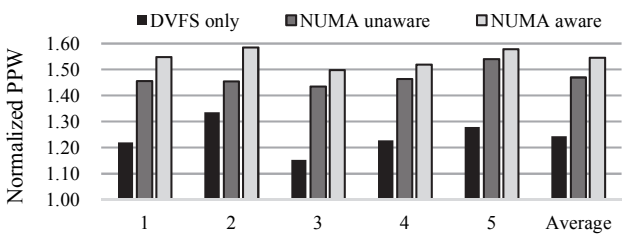

Figure 1. Comparison of the proposed technique, `NUMA aware`, against `DVFS only` [6] and a `NUMA unaware` [7] approach.

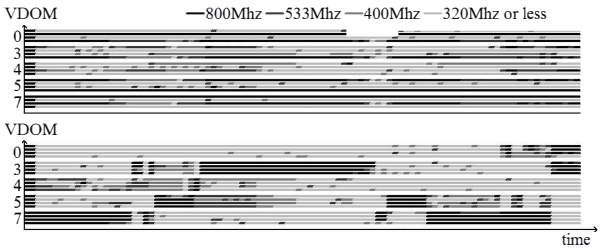

Figure 2. Operating frequency per frequency domain of `DVFS only` (upper half) versus the proposed `NUMA aware` (lower half).

controller where aggregated information is used to compute a better workload distribution for the next epoch.

Crucial for the success of the proposed method is the computation of the workload distribution. Workload migration and DVFS come with some overhead. The former is similar to a process switch between physical cores on SMP systems; performance degradation is caused by the process interruption and cold cache misses. Changing the voltage requires all cores in the affected domain to be stopped until the new voltage has stabilized. A migration algorithm has to balance these overheads against potential gains. Starting with the highest supply voltage v, the proposed greedy algorithm minimizes the number of voltages domains running at v by first assigning all cores requiring this level of computational power in order not to incur a performance loss. For each domain, the total distance of all workloads to their memory controllers is computed and weighted with the workloads' memory affinity. This process is repeated for each voltage level down to the minimal supply voltage. For a potential new workload allocation, the estimated benefit considers both the power savings and the performance degradation caused by the migration.

III. Evaluation

The proposed NUMA-aware power management technique is implemented and evaluated on the 48-core Intel Single Chip Cloud Computer (SCC) [1]. The core, frequency, voltage, and chip controllers all run on the chip and their overhead in terms of additional computational load or power consumption are included in the power measurements. Also included are all overheads caused by workload migration. We assign the workload of a physical machine observed in a Google data center [5] to one core of the SCC. One evaluation run comprises of 40 randomly selected distinct workloads that are assigned to 40 cores of the SCC.

Figure 1 shows the PPW of five distinct evaluation runs normalized to no power management. Compared to the DVFS-only [6] and a NUMA-unaware [7] approach, the proposed NUMA-aware technique achieves a 30 and 8 percent better PPW. Figure 2 visualizes the effect of workload migration. With only DVFS, high performance workloads remain distributed to the different voltage domains, forcing them to run at maximal supply voltage. Migration groups workloads with similar performance characteristics into domains, allowing us to run the domains at a more optimal frequencies without sacrificing performance.

Acknowledgments

This work was supported in part by BK21 Plus for Pioneers in Innovative Computing(Dept. of Computer Science and Engineering, SNU) funded by the National Research Foundation of Korea (NRF) (21A20151113068), the Basic Science Research Program through NRF funded by the Ministry of Science, ICT & Future Planning (Grant NRF-2015K1A3A1A14021288), and by the Promising-Pioneering Researcher Program through Seoul National University in 2015. ICT at Seoul National University provided research facilities for this study.

References

[1] J. Howard et al., "A 48-core IA-32 Message-Passing Processor with DVFS in 45nm CMOS," in *2010 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2010.

[2] A. Olofsson, "Epiphany-V: A 1024 processor 64-bit RISC System-On-Chip," https://arxiv.org/abs/1610.01832, Oct 2016.

[3] B. Bohnenstiehl, A. Stillmaker, J. J. Pimentel, T. Andreas, B. Liu, A. T. Tran, E. Adeagbo, and B. M. Baas, "KiloCore: A 32-nm 1000-Processor Computational Array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 891–902, April 2017.

[4] K. K. Rangan, G.-Y. Wei, and D. Brooks, "Thread Motion: Fine-grained Power Management for Multi-core Systems," in *Proceedings of the 36th Annual International Symposium on Computer Architecture (ISCA'09)*, June 2009.

[5] J. Wilkes, "More Google Cluster Data," November 2011, posted at http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html.

[6] N. Ioannou, M. Kauschke, M. Gries, and M. Cintra, "Phase-Based Application-Driven Hierarchical Power Management on the Single-chip Cloud Computer," in *Proceedings of the 2011 International Conference on Parallel Architectures and Compilation Techniques (PACT'11)*, October 2011.

[7] C. Kang, S. Lee, Y.-J. Lee, J. Lee, and B. Egger, "Scheduling for Better Energy Efficiency onMany-Core Chips," *LNCS 10353: Job Scheduling Strategies for Parallel Processing: 19th and 20th International Workshops, JSSPP 2015, Hyderabad, India, May 26, 2015 and JSSPP 2016, Chicago, IL, USA, May 27, 2016, Revised Selected Papers*, pp. 46–68, July 2017.