早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

A Study on Hierarchical Cache System Control based on Access Pattern Analysis for Chip Multiprocessor Systems

申 請 者
Huatao ZHAO

情報生産システム工学専攻
ASIC 自動設計研究

2020 年 7 月

Multilevel cache hierarchy is used to buffer the huge gap on processing speed between on-chip multi-core processor level and off-chip large-scale memory level. As the number of cores integrated on a single chip die recently tends to be dozens or even hundreds, the cache hierarchy loses its ability to cover over the speed gap and fails to optimally interconnect across on-chip components. More seriously, stacked multi-layer systems which have high integration density require a cache hierarchy with higher throughput to fully meet cache access demands among many concurrent threads, and also require more efficient method to guarantee data coherence in cache hierarchy.

To meet the throughput requirement in cache hierarchy, multi-level and private-shared cache structures are used in the recent chip multiprocessor systems (CMPs). Those cache hierarchies can even take a half of overall on-chip area and energy consumption of CMPs. However, many components of a cache hierarchy rarely contribute access hits in the most of execution time, but they waste too much energy for standby. Moreover, shared data which serve to concurrent threads are existing in a cache hierarchy, and coherence maintenances on those data waste too many clock cycles. As a consequence, the current cache hierarchy induces serious issues as follows: (1) Power issue on misallocating cache resources and (2) Data sharing issue. In the issue (1), On-chip caches suffer from high energy consumption overhead and a chip area overhead while such cache hierarchy is failed to satisfy cache resource demands of many threads in a large scope. In the issue (2), cache accesses to shared data among concurrent threads cause extremely expensive coherence maintenance.

The power issue (1) is caused by two reasons. First, increasing of cores requires a large scale of cache hierarchy for ensuring enough spare resources. Second, a demand for cache resource during runtime is locally changed along with processing at cores, which leads to allocation inequality that some threads tend to be in rush traffic but some threads are uncrowded with redundant cache resources. Rawlins, M. [IEEE T COMPUT, 2013] and Chen, G. [Microproc.Microsyst., 2016] proposed cache tuning based methods to explore optimal allocations on cache resources for each length-fixed interval of instructions, where during runtime, the energy-lowest cache size is explored for the next intervals once behavior changed (i.e., miss rate changed). Adegbija, T. [IEEE T VLSI SYST, 2018] and Wei, W [ACM T ARCHIT CODE OP, 2017] proposed a phase based exploration method on cache resources to save access energy, where firstly a number of phases are classified for each application, and then the optimal phase is explored if miss rate is larger than threshold. However, a demand for the cache resource in any executing period and in any thread could not be optimally satisfied with allocated cache resources appropriately. Thus, the first objective of this thesis is to dynamically allocate

cache resources to meet each demand for cache resource for each thread in any executing period, in other words, tune cache bank supply to concurrent threads' demand dynamically in intervals of selected subroutine calls, thereby saving energy by making the utmost of cache utilization.

The data sharing issue (2) is highly related to the fact that some kinds of cache access patterns (i.e., write accesses to shared data) cause expensive maintaining operations among many cores. Shared accesses generated in many concurrent threads may result in serious data inconsistency, and crossed accesses whose target data are existing in other threads will lead to access misses. The access pattern analysis shows that distributions on those harmful patterns can possess a considerable percent of total accesses. Lotfikamran, P. [IEEE HPCA, 2017] proposed a proactive resource allocation method to improve system performance based on shared data traffic profiling, which firstly predict that hot threads require more resources, and then allocate required resources to the threads in each time interval, thereby reducing router stall time and improving performance. Gupta, S. [ICPP, 2015] proposed a spatial locality-based cache partitioning method, which firstly exploits spatial locality in partitioned shared cache, and then, for memory-intensive thread, increases its block size to enlarge shared data re-usage, and save some capacity to other threads. However, it is still difficult in conventional access paths to detect and deliver shared data, as those paths waste plenty of clock cycles to handle harmful cache accesses. Thus, the second objective of this thesis is to efficiently handle those harmful cache accesses in the proposed concurrent path, which acts as a shortcut path on data sharing accesses among private caches to detect and route shared data in advance.

This thesis is organized as follows:

Chapter 1 [Introduction] introduces the research background of cache hierarchy designs and previous works on the cache optimization, and then describes the outline of the proposed methods.

Chapter 2 [Access Pattern Analysis] represents the detailed experiments on cache access patterns and statistically classifies cache access distributions. Then, those patterns are analyzed to reveal the internal relationships among cache resource demands, locality features and access distributions.

Chapter 3 [Controllable Cache Resource Allocation] proposes a low-power cache hierarchy scheme applying the control theory to give a hardware-based solution for an optimal cache resource allocation in some interval granularity. Firstly, an effective bank allocation policy is proposed for adaptive cache resource allocation. Secondly, preferable intervals which mean proper timing to change resource

allocation are designed in fine granularity of per-subroutine. Finally, the controllable cache resource allocation policy is proposed combining the discrete control theory by the PID based controller and cache resource allocation at subroutine-based interval. Experimental results using SPEC benchmark data and Gem5 simulator show that energy consumption on shared cache can be saved by 39.7 % on average compared with the conventional equi-interval method, and saved by 11.6% and 18.2% compared with Chen's method [Microproc. Microsyst., 2016] and Adegbija's method [IEEE T VLSI SYST., 2018], respectively.

Chapter 4 [Stacked 3D On-chip Cache Network] proposes a stacked 3D three-layer on-chip architecture consisted of enhanced global- and local- router networks. The router is improved in cache access detection, sharing data replacement and target data delivery functions. The proposed interaction path in the network can support fast shared data, in which "crossed read" can achieve target data by routing from other caches and both "shared write" and "crossed write" can directly update all copies in virtue of the routing network. Moreover, VLSI layout design of the proposed router architecture is implemented to verify the placement & routing details. And the on-chip design of a stacked 3D structure is evaluated from the viewpoint of thermal affection and estimated chip size. Simulation results indicate that the proposed router-integrated cache hierarchy design of a CMP system improves the system performance by 31.9% and on-chip energy by 17.6 %, compared with the base system without a cache network.

Chapter 5 [Conclusion and Future Work] sums up this thesis on achievements and contributions, that is, much energy savings is achieved by the proposed self-adapting cache resource allocation method, and both performance and energy consumption are improved by the proposed router-integrated cache optimization method. Finally, further optimizations on proposed designs are expected in future work.

As a consequence, this thesis represents optimization techniques on cache hierarchy including shared cache level and private cache level. For power issue on misallocating cache resources, a discrete PID based controller is integrated to form a self-adaptive cache allocation method in a novel granularity of per-subroutine based interval. For data sharing issue, a shortcut path on harmful accesses is designed to improve coherence-maintenance efficiency by integrating enhanced router networks. The experimental results show the substantial improvements on both performance and energy consumption.