CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Gyandhara International Academic Publication (GIAP): Journals

GIAP
International Journal of Students' Research in Technology & Management
Vol 3 (02), February 2014, ISSN 2321-2543, pg 264-268

3-D STACKED CACHE DATA MANAGEMENT FOR ENERGY MINIMIZATION OF 3-D CHIP-MULTIPROCESSOR

Mr. K. Suresh Kumar, Ms. S. Anitha, Mrs. M. Gayathri
M.E-VLSI Design Knowledge Institute of Technology Salem, India
Assistant Prof. Dept. Of ECE Knowledge Institute of Technology Salem, India
Assistant Prof. Dept. of ECE Knowledge Institute of Technology Salem, India
Sureshkumar4290@gmail.com
saece@kiot.ac.in
mg@kiot.ac.in

Abstract*- In a 3-D processor-memory system, multiple cache dies can be stacked onto multi-core die to reduce latency and power of the on-chip wires connecting the cores and the cache, which finally increases the power efficiency. However, there are two challenging issues. The principle is high power density (resulting from multiple die stacking) that incurs many temperature-related problems including temperature dependent leakage power and another is the processor-cache traffic congestions that occur at through-silicon vias (TSVs) shared by multiple stacked caches. In this model a runtime cache data mapping is discussed for 3-D stacked L2 caches to minimize the overall energy of 3-D chip multiprocessors (CMPs). The suggested method considers both temperature distribution and memory traffic of 3-D CMPs. Experimental result shows energy reduction achieving up to 22.88% compared to an existing solution which considers only the temperature distribution. New tendencies envisage 3D Multi-Processor System-On-Chip (MPSoC) design as a promising solution to keep increasing the performance of the next-generation high performance computing (HPC) systems. However, as the power density of HPC systems increases with the arrival of 3D MPSoCs with energy reduction achieving up to 19.55% by supplying electrical power to the computing equipment and constantly removing the generated heat is rapidly becoming the dominant cost in any HPC facility. Thus, both power and thermal/cooling implications play a major role in the design of new HPC systems, given the energy constraints in our society. Therefore, EPFL, IBM and ETHZ have been working within the CMOSAIC Nano-Tera.ch program project in the last three years on the development of a holistic thermally-aware design.

Keywords: MPSoC, HPC, CMP, Interconnect, TSV, Cache Management.

I. INTRODUCTION

Three-dimensional integrated circuits (3-D ICs), where two or more layers of active electronic components are integrated vertically into a single chip, significantly reduces the on chip wire length that often becomes a major bottleneck of performance and/or power dissipation in 2-D ICs. Particularly, 3-D memory stacking has received a great attention since it resolves the memory bandwidth challenges of 2-D ICs by stacking cache memory onto a multi-core die. However, the high power density resulting from multiple (memory) die stacking may lead to the temperature-related problems in reliability (e.g., NBTI), power, performance, and cooling cost. Especially, the exponential dependence of leakage power on temperature, in conjunction with the large amount of cache stacked onto a multi-core die, might aggravate the energy efficiency of 3-D processor-memory systems, when considering that on-chip SRAM cache often consumes almost half of total energy in a microprocessor system.

Dynamic cache reconfiguration (DCR) is an effective method to reduce cache energy by configuring capacity, line size, and associatively of cache according to workload characteristics, and turning off unused parts of the cache. For example, the amount of turned on cache blocks (i.e., capacity) can be optimally determined and assigned to each core based on the memory access demands of applications and, then, the unassigned cache blocks can be turned off to reduce the operating temperature and the temperature-induced leakage energy. However, excessive power gating of cache blocks may incur performance degradation due to the increase in cache misses [2].

Since cache banks directly stacked on a core share the same TSVs, traffic collision might occur even when cores access different cache banks if the cache banks are directly stacked on the same core (and share the same TSVs).

In this paper, we propose a dynamic cache reconfiguration (DCR) scheme that minimizes the energy consumption of 3- D CMPs with temperature and time-to-deadline constraints. Given the time-varying temperature profile of cores and L2 cache banks, the proposed solution determines the number of L2 cache banks (i.e., the amount

International Journal of Students' Research in Technology & Management
Vol 3 (02), February 2014, ISSN 2321-2543, pg 264-268

of cache capacity) logically allocated to each core and the physical placement of the allocated L2 cache banks, considering both temperature distribution and memory traffic of the 3-D CMPs. To the best of our knowledge, this is the first work on online DCR schemes for real-time 3-D CMPs that considers both temperature and memory traffic. Considering both temperature distribution and cache traffic congestion gives more energy reduction than considering only without the other. A key challenge in 3D memory stacking is the heat generated from the 3D chip with its increased power density. In case of memory-stacked CMP systems, temperature of each core directly affects the temperature of cache memory blocks stacked on the core. There are prior works on the temperature aware management for 3D CMP. Since workload characteristics such as memory access behavior change dramatically at runtime, online adaptive configuration of cache memory is paramount for energy reduction. We also investigate the impact of non-uniform cache access latency, cache traffic congestion, and temperature distribution on the energy consumption of 3- D CMPs.

II. WORKING PRINCIPLE

The basic idea of our method is to exploit a trade-off between the leakage energy induced by the higher temperature and the leakage energy owing to the longer execution time resulting from core's stall. Core 1 has lower operating temperature than Core 0 because of lower power consumption of Core 1. However, Core 1 running Gzip accesses its cache banks more frequently than Core 0 running Mesa does. Compared with TA-DCR (Temperature-Aware - Dynamic Cache Reconfiguration), TCA-DCR (Temperature- and Congestion-Aware - Dynamic Cache Reconfiguration) reduces memory traffic congestion at Core 1's TSVs and, thus, reduces the stall time of both cores by mapping Core 0's data onto its two local cache banks, instead of only one local cache bank, while sacrificing the temperature distribution that incurs more leakage energy induced by the higher temperature. TCA-DCR yields additional energy reduction of 5.2% compared with TADCR. Compared with DCR, TCA-DCR reduces the energy consumption by 13.8%. Each cache layer consists of two cache banks. The capacity of each cache bank is 256KB, assuming a cache bank has the same area/shape as that of a core. Cache banks directly stacked on a core local cache banks are connected to the core through shared TSVs. Cores are also able to access cache banks stacked on the other cores i.e., remote cache banks through the crossbar switch with longer latency. The core layer is located next to the heat sink. Let us assume that dynamic power gating is performed at the granularity of cache bank. Two threads, e.g., Mesa and Gzip in SPEC2000 benchmark, are mapped onto Core 0 and Core 1, respectively. The average power consumption of Mesa and Gzip are 4.76W and 1.91W, respectively. The numbers of L2 cache accesses per cycle of Mesa and Gzip are 0.0024 and 0.0391, respectively.

III. DESCRIPTION

A key challenge in 3D memory stacking is the heat generated from the 3D chip with its increased power density. In case of memory-stacked CMP systems, temperature of each core directly affects the temperature of cache memory blocks stacked on the core. There are prior works on the temperature aware management for 3D CMP. A dynamic voltage and frequency scaling scheme for 3D-stacked L2 DRAM with taking account of both DRAM error-rate and temperature-induced power consumption and a thermally aware thread migration among processor cores to reduce temperature variance and peak temperature of stacked DRAM is also proposed. To reduce energy consumption, heavily communicating tasks are allocated within the same vertical stack by taking account of shorter interconnect distance between vertical adjacent cores. To consider workload characteristics such as memory access behavior changes during runtime, we propose a run-time solution to minimize the system energy consumption (including the energy consumption of core, cache, and off-chip memory), using a multi-core system with stacked L2 cache as an example.

IV.DETAILED STUDY AND ANALYSIS

A. 3-D ON-CHIP INTERCONNECTION ARCHITECTURES

Network-on-chip (NoC) is considered to be one promising optionfor future CMP and SoC designs to mitigate the interconnect scaling problem. In the following, we give a brief introduction on the architectural designs for the combination of both 3-D integration and NoC, i.e., 3-D NoC. Please note that, in this work we only consider the processing element, which could be a core or a cache bank itself, is a 2-D planar design. Generally, the 3-DNoC architecture falls into the following three categories.

The simplest 3-D NoC design is to extend a 2-D router with two additional ports, up and down, within 3-D mesh topology. Although this architecture introduces the least modifications to the traditional design paradigm, it does not take advantage of the benefits of3-D integration. Vertical and horizontal data transmissions are indistinguishable as both of them bear identical characteristics as hop-by-hop traversal. More importantly, the symmetric NoC router design incurs significant area and power overhead.

To better explore the benefits of3-D integration, NoC-Bus hybrid architecture is proposed in which a low latency multi-drop shared bus is used to connect cores within the same stack. A centralized used to resolve the contention among cores, and cores within the same stack can be accessed in a single hop.

True 3-D router architecture is proposed in, in which all major components of a router, including crossbar, are partitioned into different layers such that the vertical link can be used more efficiently[1].

B. 3-D NoC Energy Model

The communication energy is mainly consumed by two parts, i.e., routers and links. In this work, we mainly explore the beneficial attribute in 3-D integration, that is vertical links consume much less power than horizontal links because of differences in wire lengths between neighboring cores. However, for symmetric 3-D NoC design, energy reduction for data transmission along the vertical dimension is not remarkable due to the per hop router energy. Therefore, the above latter two architectures are adopted as the target platforms in this paper. Note that the choice between the later two designs has no essential impact on the results of our methods.

Sun proposed a 3-D MPSoC thermal model to evaluate thermal impact of their task scheduling algorithm. Hung proposed a thermal-aware task allocation scheme using Hotspot to compute the temperature of the 3-D chip. Although their models are accurate, using them in the task allocation algorithm to predict thermal impacts will be very time-consuming, as the solution space is huge and a lot of trials need to be conducted. In this work, we assume tasks running for a sufficient longtime period, thus the temperature impact can be approximated by power consumption of the core running that task. For 3-DMPSoCs, as cores in the same stack have strong thermal correlation, we assume all the cores within the same stack have similar temperatures. Therefore, thermal evaluation is performed via power gradient computation in our algorithm [3].

In 3-D integration, except the bottom layer, other layers are thinned to only tens of micrometers for integration in contrast to hundreds of micrometers thickness of silicon substrate in the 2-D chip. Since the length of TSV is much smaller than that of the horizontal link (tens versus thousands of micrometers), for the same amount of data, traverse between vertical adjacent cores consumes much less energy than horizontal adjacent cores. Prior work on thermal-aware scheduling in 3-D MPSoC assumes tasks are independent from each other without any communications among them or generally ignores this beneficial property. The following example illustrates this potential in which both thermal and interconnect energy are taken into considerations [4].

Based on the above, communication energy can be reduced by restricting tasks within the same stack to take advantages of TSVs. Unfortunately, though this scheme is beneficial from the thermal problem. If tasks cluster in the same stack, the power density of this stack will increase sharply with a high possibility of generating hot spot. Thus, a tradeoff should be made between energy consumption and thermal dissipation which motivates the research of this paper.

The right hand side a homogeneous NoC-Bus hybrid 3-D MPSoC. The deterministic routing algorithm is adopted to avoid the live lock and dead lock. The task graph of the application is illustrated in the left part of Task graph is an acyclic directed graph derived from application profiling in advance. Every node in the graph represents a task in the application to be assigned to the core. Every edge represents

the communication requirement between the two corresponding nodes (tasks). The weight on the edge denotes the communication data volume between them. The power consumptions of cores running corresponding tasks, and communications between the min this hypothetical example are listed in the task graph.

First, a thermal-balanced task allocation method which is similar with that in is used. The method minimizes the stack power gradient across the 3-D Masco. The obtained solution, Tasks assigned in the same stack is grouped in a rectangular box. Communications across the stack are denoted using solid lines and those within the same stack are represented as dashed lines. By adjusting locations of tasks solution.

Reduces inter-stack communications while the power gradient remains the same as in solution. We relax the constraint of power gradient by half, and achieve the top module design which turns more inter-stack communications into intra-stack. To quantitatively evaluate these solutions, their stack power distributions and peak temperatures are shown and the corresponding interconnect energy is illustrated.

First, a thermal-balanced task allocation method which is similar with that is used. The method minimizes the stack power gradient across the 3-D MPSoC. The obtained solution is Tasks assigned in the same stack are grouped in a rectangular box. Communications across the stack are denoted using solid lines and those within the same stack are represented as dashed lines. By adjusting locations of tasks, temperature range reduces inter-stack communications while the power gradient remains the same as in MPSoC. We relax the constraint of power gradient and achieve the top module design, which turns more inter-stack communications into intra-stack. To quantitatively evaluate these solutions, their stack power distributions and peak temperatures are shown and the corresponding interconnect energy is illustrated.

We can see that interconnect energy consumption varies significantly among different solutions. With the same stack power gradient, energy can be reduced by temperature range compared with MPSoC. Meanwhile, with 1.6 W power gradient in top module design, we can gain another energy saving. The peak temperatures of three solutions are almost the same.

Xilinx is disclosing this document and intellectual property (hereinafter "the design") to you for use in the development of designs to operate on, or interface with Xilinx FPGAs. except as stated herein, none of the design may be copied, reproduced, distributed, republished, downloaded, displayed, posted, or transmitted in any form or by any means including, but not limited to, electronic, mechanical, photocopying, recording, or otherwise, without the prior written consent of Xilinx. Any unauthorized use of the design

may violate copyright laws, trademark laws, the laws of privacy and publicity, and communications regulations and statutes.

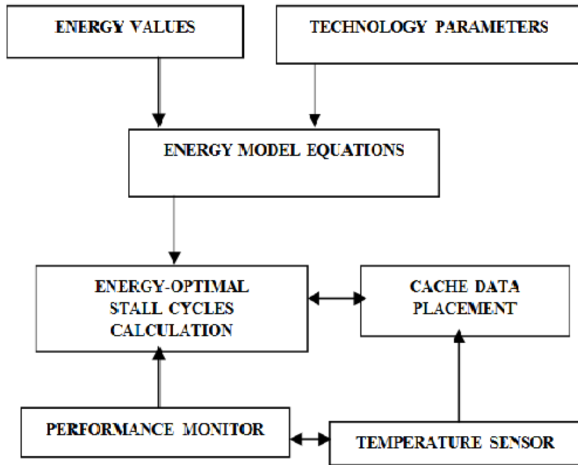

Fig 1. Block Diagram of Proposed System

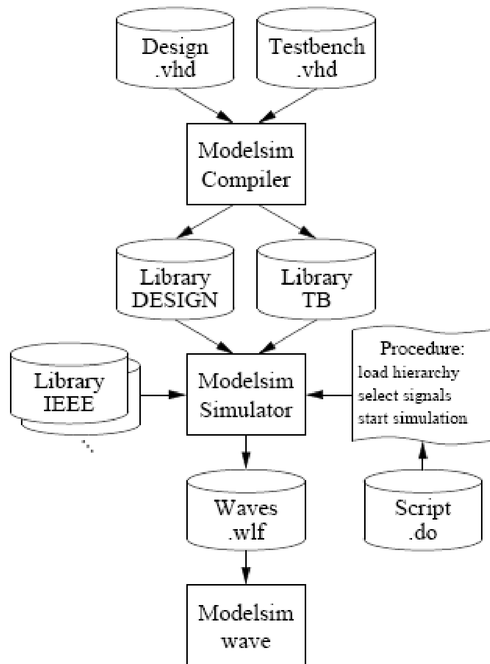

Fig 2. Simulation Flow

IV. XILINX - ISE

Xilinx does not assume any liability arising out of the application or use of the design; nor does Xilinx convey any license under its patents, copyrights, or any rights of others. You are responsible for obtaining any rights you may require for your use or implementation of the design. Xilinx reserves the right to make changes, at any time, to the design as deemed desirable in the sole discretion of Xilinx. Xilinx assumes no obligation to correct any errors contained herein or to advise you of any correction if such be made. Xilinx will not assume any liability for the accuracy or correctness of any engineering or technical support or assistance provided to you in

connection with the design. The design is provided "as is" with all faults and the entire risk as to its function and implementation is with you. You acknowledge and agree that you have not relied on any oral or written information or advice, whether given by Xilinx, or its agents or employees. Xilinx makes no other warranties, whether express, implied, or statutory, regarding the design, including any warranties of merchantability, fitness for a particular purpose, title, and non infringement of third-party rights.

V. VHDL

VHDL is an acronym which stands for VHSIC Hardware Description Language. VHSIC is yet another acronym which stands for Very High Speed Integrated Circuits. If you can remember that, then you're off to a good start. The language has been known to be somewhat complicated. The acronym does have a purpose, though; it is supposed to capture the entire theme of the language that is to describe hardware much the same way we use schematics.

VHDL can wear many hats. It is being used for documentation, verification, and synthesis of large digital designs. This is actually one of the key features of VHDL, since the same VHDL code can theoretically achieve all three of these goals, thus saving a lot of effort. In addition to being used for each of these purposes, VHDL can be used to take three different approaches to describing hardware.

These three different approaches are the structural, data flow, and behavioral methods of hardware description. Most of the time a mixture of the three methods is employed.

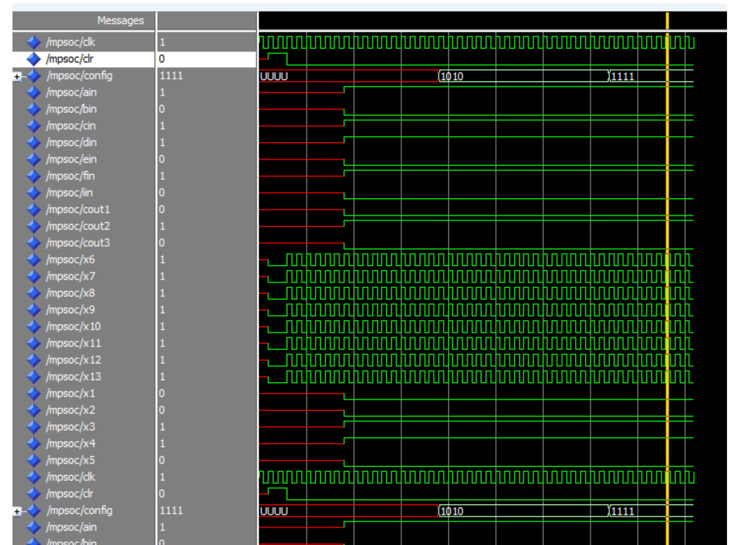

Fig3. Simulation Output for MPSoc

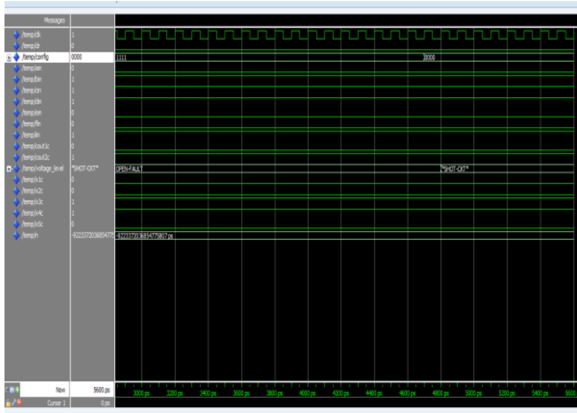

Fig4. Simulation Output for Temperature Range

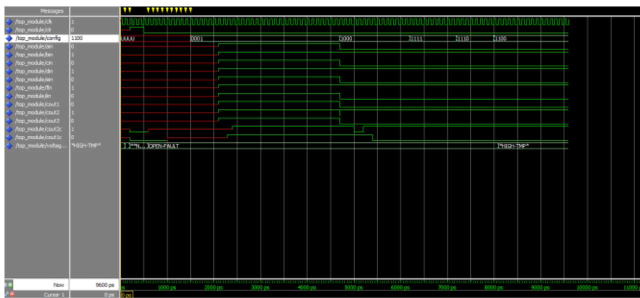

Fig5. Simulation Output for Top Module Design

VI. CONCLUSION

In this paper, we address the problem of cache data mapping for a multi-core architecture with 3D-stacked L2 cache to minimize the overall system energy. Unlike the classical approaches, our method considers memory traffic congestion as well as temperature distribution in the 3-D CMP. The proposed method places cache data to the specific physical position by exploiting the trade-off between the energy induced by memory traffic congestion and the energy induced by temperature of the cache block. In particular, we have shown that our proposed scheduling mechanism, Balancing-by-stack, outperforms other intuitive algorithms in the thermally homogeneous floor plan because of the following three properties. First, our scheduler takes into account the high thermal correlations among the layers in one core stack, and schedules tasks in bundles. Second, within every stack of cores, hot tasks are allocated to the layers that are closest to the heat sink for best heat dissipation. Third, upon a thermal emergency, power scaling is engaged in a core stack whose temperature exceeds the threshold, and to the core that generates the largest power in this stack. This can quickly cool down the core stack, reducing the performance penalty imposed to the task. We solved this problem with a runtime solution and the experiment results show that the proposed method yields up to 22.88% improvement in energy reduction compared to an existing runtime cache configuration method which only considers the temperature distribution.

REFERENCES

[1] B. Black, et al., "Die Stacking (3D) Micro architecture" in *Proc. The 39th Intl. Symp. On Micro architecture*, pp. 469-479, Dec. 2006.

[2] W. Liao, L. He, and K. Lepak, "Temperature-aware performance and power modeling" in Technical Report UCLA Engr. 04-250, UCLA, Los Angeles, CA, 2004.

[3] C. Zhang, F. Vahid, and W. Najjar, "A highly configurable cache for low energy embedded systems" in *ACM Trans. Embed. Comput. Syst.*, vol. 4, no. 2, pp. 363-387, May. 2005.

[4] D.H. Albonesi, "Selective cache ways: On-demand cache resource allocation" in *Proc. the 32nd Intl. Symp. On Micro architecture*, pp. 248-259, Nov. 1999.

[5] M. K. Qureshi and Y. N. Patt, "Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches" in *Proc. the 39th Intl. Symp. On Micro architecture*, pp. 423-432, 2006.

[6] M. Powell, et al., "Gated-Vdd: A circuit technique to reduce leakage in deep-submicron cache memories" in *Proc. the ACM/IEEE Intl. Symp. On Low Power Electronics and Design*, pp. 90-95, July. 2000.

[7] H. Noori, et al., "Temperature-Aware Configurable Cache to Reduce Energy in Embedded Systems" in *IEICE Trans. Electronics*, vol. 91, no. 4, pp. 418-431, 2008.

[8] W. Yun, et al., "Temperature-Aware Energy Minimization of 3D-Stacked L2 DRAM Cache through DVFS" in *Proc. ISOCC*, 2012, pp. 475–478.

[9] D. Zhao, H. Homayoun, and A. V. Veidenbaum, "Temperature Aware Thread Migration in 3D Architecture with Stacked DRAM" in *Proc. ISQED*, 2013, pp. 80–87.

[10] Y. Cheng, et al., "Thermal-Constrained Task Allocation for Interconnect Energy Reduction in 3-D Homogeneous MPSoCs" in *IEEE Trans. On Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 2, pp. 239-249, Feb. 2013.