Thermal Characterization of Cloud Workloads on a Power-Efficient Server-on-Chip

Dragomir Milojevic*, Sachin Idgunji‡, Djordje Jevdjic†, Emre Ozer‡, Pejman Lotfi-Kamran†, Andreas Panteli§, Andreas Prodromou§, Chrysostomos Nicopoulos§, Damien Hardy§, Babak Falsafi† and Yiannakis Sazeides§
*IMEC, †EPFL, ‡ARM, §University of Cyprus

Abstract*—[1] *We propose a power-efficient many-core server-on-chip system with 3D-stacked Wide I/O DRAM targeting cloud workloads in datacenters. The integration of 3D-stacked Wide I/O DRAM on top of a logic die increases available memory bandwidth by using dense and fast Through-Silicon Vias (TSVs) instead of off-chip IOs, enabling faster data transfers at much lower energy per bit. We demonstrate a methodology that includes full-system microarchitectural modeling and rapid virtual physical prototyping with emphasis on the thermal analysis. Our findings show that while executing CPU-centric benchmarks (e.g. SPECInt and Dhrystone), the temperature in the server-on-chip (logic+DRAM) is in the range of 175-200°C at a power consumption of less than 20W, exceeding the reliable operating bounds without any cooling solutions, even with embedded cores. However, with real cloud workloads, the power density in the server-on-chip remains much below the temperatures reached by the CPU-centric workloads as a result of much lower power burnt by memory-intensive cloud workloads. We show that such a server-on-chip system is feasible with a low-cost passive heat sink eliminating the need for a high-cost active heat sink with an attached fan, creating an opportunity for overall cost and energy savings in datacenters.

I. Introduction

In the era of multi-core architectures, Systems-on-Chip (SoCs) designed with power-efficient cores are an alternative to high performance ILP-intensive ones [1]. Data-intensive server workloads such as web servers, databases and application servers provide better performance per watt as well as comparable absolute performance with power-efficient cores [2], [3], [4], [5], [6]. Increased parallelism in many-core architectures using small, power-efficient cores requires high-bandwidth, low-latency memory systems. Traditional off-chip DDR3 memory interfaces are expensive in area per bit as well as the energy per bit consumed during reads from and writes to the memory.

3D-stacking using Through Silicon Vias (TSVs) is a pervasive technology in applications such as image sensors. This technology is now emerging as a viable candidate to address several challenges in the computing space from mobile computing through high-performance enterprise/server applications. In general-purpose computing this technology has now been integrated into next generation FPGAs [7] and is also being adopted by memory vendors to build stacked DDR3 modules [8]. The main driver of 3D-integration is to address the interconnect delays and the interface energy in advanced technology nodes. 3D-stacked systems reduce global interconnect delays significantly by reducing the number of repeaters in the design as well as improving performance at reduced power. An equally compelling use of 3D-ICs is heterogeneous integration and one of the obvious and a widely accepted form of this integration is a 3D-stack integration of DRAM on a multi-core logic die. The energy per bit for a TSV-based interface is an order of magnitude lower than the contemporary low-power DRAM interfaces (LPDDR2) and two orders lower against existing DDR3 interfaces, making the communication interface vastly energy-efficient. An emerging mobile DRAM standard is the JEDEC Wide I/O [9] that defines the memory interface in four 128-bit channels, with each channel giving a peak throughput of 3.2GB/s in the first generation, operating at 1.2V. Besides the obvious advantages, 3D-integration is accompanied with challenges associated with manufacturing such as die thinning, TSV filling, strata integrity. Additionally, as a part of the design process, 3D-driven floorplanning, TSV-μbumps co-design and the impact from the TSV-induced stress on the circuits must be considered for an efficient implementation.

In this paper, we propose a server-on-chip architecture tuned for the datacenter market, i.e., traditional server and cloud workloads. The server-on-chip consists of two distinct layers: the bottom layer contains a many-core compute engine and the top layer has Wide I/O DRAMs. We design the bottom layer (the logic die) to be aware of the available, TSV-enabled memory bandwidth, aiming to get the maximum performance out of the given die area. For that purpose, we try to maximize the number of cores on the chip at the expense of the second-level caches. We dedicate a small amount of area to the second-level cache, just enough to capture the instruction and hot data working sets of the data-intensive commercial workloads [3], [4], [5]. A bigger fraction of the area is dedicated to many processor cores in order to optimize for the total chip throughput, without sacrifying the single-thread performance. The 3D-stacked Wide I/O DRAMs are used as a last-level cache (LLC) shared by all cores.

The contributions of this paper are three-fold:
1) To the best of our knowledge, this is the first study that shows the temperature profile of a 3D DRAM-on-logic stack or server-on-chip that targets datacenter workloads. From the hardware point of view, our study considers a chip customized for the server market. From the software point of view, we consider a representative set of real-world cloud workloads

[1]This work was supported by "EuroCloud, Project No 247779" of the European Commission 7th RTD Framework Programme - Information and Communication Technologies: Computing Systems.

978-1-4673-3052-7/12/$31.00 ©2012 IEEE

running on such a system.
2) We find that CPU-centric applications (i.e., *SPECInt* and *Dhrystone*) can raise the temperature in the stack up to 200°C, exceeding the reliable operating bounds, even with embedded cores. On the other hand, we show that real datacenter workloads, which are memory-bound, tend to burn less power in the processing cores and, therefore, decrease the power density at critical hotspots.
3) We also show that the temperature in the server-on-chip while running such workloads remains within the operating bounds of the stacked DRAM using low-cost passive heat sinks. Considering the fact that a typical datacenter contains thousands of server chips, a small saving in the cooling equipment will have a drastic impact in the total server equipment cost as well as the energy savings from the thousands of attached fans.

The rest of the paper is organized as follows: *Section II* discusses the related work. *Section III* motivates the server-on-chip architecture with 3D-stacked DRAM. Then, *Section IV* describes the methodology of architectural exploration, virtual prototyping and the 3D-stack implementation. *Section V* describes the power and thermal modeling, and presents the power, power density and thermal results in the server-on-chip with a discussion of chip cooling options. Finally, *Section VI* concludes the paper.

II. Related Work

Several studies have explored the efficiency of using TSVs to implement a high-density memory interface that connects DRAM strata to processing systems [6], [10], [11], [12]. *Liu et al.* [10] propose to stack DRAM on a single-core system to improve memory latency and bandwidth. *Woo et al.* [11] propose a way to exploit the available bandwidth offered by the TSV technology by using large cache blocks. These studies do not address the thermal behavior and challenges associated with the elevated power density. However, the thermal behavior of 3D-stacked DRAM-on-logic systems has been explored by *Loi et al.* [12]. They investigate the impact of 3D-stacking on temperature of a single Alpha processing core built in the 130nm technology and 64MB 150nm DRAM. *Black et al.* [13] demonstrate a DRAM-on-logic exercise on 64MB DDR3 DRAM dies on a dual-core Intel Core Duo 2 processor using CPU-intensive benchmarks. They use an in-house 3D thermal modeling tool to analyze the thermal behavior of the chip. The 3D-stack chip uses desktop-class CPU cooling system (i.e. an active heat sink with a fan) because the CPU consumes about 92W. *Loh* investigates an 8-layer hypothetical 8GB DRAM stacked on top of a quad-core Intel Penryn-like processor running at 3.3GHz [14]. Similarly, *Sun et al.* [15] analyze the performance of an 8-layer hypothetical 1GB DRAM stacked on top of a quad-core 4GHz processor. Both studies rely on the assumptions of hypothetical, specially designed DRAM rather than commodity DRAMs, and no thermal analysis has been made in their studies. A multi-core 3D-stacking proposal was PicoServer [6] in the context of a power-efficient server-on-chip system, with design space exploration using small, power-efficient cores and larger, performance-optimized cores. A simple thermal modeling of the 3D stacks has been demonstrated using estimated power density map, as there was no 3D chip floorplanning.

Our study characterizes the thermal behavior of a 3D DRAM-on-logic stack containing a logic die with 16 high-performance power-efficient 2GHz ARM Cortex-A9 [16] cores, and DRAM dies using a four-channel JEDEC Wide I/O DRAM similar to the Samsung Wide I/O [9] using the scale-out cloud workloads. The 3D-stacked chip is implemented using virtual physical prototyping flow targeting an industrial 40nm G process technology with detailed 3D floorplanning and optimal TSV placement using an industry-class EDA flow. The thermal behavior of the server-on-chip is modeled by an accurate 3D compact thermal tool. Unlike the previous studies, we show that even with a power envelope of sub-20W, the 3D-IC stack can result in high power densities and high thermal operating points but the temperature in the stack can be kept under control with low-cost cooling options.

III. Server-on-Chip Architecture

The server-on-chip system consisting of a logic die, and Wide I/O DRAM dies stacked on top of the logic die is depicted in Figure 1. The logic die is populated with many cores to use up the available memory bandwidth and with sufficient second-level cache to capture the shared data and the instruction working set of our cloud workloads (unlike PicoServer [6], which advocates complete removal of second-level caches). Thus, the system is architected to provide maximum throughput without sacrifying quality of service [5]. As we dedicate the available area to cores and caches, we favor cores to caches as we try to improve the overall throughput by maximizing the logic die area used for cores and parallelism [5], [6], [17]. Although larger L2 caches can improve single-thread performance in desktop applications, our workloads observe limited benefit due to their huge data footprints, which are beyond the reach of today's SRAM caches, leading to a marginal improvement in the overall system throughput. Moreover, the latency incurred by large L2 caches limits both the single-thread and multi-thread performance [3], [5].

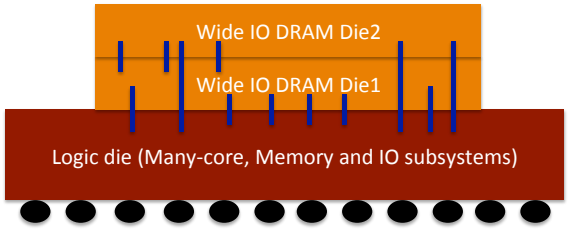

Fig. 1. The server-on-chip with the logic and Wide I/O DRAM dies

The second layer of the server-on-chip system consists of two Wide I/O DRAM dies. We limit the number of DRAM dies to two because the current state-of-the-art commercial Wide I/O DRAM technology provides 2-die stacked Wide I/O DRAM solutions. For example, Samsung [9] announced a 2-die stack of each die having a capacity of 1Gbit manufactured

in the 50nm technology in 2011. The 1Gbit Samsung Wide I/O DRAM occupies a die area of $64mm^2$. We project that the DRAM die capacity will double with each shrinking technology node. Thus, we expect that the Wide I/O DRAM die density should increase to 2Gbit and 4Gbit in 40nm and 30nm technology nodes respectively. In fact, industrial solutions already offer 4Gbit Wide I/O DRAM dies manufactured in 30nm technology in 2012 [18]. Therefore, it is reasonable to assume that two 4Gbit DRAM dies stacked together occupy the same die area (i.e., $64mm^2$), but providing a total DRAM capacity of 1GB.

Due to its limited capacity, the Wide I/O DRAM is architected to serve as a high-bandwidth last-level cache for all processor cores on the logic die, rather than serving as main memory, as assumed by the previous work [10], [14]. This is because the main memory capacity required by many processor cores in the logic die is in the order of tens of gigabytes. To motivate the use of Wide I/O DRAM as an LLC, we have measured the impact of 1GB Wide I/O DRAM LLC on the performance of the logic die having 16 cores while running several datacenter applications (the applications are detailed in *Section IV*) and the results can be seen in Figure 2. The results are normalized to the baseline model that has dual-channel DDR3 controllers (the first bar in Figure 2). The second bar represents the server-on-chip model that has dual-channel 6.4GB/s DDR3 controllers and 1GB Wide I/O DRAM LLC. The datacenter applications generate a lot of data traffic that cannot be handled by available on-chip DDR3 controllers, and therefore, the application runs slower to fit in the available bandwidth envelope. On the other hand, the applications run smoothly with a high-bandwidth Wide I/O interface of 12.8 GB/s, delivering 40% more throughput on average.

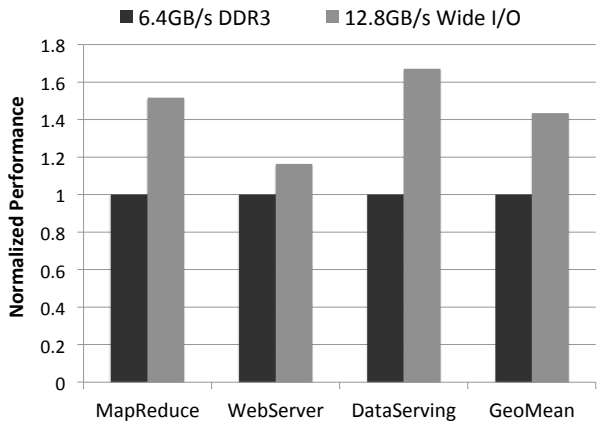

Fig. 2. Performance impact of Wide I/O DRAM

Figure 3 illustrates the architecture of the logic die of the proposed system. The logic die contains 16 processing tiles interconnected using a 4×4 mesh with 16 six-port routers. Each tile is composed of an ARM Cortex-A9 core, and an L2 cache bank. The cache coherence is maintained by a non-inclusive MESI-based invalidation protocol which uses a distributed coherence directory for scalability. The routers implement three virtual channels required by the coherence protocol, have 128-bit wide links matching the Wide I/O channel width. The four Wide I/O DRAM controllers are located at the center of the logic die connected to the four tiles in the center of the mesh. The Wide I/O DRAM controllers are placed in the center to achieve the optimal TSV placement because the TSV arrays on the DRAM dies are also located in the die center. The 1GB 3D-stacked Wide I/O DRAM LLC is split into four banks each connected to an independent Wide I/O memory channel. The LLC is organized as a page-based cache, caching 2KB pages. The cache bank interleaving happens at 2KB boundaries, matching the most commonly used DRAM row buffer size. The Wide I/O DRAM controllers are tightly coupled to the SRAM tag arrays for the corresponding on-chip DRAM banks to form the LLC controllers. The SRAM tag arrays are organized as 32-way set-associative structures. The top left and the bottom right nodes of the

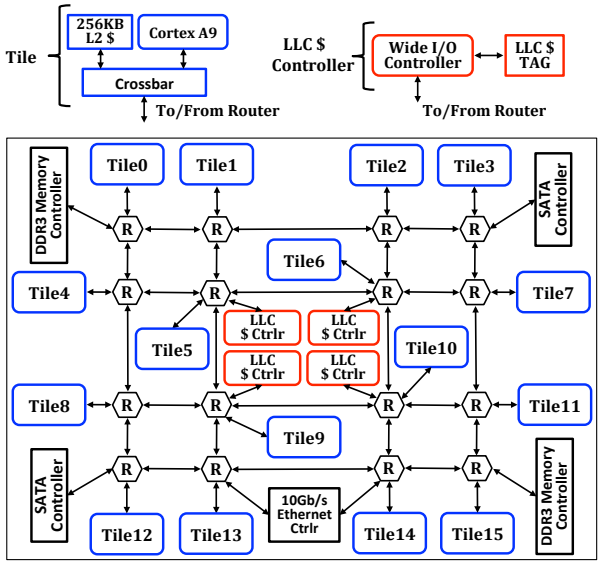

Fig. 3. 16-tile logic die architecture

mesh are connected to an on-chip DDR3 memory controller each. These controllers communicate with the off-chip DRAM (main memory), delivering 6.4GB/s of bandwidth. All memory controllers are on the chip, but they control either the on-chip or off-chip DRAM. Finally, the server die has two SATA controllers and a 10Gbit/s Ethernet port to access the hard disk drives and network.

IV. Modeling the Server-on-Chip

Our power-efficient server-on-chip architecture is modeled through a virtual prototyping design flow, illustrated in Figure 4. The flow is divided into two major steps: a) architectural exploration and b) virtual physical prototyping. In the architectural exploration phase, we focus on the application-driven microarchitectural design without any concerns about the physical properties of the circuit. In the second phase, we

model the system at the physical level (RTL, gate-level and layout) to produce an actual physical prototype of the design.

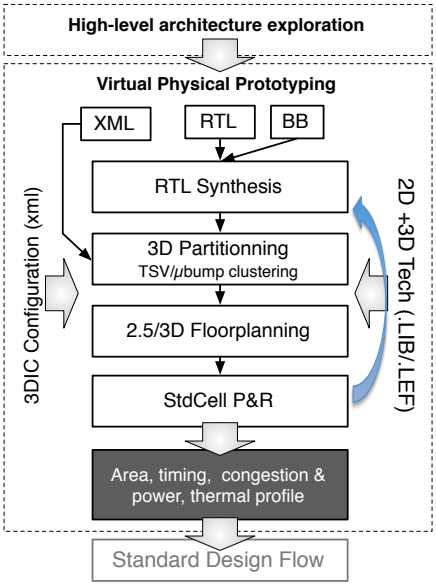

Fig. 4. Prototyping design flow

A. Architectural Exploration

We analyze the chosen set of datacenter applications using a combination of trace-based and cycle-accurate full-system simulations using the *Flexus* simulation framework [19]. *Flexus* models a RISC ISA and can execute unmodified commercial applications and operating systems. *Flexus* extends the Virtutech Simics functional simulator with models of processing tiles with out-of-order cores, NUCA cache, on-chip memory controllers, on-chip interconnect and IO interfaces. The micro-architectural parameters are chosen to match the Cortex-A9 behavior. The rest of the architectural parameters are determined after an exhaustive simulation of whole system. these parameters include the L2 cache size requirements, DRAM LLC capacity, the number of DDR3 controllers and IO interfaces based on off-chip memory and IO bandwidth requirements.

We rely on the CloudSuite benchmarks [17] as a representative set of real-world applications that dominate the use of today's datacenter infrastructure. In this work, we analyze only the applications that exhibit significantly different behavior in terms of on-chip activity.

$WebServer$: Web servers have always been an omnipresent datacenter application and as such are widely present in the cloud. We use the industrial benchmark SPECWeb2009 running the e-banking workload. The benchmark runs nginx 1.0.1, a highly scalable web server, with a built-in PHP 5.2.6 module and APC 3.0.19 PHP opcode cache.

$DataServing$: Cloud operators, such as Facebook and Google, rely on NoSQL data stores for fast and scalable storage with varying and rapidly evolving storage schemas. NoSQL systems split hundreds of terabytes of data into shards and horizontally scale to large cluster sizes, typically using various indexing schemes that support fast lookup and key range scans to retrieve the set of requested objects. We benchmark one node running the Cassandra 0.7.3 NoSQL data store using the YCSB 0.1.3 client and a dataset that exceeds the memory capacity in order to mimic a realistic setup.

$MapReduce$: The map-reduce paradigm has emerged as a scalable approach to handling large-scale data analysis, farming out requests to a cluster of nodes that first perform filtering and transformation of the data (map) and then aggregate the results (reduce). We benchmark one node running the WordCount workload on a 30 GB set of Wikipedia pages on top of a Hadoop 0.20.2 cluster.

The server-on-chip architecture parameters, shown in Table I, are chosen to meet the requirements of the cloud workloads, after several iterations of exhaustive full-system simulations running the selected workloads.

TABLE I
THE SERVER-ON-CHIP MICROARCHITECTURAL PARAMETERS

Processing subsystem	16 out-of-order cores
Processing Cores	2GHz, 8-stage pipeline, 2-wide dispatch/retirement
L1 Caches	Split I/D, 32KB 2-way, 64-byte blocks
L2 NUCA Caches	256KB per core, 16-way, 64-byte blocks, 32 MSHRs
Interconnect	4x4 2D mesh, 3 VC per port, 128-bit flits, 3 cycles per hop
DRAM LLC	1GB, 4 banks, 40ns latency, 2KB blocks, 2KB bank-interleaving
DRAM Controller	4 Wide I/O DRAM controllers and 2 on-chip DDR3 DRAM controllers
IO Controller	2 on-chip SATA and 1 on-chip 10Gbit/s Ethernet controllers

B. Virtual Physical Prototyping

Virtual physical prototyping [20] is a design practice that allows computer architects to plan advanced packaging ICs in a holistic fashion before the actual design flow. During this design phase, we typically perform the following steps: 3D design partitioning; TSV/μbumps array partitioning, clustering, place and route; 2D and 3D technology parameters choices and their co-optimization with the architectural design; 3D floorplanning with a standard cell placement and route; and an early mechanical, thermal and reliability assessment. Virtual prototyping does not aim to modify the current design methodology, but it is rather used before standard industrial design flow tools. The input to the flow is an architecture described using synthesizable RTL (VHDL or Verilog) or in black-box RTL stubs. Since we use hard macros for the CPU core and L2 cache, these components are described as high-level, black-box (BB) models as shown in Figure 4. The black-box description provides a minimal set of information required to perform one full design flow iteration, and includes: the interface definition, the area, timing and power models. Besides the design data, the virtual prototyping system relies on library information from IP vendors (.LEF/.LIB)

and technology, electrical, physical and design rule parameters from the foundry. The design is partitioned and floorplanned on a per-tier basis. The router logic goes through a place and route (P&R) at the standard cell level. After P&R, we extract the parasitics and run performance/area analyses to assess and verify the architectural and technology choices for a given design configuration. The choices are based on metrics such as area, timing and congestion analysis both for the front- and back-side. The power distribution is fed to the virtual prototyping infrastructure to generate power density maps. This information, together with the 3D stack configuration (die, BEOL and the interface thickness and properties,the package/cooling thermal resistance, etc.) is forwarded to a thermal modeling tool that generates a thermal profile for each die in the system. In our analysis, we consider average power and power density distributions to generate the thermal profiles.

The 3D SpyGlass Physical tool from *Atrenta* is used to model the virtual physical prototyping of the 3D-stack. The standard 2D version of the tool has been extended to support the 3D integration requirements: the backside routing capability (including the congestion analysis and on-the-fly technology exploration of the TSV diameter, pitch and the Keep-Out Zone area size), support for easy TSV and μbump array partitioning, clustering and placement constraining. For the thermal analysis, we use the Compact Thermal Model [21], developed at IMEC that has been silicon-validated using similar 3D stack configurations [22]. The overall run-time of the flow is short, even for complicated designs that are fully described at the RTL level. The design set-up time is typically measured in hours and an average iteration time is measured in tens of minutes (including floorplanning, P&R and parameter extraction). A temperature profile, for a given floorplan or 3D stack configuration is extracted in just a few minutes.

C. 3D-die Stack Model

The logic die is implemented using the 40nm process, with a 0.9V nominal operating voltage. The area and peak power for the logic die components are derived from several sources. The area and power numbers for the ARM Cortex-A9s are provided by ARM [23], while the numbers for the Wide I/O DRAM LLC controllers and L2 caches are estimated using in-house tools. The router code is written in Verilog, synthesized and implemented using the same technology, after which the die area and power numbers are calculated. The parameters for DDR3 memory, 10Gbit/s Ethernet and SATA controllers are estimated using the *Cadence* InCyte Chip Estimator tool [24]. On the other hand, the DDR3 PHY area and power numbers are gathered from the Synopsys DesignWare Digital IP reference guide [25]. Finally, the area and power numbers for the on-chip DRAM dies are derived from the Samsung Wide I/O DRAM specification [9]. After floorplanning, the logic die size is $100mm^2$, slightly bigger than the stacked DRAM ($64mm^2$).

The logic die is thinned to the height dictated by the aspect ratio constraints for the TSVs and it is oriented face down and connected to the package using C4 bumps. The logic and Wide I/O DRAM dies are stacked with μbumps using the back side of the logic die and the front side of Wide I/O DRAM. The interconnect between the TSVs and μbumps uses one redistribution layer (RDL). The logic die is synthesized and then partitioned into macros for design planning. The Wide I/O layout models with μbump positions are created for iterative 3D floorplanning. The locations of the TSVs are chosen based on: a) TSV technology parameters: the diameter, pitch and depth; b) TSV array partitioning and c) TSV array placement with respect to the rest of the system. The ideal approach would be to place the TSVs at exactly the same position as μbumps. However, the server-on-chip needs to be floorplanned according to rigid area/utilization constraints as well as performance/power and mechanical constraints. The design planning considers the entire TSV array as a placement blockage with a keep-out margin. The TSV partitions are made on a per-channel basis in square areas with a 20μm pitch and with an RDL pitch of 2μm. The floorplan and RDL routing are illustrated in Figure 5.

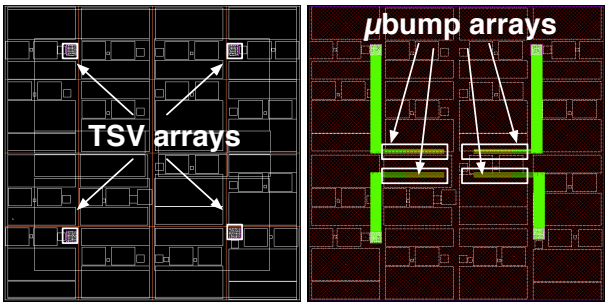

Fig. 5. The floorplan (left) and RDL routing (right) of the logic die

V. Thermal Characterization and Results

A. Power Modeling

Using full-system simulations with *Flexus* [19], we first measure the number of committed instructions per cycle for each core, the number of read and write accesses to each L2 cache bank, the number of the DRAM LLC tag-array lookups, the number of read/write requests served by each on-chip memory controller, the number of packets processed by the network-on-chip routers and number of requests going off-chip. This activity data is used for accurate power modeling to calculate the effective power consumption for the server-on-chip components. As expected, the data-centric cloud workloads and traditional CPU-centric applications exhibit very different activity distributions across the system components, which is, as we will later see, reflected on the corresponding power maps.

The effective power consumed by the logic die is estimated using *McPAT* [26], an integrated modeling infrastructure that estimates power, area and timing for SoC designs at the microarchitectural level. We select 45nm (which is the closest to our 40nm G implementation technology node) as the

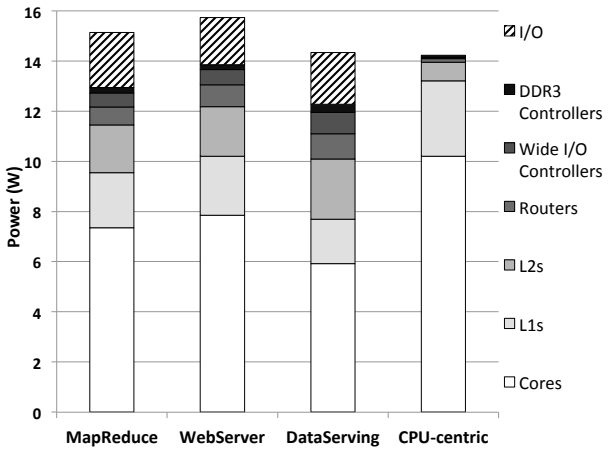

Fig. 6. The logic die power breakdown

TABLE II
THE POWER CONSUMPTION OF THE 2-DIE 1GB WIDE I/O DRAM

Workload	Wide I/O DRAM Die Power
MapReduce	1.2W
WebServer	1.1W
DataServing	1.4W
CPU-centric	0.02W

underlying technology using a 0.9V nominal operating voltage. The system architecture configuration and the activity data for each of the components in the logic die are passed to *McPAT* using an XML interface that decouples the system simulator statistics generated by *Flexus*. For the router power estimates, we use the ORION power modeling infrastructure for network-on-chip systems [27]. We generate statistical samples based on the traffic in the network to measure power consumed by each of the 16 routers and the power in the links that connect the nodes. Our observation is that the average power dissipation across the cores is fairly uniform for each of the workloads.

Figure 6 shows the power breakdown of the logic die for the cloud workloads. We also report the geometric mean power for the CPU-centric workloads (*SPECInt* and *Dhrystone*) in the last bar as *CPU-centric*. Between 40-50% of the total power is dissipated by the cores, and 25-27% of the total power goes to the caches (L1+L2). The IO interface has the third highest share of the total power, between 11-14%, while the remaining power is distributed among the rest of the system components. Overall, the maximum power dissipation in the logic die does not exceed 16W. We also provide the effective power consumption numbers of the DRAM dies when running cloud workloads in Table II. The maximum Wide I/O DRAM power consumption is close to 1.5W. This allows a sub-20W server-on-chip design consisting of 16-core Cortex-A9 and 1GB Wide I/O DRAM dies on top. The CPU-centric workload set has a lower total power profile because it does not exercise the Wide I/O DRAM, off-chip memory and IO interfaces, and therefore these components consume near-zero dynamic power.

B. Power Density, Thermal Profile and Chip Cooling

The power, power density, thermal modelling methodology is illustrated in Figure 7. We use the *Flexus* infrastructure to generate workload-specific activity numbers for the server-on-chip components. The McPAT and ORION tools produce the power map of the SoC. The 3D chip virtual prototyping tool generates power density maps from the combination of the SoC and Wide I/O DRAM power maps. The power density maps and 3D-IC stack configuration serve as inputs to the Compact Thermal Model [21] to generate the thermal profile for the whole chip. To model various chip packaging and cool-

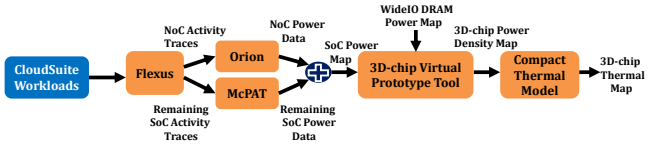

Fig. 7. Overview of power, power density and thermal modeling methodology

ing options, we select three different junction-to-air thermal resistance values (10, 3 and 1°C/W.). The server-on-chip has a die area of $100mm^2$ with approximately 700 pins, as derived from *Cadence* InCyte Chip Estimator tool [24]. The recommended chip package from InCyte Chip Estimator,considering the die size and the number of pins, is a ceramic ball grid array or fine ball grid array (FBGA). Hence, an FBGA package of 30x30mm or 35x35mm will be a reasonable package size for the server-on-chip. This is also supported by the *Altera APEXII EP2A40* chip [28] that has 672 pins packaged in 27x27mm dimensions. The junction-to-ambient resistance of the *EP2A40* package is given as 10°C/W, and its junction-to-case thermal resistance is 0.2°C/W. Thus, the selected junction-to-ambient thermal resistance value of 10°C/W represents the *package-only* solution without any cooling. The junction-to-ambient thermal resistances of 3 and 1°C/W represent *packaging + passive heat sink* and *packaging+active heat sink*, respectively. The passive heat sink dissipates heat through its fins on top. On the other hand, the active heat sink uses forced air cooling through a fan on the top. The thermal resistance of heat sinks depends on the size, height and the attached fan. The thermal resistance of passive heat sinks are an order of magnitude higher than active heat sinks. For example, the thermal resistance of a passive heat sink with a size of 35x35mm is around 2.5°C/W for heights higher than 20mm [29]. The thermal resistance of an active heat sink is normally smaller than 0.5°C/W [30]. The thermal resistance values and their associated package and cooling solutions are summarized in Table III.

A full thermal analysis using the compact thermal model is performed for all the workloads with three different junction-to-ambient thermal resistances. In Table IV, Table V and Table VI, we report the maximum temperatures in the stack for the three thermal resistance values. The ambient temperature is 25°C in all experiments. The nominal operating temperature

TABLE III
JUNCTION-TO-AMBIENT THERMAL RESISTANCES AND THE CORRESPONDING PACKAGE AND COOLING SOLUTIONS

30x30mm FBGA Package	Cooling	Junction to ambient thermal resistance
0.2°C/W	No heat sink (9.8°C/W)	10°C/W
0.2°C/W	Passive heat sink (2.5°C/W)	∼3°C/W
0.2°C/W	Active heat sink (0.5°C/W)	∼1°C/W

for logic die lies between 85°C and 125°C while the DRAM temperature must be kept under 90°C to sustain the nominal refresh rate.

We observe that the chip temperature for the CPU-centric workloads is 7-10% higher than the maximum temperature in the chip when running cloud workloads, despite the fact that the CPU-centric workloads consume about 5% lower total power compared to the nearest cloud workload. The key reason is that the power density across the logic die is not as uniform as the power density seen with the cloud workloads. The power dissipated by the cores for CPU-centric workloads is much higher due to high switching activity in the cores and L1 caches leading to a higher degree of non-uniformity in the power density across the cores. This results in higher temperatures in the areas where the cores are physically located. As an example, we show the power density and thermal profile maps of the logic and DRAM dies for a package having a thermal resistance of 3°C/W when running MapReduce in Figure 8. The power density of the DRAM dies is uniform for all workloads, and therefore there is no power density variation. On the other hand, the power density varies on the logic die with hot spots in and around the routers and the LLC tags in the center.

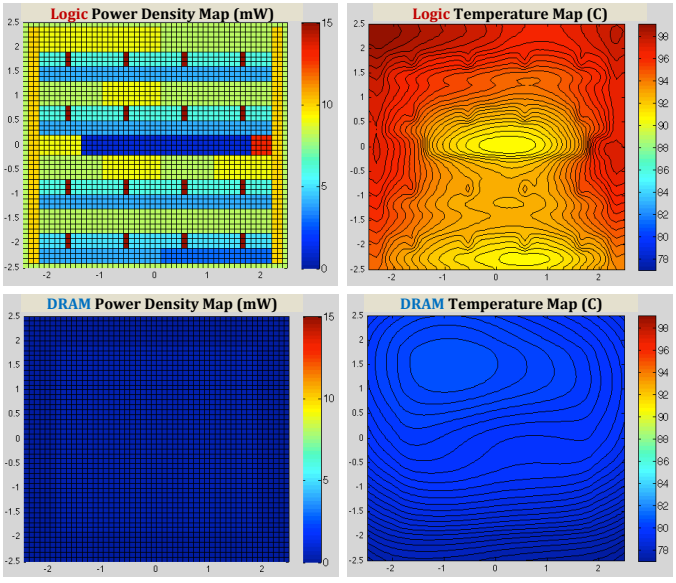

Fig. 8. Power density and thermal profile maps of the logic and DRAM dies for a package solution having a thermal resistance of 3°C/W when running MapReduce

Table IV shows that a package-only solution is not sufficient to keep the die temperatures under the nominal operating temperatures. Both the logic and DRAM die temperatures are beyond the nominal operating temperature ranges. The die temperatures when running real cloud workloads are indeed 25-30°C lower than the worst case, which supports our hypothesis that cloud workloads have a lower thermal profile than CPU-centric workloads. When moving to 3°C/W,

TABLE IV
MAXIMUM TEMPERATURES IN LOGIC AND DRAM DIES FOR A PACKAGE-ONLY SOLUTION

	Logic Die	DRAM dies
MapReduce	177.4°C	154.7°C
WebServer	183.3°C	160.5°C
DataServing	167.6°C	147.2°C
CPU-centric	201°C	175.4°C

TABLE V
MAXIMUM TEMPERATURES IN LOGIC AND DRAM DIES FOR A PACKAGE WITH A PASSIVE HEAT SINK

	Logic Die	DRAM dies
MapReduce	99.1°C	80.7°C
WebServer	101.5°C	82.9°C
DataServing	94.1°C	77.3°C
CPU-centric	110.3°C	89.7°C

which represents a package with a passive heat sink cooling solution, the die temperatures settle down to the allowed range, as shown in Table V. For the cloud workloads, the DRAM die temperatures are below the nominal operating temperature of 90°C while the temperature for the CPU-centric workloads is on the border. So, even a low-cost passive heat sink is sufficient to cool the 16-core server-on-chip with a 1GB Wide I/O DRAM, leaving enough room for short CPU-intensive bursts. However, in case of longer CPU-intensive bursts (several seconds), which are highly unlikely to happen in cloud workloads, techniques similar to thermal buffering can be used [31]. In Table VI, we also show the die temperatures

TABLE VI
MAXIMUM TEMPERATURES IN LOGIC AND DRAM DIES FOR A PACKAGE WITH AN ACTIVE HEAT SINK

	Logic Die	DRAM dies
MapReduce	73.6°C	56.2°C
WebServer	74.7°C	57°C
DataServing	69.6°C	56.4°C
CPU-centric	80.5°C	61.3°C

with a forced air cooling solution using an active heat sink. The die temperatures for both workload classes are comfortably below the nominal operating temperatures.

Although the active heat sink solution can be used in our server-on-chip design, it would be an overkill in terms of cooling equipment costs and energy. The choice between the active and passive heat sink may not be so distinct from a single server perspective. However, in datacenters that host thousands of servers, this choice will have a huge impact on the total cost of ownership (TCO) of the datacenter, because the TCO is driven by the energy bills and server/cooling

equipment cost [32]. For example, an Internet-service based datacenter may have the number of servers between 5,000 and 50,000. Today, a typical server blade has two sockets, but with the increasing popularity of power-efficient micro-servers, future blade servers will have more than two chips per blade [33]. This amounts to from 20,000 to 200,000 server chips for an typical datacenter. Thus, it becomes obvious that even a small saving in the cooling equipment will have a drastic impact in the total equipment cost as well as the energy savings from the thousands of attached fans, let alone the reduction in the datacenter noise levels.

VI. Conclusion

We studied the thermal behavior of a server-on-chip designed with 16 power-efficient ARM Cortex-A9 cores, and 1GB Wide I/O DRAM serving as a last-level cache. The Wide I/O DRAM subsystem is stacked on top of the logic die using the 3D-IC technology. We modeled the 3D-stacked chip using a set of design tools in a virtual physical prototyping environment. The server-on-chip implementation is simulated using real-world datacenter applications for obtaining the power density and thermal maps. Our results showed that the temperatures of a 16-core server-on-chip designed with power-efficient cores and 3D-stacked DRAM dies can exceed the acceptable temperature boundaries while running CPU-centric applications, typically used as representative workloads both by the industry and academia. However, while running the real-world data-intensive cloud workloads used in datacenters, the die temperatures in the server-on-chip can be kept under nominal operating temperatures using low-cost passive heat sink cooling solutions.

References

[1] V. J. Reddi et al., "Web search using mobile cores: quantifying and mitigating the price of efficiency," in *Proceedings of the 37th International Symposium on Computer Architecture (ISCA)*, Jun 2010.

[2] J. D. Davis et al., "Maximizing CMP throughput with mediocre cores," in *Proceedings of the 14th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, Sep 2005.

[3] N. Hardavellas et al., "Database servers on chip multiprocessors: limitations and opportunities," in *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR)*, 2007.

[4] N. Hardavellas et al., "Toward dark silicon in servers," *IEEE Micro*, vol. 31, no. 4, pp. 6–15, Jul-Aug 2011.

[5] P. Lotfi-Kamran et al., "Scale-out processors," in *Proceedings of the 39th International Symposium on Computer Architecture (ISCA)*, Jun 2012.

[6] T. Kgil et al., "PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor," in *Proceedings the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2006)*, Oct 2006.

[7] K. Saban, "Xilinx stacked silicon interconnect technology delivers breakthrough FPGA capacity, bandwidth, and power efficiency," http://www.xilinx.com/support/documentation/white_papers/wp380_Stacked_Silicon_Interconnect_Technology.pdf, Oct 2011.

[8] U. Kang et al., "8Gb 3D DDR3 DRAM using through-silicon-via technology," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2009.

[9] J. Kim et al., "A 1.2V 12.8GB/s 2Gb mobile Wide-I/O DRAM with 4×128 I/Os using TSV-based stacking," in *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2011.

[10] C. C. Liu et al., "Bridging the processor-memory performance gap with 3D IC technology," *IEEE Design and Test of Computers*, vol. 22, no. 6, pp. 556–564, Nov-Dec 2005.

[11] D. H. Woo et al., "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," in *Proceedings of the 16th International Symposium on High Performance Computer Architecture (HPCA)*, Feb 2010.

[12] G. L. Loi et al., "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *Proceedings of the 43rd Design Automation Conference (DAC)*, Jul 2006.

[13] B. Black et al., "Die-stacking (3D) microarchitecture," in *Proceedings of the 39th International Symposium on Microarchitecture (MICRO)*, Dec 2006.

[14] G. Loh, "3D-stacked memory architectures for multi-core processors," in *Proceedings of the 35th International Symposium on Computer Architecture (ISCA)*, Jun 2008.

[15] H. Sun et al., "3D DRAM design and application to 3D multicore systems," *IEEE Design and Test of Computers*, vol. 26, no. 5, Sep-Oct 2009.

[16] ARM Information Center, *Cortex-A9 Technical Reference Manual*, ARM, 2008.

[17] M. Ferdman et al., "Clearing the clouds: A study of emerging workloads on modern hardware," in *Proceedings the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2012)*, Mar 2012.

[18] Elpida, "Elpida memory starts sample shipments of next-generation mobile ram products," http://www.elpida.com/en/news/2011/12-28.html, Dec 2011.

[19] T. Wenisch et al., "SimFlex: statistical sampling of computer system simulation," *IEEE Micro*, vol. 26, no. 4, pp. 18–31, Jul-Aug 2006.

[20] A. D. Milojevic et al., "Pathfinding: a design methodology for fast exploration and optimisation of 3d-stacked integrated circuits," in *Proceedings of the 11th international conference on System-on-chip (SOC'09)*, 2009.

[21] C. Torregiani et al., "Compact thermal modeling of hot spots in advanced 3d-stacked ics," in *11th Electronics Packaging Technology Conference (EPTC'09*, Dec 2009.

[22] D. Milojevic, H. Oprins, J. Ryckaert, P. Marchal, and G. V. der Plas, "Dram-on-logic stack: Calibrated thermal and mechanical models integrated into pathfinding flow," in *IEEE CICC, Proceeding on*, September 2011.

[23] "ARM Cortex-A9 Performance Power and Area," http://www.arm.com/products/processors/cortex-a/cortex-a9.php.

[24] "Cadence InCyte Chip Estimator Data Sheet," http://www.cadence.com/rl/Resources/datasheets/incyte_chip_estimator_ds.pdf.

[25] "Synopsys DesignWare Digital IP Quick Reference Guide," https://www.synopsys.com/dw/doc.php/doc/dwf/manuals/dw_digital_ip_quickref.pdf.

[26] S. Li et al., "Mcpat: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proceedings of the 42nd Annual International Symposium on Microarchitecture*, Dec 2009.

[27] H.-S. Wang et al., "Orion: A power-performance simulator for interconnection networks," in *Proceedings of the 35th annual International Symposium on Microarchitecture*, Nov 2002.

[28] Altera, "Package information datasheet for mature altera devices," http://www.altera.com/literature/ds/pkgds.pdf, Dec 2011.

[29] Alpha Company Ltd, "Passive heat sinks," http://www.micforg.co.jp/en/cat_pass.html, 2012.

[30] "Active heat sinks," http://www.micforg.co.jp/en/cat_fe.html, 2012.

[31] A. Raghavan et al., "Computational sprinting," in *Proceedings of the 18th International Symposium on High Performance Computer Architecture (HPCA)*, 2012.

[32] D. Hardy et al., "EETCO: A tool to estimate and explore the implications of datacenter design choices on the tco and the environmental impact," in *Workshop on Energy-efficient Computing for a Sustainable World in conjunction with the 44th Annual IEEE/ACM International Symposium on Microarchitecture (Micro-44)*, Dec 2011.

[33] Dell, "Copper enables the ARM server ecosystem," http://content.dell.com/us/en/enterprise/d/campaigns/project-copper, 2012.