This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2015.2495125, IEEE Computer Architecture Letters

Thermally Adaptive Cache Access Mechanisms for 3D Many-core Architectures

He Xiao, Wen Yueh, Saibal Mukhopadhyay, *Member, IEEE,* and Sudhakar Yalamanchili, *Fellow, IEEE*
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332
{hxiao, wyueh3}@gatech.edu, {saibal.mukhopadhyay, sudha.yalamanchili}@ece.gatech.edu

Abstract—A compelling confluence of technology and application trends in which the cost, execution time, and energy of applications are being dominated by the memory system is driving the industry to 3D packages for future microarchitectures. However, these packages result in high heat fluxes and increased thermal coupling challenging current thermal solutions. Conventional design approaches utilize design margins that correspond to worst case temperatures and process corners leading to a significant impact on system level performance. This paper advocates a design approach based on microarchitecture adaptation to device-level temperature-dependent delay variations to realize average case performance that is superior to which can be achieved by using worst case design margins. We demonstrate this approach with adaptation principles for the last level cache (LLC) in a 3D many-core architecture. We propose and evaluate two adaptation mechanisms. In the first case, the access time to the LLC from the L1 tracks the LLCs temperature-delay variations. In the second case, the processor DVFS state tracks the LLC temperature as a negative feedback. Compared to a worst case design baseline, the full system simulation results show that both approaches increase the IPC by over 20%, and improves the energy efficiency by up to 3%.

1 Introduction

As CMOS technology advances, we are observing a confluence of technology and application trends in which the cost ($), execution time, and energy of applications are being dominated by the memory system. This is driving the industry to 2.5D and 3D packages for processor and memory systems. However, these packages also lead to higher heat fluxes and increased thermal coupling between the die challenging thermal solutions [1] [2]. The key issue addressed in this paper is that conventional design approaches for 3D systems utilize design margins that correspond to worst case temperatures and process corners. While such physical conditions may not occur often, the use of worst case design margins leads to a significant impact on average and peak system level performance. This paper advocates for microarchitecture operational principles based on adaptation to thermal effects to improve performance over that achievable with designs based on worst case margins and demonstrate that this approach has considerable promise. The thermally adaptive mechanism is presented using the multi-physics methodology, interacting with the available thermal headroom and circuit critical path delay during operation. This approach differs from past approaches focused on adaptation to maintain temperatures below a peak value. In contrast, our techniques extend the dynamic operating range (voltage and temperature) of the processor and view thermal headroom also as a resource to be consumed for performance.

We consider a 3D many-core architecture [3] that integrates a 16-core logic die, an LLC die and a 3D DRAM stack from Micron's Hybrid Memory Cube (HMC) [4]. This 3D model implies a future direction of the memory hierarchy, which enables high communication bandwidth between processors and main memory. Applications can exhibit a wide range of thermal behaviors affecting the temperature-dependent delay characteristics of the SRAM-based LLC producing temperature dependent access times. We provide a characterization of this delay behavior and propose two mechanisms for adapting to these delay variations. The first mechanism adapts the L1-LLC interface to vary the LLC access time as a function of temperature. The second mechanism adapts the core speed and scales the LLC frequency to match the time-varying LLC hit time. Using a full system simulator executing stock 32-bit x86 applications, we quantify the feasible performance gains and share some insights into the potential of this approach seeking to establish the need for, and value of, a multi-physics co-design approach for 3D microarchitectures for future processor design.

2 System Model

The target system is a 16-core homogenous x86 processor in a 3D stack organization illustrated in Figure 1. The cores reflect a typical out-of-order core design with the floorplan as shown in Figure 1. Guidelines for pipeline parameters are derived from a Intel Nehalem processor description [5]. The processor is modeled at the 16nm technology node and is the bottom die of the stack, followed by the LLC and a HMC-style DRAM stack in terms of internal concurrency. In this paper we only model the stack and not the next level of the memory hierarchy.

The cache hierarchy includes a 16KB private L1 data cache with a 1 cycle hit time and a shared LLC divided into 16, 2MB banks. The cache coherent protocol is directory-based MESI co-located in the LLC cache. The on-chip network is a 2D torus with 128-bit channels, which connects the LLC and the DRAM stack. The virtual channel wormhole switched routers are on the processor die. Memory controllers are integrated in the LLC tier. Each die is modeled as three layers, indicated in Figure 1. Electrical interconnects between different dies is realized with through silicon via (TSVs). The heat sink on top of the package is configured as the forced air convection cooling with heat transfer coefficient of 100 W/m^2-$^\circ C$.

3 Thermal Adaptation: Model

The temperature-delay characterization in an SRAM bank is simulated with a Hspice model in 16nm technology, depicted in Figure 2. The transistor sizing and cell configurations are optimized for the predictive model from Sinha et al. [6]. The critical path of a conventional SRAM bank is limited by the word-line driver, cell drive bit-line, sensamp sensing, and bit-line

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2015.2495125, IEEE Computer Architecture Letters

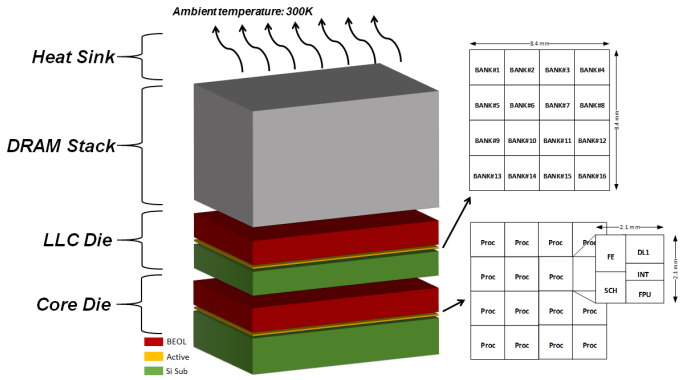

Fig. 1. The physical structure of the stacked 3D chip.

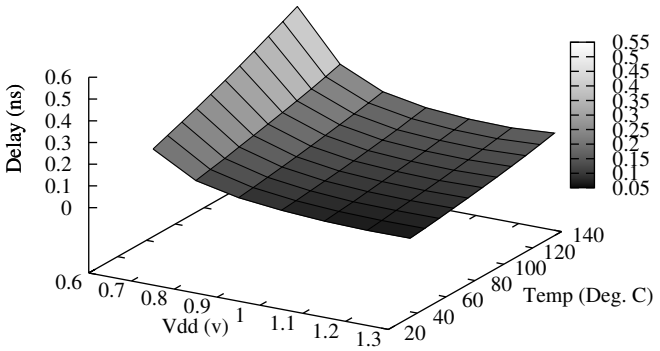

Fig. 2. The SRAM static timing delay model in terms of supply voltage and temperature

precharge/sensamp reset. This model assumes the wordline-rest is masked during sensamp evaluation with a divided-bitline multiplexing architecture. A latch-based sense amplifier architecture is considered for simulation of sense-amp delay [7]. Due to the regularity of the SRAM array, the extracted critical path of the sub-array is deterministic, and is defined as:

$$T_{random-cycle} = T_{wordline-driver} + T_{cell-drive-bitline} + T_{sensamp} + T_{sensamp-precharge} \quad (1)$$

According to Figure 2, the LLC bank access delay at $20^{\circ}C$ is 54% of that at $85^{\circ}C$. Figure 3 illustrates the IPC difference of the system with LLC operating delays corresponding to $20^{\circ}C$ and $85^{\circ}C$. The baseline uses SRAM delay at $85^{\circ}C$ as the worst-case design, while the ideal case keeps the SRAM delay corresponding to $20^{\circ}C$. The IPC measurements are taken over 250M cycles in the region of interest for each benchmark selected from SPLASH-2 [8] shared memory application suite. The geometric mean of the system IPC is improved by an average of 11%. $Barnes$ and $raytrace$ experience over 20% speed-up as they have a relatively lower L1 hit rate, but higher L2 hit rate. All the other applications achieve over 7% performance improvement except for $radix$ (2%). A closer look reveals that it is bounded by the memory latency, as it has the highest LLC miss rate. The results indicate the performance achievable with delay-dependent adaptation mechanisms.

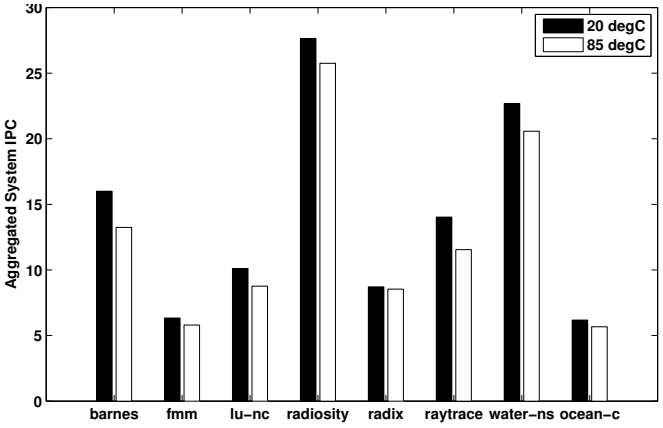

Fig. 3. System IPC comparison between $20^{\circ}C$ and $85^{\circ}C$

4 Thermal Adaptation: Algorithms

The basic idea of thermal adaptation here is to consistently convert thermal headroom into performance improvement. In this section, we discuss two LLC adaptation models in details. The sampling rate is a critical factor for both models, as we need to make sure that the SRAM timing properties does not change significantly within the sampling period. For a 2-tier 3D structure demonstrated in this paper, a typical silicon thickness is 0.03cm. The thermal time constant [9] is computed as:

$$\tau = (\frac{2 * th}{\pi})^2 \frac{\rho * C_p}{K} = (\frac{2 * 0.03}{\pi})^2 \frac{2.33 * 0.7}{1.0} = 609\mu s \quad (2)$$

where ρ, C_p and K are the density, specific heat and thermal conductivity of the silicon. The temperature change [9] in $10\mu s$ (the sampling period chosen in the simulation) for a single core with TDP of 20W is computed as:

$$\Delta T = P * R(\frac{4}{\pi^{1.5}})(\frac{t}{\tau})^{0.5} = 20 * 0.2\frac{4}{\pi^{1.5}}(\frac{10}{609})^{0.5} = 0.37^{\circ}C \quad (3)$$

where P and R are the power consumption and thermal resistance of a single core. As shown, the temperature variation within $10\mu s$ is less than $0.5^{\circ}C$.

Algorithm 1 Thermal Adaptaion Framework

```
1: update_power(core[], cachebk[]);
2: update_temperatre();
3: synchronization_barrier();
4: for i = 0 to cache.banknum-1 do
   ▷ Reduced Cycle Model
5:    cachebk[i].cycle = cycle_tbl(cachebk[i].temp);

   ▷ Partial Boosting Model
6:    new_freq = core_boost(core[i].ipc,
7:              cachebk[i].temp);
8:    if power_avail(new_freq) > 0 then
9:       core[i].freq = new_freq;
10:   end if
11: end for
12: synchronization_barrier();
```

4.1 Reduced Cycle Model (RCM)

RCM focuses on the interface between the core and adjacent LLC cache bank. The RCM algorithm reduces the number of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2015.2495125, IEEE Computer Architecture Letters

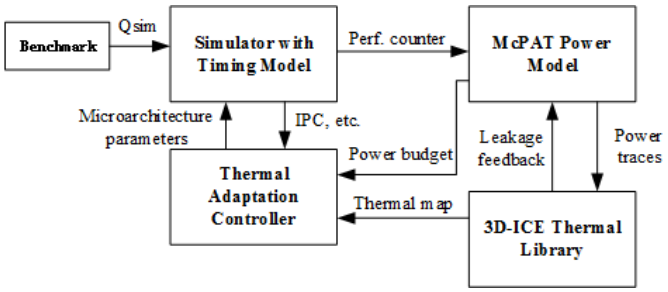

Fig. 4. System simulation framework with thermal feedback loop

TABLE 1
SPLASH-2 benchmark characterization on a 16-core machine. L1 HR - L1 hit rate. LLC MR - last-level cache miss. rate.

APP	uops	flops	memR	memW	L1 HR	LLC MR
barnes	2437M	11.9%	20.3%	15.6%	96.86%	16.97%
fmm	2624M	33.9%	18.1%	3.1%	98.13%	40.57%
lu-nc	415.9M	18.7%	21.1%	9.7%	93.55%	43.17%
radiosity	2891M	-	17.6%	10%	99.17%	17.36%
radix	325.8M	-	23.7%	13.8%	97.40%	44.65%
raytrace	719.6M	-	25.2%	9.6%	96.48%	24.70%
water-ns	675.1M	21.3%	17.6%	7.7%	98.62%	25.25%
ocean-c	665.4M	26.7%	21.6%	4.9%	93.55%	44.28%

cycles to access the bank in proportion to the temperature drop during execution, and thus improves the cache performance.

As the temperature of the cache banks does not have significant changes within the sampling period, the new bank access time in number of cycles is updated as a function of the temperature at the end of the sampling period by indexing from a pre-computed cache cycle lookup table. Support for the RCM is at the cache interface and does not affect the core.

The performance gain of RCM comes from the reduced miss penalty in the L1 cache. RCM is suitable for memory bounded applications, as the applications have more cache interactions.

4.2 Partial Boosting Model (PBM)

Unlike RCM, PBM scales up the core frequency according to the temperature of its adjacent cache bank and the power budget, and tries to boost the frequency (and therefore voltage) of a core when the vertically adjacent LLC bank temperature is low. The voltage of LLC does not change during the period to keep a constant access based on the SRAM temperature-delay curve. Compared to conventional sprinting techniques, PBM uses the LLC bank temperature as a negative feedback to prevent system degradation from overheating.

At first, the core frequency is pre-set with respect to the IPC and temperature of its associated cache bank. We construct a compact model of the upper bound on power in core and cache as a function of frequency and IPC. If the power budget (TDP minus estimated power at new frequency) is greater than 0, the core frequency will change to the new value. The maximum frequency is set to 4.5GHz to prevent system failure.

As the PBM improves the performance of cores, computational bounded applications will get more performance gain.

5 Simulation Results

5.1 Simulation Framework

The full system simulator is based on the cycle-based Manifold infrastructure [10]. Manifold boots Linux and is integrated with the Energy Introspector [11] multi-physics modeling library which includes interactions between models for energy/power (McPAT) [12] and temperature (3D-ICE) [13], as shown in Figure 4. Our baseline has all 16 cores running at 3GHz.

We characterized 8 applications from SPLASH-2 with baseline configuration as shown in Table 1. The hit rate of L1 cache and the miss rate of the last-level cache are the geometric means of the 16 cache banks.

5.2 Performance Comparison

Figure 5.a presents the IPC comparison of the SPLASH-2 benchmark. RCM has the best IPC, as its cache performance is improved. However, the IPC of PBM is worse than the baseline, as the cache miss penalty is increased as measured in *number* of clock cycle when the core boosts up.

Both RCM and PBM improve the system throughput shown in Figure 5.b. RCM speeds up the cache system by reducing the access cycle while PBM gains better throughput by boosting up the system clock. For typical computational bounded applications such as $radiosity$, PBM outperforms RCM by around 9% as more instructions can be executed from a faster core, yet for memory bounded application such as $lu\text{-}nc$, the performance of RCM is better than PBM by 5.2%.

For applications that fall in between the computational and memory bounded categories, the situation is not that straight forward. The system throughput of $barnes$ is higher than that of $radix$, yet RCM outperforms PBM in $barnes$ compared to $radix$. The reason is that the L1 hit rate of $radix$ is higher than $barnes$, so $barnes$ benefits more from improving cache performance and $radix$ gains more benefit from clock boosting.

5.3 Power/Energy Consumption

The power consumption of the adaptive system is proportional to the system performance as shown in Figure 5.c, where $radiosity$ has the highest runtime power and fmm has the lowest value. For $barnes$, $lu\text{-}nc$ and $raytrace$, the RCM power is higher than PBM as the system performance outperforms the PBM model. Generally, the PBM system consumes more power than RCM as both the core and cache run faster as shown in the other 5 applications.

Although the average power increases, the total energy consumption of the RCM and PBM reduces, as shown in Figure 5.d. The dynamic power remains constant, since the workload remains the same. However, the leakage power decreases significantly, as the total execution time shrinks, as depicted in Figure 5.e. The only exception is $radix$. $Radix$ has the highest LLC miss rate (44.65%) and its performance is constrained by the memory system. As a result, the performance improvement brought by RCM and PBM does not compensate for the power increase. The total energy is thus increased.

5.4 Energy Efficiency

The energy efficiency is measured by energy per instruction (EPI), showing the average energy for a single instruction. Figure 5.f gives the comparison results. For typical computational bounded application $radiosity$, the PBM achieves the best EPI, while typical memory bounded application $lu\text{-}nc$ gets the best EPI when applying RCM.

The LLC miss rate of $barnes$ is as small as 17%, but the energy efficiency of RCM outperforms PBM, since the L1 hit rate of barnes is relatively small. $Barnes$ is thus sensitive to the LLC cache delay. As RCM provides the best LLC performance among the three models, it also provides the best energy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCA.2015.2495125, IEEE Computer Architecture Letters

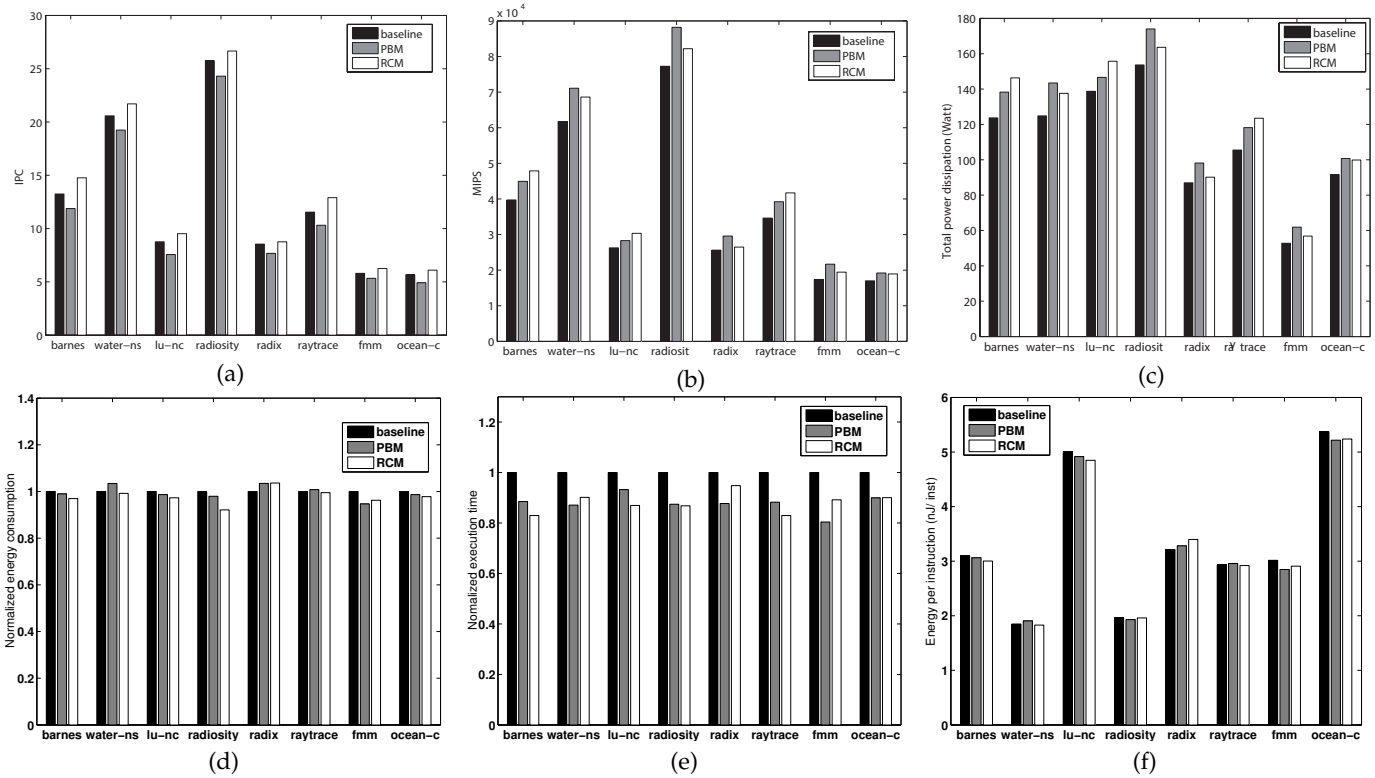

Fig. 5. The comparison of RCM and PBM running SPLASH-2 on: (a) instruction per cycle (b) system throughput in terms of MIPS (c) runtime power dissipation (d) normalized energy consumption (e) normalized execution time (f) energy efficiency in terms of EPI

efficiency. Moreover, the front-end temperature of the cores are high when running $barnes$, preventing PBM from running at a faster speed. On the contrary, fmm has LLC miss rate of 40.6%, yet PBM outperforms RCM in this application. This is because fmm contains a large amount of float point operations, and PBM provides more benefit by boosting up the cores.

The only exception in our experiments is the $radix$ application. $Radix$ is bounded by the memory latency instead of the cache latency, as it suffers from high miss rate of the last level cache and contains no float point operations. For this reason, both adaptive models will not gain much benefit in performance, and the increase in power consumption in both cases will lead to energy inefficiency.

6 Conclusion

In this paper, we argue for multi-physics as a driver for the design of 3D processors presenting a use case of a thermally adaptive LLC. Unlike previous efforts, the goal here is to consistently utilize all of the thermal headroom across the chip. Thermal headroom is a resource to be mined for performance and not a constraint to be met. We presented two thermally adaptive models for the LLC cache in a 3D stacking environment, RCM and PBM, to improve the system performance compared to conventional worst-case design operation. The RCM adapts the access time (in cycles) of the LLC cache to the temperature, while PBM modifies the core frequency based on the temperature of the vertically adjacent cache bank. Both models improves the overall system performance by over 20% and energy efficiency by up to 3%.

The thermal adaptation model presented in the paper utilizes the circuit simulator to estimate a realistic temperature-delay model to trade off between thermal headroom and performance gain. We foresee to understand theses effects across new device technologies e.g., FinFET vs. Planar, or eDRAM vs. SRAM. As the physical phenomena increasingly manifests itself at the system level, this visibility across thermal modeling, circuit behaviors and microarchitecture design will become increasingly critical to fine-grained optimizations.

References

[1] Z. W. et al, *Co-design of multicore architectures and microfluidic cooling for 3D stacked ICs*,Therminic, Berlin, German, pp. 237-242, 2013.
[2] H. Xiao et al, *Leakage power characterization and minimization in 3D stacked multi-core chips with microfluidic cooling*, SEMI-THERM, pp. 207-212, 2014.
[3] H. Xiao et al, *Multi-physics Driven Co-design of 3D Multicore Architectures*, InterPACKICNMM, 2015.
[4] Joe Jeddeloh et al, *Hybrid Memory Cube New DRAM Architecture Increases Density and Performance*, VLSIT, pp. 87-88, 2012.
[5] R. Singhal, *Inside intel next generation nehalem microarchitecture*, Intel Developer Forum, San Francisco, Mar. 2008.
[6] S. Sinha et al, *Exploring sub-20nm finfet design with predictive technology models*, DAC 49, pp. 283-288, 2012.
[7] B. Wicht et al, *Yield and speed optimization of a latch-type voltage sense amplifier*, IEEE JSSC, vol. 39, no. 7, 2014.
[8] C. Bienia et al, *PARSEC vs. SPLASH-2: A quantitative comparison of two multithreaded benchmark suites on Chip-Multiprocessors*, IISWC, pp. 47-56, 2008.
[9] *High Frequency Transistor Primer Part 3: Thermal Properties*, Hewlett-Packard Co., 5968-1410E.
[10] J. Wang et al, *Manifold: A Parallel Simulation Framework for Multicore Ssytems*, ISPASS, pp. 106-115, 2014.
[11] W. J. Song et al, *Energy Introspector: A parallel, composable framework for integrated power-reliability-thermal modeling for multicore architectures*, ISPASS, pp. 145-144, 2014.
[12] S. Li et al, *McPAT: Anf integrated power, area, and timing modeling framework for multicore and manycore architectures*, MICRO 42, pp. 469-480, 2009.
[13] A. Sridhar et al, *3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling*, ICCAD, pp. 463-470, 2010.