Thermal Requirements in Future 3D Processors

Paul Franzon,
Department of ECE, North Carolina State University, Raleigh, NC, USA, paulf@ ncsu.edu,

Avi Bar-Cohen,
Defense Advanced Research Projects Agency, Arlington, VA, USA
Avram.bar-cohen@darpa.mil

***Abstract*— This paper reports on a study in which the projected thermal load of future 3D optimized embedded computers was explored. The approach taken was to project the performance, power consumption and area of a reasonably power-efficient 7 nm, 6672 core baseline conventionally packaged ("2D") design, and 3D alternatives to this design. The 3D alternatives have improved power efficiency over the baseline 2D design, due to their reduced interconnect power consumptions and reduced processing overhead. The most efficient set of designs use more aggressive 3D-specific strategies to increase power efficiency at the expense of increased heat flux. The most efficient design is 38% more power efficient than the baseline 2D design, but has 13x the heat flux of that design. The value of that heat flux was 5.4 W/mm^2. Further optimizations increase the thermal flux even further. Specifically creating architecture optimized to floating point operations, increases the heat flux to 17 W/mm^2 while improving the computing efficiency by another 2x.**

Keywords—3DIC; 3D processor; thermal design;

I. Introduction

It is well known that 2.5D (interposers) and 3D technology provide more interconnect bandwidth with lower interconnect power consumption, thus permitting an improvement in power efficiency for any processor exploiting these technologies [1, 2, 3]. At the same time, it is widely acknowledged that one "cost" of this improved power efficiency is an increase in the heat flux that has to be dissipated. However, this tradeoff has never been quantified.

The goal of this paper is to project what this tradeoff might be for a set of plausible multi-core designs for large-scale embedded computers implemented at the 7 nm node. By "large-scale" we mean "more the one CPU chip" potentially. "Embedded" implies it is aimed at a specific range of applications "embedded" in some platform, such as radar processing.

This paper presents power efficiency vs. heat flux tradeoffs for a set of projected 2D and 3D designs at the 7 nm node. The 7 nm node was chosen on the basis that it is likely to be the last purely CMOS transistor node. It thus represents the end of the ITRS road map and is expected to be reached sometime around the turn of the decade.

A specific goal in this paper is NOT to state that high thermal fluxes will limit the exploitation of 3D but that it will justify the employment of advanced cooling paradigms, such as liquid cooling that can cope with these heat fluxes. In essence, the message is that the power advantages gained through aggressive use of 3D technologies justify the employment of advanced cooling techniques.

This paper is structured as follows. The next section describes the methodology, while section III presents the baseline designs used at 32 nm through 130 nm nodes. Section IV describes the scaling factors used to project the power, performance and area of these designs at the 7 nm node and summarizes the interconnect bandwidths assumed in these designs, and how they were arrived at. Section V presents the baseline logical organizations that were used and discusses the alternative 2D and 3.5D designs. Section VI presents and discusses the results, before concluding.

II. Methodology

This paper presents the results of a study that postulates a set of plausible 7 nm large scale multiprocessor designs based on extensions of how such a processor would be designed today. It includes the CPUs, FPUs, circuit-switched NOC, cache hierarchy and main memory but does not include non-volatile storage, or the interface to such. No unpublished optimizations, beside 3D-driven ones were explored. The steps taken to get to these designs were as follows:

1. Obtain baseline designs, together with their performance and power consumption figures, from at-hand NCSU designs, and published data. These designs were at various nodes, from 32 nm to 130 nm.
2. Determine reasonable scaling factors, for power, performance and area, so that the performance, power consumption and area of these designs could be projected at the 7 nm node.
3. Determine memory and interconnect bandwidths based on part of the Specmark benchmark suite, and memory hierarchy taper factors used in the DARPA Exascale Study [4].
4. Determine the power, performance and heat flux for a plausible 2D design based on the above projections.
5. Using known 3DIC optimizations, determine how 2.5D (interposer) and 3D technology can be used to improve power efficiency, while requiring higher heat fluxes.

III. Baselines

Baseline designs were needed for the CPU, FPU, memory and interconnect elements. A brief description of the choices, and their properties, is given in Tables I, II and III.

U.S. Government work not protected by U.S. copyright

Table I presents the baseline processor choices. The NCSU core is a fairly minimal RISC core that can issue two instructions per cycle [5]. It was chosen over commercial published designs because (1) the properties can be gauged accurately from simulation and (2) it represents the sort of stripped down core likely to be present in an energy efficient design. The baseline Floating Point Unit (FPU) was taken from reference [6]. It is an energy efficient design reported recently by IBM.

Table II presents baseline memory choices. In order to focus on power-efficient designs, fairly small cache sizes were assumed. The L1 and L2 cache properties were calculated using CACTI, an HP produced memory calculator [7]. The L3 cache was taken from published 32-nm data from IBM [8]. That design was chosen with its high energy efficiency in mind. The energy per access for the DRAMs was interpreted from data in reference [9] and the Tezzaron data sheet [10] respectively. Refresh power is generally small and was ignored in this study, since it would simply represent a small and constant overhead. Read and write powers were assumed to be the same, so as to simplify the calculations. In practice they are different from each other, but not radically so.

Table III presents assumed interconnect power properties.. When standard memories are used, it was assumed that a memory standard interface was also used. Simulations of a DDR3 interface showed it has a power efficiency of 14.5 pJ/bit (or mW/Gbps) [3]. An efficiency scaling factor of 75% was assumed for a plausible DDR4 interface. The on-chip and TSV interconnect was calculated using a normal CV^2 formulae, assuming a voltage swing of 0.6 V. Aggressive thinning of 2 µm thick wafers was assumed for the TSVs. Tezzaron uses wafers thinned this aggressively, but 25 – 50 µm thick is more likely to be seen in the short term. The results are not highly sensitive to the wafer thickness. The interposer power levels were also obtained from simulations performed at NCSU. Some sort of Network On Chip (NOC) is needed in order to support L3 cache sharing (it was assumed that one L3 cache was shared per four CPUs) and for any message passing. Data published for the Sun (Oracle) Niagara crossbar was used and extrapolated to 7 nm using the tables in the next section.

TABLE I. BASELINE LOGIC CHOICES

Component	Clock (GHz)	Performance (GFLOPS/ GOPS)	Power (mW)	Area (sq.mm)
130 nm CPU	0.25	0.5	404	2.4
32 nm FPU	1	3.6	58.5	0.04

TABLE II. BASELINE MEMORY CHOICES

	Clock (GHz)	Energy (pJ/32-bits)	Leakage (mW)	Area (sq.mm)
8 kB L1 cache	1	55	3.5	0.37
1.25 MB L2 cache	1.25	66.7	3	3.6
DRAM core	64 ns random access	144		80
Fast DRAM	6 ns for 128 bits	96		160

TABLE III. INTERCONNECT POWER PROPERTIES

Interconnect	Energy (pJ/bit)	Notes
Off-chip	10.9	Plausible DDR4 = 75% of DDR3
On-chip	0.2	per mm
TSV	0.004	2 mm barrel TSV
Interposer (7 mm)	0.6	
Interposer (14 mm)	0.8	
Crossbar	0.283	

IV. SCALING FACTORS AND BANDWIDTH CALCULATIONS

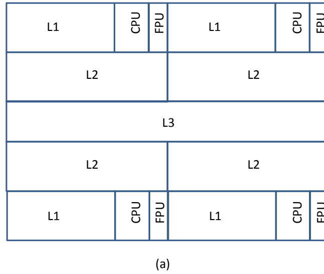
All the designs are scaled to the 7 nm node for the purposes of this study. Thus it is necessary to predict scaling factors for power, speed and area. The scaling factors presented in the ITRS are generally considered too aggressive, so "conservative" dynamic power and speed scaling factors were used as taken from Esmailzadeh et.al. [11], as modified in [12]. Leakage power scaling was taken from the ITRS. Area scaling factors could not be found. It was assumed that area scaled almost but not quite with $node^2$ (reflecting layout inefficiencies arising from multiple patterning lithography).

In order to estimate the power consumption it is necessary to know the interconnect and memory bandwidths required in practice. In particular, we need to know the hit rates on the memories and the data rates on interfaces. The hit rate on the L1 caches were obtained from a simulation run performed using the SPECMARK 2000 benchmark set performed at NCSU. The result from this simulation gave an average of 3.039 cache hits per CPU clock cycle. This was multiplied by the FPU clock rate (1.2 GHz) and 4 B/word in order to obtain the L1 memory bandwidth. Since this simulation was done on a unified cache, it was assumed that the hits were equally split between the instruction cache and data cache.

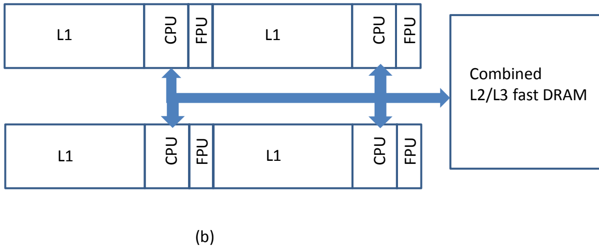
To obtain the hit rates on the other levels, miss rates of 12.5% were assumed on L1, 50% on L2, and 30% on the combined L3. These tapers were taken from the DARPA Exascale study [4] as being typical for scientific workloads and small cache sizes.

V. IMPLEMENTATIONS

Two different overall organizations were considered and these are shown in Figure 1. The first (Figure 1(a)) is a conventional organization in which each CPU/FPU has private L1 and L2 caches, and an L3 cache is shared between each four cores, and is referred to as the "CPU+L1+L2+L3" organization. The areas are drawn approximately to scale. The second (Figure 1(b)) is an unconventional organization in which a fast DRAM is used as a combined L2/L3 cache, and is referred to as the "CPU+L1" organization. This is not drawn to scale. In this organization, the fast DRAM is built in a different process technology than the core logic and thus has to be on a separate chip. Because of the high L2 bandwidth, the only practical way to build this multichip unit is with 3DIC technologies. Other technologies do not provide enough chip to chip bandwidth.



(a)



(b)

Figure 1. The two (repeating) baseline organizations considered. (a) "CPU+L1+L2+L3": A conventional organization with a 3-level SRAM cache hierarchy. (b) "CPU+L1": Using a stacked fast DRAM as a combined L2/L3 cache.

To normalize the implementations against each other, it was decided that each organization would consist of 6672 cores. This corresponds to the number of cores using the "CPU+L1" organization that fit onto one 22 x 22 mm chip. This number of cores provides a peak computing capacity of 29.3 TFLOPS. When running the Specmark benchmarks, it can only sustain an average of 320 GFLOPS, but the overall Specmark benchmark suite is not very FP intensive. Using the peak throughput number and a "typical" 0.1 B/FLOPS, this number of cores requires 2.93 TB of DRAM. Assuming 16 Gb memory chips, this corresponds to 144 DRAM chips. When using the "conventional" organization, eleven 22x22 mm chips would be required to implement 6672 cores.

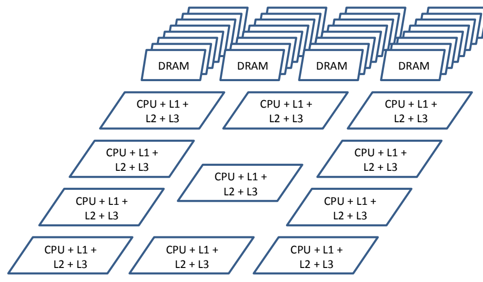
Five different physical implementations were investigated. These are summarized in Figure 2 and as follows:

A. Version A, referred to as the "2D" implementation is a conventionally packaged implementation of eleven chips consisting of arrays of CPU+L1+L2+L3 and conventional DRAMs packaged in DIMMs.
B. Version B is the most straightforward way to use 3D technologies. Referred to as "Basic Memory On Logic", it consists of the same CPU+L1+L2+L3 chips as in version A, but now they are stacked with the memories on top. 14 memories would have to be stacked with each CPU chip. The main power savings offered in this version is that the high power DDR4 memory interfaces are replaced with low power 3D and interposer interfaces.
C. Version C makes more aggressive use of 3D technologies. Referred to as "Memory on Logic on Logic", now the CPU chip is partitioned into two intimately stacked chips. Aggressive logic partitioning in this version offers potential for more than a 20% savings in the power consumed in the core logic and memories [13]. This savings is in addition to the interface power savings in Version B.
D. In Version D, referred to as "Fast DRAM on CPU + L1", the L2 and L3 caches are replaced with a Fast DRAM which has the speed of the L2 cache but the capacity of an L3 cache. It was estimated that 4-6 specialized fast DRAM chips could give the same memory capacity as the 8 Gb of L3 cache in the original design, as well as the same number of ports as the original L2 cache. By replacing SRAM with DRAM, and removing one level of memory hierarchy, significant power savings can be gained. Since DRAM is a specialized technology, it is best built on a separate substrate to the logic, requiring two or more chips. Since there is a high bandwidth to the L1 cache, the only practical way to provide that bandwidth is 3DIC vertical integration. Even an interposer could not provide enough connection capacity.

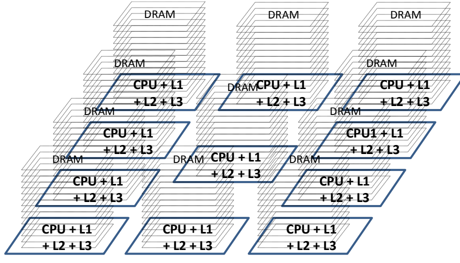
   However, there are two costs. First, the NOC is now feeding the ports of the L2 caches, instead of the L3 caches, so it is carrying more traffic. Also, now the DRAM could not be stacked with the CPU (stacking 144 DRAMs was viewed as excessive), so an off-chip interface was used. An interposer interface power level was assumed.
E. In Version E, referred to as "Fast DRAM on Split Logic", concepts in versions C and D are combined, in that the L2/L3 cache is replaced with a fast DRAM AND, the logic is aggressively partitioned in the third dimension, to obtain power savings in the CPU, FPU and cache.

VI. RESULTS

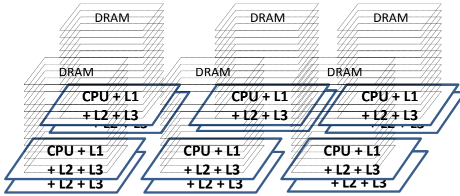
The power consumption of the different implementations above was analyzed using the assumptions outlined earlier and a spreadsheet. The core results are given in Table IV. For each implementation scenario, the following parameters are summarized:



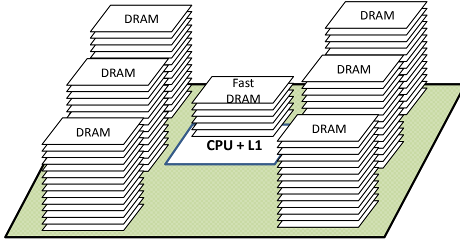
A. 2D Implementation
11 CPU chips + 144 DRAMs in conventional packaging



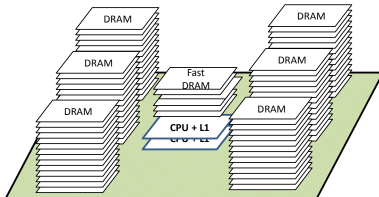
B. Basic Memory On Logic
11 CPU chips + 144 stacked DRAMs



C. Memory On Logic On Logic
6 CPU chips + 144 stacked DRAMs



D. Fast DRAM on CPU+L1
1 CPU (CPU+L1) chips + 6 stacked Fast DRAMs, With 144 DRAMs on an interposer



E. Fast DRAM on Split Logic
CPU (CPU+L1) as a stack of 2 chips + 6 stacked Fast DRAMs, With 144 DRAMs on an interposer

Figure 2. Different physical implementations investigated in this study.

- Power and power (heat) flux of the central chip stack or stacks, as appropriate.
- The internal flux between the first and second chips in the stack – this must normally be passed through the first chip in order to reach any external heatsink.
- The power flux at the worst hot-spot.
- The total power of the CPU + memories, including DRAMs.
- The "peak" power efficiency if the FPUs were operating at full capacity, 100% of the time.

TABLE IV. KEY RESULTS

	A	B	C	D	E
Chip stack power (W)	139	128	192	1068	899
Power Flux (W/mm^2)	0.29	0.26	0.45	2.2	3.7
Internal flux (W/mm^2)		0.15	0.25	0.25	2.5
Hot spot (W/mm^2)	6.5	6.6	10.6	6.7	8.8
Total Power (W)	1659	1403	1149	1200	1030
Peak Eff. (mW/GFLOPS)	56	47	39	40	35

To illustrate the distribution of power by function, this distribution is shown in Figure 3 for implementation A, the 2D implementation. This distribution is for each of the CPU+L1+L2+L3 chips. For completeness, one-eleventh of the total DRAM power is also included – representing each CPU chip's share of the DRAM power.

This figure can be used to illustrate the potential value of 3D technologies. Note that the NOC and DRAM interfaces each consume more power than the FPUs. All the 3D versions substantially reduced this power. The logic-on-logic and split-logic implementations reduced the power of CPU, FPU and L1 by 20%, while the Fast DRAM implementation reduced the power of the L2 caches. All of these are 3D specific optimizations.

Finally to illustrate the tradeoffs illustrated in this study, Figure 4 shows a scatter plot of the heat flux of the chip stack vs. the peak power efficiency for each alternative. 3D technologies are used more aggressively as we proceed left to right. It is clear that more aggressive use of 3D technologies does improve power efficiency, at the expense of increased heat flux. Moving from the 2D case to the memory-on-logic-on-logic (case C) case improves power efficiency by 31% at the "cost" of increasing heat flux by 55%. However, diminishing returns apply after that. Moving from memory-on-logic-on-logic to Fast-DRAM-on-split-logic offers a further improvement in power efficiency of 10% while needing a

further 7x increase in heat flux. However, proceeding from the 2D design to the most power efficient 3D design results in an improvement in power efficiency of 38%, at the expense of a 13x increase in heat flux.

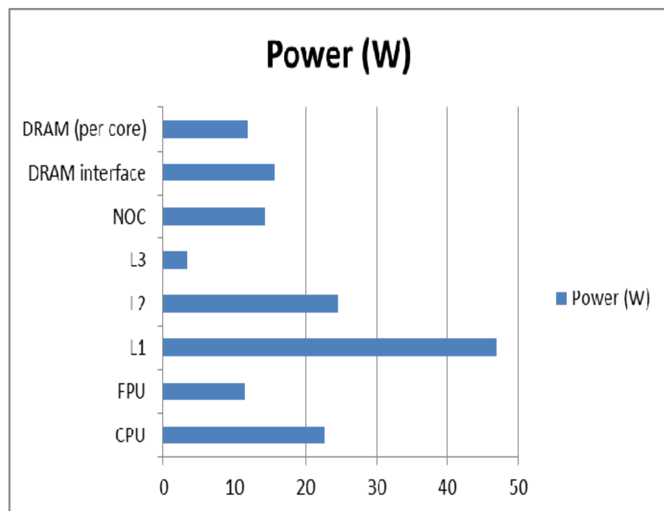


Figure 1. Distribution of power consumption for one CPU+L1+L2+L3 chip in Version A, 2D implementation. The “DRAM power” bar is that one cores share of the total DRAM power.

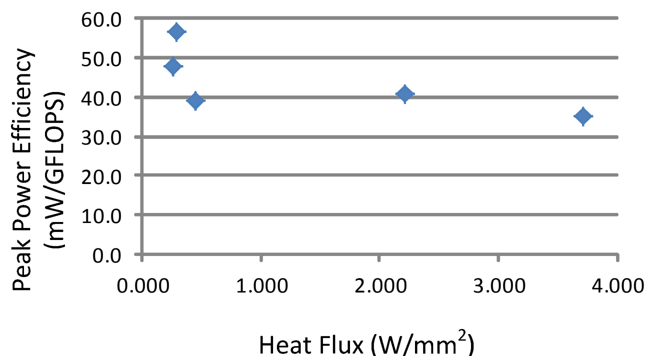


Figure 2. Scatter plot of Power Efficiency vs. Heat Flux for the different implementations.

One way to improve power efficiency is to increase the use of accelerators to perform specific ranges of task. For floating point heavy applications, such as radar algorithms, increasing the ratio of FPUs to CPUs is an appropriate way of accelerating the computing. The example above was rerun with 16 FPUs associated with each CPU. The results are presented in Table 11. This table assumes the conservative area scaling rate The results show a dramatic improvement in processing efficiency at the expense of a dramatic increase in heat flux.

TABLE V. RESULTS ASSUMING ACCELARATORS ARE USED, GREATLY INCREASING THE HEAT FLUX

	A	B	C	D	E
Chip stack power (W)	890	774	791	4988	4130
Power Flux (W/mm^2)	1.84	1.6	2.9	10.1	17
Internal flux (W/mm^2)		0.76	1.7	10.1	17
Hot spot (W/mm^2)	6.5	7.2	12	7.7	17.5
Total Power (W)	7786	8514	4746	5562	4795
Peak Eff. (mW/GFLOPS)	26	28	16	19	16

VII. CONCLUSION

In this study, a range of power efficient designs are presented for a plausible 7 nm ~6000 core, single board, embedded computer design. The baseline 2D design dissipates 1658 W with a thermal heat flux of up to 0.42 W/mm2 (42 W/cm2). A simple 3D alternative – stacking the DRAMs on top of the CPU, increases the power efficiency by 15% while actually decreasing the heat flux! However, more aggressive 3D implementations provide further improvements in power efficiency while also increasing the resulting heat flux. The most power efficient 3D design presented is 38% more power efficient than the baseline 2D design, while the heat flux is 13x that of the baseline design, at 5.4 W/mm2. Customizing the design further to optimize it to floating point operations doubles the power efficiency at the expense of a tripling of heat flux.

ACKNOWLEDGMENT

The research was supported by DARPA.

REFERENCES

[1] Davis, W.R.; Wilson, J.; Mick, S.; Xu, J.; Hua, H.; Mineo, C.; Sule, A.M.; Steer, M.; Franzon, P.D.; , "Demystifying 3D ICs: the pros and cons of going vertical," Design & Test of Computers, IEEE , vol.22, no.6, pp. 498- 510, Nov.-Dec. 2005

[2] P.Franzon, S. Priyadarshi, S. Lipa, W.R. Davis and T. Thorolfsson, “Exploring Early Design Tradeoffs for 3DIC,” in Proc. ISCAS 2013, May 2013.

[3] A. Karim, P. Franzon, A. Kumar, “Power Comparison of 2D, 2.5D, and 3D interconnect solutions and Power Optimization of Interposer Interconnect,” in Proc. IEEE ECTC 2013, May 2013.

[4] DARPA Exascale Computing Study, http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/exascale_final_report_100208.pdf

[5] E. Rotenberg, etl.al., “Rationale for a 3D Heterogeneous Mulit-core Processor,” to appear in ICCD 2013

[6] H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, R. Krisamurth, S. Borkar, "A 1.45 GHz 52-162 GFLOPS/W variable precision floating point", in Proc. ISSCC 2012, pp. 182-184

[7] CACTI, http://www.hpl.hp.com/research/cacti/

[8] J. Kuang, et.al., "The design and characterization of a half-volt 32 nm dueal-read 6T SRAM", IEE TCAS, Vol. 58, No. 9, Spe.t 2011, pp20102016.

[9] T. Vogelsang, "Understanding the energy consumptio of Dynamic Random Access Memories," in Proc. 2010, IEEE/ACM Int. Symp on Microarchitecture, pp. 363-373

[10] R.Patti, Tezzaron, private communications

[11] H. Esmailzadeh, E. Blem, R.St Amant, K. Sankaringam, D. Burger, "Dark Silicon and the End of Multicore Scaling," in IEEE Micro 2012, pp. 122-134

[12] J. Park, NCSU, Private Communications

[13] T. Thorolfsson, L Guojie, J. Cong, and P. Franzon, “Logic-on-logic 3D Integration and Placement,” in IEEE 3DIC 2010, pp. 1-4.