A Thermally-Aware Performance Analysis of Vertically Integrated (3-D) Processor-Memory Hierarchy

Gian Luca Loi, Banit Agrawal*, Navin Srivastava, Sheng-Chih Lin, Timothy Sherwood* and Kaustav Banerjee
Department of Electrical and Computer Engineering; *Department of Computer Science
University of California, Santa Barbara, CA 93106
{gianluloi, navins, sclin, kaustav}@ece.ucsb.edu; {banit, sherwood}@cs.ucsb.edu

ABSTRACT

Three-dimensional (3-D) integrated circuits have emerged as promising candidates to overcome the interconnect bottlenecks of nanometer scale designs. While they offer several other advantages, it is expected that the benefits from this technology can potentially be off-set by thermal considerations which impact chip performance and reliability. The work presented in this paper is the first attempt to study the performance benefits of 3-D technology under the influence of such thermal constraints. Using a processor-cache-memory system and carefully chosen applications encompassing different memory behaviors, the performance of 3-D architecture is compared with a conventional planar (2-D) design. It is found that the substantial increase in memory bus frequency and bus width contribute to a significant reduction in execution time with a 3-D design. It is also found that increasing the clock frequency translates into larger gains in system performance with 3-D designs than for planar 2-D designs in memory intensive applications. The thermal profile of the vertically stacked chip is generated taking into account the highly temperature sensitive leakage power dissipation. The maximum allowed operating frequency imposed by temperature constraint is shown to be lower for 3-D than for 2-D designs. In spite of these constraints, it is shown that the 3-D system registers large performance improvement for memory intensive applications.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles – *advanced technologies, VLSI.*

General Terms: Design, Performance, Reliability.

Keywords: 3D ICs, processor-memory, performance modeling, thermal analysis, three dimensional, vertical integration, VLSI.

1. INTRODUCTION

The classical solution to achieve the ever-standing requirements of faster and smaller chips has been device scaling as per Moore's Law. However, continued device scaling is slowing down due to severe short-channel effects [1], increasing variability [2, 3] and power dissipation [4]. Also, the increasing number of devices and functionality on a single chip leads to increasing complexity in interconnecting devices with a large number of metal layers. Hence, performance improvement due to device scaling cannot be fully exploited because of the degraded performance of interconnects.

Three-dimensional integration to create multiple active layer chips is a concept that can extend Moore's law in the vertical (z-direction) and promises to significantly alleviate the growing challenges of interconnects in nanometer scale VLSI circuits [5]. A true 3-D integration scheme involves monolithic stacking of multiple active layers and leads to a considerable reduction in the number and average lengths of the longest global wires seen in traditional planar (2-D) chips by providing shorter "vertical" paths for connection (**Fig. 1**). Besides the benefits of interconnect performance [5, 6], this scheme leads to increased transistor packing density, smaller chip area, lower power dissipation, and provides means to integrate dissimilar technologies (digital, analog, RF circuits, etc) in the same chip, but on different active layers.

Alongside research into developing processing technology for 3-D ICs [5, 7], several works in the literature have explored possible applications for this revolutionary technology [5, 7, 8, 9]. One of the most promising applications is that of integrating a processor-and-memory system on a single 3-D chip. In traditional 2-D design, access to the main memory (off-chip) is limited by extremely slow (< 200 MHz) off-chip buses. Moreover, the limited number of input/output pads constrains the width of these off-chip buses (typically 64 bits). On the other hand, 3-D ICs offer the unique possibility of having large memory on a single chip. As shown later in this work, the vertical on-chip buses in 3-D have a much higher frequency (~2 GHz) and also eliminate the need for external input/output pads thus removing the limitation on the width of these buses. We show that the increased bus frequency and bus width leads to significant improvement in the system performance for a memory intensive architecture.

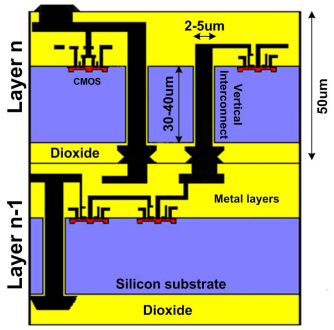

Fig. 1: Cross-sectional view of an n-layer vertically stacked active layers interconnected through vias.

However, this technology introduces new considerations in the design of high performance systems. Although it offers added flexibility in terms of bus widths and interconnect buffering schemes, the large power density (and related thermal effects) can limit the operating frequencies of such a chip. Hence, there is a pressing need to realistically quantify the performance improvement in such a system while taking thermal effects into account and re-examine traditional design paradigms in the light of this new technology.

1.1 Prior Literature and Scope of This Work

The work in [9, 10] shows that significant improvement in the

DAC 2006, July 24–28, 2006, San Francisco, California, USA.

performance of a RISC processor/cache system can be achieved with 3-D technology as compared to conventional designs. However, they claim that thermal management of such systems will not be a significant issue as their analysis shows a very small vertical temperature gradient (less than 1.3ºC) possibly due to the fact that leakage power (which is highly temperature sensitive) was not a big concern at the old 0.5 um technology node. In current technologies this does not hold true as is shown through our analysis. A similar performance analysis for 3-D architecture has been performed in [11], but this work has limited applicability because their assumptions for bus delay, memory and cache latencies are not explained clearly. These works do not fully exploit the benefits of 3-D architecture as the advantages of using wider and high-speed memory buses are not considered. Although [12] addresses such possibilities, their analysis is focused only on cache design for very specific applications using SiGe hetero-junction bipolar transistors. [12] also neglects cache misses beyond the L2 cache, which means accesses to the main memory, which often become the performance bottlenecks, are not taken into account.

Furthermore, none of the works to date addressing 3-D chip performance account for the thermal effects [13] which have a significant impact on the performance and reliability aspects of 3-D ICs. Due to the stacking of several active layers, which are all sources of power dissipation, the power density of a 3-D chip can be much higher than that of traditional chips. Moreover, the stacked active layers are far away from the heat sink (which is attached to the chip substrate of layer 1) leading to large temperature gradient in the vertical direction [13, 14]. Additionally, considering the importance of leakage power dissipation (which is exponentially dependent on temperature) in nanometer scale technologies, temperature dependent evaluation of leakage power of every active layer in a 3-D chip while taking the interlayer thermal couplings [5], [13] into account can be a key factor in determining the practical applicability of 3-D technology.

The analysis presented in this work addresses all these drawbacks in the existing literature while presenting a realistic analysis of the performance improvement that can be achieved in a processor-cache-memory system implemented in 3-D technology. Section 2 provides a detailed discussion of the possible bus configurations and associated delay modeling for interconnecting the processor layer with the vertically stacked cache and memory layers and quantifies the bus delay unlike most other works in the literature. Using this delay, we present the improvement in system performance achieved with 3-D architecture. The experiments for evaluating performance in this work encompass different memory behaviors to give a realistic picture of the improvement in performance for different applications. Section 3 describes the methodology used to evaluate the vertical thermal profile of the 3-D chip considering the thermal coupling between each layer as well as the impact of temperature on leakage power at each layer. Finally, Section 4 discusses the limits imposed by temperature on system performance. Hence this work presents a realistic comparison of 2-D and 3-D architectures considering both the advantages of performance enhancement as well as the constraints arising from thermal considerations which limit the maximum operating frequency. Concluding remarks are made in Section 5.

2. SYSTEM PERFORMANCE ANALYSIS

Although aggressive technology scaling and micro-architectural innovations have continued to improve microprocessor performance, the system performance is limited mainly because of two reasons. First, the main memory is off-chip and access latency to the memory is very high (~200 cycles). Second, the main memory bus has limited bandwidth because of the high capacitance of the external bus and the limited number of input/output (I/O) pads which limit the bus width. **Fig. 2** shows that using a 3-D architecture allows us to keep the main memory on-chip and effectively reduce the latency for accessing it. This is because the on-chip interconnections that replace the off-chip buses have much smaller delay and hence increase the memory bus frequency. Also, since there is no need for large I/O pads for the memory bus, the bus width can be increased to the maximum block size of the L2 cache so that an entire block of data can be transferred in a single clock cycle.

We consider a processor-cache-memory system with the core implemented in a standard 130 nm process, while the main memory uses 150 nm technology (3-D integration allows the possibility of integrating different technologies on the same chip [5]). The chip area is assumed to be 1 cm^2. The L1 cache (including data and instruction cache) is on-chip in the same layer as the processor for the 2-D as well as the 3-D chip. It is assumed that up to 2MB L2 cache can be implemented on-chip with the processor in a 2-D chip, while in the case of 3-D the L2 cache is on layer 2 (see **Fig. 2**). The SDRAM main memory of size 64MB is off-chip for the 2-D system, while the same SDRAM is divided into 16 layers of area 1cm^2 each in order to accommodate 64MB on the same chip. The details of the system configurations in 2-D and 3-D are shown in **Table 1** along with the memory access times for the two scenarios. The conservative timing parameters of DRAM are taken from [15] and [16] and translated into the corresponding number of DRAM cycles for both 2-D and 3-D systems.

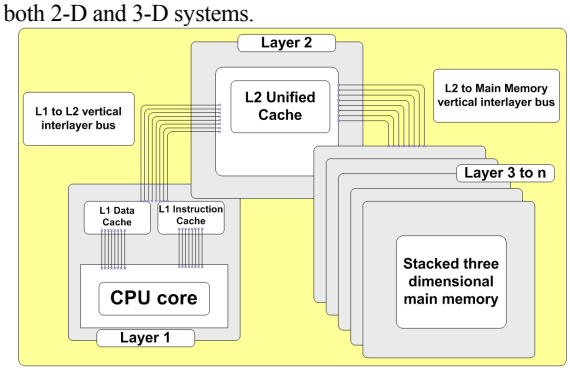

Fig. 2: Schematic view of processor-cache-memory system implemented in 3-D architecture.

Table 1: System configurations used for simulating processor-cache-memory system in 2-D and 3-D.

Alpha 21264 processor core		
Processor frequency	Frequency analysis	From 800MHz to 3GHz
	Bus analysis	2 GHz
Memory subsystem		
L1 Cache	Instruction	64KB, 4 way, 16 Byte line size 1ns latency
	Data	64KB, 4 way, 16 Byte line size 1ns latency
L2 unified cache	128 byte line size, 8 way. Fix to 1MB in freq analysis	
	Cache size (KB)	0 32 64 128 256 512 1024 2048
	Latency (ns)	0 1 1.46 1.49 1.86 3.66 4.28 4.84
Main memory	2-D design	Off-chip SDRAM 200 MHz, 30ns ras delay 30ns cas delay 20ns precharge 10ns chipset request/return
	3-D design	On-chip SDRAM 500 MHz, 30ns ras delay 30ns cas delay 20ns precharge
Bus configurations		
L1 to L2 on-chip bus	16 byte wide, core clock frequency	
L2 memory bus	2-D design	200 MHz external bus, 8 byte wide
	3-D design	Core clock frequency, 8 byte to 128 byte wide

It must be noted that this analysis assumes identical cache and memory designs for 2-D as well as 3-D architectures in order to

perform a fair comparison between the two technologies. While the memory hierarchy can be optimized (for the purpose of memory latency reduction, chip area reduction or wire length reduction) to further exploit the advantages of 3-D architecture, such considerations are outside the scope of this work. It is important to note that this analysis provides a lower bound of the performance improvement that is achievable with 3-D ICs since it does not account for the additional improvement that can be achieved with the use of architectures such as Simultaneous Multi-Threading (SMT) and multi-core chips.

2.1 3-D Vertical Bus Delay Modeling

The major advantage from the 3-D architecture is that the off-chip memory can now be brought on-chip and stacked directly on top of the processor core. In this section, we quantify the delay in a bus (connecting the processor core to the memory) formed by a 3-D vertical via configuration.

The geometry of a 3-D bus connecting successive active layers is shown schematically in **Fig. 3**. The configuration shown here (densely packed array) is preferred over the other possible configuration where all vias are spread out over the silicon area as the dense array makes it significantly easier to route the entire bus from the processor core to the memory layers, even though the capacitance of such a bus is higher. **Fig. 4** shows that in spite of the high capacitance (calculated using [17]), the RC delay of a line in such a 3-D vertical bus is comparable to that of a global interconnect in a typical microprocessor chip even for long via lengths considering up to 20 stacked layers. This is because the vertical interconnect dimensions [7] make its resistance per unit length of the 3-D via much smaller than the typical global interconnect in a 2-D chip, although its capacitance is much larger.

We now consider the design of the longest bus (worst case) that connects the devices in the microprocessor core to the memory in the top-most layer. While the longest interconnect length for a 2-D chip is twice its edge length (assuming a square footprint area), for a 3-D chip there is an added dimension in the vertical direction.

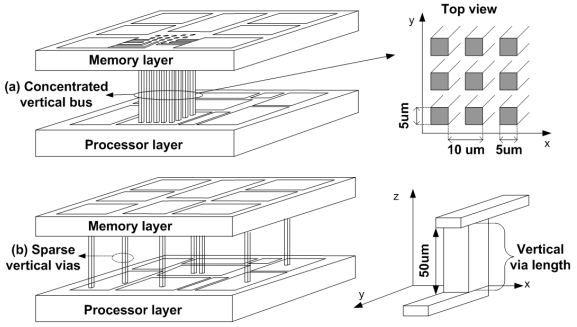

Fig. 3: Schematic view of a vertical bus interconnecting two layers stacked vertically: (a) Concentrated vertical memory bus is easier to route but has large capacitance due to small distance between adjacent vias, (b) Sparse bus has lower capacitance but makes routing complicated.

Fig. 5(a) and **(b)** depict two possible buffering options for a 3-D bus spanning a large number of layers (worst case interconnect length). The first option is to introduce buffers on the global interconnect at the core level and at the memory where the bus terminates, with no buffers in intermediate layers (**Fig. 5(a)**). Since the distance to the top-most active layer can be about 800 um (for 18 layers), the delay in the 3-D via can be large. In order to control this delay, we may introduce buffers at intermediate layers as shown in **Fig. 5(b)**. This will not only serve to reduce the bus delay but will also provide more flexibility in routing the bus through intermediate layers. A comparison between these two scenarios shows that the RC delay is nearly identical for both cases. The reason for the small difference in delay can be understood from Fig. 4 where it is observed that the 3-D via delay is nearly linear for via lengths up to 800 um. Hence the introduction of buffers for optimal delay (which serves to make delay linear with length) does not give much improvement. Hence we conclude that the choice between the two options does not have a large impact on performance. For the purpose of ease of design, the rest of this analysis assumes the via configuration shown in **Fig. 5(a)**. The bus delay in this case was found to be 0.290 ns (from A to D).

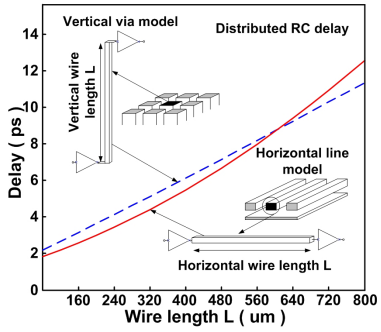

Fig. 4: RC delay comparison for horizontal global interconnect (typical dimension in a 2-D chip [1]) and vertical concentrated bus (see Fig 3(a)) of equivalent lengths. The longest length shown here (800 um) corresponds to a long stacked via spanning 20 layers. Driver and load device parameters are obtained from [1]. Chip temperature = 105ºC.

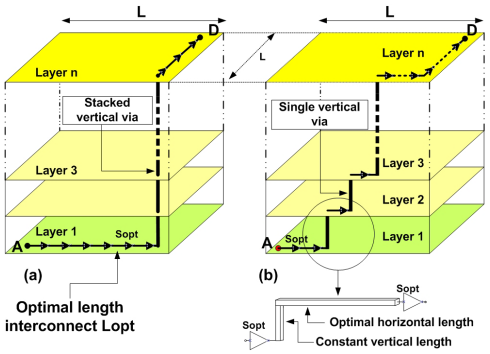

Fig. 5: Schematic showing possible buffering options for the longest interconnect spanning all the vertical layers using through vias. **(a)** Stacked via with no intermediate buffers. The horizontal sections on top and bottom layers are buffered for optimal delay (buffer size s_{opt}, inter-buffer length L_{opt}). **(b)** Staggered via with buffers inserted at each layer. Horizontal sections on top and bottom layers are buffered for optimal delay. Horizontal section lengths on intermediate layers are optimized separately for minimum delay.

2.2 Simulation and Experimental Set-up

To evaluate the performance of our selected system, we use *sim-alpha* simulator [18]. *Sim-alpha* is a cycle-approximate performance simulator for Alpha-21264, which was extended from SimpleScalar simulator [19]. While we use the baseline parameters for Alpha-21264 core, we use an updated version of Cacti (version 3.2) [20] to find the access time of L1 and L2 cache. We pick three applications (*mcf*, *parser*, *twolf*) from SPEC2000 integer benchmark suite [21] that have very different memory behaviors. In these three applications, *mcf* is the most memory intensive one, whereas *twolf* is the least memory intensive. Minnespec [22] provides a reduced version of the reference data set for SPEC benchmarks which effectively shorten the execution time from days to hours with very small difference in memory behavior. We make use of this reduced reference data set for all the three benchmarks.

2.3 Impact of Memory Bus Width and Frequency

As we already talked about some potential improvements due to on-

chip memory bus, now we characterize this improvement quantitatively for *mcf*, which is a memory intensive application. While the improvement from on-chip memory bus comes in two folds (*fast* and *wider* buses), we also investigate how this improvement may change for different L2 cache sizes, with fixed L1 instruction and data caches of 64 KB each. To understand the impact of these parameters, we find the total execution time of *mcf* application for different bus widths and different L2 cache sizes. The results are shown in **Fig. 6**. It is clear that almost all the configurations show improvement with increasing L2 cache size and bus widths. But the reduction in execution time for 2-D system is much sharper for increasing L2 cache size (especially, between 256KB-1MB). This is because the number of L2 cache-misses decrease with increasing L2 cache size, which reduces the number of accesses to the off-chip bus and off-chip main memory.

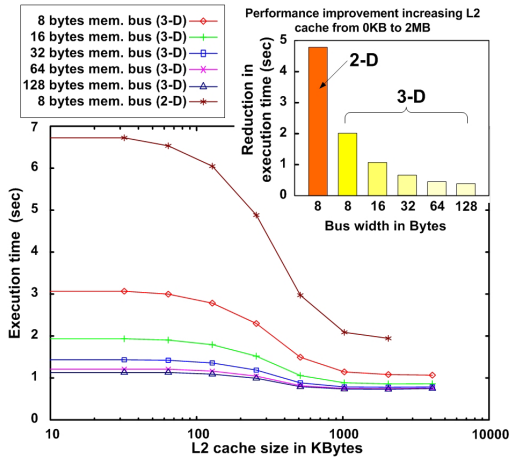

Fig. 6: Execution time as a function of cache size for 2-D system (8-byte memory bus) and 3-D system with varying memory bus width. Inset: Reduction in execution time when the L2 cache size is increased from 0KB to 2MB with 2-D system and with 3-D system for different memory bus widths.

The impact of memory bus frequency can also be seen from **Fig. 6**. When the bus width is same (8 bytes) for both 3-D and 2-D system, we see the performance gap between the two cases due to higher bus frequency in 3-D. We find that the execution time gap between 3-D system and 2-D system narrows down with increasing L2 cache sizes, but it can still provide 45% performance improvement for 1 MB L2 cache. This improvement mainly comes from the faster on-chip memory bus in 3-D system.

For a typical 1 MB L2 cache configuration, the execution time improvement over 2-D system is found to be 45% for 8-bytes 3-D system, 57.7% for 16-bytes 3-D system, and about 62-65% for 3-D system with bus width greater than 16. We also find that the effect of L2 cache size on the total execution time decreases with increasing bus width. To further illustrate this finding, we provide a bar-chart along with the figure that depicts the execution time reduction for different bus widths as we increase the L2 cache size from 0 KB to 2MB. The continual decrease in execution time in this bar-chart illustrates that memory bus can be the bottleneck for memory intensive applications. But after a certain point, the bus width does not remain the bottleneck when more percentage of the processor time is spent on executing ALU instructions. We can also see this effect in the figure as the curves for bus width of 32, 64 and 128 bytes are very close. The performance improvement shown here demonstrates that the memory bus width and frequency play a significant role in the overall performance of the system.

2.4 Impact of Processor Frequency

In this subsection, we describe the effect of processor frequency on the overall 3-D system performance. Contrary to general intuition, it is shown that increasing the processor clock speed does not necessarily render a similar increase in the system performance. Instead, it mainly depends upon the memory behavior of the selected application. For a highly memory intensive application, the overall system performance will not increase at a similar rate with increasing clock speed mainly due to the bottleneck on the memory side (memory-wall [23]). On the other hand, applications with less memory intensive behavior really reap the benefits from increasing processor clock speed.

Now, we describe how we evaluate the performance improvement of the 3-D system over 2-D system with increasing processor frequency. We keep the L2 cache size fixed to a typical value of 1 MB. We run all the three benchmarks (*mcf*, *parser*, *twolf*) for different CPU frequencies. We normalize the total execution time with the total number of instructions executed and plot the results in **Fig. 7**. This figure shows the variation of normalized instruction execution time for all the three benchmarks with processor frequency increasing from 800 MHz to 3 GHz. For all the three applications, there is a reduction in the normalized instruction execution time with increasing frequency, but the decreasing trend is much sharper in case of *parser* and *twolf*. We also see that the value of instruction execution time is much higher in *mcf* compared to *parser* and *twolf*. To better understand both the behavior, we find the percentage of memory accesses per instruction for each benchmark and list in **Table 2(a)**. As we can see from the table that the normalized memory accesses for twolf is much less than the accesses for *parser* and *mcf*. Similarly, the memory accesses for *mcf* are almost seven times more than the accesses of *parser*. The memory bottleneck and this large variation in memory access behavior explains the flatness of *mcf* curve and sharpness of *parser* and *twolf* curves with increasing frequency.

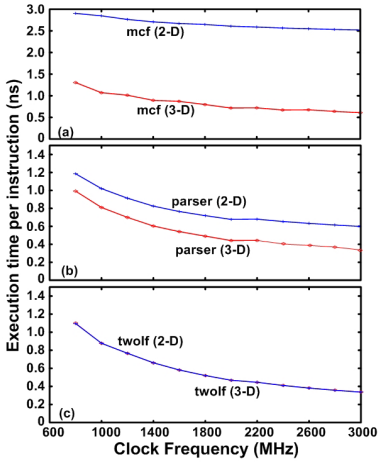

Fig. 7: Execution time per instruction for 2-D versus 3-D chip for (a) *mcf* benchmark (highly memory intensive), (b) *parser* benchmark (less memory intensive and (c) *twolf* benchmark (least memory intensive), as a function of clock frequency.

We can also observe that while the instruction execution time gap between the 3-D system and 2-D system is much larger in case of *mcf* compared to *parser*, it is almost negligible in case of *twolf*, which is again due to the different memory behavior of the applications. To understand this time gap even better, we normalize the percentage decrease in instruction execution time (when frequency increases from 800 MHz to 3 GHz) with respect to the percentage decrease in *twolf* application. Normalization is done with respect to *twolf* because it is the least memory intensive and provides the largest improvement in performance as clock frequency is

increased. We provide these results for all the three applications and for both 2-D and 3-D system in **Table 2(b)**. It can be clearly seen that increasing the frequency has the best effect on execution time in case of *twolf* (note that 2-D and 3-D show the same improvement in this case). For the same frequency increase, the 3-D system gives 9% larger improvement than the 2-D system in case of *parser* and 39% in case of *mcf*. Hence, increasing clock frequency in 3-D system can be more effective as compared to 2-D system in case of memory-bound applications rather than in case of cpu-bound applications.

Table 2: (a) Percentage of main memory accesses per instruction to characterize the three benchmarks. (b) Percentage reduction in execution time when frequency is increased from 800 MHz to 3GHz normalized to the improvement for *twolf* (benchmark which shows maximum improvement).

(a) % main memory access per instruction	
mcf	1.7%
parser	0.258472%
twolf	0.00062%

(b) % reduction in instruction execution time		
twolf	2D	100%
	3D	100%
parser	2D	77.9%
	3D	87%
mcf	2D	46.7%
	3D	85.7%

3. ELECTROTHERMAL ANALYSIS OF 3D IC

Thermal management of 3D ICs is known to be a major issue due to their high power densities [5, 13]. Thermal considerations are critical even for conventional chips because most system failures and reliability mechanisms for VLSI chips are strongly temperature sensitive [24, 25]. With the increasing power density of nanometer scale chips [1], die temperatures and on-chip thermal gradients are expected to rise substantially. The thermal problem is further aggravated by the fact that leakage power, which is known to increase significantly as technology scales, is exponentially dependent on temperature [26]. Hence rising temperatures lead to larger leakage power dissipation and vice versa. This fact is effectively captured in Equations 1-3, where T_j is average die temperature, T_{amb} is ambient temperature, P_{chip} is the total chip power dissipation and θ_j is the effective thermal resistance of the chip packaging. The constant K in Equation 2 characterizes the temperature dependence of leakage power and its value is obtained from BSIM model for 130 nm node ($\mu \cong -1.1, k \cong 0.0487\ K^{-1}$)

$$P_{active}(T) = P_{active}(T_0) \bullet (T/T_0)^{\mu} \quad (1)$$

$$P_{leakage}(T) = P_{leakage}(T_0) \bullet \exp(k(T - T_0)) \quad (2)$$

$$T_j = T_{amb} + P_{chip}(T) \bullet \theta_j \quad (3)$$

Hence, there exist strong electro-thermal couplings which must be accounted for any meaningful thermal analysis especially for nanometer scale circuits [26]. Such electro-thermal considerations also place constraints on the design space or operating conditions of circuits [27] and their thermal management [28].

While this is true for conventional (2-D) chips and 3-D chips alike, the situation gets much worse with 3-D integrated circuits. Existing packaging technologies constitute a heat sink attached to the substrate which is the primary means of cooling the chip. The process of vertical integration of active layers not only adds to the power density that needs to be dealt with by the heat sink, but also increases the distance of the additional layers from the heat sink, as shown in **Fig. 8**. Moreover, junction temperature of an upper active layer will be controlled not only by its own power dissipation, but also by the temperature of the other layers as well as the larger thermal resistance in the path towards the heat sink [5, 13].

Let us consider the system under investigation in this work. In conventional designs, the memory chips which dissipate much less power than a microprocessor core are not exposed to a high temperature environment as they are physically away from the "hot" microprocessor core. In the 3-D chip that is the subject of this study, the memory chips which are stacked on top of the micro-processor will operate at a much higher temperature.

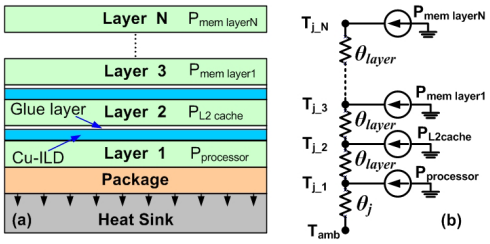

Fig. 8: (a) Schematic diagram of multiple stacked layers showing the heat sink attached to the substrate and (b) equivalent thermal circuit used for evaluating temperature profile of the 3-D stack.

In order to calculate the vertical thermal profile, the active and standby power dissipation for each layer and the thermal resistance values between active layers as well as for the entire 3-D stack are needed. We use a standard one-dimensional heat flow model [13] and equation 1-3 to calculate the layer-dependent temperature as shown in **Fig. 8(b)**. The power dissipation (active and leakage) of the processor layer (Alpha 21264) is found from [29] which provides the power dissipation of an Alpha 21264 microprocessor at 350 nm technology node by assuming a full scaling scheme down to 130 nm node. The active power dissipation of the L2 cache is found from Cacti [20], while leakage power is calculated using [30]. The active power dissipation for 64MB SDRAM main memory is obtained from [15], while the leakage power dissipation is calculated using the methodology outlined in [31] and normalized to the area suggested in [1]. The geometry for the stacked layers in the 3-D chip is also shown in **Fig. 8(a)**. Each stacked layer is assumed to have 50um thickness [32] including 5 metallization layers and a thin polyimide glue layer. Assuming HSQ dielectric (thermal conductivity K_{th} = 0.4 W/mK), Cu metal (K_{th} = 385 W/mK) with 50% metallization density, and interconnect dimensions from ITRS [1] specifications for 130 nm technology node, the effective thermal resistance for the back-end corresponding to each layer is θ_{CuILD} = 0.048 K/W, while that for the thinned Si active layers is θ_{Si} = 0.0037K/W. The effective thermal resistance corresponding to each layer (see **Fig. 8(b)**) is given by:

$$\theta_{layer} = \theta_{CuILD} + \theta_{Si} \quad (4)$$

The typical value of thermal resistance between junction and ambient for a high performance microprocessor (0.7K/W [26]) is used and ambient temperature is 45°C.

4. IMPACT OF THERMAL CONSTRAINT ON PERFORMANCE

While 3-D integration of the processor-memory hierarchy can lead to significant improvements in the performance of the system as described in Section 3, the high power dissipation of the processor (layer 1) can lead to significant temperature rise on the upper layers which are further away from the heat sink. Hence thermal constraints may impose stricter limits on the performance realizable in 3-D designs. This section combines the performance analysis of Section 3 with the thermal model described in Section 4 to show the realistic system performance gains that can be achieved with this 3-D design while taking into account the limitations imposed by thermal considerations.

As shown in **Fig. 9**, the temperature rise in the case of the 3-D chip is significantly higher than that in a 2-D chip. This is because of the higher power density as well as the large thermal resistance from the

upper layers to the heat sink. The temperature constraint on a chip (which arises from reliability concerns) limits the maximum operating frequency of the system. For the memory intensive system shown in **Fig. 9(a)**, even though thermal considerations place a lower limit on the operating frequency in 3-D, better performance can still be achieved as compared to the 2-D chip running at a higher frequency. This is because the 2-D system cannot overcome the memory interface bottleneck. On the other hand, for applications that are not memory intensive (**Fig. 9(b)**), the system performance is not dominated by memory accesses. Hence, in this case, the 2-D chip, which has a higher limit of operating frequency, has a system performance better than the 3-D chip which is constrained to operate at a lower frequency.

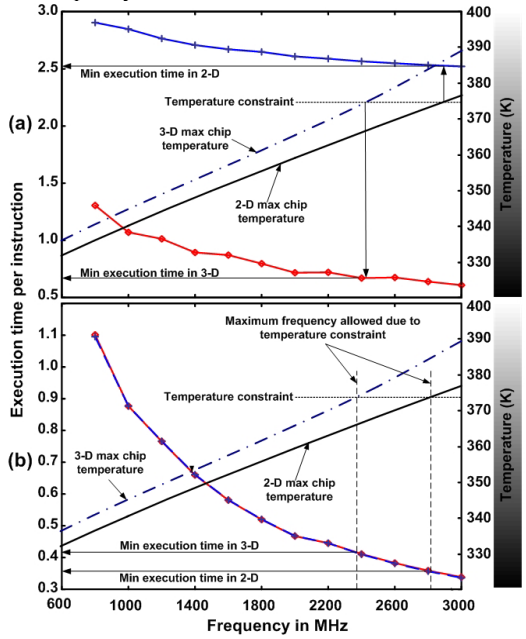

Fig. 9: Execution time per instruction and maximum chip temperature as a function of operating frequency for 2-D and 3-D chip for (a) highly memory intensive (*mcf*) application and (b) less memory intensive (*twolf*).

5. CONCLUSION

A realistic processor-memory hierarchy implemented in 3-D technology has been analyzed from a performance perspective. A detailed derivation of the memory bus delay has been presented which shows tremendous improvement in bus delay as compared to the off-chip memory bus in the case of planar (2-D) chips. Even for a large number of stacked layers (18 layers), our RC analysis shows that the bus delay is only 0.29 ns. By simulating carefully chosen applications with different memory behaviors, it is found that the major advantage of 3-D integration is the large bandwidth of the (on-chip) memory bus. It is also found that due to large bandwidth of memory bus, there is minimal performance improvement on increasing L2 cache size in a 3-D chip. Although 3-D architecture eliminates the bottleneck of memory bus and increases overall system performance, the thermal analysis shows that chip temperature rise places a much tighter constraint on the operating frequency. Most interestingly, it is found that in spite of the lower operating frequency of a 3-D chip (as imposed by thermal concerns), the overall system performance can still be significantly better than conventional planar designs, especially for memory intensive applications.

ACKNOWLEDGEMENTS

This work was supported in part by a seed grant from the Boeing Corporation, UC-MICRO/Intel grant, NSF Grant CNS-0524771, and NSF Career Grant CCF-0448654.

REFERENCES

[1] International Technology Roadmap for Semiconductors (ITRS), 2004 edition, (http://public.itrs.net/)
[2] P. Gelsinger, 41st DAC Keynote, *Design Automation Conference*, 2004. (http://www.dac.com)
[3] S. Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture," *DAC*, 2003, pp. 338-342.
[4] S. Borkar, "Low-Power Design Challenges for the Decade," *ASP-DAC*, 2001, pp. 293-296.
[5] K. Banerjee et al. "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proc of the IEEE*, Vol. 89, pp. 602– 633, 2001.
[6] A. Rahman and R. Reif, "System-level Performance Evaluation of Three-dimensional Integrated Circuits", *TVLSI*, Vol. 8, pp. 671-678, 2000.
[7] W. R. Davis, et al., "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *IEEE Design & Test of Computers,* Vol. 22, pp. 498-510, 2005.
[8] A. Zeng et al., "First-Order Performance Prediction of Cache Memory with Wafer-Level 3D Integration," *IEEE Design & Test of Computers,* Volume 22, Number 6, pp. 548 – 555, 2005.
[9] S.A. Kuhn et al., "Performance modeling of the interconnect structure of a three-dimensional integrated RISC processor/cache system," *IEEE Trans. CPMT, Part B: Adv. Packag.*, Vol. 19, pp. 719–727, 1996.
[10] M.B. Kleiner et al., "Performance improvement of the memory hierarchy of RISC-systems by application of 3-D technology, "*Trans. CPMT, Part B: Adv. Packaging*, Vol. 19, pp. 709 – 718, 1996.
[11] C. Liu et al., "Bridging the Processor-Memory Performance Gap with 3D IC Technology, "*IEEE Design & Test of Computers,* Vol. 22, pp. 556 – 564, 2005.
[12] P. Jacob et al., "Predicting the Performance of a 3D Processor-Memory Chip Stack, "*IEEE Design & Test of Computers,* Vol. 22, pp. 540 – 547, 2005.
[13] S. Im and K. Banerjee, "Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High Performance ICs," *IEDM, 2000*, pp. 727-730.
[14] A. Akturk et al., "Self-Consistent Modeling of Heating and MOSFET Performance in 3-D Integrated Circuits," *IEEE TED,* Vol. 52, pp. 2395–2403, 2005.
[15] Micron™ 128Mb synchronous DRAM datasheet, 2001.
[16] V. Cuppu et al., "A Performance Comparison of Contemporary DRAM Architectures," *ISCA*, 1999, pp. 222-233.
[17] K. Nabors and J. White, "FastCap: A Multipole Accelerated 3-D Capacitance Extraction Program", *IEEE TCAD*, Vol. 10, pp. 1447-1459, 1991.
[18] R. Desikan et al., "Sim-alpha: a validated execution driven alpha 21264 simulator," *Technical Report TR-01-23*, Department of Computer Sciences, University of Texas at Austin, 2001.
[19] D. Burger and T. M. Austin. "The SimpleScalar Tool Set Version 2.0," *Technical Report 1342*, Computer Sciences Department, University of Wisconsin--Madison, 1997.
[20] P. Shivakumar and N. Jouppi. "CACTI 3.0: *An integrated cache timing*, power and area model," *Technical Report*, Compaq WRL, 2001.
[21] J. L. Henning, "SPEC CPU2000".
[22] A. J. KleinOsowski and D. J. Lilja, "MinneSPEC: A new SPEC benchmark workload for simulation-based computer architecture research," *IEEE Computer Architecture Letters*, Vol. 1, 2002.
[23] W. A. Wulf and S.A. McKee, "Hitting the Memory Wall: Implication of the Obvious," *ACM Computer Architecture News*, 23(1):20--24, 1995.
[24] K. Banerjee et al., "Analysis and Optimization of Thermal Issues in High-Performance VLSI," *ISPD*, 2001, pp. 230–237.
[25] A. H. Ajami et al., "Modeling and Analysis of Non-Uniform Substrate Temperature Effects on Global ULSI Interconnects," *TCAD*, Vol. 24, pp. 849-861, 2005.
[26] K. Banerjee et al., "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," *IEEE IEDM*, 2003, pp. 887-890.
[27] S-C. Lin et al., "A Thermally Aware Methodology for Design-Specific Optimization of Supply and Threshold Voltages in Nanometer Scale ICs," *ICCD*, 2005, pp. 411-416.
[28] S-C. Lin et al., "Analysis and Implications of IC Cooling for Deep Nanometer Scale CMOS Technologies," *IEEE IEDM*, 2005, pp. 1041-1044.
[29] M. K. Gowan et al., "Power Considerations in the Design of the Alpha 21264 Microprocessor," *DAC*, 1998, pp. 726-731.
[30] Y. Zhang et al., "HotLeakage: A Temperature-Aware Model of Subthreshold and Gate Leakage for Architects," *Technical Report CS-2003-05*, Dept. of Computer Science, Univ. of Virginia, 2003.
[31] Y. Joo et al., "Energy Exploration and Reduction of SDRAM Memory Systems," *DAC*, 2002, pp. 892–897.
[32] P. Benkart et al., "3D Chip Stack Technology Using Through-Chip Interconnects," *IEEE Design & Test of Computers,* Vol. 22, pp. 512-518, 2005.