Exploring Serial Vertical Interconnects for 3D ICs

Sudeep Pasricha

Department of Electrical and Computer Engineering
Colorado State University, Fort Collins, CO
sudeep@engr.colostate.edu

ABSTRACT

Three-dimensional integrated circuits (3D ICs) offer a promising solution to overcome the on-chip communication bottleneck and improve performance over traditional two-dimensional (2D) ICs. Long interconnects can be replaced by much shorter vertical through silicon via (TSV) interconnects in 3D ICs. This enables faster and more power efficient inter-core communication across multiple silicon layers. However, 3D IC technology also faces challenges due to higher power densities and routing congestion due to TSV pads distributed on each layer. In this paper, serialization of vertical TSV interconnects in 3D ICs is proposed as one way to address these challenges. Such serialization reduces the interconnect TSV footprint on each layer. This can lead to a better thermal TSV distribution resulting in lower peak temperatures, as well as more efficient core layout across multiple layers due to the reduced congestion. Experiments with several 3D multi-core benchmarks indicate clear benefits of serialization. For instance, a 4:1 serialization of TSV interconnects can save more than 70% of TSV area footprint at a negligible performance and power overhead at the 65nm technology node.

Categories and Subject Descriptors:

B.7.1 [**Integrated Circuits**]: Types and Design Styles—*Advanced technologies, VLSI*;

General Terms: Performance, Design

Keywords: 3D ICs, Serial Interconnect, Networks on Chip, VLSI

1. MOTIVATION

In recent years, the rapid scaling of semiconductor technology has led to more and more processing cores and memories being integrated on a single chip. These highly integrated chip multiprocessors (CMPs) have provided the high performance needed for supporting complex emerging applications particularly in the multimedia and networking domains. However, these planar CMP architectures are now facing fundamental challenges due to on-chip interconnects not scaling well with technology [1][9]. Interconnect delay has increased significantly compared to gate delay in ultra deep submicron (UDSM) technologies as a result of increased crosstalk coupling noise and parasitic resistivity [2]. According to the International Roadmap for Semiconductors (ITRS) [3], delay on global interconnects has become a major source of performance bottlenecks and is one of the semiconductor industry's topmost challenges.

One promising solution to overcome the interconnect bottleneck and continue the pace of growth of CMP systems is the use of three-dimensional (3D) integration, in which multiple active device layers are vertically stacked and interconnected [4]-[7]. Such 3D ICs not only allow more cores to be integrated on a chip, but also provide potential performance advances, as each core can access a greater number of nearest neighbors, and thus has a higher supportable communication bandwidth. Most importantly, since the inter-layer distance is small, there is an opportunity to replace long global interconnects (~several mm) between communicating cores in a horizontal plane, by stacking the cores on adjacent layers and connecting them with shorter vertical interconnects. Since wire delay depends on the square of the wire length (or has a linear dependence if repeaters are used), this results in a reduction in inter-core latency. Wire length reduction in 3D ICs also translates into lower power dissipation in interconnects and repeaters.

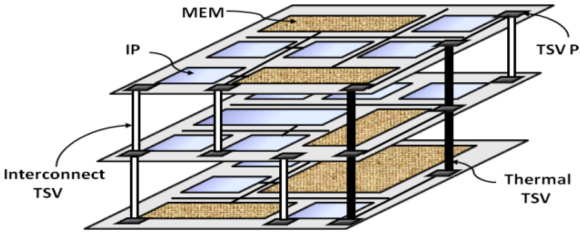


Figure 1. 3D IC with three layers

A number of different 3D IC manufacturing technologies have been explored in recent years, including transistor stacking, die-on-wafer stacking, chip stacking, and wafer stacking. Out of these, wafer stacking is one of most promising high-performance yet inexpensive implementation technology for 3D ICs [7], and the focus of this paper. Figure 1 shows an example of such a 3D IC with three active silicon layers. Each layer can have potentially multiple processing cores (IPs) and memories (MEM). Inter-layer communication is facilitated by vertical through silicon via (TSV) interconnects. A major concern in the adoption of 3D ICs is the increased power densities that can result from placing a core on top of another core. Since high peak temperatures due to increasing power densities can cause catastrophic IC failure and this is already a major concern in 2D architectures, the move to 3D will accentuate the thermal problem. Strategically placed thermal TSVs (Figure 1) can help establish a thermal path from the core of a chip to the heat sink and are one possible solution for cooling 3D ICs.

Traditional 2D network-on-chip (NoC) topologies such as mesh, torus and butterfly will most likely be extended to create scalable networks that can handle intra- and inter-layer communication requirements in 3D ICs [8]-[10]. As the number of cores increases in each layer to support rising application complexity, the amount of communication between layers is also expected to increase. This will lead to an increase in the number of interconnect TSVs. Since each interconnect TSV requires a pad for bonding to a wafer layer, this will lead to an interesting scenario where the area footprint of interconnect TSVs in each layer can no longer be ignored.

DAC'09, July 26-31, 2009, San Francisco, California, USA
Copyright 2009 ACM 978-1-60558-497-3/09/07....5.00

As an example, consider a 3D NoC with a hundred 64-bit vertical TSV links between layers (for CMPs with hundreds of cores, such a large number of vertical links is to be expected). Assuming TSV pad dimensions of 10μm×10μm and a pitch of 16μm, the TSVs will take up an area of approximately 1.6mm^2 in each layer, which is equivalent to the size of a computation core! Unlike a computation core however, these interconnect TSVs are spread out (uniformly or non-uniformly) in each layer, which will make floorplanning and routing extremely challenging. Furthermore, the need to use thermal TSVs (which can take 10-20% of total chip area [11][12]) to create thermal-efficient 3D ICs will lead to an even greater TSV footprint on each layer. This will further complicate efficient chip layout, and overall performance.

In this paper, serialization of TSV interconnects is proposed to overcome the abovementioned challenge for 3D ICs. Serialization of TSV interconnects will have the benefit of reducing the number of TSV interconnects and line drivers, which in turn will reduce the TSV interconnect area footprint in each layer. This will make it easier to obtain a more efficient chip layout. Additionally, since TSV density is limited by fabrication cost factors, fewer interconnect TSVs can make way for more thermal TSVs, which will lead to more thermal-efficient IC designs. To the best of the authors' knowledge, this is the first work to explore the impact of using serial vertical interconnects in 3D ICs. Experimental results with several CMP applications indicate that using serial TSV interconnects can significantly reduce TSV area footprint, at a negligible performance and power overhead.

2. RELATED WORK

In the last few years, there has been a growing interest in 3D ICs from academia and industry as a means to alleviate the interconnect bottleneck problem currently facing 2D ICs. IBM [4] [5] and Tezzaron [6] have recently presented promising preliminary results and test chips with 3D IC technology. Here, research on 3D ICs and on-chip serialization is briefly reviewed.

Several researchers have proposed thermal-aware floorplanning techniques for 3D ICs [12]-[16]. In particular, [14] proposed inserting thermal vias to reduce temperature hotspots during floorplanning. A few researchers have explored interconnect architecture design for 3D ICs [8][10][17]. In [8], 2D mesh and 2D folded torus topologies were compared with 3D mesh and 3D stacked mesh topologies. It was shown that 3D NoCs have more complex switches but offer better performance and lower energy for communication. In [10] a hybrid bus-NoC 3D interconnect architecture was proposed. In [17] circuit level models for TSVs were explored. Some recent work has looked at decomposing cores (processors [18][19], NoC routers [20], and on-chip cache [21]) into the third dimension which allows reducing wire latency at the intra-core level, as opposed to the inter-core level.

Serialization has been explored for long parallel on-chip global interconnects as a way to overcome UDSM artifacts such as crosstalk, wiring congestion, skew and high power dissipation [22]-[30]. In [22][23] it was shown that serial links can reduce communication area and power dissipation for not only long global links but also for shorter links in future technologies. In [24][25] a serial bus architecture was proposed to reduce on-chip bus energy. In [26]-[28] high speed ring oscillators and shift registers for communication serialization were described. In [29][30] fast asynchronous serial links were explored. None of the above works have explored the impact of using serial vertical interconnects in 3D ICs. As will be shown later, vertical TSV serialization is a powerful means of reducing TSV area footprint, and thus improving 3D IC cost, routability, and thermal efficiency.

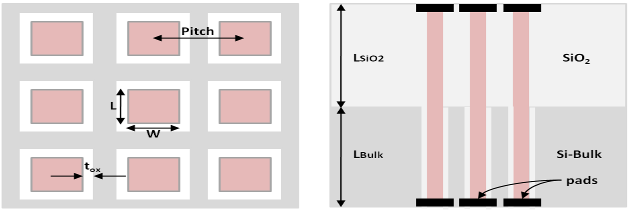


Figure 2. 3D TSV bundle schematic

3. VERTICAL TSV OVERVIEW

3.1 Interconnect TSVs

Vertical interconnects implemented as through silicon vias (TSVs) provide the highest interconnect bandwidth in 3D ICs, compared to wire bonding, peripheral vertical interconnects, and solder ball arrays. To be useful in 3D CMP communication (in NoCs for instance), TSV interconnects should not be used in isolation, but rather as a group or bundle. Figure 2 shows a bundle of 9 TSVs placed in a 3×3 grid structure, in the bulk-silicon manufacturing technology [17]. Each TSV has a length *L*, width *W*, and an oxide coating of thickness t_{ox} around it in the bulk-silicon. Pads on the wafer surfaces are needed to bond to the vertical TSVs, using mechanical thermo-compression [31]. Typically, the pads tend to be larger than via cross-section to account for the oxide coating. For instance, for a via cross section of 4μm×4μm, and an oxide thickness of 1μm, the pad thickness has to be at least 5μm×5μm.

Interconnect TSVs have inherent resistance, capacitance and inductance, and can be modeled as RLC interconnects. As discussed in [7], the impact of inductance on delay and power dissipation for frequencies of a few GHz can be ignored for vertical TSV interconnects. In such a scenario, TSV interconnect resistance can be described as a function of via length l, cross-section σ and resistivity ρ:

$$R = \frac{\rho \times l}{\sigma} \qquad (1)$$

For example, a copper TSV with a 4μm×4μm cross section has a resistance of about 1.18mΩ/μm in the 130nm technology node. The skin effect for TSVs at frequencies of a few GHz is negligible. Since TSVs are interconnected using metal bonding, an appropriate contact resistance must also be considered. In this work, a contact resistance of 100mΩ per layer is considered [31].

The capacitance of TSVs must account for coupling between adjacent TSVs in a bundle. The following capacitance matrix is used for this purpose:

$$\tilde{C} = \begin{pmatrix} C_{1,1} & -C_{1,1} & \dots & -C_{1,n} \\ -C_{2,1} & C_{1,1} & \dots & -C_{1,n} \\ \dots & \dots & \dots & \dots \\ -C_{n,1} & -C_{n,2} & \dots & C_{n,n} \end{pmatrix} \qquad (2)$$

where the elements outside of the diagonal represent inter-via coupling (with inverted signs), while the elements along the diagonal are the sum of the capacitances towards the ground plane ($C_{i,0}$ not explicitly shown in the matrix) plus the coupling capacitances:

$$C_{ii} = C_{i,0} + C_{i,1} + \dots + C_{i,i-1} + C_{i,i+1} + \dots + C_{i,n} \qquad (3)$$

The capacitance for TSVs in bulk-silicon was extrapolated from extraction results in [17] for TSV bundle densities corresponding to commonly used on-chip bus/link sizes (32, 64, 128 bits).

3.2 Thermal TSVs

In addition to acting as vertical interconnects, TSVs can also be used in a non-electrical capacity to conduct heat and alleviate hot spots in 3D ICs [12][14][33]. The idea of using thermal TSVs to overcome thermal problems was first utilized in the design of packaging and printed circuit boards (PCBs). Lee et al. [34] studied arrangements of thermal vias in the packaging of multichip modules (MCMs) and found that as the size of thermal via islands increased, more heat removal was achieved but less space was available for routing. This observation holds for 3D ICs as well, where thermal problems are greater than in 2D ICs because of the many dielectric layers. Thermal TSVs create efficient thermal conduits and greatly reduce chip temperatures, but have been shown to take up to 10-20% of total chip area to achieve a reduction in maximal chip temperature of up to 47% [11][12].

3.3 TSV Reliability Issues

Since TSV fabrication technology is not yet mature, the reliability of TSV interconnects is expected to be a limiting factor for 3D IC performance and yield for the near future [7]. Unsuccessful wafer alignment prior to and during the wafer bonding process is one of the primary mechanisms of failure for TSVs. To improve yield, hardware redundancy is often used. A simple and effective way to add redundancy and improve yield is to use double pads [35]. Since misalignments are caused by the unavoidable shift of bonding pads with respect to their nominal position, using large square pads twice as wide as standard pads can improve misalignment tolerance by an order of magnitude [31][36].

4. TSV SERIALIZATION SCHEME

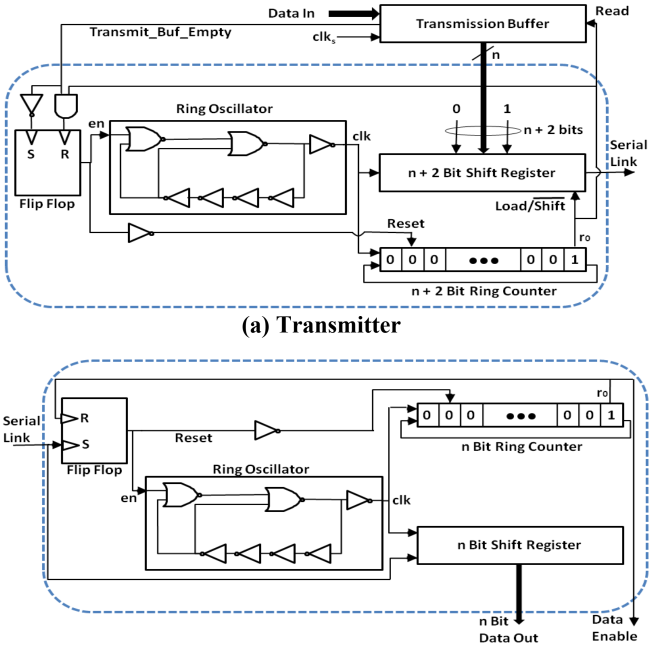
To reduce the number of interconnect TSVs in 3D ICs, a shift-register based serialization scheme is proposed, similar to [26]-[28]. A single serial line is used to communicate both data and control signals between the source and destination nodes. A frame of data transmitted on the serial line using this scheme consists of *n*+2 bits, which includes a start bit ('1'), *n* bits of data, and a stop bit ('0').

Figure 3(a) shows the block diagram of the transmitter (or serializer) at the source. When there is no transmission, the output of the flip-flop is zero, the ring oscillator and shift registers are disabled, and the *n*+2 bit counter is in the reset state, with a '1' in its least significant bit output (*r0*). When a word becomes available for transfer in the transmission buffer, the R-S flip-flop is enabled, thereby enabling the ring oscillator, which generates a local clock signal and can oscillate above 2 GHz to provide high transmission bandwidth. At the first positive edge of this clock, an *n*+2 bit data frame is loaded in the shift register. In the next *n*+1 cycles, the shift register shifts out the data frame bit by bit. The stop bit is eventually transferred on the serial line after *n*+2 cycles, and *r0* becomes '1'. At this time, if the transmission buffer is empty, the ring oscillator and shift registers are disabled, and the serial line goes into its idle state. Otherwise, the next data word is loaded into the shift register from the transmission buffer on the next positive clock edge. Thus data transmission continues without interruption.

Figure 3(b) shows the block diagram of the receiver (or de-serializer) at the destination. The R-S flip-flop in the receiver is activated when a low-to-high transition is detected on the input serial line (the 'low' corresponds to the stop bit of the previous frame, while the 'high' corresponds to the start bit of the current frame). After being activated, the flip-flop enables the receiver ring oscillator (which has a circuit similar to the transmitter ring oscillator) and the ring counter. The *n*-bit data word is read bit by bit from the serial line into a shift register, in the next *n* clock cycles. Thus, after *n* clock cycles, the *n* bit data will be available on the parallel output lines, while the least significant bit output of the ring counter (*r0*) becomes '1' to indicate data word availability at the output. With the assertion of *r0*, the R-S flip-flop is also reset, disabling the ring oscillator. At this point the receiver is ready to start receiving the next data frame. Note that in case of a slight mismatch between the frequencies of the transmitter and receiver ring oscillators, correct operation can be ensured by adding a small delay in the clock path of the receiver shift register.

The preceding discussion assumed *n:1* serialization, where *n* data bits are transmitted on one serial line (i.e., a serialization degree of *n*). If wider links are used, this scheme can be easily extended. For instance, consider the scenario where 4*n* data bits need to be transmitted on four serial lines. In such a case, the number of shift registers in the transmitter must be increased from 1 to 4. However the control circuitry (flip-flop, ring oscillator, ring counter) can be reused among the multiple shift registers and remains unchanged. At the destination, every serial line must have a separate receiver to eliminate jitter and mismatch between parallel lines.

Note that it is possible to modify the proposed scheme by using an additional strobe line for synchronization, similar to [26]. This would reduce the number of bits transferred on the serial data line, and the sizes of the shift register and ring counter (which must be rewired to disable the oscillator after the desired number of clock cycles have been generated) from *n*+2 to *n*. However, for the case of TSV serialization, the overhead of the extra TSV line and pads are prohibitive enough to overshadow the very slight improvement in energy and performance that may be obtained for this case.



(a) Transmitter

(b) Receiver

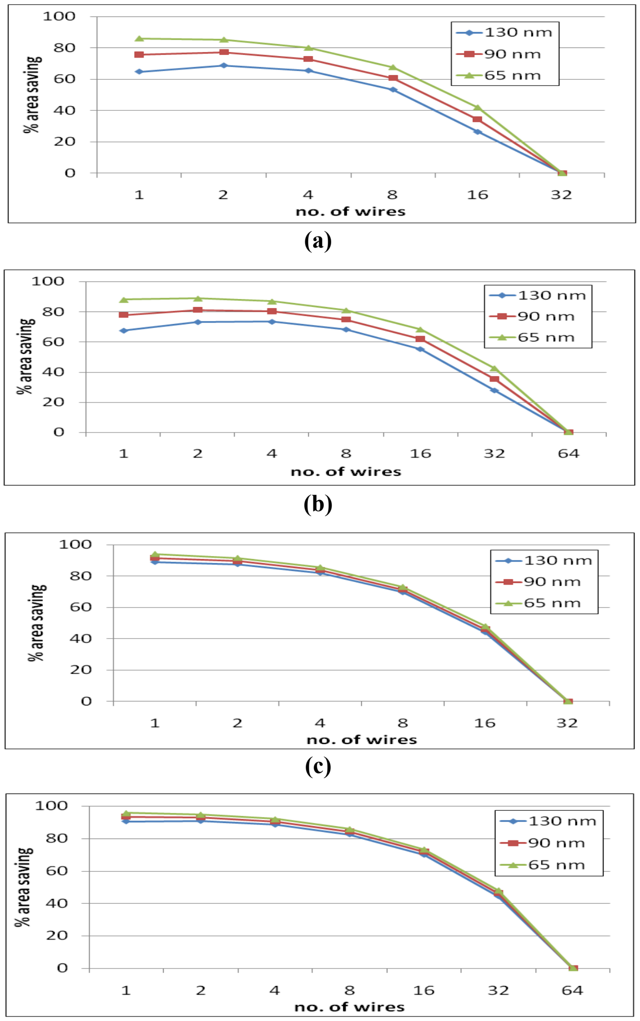
Figure 3. TSV Serialization scheme

5. EXPERIMENTS

To explore the impact of using the proposed serialization scheme for vertical TSV interconnects in 3D ICs, several experiments were conducted. First the area overhead of using the proposed scheme was explored. Subsequently, the power and performance impact of using the serialization scheme was explored, in the context of several CMP applications.

5.1 Impact on Area

The goal of the first experiment was to quantify the impact of using serialization on TSV area footprint. The serialization scheme was implemented at the RTL level and synthesized down to a gate-level net-list using Synopsys Design Compiler [38]. The synthesis was performed for the 130, 90, and 65nm TSMC standard cell libraries. This enabled a determination of the area overhead of the transmitter and receiver logic used in the serialization scheme.



(a)

(b)

(c)

(d)

Figure 4. % area saving with serialization for 130–65 nm (a) 8µm TSV pitch, 32-bit link, (b) 8µm TSV pitch, 64-bit link, (c) 16µm TSV pitch, 32-bit link, (d) 16µm TSV pitch 64-bit link

Figure 4(a) and (b) show the area savings when using the proposed serialization scheme for a 4µm×4µm interconnect TSV cross section with 1µm oxide thickness (i.e., 5µm×5µm TSV pads) and an 8µm pitch. Figure 4(a) shows the area savings for a 32-bit link. The x-axis shows the number of wires in the link with varying degrees of serialization and the y-axis shows the percentage savings in area compared to the base case (32 wires) with no serialization used. It can be seen that the proposed serialization scheme can significantly reduce the area footprint of TSVs in 3D ICs. Even a 4:1 serialization can result in as much as a 55-70% area savings. As technology scales, the savings in area increase, as can be seen with the curves for the 130, 90, and 65nm implementations. This is a result of lower area footprint of the transmitter and receiver logic with shrinking sizes of gates with technology. A similar analysis for area saving is presented for a 64-bit link in Figure 4(b).

An interesting observation from both Figure 4(a) and (b) is that the saving in area reaches a critical point beyond which serialization does not lead to further area savings. For instance, in Figure 4(b), the most area saving for a 130nm implementation is achieved with four wires. Any further serialization (i.e., reduction to 2 or 1 wires) causes a reduction in area savings. Similarly, for 90 and 65nm implementations, the most area saving is obtained for a two wire solution. This phenomenon is a result of the overhead of the transmission and receiver circuitry taking up more area as the level of serialization is increased. Serialization beyond a critical point is thus not beneficial for area savings.

As discussed in Section 3.3, often the TSV pad area is increased to ensure fault tolerance, since misalignment of wafers is not uncommon in 3D ICs. To explore area savings due to serialization under this phenomenon, TSV pad dimensions and pitch were increased to 10µm×10µm and 16µm respectively, while TSV dimensions were kept the same. Figure 4(c) and (d) show the area savings for 32-bit and 64-bit links when the proposed serialization scheme. The savings in area are more in this case, as compared to the previous non fault-tolerant case. For instance with a 4:1 serialization, the area savings are greater than 70% for implementations across 130-65nm libraries. For a TSV footprint of 1.6 mm^2 in each layer of a 3D CMP design (as explained in Section 1), the results show an area saving ranging from 45-95% depending upon the level of serialization. Such a reduction in area has significant benefits for reducing routing congestion and fabrication cost, and potentially improving thermal TSV distribution that can lead to lower peak chip temperatures.

5.2 Impact on CMP Power and Performance

Of course the area savings due to serialization come at a cost: a reduction in performance. A serialization degree of *n* theoretically reduces link throughput by a factor of *n*. There is additionally also an impact on the power dissipation of the communication fabric due to the additional serialization transmitter and receiver circuitry used. The next set of experiments attempt to quantify the power and performance impact of using the proposed serialization scheme for multi-core CMP designs implemented in 3D ICs.

5.2.1 Experimental Setup

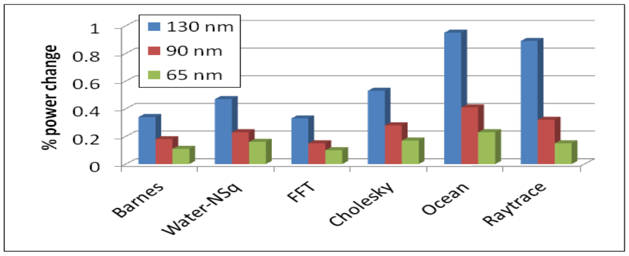
Six applications from the well known SPLASH-2 benchmark suite (*Barnes, Water-NSq, FFT, Cholesky, Ocean, Raytrace*) [39] were selected, then parallelized and implemented on multiple cores that were mapped to a 3D IC. The die size was assumed to be 2cm×2cm. The cores were connected with a packet switched 2D mesh NoC communication fabric with 64-bit wide vertical and planar links, clocked at a frequency of 1 GHz. Table 1 summarizes the implementation details of the CMP applications, such as number of cores (including processors and on-chip memories), the number of (64-bit) vertical links and the number of layers on which the cores were mapped in the 3D IC implementation.

Table 1. 3D CMP Implementation Details

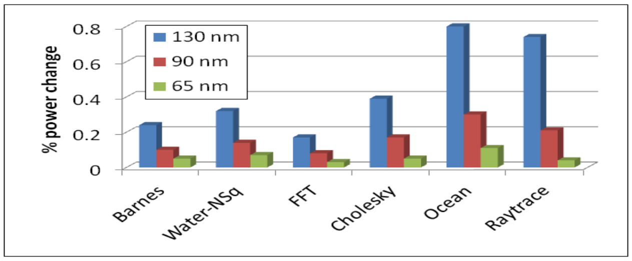
CMP applications	Description	Cores	Vert. links	Layers
Barnes	Galaxy evolution	32	16	2
Water-NSq	Forces/potentials of H_2O molecules	38	18	2
FFT	FFT kernel	44	20	2
Cholesky	Cholesky factorization kernel	76	54	4
Ocean	Ocean movements	88	66	4
Raytrace	3-D ray tracing	112	82	4

The 64-bit vertical TSV links were implemented as vertical buses with an interface at each router. There are two reasons for using vertical buses instead of vertical packet switched links. Firstly, vertical packet switched links would require the addition of two more ports and links (up and down) to each router, which would increase its complexity. In contrast, a vertical bus requires the addition of only a single new port to a router. Secondly, the distance between layers is relatively small (~20-100µm) in 3D ICs compared to inter-router distance in 2D ICs (~1000µm or more). As a consequence, multi-hop and router delay for vertical packet switched links would dominate the vertical propagation time and reduce performance [10]. Therefore a vertical bus with a dynamic TDMA arbitration to support programmable quality of service (QoS) is considered instead of vertical packet switched links.

The CMP applications were modeled in SystemC [40] using a fast and accurate transaction-based bus cycle accurate (T-BCA) modeling abstraction [41]. The cores were modeled at the behavioral level granularity, while the inter-core communication was modeled at a cycle accurate granularity. Each of the applications was simulated with testbench traffic (~few hours) to estimate performance of the implementations. A high level simulated annealing floorplanner for 2D ICs [32] was used to create a thermal-aware layout of the CMP application on the 3D-IC, and Manhattan distance based wire routing estimates were used to determine wire lengths. The wire length information, together with application traffic profiles obtained from simulation were plugged into an on-chip communication architecture power estimation framework [37] to determine link and communication-centric logic component (routers, NoC interfaces) power dissipation for the target technology library.



(a)



(b)

Figure 5. % power change for CMP applications for 130–65nm nodes when using (a) 2:1 serialization (64 to 32 bit links), and (b) 4:1 serialization (64 to 16 bit links)

5.2.2 Impact on Power Dissipation

Figure 5 shows the percentage change in power dissipation of the on-chip communication architecture fabric when the proposed serialization scheme is used, for the CMP applications implemented in the 130, 90, and 65 nm technology libraries. Figure 5(a) shows the case of 2:1 serialization (64-bit links reduced to 32-bit links) while figure 4(b) shows the case of 4:1 serialization (64-bit links reduced to 16-bit links). Higher degrees of serialization are not as important because they can cause a notable degradation in application performance (as discussed in the next subsection). It can be seen that serialization causes a slight increase in power dissipation. This is an interesting result because a lot of previous work on serialization for planar interconnects has indicated the potential for power savings with serialization [22][24][25]. The reason for the reduction in power dissipation in those schemes is because of two reasons: *(i)* the schemes allow aggressive reduction in crosstalk capacitance due to greater freedom with wire spacing and sizing after serialization reduces the number of wires, and *(ii)* the length of the wires considered are long enough for the saving in crosstalk to mitigate power dissipation in the serialization circuitry. However, for the proposed vertical TSV serialization case, while there is a reduction in crosstalk capacitance with a reduction in number of TSVs, the width and spacing (pitch) of the remaining TSVs is not altered. In addition, the length of the TSVs is much smaller, which limits the savings due to reduction in crosstalk capacitance. Thus, the power dissipation overhead of the serialization transmitter and receiver circuitry dominates, leading to an increase in power dissipation.

Note that the power dissipation for the 4:1 serialization case is less that for the 2:1 case, because of the greater sharing of resources in the transmitter and receiver circuitry (Section 4), as well as the reduction in switching power on the fewer wires for the 4:1 serialization case. The power dissipation overhead of serialization decreases with technology scaling, and is almost negligible for the 65 nm library. Further scaling below 65 nm will potentially lead to the wire power dissipation dominating the power dissipation of the serialization circuits. In such a scenario, serialization of interconnect TSVs will lead to a reduction in power dissipation.

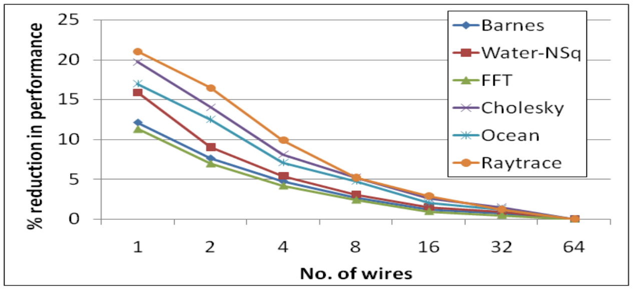


Figure 6. % performance change with varying degrees of serialization for 64-bit links in a 65nm technology implementation for CMP applications (1 GHz clk frequency)

5.2.3 Impact on Performance

Serialization of links can save area, but only at the cost of performance. Figure 6 shows the percentage reduction in overall application performance (y-axis; measured in number of cycles to finish application execution) for the CMP applications, with varying degrees of serialization (represented using number of wires; x-axis). It can be seen that as the degree of serialization increases, and the number of wires in the links are reduced from the original value of 64, the performance degrades. This performance degradation is low for low degrees of serialization, but can become high for higher degrees of serialization. For instance, the performance degradation is about 1.7% on average for a 4:1 serialization (16 wires) but reaches around 16.1% for a 64:1 serialization (1 wire). The exact value of performance degradation depends on the frequency of vertical transfers and the number of TSV interconnects, which varies across the CMP applications. It is clear however that lower degrees of serialization such as 2:1 and 4:1 are more practical because of their low performance and power overhead and an appreciable area saving.

One advantageous consequence of serialization is that it reduces crosstalk capacitance, which in turn reduces propagation delay. This can enable higher clock frequencies on the wires. Figure 7 shows the percentage performance overhead for the 4:1 serialization case, when TSV interconnects are clocked at higher clock frequencies than the baseline 1GHz, for the 65 nm technology node. The higher clock frequencies can improve performance, but only to a limited degree. Even without such frequency increase however, serialization has clear benefits for 3D IC technology. For instance, 4:1 serialization of 64-bit links can save more than 70% of the TSV area footprint on each layer, at the minimal cost of 0.06% power and 1.86% performance overhead on average for CMP applications. This is a strong motivation for considering serialization of TSV links in emerging 3D CMP architectures.

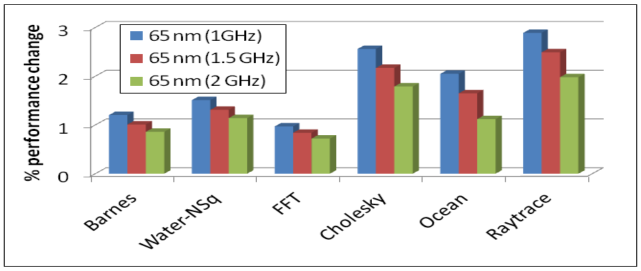


Figure 7. % performance change with 4:1 serialization (64 to 16 bit links) for CMP applications for 65 nm technology node, with scaled TSV clock frequencies

6. CONCLUSION

Vertical interconnect TSVs in 3D ICs take up significant chip area, and can cause routing congestion because of their typically spread-out distribution. In this paper, the use of serialized vertical interconnects in 3D ICs was explored to reduce the area footprint of interconnect TSVs. A shift register based serialization scheme was proposed and it was shown how varying the degree of serialization can result in significant savings in TSV area on each layer of a 3D IC. It was also demonstrated through experiments on several 3D CMP architectures that the proposed serialization scheme can significantly reduce interconnect TSV area footprint, at a nominal power and performance overhead. The extra space made available on each layer due to serialization can be used for better core placement and routing, as well as more efficient thermal TSV insertion for temperature management. Future work will quantify the temperature reduction due to serialization, as well as explore other serialization schemes for 3D ICs.

7. REFERENCES

[1] S. Pasricha and N. Dutt, “On-Chip Communication Architectures”, Morgan Kauffman, ISBN 978-0-12-373892-9, Apr 2008

[2] S. M. Rossnagel, T. S. Kuan, “Alteration of Cu Conductivity in the Size Effect Regime,” JVST, Vol. 22, Iss. 1, pp. 240-247, Jan. 2004.

[3] International Technology Roadmap for Semiconductors (ITRS), System Drivers, 2007.

[4] A. W. Topol et al., “Three-dimensional integrated circuits,” IBM J. Res. & Dev. Vol. 50 No. 4/5 Jul/Sep 2006.

[5] K. Bernstein, et al., "Interconnects in the Third Dimension: Design Challenges for 3D ICs," Proc. DAC 2007, pp.562-567.

[6] R. S. Patti, “Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs”, Proc IEEE, Vol 94, No. 6, Jun 2006.

[7] V. F. Pavlidis, E. G. Friedman, “Three-dimensional Integrated Circuit Design”, Morgan Kaufmann, Sep 2008.

[8] B. Feero, P.P. Pande, “Performance Evaluation for Three-Dimensional Networks-On-Chip”, Proc. ISVLSI 2007.

[9] S. Pasricha, N. Dutt, "Trends in Emerging On-Chip Interconnect Technologies", IPSJ Transactions on System LSI Design Methodology, Vol. 1, Sep 2008

[10] F. Li et al., “Design and Management of 3D Chip Multiprocessors Using Network-in-Memory”, Proc. ISCA 2006, pp. 130-141.

[11] B. Goplen, S. Sapatnekar, “Thermal via placement in 3D ICs”, Proc. ISPD 2005.

[12] Z. Li, et al., “Efficient thermal-oriented 3D floorplanning and thermal via planning for two-stacked-die integration”, ACM TODAES 11:2, Apr 2006, pp. 325-345.

[13] C. Addo-Quaye, “Thermal-aware mapping and placement for 3-D NoC designs,” Proc. IEEE Int. Syst.-on-Chip Conf., 2005, pp. 25–28.

[14] E. Wong, S. K. Lim, “3D Floorplanning with Thermal Vias”, Proc. DATE 2006, pp. 1-6.

[15] J. Cong, Jie Wei and Yan Zhang, “A thermal-driven floorplanning algorithm for 3D ICs”, Proc. ICCAD 2004, pp. 306-313.

[16] P. Zhou et al., “3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits”, Proc ICCAD 2007.

[17] I. Loi et al., “Supporting vertical links for 3D networks on chip: toward an automated design and analysis flow”, Proc. NanoNet 2007.

[18] K. Puttaswamy and G.H.Loh, “Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors”, Proc. HPCA 2007, pp. 193-204.

[19] Y. Liu, et al., "Fine Grain 3D Integration for Microarchitecture Design Through Cube Packing Exploration", Proc. ICCD, 2007.

[20] D. Park et al. “MIRA: A Multi-layered On-Chip Interconnect Router Architecture”, Proc. ISCA 2008, pp. 251-261.

[21] K. Puttaswamy, G. H. Loh, "Implementing caches in a 3D technology for high performance processors" Proc. ICCD 2005, pp. 525-532.

[22] A. Morgenshtein et al., “Comparative Analysis of Serial vs Parallel Links In NoC”, Proc. SSOC, 2004.

[23] R. Dobkin, et al., "Parallel vs. Serial On-Chip Communication”, Proc. SLIP 2008.

[24] M. Ghoneima et al., “Serial-Link Bus: A Low-Power On-Chip Bus Architecture”, Proc. ICCAD 2005.

[25] N. Hatta et al., “Bus Serialization for Reducing Power Consumption”, IPSJ TACS 47:3, Mar 2006.

[26] S. Kimura et al., “An On-Chip High Speed Serial Communication Method Based on Independent Ring Oscillators”, Proc. ISSCC 2003.

[27] I-Chyn Wey et al., “A 2Gb/s High-Speed Scalable Shift-Register Based On-Chip Serial Communication Design for SoC Applications”, Proc. ISCAS 2005.

[28] M. Saneei, A. Afzali-Kusha1, M. Pedram, “Two High Performance and Low Power Serial Communication Interfaces for On-chip Interconnects”, Proc. CJECE 2008.

[29] R. Dobkin, et al., "Fast Asynchronous Shift Register for Bit-Serial Communication," Proc. ASYNC, 117-126, 2006.

[30] S. Ogg et al., “Serialized Asynchronous Links for NoC”, DATE 2008.

[31] K.-N. Chen, A. Fan, and R. Reif, “Microstructure examination of copper wafer bonding,” in http://www-mtl.mit.edu/ reif/papers/2001-knchen-JEM-manuscript.pdf.

[32] S. N. Adya, I. L. Markov, "Fixed-outline Floorplanning: Enabling Hierarchical Design", *IEEE TVLSI, Dec. 2003*

[33] T. Zhang, Y. Zhan and S. Sapatnekar, “Temperature-aware routing in 3D ICs”, Proc. ASPDAC 2006.

[34] S. Lee, et al, “Analysis of Thermal Vias in High Density Interconnect Technology,” Proc. IEEE Semi-Therm Symposium, pp. 55-61, Feb. 1992.

[35] I. Loi, et al., “A Low-overhead Fault Tolerance Scheme for TSV-based 3D Network on Chip Links”, Proc. ICCAD 2008.

[36] K. N. Chen, A. Fan, and R. Reif, “Interfacial morphologies and possible mechanisms of copper wafer bonding,” in http://www-mtl.mit.edu/users/reif/papers/2002-knchen-JMS-manuscript.pdf.

[37] S. Pasricha, Y. Park, F. Kurdahi, N. Dutt, "System-Level Power-Performance Trade-Offs in Bus Matrix Communication Architecture Synthesis", IEEE/ACM CODES+ISSS 2006

[38] Synopsys Design Compiler, PrimeTime PX, www.synopsys.com.

[39] S.C. Woo et al.“The SPLASH-2 programs: Characterization and methodological considerations”, Proc. ISCAS, 1995.

[40] SystemC initiative. www.systemc.org.

[41] S. Pasricha, N. Dutt, M. Ben-Romdhane, "Extending the Transaction Level Modeling Approach for Fast Communication Architecture Exploration", IEEE/ACM DAC 2004.