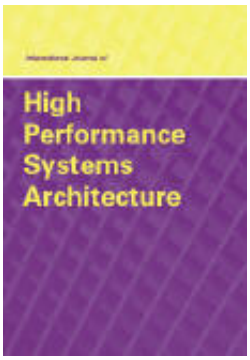


International Journal of High Performance Systems Architecture

ISSN online: 1751-6536 - ISSN print: 1751-6528
https://www.inderscience.com/ijhpsa

3D layout of the Spidergon-Donut on-chip interconnection network

Fadi N. Sibai, Abu Asaduzzaman, Ali Elmoursy

DOI: 10.1504/IJHPSA.2023.10054349

Article History:

Received:	04 June 2022
Last revised:	30 July 2022
Accepted:	05 September 2022
Published online:	06 April 2023

3D layout of the Spidergon-Donut on-chip interconnection network

Fadi N. Sibai*

Department of Electrical & Computer Engineering,
Gulf University for Science and Technology,
P.O. Box 7207, Hawally 32093, Kuwait
Email: sibai.f@gust.edu.kw
*Corresponding author

Abu Asaduzzaman

Department of Electrical and Computer Engineering,
Wichita State University,
Wichita, KS 67260, USA
Email: abu.asaduzzaman@wichita.edu

Ali Elmoursy

Department of Electrical and Computer Engineering,
University of Sharjah,
P.O. Box 27272, Sharjah, UAE
Email: aelmoursy@sharjah.ac.ae

Abstract: 3D integration promises to resolve many of the heat and die size limitations of 2D integrated circuits. A critical step in the design of 3D many-cores and MPSOCs is the layout of their 3D network-on-chip (NoC). In this paper, we explore and present multiple 3D layouts of the Spidergon-Donut (SD) NoC and estimate their longest wire lengths and cost requirements. For a total of 64 cores, the 4×2×8 and 2×4×8 placements result in the best longest wire delays, with the former higher 3D integration costs, while the second requiring larger chip area and through-silicon-vias (TSV) array costs. Such study helps in guiding 3D integration direction and weighing 3D NoC layout and placement alternatives.

Keywords: 3D integration; 3D on-chip networks; Spidergon-Donut network; many-core processors; wire delay; chip area; heat dissipation.

Reference to this paper should be made as follows: Sibai, F.N., Asaduzzaman, A. and Elmoursy, A. (2023) '3D layout of the Spidergon-Donut on-chip interconnection network', *Int. J. High Performance Systems Architecture*, Vol. 11, No. 3, pp.137–147.

Biographical notes: Fadi N. Sibai presently serves as the Associate Dean, College of Engineering and Architecture, Gulf University of Science and Technology, Kuwait. He previously served as Dean of the College of Computer Engineering and Science at PMU, Saudi Arabia, and as Program Director in the College of IT at the UAE University, UAE. He also worked for Aramco and Intel Corporation. He earned a PhD in Electrical Engineering.

Abu Asaduzzaman is an Associate Professor of Electrical and Computer Engineering and Director of the undergraduate Computer Engineering Program at Wichita State University. He received the PhD and MS degrees, both in Computer Engineering, from Florida Atlantic University. His research interests include computer architecture, high performance computing, and embedded systems.

Ali Elmoursy is Chair, Electrical and Computer Engineering, University of Sharjah, UAE. His research interests include high-performance computer architecture, multi-core multi-threaded mirco-architecture, power-aware micro-architecture, simulation and modelling of architecture performance and power, workload profiling and characterisation, cell programming, high performance computing, parallel computing, and cloud computing. He worked for Intel Corporation and received the PhD in high-performance computer architecture from the University of Rochester.

1 Introduction

Sustained scaling of metal–oxide–semiconductor field-effect transistors, on-chip multiprocessors and MPSOCs is getting more difficult and expensive. For instance, a leading chip maker recently skipped the manufacturing of desktop CPUs at 10 nm process. Also, available space for memory is shrinking on cost-conscious 2D dies. One natural solution to this problem is slowing the scaling down of devices on a 2D die and stacking devices vertically, a process known as 3D integration. 3D integration and stacking was used in the manufacturing of high bandwidth memory in the AMD R9 Fury X video card, and in the 2018 Intel Foveros 3D chip stacking technology which breaks a CPU chip down and lays the pieces into stackable chiplets inside one chip package. With Foveros technology, which was used to manufacture Intel's Lakefield processor (now discontinued), the transistor density will improve by 3D stacking rather than 2D area reduction. Owing to the feasibility of placing them very close vertically to each other, 3D-stacked transistors and devices support higher communication speeds between stacked components than between 2D devices placed on one 2D die plane. 3D stacking allows placing heterogeneous units such as CPUs, GPUs, AI processors, media cores, chipset, and FPGAs together on one chip, with these units possibly manufactured with different process technologies to optimise performance (Sibai, 2007, 2008) and power (Asaduzzaman et al., 2013). The most heat dissipating components will be placed on the top for facilitating heat dissipation. Elsewhere, Global Foundries has announced in August 2019 testing a 3D ARM chip with 12 nm FinFET process and using an ARM mesh interconnect technology in the third (vertical) dimension.

Other 3D heterogeneous solutions are possible, stacking processors, memory, and I/O on top of each other. These solutions address the memory bottleneck problem, i.e., memory speed's incapability to catch up with the processor speed. For instance, Shulaker et al. (2017) presented a nano-system combining computing and data storage with over one million resistive random-access memory (RRAM) cells and over two million carbon-nanotube field-effect transistors (CNFET) manufactured on vertically stacked layers in a single chip. 3D integration naturally matches the CNFETs and RRAMs, owing to their low-temperature fabrication capability at a couple of hundreds of degrees Celsius, and avoiding the melting of wires and transistors, on the bottom layer of the 3D stack, which start damaging at 400 degrees Celsius. The 3D wire vias permit vertical connections which are 1000x denser than traditional packaging. If a second layer of silicon circuitry is built on top, high temperature will damage the bottom layer of circuits. Because CNFETs and RRAMs can be fabricated at temperatures below 200 degrees Celsius, they can be built up in layers without damaging the underneath circuits. This 3D solution improves circuit density, improves communication speed (latency and bandwidth), and reduces energy by the use of CNFETs and RRAMs.

Recently, commercial processors with as high as 32 cores were introduced. Higher core counts are projected to continue in this decade. To interconnect such large number of IP cores in chip multiprocessors (CMPs) (Hammond et al., 1997) and multiprocessor systems-on-chip (MPSoCs) (Jerraya et al., 2004), researchers have looked at on-chip wireless communications (Deb et al., 2012; Abadal et al., 2014; Ganguly et al., 2011; Zhao and Wu, 2012), 3D interconnects (Shulaker et al., 2017), RF transmission lines (RF-I) (Chang et al., 2008; Carpenter et al., 2012; Beckmann and Wood, 2003; Hsu et al., 2012), surface wave interconnects (SWI) (Karkar et al., 2012; Karkar et al., 2014, 2015), and on-chip optical nano-photonic networks (Kirman et al., 2006; Miller, 2009), all with design and/or manufacturing challenges (Abadal et al., 2014). Karkar et al. (2016) surveyed various technologies for on-chip network technologies and concluded that wireless network-on-chip (WiNoC) and SWI are most promising for multicast and broadcast communications in many core systems. SWI beats WiNoC on power dissipation and frequency range at the expense of technological maturity. Optical interconnects suffer from higher costs, complexity, and power consumption. Metal wires remain the lowest cost type of interconnects and can overcome signal decay, latency, and bandwidth deficiencies by 3D integration. WiNOCs suffer from vulnerability to signal jamming and antenna design, while 3D integration extends known technology in a third dimension. With advances in 3D integration, when design barriers and challenges such as process control requirements, wafer thinning, low through-silicon-via (TSV) capacitance and inductance (Cueping et al., 2010) and TSV-associated reliability issues, 3D EDA tool availability, thermal issues arising from higher power densities(https://www.sciencedirect.com/topics/engineering/three-dimensional-integrated-circuits) are all overcome, 3D wire interconnects may reach preferred status in 3D integrated circuits.

TSVs are vertical copper or tungsten conducting nails passing through a thinned die, typically with a 5 μm diameter and 50 μm height, and are typically used to deliver signal/power/ground and enabling communication between different layers. TSVs allow the mixing of various process technologies, for instance, the top die can be processed with traditional 2D process technology without TSVs. The stacked dies are thinned to expose the TSVs. The TSVs go through the bottom stacked dies and connect to micro-bumps of the top die in the stack to provide electrical connections between all the stacked dies (Deutsch, 2015) in the various layers.

With shrinking circuit dimensions, signal propagation delays increase in global wires but decrease in local wires. This results from die areas tending to remain constant, so the global wires have relatively the same lengths across processes, while local wire lengths shrink with shrinking logic areas. Fat wires which reduce the cross-section area of wires improve the RC product and allow the sizing of various wires across the chip. However fat wires are not

desirable as they occupy large chip areas and increase the capacitance of fat wires in proximity to each other. Repeaters can be used to improve the signal degradation and fanout but cause the chip delay to grow linearly (instead of exponentially without repeaters) with newer process technologies and circuit scaling. Moreover, some researchers (Magen et al., 2004) have predicted that wire interconnects will consume 80% of the total IC power mainly from power dissipating in wires than gates. In addition, cache coherence requires invalidation multicasts by directory-based cache protocols and broadcasting tokens by broadcast-based cache coherence protocols and therefore require interconnects with fast and scalable 1-to-many and 1-to-all capabilities. Despite its complexity, 3D integration shows great potential for reducing the coherence network latencies and improve its reach, increase memory bandwidth, enhance 3D processor security (Gu et al., 2016), and reduce wire latencies and power dissipation due to lower inter-chip communication overhead and greater circuit densities and the possibility to mix and match by integrating layers with various process characteristics. Furthermore, security improves as a result of distributing the functions over multiple layers thereby hindering the reverse engineering of the product.

To address the above problems, the main contributions of this paper are the presentation and evaluation of the 3D structure and layouts of the Spidergon-Donut (Sibai, 2011, 2012; Sibai et al., 2012; Albughdar and Mahmood, 2015) NoC, which was shown to exhibit a number of desirable NoC properties. The 3D version of the SD NoC allows for the packing of a large number of cores while addressing the high number of core integration, density, and communication and data transfer latency requirements of modern computers. To the best of our knowledge, this is the first work related to the 3D structure and layouts of the SD NoC. The 3D layouts explored in this work also apply to similar NoCs with 4–5 node degrees. The paper is organised as follows. Thermal Challenges of 3D ICs are reviewed in Section 2. The 2D SD NoC is reviewed in Section 3. Section 4 presents the various 3D chip layouts of the SD NoC. In Section 5, the longest wire and cost are estimated and compared for the various 3D SD NoC layouts. The paper concludes in Section 6.

2 Thermal challenges of 3D NOCs

As buses cannot handle the traffic from tens of cores, network-on-chip (NoC) was proposed to support the various inter-core and inter-unit traffic of MPSoCs. TSV 3D integration technology (Black et al., 2006) allows the integration of different logic, memory, sensor, I/O units over multiple 3D layers and make 3D NoCs possible. According to Chen et al. (2015), 3D NoCs have the following advantages over 2D NoCs by virtue of vertical stacking:

i higher unit/core densities

ii shorter wires, hop count and NoC power consumption

iii higher network bandwidth of degree-6 routers in 3D NoC vs. degree-4 routers in 2D NoC).

For instance, a 2D 64-core (8x8) mesh NoC requires a maximum of 14 (i.e., 7+7) hops connecting 2 cores diagonally situated from each other. If this 2D NoC is broken down into four 4x4 cores stacked vertically on top of each other, the maximum number of hops becomes 9 (i.e., 3+3+3; with the last 3 in the vertical dimension) for a 36% reduction in end-to-end latency. Counterbalancing the improvements in latency, the higher unit stacking density results in larger power density, more hot spots, and higher temperature. These problems are exacerbated by the (vertically) longer heat conduction path and the low thermal conductivity of the oxide layers between the semiconductor layers. As the heat gets trapped in these layers, the 3D chip temperature rises more than normal 2D chips, leakage power grows, and the chip reliability and durability eventually drop. Hung et al. (2006) demonstrated that the temperature of a 3D 2-layer IC with 4 cores per layer significantly exceeds the temperature of a 2D IC with 8 cores. Known 2D IC cooling techniques such as liquid cooling and heat spreader are not cost effective for 3D ICs. Thus, various thermal solutions have been proposed for 3D NoC-based ICs among which are dynamic threat management and intelligent packet routing. The former controls the temperature by throttling and clock gating, leading to a severe drop in performance (Chao et al., 2010, 2013), or by future temperature prediction. The latter routes packets around throttled nodes to inactive nodes with low heat dissipation. In 3D, the migration could be vertical to a low temperature unit in another layer. Both techniques change the network topology as a consequence of deactivating units resulting in packets getting blocked and reduced performance. Chao et al. (2013) proposed a thermal-aware vertical throttling which throttle overheated nodes or units on top of each other (i.e., in the same column) in the upper layers while keeping the node in the bottom-most layer active to prevent packet blocking and performance reduction. Another technique migrates tasks from hot units to low temperature ones.

Other domains have also been explored to control heat dissipation in 3D ICs. For instance, Cao et al. (2019) surveyed techniques for thermal-aware processors including 3D memories, floorplanning to balance chip temperature and wire length, memory management to optimise peak temperature, energy and throughput, and task scheduling to also optimise peak temperature, energy and throughput. Liquid cooling techniques have also been proposed for 3D ICs. For instance, Sridhar et al. (2010) proposed and manufactured a thermal model for thermal simulation of 3D integrated circuits with microchannel liquid cooling. The 3D IC has microchannels etched on the back of the dies

with microchannel cavities, looking very similar to irrigation channels for irrigating dry lands. This cooling method has a fluid inlet on the side of the 3D IC and a fluid outlet on the opposite side for facilitating the flow of the coolant fluid for cooling the interlayer cavities.

In this work, TSVs are used, complemented by possibly other thermal management techniques.

3 The 2D Spidergon-Donut network

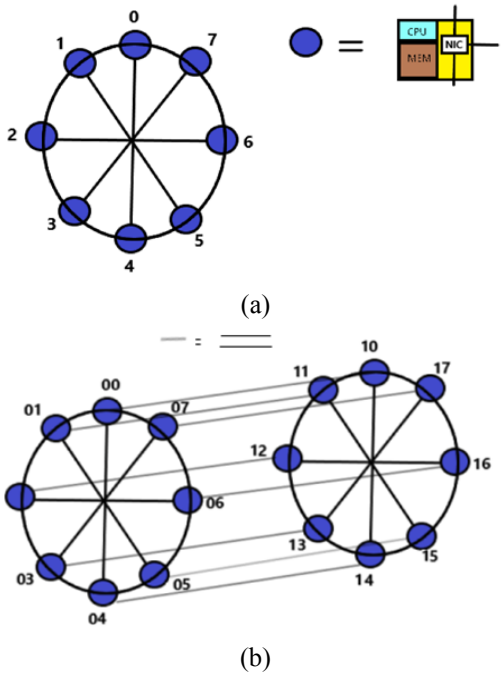
The Spidergon (Coppola et al., 2004, 2008) (SG) NoC has good advantages for a small number of cores. In many cores, the SG will not be competitive with other known NoC topologies. The Spidergon- Donut (SD) NoC (Sibai, 2012) extends the SG into the third dimension and was shown to have good properties for a large number of cores. For instance, for 1024 cores, the SD features a lower diameter than the popular Mesh and Torus networks for slightly higher node degree, and 25% additional links. With a large number of cores, concentrated or bristled versions of the SD further reduce the diameter and average inter-core distance.

Figure 1(a) shows the SG(N) NoC with $N = 8$ nodes. The bubble represents a node which could be one core or a concentrated router connecting 4 cores all with possible external cache memory. Figure 1(b) shows the SD($N_1 = 2$, $N_0 = 8$) NoC with two SG(8) instances. The nodes in the first SG(8) are labelled 00, 01, …, 07, while the nodes in the second SG(8) instance are labelled 10, 11, …, 17. A node labelled XY connects to another node labelled ZY if and only if $|X–Z| = 1$ (i.e., neighbouring instances), or (X, Z) = (0, N–1) or (X, Z) = (N–1, 0) (i.e., wraparound cases).

Some properties of the SD NoC follow. The number of nodes in a Spidergon-Donut SD(N_1, N_0) network is $N_1 \times N_0$. The number of links (link cost) in a SD(N_1, N_0) is $2.5 \times N_0 \times N_1$, as there are N_0 ring links per SG instance, $N_0/2$ cross links per SG instance, N_1 instances of SG(N_0)'s, and $N_0 \times N_1$ inter-SG instance links in the second dimension. The diameter of the SD(N_1, N_0) is $N_1/2 + N_0/4$ as the longest distance within an SG dimension is $N_0/4$, while the longest second dimension distance is $N_1/2$. The bisection width of the SD(N_1, N_0) is $2 \times N_0$. Other advantages of the SD NoC can be found in (Sibai, 2012).

Sibai et al. (2012) explored a performance queueing model for the SD. AAlbughdar and Mahmood (2015) proposed an adaptive dead-lock free routing for the SD. (Sibai, 2012) proposed a hybrid version of the SD with 8 4x concentrated SD(4,8)s connected to an 8×8 crossbar or multistage interconnection network (MIN) to interconnect the cores of a 1000 core processor. The MIN will reduce the diameter compared to a SD(32, 8). However, 3D integration shows further potential to reduce the inter-core delays of an SD or a similar NoC. We therefore explore a 3D version of the SD in the next section.

Figure 1 Spidergon-Donut NoC: (a) 1D SG(8) and (b) 2D SD(2, 8) (see online version for colours)



4 3D Spidergon-Donut NOC

4.1 Spidergon SG(8) network

The 2D SD NoC was described in Section 3. In this section, we describe the 3D NoC which improves the density, latencies, and the overall number of integrated cores over its 2D counterpart. The 8-node Spidergon (SG) is shown in Figure 2(a) where a node floorplan (Sibai, 2008) is pictured as one core or a concentrated router (Bahn et al., 2007) connecting 4 cores all with possible external cache memory. The 2D layout of the SG is shown in Figure 2(b) where the wraparound links are shown as an X. The average inter-core connections can be shortened by folding and shuffling. The 3D layout is shown in Figure 2(c) with a set of 4 nodes on one level and another set below them with one inter-node dimension implemented with vertical links (0⇔1, 4⇔5, 2⇔3, 6⇔7). Note that the 3D layout benefits from the 3D stacking of the 2 sets of 4 nodes on top of each other resulting in shorter link lengths. The main issue with the SG is its diameter scalability with increasing number of nodes.

4.2 Cubic(8) network

To get rid of the crossed wraparound links, these links can be deleted to produce the cubic network as shown in Figure 3(a). The 2D layout of Figure 3(b) can be reshuffled as in (Sibai, 2012) to reduce average inter-core links. The 8-node cubic network has a cubic 3D layout as shown in Figure 3(c).

Figure 2 Spidergon-Donut NoC: (a) topology; (b) 2D layout and (c) 3D layout (see online version for colours)

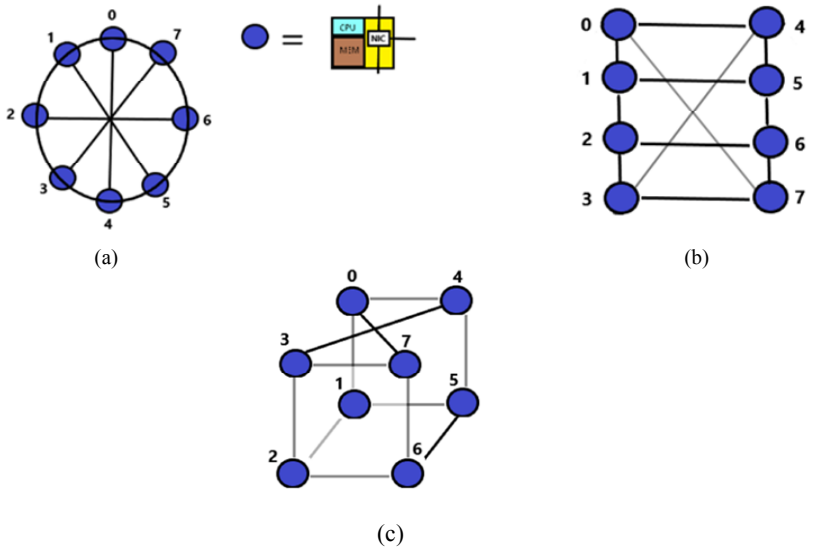
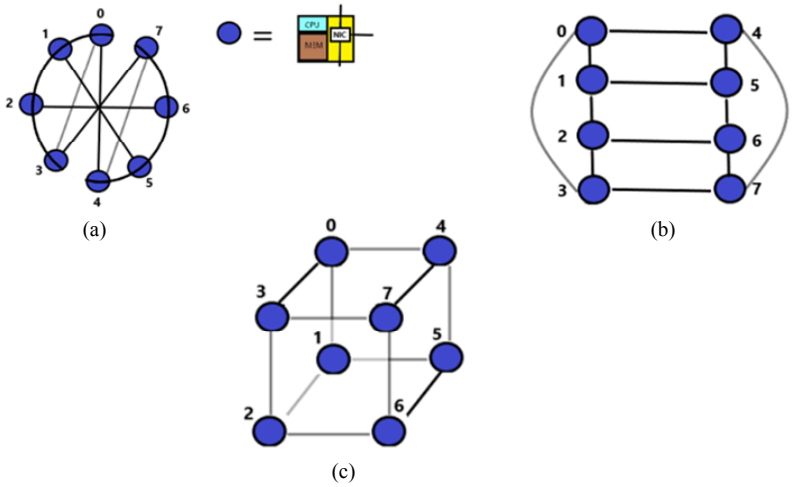


Figure 3 SG(8) network layout: (a) topology; (b) 2D layout and (c) 3D layout (see online version for colours)



4.3 Asymmetric SG(8) network

Another way to get rid of the crossed wraparound links, is to delete the 0⇔7 and 3⇔4 links to produce the asymmetric SG(8) structure shown in Figure 4(a). The asymmetric SG removes the wraparound links thereby simplifying the 2D layout, at the expense of losing the NoC symmetry. Note that the node degree varies between 2–3 leading to the network's asymmetry. The 2D and layouts of this NoC are depicted in Figures 4(b) and (c), respectively. The 3D layout of the asymmetric SG with 16 nodes in shown in Figure 4(d).

4.4 Spidergon cylinder network SC(2, 8)

To support higher node numbers, the Spidergon Cylinder SC(2, 8) network connects two SG(8)s in a cylinder fashion as shown in Figure 5(a). The degree of 4 and higher number of links starts to complicate the link layout. The 2D layout is shown in Figure 5(b) while the 3 D layout is shown in Figure 5(b). Note that the 2D layout increases the planar layout area, while the 3D layout cuts the required planar layout area at the expense of vertical space.

4.5 Spidergon-Donut SD(2, 8) network

To improve the bisection width, wraparound links can be added to the nodes of the SC(2, 8) to produce the SD(2,8) as pictured in Figure 6(a). The 2D and 3D layouts are shown in Figure 6(b) and (c), respectively.

4.6 3D SD(8, 8) network

With increasing number of nodes, the 3D layout of Figure 6(c) must remain scalable. For instance, with 64 cores, the 3D SD(8, 8) is displayed in Figure 7(b) (right) and its 3D layout front view is displayed in Figure 7(a) (left). The network skeleton of Figure 7(a) depicts how the four SD(2, 8) are connected and laid out together to form the SD(8, 8), where each horizontal bar represents a SG(8). In this layout, each SG(8) node has 3 link connections on the same plane, and 2 other vertical links (vertical blue and

green, or vertical red and blue) for connecting the 8 SG(8)s in a donut or torus fashion. With a 2-level (layer) layout, 2 sets of 2 vertically stacked SG(8)s horizontally face two similar sets. The 2-layer horizontal green and red links require 2 more levels of 3D integration, one at the top, and the other on the bottom. Four vertical levels are required in total but 4 long sets of (green and red) horizontal links are necessary to build the SD.

Figure 4 Asymmetric SG(8) and SG(16) network layouts: (a) topology; (b) 2D layout;(c) 3D layout and (d) 3D layout (see online version for colours)

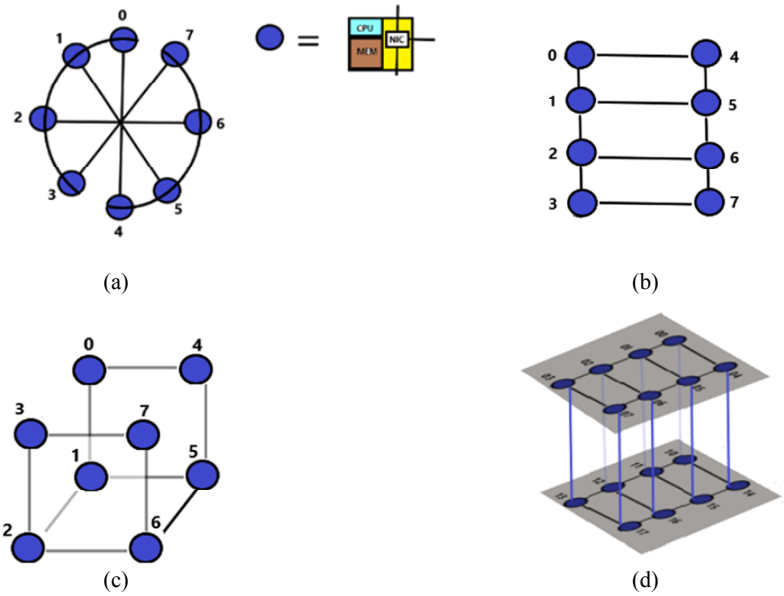


Figure 5 Spidergon Cylinder(2, 8) network layout: (a) topology; (b) 2D layout and (c) 3D layout (see online version for colours)

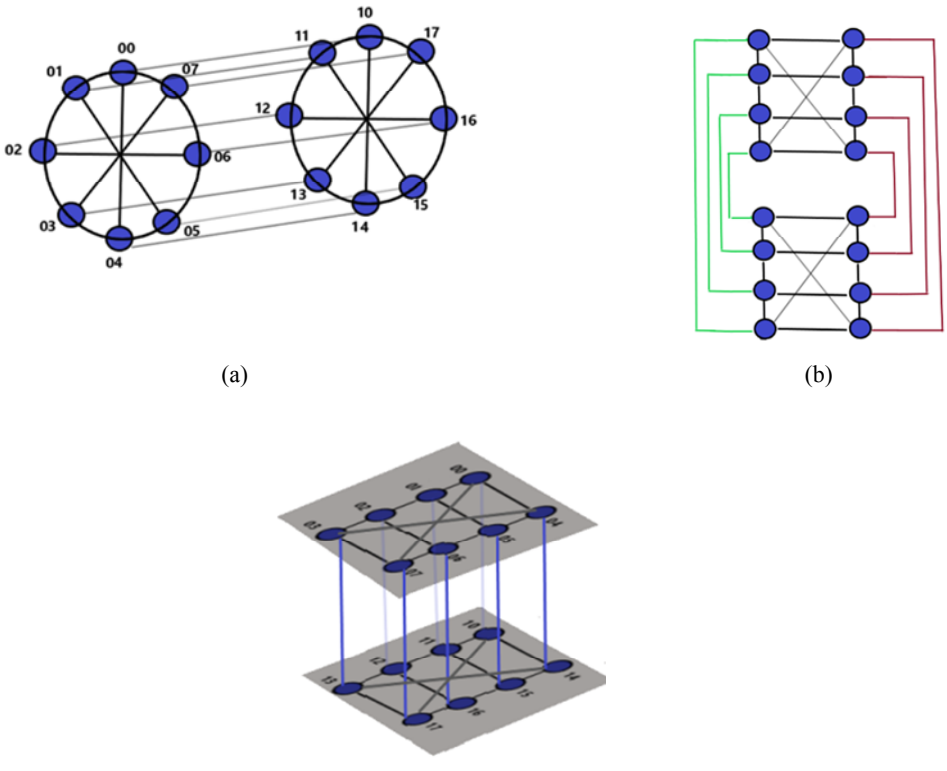
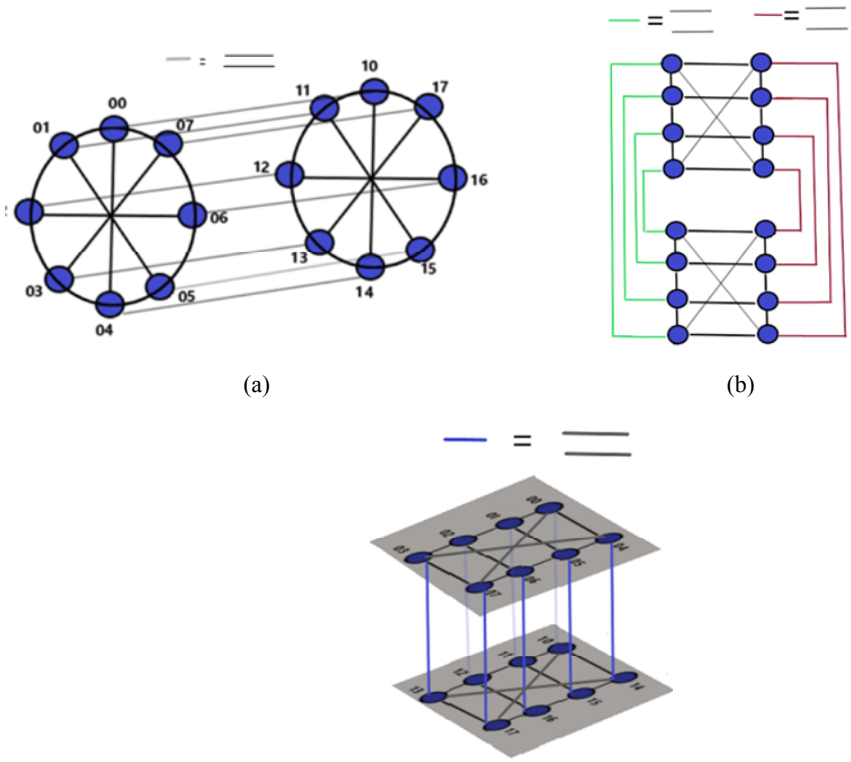


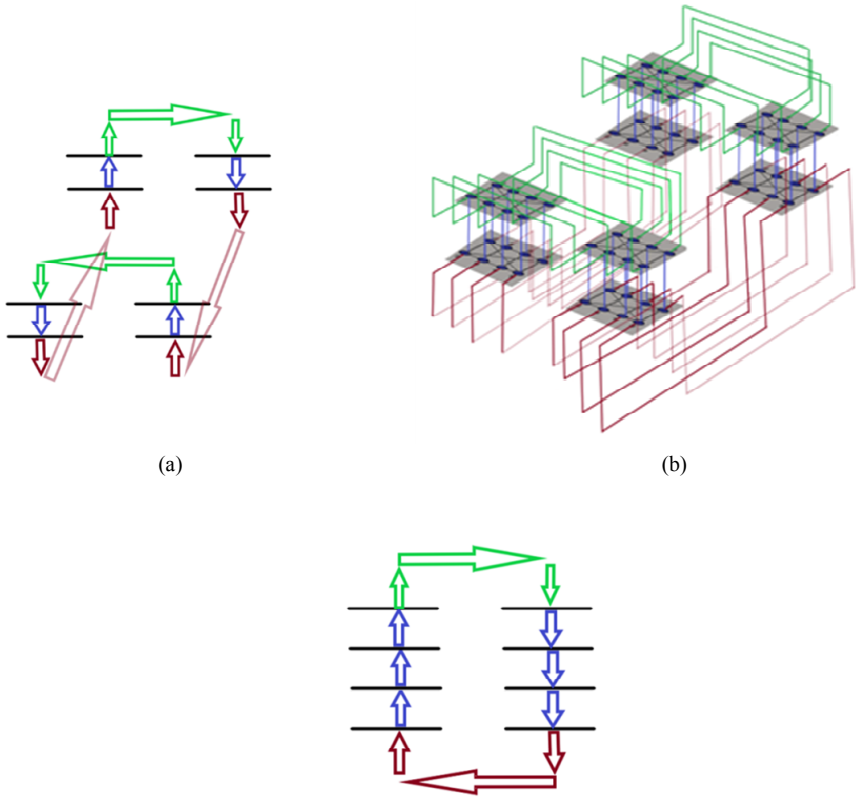
Figure 6 SD(2, 8) network layout: (a) topology; (b) 2D layout and (c) 3D layout (see online version for colours)



(a) (b)

(c)

Figure 7 3D split SD(8, 8) network layout: (a) SD(8, 8) 3D layout – front view; (b) SD(8, 8) 3D layout – side view and (c) another SD(8, 8) higher-integrated 3D layout (see online version for colours)



(a) (b)

(c)

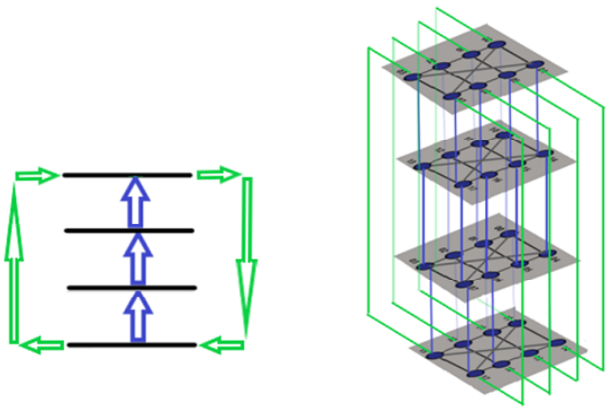
A simpler 3D layout, show in Figure 7(c), vertically stacks 2 sets of 4 SG(8)s each as shown in Figure 7(c). This layout eliminates the long horizontal links connecting the different SG(8) stacks of Figure 7(b). However, this layout increases the process integration complexity and cost given the higher number (precisely, 6) of integration levels rendering the 3D layout of Figures 7(a) and (b) more favourable.

The number of integration levels of Figure 7(a) and (b) is 4, 2 of which are for wraparound links, reducing the process integration cost compared to the layout of Figure 7(c).

4.7 3D SD(4, 8) network

A more compact 3D layout than the one in Figure 7(c) but applicable to networks with fewer nodes, stacks four SG(8)s vertically as shown in Figure 8. This SD(4, 8) layout eliminates the top and bottom wraparound link integration levels of Figure 7, by using the top and bottom core planes to initiate the wraparound connections. This reduces the 4 long sets of (green and red) horizontal links and latencies and reducing the integration levels from 6 to 4.

Figure 8 3D all-vertically stacked SD(4, 8) network layout: (a) SD(4, 8) 3D layout – front view and (b) SD(4, 8) 3D layout – side view (see online version for colours)



(a) (b)

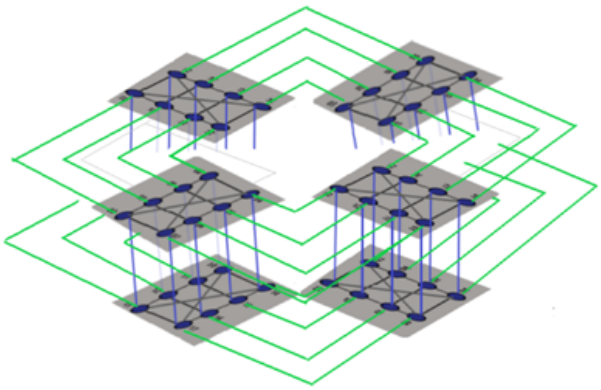
4.8 Ring of SG(8)s

Vertical SD network stacks can be connected in a ring fashion at each core-plane integration level in a ring fashion as shown in Figure 9. This layout brings the SD architecture to a hybrid ring of SDs topology. Another problem with this layout further to breaking the SD topology, is that the ring connections have unequal lengths and consume a large horizontal (2D planar) area.

4.9 3D SD with broadcast/Multicast channel

In 3D integrated circuits, as multiple semiconductor power-hungry layers are stacked vertically, the power density is higher than when components are placed together on a 2D plane, leading to potential thermal issues. Because the thermal conductivity of the oxide layer between semiconductor layers is low, the heat transfer towards the ambient temperature is diminished in comparison to 2D chips. As a result, new cooling solutions for 3D chips may be necessary.

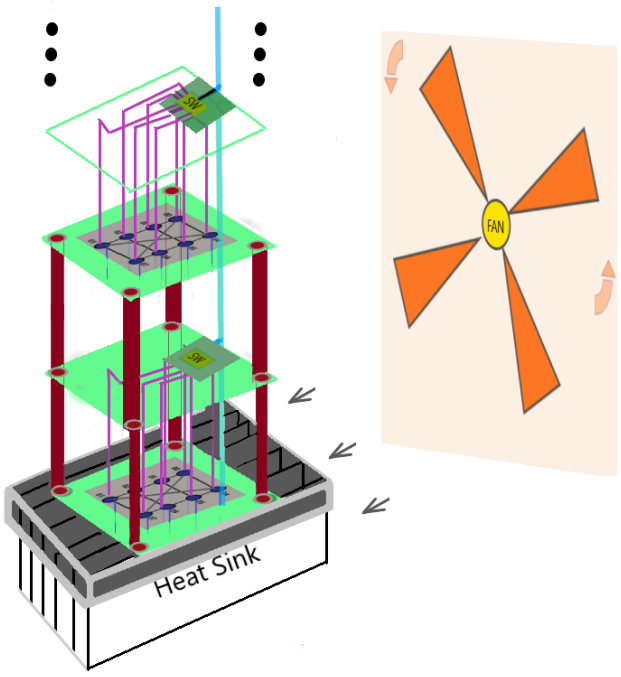
Figure 9 3D layout of rings of SDs (see online version for colours)



One intuitive idea is to construct the 3D SD integrated circuit as a building with columns, that is, layers on top of layers with spacing between them and with columns (in red), as shown in Figure 10 (not drawn to scale), to hold the structure together, as for example in commercial buildings. More columns are typically required but omitted in Figure 10 to improve readability. A fan can be added to dissipate the heat from the heat sink usually placed at the bottom of the 3D IC structure, with other heat sinks also possible. The inter-layer space may contain heat conducting material and could host the switches and wires for the cache coherence broadcast and multicast network (in pink and light blue in Figure 10) as shown below. In addition, these columns' role is not to solely hold the 3D multi-layer structure but have a dual purpose; these columns can be constructed of heat conducting material to route the heat to heat sinks at the edges of the 3D structure. For instance, Iqbal et al. (2019) proposed a heat junction to extract heat from a selected region in 3D and conducting the heat elsewhere. With the junction, they propose thermal pillars to facilitate heat dissipation through the substrate. These pillars resemble TSVs/Vias but only carry heat. Such pillars and junction were shown to reduce temperature by 50%.

Therefore, in Figure 10, the red columns could serve as structure columns to hold the structure together and to facilitate heat conduction when manufactured with heat conducting material. Multiple such columns can be designed for these purposes. The fan can also drive the heat reaching the heat sink(s) away from the 3D chips. Other possible cooling techniques include power delivery networks (Wei et al., 2012) which have appeal in the upper thin silicon layers with reduced heat conduction.

Figure 10 3D SD layout with inter-layer broadcast/multicast layer (see online version for colours)



5 Longest wire length and cost analysis

It was previously reported that reductions of 31% to 56% in the longest wire length are possible with 3D integration (Das et al., 2003), compared to 2D integration. The cost of a 3D integrated circuit was formulated in (Hsueh et al., 2011) who minimised the 3D integrated circuit cost using a cost prediction and partitioning method. The total TSV area is the product of the number of TSVs and the TSV area, the former depending on the chip connectivity, while the latter depending on the used process. Hsueh et al. (2011) found that as the number of TSV IOs (on the bottom layer) increases, the minimum 3D chip cost is reached by reducing the number of integration layers. As the ratio of the IO area over the TSV cell area rises (reflecting fewer and thicker IO channels), the minimum 3D chip cost was achieved with a larger number of layers. This means that with more and thinner IO channels, the minimum 3D chip cost is reached with a smaller number of layers. Moreover, chips with larger die areas generally reach minimum cost with a smaller number of layers.

Table 1 displays the longest horizontal (per layer) and vertical (depth) wire lengths for a 64-core SD NoC. The 64-core SD follows 4 possible considered layouts: 2 layers × 1 × 32 (similar to Figure 6(c) but with 32 cores per layer), 2 layers × 4 × 8 (similar to Figure 7(b)), 4 layers × 2 × 8 (similar to Figure 7(c)), and 4 layers × 1 × 16 (similar to Figure 8(a)). The layouts vary in number of layers (2–4) and in number of cores per layer, and in SD(N_1, N_0) network configuration. The longest horizontal and vertical wire lengths are combined to estimate the longest wire lengths in NoC-connected 3D chips. The chip areas, TSV height and area, and number of TSB arrays values are borrowed or interpolated from (Shukla et al., 2019). The 3D chip costs are qualitatively stated and broken down into 3D integration costs, die area cost, and TSV array cost. It is assumed that the individual length of the 2-layer flipper wires (red and green in Figure 7(b)) and inter-SD spacings in each layer are 20 µm, each.

Table 1 Longest horizontal and vertical wire lengths and 3D chip costs with 3D SD NoC

64 cores placement	*2×1×32* (*Figure 6*(*c*))	*2×4×8* (*Figure 7*(*b*))	*4×2×8* (*Figure 7*(*c*)	*4×1×16* (*Figure 8*(*a*))
# Layers	2	2	4	4
# TSV arrays	64	64	32	32
Total TSV height (µm)	140	140	280	280
TSV area per layer (µm^2)	64000000	64000000	32000000	32000000
Chip area (µm^2)	100000000	100000000	50000000	50000000
Chip area (cm^2)	1	1	0.5	0.5
Chip area -TSV area (µm^2)	36000000	36000000	18000000	18000000
Chip area -TSV area (cm^2)	0.36	0.36	0.18	0.18
Width (µm)	6000	6000	4242.6	4242.6
Length (µm)	6000	6000	4242.6	4242.6
Longest horizontal length (µm)	6820	2900	2060	3660
Longest vertical length (µm)	140	140	280	280
Total longest (wire) distance	6960	3040	2340	3940
Diameter (hops)	10	6	6	6
Integration cost (# layers)	$	$	$$	$$
Chip area cost	$$	$$	$	$
TSV array cost	$$	$$	$	$

The total longest distance, in μm, is the sum of the longest horizontal length (longest wire per 2D layer) and the longest vertical length (3D IC height). The longest distance is achieved by the 4×2×8 layout followed by the 2×4×8 layout. Table 1 numbers indicate that the longest distance for the first layout is worst, among all explored layouts, given the high number of cores (32) to be crossed to reach one side of the SD to the other. Considering the 3D IC cost, the 4-layer integration cost of the 4x2x8 layout exceeds that of the 2-layer 2×4×8 placement, but the die area cost (layer) and TSV array cost are higher for the 2×4×8 layout.

6 Conclusion

3D NoCs are necessary for communicating data and signals in 3D integrated circuits. As 2D chip implementations have run out of room for solving heat problems and 2D die size of many-core chips and MPSoCs, 3D integration of such complex integrated circuits offers a bright hope. In this paper, different 3D chip layouts of the Spidergon-Donut (SD) NoC are explored and compared varying in the number of integration layers and number of SD(N_0)s per layer. For a total of 64 cores, the 4×2×8 and 2×4×8 layouts result in the best longest wire delays, with the former incurring higher 3D integration costs, while the second requiring larger chip area and TSV array costs. Similar analysis can be repeated for many-core chips and MPSoCs of various core numbers. Such study helps in guiding 3D integration direction and weighing 3D NoC layout and placement alternatives.

Future work involves a power and heat dissipation study for the various 3D NoC placements.

References

Abadal, S., Iannazzo, M., Nemirovsky, M., Cabellos-Aparicio, A., Lee, H. and Alarcon, E. (2014) 'On the area and energy scalability of wireless network-on-chip: a model-based benchmarked design space exploration', *IEEE/ACM Trans. Netw.*, Vol. 23, No. 5, p.1.

Albughdar, M. and Mahmood, A. (2015) 'Maximally adaptive, deadlock-free routing in Spidergon-Donut network for large multicore NOCs', *Proc. 14th International IEEE Symposium on Parallel and Distributed Computing*, IEEE, Limassol, Cyprus.

Asaduzzaman, A., Sibai, F.N. and El-Sayed, H. (2013) 'Performance and power comparisons of MPI vs pthread implementations on multicore systems', *Proc. 9th International IEEE Conference on Innovations in Information Technology*, Al Ain, UAE, pp.1–6.

Bahn, J.H., Lee, S.E. and Bagherzadeh, N. (2007) 'Design of a router for network-on-chip', *Int. J. High Perform. Syst. Archit.*, Vol. 12, pp.98–105.

Beckmann, A. and Wood, D. (2003) 'Tlc: Transmission line caches', *Proc. 36th IEEE/ACM Int. Symp. Microarchitecture*, IEEE/ACM, San Diego, CA, pp.43–54.

Black, M.A., Brekelbaum, N., DeVale, J., Jiang, L., Loh, G.H., McCauley, D., Morrow, P., Nelson, D.W., Pantuso, D., Reed, P., Rupley, J., Shankar, S., Shen, J. and Webb, C. (2006) 'Die stacking (3D) 'microarchitecture', *Proc. IEEE/ ACM Int. Symposium on Microarchitecture*, IEEE/ACM, December, Orlando, Florida, pp.469–479.

Cao, K., Zhou, J., Wei, T., Chen, M., Hu, S. and Li, K. (2019) 'A survey of optimization techniques for thermal-aware 3D processors', *Journal of Systems Architecture*, Vol. 97, pp.397–415.

Carpenter, A., Hu, J., Xu, J., Huang, M., Wu, H. and Liu, P. (2012) 'Using transmission lines for global on-chip communication', *IEEE J*, *Emerging Sel. Top. Circuits Syst.*, Vol. 2, June, pp.183–193.

Chang, M.C.F., Cong, J., Kaplan, A., Liu, C., Naik, M., Premkumar, J., Reinman, G., Socher, E. and Tam, S-W. (2008) 'Power reduction of CMP communication networks via RF-interconnects', *Proc. 41st IEEE/ACM Int. Symp. Microarchitecture*, IEEE/ACM, November, Lake Como, Italy, pp.376–387.

Chao, C-H., Chen, K-C., Yin, T-C., Lin, S-Y. and Wu, A-Y. (2013) 'Transport-layer-assisted routing for runtime thermal management of 3D NoC systems', *ACM Trans. Embedded Comput. Syst* (*TECS*), Vol. 13, No. 1, August, pp.1–22.

Chao, C-H., Jheng, K-Y., Wang, H-Y., Wu, J-C. and Wu, A-Y.T. (2010) 'Thermal-aware run-time thermal management scheme for 3D NoC systems', *Proc. ACM/IEEE Int. Symp. Network-on-Chip (NOCS*), ACM/IEEE, May, Grenoble, France, pp.223–230.

Chen, K-C., Chao, C-H. and Wu, A-Y. (2015) 'Thermal-aware 3D network-on-chip (3D NoC) designs: routing algorithms and thermal managements', *IEEE Circuits and Systems Magazine*, pp.45–69.

Coppola, M., Locatelli, R., Maruccia, G. and Pieralisi, L. (2004) 'Spidergon: A novel on chip communication network', *Proc. Int. Symp. on System on Chip*, Finland, pp.15–18.

Coppola, M., Mafie, F., Grammatikakis, M., Locatelli, R., Maruccia, G. and Pieralisi, L. (2008) *Design of Cost-Efficient Interconnect Processing Units: Spidergon STNoC*, CRC Press.

Cueping, C., Peng, J., Lei, Y., Yue, H. and Zan, L. (2010) 'Through-silicon via (TSV) capacitance modeling for 3D noC energy consumption estimation', *Proc. 2010 10th IEEE Int. Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, IEEE, Shanghai, PRC, pp.815–817.

Das, S., Chandrakasan, A. and Reif, R. (2003) 'Three-dimensional integrated circuits: performance, design methodology, and CAD tools', *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, *ISVLSI*, IEEE, Tampa, Florida, pp.13–18.

Deb, S., Ganguly, A., Pande, P., Belzer, B. and Heo, D. (2012) 'Wireless NoC as interconnection backbone for multicore chips: promises and challenges', *IEEE J. Emerg. Sel. Top. Circuits Syst*., Vol. 2, June, pp.228–239.

Deutsch, S. (2015) *Test and Debug Solutions for 3D-Stacked Integrated Circuits,* PhD Dissertation, Department of Electrical and Computer Engineering, Duke University.

DiTomaso, D., Kodi, A., Kaya, S. and Matolak, D. (2011) 'iwise: inter-router wireless scalable express channels for network-on-chips (nocs) architecture', *Proc. IEEE 19th Annual Symp. High Performance Interconnects (HOTI) IEEE*, August, Santa Clara, California, pp.11–18.

Ganguly, A., Chang, K., Deb, S., Pande, P., Belzer, B. and Teuscher, C. (2011) 'Scalable hybrid wireless network-on-chip architectures for multicore systems', *IEEE Trans. Comput.*, Vol. 60, October, pp.1485–1502.

Gu, P., Li, S., Stow, D., Liu, L., Xie, Y. and Kursun, E. (2016) 'Leveraging 3D technologies for hardware security: opportunities and challenges', *Proc. of the IEEE International Great Lakes Symposium on VLSI*, Boston, MA, pp.347–352.

Hammond, L., Nayfeh, B. and Olukotun, K. (1997) 'The case for a single chip multiprocessor', *IEEE Computer*, Vol. 30, No. 9, September, pp.79–85.

Hsu, H-M., Lee, T-H. and Hsu, C-J. (2012) 'Millimeter-wave transmission line in 90-nm CMOS technology', *IEEE J, Emerging Sel. Top. Circuits Syst.*, Vol. 2, June, pp.194–199.

Hsueh, T-Y., Yang, H-H., Wu, W-C. and Chi, M.C. (2011) 'A layer prediction method for minimum cost three dimensional integrated circuits', *Proc. 12th Int'l IEEE Symposium on Quality Electronic Design*, Santa Clara, California, pp.1–5.

Hung, W., Link, G., Xie, Y., Vijaykrishnan, N. and Irwin, M. (2006) 'Interconnect, and thermal-aware floorplanning for 3D microprocessors', *Proc. IEEE International Symposium on Quality Electronic Design*, IEEE, San Jose, California, pp.98–104.

Iqbal, M.A., Macha, N.K., Danesh, W., Hossain, S. and Rahman, M. (2019) 'Thermal management and mitigation techniques in fine-grained 3-D. Integrated circuits', *Microelectronics Journal*, Vol. 91, pp.61–69, https://arxiv.org/ftp/arxiv/papers/1803/1803.03727.pdf

Jerraya, A. and Wolf, W (Eds.) (2004) *Multiprocessor SoCs*, M. Kaufmann.

Karkar, A., Al-Dujaily, R., Yakovlev, A., Tong, K. and Mak, T. (2012) 'Surface wave communication system for on-chip and off-chip interconnects', *Proc. 5th Int. Workshop on Network on Chip Architectures (NoCArc'12)*, New York, NY, pp.11–16.

Karkar, A., Dahir, N., Al-Dujaily, R., Tong, K., Mak, T. and Yakovlev, A. (2014) 'Hybrid wire-surface wave architecture for one-to-many communication in networks-on-chip', *Proc. Design, Automation and Test in Europe Conf. and Exhibition*, IEEE, March, Dresden, Germany, pp.1–4.

Karkar, A., Mak, T., Tong, K-F. and Yakovlev, A. (2016) 'A survey of emerging interconnects for on-chip efficient multicast and broadcast in many-cores', *IEEE Circuits and Systems Magazine*, Vol. 16, No. 1, pp.58–72.

Karkar, A., Tong, K., Mak, T. and Yakovlev, A. (2015) 'Mixed wire and surface-wave communication fabrics for decentralized on-chip multicasting', *Proc. Design, Automation and Test in Europe Conf. and Exhibition*, March, San Jose, California, pp.794–799.

Kirman, N., Kirman, M., Dokania, R., Martinez, J., Apsel, A., Watkins, M. and Albonesi, D. (2006) 'Leveraging optical technology in future bus based chip multiprocessors', *Proc. 39th IEEE/ACM Int. Symposium on Microarchitecture*, IEEE/ACM, December, Paris, France, pp.492–503.

Magen, N., Kolodny, A., Weiser, U. and Shamir, N. (2004) 'Interconnect-power dissipation in a microprocessor', *Proc. International Workshop on System Level Interconnect Prediction*, Paris, France, pp.7–13.

Miller, D. (2009) 'Device requirements for optical interconnects to silicon chips', *Proc. IEEE*, Vol. 97, July, pp.1166–1185.

Shukla, P., Coskun, A., Pavlidis, V. and Salman, E. (2019) 'An overview of thermal and opportunities for monolithic 3D ICs', *Proc. ACM Great Lakes Symposium on VLSI*, ACM, May, Tysons Corner, Virginia, 6 pages.

Shulaker, M.M., Hills, G., Park, R.S., Howe, R.T., Saraswat, K., Wong, H-S.P. and Mitra, S. (2017) 'Three-dimensional integration of nanotechnologies for computing and data storage on a single chip', *Nature*, Vol. 547, pp.74–78.

Sibai, F.N. (2007) 'Performance analysis and workload characterization of the 3dmark05 benchmark on modern parallel computer platforms', *ACM SIGARCH Computer Architecture News*, Vol. 35, No. 3, pp.44–52.

Sibai, F.N. (2008) 'Area-efficient floorplans and interconnects for homogeneous multi-core architectures', *Int. J. High Perform. Syst. Archit.*, Vol. 1, No. 3, pp.155–162.

Sibai, F.N. (2008) 'Evaluating the performance of single and multiple core processors with PCMARK®05 and benchmark analysis', *ACM SIGMETRICS Performance Evaluation Review*, Vol. 35, No. 4, pp.62–71.

Sibai, F.N. (2011) 'Design and evaluation of low latency interconnection networks for real-time many-core embedded systems', *Computers and Electrical Engineering*, Vol. 37, No. 6, pp.958–972.

Sibai, F.N. (2012) 'A two-dimensional low-diameter scalable on-chip network for interconnecting thousands of cores', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 23, No. 2, February, pp.193–201.

Sibai, F.N., El-Moursy, A. and Mohamed, N. (2012) 'Throughput and latency analysis of the Spidergon-Donut interconnection network', *Proc. International IEEE Conference on Innovations in Information Technology (IIT) IEEE*, pp.356–360.

Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschwiler, T. and Atienza, D. (2010) '3D-iCE: fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling', *Proc. IEEE ICCAD*, IEEE, San Jose, California, pp.463–470.

Wei, H., Wu, T.F., Sekar, D., Cronquist, B., Pease, R.F. and Mitra, S. (2012) 'Cooling three-dimensional integrated circuits using power delivery networks', *Proc. 2012 International Electron Devices Meeting*, San Francisco, California, pp.14.2.1–14.2.4.

Zhao, D. and Wu, R. (2012) 'Overlaid mesh topology design and deadlock free routing in wireless network-on-chip', *Proc. 6th IEEE/ACM Int. Symp. Networks on Chip (NoCS)*, IEEE, May, Copenhagen, Denmark, pp.27–34.

Website

Three Dimensional Integrated Circuits, Elsevier, https://www.sciencedirect.com/topics/engineering/three-dimensional-integrated-circuits