MIRA: A Multi-Layered On-Chip Interconnect Router Architecture*

Dongkook Park, Soumya Eachempati, Reetuparna Das, Asit K. Mishra,
Yuan Xie, N. Vijaykrishnan, Chita R. Das
Dept. of Computer Science & Engineering, The Pennsylvania State University, University Park, PA, USA
{dpark, eachempa, rdas, amishra, yuanxie, vijay, das}@cse.psu.edu

Abstract

Recently, Network-on-Chip (NoC) architectures have gained popularity to address the interconnect delay problem for designing CMP / multi-core / SoC systems in deep sub-micron technology. However, almost all prior studies have focused on 2D NoC designs. Since three dimensional (3D) integration has emerged to mitigate the interconnect delay problem, exploring the NoC design space in 3D can provide ample opportunities to design high performance and energy-efficient NoC architectures. In this paper, we propose a 3D stacked NoC router architecture, called MIRA, which unlike the 3D routers in previous works, is stacked into multiple layers and optimized to reduce the overall area requirements and power consumption. We discuss the design details of a four-layer 3D NoC and its enhanced version with additional express channels, and compare them against a (6×6) 2D design and a baseline 3D design. All the designs are evaluated using a cycle-accurate 3D NoC simulator, and integrated with the Orion power model for performance and power analysis. The simulation results with synthetic and application traces demonstrate that the proposed multi-layered NoC routers can outperform the 2D and naïve 3D designs in terms of performance and power. It can achieve up to 42% reduction in power consumption and up to 51% improvement in average latency with synthetic workloads. With real workloads, these benefits are around 67% and 38%, respectively.

1. Introduction

Design of Chip Multi-Processors (CMP) / multi-cores and System-on-Chip (SoC) architectures by exploiting the increasing device density in a single chip is quite complex mainly because it needs a multi-parameter (performance, power, temperature, and reliability) design space exploration. The core of this design lies in providing a scalable on-chip communication mechanism that can facilitate the multi-objective design space tradeoffs. The Network-on-Chip (NoC) architecture paradigm, based on a modular packet-switched mechanism, can address many of the on-chip communication design issues, and thus, has been a major research thrust spanning across several design coordinates. These include high performance [1-4], energy-efficient [5-7], fault-tolerant [8-10], and area-efficient designs [7, 11, 12]. While all these studies, except a few [4, 7, 11, 12], are targeted for 2D architectures, we believe that the emerging 3D technology provides ample opportunities to examine the NoC design space.

3D integration has emerged to mitigate the interconnect wire delay problem by stacking active silicon layers [13, 14]. 3D ICs offer a number of advantages over the traditional 2D design [15-17]: (1) shorter global interconnects; (2) higher performance; (3) lower interconnect power consumption due to wire-length reduction; (4) higher packing density and smaller footprint; and (5) support for the implementation of mixed-technology chips. In this context, several 3D designs, from distributing different logical units among different layers to splitting a unit (such as a processor) into multiple layers, have appeared recently [16]. However, 3D stacking may result in temperature hotspots due to increased power density. Thus, any 3D design should consider the thermal issue in addition to other design parameters.

In this paper, we investigate various architectural alternatives for designing a high performance, energy-efficient, and 3D stacked NoC router, called MIRA. The design is based on the concept of dividing a traditional 2D NoC router along with the rest of the on-chip communication fabric into multiple layers, with the objective of exploiting the benefits of the 3D technology in enhancing the design of the router micro-architecture for better performance and power conservation. Our multi-layer NoC design is primarily motivated by the observed communication patterns in a Non-Uniform Cache Architecture (NUCA)-style CMP [44, 45]. The NoC in a NUCA architecture supports communication

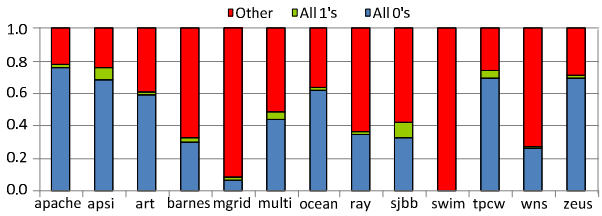


Figure 1. Data Pattern Breakdown

* This work was supported in part by NSF grants, 0702617, EIA-0202007, CCF-0429631, CNS-0509251, and a grant from MARCO / DARPA Gigascale Systems Research Center.

1063-6897/08 $25.00 © 2008 IEEE
DOI 10.1109/ISCA.2008.13

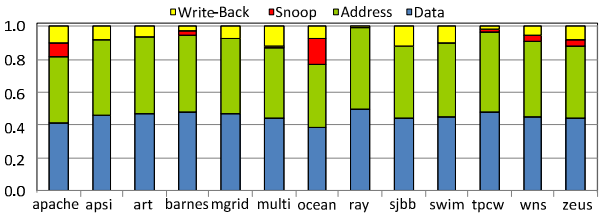


Figure 2. Packet Type Distribution

between the processing cores and the 2nd-level cache banks. The NUCA traffic consists of two kinds of packets, *data* and *control*. The data packets contain certain types of frequent patterns like all 0's and all 1's as shown in Figure 1 [18]. A significant part of the network traffic also consists of short address / coherence-control packets as shown in Figure 2. Thus, it is possible to selectively power down the bottom layers of a multi-layer NoC that have redundant or no data (all 0 word or all 1 word or short address flits), helping in energy conservation, and subsequently mitigating the thermal challenges in 3D designs. Furthermore, such a multi-layered NoC design complements a recent study, where a processor core is partitioned into multiple (four) layers [16]. It was shown that by switching off the bottom three layers based on the operand manipulation characteristics, it is possible to achieve significant power savings. These savings in power result in minimizing the thermal impacts compared to a standard 3D stacking. We believe that if future CMP cores are designed along this line, the underlying interconnect should also exploit the concept of a stacked multi-layered NoC architecture, as proposed here.

In this paper, we investigate three 3D design options for a mesh interconnect:

(1) **3D baseline router (3DB)**: First, we study the direct extension of a 2D NoC into 3D by providing two additional ports in each router for connections in the third dimension. We call this as the 3D baseline router (3DB).

(2) **3D multi-layered router (3DM)**: In this approach, we analyze the design implications of splitting the router components such as the crossbar, virtual channel allocator (VA) and buffer in the third dimension, and the consequent vertical interconnect (via) design overheads. We show that the multi-layered design renders several advantages in terms of reduced crossbar size and wire length with respect to the 3DB design.

(3) **3D multi-layered router with express paths (3DM-E)**: The saving in chip area in 3DM approach can be used for enhancing the router capability, and is the motivation for the third design, which is called a 3D router with express paths (3DM-E). These express paths between non-adjacent nodes reduce the average hop count, and helps in boosting the performance and power behavior.

Using a cycle-accurate simulator, power models from Orion [19], and the Hotspot tool [20] for thermal analysis, we have conducted a comprehensive performance, power, and temperature analysis of the above three architectures and also that of a 2D mesh network using synthetic workloads and application traces. Our analysis reveals that the 3DM and 3DM-E designs are the winners in the multi-objective design space. Due to the structural benefits obtained by utilizing the third dimension, the link traversal stage can be squeezed into the crossbar stage without increasing the cycle time, leading to reduced latency. The 3DM-E shows the best performance with up to 51% latency reduction and up to 42% power savings over the 2D baseline router. Also, it has up to 26% and 37% improvement in latency and power with respect to the baseline 3D router under synthetic workloads. With the application traces, the performance and power benefits are around 38% and 67%, respectively. In addition, our thermal analysis shows that by utilizing the shutdown technique, the average temperature can drop by 1.3K with 50% short flit traffic in the network.

The rest of the paper is organized as follows. The related work is discussed in Section 2 and Section 3 describes three design options for 3D router architectures. The experimental platform and the performance results are presented in Section 4, followed by the concluding remarks in Section 5.

2. Related Work

The prior work can be categorized into two sub-sections: NoC Router Architectures and 3D Integration Techniques.

2.1 NoC Router Architectures

Due to the resource constrained nature of an NoC router, many researchers have focused on two major themes; improving the performance and reducing power consumption. Performance can be improved by smart pipeline designs with the help of advanced techniques such as look-ahead routing [21] and path speculation [1, 2]. Also, dynamic traffic distribution techniques [3, 22] can be used to reduce the contention during the switch arbitration, reducing overall latency. These architectures can also save the crossbar power consumption by decomposing a monolithic crossbar into smaller sub-crossbars. Recently, [23] proposed a shared buffer design that achieves improved performance by dynamically varying the number of virtual channels. This can improve buffer utilization of a router and consequently maintain similar performance even with half the buffer size, significantly reducing area and power consumption.

3D on-chip interconnection architectures have been investigated by a few researchers [4, 7, 24]. An NoC-Bus Hybrid structure for 3D interconnects was proposed in [7], and [4] proposed a dimensionally decomposed crossbar design for 3D NoCs. The design of on-chip caches can take advantage of the characteristics of 3D interconnection fabrics, such as the shorter vertical connections [25] and increased bandwidth between layers. A CMP design with stacked memory layers was proposed

in [14], where the authors claim to remove L2 cache and connect the CPU core layer directly to the DRAM layer via wide, low-latency inter-layer buses. In addition, [26] has proposed a multi-bank uniform on-chip cache structure using 3D integration and [27] performed an exploration of the 3D design space, where each logical block can span more than one silicon layers.

However, all these approaches assume that planar 2D cores are simply distributed on multiple layers and then a 3D-aware interconnection fabric connects them. None of them have considered the actual 3D design of the interconnect spanning across multiple layers. On the other hand, as proposed in [16], processor cores can benefit from three-dimensional multi-layered design, and the 3DM routers proposed in this work can directly work with such true 3D cores.

2.2 3D Integration Techniques

In a three dimensional chip design, multiple device layers are stacked on top of each other and connected via vertical interconnects tunneling through them [28, 29]. Several vertical interconnect technologies have been explored, including wire bonding, microbump, contactless (capacitive or inductive), and through-silicon-via (TSV) vertical interconnect [30]. Among them, the through-silicon-via (TSV) approach offers the best vertical interconnection density, and thereby, has gained popularity. Wafers can be stacked either Face-to-Face (F2F) or Face-to-Back (F2B) and both have pros and cons. In this work, we assumed the F2B approach with TSV interconnects since it is more scalable when more than two active layers are used.

The move to a 3D design offers increased bandwidth [31] and reduced average interconnection wire length [32], which leads to a saving in overall power consumption. It has been also demonstrated that a 3D design can be utilized to improve reliability [11]. However, the adoption of a 3D integration technology faces the challenges of increasing chip temperature due to increasing power density compared to a planar 2D design. The increased temperature in 3D chips has negative impacts on performance, leakage power, reliability, and the cooling cost. Therefore, the layout of 3D chips should be carefully designed to minimize such hotspots. To mitigate the thermal challenges in 3D designs, several techniques, such as design optimization through intelligent placement [33], insertion of thermal vias [34], and use of novel cooling structures [35] have been proposed. The proposed 3DM / 3DM-E architectures can minimize the formation of hotspots by placing highly active modules closer to heat sink to avoid severe thermal problems.

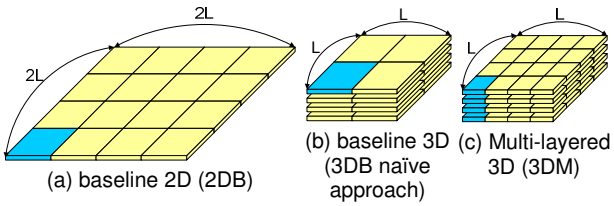


Figure 3. Area Comparison of 2D & 3D architectures with 16 nodes (4×4 in 2DB/3DM and 2×2×4 in 3DB)

3. 3D NoC Router Architectures

On-chip routers are major components of NoC architectures and their modular design makes them suitable for 3D architectures as well. Such a 3D design of an NoC router can help reduce the chip footprint and power consumption, leading to an optimized architecture. In this section, we investigate the three 3D design options for NoC routers that are using mess interconnects: (1) a baseline 3D router (3DB); (2) a 3D multilayered router (3DM); and (3) a 3D multilayered router with express paths (3DM-E). Micro-architectural details of each 3D router design approach are described in the following sub-sections.

3.1 A Baseline 3D Router (3DB)

Towards integrating many nodes into a 3D chip, a naïve approach is to group the nodes into multiple layers and simply stack them on top of each other as shown in Figure 3 (b), which shows 4 layers stacked together, each with 4 nodes, totaling 16 nodes. We define such approach as the baseline-3D (3DB) architecture.

The nodes considered here can be any type of IP blocks (e.g. CPUs or cache banks). However, if a node consumes significant power by itself, for instance a CPU node, it is not desirable to stack them on top of each other since such a design would significantly increase the on-chip temperature due to higher power density. Therefore, a good design approach would be to put all the power-hungry nodes in the top layer (which is closer to the heat sink), while accommodating the other relatively low-power consuming nodes in the lower layers. Thus, if all nodes are power-hungry, then the naïve approach may not be a good design choice. In this paper, we use both CPU and cache nodes and place the CPU nodes only on top layers to avoid thermal problems.

A typical on-chip router consists of six major functional modules; routing computation logic (RC), virtual channel allocation logic (VA), switch allocation logic (SA), crossbar, input buffer, and inter-router link. In comparison to a 2DB design, the 3DB design requires two more physical ports for inter-layer transfers (one to reach an upper layer and one to reach a lower layer). Consequently, the crossbar size increases from 5×5 to 7×7, incurring more wiring overheads. In addition, the size of buffer space and the complexity of routing logic and arbitration logics (in both VA and SA) also increase due to these additional ports. Considering the fact that crossbar and buffer are two major power consuming modules in a router [5], the increase in their power consumption can have significant impact on the overall network power. Therefore, a more optimal 3D router

design is required that can address such problems. The distributed router architecture described in the following section is one of such alternatives.

3.2 A Multi-layered 3D Router Architecture (3DM)

Puttaswamy and Loh [16] proposed a 3D processor design, where the design of individual functional modules spans across multiple layers. Although such a multi-layer stacking of a processor is considered aggressive in current technology, we believe that such stacking will be feasible as the 3D technology matures. We propose a similar on-chip network based on a 3D multi-layered router (3DM) that is designed to span across the multiple layers of a 3D chip. Logically, the 3DM architecture is identical to the 2DB case (see Figure 3 (a)) with the same number of nodes albeit the smaller area of each node and the shorter distance between routers (see Figure 3 (c)). Consequently, the design of a 3DM router does not need additional functionality as compared to a 2D router and only requires distribution of the functionality across multiple layers.

We classify the router modules into two categories – *separable* and *non-separable*, based on the ability to systematically split a module into smaller sub modules across different layers with the inter-layer wiring constraints, and the need to balance the area across layers. The input buffers, crossbar, and inter-router links are classified as separable modules, while arbitration logic and routing logic are non-separable modules since they cannot be systematically broken into sub-modules. The following subsections describe the detailed design of each component.

3.2.1 Input Buffer

Assuming that the flit width is W bits, we can place them onto L layers, with W/L bits per layer. For example, if W=128bits and L=4 layers, then each layer has 32bits starting with the LSB on the top layer and MSB at the bottom layer.

Typically, an on-chip router buffer uses a register-file type architecture and it is easily separable on a per-bit basis. In our approach as shown in Figure 4 (b) and Figure 4 (c), the word-lines of the buffer span across L layers, while the bit-lines remain within a layer. Consequently,

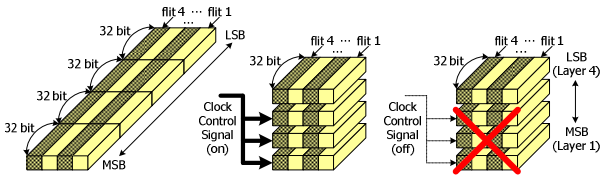


Figure 4. Input Buffer Distribution

the number of inter-layer vias required for this partitioning is equal to the number of word-lines. Since this number is small in the case of on-chip buffers (e.g., 8 lines for 8 buffers), it makes it a viable partitioning strategy. The partitioning also results in reduced capacitive load on the partitioned word-lines in each layer as the number of pass transistors connected to it decreases. In turn, it reduces the driver sizing requirements for the partitioned word-lines [15].

Since buffers contribute about 31% of the router dynamic power [5], exploiting the data pattern in a flit and utilizing power saving techniques (similar to the scheme proposed in [16]) can yield significant power savings.

The lower layers of the router buffer can be dynamically shutdown to reduce power consumption when only the LSB portion has valid data [36]. We define a short-flit as a flit that has redundant data in all the other layers except the top layer of the router data-path. For example, if a flit consists of 4 words and all the three lower words are zeros, such a flit is a short flit. Our clock-gating is based on a short-flit detection (zero-detector) circuit, one for each layer. The overhead of utilizing this technique in terms of power and area is negligible compared to the number of bit-line switching that can be avoided. Figure 4 (b) shows the case, where all the four layers are active and Figure 4 (c) depicts the case, where the bottom three layers are switched off.

3.2.2 Crossbar

In the proposed 3DM design, a larger crossbar is decomposed into a number of smaller multi-bit crossbars positioned in different layers. As shown in Figure 5, the crossbar size and power are determined by the number of input/output ports ("P") and flit bandwidth ("W") and

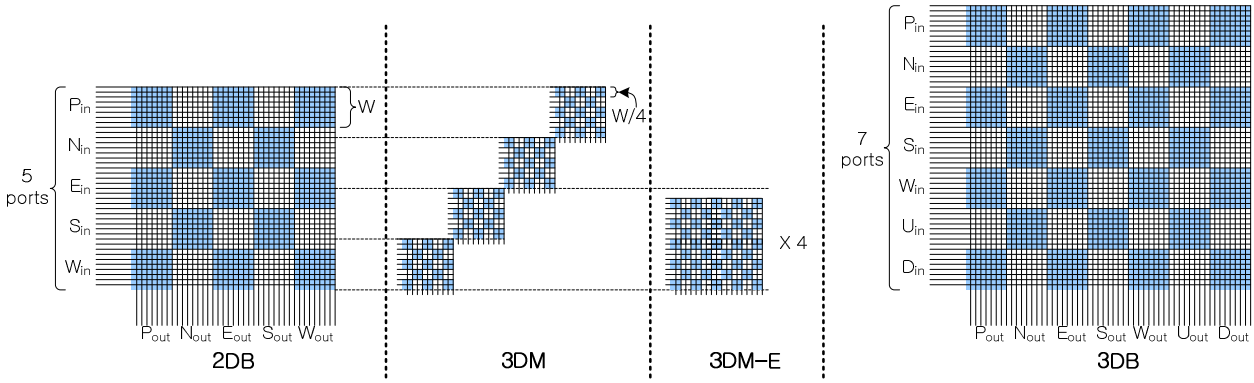
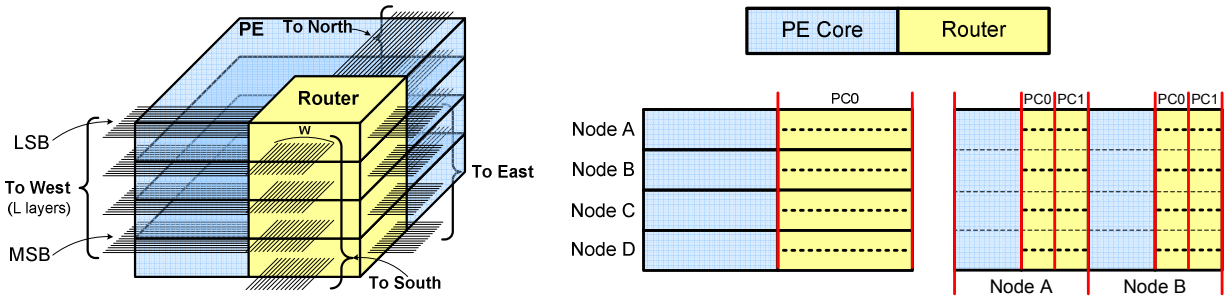


Figure 5. Crossbar Distribution / Relative Area Comparison



(a) Wire distribution (b) 3DB bandwidth usage (c) 3DM bandwidth usage

Figure 6. Inter-router Link Distribution

therefore, such a decomposition is beneficial. In the 2DB case, P=5 and the total size is 5W×5W, whereas in the 3DM case with 4 layers, the size of the crossbars for each layer is (5W/4) × (5W/4). If we add up the total area for the 3DM crossbar, it is still four times smaller than the 2DB design.

We use the matrix crossbar for illustrating the ideas in this paper. However, such 3D splitting method is generic and is not limited to this structure. In our design, each line has a flit-wide bus, with tri-state buffers at the cross points for enabling the connections from the input to the output. In the 3D designs, the inter-layer via area is primarily influenced by the vertical vias required for the enable control signals for the tri-state buffers (numbering P×P) that are generated in the topmost layer and propagated to the bottom layers. The area occupied by these vias is quite small (see Table 1) making our proposed granularity of structure splitting viable.

Table 1. Router component area

Area (μm2)	2DB	3DB	3DM*	3DM-E*
RC	1,717	2,404	1,717	3,092
SA1	1,008	1,411	1,008	1,814
SA2	6,201	11,306	6,201	25,024
VA1	2,016	2,822	2,016	3,629
VA2	29,312	62,725	9,770	41,842
Crossbar	230,400	451,584	14,400	46,656
Buffer	162,973	228,162	40,743	73,338
Total area	433,628	760,416	260,829	639,063
Total vias	0	W= 128	2P+PV+Vk	2P+PV+Vk
Via overhead per layer	0%	0.4%	1.6%	0.6%

** Maximum area in a single layer. k: buffer depth in flits per VC*

3.2.3 Inter-router Link

Inter-router links are a set of wires connecting two adjacent routers, and therefore, they can also be distributed as in the crossbar case above. Assuming that the link bandwidth is W and the number of layers is L, the cross-section bandwidth across L layers is W×L in the 3DB case as shown in Figure 6 (a). To maintain the same cross-section bandwidth in the proposed 3DM case for fair comparison, this total bandwidth should be distributed to multiple layers and multiple nodes. For example, if we assume 4 layers (L=4), the 3DB architecture has 4 separate nodes (A,B,C, and D), with one node per layer, as shown in Figure 6 (b), whereas in the 3DM case, we have only 2 separate nodes (A and B in Figure 6 (c)), since now the floor-plan is only half of the size of a 3DB node. Consequently, in the 3DB design, 4 nodes share 4×W wires; while in the 3DM design, 2 nodes share 4×W wires. This indicates that the available bandwidth is doubled from the perspective of a 3DM node. For example, node A in Figure 6 (c) has 4×(W/2) = 2×W wires available, which doubles the bandwidth over the 3DB case. This extra bandwidth can be used to support topologies with higher connectivity, such as adding one more PC, as will be explored latter in this work.

3.2.4 Routing Computation (RC) Logic

A physical channel (PC) in a router has a set of virtual channels (VCs) and each VC is associated with a routing computation (RC) logic, which determines the output port for a message (packet); however, a RC logic can be shared among VCs in the same PC, since each PC typically takes at most one flit per cycle. Hence, the number of RC logic blocks is dependent on the number of VCs (or PCs if shared) per router. Since the RC logic checks the message header and is typically very small compared to other logics such as arbiters, it is best to put them in the same layer where the header information resides; we can avoid feeding the header information across layers, thereby eliminating the area overheads from inter-wafer vias. In this paper, we fix the number of VCs per PC to be 2 and evaluate the increase in RC logic area and power. The choice of 2 VCs is based on the following design decisions: (i) low injection rate of NUCA traffic (ii) to assign one VC per control and data traffic, respectively (iii) to increase router frequency to match CPU frequency, and (iv) to minimize power consumption. However, the proposed technique is not limited to this configuration.

3.2.5 Virtual-Channel Allocation (VA) Logic

The virtual channel allocation (VA) logic typically performs a two-step operation [37]. The first step (VA1) is a local procedure, where a head flit in a VC is assigned an output VC. If the RC logic determines the output VC in addition to the output PC, this step can be skipped. In

our case, we assume that the RC logic assigns only the output PC and requires P×V V:1 arbiters, where P is the number of physical channels and V is the number of virtual channels. The second step (VA2) arbitrates among the requests for the same output VC since multiple flits can contend for the same output VC. This step requires P×V PV:1 arbiters. As the size of VA2 is relatively large compared to VA1, we place the VA1 stage arbiters entirely in one layer and distribute the P×V arbiters of the VA2 stage equally among different layers. Consequently, we require P×V inter-layer vias to distribute the inputs to the PV:1 arbiters on different layers. It should also be observed that the VA complexity for 3DM is lower as compared to 3DB since P is smaller. Hence, 3DM requires smaller number of arbiters and the size of the arbiters is also small (14:1 vs. 10:1).

3.2.6 Switch Allocation (SA) Logic

In our design, since the SA logic occupies a relatively small area, we keep it completely in one layer to help balance the router area in each layer. Further, the SA logic has a high switching activity due to its per-flit operation in contrast to the VA and RC logics that operate per-packet. Hence, the SA logic is placed in the layer closest to the heat sink.

3.2.7 3DM Router Design Summary

In summary, the 3DM router design has the RC logic, the SA logic and the VA stage1 logic in the layer closest to the heat sink and the VA stage2 logic is distributed evenly among the bottom 3 layers. The crossbar and buffer are divided equally among all the layers. All these designs were implemented in HDL and each of the modules was synthesized using a 90nm TSMC standard cell library. The pitch size for the through-silicon via (TSV) is assumed to be 5×5 μm^2 in dimension (based on technology parameters from [38]). The resulting area of each of the modules is shown in Table 1. Note that in the 3DM design via overhead is less than 2%.

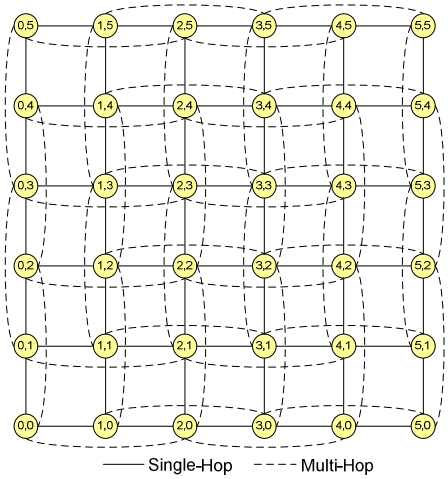


Figure 7. A 6x6 MESH topology with multi-hop links

3.3 An Enhanced Multi-layered 3D Architecture (3DM-E)

The extra wire bandwidth available in the 3DM architecture (refer to Section 3.2.3 and Figure 6 (c)) can be used to support an additional physical port per direction. This extra physical port can be used for purposes such as QoS provisioning, for fault-tolerance, or for express channels. In this work, we use the extra bandwidth to support multi-hop express channels as shown in Figure 7 to expedite the flit transfer [39]. The express topology (3DM-E) requires each router to support 9 physical ports (4×2 ports on cardinal directions and one port to local node). Consequently, it requires additional buffers and a larger crossbar in comparison to the 3DM design. Since 3DM-E components are distributed across multiple different layers, the area in a single layer is still much smaller than that of the 2DB and 3DB designs. For example, the crossbar size in a single layer is 7W×7W for 3DB and (9W/4)×(9W/4) for 3DM-E. Overall, the router area for the 3DM-E case is 2.4 times that of the 3DM case and 0.7 times that of the 2DB case.

3.4 Design Metrics

As discussed earlier, the proposed 3DM architecture has inherent structural benefits over the traditional 2DB and the baseline 3D cases. These benefits can be leveraged for better performance and reduced power consumption. In this section, we explore these benefits and propose architectural enhancements.

3.4.1 Performance Optimization

The performance of a router depends on many factors such as traffic patterns, router pipeline design, and network topology. Among these, we have less control over traffic patterns compared to router pipeline design and network topology. Therefore, optimizing the router in terms of these two aspects will lead to improved performance.

A typical on-chip router pipeline consists of four stages, RC, VA, SA, ST, and one inter-router link-traversal stage ("LT") as shown in Figure 8 (a). Many researchers have proposed techniques to reduce the router pipeline using techniques such as speculative SA (Figure 8 (b)), look-ahead routing (Figure 8 (c)).

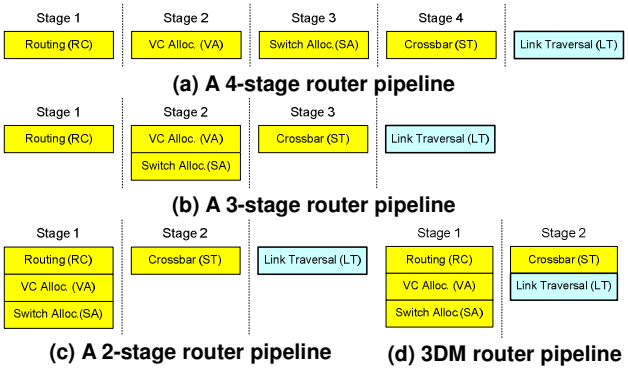


Figure 8. Router Pipeline

In the proposed 3DM router using 4 layers, the distance between two adjacent nodes is halved as compared to the 2DB or 3DB cases, leading to reduced inter-router link delay. Also, the crossbar length is shortened by 1/4 as described in the previous section. This reduces the crossbar wire delay, which is a significant portion of the crossbar delay, and enables the combination of ST and LT stage together (see Figure 8 (d)). Therefore, basically, each hop of the transfer will take one less cycle than the comparable designs using 2DB/3DB. The viability of this design is demonstrated for a 90nm router based on the switch design from [40] scaled to 90nm and a link with optimal buffer insertion using parameters from [41]. We use the design parameters as shown in Table 2 and the resulting delays are shown in Table 3 for a 2GHz router that has a maximum per stage delay of 500 ps.

Table 2. Design Parameters

Link delay per mm	254ps	**Inter-router Link length**	**2DB**	3.1mm
Inverter delay (HSPICE)	9.81ps		**3DM**	1.58mm

Table 3. Delay Validation for pipeline combination

	XBAR (ps)	Link (ps)	Combined Delay (ps)	ST and LT combined
2DB	378.57	309.48	688.05	No
3DM	142.86	154.74	297.60	Yes
3DM-E	182.85	309.48	492.33	Yes

3.4.2 Energy Behavior

The dynamic energy breakdowns of the different router designs were evaluated using the Orion power model [19], and are reported in Figure 9. The 3DM design exhibits the lowest energy consumption due to reduced dimensions (and associated capacitance) of its structures. The biggest savings for 3DM comes from the link energy due the length reduction as described earlier. Further, decomposing the crossbars to smaller parts provides significant energy savings. We observe a 35% reduction in energy for the 3DM case over 2DB. In contrast to 3DM, the energy consumed by 3DB is higher primarily due to the increased number of ports to support communication in the vertical dimension. In the 3DM-E design, the extra ports to support the express links incur energy overheads compared to the 3DM design.

4. Performance Evaluation

In this section, we conduct an in-depth evaluation of the six architectures 2D-Base (2DB), 3D-Base(3DB), 3D-Multilayer (3DM) with switch traversal and link traversal combined into one stage, 3DM without switch traversal and link traversal combining ((3DM(NC)), 3DM with Express paths (3DM-E), and (3DM-E (NC)) with respect to average latency, average power consumption, and thermal behavior. We use a cycle-accurate NoC simulator for the performance analysis. The simulator simulates the router pipeline and adopts worm-hole flow control. We

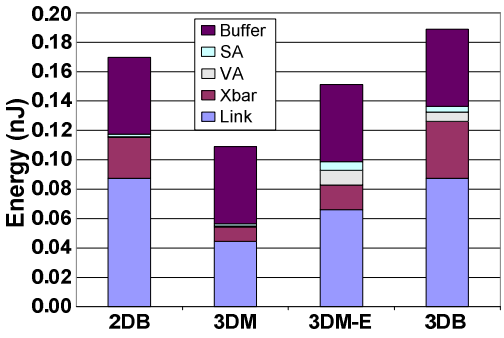


Figure 9. Flit Energy Breakdown

use the X-Y deterministic routing algorithm in all our experiments, and analyze the metrics with both synthetic workloads (uniform random injection rate and random spatial distribution of source and destination nodes) and application traces. The energy consumption of the router modules are obtained from Orion [19] and fed into the cycle-accurate NoC simulator to estimate overall power consumption. For the thermal analysis, we use HotSpot 4.0 [20]. All the cores and routers are assumed to operate at 2GHz. For fair comparison, we keep the bisection bandwidth constant in all configurations.

4.1 Experimental Setup

4.1.1 Interconnection network configuration

For our experiments, we use a 36-node network configuration and used four layers in all 3D cases. Out of the 36 cores, 8 cores are assumed to be processors and other 28 cores are L2 caches. For the CPU, we assume a core similar to Sun Niagara [42] and use SPARC ISA in our Simics simulation [43]. Each L2 cache core is 512KB, and thus, the total shared L2 cache is 14MB.

For the 2DB, 3DM, and 3DM-E cases, we configure a 6x6 2D MESH topology, where the processor cores are spread in the middle of the network as shown in Figure 10 (a) and Figure 10 (b). Also, in 3DM and 3DM-E, we assume that the processor cores can be implemented in a multi-layered (four layers in this study) fashion as proposed in [16]. Although we do not implement the power saving techniques proposed in [16] in this paper, such techniques can also be applied to the cores in 3DM and 3DM-E cases, so that the overall power consumption will be less and temperature reduction will be greater. In our experiments, we assume that all four layers in each

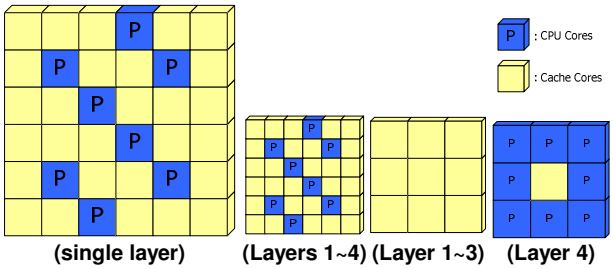


Figure 10. Node Layouts for 36 cores

processor and cache core statically consume the same amount of power. In the 3DB case, we form a 3x3x4 topology and place most of the cache cores in the bottom three layers, while all the processor cores and one cache core are placed in the top (4th) layer so that more active cores stay closer to a heat sink, as shown in Figure 10 (c).

4.1.2 Cache configuration

The memory hierarchy in our experiments consists of a two-level directory cache coherence protocol. While each core has a private write-back first-level (L1) cache, the second-level (L2) cache is shared among all cores and split into banks. These banks are interconnected via the NoC routers. The cache coherence model includes a detailed timing model of the MESI protocol with distributed directories, where each bank maintains its own local directory and the L2 caches maintain inclusion of L1 caches. The memory model is implemented as an event driven simulator to speed up the simulation and have tractable simulation time. The simulated memory hierarchy mimics SNUCA [44, 45] and the sets are statically placed in the banks depending on the low order bits of the address tags. The network timing model simulates all kinds of messages such as invalidates, requests, response, write backs, and acknowledgments. The memory traces were generated by executing the applications on the Simics full system simulator [43]. The memory configurations for our experiments are summarized in Table 4. The workloads used are:

- ***TPC-W:*** *We use an implementation of the TPC-W benchmark [46] from New York University. It consists of two tiers - a JBoss tier that implements the application logic and interacts with the clients and a Mysql-based database tier that stores information about items for sale and client information - under conditions of high consolidation. It models an online book store with 129,600 transactions and 14,400 customers. (tpcw).*
- ***Java Server Workload:*** *SPECjbb. SPECjbb2000 [47] is a Java based benchmark that models a 3-tier system. We use eight warehouses for eight processors; we start measurements 30 seconds after the ramp up time. (sjbb)*
- ***Static Web Serving:*** *Apache. We use Apache 2.0.43 for SPARC/Solaris 10 with the default configuration. We use SURGE [48] to generate web requests. We use a repository of 20,000 files (totaling 500MB) and simulate 400 clients, each with 25 ms think time between requests. (apache)*

Table 4. Memory Configuration

Private L1 Cache: Split I and D cache, each cache is 32KB and 4-way set associative and has 64 bit-lines and 3-cycle access time
Shared L2 Cache: Unified 14MB with 28 512KB banks and each bank has 4 cycle access time (assuming 2GHz clock)
Memory: 4GB DRAM, 400 cycle access time. Each processor can have up to 16 outstanding memory requests.

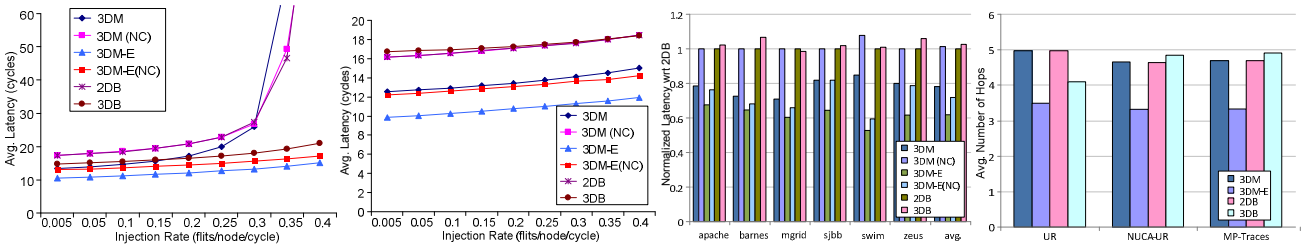
- ***Static Web Serving:*** *Zeus. Zeus [49] is the second web server workload we used. It has an event-driven server model. We use a repository of 20,000 files of 500MB total size, and simulate with an average 400 clients and 25 ms think time. (zeus)*
- ***SPEComp:*** *We used SPEomp2001 [50] as another representative workload. (apsi, art, swim mgid)*
- ***SPLASH 2:*** *SPLASH [51] is a suite of parallel scientific workloads. (barnes, ocean)*
- ***Multimedia:*** *Mediabench. We use eight benchmarks (cjpeg, djpeg, jpeg2000enc, jpeg2000dec, h263enc, h263dec, mpeg2enc, and mpeg2dec) from the Mediabench II [52] suite to cover a wide range of multimedia workloads. We used the sample image and video files that came with the benchmark. To compute throughput, we simultaneously executed all benchmarks on each individual core for 30 back-to-back runs. (multimedia)*

Although we have experimented with the entire set of applications, for clarity, we present results using only six of them that represent different categories of data patterns observed from Figure 1 (shown for all applications); the applications shown in Figure 1 demonstrate diverse percentile of short flits (very low to very high).

4.2 Simulation Results

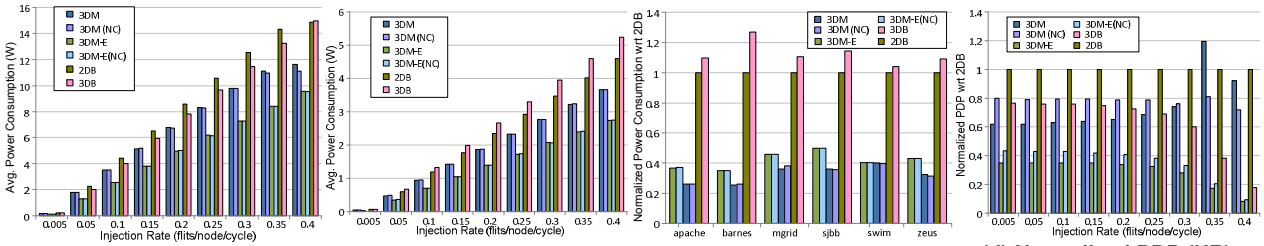
4.2.1 Latency Analysis

We start with the performance analysis by measuring the average latency for the six configurations with the synthetic Uniform Random Traffic (UR), simulated Uniform Random traffic with the NUCA layout constraints (NUCA-UR-Traffic), and six Multiprocessor Traces (MP-Traces). The results are shown in Figure 11 (a) through Figure 11 (c). The UR traffic represents the most generic case, where any node can make requests to any other nodes with uniform probability. This does not capture any specific layout of the cores and is the baseline configuration for all of our comparisons. From the performance perspective for UR traffic, 3DM-E is the best architecture since it has the least average hop counts as shown in Figure 11 (d). Our simulation results corroborate this; 3DM-E has about 26% saving on an



(a) Uniform Random (UR) (b) NUCA-UR (c) Normalized MP Traces (d) Avg. Hop Count

Figure 11. Average Latency Results



(a) Uniform Random (UR) **(b) NUCA-UR** **(c) Normalized MP-Traces** **(d) Normalized PDP (UR)**

Figure 12. Average Power Consumption Results

average over 3DB, 51% over 2DB, and 49% over 3DM at 30% injection rate. Also, the pipeline combination (merging ST and LT stages to a single stage) in 3DM and 3DM-E reduces the latency significantly before the network saturates. Thus, the pipeline combination makes 3DM/3DM-E architectures attractive at low injection rates, which is typical of NUCA networks. Also, for 3DM-E, since the average latency is lower as pointed out before, the network saturates at higher injection rates compared to other architectures, making it more robust even in the saturation region.

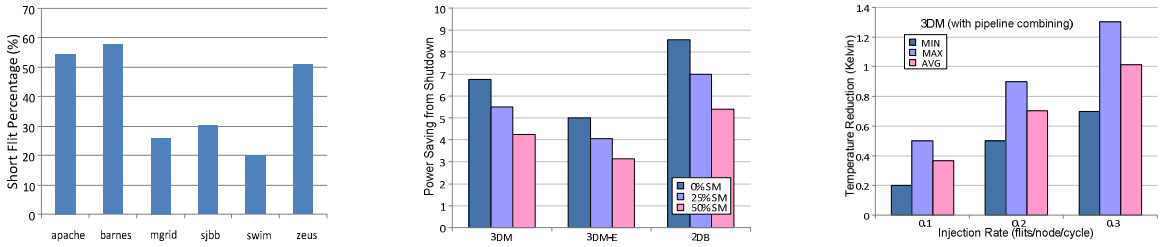
NUCA traffic is typically different from uniform random traffic in a sense that the source and destination sets are constrained. A CPU needs to communicate only with cache nodes and a cache node only needs to communicate with CPU nodes. Thus, the average hop counts will be layout dependant. Hence, we should have different results with MP traces. To capture this layout specific traffic pattern, we also run experiments that model the request-response type bi-modal traffic, where the eight CPU nodes generate requests to the 28 cache nodes with uniform random distribution. Every request is matched with a response to a CPU. The results are shown in Figure 11 (b).

Interestingly, 3DB performance takes a hit for layout dependant traffic simulations. This can easily be explained by analyzing the traffic patterns between different layers. Because of thermal constraints, all the CPU nodes need to be placed in the top layer and most of the cache nodes in the lower layers (refer to Figure 10 (c)). Thus, the average hop count increases as most requests go from the top layer to the bottom layers and similarly, from the bottom layers to the top layer for responses. The trend can be seen in Figure 11 (d), which shows the average hop count for the three kinds of simulations. 3DM-E has the minimal hop count, while both 2D and 3DM have the same hop counts as expected. All these three architectures are almost agnostics to the traffic patterns, while the 3DB configuration suffers with NUCA-UR and MP traces. MP trace simulation results present similar trend in latency behavior. 3DM-E performs around 38% better than 3DB and 2DB on an average. 3DM exhibits on an average 21% and 23% lower latency than 3DB and 2DB architectures, respectively. As expected, the 2DB and 3DM (NC) configurations have similar performance since they have the same logical network layout. Pipeline combination gives 3DM latency reduction up to 14% over 3DM (NC), and 3DM-E results around 23% latency reduction compared to 3DM-E (NC). Thus, the pipeline combination, possible due to smaller crossbars and shorter wires in the proposed 3D architectures, helps in performance improvement.

4.2.2 Power Analysis

Figure 12 (a) shows the average power consumption of the six architectures with 0% short flits. This result shows purely the power improvements of designing the self stacked multi-layered router without any impact of layer shutdown techniques. The 3DM, 3DM-E designs show lower power consumption than 2DB and 3DB. This can be explained by looking at the per flit energy in case of 3DM, 3DB and 2DB as shown in Figure 9. Due to reduced router footprints, crossbar size and link length reduction, the 3DM router has lower power consumption than the 2DB and 3DB architectures. In the 3DM-E case, although the individual router power is higher due to increased radix, because of reduced number of hops via express channels, the overall power consumption has decreased, on an average, by 37% and 42% over the 3DB and 2DB, respectively. 3DM design offers around 22% and 15% power savings over 2DB and 3DB, respectively. Unlike the latency case above, in general, the pipeline combining does not have significant impact on power consumption.

We also evaluate the power saving due to the layer shutdown technique. Figure 13 (b) shows the power savings with 25% and 50% layer shutdown. We save up to 36% power when 50% of the flits are short (32 bits and thus, use only one layer) on an average for the 3DM/3DM-E/2DB configurations. This demonstrates the potential of the shutdown technique. Comparing 3DM-E with 50% short flits with 2DB (0% short flits), we observe a power reduction of 63%. Similarly comparing 3DM with 50% short flits with 2DB (0% short flits), we see a power reduction of 50%. The MP traces (Figure 13 (a)) show up to 58% short flits and on an average 40% of the flits are short. This results in significant power savings (Figure 12 (c)); 3DM-E and 3DM both reduce power consumption around 67% and 70% with respect to 2DB and 3DB, respectively, with no layer shut down in the base cases. The power savings obtained with traces are due to the structural benefits of the self stacked routers and due to the layer shutdown techniques. 3DB



(a) Short Flit Percentage **(b) Power saving Techniques** **(c) Temperature Reduction**

Figure 13. Power and Temperature Simulation Results

exhibits the worst power behavior because of increased hop count / latency per flit.

As a combined metric of both performance and power consumption, we measured the power delay product (PDP), which is the product of delay and power consumption. The normalized PDP result with respect to 2DB is shown in Figure 12 (d), and it shows that 3DM-E and 2DB are the best and worst choices, respectively. At lower injection rates, all 3DM-based techniques performed better than the 2DB and 3DB architectures.

4.2.3 Thermal Analysis

For the thermal analysis, we use the Hotspot simulator [20] which uses the complete layout of the entire chip along with the power numbers for each component. We use the power numbers from the SUN Niagara design (90nm) [42] for the processors and power numbers for the cache memory from CACTI [53]. In our experiments, a processor core consumes 8W and a 512KB cache bank consumes 0.1W. The network power numbers for the six configurations are obtained from Orion and then, fed into our NoC simulator. The NoC simulator generates power trace for Hotspot and for processor cores and cache banks, we assumed static power consumption in power trace. For the multi-layered configurations, the processor and memory powers are divided equally among the four layers.

The shutdown technique can be applied to all four architectures, but the 3DM and 3DM-E architecture would benefit more from reduced overall temperature since lower layers will not be active for short messages, thereby reducing the overall power density. Figure 13 (c) shows the temperature difference between 50% short message and 0% (none) short message cases in the 3DM (with pipeline combining) at three different injection rates. The temperature drops up to 1.3K and on an average, we get up to 1K temperature reduction. Also, as the injection rate increases, we tend to get more temperature reduction. We speculate that this is due to increased number of flit activities in the router, which triggers more activities in separable modules (buffer, crossbar, inter-router links), where we can benefit from short flits. Although the temperature saving is small, it is reasonable since the average power each router consumes is relatively small. Overall chip temperature will depend heavily on CPU/cache core power consumption rather than router power.

We conclude the performance evaluation with a qualitative comparison with the 3D work, most closely related to ours. In [4] the authors proposed a dimensionally decomposable 3D crossbar for designing generic 3D routers, which are suitable for our 3DB layout. Our work differs in that we divide the router data path into multiple layers and provides better opportunity for performance, power and thermal optimizations. In addition, our design is better customized for CMP-NUCA architectures.

5. Conclusions

As the 3D technology is envisioned to play a significant role in designing future multi-core / SoC architectures, it is imperative to investigate the design space of one of the critical components of multi-core systems, the on-chip interconnects, in the 3D setting. In this paper, we propose the design of a multi-layered 3D NoC router architecture, called MIRA, for enhancing the performance, energy efficiency, and thermal behavior of on-chip interconnects. The design of the proposed router is based on the observation that since a large portion of the NUCA communication traffic consists of short flits and frequent patterns, it is possible to dynamically shut down the bottom layers of the multi-layer router to optimize the power consumption while providing better performance due to smaller footprints in the 3D design. Three 3D design alternatives, called 3D-base (3DB), 3D multi-layer (3DM), and 3DM with Express channels (3DM-E) are discussed along with their area and power analysis, using the TSMC 90nm standard cell library and Orion power model. The Hotspot thermal model is used to study thermal behavior of the entire chip with respect to the on-chip interconnect.

A comprehensive evaluation of the performance, power, and thermal characteristics of the three 3D router designs along with the standard 2D design was conducted with synthetic and real traces from scientific and commercial workloads. The experimental results show that the proposed 3DM and 3DM-E designs can outperform the 3D base case and the 2D architectures in terms of performance and power (up to 51% reduction in latency and 42% power savings with synthetic workloads, and up to 38% reduction in latency and 67% power savings with traces). These benefits are quite significant

and make a strong case for utilizing the 3D technology in designing future NoCs for CMP architecture.

We plan to expand our investigation by combining the true 3D processors architecture as proposed in [16] with the proposed 3D router architecture. Also, we will look into other power saving techniques that will help improve the overall power and thermal profiles.

6. References

[1] L. S. Peh and W. J. Dally, "A delay model and speculative architecture for pipelined routers," in Proc. of the High Performance Computer Architecture (HPCA), pp. 255-266, 2001.

[2] R. Mullins, A. West, and S. Moore, "Low-latency virtual-channel routers for on-chip networks," in Proc. of the International Symposium on Computer Architecture (ISCA), pp. 188-197, 2004.

[3] J. Kim, C. Nicopoulos, D. Park, N. Vijaykrishnan, M. S. Yousif, and C. R. Das, "A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip Networks," in Proc. of the ISCA, 2006.

[4] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," in Proc. of the ISCA, 2007.

[5] H. Wang, L.-S. Peh, and S. Malik, "Power-driven Design of Router Microarchitectures in On-chip Networks," in Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture, 2003.

[6] V. Soteriou and L.-S. Peh, "Dynamic power management for power optimization of interconnection networks using on/off links," in Proceedings of the High Performance Interconnects, pp. 15-20, 2003.

[7] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan., and M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," in Proceedings of the ISCA, pp. 130-141, 2006.

[8] T. Dumitras, S. Kerner, and R. Marculescu, "Towards on-chip fault-tolerant communication," in Proc. of the Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 225-232, 2003.

[9] J. Duato, "A new theory of deadlock-free adaptive routing in wormhole networks," IEEE TPDS, vol. 4, pp. 1320-1331, 1993.

[10] D. Park, C. Nicopoulos, J. Kim, N. Vijaykrishnan., and C. R. Das, "Exploring Fault-Tolerant Network-on-Chip Architectures," in Proc. of the Dependable Systems and Networks (DSN), pp. 93-104, 2006.

[11] N. Madan and R. Balasubramonian, "Leveraging 3D Technology for Improved Reliability," in Proc. of MICRO, 2007.

[12] V. F. Pavlidis and E. G. Friedman, "3-D Topologies for Networks-on-Chip," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 15, pp. 1081-1090, 2007.

[13] B. Black, D. W. Nelson, C. Webb, and N. Samra, "3D processing technology and its impact on iA32 microprocessors," In Proc. of the International Conference on Computer Design (ICCD), pp. 316-318, 2004.

[14] T. Kgil, S. D'Souza, A. Saidi, N. Binkert, R. Dreslinski, S. Reinhardt, K. Flautner, and T. Mudge, "PICOSERVER: Using 3D Stacking Technology to Enable a Compact Energy Efficient Chip Multiprocessor," in Proc. of the ASPLOS-XII, 2006.

[15] Y.-F. Tsai, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Three-dimensional cache design exploration using 3DCacti," In Proc. of the International Conference on Computer Design (ICCD), pp. 519-524, 2005.

[16] K. Puttaswamy and G. H. Loh, "Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors," in Proc. of the 13th HPCA, pp. 193-204, 2007.

[17] C. Ababei, Y. Feng, B. Goplen, H. Mogal, T. Zhang, K. Bazargan, and S. Sapatnekar, "Placement and Routing in 3D Integrated Circuits," IEEE Design & Test, vol. 22, pp. 520-531, 2005.

[18] A. R. Alameldeen and D. A. Wood., "Frequent Pattern Compression: A Significance-Based Compression Scheme for L2 Caches," Technical Report 1500, Dept. of CS, Univ. of Wisconsin-Madison, 2004.

[19] H.-S. Wang, X. Zhu, L.-S. Peh, and M. Sharad., "Orion: a power-performance simulator for interconnection networks," in Proceeding of the International Symposium on Microarchitecture, pp. 294-305, 2002.

[20] K. Skadron, M. R. Stan, W. Huang, V. Sivakumar, S. Karthik, and D. Tarjan, "Temperature-aware microarchitecture," in Proc. of ISCA, 2003.

[21] M. Galles, "Scalable Pipelined Interconnect for Distributed Endpoint Routing: The SGI SPIDER Chip," in Proc. of the Hot Interconnects, 1996.

[22] J. Kim, D. Park, T. Theocharides, N. Vijaykrishnan, and C. R. Das, "A low latency router supporting adaptivity for on-chip interconnects," in Proc. of the Design Automation Conference (DAC), pp. 559-564, 2005.

[23] C. A. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M. S. Yousif, and C. R. Das, "ViChaR: A Dynamic Virtual Channel Regulator for Network-on-Chip Routers," in Proceeding of the MICRO, pp. 333-346, 2006.

[24] S. Strickland, E. Ergin, D. R. Kaeli, and P. Zavracky, "VLSI design in the 3rd dimension," VLSI Journal, vol. 25, pp. 1-16, 1998.

[25] A. Zeng, J. Lu, K. Rose, and R. J. Gutmann, "First-order performance prediction of cache memory with wafer-level 3D integration," Design & Test of Computers, IEEE, vol. 22, pp. 548-555, 2005.

[26] K. Puttaswamy and G. H. Loh, "Implementing caches in a 3D technology for high performance processors," Computer Design, 2005. Proceedings. 2005 International Conference on, pp. 525-532, 2005.

[27] Y. Liu, Y. Ma, E. Kursun, J. Cong, and G. Reinman., "Fine Grain 3D Integration for Microarchitecture Design Through Cube Packing Exploration," IEEE International Conference on Computer Design, 2007.

[28] D. Shamik, F. Andy, C. Kuan-Neng, T. Chuan Seng, C. Nisha, and R. Rafael, "Technology, performance, and computer-aided design of three-dimensional integrated circuits," in Proc. of the International symposium on Physical design, 2004.

[29] P. Morrow, M. Kobrinsky, S. Ramanathan, C. M. Park, M. Harmes, V. Ramachandrarao, H. Park, G. Kloster, S. List, and S. Kim, "Wafer-Level 3D Interconnects Via Cu Bonding," in Proc. of the 21st Advanced metallization Conference, 2004.

[30] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," Design & Test of Computers, IEEE, vol. 22, 2005.

[31] C. C. Liu, I. Ganusov, M. Burtscher, and T. Sandip, "Bridging the processor-memory performance gap with 3D IC technology," Design & Test of Computers, IEEE, vol. 22, pp. 556-564, 2005.

[32] J. W. Joyner, P. Zarkesh-Ha, and J. D. Meindl, "A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC)," in Proceedings of the International ASIC/SOC Conference, 2001.

[33] B. Goplen and S. Sapatnekar, "Efficient Thermal Placement of Standard Cells in 3D ICs using a Force Directed Approach," in Proceedings of the ICCAD, 2003.

[34] J. Cong and Z. Yan, "Thermal via planning for 3-D ICs," in Proceedings of the ICCAD, 2005.

[35] D. Bing, P. Joseph, M. Bakir, T. Spencer, P. Kohl, and J. Meindl, "Wafer-level microfluidic cooling interconnects for GSI," in Proceeding of the Interconnect Technology Conference, pp. 180-182, 2005.

[36] M. V. Wilkes, "The best way to design an automatic calculating machine," in The early British computer conferences: MIT Press, 1989, pp. 182-184.

[37] W. J. Dally, "Virtual-channel flow control," in Proceedings of the 17th annual international symposium on Computer Architecture, 1990.

[38] "TSMC Manuals," Synopsys Inc.

[39] W. J. Dally, "Express cubes: improving the performance of k-ary n-cube interconnection networks," IEEE Transactions on Computers, vol. 40, pp. 1016-1023, 1991.

[40] A. Kumar, P. Kundu, A. P. Singh, L.-S. Peh, and N. K. Jha, "A 4.6Tbits/s 3.6GHz Single-cycle NoC Router with a Novel Switch Allocator in 65nm CMOS " in Proc. of the ICCD, 2007.

[41] ITRS, http://www.itrs.net/

[42] P. Kongetira, K. Aingaran, and K. Olukotun, "Niagara: a 32-way multithreaded Sparc processor," Micro, IEEE, vol. 25, pp. 21-29, 2005.

[43] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner, "Simics: A full system simulation platform," IEEE Computer, vol. 35, pp. 50-58, 2002.

[44] B. M. Beckmann and D. A. Wood, "Managing Wire Delay in Large Chip-Multiprocessor Caches," in Proc. of the MICRO, pp. 319-330, 2004.

[45] K. Changkyu, B. Doug, and W. K. Stephen, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," in Proc. of the ASPLOS-X, pp. 211-222, 2002.

[46] TPC-W, http://cs.nyu.edu/totok/professional/software/tpcw/tpcw.html.

[47] "SPECjbb2000 Java Benchmark", http://www.spec.org/osg/ jbb2000/.

[48] B. Paul and C. Mark, "Generating representative Web workloads for network and server performance evaluation," in Proceedings of the ACM SIGMETRICS, 1998.

[49] ZEUS,http://www.zeus.com/products/zws/.

[50] V. Aslot, M. J. Domeika, R. Eigenmann, G. Gaertner, W. B. Jones, and B. Parady, "Specomp: A new benchmark suite for measuring parallel computer performance," in Proceedings of the WOMPAT, pp. 1-10, 2001.

[51] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta,"The SPLASH-2 programs: characterization and methodological considerations," in Proceedings of the ISCA, 1995.

[52] MediaBench II, http://euler.slu.edu/fritts/mediabench//.

[53] D. Tarjan, S. Thoziyoor, and N. P. Jouppi, "CACTI 4.0," Technical Report, HPL-2006-86, 2006.