On-Chip Interconnection Networks: Why They are Different and How to Compare Them

D. N. Jayasimha, Platform Architecture Research, Intel Corporation (jay.jayasimha@intel.com)
Bilal Zafar*, Dept.of Electrical Engineering-Systems, Univ. of Southern California (bzafar@usc.edu)
Yatin Hoskote, Circuits Research Lab, Intel Corporation (yatin.hoskote@intel.com)

Abstract

The on-chip interconnect network (OCIN) is the primary "meeting ground" for various on-die components such as cores, memory hierarchy, specialized engines, etc. of a chip multiprocessor (CMP). In this paper, we argue that there are essential differences between on-die and the well-studied off-die networks. These differences necessitate the introduction of a new set of metrics related to wiring and embedding, which affects delay, power and overall design complexity. We apply these metrics to analyze various topologies and show interesting tradeoffs and non-intuitive results relating to the choice of topologies, such as, higher-dimensional networks may not be suitable for on-die implementation, on-chip wire bandwidth may not be plentiful as thought, etc. Since this is the first investigation which proposes a new methodology needed for analyzing the emerging area of on-die interconnects for CMPs, the paper concludes with a rich set of open issues.

1 Introduction

The opportunities afforded by Moore's Law (doubling transistor density with each generation) are increasingly challenged by three problems: on-die power dissipation, wire delays, and design complexity [1]. The "tiled" architecture" approach to designing CMPs proposes to alleviate these problems: each die is divided into a large number of identical or close-to-identical tiles [13,18]. Global wires spanning a significant portion of the chip are usually avoided to mitigate wire scaling problems. Each tile has relatively low power dissipation and can, for example, be a CPU core tile, a cache tile, a specialized engine tile or some combination of the three. These tiles are laid out in horizontal and vertical dimension by abutment using some interconnect topology – an example tiled architecture is shown in Figure 1. This modular approach enables ease of layout and rapid integration of different IP blocks. These overwhelming advantages mandate the use of a tiled architecture over other approaches.

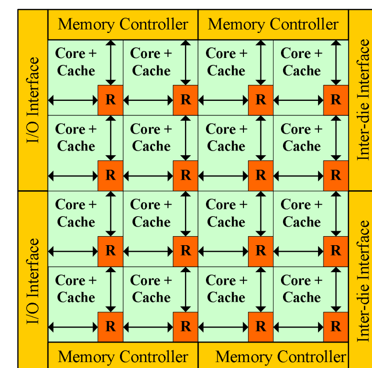

Figure 1: A 16-core tiled CMP with router (R) at each tile

In this paper, the design space of OCIN topologies for a tiled many-core architecture supporting tens to low hundreds of cores is explored. Off-chip interconnects have been extensively studied and there are standard textbooks on this topic [2-4]. We argue, however, that there are essential differences between OCINs and their off-chip counterparts. These differences, discussed in Section 2, necessitate the introduction of *new metrics* related to wiring, embedding, and power which are discussed in Section 3. In Section 4 several well known topologies are explored and compared using the quantitative metrics defined in Section 3 as well as certain qualitative measures important for OCINs. The analysis using the new metrics shows interesting tradeoffs and non-intuitive results relating to the choice of topologies for on-die interconnect, and shows why certain topologies can be weeded out while others merit more detailed investigation.

We believe that this is one of the first efforts to systematically look at *how* OCINs are different from off-chip networks and propose a methodology based on a new set of metrics, in addition to the traditional ones, to analyze them. We expect that this methodology will evolve as the emerging area of OCINs for CMPs matures. Hence, Section 5 not only summarizes our study but poses a rich set of open issues.

*Work done when author was an intern with Platform Architecture Research, Intel Corp.

** Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Other names and brands may be claimed as the property of others.
Copyright © 2006, Intel Corporation. All rights reserved.

2 OCIN: What is different?

2.1 Wiring

In off-chip networks, wires are intra-board (using printed circuit boards), inter-board (through backplanes), and across cabinets (through cables). Links are not wide primarily because of pin-out limitations (Table 1). There is flexibility in the wiring layout arising from physical reality (3D space, many layer boards, multiple boards, relative freedom of wiring with backplanes and cabinets) and wire length is not a first order concern. In OCINs, on the other hand:

- Wiring is constrained with metal layers laid out in horizontal and vertical directions only. Typically, only the two upper layers are available for interconnect wiring (in the future, this could increase to 4, which nonetheless still imposes limitations). These layers are also shared with the power and clock grids.
- Wire lengths need to be short and fixed (spanning only a few tiles) to keep the power dissipation and the resistive-capacitive (RC) product low – this allows wires to get the same benefits as logic from technology scaling [9].
- Wires should not "consume space", i.e., they should not render the die area below unusable – silicon continues to be at a premium. This implies that only a portion of the tile (over the last level cache, for example) is available for OCIN wiring. Dense logic (CPU, first level cache, etc.) provide less opportunity for routing global interconnect wires and, especially, for repeater insertion. Hence, wiring density or "wires per tile edge" becomes an important measure. As will be seen, wiring density can preclude higher dimensional topologies as OCINs, even though they have nice abstract properties (e.g., low diameter).
- There is a notion of directionality: topologies that distribute wire bandwidth per node more evenly across all four edges of each tile are more attractive, since they avoid unbalanced wiring congestion in any specific direction. For example, with a ring topology, wires tend to get used primarily in one direction only.

Off-die Multiprocessors	Link (bits)	On-die Multicore	Link (bits)
Intel ASCI Red Paragon	16	MIT RAW (Scalar, Dynamic)	256, 32
IBM ASCI White (SP Power3)	9	STI Cell BE	144
Cray XT3 (SeaStar)	12	TRIPs (Operand, non-Operand)	110, 128
IBM Blue Gene/L	1	Sun UltraSparc T1	128

Table 1: Comparison of link widths in off- and on-chip interconnects.

2.2 Embedding in 2D

Every topology needs to be laid out in 2D on die (we do not consider "3D stacking" since this area of research is still at a nascent stage). This means that with higher (greater than 2D) dimensional networks, *topological adjacency does not lead to spatial adjacency*. This has significant implications both on the wire delay and the wiring density. Consider 2D embedding of a 3D mesh, as shown in Figure 2. For the longest path (from 0,0 (brown) to 3,3(green)) the topological distance is 9 hops but 3 of these hops *span half the length of the chip*! Therefore, the distance in tile-span units is 18 (3*4 + 6*1). Furthermore, long wires could affect the frequency of operation and will impact the link power. Finally, some of the tiles towards the center of the embedded graph (Figure 2b) have up to four bidirectional links crossing at least one of their edges, while tiles around the edges have one. Large number of links crossing a tile edge may force the link width to be less than that required by the architecture.

2.3 Latencies are of a different order

Consider the ratio of the latency of the link (setup, transmitter drive, actual transfer on wires, receiver and handoff) to the router (including buffering, switching and flow control) for a flit: In classical MP systems, it is of the order of 2-4 for single board systems and increases with the introduction of the backplane and cabinets. In OCINs, this ratio is 1 at best (in unbuffered networks) and is likely to be 0.2 to 0.4. In addition, in a single-die tiled architecture with tens or hundreds of cores, the latency through the interconnect (roughly, number of hops × (router delay + wire delay)) will be a significant portion of the total protocol message latency. For these two reasons, *it is critically important to minimize the interconnect latency through proper router design under both no-load and loaded scenarios*.

2.4 Power

Power is a major factor to consider in OCINs since the power consumption of the interconnect relative to the rest of the components on-die is a much higher than in classical MP systems. We expect that at least 10-15% of the die power budget will be devoted to the interconnect. OCINs for research prototypes, where power optimization is not a first-order objective, have shown even higher percentages. For example, in the MIT Raw processor, interconnection

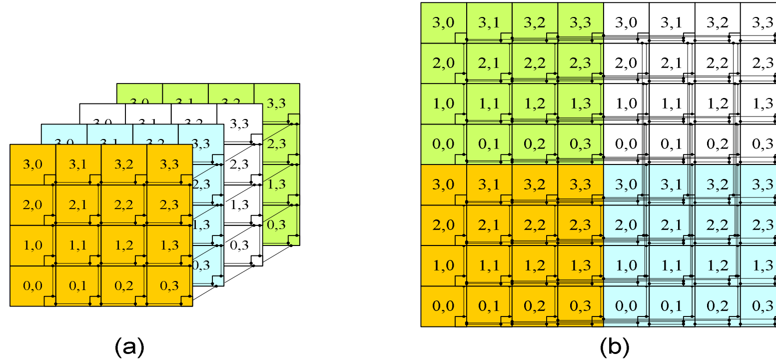

Figure 2: 2D embedding of a 64-node 3D-Mesh network

network consumes almost 35% of the total chip power [20]. Note that with OCINs, interconnect power grows with the number of tiles. Studies have shown that for many configurations, the interconnect power is roughly evenly distributed amongst link, crossbar, and buffers in ODIs [12], while link power is the overwhelming component in off-die networks. Combined with the fact that OCIN power dissipation happens within the die, requiring efficient and costly heat removal, this calls for new techniques to deal with OCIN power.

3 Metrics

While comparing different network topologies it is often difficult to find suitable metrics or weighted set of metrics that a designer must use [5,6]. In this paper, in addition to traditional metrics used to compare topologies, we present a new set of metrics that can be used to compare topologies for on-die networks. The requirements for this set of measures are the following:

- Various topologies of interest, with appropriate parameters, can be meaningfully compared without requiring extensive calculations or simulations to evaluate. The expectation is to use these quantitative measures to narrow down the set of interesting topologies.
- Conclusions based on these measures are relatively robust and do not change significantly the choice of topology, for example, with changes to the underlying micro-architecture
- Finally, using additional qualitative measures, requirements and judgment, it should be possible to further prune the set down to 1 or 2 topologies for detailed investigation.

3.1 Assumptions

- To make the analysis fair in terms of the total bandwidth available at a router node, the total number of wires entering or leaving a router is held constant across all topologies (as shown in Figure 3).

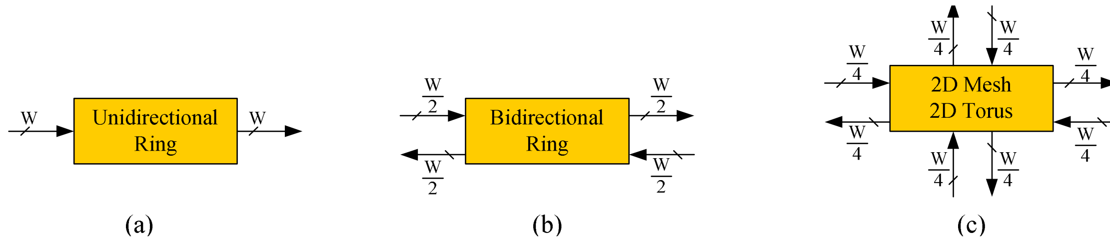

Figure 3: Illustrating equivalent wiring budget across (a) unidirectional ring, (b) bidirectional ring and (c) 2D Mesh/Torus topologies

- With regards to tile size, it is assumed that in the 45nm process, each tile with a low power core (and with L1, L2 caches) will be approximately 4 - 6 sq mm in area (other researchers have made similar assumptions [17]).

This yields 40 - 60 tiles for a die size of 256 sq mm. Using this data point, we calculate that a tile edge will be approximately 90Kλ to140Kλ in size, where λ is a technology independent parameter equal to half the nominal gate length. The number of tiles will approximately double in successive process generations. We assume that tiles are square and all tiles are of the same size.

- Finally, it is assumed that the interconnect operates at a single frequency (but not necessarily synchronously).

3.2 Wiring Density

Definition: *Tile span* is the length of a single tile. Tile-span is used throughout this paper as a fundamental unit for wire length.

Definition: *Wiring density* is the maximum number of tile-to-tile wires routable across a tile edge. In some contexts, the equivalent measure of the number of links or channels crossing a tile edge will be used.

Wiring density is a quantitative measure of the wiring complexity of a topology. It relies on the estimate of the pitch (width + spacing) for each wire in a particular metal layer. The upper metal layers are typically used for OCIN. We apply two factors to the density calculation:

A) Since the global interconnect is usually not routed over dense logic (e.g., core logic and first level cache), the routable length per edge is a fraction of the total tile edge. This fraction is assumed to be 40-60% (40% (60%) usability implies dense logic occupies 36% (16%) of the tile area).

B) To make allowances for via density, power and clock grids, the effective pitch is increased by another fraction (60%). Consequently, the wiring density is calculated for wire pitches ranging from 16λ to 32λ based on pitch estimates made in prior publications [9, 10, 11]. In Table 2 we consider the two tile edges of 90Kλ and in 140Kλ, respectively, with 40-60% of the tile edge length being used for interconnect wiring.

	Tile Size = 90Kλ × 90Kλ					**Tile Size = 140Kλ × 140Kλ**				
	16 λ	***18 λ***	***24 λ***	***28 λ***	***32 λ***	***16 λ***	***18 λ***	***24 λ***	***28 λ***	***32 λ***
40 %	1125	900	750	643	563	1750	1400	1167	1000	875
50 %	1406	1125	938	804	703	2188	1750	1458	1250	1064
60 %	1688	1350	1125	964	844	2625	2100	1750	1500	1313

Table 2: Maximum number of wires per tile edge

Table 2 shows that, in the worst case, we will be able to route over 550 wires per tile edge. *Thus, for 16B bidirectional links (256 bits), a topology requiring more than 2 channels/tile edge will be problematic* (assuming 1 wire/bit). Thus, contrary to the usual belief, we see that on-chip wire bandwidths could be at a premium. Our analysis assumes relatively small percentage areas for the dense logic. A higher percentage implies an exacerbation of the available wires for the global interconnect. On the other hand, other factors such as changing the aspect ratio of the tile or restricting underlying logic to lower metal layers to enable routing over the logic hierarchy could improve the routability.

3.3 Wire Segments: Measure of Delay, Power

Definition: *Wire segments* (or *wire lengths*) needed to implement a topology are classified as *short*: spanning 1 tile, *medium*: spanning a few tiles (typically 2 or 3), *long*: spanning a fraction of the die (typically 1/4 or more of the die).

The *longest segment* determines the wire delay and, hence, is a determinant of the frequency of operation.

The *total length of wire segments* (in tilespan units) yields the metal layer requirements of the interconnect .

Definition: The *average length of wire segment* = (total length of wire segments needed by the interconnect) / (number of router nodes × number of wires entering a router node)

The *average length of a wire segment* (in tile span units) is an indicator of the link power of the interconnect.

3.4 Router Complexity

The router implementation complexity impacts the cost of design and validation, as well as the area and power consumption. It is primarily determined by whether the network is buffered or not, the number of input and output ports, the number of virtual channels necessary or desired, the crossbar complexity and the routing algorithm. Some of these factors are not easily quantifiable. Additionally, a design may be able to absorb the complexity in favor of desirable properties offered (such as routing path diversity, etc.). *Since the router complexity is thus largely*

implementation- dependent, we believe an important abstract measure to capture this complexity is the degree of the crossbar switch K.

The crossbar switch is the centerpiece of the router in terms of area and functionality. The cost of the crossbar can be measured in terms of:

(i) the arbitration control logic, which increases *quadratically* with degree K because the number of arbiters and the complexity of each arbiter both increase linearly with K,

(ii) the area of the crossbar, which also increases *quadratically* with K because both the X and Y dimensions of the crossbar grow linearly,

(iii) the latency through the crossbar, which grows *linearly* with K since the distance and wire length through the crossbar from an input port to an output port grows linearly with the degree,

(iv) the power consumption, which grows *quadratically* with K because the total wire capacitance within the crossbar grows quadratically with K

Thus, the cost or complexity of the crossbar, and by extension that of the router, is closely tied to its degree.

3.5 Power

In OCINs, the chief sources of power consumption are: (1) the switching activity on the links, (2) the buffers in the router and (3) the crossbar switch. The average wire segment length is an indicator of the link power (Section 3.3). Figure 4 compares link power for various topologies assuming the same switching activity for all topologies. The crossbar degree is an indicator of the switch power (Section 3.4). We (arbitrarily) normalize the switch power to be 1.0 for a 3-port switch (one that would be used in a bidirectional ring). Thus a 5-port router (used in a 2D mesh or 2D torus) would have a switch power of $(5/3)^2 = 2.8$. Another factor to note is that most of this power dissipation occurs due to switching activity in the crossbar and links, not device leakage. Thus, the effect of topology on activity can make a big difference on the total power. Characterizing buffer power is more difficult. For a buffered interconnect, the number of buffers and corresponding power consumption increases linearly with the number of input ports or degree of the router. The number of buffers and how they are managed is extremely design-dependent[1]. Finally, the relative distribution of power among the three sources needs to be taken into account and the proper weighted average computed. Assuming ring as the baseline, the relative trend in power growth is calculated for various topologies based on the above observations. Higher-dimension networks suffer from severe growth in either crossbar or link power. While clever design optimizations could reduce the power for any specific implementation, the chart shows that more effort is required as we increase the degree.

On the other hand, if we compare energy consumption, which is the work done to move packets through the network, the higher average hop count of a network will adversely impact its energy consumption. This factor counters the higher power dissipation for higher dimension networks.

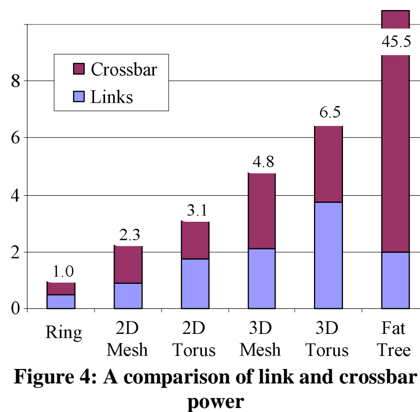

Figure 4: A comparison of link and crossbar power

3.6 Performance Measures

Performance metrics most commonly used in comparing network topologies are average and maximum number of hops and bisection bandwidth. Average and maximum number of hops serve as measures of network latency, even though they are only a partial indication of the actual message latency. The total message latency is the latency through the router times the number of hops. The latency through the router includes the link (wire) latency and it varies with the load: typical conditions are no-load (especially with bypassing of the router pipeline), light load and high load. Under the constant wiring budget assumption (Section 3.1), simpler topologies such as rings could have wider links than those that use richer topologies with high degree crossbars.

[1] Computing a meaningful abstraction is an ongoing area of work and will be addressed in the final version of the paper.

4 Comparing Topologies

In this section, we consider several well known topologies for OCINs. The main characteristics are briefly given for completeness and issues with their on-die implementation are considered. They are then compared quantitatively using the cost metrics introduced in this paper and well known performance metrics (Tables 4 and 5 respectively). With buffered networks, we also assume that each topology has the same virtual channel resources so that they can be used either for deadlock-freedom or for performance enhancements (to prevent head-of-line blocking).

4.1 Bi-directional Ring

Bi-directional ring, which is a 1D torus, is the simplest topology considered in this paper. This simplicity comes at the cost of high average hop count (N/4, where N is the number of nodes) and low bisection bandwidth that remains constant at four uni-directional channels. In case of un-buffered rings routing at each node can be accomplished in a cycle. Since the ring channels are wider (2x that of 2D mesh/torus) and the per-router delay is low the ring is a worthwhile candidate for small N. From the crossbar and wiring complexity issues, the ring has a clear advantage as shown in Table 4.

The main drawbacks of the ring are its behavior as N increases: A) High average hop count- with un-buffered rings this leads to throttling at the source and high latencies, which could further be exacerbated if packets that are unable to sink are misrouted. B) Absence of routing path diversity has the potential to degrade performance under loaded conditions. C) The topology is inherently not fault tolerant to route around router or link failures.

4.2 2D Mesh

2D mesh is an obvious choice for tiled architectures, since it matches very closely with the physical layout of the die. It is a well-understood topology with relatively simple implementation and wiring density of 1 channel per tile edge. A low wiring density means that no stringent constraints on channel width are placed.

Router for 2D mesh requires a careful design since high frequency designs are usually pipelined. Pipeline bypassing at low loads is necessary to achieve low delay through a router [12]. The crossbar power is relatively high compared to the ring – split crossbars (quadratic reduction in crossbar power traded for a linear increase in link power) could be a way to address this problem [15].

One of the main drawbacks of 2D mesh is the non-uniform topology view from node standpoint. That is, less bandwidth is available to nodes at corners and edges (i.e., fewer wiring channels enter/leave node) while these nodes have a higher average distance from other nodes.

4.3 2D Torus

Adding wrap-around links to a mesh creates a torus topology which decreases the average and maximum hop counts and doubling the bisection bandwidth. The wrap-around links, however, also double the number of wiring channels per tile edge to 2. The disadvantage of long wires which span the length of the die is overcome by the technique of "folding" which yields a maximum wires length spanning only 2 tiles.

4.4 3D Mesh

Unlike 2D mesh, the 3D mesh and other higher dimensional topologies do not satisfy the adjacency property. Although, 3D mesh has lower average and maximum hop counts, the topology requires long wires (spanning half the die) and the wiring density is much higher near the center than along the edges, i.e., 4 channels per tile edge as shown in Figure 2. Finally, the crossbar degree of 7 has implications on router complexity, no-load latency and power.

4.5 3D Torus and Higher-Dimensional k-ary n-cube Topologies

3D torus shares some of the advantages of the 2D torus such as symmetry in traffic distribution and improved distance measures. It has disadvantages similar to those for 3D mesh, such as crossbar degree of seven and long wires. The wiring density for 3D torus is even worse than 3D mesh with 5 channels per tile edge.

The wiring density and crossbar degree only gets worse in higher dimensional meshes and tori (k-ary n-cube with n > 3). On these accounts, these topologies have been weeded out. However, there is one variant of the hypercube (2-ary n-cube) – cube connected cycles – which merits further discussion (Section 4.7).

4.6 Direct Fat Tree

Fat tree topology has a unique characteristic that allows all non-conflicting end-to-end connections to have full wire bandwidth. That is, any node can communicate with any other node in the network while having 100% per channel bandwidth available for this communication (as long as no other nodes are communicating with the two nodes involved). Interestingly, implementation of a direct fat tree network does not require long wires-, the maximum wire length is two tilespans. This is possible because tiles can be connected so that the routing resources can be shared.

Figure 5 shows a 64-node direct fat tree network. Of the 16 routers in this arrangement, four have two 4× ports, eight have two 4× and one 8× port, while four have two 4× and two 8× ports to them (all of this, in addition to the ports connecting the four adjacent nodes). Figure 5(c) shows one of the eight routers with one 8x port. It is obvious that the implementation cost of the crossbar in this case is prohibitive. Further, as the network size grows so does the

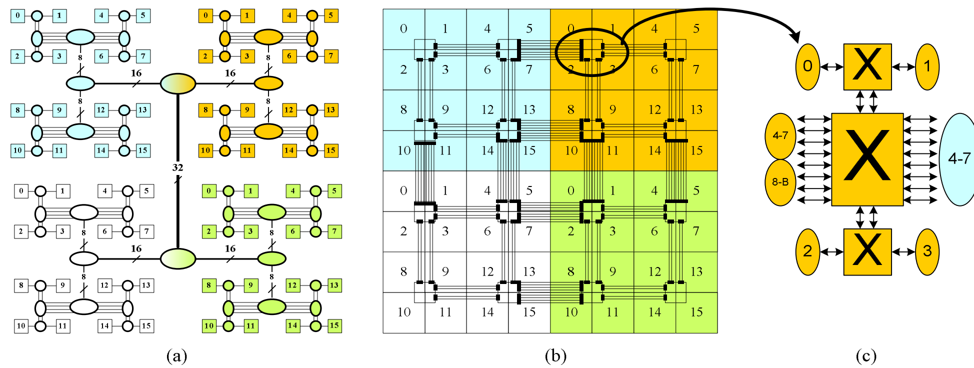

Figure 5: (a) 64-node Fat Tree network, and (b) its corresponding layout, (c) One of the eight routers with one 8x port

complexity (degree) of the router. Hence, this topology is also weeded out.

4.7 Cube Connected Cycles (CCC)

The CCC topology is derived from the hypercube with each node in an n-dimensional hypercube being replaced by a ring of size n. Hence, an n-dimensional CCC has ($n \times 2^n$) nodes.

CCC exhibits some favorable qualities: good bisection bandwidth, acceptable average and maximum latency and moderate crossbar degree of 4. The cons for this topology include unevenly distributed channel bandwidth and the need for some long wires spanning almost half of the chip length. However, the biggest drawback of this topology is that, unlike the other networks discussed so far, adding nodes to the network requires the rewiring and layout of the entire chip, since a node has to be added to each cycle in the hypercube. Finally, the ratio of nodes in an (n+1) dimensional CCC to an n-dimensional CCC is $2(1+1/n)$ - an inconvenient and high ratio for network growth (e..g, n = 4 yields 64 nodes; n= 5 yields 160 nodes), necessitating incomplete networks for configurations of interest.

4.8 Hierarchical Ring Topology

Topologies compared in Sections 4.1 through 4.7 are quite typical in off-chip interconnects and some of them have been commonly proposed for on-die implementation as well. There are, however, variations of these basic topologies that can provide good cost-performance tradeoffs [16]. One such topology is the hierarchical ring topology. In this section, the hierarchical ring topology is discussed to illustrate how such non-traditional topologies can offer interesting alternatives for on-chip implementation.

Arranging rings in a hierarchical manner has the potential to overcome some of the drawbacks of simple bidirectional rings discussed in Section 4.1 while retaining the key advantages of that topology. Hierarchical ring (H-Ring) topologies can be arranged in several different ways, based primarily on the number of levels in the hierarchy (usually two, three or four for network sizes up to 256 nodes). In this paper, we introduce two new ideas that can be used to improve the bisection bandwidth and reduce average latency of the hierarchical ring topology. We also introduce a 3-tuple representation to uniquely describe various H-Ring configurations.

A typical H-Ring network, as proposed in [7][8], consists of three or more *local* rings, where each local ring connects n endnodes. These local rings are connected together in a hierarchical arrangement, culminating at a single *global* ring. Previous studies have found that in off-die multiprocessor systems organization of the memory

hierarchy as well as memory access pattern impacts the optimal hierarchy [8]. In this work we do not attempt to find the optimal arrangement but focus on the cost and performance implications of embedding H-Ring topologies in 2D.

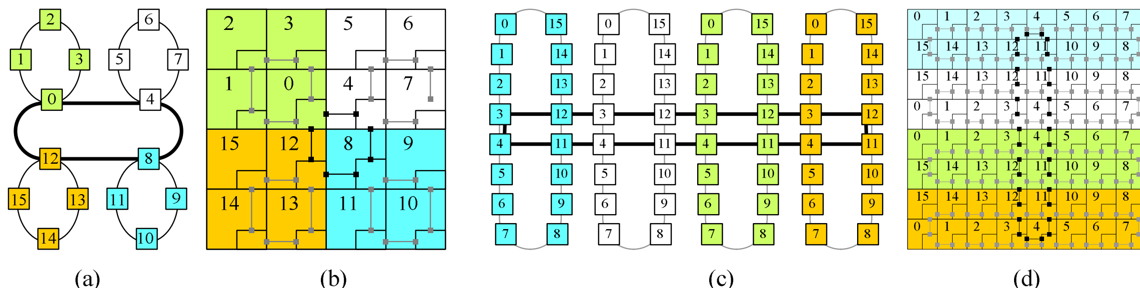

Figure 6: (a) 16-node H-Ring with one global ring and (b) its corresponding layout. (c) 64-node H-Ring with one global ring that has four taps per local ring and (d) its corresponding layout.

Figure 6(a) shows a two-level H-Ring with four nodes in each of the four local rings. Figure 6(b) shows how this 16-node H-Ring can be embedded on a 2D die. Note that maximum length of a wire segment is one tile. This layout can be extended for 32, 64 and higher network sizes. Wire segments longer than one tilespan are not needed so long as the topology is organized as four adjacent local rings.

To improve performance and fault tolerance, we propose two new design parameters to the basic H-Ring topology. The first parameter is the number of "taps" that a non-local ring has on the ring at the next lower level. For example, Figure 6(a) shows a 16-node H-Ring where the global ring has one tap on each global ring. Figure 6(c) shows a 64-node H-Ring where the global ring has four taps on each local ring. Figure 6(d) shows a wire-efficient layout of this topology that does not require wire segments longer than one tile-span. By placing the two pairs of taps at diametrically opposite ends of the local rings, the average distance from an endnode to a tap is reduced, but having four taps increases the average distance on the global ring.

The second parameter we propose is the number of ring in each level of hierarchy. Networks shown in Figures 6(a) and 6(c) both have one global ring connecting four local rings. This may cause the global ring to become a performance bottleneck – a problem that having additional global ring can mitigate.

Finally, in order to uniquely represent the various configurations of H-Ring topologies, we propose a 3-tuple representation for each level of the hierarchy. The 3-tuple is {n, r, t}, where *n* = number of nodes at a particular level, *r* is the number of rings connecting this level to its next lower level and *t* is the number of taps. For example, the 64-node H-Ring network shown in Figure 6(c) is denoted as {16,1,1}: {4,1,4}.

	HR{16,1,1}:{4,1,1}	HR{16,1,1}:{4,1,2}	HR{16,1,1}:{4,1,4}	HR{16,1,1}:{4,2,2}
Maximum Hop Count	18	12	12	10
Average Hop Count	9.5	8.6	5.7	6.8
Bisection Bandwidth	4	4	4	8
No. of Routers w/ Degree 5 & 3	4 & 60	8 & 56	16 & 48	16& 48
Number of Wire Segments	68	68	72	72
Total Wire Segments Length	68	78	80	92
Average Wire Length	1.1	1.2	1.25	1.44

Table 3: Comparison of four 64-node H-Ring topologies

Table 3 presents a comparison of four H-Ring configurations using basic cost and performance metrics. While the metrics used in Table 3 do not capture the improved performance within a local ring, it is clear that for a modest increase in cost, even the simplest H-Ring topologies (namely, {16,1,1}:{4,1,1} configuration) offers modest performance improvement over the simple bi-directional ring. More detailed studies on the performance of H-Ring topologies have shown that for limited switch area budget and high locality in traffic, optimally configured H-Rings can outperform 2D Mesh networks by up to 30% for network sizes of 128 nodes and under [7, 8].

Qualitatively, H-Ring topologies can offer better fault isolation and avoid single point of failure – two of the main drawbacks of the simple ring topology. Furthermore, H-Ring topologies with more than one global rings and taps per local ring can provide marginally higher bisection bandwidth and some path diversity. On the whole, H-Rings have improved performance metrics as compared to the simple ring and compare well with other topologies. However, if applications are not mapped properly to the network's hierarchy, the inter-ring interfaces can become

major performance bottlenecks. Furthermore, since traffic must be buffered at least at the inter-ring interface, ensuring starvation freedom and deadlock avoidance is not as easy in H-Ring topologies as it is in simple unbuffered rings.

4.9 Results

To make the discussions concrete, Tables 4 and 5 present the values for metrics discussed earlier in the paper for an example 64 node network. Table 4 compares the cost of various topologies while Table 5 presents a comparison of the same 64-node example network using commonly used performance metrics. Each entry is color coded with the usual meaning (green: desirable, red: undesirable, yellow: borderline).

	Ring	2D Mesh	2D Torus	3D Mesh	3D Torus	Fat Tree	CCC
Number of Wire Segments: L,M,S	0,0,64	0,0,112	0,96,32	0,128,16	64,64,64	0,128,0	26,8,62
Total Number of Wire Segments	64	112	128	144	192	128	96
Total Wire Segment Length	64	112	224	272	448	256	182
Average Wire Segment Length	1	1.75	3.5	4.25	7	4	2.8
Wiring Density	1	1	2	4	5	4	4
Number of Switches	64	64	64	64	64	16	64
Switch Degree	3	5	5	7	7	28	4
Relative Link & Crossbar Power	1	2.3	3.1	4.8	6.5	44.5	5

Table 4: Comparing the cost of topologies with 64 endnodes.

In addition to the standard performance metrics, Table 5 shows the latency measure for each network under no-load condition. Also, a recently-proposed measure of "effective bandwidth" [4] is used to compare the upper-limit on the effective bandwidth of each of the network under diverse traffic patterns.

	Ring	2D Mesh	2D Torus	3D Mesh	3D Torus	Fat Tree	CCC
Average Hop Count	16	7	4	14	3	9	5.4
Maximum Hop Count	32	14	8	7	6	11	10
Average Latency	16	10.6	8	12	9	38	10.8
Bisection BW	8	16	32	32	43	21	21
Effective BW: Uniform	16	32	64	64	64	42	42
Effective BW: Hot-spot	1	1	1	1	1	1	1
Effective BW: Bit Complement	8	16	32	32	43	21	21
Effective BW: NEWS	22.4	64	64	64	64	64	64
Effective BW: Transpose	7	28	56	56	64	36.8	36.8
Effective BW: Perfect-Shuffle	8	32	64	64	64	42	42

Table 5: Comparison of performance of topologies with 64 endnodes

The following conclusions can be drawn from this investigation:

- The fat tree topology is weeded out, because of the prohibitive crossbar cost among other issues.
- The cube connected cycles has relatively good measures, except for the wiring density. The latter combined with some of the qualitative cons mentioned earlier (scalable design) makes this topology unattractive.
- The 3D torus and 3D mesh have attractive metrics, except for wire density and a high crossbar degree. It is possible that these may become attractive candidates for very large number of cores (high hundreds, low thousands).
- The ring has poor measures, has the potential for interconnect bandwidth wastage, and suffers from other drawbacks (no path diversity, poor fault tolerance). Hence we find this topology not suitable for high node counts.
- The H-Ring, 2D mesh, and 2D torus have attractive topological properties as the table summary shows, though the H-Ring's bisection bandwidth is borderline. The hierarchical ring has some qualitative cons mentioned earlier (poor fault tolerance, susceptibility to performance bottlenecks at interface between the rings, poor path diversity, etc).

The 2D mesh and 2D torus are clear candidates for a more detailed study. While the mesh has simplicity on its side, the torus has superior topological and performance advantages, which could potentially be offset by wiring complexity. The H-ring's organization merits a deeper study requiring the impact of the coherence protocols on the overall complexity, which is outside the scope of this investigation.

5 Conclusions and Open Issues

In this paper, we have argued that OCINs for tiled architectures necessitates the introduction of new metrics related to wiring and embedding which affects delay, power, and overall design complexity. The application of these metrics has shown several interesting results: A) Wires for interconnects are not plentiful and hence bandwidth is not cheap, especially, for higher dimensional networks – this is contrary to normal belief. B) Higher dimensional networks, while topologically appealing, have serious limitations arising from long wire lengths needed to embed them on the 2D Silicon substrate: they could limit the frequency of operation and they consume power. Further, they typically need high degree crossbars which also exacerbates the wiring and power issues. C) The wiring analysis gives an indication of the wiring complexity of each topology. This could be used, for example, in analyzing if additional metal layers are needed to provide a richer interconnect considering the attendent costs and design decisions.

The methodology presented here can not only be used to weed out topologies but also to examine if microarchitectural or circuit optimizations can be done to improve the cost metrics when a topology has other desirable features. For example, if the crossbar power is excessive in a 2D mesh, could alternative designs (e.g., partitioned crossbar) be used to reduce the power? Since this is the first investigation which proposes a new methodology needed for analyzing the emerging area of on-die interconnects for tiled chip multiprocessors, there are a number of interesting areas for further study: A) The buffer component of power in the OCIN has not been considered in the paper. We are working on a proper abstraction of this component to complete the power metrics. B) The energy consumption of networks is another important area which has not been considered in this paper. C) An important class of networks which we have not considered are indirect networks. At first blush, these seem to require long wires spanning half the chip and may have issues with layout since they seem to need skewed aspect ratios. A deeper analysis of indirect networks is merited, however. Additionally, there are a number of other interesting variants of topologies we have not considered. For example, one interesting topology – a variant of the 2D mesh [16] - which reduces the average distance merits a study in accordance with the methodology we have outlined. D) Non-uniform sized tiles will affect router layout and wiring. Whether such aspects can be incorporated as a meaningful metric remains to be seen. E) Our analysis assumes that two metal layers are available for OCINs – the analysis can be extended in a straightforward manner if additional metal layers become available. If they do, different tradeoffs may exist and higher dimensional topologies could become appealing. F) A wire efficient embedding of the 3D mesh is possible (see Figure 7) where wires are at most 2 tilespans long. This embedding has an issue with design re-use. When a topology grows, the layout and wiring of the entire chip needs to be redone. Wire efficient embedding is a topic for further research.

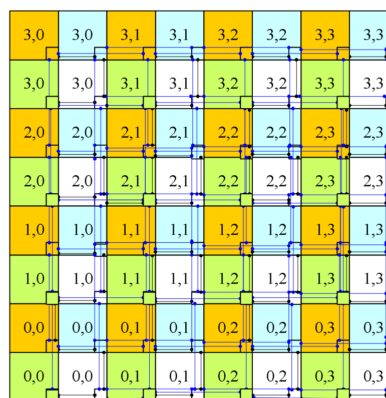

Figure 7: A wire efficient layout of the 3D mesh network shown in Figure 2.

ACKNOWLEDGMENT

We thank Dennis Brzezinski, Mani Azimi, Akhilesh Kumar, Partha Kundu, and Aniruddha Vaidya (Intel Corp.) and Timothy Pinkston (University of Southern California) for helpful comments and suggestions.

References

[1] Burger, Goodman: "Billion Transistor Architectures: There and Back Again," IEEE Computer, March 2004

[2] Dally, Towles, "Principles and Practices of Interconnection Networks", Morgan Kaufmann, 2004.

[3] Duato, Yalamanchili, Ni, "Interconnection Networks: An Engineering Approach", Morgan Kaufmann, 2003

[4] Timothy M. Pinkston and Jose Duato. "Appendix E of Computer Architecture: A Quantitative Approach", 4th Ed., Elsevier Publishers, 2006.

[4] Li-Shiuan Peh, William J Dally, "A delay model and speculative architecture for pipelined routers" Proc. Seventh International Symposium on High Performance Computer Architecture (HPCA-7), January 2001

[5] A. Chien and M. Konstantinidou. Workload and performance metrics for evaluating parallel interconnects. IEEE Computer Architecture Technical Committee Newsletter, Summer- Fall:23 – 27, 1994.

[6] Kathy J. Liszka, John K. Antonio, and Howard Jay Siegel, "Problems with Comparing Interconnection Networks: Is An Alligator Better Than an Armadillo", IEEE Parallel & Distributed Technology: Systems & Technology, Vol. 5, Issue 4, October 1997.

[7] Hamacher, V.C.; Hong Jiang, "Hierarchical ring network configuration and performance modeling ," Computers, IEEE Transactions on , vol.50, no.1pp.1-12, Jan 2001

[8] Ravindran, G.; Stumm, M., "A performance comparison of hierarchical ring- and mesh-connected multiprocessor networks," High-Performance Computer Architecture, 1997., Third International Symposium on , vol., no.pp.58-69, 1-5 Feb 1997

[9] Ho, R.; Mai, K.W.; Horowitz, M.A., "The future of wires," Proceedings of the IEEE , Vol.89, No.4, pp:490-504, Apr 2001

[10] Liqun Cheng, Naveen Muralimanohar, Karthik Ramani, Rajeev Balasubramonian, and John Carter, “Interconnect-Aware Coherence Protocols for Die Multiprocessors”, 33rd International Symposium on Computer Architecture , Boston, June 2006.

[11] Intl’ Technology Roadmap for Semiconductors: 2005 Edition. Semiconductor Industry Association, http://public.itrs.net/home.htm, 2005.

[12] H.S. Wang, L. S. Peh, and S. Malik., “Power-driven Design of Router Microarchitectures in On-chip Networks”, International Symposium on Microarchitecture, pages 105--116, Nov. 2003.

[13] William J. Dally and B. Towles, "Route Packets, Not Wires: On-chip Interconnection Networks," DAC, 2001

[14] Timothy Mark Pinkston and Jeonghee Shin, “Trends toward On-chip Networked Microsystems”, International Journal of High Performance Computing and Networking, 3(1): pp: 3-18, 2005

[15] Kim, J., Nicopoulos, C., and Park, D. “A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip Networks”, SIGARCH Computer Architecture News 34(2) (May. 2006), pp 4-15.

[16] Balfour, J. and Dally, W. J., “Design Tradeoffs for Tiled CMP On-chip Networks”, 20th Annual International Conference on Supercomputing, June, 2006.

[17] Kumar, R., Zyuban, V., and Tullsen, D. M., “Interconnections in Multi-Core Architectures: Understanding Mechanisms, Overheads and Scaling”, In Proceedings of the 32nd Annual international Symposium on Computer Architecture (June 04 - 08, 2005).

[18] Luca Benini, Giovanni De Micheli, "Networks on Chips: A New SoC Paradigm," Computer ,vol. 35, no. 1, pp. 70-78, January, 2002.

[19] Preparata, F. P. and Vuillemin, J. 1981. The cube-connected cycles: a versatile network for parallel computation. Commun. ACM 24, 5 (May. 1981), 300-309.

[20] J. S. Kim, M. B. Taylor, J. Miller, and D. Wentzlaff., “Energy characterization of a tiled architecture processor with on-chip networks”, International Symposium on Low Power Electronics and Design, pp 424–427, 2003.