Design Tradeoffs of Long Links in Hierarchical Tiled Networks-on-Chip

Authors Name/s per 1st Affiliation *(Author)*

Dept. name of organization *(Line 1 of Affiliation - optional)*
Name of organization - acronyms acceptable *(line 2)*
City, Country *(line 3)*
name@xyz.com – optional *(line 4)*

Authors Name/s per 2nd Affiliation *(Author)*

Dept. name of organization *(Line 1 of Affiliation - optional)*
Name of organization - acronyms acceptable *(line 2)*
City, Country *(line 3)*
name@xyz.com – optional *(line 4)*

Abstract — **Hierarchical topologies are frequently proposed for large Networks-on-Chip (NoCs). Hierarchical architectures utilize, at the upper levels, long links of the order of the die size. RC delays of long links might reach dozens of clock cycles in advanced technology nodes, if delay reduction techniques (e.g. wire sizing and repeater insertion) are not applied. Some proposals assume that long links can be adjusted to satisfy timing requirements but lack a deep evaluation of the tradeoffs and costs. Other proposals assume that long links must be pipelined, but do not provide a comprehensive justification.**

In this paper we evaluate the efficiency and the system costs of wire sizing and repeater insertion as methods to reduce link delays in hierarchical NoCs. We present a unified interconnect cost function that accounts for power and wiring overheads of these methods. Then, we quantify the costs of modifying long links in typical hierarchical NoCs for different target clock frequencies and technology nodes. Although long links might undergo aggressive adjustments, we find these overall costs to be low at the system level for many typical cases, taking into account that there are only a few long links in most proposed hierarchical NoC architectures.

Index Terms—**Global interconnect, hierarchical networks on chip, NoCs, long links design tradeoffs.**

I. INTRODUCTION

As the number of modules in System-on-Chip increases, the latency and throughput of pure planar topology NoCs (e.g. 2D mesh) degrade due to the increasing hop distance (number of routers) incurred by long distance (global) packets [1]. Hierarchical schemes [1,2,3,4,5,6] reduce the number of nodes traversed by global packets and therefore provide better scalability. Variable link length is inherent in such topologies and higher hierarchy levels are comprised of longer links. Such long links can reach several millimeters in length. As the RC delay of a link grows rapidly with length [7], RC delays of such links with minimum-size global wires might grow up to many clock cycles.

The approaches to address delays of long links can be divided to three classes: wire sizing and repeaters insertion [8,9,10,11,12]; buffering and pipelining [6], and utilization of techniques such as RF [13], photonic [14] or wave-pipelined fast serial links [15]. Solutions of the third class require radical changes in technology. Although pipelining
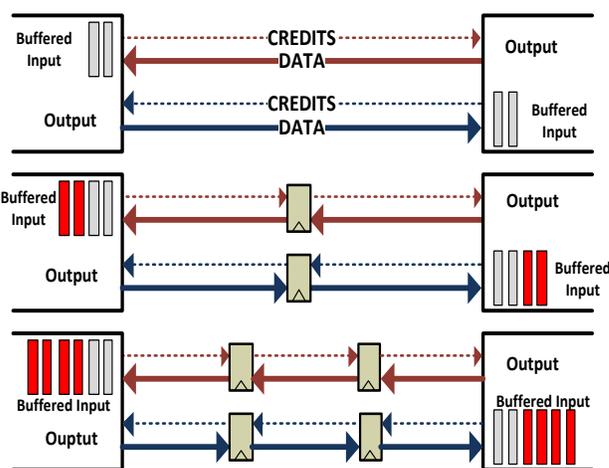

Figure 1. Extra buffering (red bars) due to pipelining of long links.

(class 2) is perceived as a low-cost mainstream approach to cope with excessive delay of combinatorial paths, its utilization in NoC links has two drawbacks. First, pipelining increases the absolute delay of long links since they are divided into N single cycle segments such that $N \cdot T_{Clock} \geq Delay_{Link}$. Second, pipelining incurs excessive buffering and area overheads as extra buffers are needed to compensate for the increased round-trip delay of the flow control mechanism (e.g. credit based, on/off, etc.) among adjacent router ports (Figure 1). Using FIFOs instead of pipelining registers, as proposed in [6], requires larger area compared to pipelining and results in a higher overall link delay. Techniques which reduce the link delay, if possible and cost-effective, are therefore preferable over pipelining. Wire sizing and repeater insertion (class 1) address the source of the problem by decreasing the absolute delay of long wires. In many architectures that comprise long wires [1]– [5] researchers assume that these techniques can be adopted to reduce the delay of long links. Usually, these assumptions are not backed up by a thorough evaluation of the respective costs and the design tradeoffs. In other cases, researchers propose the more complex solutions (i.e. classes 2 and 3) even though wire sizing and repeater insertion could suffice (e.g. [6]).
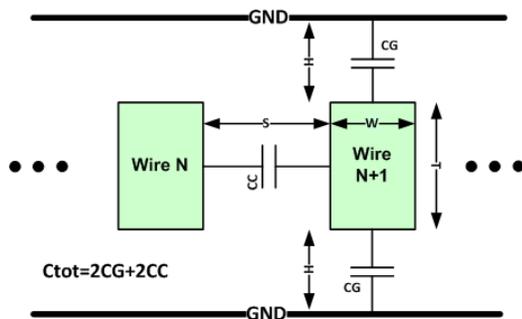
Figure 2. Interconnect physical model.

In this paper we evaluate the effectiveness and the system costs of tailoring long links to the timing constraints in hierarchical tiled NoCs using wire sizing and repeater insertion. Applying these techniques might incur considerable power, silicon area and wiring resources overheads. In the first part, we devise a unified cost function for power and wiring overheads of the adjusted long links. This function allows designers to evaluate the interconnect system costs associated with class 1 operations. We further present a technique to find the lowest-cost long wire configurations (defined by wire sizing parameters, density of repeaters and their size) which satisfy the delay requirements of the system. In the second part of the paper we use PyraMesh [1] hierarchical NoCs at different technology nodes to present the relation between target clock frequency and the adjustment cost of long links. We also provide estimations of the silicon area occupied by the repeaters in each case. Our research introduces a methodology to evaluate the feasibility and estimate the system costs of using parallel single cycle links in hierarchical NoCs. Note that although our work is focused on synchronous designs, asynchronous designs will benefit from fast, long links as well.

The paper is organized as follows: Section II summarizes the related work and outline the contributions of this paper; a universal system cost function for delay reduction modifications of parallel links is presented in Section III; system costs of long link delay reduction in typical hierarchical NoCs are calculated in Section IV; in Section V we discuss the results and conclude the paper.

II. RELATED WORK AND OUR CONTRIBUTIONS

Extensive research has been done on global interconnect optimization (e.g. [8]– [12]). Researchers address timing and power optimization of both individual interconnect trees [8,9,11], and parallel links [12,16]. Among the popular interconnect performance optimization methods are wire sizing [8], [9] repeaters insertion [10,11], and net-ordering [16]. In many works, the design space includes more than a single optimization method. Li et al. in [12], for instance, discuss the influence of both wire sizing and repeaters insertion on the latency, power and bandwidth of parallel links. Wire delay is usually described by Elmore's delay model [7].

Heterogeneous link length is an inherent property of hierarchical network topologies. Hierarchical NoCs are likely to include long parallel links of the order of the size of the die. RC delays of such links may exceed dozens of clock cycles if delay reduction techniques are not applied. Although timing constrains associated with long links have to be taken into account in every hierarchical NoC topology, many researchers tend to overlook this issue and neglect the costs of long links in their hardware costs evaluations [1]– [5]; others propose overdesigned solutions without a proper evaluation of simple alternatives [6]. Hierarchical topologies were proposed in [2] and [3], without addressing the issue of long links. In [1,4,5] it is mentioned that long links might suffer from excessive delays and assumed that the appropriate measures (e.g. wire sizing, repeaters insertion, usage of high metal layers and pipelining) are taken to reduce their delay. However, the costs and the effectiveness of these measures are not evaluated. In [6], long links are divided to short segments by 2-port FIFOs despite the fact that the overhead of such links is even higher than simple pipelining (Figure 1). The authors of [17,18] present difficulties in implementing high speed NoCs using standard CAD tools. In [17] the critical path stems from the router's logic, and in [18], although the authors focus on link delay in their synthesis, they only use repeater insertion for wire optimization. Special wire layouts such as on-chip transmission lines and special driver circuits can be used for achieving very fast links, e.g. [19,20,21].

Our paper bridges between the issues of global interconnect optimization and the design of hierarchical NoC architectures. We explore the effectiveness of wire sizing and repeaters insertion[1] to reduce delay of long links in hierarchical topologies at present and future technology nodes. Utilizing our methodologies and findings, researchers and designers can estimate the feasibility and the system costs of using long parallel links in hierarchical tiled NoCs. In a case of links which cannot achieve the required speed using those techniques, link pipelining might be needed. However, by using our techniques, a designer can reduce significantly the number of pipelined links. The number of pipe stages for those links, whose pipelining is inevitable, will be reduced as well.

III. DELAY REDUCTION OF GLOBAL WIRES – METHODS AND COSTS

Wires at the top metal layers are usually used for long links to ensure low RC delay. Wire sizing and repeater insertion are the most common methods to further reduce the delay of long links. In this section we devise a unified cost function for power and wiring costs of long links that were adjusted to correspond with system timing requirements. In addition, we introduce a methodology to obtain the lowest cost link configuration (i.e. wires and repeaters parameters) that satisfies the timing constrains.

---

[1] We do not use net-reordering since we assume that the sizes of all the drivers are similar.

A. Modeling Delay of Global Wires

We model parallel links as presented in Figure 2 and describe wire sizing with the following parameters:

- Λ_W – Scaling factor of wire's width (W) with respect to minimum size global wire [22].
- Λ_S – Scaling factor of spacing between wires (S) with respect to minimum size global wire [22].

We use Elmore's delay model [7] to calculate wire delay with no repeaters:

$$d_{wire} = 0.5 \tilde{R}_{int} \tilde{C}_{int} l^2 , \qquad (1)$$

where $\tilde{R}_{int}, \tilde{C}_{int}$ are the resistance and capacitance per unit length of the wire and l is its length. We use the wire capacitance model described by Wong et al. in [23].

Repeater insertion is described with the following parameters:

- ρ - Density of repeaters per millimeter assuming homogeneous distribution of identical repeaters along the wires.
- S_R – Repeaters' size normalized to Bakoglu's [10] delay-optimal scale factor h as defined in (3).

As shown in [11], optimal-sized repeaters have high power consumption. Their size can be reduced to significaly decrease the power dissipation, while keeping the delay close to optimal. We use Bakoglu's delay model of a repeated wire [10]:

$$d_{repated} =$$
$$= \rho l \left[0.7 \frac{R_0}{h S_R} \left(\frac{\tilde{C}_{int}}{\rho} + h S_R C_0 \right) + \frac{\tilde{R}_{int}}{\rho} \left(0.4 \frac{\tilde{C}_{int}}{\rho} + 0.7 h S_R C_0 \right) \right], \qquad (2)$$

where R_0 and C_0 are the output capacitance and the input resistance of a minimal-size inverter and h is Bakoglu's [10] delay-optimal scale factor for repeater size, relative to a minimal inverter, given by:

$$h = \sqrt{\frac{R_0 \tilde{C}_{int}}{\tilde{R}_{int} C_0}} . \qquad (3)$$

Throughout the paper, we label technology nodes with their characteristic MPU physical gate length, according to ITRS [22]. We show results for nodes starting from 29nm to 7nm (predicted for 2024). Figures 3.a-3.c present wire delay vs. length for different wire configurations at 29, 17 and 10 nm technology nodes, respectively. It is evident that as the technology advances, delay reduction of long wires becomes more challenging.

B. Unified Cost Function of Long Links Adjustments

Long links that were adapted to system timing requirements may be much more expensive than minimum size global wires (per unit length) in terms of power and wiring resources. To allow designers to compare different delay reduction configurations and evaluate the associated system costs we devise a unified long links adjustments cost function. Our unified cost function is a combination of wiring and power costs. We define wiring costs W_C as the
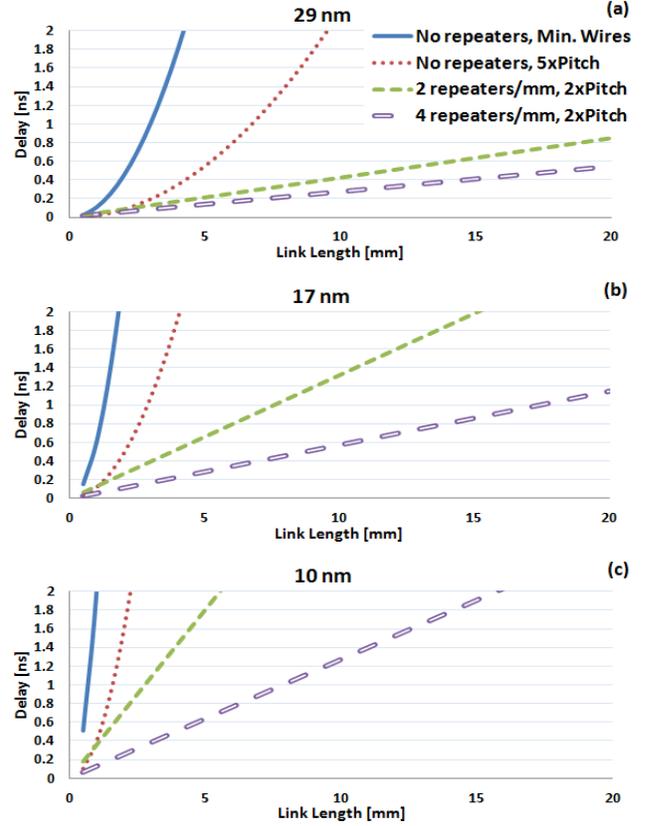

Figure 3. The four configurations presented are: I-Minimal size wires without repeaters, II-Wires with $\Lambda_W=\Lambda_S=5$ and no repeaters, III- $\Lambda_W=\Lambda_S=2$ with 2 repeaters/mm, IV- Wires with $\Lambda_W= \Lambda_S=2$ with 4 repeaters/mm. (a)-(c) present 29, 17 and 10 nm technology nodes respectively.

wire's pitch (i.e. $(S+W)/2$ in Figure 2) normalized to the pitch minimal global wire [22]. Power cost of an adjusted wire is defined as the ratio between its power and the power of minimum size wire. Power values are obtained using the following equation that describes power dissipation of a repeated wire:

$$P_{Repeated\ wire} = P_{wires} + P_{repeaters}^{dynamic} + P_{repeaters}^{static} =$$
$$= \alpha f \tilde{C}_{int} l V_{dd}^2 + \alpha f \rho l C_0 S_R h V_{dd}^2 + \rho l S_R h I_{leak} V_{dd} , \qquad (4)$$

where α is the switching probability factor (we use $\alpha = 0.125$, which matches a relatively high activity of 25%), f is the clock frequency, and I_{leak} is the repeaters leakage current that is obtained from ITRS [22]. Therefore, power cost (P_C) is defined as:

$$P_C = \frac{P_{wire}(Adjusted\ Wire)}{P_{wire}(Min.\ Global\ Wire, ITRS)} . \qquad (5)$$

We define the unified interconnect cost function as follows:

$$CF = W_C^\alpha P_C^\beta . \qquad (6)$$

CF reflects the per-unit-length power and wiring costs of modifying interconnect using wire sizing and repeater
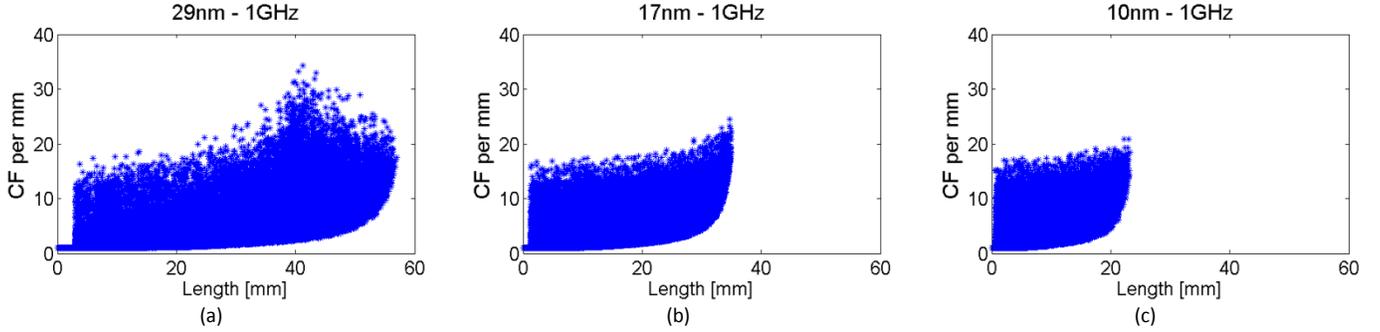
Figure 4. Results of Monte-Carlo analysis of interconnect cost, subject to CF (7) with α=β=0.5, for 29nm, 17nm and, 10nm technology nodes (a,b,c respectively) and a target clock frequency of 1GHz.

insertion. Suppose that pitch and capacitance of wire X is twice the minimum size. The relative cost of this wire is 2; accordingly, α and β satisfy:

$$\alpha + \beta = 1, \quad \alpha, \beta \geq 0, \qquad (7)$$

such that CF(X) = 2. Under the constraints in (7), designers can modify the values of α and β to tune the weights of wiring and power costs to best describe their NoC design goals. In this work we are equally concerned regarding power consumption and wiring area. Therefore we choose to set $\alpha = \beta = 0.5$. Note that CF does not account for the additional silicon area required by the repeaters since we did not find a convenient way to express this area as a per-unit-length cost normalized to the minimum size unrepeated wire. We count this area separately and provide the total area of the repeaters across the entire die for the use cases presented in Section IV.

C. Finding the Lowest Cost Links

We use the Monte Carlo method in order to find the lowest cost parallel links, subject to CF (6), that satisfy system timing requirements. Links design parameters and their ranges are summarized in Table I. For each Monte Carlo combination, we calculate the cost (CF) and use (2) and (3) to find the link length that matches a delay of a single clock cycle. Afterwards, (length, CF) pairs of each combination are plotted on a 2D plane (Figure 4). The lowest cost solutions, for a given target clock cycle and technology node, are found at the bottom edge of the full shape formed by all the Monte Carlo simulation results. We present these lowest-cost solutions curves for several technology nodes and clock cycles in Figure 5 and henceforth use them to quantify long links adjustment costs. It's evident (Figure 4) that for each wire length there are numerous working points with different costs. The curves in Figure 5 saturate at the

TABLE I. MONTE CARLO PARAMETERS

System Parameter	Architecture Parameters
Λ_W	[1..50]
Λ_S	[1..50]
ρ	[0..10]
S_R	[0..1]

maximum achievable length (for a given technology node and target clock cycle). Although this length degrades with technology nodes, length of several millimeters can still be achieved for reasonable clock frequencies in the most advanced nodes. Choosing different α and β values which satisfy (7) will produce slightly different CF values and the lowest cost configurations will be different. Nevertheless it won't change the general behavior of the cost function. Note that the saturation length is oblivious to the values of α and β exponents and the definition of CF itself. It's simply the single-cycle length of the fastest link that can be obtained with design parameters in the legal ranges (Table I).

IV. DESIGN OF LONG LINKS IN HIERARCHICAL TILED NOCS

Links at the high levels in hierarchical NoCs can reach lengths of several millimeters. In the previous section we described how long links can be modified to satisfy system timing requirements and introduced a unified cost function that quantifies the system interconnect costs of these adjustments. We observed that system costs of reducing the RC delay of long links increase with the target clock frequency and the progress in technology nodes. For a reasonable NoC clock frequency of 2GHz, a 5 mm long
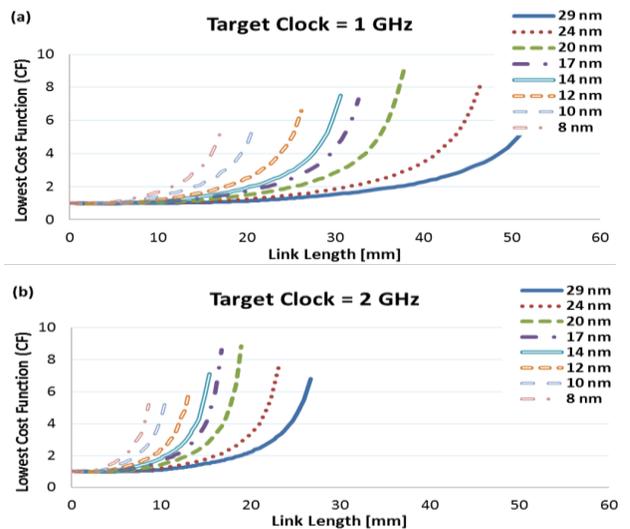

Figure 5. Minimal cost function for different technologies, (a)-for 1GHz and (b)-2GHz operating frequencies.
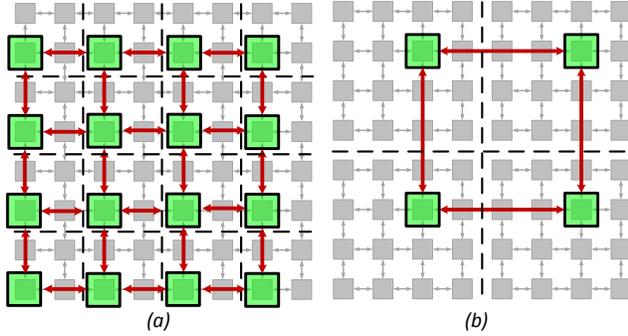
Figure 6. Hierarchical clustered 2D mesh. (a) – AHC with 2x2 clusters. (b) – Real clusters with 4x4 cluster size.
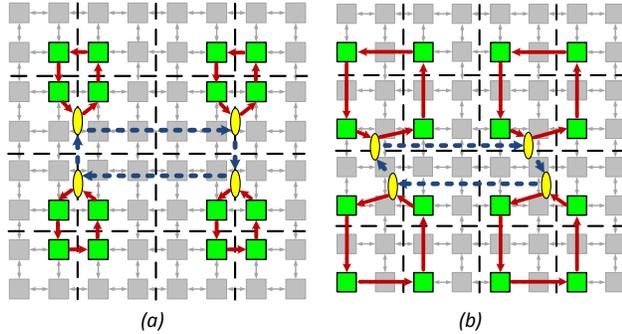

Figure 7. Ring of meshes. (a) – corner tile used to connect to the ring, (b) – center tile used. Red and blue arrows represent the lower and the upper ring respectively. The green squares represent mesh-lower ring connections and the yellow ellipses – the connection between lower and upper level rings.

link's cost might exceed 150% (i.e. 50% more than a minimal wire) per unit length in near future technology nodes. Fortunately, in many cases, only a few top level links have to be adjusted.

In this section we quantify the overall overheads of the tuning of long links in hierarchical NoCs to system timing requirements (i.e. the ratio between total costs of all NoC links before and after timing adjustments and the additional area required for repeaters). We show that in many typical present and future hierarchical tiled NoCs, most of the links are lowest-level links which are short enough to meet timing constraints without any modifications. Moreover, we show that long links are a minority, and therefore the overall overhead of their speed-up adjustments, is low in many typical cases. We evaluate several hierarchical network topologies described below.

A. Hierarchical Clustered 2D Mesh

Hierarchical clustered 2D Mesh is presented in [5]. It adds an additional interconnection level to a simple 2D mesh. The bottom level can be left untouched to create, as described in [5], additional highway connections (AHC), or divided to create real clusters as presented in Figure 6. The AHC, in fact, is a particular case of PyraMesh architecture, described below, with *NP=1, NL=2, C=1* and α is derived from mesh and cluster sizes. In the real cluster scheme, each cluster is fully connected using a bottom level 2D mesh. The

clusters themselves are connected using only an upper level 2D mesh. All the intra cluster links on the bottom level don't exist. In our work we use AHC, since it achieves better performance.

B. Ring of Meshes

The authors of [3] present an architecture which is similar to the hierarchical mesh with 16 clusters. However, instead of using 2D mesh for inter-cluster connections, they use two level hierarchical unidirectional rings as presented in Figure 7. The lower level contains four rings (red solid arrows) with four connections (highlighted square) to the 2D mesh and one connection (yellow ellipses) to the top level ring. The top level ring (blue dashed arrows) has four connections. Two locations for the mesh-ring connection are suggested: the center of the cluster or the corner.

C. PyraMesh

The PyraMesh family of pyramid-like hierarchical 2D mesh topologies is presented in [1]. PyraMesh NoCs are described by the following parameters:

- *K* - Size of the baseline mesh (i.e. *K* describes *KxK* mesh).
- *NL* - Number of levels, including the base mesh.
- *NP* - Number of pyramid structures on-top the baseline mesh.
- α_i - Ratio between sizes of levels *i* and *i+1*.
- C_i - Concentration of level *i*, i.e. how many routers in level *i* are connected to a router in level *i+1* along a single dimension.

Examples of two PyraMeshes with K = 8 (i.e. 8x8 baseline mesh) are presented in Figure 8. Hierarchical topologies were found in [1] to be more cost-effective than planar 2D mesh starting from systems size of 16x16. Ref. [1] also devised latency and throughput optimized PyraMesh configurations for 16x16–128x128 systems (Tables II and III) assuming single clock cycle delay for all the links in the scheme.

TABLE II. THROUGHPUT OPTIMIZED PYRAMESHES [1]

System Size	Architecture Parameters
16x16	NP = 1, NL=3, α_i=[2,4], C_i=[1,1]
32x32	NP=1, NL=5, α_i=[2,2,2,4], C_i=[1,1,1,1]
64x64	NP=4, NL=2, α_i=4, C_i=1
128x128	NP=4, NL=3, α_i=[4,4], C_i=[1,1]

TABLE III. LATENCY OPTIMIZED PYRAMESHES [1]

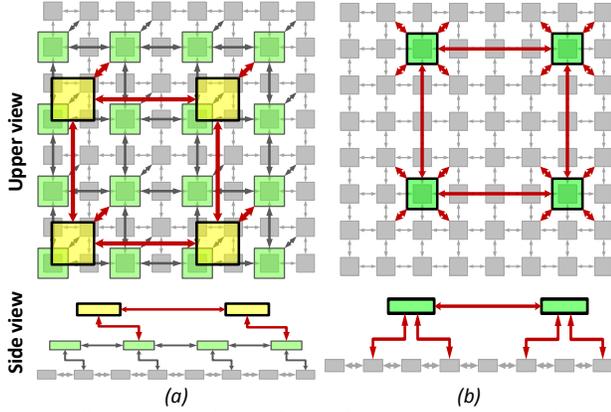| System Size	Architecture Parameters
16x16	NP=1, NL=3, α_i=[4,4], C_i=[2,4]
32x32	NP=1, NL=4, α_i=[4,4,2], C_i=[2,4,2]
64x64	NP=1, NL=4, α_i=[8,4,2], C_i=[4,4,2]
128x128	NP=1, NL=3, α_i=[8,4], C_i=[4,4]

Figure 8. Two Examples of PyraMesh, upper view (top) and side view (bottom). (a) – 3-levels PyraMesh with (K = 8, NP = 1, NL = 3, α = 2, C = 1). (b) – 2-levels PyraMesh with (K = 8, NP = 1, NL = 2, α = 4, C = 2). The upper view figures are taken from [1].

D. System Costs of Adjusting Long Links

We claim that system costs of adapting long links to the target clock frequency are low in many typical scenarios, since most of the links in the system are short enough to satisfy timing constraints with no delay reduction applied. To validate this assumption, we generated links delay histograms of the systems described above at each technology node between 29nm and 14nm. Selected histograms are presented in Figure 9. We found, for instance, that only 11.3% and 16.1% out of all links, on average, have to be modified for targets of 1GHz and 2GHz, respectively, in 32x32 PyraMesh systems across all technology nodes.

For the same frequencies, the number of links that need adjustment in hierarchical clustered 2D 32x32 mesh are 5.6% and 9.2%. In the ring of meshes architecture, this number goes as low as 0.7% for both frequencies due to the small number of high level links. We observe that in all the reviewed hierarchical architectures, the majority of the links meet the clock constraints, and therefore do not need to be adjusted. For technology nodes lower than 14 mm, the lowest 2D mesh level need to be adjusted as well. However, as we will show further, those adjustments and, therefore the cost are minor. Hereafter we focus on the PyraMesh family, since it has the most links needed adjustment.

We use the cost function CF from (6) to calculate the overall interconnect costs of adjustments of long links. We define **N**oC **I**nterconnect **O**verhead (*NIO*) as follows:

$$NIO = \frac{\sum_{All\ NoC\ Links} CF(\text{adjusted link}) \cdot Length(\text{link})}{\sum_{All\ NoC\ Links} Length(\text{link})}. \quad (8)$$

NIO is defined as the ratio between the overall links cost after and before the delay reduction adjustments. We have calculated NIO for the systems from Tables II and III at different technology nodes and target clock cycles. In Figure 10, we present NIO values (through a tones-map) over the technology nodes – clock cycle 2D plane. White indicates
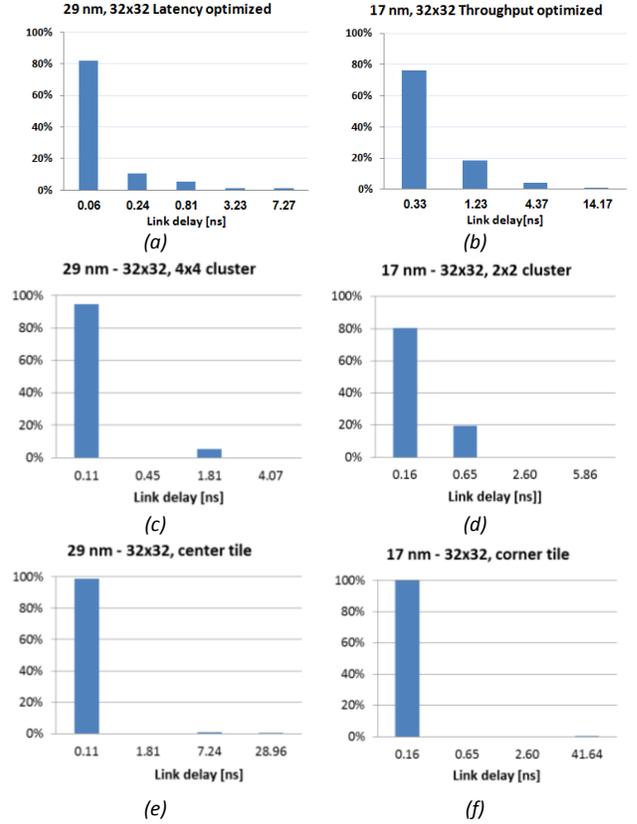

Figure 9. Six examples of minimum size global links delay histograms: (a) – 32x32 delay optimized PyraMesh at 29nm. (b) – 32x32 throughput optimized PyraMesh at 17nm. (c) – 32x32 hierarchical NoC with 4x4 clusters at 29nm. (d) – 32x32 hierarchical NoC with 2x2 clusters at 17nm. (e) – 32x32 Ring of meshes with a center located ring interconnect. (f) – 32x32 Ring of mesh with a corner located ring interconnect. Each bin indicates the ratio in % out of all NoC links.

that NIO is greater than 2 which reflects that either the required NoC interconnect overhead is higher than 100% or the target clock can't be reached using wire sizing and repeater insertion. The results indicate that it is easier to adapt throughput optimized than latency optimized PyraMeshes to high frequencies. The reason is that latency optimized PyraMeshes include much longer wires at the highest hierarchy levels. We used both throughput and latency optimized systems in order to study a variety of typical hierarchical NoCs.

In addition to NIO, we have estimated the fraction of the die's silicon area required by the repeaters of the adjusted wires. We approximated the area of a minimum size transistor (A_{MT}) as 1/6 of the net area of 6T SRAM cell provided by ITRS [22]. A single repeater is composed of two transistors (NMOS and PMOS); therefore, assuming A_{MT} is the average area of NMOS and PMOS minimum size transistors, the fraction of the die's area required by the repeaters is given by:

Latency Optimized Pyrameshes — (a), (c), (e), (g)
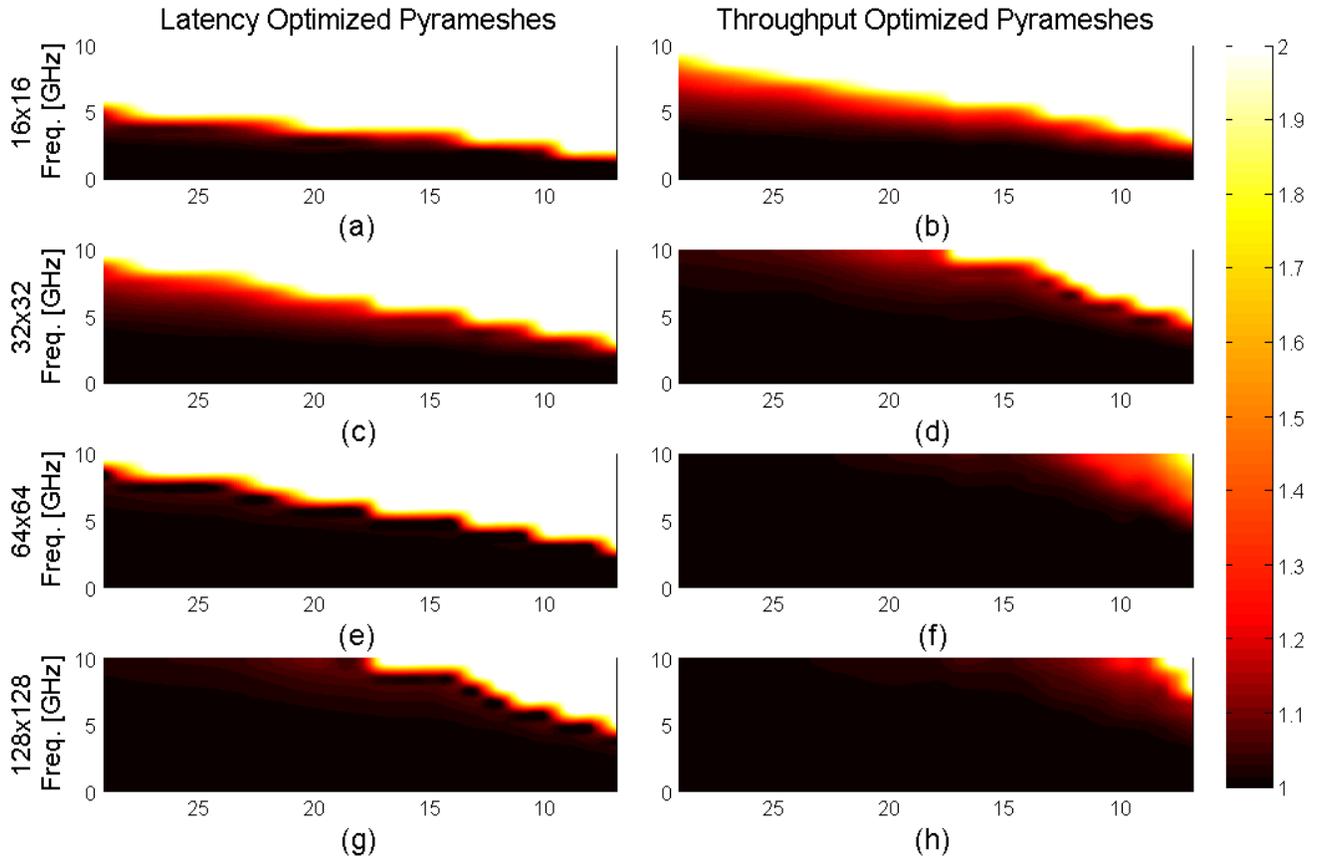**Throughput Optimized Pyrameshes** — (b), (d), (f), (h)

Figure 10. NIO for different target clocks at different technologies. Latency optimized and throughput optimized PyraMeshes are presented in figures (a,c,e,g) and (b,d,f,h) respectively; Baseline mesh sizes: 16x16 (figures a and b), 32x32 (c, d), 64x64 (e, f) and 128x128 (g, h).

$$RArea_{Ratio} =$$

$$= \frac{\sum\limits_{\text{All NoC Repeaters}} 2A_{MT} \cdot h(repeater) \cdot SR(repeater)}{die\ area} \quad (9)$$

In each technology node, we calculated the average $RArea_{Ratio}$ of the systems from Tables II and III for clock frequencies of 1GHz - 3GHz assuming 64 bits wide links. The results imply that the area required for repeaters within a whole system is negligible, as presented in Figure 11.

V. DISCUSSION AND CONCLUSIONS

Clock frequencies in present tiled commercial CMP NoCs rarely exceed 1GHz, even though CMPs are implemented in a wide spectrum of technology nodes (e.g. Tilera's CMP – 1GHz @ 90nm [24], Adapteva's CMP – 1GHz @ 28nm [25]). Assuming consistency in future technology nodes, we expect clock frequency in tiled NoCs not to exceed a few GHz.

The results in Figures 10 and 11 indicate that using long parallel links is feasible in a wide span of present and future hierarchical NoCs. 1GHz is reachable using wire sizing and repeater insertion (with the constraints of Table I) in all the systems and nodes that we have evaluated. 2GHz and 3GHz can also be reached in almost all the systems, except for few configurations with very advanced technology nodes. In systems where target clock can't be reached, techniques such as pipelining should be used in very few long top-level links. The average NoC interconnect overheads (i.e. power and wiring, based on CF) of adjusting the system to frequencies of 1GHz, 2GHz and 3GHz are 0.2%, 1.8% and 4.5% respectively across all the system configurations and technology nodes that we have evaluated (except where 2GHz and 3GHz can't be achieved). The average area required by repeaters in systems adapted to 3GHz is only 0.09% of the die. Higher frequencies can also be achieved with very low overheads mainly in large NoCs (i.e. ≥32x32 nodes).
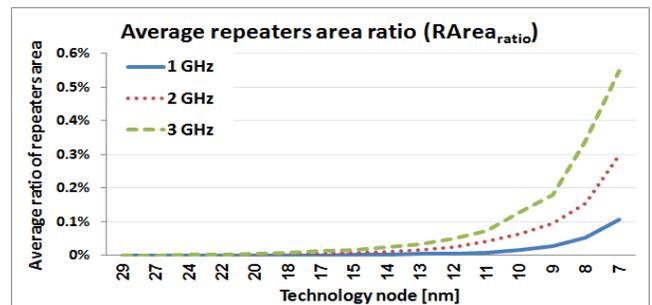

Figure 11. Average $RArea_{Ratio}$ for 1, 2 and 3 GHz target clocks, for various technology nodes.

In conclusion, our evaluation shows that parallel single-cycle links are feasible at a reasonable cost in present and future hierarchical on-chip NoCs. Our work provides a methodology to evaluate the feasibility and the system costs for using long parallel links in hierarchical NoCs and to evaluate the hardware cost of present and future hierarchical Networks-on-Chip.

REFERENCES

[1] R. Manevich, I. Cidon, and A. Kolodny, "Handling global traffic in future CMP NoCs," in *proceeding of international workshop on system level interconnect prediction (SLIP)*, 2012, pp. 40-47.

[2] C. Puttmann, J.-C. Niemann, M. Porrmann, and U. Ruckert, "GigaNoC - a hierarchical network-on-chip for scalable chip-multiprocessors," in *proceedings of the 10th Euromicro conference on digital system design (DSD)*, 2007.

[3] S. Bourduas and Z Zilic, "A hybrid ring/mesh interconnect for network-on-chip using hierarchical rings for global routing," in *Proceeding of the first international symposium on networks-on chip (NOCS)*, 2007, pp. 195-204.

[4] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly topology for on-chip networks," in *proceeding of the 40th annual international symposium on microarchitecture*, 2007, pp. 172-182.

[5] M. Winter, S. Prusseit, and P.F. Gerhard, "Hierarchical routing architectures in clustered 2D-mesh Networks-on-Chip," in *proceedings of international SoC design conference (ISOCC)*, 2010, pp. 388-391.

[6] U.Y. Ogras and R. Marculescu, ""It's a small world after all": NoC performance optimization via long-range link insertion," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 7, pp. 693-706, 2006.

[7] W.C. Elmore, "The transient response of damped linear network with particular regard to wide-band amplifiers," *Applied Physics*, vol. 19, no. 1, pp. 55-63, 1948.

[8] S.S. Sapatnekar, "RC interconnect optimization under the Elmore delay model," in *proceedings of the 31st annual design automation conference (DAC)*, 1994.

[9] J Cong and K-S. Leung, "Optimal wire sizing under Elmore delay model," *IEEE Transactions on Computer-Aided Design of Inegrated Circuits System*, vol. 14, no. 3, pp. 321-336, 1995.

[10] H.B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*.: Addison-Wesley, 1990.

[11] K Banerjee and A Mehrotra, "A power optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Transactions on Electron Devices*, vol. 49, no. 11, pp. 2001-2007, 2002.

[12] X.C. Li, J.F. Mao, H-F Huang, and Y Liu, "Global interconnect width and spacing optimization for latency, bandwidth and power dissipation," *IEEE Transactions on Electron Deivces*, vol. 52, no. 10, pp. 2272-2279, 2005.

[13] M.F. Chang et al., "CMP network-on-chip overlaid with multi-band RF-interconnect," in *Proceeding of the international symposium on high performance computer architecture (HPCA)*, 2008, pp. 191-202.

[14] A Shacham, K Bergman, and L. P. Carloni, "Photonic networks-on-chip for future generations of chip multiprocessors," *IEEE Transactions on Computers*, vol. 57, no. 9, pp. 1246-1260, 2008.

[15] R. Dobkin, Y. Perelman, T. Liran, R. Ginosar, and A. Kolodny, "High rate wave-pipelined asynchronous on-chip bit-serial data link," in *proceeding of the international symposium on asynchronous circuits and systems (ASYNC)*, 2007, pp. 3-14.

[16] K. Moiseev, S. Wimer, and A. Kolodny, "Timing optimization of interconnect by simultaneous net-ordering, wire sizing and spacing," in *proceeding of the international symposium on circuits and systems (ISCAS)*, 2006, pp. 329-332.

[17] A. Pullini et al., "Bringing NoCs to 65 nm," *Micro IEEE*, vol. 27, no. 5, pp. 75-85, 2007.

[18] M. Ferraresi, G. Gobbo, D. Ludovici, and D. Bertozzi, "Bringing Network-on-Chip links to 45nm," in *proceeding of the international symposium on system-on-chip (SOC)*, 2011, pp. 122-127.

[19] D. Goren et al., "An interconnect-aware methodology for analog and mixed signal design, based on high bandwidth (over 40 GHz) on-chip transmission line approach," in *proceeding of design, automation and test in Europe conference and exhibition (DATE)*, 2002, pp. 804-811.

[20] A. Deutsch et al., "On-chip wiring design challenges for gigahertz," *proceedings of the IEEE*, vol. 89, no. 4, pp. 529-555, 2001.

[21] R. Ho, K. Mai, and M. Horowitz, "Efficient on-chip global interconnects," in *proceeding of the symposium on VLSI circuits*, 2003, pp. 271-274.

[22] (2012) The international technology roadmap for semiconductors (ITRS). [Online]. http://www.itrs.net/

[23] S. Wong, G. Lee, and D. Ma, "Modeling of interconnect capacitance, delay, and crosstalk in VLSI," *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 1, 2000.

[24] D. Wentzlaff et al., "On-chip interconnection architecture of the tile processor," *IEEE Micro*, vol. 27, no. 5, pp. 15-31, 2007.

[25] A. Olofsson, "A 1024-core 70 GFLOP/W floating point manycore microprocessor," in *Poster on high performance embedded computing (HPEC)*, 2011.