Designing Single-Cycle Long Links in Hierarchical NoCs

Ran Manevich[1] Leon Polishuk[1] Israel Cidon[2] Avinoam Kolodny[2]
Electrical Engineering Department, Technion – Israel Institute of Technology, Haifa, Israel.
{ranman, leonpol}@tx.technion.ac.il[1], {cidon, kolodny}@ee.technion.ac.il[2]

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 typical cases, taking into account that there are only a few long links in most proposed hierarchical NoC architectures.[1]

1. 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, as 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 repeater insertion [8, 9, 10, 11, 12], buffering and pipelining [6, 13], and utilization of techniques such as RF [14], photonic [15] or wave-pipelined fast serial links [16]. Solutions of the third class require radical changes in technology. In pipelining (class 2) [13], long links are split into single-cycle segments. Pipelining requires extra buffering resources, and increases the latency.

Wire sizing and repeater insertion (class 1) address the source of the problem by decreasing the absolute delay of long wires. In many architectures which employ long wires [1, 2, 3, 4, 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 design tradeoffs. In other cases,

[1] *A preliminary short version of this work entitled "Design Tradeoffs of Long Links in Hierarchical Tiled Networks-on-Chop (NoCs)" was presented in the 16th Euromicro Conference on Digital System Design (DSD), 2013.*

researchers propose the more complex solutions (i.e. classes 2 and 3) even though wire sizing and repeater insertion could suffice (e.g. [6]).

In this paper we evaluate the effectiveness and the system costs of tailoring long links to the timing constraints in hierarchical NoCs using wire sizing and repeater insertion. Applying these techniques might incur considerable power, silicon area and wiring resources overhead. In the first part, we present our methodology and 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] and hybrid ring/mesh [3] hierarchical NoCs at different technology nodes to explore 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.

Clock frequencies in present commercial CMP NoCs rarely exceed 1 GHz, even though CMPs are implemented in a wide spectrum of technology nodes (e.g. Tilera's CMP – 1 GHz @ 90nm [17], Adapteva's CMP – 1 GHz @ 28nm [18], Kalray's CMP – 400MHz @ 28nm [19]). Assuming consistency in future technology nodes, we expect clock frequency in NoCs not to exceed a few GHz and therefore concentrate on those frequencies.

The paper is organized as follows: section 2 summarizes the related work and outlines the contributions of this paper; architecture and traffic modeling of hierarchical NoCs are briefly discussed in section 3 for further use. An analysis of distribution of link lengths in hierarchical NoCs is presented in section 4. Section 5 introduces a universal cost function for delay reduction modifications of parallel links and introduces a methodology to find lowest cost links subject to timing and design constraints. Overall system costs of adjusting long wires to practical target clock frequency in hierarchical NoCs are analyzed in section 6. Section 7 summarizes the results and concludes the paper.

2. RELATED WORK AND OUR CONTRIBUTIONS

Extensive research has been done on global interconnect optimization (e.g. [8, 9, 10, 11, 12, 20]). Researchers address timing and power optimization of both individual interconnect trees [8, 9, 11], and parallel links [12, 20]. Among the popular interconnect performance optimization methods are wire sizing [8, 9] repeater insertion [10, 11], and net-ordering [20]. 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 repeater insertion on the latency, power and bandwidth of parallel links. Wire delay is usually described by Elmore's 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. 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, 2, 3, 4, 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. The authors of [21, 22] present difficulties in implementing high speed NoCs using standard CAD tools. In [21] the critical path stems from the router's logic, and in [22], 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. [23, 24, 25].

Our paper bridges between the issues of global interconnect optimization and the design of hierarchical NoC architectures. We analyze the distribution of lengths of links in hierarchical NoCs and conclude that links in hierarchical NoCs should not exceed lengths of a few millimeters. We explore the effectiveness of wire sizing and repeater insertion to reduce delay of long links in hierarchical topologies in present and future technology nodes. Utilizing our methodology, researchers and designers can estimate the feasibility and the system costs of using long parallel links in hierarchical 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 significantly reduce the number of pipelined links. The number of pipe stages for those links, whose pipelining is inevitable, will be reduced as well.

3. HIERARCHICAL NOCS – ARCHITECTURE AND TRAFFIC MODELING

Hierarchical topologies are usually comprised of a bottom hierarchy level that includes the network interfaces to the processing elements, and one or more upper levels. The upper hierarchy levels are most often more sparse than the bottom level as they span longer physical distances per hop and provide shortcuts in terms of hop-distance for global packets.

3.1 Baseline Hierarchical Architectures

We base our analysis on two hierarchical architectures: PyraMesh [1] and hybrid ring/mesh [3]. PyraMesh is a family of pyramid-like hierarchical 2D mesh topologies. The PyraMesh is a low-cost upgrade of the well known 2D-Mesh topology where smaller mesh networks are added on the top of the baseline mesh in a single or several pyramid-like structures. While the original mesh provides full connectivity, the upper levels of the pyramid(s) are used for reducing the hop distance traversed by global packets. PyraMesh preserves packet structure, flow control and routing mechanisms across all its levels. 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 1. We define the wiring overhead as the ratio between the accumulated length of global links of PyraMesh and the baseline 2D mesh NOC. For simplicity, we assume that the routers in each hierarchy level are distributed homogeneously across the die and model lengths of links as illustrated in Figure 2. Various sets of

PyraMesh design parameters describe completely different architectures and optimize different design goals [1].
The hybrid ring/mesh topology resembles PyraMesh in many aspects but offers less flexibility in design and routing [1, 26]. We use PyraMesh and hybrid ring/mesh as the baseline architectures for analysis of lengths of long links and their percentage among all the links in the network. These architectures represent both clustered (e.g. [2]) and non-clustered (e.g. [5]) approaches for hierarchical NoC design. Topologies such as Flattened-Butterfly [4] and fat-tree [27] utilize many more long links but are not scalable and are not feasible as module counts exceed a few hundreds [1], hence they are seldom considered for NoC applications.

3.2 Modeling NoC Traffic Locality with Rent's Rule

The maximum hop-distance in flat 2D topologies is proportional to the square root of the number of routers. Hierarchical topologies provide much better maximum hop-distance scalability due to their tree-like structure (it scales as log[√(number of routers)]). Usually, the average hop-distance metric is more important than the maximum. Traffic patterns with distinct locality (i.e. where most packets are exchanged between neighboring nodes) are desirable in large NoCs since they are likely to yield lower latency and power, better system performance, and higher scalability. The degree of traffic locality has a direct effect on average hop-distance. Hence, modeling traffic locality might be very helpful for design space exploration of large NoCs. The bandwidth version of Rent's rule [28] relates the communication bandwidth (B) between a cluster of modules and the rest of the system with the number of modules in the cluster (G) (Eq. 1, k – average bandwidth of a single module, R – Rent's exponent):

$$B = kG^R \tag{1}$$

Heirman et al. [29] showed experimentally that traffic patterns of popular CMP benchmarks follow the bandwidth version of Rent's rule with Rent's exponent R of ~0.7 on average. Moreover, [29] showed that R varies among different phases of the application over time. Based on these results, we use synthetic traffic patterns that follow the bandwidth version of Rent's rule (a.k.a Rentian traffic) to model traffic in CMP NoCs and analyze the design tradeoffs of long links in hierarchical NoCs, using the average hop-distance metric as one of the considerations. Synthetic patterns are formed similarly to [1].

4. LENGTH OF LONG LINKS IN HIERARCHICAL NOCS

In [1], we introduced a methodology to obtain hierarchical NoCs that optimize average hop-distance, having an average degree of traffic locality (described by Rent's exponent R) and subject to hardware and wiring cost constraints. We used a similar methodology to obtain NoCs with optimized hop-distances for different system sizes, and calculated their maximum and average hop distance assuming the average traffic locality of CMP benchmarks that was observed in [29] (i.e. Rent's exponent R = 0.7). We limited the radix of routers in the network to 9x9 (i.e PyraMesh design parameter $C_i \leq 2$) to ensure that the in-router switch does not become a timing bottleneck. Design parameters and the radix of each of the levels for 8x8-128x128 systems, optimized for average-hop-distance, are presented in Table I. The length of the longest link is also indicated in Table I assuming a ~300mm^2 die (i.e. 17x17 mm). Average and maximum hop distance vs. number of modules of a flat system and the hierarchical systems from Table I are presented in Figures 3.a and 3.b. Figures 3.a

and 3.b illustrate how hierarchical topologies can improve the hop-distance scalability of NoCs. Moreover, these figures show that hierarchical NoCs provide significant benefit in both hop-distance metrics, starting from systems of a few hundreds of modules (i.e. between 16x16 and 32x32).
The radix of the highest hierarchy level, of the hop-distance hierarchical optimized systems in Table I, is 2x2. According to our link-length model (Figure 2.a), 2x2 mesh utilizes quite long links (5.67mm @ 17x17 mm die) that would probably be a timing bottleneck in many systems. The benefit of this 2x2 level in terms of hop-distance is likely to be very limited, particularly in large systems. We performed hop-distance optimization again, but now limiting the smallest radix of the upper level to 4x4, 8x8 , or 16x16. The architecture parameters of these systems are presented in Table II. Figures 4.a and 4.b provide the maximum and average hop-distance of these hierarchical NoCs for 8x8-128x128 systems, compared to flat 2D mesh and the hop-distance optimized hierarchical NoCs with unconstrained radix of the upper level. The figures show that there is practically no difference between hierarchical NoCs with highest levels of 2x2 and 4x4. Moreover, restricting the highest hierarchy level to radix of 8x8, or even 16x16 might be enough in large systems (e.g. 32x32, 64x64 and 128x128). Our discussion reveals that long links in hierarchical NoCs do not have to be very long and can be limited to a few millimeters. Consequently, we use the hierarchical hop-distance optimized systems with upper levels of 4x4 (Table II) as representative use-cases for our analysis.

5. 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.

5.1 Modeling Delay of Global Wires

We model parallel links as presented in Figure 5 and describe wire sizing with the following parameters:

Λ_W – Scaling factor of wire's width (W) with respect to minimum size global wire [30].
Λ_S – Scaling factor of spacing between wires (S) with respect to minimum size global wire [30].

Wire delays are modeled using Bakoglu's delay model of a repeated wire [10], based on Elmore's [7] distributed RC wire delay for each of the wire segments. This is a closed-form model which had been validated extensively in the literature [31]. It was shown to have high-fidelity as an objective function for interconnect optimization [32], and its parameters can be scaled or fitted to obtain absolute accuracy in the range of 2-10% [33]. Therefore, it is most appropriate for an analytical study comparing design strategies over several technology generations.

$$d_{repated} = \rho l \left[0.7 \frac{R_0}{hS_R} \left(\frac{\tilde{C}_{\text{int}}}{\rho} + hS_R C_0 \right) + \frac{\tilde{R}_{\text{int}}}{\rho} \left(0.4 \frac{\tilde{C}_{\text{int}}}{\rho} + 0.7 hS_R C_0 \right) \right] \quad (2)$$

$\tilde{R}_{int}, \tilde{C}_{int}$ are the resistance and capacitance per unit length of the wire and l is its total length. R_0 and C_0 are the output capacitance and the input resistance of a minimal-

size inverter and h is Bakoglu's delay-optimal scale factor for repeater size, relative to a minimal inverter, given by:

$$h = \sqrt{\frac{R_0 \tilde{C}_{\text{int}}}{\tilde{R}_{\text{int}} C_0}} \tag{3}$$

As shown in [11], optimal-sized repeaters have high power consumption. Their size can be reduced to significantly decrease the power dissipation, while keeping the delay close to optimal.

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 delay-optimal scale factor h as defined in (3).

Delay of an un-repeated wire is calculated using (2) assuming identical drivers at both wire ends:

$$d_{un-repated} = 0.7\frac{R_0}{hS_R}\left(l\tilde{C}_{\text{int}} + hS_R C_0\right) + l\tilde{R}_{\text{int}}\left(0.4\tilde{C}_{\text{int}} + 0.7hS_R C_0\right) \tag{4}$$

We use RC based delay modeling since the inequality in (5) is valid for all the wires and repeated wire segments that we analyze in this paper and therefore wire inductance effects are not significant [34].

$$l_{\text{Wire or Repeated Wire Segment}} > \frac{2}{\tilde{R}_{\text{int}}}\sqrt{\frac{\tilde{C}_{\text{int}}}{\tilde{L}_{\text{int}}}} \tag{5}$$

We use the wire capacitance model described by Wong et al. in [35] and take into account the worst case delay capacitance, as presented in Figure 5. Even though [35] has been published over a decade ago, the total wire capacitance of minimum dimension global wires obtained using [35] falls right between the lower and higher capacitance values in IRTS [30] for all the technology nodes. Consequently, we believe that our wire-delay modeling is useful for architectural planning in advanced technology nodes.

Throughout the paper, we label technology nodes with their characteristic MPU physical gate length, according to ITRS [30]. We show results for nodes starting from 29nm to 8nm (predicted for 2023). Dimensions and attributes of minimum global wires that were obtained from ITRS [30] and are summarized in Table III. Figures 6.a-6.b present wire delay vs. length for different wire configurations at 29nm and 10nm technology nodes, respectively. It is evident that as the technology advances, delay reduction of long wires becomes more challenging. We utilized stimulated annealing to find the lowest achievable link delay using wire sizing and repeater insertion vs. link length (Figure 7.b). We restricted sizing and repeaters design parameters to practical ranges (Table IV). Delays of minimum size global wires vs. length are presented in Figure 7.a. Figures 7.a and 7.b illustrate how wire sizing and repeaters insertion can be useful to adapt the delay of long links to practical target clock frequencies at present and future technology nodes.

5.2 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 wire's pitch (i.e. S+W in Figure 5) normalized to the pitch of a minimal global wire. 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:

$$P_{Repeated\ wire} = P_{wires} + P_{repeaters}^{dynamic} + P_{repeaters}^{static} = \alpha f \tilde{C}_{int} l V_{dd}^2 + \alpha f \rho l C_0 S_R h V_{dd}^2 + \rho l S_R h I_{leak} V_{dd} \quad (6)$$

where α is the switching probability factor (we use α = 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. Therefore, power cost (P_C) is defined as:

$$P_C = \frac{P_{wire}\left(Adjusted\ Wire\right)}{P_{wire}\left(Min.\ Global\ Wire, ITRS\right)} \quad (7)$$

We define the unified interconnect cost function as follows:

$$CF = W_C^{\alpha} P_C^{\beta} \quad (8)$$

CF reflects the per-unit-length power and wiring costs of modifying interconnect using wire sizing and repeater 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 ß satisfy:

$$\alpha + \beta = 1, \quad \alpha, \beta \geq 0 \quad (9)$$

such that CF(X) = 2. Under the constraints in (9), designers can modify the values of α and ß 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 α=ß=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 6.

5.3 Finding the Lowest Cost Links

We use the Monte Carlo method in order to find the lowest cost parallel links, subject to the cost function CF (8), that satisfy system timing requirements. Links design parameters and their ranges are summarized in Table IV. 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 8).

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 (i.e the Pareto curve). We present these lowest-cost solutions curves for several technology nodes and clock cycles in Figure 9 and henceforth use them to quantify long links adjustment costs. It's evident (Figure 8) that for each wire length

there are numerous working points with different costs. The Parto Curves in Figures 8, 9 saturate at the 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 ß values which satisfy (9) 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 ß exponents and the definition of CF itself. It is simply the single-cycle length of the fastest link that can be obtained with design parameters in the legal ranges (Table IV).
We devised an on-chip parallel links calculator [36] that implements our methodology to find the lowest-cost parallel links for 29nm-7nm technology nodes subject to various performance and design constraints (Figure 10). The calculator is available online.

5.4 Pipelined Links

Link pipelining stands for splitting long links into single-cycle short segments with flip-flops [13]. Although pipelining 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 $N \cdot T_{Clock} > Delay_{Link}$. 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 11). Using FIFOs instead of pipelining registers, as proposed in [6], requires larger area compared to pipelining and results in a higher overall link delay.
From the latency perspective, delay reduction of long links using wire sizing and repeater insertion is clearly preferable over pipelining. While pipelining increases the absolute delay of multi-cycle links, wire-sizing and repeaters insertion can reduce the delay to less than a single cycle. For instance, we present the effect of pipelining on the longest path in a 16x16 PyraMesh in Figure 12 (assuming router delay of four clock cycles). Moreover, the addition of flip-flops in pipelining increases the power dissipation of the system, as each of them requires continuous clocking. The quantitative comparison between the power overheads of pipelining and straight-forward delay reduction (i.e. wire sizing and repeater insertion) approaches is out of the scope of this paper. However, we show that the straight-forward delay reduction has very limited power overheads [2] in many practical cases and believe that pipelining should be used only when other link-improvement techniques, which reduce the wire delay, are not cost-effective or impossible.

6. DESIGN OF LONG LINKS IN HIERARCHICAL 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.

[2] Wire sizing can even reduce power consumption if the overall capacitance per unit length is decreased due to extension of the spacing between the wires (S in Figure 5).

In this section we quantify the overall overhead of tuning 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 typical hierarchical NoCs in present and future technology nodes, most of the links are lowest-level links which are short enough to meet timing constraints with insignificant 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. Overall cost of adjusting long links in hierarchical NoCs is not dependent on the particular topology, but on the distribution of lengths of links. Following our discussion in section 4, we understand that distribution of links length should not vary among different hierarchical topologies. In this section, we use the 16x16 PyraMesh from Table II (Figure 13.a) and equivalent 16x16 hybrid ring/mesh [3] (Figure 13.b) as representative hierarchical NoCs for our overall costs analysis. We model lengths of links for both the topologies as presented in Figure 2. We assume that the lengths of links in the hybrid ring/mesh topology are similar to the equivalent levels in PyraMesh. The lengths of the links at the top ring are assumed to be as those of the level beneath since it can be placed in the middle of the die as illustrated in Figure 13.b.

We predict 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 minimum or even no delay reduction applied. We have generated link length histograms of the systems described above (Figure 13). The die size is assumed to be 17x17mm (i.e. ~300mm^2). The weight of each link was normalized to its length. The histograms are presented in Figure 14. Figure 15.a presents the delay of each link (1mm, 1.89mm and 3.4mm) vs. technology node, assuming minimum size un-repeated global links (Table III); the cost-function per mm (Eq. 8) of adjusting each of the links for target frequencies for 1-5 GHz vs. technology node are presented in Figures 15.b-d. Figures 14-15 confirm our assumptions. Short links, which are the vast majority (Figure 14), can be adjusted to satisfy frequencies of 1-5 GHz with almost no overhead (Figure 15.b). The cost of adjusting the longer links becomes significant only at long-term technology nodes (≤14nm) and high target frequencies (≥4 GHz). Cost function is marked as ∞ where target frequency is not achievable in one cycle under the design constraints from Table IV. We use the cost function CF from (8) to calculate the overall interconnect costs of adjustments of long links. We define NoC Interconnect Overhead (NIO) as follows:

$$NIO = \frac{\sum_{All\ NoC\ Links} CF(\text{adjusted link}) \cdot \text{Length(link)}}{\sum_{All\ NoC\ Links} \text{Length(link)}} \qquad (10)$$

NIO is defined as the ratio between the overall links cost after and before the delay reduction adjustments. We have calculated NIO for the 16x16 systems from Figure 13 at different technology nodes and target frequencies. The results (Figure 16) indicate that the cost (in terms of wiring and power) of adjusting hierarchical NoC links to satisfy clock frequency in a single cycle is negligible unless technology node is very advanced (≤12nm) and the target frequency is high (≥4 GHz).

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 (AMT) as 1/6 of the net area of 6T SRAM cell provided by ITRS [30]. A single repeater is composed of two transistors (NMOS and PMOS); therefore,

assuming AMT 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:

$$RArea_{Ratio} = \frac{\sum_{\text{All NoC Repeaters}} 2A_{MT} \cdot h(repeater) \cdot SR(repeater)}{die\ area} \quad (11)$$

In each technology node, we calculated the average $RArea_{Ratio}$ of the systems from Figure 13 for clock frequencies of 1–5 GHz assuming 64 bits wide links. The results imply that the area required for repeaters within a whole system is negligible, as presented in Figure 17.

In hierarchical systems with localized traffic, the hierarchy levels provide shortcuts for the few packets that traverse long distances. As global packets are only a minority in large (hundreds to thousands of modules) hierarchical NoCs [1], we have concentrated our evaluation and discussion on hierarchical NoCs with uniform bandwidth links, using a reasonable range of link allocations at the high hierarchy levels. The potential bottlenecks at the upper hierarchy levels in these NoCs can be addressed with dynamic traffic distribution among the hierarchy levels [26]. Although uniform-bandwidth links are the typical design choice used in NoCs, our methodology provides the tools to estimate overheads of long-links for non-uniform link-bandwidth as well. Note that such heterogeneous NoCs incur higher latency in routers, due to the complexity of ingress-egress mismatch [37].

7. DISCUSSION AND CONCLUSIONS

We have analyzed the distribution of length of links in hierarchical NoCs. We have shown that there is almost no benefit in utilizing hierarchy levels with radix that is smaller than 4x4. As a consequence, we have concluded that longest links in hierarchical NoCs should not exceed lengths of a few millimeters. Thereupon, we have evaluated power, wiring and silicon area costs of adjusting the delay of long links in hierarchical NoCs to meet practical target clock frequencies using wire sizing and repeater insertion. We have defined a unified cost function that takes into account wiring and power costs (silicon area cost of the repeaters was accounted separately) and introduced a methodology to find the lowest cost parallel links that satisfy the required delay under design constraints. Two 16x16 hierarchical NoCs (PyraMesh [1] and hybrid ring/mesh [3]) were used as representative use-cases to evaluate the costs. The results in Figures 16 and 17 indicate that using long parallel links is feasible in a wide span of present and future hierarchical NoCs. Frequencies of 1-4 GHz are reachable using wire sizing and repeater insertion (with the constraints of Table IV) in all the systems and technology nodes that we have evaluated. 5 GHz can also be reached in almost all the systems, except for few configurations at very advanced technology nodes. In systems where target clock can't be reached, techniques such as pipelining should be used in the very few long top-level links. The average NoC interconnect overheads (i.e. power and wiring, based on CF, NIO) of adjusting the system to frequencies of 1–4 GHz are 0.05%, 0.69%, 2.85% and 10.53% respectively across all the system configurations and technology nodes that we have evaluated. The average overhead for 5 GHz, except the configuration where 5 GHz can't be achieved, is 5.05%. The average area required by repeaters in systems adapted to 1-5 GHz (again except the unfeasible configurations for 5 GHz at 10 and 8 nm nodes) is only 0.16% of the die.

In conclusion, our evaluation shows that parallel single-cycle links are feasible at a reasonable cost in present and future hierarchical on-chip networks. We have studied that single-cycle links can be adapted to frequencies as high as 5 GHz and beyond in

present and near future technology nodes. Having in mind that present CMP's rarely exceed frequencies of 1 GHz (e.g. Tilera's CMP – 1 GHz @ 90nm [17], Adapteva's CMP – 1 GHz @ 28nm [18], Kalray's CMP – 400 MHz @ 28nm [19]), we believe that single-cycle parallel links will be feasible in the vast majority of future hierarchical NoCs. Our work provides a methodology to estimate the feasibility and the system costs for using long single-cycle parallel links in present and future hierarchical NoCs.

References

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

[2] C. Puttmann, J. Niemann, M. Porrmann and U. Ruckert, "GigaNoC - a hierarchical network-on-chip for scalable chip-multiprocessors," in *proceedings of the 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 *proceedings of the international symposium on networks-on-chip (NoCs)*, 2007.

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

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

[6] U. Ogras and R. Marculescu, ""It's a small world after all": NoC performance optimization via long-range link insertion," *transactions on very large scale integration (VLSI) systems*, vol. 14, no. 7, pp. 693-706, 2006.

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

[9] J. Cong and K. S. Leung, "Optimal wire sizing under Elmore delay model," *transactions on computer-aided design of inegrated circuits and systems*, vol. 14, no. 3, pp. 321-336, 1995.

[10] H. 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," *transactions on electron devices*, vol. 49, no. 11, pp. 2001-2007, 2002.

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

[13] L. Scheffer, "Methodologies and tools for pipelined on-chip interconnect," in *proceedings of the international conference on compute design: VLSI in copmuters and processors (ICCD)*, 2002.

[14] M. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher and S. W. Tam, "CMP network-on-chip overlaid with multi-band RF-interconnect," in *proceedings of the international symposium on high performance computer architecture (HPCA)*, 2008.

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

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

[17] D. Wentzlaff, P. Griffin, H. Hoffmann, L. Bao, B. Edwards, C. Ramey, M. Mattina, C. C. Miao, J. F. Brown and A. Agarwal, "On-chip interconnection architecture of the tile processor," *IEEE Micro*, vol. 27, no. 5, pp. 15-31, 2007.

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

[19] "Kalray MPPA 256," [Online]. Available: http://www.kalray.eu/products/mppa-manycore/mppa-256/.

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

[21] A. Pullini, F. Angiolini, S. Murali, D. Atienza, G. De Micheli and L. Benini, "Bringing NoCs to 65 nm," *Micro IEEE*, vol. 27, no. 5, pp. 75-85, 2007.

[22] M. Ferraresi, G. Gobbo, D. Ludovici and D. Bertozzi, "Bringing network-on-chip links to 45nm," in *proceedings of the international symposium on system-on-chip (SoC)*, 2011.

[23] D. Goren, M. Zelikson, T. Galambos, R. Gordin, B. Livshitz, A. Amir, A. Sherman and I. Wagner, "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.

[24] A. Deutsch, P. W. Coteus, G. V. Kopscay, H. H. Smith, C. W. Surovic, B. L. Krauter, D. C. Edelstein and P. J. Restle, "On-chip wiring design challenges

for gigahertz," *proceedings of the IEEE,* vol. 89, no. 4, pp. 529-555, 2001.

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

[26] R. Manevich, I. Cidon and A. Kolodny, "Dynamic traffic distribution among hierarchy levles in hierarchical networks-on-chip (NoCs)," in *proceedings of the international conference on networks-on-chip (NoCs)*, 2013, 2013.

[27] D. Ludovici, F. Gilabert, S. Medardoni, C. Gomez, M. E. Gomez, P. Lopez, G. N. Gaydadjiev and D. Bertozzi, "Assessing fat-tree topologies for regular network-on-chip design under nanoscale technology constraints," in *proceedings of the design, automation and test in Europe conference and exhibition (DATE)*, 2009.

[28] D. Greenfield, A. Banerjee, J. G. Lee and S. Moore, "Implications of Rent's rule for NoC design and its fault tolerance," in *proceedings of the international symposium on networks-on-chip (NoCs)*, 2007.

[29] W. Heirman, J. Dambre, D. Stroobandt and J. Campenhout, "Rent's rule and parallel programs: characterising network trafic behaviour," in *proceedings of the international workshop on system level interconect prediction (SLIP)*, 2008.

[30] "The international technology roadmap for semiconductors (ITRS)," [Online]. Available: http://www.itrs.net/.

[31] D. A. Papa and I. L. Markov, "State of the art in physical synthesis," *Multi-objective optimization in physical synthesis of integrated circuits,* vol. 166, pp. 11-18, 2013.

[32] G. Santos, T. Reimann, M. Johann and R. Reis, "The fidelity property of the Elmore delay model in actual comparison of routing algorithms," in *proceedings of the interntional comference on computer design (ICCD)*, 2010.

[33] A. I. Abou-Seido, B. Nowak and C. Chu, "Fitted Elmore delay: a simple and accurate interconnect delay model," *transactions on very large scale (VLSI) systems,* vol. 12, no. 7, pp. 691-696, 2004.

[34] Y. I. Ismail, E. G. Friedman and J. L. Neves, "Figures of merit to characterize the importance of on-chip inductance," *transactions on VLSI systems,* vol. 7, no. 4, pp. 442-449, 1999.

[35] S. Wong, G. Lee and D. Ma, "Modeling of interconnect capacitance, delay, and crosstalk in VLSI," *transactions on semiconductor manufacturing,* vol. 13, no. 1, pp. 108-111, 2000.

[36] "On-chip parallel links calculator," [Online]. Available: http://ranman.eew.technion.ac.il/on-chip-parallel-links-calculator/.

[37] Y. Ben-Itzhak, I. Cidon and A. Kolodny, "Optimizing hetrogeneous NoC design," in *proceedings of the international workshop on system level interconnect prediction (SLIP)*, 2012.

TABLE I. AVERAGE HOP-DISTANCE OPTIMIZED HIERARCHICAL NOCS

Size of Baseline Mesh	Upper Levels	C_i	Longest Link [mm]
8x8	[4x4], [2x2]	[2,2,1]	5.67
16x16	[8x8], [4x4], [2x2]	[2,2,2,1]	5.67
32x32	[16x16], [8x8], [4x4], [2x2]	[2,2,2,2,1]	5.67
64x64	[32x32], [8x8], [4x4], [2x2]	[2,2,2,2,1]	5.67
128x128	[64x64], [16x16], [8x8], [4x4], [2x2]	[2,2,2,2,2,1]	5.67

TABLE II. AVERAGE HOP-DISTANCE OPTIMIZED HIERARCHICAL NOCS WITH UPPER LEVEL RADIX LIMITED TO 4X4

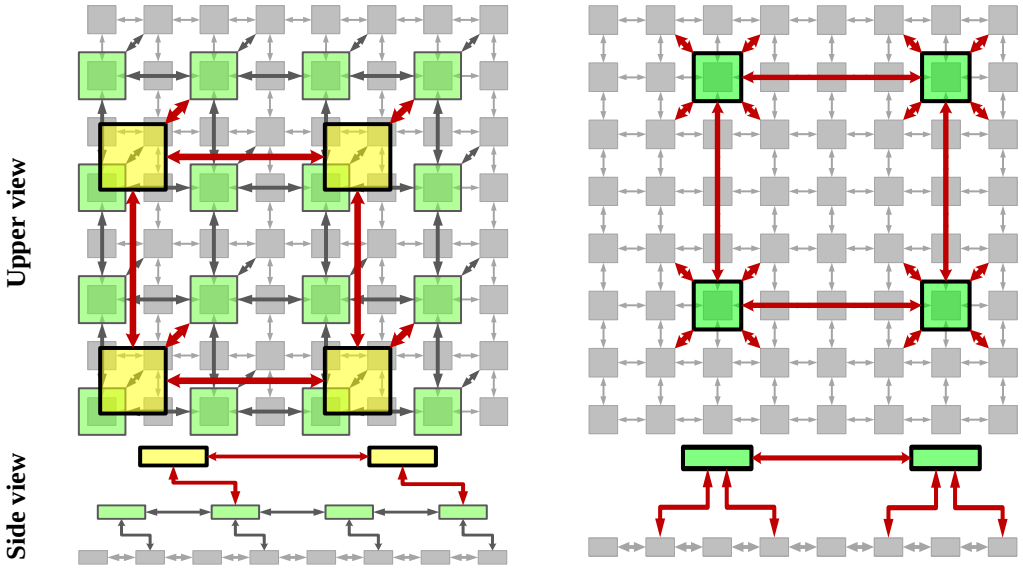
Size of Baseline Mesh	Upper Levels	C_i	Longest Link [mm]
8x8	[4x4]	[2,1]	3.4
16x16	[8x8], [4x4]	[2,2,2,1]	3.4
32x32	[16x16], [8x8], [4x4]	[2,2,2,1]	3.4
64x64	[32x32], [16x16], [8x8], [4x4]	[2,2,2,2,1]	3.4
128x128	[32x32], [16x16], [4x4]	[2,2,2,1]	3.4

TABLE III. INTERCONNECT DIMENSIONS AND ATTRIBUTES OF MINIMUM SIZE GLOBAL WIRES [30]. W,S,T, AND H ARE ILLUSTRATED IN FIGURE 5. "RES." STANDS FOR CONDUCTOR EFFECTIVE RESISTIVITY OF CU WIRE INCLUDING THE EFFECT OF WIDTH DEPENDANT SCATTERING. C WAS CALCULATED ACCORDING TO THE MODEL IN [35].

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
L_{eff}[nm]	29	26.5	24.2	22.2	20.2	18.4	16.8	15.3	14	12.8	11.7	10.7	9.7	8.9	8.1
W[nm]	91.5	77	64.5	54	45.5	40.5	36	32	28.5	25.5	22.5	20	18	16	14
S[nm]	91.5	77	64.5	54	45.5	40.5	36	32	28.5	25.5	22.5	20	18	16	14
T[nm]	214	180	151	126	106	95.8	84.2	74.9	66.7	59.7	52.7	46.8	42.1	37.4	32.8
H[nm]	137	116	96.8	81	68.3	60.8	54	48	42.8	38.3	33.8	30	27	24	21
$K_{dialect}$	3.1	2.75	2.75	2.75	2.6	2.6	2.6	2.3	2.3	2.3	2.15	2.15	2.15	1.85	1.85
Res. [μΩ-cm]	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06
R [Ω/mm]	1051	1484	2116	3019	4252	5367	6792	8597	10838	13539	17389	22009	27171	34388	44915
CG [fF/mm]	36.14	32.06	32.06	32.06	30.31	30.31	30.31	26.81	26.81	26.81	25.06	25.06	25.06	21.56	21.56
CC [fF/mm]	78.02	69.21	69.21	69.21	65.43	65.43	65.43	57.88	57.88	57.88	54.11	54.11	54.11	46.56	46.56

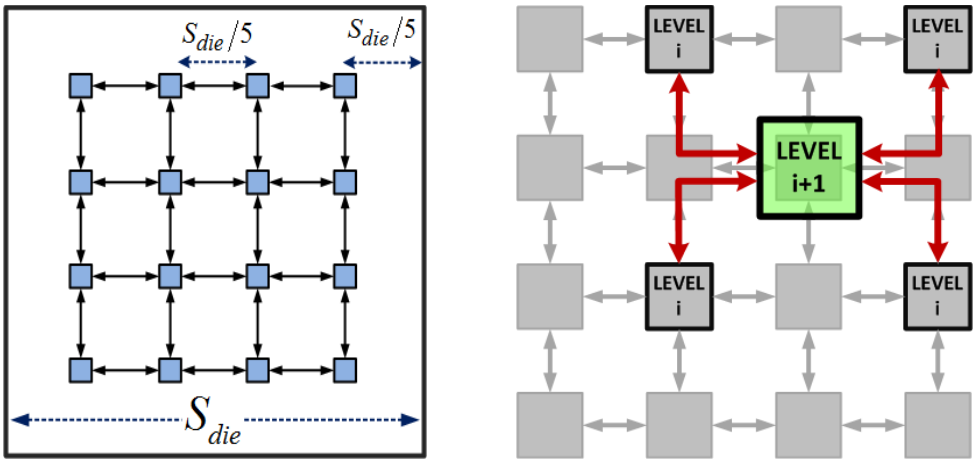
TABLE IV. RANGES OF DESIGN PARAMETERS

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



(a) (b)

Figure 1. 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].



(a) (b)

Figure 2. (a) - Relation between length of mesh links and the size of the die S_{die}. (b) -

Illustration of global wiring between levels i and i+1 for C_i>1.

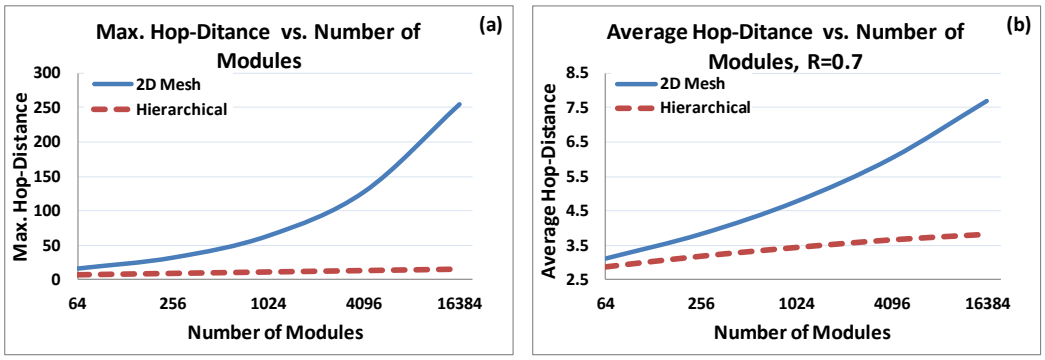


Figure 3. Maximum (a) and average (b) hop distance of flat 2D Mesh and hierarchical PyraMesh NoCs vs. number of modules (i.e. nodes at the bottom level).

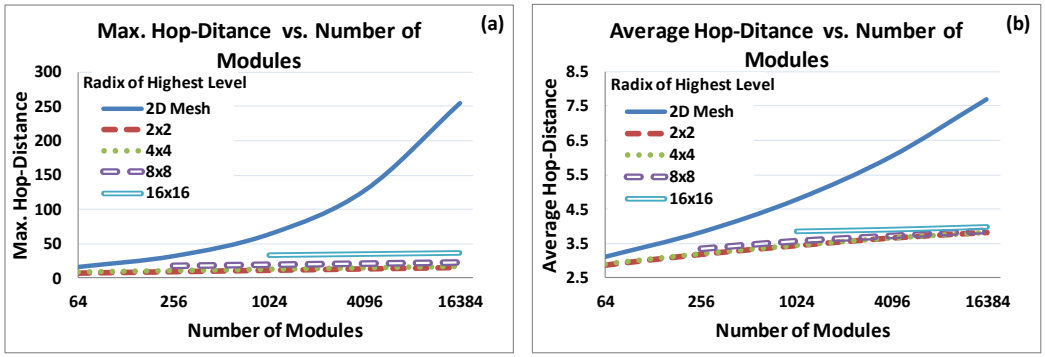


Figure 4. Maximum (a) and average (b) hop distance with restricting the radix of the upper hierarchy level.

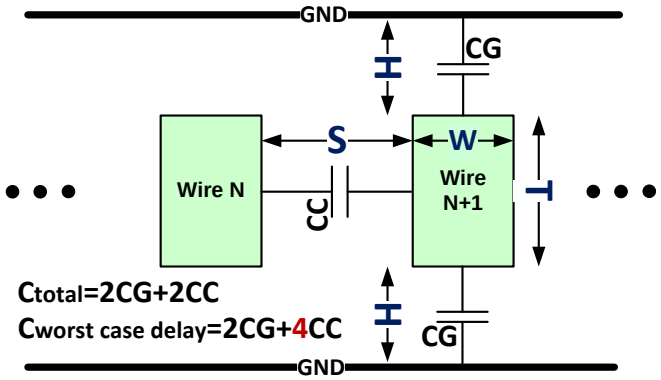


Figure 5. Interconnect physical model.

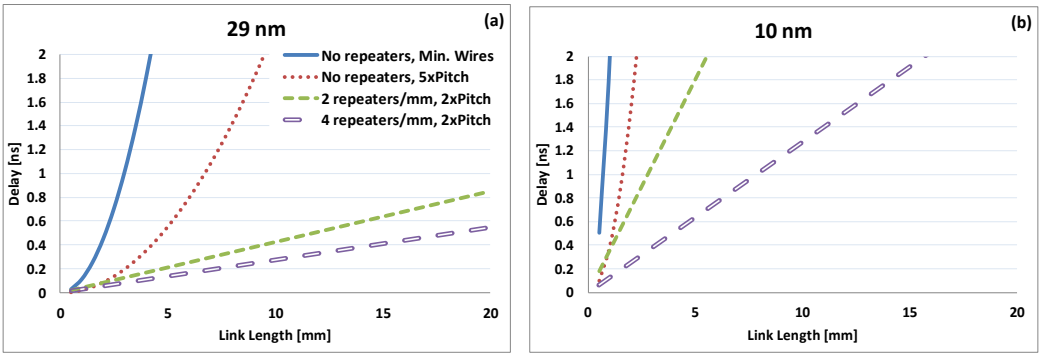


Figure 6. The four configurations presented are: I-Minimal size wires without repeaters, II-Wires with $\Lambda_W=\Lambda_S=5$ and no repeaters, III- $\Lambda_W=\Lambda_S=2$ with 2 repeaters/mm, IV- Wires with $\Lambda_W=\Lambda_S=2$ with 4 repeaters/mm. (a)-(b) present 29nm and 10nm technology nodes respectively.

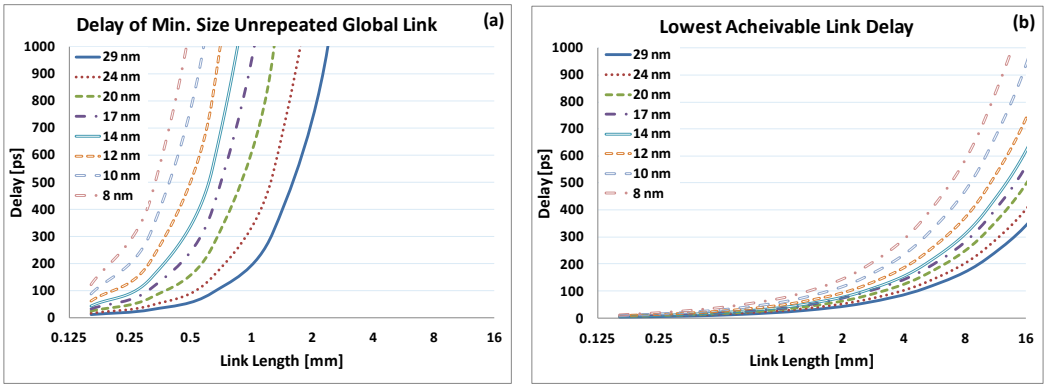


Figure 7. (a) – Delays of minimum size un-repeated global links. (b) – The lowest achievable delay of links vs. length using wire sizing and repeaters insertion subject to design parameters ranges presented in Table IV.

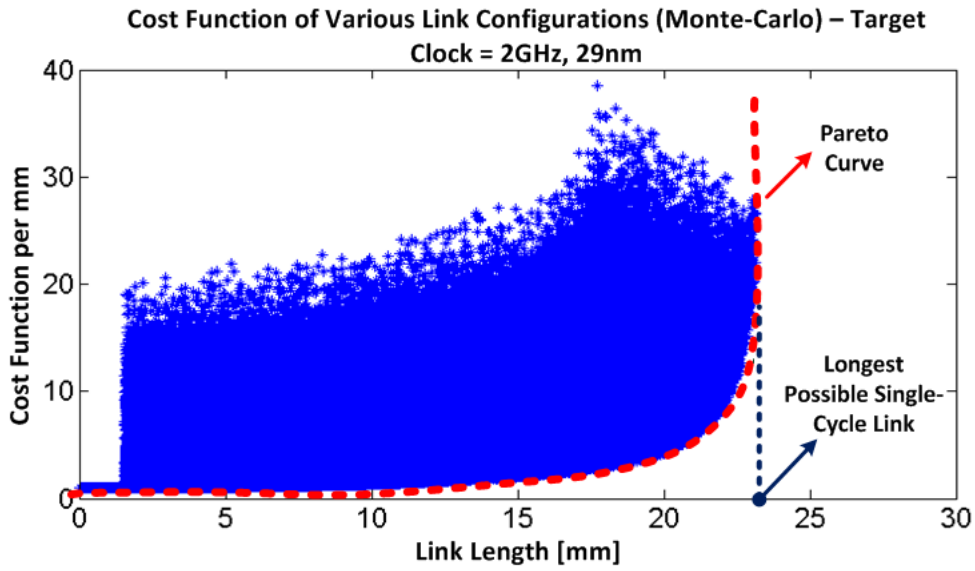


Figure 8. Results of Monte-Carlo analysis of interconnect cost, subject to CF (7) with α=β=0.5, for 29nm technology node and a target clock frequency of 2 GHz.

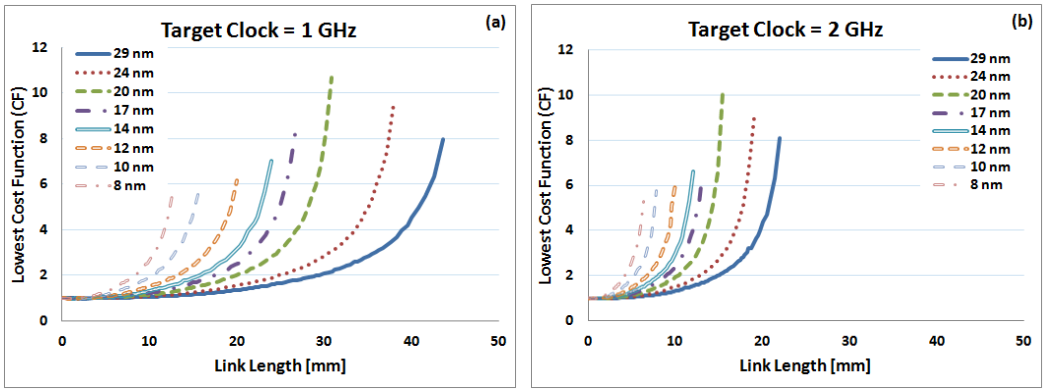


Figure 9. Minimal cost function for different technologies, (a)-for 1 GHz and (b)-2 GHz operating frequencies.

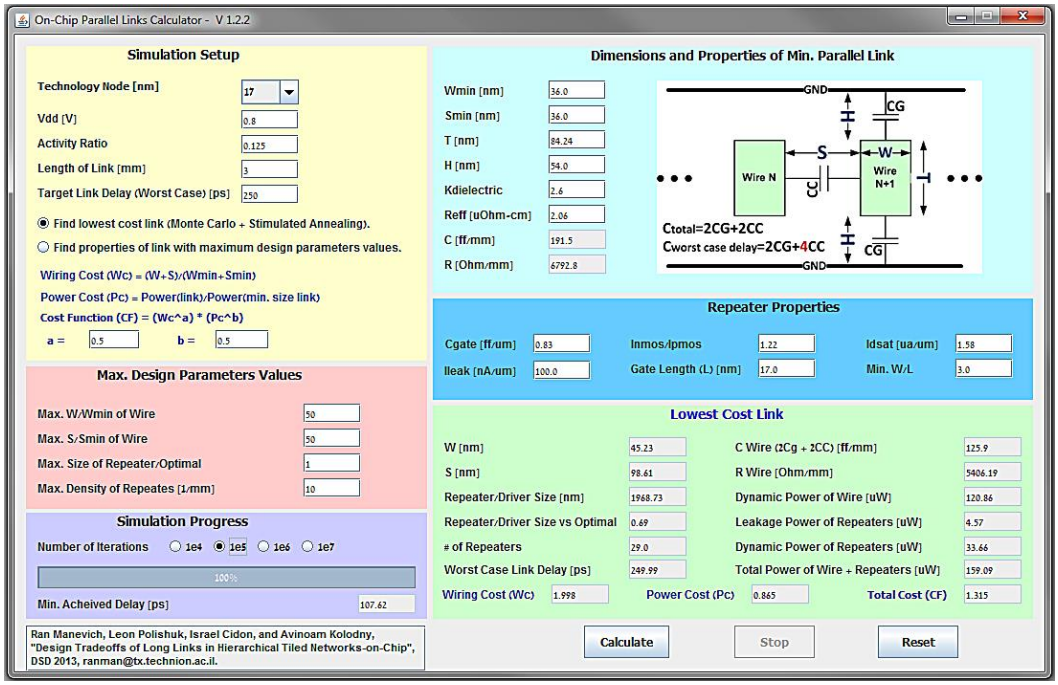


Figure 10. A screenshot of our parallel links calculator available at http://ranman.eew.technion.ac.il/on-chip-parallel-links-calculator [36].

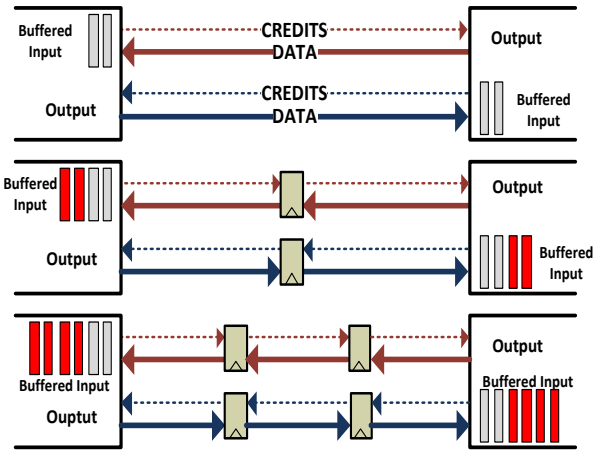


Figure 11. Extra buffering (red bars) due to pipelining of long links.

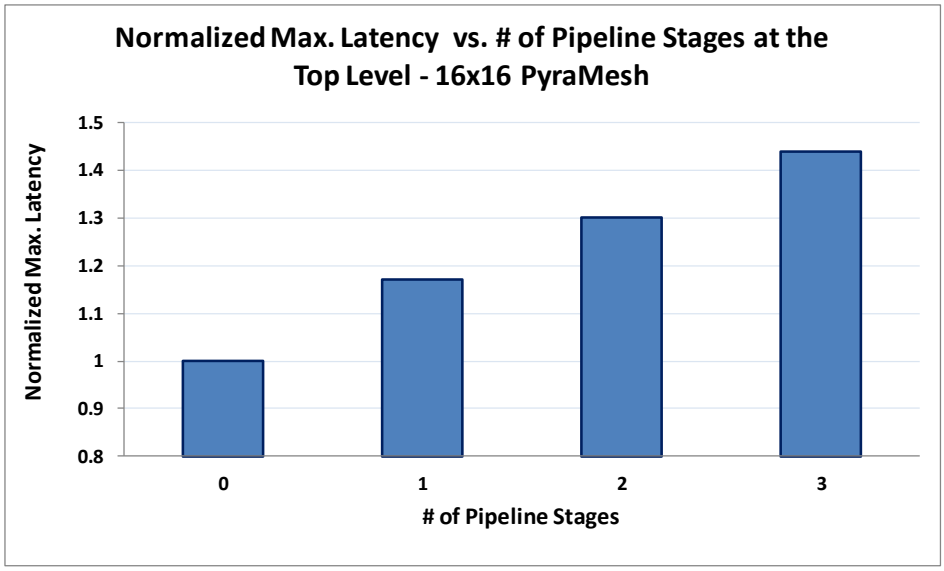
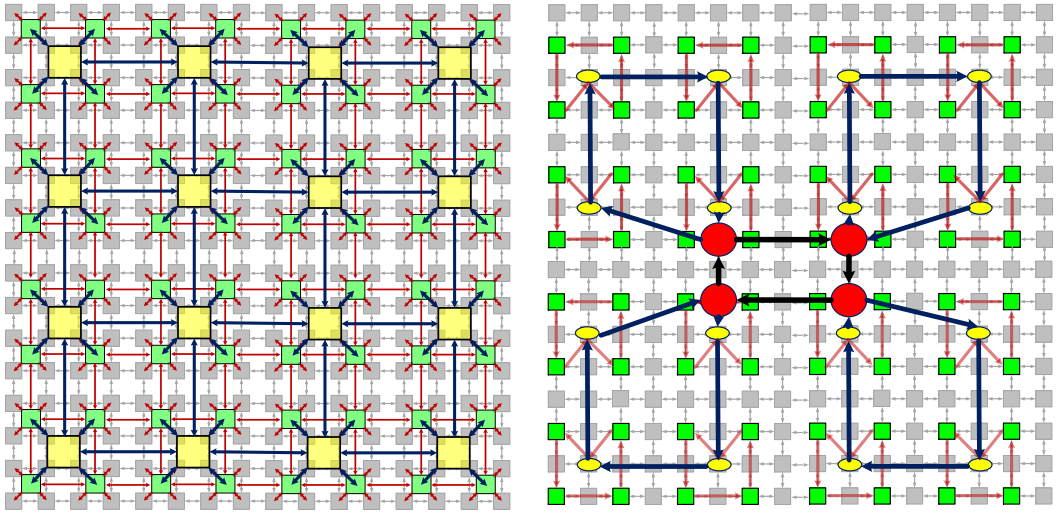


Figure 12. Normalized longest-path latency in 16x16 PyraMesh vs. number of pipeline stages at the top level.



(a) (b)

Figure 13. (a) – 16x16 PyraMesh with 8x8 and 4x4 upper levels. (b) – 16x16 hybrid ring/mesh [3] with 16 rings at the second level, 4 at the third and 1 at the fourth.

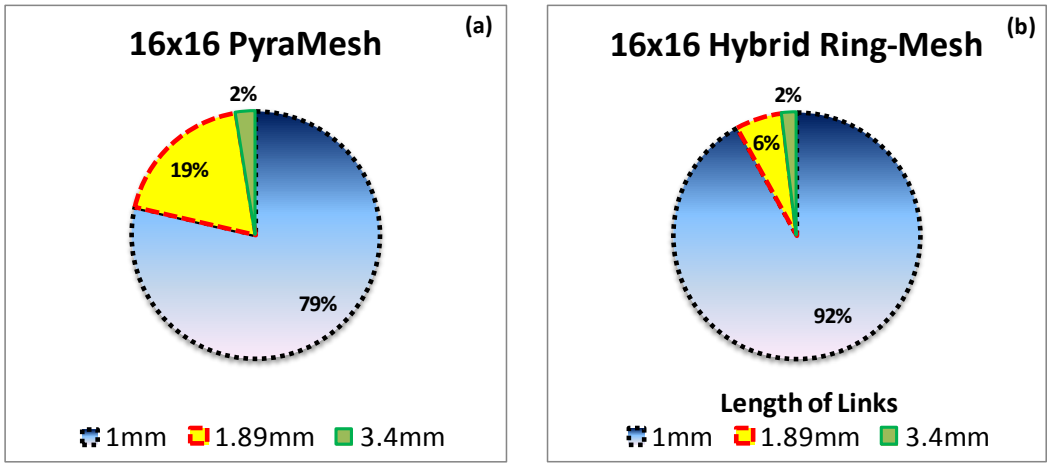


Figure 14. (a) – Links lengths histogram – 16x16 PyraMesh [1]. (b) – Links lengths histogram – hybrid ring/mesh [3].

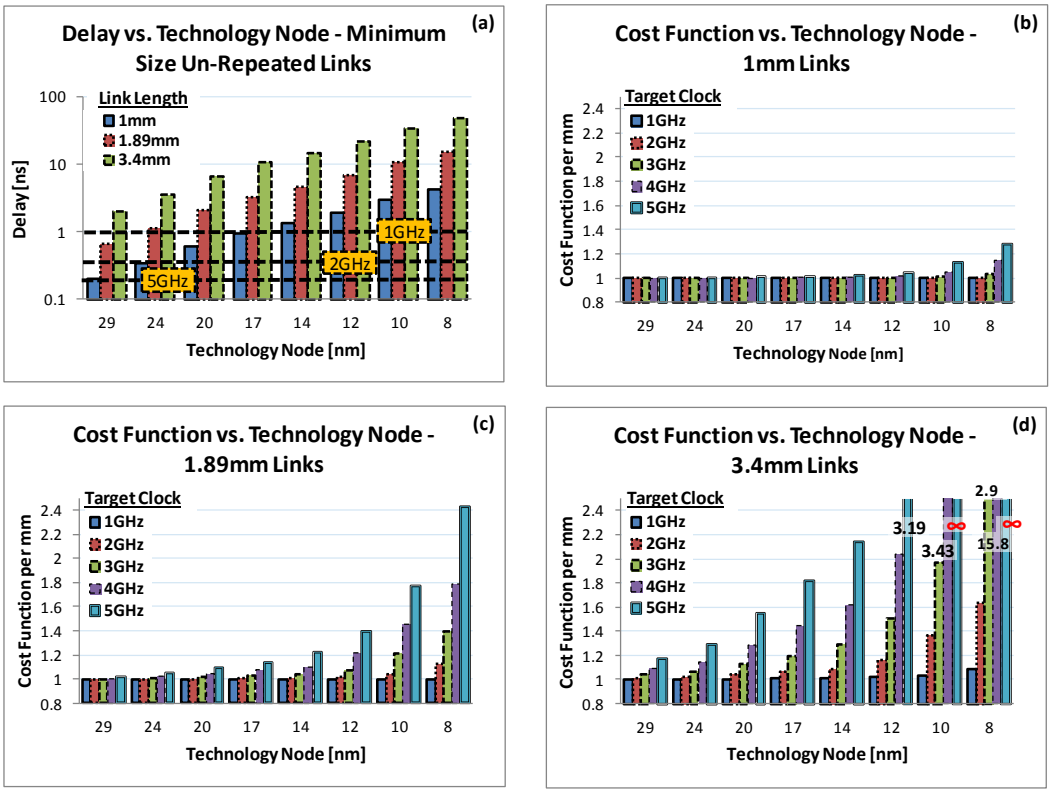


Figure 15. (a) – Delay of minimum size global parallel links vs. technology node. (b)-(d) – Cost function of adjusting 1mm (a), 1.89mm (b) and 3.4mm (d) links to target clock frequencies of 1-5 GHz vs. technology node. The numbers that appear inside the data area in (d) indicate the values of the out-of-range bars. 5 GHz cannot be achieved in 3.4 mm links @ 8 nm and 10 nm nodes, the respective bars are marked with ∞.

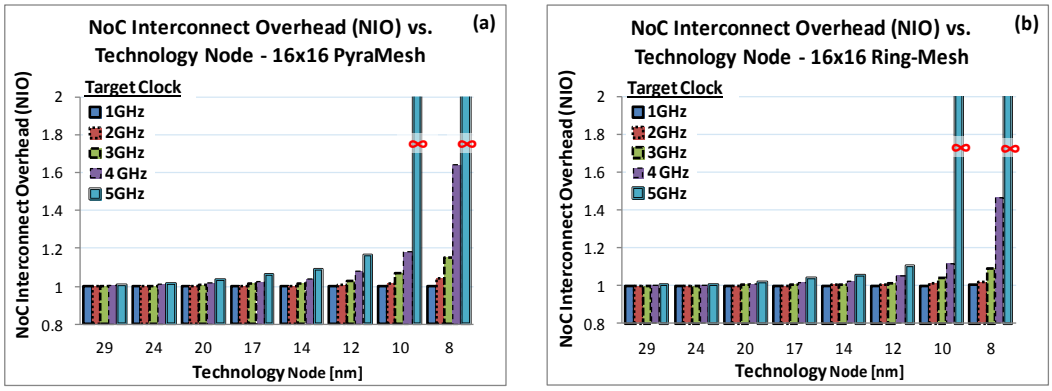


Figure 16. NoC Interconnect Overhead (NIO) of adjusting parallel links to meet clocks of 1-5 GHz in 16x16 PyraMesh [1] and hybrid ring/mesh [3] NoCs. Not-feasible configurations are marked with ∞ (Figure 15).

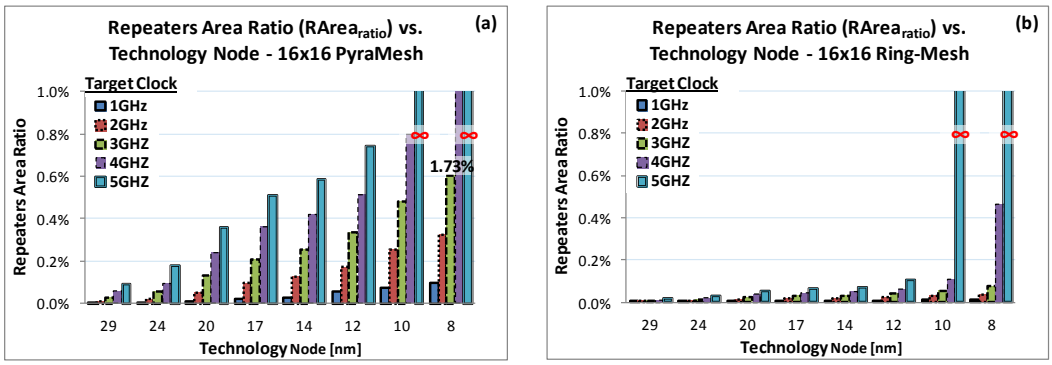


Figure 17. The percentage of area required by repeaters of adjusted links in PyraMesh (a) and hybrid ring/mesh (b) assuming a 17x17 mm die.