

# Throughput Driven Unidirectional Bus Design for NoC Applications

Giuseppe S. Garcea, Nick P. van der Meijs  
Delft University of Technology, Faculty of EEMCS, The Netherlands  
giuseppe@cas.et.tudelft.nl

**Abstract—** At the physical layer of Network-on-Chip (NoC) implementations, blocks and switches are already assigned and inter block connections have to be planned. Most of the inter block connections will consist of a large number of long parallel wires. Such wires usually have significant resistance and (coupling as well as self) capacitance. As such, these wires may severely limit the achievable performance. The performance may (and usually is) improved by buffering, i.e. the insertion of restoring electronics to avoid the quadratic dependence (caused by the product of  $r$  and  $c$  per unit length) of the delay on wire length.

The main contribution of this paper is in providing an optimization approach for the throughput on such buffered NoC wires. By extending our work of [2], we provide a methodology to optimize the throughput under buffer area constraints. It appears that in that case the highest throughput can be obtained using sub-maximal density configurations with greater than minimum wire spacing and width. Moreover, the results show that a significant saving on Si area for buffering need only cost a moderate decrease of throughput. For example, a specific experiment for a  $0.18\mu\text{m}$  technology yields 90% of the absolute maximum throughput using only 30% of the corresponding buffer area.

## I. INTRODUCTION

In order to cope with the increasing complexity of the systems and to boost design productivity it is important to be smart in using, reusing and/or adding new design parts. A wide part of emerging microelectronics research is dedicated to this goal. During the 1990s, the research aimed at integration of different large components and/or processing cores on a single die. The resulting systems became known under the name *System on Chip (SoC)*. As silicon technology has advanced, several problems have emerged, especially in the field of the communication part of the SoCs. The problem of finding suitable communication structures has to be addressed at all levels from physical to architectural to the operating systems and application level [3]. For the branch of research which focus simultaneously on integration of different design parts and communications structures the term *Network on Chip (NoC)* [4], [5] is used.

A common way to implement such NoC architecture is by using a regular mesh of switches and resources [6]. Re-

sources can be processor cores, memories, custom hardware blocks or any generic IP. Switches are used to route and buffer messages and data between resources.

Switches and resources are connected by input and output channels. A channel consists of an unidirectional point-to-point bus. Such NoC architectures have been proposed as solution in order to improve the scalability and the modularity such that the complexity and functional diversity can be handled systematically.

It seems obvious that this approach introduces challenging problems. The first problem is to identify the real nature of applications that are possible to implement optimally or near optimally on an NoC architecture. We are not attacking this problem because it has to be solved at a higher system level, while the work presented here remains more at the physical level. However, an important physical level problem is timing closure. Indeed, even small changes in adding gates to the netlist can result in a change of timing of the complete system with a consequent unexpected change of placement and routing. In this context, a systematic approach to predict and to select (by tuning the dimensions) the appropriate communication system, becomes essential.

For the type of on-chip buses that we consider, the wire parasitics are important. Coupling is becoming critical because the side wall capacitance is increasing with the wire aspect ratio. This has an effect on the signal propagation properties, and has to be accounted for during design. Thus, buffering can be required to increase the performance of the wires and to decrease the adverse affects of coupling.

In this paper we will focus on unidirectional bus design connecting the system blocks. We model our bus by using a channel of uniform parallel wires that has fixed width  $W_{ch}$  and fixed length  $L$  as presented in Figure 1. We propose a wire sizing approach combined with a repeater planning for the channel. We assume that wire width and spacing are uniform for every wire and also that the buffer insertion is uniform.

We will now first discuss some preliminaries in Section II and in particular some qualitative trends using a simplified capacitance model in Section II-B. Then, in Section

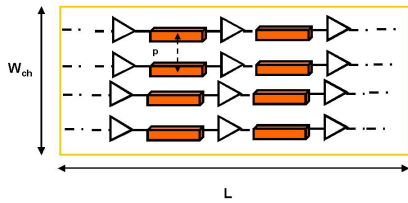


Fig. 1. Bus Model.

III we discuss the buffer and wire sizing solution, first for the unbuffered and the unconstrained case in Section III-B and subsequently for the constrained case in Section III-C. We then conclude in Section IV.

## II. PRELIMINARIES

### A. Introduction

Optimizing buses in fact entails selecting the optimal spacing and width of the wires, but also the proper size and distance of the buffers. For the buffering, we consider in the sequel of this paper three different policies.

- *No buffering.* This is normally not a good policy, because of the quadratic dependence of the delay on the distance. However, we include it in our analysis because it helps in understanding the problem more fully.
- *Unconstrained buffering.* This policy refers to a dimensioning of the buffers (distance and size) for maximum performance. Given wire resistance and capacitance, Bakoglu [1], [2] has derived buffer size and distance that minimize the latency for a single line. However, the Bakoglu model only considers single lines.
- *Constrained buffering.* This policy refers to buffer size and distance such that the performance is maximal under total buffer area constraints. An analytic model for the area-latency trade-off for single lines was presented in [2]. Sometimes, we also use the adjective "optimal" in relation to buffering, but in fact it is redundant. For both constrained and unconstrained buffering we only consider the optimal case. When we would apply buffering methodologies introduced above for regular pipelined bus structures as the one described in Figure 1, a measure of performance can not be exclusively the single wire latency. Instead, it has to be influenced also by the wire density. Indeed, design strategies (such as clocked and pipelined interconnect) might be able to tolerate latency, and the design objective could become throughput. Here, we define throughput as the amount of information that can be sent per unit time and per unit bus width (in microns, not in number of wires). In particular, it might be better to use many wires with small pitch than to use few wires with a large pitch. Wide wires have lower latency but more narrow wires fit in the same space. As a result, the throughput

can be higher.

A throughput-centric design strategy has already been introduced in [7], but wire sizing is derived only for optimal unconstrained buffer insertion. In [8], a first attempt to maximize throughput under controlled buffer area usage is presented, but this work has the limitation that buffer area reduction is obtained by resizing only the buffer size without changing their distance and number. This is a suboptimal buffering scheme.

In [9], the throughput is maximized under constrained buffering, by using iterative algorithms based on closed form expressions that are more accurate than the Elmore delay. Our approach also optimizes throughput under constrained buffering. We use an analytic model as presented in [2] to arrive at optimal constrained buffer distance and size for a certain wire geometry, while [9] uses a global numeric optimization approach. It is in fact not clear how [9] finds the optimal number of buffers along a line, while in our approach this number unambiguously follows analytically. In addition, the increased insight gained by our analytic approach helps in understanding the exact nature of the bus optimization problem.

We will now first introduce our design resources. In particular we will present the relations between throughput and buffer area. Consider a bus of wires introduced above and in Figure 1.  $L$  and  $W_{ch}$  denote the total physical length and width of the channel, respectively. Furthermore,  $N_W$  denotes the number of wires, which have width  $w$  and spacing  $s$  where the pitch  $p$  is given by  $p = w + s$ . Thus, we have  $N_W = W_{ch}/(w + s)$ . The wires are uniformly buffered, with  $l$  the distance between buffers and  $N_b$  the number of buffers. For simplicity we take  $N_b = L/l$  (in reality the number of buffers is one plus the number of segments). Furthermore, we define  $\tau$  as the delay of a single buffered segment and  $v = l/\tau$  as the signal propagation velocity. Now we can express the throughput as

$$T = \frac{1}{N_b \cdot \tau} \times N_w = \frac{l}{L \cdot \tau} \times \frac{W_{ch}}{w + s} = \frac{v}{w + s} \times \frac{W_{ch}}{L}. \quad (1)$$

While shielding can improve the delay predictability, it usually does not improve the throughput as defined above. Although the signal line density is reduced, the reduced effective (Miller) capacitance does not sufficiently improve the performance of a single line. Thus, other techniques to reduce coupling effects, like signal alignment of adjacent wires are preferred to shielding. However, if necessary, our models can be extended to include shielding. For simplicity we will not do so in this paper.

Motivated by the fact that  $T$  is proportional to  $W_{ch}$  and inversely proportional to  $L$  (if  $v$  is constant), we will in the rest of this paper use the normalized or square throughput,

$T_{\square}$ , as the throughput for a square channel:

$$T_{\square} = \frac{v}{w+s} \quad (2)$$

By using the same bus structure and the same notation as introduced in defining the throughput, we can now define the total buffer area required for buffering in the channel, where  $w_b$  is the size of a single buffer:

$$A_{ch} = N_W N_b w_b = \frac{LW_{ch}}{l \cdot (w+s)} w_b \quad (3)$$

Note that  $w_b$  is actually only proportional to the buffer area. In this paper we will ignore the constant multiplicative constant that relates the transistor width to this buffer area. This constant is only layout and technology dependent and does not change our model.

We present in the following subsection, by using qualitative trends, how to select the wire density depending on the resources just introduced. We will illustrate the cases of no buffering and of unconstrained buffering. This will help in understanding that the density has to be selected considering a trade-off between throughput and buffer area.

### B. Simplified qualitative trends

In this subsection exclusively, we use a parallel plate capacitance model. Here this simplistic formulation gives clear qualitative insight in the factors involved in selecting the appropriate wire density. We postpone the quantitative results to the next section, using more accurate capacitance models. For simplicity, we will assume here that  $w = s = p/2$  but in the rest of this paper this assumption will be relaxed.

We use a simple first order parallel plate model for capacitance  $c$  per unit wire length, that includes the coupling to the layers above and below the wire as well as to the neighboring wires, as follows:

$$c = \epsilon w / t_{ox} + 2\epsilon SF h / s \quad (4)$$

Here,  $\epsilon$  is the permittivity of the medium,  $w$ ,  $s$ ,  $h$  and  $t_{ox}$  are wire width, spacing, wire thickness and interlayer spacing, respectively and  $SF$  is the so-called switch factor accounting for the neighbor line activity [11].

Resistance per unit wire length will be given by

$$r = \rho / wh. \quad (5)$$

For long lines without buffers, the delay is of course proportional to  $rcL^2$ . Thus, the signal propagation velocity is proportional to  $v = 1/rcL$ . Using this and  $w + s = p$ , the normalized throughput (2) can be written as  $T_{\square} \propto 1/rcp$ .

Thus, maximizing throughput is now equivalent to minimizing  $rcp$ . Using (4) and (5) with  $w = s = p/2$ , this is equivalent to

$$\text{minimize } f_1(p) = \frac{p}{ht_{ox}} + 8\frac{SF}{p}. \quad (6)$$

Then, an analytical value of the wire pitch which maximizes throughput is obtained by setting the first derivative with respect to  $p$  equal to zero. Thus, the wire pitch that maximizes the throughput for unbuffered buses is proportional to  $p = 2\sqrt{2SFht_{ox}}$ .

However, buffering can improve the single wire latency and then, implicitly, also the throughput. In this situation, we try to maximize the throughput for unconstrained total buffer area. It is possible to show [1], [2] that the reciprocal of the signal propagation velocity  $v^{-1}$  is then proportional to  $\sqrt{rc}$ . Thus, (2), can be rewritten as  $T_{\square} \propto 1/p\sqrt{rc}$ . Note the square root when comparing to the equivalent expression above for the unbuffered case. Thus, maximizing the throughput is now equivalent to minimizing  $p\sqrt{rc}$ . By combining this with (4) and (5) and again using  $w = s = p/2$ , we find that maximizing throughput becomes equivalent to

$$\text{minimize } f_2(p) = \sqrt{\frac{p^2}{ht_{ox}} + 8SF}. \quad (7)$$

Thus, in case of unconstrained buffering, the pitch should be chosen as small as possible.

In the first case, for unbuffered buses, the optimal pitch, actually balances between high density and small capacitance situations. In the second case, for unconstrained buffering, the optimal pitch is the smallest possible. Indeed, the buffers can compensate for the capacitive loading associated with the high density.

However, the increased throughput of unconstrained buffering is not always necessary and/or the cost (in terms of silicon area and power) of such buffering is too high. Then, it is natural to search for the best trade-off: either the minimum buffer area for a certain throughput or the maximum throughput achievable with a certain buffer area.

It is clear that throughput and buffer area can be traded by increasing the wire spacing. Informally, for constant latency of a line the increased spacing allows less buffering because the capacitance of a line is reduced. With constant latency but increased spacing, the density and hence the throughput is reduced. This reasoning, however, does not give a clue towards calculation of optimal trade-off, which might even be less clear when also the width of the line is allowed to change. However, in the next section we will investigate its solution.

### III. THROUGHPUT DRIVEN BUFFER INSERTION

#### A. Introduction

Until now we have considered a parallel plate capacitance model for the wire. However, it is known that fringing components of the capacitances become important for deep-submicron technologies where wires have a high aspect ratio. Therefore, we will use in the rest of this paper the analytical model from [10], which can accurately model the capacitance including the fringing terms.

Given the switching factor  $SF$ , we can find for each technology and each channel (total width and length) in the maximum density configuration the Bakoglu solution [1] for the buffer area. We will denote this value as  $A_{ch}^*$ . It is the total unconstrained buffer area required by the bus for minimum latency in the maximum density wire configuration for that switching factor. For the specific case of  $SF = 1$ , we will denote the corresponding normalized throughput as  $T_{\square}^*$ . If we exclude cases of  $SF < 1$ , based on the theoretical analysis of the previous section, this throughput is actually the maximum possible, which will indeed be confirmed below for the  $0.18\mu m$  technology example.

$A_{ch}$  will denote our design constraint for the buffer area available in the channel. The maximum useful buffer area is equal to  $A_{ch}^*$ , since this was determined for the most dense buses with the greatest capacitive loading, given the switching factor. The optimal buffer area for buses with a lower density is always less than  $A_{ch}^*$ . Then, the cases of no buffering, unconstrained buffering and constrained buffering correspond to  $A_{ch} = 0$ ,  $A_{ch} = A_{ch}^*$  and  $0 < A_{ch} < A_{ch}^*$ , respectively.

We will use  $A_{ch}^*$  and  $T_{\square}^*$  as normalization values. In particular, we define the normalized total buffer area constraint as  $\gamma_{ch} = A_{ch}/A_{ch}^*$  with  $0 < \gamma_{ch} < 1$ , and we use  $T_{\square}^*$  to normalize the throughput graphs below so that they can be compared.

Now, our optimization procedure works by sampling the space of allowable  $(w, s)$ , and for each sample we will optimize the latency using the procedure from [2] as constrained by  $A_{ch}$ . For some cases with a low density, the  $A_{ch}$  actually exceeds the optimally needed buffer area. In that case, the procedure from [2] simplifies to the Bakoglu sizing [1]. After the sampling, we can just select the design point with the greatest throughput.

This sampling method is efficient, since the whole formulation is analytic. In most cases, it would not be necessary to apply a more efficient search method for the optimum but otherwise it could be implemented if that would be desirable.

Since our intent is only to explain a methodology to

trade different resources in bus design, we will limit ourselves in the examples below to a particular technology. That is, in our examples we will use parameters typical for a  $0.18\mu m$  technology. For the value of those parameters we refer to [2]. However, our methodology is completely generic and can also be applied to other processes.

For our sampling procedure we need to specify the minimum width and spacing, which are denoted by  $w_{min}$  and  $s_{min}$ , respectively. In this paper, we take them representative for the top most metal layers in the  $0.18\mu m$  example technology and assume a value of  $0.6\mu m$ .

Furthermore, since the delay of unbuffered lines is quadratic in the length and that of buffered lines is linear in the length, we need to specify the length of the bus when we intend to compare different scenarios. This is true even while we use normalized values of throughput, see (2). For our examples, we assume that the length  $L = 2cm$ . Also, we take  $W_{ch} = 60\mu m$ , allowing 50 minimum pitch wires.

#### B. Unbuffered and Unconstrained Bus Design

Before we actually discuss constrained buffering, we will for comparison first consider the two limiting cases for bus design, namely that of no buffering ( $A_{ch} = 0$ ) and of unconstrained buffering ( $A_{ch} = A_{ch}^*$ ). The former case is trivial, and for the normalized throughput it follows that

$$T_{\square} \propto \frac{1}{Lrc(w+s)}. \quad (8)$$

Note that since the latency of unbuffered lines is quadratic in  $L$ , the normalized throughput  $T_{\square}$  as in (2) actually depends on  $L$ .

For unconstrained buffering, we can apply the Bakoglu procedure. Thus, for each sample from our  $(w, s)$  space, we get values of  $l_{crit}$  and  $w_{opt}$ , being the optimal buffer distance and size, respectively. The resulting buffer area per line will be denoted by  $A^U$ , and is given by

$$A^U = \frac{L}{l_{crit}} w_{opt}. \quad (9)$$

since  $L/l_{crit}$  is the number of segments which approximates the number of buffers along a line if  $L$  is sufficiently large. Please note the difference between  $A^U$  as defined above, which is actually a function of  $w$  and  $s$ , and  $A_{ch}^*$ , which actually is the value of  $A^U$  when  $w = w_{min}$  and  $s = s_{min}$ .

For the normalized throughput,  $T_{\square}$ , it then follows that

$$T_{\square} \propto \frac{1}{\sqrt{rc}(w+s)}. \quad (10)$$

Figures 2 and 3 illustrate the throughput of a bus for different wire densities for the two extreme cases considered in this subsection. Note that in these and the following throughput graphs, we actually show the ratio of  $T_{\square}$  to  $T_{\square}^*$ . Thus, they can directly be compared to evaluate the benefit of buffering. (But this benefit of course depends on the particular value of  $L$  considered.)

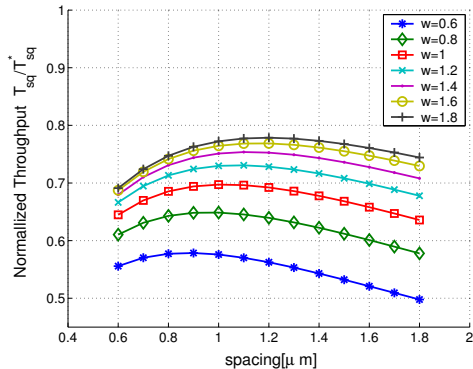


Fig. 2. Throughput without buffering.

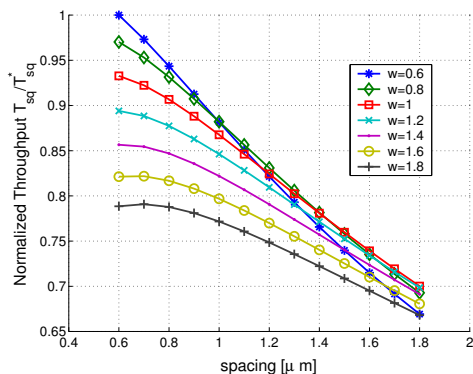


Fig. 3. Throughput with optimal unconstrained buffer insertion.

The results, as presented in Figures 2 and 3, seem to confirm the trends illustrated in the previous section. If no buffering is considered, the wire density which maximizes the throughput has an intermediate value, balancing between wire latency and density. If we allow unconstrained buffering, the maximum wire density solution clearly is the optimal solution. We will show in the next subsection that the choice of the appropriate wire density for the highest throughput will depend on the specific buffer area constraint.

### C. Constrained Bus Design

Here we consider the case of  $A_{ch} < A_{ch}^*$ . For each sample in our  $(w, s)$  space, we will optimize the latency using the procedure from [2] as constrained by  $A_{ch}$ .

However, the procedure of [2] effectively works on single lines, with the interaction to other lines only modeled

via the coupling capacitance. Therefore, we need to translate the overall  $A_{ch}$  into a specification per line. Moreover, the optimization procedure takes as input the buffer area constraint for a single line, say  $A$ , normalized to  $A^U$ . That is, the buffer area constraint is specified as

$$\gamma = \frac{A}{A^U}. \quad (11)$$

For bus design, we can relate this to  $A_{ch}$  by using  $A = A_{ch}/N_W$ , since each line will be buffered identically. Then,

$$\gamma = \frac{A_{ch}}{N_W A^U} \quad (12)$$

and because of (9) we get

$$\gamma = \frac{A_{ch} l_{crit}}{N_W L w_{opt}}. \quad (13)$$

Now, we can use the Bakoglu procedure to find  $l_{crit}$  and  $w_{opt}$  so that we can determine  $\gamma$ . Given  $\gamma$  and the relevant technology parameters, including wire resistance and capacitance, [2] calculates the buffer distance  $l$  and the size  $w_b$  for lowest latency under the area constraint specified by  $\gamma$ . We will denote this corresponding latency of a single segment by  $\tau(\gamma)$ . Then, the throughput (2) in the presence of a buffer area constraint  $\gamma$  can be written as follows:

$$T_{\square} = \frac{l}{\tau(\gamma)} \frac{1}{w + s} \quad (14)$$

In the examples below, we will normalize  $l$  and  $w$  to the optimal unconstrained buffer distance  $l_{crit}$  and buffer size  $w_{opt}$  as  $\alpha = l/l_{crit}(w, s)$  and  $\beta = w_b/w_{opt}(w, s)$ .

We show an example for  $\gamma_{ch} = 0.3$ . It means that the total area available for buffering in the channel is 30% of  $A_{ch}^*$ . We derive the normalized buffer area for a single wire,  $\gamma$ , from (13) and we evaluate the throughput using (14) in the space  $(w, s)$ . That is, we sample this space and for each sample we find the optimal buffering. This is fast, because the whole formulation is analytic. The result is in Figure 4.

As shown in Figure 4, the maximum throughput possible for a given area constraint is achieved for a wire topology which differs from the minimum one. Higher densities require an aggressive buffering to achieve low enough latency, which is not possible due to the area constraint. Also lower density solutions can not achieve a low enough latency to compensate for the reduced parallelism. Note that the highest throughput for 30% of the maximum area is only 13% below the absolute maximum throughput. It is obtained by using  $w = 0.8\mu m$  and  $s = 0.8\mu m$ , together with  $w_b = 35\mu m$  and  $l = 4.2mm$ .

The optimal normalized buffer size  $\beta$  (in dashed line) and buffer distance  $\alpha$  (in solid line) are presented as function of the wire topology in Figure 5. The values on the

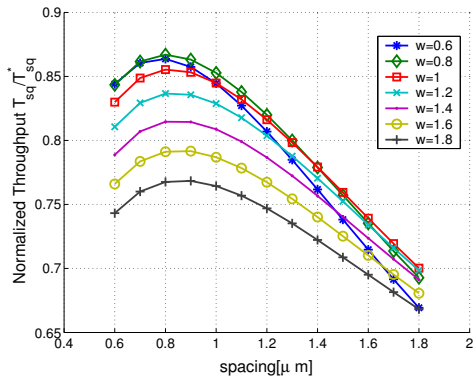


Fig. 4. Wire sizing for optimal area-constrained buffer insertion  $A_{ch}(w, s) = 0.3A_{ch}^*$ .

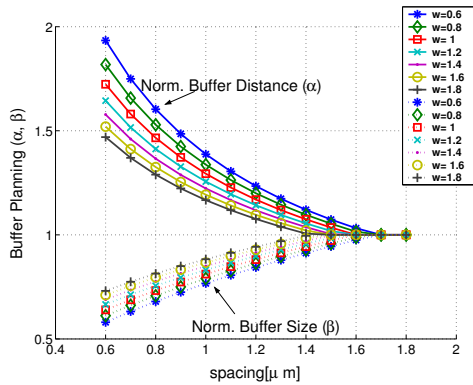


Fig. 5. Normalized buffer size and distance for optimal area-constrained buffer insertion  $A_{ch} = 0.3A_{ch}^*$ .

y-axis equal to 1 represents the area unconstrained solution [1]. For large spacing (larger than  $1.6\mu m$  in this example), the area constrained buffer insertion solution is more and more similar to the area unconstrained solution. This is because only few and relatively small buffers are necessary or in other terms the area constraints for those topologies become less stringent.

Figures 6 and 7 show contours of constant throughput as a function of the wire width and spacing for the case of  $A_{ch} = A_{ch}^*$  (unconstrained buffering) and for the case of  $A_{ch} = 0.3A_{ch}^*$  (constrained buffering), respectively. Note that for each  $w, s$  configuration shown, specific  $w_b$  and  $l$  have been computed that maximize the throughput for the given area constraint. Note that both graphs actually have a single optimal point from the perspective of throughput, as marked by the black square. However, these graphs illustrate what happens to the throughput if suboptimal  $w$  and  $s$  are chosen.

If we compare for example the curve at  $0.85T_{\square}^*$  of Figure 6 and 7, we notice that the same throughput requires higher density in the constrained case. This is to be expected considering the fact that higher density compen-

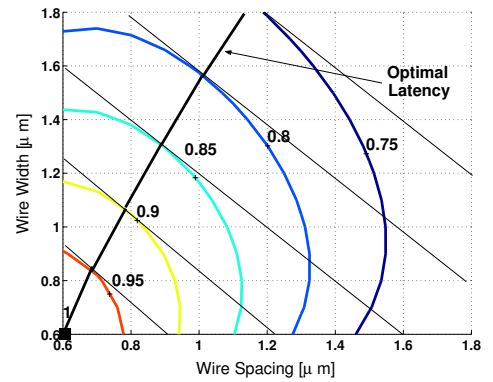


Fig. 6. Constant throughput contours for unconstrained buffer area ( $A_{ch}^*$ ).

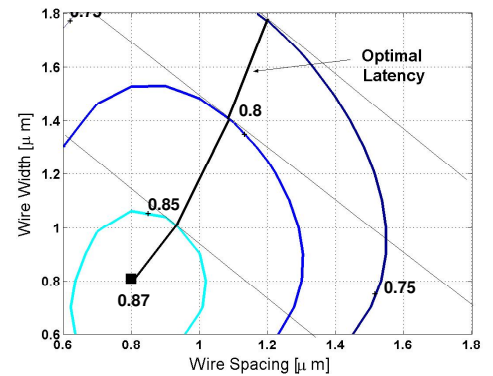


Fig. 7. Constant throughput contours for constrained buffer area ( $A_{ch} = 0.3A_{ch}^*$ ).

sates, in the throughput budget, for the increased latency of a single wire. The increased latency of a single wire is a result of the limited available buffer area.

While the methodology in this paper as developed thus far optimizes throughput, these graphs actually also offer a perspective on latency. In fact, the latency is what is actually optimized for a single wire. Now, as the wire spacing is increased, the coupling capacitance decreases and a better latency can be achieved for the same buffer area.

Thus, given a certain throughput contour, the same throughput is achieved with the lowest possible latency for the configuration with the lowest density. This configuration corresponds to the point where the constant contour curve is tangential to the lines with constant density. In the graph, these constant density are the diagonal  $-45^\circ$  lines, and the loci of optimal latency for a certain throughput are along the lines labeled "optimal latency". While not explicitly presented in this paper, the actual latency in each design point can easily be obtained from the model. Thus, the model also offers a solution for latency constrained designs.

In Figure 8, we plot the maximum throughput that is achievable using our model as a function of the normal-

ized buffer area constraint  $\gamma_{ch}$ . The figure shows two sets of points, for different switching factors. Note that the achievable throughput strongly depends on the switching factor since it relates to the capacitive coupling effect. Also note that with  $SF = 2$ , the buffer area maximum throughput is actually 80% larger than for the  $SF = 1$  case. However, for either switching factor, the throughput is only degrading relatively little if the buffer area becomes small. For example, for 80% reduction of the buffer area compared to the buffer area for maximum throughput, the performance is only reduced by less than 25% for both extreme values of the switching factor. Thus, the effectiveness of proper wire and repeater sizing is clearly demonstrated.

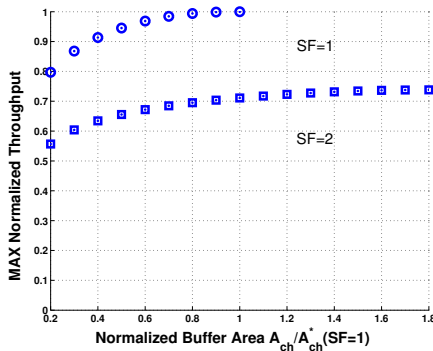


Fig. 8. Maximum throughput as a function of the buffer area constraint  $\gamma_{ch}$ .

Finally, Figure 9 is a companion figure to Figure 8. It again presents 2 sets of points for the two extreme switching factors. On the y-axis, the wire spacing and width are given that are found to be the optimal ones for the specific buffer area and can realize the throughput in Figure 8.

Thus, if such graphs are created once for a technology, possibly using unnormalized axis, they can be used for design. First, a suitable buffer area is selected that achieves the required throughput, by reading it off from the x-axis in Figure 8. Subsequently, this value can be used to get the

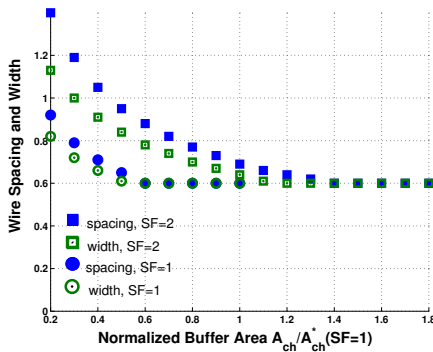


Fig. 9. Relative wire density as function of the buffer area constraint of the channel.

wire geometry from Figure 9. Finally, the right buffering is determined from [2] or from corresponding graphs that could be prepared but are omitted in this paper for brevity.

#### IV. SUMMARY

At the physical layer of Network-on-Chip (NoC) implementations, blocks and switches are already assigned and inter block connections have to be planned. Most of the inter block connections will consist of a large number of long parallel wires. Such wires usually have significant resistance and (coupling as well as self) capacitance.

This paper illustrates methodologies on how to perform buffering and wire sizing to obtain the maximum throughput. The results show that the appropriate wire density has to be chosen depending on the total buffer area available in the channel. High wire density is not beneficial if the buffer area is small. Low wire density on the other hand can result in a low latency but not in a high throughput. Our methodology can explore this trade-off and produce optimal buffering and wire sizing for a given area constraint.

#### REFERENCES

- [1] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley, 1990.
- [2] G.S. Garcea, N.P. van der Meijs and R.H.J.M. Otten. *Simultaneous analytic area and power optimization for repeater insertion*. in Proc. International Conf. on Computer Aided Design, Nov. 2003, pp. 568–573.
- [3] L. Benini and G. De Micheli, *Networks on chips: a new soc paradigm*, Computer, vol. 35, no. 1, pp. 70–78, Jan 2002.
- [4] A. Jantsch H. Tenhunen, Ed., *Networks on Chip*, Kluwer Academic Publishers, 2003.
- [5] S. Kumar, A. Jantsch, M. Soiminen, J.-P.; Forsell, M. Millberg, J. Oberg, K. Tiensyrja, and A. Hemani, *A network on chip architecture and design methodology*, in Proc. IEEE Computer Society Annual Symp. on VLSI, 2003, pp. 105–112.
- [6] J. Liu, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, *A global wire planning scheme for network-on-chip*, in Proc. of 2003 Int. Symp. on Circuits and Systems, ISCAS '03, May 2003, vol. 4, pp. 892–895.
- [7] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. Davis, *Repeater insertion and wire sizing optimization for throughput-centric VLSI global interconnects*, in Proc. International Conference on Computer Aided Design, Nov 2002, pp. 280–284.
- [8] Tao Lin and L.T. Pileggi, *Throughput-driven IC communication fabric synthesis*, in Proc. International Conference on Computer Aid Design, 2002, pp. 274–279.
- [9] D. Pamunuwa, Li-Rong Zheng and H. Tenhunen, *Maximizing throughput over parallel wire structures in the deep sub-micrometer regime*, Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 11, no. 2, pp. 224–243, April 2003.
- [10] BPTM web site:  
<http://www-device.eecs.berkeley.edu/~ptm/introduction.html>
- [11] A.B. Khang, S. Muddu and E. Sarto. *On switch factor based analysis of coupled RC interconnects*. in Proc. Design Automation Conference, pp. 79–84. Jun. 2000.