Available online at www.sciencedirect.com
SCIENCE DIRECT®
INTEGRATION, the VLSI journal 38 (2004) 19–42
INTEGRATION the VLSI journal
www.elsevier.com/locate/vlsi

Cost considerations in network on chip

Evgeny Bolotin*, Israel Cidon, Ran Ginosar, Avinoam Kolodny

Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa 32000, Israel

Received 1 August 2003; received in revised form 26 January 2004; accepted 19 March 2004

Abstract

Systems on Chip (SoCs) require efficient inter-module interconnection providing for the required communications at a low cost. We analyze the generic cost in area and power of Networks on Chip (NoCs) and alternative interconnect architectures: a shared bus, a segmented bus and a point-to-point interconnect. For each architecture we derive analytical expressions for area, power dissipation and operating frequency as well as asymptotic limits of these functions. The analysis quantifies the intuitive NoC scalability advantages.

Next we turn to NoC cost optimization. We explore cost tradeoffs between the number of buffers and the link speed. We use a reference architecture, termed QNoC (Quality-of-Service NoC), which is based on a grid of wormhole switches, shortest path routing and multiple QoS classes. Two traffic scenarios are considered, one dominated by short packets sensitive to queuing delays and the other dominated by large block-transfers. Our simulations show that network cost can be minimized while maintaining quality of service, by trading off buffers with links in the first scenario but not in the second.

Keywords: Network on chip; Scalable interconnect; Wormhole buffering; Cost minimization

1. Introduction

Networks on Chip (NoCs) can help to solve major design challenges of integrated Systems on Chip (SoCs) [1–12], including modularity and reuse, design productivity, global wire speed/power optimization, synchronization, and communication error recovery. However, since VLSI is

*Corresponding author. Tel.: +972-4829-4711; fax: 972-4829-5757.
E-mail address: bolotin@tx.technion.ac.il (E. Bolotin).

0167-9260/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.vlsi.2004.03.006

extremely cost-sensitive, the required communication Quality-of-Service (QoS) must be provided at a minimal cost [9,11,26,27]. QoS is associated primarily with latency and throughput. Cost is measured by chip-area and power dissipation. The purpose of this paper is to make a quantitative comparison between the generic cost of a NoC and the cost of other interconnection schemes, and to explore cost minimization options within a specific NoC architecture.

Traditional solutions for on-chip global communication include various shared-bus structures [13–15] and ad-hoc point-to-point interconnections. The lack of scalability of these approaches was qualitatively discussed in [2,7,10]. Advantages of spatial-reuse packet/wormhole switched networks were reported and explored in comparison with buses in [1,2,4,6,9,10]. However, no quantitative cost analysis has been conducted so far. This paper analyzes and quantifies the cost and performance advantages of a network based interconnection scheme over other interconnection alternatives for future SoCs. In particular, we analyze the area and power cost of a packet-switched NoC in comparison with non-segmented (shared) system bus (NS-Bus), segmented system bus (S-Bus) and point-to-point (PTP) interconnect. Assuming a given set of Quality-of-Service requirements we derive analytical expressions for the wire area, power and operating frequency of each interconnection scheme. With an increasing number of system modules, simple asymptotic limits of these expressions are derived. The results clearly quantify the scalability advantage of NoC over the traditional alternatives.

Switched networks and techniques for their design have been developed for computer networks and for multiprocessor systems, for example [16–22]. However, a unique set of resource constraints and design considerations exists for an on-chip environment. As described in [2,10], memory and computing resources are relatively more expensive on-chip, while relatively more wires are available. As a result, many NoC architectures are based on wormhole packet routing [1,2,7,8], since wormhole routing reduces latency and buffer requirements in the routers [2,22,23]. Thus, the area of a generic NoC can be approximated by the wiring area used for the NoC links. Shortest-path routing guarantees minimal wire length and power dissipation in the links.

Some studies investigated optimum wormhole buffering for increased router performance in general computer networks [28,29]. Performance-power cost tradeoff was explored by selecting appropriate packet size in [26]. Unlike computer networks which are built for on-going expansion, future growth and standards compatibility, on-chip networks can be designed and customized for an a priori known set of computing resources and pre-characterized traffic patterns among them. These imply that various design parameters of the network architecture such as buffer size and link bandwidth allocation can be designed for specific implementations in order to provide a required QoS for known traffic patterns. Moreover, one can apply a tradeoff between these parameters to achieve a more cost-effective NoC implementation at a given QoS specification.

Based on the above considerations, we present a NoC cost minimization process by exploring the influence of increasing the number of wormhole buffers versus decreasing link bandwidth (by reducing the number of wires). For this tradeoff study we use a specific NoC architecture termed QNoC (Quality-of-Service NoC) [1], which is based on a planar grid of switches that route the traffic according to a fixed shortest path (X–Y based) discipline. It uses input buffering scheme and employs multi-class wormhole forwarding to support multiple service priority classes. The optimization process attempts to reduce the cost while supporting the different QoS classes and the QoS requirements for each class. We study two different system traffic scenarios. The first scenario is dominated by short packets that are sensitive to queuing delays. The second scenario is

dominated by long block-transfers consisting of long packets. We show by simulations that in the first case NoC area cost minimization is achieved by adding wormhole buffers and decreasing link bandwidth up to an optimal value. However, this is not true in the block-transfer dominated traffic scenario where there is no cost advantage in increasing the number of buffers above the minimum. The total area cost is estimated by calculating total area occupied by wires, and adding to it the estimated area occupied by the packet switch logic (buffers, tables, etc.). The power cost is based on summation of the traffic that traverses each wire length and is received by input stages.

The rest of this paper is organized as follows: Section 2 describes the example QNoC architecture, Section 3 presents an analytical comparison between a generic NoC and alternative architectures, Section 4 presents QNoC cost minimization process and provides cost minimization examples for several system traffic scenarios along with simulation results, and finally Section 5 concludes.

2. QNoC architecture and design process

The QNoC architecture and design process were presented in [1]. In this section we first present a brief overview and then develop additional architecture details (Section 2.4). The QNoC architecture is based on a grid topology and wormhole packet routing. Links are assumed reliable[1] and backpressure is applied between stages resulting in a loss-less network. Packets traverse the network along the shortest route, thus minimizing power dissipation and maximizing network resource utilization.

2.1. QNoC topology

QNoC comprises routers interconnected by point-to-point links. Network topology can vary depending on system needs and module sizes and placement. Each system module is connected to a router (Fig. 1) via a standard interface, whose bandwidth is adapted to the communication needs of that module. The bandwidth of each inter-router link is similarly adjusted to accommodate the expected traffic and fulfill QoS requirements at the specific link. Link and interface bandwidth are adjustable by changing either the number of wires or the data frequency, or both. In addition, a module may be connected to the network through more than one interface.

Routing is performed over fixed shortest paths, employing a symmetric X–Y discipline whereby each packet is routed first in an "X" direction and then along the perpendicular dimension or vice versa[2]. Network traffic is thus distributed non-uniformly over the mesh links, but each link's bandwidth is adjusted to its expected load, achieving an approximately equal level of link utilization across the chip.

2.2. QNoC service levels

We identify four different types of communication requirements and define appropriate service levels (SL) to support them:

[1] Or made reliable using error correction.

[2] Simple "around the block" modification is employed where needed.

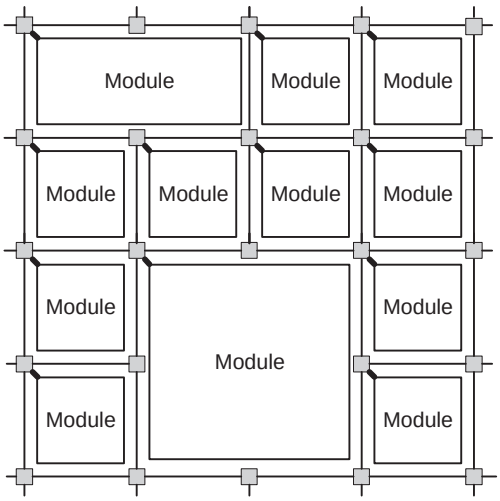


Fig. 1. QNoC custom topology example—irregular mesh.

Signaling covers urgent messages and very short packets that are given the highest priority in the network to assure shortest latency. This service level represents interrupts and control signals and alleviates the need for dedicated wires.

Real-Time service level guarantees bandwidth and latency to real-time applications, such as streamed audio and video processing. This service is packet based; a maximal level of guaranteed bandwidth is allocated to each real-time link and should not be violated.

Read/Write (RD/WR) service level provides bus semantics and is designed to support short memory and register accesses.

Block-Transfer service level is used for the transfer of long messages and blocks of data, such as cache refill and DMA transfers.

We establish a priority ranking, where Signaling is given the highest priority and Block-Transfer the lowest. QNoC employs preemptive communication scheduling where data of a higher priority packet is always transmitted before that of a lower service level (a round-robin is employed within service levels). Additional service levels may be defined if desired. For instance, the RD/WR service level may be split into normal and urgent RD/WR sub-levels.

2.3. QNoC communication

Packets carry routing information, command and payload. The command field identifies the payload, specifying the type of operation. The packet is divided into multiple flits following [22]. Flit transfer over the inter-router link is controlled by handshake.

2.4. QNoC routers

Routers connect to up to five links (Fig. 2), designed for planar interconnect to four mesh neighbors and to one SoC module. The router forwards packets from input to output ports. Every

arriving flit is first stored in an input buffer. On the first flit of a packet, the router invokes a routing algorithm to determine to which output port that packet is destined. The router then schedules the transmission for each flit on the appropriate output port.

The routing algorithm uses a simple routing function. For example, relative routing is employed for X–Y routing, leading to a minimal VLSI implementation. Routing information per each service level and per each input port is retained until the tail flit of the packet is delivered. When a flit is forwarded from an input to an output port, one buffer becomes available and a *buffer-credit* is sent back to the previous router on separate wires.

Each output port of a router is connected to an input port of a next router via a communication link. The output port maintains the number of available flit slots per each service level in the buffer of the next input port. The number is decremented upon transmitting a flit and incremented upon receiving a buffer-credit from the next router. When a space is available, the output port schedules transmission of flits that are buffered at the input ports and waiting for transmission through that output port, as detailed below.

Flits are buffered at the input ports, awaiting transmission by the output ports (Fig. 3). There are separate buffers for each of the four service levels (''direct buffer mapping''). Relatively small buffers are allocated to each service level, capable of storing only a few flits. For example, a buffer capable of storing four flits is the minimum required to avoid stalls in the wormhole pipeline caused by waiting for buffer credits from the next node. This number is calculated using the following considerations: One cycle is required for transmitting the flit, one cycle for latching incoming flit and routing decision in the router, one cycle for the transmission delay of credit-buffer information from the next router and an additional cycle for latching the credit-buffer information in the scheduling logic of the output port, see Fig. 4.

Each output port schedules transmission of flits according to the availability of buffers in the next router and the service level priority of the pending flits. A packet based round-robin arbitration is performed on input ports within the same service level. This scheduling discipline implies that a particular flit gets transmitted on an output port as long as there is buffer space available on the next router and there is no packet with a higher priority pending for that particular output port. Once a higher priority packet appears on one of the input ports, transmission of the current packet is preempted and the higher priority packet gets through. Transmission of the lower priority packet is resumed only after all higher priority packets have been serviced.

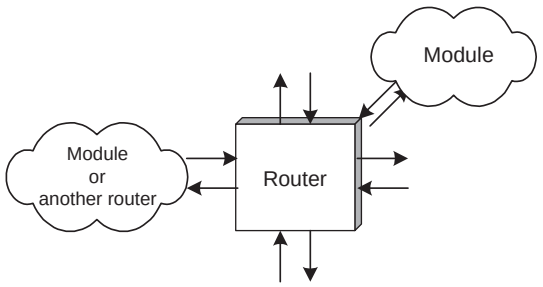


Fig. 2. The router has up to five links and may connect to neighbor mesh routers or to chip modules.

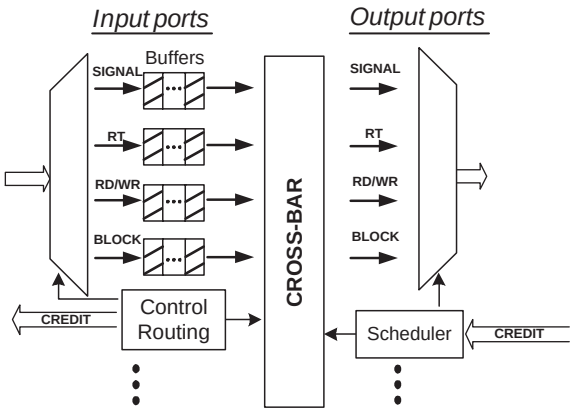


Fig. 3. QNoC router architecture.

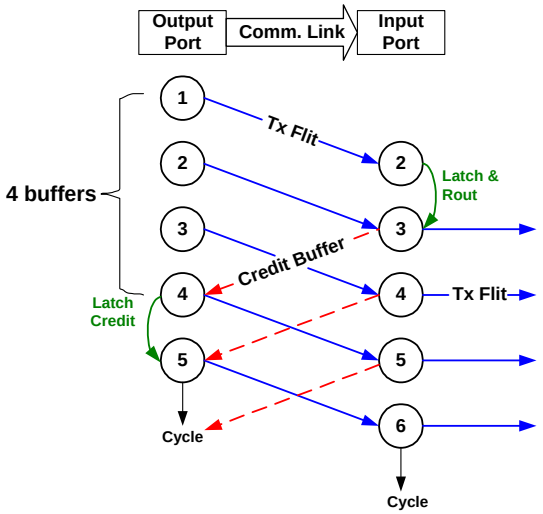


Fig. 4. QNoC transmission time-diagram demonstrating minimal buffering requirements preventing bubbles in a wormhole pipeline.

2.5. QNoC design process

The QNoC design process [1] is employed to construct a specific cost-effective QNoC based on the general architecture described above. It characterizes and verifies the inter-module traffic, places the modules on a generic network grid so as to minimize spatial traffic density, and optimizes the grid by trimming links, routers and buffers while maintaining the required QoS. The layout of the network is customized and bandwidth is allocated to links according to their relative load so that the utilization of links in the network is balanced and QNoC cost is reduced.

3. Cost of NoC versus other interconnection architectures

In this section we compare hardware and power costs of the most common on-chip communication architectures: NoC, a Non-Segmented Bus (NS-Bus), a Segmented Bus (S-Bus) and a direct Point-To-Point (PTP) interconnect, and explore the effect of an increasing number of system modules on the cost of each interconnection scheme. We consider an n-module SoC. The area of each module is $d \times d$, and they are arranged in a regular mesh (Fig. 5). We assume a uniform traffic distribution among the modules. Load capacitance of the interconnection architecture is assumed to depend only on the link length (neglecting the capacitance of module input ports). We derive analytical expressions for area, power and operating frequency of each interconnection scheme, and assuming fixed QoS we compare the cost as the number of system modules increases.

We define QoS as the throughput and end-to-end (ETE) delay provided by the interconnection architecture. Throughput depends on the level of parallelism available in the architecture and the bandwidth of the interconnecting links. ETE delay can be tuned by increasing or reducing the link bandwidth through changing link width or frequency. Such variations in link bandwidth for given throughput and ETE delay are reflected in link utilization. For example, an architecture designed for a given set of source rates (throughput), whose link bandwidth is increased in order to meet a stricter ETE requirement, will demonstrate a lower link utilization. In order to compare different architectures that provide the same QoS, we define an Effective Bandwidth as the actual communication bandwidth or throughput carried by the given architecture (*arch*), given that the link bandwidth is already adjusted to provide ETE delay requirements:

$$BW_{\text{eff, arch}} \triangleq \frac{U_{\text{arch}} \sum_{i \in \{\text{Arch. links}\}} w(i) f(i)}{Av\ Dist_{\text{arch}}}, \tag{3.1}$$

where "arch" is the interconnection architecture, such as NoC, NS-Bus, etc., U_{arch} is link utilization, $w(i)$ is number of wires in link i, $f(i)$ is its frequency and $Av\ Dist_{\text{arch}}$ is the average number of hops between any two interconnected modules.

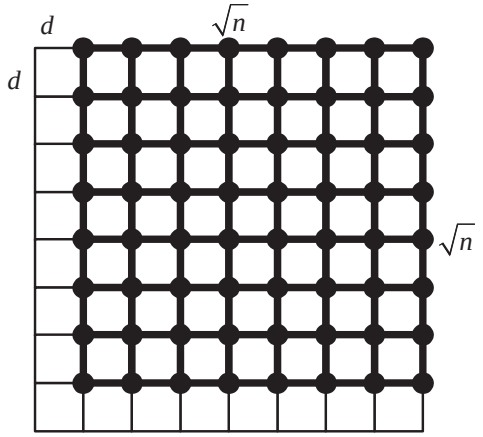


Fig. 5. NoC-interconnecting n communicating modules arranged in regular mesh, size of each is $d \times d$.

We first consider a given NoC and calculate its cost. Other architectures having the same effective bandwidth are then examined, and their cost functions are computed. Cost is estimated by analyzing the total wire length of each architecture. For the sake of simplicity, we assume that a single metal layer is used for links in all architectures.[3]

$$TL_{\text{arch}} = \sum_{i \in \{\text{Arch. links}\}} w(i)l(i), \tag{3.2}$$

where $w(i)$ is number of wires in link i and $l(i)$ is its length. The area cost of an architecture is

$$A_{\text{arch}} = W_{\text{p}} TL_{\text{arch}}, \tag{3.3}$$

where W_{P} is the global wire pitch (constant for a given technology). Total load capacitance is

$$C_{\text{L,arch}} = C_0 TL_{\text{arch}}, \tag{3.4}$$

where C_0 is wire capacitance per unit length. The delay over a link is estimated by the wire-delay model, $T = \delta R_{\text{link}} C_{\text{link}}$. Here, δ is the Elmore delay coefficient, $R_{\text{link}} = R_0 L_{\text{link}}$ (R_0 is wire resistance per unit length, L_{link} is the link length) and $C_{\text{link}} = C_0 L_{\text{link}}$. Thus, the switching frequency can be derived as

$$f_{\text{arch}} = \frac{1}{T} = \frac{1}{\delta R_0 C_0 L_{\text{link}}^2}. \tag{3.5}$$

The power cost function is calculated assuming that the dynamic power consumed by wires is proportional to the wire length and thus the wire length is a good estimator of power dissipated on wires. Dynamic power dissipation in switching circuits is:

$$P_{\text{arch}} = C_{L,\text{arch}} V_{dd}^2 f_{\text{arch}} U_{\text{arch}}, \tag{3.6}$$

where $C_{L,\text{arch}}$ is the total load capacitance, V_{dd} is the supply voltage, $f_{\text{arch}}$is the switching frequency and U_{arch} is link utilization, which serves as an activity factor for the links. Thus, the switching frequency of a link is its frequency multiplied by the link utilization. C_L, the total load capacitance, consists of link capacitance (C_{link}) and gate capacitance of the transistors driven by that link (C_{gate}). We assume that C_{gate} can be neglected and the dominant factor is C_{link}.

In the following sections we derive the explicit cost functions for each of the alternative architectures; the results are summarized in Section 3.5 below.

3.1. NoC cost functions

Consider n system modules interconnected by a NoC (Fig. 5). Each module is connected to a router using a standard interface, and the routers are interconnected in a mesh topology. For the NoC case, we assume that the silicon cost of minimal buffer routers and simple module interfaces are comparable to similar costs of other solutions (such as bus multiplexers, bus interfaces, etc.). Moreover, these costs are linear with the number of modules and therefore do not change the asymptotic comparison. The length of each inter-router link is $L_{\text{link}} = d$. Assuming that the number of wires in each link has been adjusted in order to equalize the expected utilization of all links [1], we define $\bar{w}$, the average number of wires in each link. The total link length in the NoC is

[3]Additional metal layers can easily be accounted for and would not change the asymptotic results.

$L_{\mathrm{NoC}} = 2d\sqrt{n}(\sqrt{n}-1)$, and the total wire length of the NoC is

$$TL_{\mathrm{NoC}} = 2d\bar{w}\sqrt{n}(\sqrt{n}-1). \tag{3.7}$$

Combining Eqs. (3.3) and (3.7), the NoC wiring area is

$$A_{\mathrm{NoC}} = 2W_p d\bar{w}\sqrt{n}(\sqrt{n}-1). \tag{3.8}$$

The effective bandwidth of the NoC is

$$BW_{\mathrm{eff,NoC}} = \frac{\sum w(i)f(i)U_{\mathrm{noc}}}{Av\ Dist_{\mathrm{NoC}}}. \tag{3.9}$$

$A_v Dist_{\mathrm{NoC}}$ is the average distance between every two nodes in the mesh and equals $(2/3)\sqrt{n}$ [24], leading to the following result for NoC:

$$BW_{\mathrm{eff,NoC}} = 3\bar{w}(\sqrt{n}-1)f_{\mathrm{NoC}}U_{\mathrm{NoC}}. \tag{3.10}$$

Eq. (3.10) reflects the actual bandwidth carried by the NoC. Note that it is directly proportional to link width, link utilization and frequency. For instance, if link width is increased (in order to reduce ETE delay) while frequency and total bandwidth are fixed then the link utilization is consequently reduced. The total load capacitance of a NoC is calculated using Eqs.(3.4) and (3.7),

$$C_{\mathrm{L,NoC}} = C_0 2d\bar{w}\sqrt{n}(\sqrt{n}-1). \tag{3.11}$$

The NoC operating frequency is computed using Eq. (3.5),

$$f_{\mathrm{NoC}} = \frac{1}{\delta R_0 C_0 d^2}. \tag{3.12}$$

Substituting the above results into Eq. (3.6), leads to NoC power dissipation

$$P_{\mathrm{NoC}} = \frac{P_0 2\bar{w}U_{\mathrm{NoC}}}{\delta}\sqrt{n}(\sqrt{n}-1), \tag{3.13}$$

where $P_0 \triangleq V_{dd}^2/R_0 d$. In conclusion, asymptotic power and area cost functions for NoC (including the cost incurred by the routers) are both O(n).

3.2. Non segmented BUS (NS-bus) cost functions

The NS-Bus is a simple shared bus, connecting all modules in the system and laid out as a minimal spanning tree (Fig. 6). It consists of a single segment and has no parallelism (only one transaction is active at a time). The total length of such a bus is $L_{\mathrm{NS-Bus}} = (1/2)d(n-4)$.

The NS-Bus effective bandwidth, following Eq. (3.1), is:

$$BW_{\mathrm{eff,NS-Bus}} = W_{\mathrm{NS-Bus}} f_{\mathrm{NS-Bus}} U_{\mathrm{NS-Bus}}. \tag{3.14}$$

The operating frequency is calculated using Eq. (3.5),

$$f_{\mathrm{NS-Bus}} = \frac{4}{\delta R_0 C_0 d^2 (n-4)^2}. \tag{3.15}$$

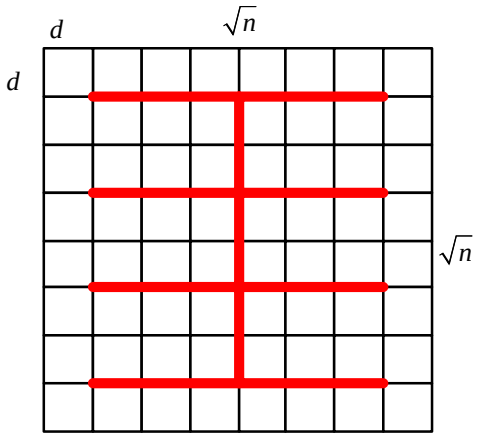


Fig. 6. Non-Segmented Bus-interconnecting n communicating modules arranged in regular mesh, size of each is $d \times d$.

We obtain the width of NS-Bus by equating the effective bandwidth of the NS-Bus with that of the NoC

$$W_{\mathrm{NS-Bus}} = \left\lceil \frac{3\bar{w}U_{\mathrm{NoC}}(\sqrt{n}-1)(n-4)^2}{4U_{\mathrm{NS-Bus}}} \right\rceil. \tag{3.16}$$

Note that the NS-bus requires an excessive bus width of $O(n^2\sqrt{n})$ in order to compensate for the lack of parallelism and for the low operating frequency due to its larger load capacitance. The Total wire Length of the NS-Bus is thus

$$TL_{\mathrm{NS-Bus}} = \frac{3d\bar{w}}{8}\frac{U_{\mathrm{NoC}}}{U_{\mathrm{NS-Bus}}}(\sqrt{n}-1)(n-4)^3 \tag{3.17}$$

and the NS-Bus area is

$$A_{\mathrm{NS-Bus}} = \frac{3W_p d\bar{w}}{8}\frac{U_{\mathrm{NoC}}}{U_{\mathrm{NS-Bus}}}(\sqrt{n}-1)(n-4)^3. \tag{3.18}$$

Using the same method as in the previous section, and applying total wire length and frequency of the NS-Bus, we compute the average dynamic power dissipated in this architecture following Eq. (3.6):

$$P_{\mathrm{NS-Bus}} = \frac{3P_0\bar{w}U_{\mathrm{NoC}}}{2\delta}(\sqrt{n}-1)(n-4). \tag{3.19}$$

The asymptotic area of the NS-Bus is of $O(n^{3.5})$ while its asymptotic power is of $O(n^{1.5})$.

3.3. Segmented BUS (S-bus) cost functions

The S-Bus is the most common SoC interconnection architecture, since a long shared bus that interconnects all system modules is not feasible in systems consisting of many communicating nodes (as can also be deduced from the results of the previous section). We assume that S-Bus has

the same topology as NS-Bus, but it is segmented into $\sqrt{n}/2$ identical sections (of the same length, width and frequency) interconnected by bridges, as in Fig. 7. The S-Bus has more parallelism, and the capacitance of each segment is substantially reduced relative to that of the NS-Bus, allowing the S-Bus to operate at higher frequencies. This structure can also be interpreted as a step in the evolution from shared-bus architectures towards networked system interconnect.

The total length of the S-Bus is the same as that of the NS-Bus: $L_{\mathrm{S-Bus}} = L_{\mathrm{NS-Bus}} = (1/2)d(n-4)$. As in the previous section, we calculate bus width by equating the effective bandwidth

$$BW_{\mathrm{eff,S-Bus}} = \frac{W_{\mathrm{S-Bus}} f_{\mathrm{S-Bus}} U_{\mathrm{S-Bus}}(\#\ segments)}{Av\ Dist_{S-\mathrm{Bus}}} = BW_{\mathrm{eff,NoC}},$$

where $Av\ Dist_{\mathrm{S-Bus}} = Av\ Dist_{\mathrm{1D-array}} = (k+1)/3$ (see the Appendix). The operating frequency and total wire length are

$$f_{\mathrm{S-Bus}} = \frac{1}{\delta R_0 C_0 d^2 n}, \tag{3.20}$$

$$TL_{\mathrm{S-Bus}} = \frac{\bar{w}d}{2} \frac{U_{\mathrm{NoC}}}{U_{\mathrm{S-Bus}}} \sqrt{n}(n-4)(\sqrt{n}-1)(\sqrt{n}-2). \tag{3.21}$$

Thus, the S-Bus area cost function is

$$A_{\mathrm{S-Bus}} = \frac{W_p d\bar{w}}{2} \frac{U_{\mathrm{NoC}}}{U_{\mathrm{S-Bus}}} \sqrt{n}(n-4)\left(\sqrt{n}-1\right)\left(\sqrt{n}+2\right). \tag{3.22}$$

As in the previous sections, we used the total wire length to estimate the total load capacitance, leading to the power dissipation of the S-Bus

$$P_{\mathrm{S-Bus}} = \frac{P_0 \bar{w} U_{\mathrm{NoC}}}{2\delta} \frac{(n-4)\left(\sqrt{n}-1\right)\left(\sqrt{n}+2\right)}{\sqrt{n}}. \tag{3.23}$$

In summary, the asymptotic area of the S-Bus is of $O(n^{2.5})$ and its asymptotic power is of $O(n^{1.5})$.

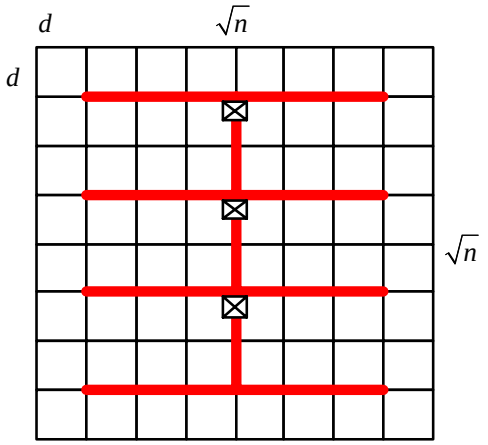


Fig. 7. Bus segmented into $\sqrt{n}/2$ segments-interconnecting n communicating modules.

3.4. Point-To-Point (PTP) cost functions

Consider n modules arranged in a mesh and interconnected point-to-point with links that are routed in an x–y fashion, similar to the NoC of Fig. 5. The total length of all PTP interconnects is $L_{\text{ptp}} = (1/3)dn\sqrt{n}(n-1)$, assuming n traffic sources having $(n-1)$ destinations each, multiplied by the average distance in a mesh and divided by two (in previous sections only one direction of communication was considered). As in the previous sections, effective bandwidths is equated to that of the NoC

$$BW_{\text{eff,ptp}} = \frac{W_{\text{ptp}} f_{\text{ptp}} n(n-1) U_{\text{ptp}}}{2} = BW_{\text{eff,NoC}}$$

leading to PTP average frequency and the width of each PTP link

$$f_{\text{PtP}} = \frac{9}{\delta R_0 C_0 4d^2} \frac{1}{n}, \tag{3.24}$$

$$W_{\text{ptp}} = \left\lceil \frac{8}{3} \frac{\bar{w} U_{\text{NoC}}(\sqrt{n}-1)}{U_{\text{ptp}}(n-1)} \right\rceil. \tag{3.25}$$

However, note that the obtained PTP link width in Eq. (3.25) is $\mathrm{O}\left(1/\sqrt{n}\right)$, therefore for n large enough

$$W_{\text{ptp}} = 1. \tag{3.26}$$

Since we are interested in asymptotic cost functions, we assume that $W_{\text{ptp}} = 1$, and hence the total wire length of the PTP interconnect architecture becomes

$$TL_{\text{ptp}} = \frac{d}{3} n\sqrt{n}(n-1) \tag{3.27}$$

PTP area is thus

$$A_{\text{ptp}} = \frac{dW_{\text{p}}}{3} n\sqrt{n}(n-1) \tag{3.28}$$

Proceeding as above,

$$P_{\text{ptp}} = \frac{3P_0 U_{\text{ptp}}}{4\delta} \sqrt{n}(n-1). \tag{3.29}$$

Generally one could expect that the power dissipation of PTP should be similar to the NoC power dissipation, since the same communication traffic travels along the same distances. However, since the minimal width of PTP link when the number of nodes grows is one, the capacitance and consequently the power dissipation of PTP becomes higher than in NoC by a factor $\sqrt{n}$:

$$\frac{P_{\text{ptp}}}{P_{\text{NoC}}} = \frac{3U_{\text{ptp}}}{8\bar{w}U_{\text{NoC}}} (\sqrt{n}+1). \tag{3.30}$$

In other words, with growing n and uniform traffic distribution, the communication between each pair of nodes decreases, but the link in the PTP architecture cannot benefit from it since it reaches a minimal link width. The NoC architecture, on the other hand, can benefit from it by

sharing traffic of many sources over the same links. This phenomenon becomes even stronger in non-uniform (and more realistic) traffic scenarios with higher traffic locality, where less traffic traverses long distances and PTP interconnect wastes more power than the NoC.

3.5. Summary and comparison of cost functions

The preceding sections are summarized in Table 1 below.

Asymptotic cost functions are presented in Table 2. It can be observed that networked interconnection architecture requires less wiring area, dissipates less power and therefore is preferable to other architectures. From these results one can also observe the evolution of shared bus interconnection systems towards networked architectures. Clearly, NS-Bus architectures become infeasible with a growing n. S-Bus shows better performance and lower cost relative to NS-Bus, and NoC demonstrates a pronounced superiority over the other architectures from both performance and cost points of view. In our model, PTP interconnect cost and performance are similar to those of the S-Bus, due to the assumption that capacitance depends only on wire length and neglecting module port capacitance. Non-scalability of PTP becomes evident when module ports are also considered: PTP requires a port for each connection, resulting in $O(n)$ ports for each module.

Let's summarize and intuitively explain the results in Table 2:

Frequency: NoC operating frequency is O(1) thanks to utilizing short links of constant length, independent of n. The frequency of the NS-Bus decreases as $O(n^2)$ because its length grows as $O(n)$ and therefore resistance (R) and capacitance (C) grow as $O(n)$ each. In the S-Bus the length of each segment grows as $O(\sqrt{n})$, and therefore RC delay grows only as $O(n)$. We assumed that the PTP links are asynchronous and can operate at different frequencies (shorter links can operate faster than longer ones). On average, PTP link length grows as $O(\sqrt{n})$ and its RC delay grows as $O(n)$.

Table 1
Cost functions and operating frequencies for uniform traffic

	Total area	*Power dissipation*	*Operating frequency*
NS-Bus	$\frac{3W_p d\bar{w}}{8}\frac{U_{\mathrm{NoC}}}{U_{\mathrm{NS-Bus}}}(\sqrt{n}-1)(n-4)^3$ (3.18)	$\frac{3P_0\bar{w}U_{\mathrm{NoC}}}{2\delta}(\sqrt{n}-1)(n-4)$ (3.19)	$\frac{4}{\delta R_0 C_0 d^2}\frac{1}{(n-4)^2}$ (3.15)
S-Bus	$\frac{W_p d\bar{w}}{2}\frac{U_{\mathrm{NoC}}}{U_{\mathrm{S-Bus}}}\sqrt{n}(n-4)(\sqrt{n}-1)(\sqrt{n}+2)$ (3.22)	$\frac{P_0\bar{w}U_{\mathrm{NoC}}}{2\delta}\frac{(n-4)(\sqrt{n}-1)(\sqrt{n}+2)}{\sqrt{n}}$ (3.23)	$\frac{1}{\delta R_0 C_0 d^2}\frac{1}{n}$ (3.20)
NoC	$2W_p d\bar{w}\sqrt{n}(\sqrt{n}-1)$ (3.8)	$\frac{2P_0\bar{w}U_{\mathrm{NoC}}}{\delta}\sqrt{n}(\sqrt{n}-1)$ (3.13)	$\frac{1}{\delta R_0 C_0 d^2}$ (3.12)
PTP	$\frac{dW_p}{3}n\sqrt{n}(n-1)$ (3.28)	$\frac{3P_0U_{\mathrm{ptp}}}{4\delta}\sqrt{n}(n-1)$ (3.29)	$\frac{9}{4\delta R_0 C_0 d^2}\frac{1}{n}$ (3.24)

Table 2
Asymptotic cost functions

Arch	*Total area*	*Power dissipation*	*Operating frequency*
NS-Bus	$O(n^3\sqrt{n})$	$O(n\sqrt{n})$	$O(\frac{1}{n^2})$
S-Bus	$O(n^2\sqrt{n})$	$O(n\sqrt{n})$	$O(\frac{1}{n})$
NoC	$O(n)$	$O(n)$	$O(1)$
PTP	$O(n^2\sqrt{n})$	$O(n\sqrt{n})$	$O(\frac{1}{n})$

The total area: Since the NS-bus operates at a very slow frequency (decreasing as $O(1/n^2)$) and has no parallelism, it has to be made excessively wide in order to provide the same effective bandwidth as the NoC. As a result, its width grows as $O(n^2\sqrt{n})$ and its length grows as $O(n)$, so that its total area cost function grows as $O(n^3\sqrt{n})$. The S-bus is $O(n)$ faster than the NS-Bus because each segment is $O(\sqrt{n})$ shorter and it employs $O(\sqrt{n})$ segments in parallel, but since the average number of hops traversed on the segmented bus is also $O(\sqrt{n})$, it results in no parallelism. Thus, the S-bus requires $O(n)$ fewer links than the NS-bus and its total area cost function is $O(n^2\sqrt{n})$. The NoC wire-cost increases only as $O(n)$. In PTP the average link frequency is $O(n)$ slower than in the NoC (longer links with higher capacitance). The, link length grows as $O(n^2\sqrt{n})$ and since the link width is asymptotically one, its total area also grows as $O(n^2\sqrt{n})$.

The power dissipation cost function: Power dissipated by all architectures is proportional to the product of operating frequency and total wire length.

In this section we analyzed area and power cost functions of interconnection architectures assuming a given technology. We showed the advantage of NoC, assuming a uniform traffic distribution and also assuming that load capacitance depends only on the interconnect (ignoring the capacitance of system module ports). Moreover it is clear that non-uniform, mostly-local traffic favors NoC, as does the inclusion of input port capacitance. In more advanced VLSI technology generations the capacitance and delay of long interconnect wires becomes even more dominant. As the technology improves, NoC is the only communication architecture where the links become shorter and less vulnerable to delays and noise. With a growing number of system modules, for a given die size (DxD) the link length of NoC is $D/\sqrt{n}$ (decreasing as $O(\sqrt{n})$), the link length of the NS-Bus is $D(n-4)/(2\sqrt{n})$ (growing as $O(\sqrt{n})$), and the link lengths of the S-Bus and PTP are $\sim D$ and $2D/3$, respectively (independent of n). As a result, the cost and performance advantages of NoC will become even more pronounced in future technology generations.

4. Cost minimization in QNoC by trading off link-bandwith and buffer-space

In the previous section we quantified the scalability of NoC as a communication architecture for future SoCs in terms of the cost of power and wiring-area. When constructing a NoC for a specific application, the system architect can use a design process presented in [1], which characterizes and verifies the inter-module traffic, places the modules so as to minimize the system spatial traffic density on a generic network grid, then the layout of the network is customized and

bandwidth is allocated to links according to their relative load so that the utilization of links in the network is balanced and cost is minimized. Further improvements can be made in order to minimize cost, while preserving the required QoS. In particular, in this section we refine our cost model by adding also buffer-space when considering the total cost of NoC, and explore the tradeoff between increasing the wormhole buffer space in routers and decreasing the link bandwidth. Thus, we increase the utilization of the network links and may still maintain the required QoS in terms of ETE delay due to the contribution of additional buffers that resolve contentions inside the network.

As an example for such a tradeoff we use the QNoC architecture and service-level communication model described in Section 2. We simulate various communication traffic scenarios and extract the possible buffer-link tradeoff curve. For each traffic scenario, communication traffic is fixed and several sets of different network buffer and link bandwidth allocations are simulated. The output of each simulation is packets ETE delay for each service level. Different resource (buffers and link bandwidth) allocations result in different silicon area cost. Only allocations providing adequate QoS are considered.

4.1. QNoC cost minimization process

During the QNoC optimization steps we aim to minimize the cost in power and silicon area of the resulting QNoC. Detailed area cost is calculated considering both wiring and logic gates/buffers costs.

Wire cost: Since the distance between two adjacent wires is fixed, the area occupied by link wires on a chip is proportional to the total wire length. For the sake of simplicity we assume one metal layer[4] and estimate the total area occupied by the network by calculating total wire length of the network links and using Eq. (3.3).

Logic cost: QNoC logic consists of the routers and network interfaces of system modules. The cost of a router depends on several parameters: the number of ports ($\#Port$), number of service levels ($\#SL$), flit size ($FlitSize$) and buffer size for each service level ($BufSize$). We give an estimate for the cost of the router in the architecture that was presented in Section 2.4. Our experiments show that the buffers dominate the area of the router. The total number of flip-flops ($\#FF_i$) in a router include buffer storage and control memory [1],

$$\#FF_i = \#Port \cdot \#SL \cdot \left[(FlitSize + 2)BufSize + \log_2\left(BufSize(\#Port)^2\right)\right]. \tag{4.1}$$

Since the cost of network interfaces is constant and has no influence on the optimization process, the total logic area of a QNoC is the sum of all routers:

$$\text{logic-area} \sim FF_{\text{a}} \sum_{i \in \{Routers\}} \#FF_i, \tag{4.2}$$

FF_{a} is the area occupied by a single flip-flop.

[4]Assuming multiple metal layers would not change the generic conclusions we made in this section regarding buffering strategy in networks on chip.

We assume that power is a function of the rate of transmitted information and the number of hops that it traverses until it reaches destination. Thus, we can neglect the effect that increasing buffer space might have on power.[5]

We start from a network designed with the minimal number of buffers as described in Section 2.4 and apply an area cost minimization process to it. During optimization, link bandwidth (wire cost) is decreased and buffer space (logic cost) of the routers is increased. As the link bandwidth decreases, network performance drops and packet ETE delays grow. Queuing delays in a wormhole system imply that there are blocked worms in the network. Hence, increasing buffer space can free up the system and restore the required ETE delay. Naturally, only buffer and link bandwidth allocations that provide the required QoS in terms of ETE delay are considered. The total change in area (ΔArea) is then calculated. When the obtained ΔArea is negative it means that total area cost is being reduced.

QNoC architecture uses dedicated buffers for each service level, with a preemptive inter-service level priority mechanism. As a result, the delays of the highest priority packets are not affected by the load and delays of lower priority packets. Therefore, the optimization process starts from the highest priority service level, calculates the optimum buffer space and link bandwidth allocation for it, then the number of buffers at this service-level is fixed and optimization is performed for the next lower priority service-level, and so on. Since bandwidth reduction may adversely affect ETE delay, the process may have to back-track and reiterate, until all communication requirements in all service levels are met.

4.2. QNoC-design optimization examples

We present several QNoC design optimization examples. We make a distinction between system traffic scenarios dominated by many short packets that are sensitive to queuing delays, which are termed RD/WR dominated scenarios, and Block-Transfer dominated scenarios consisting of very long packets.

In our design examples we consider a system with 16 communicating modules interconnected by a QNoC arranged in a 4×4 mesh and designed using the process described in [1]. Links operate at a frequency of 1GHz (one nanosecond cycle) and the width of each link is calibrated and tuned during the design process. We assume a uniform traffic distribution among the modules. Each module contains several traffic sources that correspond to the different classes of system traffic: Signaling, Real-Time, RD/WR and Block-Transfer. Each source creates packets with a specific distribution of packet size and inter-arrival time [1]. OPNET [30] as chosen as our simulation framework. The initial QNoC is designed using a minimal buffer size of four 16-bit flits for each service level and for each input port. We assumed 0.13 μm process technology; the area occupied by one flip-flop is $FF_a = 36\,\mu m^2$ and global wire pitch is $W_p = 670$ nm, according to the ITRS [25]. Let's consider the two scenarios.

4.2.1. RD/WR dominated traffic scenario

In this scenario, communication traffic consists of only three service levels: Signaling, Real-time and RD/WR, and it is dominated by RD/WR packets that are relatively short and abundant. We

[5]When exact calculations are performed the crossbar and links parallelizer/serializer circuitry area and power costs should also be included in the metrics.

consider two design examples and try to minimize the hardware cost of the QNoC by adding buffers and cutting down the link bandwidth.

4.2.1.1. Low - utilization network (severe latency requirements). The first example considers a lightly loaded network designed to operate at a low link utilization in order to meet stringent latency requirements. Each module contains three traffic sources, one for each service level. Source rates and QoS requirements are summarized in Table 3.

The initial QNoC satisfies the QoS requirements of this example with a total link bandwidth of 853 Gbps and total wire length of 2.56 m. The total QNoC area (wires and routers) is 2.26 mm^2. We start the optimization process by adding buffers for Signaling packets and trying to reduce link bandwidth. Signaling traffic consists of very short packets and has the highest priority in the network, preempting all lower priority packets. In that way, Signaling packets experience an extremely under-utilized network and consequently they do not experience any significant queuing delays. Obviously, no buffer increase can improve performance of Signaling packets. Real-time traffic in our example uses longer packets, but total available network bandwidth is still very high, so it experiences an under-utilized network. Reduction of only 2% of network bandwidth (by removing link wires) required an increase of Real-time buffers from four to seven flits, which resulted in the increase of total area (Table 4). In RD/WR traffic, on the other hand, the optimization resulted in area reduction (Table 4). The minimum value (Fig. 8) was achieved when network bandwidth was reduced to 90% of the original while adding only one buffer to the RD/WR service level. This optimization reduced the area by 0.13 mm^2, which is 5.7% total QNoC area saving. Further increasing the buffer space provides a diminishing return, as clearly shown by the growing ΔArea function (Fig. 8). Note that, network bandwidth drop of 10% (in our example) that provides optimum tradeoff for the RD/WR service level, inevitably results in performance degradation of Signaling and Real-time packets at the same percentage (10% increase of ETE delay). But since this reduced performance still satisfies the initial QoS requirements, it is still acceptable, see Fig. 9.

4.2.1.2. High-utilization network (moderate latency requirements). In this example we check what are reduction can be achieved when our optimization process is applied to network with a

Table 3
Each module source rate and QoS requirements—Low-utilization, RD/WR dominated scenario

Service level	Traffic interpretation	Average packet length (flits)	Average inter-arrival time (ns)	Total load	Max ETE delay requirements (for 99.9% of packets)
Signaling	Each module sends interrupt to a random target every 100 cycles	2	100	320 Mbps	20 ns (Several cycles)
Real-time	Periodic real-time connections from each module to all others	40	2000	320 Mbps	500 ns (Hundreds of cycles)
RD/WR	A random target RD/WR transaction every ~25 cycles.	4	25	2.56 Gbps	100 ns (Tens of cycles)

Table 4
Optimization steps at each service level (the optimum point is indicated in *italics*)

Signaling	BufSize	4	No possible optimization		
	Network BW (%)	100			
	Delta Area [mm^2]	0			
Real-time	BufSize	4	7	No possible optimization	
	Network BW (%)	100	98		
	Delta Area [mm^2]	0	0.09		
RD/WR	BufSize	4	*5*	6	8
	Network BW (%)	100	*90*	88	85
	Delta Area [mm^2]	0	*−0.13*	−0.12	−0.09

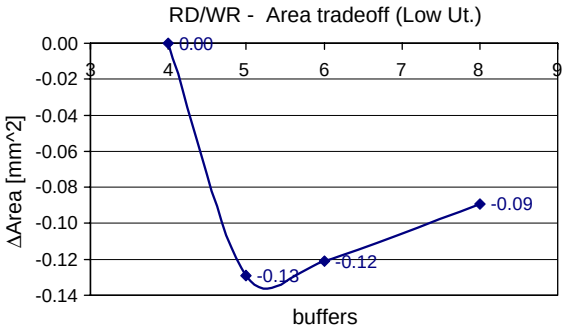


Fig. 8. ΔArea—Optimization performed on RD/WR traffic (low utilization example).

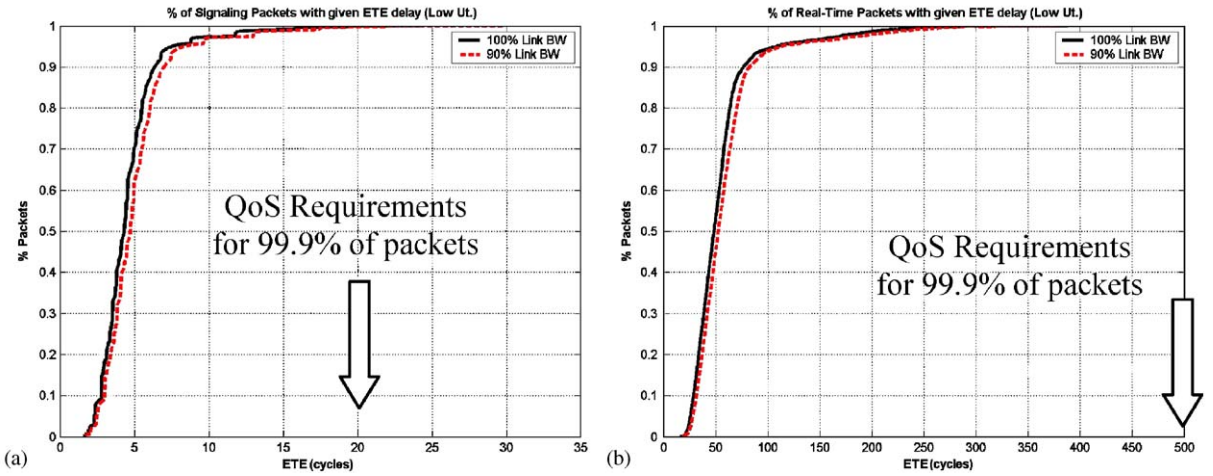


Fig. 9. Degradation in ETE delay of Signaling (a) and Real-time (b) traffic when network bandwidth is dropped 10%. QoS requirements are still satisfied, in spite of this bandwidth reduction.

higher level of utilization. This network is more sensitive to queuing delays than the previous design example, and is typically employed for more moderate QoS requirements. The Initial QNoC is identical to the one in the previous section. It is equipped with minimal buffering (four buffers for each service level, see Section 2.4), it satisfies QoS requirements and consumes a total link bandwidth of 853 Gbps and total wire length of 2.56 m. The total QNoC area (wires and routers) is 2.26 mm^2. The source rate of all service levels is increased by about 40% compared to the previous example (Table 3), leading to higher queuing delays in the network. RD/WR maximum ETE delay requirement is increased from 100 to 350 ns.

As in the previous example, no further buffer increase can improve performance of Signaling packets. On the other hand, Real-time traffic in this example suffers longer queuing delays and its packets are short enough to benefit from additional buffers. The optimization process performed on this service-level (Table 5, Fig. 10) yields an optimum point of five flit buffers for Real-time packets. Subsequently, this number is adopted and the optimization process is performed on RD/WR packets. Indeed as expected, since the source load has been increased and there is much more queuing in the network, more area can be saved by trimming the links bandwidth and increasing buffer space. The optimization process leads to ten buffers for RD/WR packets, while links bandwidth is reduced by 30% (Table 5, Fig. 11). Area is reduced by 0.22 mm^2, which is 10% area saving. Further buffer increments contribute diminishing returns, as shown by an increasing ΔArea function. As above, reduction of network bandwidth results in increased ETE delays of Signaling and Real-time service levels. However, even with this degradation, QoS requirements for Signaling packets are still satisfied; buffer increase in the Real-time service-level (from four to five) brings the optimized QNoC to a point in which Real-time service level QoS requirements are satisfied as well. See Fig. 12.

Table 5
Optimization steps at each service level (the optimum point is indicated in *italics*)

Signaling	BufSize	4	No possible optimization
	Network BW [%]	100	
	Delta Area [mm^2]	0	

Real-Time	BufSize	4	*5*	6	8	No further optimization possible
	Network BW [%]	100	*86*	85	83	
	Delta Area [mm^2]	0	*−0.20*	−0.17	−0.12	

RD/WR	BufSize	4	5	6	8	*10*	12	16	27
	Network BW [%]	100	87	82	75	*70*	68	65	60
	Delta Area [mm^2]	0	−0.138	−0.181	−0.218	*−0.220*	−0.170	−0.055	0.317

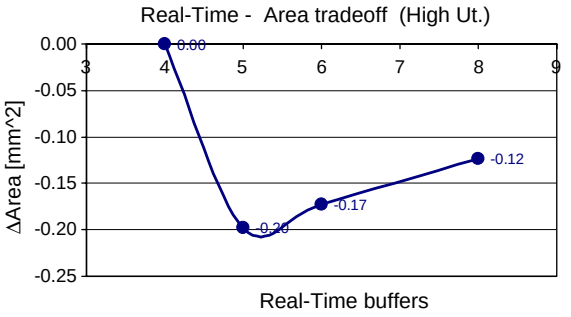


Fig. 10. ΔArea—Optimization performed on Real-time traffic (high utilization example).

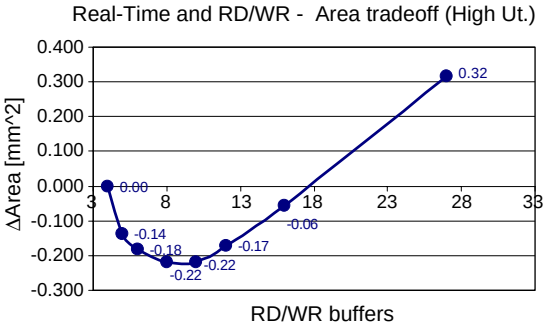


Fig. 11. ΔArea—Optimization performed on Real-Time and RD/WR traffic (high utilization example).

4.2.2. A block-transfer dominated traffic scenario

In this section we check the effect of adding buffer in the case of traffic consisting of long packets. Such communication traffic corresponds to the Block-Transfer service-level, defined in Section 2.2. Block-transfer dominated design example source rate and QoS requirements are summarized in Table 6.

The simulation results (Table 7, Fig. 13) confirmed our expectations. Since Block-Transfer packets are very long the cost of additional buffers that have to be added in order to maintain the required QoS when the link bandwidth is decreased is very high. In other words, it is impossible to decrease the cost of a QNoC designed to move large chunks of data by adding buffers and decreasing link bandwidth. For such QNoC minimal buffering should be employed in order to achieve minimal cost.

In this section we presented a cost optimization process targeted to reduce QNoC area by trading off link bandwidth and router buffer space. We presented several design scenarios, distinguishing traffic by dominating packet length. Simulation results show that only service levels characterized by short packets which are sensitive to queuing delays can benefit from increasing the buffer space. In fact, such increase results in changing the switching technique at these service levels from wormhole to virtual cut-through switching. Naturally, increasing buffer space can be afforded only for relatively short packets. Performance of long packet communications can be improved by enhancing the link bandwidth instead.

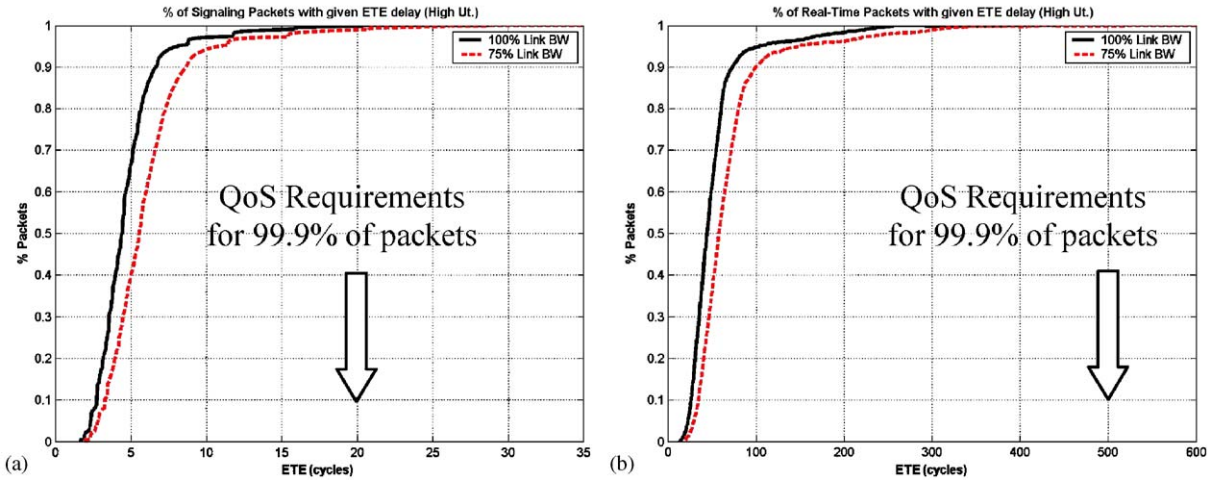


Fig. 12. Degradation in ETE delay of Signaling (a) and Real-time (b) traffic when network bandwidth is dropped 30%. QoS requirements are still satisfied, in spite of this bandwidth reduction.

Table 6
Block-transfer source rate and QoS requirements

Service level	Average packet length (flits)	Average inter-arrival time (μs)	Total load	Max ETE delay requirements (for 99% of packets)
Block-transfer	2000	8.75	3.68 Gbps	50 μs (Several times the transmission delay on 32 bit, 50 MHz bus)

Table 7
Optimization steps at block transfer dominated traffic example (no optimum achieved)

Block-transfer	BufSize	4	32	64	280	No optimum
	Network BW [%]	100	99	96	90	
	Delta Area [mm^2]	0	+1.15	+2.43	+11.31	

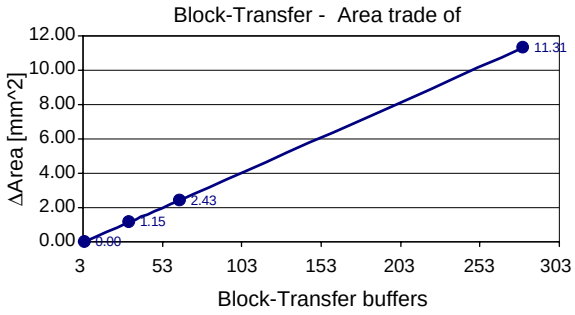


Fig. 13. ΔArea—Optimization performed on a Block-Transfer traffic—no cost minimization can be achieved.

5. Conclusions

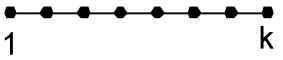
In this paper we quantified analytically the scalability of NoC as a communication architecture for future SoCs and its generic advantage over alternatives in terms of wiring area and power cost for future SoCs, when the number of communicating modules on a chip rapidly increases. We also presented a NoC cost minimization process in which we explored the influence of increasing wormhole buffers versus reducing network links bandwidth, while preserving the required QoS for all classes of service, in order to further minimize the area cost of the NoC. We defined criteria for the characteristics of system communication traffic allowing such cost minimization. We showed several QNoC cost optimization examples in which different levels of area reduction were achieved according to the nature of system traffic among the modules. In a RD/WR dominated, highly utilized network, an example of 10% reduction from original QNoC area was demonstrated. No area cost minimization could be achieved in Block-transfer dominated traffic scenario. The results clearly show the advantage of the wormhole routing technique in networks on chip, since in many cases adding network buffers beyond the minimum value is very expensive and provides a diminishing return.

Acknowledgements

This research has been partially supported by Intel Corp. and the Semiconductor Research Corp. (SRC). We would also like to thank the anonymous reviewers for their useful comments and suggestions that helped in improving this paper.

Appendix

Average distance in 1D array



$$(k-1)\begin{cases} 1+2+3+\ldots+(k-1) - \text{distance from all to node } k \\ 1+2+3+\ldots+(k-2) \quad \text{distance from all to node } k-1 \\ 1+2+3+\ldots+(k-3) \\ \vdots \\ k-(k-1). \end{cases}$$

Summation of all distances

$$\begin{aligned} &1\bullet(k-1)+2\bullet(k-2)+3\bullet(k-3)+\ldots+(k-1)\bullet(k-(k-1)) \\ &\quad = k(1+2+3+..+(k-1)) - (1+2^2+3^2+..+(k-1)^2) \\ &\quad = k\left(\frac{k(k-1)}{2}\right) - \left(\frac{(k-1)k(2(k-1)+1)}{6}\right) = \frac{k^2(k-1)}{2} - \frac{(k-1)k(2k-1)}{6}. \end{aligned}$$

Number of distances (addends)

$$1+2+3+...+(k-1)=\frac{k(k-1)}{2}.$$

Average distance = Sum of all distances/number of distances,

$$\text{Average distance}=\frac{\frac{k^2(k-1)}{2}-\frac{(k-1)k(2k-1)}{6}}{\frac{k(k-1)}{2}}=k-\frac{2k-1}{3}=\frac{k+1}{3}.$$

References

[1] E. Bolotin, I. Cidon, R. Ginosar, A. Kolodny, QNoC: QoS architecture and design process for Networks on Chip, Special issue on Networks on Chip, Journal of Systems Architecture 50 (February 2004) 105–128.

[2] W.J. Dally, B. Towles, Route packets, not wires: on-chip interconnection networks, DAC 2001, Las Vegas, Nevada, USA, June 18–22, 2001.

[3] M. Sgroi, M. Sheets, A. Mihal, K. Keutzer, S. Malik, J. Rabaey, A. Sangiovanni-Vincentelli, Addressing the system-on-a-chip interconnect woes through communication-based design, Design Automation Conference, DAC '01, June 2001.

[4] L. Benini, G. De Micheli, Networks on chips: a new SoC paradigm, IEEE Comput. 35 (1) (2002) 70–78.

[5] S. Kumar, A. Jantsch, J.-P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja, A. Hemani, A network on chip architecture and design methodology, Proceedings of the IEEE Computer Society Annual Symposium on VLSI 2002 (ISVLSI.02).

[6] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, D. Lindqvist, ''Network on a chip: an architecture for billion transistor era'', in: Proceedings of the IEEE NorChip Conference, November 2000.

[7] P. Guerrier , A. Greiner, a generic architecture for on-chip packet-switched interconnections, Design, Automation and Test in Europe Conference and Exhibition 2000, Proceedings, 2000, pp. 250–256.

[8] E. Rijpkema, K. Goosens, P. Wielage, ''A router architecture for networks on silicon'', Proceedings of Progress 2001, 2nd Workshop on Embedded Systems.

[9] K. Goossens, J. van Meerbergen, A. Peeters, P. Wielage, Networks on silicon: combining best-effort and guaranteed services, DATE 2002, Design Automation and Test Conference, March 2002.

[10] A. Radulescu, K. Goossens, Communication services for networks on silicon, in: S. Bhattacharyya, E. Deprettere, J. Teich (Eds.), Domain-Specific Processors: Systems, Architectures, Modeling, and Simulation. Marcel Dekker, New York, 2003.

[11] P. Wielage, K. Goossens, Networks on silicon: blessing or nightmare?, Euromicro Symposium On Digital System Design (DSD 2002), Dortmund, Germany, September 2002.

[12] W.J. Bainbridge, S.B. Furber, Chain: A Delay Insensitive Chip Area IEEE, Micro 22 (5) (2002) 16–23.

[13] AMBA Specification, Arm Inc., May 1999.

[14] The CoreConnect Bus Architecture, IBM, 1999.

[15] D. Wingard, MicroNetwork-based integration of SOCs, in: Proceedings of the 38th Design Automation Conference, June 2001.

[16] C.H. Sequin, R.M. Fujimoto, X-tree and Y-components,VLSI Architecture, Prentice-Hall International, Englewood Cliffs, NJ, 1983, pp 70–78.

[17] J. Rexford, J. Hall, K.G. Shin, A router architecture for real-time communication in multicomputer networks, IEEE Trans. Comput. 47 (10) (1998) 1088–1101.

[18] S.S. Mukherjee, P. Bannon, S. Lang, A. Spink, D. Webb, Compaq Computer Corp., The alpha 21364 network architecture, IEEE Micro. January–February (2002) 26–35.

[19] W.J. Dally, Virtual-channel flow control, IEEE Trans. on Parallel and Distributed Systems 3(2) (1992) 194–205.

[20] InfiniBandTM Architecture Specification, vol. 1, Release 1.0, October 24, 2000.
[21] C.B. Stunkel, J.Herring, B. Abali, R.Sivaram, A new switch chip or IBM RS/6000 SP systems, Proceedings of the 1999 Conference on Supercomputing, January 1999.
[22] W.J. Dally, A VLSI Architecture for Concurrent Data Structures, Kluwer Academic Publishers, Dordrecht, 1987.
[23] L.M. Ni, P.K. McKinley, A survey of wormhole routing techniques in direct networks, IEEE Comput. February (1993), 62–75.
[24] D. Stroobandt, A Priory Wire Length Estimates for Digital Design, Kluwer Academic Publishers, Dordrecht, 2001, pp. 261–262.
[25] The International Technology Roadmap for Semiconductors (ITRS) 2001 ed., Interconnect section, p. 5.
[26] Terry Tao Ye, L. Benini, G. de Micheli, Packetized on chip interconnect communication analysis for MPSoC, DATE 03, 2003.
[27] L. Benini, G. de Micheli, Powering networks on chips: energy-efficient and reliable interconnect design for SoCs, System Synthesis, Proceedings, The 14th ISSI, 2001, pp. 33–38.
[28] Chin-Yuan Chang, Ting-Wei Hou, Ce-Kuan Shieh, The performance improvement of wormhole router for multicomputer systems, TENCON '93, Proceedings, Comput. Commun. Control Power Eng. 1 (1993) 254–257.
[29] W.J. Dally, Virtual-channel flow control, IEEE Trans. Parallel Distributed Syst. 3(2) (1992).
[30] OPNET Modeler, www.opnet.com.