Layout, Performance and Power Trade-Offs in Mesh-Based Network-on-Chip Architectures

D. Pamunuwa, J. Öberg, L. R. Zheng, M. Millberg, A. Jantsch, and H. Tenhunen

Laboratory of Electronics and Computer Systems, Dept. of Microelectronics and Information Technology
Royal Institute of Technology, Stockholm, Sweden.
e-mail: {dinesh|johnny|lrzheng|micke|axel|hannu}@imit.kth.se

ABSTRACT

On-chip packet-switched networks have been proposed for future giga-scale integration in the nanometre regime. This paper examines likely architectures for such networks and considers trade-offs in the layout and wiring strategies based on full-swing CMOS signalling. A study is carried out for a future technology with parameters as predicted by the International Technology Roadmap for Semiconductors to yield a quantitative comparison of the performance and power trade-off.

1. INTRODUCTION

Expected trends in the future evolution of VLSI systems can be codified into the following:

- Moore's law will continue to hold for another ten years [1].
- Single processors will not be able to utilize the transistors of an entire chip, and a single synchronous clock region will span only a small fraction of the chip area [2], [3].
- Applications will be modelled as a large number of communicating tasks, where the tasks may have very different characteristics (e.g. control or data flow dominated) and origins (IP re-use from earlier products or external sources) [4], [5].

An architecture that enforces modularity and is suitable for this kind of heterogeneous implementation is the Network-on-Chip (NoC) architecture [6]-[8]. It eases the expected bottlenecks of complexity and wire delay in nanometre technologies, and promotes extensive re-use of design cores through standardization of on-chip communication.

This paper considers the physical layout of two likely NoC architectures, and carries out a feasibility study in a future 65 nm technology. Some of the issues covered are: layout trade-offs, wiring schemes for the network, and area, performance and power metrics for the different architectures. The cost of the network in terms of power consumption is investigated when standard rail-to-rail CMOS signalling is used.

2. NoC BACKBONE

The NoC backbone consists of *Resources* and *Switches* organised in a Manhattan-like structure with a one-to-one correspondence (Fig. 1.a). All resources are equipped with a *Network Interface (NI)* to communicate between the resource core and the network. The NI handles all communication protocols to make the network as transparent as possible to the resources. To accommodate a reasonably sized network (more than 25 resources), a bus width of 128 bits in each direction for the switch-to-switch and switch-to-resource connection appears suitable [6].

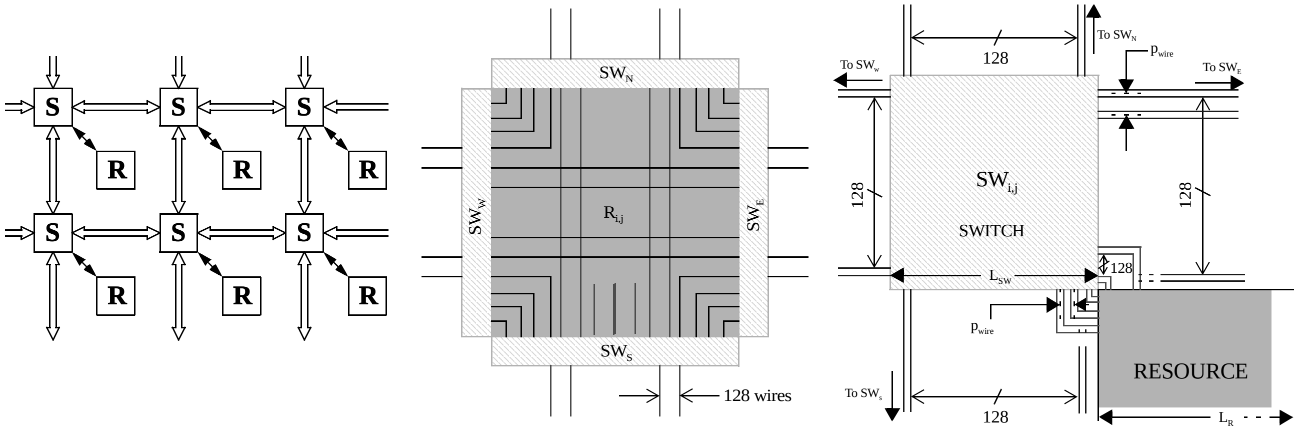
A perusal of the literature shows many works that have elucidated the NoC concept and a layered protocol [6]-[8]. Of these and others, the most attention to the physical level is paid in [6]. It describes a folded torus topology that fits well to VLSI implementations with a two-dimensional layout and limited wires. The proposed routing layout places the network wires on top of the resources in dedicated metal layers. This is the most intuitive layout, but there are in general a myriad of ways to lay out this NoC backbone. Two extremes can be identified: the first is where, as mentioned, the network interconnects are routed over the resource (the "thin-switch architecture" shown in Fig. 1.b), and the second is where the wires run in dedicated channels (the "square-switch" architecture shown in Fig. 1.c). The former has no area overhead associated with the network wires, but routing the wires over the resource does impose a few restrictions on the design methodology of the resource. The placing of repeaters for example may interfere with importing IP cores. Also, to avoid routing congestion over the resource it may be necessary to dedicate one or two metal layers to the network interconnects, which may pose problems in distributing the power and ground networks depending on the number of metal layers available. Even with dedicated metal layers, vias to I/O (power, ground and signal) pads will restrict the number of available wiring tracks. On the other hand, laying restrictions on the routing imposes area overheads, but routing the network and ensuring signal integrity over its wires is straightforward.

Both these two alternative global routing layouts are analysed here. The analysis is conducted for a simple mesh network with only direct neighbour connections but is also applicable to the folded torus topology of [6]. This simplified topology where the switches are connected to their direct neighbours only, is described in [9].

3. MODELLING ISSUES

3.1 Technology Scaling

A feasibility study for NoC implementations in the deep submicrometer (DSM) regime requires models that accurately capture the behaviour of active and passive devices in the given technology node. This is the science of *technology extrapolation* which has received a great deal of attention over the past few decades. Its importance is due to the fact that not only does predicting future trends give us an idea of what is achievable, but



a. Logical mapping of network

b. Thin-switch architecture

c. Square-switch architecture

Figure 1. Network-on-chip Backbone

also has a strong influence on the evolvement of VLSI systems. Influential technology extrapolation systems developed 10-15 years ago are [10] and [11]. More recent second-generation systems include [12] and [13]. A major collaboration of both industry and academia has resulted in the roadmapping venture of the International Technology Roadmap for Semiconductors (ITRS), the latest version of which is [1].

Typically, each system provides estimates of chip area, maximum clock frequency, power dissipation and other parameters based on a small set of descriptors spanning device and interconnect technology through system architecture. Because of the wealth of research available on technology extrapolation, appropriate models for devices and interconnects are either already available for the chosen technology node, or can easily be derived by modifying proposed models. Based upon the ITRS [1], the following properties of a representative 65nm technology available in the year 2007 have been obtained. The process supports 10 layers of metal made of a copper-alumina alloy having a resistivity of 2.5μΩcm. The wires of the lower levels are 210 nm thick, giving a sheet resistance of 0.12 Ω/square. They have a minimum width of 100nm and a minimum pitch of 200nm. The Thevenin equivalent output impedance of a minimum sized inverter (R_{drv}) driving a similar load is 6 kΩ while its input capacitance (C_{drv}) is 1 fF. It is expected that area array bonding techniques will be available for power and I/O connections, with a likely pitch of 150μm.

3.2 Switches and Inter-Switch Links

According to the communication protocol, 128 wires come into and go out of the switch in each direction. Also an additional 128 wires go into and out of the resource to handle the resource's communication with the network. For the thin switch these wires translate to two extra links (Fig. 1.b), while for the square switch they are situated on the two sides of the switch that are closest to the resource (Fig. 1.c). Each incoming and outgoing wire in the switch is latched by a flip-flop, and each outgoing wire is fed from a 4-to-1 multiplexer. Since each switch has 10 sets of 128 wires, this translates to a total of 14,720 gates. Inside the switch, there also exist some additional decision circuitry consisting of adders, subtractors and comparison units, adding up to an estimated 2000 gates. Thus the control logic for the switch is approximately 17,000 gates [15], with each gate occupying approximately 1.6 μm^2. Leaving a routing overhead of 30% for the control logic gives a total of 36,000 μm^2, which translates to an approximate switch size of 0.2 mm X 0.2 mm. Now in the square-switch architecture, the dedicated communication channels are the same width as the side of a switch. Hence the tile size has to be large enough that the area overhead is not too high. A 2mm X 2mm tile size gives an overhead of 20% for the network, which would seem to be an upper limit. In order to be able to compare between architectures, we use the same tile size for both. This choice may seem somewhat arbitrary, but it is a reasonable one, allowing good sized resources to be housed in a single tile, as shown in section 3.3.

An essential part of this analysis is the physical modelling of the network links. Many different signalling techniques have been proposed in the literature including low-swing, differential and current-mode techniques, but the most common and robust is full swing CMOS signalling with inverters as repeaters. This is the signalling convention we adopt here, with analysis techniques similar to those used in [14]. A line is modelled as a distributed *RC* line with capacitive coupling to other lines. The notation adopted is that *k* refers to the number of repeaters (inverters) on a single line including the first driver, and *h* the size of the inverter in terms of multiples of the *W/L* ratio of a minimum sized inverter. This arrangement is sketched out in Fig. 2. The inverters are modelled as resistor-capacitor combinations -with parameters as given in section 3.1- that scale linearly with size. The delay (rise time) is calculated by using Bakoglu's equation [16] with an appropriate distribution of the capacitance into ground and coupled components as given in (1). The interested reader is referred to [14] for further details.

$$t_{r,o} = k\left[0.7\frac{R_{drv_m}}{h}\left(\frac{C_s}{k} + hC_{drv_m} + 4.4\frac{C_c}{k}\right) + \frac{R}{k}\left(0.4\frac{C_s}{k} + 1.51\frac{C_c}{k} + 0.7hC_{drv_m}\right)\right] + \frac{t_{r,i}}{2} \qquad (1)$$

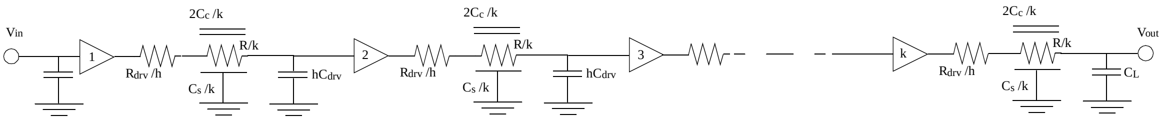


Figure 2. Wire Model for Inter-Switch Network Links

3.3 Gate Count

The number of gates that can be accommodated in a given area can be estimated by using the following scaling equation.

$$N_{new} = N_{old} \times \left(\frac{A_{new}}{A_{old}}\right) \times \left(\frac{\lambda_{old}}{\lambda_{new}}\right)^2 \times \alpha_s \qquad (2)$$

Here N denotes the number of gates, A the area, and 2λ the feature size, with the subscripts *old* and *new* referring to the technology in which the design is currently implemented, and the future technology respectively. The factor α_s ($1 \geq \alpha_s > 0$) is used to account for integration losses or gains in the scaling. A high performance ASIC in an 0.35 μm technology comprises approximately 1M gates on a 16 mm X 16 mm die [17]. Using (1) to scale to a 2 mm X 2 mm area for the same integration efficiency (α_s=1) results in a gate count of 450k. A more representative gate count for a tile is obtained by relaxing α_s to account for standard cell-based designs, to approximately 200k gates.

3.4 Power Estimations

The power consumption in the NoC is composed of the portions dissipated in the resources, and the network. Most of the power consumed in the resource is due to switching logic. The average current consumed in a cycle [17] is

$$I_{avg} = \frac{N_s C_{ld} V_p}{t_{clk}} \qquad (3)$$

where N_s is the fraction of gates switching in one direction in one clock cycle, C_{ld} the average capacitive load of a gate, V_p the positive rail voltage, and t_{clk} the clock period. A typical gate load in the 65 nm technology is estimated to be 4 fF, the power supply to be 0.9 V, and the local clock frequency to be 3 GHz. The number of switching gates can be computed by assuming that half the gates switch in any given cycle, with equal numbers in each direction. This means that for our representative 200k gate resource, N_s = 50,000, giving an average current of 0.64 A. The power is then *VI*, or 0.52 W per resource. Assuming a likely die size of 3 cm [1], the total power consumed in the resources is roughly 120 W.

The average power consumed in a single wire in an inter-switch link is calculated by the following expression

$$P_w = k(h \times C_{drv} + C/k) \times V_p^2 \times f_b \qquad (4)$$

where k, h and C_{drv} are as defined in section 3.1, and f_b is the bit period. The total power consumed in the network links is calculated by

$$P_{nw} = P_w \times N_w \times \beta \times \delta \qquad (5)$$

where β and δ are coefficients that represent the fraction of bits that switch on a given link in one direction (assumed to be 0.25, i.e. half the bits switch in both directions), and the fraction of links that are active in any given cycle (assumed to be 0.5). By using an analysis similar to that used for resources, it can be verified that the power consumed in the network switches is negligible in comparison to the power expended in the link. Hence it is neglected.

4. RESULTS

4.1 Square-Switch Architecture

Sketched out in Fig. 1.c is the switch and resource arrangement. As mentioned, the area overhead for this architecture is 20%. To reduce this, there are two possibilities:

(a). let the resource area extend under the wire channels, creating a compromise between architectures 1 and 2;

(b). let the switch extend into the wire channels (i.e., shape it like a '+' sign instead of a square).

The former will not be investigated as there are a great variety of different geometrical arrangements possible and the intent here is to investigate the two extremes. The latter trades off area overhead against link bandwidth. Consider that the switch extensions (peripheral legs of the '+') have linear dimensions equal to the square in the centre, when 5 times the area of the central square is available. For the 36,000 μm^2 total area mentioned above, this translates to a switch that is 85μm or say 100μm square, with extensions of the same size into all four channels. Now the area overhead is only 10% but the channel width for the communication link is halved, resulting in a lower bandwidth. However because of the non-linear scaling of the parasitics with wire width, this is only a small percentage decrease [14].

All metal layers can be utilised to route the network wires. If it is assumed that 10 metal layers are available for example, and that the top three are reserved for power, ground and clock distribution (which is conservative), 7 layers remain for routing the communication link. One possible implementation would be to route the 256 inter-switch wires in 4 layers, with explicit signal return planes between each signal layer. This also provides for shielding between metal layers, which is important as the wires on all layers are parallel. Then for the western and eastern sides of the switch, 64 wires need to be routed on a metal layer, giving a wire pitch of approximately 3μm. Since the 128 wires between the resource and switch on the southern and eastern sides run only for a short fraction of the resource length, they can either occupy one of the shielding layers briefly, or be routed on the four signal layers along with the inter-switch wires at less than maximum pitch. Once all the wires have gone into the resource,

Table 1. Slew rates for wiring schemes of square switch architecture

Switch shape	Area overhead	k	h	t_r (psecs)
square	20%	4	140	70
+	10%	4	85	80

the remaining wires (from the inter-switch link) spread out to take maximum advantage of the space available. The distance for which wires are congested will be a very small fraction of the total length of 2mm, and can be neglected for timing purposes.

Due to lack of space, the modelling details related to delay analysis are omitted. The interested reader can refer to [14] for a detailed treatment. Shown in Fig. 3 are plots detailing the variation of slew rates for individual wires and the power consumption for the entire network. Given that the maximum pitch is 3 µm for 64 wires on a single layer, an arrangement that appears to maximise the bandwidth for the square switch is the following: the signal wires are 1.2 µm wide on 3 µm centres, and between the signal lines are thinner shielding lines of the minimum width of 0.1 µm, on the same 3 µm pitch. The signal wires are also vertically shielded with clearly defined return paths. For the '+' shaped switch, the signal wire width halves to 0.6 µm, while the pitch of the signal and shielding wires changes to 1.5 µm. Locating repeater stations in the channel is ideal in terms of utilising space. Since the wire pitch is not sufficient to fit all repeaters horizontally, they can be placed in zig-zag fashion, so that the channel is packed with repeaters.

Given in Table 1 are the delays (rounded to the nearest 10th digit) corresponding to the two cases along with repeater details (rounded to the nearest 5th digit) for the maximum possible bandwidth. As a point of comparison, the line delay at the speed of light in SiO_2 is 13.3p seconds.

4.2 Thin-switch Architecture

The physical layout is that a resource block is surrounded by four thin switches, each of which is connected to four other switches (Fig. 1.b). In addition one switch has wires going into the resource. The shorter dimension of the switch is defined by the layout of the logic contained in it. Since the control logic will be the same for both architectures, the total area is also the same, but as the logic will be distributed along the full side of the region, the area overhead for the switches is negligible.

Instead the main consideration will be routing congestion over the resource. The vertical and horizontal wires of the network need to be wired in two metal layers since the lines cross as shown, restricting either the number of metal layers to be used by the resource blocks or the wiring freedom in terms of the available fraction of a metal layer. Although the wires may occupy less than a quarter of each metal layer, having the resource share them for local signal wiring would impose fairly severe restrictions on the design methodology of the resource. It would appear to be more viable to share the metal layer between the network and power distribution grid, which would also provide some form of mutual decoupling. Devoting two entire layers to the network if possible would of course be the ideal solution.

Table 2. Slew rates for wiring schemes of thin switch architecture

Scheme	k	h	t_r (psecs)
Dedicated metal layers	1	40	140
	1	55	120
	2	40	110
	2	55	100
Shared metal layers	1	20	190
	1	25	170
	2	20	150
	2	25	130

Now pins to power, ground and signal I/O pads over the chip will limit the number of wiring tracks for all metal layers. Practice shows that between 20% and 30% of a metal layer will not be wireable. Two cases are considered:

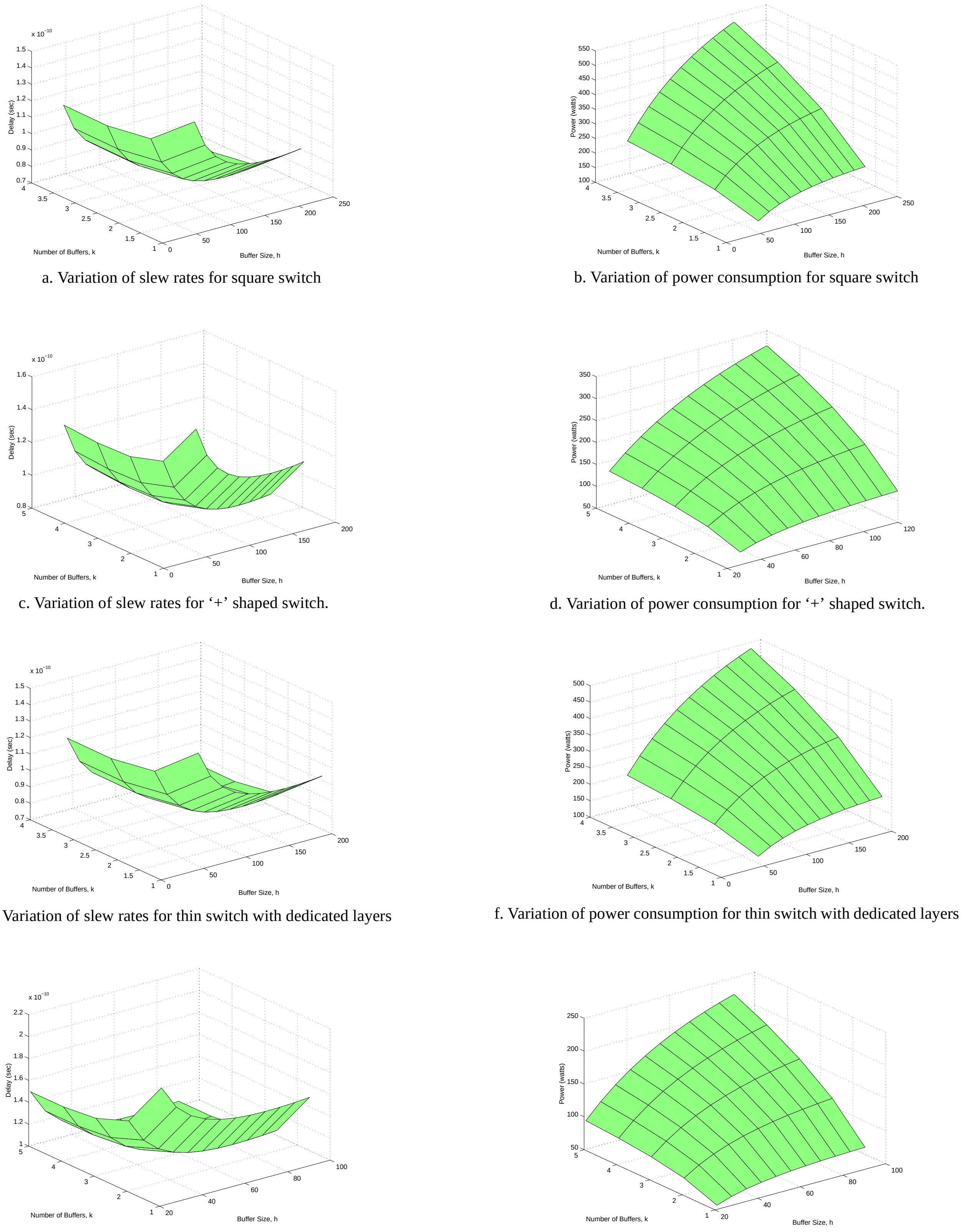
(a). two metal layers are devoted to routing the network;

(b). only a certain fraction of each metal layer is utilized for the network.

In the first case it is estimated that signal and power and ground pins (including connections to on-chip decoupling capacitance) to pads will render 20% of a metal layer unusable, while in the second that 70% is unusable due to additional utilisation for the power distribution grid. In both cases 512 wires need to fit on one side (of 2 mm length) on a single metal layer, comprising the links to two switches as can be seen from Fig. 1.b. As in the dedicated channel architecture, the wires into the resource may either occupy the same metal layer at the minimum pitch or a different metal layer as they are very short in comparison with the other wires.

In the first case, the wire pitch is 3.1 µm, while it is 1.1 µm in the second case. In comparison to the overhead of the NI, that imposed by repeater stationing inside resource IPs would be relatively minor. However it is possible to obtain fairly good timing figures without extra repeaters, and the following layout is a reasonable option for the first case: the signal wires are 1.1µm wide on 3 µm centres, and between the signal lines are thinner shielding lines of 0.2 µm width, on the same pitch. For the second case a reasonable layout would be that the signal wires are 0.4 µm wide on 1.1 µm centres. Now there is no need to provide explicit shielding as the signal wires are interspersed with power and ground lines. Fig. 3 and Table 3 gives plots and data for likely timing and power consumption for both schemes.

In this communication architecture, the header information of all received messages needs to be forwarded to the side that takes the decision (the western side in this case) to resolve its destination [6]. Then the packet is forwarded to the side where the data is actually multiplexed into the outgoing flip-flops. This means that the actual delay is twice the line delay, unless the decision procedure is pipelined into two stages, which would increase the



a. Variation of slew rates for square switch

b. Variation of power consumption for square switch

c. Variation of slew rates for ‘+’ shaped switch.

d. Variation of power consumption for ‘+’ shaped switch.

e. Variation of slew rates for thin switch with dedicated layers

f. Variation of power consumption for thin switch with dedicated layers

g. Variation of slew rates for thin switch with shared layers

h. Variation of power consumption for thin switch with shared layers

Figure 3. Performance and Power Consumption for Different Wiring Schemes

switch area by a factor of two. Since the area overhead is negligible, a pipelined structure is indeed assumed to allow a direct comparison with the square-switch architecture.

5. CONCLUSIONS

One interesting conclusion of this study is that the power consumption of the NoC seems to be dominated by the power consumed in the network, when rail-to-rail signalling is used. As can be expected, it is possible to trade-off power for bandwidth, and for the maximum possible bandwidth (in the square-switch architecture) the power consumed in the network is more than 4 times as much as the power consumed in the resource logic. Some of the advantages and disadvantages of the two architectures are given in Table 3.

The intuitive and obvious layout is the thin-switch architecture, but the square-switch architecture also recommends itself in certain aspects. In summary, the square switch architecture guarantees signal integrity and provides a higher link bandwidth at the cost of a fairly high area overhead. The thin-switch architecture is more difficult to wire and has a slightly lower link bandwidth, but has negligible area overhead. In both, the power consumption is dominated by the power expended in the network. Some possibilities to reduce the power in the network are usage of low-swing or current-mode signalling techniques, and encoding of the data packets to minimise transitions. These possibilities will be investigated in future work.

The choice of architecture, whether one of the above or a hybrid of the two, will of course depend on the application. However this study has shown the feasibility of implementing the NoC concept under the physical constraints of interconnections in the DSM regime, and cost and performance estimates were extracted by considering the physical implementation in as much detail as possible.

Table 3. Pros and Cons of the Two Schemes

Square Switch	Thin Switch
Area overhead of between 10% and 20% for network	Virtually no area overhead
All metal layers can be freely utilized for resource	No. of available metal layers or available fraction of two metal layers reduced for resource
No routing/pin congestion over resource due to network	Routing/pin congestion introduced by network
Dedicated channel allows repeater insertion, shielding and explicit signal return planes, guaranteeing signal integrity	Repeater insertion and shielding more of a problem. More susceptible to noise coupling from above and below
Max. link bandwidth of 1.22 Tbits/sec in any direction	Max. link bandwidth of 0.85 Tbits/sec and 0.65 Tbits/sec with no repeaters on dedicated and shared metal layers respectively
Power consumption dominated by network	Power consumption dominated by network

6. REFERENCES

[1] International Technology Semiconductor Roadmap (ITRS) 2001. [Online]. Available: http://public.itrs.net/Files/2001ITRS/Home.htm

[2] D. Sylvester and K. Keutzer, “Getting to the Bottom of Deep Submicron”, in *Proc. ICCAD*, 1998, pp. 203-211.

[3] A. Hemani et. al., “Lowering power consumption in clock by using Globally Asynchronous Locally Synchronous Design style”, in *Proc. DAC*, 1999, pp. 21-25.

[4] C. Szyperski, *Component Software: Beyond Object Oriented Software*, Reading, MA, ACM/Addison Wesley, 1998.

[5] D. Gajski, R. Dömer and J. Zhu, “IP-Centric Methodology and Design with the SpecC Language”, *System Level Design*, Ed. A. A. Jerraya and J. Mermet, Nato Science Series, Vol. 357, 1999.

[6] W. J. Dally and B. Towles, “Route packets, not wires: on-chip interconnection networks,” in *Proc. DAC*, 2001, pp. 684-689.

[7] M. Sgroi, M. Sheets, A. Mihal, K. Keutzer, S. Malik, J. Rabaey, and A. Sangiovanni-Vincentelli, “Addressing the system-on-a-chip interconnect woes through communication based design,” in *Proc. DAC*, 2001, pp. 667-672.

[8] L. Benini and G. DeMicheli, “Powering Networks on Chip,” in *Proc ISSS*, 2001, pp. 33-38.

[9] M. Millberg, “The Nostrum protocol stack and suggested services provided by the Nostrum backbone”, Technical Report TRITA-IMIT-LECS R 02:01, Laboratory of Electronics and Computer Systems, Department of Micro-Electronics and Information Technology, Royal Institute of Technology, Stockholm, Sweden, 2003.

[10]H. B. Bakoglu and J. D. Meindl, “A system-level circuit model for multi- and single-chip CPUs,” in *Proc. ISSCC*, 1987, pp. 308-9.

[11]G. A. Sai-Halasz, “Performance trends in high-end processors,” in *Proc IEEE*, vol. 83, pp. 20-36, Jan. 1995.

[12]RIPE: Rensselaer Interconnect Performance Estimator. [Online]. Available: http://latte.cie.rpi.edu/ripe.html

[13]BACPAC: Berkeley Advanced Chip Performance Calculator. [Online]. Available: http://www.eecs.umich.edu/~dennis/bacpac/

[14]D. Pamunuwa, L. R. Zheng and H. Tenhunen, “Maximising Throughput over Parallel Wire Structures in the Deep Submicron Regime,” *IEEE Trans. VLSI Systems*, vol. 11, no. 2, pp. 224-243, April, 2003.

[15]E. Nilsson, “Design and implementation of a hot-potato switch in a network on chip,” MSc Thesis, Royal Institute of Technology, Department of Micro-Electronics and Information Technology, Laboratory of Electronics and Computer Systems, Stockholm, Sweden, Jun. 2002.

[16]H. B. Bakoglu and J. D. Meindl, “Optimal Interconnection Circuits for VLSI,” *IEEE Trans. Electron Devices*, vol. ED-32, no. 5, pp. 903-909, May 1985.

[17]W J Dally and J W Poulton, *Digital Systems Engineering*, CUP, 1998.