Replacing Global Wires with an On-Chip Network: A Power Analysis *

Seongmoo Heo and Krste Asanović
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139
{heomoo,krste}@csail.mit.edu

ABSTRACT

This paper explores the power implications of replacing global chip wires with an on-chip network. We optimize network links by varying repeater spacing, link pipelining, and voltage scaling, to significantly reduce the energy to send a bit across chip. We develop an analytic model of large chip designs with an on-chip two-dimensional mesh network and estimate the power savings possible in a 70 nm process for two different design points: a circuit-switched ASIC or FPGA design, and a dynamic packet-switched tiled architecture. For circuit-switched networks, achievable power savings are 35–50% for a mesh with 1 mm links. The packet switched designs use multiplexing and signal encoding to reduce the number of link wires required, but the router overhead limits peak wire power savings to around 20% with optimal tile sizes of around 2 mm.

Categories and Subject Descriptors: B.7.1 [Integrated Circuits]: Types and Design Styles–*Advanced Technologies, Microprocessors and Microcomputers, VLSI*

General Terms: Performance, Design, Theory

Keywords: On-Chip Network Power Model, Pipelining, Router, Tile Size, Tiled Architecture, Wire Power Model

1. INTRODUCTION

Cross-chip global wires are becoming increasingly problematic as feature sizes shrink, with their delay and power consumption increasing rapidly relative to individual logic gates [13]. Long global wires also cause many design problems including routing congestion, noise coupling, and difficult timing closure. These worsening trends have led to proposals to replace design-specific global wiring with structured on-chip networks in large ASIC designs [17, 7]. Circuit blocks or *tiles* now communicate by sending packets across the on-chip network instead of driving signals across dedicated global circuit wires. The on-chip network links can be highly optimized by controlling their electrical environment to allow the use of optimized signaling techniques. Wiring efficiency can be improved by replacing a large number of low activity dedicated wires with fewer multiplexed communication links. Routers represent the main area and power overhead in an on-chip network.

In this paper, we explore the power implications of replacing global wires with an on-chip network. We first develop detailed power models for power-optimized wires, including the effects of leakage current in a 70 nm process. Deep wire pipelining can reduce communication power by using fewer and smaller repeaters and lower interconnect supply voltages. We examine the tradeoffs between pipeline depth, repeater sizing, repeater spacing, supply voltage, threshold voltage, and total power. We show how small increases in cross-chip latency can significantly lower the energy required to send a bit across chip. Many applications for which custom chips are developed have ample parallelism, which can be used to tolerate the increased interconnect latency of power-optimized pipelined global wires.

We next examine the use of power-optimized wires in two contexts: 1) conventional ASIC or FPGA designs where dedicated global wires are replaced with dedicated but power-optimized wires, 2) tiled architectures, where all inter-tile global communication is via a dynamic packet-routed on-chip network using power-optimized links. We vary tile size and use Rent's Rule [5] to estimate interconnect density. Smaller tiles put more connections on the power-optimal wires of the on-chip network, but require more routers. Our results show that although power-optimized wires can reduce global wire power significantly (35–50%) in wire-routed ASIC or FPGA designs, it is difficult to achieve significant power savings in packet-routed tiled designs due to the energy expended in routers even for highly multiplexed inter-tile traffic. Tile sizes of around 2 mm on a side appear to provide the lowest total communication power.

2. RELATED WORK

Bakoglu [1] reported the delay-optimal repeater sizing and spacing for a repeated wire. Ho et al. [13] pointed out that the power consumption of the delay-optimal repeated wire is prohibitively large, and suggested increasing repeater spacing and decreasing repeater size to save power while sacrificing some speed. Kapur et al. [14] and Banerjee and Mehrotra [2] calculated the power-optimal repeater sizing and spacing for global interconnects using a simple first-order RC repeated wire model. Gupta et al. [11] described a high-level interconnect power model for wires of a single core chip.

Chandrakasan et al. [3] first suggested the use of pipelining for power reduction in digital circuits. Heo and Asanovic [12] examined power-optimal pipelining for logic datapaths in deep submicron technology both analytically and through circuit simula-

*This work was partly funded by NSF CAREER award CCR-0093354, NSF ITR award CCR-0219545, and a donation from Intel Corporation.

ISLPED'05, August 8–10, 2005, San Diego, CA

tion. Cocchni [6] estimated the effect of concurrent flipflop and repeater insertion while considering routing tree topology, but focused only on minimizing wire latency. Liao and He [15] modeled full-chip interconnect power using a more sophisticated concurrent repeater and flipflop insertion scheme, and showed how increased wire pipelining could reduce communication power.

Sgroi et al. [17] and Dally and Towles [7] proposed replacing design-specific global on-chip wiring with a general-purpose on-chip interconnection network. Eisley and Peh [9] first provided a high-level network power analysis with link utilization as the abstraction of network power. Some previous work focused on the low-level power estimation of routers. Wang et al. [18] provided a low-level general framework for different types of routers, Orion, and verified the simulator with Alpha 21364 and Infiniband router examples [19]. Chen and Peh [4] added a leakage power model to the Orion simulator. Ye et al. [22] focused only on the power consumption of the switch fabric in a router.

In this paper, we build a complete system-level power model for on-chip networks including routers and power-optimal link wires. We examine the trade off in tile size versus communication power, using Rent's Rule to estimate inter-tile and intra-tile interconnect.

3. WIRE POWER MODEL

In this section, we present an analytical latency and power model for a pipelined and repeated wire where throughput is fixed. We include the active and leakage power consumed by repeaters and flipflops in addition to the switching power of the wire capacitance.

3.1 Methodology

We choose BPTM 70 nm technology as our deep submicron process technology [8]. Base supply voltage is 0.9 V and V_{th} is 0.20(-0.22) V. V_{th} is set quite high to reduce leakage power of repeaters and flipflops. We assume that clock period is fixed at 24 FO4 delays (clock frequency of 2 GHz), representing a high-performance digital circuit [12]. For most designs, clock frequency will be set by logic within a tile, and we assume the network links run at the same frequency with a fixed throughput requirement.

We assume three categories of metal interconnect: local, semi-global, and global. Table 1 shows the characteristic dimensions and RC components of these wires.

70 nm Cu tech	Local	Semi-global	Global
Width (μm)	0.10	0.14	0.45
Spacing (μm)	0.10	0.14	0.45
Thickness (μm)	0.20	0.35	1.20
Height (μm)	0.20	0.20	0.20
Resistivity (Ωcm)	2.2	2.2	2.2
$C_{wire}(fF/mm)$	152	178	228
$R_{wire}(\Omega/mm)$	1100	449	41
$R_{wire}C_{wire}$ (FO4)	8.03	3.84	0.45
Max. distance in 24 FO4 (mm)	2.10	3.04	8.88
Max. distance in 24 FO4 (mm)	2.10	3.04	8.88
Latency-optimal repeater spacing (mm)	0.29	0.42	1.22

Table 1: Wire characteristics of our example 70 nm technology.

3.2 First-Order RC Wire Model

Figure 1 shows a first-order RC model of a wire [13]. We assume minimum-sized flipflops. In the RC circuit, the wire segment is modeled inside the dotted box and the repeaters are outside.

Wire delay is represented as a function of the repeater sizing, r. We define r as the ratio of the repeater gate cap and wire cap within a wire segment.

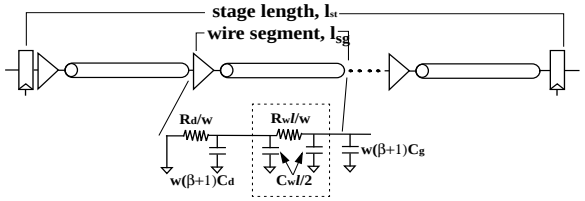

Figure 1: First-order RC model of wire. l, l_{st}, and l_{sg} are the lengths of the whole wire, one stage, and wire segment respectively. w and βw are the widths of the repeater NMOS and PMOS transistors (normalized by the minimum width). C_w, C_d, and C_g are the unit-length wire cap and drain and gate caps of the minimum-sized NMOS transistor respectively. R_w and R_d are the unit-length wire resistance and the resistance of the minimum-sized repeater respectively.

$$r = \frac{w(\beta+1)C_g}{C_w l_{sg}}$$

When interconnect supply voltage is scaled, the scaling factor,

$$S = \frac{V_{dd}}{V_{dd0}} \times (\frac{V_{dd0} - V_t}{V_{dd} - V_t})^{\alpha}$$

is used, where α is career velocity saturation index, V_t is threshold voltage, and V_{dd0} and V_{dd} are baseline and scaled supply voltages respectively.

Wire delay and link latency are calculated as:

$$\begin{aligned} Delay &= 0.7\frac{l_{st}}{l_{sg}}\frac{R_d}{w}(w(\beta+1)(C_d+C_g) + l_{sg}C_w) \quad (1)\\ &\quad +0.7\frac{l_{st}}{l_{sg}}(l_{sg}^2\frac{R_wC_w}{2} + l_{sg}R_w w(\beta+1)C_g) \quad (2)\\ &= 0.7l_{st}(\frac{1}{3l_{sg}} + \frac{2}{9rl_{sg}} + \frac{kl_{sg}}{2} + krl_{sg})FO4 \quad (3)\\ Latency &= \frac{0.7L}{T - D_{ff}}(\frac{1}{3l_{sg}} + \frac{2}{9rl_{sg}} + \frac{kl}{2} + krl_{sg}) \quad (4) \end{aligned}$$

where T is the clock period, D_{ff} the flipflop delay, and k is the unit-length wire RC delay in FO4.

We calculate power as:

$$\begin{aligned} Power &= \frac{1}{2}AF((1+\frac{3}{2}r)C_w l + C_{ff}latency)fV_{dd}^2 \quad (5)\\ &\quad + (k_{rep}(\frac{3}{2}r)C_w l + k_{ff}C_{ff}latency)V_{dd}^{1+\gamma} \quad (6) \end{aligned}$$

where AF is the activity factor and C_{ff} is the flipflop cap. The repeater drain cap is assumed to be half of the repeater gate cap. k_{rep} and k_{ff} are leakage power coefficients for repeaters and flipflops. The first term of the equation is the switching power component and the latter is the leakage power component. We assume leakage current remains constant regardless of input patterns or internal states. Leakage power scales super-linearly in deep submicron technology ($1 < 1 + \gamma < 2$) [12].

3.3 Pipelining Wire

Latency is minimized when r^* is $\frac{1}{\sqrt{3}}$ and l^* is $\sqrt{\frac{2}{3k}}$, where wire delay and repeater delay are equal. However, this minimum latency point requires very large, power-consuming repeaters [13].

We can save power by using deeper pipelining to provide additional time slack in each wire stage. Although there are many ways of exploiting the time slacks obtained from pipelining, we focus on two variables: repeaters and supply voltage. We can either reduce

the size and increase the spacing of repeaters, or scale down supply voltages, or both.

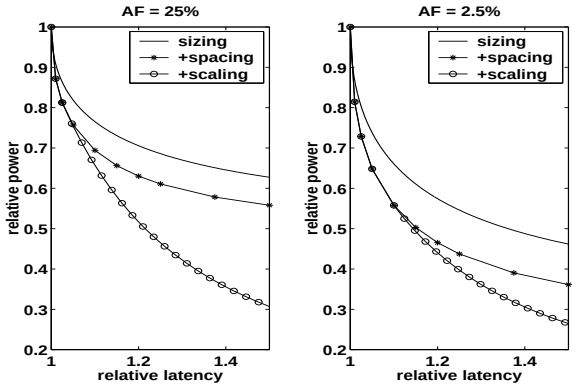

Figure 2: Wire latency-power curves for activity factors of 25% and 2.5%, while changing repeater sizing, spacing, and supply voltage. Both axes are normalized to the minimum latency point.

Figure 2 shows latency-power curves of wires while varying repeater sizing, spacing, and supply voltage. In the figure, the `sizing` curve shows latency-power tradeoff through repeater sizing only, while repeater spacing is fixed at the minimum latency point and supply voltage is constant at the nominal voltage. Increasing repeater size over the minimum latency point results in larger latency and power. The `+spacing` curve shows latency-power tradeoff through repeater sizing and spacing. Repeaters are power-optimally sized and placed while supply voltage is fixed.

Finally, the `+scaling` curve adds supply voltage scaling to the optimally sized and spaced repeaters. Supply voltage scaling is by far one of the most effective techniques for trading time slack for power. Supply voltage reduction leads to a quadratic reduction in active power and also a super-linear reduction in leakage power, as leakage current has a strong dependency on drain voltage in deep submicron processes [12]. Adding supply voltage scaling enables much greater power saving compared to optimal repeater sizing and spacing alone, especially when latency is allowed to increase by greater than 10%, but requires a second power supply to be distributed to the interconnect network.

Repeater sizing and spacing is more effective when the activity factor is low, as it can significantly reduce leakage power. For higher activities, wire cap switching is a significant portion of active power and is unaffected by repeater sizing and spacing.

Overall, the combined techniques achieve a factor of 3–4 reduction in communication power when latency is allowed to increased by 50%.

4. ON-CHIP INTERCONNECT NETWORK POWER MODEL

In this section, we develop a system-level interconnect power model for large digital designs. We assume the design is for a highly parallel system, such as a DSP engine or network processor, and assume the design can be divided easily and flexibly into any number of smaller tiles. Each tile represents a computation module including local memory. We examine three design points: 1) a single tile containing the whole design, 2) a tiled design where tiles are connected with wire-routed power-optimal pipelined wires, and 3) a tiled design with a packet-routed inter-tile network.

We focus only on power consumption for this analysis and assume the performance impact from additional inter-tile latency is not significant. When we divide the chip, we keep the logic and local memory ratio the same regardless of the tile size. We assume the total power consumed by logic and memory transistors remains roughly the same regardless of network configuration and focus on power consumed by communication wires (intra-tile or inter-tile) and supporting transistors (repeaters or registers). Communication power is already comparable to logic and memory power and increasing as digital systems become more communication-centric than computation-centric.

Communication power can be divided into three parts: inter-tile wire power, router power, and intra-tile wire power. The intra-tile wire is the power consumed by design-specific wires connecting logic and memory transistors within tiles. As tile sizes shrink, more intra-tile wires move onto the inter-tile wires depending on tile and network architecture.

Table 2 summarizes the dimensions of our example chips. Estimates are based on ITRS 2004 [10]. We assume that gates are uniformly distributed on a chip. Although Rent's rule provides the estimates of number of wires, it does not give the bandwidth requirements. For simplicity, we assume a uniform activity factor (AF) regardless of wire length.

Chip length (mm), M	20
Tile length (mm), L	0.5 – 20
Gate density (gates/mm^2)	3.2×10^5
Gate pitch (μm)	1.8

Table 2: Dimensions of our example chips.

4.1 Single Tile Baseline

We first use Donath's method to estimate the wire distribution for the whole chip treated as a single tile. Estimation of wire distribution is an important application of Rent's rule, which is an empirical rule stating that the number of wires leaving a circuit block is exponentially proportional to the number of gates in the block. We assumed a nominal value, $\frac{2}{3}$ for the Rent exponent, p which agrees with the wire distribution of Intel microprocessors [20].

We chose the following wire distribution equations, which divide wires into two regions [5]:

$$N(l) \propto (\frac{l^3}{3} - 2Ml^2 + 2M^2 l) l^{(2p-4)} (l \leq M) \quad (7)$$

$$\propto (2M - l)^3 l^{(2p-4)} (M < l \leq 2M) \quad (8)$$

where l is the length of a wire, M is the chip length, and $2M$ is the length of the longest wire within the chip assuming Manhattan wiring.

We assumed that a local wire is used if the wire length is less than 100 gate pitches, a semi-global if less than 1400, and a global for other longer wires. Figure 3 shows the wire distribution of our base chip. We can see that the number of wires decreases drastically as we reach the very longest wires in region 2. Figure 4 shows the wire power of the base chip, at the point where the tile size is 20 mm (that is, when the whole chip is one tile).

4.2 Wire-Routed Tiles

We now divide the chip into multiple tiles, and replace inter-tile wires with power-optimal pipelined wires. The number and total length of wires does not change. We assume that all the wires longer than twice the tile size ($2L$) are replaced with inter-tile wires, while the rest remain intra-tile wires.

We assume that all the intra-tile wires regardless of dimensions, are latency-optimized. In particular, compared to inter-tile wires,

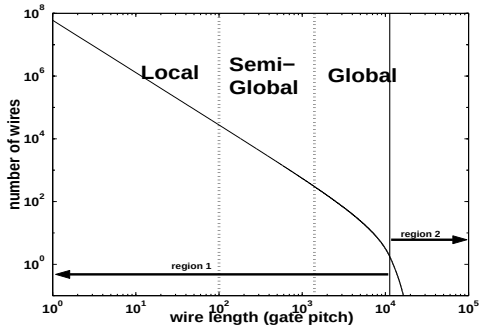

Figure 3: Wire distribution of our base chip. Two dotted vertical lines divide wires into local, semi-global, and global wires. The vertical solid line shows the boundary between region 1 and 2 wires.

they are more performance-critical. The following equations show that the total power of the intra-tile wires.

$$N(l) \propto l^{2p-3} (l \leq 2M) \quad (9)$$

$$P_{intra} \propto \sum_{l=1}^{2L} N(l) l \quad (10)$$

$$\propto L^{2p-1} \quad (11)$$

where L is the tile size. Eq. (9) is a simplified and combined form of Eq. (7) and (8).

Figure 4 shows the wire power consumption of a tiled wire-routed chip. As the tile size decreases (smaller than half of the chip), intra-global wire power decreases exponentially while intra-semi-global or intra-local wire power remain roughly unchanged. However, the increase in power on inter-tile wires matches the power loss of the intra-global wires and so the total wire power stays roughly the same.

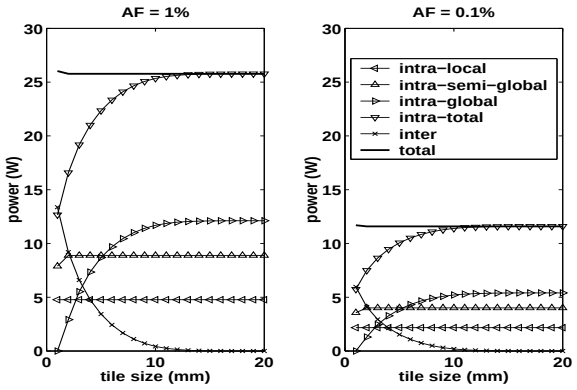

Figure 4: Wire power consumption of tiled wire-routed design for varying tile sizes, assuming uniform activity factors of 1% and 0.1%.

Figure 5 shows the power saving of a tiled wire-routed chip when inter-tile latency is increased by 25% and the network wires are pipelined. Smaller tiles result in greater power saving as more signals are pipelined. In particular, when leakage power is significant (AF = 0.1%), pipelining through repeater optimization and voltage scaling is more effective at reducing the inter-tile wire power, since pipelining is more effective at saving leakage power (Section 3.3). When AF is 0.1%, almost half of the total power can be saved.

4.3 Packet-Routed Tiles

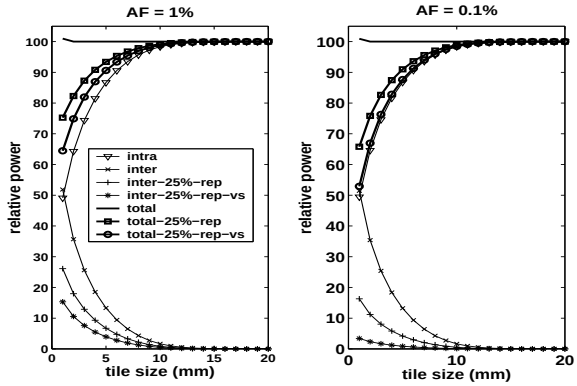

Figure 5: Tiled wire-routed design: power saving by pipelining. `inter-25%` means 25% increased latency requirements for the inter-tile wires. `rep` represents repeater sizing and spacing and `vs` includes voltage scaling as well as repeater sizing and spacing.

We finally consider a packet-routed tiled architecture. Figure 6 shows an example tiled ASIC architecture (4 by 4) and a mesh interconnect network. A mesh interconnect network was chosen since it is simple to design, power-efficient, and scalable. Tiles communicate with others only through routers and links between routers.

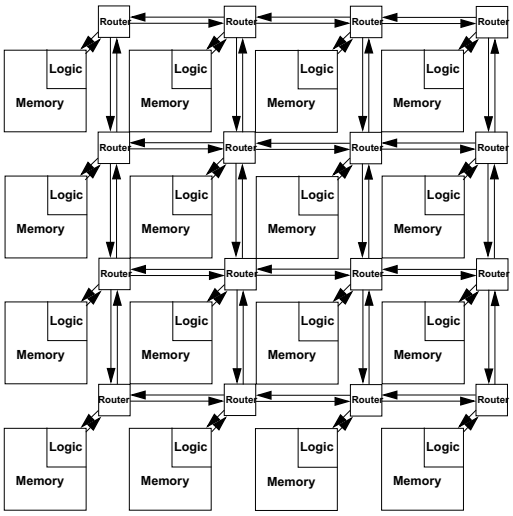

Figure 6: Tiled packet-routed ASIC design.

On-chip network links are much cheaper in terms of area and power, than traditional off-chip network links. Thus it is natural that tiles exploit wider links than a single-tile chip does. However, an excessive number of on-chip IO wires result in a huge power and area overhead for routers. Usually, some degree of multiplexing and packet encoding are employed to reduce the number of IO wires while increasing the activity on link wires. We define the multiplexing factor, MF, as the ratio between activity on packet-routed link wires (LAF) and that on the inter-tile wire-routed links they replace (AF). Higher MF results in area and leakage reduction, but increased complexity. MF varies according to application and tile architecture.

We assume the total chip bandwidth (BW) is conserved and the

total sum of global wire length times activity factor remains the same regardless of the tile size. The left side of the following equality shows the BW of a packet-routed tiled chip, and the right side shows that of the base single tile chip.

$$LAF \times (N_{link} \times N_{tile} \times L) \quad = \quad AF \times \sum_{l=2L}^{2M} N(l)l \quad (12)$$

where LAF is the activity factor on link wires, and N_{link} is the number of wires between two adjacent routers.

The following equations show the total inter-tile wire power. The total inter-tile wire power increases as the tile size decreases at the same rate as the decrease of the intra-tile wire power.

$$P_{inter,switching} \propto LAF \times N_{tile} \times N_{link} \times L \quad (13)$$
$$\propto AF \times (M^{2p-1} - L^{2p-1}) \quad (14)$$
$$P_{inter,leakage} \propto N_{tile} \times N_{link} \times L \quad (15)$$
$$\propto \frac{M^{2p-1} - L^{2p-1}}{MF} \quad (16)$$

Figure 7 shows the number of IO wires per tile (N_{IO}) and N_{link} and the total length of inter-tile wires while varying the tile size. While the total length is exponentially increasing as the tile size decreases, N_{link} and N_{IO} are maximized when the tile size is 5 mm.

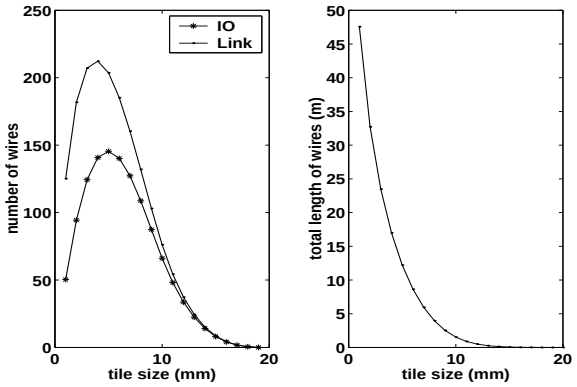

Figure 7: Number and total length of inter-tile wires.

We assume a low-latency virtual channel router [16]. Virtual-channel flow control maintains high throughput even when the packet traffic is high. The inter-tile wire latency becomes relatively significant since the router has a low-latency. A router design can be divided into three main components: input and output packet storage, switch fabric, and arbiters. The power consumption of arbiters is relatively insignificant and thus is ignored here [19]. We choose a matrix crossbar for a switch fabric implementation because it is more common and low power. Since the number of ports for the switch is rather large, the wires dominate power consumption of the crossbar [22], and we ignore power consumed by the internal switches of the crossbar. Table 3 describes the router parameters we assumed. Buffers are implemented with SRAM cells. We fix phit size and the size of routers, and instead allow multiple routers per a tile rather than one large router. In case that the number of IOs between routers or between a tile and the router connected exceeds the phit size, multiple phits are sent simultaneously through multiple fixed-sized routers.

The following equations show that the total power of routers, P_{router} grows even faster than the total power of inter-tile wires, P_{inter} as the tile size (L) decreases.

Phit (bits)	32
Number of input ports	5
Number of output ports	5
Number of virtual channels per physical channel	2
Number of input buffers per virtual channel	4

Table 3: Virtual channel router parameters. Phit is the physical transfer size of the link.

$$P_{router} \propto N_{tile} \times N_{link} \quad (17)$$
$$\propto \frac{M^{2p-1} - L^{2p-1}}{MF \times L} \quad (18)$$

Figure 8 shows the power consumption of a packet-routed tiled chip. We assume zero-power routers and vary multiplexing factor (MF) and activity factor (AF). We can see that smaller and hence more numerous tiles results in large achievable power reduction, as more global wires are replaced with fewer and thus lower-leakage network wires. When the leakage is more dominant (AF=0.1%) and more multiplexing is employed (MF=25), more inter-tile wire power is saved.

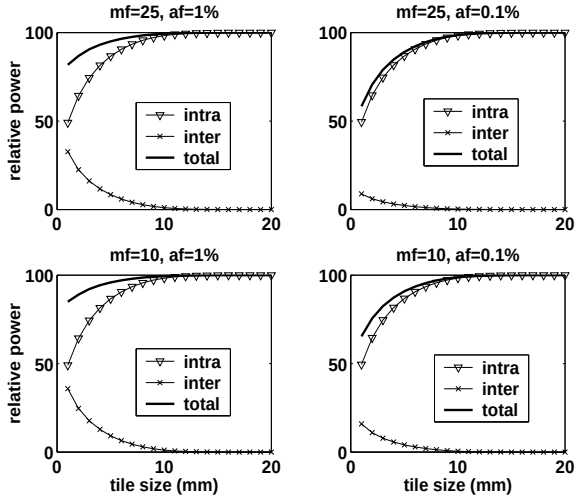

Figure 8: Packet-routed ASIC with ideal zero-power routers.

Figure 9 shows the power saving by pipelining. When MF is 25 and AF is 1%, more than 35% of power can be saved through pipelining the network wires.

Figure 10 and Figure 11 show the power consumption when considering the power overhead of routers. In all cases, the power saving is limited by the energy cost of the routers. The router overhead limits peak wire power savings to around 0–20% with optimal tile sizes of around 2 mm.

5. CONCLUSIONS

We have developed a system-level interconnect power model that predicts the power savings possible by moving global traffic onto a power-optimized on-chip network. The switch to packet-routed on-chip networks has many advantages over wire-routed circuits, but we show that large power reductions are unlikely due to router power overheads. A tile size of around 2 mm is optimal in a 70 nm technology, balancing global wire power reduction with

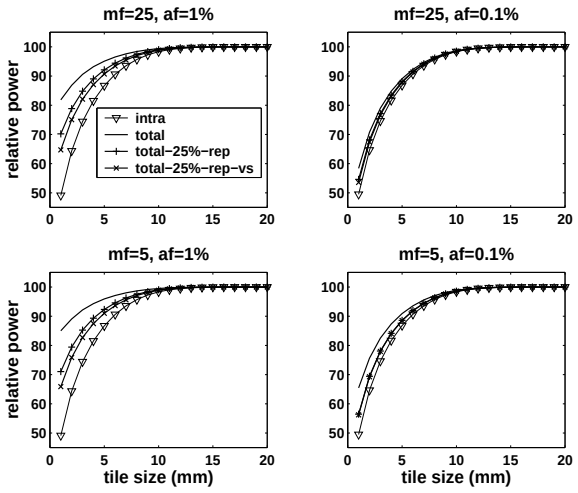

Figure 9: Packet-routed ASIC with ideal routers: power saving by pipelining.

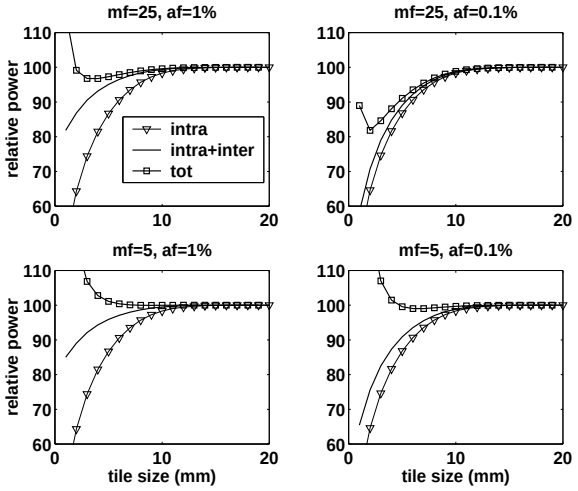

Figure 10: Packet-routed ASIC with real routers: power consumption.

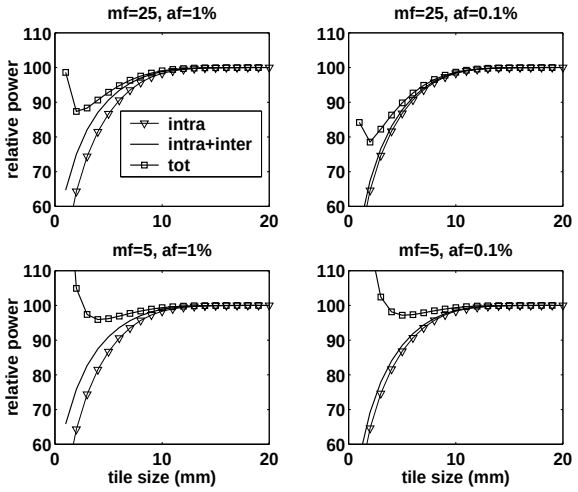

Figure 11: Packet-Routed ASIC with real routers: power saving by pipelining.

router overhead. Additional work is needed to develop low-power on-chip router units.

6. REFERENCES

[1] H. B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. Addision-Wesley, 1990.

[2] K. Banerjee and A. Mehrotra. Power dissipation issues in interconnect performance optimization for sub-180 nm designs. In *Symposium on VLSI circuits*, pages 12–15, June 2002.

[3] A. Chandrakasan et al. Low-power CMOS digital design. *IEEE JSSC*, 27(4):473–484, Apr. 1992.

[4] X. Chen and L. Peh. Leakage power modeling and optimization in interconnection networks. In *ISLPED*, pages 90–95, 2003.

[5] P. Christie and D. Stroobandt. The interpretation and application of rent's rule. *IEEE TVLSI*, 8(6):639–648, Dec. 200.

[6] P. Cocchini. Concurrent flip-flop and repeater insertion for high performance integrated circuits. In *ICCAD*, pages 268–273, Nov 2002.

[7] W. Dally and B. Towles. Route packets, not wires: On-chip interconection networks. In *DAC*, pages 684–689, 2001.

[8] Device Group at UC Berkeley. Predictive technology model. Technical report, UC Berkeley, 2001. http://www-device.eecs.berkely.edu/ ptm/.

[9] N. Eisley and L. Peh. High-level power analysis for on-chip networks. In *CASES*, Sept. 2004.

[10] International Technology Roadmap for Semiconductors. 2004 update. Technical report, ITRS, 2004.

[11] P. Gupta et al. A high-level interconnect power model for design space exploration. In *ICCAD*, pages 551–558, Nov 2003.

[12] S. Heo and K. Asanovic. Power-optimal pipelining in deep submicron technology. In *ISLPED*, pages 218–223, 2004.

[13] R. Ho et al. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, Apr. 2001.

[14] P. Kapur et al. Power estimation in global interconnects and its reduction using a novel repeater optimization methodology. In *DAC*, pages 461–466, 2002.

[15] W. Liao and L. He. Full-chip interconnect power estimation and simulation considering concurrent repeater and flip-flop insertion. In *ICCAD*, pages 574–580, Nov 2003.

[16] R. Mullins et al. Low-latency virtual-channel routers for on-chip networks. In *ISCA 31*, pages 188–197, June 2004.

[17] M. Sgroi et al. Addressing the system-on-a-chip interconnect woes through communication-based design. In *DAC*, 2001.

[18] H. Wang et al. Orion: A power-performance simulator for interconnection networks. In *MICRO*, pages 294–305, Nov. 2002.

[19] H. Wang et al. A power model for routers: Modeling alpha 21364 and infiniband routers. *IEEE Micro*, 23(1):26–35, Jan/Feb 2002.

[20] S. Yang et al. Scaling and integration of high performance interconnects. In *MRS Symposium on Advanced Interconnect*, Apr. 1998.

[21] M. Yazdani et al. Microprocessor pin predicting. *IEEE Circuits and Devices Magazine*, 13(2):28–31, Mar. 1997.

[22] T. Ye et al. Analysis of power consumption on switch fabrics in network routers. In *DAC*, pages 524–529, 2002.