A Survey of Techniques for Power Aware On-Chip Networks.

Samir Chopra Ji Young Park

May 2, 2005

1. Introduction

On-chip networks have been proposed as a solution for challenges from process technology scaling. As technology scaling lowers signal integrity and increases delays in on-chip interconnect [6], global wiring and buses are becoming unfeasible as on-chip interconnect fabric. On-chip networks are fast replacing them because of its advantages of structure, performance, and modularity [4].

While performance has been the primary concern in the design of on-chip networks due to the tight delay requirements of communication, increasing requirements for system power consumption due to fast growing mobile platforms and increasing processor power are imposing constraints on power consumption. As the demand for interconnection bandwidth increases, on-chip networks are taking a considerable portion of total system power. In a system like RAW, which is heavily dependent on on-chip interconnect communication, on-chip networks consume up to 36% of total power. Hence, the power-driven design of on-chip network is gaining increasing importance.

Although there is a compelling reason for power efficient on-chip networks, there has not been much research about power models for on chip networks besides the circuit level techniques due to the relatively recent emergence of on-chip networks. Most research is done at the circuit level such as low-swing signaling and bus encoding schemes, which have been actively researched in other fields such as off-chip networking, and global wiring. Therefore, in this paper, we focus on the research that investigates power consumption of an on-chip network from other perspectives: topologies and architectures.

2. Architecture level techniques

Wang et al. present and analyze a set of microarchitectures from power driven perspective [1]. They first introduce an equation that characterizes the energy consumed when a data flit is transmitted into two components as:

$$E_{flit} = E_R \cdot H + E_{wire} \cdot D$$

where E_R is average router energy which is composed of input buffer energy, crossbar traversal energy, and arbitration energy, H is the number of hops traversed by the flit, E_{wire} is average link wire transmission energy per unit lengths with optimally placed repeaters, and D is the Manhattan distance between source and destination. By characterizing the power consumption of existing on-chip networks, RAW and TRIPS, they point out that unlike off-chip networks, where link power dominates, router power contributes a significant portion of on-chip network power. Therefore, as architectural techniques that reduces E_R and H, which compose the router power, they propose segmented crossbar, cut-through crossbar, and write through buffer and also study the power saving potential of an existing architecture; express cube.

Segmented crossbar, a simplified application of segmented buses, reduces the switching capacitance of input and output lines of the crossbar by dividing the lines into segments using tri-state buffers. The power saving by reduced capacitance reaches 37.5% of the total power consumed by the crossbar when the lines are divided into 4 segments if we ignore overheads due to segment tri-state buffers that split the lines. While increasing the number of segments per line reduces the capacitance of each segmented line, it trade off the reduced capacitance for the increased overhead per segment; the capacitance of a tri-state buffer becomes more significant compared to the capacitance of the segmented line. Therefore, they conclude that adding more segments beyond 2 or 3 does not give you more power savings.

Cut-through crossbar, on the other hand, focuses on network traffic patterns and routing protocols in on-chip networks, and it optimizes for the common case. They assume that straight-through traffic is the common case for on-chip networks, and cut-through crossbar provides reduced power and improved performance for the

straight-through traffic by providing direct connection between each input port and the output port of the opposite direction at the expense of the connectivity to other directions. Cut-through crossbar incurs additional delays when two flits from different inputs use a common line segment, which happens when the inputs are from different dimension and they both want to turn. However, with a simple dimension-ordered routing protocol, the conflict on the same line segment does not occur and, thus, there is no penalty using cut-through crossbars. Even with a worst-case routing algorithm the latency of the crossbar increases by 53%, which is a tolerable trade-off for energy-delay product assuming that the latency can be amortized over pipeline of the router.

Write-through buffer targets another component in router energy – input buffer energy. It is a revised architecture of a bypass input buffer and removes buffer read energy when bypassing can be done. Although the ideal buffer power saving can reach 60% of the total power consumed by the buffer, it can be only achieved with very low flit arrival rate, which is unlikely to be the common case. Therefore, in reality write-through buffer achieves much less power savings than the upper bound. Write-buffer incurs negligible performance overhead due to the muxes, and area overhead due to the additional registers and muxes.

Finally, express cube is considered as an architecture that saves power by reducing average hop count. Express cube is a k-ary, n-cube augmented by one or more levels of express channels that connect non-neighboring nodes and allow non-local messages to bypass nodes. Since non-local messages can bypass nodes, the hop count can be reduced compared to 2D torus architecture. Express cube presented in the paper has one level of express channels, and the reduction of the average hop count varies from 10% to 60% according to network size and express channel interval. Although an express node in an express cube consumes more power and takes more areas since it has twice as many ports as a local node, the increased power is compensated by the reduction of hop count, and the area overhead can be compensated by reducing flit size.

2.1 Critique

This paper introduces an approach to reduce power consumption of on-chip networks and is successful to present some router microarchitectures with substantial power savings. While the overall analysis is well designed, there are a few problems.

The authors evaluate power efficiency of the proposed microarchitectures with uniform random traffic. Although uniform random traffic simplifies the model, the analysis may not apply to real traffics. One example is locality of the network traffic. To minimize the latency, applications try to exploit locality in their traffic and, thus, real traffic tends to have more locality than uniform random traffic, and this locality may degrade the efficiency of the express cube, which has the greatest improvement among all microarchitectures. As a model for the real traffic, they provide the simulation result of TRIPS. However, since TRIPS is a fine-grained CMP, which generates small, frequent traffics due to the fact that it transports operands instead of cache lines or user messages, its traffic pattern may differ from that of other system-on-a-chip or coarse-grained CMPs, and lack of the simulation result of RAW without any proper explanation lowers the credibility of using TRIPS as a model for real applications.

While analyzing the power savings of their proposed microarchitectures and the express cube, the authors ignore the overheads due to increased complexity of the router architecture to simplify their models and make the power savings obvious. Although they mention some of the overhead and in what cases they should be considered, they fail to provide any detailed analysis about the effects of the overheads on their models, and some overheads such as increased buffer energy for cut-through crossbar due to the increased latency of the crossbar, performance impact of the muxes on the write-through buffer, and increased complexity of the router architecture in the express cube are not considered.

The authors seem to make a false judgment while they are defining the simulation model for express cube. They argue that using a flit size of 2/3 of the flit size used by other models keeps the bisection bandwidth of express cube equal. However, since an express channel spans over 2 local channels, every channel, instead of every other channel, is augmented by an express

channel. Therefore, the flit size should be reduced to half the size of the flit size of other models. Due to the failure in making a fair adjustment of the flit size, express cube has larger throughput than the baseline configuration model and shows larger saturation throughput in the simulation. When the flit size is adjusted to equalize the bisection bandwidth, express cube can have even smaller throughput than the baseline configuration model due to the load imbalance between express channels and local channels.

3. Topology

Interconnect network topology choices have a high impact on overall power consumption of the network. When an on-chip network is thoughtfully designed with low power in mind, the designers may overlook how technology scaling may change what the most energy efficient network configuration is. Technology scaling leads to an overall decrease in network power, but changes the parameters of the most optimal topology.

This paper uses an analytical model to predict the most energy efficient topology at three different technologies.

The metric used throughout the paper is E_{flit} – the average flit traversal energy, as mentioned earlier. This metric incorporates dynamic power, including both wire and transistor capacitance effects, and static power.

$$E_{flit} = E_R \cdot H + E_{wire} \cdot D$$

Static power is included by evenly distributing the amount of static power dissipated in a router among all flits which pass through that router in a given cycle.

The study in this paper hinges on hop count, as this is the parameter which can be adjusted with different topologies. All of the analytical models assume uniform random traffic, and two different types of workloads: constant and linear. With a constant workload, the number of flits does not vary with increasing number of router ports, and with a linear workload, the number of flits increases as the number of router ports increase.

The three types of topologies considered are high-dimensional meshes/tori, hierarchal meshes/tori, and express cubes. Investigations into the topologies were done in several steps: Derivation of the average hop count of the network, average flit traversal energy, and minimal network size at which the topology becomes more energy efficient than a two dimensional tori. Then the most energy efficient configuration of the particular topology is found, and how this optimal configuration varies with process technology.

	Linear load		Constant load	
buffer size	4-flit	16-flit	4-flit	16-flit
0.1μm	53, 8	44, 7	61, 8	51, 8
70nm	49, 7	42, 7	64, 8	55, 8
50nm	46, 7	39, 7	82, 10	73, 9
35nm	46, 7	36, 6	165, 13	140, 12

Table 1: Minimal network size for high-dimensional tori, in the form of (M, Nmin)

	Linear load		Constant load	
buffer size	4-flit	16-flit	4-flit	16-flit
0.1μm	2, 16	2, 13	2, 20	2, 16
70nm	2, 14	2, 12	2, 22	2, 18
50nm	2, 13	2, 11	2, 46	2, 32
35nm	2, 13	2, 10	3, 28	3, 23

Table 2: Minimal express interval and corresponding minimal network size for hierarchical tori, in the form of (v, Nmin)

	Linear load		Constant load	
buffer size	4-flit	16-flit	4-flit	16-flit
0.1μm	3, 11	3, 10	3, 12	3, 10
70nm	3, 10	3, 9	3, 12	3, 11
50nm	3, 10	3, 9	4, 14	4, 13
35nm	3, 10	3, 9	4, 21	4, 19

Table 3: Minimal network size and corresponding express interval for hierarchical tori, in the form of (v, Nmin)

First two-dimensional tori and meshes are analyzed, and then higher dimensional topologies are compared to the two-dimensional configurations. When comparing a two-dimensional torus with a mesh of the same size, the two dimensional torus consumes 25% less router energy but has 50% more link energy compared to an equal size 2-D mesh. Extending the analysis to higher dimensions, topologies where three dimensional tori outperform two-dimensional are found and noted on table 2. The same table also shows that the workload analysis is important. A linearly loaded network requires a smaller size to benefit from high dimensional tori. Also, larger buffer size requires smaller network size to benefit.

Next, hierarchal meshes and tori are analyzed, and the nominal express interval and

	Linear load		Constant load	
buffer size	4-flit	16-flit	4-flit	16-flit
0.1μm	2, 20	2, 17	2, 24	2, 20
70nm	2, 18	2, 16	2, 26	2, 22
50nm	2, 17	2, 15	2, 50	2, 36
35nm	2, 17	2, 14	3, 25	3, 22

Table 4: Minimal express interval and corresponding minimal network size for express cubes, in the form of (v, Nmin)

	Linear load		Constant load	
buffer size	4-flit	16-flit	4-flit	16-flit
0.1μm	3, 13	3, 13	3, 14	3, 13
70nm	3, 13	3, 12	3, 14	3, 14
50nm	3, 13	3, 12	4, 15	4, 15
35nm	3, 13	3, 12	4, 19	4, 18

Table 5: Minimal network size and corresponding express interval for express cubes, in the form of (v, Nmin)

torus dimension	2	3	4	5	6
70nm	**1.59**	1.79	1.98	2.44	N/A
50nm	1.98	2.14	2.21	2.67	2.83
35nm	3.53	3.88	3.40	4.03	4.27
topology	H-2	H-3	H-4	E-2	E-3	E-4
70nm	1.82	1.78	1.88	1.96	1.98	2.15
50nm	2.07	**1.94**	1.97	2.19	2.12	2.22
35nm	3.47	3.14	**3.01**	3.63	3.38	3.42

Table 6: Simulated average flit traversal energy(10^{-10}J) of tori, hierarchical tori and express cubes, with the minimal energy at each technology highlighted.

corresponding minimal network size for the hierarchal tori to achieve better energy efficiency than a two dimensional torus. The results are given in table 2 and table 3.

Finally, express cubes are analyzed and corresponding results are given in table 4 and table 5.

After digesting the results of the analysis, a few results are found. Both hierarchal tori and express cubes perform well however the choice between them depends on which load model is closer to reality. In most cases hierarchal tori outperforms high dimensional tori due to longer average link length. Express cubes can save energy over high dimensional tori; however it depends on the amount of increased router complexity which is added in this topology.

A real design is studied in order to test topology configurations predicted by the analytical model. TRIPS is a 5x5 mesh-like network connecting ALUs and memory components and is designed with 0.1 micron technology. Table 6 summarizes the average flit traversal energy with several different topologies and technology scales. The results show that the analytical model is accurate.

A more realistic traffic pattern is used to test the analytical model, including LU, water, and mp3. It was noted earlier that a hierarchal three-dimensional torus is the optimal topology at 50nm with random traffic, but the results show that a hierarchal two-dimensional torus is more optimal for LU and water, and a regular two-dimensional tours results in the smallest flit traversal energy for mp3.

3.1 Critique

The problem of finding the right interconnect topology for low-power is important as network sizes scale. A lot of research is done to find the optimal balance between throughput, number of nodes, and power consumptions. It is important that this work does not go to waste as technology scales to the next generation. Scaling effects physical properties of devices and can lead to deviation from the optimal power-efficient topology. The problem is relevant to low-power on-chip network design. However it does have a few flaws.

The paper uses uniform random traffic in their analytical models. However, real traffic patterns are not uniform random. In fact, it is very likely that most traffic from a node will be sent to one its neighbors. We have already seen real results deviate from the predicted model with real-life patterns presented by LU, water, and mp3. It would be nice to see some worst case traffic patterns as well and how the model predictions change given these traffic patterns.

The scope of the paper is limited to a few basic topologies: only including meshes, tori, and variants thereof. Perhaps a few other topologies such as butterfly networks and fat trees should be included as well. The paper mentions that meshes and tori are easily routed on a two dimensional board, but some networks may require the load balancing offered by fat trees and butterfly networks. Models for these networks should be created and tested as well.

There seems to be some a few discrepancies in Table 6. This table shows the results of scaling the TRIPS network to 70nm, 50nm, and 35nm. The average flit traversal energy increases as technology scales. It is mentioned that the size of the network is increased as technology scales.

However, nothing is mentioned about the new traffic patterns or some explanation as to why there is an increase in average flit traversal energy. It also scales the link traversal energy using the equation EL = 0.53Er5, where Er5 is the average five port router traversal energy. It is mentioned throughout the paper that energy does not scale linearly with technology. Perhaps the 0.53 parameter should be revised for each new technology.

Despite a few shortcomings in the results, this paper does a good job in presenting a new problem in low-power interconnect design. Topologies which are low power in one technology may not be low power as we move to the next generation of devices. The analytical model presented gives designers a useful tool in predicted how network designs will need to be modified in order to maintain an optimal low-power topology.

I would address the problem in a similar way, but also include some of the additional elements mentioned above. The paper has extensive results and brings into light a new type of problem that may not have been thought about before. The mathematical analysis is useful for a designer to would like to predict how his topology choice will perform in future technology generations. Including some different traffic patterns and topologies would be the only additions I would add to this work.

4. Power & Performance

Networks-on-chip are a solution for future high performance multiprocessor chip communication. In order to design networks, which have both high performance and optimally low power consumption, a useful model needs to be created and evaluated. The paper introduces a VHDL based model for evaluating latency, throughput, and power consumption of a network-on-chip.

The power-performance tradeoff analysis of on-chip networks can be performed by varying the type of topology, length and width of wires, buffer sizes, switching techniques, router algorithms, and quality of service. Past models, such as Orion, do not consider power consumption due to coupling capacitance and internal control units. This model addresses these shortcomings.

Switching	Energy (in fJ)		
	100 μm	1000 μm	5000 μm
(000-000), (001-001),(010-010), (011-011), (100-100), (101-101),(110-110), (111-111)	0	0	0
(000-001), (000-100),(001-101), (010-011), (010-110), (011-111),(100-101), (110-111),	0.122	5.25	99
(000-010), (001-011),(100-110),(101-111)	0.189	12.48	213
(000-011), (000-110),(001-111), (100-111)	0.1914	5.94	121
(000-101), (010-111)	0.258	13.17	235
(000-111)	0.2075	20.46	66
(001-010), (010-100),(011-101), (101-110)	0.4314	29.54	504
(001-100), (011-110)	0.2309	7.83	165
(001-110), (011-100)	0.42	20.31	378
(010-101)	0.6864	48.9	830

Table 7: 3 wire-set characterization

Individual wires are modeled by a distributed RLC analysis. Cross coupling is measured by pairing two and three wires together, while distributed RLC effects are measured with a single wire model. Table 7 summarizes the energy consumption for the different wire models at three different lengths.

The network used in the analysis is a 4x4 mesh based dimension-routing architecture. Each router consists of five unit routers to forward traffic in two X directions, two Y directions, and the processing node. Messages are partitioned into flits: a head flit, a tail flit, and body flits to ensure efficient data transfer. Two switching techniques, virtual cut through and wormholing were used to compare the performance of each method. Each component is modeled separately using RTL models and larger components are characterized in terms of smaller functional units.

For the performance/power tradeoff analysis, uniform random traffic was injected into the network. The network was allowed to stabilize before measurements were taken.

Figure eight shows virtual cut through generally has lower latency and higher packet acceptance rate than wormholing. Virtual cut through consumed more power as shown by figure nine; however the performance gains outweigh the additional power needed for this switching technique.

The component which consumes the most power in their network is the virtual channel buffers followed by header decoder and link controllers. These components also show larger increases in power dissipation as packet injection rate increases. Average leakage power decreases as packet injection rate increases because the

power lost is amortized over more packets.

4.1 Critique

Since this is a paper describing a model rather than solving a problem, it is understandable that no suggestions are made for low-power/performance tradeoffs. However, the paper did not use any mathematical analysis to justify their results found in simulations. It is difficult to justify the results found in this paper without some concrete hand analysis to reinforce the results found through simulation.

The paper also did not try to adjust their topology or devices in order to observe how this would effect the change in power consumption. The paper mentions several different parameters which can be adjusted for a power-performance tradeoff, including topology, length and width of physical links, buffer allocation, switching techniques, routing algorithms, and levels of service. However, the only tradeoff made was routing algorithm (wormholing/virtual cut through). Additional tradeoffs should be compared and documented.

Uniform random traffic is used in their analytical models. However, real traffic patterns are generally non-uniform. In fact, it is very likely that most traffic from a node will be sent to one its neighbors in a mesh topology. Different traffic patterns should be looked at as well, including worst case scenarios and real-life traces. The results of the analysis done may be very different under a non-ideal traffic pattern.

Despite the lack of additional analysis other than what is presented in the results, this paper does have a better wire model than what has been done in other modeling and analysis. It includes more detail including cross-coupling effects, distributed RLC modeling, and observes the effects of varying the length of the wires. This is a good contribution to the area of low-power interconnects since wire energy contributes to the overall power dissipation.

I don't think I would have addressed the problem of creating a good power/performance model as done in this paper. It lacked a solid mathematical analysis which is important in understanding the reasoning behind the results observed. The paper's only strong point was the extensive wire modeling. I would have used a different approach of extending an existing model rather than creating an entirely new one. Very little is gained by repeating work done my others only to include a few more features.

5. Conclusion

In this paper, we reviewed some power analysis models for on-chip networks, and techniques for power efficient on-chip networks. The models presented address tradeoffs of making certain design choices from power driven perspective, and one can use these models for designing a power efficient on-chip network. Wang et al. shows in [1] that router energy takes a substantial portion of total network power, and presents four network microarchitectures as techniques that significantly reduce the router energy and represent power efficient architectures. In [2], Wang et al. points out the impact of technology scaling on selecting topologies, and present a model for analysis of topology with technology scaling in mind. Banerjee et al. present a cycle accurate performance and power model for on-chip interconnection network in [3].

These models, however, do not provide complete roadmap for on-chip networks. There are still other topologies, traffic patterns, router architectures, and other aspects of network design to be considered, and routing protocols is also one of the aspects to be considered.

References

[1] H. Wang, L.-S. Peh, and S. Malik. Power-driven design of router microarchitectures in on-chip networks. *In Proc. International Symposium on Microarchitectures*, November 2003.

[2] H. Wang, L.-S. Peh, and S. Malik. A technology-aware and energy-oriented topology exploration for on-chip networks. *In Proceedings of the Design Automation and Test in Europe Conference*, March 2005.

[3] N. Banerjee, P. Vellanki, and K. S. Chatha. A power and performance model for network-on-chip architectures. *Proceedings of Design Automation and Test in Europe Conference*, pages 1250-1255, February 2004.

[4] W.J.Dally and B. Towles. Route packets, not wires: On-chip interconnection networks. *In Proc. Design Automation Conference*, pages 684-689, 2001.

[5] W.J.Dally. Express cubes: Improving the performance of k-ary n-cube interconnection networks. *IEEE Transactions on Computers*, 40(9):1016-1023, 1991.
[6] D.Sylverster and K.Keutzer. Impact of small process geometries on microarchitectures in systems on a chip. *Proceedings of the IEEE*, pages 467-484, April 2001.