

- straints,” *Journal of Parallel and Distributed Computing*, no. 12, 1991.
- [2] A. Agarwal, “Limits on interconnection network performance,” *IEEE Transactions on Parallel and Distributed Systems*, October 1991.
  - [3] Abdellatif Bellaouar, Mohamed Elmasry, *Low-Power Digital VLSI Design: Circuits and Systems*, Kluwer Academic Publishers, Norwell, Massachusetts, 1995.
  - [4] R. W. Broderson and A. P. Chandrakasan, *Low power digital CMOS design*, Kluwer Academic Publishers, Massachusetts, 1995.
  - [5] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, S. M. Watt, *Maple V*, Waterloo Maple Publishing, Virginia, 1991.
  - [6] W. J. Dally, “Performance Analysis of k-ary n-cube interconnection networks,” *IEEE Transactions on Computers*, June 1990.
  - [7] W. J. Dally and C. L. Seitz, “Deadlock-free message routing in multiprocessor interconnection networks,” *IEEE Transactions on Computers*, Vol. C-36, no. 5, pp. 547-553, May 1987.
  - [8] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks, An Engineering Approach*, IEEE Press, 1997.
  - [9] D. Lin and C. Svensson, “Trading Speed for Low Power by Choice of Supply and Threshold Voltage,” *Journal of Solid State Circuits*, Vol. 28, No. 1, Jan. 1993.
  - [10] S. L. Scott and G. M Thorson, “The Cray T3E networks: adaptive routing in a high performance 3D torus”, Proceedings of Hot Interconnects, August 1996.
  - [11] S. L. Scott and J. R. Goodman, “The Impact of Pipelined Channels on k-ary n-cube networks”, *IEEE Transactions on Parallel and Distributed Systems*, vol. 5, no. 1, pp. 2-16, January 1994.
  - [12] J.P. Uyemura, *Circuit Design for CMOS VLSI*, Kluwer Academic Publishers, Norwell, Massachusetts, 1995.
  - [13] C. S. Patel, S. M. Chai, S. Yalamanchili, D. E. Schimmel, “Power/Performance Trade-offs for Direct Networks,” *Parallel Computer Routing and Communication Workshop*, June 1997.

dimension and then find the optimum power distribution. Alternatively, one can fix the power distribution and find optimum network dimension.

Both of these approaches are equally valid in achieving maximum network performance and they address questions of the degrees of freedom permitted to the system and the hardware designers of the network. For example, from our results we find that small systems are insensitive to the topology if  $P_w$  is very small providing greater flexibility in the design of network topology. On the other hand, the network performance is insensitive to  $P_w$  at small values of dimension which permits increased flexibility to the hardware designer in distributing the total available power between switches and the wire. In other words, placing a constraint on power distribution provides flexibility in the choice of network topology and placing a constraint on the topology can realize greater flexibility in designing for the power distribution between the switch and wire. It is often the case in the practical design of systems that we are not only limited by the overall power constraint, but also by the design choices we can make. In such cases the above analysis can be used as a guide to high performance network design.

## 6.0 Concluding Remarks and Directions for Future Research.

This paper has reported on our study of the relationships between a fixed power budget, and the network topology from the point of view of message latency. When a higher percentage of the power dissipation is in the wire, the reduced per node switching power penalizes small to modest sized networks (up to 1K-2K nodes). For larger networks, the reduced available power/node begins to favor larger dimensions since the number of intermediate nodes is substantially reduced. We see a general trend favoring higher dimensional networks as available power becomes more constrained and system sizes increases. These models only begin to address the major issues in the power constrained design of multiprocessor interconnection networks. The major issues that we will address in the future include the following.

- The development and incorporation of workload models will enable the analysis of the effect of various applications characteristics such as message injection rate, distribution of the destination addresses, etc.
- We will instrument a detailed network simulator to incorporate power estimation models. This simulator will then be used to host applications (or communication traces derived from these applications) to more accurately estimate power savings due to proposed architectural techniques.
- We are interested in generalizing the models to constraints on energy rather than power. Such models will prove useful when there is a limited amount of energy available (e.g., battery) vs. a virtually unlimited amount of energy at a much lower rate (e.g., solar energy). These extensions are focused on applications to onboard computations for deep space missions.

The expectation is that these models can aid the designers of embedded multiprocessor systems in power constrained environments.

## 7.0 References

- [1] S. Abraham and K. Padmanabhan, "Performance of multicomputer networks under pin-out con-

Note that the vertical axis represents decreasing latency. Thus a maxima on the surface corresponds to minimum latency. The overall latency as a function of network dimension and power dissipated in the wire is illustrated in Figure 8. For a small size system of 16 nodes, we found that low (two) dimensional networks are preferred to achieve minimum latency operation. As the size of the network grows from 16 nodes to 256 nodes, the minima in the network latency plot is shifted towards higher (four or five) dimensions. As a general trend, higher dimensional networks are preferred to yield minimum latency as we increase the size of the system. The reason for this appears to be as follows.

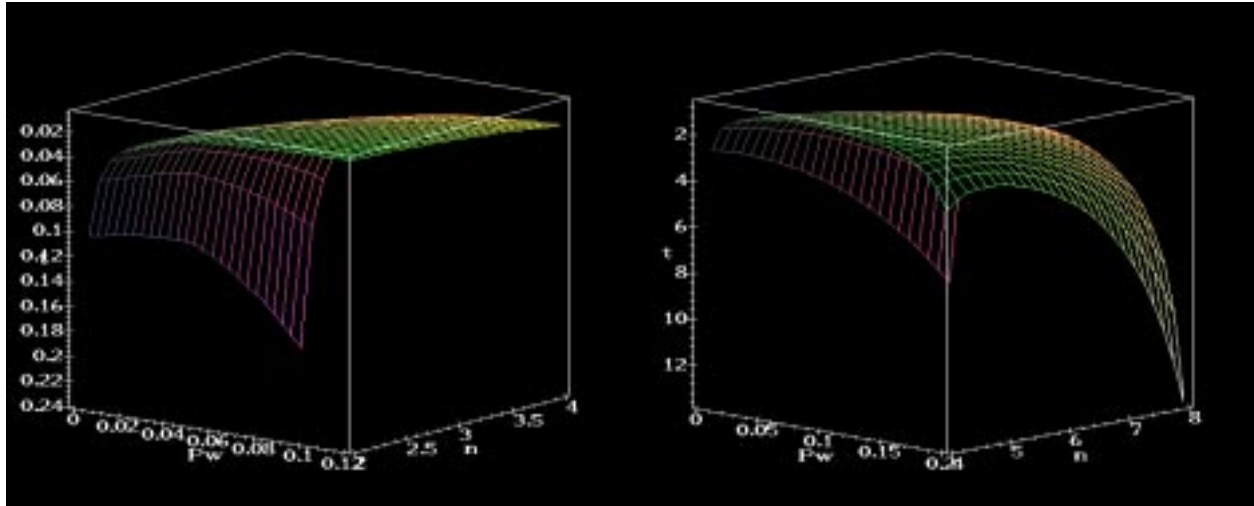
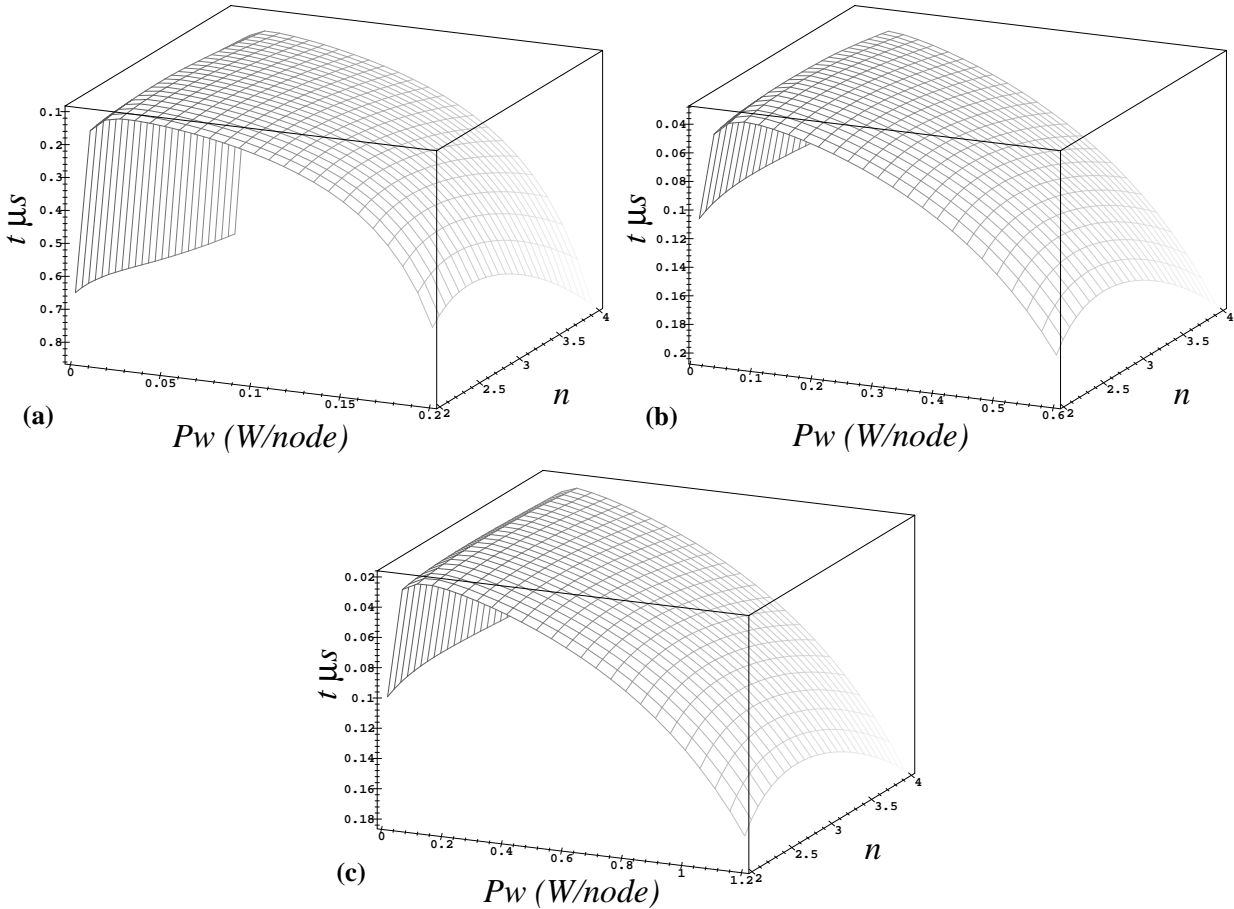


Figure 8. Message latency for (a) 16 nodes (b) 256 nodes system sizes

Messages require a fixed amount of energy to be transmitted between a pair of nodes. With decreasing dimension, the average distance between nodes grows exponentially as  $O(N^{1/n})$ . This naturally increases the switching delays experienced by a message. Furthermore, with a fixed amount of power available per node and longer average distances, the power available at each node is less. Therefore channel widths must become smaller and thus latency increases. A closer look at the graphs for larger networks will show that, when a larger percentage of the power is spent in the wire, the latency penalty for low dimensional networks is particularly severe. This is due to the fact that less switching power is available in the nodes and the preceding phenomenon becomes much more pronounced. As power constraints become more stringent and systems grow in size we expect to see a pronounced shift towards higher dimensional networks. Other observations and design considerations due to power distribution is discussed in the following section.

### 5.3 Distribution of Power Between the Switch and Wire

The preceding analysis focused on minimizing latency, while we noted the distribution of power between switching messages vs. driving them across the channels. What guidelines can we infer for designers of the channel drivers and internal router switch designs? A closer examination of the surfaces in Figure 7 and Figure 8 reveals several options for achieving minimum latency. One approach is to choose a particular



**Figure 7. Message Latency under power constraint of (a) 0.25 watts/node, (b) 1 watts/node, (c) 2 watts/node**

stantial as increased power dissipation takes place in the physical channels. This important observation can be explained as follows. The overall latency is a function of the delay through switch and wire, which in turn are functions of how power is distributed among them. If power consumed in the switch is larger, under a fixed power constraint the wire delay will dominate. Alternatively, if the power consumed in the wire dominates, the switch delay will dominate the latency. This trade-off in power distribution is clearly described by the model and represented by the surfaces in Figure 7. The exact representation of the distribution of total power between the switch and the wire will certainly depend on application parameters such as message injection rates. In an extended study of including message injection rate into this model [13], we found that the minimal latency consistently occurs when approximately 60% of the available power is dedicated to switching power dissipation and 40% to wire power dissipation. Clearly more analysis is required, but such observations are extremely useful for both the system and hardware designers as a rule of thumb in designing networks under fixed power constraints.

## 5.2 Impact of System Size

With a fixed per node power budget, we studied the impact of system sizes such as small (16 node) and medium (256 node) systems on the topology. This analysis utilized a constraint of 0.25 watts per node.

through the switch. We are now equipped with all of the necessary models to be substituted in (1) to study the overall latency through the router as a function of network dimension and power distribution under a power constraint. Thus, the complete expression for the no load latency of a message through a k-ary n-cube network under a maximum power constraint is as given in (1)

$$t_{wormhole} = [D(t_r + t_s + t_w)] + \left[ \max(t_s, t_w) \left[ \frac{L}{W} \right] \right] \quad (18)$$

where  $t_s$  is defined in (12),  $t_w$  is defined in (17),  $W$  is defined in (9), and  $D = (nN^{1/n})/4$  [2,8].

## 5.0 Performance Analysis

The goal of our analysis is to understand the relationship between the topology of the network and message latency under a fixed power budget. The preceding analysis has identified the relationships between power and features of the topology such as switching speeds, channel delays, and channel widths, in addition to numerous physical parameters. We utilized the Maple V symbolic computational package to study these relationships [5]. Fixed power budgets are represented on a per node basis in watts per node. This study focuses on the relationship between three variables: i) message latency, ii) distribution of power between the switches and wire (channels), and iii) network dimension. We are interested in the network dimension that minimizes message latency. The preceding relationship is examined under increasing power budgets and for various system sizes. These results are discussed in the following subsections.

### 5.1 Impact of Available Power

To analyze the impact of various power budgets we examined the behavior of network latency with available power constrained to 0.25, 1, and 2 watts per node. The overall latency for a 256 node network as a function of network dimension and power dissipated in the wire is illustrated in Figure 7. Note that the vertical axis represents decreasing latency. Thus a maxima on the surface corresponds to minimum latency.

As we increase the available power per node for a fixed system size, the dimension at which message latency is minimized shifts towards higher dimensional networks. Higher dimensional networks embedded in the plane lead to longer wire lengths for the inter-router physical channels. As available power is increased, the negative effect (larger power consumption) of these longer wires is reduced. Moreover, the smaller number of switches that are traversed reduces the power consumption in the switches. The net effect is to favor low dimensional networks in low power applications and higher dimensional networks when power is not so constrained.

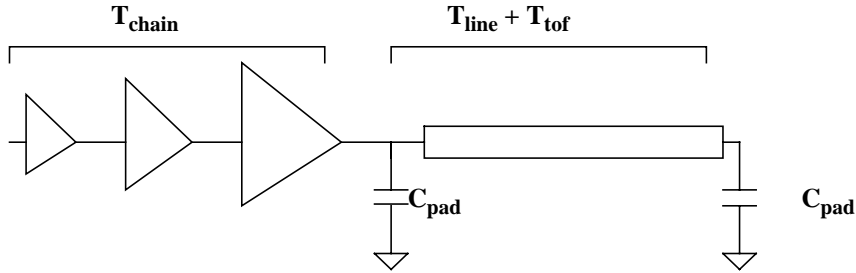
Furthermore, increasing the available power per node produces clear definitions of the distribution of the power between the wire and switch to minimize the overall latency. Analyzing latency as a function of  $P_w$  under a fixed value of dimension, we note that the latency starts to decrease as  $P_w$  increases. However, as an increased amount of power is used to drive the wire, the latency begins to increase and becomes sub-

where  $\beta = \mu C_{ox}(W_1/L_1)$  is the minimum transistor size transconductance parameter [4,12]. (Please refer to Table 1 for a reference to the notation). The total power dissipated in the driver chain is simply  $P_{cap} + P_{sc}$  and is a function of the number of drivers. We can therefore solve the total power equation in terms of the number of drivers, and obtain the following expression.

$$X = \frac{\ln(1 + (\alpha - 1)P_w/\Gamma)}{\ln(\alpha)} \quad (15)$$

where  $\Gamma = (Vdd^2 f(\alpha C_{in} + C_p)) + ((\beta/12)(Vdd - 2Vt)^3(\tau/T_{clk}))$  is the coefficient of the total power term. From (15),  $X$  is function of both the scaling factor,  $\alpha$ , the maximum available power, and the power dissipated in the wire,  $P_w$ . The motivation for this simplification is that given a power budget, we can distribute the power in several ways: a) completely within the switch, b) completely within the wire c) across the switch and the wire. By leaving  $P_w$  as a free variable, we can study all three cases with the number of drivers be limited by the power allocated to driving the wires.

The delay through the wire is modeled as the sum of the delay through the driver chain, the RC line and the time of flight. The model schematic is shown in Figure 6.



**Figure 6. Wire Delay Model**

Assuming the pFET and nFET in the driver are sized to provide equal performance (i.e. nFET = 2pFET), the delay through the driver chain is simply  $T_{chain} = 3\alpha(X - 1)R_o C_o$ , the delay through the line is

$$T_{line} = S_n \left( \frac{R_o}{\alpha^X - 1} \right) (2C_{pad} + C_w + (\alpha^X - 1)C_o), \quad (16)$$

and the time of flight delay is simply  $T_{tof} = w_{avg}/v_p$ . The total delay through the wire is simply the sum of these individual delays and can be written as

$$t_w = 3\alpha(X - 1)R_o C_o + S_n \left( \frac{R_o}{\alpha^X - 1} \right) (2C_{pad} + C_w + (\alpha^X - 1)C_o) + \frac{w_{avg}}{v_p} \quad (17)$$

The delay through the RAU will be assumed to be equal to the delay through the switch ( $t_r = t_s$ ). This is reasonable since modern routers attempt to perform routing decisions in the time it takes to drive a flit

In this manner the analysis is relative to the best that can be achieved for 2- $D$  tori in any given technology. The delay along the wire may follow one of several models depending upon the length and the implementation medium. The three common wire delay models include the linear, logarithmic, and constant delay models. The above expression is based on a linear delay model, and is therefore conservative.

### 4.3 Time delay models

The next step is to derive the expressions for the delays associated with the switch, RAU, and the physical channel between routers, i.e., wire delay. The delay through the switch can simply be found by rearranging (8) ( $t_s = 1/f_s$ ) to obtain,

$$t_s = \frac{0.5K_s n^2 W C_t V_{dd}^2}{P_s} \quad (11)$$

Substitution of  $W$  from (9) into (11) yields,

$$t_s = \left( \frac{P_w}{P_s} \right) \frac{0.5K_s C_t V_{dd}^2 n^2 t_w}{K_l C_w V_{dd}^2 w_{avg}} \quad (12)$$

The wire delay associated with the physical channel between adjacent routers must be obtained very carefully. Normally the connection between two routers are off-chip signal paths and we have to treat this interconnect as a transmission line rather than a  $RC$  line. The geometry of these off-chip wires is “fat” to reduce the resistance of the line. However, the inductance will dominate the signal propagation and it must be taken into consideration. In addition, if we are not otherwise constrained, the number of drivers in a cascaded driver chain is determined by the need to achieve minimum latency across the channel. This is a valid approach for designs where we are not constrained by power. Under power constraints, the maximum number of available drivers may be less than the optimum number of drivers needed to achieve minimum latency. Therefore, in our analysis of low power network designs, we have to constrain the drivers by the maximum allowable power to the network.

There are primarily four types of power consumption mechanisms in the drivers: a) capacitive power, b) short circuit power, c) leakage power and d) static power. In our analysis of the driver chain we will ignore the last two power consumption mechanisms and focus on the capacitive and short circuit power. Assuming monotonic scaling of drivers in the driver chain, the total capacitive and short circuit power dissipation in the driver chain is given as:

$$P_{cap} = V_{dd}^2 f_i (\alpha C_{in} + C_p) \frac{(\alpha^X - 1)}{\alpha - 1} \quad (13)$$

$$P_{sc} = \frac{\beta}{12} (V_{dd} - 2V_t)^3 \frac{\tau}{T_{clk}} \frac{(\alpha^X - 1)}{\alpha - 1} \quad (14)$$

The preceding equations were derived from a system architecture stand point. In order to relate the system architecture parameters to device and technology parameters, we must develop models in terms of power dissipation within the switches and physical channel interfaces, e.g., drivers. We follow the same modeling technique of viewing power dissipation as simply charging and discharging of the total capacitance. In doing so we arrive at the following equations.

$$P_s = \frac{1}{2} K_s M_s C_t V_{dd}^2 f_s \quad (8)$$

$$P_w = K_l C_w V_{dd}^2 f_w w_{avg} W \quad (9)$$

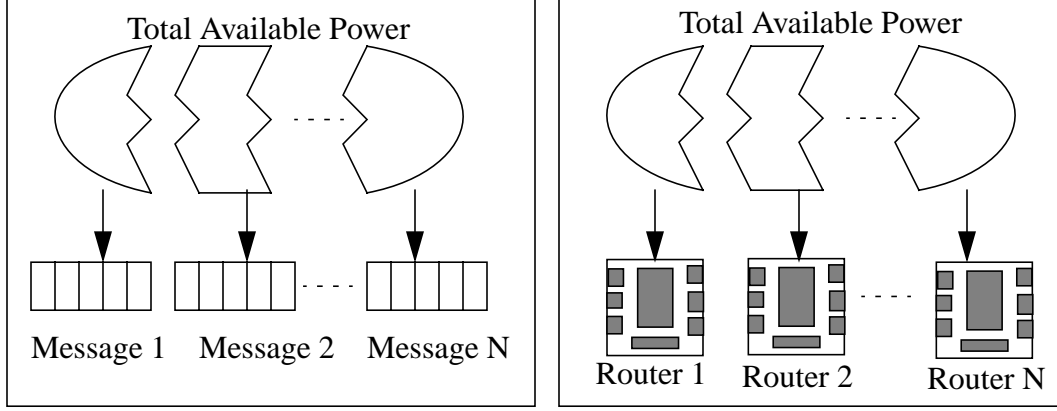
Here,  $M_s = n^2 W$  is the number of transistors in a crossbar switch and  $w_{avg}$  is the average wire length between two adjacent routers.

In VLSI systems the cycle time of the network is typically determined by the maximum wire delay, while the majority of the power is expended in driving these wires. Thus, topologies that can be embedded in two and three dimensions with short wires are generally preferred. While even low dimensional networks may logically appear to have long wires due to the wrap around connections, these can be avoided by interleaving nodes in the physical embedding. However, when higher dimensional networks are embedded in two and three dimensions, longer long wires do result. For the purpose of analyzing the effect of wire length we can determine this increase in wire length as follows (following the approach in [2]). Our analysis is based on an embedding in two dimensions, although it can be extended to three dimensions in a straightforward manner [11].

A  $k$ -ary  $n$ -cube is embedded in two dimensions by embedding  $n/2$  dimensions of the network into each of the two physical dimensions (assuming an even number of dimensions). After embedding the first two dimensions each additional dimension increases the number of nodes in the network by a factor of  $k$ . Embedding in two dimensions provides an increase in the number nodes in each physical dimension by a factor of  $\sqrt{k}$ . If we ignore wire width and wiring density effects, the length of the longest wire in each physical dimension is also increased by a factor of  $\sqrt{k}$ . Note that if we had to account for the thickness of the wires, the length of the longest wire would actually grow faster than  $\sqrt{k}$  as we increase the number of dimensions (see [2] for a good discussion). Ignoring wire width effects, the length of the longest wire in a two dimensional embedding of an  $n$  dimensional network increases by a factor of  $k^{(n-2)/2}$  over the maximum wire length of a two dimensional network. To obtain the average wire length, these distances are averaged over all dimensions and we have

$$w_{avg} = 2l_2(N^{1/2} - 1)/nk \quad (10)$$

where  $N$  is number of nodes,  $n$  is the dimension of the network, and  $l_2$  is the distance between adjacent nodes. Thus the expression is normalized to the distance between two routers in a two dimensional torus.



**Figure 5. Distribution of Power by (a) message and (b) by router**

The power dissipation in the RAU and switch is governed by the mechanism of charging and discharging the total capacitance. Therefore in our analysis of the router we assume that the power dissipated in the RAU, i.e.  $P_r$ , is related to the power dissipated in the switch as follows.

$$P_r = \frac{M_r}{M_s} P_s = K_r P_s \quad (5)$$

Substitution of (4) and (5) into (3) yields the following expression for the total power dissipated in the network:

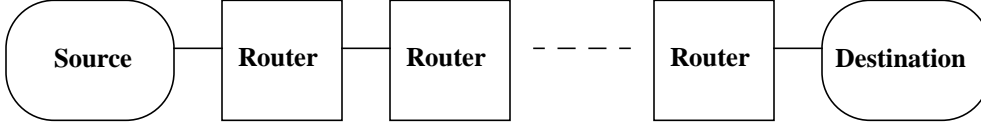
$$P_t = NG_n \left[ \left( \frac{K_r}{D} + 1 \right) P_s + P_w \right] \quad (6)$$

This equation expresses the distribution of total power between the switches and the wire in the network. The trade-offs related to power distribution between the switch and the wire is readily seen from this expression. An important question is the nature of this power distribution. What is the desired ratio? Should we build faster drivers or should we design faster switches? These are very important questions and must be addressed for future low power network designs. Generally, the design of a crossbar switch and cascaded drivers for the wire determines the values of power dissipation. However, we are interested in relationships between  $P_s$  and  $P_w$  that are determined by network topology. Therefore we leave them as free variables and study how total power dissipation is distributed between driving messages across the wires and through the switches. Now (6) can be rearranged to provide an expression for  $P_s$  :

$$P_s = \frac{D((P_t/K_n) - P_w)}{(K_r + D)} \quad (7)$$

This expression permits us to study the distribution of power between the wire and through the switches. Note that under a fixed constraint on total power, minimum wire power corresponds to maximum switch power dissipation and vice a versa.

$W$ ,  $t_r$ ,  $t_s$ , and  $t_w$ , and therefore the message latency. For a fixed number of nodes, the choice of topology also affects  $D$ ,  $W$ ,  $t_s$  and  $t_w$  (assuming  $t_r$  is fixed here). We can now establish a relationship between available power and message latency as a function of topology.



**Figure 4. Interconnected Sequence of Routers in a Network**

## 4.2 Power Dissipation Models

As described in Section 3.0, the available power can be easily obtained from the available energy used in a finite amount of time. Power dissipation in the network can be viewed in two ways: a) *message centered* view and b) *router centered* view. *Message centered* (Figure 5 a) view focuses on power dissipated in driving a message through the network. Imagine a message spreaded across  $D$  number of links (Figure 4). In a steady-state view of the network and at a particular instant in time, the power required to drive the message is equal to the power required to make a routing decision,  $P_r$ , and since the message is spreaded over  $D$  number of links and each link is concurrently switching, we have an additional term in the power required to drive the message. Expressing the above in mathematical form we have,

$$P_{msg} = P_r + D(P_s + P_w) \quad (2)$$

Alternatively, a *router centered* (Figure 5 b) view focuses on power dissipated in each router (or node) per message. The power per node,  $P_{node}$ , is simply the power it takes to process a message. This results into the following expression,

$$P_{node} = (P_r/D) + P_s + P_w \quad (3)$$

which is power dissipated in the switch,  $P_s$ , the wire,  $P_w$ , and power dissipated in the RAU,  $P_r$ , amortized over  $D$  number of links traversed by a message. These conceptually different views capture power dissipation in the network at two distinct levels. The message centered view of power dissipation is helpful if one is analyzing the network from application perspective. The router centered view is useful when the network is analyzed from technology standpoint. It is important to note that (2) and (3) are not independent equations; they are related to total available power as

$$P_t = G_n N P_{node} = G_m P_{msg} \quad (4)$$

where  $G_n$  is the number of messages handled by a processing node and  $G_m$  is the number of messages injected into the network which is an application oriented parameter subject to power constraint.

**Table 1. List of Symbols and their description**

Symbol	Description	Symbol	Description
$N$	Number of nodes	$X$	Number of inverters in cascaded driver chain
$n$	Network dimension	$\alpha$	Driver scaling factor
$P_{cap}$	Capacitive power consumption	$\beta$	Minimum size transistor transconductance
$P_{msg}$	Power dissipation per message	$\Gamma$	Total power constant coefficient
$P_{node}$	Power dissipation per node/router	$\mu$	Mobility of electrons
$P_r$	Power dissipated in the RAU	$\tau$	Rise time of the input clock pulse to the driver

buffers of a router and requests an output node. The routing and arbitration logic assigns an output buffer to the message. The message then travels through the switch, and eventually through the wire to the adjacent router. Therefore in our model, the critical path for a message involves the routing & arbitration unit, switching logic and inter-router wire delays. We will examine the power and time delays associated with each component. The assumptions made in the derivation of the model are listed below:

1. Wormhole switching for message transmission.
2. No load latency and contention free network. A model of contention is necessary to accurately reflect power consumption in active networks. However, as a first step we study the basic relationships between topology, power and physical constraints. We will then seek to extend this model to incorporate the effects of contention within the network.
3. Power dissipated in the RAU is related to the power dissipated in the switch. We use a simple model where this relationship is given by the ratio of number of transistors in the RAU to that in the switch.
4. The RAU delay is equal to the switching delay. This is generally the case in most modern high speed routers.

## 4.1 Model Development

Table 1 provides a summary of all of the parameters used in the model development. The timing delay through the switch is referred to as the switching delay and is characterized by  $t_s$ . The delay through the RAU is given by  $t_r$ , and the wiring delay is given by  $t_w$ . Using a wormhole switching implementation, the no load latency in a contention free network can be expressed as the sum of distance and message length components,

$$t_{wormhole} = [D(t_r + t_s + t_w)] + \left[ \max(t_s, t_w) \left[ \frac{L}{W} \right] \right] \quad (1)$$

where  $D$  is the average number of links or distance traversed by a message through the network,  $L$  is length of the message in bits and  $W$  is the width of the physical channel [8]. The distance component is the time it takes for the header flit to reach from the source to destination and the message length component is the time it takes to transfer the rest of the message. We first determine expressions for  $t_s$  and  $t_w$  constrained by the available power. By imposing this constraint, we are implicitly constraining the values of

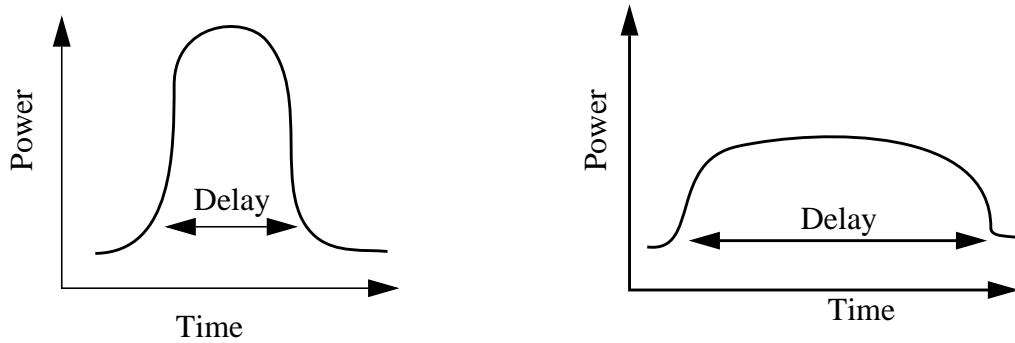
satellite comes to an area of interest and images are taken for processing and transmission back to Earth. There is a choice of accommodating this workload immediately with low latency at the price of high power dissipation, or a higher latency with lower power dissipation..

## 4.0 Power and Latency Models

From a system performance point of view we have three distinct delays associated with driving the message through the network: switching delay, wire delay, and routing delay. A message enters the input

**Table 1. List of Symbols and their description**

Symbol	Description	Symbol	Description
$C_{in}$	Input gate capacitance	$P_s$	Power dissipated in the switch
$C_o$	Output capacitance of minimum transistor in the cascaded drivers	$P_{sc}$	Short circuit power consumption
$C_{ox}$	Gate oxide capacitance	$P_s$	Power dissipated in the wire
$C_p$	Next stage capacitance in the driver chain	$P_t$	Total power dissipated in the system
$C_{pad}$	Output pad capacitance	$R_o$	Output resistance of minimum transistor in cascaded drivers
$C_t$	Total capacitance of the node	$R_w$	Wiring resistance
$C_w$	Total wiring capacitance	$S_n$	Delay constant of nFET transistor
$D$	Average distance or number of hops	$T_{clk}$	Period of the clock in the driver chain
$f_i$	Frequency of the driver	$T_{chain}$	Time delay in the driver chain
$f_s$	Frequency of the switch	$T_{line}$	Time delay in the wire or the line
$f_w$	Frequency of the wire	$T_{tof}$	Time of flight
$G_m$	Messages injected into the network	$t_r$	RAU delay
$G_n$	Messages handled by the node	$t_s$	Switch delay
$K_1$	Duty cycle	$t_w$	Wire delay
$K_r$	Ratio of number of transistors in the RAU to switch	$V_{dd}$	Supply voltage
$K_s$	Fraction of active gates during switching	$V_t$	Threshold voltage
$L$	Message bit length	$v_p$	Propagation velocity of signal
$l_2$	Base latency	$W$	Message channel width
$M_r$	Number of transistors in the RAU	$W_1/L_1$	The size of the minimum transistor in the driver chain
$M_s$	Number of transistors in the switch	$w_{avg}$	Average wire length between processors



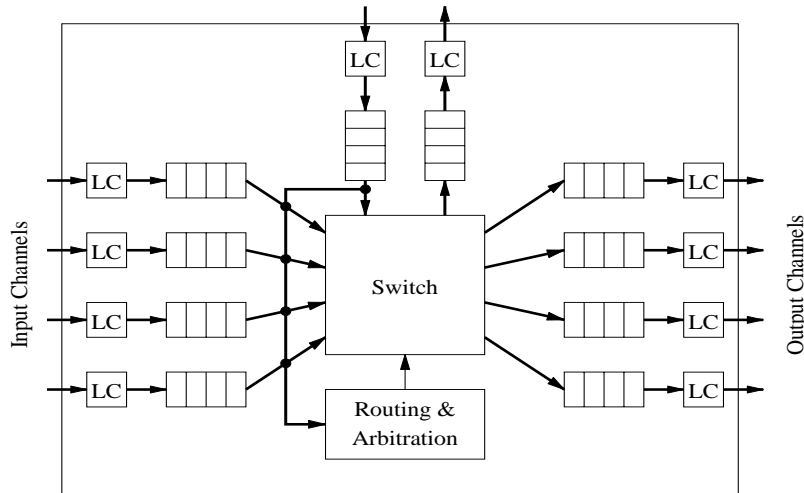
**Figure 3. Energy Profiles: (a) high rate, low latency dissipation and (b) low rate, high latency dissipation**

### 3.0 Power Dissipation in Pipelined Networks

Power is the rate of energy dissipation in a system. In a typical VLSI CMOS circuit, this energy is consumed in switching transistors, in leakage due to circuit device structures, and in holding the logic state of a circuit. The rate at which this energy is used by a system determines its power dissipation. Power is measured in watts [joules/sec] and is the dissipation rate of energy which is measured in joules. A network consumes approximately the same quantity of energy to transport a message to its destination independent of the switching technique used (i.e. either packet switching or virtual cut-through switching may be employed). The latency of message delivery and the dissipated power are related by the architecture of the network. For example, a network architecture with wide channel widths can transport data packets in less time compared to a network architecture with smaller channel widths. However, the wider channel network will consume energy at a higher rate because of the larger data buffers and wider busses. Figure 3(a) illustrates the power utilization profile of a network with a sharp peak, indicating high power dissipation, and Figure 3(b) illustrates the power dissipation of a system with lower plateau to indicate low power dissipation. The total energy consumed, indicated by area under the curve may be the same. However, average message latency is lower in the former case. The trade-off for the higher power consumption is the lower latency of the system.

In addition to understanding the relationship between power dissipation and latency, we must also consider some related constraints. In spaceborne and battery powered portable systems a limited amount power can be delivered instantaneously. In other words, the power supply is current limited. This power constraint may be derived from power budgets where the current supply has been limited so that other components may utilize the available power for mission critical computations. Other issues such as cooling and reliability may also limit the system power consumption. Thus while available energy may not be constrained, the rate at which it can be used may be constrained.

It is also important to consider the demands of a bursty workload on the network, especially in a power constrained environment. This bursty workload often rises sharply at a short period of time, e.g. when a



**Figure 2. Generic Router Architecture**

- **Buffers:** These are first-in-first-out (FIFO) buffers for storing messages in transit. In the above model, a buffer is associated with each input physical channel and each output physical channel. The buffer size is an integral number of flow control units.
- **Switch:** This component is responsible for connecting router input buffers to router output buffers. High speed routers will utilize crossbar networks with full connectivity, while lower speed implementations may utilize networks that do not provide full connectivity between input buffers and output buffers.
- **Routing and Arbitration Unit (RAU):** This component implements the routing algorithms, selects the output link for an incoming message, and accordingly sets the switch. If multiple messages simultaneously request the same output link, this component must provide for arbitration between them. If the requested link is busy, the incoming message remains in the input buffer. It will be routed again after the link is freed and if it successfully arbitrates for the link.
- **Link Controllers (LC):** The flow of messages across the physical channel between adjacent routers is implemented by the link controller. The link controllers on either side of a channel coordinate to transfer units of flow control.
- **Processor Interface:** This component simply implements a physical channel interface to the processor rather than to an adjacent router.

Given the architectural abstractions we have defined above, our model of power dissipation is comprised of three components:

- *switching power dissipation:* This component models the power dissipated when a flit is driven through the router switch from input buffers to output buffers.
- *wire power dissipation:* Power is dissipated in driving messages through on-chip wires as well as inter-router wires. i.e., the physical channel.
- *routing power dissipation:* This component is the amount of power required to make a routing decision when examining a header flit.

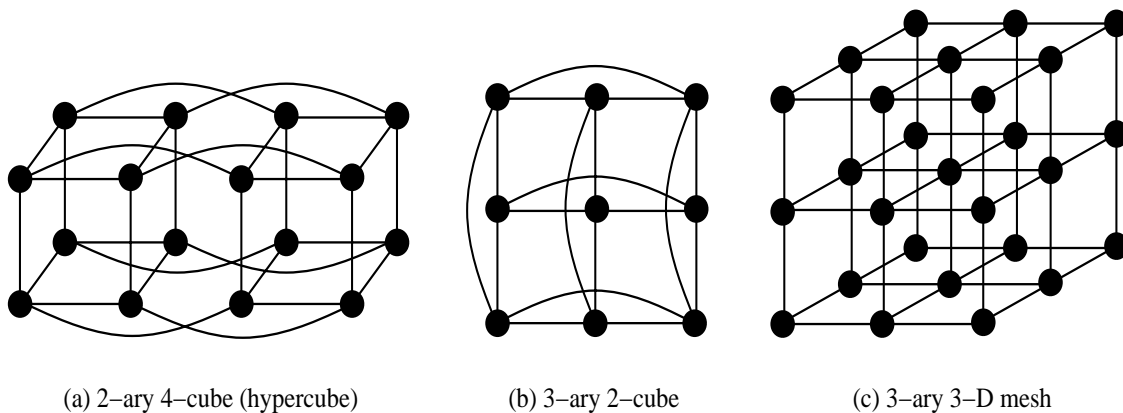
tions for the construction of low power networks are discussed. The paper concludes with a summary of our future research directions.

## 2.0 Network and Router Model

The class of networks considered in this paper are the torus connected, bidirectional,  $k$ -ary  $n$ -cubes. Parallel computers employing this type of network include, for example, the CRAY T3/E [10]. A  $k$ -ary  $n$ -cube is a network with  $n$  dimensions and  $k$  processors in each dimension. In torus connected  $k$ -ary  $n$ -cubes, each processor is connected to its immediate neighbors modulo  $k$  in every dimension. Routing within each dimension is orthogonal to routing in other dimensions. The mesh networks differ from tori in that there are no wrap around links in each dimension. Some common topologies are illustrated in Figure 1.

A message is broken up into small units referred to as *flow control digits* or *flits*. A flit is the smallest unit on which flow control is performed, and represents the smallest unit of synchronized communication between adjacent routers. Messages are pipelined through the network at the flit level. Each network node is comprised of a router (see Figure 2) and associated processor. This paper focuses on the network routers and interconnecting links. The network communication links are full-duplex links, and the physical channel width and flit size are assumed to be equivalent. Flits are moved from input channel buffers to output channel buffers within a node by an internal crossbar switch. Buffers are not large enough to store complete message packets. Therefore, if an output channel at a router is busy, the message blocks in place occupying buffers over several routers. This is referred to as wormhole switching [7]. The generic router architecture utilized in this study is illustrated in Figure 2.

The basic components include:



**Figure 1. Examples of Orthogonal, Direct Network Topologies**

capable of processing real time video and audio streams, concurrently handle communications, displays and computations, and do so in portable, mobile environments. Another important area for low power designs is the mission critical systems, such as a space-borne system. Deep space missions require substantial increases in onboard computations to make efficient use of the low bandwidth downlinks, but must do so in a severely power constrained environment. Multiprocessor architectures will become the essential building blocks to meet the computational demands of increased autonomy and functionality of these applications. The interconnection network is a major consumer of power in these systems. While the performance of various interconnection topologies has been studied in depth, power dissipation characteristics have not been reported in the literature. In such systems, the power consumed by the interconnection network must come under the same scrutiny as the rest of the system if we are to be able to design networks that can be efficient consumers of power.

This paper analyzes the performance of direct interconnection network topologies under a fixed power constraint. The research is focused on architectures employing hardware support for low latency, fine grained communication. In such networks, the raw hardware latency is a significant component of the overall message latency. Our goal is to understand how system design decisions such as choice of network topology, switching techniques and power distribution are influenced by constraints on the available power. We develop an analytic model of the average message latency as a function of several architectural parameters such as network dimensions, channel widths, bisection bandwidth; technological parameters such as switching delays, wire delays, design of switch, wire and routing arbitration unit (RAU). These parameters are placed under a fixed power constraint to study their effect on the overall design of the network. Thus, the model captures relationships between message latency, total available power, and various architectural parameters. The model provides useful insights into the design of network topologies for low power applications. For example, we have found that the distribution of power between driving inter-router channels and switching through the routers has a strong influence on the network topology, particularly for larger systems. Important information on the total power distribution between the switch and the wire is obtained from the analysis to set a guideline for the hardware design. The model supports the incorporation of application parameters such as message injection rate [13], contention, and pipelining of messages in the network. It also supports the incorporation of the effect of packaging technologies. e.g., PCB vs. multi-chip modules. Different packages possess distinct physical characteristics which in turn affects the distribution of power among network components.

The following section introduces the class of networks that are addressed in our analysis; Section 3.0 describes power dissipation concepts in these networks. Section 4.0 develops a detailed model of the average message latency under a fixed power constraint. The model is analyzed in Section 5.0 and the implica-

# Power Constrained Design of Multiprocessor Interconnection Networks<sup>†</sup>

Chirag S. Patel, Sek M. Chai, Sudhakar Yalamanchili, David E. Schimmel

School of Electrical and Computer Engineering  
Georgia Institute of Technology  
Atlanta, Georgia 30332-0250  
e-mail: sudha.yalamanchili@ece.gatech.edu  
Fax: (404) 894-9959

## Abstract

*This paper considers the power constrained design of orthogonal multiprocessor interconnection networks. We present a detailed model of message latency as a function of topology, technology, architecture, and power. This model is then used to analyze a number of interesting scenarios, providing a sound engineering basis for interconnection network design in these cases. For example, we have observed that under a fixed power constraint, the network dimension which achieves minimal latency is a slowly growing function of system size. In addition, as we increase the available power per node for a fixed system size, the dimension at which message latency is minimized shifts towards higher dimensional networks.*

## 1.0 Introduction

The design of modern high performance multiprocessor interconnection networks is a process of optimizing the performance of the network in the presence of implementation constraints. Common constraints that have formed the basis of such an analysis in the past have included wiring constraints [6], pin out constraints[1], wire delays [11], and switching speeds [2]. This paper focuses on a similar analysis of network performance under a constraint that has been ignored to date: a constraint on the available power.

The explosive growth of telecommunications applications and portable computing has produced a new and often overriding system design constraint: limited available power. Embedded computers, laptops, palmtops, pagers, personal digital assistants, etc. have been the main focus of the application for low power technologies. The majority of low power research is focused on device technology and circuit design, and there has been by comparison, not as much attention devoted to architectural techniques for managing power more effectively. Efforts that have developed architectural and design techniques have been quite successful in uniprocessor architectures, and embedded digital logic systems (c.f. [4,9]). However as we look to future applications, we can see enormous growth in arenas where small, compact systems must be

---

<sup>†</sup>. This research was partially supported by NASA Jet Propulsion Laboratory under contract 960586.