Proc. Int'l Conf. Communications in Computing (CIC-2000)
Las Vegas, Nevada, USA, 2000 June 26-29

Why Network Diameter is Still Important

Behrooz Parhami
Department of Electrical and Computer Engineering
University of California
Santa Barbara, California 93106-9560, USA
E-mail: parhami@ece.ucsb.edu

Chi-Hsiang Yeh
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario K7L 3N6, Canada
E-mail: chi-hsiang.yeh@ece.queensu.ca

Abstract

It has become fashionable to argue against research on new interconnection networks by claiming that with wormhole switching, the dominant communication method in modern parallel computers, routing latency is insensitive to the diameter and other topological parameters of the network; hence, the inference that network topology is unimportant and the edict that we should adhere to 2D and 3D mesh networks that offer low-cost implementations, simple algorithms, and ease of expansion. This view, if left unchallenged, could be harmful to further developments in the field of parallel computer architecture. In this paper, we present various types of evidence that show diameter to be quite important when networks are compared under realistic and fair conditions that include some form of aggregate bandwidth and/or cost normalization.

Keywords: adaptive routing, congestion, deadlock, diameter, interconnection network, network topology, oblivious routing, packet or wormhole switching.

1. Introduction

This paper is motivated by the following common observation, offered by some researchers in parallel processing, to decry proposals for new interconnection networks that purport to offer certain topological advantages over existing architectures:

> With wormhole switching, which is dominant in modern parallel computers, message delay is almost independent of the routing path length; this makes the diameter, or the whole network topology for that matter, completely irrelevant.

The foregoing statement, and milder variants thereof, have been repeated so many times that it is now taken for granted by some that network diameter and other topological parameters are truly unimportant.

We strive to dispel this misconception by showing that topological parameters in general, and network diameter in particular, are still quite important. Toward this goal, we offer intuitive arguments and rough analyses based on a number of architecture-independent parameters. While architecture can be taken into consideration to make the analyses more detailed and the estimates more accurate, we believe that this rough treatment is adequate to support our thesis.

We demonstrate, in the various sections of the paper, that network diameter does not only influence routing distances but is also intimately related to unroutability due to conflicts, congestion, deadlock probability, and realizability in a pin-limited context. Due to space limitation, formal proofs have been left out.

2. Topological Parameters

Topological parameters of a network that potentially affect routing performance include:

- N Number of nodes in the network
- D Diameter (maximum among the internode shortest distances)
- Δ Average internode distance
- d Node degree (indicator of node bandwidth with all-port communication)
- C Total number of channels or links (equal to $Nd/2$ for a regular network)
- B Bisection bandwidth (measure of aggregate bandwidth with random messages)
- L Maximum wire length or wire latency (favors local connectivity, as in mesh)
- Λ Average wire length or wire latency (related to L as Δ is related to D)

Let's agree to always compare networks with the same number N of nodes. Any global cost-performance optimization for a given application must of course include architectures with fewer and more nodes, but this is beyond our scope here. The following theorem helps us establish that diameter and average internode distance are practically equivalent for approximate performance comparison of competing topologies.

Theorem 1: For any node-symmetric network, the average internode distance Δ and the diameter D are related by the inequalities $D/2 \leq \Delta \leq D$.

Note that the bounds given in Theorem 1 are tight: binary nD hypercube exemplifies the case of $\Delta \approx D/2$; nD radix-r generalized hypercubes [Bhu84], with nonconstant n and r, and macro-star networks [Yeh98] are examples where $\Delta = D - o(D)$.

Most networks of theoretical and practical interest are node-symmetric: e.g. torus, hypercube, cube-connected cycles, and star. In fact, the inequality of Theorem 1, or a slightly laxer form of it, holds for many popular node-asymmetric networks such as meshes, balanced binary trees, pyramids, and meshes of trees. It follows that demonstrating the importance of Δ should lead to the conclusion that the diameter D is important also.

The total number C of channels or links has been used as a rough indicator of network cost in the absence of more detailed information about the architecture. Given a fixed per-channel capacity, the parameter C is also proportional to the aggregate bandwidth of the network. A large C translates to good communication performance only if the routing algorithms can utilize all the links with spatial and temporal uniformity.

The bisection bandwidth B is a measure of the available network resources for communication-intensive applications with random data exchange patterns. Generally speaking, a large bisection implies dense connectivity and higher performance for communication-intensive problems, albeit at the cost of more difficult wiring (larger layout area in the chip/board realm). Unlike network diameter D, which is related to N and d through Moore bounds [Parh99], the numbers of nodes and links in a network have no relation to its bisection bandwidth; it is fairly easy to construct large, dense, node-symmetric networks that exhibit relatively small bisections.

Dimensionizable node- and edge-symmetric networks are an important subclass of node- and edge-symmetric networks. We can map a node of a dimensionizable node- and edge-symmetric network to any node of the same network so that every dimension-i link, where i is a unique label in $[1, d]$ given to each edge incident to a node, is mapped to another dimension-i link for all i. Hypercubes, k-ary n-cubes, radix-r generalized hypercubes, and star graphs all belong to the class of dimensionizable node- and edge-symmetric networks. The following result, provable by combining results from [Yeh88b] with Theorem 1, shows that in dimensionizable node- and edge-symmetric networks, small diameter guarantees large bisection.

Theorem 2: Bisection width of a dimensionizable node- and edge-symmetric interconnection network is at least $(d/D)\lfloor N/2 \rfloor \lceil N/2 \rceil / (N-1) \approx dN/(4D)$.

As in the case of Theorem 1, the bound of Theorem 2 is tight: nD radix-r generalized hypercubes [Bhu84], with nonconstant n and r, have bisection widths that are within a factor of $1 + o(1)$ from this lower bound.

The maximum wire length L captures interconnection locality. Mesh and torus networks can be laid out (or wired) with short links, whereas the 2D VLSI layout of an N-node hypercube, e.g., leads to wires of length $\Omega((N/\log N)^{1/2})$. Compared to L, the average wire length Λ is a better indicator of latency due to signal propagation, provided that communication is at least in part asynchronous and can take advantage of the faster propagation times when encountered. With a suitably designed routing algorithm that utilizes the available channels efficiently and fairly, network performance increases with C and decreases with L (or Λ) when links are not pipelined.

There are other aspects of network topology, not quantifiable by simple numerical measures as the ones discussed above, that have significant effects on routing performance and implementation cost. The most important of these concerns modularity or packageability. Current VLSI implementation technology imposes severe I/O or pin limitations at various levels of the packaging hierarchy (chip, board, chassis, cabinet, etc.). Networks that can be efficiently packaged within these constraints due to their hierarchical or recursive structures invariably lead to smaller L, smaller Λ, reduced chip area for drive electronics, and lower per-hop power dissipation, by reducing intermodule communications at the higher packaging levels (see Section 7).

3. Hidden Assumptions

It is by no means certain that all future parallel computer systems will use wormhole switching exclusively [Magg96]. In fact, packet switching, whose latency is directly proportional to the hop distance from source to destination, is still quite competitive under some conditions, and will remain so, in one form or another, for the foreseeable future. However, let us assume the use of wormhole switching and focus on the remaining aspects of the misconception.

Note that routing latency with wormhole switching is insensitive to hop distance or network diameter only if the network is very lightly loaded AND messages are rather long. Unless applications of interest are not communication-intensive, light message loading can be ensured only by providing an aggregate network bandwidth that is significantly higher than required; this implies a much greater cost that may render the resulting network solution cost-ineffective.

As for message length, clearly, the latency of a message consisting of a few flits increases sharply with any increase in the hop distance. Less obvious is the fact that for the very same long messages where distance assumes a secondary role, congestion or conflicts become quite important, since such wormhole messages occupy the links for extended periods. Of course with more conflicts comes not only increased delay due to waiting, but the heightened possibility of deadlock. When deadlock is a serious threat, we are forced to sacrifice system resources to detect or recover from deadlock, or else restrict ourselves to less flexible, and often lower-performing, deadlock-free routing algorithms. It is well known that when deadlock is extremely unlikely, an aggressive routing strategy coupled with deadlock detection and recovery can perform much better than a conservative strategy based on strict deadlock avoidance [Pink97].

A common assumption in analyzing networks with regard to their communication capacity is that of random message traffic. Specific communication patterns with various regularities allow some networks to perform much better than in the case of random traffic. The flip side of this coin is the fact that one can often construct worst-case communication patterns that strain any given network's ability to efficiently route the messages. This is particularly true with wormhole switching; and it tends to be easier to do for networks that have large diameters. For these reasons, we stick with the assumption of random traffic.

Let us continue our discussion in the case of long messages. In a network with a fixed total number C of links, the average internode distance Δ has a direct effect on congestion. For example, with $C = 1000$ and $\Delta = 10$, on the average no more than 100 long wormhole messages can be in transit at any given time (probably far less, due to conflicts and uneven distribution of traffic). If Δ drops to 8, say, the number of messages that can be handled increases, leading to enhanced network bandwidth. In the following sections, we proceed to quantify certain aspects of the foregoing observation.

4. Oblivious Routing

We would like to compare two networks that have the same number C of links but different average internode distances Δ and Δ'. The number C of links is a very rough measure of network cost and it makes sense to compare equal-cost networks. As stated earlier, other aspects of a network, such as VLSI layout area and the length of the longest wire affect its cost and per-hop latency, these cannot be taken into account in an architecture-independent discussion and we proceed by equating the network cost with C; let us say that we strive to compare only network architectures with comparable layout or packaging complexity.

In the following analysis, we assume that oblivious routing is used; i.e., the same unique routing path $P_{i,j}$ is always chosen when sending a message from node i to node j. In our rough analysis, we assume all routing paths to have the common length Δ.

Consider the probability p_i of being able to establish an ith routing path, given that $i - 1$ paths are already in place. This requires that all the required Δ links in the new path be available. Hence:

$$p_i = \binom{C-(i-1)\Delta}{\Delta} \Big/ \binom{C}{\Delta}$$

For most values of Δ, the probability p_i is a sharply decreasing function of i. To get a feel for the numbers involved, let's take $C = 100$ and compute p_i for different values of i and Δ (Fig. 1).

The data used in constructing Fig. 1 also suggests that the expected number of routing paths that can be established before conflicts make additional paths impossible is rather small. The foregoing observations are basically the routing counterpart to the "birthdays" problem, often discussed in elementary probability courses; here, as in the birthdays problem, the probability of an event's occurrence, namely that of multiple paths requiring the use of the same link (two or more people sharing the same birthday), is much larger than intuition would lead us to believe.

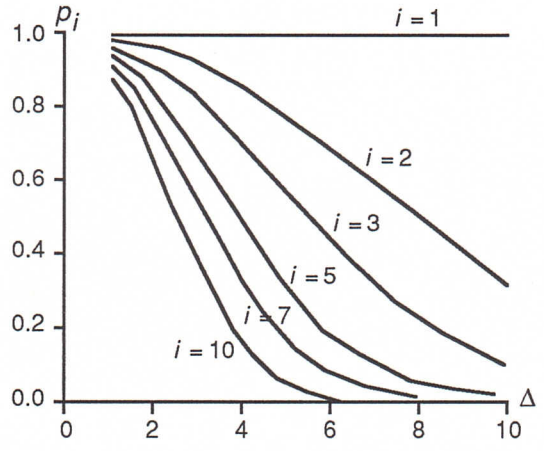

Fig. 1. Probability p_i of being able to establish an ith routing path, given that $i - 1$ paths are already in place, in a 100-link network with average internode distance Δ.

5. Adaptive Routing

One may think that the pessimistic conclusions of Section 4 stem from the rigidity of oblivious routing and that the flexibility of adaptive routing may significantly improve the situation.

First, we note that adaptive routing comes with nontrivial overhead in computing and establishing the routing paths, both due to the more complex routing decisions and due to the greater difficulty of deadlock avoidance. This is why, in practice, adaptive routing is not as popular as one might think [Duat94].

Arguing for or against the benefits of adaptive routing is outside the scope of this paper. However, we note that any benefits offered by adaptive routing are likely to be more pronounced for a network with smaller Δ. Discussing the effects of adaptive routing is difficult, if not impossible, without reference to a specific architecture. However, various intuitive arguments can be used to justify the preceding claim.

The first argument goes like this. Suppose we are trying to establish an ith routing path after $i - 1$ paths are already in place. The larger the average internode distance Δ, the longer the paths that are already in place, and the harder it is to circumvent them to establish the new path. This is particularly true for the majority of implemented or proposed adaptive routing algorithms that are restricted in their routing decisions; e.g., must select a shortest path. Unrestricted adaptive algorithms present a different type of problem. They may be able to proceed in a situation where an oblivious or more restricted adaptive routing algorithm fails. However, they do so by establishing increasingly longer routing paths to get around already occupied links. This, however, is self-defeating in that greater network bandwidth tends to be tied down for each additional message, particularly if messages are very long. Of course, the preceding shortcoming is in addition to greater overhead (in routing decisions and deadlock avoidance/recovery techniques) implied by unrestricted adaptive routing.

Here is a second intuitive argument: Any routing path of length l or less between the nodes i and j must visit exclusively nodes that are $l/2$ or fewer hops away from one of the end points i and j. The smaller the network diameter, the larger the number of nodes satisfying the condition above, and thus the greater the potential benefits of adaptive routing. Note that because the benefit accrues for each l, the preceding observation is valid for both unrestricted adaptive routing and for adaptive routing algorithms that place an absolute or shortest-distance-related upper bound on the length of the routing path used.

A third argument can be added for a special class of adaptive routing algorithms known as deflection or hot-potato routing. Intuitively, any decrease in Δ tends to decrease the probability of deflection and the additional travel time of deflected messages. Both of these tend to improve performance in a nonlinear fashion; i.e., a 5% decrease in deflections is likely to have a significantly higher than 5% impact on the average latency and network throughput.

6. General Routing

There are also bounds on routing performance that hold regardless of the routing method used (oblivious or adaptive). Here, we provide two such results that help establish the importance of network diameter.

Random routing, where packets are generated randomly at network nodes and packet destinations are uniformly distributed, is a fundamental communication problem. Total exchange (all-to-all personalized communication) is also deemed quite important. The following results relate the performance of a network during random routing and total exchange to network diameter.

Theorem 3: The maximum achievable throughput for random routing in a dimensionizable node- and edge-symmetric network is at lease d/D.

Theorem 4: In a dimensionizable node- and edge-symmetric network, a set of d total exchange tasks can be executed in $(N - 1)D$ or fewer communication steps under the all-port communication model.

7. Packaging Considerations

A common problem in practical systems involving multiple packaging levels is I/O pin limitations that tend to favor architectures with smaller diameters. Therefore, whereas low-dimensional k-ary n-cubes outperform higher dimensional ones when bisection width is kept constant [Dall90], the situation reverses under pin-out constraints [Abra91].

We have previously shown [Yeh98] that certain small-diameter architectures can offer high performance when packaging constraints are taken into account. Such architectures possess hierarchical constructions and can be made to satisfy pin-out constraints through judicious choice of their structural parameters. Suitable choices, in effect, trade off intermodule connectivity for intramodule links. This allows us not only to meet any hard constraints on intermodule connectivity but also to use the aforementioned tradeoff to reduce system cost and/or power dissipation via fewer pin-outs and associated buffering and drive circuitry.

Recent studies of pruned networks also offer evidence for the suitability of low-diameter architectures in pin-limited implementation contexts. For example, a 3D torus can be pruned to use degree-4 nodes without a notable increase in its D or Δ parameter. The resulting network is close to 2D torus in layout complexity but offers performance comparable to 3D torus when cost is normalized [Kwai99]. Similar benefits accrue from pruning strategies applied to other densely connected low-diameter networks [Parh99a].

Imagine a million-processor system implemented with 1K or so nodes per printed-circuit board (requires the use of multiple-processor chips, as in the recently announced IBM Blue Gene project [Clar00]). If boards were put in eight cabinets, each holding one-eighth of a 3D torus, say, we would need 15,000 channels per cabinet. In such a scenario, one cannot dismiss more pin-efficient, low-diameter architectures as one does for today's top-of-the-line parallel computers that do not go beyond several thousands of processors.

8. Conclusion

Network diameter and other topological properties are not as unimportant as some would lead us to believe. For one thing, with implementation cost normalized, even a small reduction in network diameter should be welcome rather than dismissed as insignificant. For another, the space of possibilities for network architectures is vast; the choice is not limited to low- versus high-dimensional k-ary n-cubes. Furthermore, it is quite dangerous to generalize from a small number of high-level studies. It is even more dangerous to base the evaluation of research papers and proposals on practices that may have been derived from nontechnical considerations. If a similar mentality prevailed in operating systems, for example, only research on Microsoft Windows would be deemed appropriate.

References

[Abra91] Abraham, S. and K. Padmanabhan, "Performance of Multicomputer Networks under Pin-out Constraints", *J. Parallel and Distributed Computing*, Vol. 12, pp. 237-248, July 1991.

[Basa96] Basak, D. and D.K. Panda, "Designing Clustered Multiprocessor Systems under Packaging and Technological Advancements", *IEEE Trans. Parallel and Distributed Systems*, Vol. 7, pp. 962-978, Sep. 1996.

[Bhuy84] Bhuyan, L.N. and D.P. Agrawal, "Generalized Hypercube and Hyperbus Structures for a Computer Network", *IEEE Trans. Computers*, Vol. 33, pp. 323-333, Apr. 1984.

[Clar00] Clark, D., "Blue Gene and the Race Toward Petaflops Capacity", *IEEE Concurrency*, Vol. 8, pp. 5-9, Jan.-Mar. 2000.

[Dall90] Dally, W.J., "Performance Analysis of k-ary n-Cube Interconnection Networks", *IEEE Trans. Computers*, Vol. 39, pp. 775-785, June 1990.

[Duat94] Duato, J., "Why Commercial Multicomputers Do Not Use Adaptive Routing", *IEEE Computer Arch. Tech. Committee Newsletter*, pp. 20-22, Summer/Fall 1994.

[Kwai99] Kwai, D.M. and B. Parhami, "Comparing Torus, Pruned Torus, and Manhattan Street Networks as Interconnection Architectures for Highly Parallel Computers", *Proc. Int'l Conf. Parallel and Distributed Computing and Systems*, Nov. 1999, pp. 19-22.

[Magg96] Maggs, B.M., "A Critical Look at Three of Parallel Computing's Maxims", *Proc. Int'l Symp. Parallel Arch's, Algor's and Networks*, 1996, pp. 1-7.

[Parh99] Parhami, B., *Introduction to Parallel Processing: Algorithms and Architectures*, Plenum Press, 1999.

[Parh99a] Parhami, B. and D.-M. Kwai, "Periodically Regular Chordal Rings", *IEEE Trans. Parallel and Distributed Systems*, Vol. 10, pp. 658-672, June 1999.

[Pink97] Pinkston, T.M. and S. Warnakulasuriya, "On Deadlocks in Interconnection Networks", *Proc. Int'l Symp. Computer Architecture*, 1997, pp. 38-49.

[Yeh98] Yeh, C.-H. and B. Parhami, "Synthesizing High-Performance Parallel Architectures under Inter-Module Bandwidth Constraints", *Proc. Int'l Conf. Parallel and Distributed Computing & Systems*, 1998, pp. 414-416.

[Yeh98a] Yeh, C.-H. and E.A. Varvarigos, "Macro-Star Networks: Efficient Low-Degree Alternatives to Star Graphs", *IEEE Trans. Parallel and Distributed Systems*, Vol. 9, pp. 987-1003, Oct. 1998.

[Yeh98b] Yeh, C.-H., "Efficient Low-Degree Interconnection Networks for Parallel Processing: Topologies, Algorithms, VLSI Layouts, and Fault Tolerance", PhD Dissertation, ECE Dept., UC Santa Barbara, Mar. 1998.

[Yeh00] Yeh, C.-H. and B. Parhami, "A Unified Model for Hierarchical Networks Based on an Extension of Cayley Graphs", *IEEE Trans. Parallel and Distributed Systems*, to appear.