Some Properties of Swapped Interconnection Networks

Behrooz Parhami

Dept. of Electrical & Computer Engineering
University of California
Santa Barbara, CA 93106-9560, USA

parhami@ece.ucsb.edu

Abstract

Interconnection architectures range from complete networks, that have a diameter of $D = 1$ but are impractical except when the number n of nodes is small, to low-cost, minimally connected ring or loop networks whose diameter $D = \lfloor n/2 \rfloor$ is unacceptable for large n. In this paper, our focus is on swapped interconnection networks that allow systematic construction of large, scalable, and highly modular parallel architectures, while maintaining desirable properties of underlying basis networks. We show how key parameters of a swapped interconnection network are related to the corresponding attributes of its basis network and demonstrate applications of these results to synthesizing large networks with desirable modularity, packageability, performance, and fault tolerance attributes.

Keywords – Average distance, Bisection width, Complete graph, Fault diameter, Fault tolerance, Interconnection network, Modularity, OTIS Network, Packageability, Routing, Survivability.

1. Introduction

Low latency, high bandwidth, modularity, energy efficiency, and robustness are some of the properties that are sought in networks for parallel and distributed computing. Given that network performance parameters depend not only on the network architecture but also on a number of factors relating to applications and data exchange characteristics, the challenge in interconnection network design is finding the right match between communication needs on one side and capabilities and limitations inherent in each architecture on the other. This, in turn, explains the proliferation of implemented and proposed connectivity schemes among multiple processors, sometimes characterized as the sea of interconnection networks [Parh99], [Parh05].

Ideally, each network node is directly connected to every other node, thus allowing one-hop communication between any pair of nodes. This is modeled by the n-node *complete graph* K_n. Physically, however, complete-graph connectivity is difficult to provide for large systems that are of practical interest. At the other extreme from K_n, the simplest possible physical connectivity pattern is that of n-node *ring* R_n. Here, each node has only two communication channels. Data exchange is direct only between each node and one or two neighbor(s) and indirect in all other cases. Intermediate architectures between K_n and R_n can be obtained in a variety of ways, providing tradeoffs in cost and performance. Network cost is affected, among other factors, by the (maximum) *node degree* d, while indicators of network performance include *diameter* D and *bisection width* B. The degree-diameter product dD is sometimes used as a composite measure of cost-effectiveness in network comparisons.

Many of these intermediate architectures can be viewed as *chordal rings* [Arde81], rings to which *bypass links* or *chords* have been added to reduce the network diameter, or richly connected graphs from which certain links are systematically removed via *pruning* [Kwai98], [Parh01], [Parh04] so as to reduce the node degree, wiring density, and network cost. Other mechanisms for deriving new interconnection networks from other (basis) networks include cross-product composition, recursive substitution, and hierarchical composition. These combining strategies lead to families of networks that are all based on the same component networks and thus share a number of common topological, performance, and robustness attributes.

Our focus of discussion in this paper is swapped interconnection networks that allow systematic construction of large, scalable, and modular parallel architectures, while maintaining desirable properties of an underlying nucleus or basis network. We show how key parameters of a swapped network are related to the corresponding attributes of its basis network and demonstrate applications of these results to synthesizing large interconnection networks with desirable packageability, performance, and fault tolerance attributes.

2. Swapped Networks

Symmetric interconnection networks, in which node degree is uniformly equal to d and the network "looks the same from every node," are of particular interest due to their algorithmic simplicity and greater resilience resulting from the absence of weak spots. A symmetric network is characterized by its number n of nodes, along with one or more other parameters or rules that define its connectivity pattern. Focusing on the number n of nodes for the moment (e.g., by fixing other independent variables at default values), topological parameters can be expressed as conventional or asymptotic functions of n. This is shown in the case of network diameter in Fig. 1.

Practical interconnection networks used in parallel computers fall on the right side of Fig. 1, where networks tend to be scalable, readily packageable, and low-cost. Interconnection network research over the past two decades, on the other hand, has focused on lower-diameter networks constituting the left half of Fig. 1. Swapped networks offer some of the advantages of each group in that they can have sublogarithmic diameters while remaining both packageable and relatively scalable.

Definition 1: The *swapped network* $Sw(G)$, derived from the n-node *nucleus* or *basis* graph G, is a graph with n copies of G numbered 0 to $n - 1$, so that v_{ij}, node j in copy i, is connected to v_{ji}, node i of copy j, for all $i \neq j$ and $0 \leq i, j \leq n - 1$. For the sake of uniformity, we can assume that node v_{ii} possesses an external I/O link. ■

Algorithm R^{sw}: Routing in a swapped network with basis network G from v_{ij} to v_{kl} using at most one intercluster hop – If $i = k$, then use the routing algorithm for G to route in $\delta_G(j, l)$ hops. Otherwise, route from v_{ij} to v_{ik} within the source cluster i in $\delta_G(j, k)$ hops, then from v_{ik} to v_{ki} in a single intercluster hop, and finally from v_{ki} to v_{kl} within the destination cluster k in $\delta_G(i, l)$ hops. ■

Note that R^{sw} is not a shortest-path routing algorithm. The path $v_{ij} \rightarrow v_{ik} \rightarrow^{sw} v_{ki} \rightarrow v_{kl}$, which is of length $\delta_G(j, k) + \delta_G(i, l) + 1$, may be longer than $v_{ij} \rightarrow v_{im} \rightarrow^{sw} v_{mi} \rightarrow v_{mk} \rightarrow^{sw} v_{km} \rightarrow v_{kl}$, via intermediate cluster m, which is of length $\delta_G(j, m) + \delta_G(i, k) + \delta_G(m, l) + 2$ (see Fig. 3). The alternate path may be shorter, for example, if the pairs of nodes j and m, i and k, and m and l are neighbors in G whereas $\delta_G(j, k) + \delta_G(i, l) > 4$. Consider, for instance, a 4×4 torus as G, with i, j, k, l, and m located at (2, 1), (0, 0), (2, 2), (0, 2), and (0, 1). Despite the observation above, the fact that R^{sw} uses at most one intercluster link makes it preferable in a modular system with hierarchical packaging where intracluster communication is much faster than intercluster data exchange.

3. Historical Perspective

One of the referees of this paper pointed out that swapped networks are the same as OTIS (optical transpose interconnection system) architectures which have been extensively studied by other researchers. Tracing the history of OTIS, the author discovered that its roots go back to 1993, when Marsden et al published a 3-page note in *Optics Letters* [Mars93] suggesting a topology in which nodes (i, j) and (j, i) are linked via an optical channel. It appears that transfer of the OTIS idea to the computer architecture and parallel processing community occurred, in part, due the 1998 PhD dissertation of C.-F. Wang at University of Florida, under Sartaj Sahni, and publication of its results beginning in 1997 (see, for example, [Raja98], [Wang98], [Oste00], [Wang00], [Wang01]). Architectural and some topological considerations for OTIS networks have been studied by Zane et al [Zane00] and Day and Al-Ayyoub [Day02], among others.

Concurrent with the developments cited above, and before any reference to OTIS appeared in the computer architecture or parallel processing literature, Chi-Hsiang Yeh (a former doctoral student of the author) proposed swapped networks [Yeh96a], [Yeh96b], [Yeh96c] as tools for unifying and extending a number of known hierarchical networks. Prominent among these prior architectures were the two-level special case of WK-recursive networks [Dell87], [Dell88] (beyond two levels, WK-recursive and swapped networks diverge in structure and do not have much in common), hierarchical cubic networks [Ghos95], and recursively fully connected networks [Yeh96]. The unification was due to the replacement of complete-graph or hypercube component networks of the prior architectures with an arbitrary graph.

It thus seems that it is the OTIS community that has failed to take note of our prior work on swapped networks, and not the other way around. Incidentally, “swapped network” is a much more descriptive, as well as more appropriate, name than OTIS because many of the algorithms and topological properties discussed in the literature are independent of the implementation technology. Reference to optical implementation may in fact serve to discourage researchers not involved in optical computing or communications to overlook the results which may be useful in other contexts.

4. Topological Parameters

In this section, we relate some of the topological parameters of a swapped network to the respective parameters of the basis network G.

Theorem 1: Degree and diameter of swapped networks – If G has node degree d and diameter D, the degree and diameter of $Sw(G)$ are $d^{sw} = d + 1$ and $D^{sw} = 2D + 1$, respectively [Yeh96a]. ■

An implication of Theorem 1 is that if the basis network has logarithmic/sublogarithmic diameter, then so does the resulting swapped network.

Theorem 2: If the n-node graph G is node-symmetric and has an average internode distance Δ, then $Sw(G)$ has an average internode distance Δ^{sw} satisfying $\Delta^{sw} < 2\Delta + 1 - (\Delta + 1)/n$, where the right-hand-side expression is the average internode distance with respect to the routing algorithm R^{sw}.

Proof: Let v_{ij} denote node j in the ith component graph G and $\delta(v, v')$ denote the distance between nodes v and v' when the routing algorithm R^{sw} is used. Then, the average internode distance $\Delta^{R{sw}}$ of $Sw(G)$ with respect to R^{sw} is derived as follows:

$$
\begin{aligned}
n^4\Delta^{R{sw}} &= \textstyle\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\delta(v_{ij}, v_{kl}) \\
&= \textstyle\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{l=0}^{n-1}\delta(v_{ij}, v_{il}) \; + \\
&\quad \textstyle\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{k=0(k\neq i)}^{n-1}\sum_{l=0}^{n-1}[\delta(v_{ij}, v_{ik}) + 1 + \delta(v_{ki}, v_{kl})] \\
&= n^3\Delta + n\textstyle\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\sum_{k=0(k\neq i)}^{n-1}\delta(v_{ij}, v_{ik}) \\
&\quad + n^3(n-1) + n\textstyle\sum_{i=0}^{n-1}\sum_{k=0(k\neq i)}^{n-1}\sum_{l=0}^{n-1}\delta(v_{ki}, v_{kl}) \\
&= n^4 + n^3(\Delta - 1) + n^2\textstyle\sum_{j=0}^{n-1}\sum_{k=1}^{n-1}\delta(v_{0j}, v_{0k}) \\
&\quad + n(n-1)\textstyle\sum_{i=0}^{n-1}\sum_{l=1}^{n-1}\delta(v_{0i}, v_{0l}) \\
&= n^4 + n^3(\Delta - 1) + n^2\textstyle\sum_{j=0}^{n-1}\sum_{k=1}^{n-1}\delta(v_{0j}, v_{0k}) \\
&\quad + n^3(n-1)\Delta \\
&= n^4(\Delta + 1) - n^3 \\
&\quad + n^2\,[\,\textstyle\sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\delta(v_{0j}, v_{0k}) - \sum_{j=0}^{n-1}\delta(v_{0j}, v_{00})\,] \\
&= n^4(\Delta + 1) - n^3 + n^2[n^2\Delta - n\Delta] \\
&= n^4(2\Delta + 1) - n^3(\Delta + 1) \quad ■
\end{aligned}
$$

Theorem 3: Bisection width B^{sw} of the swapped network whose basis network G has bisection width B is upper bounded by $n \times min(n/4, B)$.

Proof: An upper bound u for the bisection width of a network can be established by showing a particular bisection cut of size u. Let n be even (the case of odd n is more involved, but similar). A bisection cut placing basis networks 0 through $n/2 - 1$ on one side and $n/2$ through $n - 1$ on the other is of size $n^2/4$. Bisecting each basis network, so that nodes 0 through $n/2 - 1$ are on one side and nodes $n/2$ through $n - 1$ on the other, leads to the second bound nB. Note that in the latter case, all swap links are confined to the same side of the bisected network. ■

We conjecture that the bound in Theorem 3 is tight for a wide class of symmetric basis networks, but have been unable to derive a formula for the exact bisection width.

Swapped networks can be built up recursively beginning with a fairly small basis network. The number of nodes increase from n to n^2 to n^4 and so on. It is fairly easy to prove that when so constructed, the node degree of a recursive swapped network is a double-logarithmic function of its size. In this paper, we continue to focus on two-level swapped networks.

5. Structural Properties

In this section, we cover some structural properties of swapped networks. These shed light on the relationships between classes of swapped networks and facilitate not only systematic studies of such networks, but also comparison with competing networks.

Theorem 4: If H is a subgraph of G, then $Sw(H)$ is a subgraph of $Sw(G)$.

Proof: Number nodes of G from 0 through $n - 1$. Let the nodes of H be $i_1, i_2, \ldots, i_k$. Then, the subgraphs H in clusters $i_1, i_2, \ldots, i_k$ of $Sw(G)$, along with their swap links, form $Sw(H)$. ■

Theorem 4 can be easily extended to the case of r disjoint subgraphs $H_1, H_2, \ldots, H_r$, leading to the associated r swapped networks $Sw(H_i)$ forming disjoint subgraphs of $Sw(G)$. This is a generalized version of Theorem 1 in [Day02] which pertains to the case of all the H_i being identical.

Hamiltonicity and Hamiltonian connectivity are useful properties of interconnection networks.

Theorem 5: If G is Hamiltonian-connected, then so is $Sw(G)$. Also, Hamiltonian connectivity of G ensures the Hamiltonicity of $Sw(G)$.

Proof outline: Consider nodes v_{ij} and v_{kl} in clusters i and k of $Sw(G)$ and a Hamiltonian path between them in the n-node complete graph formed by the clusters (supernodes). This path can be converted to a Hamiltonian path between v_{ij} and v_{kl} by simply replacing the $x \to y \to z$ segment of the Hamiltonian path through the clusters by $v_{xy} \to v_{yx} \sim^{\mathrm{H}} v_{yz} \to v_{zy}$, where $v_{yx} \sim^{\mathrm{H}} v_{yz}$ represents a Hamiltonian path between nodes x and z within cluster y. ■

Modular construction of networks is of great significance. Given limitations of the currently available and forthcoming packaging technologies for digital systems, a hierarchy of packaging levels is imposed so that crossing the packaging boundaries is undesirable, not only in terms of implementation cost, but also with regard to communication performance [Parh00]. Swapped networks are naturally modular.

Theorem 6: For an integer m that divides n, an n^2-node swapped network can be partitioned into m modules having $(n^2/m)(1 - 1/m)$ external links.

Proof: Each of the m modules holds n/m clusters, or n^2/m nodes. Within a cluster, n/m of the nodes do not have links to outside the module. So, the number of intermodule links is $(n/m)(n - n/m)$. ■

The modularity suggested by Theorem 6, requiring roughly 1 external link per node in a module of n^2/m nodes, is better than those of the hypercube that requires $(n^2/m)\log_2 m$ links per module. It also compares favorably with other networks of similar performance.

6. Fault Tolerance

Swapped networks are quite robust. This is intuitively justified by noting that the complete connectivity among clusters is not much affected even if all nodes in a cluster, or parts of several clusters, fail. As exemplified by the pair of paths depicted in Fig. 3, there are typically many node- and edge-disjoint paths between pairs of nodes in various clusters. The existence of intercluster links also has a positive effect on fault tolerance within the same cluster in the sense that nodes becoming inaccessible in the cluster due to faults may be reachable through intermediaries in other clusters. The following three theorems collectively establish the strong fault tolerance features of swapped networks.

Theorem 7: If G is h-connected, then $Sw(G)$ is also h-connected (Theorem 6 of [Day02]). ■

Theorem 8: If the fault diameter of the basis network G is $D + \varepsilon$, the fault diameter of $Sw(G)$ is no greater than $2D + 3 + \varepsilon$.

Proof outline: Let d be the minimum node degree of G. Then, the fault diameter of G being $D + \varepsilon$ means that for $d - 1$ or fewer faults in G, the diameter of the remaining (connected) network is $D + \varepsilon$ or less. Because the connectivity of $Sw(G)$ is also d, we need to consider the distance from a source node v_{ij} to a destination node v_{kl} in the surviving network when $Sw(G)$ has $d - 1$ or fewer faults. Consider the source node j along with d of its neighbors in the source cluster i. Let these neighbors be m_1, $m_2, \ldots, m_d$. These $d + 1$ nodes have at least d swap links to different clusters. Consider also nodes $m_1, m_2, \ldots, m_d$ in the destination cluster k. There should be some m_g for which both the cluster m_g and node m_g in cluster k are fault-free. This is because at least d faults are needed to have a fault in cluster m_r or node m_r in cluster k for every r. Then, it is easy to see that the path from v_{ij}, perhaps via one of its neighbors in cluster i, to some node x in cluster m_g, then to node k of cluster m_g ($\leq D$ hops), to node m_g of cluster k, and finally to node v_{kl} ($\leq D + \varepsilon$ hops) is of length no greater than $2D + 3 + \varepsilon$. ■

We conjecture that the tighter bound of $2D + 2 + \varepsilon$ may be provable for the fault diameter of a swapped network. Also, placing mild restrictions on the structure of G may allow the establishment of the optimal $2D + 1 + \varepsilon$ as the fault diameter. These improvements are being investigated.

Swapped networks are also provably survivable in the sense that they do not contain any obvious vulnerability points [Hobb91]. Consider, for example, the behavior of swapped networks under complete cluster failures, as opposed to a small number of node failures. The following result is established in a manner similar to Theorem 8.

Theorem 9: The diameter of an incomplete swapped network, with $d - 1$ or fewer clusters completely removed, where d is the minimum node degree of G, is no more than $2D + 2$. ■

7. Conclusion

We derived some general properties of a swapped network $Sw(G)$ based on parameters and structure of the basis network G. In particular, we showed that swapped networks are cost-effective, modular, packageable, and quite robust. Table 1 indicates how the main topological parameters of certain classes of interconnection networks change as the network size is scaled from n to n^2. The various parameters are formulated in terms of the respective parameters for the network of size n. As evident from Table 1, swapped networks offer a mechanism for increasing the size of a network with relatively small cost increase, while limiting the deterioration of topological parameters and ensuring strong fault tolerance.

Work in progress includes expanding our results on the properties of swapped network to include other topological attributes as well as a deeper analysis of their performance parameters and robustness attributes. In particular, a fault-tolerant routing algorithm is required if the existence of multiple node- or edge-disjoint paths is to be practically exploited. Other fault tolerance notions, such as fault-survivability and fault-Hamiltonicity, are also under investigation.

References

[Arde81] Arden, B.W. and H. Lee, "Analysis of Chordal Ring Networks," *IEEE Trans. Computers*, Vol. 30, pp. 291-295, April 1981.

[Day02] Day, K. and A. Al-Ayyoub, "Topological Properties of OTIS-Networks," *IEEE Trans. Parallel and Distributed Systems*, Vol. 13, No. 4, pp. 359-366, April 2002.

[Dell87] Della Vecchia, G. and C. Sanges, "Recursively Scalable Networks for Message Passing Architectures," *Proc. Conf. Parallel Processing & Applications*, pp. 33-40, 1987.

[Dell88] Della Vecchia, G. and C. Sanges, "A Recursively Scalable Network for VLSI Implementation," *Future Generation Computer Systems*, pp. 235-243, 1998.

[Ghos95] Ghose, K. and K.R. Desai, "Hierarchical Cubic Networks," *IEEE Trans. Parallel and Distributed Systems*, Vol. 6, No. 4, pp. 427-435, April 1995.

[Hobb91] Hobbs, A.M., "Network Survivability," in *Applications of Discrete Mathematics*, ed. by J.G. Michaels and K.H. Rosen, McGraw-Hill, pp. 332-353, 1991.

[Kwai98] Kwai, D.-M. and B. Parhami, "Pruned Three-Dimensional Toroidal Networks," *Information Processing Letters*, Vol. 68, No. 4, pp. 179-183, November 30, 1998.

[Mars93] Marsden, G., P. Marchand, P. Harvey, and S. Esener, "Optical Transpose Interconnection System Architectures," *Optics Letters*, Vol. 18, No. 13, pp. 1083-1085, July 1, 1993.

[Oste00] Osterloh, A., "Sorting on the OTIS-Mesh," *Proc. 14th Int'l Parallel and Distributed Processing Symp.*, pp. 269-274, 2000.

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

[Parh00] Parhami, B. and D.-M. Kwai, "Challenges in Interconnection Network Design in the Era of Multiprocessor and Massively Parallel Microchips," *Proc. Int'l Conf. Communications in Computing*, pp. 241-246, 2000.

[Parh01] Kwai, D.-M. and B. Parhami, "A Unified Formulation of Honeycomb and Diamond Networks," *IEEE Trans. Parallel and Distributed Systems*, Vol. 12, pp. 74-80, January 2001.

[Parh04] Parhami, B. and D.-M. Kwai, "Incomplete *k*-ary *n*-cube and Its Derivatives," *J. Parallel and Distributed Computing*, Vol. 64, pp. 183-190, February 2004.

[Parh05] Parhami, *Computer Architecture: From Microprocessors to Supercomputers*, Oxford, 2005.

[Raja98] Rajasekaran, S. and S. Sahni, "Randomized Routing, Selection, and Sorting on the OTIS-Mesh," *IEEE Trans. Parallel and Distributed Systems*, Vol. 9, No. 9, pp. 833-840, September 1998.

[Wang98] Wang, C.-F. and S. Sahni, "Basic Operations on the OTIS-Mesh Optoelectronic Computer," *IEEE Trans. Parallel and Distributed Systems*, Vol. 9, No. 12, pp. 1226-1236, December 1998.

[Wang00] Wang, C.-F. and S. Sahni, "Image Processing on the OTIS-Mesh Optoelectronic Computer," *IEEE Trans. Parallel and Distributed Systems*, Vol. 11, pp. 97-109, February 2000.

[Wang01] Wang, C.-F. and S. Sahni, "Matrix Multiplication on the OTIS-Mesh Optoelectronic Computer," *IEEE Trans. Computers*, Vol. 50, pp. 635-646, July 2001.

[Yeh96] Yeh, C.-H. and B. Parhami, "Recursively Fully-Connected Networks: A Class of High-Performance Low-Degree Interconnection Networks," *Proc. 11th Int'l Conf. Computers and Their Applications*, pp. 227-230, 1996.

[Yeh96a] Yeh, C.-H. and B. Parhami, "Swapped Networks: Unifying the Architectures and Algorithms of a Wide Class of Hierarchical Parallel Processors," *Proc. Int'l Conf. Parallel and Distributed Systems*, pp. 230-237, 1996.

[Yeh96b] Yeh, C.-H. and B. Parhami, "Hierarchical Swapped Networks: Efficient Low-Degree Alternatives to Hypercube and Generalized Hypercube," *Proc. Int'l Symp. Parallel Architectures, Algorithms, and Networks*, pp. 90-96, 1996.

[Yeh96c] Yeh, C.-H. and B. Parhami, "Recursive Hierarchical Swapped Networks: Versatile Interconnection Architectures for Highly Parallel Systems," *Proc. 8th IEEE Symp. Parallel and Distributed Processing*, pp. 453-460, 1996.

[Zane00] Zane, F., P. Marchand, R. Paturi, and S. Esener, "Scalable Network Architectures Using the Optical Transpose Interconnection System," *J. Parallel and Distributed Computing*, Vol. 60, No. 5, pp. 521-538, 2000.

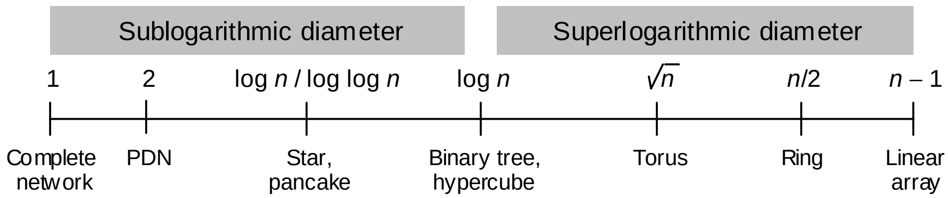
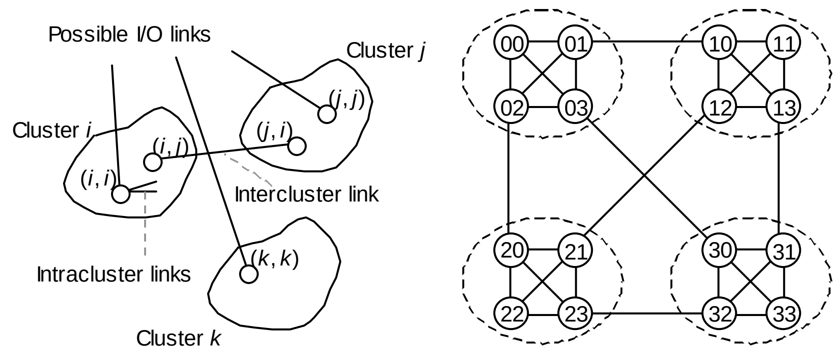


Fig. 1. The spectrum of interconnection networks in terms of diameter for size *n*.



(a) Intracluster, intercluster, and I/O links

(a) An example 16-node swapped network

Fig. 2. The general structure of a swapped network and an example network with the 4-node complete graph as its basis.

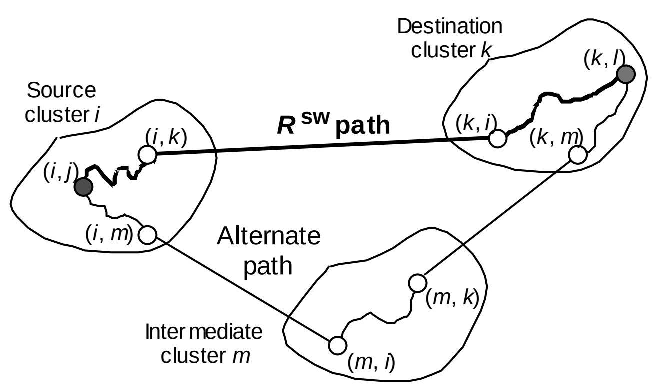


Fig. 3. The path prescribed by R^{sw} and an alternate path from v_{ij} to v_{kl}.

Table 1. Change in topological parameters for various ways of increasing network size from *n* to n^2. For example, an entry of *D* for diameter means that there is no change when the network size is squared and 2*D* means that the diameter is doubled. Squared network refers to the cross product of a network with itself.

Network	Degree	Diameter	Avg. dist.	Bisection
Complete	$d^2 + 2d$	D	Δ	$4B^2$
Squared network	$2d$	$2D$	2Δ	nB
Hypercube	$2d$	$2D$	2Δ	$2B^2$
2D square torus	d	D^2	$4\Delta^2/3$	B^2
Swapped network	$d + 1$	$2D + 1$	$2\Delta + 1$	?