Robustness Attributes of Interconnection Networks for Parallel Processing

Behrooz Parhami
ECE Dept., Univ. California, Santa Barbara, USA

Int'l Supercomputing Conf. in Mexico (ISUM2010)
Keynote talk: March 4, 2010

Presentation Outline

Interconnection Networks

The Reliability Problem

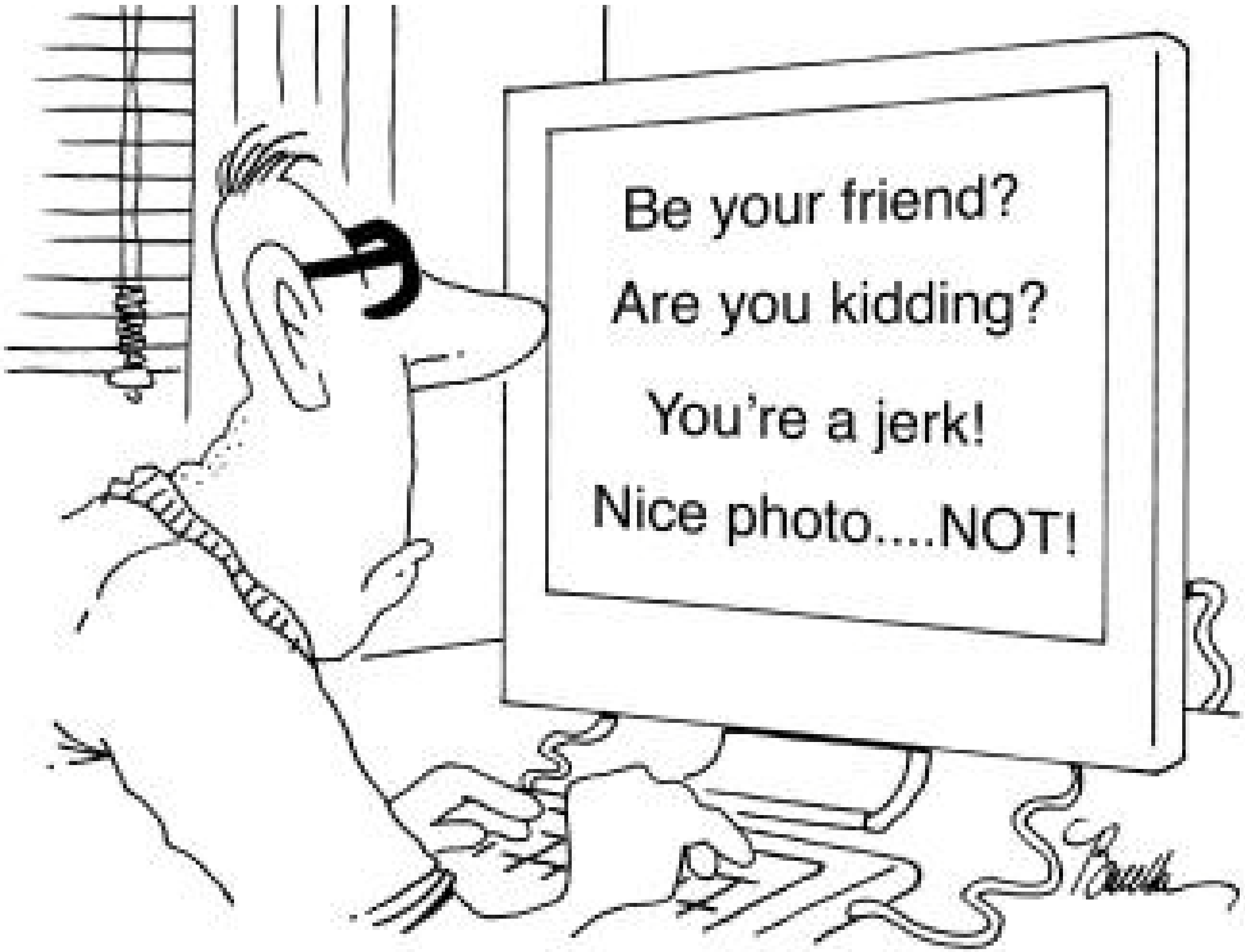
Robustness Attributes

Deriving New Networks

Problems and Challenges

Abstract of talk and speaker's biography are on the last slide

My Talk's Most Interesting Part



I wanted to start my talk with something funny, but I could not find any funny stories related to "network robustness" or plain "interconnection networks." My topic isn't funny, I guess!

This cartoon with the caption "unsocial networking" was as close as I could get to today's topic

Presentation Outline

Interconnection Networks

A sea of choices
Evaluation criteria

The Reliability Problem

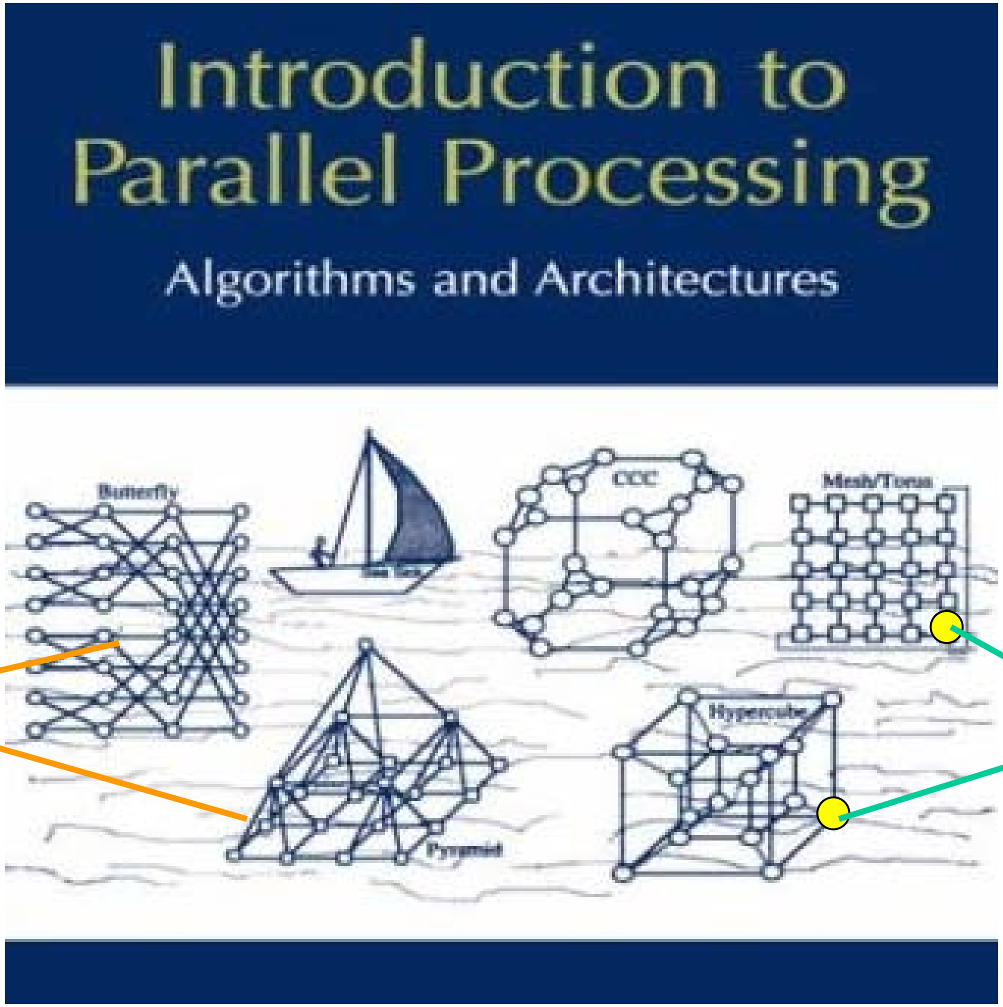
Robustness Attributes

Deriving New Networks

Problems and Challenges

Parallel Computers

Parallel computer = Nodes + Interconnects (+ Switches)

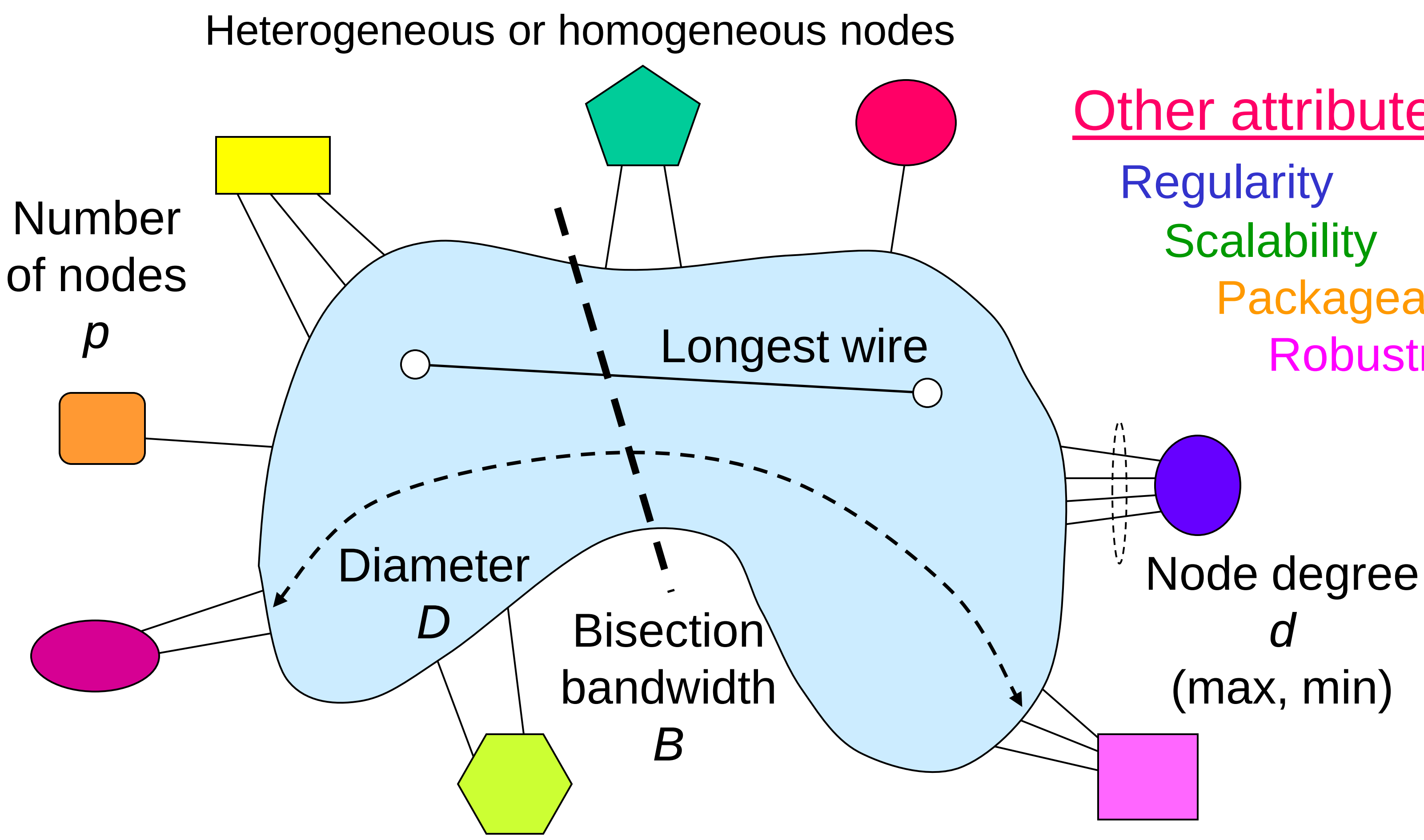


B. Parhami, Plenum Press, 1999

Interconnects, communication channels, or links

Nodes or processors

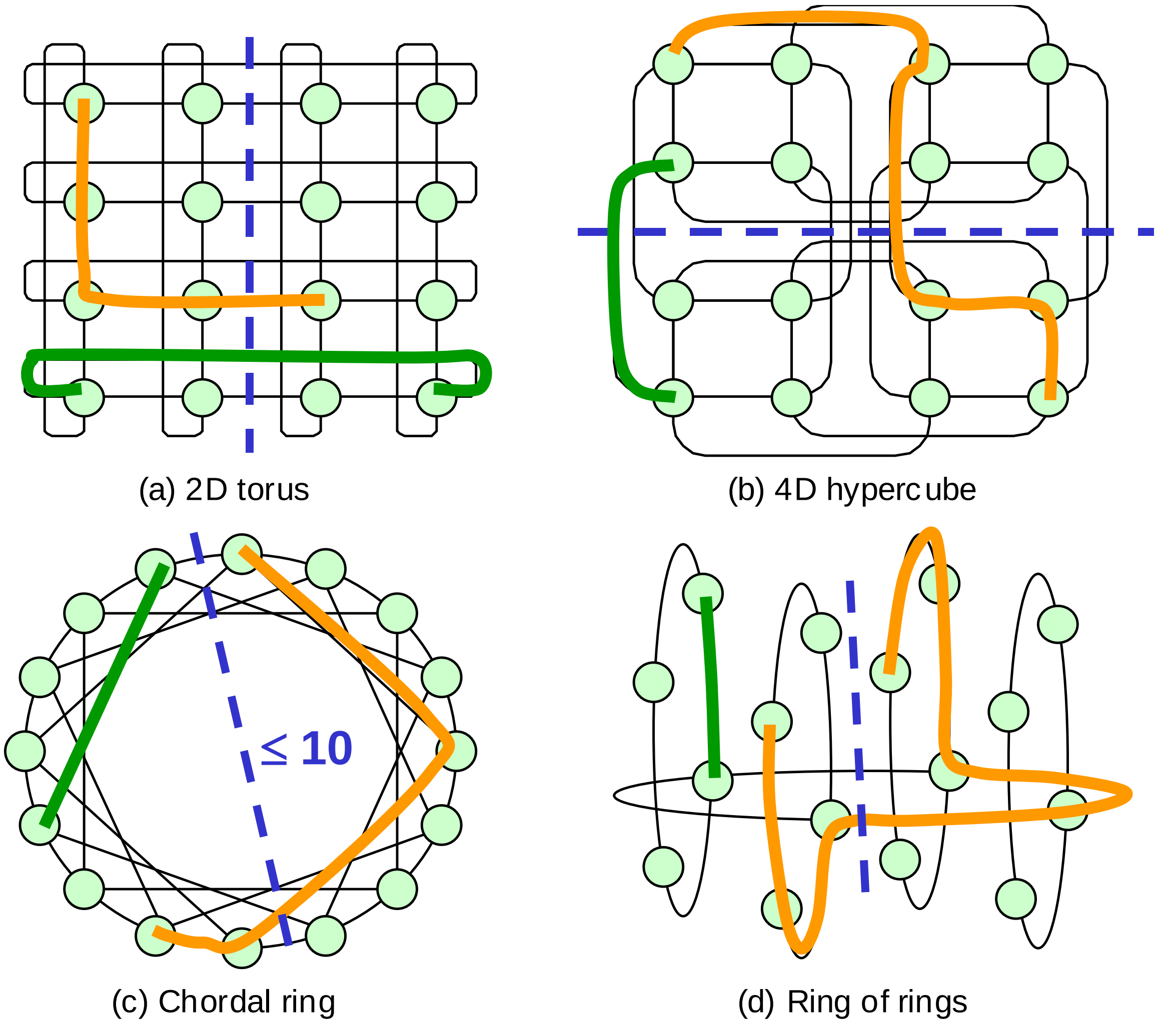
Interconnection Networks



Other attributes:

- Regularity
- Scalability
- Packageability
- Robustness

Four Example Networks



Nodes $p = 16$

Degree $d = 4$

Diameter D

Avg. distance Δ

Bisection B

Longest wire

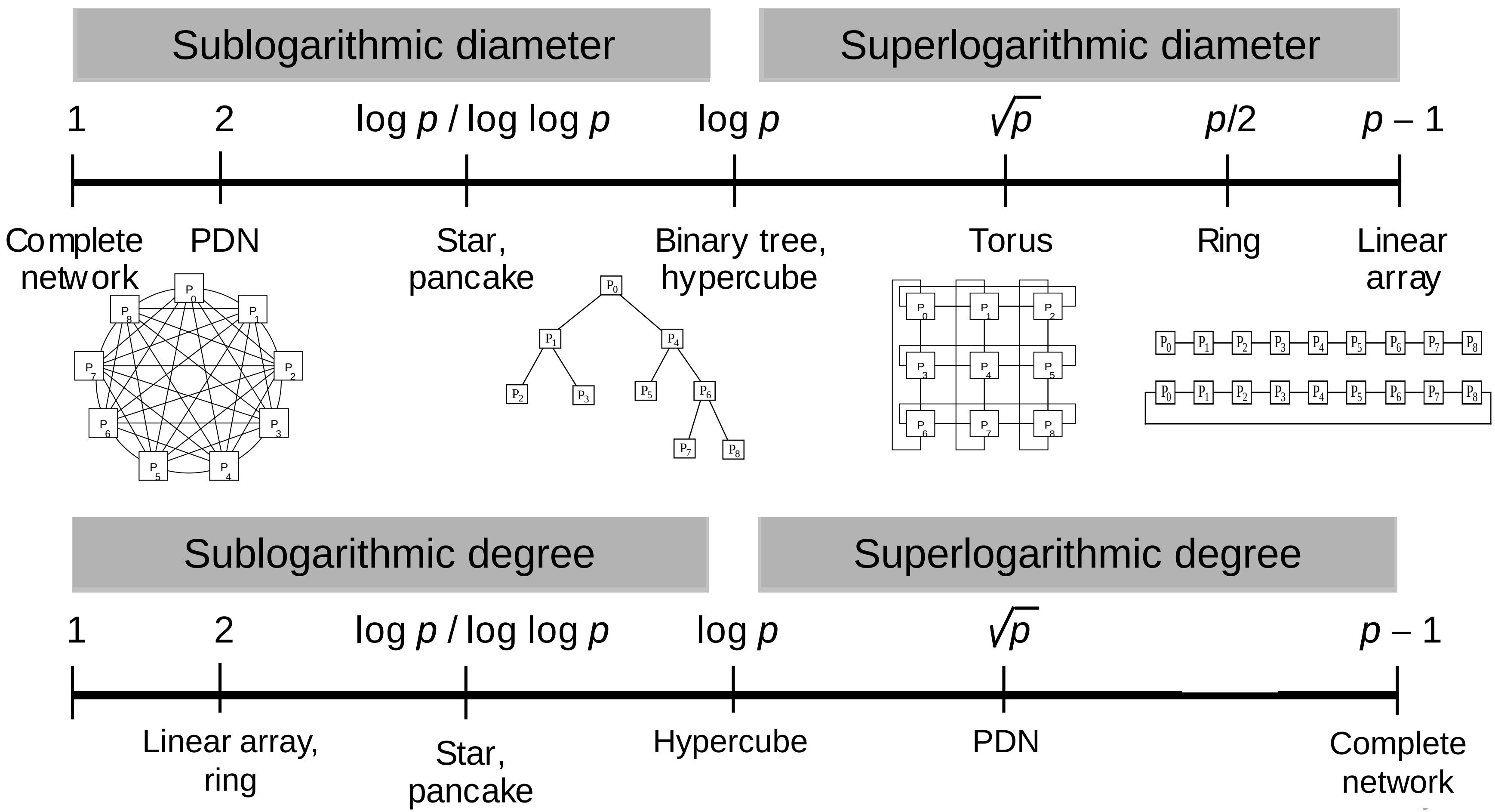
Regularity

Scalability

Packageability

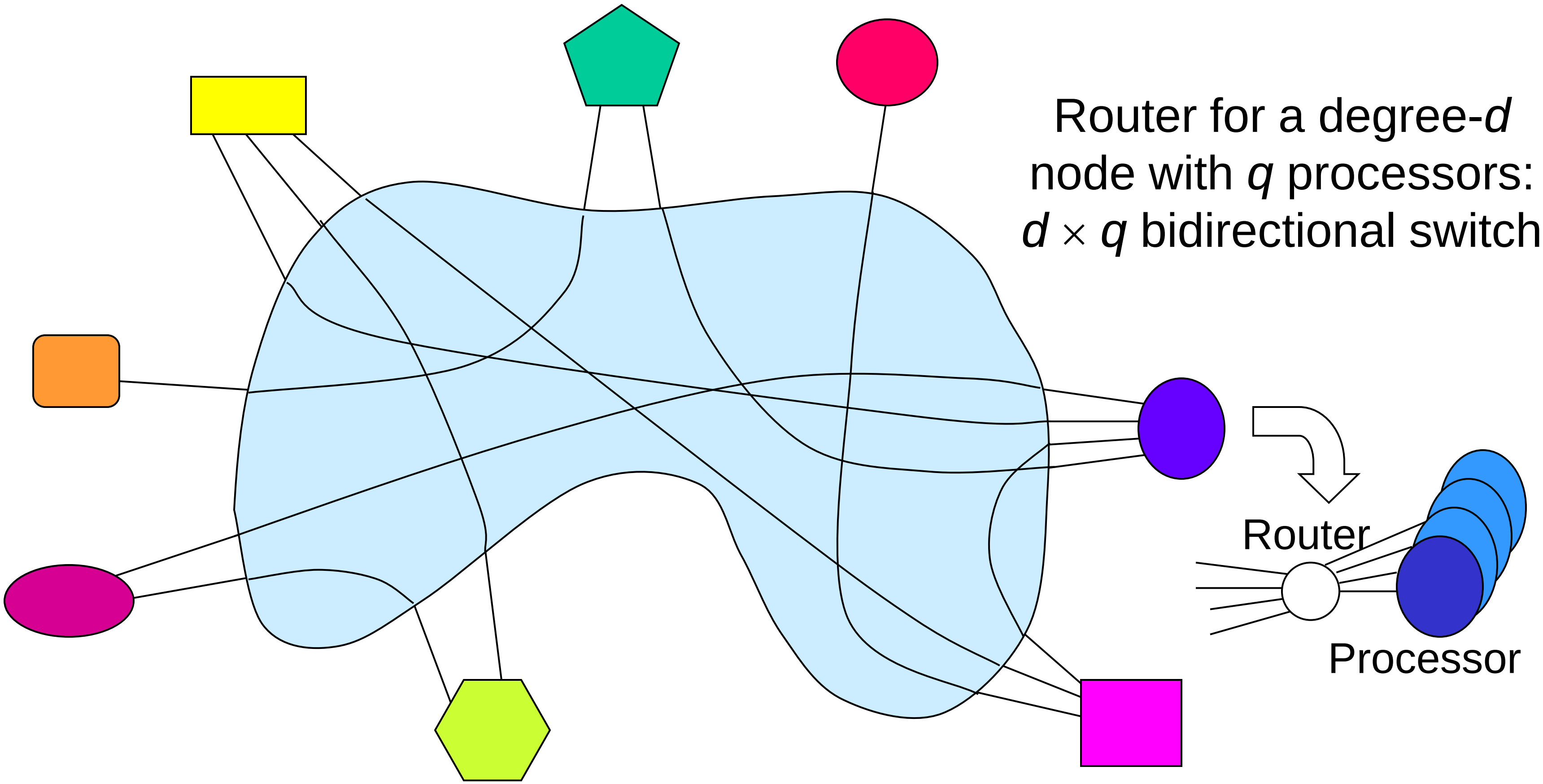
Robustness

Spectrum of Networks



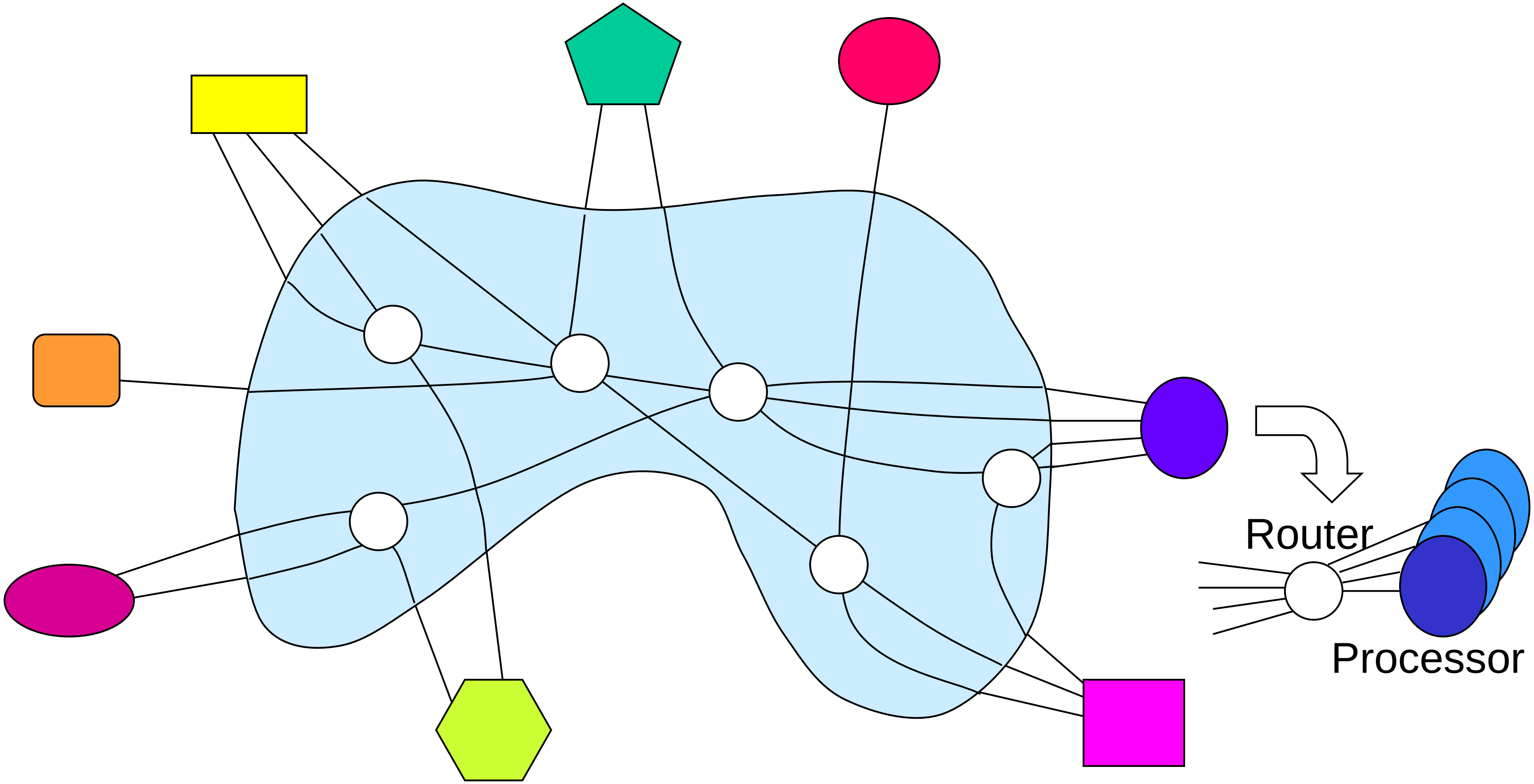
Direct Networks

Nodes (or associated routers) directly linked to each other

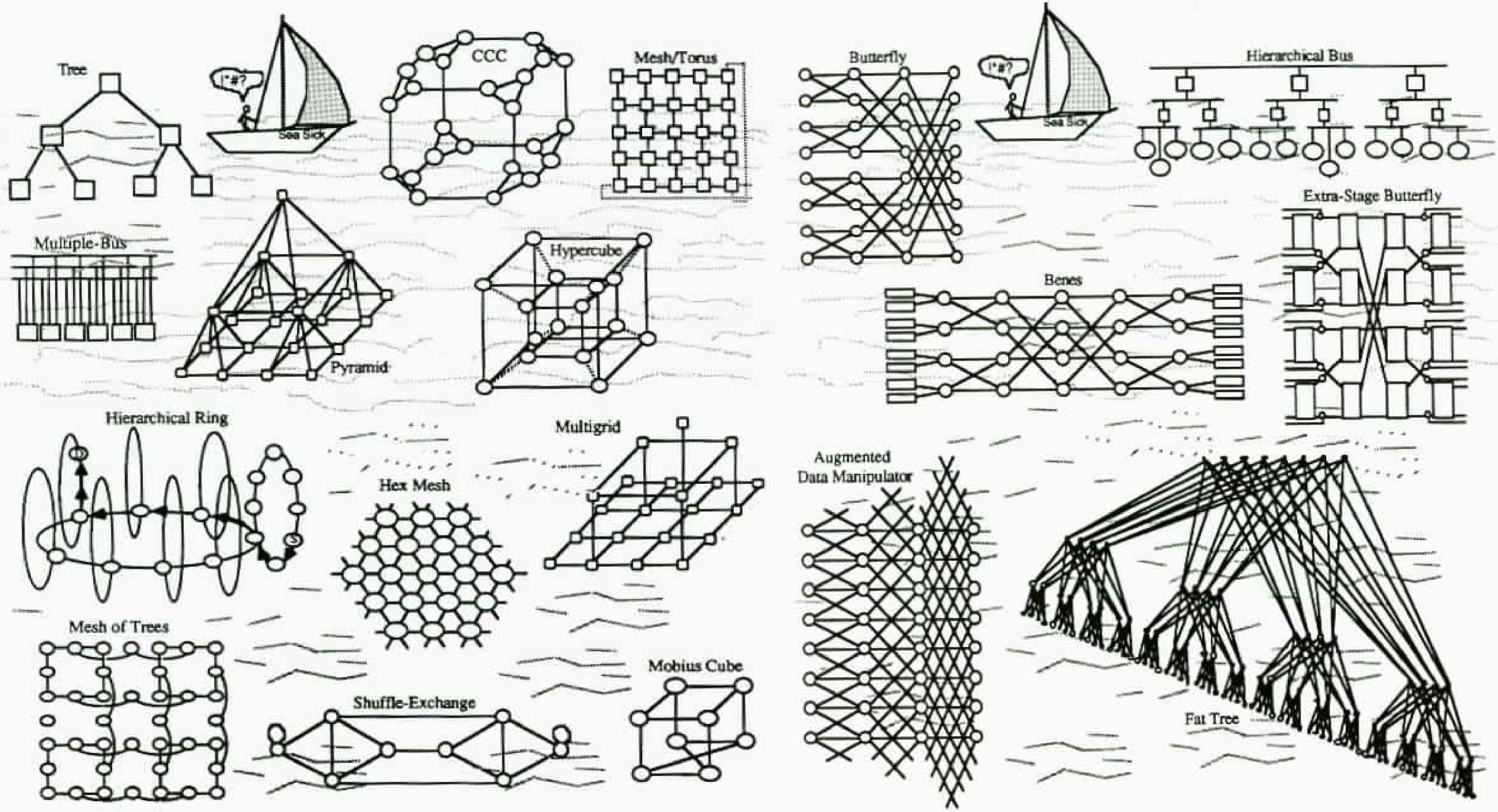


Indirect Networks

Nodes (or associated routers) linked via intermediate switches



A Sea of Networks



Moving Full Circle

2000s
1960s
Mesh-based
(ILLIAC IV)

Direct to indirect,
shared memory

1970s
Butterfly,
other MINs

Lower diameter,
message passing

1980s
Hypercube,
bus-based

Greater
bandwidth

1990s
Fat tree,
LAN-based

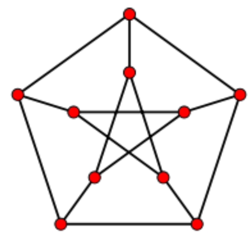
Scalability,
local wires

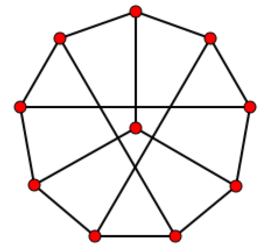
So, only a small portion of the sea of networks has been explored in practical parallel computers

The (d, D) Graph Problem

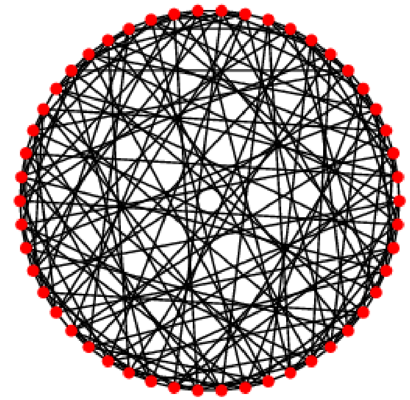
Suppose you have an unlimited supply of degree-d nodes
How many can be connected into a network of diameter D?

Example 1: $d = 3$, $D = 2$;
10-node Petersen graph





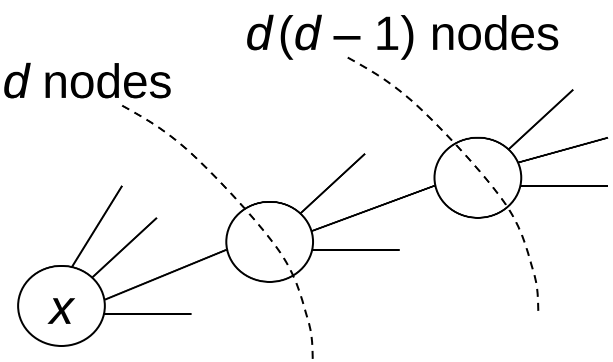
Example 2: $d = 7$, $D = 2$;
50-node Hoffman-Singleton graph



Moore bound (undirected graphs)

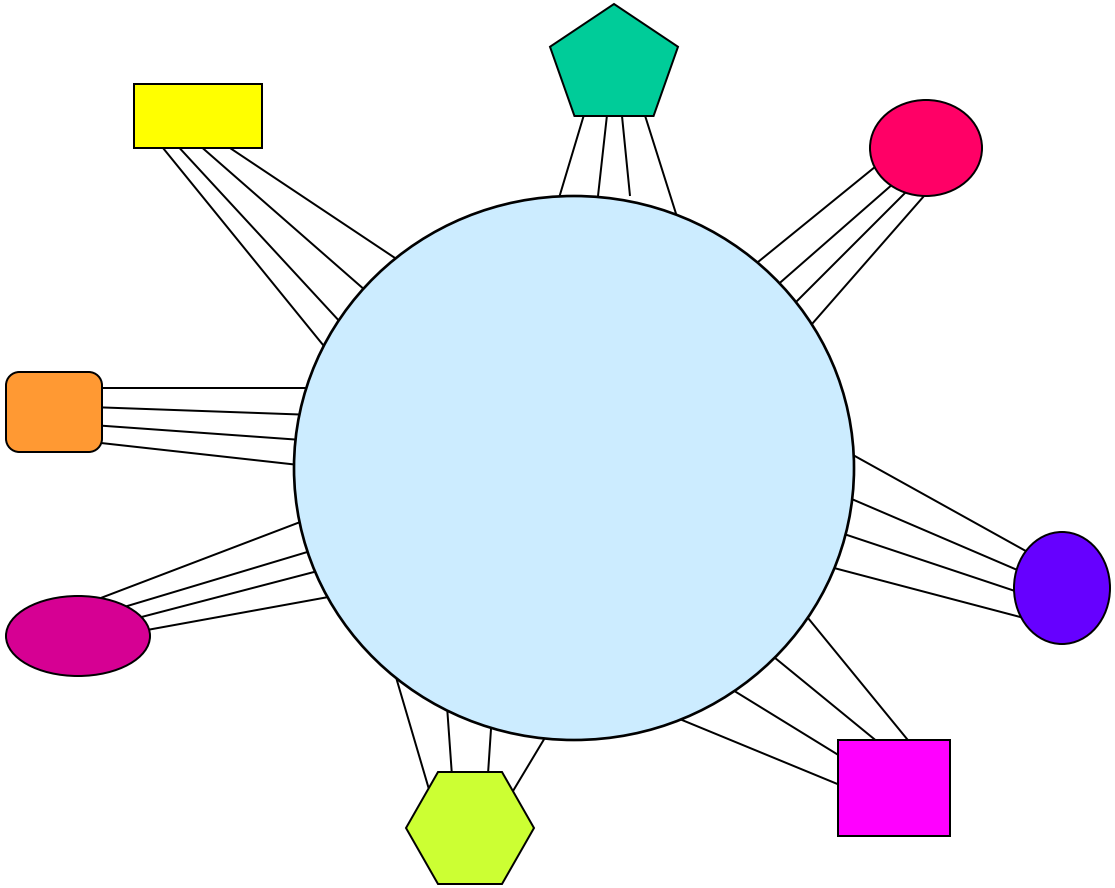
$$p \le 1 + d + d(d-1) + \ldots + d(d-1)^{D-1}$$
$$= 1 + d[(d-1)^D - 1]/(d-2)$$

Only ring with odd p and a few other networks match this bound

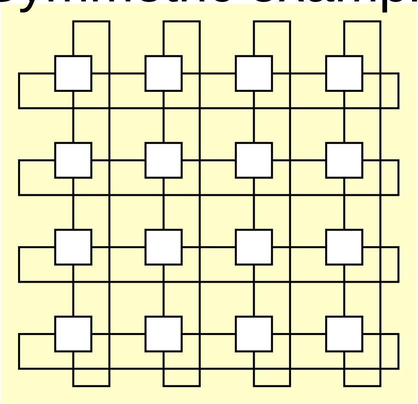


Symmetric Network

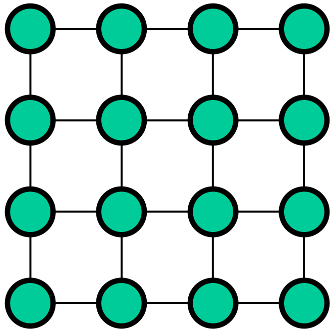
Viewed from any node, it looks the same



Symmetric example

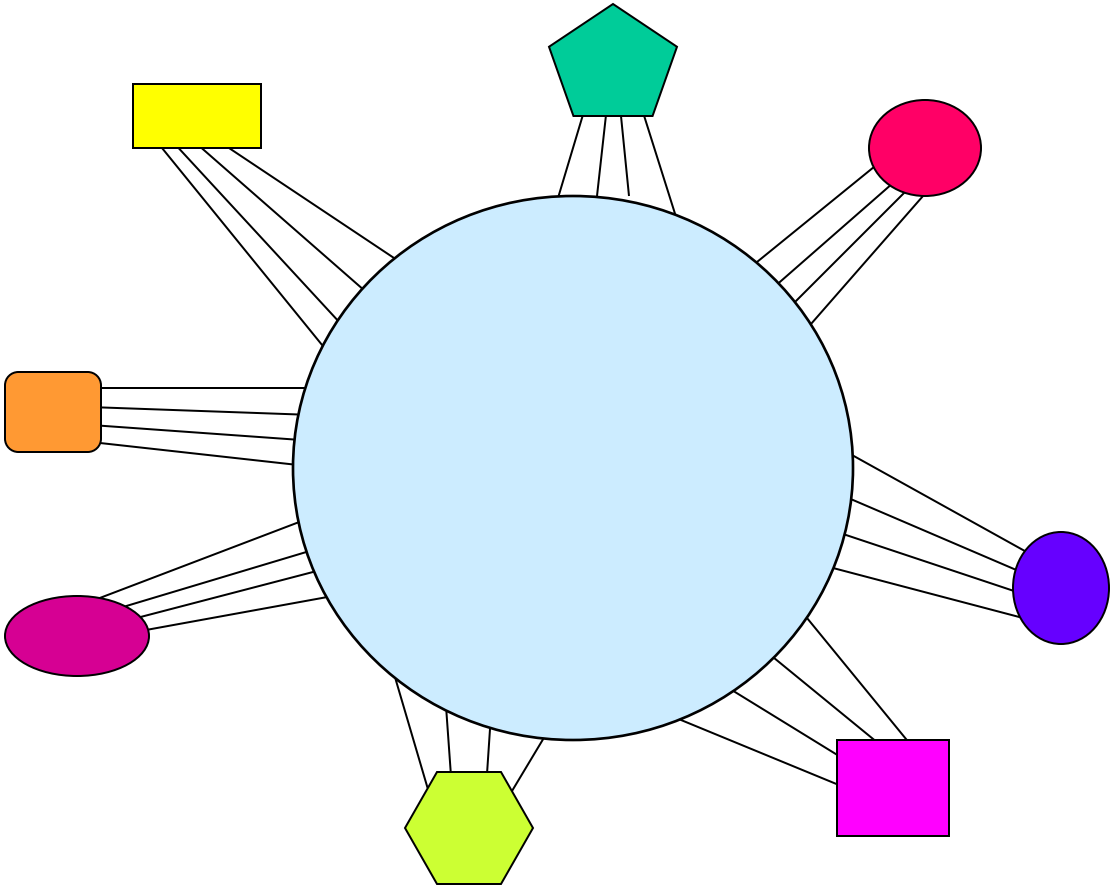


Asymmetric example



Implications of Symmetry

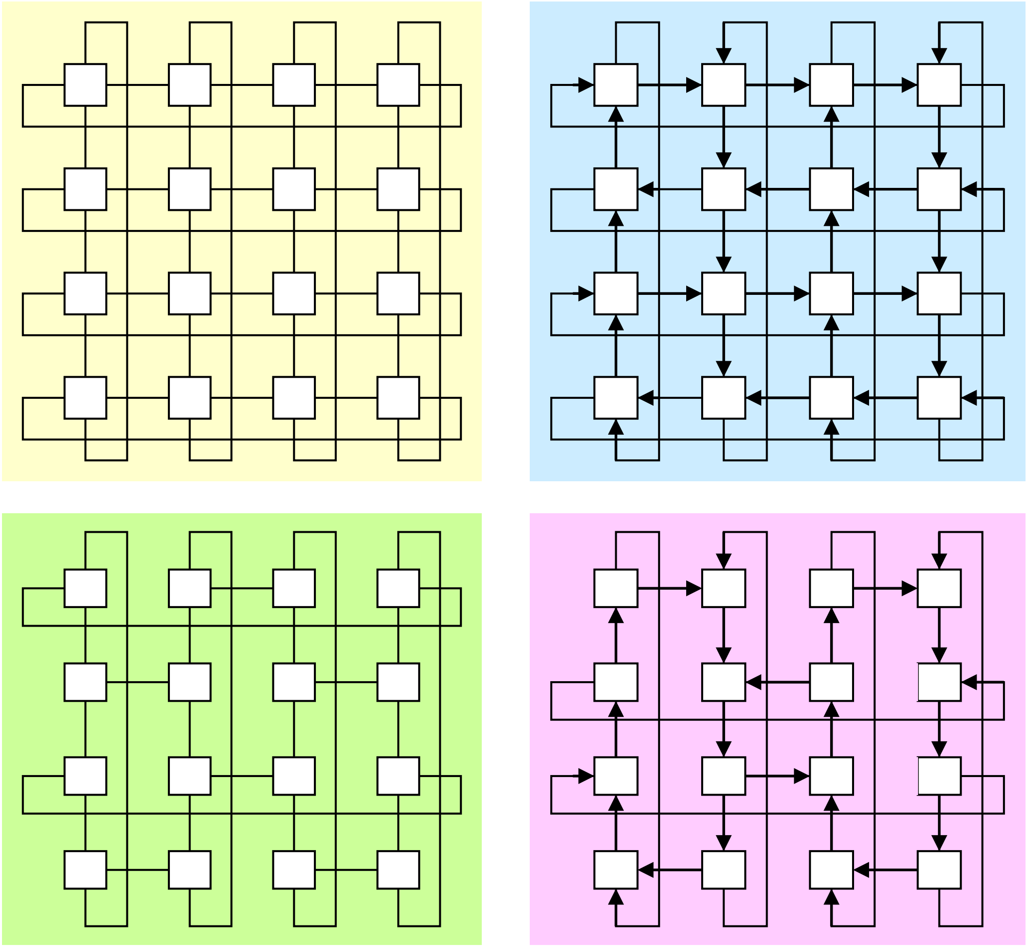
A degree-4 network



- Routing algorithm the same for every node
- No weak spots (critical nodes or links)
- Maximum number of alternate paths feasible
- Derivation and proof of properties easier

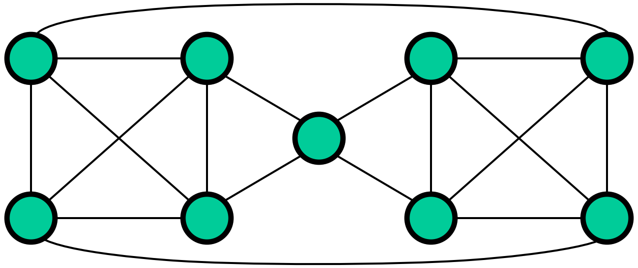
We need to prove a particular topological or routing property for only one node

A Necessity for Symmetry



Uniform node degree:
$d = 4;\ \ d_{in} = d_{out} = 2$

An asymmetric network
With uniform node degree



Uniform node degree
is necessary but not
sufficient for symmetry

Presentation Outline

Interconnection Networks

The Reliability Problem

Outage detection/diagnosis
Building reliable networks

Robustness Attributes

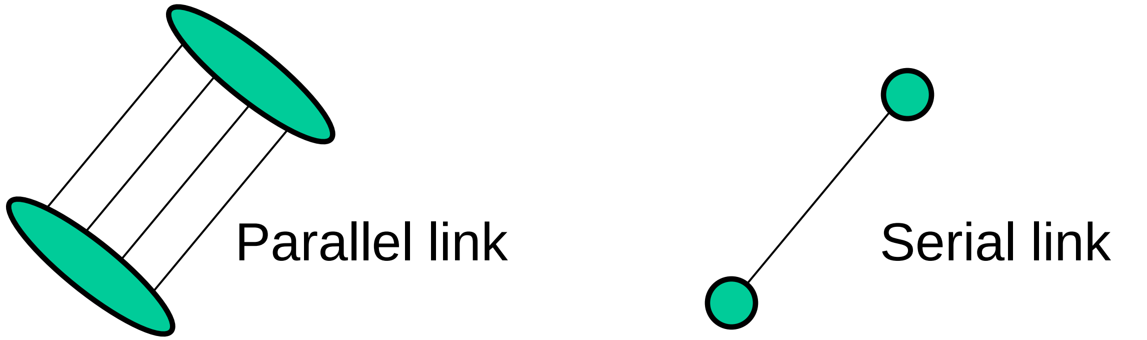
Deriving New Networks

Problems and Challenges

Link Malfunctions

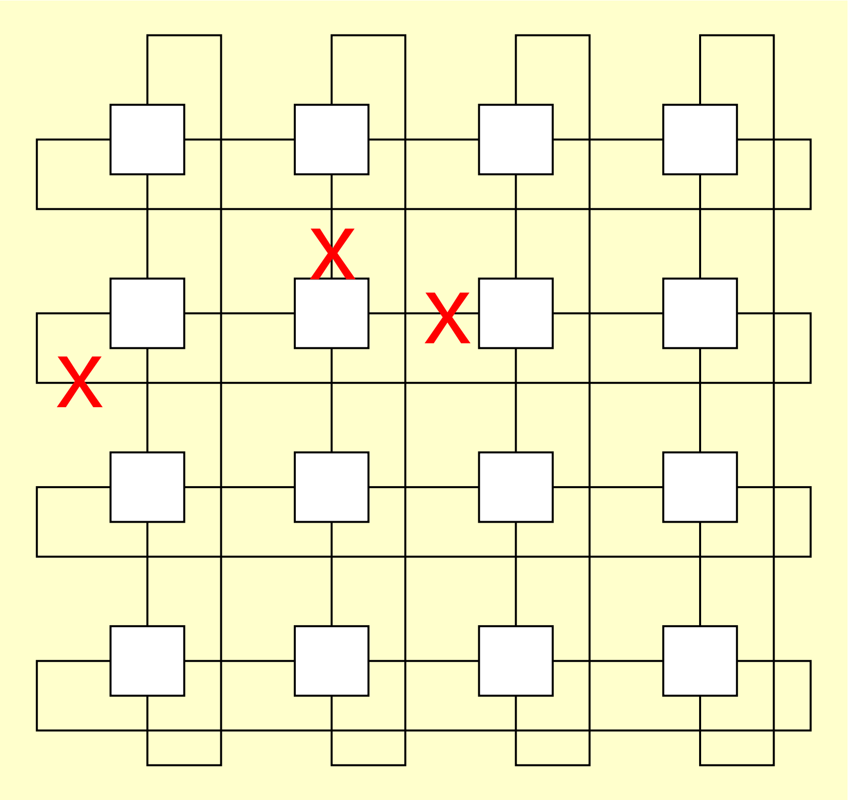
Link data errors or outage

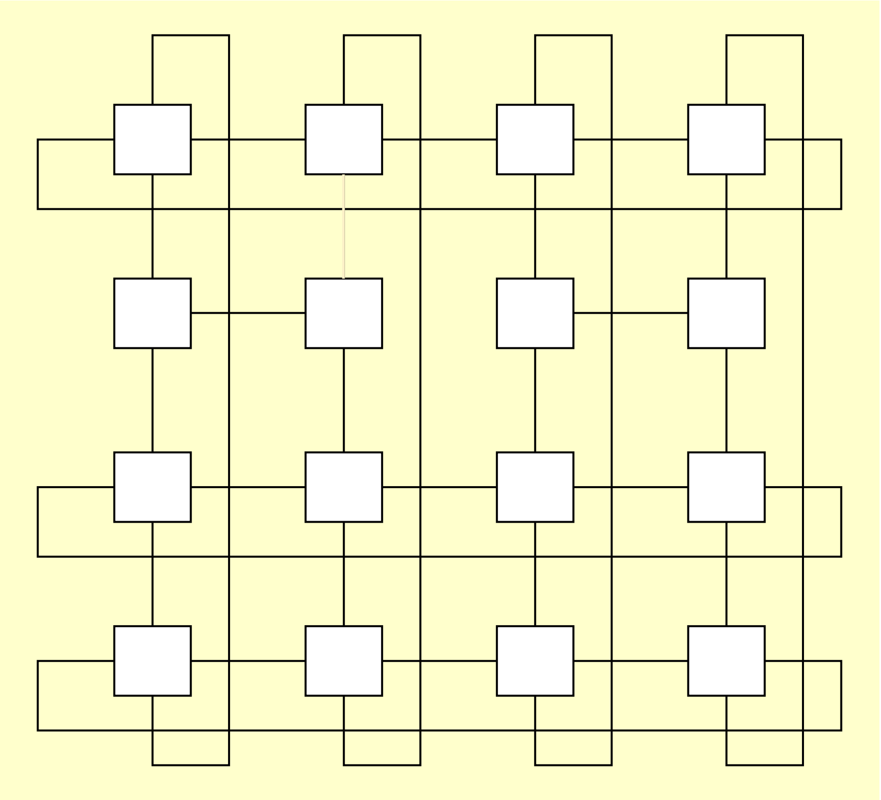
- Use of error-detecting/correcting codes (redundancy in time/space)
- Multiple transmissions via independent paths (redundancy in space)
- Retransmission in the same or different format (time redundancy)
- Message echo/ack in the same or different format (time redundancy)
- Special test messages (periodic diagnostics)



Link Outage Example

Three links go out in this torus





Malfunction-Tolerant Routing

1. Malfunctioning elements known globally (easy case; precompute path)
2. Only local malfunction info available (distributed routing decisions)

Distributed routing decisions are usually preferable, but they may lead to:

Suboptimal paths—Messages not going through shortest paths possible

Deadlocks—Messages interfering with or circularly waiting for each other

Livelocks—Wandering messages that never reach their destinations

Vast amount of literature on malfunction-tolerant (adaptive) routing:

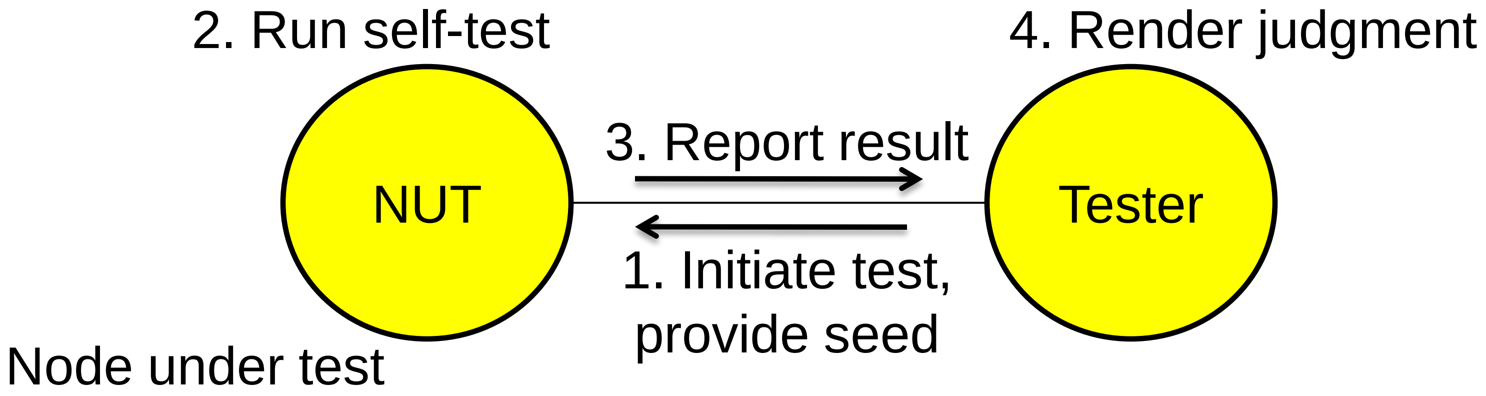
For nearly all popular interconnection networks

With many different assumptions about malfunctions and their effects

Node Malfunctions

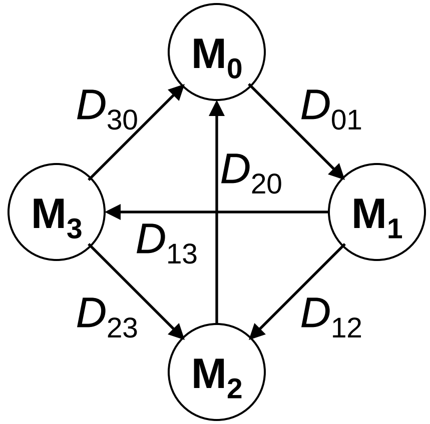
Node functional deviations or outage

- o Periodic self-test based on a diagnostic schedule
- o Self-checking design for on-line (concurrent) malfunction detection
- o Periodic testing by neighboring nodes
- o Periodic self-test with externally supplied seed



Malfunction Diagnosis

Consider this system, with the test outcomes shown



Diagnosis syndromes

Malfn	D_{01}	D_{12}	D_{13}	D_{20}	D_{30}	D_{32}
None	0	0	0	0	0	0
M_0	0/1	0	0	1	1	0
M_1	1	0/1	0/1	0	0	0
M_2	0	1	0	0/1	0	1
M_3	0	0	1	0	0/1	0/1
M_0,M_1	0/1	0/1	0/1	1	1	0
M_1,M_2	1	0/1	0/1	0/1	0	1

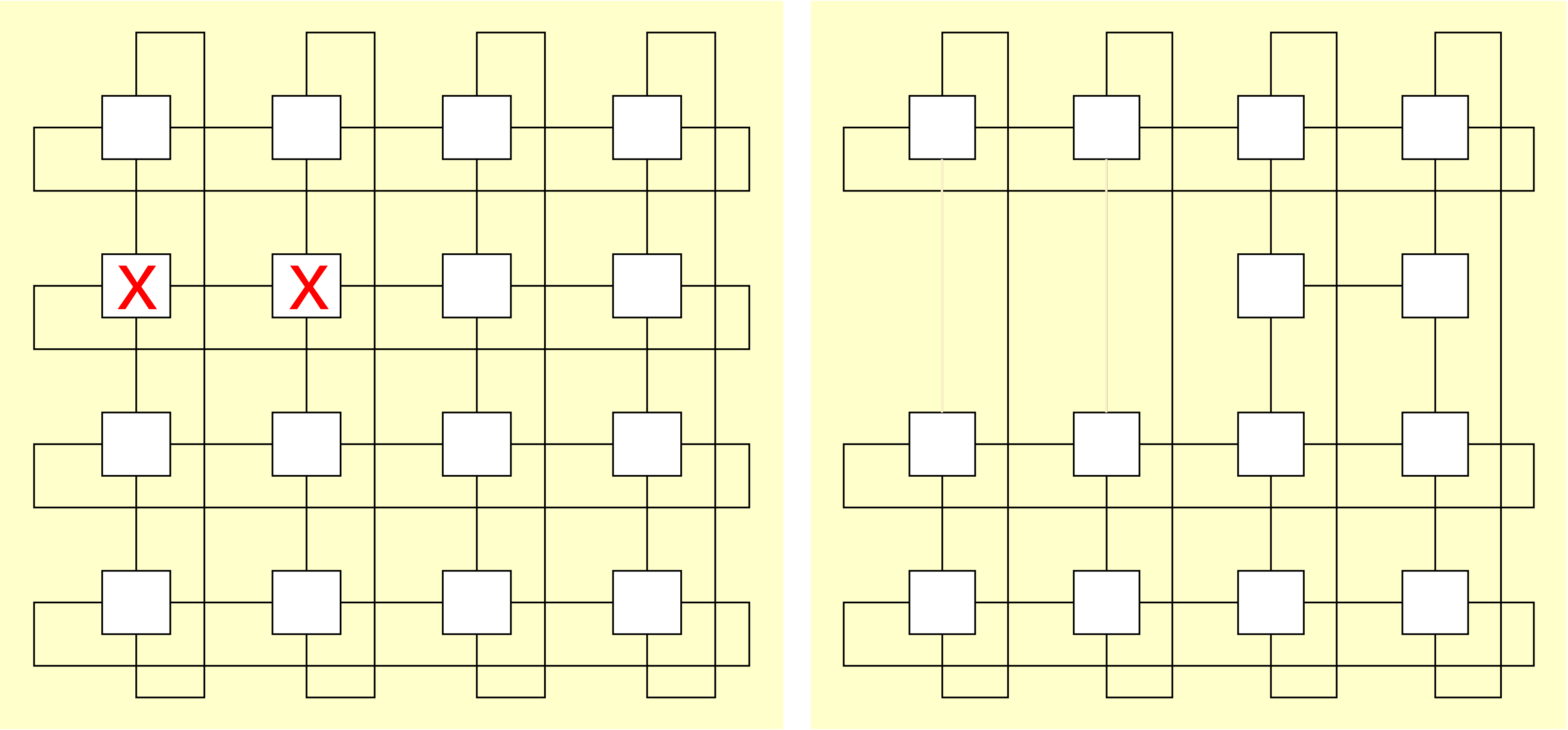
Syndrome dictionary

0	0	0	0	0	0	OK
0	0	0	1	1	0	M_0
0	0	1	0	0	0	M_3
0	0	1	0	0	1	M_3
0	0	1	0	1	0	M_3
0	0	1	0	1	1	M_3
0	1	0	0	0	1	M_2
0	1	0	1	0	1	M_2
1	0	0	0	0	0	M_1
1	0	0	1	1	0	M_0
1	0	1	0	0	0	M_1
1	1	0	0	0	0	M_1
1	1	1	0	0	0	M_1

Malfunction diagnosis is also called “system-level fault diagnosis”

Node Outage Example

Two nodes go out in this torus



Presentation Overview

Interconnection Networks

The Reliability Problem

Robustness Attributes

Network connectivity
Performance degradation

Deriving New Networks

Problems and Challenges

Dependable Parallel Processing

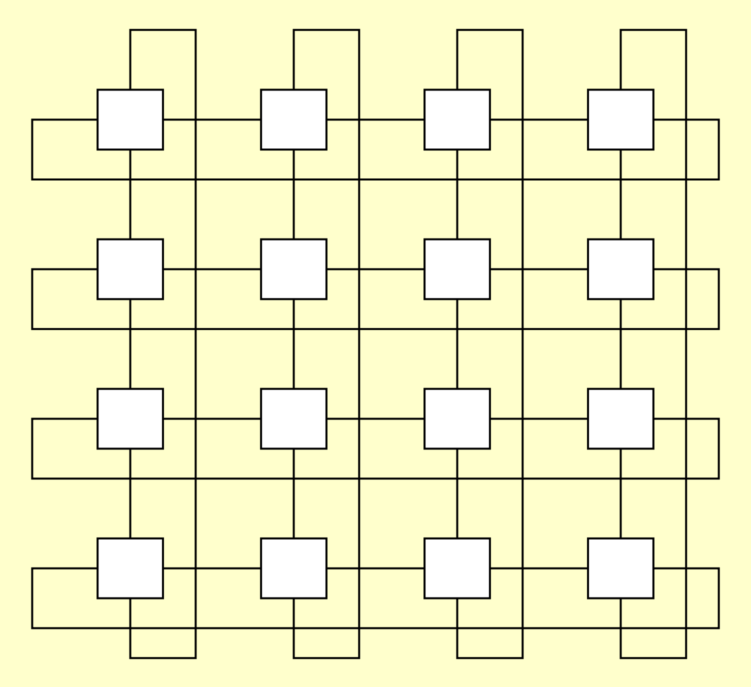
A parallel computer system consists of modular resources (processors, memory banks, disk storage, . . .), plus interconnects

Redundant resources can mitigate the effect of module malfunctions

An early approach: Provide shared spares (e.g., 1 for every 4 nodes)

The switching requirement of massive sparing is prohibitive

Furthermore, interconnects cannot be Dealt with in the same way



The modern approach to dependable parallel processing:

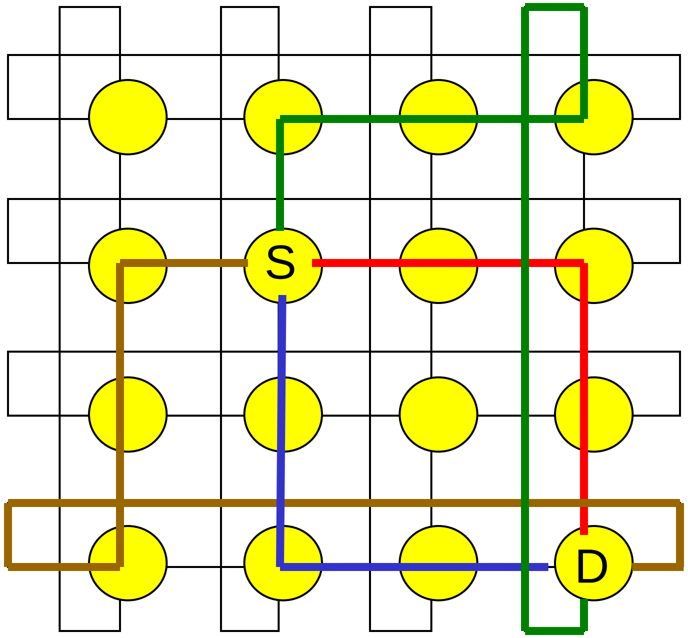
Provide more-than-bare-minimum nodes and interconnects, but do not label them as ordinary and spare

Multiple Disjoint Paths

Connectivity $\kappa \leq d_{\min}$ (min node degree)
If equality holds, the network is optimally/maximally malfunction-tolerant
(I will use k instead of the standard κ)

Network connectivity being k means there are k "parallel" or "node/edge-disjoint" paths between any pair of nodes

Parallel paths lead to robustness, as well as greater performance



1. Symmetric networks tend to be maximally malfunction-tolerant
2. Finding the connectivity of a network not always an easy task
3. Many papers in the literature on connectivity of various networks

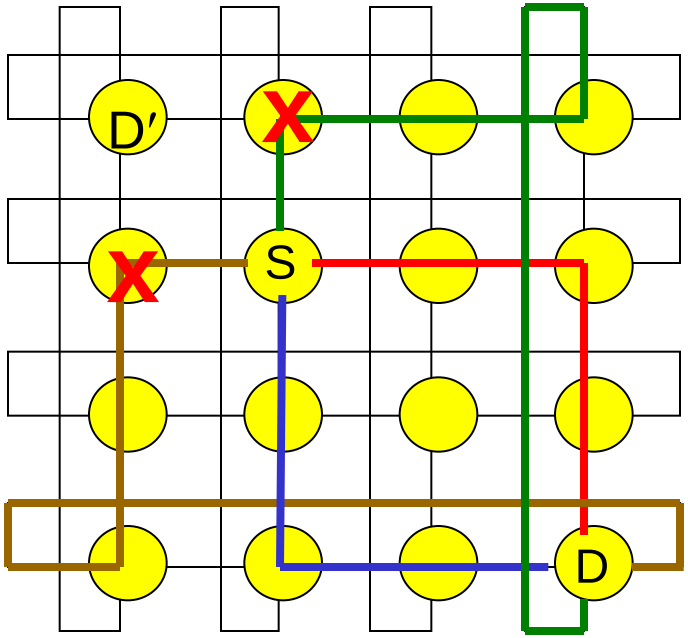
Dilated Internode Distances

When links and/or nodes malfunction:
Some internode distances increase;
Network diameter may also increase

Consider routing from S to D′

Two node malfunctions can disrupt both available shortest paths

Path length increases to 4
(via wraparound links to D′)

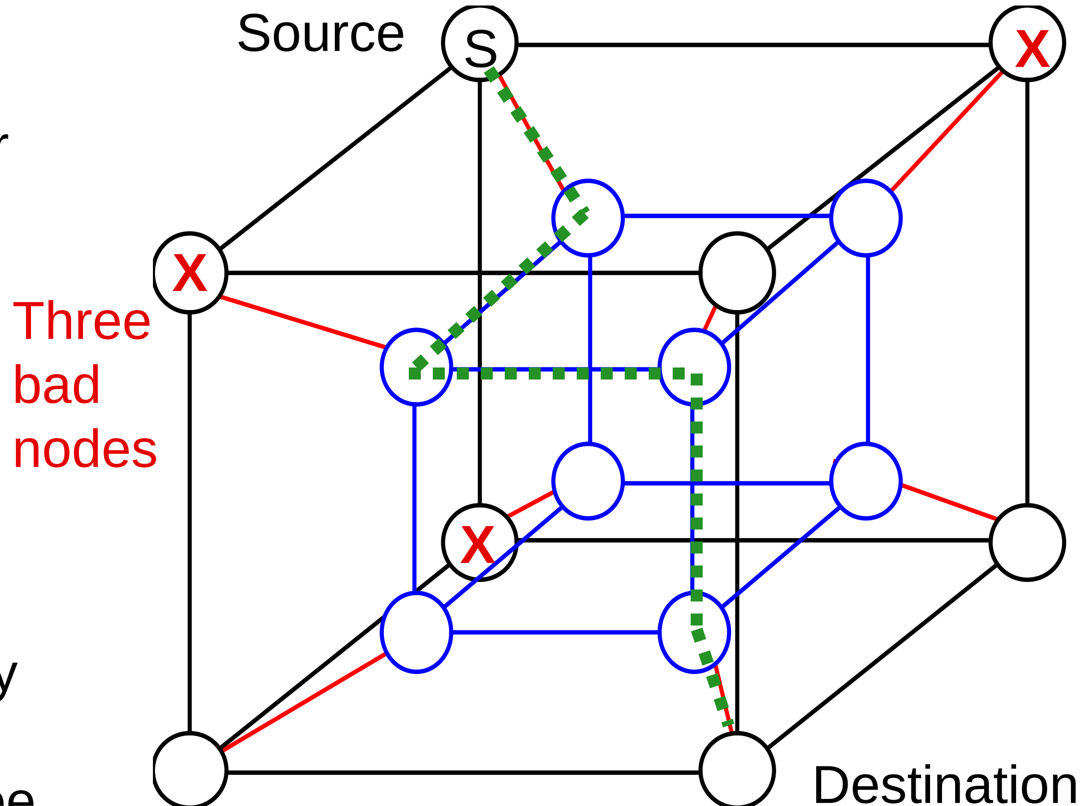


Malfunction diameter: Worst case diameter for *k* – 1 malfunctions

Wide diameter: Maximum, over all node pairs, of the longest path in the best set of *k* parallel paths (quite difficult to compute)

Malfunction Diameter

Rich connectivity provides many alternate paths for message routing



The node that is furthest from S is not its diametrically opposite node in the malfunction-free hypercube

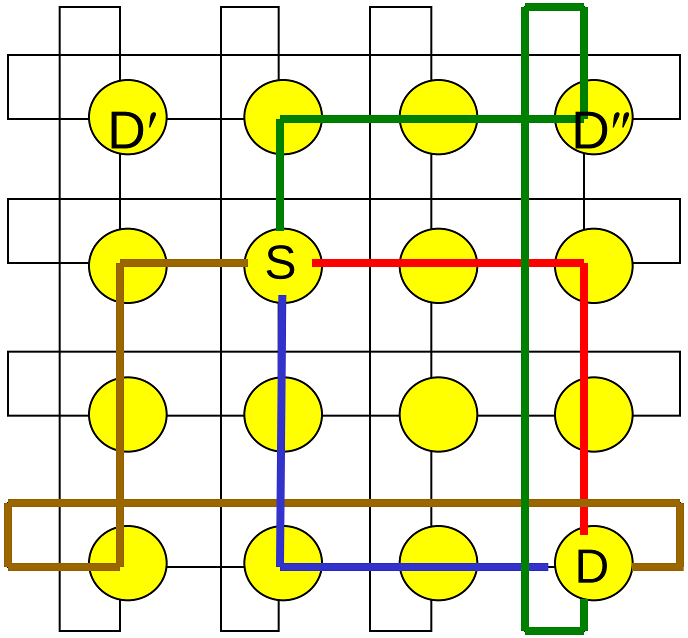
Malfunction diameter of the *q*-cube is *q* + 1

Wide Diameter

Consider parallel paths between S and D
All four paths are of length 4
So, the wide distance is 4 in this case

Now consider parallel paths from S to D′
Two are of length 2
Two are of length 4
So, the wide distance is also 4 here

Thus $D_W \geq 4$ for this network



To determine D_W, we must identify a worst-case pair of nodes

S and D″ constitute such a worst-case pair ($D_W = 5$)

Deriving D_W is an even more challenging task than determining D_M

Presentation Overview

Interconnection Networks

The Reliability Problem

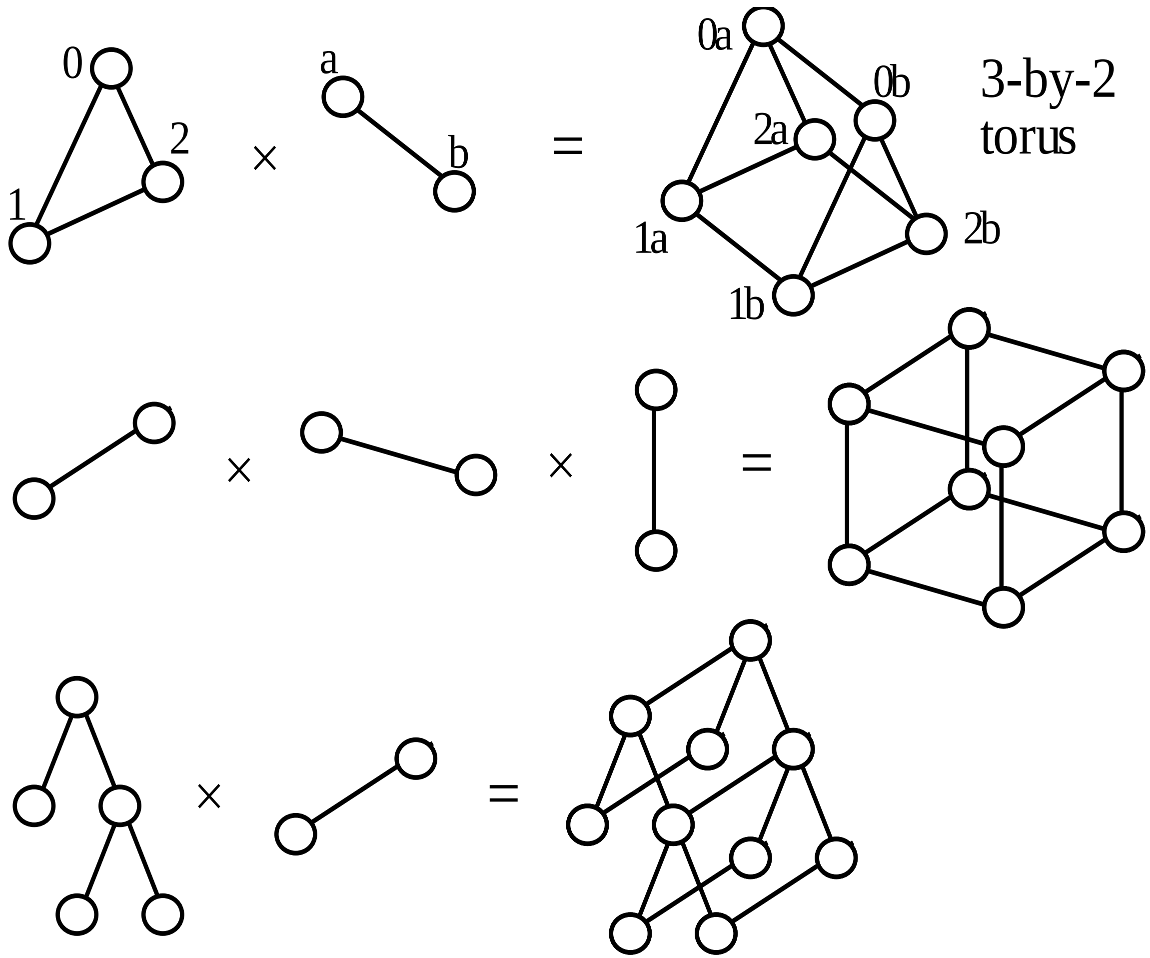
Robustness Attributes

Deriving New Networks

Cartesian product
Swapped/OTIS structure
Pruning of networks

Problems and Challenges

Cartesian Product Networks



Properties of product graph $G = G_1 \times G_2$:

Nodes labeled (x_1, x_2), $x_1 \in V_1$, $x_2 \in V_2$

Two nodes in G are connected if either component of the two nodes were connected in component graphs

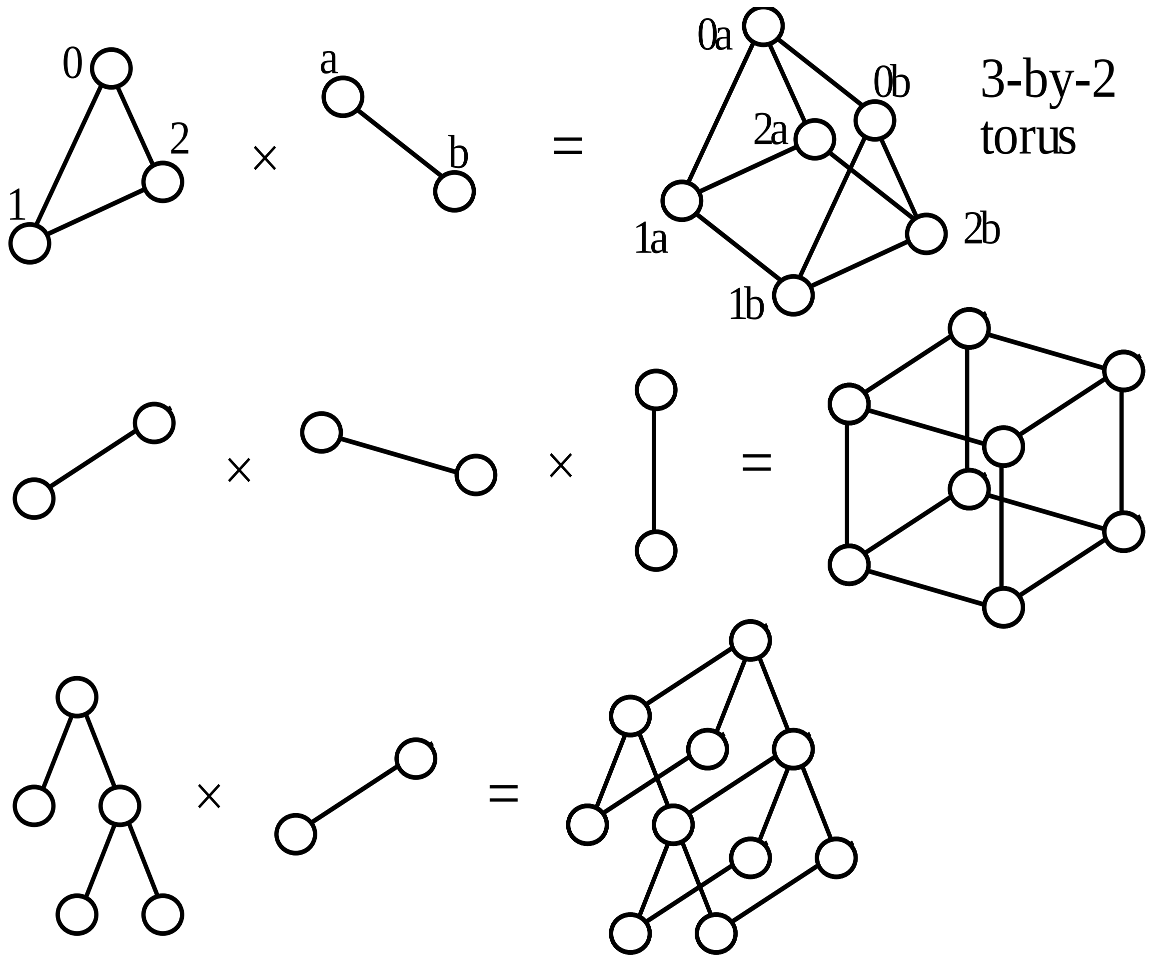
$p = p_1 p_2$

$d = d_1 + d_2$

$D = D_1 + D_2$

$\Delta = \Delta_1 + \Delta_2$

Product Network Robustness



Robustness attributes of $G = G_1 \times G_2$:

Connectivity

$k \geq k_1 + k_2$

Scalable in connectivity for logarithmic or sublogarithmic k_1 and k_2

Malfunction diameter

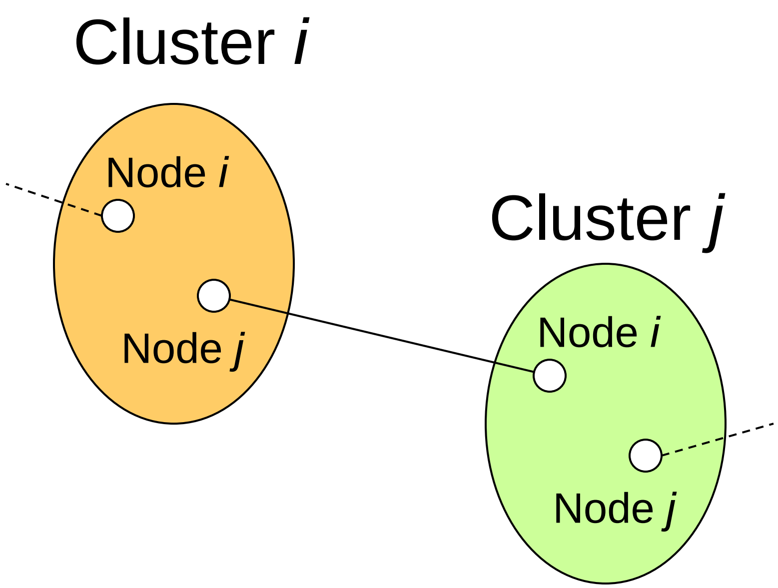
No general result

Wide diameter

No general result

Swapped (OTIS) Networks

Swapped network
OTIS (optical transpose interconnect system) network
Built of m clusters, each being an m-node "basis network"
Intercluster connectivity rule: node j in cluster i linked to node i in cluster j

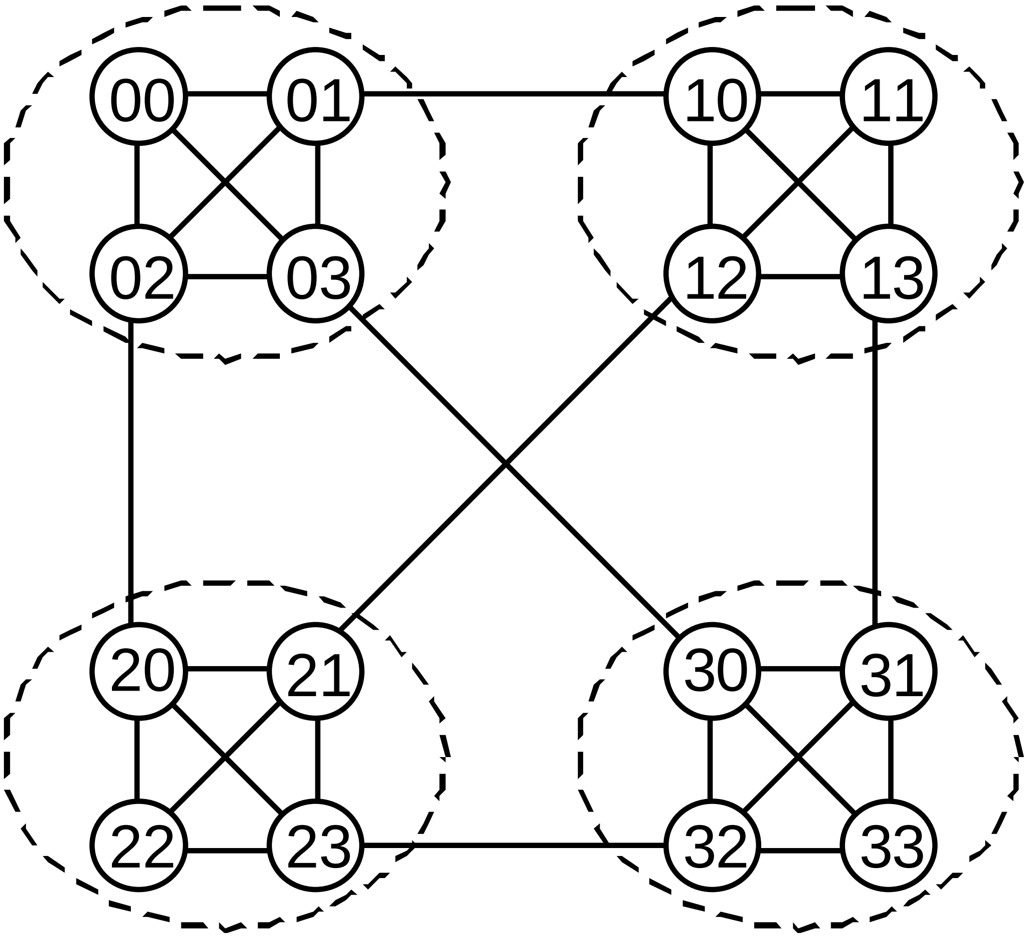


Two-level structure
- Level 1: Cluster (basis network)
- Level 2: Complete graph

Number of nodes: $p = m^2$
Diameter: $D = 2D_{\text{basis}} + 1$

Nucleus K_m: WK Recursive
Nucleus $Q_{\log m}$: HCN

Swapped Network Robustness



Robustness of Sw(*G*):

Connectivity

d(*G*), regardless of *k*(*G*)

Sw(*G*) provides good connectivity even when the basis network is not well-connected

Malfunction diameter

At most *D*(Sw(*G*)) + 4

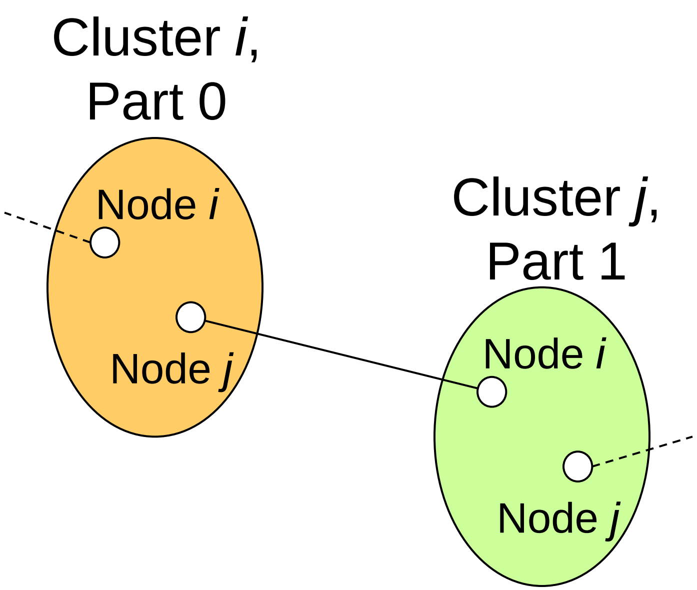
Wide diameter

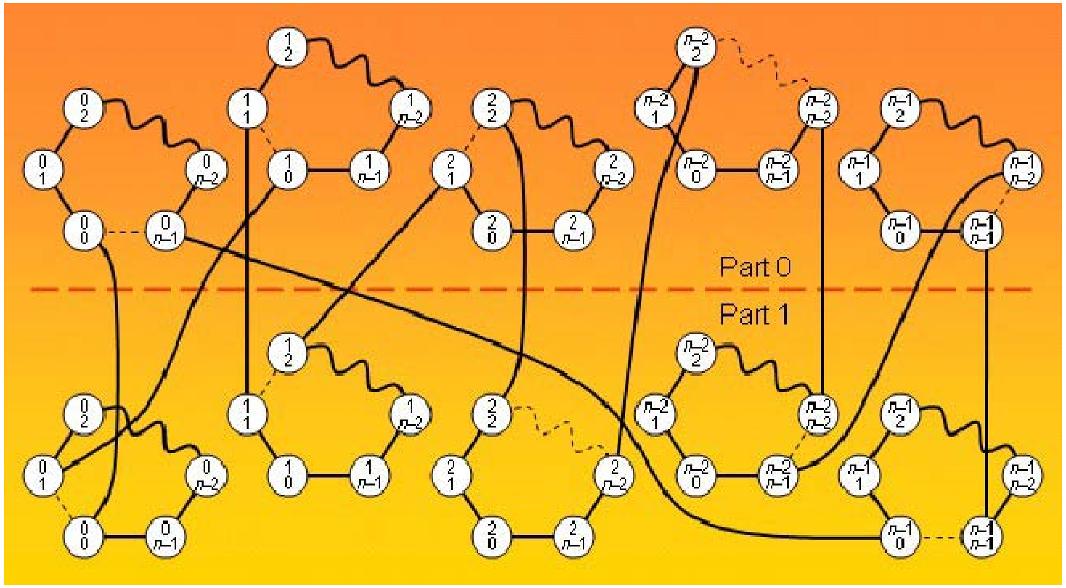
At most *D*(Sw(*G*)) + 4

Biswapped Networks

Similar to swapped/OTIS but with twice as many nodes, in two parts
Nodes in part 0 are connected to nodes in part 1, and vice versa

Biswapped networks with connected basis networks are maximally malfunction-tolerant (connectivity = node degree)

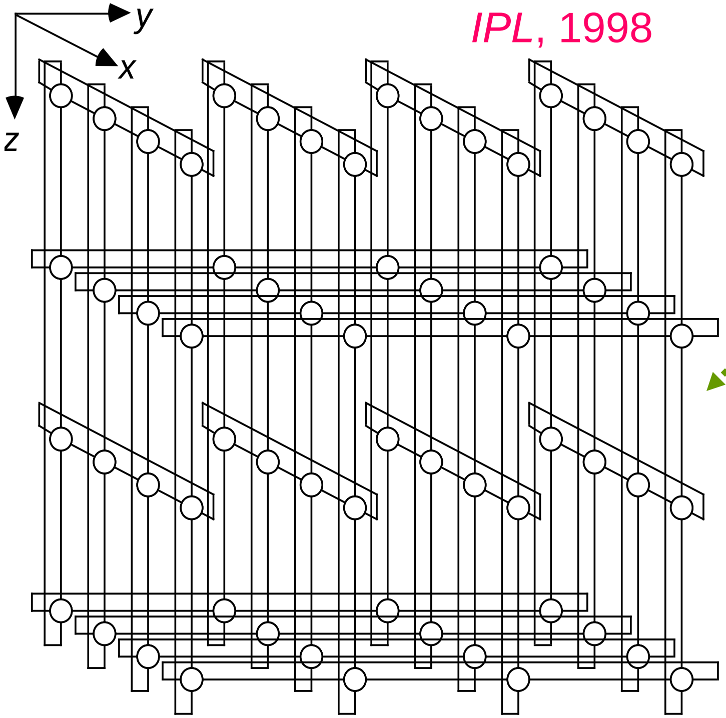




Systematic Pruning

3D torus pruned along *Z*

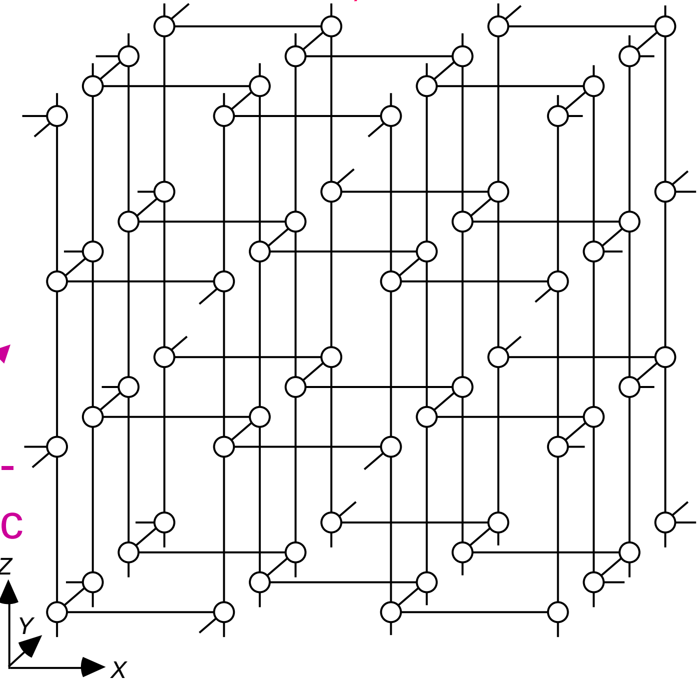
IPL, 1998



Cayley Graph and edge-symmetric

Diamond net = pruned torus

IEEE TPDS, Jan. 2001



Not edge-symmetric

Must have simple and elegant pruning rules to ensure:

- Efficient point-to-point and collective communication
- Symmetry, leading to “blandness” and balanced traffic

Pruned Network Robustness

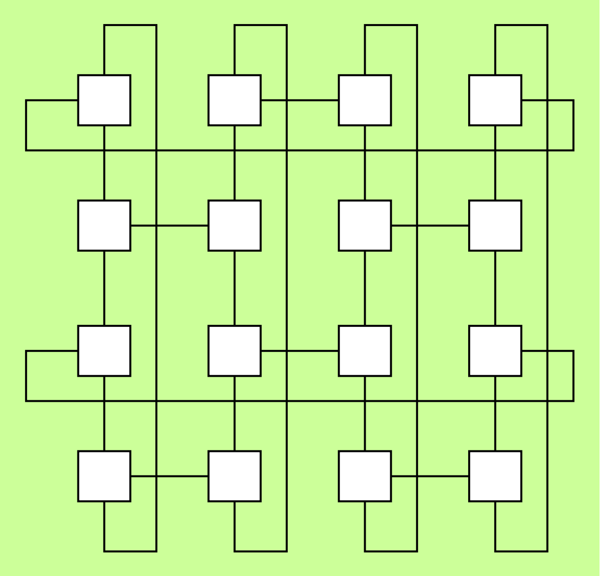
Robustness is in general adversely affected when a network is pruned

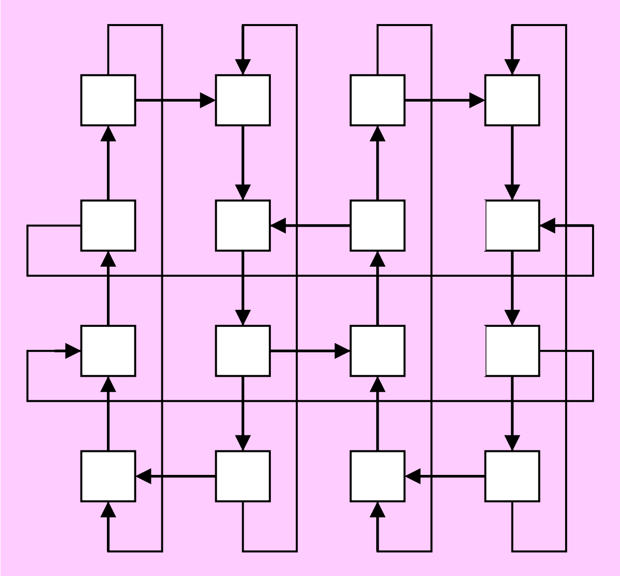
Systematic pruning can ensure maximal robustness in the resulting network

General strategy:

Begin with a richly connected network that is a Cayley graph

Prune links in such a way that the network remains a Cayley graph





We have devised pruning schemes for a wide variety of networks and proven resulting networks to be robust & efficient algorithmically

Presentation Overview

Interconnection Networks

The Reliability Problem

Robustness Attributes

Deriving New Networks

Problems and Challenges

Where do we go from here?

On the Empirical Front

Which hybrid (multilevel, hierarchical) network construction methods yield robust structures?

Given different robustness attributes, is there a good way to quantify robustness for comparison purposes?

What would be a good measure for judging cost-effective robustness?

Of existing “pure” networks, which ones are best in terms of the measure above

Are there special considerations for robustness in NoCs?

On the Theoretical Front

The (d, D) graph problem: Given nodes of degree d, what is the maximum number of nodes that we can incorporate into a network if diameter is not to exceed D? **aka (d, k) problem**

The (d, D) graph problem is very difficult
Answers are known only for certain values of d and D

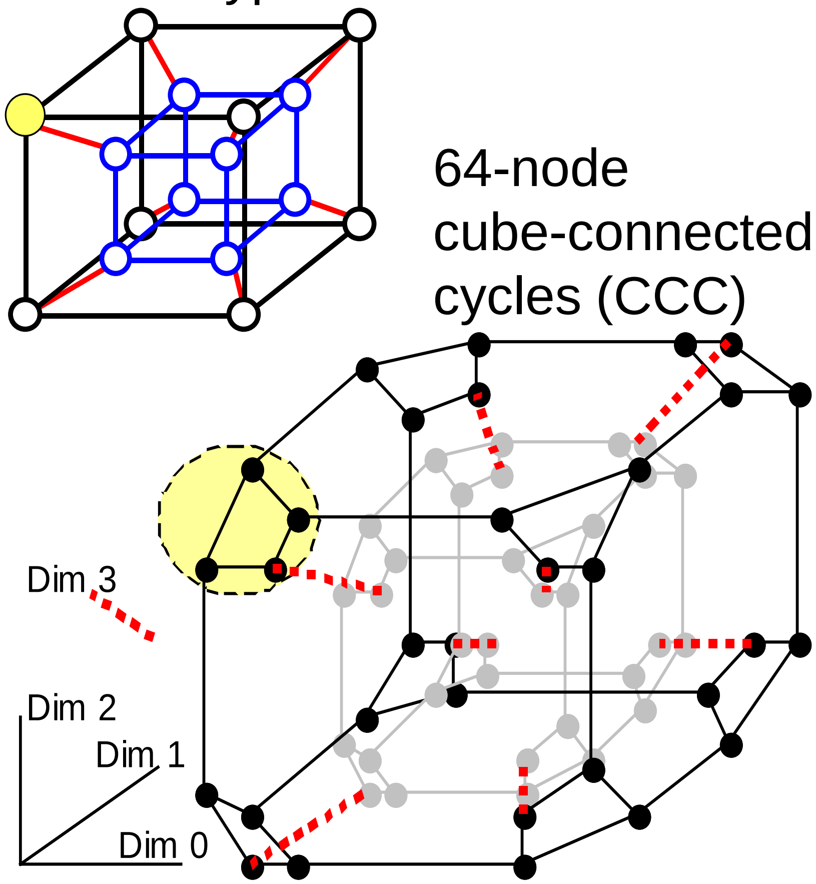
Malfunction diameter: **aka fault diameter**
Can we solve, at least in part, the (d, D_M) graph problem?
How much harder is this problem compared with (d, D)?

Wide diameter:
Can we solve, at least in part, the (d, D_W) graph problem?
How much harder is this problem compared with (d, D)?

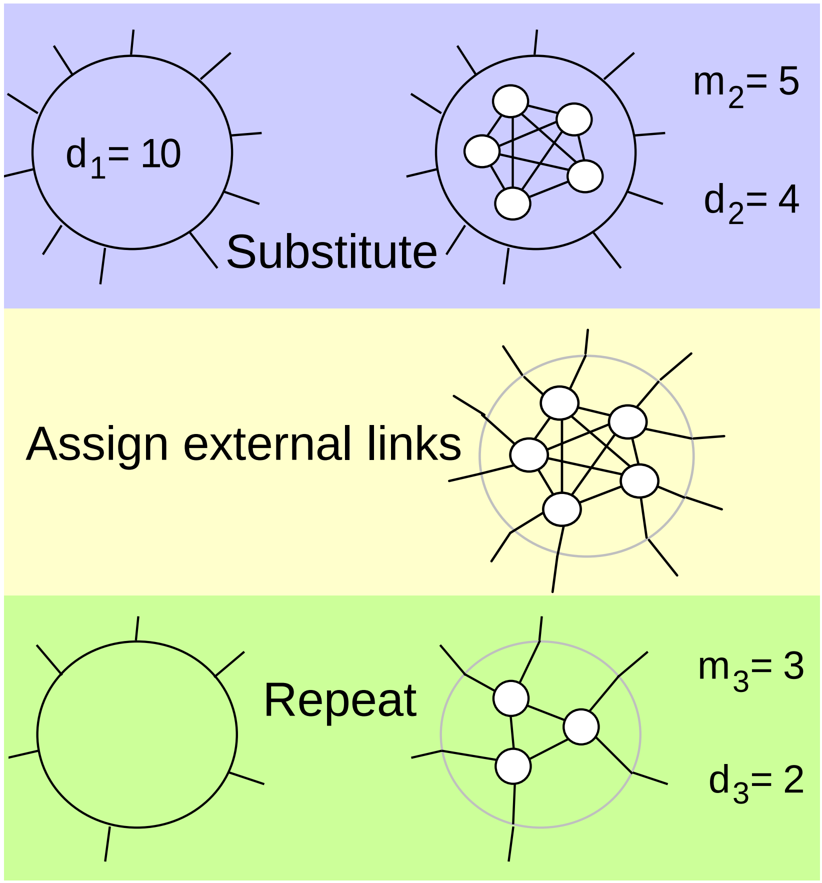
Recursive Substitution

16-node hypercube

64-node cube-connected cycles (CCC)



The general approach



Questions or Comments?

parhami@ece.ucsb.edu
http://www.ece.ucsb.edu/~parhami/

Robustness Attributes of Interconnection Networks for Parallel Processing

Additional Slides

Importance of Diameter

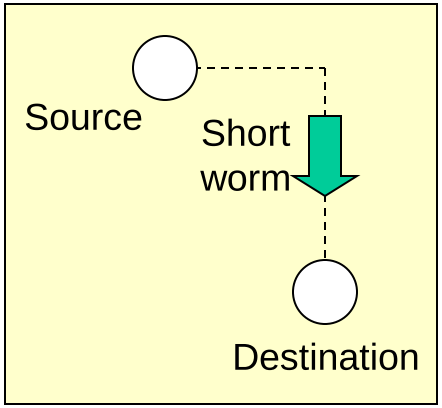
Average internode distance Δ is an indicator of performance

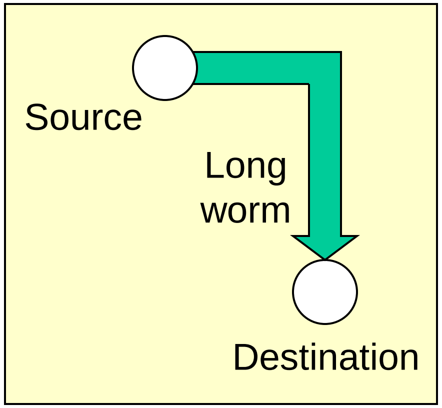
Δ is closely related to the diameter D

For symmetric nets: $D/2 \le \Delta \le D$

Short worms: hop distance clearly dictates the message latency

Long worms: latency is insensitive to hop distance, but tied up links and waste due to dropped or deadlocked messages rise with hop distance





Diagnosis Challenges

Analysis problems:

1. Given a directed graph defining the test links, find the largest value of t for which the system is 1-step t-diagnosable (easy if no two units test one another; fairly difficult, otherwise)
2. Given a directed graph and its associated test outcomes, identify all the malfunctioning units, assuming there are no more than t

Vast amount of published work dealing with Problems 1 and 2

Synthesis problem:

3. Specify test links (connection assignment) that makes an n-unit system 1-step t-diagnosable; use as few test links as possible

A degree-t directed chordal ring, in which node i tests the t nodes $i + 1, i + 2, \ldots, i + t$ (all mod n) has the required property

There are other problem variants, such as sequential diagnosability

Mesh Adaptive Routing

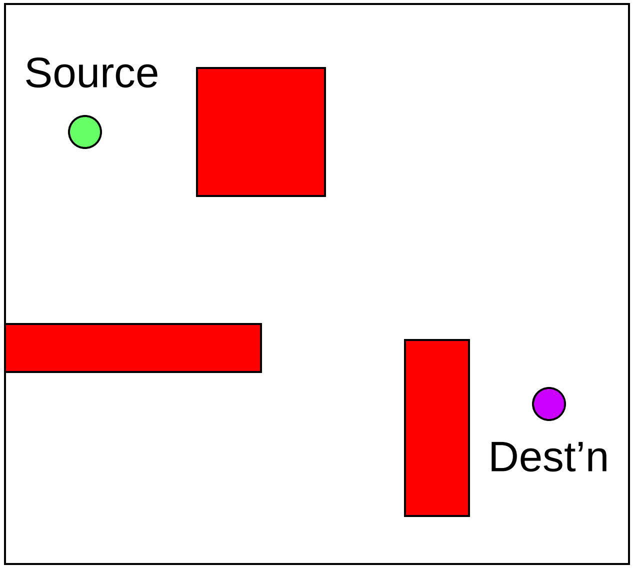
With no malfunction, row-first or column-first routing is simple & efficient

Hundreds of papers on adaptive routing in mesh (and torus) networks

The approaches differ in:
Assumptions about malfunction types and clustering
Type of routing scheme (point-to-point or wormhole)
Optimality of routing (shortest path)
Details of routing algorithm
Global/local/hybrid info on malfunctions

Of the proposed routing strategies:
Some are specific to meshlike networks
Others can be extended to many networks

Meshes/tori are surprisingly robust if you don't mind losing a few of the good nodes



Product Network Scalability

A. Logarithmic-diameter networks

$D = \log p_1 + \log p_2 = \log(p_1 p_2)$ → Perfect diameter scaling in this case

But diameter scaling achieved at the cost of much more complex nodes

B. Sublogarithmic-diameter networks

$D = \log\log p_1 + \log\log p_2 = \log(\log p_1 \log p_2) = \log\log(p_1^{\log p_2})$

$= \log\log(p_1 p_2 (p_1^{\log p_2 - 1}/p_2))$

In the special case of $p_1 = p_2 = p$, the parenthesized factor multiplied by $p_1 p_2$ will be greater than 1 for $p > 4$ → Poor diameter scaling

C. Superlogarithmic-diameter networks

Similar analysis shows good diameter scaling

Unfortunately, B is the most important case for massive parallelism

Swapped Network Scalability

A. Logarithmic-diameter basis network

$D = 2 \log m + 1 = \log(2m^2)$ → Near-perfect diameter scaling in this case

Good diameter scaling achieved at minimal added cost (d → $d + 1$)

B. Sublogarithmic-diameter networks

$D = 2 \log \log m + 1 = \log(2 \log^2 m) = \log \log(m^2\, m^{2(\log m - 1)})$

The factor multiplied by m^2 in the final result is always greater than 1, leading to poor diameter scaling

$D = 2 (\log m)^{1/2} + 1 = 1.414(\log m^2)^{1/2} + 1$

C. Superlogarithmic-diameter networks

Similar analysis shows good diameter scaling

Unfortunately, B is the most important case for massive parallelism

Analogy for Adaptive Routing

This slide was added after the talk: *During our informal discussions, an ISUM2010 participant used the word "fire," thinking that it meant "failure," thus inadvertently creating the following interesting analogy.*

A graph that models an interconnection network can be interpreted as the floorplan of a building, with nodes representing rooms, and links standing for hallways that interconnect rooms.

Suppose there are fires raging in the building and you want to go from your current room S to an exit located in room D. Let's say you know the exact floorplan of the building (the analog of the network topology).

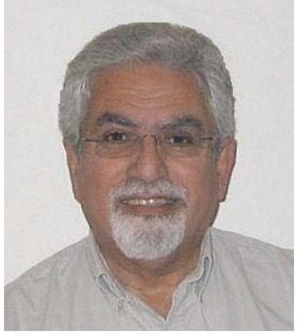
If you have complete knowledge of where the fires are located, you can easily plan an escape route, assuming one exists (precompute your path).

If you know nothing about fire locations, you try to move in the direction of the exit, taking detours whenever you hit an unpassable hallway or room.

Abstract and Speaker’s Bio

Abstract: Large-scale parallel processors, with many thousands or perhaps even millions of nodes and links, are prone to malfunctions in their constituent parts. Thus, even under a best-case scenario of prompt malfunction detection to prevent data contamination, such systems tend to lose processing and communication resources over time. Whether they can survive such inevitable losses is a function of the way computational tasks and their attendant information exchanges are organized and on certain intrinsic properties of the interconnection topology.

This talk begins with an overview of robustness features, as they pertain to interconnection architectures. Next, a number of well-known interconnection structures are viewed from the robustness angle. Finally, it is shown how large-scale hierarchical or multilevel networks can be synthesized for robustness, while keeping implementation cost, power dissipation, and routing overhead in check.



Very brief bio: Behrooz Parhami (PhD, University of California, Los Angeles, 1973) is Professor of Electrical and Computer Engineering, and Associate Dean for Academic Affairs, College of Engineering, at University of California, Santa Barbara, where he teaches and does research in computer arithmetic, parallel processing, and dependable computing. A Fellow of IEEE and British Computer Society and recipient of several other awards, he has written six textbooks and more than 260 peer-reviewed technical papers. Professionally, he serves on journal editorial boards and conference program committees and is also active in technical consulting.