Interconnection Networks for Parallel Processors and Data Centers

Behrooz Parhami
University of California, Santa Barbara, USA

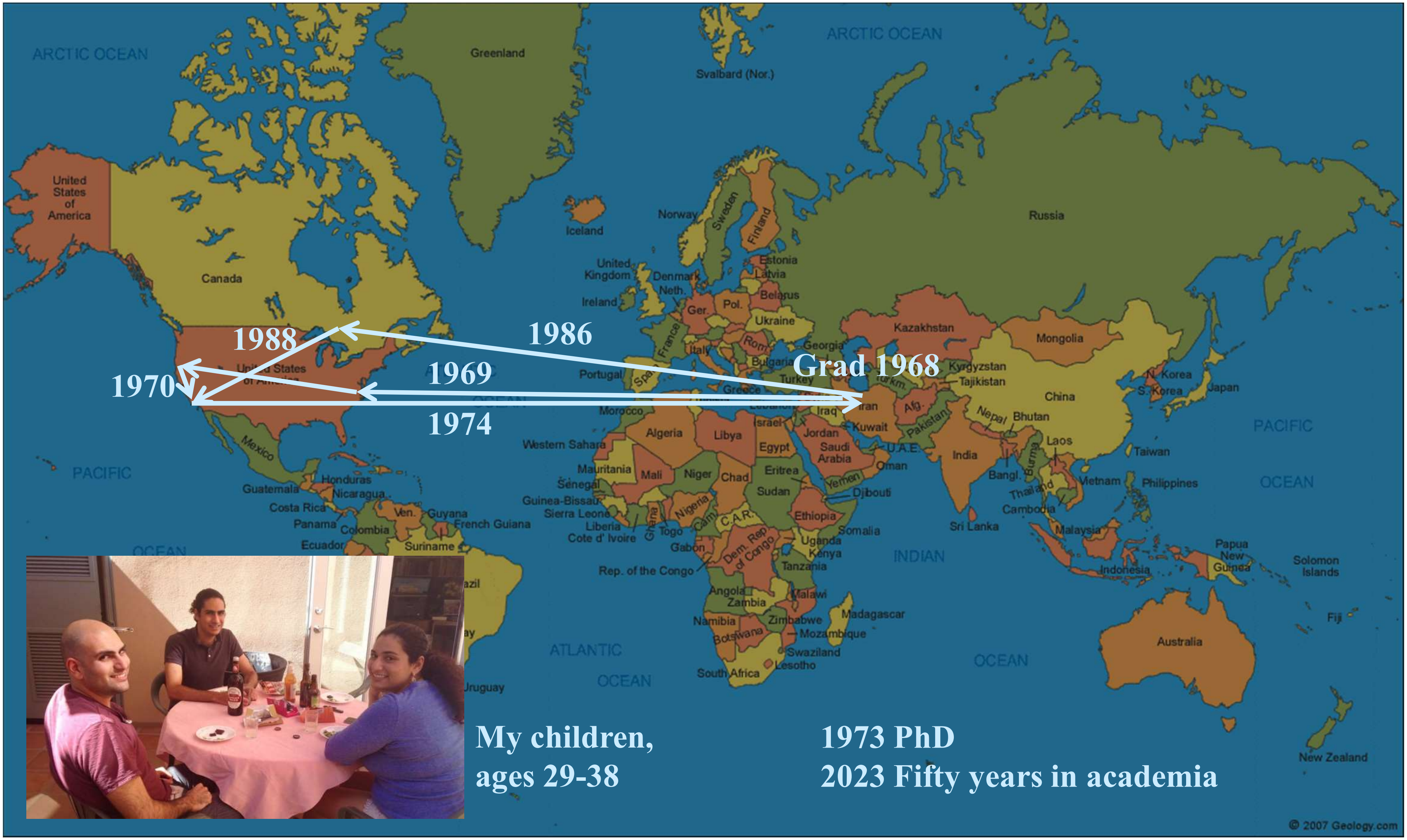
About This Presentation

This slide show was developed as part of B. Parhami's suite of seven lectures for IEEE Computer Society's Distinguished Visitors Program, 2021-2023 (3-year term).

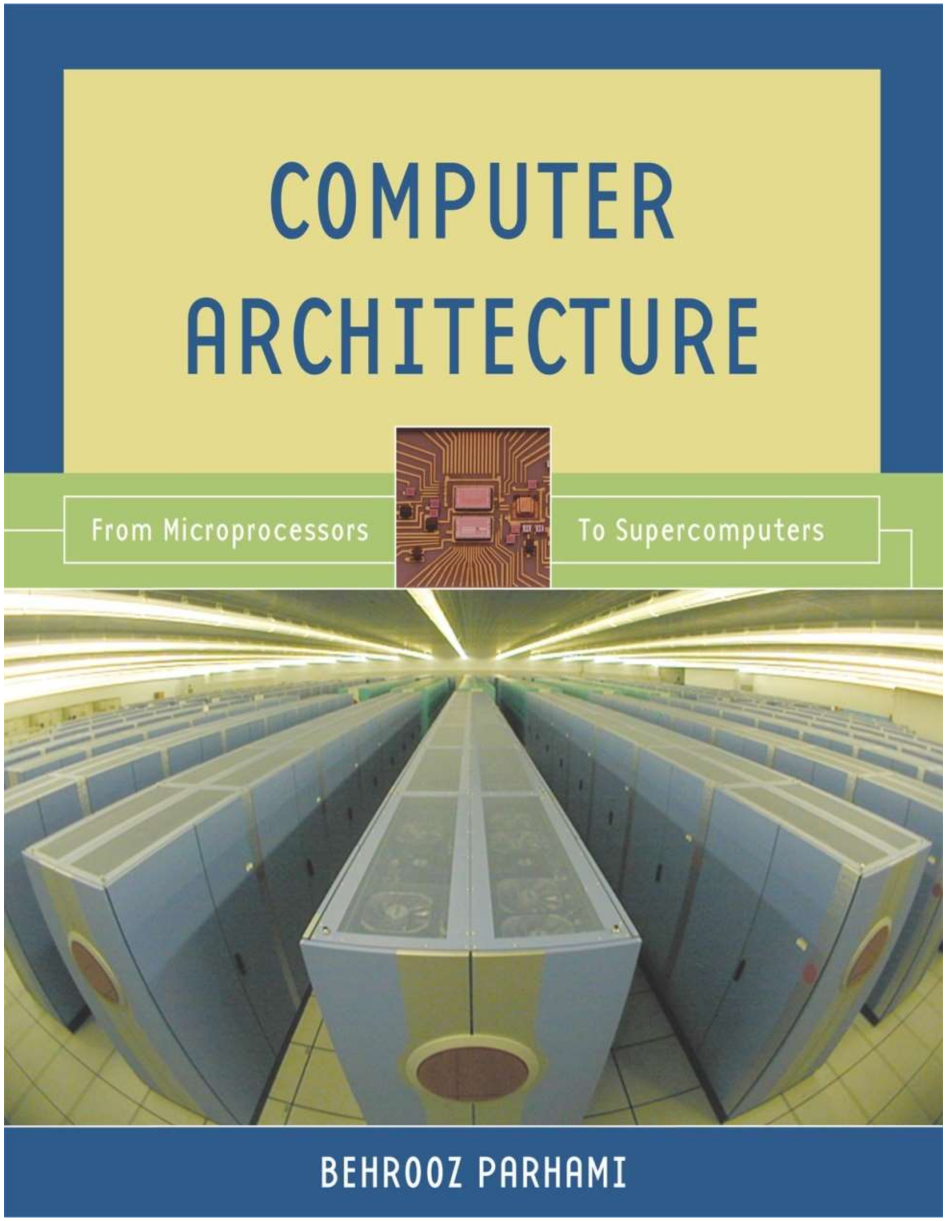
Edition	Released	Revised	Revised	Revised
First	Jan. 2022	Sep. 2022	Nov. 2023	

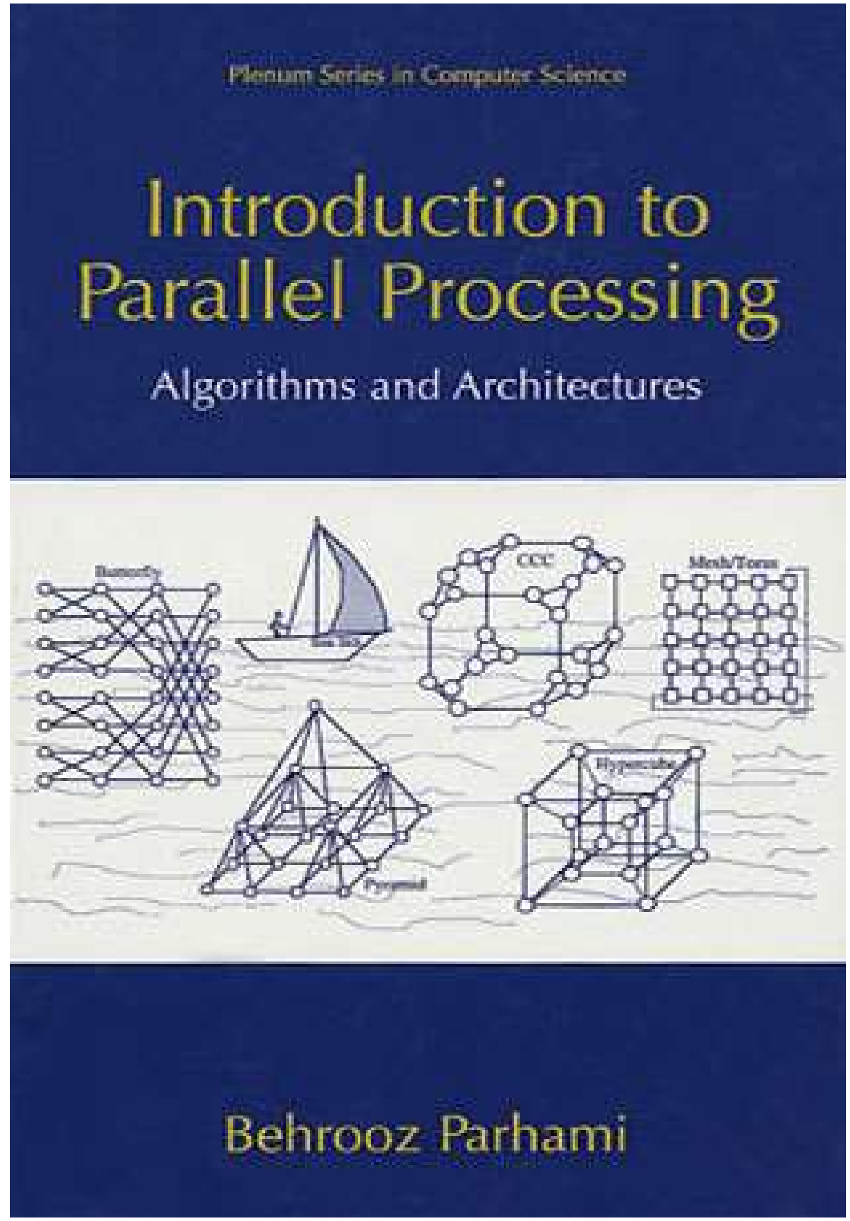
File: http://www.ece.ucsb.edu/~parhami/pres_folder/parh23-interconnection-networks-231117-dvp.pdf

My Personal Academic Journey



Some of the material in this talk come from, or will appear in updated versions of, my two computer architecture textbooks





Interconnection Networks for Parallel Processing and Data Centers

Interconnecting multiple processors in a parallel supercomputer or servers in a data center constitutes a challenging problem. There are so many ways to interconnect the computing nodes that the range of options has come to be known as "the sea of interconnection networks." In this talk, I will outline the theoretical underpinnings of interconnection network design in a way that exposes the challenges. I will then review desirable network properties and relate them to various network classes that have been used or proposed. Emphasis will be placed on robustness attributes of networks, given that large networks with many thousands or perhaps even millions of nodes are bound to experience malfunctions in nodes and links.

Presentation Topics

Theoretical Foundations

Graph theory; the (*d*, *D*) problem

Types of Networks

Direct; Indirect; Mesh/Torus; Fat-tree

Performance Attributes

Latency; Bandwidth; Power/Energy

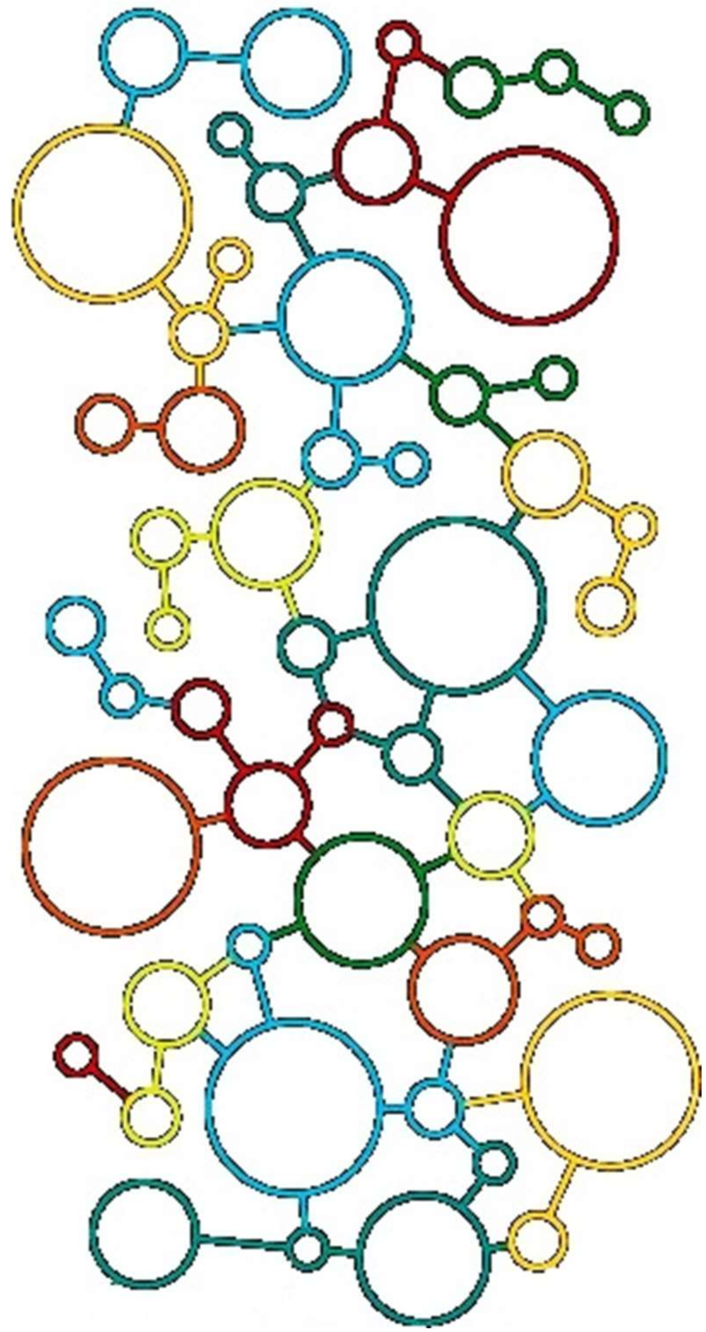
Reliability and Robustness

Survival; Tolerance to node/link loss

Design Contributions

Swapped; Chordal rings; Pruning

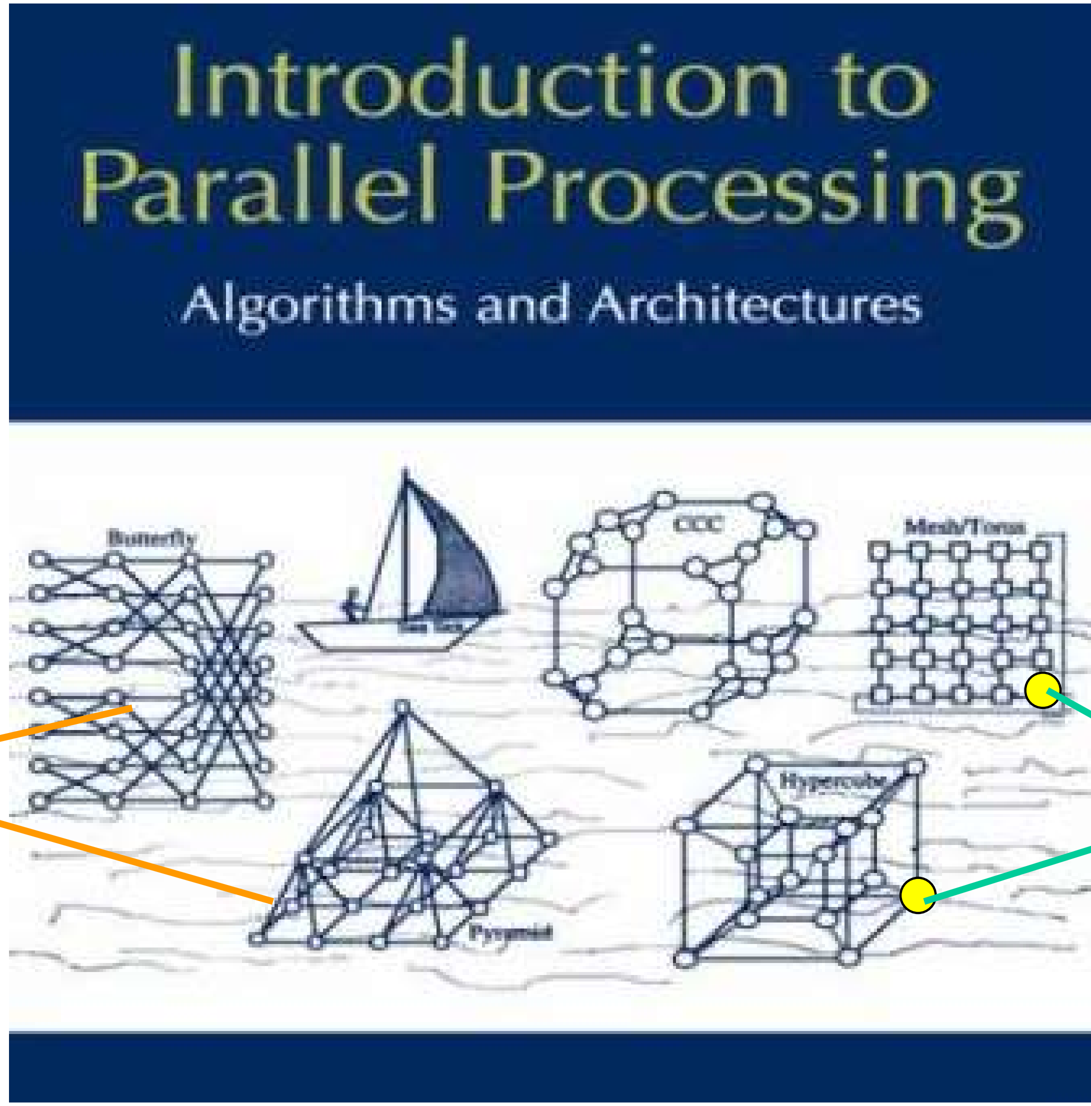
Summary and Future Work



Parallel Computer = Nodes + Interconnects
Data Center = Servers + Interconnects

Nodes + ICs = The Internet (not discussed)

Cores + ICs = Chip-multiproc. (NoC)



B. Parhami, Plenum Press, 1999

Built into nodes are switches or routers

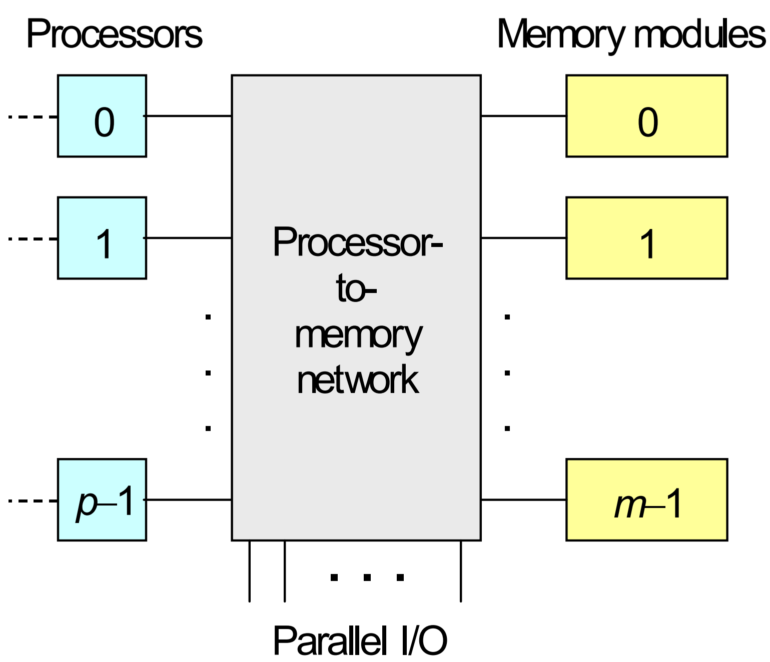
Interconnects, communication channels, or links

Nodes, processors, or servers

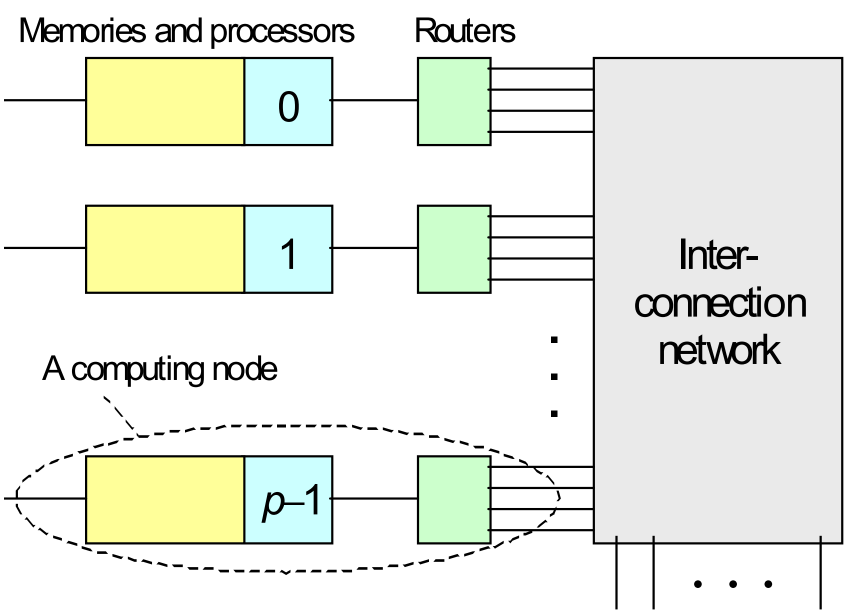
Shared-Memory Parallel Computing

Control parallelism: executing several instruction streams in parallel

GMSV: Shared global memory – symmetric multiprocessors
DMSV: Shared distributed memory – asymmetric multiprocessors
DMMP: Message passing – multicomputers

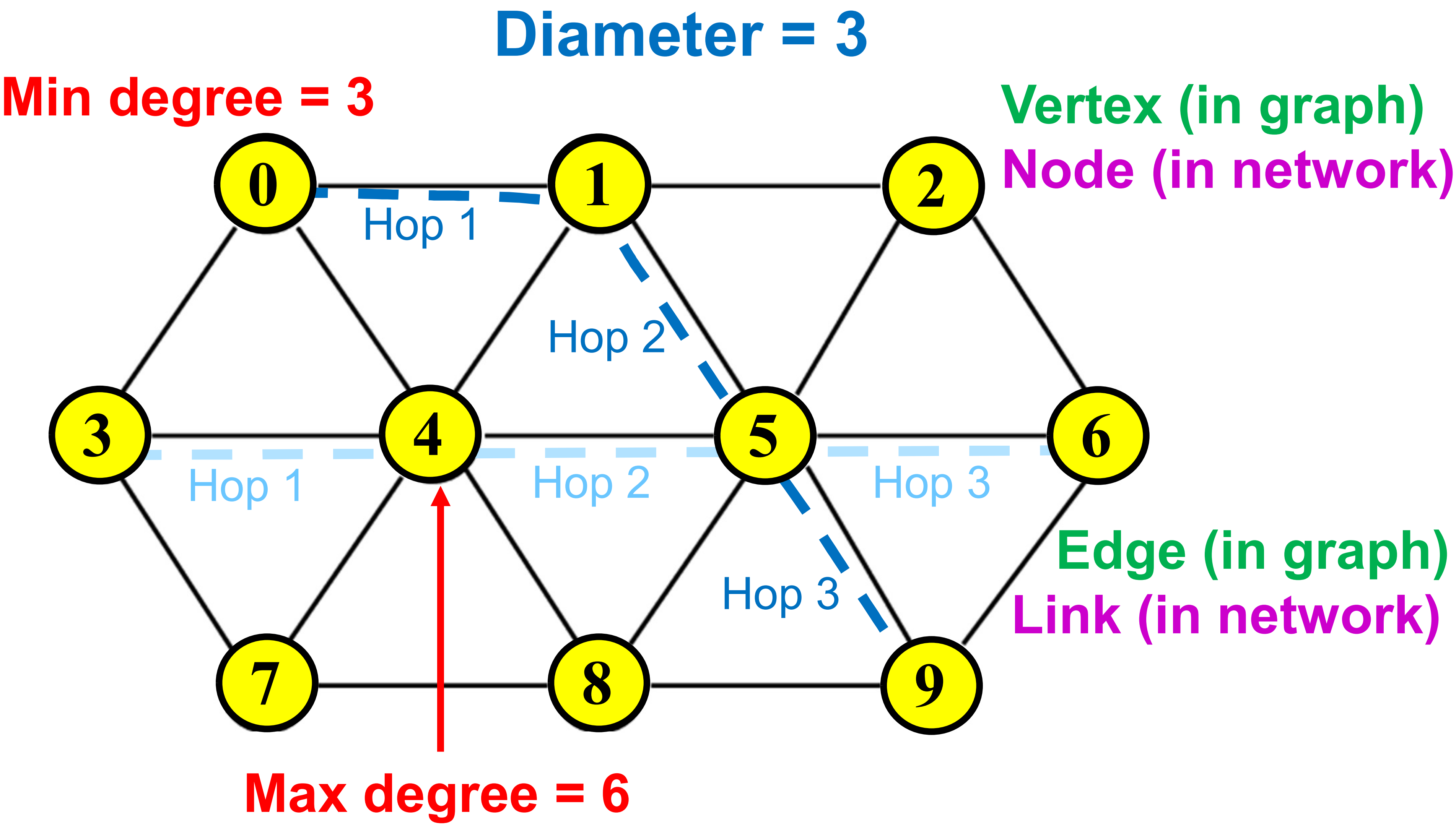


Global shared memory

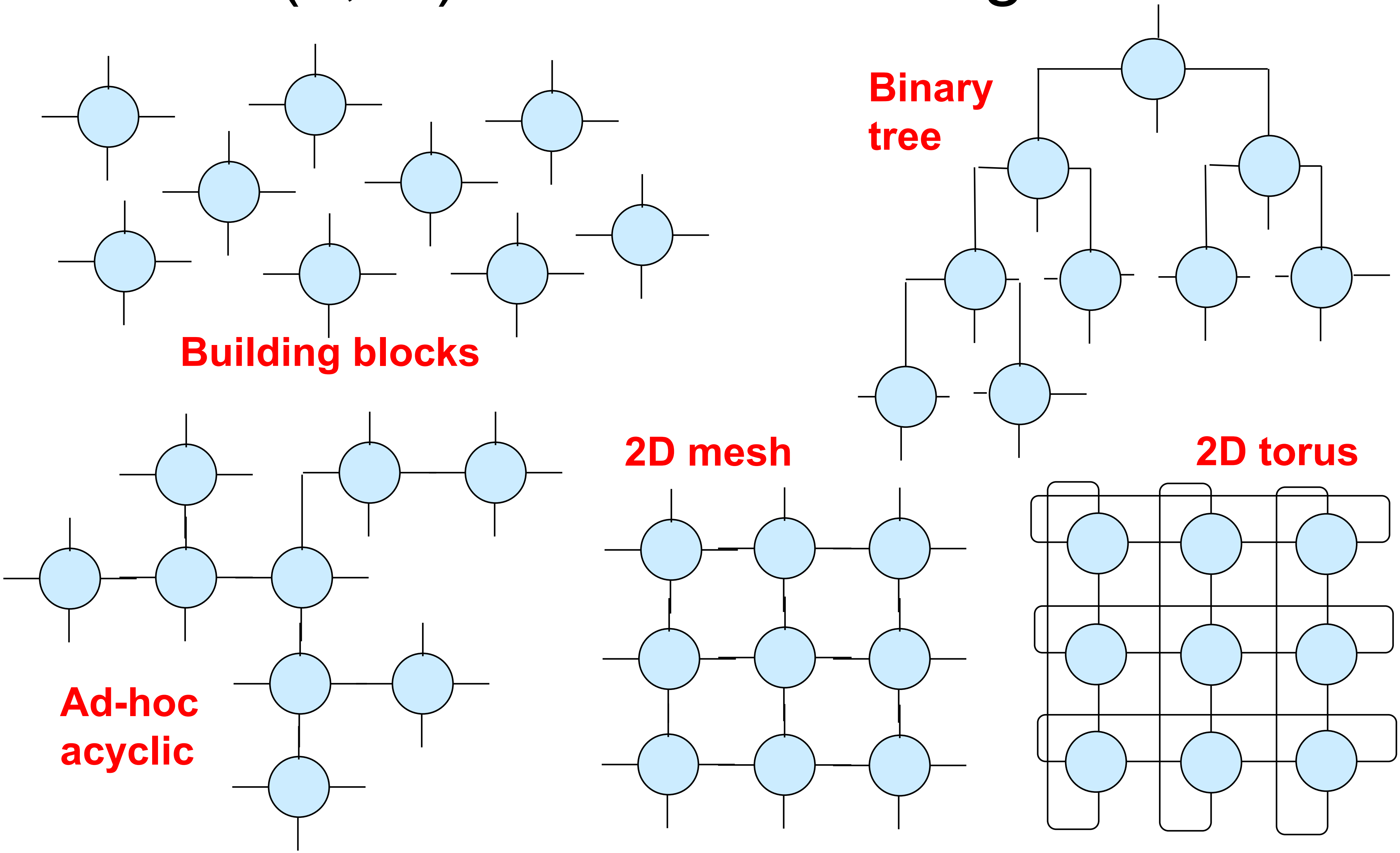


Distributed shared memory

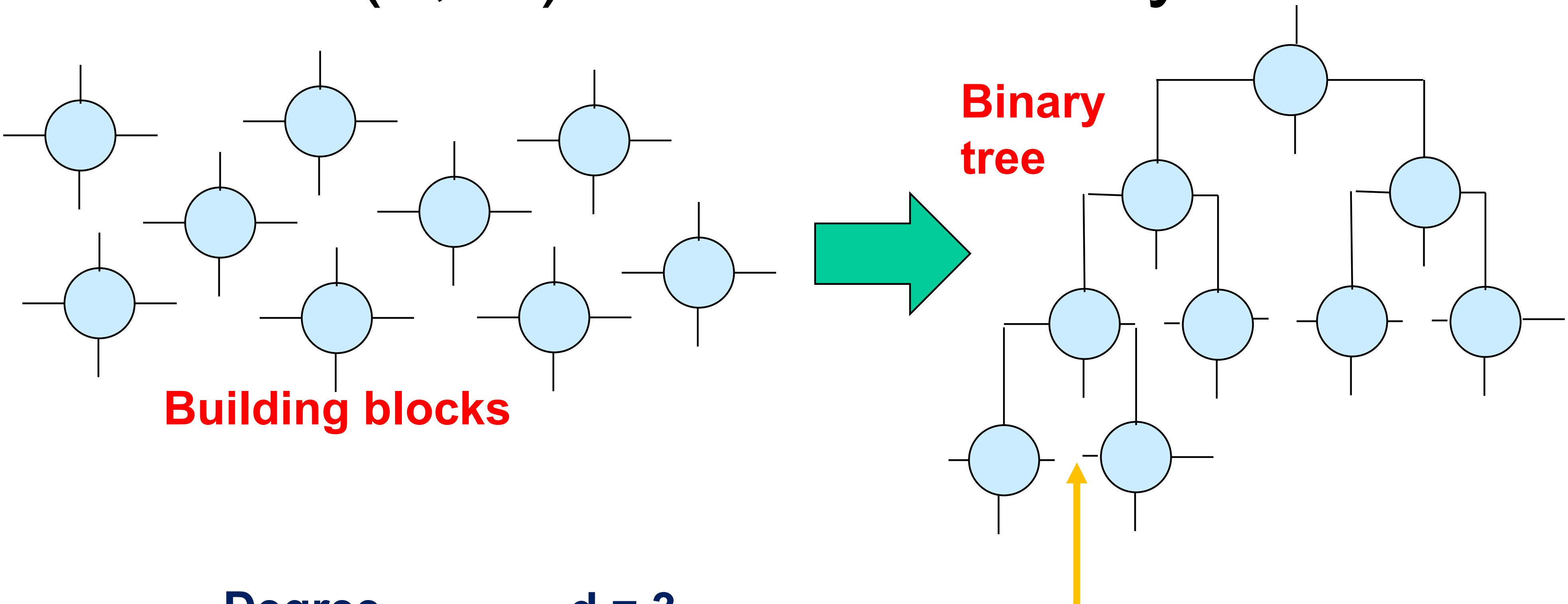
Node Degree and Network Diameter



The (*d*, *D*) Problem: A Tough Puzzle



The (*d*, *D*) Problem: Binary Tree

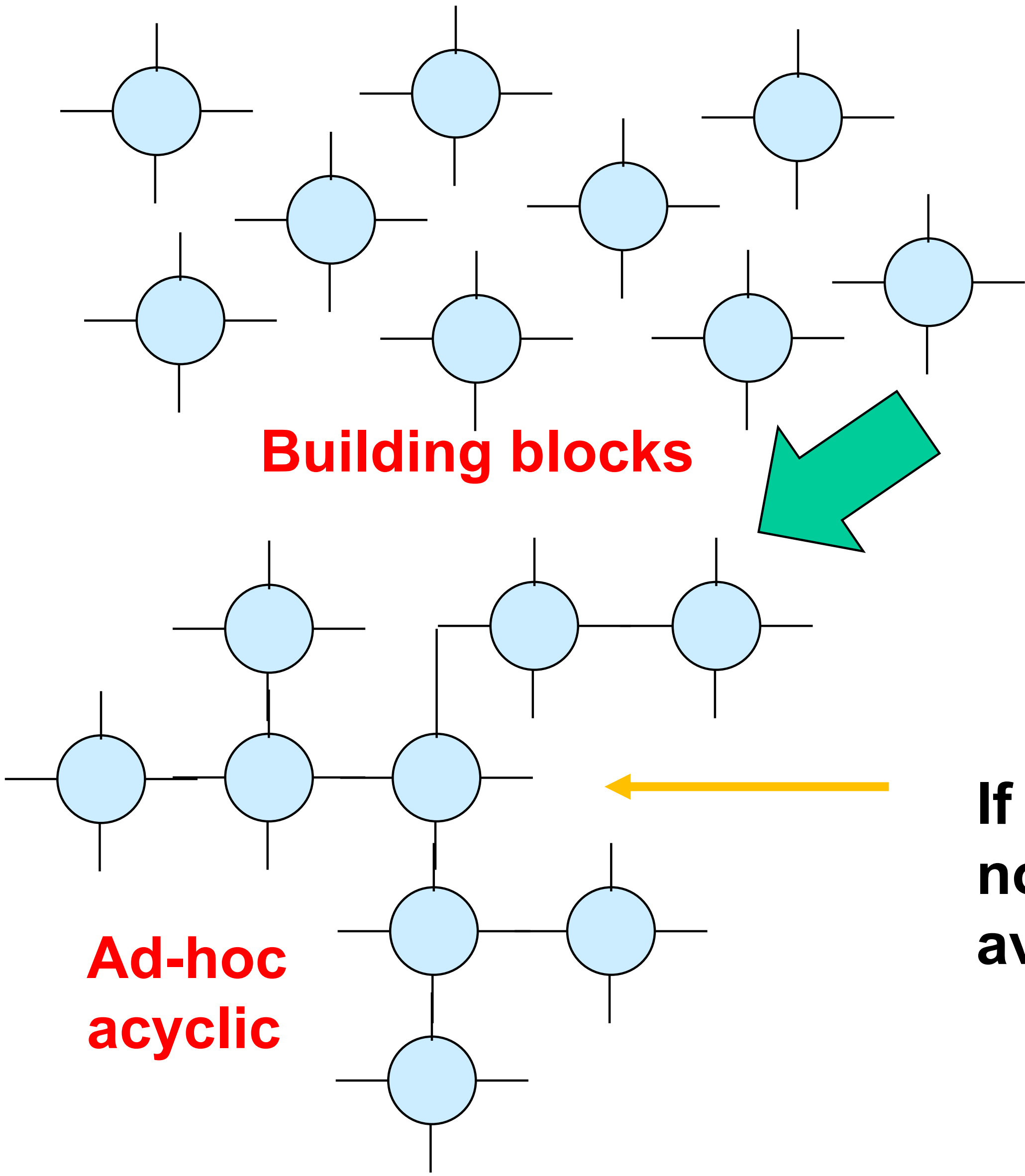


If you connect these links, nothing changes, except for average internode distance

Degree	**d = 3**
Diameter	**D = 5**
Bisection	**B = 1**
Connectivity	**C = 1**

Average internode distance?

The (*d*, *D*) Problem: Ad-hoc Acyclic

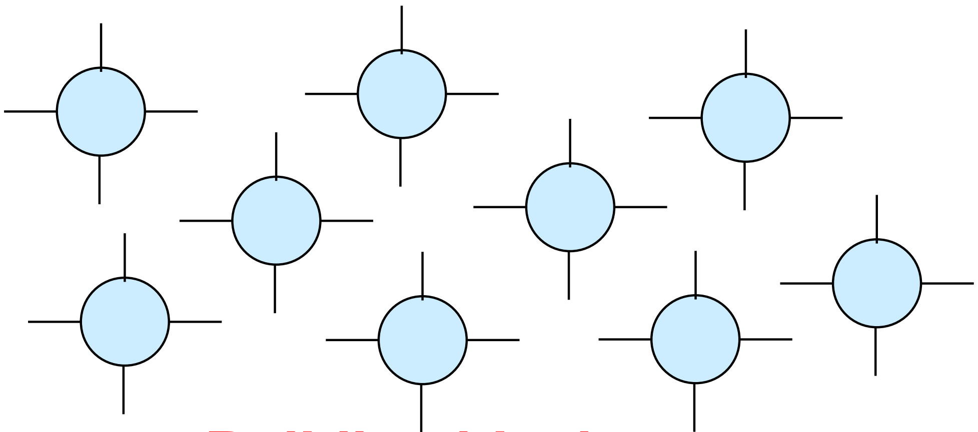


Degree d = 3
Diameter D = 4
Bisection B = 2
Connectivity C = 1

Average internode distance?

If you connect these links, nothing changes, except for average internode distance

The (d, D) Problem: 2D Mesh

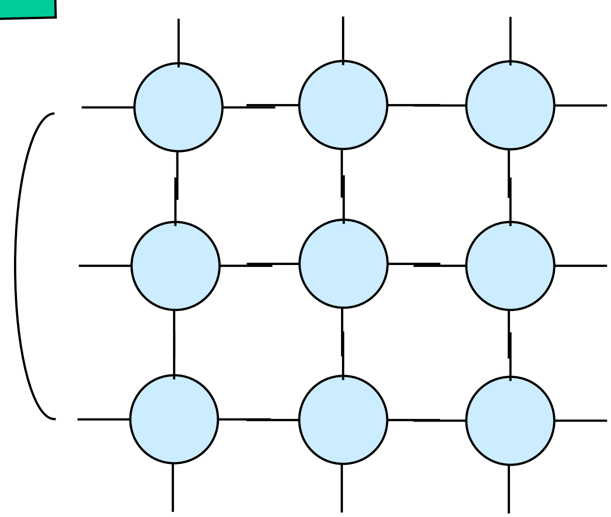


Building blocks

Degree	**d = 4**
Diameter	**D = 4**
Bisection	**B = 4**
Connectivity	**C = 2**

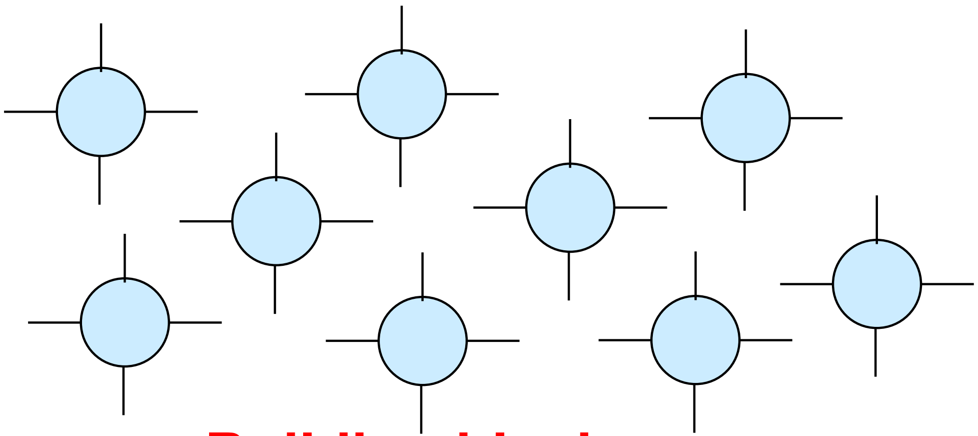
Average internode distance Δ?

2D mesh



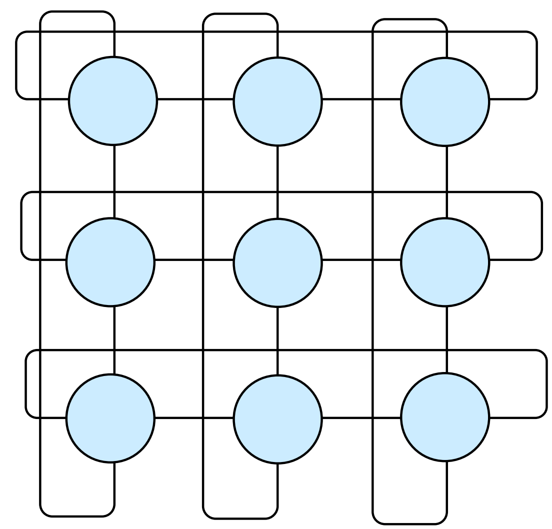
What happens if you connect these links?

The (*d*, *D*) Problem: 2D Torus



Building blocks

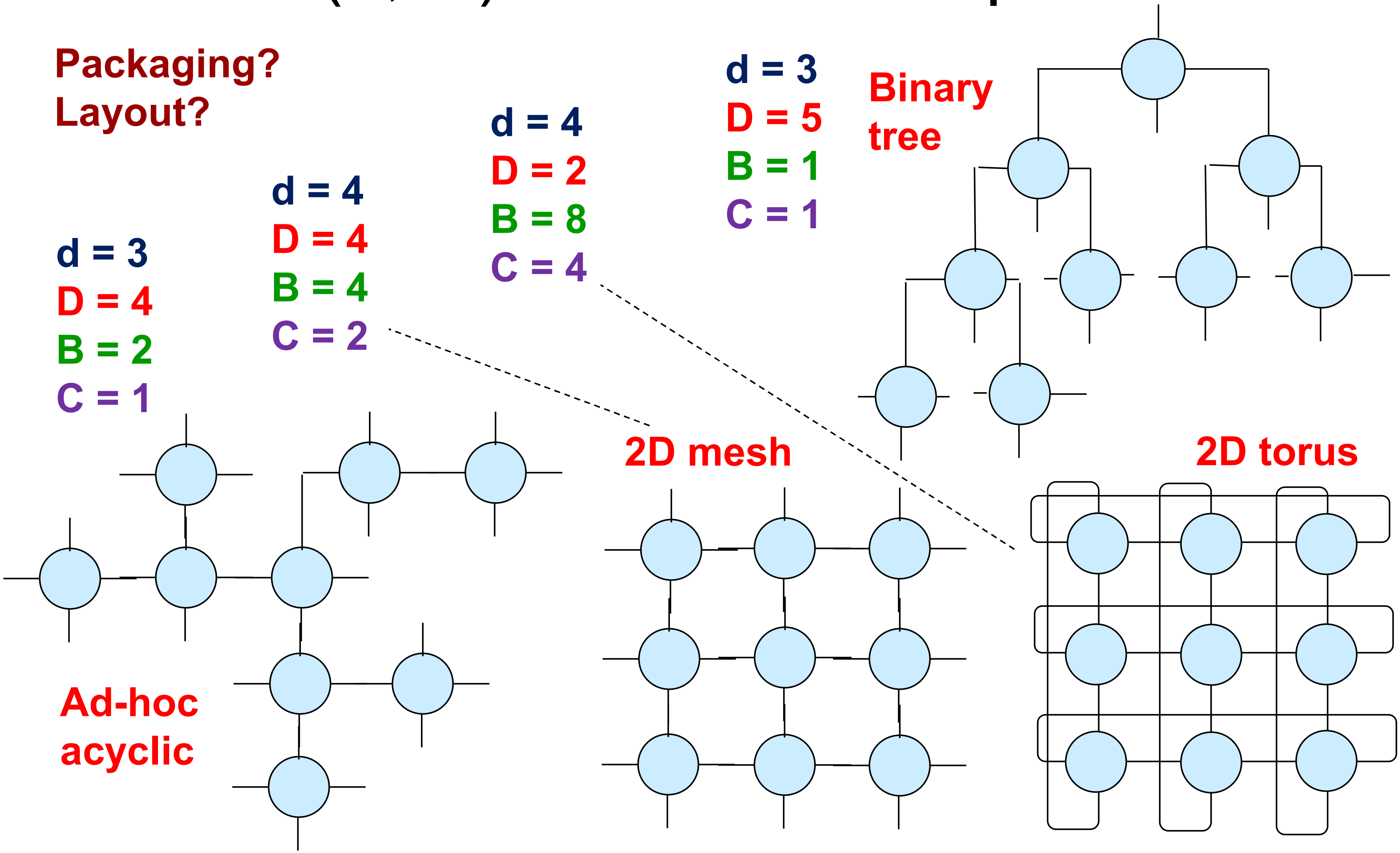
2D torus



Degree	**d = 4**
Diameter	**D = 2**
Bisection	**B = 8**
Connectivity	**C = 4**

Average internode distance Δ?

The (*d*, *D*) Problem: Comparisons

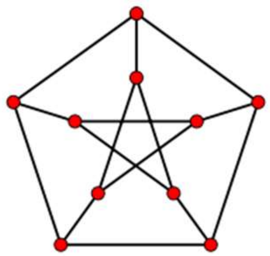


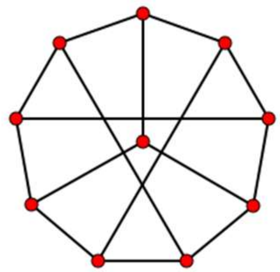
The (d, D) Problem: Moore Bound

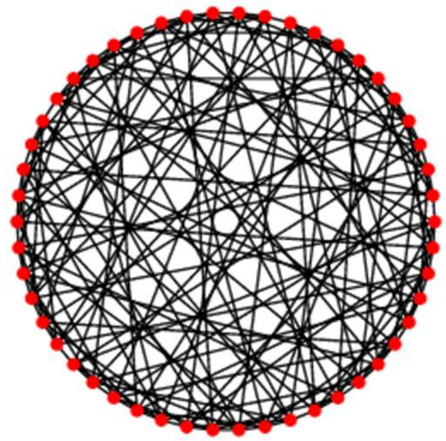
Suppose you have an unlimited supply of degree-d nodes
How many can be connected into a network of diameter D?

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

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



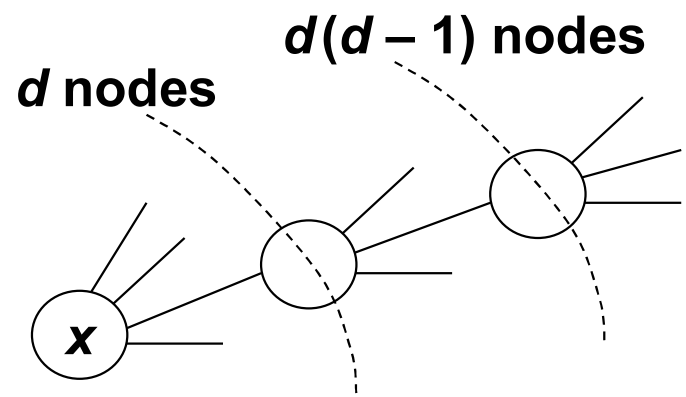




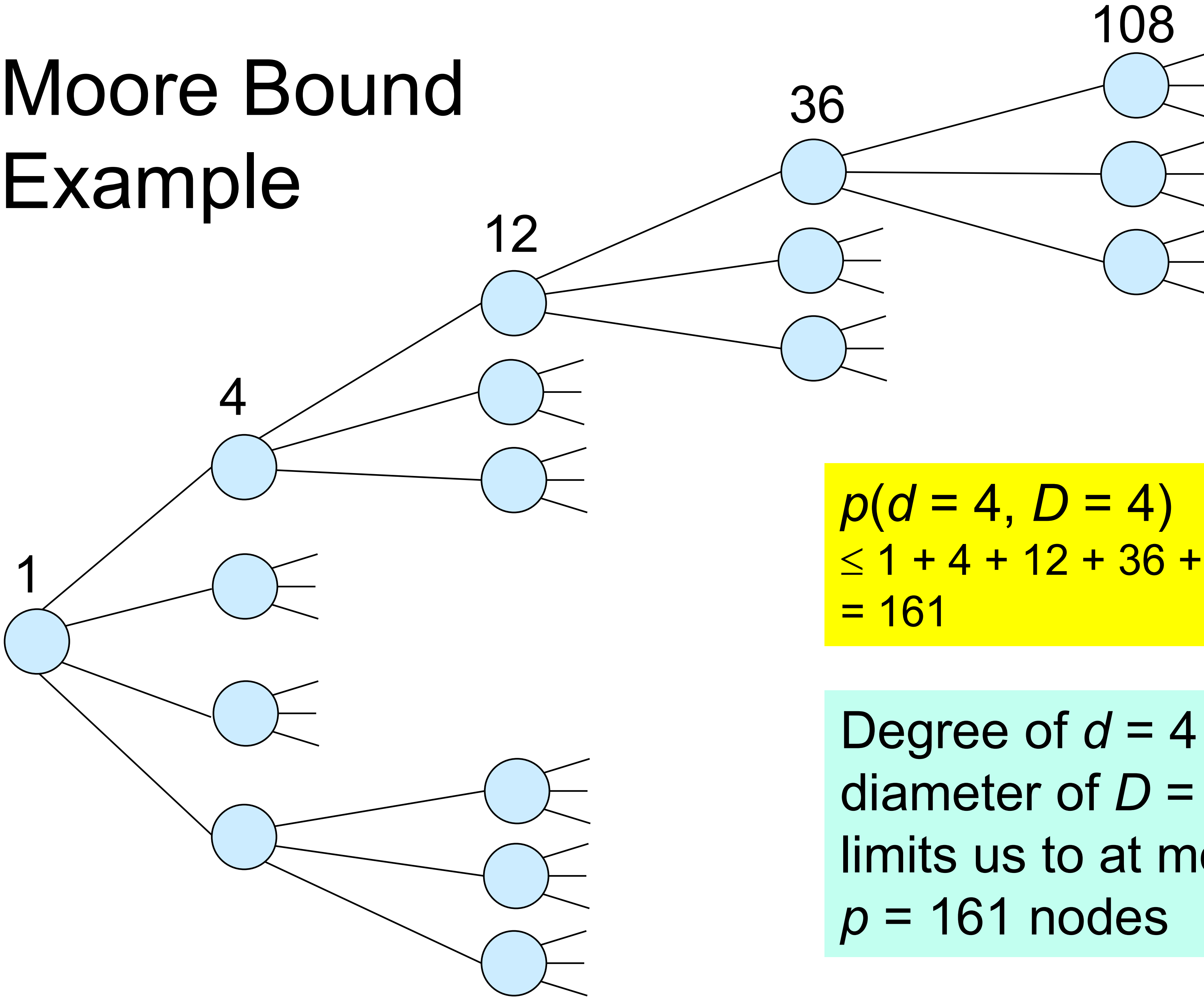
Moore bound (undirected graphs)

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

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



Moore Bound Example

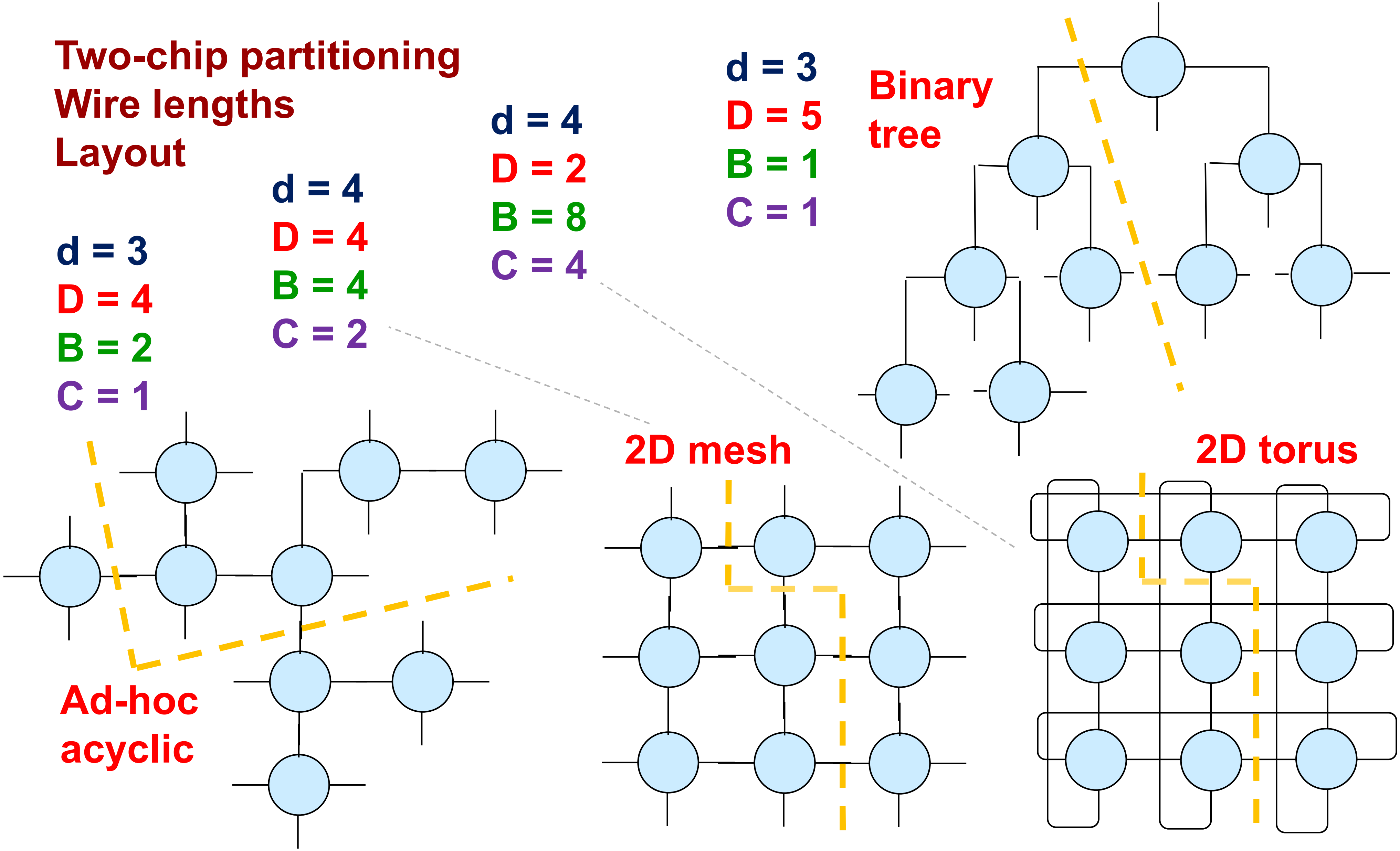


$p(d = 4, D = 4)$
$\leq 1 + 4 + 12 + 36 + 108$
$= 161$

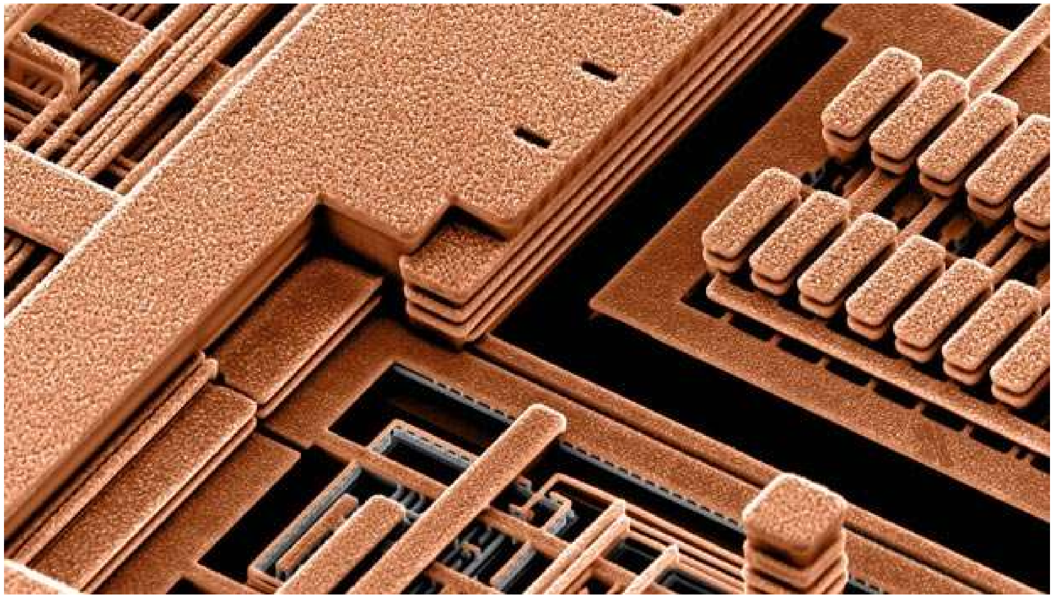
Degree of $d = 4$ and diameter of $D = 4$ limits us to at most $p = 161$ nodes

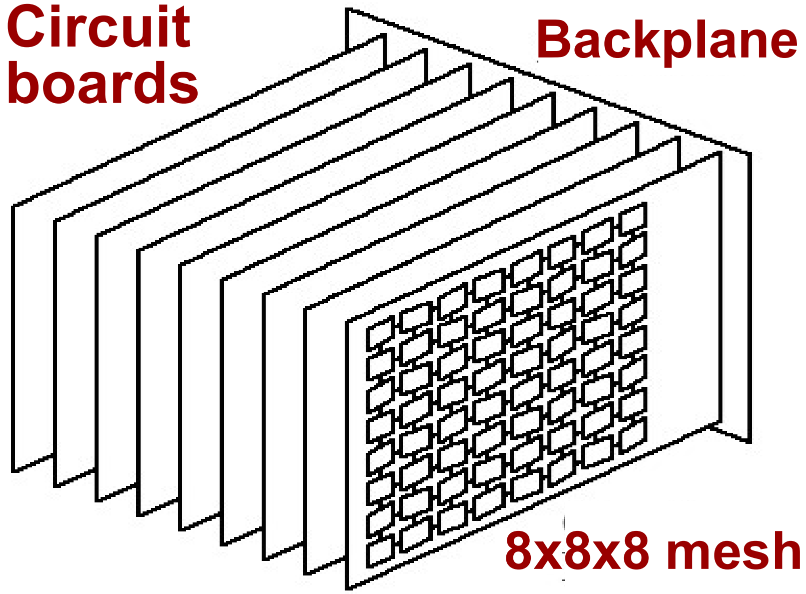
The (*d*, *D*) Problem: Beyond the Math

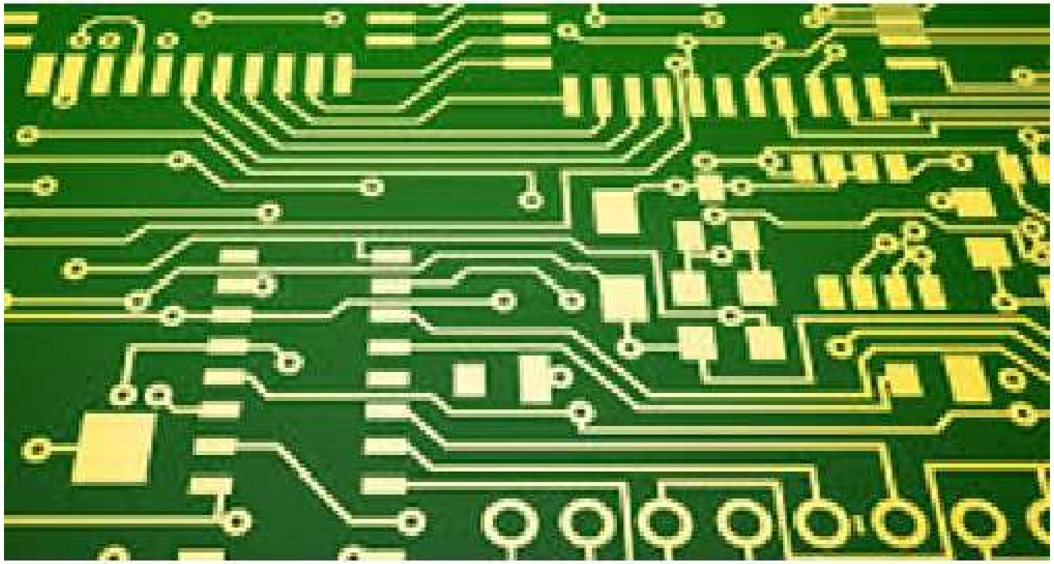
Two-chip partitioning
Wire lengths
Layout

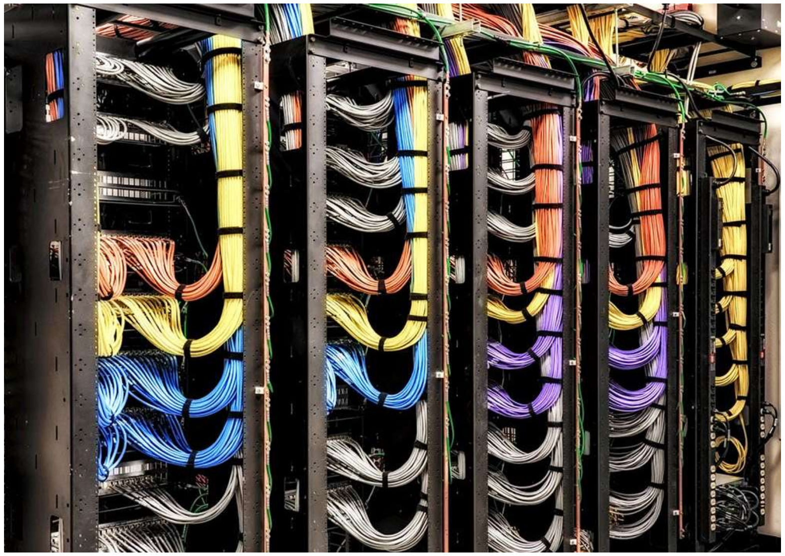


On-Chip, Off-Chip, Inter-Cabinet Wires







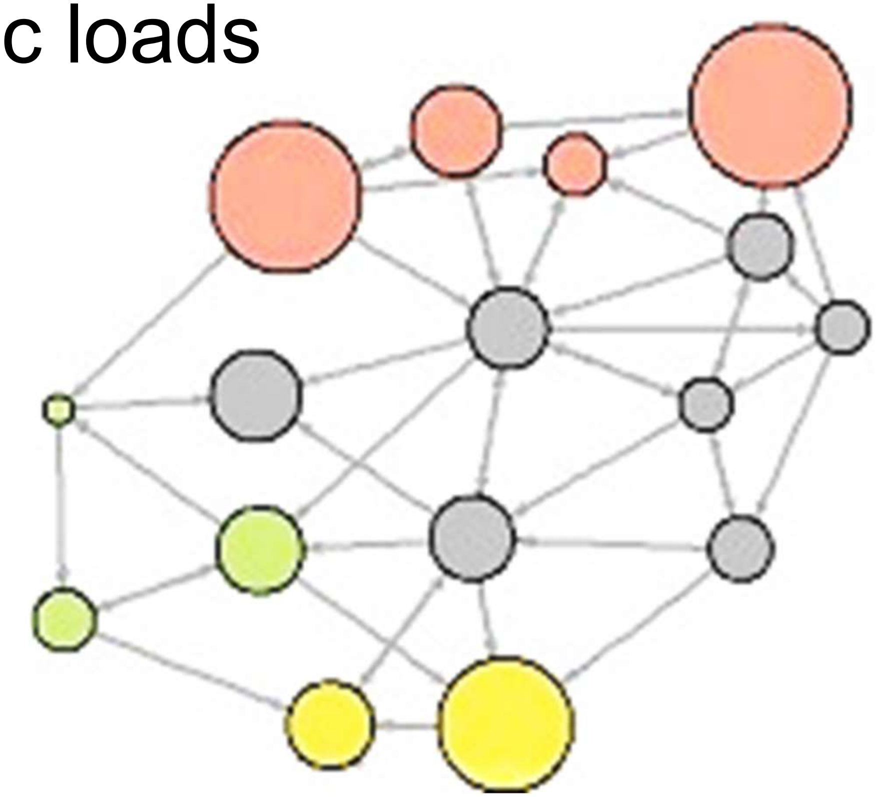


The Network-Design Challenge

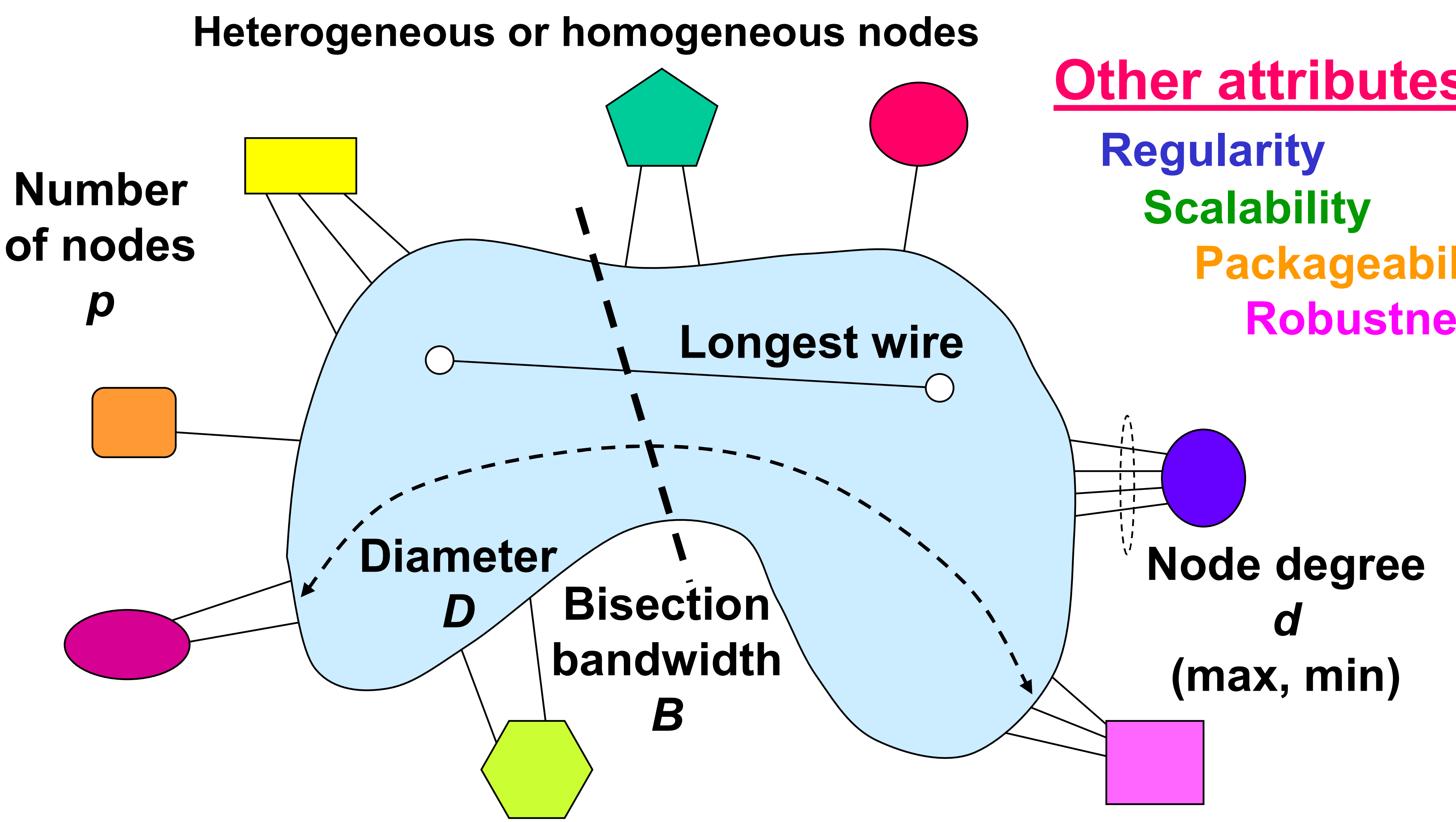
The underlying math problem is already difficult

Now consider these added considerations:

Power frugality
P Performance under realistic loads
Packageability
Q Quality of service
Robustness
R Reliability
Symmetry/Regularity
S Scalability
Serviceability



Interconnection Network Attributes

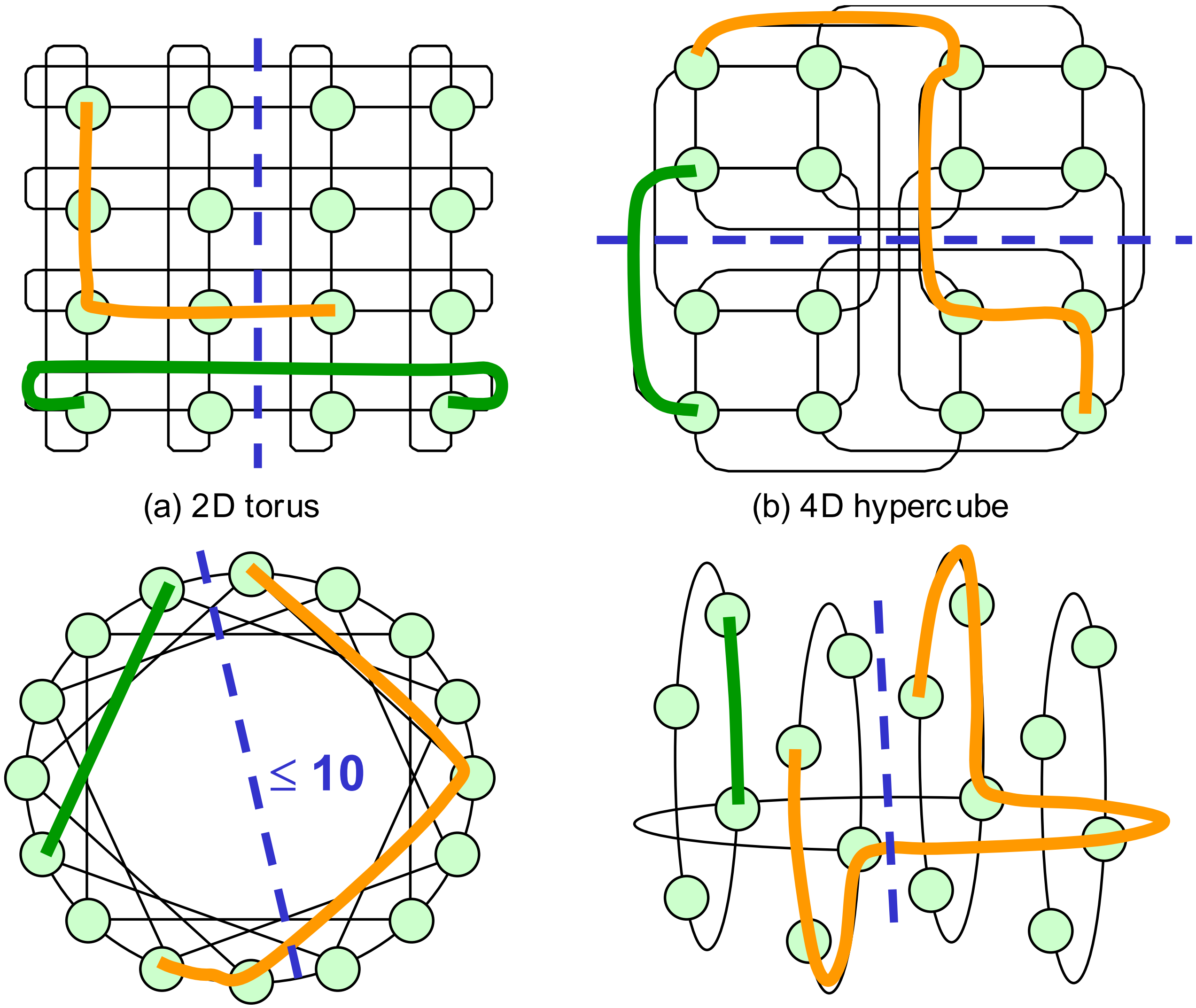


Other attributes:

- Regularity
- Scalability
- Packageability
- Robustness

Average internode distance Δ?

Four Example Interconnection Networks



(a) 2D torus

(b) 4D hypercube

(c) Chordal ring

(d) Ring of rings

Nodes $p = 16$

Degree $d = 4$

Diameter D

Avg. distance Δ

Bisection B

Longest wire

Regularity

Scalability

Packageability

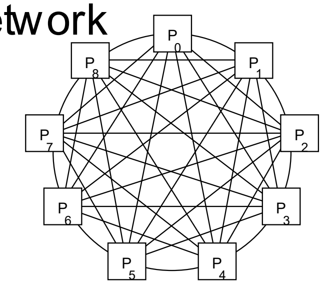
Robustness

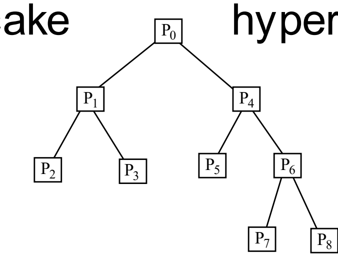
Degree and Diameter Spectrums

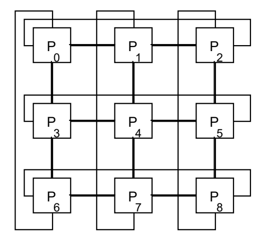
Sublogarithmic diameter | **Superlogarithmic diameter**

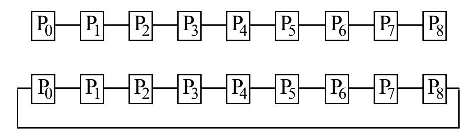
1 — 2 — $\log p / \log\log p$ — $\log p$ — $\sqrt{p}$ — $p/2$ — $p - 1$

- 1: Complete network
- 2: PDN
- $\log p / \log\log p$: Star, pancake
- $\log p$: Binary tree, hypercube
- $\sqrt{p}$: Torus
- $p/2$: Ring
- $p - 1$: Linear array









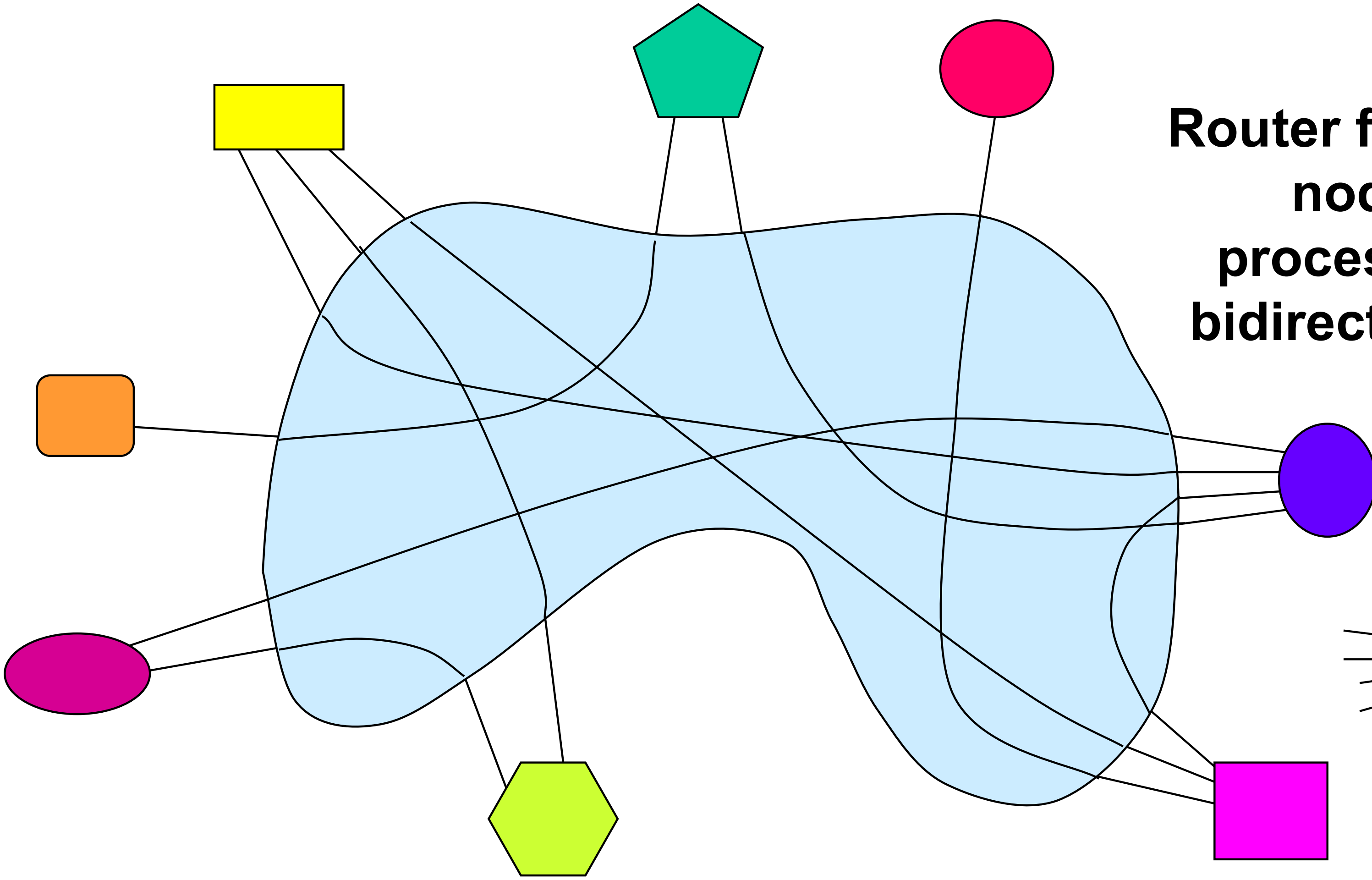
Sublogarithmic degree | **Superlogarithmic degree**

1 — 2 — $\log p / \log\log p$ — $\log p$ — $\sqrt{p}$ — $p - 1$

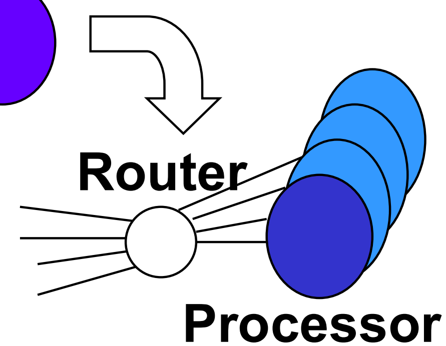
- 2: **Linear array, ring**
- $\log p / \log\log p$: Star, pancake
- $\log p$: **Hypercube**
- $\sqrt{p}$: **PDN**
- $p - 1$: **Complete network**

Direct Interconnection Networks

Nodes (or associated routers) directly linked to each other

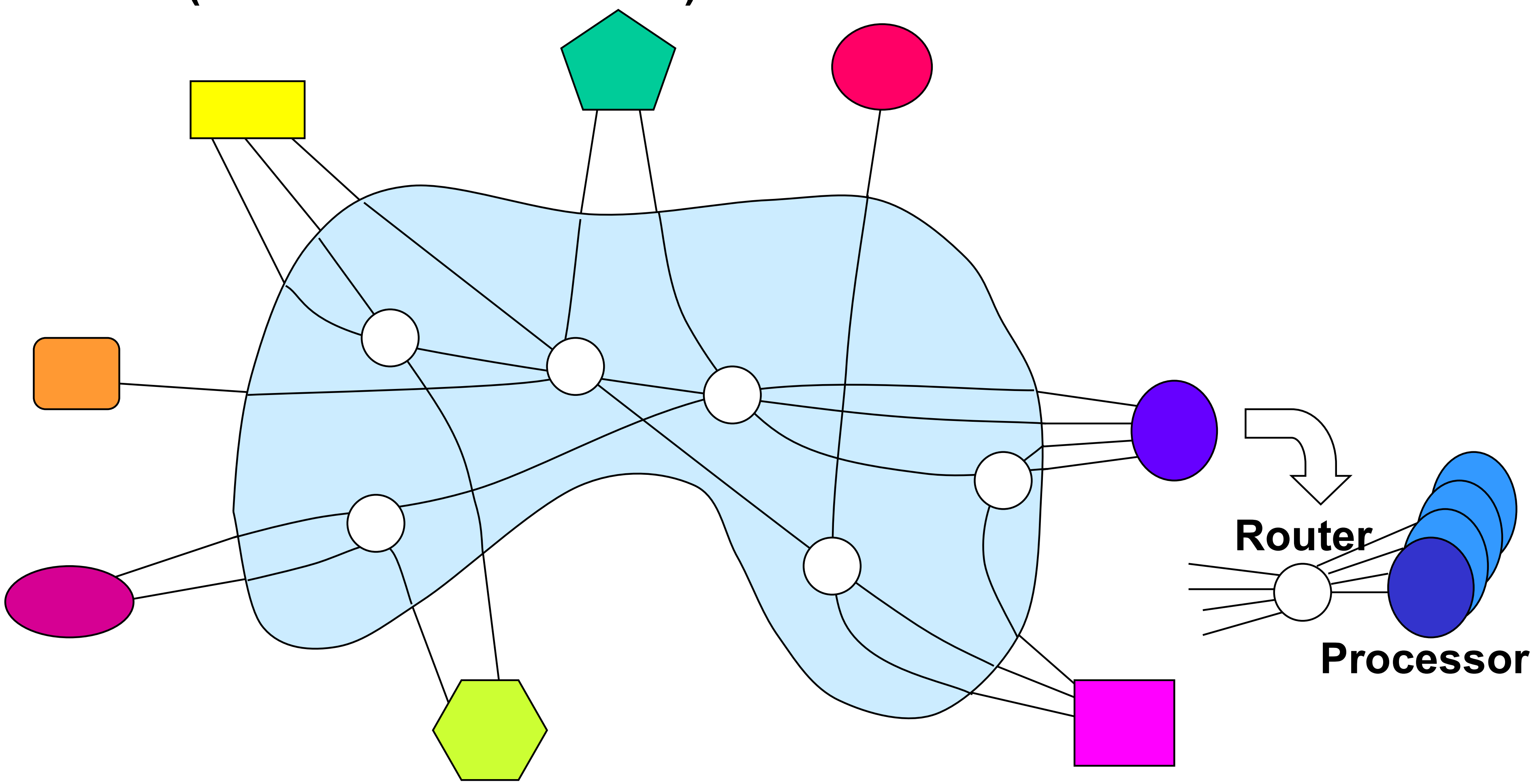


Router for a degree-*d* node with *q* processors: $d \times q$ bidirectional switch

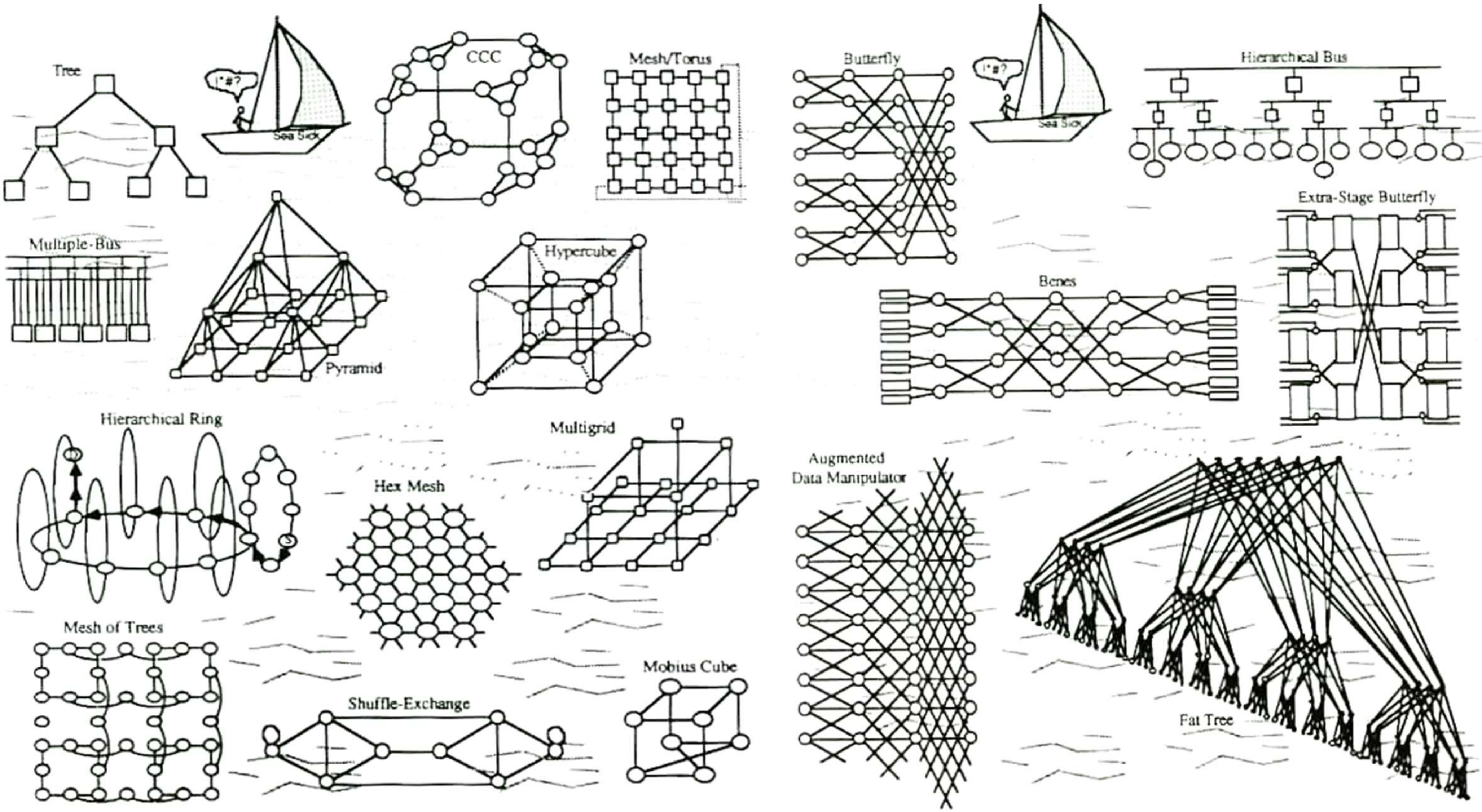


Indirect Interconnection Networks

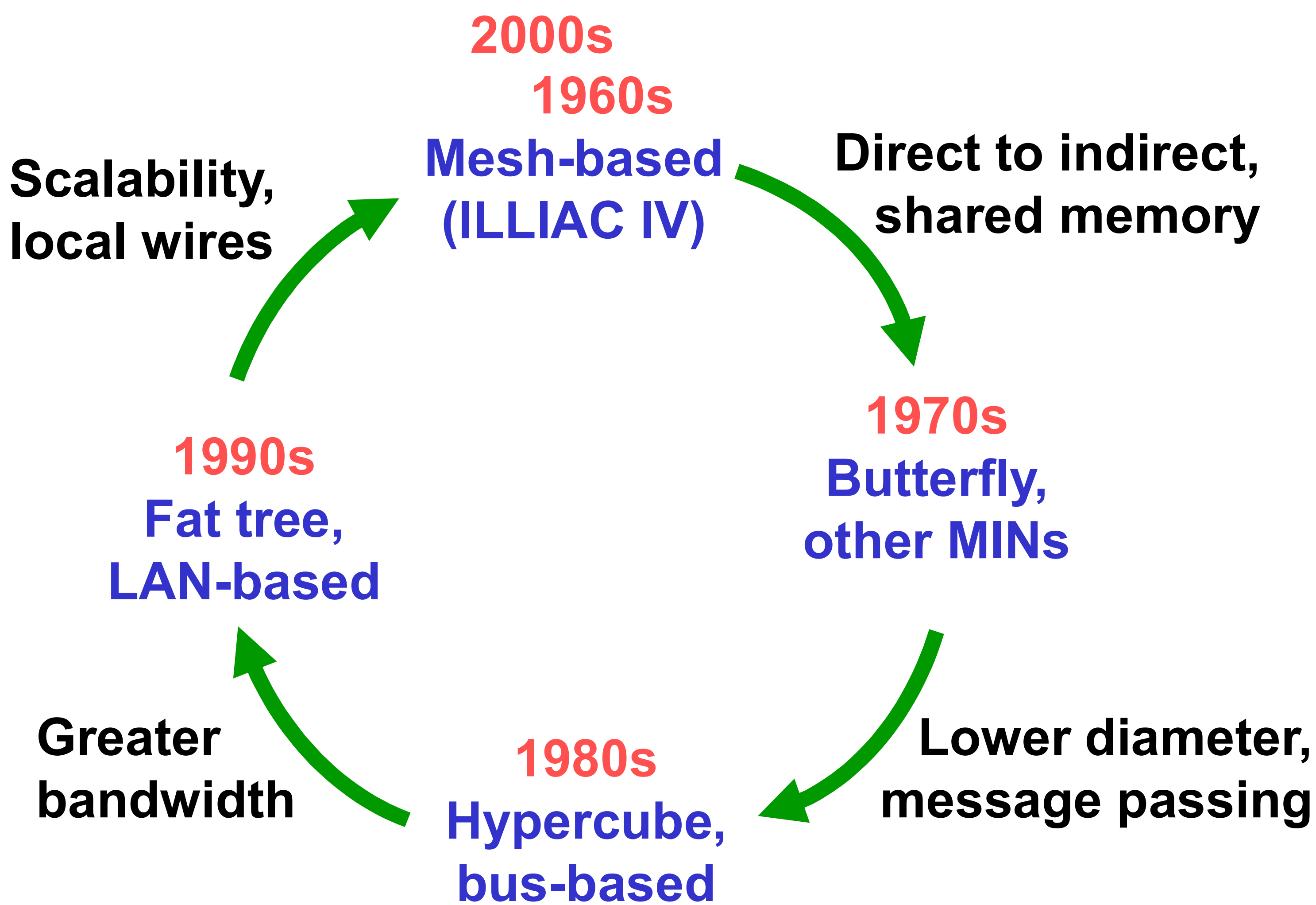
Nodes (or associated routers) linked via intermediate switches



Sea of Interconnection Networks



A Bit of History: Moving Full Circle

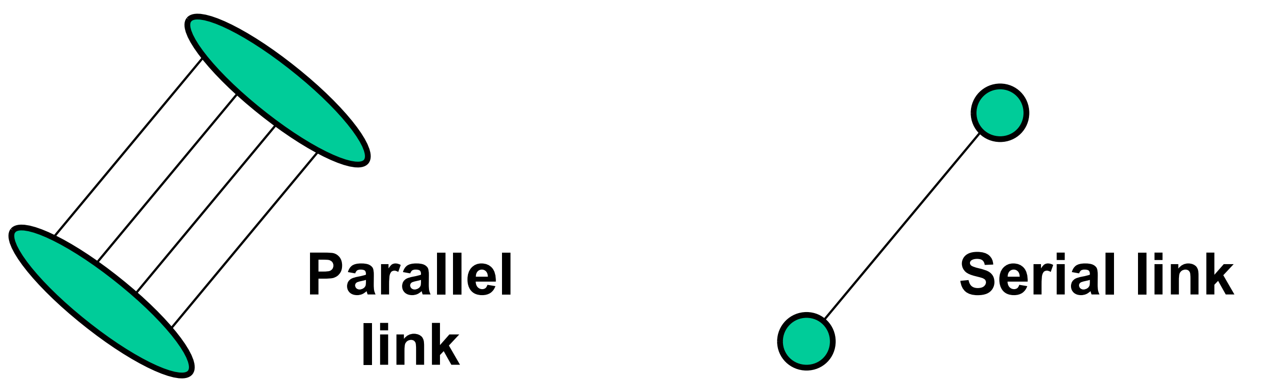


So, only a small portion of the sea of networks has been explored in practical parallel computers

Link Malfunctions

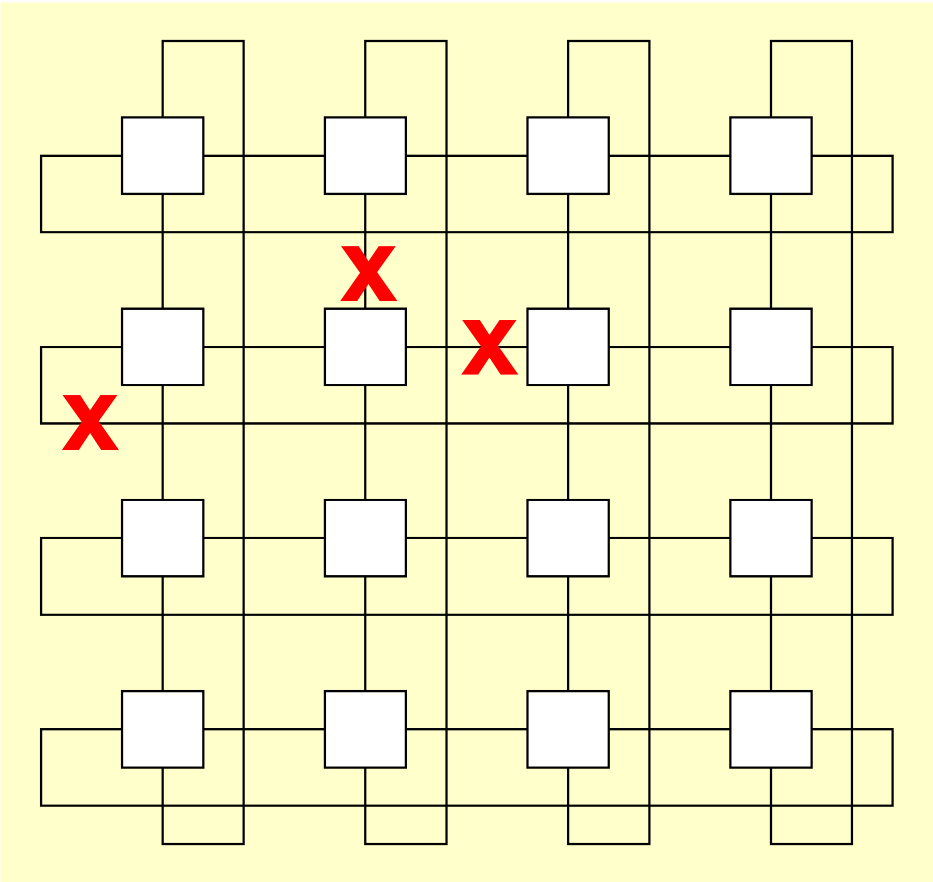
Link data errors or outage

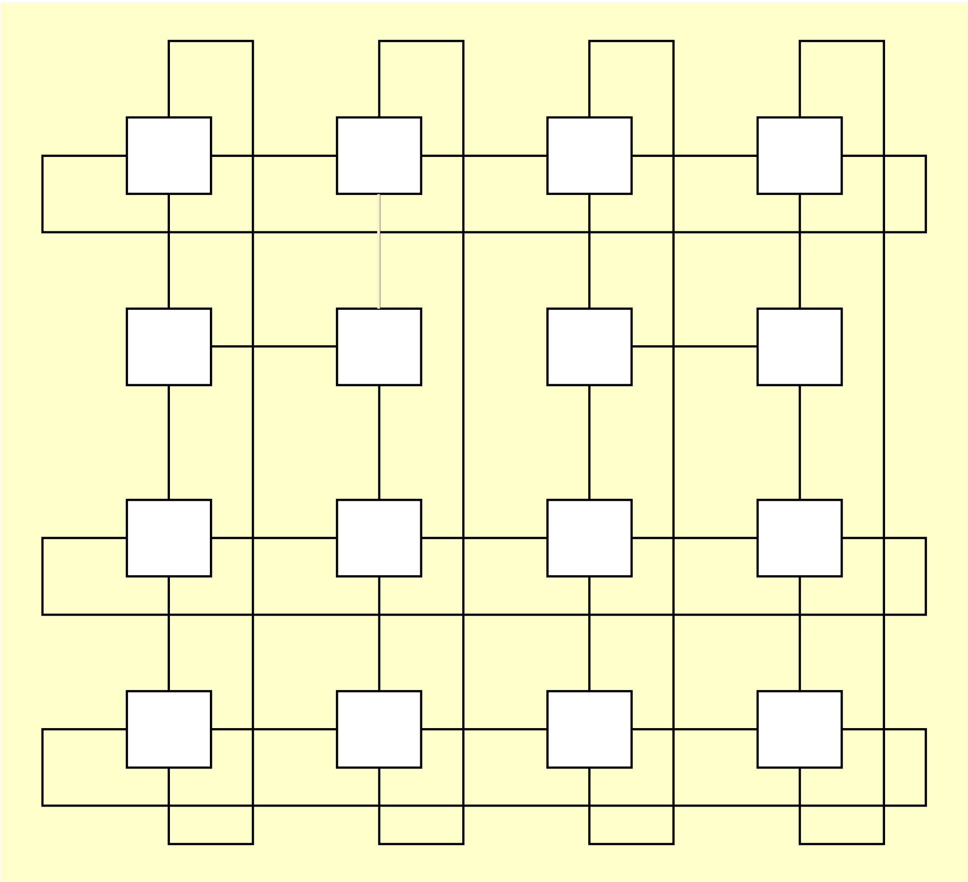
- **Use of error-detecting/correcting codes (redundancy in time/space)**
- **Multiple transmissions via independent paths (redundancy in space)**
- **Retransmission in the same or different format (time redundancy)**
- **Message echo/ack in the same or different format (time redundancy)**
- **Special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



Link Outage Example

Three links go out in this torus

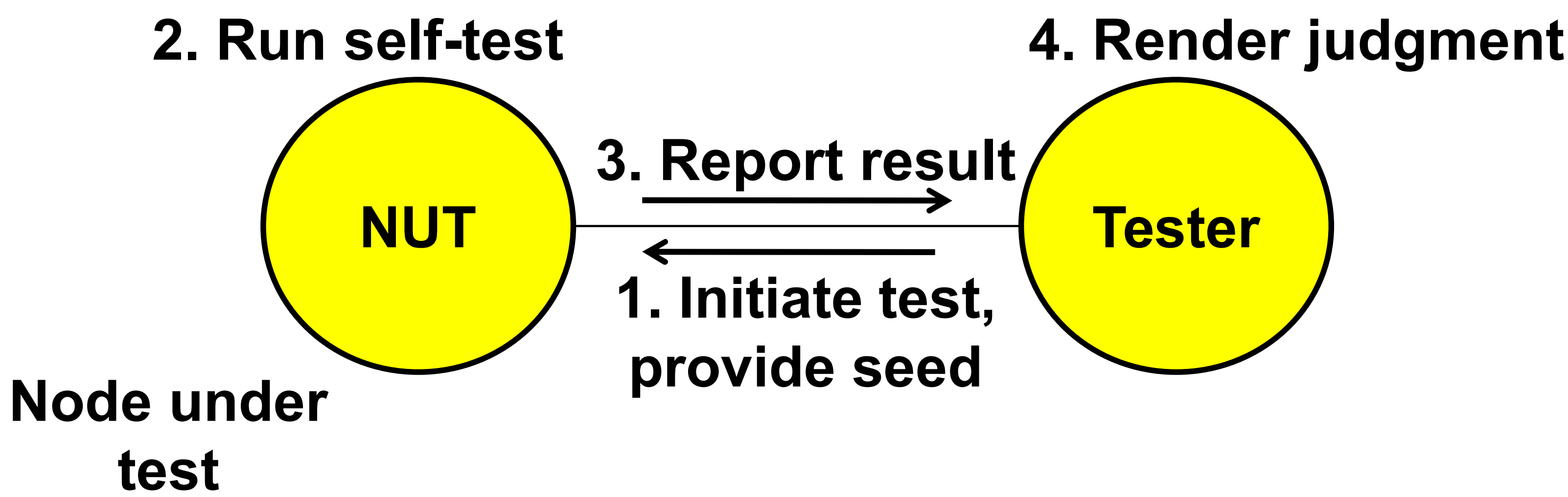




Node Malfunctions

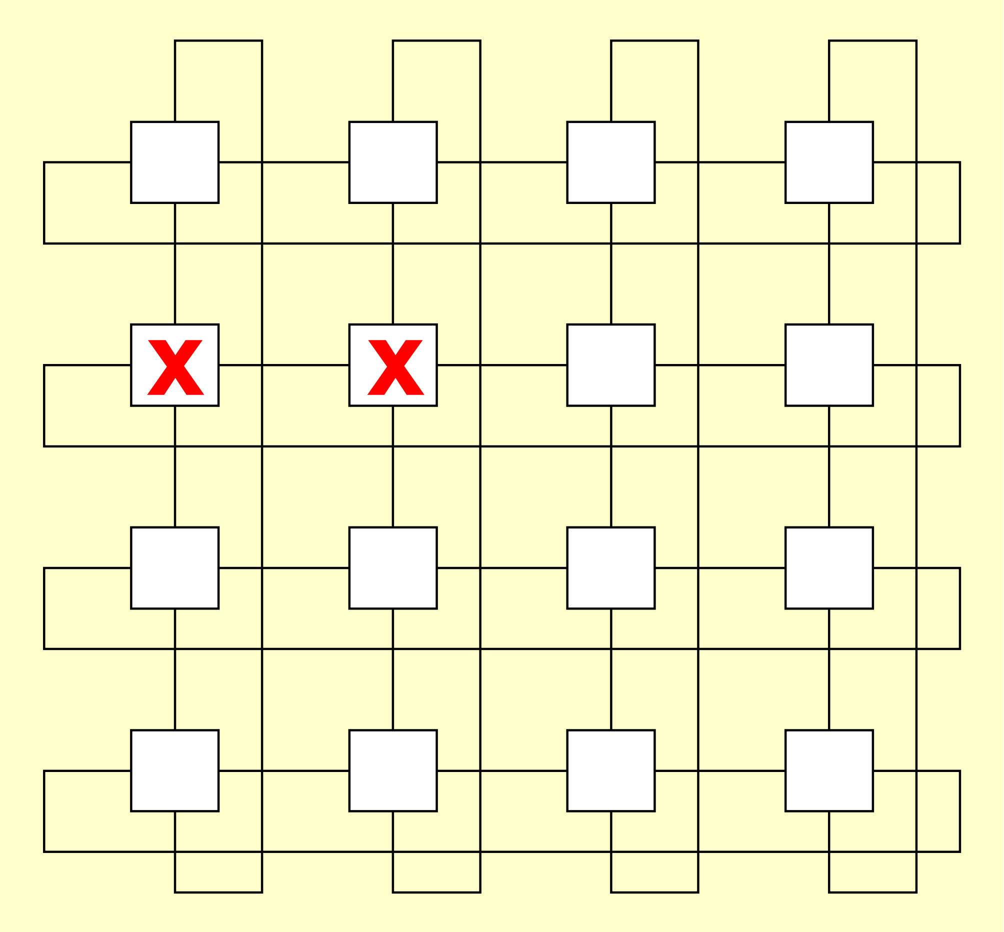
Node functional deviations or outage

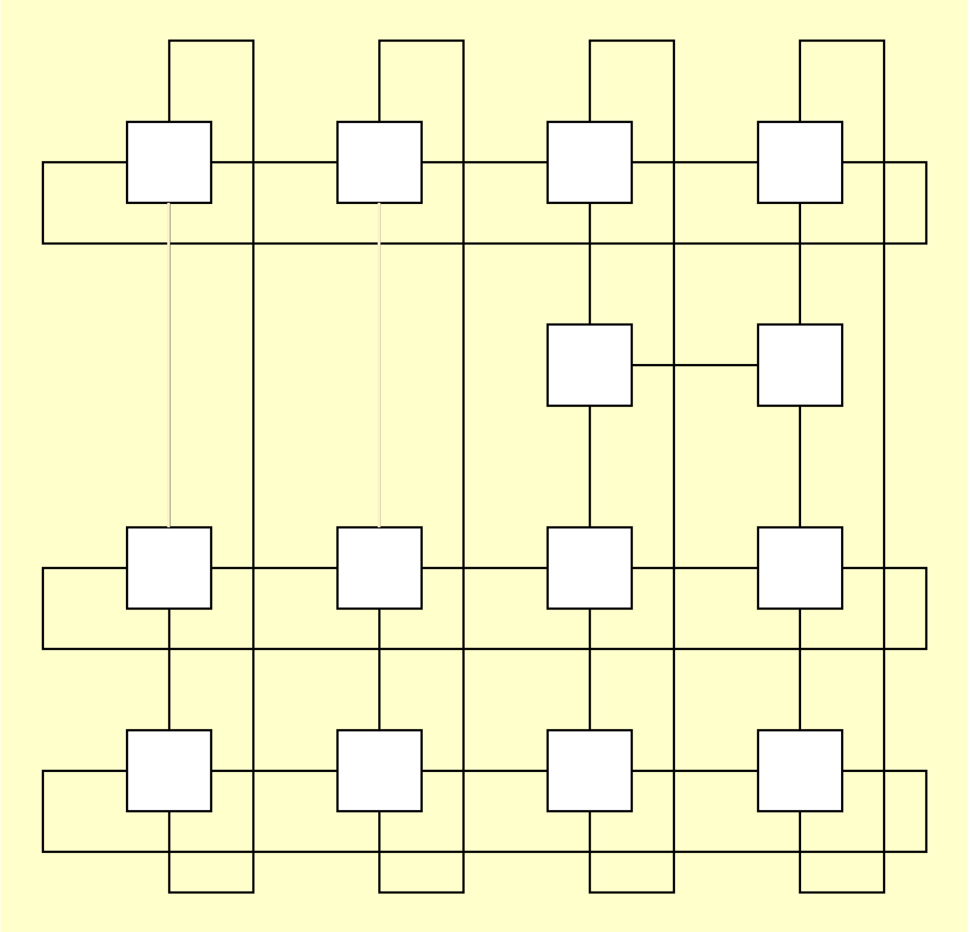
- **Periodic self-test based on a diagnostic schedule**
- **Self-checking design for on-line (concurrent) malfunction detection**
- **Periodic testing by neighboring nodes**
- **Periodic self-test with externally supplied seed**



Node Outage Example

Two nodes go out in this torus



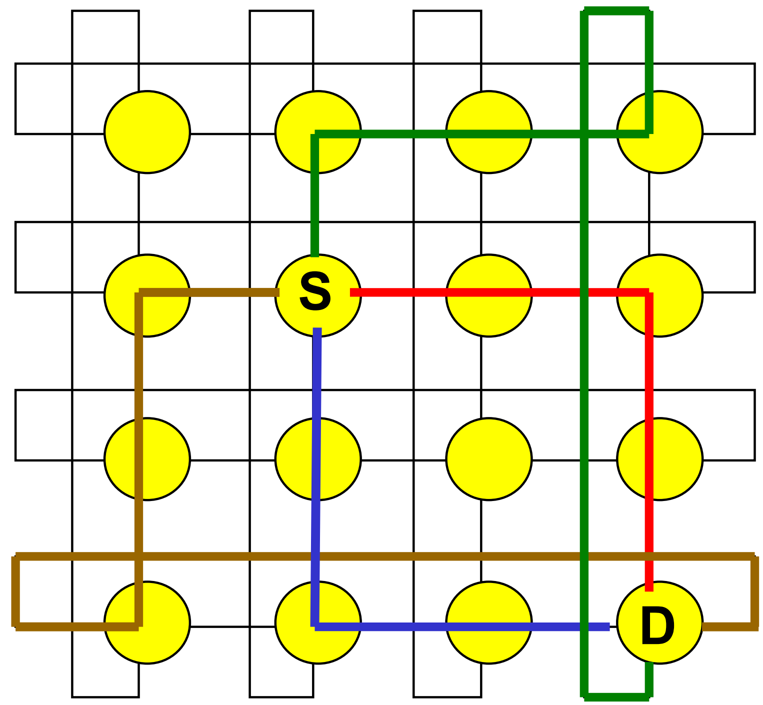


Multiple Disjoint Paths

Connectivity $\kappa \le d_{min}$ (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

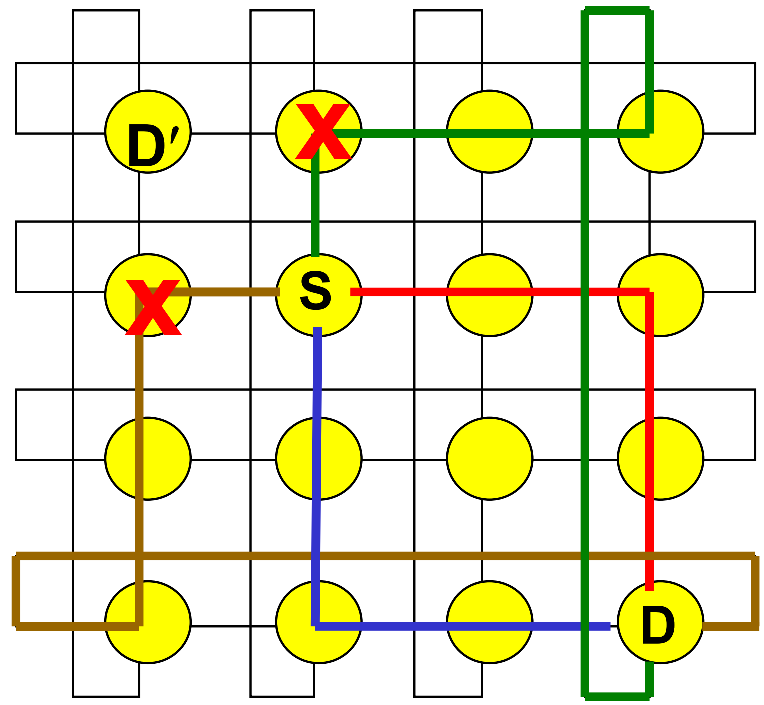
Parallel paths lead to robustness, as well as greater performance



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

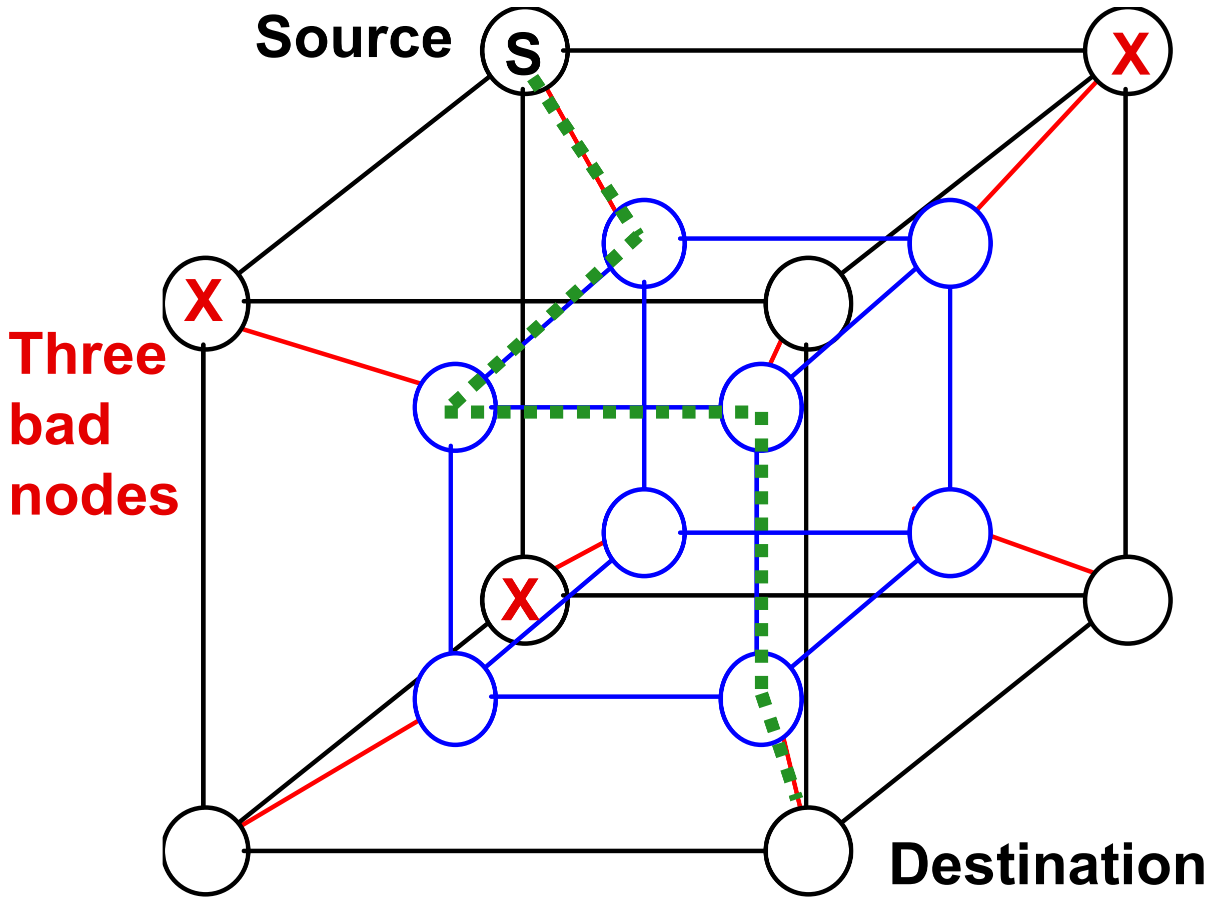
Path length increases to 4 (via wraparound links to D′)

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 (Fault) 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

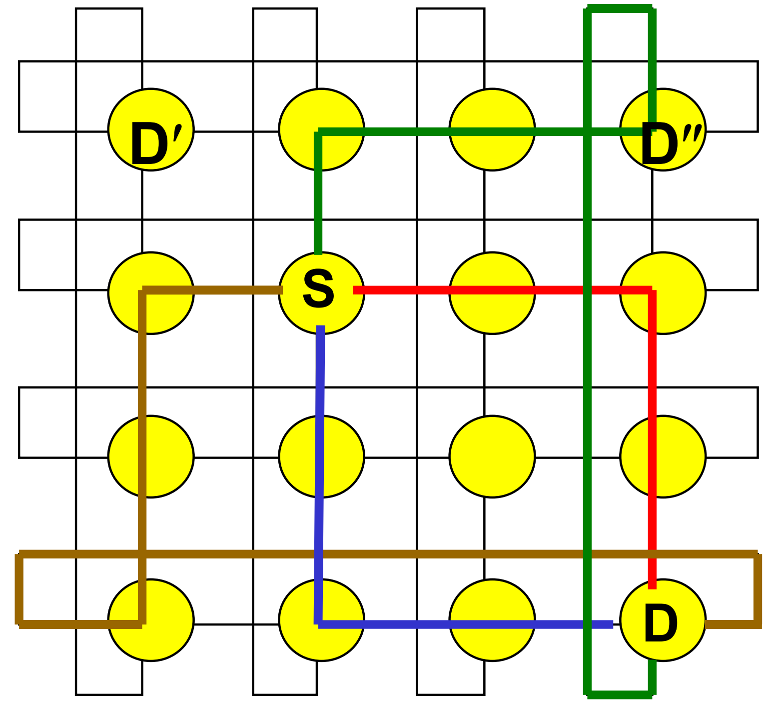
Malfunction diameter of the *q*-cube is *q* + 1

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

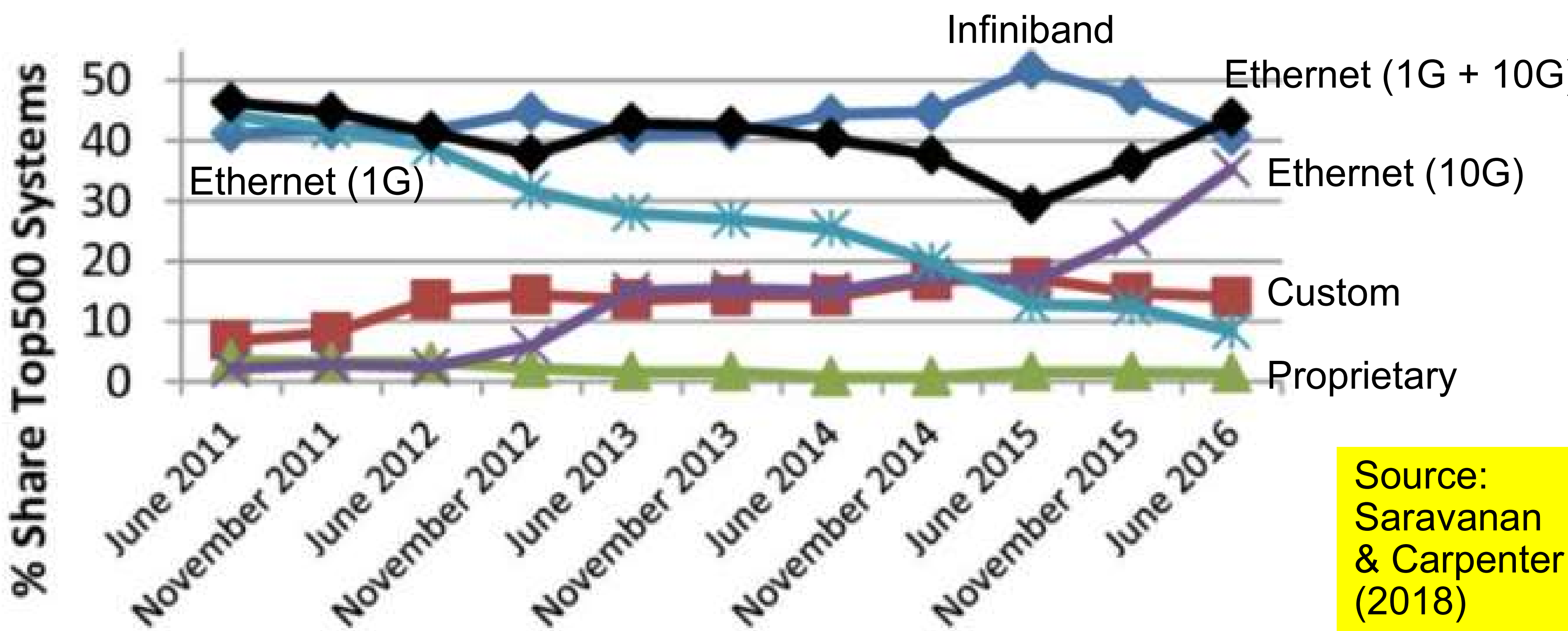
Classes of Interconnection Networks

Buses / Ethernet

Meshes & Tori (2D and higher-dimensional arrays)

Fat Trees

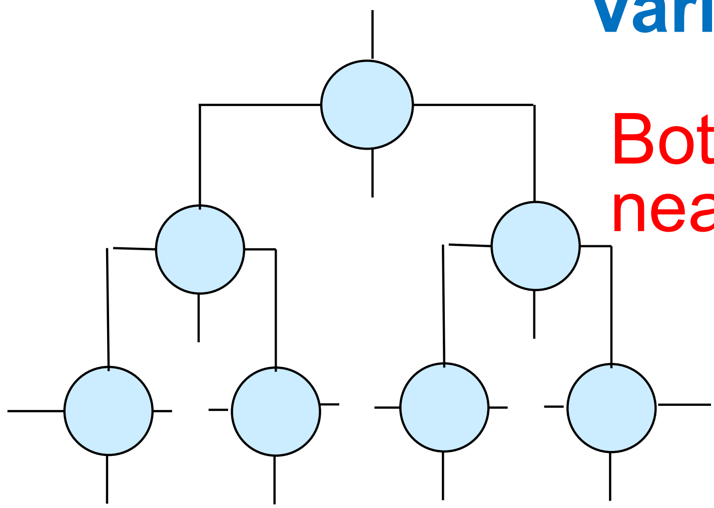
Many others (sea of interconnection networks)



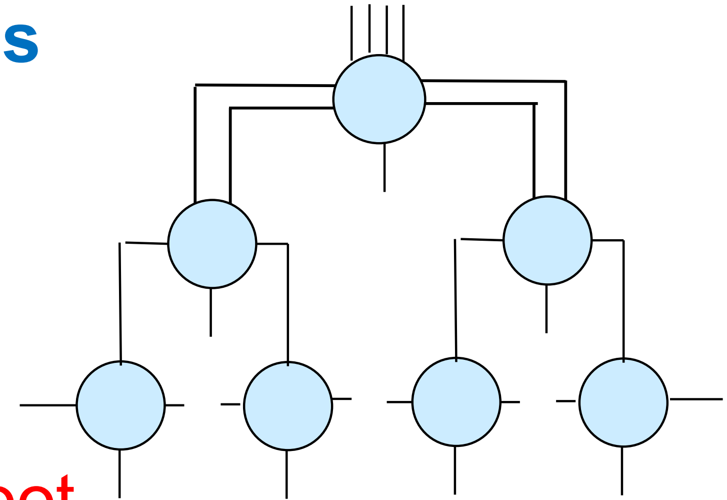
Source: Saravanan & Carpenter (2018)

Modern Data-Center Networks

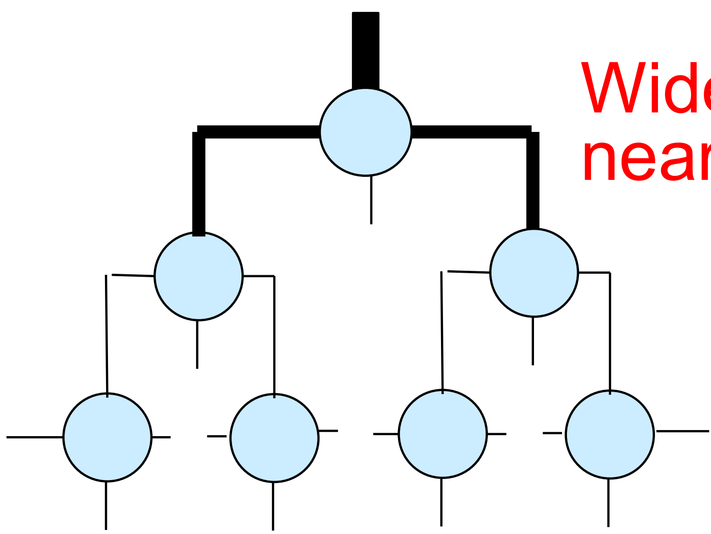
Variations on fat trees



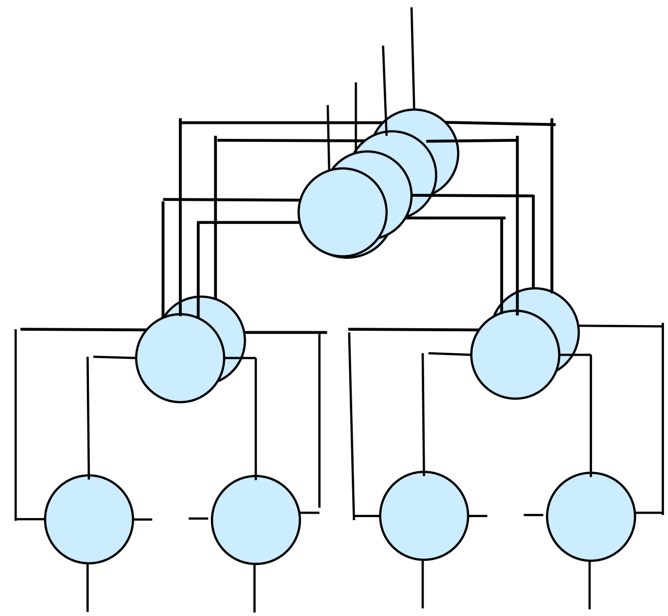
Bottleneck
near the root



More links
near the root



Wider links
near the root



More nodes
and links
near the root

Facebook & Google Data-Center Networks

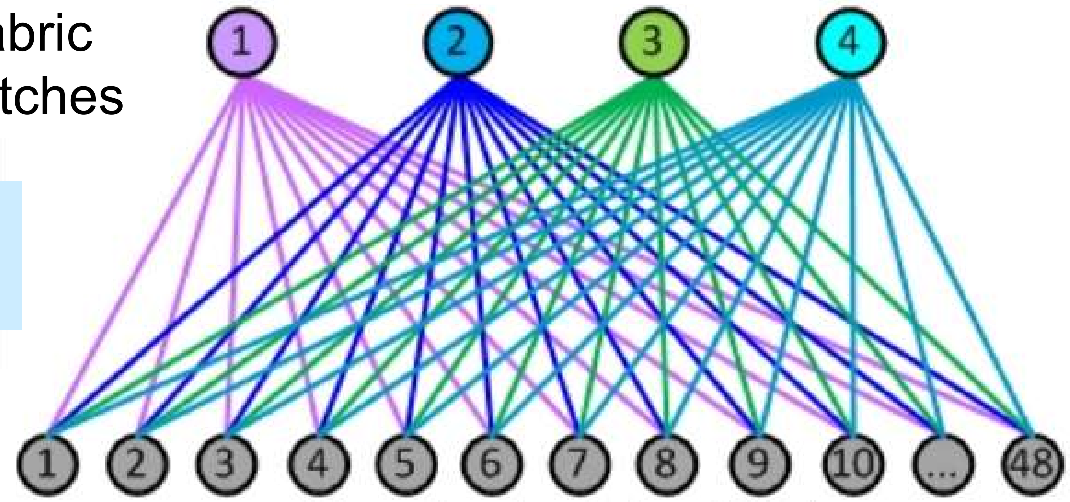
Nearly all modern data-center networks are based on some variation of fat-tree topology

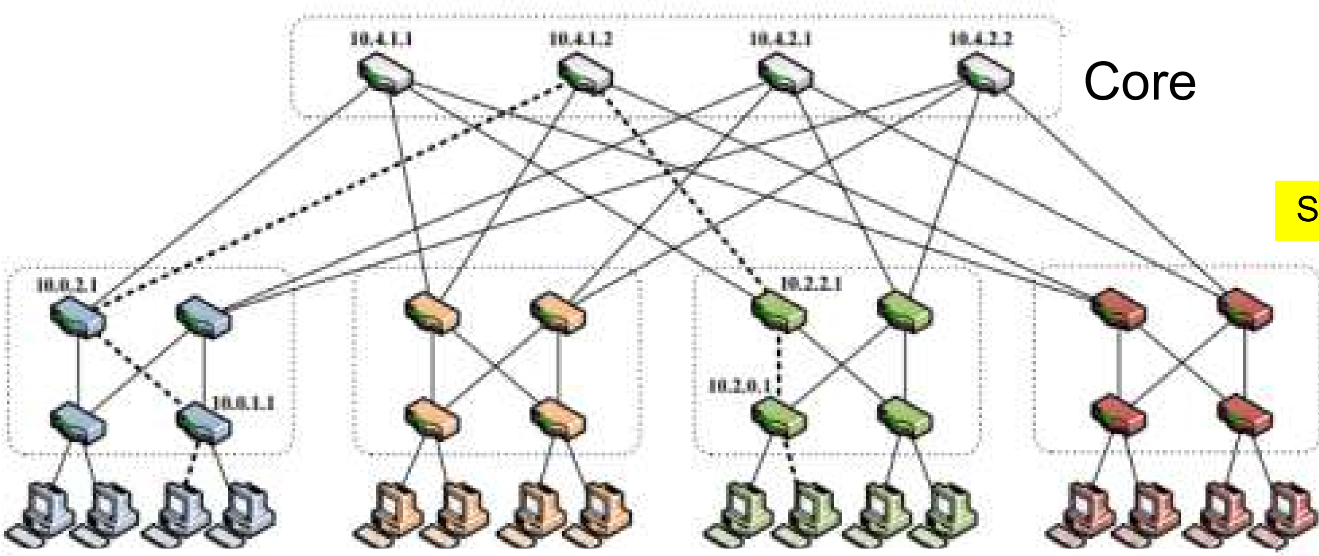
Facebook, 2014

Ten-year review:
Tom Furlong, 2021

Fabric Switches

Top-of-Rack Switches





Core

Aggregation

Edge

Pods

Google, 2008
(since updated)

Source: Lebiednik, Mangal, & Tiwari (2016)

Recent:
Jupiter (see 2022 paper by A. Vahdat)

Swapped (OTIS) Networks

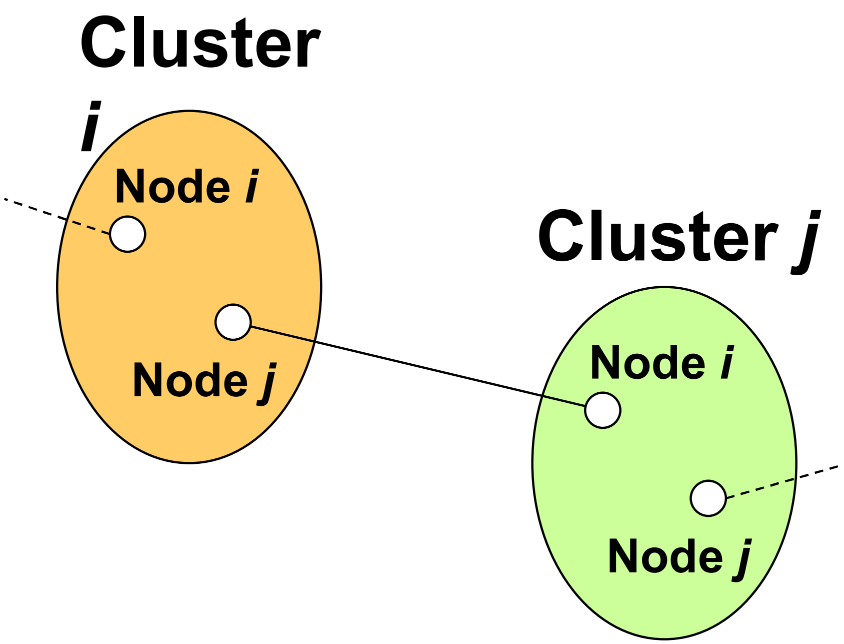
Swapped network

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 Scalability

A. Logarithmic-diameter basis network

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

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

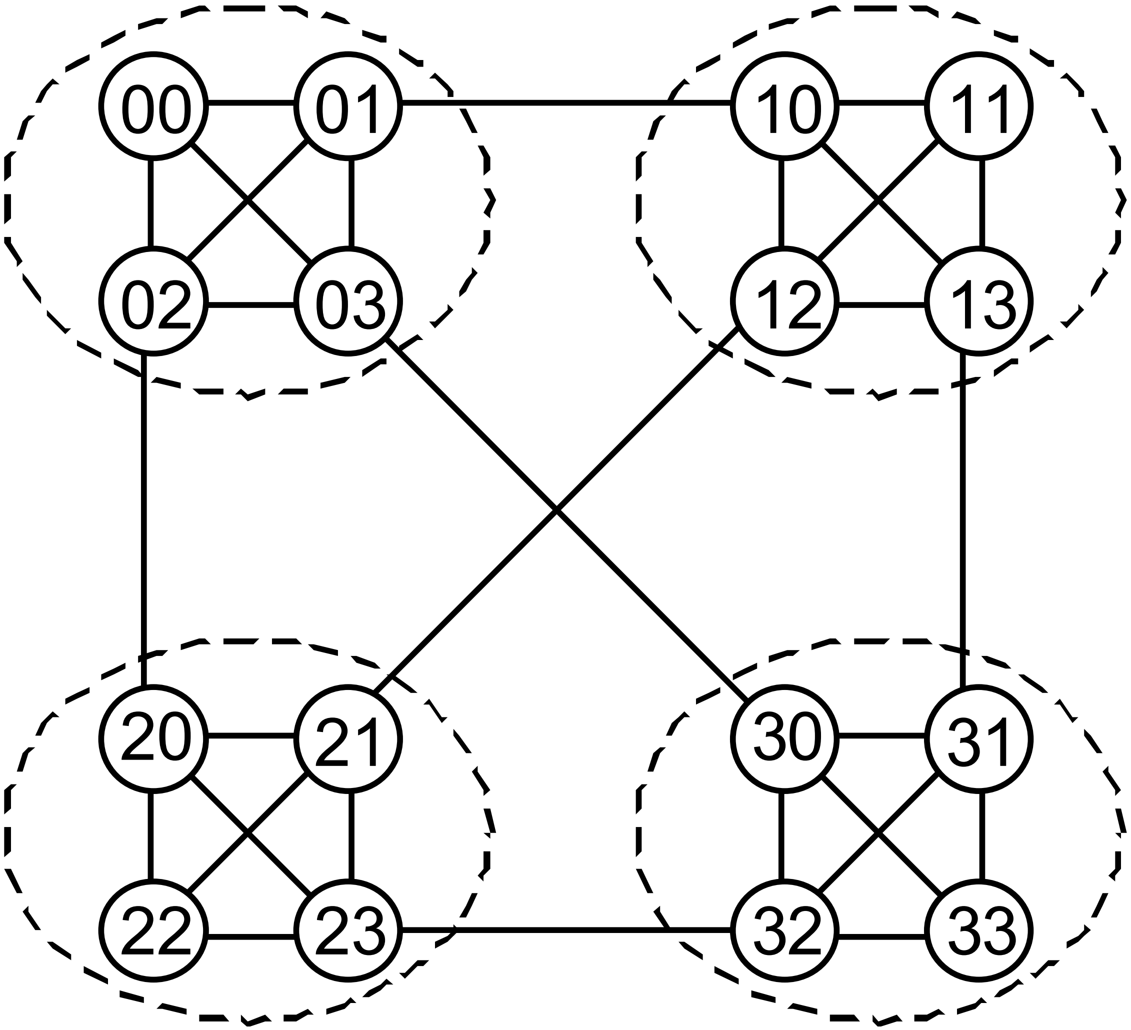
$D = 2 (\log m)^{1/2} + 1 = 1.414(\log m^2)^{1/2} + 1$

Unfortunately, B is the most important case for massive parallelism

C. Superlogarithmic-diameter networks

Similar analysis shows good diameter scaling

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

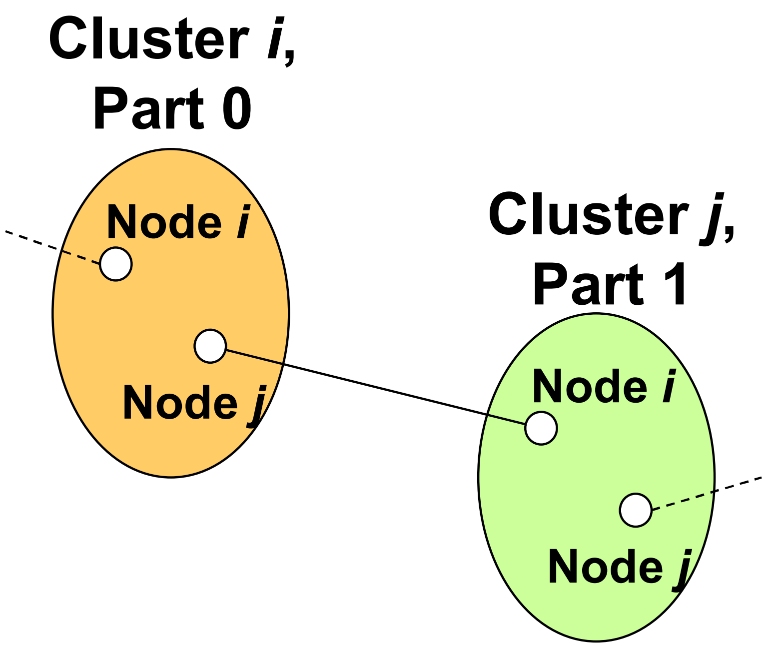
Wide diameter

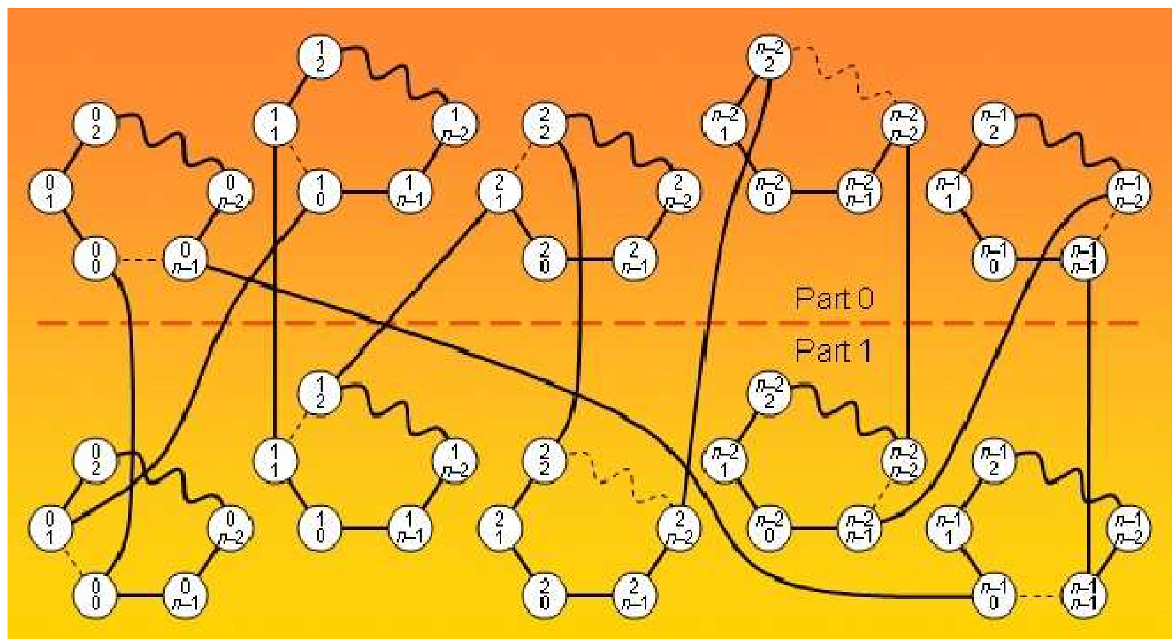
At most *D*(Sw(*G*)) + 4

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)

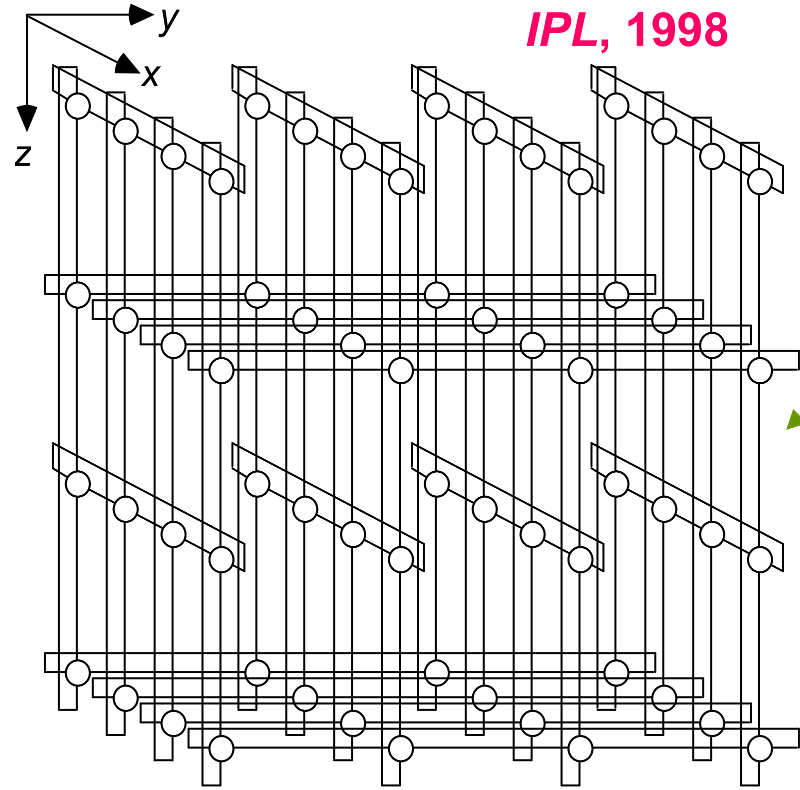




Pruning of Interconnection Networks

3D torus pruned along *Z*

IPL, 1998

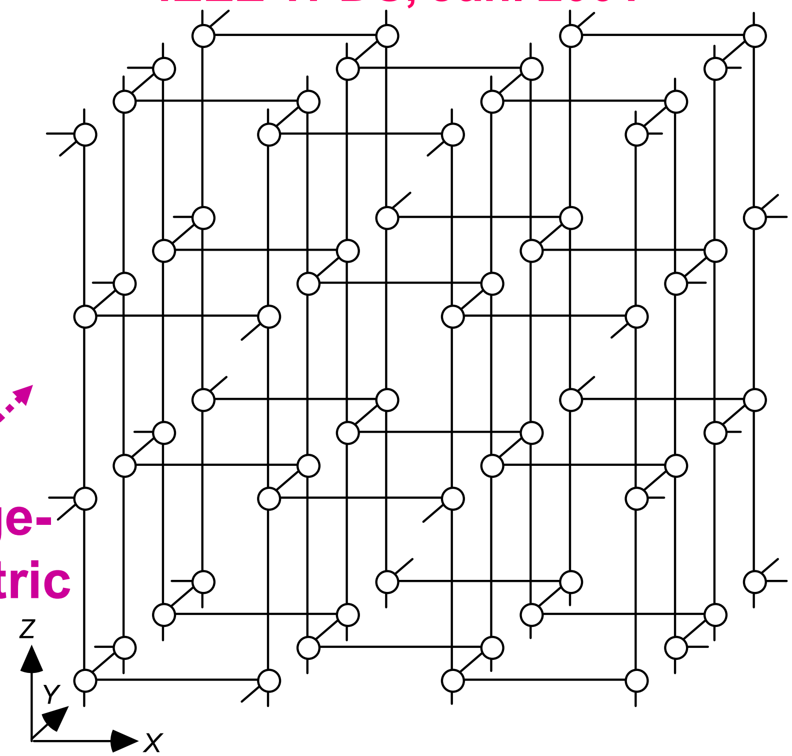


Cayley Graph and edge-symmetric

Diamond net = pruned torus

IEEE TPDS, Jan. 2001

Not edge-symmetric



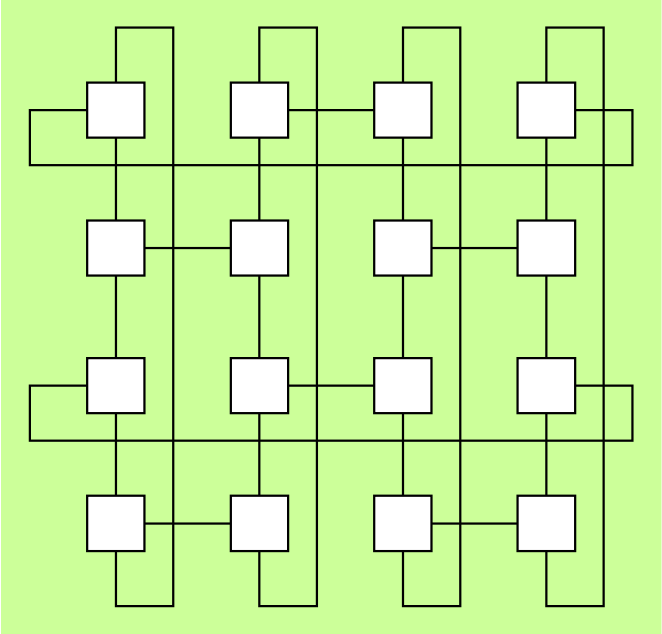
Must have simple and elegant pruning rules to ensure:

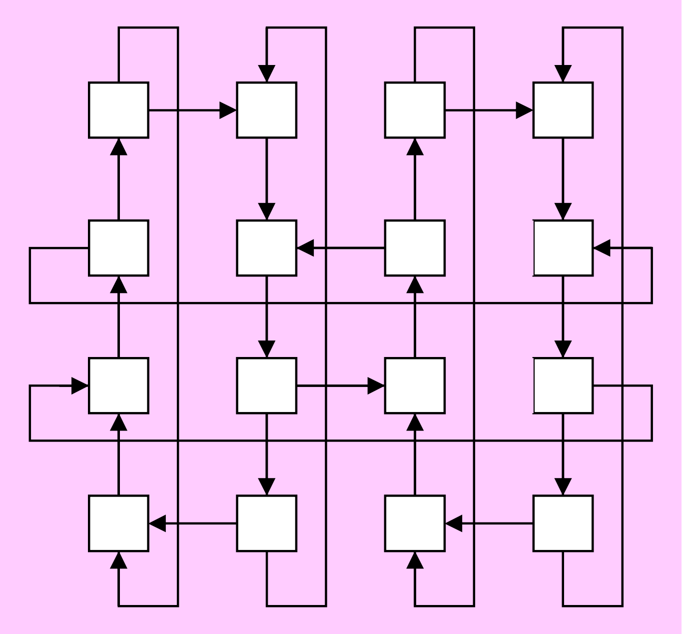
- **Efficient point-to-point and collective communication**
- **Symmetry, leading to “blandness” and balanced traffic**

Pruned Network Robustness

Robustness is in general adversely affected when a network is pruned
Systematic pruning ensures max 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

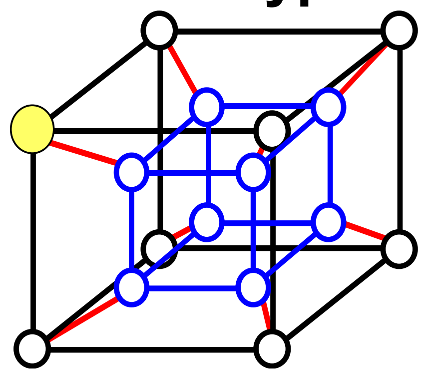




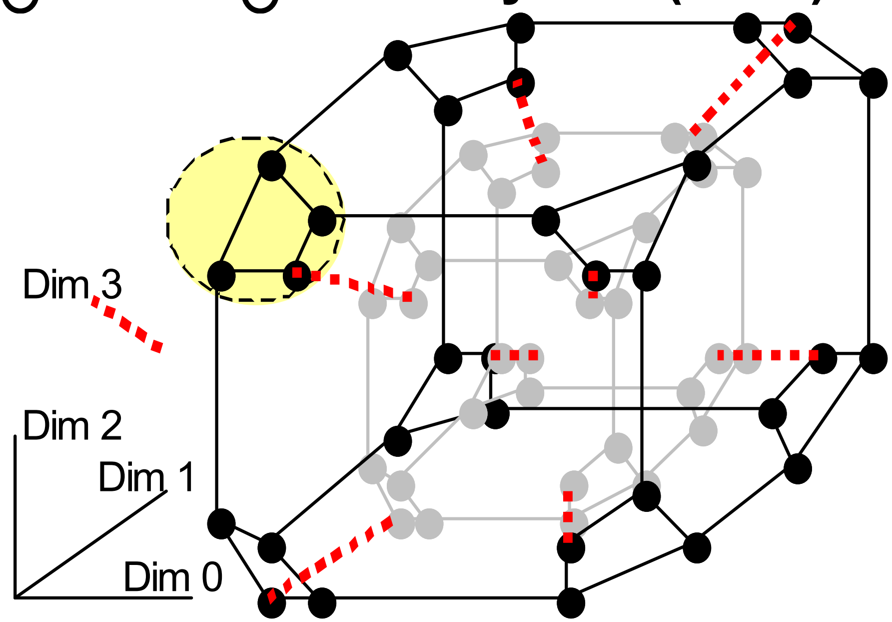
We have devised pruning schemes for a wide variety of networks and proven resulting networks to be robust & efficient algorithmically

Recursive Substitution

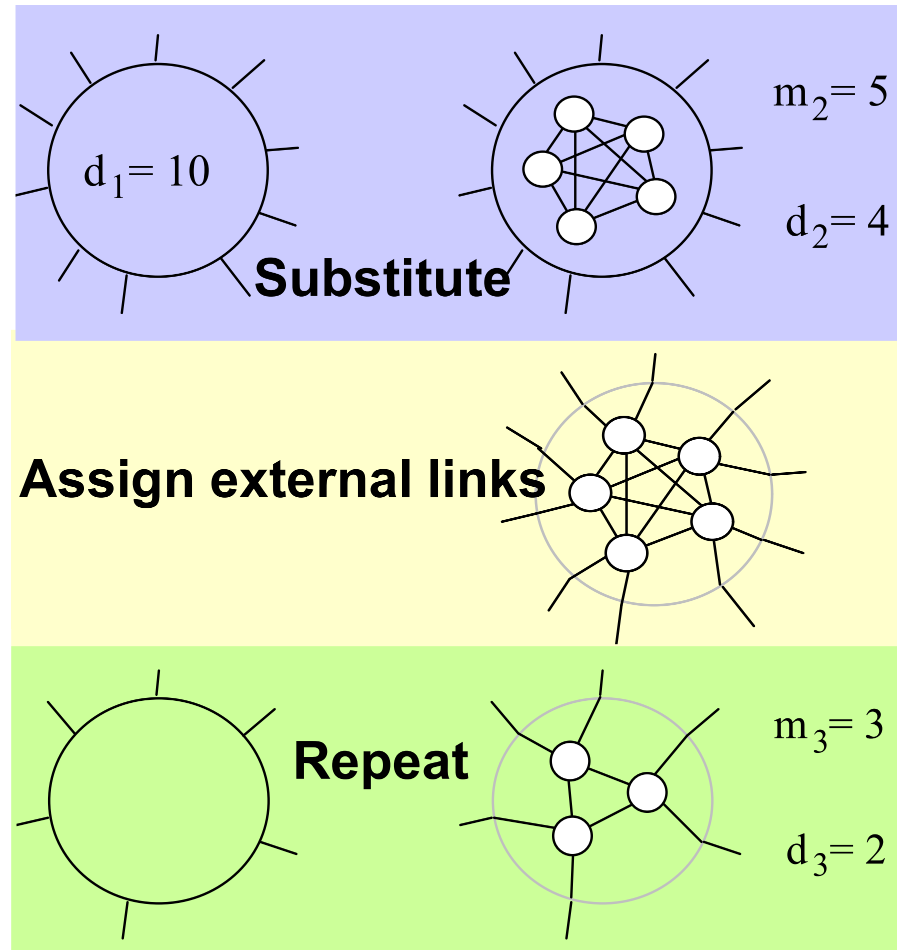
16-node hypercube



64-node cube-connected cycles (CCC)

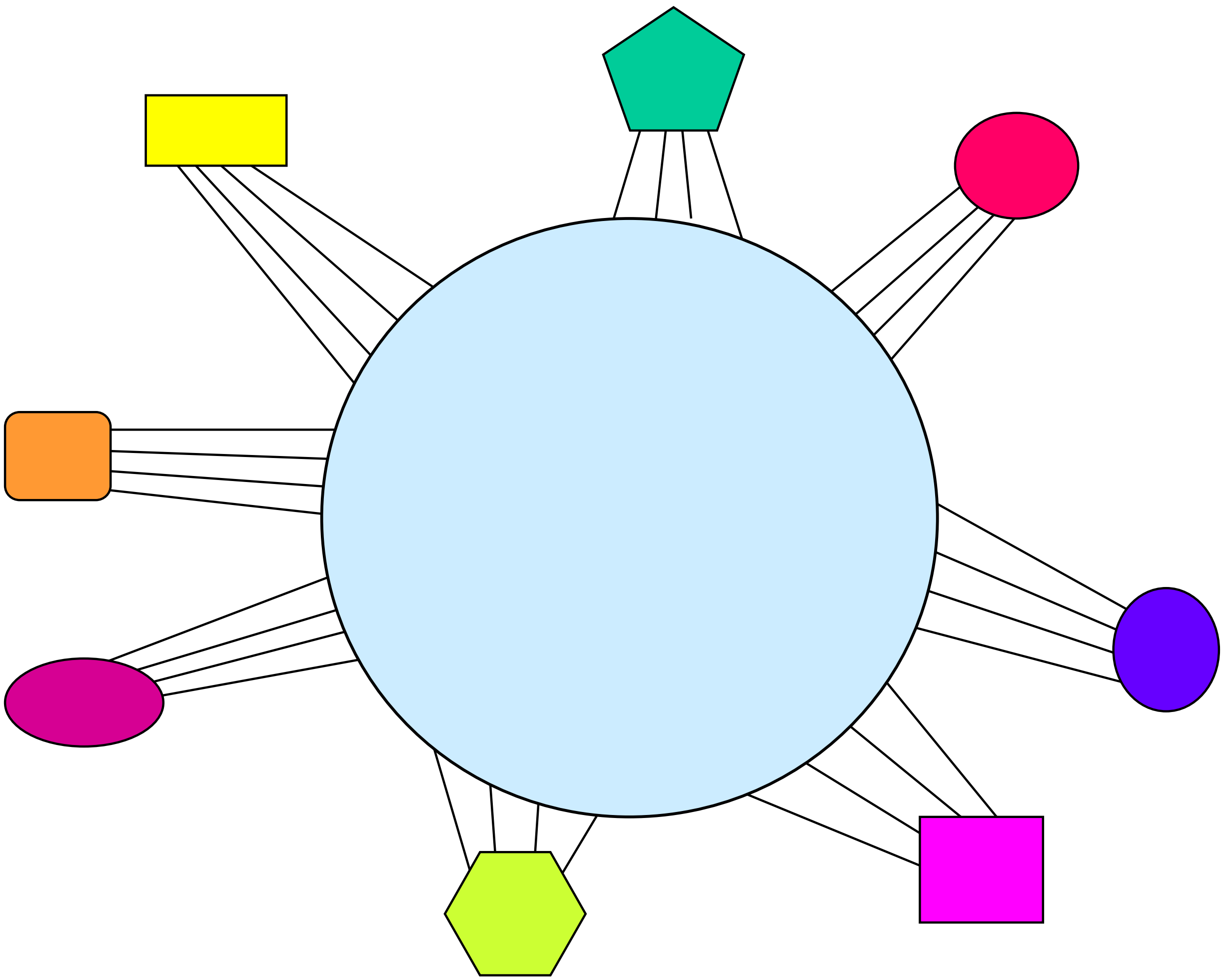


The general approach



Symmetry as a Desirable Property

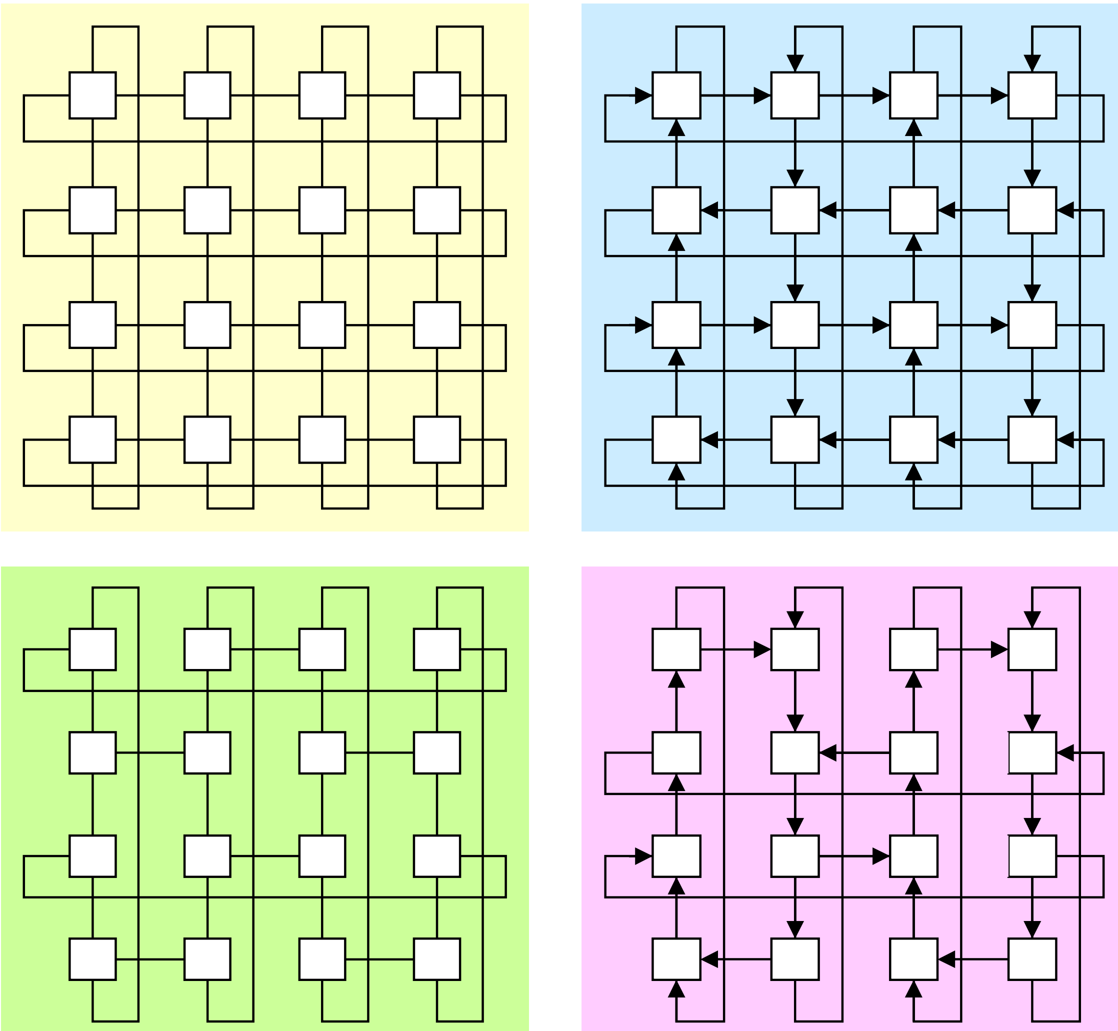
A degree-4 network



- **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**

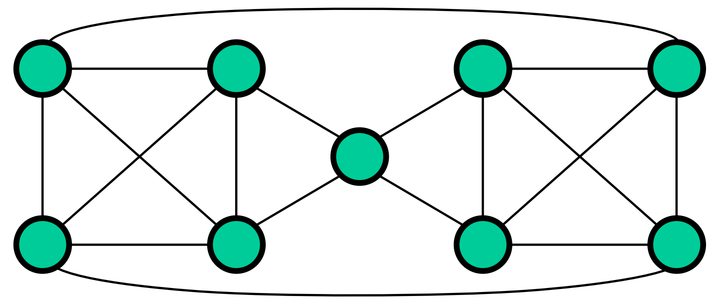
We need to prove a particular topological or routing property for only one node

A Necessary Condition for Symmetry



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

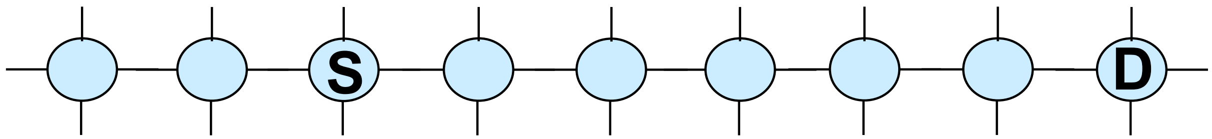
An asymmetric network with uniform node degree

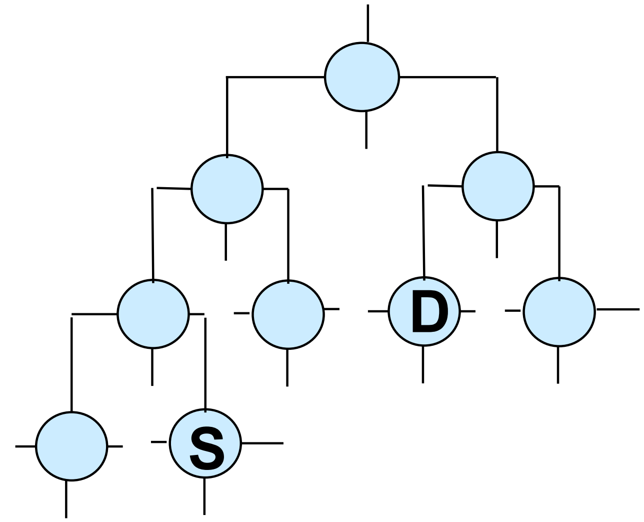


Uniform node degree is necessary but not sufficient for symmetry

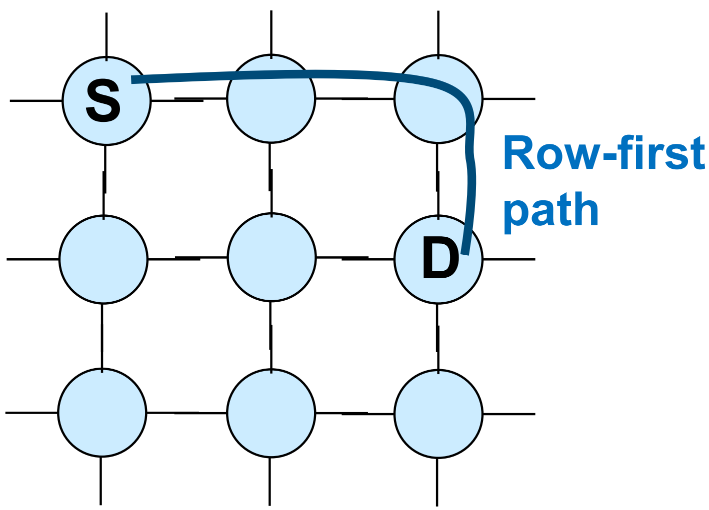
Oblivious vs. Adaptive Routing

When paths are unique, we have no choice

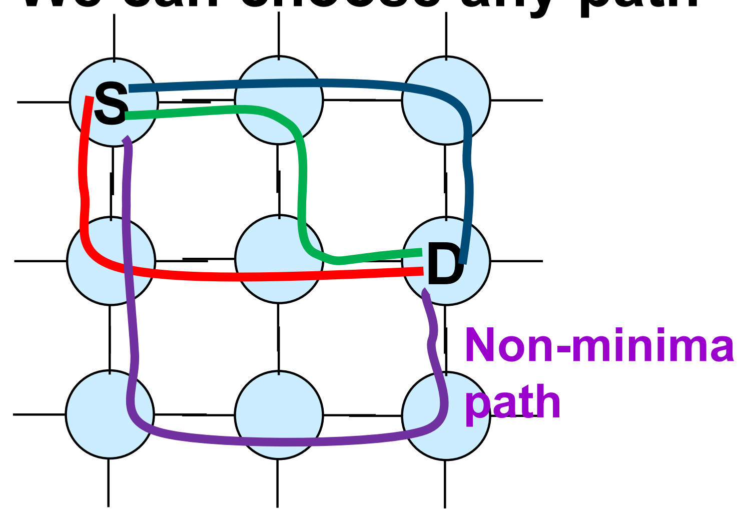




Oblivious routing:
Path is pre-determined



Adaptive routing:
We can choose any path



Analogy for Adaptive Routing

Graph: Models an interconnection network (nodes & links)
Floorplan: Represents a building floor (rooms & pathways)

Pre-planned escape route: Designed and posted for occupants
Adaptive route: Computed from floorplan & flame/smoke locations

The Bottom Line

Interconnection networks: Key parts of parallel computers and data centers (perhaps more important than the nodes)

Motivation for inventing new interconnection architectures

Hard to convince designers to abandon proven schemes (due to design turnaround-time & maintenance benefits)

Economy of scale favors existing, off-the-shelf technology

Top-of-the-line systems more likely to use new networks (prestige of being at the top motivates greater investment)

Inventing new networks is like inventing new tools (sometimes they catch on; otherwise, they are added to the toolbox in hopes of being used in future)

Future Work: 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?

Future Work: 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?

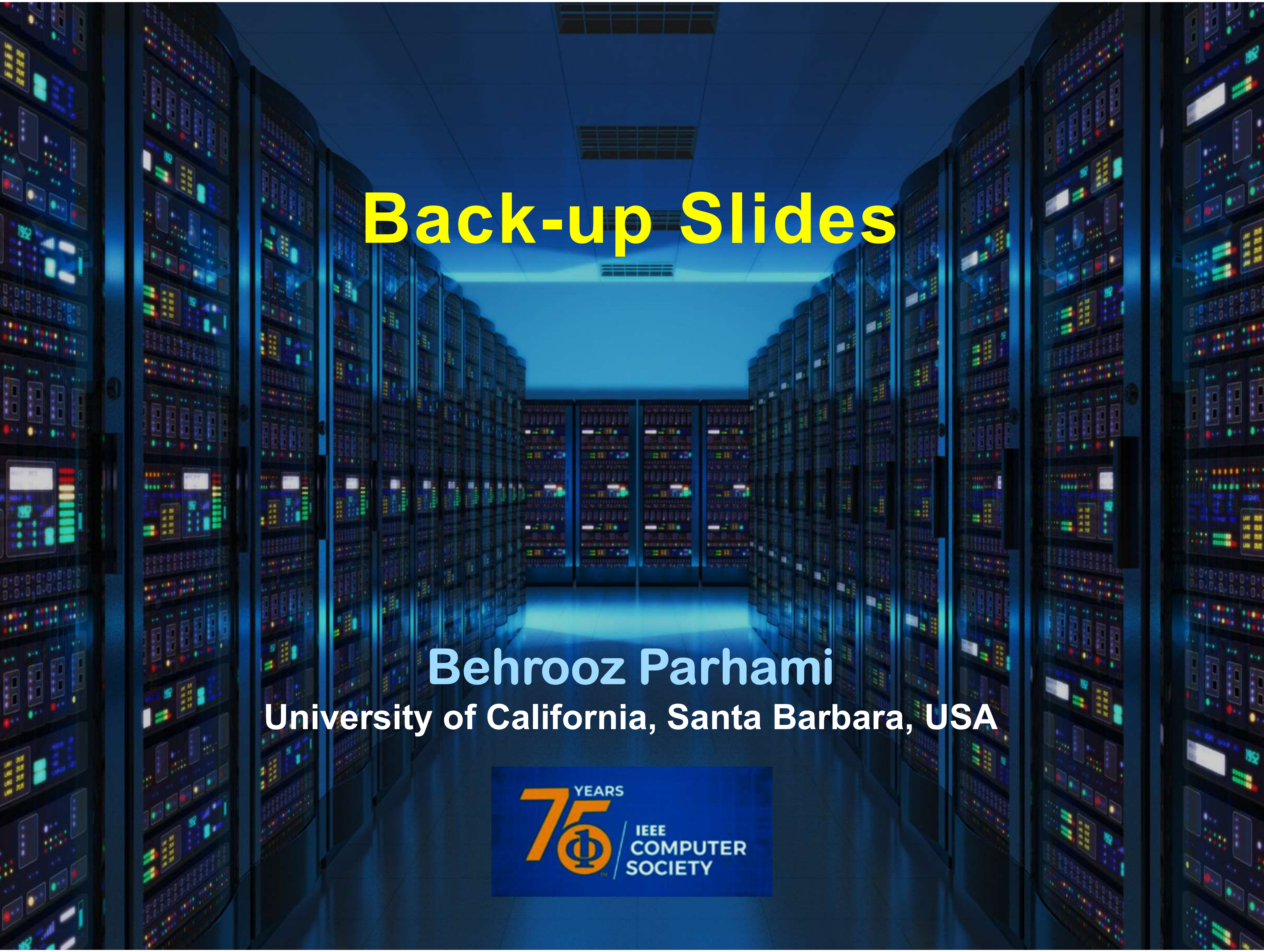
The (d, D) graph problem is very difficult
Answers are known only for certain values of d and D

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)?

Questions or Comments?

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

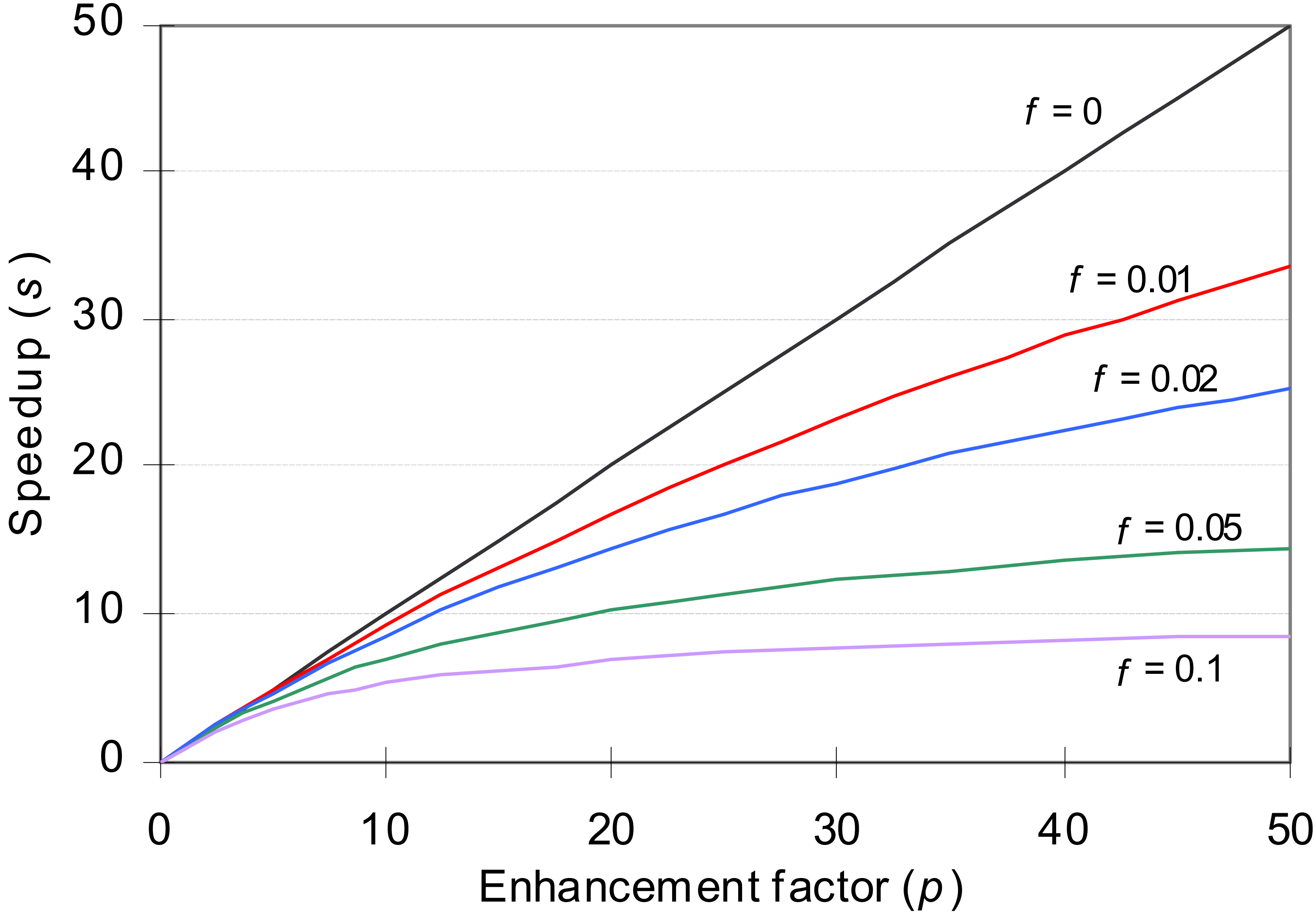


Back-up Slides

Behrooz Parhami

University of California, Santa Barbara, USA

Amdahl's Law



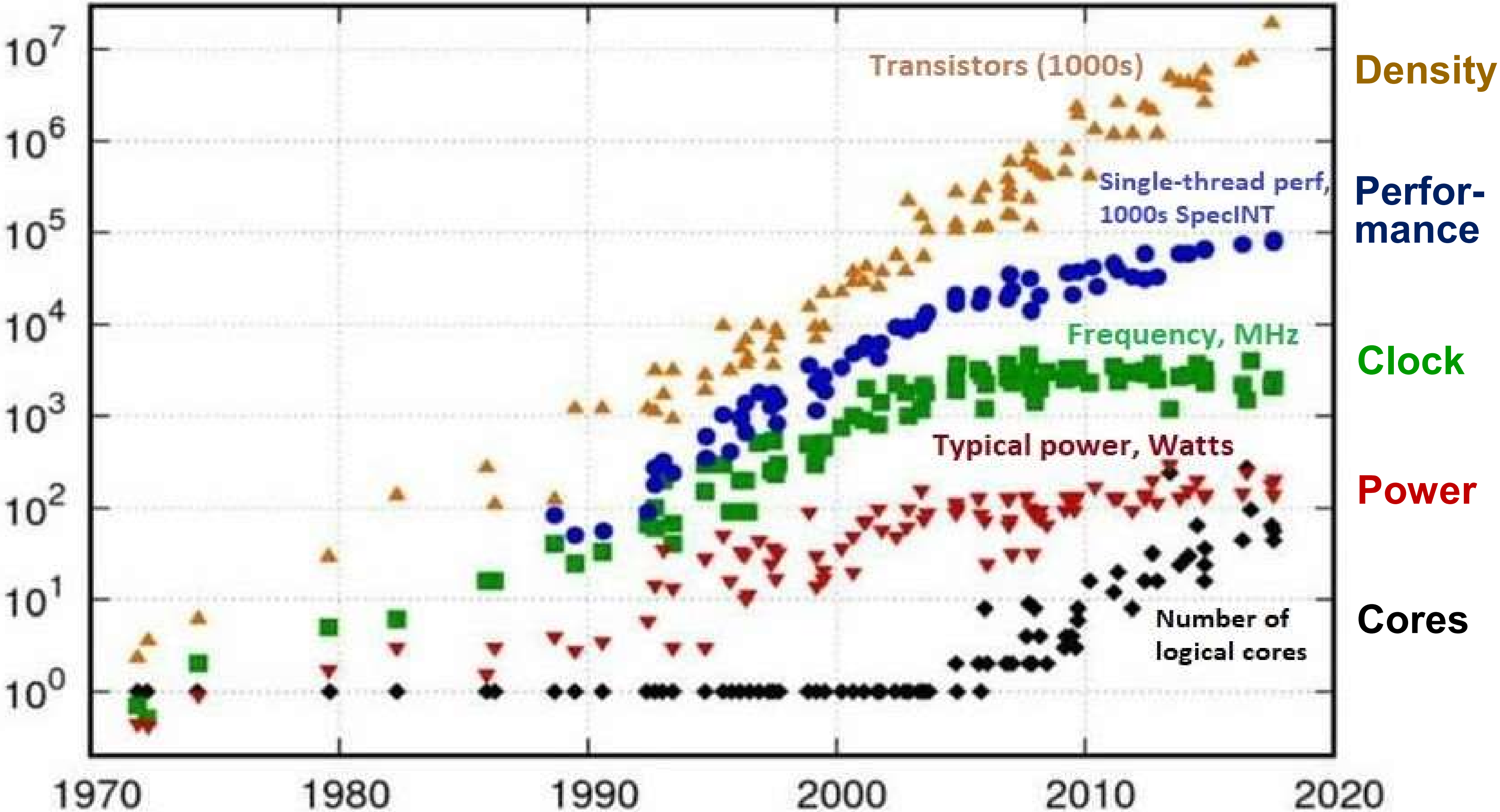
f **= fraction unaffected**

p **= speedup of the rest**

$$s = \frac{1}{f + (1 - f)/p}$$

$$\le min(p, 1/f)$$

Trends in Processor Chip Density, Performance, Clock Speed, Power, and Number of Cores



Original data up to 2010 collected/plotted by M. Horowitz et al.; Data for 2010-2017 extension collected by K. Rupp

The Flynn/Johnson Classification

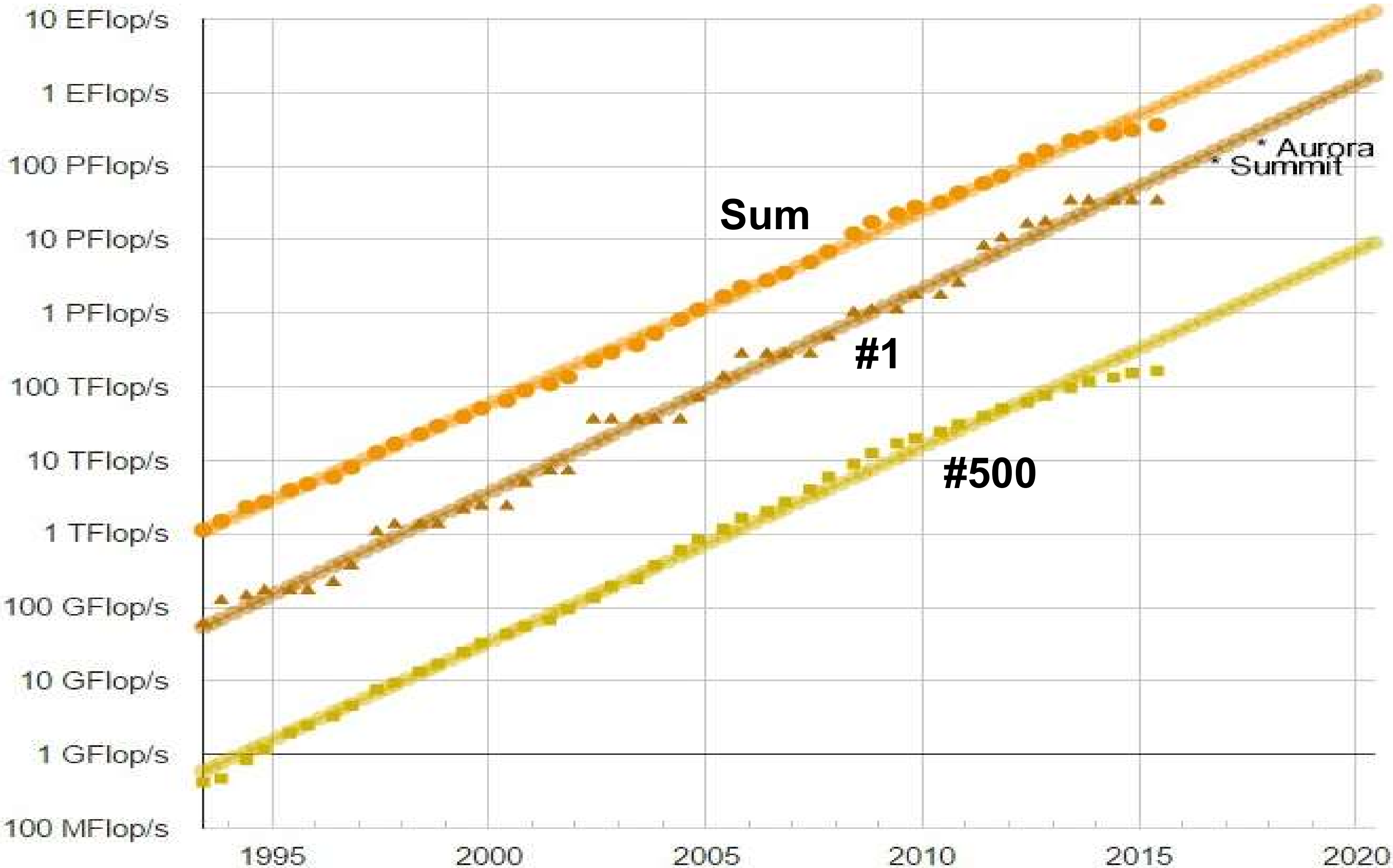
Flynn's categories

	Single data stream	Multiple data streams
Single instr stream	SISD Uniprocessors	SIMD Array or vector processors
Multiple instr streams	MISD Rarely used	MIMD Multiproc's or multicomputers

Johnson's expansion

	Shared variables	Message passing
Global memory	GMSV Shared-memory multiprocessors	GMMP Rarely used
Distributed memory	DMSV Distributed shared memory	DMMP Distrib-memory multicomputers

Top 500 Supercomputers in the World



The Quest for Higher Performance

Top-Five Supercomputers in November 2020 (http://www.top500.org)

Rank (previous)	Rmax Rpeak (PFLOPS)	Name	Model	CPU cores	Accelerator (e.g. GPU) cores	Interconnect	Manufacturer
1	442.010 537.212	Fugaku	Supercomputer Fugaku	158,976 × 48 A64FX @2.2 GHz	0	Tofu interconnect D	Fujitsu
2▼ (1)	148.600 200.795	Summit	IBM Power System AC922	9,216 × 22 POWER9 @3.07 GHz	27,648 × 80 Tesla V100	InfiniBand EDR	IBM
3▼ (2)	94.640 125.712	Sierra	IBM Power System S922LC	8,640 × 22 POWER9 @3.1 GHz	17,280 × 80 Tesla V100	InfiniBand EDR	IBM
4▼ (3)	93.015 125.436	Sunway TaihuLight	Sunway MPP	40,960 × 260 SW26010 @1.45 GHz	0	Sunway[26]	NRCPC
5▲ (7)	63.460 79.215	Selene	Nvidia	1,120 × 64 Epyc 7742 @2.25 GHz	4,480 × 108 Ampere A100	Mellanox HDR Infiniband	Nvidia

The Quest for Higher Performance

June 2022 update (http://www.top500.org)

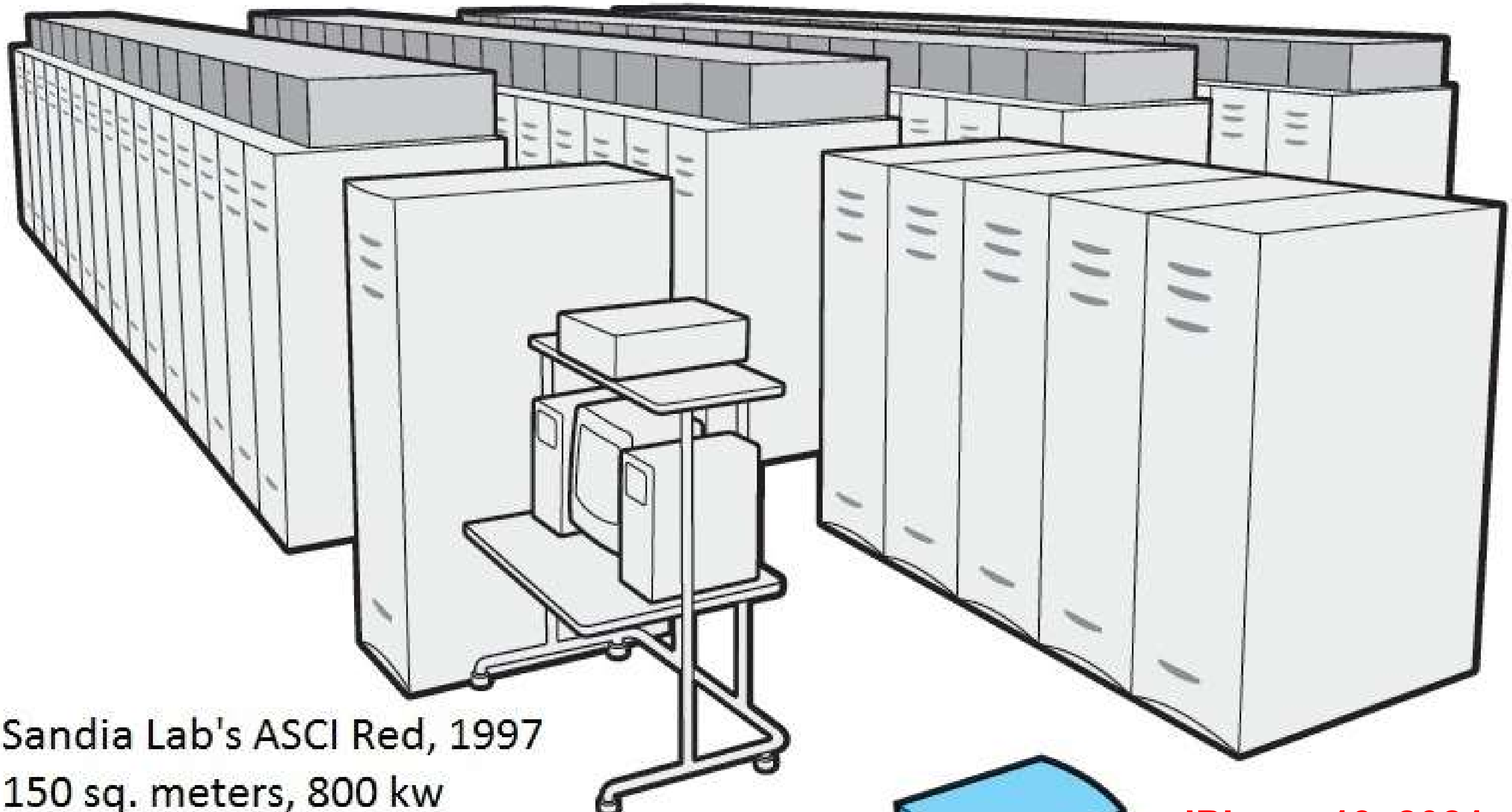
Top Supercomputer: 1+ exaflops performance

- Frontier system at US Oak Ridge National Lab.
- Based on the latest HPE Cray EX235a architecture
- Equipped with AMD EPYC 64C 2 GHz processors
- Number of Cores ~8.7 million
- Power efficiency rating of ~52 gigaflops/W

Top "Green" Supercomputer: ~20 petaflops

- Frontier Test & Development System at ORNL
- A subset of the top supercomputer above
- Number of cores ~120K
- Power efficiency rating of ~63 gigaflops/W
- The top supercomputer above is #2 on the Green500 list

The Shrinking Supercomputer



Sandia Lab's ASCI Red, 1997
150 sq. meters, 800 kw

Both perform at ~2 TFLOPS

Sony Playstation, 2006
0.08 sq. meter, < 0.2 kw

iPhone 13, 2021
6 CPU + 5 GPU + 16 NN
0.01 sq. meter, < 0.5 w

Warehouse-Sized Data Centers

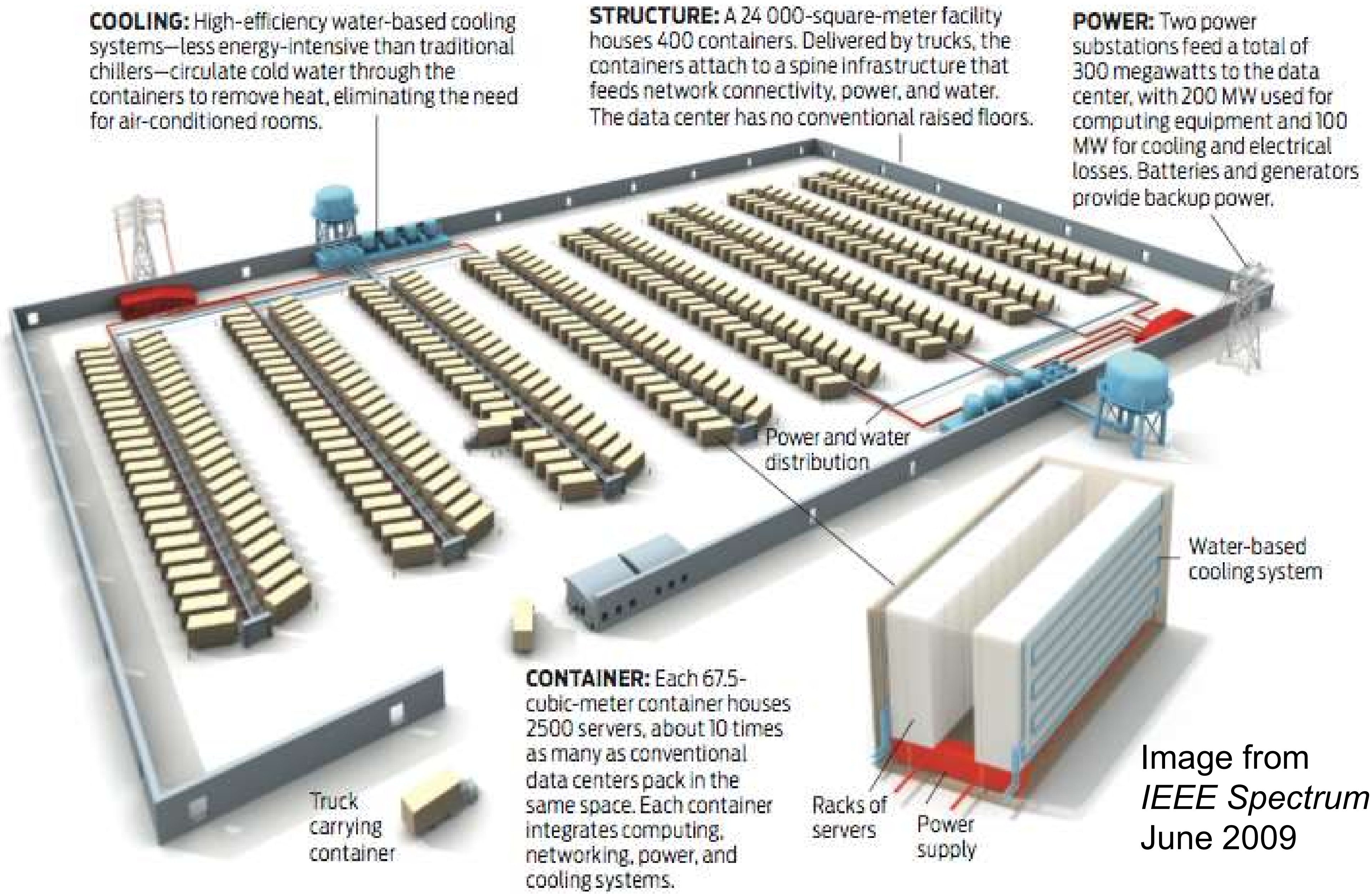


Image from *IEEE Spectrum*, June 2009

Computing in the Cloud

Computational resources, both hardware and software, are provided by, and managed within, the cloud

Users pay a fee for access

Managing / upgrading is much more efficient in large, centralized facilities (warehouse-sized data centers or server farms)

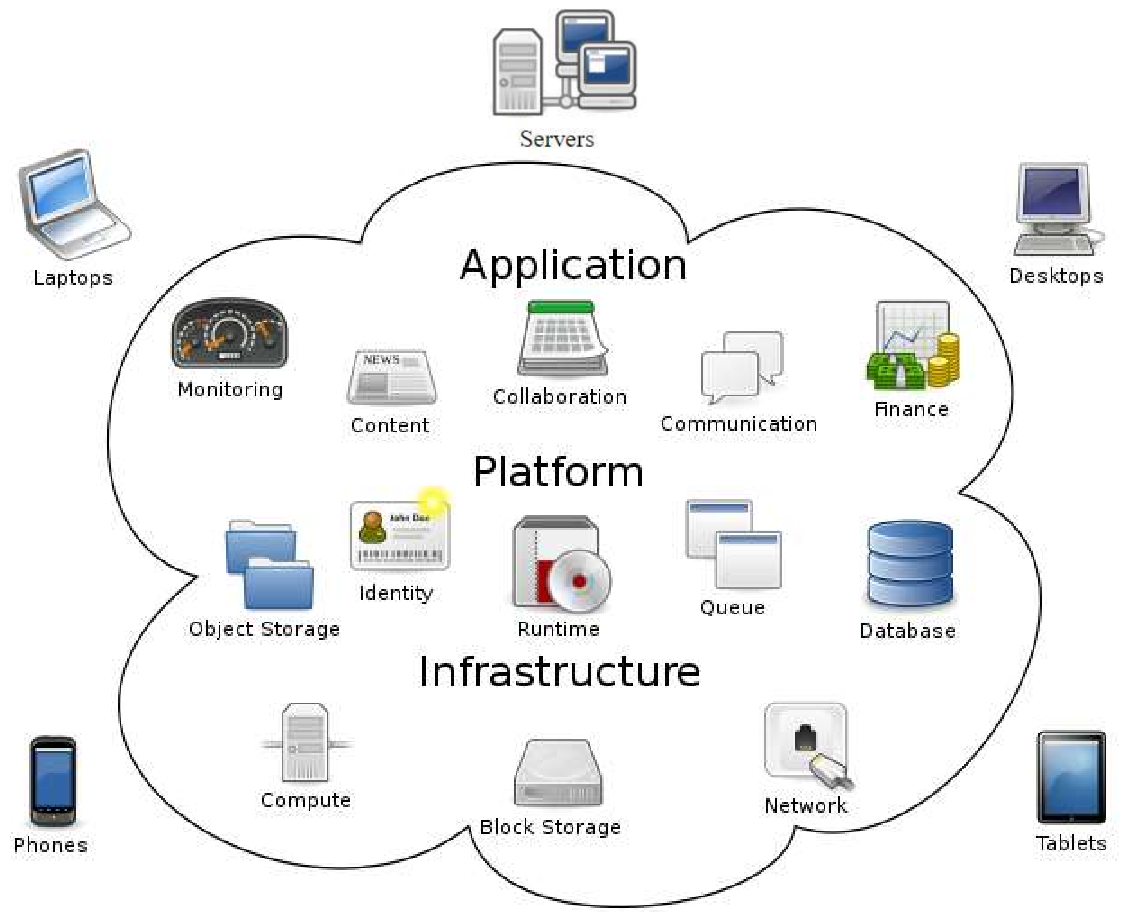


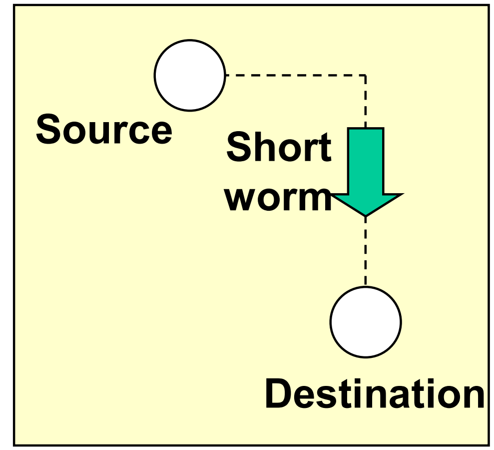
Image from Wikipedia: Created by Sam Johnston

This is a natural continuation of the outsourcing trend for special services, so that companies can focus their energies on their main business

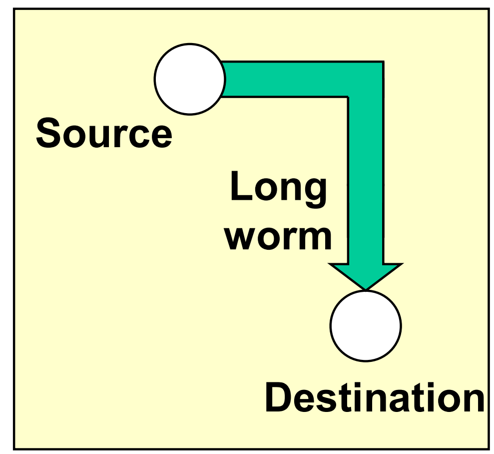
Importance of Diameter

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