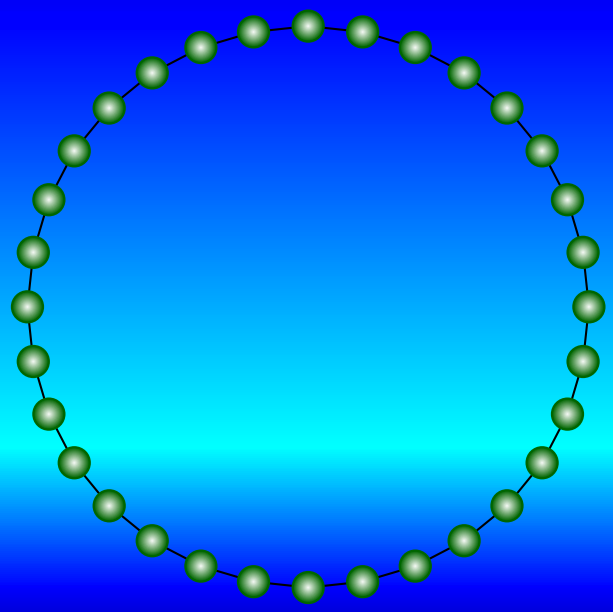
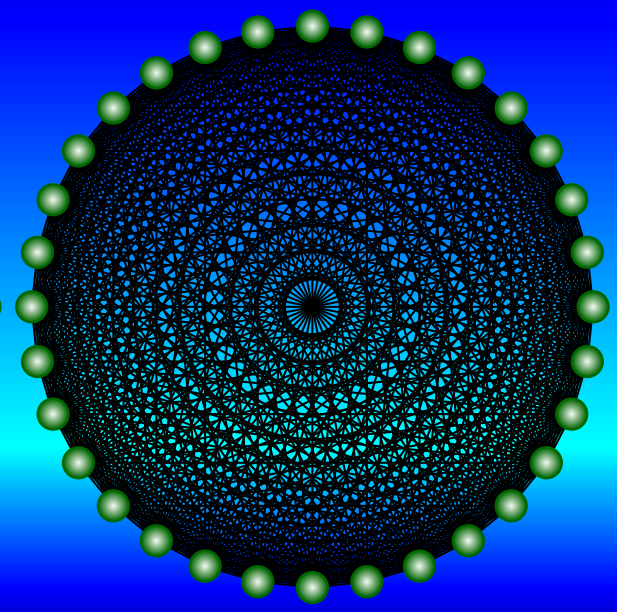
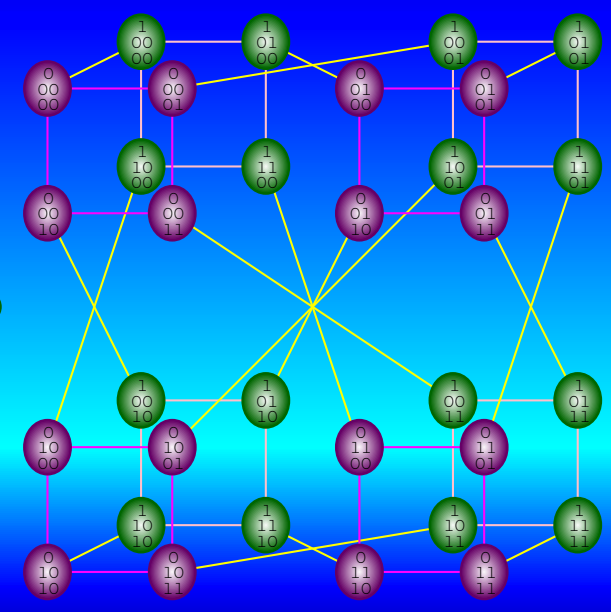
Cost/Performance Efficient Interconnection Networks for Supercomputers



Low performance



High cost



High performance at low cost

Yamin Li, Graduate School of CIS, Hosei University, yamin@hosei.ac.jp

Outline

- Top 10 supercomputers
- Multiprocessor shared-memory parallel systems
- Multicomputer distributed systems
- Problems of interconnection network
- Low-degree and short-diameter topologies
  - Dual-Cube
  - Metacube and KMS-cube
  - RDN — Recursive Dual-Net
- Dynamic interconnection network — MiKANT
- Summary and exercise

Top 10 Supercomputers

TOP500 Supercomputer Sites: https://www.top500.org/

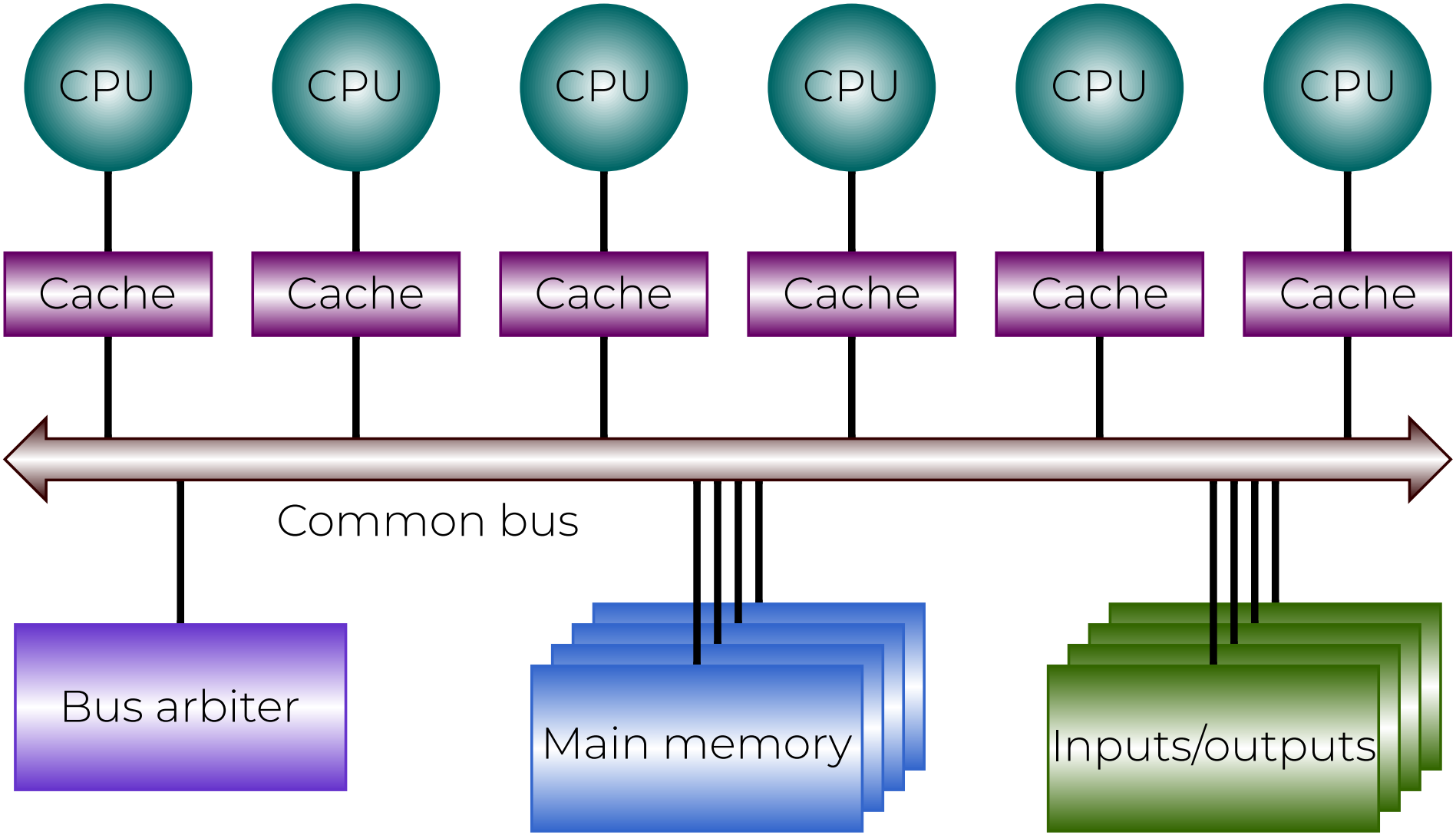
TOP 10 Supercomputers in June 2023

Rank	System	Country	Maker	Cores	TFlop/s
1	Frontier	USA	HPE	8,699,904	1,194.00
2	Fugaku	Japan	Fujitsu	7,630,848	442.01
3	LUMI	Finland	HPE	2,220,288	309.10
4	Leonardo	Italy	Atos	1,824,768	238.70
5	Summit	USA	IBM	2,414,592	148.60
6	Sierra	USA	IBM	1,572,480	94.64
7	Sunway TaihuLight	China	NRCPC	10,649,600	93.01
8	Perlmutter	USA	HPE	761,856	70.87
9	Selene	USA	Nvidia	555,520	63.46
10	Tianhe-2A	China	NUDT	4,981,760	61.44

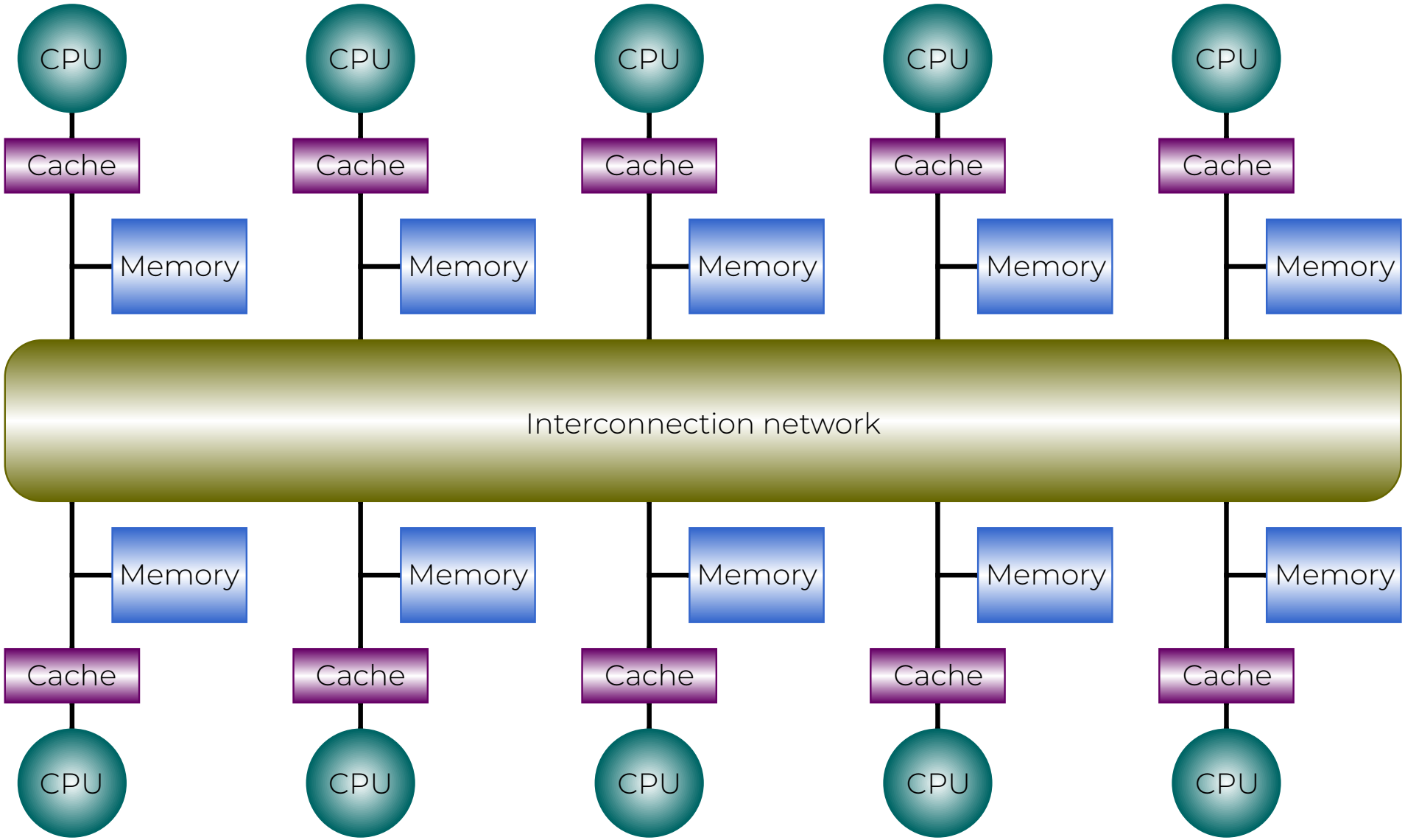
Parallel and Distributed Systems

- In general, parallel systems are referred to as shared-memory multiprocessors.
  - SMP — Symmetric multiprocessors
    - Uniform memory access (UMA)
    - Example: Servers
  - DSM — Distributed shared memory
    - Non-uniform memory access (NUMA)
    - Example: Supercomputers
- In contrast, distributed systems are referred to as message-passing multicomputers.
  - Cannot access other computer's memory directly
  - Example: Infrastructure of Cloud Computing

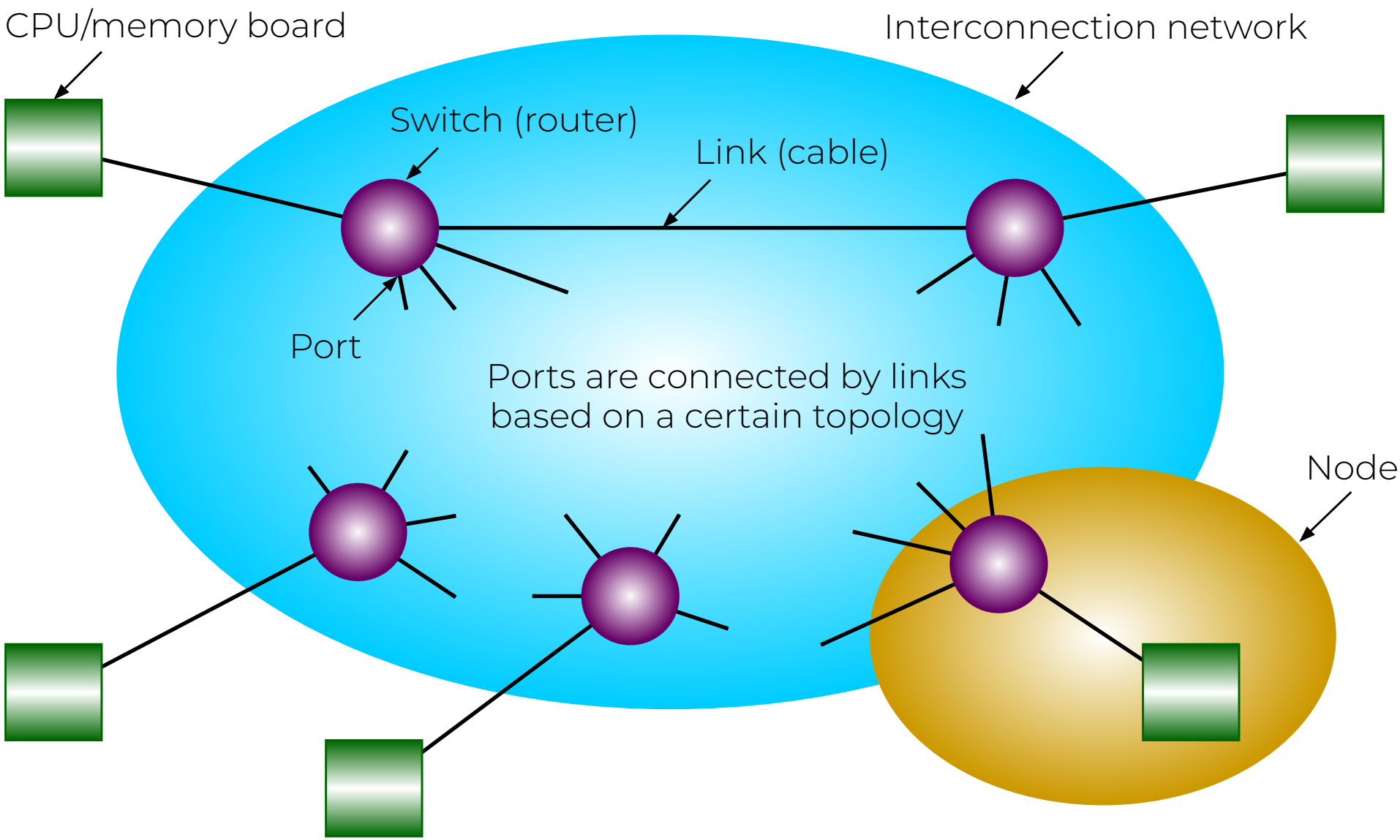
Symmetric Multiprocessors (Servers)



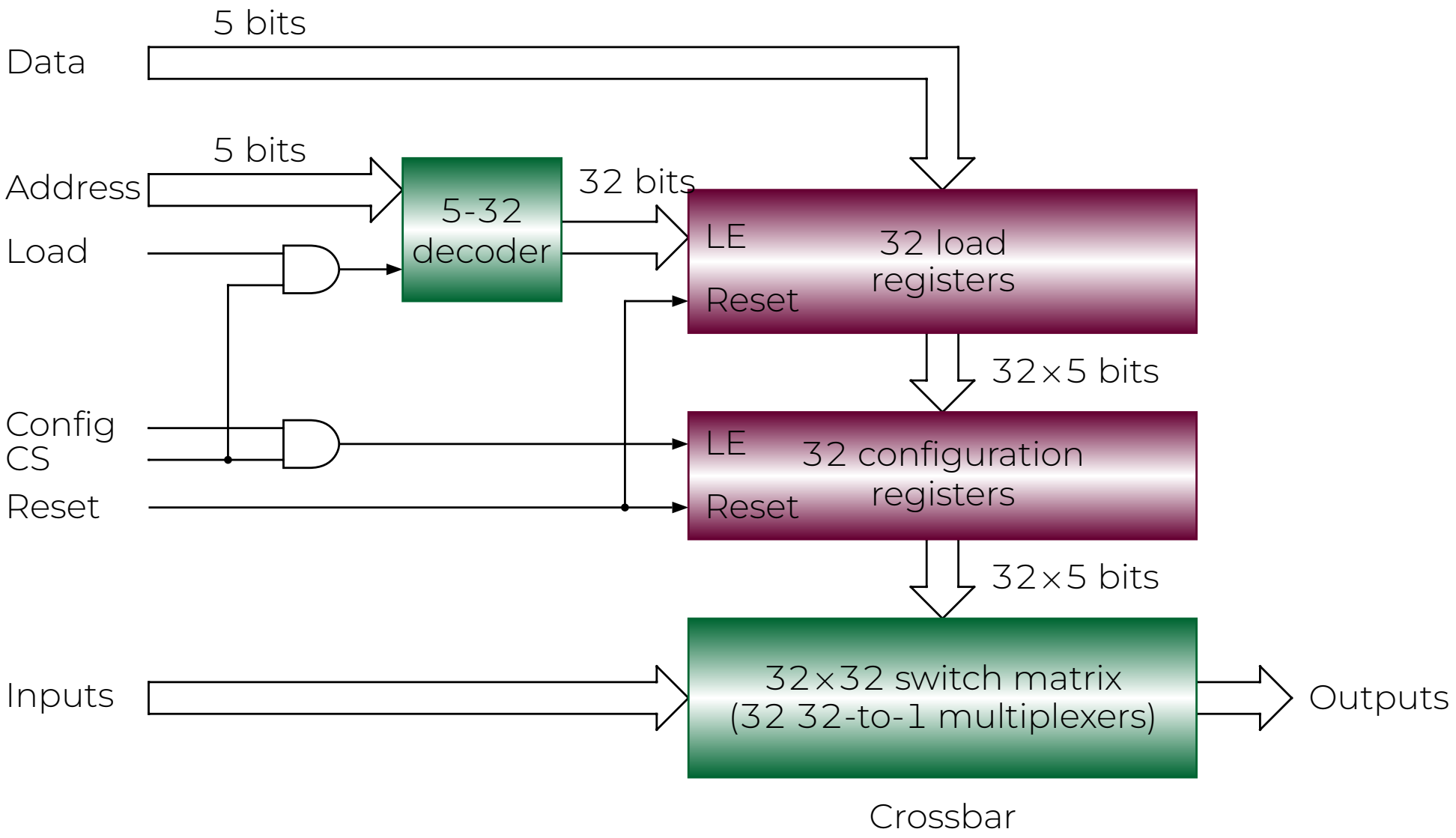
Distributed Shared Memory (Supercomputers)



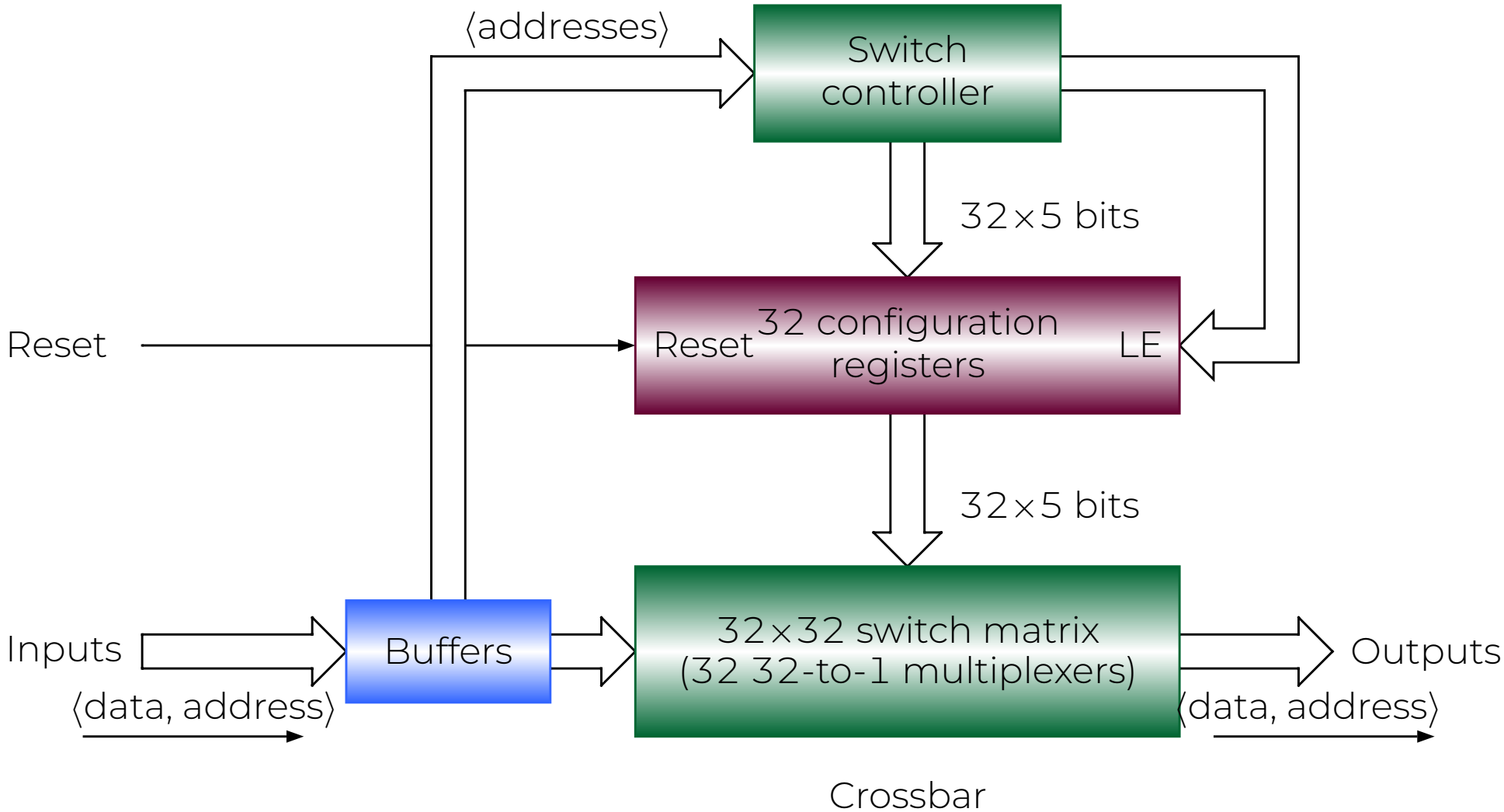
Interconnection Networks



An Implementation of Switch



A Switch with Self-Routing Function



It can support wormhole routing.

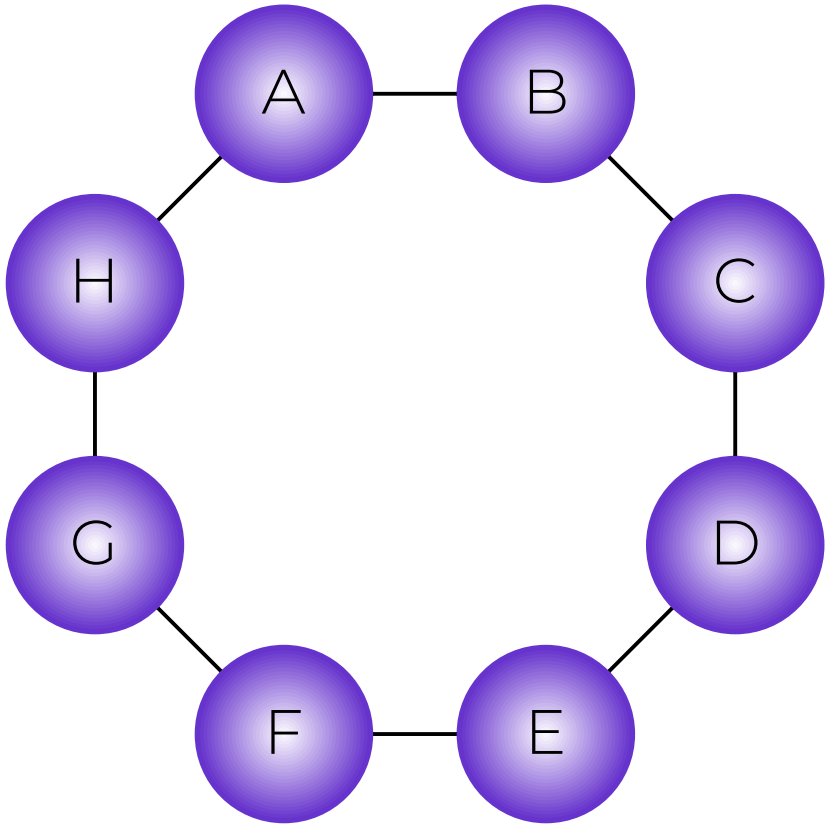
Properties of Interconnection Networks

- Degree
  - Two nodes are neighbors if there is a link connecting them.
  - The degree of a node is defined to be the number of its neighbors.
  - Affect hardware cost.
- Diameter
  - The diameter of a network is defined as the maximum of the shortest distances between any two nodes.
  - Affect communication time.

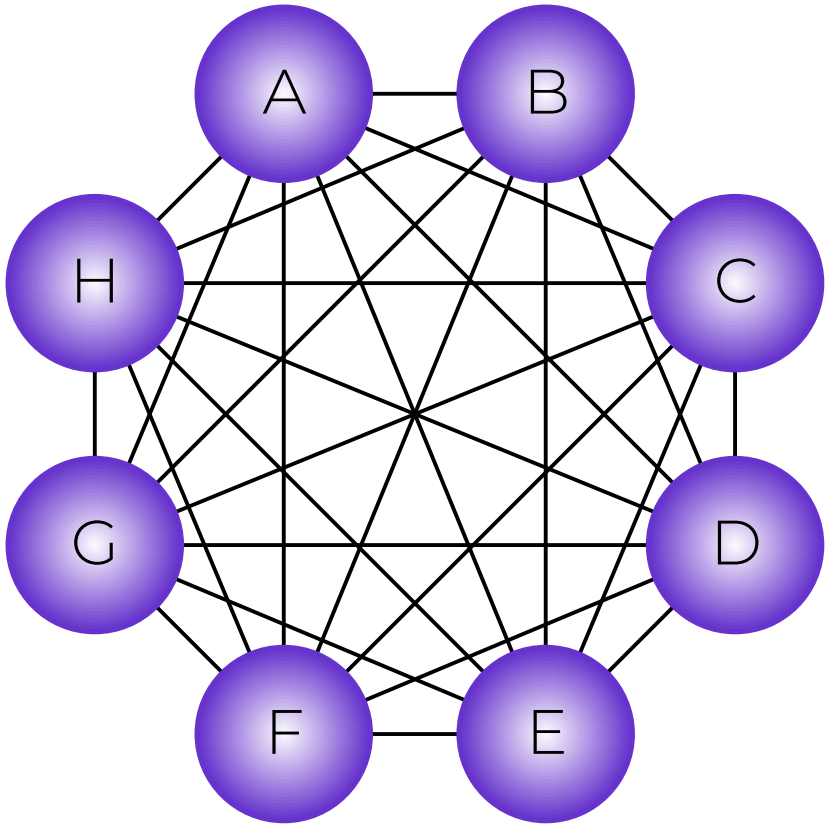
Popular Topologies

- Fat-tree
  - Almost all systems
- Torus
  - Fujitsu
  - IBM
  - HPE/Cray
- Dragonfly
  - HPE/Cray
- Hypercube
  - SGI
  - Intel

Ring and Completely Connected Networks



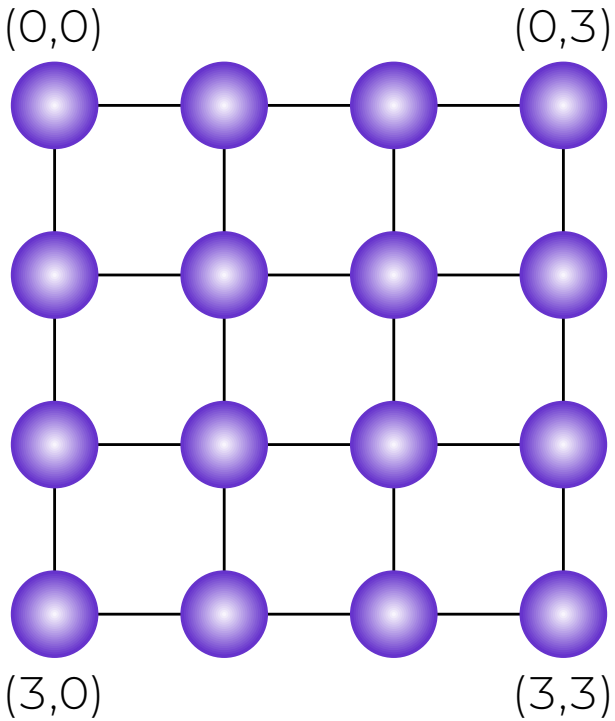
(a) Ring



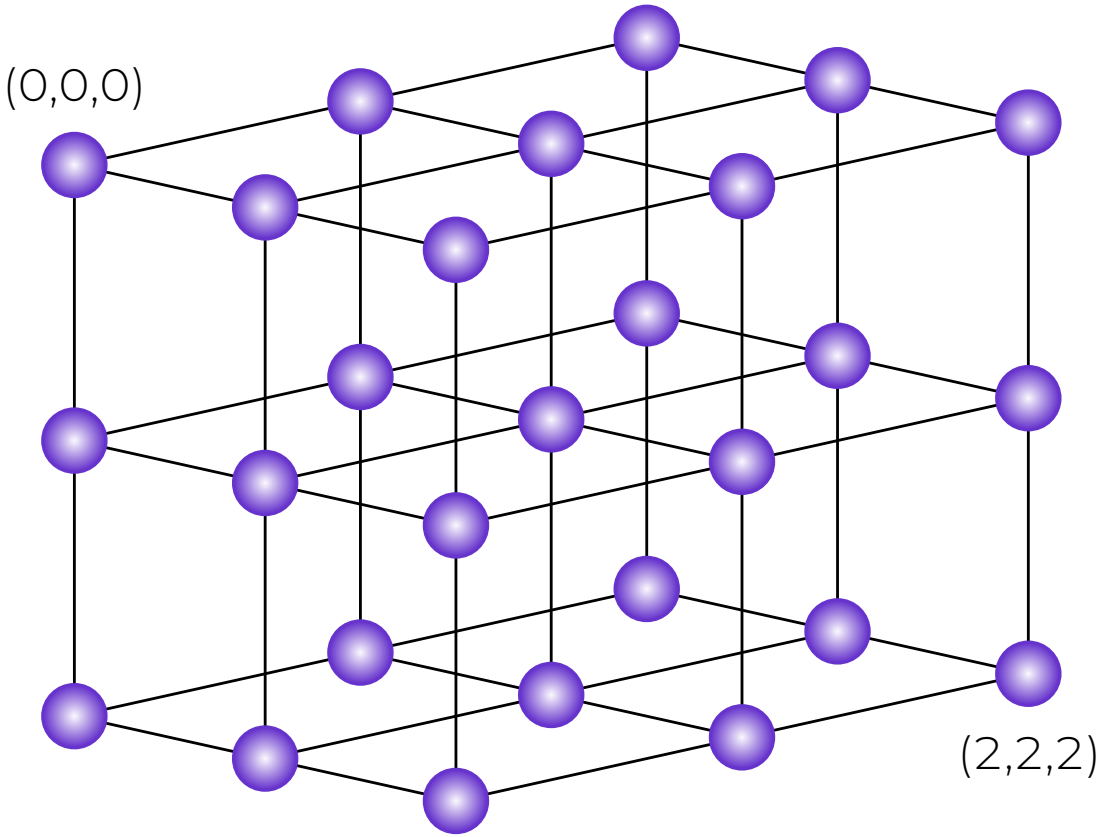
(b) Completely connected

Symmetric

Mesh Interconnection Networks



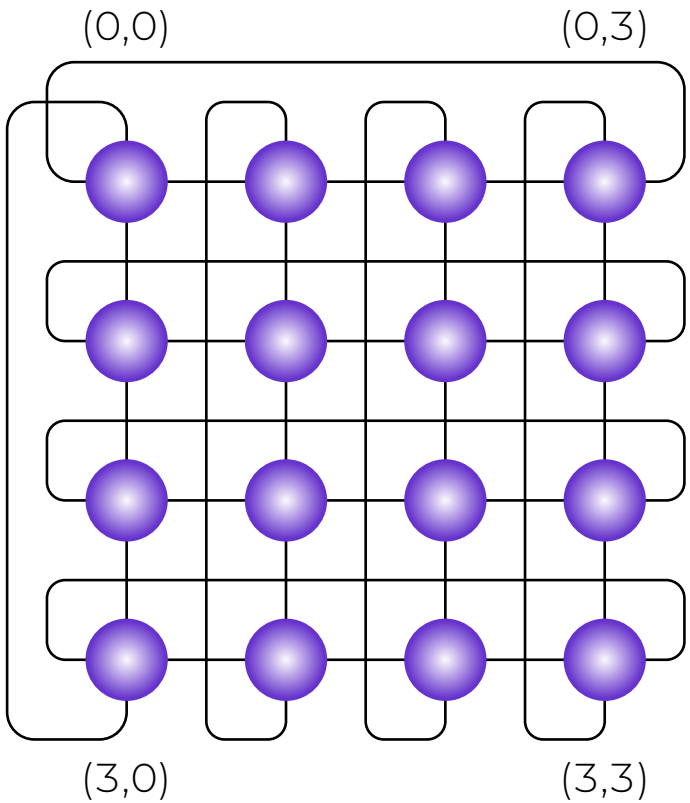
(a) 2D mesh



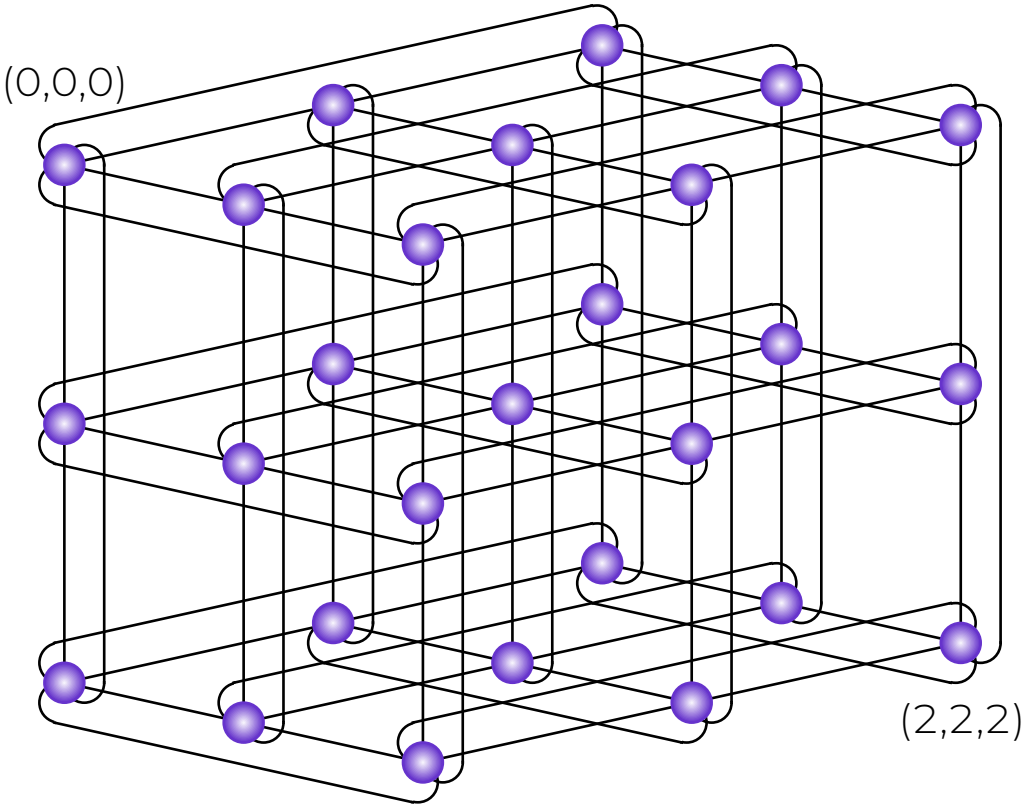
(b) 3D mesh

Not symmetric

Torus Interconnection Networks



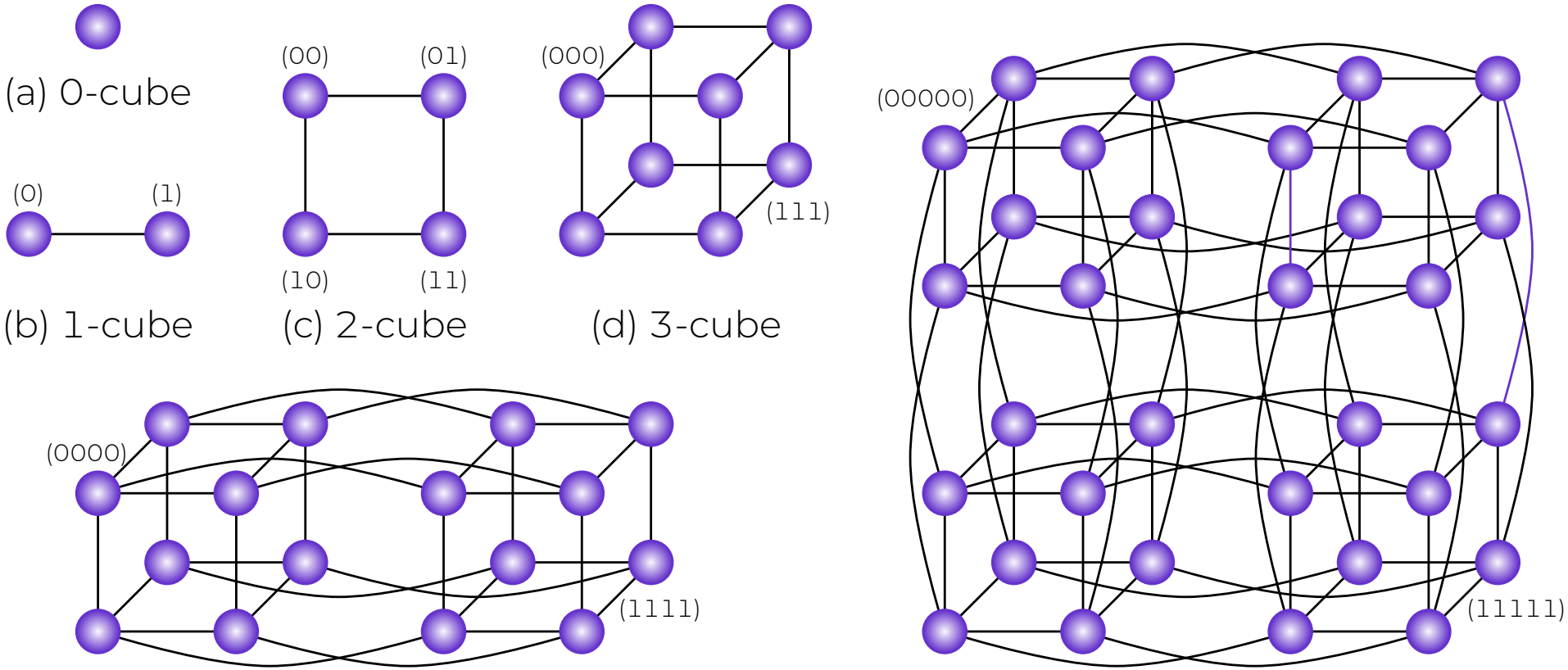
(a) 2D torus



(b) 3D torus

Symmetric

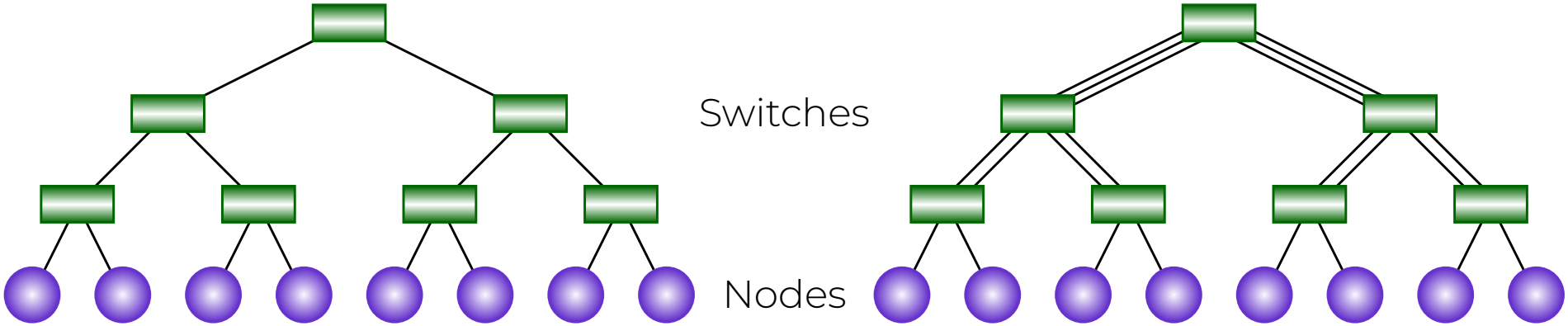
Hypercube Interconnection Networks



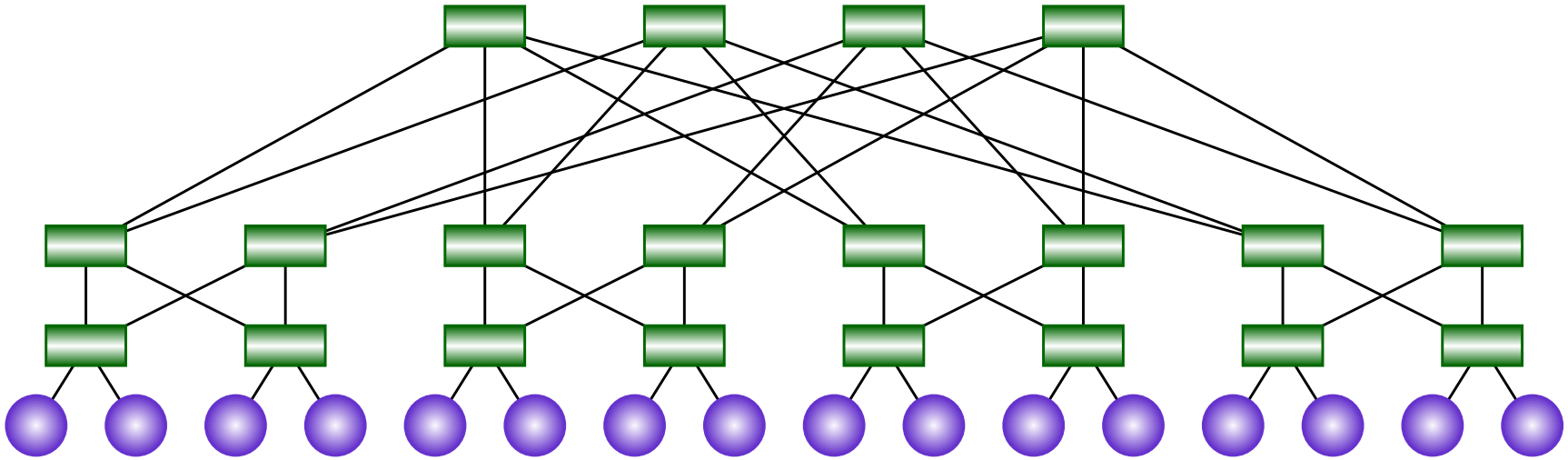
(a) 0-cube (b) 1-cube (c) 2-cube (d) 3-cube (e) 4-cube (f) 5-cube

Symmetric

Tree Interconnection Networks

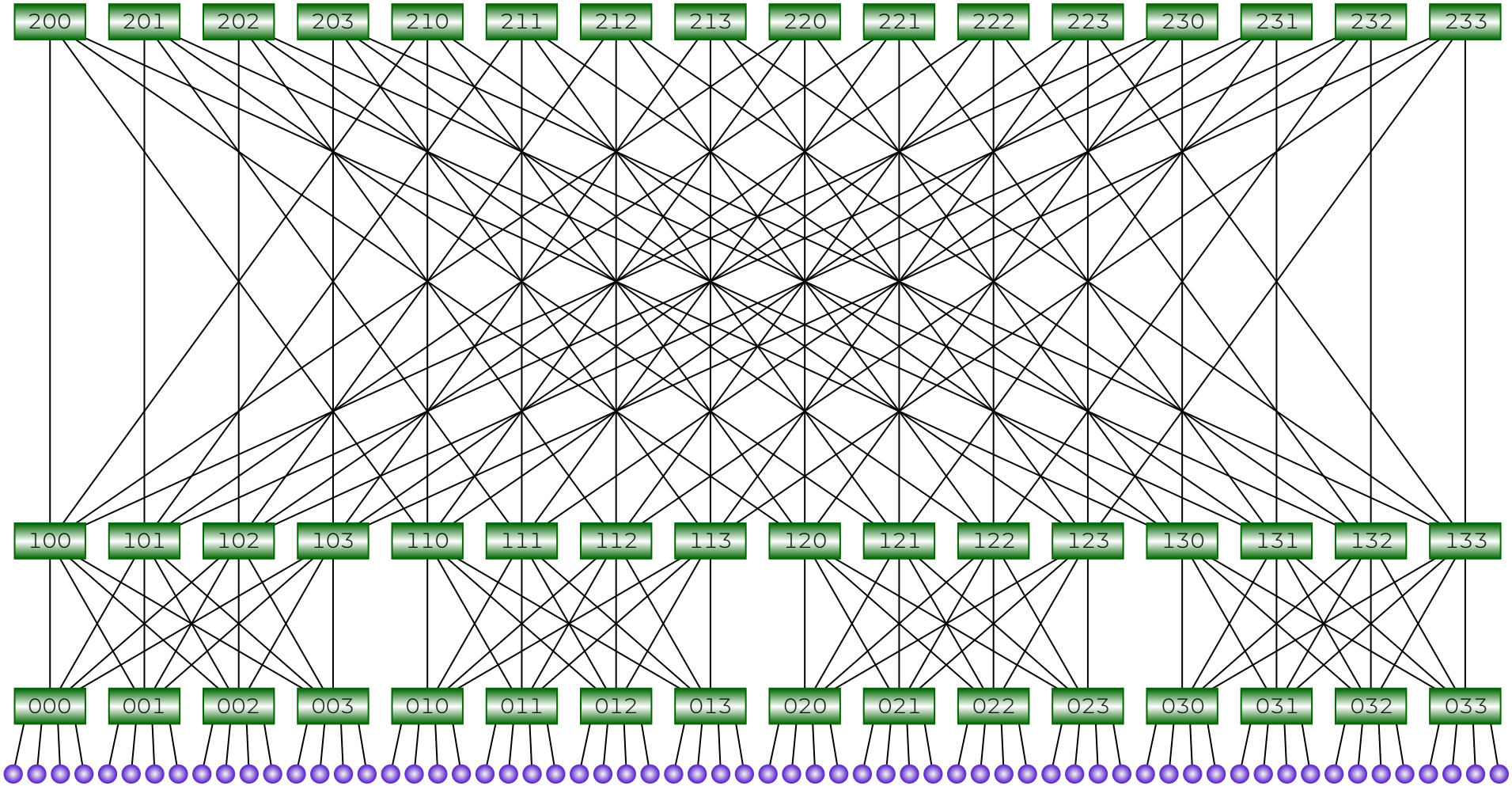


(a) Tree (b) Fat-tree



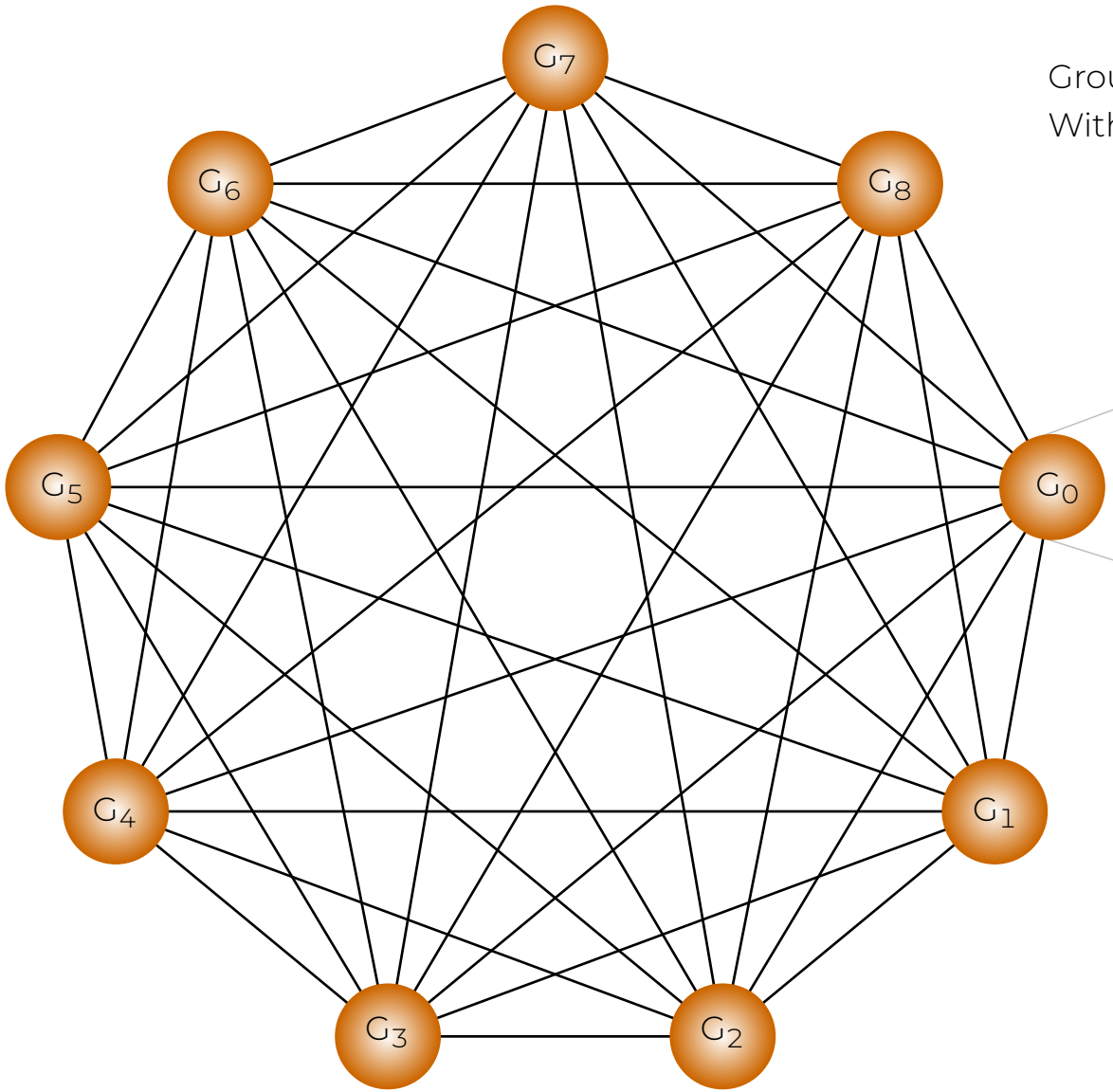
(c) Fat-tree

K-ary N-tree Interconnection Networks

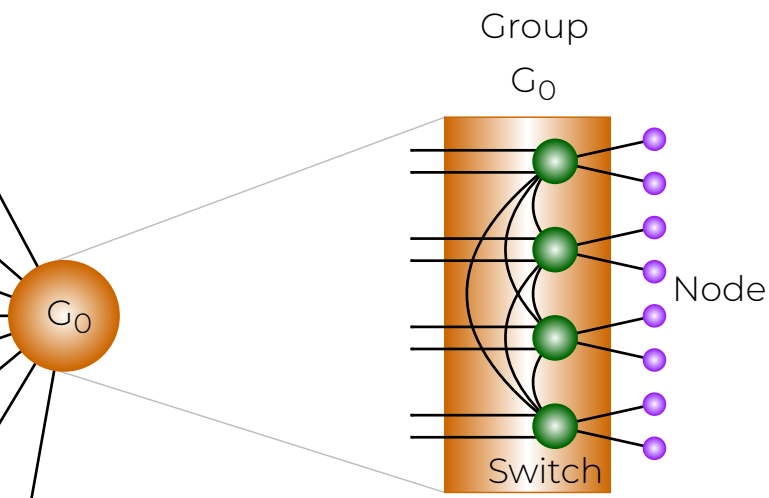


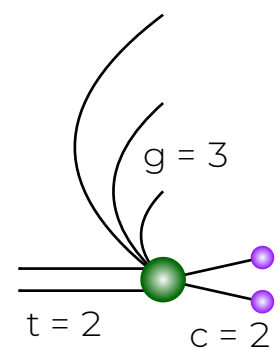
k = 4, n = 3 Not symmetric

Dragonfly Interconnection Networks



Groups are fully connected.
Within a group: switches are fully connected.





Cost Ratio (CR)

$$CR(G) = \frac{d(G) \times w_1 + D(G) \times w_2}{\log_2|(G)|}$$

d(G): the node degree of G (the number of links of a node)
D(G): the diameter of G
|(G)|: the total number of nodes in G
w_1 and w_2: weights, $w_1 + w_2 = 100\%$

INs	# Nodes	Degree	Diameter	CR
3D-Torus (10)	1,000	6	15	1.05
10-cube	1,024	10	10	1.00
3D-Torus (128)	2,097,152	6	192	4.72
21-cube	2,097,152	21	21	1.00

$w_1 = w_2 = 0.5$

Problems of Interconnection Networks

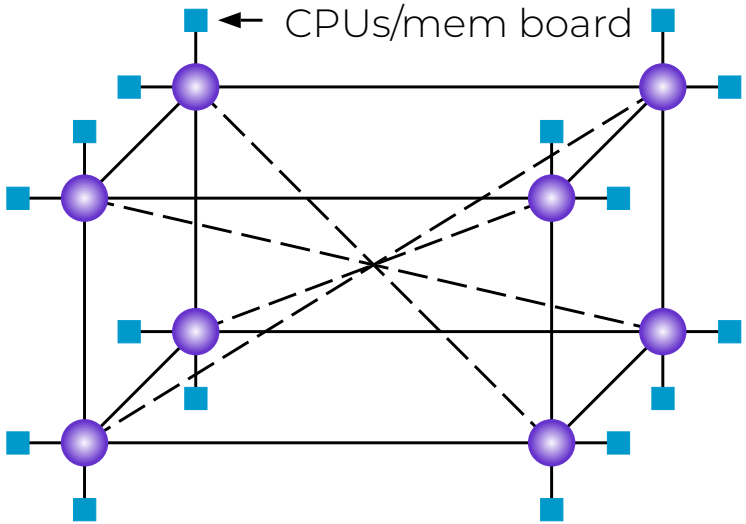
- Problems of Hypercube, 3D Torus, and Tree
  - Hypercube problem
    - The number of links increases logarithmically as the number of nodes increases: $n = \log_2 N$
  - 3D Torus problem
    - Large (long) diameter: $D = \lfloor X/2 \rfloor + \lfloor Y/2 \rfloor + \lfloor Z/2 \rfloor$
  - Tree problem
    - Not symmetric

Problems of Interconnection Networks

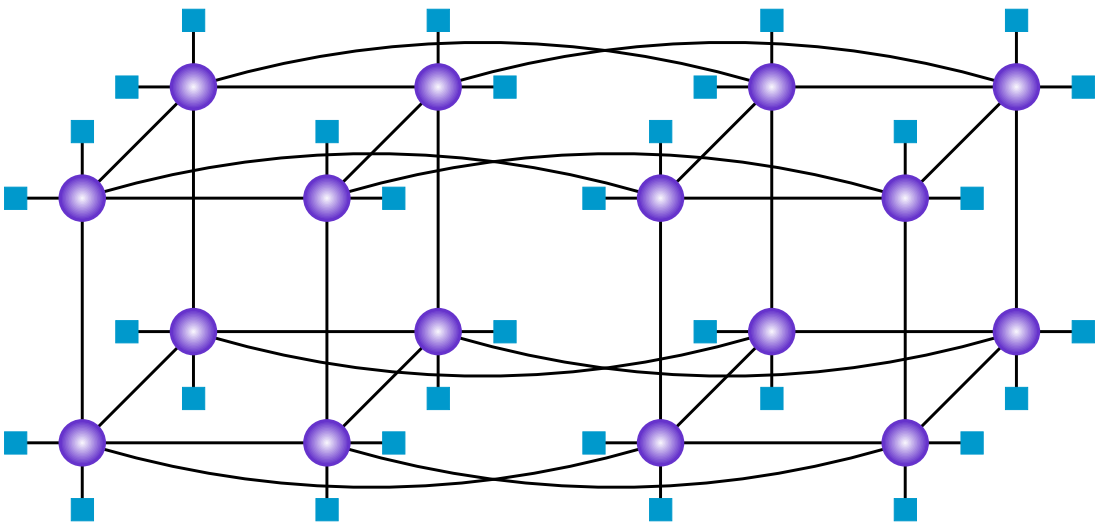
- Suppose a supercomputer has 2^{21} = 2,097,152 nodes
- By Hypercube
  - Diameter = 21
  - Node-degree = 21 (high cost)
- By $128 \times 128 \times 128$ Torus
  - Diameter = $\lfloor 128/2 \rfloor \times 3 = 192$ (too large)
  - Node-degree = 6

SGI Origin2000's Solution

Origin2000 of 3D and 4D



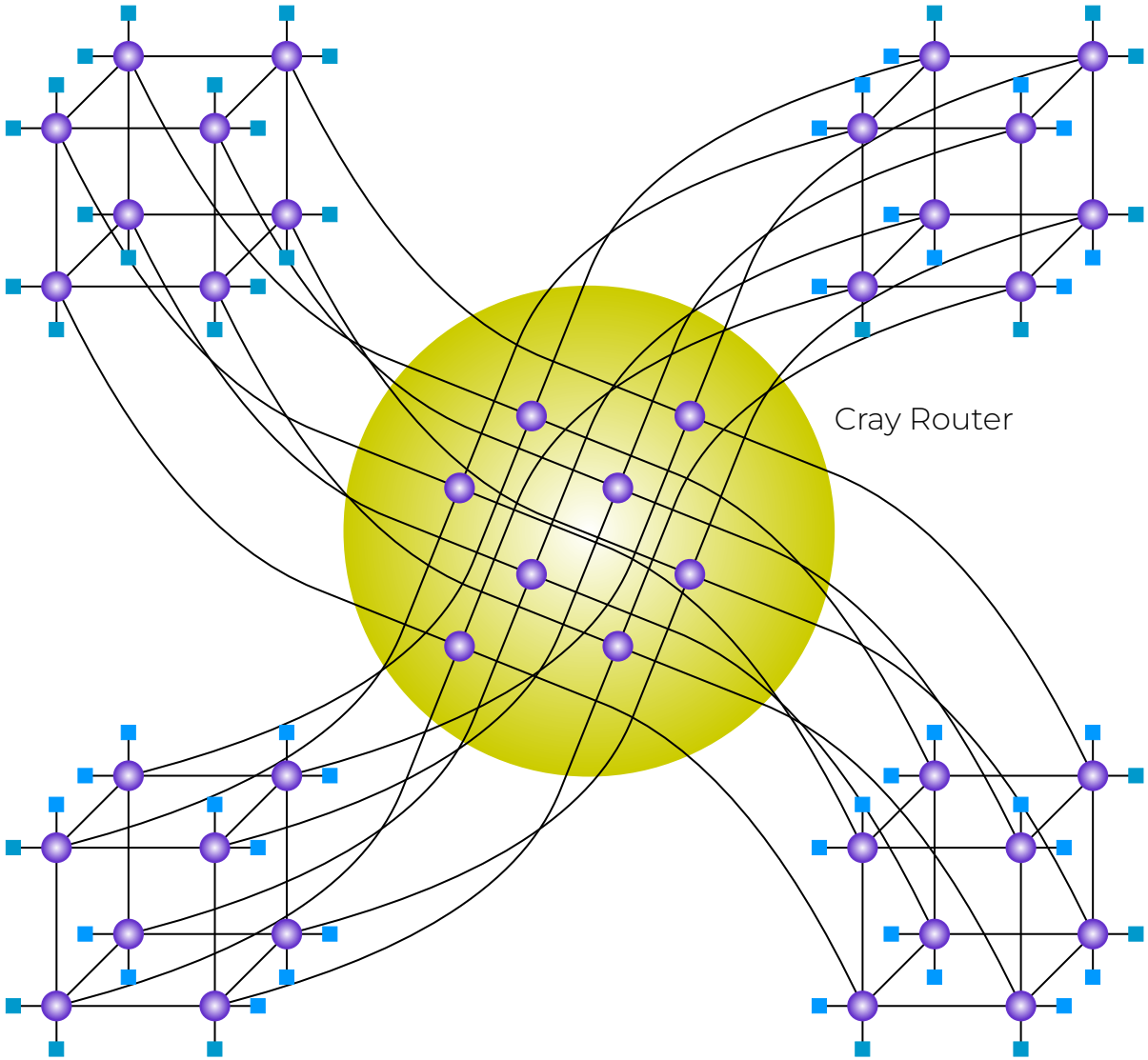
(a) 3D Origin 2000 (32 CPUs)



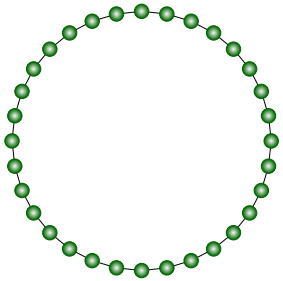
(b) 4D Origin 2000 (64 CPUs)

- An Origin2000 router has six links
- Two links connect to two CPU boards
- A board contains two CPUs

Cray Router in 5D Origin2000

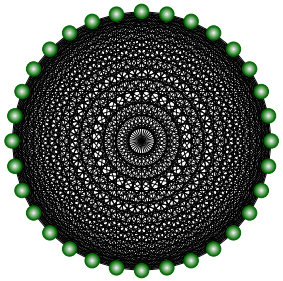


WWW — What We Want



Low node degree

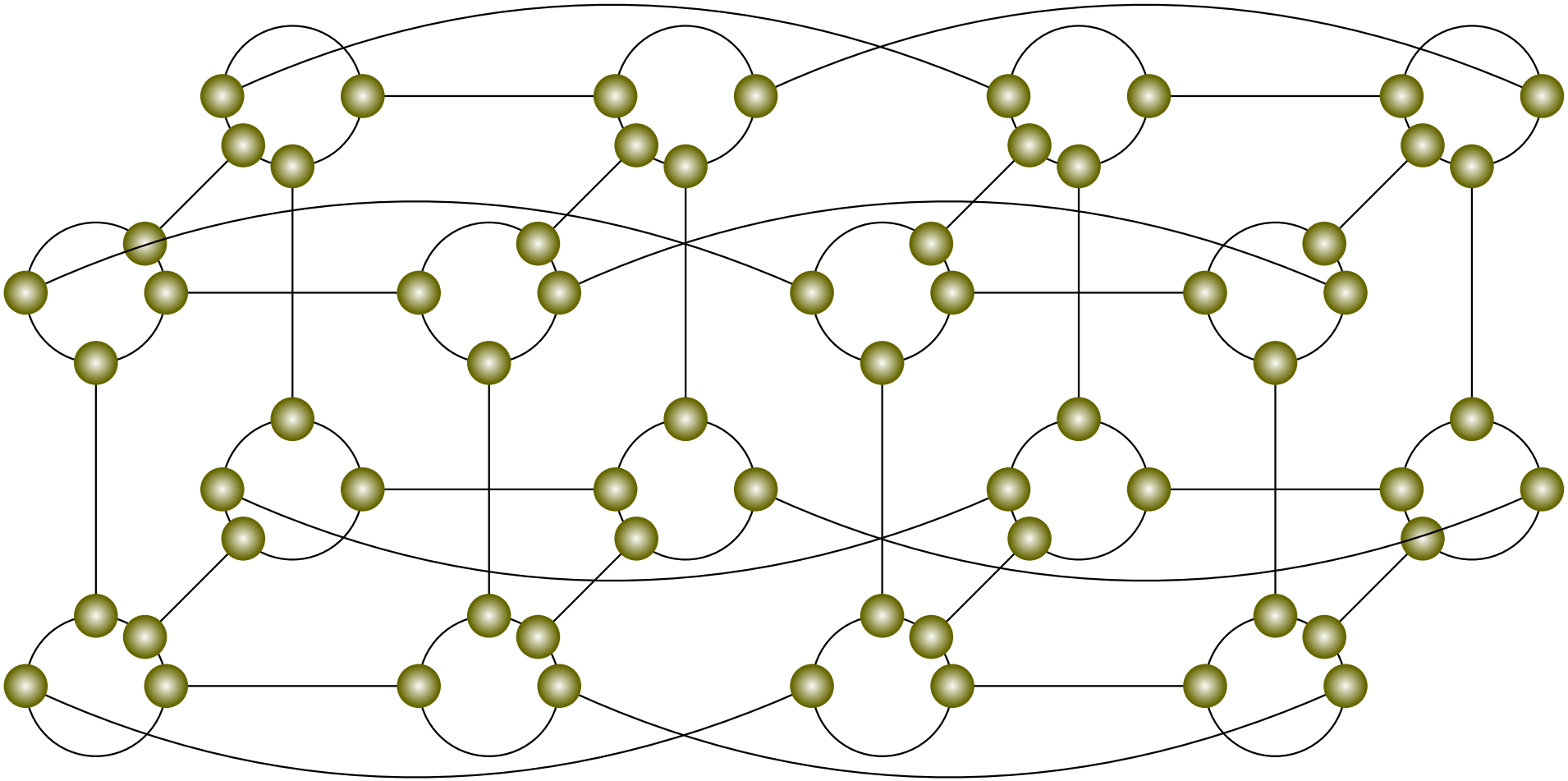
and



Short diameter

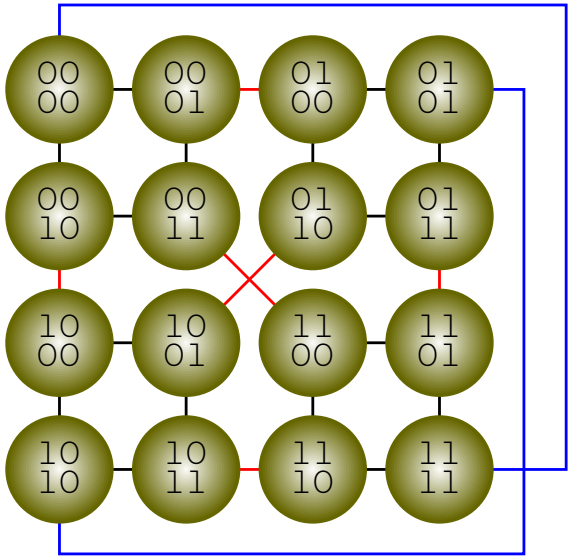
These are conflicting!

Cube-Connected Cycles



Node degree === 3

Hierarchical Cubic Network



Each node (X, Y) is adjacent to

1. $(X, Y^{(k)})$ for all $1 \leq k \leq n$, where $Y^{(k)}$ differs from Y at the kth bit position,
2. (Y, X) if $X \neq Y$, and
3. $(\overline{X}, \overline{Y})$ if $X = Y$, where $\overline{X}$ and $\overline{Y}$ are the bitwise complements of X and Y, respectively.

Dual-Cube

Dual-Cube Interconnection Network

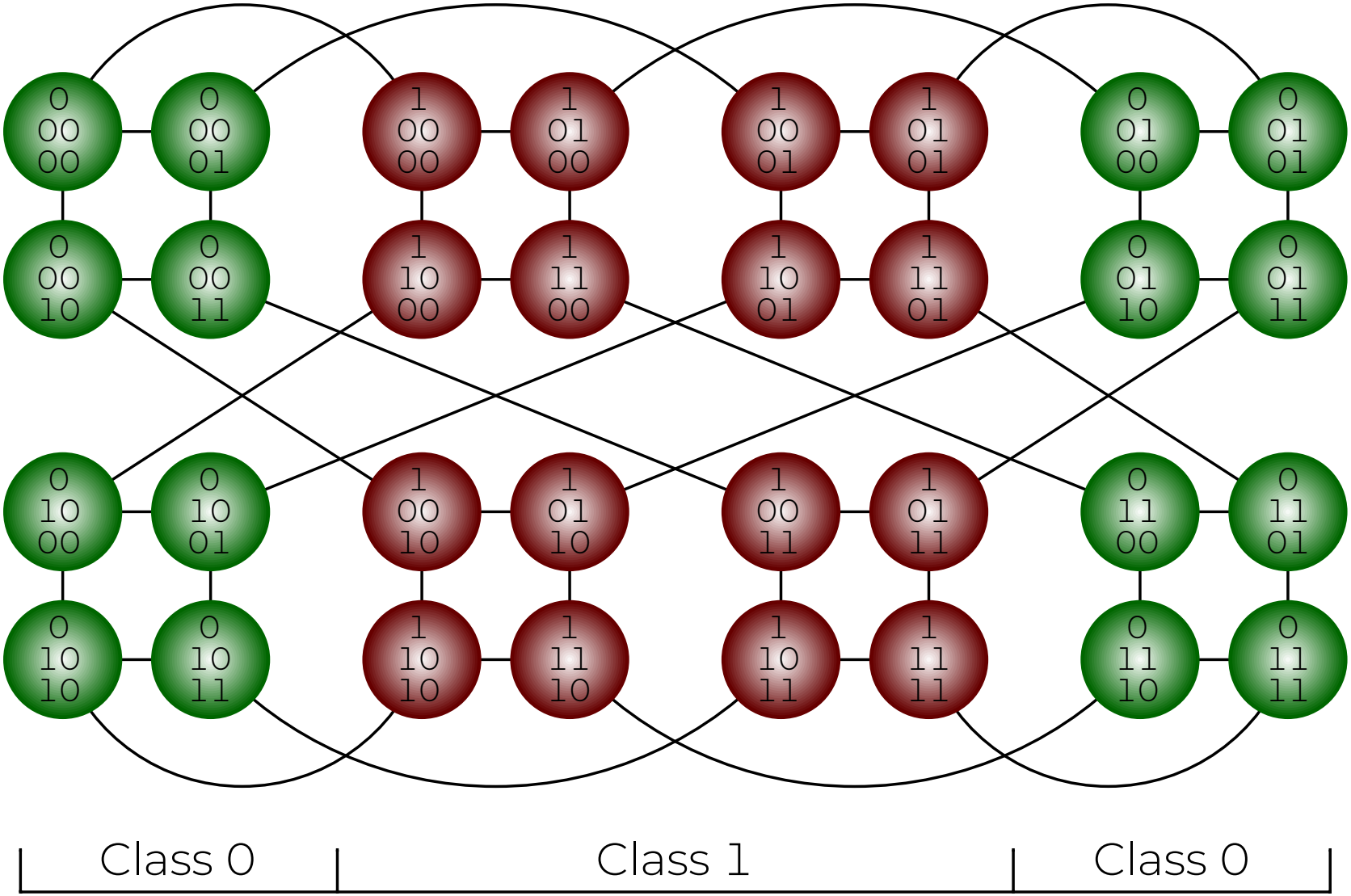
- Node degree: $m + 1$
- Can connect $N = 2^{2m+1}$ nodes
- Keeps the main properties of hypercube
- Simple routing algorithm
- Is Hamiltonian
- Performs collective communications efficiently
- Low communication cost for matrix multiplication
- Easy to build disjoint paths
- Maximum length of fault-free cycle embedding
- Efficient fault-tolerant routing

Node Address Format of Dual-Cube

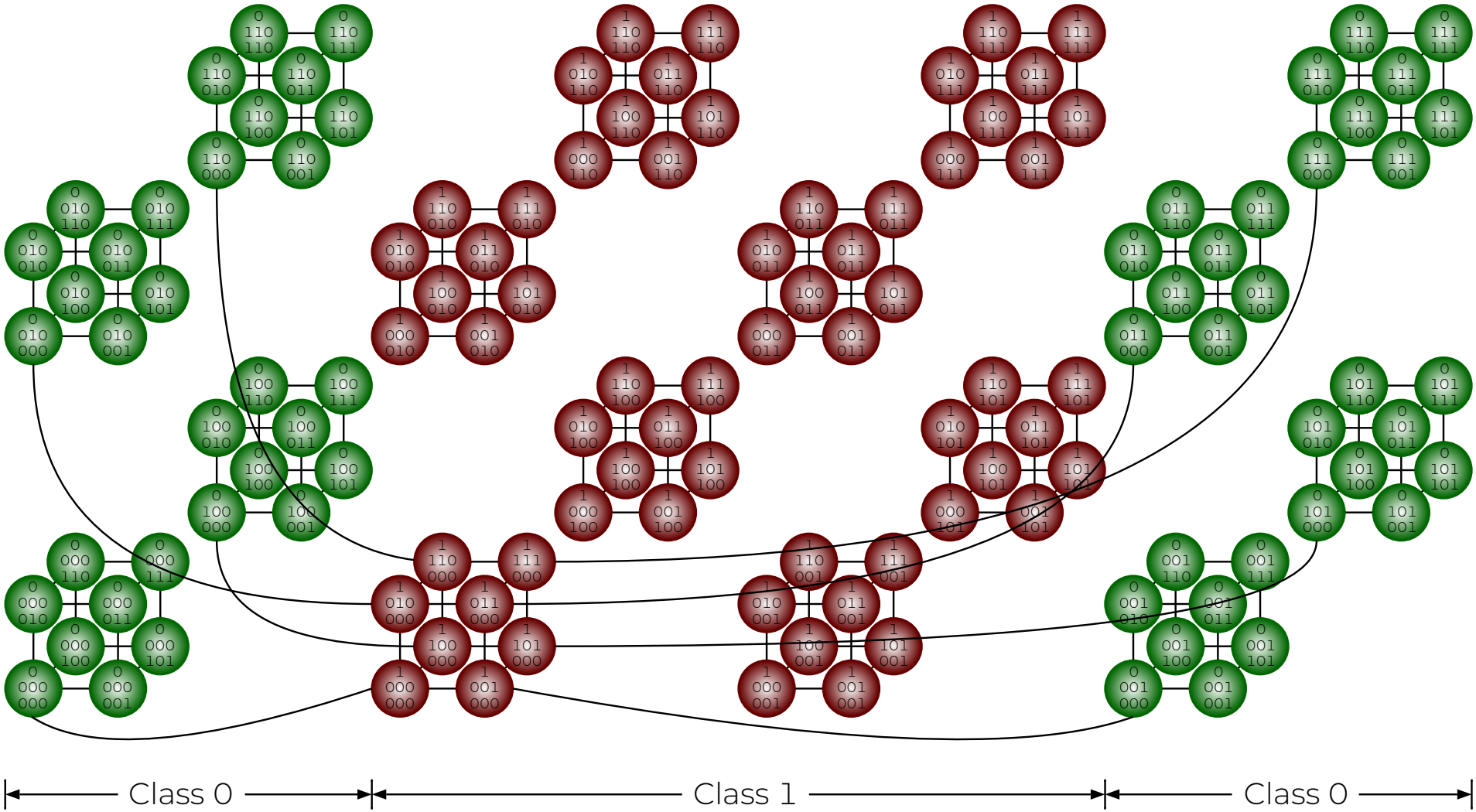
	1-bit Class ID	m bits	m bits
Class 0	0	Cluster ID	Node ID
Class 1	1	Node ID	Cluster ID

- Each node has (m + 1) links
  - The m links in Node ID builds a cluster (m-cube)
  - One link in Class ID connects to a node in a cluster of the other class
  - No links in Cluster ID
- A DC(m) can connect 2^{2m+1} nodes

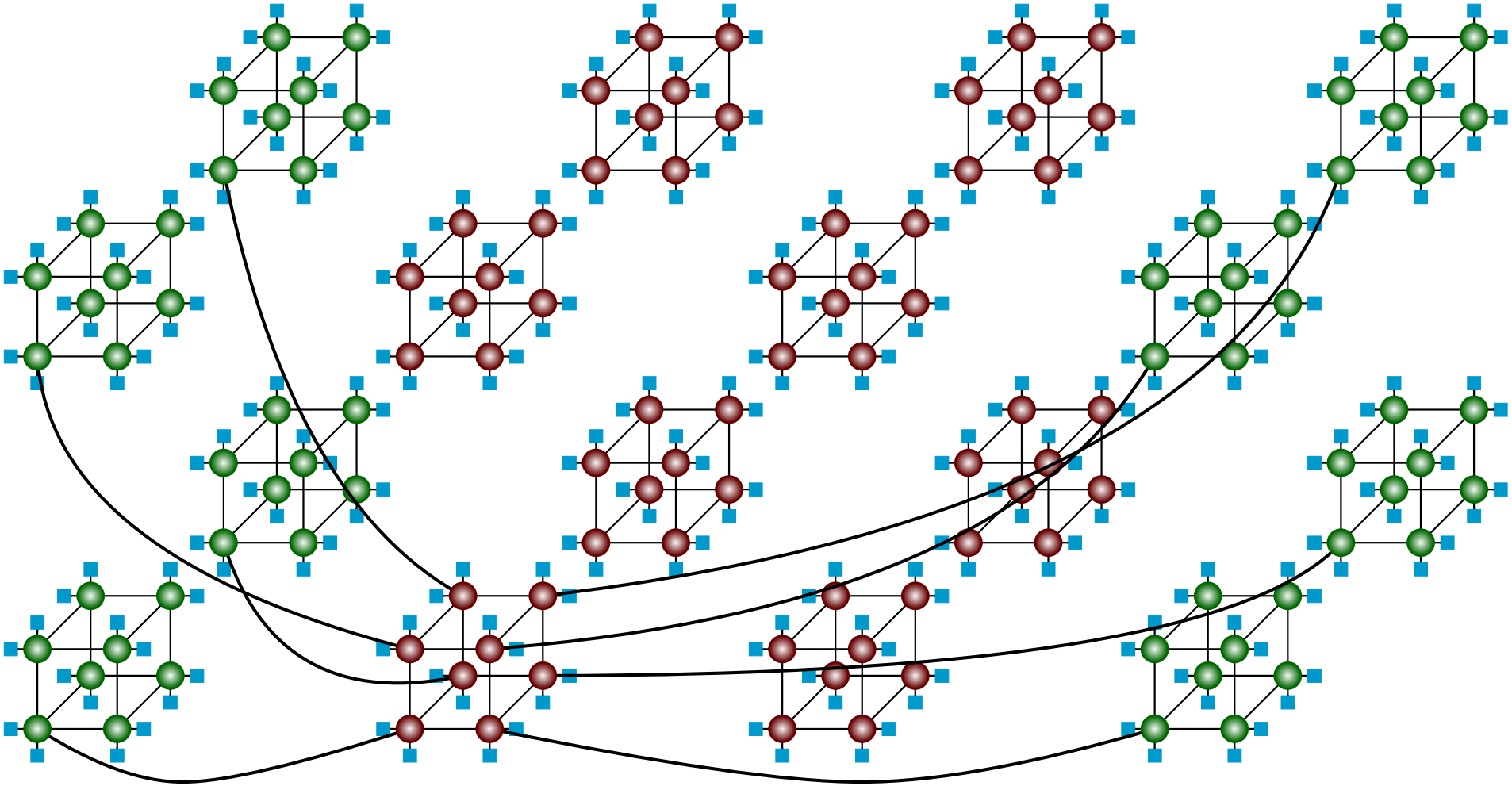
A Dual-Cube DC(2)



A Dual-Cube DC(3)



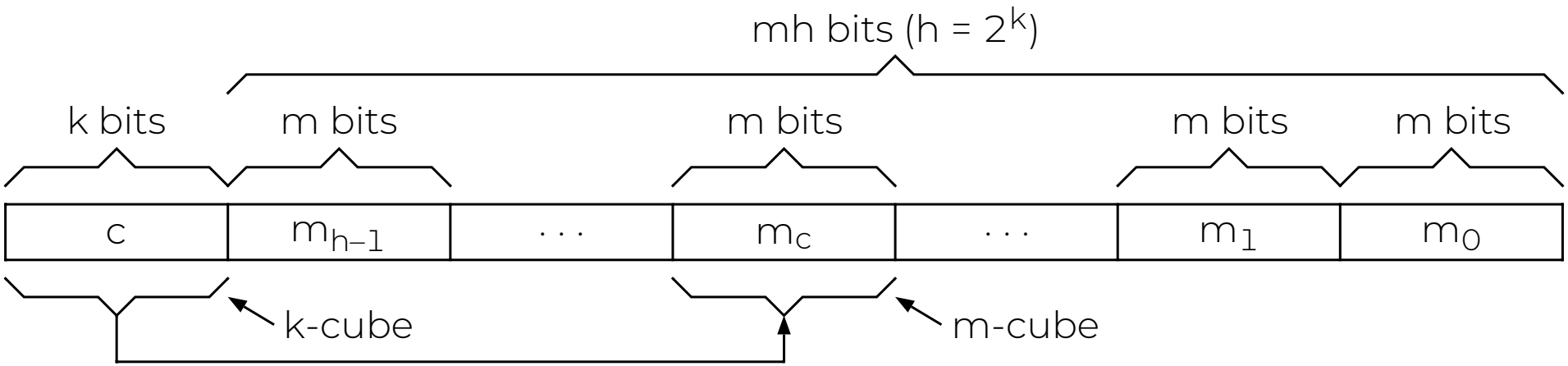
Building Origin2000 with Dual-Cube



It can connect 16 × 8 × 2 × 2 = 512 CPUs

Metacube

Node Address Format of Metacube



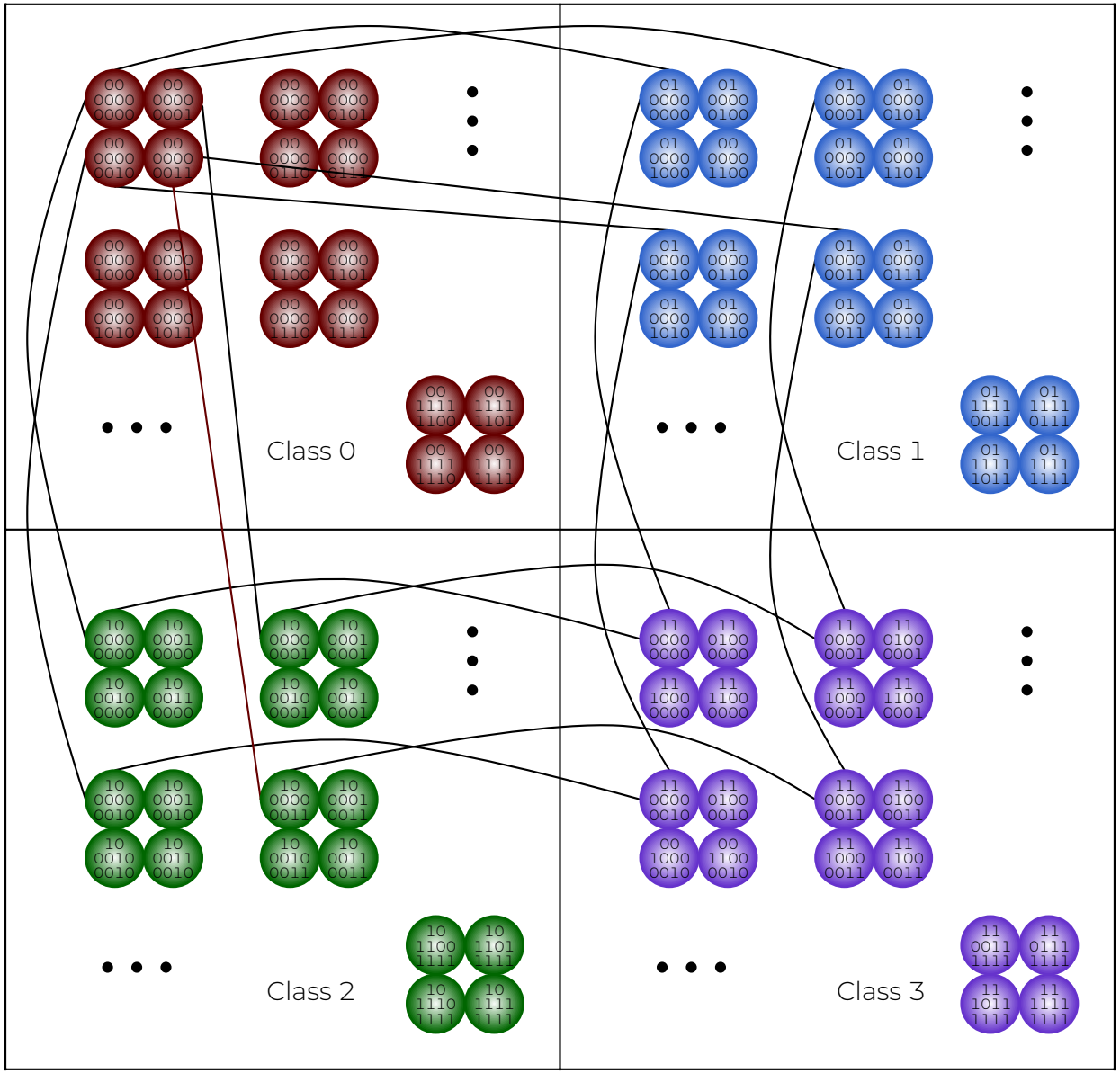
class_id : c

node_id : m_c

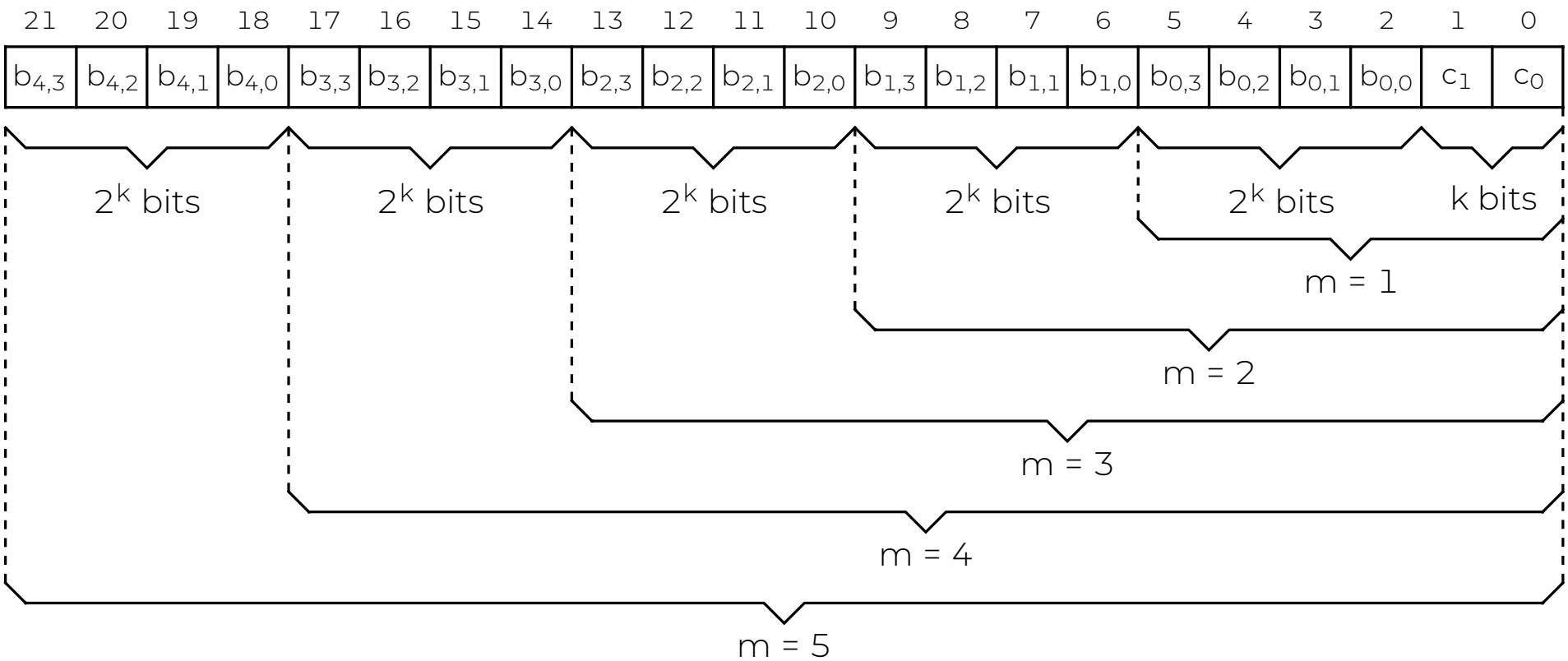
cluster_id : $m_{2^k-1}, \ldots, m_{c+1}, m_{c-1}, \ldots, m_0$

- An MC(k, m) can connect 2^{m2^k+k} nodes
- Each node has (m + k) links
- Links in c field form high-level k-cubes
- Links in m_c field form low-level m-cubes

A Metacube MC(2,2)

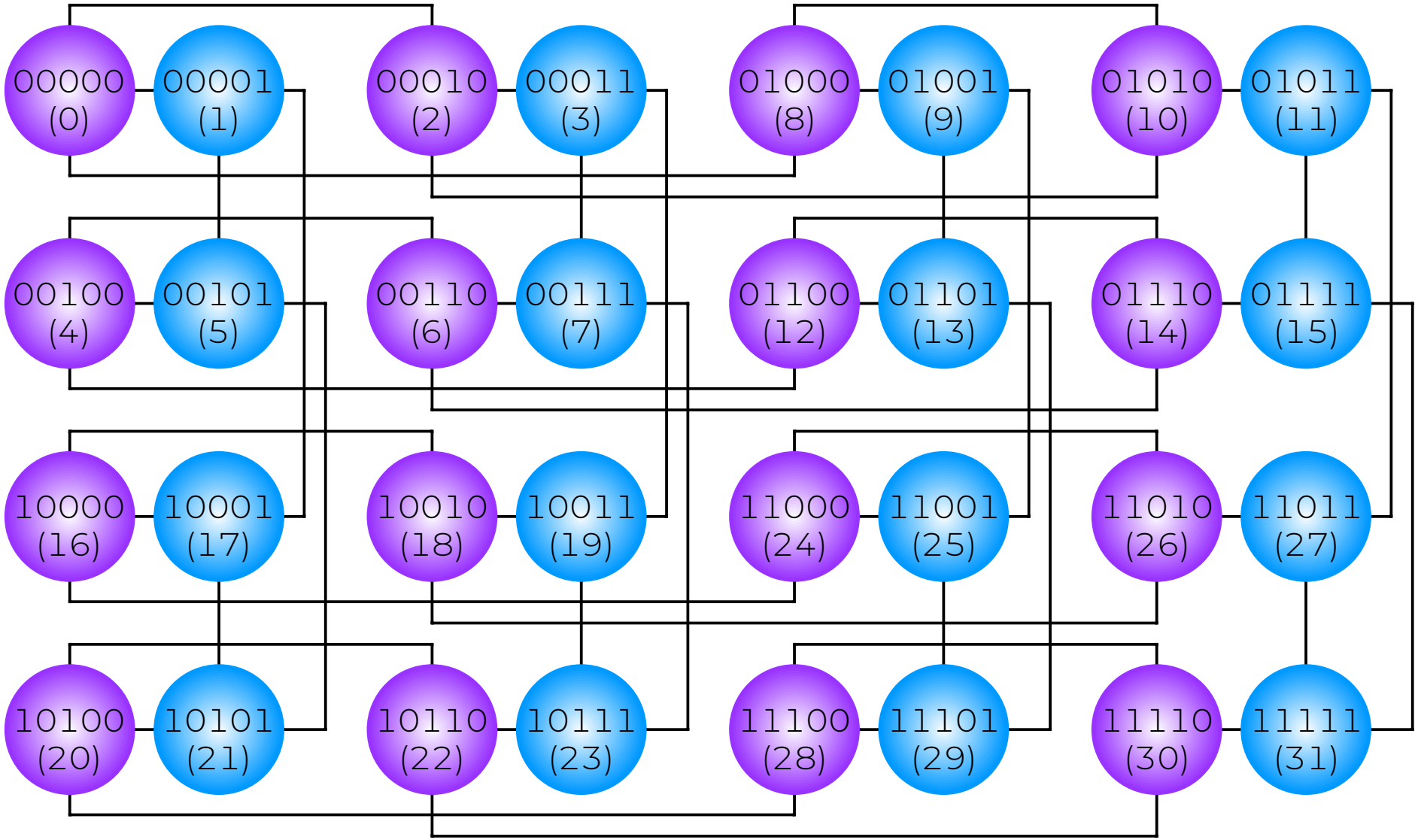


An Alternative Address Format

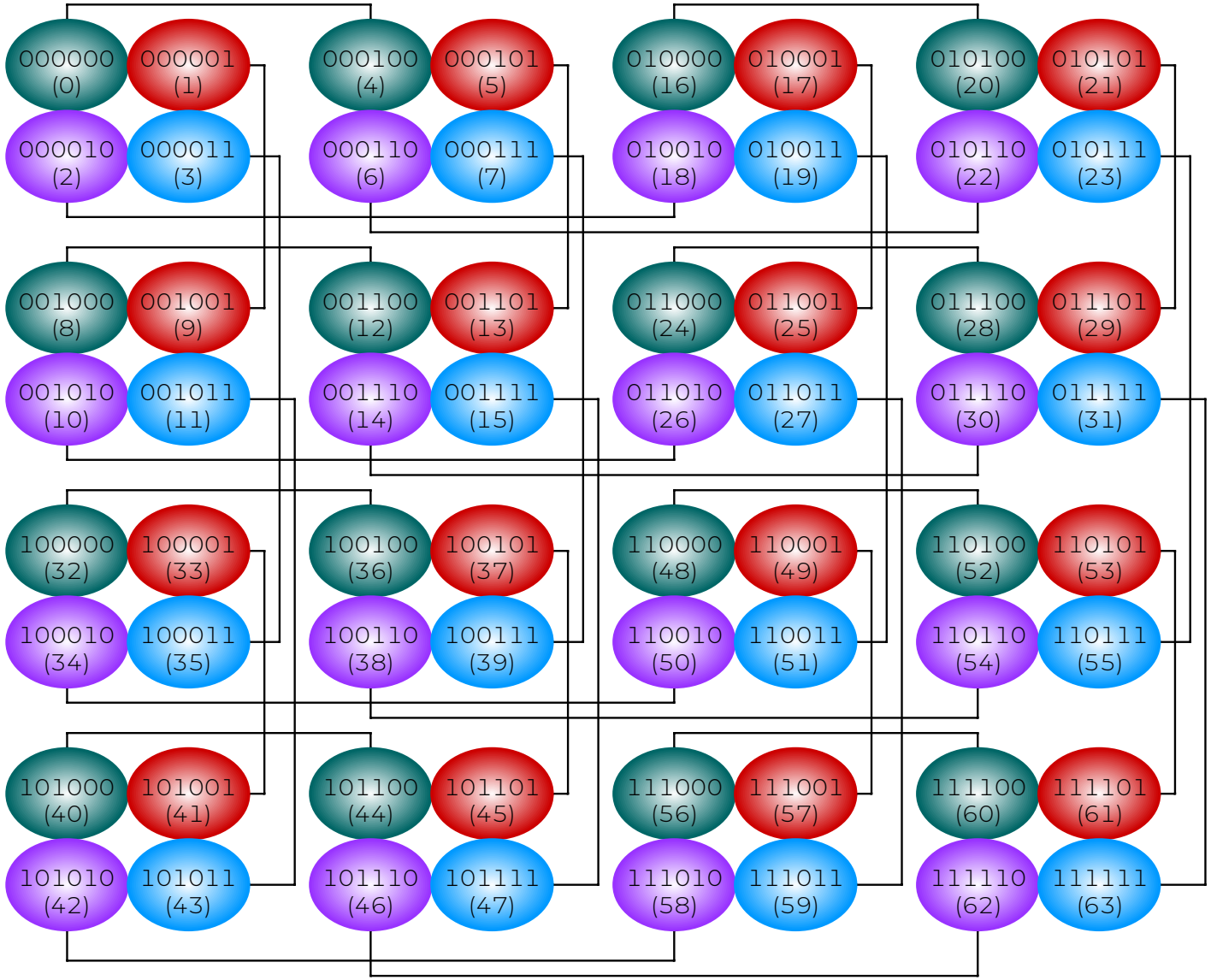


Address of k-Cube Oriented Metacube

A k-Cube Oriented Metacube MC(1,2)



A k-Cube Oriented Metacube MC(2,1)



Number of Nodes in Metacubes

An MC(k, m) can connect 2^{m2^k+k} nodes. Degree: $m + k$

Links/node	3	4	5	6	7	8
Hypercube	8	16	32	64	128	265
MC(1, m)	32	128	512	2,048	8,192	32,768
MC(2, m)	64	1,024	16,384	2^{18}	2^{22}	2^{26}
MC(3, m)	–	2,048	2^{19}	2^{27}	2^{35}	2^{43}
MC(4, m)	–	–	2^{20}	2^{36}	2^{52}	2^{68}

$2^{27} = 134{,}217{,}728$

16,384 Nodes: 14-Cube vs MC(2,3)

- Hypercube
  - Links: 14 links per node
    $2^{14} \times 14/2 = 114{,}688$ links in total
  - Diameters: 14
- Metacube
  - Links: 5 links per node
    $2^{14} \times (3 + 2)/2 = 40{,}960$ links in total
  - Diameters: 16
- The reduction in the total number of links for this example is 73,728 links or about 64%
- Diameters: only 2 more than that of hypercube

MC(2,4) vs 18-Cube vs 3D Torus

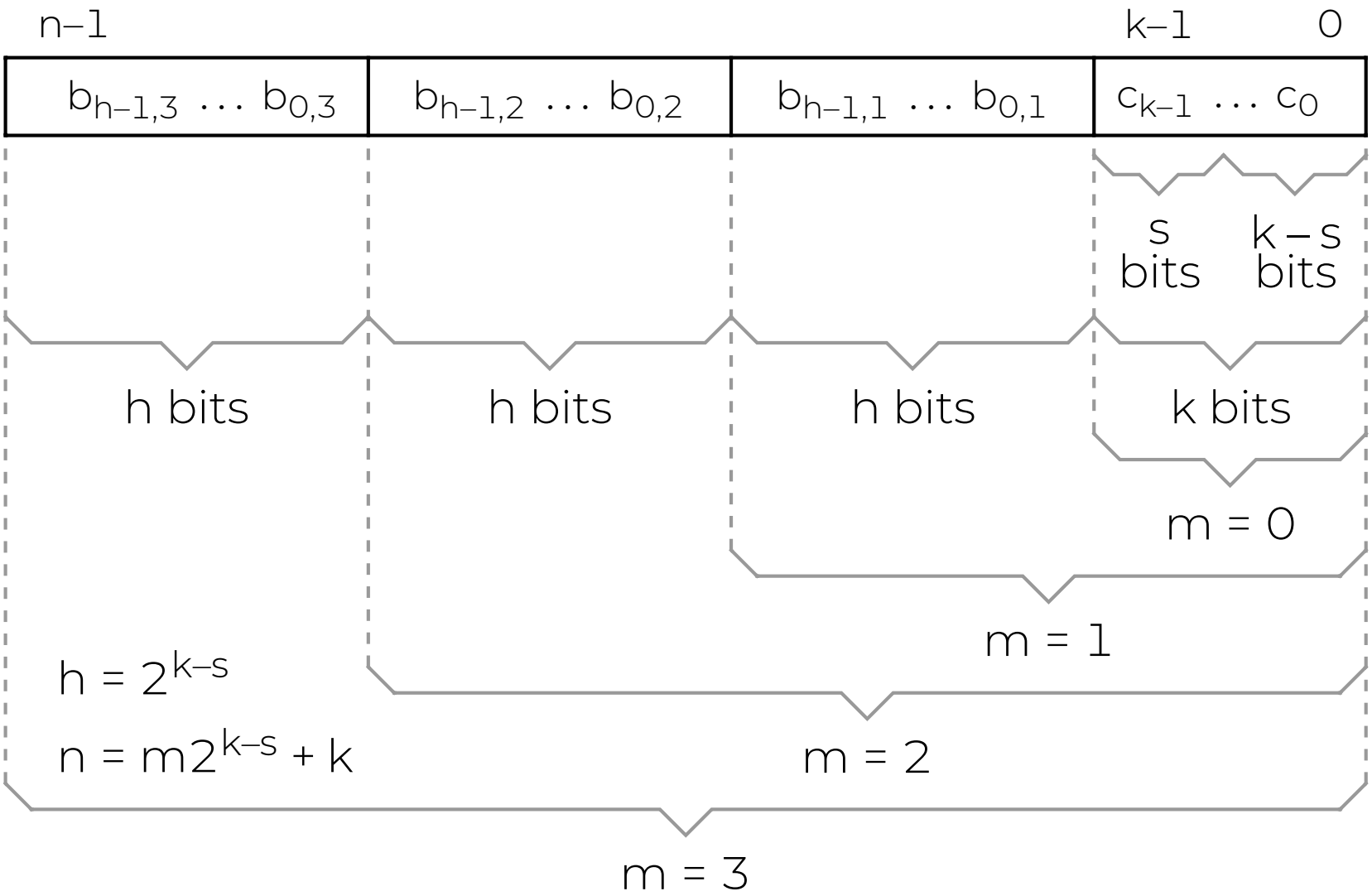
- Metacube(2,4) $k = 2,\ m = 4$
  - The number of nodes: $2^{18} = 262{,}144$
  - Node degree: 6
  - Diameter: 20
- Hypercube (18-Cube)
  - The number of nodes: $2^{18} = 262{,}144$
  - Node degree: 18
  - Diameter: 18
- 3D Torus ($64 \times 64 \times 64$)
  - The number of nodes: $2^{18} = 262{,}144$
  - Node degree: 6
  - Diameter: $32 + 32 + 32 = 96$

Building Origin2000 with MC(2,2)

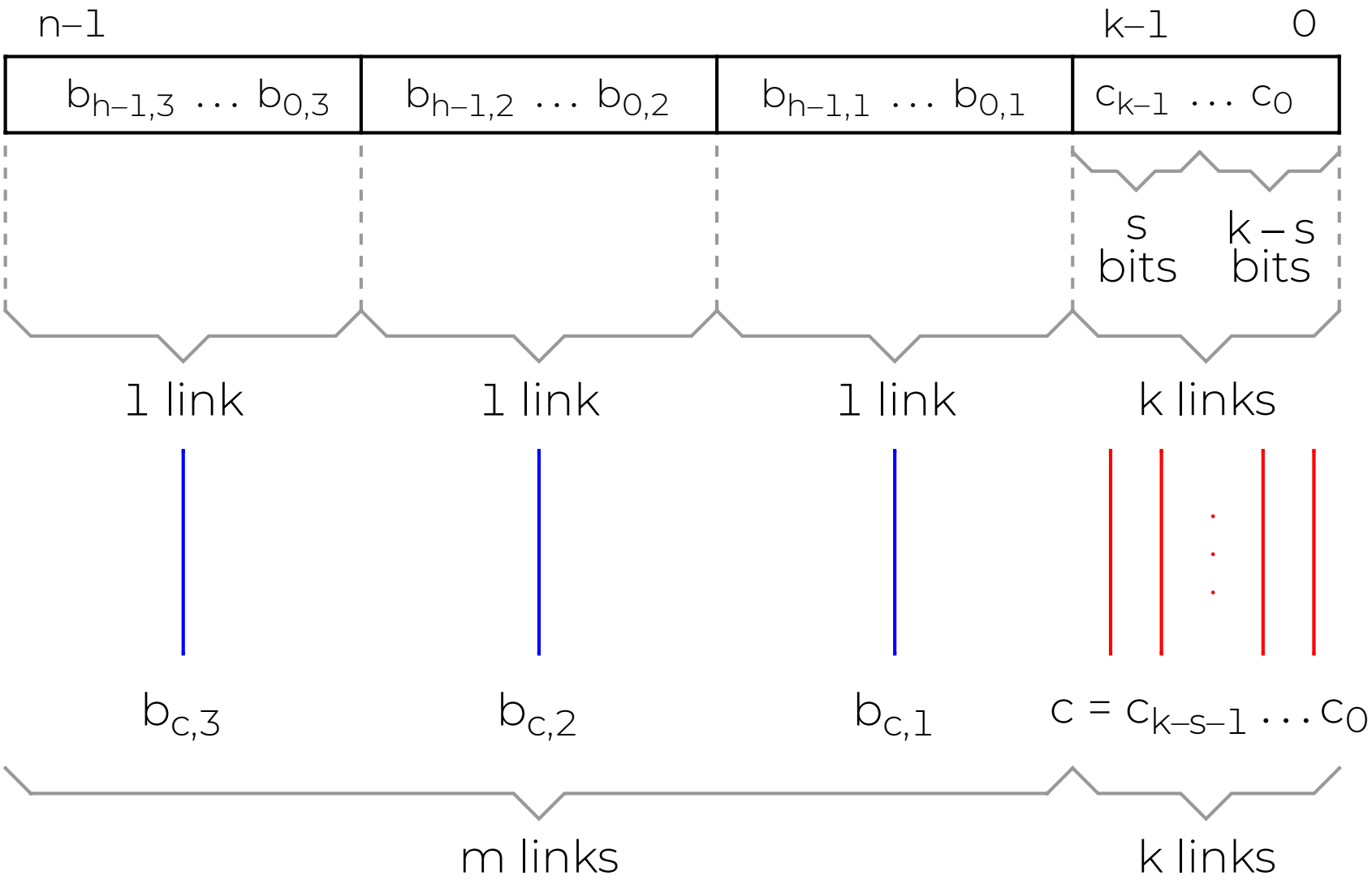
- An Origin2000 router has six links
- Two links connect to two CPU boards
- A board contains two CPUs
- Four links are used to build an MC(2,2)
  - $k = 2$ and $m = 2$ ($k + m = 4$)
  - There are $2^{m2^k+k} = 2^{10} = 1024$ nodes
- It can connect $1024 \times 2 \times 2 = 4096$ CPUs
  - Does not use the Cray Router anymore
- In contrast, the Origin2000 using Cray Router can only connect $32 \times 2 \times 2 = 128$ CPUs

KMS-Cube

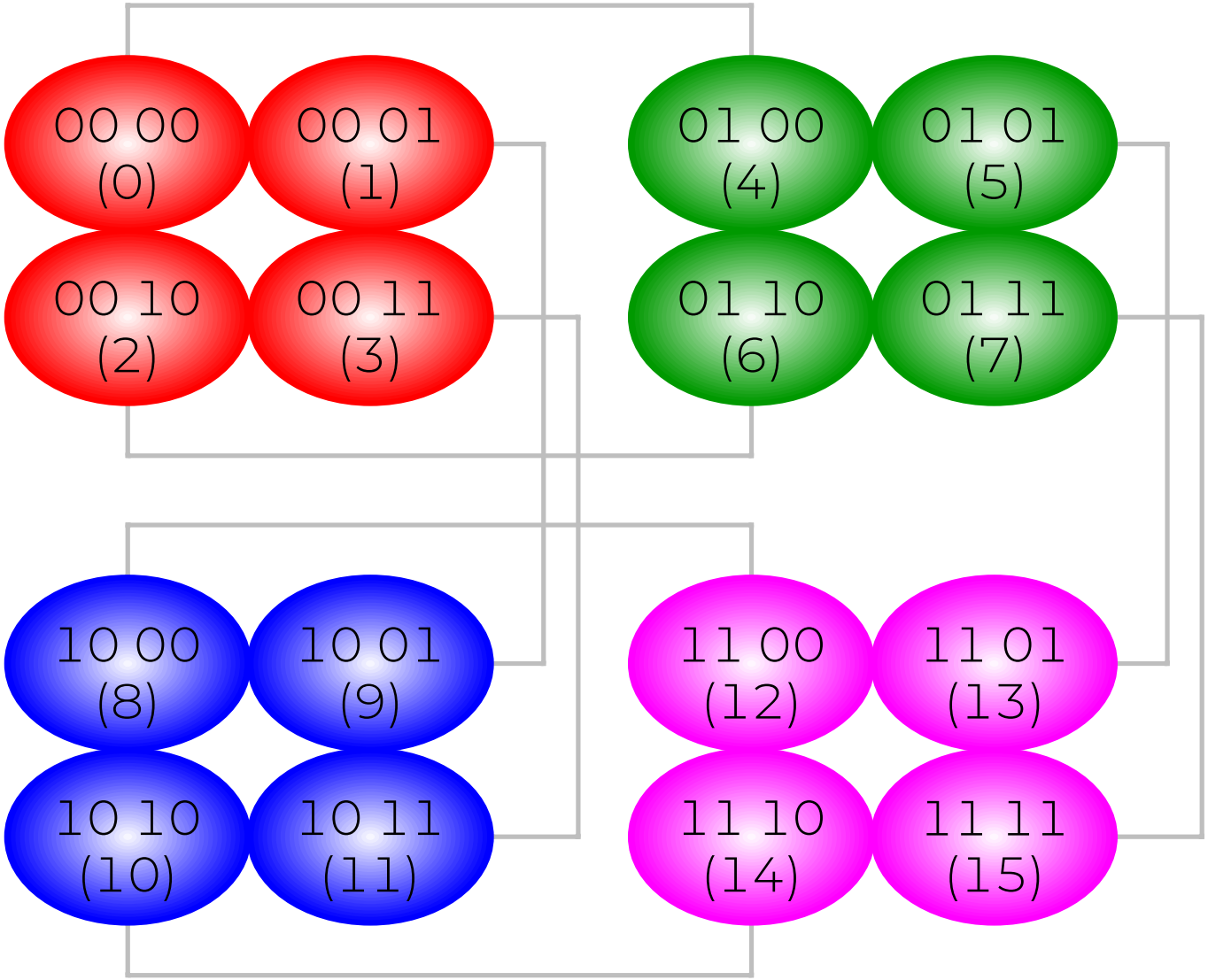
Node Address of KMS-Cube



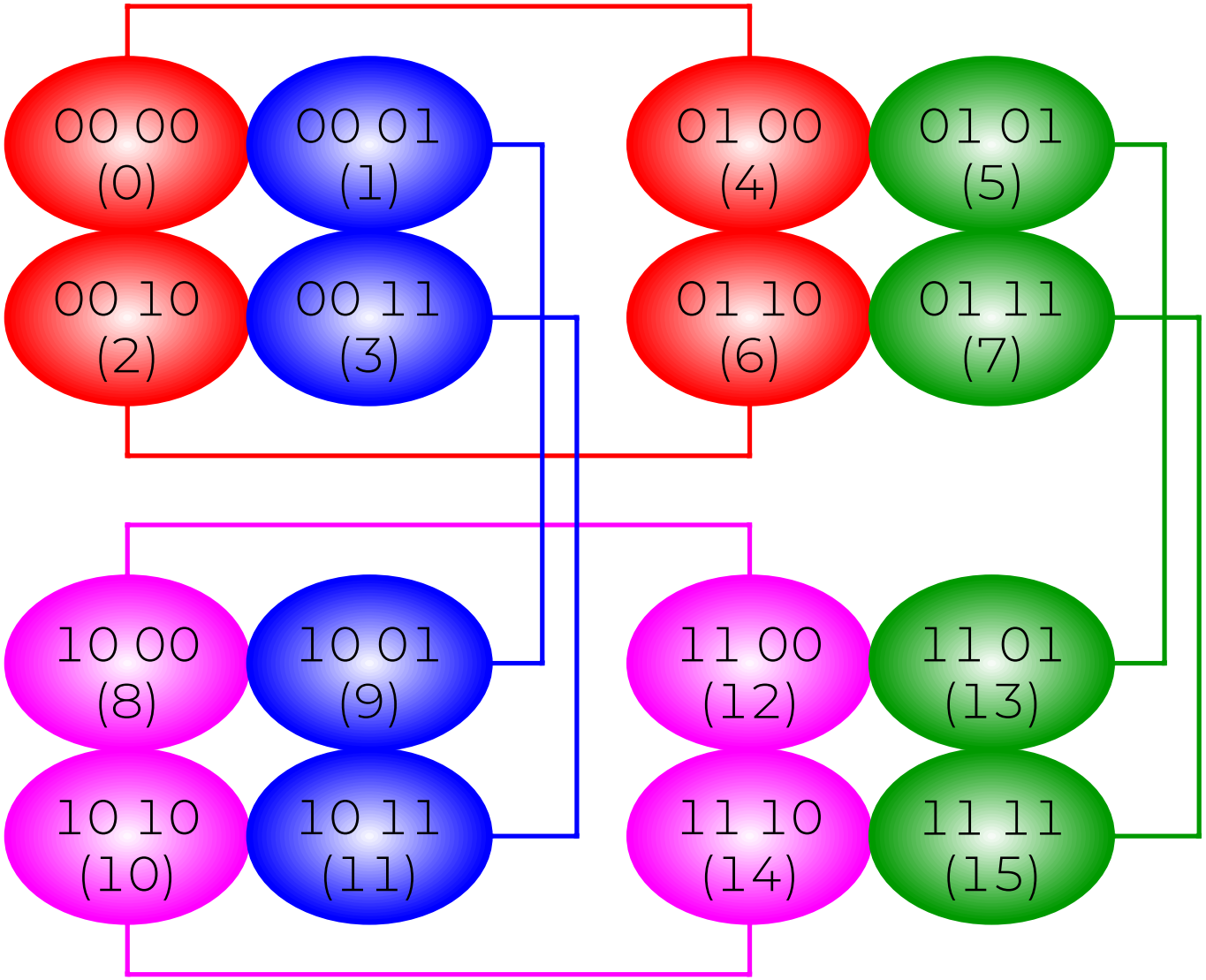
Links of KMS-Cube



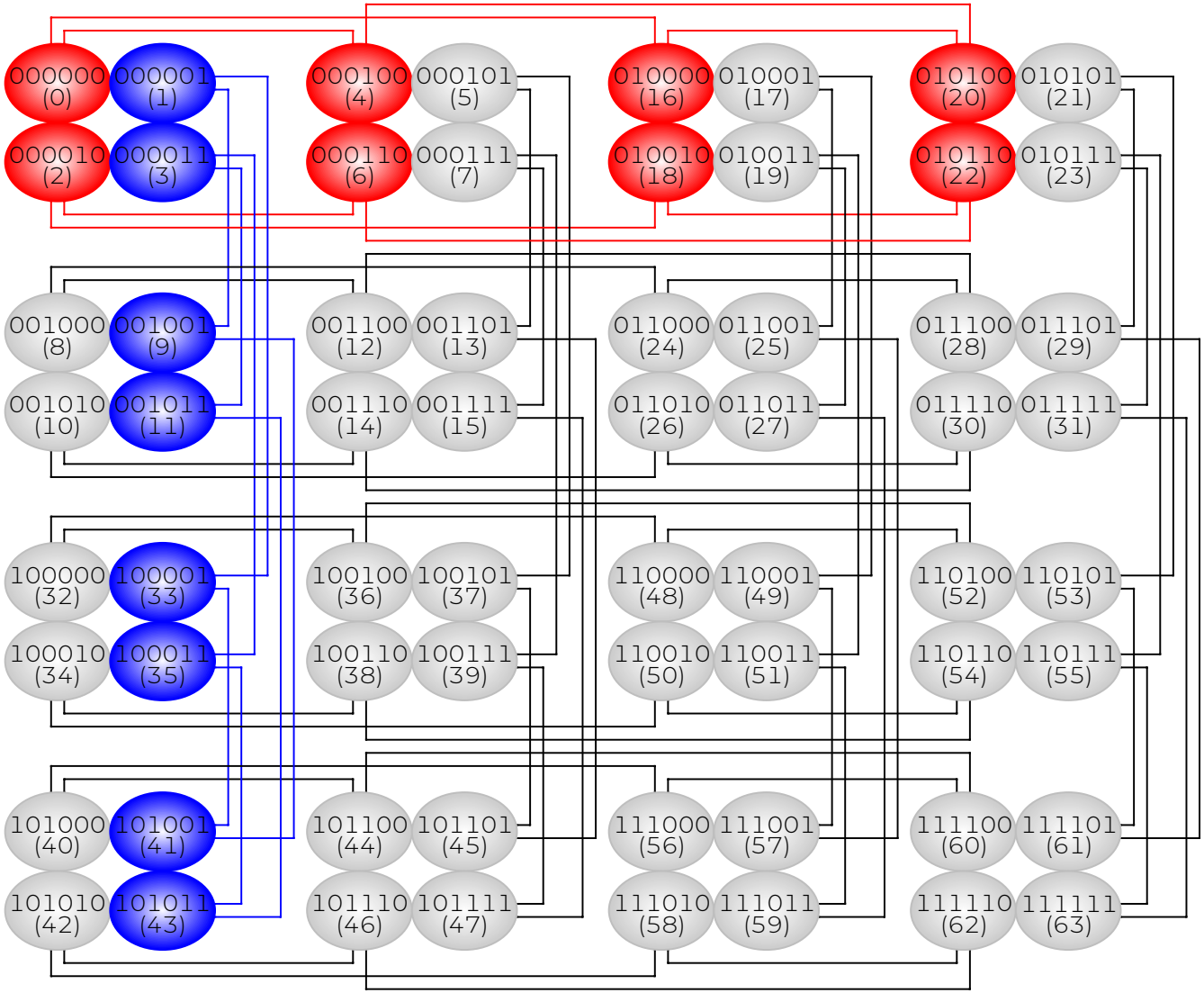
KMS-Cube with K = 2, M = 1, and S = 1



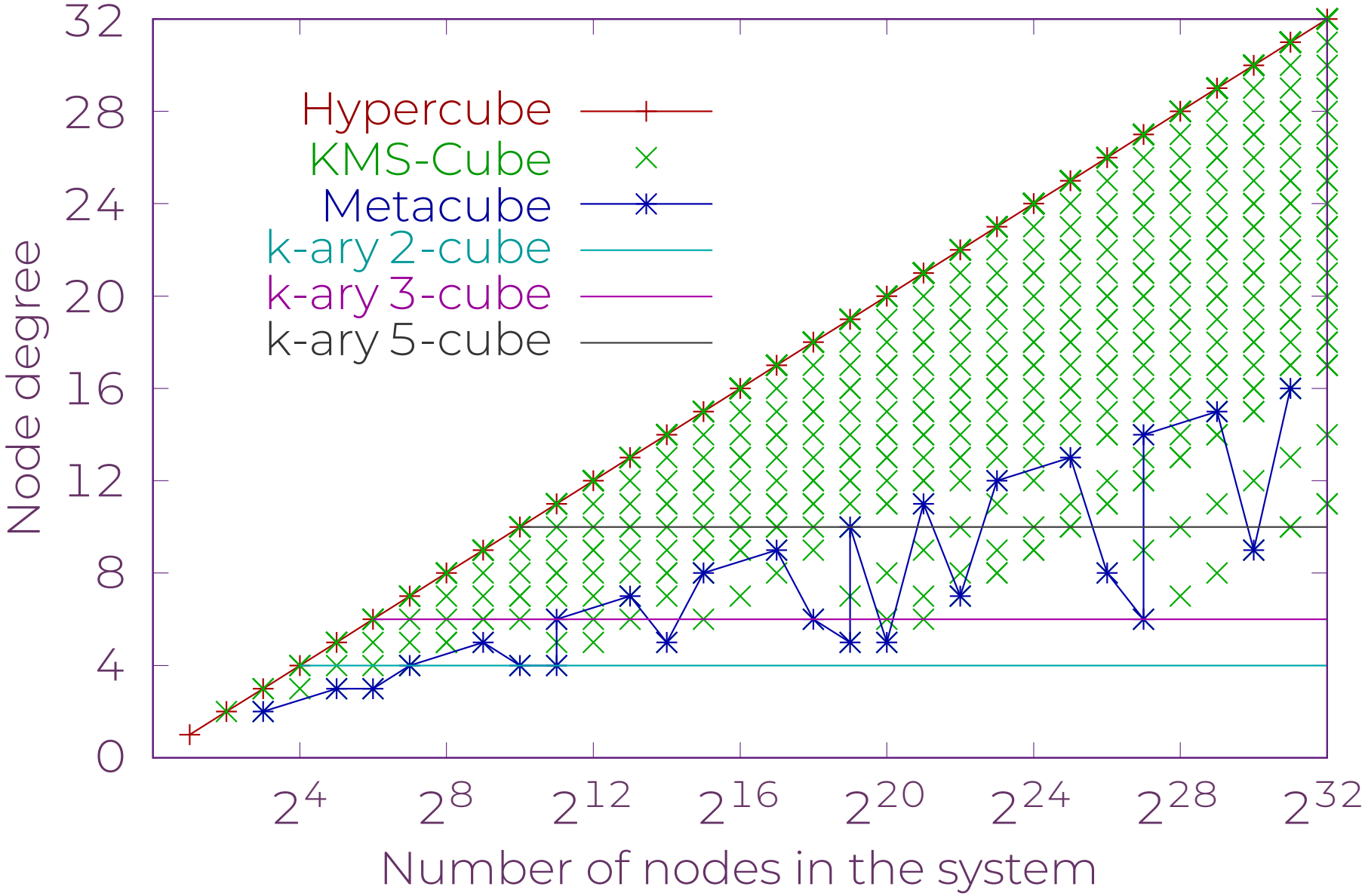
KMS-Cube with K = 2, M = 1, and S = 1



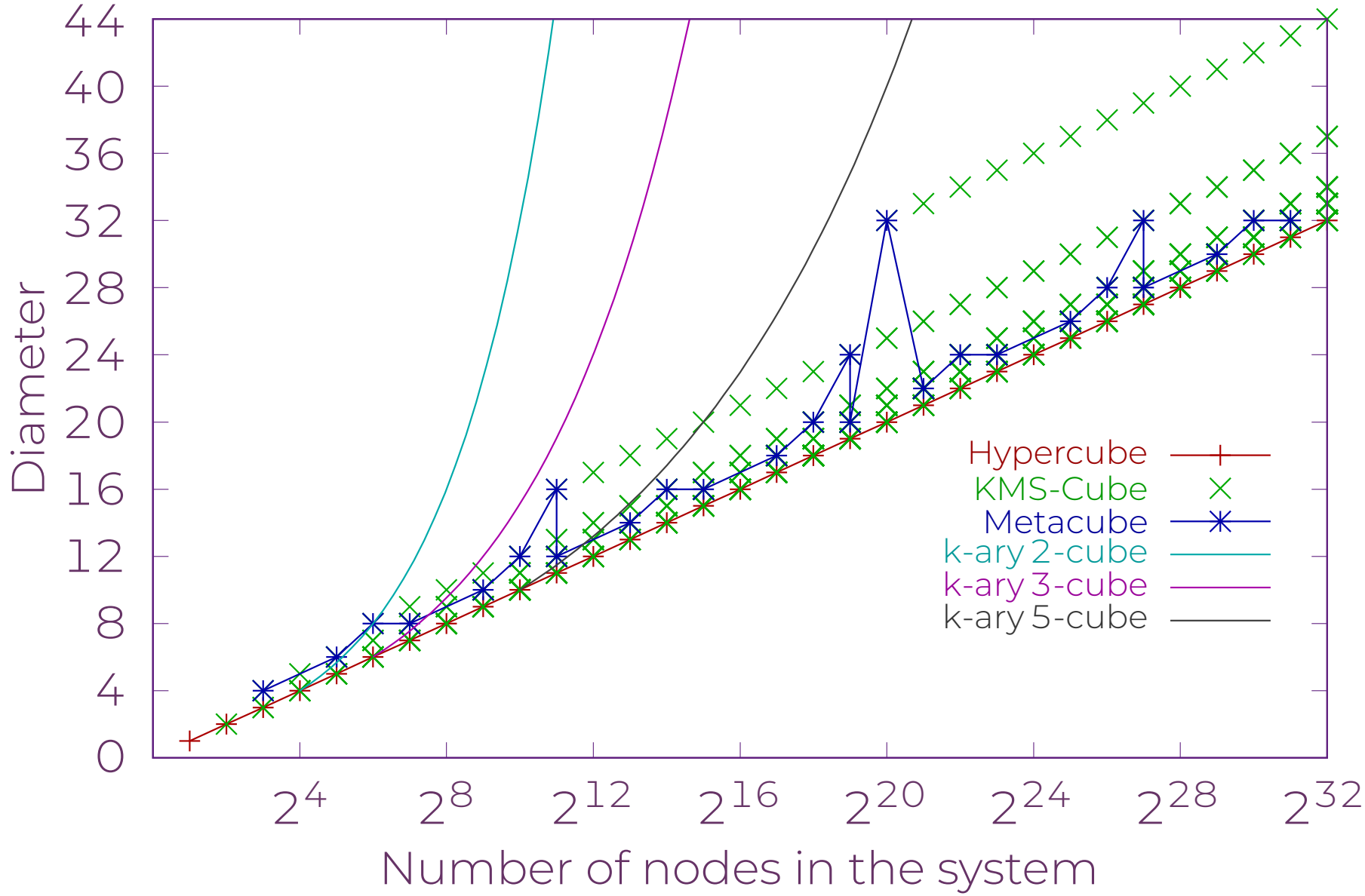
KMS-Cube with K = 2, M = 2, and S = 1



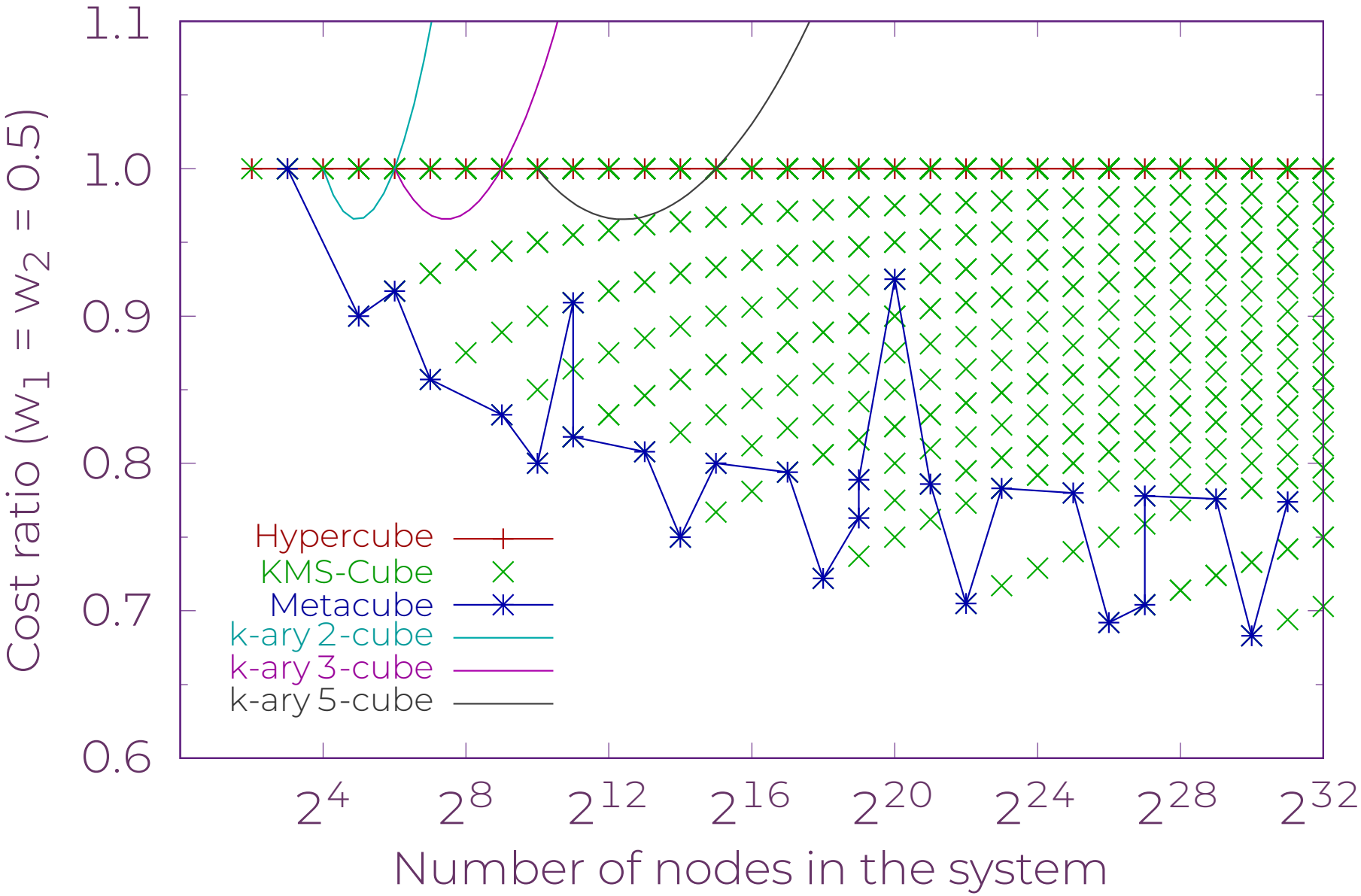
Degree Comparison



Diameter Comparison



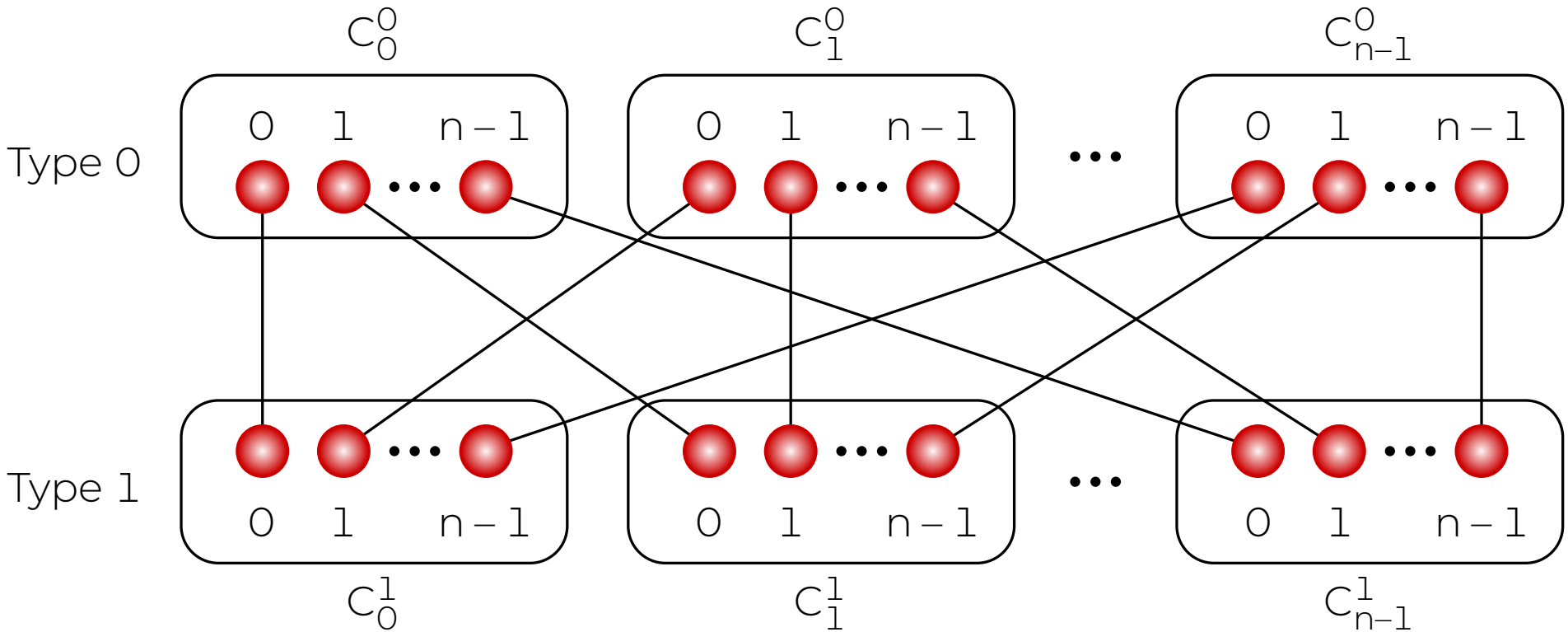
Cost Ratio Comparison



RDN — Recursive Dual-Net

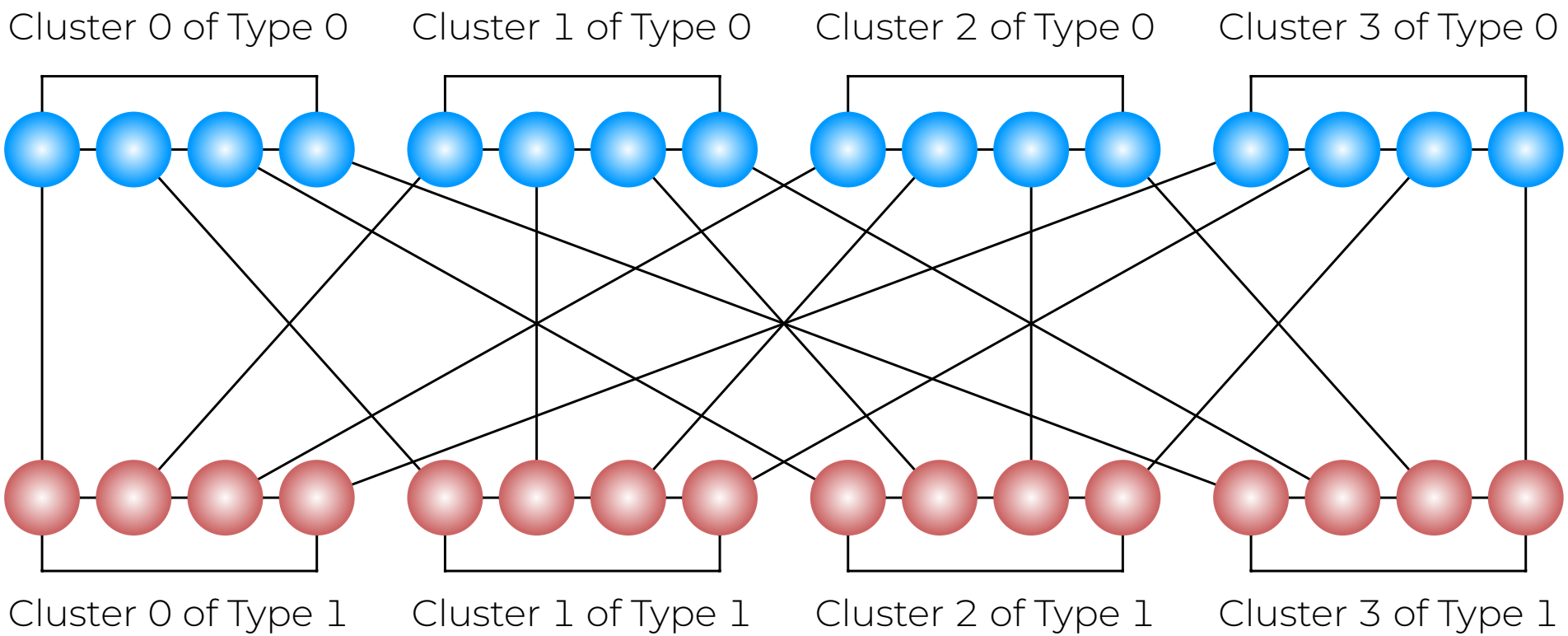
RDN

Recursive Construction of RDN



- $n_i = n_{i-1} \times n_{i-1} \times 2 = 2n_{i-1}^2$
- $N = (2m)^{2^k}/2$
  where m is the number of nodes in a base network; k: level

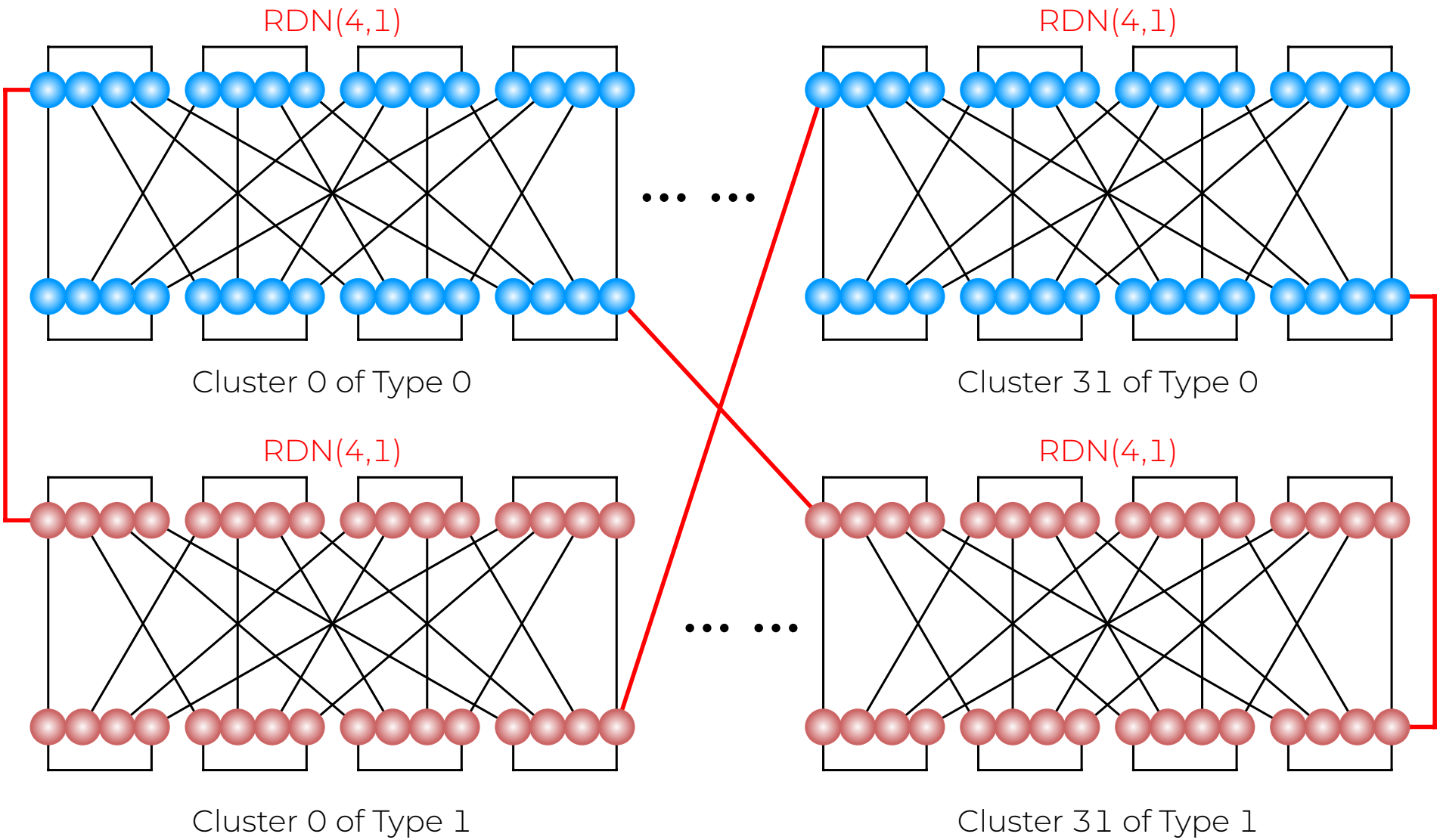
A Recursive Dual-Net RDN(4,1)



RDN(m, k)

m: the number of nodes in the base network; k: level
The base network can be an any symmetric network.

A Recursive Dual-Net RDN(4,2)



RDN(27,2)

- Base network: 3D Torus ($3 \times 3 \times 3$), $k = 2$
  - $m = 27$ (nodes)
  - $d_0 = 6$ (degree)
  - $D_0 = 1 + 1 + 1 = 3$ (diameter)
- The number of nodes
  - $k = 1$: $N_1 = 27 \times 27 \times 2 = 1{,}458$
  - $k = 2$: $N_2 = 1{,}458 \times 1{,}458 \times 2 = 4{,}251{,}528$
- $d = d_0 + k = 6 + 2 = 8$ (degree) ($k = 2$)
- $D = 2(2D_0 + 2) + 2 = 18$ (diameter) ($k = 2$)
- 3D Torus ($162 \times 162 \times 162$): 4,251,528 nodes; $d = 6$ (degree)
- 3D Torus ($162 \times 162 \times 162$): $D = 81 + 81 + 81 = 243$ (diameter)

RDN(5,3)

- Base network: 5-node-Ring, $k = 3$
  - $m = 5$ (nodes)
  - $d_0 = 2$ (degree)
  - $D_0 = 2$ (diameter)
- The number of nodes
  - $k = 1$: $N_1 = 5 \times 5 \times 2 = 50$
  - $k = 2$: $N_2 = 50 \times 50 \times 2 = 5{,}000$
  - $k = 3$: $N_3 = 5{,}000 \times 5{,}000 \times 2 = 50{,}000{,}000$
- $d = d_0 + k = 2 + 3 = 5$ (degree) ($k = 3$)
- $D = 2[2(2D_0 + 2) + 2] + 2 = 30$ (diameter) ($k = 3$)
- 3D Torus ($500 \times 500 \times 200$): 50,000,000 nodes; $d = 6$ (degree)
- 3D Torus ($500 \times 500 \times 200$): $D = 250 + 250 + 100 = 600$ (diameter)

Cost Ratio (CR)

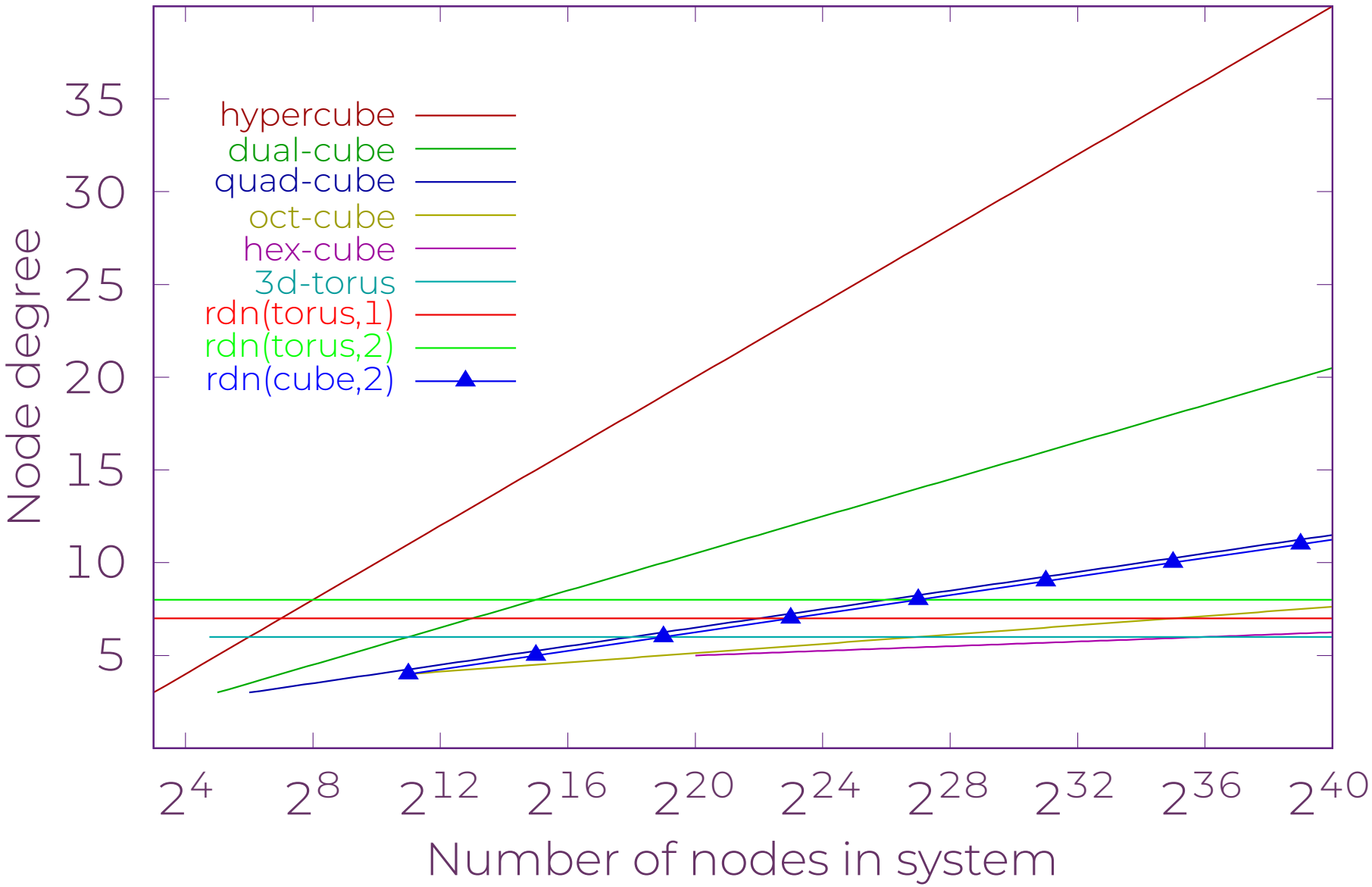
$$CR(G) = \frac{d(G) \times w_1 + D(G) \times w_2}{\log_2|(G)|}$$

INs	# Nodes	Degree	Diameter	CR
3D-Torus (10)	1,000	6	15	1.05
10-cube	1,024	10	10	1.00
RDN(5^2, 1)	1,250	5	10	0.73
RDN(3^3, 1)	1,458	7	8	0.71
3D-Torus (128)	2,097,152	6	192	4.72
21-cube	2,097,152	21	21	1.00
DC(10)	2,097,152	11	22	0.79
RDN(5^2, 2)	3,125,000	6	22	0.65
RDN(3^3, 2)	4,251,528	8	18	0.59
RDN(5, 3)	50,000,000	5	30	0.68

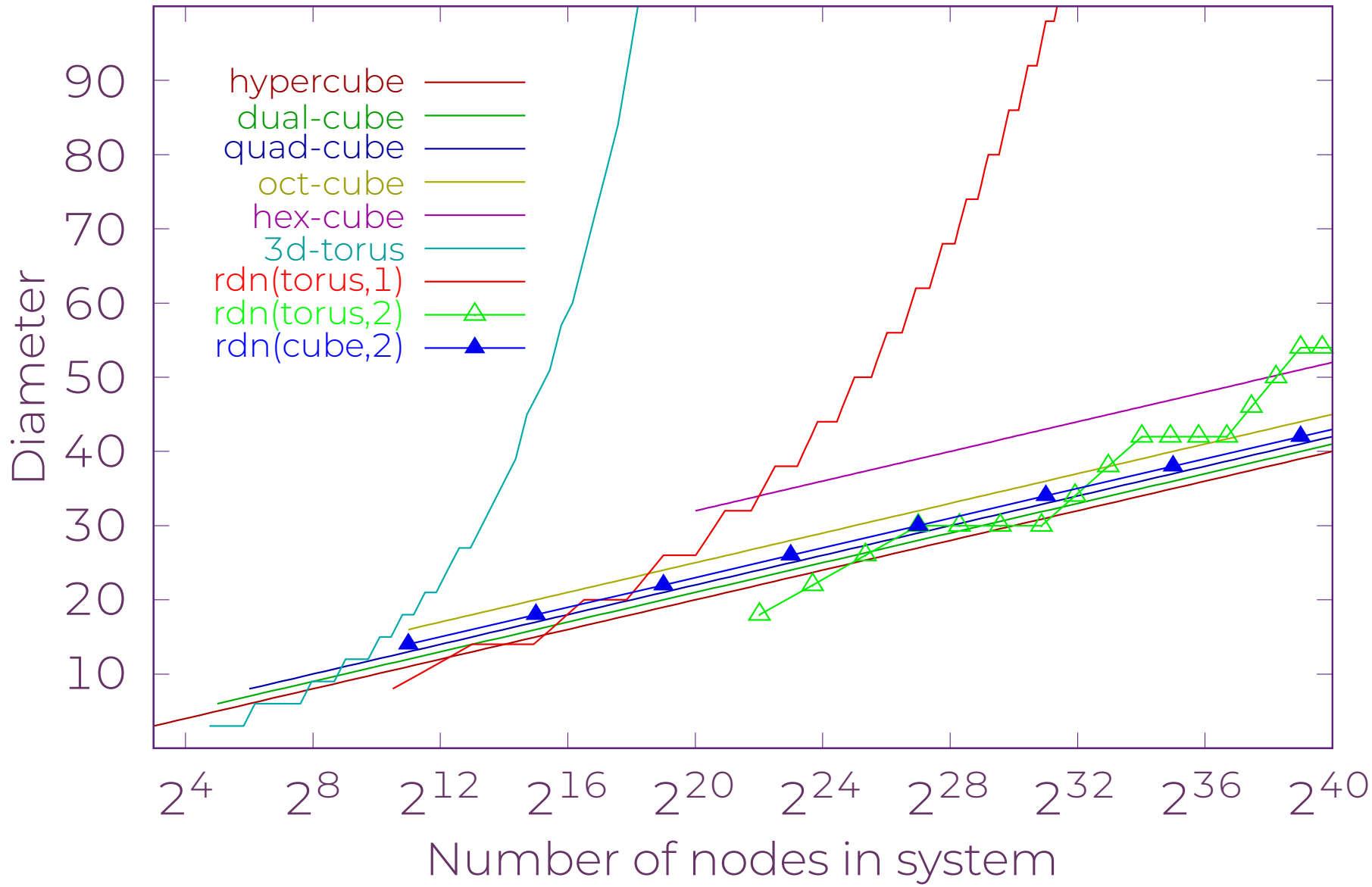
Topological Properties of INs

Network	Nodes	Degree	Diameter
3D Torus	n^3	6	$3\lfloor n/2 \rfloor$
n-cube	2^n	n	n
CCC(n)	$n2^n$	3	$2n + \lfloor n/2 \rfloor - 2$
HCN(n)	2^{2n}	$n + 1$	$n + \lfloor (n+1)/3 \rfloor + 1$
Dual-Cube(m)	2^{2m+1}	$m + 1$	$2(m + 1)$
Metacube(k, m)	2^{m2^k+k}	$m + k$	$(m + 1)2^k$
RDN(m, k)	$(2m)^{2^k}/2$	$d_0 + k$	$2^k D_0 + 2^{k+1} - 2$

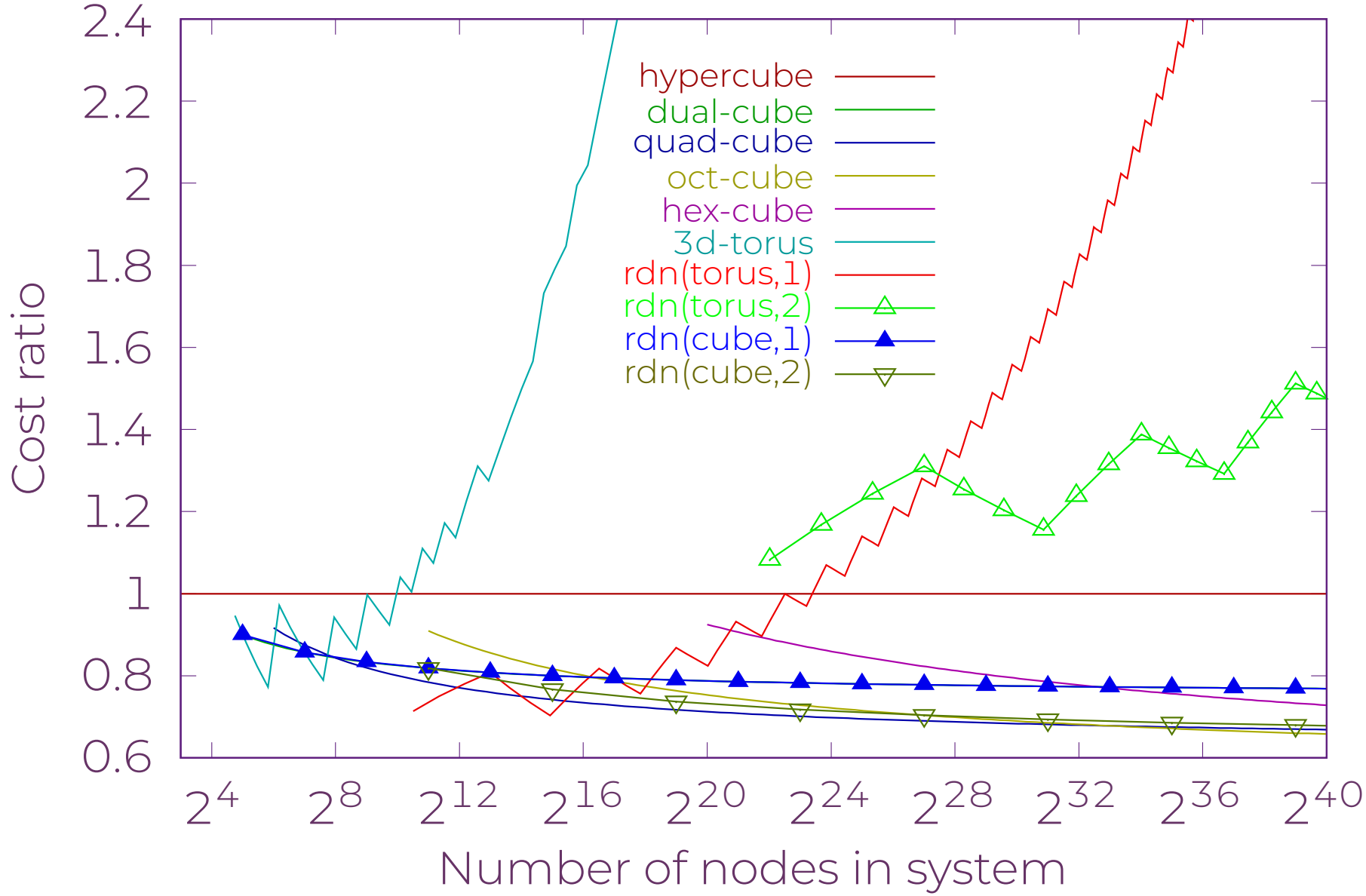
Degree Comparison



Diameter Comparison



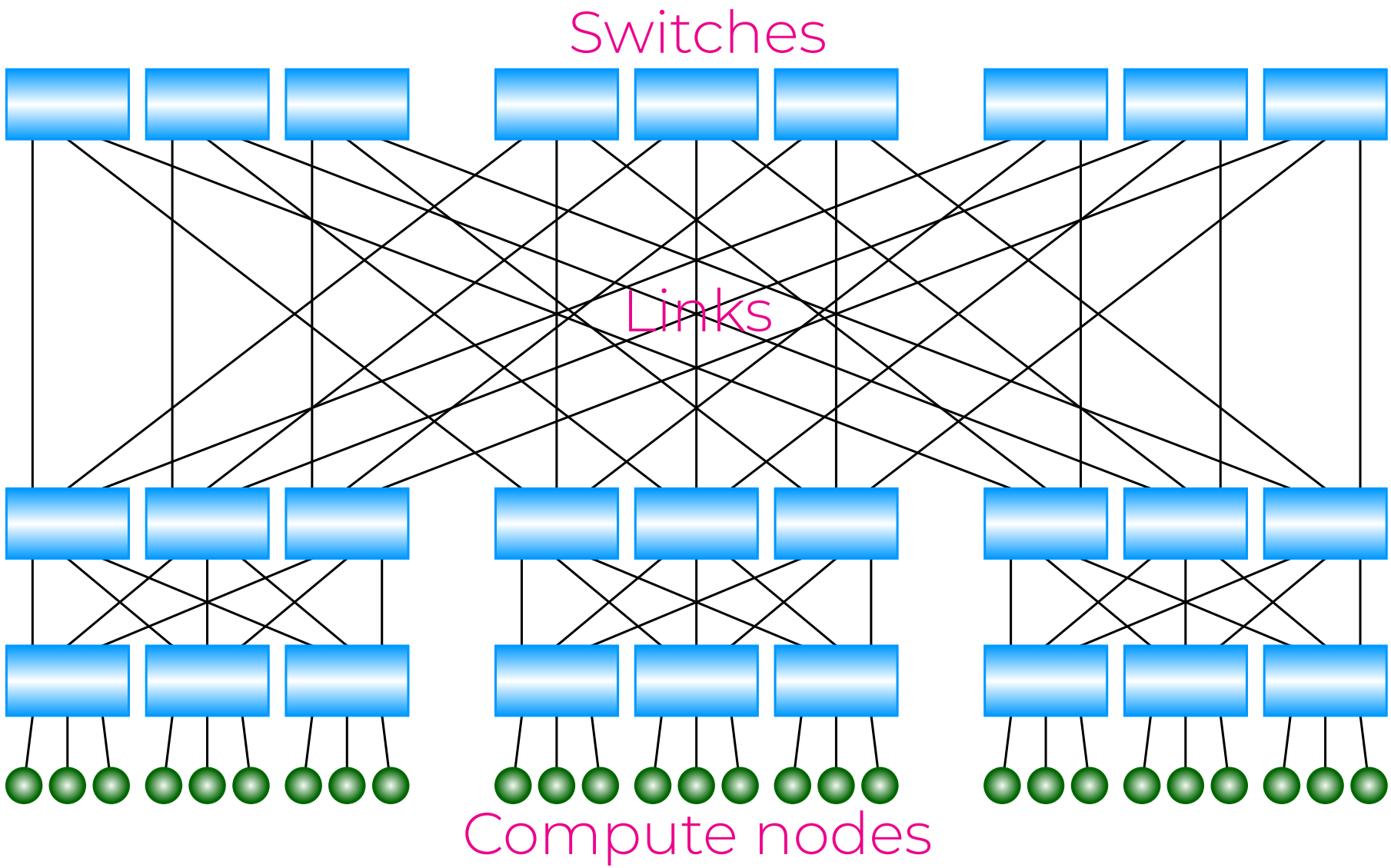
Cost Ratio Comparison



MiKANT — Mirrored K-Ary N-Tree

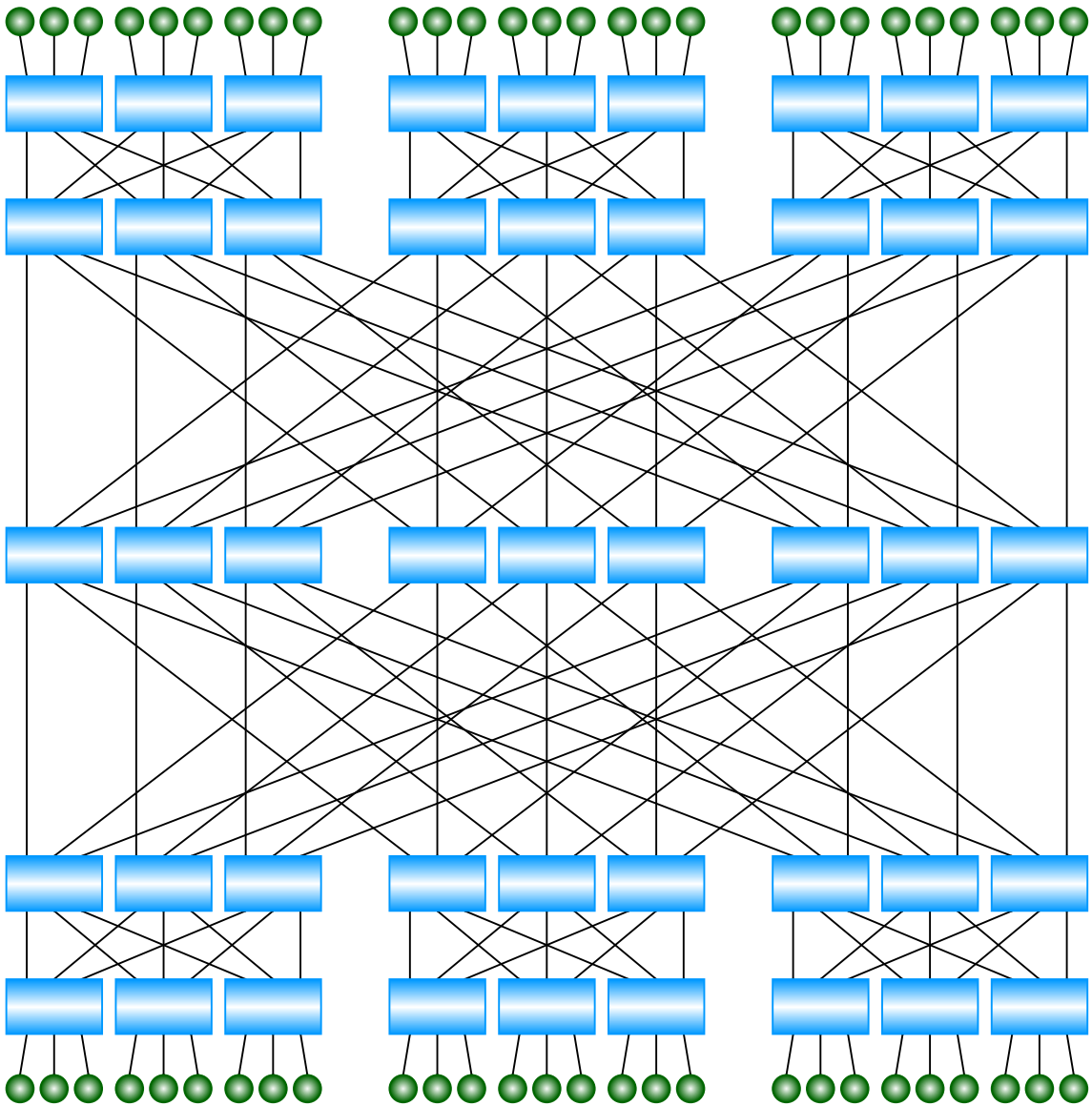
MiKANT

Dynamic IN: Fat Tree (3-ary 3-tree)



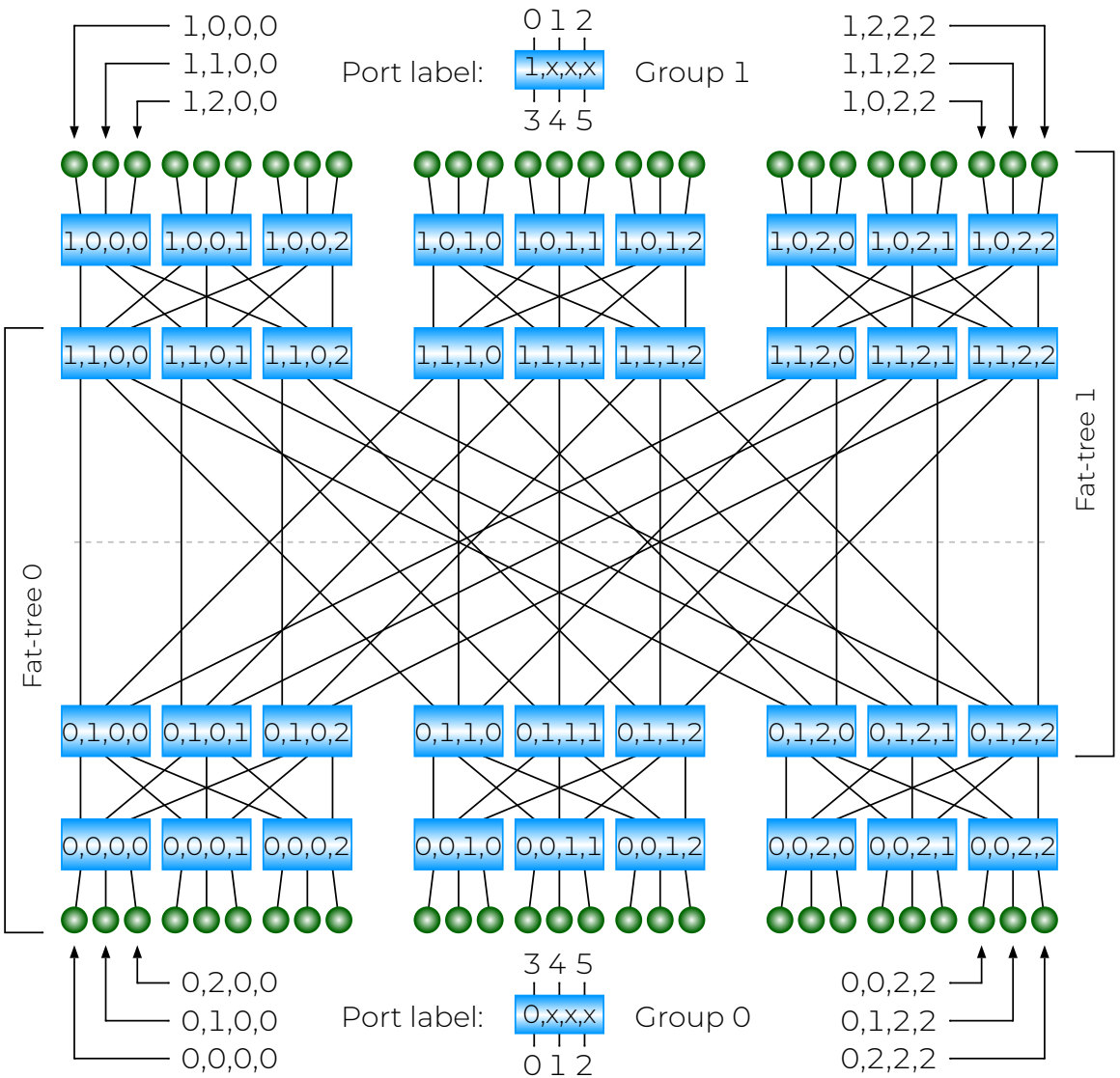
Root switches have fewer ports than others

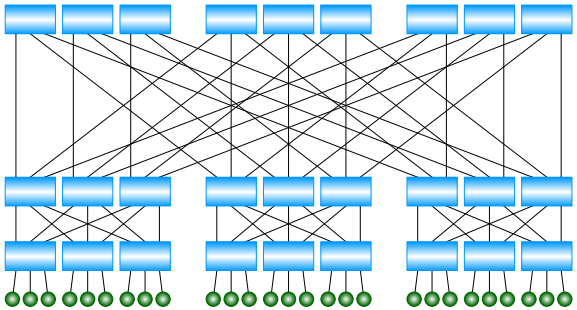
Bidirectional Clos Network



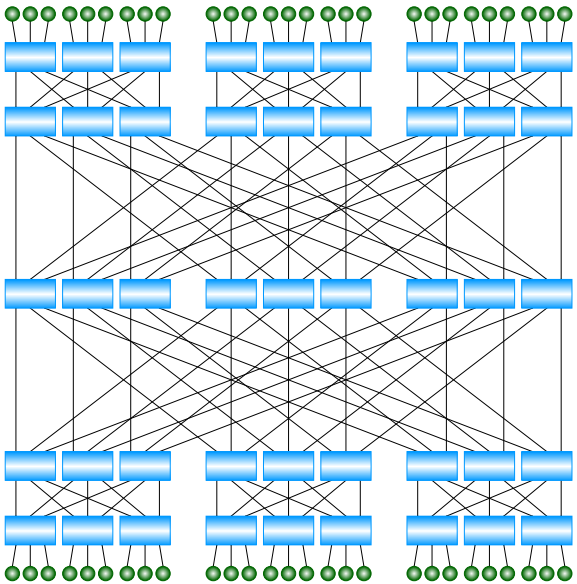
High switch cost
High link cost
Long distance

Mirrored K-Ary N-Tree MiKANT(3,3)



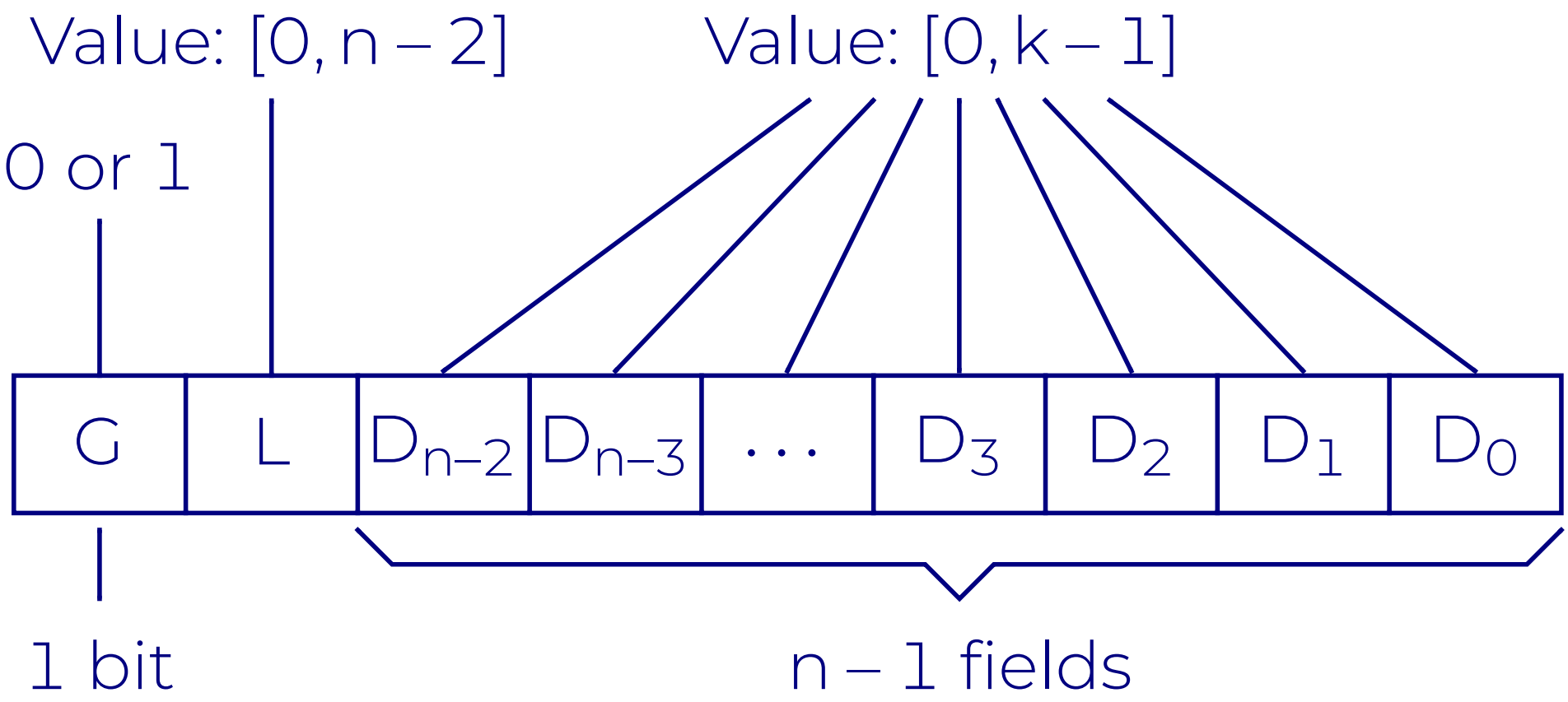


3-ary 3-tree



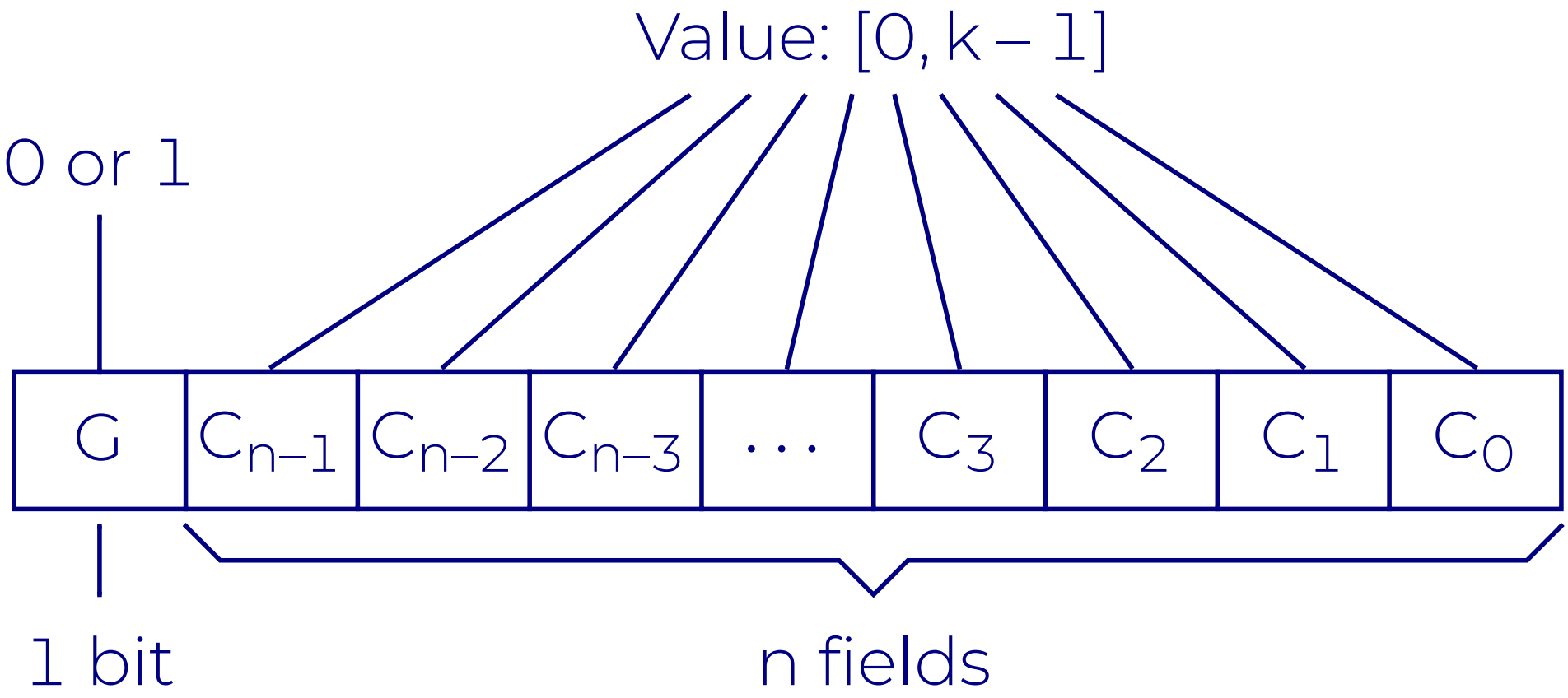
Clos 3-ary 3-tree

Switch Address Format of MiKANT(n, k)



G: Group ID; L: Level (Stage) ID; D: Switch ID

Node Address Format of MiKANT(n, k)



G: Group ID; C: Node ID; $C_{n-2} \ldots C_0$: Switch ID

Switch-Switch Connections

A switch

$$\langle G, L, D_{n-2}, \ldots, D_{L+1}, D_L, D_{L-1}, \ldots, D_0 \rangle$$

will connect to switches

$$\langle G, L+1, D_{n-2}, \ldots, D_{L+1}, *, D_{L-1}, \ldots, D_0 \rangle$$

if $0 \leq L \leq n-3$; otherwise ($L = n-2$) to switches

$$\langle \overline{G}, L, *, D_{n-3}, \ldots, D_1, D_0 \rangle$$

Switch-Compute Node Connections

There is a link between a switch

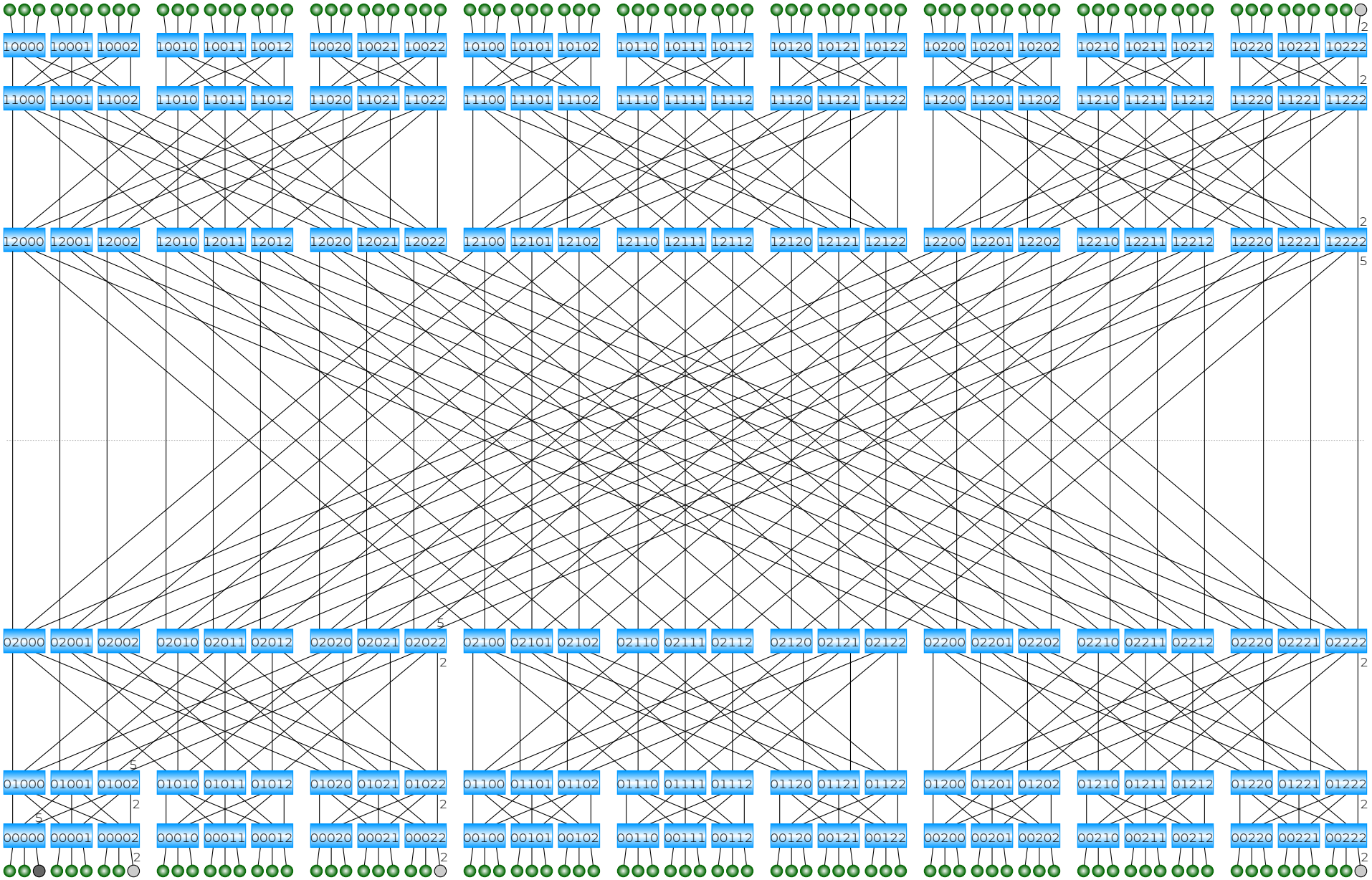
$$\langle G, 0, D_{n-2}, \ldots, D_1, D_0 \rangle$$

and a compute node

$$\langle G, C_{n-1}, C_{n-2}, \ldots, C_1, C_0 \rangle$$

if $D_i = C_i$ for all $i \in \{n-2, \ldots, 1, 0\}$.

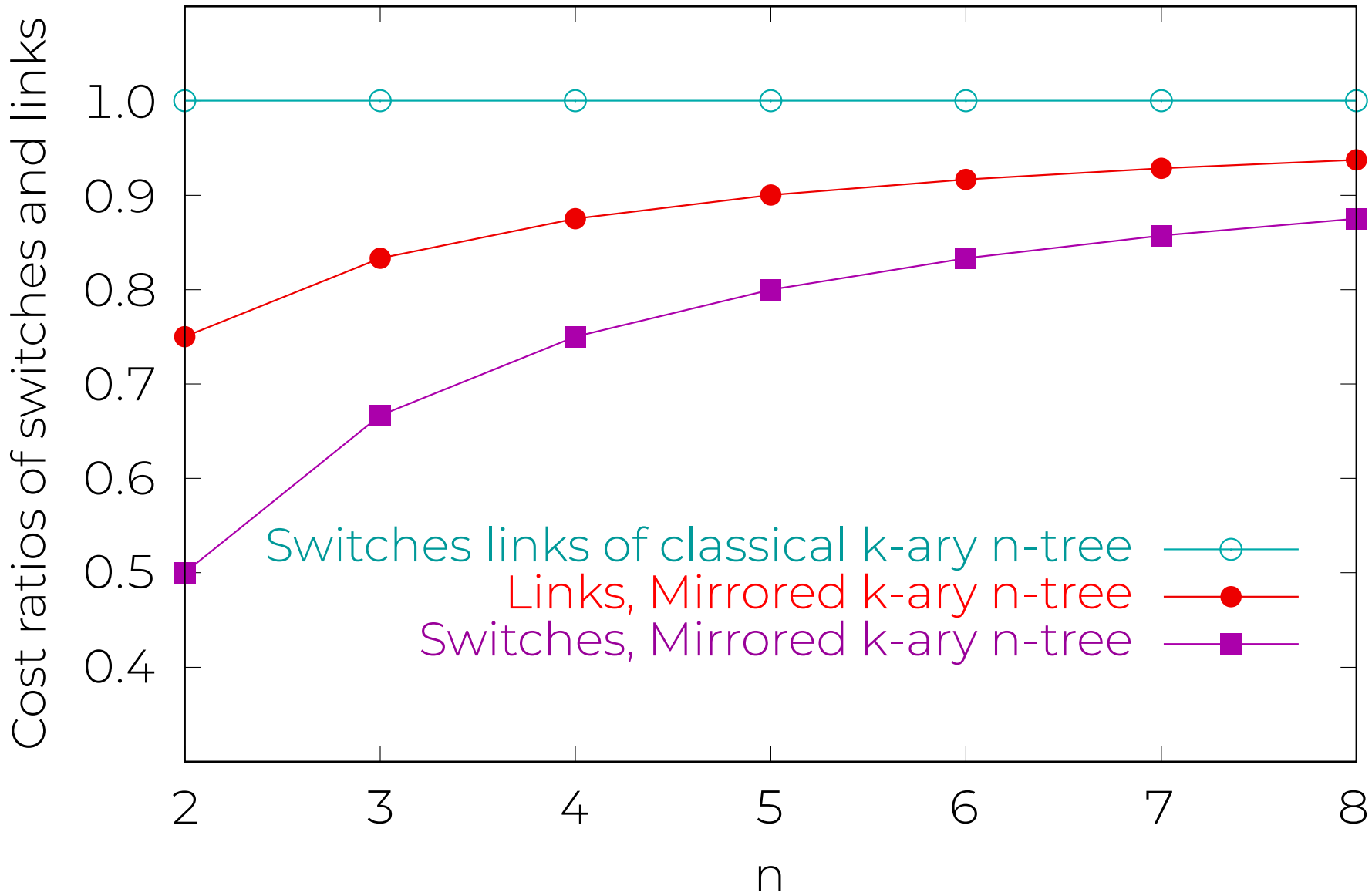
Mirrored K-Ary N-Tree MiKANT(3,4)



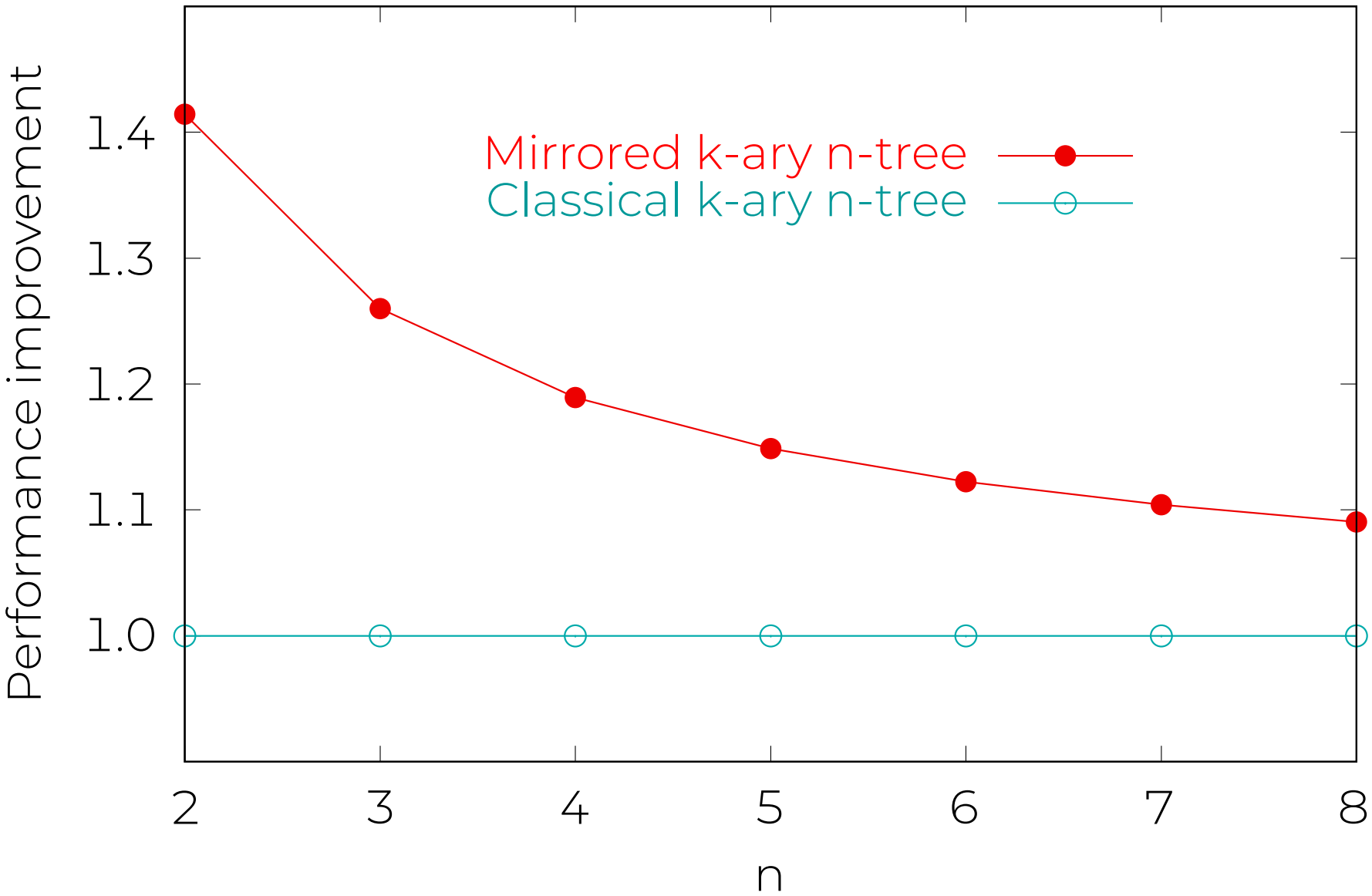
Comparison of Topological Properties

	Classical k-ary n-tree	Clos k-ary n-tree	Mirrored k-ary n-tree
# nodes	k^n	$2k^n$	$2k^n$
# switches	nk^{n-1}	$(2n-1)k^{n-1}$	$(2n-2)k^{n-1}$
# links	nk^n	$2nk^n$	$(2n-1)k^n$
Degree	$2k$	$2k$	$2k$
Diameter	$2n$	$2n$	$2n$
Bisection	$k^n/2$	$k^n/2$	$k^n/2$
Ave. dist	$2n - \frac{2}{k-1} + \frac{2}{(k-1)k^n}$	$2n - \frac{1}{k-1} + \frac{1}{(k-1)k^n}$	$2n - \frac{1}{k-1} + \frac{1}{(k-1)k^n} - \frac{1}{2}$

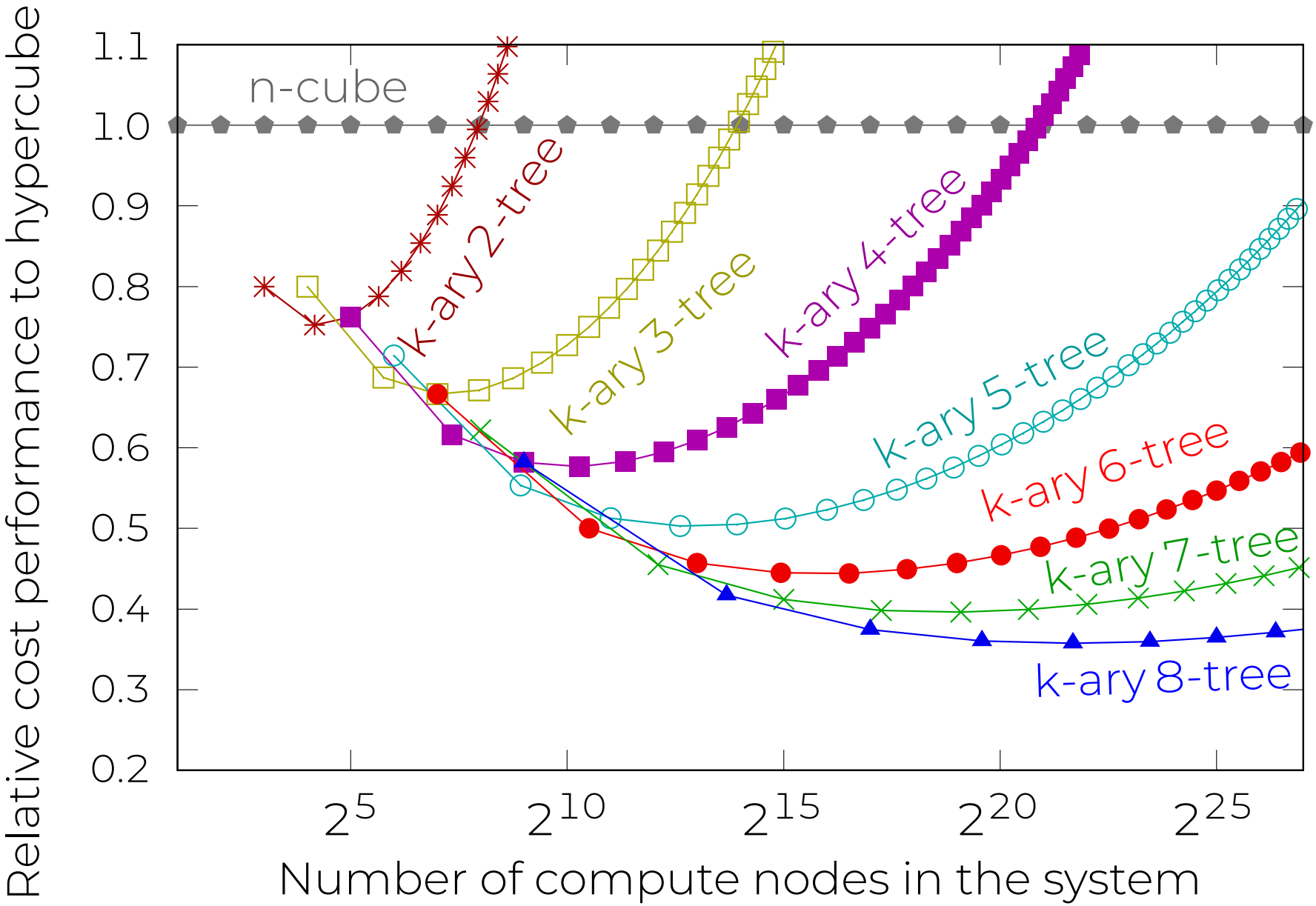
Cost Ratios of Links and Switches



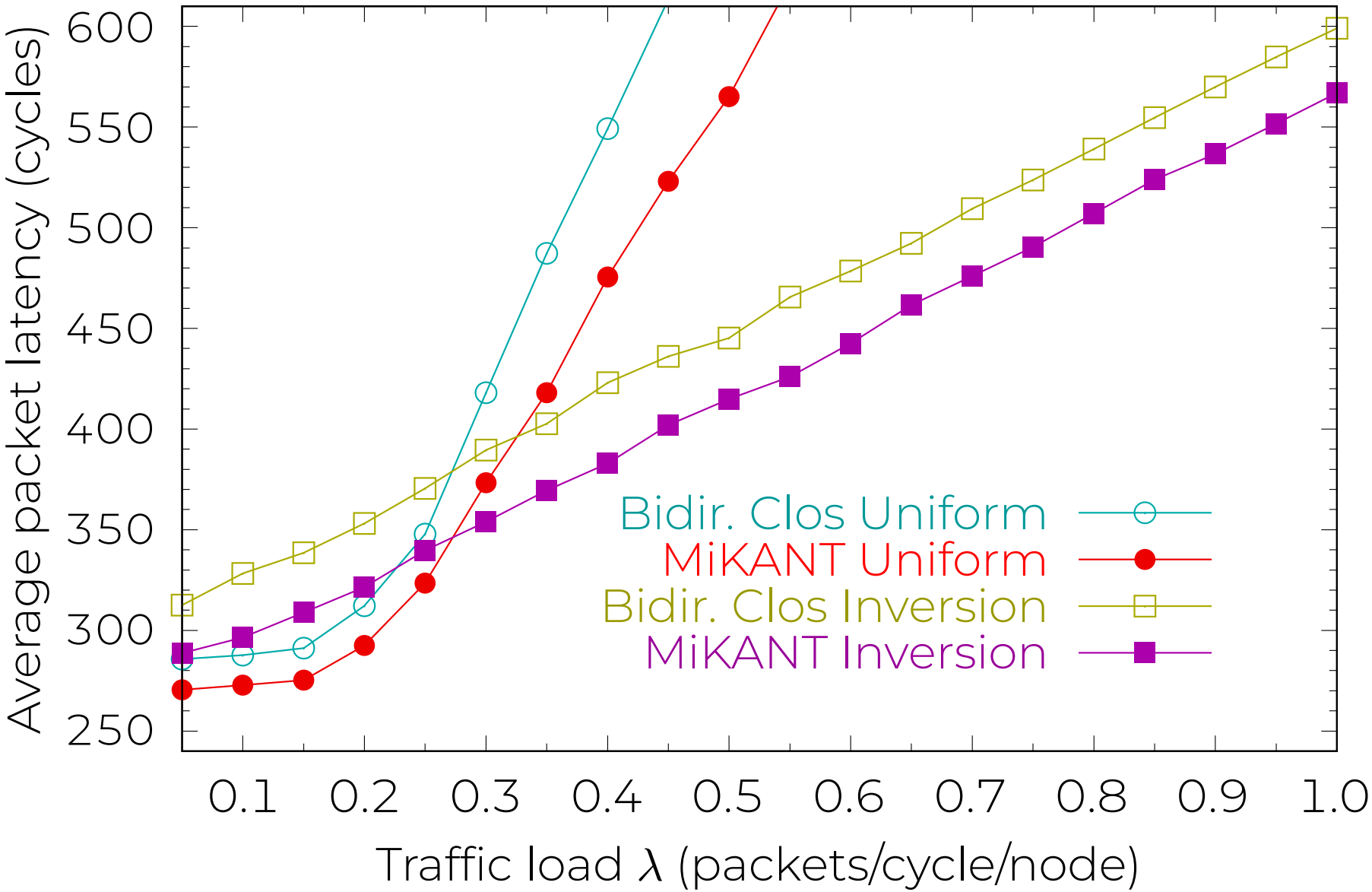
Performance Improvement of MiKANT



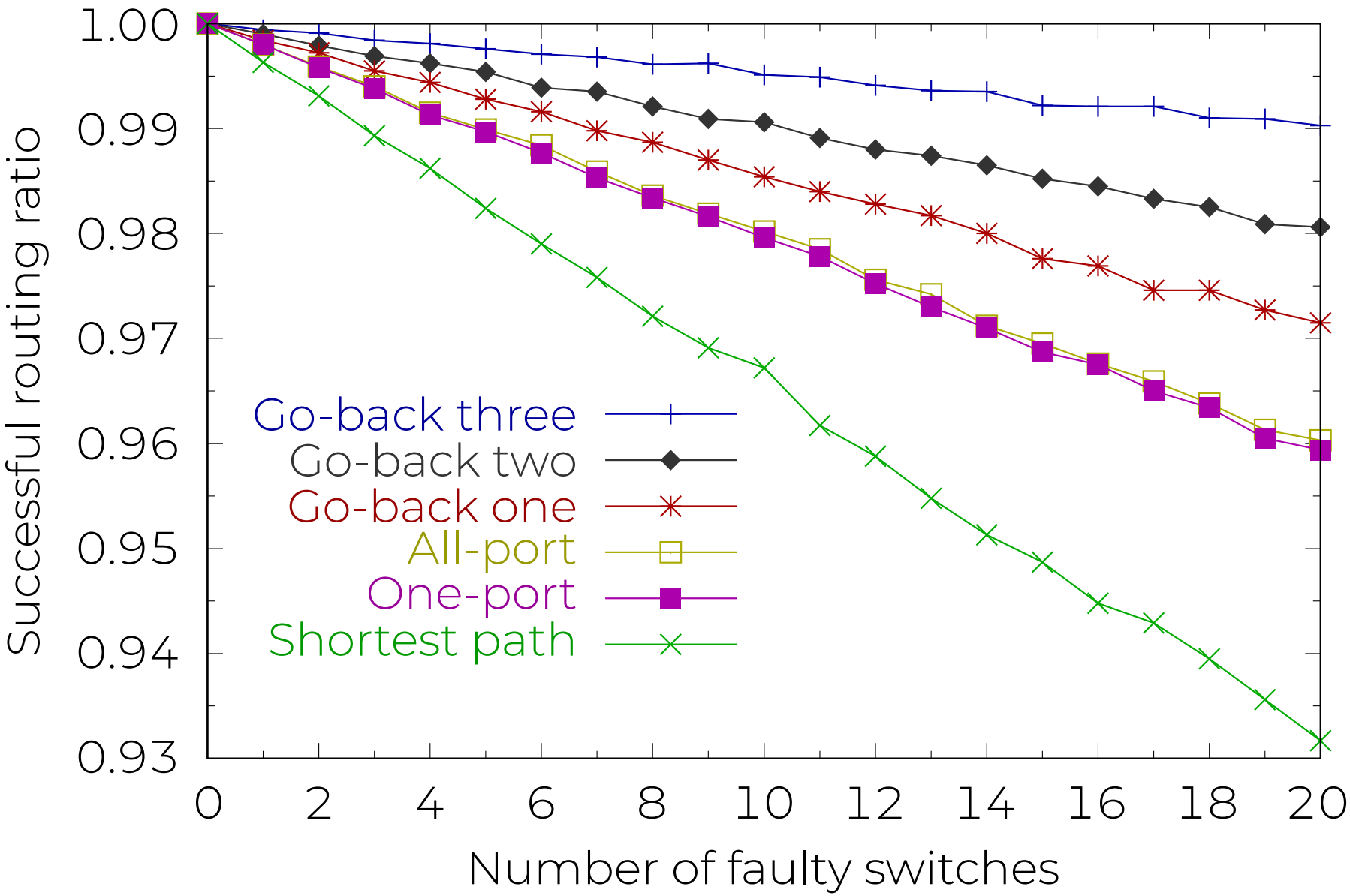
Relative Cost Performance to Hypercube



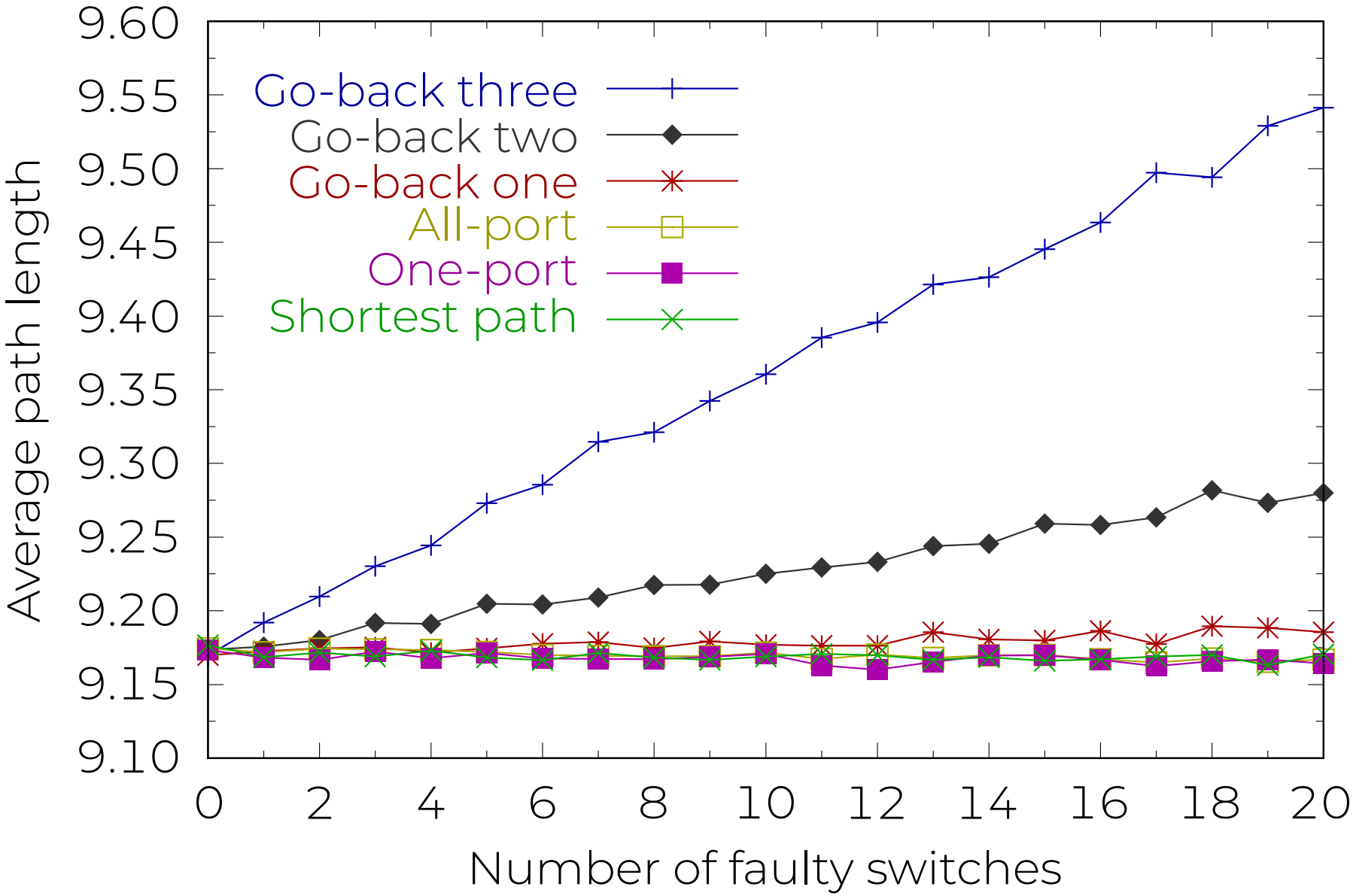
Average Packet Latencies (Clock Cycles)



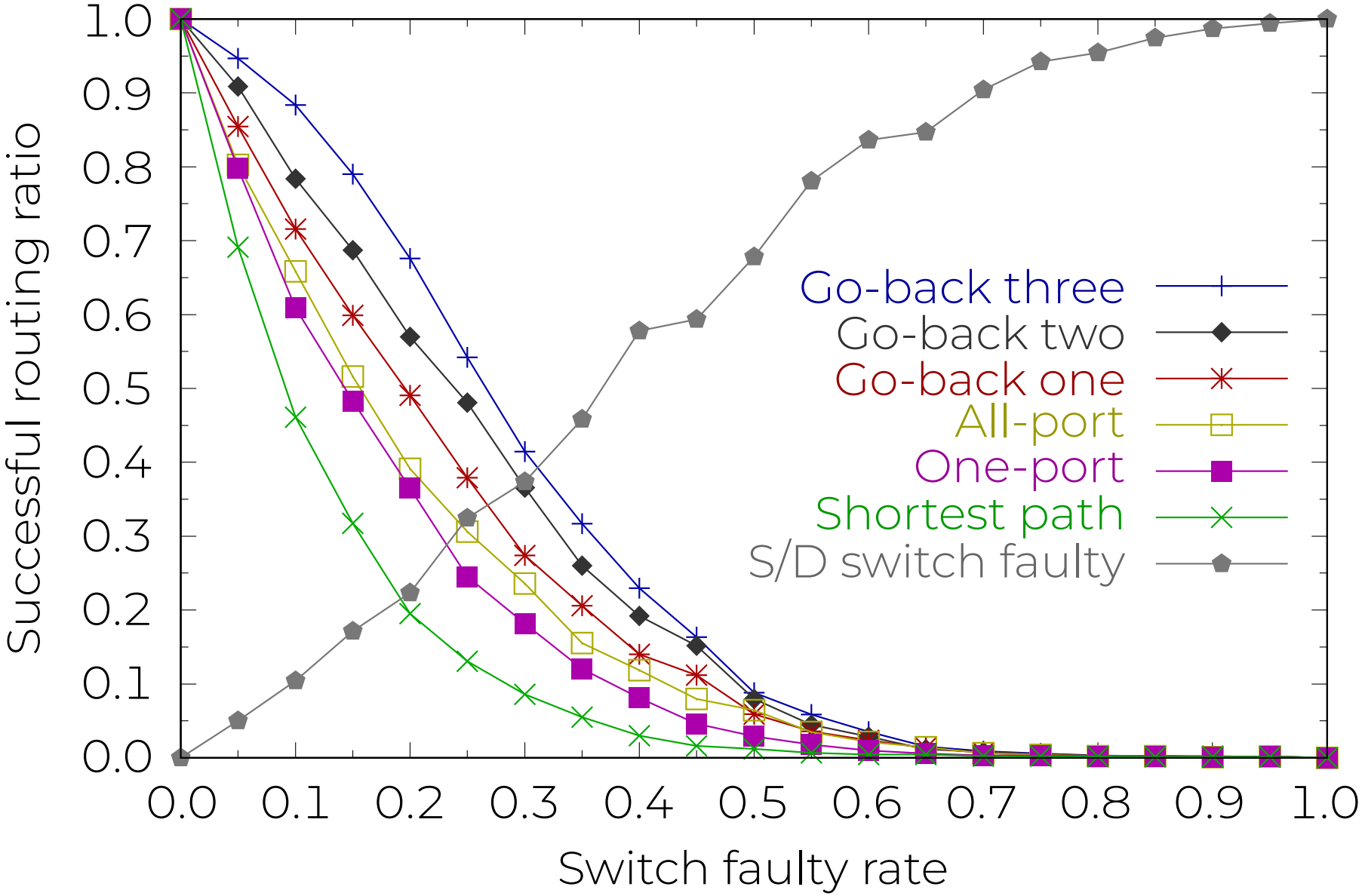
Fault Tolerance — Performance



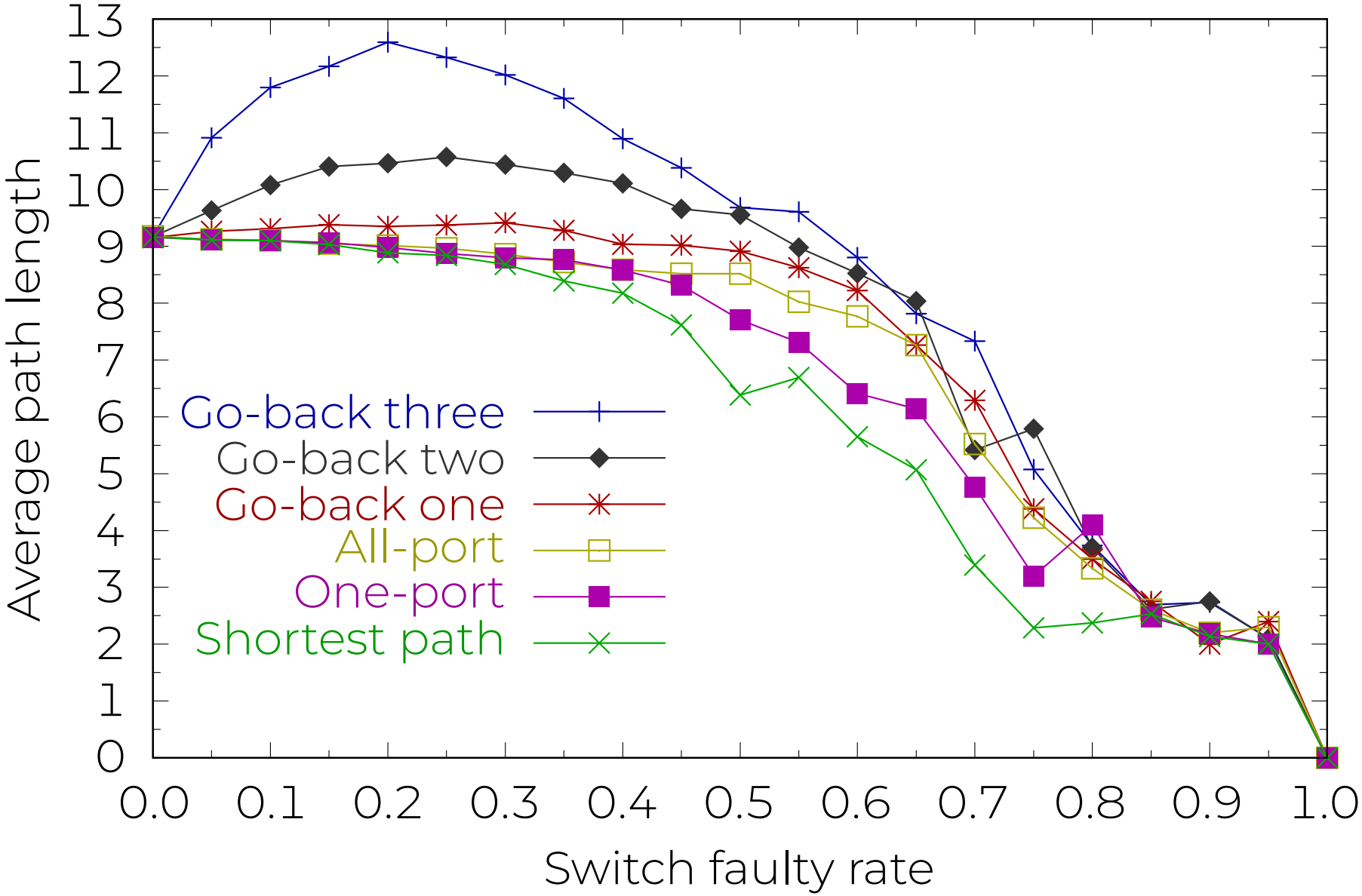
Fault Tolerance — Path Length



Fault Tolerance — Performance



Fault Tolerance — Path Length



Summary

- The IN is at the center of supercomputers.
- It is built with switches (routers) and cables that connect ports of switches by following some topologies.
  - Switches: Infiniband, Gigabit Ethernet, or custom
- A good interconnection network should use a small number of links (low cost of switch) and have a short diameter (fast communication).
- IN topologies for building supercomputers
  - Dual-Cube, Metacube, and KMS-Cube
  - RDN — Recursive Dual-Net
  - MiKANT — Mirrored K-Ary N-Tree

Research Topics on Parallel Systems

- Shortest-path routing algorithm
- Multicast and Broadcast algorithms
- Collective communication
- Disjoint paths and Hamiltonian
- Fault-tolerant routing
- Algorithmic design
  - Parallel prefix computation
  - Parallel sorting
- Matrix multiplication and Linpack (Linear algebra)
- Parallel (programming languages and) compilers

Performance Evaluation

Exercise

Define Cost Ratio $CR(G) = \frac{d(G) \times w_1 + D(G) \times w_2}{\log_2 |(G)|}$

d(G): the node degree of G (the number of links of a node)
D(G): the diameter of G
|(G)|: the total number of nodes in G

Question: Let $w_1 = w_2 = 50\%$. Calculate CRs for

(a) A 256-node Ring
(b) A 256-node Completely Connected Network
(c) A 256-node 8-Cube
(d) A 256-node 16×16 Two-Dimensional Torus
(e) A 512-node $8 \times 8 \times 8$ Three-Dimensional Torus
(f) A 512-node Dual-Cube DC(4)
(g) A 512-node Mirrored 4-ary 4-tree MiKANT(4, 4)

Report submission: **https://hoppii.hosei.ac.jp/portal**

Recommendations and CFP (CANDAR23)

- Format of conference/journal papers
  - Use LaTeX to prepare the manuscript
  - Draw block diagrams with tgif
  - Generate graphs with Gnuplot
- Related international conference
  - CANDAR 2023, Nov. 28 - Dec. 1, Matsue

    The Eleventh International Symposium on Computing and Networking

    Abstract submission due: July 20, 2023
    Paper (PDF) submission due: July 25, 2023