Centralized (Indirect) switching networks

Computer Architecture

AMANO, Hideharu

Textbook pp.92～130

Centralized interconnection networks

- Symmetric:
  - MIN (Multistage Interconnection Networks)
  - Each node is connected with equal latency and bandwidth
- Asymmetric:
  - Fat-tree, base-m n-cube, etc.
  - Locality of communication can be used.

Properties of MIN

- Throughput for random communication
- Permutation capability
- Partition capability
- Fault tolerance
- Routing

MIN（Multistage Interconnection Network)

- Multistage connected switching elements form a large switch.
- Symmetric
- Smaller number of cross-points, high degree of expandability
- Bandwidth is often degraded
- Latency is stretched

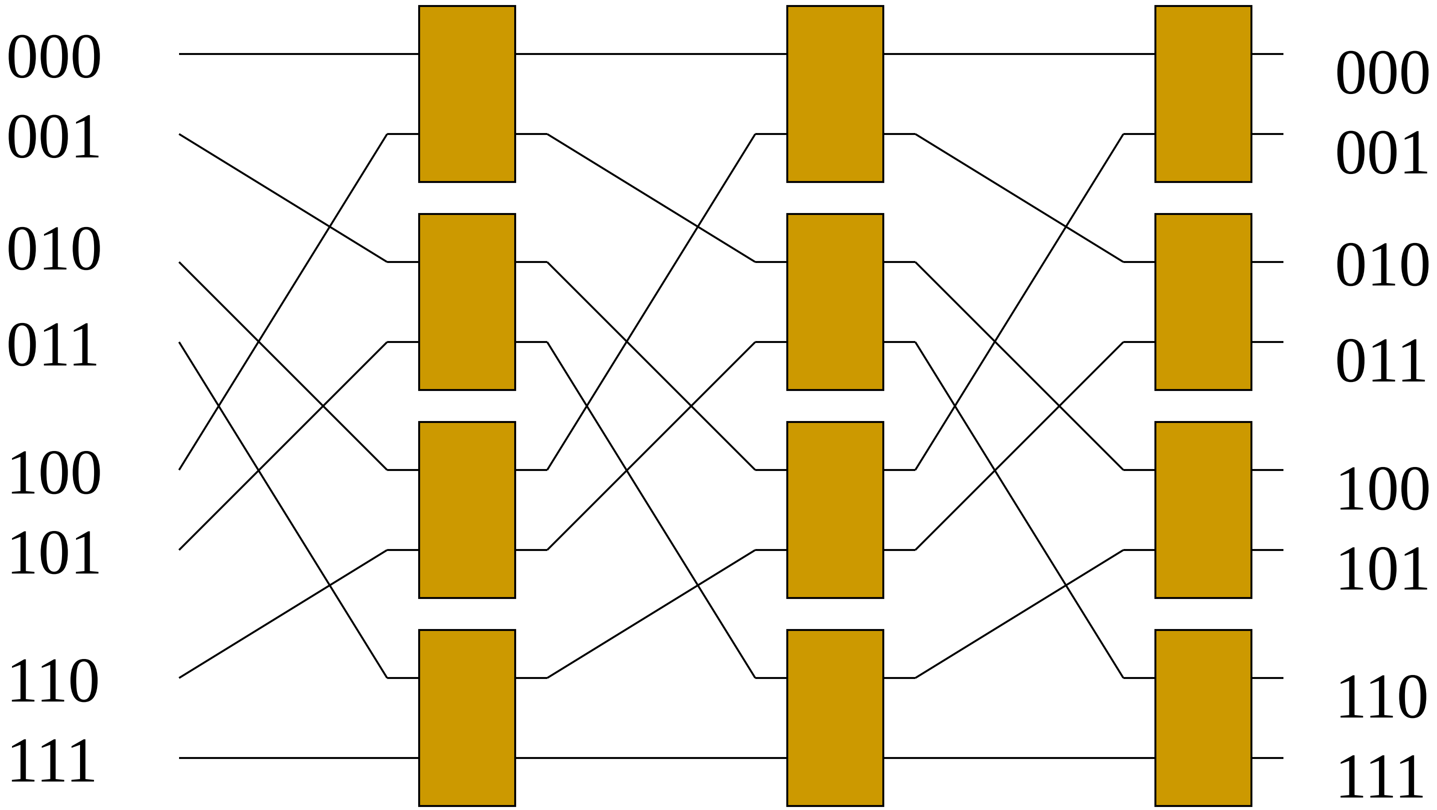
Classification of MIN

- Blocking network：Conflict may occur for destination is different：NlogN type standard MIN,πnetwork,
- Re-arrangeable：Conflict free scheduling is possible：Benes network、Clos network（rearrangeable configuration）
- Non-blocking：Conflict free without scheduling：Clos network (non-blocking configuration)、Batcher-Banyan network

Blocking Networks

- Standard NlogN networks
  - Omega network
  - Generalized Cube
  - Baseline
- Pass through ratio (throughput) is the same.
- Π network

Omega network



- The number of switching element（2x2, in this case）is 1／2NxLogN

Perfect Shuffle

- Rotate to left

000	000
001	010
010	100
011	110
100	001
101	011
110	101
111	111

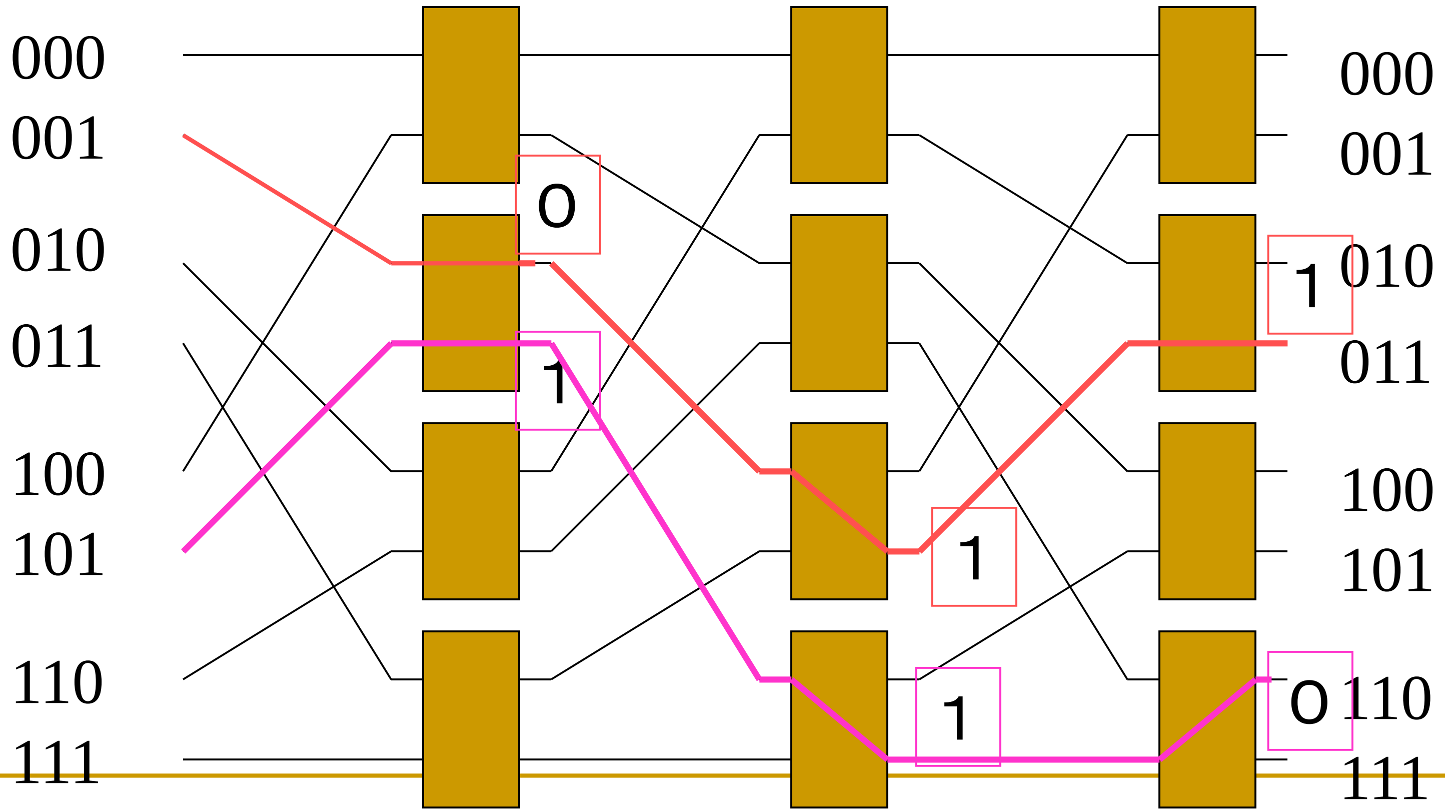
Inverse Shuffle
Rotate to right

Destination Routing

Check the destination tag from MSB
If 0 use upper link, else use lower link.

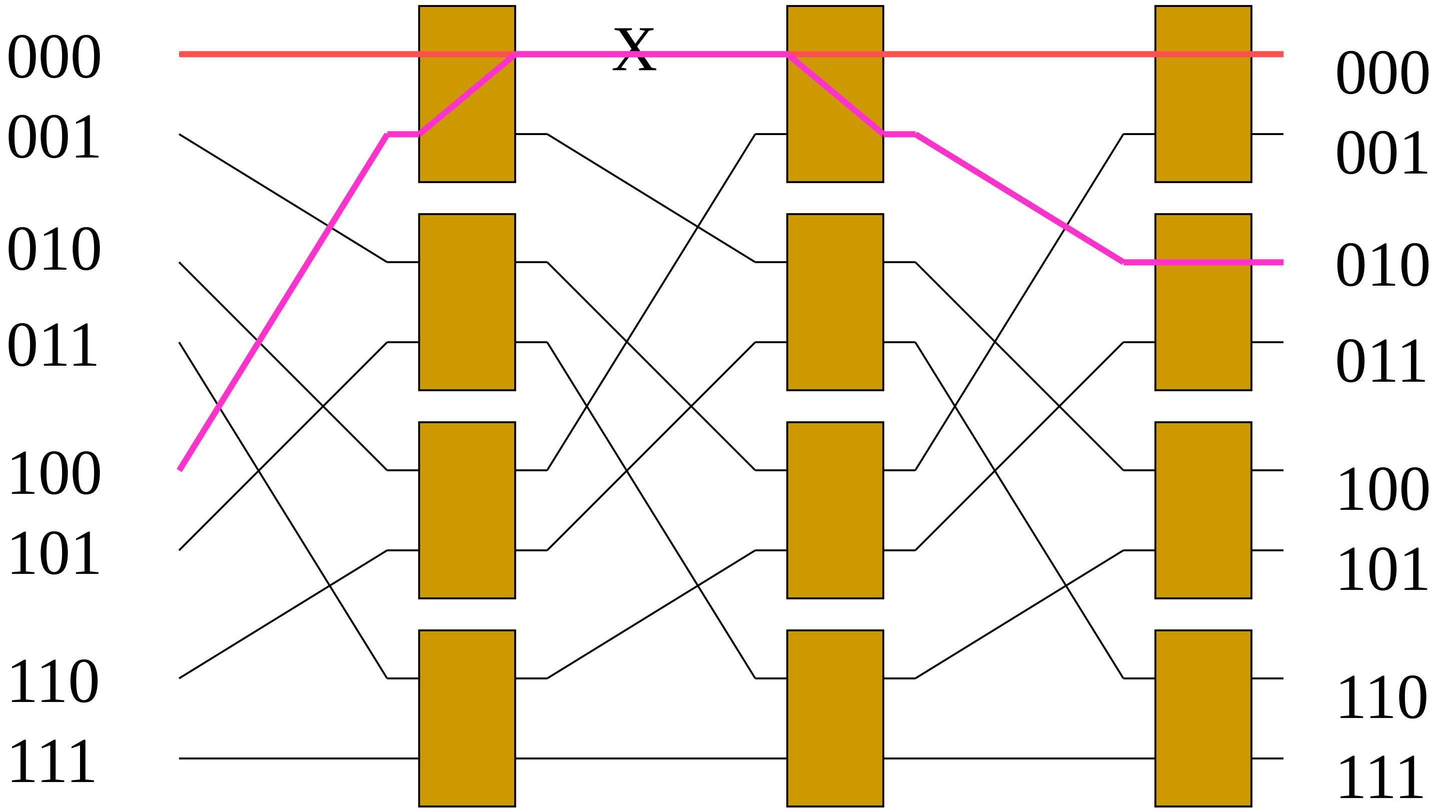
1→3

5→6



Blocking Property

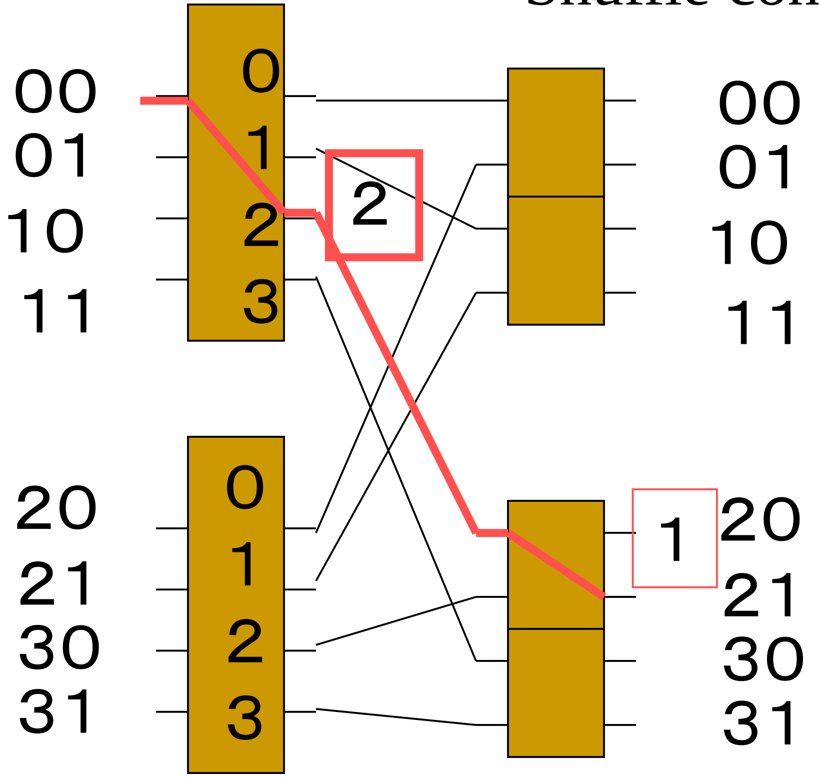
0→0
4→2



- For different destination, multiple paths conflict

For using large switching elements (Delta network)

Shuffle connection is also used.

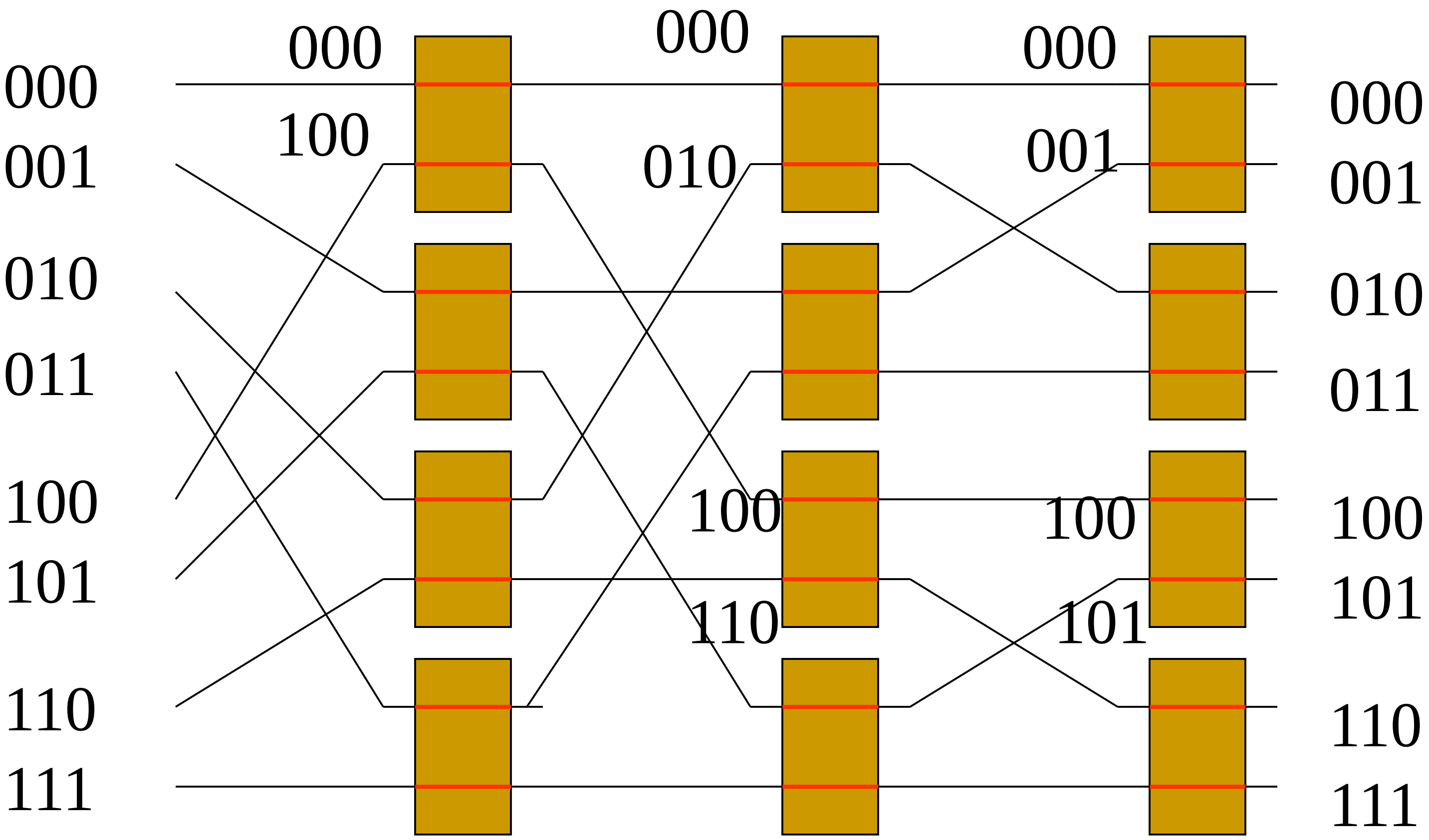


- In the current art of technology, 8x8 (4x4) crossbars are advantageous.

Omega network

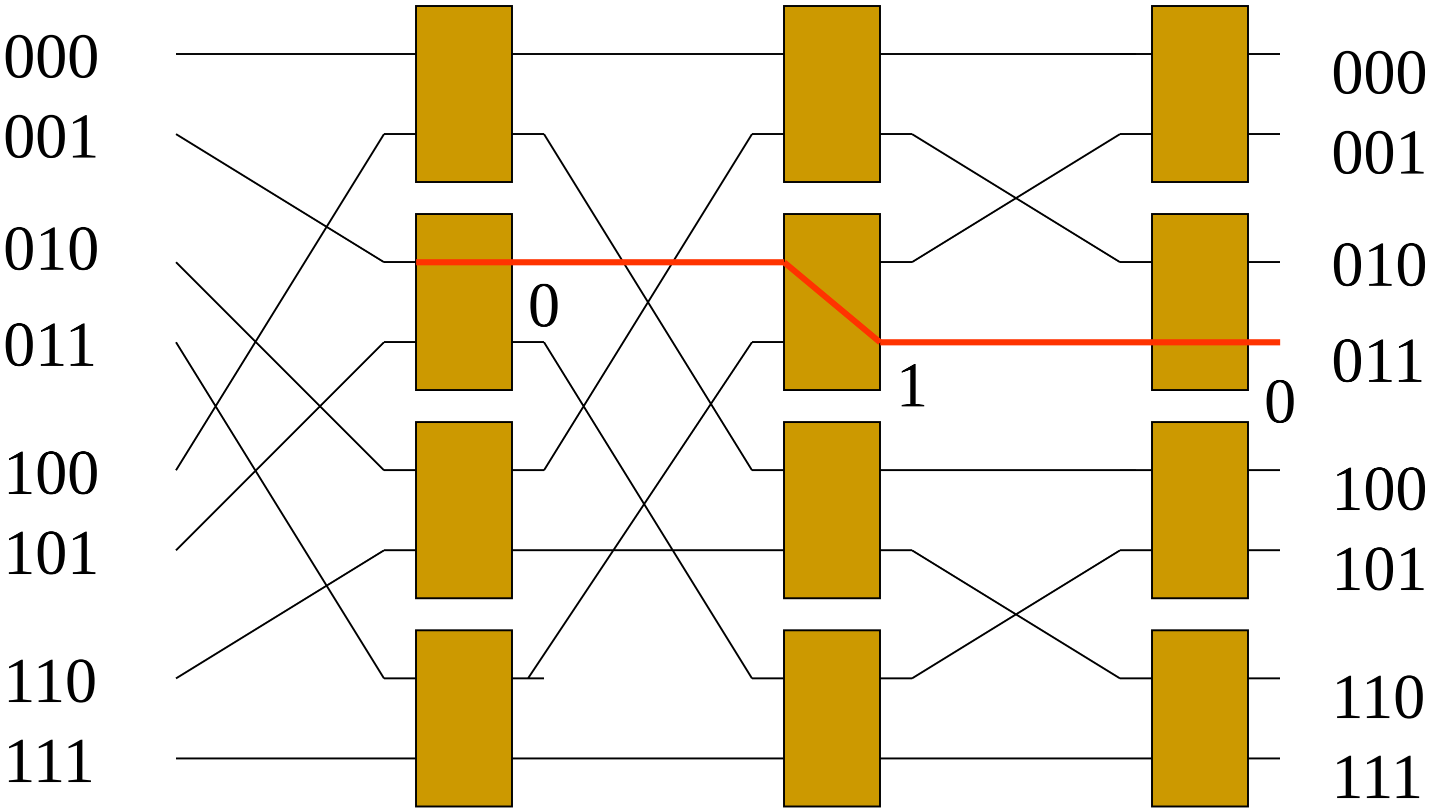
- The same connection is used for all stages.
- Destination routing
- A lot of useful permutations are available.
- Problems on partitioning and expandability.

Generalized Cube



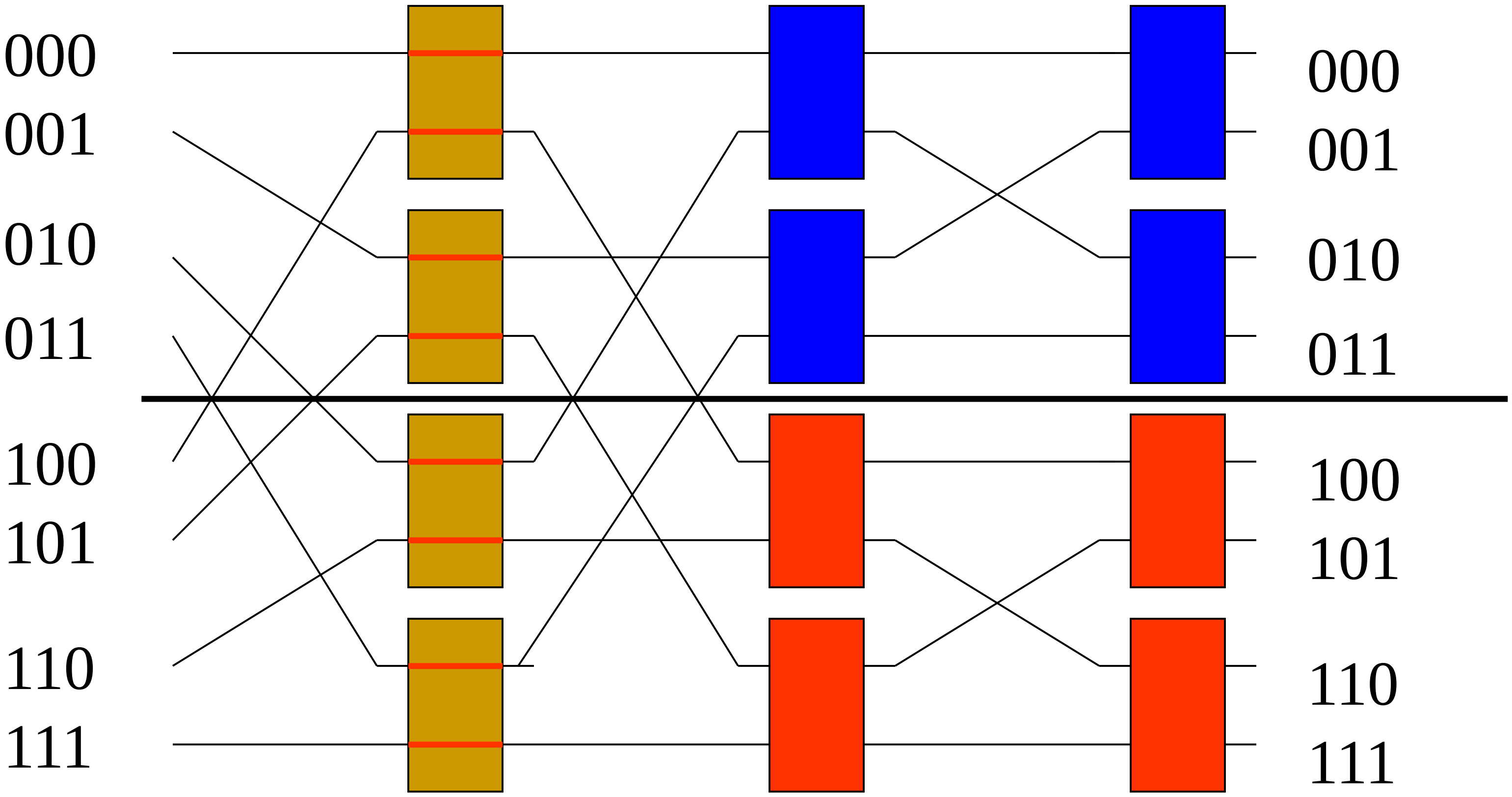
Links labeled with 1bit distance are connected to the same switching element.

Routing in Generalized Cube



The source label and destination label is compared (Ex-Or) : 001→011
Same(0) : Straight   Different (1) : Exchange   010

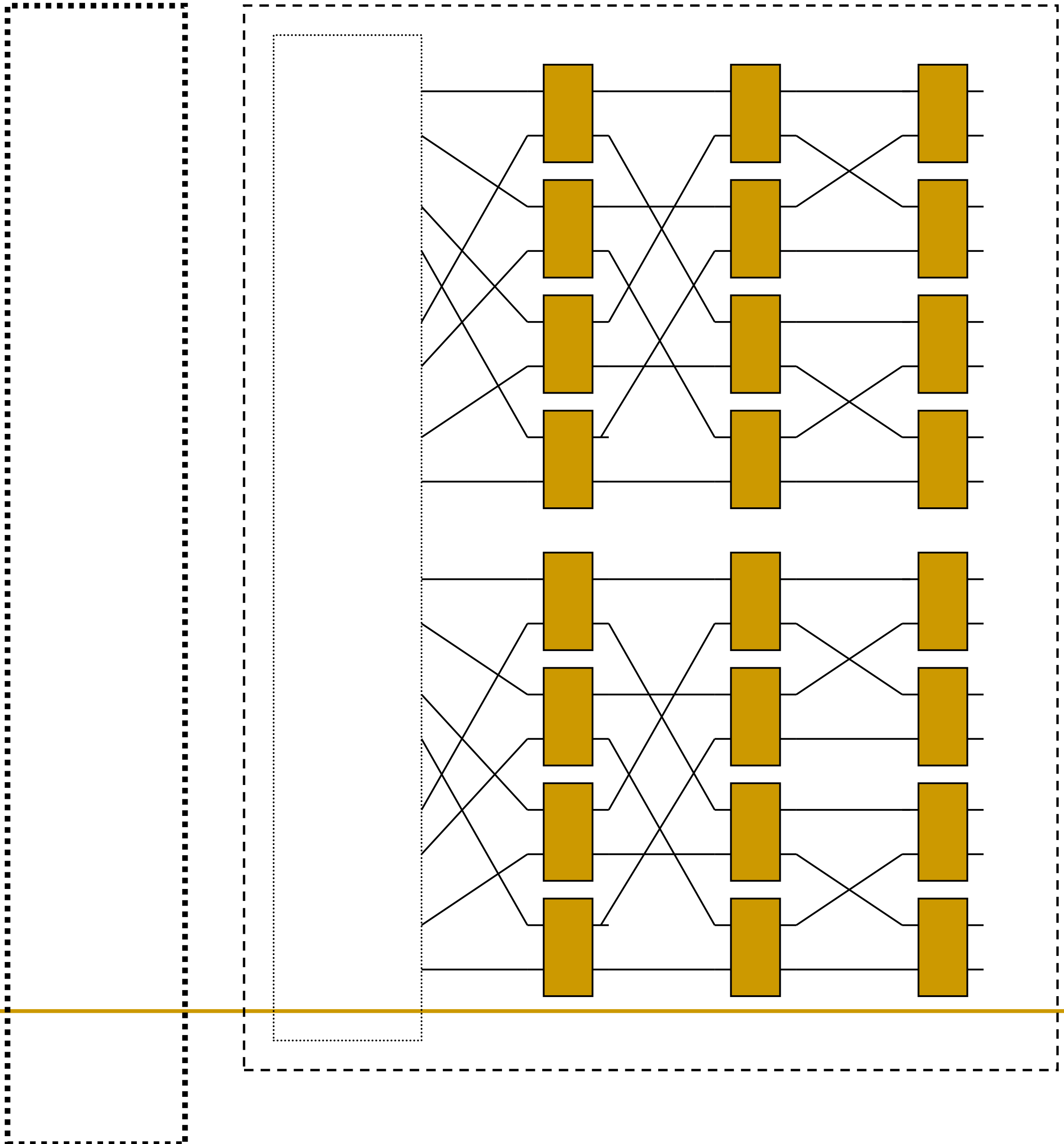
Partitioning



The communication in the upper half never disturbs the lower half.

Expandability

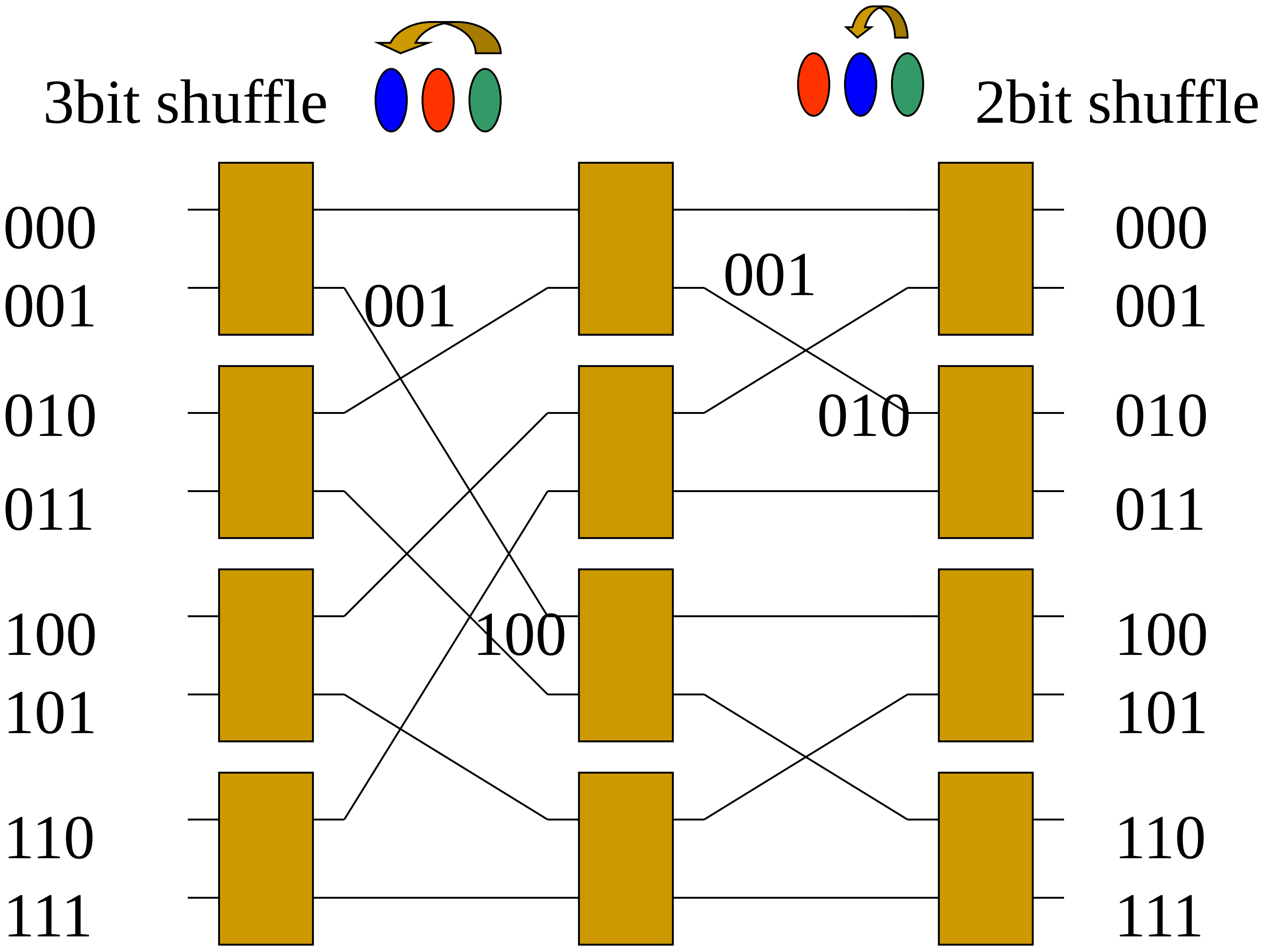
A size of network can be used as an element of larger size networks



Generalized Cube

- Destination routing cannot be applied.
- The routing tag is generated by exclusive or of source label and destination label.
- Partitioning
- Expandability

Baseline Network



3bit shuffle

2bit shuffle

000
001
010
011
100
101
110
111

001

001

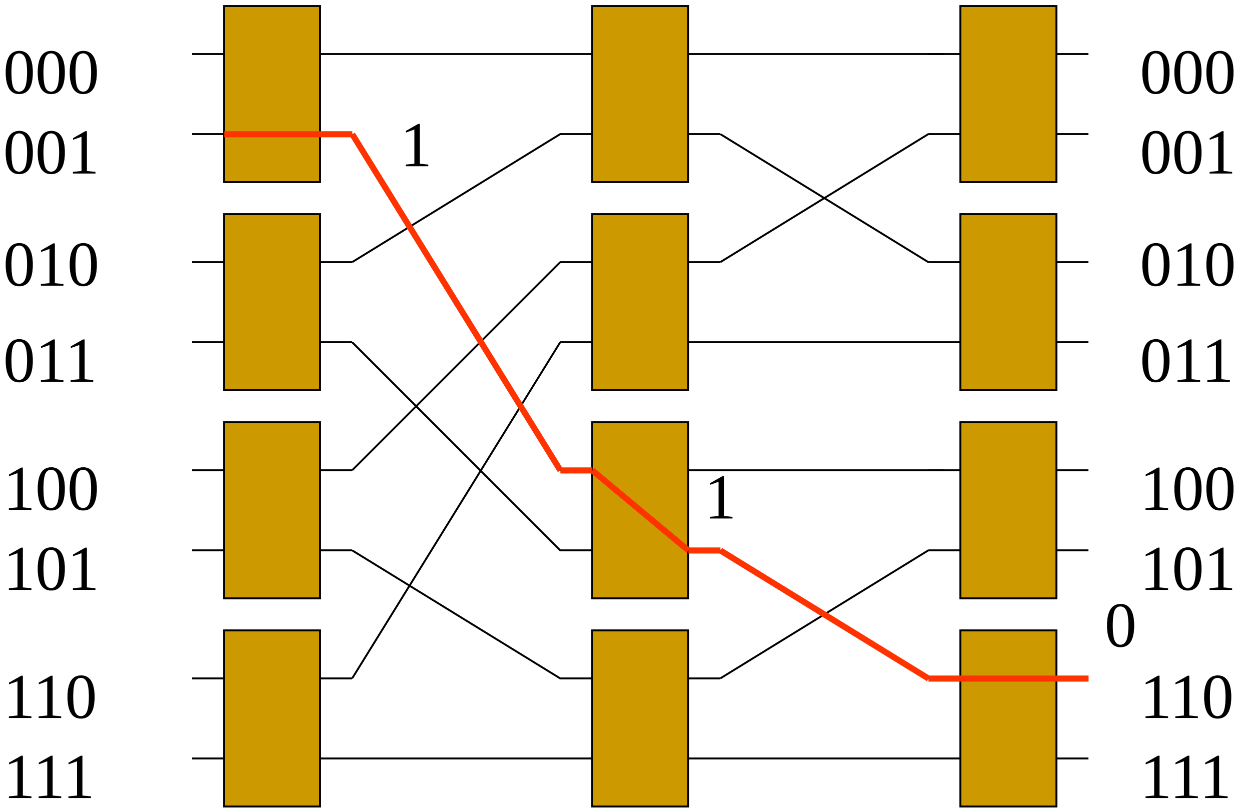
010

100

000
001
010
011
100
101
110
111

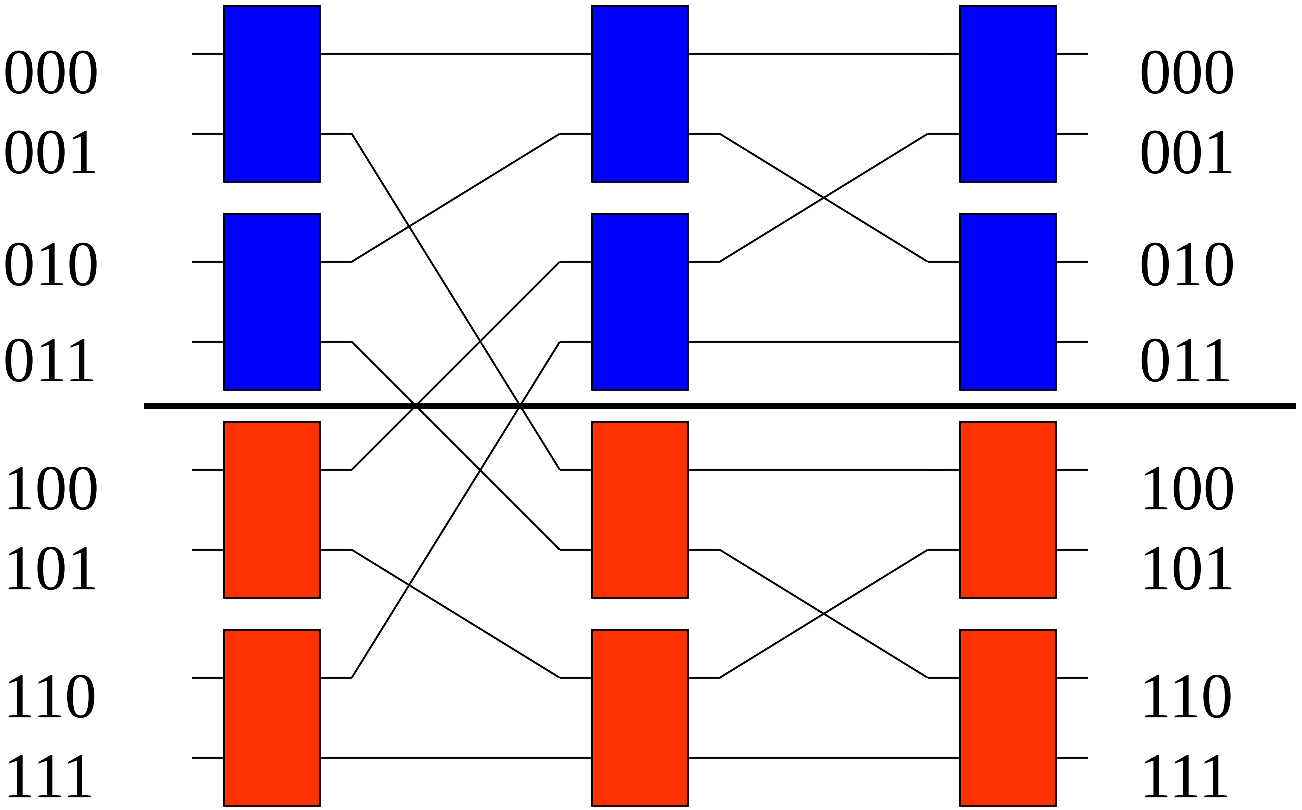
The area of shuffling is changed.

Destination Routing in Baseline network



Just like Omega network

Partitioning in Baseline



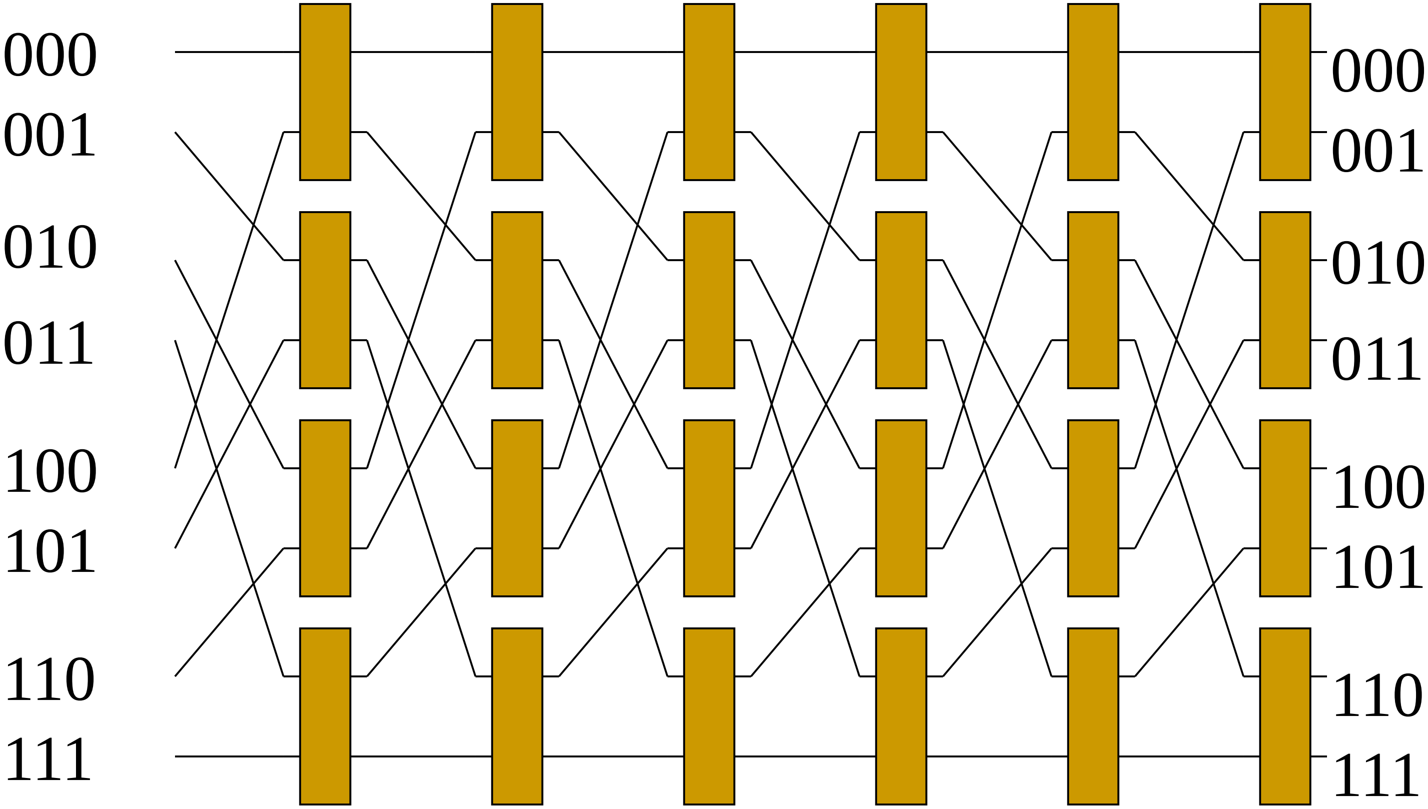
Baseline network

- Providing both benefits of Omega and Generalized Cube
  - Destination Routing
  - Partitioning
  - Expandability
- Used in NEC's Cenju

Quiz

- Assume that 2x2 crossbar Is used for a switching element of 32inputs Omega network. For making the calculation simple, only 1bit is used for each input. Calculate the number of cross-points used in the network, and compare with 32inputs crossbar switch.

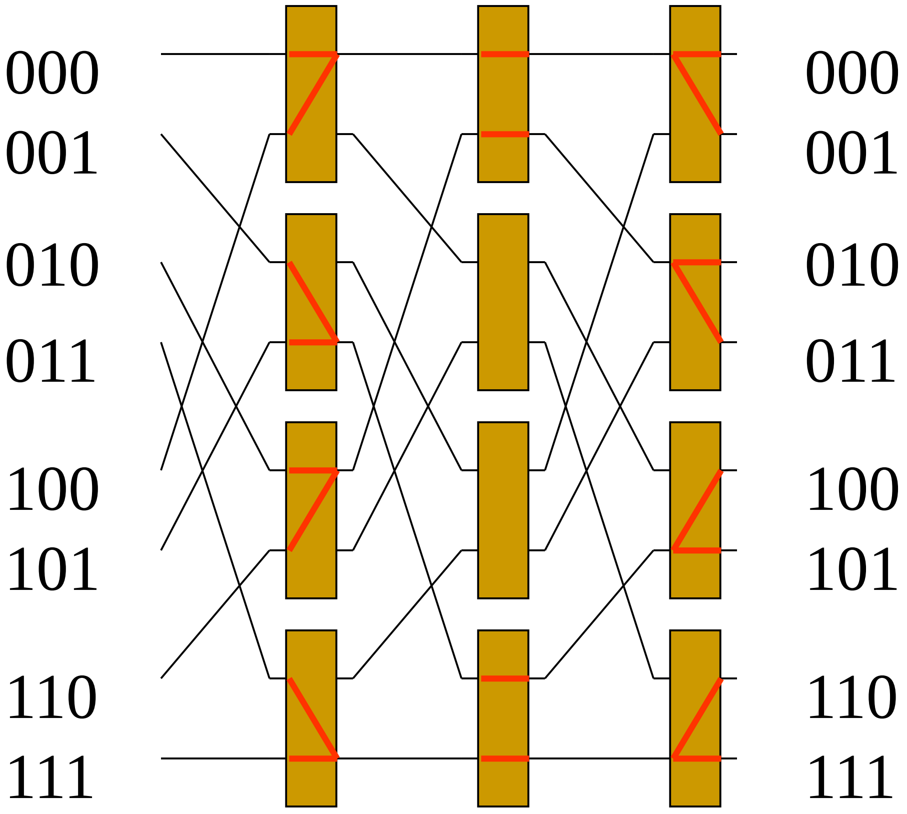
Π network



- Tandem connection of two Omega networks

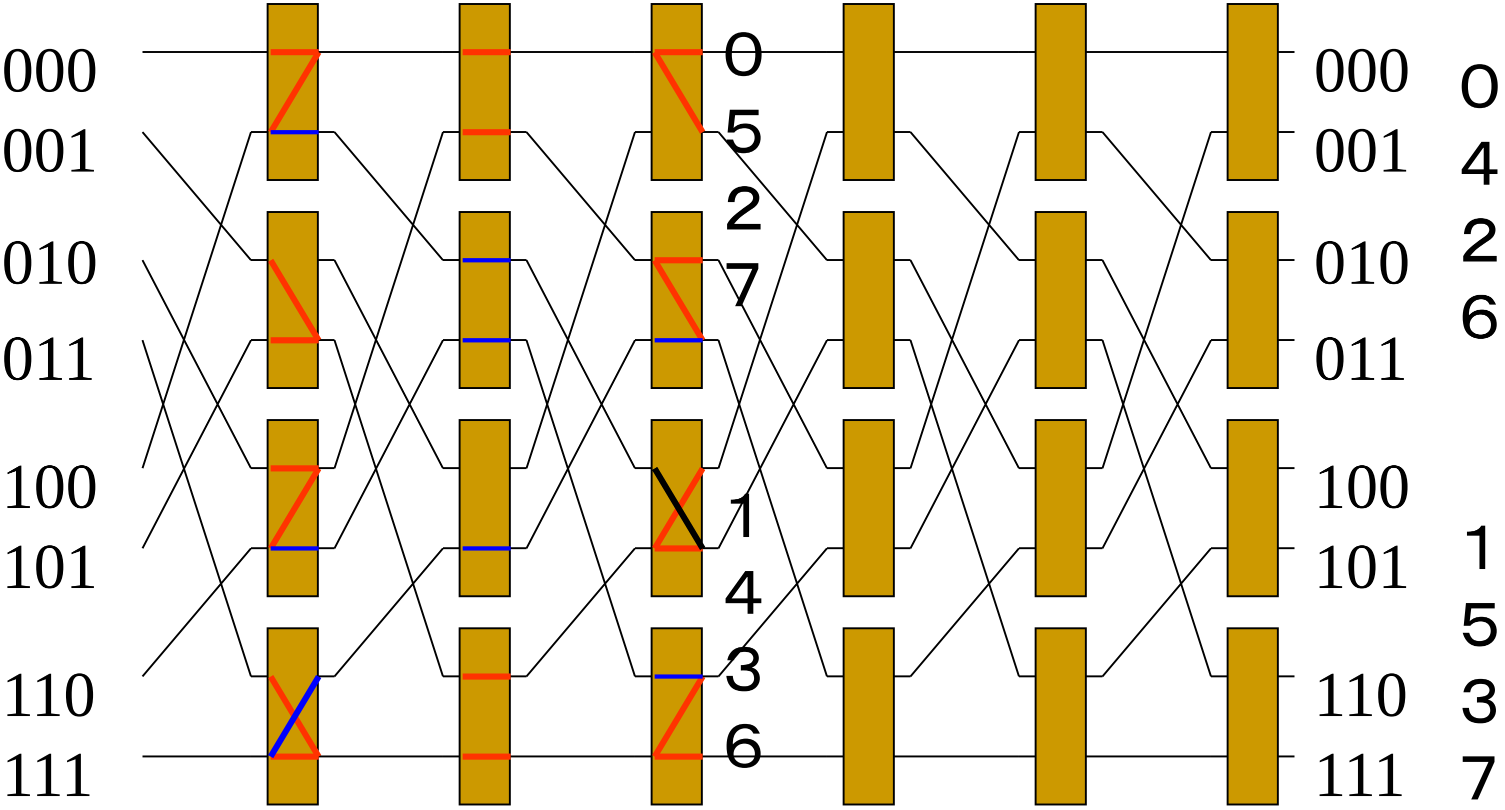
Bit reversal permutation (Used in FFT)

0	0
1	4
2	2
3	6
4	1
5	5
6	3
7	7



- Conflicts occur in Omega network.

Bit reversal permutation in Π network



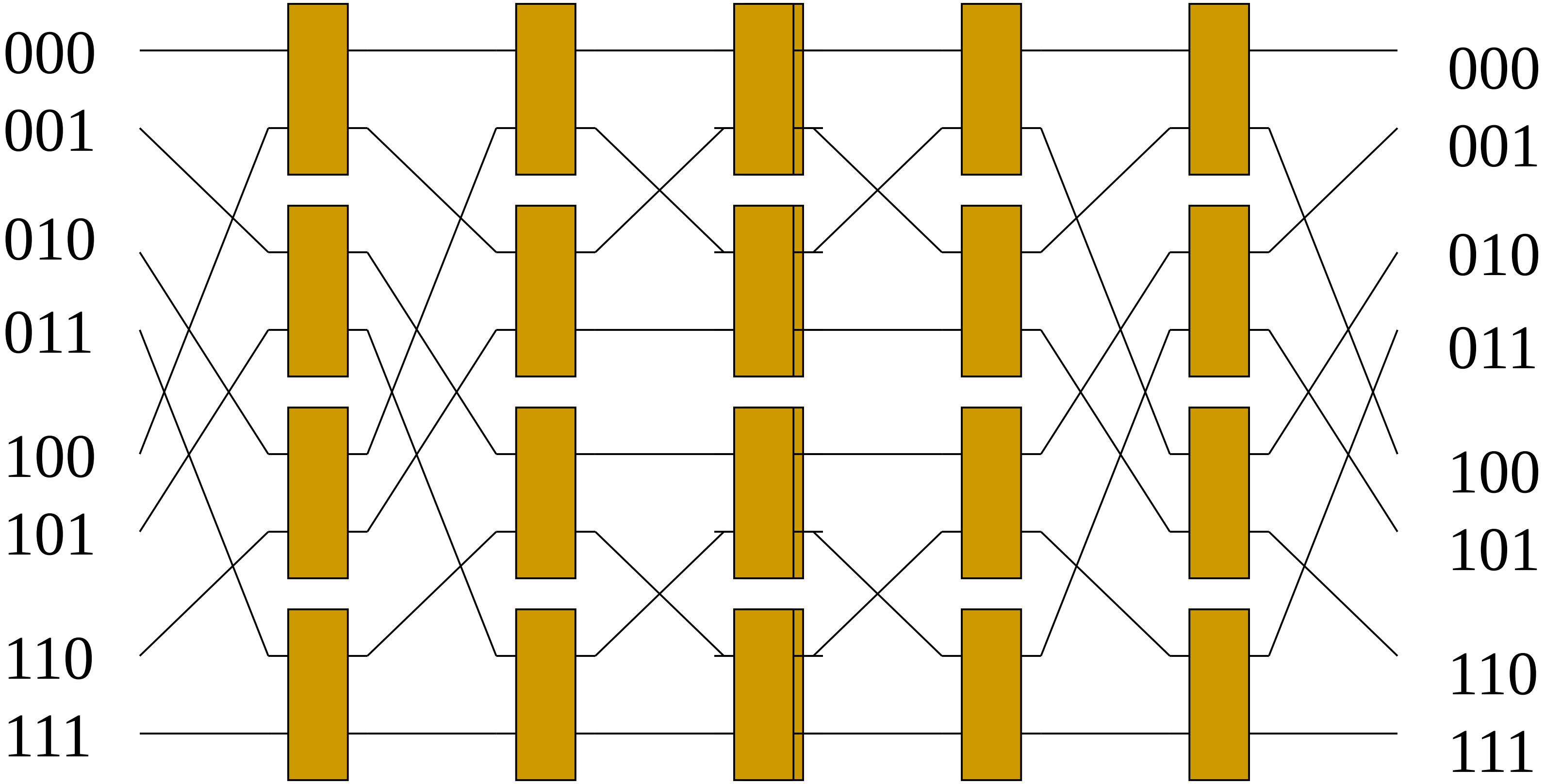
The first Omega : Upper input has priority.
The next Omega : Destination Routing

➡ Conflict free

Permutation capacity

- All possible permutation is conflict free ＝ Rearrangeable networks
- Three tandem connection of Omega network is rearrangeable.
- The tandem connection of Omega and Inverse Omega (Baseline and Inverse Baseline) is rearrangeable. Benes network

Benes Network



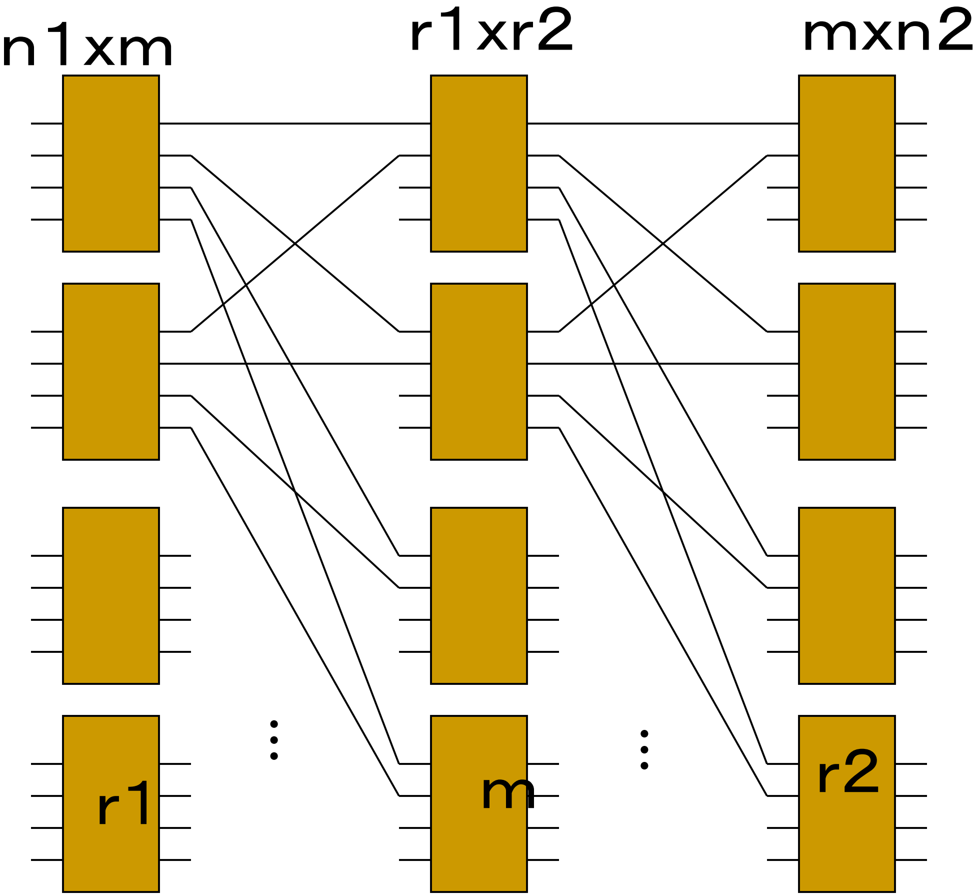
- Note that the center of stage is shared.
- The rearrangeable network with the smallest hardware requirement.

Non-blocking network

- Clos network
  - m＞n1＋n2－1: Non-blocking
  - m＞＝n2: Rearrangeable
  - Else: Blocking

Clos network

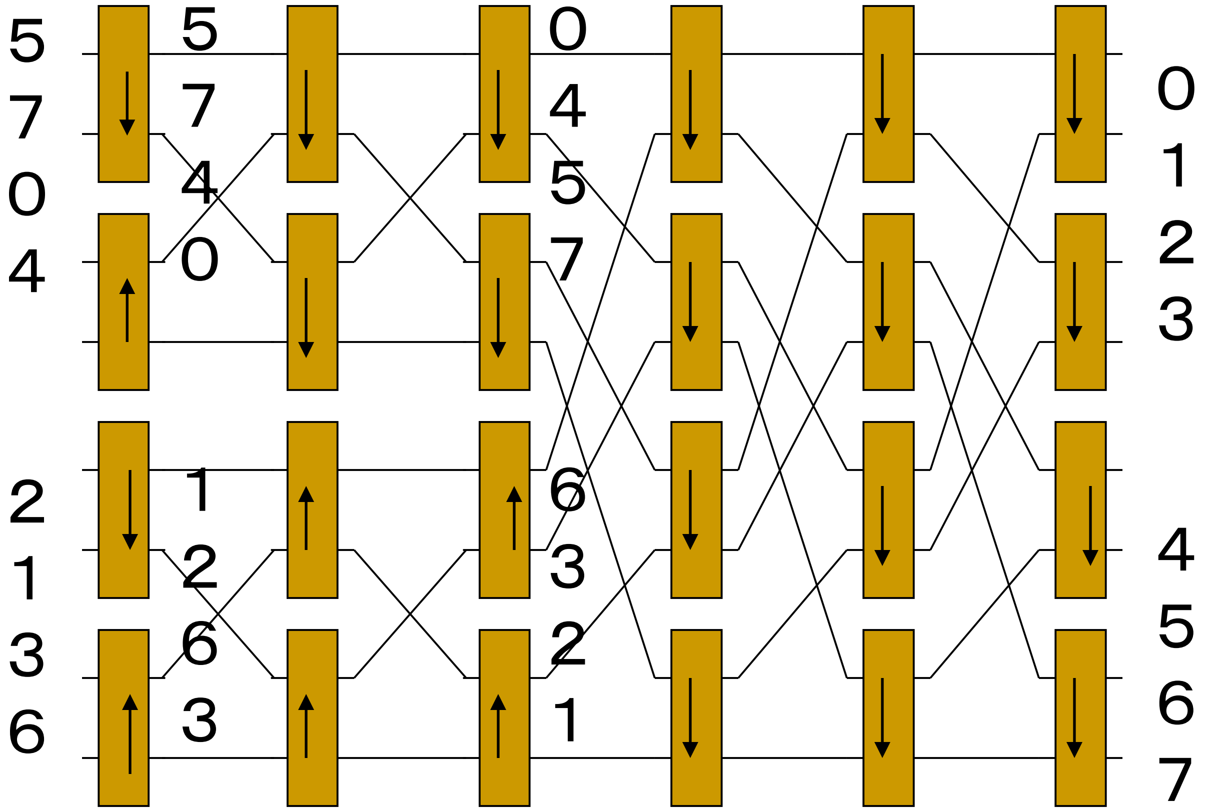
3-stage



m=n1+n2-1 : Non-blocking
m=n2 : Rearrangeable
m<n2 : Blocking

The number of intermediate stage dominates the permutation capability.

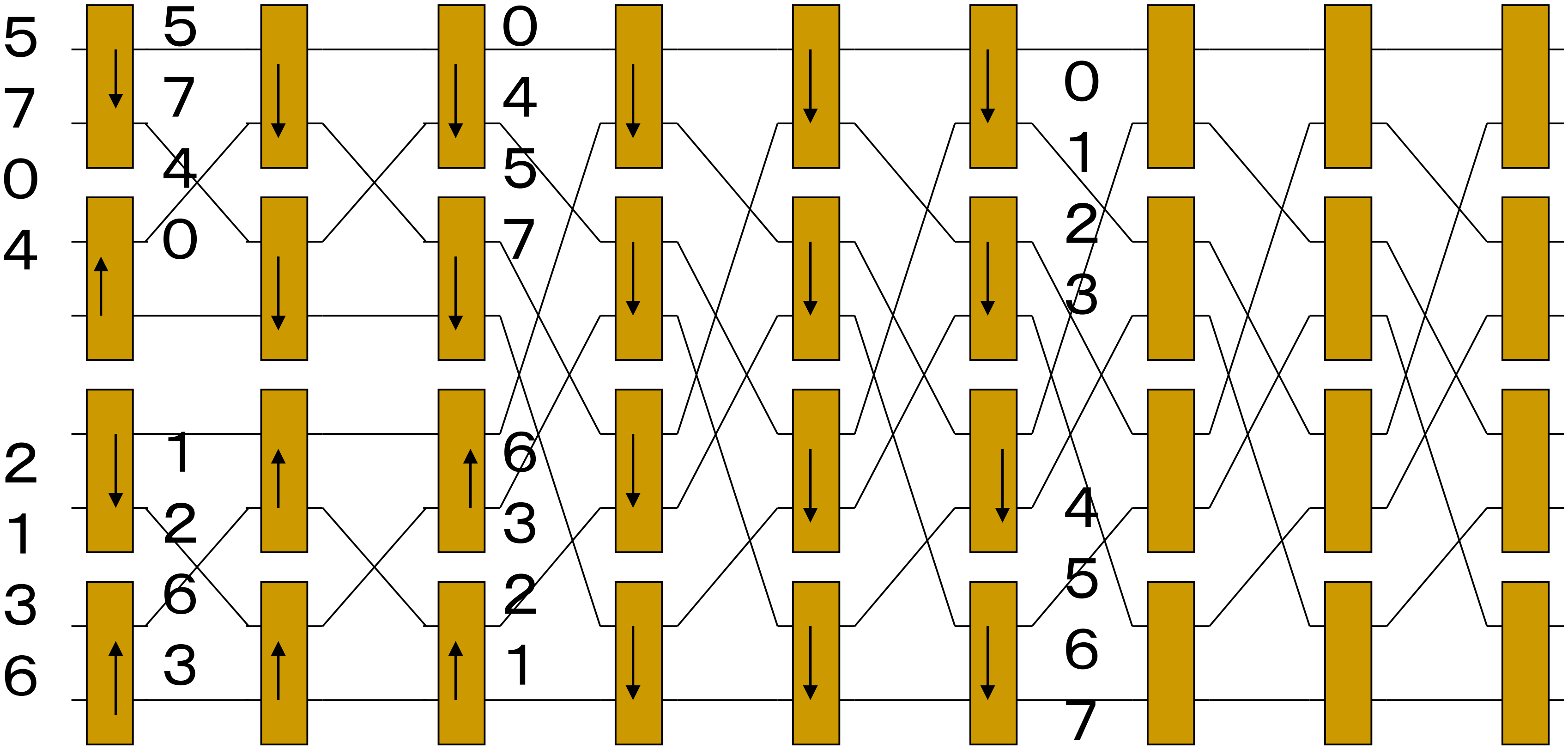
Batcher network



Bitonic sorting network

Batcher-Banyan

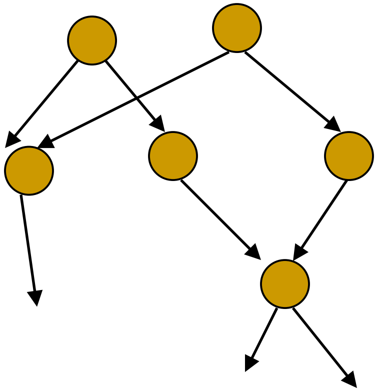
Omega
Baseline



Sorted input is conflict free in the banyan network

Banyan networks

- Only a path is provided between source and destination.
- The number of intermediate stages is flexible.
- Approach from graph theory
- SW-Banyan, CC-Banyan, Barrel Shifter



Irregular structure is allowed.

Batcher-banyan

- If there are multiple packets to the same destination, the conflict free condition is broken

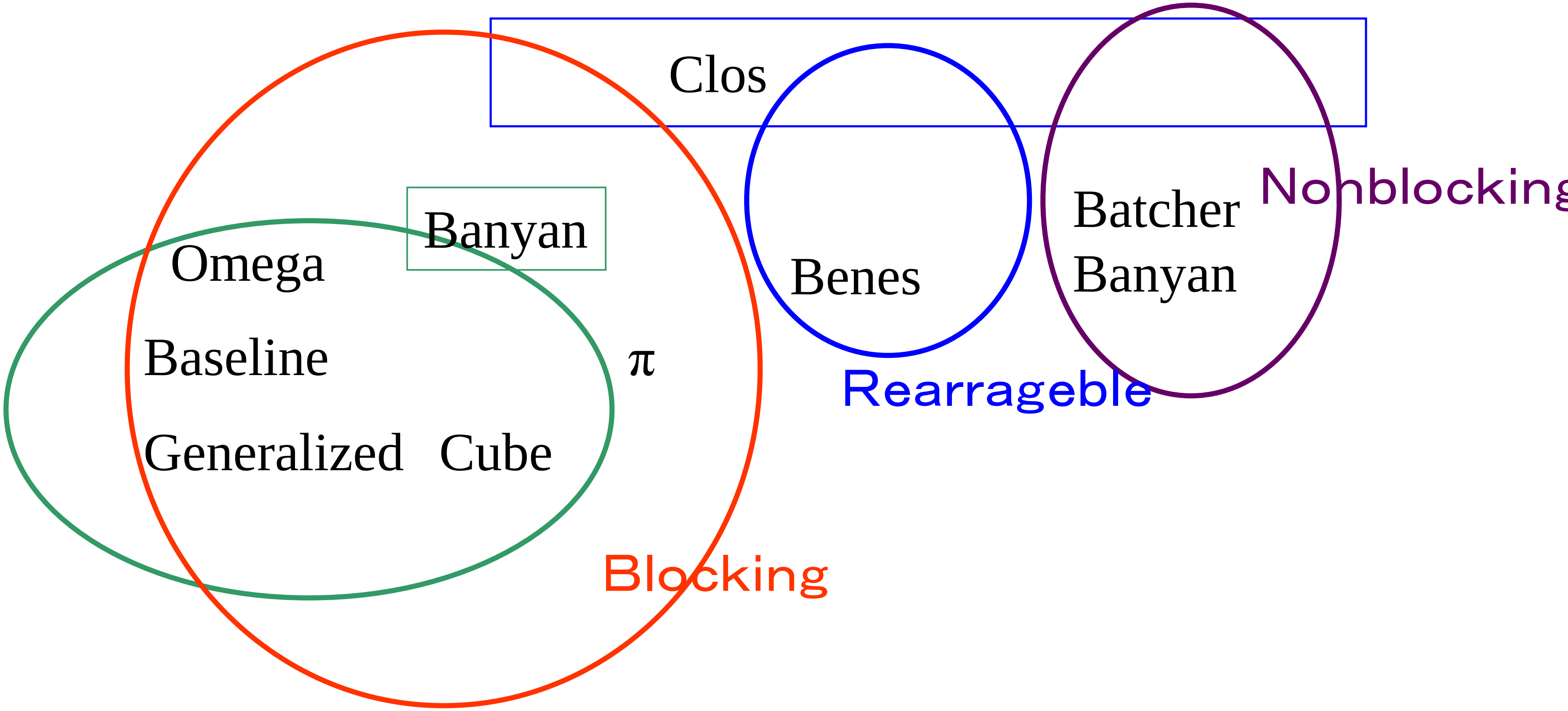
  → The other packets may conflict.

  - The extension of banyan network is required.
- The number of stages is large.

  → Large pass through time

  - The structure of sorting network is simple.

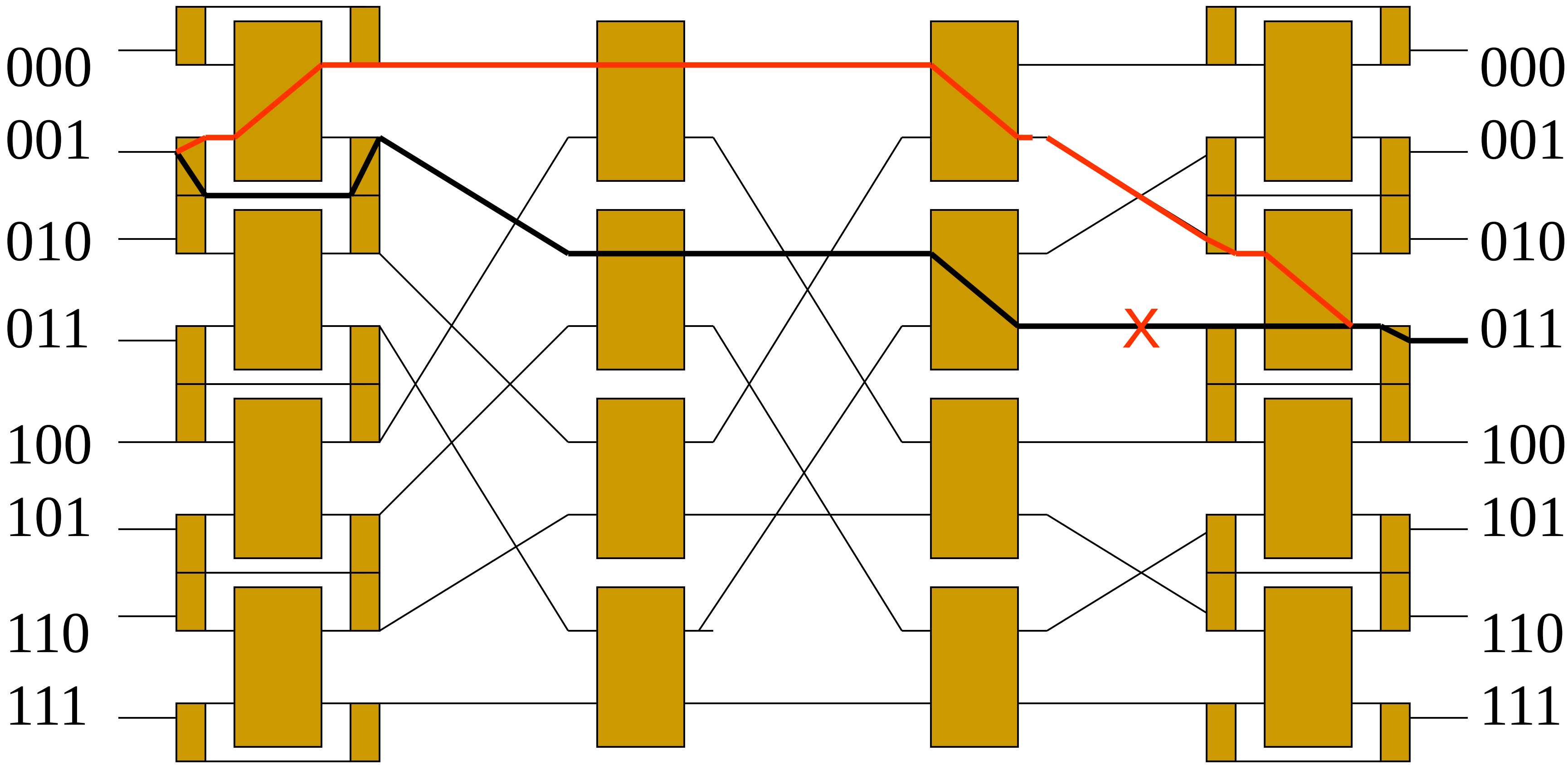
Classification of MINs



Fault tolerant MINs

- Multiple paths
- Redundant structure is required.
- On-the-fly fault recovery is difficult.
- Improving chip yield.

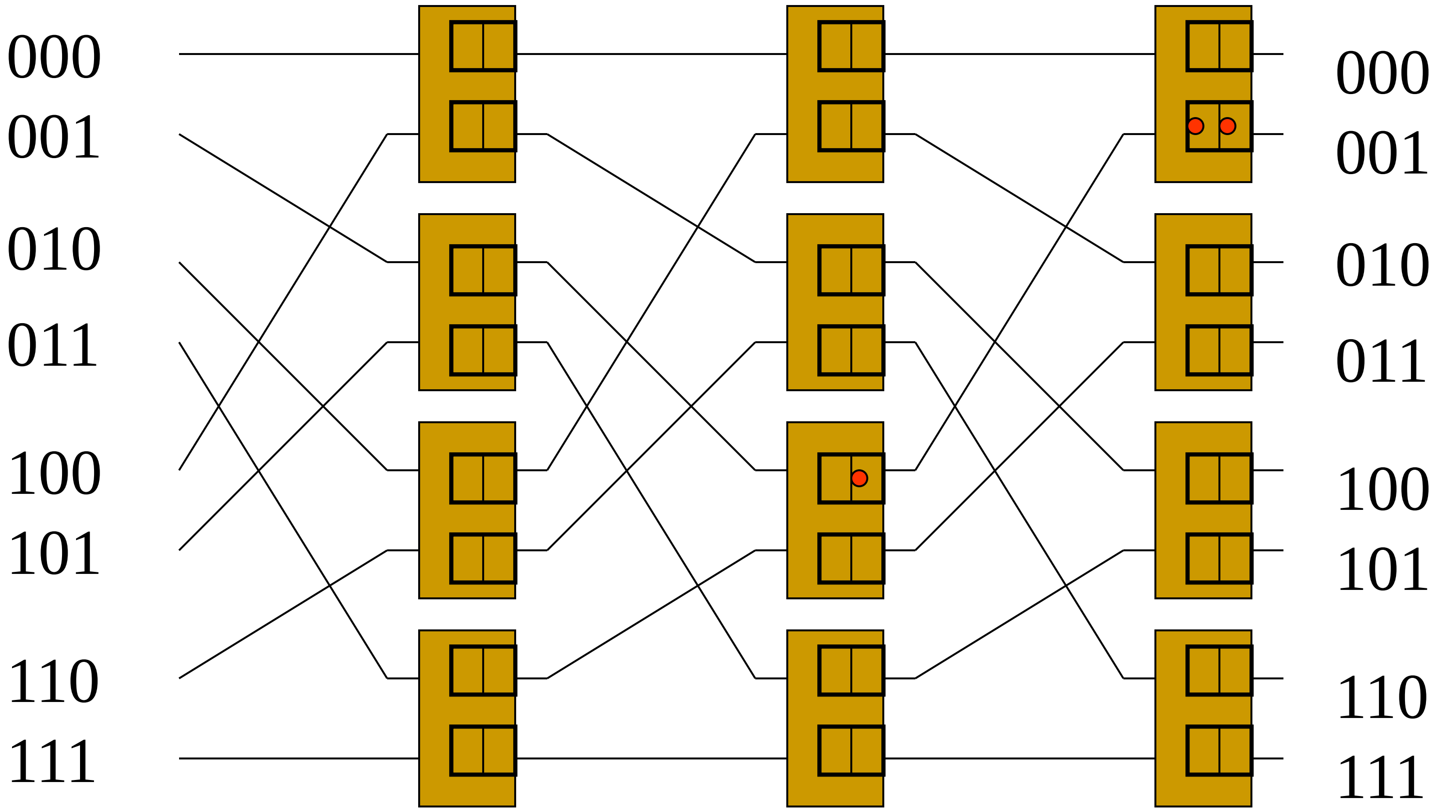
Extra Stage Cube (ESC)



- An extra stage＋Bypass mechanism

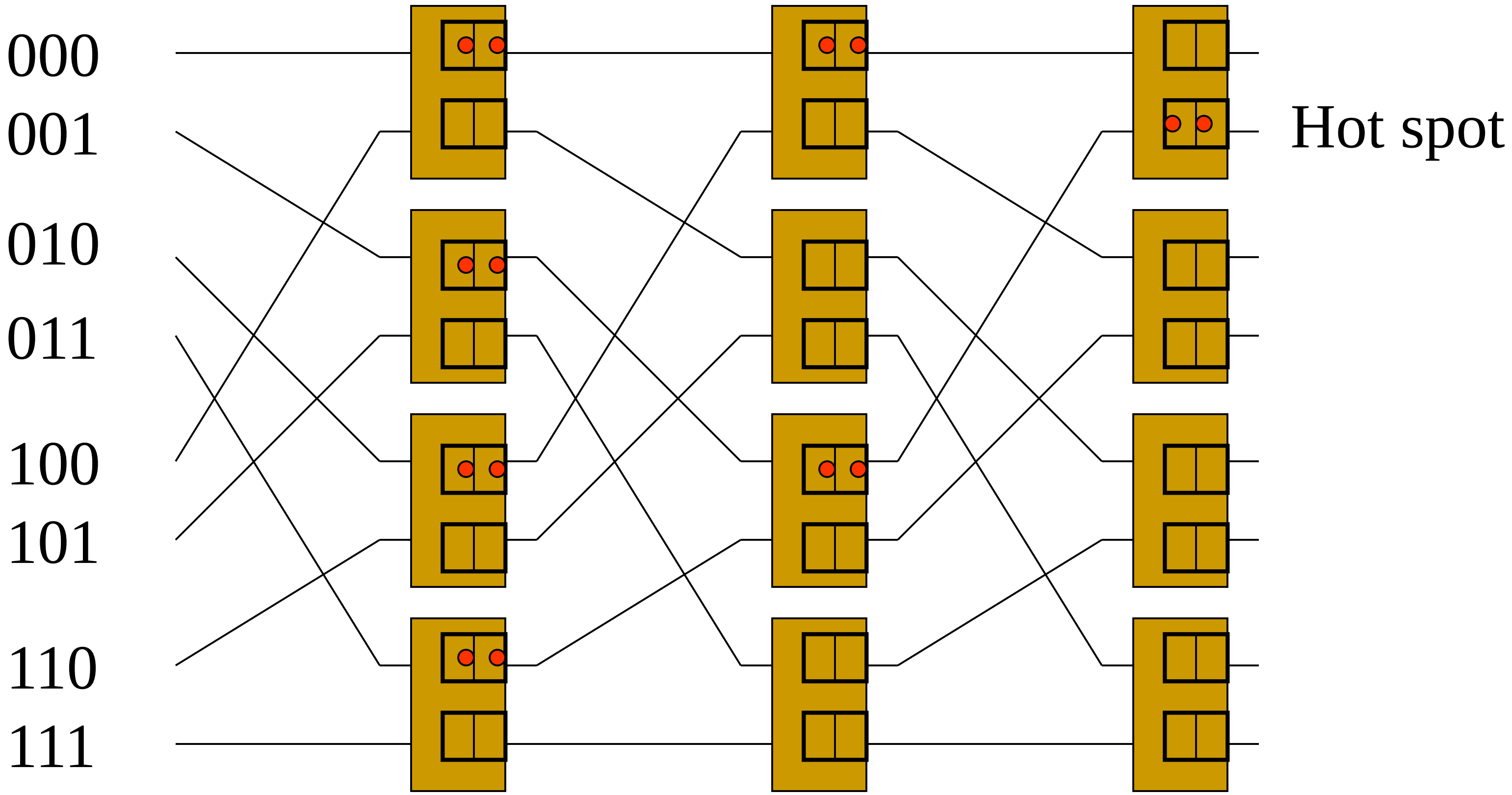
If there is a fault on stages or links, another path is used.

The buffer in switching element



- Conflicting packets are stored into buffers.

Hot spot contention



- Buffer is saturated in the figure of tree (Tree　Saturation)

Relaxing the hot spot contention

- Wormhole routing with Virtual channels → Direct network
- Message Combining
  - Multiple packets are combining to a packet inside a switching element (IBM RP3)
  - Implementation is difficult (Implemented in SNAIL)

Other issues in MINs

- MIN with cache control mechanism
  - Directory on MIN
  - Cache Controller on MIN
- MINs with U-turn path → Fat tree

Glossary 1

- Rearrange-able: スケジュールすることにより、出力が重ならなければ内部で衝突しないようにできる構成
- Perfect shuffle:シャッフルは、トランプの札を切る時に使う単語だが、ここでは、配線のつなぎ方の方式のひとつ。Inverse shuffleは逆シャッフルと呼ばれ、逆接続方式。
- Destination routing : 目的地のラベルだけで経路を決める方法
- Permutation:並び替え、順列のことだが、ここでは目的地ラベルが重ならない経路を無衝突で生成することができる能力のこと
- Partitioning:ネットワークを分離して独立に使える能力のこと
- Fault tolerance:耐故障性。一部が故障しても全体がダウンしないような性質、Fault tolerant MINは複数経路を持たせたMIN
- Expandability:拡張性、小さなものからサイズを大きくしていくことのできる性質
- Hot spot contention:　局所的に交信が集中して、これが全体に波及すること。
- Tree saturation:　Hot spot contentionによりネットワークが木の形で飽和していく現象。特にMINで起きる。Message Combiningは、メッセージをくっつけてまとめることによりこれを防止する方法の一つ

Summary

- Recently, practical new topologies are not proposed.
- A lot of “made-in-Japan” networks
- Asymmetric indirect networks will be widely used.

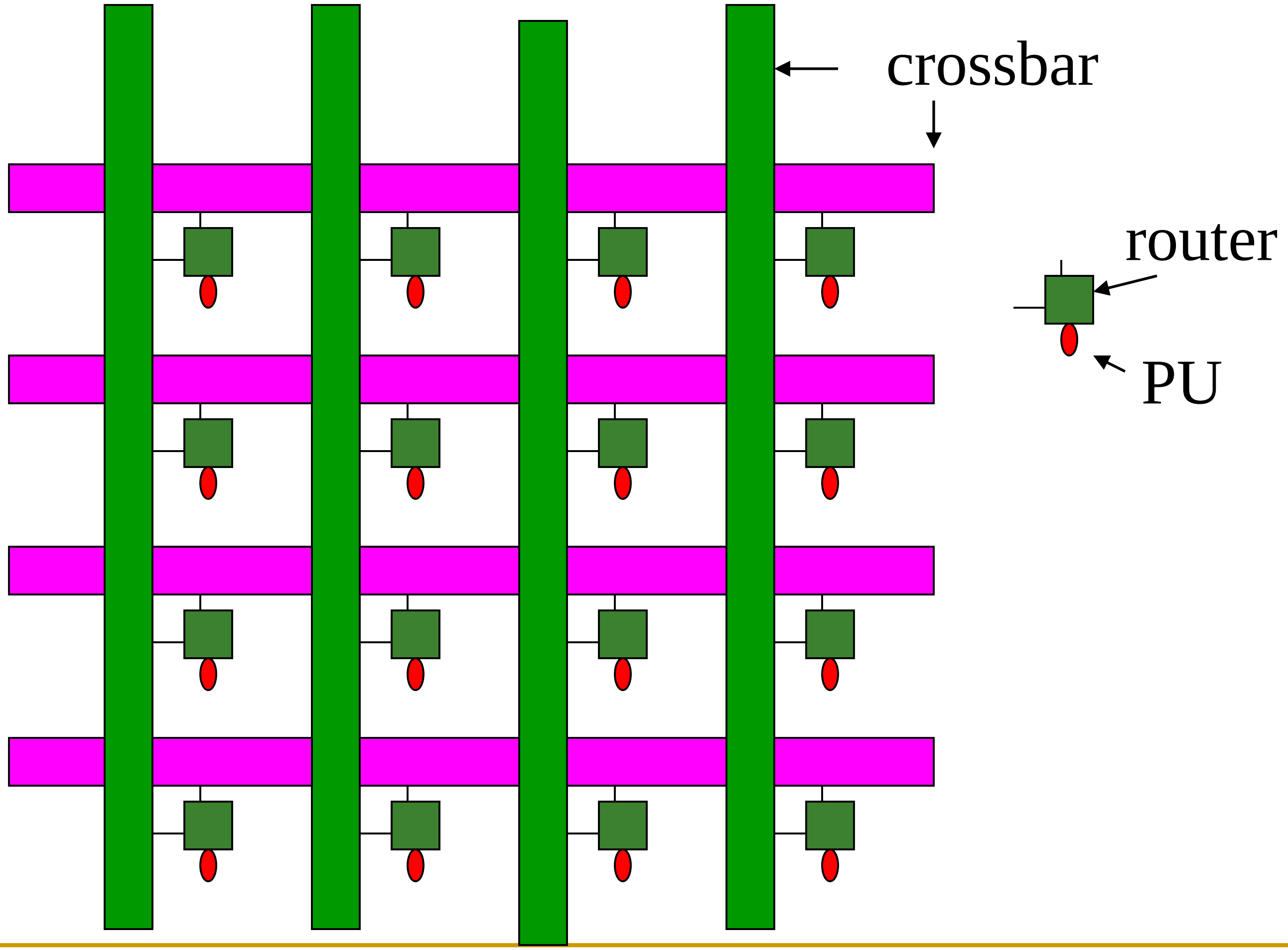
Centralized interconnection networks

- Symmetric:
  - MIN (Multistage Interconnection Networks)
  - Each node is connected with equal latency and bandwidth
- Asymmetric:
  - Fat-tree, base-m n-cube, etc.
  - Locality of communication can be used.

Asymmetric indirect networks

- Intermediate position between direct and indirect networks
- High communication capability considering cost
  - base-m　n-cube（Hyper crossbar）
    - SR2000、CP-PACS
  - Fat　Tree
    - CM-5, Some WS　Clusters
  - Hyper-cross
    - ADENART

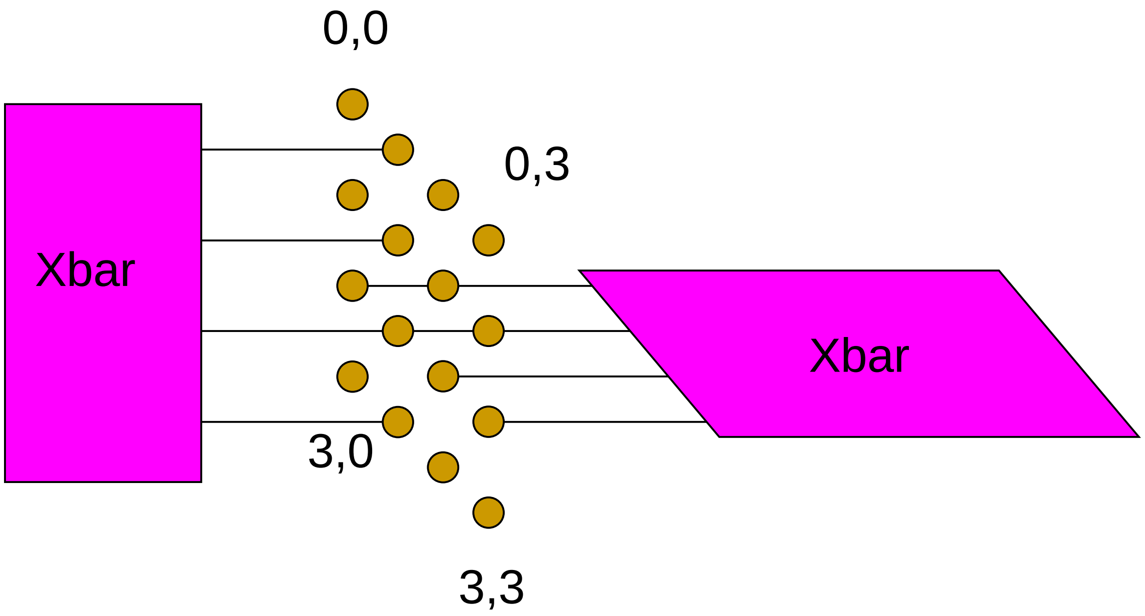
base-m n-cube (Hyper crossbar)



Used in Toshiba's Prodigy and Hitachi's SR8000

HyperCross

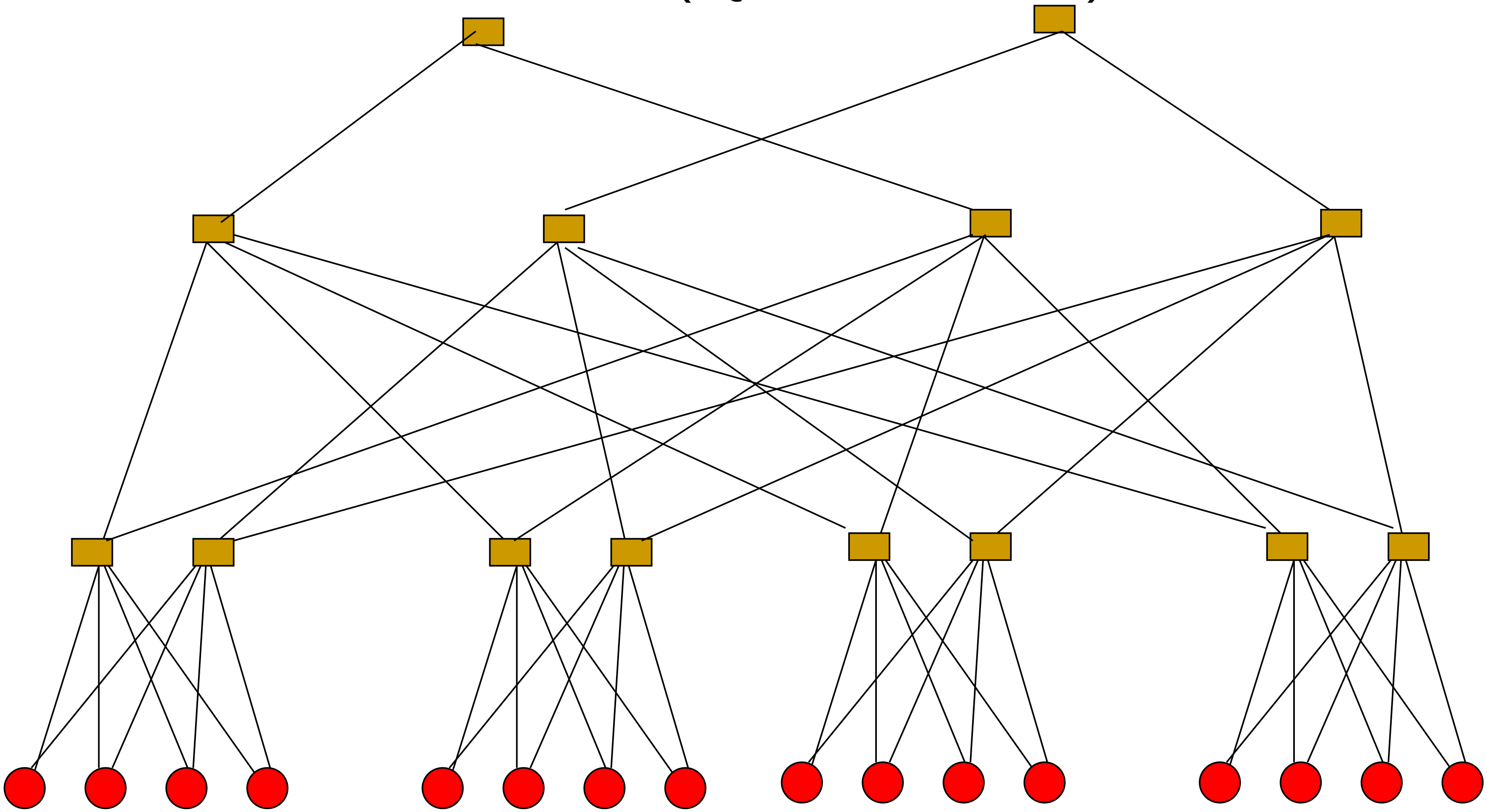
(pi,pj)→ (pj,*),(*,pi)



Used in ADENART by Matsushita

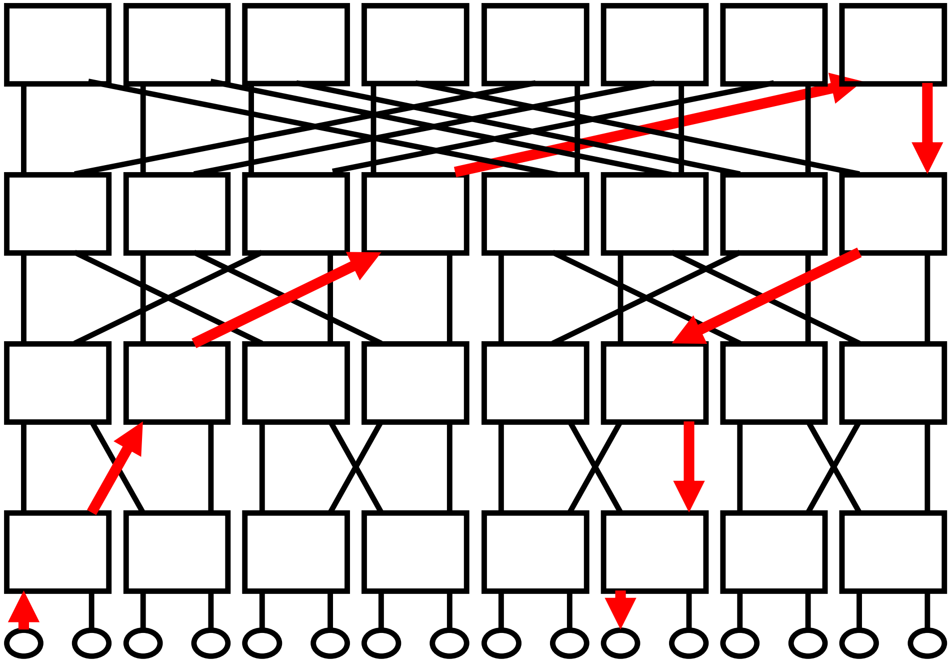
Fat Tree

Used in CM-5 and
PC Clusters(QsNet, Autonet)

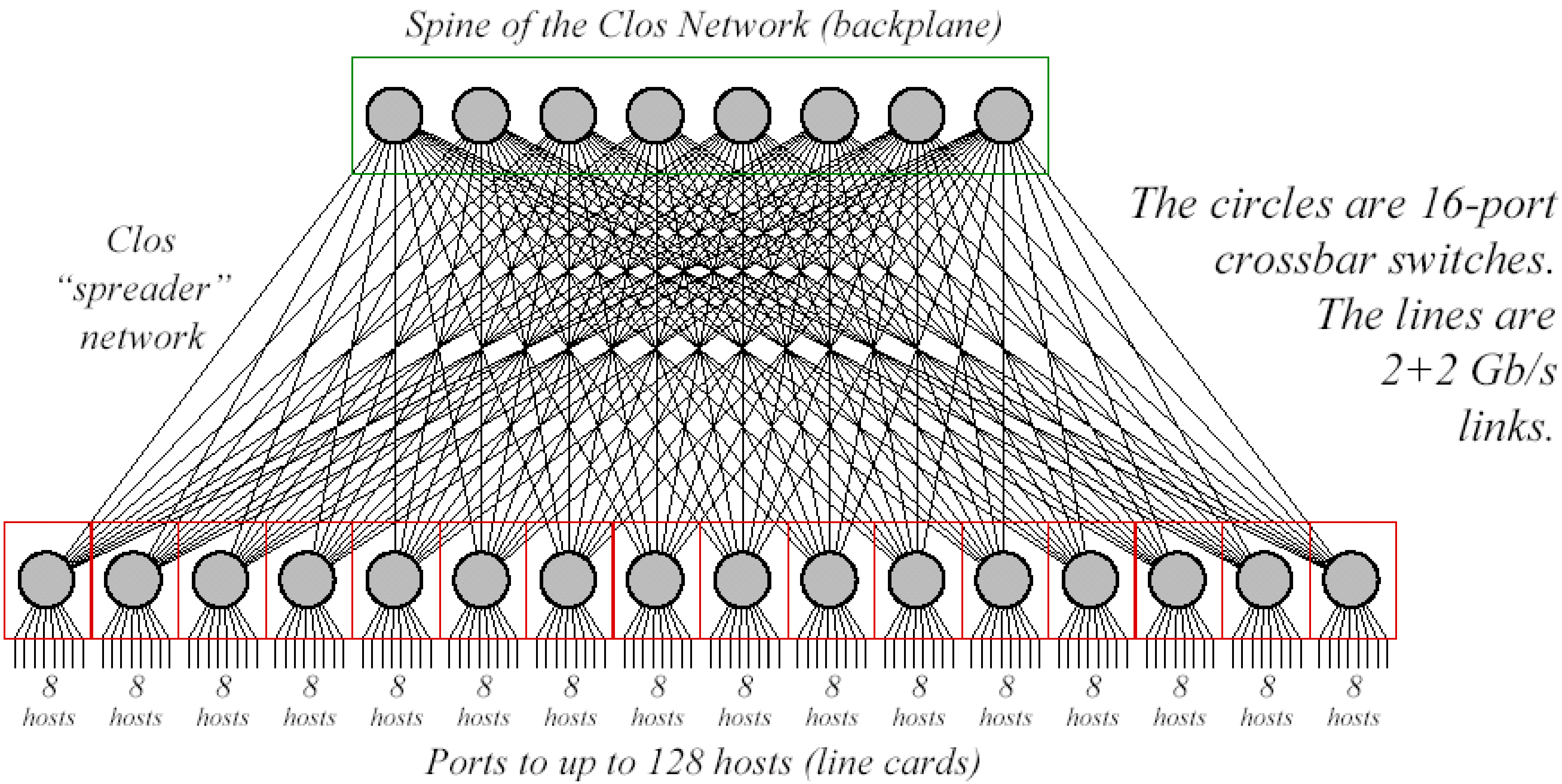


Myrinet-Clos is actually a type of Fat-tree

Return paths with MIN = Fat tree

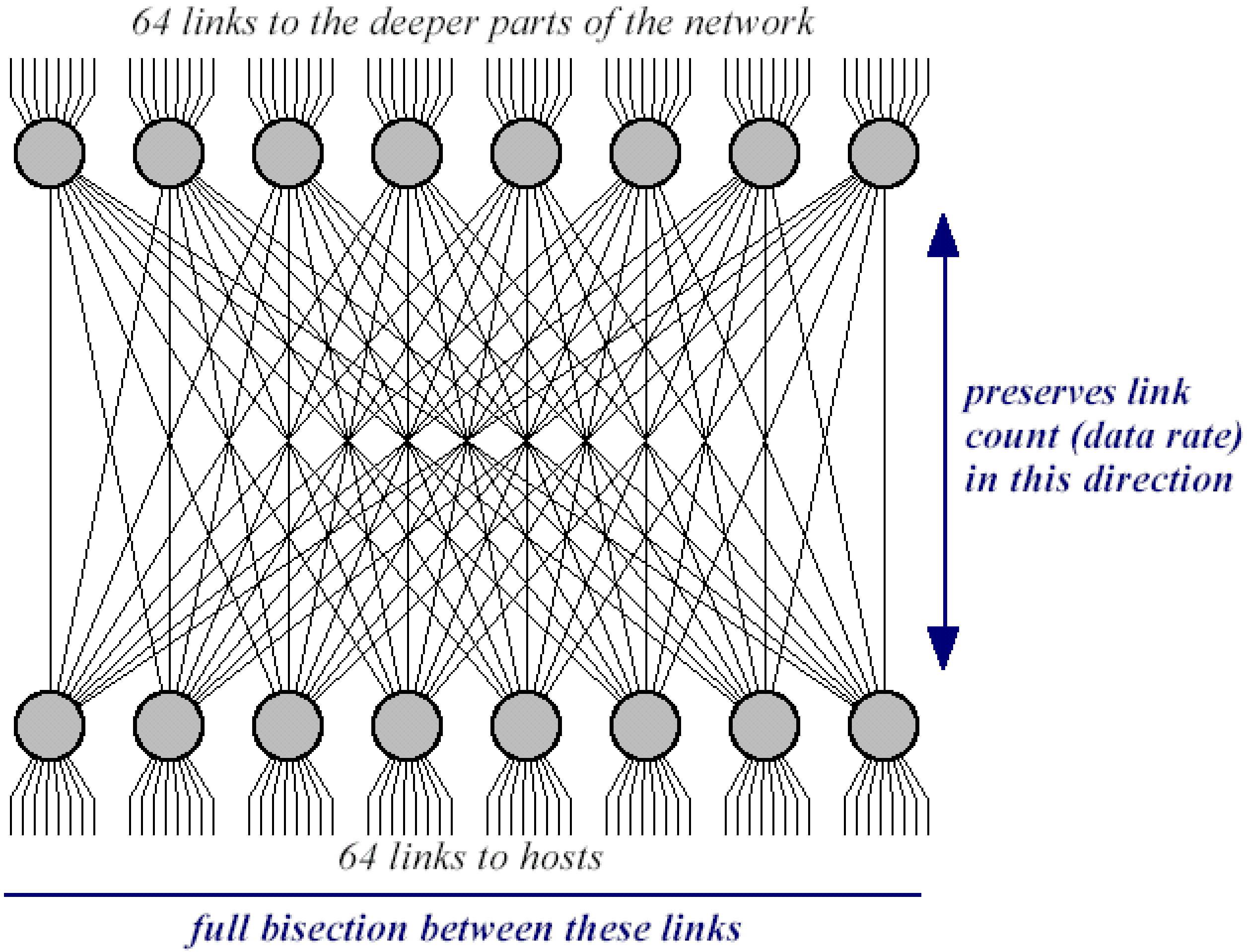


Myrinet-Clos（1/2)


Spine of the Clos Network (backplane)
Clos "spreader" network
The circles are 16-port crossbar switches.
The lines are 2+2 Gb/s links.
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
8 hosts
Ports to up to 128 hosts (line cards)

- 128nodes(Clos128)

Clos64+64



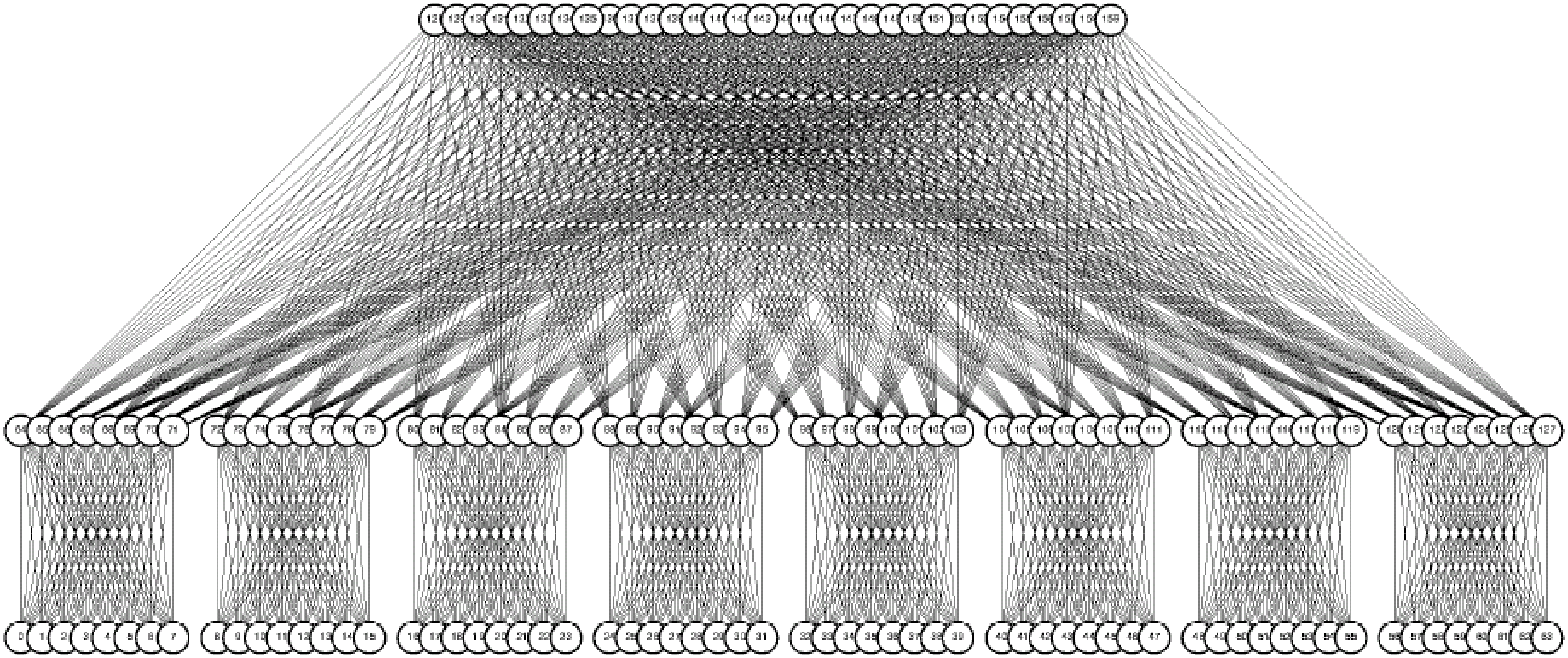
64 links to the deeper parts of the network

preserves link count (data rate) in this direction

64 links to hosts

full bisection between these links

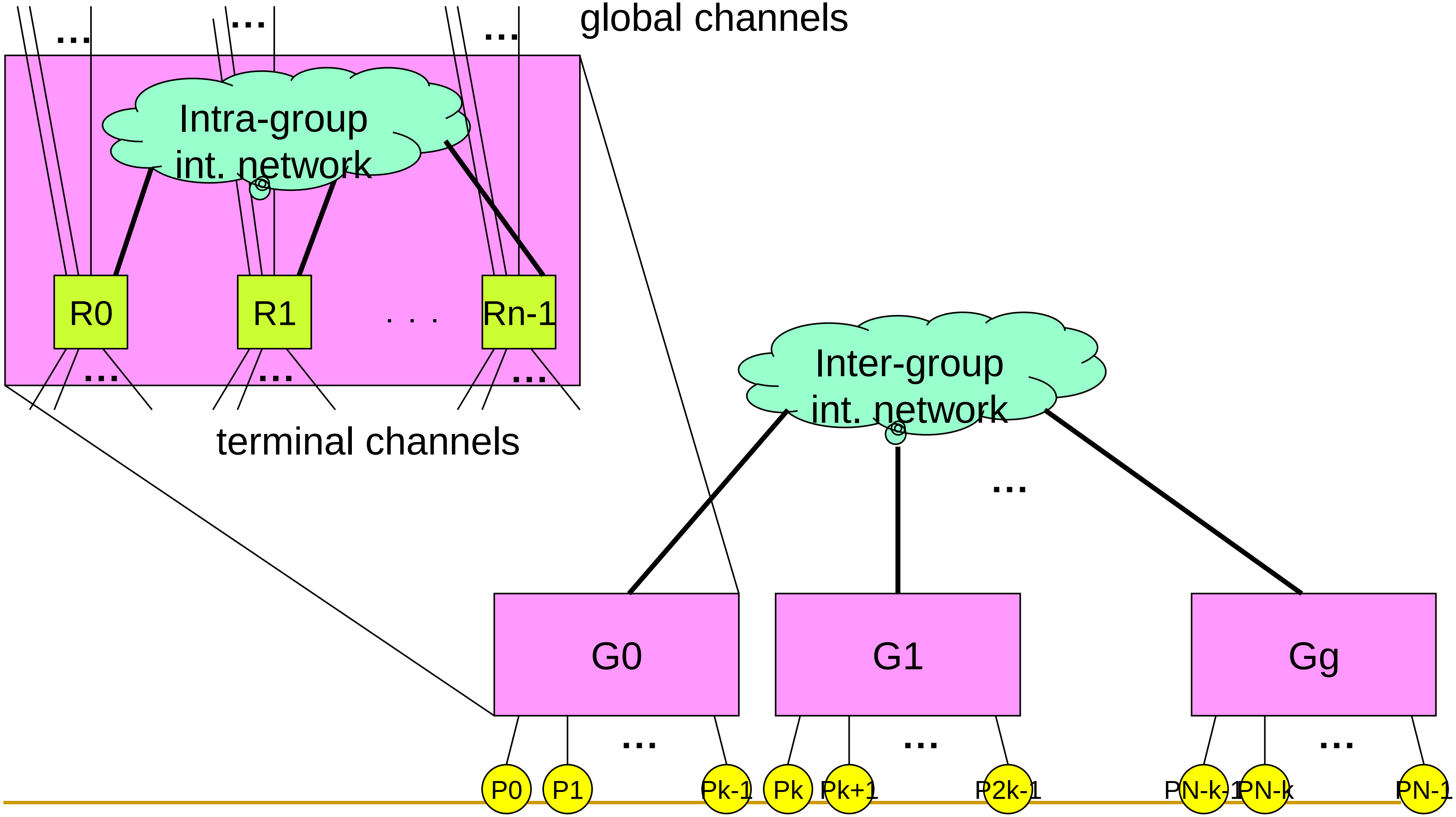
Myrinet-Clos(2/2)

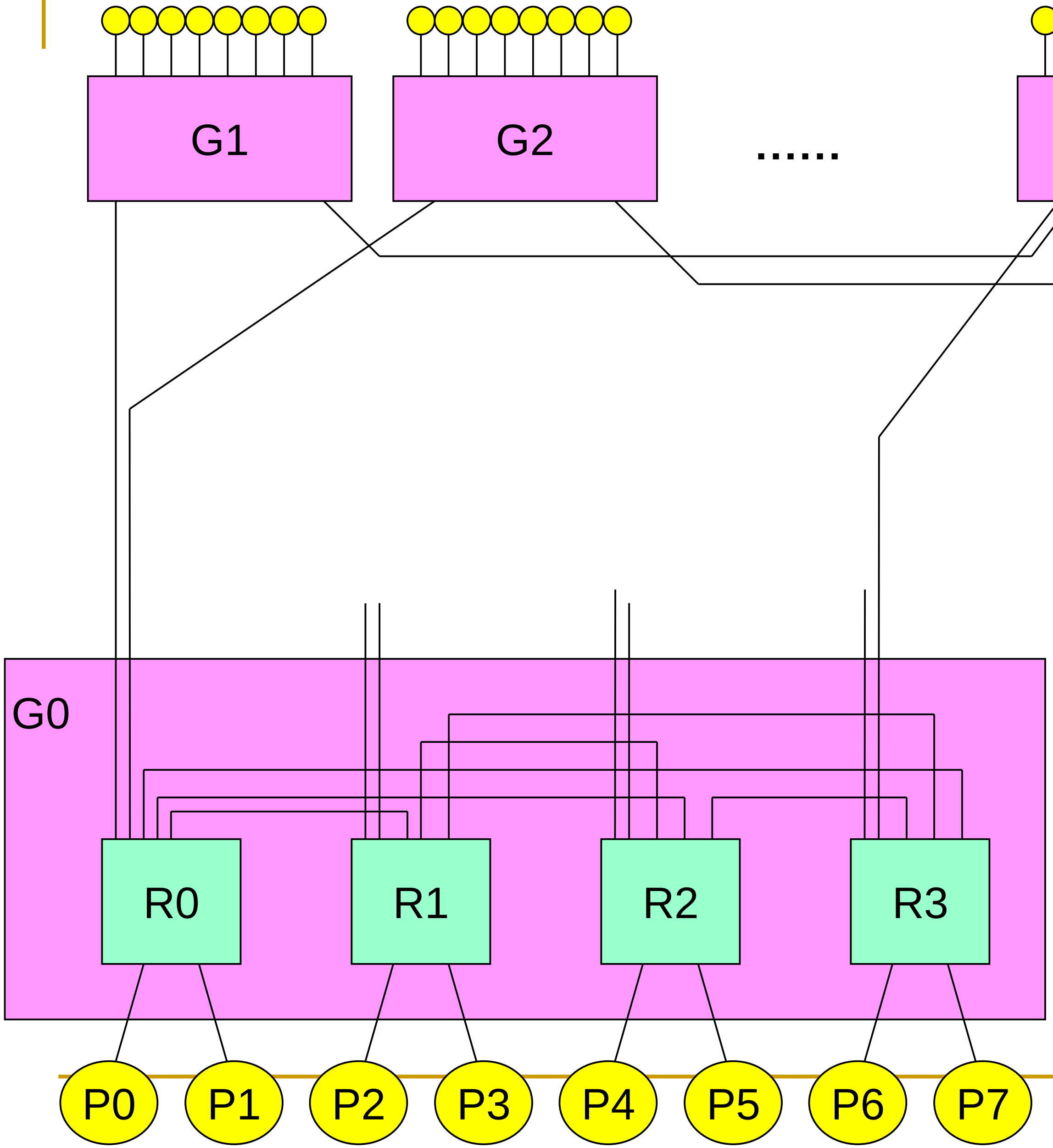


The 512 hosts connect to 8 ports on each of these 64 "leaf" switches

- 512nodes

Dragonfly


global channels
…
…
…
Intra-group
int. network
R0
R1
. . .
Rn-1
…
…
…
terminal channels
Inter-group
int. network
…
G0
G1
Gg
…
…
…
P0
P1
Pk-1
Pk
Pk+1
P2k-1
PN-k-1
PN-k
PN-1

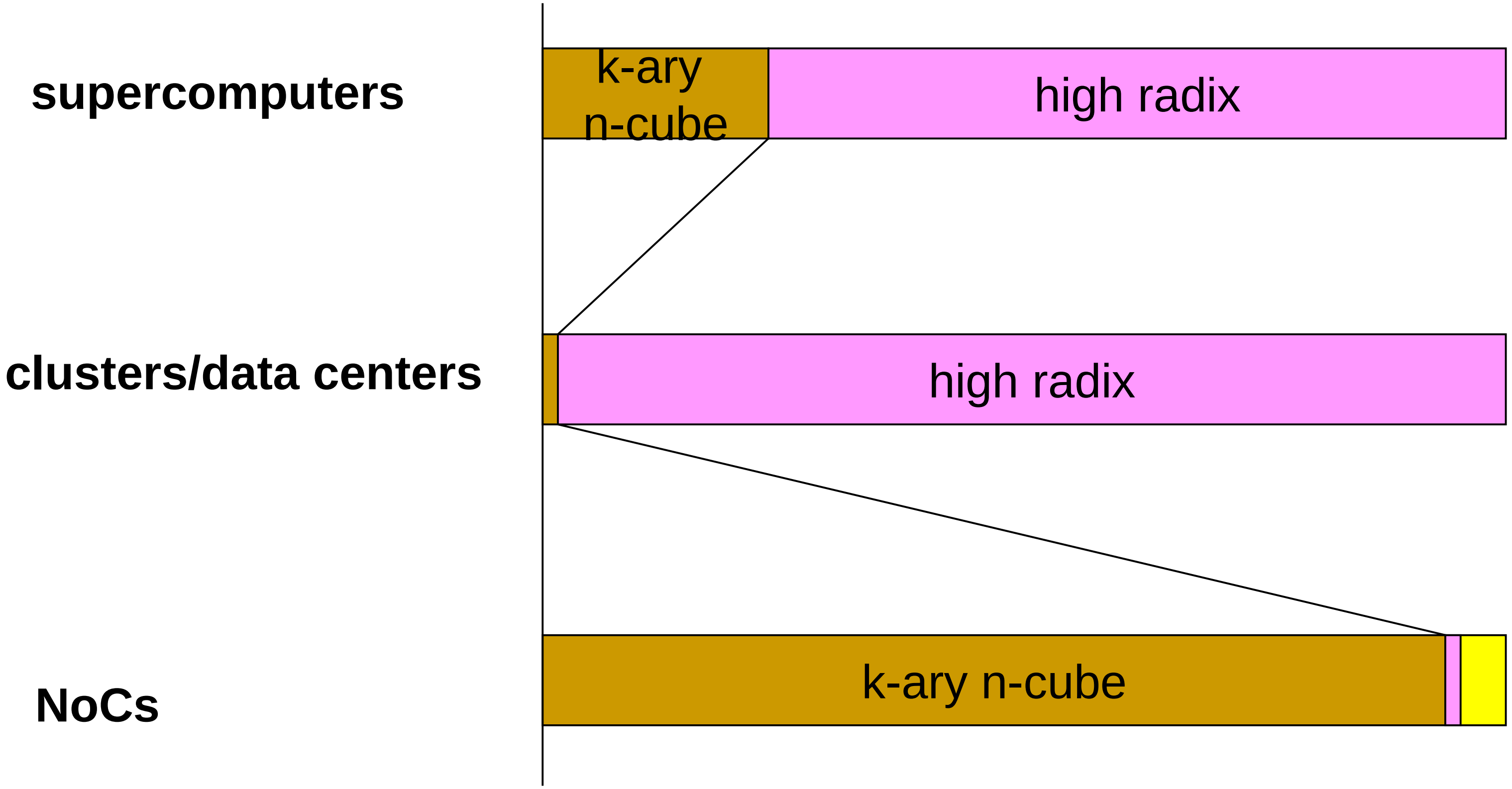


An example of Dragonfly
(N=72)

The interconnection of
this part can be Flatten Butterfly

k-ary n-cube vs. high radix

Sorry. This figure is not based on accurate data.



Now, they dominate the world of Interconnection Networks

Exercise

- Every path between source and destination is determined with the destination routing in Omega network. Prove (or explain) the above theory in Omega network with 8-input/output.