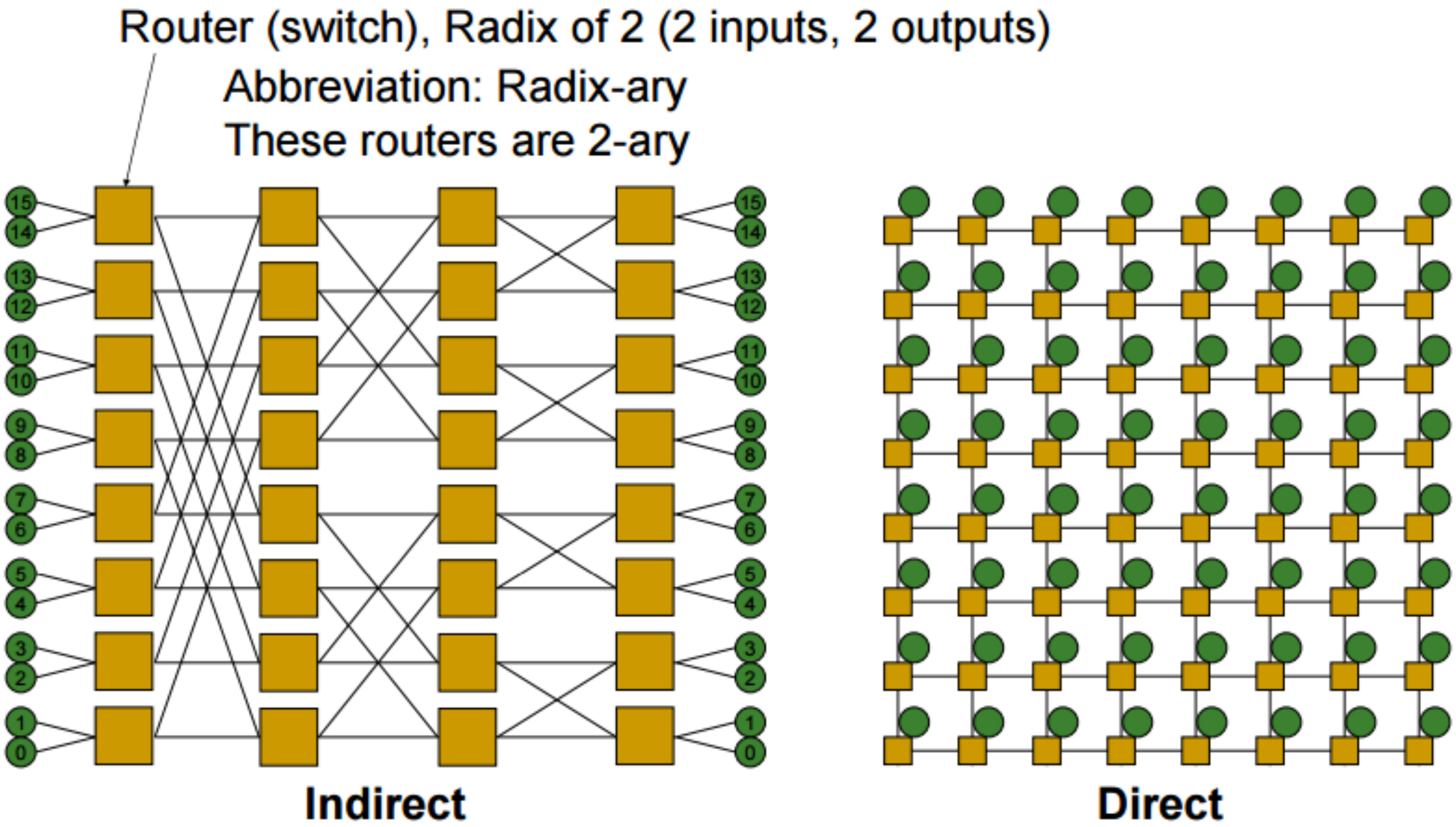
University of BRISTOL

Lecture 5

Interconnection networks - Non-blocking networks

Direct & Indirect Networks

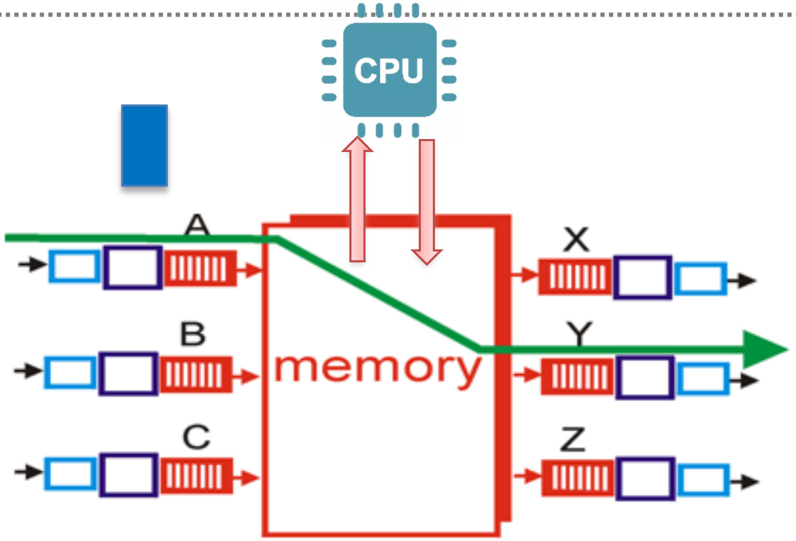
- Direct: Every switch also network end point
  - Ex: mesh, torus, and hypercubes
- Indirect: Not all switches are end points
  - Ex: Butterfly, Fat Tree, Clos



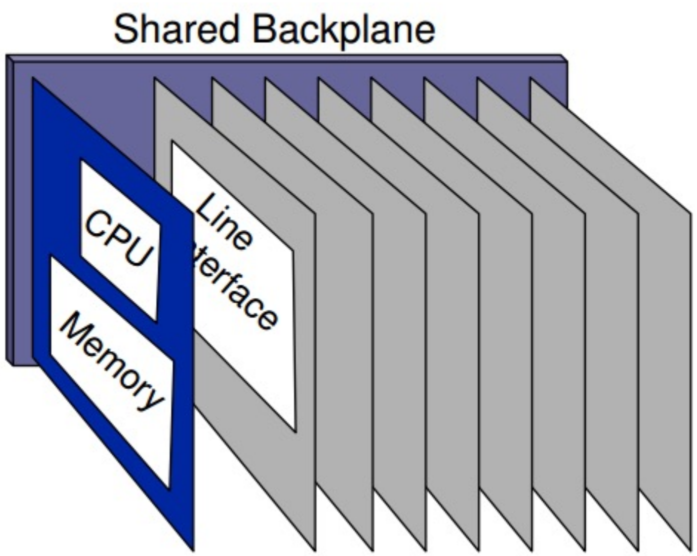
Direct Networks vs. Indirect Networks
To a fully connected network

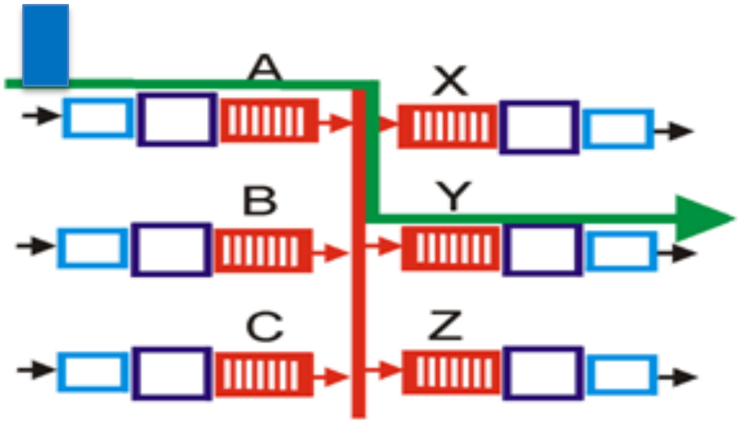
- Direct networks
  - All network nodes have processor or memory attached (direct connection between processors)
  - Difficult to scale up with limited bisection bandwidth
  - Impractical ports/switch
- Indirect networks
  - Intermediate routing-only nodes
  - Few pins per switch node
  - Better network throughput
  - Decoupling computing and switching

Switching via memory or Switching via a bus



Switching via memory Typically < 0.5Gbps



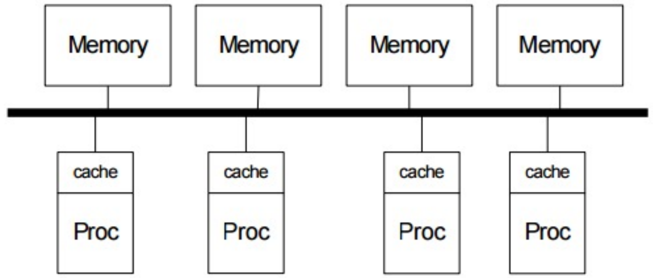


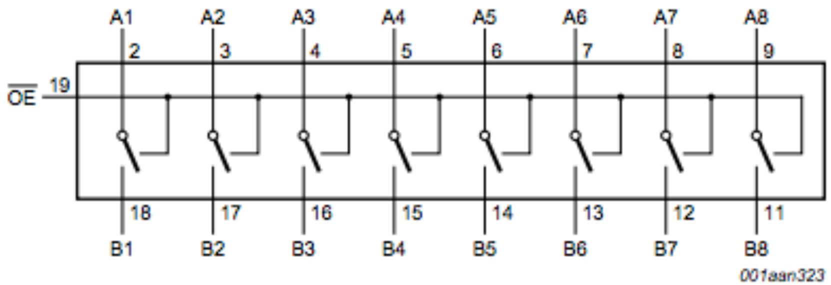
Typically < 5Gbps

Switching via a bus

Bus

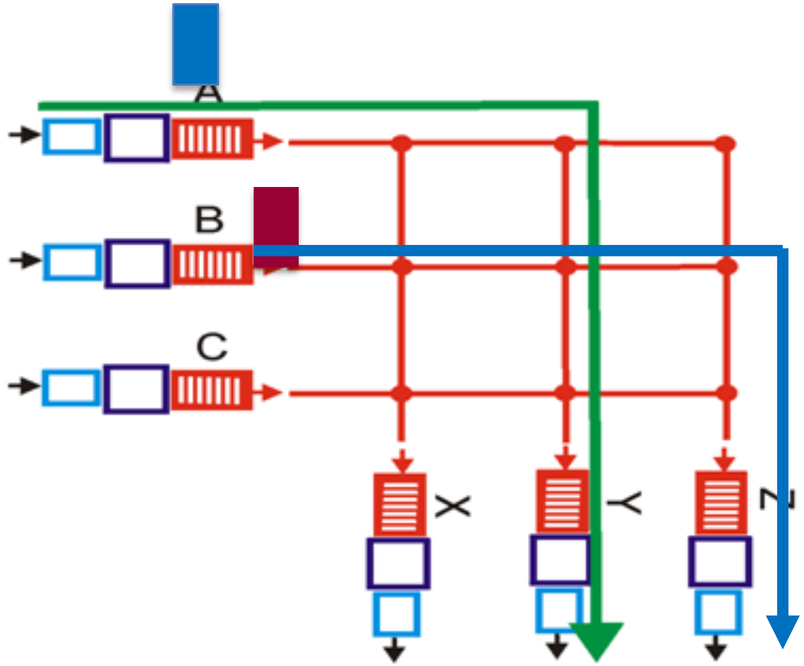
- Pros:
  - Simple
  - Cost effective for a small number of nodes
  - Easy to implement coherence (snooping)
- Cons:
  - Not scalable to large number of nodes (limited bandwidth, electrical loading)
  - High contention Memory



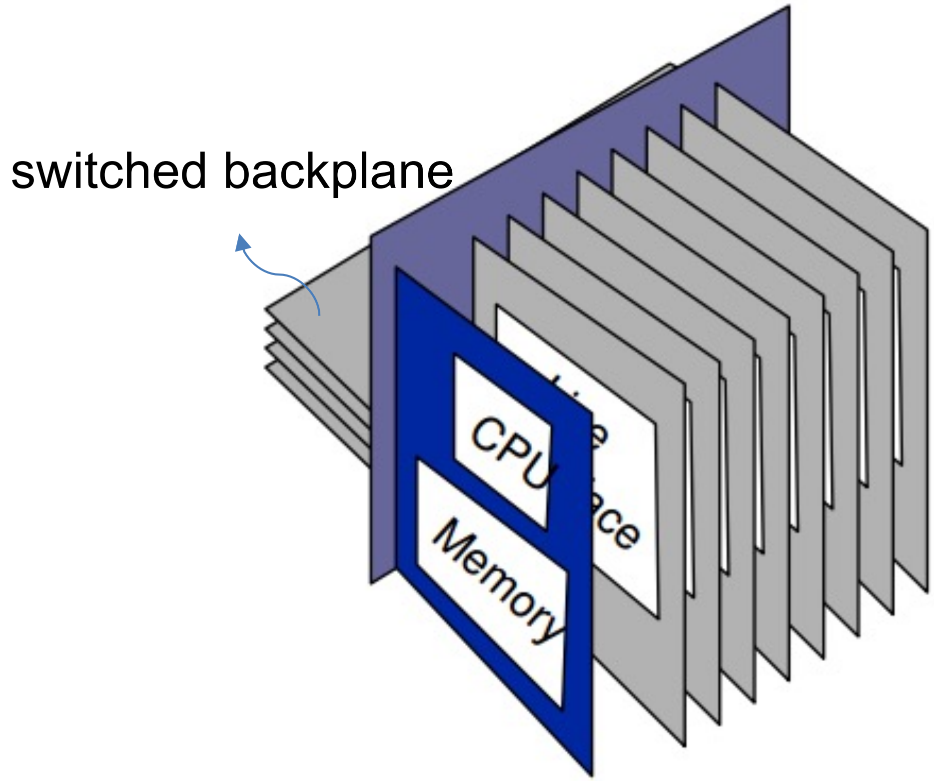


8-bit Bus Switch

Three types of switching fabrics

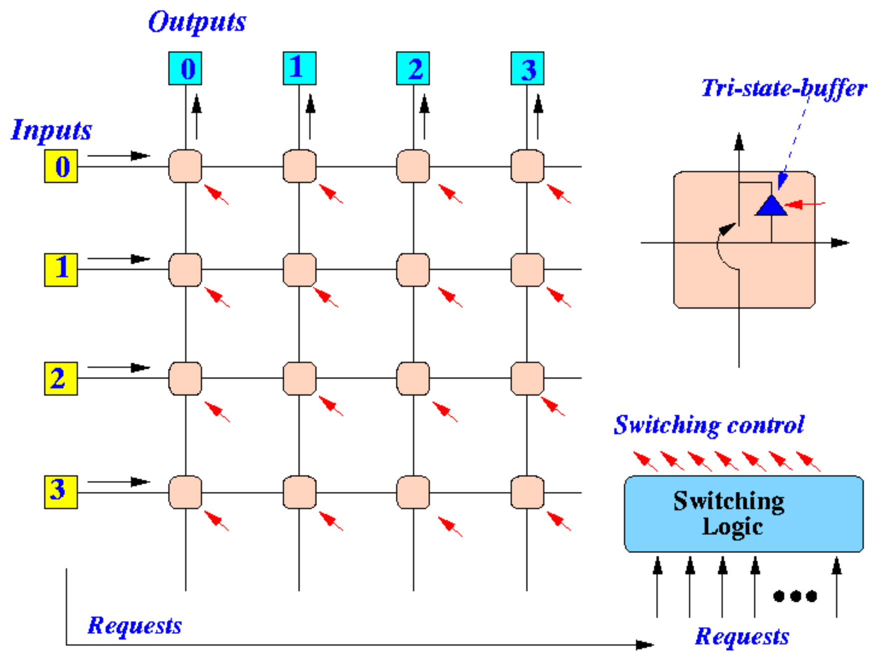


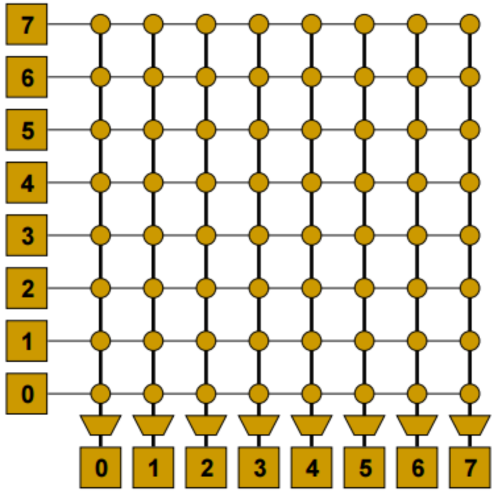
Switching via an interconnection network

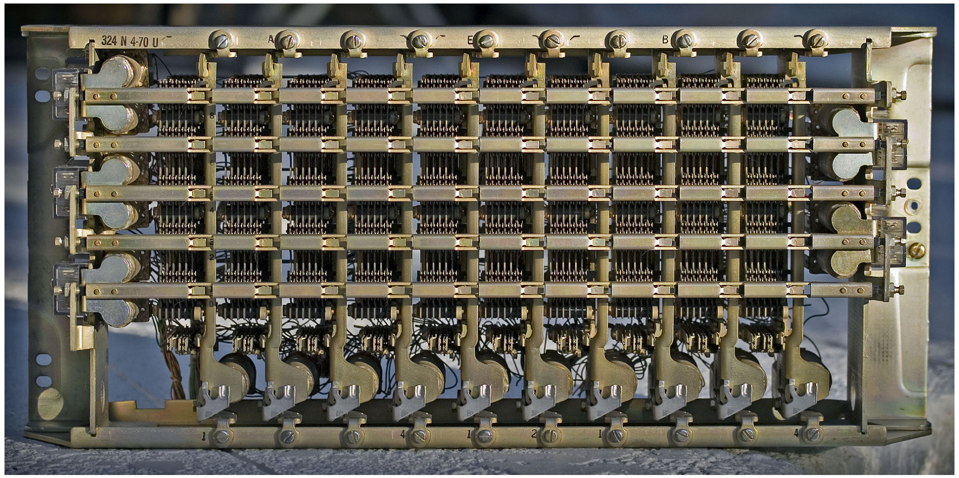


Crossbar

- Every node connected to all others (non-blocking)
- Good for small number of nodes
- Pros:
  - Low latency and high throughput
- Cons:
  - Expensive
  - Not scalable O(N^2) cost

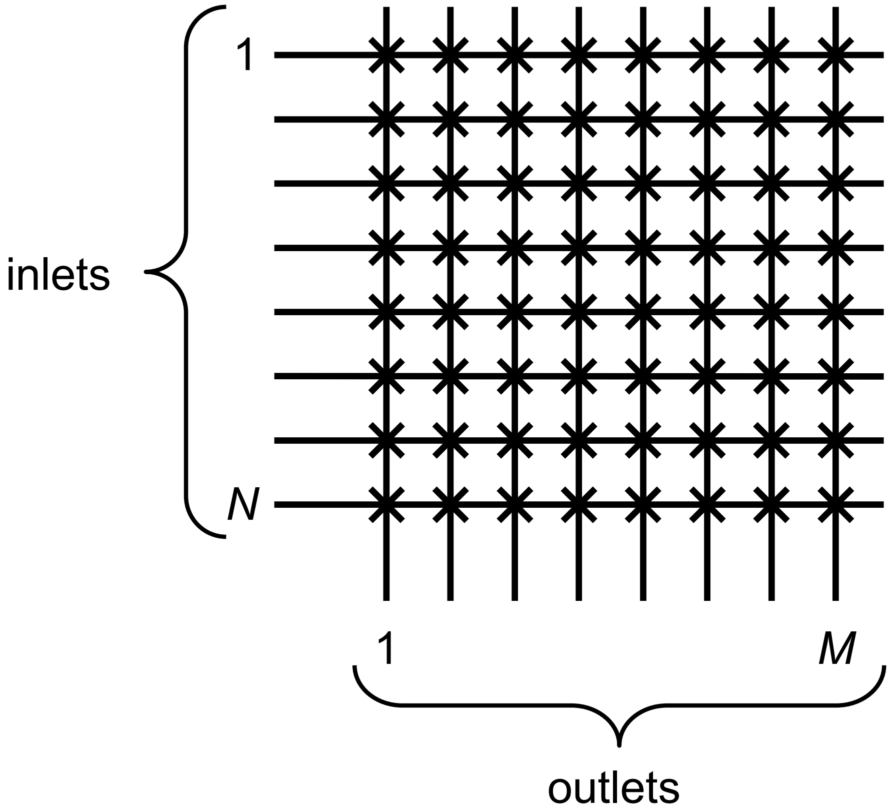


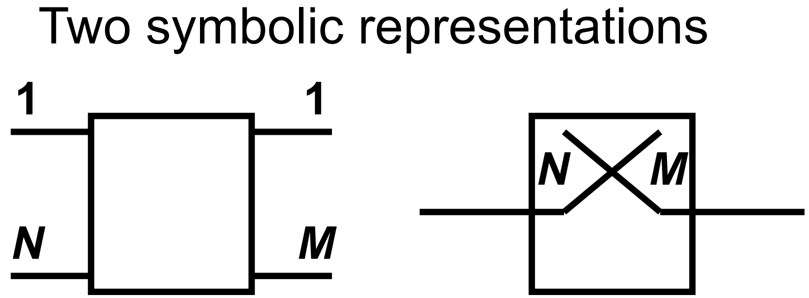


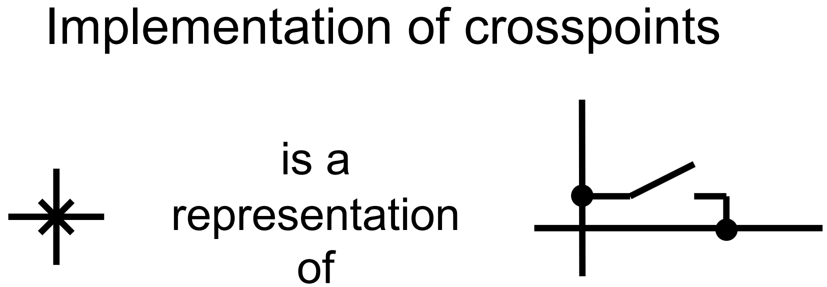


Western Electric 100-point six-wire Type B crossbar switch

N×M Crossbar switch





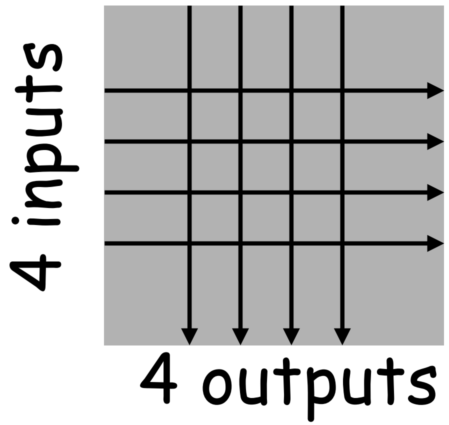


- Originally electromechanical but now commonly electronic
  - N > M: concentration (blocking if more than M sources active)
  - N < M: expansion
  - N = M: non-blocking square array
- Shortest path: 1; Longest path: N+M-1 hops
- Bisection bandwidth: N or M. R×R crossbar switch: R

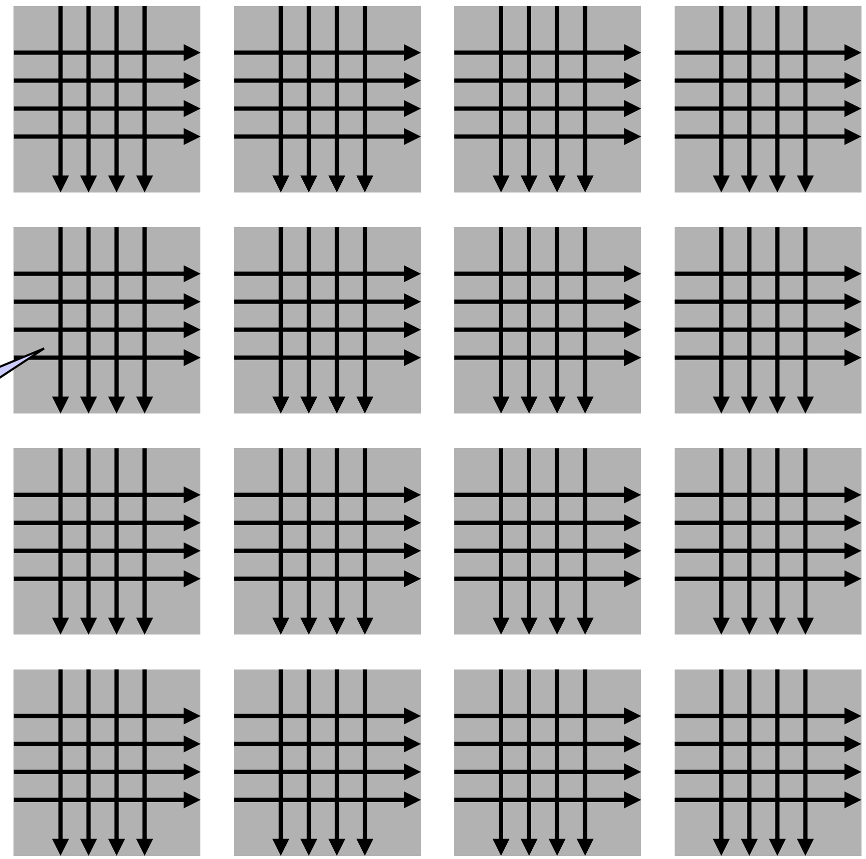
Scaling number of outputs:

Trying to build a crossbar from multiple chips

Building Block:



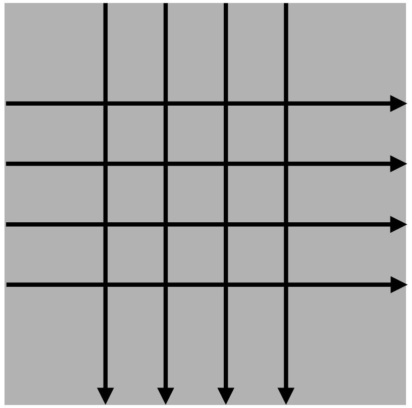
16x16 crossbar switch:



Eight inputs and eight outputs required!

N×N Crossbar switch Limiting factors

1. $N \times N$ crosspoints per chip
2. It's not obvious how to build a crossbar from multiple chips,
3. Capacity of "I/O"s per chip.



- State of the art: About 300 pins each operating at 3.125Gb/s ~= 1Tb/s per chip.
- About 1/3 to 1/2 of this capacity available in practice because of overhead and speedup.
- Crossbar chips today are limited by "I/O" capacity.

Scaling a crossbar

- Scaling the capacity is relatively straightforward (although the chip count and power may become a problem).
- What if we want to increase the number of ports?
- Can we build a crossbar-equivalent from multiple stages of smaller crossbars?
    - If so, what properties should it have?

Non-blocking network

A circuit-switching network is said to be *non-blocking* if it can handle all circuit requests that are a permutation of the input and outputs.

- Strictly non-blocking

  A network is strictly non-blocking if any permutation can be set up incrementally, one circuit at a time, without the need to reroute (or rearrange) any of the circuit that are already set up.

- Rearrangeable non-blocking

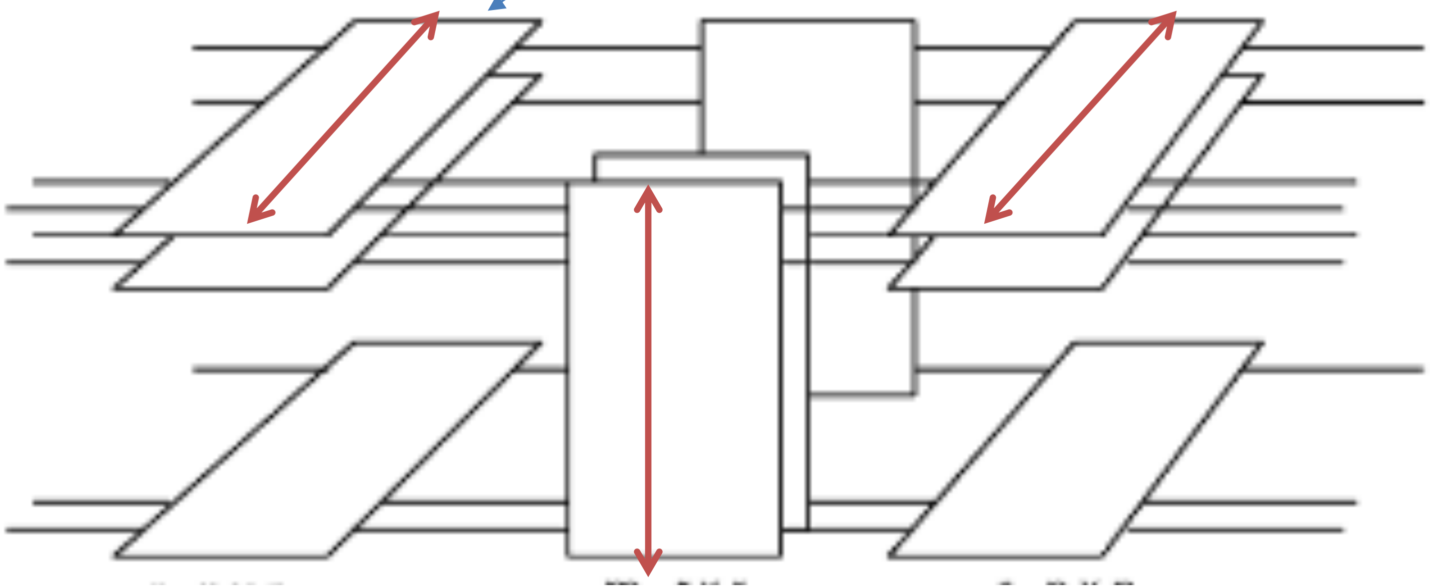
  A network can route circuits for arbitrary permutations, but incremental construction of a permutation may require rearranging some early circuits.

Clos network (m, n, r)

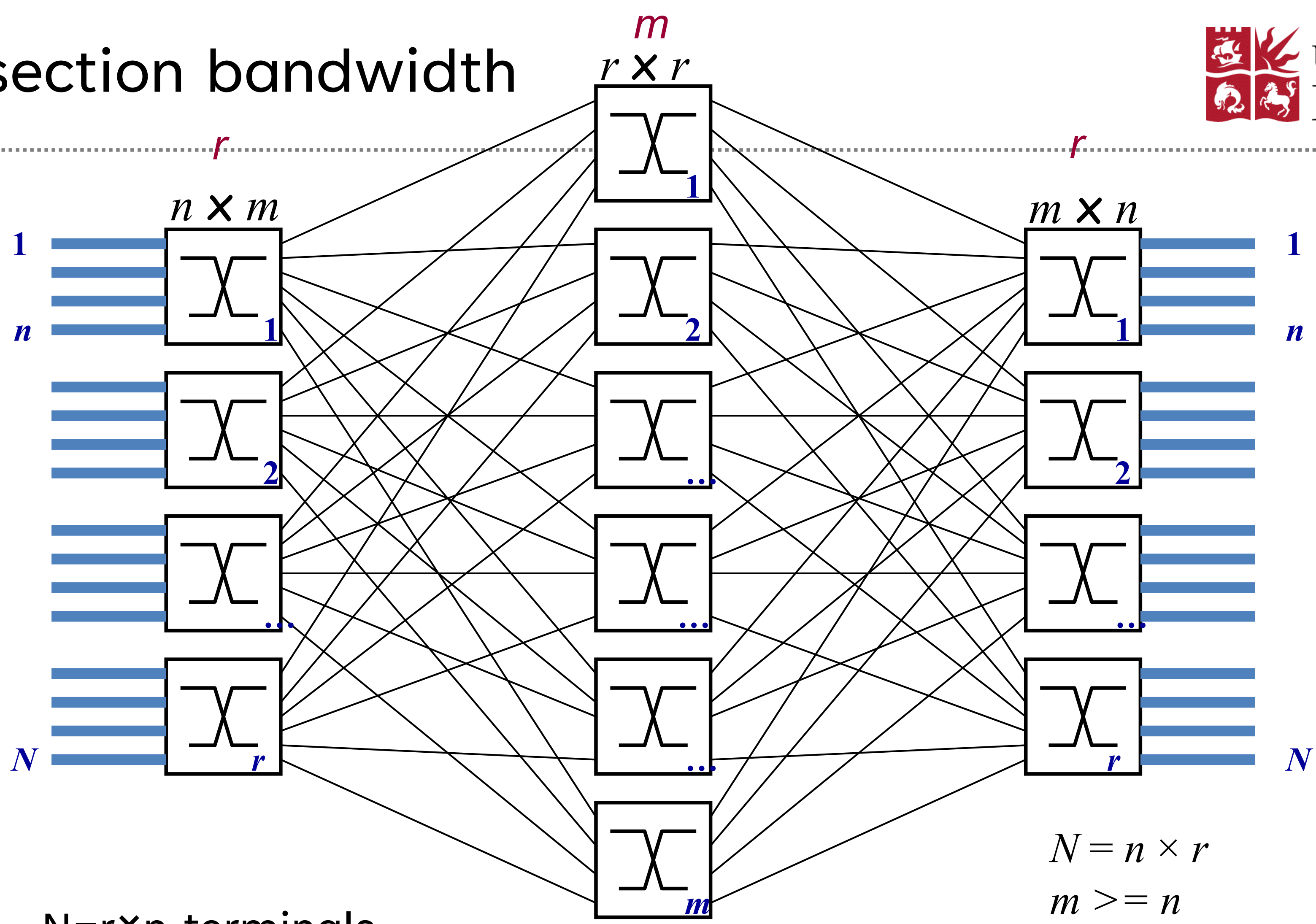
- Clos networks consists of :
  - r n×m input switches
  - m r×r middle switches
  - r m×n output switches

Crossbar switch

Total node number: N =*n*×*r*



Bisection bandwidth



$N = n \times r$

$m >= n$

N=r×n terminals

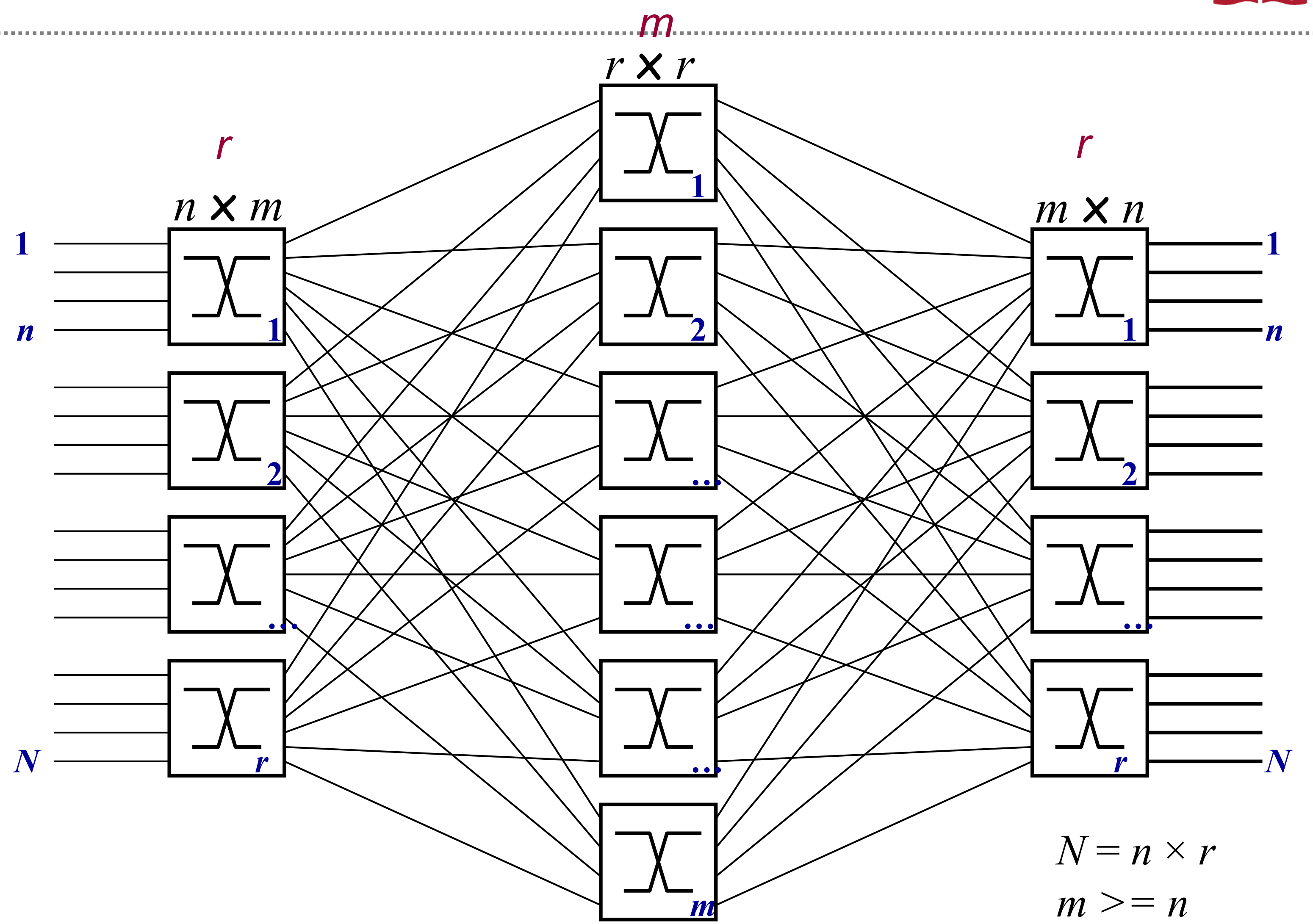
Horizontal cut:

$B_c = mr$

Vertical cut:

$B_c = 2nr = 2N$

Number of Crosspoint



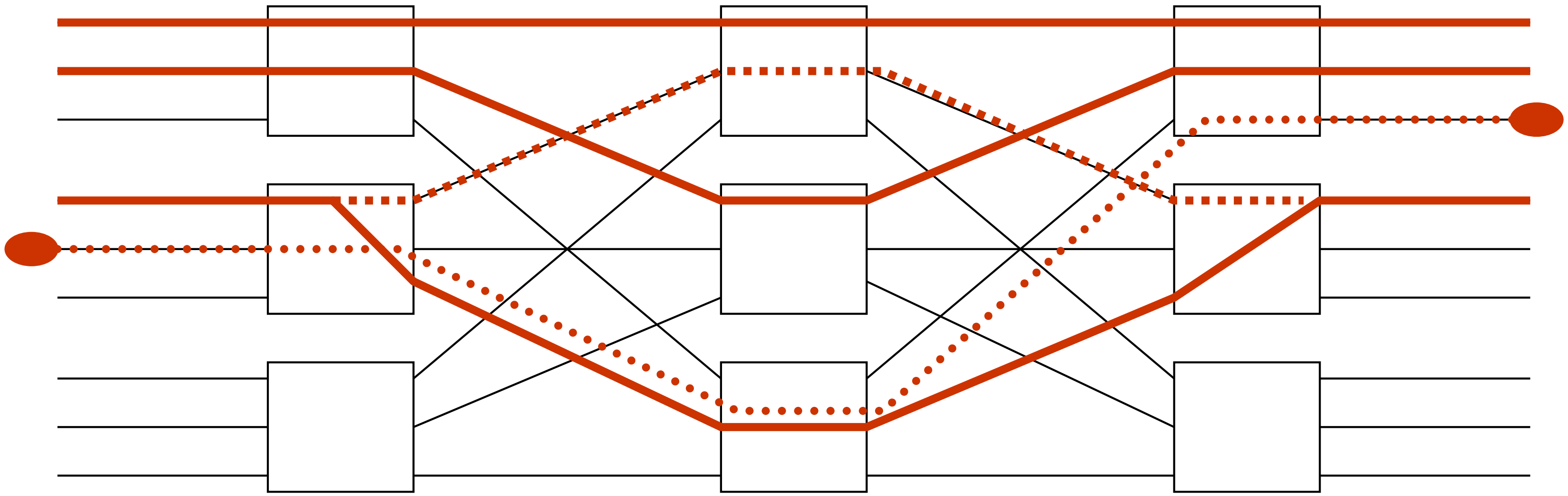
$N = n \times r$

$m >= n$

Calculate the number of Crosspoint: $2m \times n \times r + mr^2$

With $m = n$ is a Clos network non-blocking like a crossbar?

Consider the example: scheduler chooses to match
(1,1), (2,2), (4,4), (5,3), ...



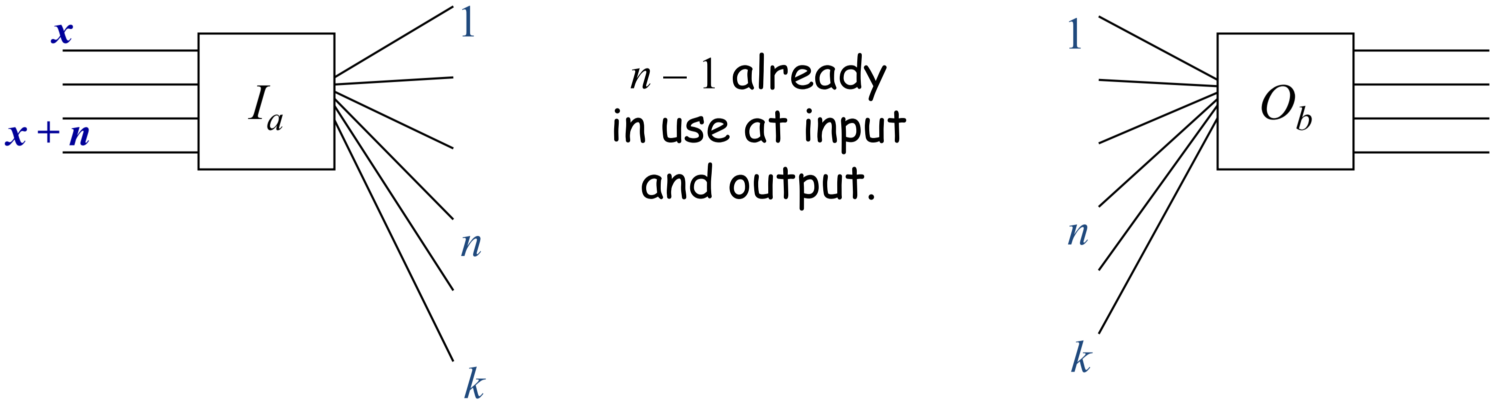
By rearranging matches, the connections could be added.
Q: Is this Clos network "rearrangeably non-blocking"?

Strictly non-blocking Clos Network

Clos' Theorem:

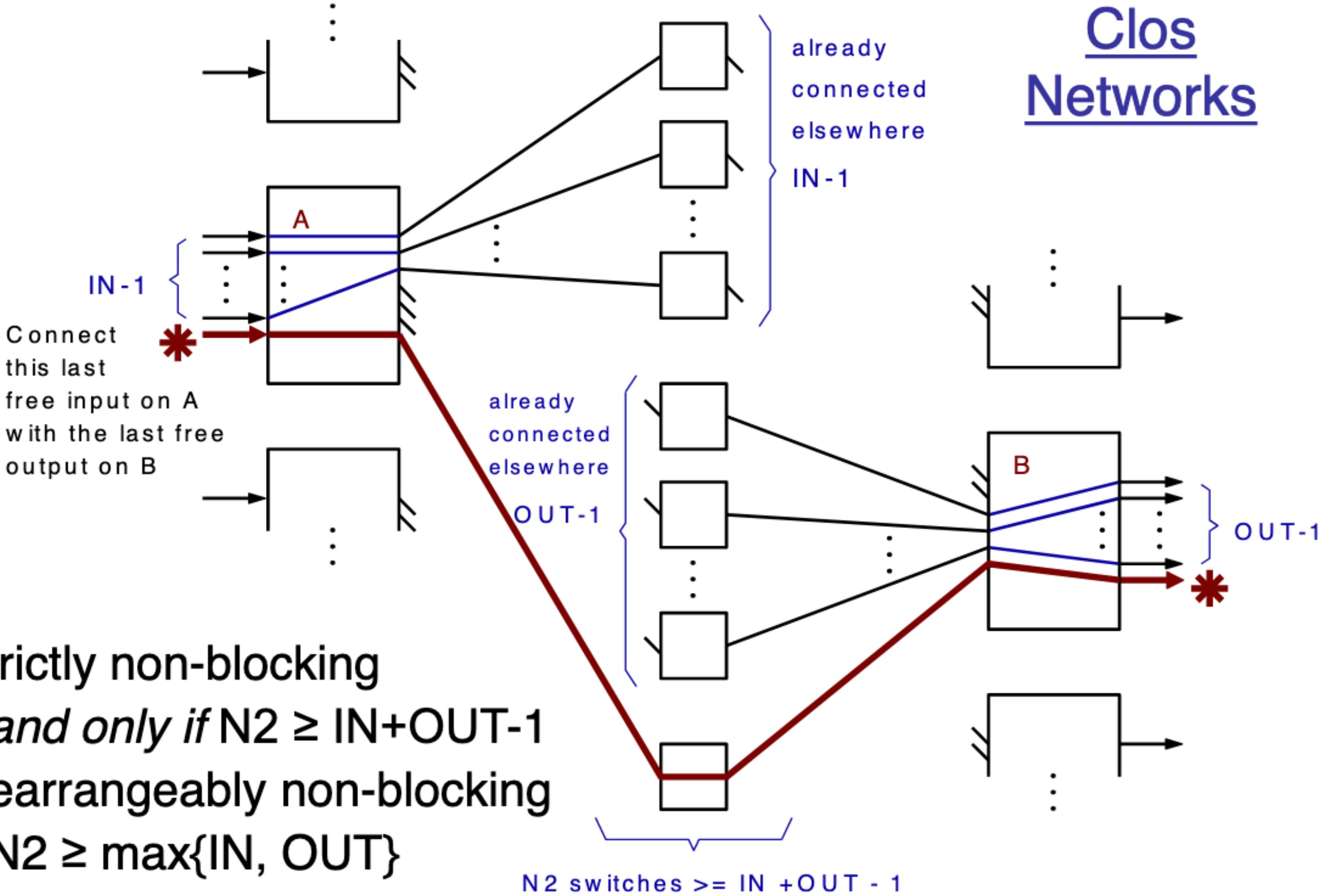
If $m >= 2n - 1$, then a new connection can always be added without rearrangement.

Clos Theorem



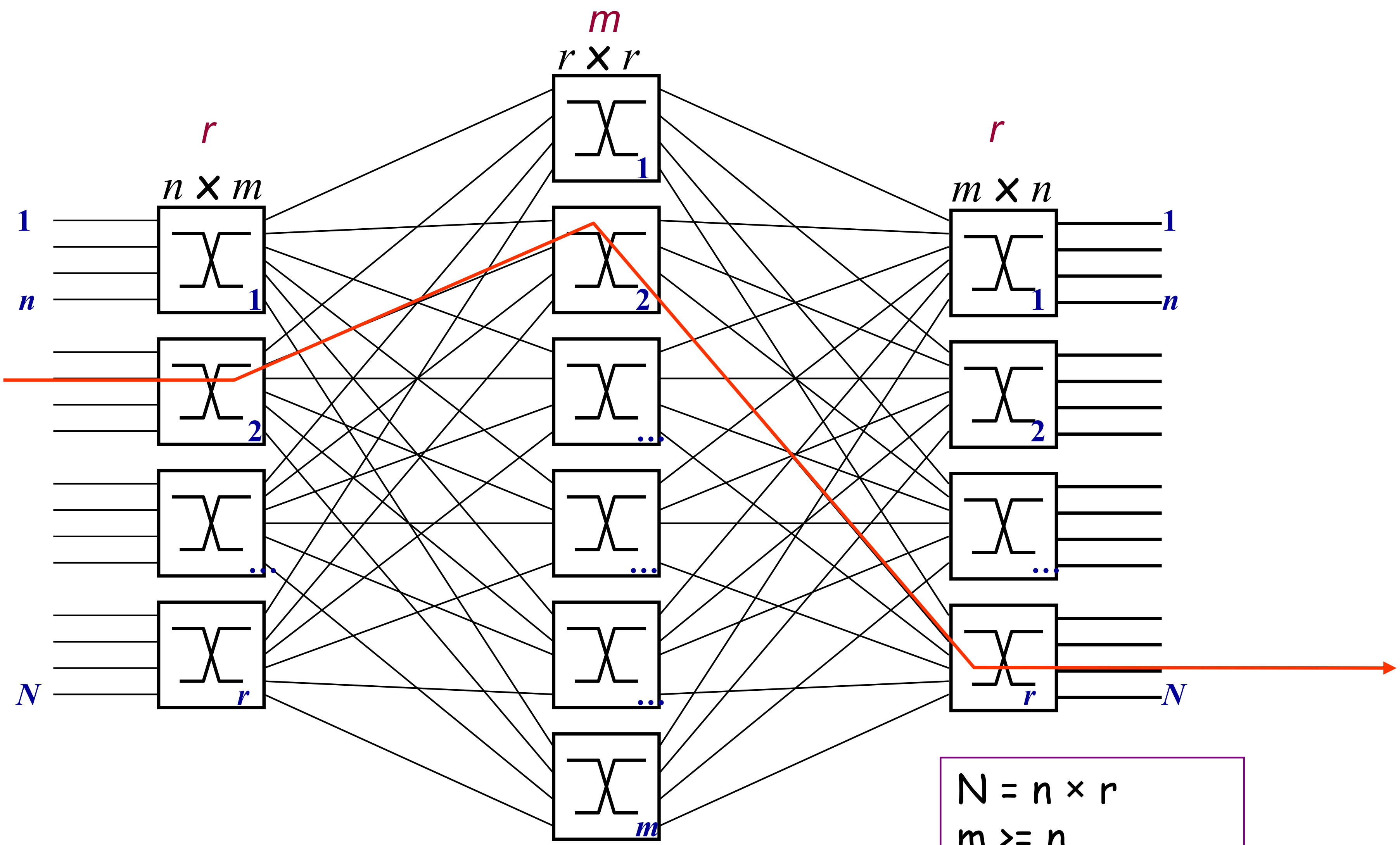
1. Consider adding the n-th connection between 1st stage I_a and 3rd stage O_b.
2. We need to ensure that there is always some center-stage M available.
3. If $k > (n - 1) + (n - 1)$, then there is always an M available. i.e. we need $k >= 2n - 1$.

Proof



- Strictly non-blocking
  if and only if N2 ≥ IN+OUT-1
- Rearrangeably non-blocking
  if N2 ≥ max{IN, OUT}

3-stage Clos Network (m, n, r) Vs. N×N crossbar switch



N = n × r
m >= n

Rearrangeable non-blocking Clos Network

Clos' Theorem:

A Clos network with *m >= n is rearrangeable.*

Action: Check Hall's Theorem for the proof

Clos Three-Stage Non-Blocking Switch

Minimum number of Crosspoints

- Since xpt = r(nm)+ m(rxr)+r(mn)=2rmn+mr^2=2Nm+m(N/n)^2,

  In 3-stage with no blocking:

  xpt = 2N(2n-1)+(2n-1)(N/n)^2 {since m = (2n-1) }

For large n, xpt = 4Nn + 2N^2/n-2N- N^2/n^2

If we differentiate with respect to n,

$$\frac{\partial Xpt}{\partial n} = 0 \Rightarrow 4N - \frac{2N^2}{n^2} = 0 \Rightarrow n = \sqrt{\frac{N}{2}}$$

If N is given, then best choice that minimizes the number of crosspoints is n=sqrt(N/2)

$$\text{and Min } xpt = 4N\sqrt{\frac{N}{2}} + 2\sqrt{2} \times \frac{N^2}{\sqrt{N}} = 4\sqrt{2} \times N^{\frac{3}{2}}$$

{Substitute for n)

Example

Design a three-stage, 500 × 500 switch with m = 25 and n = 25. Compute the number of crosspoints.

Solution

N = 500;

r = N/n = 500/25 = 20;

In the first stage we have N/n or r=20 crossbars, each of size n×m (25 × 25). In the second stage, we have m=25 crossbars, each of size r×r (20 × 20). In the third stage, we have r=20 crossbars, each of size 25 × 25. The total number of crosspoints is

$$2rnm + mr^2 = 35000$$

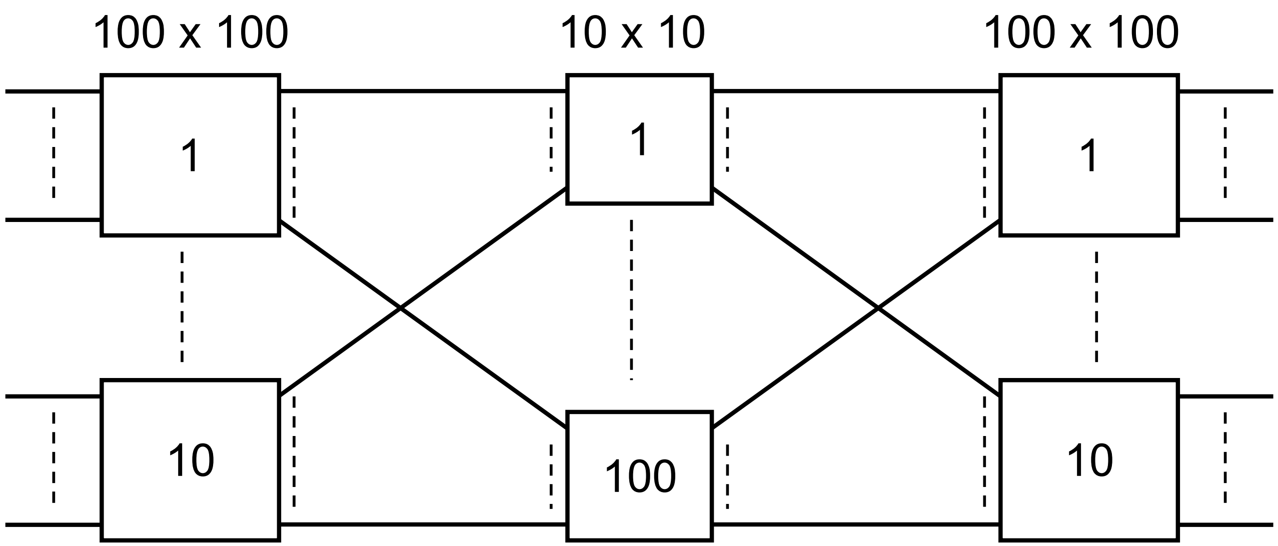
crosspoints. This is 14 percent of the number of crosspoints in a single-stage switch (500 × 500 = 25,0000).

Question

Design a three-stage, 500×500 switch fabric for strictly non-blocking clos network. $n = 25$;

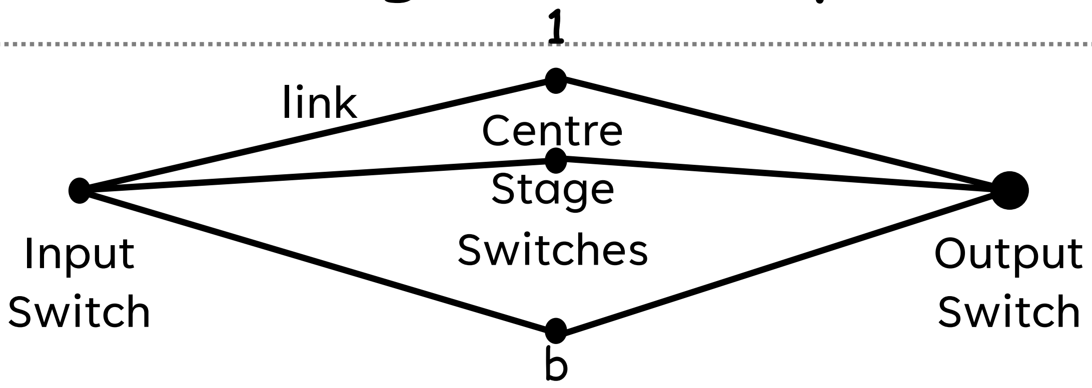
A 1000 x 1000 three-stage Clos (100, 100, 10):

- Number of crosspoints = 10(100 x 100) + 100(10 x 10) + 10(100 x 100)
  $= 10^5 + 10^4 + 10^5 = 2.1 \times 10^5$
- Less than 1000 x 1000 = 10^6, as required for a single non-blocking crossbar
- 3-stage switch has blocking probability depending on link occupancy p
  - p is often measured in Erlangs, a dimensionless unit



Evaluation of Blocking Probability

Channel graph for 3-stage switch



A route is two links connected in series, thus connecting the input switch and the output switch.

- The channel graph shows all possible paths from any input to any output
  - p is the probability that a single link within the 3-stage switch is busy
  - q=1-p is the probability that the link is idle.
- If b is the number of centre stage switches, a call is blocked if all b parallel routes are busy.
- An individual route is busy if either of the links forming that route are busy
  - Prob(link busy) = p
  - Prob(link free) = 1 – p
  - Prob(route free) = Prob(both links free) = $(1 – p)^2$
  - Prob(route busy) = $1 – (1 – p)^2$
  - Prob(blocking) = Prob(all routes busy) = $(1 – (1 – p)^2)^b$

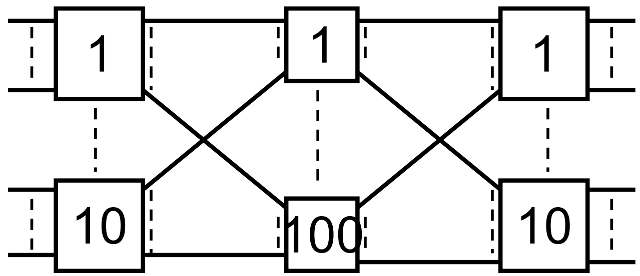
Blocking probabilities for 3-stage 1000 x 1000 switch

- In the previous 3-stage 1000x1000 switch example, b=m = 100
- If a is the average line occupancy on each input or output link, internal link occupancy, p = a (square array of switches)

p	blocking probability
0.1	10^{-73}
0.5	10^{-13}
0.6	2.4×10^{-7}
0.7	8×10^{-5}
0.8	1.7×10^{-2}
0.9	0.37

Virtually non- blocking

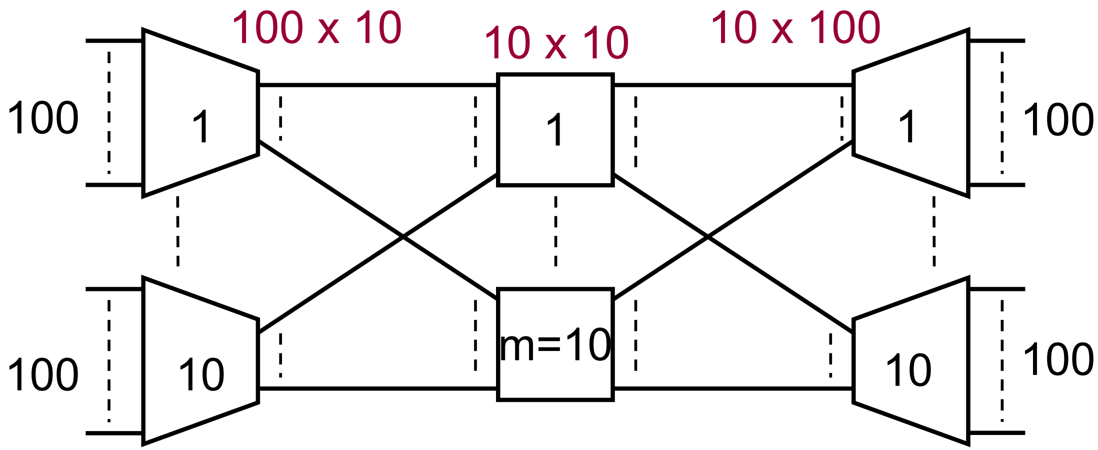
100 x 100 10 x 10 100 x 100



- Blocking probability rises as the traffic load is increased
- If the blocking probability is less than 0.01 (1%), the switch is referred to as "virtually non-blocking"
  - Commercial telephone systems are usually virtually non-blocking

Reducing crosspoints using concentrators (small load)

- Small line occupancy leads to negligible blocking probability – switching complexity can be reduced by concentrating traffic at first stage.



- Number of crosspoints = 10(100 x 10) + 10(10 x 10) + 10(10 x 100)

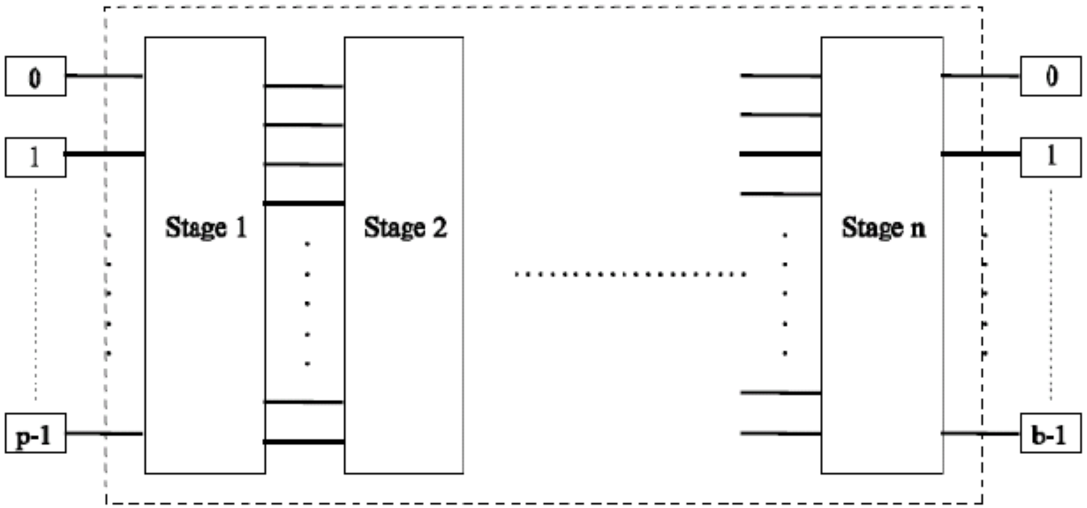
  $= 10^4 + 10^3 + 10^4 = 2.1 \times 10^4$

- Internal link occupancy, p = 10a (10:1 concentration)
- Prob(blocking) = $(1 - (1 - p)^2)^{10}$

a	0.03	0.04	0.05	0.06
p	0.3	0.4	0.5	0.6
blocking	1.2×10^{-3}	1.15×10^{-2}	5.6×10^{-2}	0.17

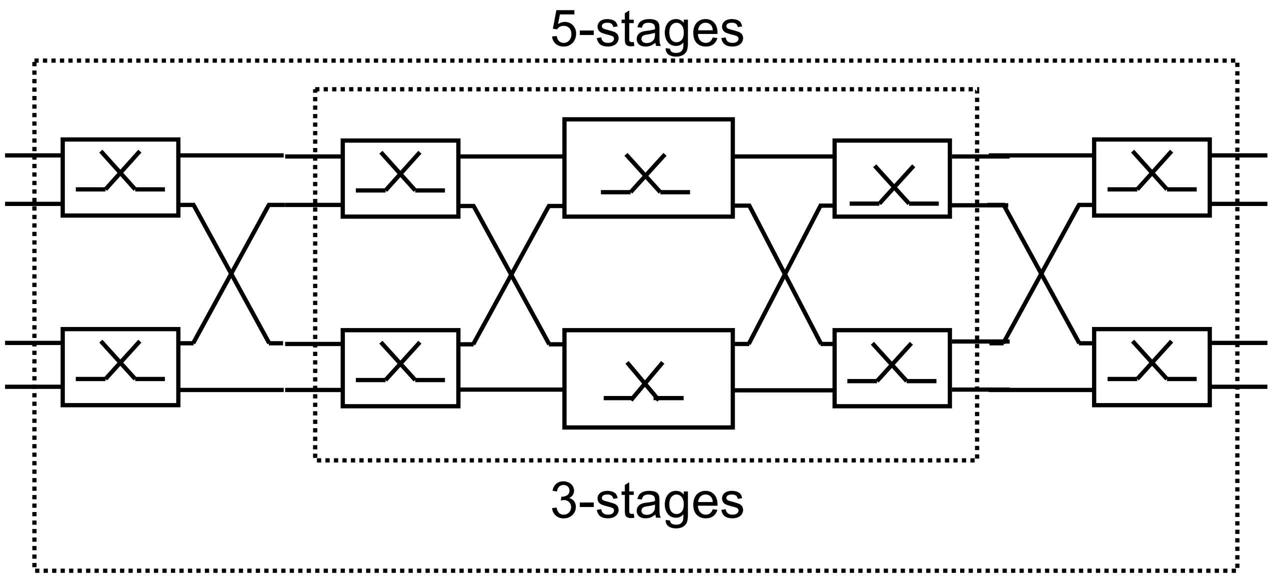
Multistage interconnection networks (MIN)

- Try to emulate the cross-bar connection.
  - Realizing permutation without blocking
  - Using smaller cross-bar(2x2, 4x4) switches as the building block. Usually O(Nlg(N)) switches (lg(N) stages.



Multi-Stage Switches

- The cost of space switches is proportional to the number of crosspoints
- By splitting the crossbar switch into smaller chunks and interconnecting them, it is possible to build multistage switches with fewer crosspoints
- The number of stages reduces the number of necessary crosspoints, but requires greater interconnectivity

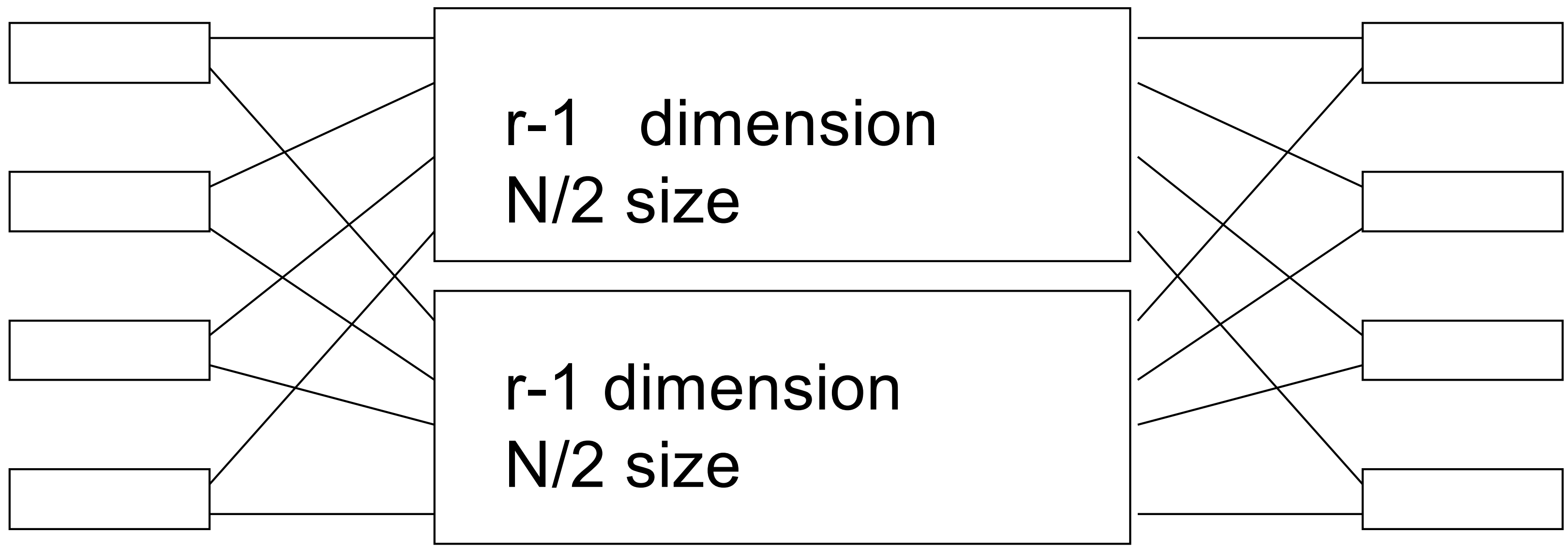


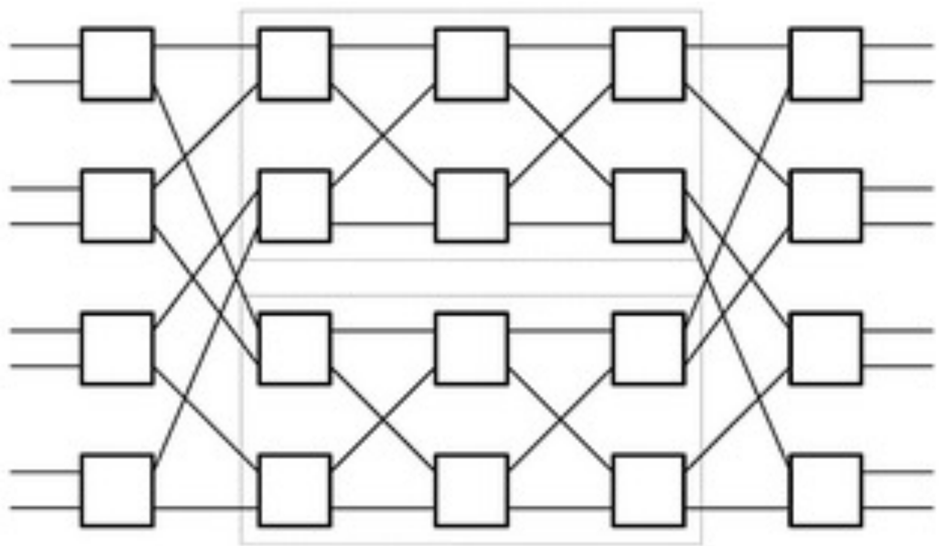
Recursive Construction: basis

The basic element: 2×2 switches

The two states:

Recursive Construction: Benes Network





a (2,2,4) rearrangeable clos network, using two (2,2,2) Clos Networks as Middle switches

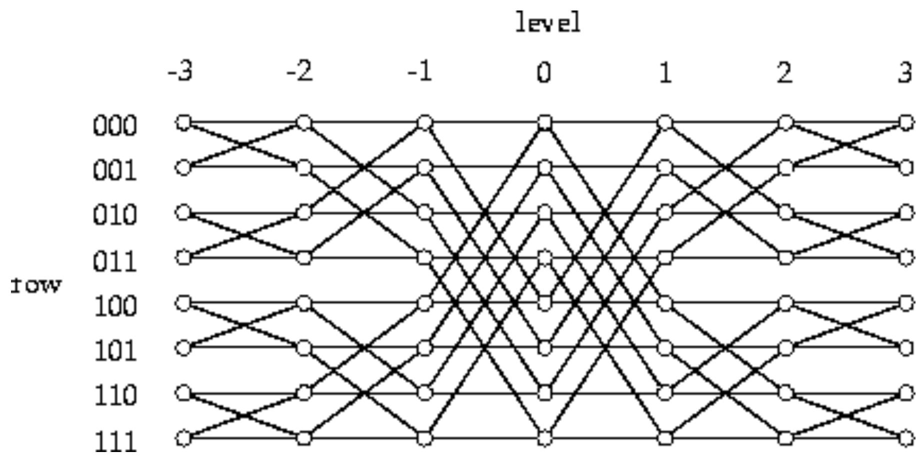
Benes Networks

- A network constructed from 2 $\times$ 2 switches
  - Minimum number of crosspoints

Network: $N = 2^i$

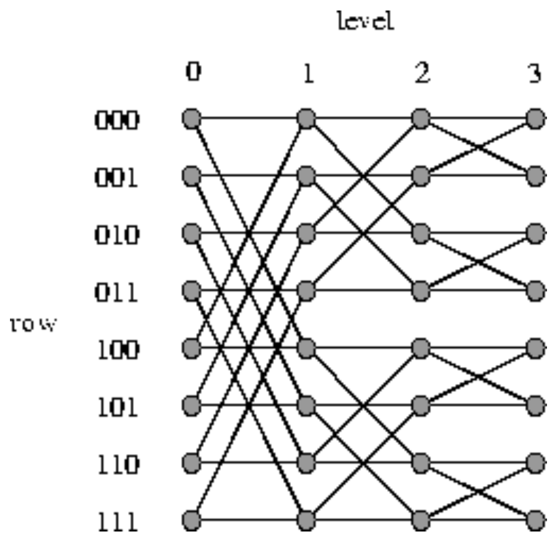
$2i - 1$ Stages of :

2^{i-1} 2$\times$2 switches

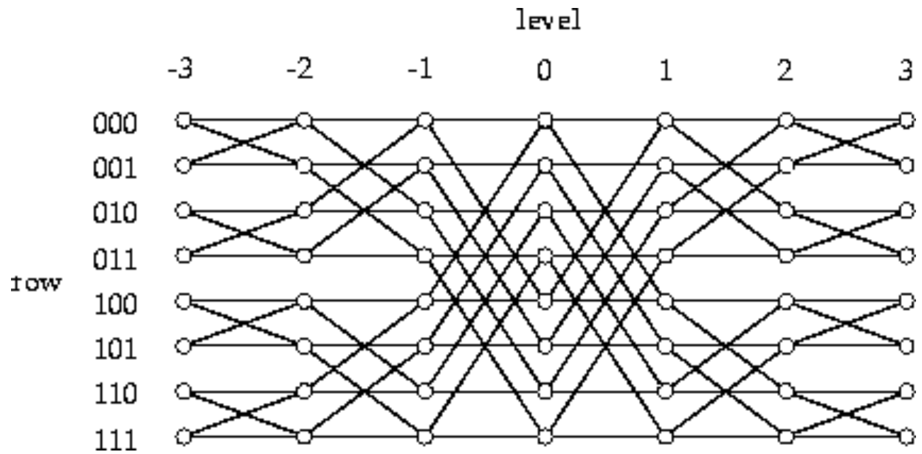


Number of crosspoints: $4(2i-1)\,2^{i-1}$

Multi-stage Interconnection Networks (MINs)



(a) An 8-input butterfly network

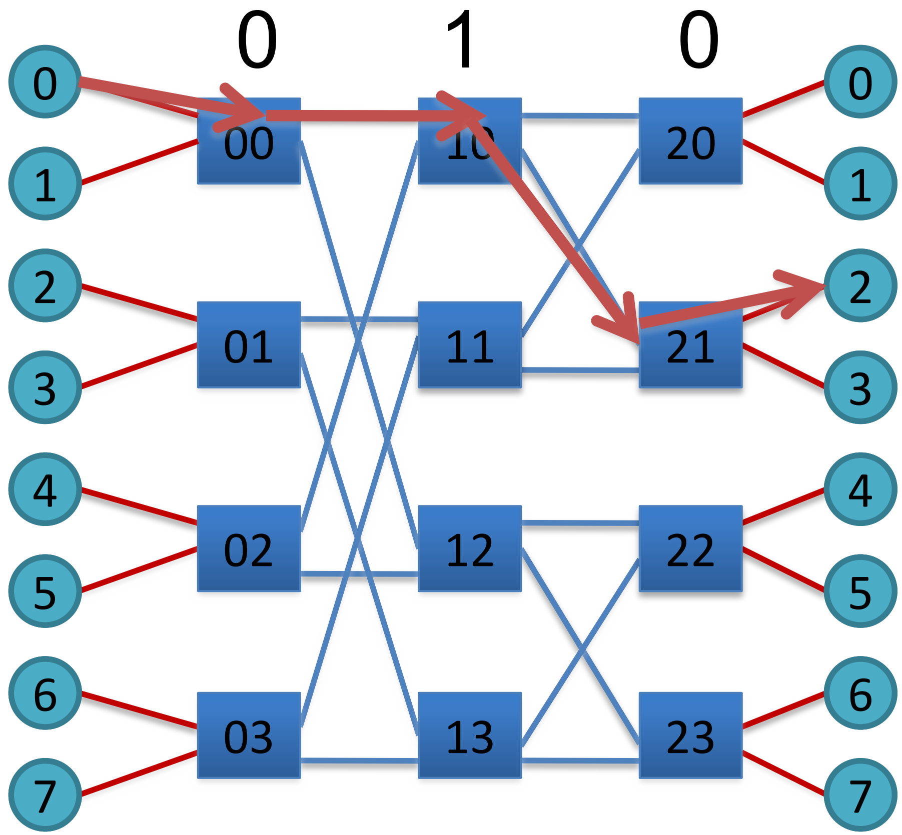


(b) An 8-input Benes network

- MINs can be blocking or non-blocking
  - Blocking: there exist some permutation that results in link contention.
  - Non-blocking: any permutation can be realized without link contention
- Butterfly network is blocking.
- Benes network is non-blocking.

Butterfly network

- k-ary n-fly: nk^{n-1} switch nodes
- K^n Input terminal nodes
- Degree = 2k
- Diameter: n+1
- Bisection bandwidth = 2^k
- Routing from 000 to 010
  - Dest. address used to directly route packet
  - Bit n used to select output port at stage

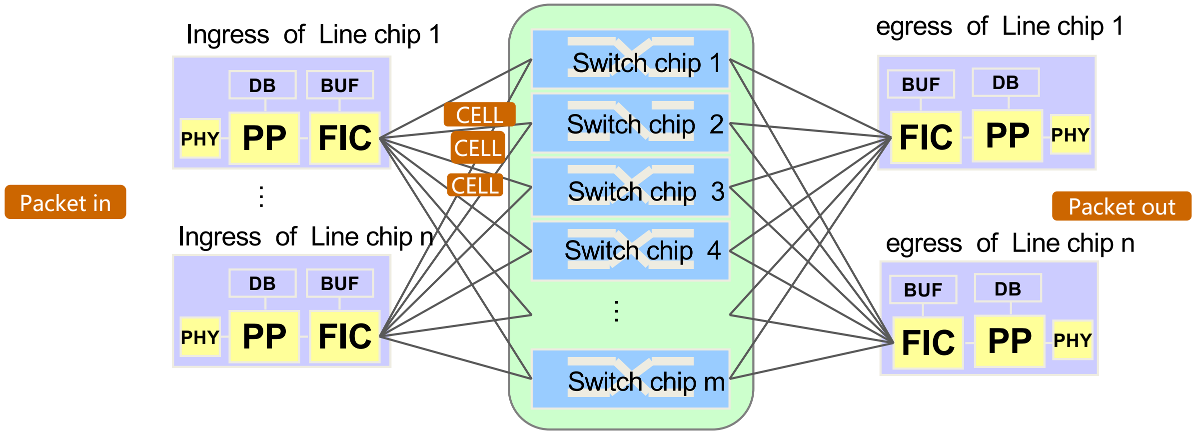


2-ary 3-fly butterfly network

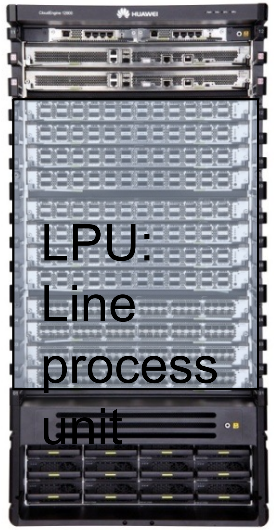
K-ary n-fly butterfly networks

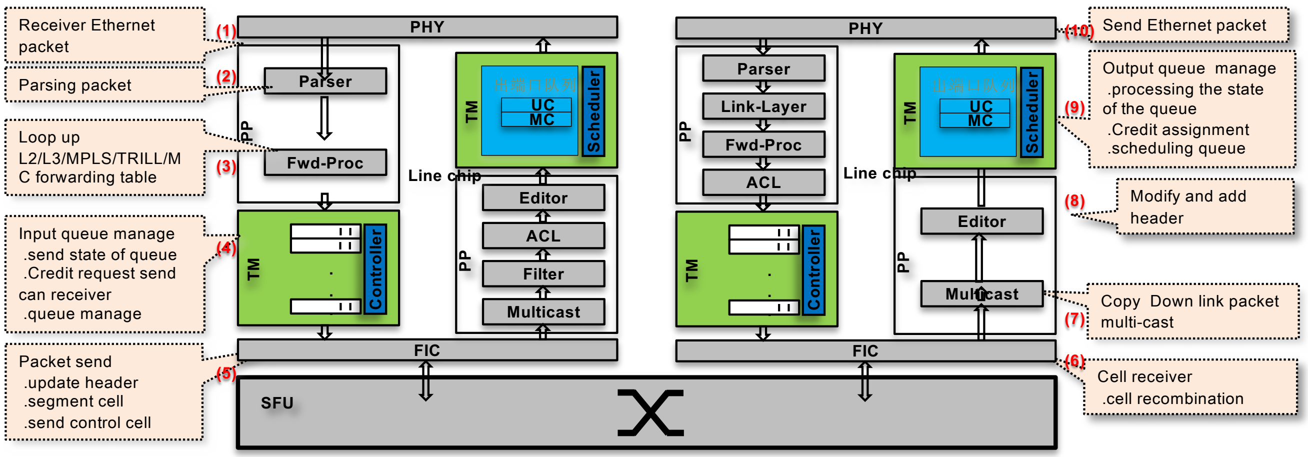
- Cons:
  - No path diversity: exactly on route from each source node to each destination node
  - Lone wires are needed to traverse at least half the diameter of the machine
- Pros:
  - Simple routing
  - Logarithmic diameter: $H=\log_k N+1=n+1$
  - Small scale switches

Architectures of large capacity switch



CLOS architecture for large capacity switch



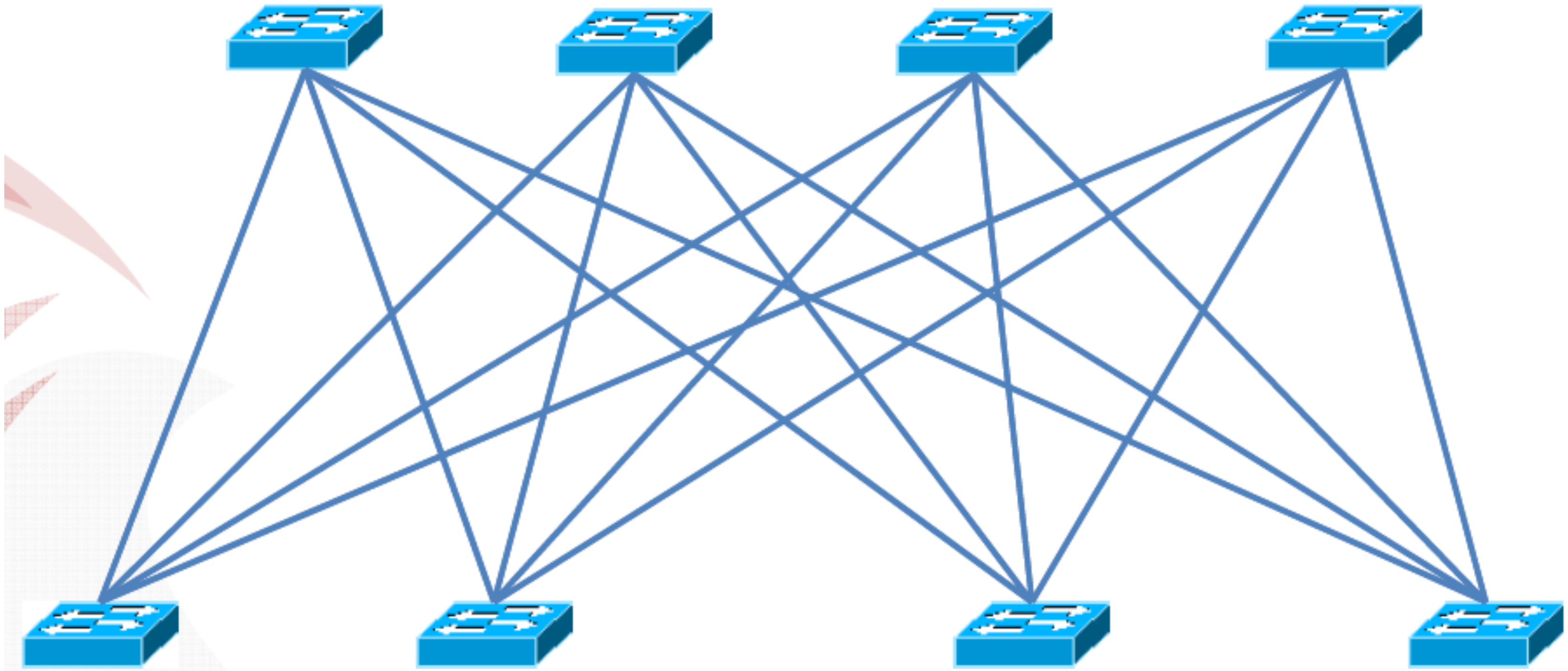


Process flow of forwarding packet

- For large capacity switch, it consist of many Line chips and switch chips, which form 3-stage-CLOS network.
- Switch chip : forward packet
- Line chip : consist of PP, TM and FIC.
- P P: process and edit packet。
- TM: scheduling and manage buffer
- FIC: send and receive CELL, and

Clos Networks

- Folded Clos: Leaf and Spine



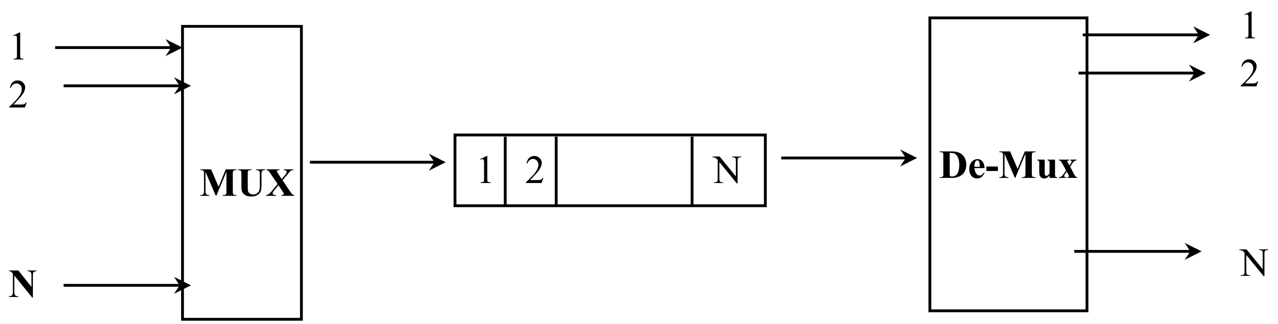
Summary

- Introduce indirect networks
- Clos network
  - Strictly non-blocking network
  - Rearrangeable non-blocking network
- Blocking analysis
- Multi-stage network

TDM switching

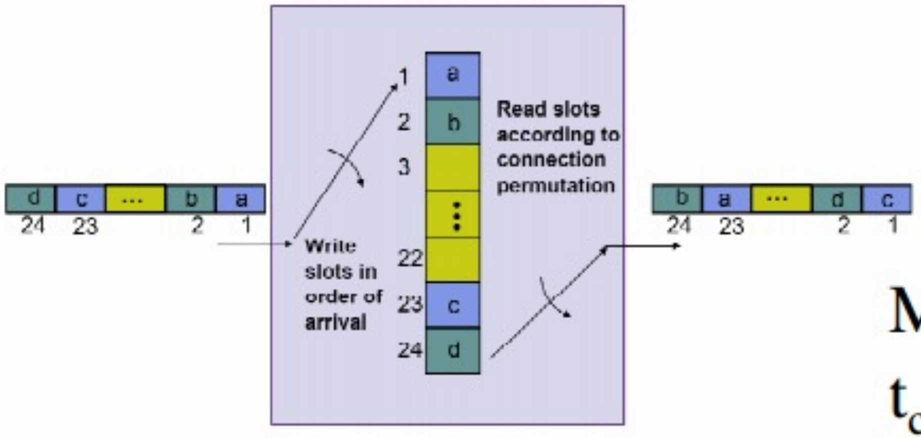
Multiplexers and Demultiplexers

- Multiplexer: aggregates sessions
  - N input lines
  - Output runs N times as fast as input
- Demultiplexer: distributes sessions
  - one input line and N outputs that run N times slower
- Can cascade multiplexers



Time Division Multiplexing Switching

- Mostly all modern circuit switches are time-division switches.
  - Time-slot interchanger (TSI)
  - Synchronous TDM
- Multiple low speed inputs share a high-speed line
- There is no need for address bits in each slot (Synchronous)
  - The slot could be a bit, a byte, or a longer block (Frame)
- Switching complexity can be greatly simplified if each inlet carries several telephone channels through the use of multiplexing



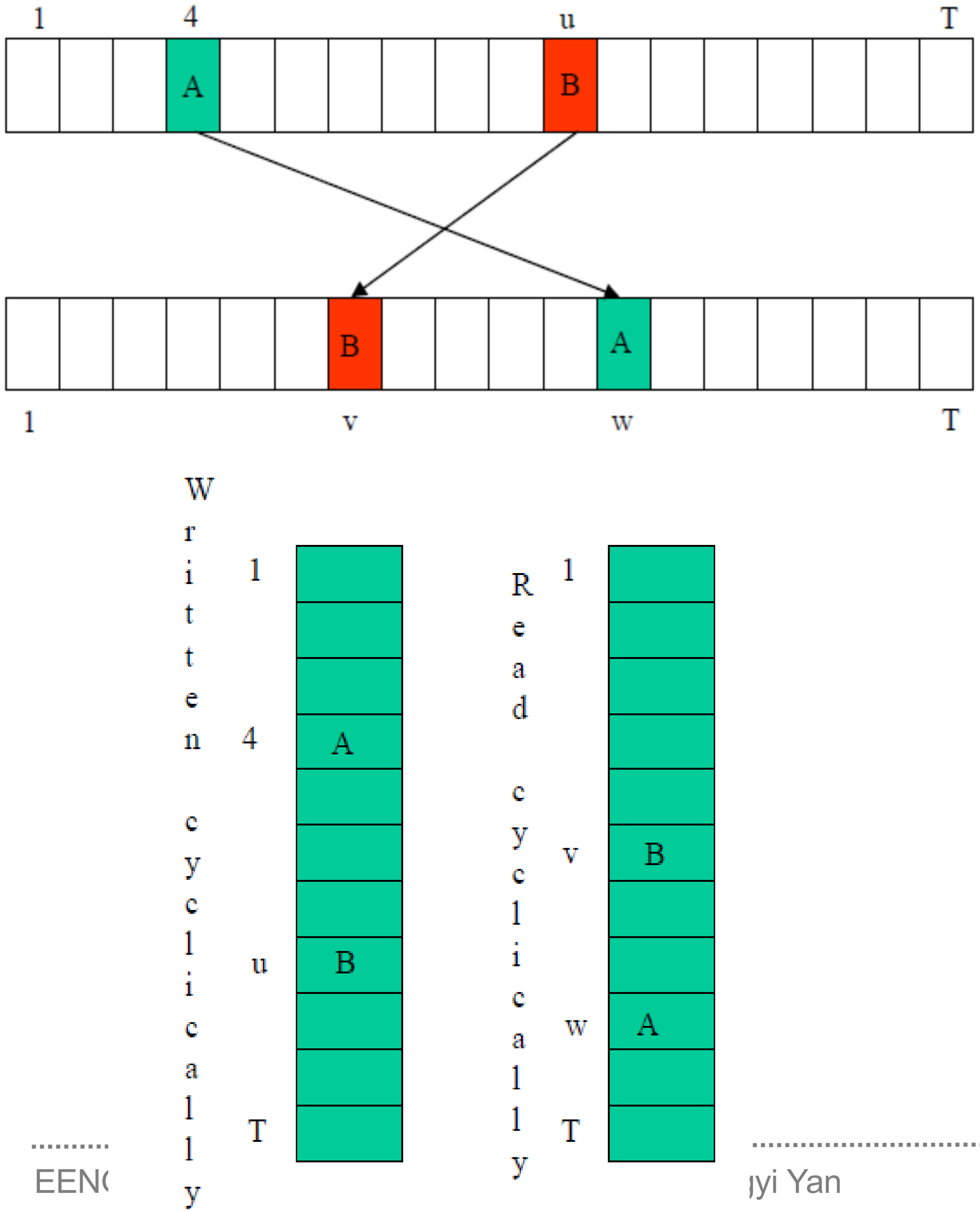
Time-slot interchange

Maximum # of slots=$125/(2\times t_c)$

t_c=memory cycle time (μ sec)

Read and write to shared memory in different order.

Time Slot Interchange (TSI) Switches

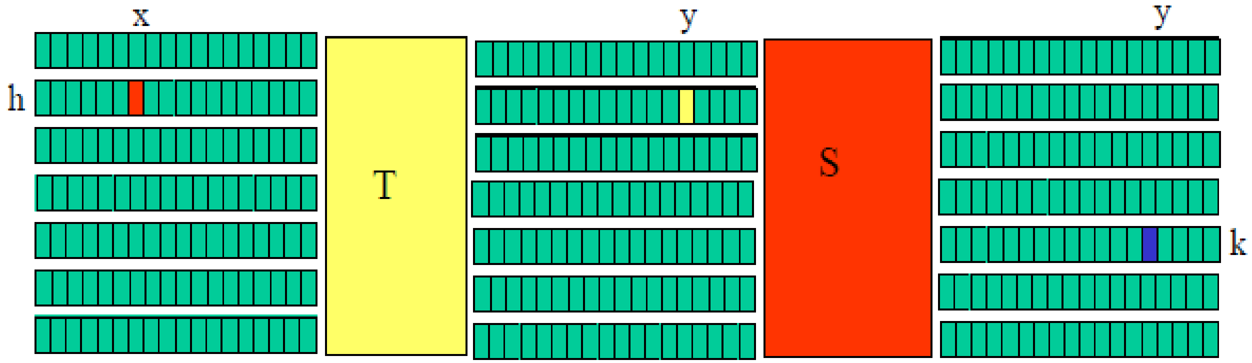


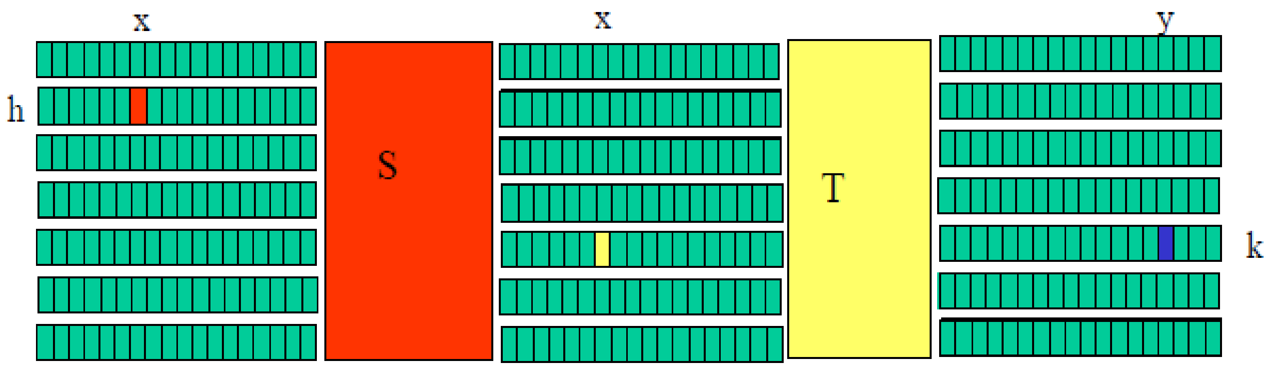
- Interchange is done by writing the whole frame into a memory in the normal order, but reading out in the required sequence
- Like the SPACE switch, a connection store holds the address of the slots required to be interchanged. They are held as long as connection is required and periodically executed.
- Random Access Memories (RAM) store the address of the input channels at the location of output channels
- During 125µs, T slots in the frame should be written and read from memory. Thus the memory access time should be faster than:

$$\tau \le \frac{125\mu s}{2T}$$

Mixed Time and Space Switches

To exchange slots between highways, we need a combination of Time (T) and Space (S) switches. Can use only one T and one S switch (TS) or (ST)





- In TS: the T switch transfers time slot x of input highway h, into time slot y at its output. The S switch opens its gate at time slot y to transfer its content from highway h to highway k.
- In TS, the transfer cannot be done if the output time slot of the T switch has already been used.

In ST: the S switch during time slot x transfers input from highway h, into highway k. The T switch of highway k transfers data from time slot x into time slot y.

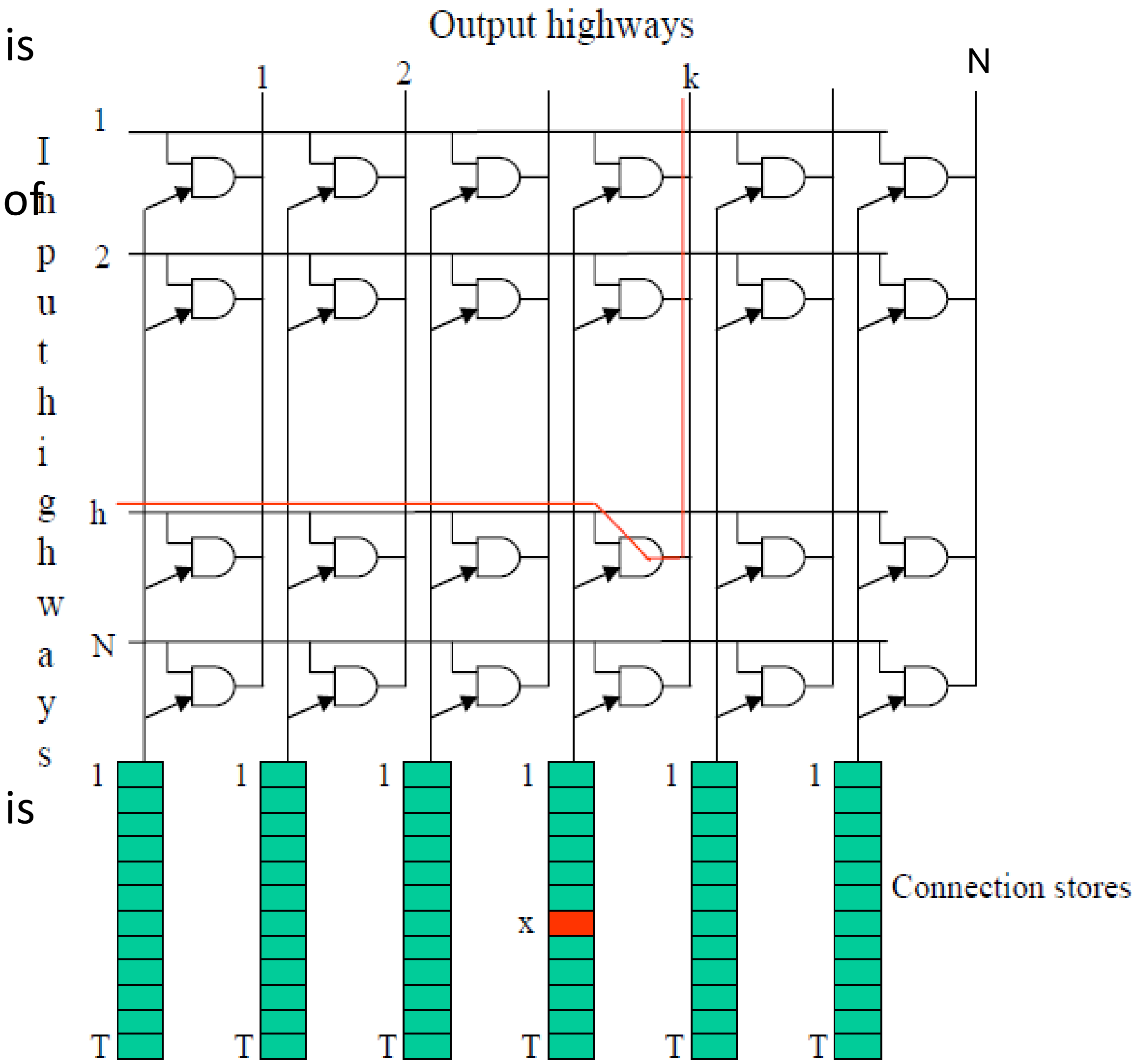
In ST, the transfer cannot be done, if at the time slot x, the output highway of the S switch (input highway of the T switch) has already been occupied.

In both TS and ST, the switching cannot be carried out if the wanted time slot is busy

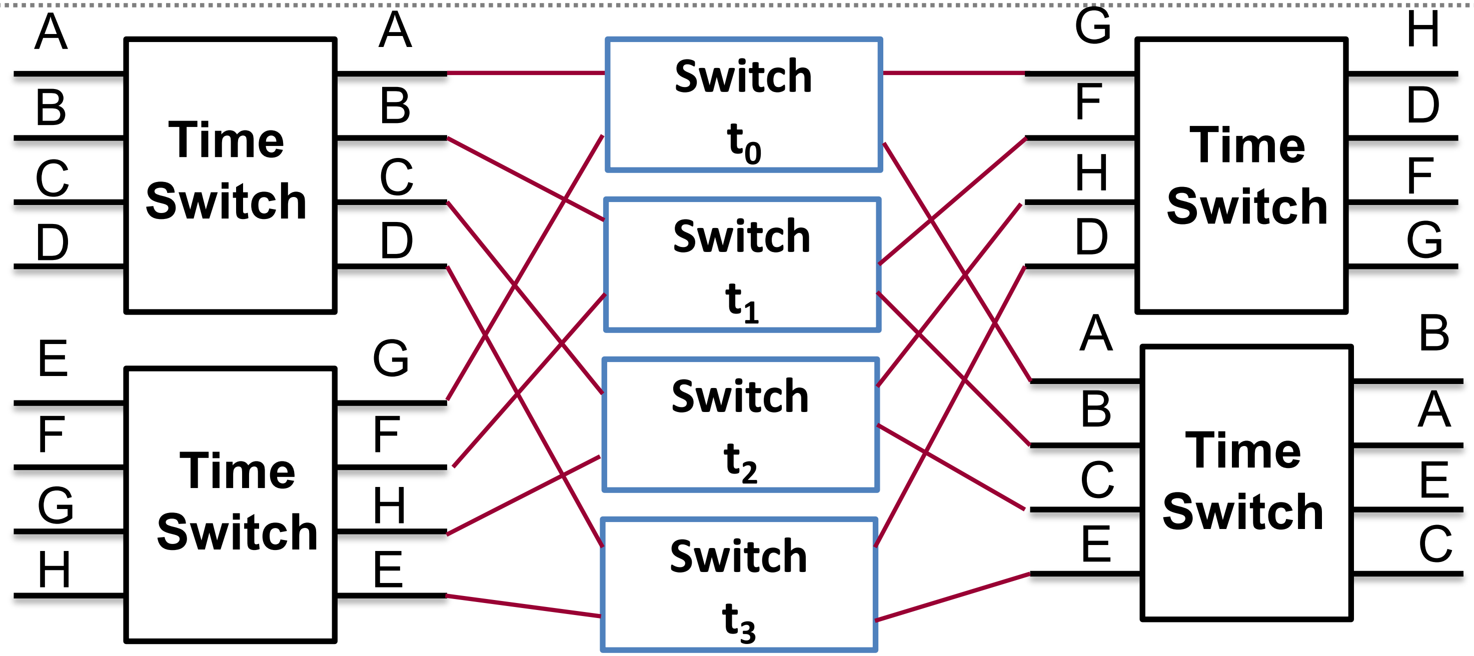
Thus the *Blocking probability* is equal to time slot occupancy

Connection Store for Space Switch
Switching timeslot x from input h to output k

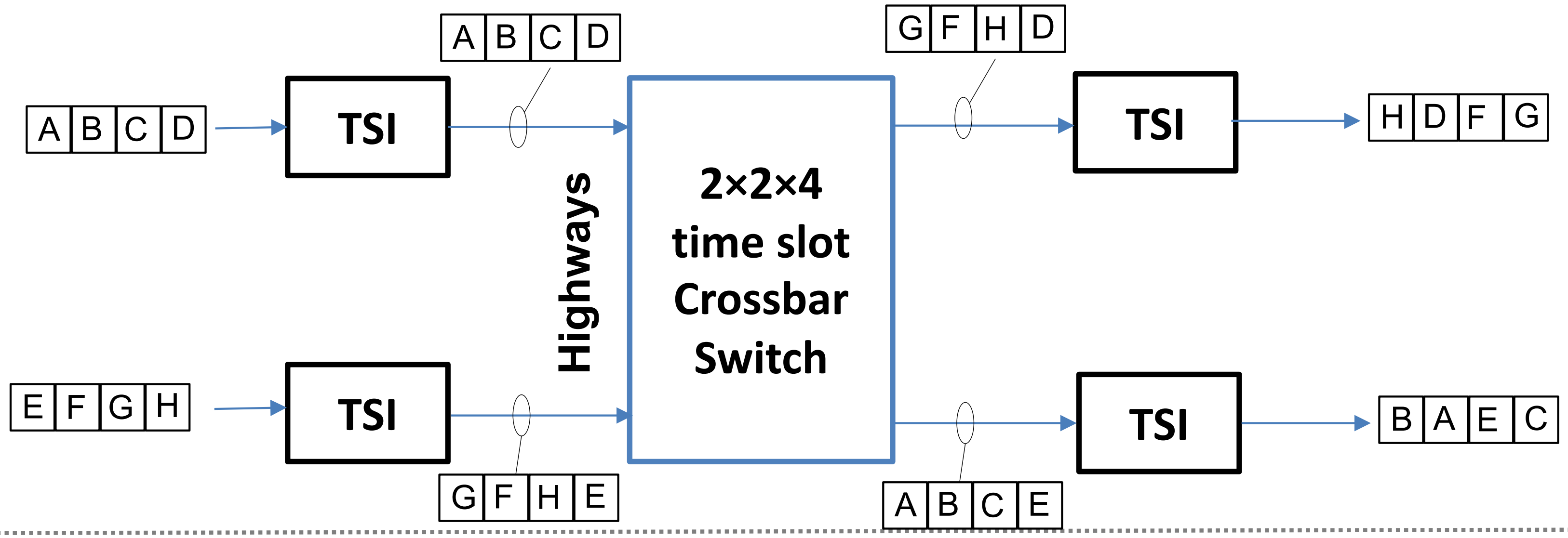
- At the timeslot x, the input highway is connected to output highway k
- The gate is only open for the period of ONE timeslot
- This gate is periodically opened and closed as long as connection held
- During one time slot only one of the inputs can be connected to one output
- At call setup, the connection stores hold the information of which input is to be connected to which output
- The switching speed is 125ms/T
    - T= no. of slots per frame



Time-space-time(TST) TDM switch

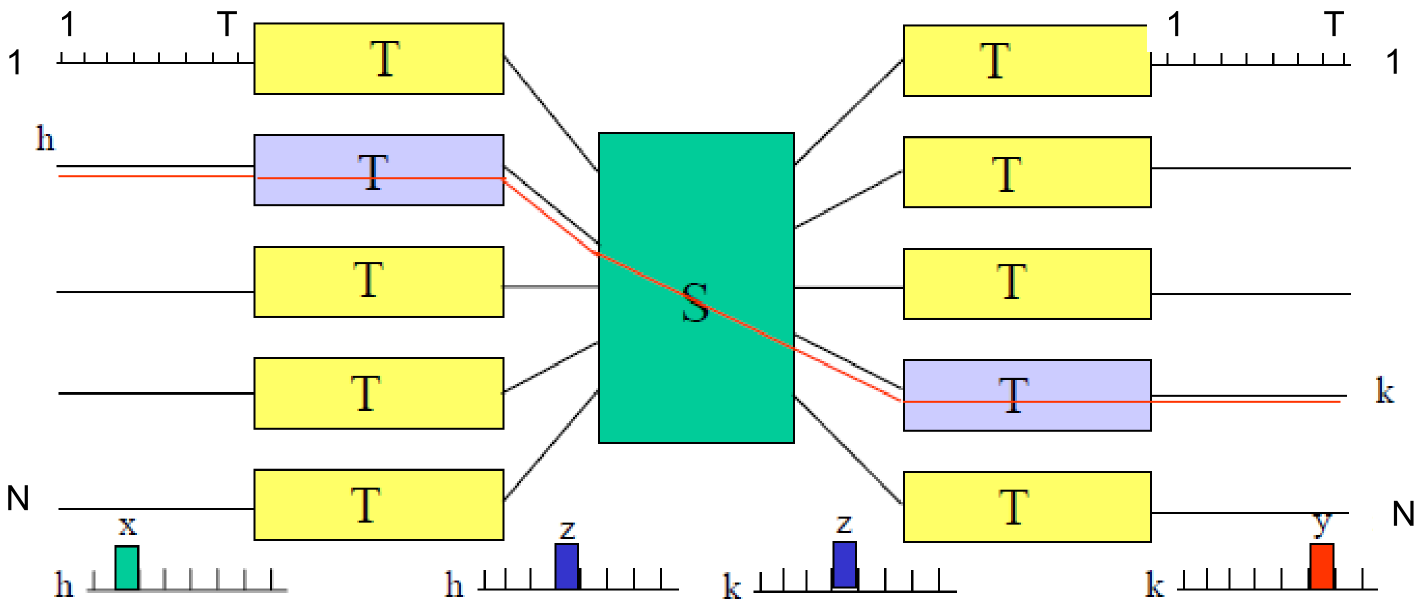


m = 4
n = 4
r = 2



Time - Space - Time (T-S-T) Switches

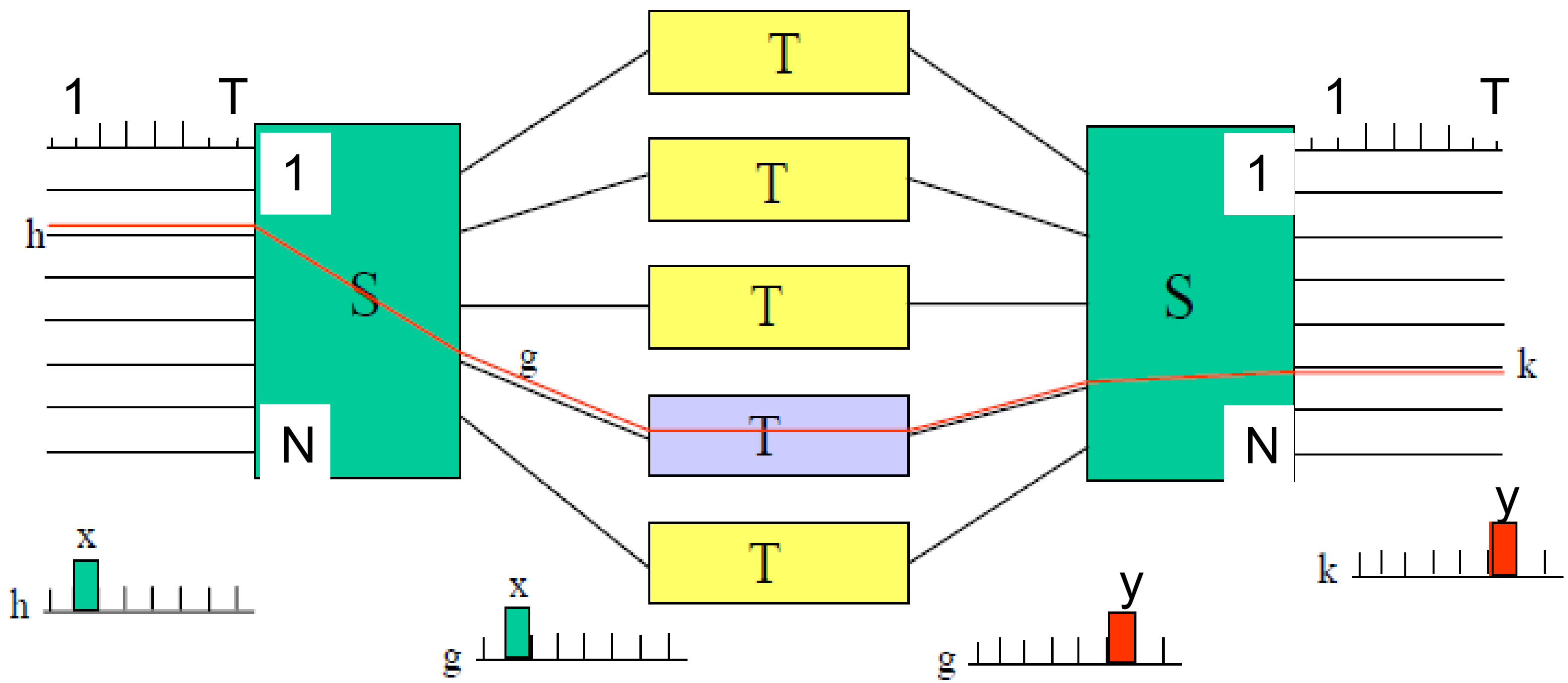
- T-S-T Configuration: Find a time slot, like z which is free at both the output of the incoming highway and input to the outgoing highway.



- Early designs of digital systems used S-T-S networks since storage of speech samples was expensive
- Memory is not so expensive today so most current systems use T-S-T networks

Space – Time – Space (S-T-S) Switches

- To reduce blocking, parallel paths should be created. This is done by three stage switching
- S-T-S Configuration: Find an output highway in first S switch (like g) which has a free time slot x at the input and free time slot y at the output of the T switch

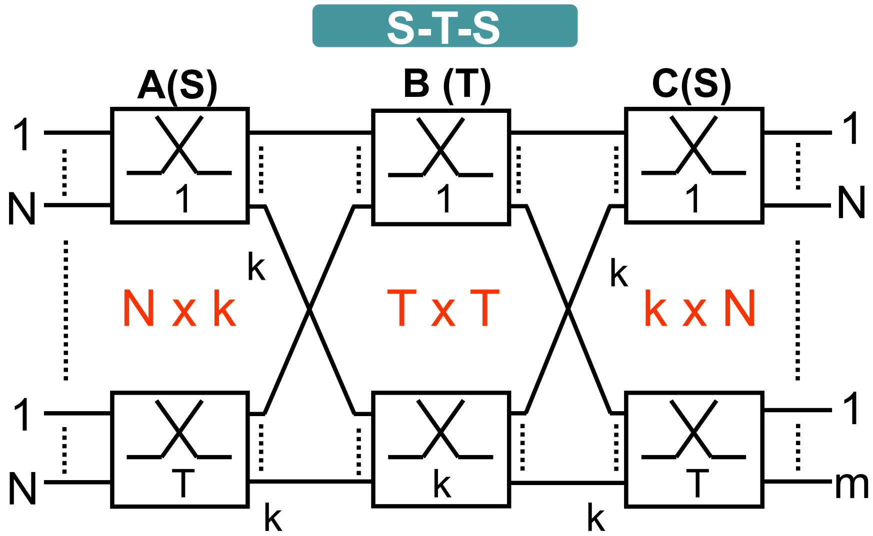


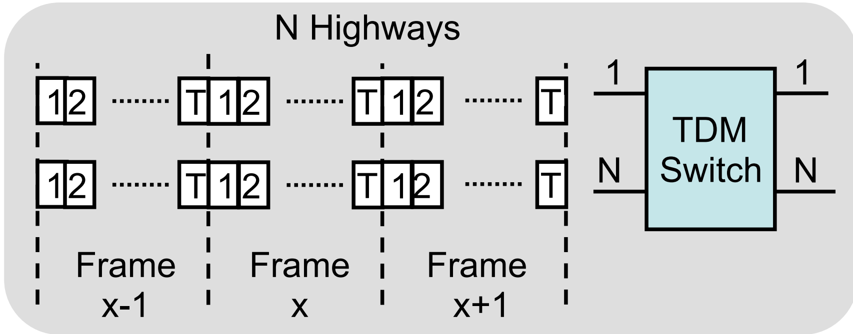
Blocking in TDM Networks

S-T-S Switch

- In a three-stage switch, blocking probability,

  Prob(blocking) = $(1 - (1 - p)^2)^b$ {if k<2n-1}





- In the S-T-S network, each crosspoint of the switch is time shared by T channels
- The A (primary) switch is equivalent to T space-division switches of size N x k and the C switch is equivalent to T space-division switches of size k x N. Each of the k time switches is equivalent to a space division switch of size T x T
- Prob(blocking) = $(1 – (1 – p)^2)^k$

  k = no. of time-switch links

Traffic Modelling

Speech traffic is statistically strongly peaked. The highest peak, when most calls in progress occur at any one time is called the **'busy hour'** (often between 10-11am). The exchange has to be dimensioned to cope with the maximum traffic intensity, which is measured in erlang (E):

$$\textbf{Traffic Intensity (erlang) = A =} \frac{C * h}{T}$$

where C is the average number of call arrivals during time T, and h is the average call **holding time**. T and h should be in the same unit of measurement, e.g. minutes or seconds.

If the measurement is made over just one trunk then A is the **occupancy** (*p*) and indicates the blocking probability or probability that the line will be engaged.

Traffic Modelling

a) 5 trunks carry 54, 65, 74, 76, and 80 minutes of calls respectively during 2 hours. The average call holding time is 1minute.

Total call duration = 349 min.
Traffic intensity (A) = 349x1/(2 x 60) = 2.9 E
Traffic intensity per trunk = 2.9/5 = 0.58 E

b) On average, during the busy hour a company makes 120 outgoing calls of average duration 2 mins. It receives 200 incoming calls of average duration 3 mins.

Outgoing traffic 120 x 2/60 = 4 E
Incoming traffic 200 x 3/60 = 10 E
Total traffic 4 + 10 = 14 E.

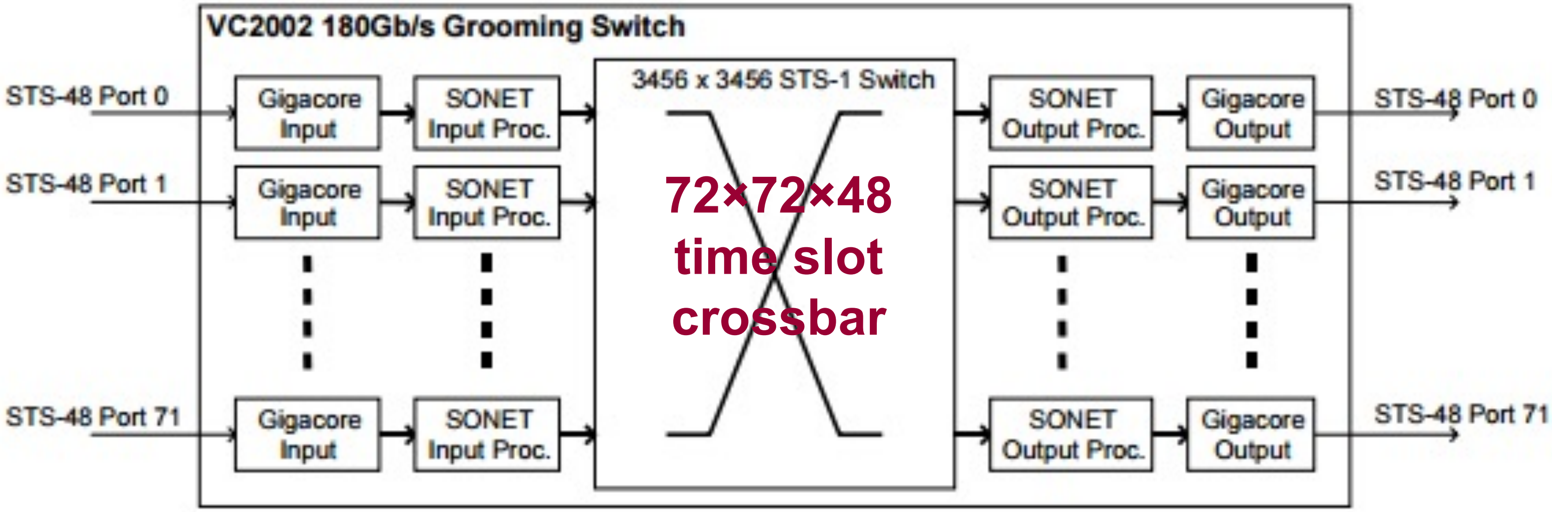
c) During the busy hour a customer with a single telephone line makes 3 calls and receives 3 calls. The average call duration is 2 minutes. What is the prob. that a caller will find the line engaged?

Occupancy = (3 + 3) x 2/60 = 0.1 E = prob. of line engaged.

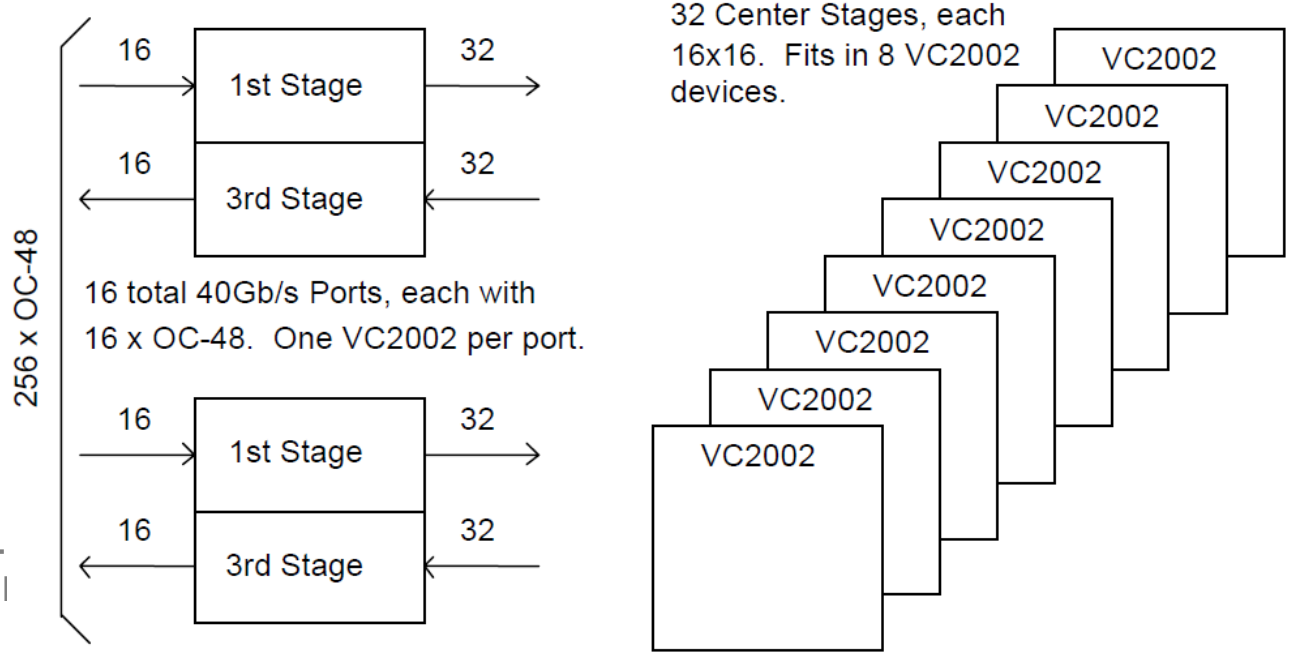
(same formula above, but for one trunk)

The Velio VC2002 Grooming Switch

A single-chip 72×72 STS-48 grooming switch (37.5mm×37.5mm package).



a multistage Digital Cross Connect system at 256 STS-48 is shown in Figure at right.



Reference

- Book: Principles and practices of interconnection networks, by William James Dally & Brian Towles.
- https://www.networkworld.com/article/2226122/cisco-subnet/clos-networks--what-s-old-is-new-again.html