

# Switch Fabrics

**Switching Technology S38.165**  
<http://www.netlab.hut.fi/opetus/s38165>

## Switch fabrics

- Basic concepts
- Time and space switching
- Two stage switches
- Three stage switches
- Cost criteria
- Multi-stage switches and path search

## Switch fabrics (cont.)

- Multi-point switching
- Self-routing networks
- Sorting networks
- Fabric implementation technologies
- Fault tolerance and reliability

## Basic concepts

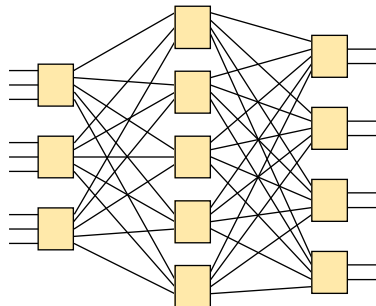
- Accessibility
- Blocking
- Complexity
- Scalability
- Reliability
- Throughput

## Accessibility

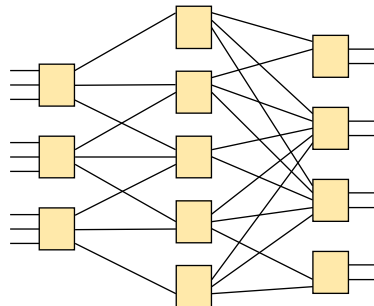
- A network has **full accessibility (=connectivity)** when each inlet can be connected to each outlet (in case there are no other I/O connections in the network)
- A network has a **limited accessibility** when the above given property does not exist
- Interconnection networks applied in today's switch fabrics usually have full accessibility

## Accessibility (cont.)

Example of full accessibility



Example of limited accessibility



## Blocking

- Blocking is defined as failure to satisfy a connection request and it depends strongly on the combinatorial properties of the switching networks

Network class	Network type	Network state
Non-blocking	Strict-sense non-blocking	Without blocking states
	Wide-sense non-blocking	With blocking state
	Rearrangeably non-blocking	
Blocking	Others	

## Blocking (cont.)

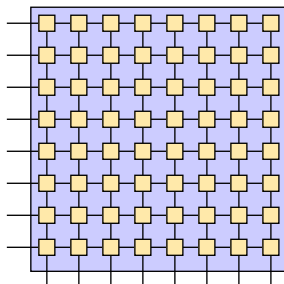
- **Non-blocking** - a path between an arbitrary idle inlet and arbitrary idle outlet can always be established independent of network state at set-up time
- **Blocking** - a path between an arbitrary idle inlet and arbitrary idle outlet cannot be established owing to internal congestion due to the already established connections
- **Strict-sense non-blocking** - a path can always be set up between any idle inlet and any idle outlet without disturbing paths already set up
- **Wide-sense non-blocking** - a path can be set up between any idle inlet and any idle outlet without disturbing existing connections, provided that certain rules are followed. These rules prevent network from entering a state for which new connections cannot be made
- **Rearrangeably non-blocking** - when establishing a path between an idle inlet and an idle outlet, paths of existing connections may have to be changed (rearranged) to set up that connection

## Complexity

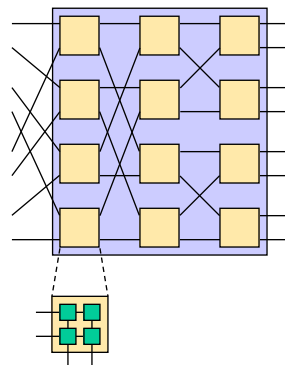
- Complexity of an interconnection network is expressed by **cost index**
- Traditional definition of cost index gives the **number of cross-points in a network**
  - used to be a reasonable measure of space division switching systems
- Nowadays cost index alone does not characterize cost of an interconnection network for broadband applications
  - VLSIs and their integration degree has changed the way how cost of a switch fabric is formed (number of ICs, power consumption)
  - management and control of a switching system has a significant contribution to cost

## Complexity (cont.)

Cost index of an 8x8 crossbar is 64 (cross-points)



Cost index of an 8x8 banyan is  $12 \times 4 = 48$  (cross-points)



## Scalability

- Due to constant increase of transport links and data rates on links, scalability of a switching system has become a key parameter in choosing a switch fabric architecture
- Scalability describes ability of a system to evolve with increasing requirements
- Issues that are usually matter of scalability
  - number of switching nodes
  - number of interconnection links between nodes
  - bandwidth of interconnection links and inlets/outlets
  - throughput of switch fabric
  - buffering requirements
  - number of inlets/outlets supported by switch fabric

## Scalability (cont.)

### Example of scalability

- a switching equipment has room for 20 line-cards and the original design supports 10 Mbit/s interfaces (one per line card)
- throughput of switch fabrics is scalable from 500 Mbit/s to 2 Gbit/s
- original switch fabric can support new line cards that implement two 10 Mbit/s interfaces each
- when line interfaces are replaced with 100 Mbit/s rates(one per line-card), the switch fabric has to be updated (scaled up) to 2 Gbit/s speed
- buffering memories need to be replaced by faster (and possible larger) ones
- larger number of line cards implies at least new physical design
- increase of line rates beyond 100 Mbit/s means redesign of switch fabric

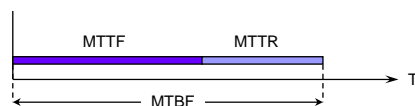
## Reliability

- Reliability and fault tolerance are system measures that have an impact on all functions of a switching system
- Reliability defines probability that a system does not fail within a given time interval provided that it functions correctly at the start of the interval
- Availability defines probability that a system will function at a given time instant
- Fault tolerance is the capability of a system to continue its intended function in spite of having a fault(s)
- Reliability measures:
  - MTTF (Mean Time To Failure)
  - MTTR (Mean Time To Repair)
  - MTBF (Mean Time Between Failures)

## Reliability (cont.)

Relation of reliability  $R(t)$  to availability  $F(t)$  is given by  
 $F(t) = 1 - R(t)$

Relation of MTTF, MTTR and MTBF



## Throughput

- Throughput gives forwarding/switching speed/efficiency of a switch fabric
- It is measured in bits/s, octets/s, cells/s, packet/s, etc.
- Quite often throughput is given in the range (0 ... 1.0], i.e. the obtained forwarding speed is normalized to the theoretical maximum throughput

## Switch fabrics

- Basic concepts
- **Time and space switching**
- Two stage switches
- Three stage switches
- Cost criteria
- Multi-stage switches and path search

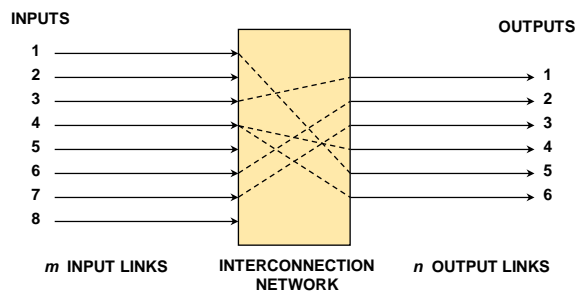


## Switching mechanisms

- A switched connection requires a mechanism that attaches the right information streams to each other
- Switching takes place in the switching fabric, the structure of which depends on network's mode of operation, available technology and required capacity
- Communicating terminals may use different physical links and different time-slots, so there is an obvious need to switch both in time and in space domain
- **Time and space** switching are basic functions of a switch fabric

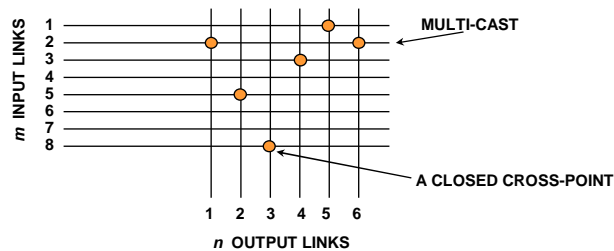
## Space division switching

- A space switch directs traffic from input links to output links
- An input may set up one connection (1, 3, 6 and 7), multiple connections (4) or no connection (2, 5 and 8)



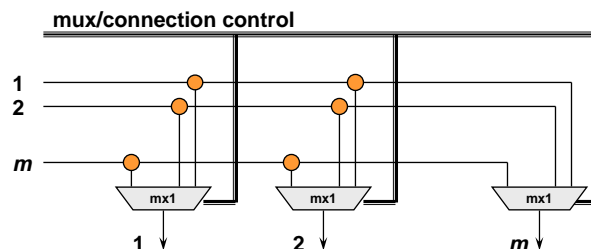
## Crossbar switch matrix

- Crossbar matrix introduces the basic structure of a space switch
- Information flows are controlled (switched) by opening and closing cross-points
- $m$  inputs and  $n$  outputs  $\Rightarrow mn$  cross-points (connection points)
- Only one input can be connected to an output at a time, but an input can be connected to multiple outputs (multi-cast) at a time



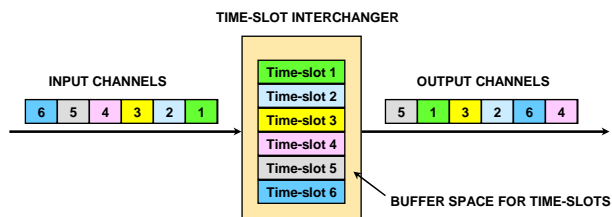
## An example space switch

- $m \times 1$  -multiplexer used to implement a space switch
- Every input is fed to every output mux and mux control signals are used to select which input signal is connected through each mux

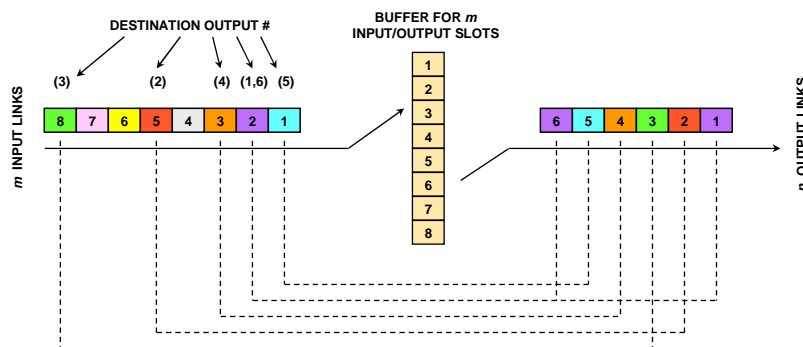


## Time division multiplexing

- Time-slot interchanger is a device, which buffers  $m$  incoming time-slots, e.g. 30 time-slots of an E1 frame, arranges new transmit order and transmits  $n$  time-slots
- Time-slots are stored in buffer memory usually in the order they arrive or in the order they leave the switch - additional control logic is needed to decide respective output order or the memory slot where an input slot is stored

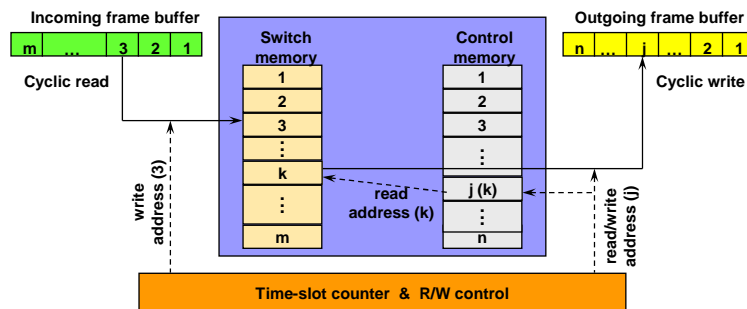


## Time-slot interchange



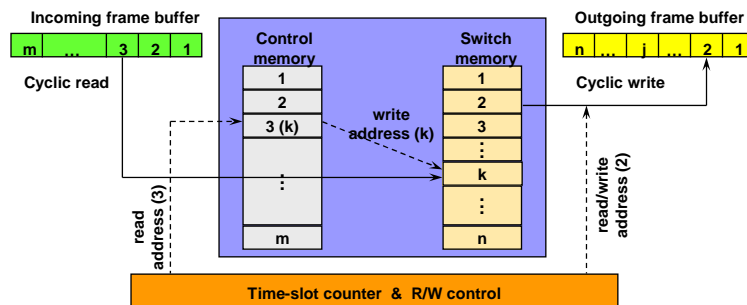
## Time switch implementation example 1

- Incoming time-slots are written cyclically into switch memory
- Output logic reads cyclically control memory, which contains a pointer for each output time-slot
- Pointer indicates which input time-slot to insert into each output time-slot



## Time switch implementation example 2

- Incoming time-slots are written into switch memory by using write-addresses read from control memory
- A write address points to an output slot to which the input slot is addressed
- Output time-slots are read cyclically from switch memory

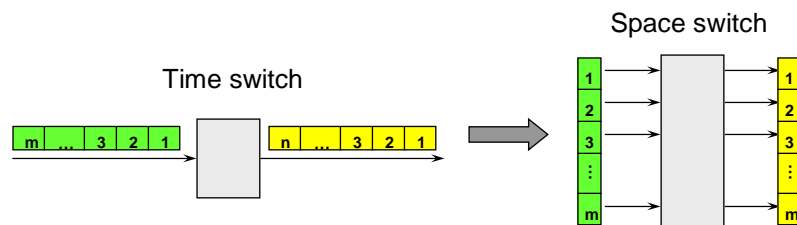


## Properties of time switches

- Input and output frame buffers are read and written at wire-speed, i.e.  $m$  R/Ws for input and  $n$  R/Ws for output
- Interchange buffer (switch memory) serves all inputs and outputs and thus it is read and written at the aggregate speed of all inputs and outputs  
=> speed of an interchange buffer is a critical parameter in time switches and limits performance of a switch
- Utilizing parallel to serial conversion memory speed requirement can be cut
- Speed requirement of control memory is half of that of switch memory (in fact a little more than that to allow new control data to be updated)

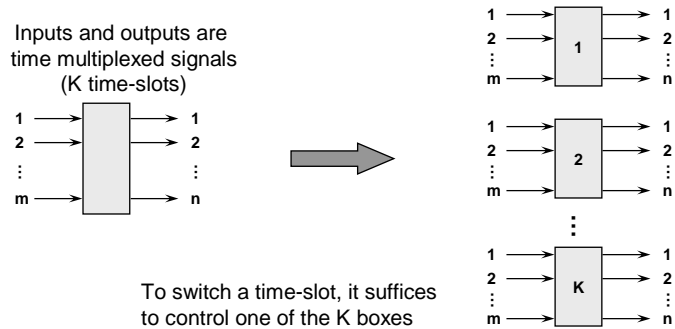
## Time-Space analogy

- A time switch can be logically converted into a space switch by setting time-slot buffers into vertical position => time-slots can be considered to correspond to input/output links of a space switch
- But is this logical conversion fair ?

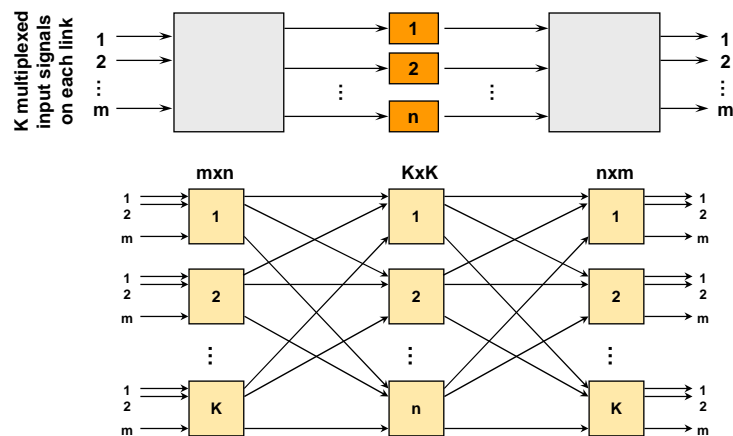


## Space-Space analogy

- A space switch carrying time multiplexed input and output signals can be logically converted into a pure space switch (without cyclic control) by distributing each time-slot into its own space switch



## An example conversion



## Properties of space and time switches

### Space switches

- number of cross-points (e.g. AND-gates)
  - $m$  input  $\times$   $n$  output =  $mn$
  - when  $m=n \Rightarrow n^2$
- output bit rate determines the speed requirement for the switch components
- both input and output lines deploy "bus" structure  $\Rightarrow$  fault location difficult

### Time switches

- size of switch memory (SM) and control memory (CM) grows linearly as long as memory speed is sufficient, i.e. SM + CM + input buffering + output buffering =  $2 \times 2 \times$  number of time-slots
- a simple and cost effective structure when memory speed is sufficient
- speed of available memory determines the maximum switching capacity

## Switch fabrics

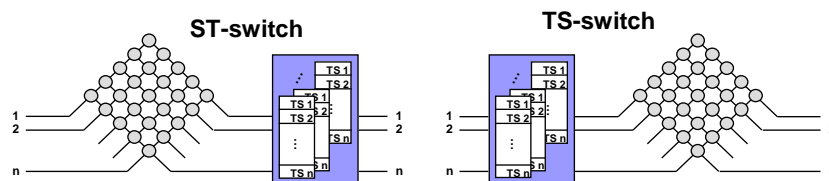
- Basic concepts
- Time and space switching
- **Two stage switches**
- Three stage switches
- Cost criteria
- Multi-stage switches and path search

## A switch fabric as a combination of space and time switches

- Two stage switches
  - Time-Time (TT) switch
  - Time-Space (TS) switch
  - Space-Time (ST) switch
  - Space-Space (SS) switch
- TT-switch gives no advantage compared to a single stage T-switch
- SS-switch increases blocking probability

## A switch fabric as a combination of space and time switches (cont.)

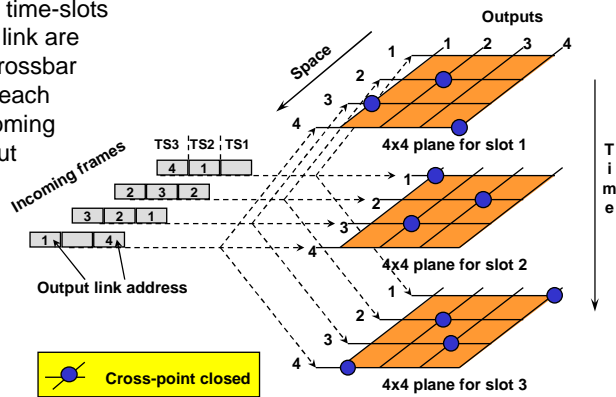
- ST-switch gives high blocking probability (S-switch can develop blocking on an arbitrary bus, e.g. slots from two different buses attempting to flow to a common output)
- TS-switch has low blocking probability, because T-switch allows rearrangement of time-slots so that S-switching can be done blocking free





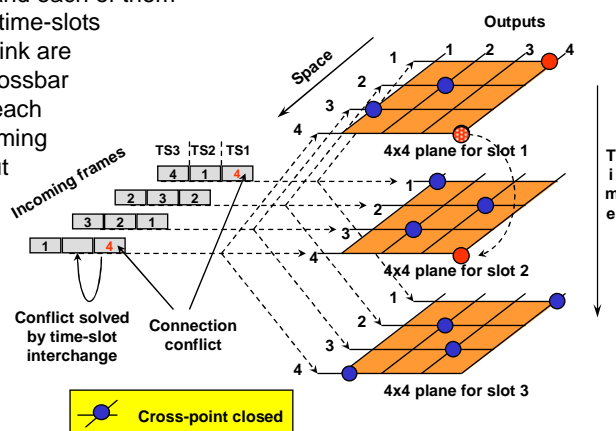
## Time multiplexed space (TMS) switch

- Space divided inputs and each of them carry a frame of three time-slots
- Input frames on each link are synchronized to the crossbar
- A switching plane for each time-slot to direct incoming slots to destined output links of the corresponding time-slot



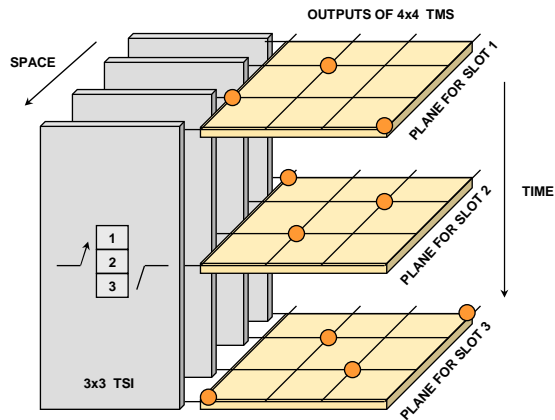
## Connection conflicts in a TMS switch

- Space divided inputs and each of them carry a frame of three time-slots
- Input frames on each link are synchronized to the crossbar
- A switching plane for each time-slot to direct incoming slots to destined output links of the corresponding time-slot

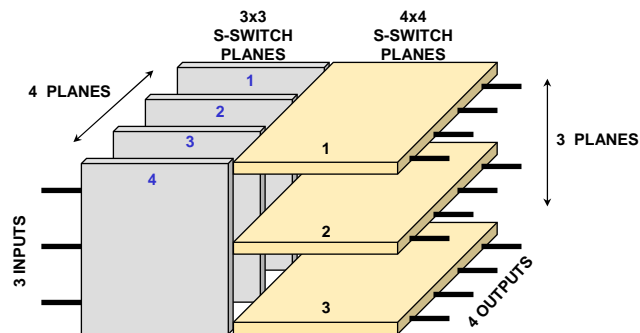


# TS switch interconnecting TDM links

- Time division switching applied prior to space switching
- Incoming time-slots can always be rearranged such that output requests become conflict free for each slot of a frame, provided that the number of requests for each output is no more than the number of slots in a frame



# SS equivalent of a TS-switch

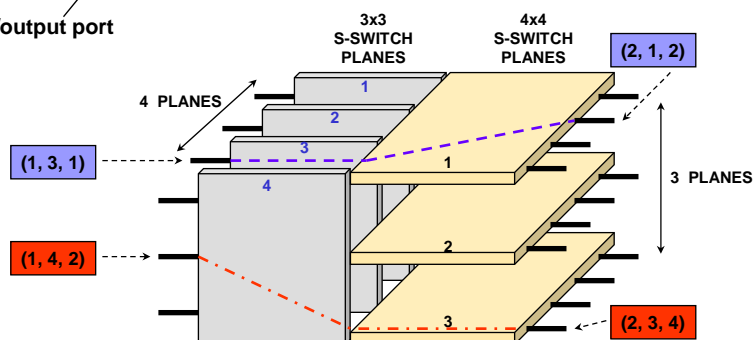


## Connections through SS-switch

Coordinate (X, Y, Z)  
stage  
plane  
input/output port

Example connections:

- (1, 3, 1) => (2, 1, 2)
- (1, 4, 2) => (2, 3, 4)



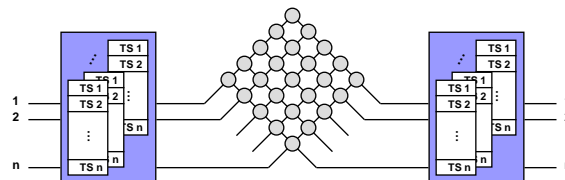
## Switch fabrics

- Basic concepts
- Time and space switching
- Two stage switches
- **Three stage switches**
- Cost criteria
- Multi-stage switches and path search

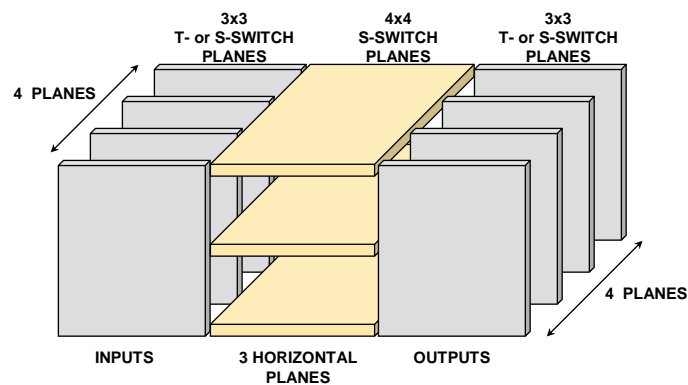
## Three stage switches

- Basic TS-switch sufficient for switching time-slots onto addressed outputs, but slots can appear in any order in the output frame
- If a specific input slot is to carry data of a specific output slot then a time-slot interchanger is needed at each output
  - => any time-slot on any input can be connected to any time-slot on any output
  - => blocking probability minimized
- Such a 3-stage configuration is named TST-switching (equivalent to 3-stage SSS-switching)

TST-switch:



## SSS presentation of TST-switch

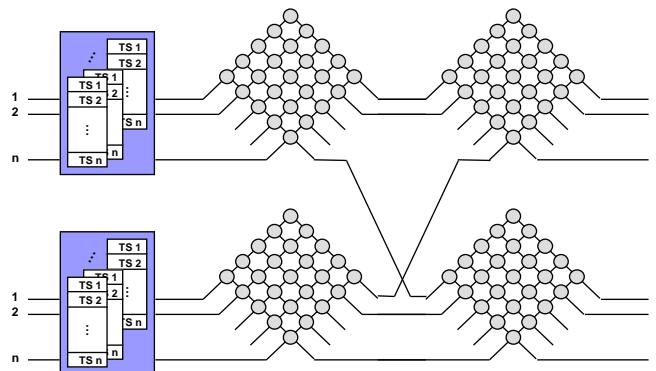


## Three stage switch combinations

- Possible three stage switch combinations:
  - Time-Time-Time (TTT) ( not significant, no connection from PCM to PCM)
  - Time-Time-Space (TTS) (=TS)
  - Time-Space-Time (TST)
  - Time-Space-Space (TSS)
  - Space-Time-Time (STT) (=ST)
  - Space-Time-Space (STS)
  - Space-Space-Time (SST) (=ST)
  - Space-Space-Space (SSS) (not significant, high probability of blocking)
- Three interesting combinations TST, TSS and STS

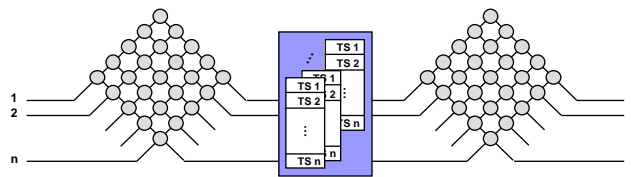
## Time-Space-Space switch

- Time-Space-Space switch can be applied to increase switching capacity



## Space-Time-Space switch

- Space-Time-Space switch has a high blocking probability (like ST-switch) - not a desired feature in public networks



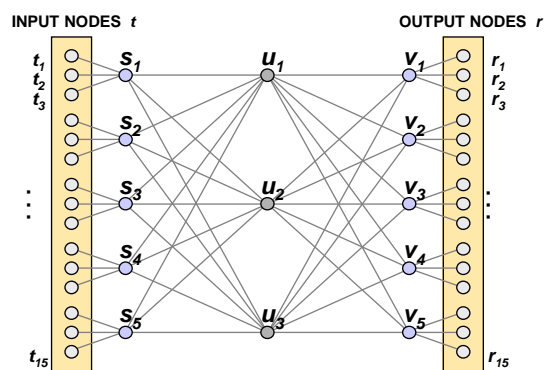
## Graph presentation of space switch

- A space division switch can be presented by a graph  $G = (V, E)$ 
  - $V$  is the set of switching nodes
  - $E$  is the set of edges in the graph
- An edge  $e \in E$  is an ordered pair  $(u, v) \in V$ 
  - more than one edge can exist between  $u$  and  $v$
  - edges can be considered to be bi-directional
- $V$  includes two special sets ( $T$  and  $R$ ) of nodes not considered part of switching network
  - $T$  is a set of transmitting nodes having only outgoing edges (input nodes to switch)
  - $R$  is a set of receiving node having only incoming edges (output nodes from switch)

## Graph presentation of space switch (cont.)

- A connection requirement is specified for each  $t \in T$  by subset  $R_t \in R$  to which  $t$  must be connected
  - subsets  $R_t$  are disjoint for different  $t$
  - in case of multi-cast  $R_t$  contains more than one element for each  $t$
- A path is a sequence of edges  $(t,a), (a,b), (b,c), \dots, (f,g), (g,r) \in E$ ,  $t \in T, r \in R$  and  $a,b,c,\dots,f,g$  are distinct elements of  $V - (T+R)$
- Paths originating from different  $t$  may not use the same edge
- Paths originating from the same  $t$  may use the same edges

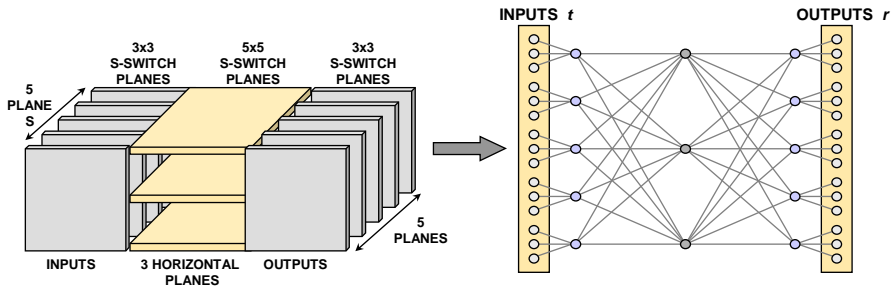
## Graph presentation example



$$V = (t_1, t_2, \dots, t_{15}, s_1, s_2, \dots, s_5, u_1, u_2, u_3, v_1, v_2, \dots, v_5, r_1, r_2, \dots, r_{15})$$

$$E = \{(t_1, s_1), \dots, (t_{15}, s_5), (s_1, u_1), (s_1, u_2), \dots, (s_5, u_3), (u_1, v_1), (u_1, v_2), \dots, (u_3, v_5), (v_1, r_1), (v_1, r_2), \dots, (v_5, r_{15})\}$$

# SSS-switch and its graph presentation

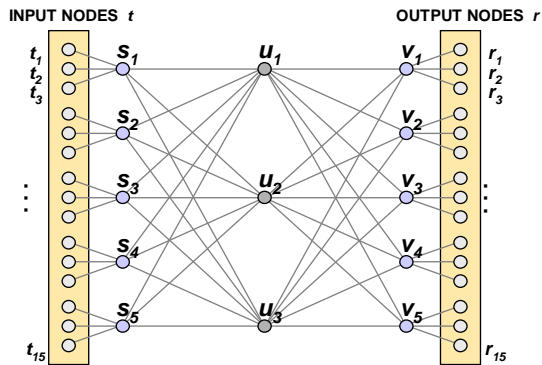


# Graph presentation of connections

**Establish connections:**

$$\text{Path 1} = \{(t_1, s_4), (s_4, u_1), (u_1, v_2), (v_2, r_5)\}$$

$$\text{Path 2} = \{(t_4, s_2), (s_2, u_2), (u_2, v_1), (v_1, r_2), (u_2, v_4), (v_4, r_{11})\}$$





## Graph presentation of connections (cont.)

