RW354
Principles of Computer Networking

A.E. Krzesinski and B.A. Bagula
Department of Computer Science
University of Stellenbosch

Last updated: 17 August 2004

The material presented in these slides is used with permission from

- Larry L. Peterson and Bruce S. Davie. Computer Networks: A Systems Approach (Second Edition). Morgan Kaufmann Publishers. ISBN 1-55860-577-0.
- William Stallings. Data and Computer Communications (Sixth Edition). Prentice-Hall Inc. ISBN 0-13-571274-2.
- Andrew S. Tannenbaum. Computer Networks (Fourth Edition). Prentice Hall Inc. ISBN 0-13-349945-6.

Packet Switching

Directly connected networks have two limitations

- *a limited number of attached hosts*
- *a limited geographic area.*

Hosts that are not directly connected must be able to communicate if networks are to be global.

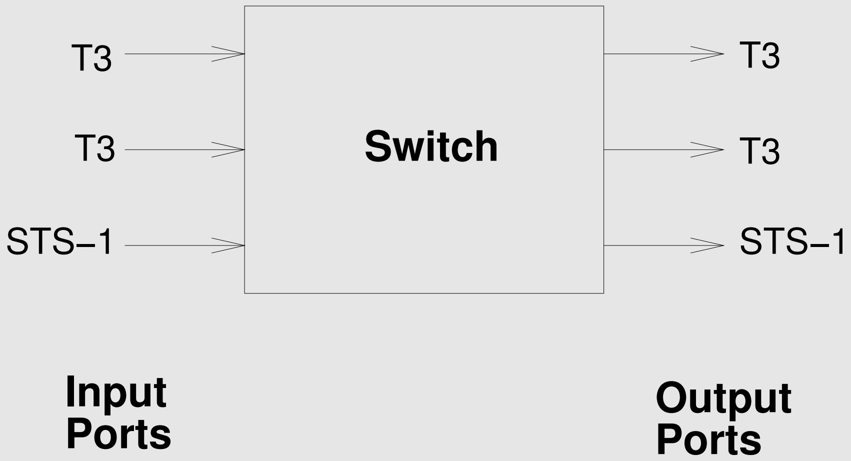
Packet switches enable packets to travel between hosts when the hosts are not directly connected.

Packet switching issues include

- *routing & forwarding*
- *contention - packets are buffered*
- *congestion - packets are dropped.*

Scalable Networks

A switch forwards packets from the input port to the output port. The output port is selected based on the destination address in the packet header.



- *can build networks that cover a large geographic area & support a large numbers of hosts*
- *can add new hosts without affecting the performance of the existing hosts.*

Circuit Switched Networks

The telephone network is a circuit switched network & is an example of a connection oriented network.

Communication in such a network proceeds in 3 steps

- *circuit connection: an end-to-end connection is set up from the source switch via intermediate switches to the destination switch before any data are sent*
- *data transfer*
- *circuit termination: the end-to-end connection is released.*

Circuit Switched Networks

Users (subscribers) are connected to a switch by a local loop: usually copper wire. Some of the switches are connected by fibre optic links.

The signalling system is used to instruct the switches to setup, monitor & terminate the communication path.

The SS7 (signalling system number 7) signalling network is a connectionless network entirely separate from the telephone network.

Nyquist's Sampling Theorem

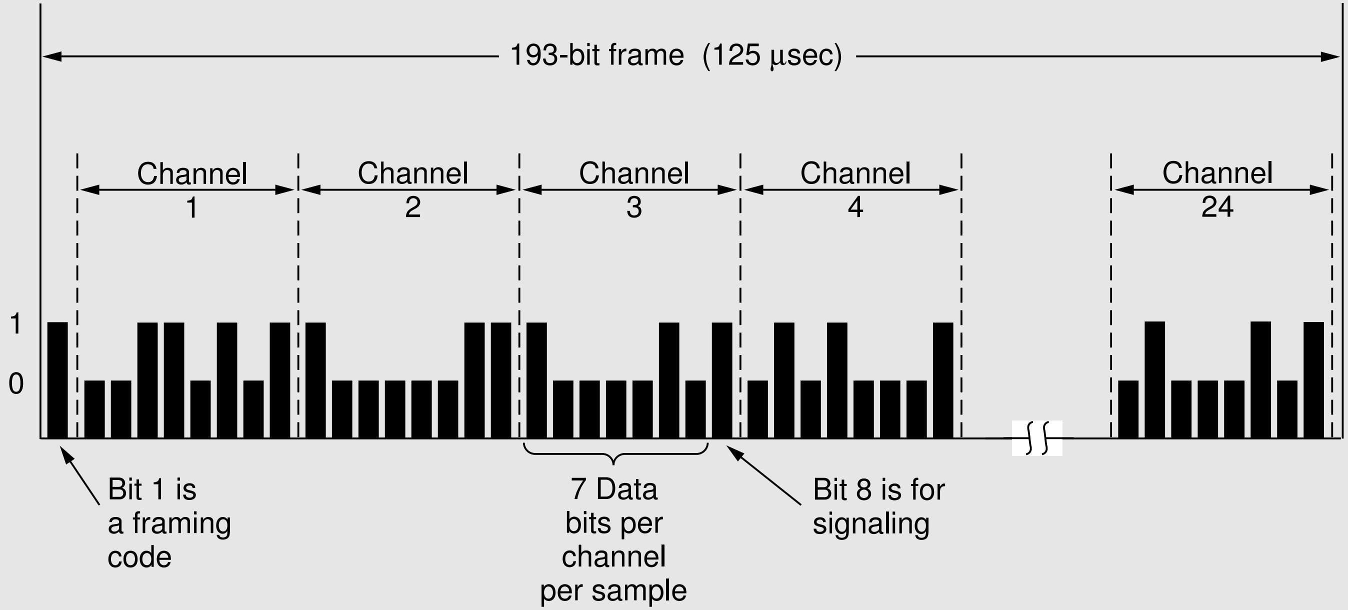
If a signal $f(t)$ is sampled at regular time intervals & at a rate higher than twice the highest significant signal frequency, then the samples contain all the information of the original signal.

The function $f(t)$ may be reconstructed from these samples.

Digital Transmission

- *the calling party uses the local loop to transmit voice signals in the frequency range 400 to 3400 Herz to the source switch*
- *Pulse Amplitude Modulation: the source switch samples the voice signal 8,000 times per second*
- *Pulse Code Modulation: the PAM samples are quantized: the amplitude of each PAM sample is approximated by an 8-bit digit*
- *the PCM samples are transmitted from the source switch to the destination switch (via intermediate switches) using the DS-1 synchronous time division multiplexing scheme (USA)*
- *the destination switch converts the PCM samples to analog signals and transmits these signals on the local loop to the called party.*

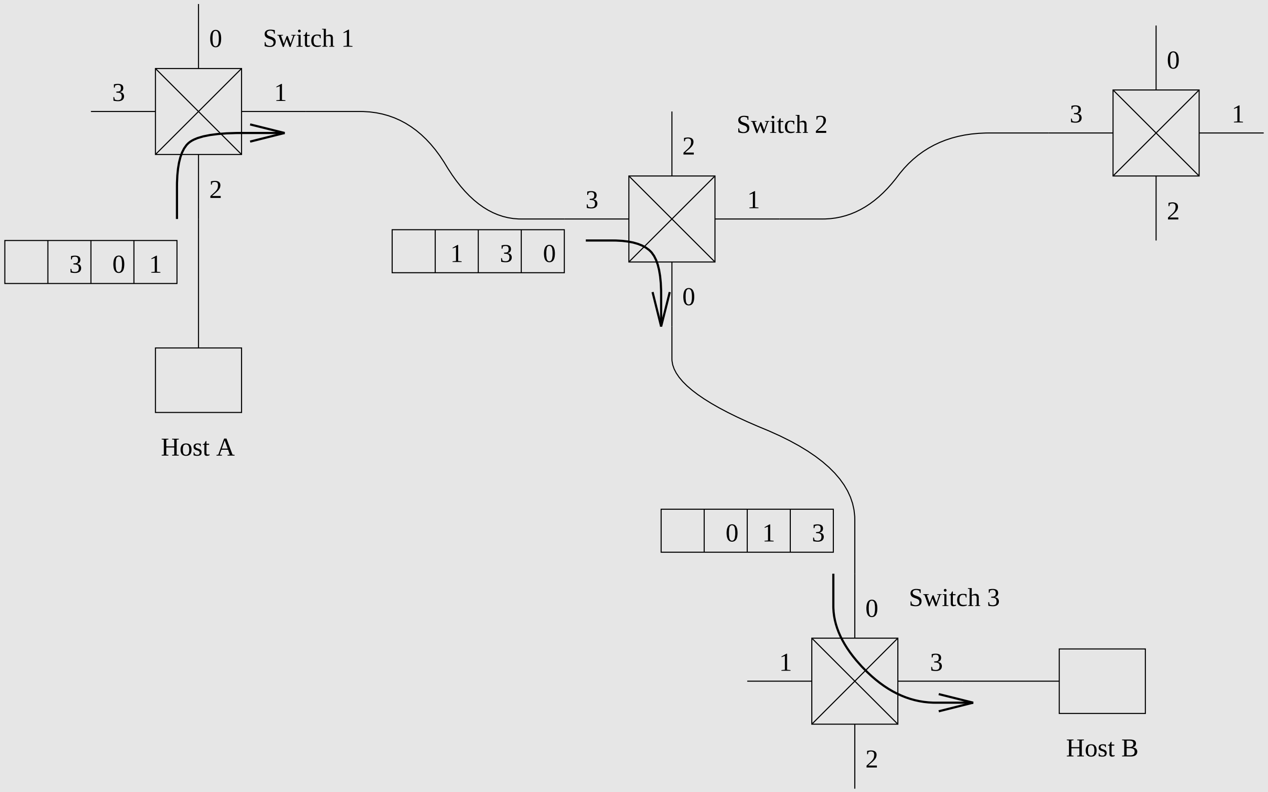
DS-1 Transmission Format



- *bit 1 is a framing bit used for synchronization*
- *voice channels*
  - *8-bit PCM used on 5 of 6 frames*
  - *7-bit PCM used on every 6th frame where bit 8 of each channel is a signalling bit*

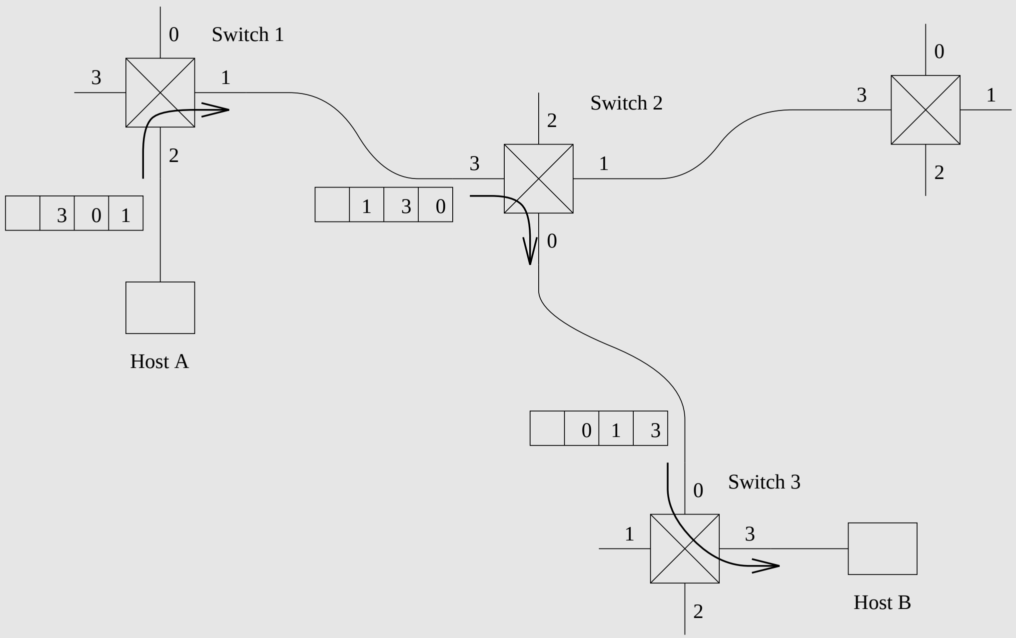
Source Routing

The address contains a sequence of ports on the path from the source to the destination.



The port list in the header is rotated so that the next switch in the path is at the head of the list.

Source Routing

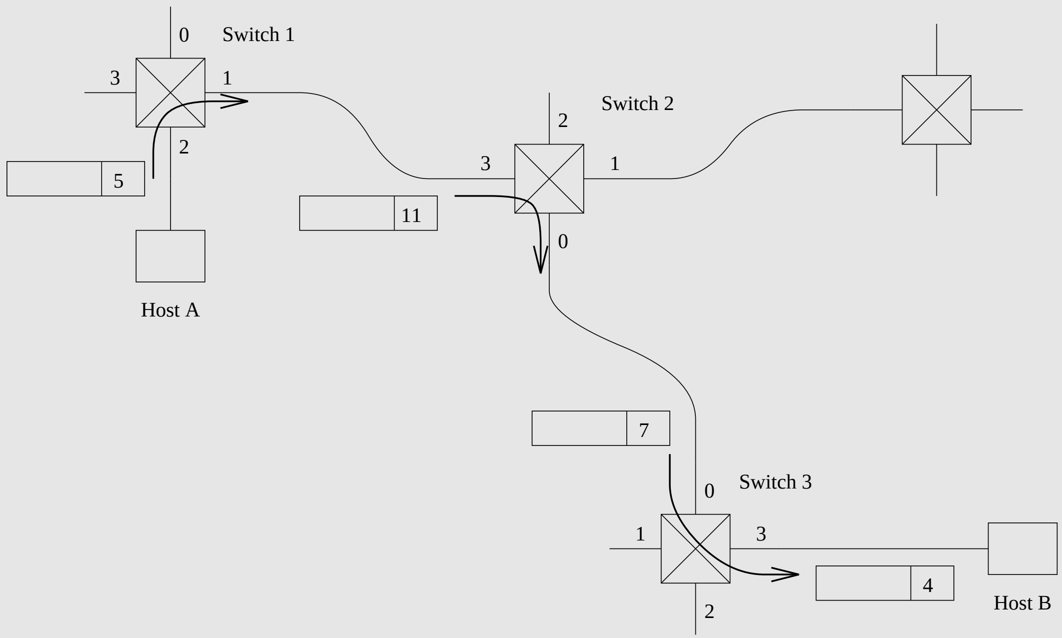


Some issues

- *scaling: can the source determine the complete path?*
- *variable size headers*
- *rotated port list implies that the destination knows the reverse path to the source.*

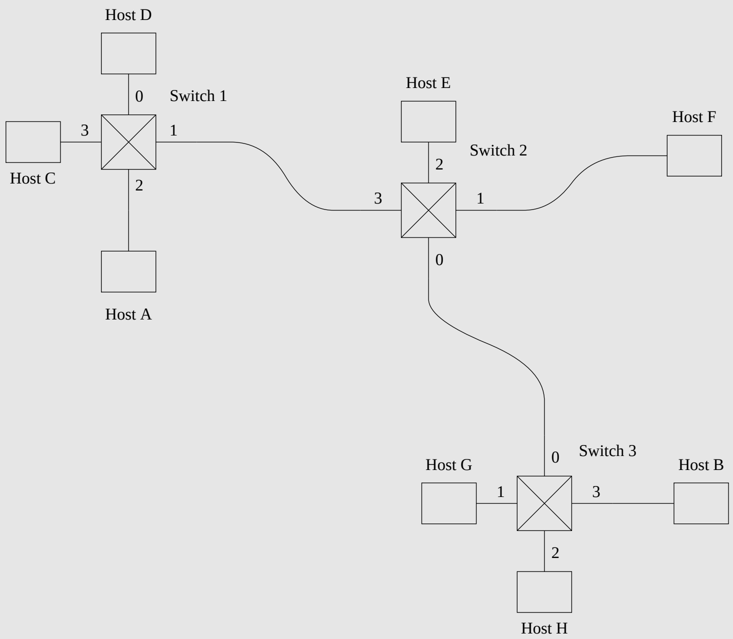
Virtual Circuit Switching

- *an explicit connection setup (& tear-down) phase*
- *subsequent packets follow the same circuit*
- *analogy: phone call*
- *sometimes called the connection-oriented model*
- *each switch maintains a VC table: the VCI's have link-local scope.*



Datagrams

- *no connection setup phase*
- *each packet is forwarded independently*
- *analogy: postal system*
- *sometimes called the connectionless model*
- *each switch maintains a forwarding (routing) table.*



Virtual Circuit versus Datagram

The virtual circuit model

- *There is a full RTT delay waiting for the connection setup before sending the first data packet.*
- *While the connection request contains the full address for the destination, each data packet contains only a small identifier, making the per-packet header overhead small.*
- *If a switch or a link in a connection fails, the connection is broken & a new connection needs to be established.*
- *The connection setup provides an opportunity to reserve resources - each VC can be assigned a Quality of Service (Qos).*

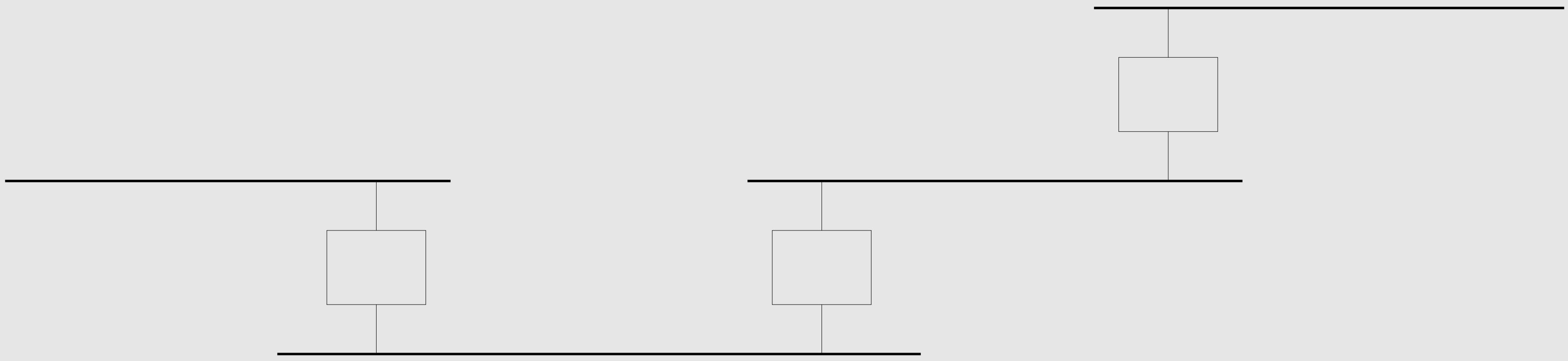
Virtual Circuit versus Datagram

The datagram model

- *There is no RTT delay waiting for the connection setup; a host can send data as soon as it is ready.*
- *The source host has no way of knowing if the network is capable of delivering a packet or if the destination host is up.*
- *Since packets are treated independently, it is possible to route around link & node failures.*
- *Since every packet must carry the full address of the destination, the overhead per packet is higher than for the connection-oriented model.*

Bridges: Overview

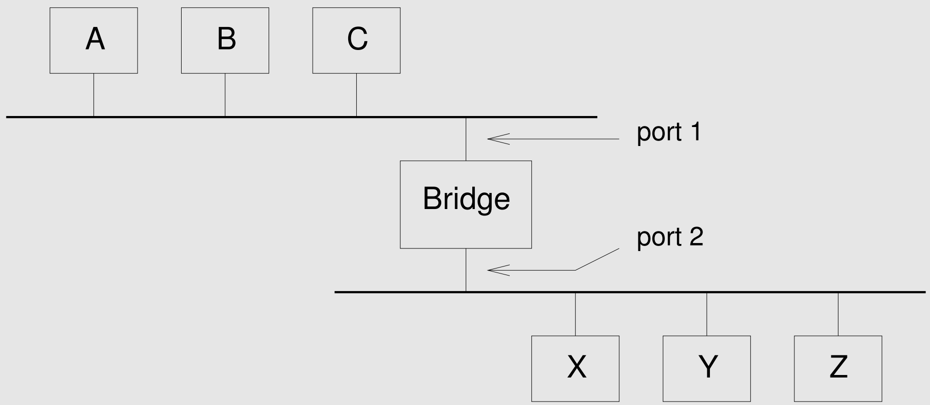
- *LANs have physical limitations (e.g. 1500m Ethernet).*
- *Connect two or more LANs with a bridge*
  - *bridges operate at layer 2*
  - *the bridge is in promiscuous mode*
  - *the bridge implements an accept & forward strategy.*



- *The collection of LANs connected by bridges called an extended LAN.*

Bridges: Learning Bridges

- *Bridges do not forward when unnecessary.*



- *Each bridge maintains a forwarding table.*

Host	*Port*
A	*1*
B	*1*
C	*1*
X	*2*
Y	*2*
Z	*2*

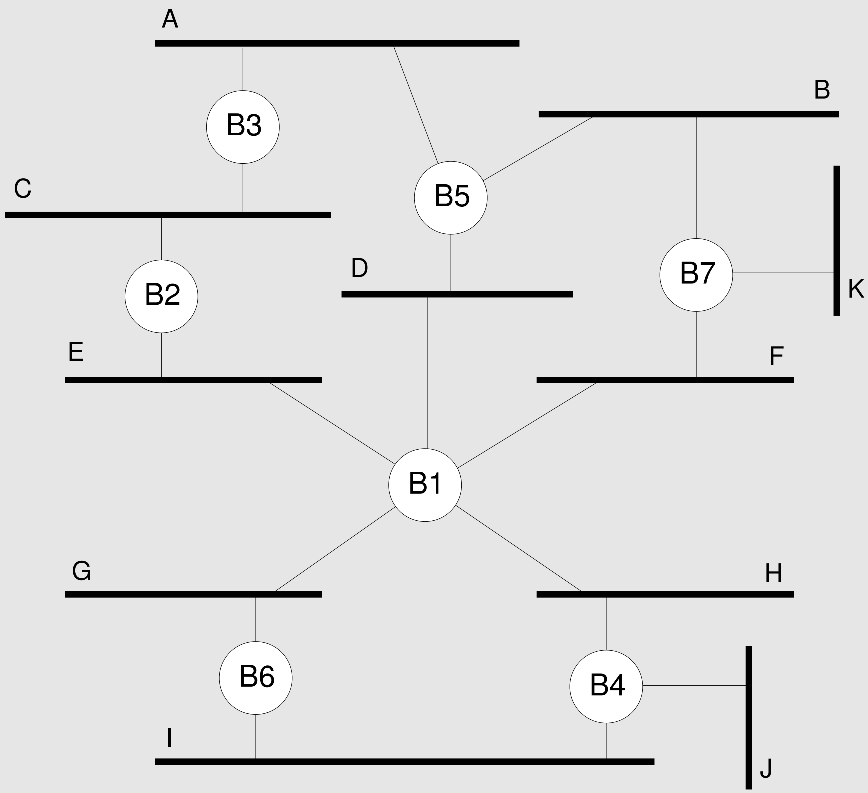
Bridges: Learning Bridges

The goal of a bridge is to transparently extend a LAN across multiple networks

- *at boot time the forwarding table is empty*
- *bridges learn the table entries based on the source addresses of forwarded packets*
- *the forwarding table is an optimization*
- *if the bridge receives a frame for a host whose address is not in the table, the frame is forwarded to all the output ports*
- *broadcast frames are always forwarded*
- *table entries have finite lifetimes: timeout.*

Bridges: Spanning Tree Algorithm

Extended LANs sometimes have loops.



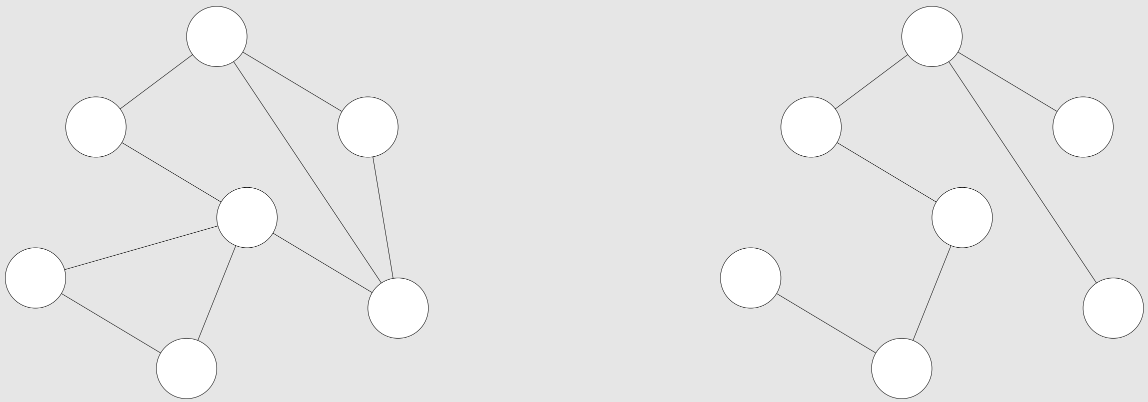
B1-B6-B4-B1.

We need to find an acyclic path (spanning tree) connecting all the LANs.

Bridges: Spanning Tree Algorithm

The bridges run a distributed spanning tree algorithm

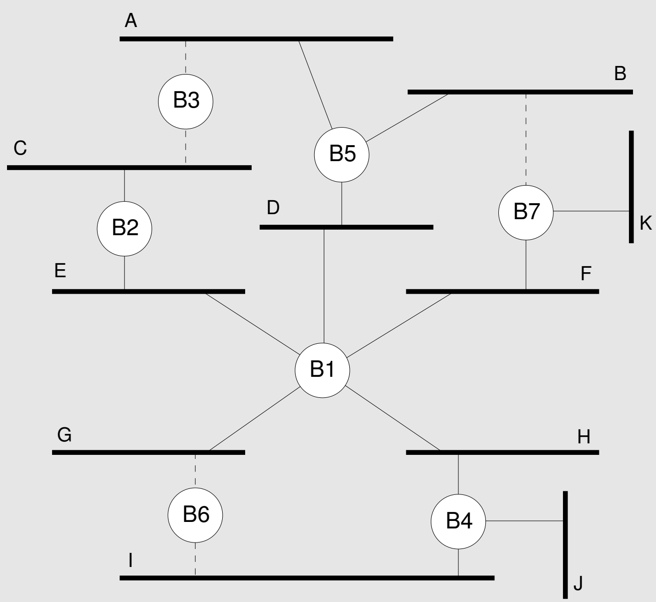
- *select which bridges actively forward frames*
- *developed by Radia Perlman at DEC*
- *now the IEEE 802.1 specification.*



A cyclic graph & a (one of many) corresponding spanning trees.

Bridges: Spanning Tree Algorithm Overview

- *Each bridge has a unique id: B1, B2, B3, . . .*
- *Select the bridge with smallest id as the root.*
- *Select the bridge on each LAN that is closest to the root as that LAN's designated bridge: use the id to break ties*
- *Example: B5 is the designated bridge for LAN's A, B & D: B5 forwards frames to LAN's A, B & D.*



Bridges: Spanning Tree Algorithm Detail

- *Bridges periodically exchange configuration messages*
    - *the id of the bridge sending the message*
    - *the id of what the sending bridge believes to be the root bridge*
    - *the distance (hops) from the sending bridge to the root bridge.*
- *Each bridge records the current best configuration message for each port.*

Bridges: Spanning Tree Algorithm Detail

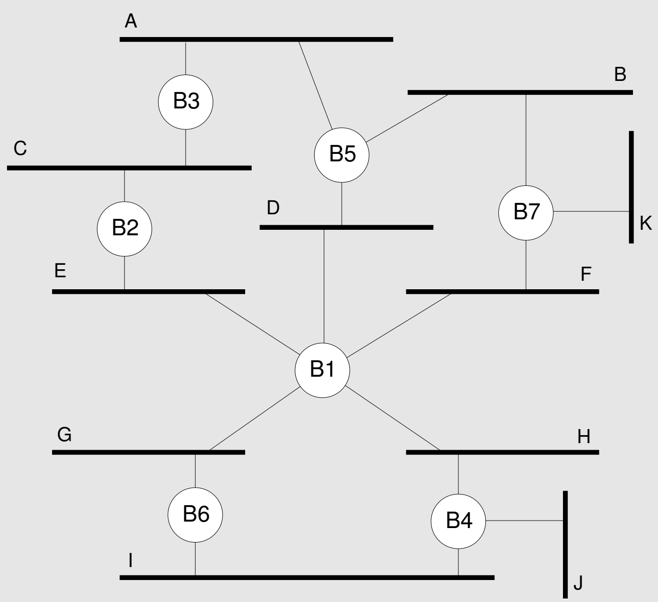
- *Initially each bridge believes that it is the root.*
- *When a bridge learns that it is not the root, it stops generating configuration messages*
    - *eventually only the root generates configuration messages.*
- *When a bridge learns that it is not a designated bridge, it stops forwarding configuration messages*
    - *eventually only designated bridges forward configuration messages.*
- *The root bridge continues to send configuration messages periodically.*
- *If a bridge does not receive a configuration message after a period of time, it starts generating configuration messages claiming to be the root.*

Bridges: Spanning Tree Algorithm Example

B3 sends (B3,0,B3)	*B3 accepts B3 as root*
B3 receives (B2,0,B2)	*B3 accepts B2 as root*
B3 sends (B2,1,B3) to B5	*B3 stops generating messages*
B2 receives (B1,0,B1)	*B2 accepts B1 as root*
B2 sends (B1,1,B2) to B3	*B2 stops generating messages*
B5 receives (B1,0,B1)	*B5 accepts B1 as root*
B5 receives (B2,1,B3)	*B5 retains B1 as root*
B5 sends (B1,1,B5) to B3	*B5 stops generating messages*
B3 receives (B1,1,B2)	*B3 accepts B1 as root*
B3 receives (B1,1,B5)	*B3 retains B1 as root*
	B2 & B5 are closer to B1 than B3: B3 stops forwarding messages on both its interfaces

Bridges: Broadcast & Multicast

- *The current practice is to forward all broadcast & multicast frames.*
- *It is possible for the bridge to learn when no group members are downstream*
  - *each member of group G sends a frame to the bridge multicast address with G in its source field.*



Bridges: Limitations of Bridges

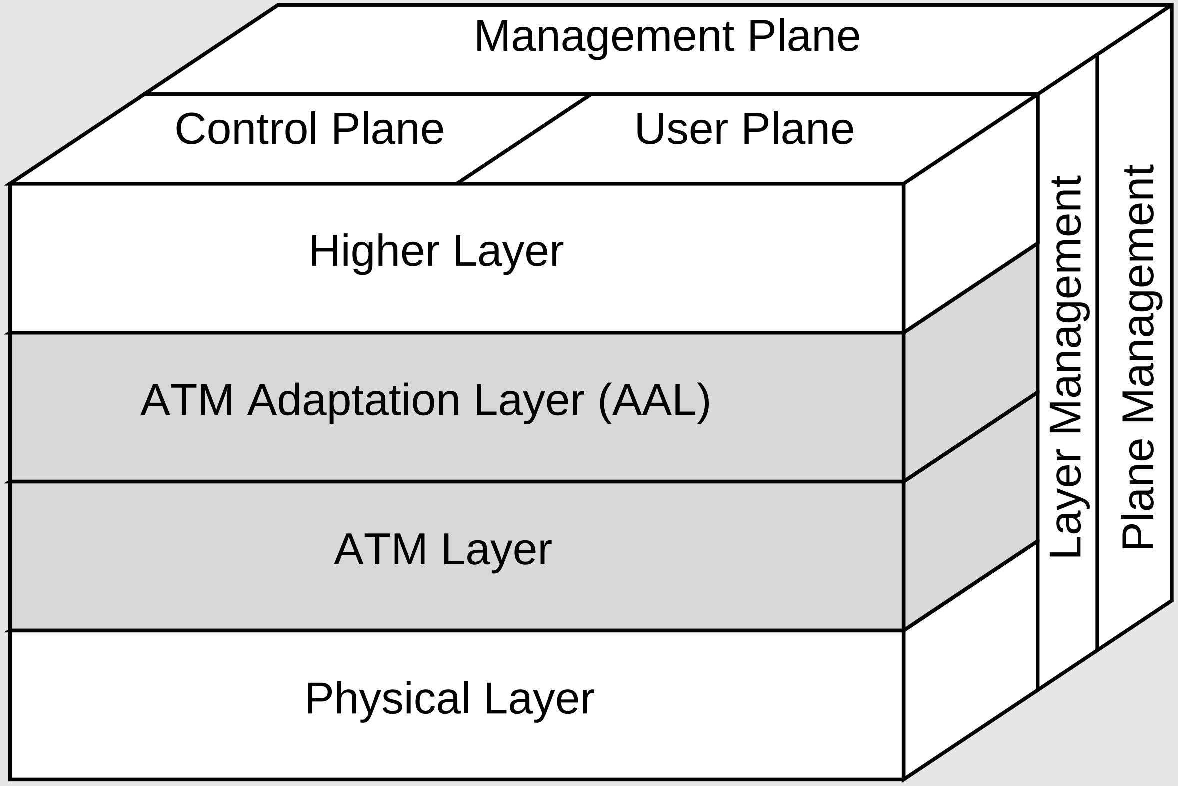
- *Extended LAN's do not scale beyond O(10) bridges*
  - *the spanning tree algorithm does not scale*
  - *broadcast does not scale.*
- *Solution: partition an extended LAN into independent Virtual LAN's: packets are confined to a VLAN.*
- *Bridges do not accommodate heterogeneity.*
- *Caution: beware of transparency*
  - *congested bridges can increase latency and drop frames.*

ATM overview

Asynchronous transfer mode

- *ATM is a connection-oriented packet-switched network.*
- *ATM is used in both WANs and LANs.*
- *The signalling (connection setup) Protocol: Q.2931.*
- *Specified by the ATM Forum.*
- *Packets are called cells: 5-byte header + 48-byte payload.*
- *Commonly transmitted over SONET (but not necessarily).*

ATM Protocol Reference Model



- *user plane: data transfer, flow control, error control*
- *control plane: call control, connection control*
- *management plane: management of the system as a whole.*

ATM cells

Variable versus fixed-length cells

- *There is no optimal fixed-length*
  - *if small: high header-to-data overhead*
  - *if large: low utilization for small messages.*
- *Fixed-length cells are easier to switch in hardware*
  - *simpler*
  - *enables parallelism.*

ATM cells

Small cell size improves the queue behavior

- *Finer-grained pre-emption of the link*
  - *maximum packet = 4KB*
  - *link speed = 100Mbps*
  - *transmission time =* $4096 \times 8/100 = 327.68\mu s$
  - *high priority packet may sit in the queue 327.68*μs
  - *in contrast,* $53 \times 8/100 = 4.24\mu s$ *for ATM.*
- *better queue behavior*
  - *two 4KB packets arrive at same time*
  - *the link is idle for 327.68*μs *while both packets arrive*
  - *at the end of 327.68*μs*, still have 8KB to transmit*
  - *in contrast, can transmit first cell after 4.24*μs
  - *at the end of 327.68*μs *only 4KB left in queue.*

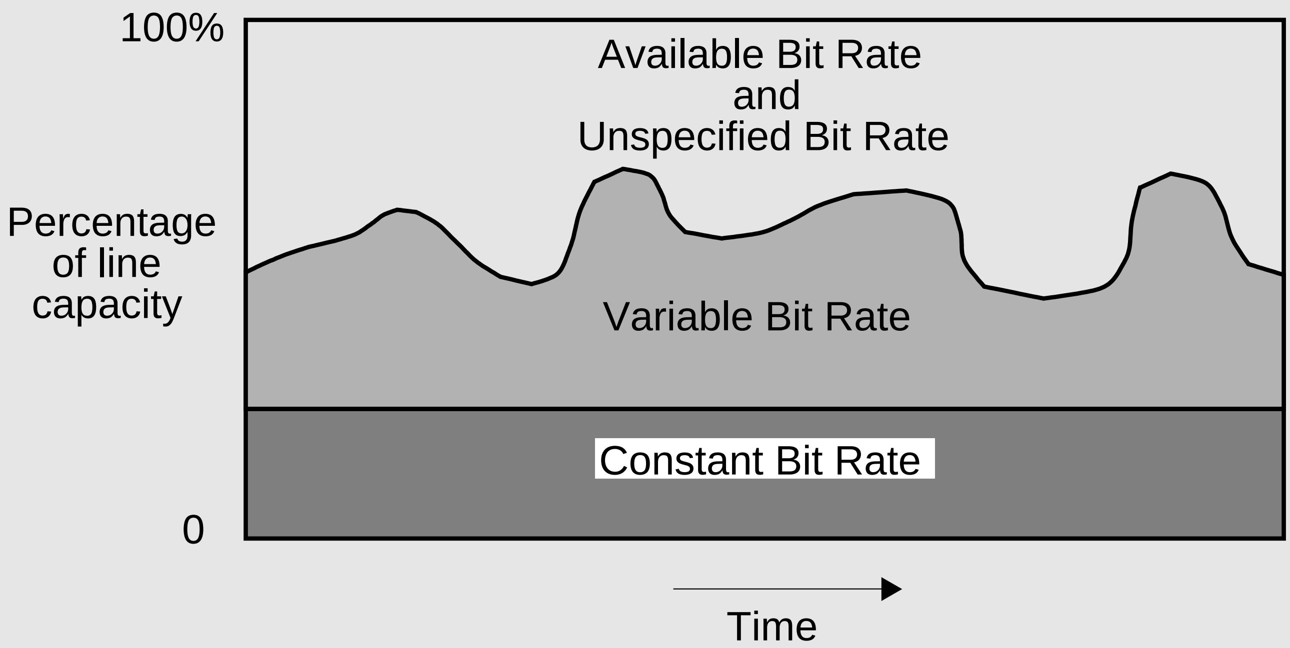
ATM cells

Carrying voice in cells

- small cells: re-assembly overhead
- large cells: padding overhead
- voice digitally encoded at 64Kbps: 8-bit samples at 8KHz, 1 byte every 125μs
- need a full cell's worth of samples before sending the cell
- example: 1000-byte cells implies 125ms per cell which is too long: this latency is noticeable for a human listener
- smaller latency implies no need for echo cancellors.

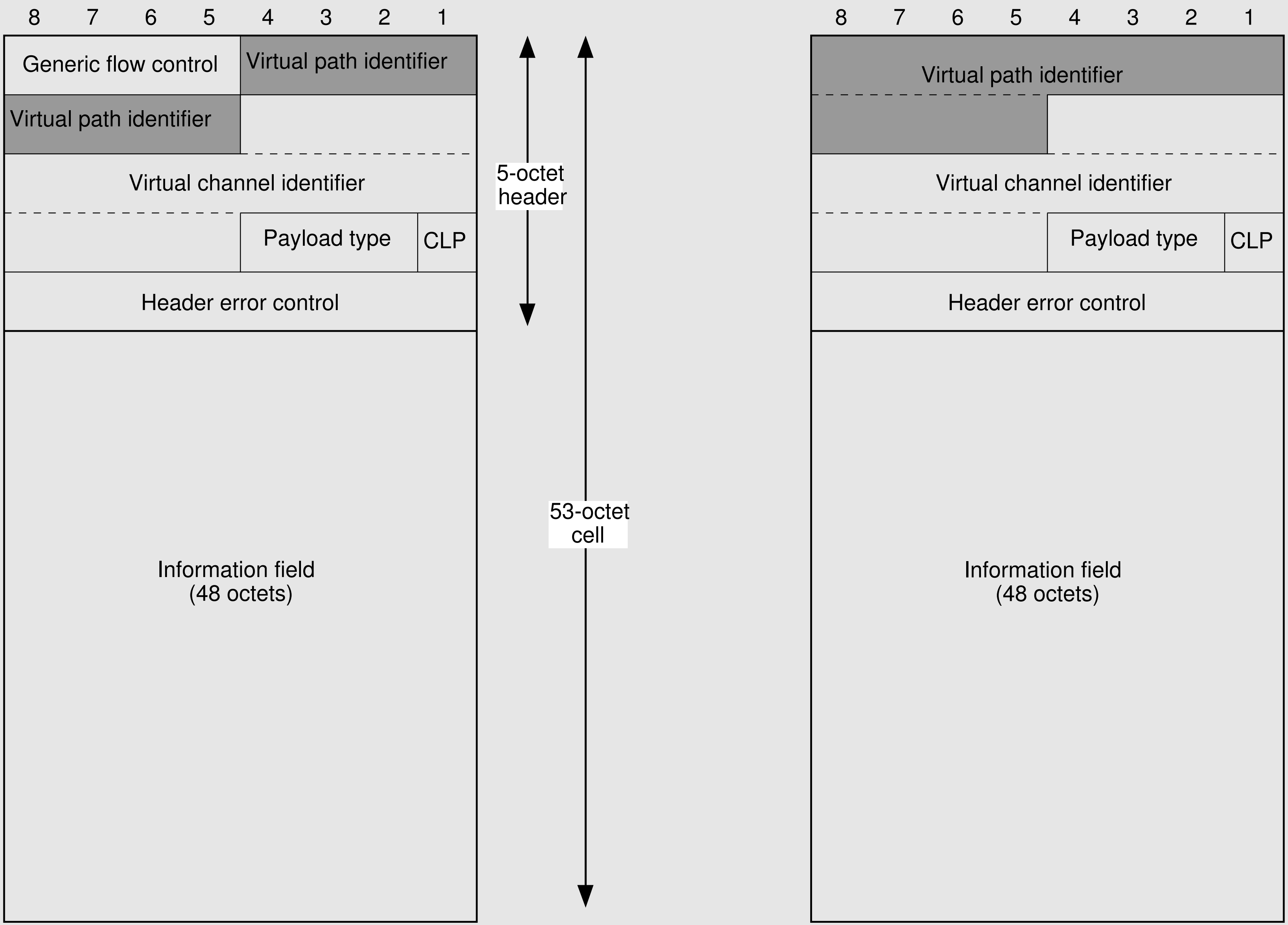
Settled on a compromise of 48 bytes: (32+64)/2 which is not a power of 2.

ATM Quality of Service



QOS is discussed in Chapter 6.

ATM cell format



(a) User-Network Interface

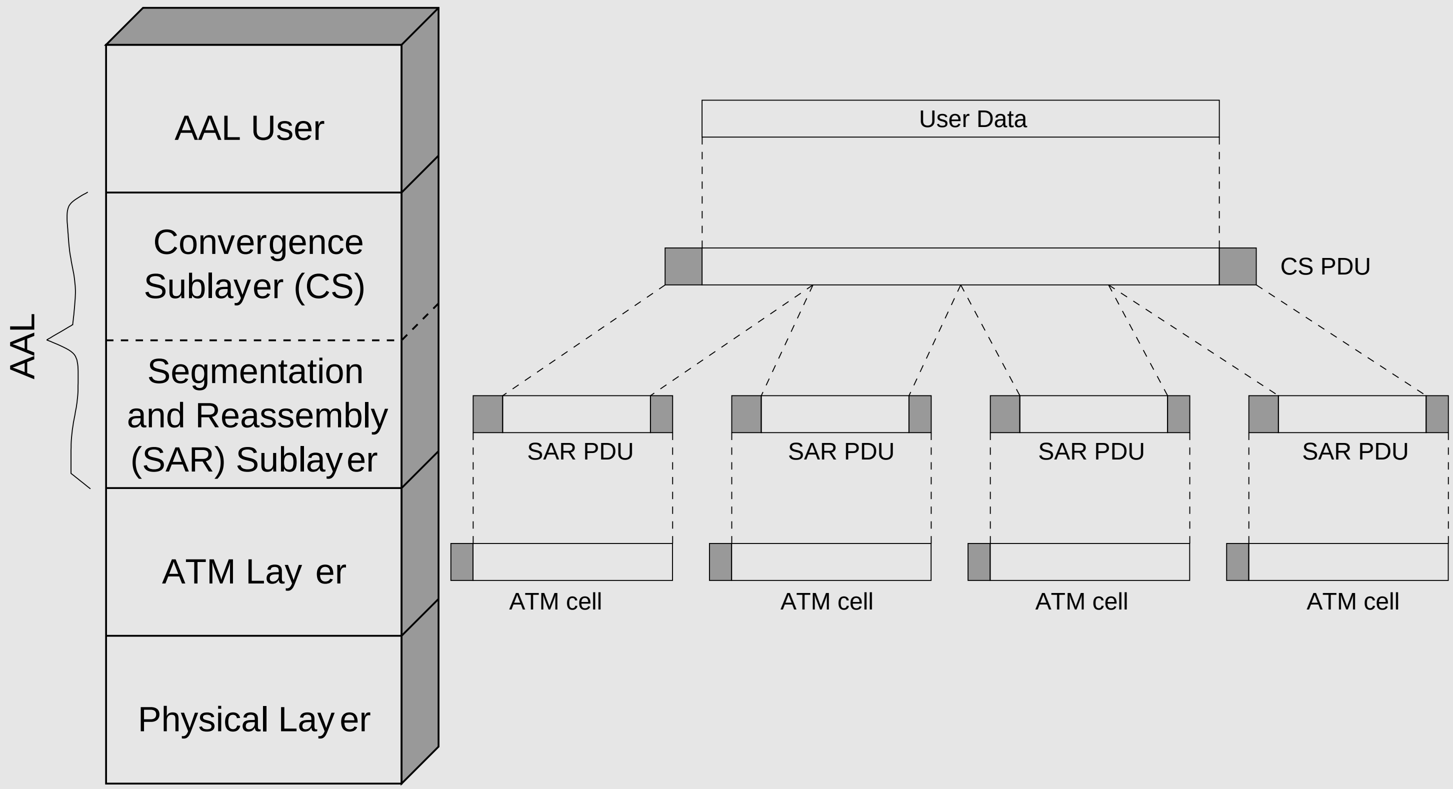
(b) Network-Network Interface

ATM cell format

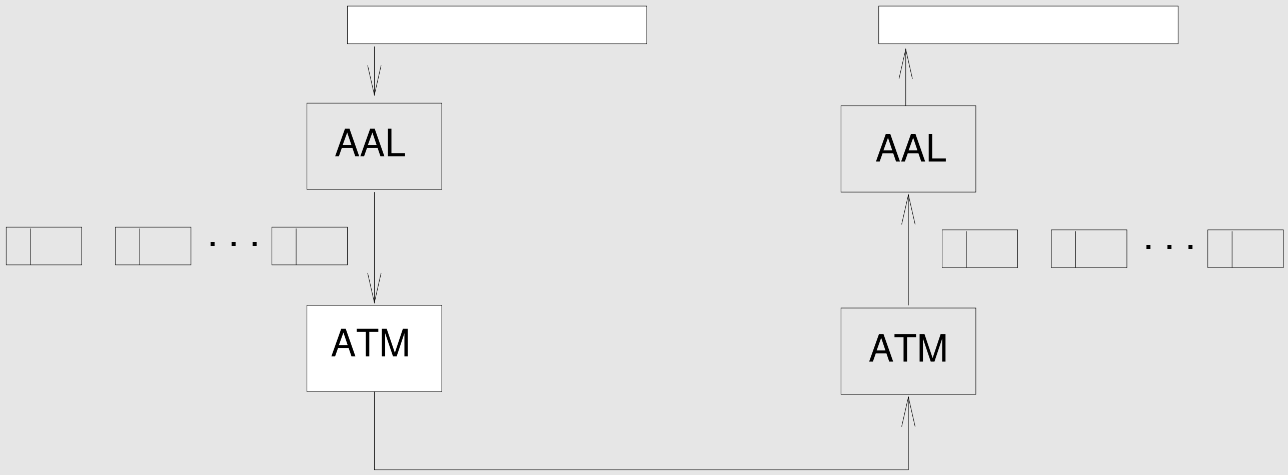
GFC	VPI	VCI	Type	CLP	HEC(CRC-8)	Payload
4	8	16	3	1	8	384 (48 bytes)

- *UNI: User-Network Interface*
  - *host-to-switch format*
  - *GFC: Generic Flow Control (still being defined)*
  - *VPI: Virtual Path Identifier*
  - *VCI: Virtual Circuit Identifier*
  - *Type: management, congestion control, AAL5, user data, . . .*
  - *CLP: Cell Loss Priority*
  - *HEC: Header Error Check (CRC-8) 1-bit error correction*
- *NNI: Network-Network Interface*
  - *switch-to-switch format*
  - *GFC becomes part of VPI field*

ATM: segmentation and re-assembly



ATM: segmentation and re-assembly



The ATM Adaptation Layer (AAL) segments/re-assembles packets/cells

- *AAL 1 and 2 are designed for applications that need guaranteed bit rates such as voice & video*
- *AAL 3/4 is designed for packet data*
- *AAL 5 is an alternative standard for packet data*

ATM: AAL 3/4

8	8	16	<64kbytes	0-24	8	8	16
CPI	Btag	BASize	User Data	Pad	0	Etag	Len

Convergence Sublayer Protocol Data Unit (CS-PDU)

- *CPI: common part indicator (version field)*
- *Btag/Etag: beginning and ending tag*
- *BAsize: hint on the amount of buffer space to allocate*
- *Length: size of the whole PDU*

ATM: AAL 3/4

40	2	4	10	352 (44 bytes)	6	10
ATM header	Type	SEQ	MID	Payload	Length	CRC-10

ATM cell format: 9/53 = 17% overhead.

- *Type*
  - *BOM: beginning of message*
  - *COM: continuation of message*
  - *EOM: end of message*
  - *SSM: single segment message*
- *SEQ: sequence number (wraps around)*
- *MID: message id for multiplexing several PDUs*
- *Length: number of bytes of PDU in this cell: 44 for BOM/COM*
- *CRC: detect errors in 48 byte payload.*

ATM: AAL5

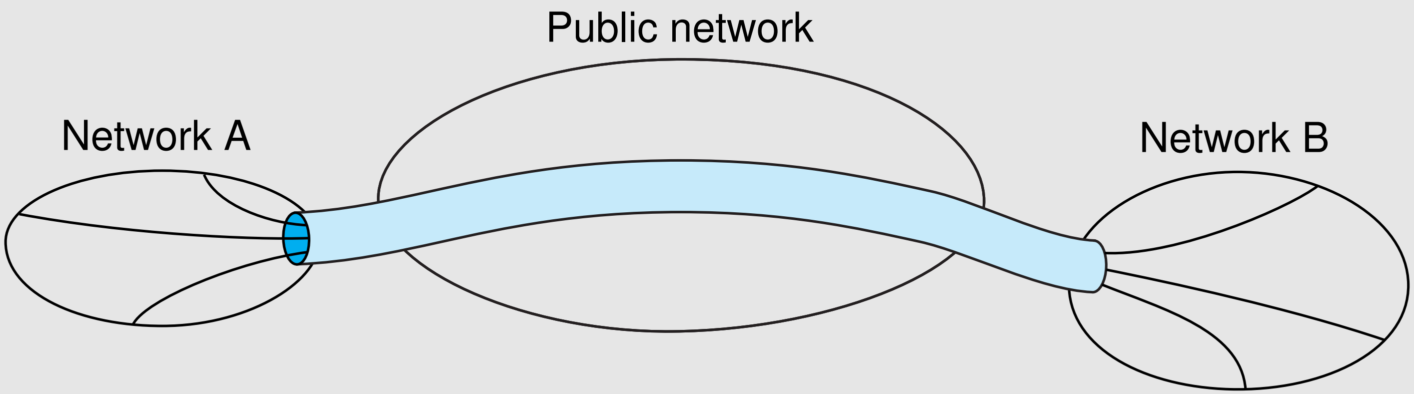
- *CS-PDU Format*

<64kB	0-47 bytes	16	16	32
Data	Pad	Reserved	Len	CRC-32

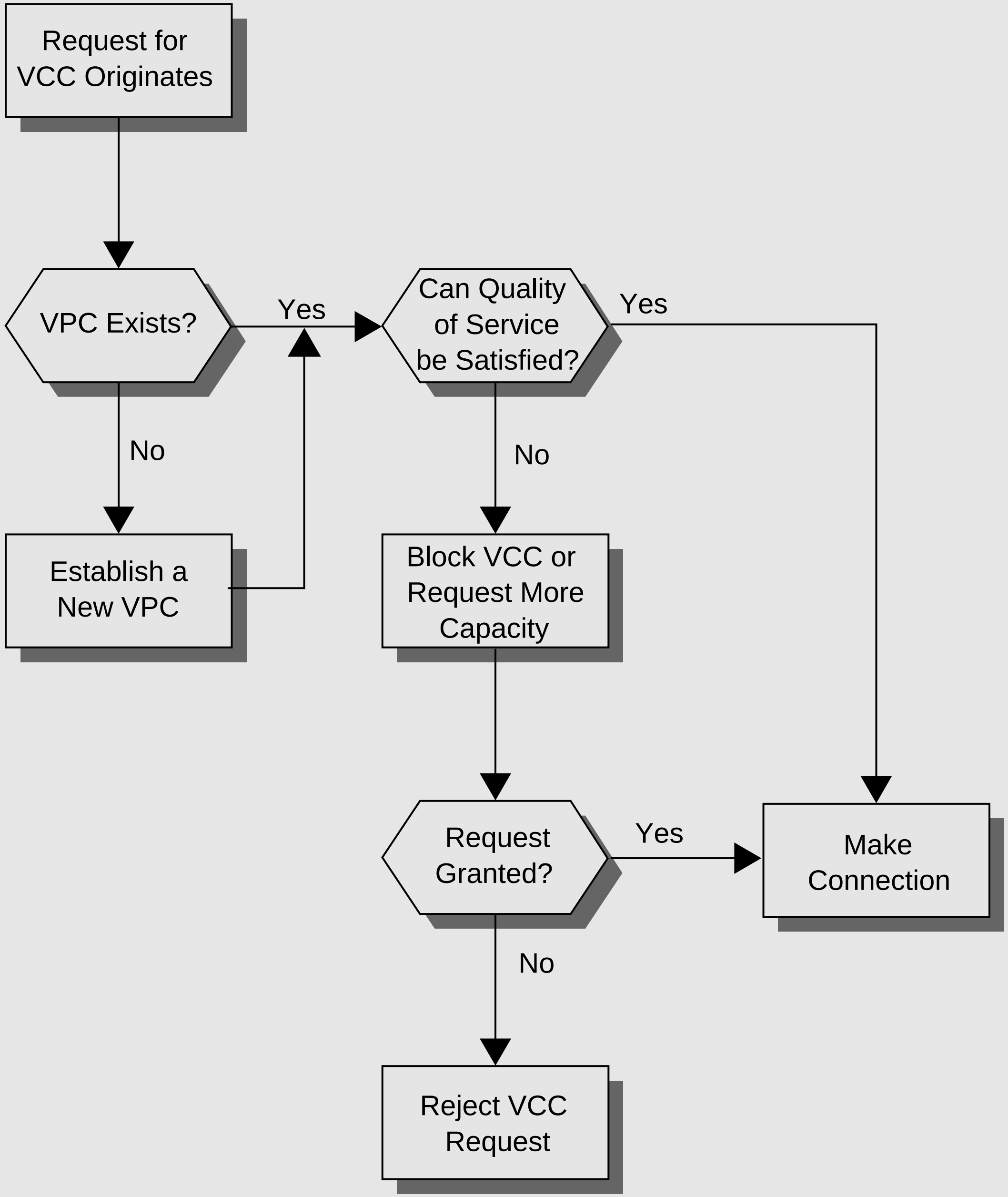
  - *pad so trailer always falls at end of ATM cell*
  - *Length: size of PDU (data only)*
  - *CRC-32 (detects missing or mis-ordered cells)*
- *Cell Format*
  - *end-of-PDU bit in Type field of ATM header*
- *no multiplexing*

ATM VPI/VCI

- *Host: treat as 24-bit circuit identifier*
  - *if cheap: one-per application, use for demultiplexing*
  - *if expensive: multiplex several connections onto one VCI*
- *Network: aggregate multiple circuits into one path*

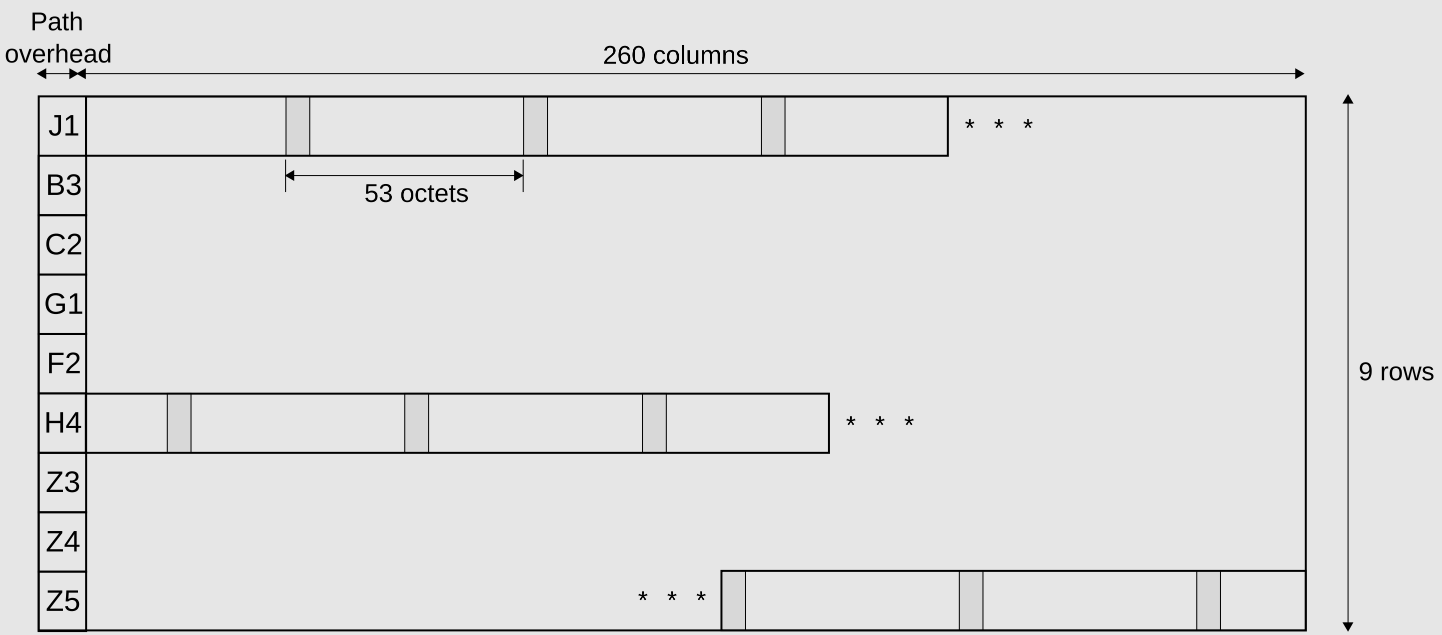


ATM: call establishment using VPs



Chapter 3.3

ATM over SONET

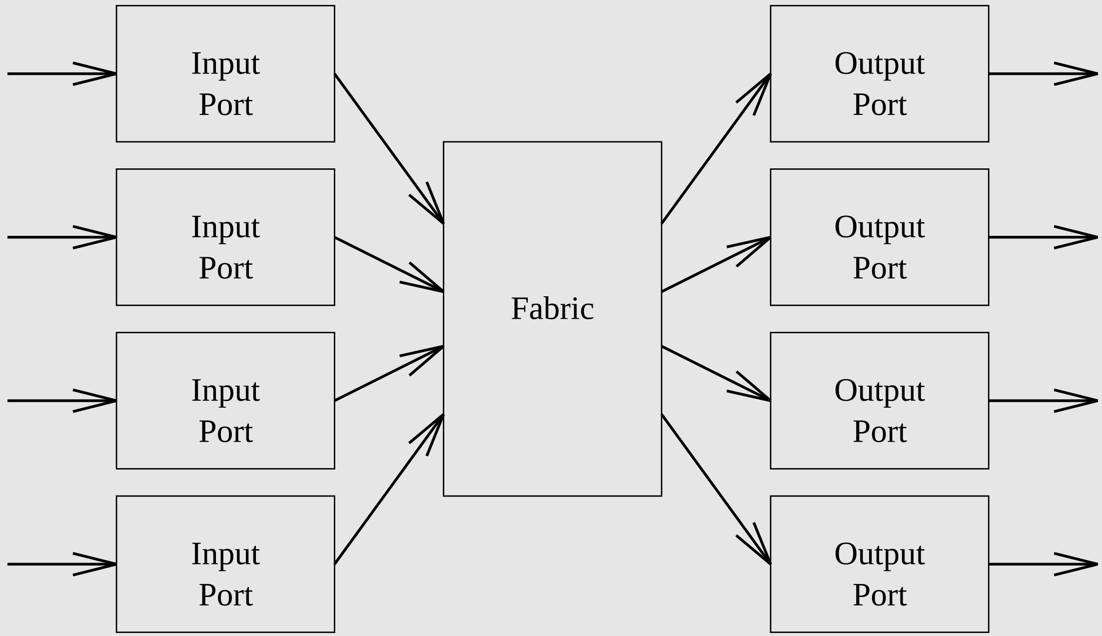


STM-1 (155.2Mbps) payload for SDH-based ATM cell transmission.

Hardware: overview

- *Terminology*
  - *an $n \times m$ switch has n inputs and m outputs.*
- *Design goals*
  - *the throughput of the switch depends on the traffic pattern*
    - *blocking may occur if all packets are switched to the same output port*
    - *switch designers use complex traffic models*
    - *good traffic models exist for telephone traffic*
    - *difficult to get accurate traffic models for IP traffic*
  - *scalability: a function of n*
  - *cost.*

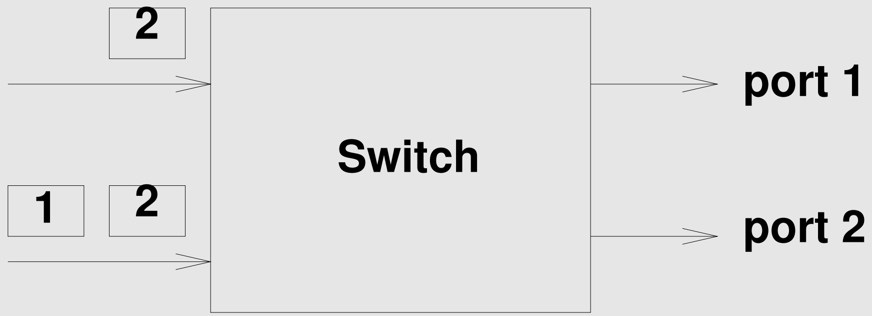
Hardware: ports and fabrics



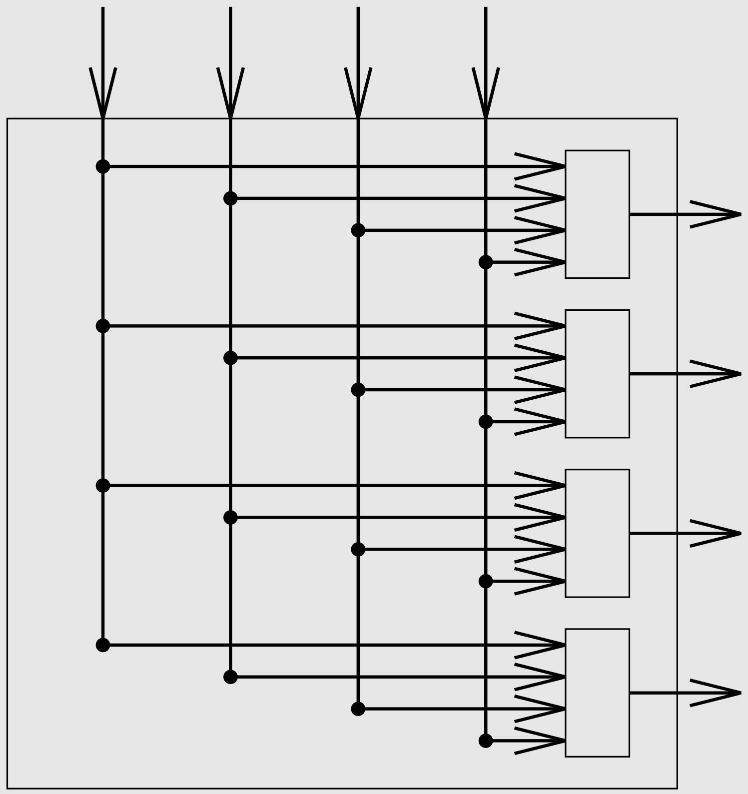
- *ports*
  - *circuit management (e.g. map VCIs, route datagrams)*
  - *buffering: input and/or output*
- *fabric*
  - *as simple as possible*
  - *sometimes do buffering in the fabric*

Hardware: buffering

- *Wherever contention is possible*
  - *input port: contend for fabric*
  - *internal: contend for output port*
  - *output port: contend for link*
- *Head-of-Line blocking*
  - *input buffering*
  - *Quality of Service*



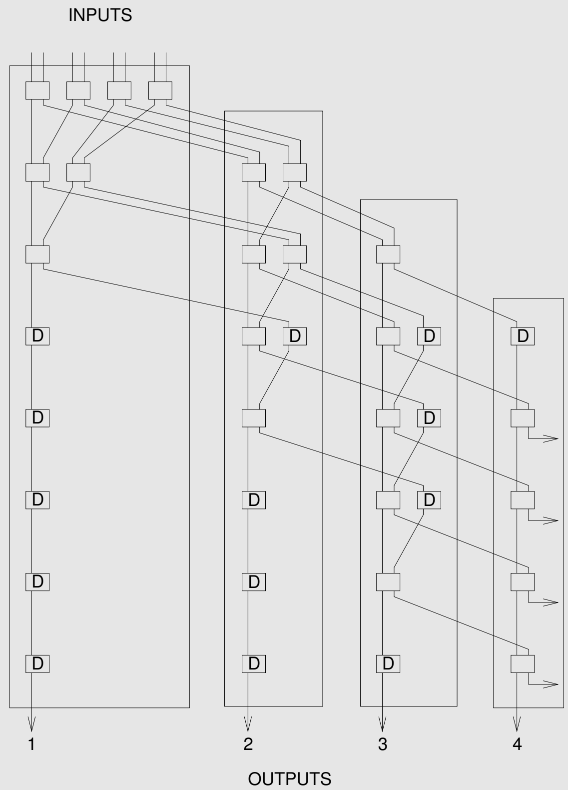
Hardware: crossbar switches



- *every input is connected to every output*
- *non-blocking*
- *complexity* n^2.

Hardware: knockout switch

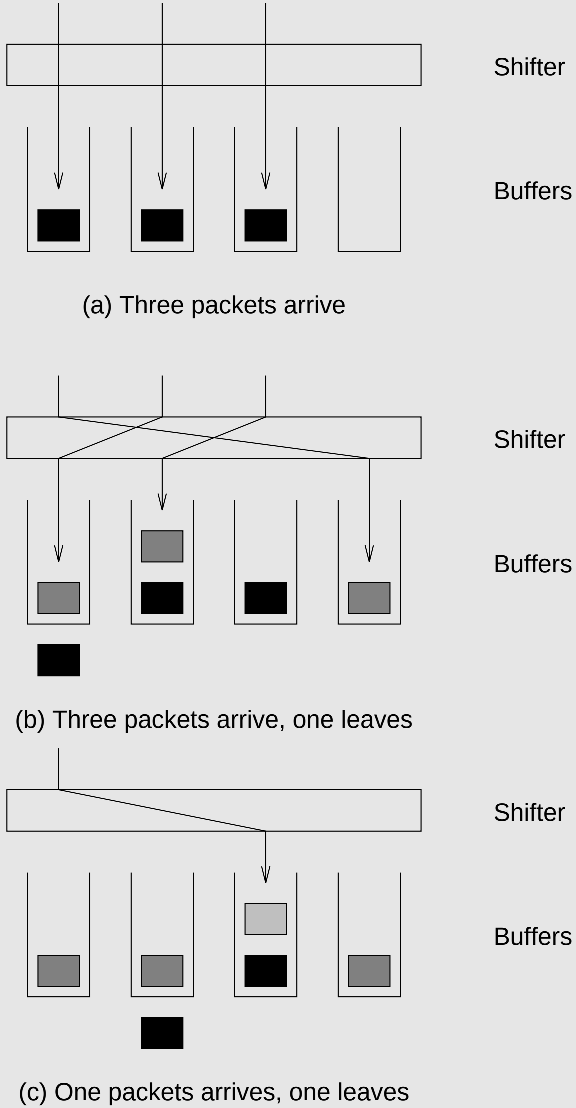
- *An example of a crossbar switch*
- *The concentrator selects ℓ of n packets destined for the same output port, the other packets are dropped*



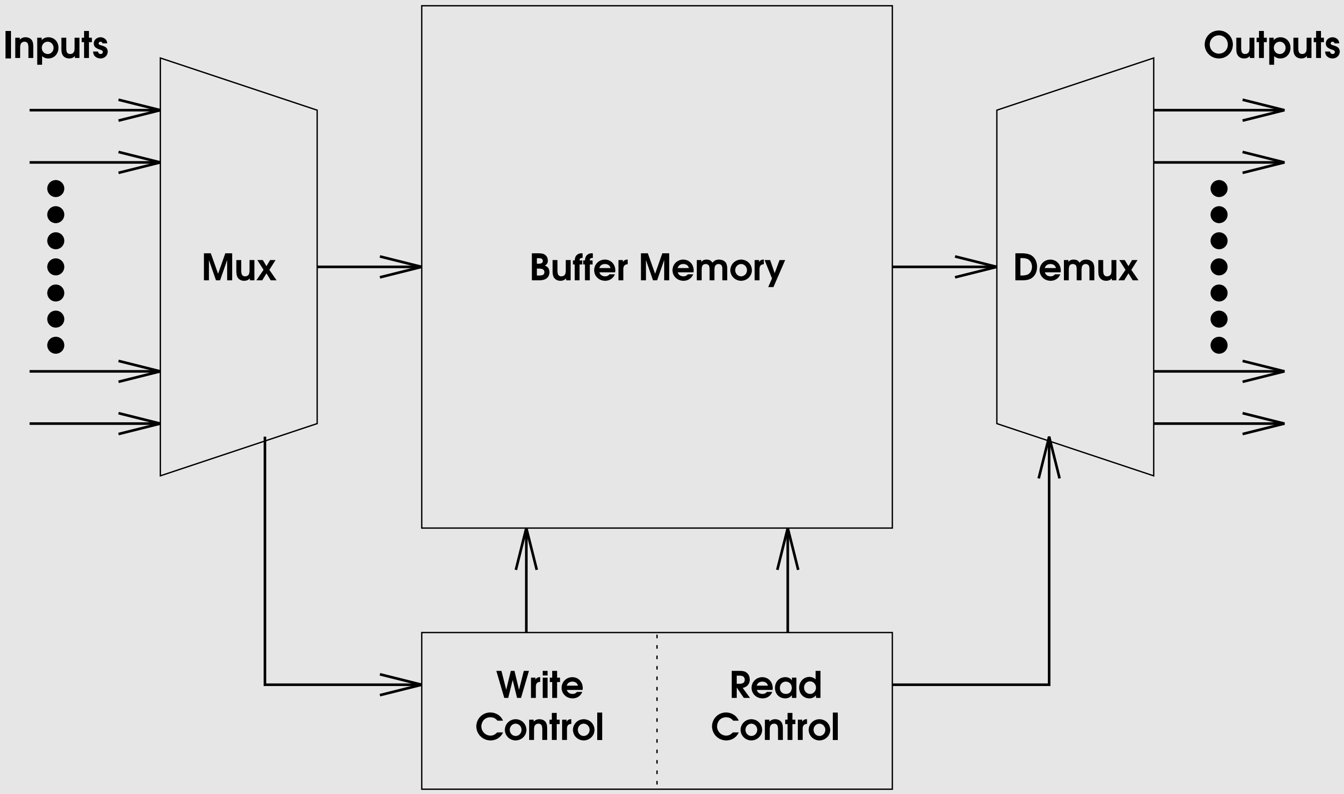
- *Complexity: n^2.*

Hardware: knockout switch

- *Each output port has a buffer which accepts up to ℓ packets/cycle & transmits 1 packet/cycle.*



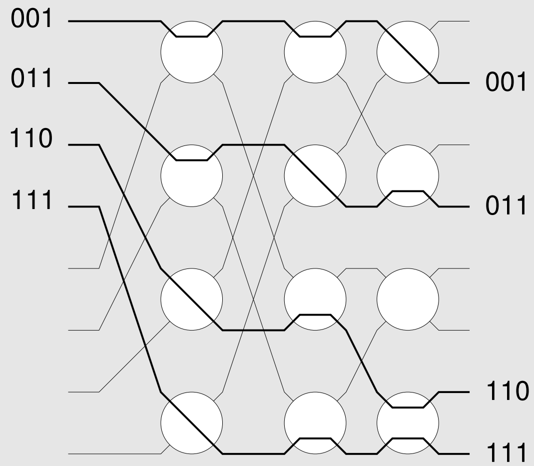
Hardware: shared media switch



Hardware: self-routing fabrics

Banyan Network

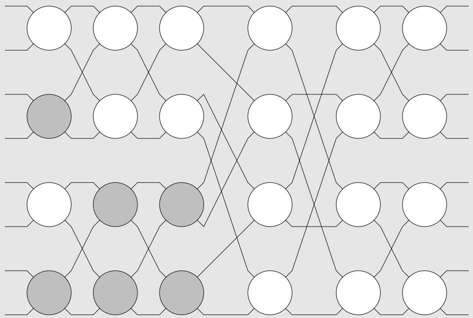
- constructed from simple 2×2 switching elements
- self-routing header attached to each packet
- elements arranged to route based on this header
- no collisions if input packets are sorted into ascending order
- complexity: $n \log_2 n$.



Hardware: Batcher network

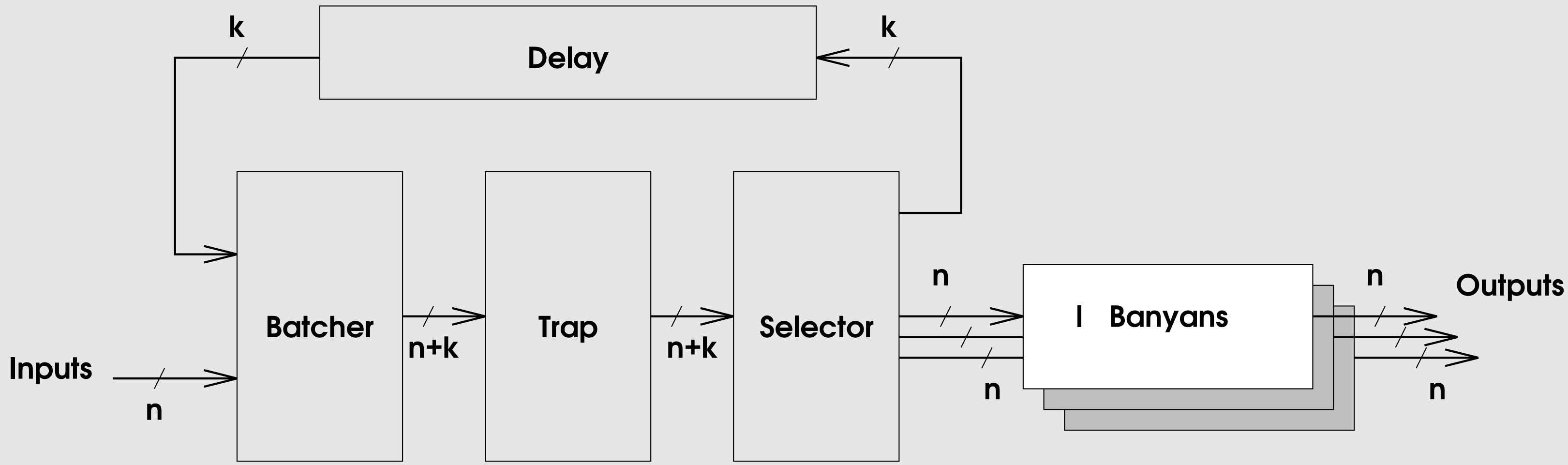
A Batcher network sorts packets into ascending order

- *switching elements sort two numbers*
  - *some elements sort into ascending (clear)*
  - *some elements sort into descending (shaded)*
- *elements arranged to implement merge sort*
- *complexity:* $n \log_2^2 n$.



Common design: Batcher-Banyan switching fabric.

Hardware: sunshine switch



Each output port accepts ℓ packets per cycle. If more than ℓ packets/cycle are sent to an output port, they are re-circulated, raised in priority & re-submitted to the switch in the next cycle.