# Gigabit Kits Course
# Switch Architecture

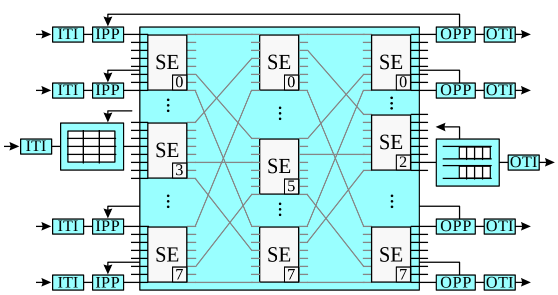
## Summer 1998

Jonathan Turner
Washington University
Computer Science Department

http://www.arl.wustl.edu/~jst/gigatech/kits.html

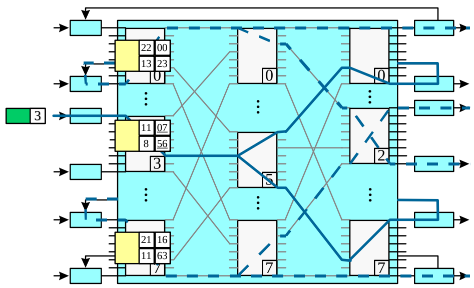


# WUGS Architecture




- Scalable switch architecture
- Multistage interconnection network
  - » 8 port, shared buffer *Switch Elements* (SE)
  - » interstage flow control
  - » dynamic routing
  - » generalized Benes topology
  - » support for binary multicast and range-copy multicast
- *Input and Ouput Transmission Interfaces* (ITI,OTI) include optoelectronics and transmission line coding, synchronization, etc.
  - » an interface may support multiple external links
- *Input Port Processor* (IPP) performs routing table lookup for received cells
- *Output Port Processor* (OPP) queues cells awaiting transmission
- *Recycling Paths* connect OPPs to corresponding IPPs
  - » used for multicast virtual circuits and for in-band configuration

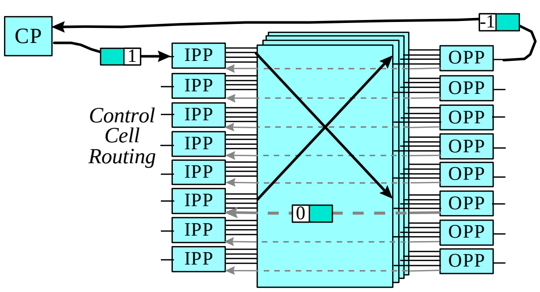
# Basic Switching Operation

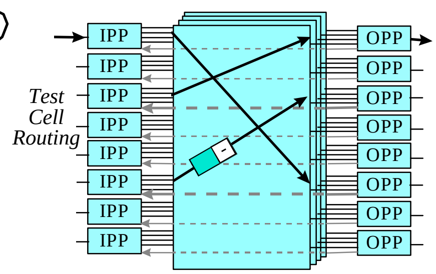


- Routing lookup at IPP yields output port number & VPI/VCI
  - » binary multicast cell gets pair of port numbers and VPI/VCI
- First stage switch elements distribute traffic to balance load
  - » in general, first $k$ stages of $2k+1$ stage network
  - » ensures traffic on internal links cannot exceed external traffic
- Second and third stages route cells using destination port number
  - » first octal digit of port number used in second stage, second digit in third stage
  - » binary multicast cell is copied at first stage where the octal digits of ouput port numbers differ
  - » after copy point, cell treated as unicast
- One or both copies of multicast cells can be *recycled* back to input side
  - » VPI/VCI used for new table lookup, yielding new routing information
- Can produce $f$ copies of cell in $\log_2 f$ passes



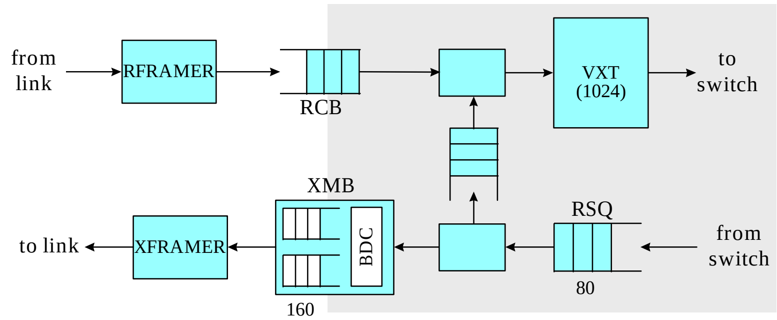
# In-Band Configuration and Management




- Switch configuration and management cells from remote processors are forwarded through switch to target IPP or OPP.
  - » read/write VXT entries
  - » read counters (cells passed, buffer overflow, HEC errors, ...)
  - » set configuration registers (link enable/disable, queue thresholds, ...)
- Can also reset entire switch (action initiated at IPP where cell first received).
- Control offset mechanism and open cell format provide flexibility.
- Three hop cells enable path testing.
- Control cell reception can be selectively enabled on per port basis.

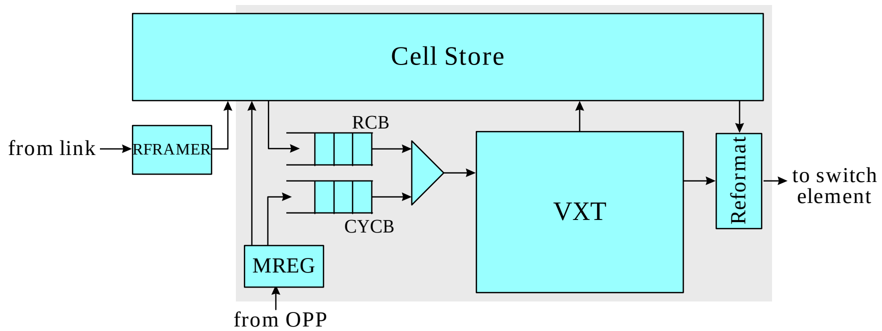
# Port Processor Logical Organization




- Framer section matches different transmission interfaces.
  - » 16 bit for OC-3C, OC-12C, G-link; 32 bit for OC-48C
- VXT handles both virtual paths and virtual circuits.
- RCB and XMB separate link and switch timing regimes.
- Resequencing buffer forwards cells in order they *entered* interconnection network
- Transmit buffer separates CBR,VBR from ABR,UBR; packet level discard using EPD with hysteresis.



# Input Port Processor Design



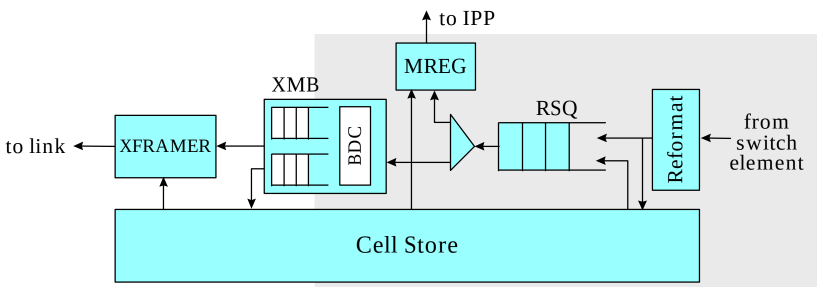

- Cells placed in common Cell Store on entry; other circuits pass pointers plus control fields.
- Cell store holds 64 cells; VXT has 1024 entries.
- Maintenance register provides access to configuration/status information.
  - » link status, cell counts, HEC error count, buffer overflow count, . . .
  - » RCB discard threshold, VXT bounds register, transitional time stamping parameter, . . .

# Input Port Processor Chip Layout


- Total area: 200 mm$^2$
  - » 80 mm$^2$ wiring (40%)
  - » 54 mm$^2$ memory (27%)
  - » 25 mm$^2$ logic (13%)
  - » 22 mm$^2$ pad ring (11%)
  - » 20 mm$^2$ empty space (10%)
- Total Transistors: 1,468K
  - » 1,105K in memory (75%)
  - » 362K in logic, pads (25%)
- VXT consumes largest share of memory
- Cell Store is next largest
  - » dominated by access registers
- Worst-case power: 5.8 W
  - » 90% core, 10% pads

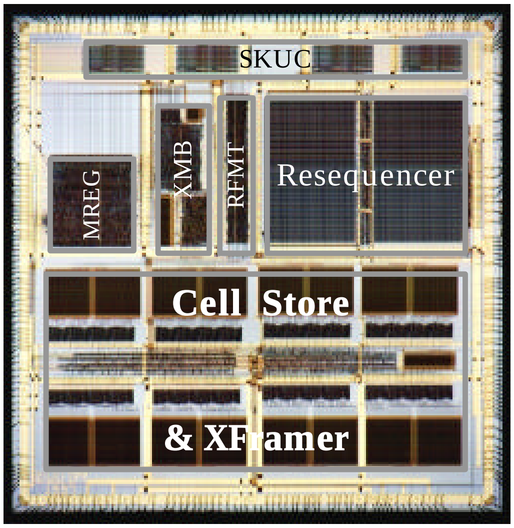


# Output Port Processor Design




- Cells placed in common RAM on entry; other circuits pass pointers plus control fields.
- Resequencer reorders pointers according to timestamp information.
- Maintenance register on recycling path provides control access to hardware registers.
  - » cell counters, buffer overflow counters, parity error register, . . .
  - » XMB configuration and discard thresholds, resequencer age threshold, . . .

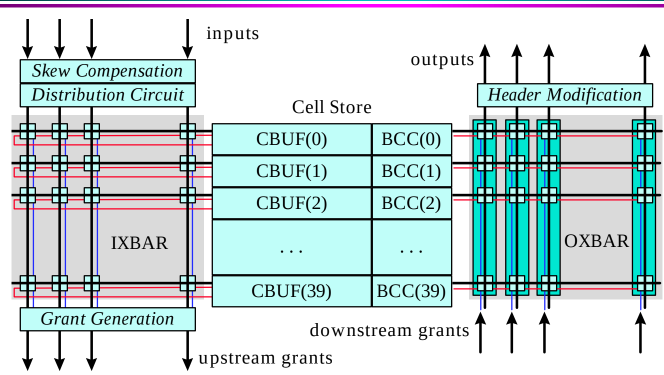
## Output Port Processor Chip Layout




- 0.7 µm CMOS
  - » 256 cells in cell store
  - » 80 cells in resequencer
- Total area: ≈180 $mm^2$
- Total transistors: 1,221K
  - » about 65% in memory
  - » about 35% in logic
- Cell Store consumes largest share of chip area
  - » dominated by overhead
- Resequencer uses about 20%

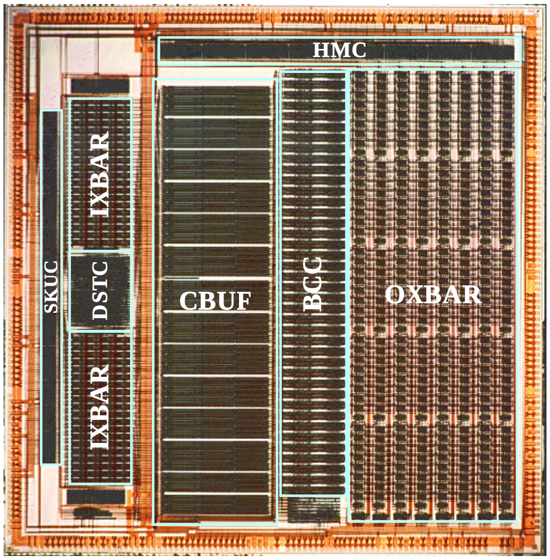


## Switch Element Organization




- Four chips implement 8 port switch element.
- 40 cell shared buffer; grant flow control.
  - » when buffer too full for 8 new cells at once, grants rotated among inputs
- Distribution circuit does round-robin assignment of arriving cells to outputs.
- OXBAR selects cells based on dynamic priority (increases with cell waiting time).
- Skew compensation allows two clock periods of clock/data skew.
  - » inserts variable delay to offset skew; tracks delay changes

# Switch Element Photo




- .7 µm CMOS
- 14.5 by 14.8 mm
- 650,000 transistors
- Oxbar consumes largest share of area
  - » control & wiring dominate
- Cell store and buffer control use comparable areas



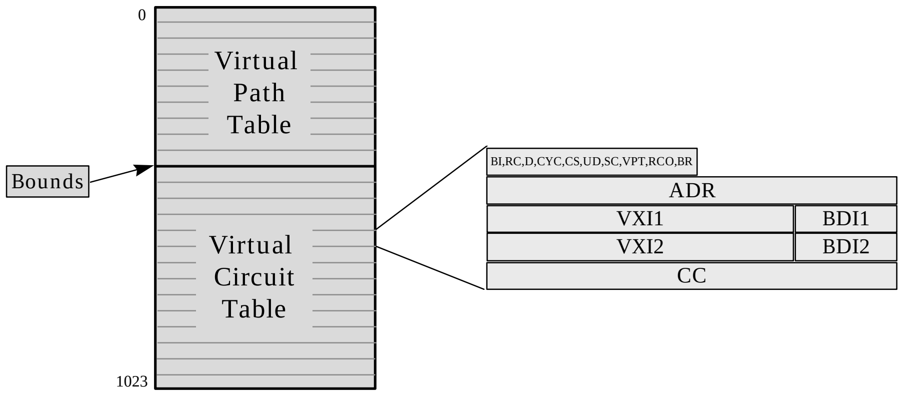
# Internal Cell Format

| | 4 | 4 | 12 | 5 | 5 | 6 |
|---|---|---|---|---|---|---|
| BI | RC | -- | STG | D,CYC,CS,BR | -- | UD,PT,CLP |
| - | ADR | VXI1 | | | | BDI1 |
| | | VXI2 | | | | BDI2 |
| | | Payload | | | | |
| | reserved | | | | | |
| TS | | | | | | |

- Busy/Idle (BI)
- Routing Control (RC)
  - » unicast 0 or 1
  - » specific path
  - » binary copy
  - » copy range
- Address (ADR)
  - » single, pair or complete path
- Time stamp (TS)
- Source (STG)
- Virtual Path/Circuit Identifier (VXI1,VXI2)
- Block Discard Index (BDI1,BDI2) for packet level discarding

- Data bit (D), Recycling bits (CYC), Continuous Stream Bit (CS), Bypass Resequencer (BR), Upstream Discard (UD), Payload Type (PT), Cell Loss Priority (CLP).

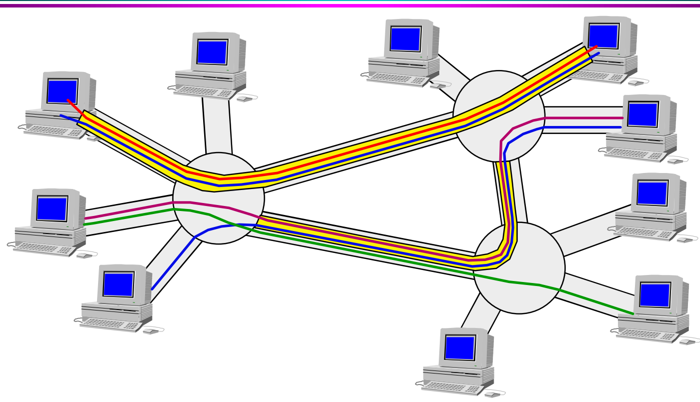
# Virtual Path / Circuit Table




- Adjustable boundary, up to 256 VP table entries.
- Shared Virtual Circuit Table means terminating VPs must have disjoint VCs.
- Cell Counter (CC), Set CLP (SC), Virtual Path Termination (VPT), Recycling Cells Only (RCO)

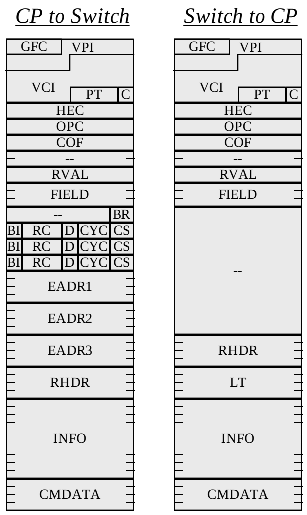


# Virtual Paths and Circuits



- Virtual paths combine collection of VCs together.
  - » intermediate switches route cells using VPI only and only translate the VPI
  - » switches at VP termination points, switch using both VPI and VCI
- Use of VPs conserves table entries in intermediate switches.
- New VCs can be established over VPs without involvement of control processors in intermediate switches.

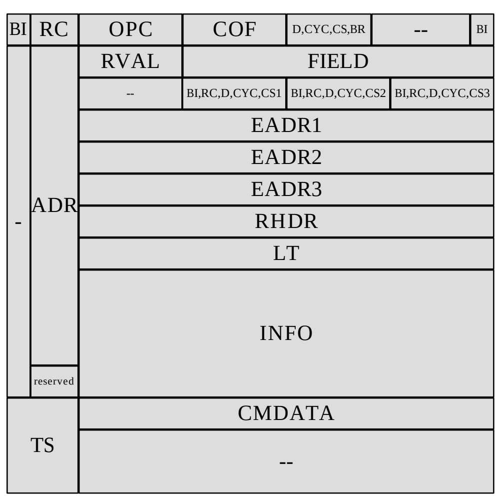
# External Control Cell Formats



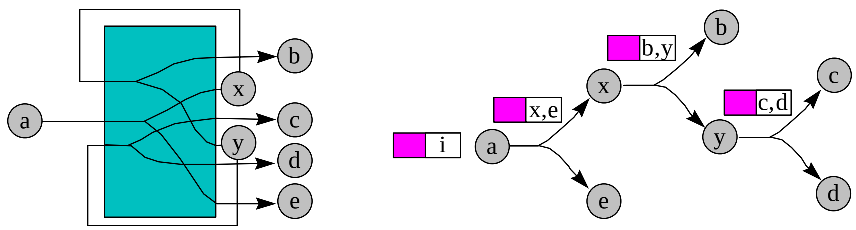

- Operation Code (OPC) specifies operation.
- Control Offset (COF) identifies target.
- Return Value (RVAL) for returning status.
- Field or table entry to be accessed (FIELD).
- BI, RC, D, CYC, CS fields for each of three hops through the switch.
- External address (EADR1,2,3) specifies internal addresses used in each of three hops.
- Return Header (RHDR) is cell header of returned cell
- Information field (INFO) contains information read from or written to table entry / register field.
- Local Time (LT) field gives switch time at which information was accessed.
- Connection Management Data (CMDATA) used to correlate responses with requests.



# Internal Control Cell Format



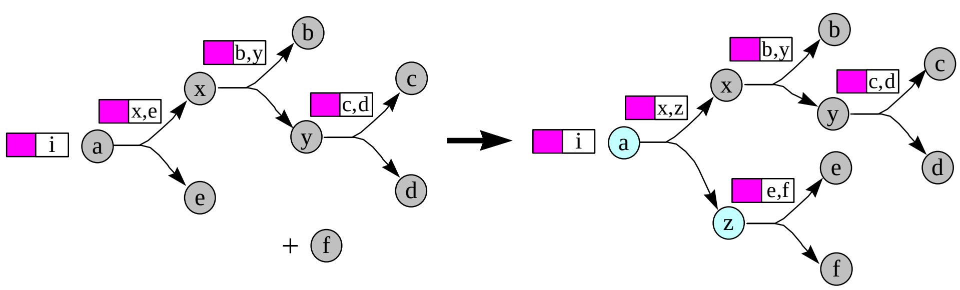
# Multicast Connection Trees




- Each multicast connection defines a binary tree.
  - » input at tree root
  - » outputs at leaves
  - » internal nodes are ***recycling ports***
- Recycling ports can be dedicated to recycling only, or can be shared between recycling traffic and external traffic.
- Note: multicast with *m* outputs uses (*m*–1) table entries.
  - » fewer entries than equivalent number of unicast connections

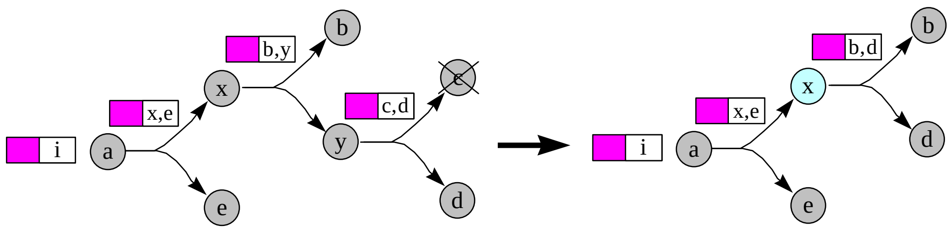


# Adding an Endpoint




- Pick a *shallow* leaf *e* and recycling port *z*.
- Make *e* and new leaf *f* children of *z*.
- Make *z* a child of *e*'s former parent.
- Note: two table entries are changed.
  - » independent of switch size and connection fanout
- Selection of shallow leaf, limits number of passes to $\log_2$(maximum fanout)
  - » per pass delay is 10 µs, so multicast with fanout of 256 can be implemented with maximum delay of 80 µs

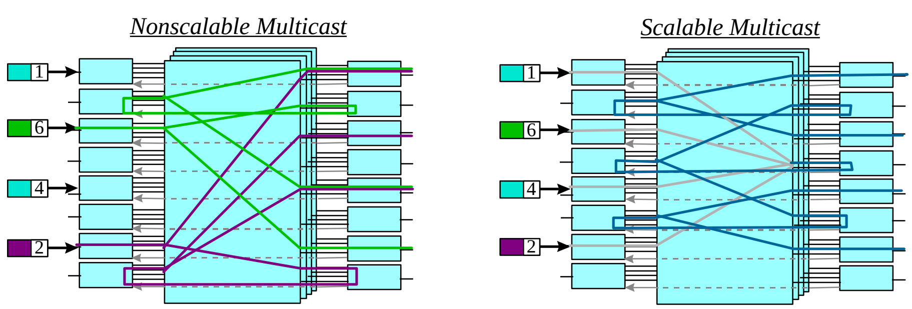
# Dropping an Endpoint




- Let *c* be leaf to be removed.
  - » if *c* has a grandparent in tree, let *y* be its parent, *x* its grandparent and *d* its sibling; in *x*'s VXT entry, replace *x* with *d*
  - » if *c* is the child of the tree root and it has a sibling with children, redirect the root's pointers to the sibling's children
  - » if *c* is the child of the tree root and it has no sibling, or its sibling is childless, simply remove it
- Note that in all cases, only one table entry changed.
- No need to balance tree after deletion.

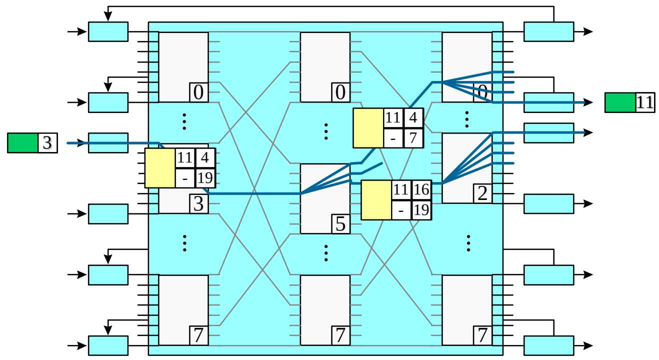


# Scalable Many-to-Many Multicast




- Overlaid one-to-many trees yields poor scaling properties.
  - » *m*-way multicast consumes *m*(*m*–1) routing table entries
  - » adding another endpoint requires changing 3*m*–1 table entries
- Common tree yields fully scalable multicast.
  - » upstream discard option prevents unwanted "return cells"
  - » *m*-way multicast consumes 2*m*–1 table entries
  - » adding another endpoint requires changing 3 table entries

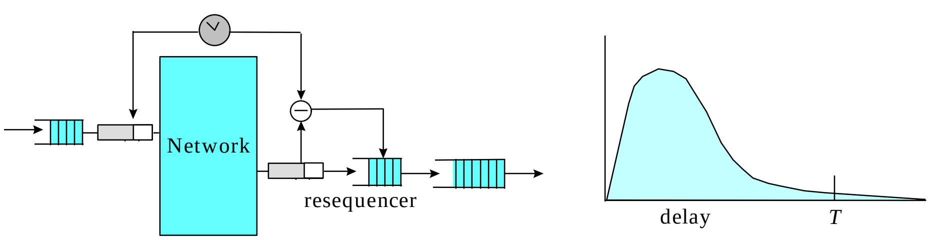
# Range-Copy for Multicast



- Address pair interpreted as defining range.
- Ranges modified as cells pass through network.
- All copies get same VCI, limiting general use.
  - » potential application for broadcast of popular video channels to mux'ed outputs
- Copies can still be recycled to obtain unique VCIs.
  - » allows general use and potential for improved average-case performance

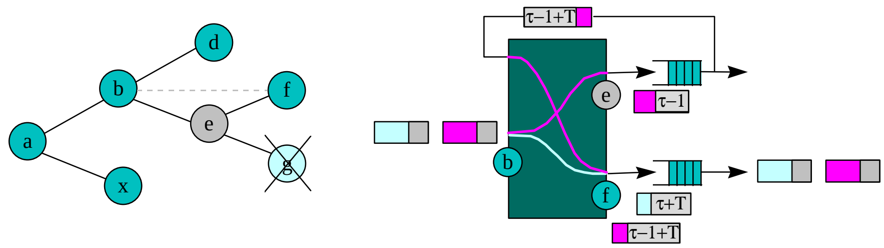


# Cell Resequencing




- Dynamic routing allows cells to get out of order.
- Time-based resequencing involves time-stamping cells at input and releasing at output in order of entry time.
- Fixed *age threshold T* equal to max delay expected in network.
- If mean and variance of per stage delay is 3 cell times, then, mean delay+10 std. dev. ≈67 cell times for 7 stages.
  - » for $d$=8, 7 stages yields 4,096 port switch
  - » with internal cell time of 133 ns (16 clock ticks at 120 MHz), 67 cell times is ≈9 µs

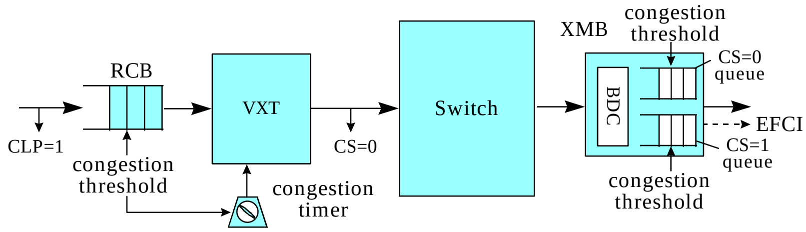
## Avoiding Misordering During Transitions




- Endpoint removal requires extra delay after change.
- *Inflate* time stamps of cells arriving just after the change.
  - » increase by *T* initially right after change
  - » reduce increment to 0 over next *T* cell times
- Let τ be time of change and *T* be resequencer delay.
  - » cells arriving between τ and τ+*T* assigned time stamp in range τ+*T* to τ+2*T*
  - » avoid *time-stamp collision* by giving clock half-step precision
  - » required resequencer size increases by ratio of maximum virtual circuit rate to link rate



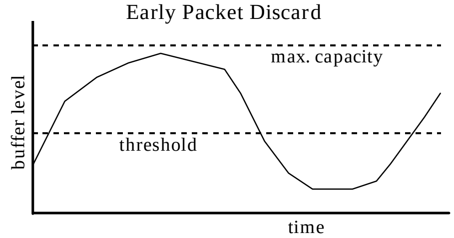
## Congestion Control Mechanisms



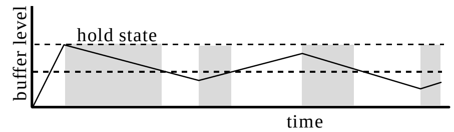

- Output side congestion control
  - » packet-level discarding of cells with BDI>0 uses combination of Partial Packet Discard and Early Packet Discard with hysteresis
  - » CLP=1 cells discarded when queues above threshold
  - » EFCI bit set in outgoing cells when queue is above threshold
- Input side congestion control
  - » congestion in switch or excess recyling traffic can cause flow control to backup into IPP causing RCB to fill beyond congestion threshold
  - » input congestion causes CLP=1 cells to be discarded at RCB input and CS=0 cells to be discarded at VXT
  - » VXT discarding action continues until RCB is below threshold for timeout period

# Packet Discard Mechanism

Early Packet Discard




with Hysteresis

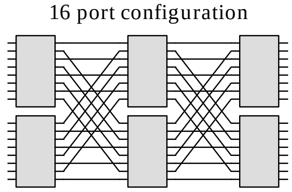



- Discard packets, not cells during overload periods to avoid *congestion collapse*.
- Partial Packet Discard (PPD) (discard remaining cells in packet, once you have discarded one) improves *goodput*, but cannot avoid congestion collapse.
- Early Packet Discard can achieve 100% goodput with large enough buffers.
  - » need about $k$ packets worth of buffering where $k$=(VC rate)/(link rate)
  - » acceptable goodput (>50%) even with moderate buffers
- Hysteresis reduces variability in buffer usage dramatically.
  - » 100% goodput with buffer capable of holding two packets
  - » yields better fairness properties than standard EPD
- Enabled for VCs with non-zero BDI; uses AAL5 framing.

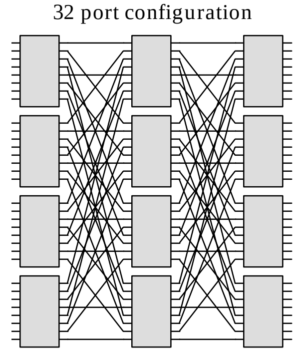


# Alternative Network Configurations

16 port configuration



32 port configuration



- Benes topology with 8 port switch elements limits network sizes to powers of 8.
- Allow power-of-2 network sizes by modifying connections at center stage.
  - » $2k$-1 stages for $8^k$, $8^k/2$ or $8^k/4$ port networks
- Middle stage does combination of traffic distribution and route-copy.
  - » for switch size of $8^k/4$ middle stage uses one address bit for routing
  - » for switch size of $8^k/2$ middle stage uses two address bits for routing
- No change to operation of switch elements before and after middle stage.

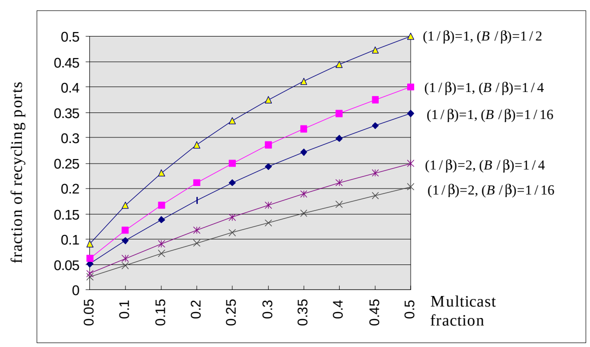
## Speed Advantage for Nonblocking Multicast

- Let $\beta$ be maximum entry/exit load on switch port (as fraction of internal data path speed).
- Number of internal nodes in multicast connection trees is less than number of leaves, so recycling bandwidth is less than output bandwidth.
- Since total exiting traffic is $\leq \beta n$, there must always be some recycling port with load $\leq \beta$.
- Result: if $2\beta+B\leq 1$, there is always a recycling port that can accommodate a new connection of rate $B$.
- If $\delta$ is fraction of exiting traffic in multicast connections, it's enough to have $(1+\delta)\beta+B\leq 1$ or equivalently, $(1/\beta)\geq 1+\delta+B/\beta$.
- Note that required speed advantage independent of $n$.
- Examples: If $\beta=B$ and $\delta=1$, a 3x speed advantage is required. If $B=\beta/16$ and $\delta=.2$, a speed advantage of 1.26 is enough.
- If instead of recycling at all ports, we dedicate $h$ ports to recycling, the system is nonblocking if

$$\frac{(n-h)\delta\beta}{h}+B\leq 1 \quad \text{or} \quad h\geq\frac{\delta n}{\delta+(1/\beta)-(B/\beta)}$$



## Recycling Ports for Nonblocking Multicast



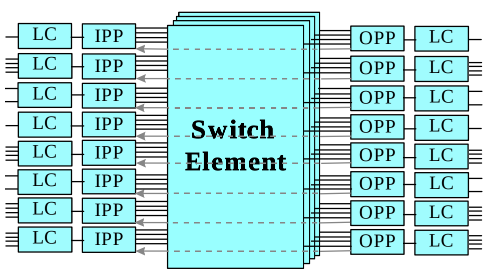

- Moderate number of recycling ports sufficient in cases of most interest.
- Can adjust capacity used for multicast as demands change.
- In systems where external interfaces for single port consume less than switch capacity, "left-over" bandwidth can be used for recycling.

## Interconnection Network Queueing Performance

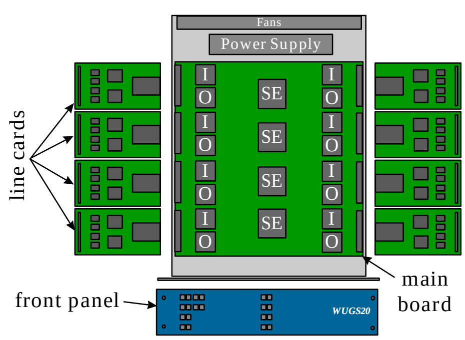
- Queueing performance of buffered multistage networks determined by:
  - » traffic characteristics
  - » type of routing (dynamic or static)
  - » switch element queueing discipline (input, output, shared)
  - » flow control (grant, ack, none)
  - » buffer capacity
  - » switch element and network dimensions
- Large WUGS configurations can support
  - » uniform random (Bernoulli) traffic up to 80% of internal link speed without congestion
  - » uniform random bursty data traffic with peak rates up to about 50 Mb/s and average utilizations of about 60% of internal link speed
  - » under most conditions, system performance determined by output queues
- Bursty data traffic with higher peak rates can lead to congestion between last stage SE and OPP
  - » dynamic routing spreads load, preventing congestion between SEs
  - » can improve performance of bursty data traffic by increasing bandwidth between last stage and OPP



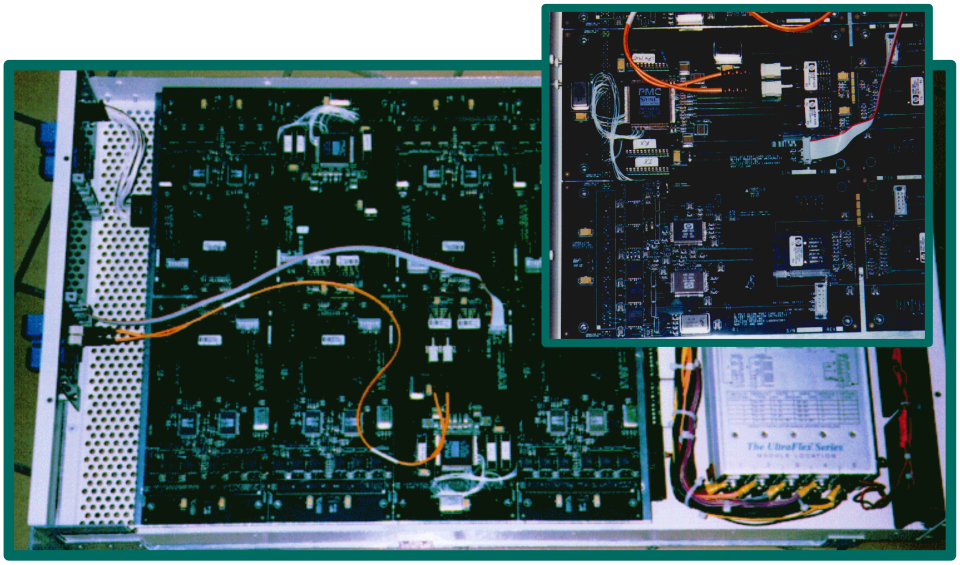
## Prototype Switch




- Eight port configuration
  - » two dual OC-3 line cards
  - » six G-link line cards (1.2 Gb/s)
  - » other line card configurations possible
- Kits will ship with reduced clock rates
  - » 75 MHz gives internal data rates equivalent to 1.48 Gb/s link rates
  - » adjustable for experimental purposes
  - » with standard line cards, still allows up to 50% multicast traffic

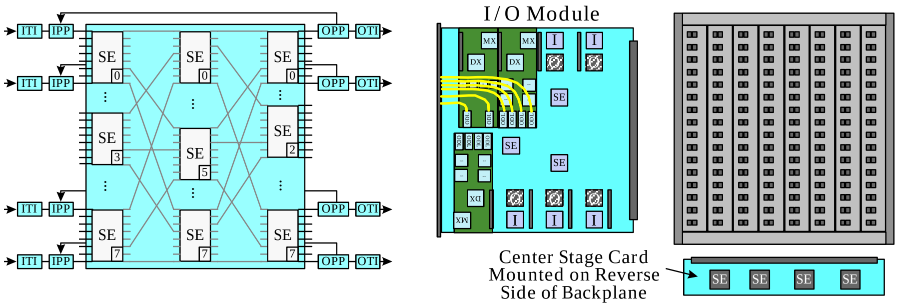


# Prototype Switch Internals





# Planned 160 Gb/s Configuration



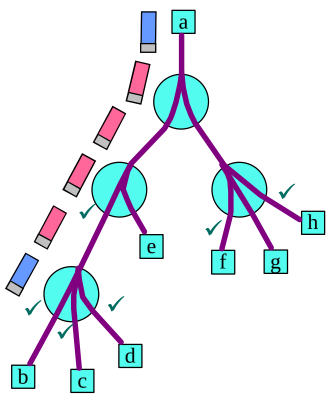

- 8 I/O modules include IPPs, OPPs, line cards, first and third stage SEs
- Horizontal network cards at top and bottom contain middle stage
- Passive midplane interconnects line modules and network cards
- Line cards on I/O modules contain transmission interfaces
  - » quad OC-12 card, dual G-link, OC-48 interfaces
  - » single G-link with FPGAs on input and output for time stamping and timed forwarding
- Will correct speed limits in current chips and provide new features.
- Potential for remote use by kit participants and/or upgrading of kits.

# SE Modifications

- .35 micron technology
- Rework OXBAR layout for symmetry, shorter signal paths.
- Increase total buffer size to 64 cells.
- Modify skew compensation control to hunt when sync. lost; use per bit skew compensation.
- Change pad ring to accommodate new pinout.
- Priority queueing for fast burst setup cell handling.
  - » four priority levels
  - » OXBAR gives strict priority to higher priority cells.
  - » grant line provides three bit code specifying lowest enabled priority class
    - – restrict priority 0 (highest priority) grants when fewer than 16 empty cell slots
    - – restrict priority 1 grants when fewer than 24 empty cell slots
    - – restrict priority 2 grants when fewer than 32 empty cell slots
    - – restrict priority 3 grants when fewer than 40 empty cell slots
  - » OXBAR blocks passage of lower priority cells than allowed by grants.



# Reliable Multicast Support




- IPP modifications provide optional support for redundant ack suppression, for scalable, reliable multicast protocols.
- Packets delineated with start and end cells and sent by source.
- Switches replicate and deliver.
- Receivers send acks.
- Switches discard all but last ack.
- Timeout at source triggers retransmission.
- Retransmitted packets sent only to receivers that need them (*targeted retransmission*).

- VC supports multiple *transmission slots*, allowing pipelining of packets.
  - » maintain ack state for each transmission slot
  - » *W* slots provides support for conventional sliding window protocol with window size of *W* packets
- Many-to-many reliable multicast can be implemented either with relay or *n*-way shared tree.

# Possible Areas for Experimentation

- Performance evaluation
  - » measure system performance under range of traffic conditions
    - – evaluate limitations of internal congestion control mechanisms
    - – evaluate impact of packet discard mechanism on goodput during sustained overloads
    - – assess system's ability to isolate high priority traffic from low priority
  - » end-to-end flow control mechanisms
    - – evaluate rate-based flow control using EFCI mechanism
    - – determine effectiveness of coupling EFCI mechanism to TCP flow control
- Modify line cards to provide new features
  - » line card with microprocessor that can access selected data streams
  - » per VC queueing subsystem for better performance with bursty traffic
  - » UPC mechanism to monitor input traffic and optionally mark/drop
  - » Traffic shaper to regulate flow of output traffic to conform to traffic spec
  - » Fast Ethernet or Gigabit Ethernet interface; IP-over-SONET interface
  - » IP address lookup and packet classification module
- IC modifications
  - » implement rate-based flow control for ABR traffic (including multicast?)
  - » implement VC merging for packet-oriented VCs
  - » adaptive resequencing
  - » switch element that implements distributed shared buffer



# Review Questions

- Explain the remote control mechanism in WUGS. How would you use it to read a VXT entry? How would you use it to measure the rate at which data is being sent on a given virtual circuit? How would you use it to check that all components in a system with a three stage network can pass data correctly?
- In a system with OC48 external links, what clock rate is needed within the switch to exactly match the cell rate of the external links (assume that the external link carry cells at exactly 2.4 Gb/s).
- How does binary replication and recycling work? Why not use three-way copying? Four-way? How do larger branching factors affect routing table requirements? Switch bandwidth requirements? Consider both worst-case and "expected case."
- What's the difference between dynamic routing and static routing? What are the trade-offs between them?
- How does time-based resequencing work? What is the role of the age threshold? How does time-based resequencing affect switch latency? What is transitional timestamping?
- Consider the following situation at one of the G-link outputs of a gigabit kit switch with a 75 MHz clock. The link carries 75 motion JPEG video streams at 15 Mb/s each, plus a bursty data channel which periodically sends 1 MB bursts at 75 Mb/s. Assuming both the video and data are carried as Continuous Stream connections (CS=1), what cell loss rate would you expect to see for the video and data connections, assuming the data channel is sending bursts 10% of the time? (Ignore the impact of end-to-end flow control and error-recovery.) What loss rates would you expect if the peak rate of the data channel was 150 Mb/s? How would you expect the loss rates to change if the data channel were changed to a discrete stream connection (CS=0)? How would you expect the loss rates to change if the peak rate of the data channel was 600 Mb/s?
- In the gigabit kits, with the standard line card configuration and a 75 MHz clock, how big can the multicast fraction ($\delta$) be without introducing the possibility of blocking for connections with bandwidths of 150 Mb/s? What about 25 Mb/s connections? 600 Mb/s connections?
- Explain how the copy-range mechanism can be used with recycling to provide general multicast capabilities. Assuming all switch ports are used for both recycling and external traffic, what speed advantage is needed to ensure that connections with bandwidths of 150 Mb/s will not block in the worst-case? Is this worst-case estimate unrealistcally pessimistic? If so, what speed advantage do you think is needed to make blocking very unlikely? How many table entries must be changed to add or remove an endpoint from a multicast connection in your scheme (consider both the worst-case and the "typical" case}. How many VXT table entries does your approach require, relative to a switch that supports only unicast?
- Explain the difference between virtual paths and circuits. In the WUGS switch, what would you do to set up a virtual path connection (non-terminating VP) from input 2 to output 6, with input VPI=12 and output VPI=25? What would you do to set up a virtual circuit connection from input 2 to output 6 with input VPI/VCI=14/221 and output VPI/VCI=30/50?