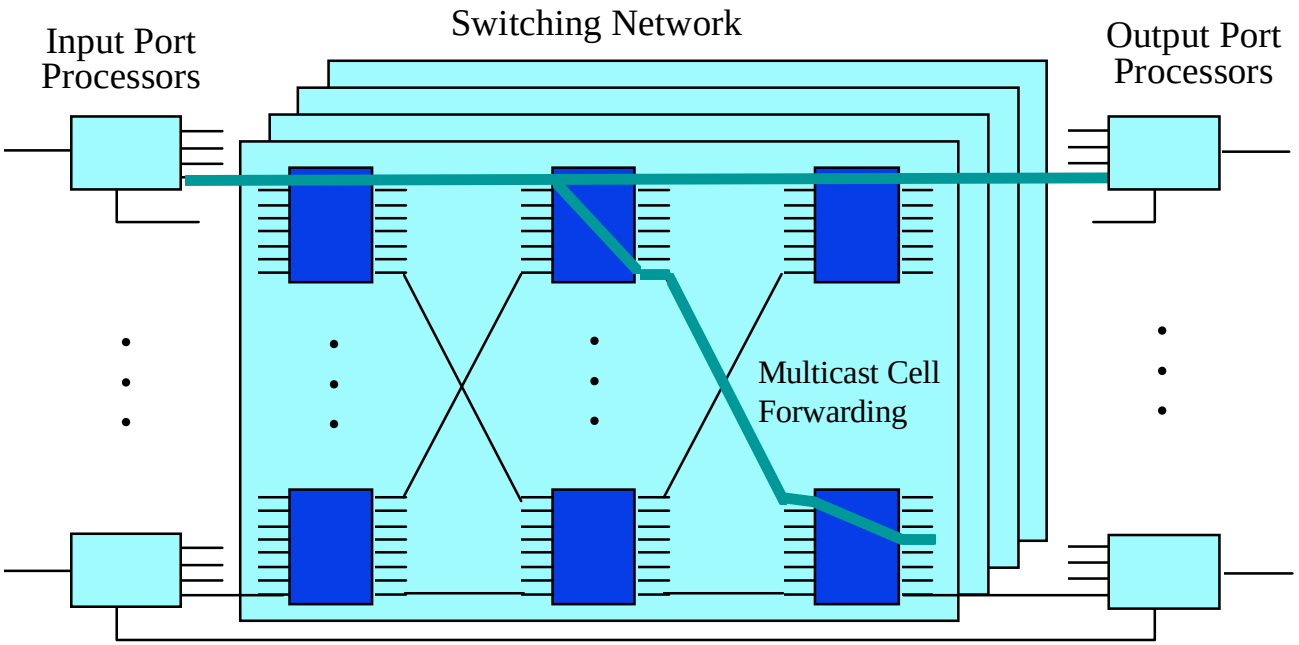
Terabit Switching

Jonathan Turner
Washington University
Computer Science Department

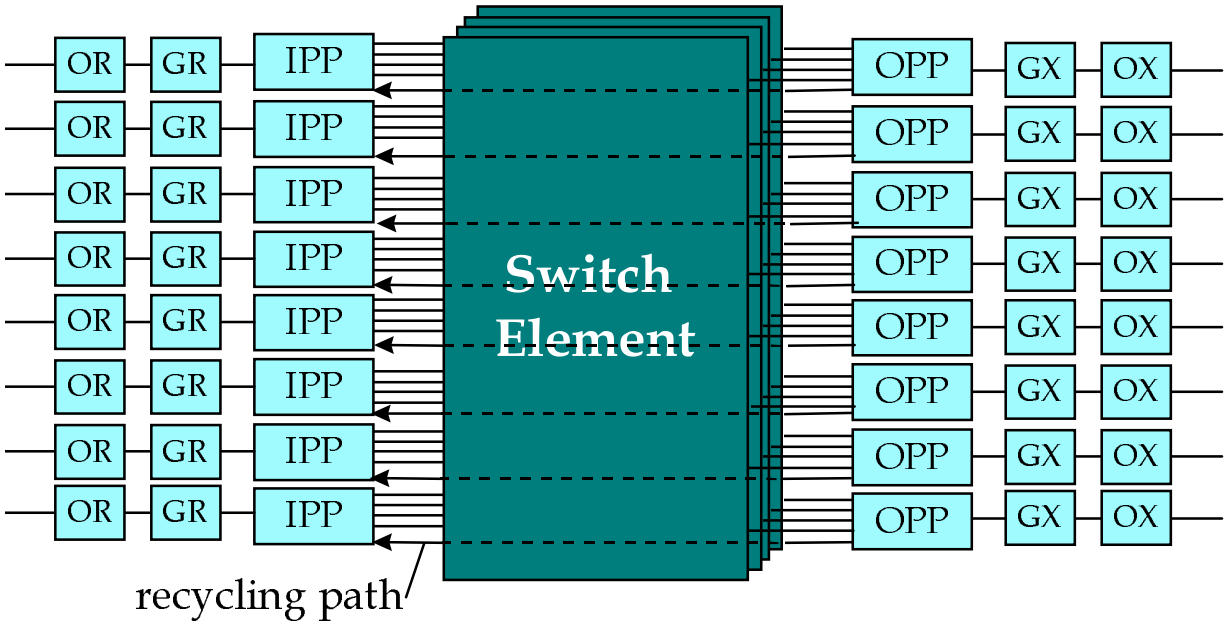
jst@cs.wustl.edu
http://www.arl.wustl.edu/~jst/

Washington University Gigabit Switch



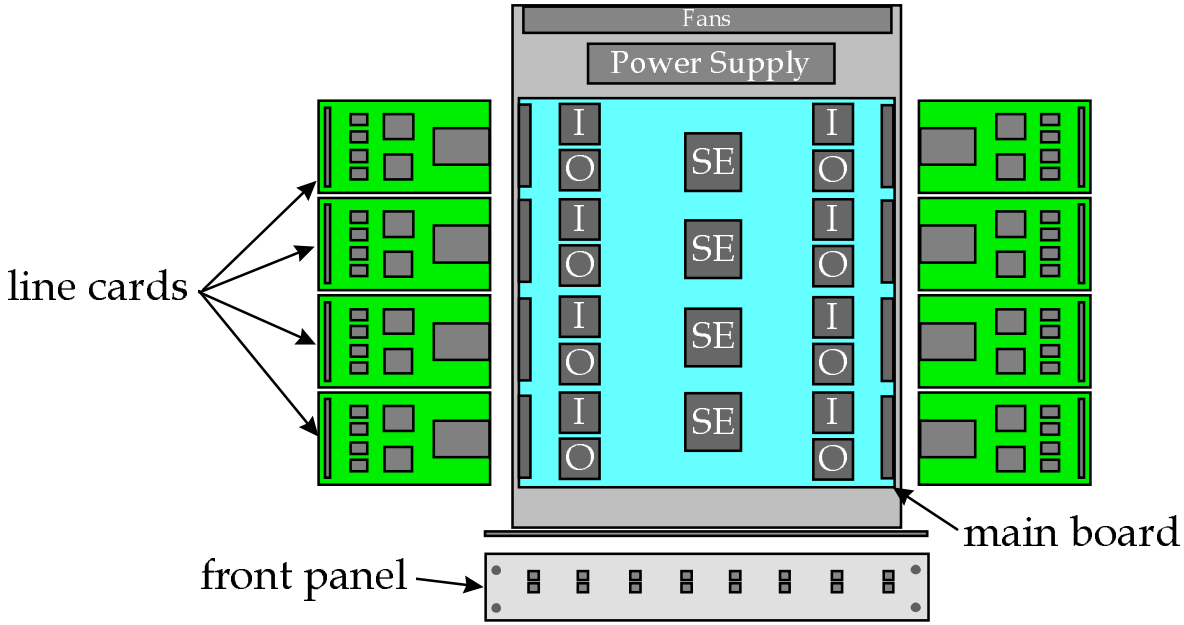
- Multipoint switching using copy-by-two and cell recycling.
- First k–1 stages of 2k–1 stage network do load balancing.
- Nonblocking operation with optimal scaling (nearly flat cost per port for systems with 8 to 4096 ports; 10 Tb/s capacity).
- 32 bit data paths and 120 MHz clock rate; up to 2.4 Gb/s links.
- Extension to 10 Gb/s is feasible in near term.

WUGS Prototype Switch



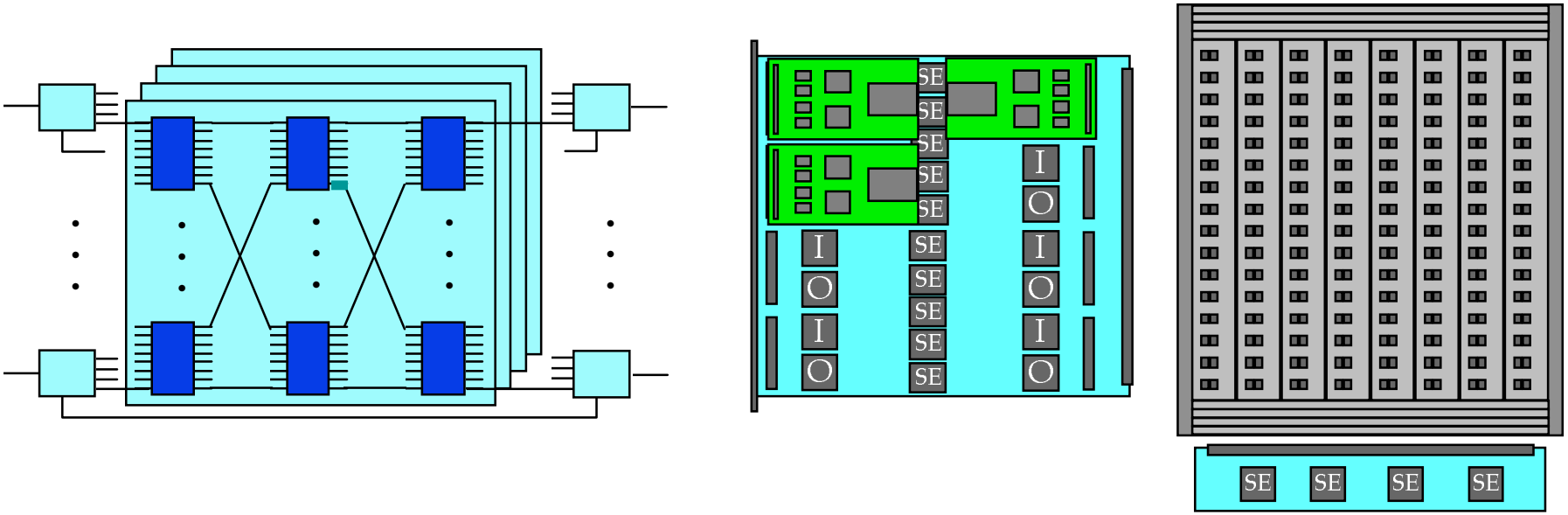
- Eight ports, all fully configurable.
  - » 1.2 Gb/s G-link, OC-3, OC-12, OC-48
- General multicast support (many-to-many); virtual paths and circuits, early packet discard with hysteresis.
- Designed to facilitate additional "out-board" port processing functions.

Prototype System Packaging



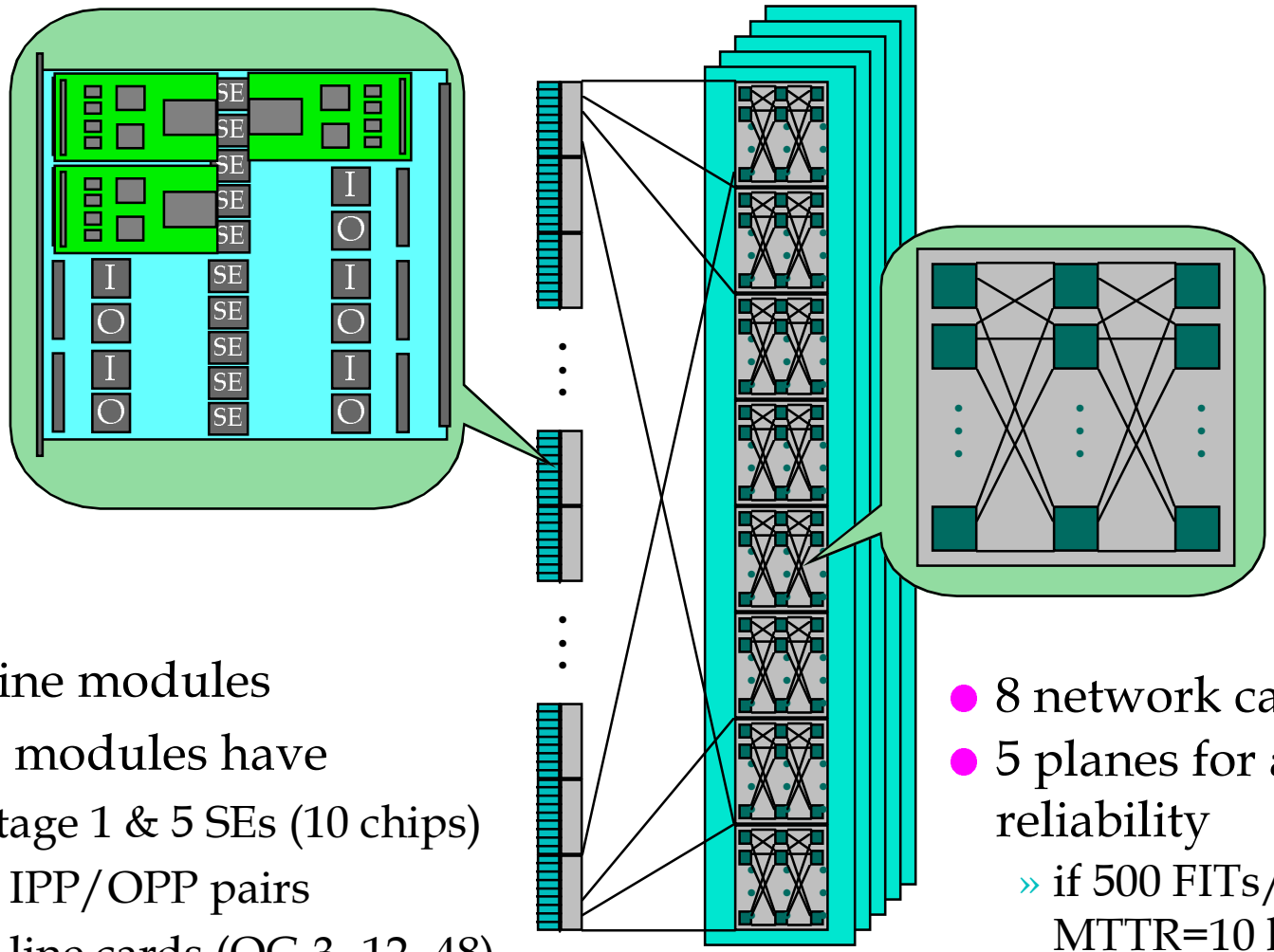
- Rack-mountable, 20 Gb/s system with 8 ports, 14 cm high.
- Pluggable line cards on all ports.
  - » single OC-3, G-link, dual G-link
  - » oversized and/or "double-decker" line cards for OC-12 and OC-48
  - » optical modules on line cards connect to front panel with pigtails
  - » different front panel configurations

160 Gb/s System



- 8 line interface modules include IPPs, OPPs, line cards, first and third stages of network
- horizontal network cards at top and bottom contain middle stage
- backplane interconnects line modules and network cards
- may augment with fifth plane for high reliability

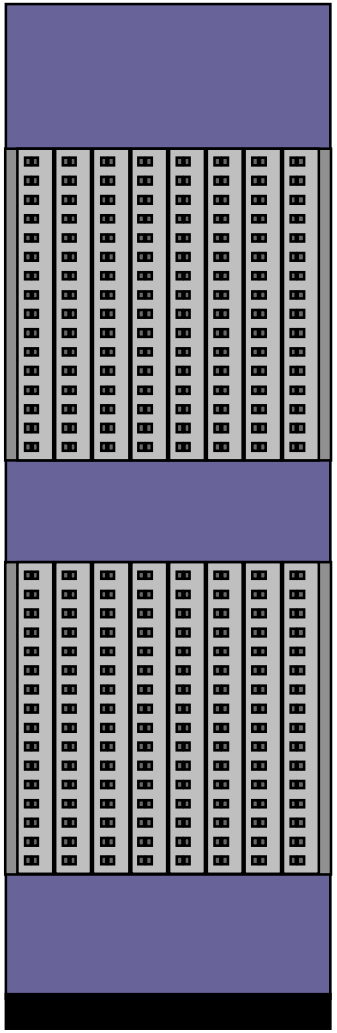
1.2 Tb/s System



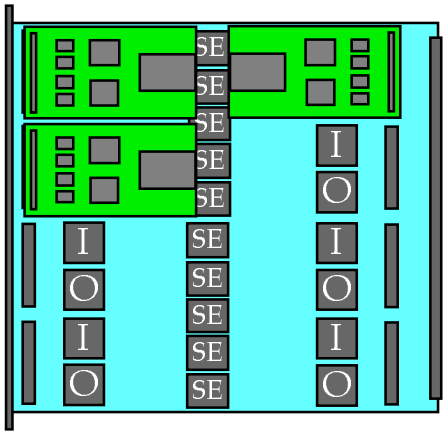
- 64 line modules
- line modules have
  - » stage 1 & 5 SEs (10 chips)
  - » 8 IPP/OPP pairs
  - » 8 line cards (OC-3,-12,-48)

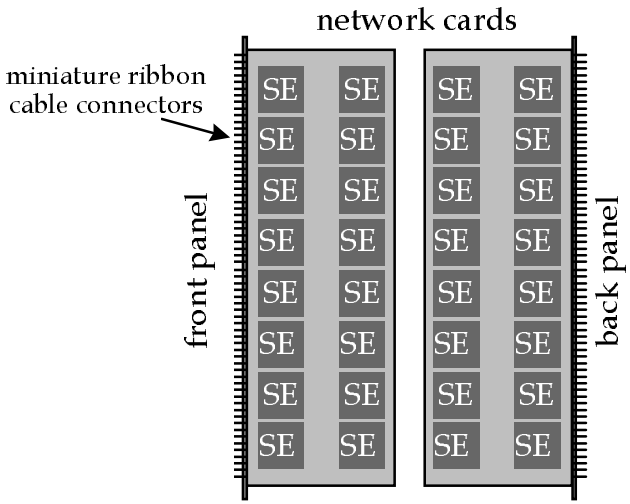
- 8 network cards/plane
- 5 planes for added reliability
  - » if 500 FITs/SE chip & MTTR=10 hrs then MTTF of core is 87 days, MTDF=62 yrs

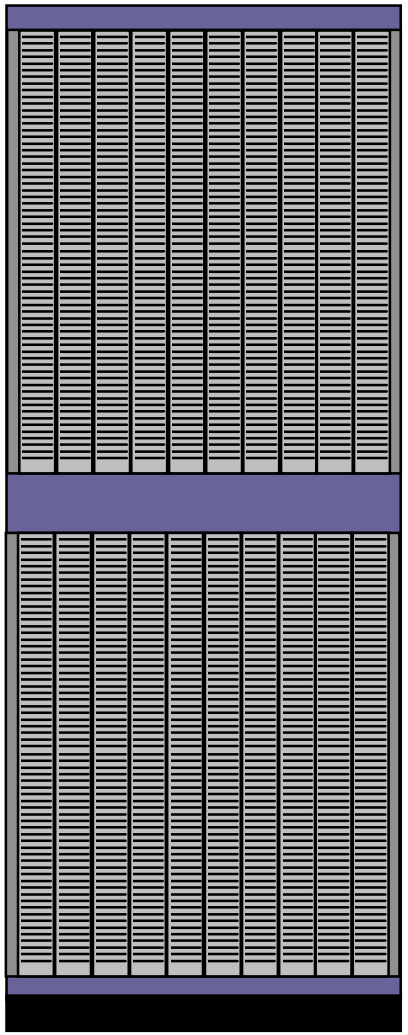
1.2 Tb/s System Packaging



Line Frame (4)

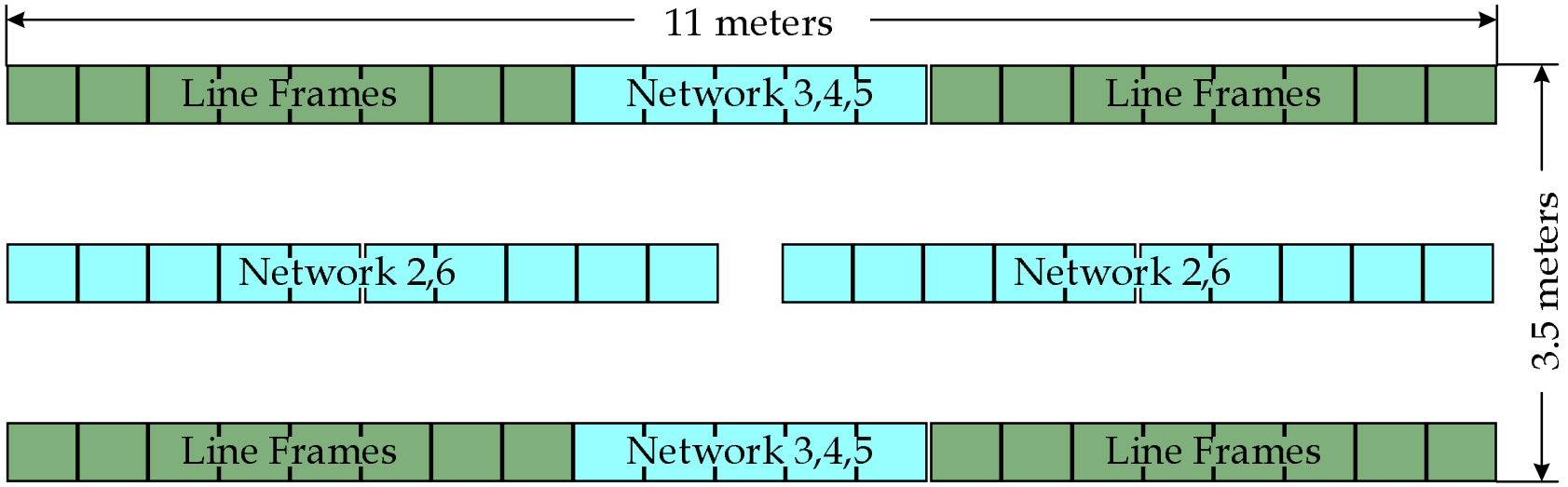






Network Frame (1)

10 Tb/s System

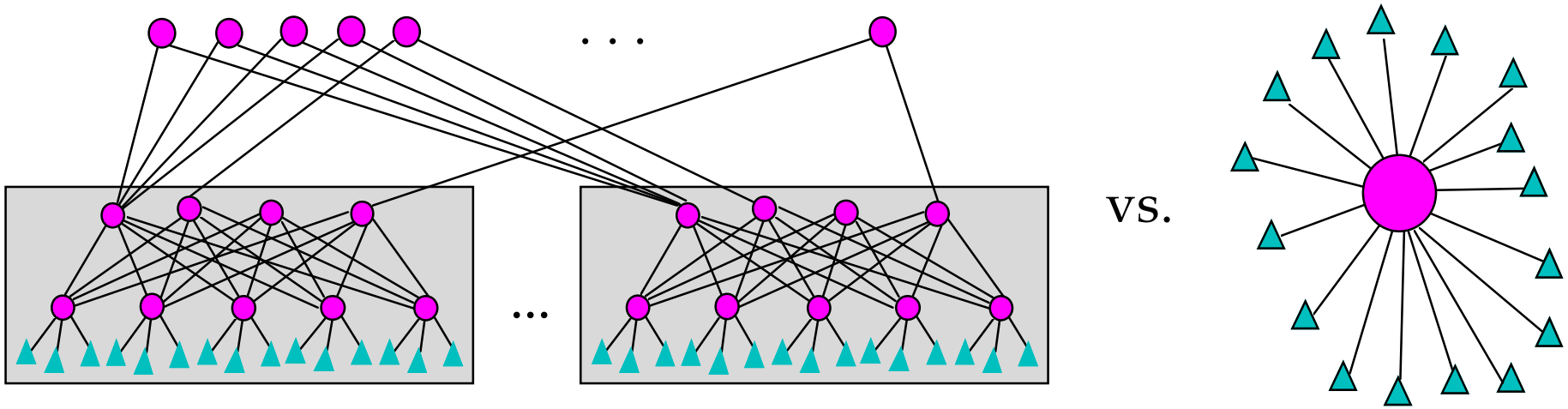


- Similar line and network frames to 1.2 Tb/s system.
- Offset timing between stages to allow for 10 meter interconnect distances.
- Five planes and tolerance of failed center section.
  - » MTTF of center section is 6.5 days, MTDF=126 days, MTTF=62 years

Optimal Scaling Properties

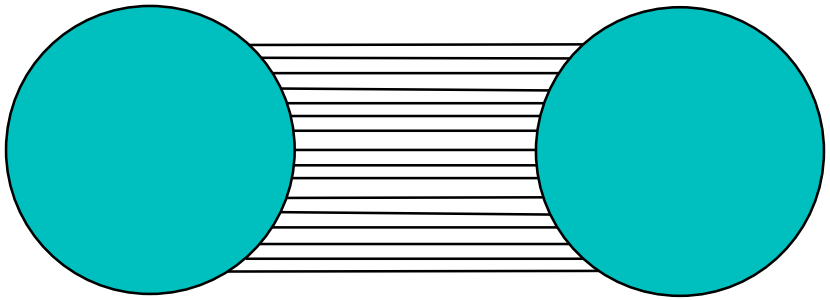
- Switching Network
  - » $n/8$ switch elements per stage, $2\log_8 n - 1$ stages
  - » bandwidth of internal data path is independent of number of ports (does depend on fraction of traffic that is multipoint)
- Routing Table Entries
  - » for m virtual circuits per output link, need mn table entries
  - » multipoint requires larger entries, but no more entries than point-to-point
- Adding and removing virtual circuits
  - » one or two table entries changed per operation
  - » computation time for control processor is constant
- Allows large systems with essentially same cost per port as small systems
  - » 4096 port system has cost per port about 20% more than 8 port system

Big Networks Mean Big Switches

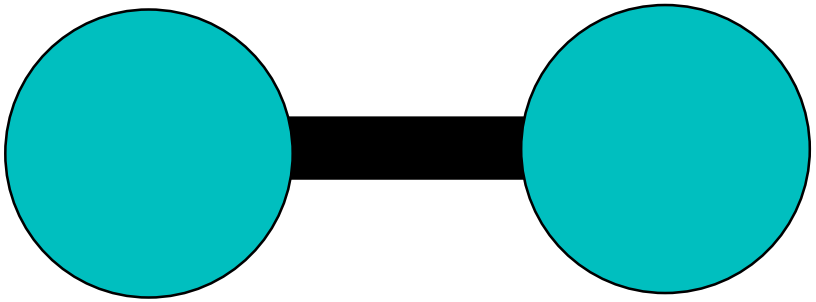


- Dominant cost of well-designed switch is port interfaces.
- Single large switch can be fraction of cost, more reliable, easier to maintain and offer better performance.
  - » Building network for n = 8,192 users with 64 port switches requires 5n switch ports with no concentration, 2.3n with 3:1 concentration.

Big Networks Mean Big Links



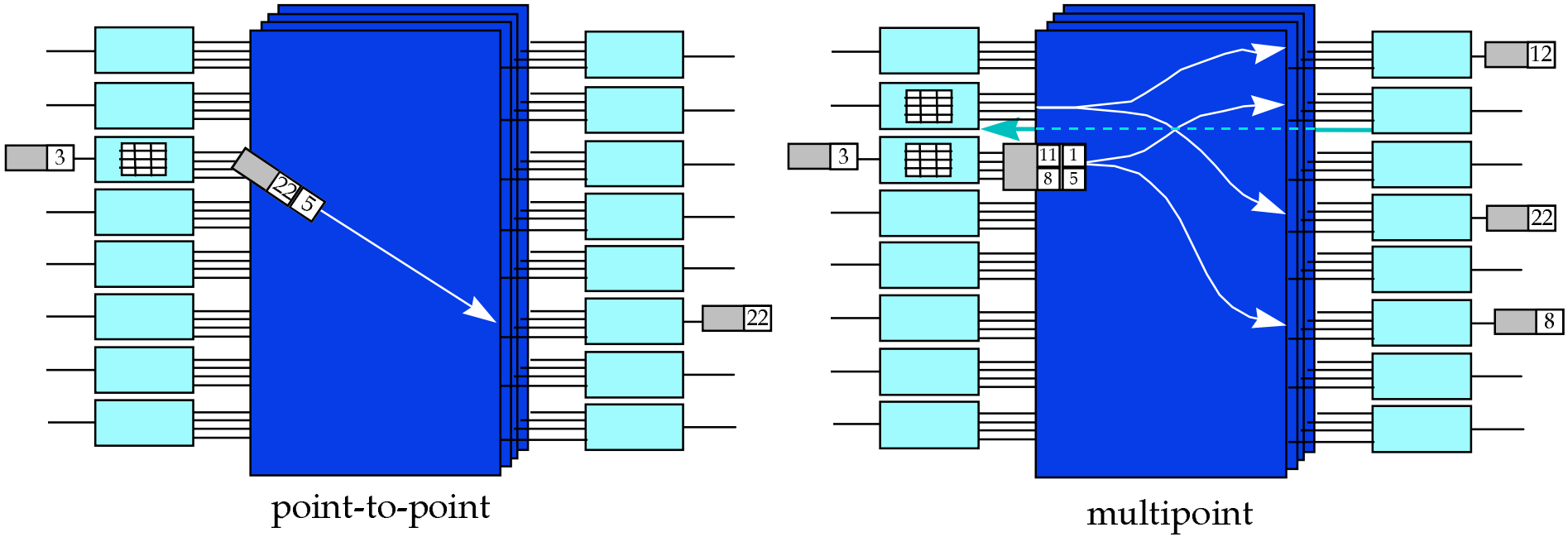
16x150Mb/s



1x2.4 Gb/s

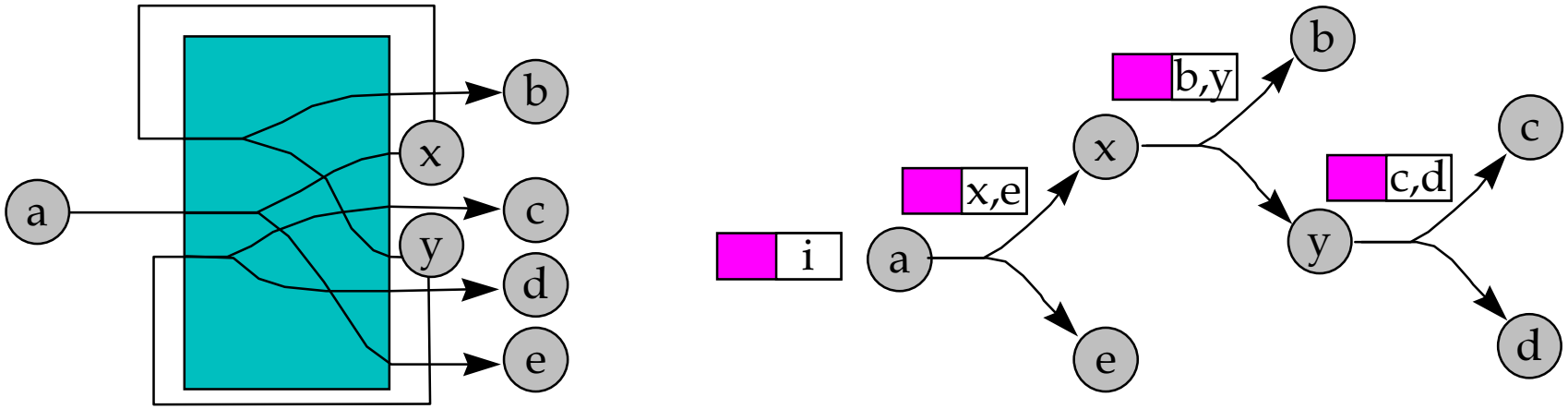
- given bursty traffic with 50 Mb/s peak, 2 Mb/s average
  - » a 150 Mb/s link can handle 12 users with 1% congestion prob.
  - » a 2.4 Gb/s link can handle 841 users with 1% congestion prob.
  - » faster link handles > four times traffic of 16 slower links
- intrinsic termination cost of fast link less than slow links
- fiber savings ($7,500/mile at 10 cents/foot for fiber pair)

Switch Operation



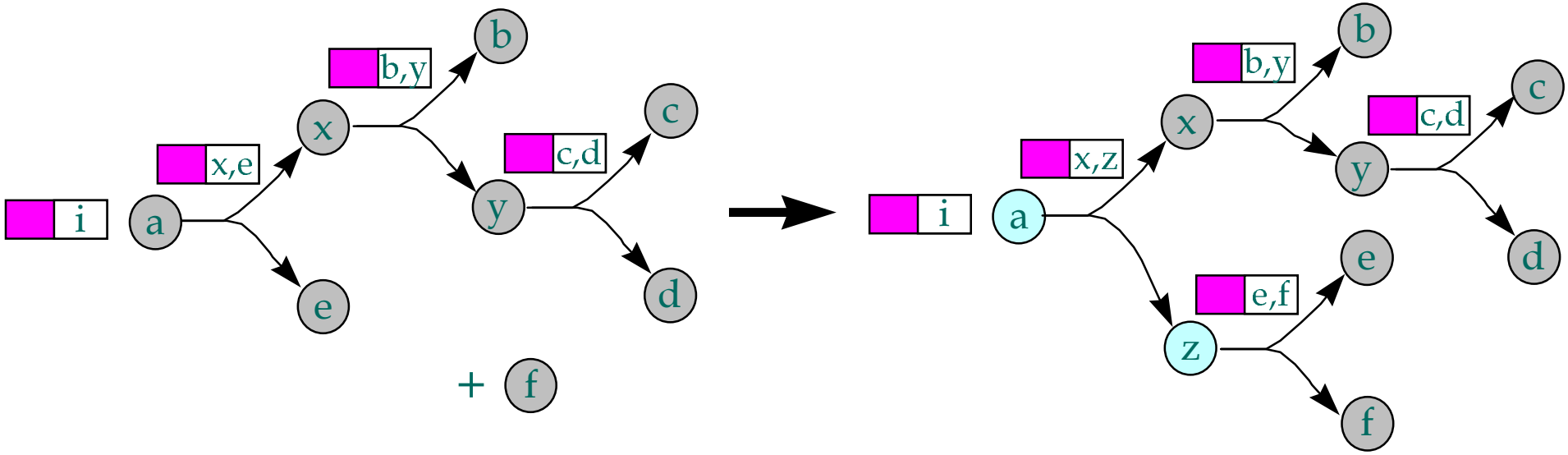
- Virtual Path/Circuit Table (VXT) in IPP holds output and new VCI for point-to-point cells.
- VXT supplies (output, VCI) pairs for multipoint cells.
- Can produce f copies in $\log_2 f$ recycling passes.
- VXT contents established by remote Control Processor during session establishment.

Multipoint Connection Trees



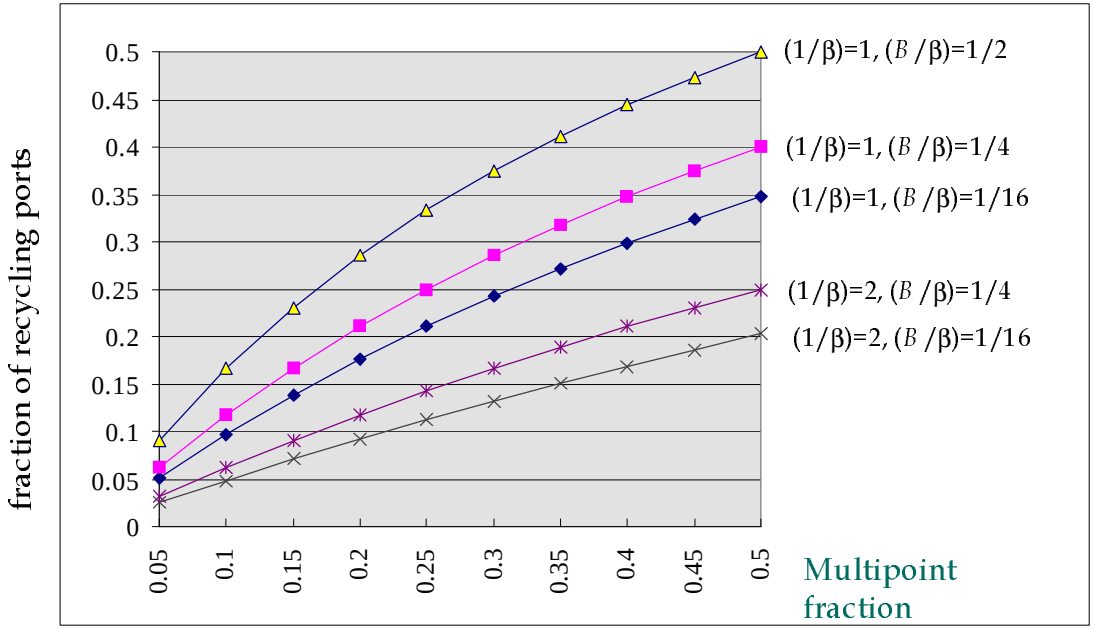
- Each multipoint connections defines a binary tree.
  - » input at tree root
  - » outputs at leaves
  - » internal nodes are ***recycling ports***
- Cells are resequenced after every pass.

Adding an Endpoint



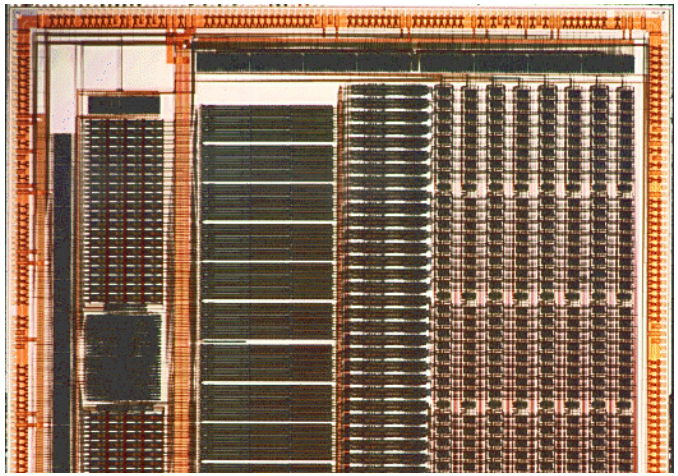
- Pick a shallow leaf *e* and recycling port *z*.
- Make *e* and new leaf *f* children of *z*.
- Make *z* a child of *e*'s former parent.
- Note: two table entries are changed.

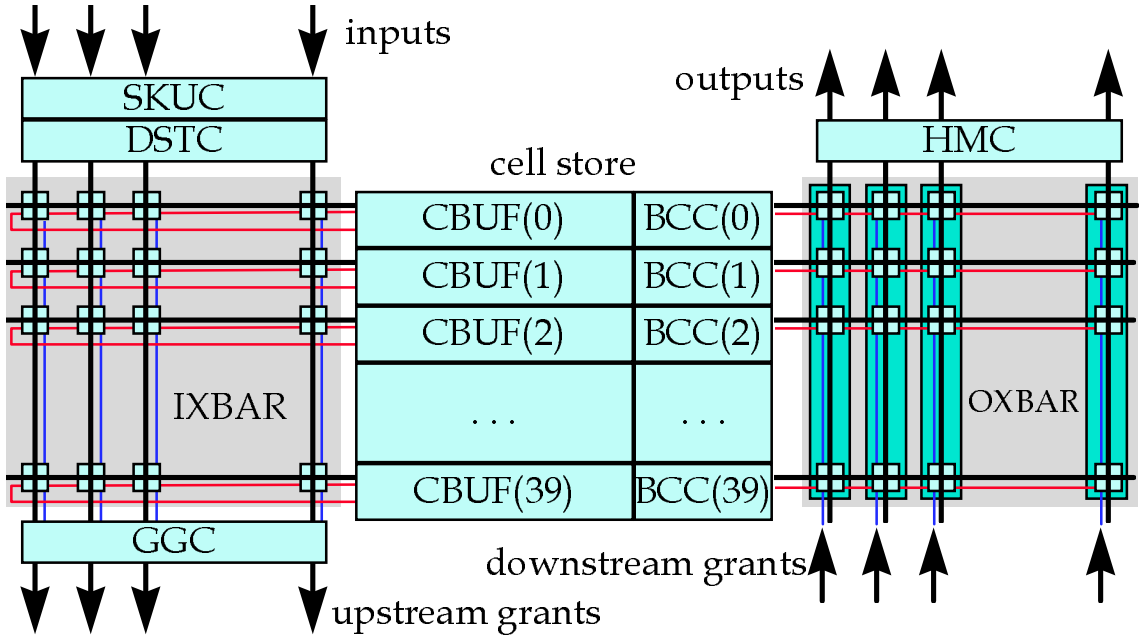
Recycling Ports Needed for Nonblocking



- Moderate speed advantage sufficient in cases of most interest.
- Can adjust capacity used for multipoint as demands change.

Switch Element Component





- eight port switch element with central buffer
- chip is about 1.5 cm per side, 640,000 transistors
- switch element throughput of 25 billion bits per second