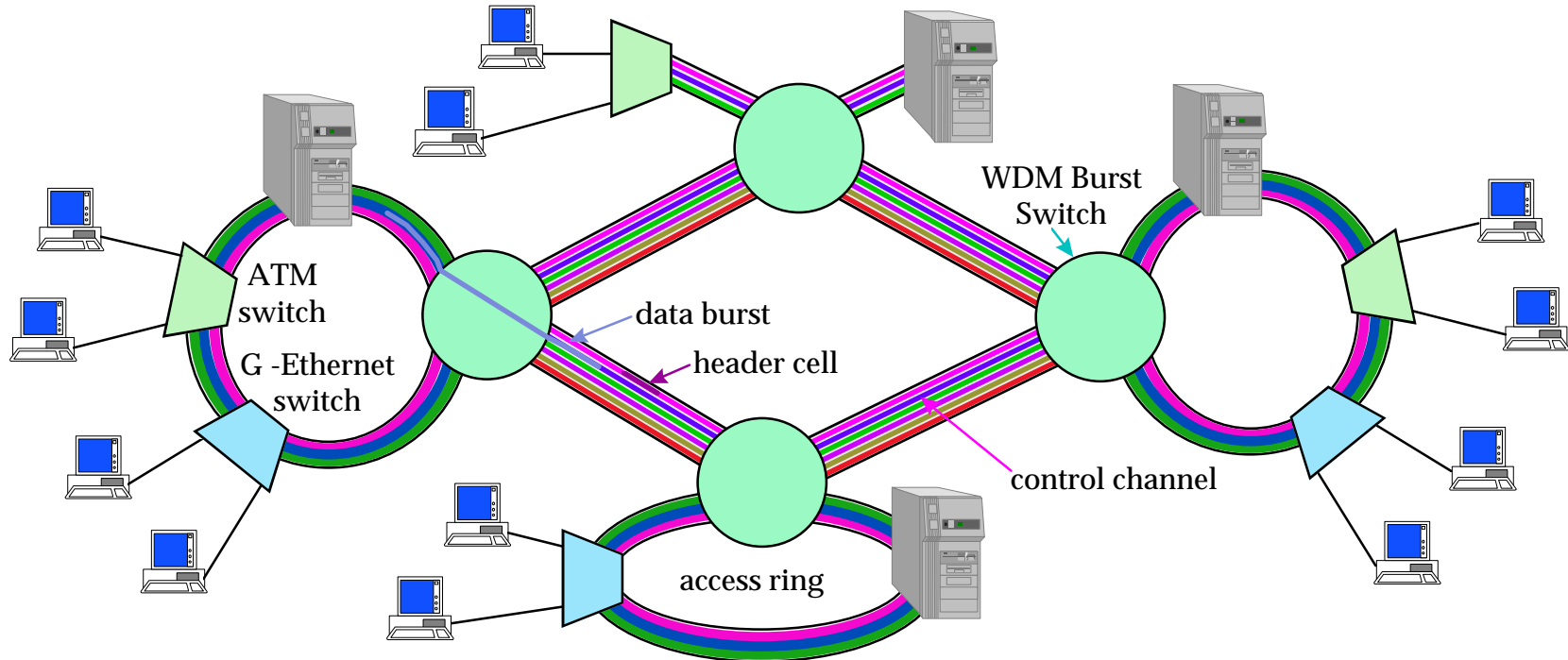

Terabit Burst Switching

Jonathan Turner
Washington University
Computer Science Department

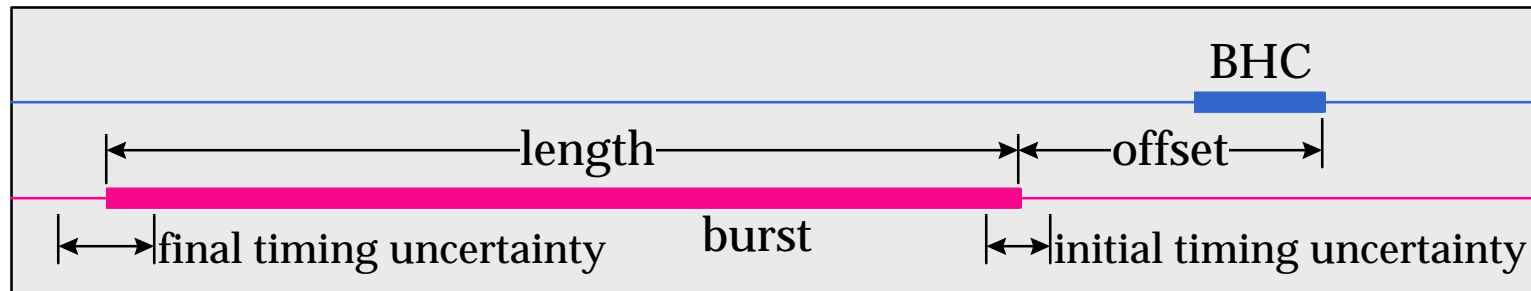
jst@cs.wustl.edu
<http://www.arl.wustl.edu/~jst>

System Concept



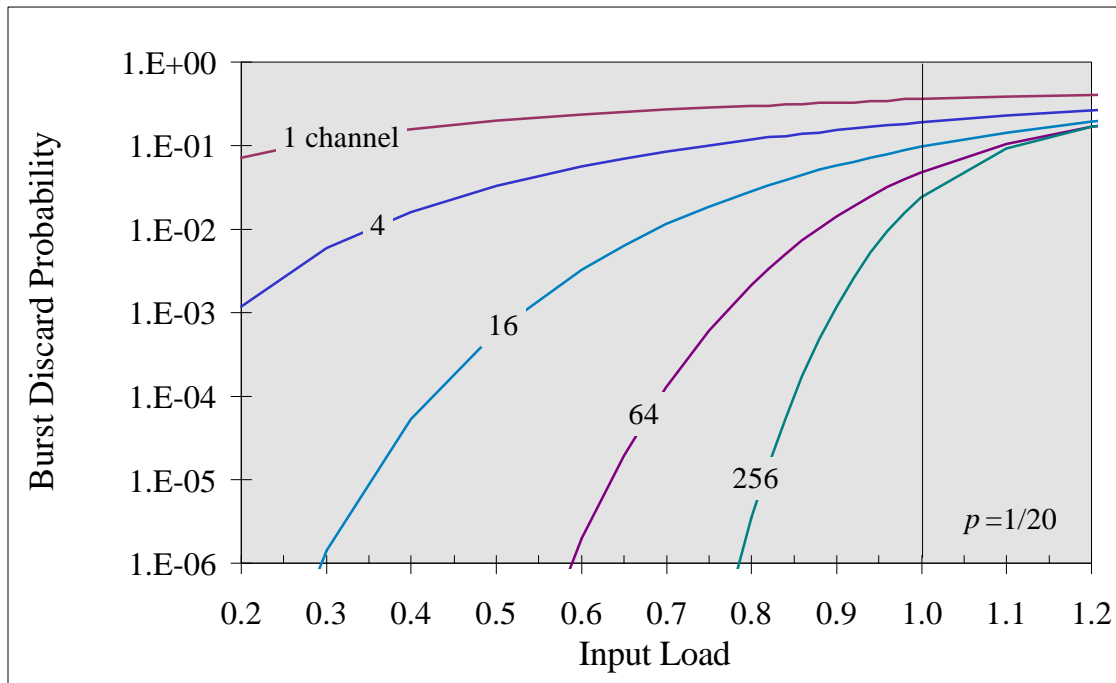
- Send *Burst Header Cell* containing VPI/VCI or IP addr. and burst duration on control channel, followed by burst on free data channel.
- Switches forward BHC and setup switch to route burst using free data channel. Store burst if necessary.
- Resources released when burst duration expires - max duration ≈ 1 ms.
- Access rings allow efficient statistical multiplexing of access facilities.
- At least 32 channels per link, individual channel rates of 2.4-10 Gb/s.

Burst Timing Issues



- bursts transmitted a short time after BHC (e.g. 2 μ s)
- data bursts subjected to fixed delay at every switch (e.g. 10 μ s) to allow for burst header processing - delay can be implemented by WDM fiber delay
- burst header cell delays are matched to fixed data delay
 - » timestamp BHCs on reception
 - » use timestamp information to time forwarding of BHCs on output
 - » forwarded BHC includes explicit offset field to allow compensation of transmit time collision at the output
 - » additionally, input ports include compensation tables to handle wavelength dependent delay variations
- timing uncertainty determines switching overhead for given burst length; with 1 μ s uncertainty, get <10% overhead for >12 KB bursts at 10 Gb/s
- lower overhead possible with explicit burst delineation in data channel

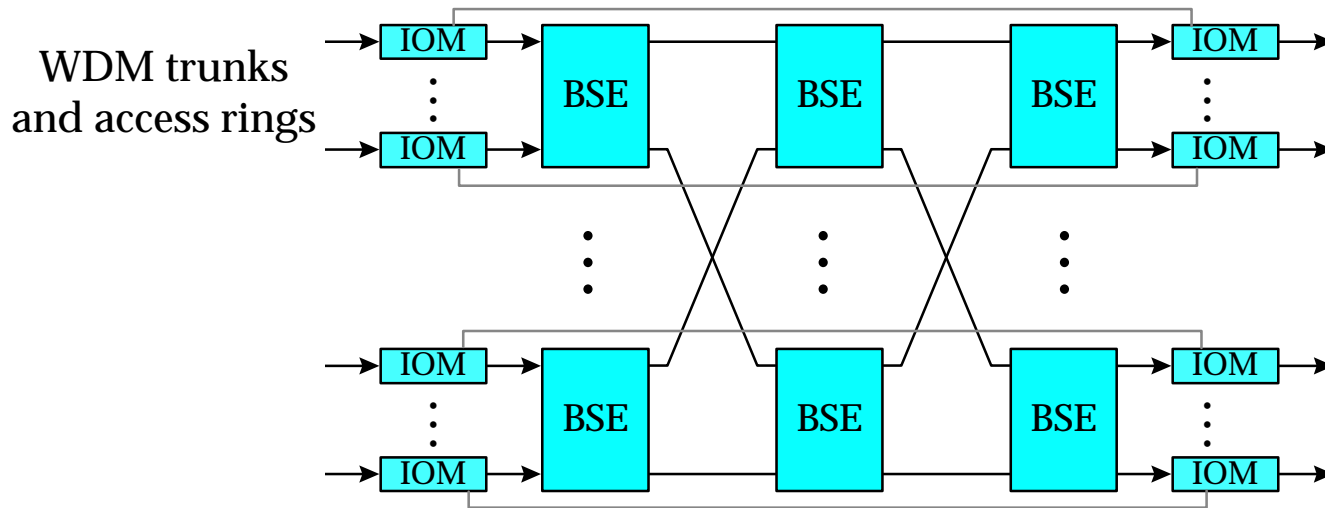
Statistical Multiplexing Issues



- With large numbers of WDM channels per link, can obtain good statistical multiplexing performance with little or no buffering.
 - » with 64 channels, obtain 10^{-6} burst discard probability at link utilizations of $>50\%$
 - » results are insensitive to burst length distribution - self-similarity no problem
 - » yields minimum network delay, since no queueing

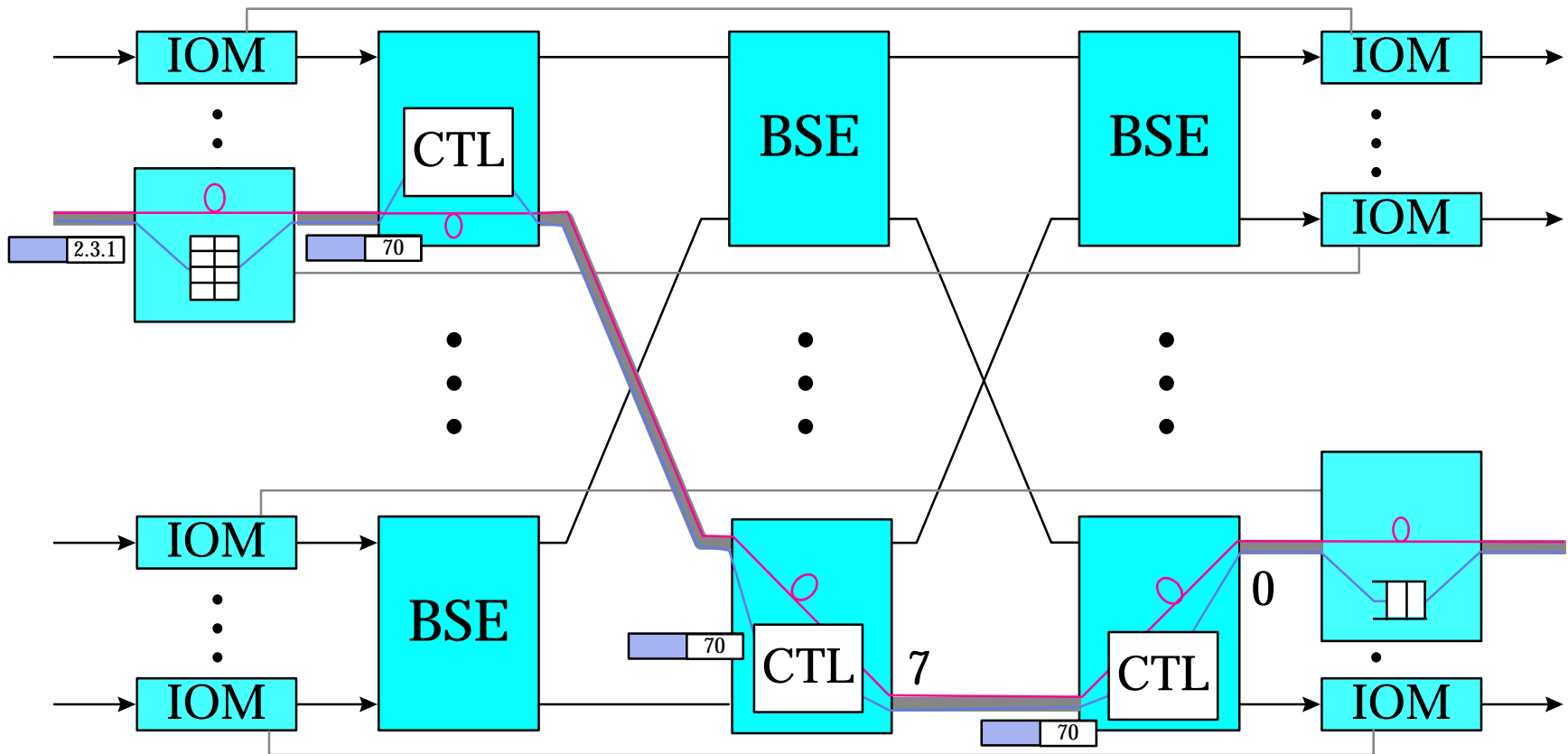
- Queueing can be added to enable higher utilization, multi-channel bursts and links with smaller numbers of channels.
 - » at high speed queueing is expensive (especially when implemented optically)
 - » use shared buffering to reduce impact on overall system cost

Burst Switch Architecture



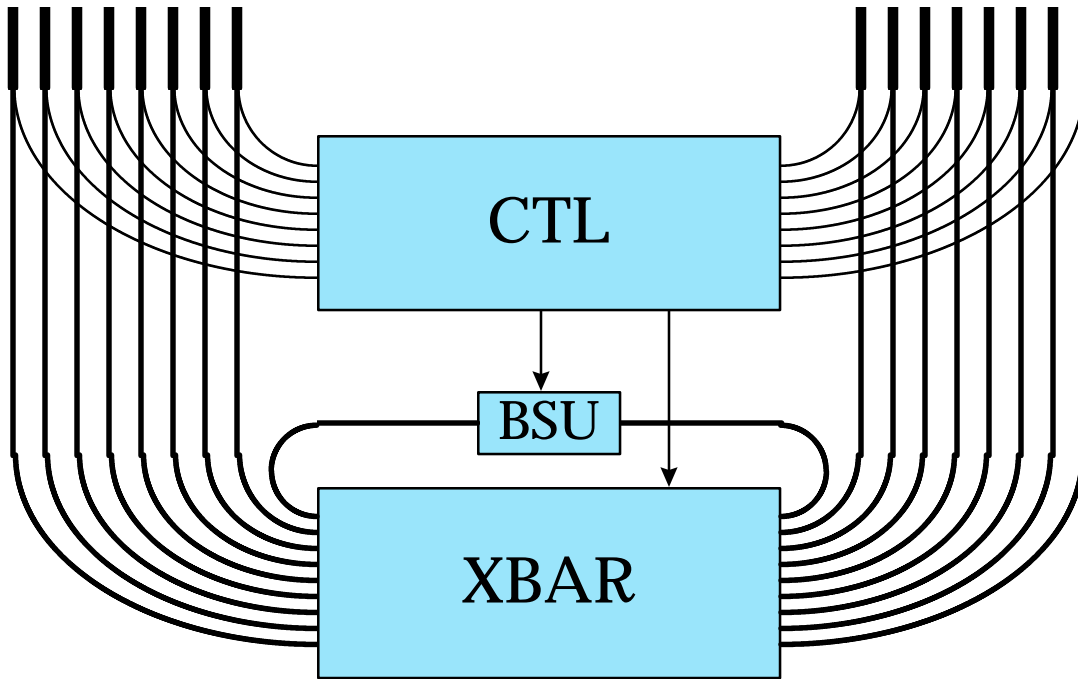
- *Input/Output Modules (IOM)*
 - » process BHCs, perform routing lookup, insert output in control cells
 - » re-synchronize outgoing BHCs to data burst, update time variance field
- *Burst Switch Elements (BSE)* route data bursts using routing information in control cells; dynamic load balancing.
- Simple data path implementable in high speed electronics or (potentially) optics.
- Supports d^k ports with $2k-1$ stages and d port BSEs.
- Multicast uses binary copy and recycling on dedicated recycling ports.

Example Burst Setup



- on input, IOM looks up output port and adds to cell
- BSE's use port number to make local routing decisions
- IOMs re-sync BHCs with data bursts at output

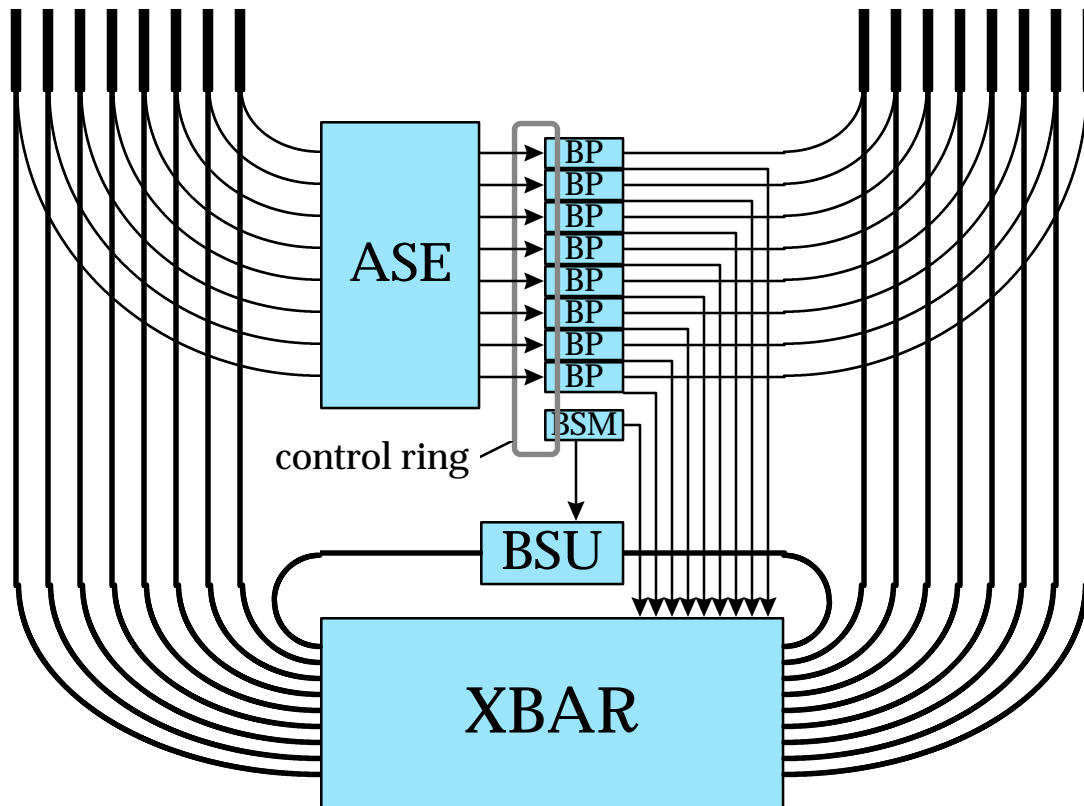
Burst Switch Element



- Control unit (CTL) processes BHCs and configures XBAR and BSU.
- Bursts switched through XBAR from input (port,chan) to output (port,chan).
- Burst Store Unit holds bursts waiting to be forwarded.
- Dimensions: d (7) ports and $h+1$ (32) channels per port -- crossbar dim. $(d+1)h \times (d+1)h$

- In distribution stages, BSE routes bursts to arbitrary outputs.
- In other stages, bursts may be routed or copied (binary or range)
- CTL sends status information to upstream neighbor.
 - » number of idle channels on each output port
 - » number of idle channels entering BS
- BSE may store bursts locally based on downstream neighbor status.

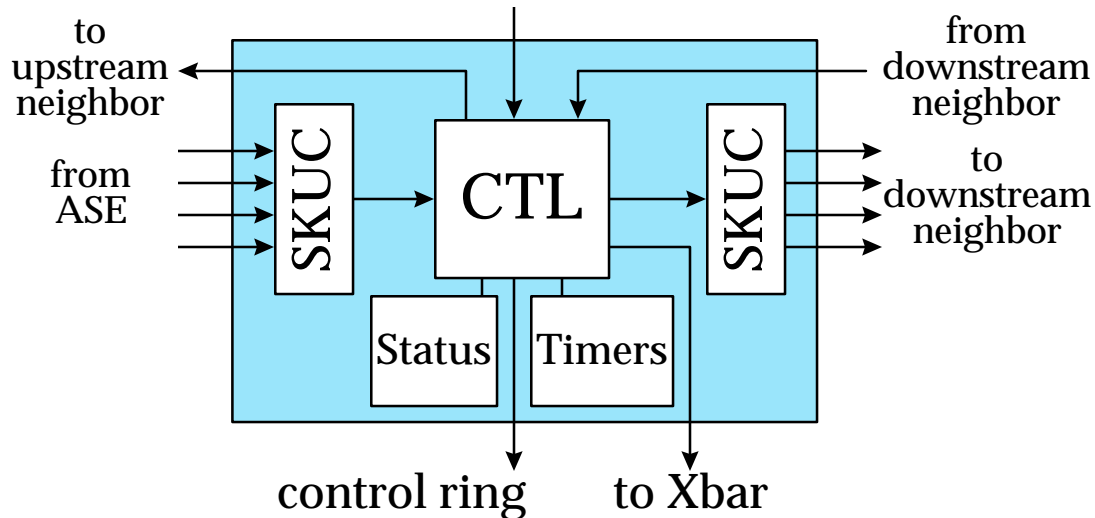
BSE Control



- ATM Switch Element (ASE) routes BSCs to proper output.
- Burst Processors (BP) handle bursts for single output.
- Burst Store Manager (BSM) manages BSU and routing of bursts to BSU.
- BPs communicate status information and storage requests using control ring.

- Each BP receives status information from downstream neighbor.
- Each BP collects status information on entire BSE
- BP_i sends status information for entire BSE to upstream neighbor i .

Burst Processor



● BP status information

- » downstream neighbor status
 - number idle channels per link
 - number idle BS channels
 - updated each cell time
- » BSE status
 - like above, sent each cell time
- » status of channels
 - busy/idle, input <port,channel>, scheduled release time
- » queue of bursts to BSU

● Processing steps for arriving burst from input i , channel j .

- » if idle channel and downstream neighbor can switch or store burst, configure crossbar to switch arriving burst to idle channel; set timer for channel release based on burst length; forward BHC
- » else, send burst-store request to BSM; store BSU channel number and related status in queue of stored requests

● When timer expires indicating channel is available

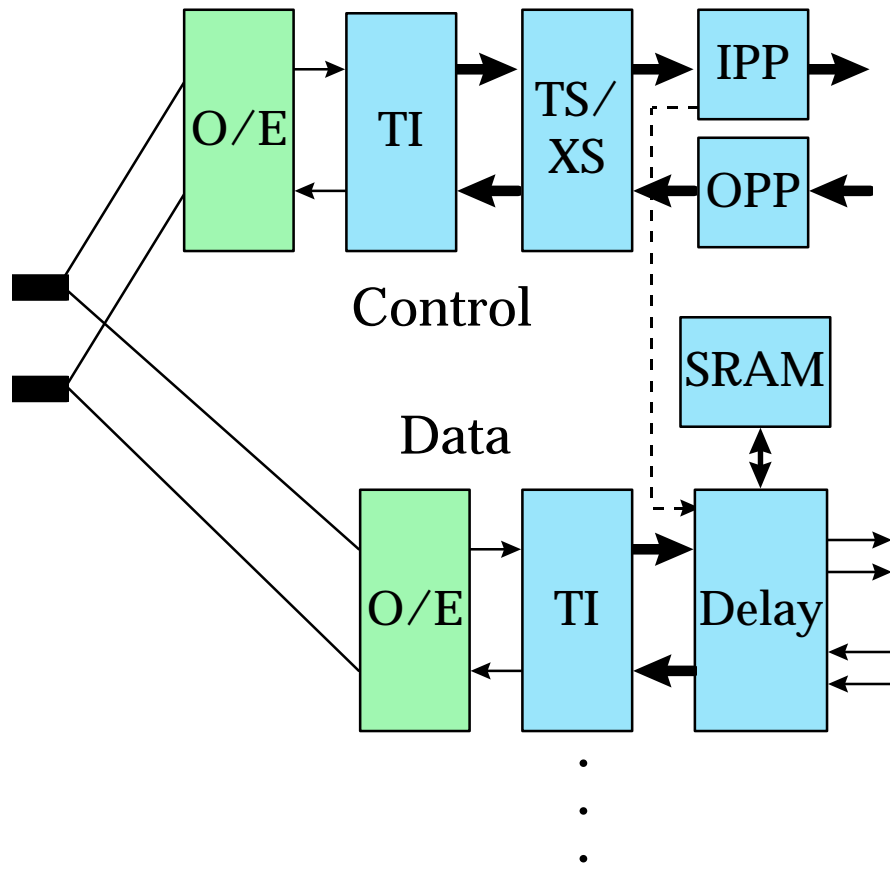
- » assign next waiting burst in BSU (if any), send BHC, configure xbar, update status

● Ensure sequentiality constraint for bursts from common input

Maintaining Burst Sequence

- Time stamp mechanism inadequate for maintaining burst sequence, given potentially long burst delays (1-10 ms possible during overload).
- Requires sequence numbers for each input/output pair (BHCs only).
 - » sequence numbers assigned by ILM
 - » binary copy cells require two sequence numbers (range copy disabled for BHCs)
 - » use time stamp field of BHC to carry sequence numbers (six bits each)
 - » if BP must discard burst, BHC is still forwarded with “delete bit” set
- BP at last stage delivers bursts in order of sequence numbers
 - » maintains sequence number for every input
 - » maintains list of waiting bursts and flushes waiting bursts when “late” burst arrives
 - » releases waiting bursts after timeout (1 second) and sets error flag
 - » rewrites timestamp when forwarding BHC to OPP to avoid resequencer delay
- In single stage configuration, can simplify -- BP just maintains FIFO order for bursts coming from same input
- Note -- order can be lost when restructuring multicast connections
- Alternative -- input-based dynamic routing with “temporally close” bursts constrained to same path (e.g. pick a new path if inter-burst gap of >10 ms).

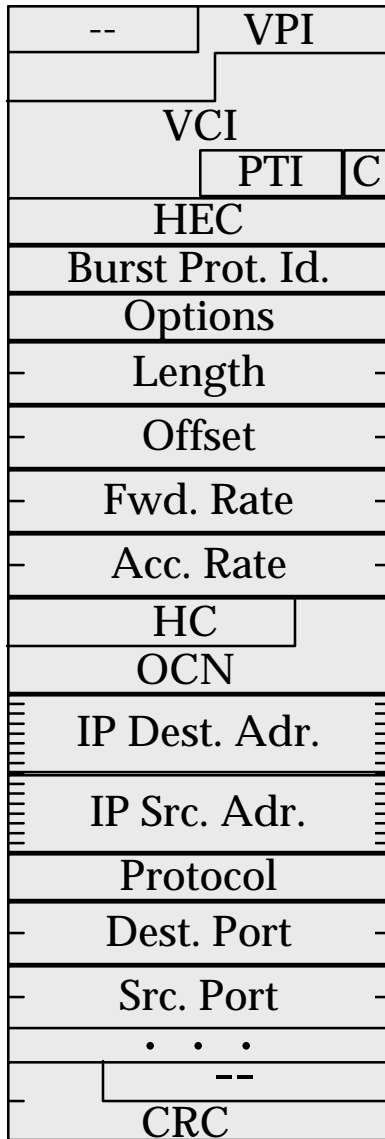
Prototype IO Module



- IO Module includes 1 control and h data sections.
- 2.5 Gb/s link rate, 8B/10B.
- Burst header cells time stamped on input and resynced on output.
- IPP does VCI translation and/or IP lookup.
- Delay chip
 - » delays data by fixed amount
 - » retimes using “idle words”
 - » delay adjustable by IPP up to 20 μ s
 - » forwards data in two bit format at 2.5 Gb/s total
 - » deskews received data separately on each bit

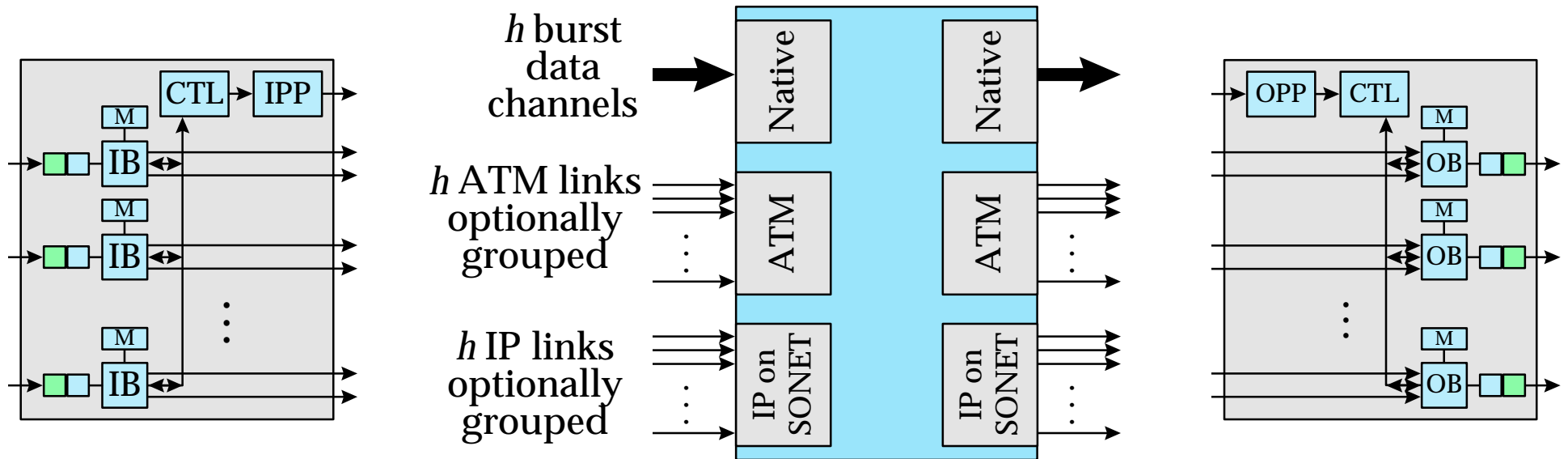
- For $h=31$, IOM requires:
 - » 160 parts (4.8 in²/chip using single two-sided PC board)
 - » 124 D-ECL sigs, 112 CMOS to backplane

Burst Header Cell Format



- Usual ATM cell header with PTI=110 for burst control cells.
- Burst protocol identifier (BPI) identified burst control cells.
- Options (OPT) field distinguishes cell types and options.
- Length (LNG) gives burst length in 16 byte words (1 MB max).
- Offset (OFS) gives number of nanoseconds from first bit of BHC to first bit of burst.
- Hop Count (HC) is allowed number of hops to destination.
- Optical Channel Number (OCN) identifies channel used by data burst.
- Forwarding rate (FR) gives rate at which burst is being sent in Mb/s.
- Access rate (AR) gives rate at which burst was received at entry to burst network.
- IP address (SA, DA), protocol (PRO) and port number (SP, DP) fields required only for IP-switched bursts.
- Address fields contain first eight bytes of IPv6 addresses.
- 13 unused bytes.

Burst Switch Interfaces



- Legacy network interfaces can be individual or grouped.
 - » bursts dynamically distributed over multiple interfaces in a group
 - » on input side, collect cells into bursts (multiple burst formation buffers)
 - » on output side, pace cell transmissions from received bursts (multiple buffers)
- Implications for burst switch core
 - » when routing to legacy interface, BHC must specify link or group to route to
 - » last stage switch element must route burst to specified link or group
 - » when routing to interface with small group size, more buffering needed to achieve acceptable burst loss rates

Summary

- Burst switching can effectively exploit high performance technologies to yield order-of-magnitude performance gains
 - » control mechanisms can handle tens of multigigabit channels per link and hundreds of links per systems
 - » control can be efficient for bursts as short as 1-10 KB
 - » excellent statistical multiplexing performance achievable with tens of channels per link and little or no buffering
 - » shared queueing provides additional performance gains without major cost impact
 - » system using electronic data path can provide 400 Gb/s of switching capacity in ten PC boards
- Immaturity of optical component technology currently makes optical data path implementation cost-prohibitive, relative to electronics
 - » to compete with electronics, need single board $(8,32) \times (8,32)$ optical WDM switch with full wavelength conversion and (<100 ns) switching times
 - » for high bit rate channels will also need optical digital signal regenerators (with retiming) and optical phase aligners with fast phase acquisition (<100 ns)