# *Labeled Optical Burst Switching and IP/WDM Integration*

Chunming Qiao

# *OVERVIEW*

– Introduction to IP/WDM

– Optical Switching Paradigms

  Circuit or Packet Switching?

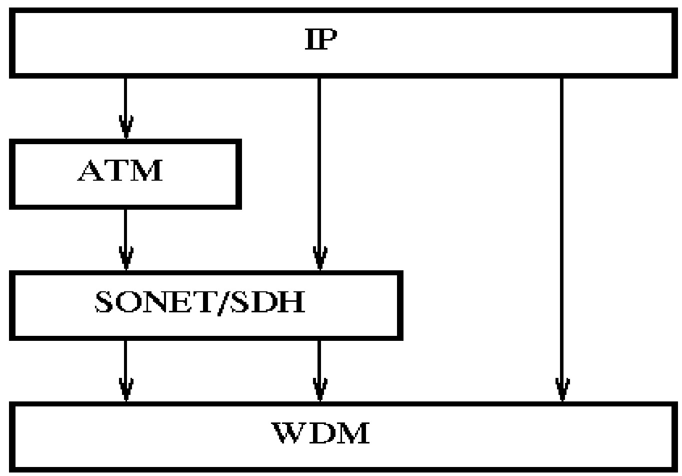
– Optical Burst Switching (OBS)

# *Just In Case ...*

- IP:  Internet Protocol
  - not *Intellectual Property*
- ATM: Asynchronous Transfer Mode
  - not *Automatic Teller Machine*
- SONET: Synchronous Optical NETwork
  - not as in *son et (lumiere)*
- WDM: Wavelength Division Multiplexing
  - or *Wha'Daya  Mean* ?

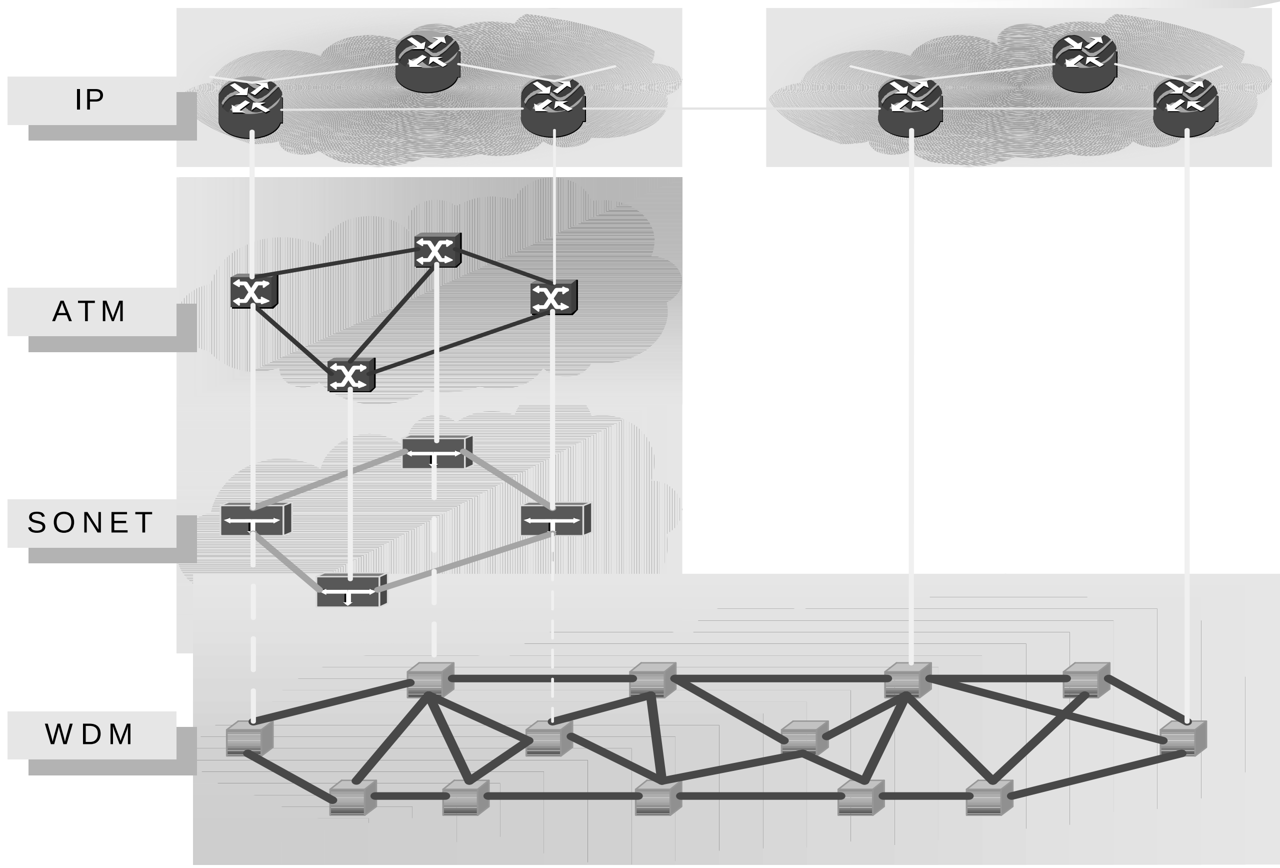
# *Network Architectures*

- today: IP over (ATM/SONET) over WDM
- trend: Integrated IP/WDM (with optical switching)




- goal: ubiquitous, scalable and future-proof

# *IP / ATM / SONET / WDM*



# *SONET/SDH*

- standard for TDM transmissions over fibers
  - basic rate of OC-3 (155 Mbps) based on 64 kbps PCM channels (primarily voice traffic)
  - expensive electronic Add-Drop Muxers (ADM) @ OC-192 (or 10 Gbps) and above
  - many functions *not* necessary/meaningful for data traffic (e.g., bidirectional/symmetric links)
  - use predominantly rings: not BW efficient, but quick protection/restoration (<= 50 *ms*)

# *Internet Protocol (IP)*

- main functions
  - break data (email, file) into (IP) packets
  - add network (IP) addresses to each packet
  - figure out the (current) topology and maintains a routing table at each router
  - find a match for the destination address of a packet, and forward it to the next hop
    - a link to a popular server site may be congested

# *Asynchronous Transfer Mode*

- break data (e.g., an IP packet) into smaller ATM cells, each having 48+5 = 53 bytes
- a route from point A to point B needs be pre-established before sending cells.
- support Quality-of-Service (QoS), e.g., bounded delay, jitter and cell loss rate
- basic rate: between 155 and 622 Mbps
  – just start to talk 10 Gbps  (too late?)

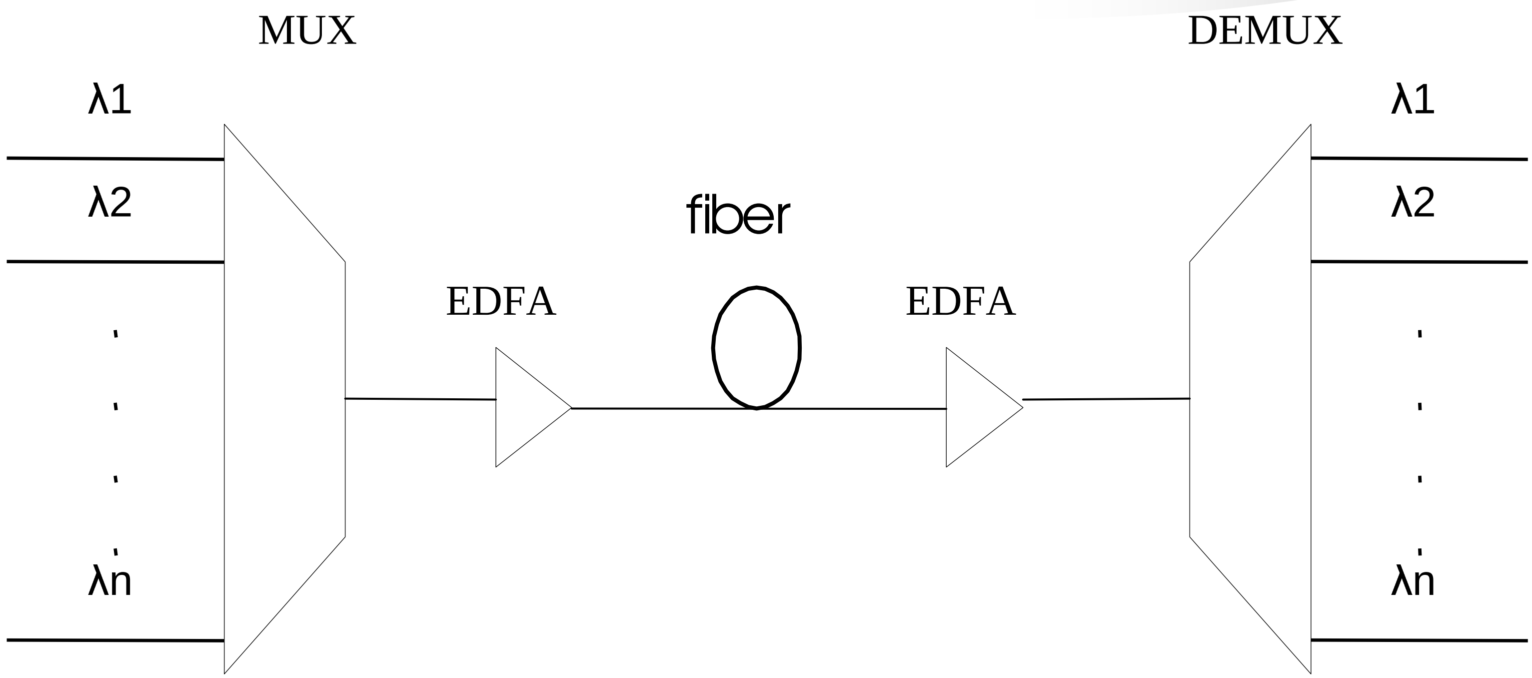
# *Data Traffic Growth*

- double every 4 (up to 12) months or so, and will increase by 1,000 times in 5 years
  - at least 10 x increase in users, and uses per user
  - at least 100 x increase in BW per use:
    - current web pages contain 10 KB each
    - *MP3* & *MPEG* files are 5 & 40 MB each, resp.
- beat Moore’s Law (growth rate in electronic processing power)
  - electronic processing, switching, and transmission cannot and will not keep up
  - need WDM transmissions *and* switching

# *Wavelength Division Multiplex*

- up to 50 THz (or about 50 Tbps) per fiber (low loss range is now 1335*nm* to 1625*nm*)
- mature WDM components
  - mux/demux, amplifier (EDFA), transceiver (fixed-tuned), add-drop mux, static λ-router,
- still developing
  - tunable transceiver, all-optical λ-conversion and cross-connect/switches, Raman amplifiers

# *WDM Pt-2-Pt Transmission*



# *Advance in WDM Networking*

- Transmission (long haul)
  - 80 λs (1530*nm* to 1565*nm*) now, and additional 80 λs (1570*nm* to 1610*nm*) soon
  - OC-48 (2.5 Gbps) per λ (separated by 0.4 *nm)* and OC-192 (separated by 0.8 *nm*)
  - 40 Gbps per λ also coming (>1 Tbps per fiber)
- Cross-connecting and Switching
  - Up to 1000 x 1000 optical cross-connects (MEMS)
  - 64 x 64 packet-switches (switching time < 1 *ns*)

# *ATM and SONET: Legacy*

- interest in ATM diminished
  - a high *cell* tax, and segmentation/re-assembly and signaling overhead
  - failed to reach desktops (& take over the world)
  - on-going effort in providing QoS by IP (e.g., *IPv6* & *Multi-protocol Label Switching* or MPLS)
- SONET/SDH more expensive than WDM
  - & IP & WDM can jointly provide *satisfactory* protection/restoration (< 99.999% reliability?)

# *Datagram (IP) or VC (ATM)*

- *datagram*-based packet switching
  - next-hop determined for each packet based on *destination* address and (*current*) routing table
    - IP finds a longest sub-string match (a complex op)
- *virtual circuit* (VC)-based packet-switching
  - determines the path (*VC*) to take before-hand
    - entry at each node: [*VCI -in*, next-hop, *VCI-out*]
  - assigns packets a VCI (e.g., Rt. 66 )

# *Benefit of VC (as in ATM)*

- faster and more efficient forwarding
  - an exact match is quicker to find than a longest sub-string match
- facilitates traffic engineering
  - paths can be explicitly specified for achieving e.g., network-wide load-balance
  - packets with the same destination address (but different *VCI's)* can *now* be treated differently

# *IP-over-ATM*

- IP routers interconnected via ATM switches
- breaks each packet into cells for switching
- a flow: consecutive packets with the same source/destination (domain/host/TCP conn.)
- Multi-protocol over ATM (MPOA)
  - ATM-specific signaling to establish an ATM VC between source/destination IP routers
  - segmentation and re-assembly overhead

# *IP-centric Control*

- Tag Switching (centralized,  control-driven)
  - the network sets up end-to-end VC's
  - each packet carries a tag (e.g., VCI)
- IP Switching (distributed, data-driven)
  - first few packets are routed at every IP router
    - up to a threshold value to filter out short "flows"
  - following packets bypass intermediate routers via a VC (established in a hop-by-hop fashion).

# *MPLS (Overview)*

- A control plane integrating network-layer (routing) and data-link layer (switching)
  - packet-switched networks with VC's
- LSP: label switched path (VC's)
  - identified with a sequence of labels (tag/VCI)
  - set up between label switched routers (LSRs)
- Each packet is augmented with a *shim* containing a label, and switched over a LSP

# *IP over WDM Architectures*

- IP routers interconnected with WDM links
  - with or without built-in WDM transceivers
- An optical cloud (core) accessed by IP routers at the edge
  - pros: provide fat and easy-to-provision pipes
  - either transparent (i.e., OOO) or opaque (i.e., O-E-O) cross-connects (circuit-switches)
  - proprietary control and non-IP based routing

# *Optical/Photonic (OOO) Switching*

- Pros:
  - can handle a huge amount of *through*-traffic
  - synergetic to optical transmission (no O/E/O)
  - transparency (bit-rate, format, protocol)
- caveats
  - optical 3R/performance monitoring are hard
  - more mature/reliable opaque (OEO) switches
  - SONET or GbE like framing still useful

# *Emerging Integrated IP/WDM*

- IP and *MPLS* on top of every optical circuit or *packet* switch :
  - IP-based addressing/routing (electronics), but data is optically switched (circuit or packet)
  - MPLS-based provisioning, traffic engineering and protection/restoration
  - internetworking of optical WDM subnets
    - with interior and exterior (border) gateway routing

# *Why IP over WDM*

- IP: the unifying/convergence network layer
- IP traffic is (& will remain) *predominant*
  - annual % increase in voice traffic is in the teens
- IP/WDM the choice if start from scratch
  - ATM/SONET were primarily for voice traffic
  - should optimize for pre-dominant IP traffic
- IP routers' port speed reaches OC-48
  - no need for multiplexing by ATM/SONET

# *Why IP/WDM (continued)*

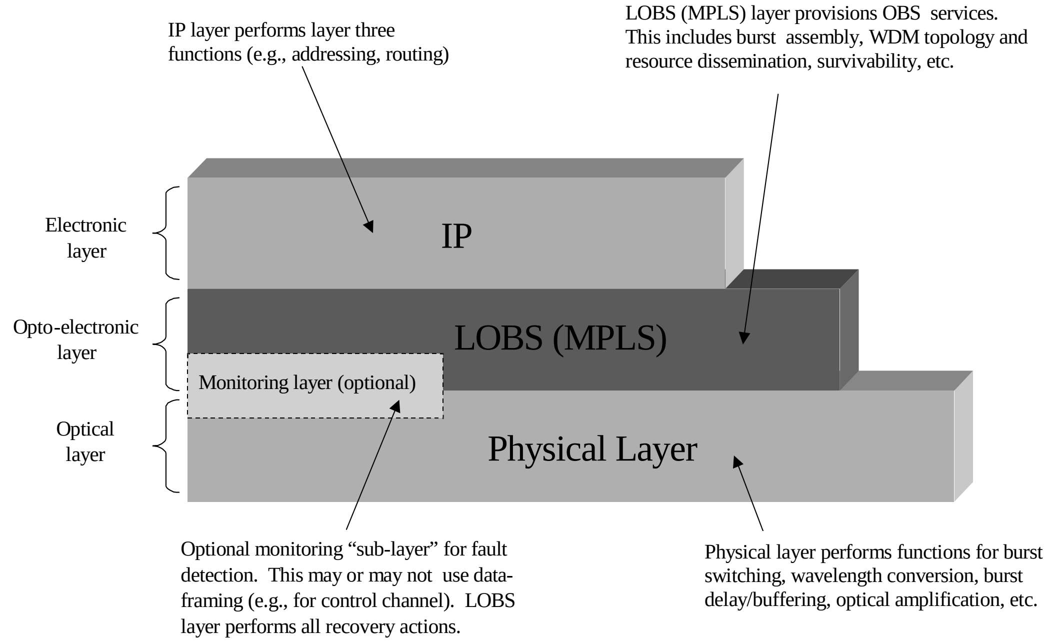
- IP is resilient (albeit *rerouting* may be slow)
- a WDM layer (with optical switches)
  - provides fast restoration (not just WDM links for transmission only)
- Why Integrated IP/WDM
  - no need to re-invent routing and signaling protocols for the WDM layers and corresponding interfaces
  - facilitates traffic engineering and inter-operability

# *MPLS-variants: MPλS and LOBS*

- optical core: circuit- or packet- switched?
- circuit-switched WDM layer
  - OXC's (e.g., wavelength routers) can be controlled by MP*Lambda*S (or MPλS)
- packet-switched or burst-switched (a burst = several packets) WDM layer
  - optical switches controlled by *Labeled Optical Burst Switching* (LOBS) or other MPLS variants.

# *Labeled Optical Burst Switching*

- similar to MPLS (e.g., different LOBS paths can share the same λ)
- control packets carry labels as well as other burst info
- unique LOBS issues: assembly (offset time), contention resolution, light-spitting (for WDM mcast), λ conversion...



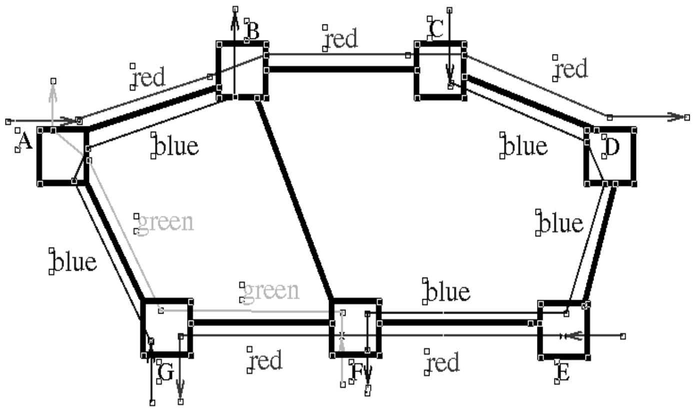
# *Observation*

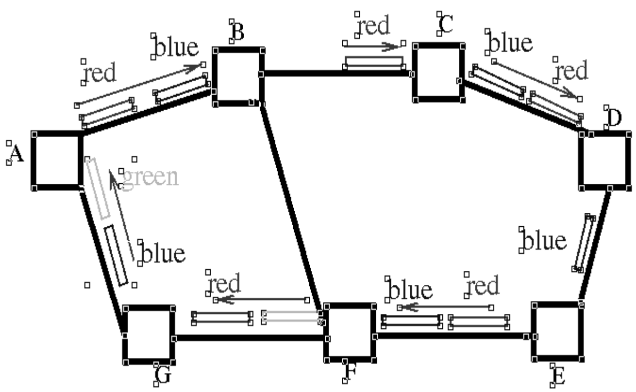
- IP over WDM has evolved:
  - from WDM links, to WDM clouds (with static virtual topology and then dynamic λ services),
  - and now integrated IP/WDM with MPλS
- to be truly ubiquitous, scalable and future-proof, a WDM optical core should also be
  - capable of OOO packet/burst-switching, and basic QoS support (e.g., with LOBS control)

# *Optical Switching Techniques*

historically, circuit-switching is for voice and packet-switching is for data

# *Optical Core: Circuit or Packet ?*

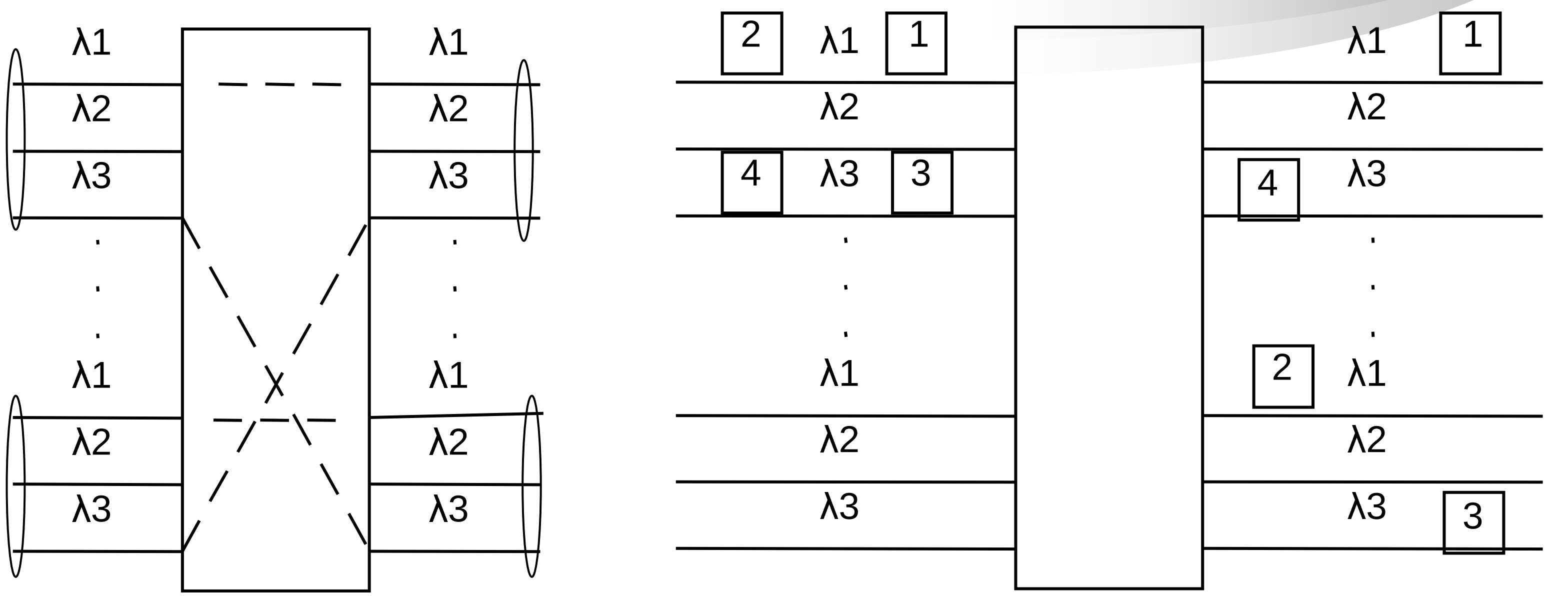



- five *src/dest* pairs
  - circuit-switching (wavelength routing)
    - **3 λs if without λ–conversion**
    - **only 2 λs otherwise**
- if data is sporadic
  - packet-switching
    - **only 1 λ needed with statistical muxing**

      **λ conversion helps too**

# *Impacts on Components*




(a) Cross-Connect (1000 by 1000, *ms* switching time)

(b) Packet-Switch (64x64, with *ns* switching time)

# *Packet Core: A Historical View (hints from electronic networks)*

- optical access/metro networks (LAN/MAN)
  - – optical buses, passive star couplers (Ethernet)
  - – SONET/WDM rings (token rings)
  - – switched networks ? (Gigabit Ethernet)
- optical core (WAN)
  - λ-routed virtual topology (circuits/leased lines)
  - – dynamic λ provisioning (circuits on-demand)
  - – optical burst (packet/flow) switching (IP)

# *Packet Core: Technology Drivers*

- explosive traffic growth
- bursty traffic pattern
- to increase bandwidth efficiency
- to make the core more flexible
- to simplify network control & management by *making the core more intelligent*

# *Circuit Switching*

- long circuit set-up (a 2-way process with Req and Ack): *RTT* = tens of *ms*
- pros: good for smooth traffic and QoS guarantee due to *fixed* BW reservation;
- cons: BW inefficient for bursty (data) traffic
  - either wasted BW during off/low-traffic periods
  - or too much overhead (e.g., delay) due to frequent set-up/release (for every burst)

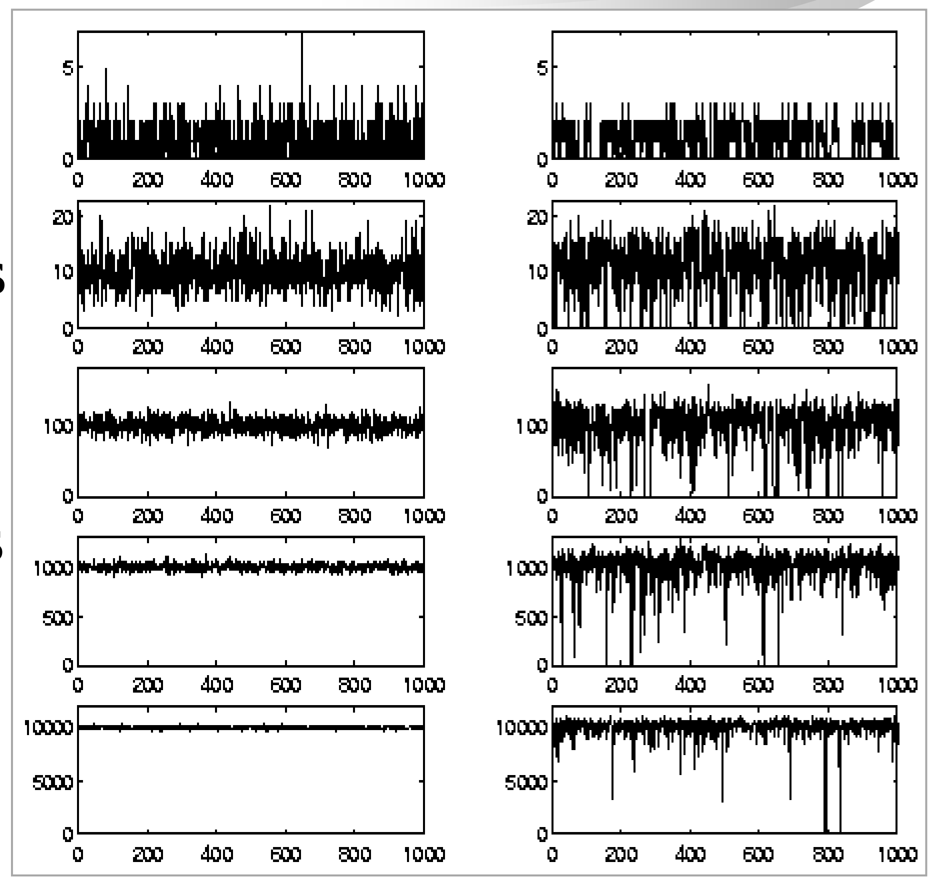
# *Wavelength Routing*

- setting up a lightpath (or λ path) is like setting up a circuit (same pros and cons)

  λ-path specific pros and cons:
  - very coarse granularity (OC-48 and above)
  - limited # of wavelengths (thus # of lightpaths)
  - no aggregation (merge of λs) inside the core
    - traffic grooming at edge can be complex/inflexible
  - mature OXC technology (*msec* switching time)

# *Self-Similar (or Bursty) Traffic*

- Left:
  - *Poisson* traffic (voice)
  - smooth at large time scales and mux degrees
- Right:
  - data (IP) traffic
  - bursty at all time scales and large mux degrees
  - circuit-switching not efficient (*max* >> *avg*)



# *To Be or Not to Be BW Efficient?*

*(don't we have enough BW to throw at problems?)*

- users' point of view:
  - with more available BW, new BW intensive (or hungry) applications will be introduced
    - high BW is an addictive drug, can't have too much!
- carriers' and venders' point of view:
  - expenditure rate higher than revenue growth
  - longer term, equipment investment cannot keep up with the traffic explosion
  - need BW-efficient solutions to be competitive

# *Packet (Cell) Switching*

- A packet contains a header (e.g., addresses) and the payload (variable or fixed length)
  – can be sent without circuit set-up delay
  – statistic sharing of link BW among packets with different source/destination
- store-and-forward at each node
  – buffers a packet, processes its header, and sends it to the next hop

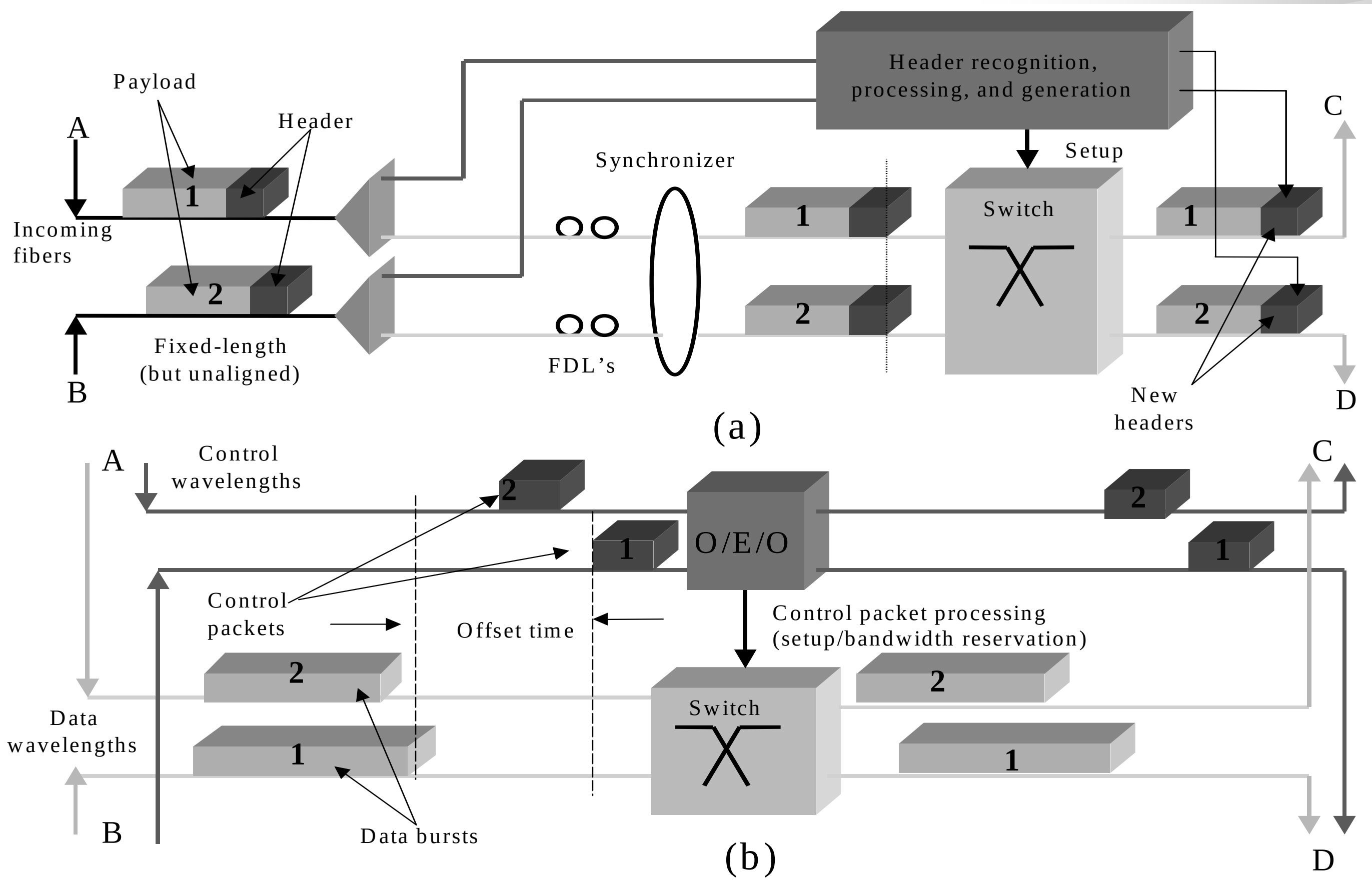
# *Optical Packet Switching: Holy Grail*

- No.1 problem: lack of optical buffer (RAM)
- fiber delay lines (FDLs) are bulky and provide only limited & deterministic delays
  - store-n-forward (with feed-back FDLs) leads to fixed packet length and synchronous switching
- tight coupling of header and payload
  - requires stringent synchronization, and fast processing and switching (*ns* or less)

# *Optical Burst Switching (OBS)*

- a burst has a long, variable length payload
  - low amortized overhead, no fragmentation
- a control packet is sent out-of-band ($\lambda_{control}$)
  - reserves BW ($\lambda_{data}$) and configures switches
- a burst is sent after an offset time *T >0 (loose coupling*), but $T << RTT$ (1-way process)
  - uses asynchronous, cut-through switching (no delay via FDLs needed)

# *Packet (a) vs. Burst (b) Switching*



# *Optical Packet or Burst Switching?*

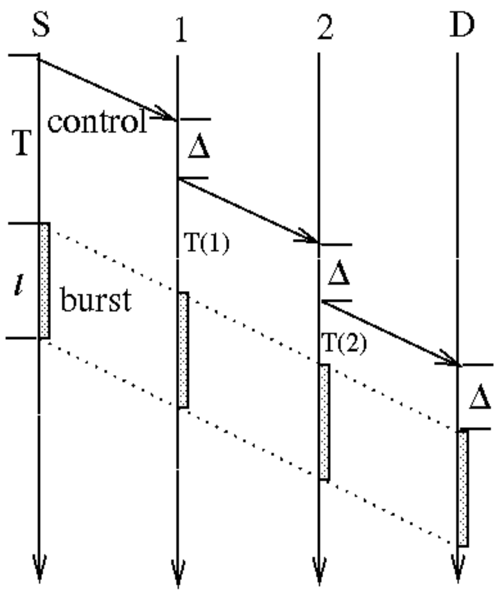
- OBS = optical packet switching with:
  - variable-length, super (or multiple) packets
  - asynchronous switching with switch *cut-through* (i.e., no store-and-forward)
    - a packet is switched before its last bit arrives
  - out-of-band control using e.g., dedicated $\lambda$s or sub-carrier multiplexing (SCM)
    - electronically processed or optically processed (with limited capability and difficult implementation)

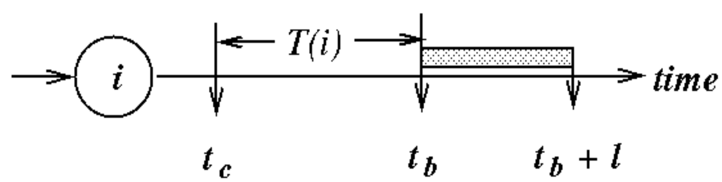
# *OBS Protocols*

- based on Reserve-Fixed-Duration (*RFD*)
  - T >= Σ (processing delay of the control packet)
    - eliminate the need for FDLs at intermediate nodes
  - same end-to-end latency as in packet-switching
    - bursts delayed (electronically) at sources only
    - use 100% of FDL capacity for contention resolution
  - auto BW release after a fixed duration (= burst length) specified by the control packet (YQ97)

# *Just-Enough-Time (JET)*

- combined use of offset time and delayed reservation (DR) to facilitate intelligent allocation of BW (and FDLs if any)



# *TAG-based Burst Switching*

- BW reserved from the time control packet is processed, and released with: (Turner'97)
  - an explicit *release* packet (problematic if lost)
  - or frequent *refresh* with time-out (overhead)
- T = 0 (or negligible)
  - without *DR*, using ***T > 0*** wastes BW
  - FDLs per node >= max {proc. + switch time}

# *Burst Switching Variations*

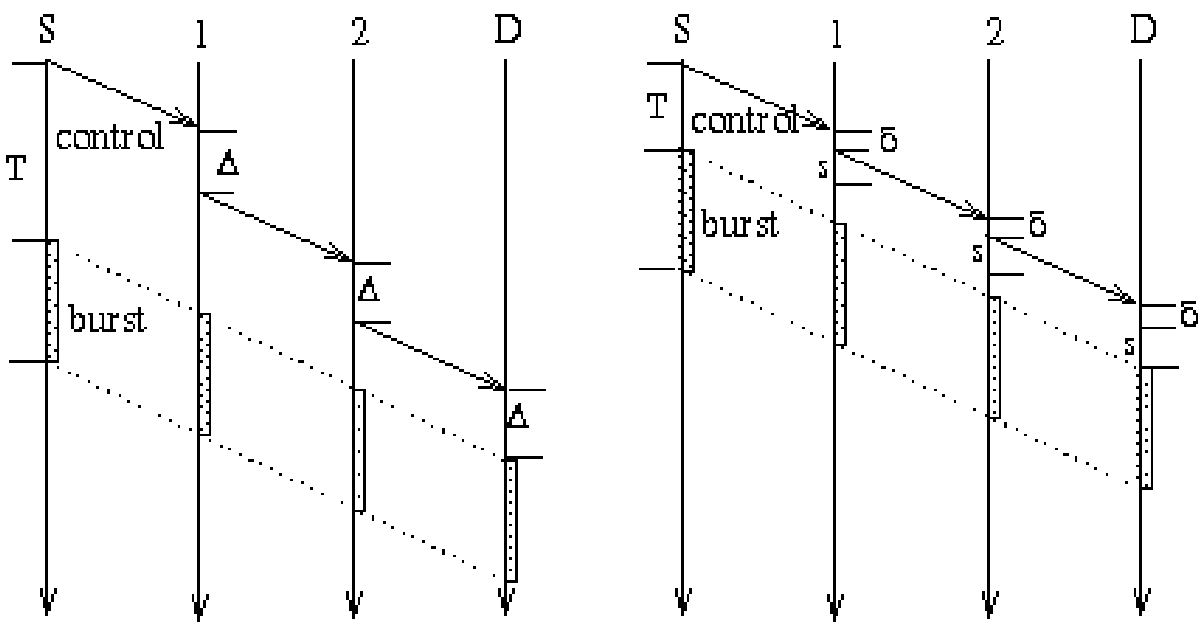
- based on Tell-And-Go (TAG)
  - BW reserved from the time control packet is processed, and released with: (Turner97)
    - either an explicit *release* packet (problematic if lost)
    - or frequent *refresh* packets with time-out (overhead)
- based on In-Band-Terminator (IBT)
  - BW released when an IBT (e.g., a period of silence in voice communications) is detected
  - optical implementation is difficult

# *More on Offset Time*

- TAG and IBT: T = 0 (or negligible)
  - without *DR*, using ***T > 0*** wastes BW
  - FDLs per node >= max. (proc. + switch) time
- *JET* buffers bursts for T > Σ (Δ: proc. delay)
  - a plenty of electronic buffer at source
  - no mandatory FDLs to delay payload
  - can also take advantage of FDLs (buffer)
    - 100 % used for (burst) contention resolution

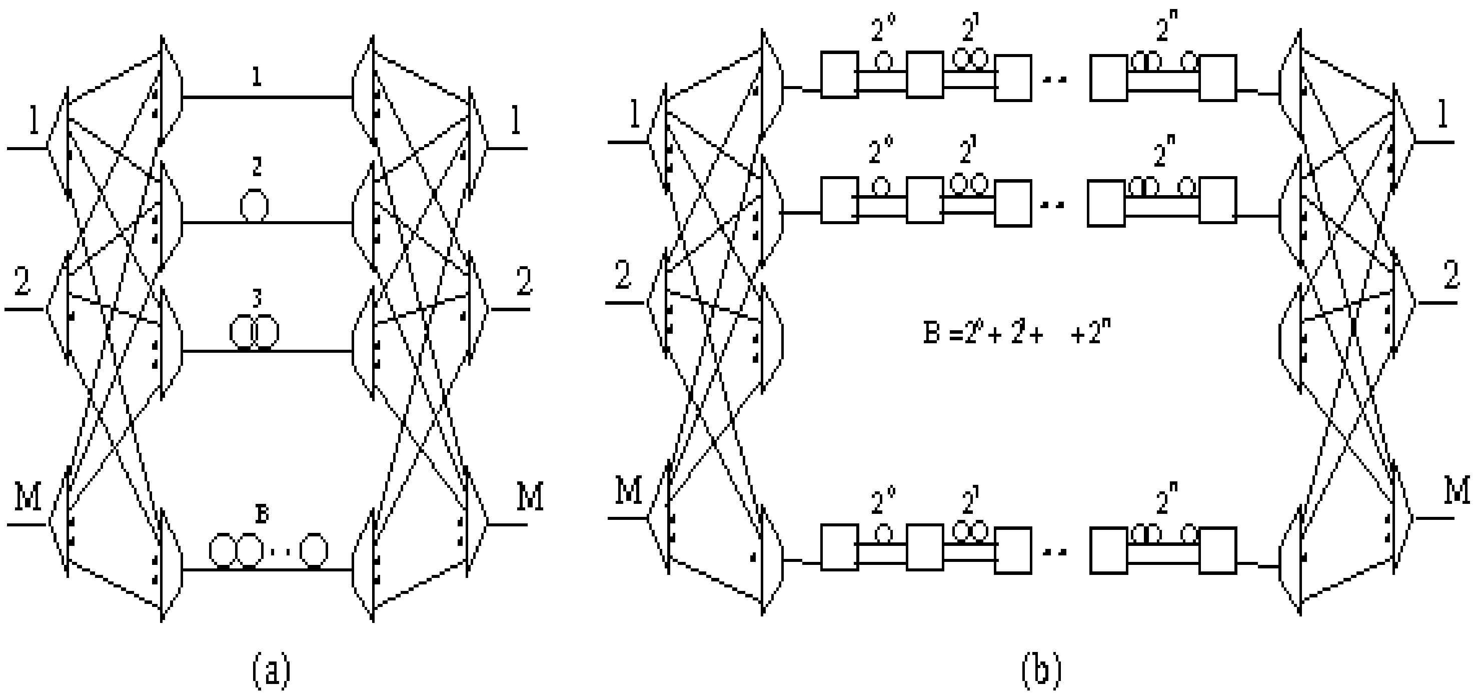
# *Tolerate Switching Delay*

- control packet can leave right after $\delta = \Delta - s$
  - where *s* is the switch setting time

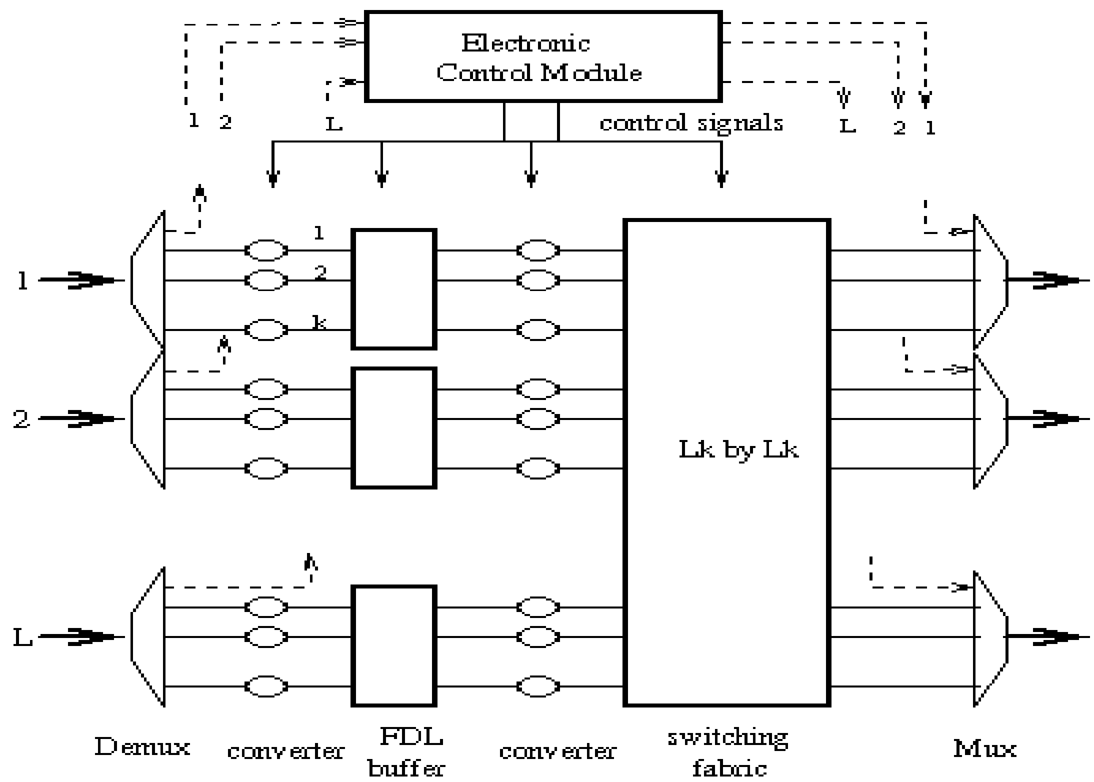


# *FDLs for Contention Resolution*

- shared (a) or dedicated (b) structure with max delay time = *B*



# *OBS Nodes with FDL*

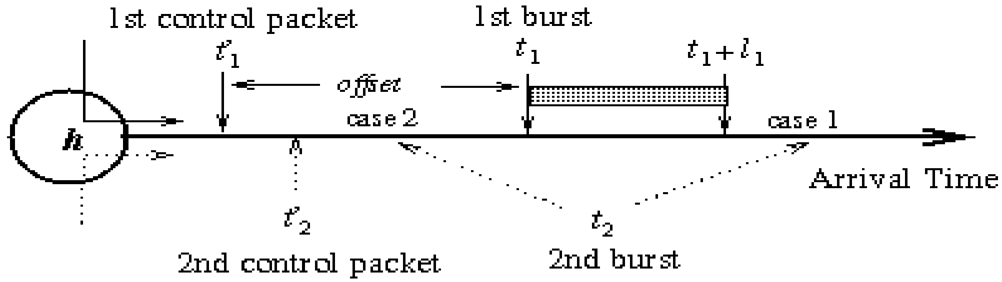


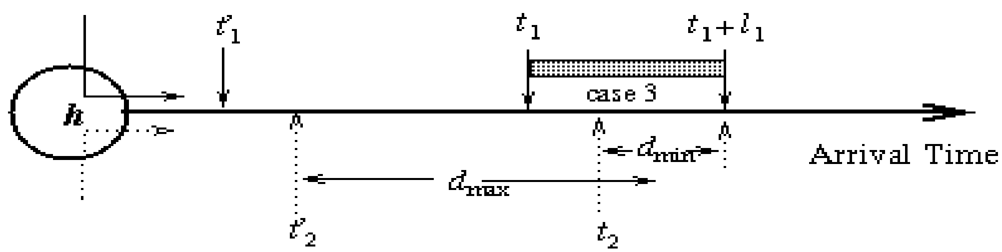
# *BW and FDL Allocation*

– intelligent BW scheduling (known durations)

– no wasted FDL capacity (known blocking time)

- max. delay time *0 < dmax <= B*



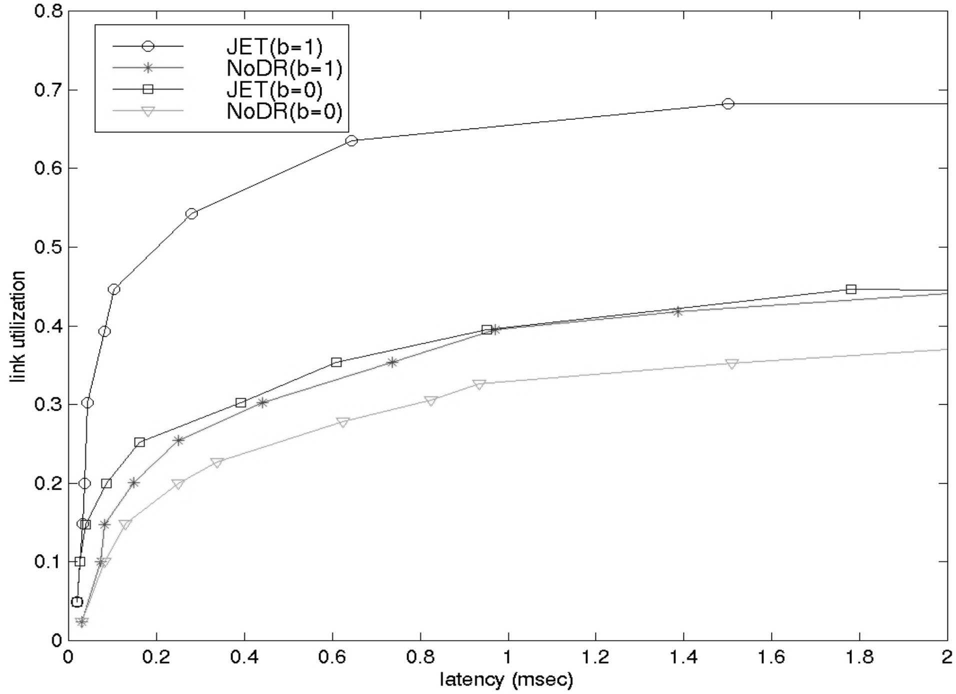
# *Performance Evaluation*

- metrics: link utilization *vs*. latency
- a 16-node mesh network (with OC-192 links)
- ave. burst length (*L*): *0.1 msec* (1 Mbits)
- relative FDL capacity $b = B/L$ is 0 or 1
- also found performance improvement of JET over other protocols scale with
  - # of λs (*k*) & relative processing speed $c = \Delta/L$

# *BW Utilization vs Latency*

– JET as good as NoDR with FDLs

– JET with FDLs 50% better NoDR with FDLs.

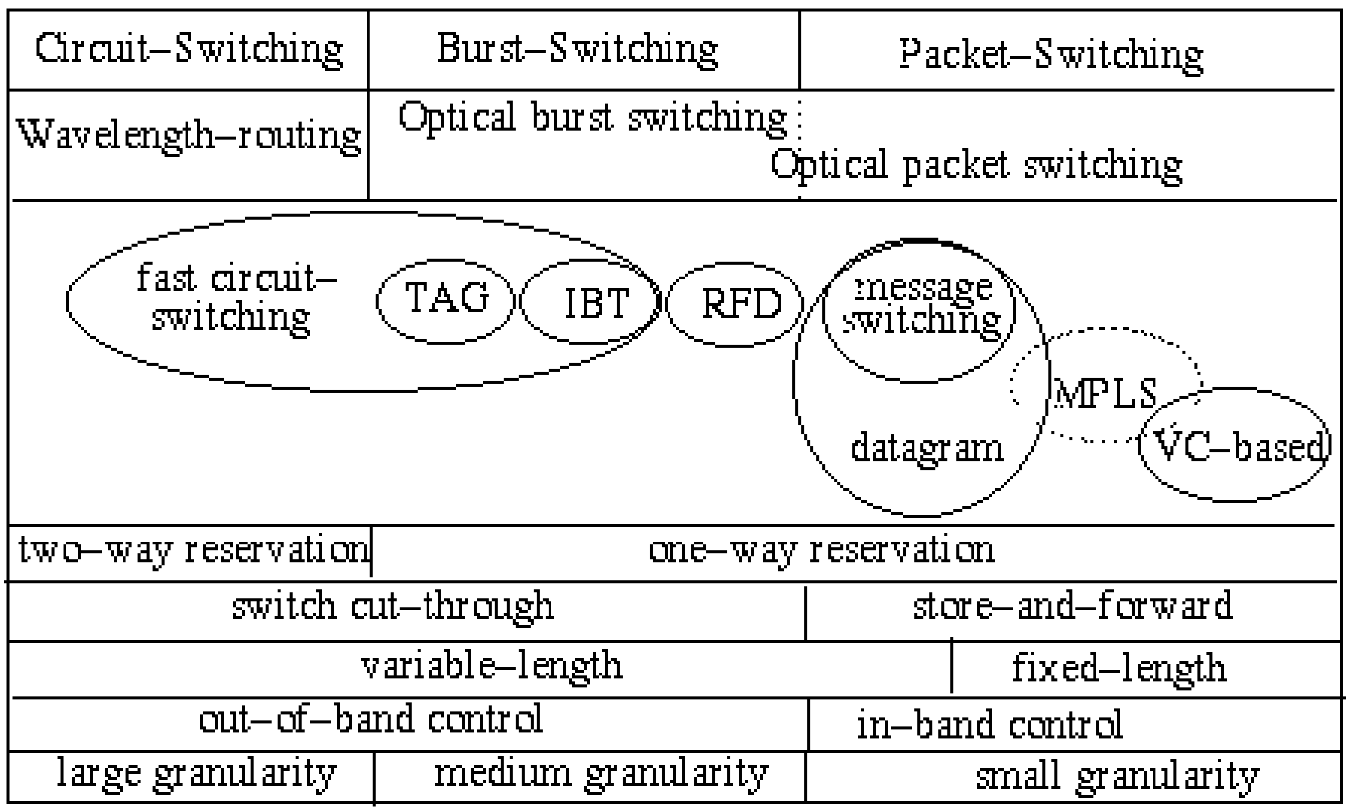


# *Why OBS? A Comparison*

| Optical switching paradigms | Bandwidth Utilization | Latency (setup) | Optical Buffer | Proc./Sync. Overhead (per unit data) | Adaptivity (traffic & fault) |
|---|---|---|---|---|---|
| Circuit | **Low** | **High** | **Not required** | **Low** | **Low** |
| Packet/Cell | **High** | **Low** | **Required** | **High** | **High** |
| **OBS** | **High** | **Low** | **Not required** | **Low** | **High** |

OBS combines the best of coarse-grained circuit-switching with fine-grained packet-switching

# *Switching Paradigms (Summary)*



# *Support QoS Using OBS*

# *QoS schemes*

- current IP: single class, best-effort service
  - *Apps*, users and ISPs need *differentiated service*
- existing schemes (e.g., WFQ) require buffer
  - so to have different queues and, service a higher priority queue more frequently
  - not suitable for WDM networks
    - no optical RAM available (FDLs not applicable)
    - using electronic buffers means E/O/E conversions

# *Why QoS at WDM layer?*

- a WDM layer supporting basic QoS will
  - support legacy/new protocols incapable of QoS and thus making the network truly ubiquitous
  - facilitate/complement future QoS-enhanced IP
  - handle mission-critical traffic at the WDM layer for signaling, and restoration

# *Prioritized OBS Protocol*

- extend *JET* (which has a *base t* > 0) by using an *extra* offset time T to isolate classes
- example:
  - – two classes (class 1 has priority over class 0)
  - – class 1 assigned an *extra T*, but not class 0

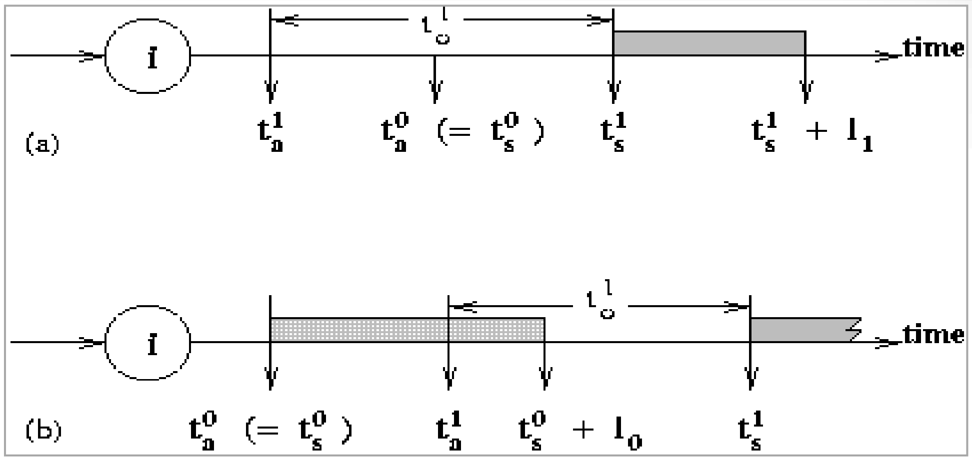
# *Prioritized OBS (continued)*

- – no buffer (not even FDLs) needed, suitable for all-optical WDM networks
- – can take advantage of FDLs to improve QoS performance (e.g., a higher isolation degree)
- – the *extra T* does introduces *additional* latency
  - • but,  only insignifcantly (e.g., <=  a few *ms*)

# *Why Extra Offset Time => Priority ?*

- assumptions:
  - a link having one available λ and no FDLs
  - two classes (class 1 has priority over class 0)
    - lost class 0 (best-effort class) bursts retransmitted
    - class 1 (critical) bursts need low blocking prob.
  - class 1 assigned an *extra T*, but not class 0
  - the difference in their base *t*'s is *negligible*

# *Class Isolation: Example*




- a class 0 burst won't block a class 1 burst
  - class 1 control packet arrives first (a)
  - class 0 control packet arrives first (b)
- extra T = right to reserve BW in advance

# (Extra) Offset Time Required

- extra T assigned to class 1: $t_1$
- class 0 burst length: $l_0$
  - expected ave: 10 *Mbits* or 1 *ms* @ OC-192
- completely isolated classes if $t_1$ >= max.{$l_0$}
- let $p$ = prob {$l_0$ <= $t_1$ }, that is, $p$% of class 0 bursts are no longer than $t_1$
  - partially isolated (with a degree of *p)*
  - e.g., 95% isolation when $t_1$ = 3 times of ave{$l_0$}

# *When Number of Classes (n) > 2*

- $L_i$: class *i*'s mean burst length
- $t_{i,i-1}$: difference in *T* between classes *i* & *i-1*
- $R_{i,i-1}$: (adjacent) class isolation degree
  - prob. {class *i* will not be blocked by class *i-1*}
- $R_{i,i-1}$= PDF{class *i*-1 bursts shorter than $t_{i,i-1}$}
  - with exponential distribution

$$PDF = 1 - e^{(-u_{i-1} \times t_{i,i-1})}, \quad u_{i-1} = 1/L_{i-1}$$

# *Isolation Degree Achieved*

| offset time difference | 0.4 $L_{l-1}$ | $L_{l-1}$ | 3 $L_{l-1}$ | 5 $L_{l-1}$ |
|---|---|---|---|---|
| Isolation degree | 0.3296 | 0.6321 | 0.9502 | 0.9932 |

- more isolated from lower priority classes
  - class *i* is isolated from class *i* - 1 with $R_{i,i-1}$
  - class *i* is isolated from class *i* - 2 with $R_{i,i-2} > R_{i,i-1}$ (since $t_{i,i-2} = t_i - t_{i-2} > t_{i,i-1} = t_i - t_{i-1}$ )
  - similarly, class *i* is isolated from all lower classes with at least $R_{i,i-1}$

# *Analysis of Blocking Probability*

- single node with $k$ $\lambda$'s and $\lambda$–conversions
- the classless OBS (for comparison)
  - blocking probability: $B(k,\rho)$ using *Erlang*'s loss formula ($M/M/k/k$) (bufferless)
- the prioritized OBS
  - $B(k, \rho) =$ ave. blocking probability over all classes (the conservation law)
  - assume complete (100%) class isolation

# *Analysis (II)*

- block prob. of class $n$ - 1 (highest priority)
  - $pb_{n-1} = B(k,\rho_{n-1})$ because of its complete isolation from all lower priority classes
- blocking prob. of bursts in classes $j$ to $n$ - 1:
  - calculated as *one* super class isolated from all lower classes: $PB_{n-1,j} = B(k,\rho_{n-1,j})$ (1)
  - where the combined load is

$$\rho_{n-1,j} = \sum_{i=j}^{n-1} \rho_i$$

# *Analysis (III)*

- blocking prob. of bursts in classes *j* to *n* - 1
  - when calculated as a weighted sum:

$$PB_{n-1,j} = \sum_{i=j}^{n-1} c_i \times Pb_i \quad where \quad c_i = \rho_i / \rho \quad (2)$$

- given blocking prob of classes *j*+1 to *n* - 1

$$pb_j = (B(k, \rho_{n-1,j}) - \sum_{i=j+1}^{n-1} c_i \times pb_i) / c_j$$

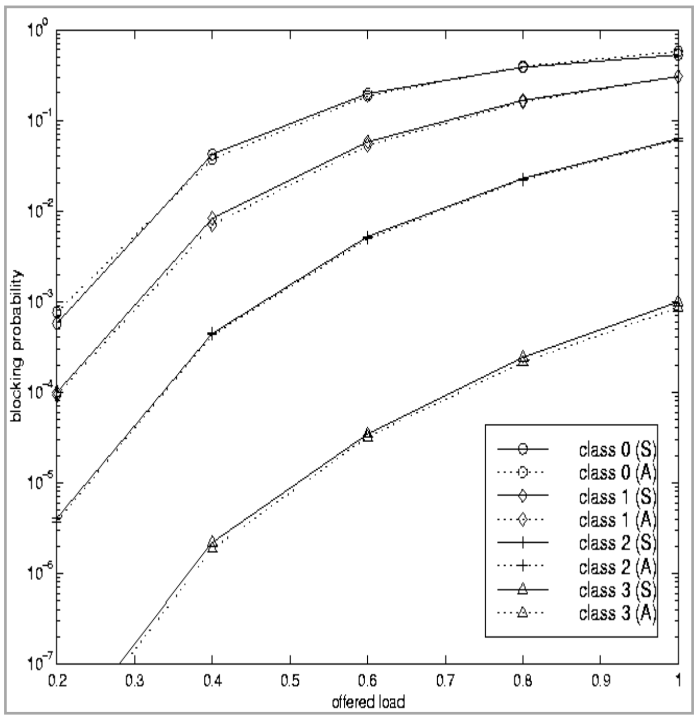
  - e.g., blocking prob. of class *n* - 1

$$pb_{n-2} = (B(k, \rho_{n-1,n-2}) - c_{n-1} \times pb_{n-1}) / c_{n-2}$$

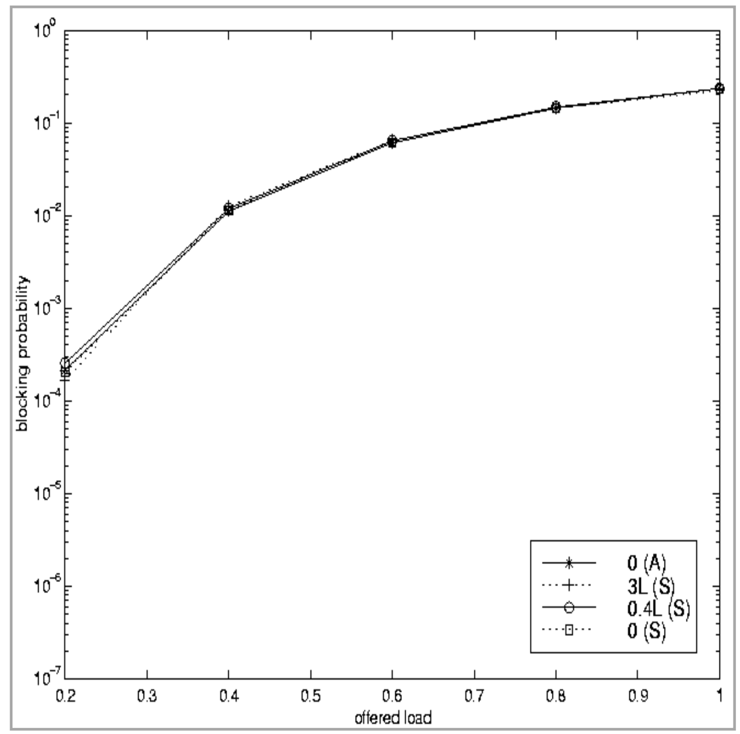
# *Loss Probability vs. Load*

- by default: $n = 4$, $k = 8$, $L_i = L$, and $t_{i,i-1}=3L$

Class Isolation

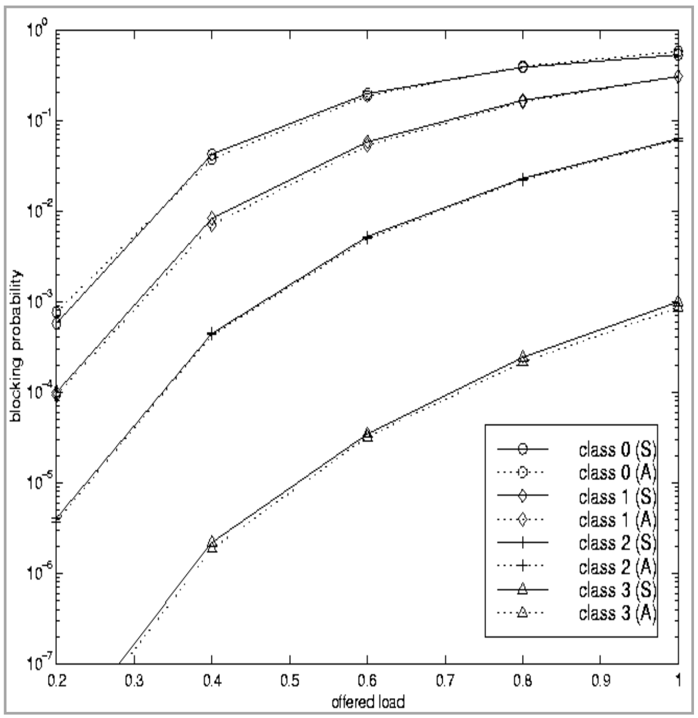



Average (Conversation Law)



# *Differentiated Burst Service*

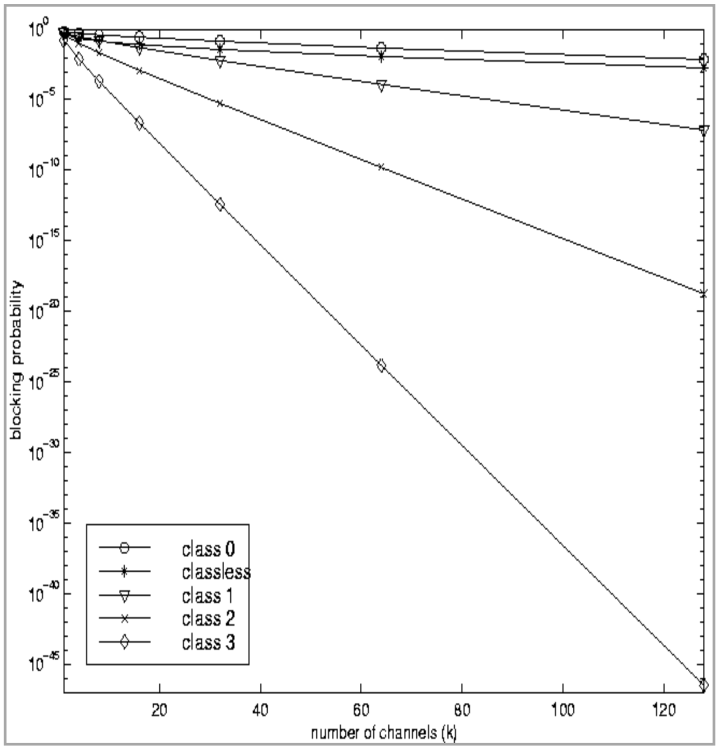
Loss Prob *vs.* Load
(four classes, 8 λs)



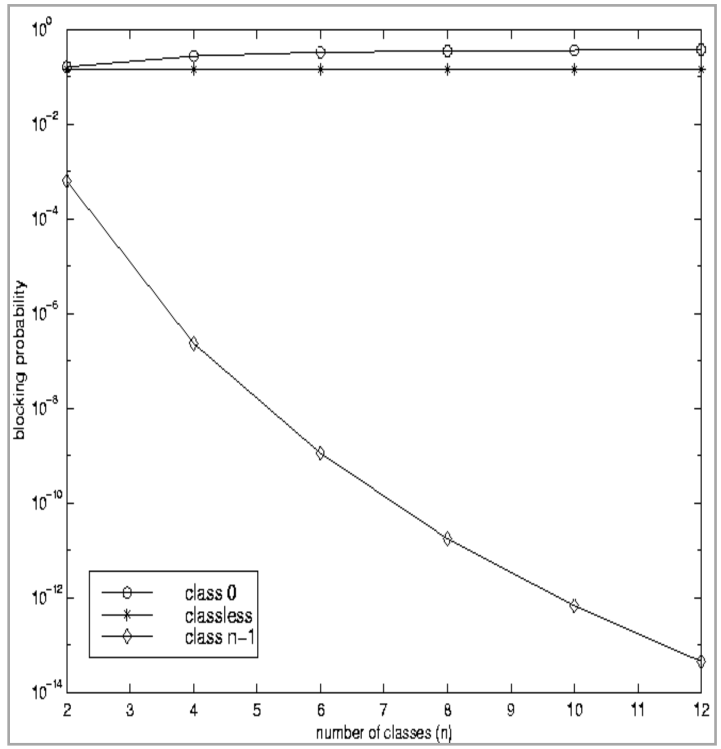

- same average over all classes (conservation law)
- FDLs (if any) improve performance of all classes
- class isolation increases with # of λs, classes and FDLs (if any)
- bounded E2E delay of high priority class

# *Scalability*

Loss prob vs. *k*

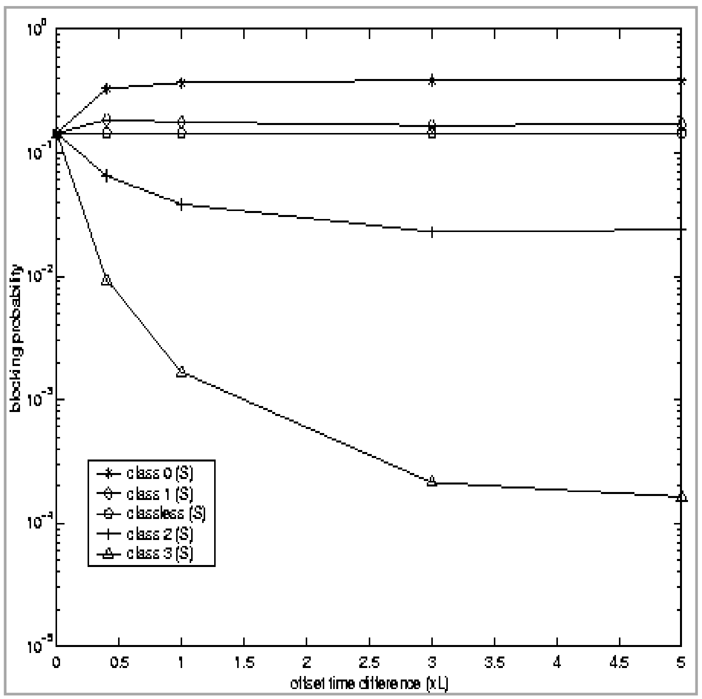



Loss prob vs. *n*

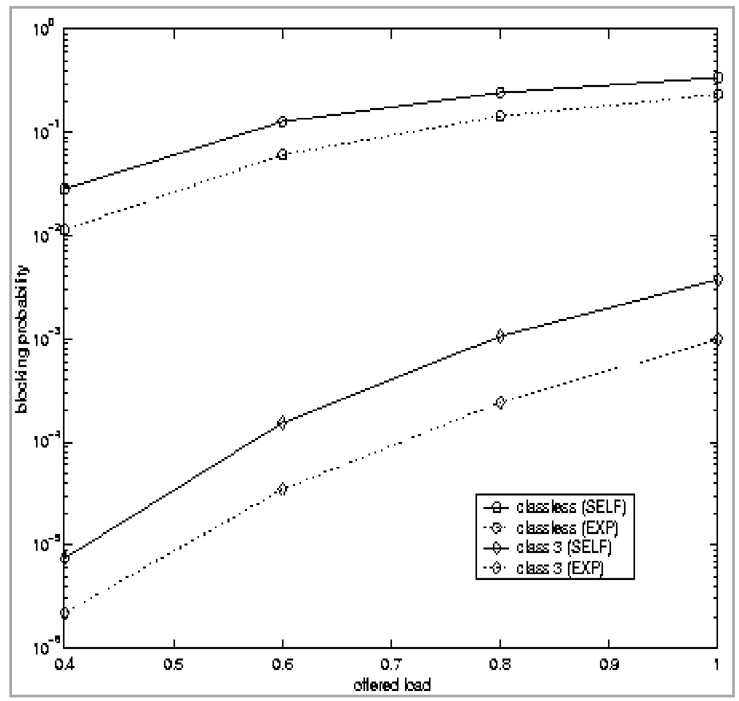


# *Some Practical  Considerations*

Loss prob. saturation when offset time difference = 3L




Loss prob under self-similar traffic



# *Application to FDLs*

- to isolate two classes for FDL reservation
  - extra offset time to class 1 > max{ $l_0$ }
- for λ reservation: *extra t* > B + max{ $l_0$ }
  - class 0 may be delayed for up to *B* units
- isolation degree differs for a given *t*

| FDL (buffer) | 0.4 $L_0$ | $L_0$ | 3 $L_0$ | 5 $L_0$ |
|---|---|---|---|---|
| Wavelength | 0.4 $L_0$ + B | $L_0$ + B | 3 $L_0$ + B | 5 $L_0$ + B |
| Isolation degree (R) | 0.3296 | 0.6321 | 0.9502 | 0.9932 |

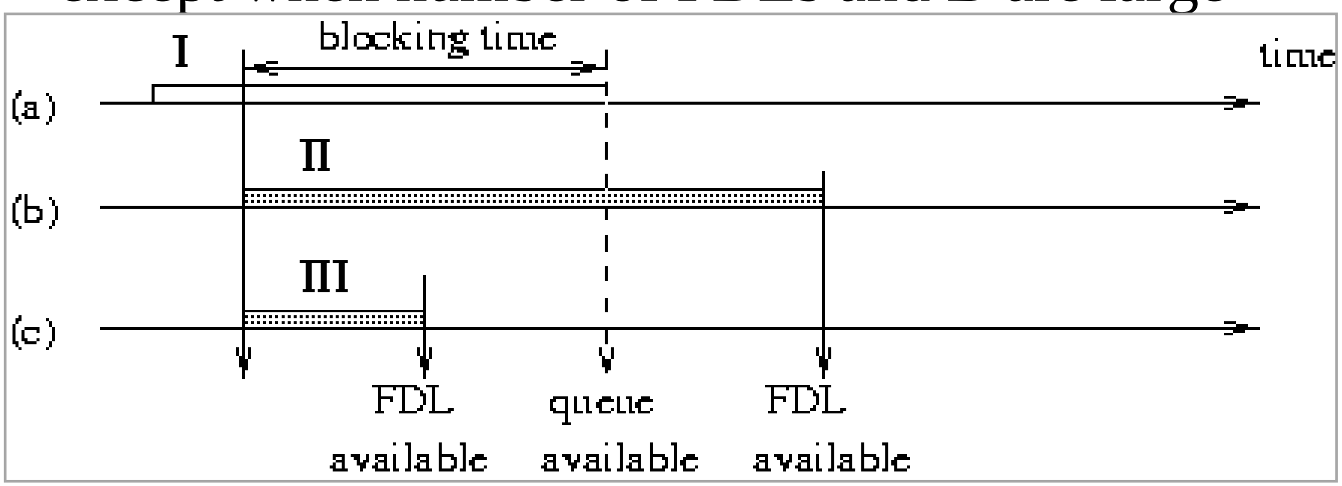
# *FDLs vs Queue*

– FDLs only store bursts with blocking time < *B*

– a queue can store any burst indefinitely

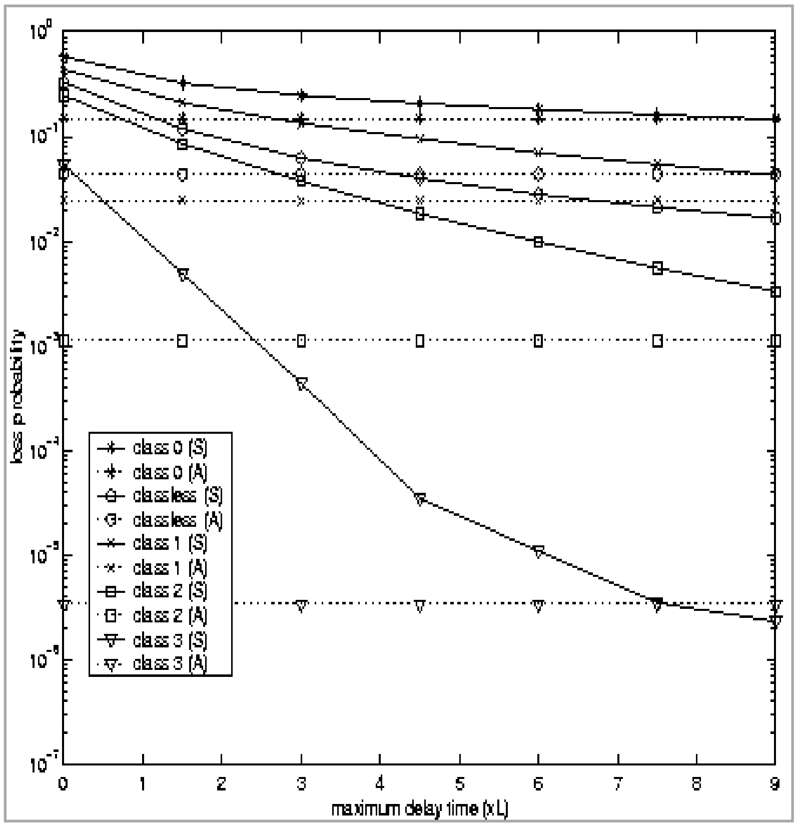
– queueing analysis (M/M/k/D) generally yields a lower bound on the loss probability

- except when number of FDLs and B are large

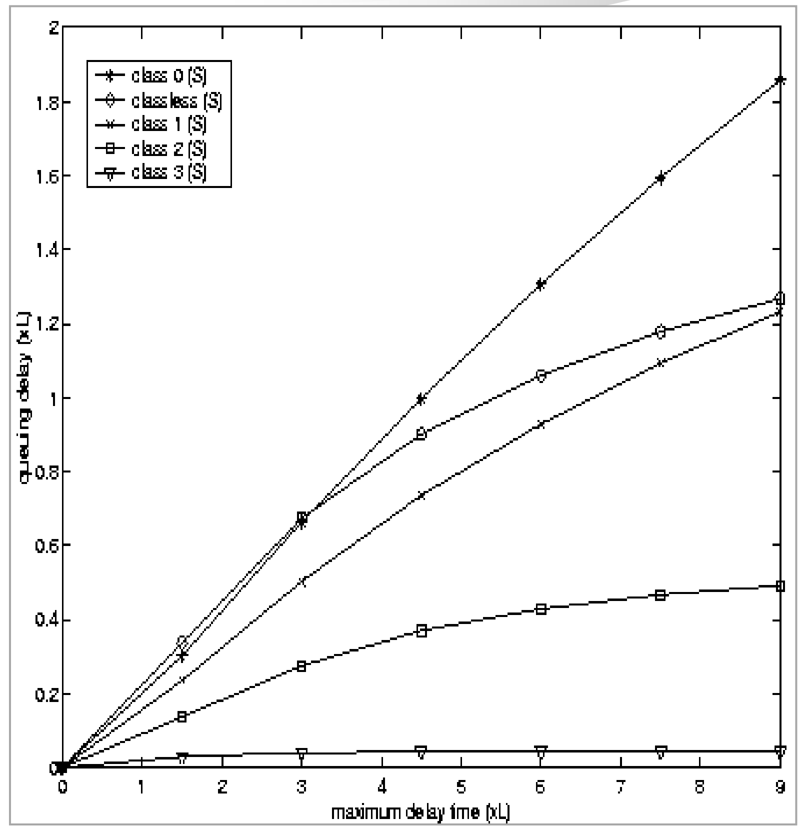


# *Effect of Max Delay Time*

Loss Prob.




Queuing Delay



# *Other Topics in OBS (I)*

- burst assembly
  - based on fixed time, min. length, or burst detection heuristics
- offset time value
  - priority vs additional pre-transmission delay
- burst route determination
  - shortest (in hop count) or least loaded
  - alternate routes & adaptive routing

# *Other Topics in OBS (II)*

- WDM multicasting
  - constrained multicast routing (e.g., multicast forests to get around mcast-incapable switches)
  - IP/WDM multicast interworking
- contention resolution & fault recovery
  - drop, re-transmission (WDM layer), buffering (via FDLs), deflection (in both space and wavelength), or pre-emption

# *End of Part I*