Optical Burst Switching – A Tutorial from e-Photon ONe

Nail AKAR and Ezhan KARASAN
Bilkent University, TURKEY

Contributors

DEIS-UniBo: Carla Raffaelli, Maurizio Casoni (Department of Information Engineering, University of Modena and Reggio Emilia)

Telenor/NTNU: Harald Øverby, Norvald Stol, Steinar Bjørnstad

UPC: Miroslaw Klinkowski, Davide Careglio, Josep Solé i Pareta

Universidad Pública de Navarra : D. Morato, M. Izal, E. Magaña and J. Aracil

Contributors

Univ of Essex : R. Nejabati, D. Simeonidou, M. O’Mahony

RACTI-Univ. of Patras: K. Christodoulopoulos, K. Vlachos

Bilkent Univ.: Kaan Dogan, Guray Gurel, Nail Akar, Ezhan Karasan

TID: Juan Fernández-Palacios, Óscar González

Contributors

Universität Stuttgart

Universität Stuttgart: Christoph M. Gauger, Guoqiang Hu

FP6 IST “Broadband for all”

Network of Excellence

e-Photon/ONe
“Optical Networks: Towards Bandwidth Manageability and Cost Efficiency”

1st phase: 2004-2006

2nd phase: 2006-2008

What is a Network of Excellence?

From Marimon report on EC IST projects: "Networks of Excellence should be designed as an instrument to cover different forms of collaboration and different sizes of partnerships"

- e-Photon/ONe aims at "integrating and focusing the rich know-how available in Europe on optical communication and networks, both in universities and in research centres of major telecom manufacturers and operators" using the following structure:
  - strong integration of a *core membership*
  - active involvement of all partners in the NoE
  - involvement of external institutions ("Collaborating Institutions")

Consortium composition - I

- Politecnico di Torino, Italy
- Università di Bologna, Italy
- Politecnico di Milano, Italy
- Fondazione Ugo Bordoni, Rome, Italy
- Scuola Superiore Sant'Anna, Pisa, Italy
- INTEC - Ghent University - IMEC, Gent, Belgium
- Technical University of Eindhoven, The Netherlands
- Faculté Polytechnique de Mons, Mons, Belgium
- COM - Technical University of Denmark, Copenhagen, Denmark
- Kista Photonics Research Centre, Kista, Sweden
- Fraunhofer Gesellschaft - Heinrich Hertz Institute, Germany
- Duisburg University, Germany
- University of Stuttgart - Institute of Communication Networks and Computer Engineering, Germany
- Technical University Berlin, Berlin, Germany
- Vienna University of Technology, Austria
- Groupe des Ecoles de Telecommunications, France

Consortium composition - II

- University of Essex, UK
- University College London (UCL), London, UK
- University of Cambridge, UK
- University of Southampton, UK
- Universitat Politècnica de Catalunya, Spain
- Universdad Carlos III de Madrid, Spain
- Universidad Pública de Navarra, Spain
- Polytecnic of Valencia, Spain
- Instituto de Telecomunicações, Aveiro, Portugal
- National Technical University of Athens, Greece
- University of Athens, Greece
- University of Patras, Greece
- Budapest University of Technology and Economics, Budapest, Hungary
- Bilkent University, Ankara, Turkey
- University of Zagreb, Zagreb, Croatia
- University of Mining and Metallurgy (AGH), Poland

Consortium composition - III

[Industrial partners]

- **Telefónica Investigación y Desarrollo, Spain**
- **T-Systems Nova GmbH, Germany**
- **Siemens, Germany**
- **Telenor R&D, Oslo, Norway**
- **France Telecom, France**
- **Alcatel R&I, France**

38 partner institutions:
- 32 academic institutions
- 4 telecom operators
- 2 manufacturers

with broad European coverage (from Portugal to Turkey)

~400 researchers actively involved in the NoE

Coordinator: Fabio Neri (Politecnico di Torino)

Objectives of e-Photon/ONe

- e-Photon/ONe is focused on **optical networks**
- Its main goals are:
  - integrate and focus the rich technical know-how available in Europe on optical networking
  - favour a consensus on the engineering choices towards the deployment of optical networks
  - understand how to exploit the unique characteristics of the optical domain for networking applications
  - promote and organize activities to disseminate knowledge on optical networks

e-Photon/ONe VDs

VD1

Core Networks: technologies, architectures, protocols

VD2

Metro & Access Networks: technologies, architectures, protocols

VD3

Home Networks & other Short-Reach Networks

VD4

Optical Switching Systems

VD5

Transmission Techniques for Broadband Networks

Tutorial Outline

- Introduction to Optical Burst Switching Networks
  - Circuit/packet/burst Switching
  - Signaling Issues
  - Just Enough Time Protocol
  - Burst Assembly Mechanisms
- Contention Resolution Mechanisms

Tutorial Outline

- Teletraffic Modeling of OBS Networks
- Quality of Service Mechanisms for OBS Networks
- TCP over OBS
- Wrap-up

OBS Tutorial

Introduction to Optical Burst Switching Networks

Electronic vs Optical Switching

- Data transmission is carried out in the optical domain today in WANs and MANs today however switching is mostly done in the electronic domain
- Electronic switching uses electronic switching fabrics
  - Converts data from optical to electronic for switching purposes, and then from electronic back to optical for transmission.
- Optical switching uses optical switching fabrics
  - Payload stays in the optical domain

Advances 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 on the way (>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*)

Caveats of OEO Switching

- Internet traffic doubles 6 months (1997-2008)
- Semiconductor performance doubles every 18 months which is known as the Moore’s Law
- The first time in history that improvements have been required faster than the improvement rate for semiconductors, Moore’s Law.
  - Complex operations are needed at a OEO router's line card for example processing the packet header, longest prefix match, packet buffering, etc.
- The cost of OEO at OC-48 (2.5Gbps) and at OC-192 is relatively high

Circuit Switching

- Two-way process with request and acknowledge
  - Round Trip Time = tens of *ms* therefore long setup delays
  - Suitable for smooth traffic and QoS guarantees due to fixed bandwidth allocation
  - Bandwidth inefficient for bursty (data) traffic
    - Wasted bandwidth during off/low-traffic periods
    - Overhead due to frequent set-up/release

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 the edge can be complex/inflexible
  - Mature OXC technology (*msec* switching time)
- Current state of the art

Packet 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 bandwidth 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
- One-way process

Optical Packet Switching

- Optical packet consists of a header and a payload
- Packet header is processed all-optically at each node and switched to the next hop

+ Statistical multiplexing of data
+ Suitable for bursty traffic

– Requires fast switching speeds (nanoseconds)
– Stringent synchronization requirements
– More viewed as a longer term solution

Motivations for a New Paradigm

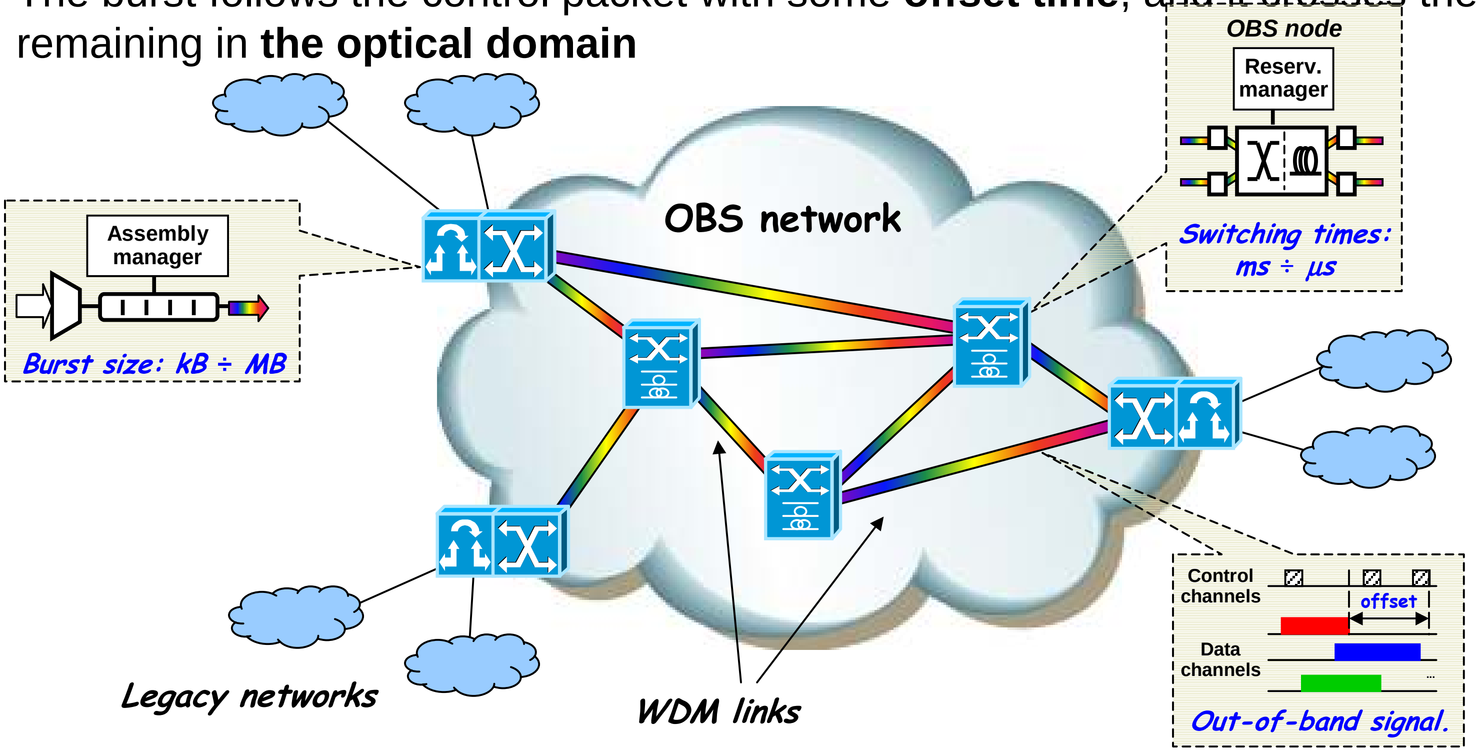
- Changes in traffic profile
  - P2P file downloading vs. multimedia streaming
  - grid networking
- Wavelength routed networks
  - low network utilization and flexibility
- Problems in optical packet switched networks
  - lack of optical buffering
  - need for fast packet switching and header processing

OBS Approach

- Main design objectives
  - decreasing complexity of OPS with still employed statistical multiplexing in optical domain
  - building a buffer-less network
  - user data travels transparently as an optical signal and cuts through the switches at very high rates
- Solution
  - sending a header to temporarily reserve a wavelength path
  - after that, sending an optical burst (a block of IP packets) through the network
- Thanks to the great variability in the duration of bursts, the OBS can be viewed as lying between OPS (one-way reservation) and WS networks (two-way reservation)

OBS Network Architecture

- Control and data information travel **separately** on different channels
- Data coming from legacy networks are aggregated into **a burst unit** in edge node
- **The control packet** is sent first in order to reserve the resources in intermediate nodes
- The burst follows the control packet with some **offset time**, and it crosses the nodes remaining in **the optical domain**



OBS Principles

- Variable-length packets, named bursts
- Asynchronous node operation
- A strong separation between the control and data planes
  - Control burst (with control information) transmitted on dedicated control channel and processed electronically
  - Data burst transmitted and switched all-optical way

Burst Signaling Protocols

- Burst transmission is preceded by a setup message to reserve resources
- Signaling packets undergo E/O conversion at every hop while burst data travels transparently
- Two different types of protocols
  - Tell-and-Wait (TAW): two-way reservation schemes
  - Tell-and-Go (TAG): one-way reservation schemes

TAW-two way reservation schemes

- The data burst is transmitted after an end-to-end connection is established
  - SETUP is sent to hard reserve resources
  - ACK packet acknowledges the reservation
  - In case of failure – setup phase can be repeated
- Main drawbacks: Long round trip time
- Solutions:
  - burst size estimation -> earlier transmission of SETUP
  - “timed” and “in advance” mechanism ->
    - increase burst acceptance probability
    - decrease the number of setup retransmissions

TAG-one way reservation schemes

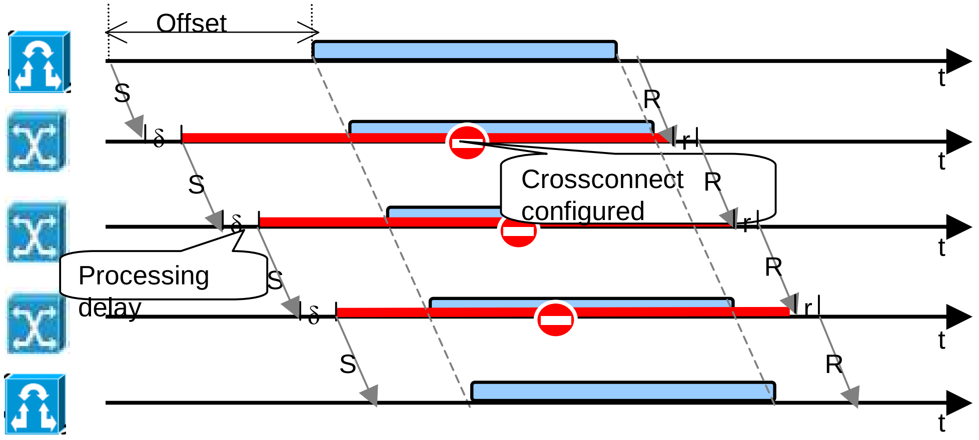
- Signaling messages travel ahead of the data burst
- Burst is transmitted after a time offset that prevents a burst from entering the switch before the configuration is finished
- Classification of TAG Variants
  - Start of Reservation
    - *Immediate-explicit:* reservation starts immediately after the reception of the SETUP
    - Simple to implement
    - *Delayed-implicit:* reservation start by the beginning of the data
    - Requires vast memory and complex to implement
  - Release mechanism (tearing down)
    - Implicit: based on burst length information.
    - Explicit: use a release control packet.

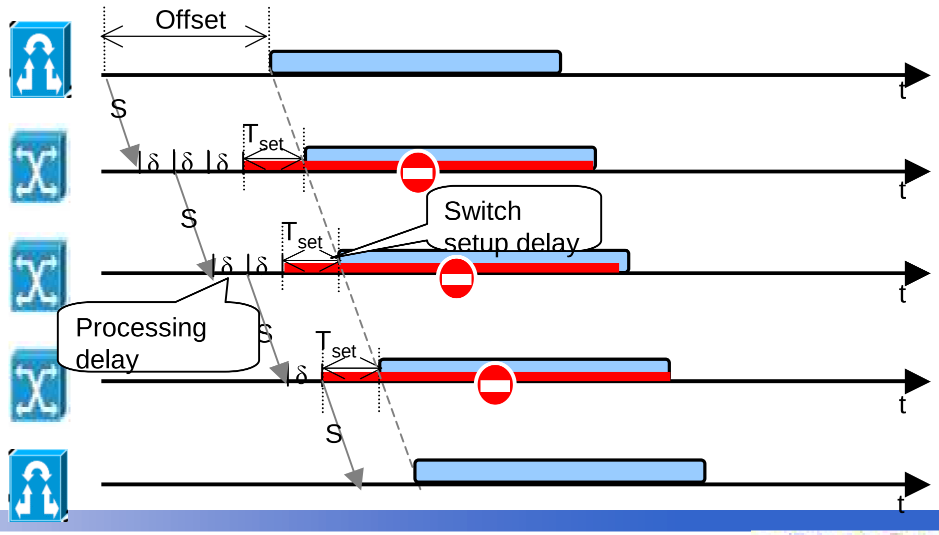
State-of-art in OBS signaling

- JIT protocol:

  explicit setup and explicit or implicit release

- Horizon and JET protocols employ estimated setup and estimated release
  - Horizon doesn't support void filling
  - JET supports void filling

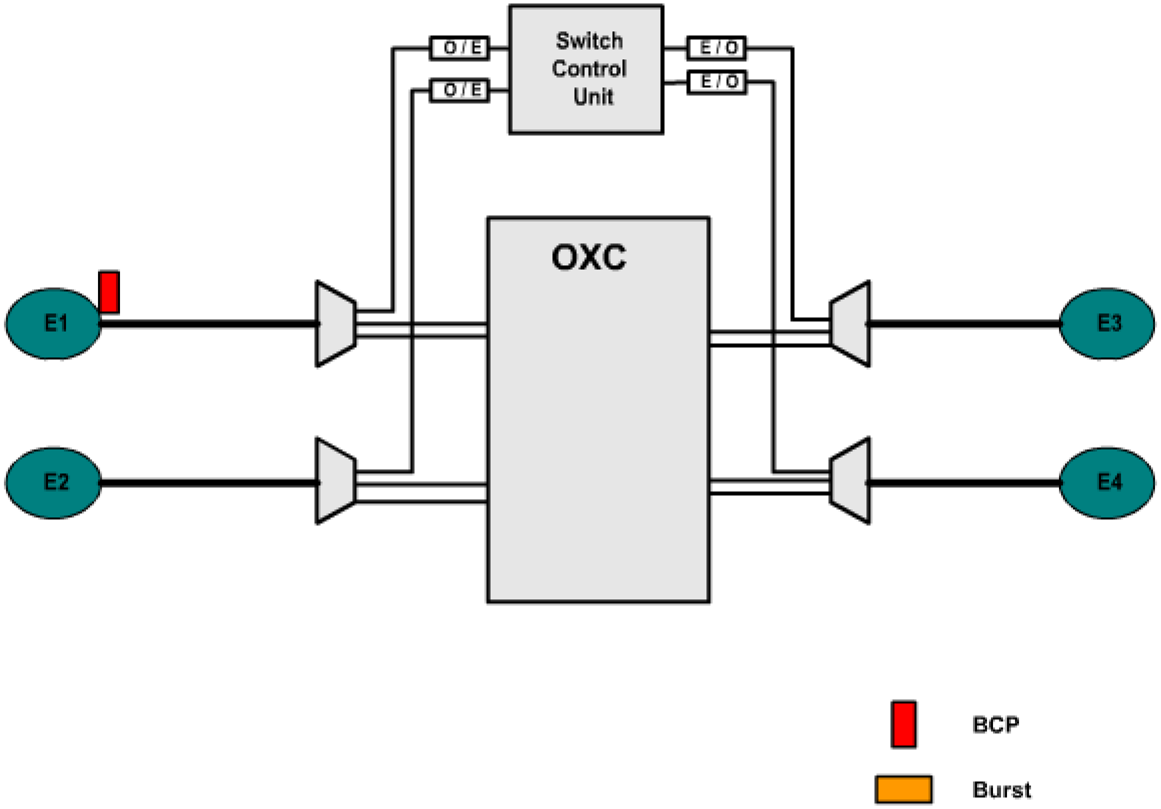




Just-Enough-Time (JET) Protocol

- Mostly adopted reservation protocol that uses the one-way reservation mechanism
  - A BCP is sent first over a separate control channel
  - Data is sent after an offset time T_{off} over the data channel
- The BCP consists of
  - The offset time information
  - The burst length information
- The offset time field is used by intermediate nodes to determine the arrival time of the burst
- The length of the burst enables the switches to make close-ended reservations for bursts

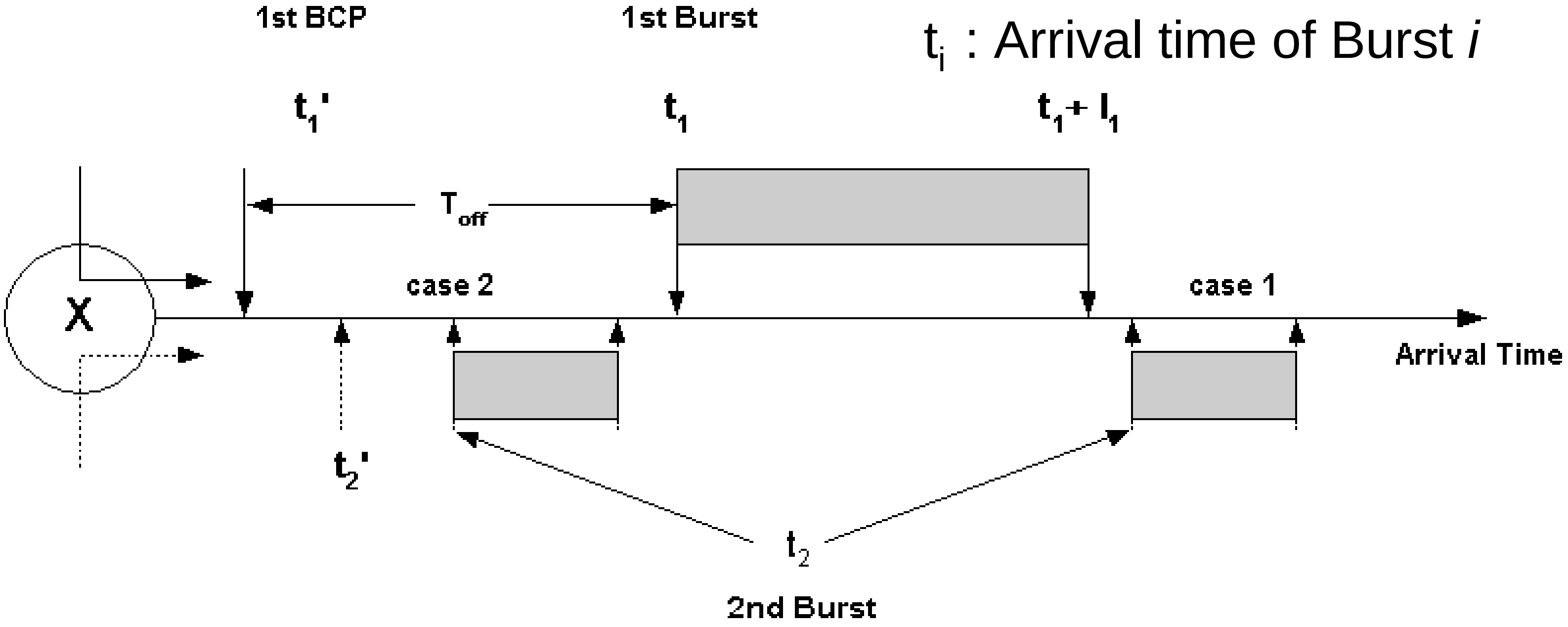
JET Overview



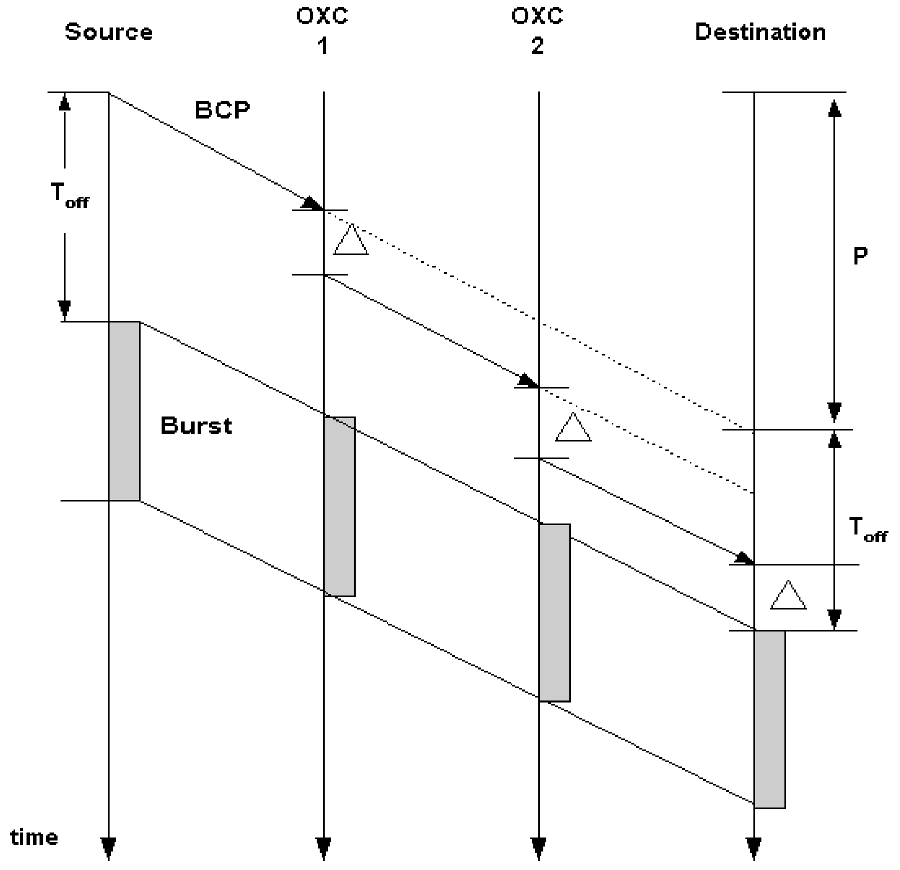
JET (Cont'd)

t_i' : BCP arrival time for Burst *i*

t_i : Arrival time of Burst *i*



JET (Cont’d) – The case of 3 hops



△ : Processing time at each node

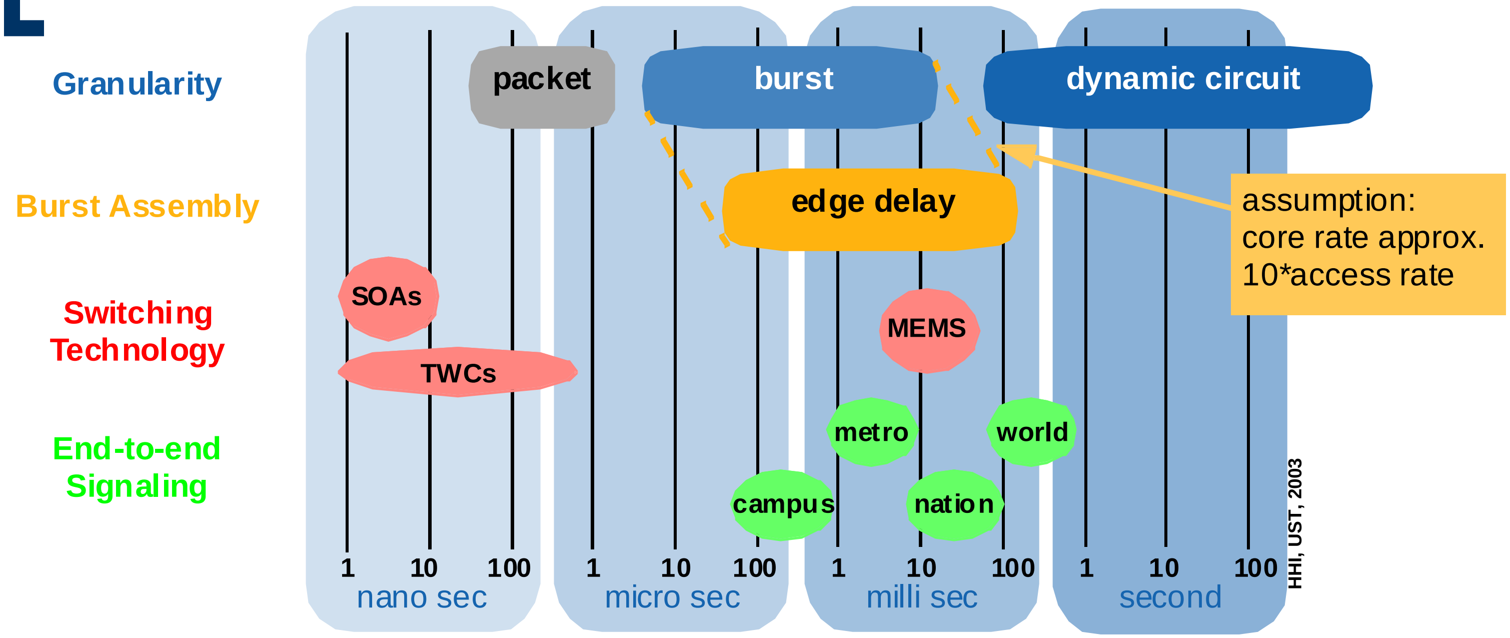
T_{off} : Offset time

P : Propagation time

$T_{off} \geq H^* \Delta$, H = 3

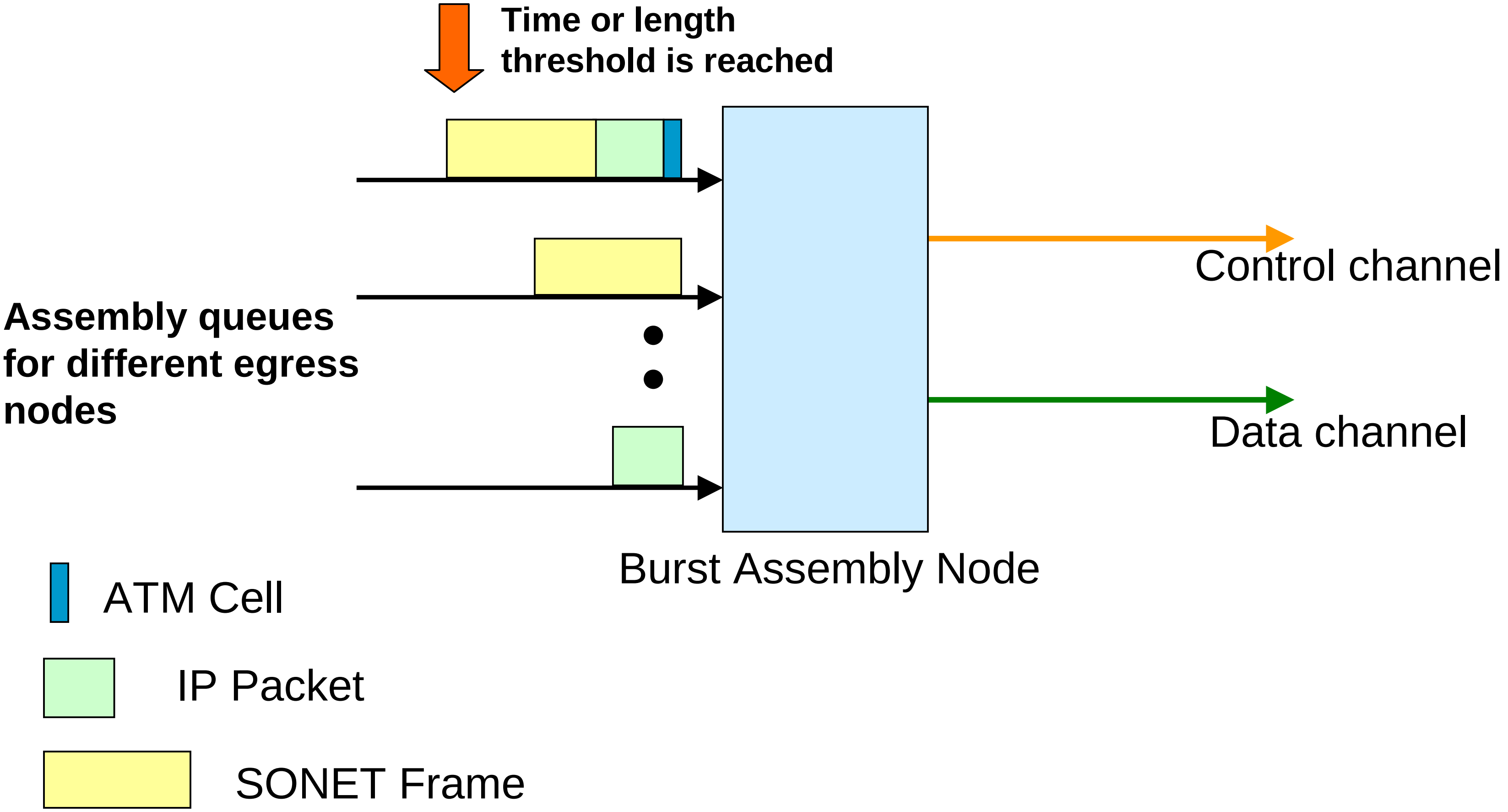
- $T_{off} \geq H\Delta$, H = 3
- T_{off} should be updated at each OXC i.e., $T'_{off} = T_{off} - \Delta$

OBS Timing Overview



- Granularity determines switching technology and vice versa
  → switching time << mean burst duration
- Tell-and-Wait: Granularity determines end-to-end signaling distance
  →end-to-end propagation < mean burst duration
- Access rate (assembly delay) determines granularity

Burst Assembly



Time or length threshold is reached

Assembly queues for different egress nodes

Control channel

Data channel

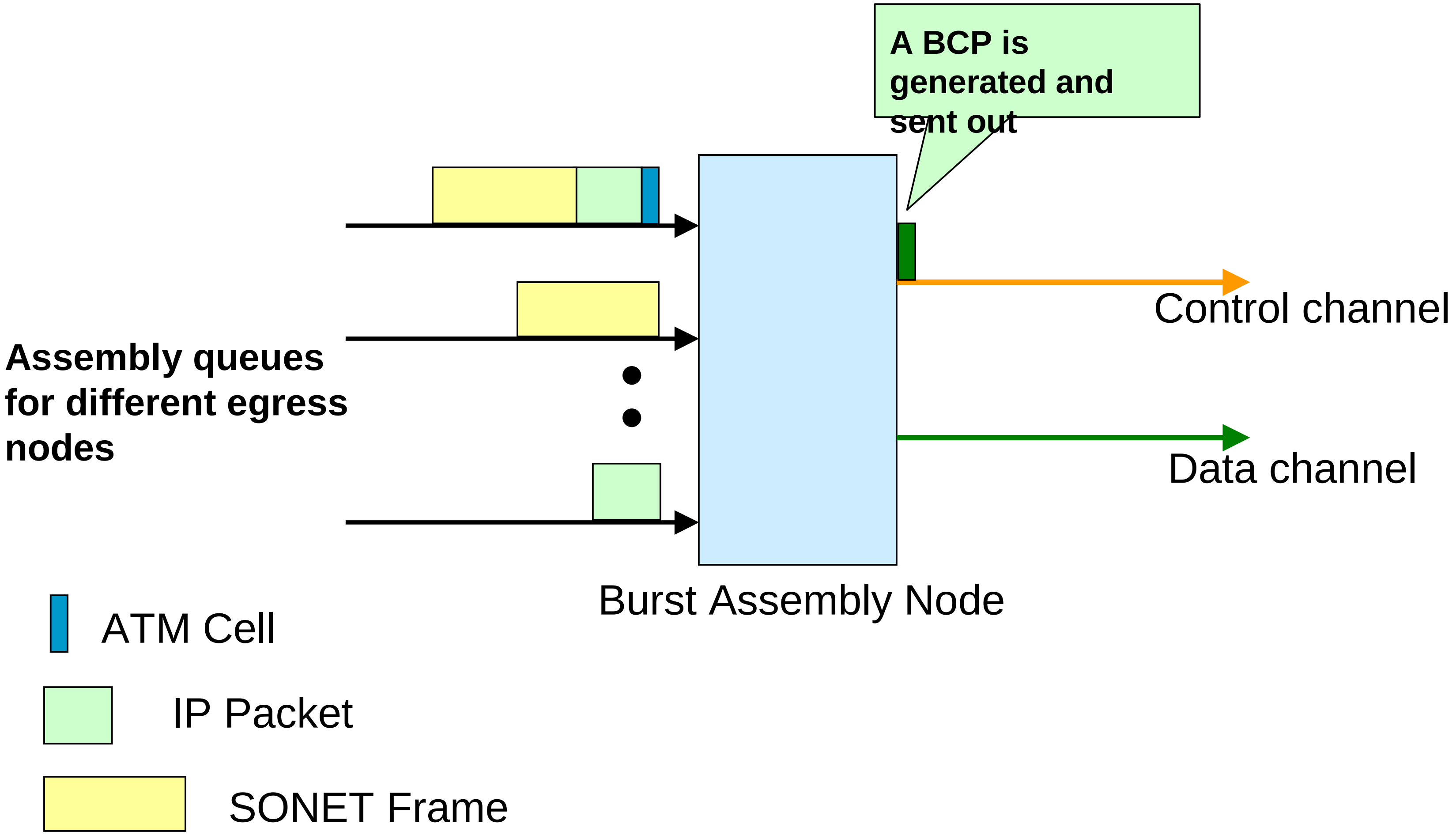
Burst Assembly Node

ATM Cell

IP Packet

SONET Frame

Burst Assembly



A BCP is generated and sent out

Control channel

Data channel

Assembly queues for different egress nodes

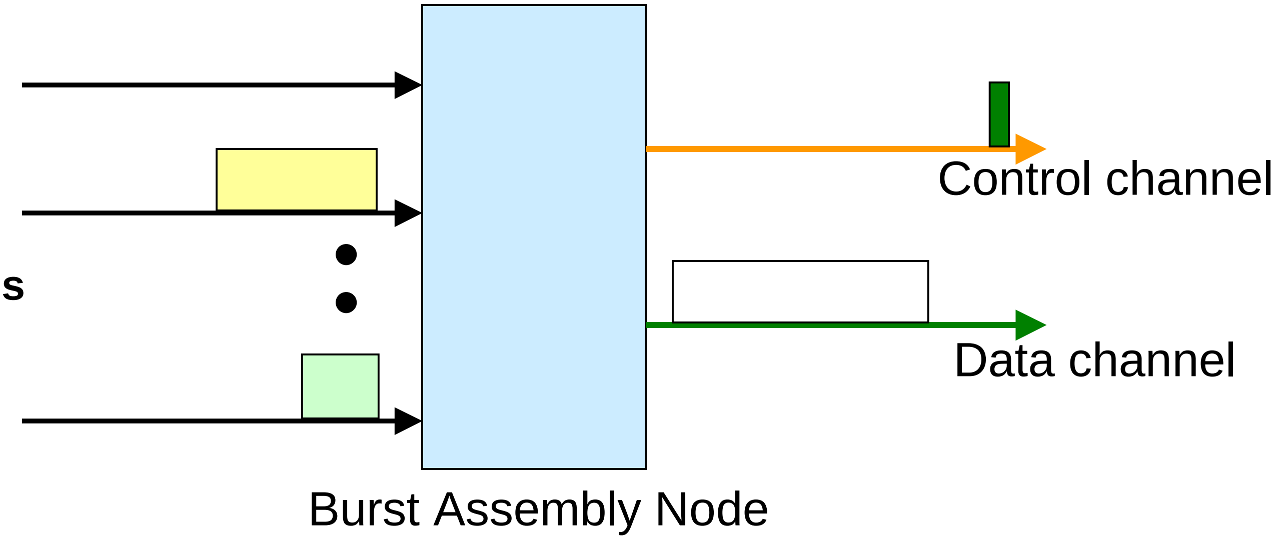
Burst Assembly Node

ATM Cell

IP Packet

SONET Frame

Burst Assembly



Assembly queues for different egress nodes

Control channel

Data channel

Burst Assembly Node

🟦 ATM Cell

🟩 IP Packet

🟨 SONET Frame

Edge Node

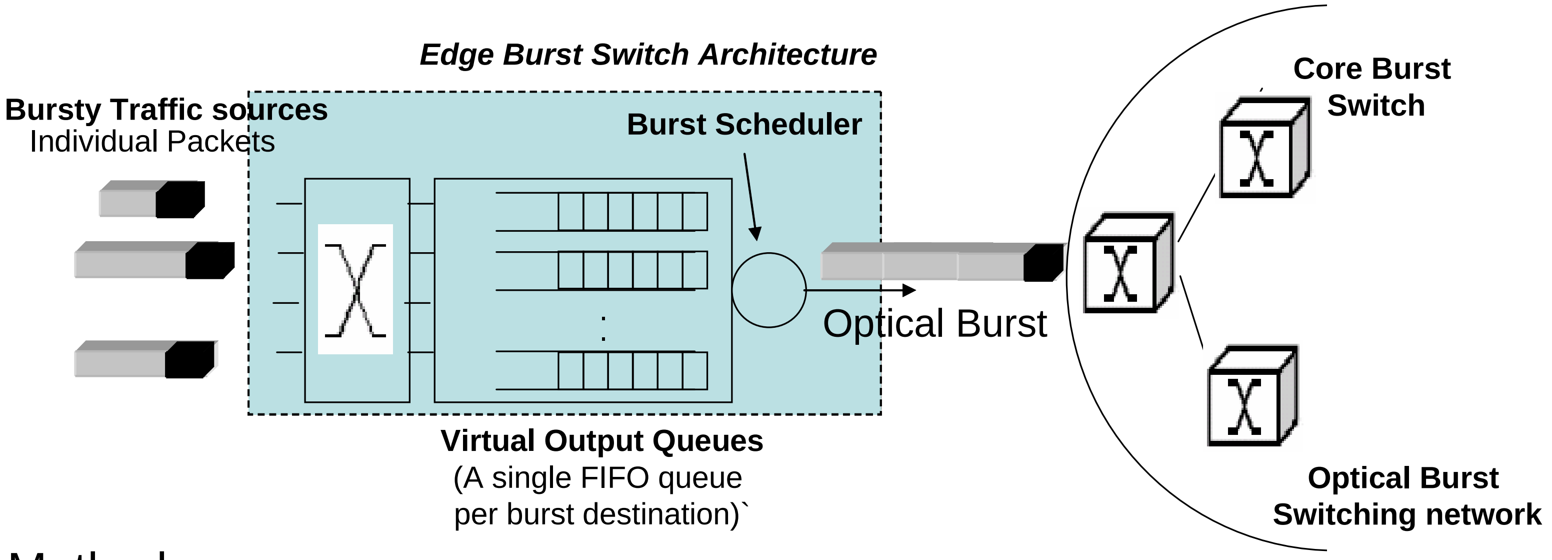
- Consists of electronic router and OBS interface
- Functions
  - Electronic data buffering and processing
  - Burst Aggregation (BA) responsible for collecting data from legacy networks and building the burst unit
    - impact on the overall network operation by the control of the burst characteristics
    - in order to reduce the burst loss probabilities in the network the aggregation function can segment data bursts for the purpose of their partially dropping in core nodes when contention occurs

Edge Node

- Setting up the pre-transmission offset time
  - in simple fixed offset scheme, the offset time is calculated as a sum of the total processing times at all the intermediate hops
  - offset time is one of the crucial OBS network parameters since its incorrect estimation has impact on data lost
- Sending the control packet
- Sending the burst

Edge Burst Switch Architectures

Architecture of a typical OBS ingress node



Assembly Methods:

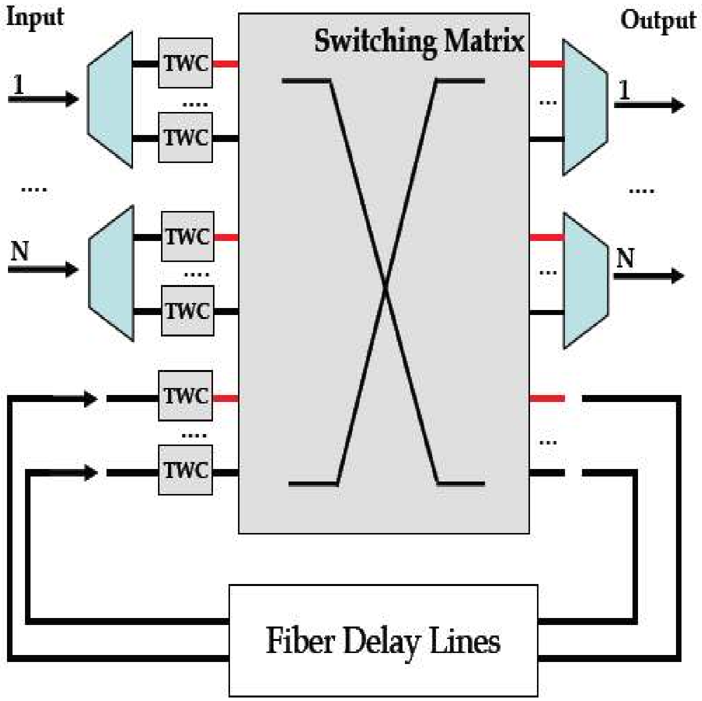
- Timer-based, where the burst is sent out after a time-out signal expires.
- Burst-length-based where the burst is sent out when it reaches a certain size
- Mixed timer / burst-length-based ones, where burst transmission time is constrained on both Time-out and Burst size criteria.
- Bursts also need to be larger than a minimum threshold so padding required

Core Node

- Hardware requirements
  - O/E/O conversion for header processing
  - λ-conversion
  - switching speeds fast enough
  - eventually optical buffering (FDLs)
- Operation
  - Processing of incoming control packets (electronically) and sending it to the next node that lays on the routing path
  - Reservation of optical resources for transferring the burst
    - Just-In-Time (JIT)
    - Horizon Reservation Mechanism (HRM)
    - Just-Enough-Time (JET) – the most efficient but of high complexity
  - Fast optical switching with wavelength conversion and optical buffering (when available and necessary)
  - Dealing with contention resolution (by a proper scheduling algorithm)

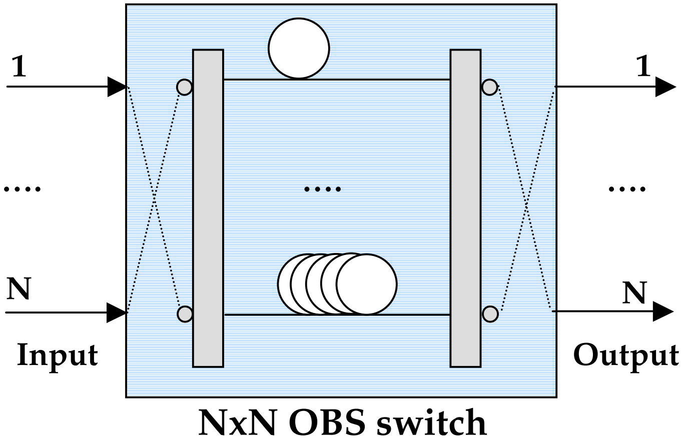
Core Burst Switch Architectures

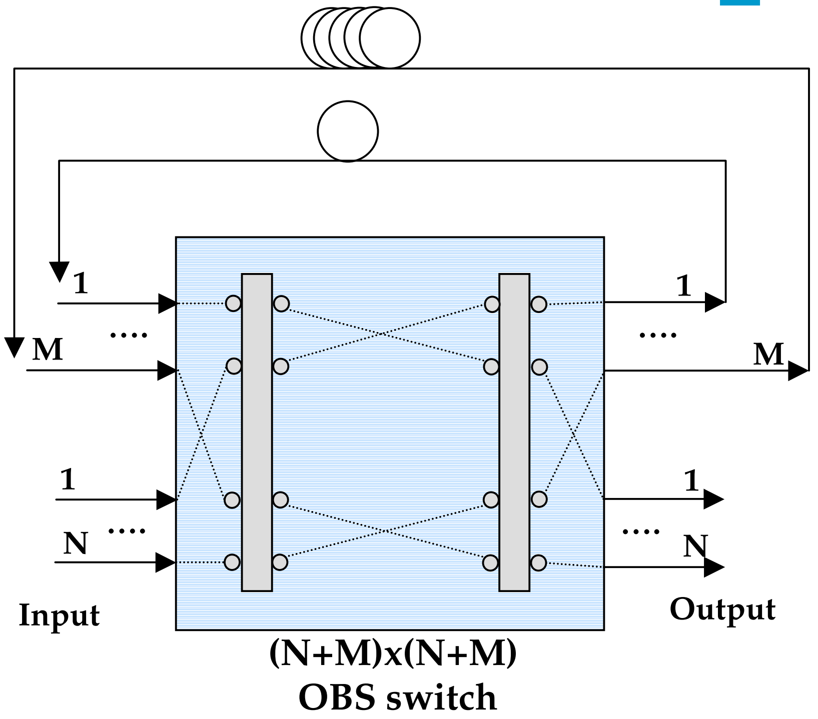
- Different nodes design: according to contention resolution mechanism
  - time domain: using Fiber-Delay-Lines (Feed-forward and/or Feedback)
  - wavelength domain via wavelength conversion
  - space domain via deflection to another fiber output



Example of OBS node using feedback FDLs and tunable wavelength converters.

FDL Architectures



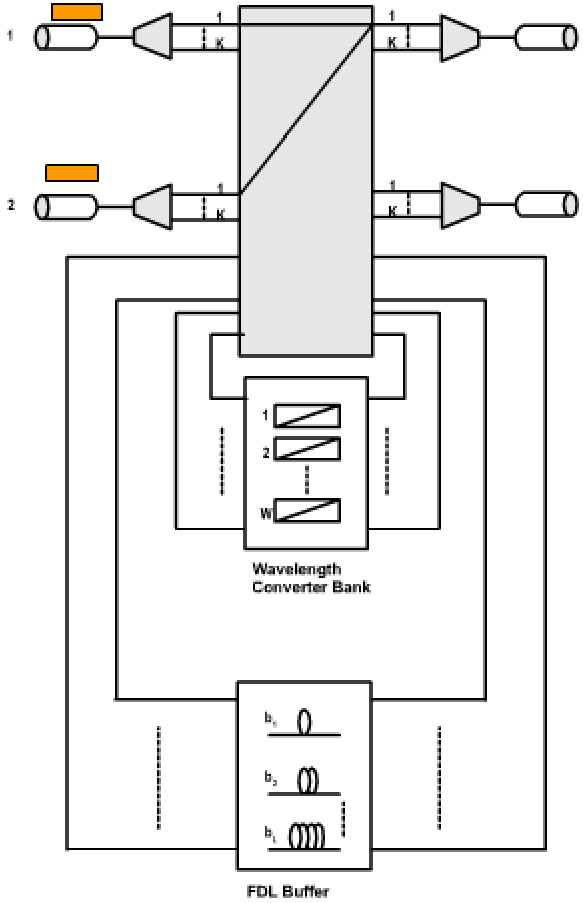


- In the feed-forward method, bursts are fed into fiber delay lines of different lengths and when they come out, they have to be switched out.
- In the feedback scheme, a burst may re-circulate as long as there is a bandwidth shortage at the output ports.

Contention Resolution

Performance evaluation of contention resolution schemes

Contention Resolution



Contention Resolution

- Burst loss possible due to statistical multiplexing
- Application of OBS in high-speed metro/core networks
  - lost data has to be retransmitted on end-to-end basis (e.g. TCP)
  - very low burst loss probability required

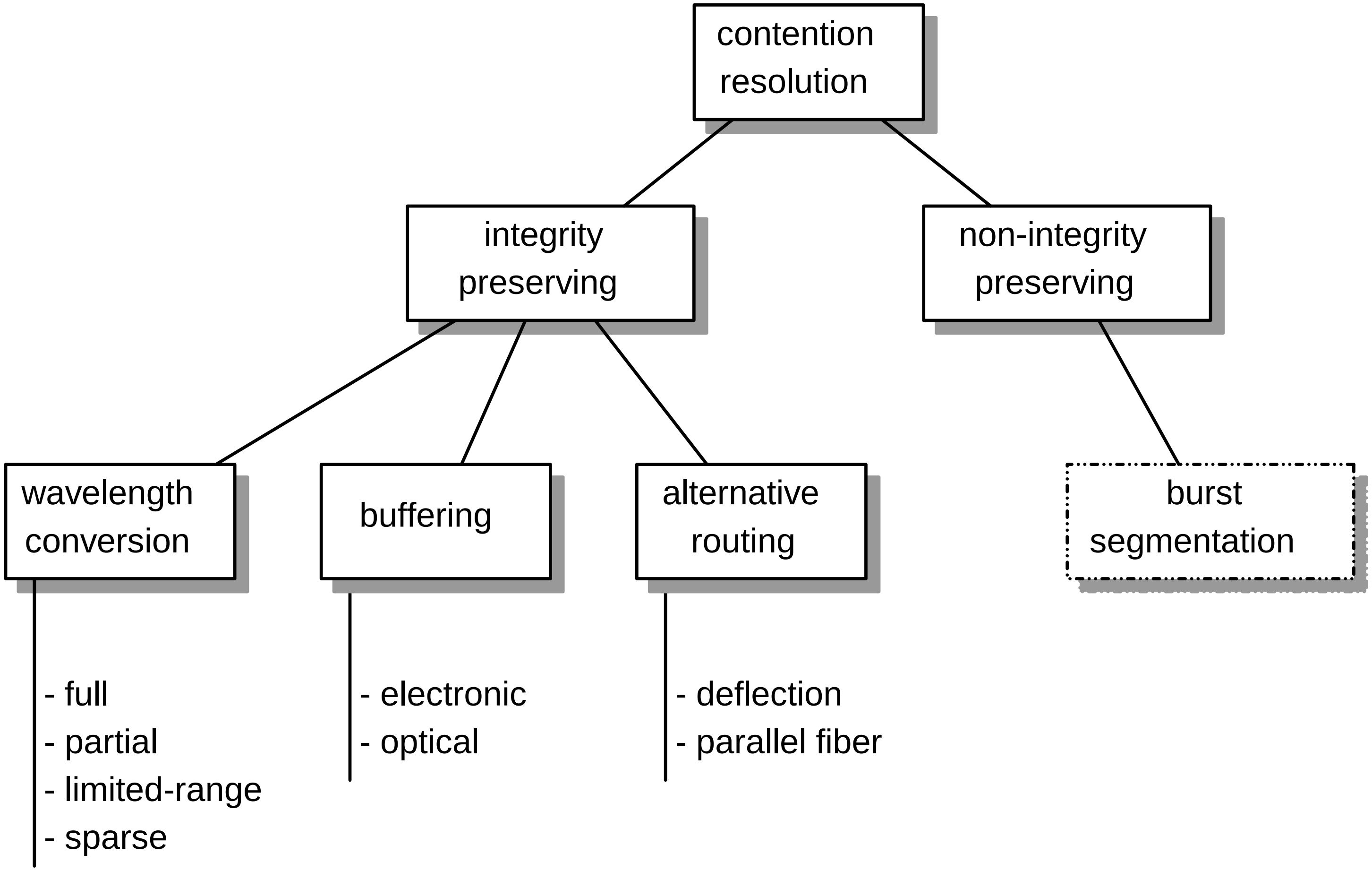
  → need for highly effective contention resolution

Domains of contention resolution

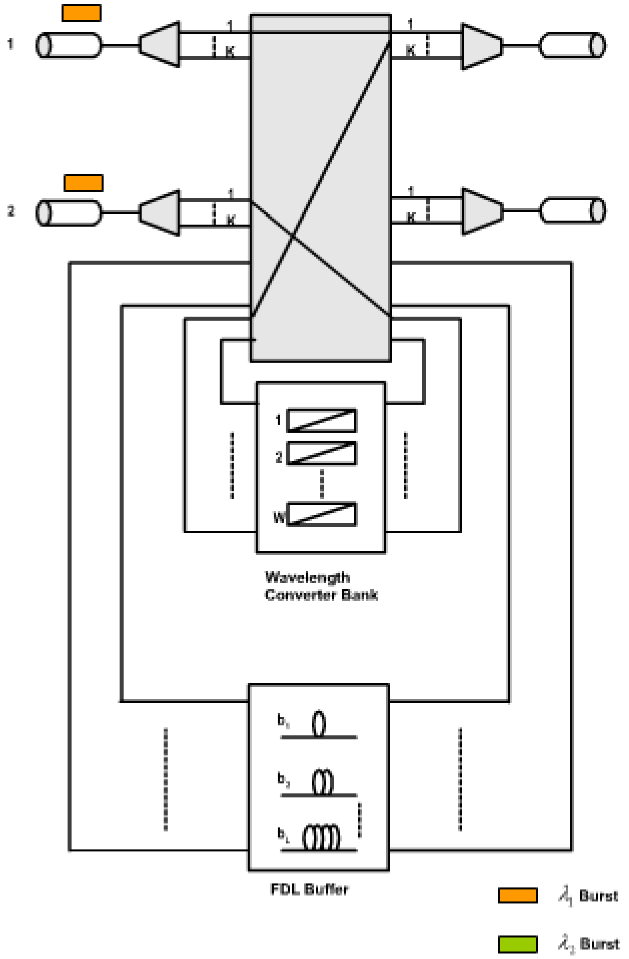
- Wavelength domain wavelength conversion
  - very effective as all WDM channels shared among all bursts
  - but: low burst loss probabilities only for many λs

  → additional schemes necessary, combinations beneficial
- Time domain fiber delay lines (FDLs)
- Space domain deflection/alternative routing
- Segmentation only conflicting part of burst dropped

Contention Resolution: Classification



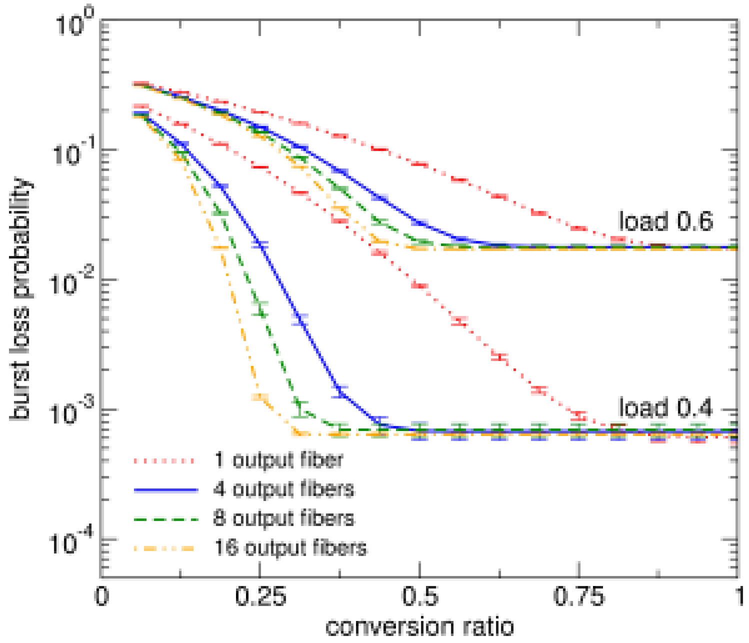
Wavelength Conversion



Full/No/Partial Wavelength Conversion

- Wavelength conversion is the most readily available contention resolution method
- In Full Wavelength Conversion (FWC), a burst arriving at a certain wavelength can be switched onto any other wavelength towards its destination
- No Wavelength Conversion (NWC)
  - Wavelength continuity constraint
- In cost-conscious Partial Wavelength Conversion (PWC), there is a limited number of converters
  - Consequently, some bursts cannot be switched towards their destination (and therefore blocked) when all converters are busy despite a free channel

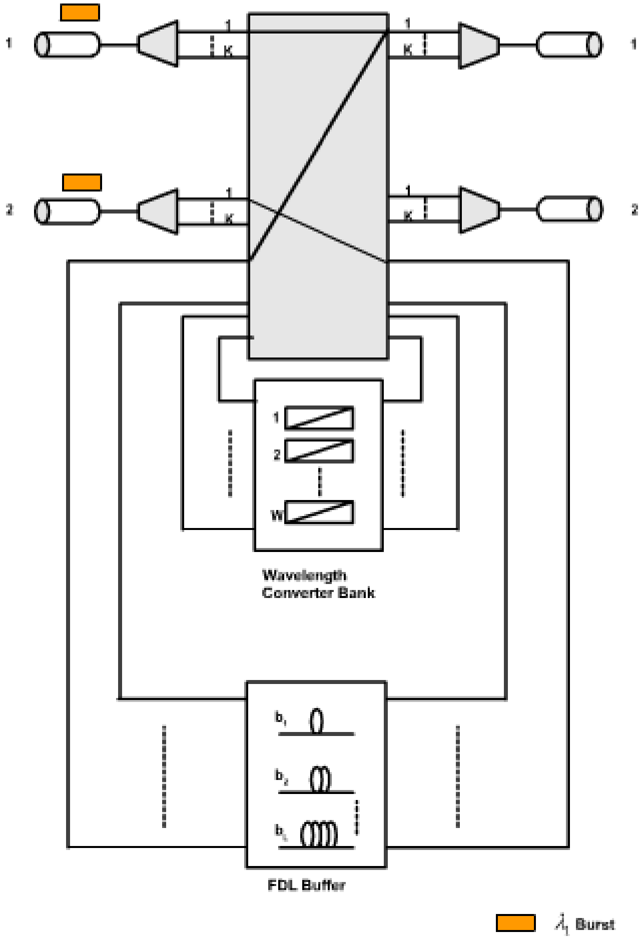
Case Study: Shared Conv Pool



- Share-per-node converter pool
- M = 16 wavelengths per fiber
- N output fibers
- C converters
- Conversion ratio = C / (M*N)

- Significant converter savings possible
  - → For realistic load values up to 50-75% savings possible
- N=1 is case of share-per-output pool

Use of FDLs

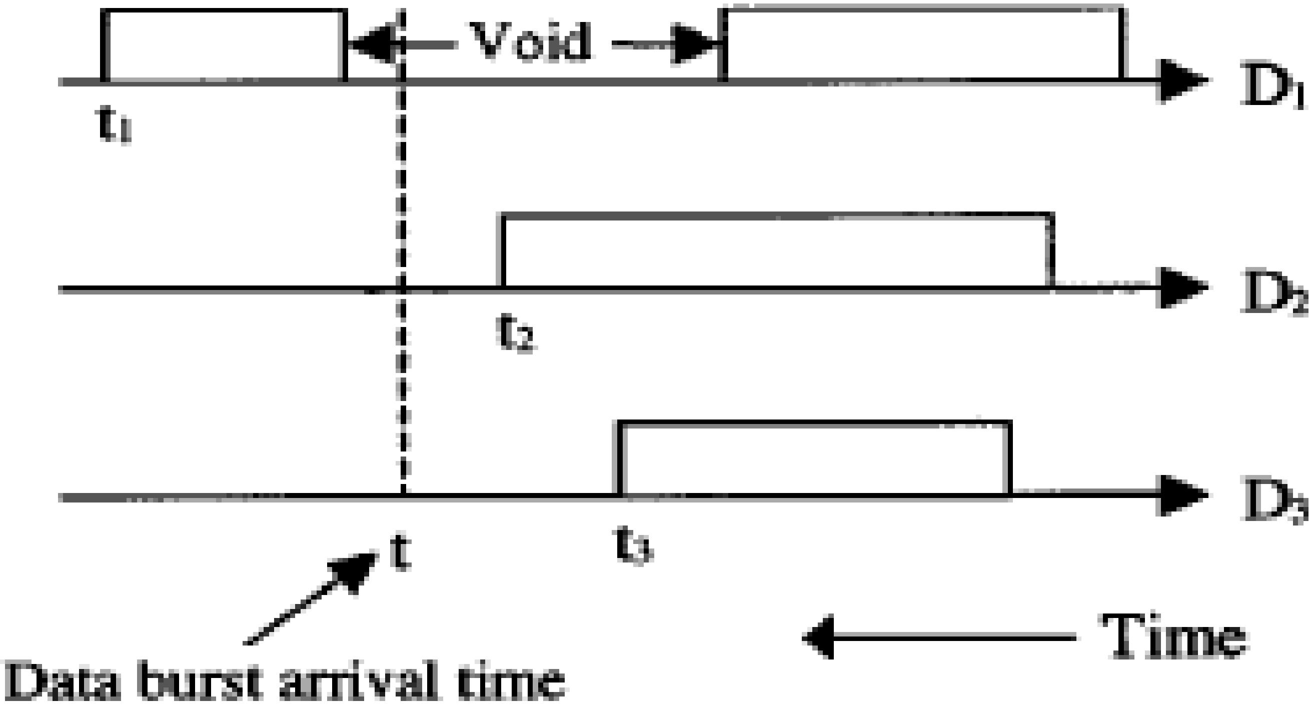

1
1
K
1
K
1
2
1
K
1
K
2
1
2
W
Wavelength
Converter Bank
b_1
b_2
b_L
FDL Buffer
λ_1 Burst

SCHEDULING ALGORITHMS

Channel Scheduling

- Problem of assigning a burst to a channel when it gets information about when it will arrive.
- Ideally we assign bursts to a channel that becomes free just before the burst arrives.
- This minimizes idle time (voids) and helps for scheduling later by maintaining maximum flexibility for later bursts.

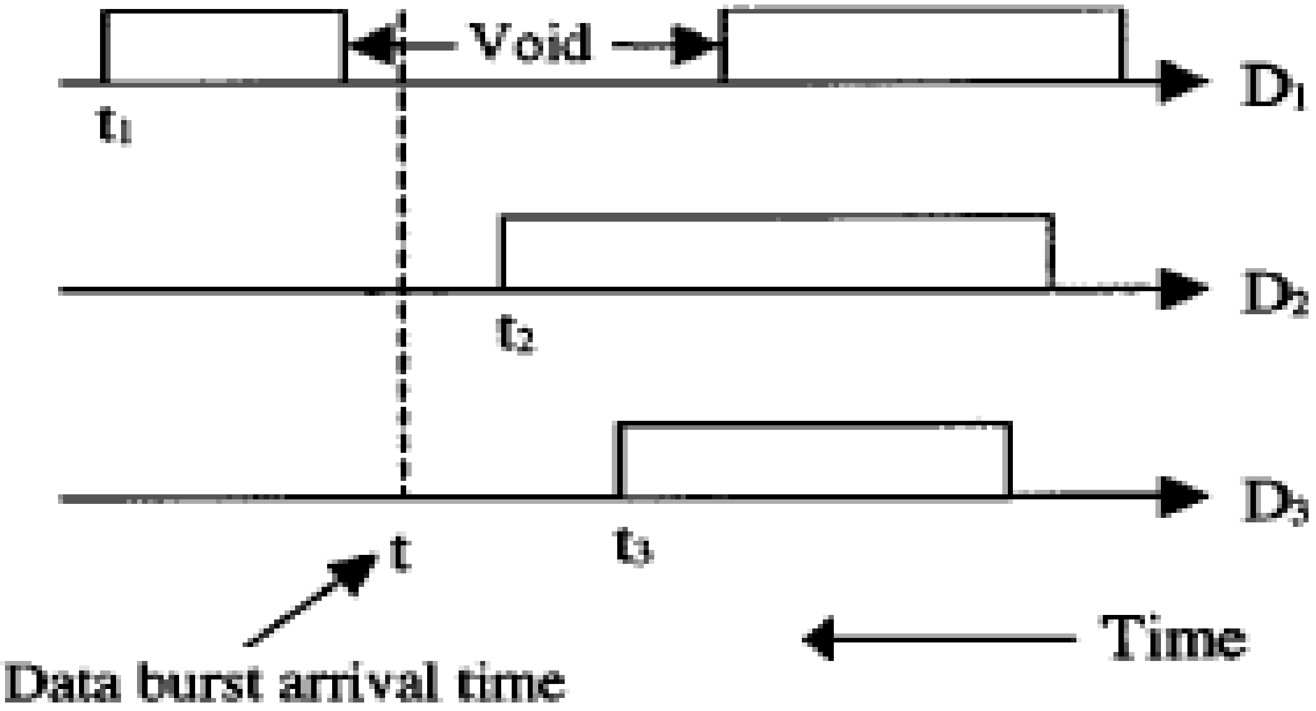
Channel Scheduling Example



Horizon Channel Scheduling

- Maintains a single horizon for each channel on the link.
- A horizon is the time after which no reservation has been made on the channel.
- Among the channels that have a horizon earlier than the arrival time of the burst, use the channel which has the latest horizon
- Find the channel with the smallest gap
- Once a channel is selected, the scheduler computes the new scheduling horizon of that channel

Example of Horizon Scheduling



Channels 2 and 3 are available.
Channel 2 is selected since it has a smaller gap

Horizon Shortcomings

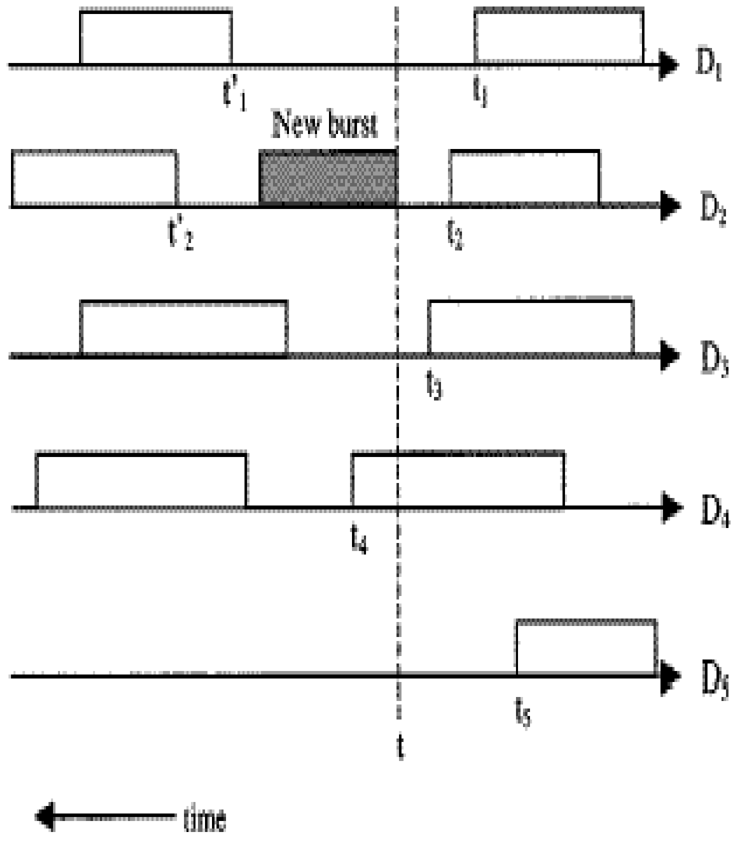
- Horizon cannot schedule any bursts within the voids since the scheduling horizon just gives the time of the end of the burst.
- There is low channel utilization and high loss rate since Horizon discards all bursts that can fit in the void intervals.

LAUC-VF

- Idea is to minimize voids by selecting the latest available unused data channel for each arriving data burst.
- The scheduler first finds all outgoing data channels that are available for the burst (t, t+L).
- If at least one channel is available, the scheduler selects the latest one. That is the channel having the smallest gap between t and the end of the last data burst before t.

LAUC-VF Example

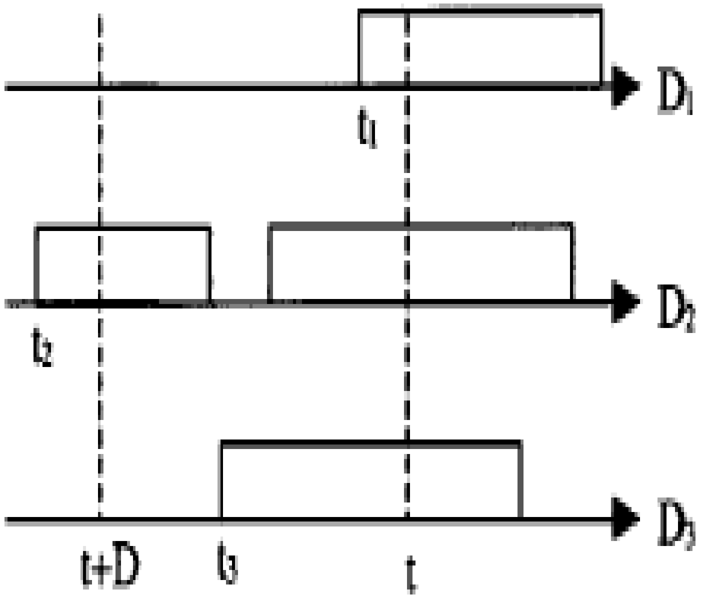
- Burst arrives at time t
- Channels D3 and D4 cannot fit the burst
- Channels D1, D2 and D5 are considered
- D2 is chosen since the gap between the last burst on the channel and the new burst is minimized
- Horizon chooses D5



No Available Channels

- If all the channels are unavailable at the time that a burst arrives, LAUC-VF scheduler checks the usage of FDLs
- D = FDL unit delay
- LAUC-VF scheduler finds the minimum i such that the channel is available for [t+iD,t+iD+L]
- If such an i cannot be found, the burst cannot be scheduled

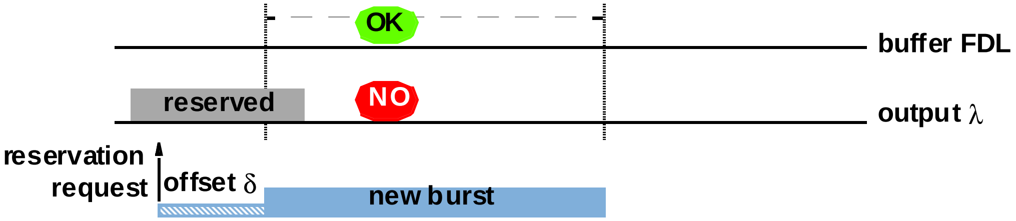
Example of use of FDLs



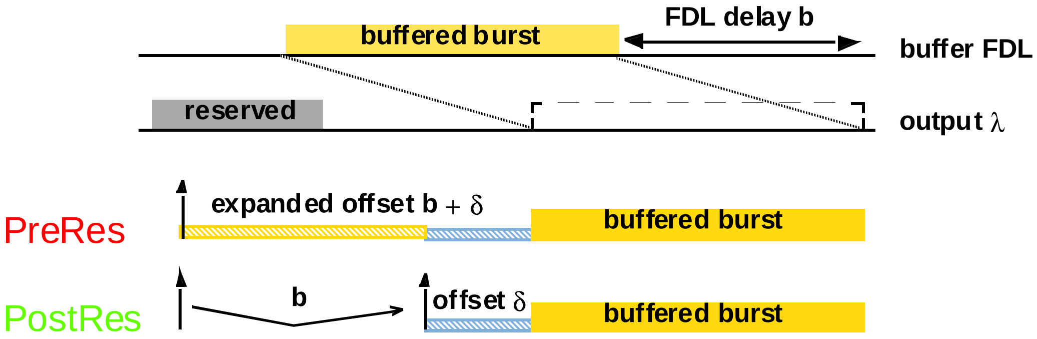
- At time t when the burst arrives, no channels are available.
- The burst is sent through an FDL and arrives for scheduling at t+D.
- Now a channel is available and a burst can be scheduled.

FDL Buffer Scheduling

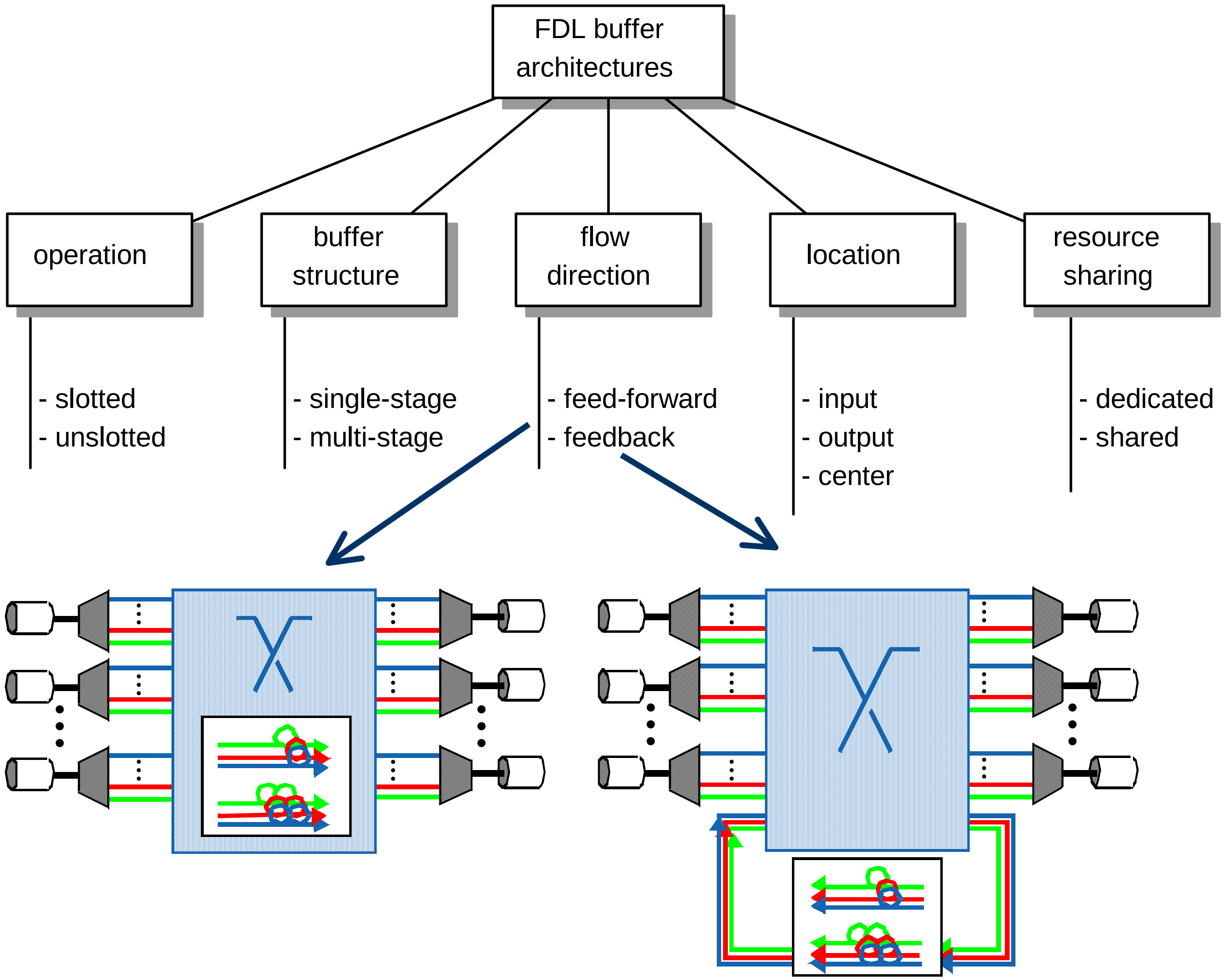
- **Use of FDL buffer in case of blocked output wavelength**



- **Reservation of buffer FDL and output wavelength**
  - PreRes: early reservation of output together with buffer, no reloop
  - PostRes: reservation of output only after buffering time b



FDL Buffer Architectures



FDL buffer architectures

operation
- slotted
- unslotted

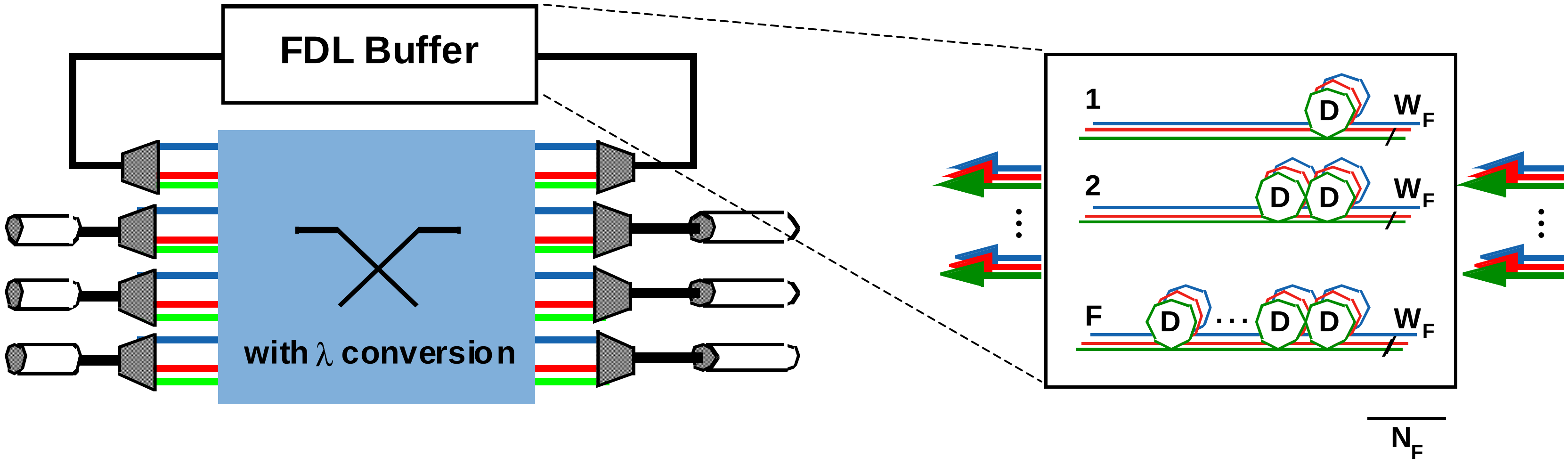
buffer structure
- single-stage
- multi-stage

flow direction
- feed-forward
- feedback

location
- input
- output
- center

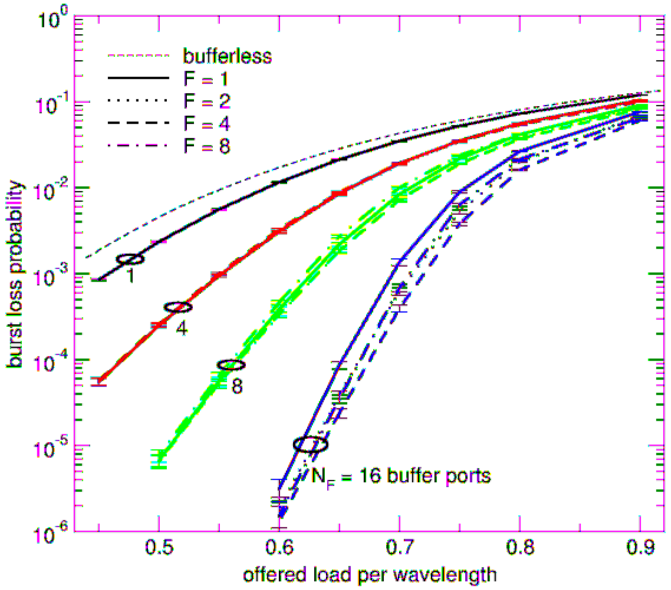
resource sharing
- dedicated
- shared

FDL Buffer Dimensioning



Number of FDLs in buffer:	F
Number of wavelengths per FDL:	W_F
Total number of FDL buffer ports:	$N_F = F * W_F$
Delay granularity:	D
Degenerate FDL buffer, delay of ith FDL:	$i * D$

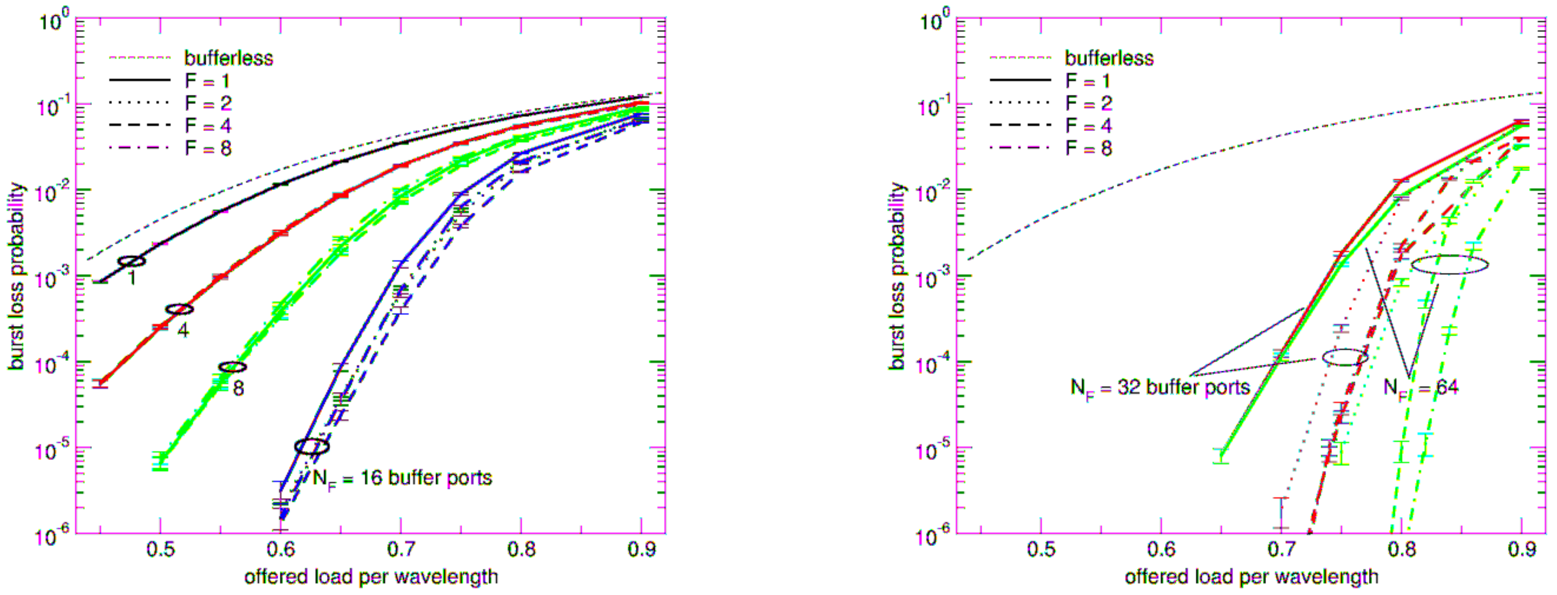
FDL Buffer Dimensioning



- **16 wavelengths per fiber**
- **4 input/output fibers**
- **Share-per-node FDL buffer**
- **Degenerate FDL buffer**
- **D = 4 mean burst transmission times**
- **Conversion always tried first**
- **Poisson arrivals**

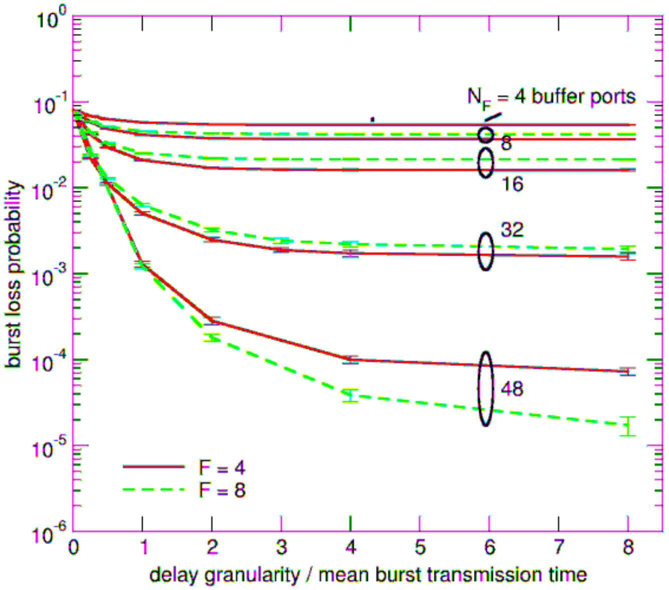
- Increasing number of buffer ports N_F substantially reduces losses
- At lower load FDLs are more efficient
- Up to N_F = 16: minor impact of FDL buffer architecture (number of buffer FDLs), mainly the total number of buffer ports decides on performance

FDL Buffer Dimensioning



- Increasing number of buffer ports N_F substantially reduces losses
- Up to N_F = 16: minor impact of FDL buffer architecture (number of buffer FDLs), mainly the total number of buffer ports decides on performance
- For N_F > 16: FDL buffer architecture becomes relevant! More buffer FDLs F (with less wavelengths W_F each) yield lower losses.

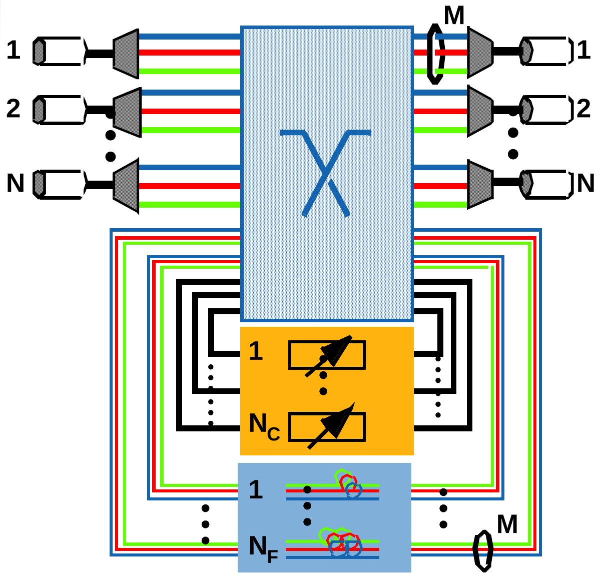
FDL Buffer Dimensioning

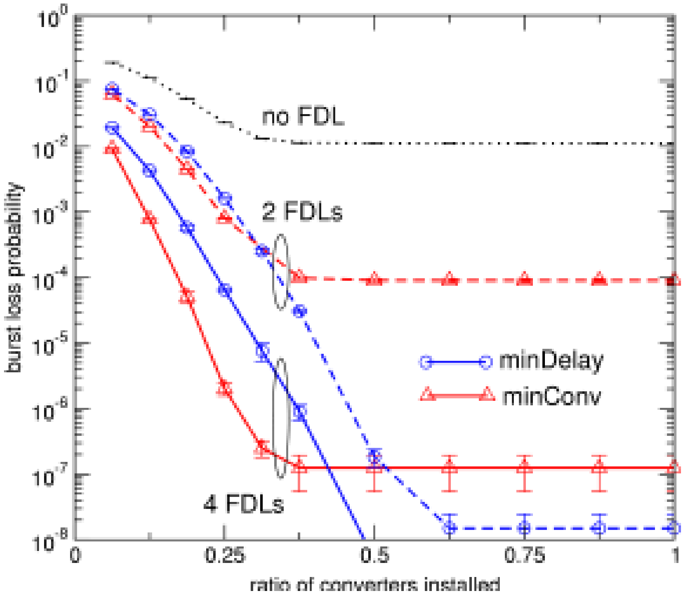


- **16 wavelengths per fiber**
- **4 input/output fibers**
- **Share-per-node FDL buffer**
- **Degenerate FDL buffer**
- **Conversion always tried first**
- **Poisson arrivals**
- **Offered load per wavelength: 0.8**

- For very few buffer ports: hardly any impact of delay granularity
- For more buffer ports: approx. 4 mean burst transmission times lead minimal loss probability
- Dependence on FDL delay quite insensitive to number of FDLs F

Case Study: Optimized CR

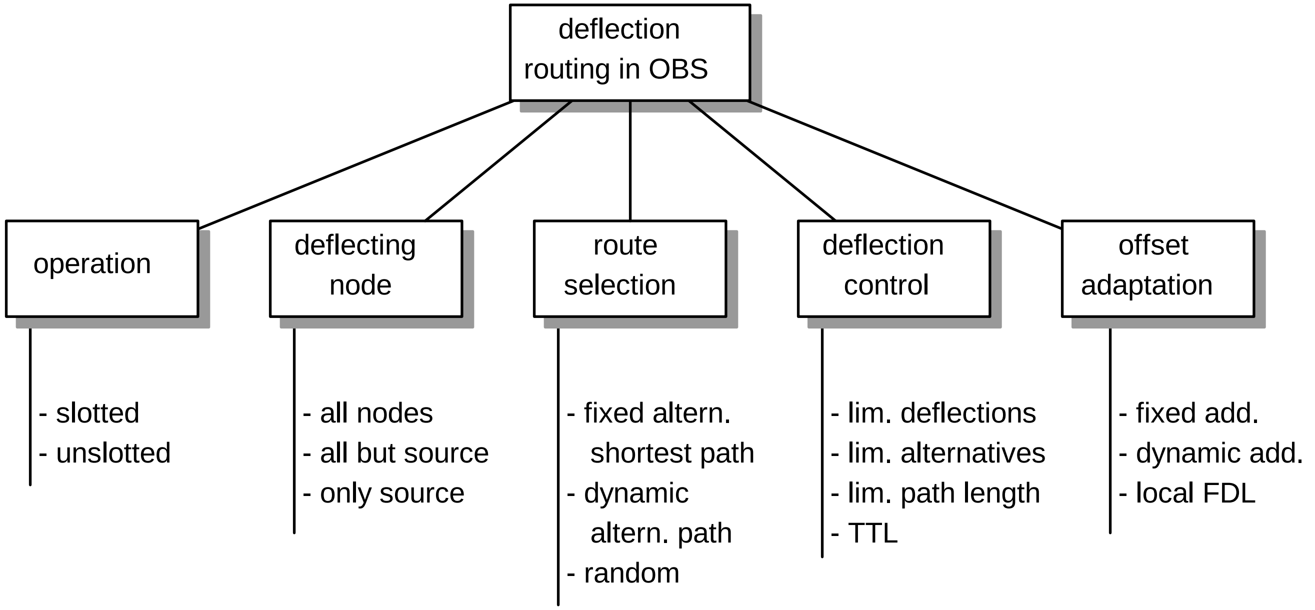




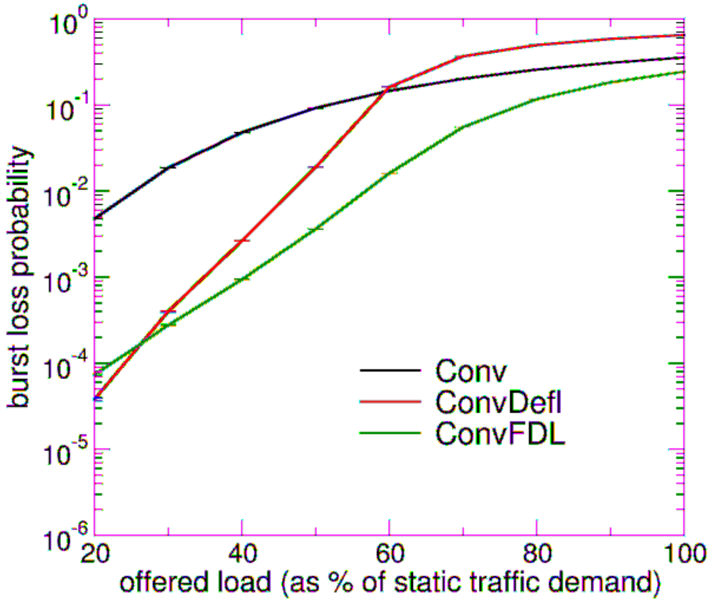
- Share-per-node
- M = 8 wavelengths
- N = 8 output fibers
- N_C tun. converters
- N_F FDLs
- $r_C=N_C/(MN)$

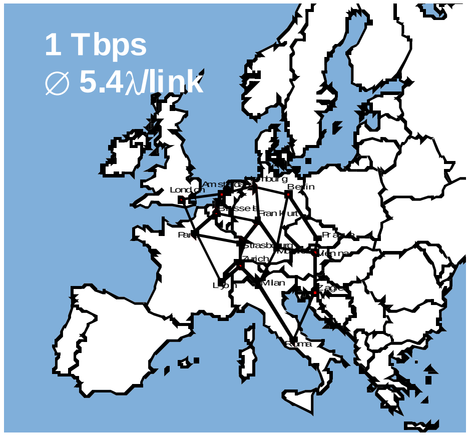
- Combination of FDL buffers and shared converter pools
- Strategy for selecting resources has significant impact
  - Red: strategy that minimizes converters (minConv)
  - Blue: strategy that minimizes FDL usage (minDelay)

Deflection Routing: Classification



Case Study: FDL—Deflection

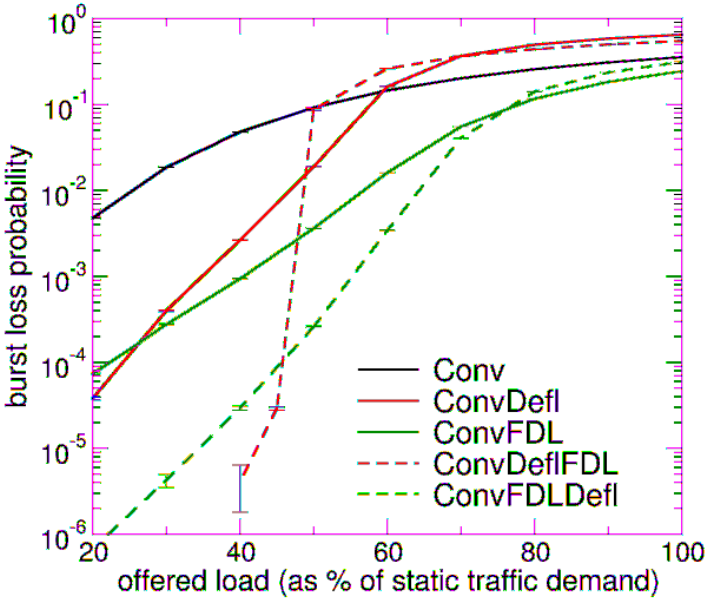


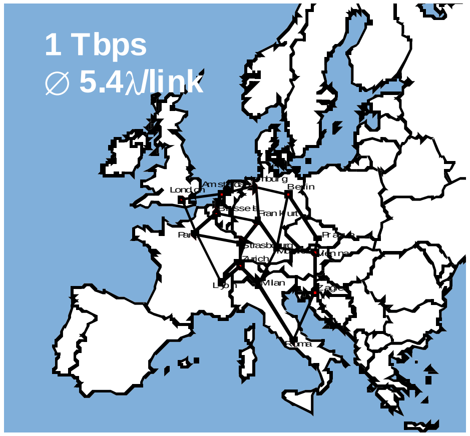


COST 266 CT dimensioned
for 2004 traffic demands:
in average 5.4 λ/link

- Comparison of conversion only (Conv) and combinations with FDL buffer (ConvFDL) and deflection routing (ConvDefl)
- ConvDefl suffers for high load from deflected bursts on longer detour routes (positive feedback on offered load)
- ConvFDL yields lower losses than Conv and ConvDefl until low loads
- For low load, improvement of ConvFDL limited by congestion in few nodes

Case Study: FDL—Deflection





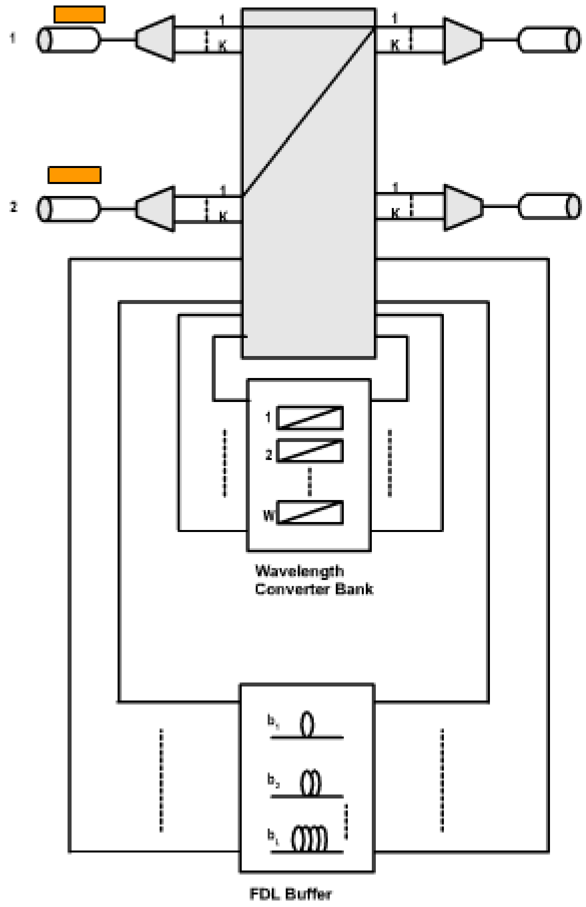
COST 266 CT dimensioned
for 2004 traffic demands:
in average 5.4 λ/link

- Comparison extended to combinations with FDL buffer and deflection routing: ConvDeflFDL and ConvFDLDefl
- ConvDeflFDL still suffers for high load; substantial improvements for lower load → behavior toggles between two states, stability?!
- ConvFDLDefl: deflection resolves residual congestion after FDL buffering

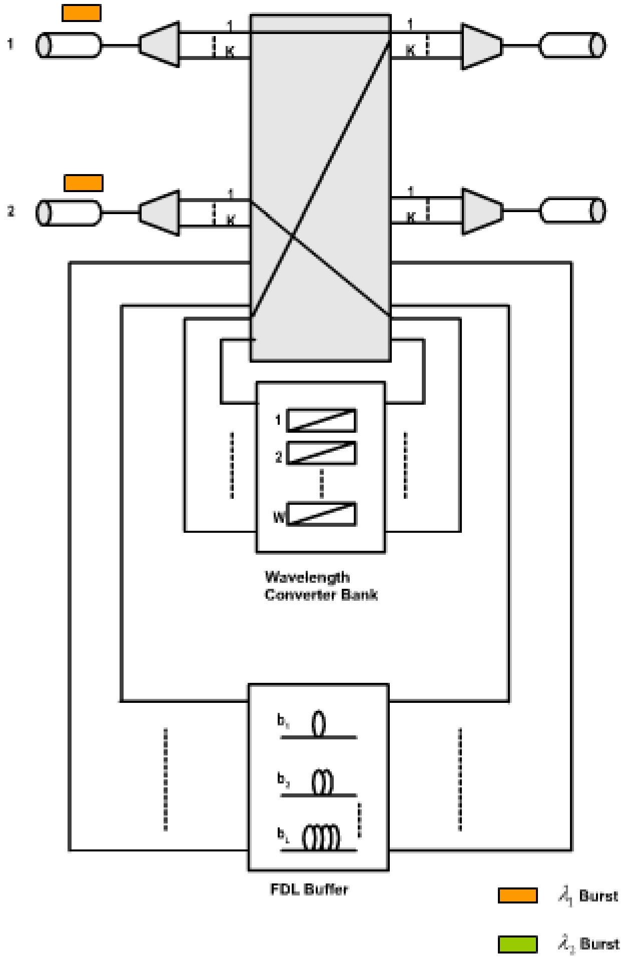
OBS tutorial

Teletraffic Modeling of OBS networks

Contention Resolution



Wavelength Conversion



Full/No/Partial Wavelength Conversion

- Wavelength conversion is the most readily available contention resolution method
- In Full Wavelength Conversion (FWC), a burst arriving at a certain wavelength can be switched onto any other wavelength towards its destination
- No Wavelength Conversion (NWC)
  - Wavelength continuity constraint
- In cost-conscious Partial Wavelength Conversion (PWC), there is a limited number of converters
  - Consequently, some bursts cannot be switched towards their destination (and therefore blocked) when all converters are busy despite a free channel

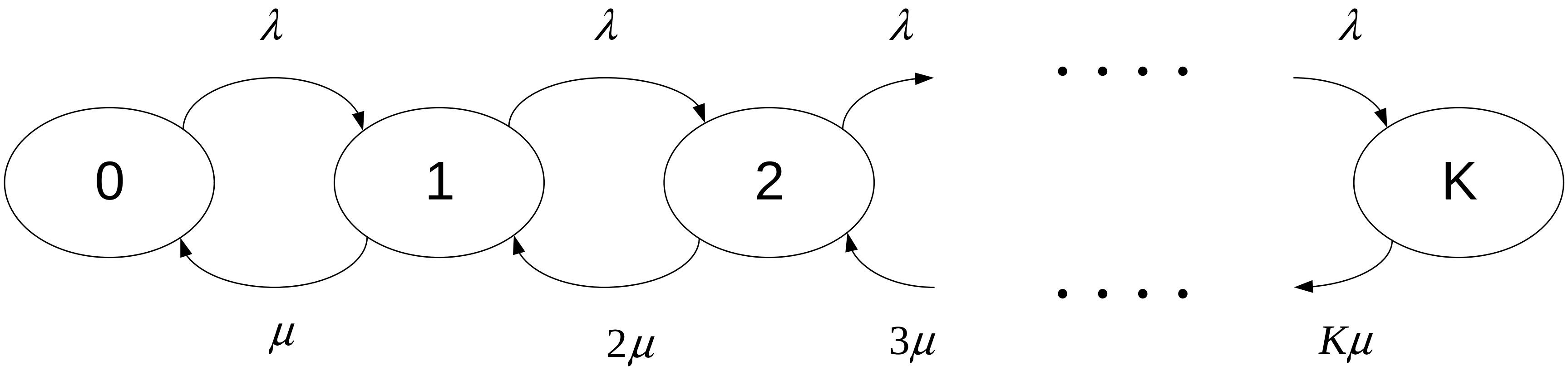
Full-range/Limited-range Wavelength Conversion

- Full Range TWCs do not have any tuning range limit and they can convert an incoming wavelength to any other wavelength.
- In limited-range wavelength conversion, a burst arriving on a wavelength can be converted to a fixed set of wavelengths probably above and below the original wavelength, i.e., limited-Range TWCs
  - Waveband switching
  - Circular conversion
  - Etc.

Full Wavelength Conversion

- *K* wavelength channels per fiber
- Burst arrival process is Poisson with rate λ
- The wavelength channel they arrive on is uniformly distributed on *(1,K)*
- Burst durations are exponentially distributed with mean $1/\mu$
- Offered load λ/μ

M/M/K/K Model



- Erlang-B formula describes the burst loss rate

$$P_{loss} = \frac{q^{c} / K!}{\sum_{i=0}^{K} q^{i} / i!}, q = \lambda / \mu$$

Implications

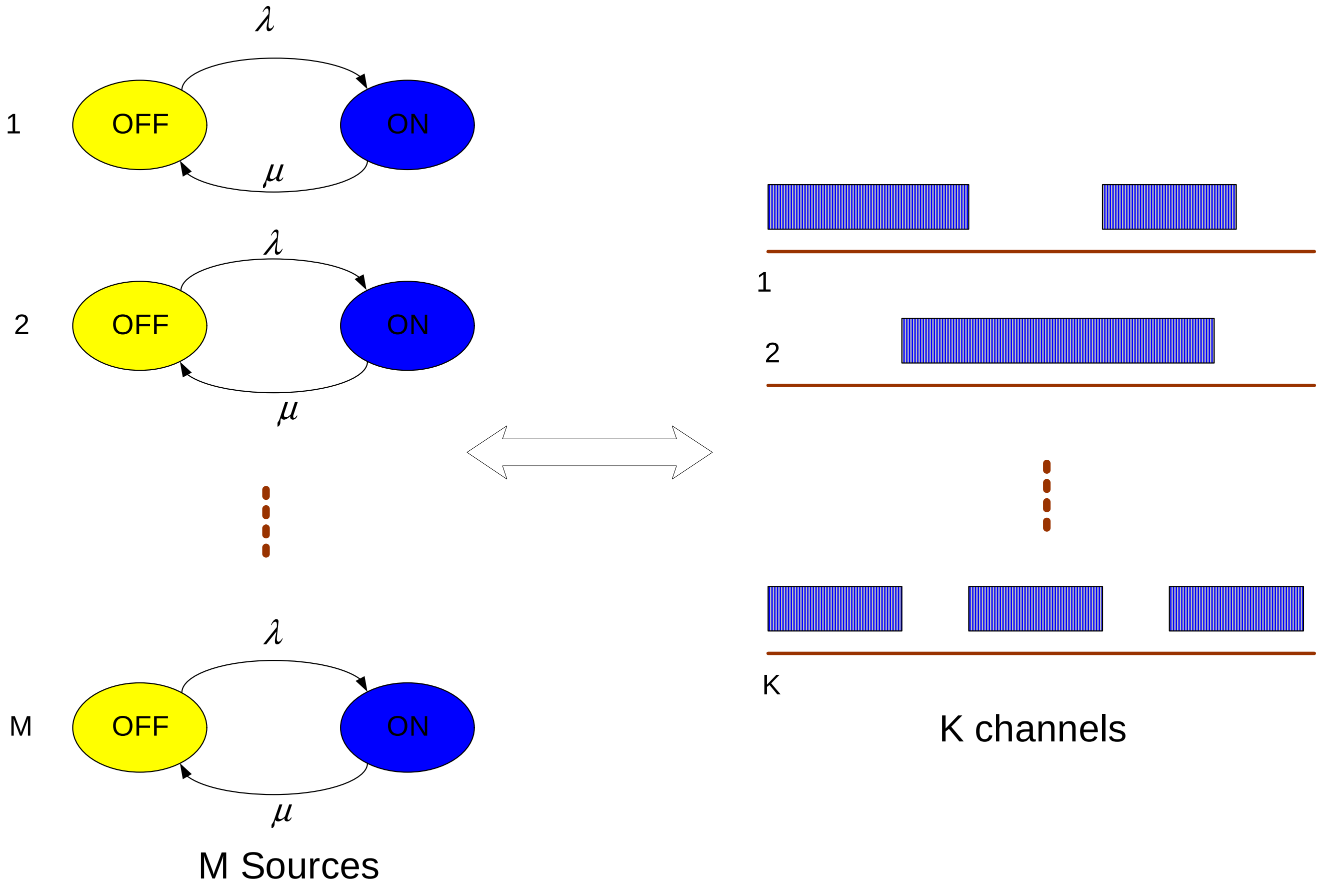
Throughput of the system for a given desired loss probability P_{loss}

P_{loss} /K	16	32	64	128
10^{-6}	22.4	37.9	52.4	64.8
10^{-5}	27.2	42.8	56.9	68.4
10^{-4}	33.4	48.9	62.2	72.7
10^{-3}	42.0	56.8	68.9	78.1
10^{-2}	54.9	68.2	78.2	85.5
10^{-1}	75.9	85.0	91.0	94.8

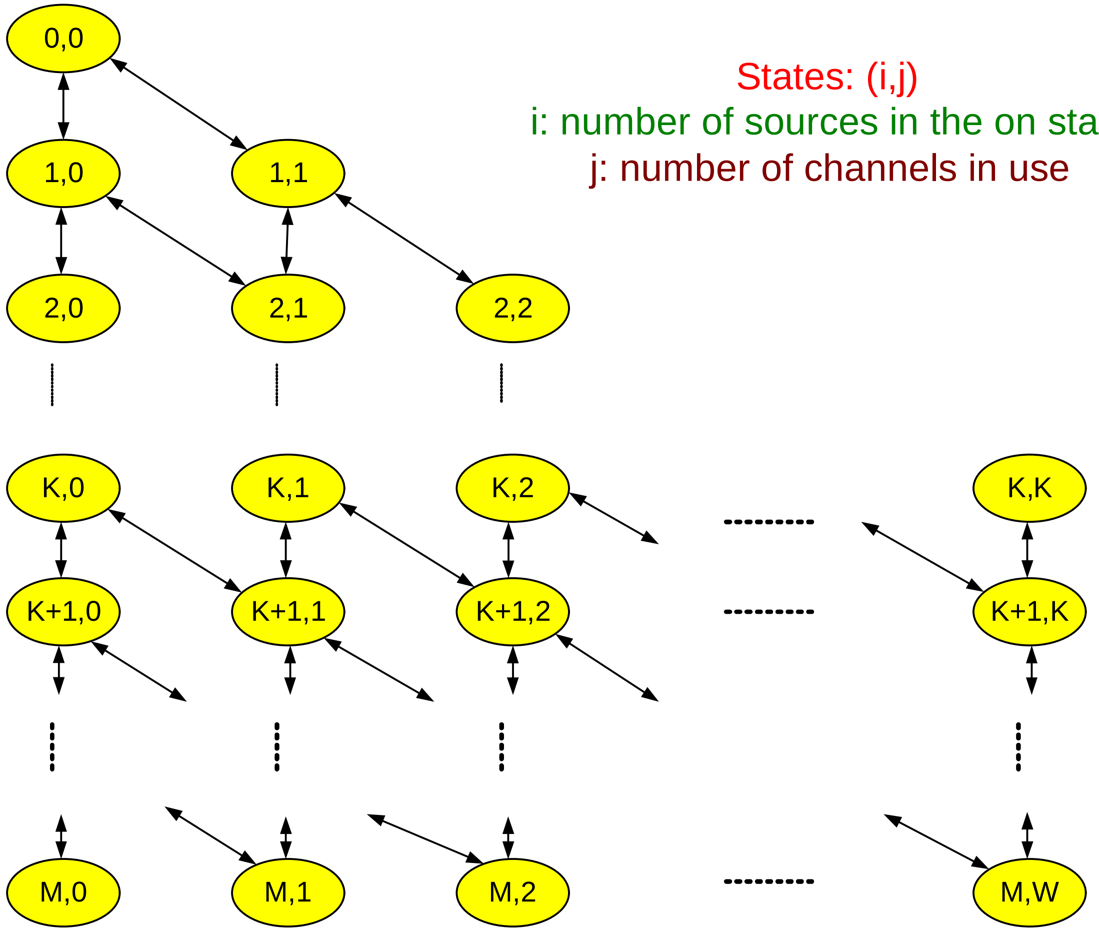
Implications

- The throughput of OBS networks is especially low for real time traffic with stringent QoS requirements
  - Around %50 throughput is achievable with current systems
  - The number of wavelength channels has to be very large to reach burst loss probabilities in the order of 10^{-5} or less, e.g., 350 wavelength channels are needed to carry a load of 0.8 Erlang per wavelength channel at this loss rate.
- The throughput of OBS networks is relatively higher for TCP traffic for which the operating loss rates are much higher.
- The situation gets to change for more realistic traffic models
  - Traffic is not generated by an infinite population but rather a finite population, e.g., on-off traffic models
  - Autocorrelation in network traffic

On-off Traffic Modeling



Two-dimensional Markov Chain



States: (i,j)

i: number of sources in the on state

j: number of channels in use

Non-homogenous Quasi Birth Death (QBD) Process

$$Q = \begin{bmatrix} A_0 & U_1 & & & \\ D_0 & A_1 & U_2 & & \\ & D_1 & A_2 & \ddots & \\ & & \ddots & \ddots & U_M \\ & & & D_{M-1} & A_M \end{bmatrix}$$

$$xQ = 0, x = \begin{bmatrix} x_0 & x_1 & \cdots & x_M \end{bmatrix}, \sum_i \sum_j x_{i,j} = 1$$

$$P_{loss} = \frac{\sum_{j=K}^{M} x_{j,K} (M-j)\lambda}{M \dfrac{\lambda\mu}{\lambda+\mu}}$$

Block-Tridiagonal LU Factorizations

$$Q = \begin{bmatrix} I & & & & \\ L_0 & I & & & \\ & L_1 & I & & \\ & & \ddots & \ddots & \\ & & & L_{M-1} & I \end{bmatrix} \begin{bmatrix} F_0 & U_1 & & & \\ & F_1 & U_2 & & \\ & & F_2 & \ddots & \\ & & & \ddots & U_M \\ & & & & F_M \end{bmatrix}$$

- Computational complexity *O(M K*^{3}*)*
- Storage requirement of *O(M K*^{2}*)*
- Gains of order *O(M*^{2}*)* (computation) and *O(M)* (storage) against the brute force approach

Example

- Poisson case
  - *K* = 32 channels
  - ρ =0.4885 → P_{loss} = 10^{-4}
- Finite population case
  - *# users M* varied
  - Overall load fixed to ρ =0.4885

M *#users*	P_{loss}
32	0
48	5.1 10^{-7}
64	3.9 10^{-6}
128	2.7 10^{-5}
256	5.9 10^{-5}
1024	8.7 10^{-5}
4096	9.6 10^{-5}
∞	1 10^{-4}

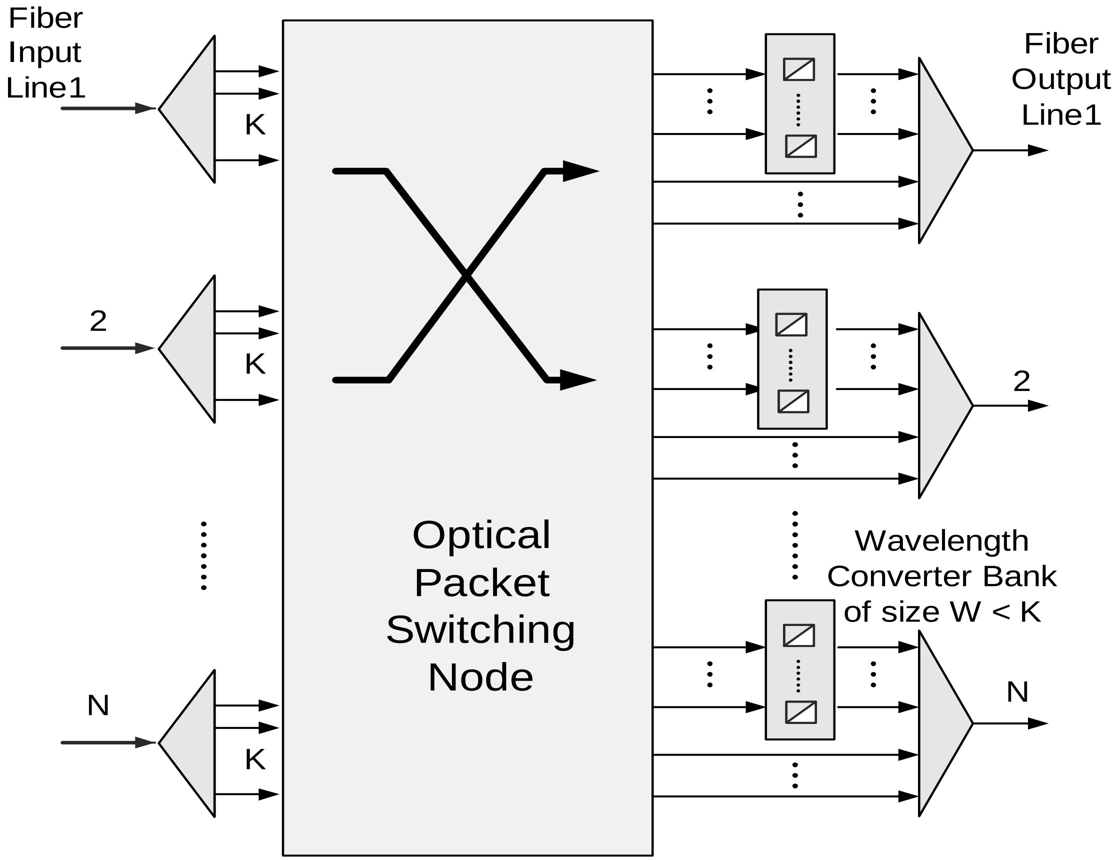
Implications

- Erlang loss formula can in general be used to dimension OBS networks when
  - The number of source-destination pairs is large enough to justify the Poisson model
- In case the number of users is far fewer, the 2D Markov chain can be solved for dimensioning purposes
- The on-off model assumes burst shaping at the edge

Converter Sharing Models

- Share-per-node
  - All converters are collected in a single pool for converter sharing across all fiber lines
  - Powerful but costly
- Share-per-link
  - A simpler and less costly architecture allows separate converter banks per fiber link
  - Share-per-input-link (SPIL)
  - Share-per-output-link (SPOL)
- Stochastic analysis of converter sharing with SPOL
  - No tuning range limit (exact solution)
  - Tuning range limit (approximation)

Partial Wavelength Conversion - SPOL


Fiber Input Line1
K
2
K
N
K
Optical Packet Switching Node
Fiber Output Line1
2
Wavelength Converter Bank of size W < K
N

SPOL Model

- *K* wavelength channels per fiber
- A wavelength converter bank of size $0 < W < K$ per output fiber
- Bursts destined to a particular output fiber line arrive at the OXC through the *N* input fibers
- Burst arrival process is Poisson with rate λ
- The wavelength channel they arrive on is uniformly distributed on $(1,K)$
- Burst durations are exponentially distributed with mean 1/μ

SPOL Model

- A new burst arriving at the switch on wavelength *w* and destined to output line *k* is forwarded to output line *k* without using a converter if
  - channel *w* is available, else
  - is forwarded to output line *k* using one of the free TWCs in the converter bank and using one of the free wavelength channels selected at random,
  - else is blocked

Markov Chain Analysis

- The process *X(t) = {(i(t), j(t)): t > 0}* is a Markov process on the state space *S = {(i,j): 0 ≤ i ≤ K, 0 ≤ j ≤ min(i,W)}*
  - *i(t):* number of wavelength channels occupied
  - *j(t):* number of converters used

$$S=\left\{\underbrace{(0,0)}_{level\ 0},\underbrace{(1,0),(1,1)}_{level\ 1},\underbrace{(2,0),(2,1),(2,2)}_{level\ 2},\cdots,\underbrace{(K,0),(K,1),\cdots,(K,W)}_{level K}\right\}$$

- Based on this enumeration, we conclude that state transitions can occur either among neighboring levels or within a level
- The resulting Markov chain is again block tri-diagonal

Non-homogenous QBD

$$Q=\begin{bmatrix} A_0 & U_1 & & & \\ D_0 & A_1 & U_2 & & \\ & D_1 & A_2 & \ddots & \\ & & \ddots & \ddots & U_K \\ & & & D_{K-1} & A_K \end{bmatrix}$$

$$P_{loss}=x_K e+\sum_{i=W}^{K-1} x_{i,W}\frac{i}{K}$$

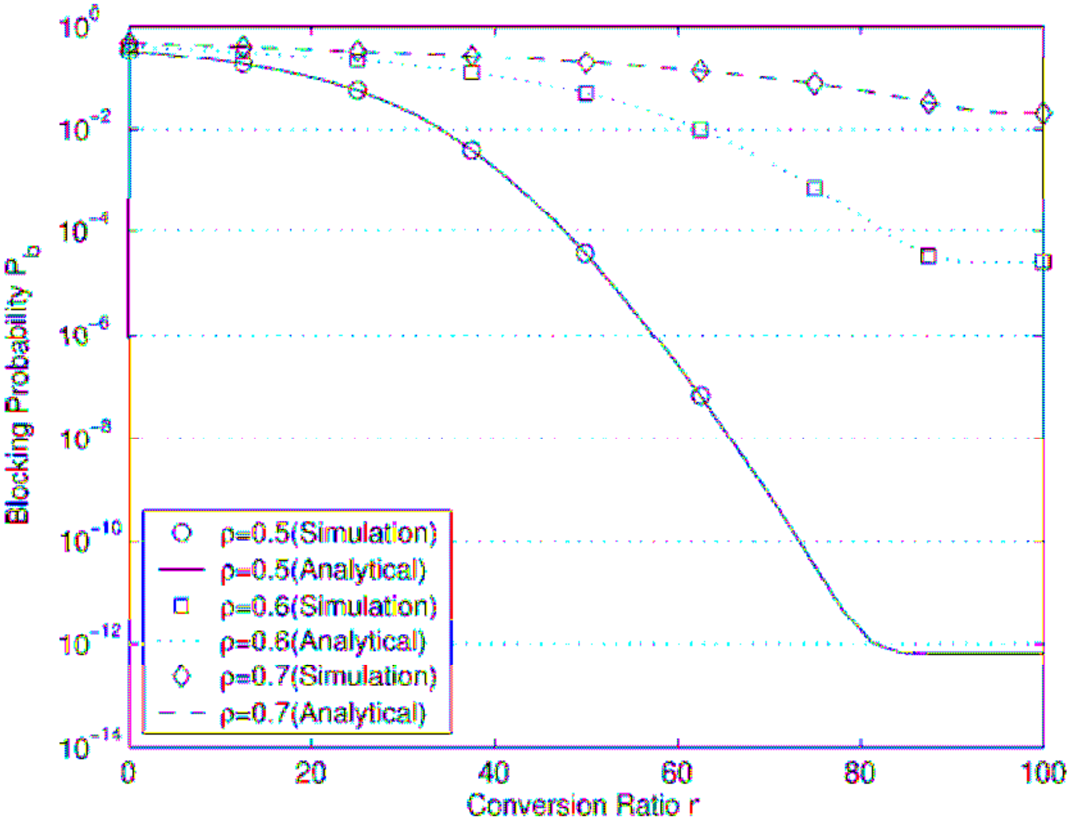
$$xQ=0,\sum_i\sum_j x_{i,j}=1$$

🟡 Loss due to lack of channels

🔵 Loss due to lack of converters

Numerical Results – Poisson Burst Arrivals

K=128



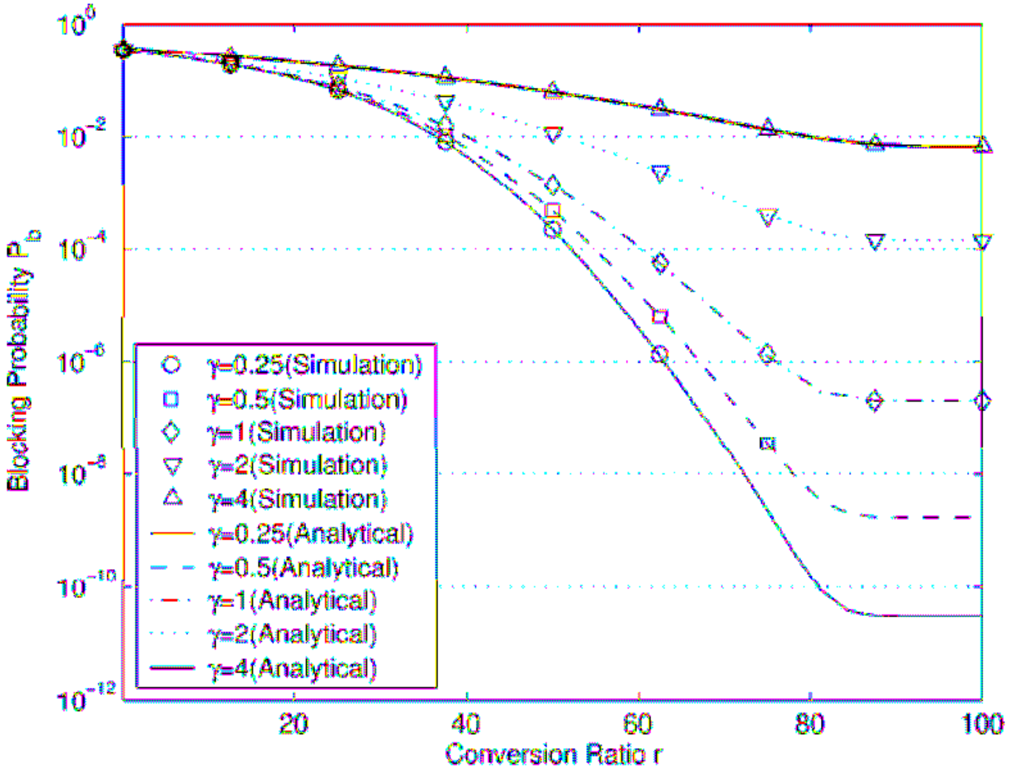
- Conversion ratio

r = # converters/ # channels

- Loss probability decreases first slowly then rapidly with increased number of converters but saturates after a while when conversion ration *r* = 80%

Numerical Results – Arrival Process Coefficient of Variation CoV

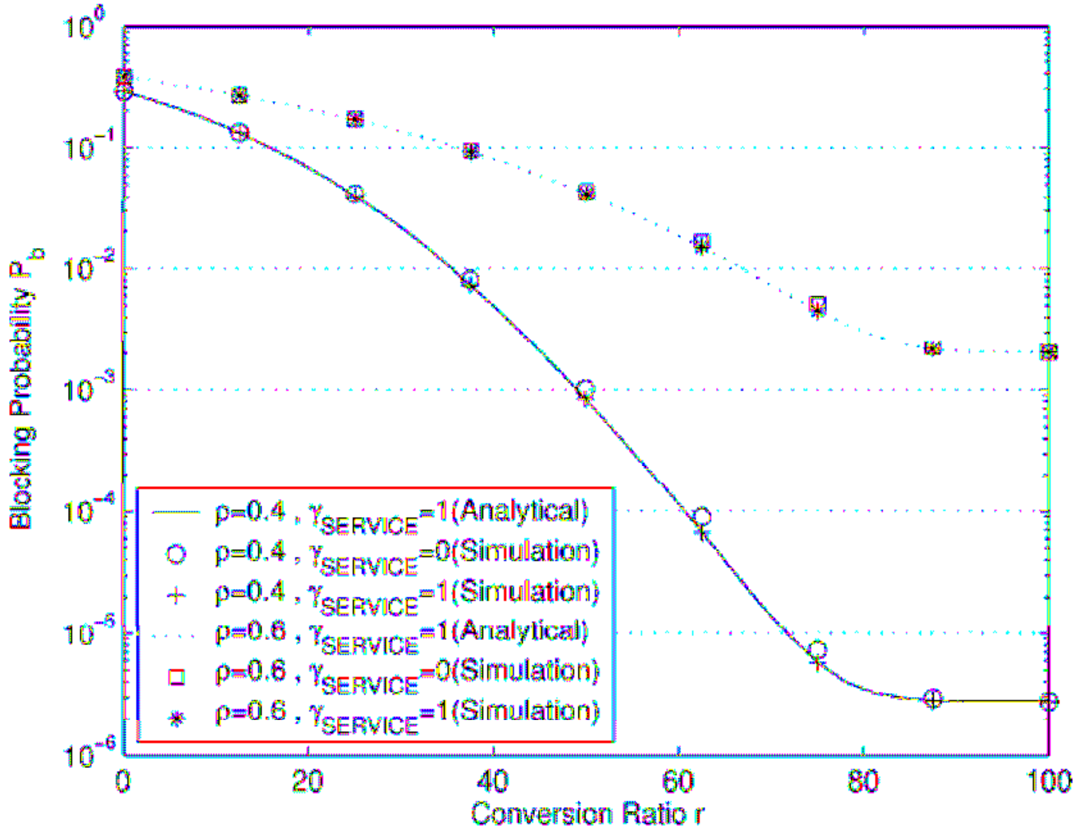
K=64, ρ=0.4



- The higher the CoV, the higher the loss probability
- Such second order traffic characteristics need to be taken into consideration for accurately modelling burst switching systems
- Burst traffic shaping at the ingress of an OBS network that can reduce the CoV would also be effective in reducing burst blocking inside the OBS core.

Sensitivity to Burst Length Distributions

K=32

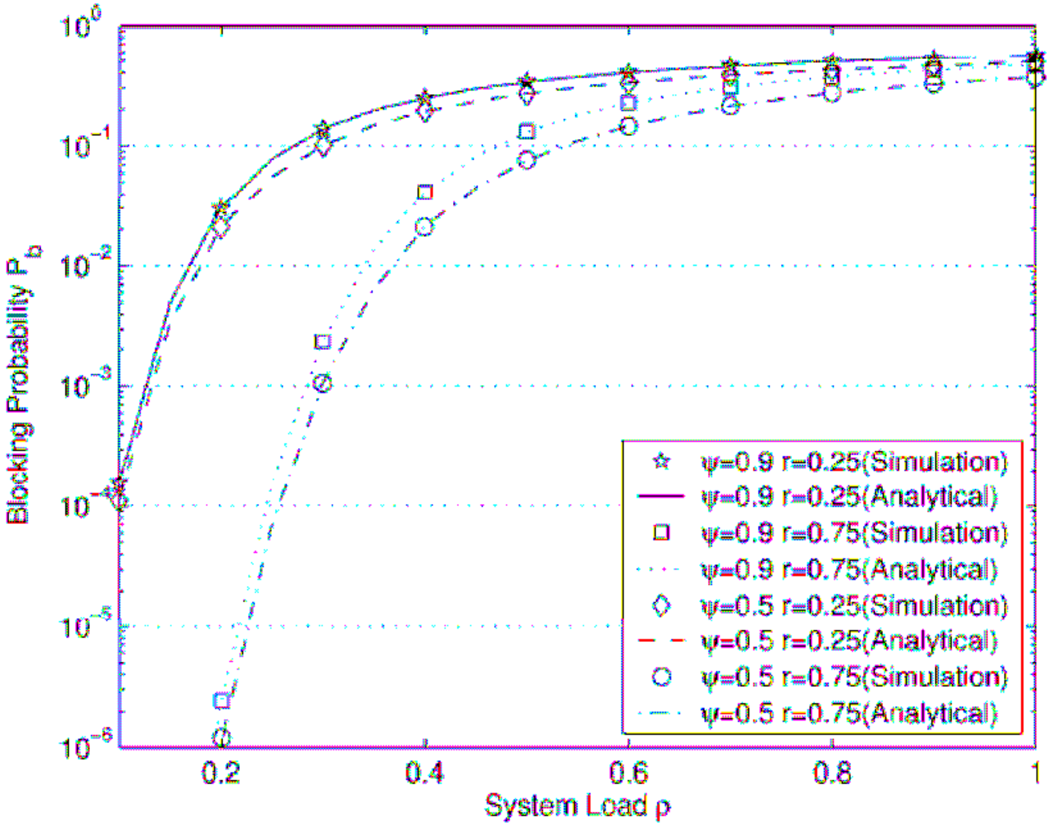


Insensitivity:

- Known to hold at the boundaries
- Slight discrepencies at the middle regime

Numerical Results – Impact of Auto-correlation

K=32



MAP obtained through

- leaving the marginal phase-type distribution invariant
- incorporating lag-k correlation of the form $\mathrm{corr}(x_0,x_k) = c\psi^k$
- The higher the correlation parameter, the larger the blocking probability

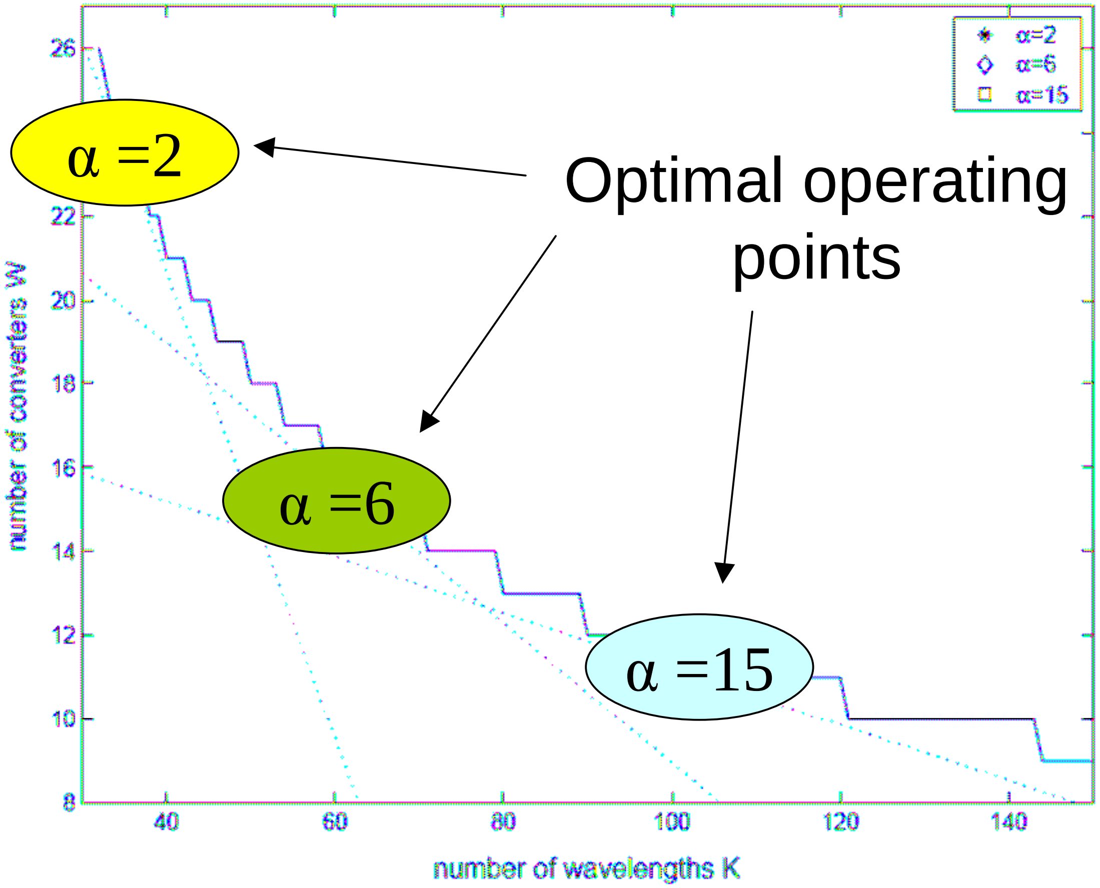
Numerical Results – Computation Times in Sec.

W/K	64	128	256
20	0.02	0.05	0.08
60	0.05	0.19	0.47
120	-	0.55	1.94
180	-	-	3.81

- **Results obtained on a 3Ghz Pentium PC**
- **For a MAP multiply with m^3**

Numerical Example – 15 Erlangs Input

$P_b = 10^{-4}$, converter_cost = α channel_cost



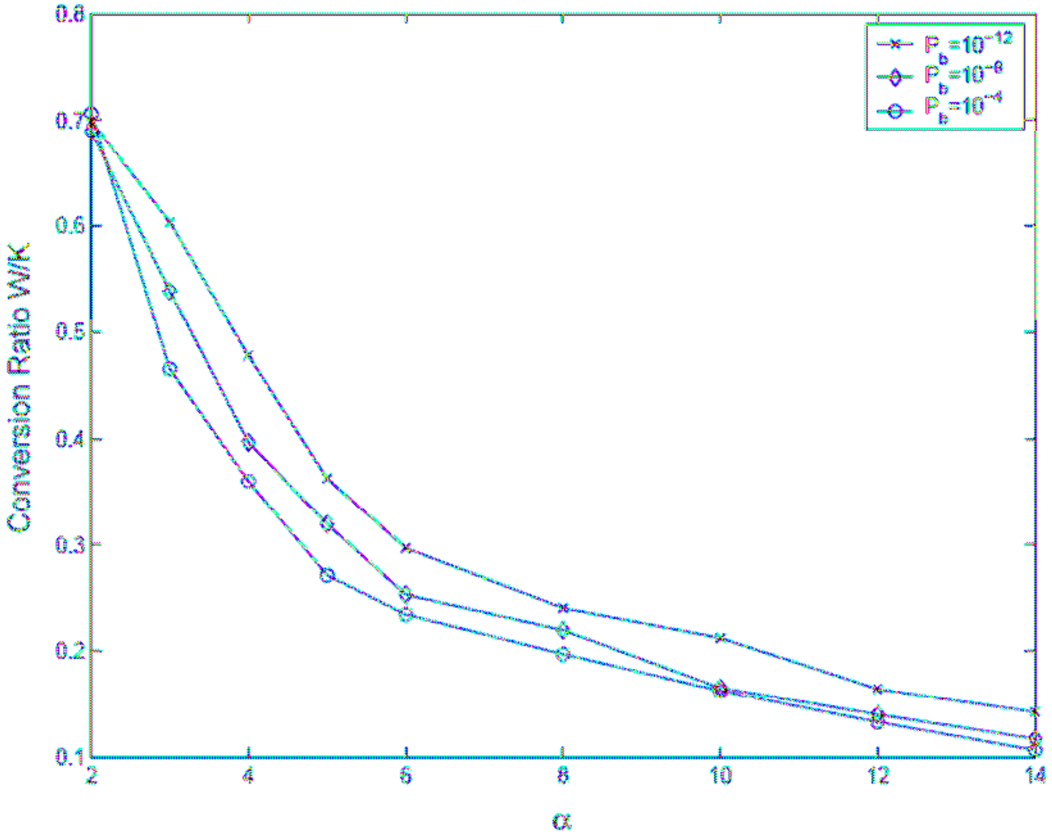
α =2

α =6

α =15

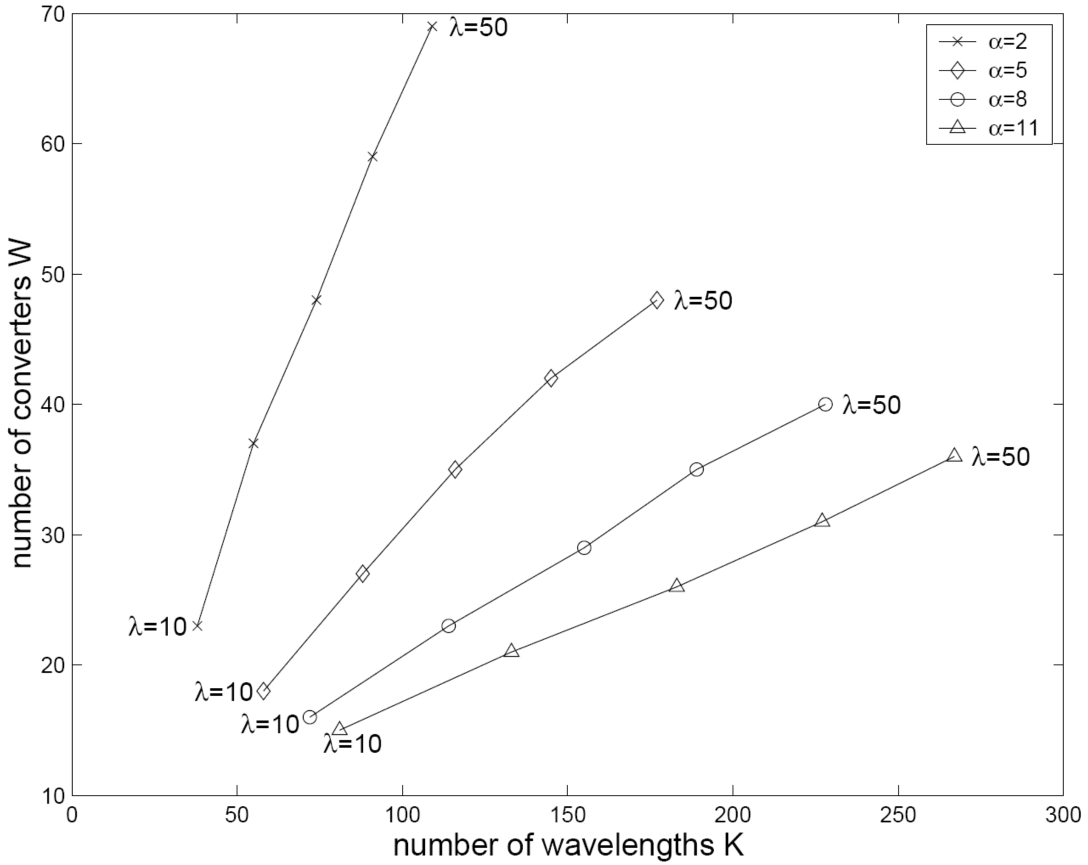
Optimal operating points

Provisioning Guidelines



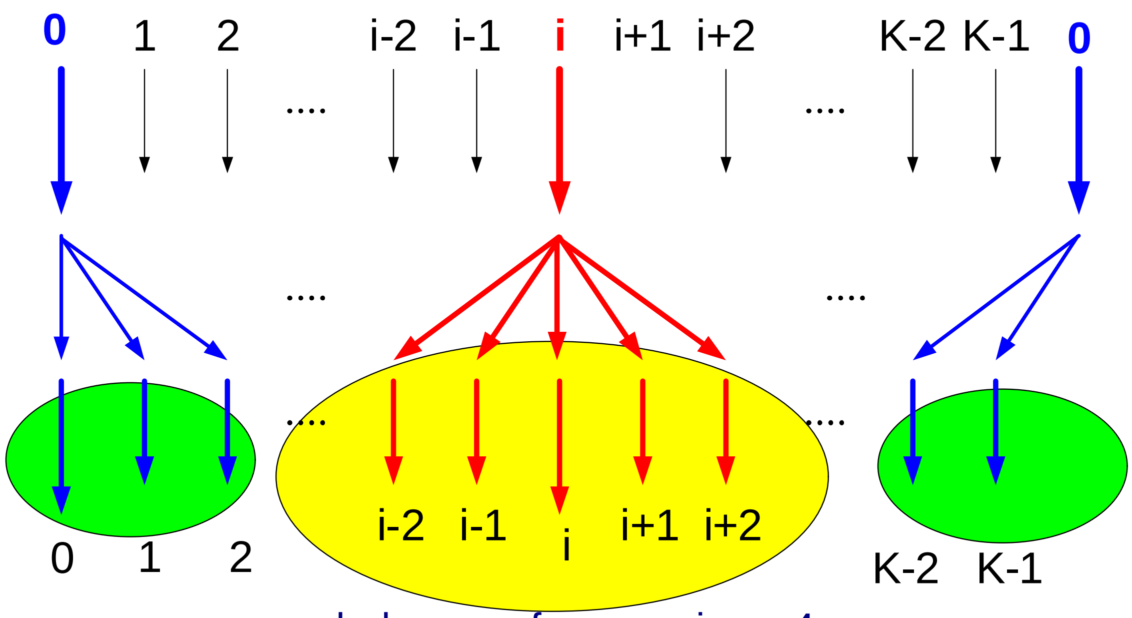
- When α is small, i.e., converters are relatively inexpensive the optimal conversion ratio is around 70%.
- When α is around 10, i.e., converters are relatively expensive, the optimal conversion ratio drops to around 20%.
- The optimal ratios slightly increase in favor of more converter use with more stringent blocking probability requirements.

Optimal Channel/Converter Pairs - Varying Load



- The optimal conversion ratio does not appear to change with load
- Use the same conversion ratio as a guideline as long as relative costs remain the same with respect to increased loads

Limited-range Wavelength Conversion – Circular type



incoming wavelength

0 1 2 .... i-2 i-1 i i+1 i+2 .... K-2 K-1 0

....

....

....

0 1 2

i-2 i-1 i i+1 i+2

K-2 K-1

d: degree of conversion = 4

outgoing wavelength

🟩 Tuning range for wavelength 0

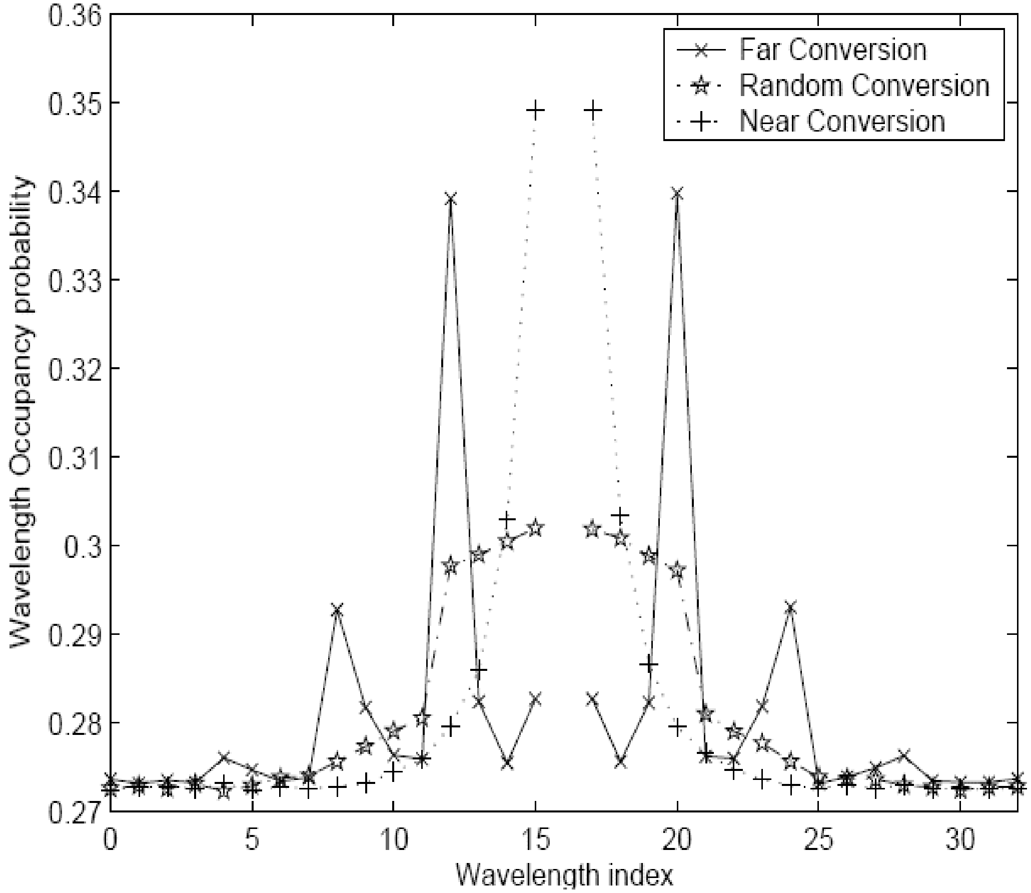
🟨 Tuning range for wavelength i

Conversion Policies

- Random Conversion Policy
  - The outgoing wavelength is selected randomly from the set of idle wavelengths in the range.
- Near Conversion Policy
  - We choose the nearest available wavelength from the set of idle wavelengths in the conversion range and if there exist two such wavelengths, one of them will be selected in random.
- Far Conversion Policy
  - In this policy, the farthest available wavelength is selected from the set of idle wavelengths in the conversion range. If there exist two such wavelengths, one of them will be selected in random.

Clustering Effect

K = 33, W = 15, d=8, load = 27%



- The occupancy probabilities of wavelengths conditioned upon an arriving packet on wavelength 16 finding this wavelength in use
- Wavelength occupancy probability histogram clustered within the conversion range
- Clustering most dominant in the near conversion policy
- Clustering least dominant in far conversion
- We propose far conversion

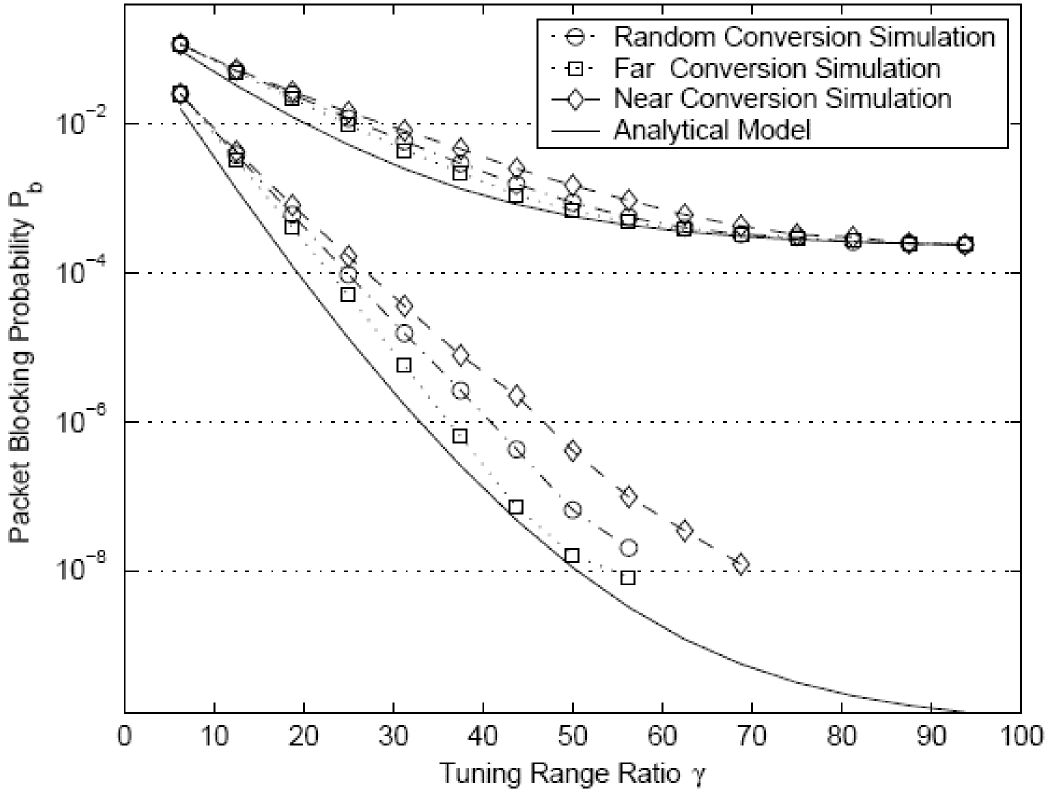
Approximate Analytical Model

- Use the same PWC model except that
  - Used wavelengths are uniformly distributed and not clustered
  - A packet arrival finding its incoming wavelength occupied and *i* channels occupied also finding its tuning range all occupied has probability

$$\frac{i-1}{K-1}\,\frac{i-2}{K-2}\cdots\frac{i-d}{K-d}$$

  - Does not capture the clustering effect

Far Conversion vs Others



- Analytical model hard to find
- The model here captures the limited-range nature but not the clustering
- Used as a lower bound
- Far conversion policy outperforms all the other conversion policies
- Far conversion is easy to implement

OBS tutorial

Quality of Service Provisioning in OBS networks

QoS in one-way reservation OBS

- Tell-and-Go (TAG) OBS performs according to the statistical multiplexing paradigm
  - need for an additional support for QoS provisioning in order to preserve HP traffic from LP traffic
- Two basic models can be distinguished
  - relative QoS
  - absolute QoS

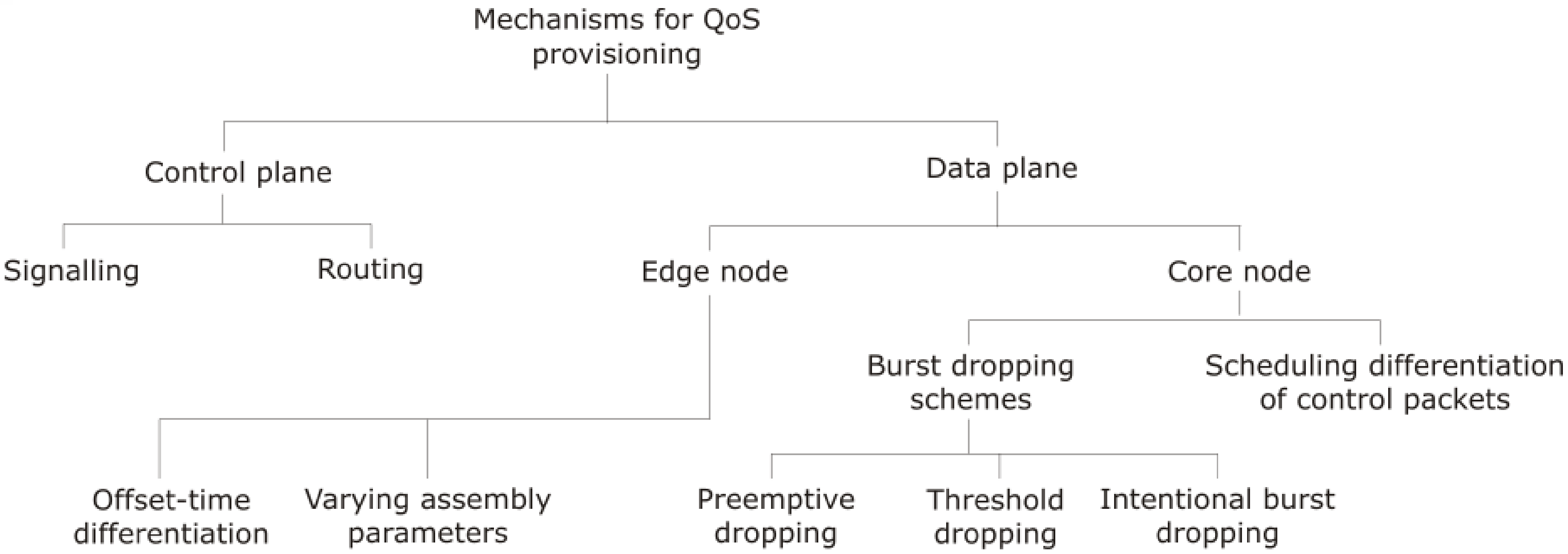
Relative QoS

- The performance of a class is defined with respect to the other classes
  - for instance it is guaranteed that loss the probability of a burst belonging to the HP class is lower than the loss probability of a burst belonging to the LP class
- Analogous to differentiated services in IP networks
- In most cases, can be easily implemented
- Performance of a given class may depend on traffic characteristics of the other classes

Absolute QoS

- An absolute performance metric of quality is defined for a class
  - for example a maximal acceptable level of burst loss
- Absolute QoS model aims at irrelative quality provisioning
- It may require more complex implementations
  - the problem is to provide desired quality levels in wide range of traffic conditions and at the same time to preserve high output link utilization

Classification of QoS mechanisms


Mechanisms for QoS provisioning
Control plane
Data plane
Signalling
Routing
Edge node
Core node
Burst dropping schemes
Scheduling differentiation of control packets
Offset-time differentiation
Varying assembly parameters
Preemptive dropping
Threshold dropping
Intentional burst dropping

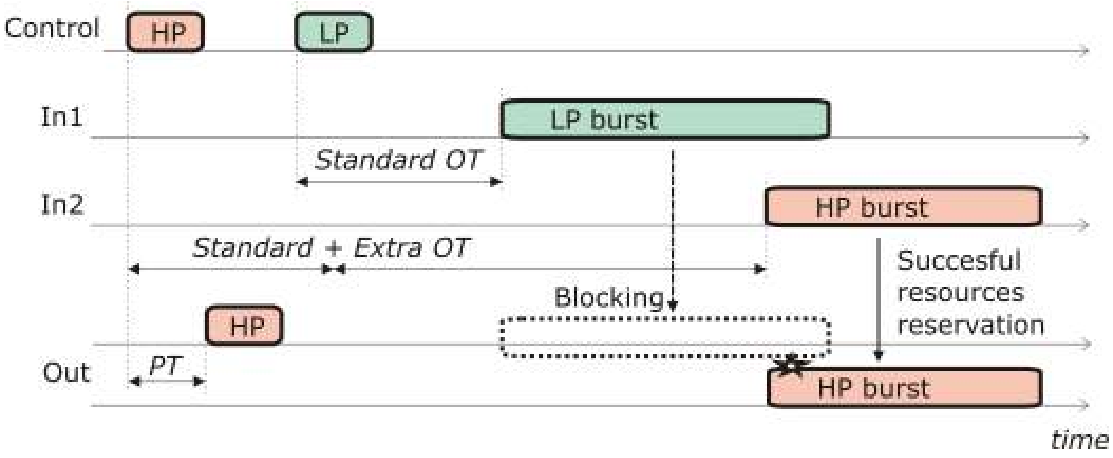
QoS in Control Plane

- Supporting an absolute QoS by a hybrid signaling protocol that consists of a co-operation of two-way and one-way resource reservation modes
  - end-to-end wavelength paths providing guarantees such as no losses and negligible delays inside the network
  - the unreserved resources used for best-effort traffic
- QoS functions of routing protocol
  - preserving the selection of overloaded parts of the network for loss-sensitive applications
  - minimizing the path lengths for delay-sensitive ones

QoS in Data Plane (Edge Nodes)

- Burst assembly according to the class and destination of client packets
- Assigning specific attributes (labels, priorities) to the bursts
  - carried by control packets
  - with the purpose of their further discrimination and processing in core nodes
- QoS mechanisms in the edge node
  - Offset-time differentiation
  - Varying burst assembly parameters

Offset-Time Differentiation



- Extra offset-time (OT) assigned to HP bursts that results in earlier reservation
- Fine class isolation if the extra OT is equal to a few burst durations of the mean length of a LP burst

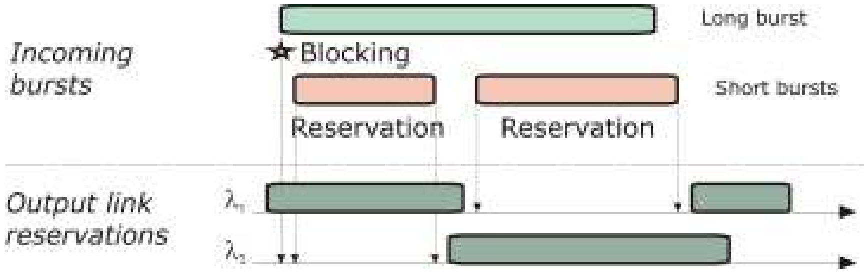
  (+) no additional differentiation mechanisms necessary in the core of network

  (–) sensitivity of HP class to burst length characteristics

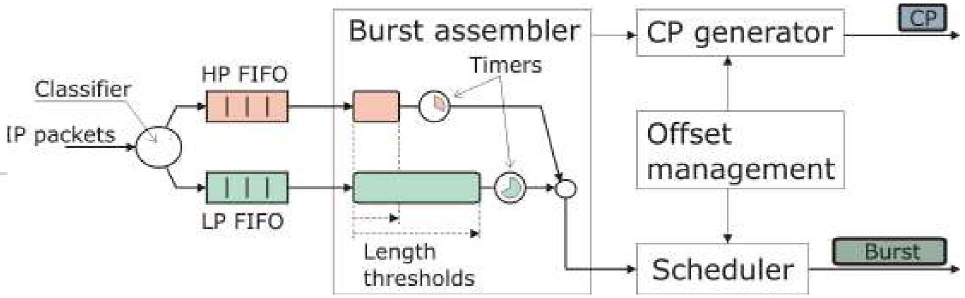
  (–) extended pre-transmission delay which can be critical especially for TCP

Burst Length Differentiation (BLD)

a) Burst lengths and the contention problem



b) Assembly unit for BLD



- Shorter bursts have more chances to fill gaps between already scheduled bursts
- In BLD, each of the QoS classes engages different assembly parameters
  - HP bursts are aggregated with lower timer (decreasing the delay) and maximum burst length thresholds than LP bursts

(+) significantly improved blocking performance of HP bursts in combined scenarios with other QoS mechanisms and with FDL buffers applied

(–) higher switching-time requirements due to shorter bursts

(–) increased signaling overhead due to increased number of control packets

QoS in Data Plane (Core Nodes)

- QoS provisioning takes place in resolving the contention with assistance of wavelength conversion, FDL buffering and deflection routing
- Burst dropping schemes:
  - Intentional burst dropping
  - Preemptive dropping
  - Threshold-based dropping

Intentional Burst Dropping

- Can be classified as an absolute QoS technique
- Maintains the performance objectives of the higher priority bursts on certain levels by intentional dropping the lower priority bursts using active discard techniques such as RED (*Random Early Detection*)

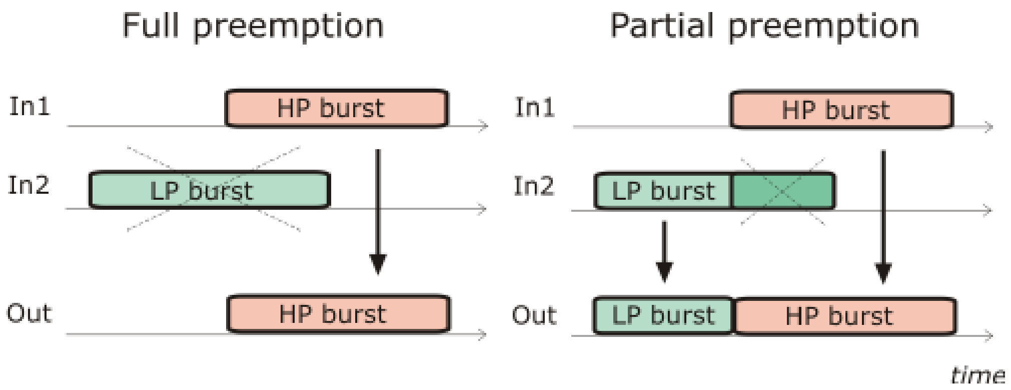
  (+) provides absolute QoS

  (–) link utilization may suffer

  (–) its implementation may be complex

  (–) loss is a rare event to measure

Burst Preemption



- In case of burst conflicts: overwrites the resources reserved for LP burst by HP one; the preempted LP burst is discarded
- Preemption concerns either a whole burst or can be partial

(+) fine class isolation and output link utilization

(–) in case of successful preemption either resources reserved for the preempted burst are wasted in consecutive nodes or a signaling protocol is necessary in order to release them

(–) additional complexity involved in the burst assembly process in partial preemption

An absolute QoS differentiation scheme

- Based on
  - Harald Øverby and Norvald Stol, “*Quality of Service in Asynchronous Bufferless Optical Packet Switched Networks*”*,* Telecommunication Systems, Oct.-Dec. 2004.
- One can extend a preemption-based QoS differentiation scheme to provide absolute QoS
- Preemption parameter *p*: Probability for a high priority burst to preempt a low priority burst when output link is congested
- The preemption parameter *p* is adjusted according to loss rate measurements at each core node

An absolute QoS differentiation scheme

- Measurements are performed within a window consisting of Q packets
  - Small Q: fast but not accurate measurements
  - Large Q: slow but accurate measurements
  - Optimal trade-off depending on required accuracy and adaption time
- Estimate the loss probability for class 0:

$$P_{0,\mathrm{est}} = \frac{\sum_{i=1}^{Q} d_i}{Q}$$

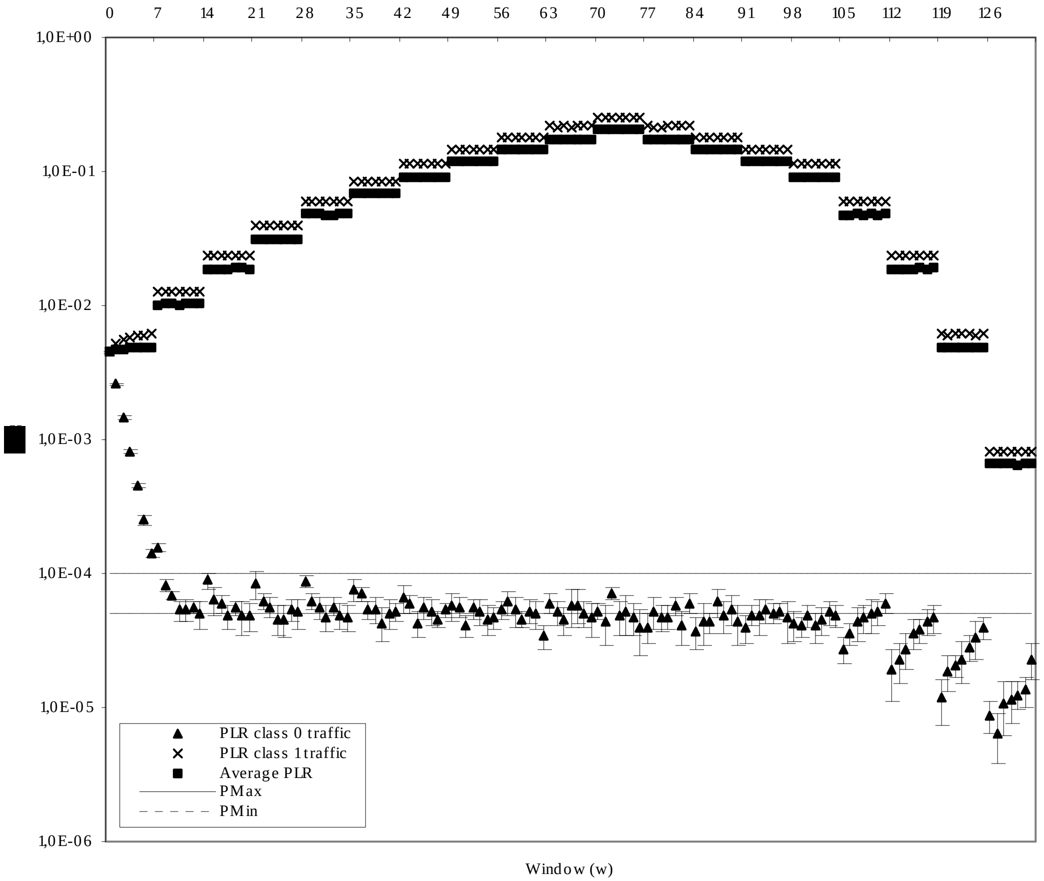
An absolute QoS differentiation scheme

- Adapt the preemption parameter *p*

$$p_{n+1} = \begin{cases} p_n - (1-p_n)\delta, & P_{0,\mathrm{est}} < P_{0,\mathrm{min}}, \\ p_n + (1-p_n)\delta, & P_{0,\mathrm{min}} < P_{0,\mathrm{est}} < P_{0,\mathrm{max}}, \\ p_n + (1-p_n)\delta^{1/2}, & P_{0,\mathrm{est}} > P_{0,\mathrm{max}}. \end{cases}$$

- Next viewgraph: We study an example assuming
  - 20% HP traffic
  - 8 fibres, 16 wavelengths per fiber
  - Full conversion, no FDLs, δ=0.2
- Shows the loss rate as a function of time (w)

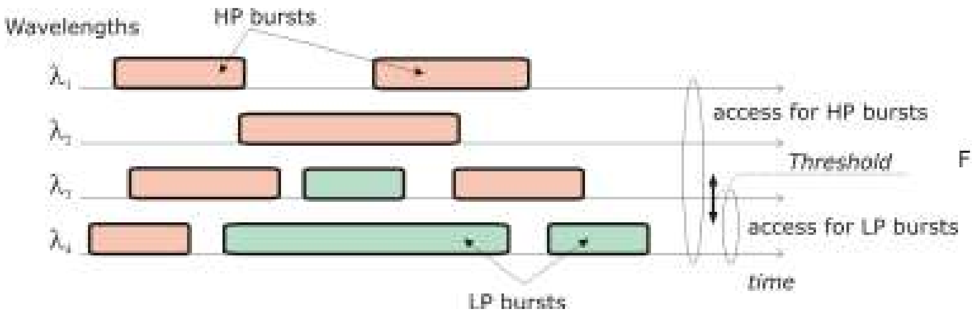
An absolute QoS differentiation scheme



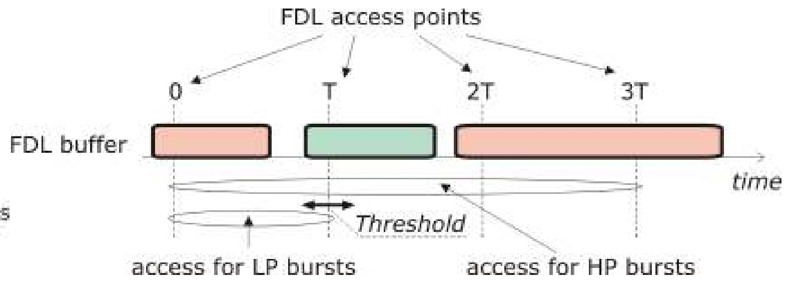
- Load is varied from 0.5 to 0.95 and back to 0.5
- Loss rate for class 0 traffic is almost constant
- Loss rate for class 1 traffic varies according to load variations

Threshold-based Dropping

a) Burst dropping with wavelength threshold



b) Burst dropping with buffer threshold



- Provides more resources to HP bursts then to LP bursts according to certain *threshold* parameter
- If the resource occupation is above the threshold then the LP bursts are discarded whilst the HP bursts are accepted

(+) easy implementation

(–) the efficiency strongly depends on the threshold adaptability to actual traffic load

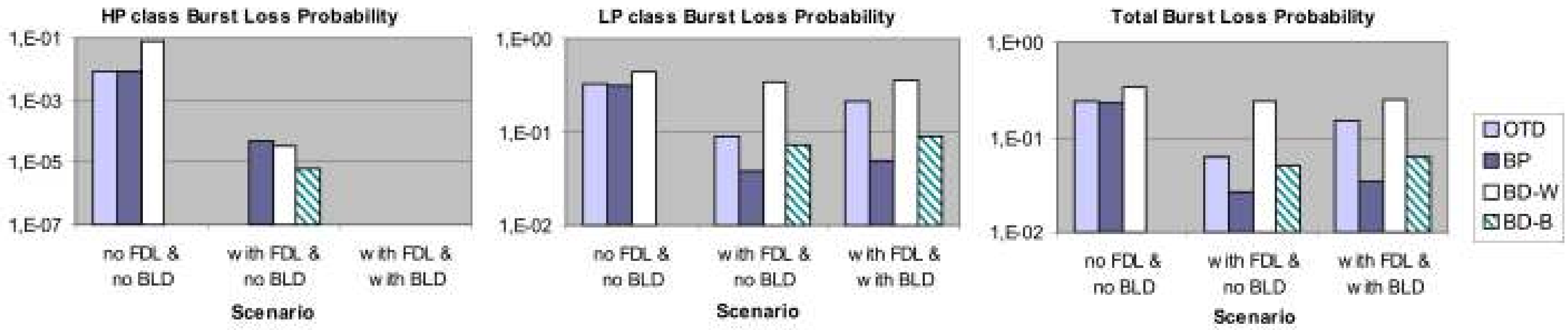
Scheduling Differentiation of Control Packets

- Reservation requests serviced earlier have more chances to encounter free transmission resources
- Scheduling of control packets based on
  - priorities
  - fair queuing techniques, which regulates access to the reservation manager

Comparison of Various QoS Schemes

- Based on
  - M. Klinkowski, et al., *“Impact of Burst Length Differentiation on QoS performance in OBS networks”,* Proc. ICTON 2005, Barcelona (Spain), July 2005.
  - Results obtained for
    - 4 wavelengths,
    - 0.8 Erlangs traffic load,
    - 30% of HP class traffic.

Comparison of QoS Mechanisms in a Single Node Scenario$^{1)}$

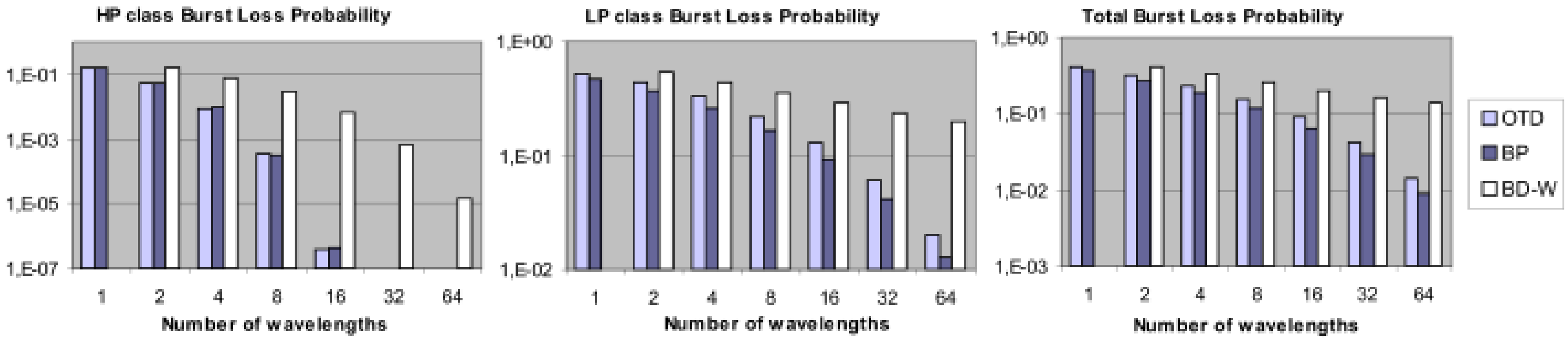


OTD – Offset-Time Differentiation
BP – Full Burst Preemption
FDL – Fiber Delay Line buffer (of 4 FDLs)

BD-W – Burst Dropping with Wavelength Threshold (50% of λs for LP class)
BD-B – Burst Dropping with Buffer Threshold (50% of FDLs for LP class)
BDL – Burst Length Differentiation (1:4 burst length ratio)

- In bufferless case, both OTD and BP offer the same performance for HP class, however BP conserves better from losses for both LP and total traffic
- FDL buffers improve the performance
- BLD improve the loss performance of HP class in BP, BD-W and BD-B mechanisms
- Variable length bursts should not be used with OTD since the total and LP class performance may be significantly impaired.

QoS and Wavelength Dimensioning



*) results obtained in buffer-less scenario

- Increasing the number of wavelengths improves the effectiveness of QoS differentiation
- The improvement of HP class performance in both OTD and BP can be really high
  - e.g. when increasing the number of wavelengths from 8 to 16 there is HP class performance gain of 3 orders of magnitude
- Poor performance of BD-W is because it has effectively less wavelengths available for burst transmissions on the output port then other mechanisms
  - it provides only 50% of wavelengths for LP class whilst it attempts to serve the same amount of input traffic as the other mechanisms

Feedback-based QoS (Edge-core)

- Uncontrolled OBS
  - Whenever bursts are formed, they are immediately sent towards the OBS domain
  - Analogous to ATM UBR service
- Non-feedback based OBS
  - Sources are allowed to inject bursts at predetermined static rates so that congestion does not arise in the OBS network
  - Analogous to ATM VBR service
- Feedback-based OBS
  - Congestion control is achieved by varying the burst injection rate at the ingress nodes so as to match the available bandwidth in the network
  - Analogous to ATM ABR service which motivates the current study
  - Most suitable for unpredictable bursty aggregate traffic
- What mechanisms are required for feedback-based OBS?

Effective Capacity

- QoS requirement: burst blocking probability P_{loss}
- The **Effective Capacity (EC)** of an optical WDM link between two OXCs is the amount of traffic in bps that can be burst switched by the link while meeting the desired QoS requirement.

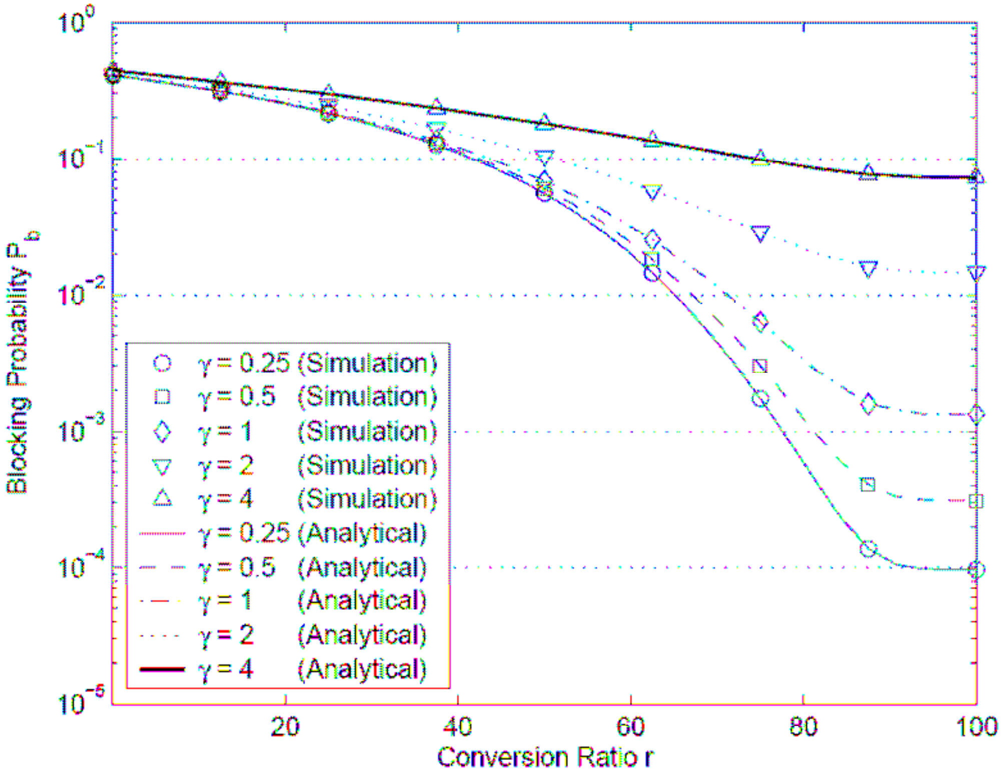
- EC depends on
  - Burst Interarrivals
  - Burst Duration
  - Contention Resolution Capability of the OXC

- *Poisson ?*
- Exponential (as a first step)
- FWC, PWC, FDLs

Effective Capacity

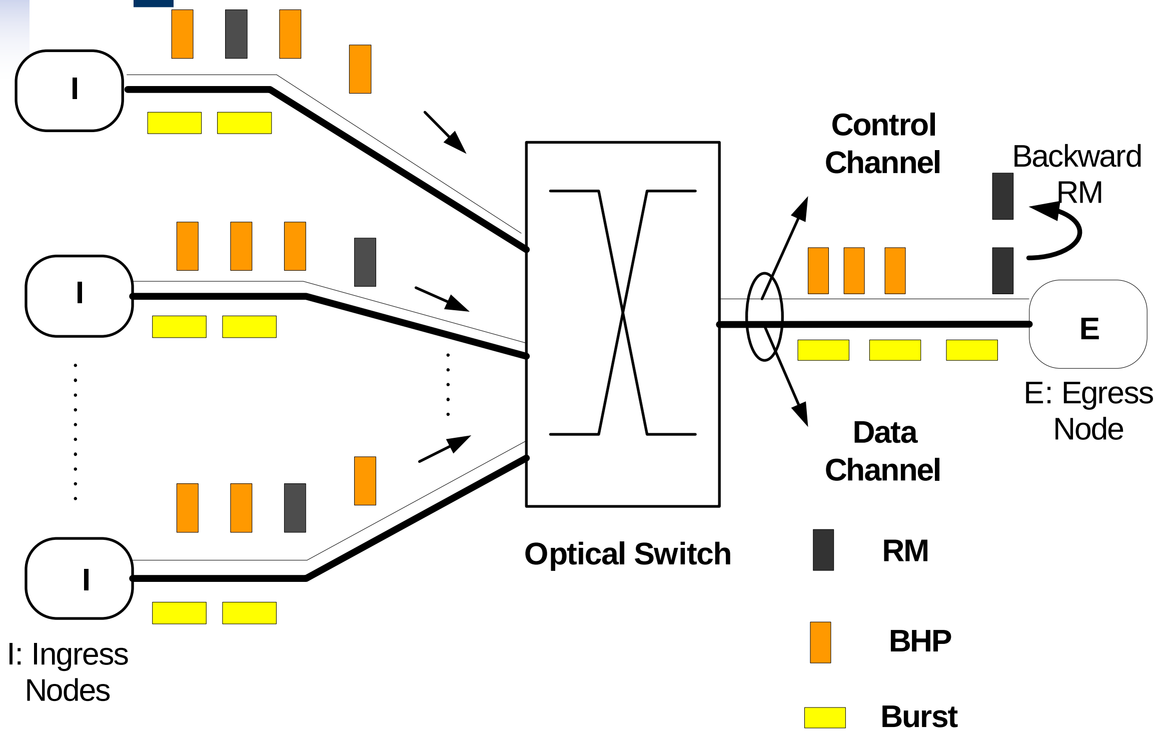
- Given the traffic model and the contention resolution capabilities of the optical link, solve for blocking probabilities using
  - off-line simulations or
  - analytical techniques
- Note the maximum λ_{max} that results in the desired blocking probability p_{loss}
- Set $EC = \lambda_{max} p/\mu$. p: link transmission rate $1/\mu$: mean burst transmission time,
  - Example: A 100-wavelength optical link
    - *$p_{loss} = 10^{-4}$, channel rate = 10 Gpbs, $1/\mu = 1\mu s$*
    - *EC = 690 Gps (full wavelength conversion)*
    - *EC = 490 Gbps (%50 wavelength conversion – share per link)*

Why Poisson?



- Rate-controlled traffic will have a smaller Coefficient of Variation (CoV), denoted by γ, than that of Poisson traffic
- **Conjecture:** Rate control using EC with Poisson assumption leads to even better performance than provisioned

Differentiated ABR Protocol



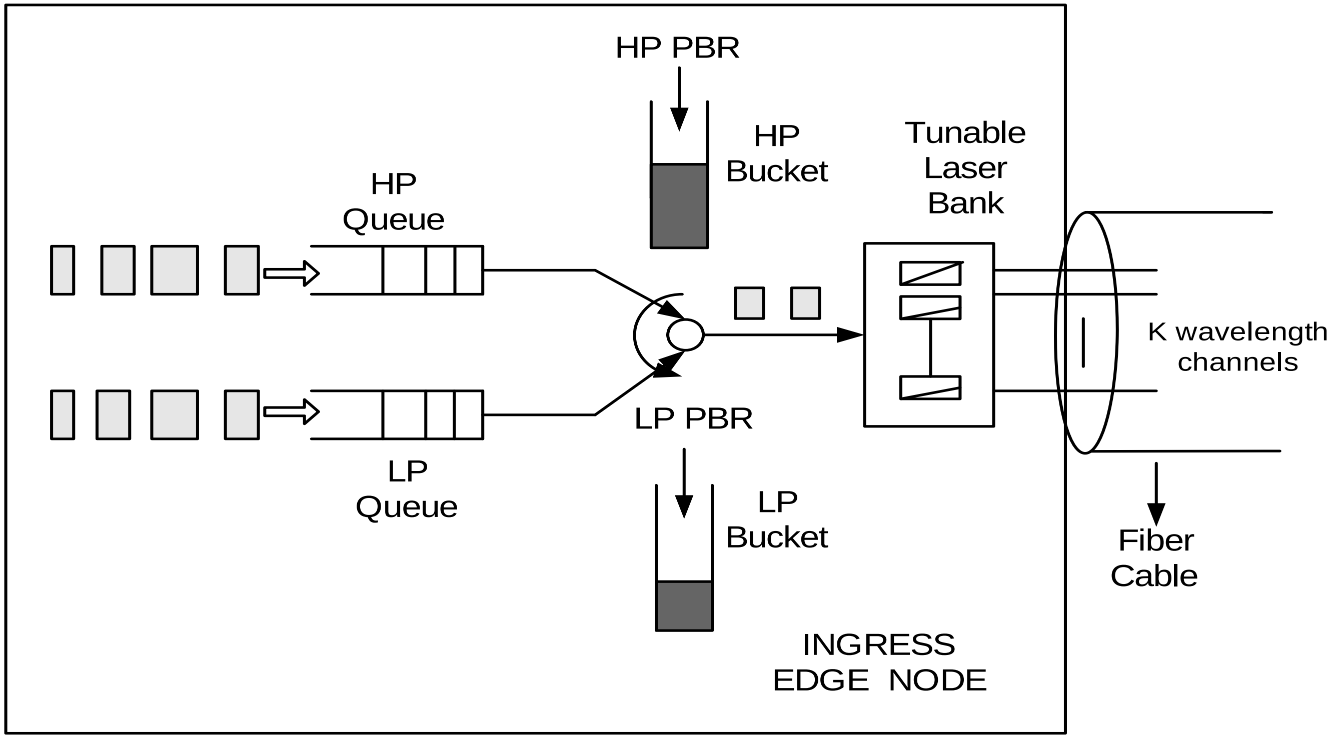
- RM packets are sent over the control channel with period T
- ER fields for High- and Low-priority traffic are written by the OXC
- We use the basic bufferless version of ERICA for ER calculation

H. Boyraz and N. Akar, "*Rate-controlled optical burst switching for both congestion avoidance and service differentiation*", Optical Switching and Networking, Dec. 2005.

Outline of the D-ABR Protocol

- The EC is distributed among HP sources on a max-min fair share basis
- Capacity remaining from HP sources is then distributed among LP sources again on a max-min fair share basis
- HP ER and LP ER fields on backward RM packets are written accordingly
- The Permitted Bit Rates HP PBR and LP PBR (the rates at which HP and LP bursts would be injected into the OBS network) are calculated by
  - HP PBR := min(HP ER, HP PBR + RIF*HP PBR),
  - LP PBR := min(LP ER, LP PBR + RIF*LP PBR)

Edge Scheduler

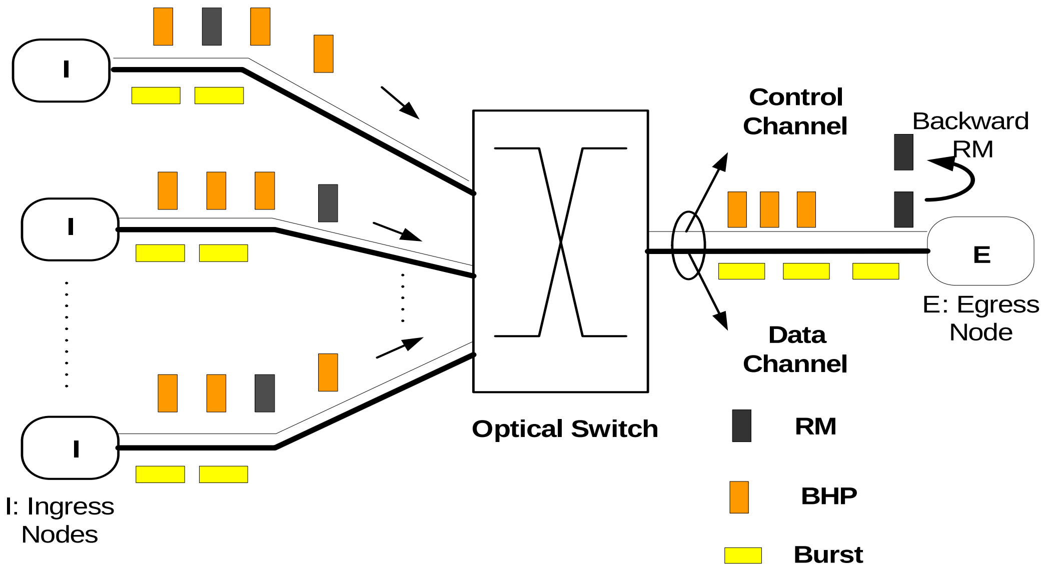


OBS Multiplexer Example

D: Propagation Delay of the links

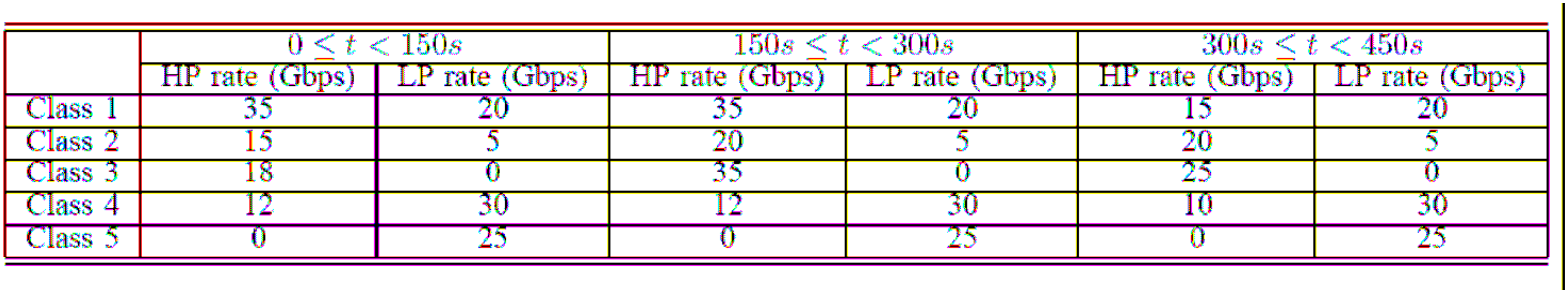
100 Wavelengths

25 sources
5 classes



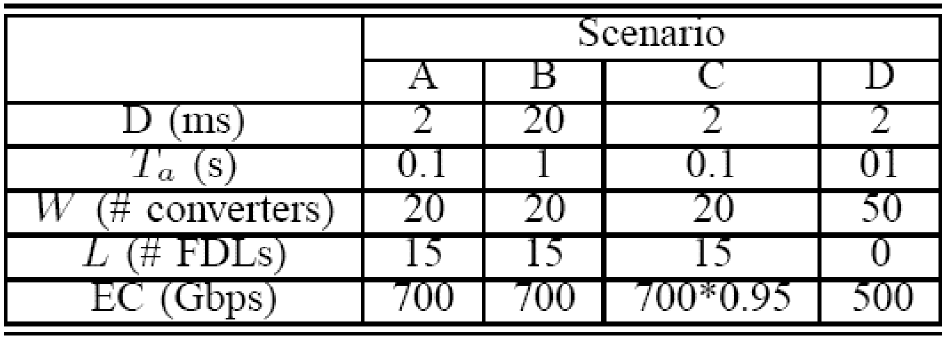
Problem Parameters

- 25 users, 5 classes with 5 users each
- 4 tuneable lasers per source



	$0 \leq t < 150s$		$150s \leq t < 300s$		$300s \leq t < 450s$	
	HP rate (Gbps)	LP rate (Gbps)	HP rate (Gbps)	LP rate (Gbps)	HP rate (Gbps)	LP rate (Gbps)
Class 1	35	20	35	20	15	20
Class 2	15	5	20	5	20	5
Class 3	18	0	35	0	25	0
Class 4	12	30	12	30	10	30
Class 5	0	25	0	25	0	25

- 4 scenarios

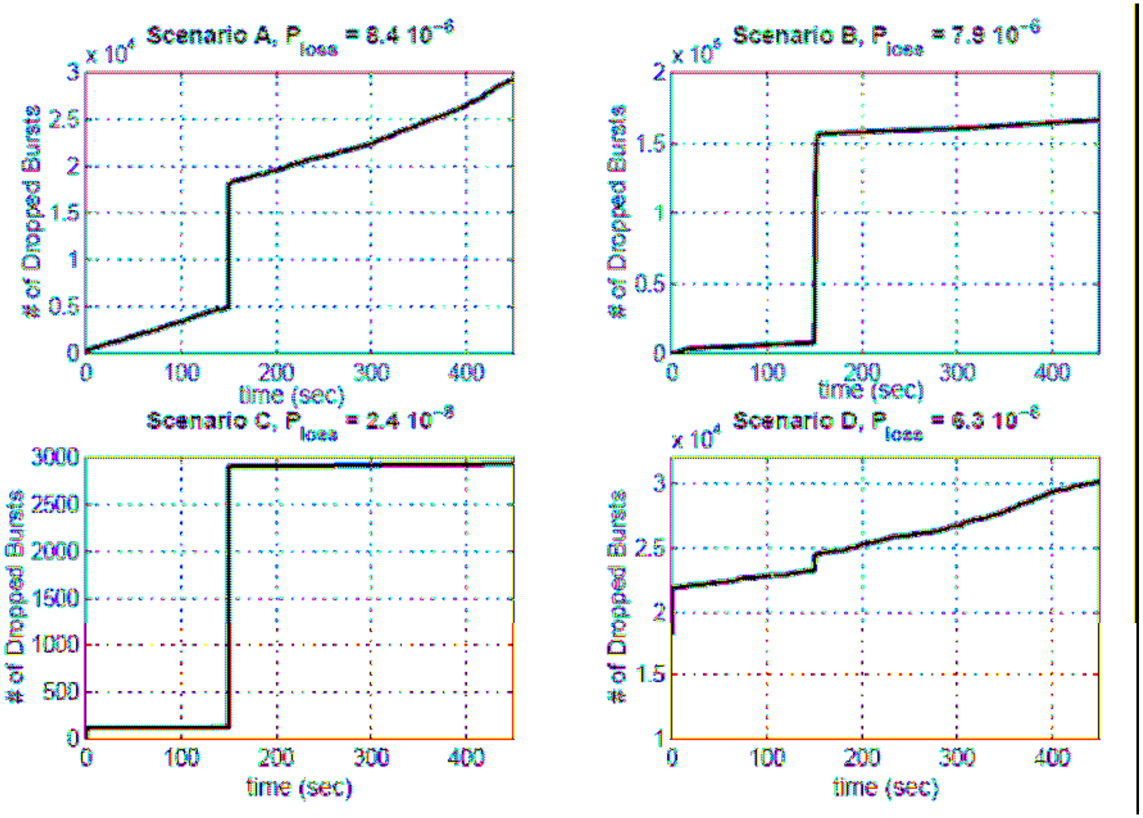


	Scenario			
	A	B	C	D
D (ms)	2	20	2	2
T_a (s)	0.1	1	0.1	01
W (# converters)	20	20	20	50
L (# FDLs)	15	15	15	0
EC (Gbps)	700	700	700*0.95	500

Burst Blocking Rate wrt Time

Provisioned P_{loss}= 3.2 10^{-5}

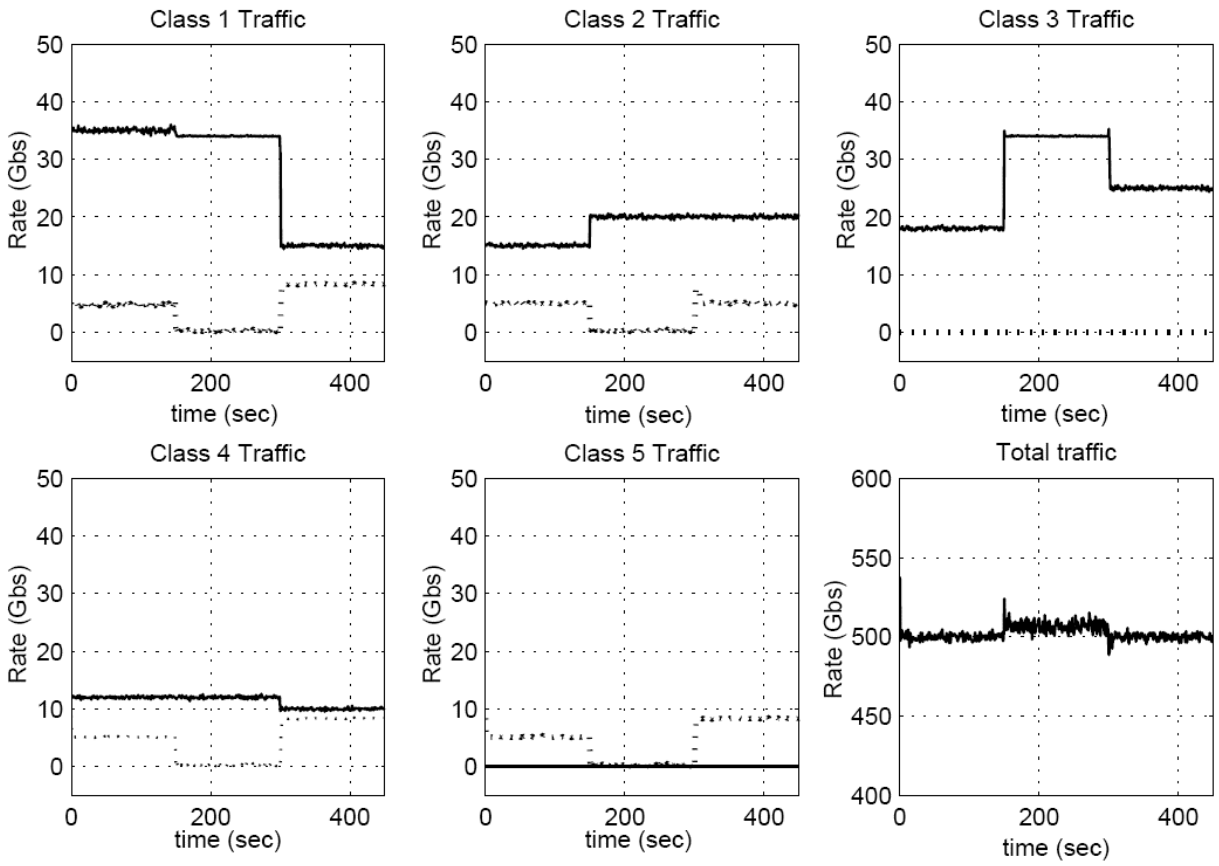
Provisioned P_{loss}= 3.2 10^{-5}



Provisioned P_{loss}=1.8 10^{-4}

- Actual steady-state losses are better than provisioned

Scenario D



- Strict isolation among HP and LP traffic without having to use large offset times or slow loss rate measurements

Numerical Example Cont.

- For $0 \le t < 150$ sec;
  - Total HP traffic demand is 400 Gbps
  - Since 400 Gbps < EC, max-min fair share vector for HP traffic is: [35, 15, 18, 12, 0].
  - Remaining capacity for LP traffic is 500 – 400 = 100 Gbps
  - It is allocated to LP flows on a max-min fair share basis, i.e.

    [5, 5, 0, 5, 5]

OBS tutorial

TCP over OBS

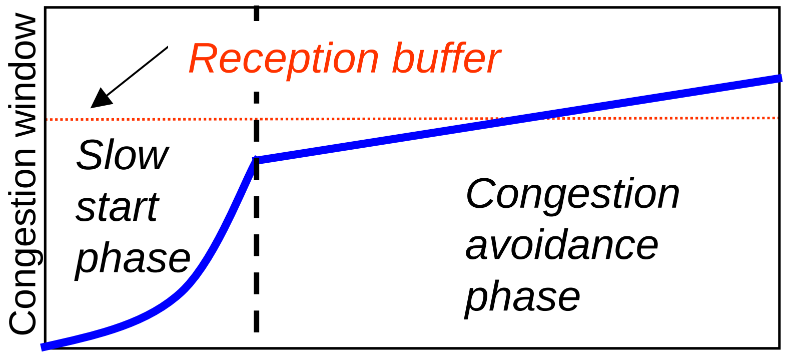
Basic TCP control functions

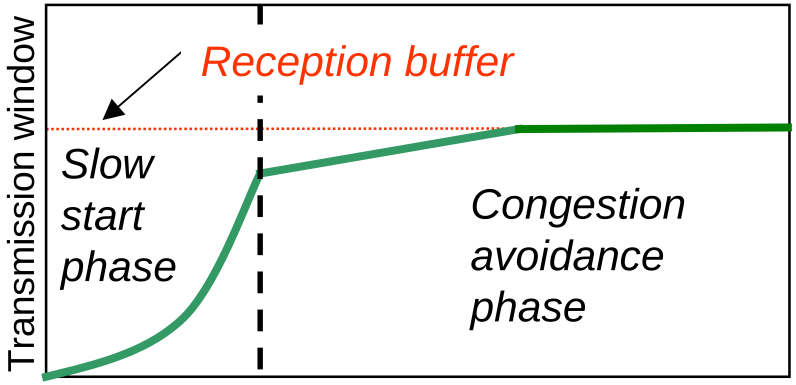
- TCP is in charge of the end to end communication
- TCP sends data in segments, which are acknowledged by the receiver.
- Flow control
  - the TCP window size is used to prevent the sender from flooding the receiver
- Congestion control
  - TCP window is dynamically updated in relation to the network state as perceived by the sender

Basic TCP control functions

- How does the TCP window work?
  - TCP sends as many segments as transmission window allows
  - Congestion window and receiver buffer determine window

    Transmission Window = min(CongWin,RcvWin)

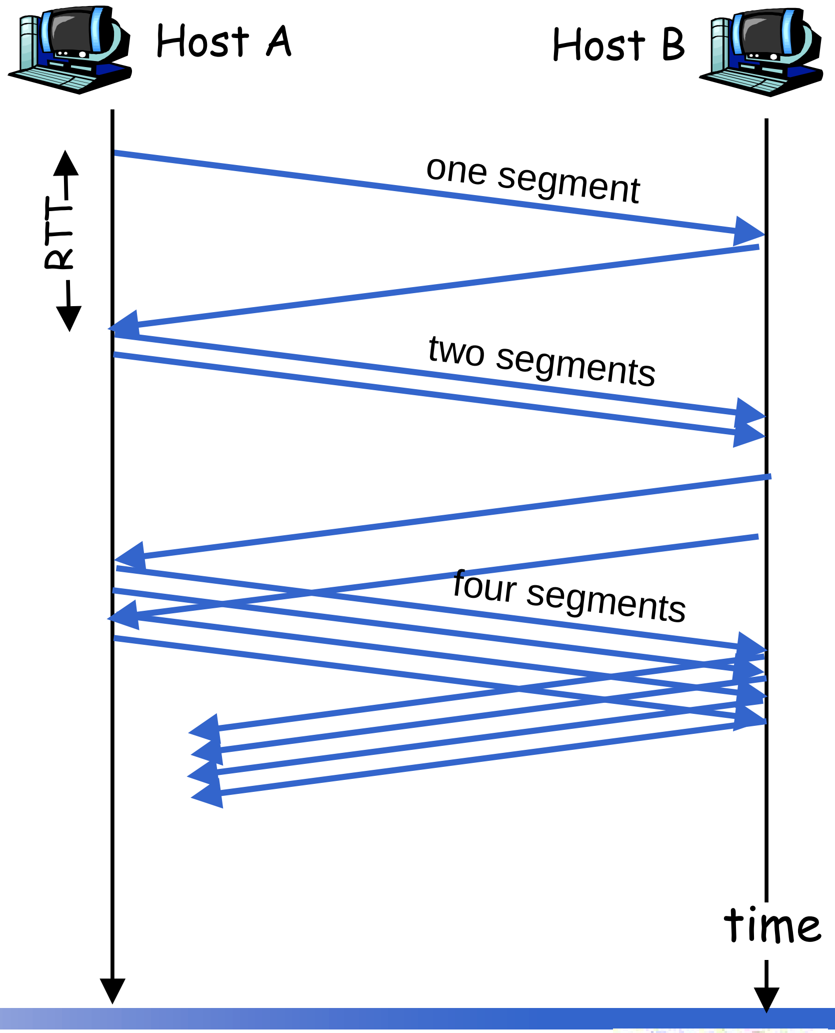




- Transmission rate: $X(t) = \dfrac{W(t)}{RTT}$

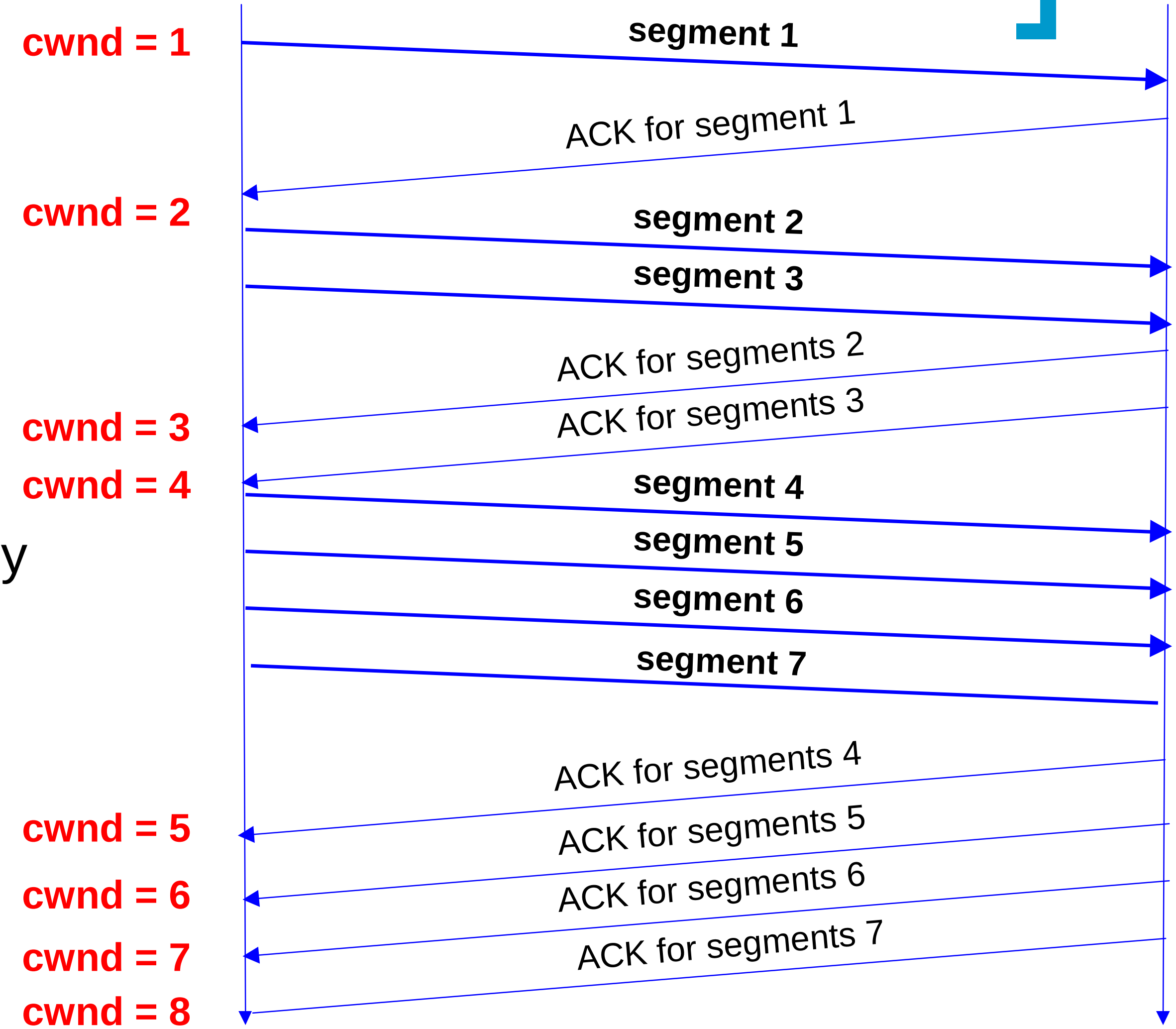
TCP Slow Start

- When connection begins, **CongWin** = 1 MSS
  - Example: MSS = 500 bytes & RTT = 200 msec
  - initial rate = 20 kbps
- When connection begins, increase rate exponentially fast until first loss event



Slow Start Example

- The congestion window size grows very rapidly
  - For every ACK, we increase CongWin by 1 irrespective of the number of segments ACK’ed
  - double **`CongWin`** every RTT
  - initial rate is slow but ramps up exponentially fast
- TCP slows down the increase of CongWin when ***CongWin > ssthresh***



cwnd = 1

segment 1

ACK for segment 1

cwnd = 2

segment 2

segment 3

ACK for segments 2

ACK for segments 3

cwnd = 3

cwnd = 4

segment 4

segment 5

segment 6

segment 7

ACK for segments 4

ACK for segments 5

ACK for segments 6

ACK for segments 7

cwnd = 5

cwnd = 6

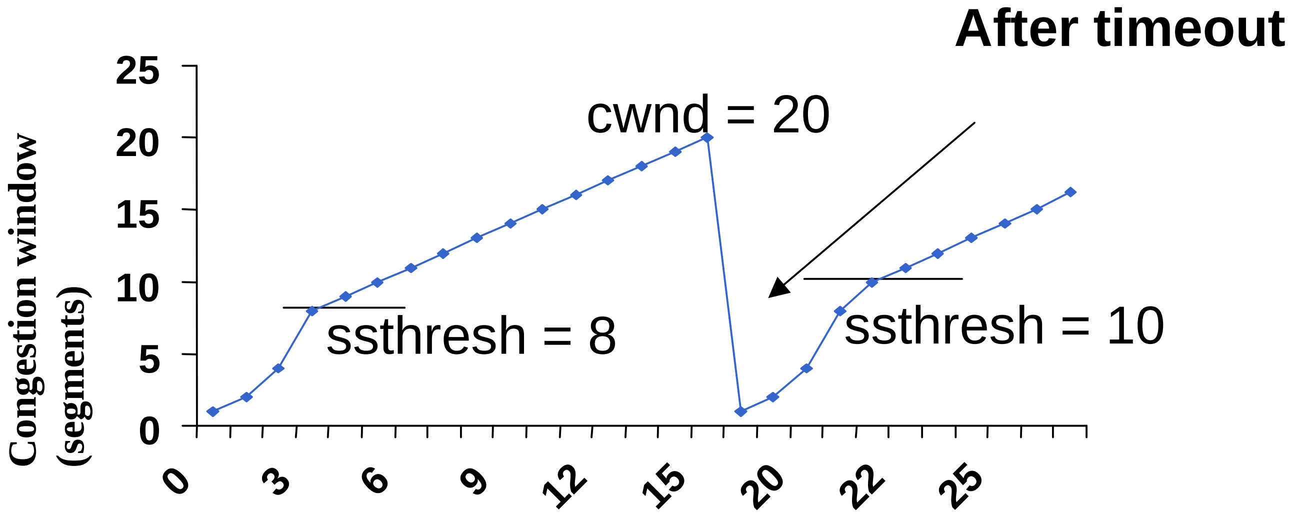
cwnd = 7

cwnd = 8

TCP loss detection and recovery

- TCP has two ways of detecting losses:
  - By Retransmission Time Out (RTO)
  - By receiving Triple duplicate ACKs

- TCP recovers from RTO by returning to slow start



Responses to Congestion

TCP interprets a Timeout as a binary congestion signal. When a timeout occurs, the sender performs:

- ssthresh is set to half the current size of the congestion window:

  ssthresh = CongWin / 2

- CongWin is reset to one:

  CongWin = 1

TCP Congestion Control

```
Initially:
    CongWin = 1;
    ssthresh = advertised window size;
New Ack received:
    if (CongWin < ssthresh)
        /* Slow Start*/
        CongWin = CongWin + 1;
    else
        /* Congestion Avoidance */
        CongWin = CongWin + 1/CongWin;
Timeout:
    /* Multiplicative decrease */
    ssthresh = CongWin/2;
    CongWin = 1;
```

Slow Start (exponential increase phase) is continued until CongWin reaches half of the level where the loss event occurred last time. CongWin is increased slowly after (linear increase in Congestion Avoidance phase).

Fast Recovery

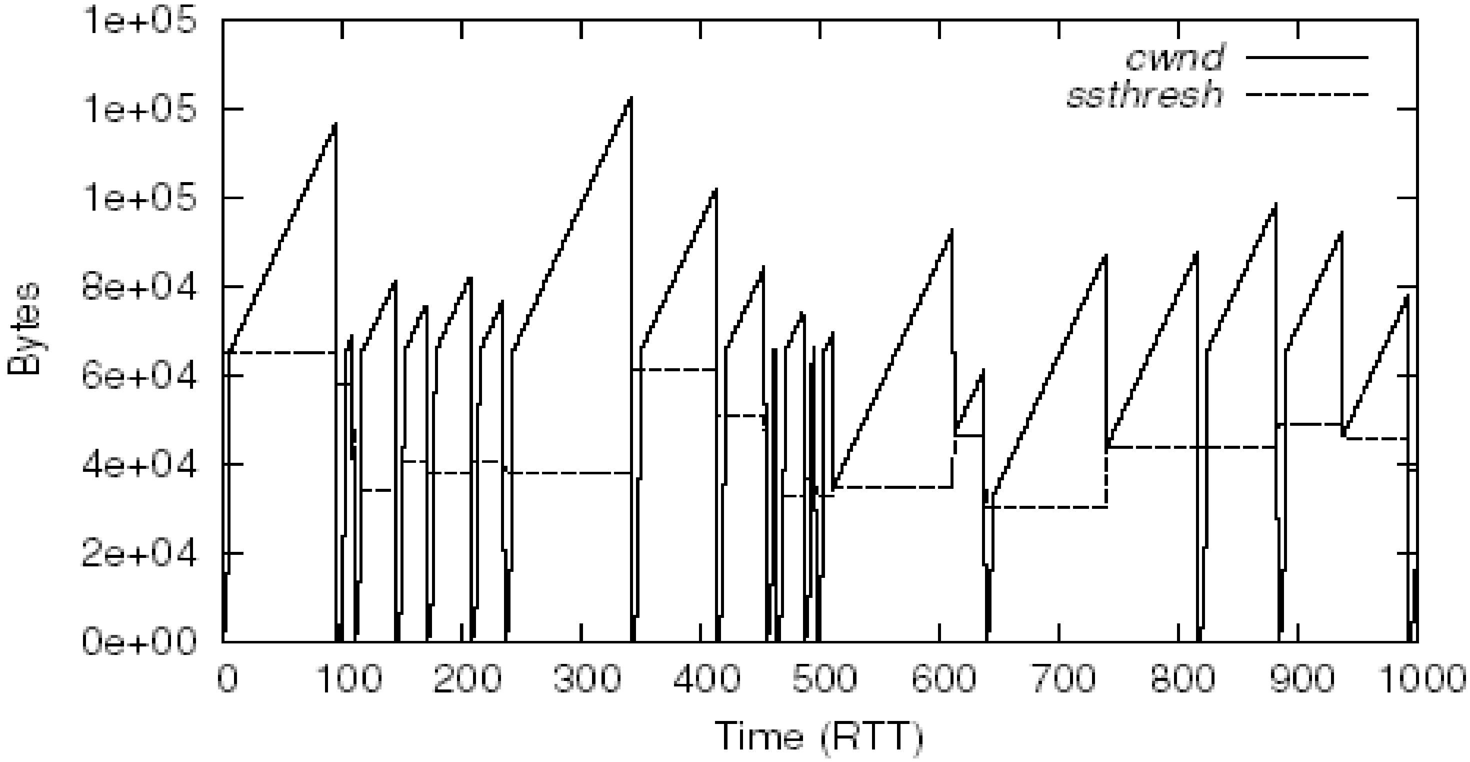
- After 3 dup ACKs (fast Retransmit):
  - ssthresh = CongWin/2
  - CongWin = CongWin/2
  - window then grows linearly
- But after timeout event:
  - CongWin = 1 MSS;
  - window then grows exponentially
  - to the threshold, then grows linearly

Philosophy:

- 3 dup ACKs indicates network capable of delivering some segments
- timeout before 3 dup ACKs is "more alarming"

TCP Congestion Control

- The evolution of CongWin and ssthresh for a TCP connection incl.
  - Slow start and Congestion avoidance
  - Fast retransmit and fast recovery, occur at time around 610, 740, 950.



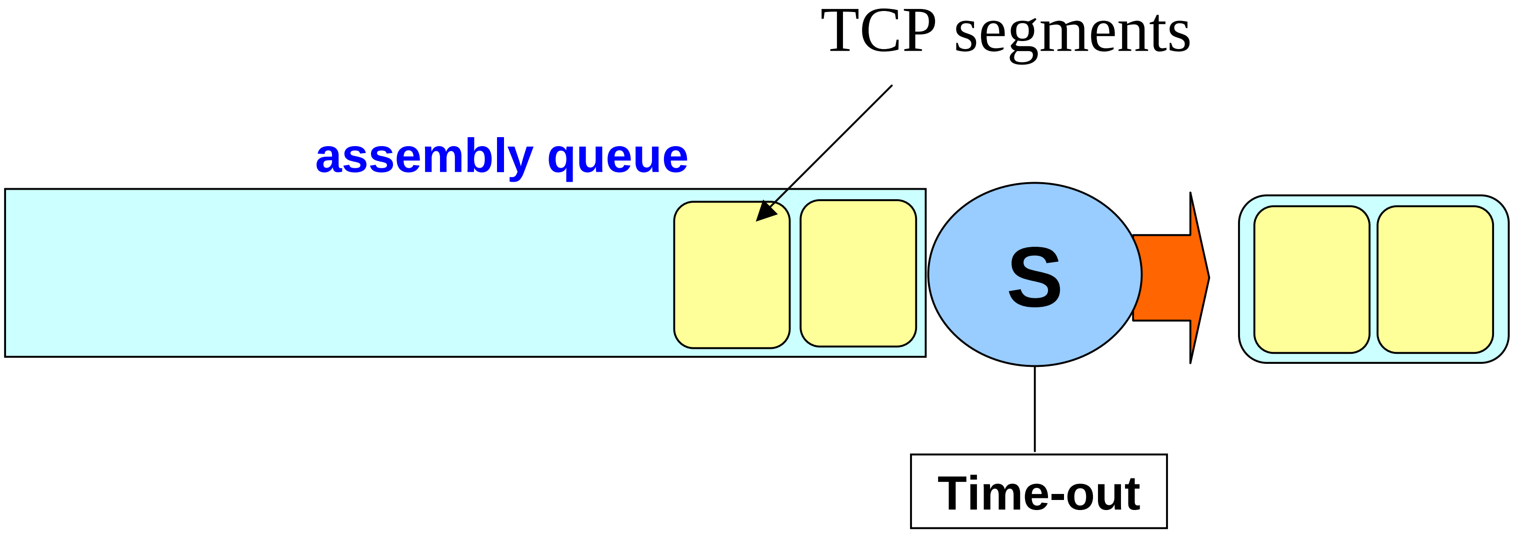
Impact of OBS network on TCP

- Edge node
  - Assembly algorithms
    - Mixed flow/ per flow
    - Time out, threshold-based
- Core node
  - Scheduling algorithms
  - Contention resolution schemes
    - Wavelength domain
    - Time domain
- Network
  - Routing algorithms
    - Deflection routing
    - QoS routing

TCP performance (throughput, fairness) is influenced by OBS networks

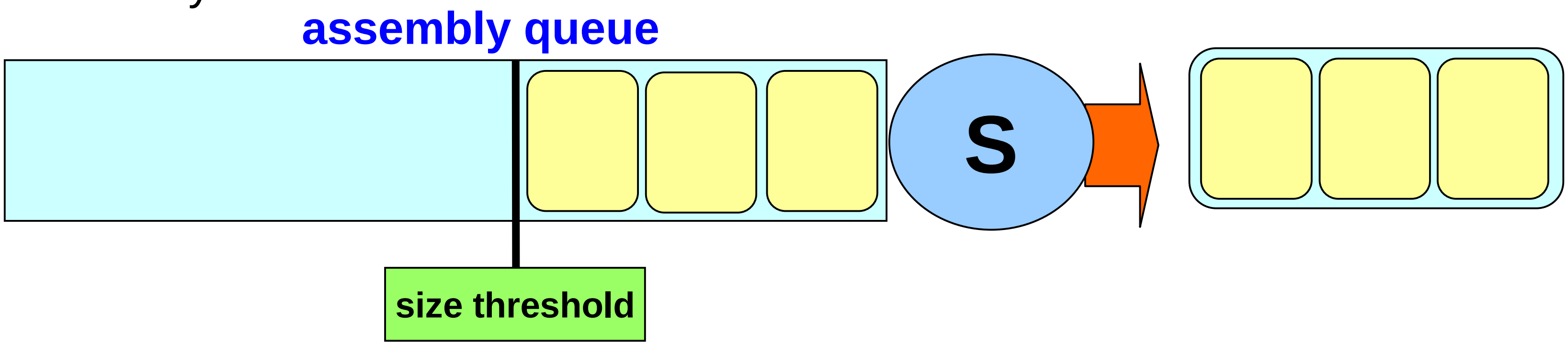
Burst assembly: timer based

- Timer-based
  - Introduction of the assembly time out, T_b, to
  - Limit the maximum assembly delay
  - reduce the negative impact on TCP performance
- When the assembly time out period expires the optical burst is sent



Burst assembly: size based

- Size-based
  - The burst is ready when L_o segments are collected
  - Typically employed with a time-out to limit assembly delays
  - A minimum burst size is also employed because of the limitations on the switching speeds
- T_b and the optical payload size, L_o must be carefully designed to
  - achieve a trade-off between assembly efficiency and assembly delay



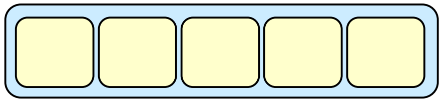
Classes of TCP sources

B_a (bit/s) access network rate, L (bit) segment length,
W_m (bit) maximum window size, T_b (s) burstification time out
A source can be classified as:

Optical bursts

- **Fast source**

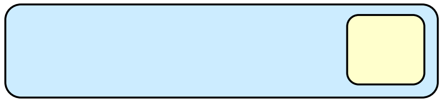
$$\frac{W_m L}{B_a} \le T_b$$



  - All segments of the maximum window are emitted in T_b

- **Slow source**

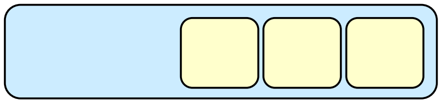
$$\frac{L}{B_a} \ge T_b$$



  - At most one segment emitted in T_b

- **Medium source**

$$\frac{L}{B_a} < T_b < \frac{W_m L}{B_a}$$



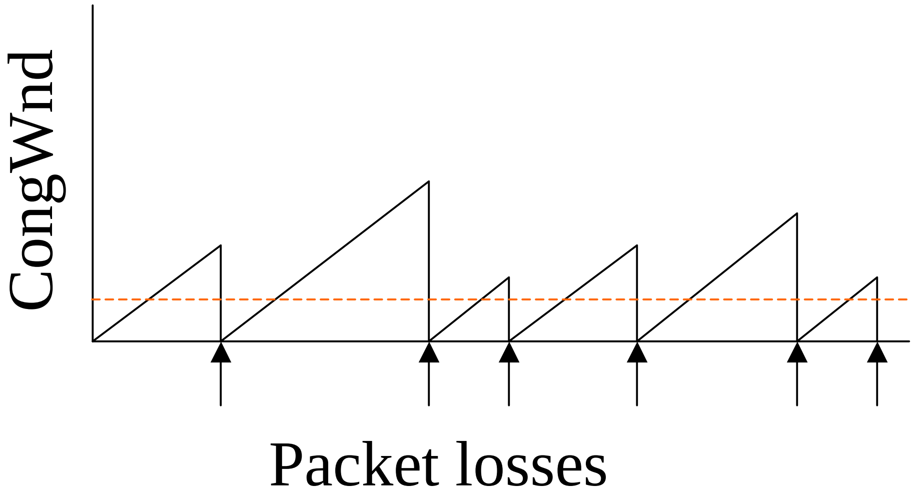
Burst loss

Burst loss is a consequence of contention in core nodes

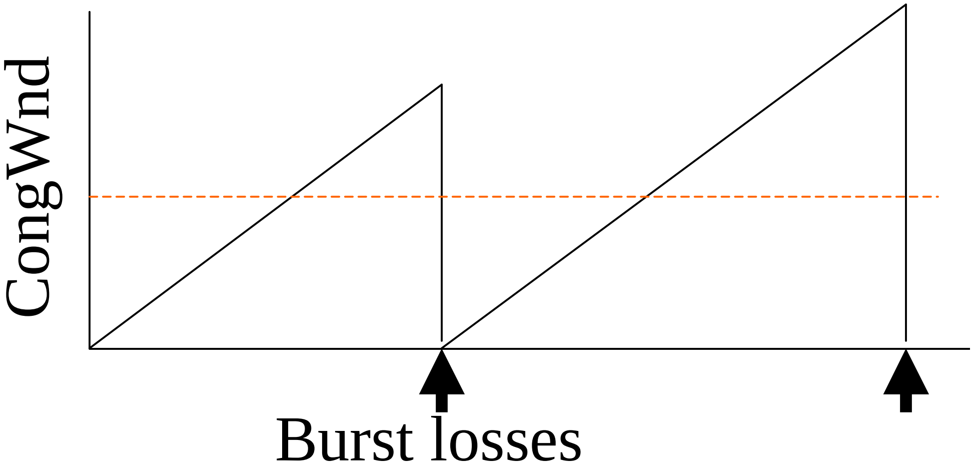
- Multiple segment losses
  - Depend on the level of aggregation of segments in a burst
- Retransmission time out is the main indication of loss for fast sources
  - Congestion window shrinks to 1 MSS when a burst is lost
- Slow sources recover mainly by means of fast recovery/fast restransmit
- Medium source
  - Recovery depends on TCP version
  - Reno recovers mainly by RTO
  - SACK recovers by fast retransmit/fast recovery

Concentrated Losses

Classical packet based network



OBS network



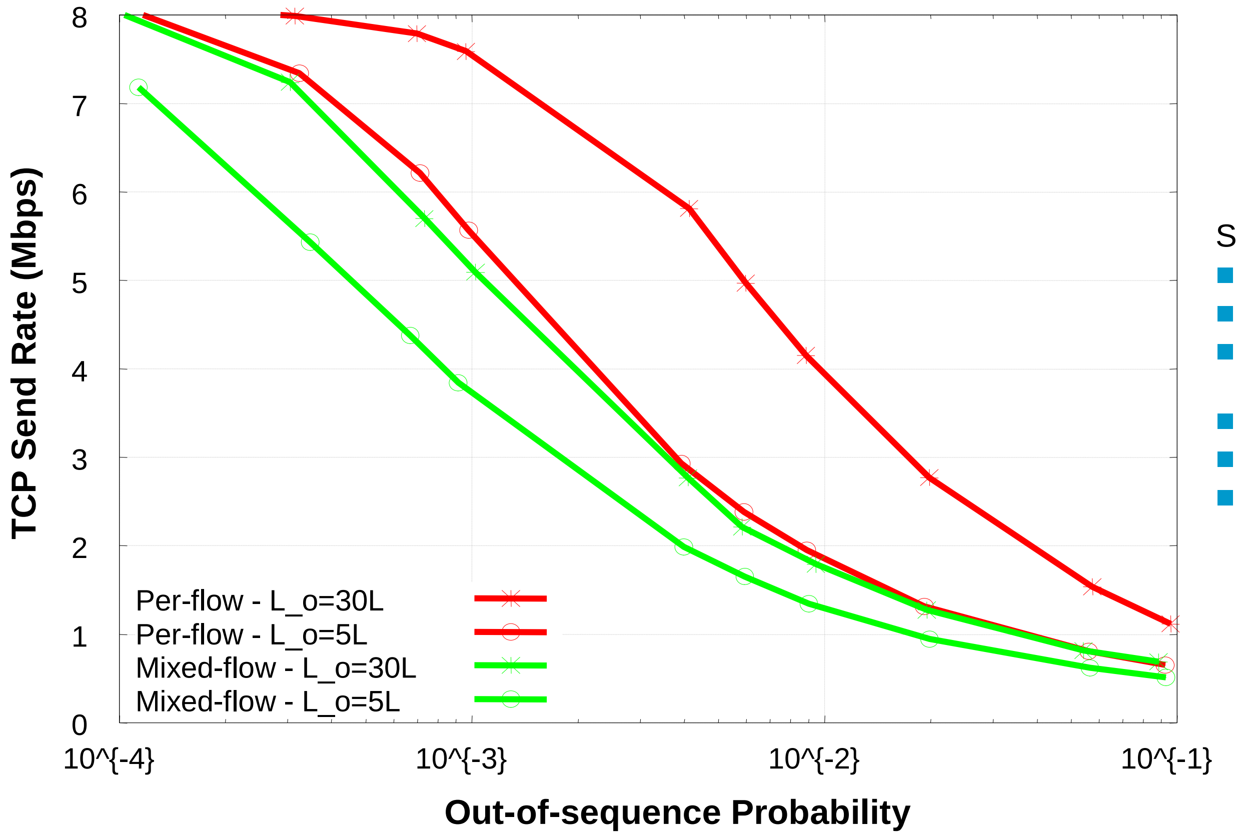
Burst assembly: key aspects

- **Delay Penalty:** Additional delay introduced by the assembler procedures may cause:
  - an increasing of both the RTT and the RTO
  - an undesirable degradation of connection throughput
- **Correlation Gain:** Concentrated losses and successful deliveries of TCP segments carried in each optical packet strongly affect end-to-end performance
  - impact on the evolution of the congestion window
  - influence on the behavior of TCP loss recovery mechanisms

Variable delay

- Delay due to burst assembly task
  - Edge architecture
  - Algorithm employed (Timer based, size-based,…)
- Delay due to the presence of FDLs
  - Core architecture
- Delay due to the scheduling algorithm

Reordering: Effects of deflection routing

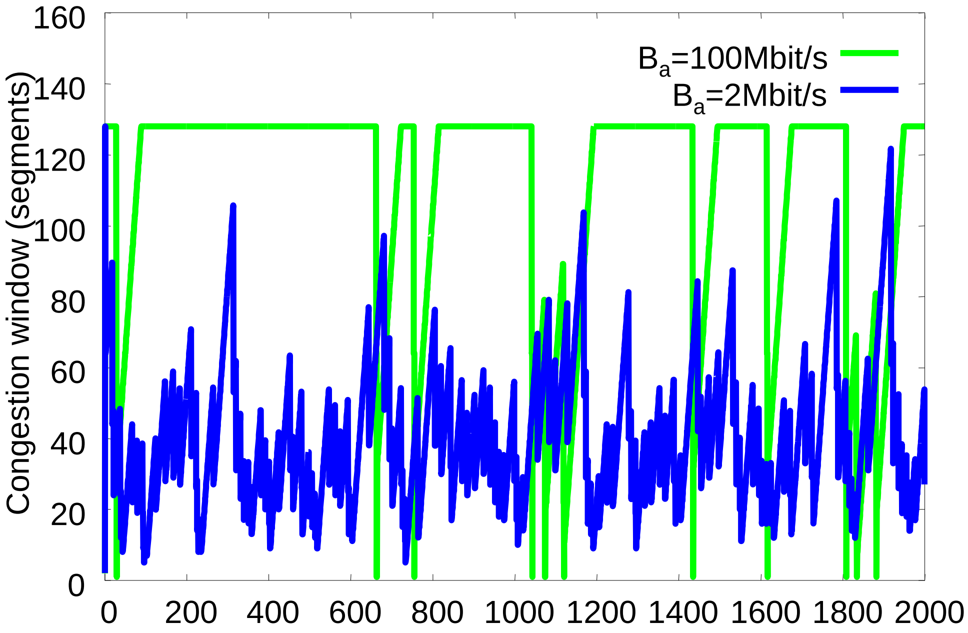


Simulated scenario:

- B_o = 2.5 Gb/sec
- RTT =600 ms
- Max window size W_{max}=128 MSS
- MSS = 512 bytes
- T_b= 3 ms
- Delay variation 30 ms

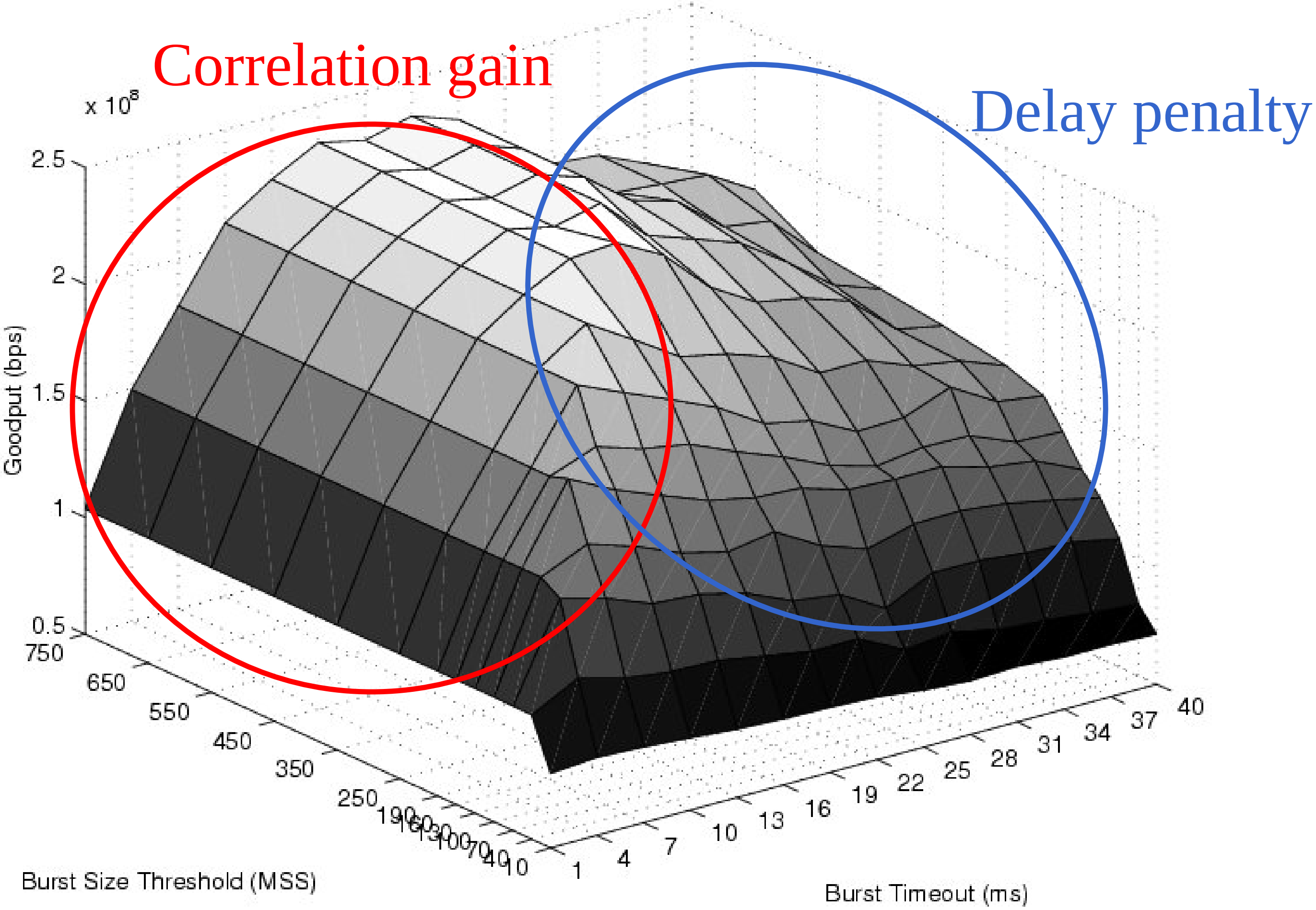
Correlation gain

- Effect related to correlated segment delivery
  - Fast/medium source
- Fast window re-opening is due to concentrated losses
- Congestion window quickly reaches its maximum value
- When present, it can significantly increase the TCP send rate



Example of congestion window evolution for two different source speed

Assembly performance: hybrid

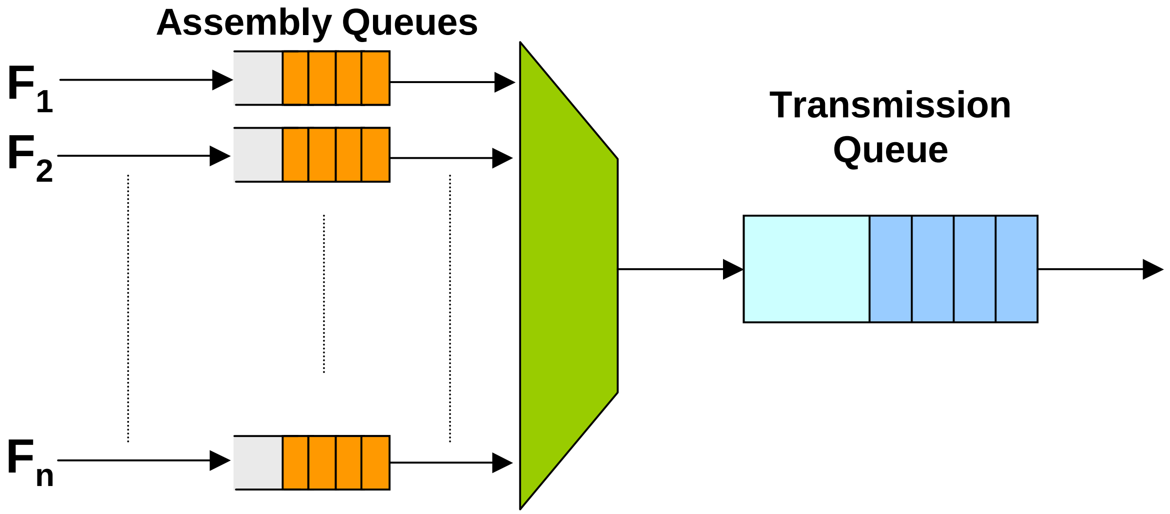


Correlation gain

Delay penalty

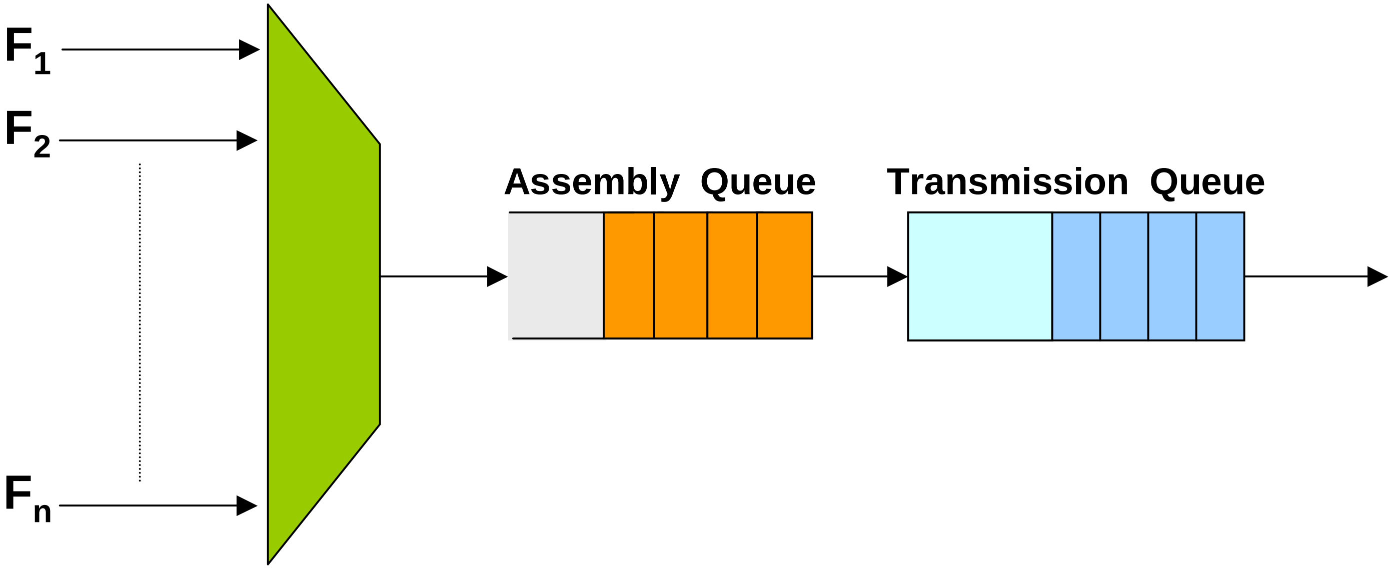
Per-flow aggregation

- Ingress per-flow queuing
- Optical bursts assembled with segments of the same flow
- An assembly time-out for each active flow is needed
- High complexity of the assembly mechanism



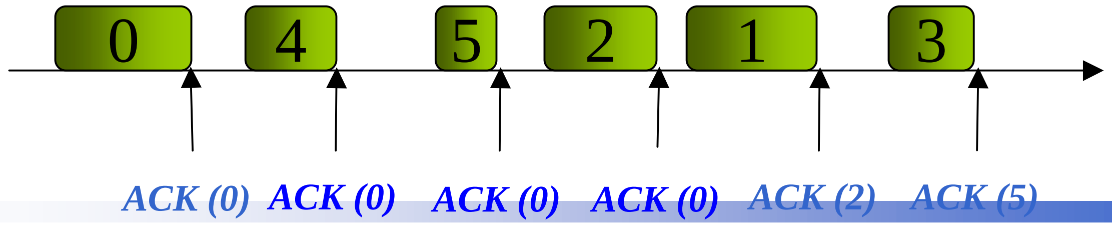
Mixed-flow aggregation

- TCP segments from different flows and with the same optical destination address aggregated in the same optical burst
- Only one assembly time-out is needed
- Lower complexity of the assembly mechanism



Burst reordering

- Burst re-ordering impacts on TCP performance
- For each *out-of-order* packet a duplicate ACK (*D-ACK*) of the last *in-order* packet is sent
- After the 3rd consecutive *D-ACK* the sender erroneously assumes the packet to be lost
  - Both *fast retransmit* and *fast recovery* are unnecessarily triggered
  - A significant drop in the available link bandwidth occurs
- In networks with large bandwidth-delay product only a small percentage of reordering can significantly affect the application throughput

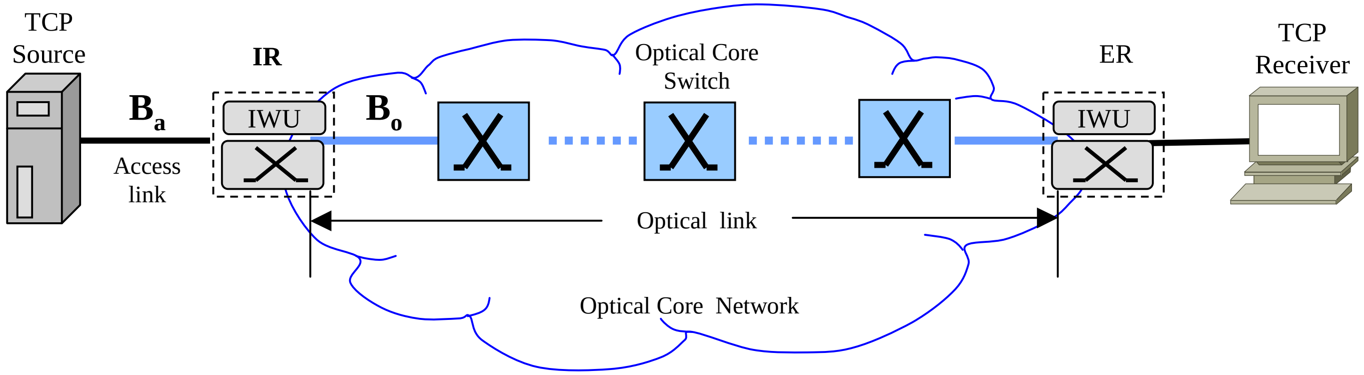


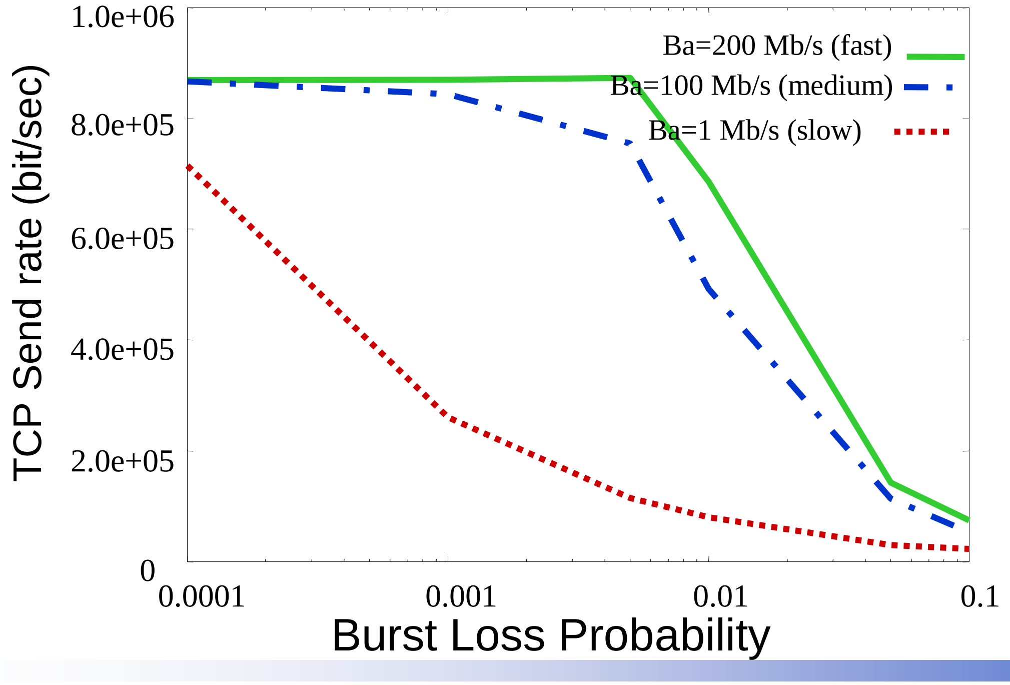
Packet reordering of 3 locations

How burst reordering arises

- Deflection routing
  - The forwarding path changes and also the propagation delay
  - Time shifts of the order of some ms
- Fiber Delay Lines in switches
  - Time shifts of the order of the average burst size (some µs)

TCP send rate results for per-flow assembly



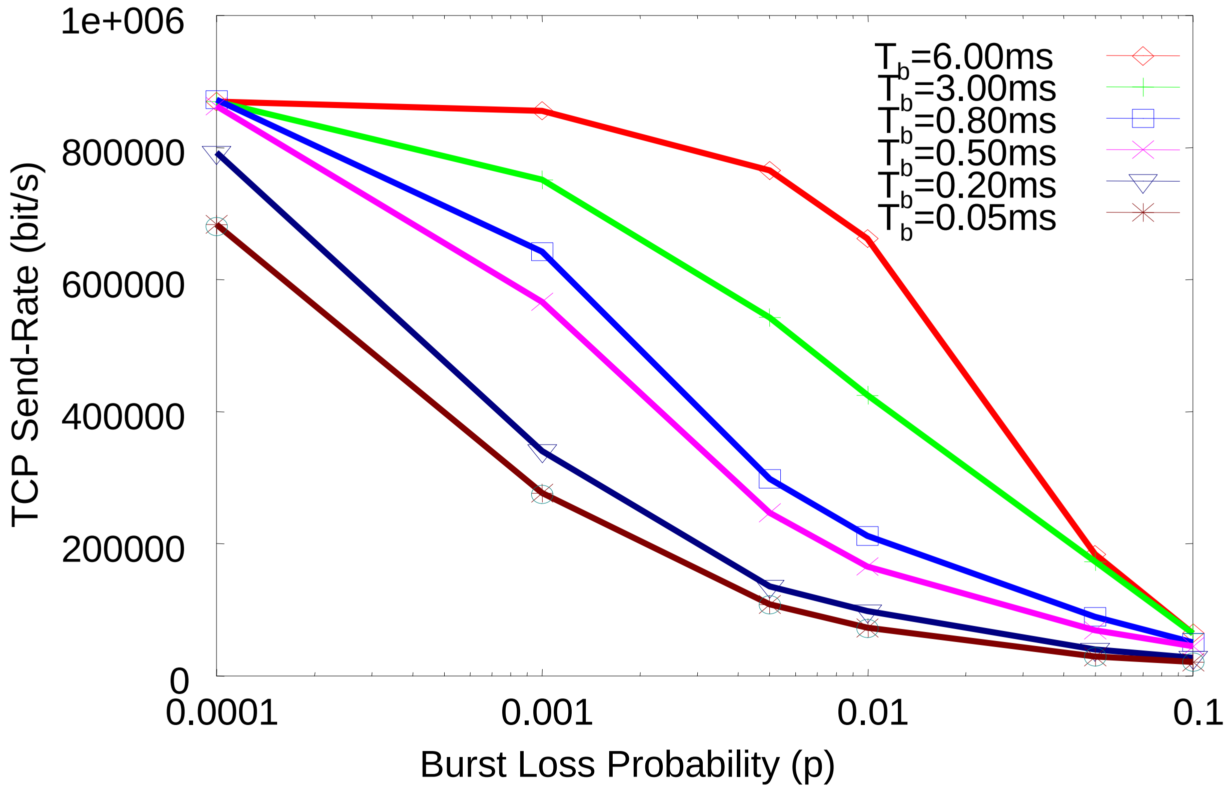


Simulated scenario:

- B_o = 2.5 Gb/sec
- RTT =600 ms
- Max window size W_{max}=128 MSS
- MSS = 512 bytes
- T_b= 3 ms

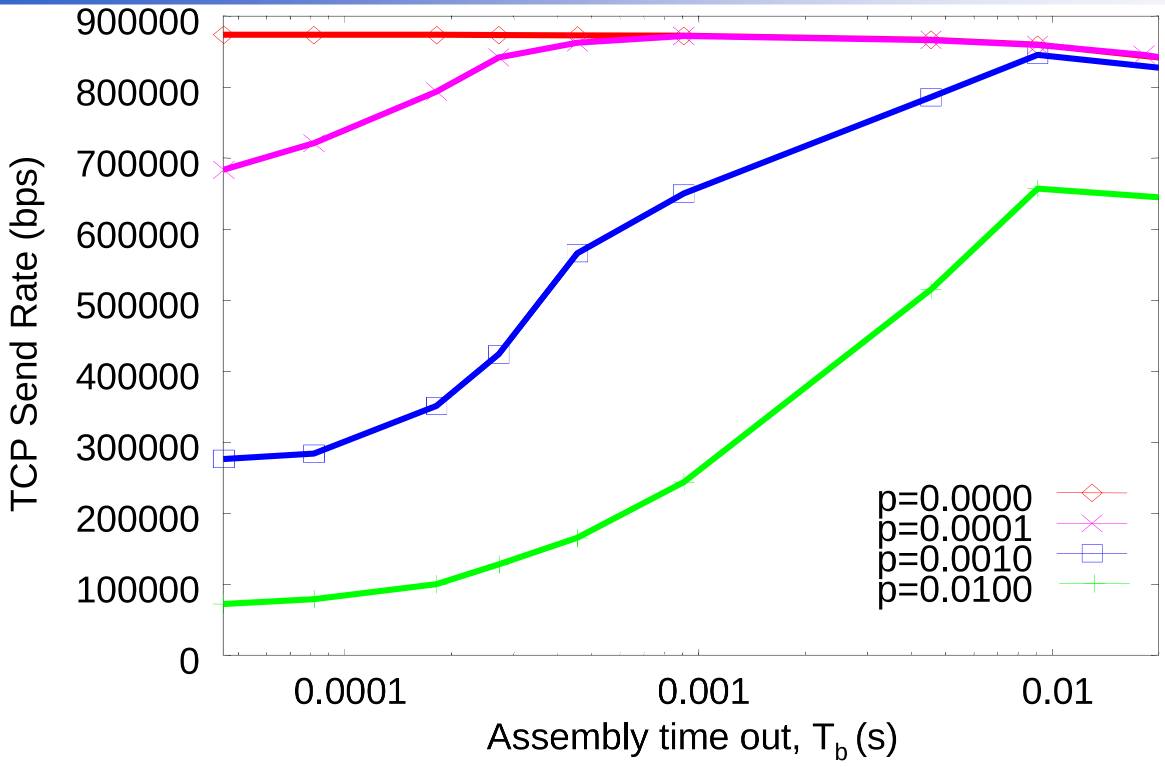
More segments are in a burst, the higher the TCP performance

B_{TCP} vs. Loss Probability varying T_b



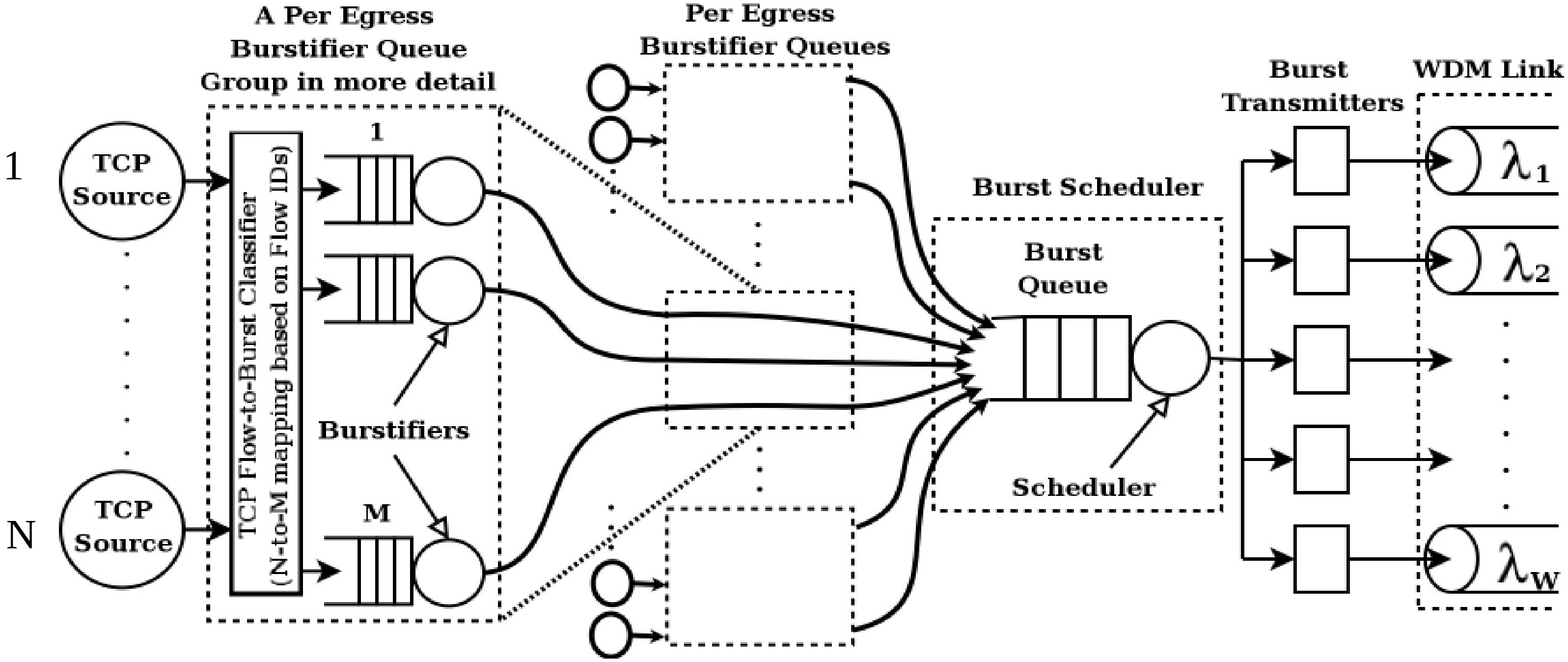
- The number of segments assembled in the optical packet depends on the value of T_b
- High correlation gain can be obtained when large number of segments can be merged within the optical payload, large T_b

B_{TCP} vs. T_b varying Loss Probability



- Lower values of T_b lead to a slow behavior of TCP source
- As T_b increases the TCP send rate grows due to the higher level of segment aggregation
- The send rate gets higher up to a maximum related to the maximum window size

Effect of # of Burst Assemblers



M = 1: Per-destination burst assembly

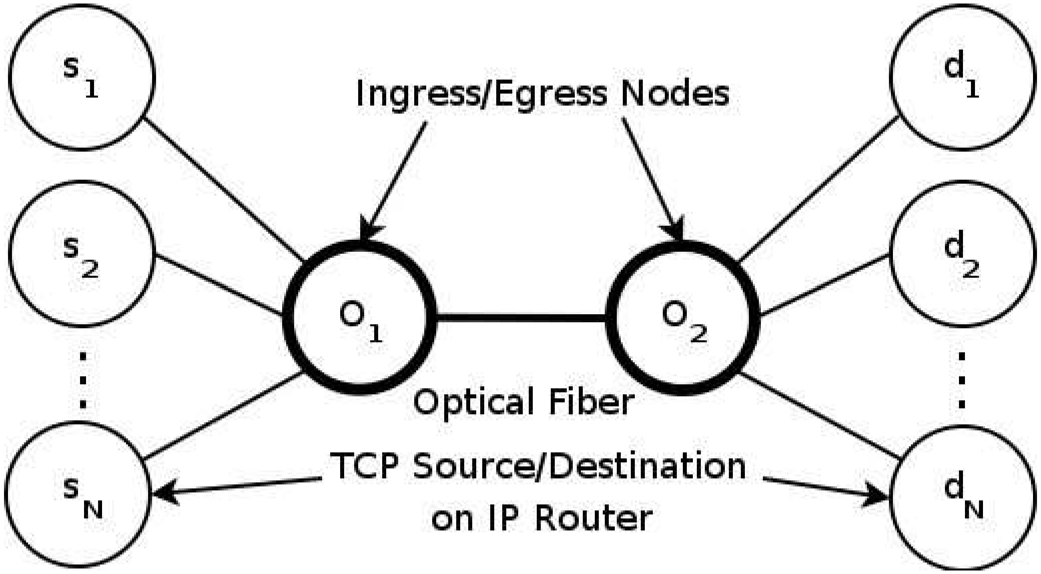
M = N: Per-flow burst assembly

1 < M < N: Tradeoff between performance and complexity

Bernoulli Loss Model

- Burst losses at a core switch occur independently with probability p
- For example, burst independent losses may occur at a switch
  - if all bursts are destined for the same egress node, or
  - if the per-hop processing delay is negligible with respect to the minimum burst size

Bernoulli Loss Model Simulation

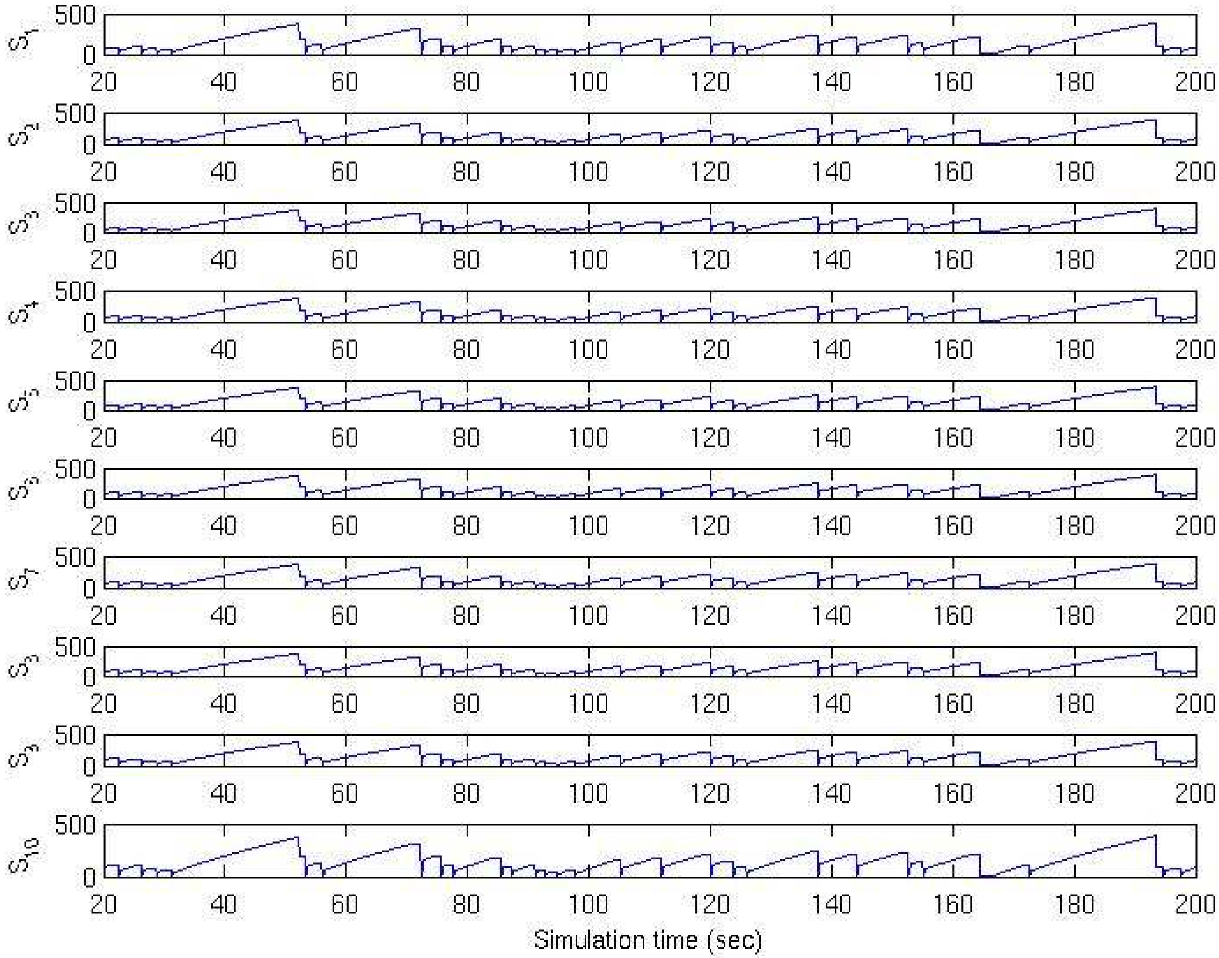


Burst losses are randomly inflicted with probability p
on the core link connecting the ingress and egress nodes

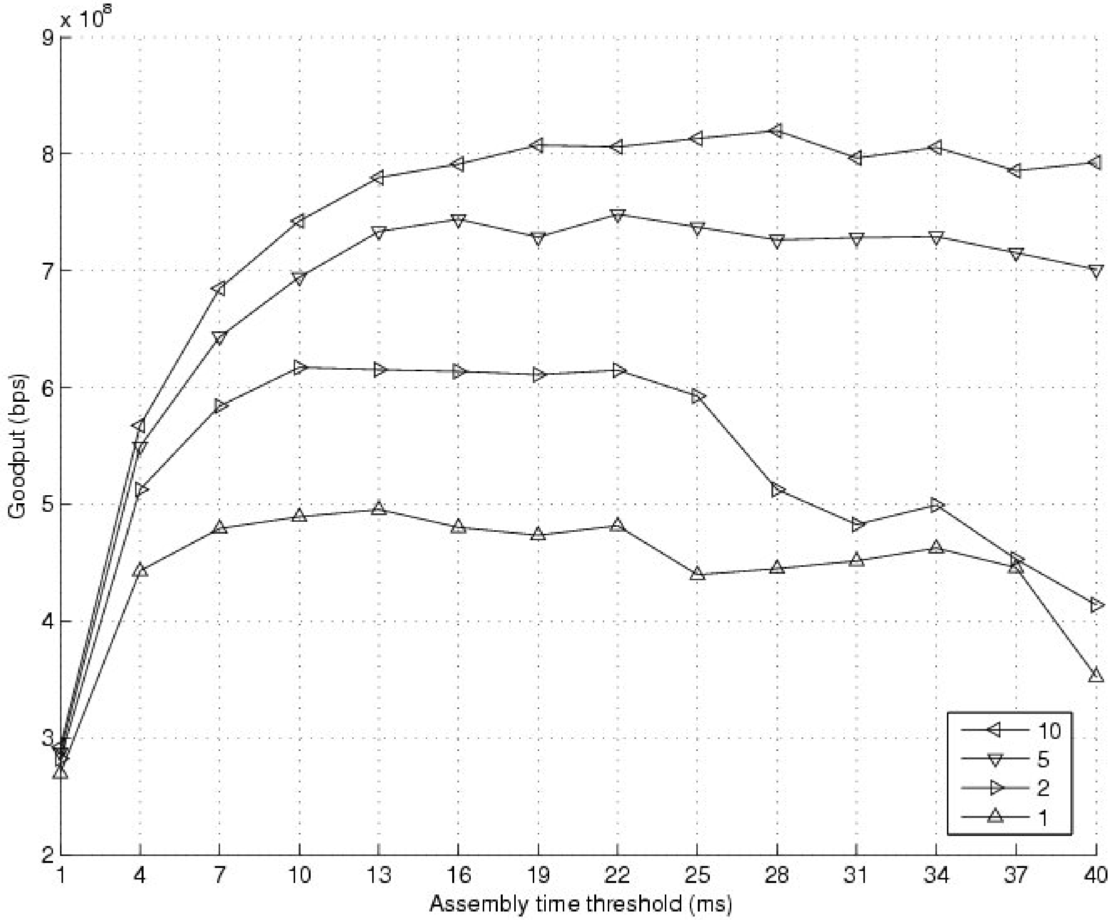
Flow Synchronization

N = 10

M = 1

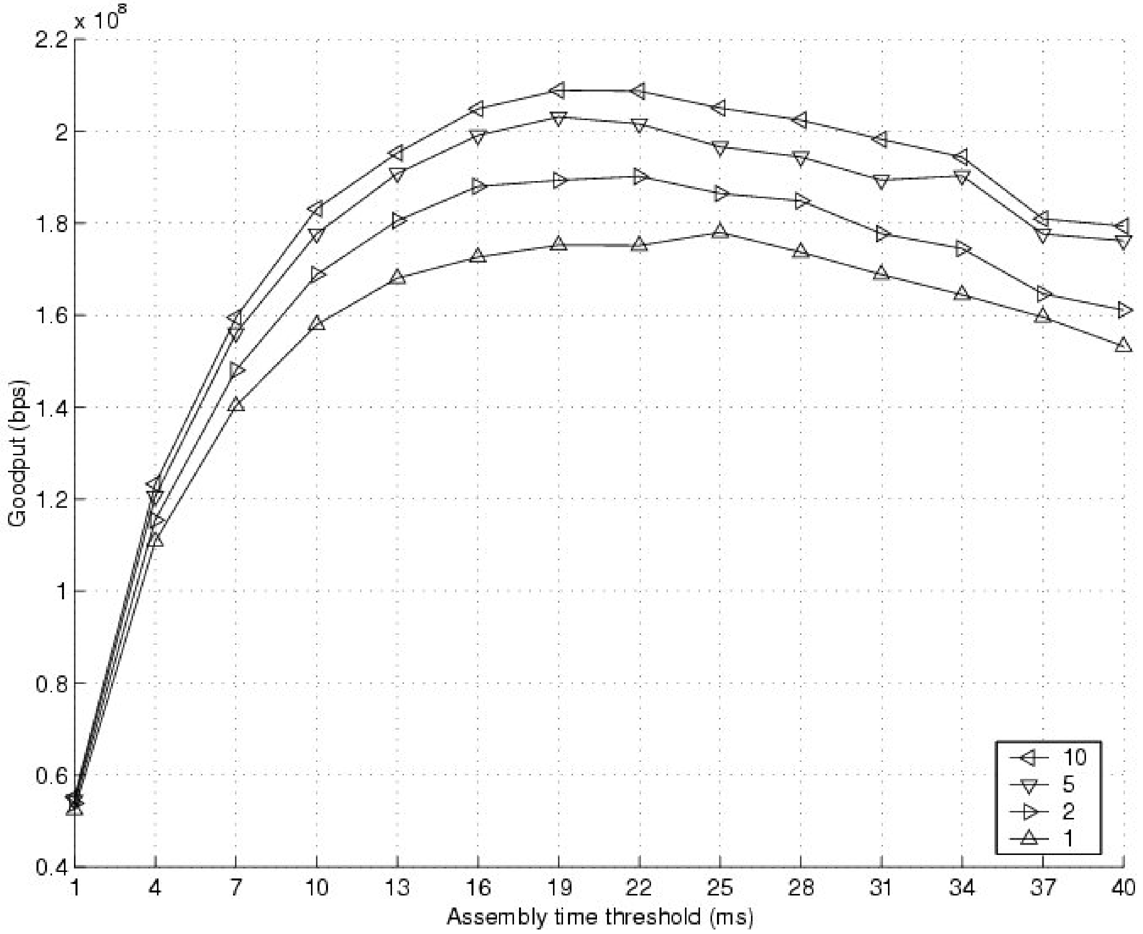


Increasing # of Burstifiers



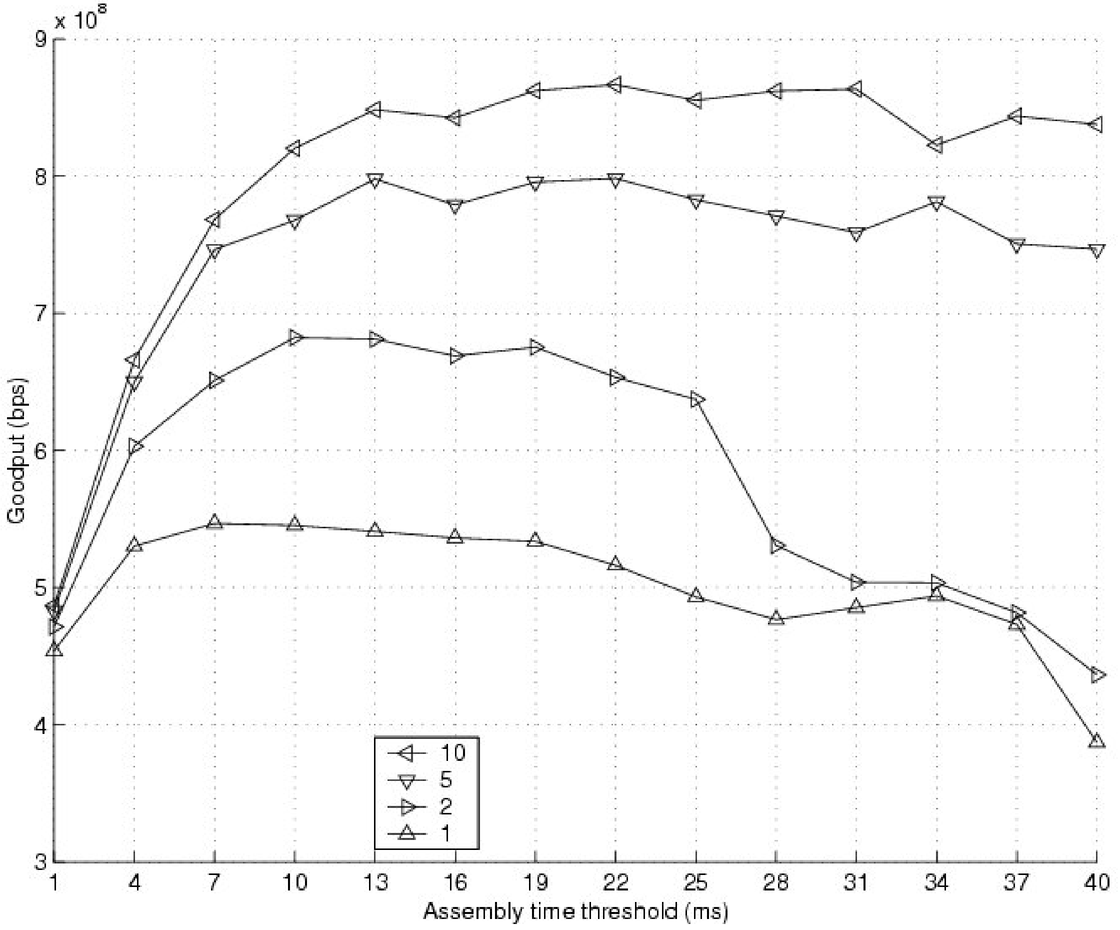
TCP Reno, N = 10, p = 1e-3

Increasing # of Burstifiers



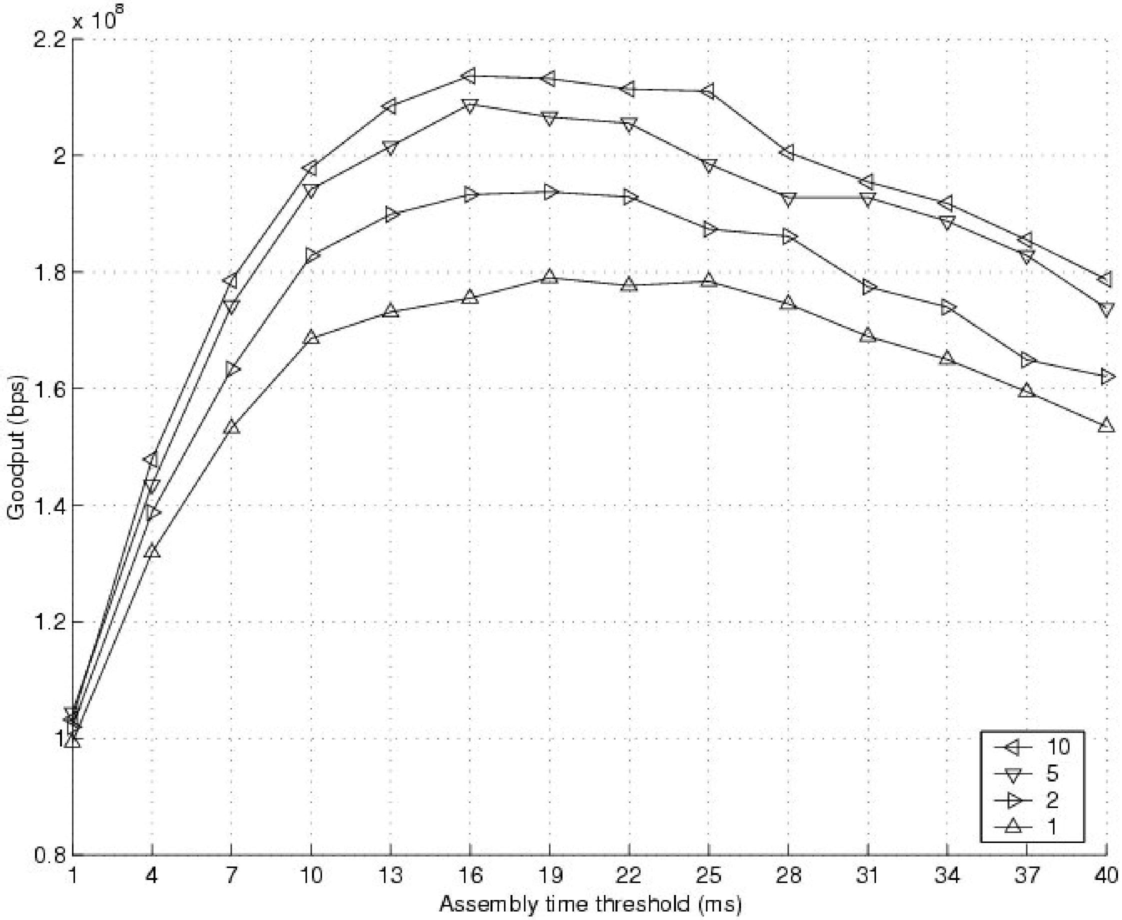
TCP Reno, N = 10, p = 1e-2

Increasing # of Burstifiers



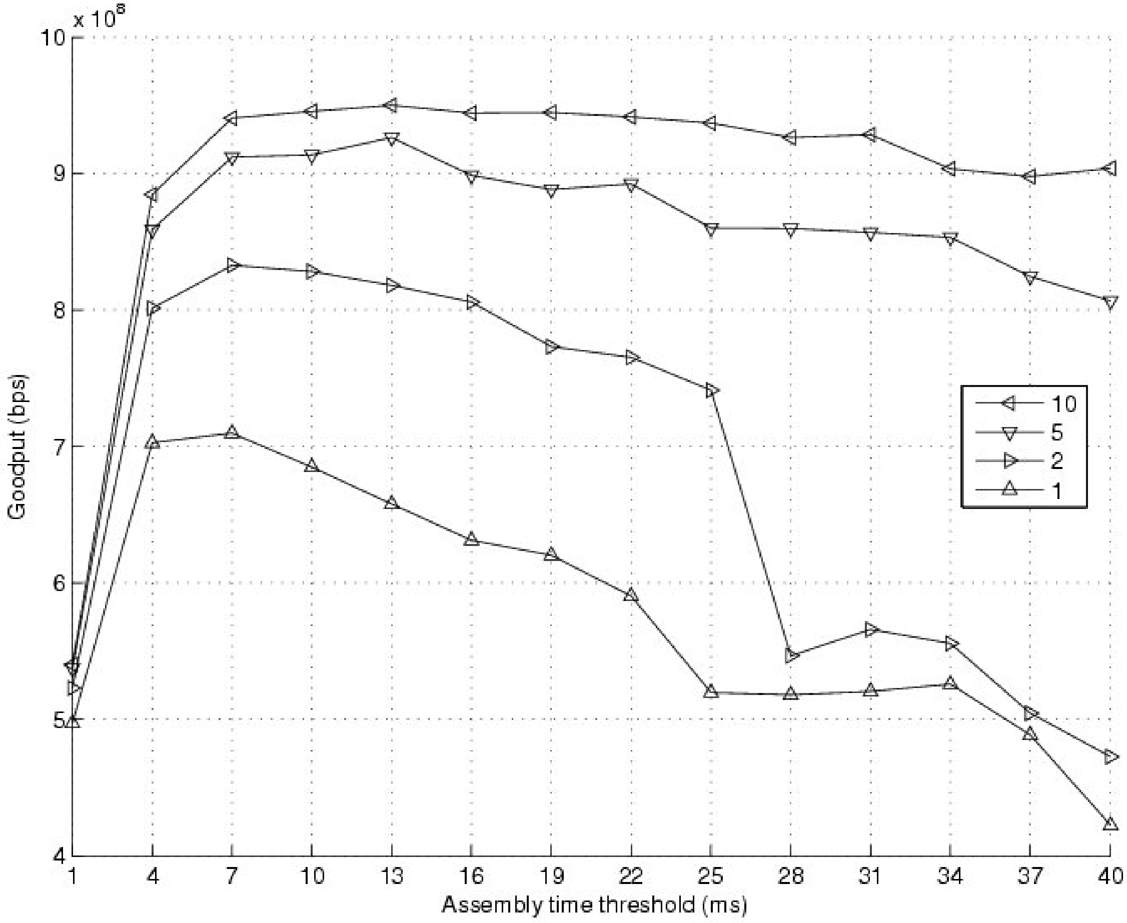
TCP NewReno, N = 10, p = 1e-3

Increasing # of Burstifiers



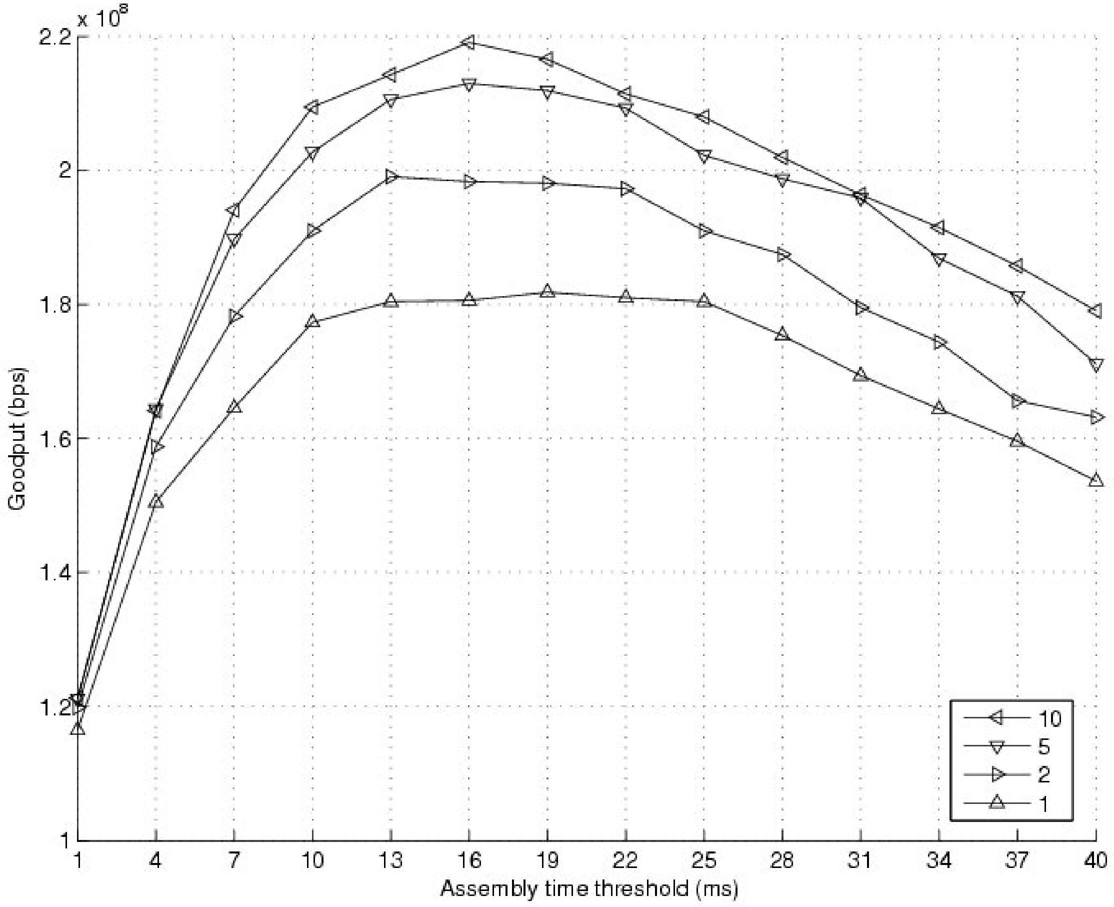
TCP NewReno, N = 10, p = 1e-2

Increasing # of Burstifiers



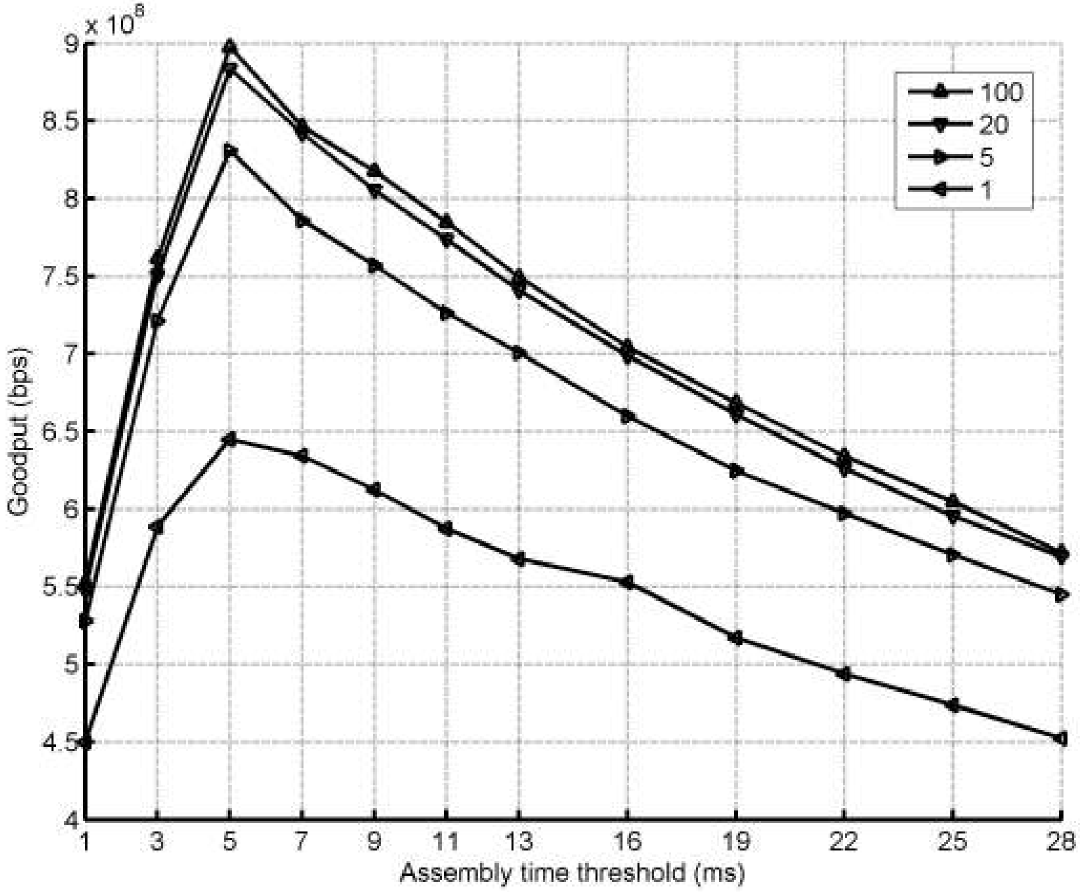
TCP Sack, N = 10, p = 1e-3

Increasing # of Burstifiers



TCP Sack, N = 10, p = 1e-2

Increasing # of Burstifiers

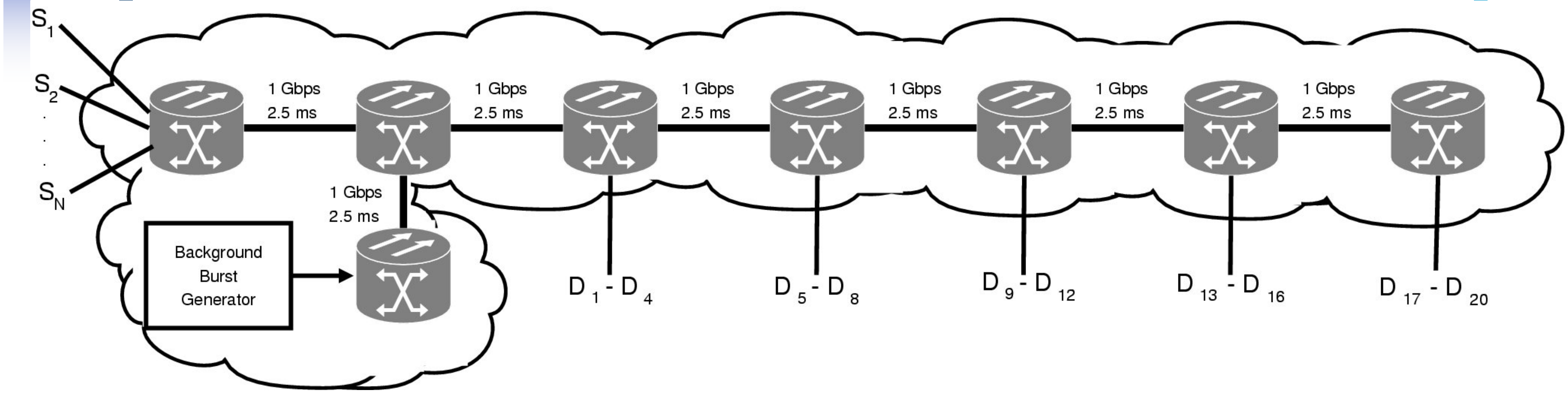


TCP NewReno, N = 100, p = 1e-3

More Realistic Loss Model

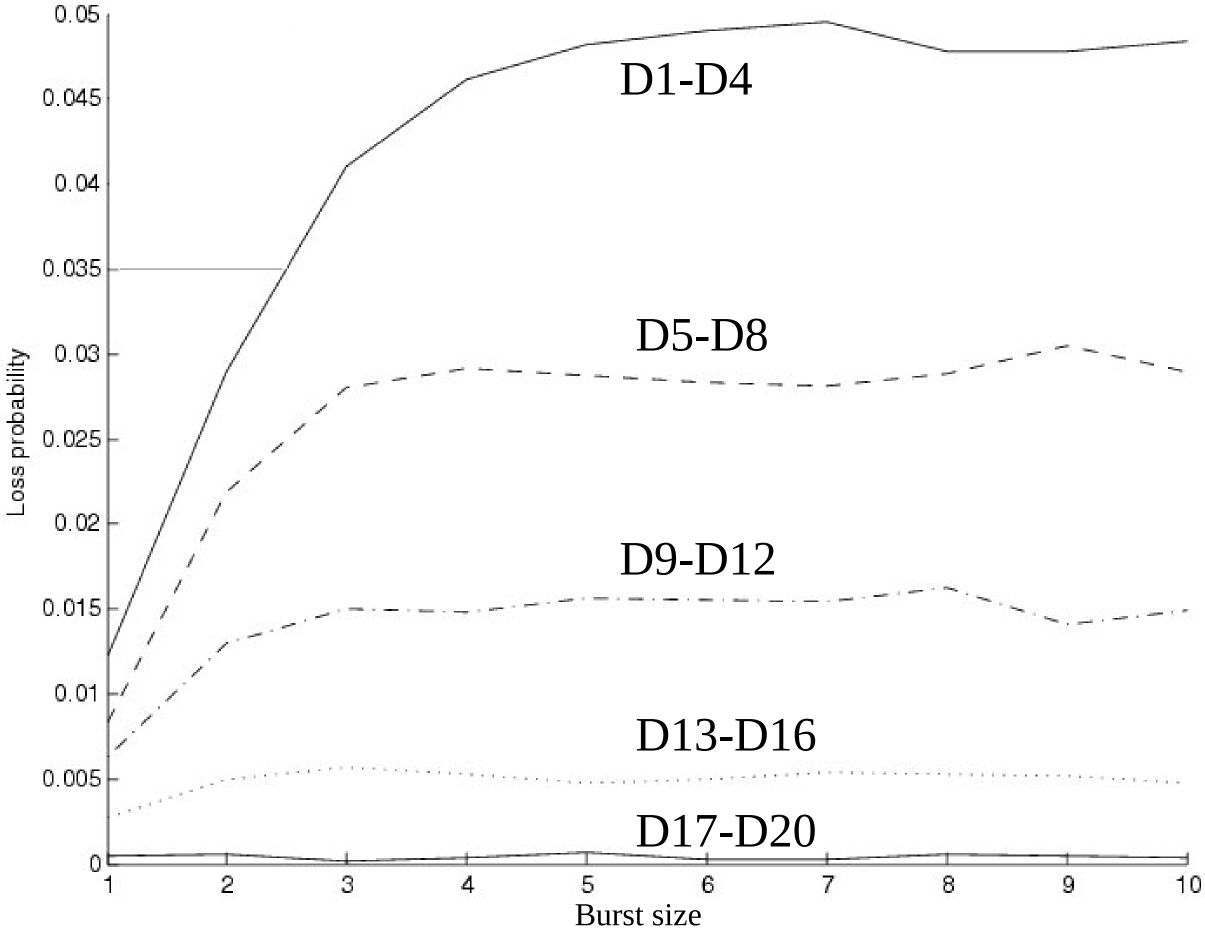
- Burst losses depend on some other factors
  - Burst length: larger bursts are more likely to be lost (assuming a void filling scheduling algorithm is used)
  - Residual offset: burst have a larger loss probability as they get closer to the egress node (assuming JET signaling scheme is used)

Realistic Loss Model Simulation



- Burst losses occur due to scheduling conflicts
- Background burst generator generates bursts according to a Poisson process with exponentially distributed burst lengths and uniformly distributed destinations
- N = 20 TCP sources generate packets destined for 5 egress nodes
- LAUC-VF scheduling algorithm is used

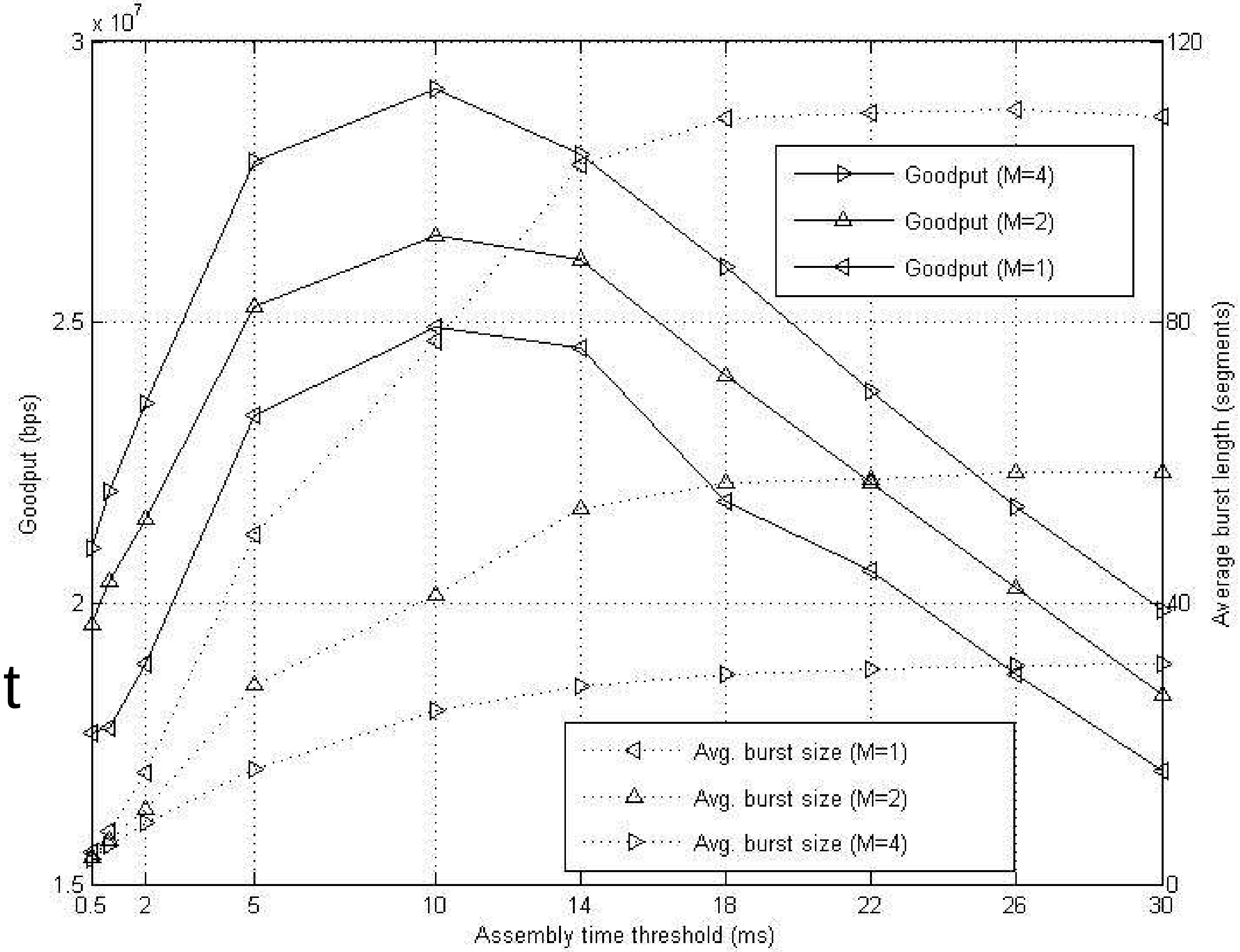
Burst Loss Probability



- Bursts destined for the furthest node experience more loss
- Loss probability increases with the burst size
- Bursts destined for the closest node experience burst length independent loss

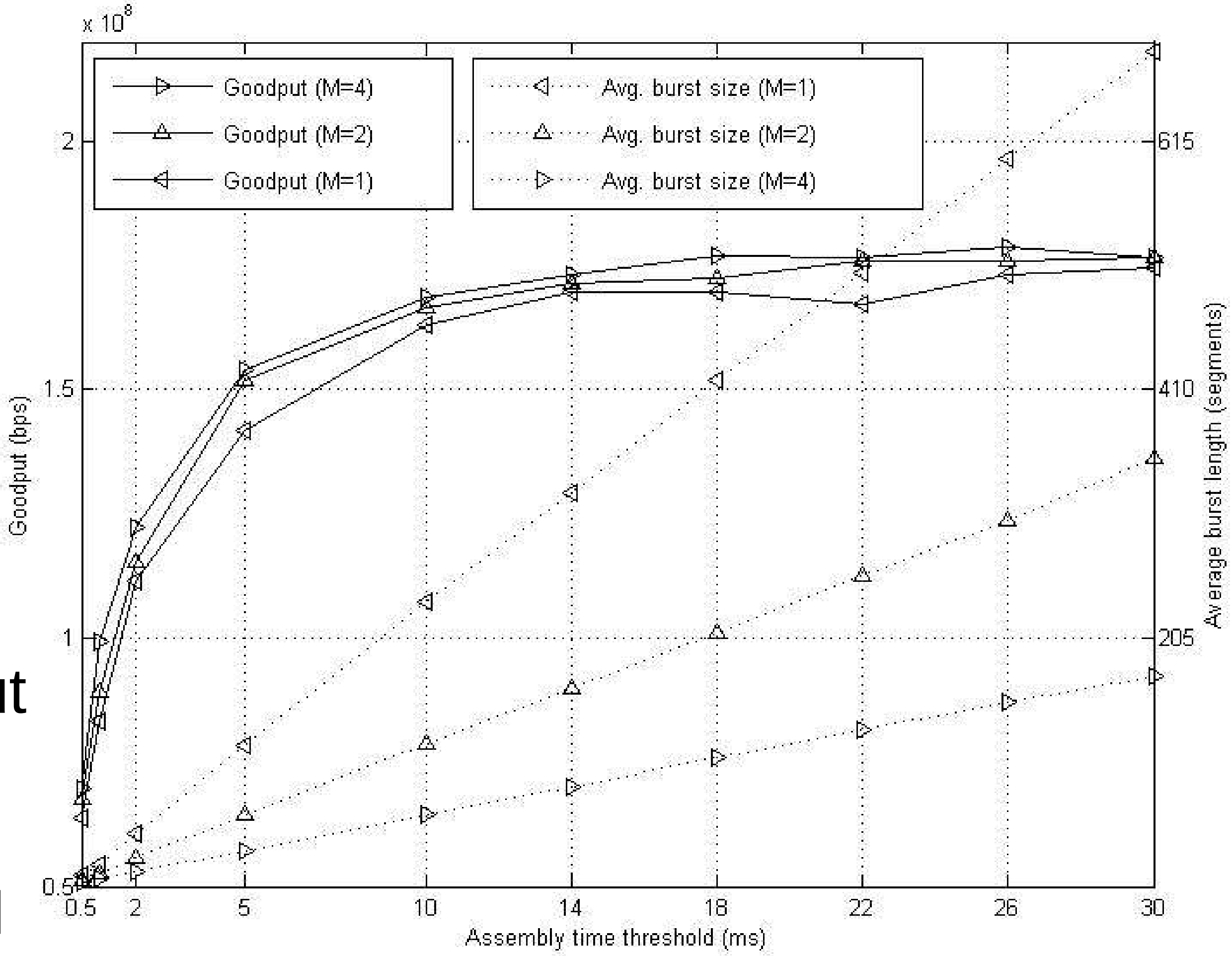
Loss Probability

- D1-D4
- Burst size saturates due to frequent losses
- Correlation gain saturates and delay penalty dominates as assembly timeout increases
- Goodput increases with M



Loss Probability

- D17-D20
- Burst size increases with assembly delay
- Correlation gain continues to increase and balances delay penalty as assembly timeout increases
- Goodput increases with M



Summary

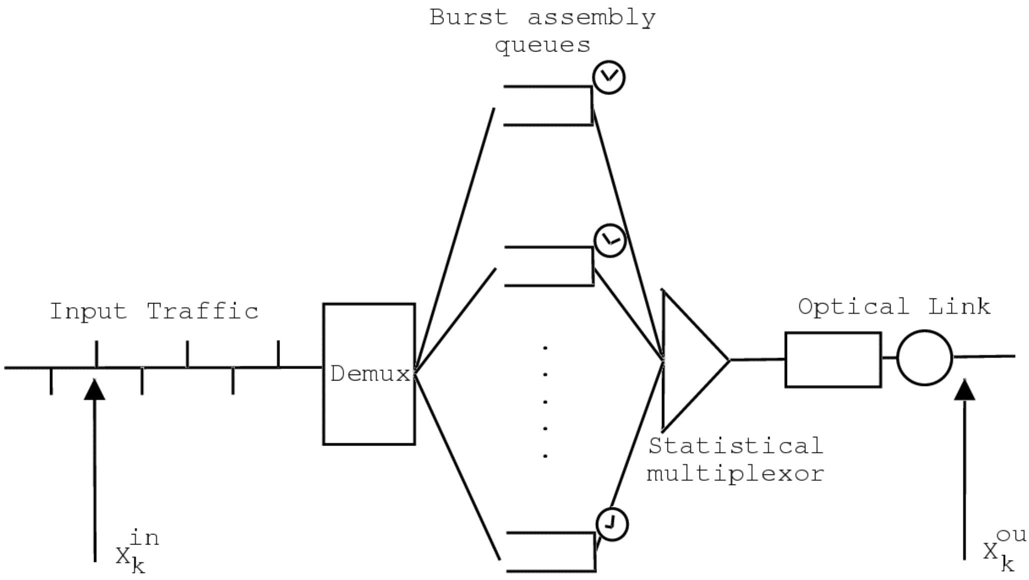
- The optical packet assembly mechanism strongly impacts TCP performance
- TCP throughput is influenced by several factors such as
  - Burst loss probability
  - correlation gain
  - assembly delay
  - # of burstification buffers
  - TCP version
  - Out-of-order packet arrivals

Wrapup

Traffic models for OBS

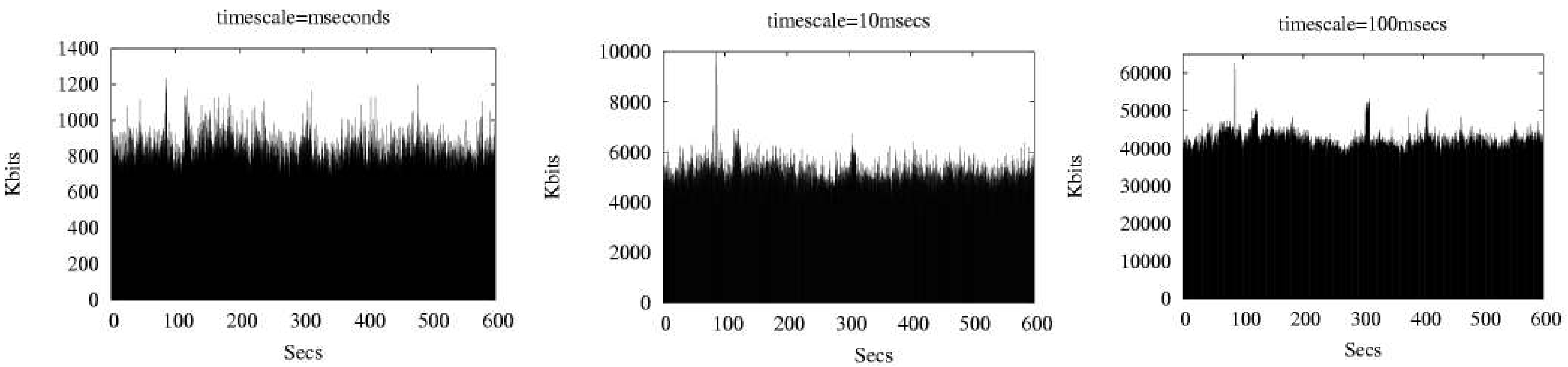
- The generated OBS traffic depends on
  - – Burstification algorithm
  - – Input traffic features:
    - Long-range dependence
    - Instantaneous burstiness
- No single traffic model can portray all possible scenarios!

Bustification algorithms



- Time-based, burst-size based or mixed-time-size based
- In all cases input traffic goes through demultiplex and then burst formation queues

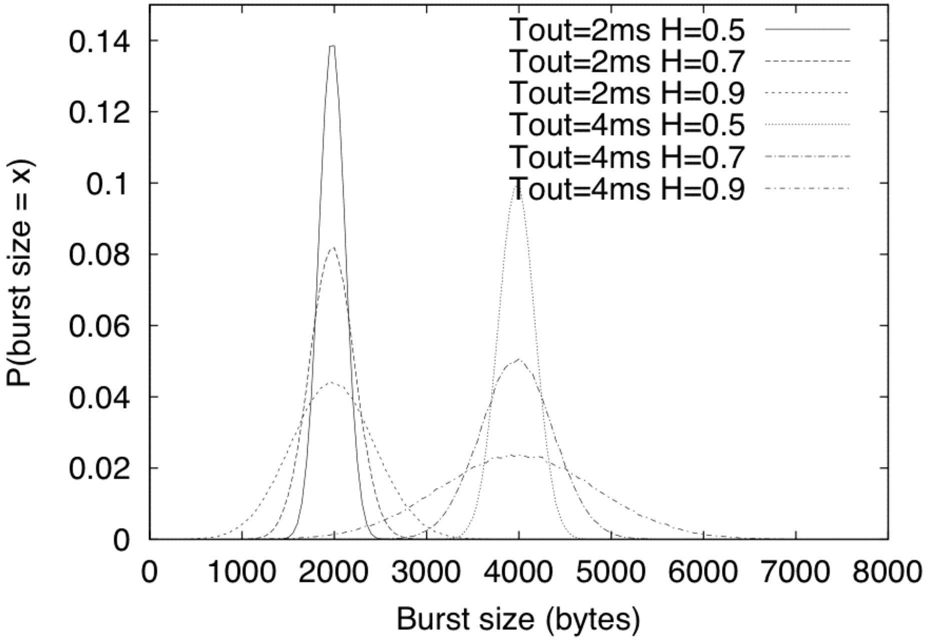
Traffic models



- Long-range dependence happens from a cutoff timescale, beyond which traffic may show independent increments
- For time-based burstifiers only the number of bytes per interval matters
- For burst-size-based burstifiers the packet arrival dynamics matter

Burst size distribution

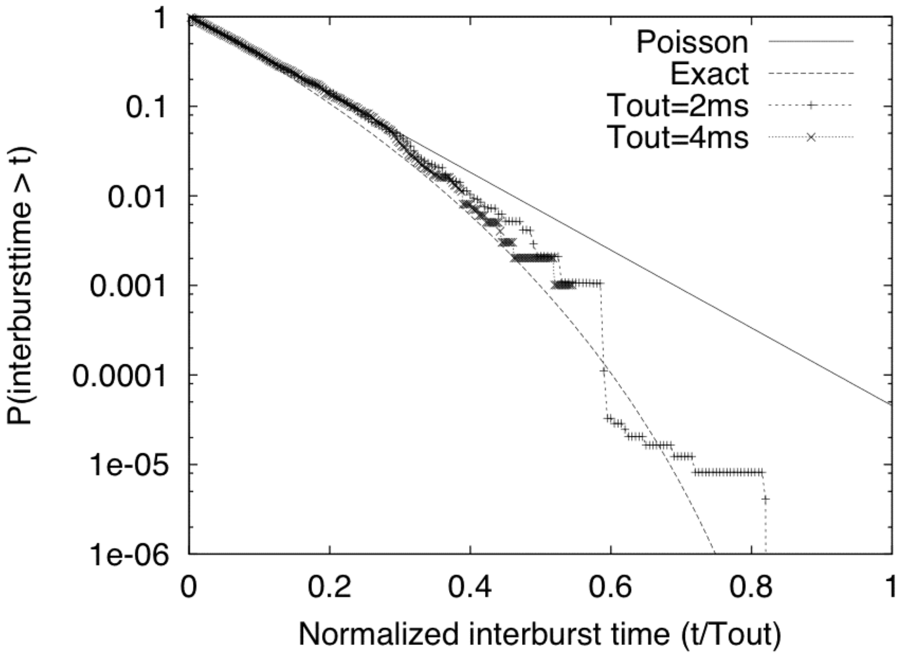
Results



- Burst size distribution is Gaussian due to the Gaussian nature of the marginal distribution at moderate timescales (analytical justification)

Burst interarrival time

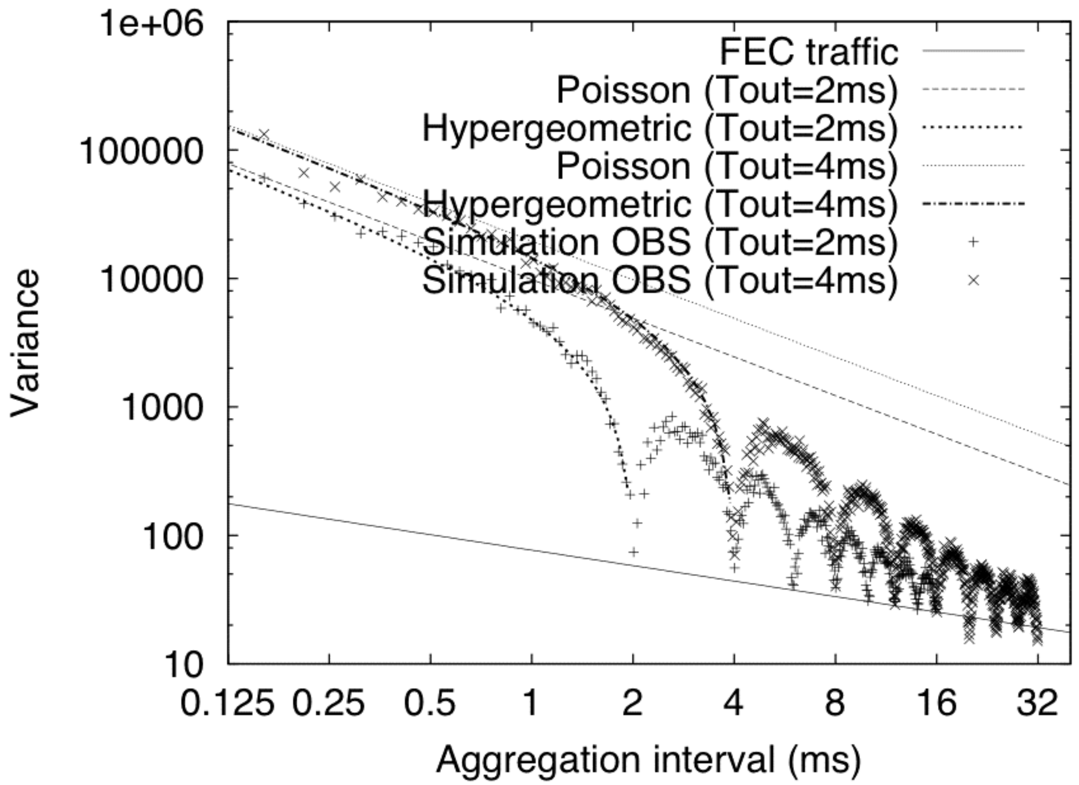
Results



- Burst arrivals per interval can be modeled by a hypergeometric variable.
- Exponential approximation is fine for small timescales.

Self-similar features

Results



Variance-aggregation plot shows:

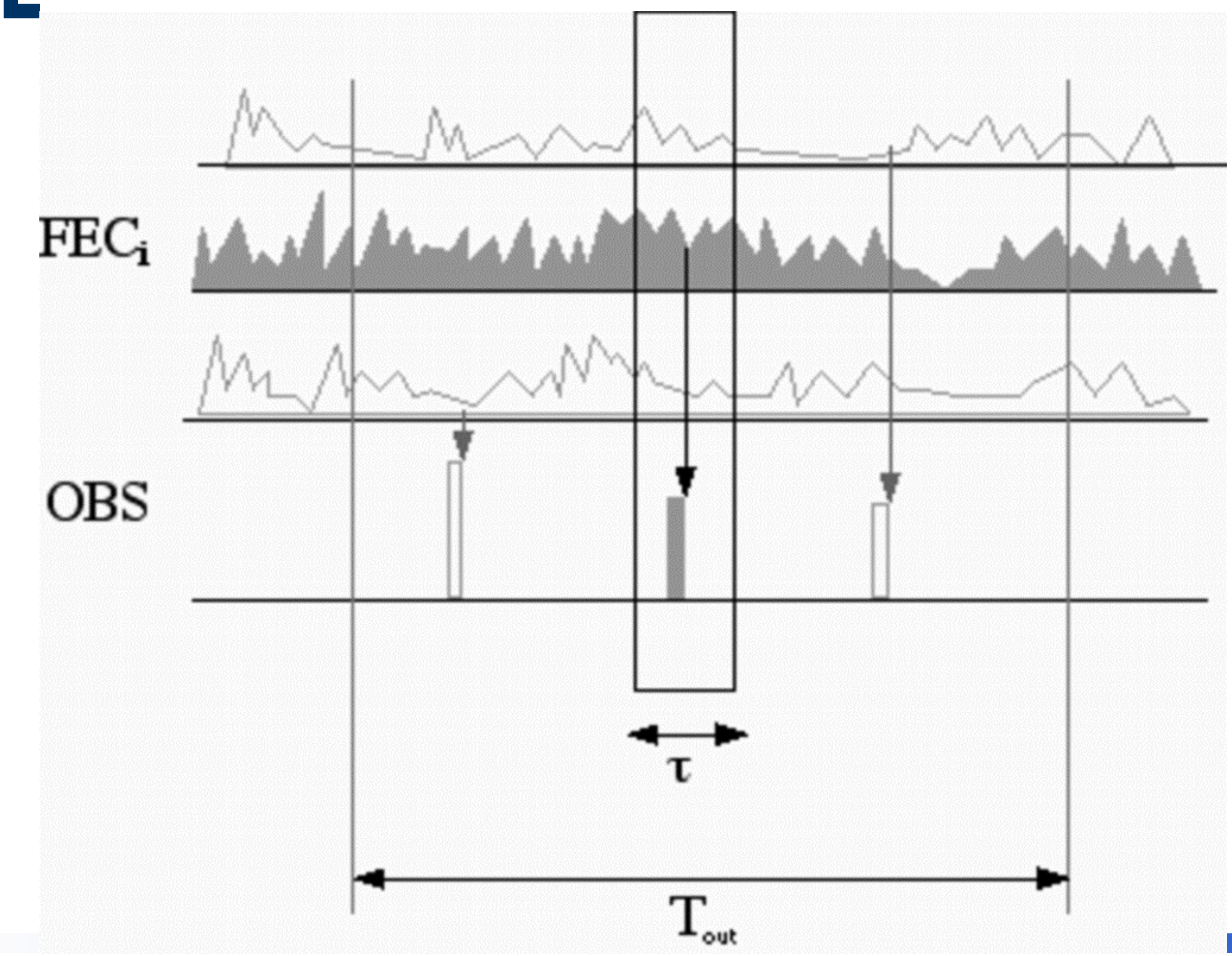
- Long-range dependence vanishes at low timescales but **not** at large timescales
- Variability increases at low timescales

Self-similar features (II)

Results

- Since burst from several independent sources are interleaved at the network edge:
  - There is a **shift of the scaling region to large timescales** since the burst assembly process makes traffic change at low timescales only
  - Variability increases at low timescales due to bursts

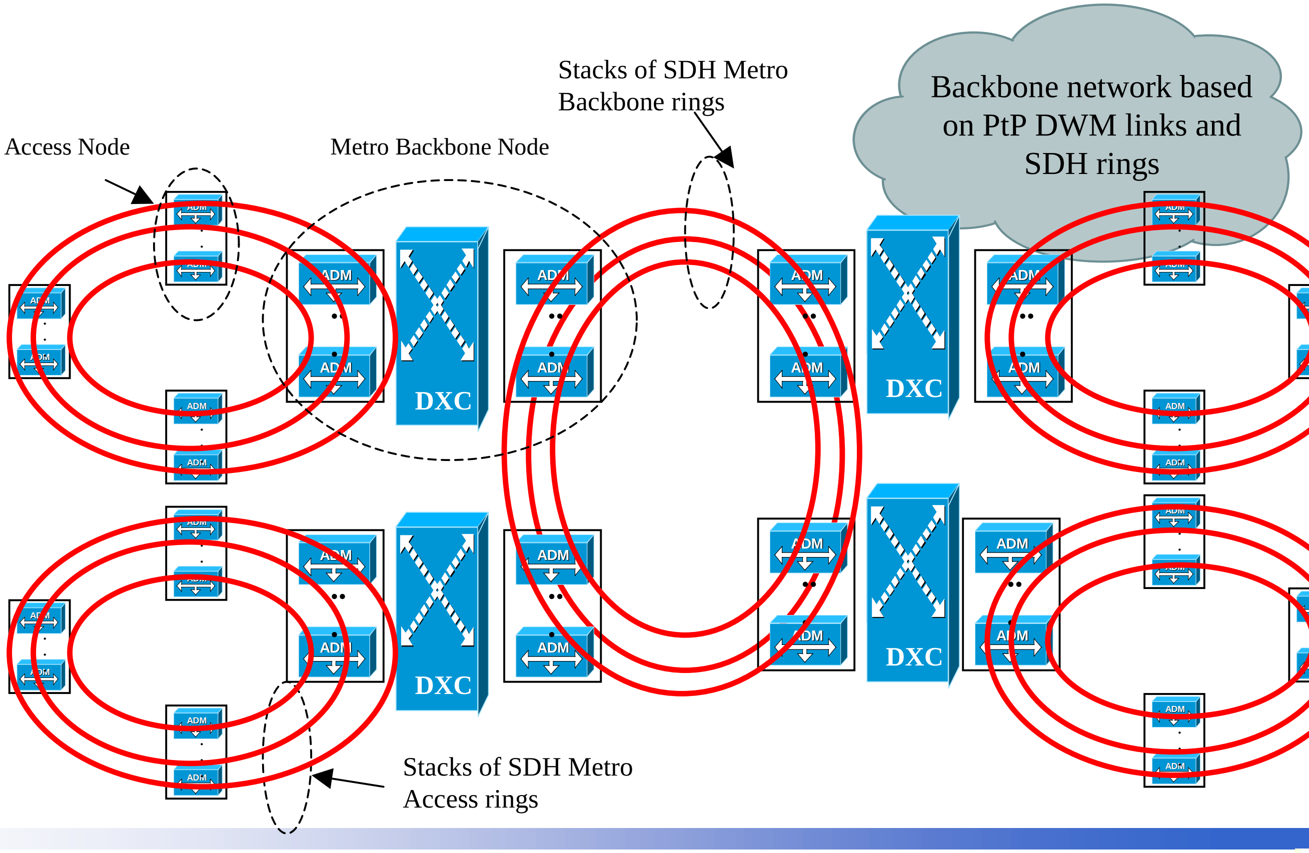
Self-similar features (III)



Network Topology:
Previous steps before OBS deployment

Scenario 0: Opaque network

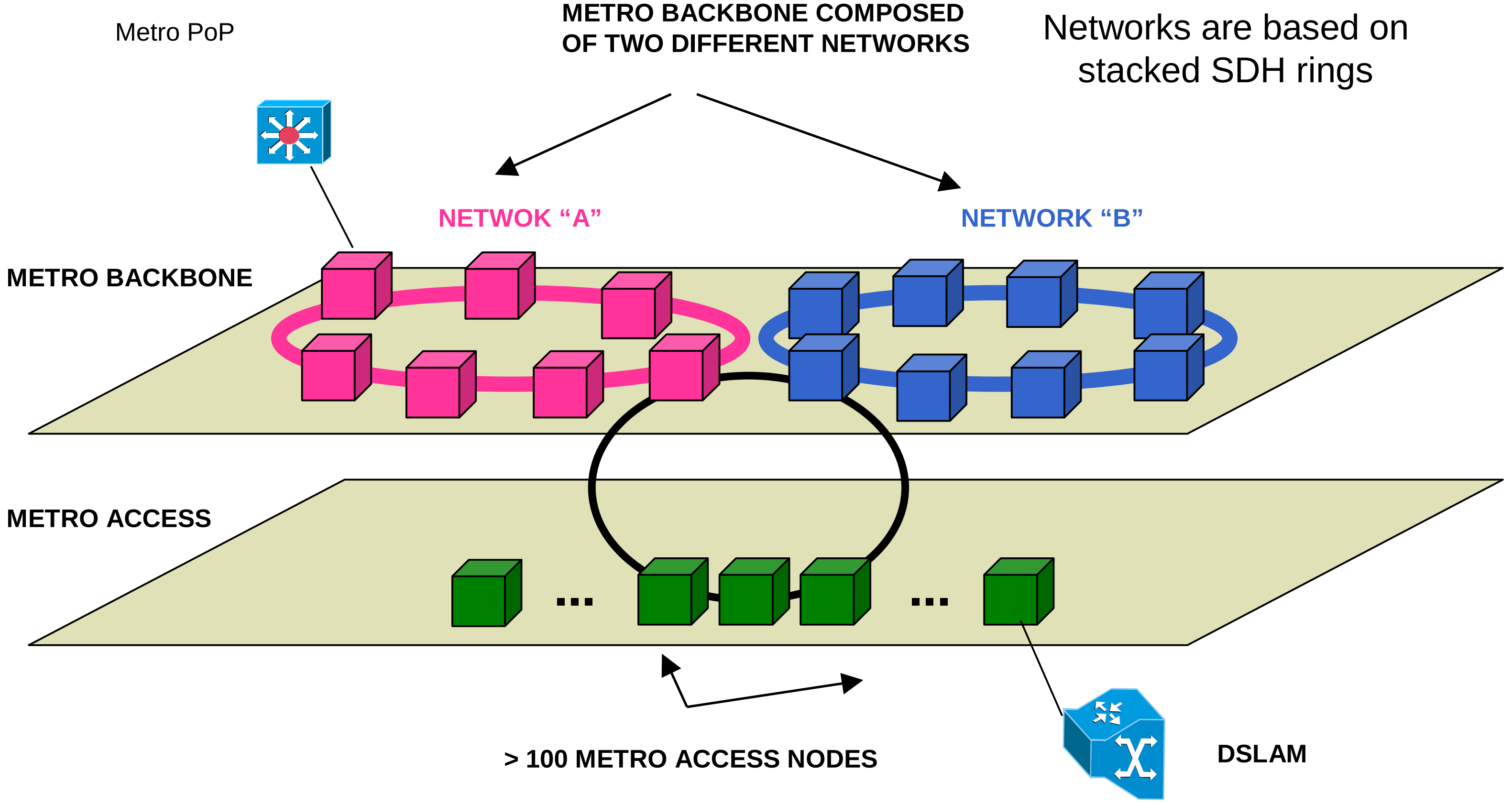
- Layer 1 networks are mainly based on SDH technology
- Ethernet is replacing ATM as main Layer 2 technology


Stacks of SDH Metro Backbone rings
Backbone network based on PtP DWM links and SDH rings
Access Node
Metro Backbone Node
ADM
DXC
Stacks of SDH Metro Access rings

Network Topology: Previous steps before OBS deployment

- **Scenario 0: Opaque network**
  - Dual Homing configuration

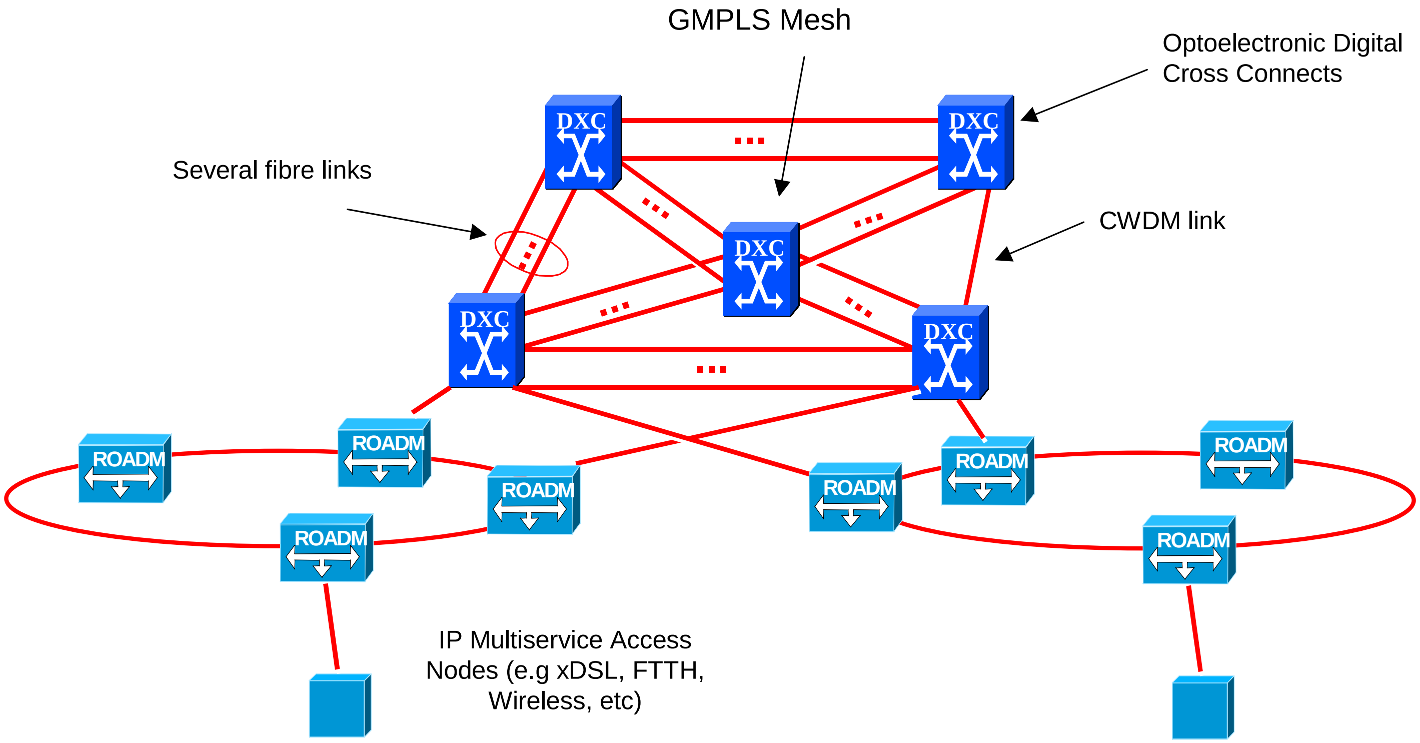
Current Metro Transmission Networks are based on stacked SDH rings



Network Topology:
Previous steps before OBS deployment

Scenario 1: Hybrid Network

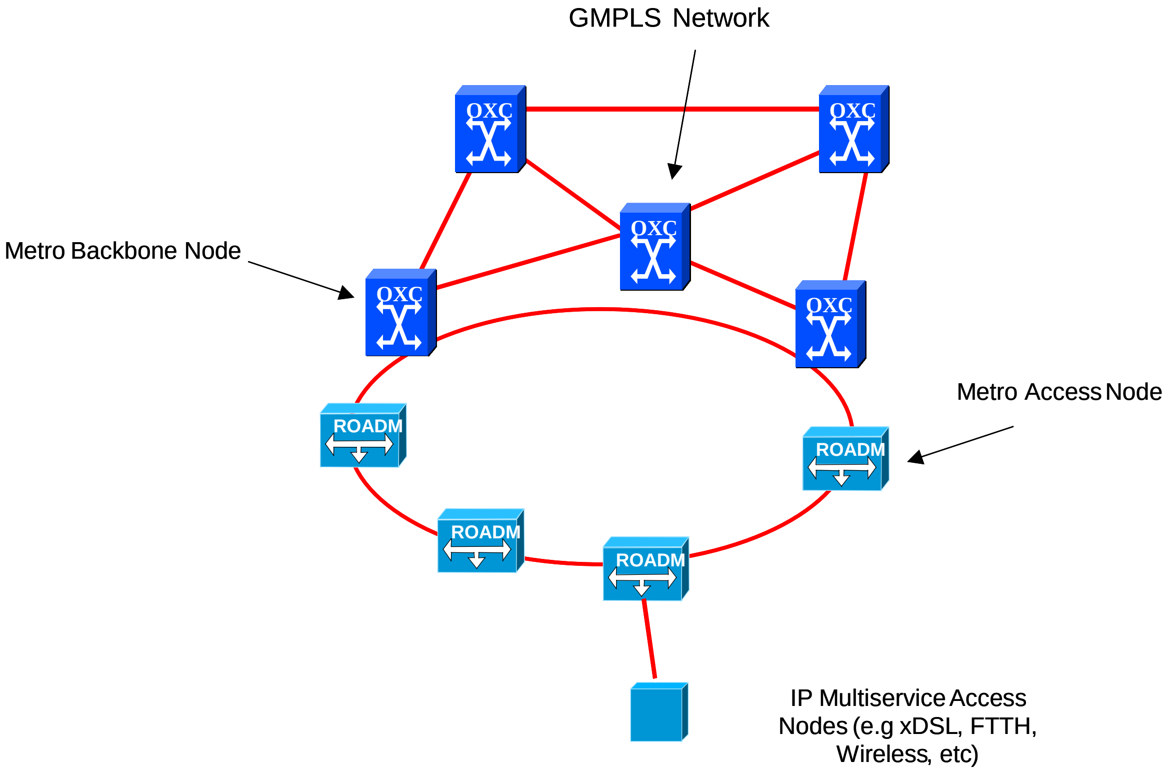
- Core and metro backbone networks are based on a GMPLS mesh composed of OEO DXCs
- Metro Access rings are based on DWDM rings composed of ROADMs



Network Topology:
Previous steps before OBS deployment

Scenario 2: WS Network

- Evolution towards an all optical Layer 1 network composed of OXC and ROADMs with GMPLS capabilities



Network Topology: First OBS deployments

- **OBS networks are expected to be deployed in long term scenarios with dramatically increased traffic demands and higher flexibility and granularity requirements**

- **A natural, simple and low cost evolution from WS to OBS scenarios may be achieved by gradually updating the ROADMs and OXCs previously used in the WS scenario in order to support optical burst transmission.**

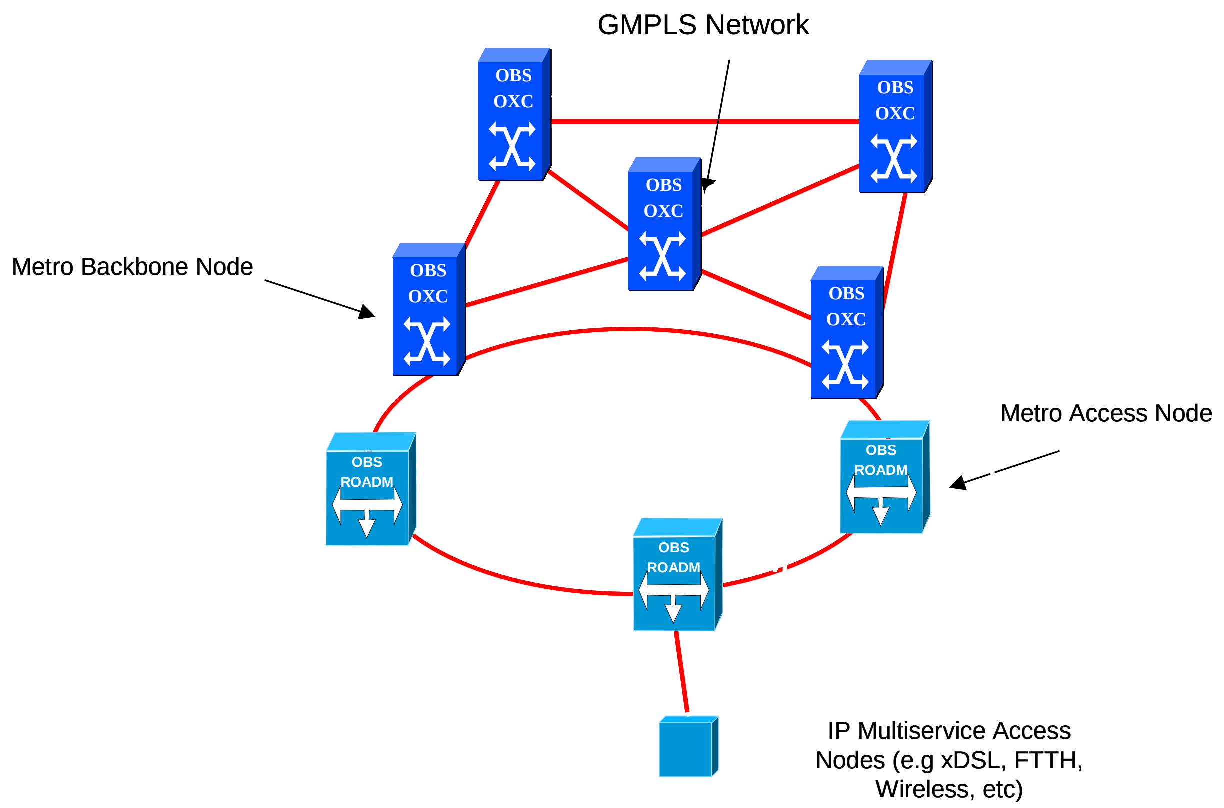
- **Therefore, in a first step, OBS networks may have similar topologies than WS (i.e metro access rings and core meshes).**

Network Topology:
Previous steps before OBS deployment

- **OBS network topologies will be strongly affected by the previous network evolution**
- **Currently, European transmission networks are mainly based on traditional SDH topologies (i.e SDH rings interconnected by DXCs).**
- **The appearance of GMPLS is favoring the migration from static SDH ring architectures with protection mechanisms towards more flexible SDH meshed backbone network architectures including GMPLS restoration**
- **In the short term, SDH technology is expected to be gradually migrated to Wavelength Switching (WS) due to the following drivers:**
  - **Technological availability (appearance of the first ROADMs and OXCs)**
  - **CAPEX and OPEX reduction, mainly due to automation and transparency, and increase of revenue coming from new services (Optical VPNs)**
- **A feasible trend could be the evolution towards metro aggregation rings based on ROADMs and connected through a core mesh composed by OXCs with full GMPLS support.**

Network Topology: First OBS deployments

- **Scenario 3: Optical Burst Switching (OBS) network**
  - –Optical equipment is updated in order to support optical burst transmission



Routing: OBS features

- Some OBS features need to be taken into account in the routing strategy:
  - Calculation of the optimal value of the offset time (time between the arrival of the control packet and the arrival of the burst)
  - Contention in nodes. Buffering is still very limited.
- Goals of routing in OBS
  - Reduce contention in nodes
  - Improve performance

Source routing

- Where is routing performed?
  - Source and hop-by-hop routing
- Source Routing
  - The routing decision is performed in the ingress router. The path is not changed in the intermediate nodes.
  - The control packet contains the information of all the hops of the path
  - The optimal value of the offset time can be calculated accurately, because the number of hops is known
  - In order to consider network state, flooding the network with congestion information is needed
  - Traffic engineering techniques can be used (GMPLS approach)

Hop-by-hop Routing

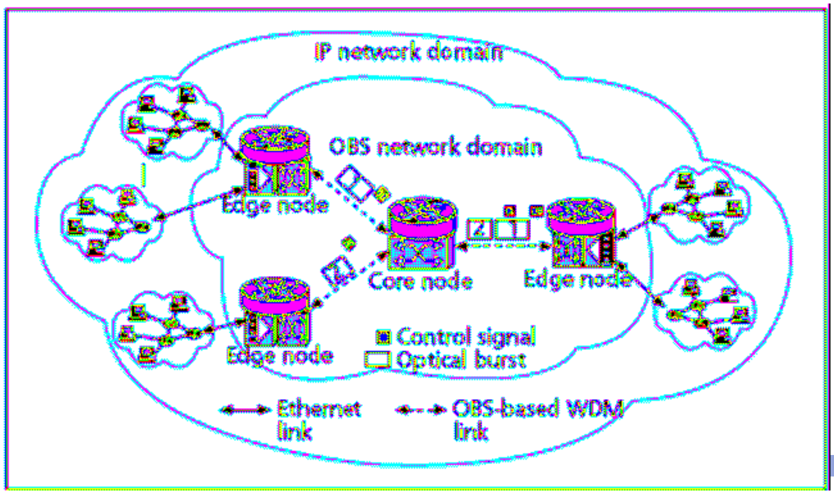
- Hop-by-hop routing
  - Routing decision in performed in every node
  - The whole path and the number of hops is unknown
  - The value of the offset time must be estimated (the number of hops is not known).
  - Possibility to use routing algorithms of IP networks
    - Need to adapt metrics to OBS
  - It is possible to use local congestion information (there is no need to flood the network)

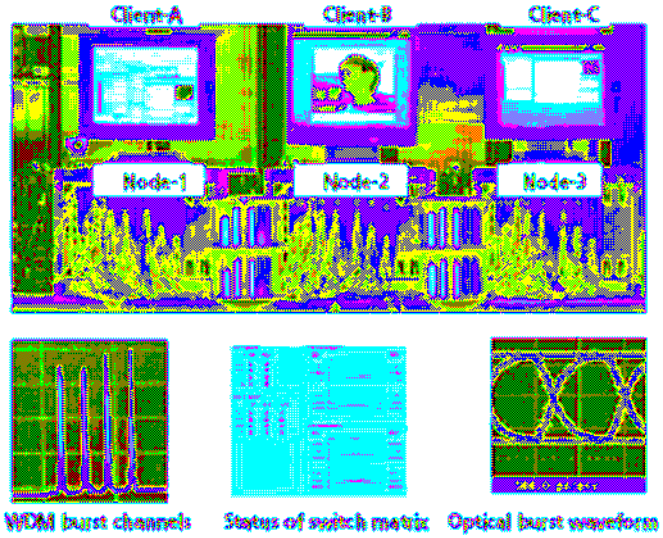
State of the art in OBS Testbeds

- **Research OBS testbeds**
  - – University of Tokio Testbed
  - – BUPT Testbed (China)
  - – JAPAN testbed

Testbed of OBS of the university of Tokio

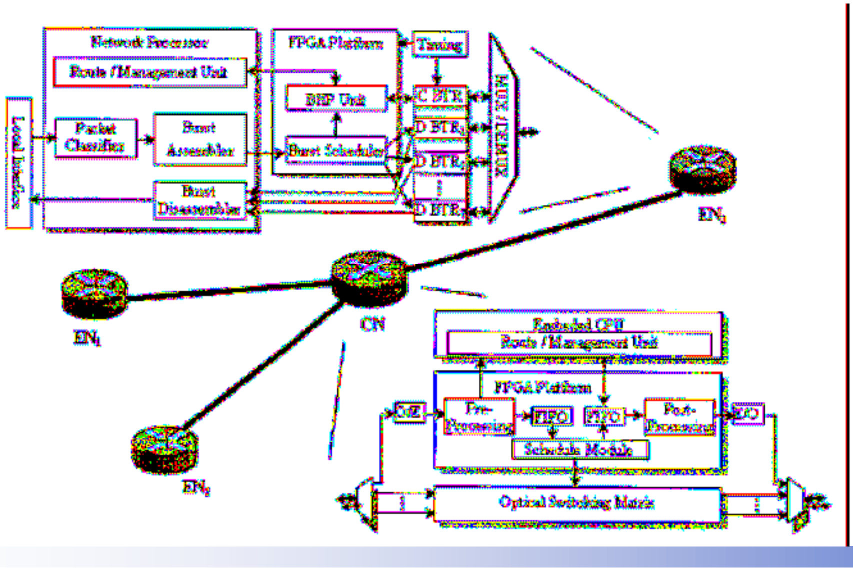
- **Y. Sun et al. “**Design And Implementation Of An Optical Burst-Switched Network Testbed**”, IEEE Communications Magazine, Nov 2005**
- **Switching technology:** PLC **(16x16) Switching time:** 3ms
- **Protocols: Signaling:** JET**, Scheduling:** PWA**, Contention resolution:** deflection routing, **Burst Assembly:** timer based **(15ms)**
- **Additional details: 3 edge nodes + 1 core node. Offset time: 13ms. Guard time 10ms (the 3ms of switching time are included there)**
- **Results: It was demonstrated the real time transmission of a video stream over the OBS testbed.**

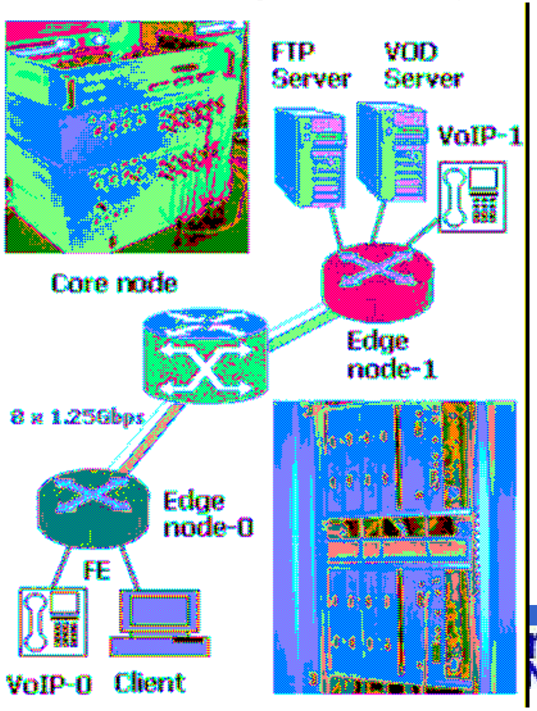




BUPT Testbed (China)

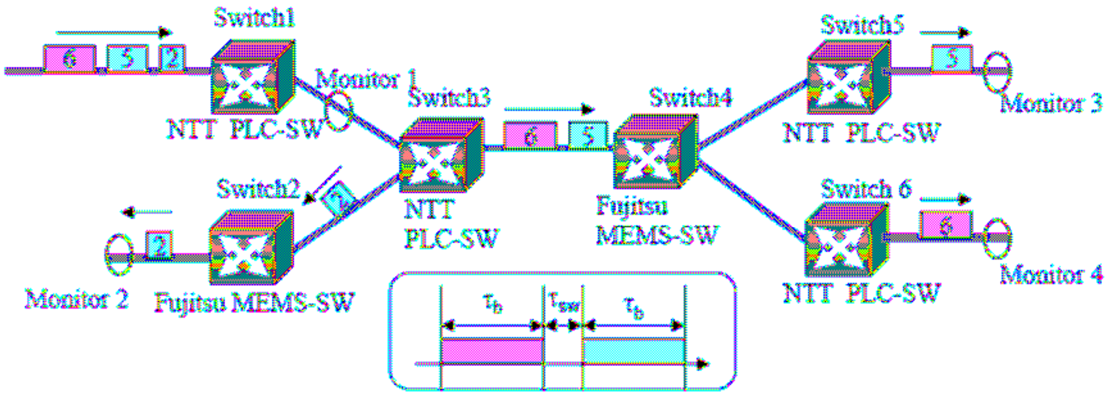
- **H. Guo et al** “A testbed for optical burst switching network”, OFC 2005
- **H. Guo et al. “**Multi-QoS Traffic Transmission Experiments on OBS Network Testbed**”, ECOC 2005**
- **Switching technology::** SOA**, 32x32, Switching time:** 100ns
- **Protocols: Signaling: (priority)** JET**, Scheduling:** LAUC-VF**, Burst Assembly: mixed** timer based **(1 ms)** and max length **(90Kb)**
- **Additional details: 3 edge nodes + 1 core node, 8 data channels + 1 control channel a 1.25 Gbps, maximum processing delay: 2.5us in edge nodes and 10us in core nodes. Offsets (50us, 900us, 1750us for each QoS)**
- **Results**: **It demonstrates the feasibility of OBS. Verifies QoS provisioning in an OBS network with pJET. Real TCP traffic was transmitted, and it was shown experimentally the delay penalty and correlation in the losses.**





Japan testbed

- **K. Kitayama et al. "**Optical Burst Switching Network Testbed in Japan**", OFC 2005**
- **M. Koga, "**Design and demonstration of connection guaranteed optical burst switching network**", APOC 2005**
- **Switching technology: 3-D** MEMS in 2 nodes **and** PLC in 4 nodes**, switching time: less than 30ms**
- **Protocols:** two-way GMPLS**-based. Burst sizes: 100, 200 and 300ms**
- **Additional details: 6 node network.**
- **Results**: **0.87 network throughput obtained. Two way signaling OBS was demonstrated.**


Switch1
6
5
2
NTT PLC-SW
Monitor 1
Switch3
NTT
PLC-SW
6
5
Switch4
Fujitsu
MEMS-SW
Switch5
5
Monitor 3
NTT PLC-SW
Switch2
1
2
Monitor 2
Fujitsu MEMS-SW
Switch 6
6
Monitor 4
NTT PLC-SW
τ_b
τ_{sw}
τ_b