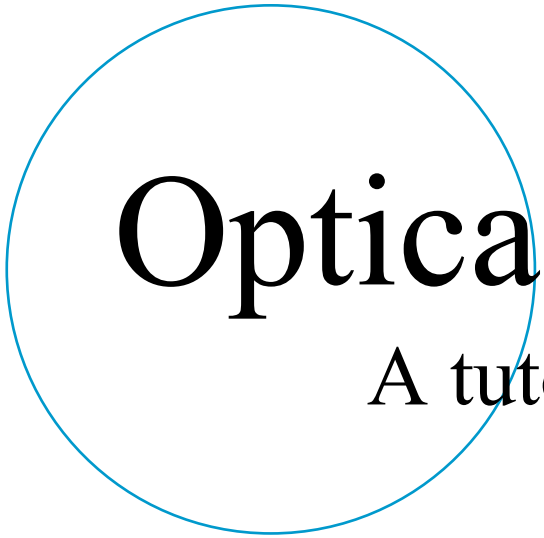


Optical Burst Switching

A tutorial from E-photon/ONe

The VD1 OBS taskforce

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, Nail Akar, Ezhan Karasan

TID: Juan Fernández-Palacios, Óscar González

Sections

- Introduction to OBS, including QoS issues – *UPC*
- OBS switching and signaling – *RACTI-Upatras*
- Network topologies and routing - *TID*
- Contention resolution schemes – *UST-IKR*
- Traffic models - *UPNA*

Sections

- TCP over OBS – *DEIS-Unibo*
- Performance evaluation of OBS networks – *Bilkent Univ.*
- OBS in the real world: testbeds – *Univ. of Essex*
- Physical components for OBS networks – *FUB*
- Dependability – *Telenor/NTNU*

Introduction to OBS

Universitat Politécnica de Catalunya
(UPC)

Motivations

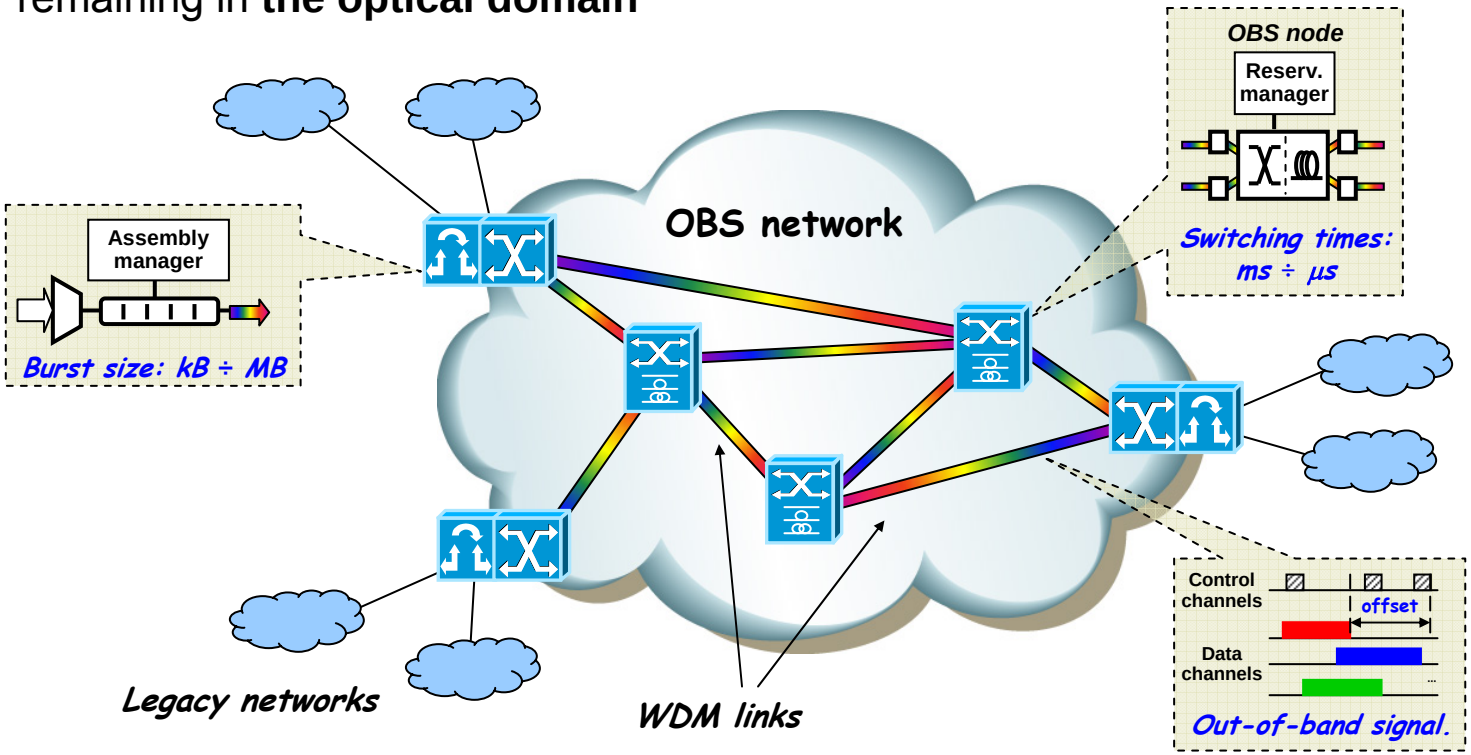
- Changes in traffic profile
  - P2P file downloading vs. multimedia streaming
  - grid networking
- Wavelength-Switched Networks
  - low network utilization and flexibility
- Problems in Optical Packet-Switched Networks
  - lack of optical buffering
  - need for fast packet switching and header processing

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 in order to temporary 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 view as lying between OPS (one-way reservation) and WS networks (two-way reservation)

OBS network

- 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

Edge Node

- Consist 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 segmentate data bursts for the purpose of their partially dropping in core nodes when contention occures
  - Setting up the pre-transmission offset time
    - in simple fixed offset scheme, the offset time is calculated as a sum of the total processing time at all the intermediate hops
    - offset time is one of the crucial OBS network parameter, since his incorrect estimation has impact on data lost
  - Sending the control packet
  - Sending the burst

Core Node

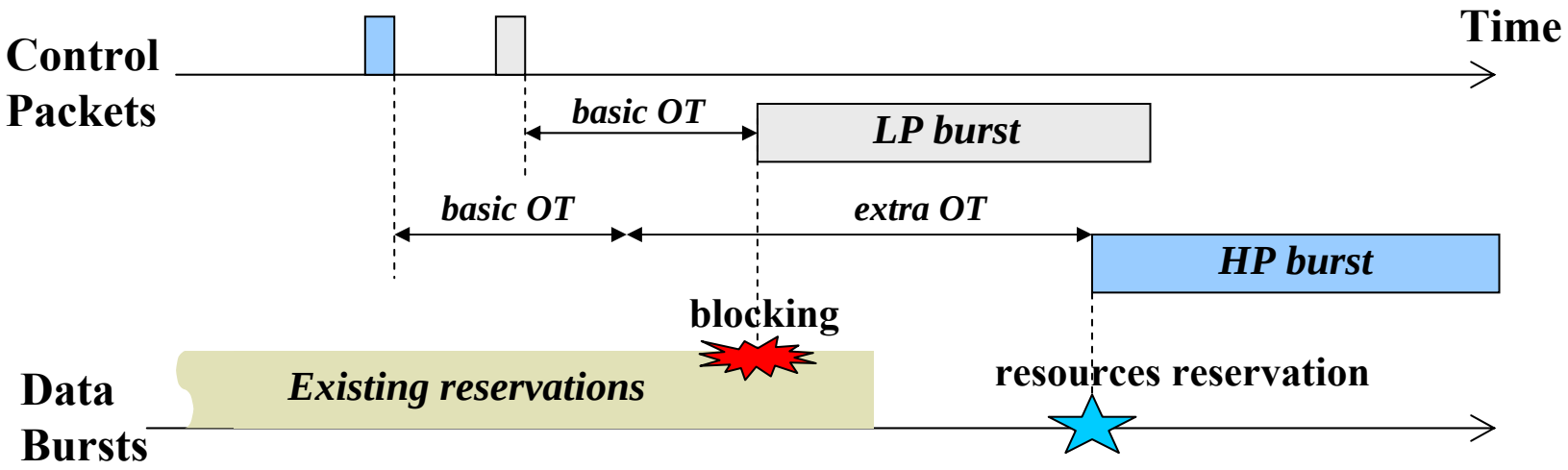
- Hardware requirements
  - O/E/O conversion for header processing
  - λ-conversion
  - switching speeds fast enough
  - eventually optical buffering (FDLs)
- Operation
  - Processing of incoming control packets (electronically) and sending it to the next node that lays on the routing path
  - Reservation 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)

Quality of Service

QoS components

- Signalling
  - *absolute* QoS by wavelength reservation in 2WR-OBS
- QoS routing
- Techniques for QoS management in nodes
  - Edge nodes
    - Offset-time differentiation
    - Varying burst assembly parameters
    - Class based queuing
  - Core nodes
    - Dropping techniques in contention resolution
    - Priority scheduling of control packets
- Admission Control
  - both in Edge and Core nodes
- Classes of Services

Offset-Time Differentiation



- An additional offset-time (OT) is assigned to high priority bursts, what results in an earlier reservation, in order to favors them while the resources reservation
- Extra QoS offset have to be in the order of a few burst durations of lower priority bursts
- This technique keeps constant the total blocking probability while reduces the loss probability of high priority bursts
- The main disadvantages are both the sensitivity of high priority class to the burst length characteristics and extended pre-transmission delay

Dropping techniques

- Threshold based – reservation of a different number of resources for High and Low priority traffic
  - with λ-threshold
  - with FDL threshold
- Intentional dropping – initiated according to either loss rate measure or traffic profile to guarantee burst loses on defined levels
- Scheduling with burst preemption
  - of all bursts
  - partially with the segmentation

Classes of Services (CoS)

- **Emergency** class, e.g. for real-time applications in emergency cases
  - has absolute priority over all other classes
  - in case of lack of resources expropriates them from other connections
- **Low delay and low loss** class for carrying loss and delay sensitive applications (like e.g. Voice over IP)
- **High priority** class, prioritized non-realtime traffic transmission for special customers and services (e.g. bank transactions)
- **Best effort** class for non-real data transmission
- **Data Container** class for transmissions of huge data volumes
  - e.g. in grid, banking data backups, data file transfer applications
  - aims in decreasing the transmission control overhead

Implementation of CoS

- Each class requires an individual treatment thus the application of appropriate QoS management rules
- **The emergency class** could be provided on dedicated wavelength by two-way reservation signaling protocol in order to give it absolute communication quality.
- **Delay sensitive applications** should carefully explore offset time differentiation as well as the deflection routing because can add undesirable delay. On the other hand, aggregation schemes with a timer which bound an assembly delay are recommended.
- **The loss sensitive** as well as **prioritized classes** should be implemented according to the services differentiation approach.
- Access to the resources for **data container class** that carries very long bursts may be ensured either by assigning a very high extra-offset time or with two-way reservation protocol.

  However high transmission delays caused by signalization procedures in both cases put additional requirements for buffering memories applied in edge nodes in order to store the data coming from legacy networks.

Comparison of QoS mechanisms

Technique	Scenario	α	BLP_{HP}	BLP_{Total}	BLP_{BE}	
Offset-Time Differentiation	No FDLs	**2.5%**	$<10^{-6}$	1.95e-001	2.00e-001	⇐ **No FDLs**
	With FDLs	**42%**	$<10^{-6}$	6.76e-002	1.16e-001	
	With FDLs & BLD (MBL_{HP} = 10kB)	**59%**	$<10^{-6}$	5.18e-002	3.50e-001	
	With FDLs & BLD (MBL_{HP} = 5kB)	**65%**	$<10^{-6}$	3.33e-002	4.33e-001	⇐ **FDLs & BLD**
Burst Preemption	No FDLs	**2%**	$<10^{-6}$	1.94e-001	1.98e-001	⇐ **No FDLs**
	With FDLs	**14%**	$<10^{-6}$	1.79e-002	2.08e-002	
	With FDLs & BLD (MBL_{HP} = 10kB)	**49%**	$<10^{-6}$	1.39e-002	6.71e-002	
	With FDLs & BLD (MBL_{HP} = 5kB)	**55%**	$<10^{-6}$	7.86e-003	8,48e-002	⇐ **FDLs & BLD**
λ threshold	No FDLs, BE with access to ½ of λs	**0.3%**	6.7e-006	5.27e-001	5.27e-001	
	With FDLs & BE with access to ¾ of λs	**2%**	$<10^{-6}$	1.98e-001	2.02e-001	
	With FDLs & BE with access to ½ of λs	**20%**	$<10^{-6}$	2.87e-001	3.59e-001	
	With FDLs & BE with access to ½ of λs & BLD (MBL_{HP} = 10kB)	**37%**	$<10^{-6}$	8.91e-002	2.98e-001	
	With FDLs & BE with access to ½ of λs & BLD (MBL_{HP} = 5kB)	**38%**	$<10^{-6}$	5.09e-002	3.00e-001	
FDL threshold	BE with access to ½ of FDLs	**22%**	$<10^{-6}$	4.27e-002	5.47e-002	
	BE with access to ½ of FDLs & BLD (MBL_{HP} = 10kB)	**38%**	$<10^{-6}$	2.58e-002	8.90e-002	
	BE with access to ½ of FDLs & BLD (MBL_{HP} = 5kB)	**40%**	$<10^{-6}$	1.59e-002	1.00e-001	
	No FDLs for BE	**57%**	$<10^{-6}$	1.31e-001	3.06e-001	

Scenario

Traffic
- load = 0.8
- Gaussian burst length distribution
- Gamma IAT distribution
- Two CoS: **High Priotity (HP)** & **Best Effort (BE)**
- MBL_{BE} = 40kB
- BitRate = 10Gbits

Node
- 4 input/output ports
- 4 wavelengths
- Buffer:
  - 4 FDLs
  - Granularity = MBL_{BE}
- LAUC-VF_MIN-SV scheduling algorithm

FDL - Fiber Delay Lines

BLD - Burst Length Differentiation
(HP bursts are aggregated with lower timer and maximum burst length thresholds than LP bursts)

α - HP traffic load

BLP - Burst Loss Probability

MBL – Mean Burst Length

QoS mechanisms - conclusions

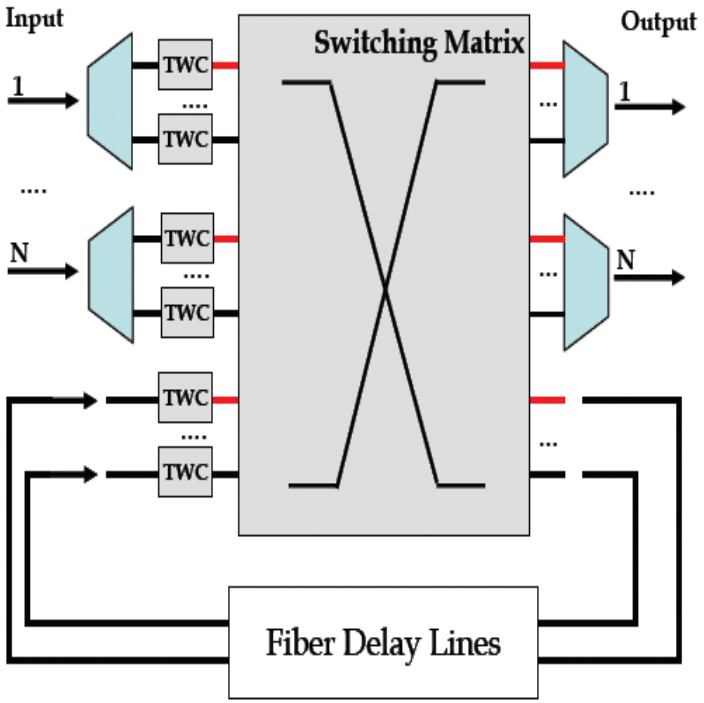
- Offset Time Differentiation (OTD) and Burst Preemption (BP) with FDL buffering and Burst Length Differentiation (BLD) applied achieve high HP traffic load ratios (α)
  - for OTD it's at the cost of high BLP of Best Effort (BE) traffic
  - BP performes better then OTD by providing higher overall BLPs
- Both OTD, BP and λ-threshold schemes achieve much better results when FDL buffering is applied
  - α doesn't exceeds 2-3% in every buffer-less scenario
- BLD in each case improves the results
  - the performance is better with shorter HP bursts

OBS switching and signaling

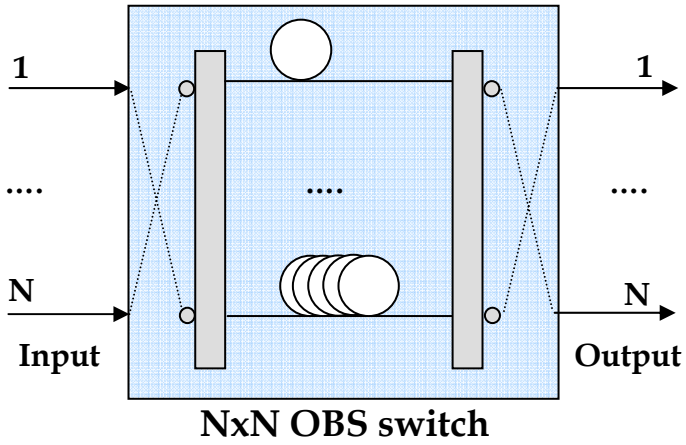
RACTI-University of Patras

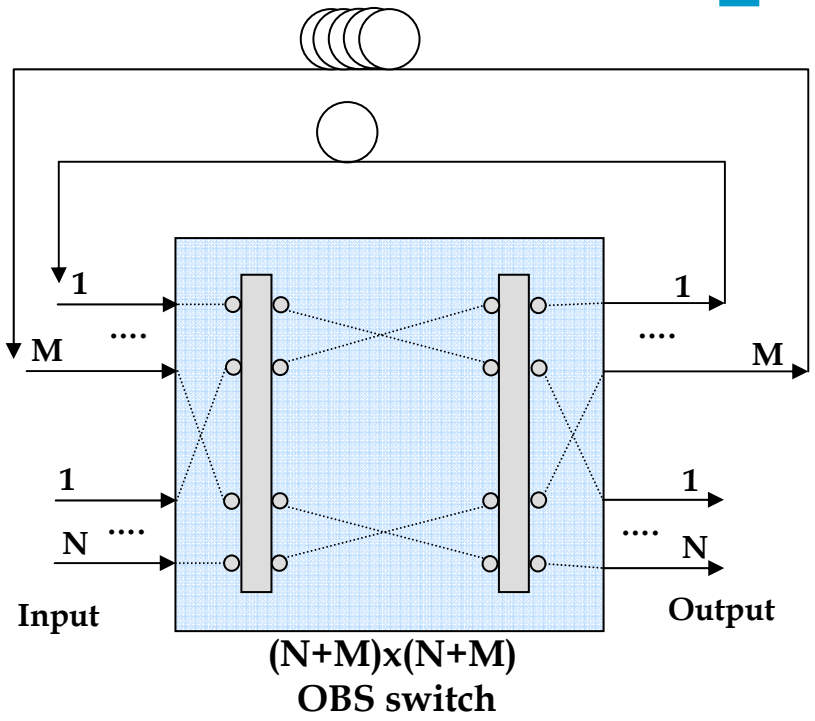
Burst Switch Architectures

- 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



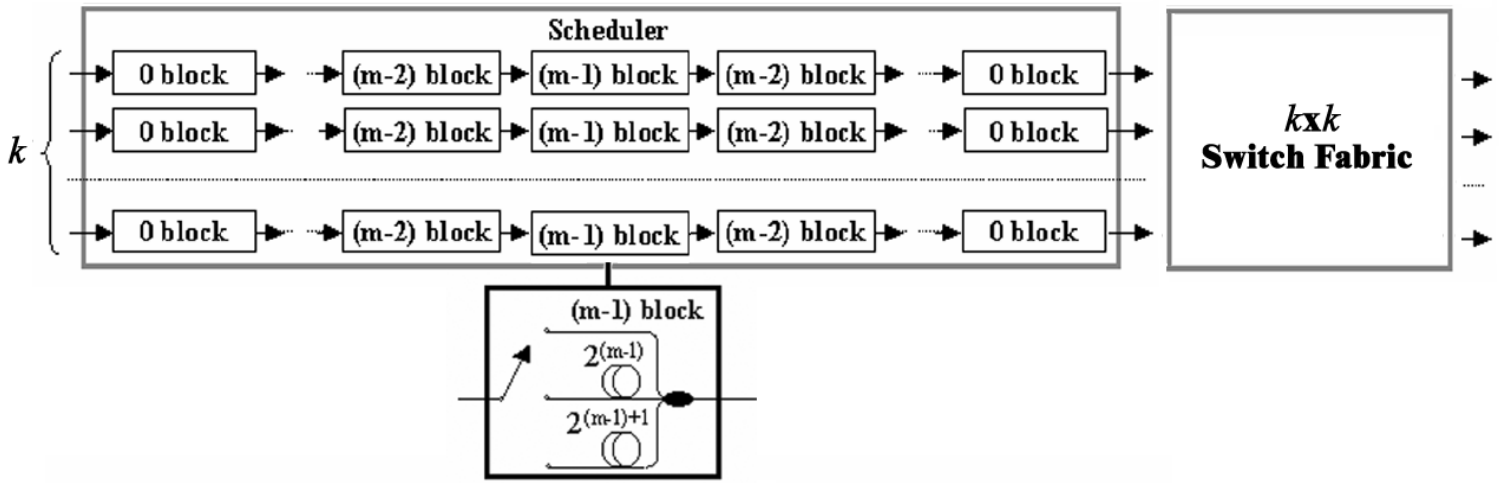
FDL burst switch architectures





- In the feed-forward method, bursts are fed into fiber delay lines of different lengths and when it comes out, it has 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.

Performance evaluation of FDL architectures



Concept: Use a branch of delays to scheduler packets in a T frame and resolve contention.

Each delay branch consist of $2m$-1 delay blocks, where $m = logT$.

T is assumed to be a power of 2 and corresponds to the maximum number of sequential packets from all incoming links that request the same output and can be served with no contention.

The ith block consists of a three-state (or two 2x2) optical switch and three fiber delay paths, corresponding to delays equal to 0, 2^i and 2^{i+1} slots.

Scheduler Performance – Pareto Analysis I

Pareto Traffic Model

Formula to generate a Pareto distribution:

$$X_{PARETO} = \frac{b}{x^{1/a}}$$

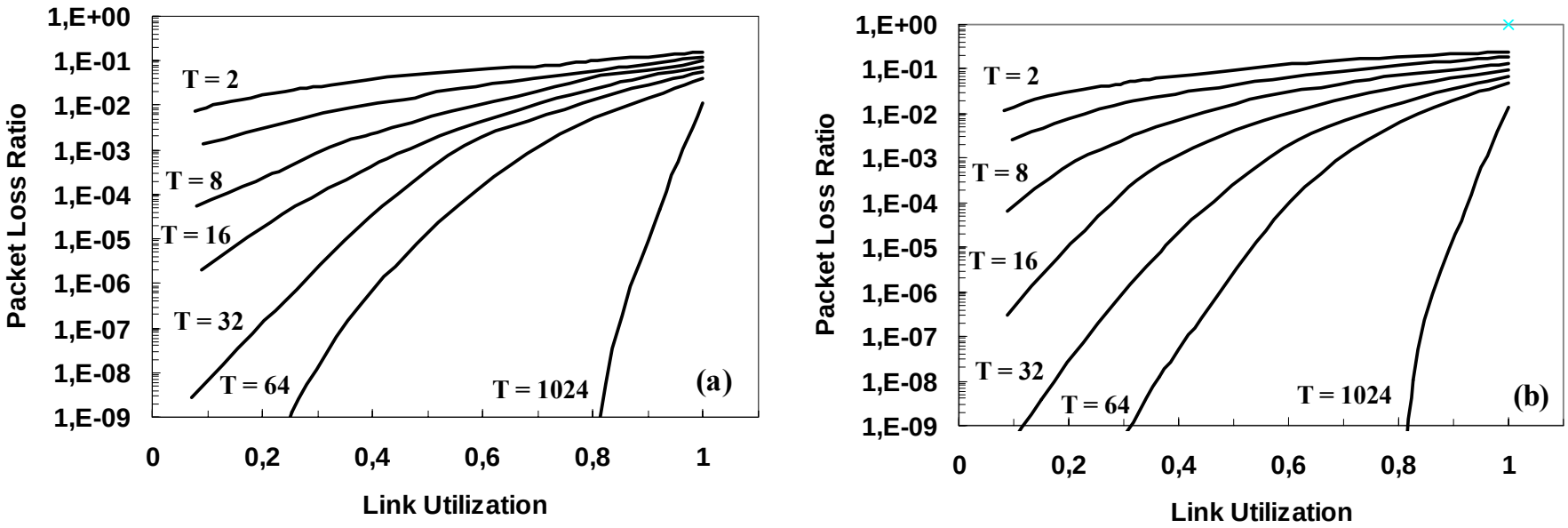
Assuming:

$$p = \frac{\overline{ON}_{period}}{\overline{ON}_{period} + \overline{OFF}_{period}}$$

We calculate the min OFF period

$$b_{off} = \frac{\frac{a_{off} - 1}{a_{off}}}{\frac{a_{on} - 1}{a_{on}}} \cdot \frac{1 - x_{\min}^{\frac{a_{on} - 1}{a_{on}}}}{1 - x_{\min}^{\frac{a_{off} - 1}{a_{off}}}} \cdot \left(\frac{1}{p} - 1 \right)$$

Scheduler Performance – Pareto Analysis II



Packet loss ratio for (a) k=2 and (b) k=4 versus link utilization for T ϵ [2…64] and T = 1024. Packet arrivals follow a Pareto distribution of ON periods with a tail index of 1.7, while OFF periods obey a truncated Pareto distributed as well, with a tail index of 1.2.

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 travel 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 bursts is transmitted after an end-to-end connections is established**
  - SETUP is send 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 determined by the start/release policy:**
  - Start of Reservation:
    - Immediate-explicit: reservation starts immediately after the reception of the SETUP.
    - Delayed-implicit: reservation start by the beginning of the data
  - Release mechanism (tearing down)
    - Implicit: based on burst length information.
    - Explicit: use a release control packet.

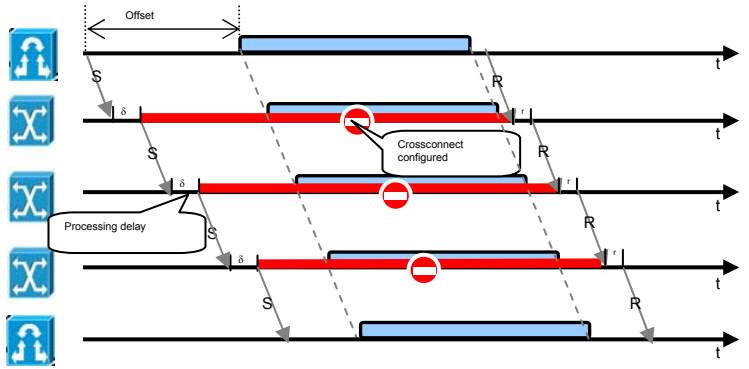
State-of-art in OBS signaling

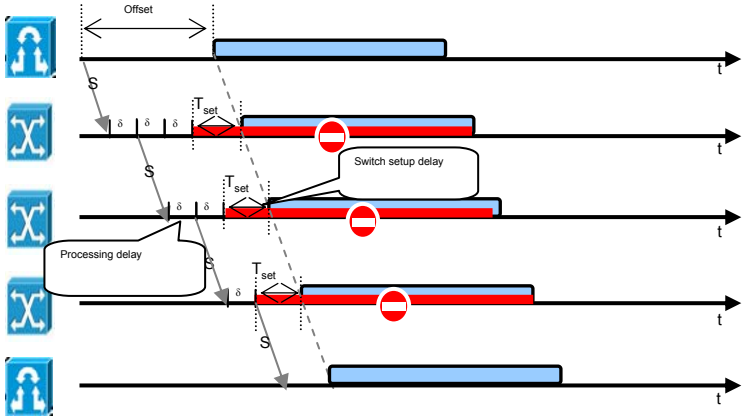
- JIT protocol :

  explicit setup and explicit or implicit release

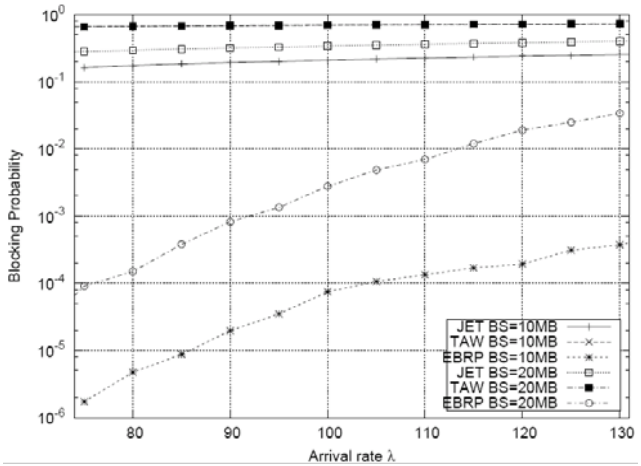
- Horizon and JET protocols

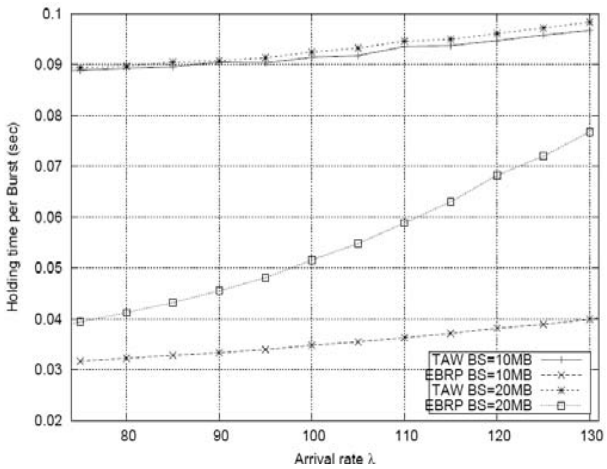
  employ estimated setup and estimated release
  - Horizon doesn't support void filling
  - JET supports void filling





Signaling Protocols Evaluation

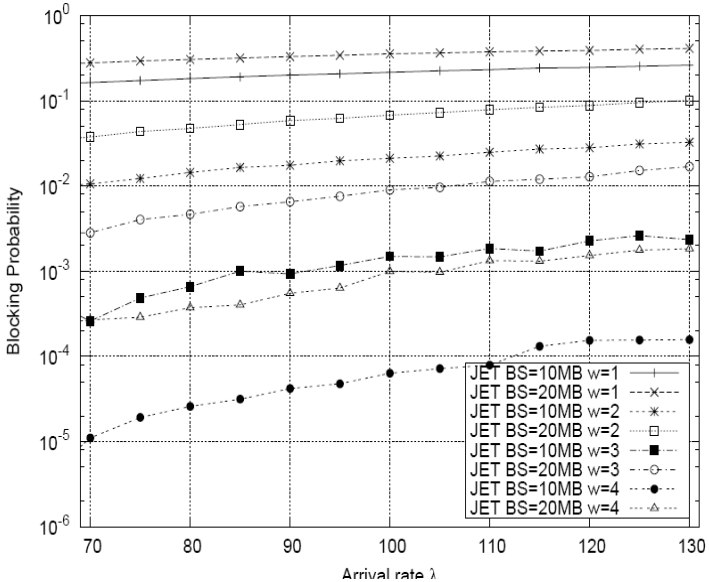




- Bufferless Network - Single channel OTDM
- JET, TAW, EBRP performance
  - NSFNET topology
  - Poisson burst arrivals
  - Burst Size exponentially distributed
- JET provides minimum holding time – high blocking probability
- EBRP reveals superior blocking performance for bursts that can tolerate the RTT

JET – full wavelength conversion

- Identical traffic characteristics
- JET performance improves as the number of supported channels increases
- JET performance is equivalent to EBRP for w=4 wavelengths per fiber



Network topology and routing

Telefónica Investigación y Desarrollo

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)

Classification of routing algorithms

- Static algorithms
  - Routing information does not change
- Dynamic algorithms
  - Routing information changes in time
  - Can be adaptative to the traffic
- Single-path
  - Only one path can be used at a time
- Multipath
  - Several paths can be used. The traffic is distributed among the paths.

Routing strategies

- Routing strategies
  - Shortest path source routing approach
    - Easy to implement
    - Links with the best metrics can concentrate all the traffic
    - Network state
  - Routing path optimization
    - Optimal paths are calculated for a given traffic matrix
    - Minimizes overall burst drop probability (only for the given traffic pattern)
  - Least-congested dynamic route calculation
    - Source routing strategy
    - Paths are calculted dynamically according to congestion information.
    - Flooding of congestion information to the edge nodes

Routing strategies

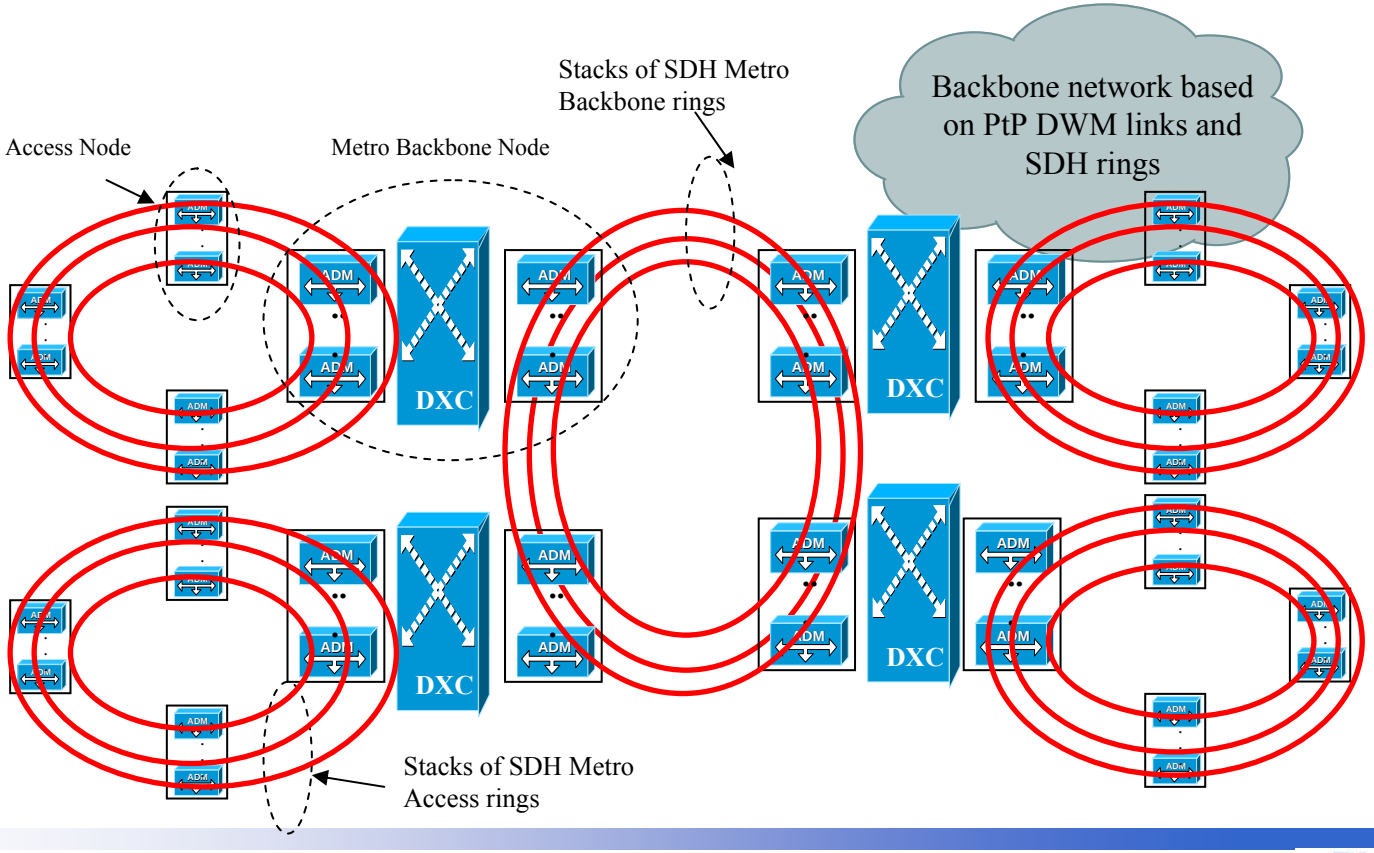
- Congestion-based static route calculation technique
  - Source routing strategy
  - Several link-disjoint paths are pre-calculated
  - The path is chosen based on congestion information
  - Flooding of congestion information to the edge nodes
- Multipath routing with dynamic variance
  - Hop by hop routing
  - Next hop is chosen dynamically in each node
  - Uses local congestion information
  - Traffic is distributed among several paths
  - Hysteresis mechanism to avoid instability
  - The load is balanced, reducing blocking probability

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.**

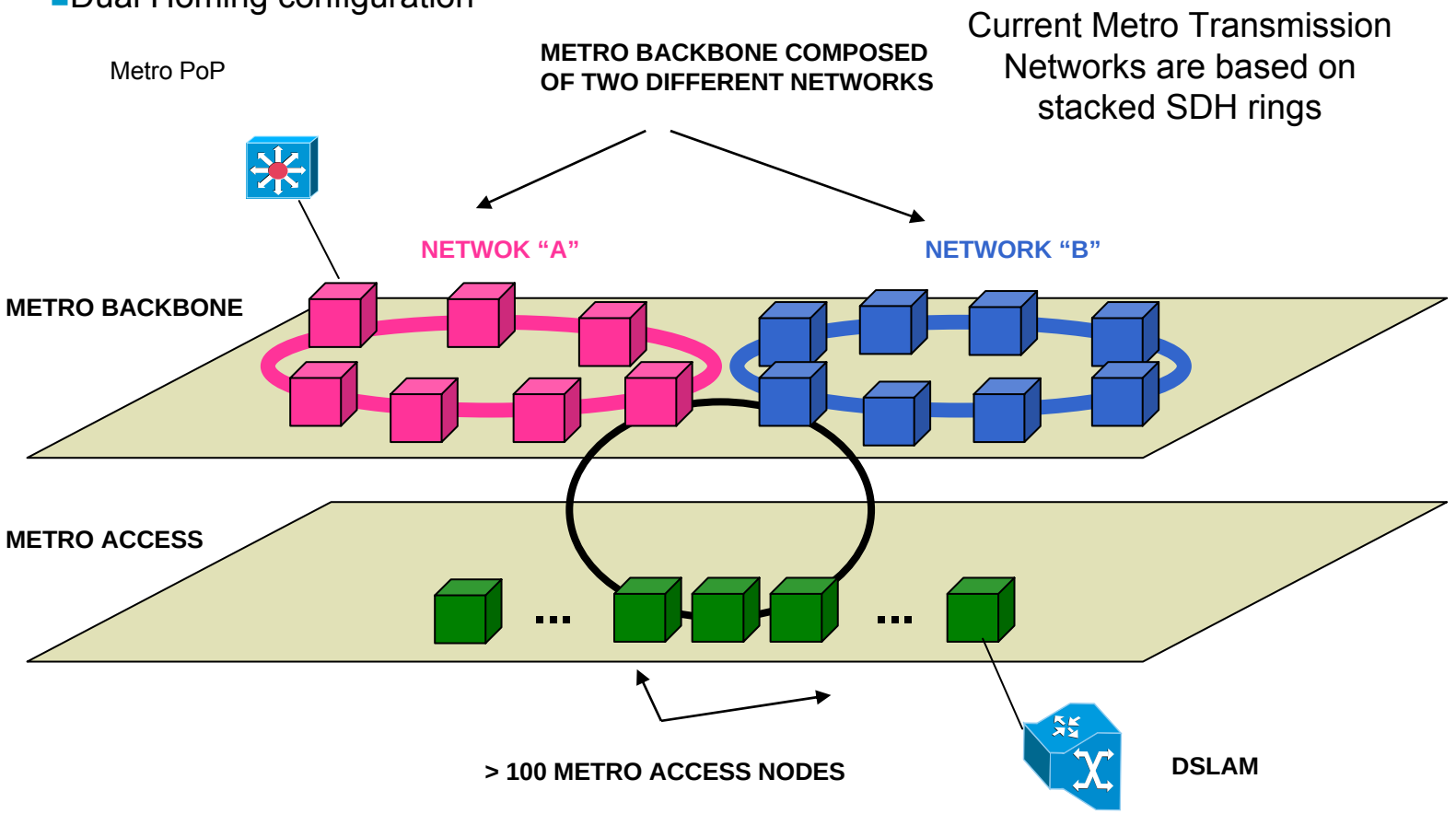
Network Topology: Previous steps before OBS deployment

- **Scenario 0: Opaque network**
  - **Layer 1 networks are mainly based on SDH technology**
  - **Ethernet is replacing ATM as main Layer 2 technology**



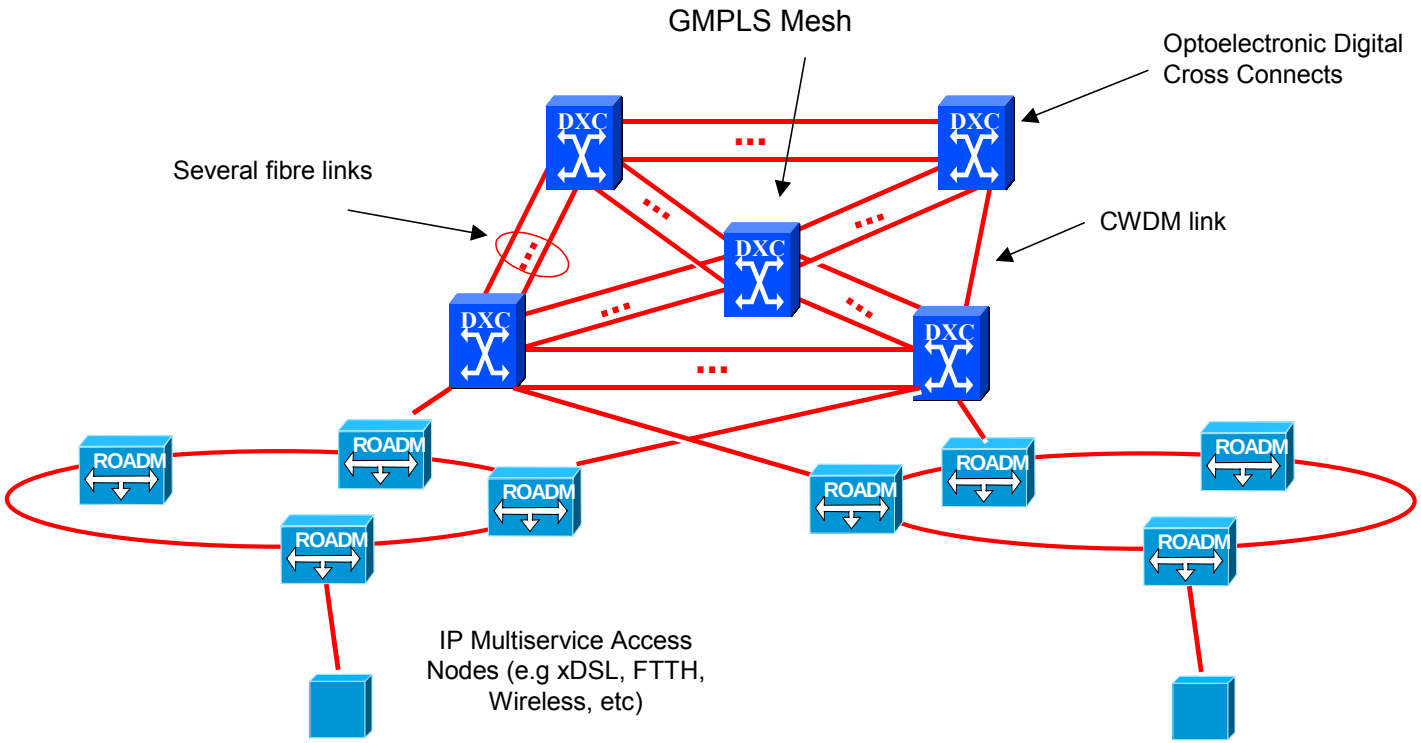
Network Topology: Previous steps before OBS deployment

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



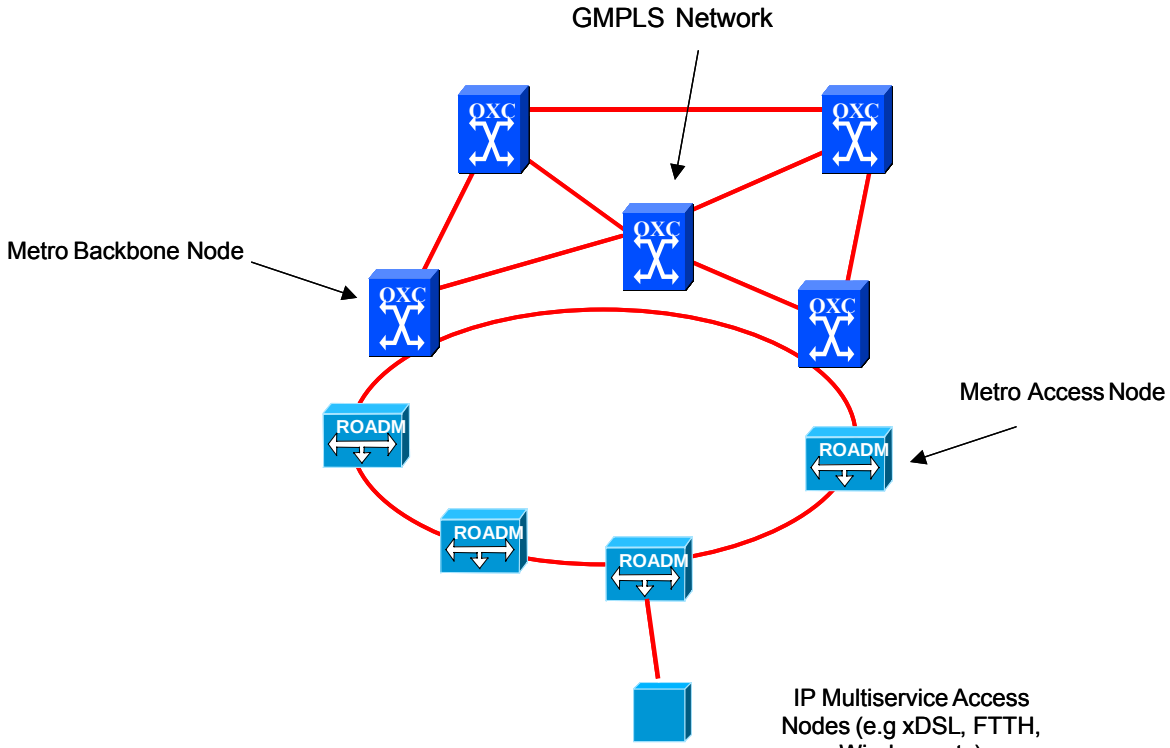
Network Topology: Previous steps before OBS deployment

- **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

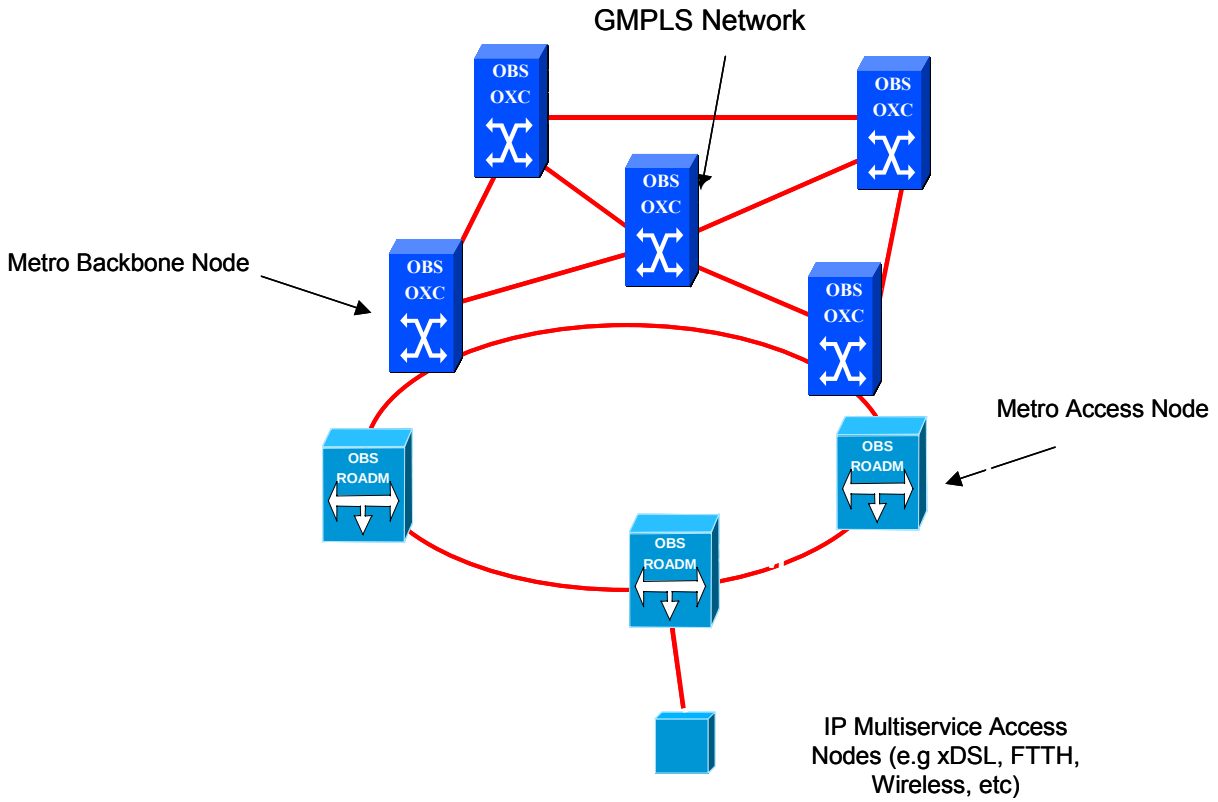
- **OBS networks are expected to be deployed in long term scenarios with dramatically increased traffic demands and higher flexibility and granularity requirements**

- **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: First OBS deployments

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



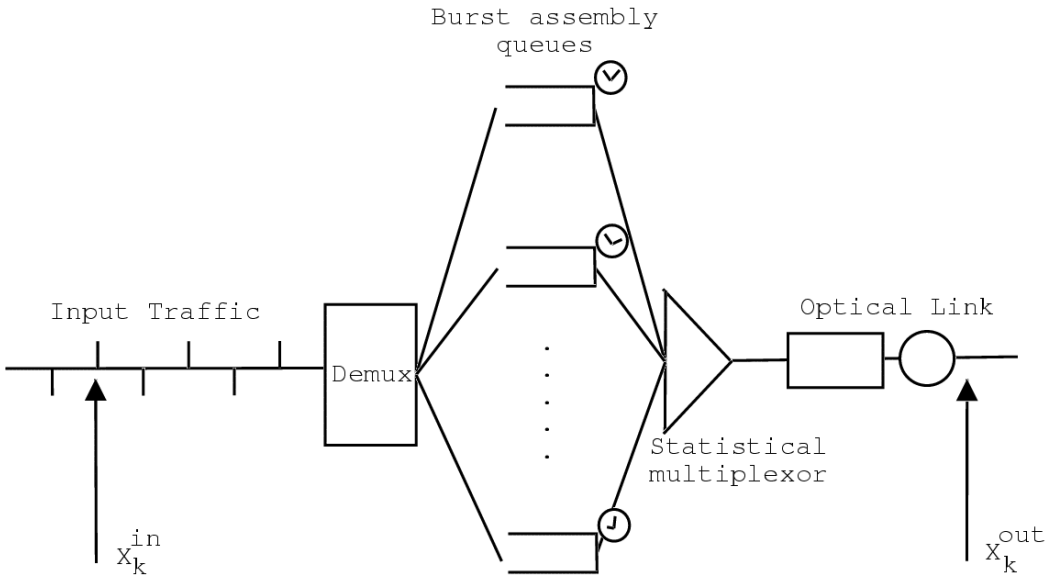
Traffic models for OBS

Universidad Publica de Navarra

Traffic models for OBS

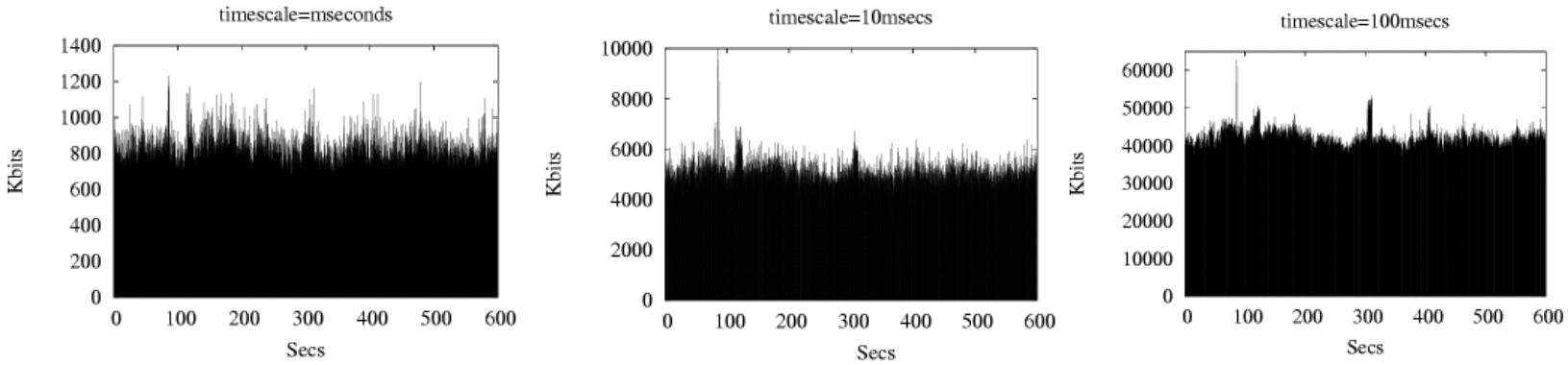
- The generated OBS traffic depends on
  - Burstification algorithm
  - Input traffic features:
    - Long-range dependence
    - Instantaneous burstiness
- No single traffic model can portray all possible scenarios!

Bustification algorithms



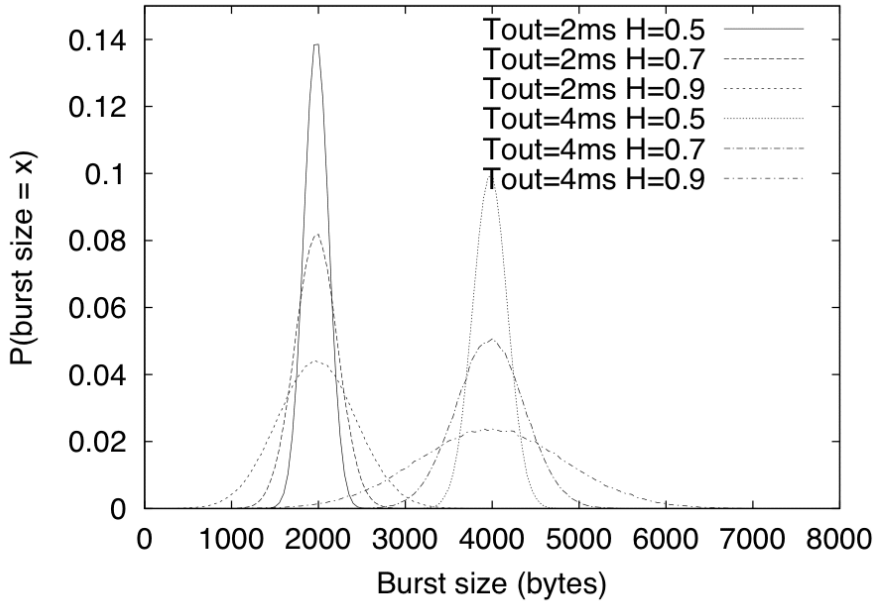
- Time-based, burst-size based or mixed-time-size based
- In all cases input traffic goes through demultiplex and then burst formation queues

Traffic models



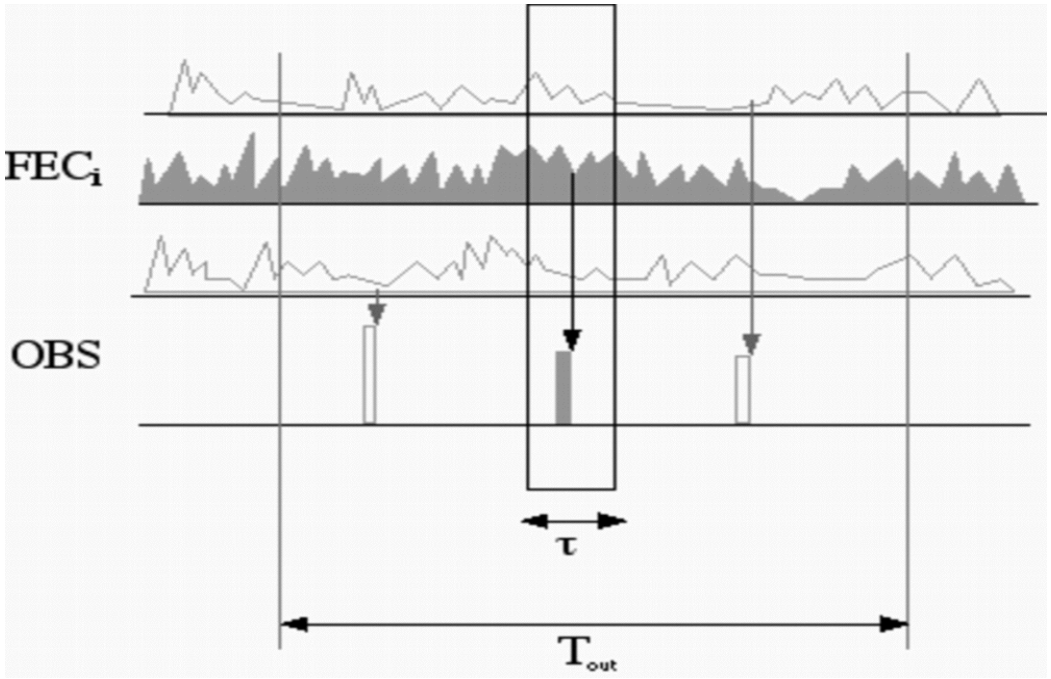
- Long-range dependence happens from a cutoff timescale, beyond which traffic may show independent increments
- For time-based burstifiers only the number of bytes per interval matters
- For burst-size-based burstifiers the packet arrival dynamics matter

Burst size



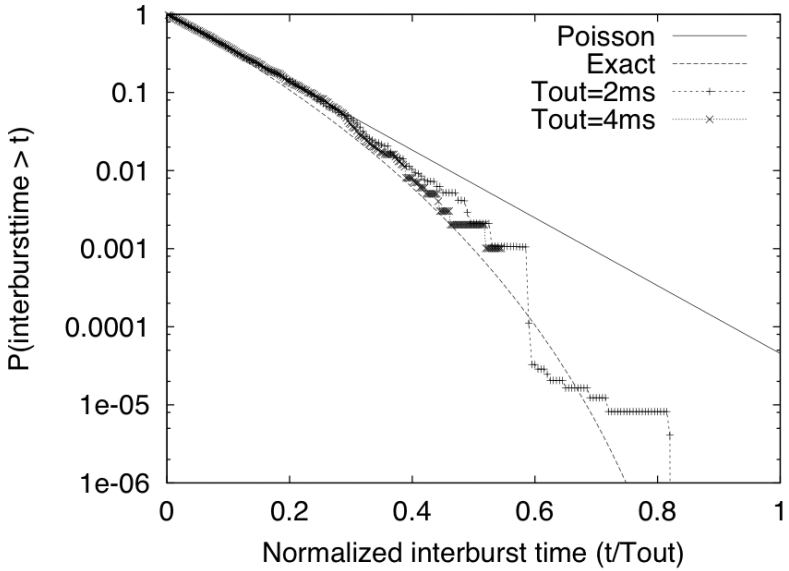
- For time-based burstifiers: Gaussian (or Gamma)
- For burst-based burstifiers: Constant

Long-range dependence



- Long-range dependence is inherited at large timescales only but at short timescales the statistical interleaving of bursts produces independent traffic

Burst interarrival time



- If many burstifier queues are statistically multiplexed to the same wavelength: Poisson at the burstifier output!

TCP over OBS

DEIS- Universita di Bologna

C. Rafaelli and M. Casoni*

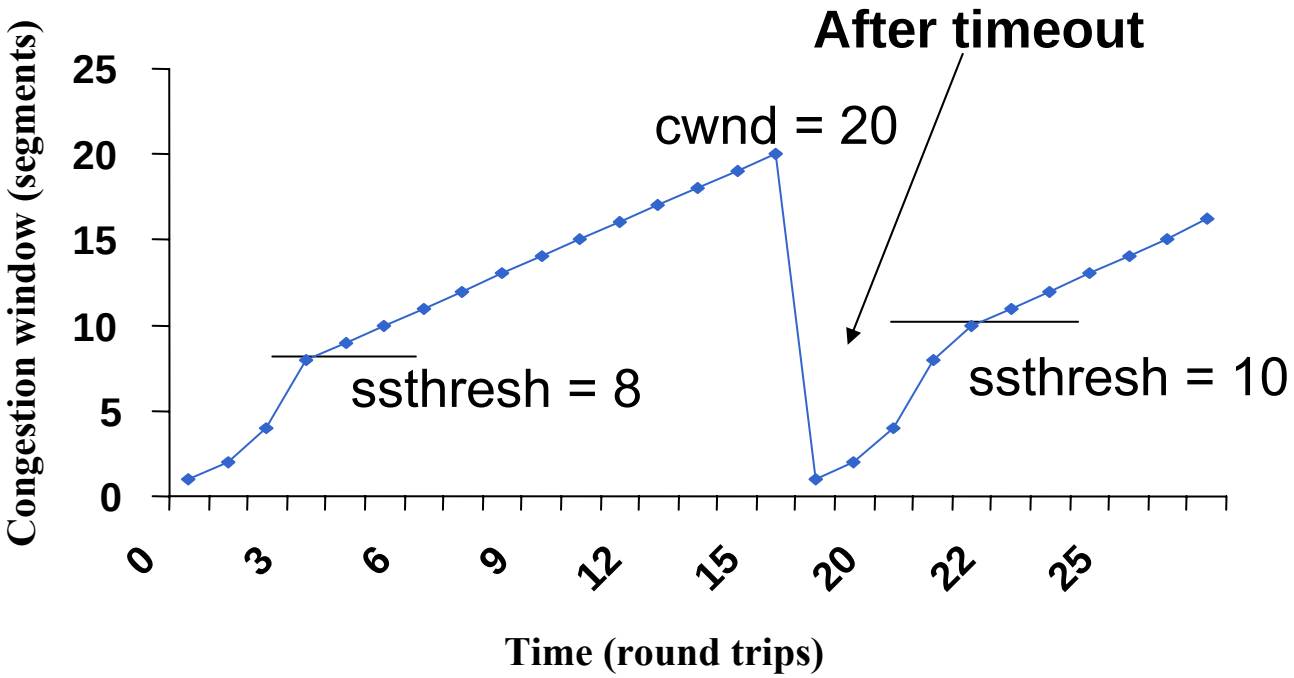
*Department of Information Engineering, University of Modena and Reggio Emilia

Basic TCP control functions

- 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

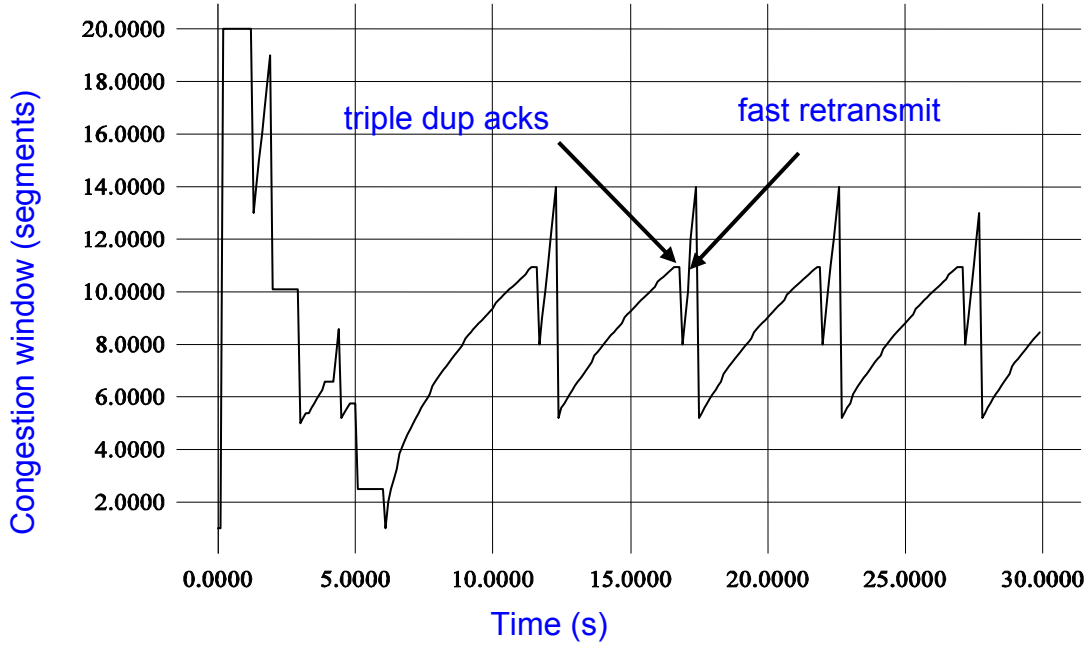
TCP Congestion Control

- Congestion window adaptation
  - Additive increase/multiplicative decrease



TCP loss detection

- Retransmission Time Out (RTO)
- Triple duplicate ACKs (TCP Reno)
  - Fast retransmit/Fast recovery



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)
are influenced
by OBS networks

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

- Fast source

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

  - All segments of the maximum window are emitted in T_b

- Slow sources

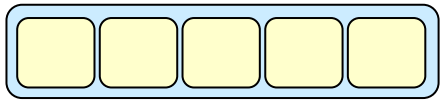
$$\frac{L}{B_a} \geq T_b$$

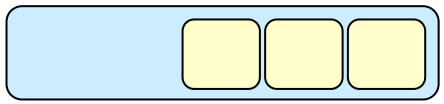
  - At most one segment emitted in T_b

- Medium source

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

Optical bursts





Burst loss

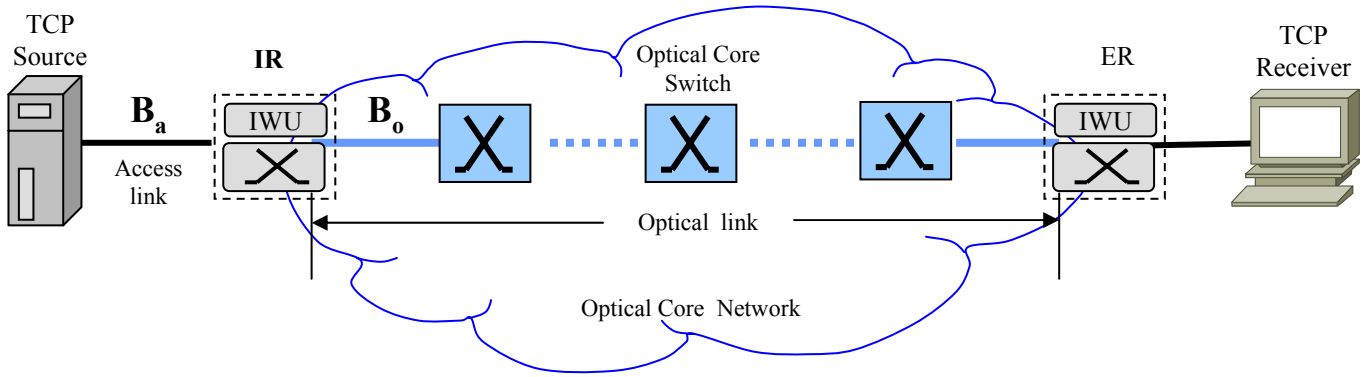
Burst loss is a consequence of contention in core nodes

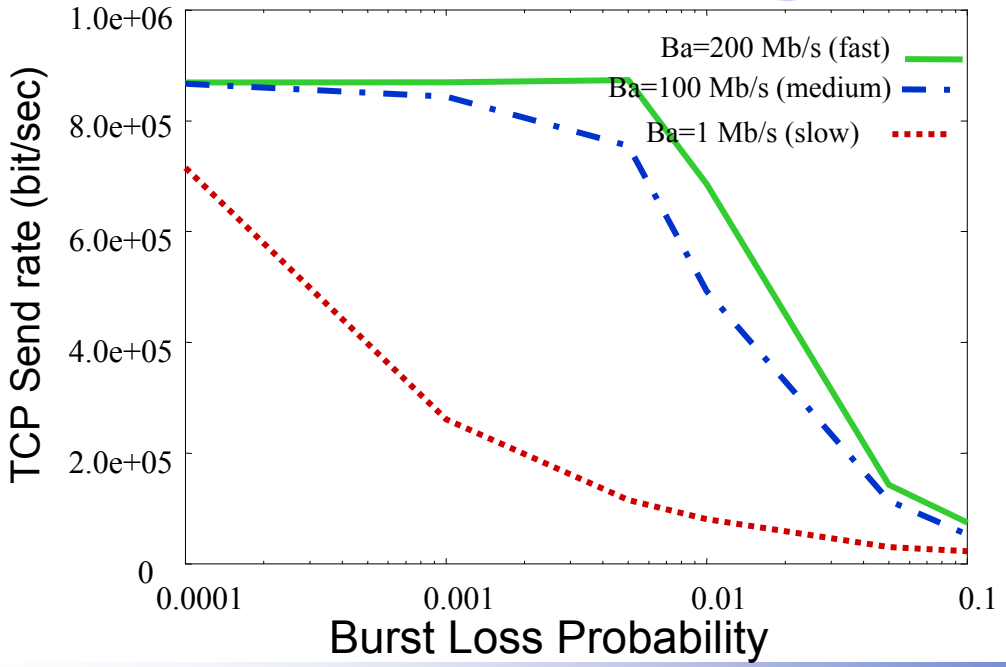
- Multiple segment losses
  - Depend on the level of aggregation of segments in a burst
- Retransmission time out is the main indication of loss for fast sources
  - Congestion window shrinks to 1 MSS when a burst is lost
- Slow sources recover mainly by means of fast recovery/fast restransmit

Correlation benefit

- Effect related to correlated segment delivery
  - Fast/medium source
- Fast window reopening is due to concentrated acks
- Congestion window quickly reaches its maximum value

TCP send rate for different sources





- 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

Variable delay *

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

Modeling TCP throughput *

- A simple model is able to calculate throughput as a function of burst loss probability p
  - p is a Bernoulli r.v.
  - Aggregation is accounted through the average number of segment in a burst E[N]
- The average TCP throuhput is calculated starting from the formula

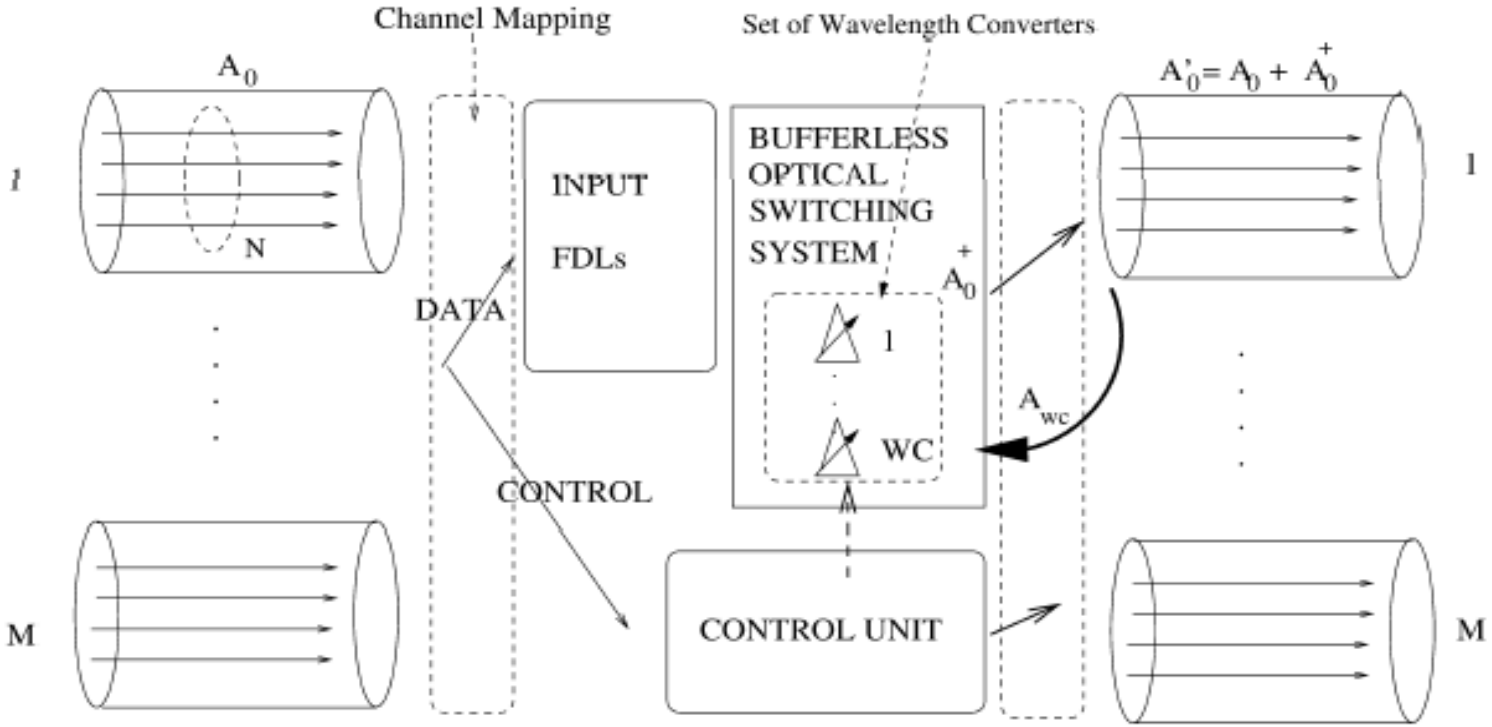
$$B_{TCP} = \frac{E[Y] + E[R]}{E[I] + E[Z^{TO}]}$$

  - where E[Y] is the average number of segments transmitted in bursts during during the interval I between two time out periods and E[R] is the average number of segments transmitted during the time out period Z^{TO}
- The result is

$$B_{TCP} = F(p)$$

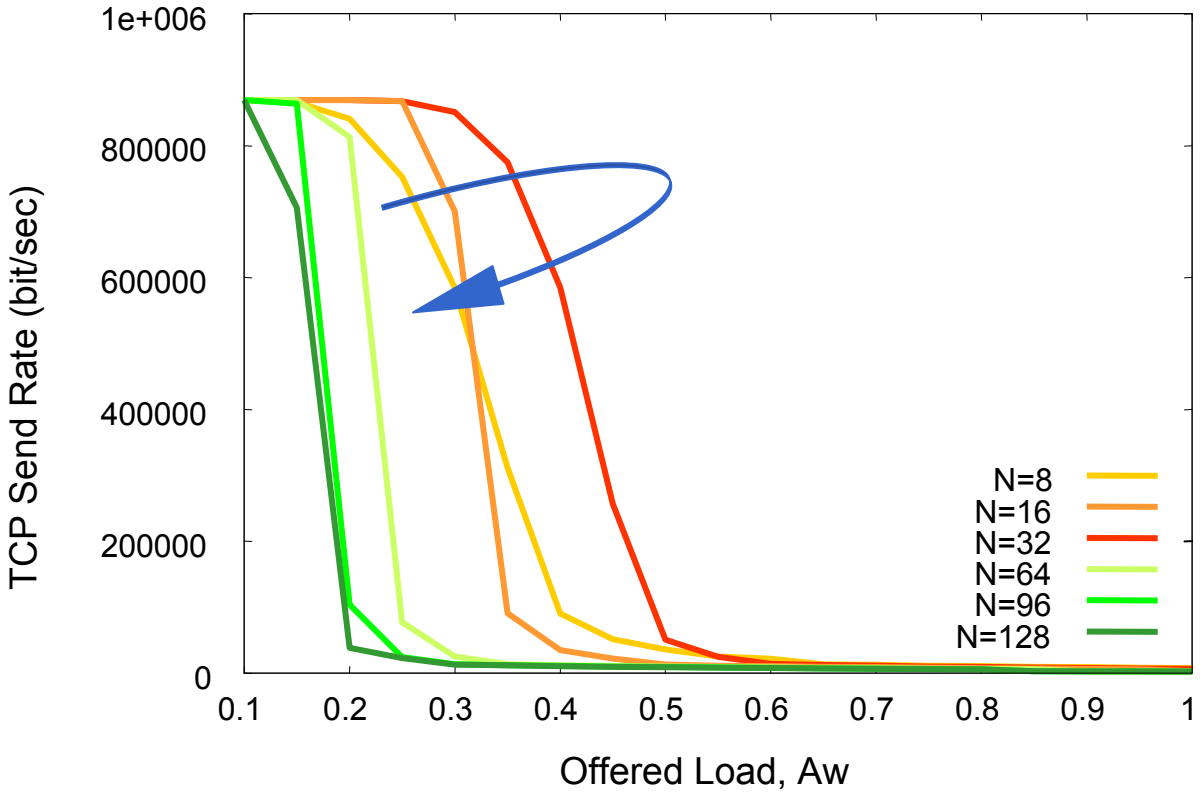
p is due to losses arising in the network and in the core nodes

Core switch architecture



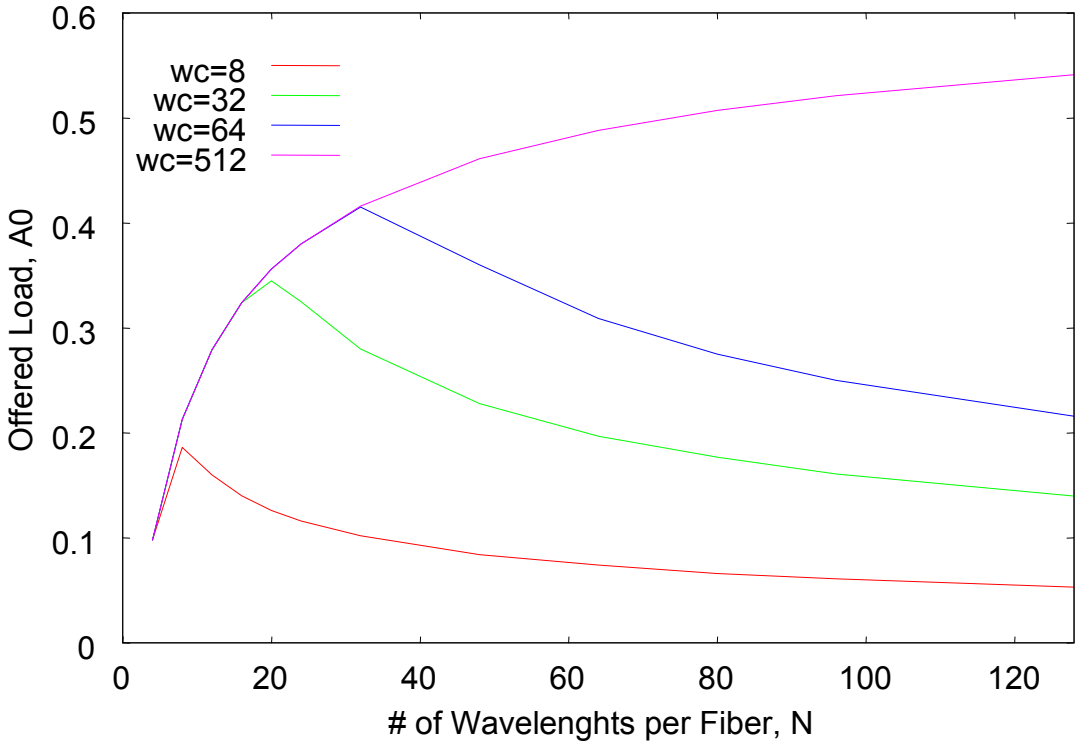
Effects of core node design

TCP send rate as a function
of offered load per wavelength



- N = wavelengths per fibre
- WC = 64
- Tb = 3 ms

Core node design optimization



- Maximum load per wavelength at 95% of maximum TCP send rate
- WC = no. of wavelength converters;
- N = wavelengths per fibre;
- T_b = 3 ms;

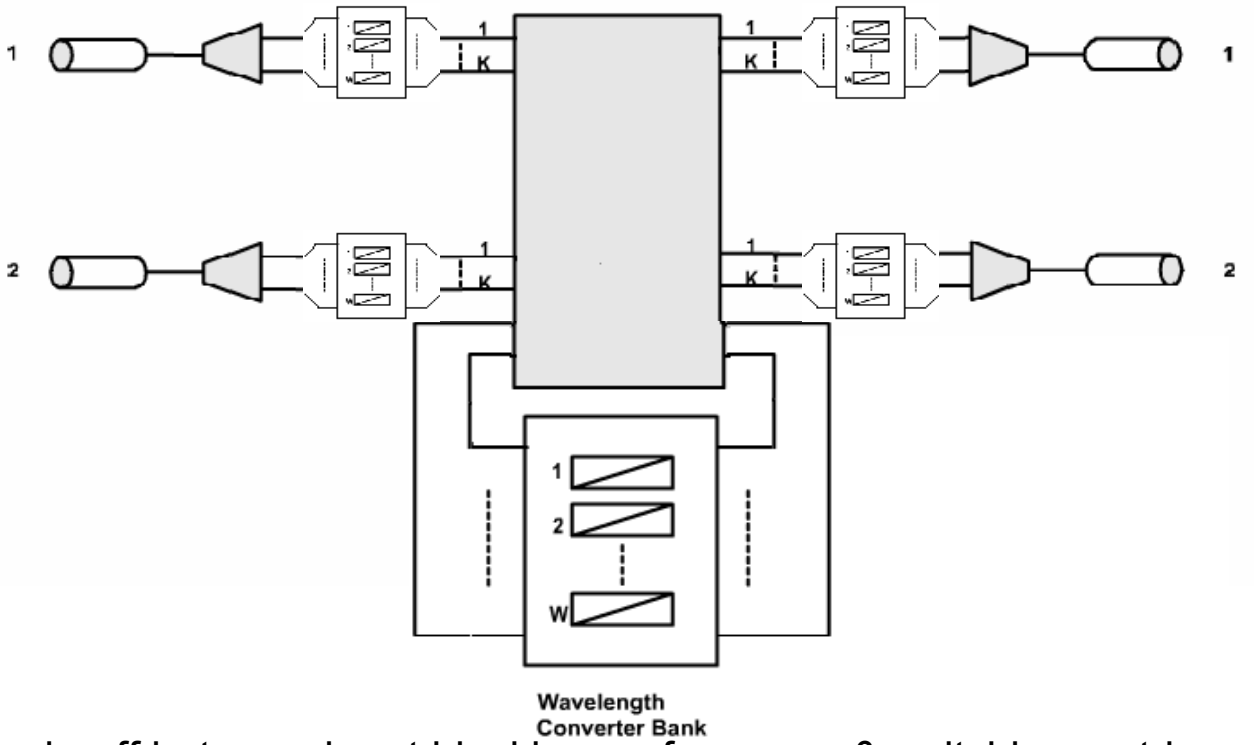
References

- A.Detti, M. Listanti, “Impact of Segment Aggregation on TCP Reno Flows in Optical Burst Switching Networks”, Proc. INFOCOM 2002, Vol.3 pp1803-1812.
- J. Padhye, V. Firoiu, D.F. Towsley, J. F. Kurose, “Modeling TCP Reno Performance: a simple model and its empirical validation”, IEEE/ACM Transaction on Networking, Vol. 8, No.2, pp.133-144, April 2000.
- J. He, S.-H. G. Chan, “TCP and UDP Performance for Internet over Optical Packet-Switched Networks”, Proc. of ICC 2003, pp.1350-1354.

Performance evaluation

Bilkent University

OBS Node - Architectures



- Trade-off between burst blocking performance & switching matrix complexity
- Share Per Node (SPN)
- Share Per Input Link (SPIL)
- Share Per Output Link (SPL)

OBS Node Model – SPL

- K wavelength channels per fiber.
- A Wavelength Converter (WC) bank of size 0 < W < K per output fiber; (Share Per output Line).
- 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.
- 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 WCs in the converter bank and using one of the free wavelength channels selected at random, else
  - is blocked

OBS Node Analysis

- W = K, Full Wavelength Conversion (FWC)
  - M/M/K/K loss system with offered load r=λ/μ
  - Erlang-B loss formula
- W=0, No Wavelength Conversion
  - K independent M/M/1/1 loss systems each with offered load r=λ/μK
- 0 < W < K, Partial Wavelength Conversion (PWC)
  - Can one have an exact analysis?
  - Seek numerically stable and efficient computational schemes

OBS Node Analysis

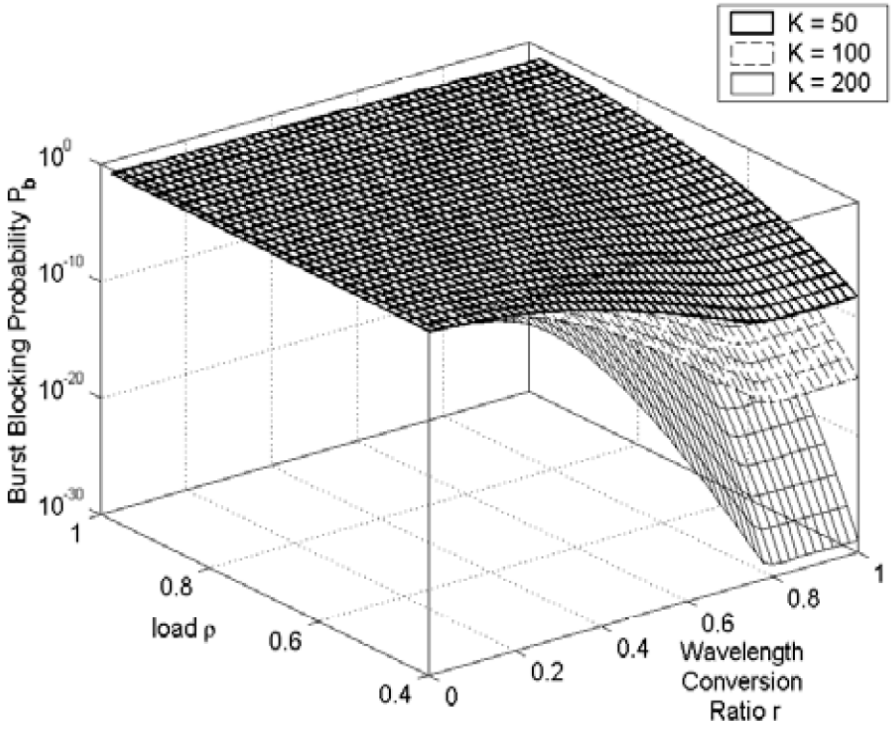
$$Q = \begin{bmatrix} A_0 & U_1 & & & \\ D_0 & A_1 & U_2 & & \\ & D_1 & A_2 & \ddots & \\ & & \ddots & \ddots & U_K \\ & & & D_{K-1} & A_K \end{bmatrix}$$

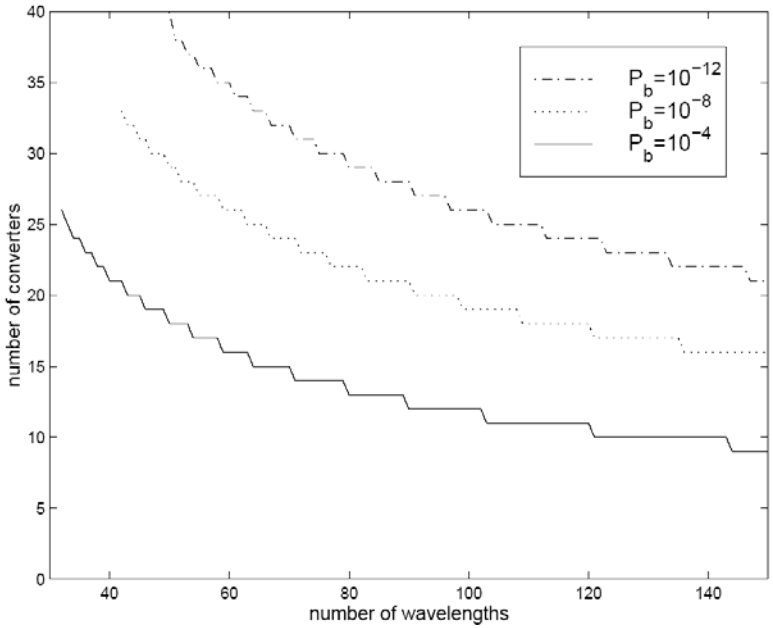
$$xQ = 0, xe = 1$$

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

- Formulated as the steady-state solution of a structured Markov chain
- Known as nonhomogeneous QBD (Quasi-Birth-Death-Process) in the applied probability literature
- Stable and efficient way to solve based on block LU factorizations
- Computational complexity less than O(K W^3) compared the brute force approach O(K^3 W^3)

SPL - PWC Results





- Model can be used for
  - Rapid production of burst loss curves
  - Iterative methods for finding cost-optimal choices for the pair (K,W) so as to satisfy a QoS requirement in terms of burst loss

Network-wide Study – Reduced Load Fixed Point Approximations

- Reduced offered loads are obtained by

$$\rho_j = \mu_j^{-1} \sum_{r \in R_j} \lambda_r \prod_{i=1}^{J} \left(1 - I(i,j,r) \times B_i\right), \quad \text{where}$$

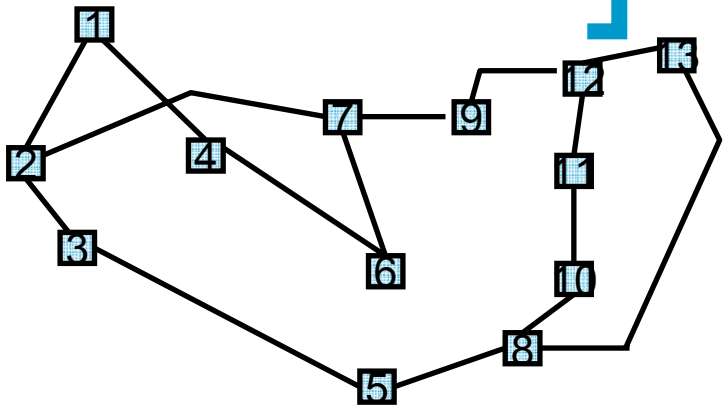
  - *I(i,j,r)* →1 or 0 whether or not *i* ∈ ***r*** and link *i* strictly precedes link *j* along route ***r***
  - B → vector of blocking probabilities
- Evaluate B_j's by using ρ_j's on each link by using the PWC model described before, denoted as PWC(ρ_j,W_j,K)

$$B_j = PWC\left(\rho_j, W_j, K\right)$$

- By successive iterations, approximate blocking probabilities could be obtained.

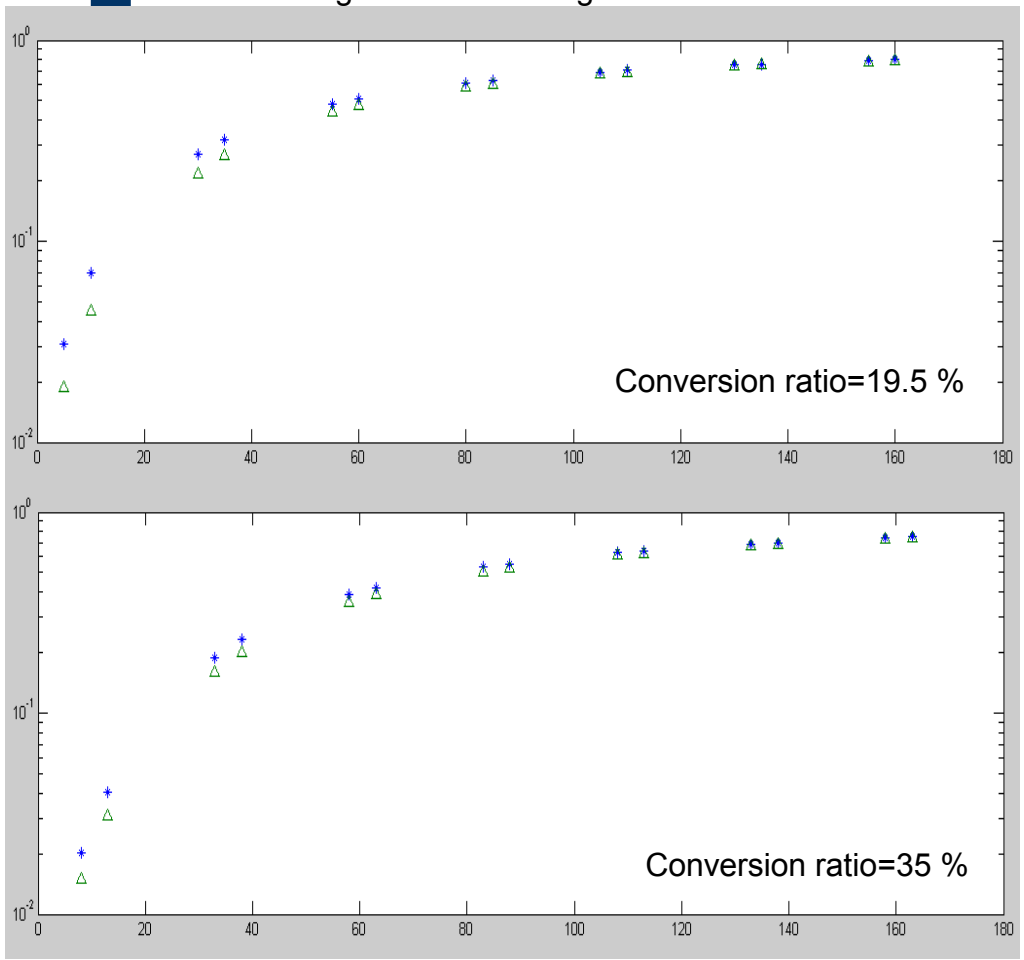
Test Network –NSFNET

- Homogeneous system, i.e., same arrival rates for all routes.
- Channel transmission speeds and burst mean lengths are adjusted to have mean service rate 25 bursts/sec.
- Nodes have various wavelength conversion capabilities distributed in a random manner.



Route	Route Hops
R1	1→4→6→5→8→10
R2	2→3→5→6→7
R3	2→7→9→12→13
R4	3→5→8→13
R5	5→6→7→9→12
R6	8→10→11→12→13
R7	10→8→5→6→4→1
R8	7→6→5→3→2
R9	13→12→9→7→2
R10	13→8→5→3
R11	12→9→7→6→5
R12	13→12→11→10→8

Results



- RLFPA underestimates blocking for low load values (Kelly Limiting Regime) & low conversion ratios (link independence assumption failing)
- The proposed OBS node analysis scheme can successfully be applied to a wide variety of network scenarios

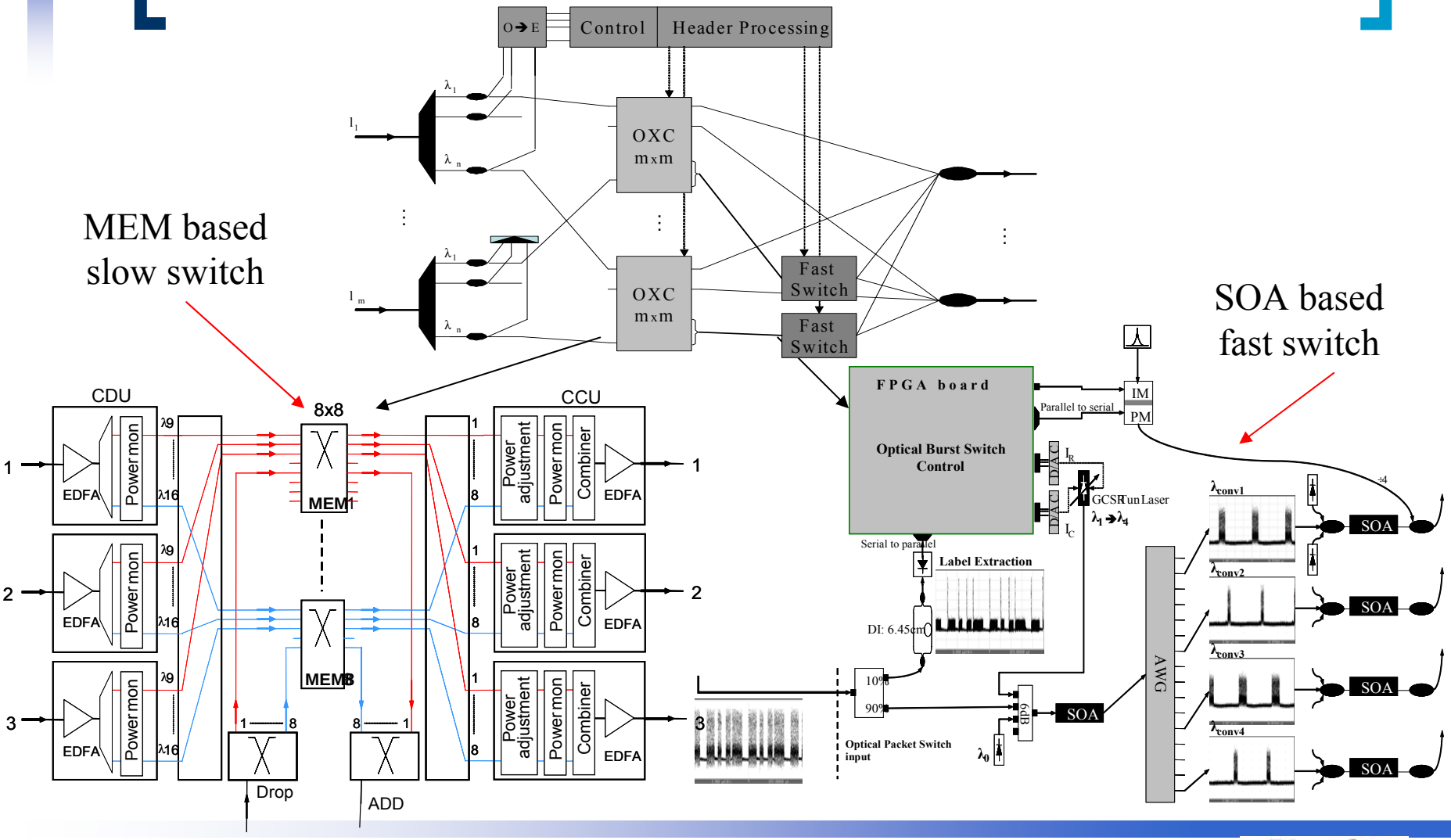
OBS in the real world: Testbeds

University of Essex

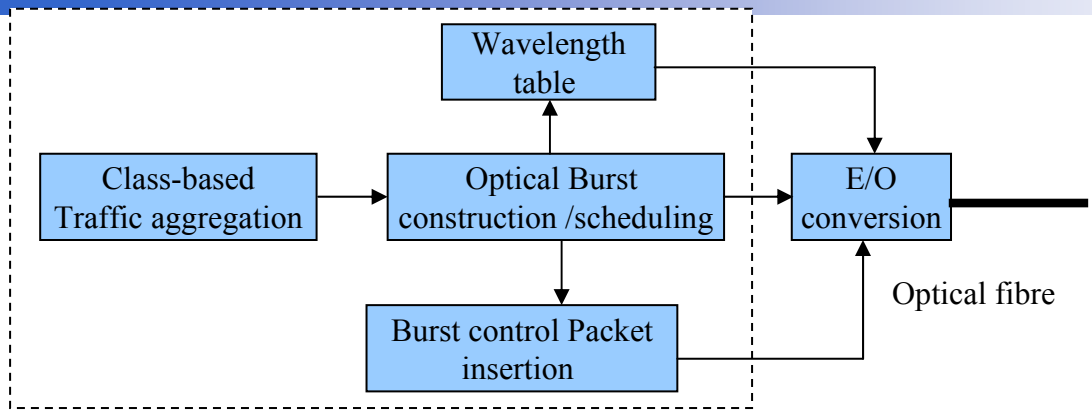
Core OBS Node Technology

- Traditionally OBS switches are based on slow switching technology (e.g. MEM)
  - Suitable for long bursts with large offset time
  - Not suitable for networks with large number of users transmitting small data bursts
    - Not efficient for short bursts with short offset time
- Combination of fast and slow optical switching technology is emerging for future OBS networks
  - SOA based switch technology for fast switching
  - MEM based switch technology for slow switching

Core OBS Node Architecture

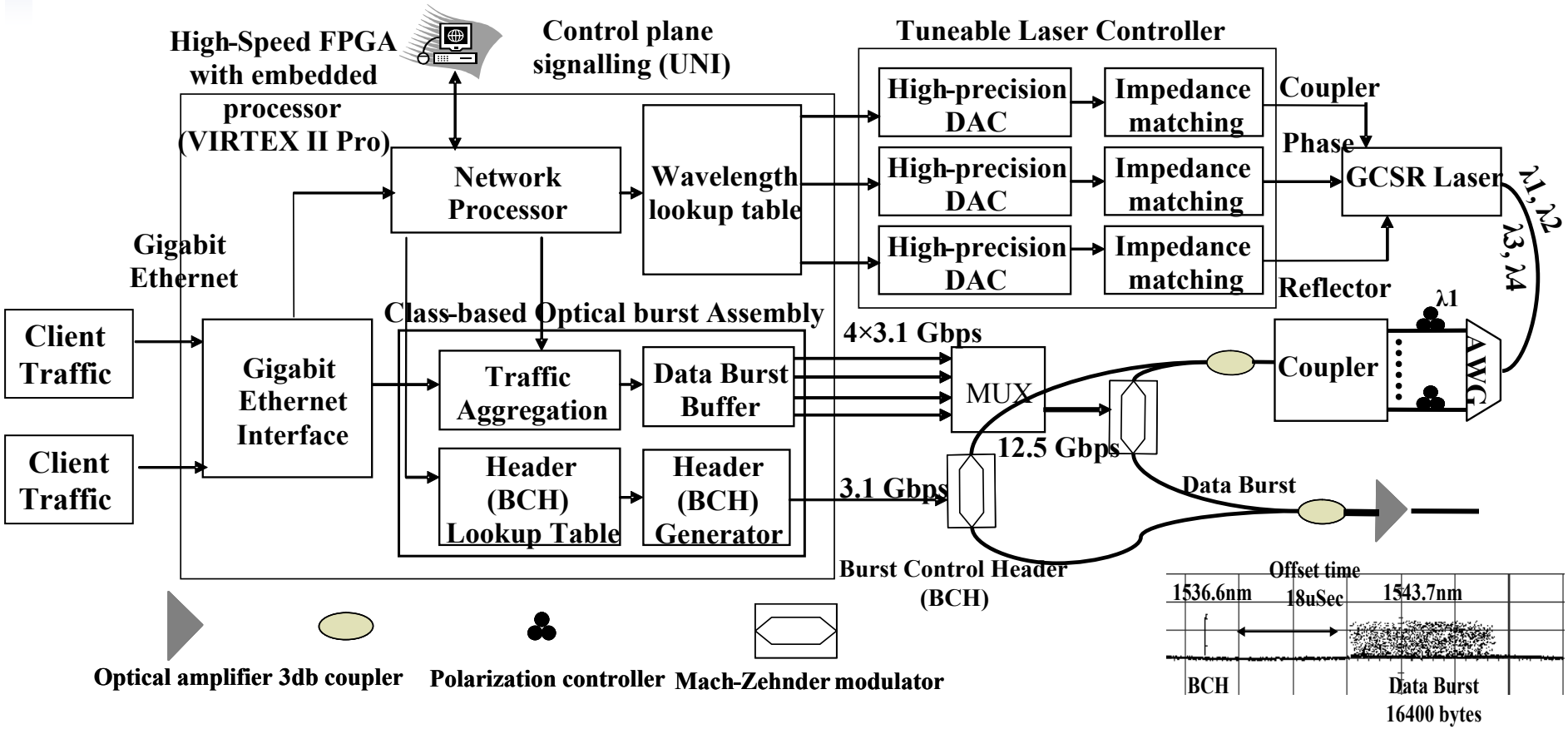


Edge OBS Node Functionality



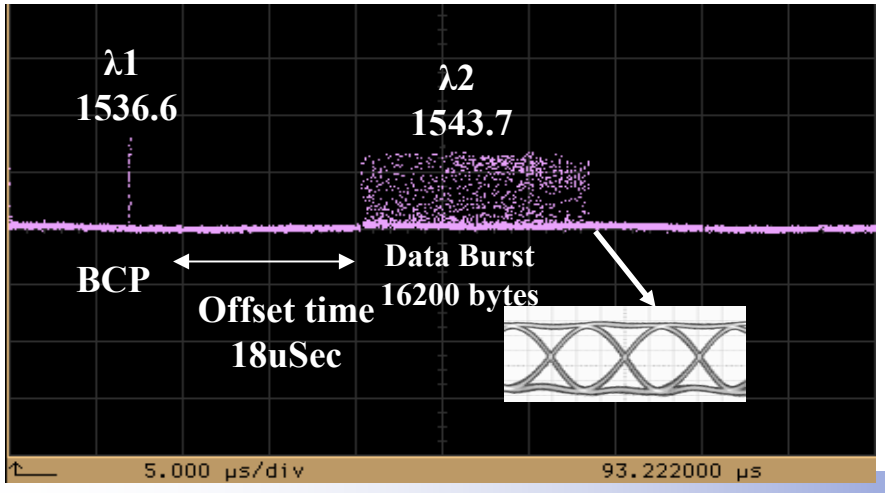
- User data (electronic packets) are aggregated in optical bursts based on:
  - Destination address
  - Class of service
- Optical bursts are scheduled for transmission based on :
  - Number of bytes per optical burst
  - Maximum experienced delay
- Optical control packet and wavelength are assigned based on:
  - Destination address and class of service
  - Information from control plane (lookup table)

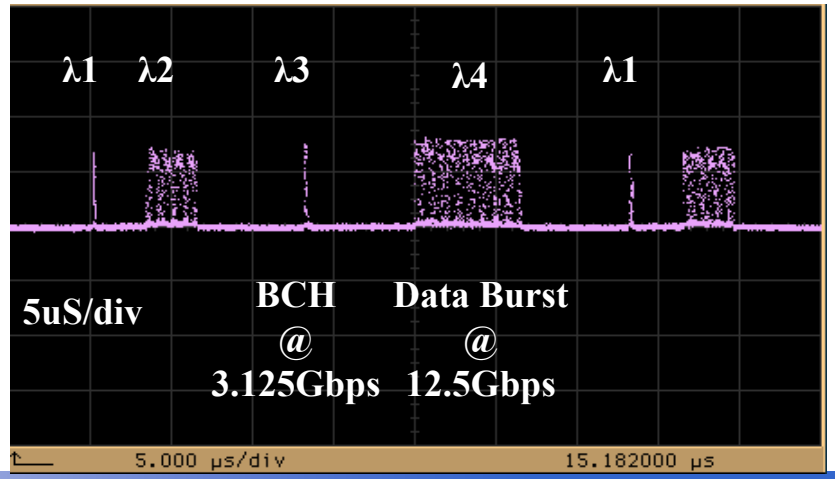
Edge OBS Node Architecture



Optical Burst Construction (Results)

- The length of payload is variable : 3KBytes, 16KBytes
- Optical Burst size are multiple integral of 100 bytes
- Optical Bursts modulate in four different wavelengths (1536.6, 1543.7, 1546.1, 1548.5 nm)
- Programmable BCP (length here =80 bytes=preamble+label+burst length+offsettime+wavelength)
- Offset time variable
- Guard band between two adjacent bursts is variable based on the traffic load

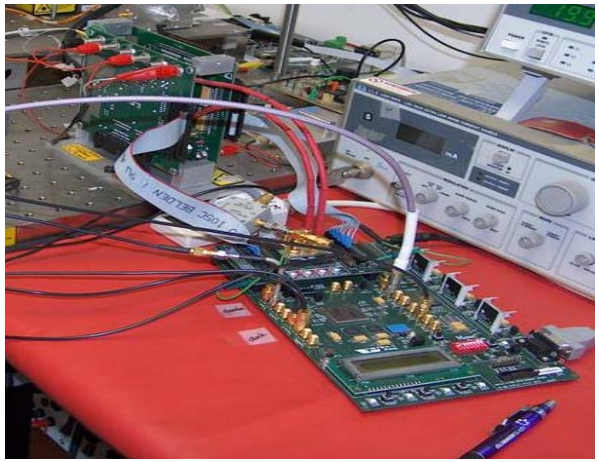




Edge & Core Routers

Fast SOA-Based Switch

Slow Mem-based Switch



Burst Generator + Tuneable laser

Dependability

Telenor/NTNU

Dependability – Definitions

- Dependability is in our context defined as
  - An optical networks ability to provide services at any random time
- Important distinction:

Availability of a service (”part of time”)

versus

Reliability of a service (”continous in time”).

Dependability – OBS challenges (1)

- Main rationale for an OBS network:
  - Use of packet based information transport is increasing.
  - Dependability functionality comparable to SDH/SONET must be present in packet based optical networks.
- Disruption of service delivery due to:
  - Fibre cuts.
  - Component/node failure.

Dependability – OBS challenges (2)

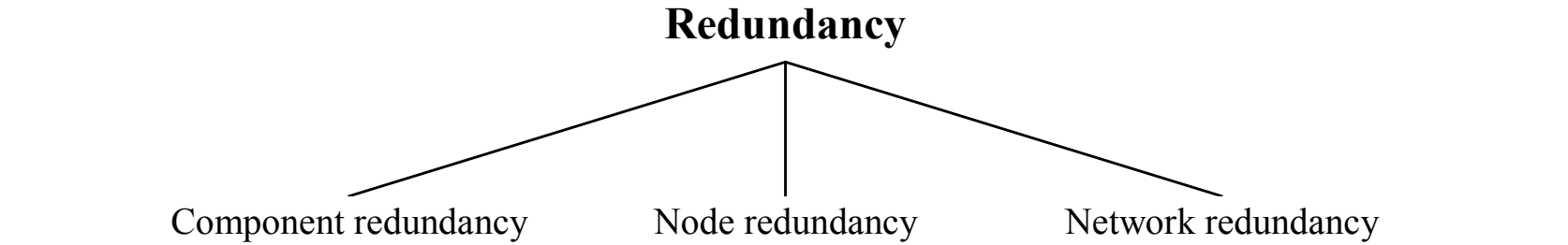
- Specific dependability challenges for OBS:
  - Use of offset time reservation protocol
    => vulnerability to loss of both control packet and data bursts.
  - Detecting failures by detecting loss of control messages
    => increased overhead and congestion due to need of frequent control message transmission.
  - No optical RAM for buffering during restoration.

Dependability - Main approaches

- Two main approaches:
  - **Fault avoidance:** minimize the probability of faults in the network.
  - **Fault tolerance:** continue to provide the intended service even when faults occur.
- Faults can never be completely avoided =>
  Goal: We should aim for building a fault-tolerant (survivable) OBS network.

Dependability - Fault-tolerance (1)

- Building fault-tolerance into a network by adding redundancy at different levels:



Component redundancy	Node redundancy	Network redundancy
■ Active or passive spare components used in a node. ■ Duplicated links between two nodes.	■ Duplication of a full switching node; active or passive operation.	■ End-to-end redundant paths used in the network; back-up may be pre-planned.

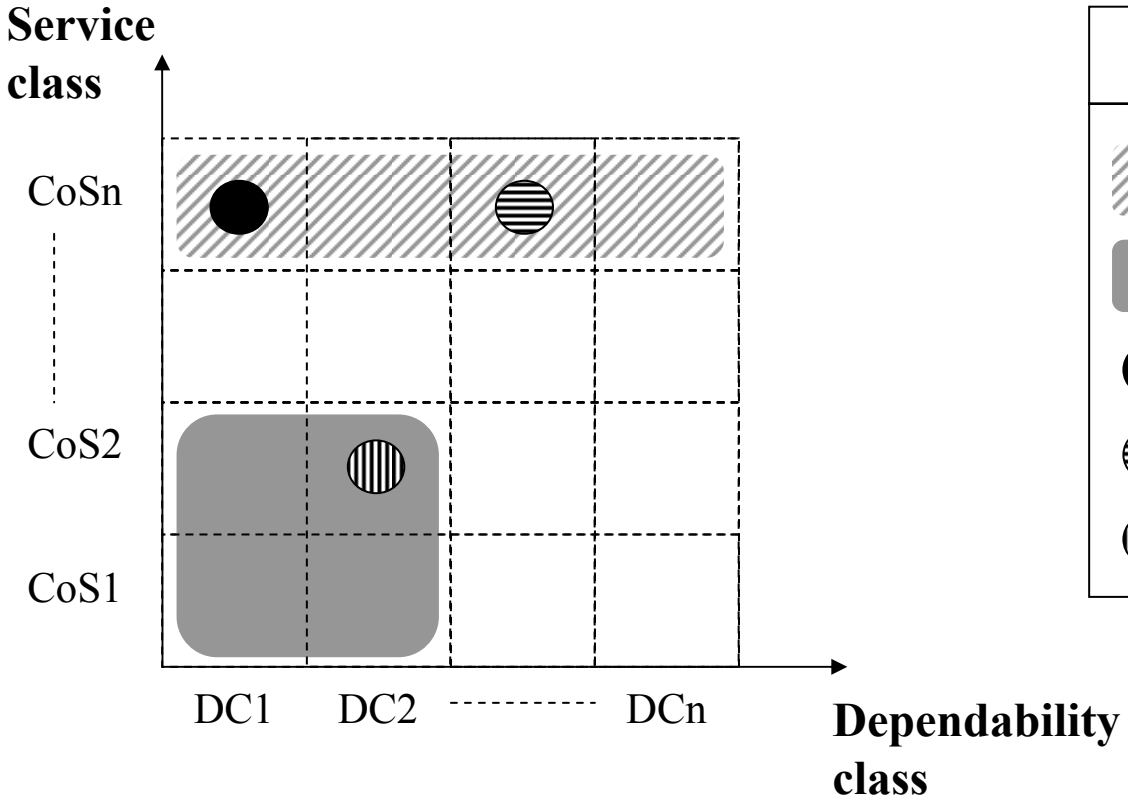
Dependability - Fault-tolerance (2)

- Mechanisms to achieve network redundancy involves control signalling and have slower response times than (local) mechanisms to achieve component or node redundancy.
- Coordination between mechanisms at different levels is important (e.g. avoid rerouting of traffic end-to-end if a local mechanism may be sufficient to handle the problem).
- Any fault must be reported to the control level for follow-up (e.g. change of spare part).

Dependability - Differentiation (1)

- We should aim for differentiated dependability:
  - To optimize resource utilization.
  - To enable price discrimination.
- The dependability level and performance level should be independently chosen:
  - A service with high performance (e.g. real-time) demands may have low dependability demands.
  - A given service may have different dependability demands in different situations or contexts.

Dependability – Differentiation (2)



- A ”service-session” is a given service used in a given situation or context.

Dependability - Differentiation (3)

Examples:

- High picture-quality video-telephony:
  - High performance demands (wrt. real-time and packet loss)
  - Moderate to low dependability demands depending on context.
- Messaging services:
  - Have in general very low performance demands.
  - May have very high dependability demands if used in a business context.
- Regular phone call vs. Emergency phone call:
  - Both have the same performance demands (high real-time demands but moderate packet loss demands)
  - Regular phone call has moderate dependability demands; Emergency phone call has very high dependability demands.

  i.e. even the same service may have very different dependability demands in different contexts.

Dependability - Implementation

- At least one service class should provide the same, or a better dependability level than traditional PSTN.
- A network operator may (for cost reasons) choose to implement only a few dependability classes in the (core) network.
- Finer differentiation (both wrt. performance and dependability) may still be achieved by additional mechanisms/procedures at the access point to the network.

Dependability - References

References

- B.E. Helvik, *Dependable Computing Systems and Communication Networks*, Department of Telematics, 2001.
- A. Fumagalli, L. Valcarenghi, “IP Restoration vs. WDM Protection: Is There an Optimal Choice?”, IEEE Network 14(6) 34-41 (2000).
- C. Metz, “IP protection and restoration”, IEEE Internet Computing 4(2) 97-102 (2000).
- O. Gerstel, R. Ramaswami, ”Optical Layer Survivability – An Implementation Perspective”, IEEE Journal on Selected Areas in Communications 18(10) 1885-1899 (2000).
- M. Yoo, C. Qiao, S. Dixit, “QoS Performance of Optical Burst Switching in IP-Over-WDM Networks”, IEEE Journal on Selected Areas in Communications 18(10) 2062-2071 (2000).

State of the Art 2.8

- Evaluation of the performances of an optical switch:
  - reliability
  - energy usage
  - port configurations and scalability
  - optical insertion loss
  - cross-talk
  - temperature resistance
  - polarization-dependent loss characteristics

State of the Art 3.8

- In the area of all-optical switches, optical switch designs can be classified into several categories:
  - optomechanical
  - liquid crystal
  - holografic
  - micro-electrical mechanical
  - thermo-optical
  - gel/oil-based
  - electro-optical
  - acousto-optic
  - semiconductor optical amplifier (SOA)
  - ferro-magnetic

State of the Art 4.8

- Micro-electrical mechanical machines (MEMs) use reflective surfaces to redirect light beams to a desired port. 3D MEMs have reflecting surfaces that pivot on axes to guide the light while 2D MEMs have reflective surfaces that "pop up" and "lay down" to redirect the light beam. MEMs easily scale to large port counts but can be a challenge to package due to the density and microscopic size of the light paths entering and exiting the substrate.

  In terms of optical insertion loss and switching speed, performance characteristics of optomechanical switches vary according to architecture.

  The drawback is the durability and cycle limitation of the mechanical actuator (for ex, an 80 x 80 microelectromechanical system switching module for use in optical cross-connect and burst-switching systems has been producted by Fujitsu-Japan).

- Planar lightwave circuit thermo optical switches are polymer-based or silica on silicon substrate. They use temperature control to change index of refraction properties of Mach-Zehnder interferometer based waveguide arms on the substrate. The light is processed by waveguide interaction and is guided through the appropriate path to the desired port.

  They are small but have high power requirements and optical performance issues. Furthermore, the size constraint, due to the fact that the paths must be quite long, restricts the technology's scalability limiting it to about 40 ports.

- Electro-optical switches use highly birefringent substrate material and electrical fields to redirect light from one port to another. A popular material used in an electro-optical switch is lithium niobate. The electrical field changes the index of refraction of the substrate, which manipulates the light through the appropriate waveguide path to the desired port.

  Are fast and reliable, but have high insertion loss and possible polarization dependence.

State of the Art 5.8

- Liquid crystal switches work by processing light polarization states by voltage applying. If a voltage is applied to the crystal, the polarization is changed to a known state and reflected via a beam splitter to an assigned output port. If no voltage is applied, the signal is passed through to another output.

  The benefits of liquid crystal technology for optical switch applications are an high reliability and a lack of moving parts but can be affected by extreme temperatures if not properly designed.
- Electroholography (EH) technology switchs wavelengths in the optical domain by appliyng a voltage to a specific crystal (potassium lithium tantalate niobate or KLTN). The applied voltage activates a pre-written hologram that works as a bragg grating deflecting the incoming wavelength.
- Acousto-optic switches receive acoustic-wave-induced pressure from a RF-fed piezoelectric transducer to generate fine gratings in optical waveguides. The gratings diffract lights to the desired port (for ex, a sub-microsecond photonic switching speed based on electroholography technology has been achieved by Trellis Photonics).

State of the Art 6.8

- Index-matching gel-and oil-based optical switches can be classified as a subset of thermo-optical technology because the switch substrate needs to heat and cool to operate. This index of refraction changed "bubble" or liquid, by thermally heating a portion of the switch, redirects the light stream through the appropriate waveguide path to the desired port.
  - There are still questions regarding long-term reliability and optical insertion loss.
- Semiconductor Optical Amplifiers technology uses SOAs that operate in the gain-clamped mode with some types of interferometric switch geometries to optically switch. They are characterized by an high degree of integration. These relatively basic elements can also be integrated with passive functions, such as splitters or wavelength multiplexers, to perform very simple wavelength add-drop functions in a metropolitan ring network.
  - Drawbacks of this technology are a high noise factor and interchannel crosstalk but whit a careful design, at the technology and system levels, it is possible to overcome these impairments.
- Ferro-magnetic technology uses magneto-optical Faraday effect in which electromagnetic waves interact each other directly to create extremely fast switches (femto-seconds).

State of the Art 7.8

- Burst Loss Probability (BLP) becomes an important measure to qualify an OBS architecture.
- Reasons for burst loss must be find out int the one way resource reservation, in the control channel congestion and in the limited resources coupled with high traffic loads.
- To reduce the BLP, wavelength conversion (wavelength domain) is the first solution if the path is quite long but becomes also quite expensive in shorter path situations where it is easier to find free wavelengths.
- Another approach in the time domain based on Fiber Delay Lines as recicling loops inside the switch fabrics.

State of the Art 8.8

- In the area of Optical Memories, until now, the only viable solution for optical buffering was to use switched fiber delay lines.
- Such optical buffers, when optimized thanks to WDM, can provide several tens of buffer positions to be shared by several tens of optical channels at 10 Gb/s or more.
- A short-term implementation for optical synchronization, required for efficient operation, could see the use optical or electronic memories in the form of a shared buffer, to save on system cost.