Advanced Lecture on Internet Infrastructure

12. Peta/Exa bps Router

Masataka Ohta

mohta@necom830.hpcl.titech.ac.jp

ftp://chacha.hpcl.titech.ac.jp/infra12e.ppt

Why High Speed Routers are Necessary?

- just for speed
  - 100Mbps*(50000 subscribers)=5Tbps
  - limit of electric interface speed is tens of Gbps

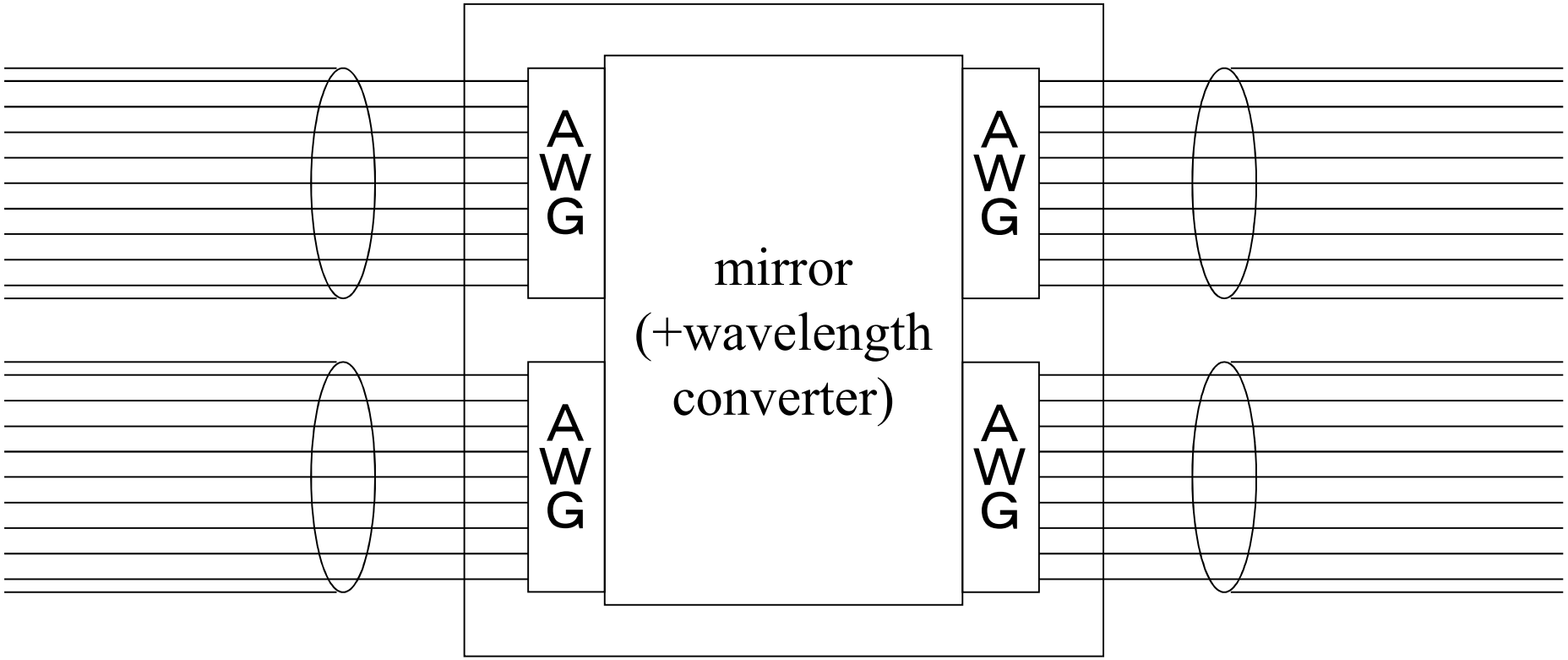
Proper Use for Optics and Electronics

- optics
  - scarcely no interference, almost no nonlinearity
    - best for transmission, logical operations almost impossible (optical computers not feasible)
  - ultra wide band (propagation speed is not very fast)
- electronics
  - strong interference
    - no good for transmission
    - good for logical operations and control

Optical Fiber Delay Line and Slow Light

- optical buffer may be made from delay lines
  - long fiber is necessary (240m for delay of duration of 1500B packet @10Gbps)
- with slow light (series of high Q resonators)
  - light intensity changes slowly
  - can construct buffer with short delay line?
  - slow change means low bps, longer packet duration, longer delay line

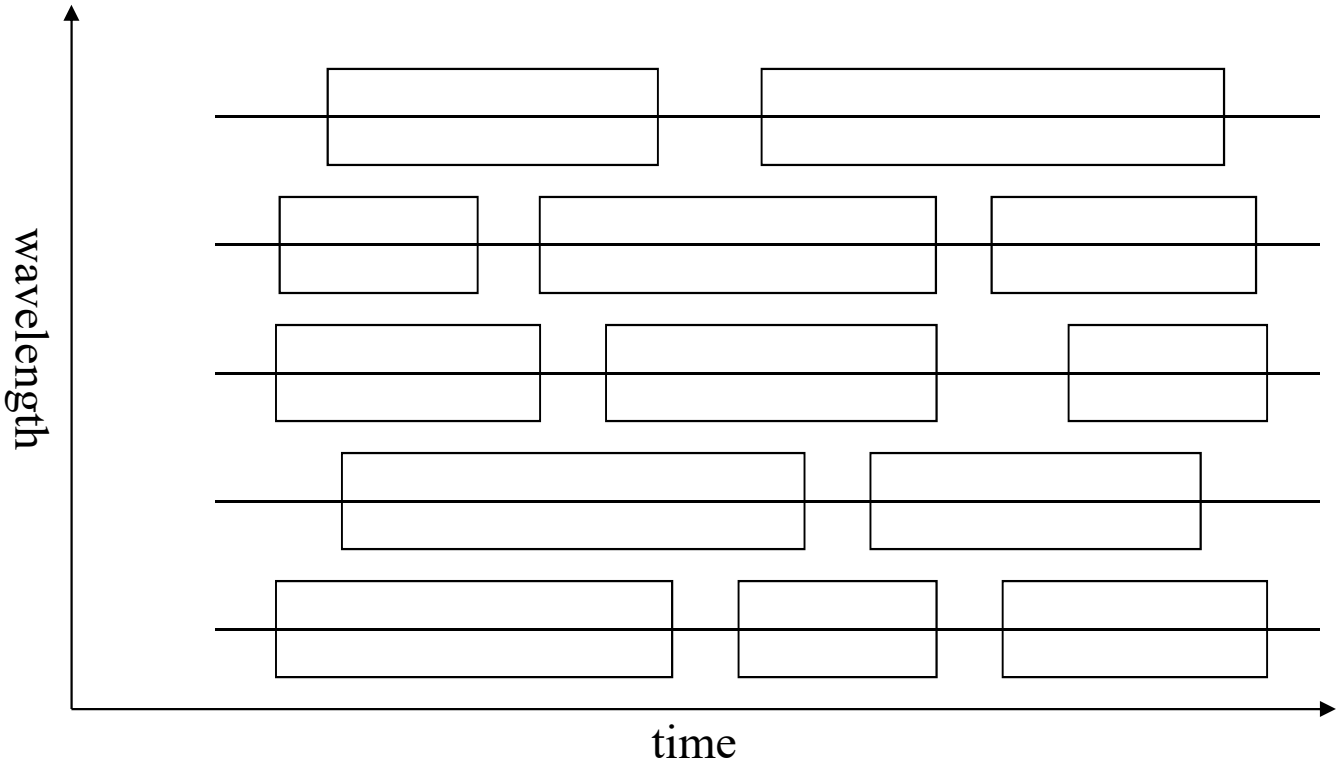
Wavelength Routing



What's Wrong with Wavelength Routing?

- Tbps scale wide BW of optical fiber is
  - divided into 10Gbps*100 or so
  - # of equipment (power) increase at least proportional to # of wavelengths
- with optical transmission, on the other hand
  - all the optical BW is amplified by single EDFA
  - the reason why WDM so successful
- WDM for transmission, not for exchange
  - exchange all the wavelength at once!

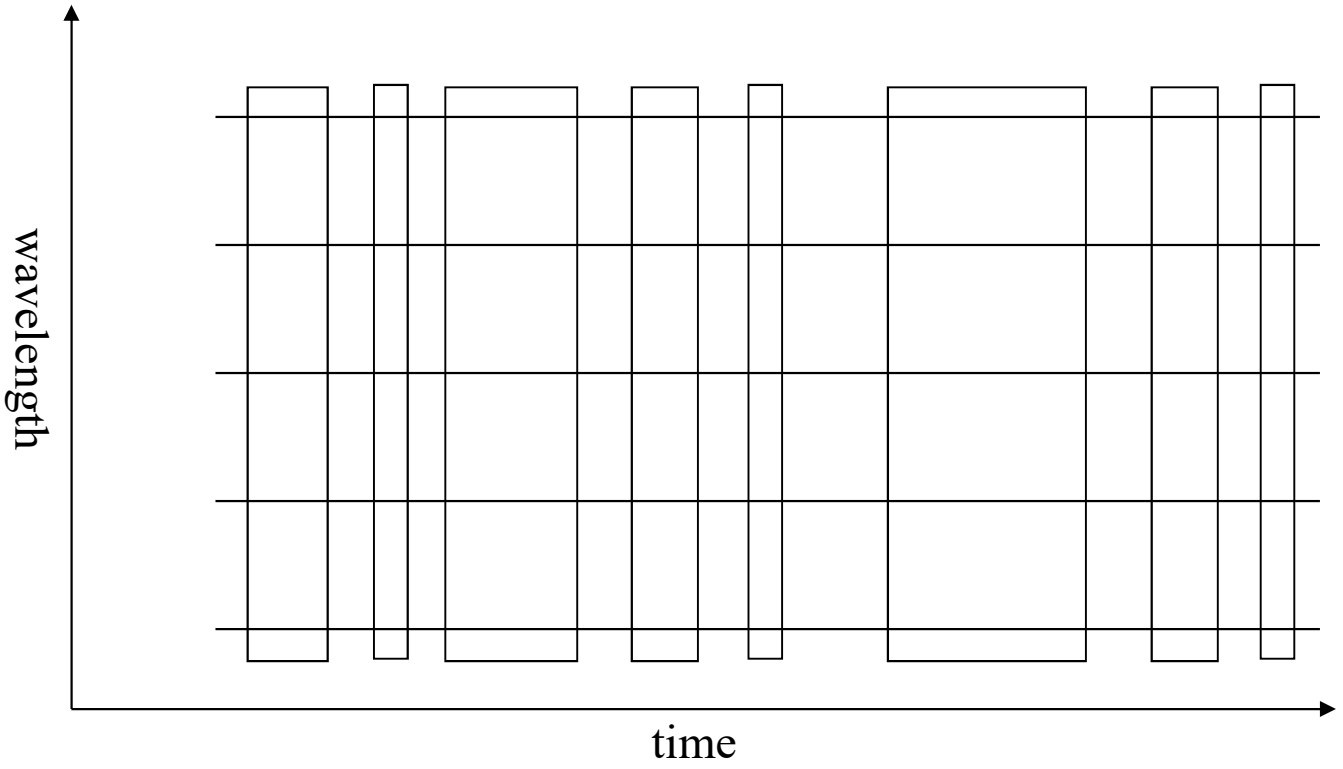
IP over WDM and Packet Multiplexing with WDM



IP uber Alles

- multiplex by packet only!!
  - all the BW should be used for transmission of each packet
- high speed (100ps) optical switches available
  - data path should be optical
  - control?
- “almost all-optical” router
  - electric control fast enough for packets@1Tbps

IP over WDM and Packet Multiplexing without WDM



High Speed Optical Switch

0.125W power consumption for ±2.5V control @50Ω

Speed of Optics and Electronics

- electrically controlled optical switch
  - can switch within 100ps
- 500(1500)B packet @ 1Tbps
  - 4(12)ns
- clock speed of recent LSIs
  - >>1GHz
- Tbps almost all optical router
  - can be implemented with electric control

Optical Packet Buffer?

- 500(1500)B @ 1Tbps
  - 4(12)ns long in time
  - 0.8(2.5)m long in optical fiber
    - loss of 0.037kT(T=300K) if bit consists of 10 photons
  - @10Gbps, 100 time longer fiber necessary
    - not very practical
    - 100 times parallelism necessary for 1Tbps
- 2.5km for 1000 packet duration
  - 15cm*15cm*4cm box for 4km fiber

Compact Time Delay Coil

Winding a large fiber spool is easy; but making compact and low loss fiber coils demands attention, precision, and skills. With specially designed & computerized machinery and proprietary manufacturing process, we can produce extremely low insertion loss fiber coils that fit your budget and tight space. No more large fiber spools to occupy your precious space and no more high loss associated with the small size! Our optical fiber coil fills a long overdue vacuum in the photonics market, where large time delay and small size are essential. Each coil is ruggedly packaged to withstand various environments in field applications. Bare coils are available for OEM applications.

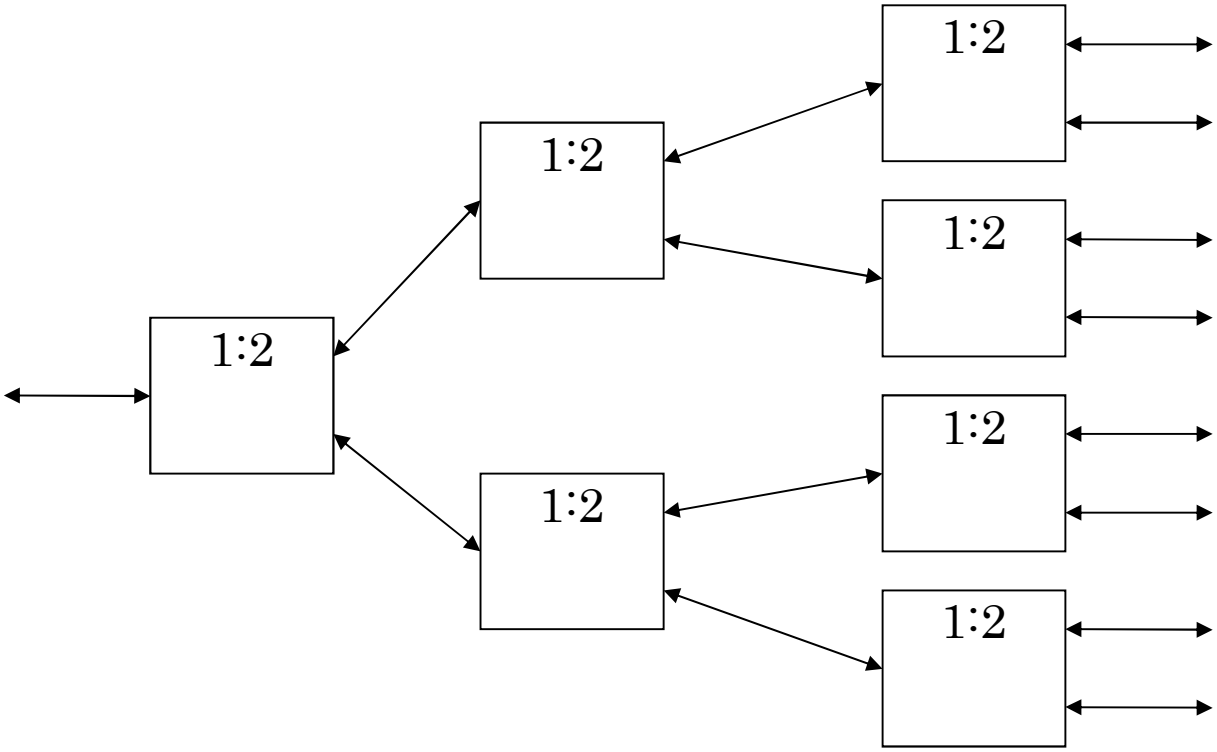
Specifications:

Insertion Loss	< 0.3 dB/km typical, < 0.5 dB/km max. (above intrinsic loss)
Fiber Length	10 m up to 4 km
Optical Delay	Nanosecond to microsecond depending on fiber length and type
Operating Wavelength	1260 ~ 1650 nm standard, others specify
Fiber Type	Corning SMF-28 standard, others specify
Operating Temperature	-40 ~ 85 °C
Storage Temperature	-40 ~ 85 °C
Dimensions	Ø 3.5" (I.D.) standard 6.00" x 6.00" x 1.59" with enclosure

(Values are referenced without connectors)

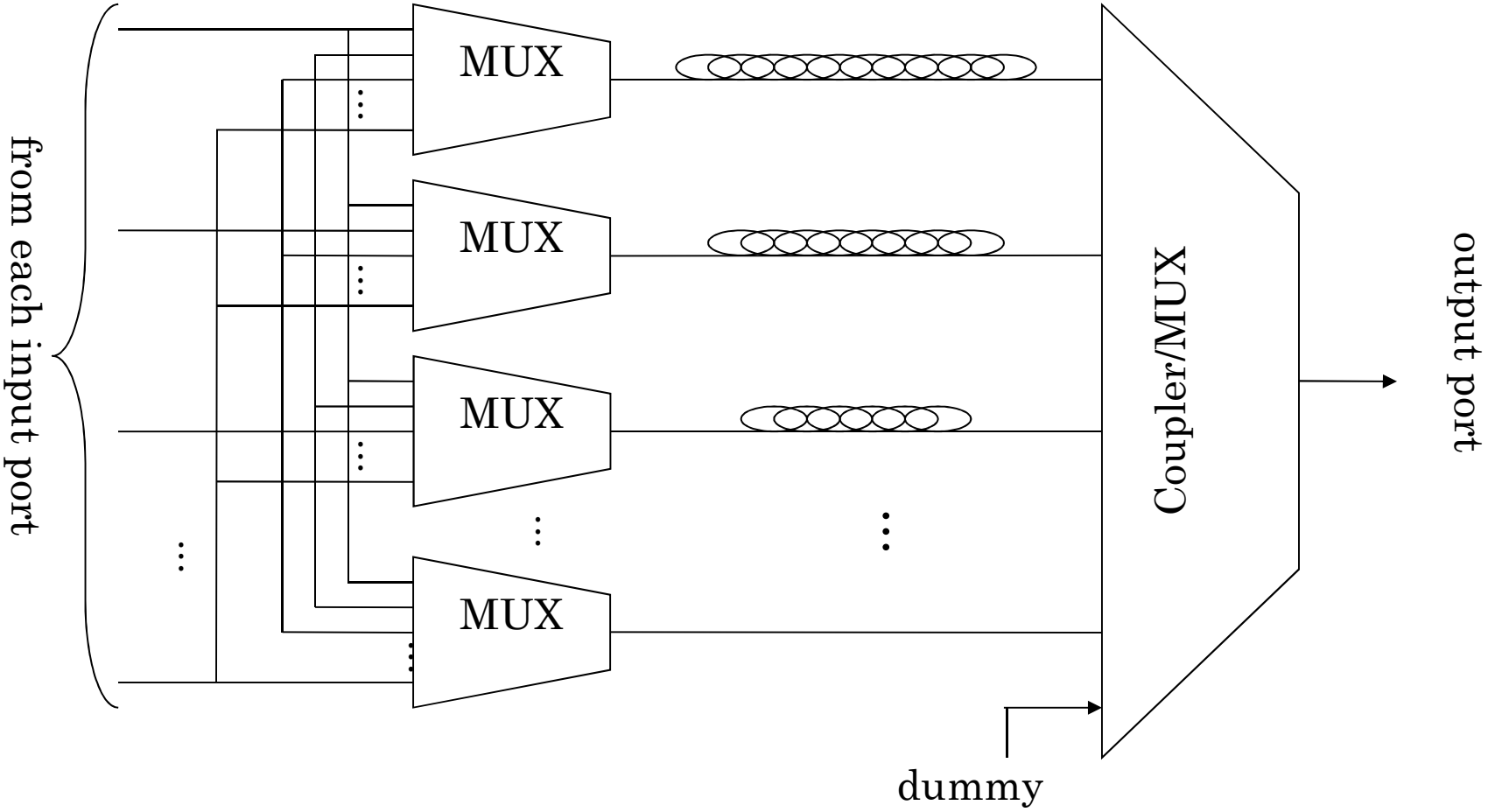
General Photonics Corporation社カタログより

MUX/DeMUX



optical MUX (⟶)and DeMUX (⟵)

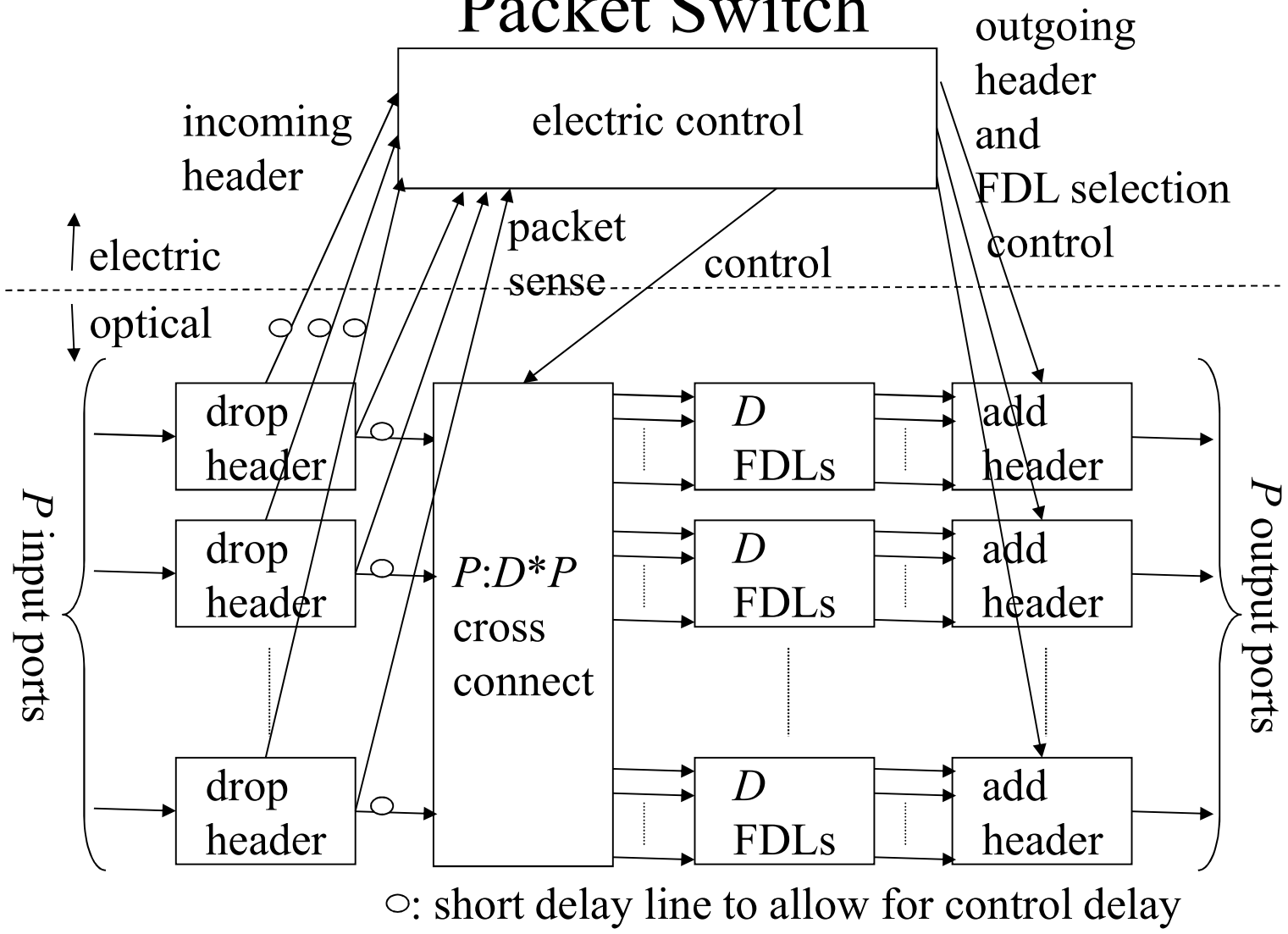
Optical Buffer with FDLs (Fiber Delay Lines)



Traffic Pattern at the Internet Backbone

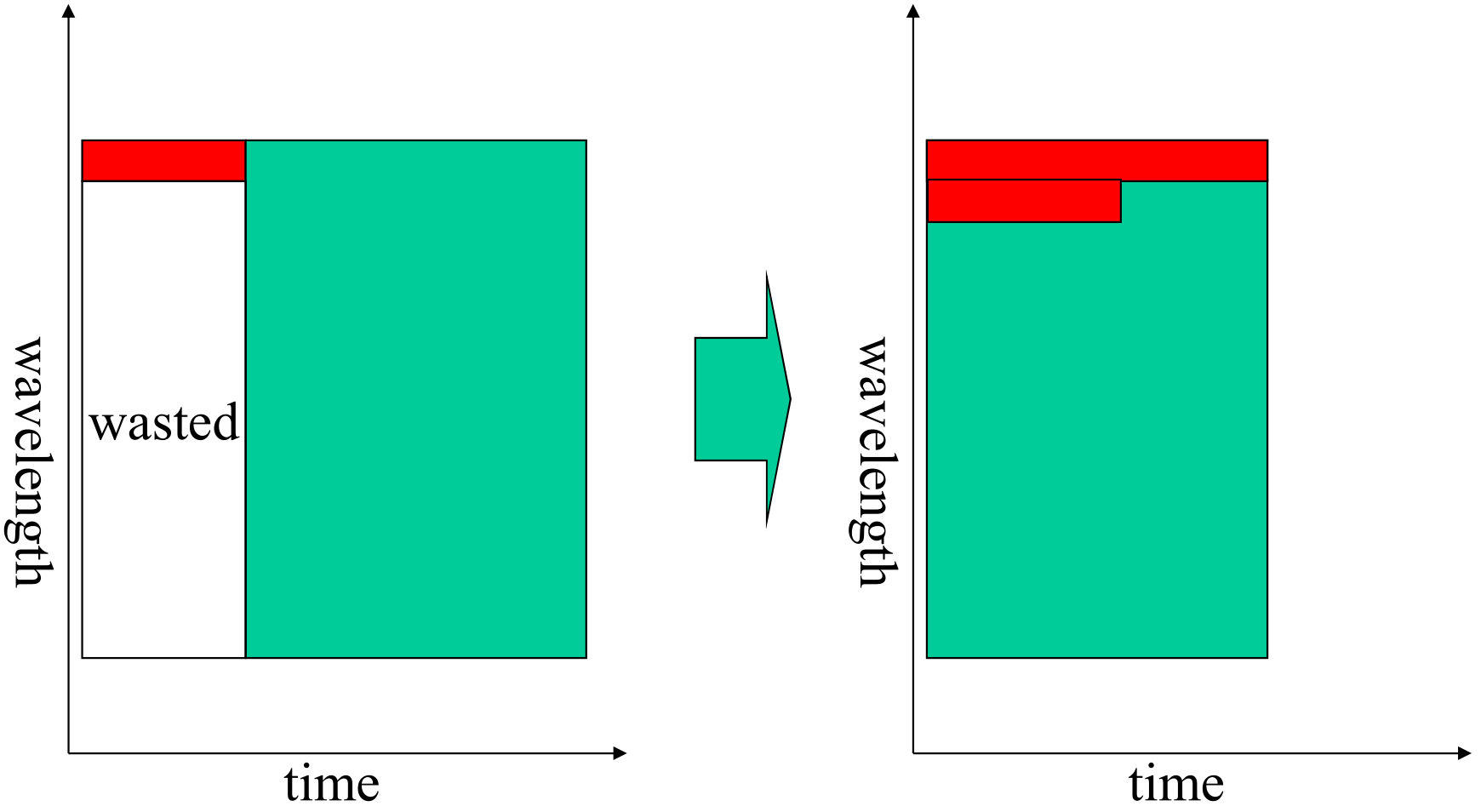
- Poisson
  - variation of each TCP is smoothed
  - buffer of several tens of capacity is enough
- average packet length
  - several hundreds of byts
- # of TCP connections
  - several tens of thousands

A Micro Architecture of A Proposed Optical Packet Switch



Packet Format

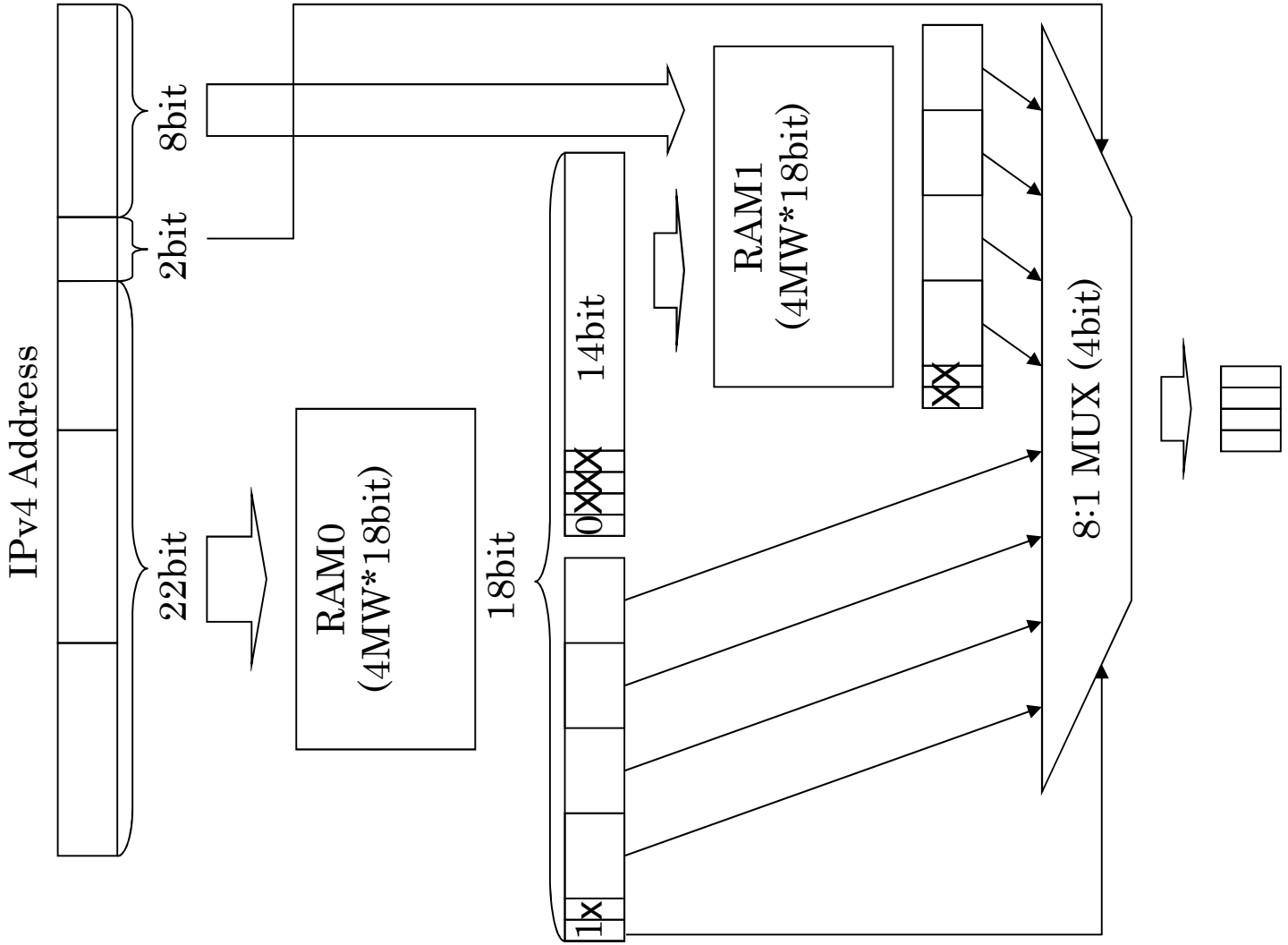
- 500B over 100 wavelength: 5B/wavelength
  - 5B may be shorter than header
- packet consists of header and payload
- if header and payload are separated by time
  - no payload can be sent while sending header
- header and payload are separated by wavelength
  - header may need multiple wavelengths

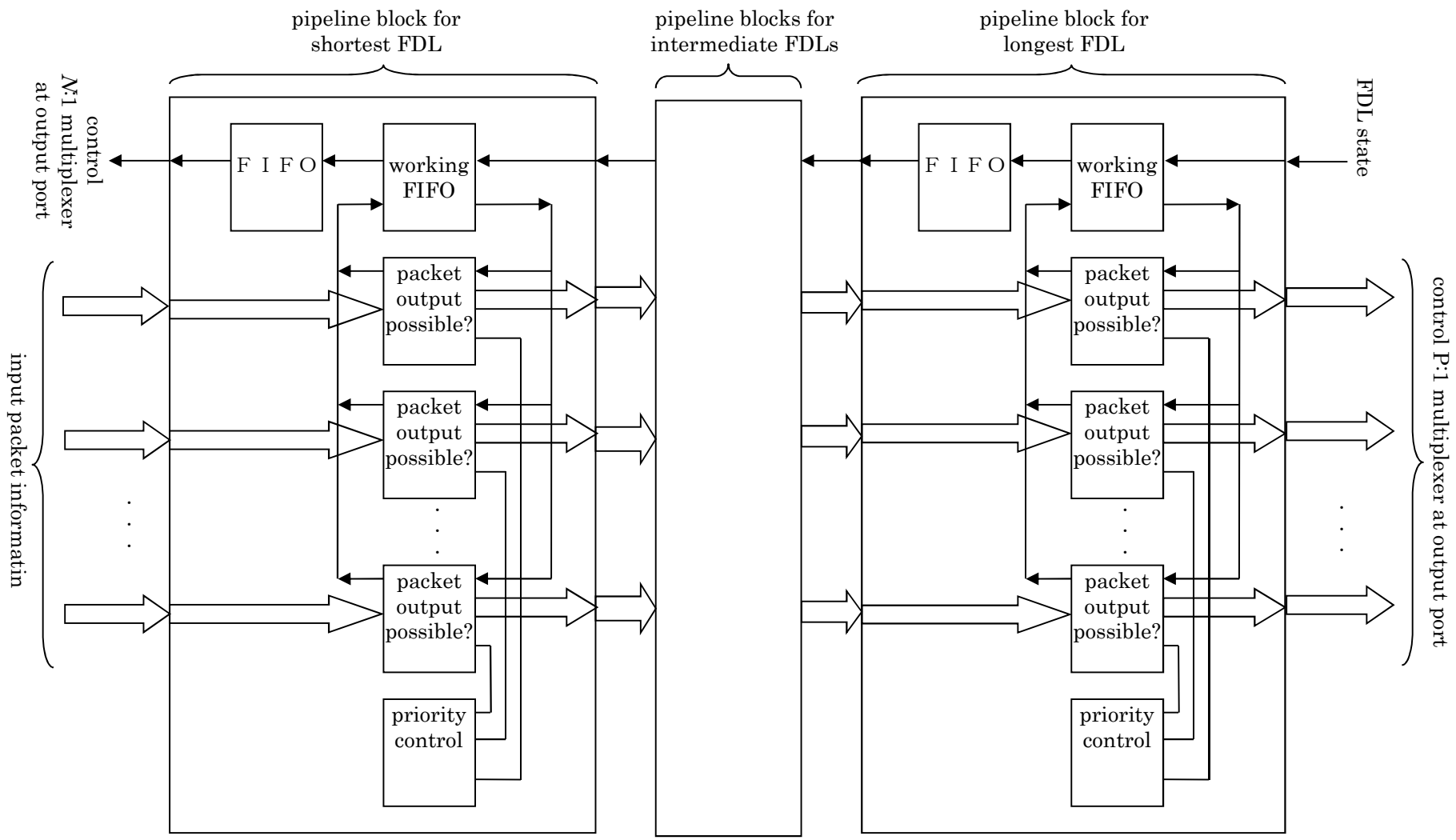

wavelength
wasted
time
wavelength
time
: header
: payload
separation of header and payload

Electric Circuit

- routing table
  - full route for /24 and host route for 16k /22
  - 2 SRAM chips pipelined with 3.3ns clock
  - IPv6 needs more pipeline stages
- FDL control
  - can be pipelined for each FDL
    - though # of input/output port cannot be large
  - 4ns pipeline with 550MHz FPGA

Pipelined Lookup of Routing Table for IPv4 Address

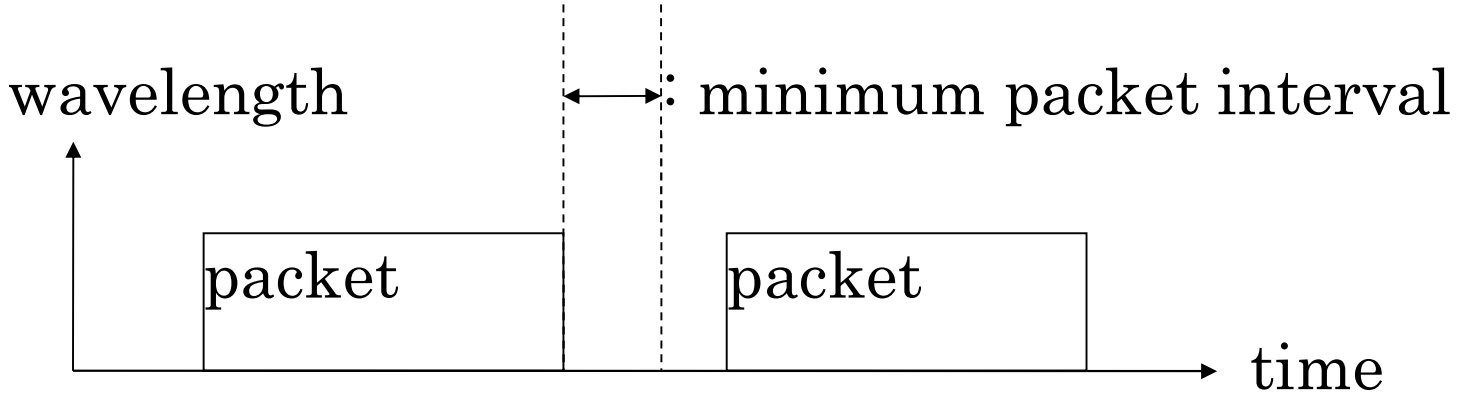




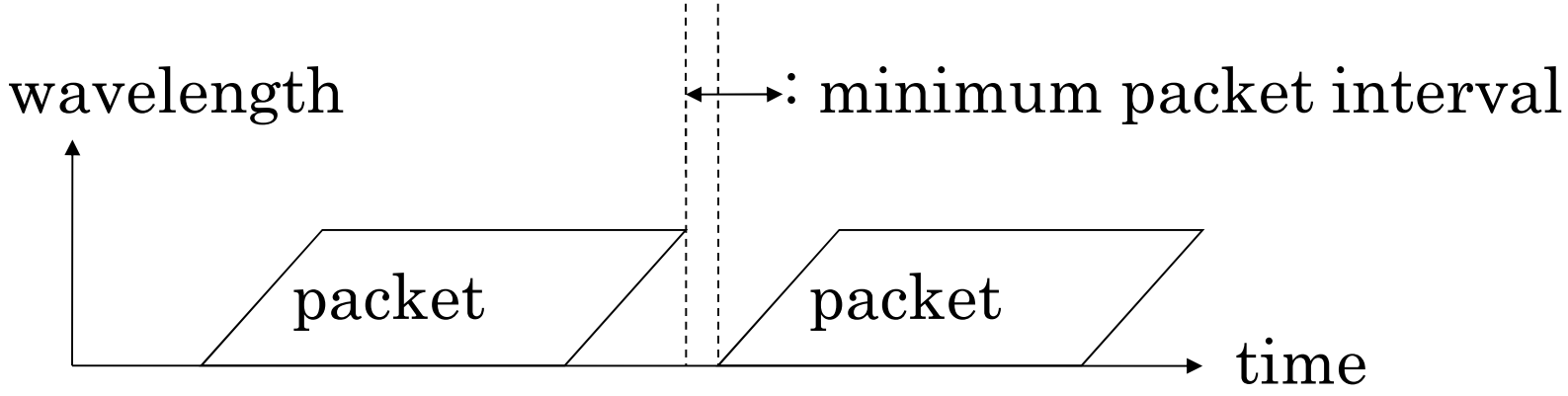
pipelined FDL control

Adverse Effect of Dispersion

- within a wavelength
  - wave form is distorted
    - tens of ps of delay variance is problematic @ 10Gbps
- between wavelengths
  - packet-wise switching may become impossible
    - 1ns of delay variance is proplematic @ 1Tbps
    - ideal dispersion managed fiber with SLA (Super Large Area fiber) and IDF (Inverse Dispersion Fiber) can achieve less than 1ns of delay variance within 2.5Thz for 5000km transmission



a) initial inter-packet gap

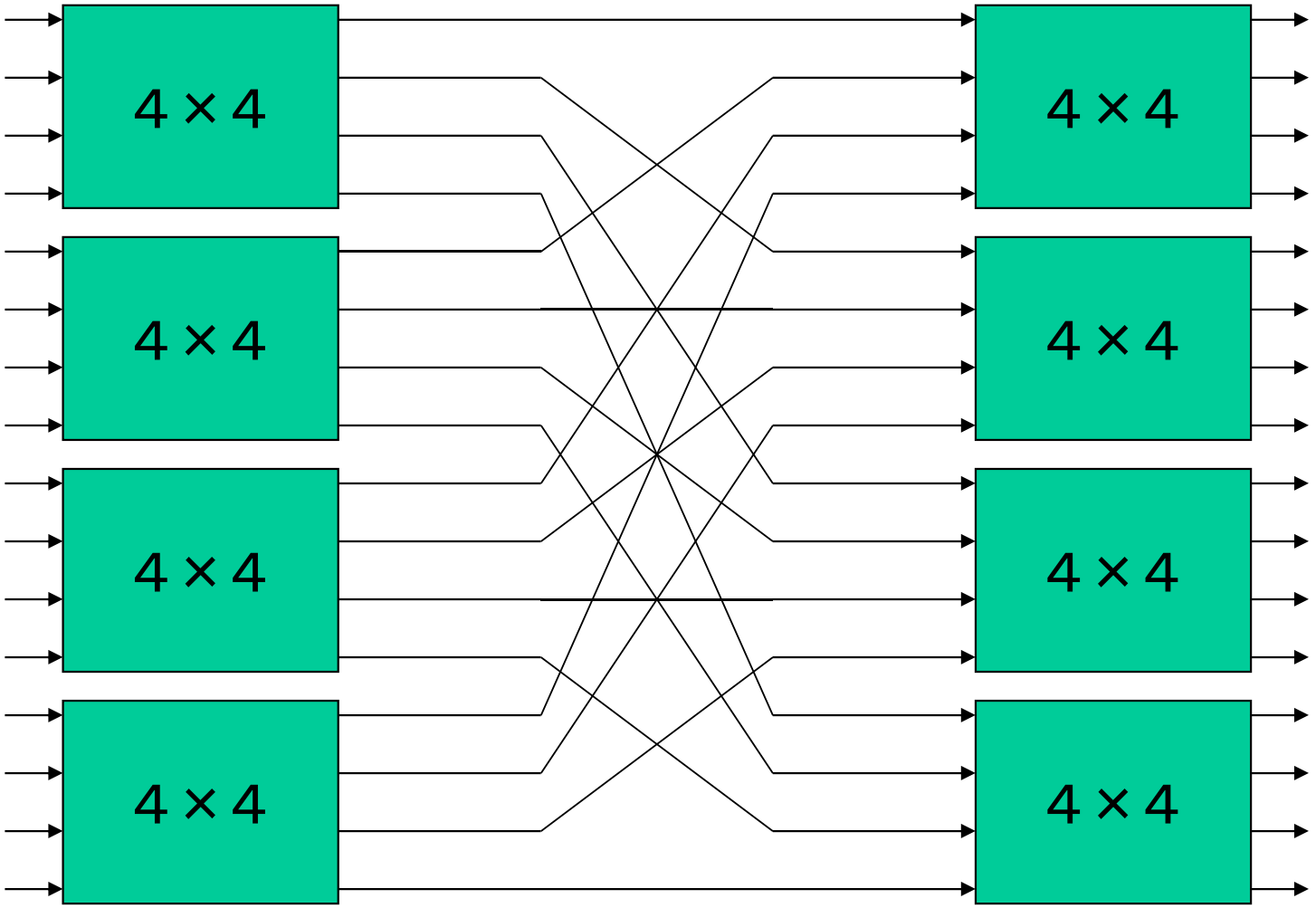


b) inter-packet gap after distortion by dispersion

inter-packet gaps and dispersion

Pbps Routing by Massively Parallel Routers

- Massively Parallel Routing
  - have multiple stages of
    - 1000 1Tbps elementary routers



constructing 16 port switch from 4 port elementary switches

For Supercomputers with Ebps Interconnection Network

- exascale supercomputers
  - should have exascale interconnection network
  - byte-per-FLOP ratio of supercomputers is decreasing
    - TOP500? GRAPH500!
  - should use optical packet switching for high speed low power consuming interconnection network
    - 8 stages of 16k elementary 4 16Tbps port optical packet switches can attain 1Ebps

Optical Switching of Many Wavelength Packets

A Conservative Approach
for an Energy Efficient Exascale Interconnection Network

Masataka Ohta
Department of Computer Science, School of Computing
Tokyo Institute of Technology
mohta@necom830.hpcl.titech.ac.jp

Background

- Exascale Era is coming
- “a long-term goal is to reach the 1mW/Gb/s (i.e., 1pJ/bit) range” [1]
- “~5mW/Gb/s for the power of an optical TX/RX pair” [1], which means EO/OE consumes 5pJ/bit
- Optical switching omitting EO/OE seems to be the MUST

OPS is Conservative but OCS is NOT!

- Data Centers and Super Computers, today, use Packets for Communication
  - We don't want to change our packet based programs or programming styles
- OCS can not Support Certain Communication Pattern such as All to All
  - At 1Ebps bisection bandwidth with 100k nodes and 100k*100k OCS
    - Average bandwidth of a circuit is 10Tbps
      - scarcely no room for wavelength routing (just switch spacially)
    - too fast for most, if not all, applications
      - Elephant (1GB) data moved in 0.8ms (or, with elasticity, faster)
      - The problem of current elephants are that they are so tiny

So, Let's Have OPS

- How?
- Isn't OPS proven to consume a lot of power and be hopeless?
  - [6] R. S. Tucker, "The Role of Optics and Electronics in High-Capacity Routers", J. of Lightwave Technology, V. 24, N. 12, Dec. 2006.
- Not necessarily, as I have been working on OPS since 2005 in a way not considered in [6] and, basically, it is confirmed to works, [2] with pipelined buffer control, [3] with 1.2Tbps DP-DQPSK encoded packets and [4] with 31 FDLs.

Photonics Experts Might Have Thought

- OPS must be hard
- OPS should need most complex photonic circuits
- Designing less complex, but still complex, components for OPS should be the first step to achieve OPS
- Complexity means Much Power Consumption
  - Instead, just make it simple and evaluate power consumption

Packet Experts (Most of US, here at HPSR) Know

- Packet Switches are Boringly Simple
  - Input a packet
  - Analyze header of the packet
  - Forward the packet to an output port
  - If the packet collides with other packets at the output port, buffer, OW, output the packet

Can Packet Experts Still Say:

- Optical Packet Switches are Boringly Simple?
  - Input a packet
  - Analyze header of the packet
  - Forward the packet to an output port
  - If the packet collides with other packets at the output port, buffer, OW, output the packet

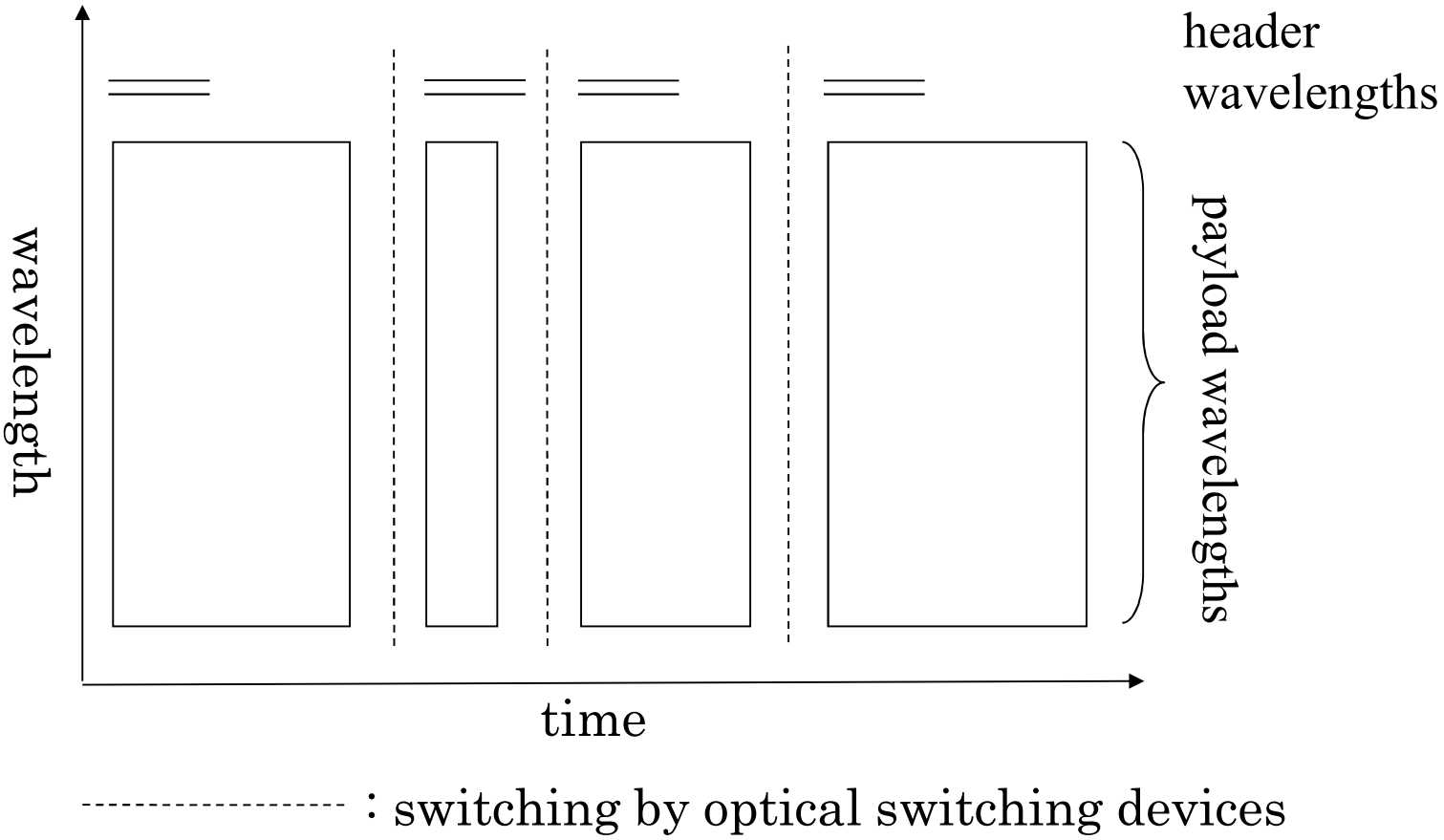
Packet Experts Knows

- Optical Packet Switches are Boringly Simple
  - Input a packet
  - Analyze header of the packet
    - may use usual electric circuits
    - bit-wise operation, but the number of bits is small
  - Forward the packet to an output port
    - must be done optically, but is a packet-wise operation
  - If the packet collides with other packets at the output port, buffer , OW, output the packet
    - buffers are to avoid collisions in time domain
      - FDLs are enough
    - **the last thing to do is to evaluate FDLs as the Buffer**

Evaluating Fiber Delay Lines (1) Aren't They Lengthy?

- Delay for Duration of a Packet needs Length of:
  - (bits of a packet)*(speed of light)/(**bps of fibers**)
- In 2005, assuming Ethernet and 1Tbps
  - (12kbits)*(2*10^8m/s)/(1Tbps)=2.4m
  - Short Enough! Slow Light? Why bother?
- Today, assuming 9kB packets and 16Tbps (40GBaud DP-QPSK with 100 Wavelengths)
  - (72kbits)*(2*10^8m/s)/(16Tbps)=0.9m
- How can we have 1 or 16 Tbps packets?
  - Obviously, with many wavelengths! (and polarization)

Many Wavelength Packets



Evaluating Fiber Delay Lines (2) How Many Delay Lines Needed?

- Packet drop probability should be small
  - but, how small should it be? 0? NOT AT ALL!
  - small enough not to degrade TCP performance
  - old theory requires amount of buffer capacity of
    - (bps of a link)*(round trip time of the TCP)
      - round trip time within LANs is still small
    - the theory applicable when the number of TCP is small
  - new theory requires buffer for tens of packets or less
    - the theory applicable when the number of TCP is large (traffic is Poisson) and small amount of bandwidth is sacrificed
- FDLs, lengths of which increases with geometric progression of common ratio 2, seems to be best

An Example of TCP Performance

- Expected TCP bandwidth is MSS/RTT/sqrt(p) [11]
- Assuming MSS (Maximum Segment Size)=8960B, RTT (in this case including buffering delay)=10μs (delay by 1km of FDLs in each direction) and p (packet drop probability) = 0.15%, it is 185Gbps.

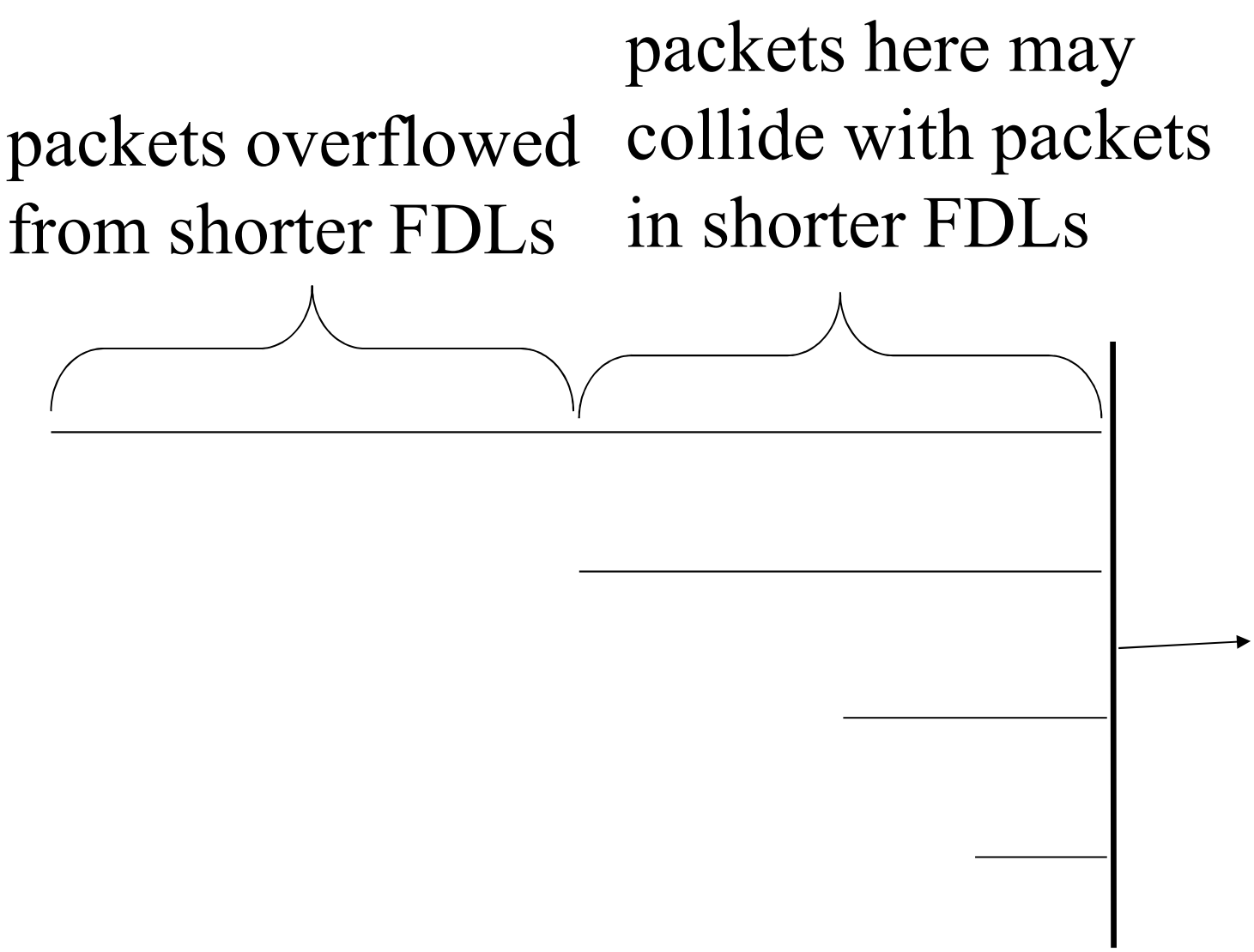
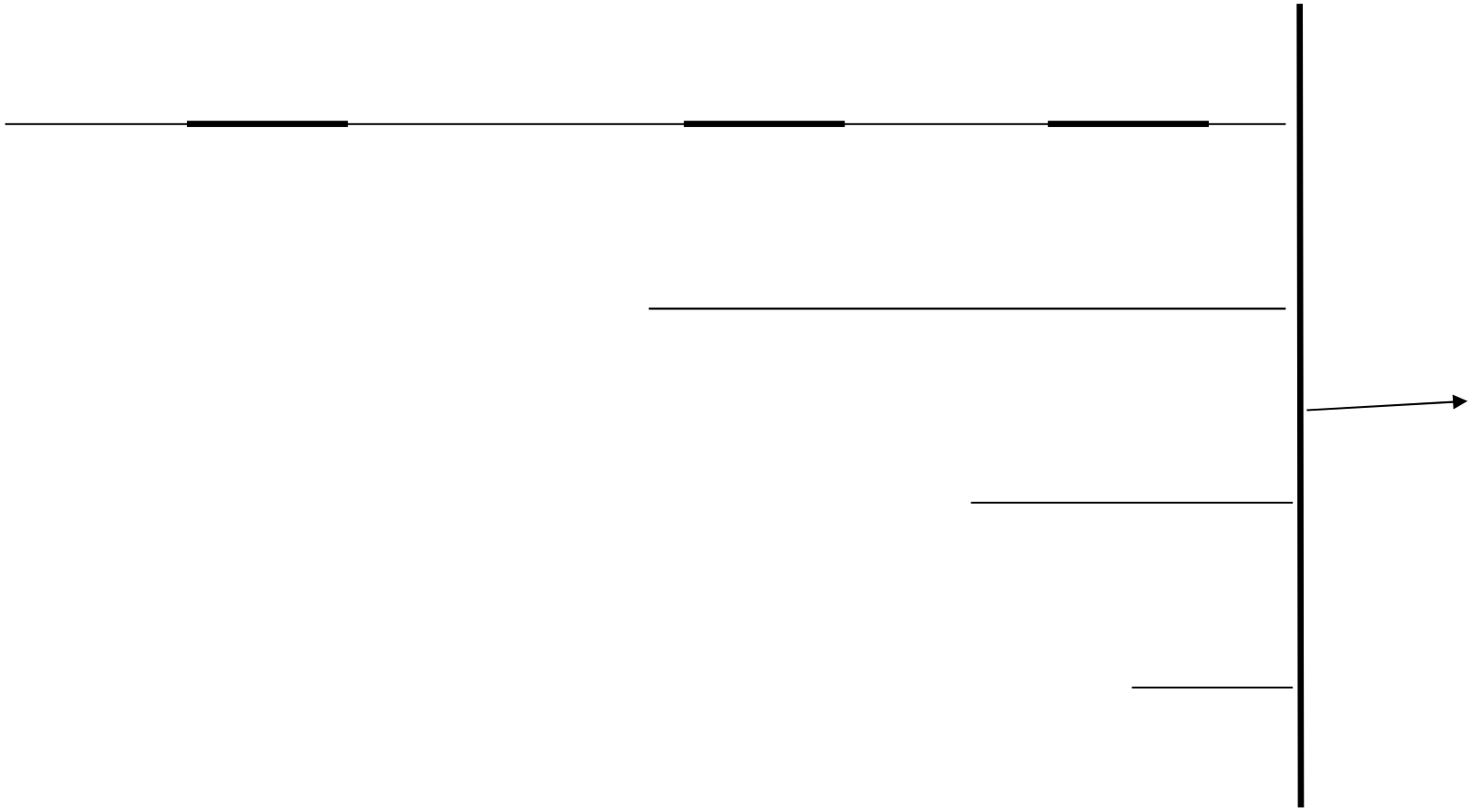


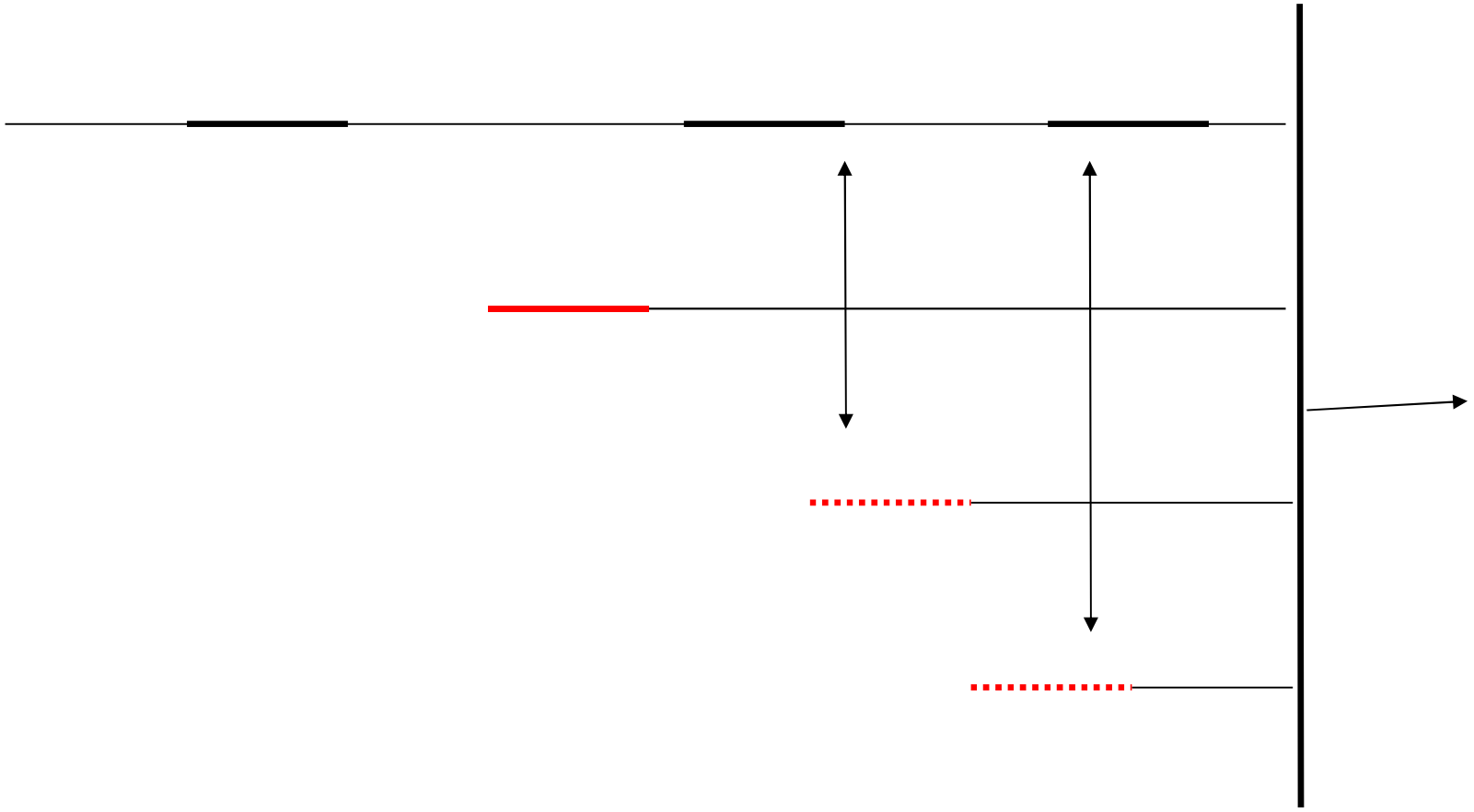
Fig. 5. FDLs with Lengths in Geometric Progression with Common Ratio of 2

Buffer Control (1)



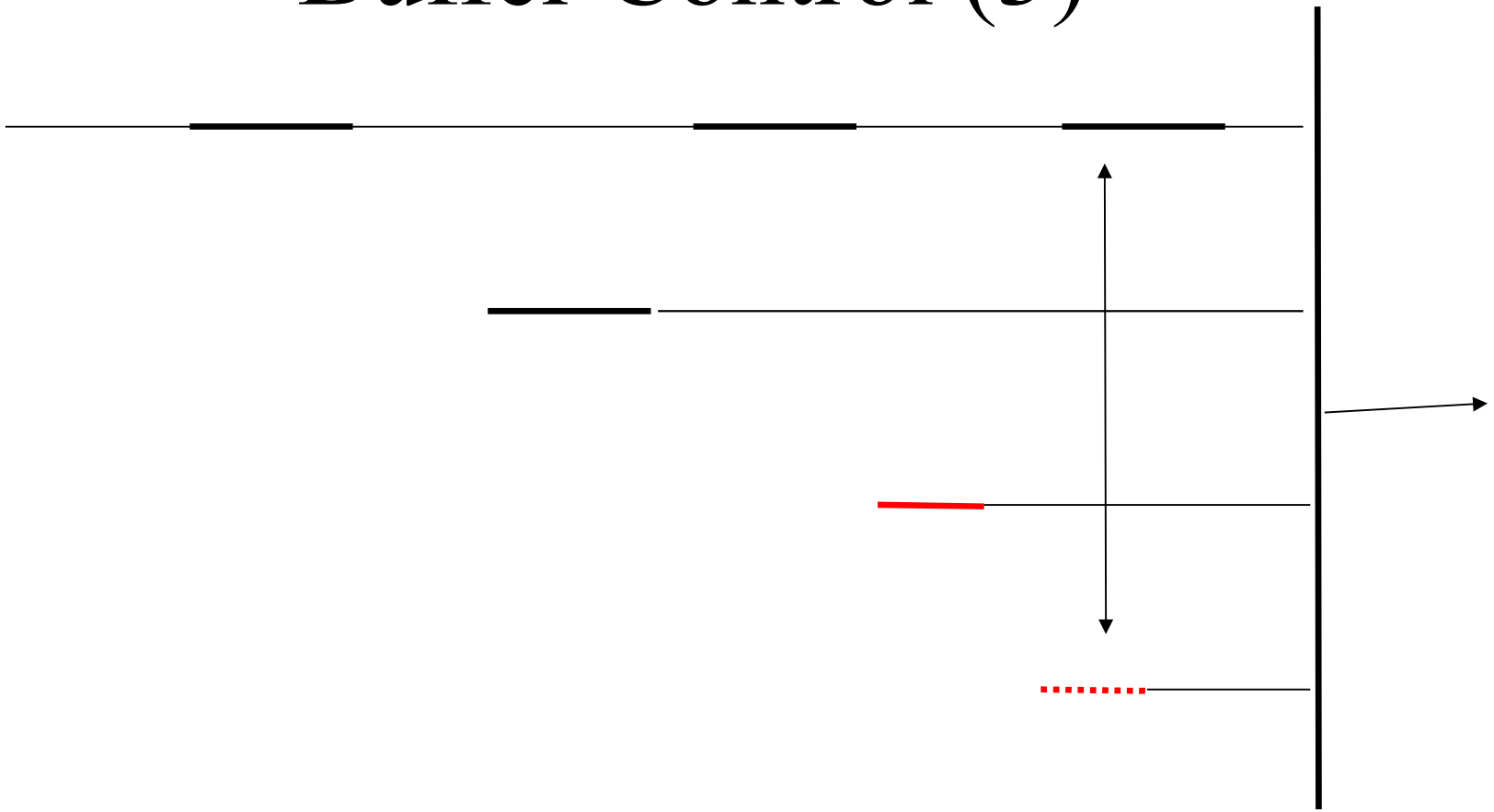
a) initial packet distribution

Buffer Control (2)



b) new packet put to the third shortest FDL

Buffer Control (3)



c) another new packet (shorter) put to the second shortest FDL

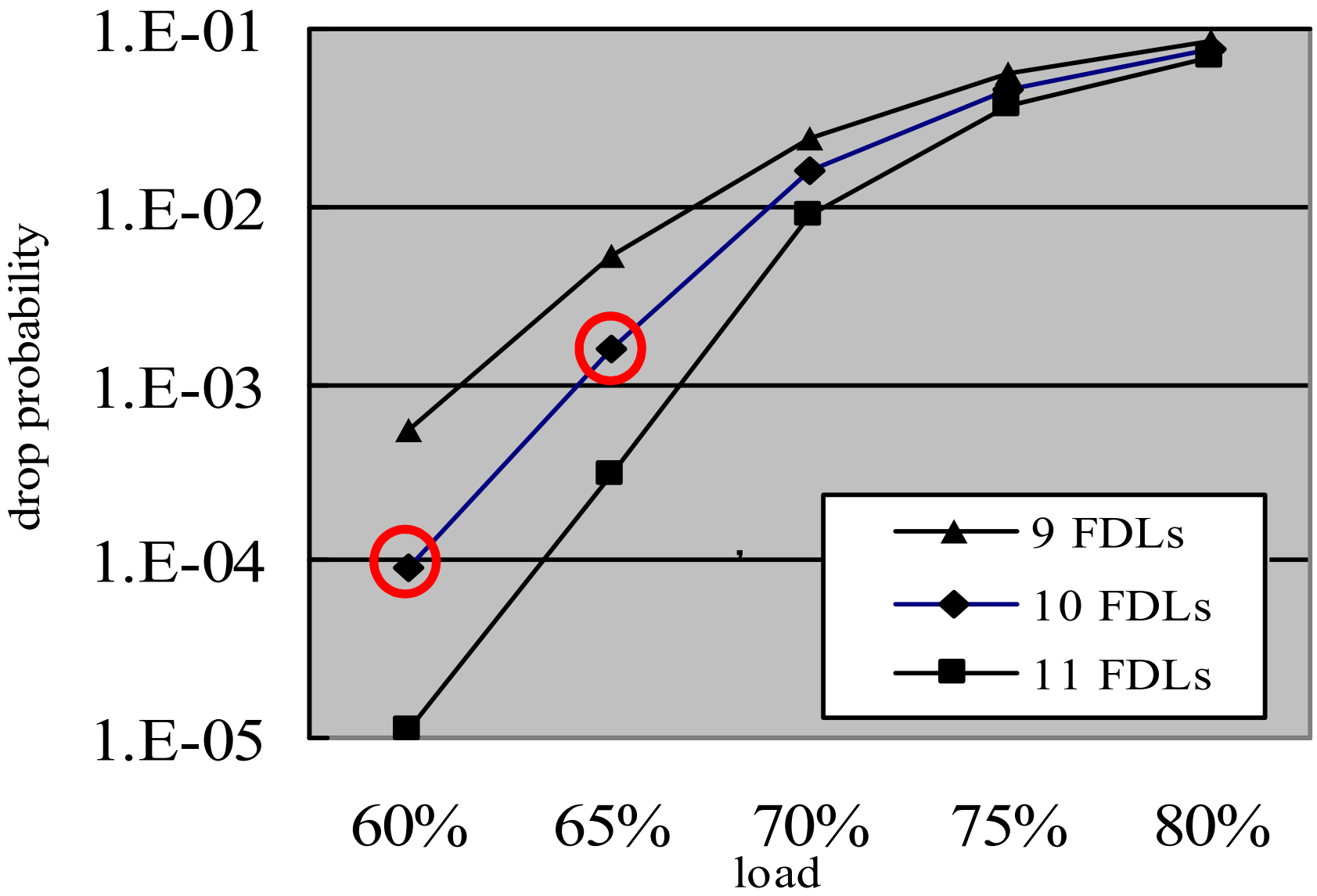
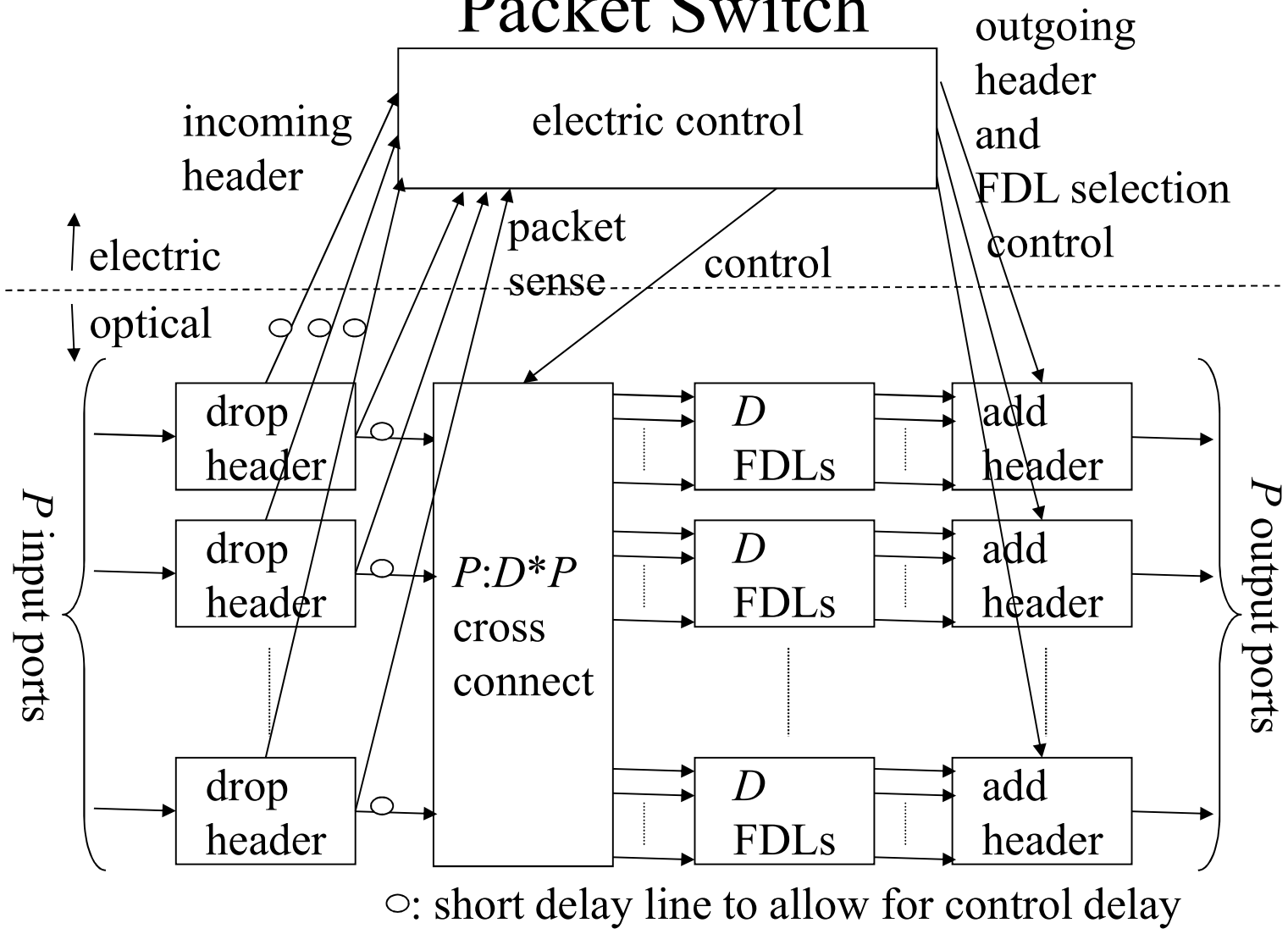
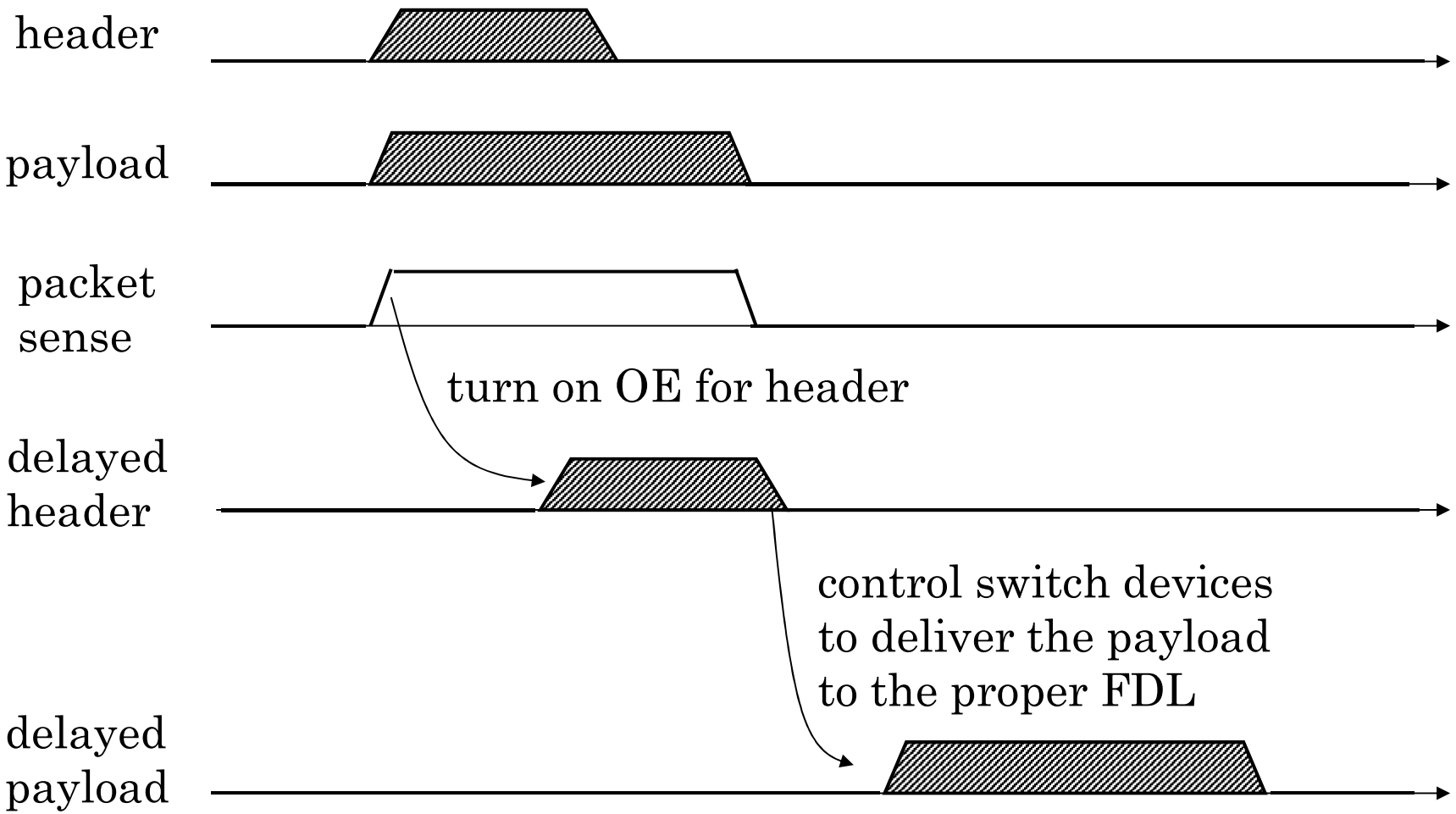


Fig. 6. Packet Drop Probability of FDL Buffers

A Micro Architecture of A Proposed Optical Packet Switch



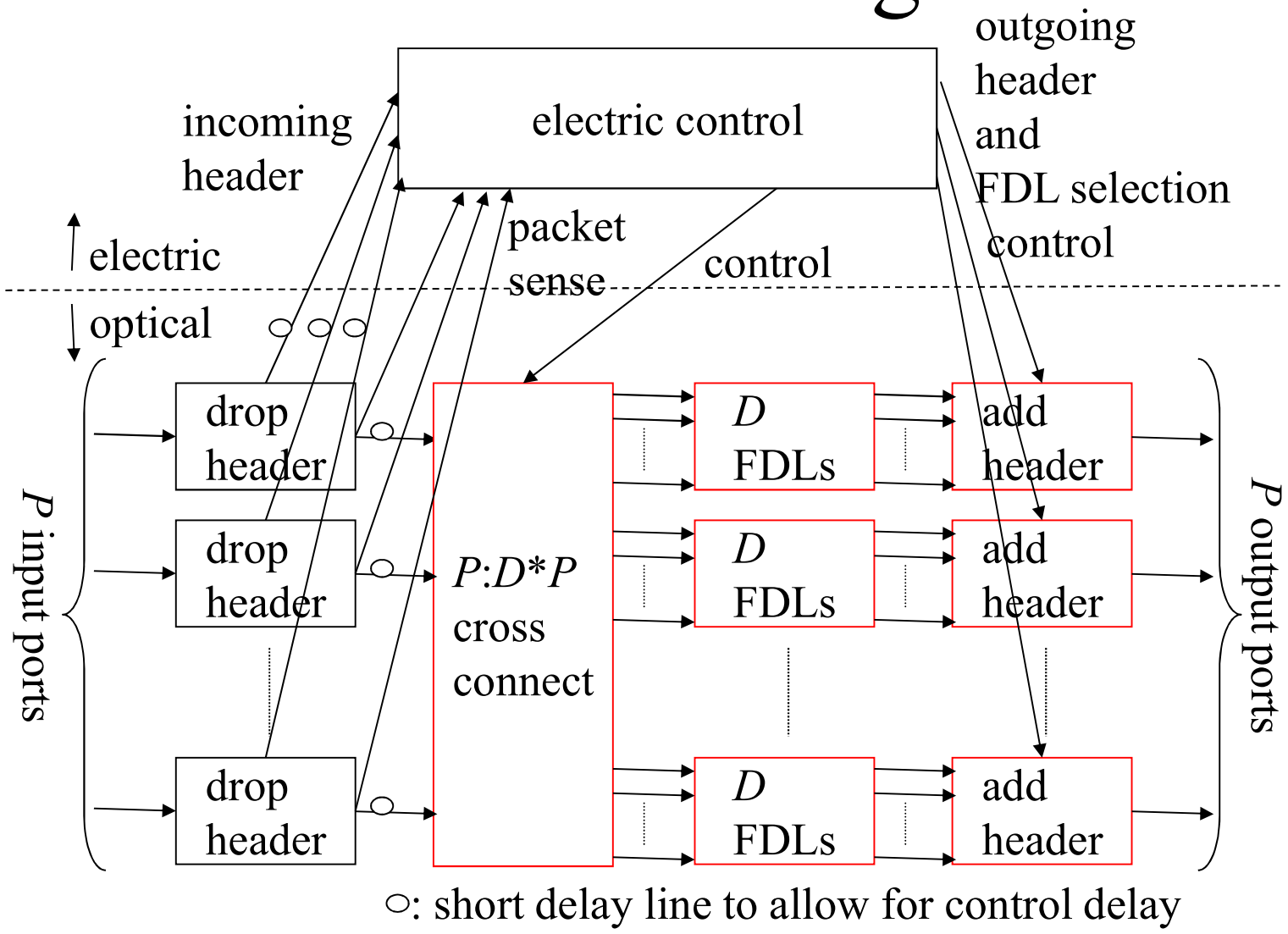
Relationships between Signals



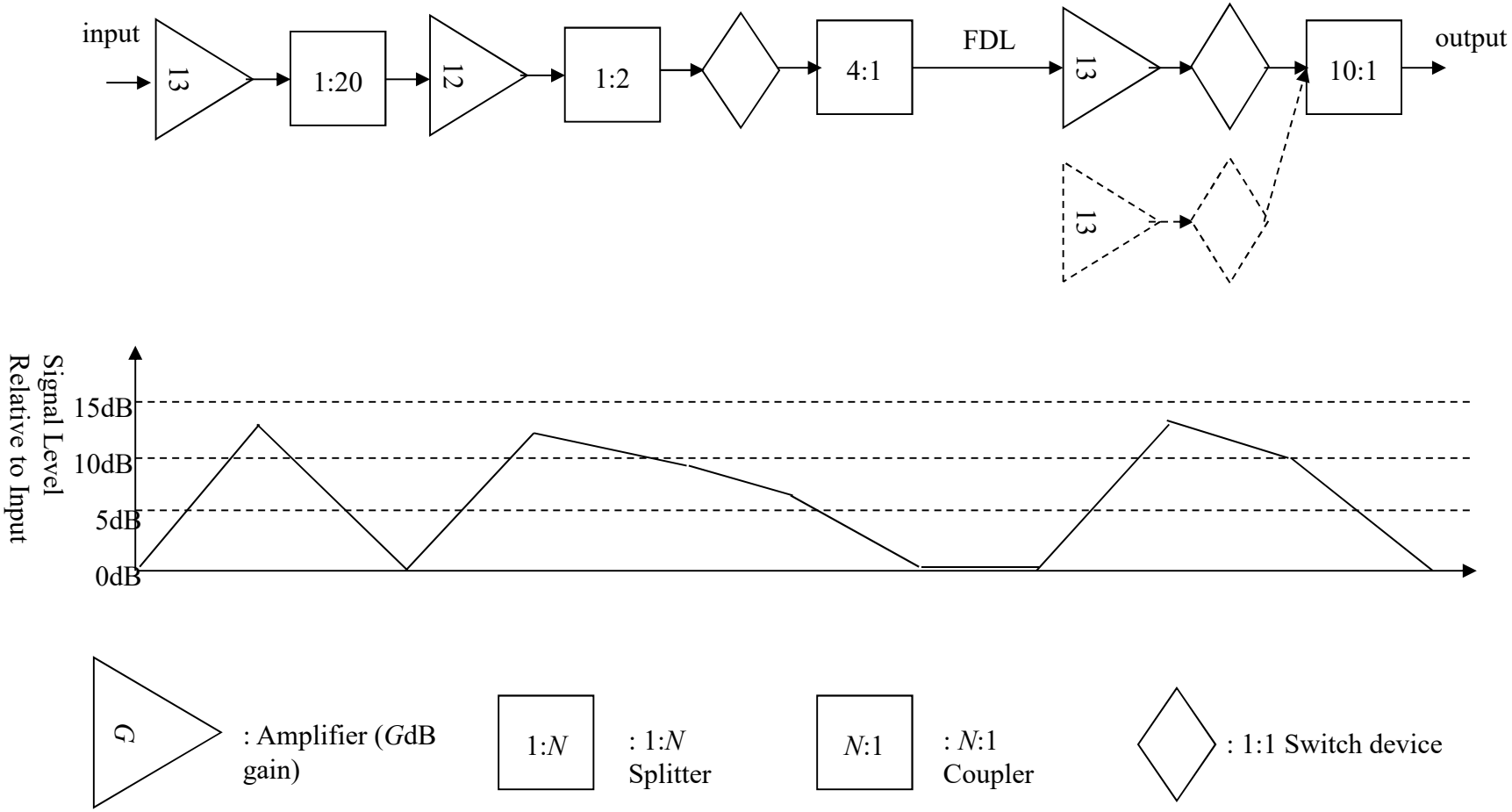
Power Consumed by Optical Packet Switches

- Optical Packet Switches are not Power Consuming
  - Input a packet
  - Analyze header of the packet
    - bit-wise operation, but the number of bits is small
      - negligible power consumed
  - Forward the packet to an output port
    - must be done optically, but is a packet-wise operation
      - negligible power consumed by **capacitive** optical switching devices without termination registers
    - **most power is consumed by optical losses here**
  - If the packet collides with other packets at the output port, buffer
    - **and here**

Power Consuming Parts



Level Diagram within a 4 Port Optical Switch with 10 FDLs



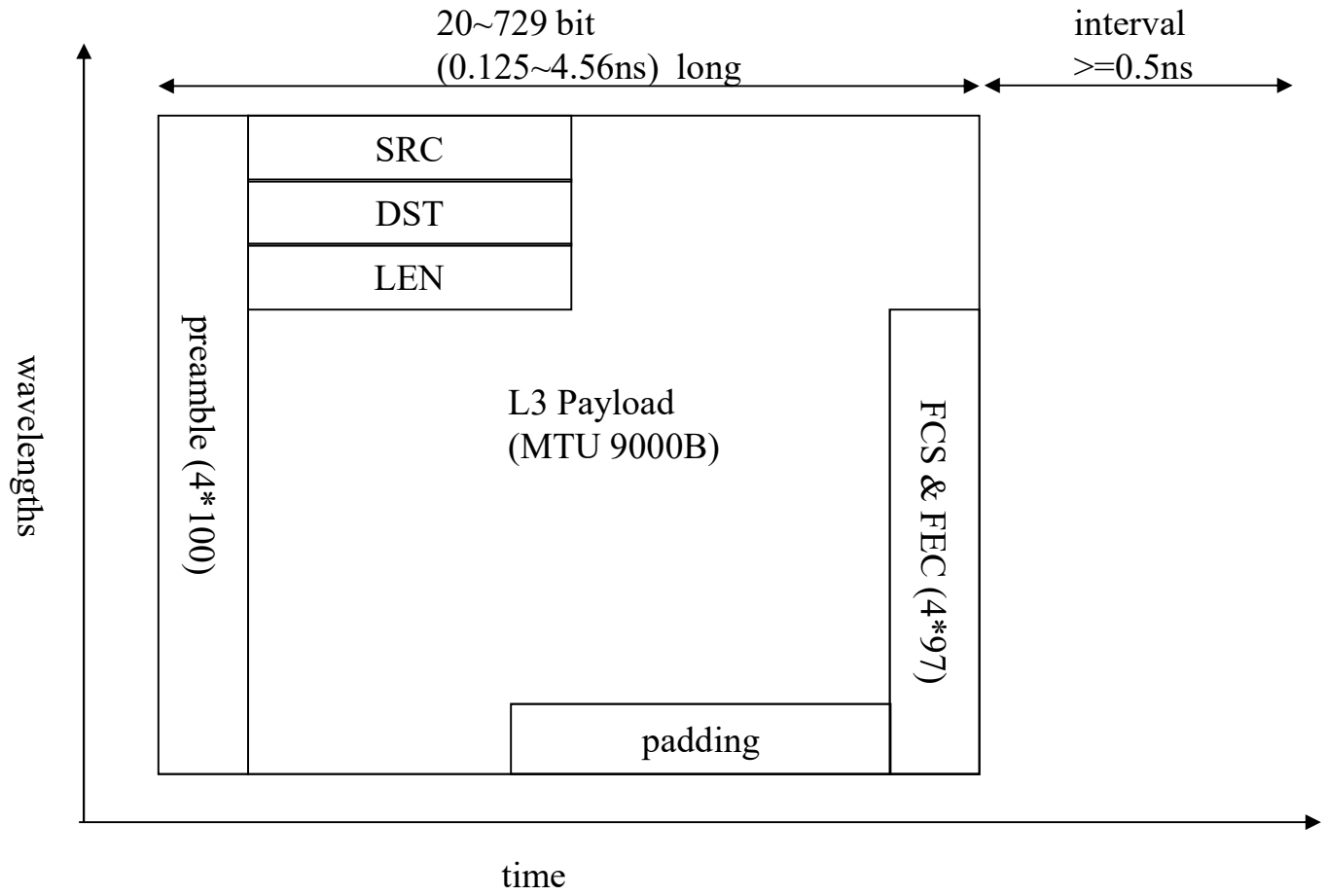
Estimating Power Consumption of An Optical Packet Switch

- Depends on Signal Energy
  - (Signal Energy)=SNR*(Noise Energy)
  - (Noise Energy)=(Photon Energy)*(# of Noise Photons)
  - (# of Noise Photons)=($10^{NF(dB)/10}$-1)*(# of EDFA Stages)
  - (# of EDFA Stages)=3*(# of Optical Switch Stages)
- With SNR=10dB, NF=3.98(!4.77)dB and 64K*64K Butterfly (8 stages of 4 port switches)
  - (Signal Energy)=$4.62*10^{-17}$J/bit
- Power Consumed by 1 14dB, 20 13dB and 10 14dB EDFAs (30% Efficiency) is $9.9*10^{-14}$J/bit

Estimating Power Consumption of Interconnection Network

- Minimum Packet Length: 0.125ns
- Minimum Packet Interval: 0.5ns
- Packetization Overhead: 0.06ns
- Load: 60%
- Traffic: TCP with two 9kB Data and one ACK
- Energy Consumed by 8 stage butterfly
  - 1.49pJ/bit @ effective bisection bandwidth of 0.53Ebps
- Energy Consumed by 15 stage Benes
  - 5.3pJ/bit @ effective bisection bandwidth of 0.53Ebps

Payload Format



Estimated Volume Occupied by a Proposed Optical Packet Switch

- A 4 port elementary switch consists from:
  - 4 1:20 and 80 1:2 splitters
  - 40 4:1 and 4 10:1 couplers
  - 200 1:1 switch devices

  Assume photonic integration with control circuits except for 1:20 splitters
  - 124 EDFAs (12.4km EDF assuming each have 100m)
    - Assume each EDFA needs additional $10cm^3$ (more integration?)
  - 40 FDLs (total length of 3.7km)
- 1.2km of fiber can be coiled in a compact bobbin (40mm diameter and 20mm height, $25.1cm^3$) [12]
- With 100% overhead, total volume is $3250cm^3$
  - smaller than a cube with 15cm edges
  - a lack storing 16 nodes stores 32 switches (butterfly)

Conclusions

- Many wavelength packets enables 16Tbps packets
  - with 100 wavelengths and 40GBaud DP-QPSK
  - 9kB@16Tbps is 4.5ns long (delay by 0.9m FDL)
  - At 60% load, an optical buffer with 10 FDLs have:
    - packet drop probability of 0.0089%
- An Exascale interconnection network for 64K nodes with 4 16Tbps port optical packet switches
  - estimated to consume 1.49pJ/bit (butterfly topology) and 5.3pJ/bit (Benes topology)
    - with effective bisection bandwidth of 0.53Ebps
  - the volume of such a switch is estimated to be 3250cm^3

Related Paper in the Workshop (this Afternoon)

- M. Ohta, “Optimal Radix for High Speed Optical Packet Switching”
  - optical packet switches in an interconnection network should have low radix such as 2, 3 or 4 to minimize power consumption of the network

Optimal Radix for High Speed Optical Packet Switching

Masataka Ohta
Department of Computer Science, School of Computing
Tokyo Institute of Technology
mohta@necom830.hpcl.titech.ac.jp

Conclusions of [1] (Presented in this Morning) assume Low Radix

- Many wavelength packets enables 16Tbps packets
  - with 100 wavelengths and 40GBaud DP-QPSK
  - 9kB@16Tbps is 4.5ns long (delay by 0.9m FDL)
  - At 60% load, an optical buffer with 10 FDLs have:
    - packet drop probability of 0.0089%
- An Exascale interconnection network for 64K nodes with 4 16Tbps port optical packet switches
  - estimated to consume 1.49pJ/bit (butterfly topology) and 5.3pJ/bit (Benes topology)
    - with effective bisection bandwidth of 0.53Ebps
  - the volume of such a switch is estimated to be 3250cm^3

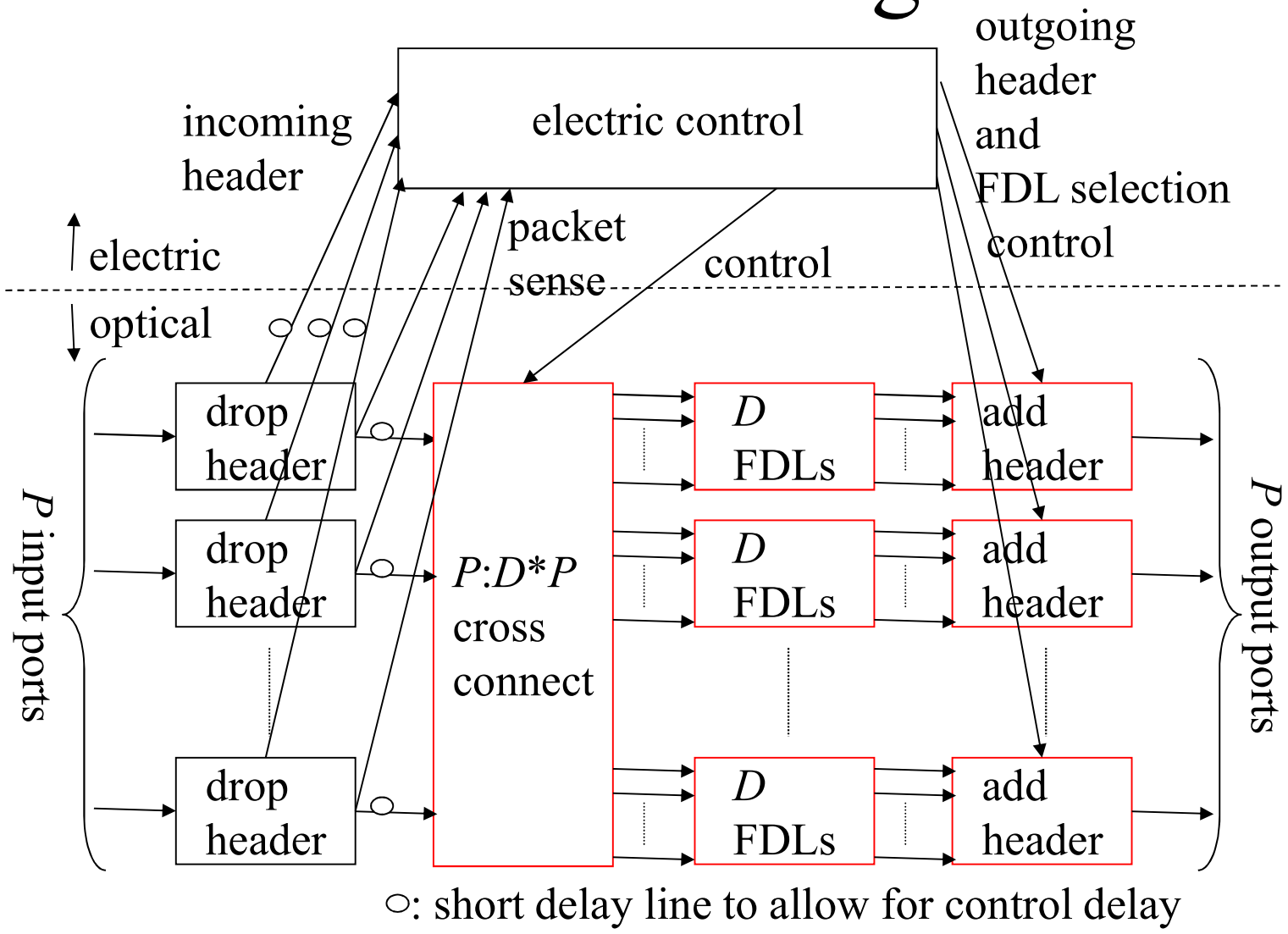
Isn't High Radix Better?

- Yes, if we want to minimize delay with a single chip switch with limited IO bandwidth of the chip
  - optimal radices are 40 and 127 assuming technology available in years 2003 and 2010, correspondingly
- Yes, if we want to minimize power consumed by EO/OE
- However, if it is "Optimal Radix for High Speed Optical Packet Switching", not necessarily, because
  - "High Speed" makes delay negligible
  - "Optical Packet Switching" means there is no EO/OE
- So, what is the optimal radix to minimize power consumption of a butterfly network?

Power Consumed by Optical Packet Switches

- Optical Packet Switches are not power consuming
  - Input a packet
  - Analyze header of the packet
    - bit-wise operation, but the number of bits is small
      - negligible power consumed
  - Forward the packet to an output port
    - must be done optically, but is a packet-wise operation
      - negligible power consumed by **capacitive** optical switching devices without termination registers
    - **most power is consumed by optical losses here**
  - If the packet collides with other packets at the output port, buffer
    - **and here**

Power Consuming Parts



Power Consumption of An Optical Packet Switch

- Depends on Signal Attenuation
  - with broadcast & select with P ports and D FDLs
    - splitting signal to $P*D$ FDLs: $P*D$ attenuation
    - merging signal from P ports and D FDLs: $P*D$ attenuation
  - energy lost is: $(P*D)^2$-1 (approximately $(P*D)^2$)
- Proportional to Signal Energy
  - (Signal Energy)=SNR*(Noise Energy)
  - (Noise Energy)=(Photon Energy)*(# of Noise Photons)
  - (# of Noise Photons) $\propto$ (# of Optical Switch Stages)
  - thus, proportional to # of Optical Switch Stages
    - with butterfly topology for N nodes, it is $\log_P N$
- Proportional to # of Switch Ports: $N*\log_P N$

The Optimal Radix

- As D and N are Constants, the Optimal Radix P Minimizes
  - $(P*D)^2*\log_P N*N*\log_P N \propto (P/\ln P)^2$
  - or, just $P/\ln P$ and $d/dP(P/\ln P)=(\ln P-1)/(\ln P)^2$
- Thus, the optimal radix is e=2.71828..., or, in integer, 3
  - 12% more power is consumed with radix 2 or 4, not bad

Wrap-up

- Tbps almost all optical routers
  - can be constructed with current technology
- massively parallel construction can achieve Peta or Exa bps speed
- not so much demand
  - hopefully except for data centers and supercomputers