インターネットインフラ特論
１２．ペタ・エクサビットルータ

太田昌孝

mohta@necom830.hpcl.titech.ac.jp

ftp://chacha.hpcl.titech.ac.jp/infra12j.ppt

超高速ルータはなぜ必用

- 速度
  - １００Mbpsを５万人が使うと5Tbps
  - 単体電気ルータは数十Gbps程度

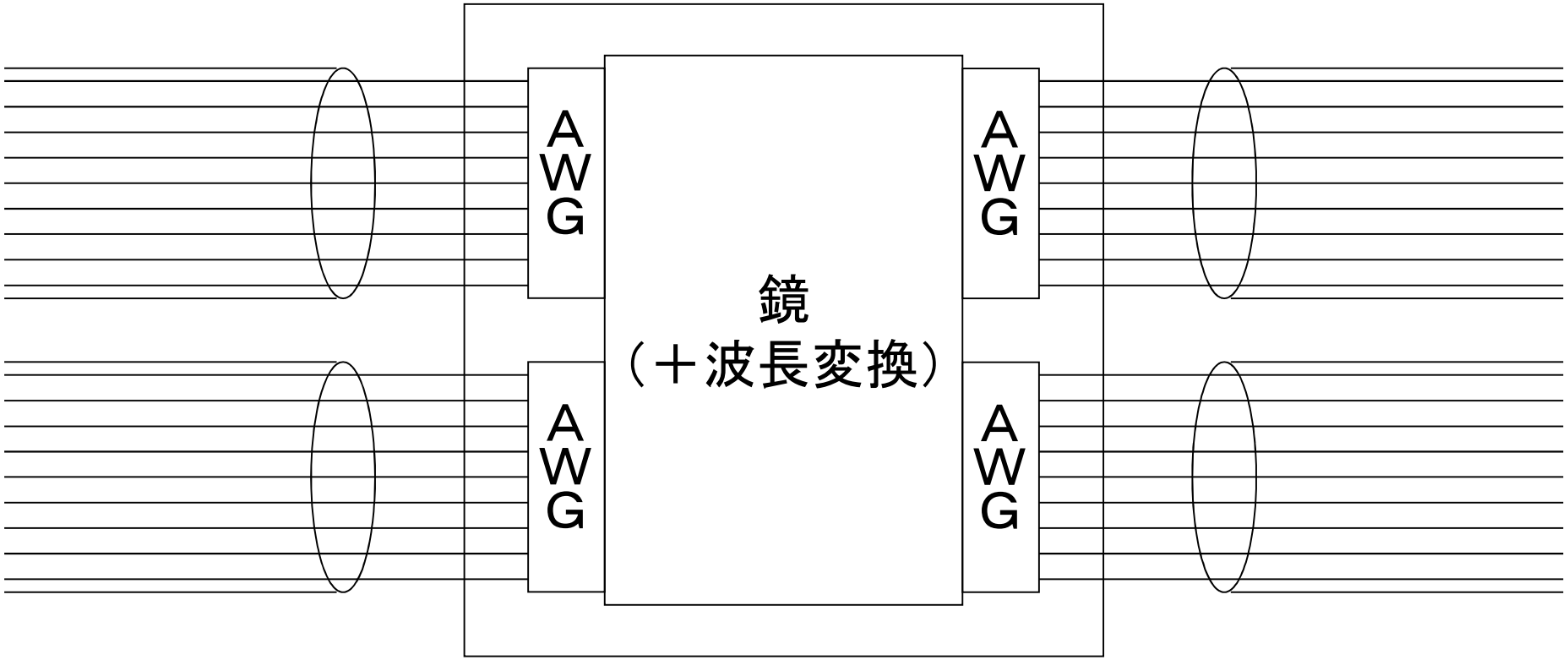
光と電気の棲み分け

- 光
  - ほとんど干渉なし（非線形性はほぼ無し）
    - 伝送に向くが、論理演算はほぼ無理
  - 超広帯域（特に速いわけではない）
- 電気
  - 干渉が大きい
    - 伝送には不向き
    - 演算、制御に向く

光ファイバ遅延線とスローライト

- 光バッファは遅延線により実現可能だが
  - 一般に、長いファイバが必要（10Gbps1500Bの遅延で、240m）
- 高Qの光共振器を並べたスローライトでは
  - 光がゆっくりとしか変化しない
  - より、短い長さで、バッファ可能？
  - 光がゆっくりとしか変化しないと、bpsが下がるので、むしろ長い距離が必要

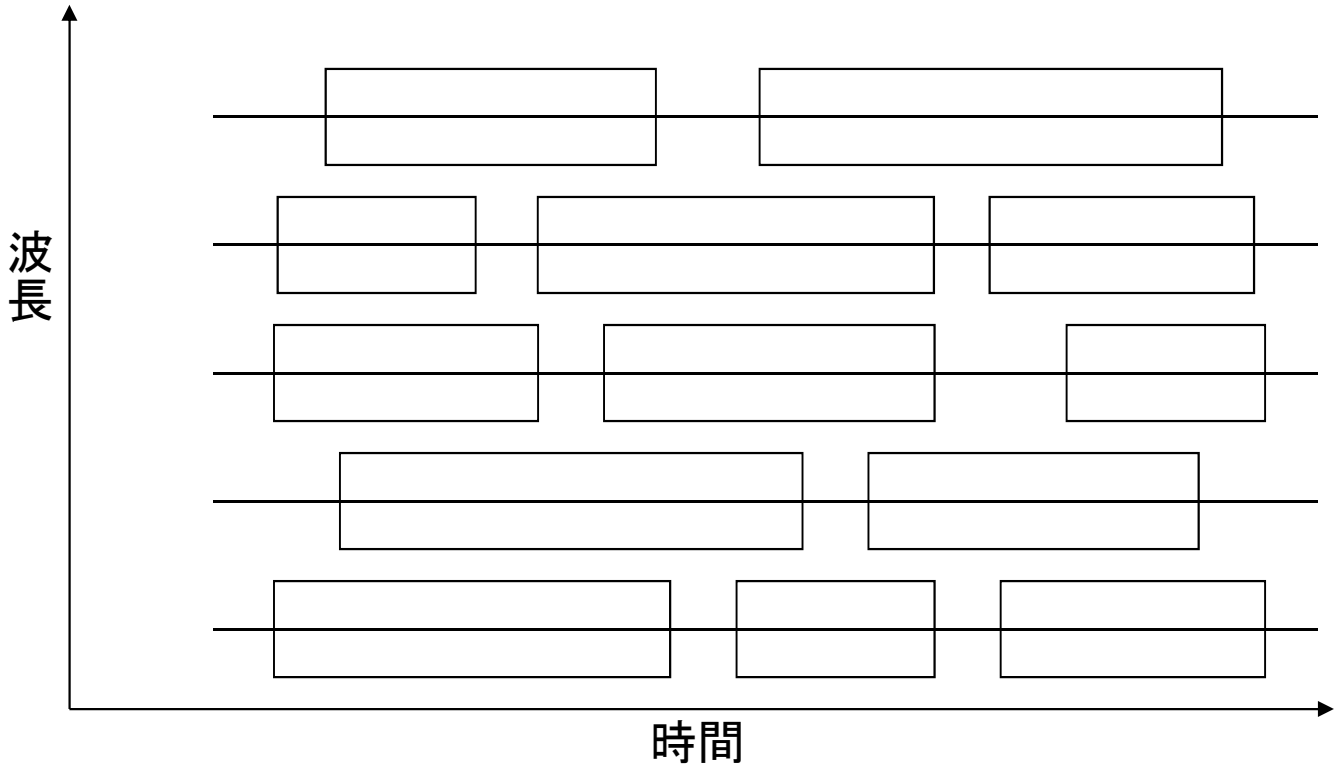
波長ルータ



波長ルーティングは何を間違えているのか？

- せっかくの光のテラビット級の伝送速度を
  - 10Gbps＊100波長程度にこまぎれで処理
  - 機器の規模（電力）は少なくとも波長数に比例
- 一方、光伝送では
  - 光の全帯域を一台のEDFA（光アンプ）で増幅
  - WDM伝送大成功の要因
- 波長多重は伝送技術、交換で使うな
  - 交換は全波長を一括で！

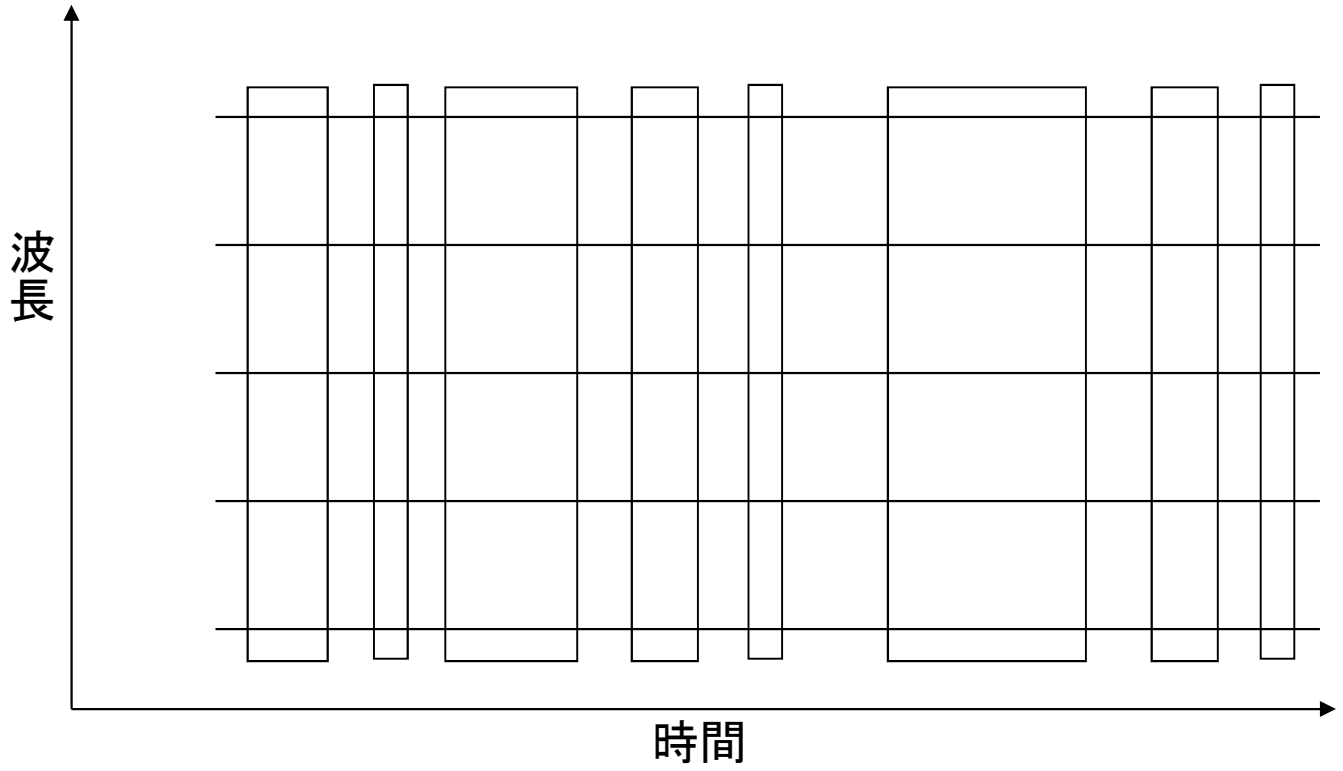
IP over WDMと
WDMによるパケット多重



IP uber Alles

- パケット多重こそすべて！！
  - 波長多重で利用できる全帯域を個々のパケット伝送に利用すべき
- 超高速（100ps以下）光スイッチの出現
  - データパスはこれでOK
  - 制御は？
- Almost All-Opticalなら簡単
  - たかが1Tbpsなら、制御は電気で楽勝

IP over WDMと
WDMによらないパケット多重



超高速光スイッチ

EOSPACE

Exceptionally-Low-Loss $LiNbO_3$ Optical Devices & ICs

Technology Originally Developed for High-Performance Aerospace Systems

8711–148th AVE N.E., Redmond, WA 98052 Tel: 425-869-8673 info@eospace.com

Custom High-Speed **Lithium Niobate** Electro-optic Switches

λ = 1550nm; Please call for other λ : 2000+, 1700, 1300, 1060, 980, 850, 700nm

Ultra-High-Speed (sub-nanoseconds) **1x2, 2x2 Optical Switches/Modulators**

(wideband traveling-wave electrode structure with internal 50-Ω termination)

2.56" x 0.35" x 0.195" (65 x 8.9 x 4.95 mm³)

1x2, 2x1, 2x2 Ultra-high-speed Switch/Modulator

- *Single polarization (SP), separate DC bias port*
- *>10GHz (>18GHz option), T_{switch}<<100ps, $V\pi$ ~ 5V*
- *Insertion loss < 4.0dB (< 3.0dB option)*

±2．5V@50Ωで、消費電力は0．125W

光と電気の速度

- 電気制御の光スイッチ
  - 100psで切り替え
- 1Tbpsで500（1500）Bパケットは
  - 4（12）ns
- いまどきのプロセッサのクロック
  ＞1GHz（クロック周期＜1ns）
- 実効速度数百Gbpsのルータは
  - 電気制御で余裕で実現可能

光パケットバッファは？

- 1Tbpsで500（1500）Bパケットは
  - 4（12）ns
  - 光ファイバ長にして0．8m（2．5m）
    - 10光子ビットあたり損失0．037kT（T＝300K）
  - 10Gbpsだと、光ファイバ長は100倍必要
    - かなり非現実的な長さに
    - 1Tbpsの性能には100並列が必要
- 1000パケット分でも2．5km
  - ≦4kmで15cm＊15cm＊4cmの機器あり

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.

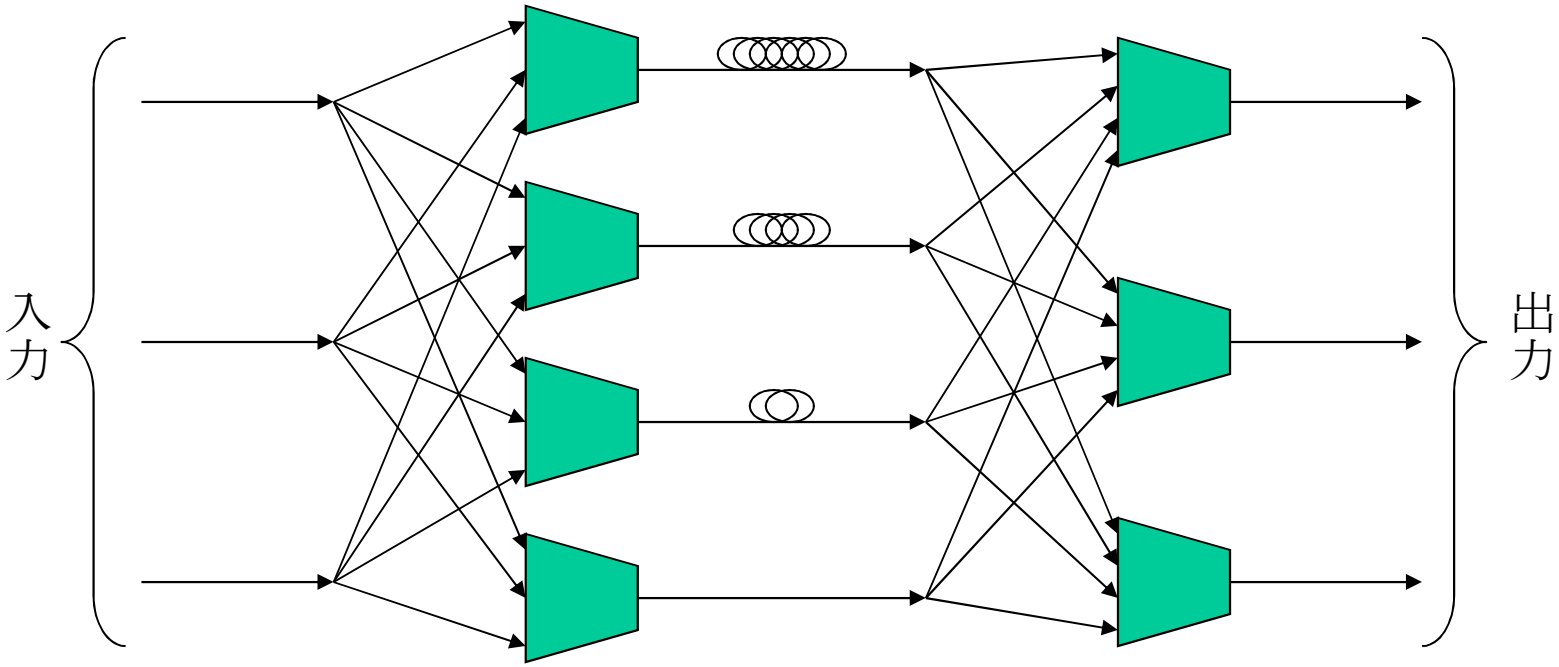
Specifications:

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

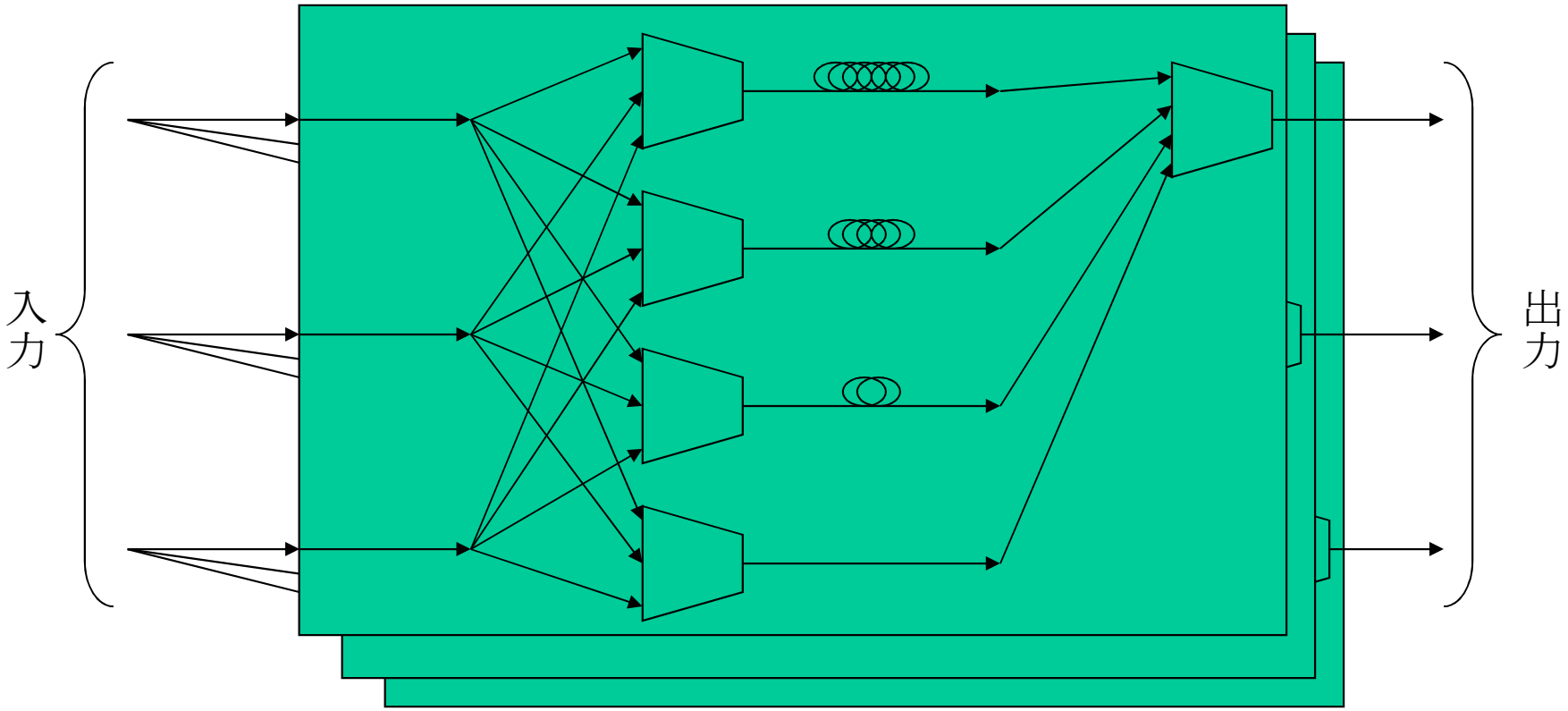
共有バッファ方式



NポートでKN本の遅延線を利用する場合の2:1光スイッチの数：

$2KN^2-(K+1)N$

個別バッファ方式



Nポートで各K本の遅延線を利用する場合の2:1光スイッチの数：

K＊N^2－N

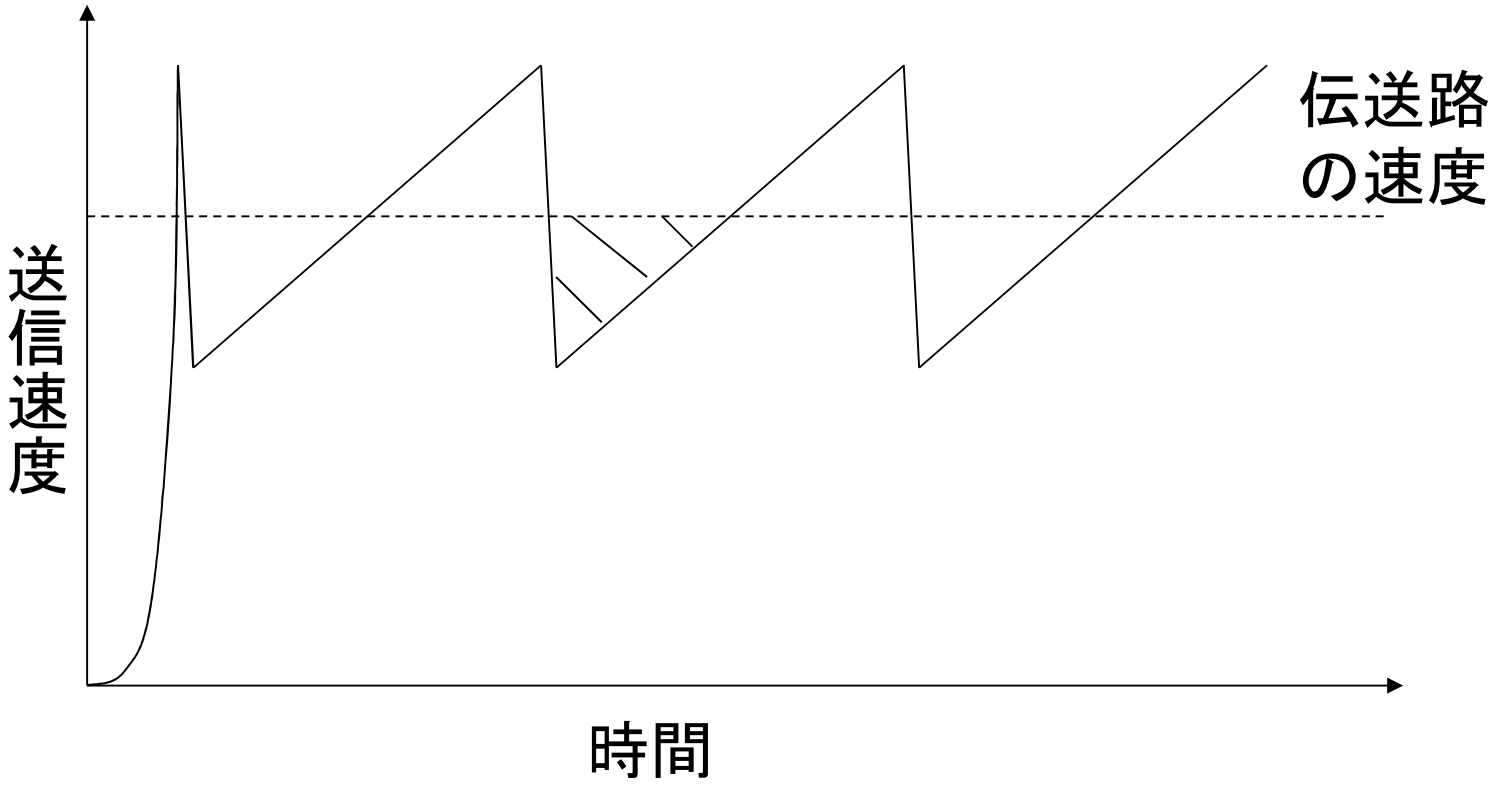
インターネットバックボーンのトラフィック

- ポワソン
  - 個々のTCPの変動は平均化して見えない
- 平均パケット長
  - 数百バイト
- TCPのフロー数は数万程度

TCPとルータのバッファ

- CAによりTCPの速度は鋸歯状に変動
- バッファしないと回線速度を使い切れない
  - （伝送遅延）＊（伝送速度）だけのバッファが必要
- 一部幹線では巨大なバッファが必要？
  - 幹線は速い
  - 幹線は長い

TCPトラフィックの変動の様子



TCPと幹線ルータのバッファ

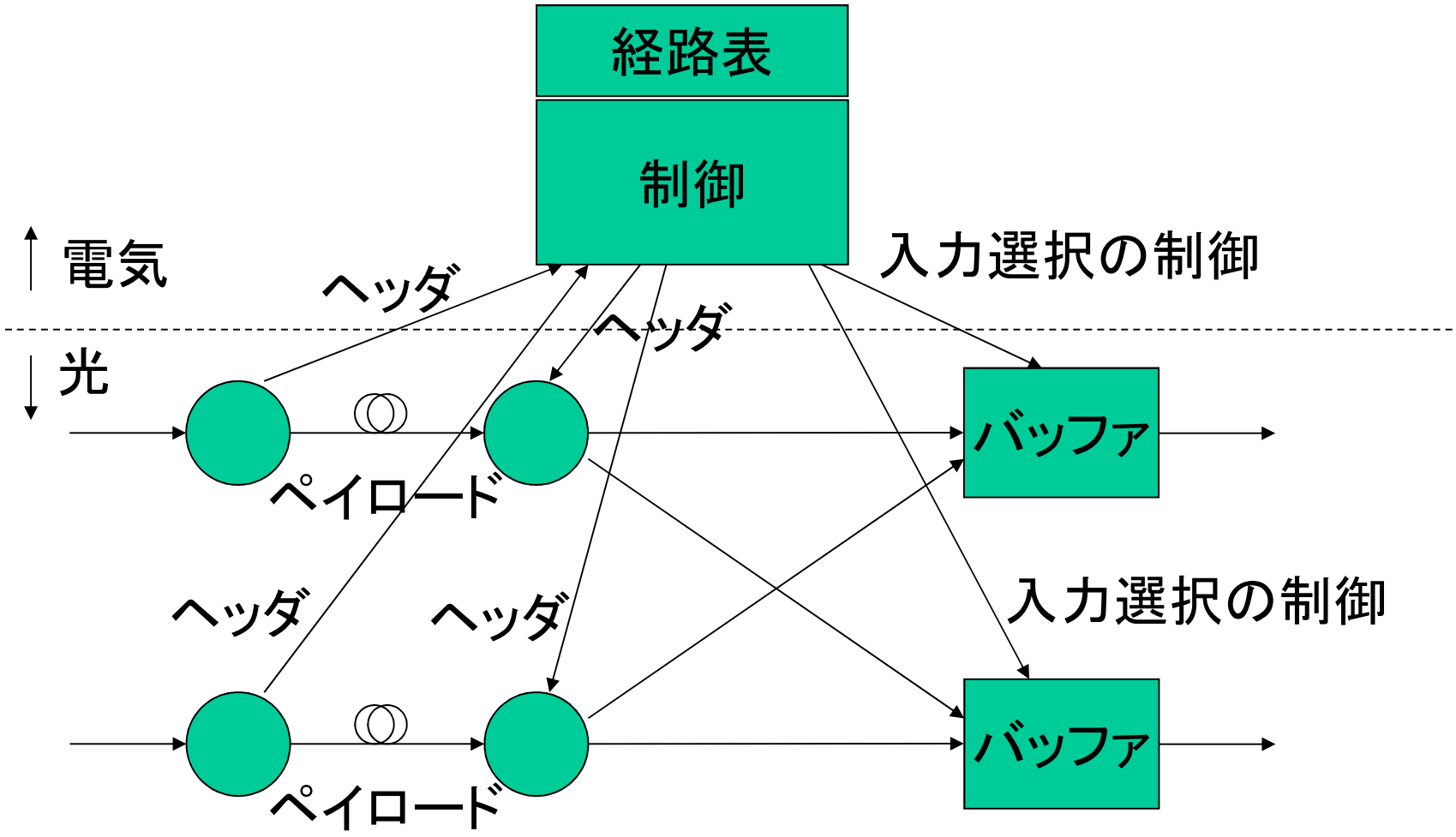
- 幹線では巨大なバッファが必要？
  - 幹線では多数（N）のTCPの変動が平均されるので（各TCPは独立）
    - 変動は1／sqrt（N）に
      - バッファは1／sqrt（N）に？
    - 回線速度を1／sqrt（N）の数倍犠牲にすれば
      - 総送信速度が回線速度を上回ることは、まずない
      - バッファは短時間変動を吸収する十数パケット分で十分
        - » 光ルータが実用的に

バックボーンルータ

- バックボーンはルータ10段くらい？
  - 全光でスイッチ
- 長期のバッファは不要
  - 偶発的パケット落ちは1段0．001%程度に
    - 15本程度の遅延線バッファで十分
- デフォールトフリーな経路表？
  - 数十（百？）万エントリー？
- 数百バイトに対して性能が出ればよい

環境の仮定

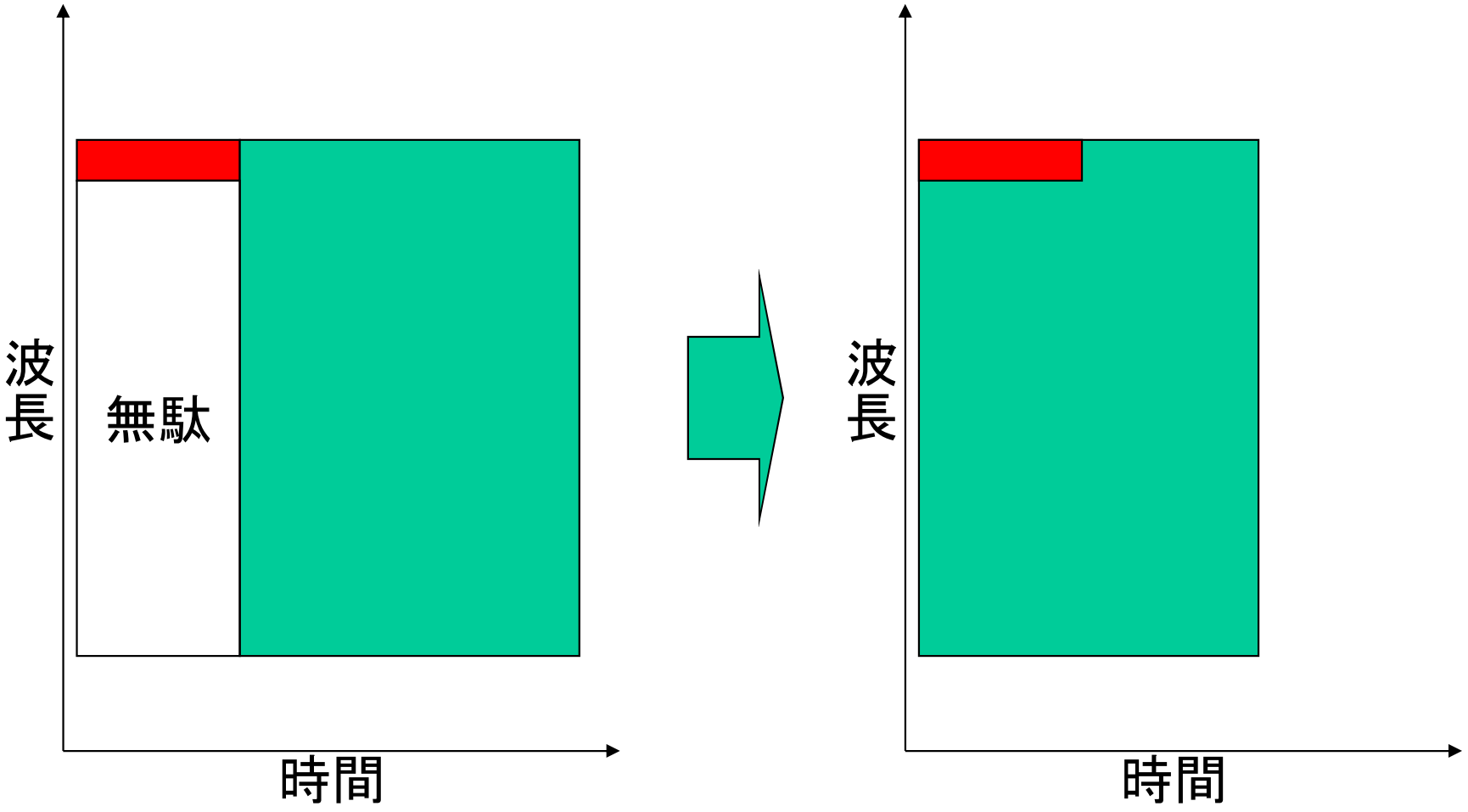
- インターネットバックボーンで利用
  - 平均パケット長500Bで、そこそこの性能を
    - 将来は、ジャンボフレームにより増えるかも
- 伝送路は10Gbpsを100波長多重
  - 長距離伝送にも困難はない
    - 短距離では10Gbaud＊6bit／baud＊200波長程度も可能



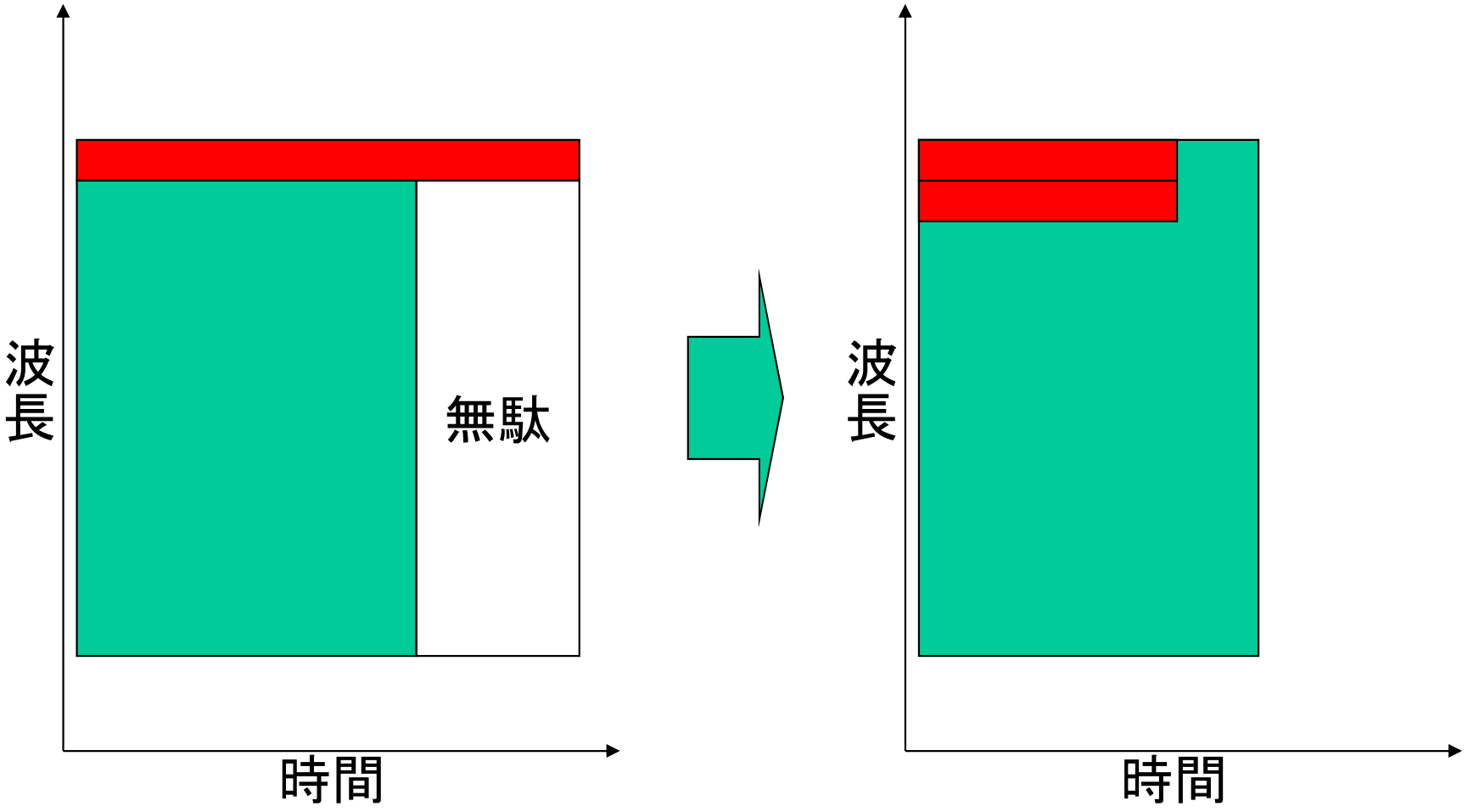
全光データパスルータの概略

基本的パケット形式

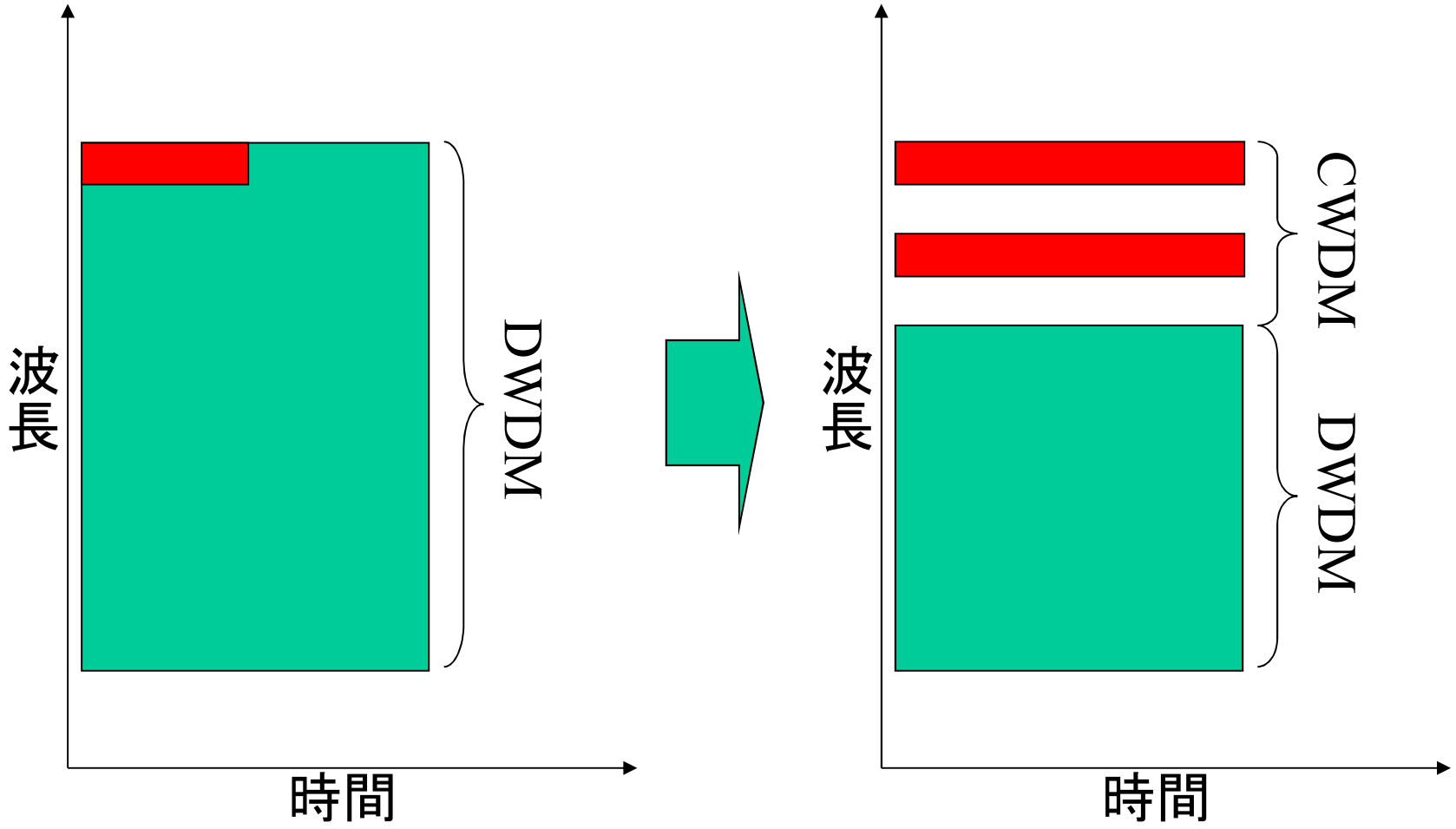
- ５００Ｂ１００波長では、波長あたり5Ｂ
- パケットはヘッダとペイロードからなる
- ヘッダとペイロードを時間軸で分離すると
  - ヘッダ伝送中はペイロードが送れない
    - 実効速度が低下
- ヘッダとペイロードは波長多重
- ヘッダには複数波長を利用する
  - ヘッダ部分をWDMにするとADMが楽



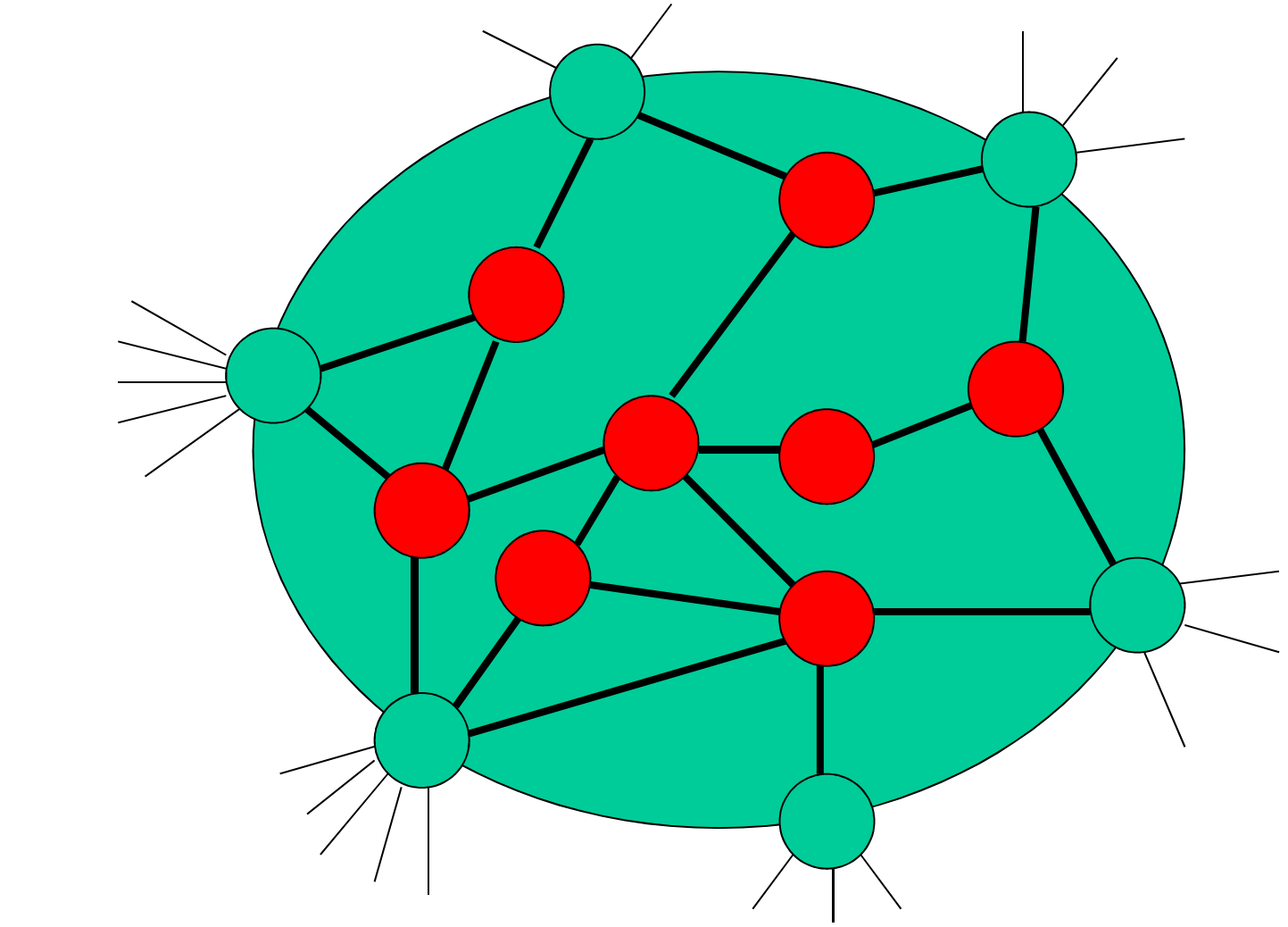
ヘッダとペイロードの時間軸上の重ね合わせ



ヘッダの波長軸への分割



ヘッダの分離をより容易に

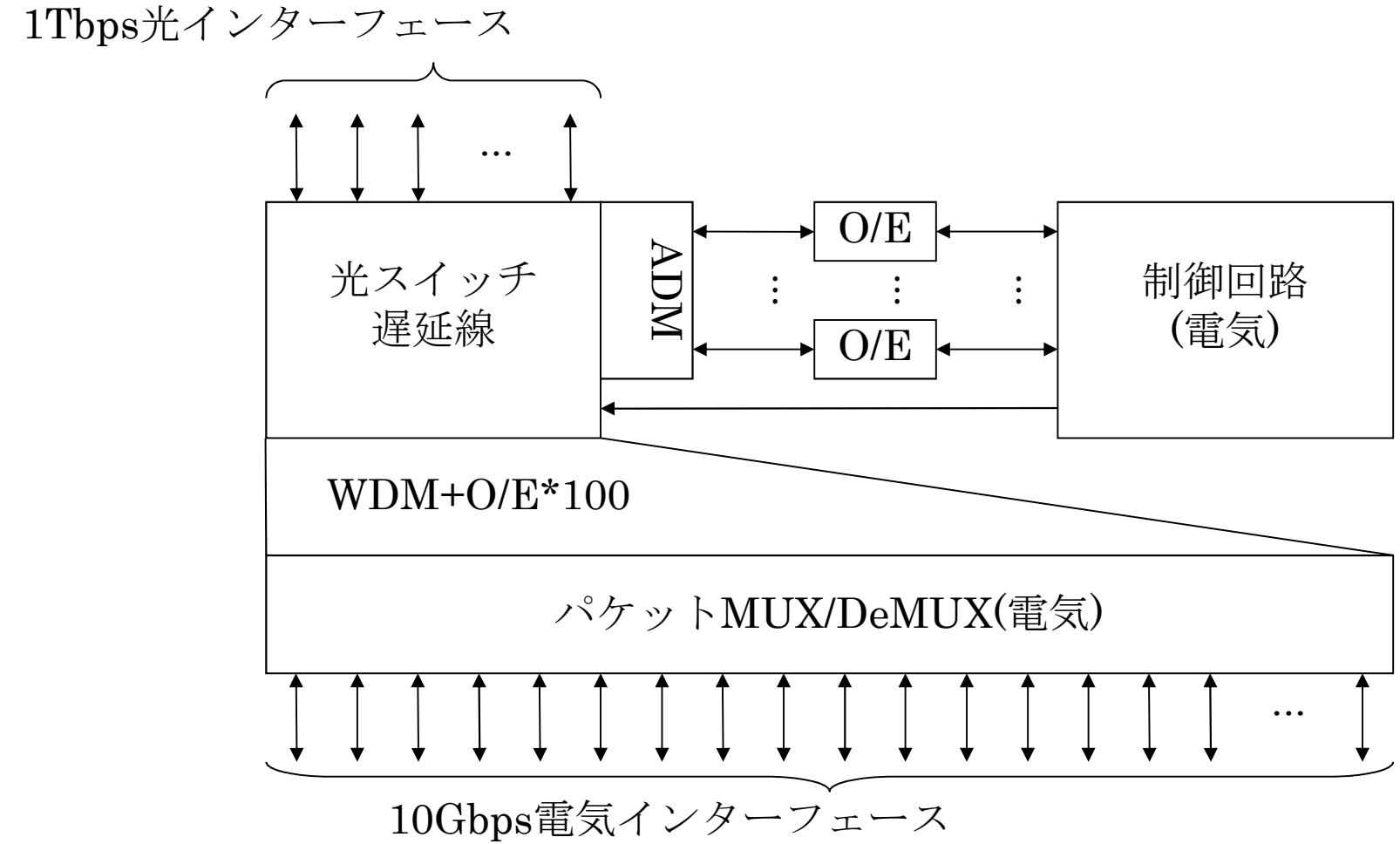


：コアルータ　：エッジルータ　：光　：電気

全光ネットワークのコアルータとエッジルータ

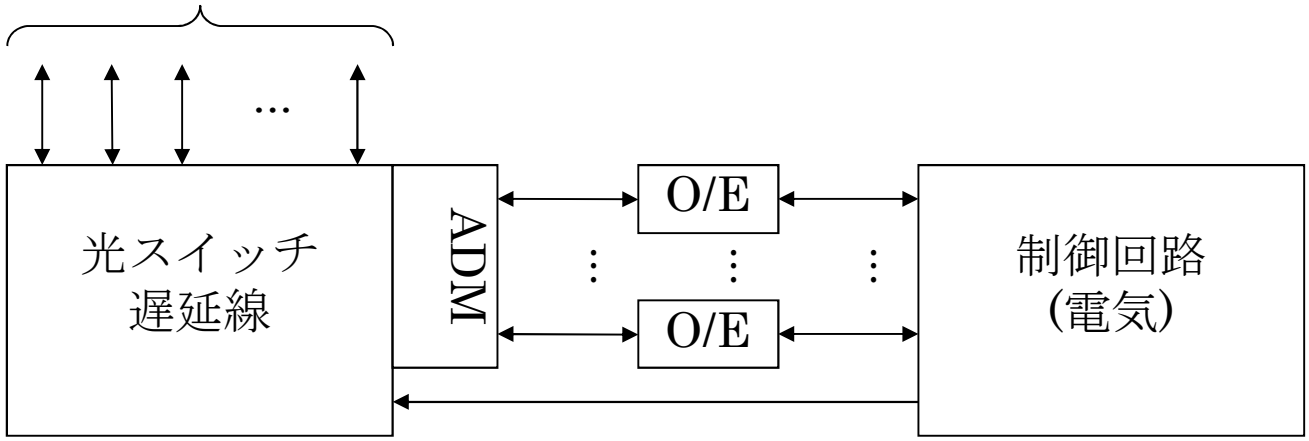
コアルータとエッジルータ

- コアではほぼ全光、エッジでは電気が必要
- エッジルータは高価、コアルータは安価
- ほぼ全光ルータの電気回路が（自らへのパケットを受信／自らパケットを発信）する（経路制御、ICMPエラー等）には？
  - 頻度が低ければ（Gbps程度）、波長時間変換回路を利用すると容易に可能
    - 高信頼化光源（＋予備）で、LDの寿命問題を回避

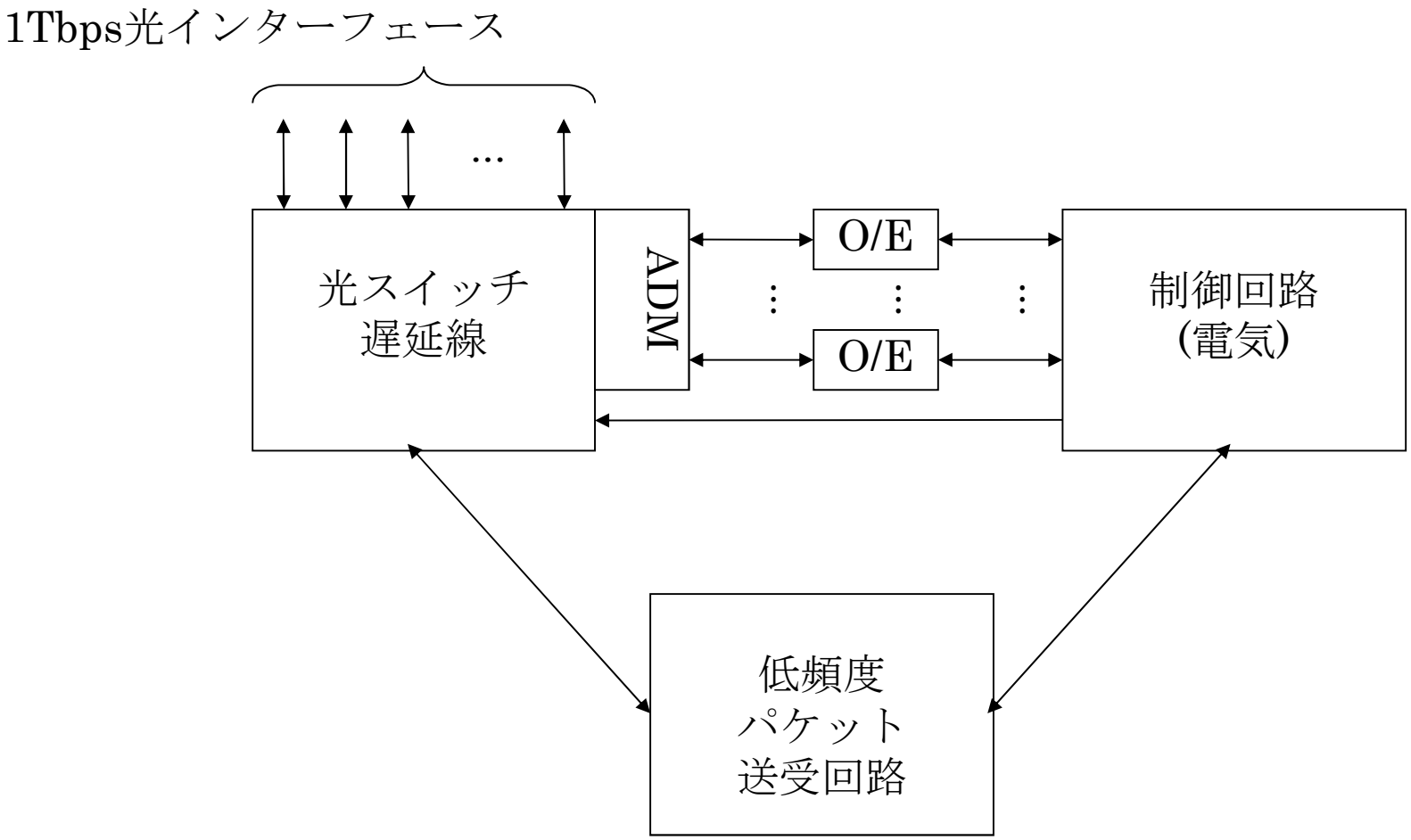


エッジルータ（電気回路部分が高価？）

1Tbps光インターフェース



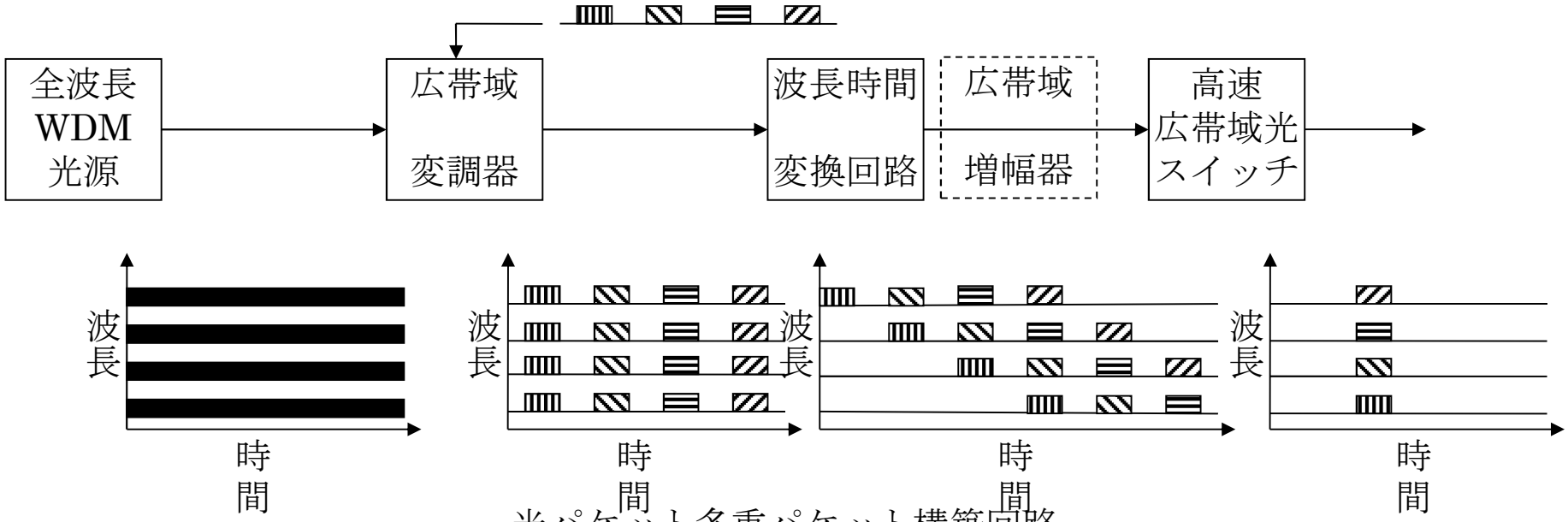
光幹線網の中枢の光ルータ（自らパケット送受信は不可）



光幹線網の中枢光ルータ(パケット送受信可能)

波長時間変換による光パケットの構築

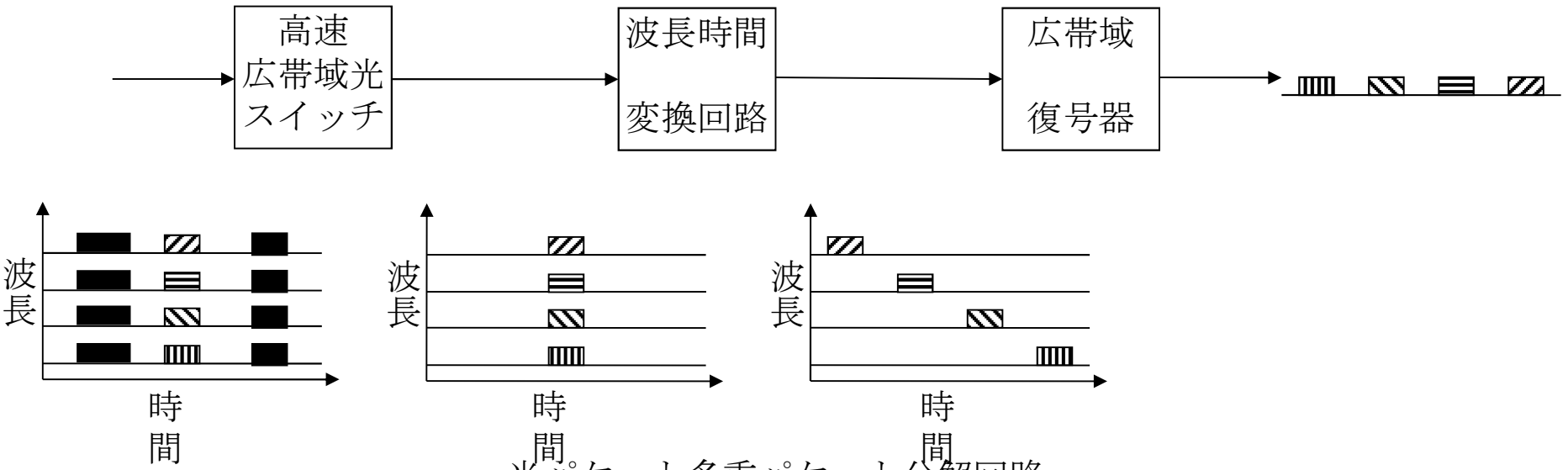
- 全波長光源（SC光源等）を
- 広帯域変調器で変調し
- 波長時間変換によるDESを施し
- （光を増幅し）
- 光パケット多重パケット部分を高消光比で切り出す



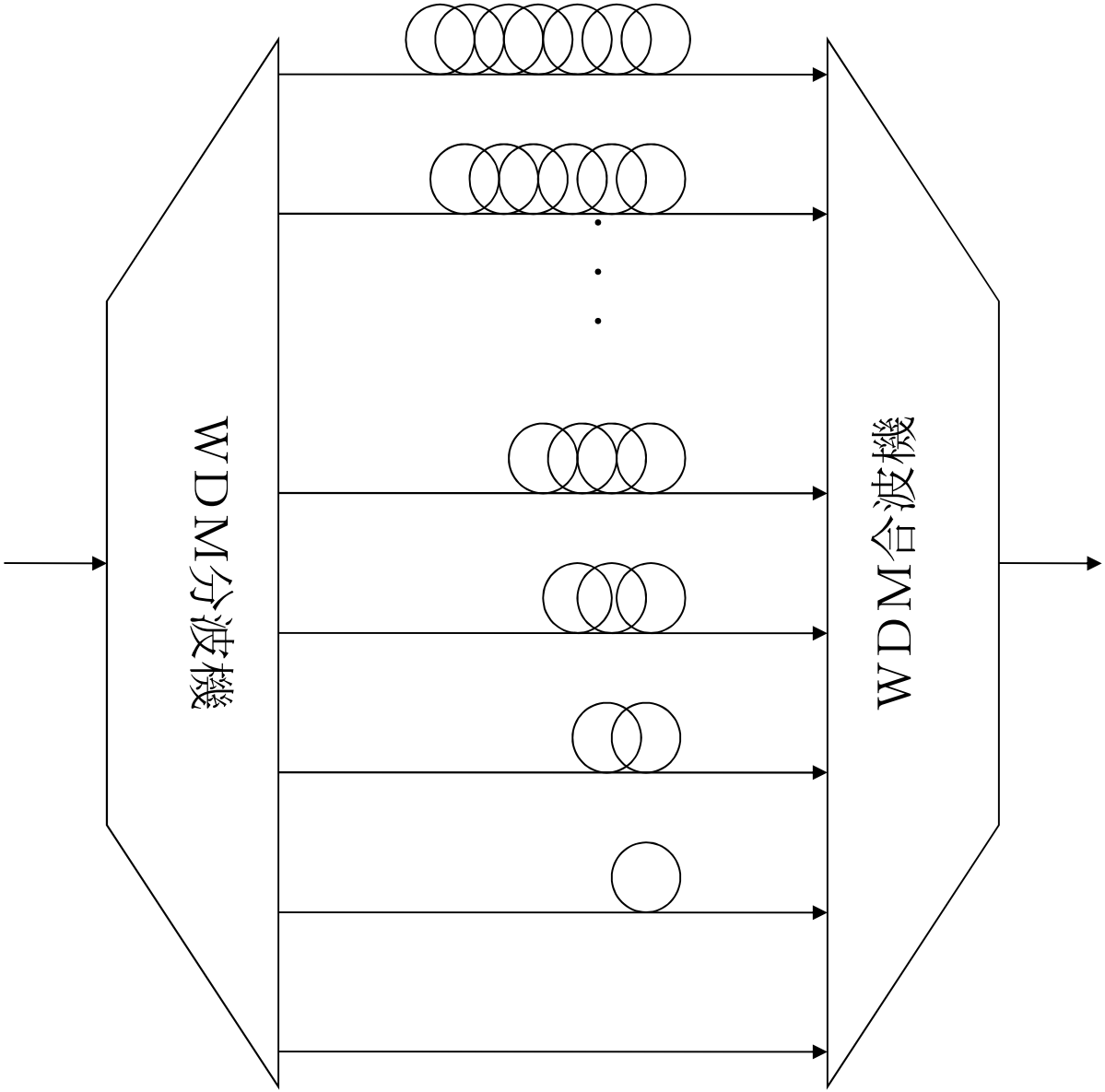
光パケット多重パケット構築回路

波長時間変換による
光パケットの分解

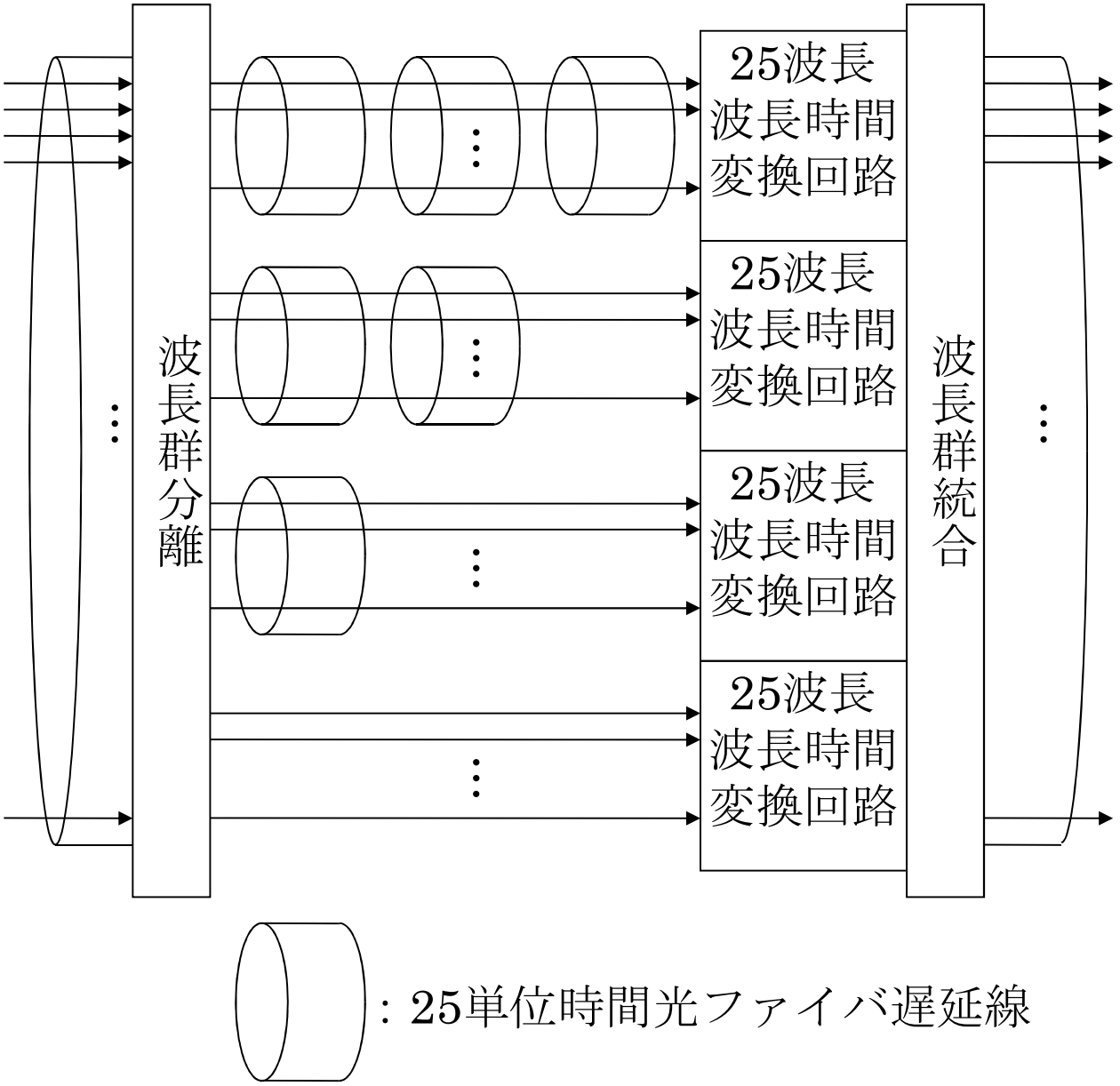
- 光パケット多重パケットを高消光比で切り出し
- 波長時間変換によるSERを施し
- 広帯域復調器で復調



光パケット多重パケット分解回路



波長時間変換回路の構成例



光ファイバ長節約型100波長波長時間変換回路

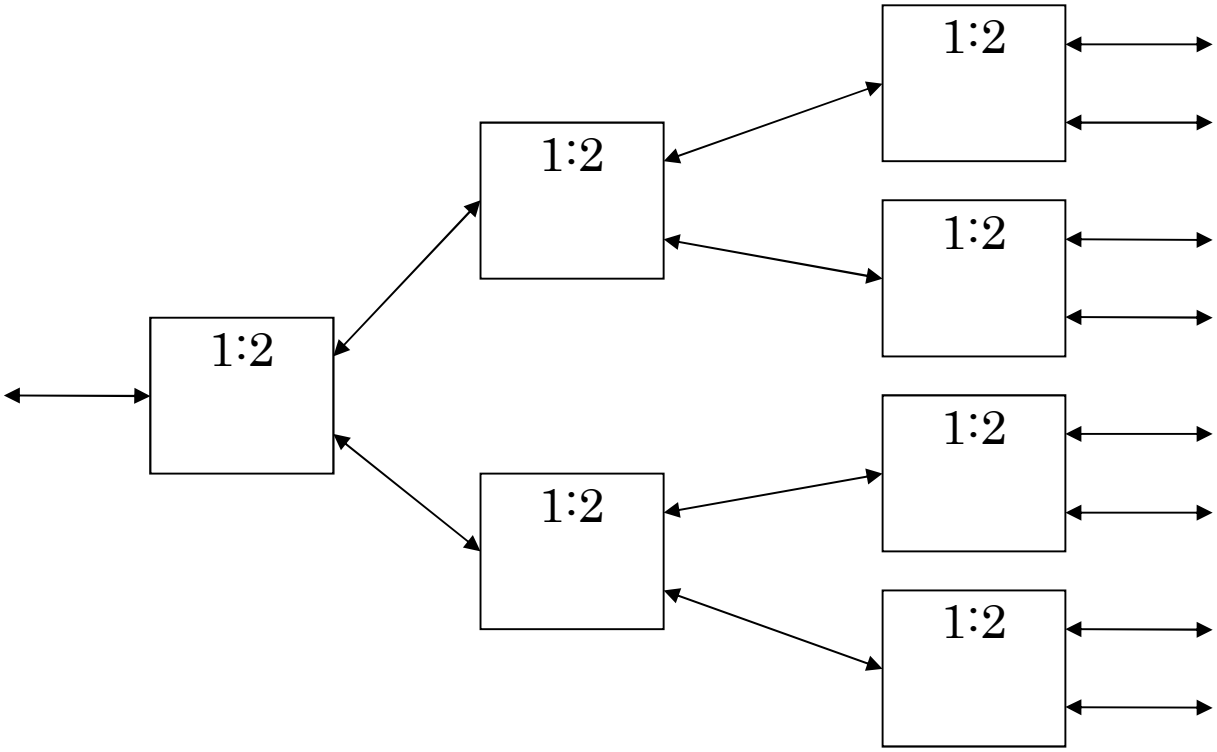
パケット形式とパケット間

- パケット間は、無光に
  - パケット間での光スイッチで信号が乱れない
- パケット間が長い（μs～ms単位）と
  - EDFAにエネルギーが溜まる
    - 次のパケットの先頭でサージが、、、
  - ダミーパケットで対処
    - 数μsの平均が一定になるように
    - ダミーは、次段のルータで無視

光パケットヘッダに含むべき情報

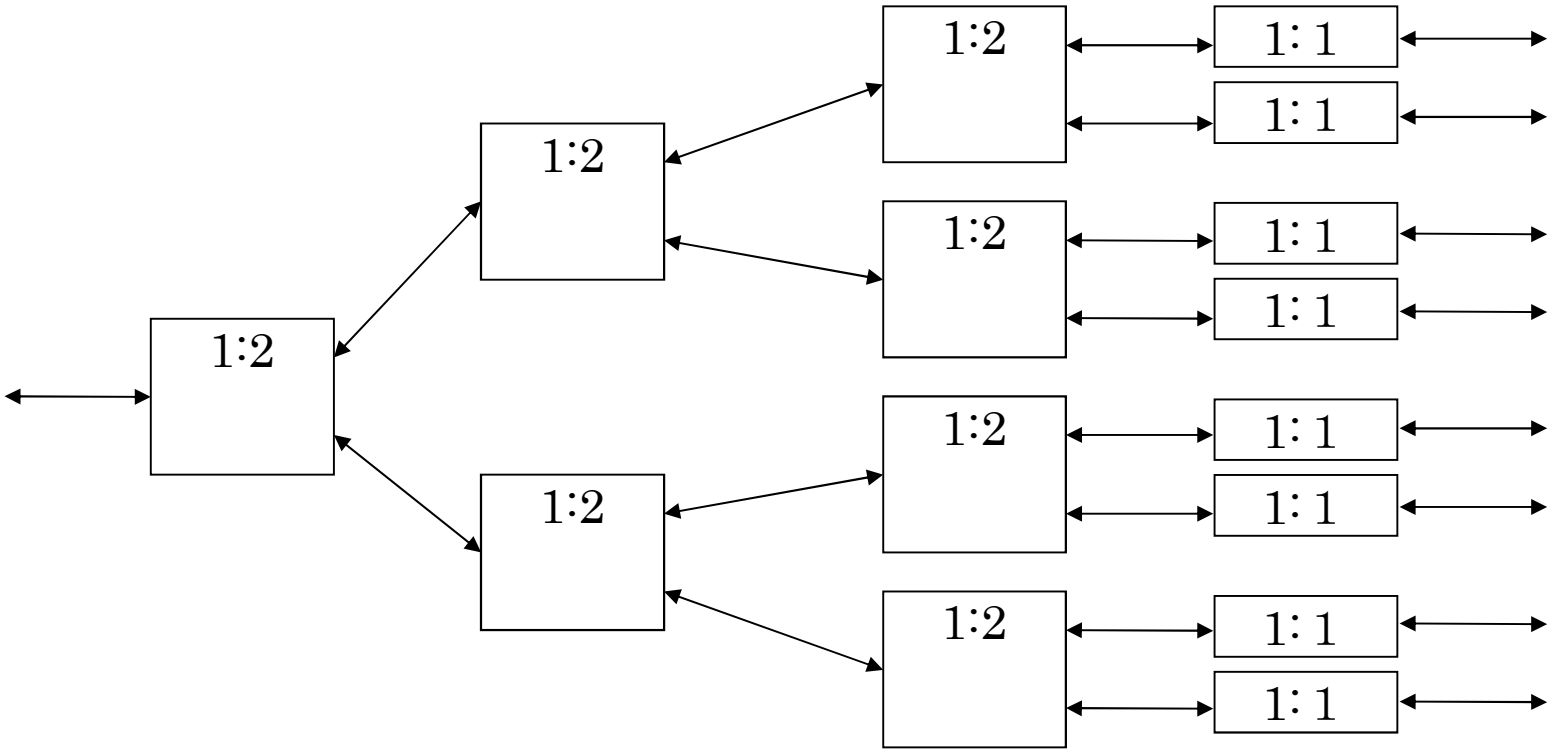
- 少ないほどよい
  - よりAll-Opticalに近づく
  - イーサネットは衰退する（ヘッダが大きすぎ）
- ディスティネーションアドレス情報
  - AF＋アドレスの上位数（４？）バイト？
- （パケット長）、TTL、ToS、（フローラベル）
- 光ネットワーク内でフラグメント化はやらない（MTUを統一）

MUX／DeMUX



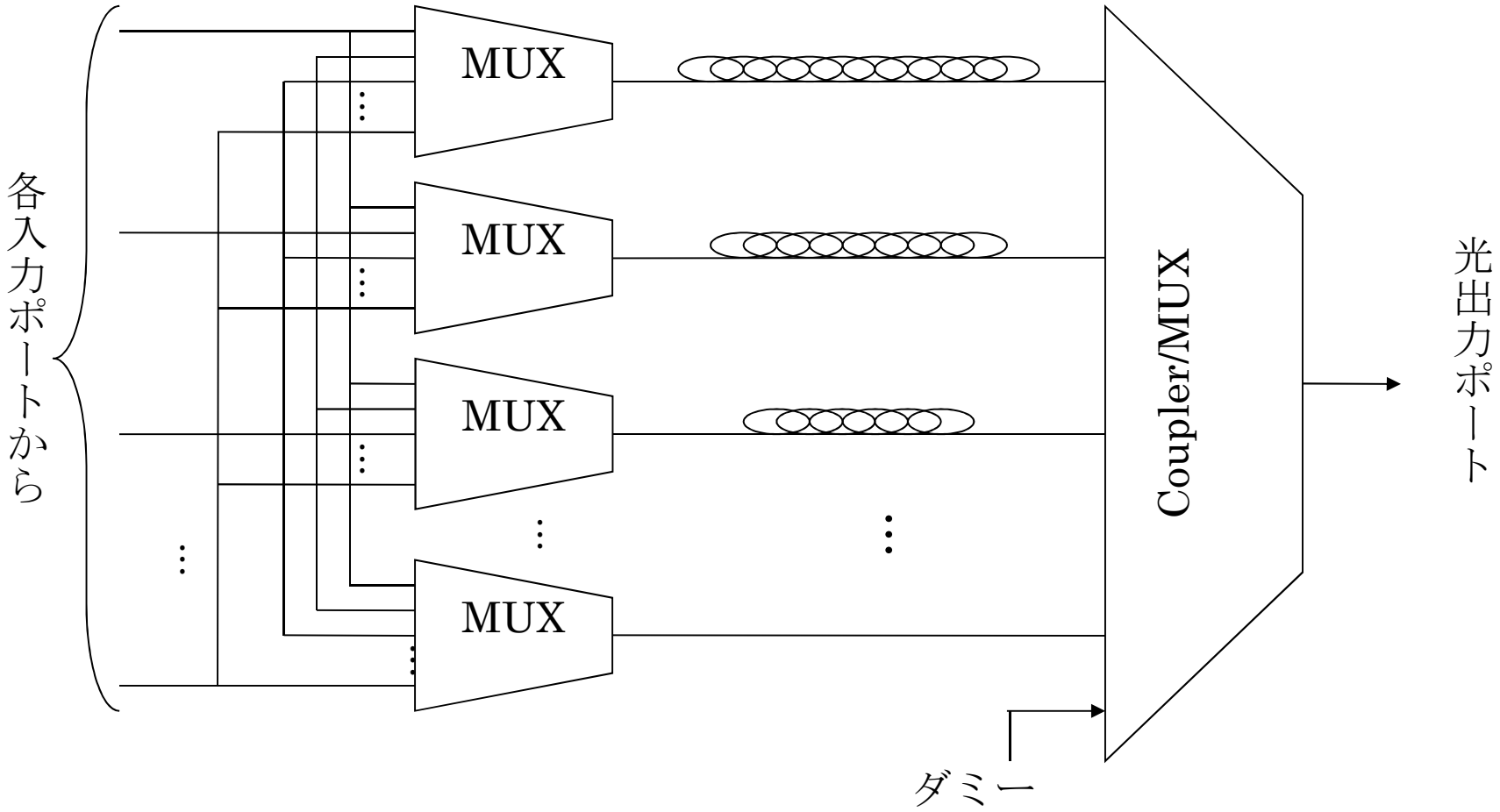
光MUX（⟶）と光DeMUX（⟵）

消光比を改良した MUX／DeMUX



光MUX（⟶）と光DeMUX（⟵）

遅延線による光バッファ



パケット落ちの確率とTCP

- 遅延線15本（等比、最長813m）の場合
  - 負荷65%（497Gbps）で0．0017%
  - 70%で0．833%、75%で4．9%（RED）
- TCPの理論性能
  - 0.97*MSS/RTT/sqrt(パケット落ち確率)
  - ルータ10段、MSS1440B、RTT0．1sで
    - 34Mbps
    - TCP1万本で340Gbps（幹線では十分）

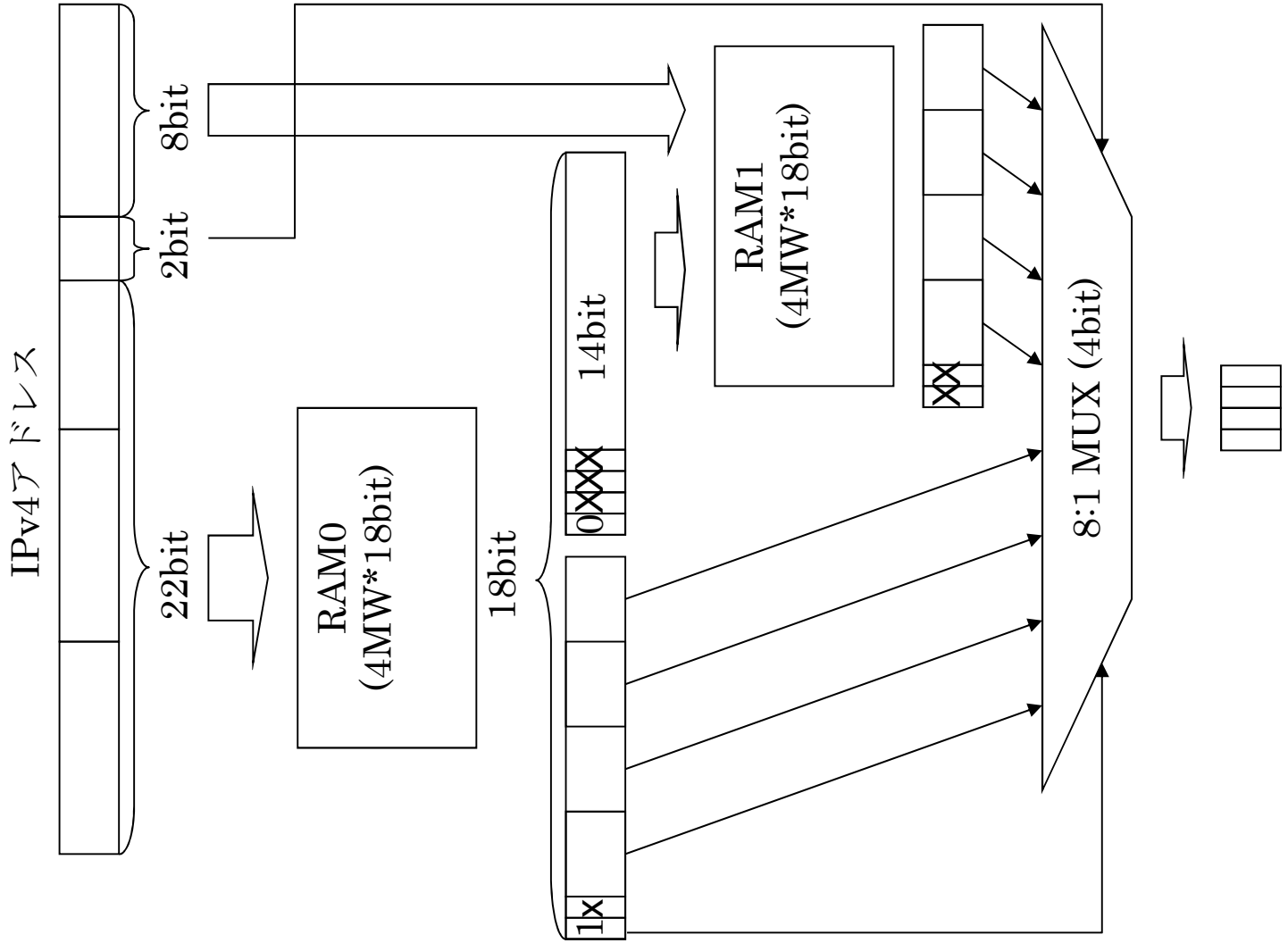
パケットの順序とTCP

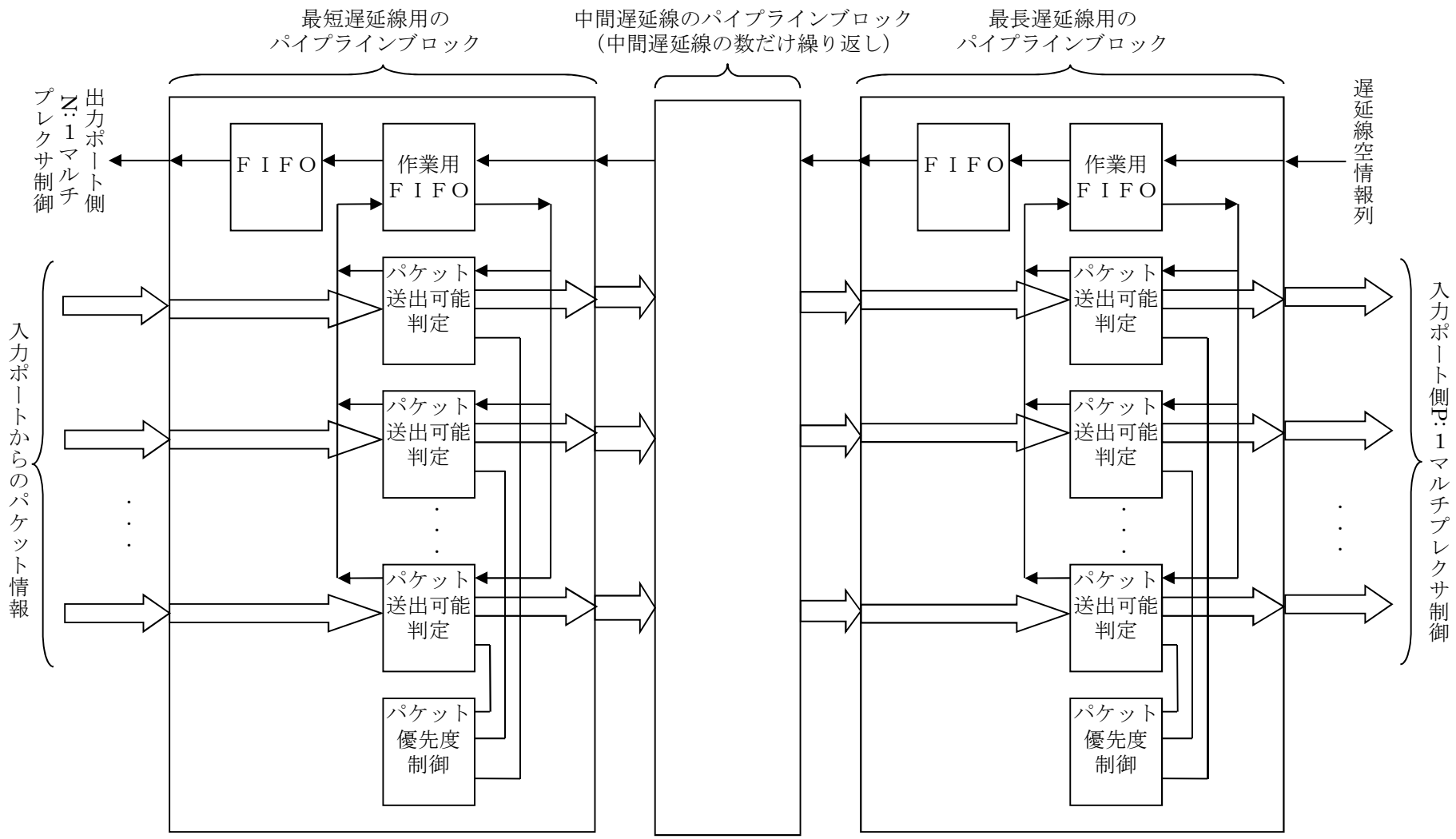
- TCPで同じシーケンス番号のACKが3個続くと
  - Fast　Retransmissionが動作
  - パケット落ちと認識される
- データパケットの順序が変ると
  - 先着パケットは無視される（再送が必要）
- よほど高速でないと、順序は変らない
  - 813mの遅延線で4μ秒（レート2．9Gbps）

電気回路の規模と速度

- 経路表検索
  - ／24までのフルルート＋16Kの／22を細分
  - SRAM2チップで実現可能
  - パイプラインクロック3．3ns
  - IPv6もパイプライン段数増やせば対応可
- 遅延線制御
  - 遅延線方向のパイプライン化が可能
  - 550MHzFPGAで4ns以下で動作

IPv4アドレスによる経路表の高速検索

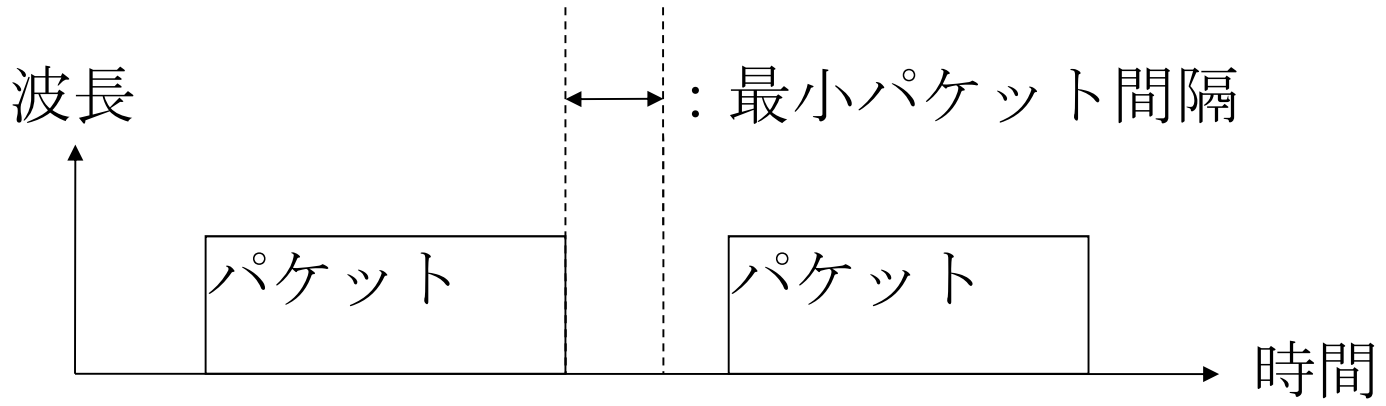




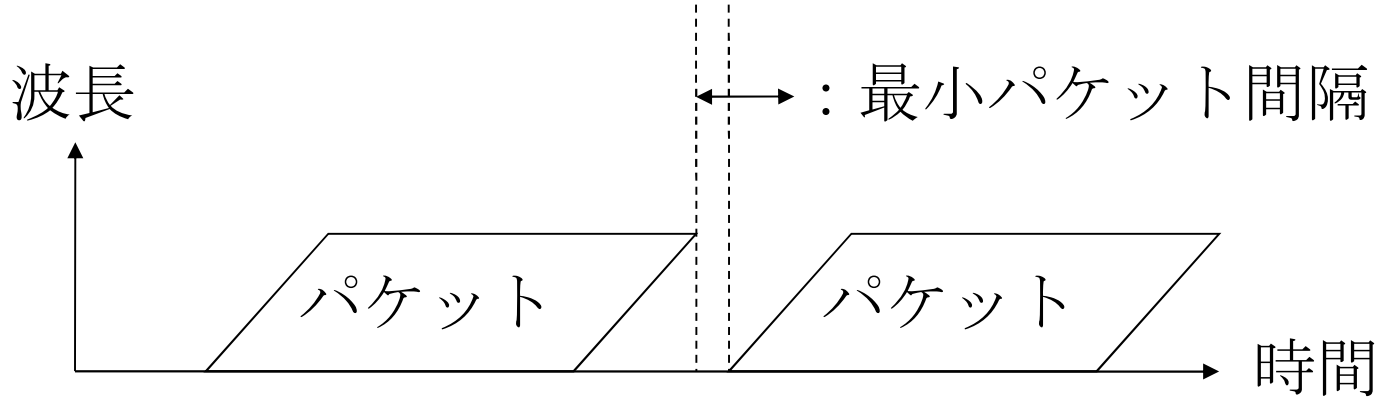
パイプライン化されたバッファ制御回路

分散の影響

- 波長内では
  - アイパターンが乱れ、復調できなくなる
  - 数十ps程度で十分問題
- 波長間では
  - パケット単位でスイッチできなくなる
    - 数nsもずれると、かなり問題
    - SLAとIDFを用いた理想的な分散マネージメント伝送路では、２．５THzの帯域内で
      - ５０００Kmの伝送で群遅延差は＜1ns



a)当初のパケット間隔



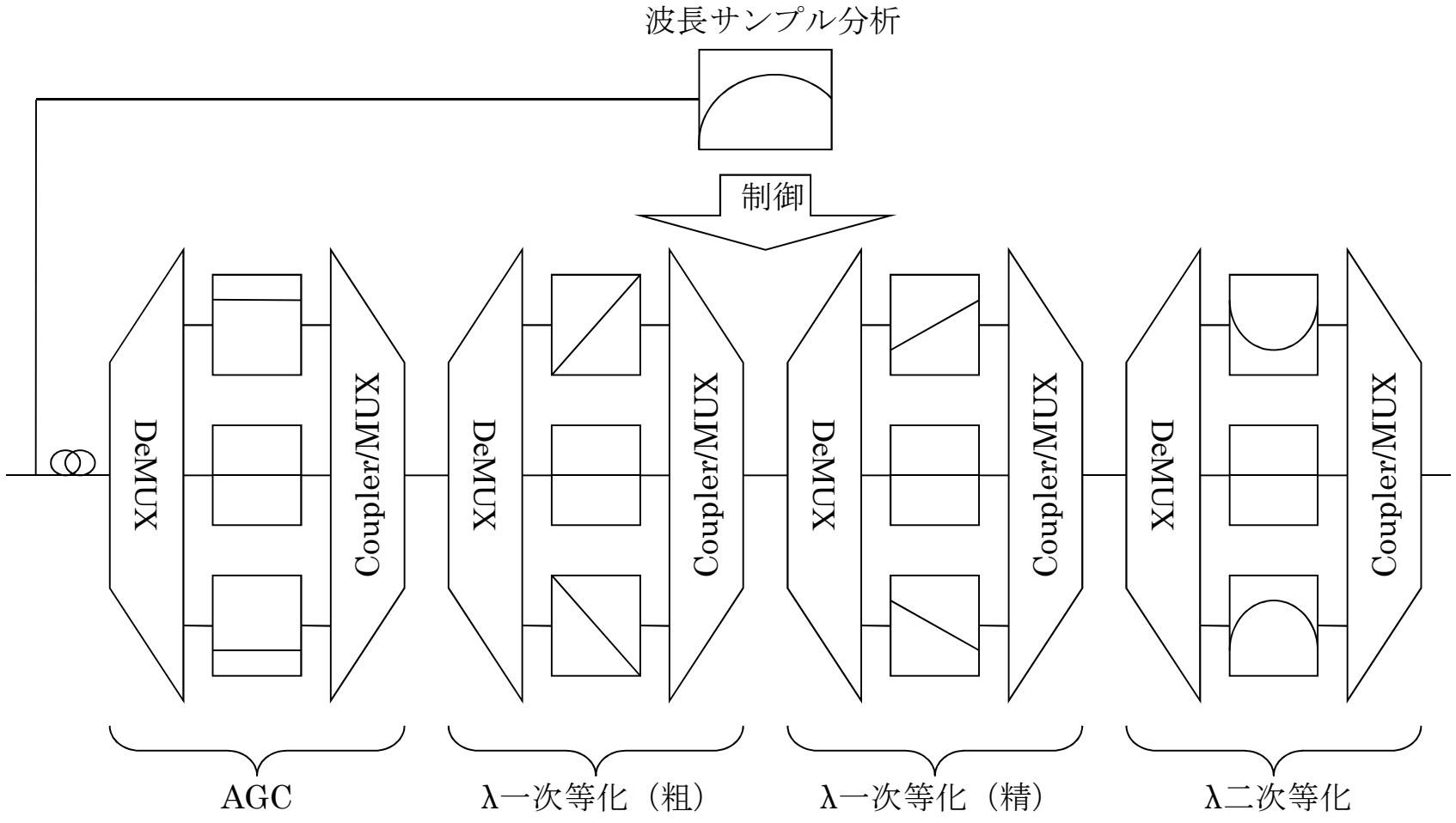
b)波長間のタイミングのずれとパケット間隔

パケット間隔と波長間タイミングのずれ

伝送特性補正

- 多段の光回路ではパケットが徐々に歪む
- 波長ルータでは歪みは波長ごとに違う
  - 補正は波長ごとに必要（波長数の補正回路）
- パケットルータでは経路はパケットごとに違う
  - 信号強度や歪みもパケットごとに違う
  - 補正はパケットごとに必要
  - 歪みは波長に対してなめらかに変化

パケット単位の
AGCとλイコライザ



期待できる速度

- ラインレート1Tbps、平均パケット長500B（4ns）、最小パケット間隔2nsで
  - 平均最高速度666Gbps
- 負荷率65%で
  - 平均実効速度433Gbps

偏波依存損失（PDL）の問題

- 通常の光スイッチ素子は、偏波状態によって損失が微妙に異なる
  - 偏波状態は、光ファイバでの伝送で、波長ごとにランダムに変化
    - PDLにより、各波長の信号強度がぶれる
- PDLが大きい（＞0．1dB？）と
  - 単一偏波で偏波保持ファイバを使うしかない
    - 既存WANのファイバは、使えない

消費電力

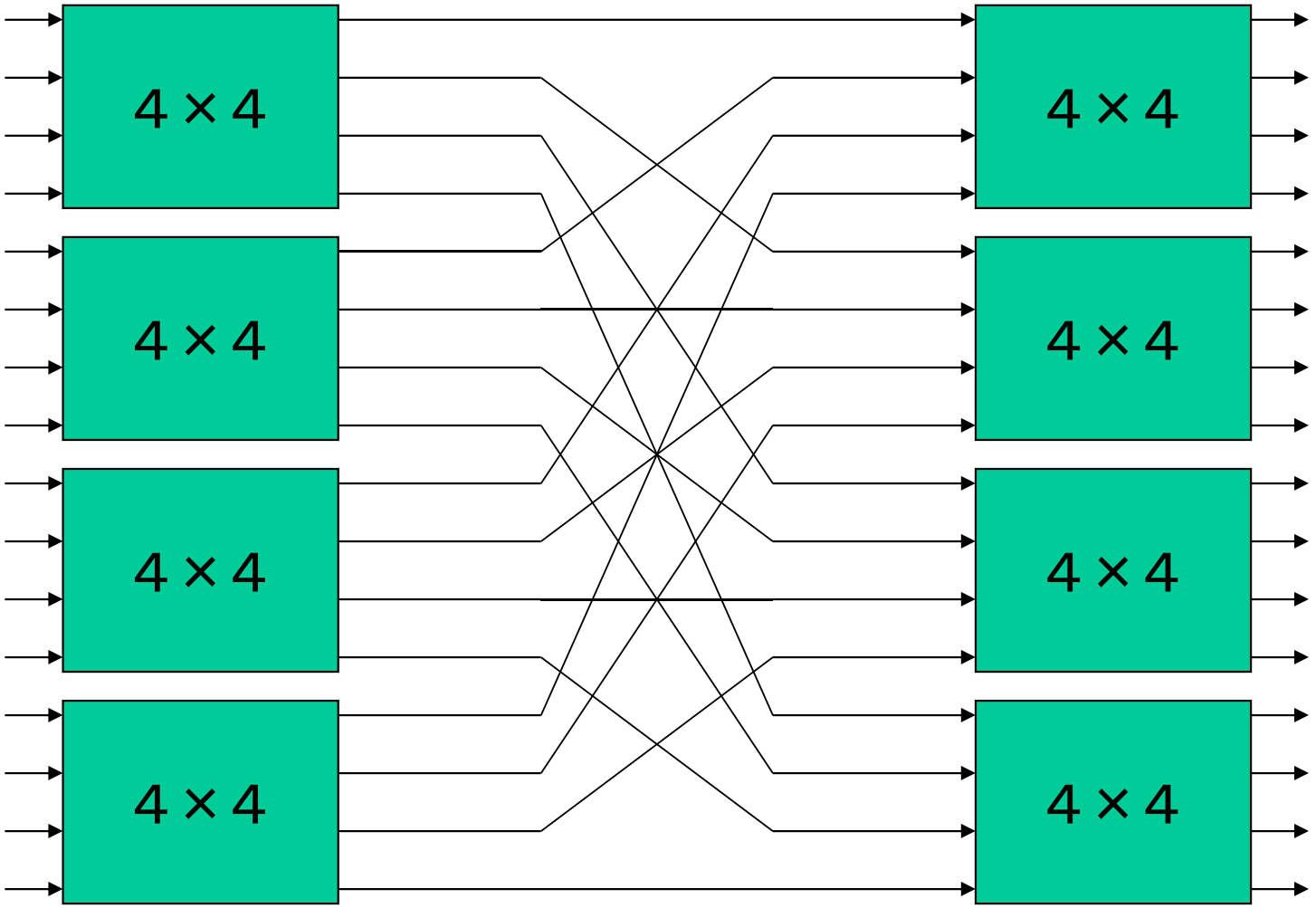
- 8ポートで個別バッファ遅延線15本として、必要な2:1(1:1)スイッチ数は
  - 出力ポートあたり15＊15＋15＝240
- スイッチとスイッチドライバの消費電力は
  - 0．25W程度、全体で480W
  - 経路表、遅延線制御、光増幅等に、＋数十W
- 4ポートなら、120W＋α

超並列ルーティングによる ペタビットルータ

- 超並列ルーティング
  - 1Tbpsの要素ルータを1000台ならべる
  - それらを多段にして相互結合

相互結合網のつくりかた

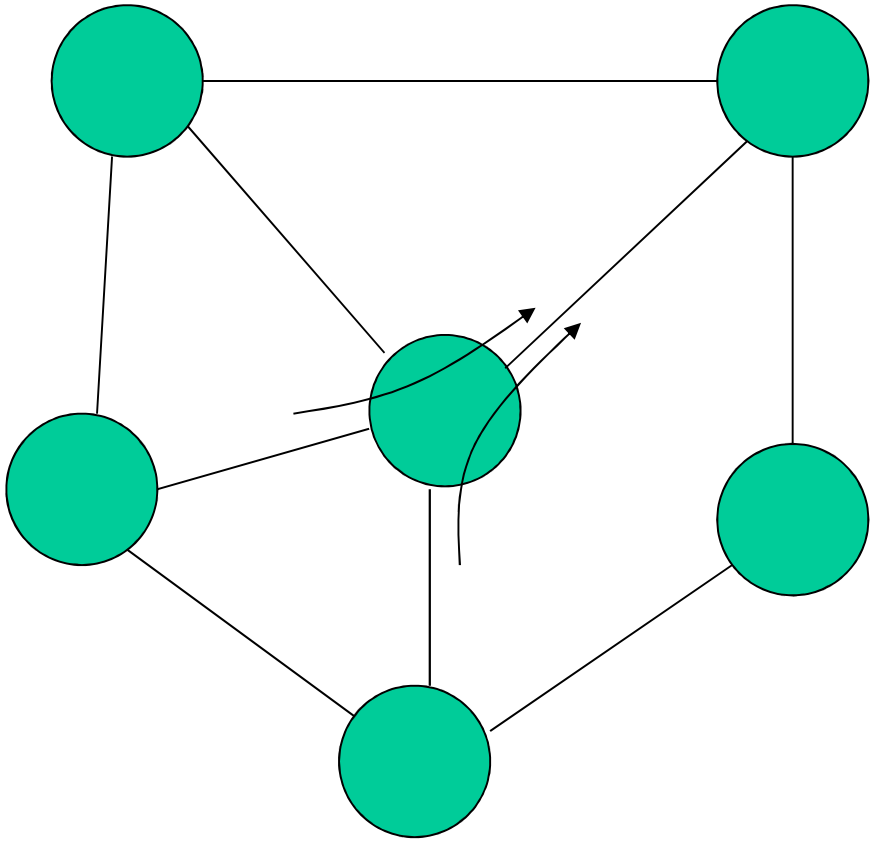
- Ｋ×Ｋの要素スイッチを多段につなぐ
- Ｎ要素の相互接続には$\log_k N$段必要
- 少なくともNlogNのハードウェア
  - ハイパーキューブは非効率的（$N\log^2 N$）
- logNの遅延は避けられない
- １２８要素ルータを４×４の要素スイッチでつなぐと、４段必要



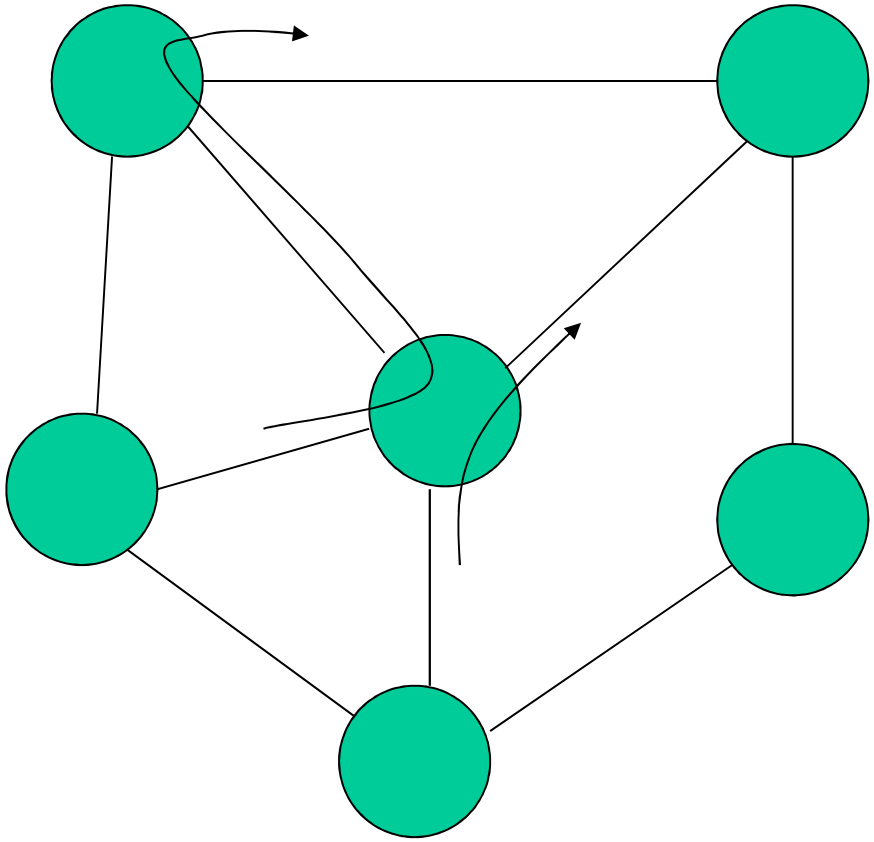
4ポートルータからの、16ポートルータの作成

そもそも衝突回避のために バッファは必要か？

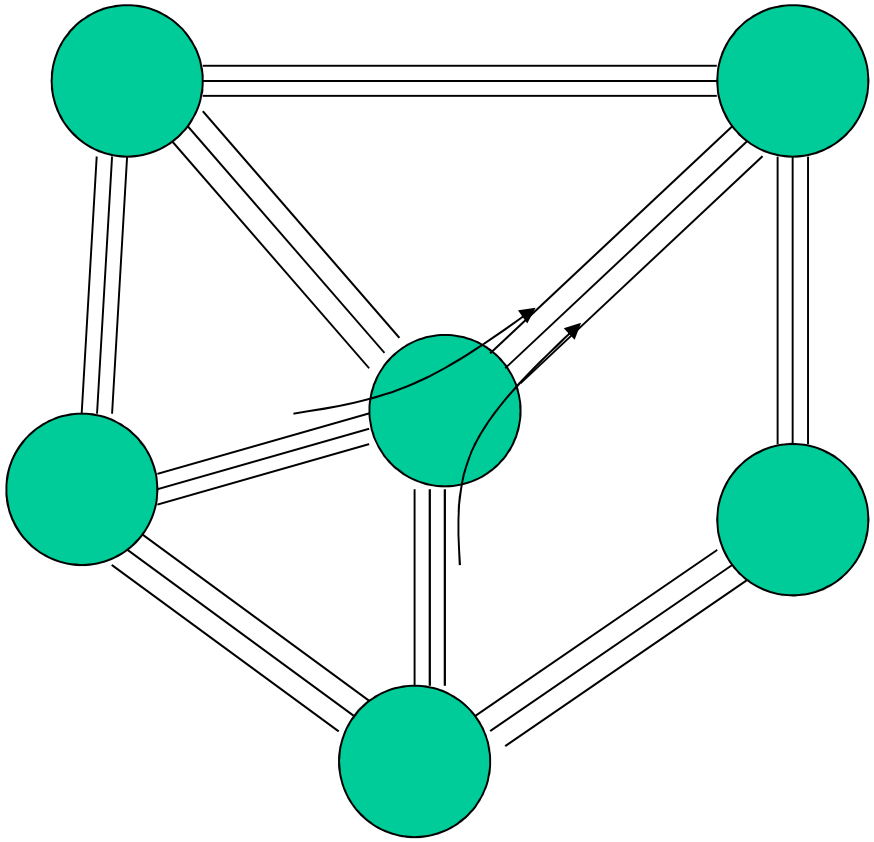
- 遅延線バッファは
  - 時間ドメインで衝突回避
- デフレクションルーティングという技法
  - 空間ドメインで衝突回避
  - ほとんど効果がない上に、パケットが劣化
- ペタビット幹線では
  - 多数の平行光ファイバが存在
  - 空間ドメインでの衝突回避が自然に可能



出力の衝突



デフレクションルーティング



出力の衝突なし

光バッファをしない場合（ポート数：４）

- 同期固定長でシミュレーション
- ファイバ数（Ｎ）２０～３０本程度から実用的
- ２：１光スイッチ素子数
  - Ｎ＝２０で４７２０個
  - Ｎ＝３０で１０６８０個

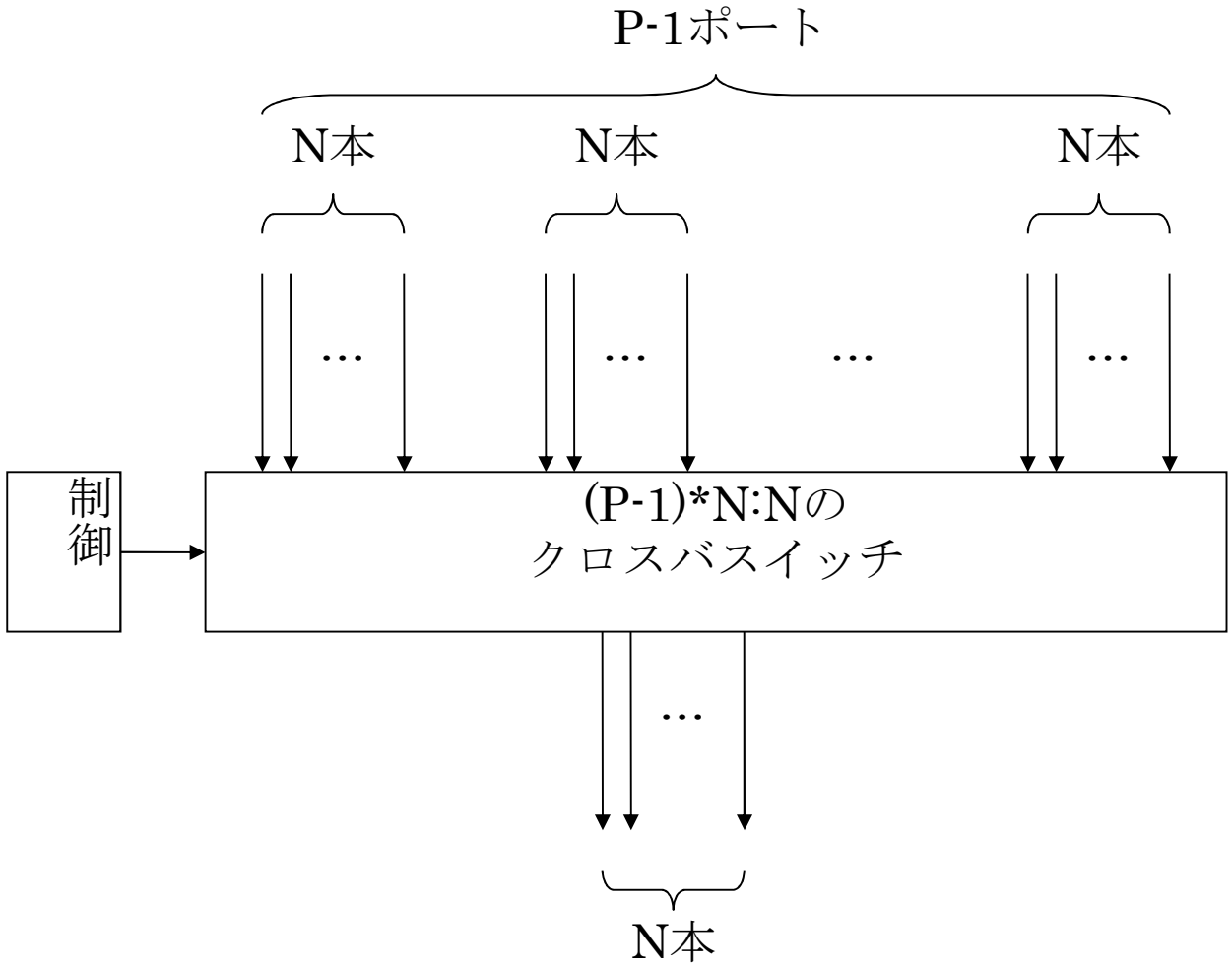


図１　空間ドメインだけで衝突回避を行う
光パケットスイッチの出力ポート

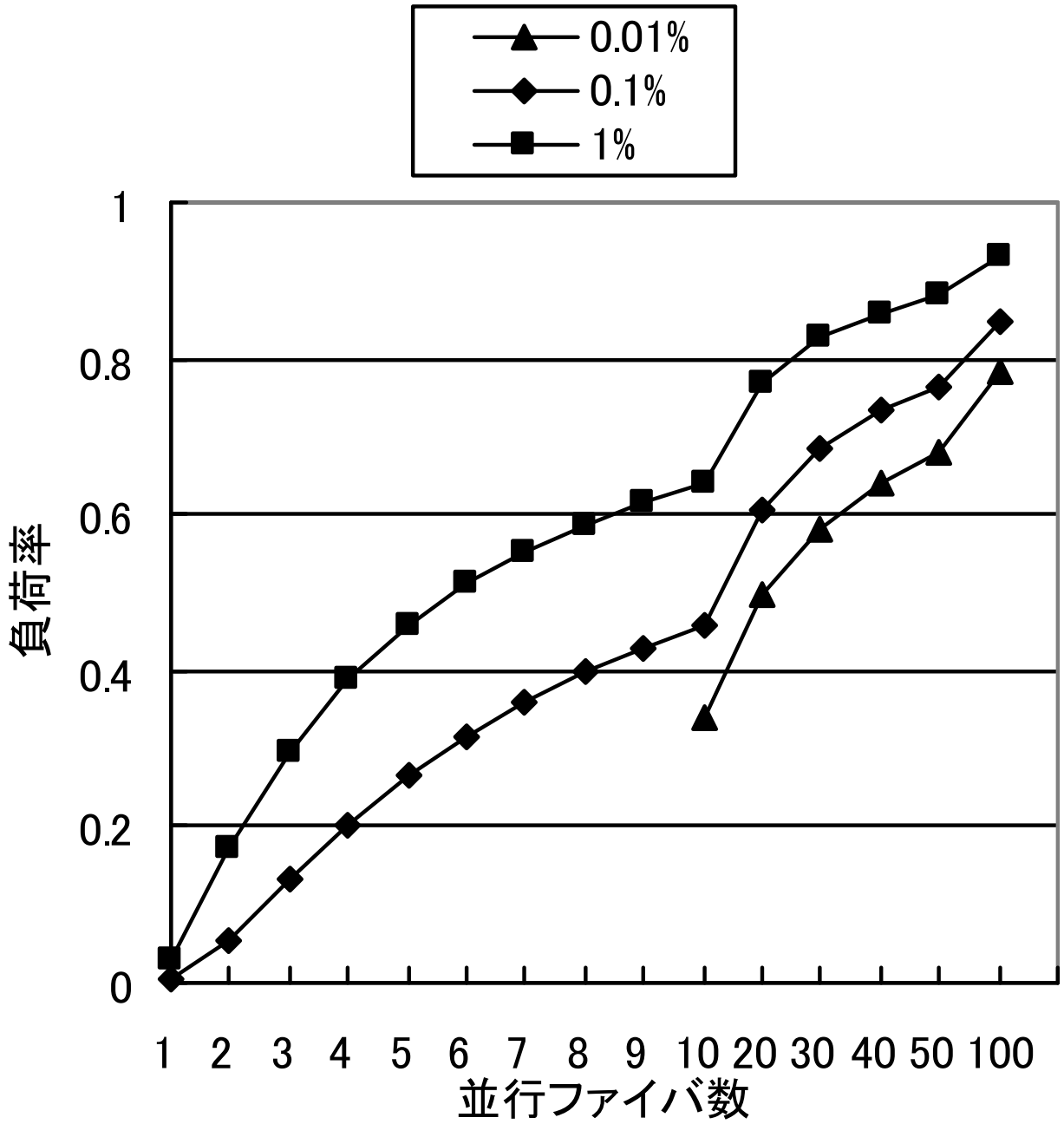


図2　並行ファイバだけでの衝突回避

光バッファもする場合（ポート数：４）

- 同期固定長でシミュレーション
  - 遅延は1パケット分
- ファイバ数４～５本程度から実用的
- ２：１光スイッチ素子数
  - Ｎ＝４で３６８個、Ｎ＝５で５８０個
    - 遅延線１６本の4ポートスイッチでスイッチ素子数１８８個（倍のスイッチ数で性能4倍）
    - 光パスと違い、並行光ファイバは必然ではないが
      - 幹線速度が増加してゆけば、時間の問題
    - 8ポートの場合も1760個と888個

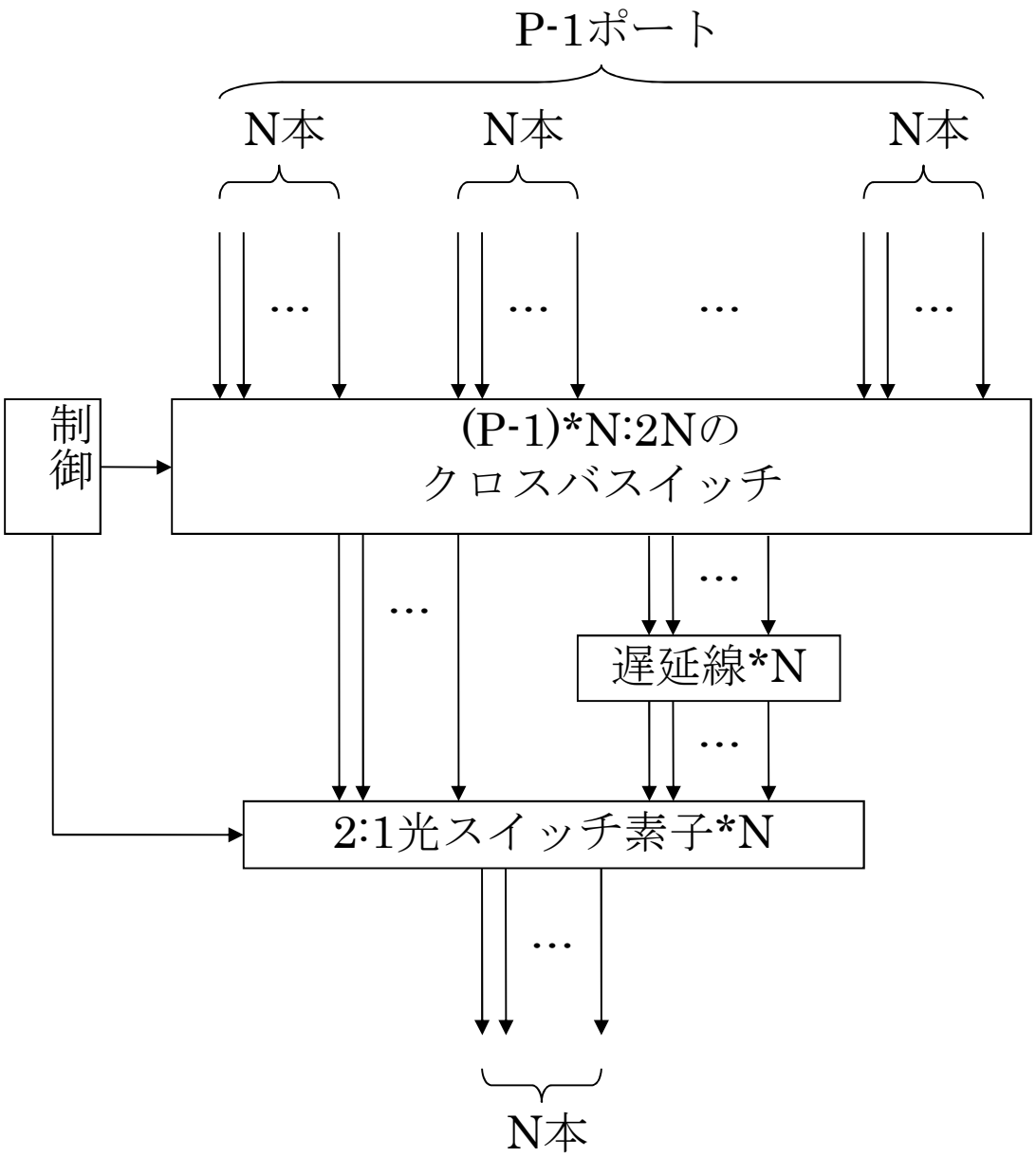


図3　空間ドメインと時間ドメインで衝突回避を
行う光パケットスイッチの出力ポート

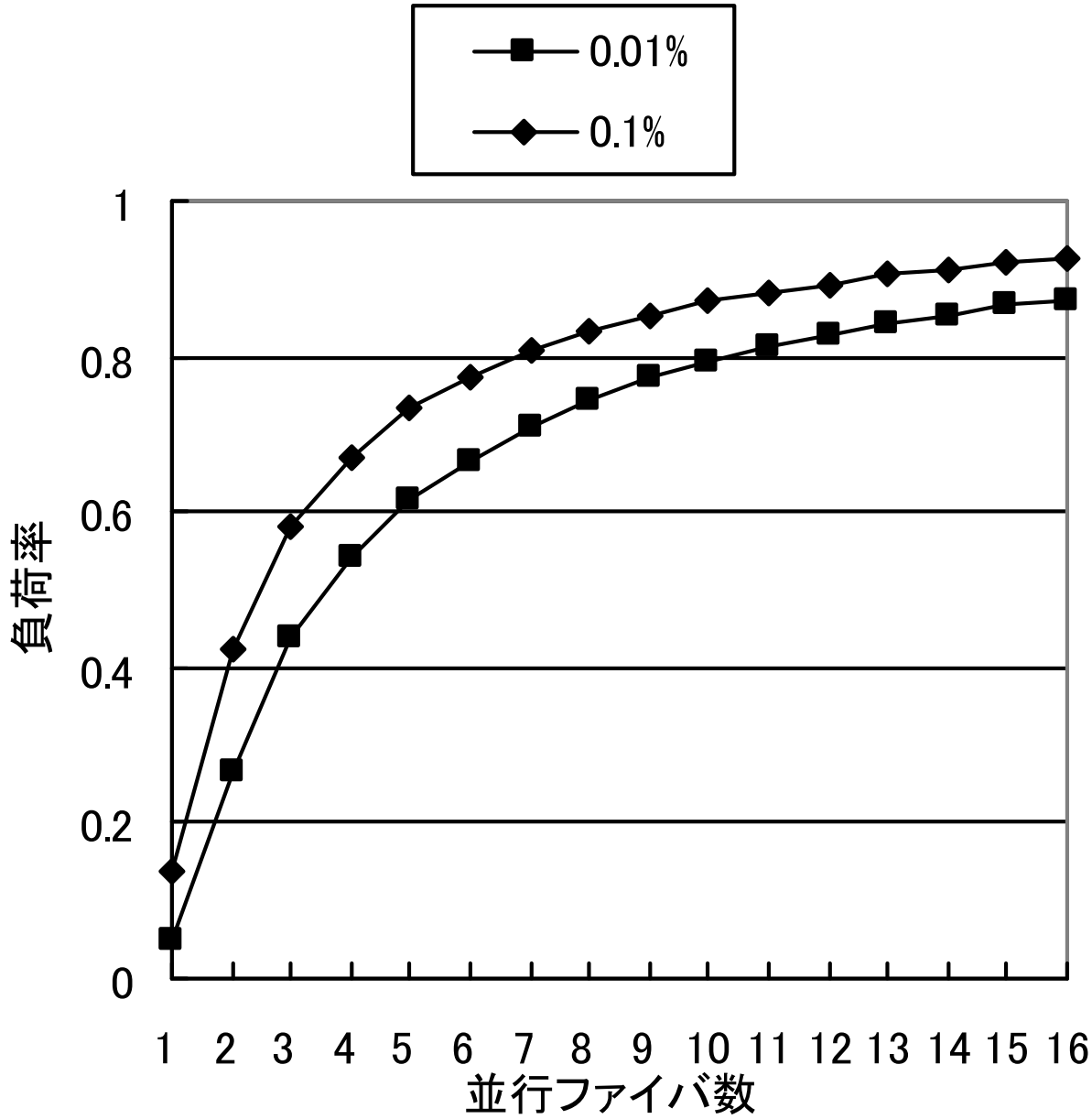


図4　並行ファイバと遅延線での衝突回避

スパコン内部での利用

- 超並列スパコン内部の相互結合網は
  - １０PFLOPSなら数Pbps程度が望ましい
  - 8ポート光ルータ（500Gbps＊８＝4Tbps）なら1000台（500kW）程度（＊２？）で済む
- 近距離なので、波長あたりの速度と波長数は増やせ、パケット間隔も詰められる
  - 例えば10Gbaud＊6bit／baud＊200波長＝12Tbpsも可、台数と消費電力が減る

おわりに

- テラビット級（ほぼ）全光ルータは
  - 現在の技術で実現可能
- 並列化によりペタビット級幹線も可能
- 需要がまだない
  - まずは、スパコンやデータセンター？

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

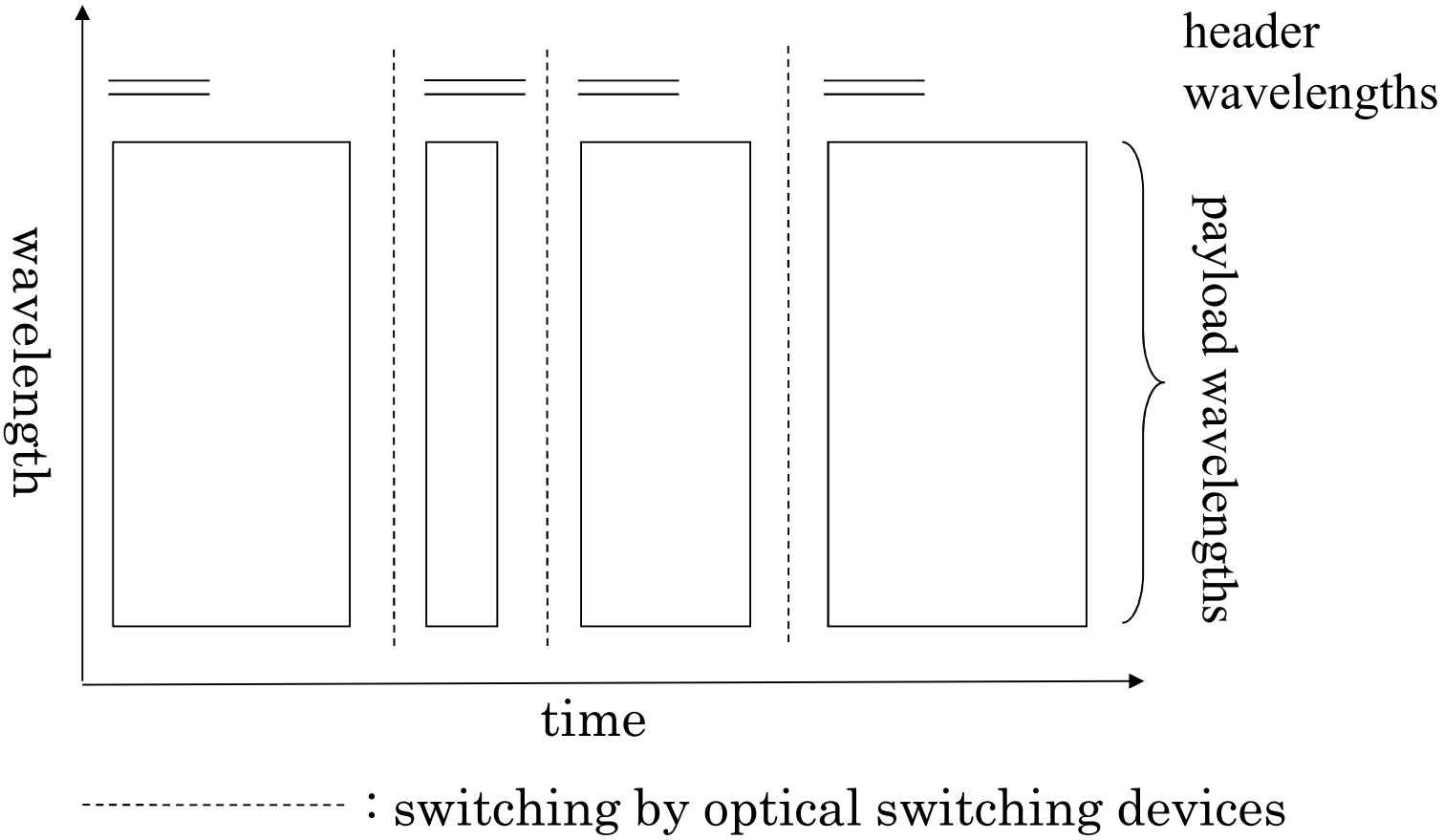
Packet Experts Knows

- 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.

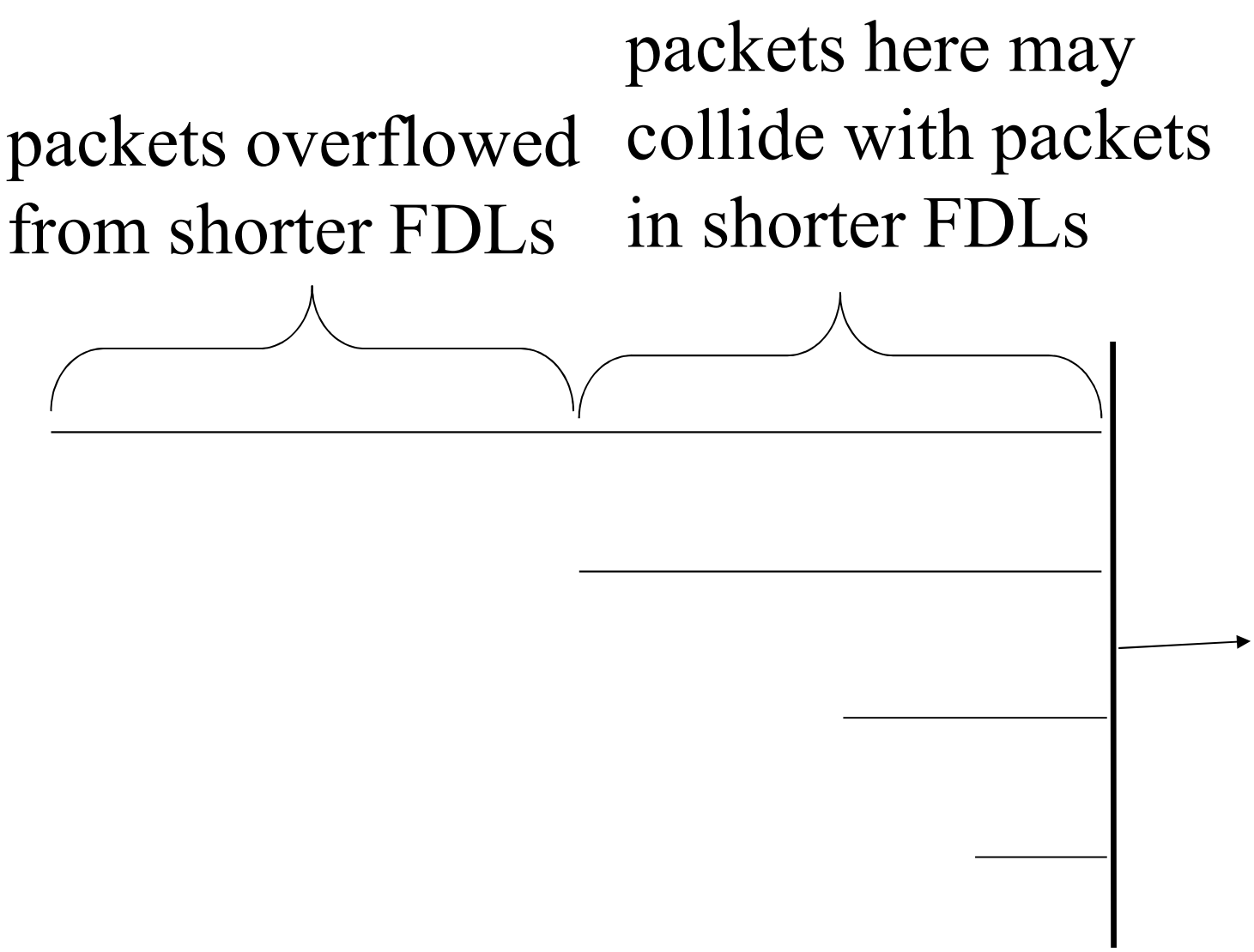
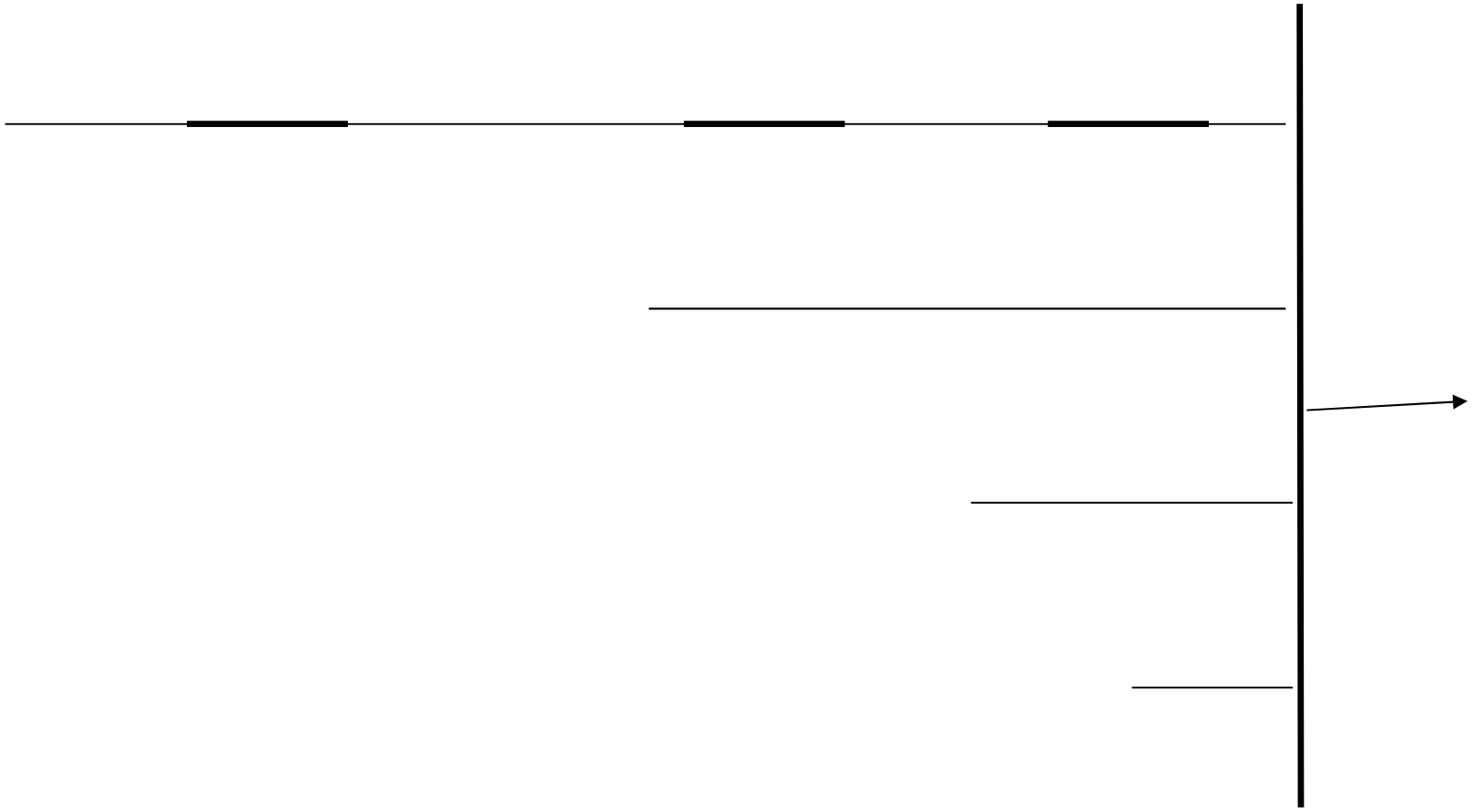


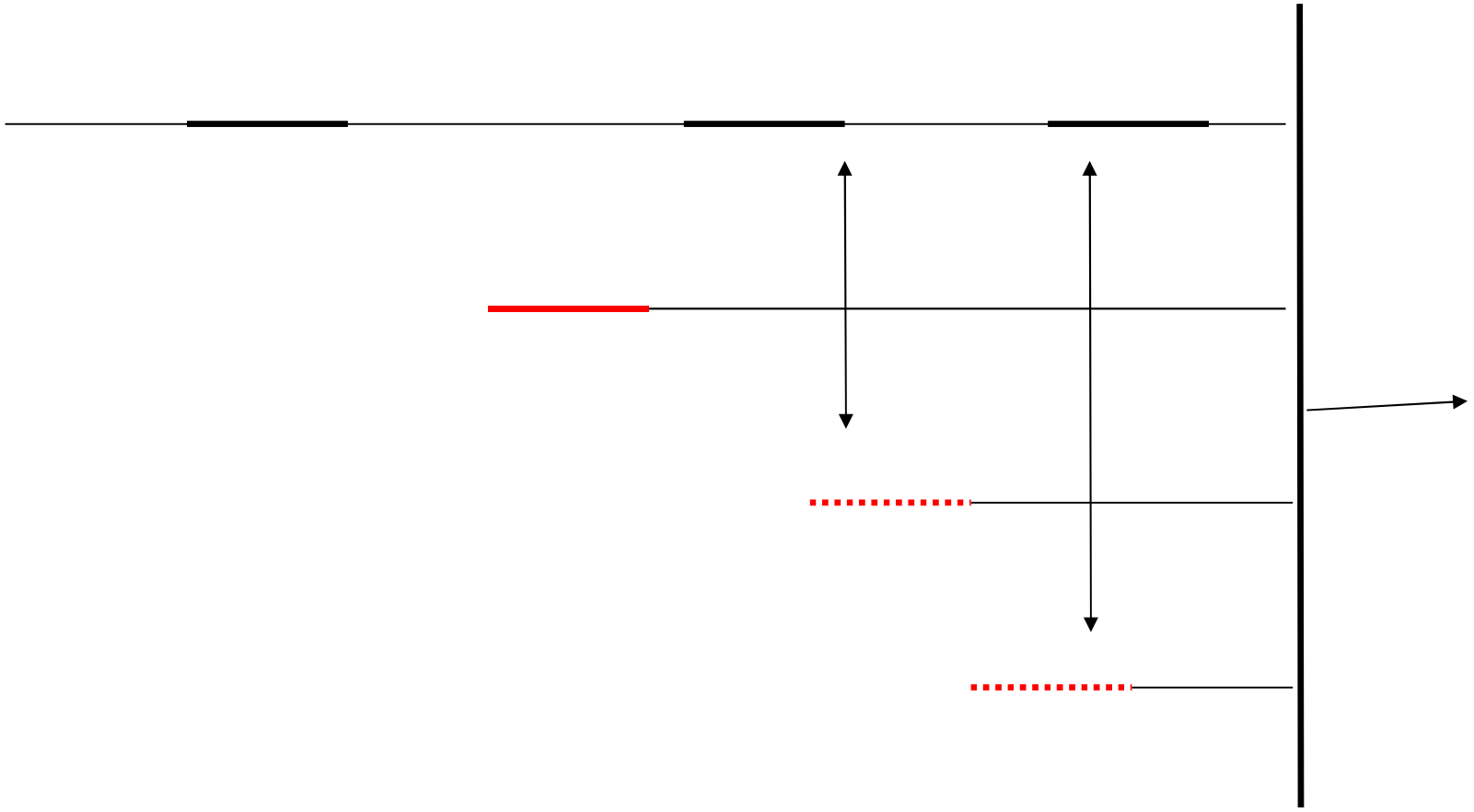
Fig. 5. FDLs with Lengths in Geometric Progression with Common Ratio of 2

Buffer Control (1)



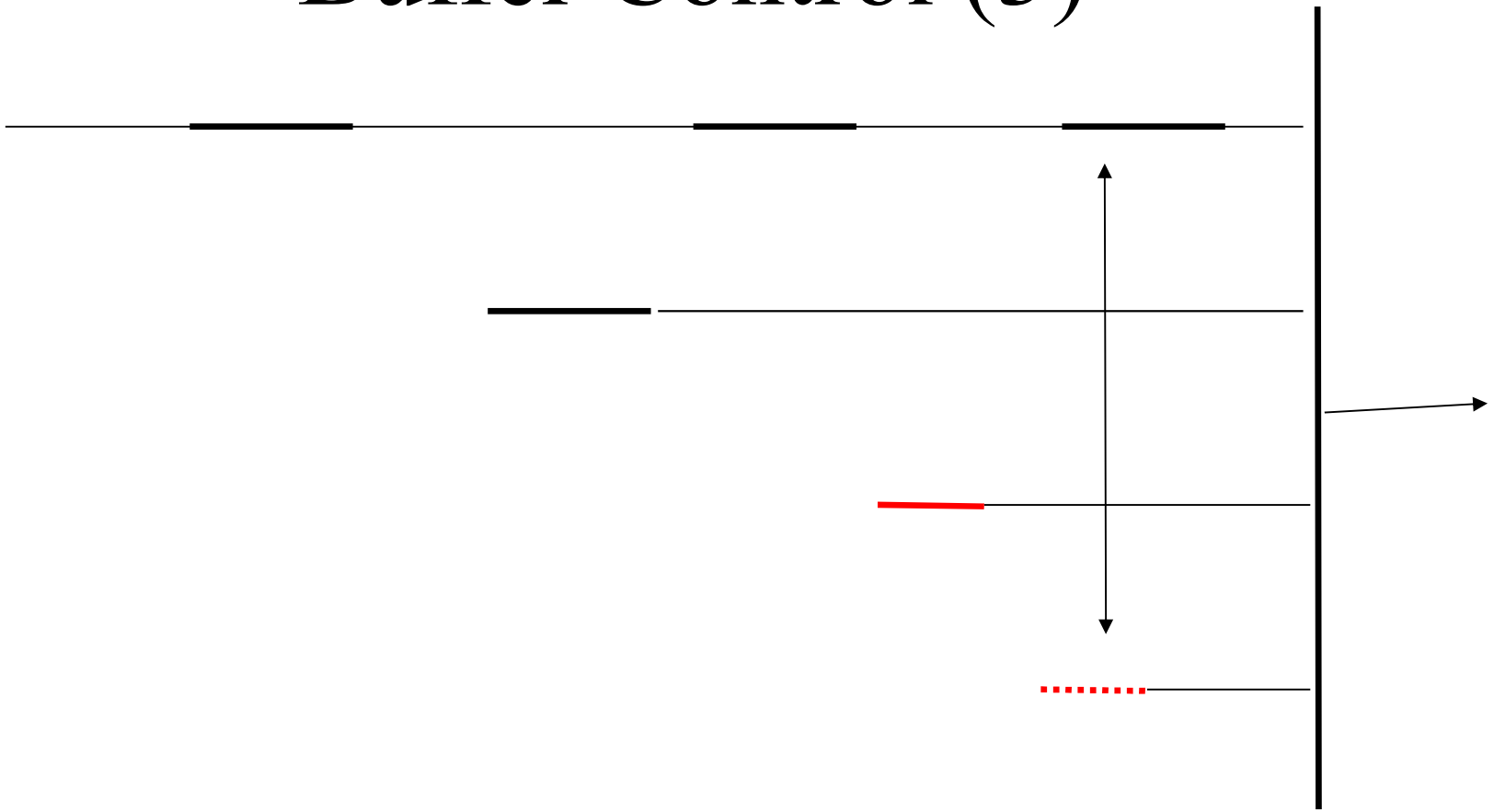
a) initial packet distribution

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

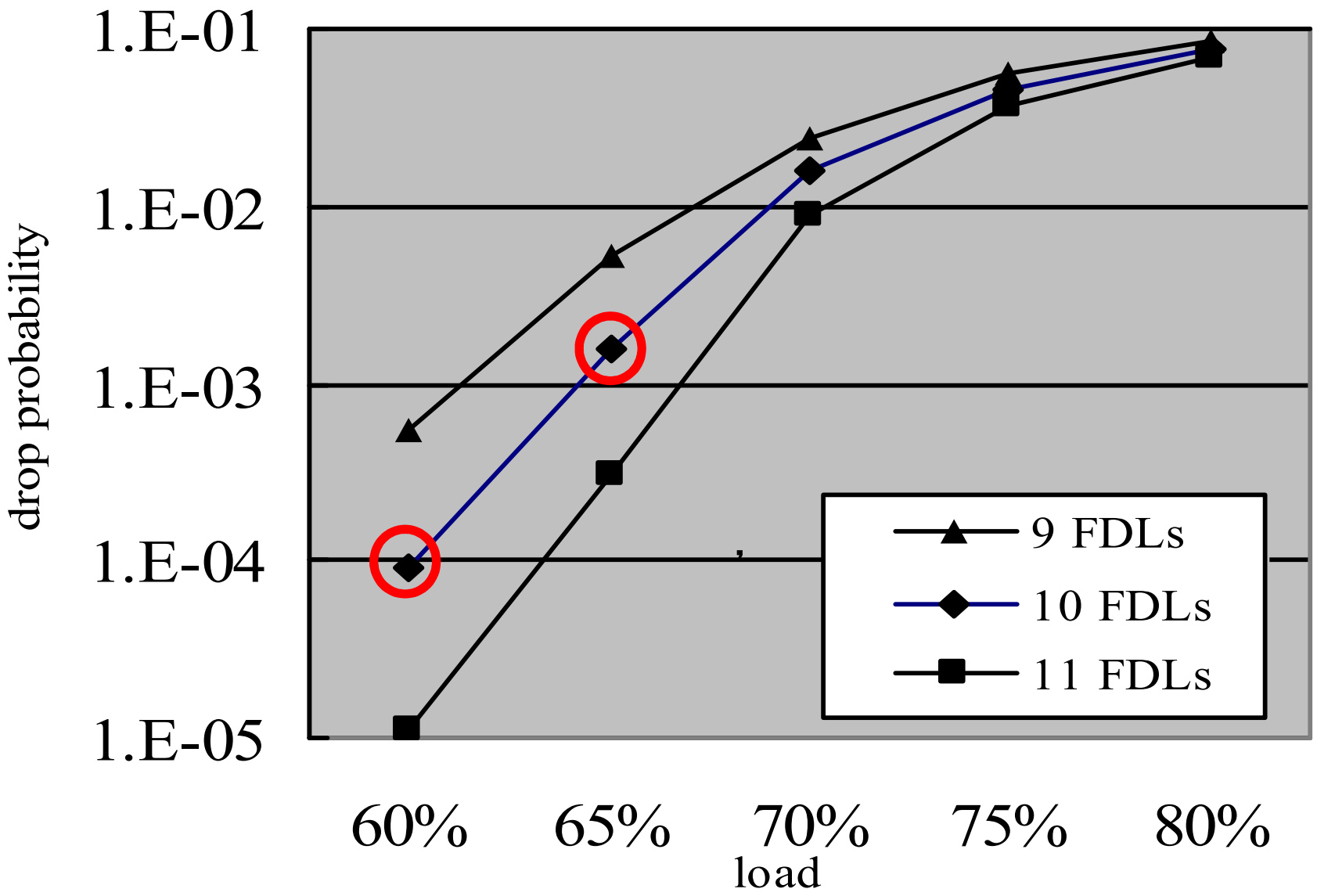
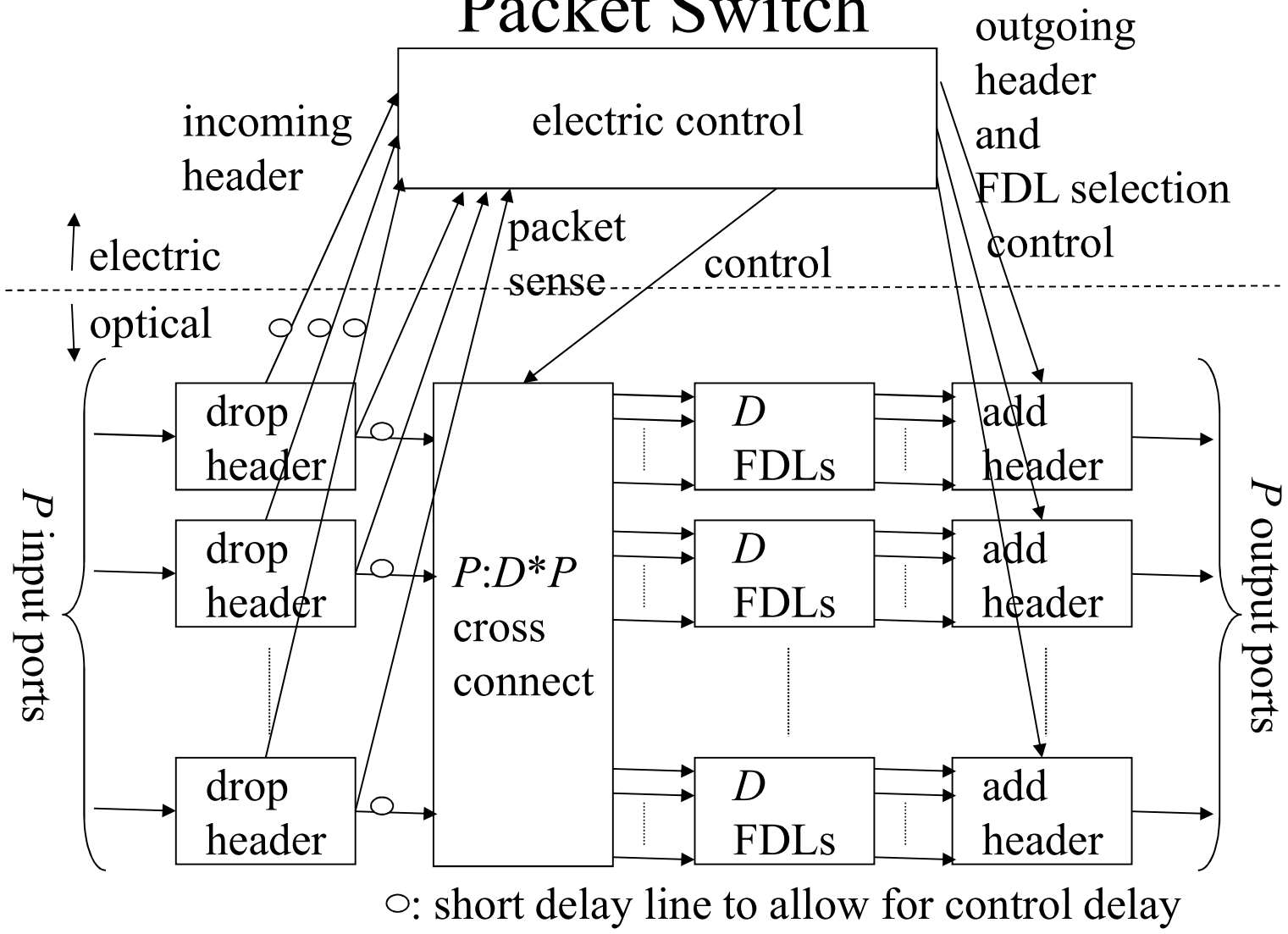
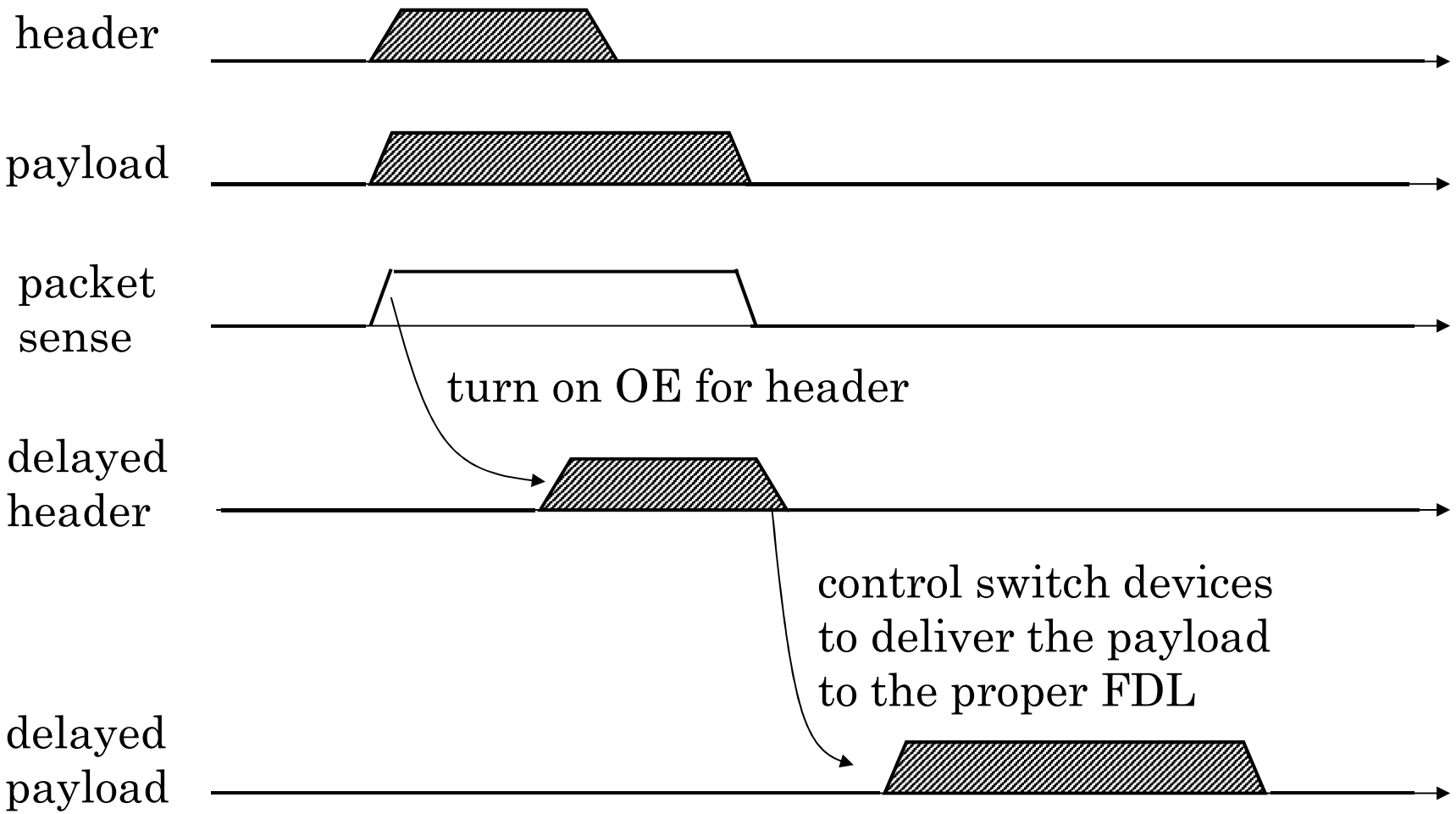


Fig. 6. Packet Drop Probability of FDL Buffers

A Micro Architecture of A Proposed Optical Packet Switch



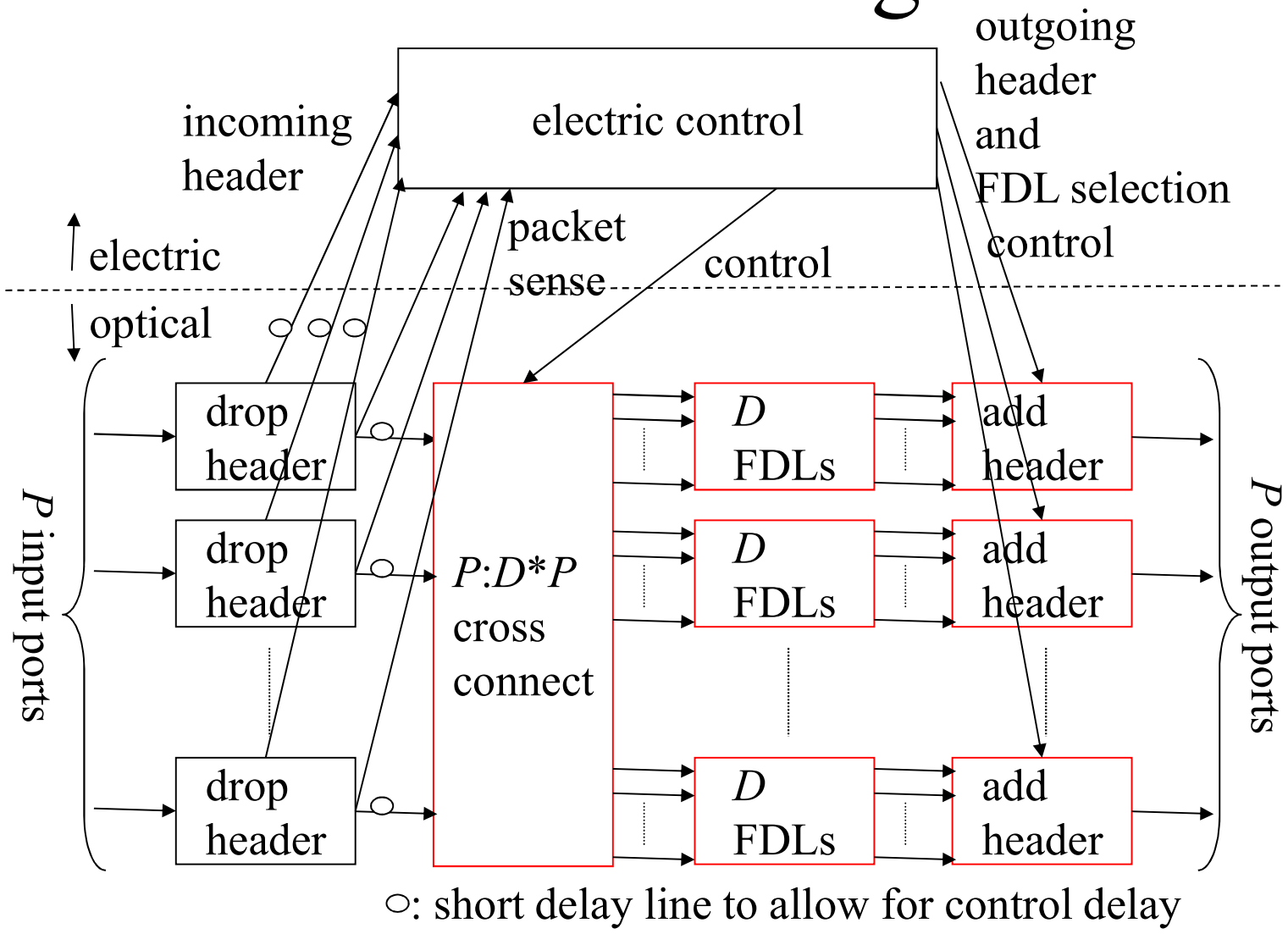
Relationships between Signals



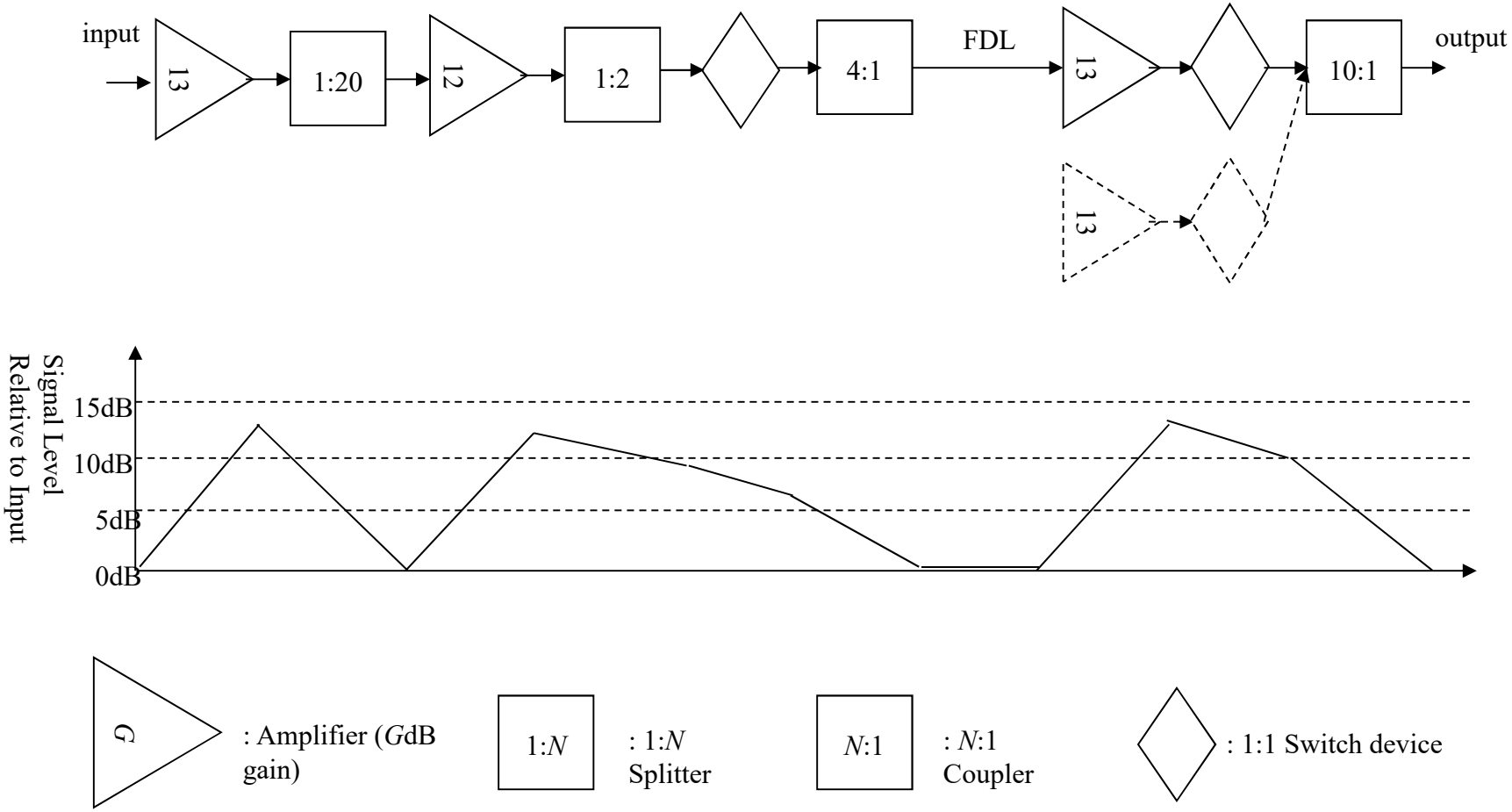
Power Consumed by Optical Packet Switches

- 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



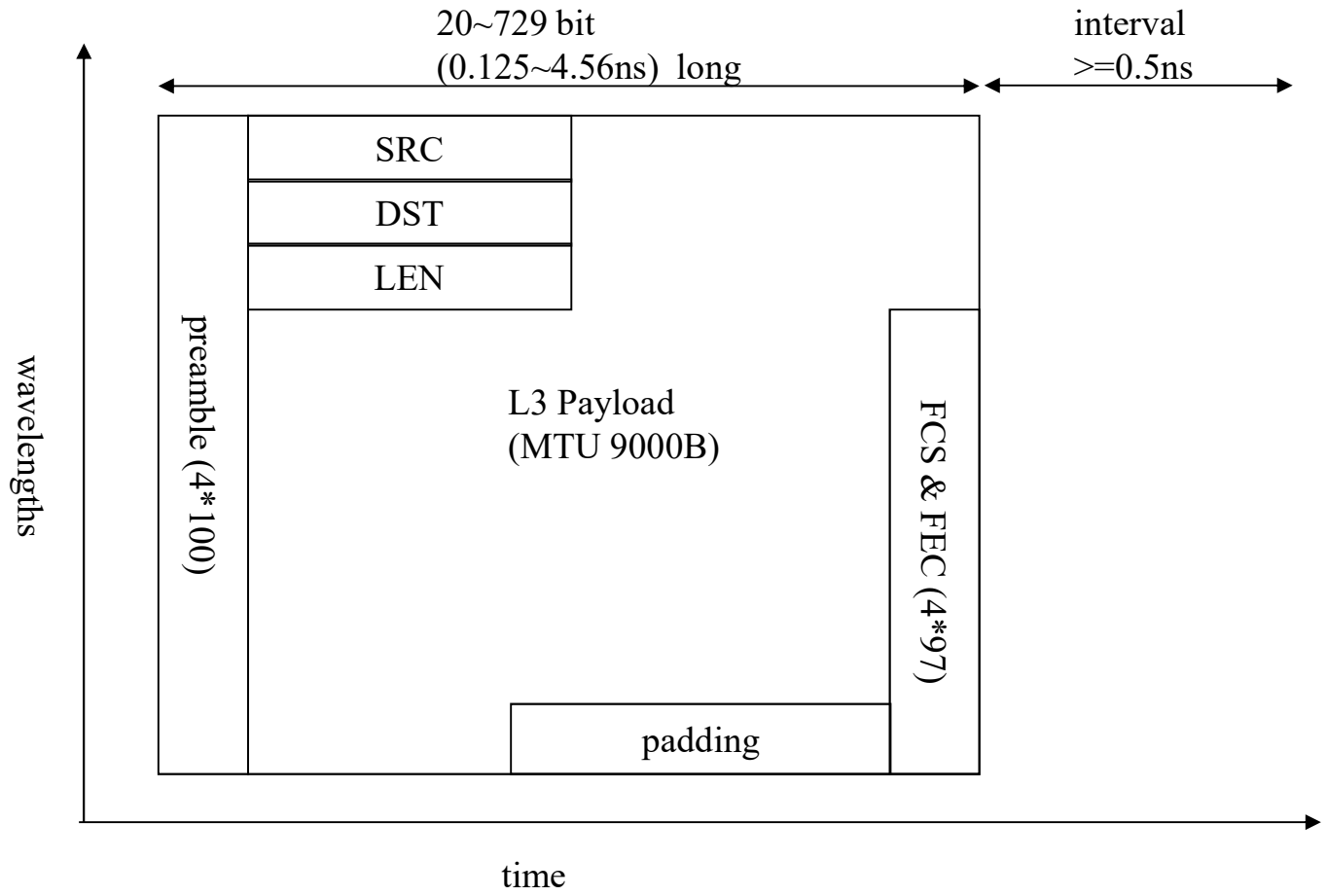
Estimating Power Consumption of An Optical Packet Switch

- 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
  - 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)

Assume photonic integration with control circuits except for 1:20 splitters

- 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

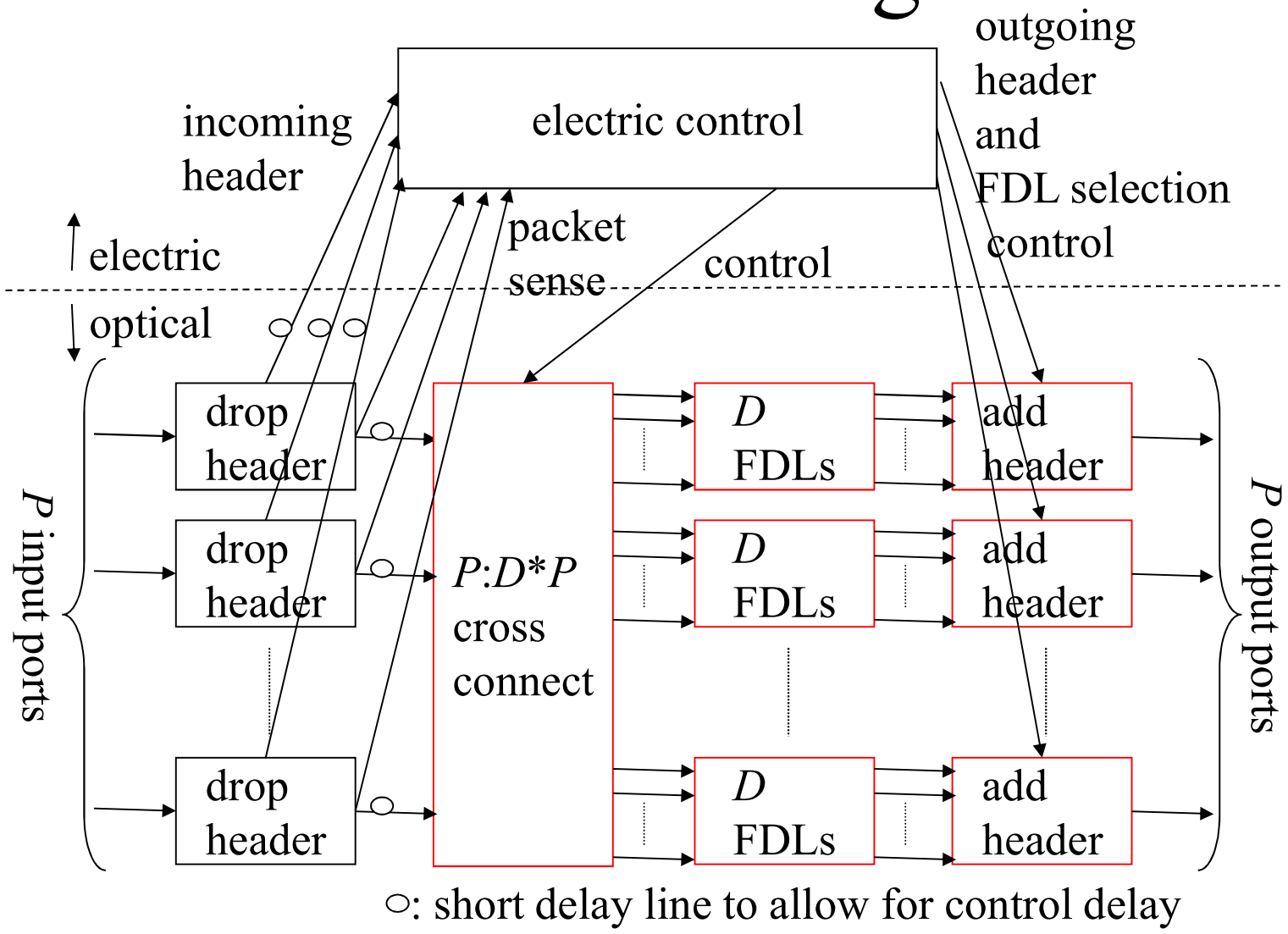
Isn't High Radix Better?

- 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