ハードウェアから見たイーサネット

名古屋大学 情報基盤センター
情報基盤ネットワーク研究部門
基盤ネットワーク研究グループ
嶋田 創

概要

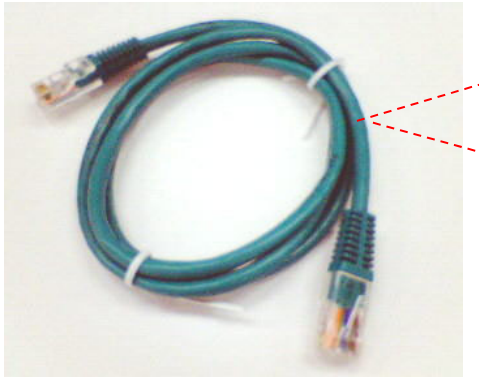
- 現在の主流: 1000BASE-T、1000BASE/10GBASE-(L/S)R
  - 派生規格: 2.5GBASE-T/5GBASE-T
- 過去の主流: 10BASE-T/100BASE-TX
- 将来の主流や現バックボーン: 25GBASE-(S/L)R/40GBASE-(S/L)R/100GBASE-(S/L/F)R/400GBASE-(S/L/F)R
  - ちょっとニッチさのある50GBASE/200GBASE
- これから?
  - 企画化はされた800GBASE、その次は1.6TBASE?
- 個人的な将来展望

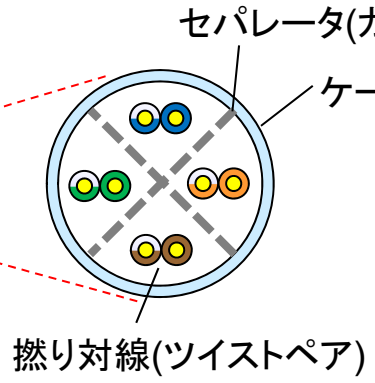
現在の皆さんがよく目にする有線イーサネットの規格

おそらく1000BASE-T

- シールド無しツイストペア(UTP: Unshielded Twist Pair)ケーブルを用いる
- 通信速度は1Gbpsで全二重通信
  - 対義語: 半二重通信
  - CSMA/CD(Carrier Sense Multiple Access Collision Detect)はもう使わない
- スイッチングハブ等を介して複数の機器を接続

UTPケーブル

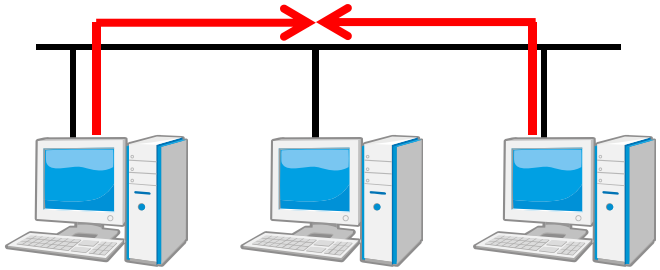




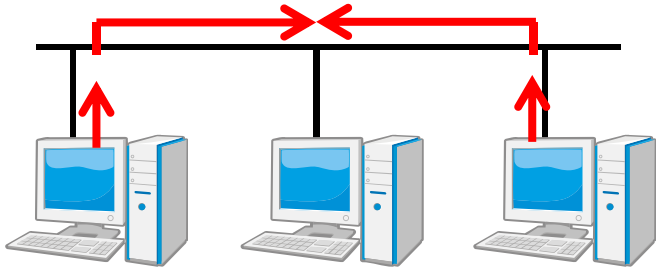
スイッチングハブ

CSMA/CDはなぜできなくなったのか?

- 最小のパケット(64バイト)の送信が完了するまでに衝突が検出されないとだめ
  - 10BASE: 5.12×10^{-5}秒 = 光速の信号(理想)で15360m
  - 100BASE: 5.12×10^{-6}秒 = 同1536m
  - 1000BASE: 5.12×10^{-7}秒 = 同153.6m ←ちょっと無理
- 1000BASEでも最短パケット長を大きくする対応策はあるが、あまり現実的でない
- 基本的にネットワークスイッチでパケットをバッファリング



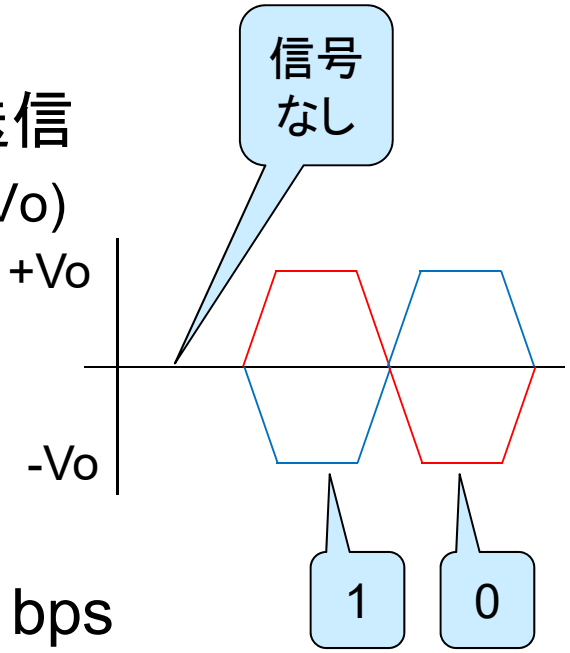
○衝突検出時にパケットがまだ送信中



×衝突検出時には次のパケットの送信が開始されている

電気信号としての1000BASE-T

- ツイストペア中を作動信号が流れる
  - 通常の信号の例: 0を+Vo、1を0Vで表現
  - 差動信号の例: 0を-Voと+Voの組、1を+Voと-Voの組で表現
- ベースクロック125MHz
- 4対のツイストペアを利用
- PAM5で1クロックあたり2ビット(シンボル)送信
  - 電圧を4段階に変更(+Vo, +0.5Vo, 0, -0.5Vo, -Vo)
  - 8b10bエンコーディングで2ビットを5値へ
- 各ツイスト・ペアで送信/受信を重畳
  - 同じ信号線に受信信号と送信信号を載せる
  - 受信電圧－出力中の電圧＝受信信号の電圧
- 通信速度: 125MHz x 2bit x 4pair = $1x10^9$ bps



10GBASE-T

- ベースクロック200MHz
- 4対のツイストペアを利用
- 1クロックあたり14ビット送信
  - 電圧を16段階に変更→4シンボル/クロック
  - 128 Double Square QAMを採用で7bit/2シンボル
- Low Density Parity Checkで1723b/2048b変換
- 各ツイストペアで送信/受信を重畳

- 通信速度: 200MHz x 14bit x 4pair x 1723/2048

$= 9.4 \times 10^9$ bps

(かなり回路面積と電力を食う仕様でなかなか安くなってくれない)

(現在10GBASE 約2.5W、5GBASE 約1.7W、2.5GBASE 約0.6Wほど[1])

[1] https://pc.watch.impress.co.jp/docs/news/event/1598156.html

2.5GBASE-Tと5GBASE-T

- 10GBASE-Tの技術をカテゴリ5e/カテゴリ6ケーブルで利用可能な所までデグレードさせたもの
  - 2016/9にIEEE 802.3bzとして承認された新しい規格
  - ベースクロックを100MHz(5GBASE)、50MHz(2.5GBASE)に低下
- 2つ合わせてマルチギガビット・イーサネット(略: mGig)とも呼ばれる
- 利点
  - すでに建物内の各部屋の情報コンセントまでの間に敷設されているカテゴリ5e/カテゴリ6ケーブルのまま1Gbps超えに対応できる
  - Power over Ethernet対応 (→10GBASE-Tも対応した)

  →実行速度数Gbpsの無線LAN規格である、802.11ax/beのアクセスポイントにUTPケーブル1or2本で電力と十分なバンド幅を供給

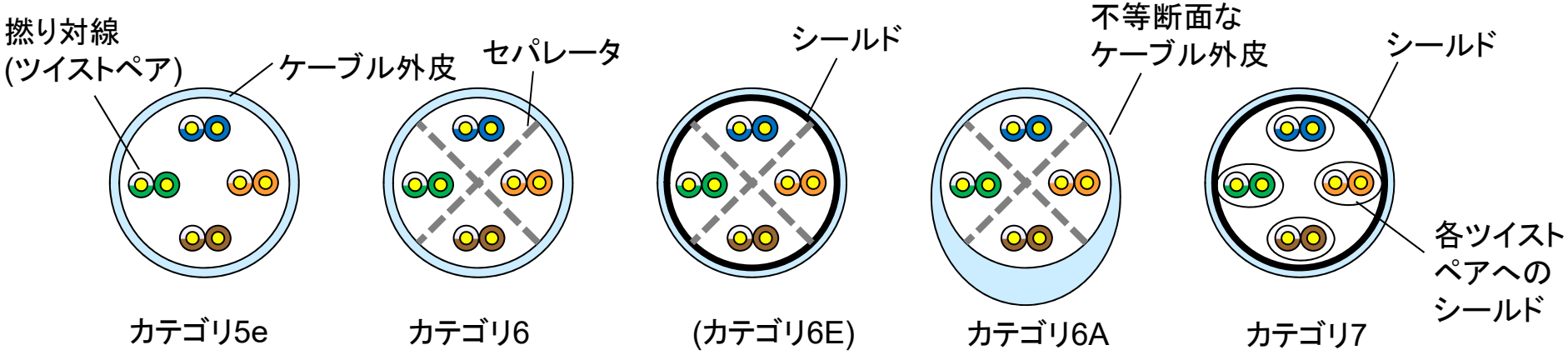
ツイストペアケーブルとその規格

- カテゴリ3: 16MHzまで
- カテゴリ5: 100MHzまで
- カテゴリ5e: 250MHzまで
- カテゴリ6: 250MHzまで
- カテゴリ6A: 500MHzまで
- (カテゴリ6E: 500MHzまで)
- カテゴリ7: 600MHzまで
- カテゴリ7A: 1000MHzまで
- カテゴリ8: 2000MHzまで

シールド無しツイストペア
(UTP: Unshielded Twist Pair)

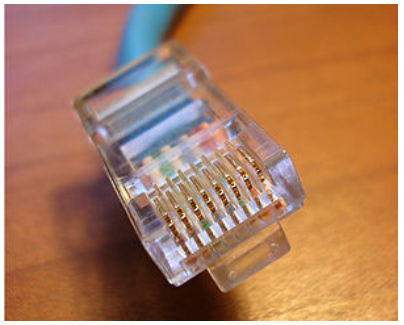
STPでないケーブルが「Cat 7相当」とか「消防署の方から来ました」的な名付けで売られている話あり

シールドつきツイストペア
(STP: Shielded Twist Pair)

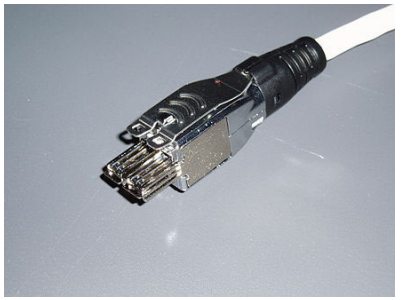


ツイストペアケーブル用コネクタとその規格

- ケーブルにシールドがあるか無いかで変わる
  - アースされていないシールドはただの電気的に浮いた導電体(帯電しまくり) →要アース接続
  - ネットワークスイッチのポート側もアース接続に対応している必要がある
    - シールドありのケーブルをシールドなし用のネットワークスイッチに接続すると…
- シールドなし: RJ-45
- シールドあり:
  - RJ-45互換: GG45, ARJ45
  - RJ-45非互換: TERA
- Cat 7はSTPなのでシールドありコネクタ必須(RJ45な偽Cat 7ケーブルが売られ…)



RJ-45コネクタ



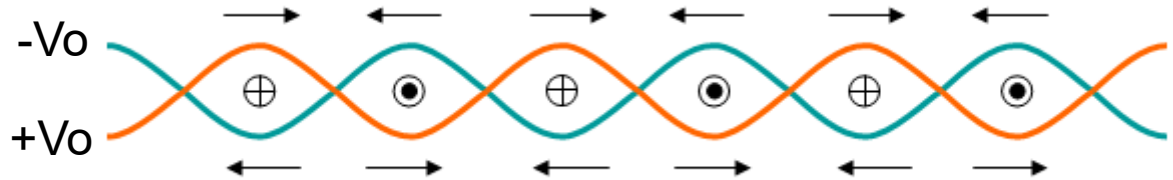
TERAコネクタ
(RJ-45と互換性無し)

ツイスト・ペア・ケーブルの特徴

ノイズに強い、ノイズを出さない

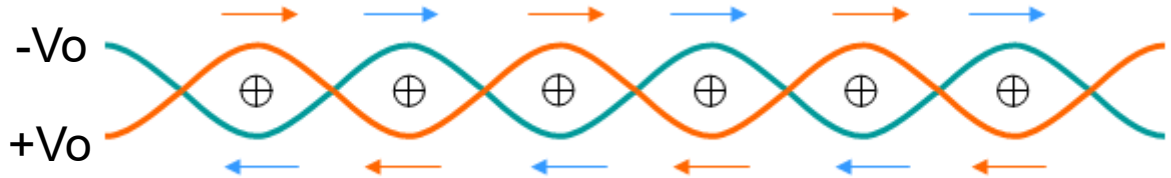
→安定して利用可能、他の機器の近くで利用可能

- 自身の発する磁束は打ち消し合う



磁束の向き: ⊕ 手前から奥　◉ 奥から手前

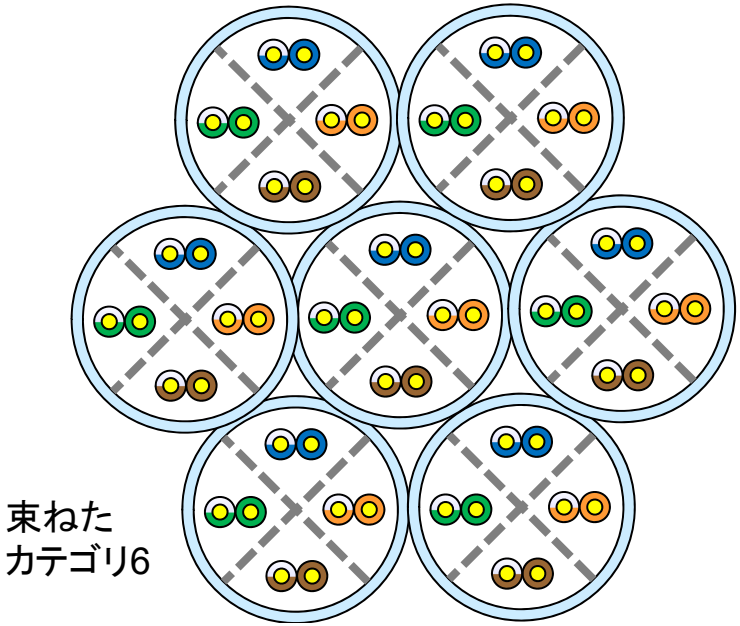
- 外部からの磁束による電流は打ち消し合う

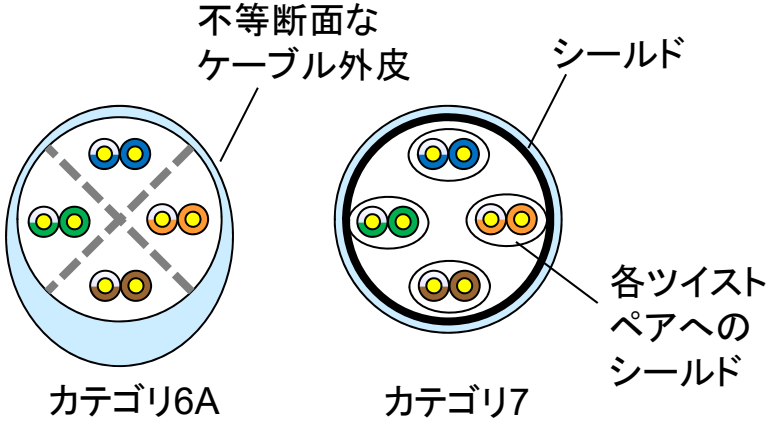


  - こちらが厳しくてUnshieldedでは限界が来ている
- さらに、作動ペアにより乗ったノイズを除去可能

ツイスト・ペア・ケーブルの限界

- サーバラックでぎゅっと束ねた時にケーブル外皮を通して隣り合うケーブルのツイストペアが近寄ってしまう
  - エイリアンクロストークによるノイズが信号に影響を及ぼす
- カテゴリ6Aでは、近寄りすぎないよう(かつ体積効率が良いように)外皮を不等断面にした
  - 根本的にはシールド付きのカテゴリ7にしないとだめ



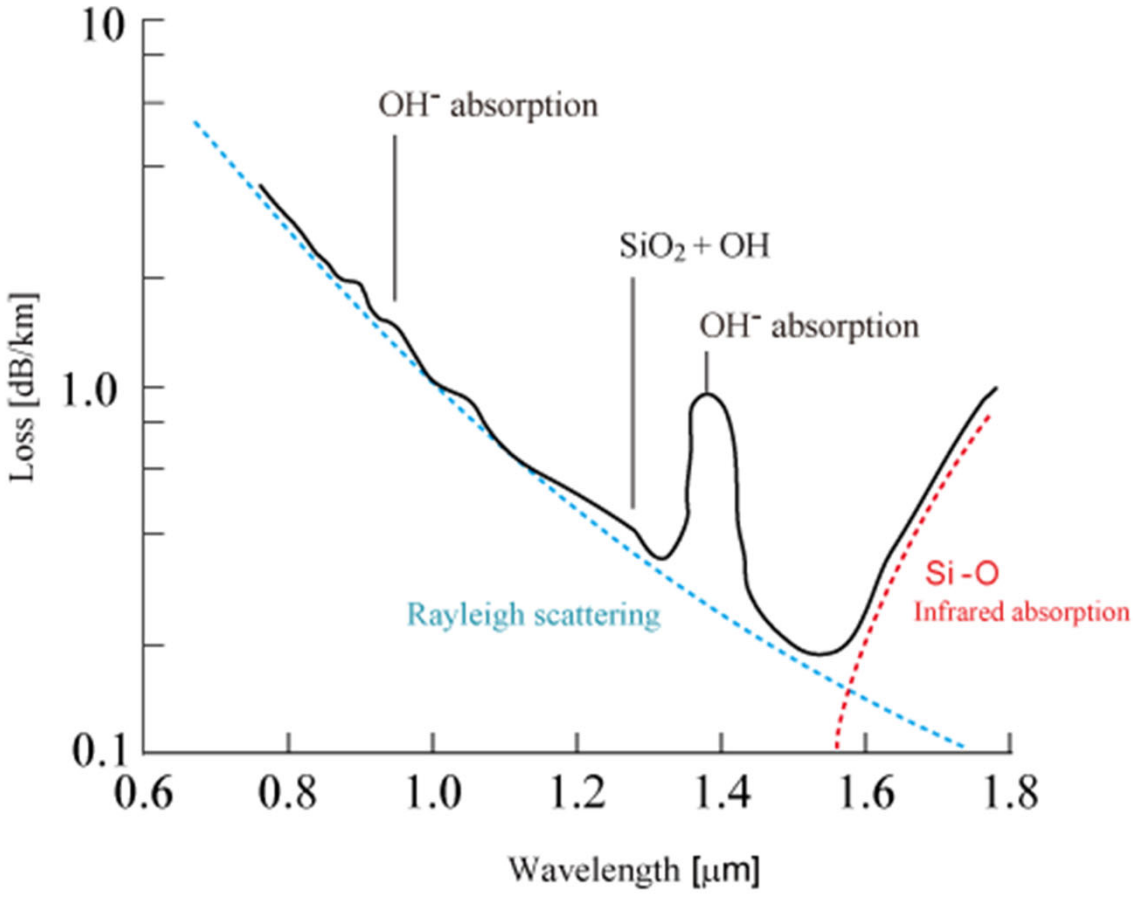


他の物理媒体を使う1000BASE規格

- 光ファイバを使うもの
  - 1000BASE-LX: 1350nmの長波長レーザー
    - シングルモードの光ファイバ: 長距離向け(～10km)
      - 高い、曲げに弱い、減衰が小さい
  - 1000BASE-SX: 850nmの短波長レーザー
    - マルチモードの光ファイバ: 短距離向け(～300m)
      - 安い、曲げに強い、減衰が大きい
  - 1000BASE-ER: 1550nmの長波長レーザー
    - シングルモードの光ファイバ: 超長距離向け(～40km)
- シールド付きツイストペアケーブルとか同軸ケーブルを使うものもあったが、1000BASE-Tの普及でほぼ絶滅
  - 同軸ケーブルはごく短距離(数m)レベルの接続には10Gbpsオーバでも多用されている

石英系光ファイバの波長と減衰率

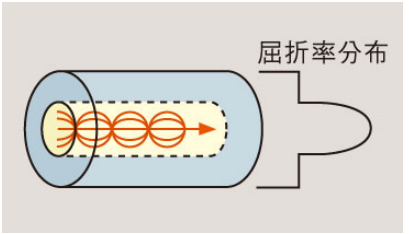
- 減衰率が小さい波長を長距離向けに利用



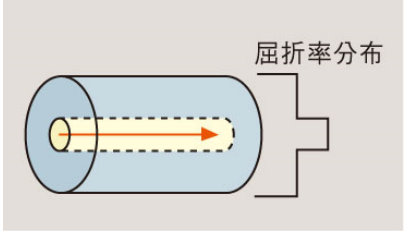
シングルモード光ファイバとマルチモード光ファイバ

- (グレーデッドインデクス)マルチモード光ファイバ
  - マルチモード: 光の伝送パスが複数
  - グレーデッドインデクス: 屈折率が段階的に変わる
  - かつてはステップドインデクスなマルチモード光ファイバもあった
    - シングルモードと同様に屈折率が階段上に変化
  - コア径が太い

    →製造が容易な上、コネクタ等の軸合わせが容易



- シングルモード光ファイバ
  - シングルモード: 光の伝送パスが1つ(反射はしつつ)
  - コア径が細い →製造が難しくて高価



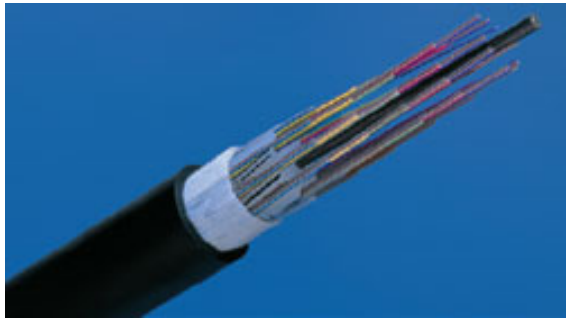
- 家庭用などではプラスチック製光ファイバなどもある

光ファイバのグレード

- シングルモードファイバ: 距離だけ
  - OS1: どの通信規格でも10km
  - OS2: どの通信規格でも200km
- マルチモードファイバ: 低い規格は通信速度高い規格で影響あるので注意
  - OM1: 10Gで33m、40G以上は無理
  - OM2: 10Gで82m、40G以上は無理
  - OM3: 10Gで300m、40G/100Gで100m
  - OM4: 10Gで550m、40G/100Gで150m
  - OM5: OM4をベースに波長多重に対応した物
    - 850nm～950nmで使うことを想定
    - 単一波長ではOM4とほぼ変わらない

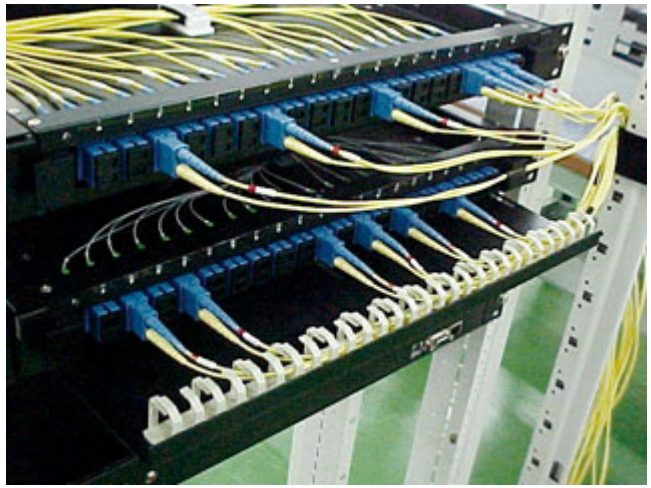
名大内の光ファイバ網

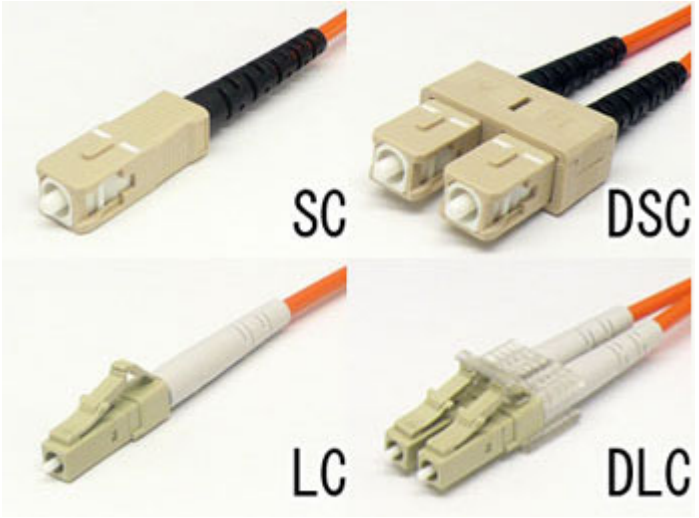
- テープスロット型ケーブルを学内の共同溝に敷設
  - 建物間の距離によってマルチモードとシングルモードを使い分け
  - 1990年代に一斉に敷設
    - 光ファイバの設計寿命は20年ぐらいのため、
      そろそろ劣化がやばい(使用不可芯線が…)
      →2022年度に基幹部を更新、他も更新予定
  - 相次ぐ建て増しなどで管路の容量も厳しい



光ファイバの引き込み

- 光ファイバの末端は光コネクタに整形される
  - 大抵は光パッチパネルに列の形で設置する
  - SC規格とLC規格が主流
    - 後述する光トランシーバ側も同様
  - パッチパネルはUTPケーブルでもよく利用される
- 通常のネットワーク構築では2本の光ファイバを組みで使う
  - 送信側と受信側で1本ずつ
  - 1本でWDM(波長分割多重)を使う方法もあるが、コスト的に2本使う方が安い
    - ただ、今は逆に既設の2本ペアのファイバを再利用する規格も出てきている





光ファイバのネットワークスイッチへの接続

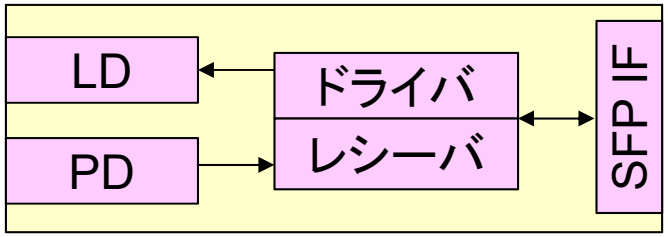
- パッチパネルから光パッチケーブルを介して接続
- 余談: 光パッチケーブルを介して別の光ファイバに乗り換えることもよくやる

最近の光ファイバ接続事情

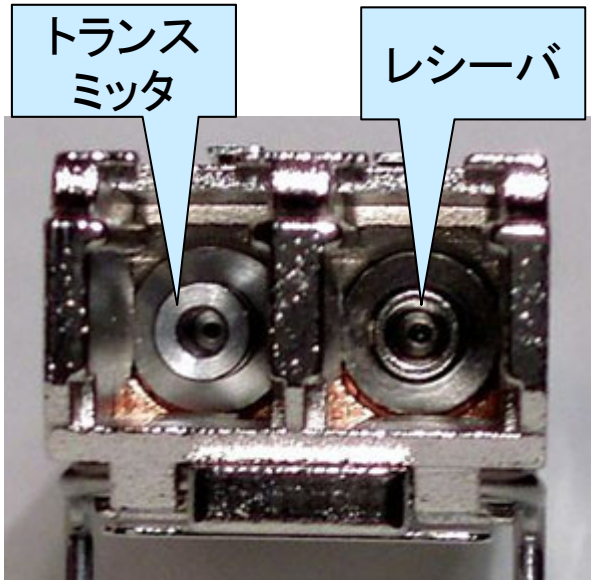
- 100G/400Gの光通信は光ケーブリングに対してかなりセンシティブ
- 光パッチケーブルを接続する時に、パッチパネル/光モジュールと光パッチケーブルの両方の端面をちゃんと清掃する
  - 微細なゴミがコアの接続面に挟まった時の信号レベル低下の影響大
  - コアの接続面が密着しないと、コアの端面での反射光がノイズに
- 光信号の出力が強すぎて受信側で信号の検知ができないエラーが出る話も(聞いた話)
  - 400Gだと基本的に複数波長の利用に加えてPAM4で1波長あたり4値を利用する
  - 長距離に対応した光モジュール(LRなど)を短距離で使うと起こり得る(?)
  - 送信側の光モジュールの送信出力を調整

光トランシーバ

- トランスミッタ: レーザーダイオード(LD)
  - LEDも用いられることはあるが、レーザーの方が波長が揃っていて好ましい
  - 駆動はドライバ回路を用いる
- レシーバ: フォトダイオード(PD)
  - 動作速度から使えるフォトダイオード構造は限定される
    - PIN-PD、アバランシェPD
  - 電流の変化をオペアンプで増幅して検出
    - ドライバ回路と含めてIC化

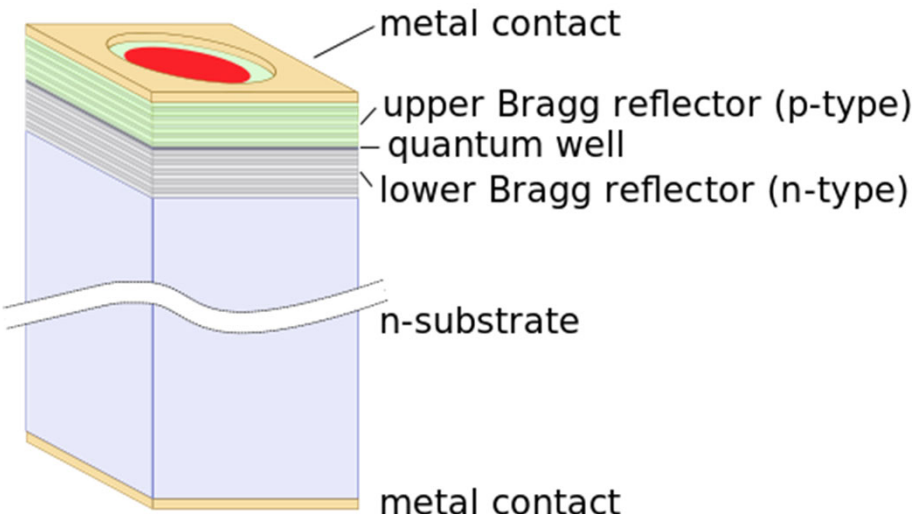


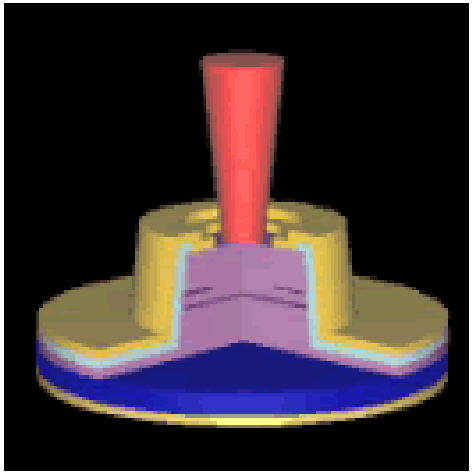
SFPモジュール内の
トランスミッタとレシーバ



面発光レーザーダイオードによる送信

- レーザーは垂直方向に発振するのが主流(VCSEL: Vertical Cavity Surface Emitted Laser)
  - 昔の半導体レーザは結晶面から横に発振していた
- ドライバ回路が発光に必要な電圧の変化を作成
  - 必要な電圧振幅を維持したまま高速化が非常に難しくボトルネック
  - 規格上で使われている最も高速な物は50Gbps/波長
    - これにPAM4変調(光の強度を4段階に)をかけて100Gbps/波長

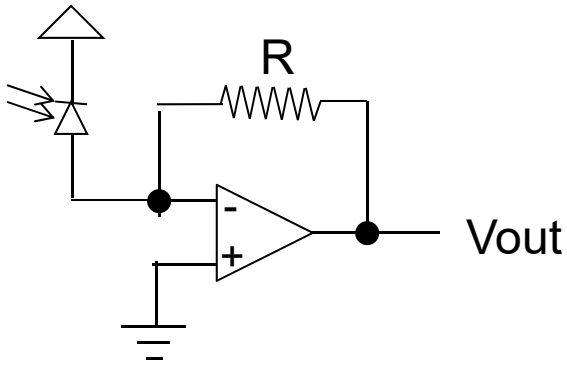




フォトダイオードによる受信

- 高速動作のために端子間容量や内部抵抗が低いフォトダイオードが必要
  - PIN-PD: PN接合に半導体のインシュレータを挟んだ構成
    - 逆バイアスをかけて空乏層を広げ、電子/ホールペアを増大可能
  - アバランシェPD: アバランシェ増幅により内部で光を増幅可能
- 微小な信号の変化をオペアンプで増幅した方が高速になる

→通常はオペアンプを併用



(光)トランシーバの規格(1/3)

基本的に、最終的にモジュールの縦横がコネクタの前投影面積になるまで新規格が作られる傾向

- SFP系
  - SFP: ほぼ1000BASE用
  - SFP+: ほぼ10GBASE用
    - 発展中のもの: XENPAK, X2, XFPなど
    - 2.5GBASE-T/5GBASE-T用の物も
  - SFP28: 25GBASE/50GBASE用
    - ここからGbps単位での物理層を名前につけるように
    - 56Gbpsの物理層を使うSFP56もちらほら

XENPAK

SFP/
SFP+/
SFP28

(光)トランシーバの規格(2/3)

- QSFP系
  - QSFP: ほぼ40GBASE用
    - 内部的にはSFP+を4本束ねているためQSFP(+)
  - QSFP28: ほぼ100GBASE用
    - こちらはまずQSFP28が主流になってからSFP28へ降りてきた
  - 発展中のもの: CFP, CFP2, CFP4
- 物理層は28Gbps, 56Gbps, 112Gbps(56GbpsのPAM4)と発展中
  - エラー訂正符号分を除くと、それぞれ25Gbps 50Gbps, 100Gbpsになる

QSFP/
QSFP28

CFP

(光)トランシーバの規格(3/3)

200Gbps/400Gbps/800Gbpsで光モジュール戦国時代化

- 後述のように、1信号が50-100Gbpsで上限 →信号を束ねて高バンド幅化 →配線数や消費電力で新たなモジュールへ
- 「そこそも高バンド幅化が進まない →実運用ではx00Gbpsをさらに束ねる →光モジュールの単価は重要」も原因そう
- 100Gbps超の光モジュール
  - QSFP-DD: -400Gbps
  - OSFP: -800Gbps
    - 内部的には8本(O = Octal)ためにQSFPより1回りサイズが大きくなっている
  - QSFP-DD800: -800Gbps

光トランシーバよもやま

- トランシーバモジュールに許容されたサイズと電力(発熱)で必要とされる性能を実現しなければならない点が難しい
  - XENPACKの電力許容量は9WだがSFP(+)は1W
- ドライバ/レシーバIC側はCMOSではなくバイポーラ等の高速動作に必要な構成
  - より高速にするためにシリコンではなくSiGeを使う場合もある
- 40G/100GBASE以上では複数波長を使うのでやっかい
  - 1波長あたりのエネルギーを落とさないとファイバ側が持たない
  - 微妙に異なる波長を使うので、基板側のノイズ対策がやっかい
    - 発振周波数が違う
    - 配線が平行することによるクロストーク
  - 400GBASEでは50Gbpsを8本(波長)使うのが最初は主流
    - マルチモードファイバを使う方では16本(波長)の規格もあった
    - 最近では100G(PAM4)を4本束ねる物の利用も増えてきている

40/100Gbpsの光ファイバ規格

- 40GBASEと100GBASEは広く使われている
  - 40GBASE-SR4: 10G x4、トランクケーブルマルチモード光ファイバ
  - 40GBASE-LR4: 10G x4のWDM、シングルモードファイバ
  - 100GBASE-SR4: 25G x4、トランクケーブルマルチモード光ファイバ
  - 100GBASE-LR4: 25G x4のWDM、シングルモード光ファイバ
  - (New)100GBASE-LR1: 400GBASE用1波長の活用、シングルモード
- トランクケーブルマルチモード光ファイバ
  - 複数波長の数だけ光ファイバを利用
  - 専用コネクタと合わせての利用が多い

12本の光ファイバを利用可能な
MPO12コネクタと12本の光ファイバ
(-SR4では8本のみ利用)

派生光ファイバ規格

- 25GBASE: 25G x1(100GBASEの光を1波長だけ利用)
  - 光モジュールのサイズもSFP+と同じなので高密度
- 50GBASE: 25G x2(100GBASEの光を2波長だけ利用)
- 正式な規格でないけど、メーカ独自で派生規格がちょこちょこある
  - 40GBASE-LR4 Lite
    - 到達距離が無印LR4の10kmから2kmや1kmへダウン
    - その代わり、光モジュールが低コスト
    - AristaやCiscoなどが提供
  - 40GBASE-SR BiDi
    - 1本のマルチモードファイバに異なる波長で送信/受信を重畳
    - 1波長で20Gbpsを実現

    →10GBASE-SRと同じ2本のマルチモードファイバで40GBASEを実現

400Gbpsの光ファイバ規格(1/2)

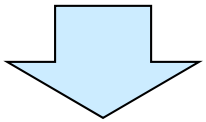
- 400GBASE(第1世代)
  - 400GBASE-SR16: 25G x16、マルチモードファイバ
  - 400GBASE-FR8: 50G x8、シングルモードファイバ
    - ただし、到達距離は2km
    - 一部メーカが出している40GBASE-LR4 Liteが規格化された感じ
  - 400GBASE-LR8: 50G x8、シングルモードファイバ
  - MII層でも50Gbpsまでの対応は目処がついている
- 400GBASE(第2世代)
  - 100Gを4本束ねるもの: 400GBASE-FR4/LR4
  - マルチモードファイバで50Gを8本束ねるもの: 400GBASE-SR8
  - マルチモードファイバで50Gx2波長を4本束ねる: 400GBASE-SR4.2
- この世代は市場には出ているが、まだ非常にコストパフォーマンスが悪い

400Gbpsの光ファイバ規格(2/2)

- 400Gの低コスト版(波長数を削減)で200GBASEも出てきた
  - 200GBASE-LR4/FR4/SR4
- 100GBASEにも新技術が降りてきた
  - マルチモード: 100GBASE-SR2、100GBASE-(e)SWDM4、100GBASE-SR BiDi
  - シングルモード: 100GBASE-LR1(10km)、100GBASE-FR1(2km)、100GBASE-DR1(500m)
- というか、100G以上から「少しでもコストを下げたい」要望からローカル規格も含めものすごく規格が乱立している

次世代の光ファイバ規格

- 800GBASEは2020/4に策定
  - 100G x8で800G
  - というか、100Gを束ねる400GBASEを2つ並べて実装
  - いずれ1波長200Gも出てくると思う
- 次は1.6TBASE
  - さすがに1波長200Gbpsが実現できないと厳しい

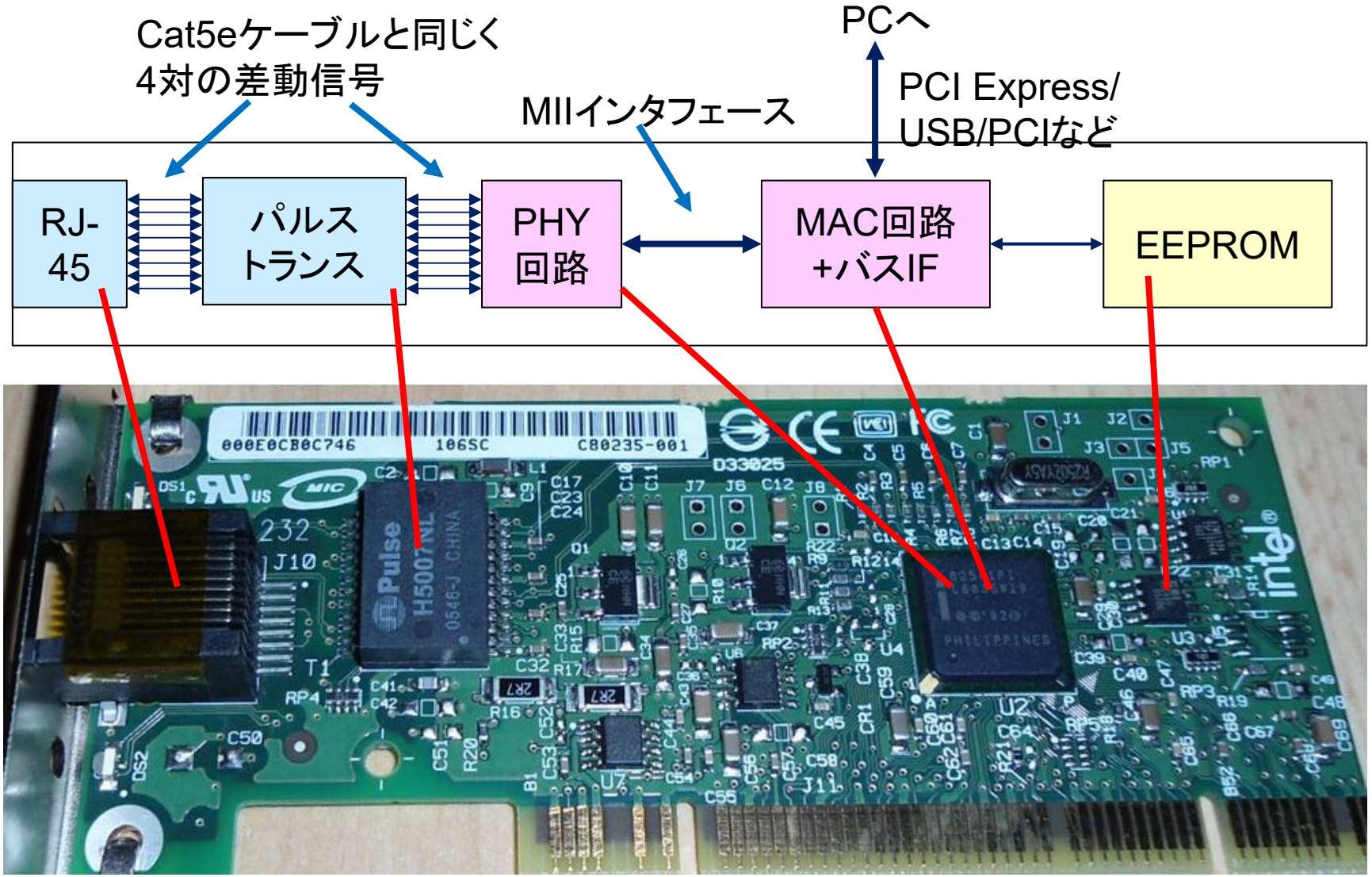


速度向上が苦しくなってきている様子が規格制定からも分かる

規格の表記について

- バンド幅: バンド幅+BASE
- 媒体や距離によるもの
  - T: ツイスト・ペア・ケーブルを使うもの
  - S: マルチモード光ファイバを用いた短距離(～300m)
  - L: シングルモード光ファイバを用いた長距離(～10km)
    - 最近ではF(～10km)やD(～500m)も
  - E: シングルモード光ファイバを用いた超長距離(～40km)
- エンコーディングによるもの
  - X: 4b/5bエンコーディング
  - R: 64b/66bエンコーディング
- 利用する波長数or光ファイバの組数: 末尾の数字
  - いずれかで多重化しているものが主流だが、最近は両方を組み合わせたSR4.2とかも

一般的なNIC(1000BASE-T)の構成

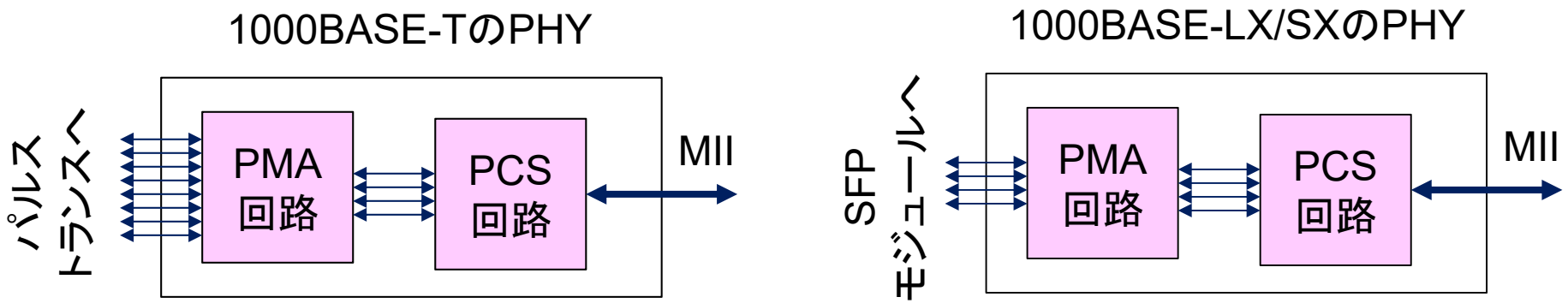


物理層(PHY)とメディアアクセス制御層(MAC)

- PHY: PHYsical layer
  - 信号のトランシーバ
  - アナログ回路の部分が多い
- MAC: Media Access Control layer
  - レイヤ2(の下位側)
  - デジタル回路な部分が多い
  - 最近では、他のデジタル回路と混載される
    - コンパニオンチップ、組み込み用チップ、など
- MII(Media Independent Interface)
  - PHYとMACを接続するインタフェース規格
  - GMII(GbE), XGMII(10GbE), SGMII, XAUIなどの派生規格も

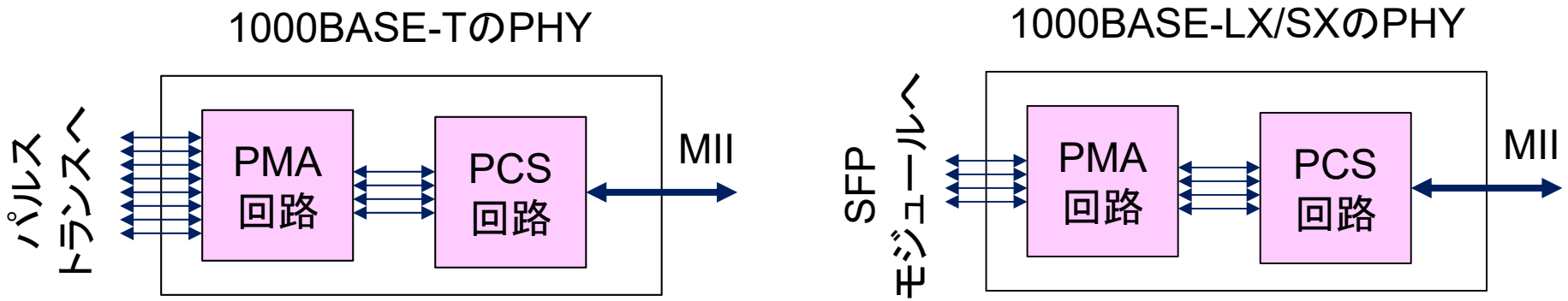
PHYの構成(1/2)

- PCS(Physical Coding Subsystem)回路
  - MIIの通信エンコーディングを物理層のエンコーディングに変更
    - 1000BASEでは8b/10b変換をする(→1.25Gbpsへ)
- PMA(Physical Medium Attachment)回路(1000BASE-T)
  - 実際に通信ケーブルに乗る信号に変換
    - 1000BASE-Tでは5値の信号の作動ペアx4
  - 1000BASE-Tでは送受信の全二重通信処理も
    - 受信電圧－送信電圧＝相手側が送信した信号
  - その他、通信ケーブル側からクロック信号の再生など



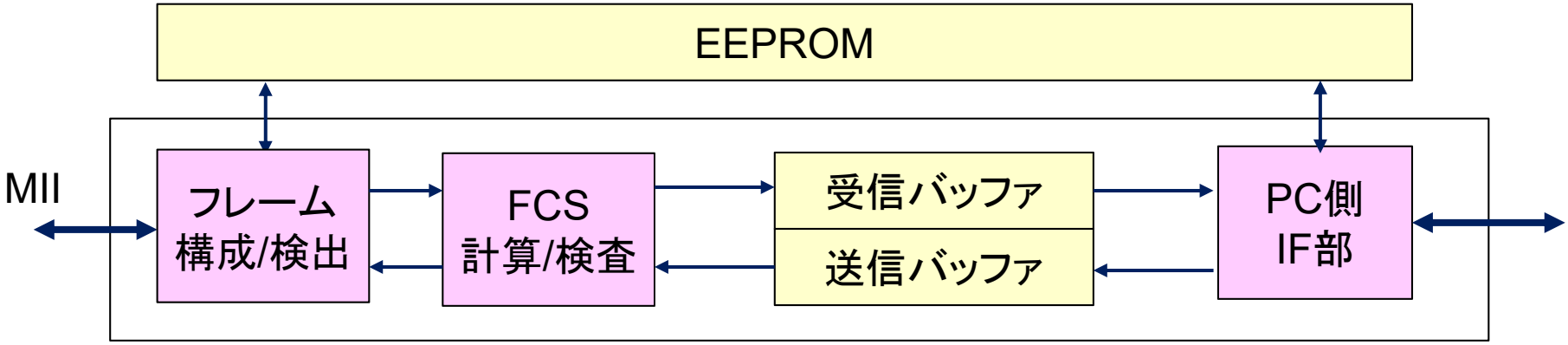
PHYの構成(2/2)

- PMA(Physical Medium Attachment)回路(1000BASE-LX/SX)
  - 実際に通信ケーブルに乗る信号に変換
    - 1.25Gbpsのシリアル信号
  - さらに、PMD(Physical Medium Dependent)回路(=SFPモジュールなどの光トランシーバ)で光信号に変換
- その他、物理層で発見したエラーのMAC層への通知など
  - MDIO(Management Data Input Output)という規格で通信
    - MIIに含まれています



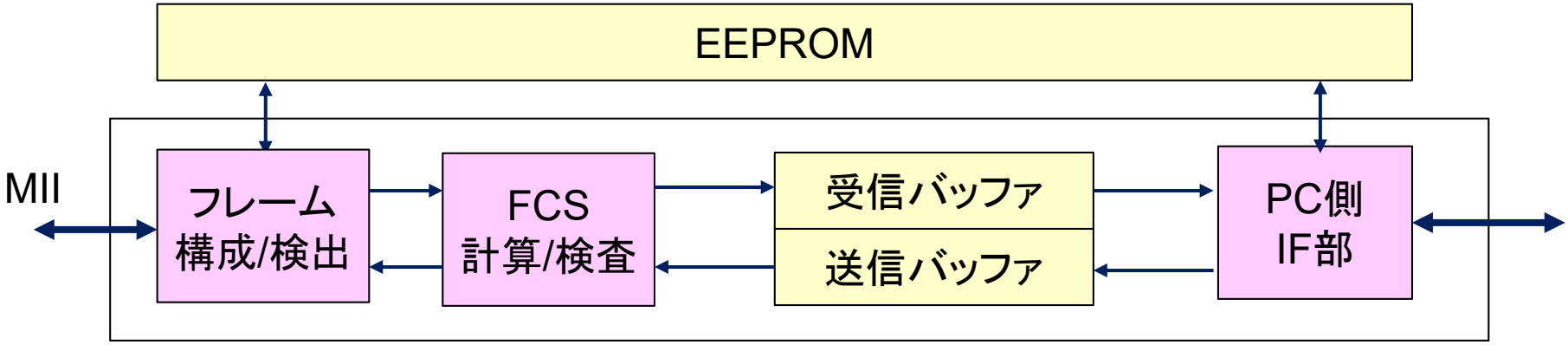
MAC層の構成(1/2)

- フレーム構成/検出
  - 送信時: プリアンプル、MACアドレス、FCSを付加してイーサネットフレームを構成
  - 受信時
    - プリアンプルでフレーム検出
    - 受信時のエラーチェック、送信時のデータを負荷
  - 受信時のMACアドレスの確認、送信時の付加
    - MACアドレスはEEPROMに書かれている



MAC層の構成(2/2)

- FCS(Frame Check Sequence)計算/検査
  - 送信時: ペイロードデータからのFCSの計算と付加
  - 受信時: 受信したペイロードデータとFCSからエラーチェック
- 送信/受信バッファ
  - 安いMAC層チップだとケチられていることも
    →PC側の負荷大、再送信による実行転送レート低下
- PC側インタフェース: PCI Express, USB, など



MAC層を触ってみる

というか、物理層と違って触る必要性が出てくることが多い

- NICのデバイスドライバを書く
  - メーカから仕様書が出ていますので、それを見て処理を考えて書く
- 組み込みプロセッサ内蔵のMAC層のデバイスドライバを書く
- FPGAにMAC層を実装する
  - MAC層がIPコアとして提供されている
    - Altera Triple Speed Ethernet, Xilinx Tri-Mode Ethernet MAC, など
    - 送受信バッファやMDIOのサポートをカスタマイズ可能
      - 論理セルの使用量に影響が出る
  - 頑って、MAC層を使わずに直接物理層と通信するのもあり

MIIで共通の処理はライブラリ等があったりする

MII

- 10BASE/100BASE時代のPHY-MAC間の接続信号線規格
  - 当時はまだまだデジタル回路/アナログ回路混在は難しい時代
    →必然的にMACとPHYは別チップになっていた
  - もちろん、MIIを使わない実装もあった
- 4bit幅のデータ信号線と各種制御信号線から構成
- 動作基準クロックはPHY側から供給される
- 通信速度
  - 10BASE: 動作基準クロック2.5MHz × 4bit = 10Mbps
  - 100BASE: 動作基準クロック25MHz × 4bit = 100Mbps

MIIの派生(1/4)

- GMII(Gigabit MII): GbE用MII
  - データ信号線が4bit→8bitに増加
  - 動作基準クロック周波数が25MHz->125MHzに増加
  - 通信速度: 125MHz × 8bit = 1Gbps
  - データをDouble Data Rateで転送して信号線数を4bitにしたRGMIIもあり
- SGMII(Serial GMII): シリアル信号のGMII
  - データ信号線を8bit→1bitに削減
  - 通信ケーブルもシリアル通信な光ファイバと相性が良い
  - ただし、1本の信号線で1Gbpsの通信を行うので、普通のレベル論理が使えないことも
    - LVDSなどの作動ペア信号線を用いたり

MIIの派生(2/4)

- XGMII(eXtended GMII): 10GBASE用MII
  - 動作基準クロック: 312.5MHz
  - データ信号線幅: 32bit
  - 通信速度: 312.5MHz × 32bit = 10Gbps
  - 信号線が多い: 72本
    - データ32bit、コントロール4bitを全二重通信で2倍
  - さらに信号線が多い実装も: 136本
    - データ信号線を64bitとし、動作基準クロックを156.25MHzへ
    - データ64bit、コントロール4bitを全二重通信で2倍
    - FPGAによっては、312.5MHzに回路がついていけないため

MIIの派生(3/4)

- XAUI: XGMIIのシリアル版
  - XGMIIは信号線が72本と多くて嬉しくない
  - データ信号線8bit(+コントロール信号線1本)を1本のシリアル信号に
    →データ信号線とコントロール信号線をまとめて削減
    - データ信号2.5Gbps(8b/10b変換して3.125Gbps)+コントロール信号
  - 4本(4対)の信号線で片方向の信号を転送
    - 全二重通信をするので、実際は8本(8対)
- SFI: XAUIのデータ信号線本数削減版
  - 10Gbpsのデータ信号線1本(1対)
    - 最近のFPGAの20Gbps超の高速IOを利用して接続可
  - ほぼSFP+規格モジュール用

MIIの派生(4/4)

- XLAUI: 40GBASE用のSFI
  - 10Gbpsの信号線4対
- CAUI: 100GBASE-SR10用のSFI
  - 10Gbpsの信号線10対
- CAUI-4: 100GBASE-LR4/SR4用のSFI
  - 25Gbpsの信号線4対

ネットワーク規格の高速化 (高バンド幅化)

- 10BASE-T(IEEE 802.3i): 1990年
  - 10BASE-5(IEEE 802.3): 1983年
  - 10BASE-F(IEEE 802.3j): 1993年
- 100BASE-TX(IEEE 802.3u): 1995年
  - 100BASE-FX(IEEE 802.3u): 1995年
- 1000BASE-T(IEEE 802.3ab): 1999年
  - 1000BASE-SX(IEEE 802.3z): 1998年
- 10GBASE-T(IEEE 802.3an): 2007年
  - 10GBASE-SR(IEEE 802.3ae): 2003年
- 100GBASE-SR4(IEEE 802.3ba): 2010年
- 400GBASE-SR16(IEEE 802.3bs): 2017年

おおむね、5年間で10倍の速度向上

10BASE-T

- ベース・クロック10MHz
- カテゴリ3のUTPを利用
  - 2対しかツイスト・ペアが無いので細かった
- 2対のツイスト・ペアを利用して送信
- 送信/受信をツイスト・ペアで分離した全二重を採用
  - つまり、1対のツイスト・ペアで10Mbpsの送信or受信
- 当時はリピータ・ハブが主流

→ CSMA/CD(Carrier Sense Multiple Access Collision Detect)による衝突検出

  - パケット送信時間>衝突検出時間だから実現可能

- 通信速度: 10MHz x 1対 = $1x10^7$ bps

100BASE-TX

- ベース・クロック125MHz
- 2対のツイスト・ペアを利用
  - 送信に1対、受信に1対
- 4b5bエンコーディングでエラー耐性を確保
- +Vo, 0, -Voの3つの電圧を使うが、データ圧縮には使っていない(MLT-3)
- 引き続き、CSMA/CDを利用

- 通信速度: 125MHz x 4/5bit x 1対 = $1x10^8$ bps

25GBASE-T/40GBASE-T

- 2016/6に標準化
- 10GBASEの仕様のまま、周波数を500MHz(25GBASE)と800MHz(40GBASE)に上げたもの
- ケーブルはCat8専用、コネクタもRJ-45から変更(GG45)
  - RJ-45だと、コネクタ部でほどかれたツイストペアがシールドされていない
- 長さも30mまで →ほぼデータセンター専用
  - 「8m」という噂もあったので、これでもましになった

40GBASE-T以降のメタルケーブルの展望

- 現状の電圧論理ではいろいろ厳しいと思う
- 既存の物とは互換性を捨てればもう少しいけるかも
  - 作動論理の電圧振幅を小さくする
- ただし、電圧小さくするとノイズ耐性が小さくなる

- 半導体製造技術進歩停滞(密度は上がるけど電力は減りにくい)で1bitのデータの転送エネルギーは減りにくくなっている
  - というか、光モジュールでも厳しくなっている
- 個人的には、40GBASEから先は光しかないのではと思う
  - というか、カテゴリ8ケーブルもお高いので、自分で組むならば10GBASE-Tより先は光で

Power over Ethernet(PoE) (1/2)

- UTPケーブルで通信と電力供給の相応を行なう規格(IEEE 802.3af)
- 受電側で最大12.95Wまでの電力を利用可能
  - 送電側は途中の電圧降下を見越して15.4Wを出力する必要あり
  - UTPに-48Vの電圧を載せる(信号の電圧は-48 +- Vo [V])になる
- 無線LANアクセスポイントやネットワークカメラへの接続が本の線で済む
- PoEインジェクタやPoEネットワークスイッチで電力を重畳
  - PoEネットワークスイッチは全ポートにフルの電力を供給できないものもある
    - 例: Catalyst 2960L 8ポートPoE対応モデル60Wまで
    - 全ポートにフルで電力を供給しようとすると高価になるため

Power over Ethernet(PoE) (2/2)

- IEEE 802.3at(PoE+)
  - 受電側で最大25.5Wまで利用可能な形に拡張(送信側は30W送信)
  - 受電側と送電側で0.1W単位での電力ネゴシエーションが可能
  - 特に、最近発展の著しい無線LANアクセスポイントで利用が多い
- IEEE 802.3bt(PoE++)
  - Type 3は送信側60W、受信側51W
  - Type 4は送信側100W、受信側71W
- 802.3afだとネゴシエーションが弱くて、適当な抵抗で線が導通すると電圧が加えられたりする可能性があるので注意
  - PoE対応スイッチだとPoEのenable/disableもスイッチ側のコマンドで制御できるので便利
- 802.3cg(10BASE-T1)という1km先に50Wの電力と10Mbpsの通信を1対のツイストペアで供給する規格が策定中

PoEと無線LANアクセスポイント

- 現状の無線LANアクセスポイントは消費電力が増加傾向
  - 2.4GHzと5GHzのデュアルバンドは当たり前
  - MIMO数も増えて重畳された信号の分離処理や行列の係数もMIMO数(の2乗)に比例

→IEEE 802.3afの12.95Wでは演算処理の電力が足りず、802.3at必須

- 最近の業務用802.11acや802.11axフルスペック対応無線LANアクセスポイントだと20W強を消費
  - 電力不足(802.3af)だと機能が制限されるアクセスポイントもある
  - 電力不足で性能が制限される例: MIMO数が減る、有線イーサネットの口が1口のみになる、オプション接続用USBポートが利用不能
  - トライバンド機種だと25W以上(PoE+ぎりぎり)利用する機種も
    - 6GHz帯を利用するWi-Fi 6EやWi-Fi 7?(802.11be)ではPoE++必須?

組み込み用100Mイーサネット物理層

イーサネットが組み込み機器に降りて来て、低コスト(+低重量)な物理層の要求が出てきた

- 100BASE-T1(802.3bw, 2015):
  - 1対のUTPで15mまで、STP(40m)もあり
  - 1000BASE-Tで使われた技術を利用
    - PAM3(+Vdd, GND, -Vdd)の3値を効果的に使ってクロックを2/3に下げる
    - 1対のツイストペアに送信と受信を重畳
  - この手のシングルペアイーサネットは車でCANの後継で使われる
- 100BASE-T2(802.3y): ツイストペア2対
- PoE化したのが802.3br(2017)

組み込み用1000Mイーサネット物理層

- 1000BASE-T1(802.3bp, 2016):
  - 100BASE-T1を750MHzで動作させている
  - 個人的には、かなり無理があるのでは…
- PoE化したのが802.3br(2017)
  - 個人的には、電力と750MHzの信号を1ツイストペアで送信は相当無理があるのでは…(1000BASE-TでもPoEはそこそこトラブル多いし)
- mGig化も802.3chで検討されているらしい
- 1000BASE-RHx(802.3bv, 2017)
  - GEPOF: Gigabit Ethernet over Plastic Optical Fiber
  - 個人的には、重量も考えると1000Mから先はこちらがベストでは

高速化という点からメタルケーブルのやっかいな点(1/2)

単純な高速化の視点から

- 100mという長い信号線のドライブ可能なドライバ
  - 高速動作と電流ドライブ能力の両立
- 長い信号線を通過して崩れた信号を受信するレシーバ
  - クロック信号を復元する時にツイストペア間の差をどう補償する?
- 高速化のために信号の電圧を落とす方向にある
  - ノイズ耐性が落ちる

個人的には、無線LANとか多値フラッシュメモリのように符号訂正てんこ盛りを利用すればなんとかなると思う →厳しいのでは

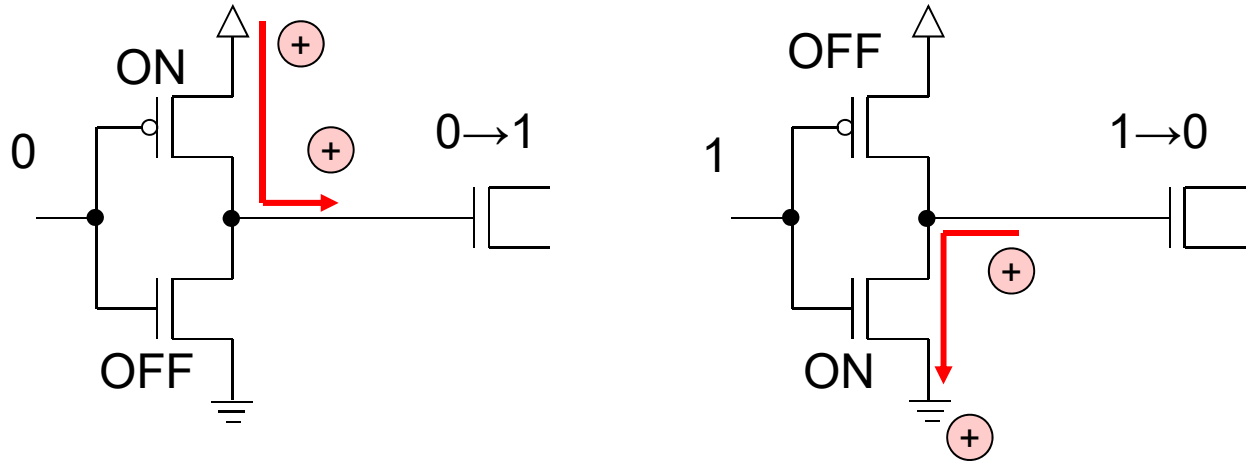
- ただし、消費電力を無視すれば

  →これ以上複雑な符号訂正は電力的に厳しいのでは

電圧論理の動作の基本 (出力側キャパシタンスの問題)

- CMOSでは次の動作で電荷が移動
  - 入力が0の時に電荷がVDDから出力ノードに移動
  - 入力が1に時に電荷が出力ノードからGNDに移動
- 電荷の移動にかかる時間で動作時間は決まる
  - 出力側のノードのキャパシタンスと電圧振幅で電荷の移動にかかる時間は決まる

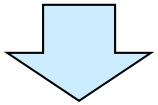
→長いケーブルはキャパシタンス大でドライブに時間がかかる



高速化という点からメタルケーブルのやっかいな点(2/2)

消費電力の視点から

- 8b/10bエンコーディングは一定間隔で0/1が変わる
- 長い信号線ドライブに対応したドライバ
- 高速かつ高精度なレシーバ
  - 送受信に使う電位は増大する傾向に
  - アナログ回路は性能を優先すると電流だだ流しになりやすい



- 28nmの半導体プロセスでも1ポートあたり2.5W電力を食う
  - 10GBASE-(S/L)Rは10年前でもSFP+モジュールで1W未満
  - 1000BASE-TのSFPはあるが、10GBASE-TのSFP+はいまだ無し

→40GBASEとか作ってもポート密度が落ちたら嬉しくない

興味を持った人へ

Internet Watchですごくしっかりした内容のコラムが連載がされている(この講義資料より精確かつ細かく解説されている)

- 「光Ethernetの歴史と発展」2020/3-連載中

https://internet.watch.impress.co.jp/docs/column/nettech/1243869.html

- 「10GBASE-T、ついに普及へ？」連載終了

https://internet.watch.impress.co.jp/docs/column/nettech/1086305.html

- 「アクセス回線10Gbpsへの道」連載終了

https://internet.watch.impress.co.jp/docs/column/nettech/1097180.html