情報ネットワーク特論
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
嶋田 創

嶋田側の講義の概要

- テーマ1: 高速ネットワークを支えるハードウェア技術
  - 私の得意分野がハードウェア側のため
  - 単純に通信を行うだけでなく、それを支える技術も利用可能か?
    - 高速通信下でセキュリティ対策は行えるか?
    - 通信速度に見合うサーバは組めるか?
  - 高速ネットワークを実現するのに必要となる、処理速度、回路技術、メモリ技術、トランシーバ技術、符号化技術
- テーマ2: 情報セキュリティ技術
  - 近年のサイバー攻撃の話も含めた情報セキュリティの話題
    - 攻撃側の様々な手口とその考え方
    - 防御側の様々な技法とその考え方
  - 情報セキュリティに関する法律について

最初のトピック

- ハードウェア規格として見たイーサネット規格
  - 現在の主流: 1000BASE-T、1000BASE/10GBASE-(L/S)R
  - 過去の主流: 10BASE-T/100BASE-TX
  - 将来の主流や現バックボーン: 10GBASE-T/40GBASE/100GBASE-(S/L)R
- 計算機内部の速度とイーサネットの速度の比較
- イーサネットの物理層の動作を高速化/低電力化するには?

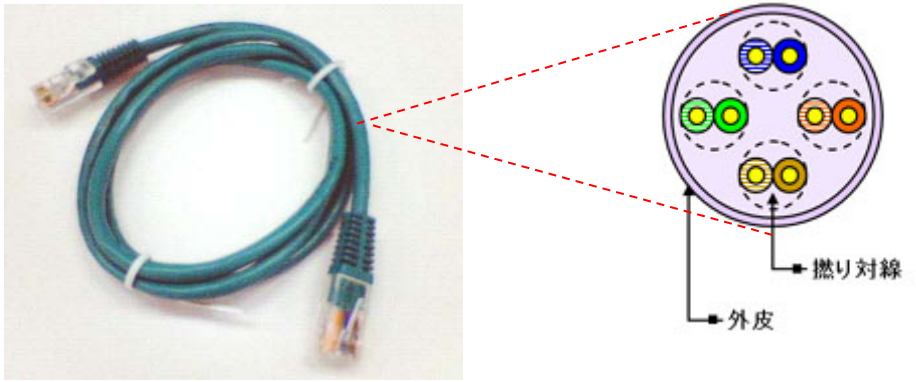
この講義資料も、後から継ぎ足される可能性があります

現在の主流のイーサネット規格

→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

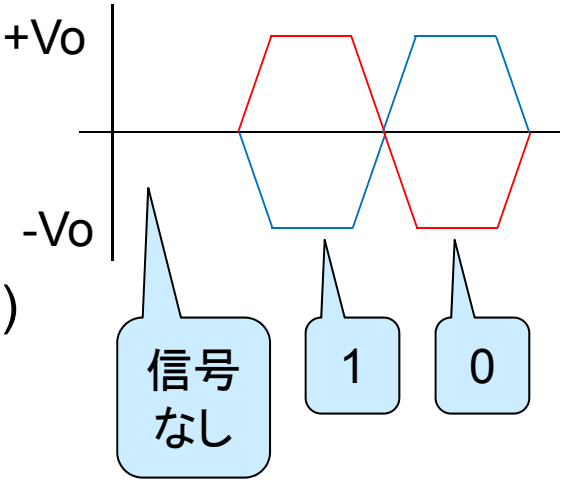
UTPケーブル



スイッチング・ハブ

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
- 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, -0.5Vo, -Vo)
- 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



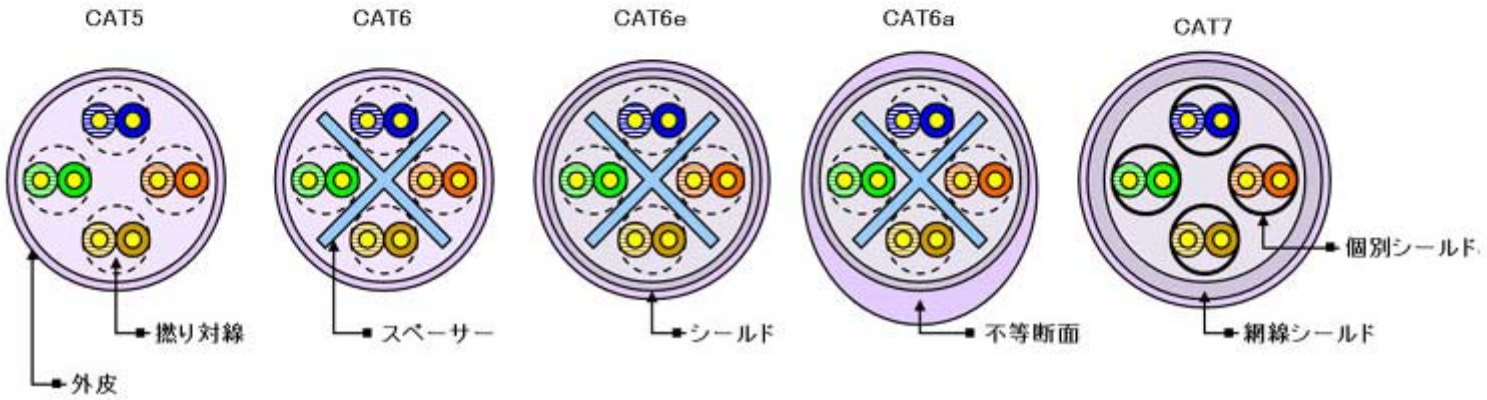
- 通信速度: 125MHz x 2bit x 4pair = $1x10^9$ bps

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

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

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

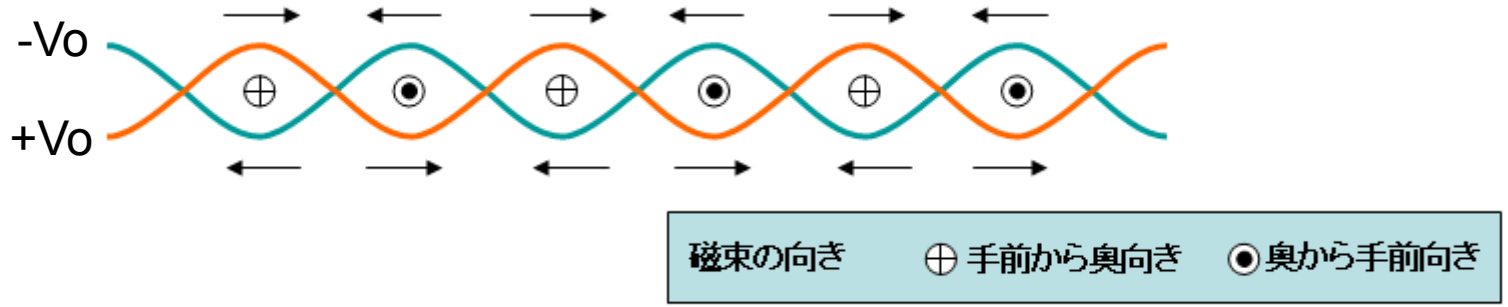


ツイスト・ペア・ケーブルの特徴

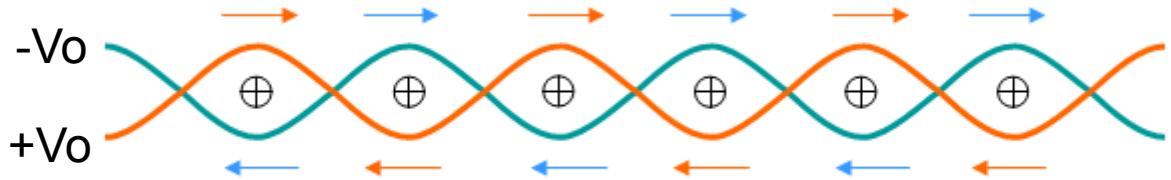
ノイズに強い、ノイズを出さない

→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



- 外部からの磁束による電流は打ち消し合う



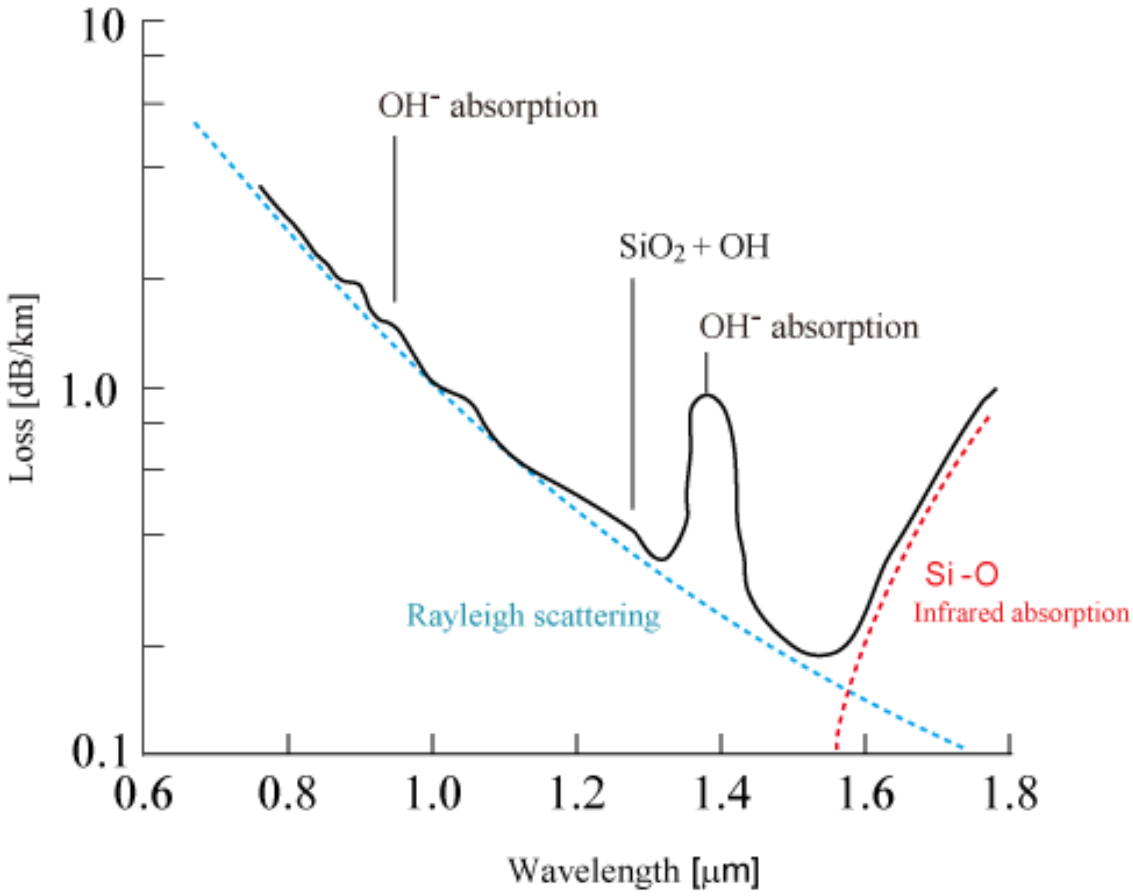
- さらに、作動ペアにより乗ったノイズを除去可能

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・ペアケーブルを使うもの
  - 1000BASE-CX: 超短距離向け(～ 25m)

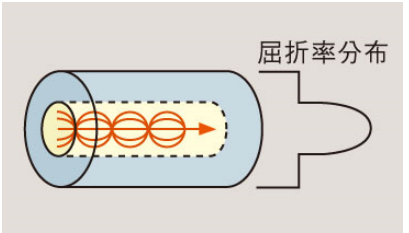
石英系光ファイバの波長と減衰率

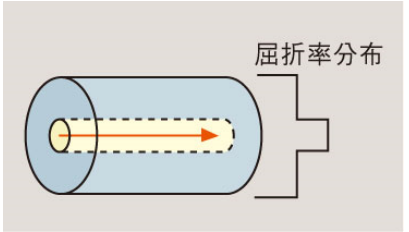
- 減衰率が小さい波長を長距離向けに利用



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
- 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



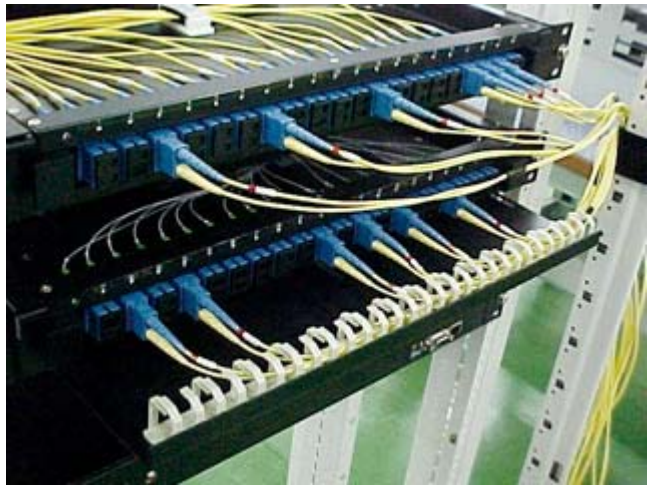


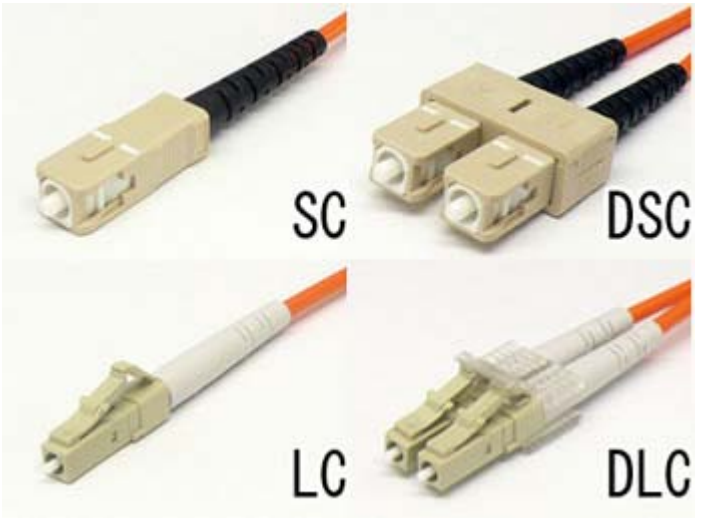
名大内の光ファイバ網

- 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グレーデッドインデクスとシングルモードを使い分け
  - 1990年代に一斉に敷設
    - そろそろ劣化が怖い
  - 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

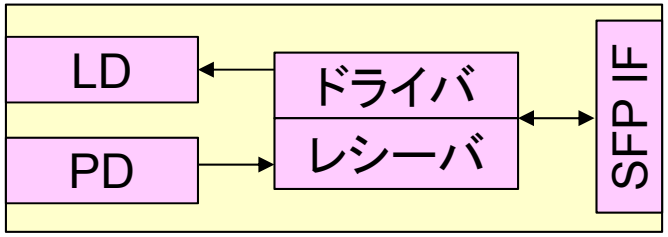
- 光ファイバの末端は光コネクタに整形される
  - 大抵は光パッチパネルに列の形で設置する
  - SC規格とLC規格が主流
    - 後述する光トランシーバ側も同様
- 通常のネットワーク構築では2本の光ファイバを組みで使う
  - 送信側と受信側で1本ずつ
  - 1本でWDM(波長分割多重)を使う方法もあるが、コスト的には2本使う方が安い



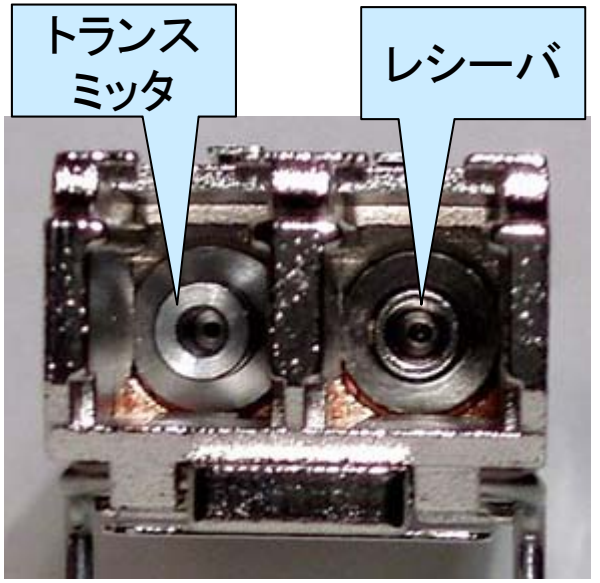


光トランシーバ

- 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

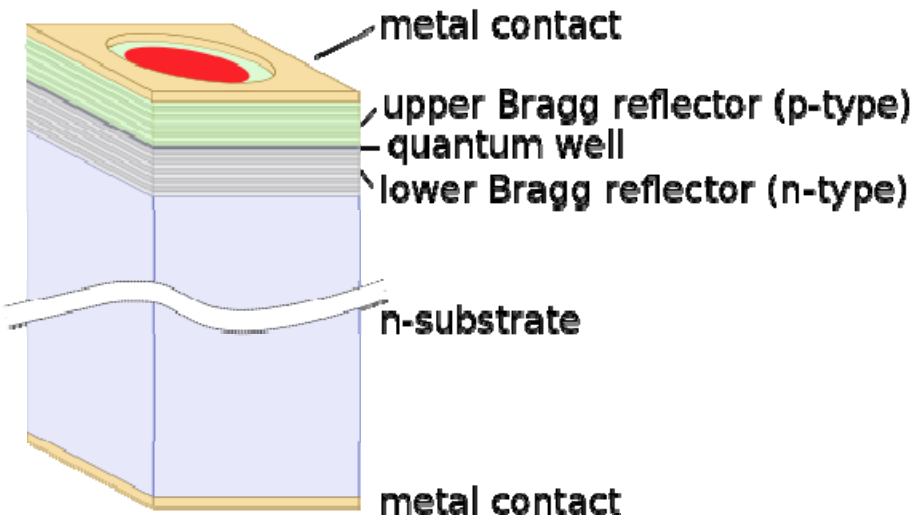


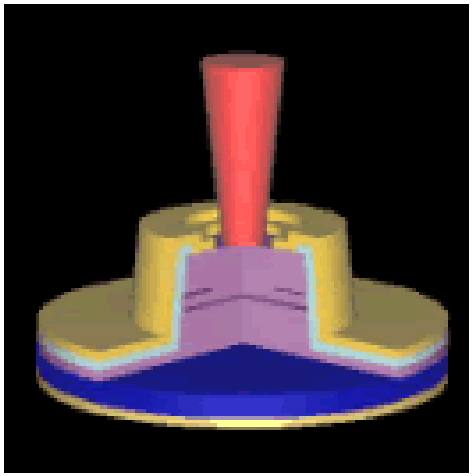
SFPモジュール内の
トランスミッタとレシーバ



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25Gbps/波長

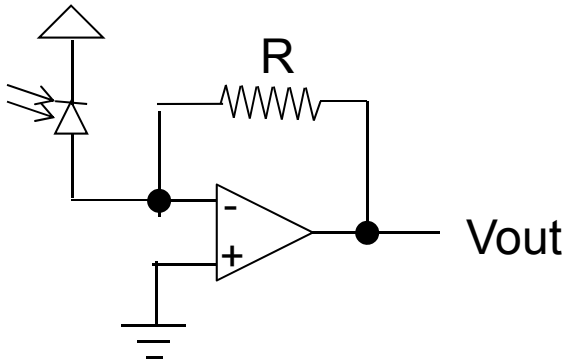




フォトダイオードによる受信

- 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



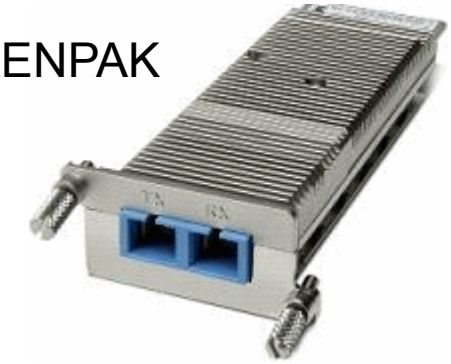
(光)トランシーバの規格

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: ほぼ1000BASE用(発展終了)
- SFP+: ほぼ10GBASE用(発展終了)
  - 発展中のもの: XENPAK, X2, XFPなど
- QSFP: ほぼ40GBASE/100GBASE用
  - 厳密には、100GBASE用はQSFP28という名前
- CFP: ほぼ100GBASE用(横方向でかすぎ)

QSFP

XENPAK



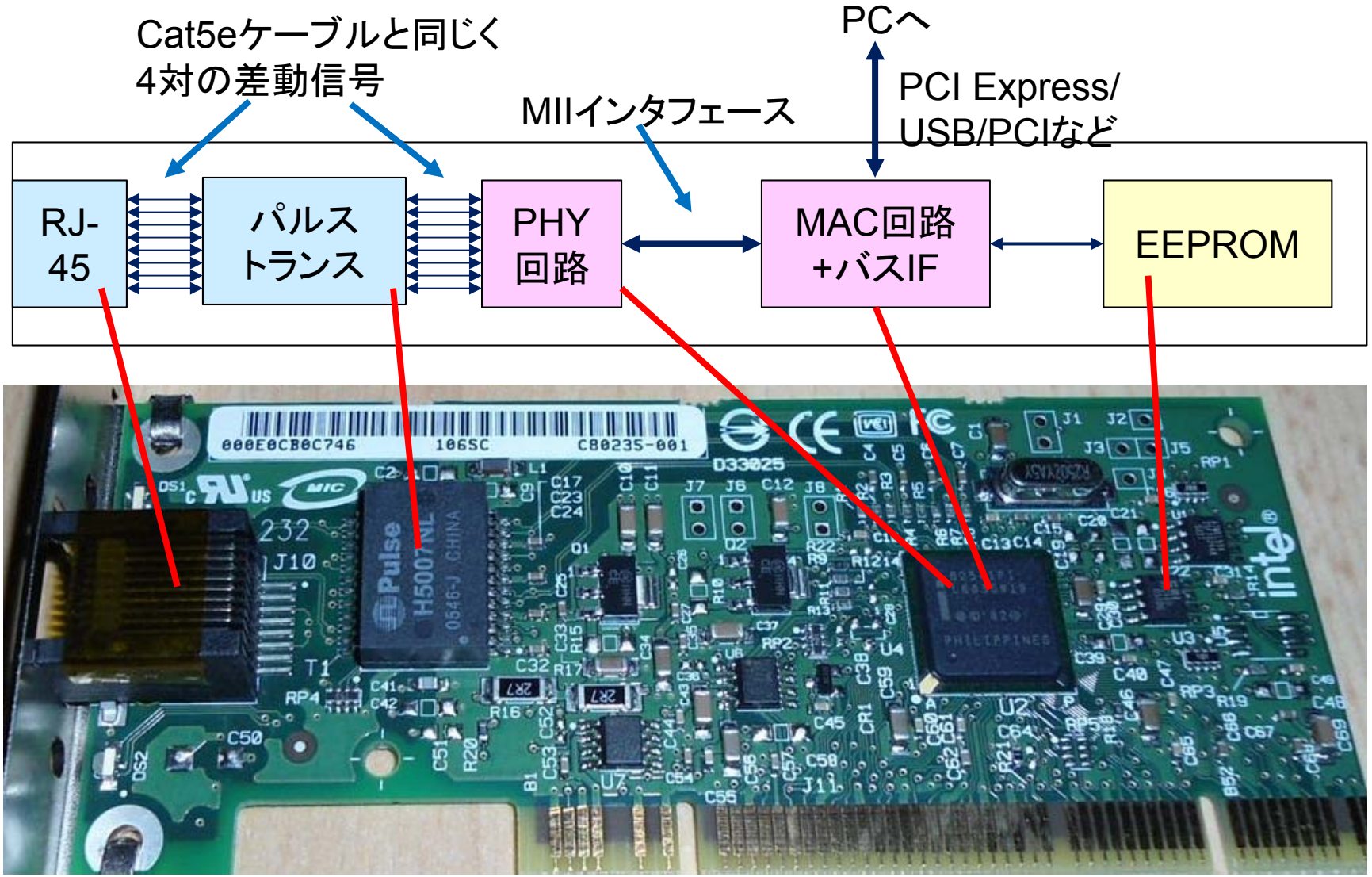
SFP/
SFP+

CFP

光トランシーバーよもやま

- 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等が持たない
  - 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

一般的なNIC(1000BASE-T)の構成

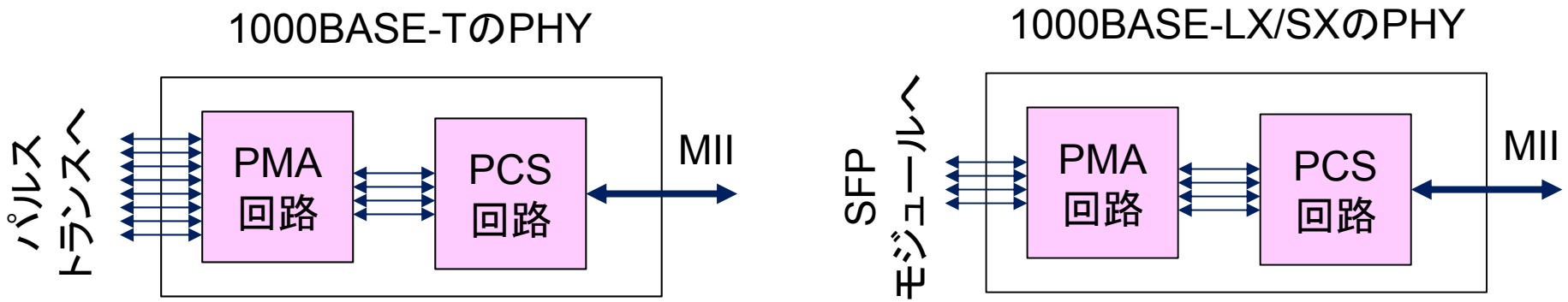


物理層(PHY)とメディアアクセス制御層(MAC)

- 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

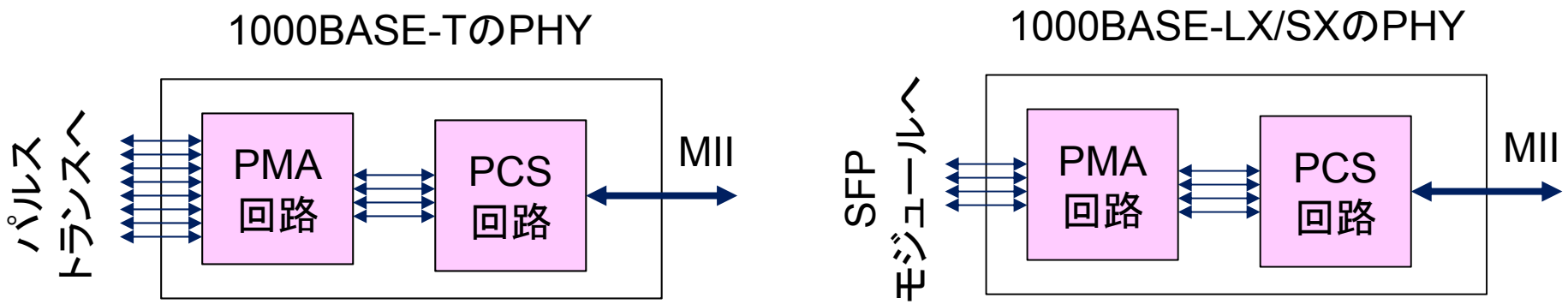
PHYの構成(1/2)

- 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



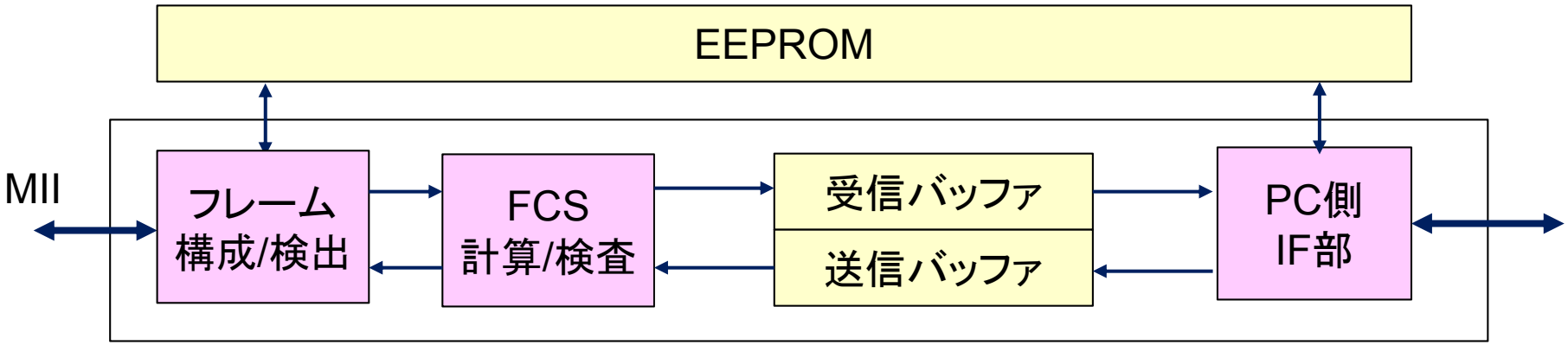
PHYの構成(2/2)

- 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



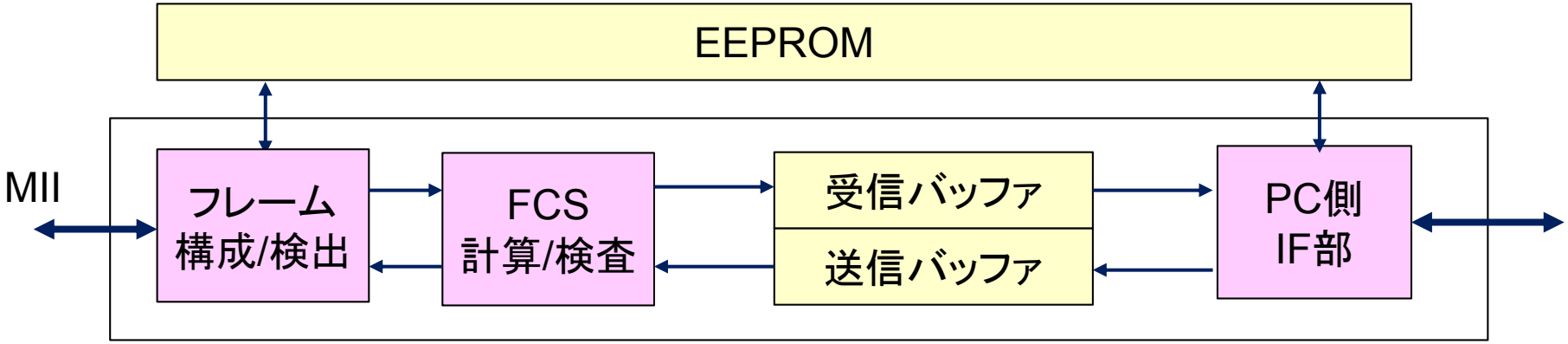
MAC層の構成(1/2)

- 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(1/3)

- 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(2/3)

- 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(2/3)

- 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

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

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

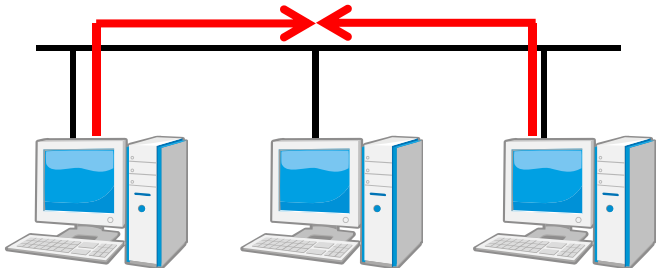
100BASE-TX

- ベース・クロック125MHz
- 2対のツイスト・ペアを利用
- 4b/5bエンコーディングを採用
  - 通信速度の実効値は4/5になる
- 引き続き、CSMA/CDを利用

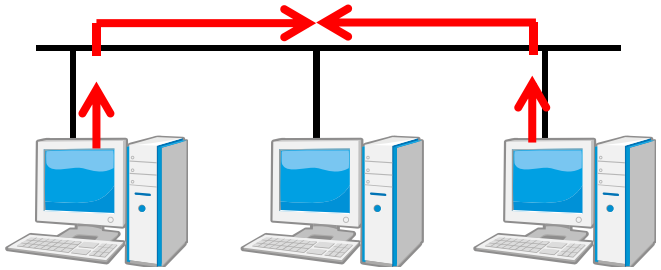
- 通信速度: 125MHz x 4/5bit x 1対 = $1x10^8$ bps

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



○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

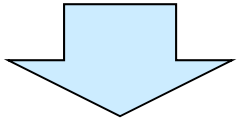
10GBASE-T

- 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 4対 x 1723/2048
  $= 9.4 \times 10^9$ bps

10GBASE-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

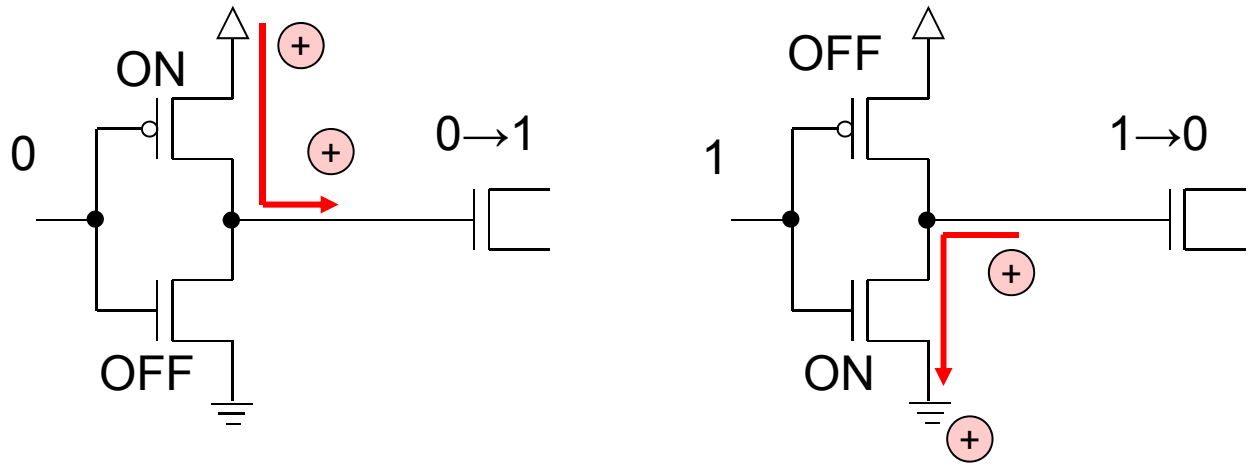


- メタルケーブルで10GBASE以降の高速化はかなり難しい
- 現在は40GBASE-Tが規格策定中だが最長で30m
  - データセンタの中ぐらい程度の距離
  - ケーブルもカテゴリ8(max 2GHz)らしいが…

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



高速化という点からメタルケーブルのやっかいな点

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?

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
- 長い信号線を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

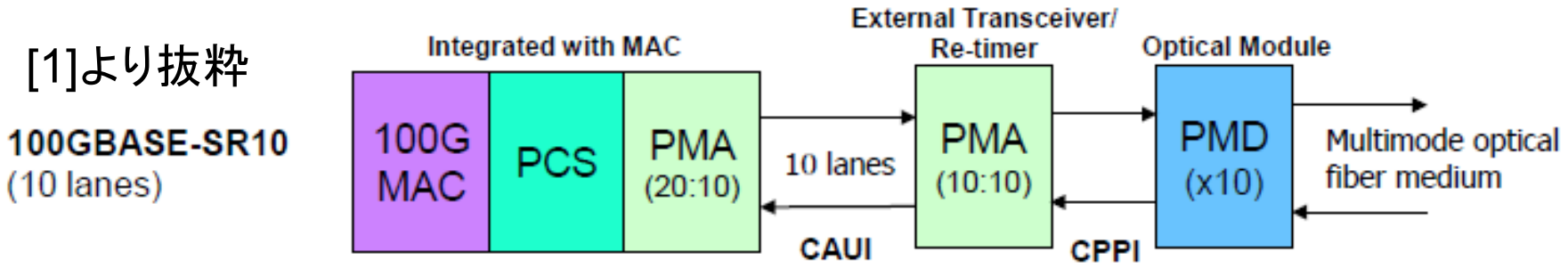
10Gbps以上の光ファイバ規格

- 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-SR10: 10G x10、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、マルチモード光ファイバ
- 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
  - 25GBASE: 25G x1(100GBASE-LR4の光を1つだけ利用)
    - 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-LR4の光を1つだけ利用)
  - 40GBASE-LR4 Lite
    - 到達距離が無印LR4の10kmから2kmへダウン
    - その代わり、光モジュールが低コスト

100GBASE-SR10

- 10Gbps/波長を物理的に10本束ねて100Gbps
  - マルチモード光ファイバを20本利用
- 技術的には物理層は面白く無い
- というか、トランシーバの構造的にも安くならないので嬉しく無い
  - 10GBASEを10本トランクリンクで束ねた方が安い
  - ただし、トランクリンクでは1TCPセッションは10Gbpsに制限

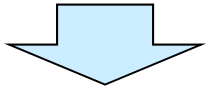
[1]より抜粋



[1] I. Ganga, “IEEE 802.3ba 40 and100 Gigabit Ethernet Architecture”, 2010.

現在規格化中の光ファイバ規格

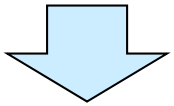
- 400GBASE
  - 400GBASE-SR16: 25G x16、マルチモード光ファイバ
  - 400GBASE-FR8: 50G x8、シングルモードファイバ
    - ただし、到達距離は2km
    - 一部メーカが出している40GBASE-LR4 Liteが規格化された感じ
  - 400GBASE-LR8: 50G x8、シングルモードファイバ
  - MII層でも50Gbpsまでの対応は目処がついている
- 低コスト版で200GBASEも考えられているらしい



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

メタルケーブルの1Gと10Gの中間速度策定話

- メタルケーブルの仕様的に10Gは厳しい
  - 距離100mにはカテゴリ7/6Aが必要
- すでに敷設されているケーブルはカテゴリ5e/6が多い
- 無線LANの規格が1Gbpsを超えたのでアクセスポイントの足回りに1Gbps以上欲しい



- カテゴリ5eで100m/2.5Gbps狙いましょう(2.5GBASE-T)
  - 10GBASE-Tのハーフクロック動作の電圧変動4段階版
- カテゴリ6で100m/5Gbps狙いましょう(5GBASE-T)
  - 10GBASE-Tのハーフクロック動作

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: 光ファイバを用いた短距離(～300m)
  - L: 光ファイバを用いた長距離(～10km)
  - E: 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
- 利用する光の(波長の)数
  - 末尾の数字

計算機内部の速度とネットワークの速度

では、ネットワークの速度向上と比較した計算機内部の速度の向上は?

- プロセッサ性能の向上
- 内部バスの速度向上
- 入出力装置の速度向上

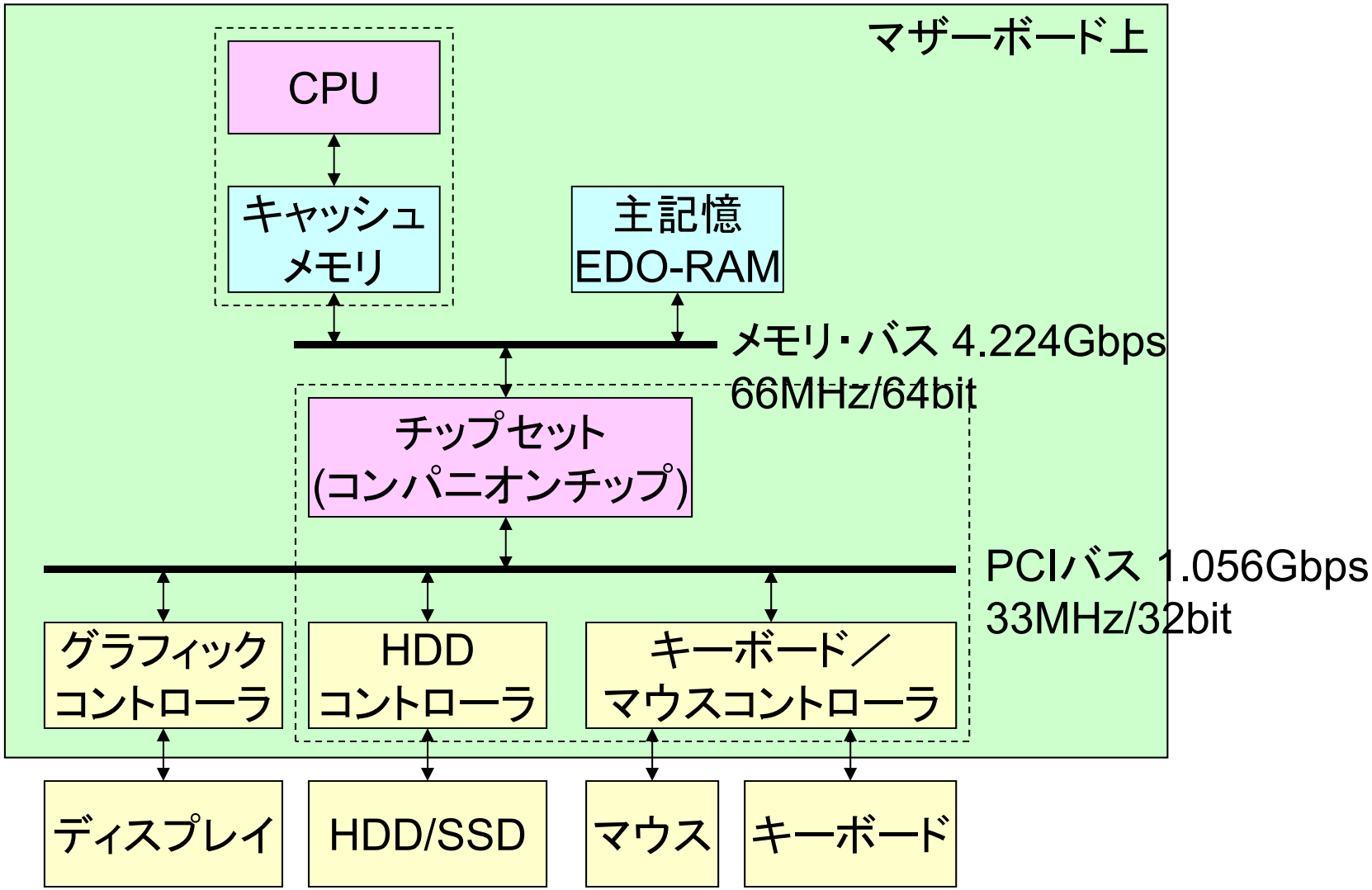
ネットワーク規格とプロセッサ性能

- 10BASE-T: 1990年
  - プロセッサに対して1/160
    = 10M/(50M*32)
- 100BASE-TX: 1995年
  - プロセッサに対して1/144
    = 100M/(450M*32)
- 1000BASE-T: 1999年
  - プロセッサに対して1/57.6
    = 1G/(1G*32)
- 10GBASE-T: 2007年
  - プロセッサに対して1/30.7
    = 10G/(9.6G*32)

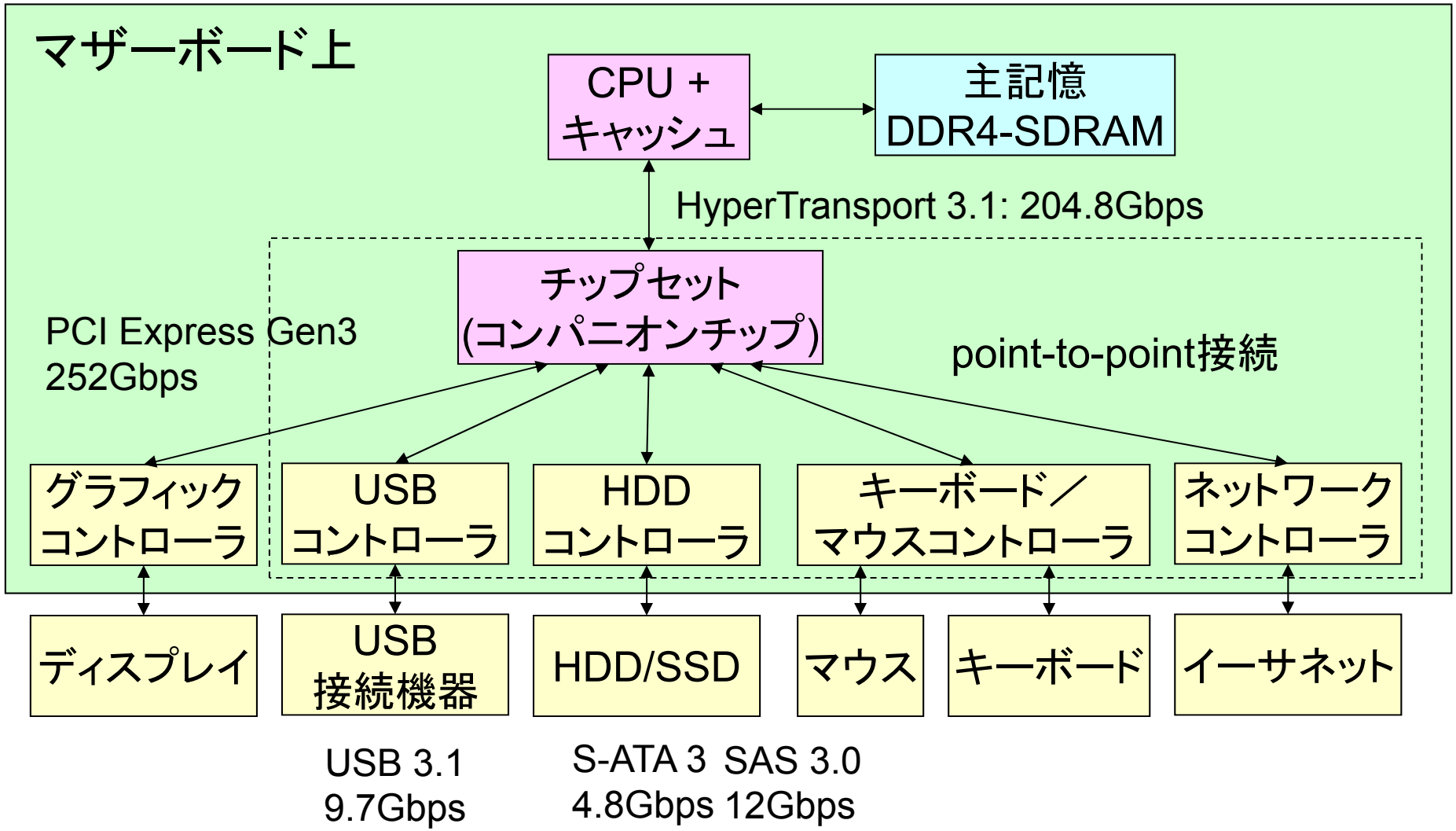
- 1990年 Intel 80486DX 50MHz
  - スカラプロセッサ
  - 最大50M演算/秒
- 1995年 同Pentium Pro 150MHz
  - 3命令発行スーパスカラ
  - 最大450M演算/秒
- 1999年 同Pentium III 600MHz
  - 3命令発行スーパスカラ
  - 最大1.8G演算/秒
- 2007年 同Core2 2.4GHz
  - 4命令発行スーパスカラ
  - 最大 9.6G演算/秒

近年を除き、おおむねムーアの法則(年率1.4倍)に従う

10BASE全盛時の計算機の構成



現在の計算機の構成


マザーボード上
CPU +
キャッシュ
主記憶
DDR4-SDRAM
HyperTransport 3.1: 204.8Gbps
チップセット
(コンパニオンチップ)
PCI Express Gen3
252Gbps
point-to-point接続
グラフィック
コントローラ
USB
コントローラ
HDD
コントローラ
キーボード／
マウスコントローラ
ネットワーク
コントローラ
ディスプレイ
USB
接続機器
HDD/SSD
マウス
キーボード
イーサネット
USB 3.1
9.7Gbps
S-ATA 3
4.8Gbps
SAS 3.0
12Gbps

計算機内外のバスの速度比較(1/2)

この速度では、2本の作動信号線(lane)で1bitを送る

- HyperTransport 3.1
  - 3.2Gbps/lane
  - 最大32lane
  - Double Data Rate
  - 最大速度: 3.2Gbps x 32lane x 2bit/clk = 204.8Gbps
- S-ATA3
  - 6Gbps/lane
  - 8b/10bエンコーディング
  - 最大速度: 6Gbps x 8/10 = 4.8Gbps
- SAS 3.0
  - 12Gbps/lane
  - 8b/10bエンコーディング
  - 最大速度: 12Gbps x 8/10 = 9.6Gbps

計算機内外のバスの速度比較(2/2)

- PCI-Express Gen3.0
  - 8Gbps/lane
  - 最大32lane
  - 128b/130bエンコーディング
  - 最大速度: 8Gbps x 32lane x 128/130 = 252Gbps
- USB 3.1
  - 5Gbps/lane
  - 2lane
  - 128b/132bエンコーディング
  - 最大速度: 5Gbps x 2lane x 128/132 = 9.69Gbps

10Gbps以上も計算機内外で余裕で取り扱える状態だが、
これ以上の速度向上ではメタルケーブルと同じ問題が…

レイテンシの問題

- 前述の議論には、レイテンシという視点が欠けている
- 理想的にデータを転送して処理できている時は問題ない(流れ作業状態)
- あるデータによって、転送するデータを変える時はどうする?

  →新たなデータ読み出し開始までの時間は通信ができない
- 記憶装置や入出力装置に要求を送り、データが読み出されるまでの時間をレイテンシと呼ぶ
  - 主記憶の読み書きレイテンシ
  - HDDの読み書きレイテンシ
  - フラッシュメモリの読み書きレイテンシ
- TCPのフロー制御に関するRound Tripのレイテンシ

各種データ保持機器の読み書きレイテンシ

読み出し/書き込み開始まで

- HDD(7200rpm前後): 数ms
  - 基本的に、ディスクが半回転する時間
- SSD(フラッシュメモリ)
  - 読み出し: 数百ns
  - 書き込み: 数十us
- 主記憶(DDR3-SDRAM): 数十ns

光速の上限によるレイテンシの問題

- ではより高速な動作をする入出力デバイスを準備すればレイテンシは減るか

  →バスの信号線のレイテンシが支配的になるだけ

- 10GHzでは1クロックで信号は3cmしか進まない
  - 信号が光速で伝播するという理想的な状態
  - 導出: 光速/クロック周波数

  →すでに、計算機の中ではこちらがバスのレイテンシが問題となっている

まとめ

- イーサネットの規格はおおよそ5年ごとに10倍のバンド幅の規格が出ているが、この先の向上は厳しそう
- 計算機内部の方もネットワークのバンド幅向上とバランスよく向上してきたが、同様に厳しそう
- レイテンシ短縮に至ってはそれほど大きくはない
- むしろ、レイテンシはどの世界でも長くなる傾向にある
  - 信号の伝達速度の上限は光の速度という物理制限
  - 転送中に制御が入ると、実効速度は落ちる

→バースト転送を行う用途以外は効果が薄い