情報ネットワーク特論
ネットワーク機器とFPGA

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

ネットワークのハードウェア周りを実装するには?

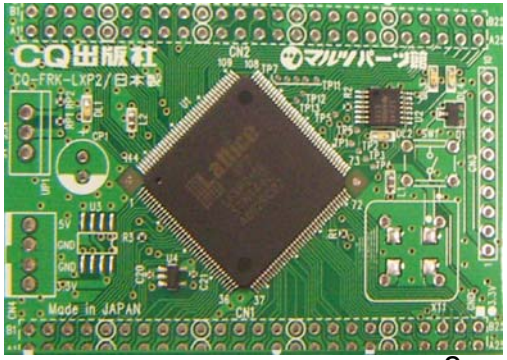
- 今までネットワークに関連するL1,L2,(L3)の世界とハードウェアの関係を見てきた
- 中身のよくわからない部分としてASICで構成されている部分がある
  - 高速化の要となっているようだが中身は細かく分からない
    - 他の企業に真似されると嫌なので、特に最近は公開されない

- ASICの部分は自分で細かく見たりすることはできない?

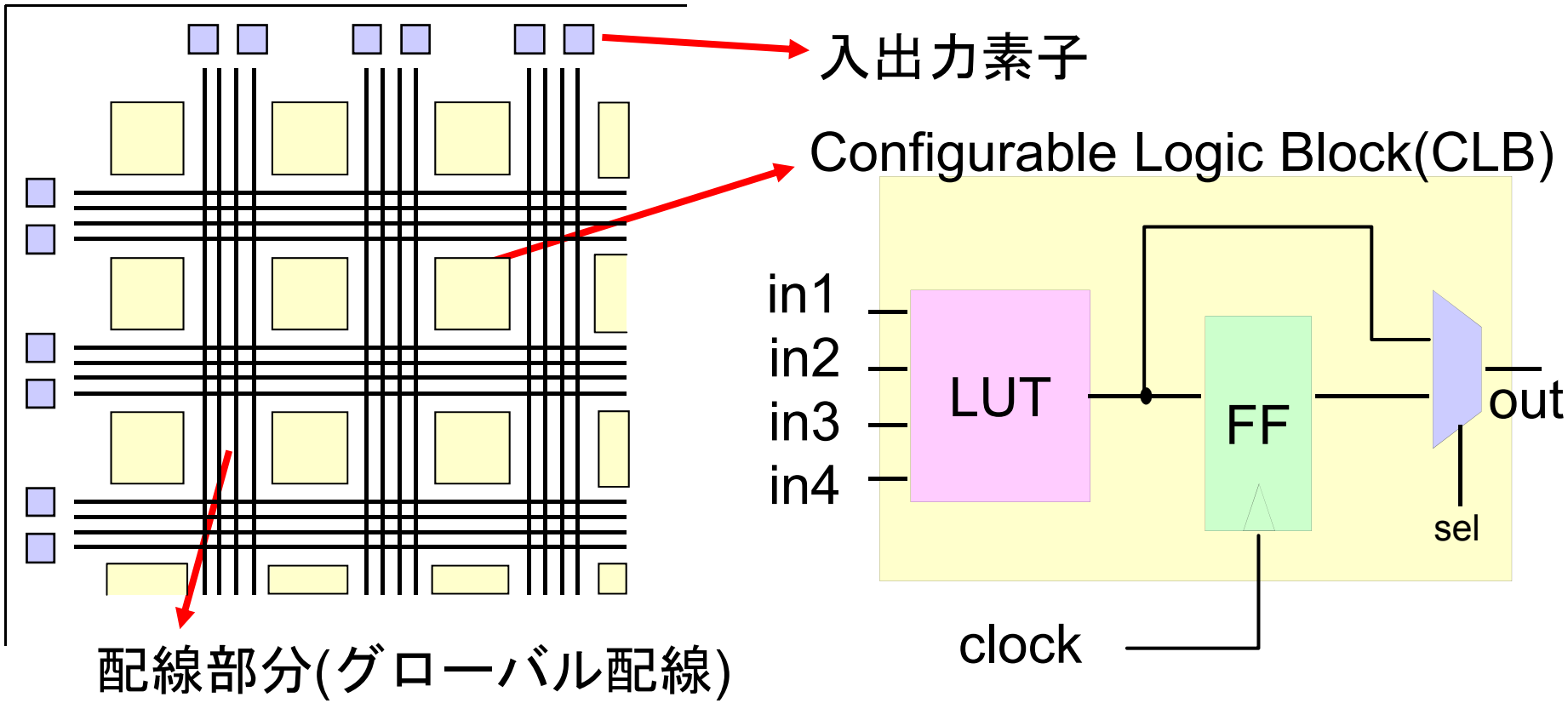
  →FPGAで実装することで確認できるかもしれない

FPGA: Field Programmable Gate Array

- 近年多用される再構成可能ハードウェア
- LUTを使った構成が主流
  - LUT(Look-Up Table): 任意の3-8入力の信号に対して任意の値を出力する論理素子
- プロトタイピングで多用される
  - もしくは少量生産
    - ネットワーク機器ではよくある
  - もしくはASICが来るまでのつなぎ

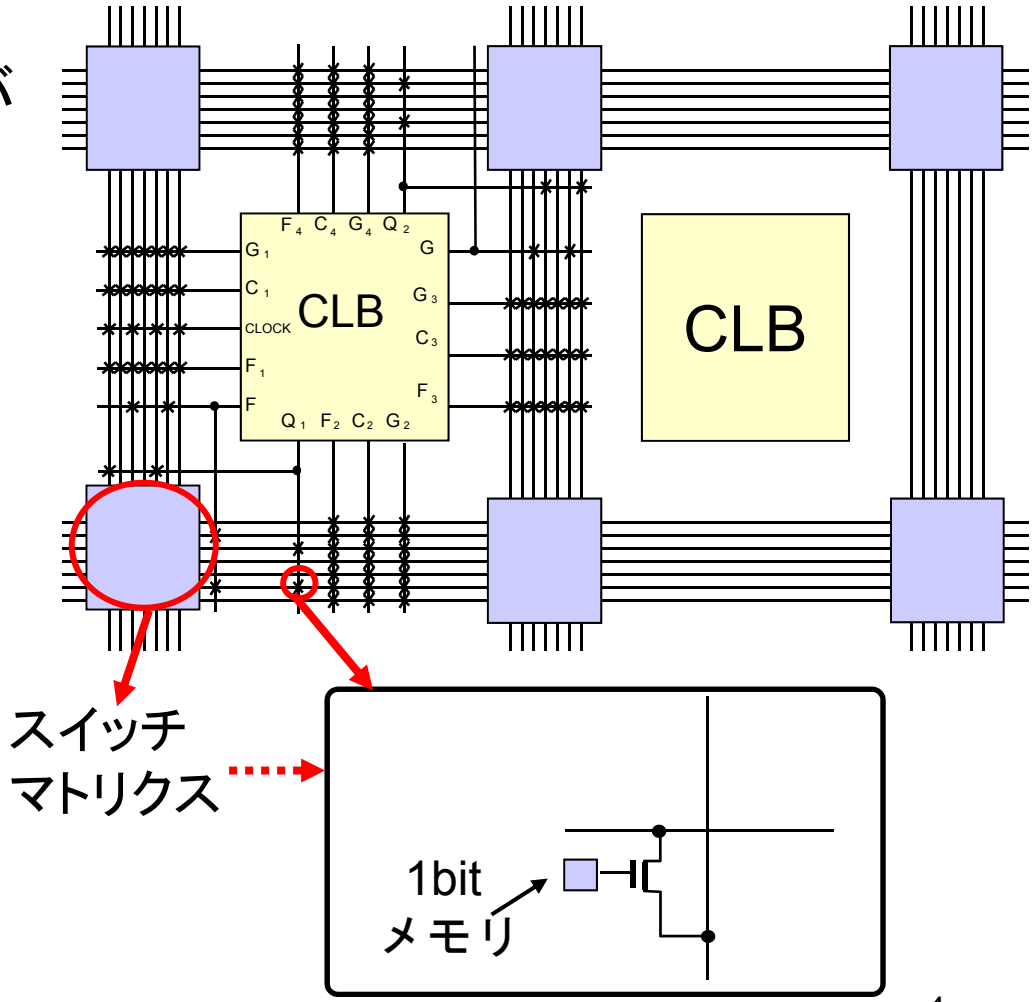


LUTを使うFPGAの概観



グローバル配線の構成

- 接続部分は2箇所
  - グローバル配線とグローバル配線(スイッチマトリクス)
  - グローバル配線とCLB
- 配線の接続はパストランジスタで制御される
  - パストランジスタに接続されたメモリに接続情報を書き込む

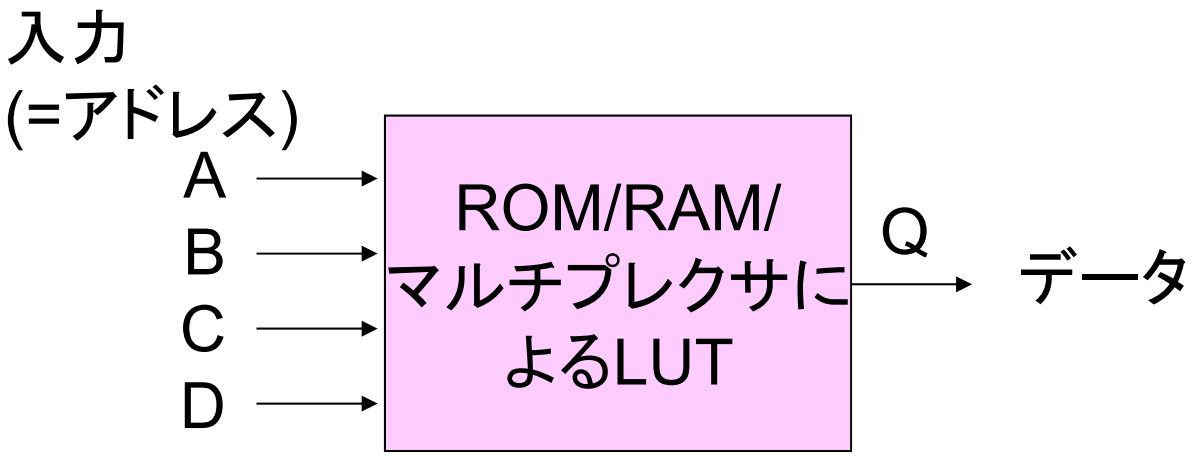


LUT(Look-Up Table): 任意の論理値を出力できる論理素子

- RAMベースのLUTを考えると考えやすい
  - e.g. 4bit入力アドレスに対して1bitを出力するRAM
- LUTはマルチプレクサやROMなどでも実現される

RAMの値

ABCD	Q
0000	0
1000	1
0100	1
1100	0
0010	1
1010	1
0110	0
1110	1
0001	1
1001	0
0101	1
1101	1
0011	0
1011	1
0111	1
1111	0

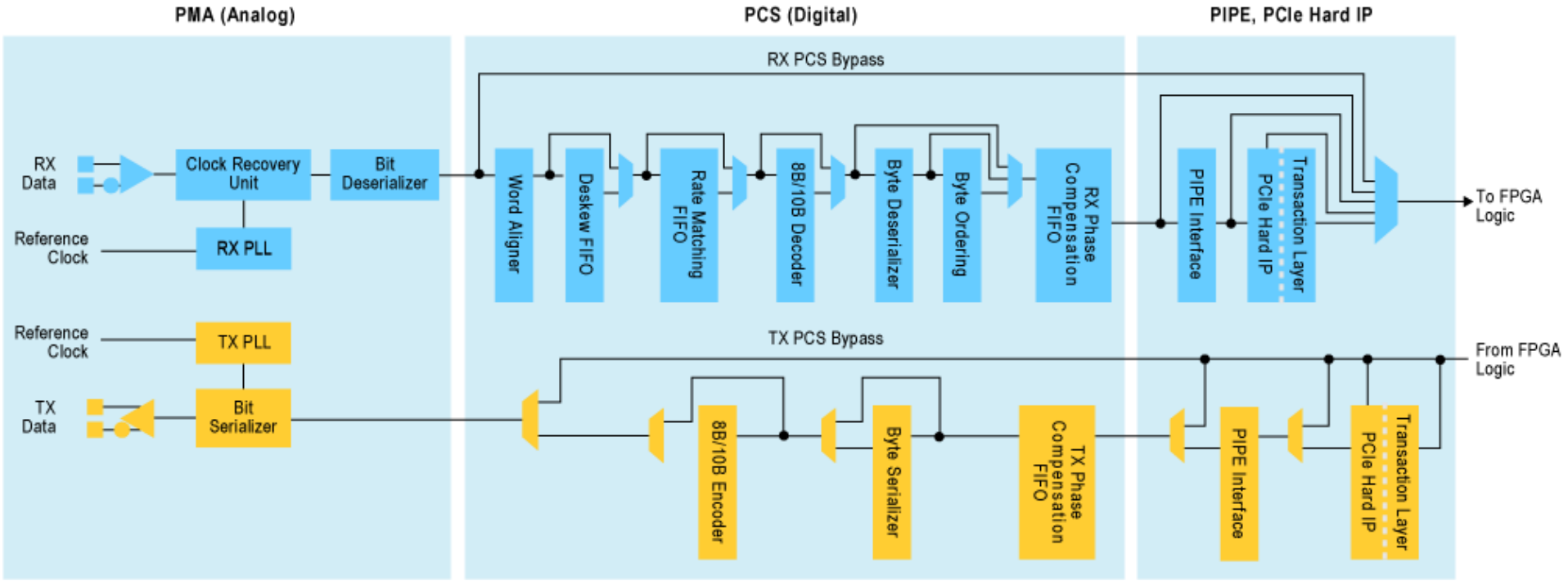


最近のFPGAはCLB以外もいろいろ搭載している

- ブロックSRAM
  - 容量重視、速度重視などバリエーションあり
- 全加算器(高速キャリー線付き)
- 乗算器
- 組み込みプロセッサ
- DSPコア
- 高速I/O
  - おおむね3Gbps以上

Alteraの高速I/Oの物理構造

- PMS/PCSはイーサネットと同様



FPGAメーカ

- AlteraとXilinxが業界大手
  - 10G以上を実用的に使おうとすると実質この2社
- Altera(2015/6にIntelに買収された)
  - 高速IO付きFPGAのバリエーションが多い
  - Intelの14nmプロセスを利用した高性能
  - Intel XeonとのMulti Chip Module版も発表された(2016/4)
- Xilinx
  - 10GのMAC IPコアを無料で使える
- その他: 1GBASEあたりまでは対応できる
  - Actel: アンチヒューズ型(高速だが書き換え回数1回をラインアップ)
  - Quicklogic: アンチヒューズ型
  - Lattice

高速IOを持つFPGA(Altera)

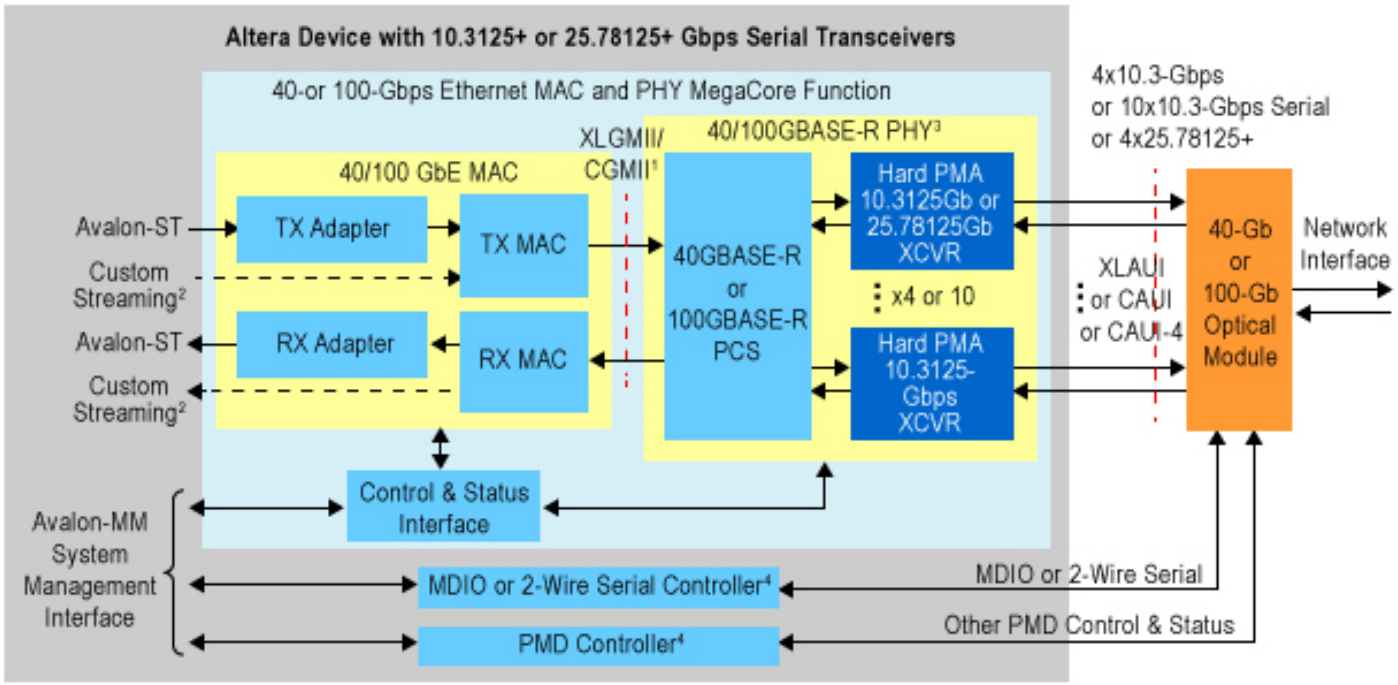
- Stratix
  - Stratix V GX(28nm): 14.1Gbps x66
  - Stratix V GT(28nm): 28.05Gbps x4, 12.5Gbps x32
  - Stratix 10 GX(14nm): 30Gbps x96
  - Stratix 10 GT(14nm): 56Gbps x?
- Arria
  - Arria V GZ(28nm): 12.5Gbps x36
  - Arria 10 GT(20nm): 25.8Gbps x 96
- Cyclone
  - Cyclone IV GX(40nm): 3.125Gbps x8
  - Cyclone V GT(28nm): 6.144Gbps x12

高速IOを持つFPGA(Xilinx)

- Virtex
  - Virtex-7(28nm): 28.05Gbps x16, 12.5Gbps x72
  - Virtex UltraScale(20nm): 30.5Gbpx x60
  - Virtex UltraScale+(16nm): 32.75Gbpx x128
- Kintex
  - Kintex-7(28nm): 12.5Gbps x32
  - Kintex UltraScale(20nm): 16.3Gbpx x64
  - Kintex UltraScale+(16nm): 32.75Gbpx x32
- Artix-7(28nm): 6.6Gbps x16
- Spartan-6 LXT(40nm): 3.2Gbps x8

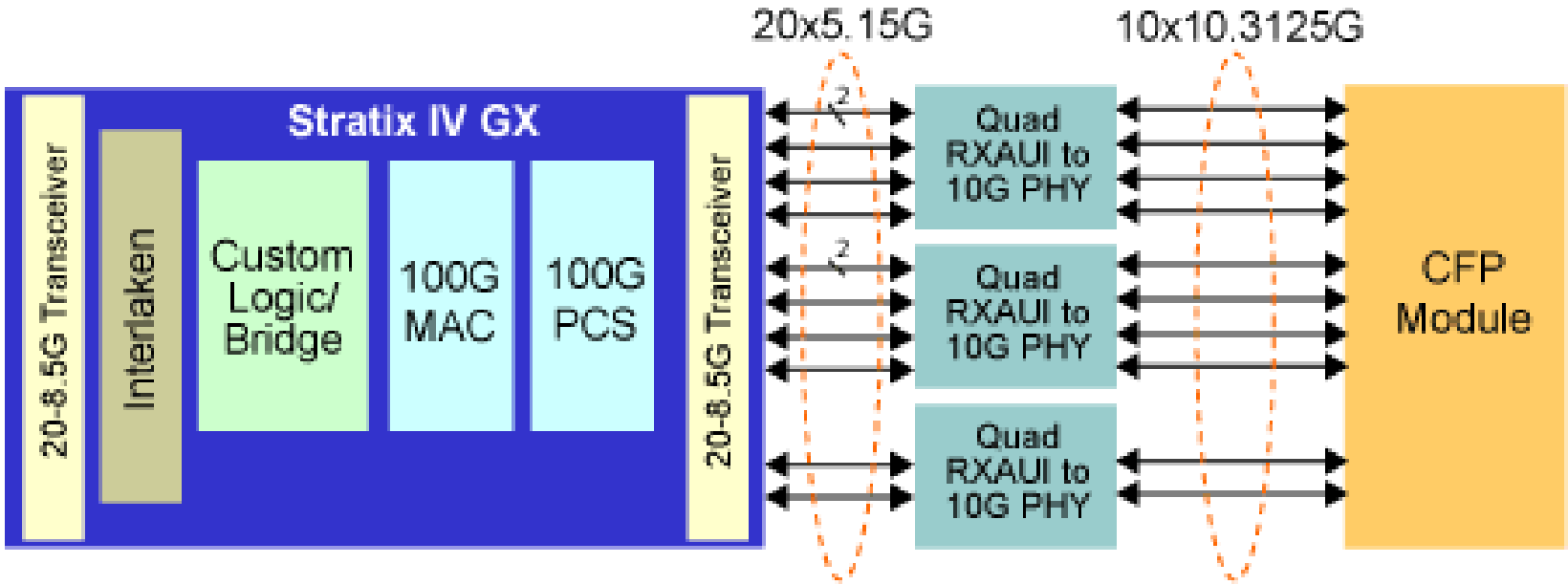
高速I/Oを使ったイーサネットのMAC層

- 通常、FPGAメーカから汎用バスインタフェースを持つMAC層がIPコアとして提供されている
- MAC部は全てFPGA内に実装可能
- Alteraの40G/100G IP Core(下図)



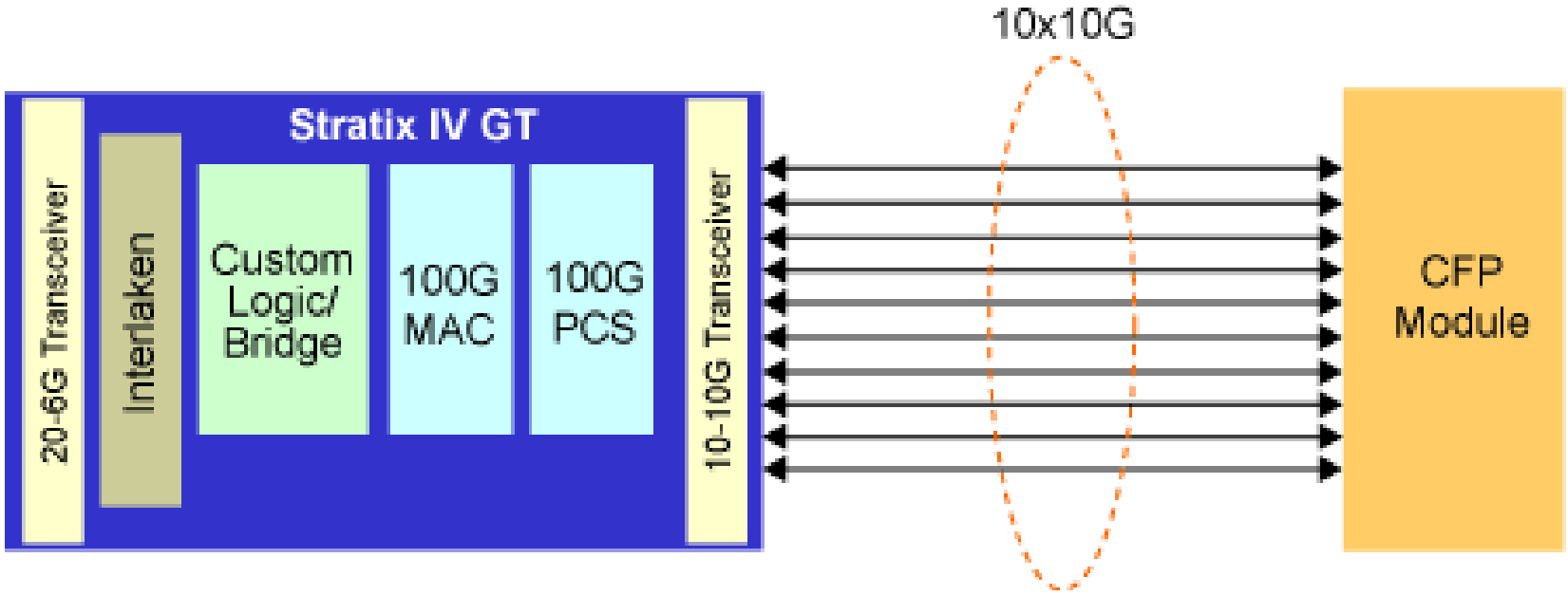
Alteraによる高速イーサネット実装例 (1/2)

- Stratix IV GXを利用した100GbEの実装
- 10G用PHYを使う構成
  - 10Gbps以上のI/Oを持たないFPGAでも対応可能



Alteraによる高速イーサネット実装例 (2/2)

- Stratix IV GTを利用した100GbEの実装
- 10Gbps以上のI/Oを持つFPGA用
  - 10Gbpsちょい上のI/Oは

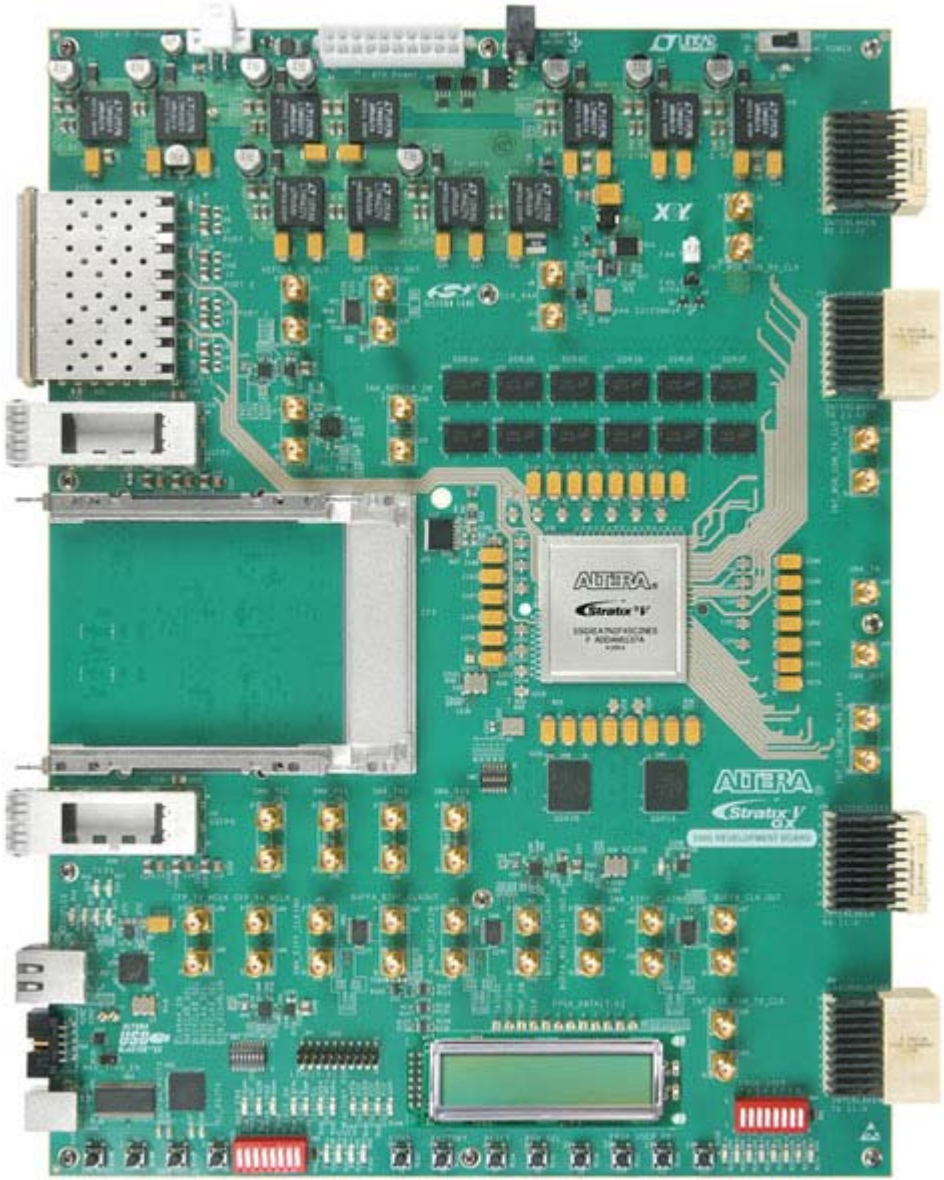


高速I/Oを利用して実装できる他の高速通信規格

- Interlaken: 3.125-6.375Gbps
- 10GbE XAUI: 3.125Gbps
- Fibre Channel: 1.0625, 2.125, 4.25, 8.5Gbps
- OTN(OTN(Optical Transport Network)-4: 9.9-11.3Gbps
- 10G FibreChannel: 10.3125Gbps
- 40GbE: 10.3125Gbps x4
- 100GbE: 10.3125Gbps x10

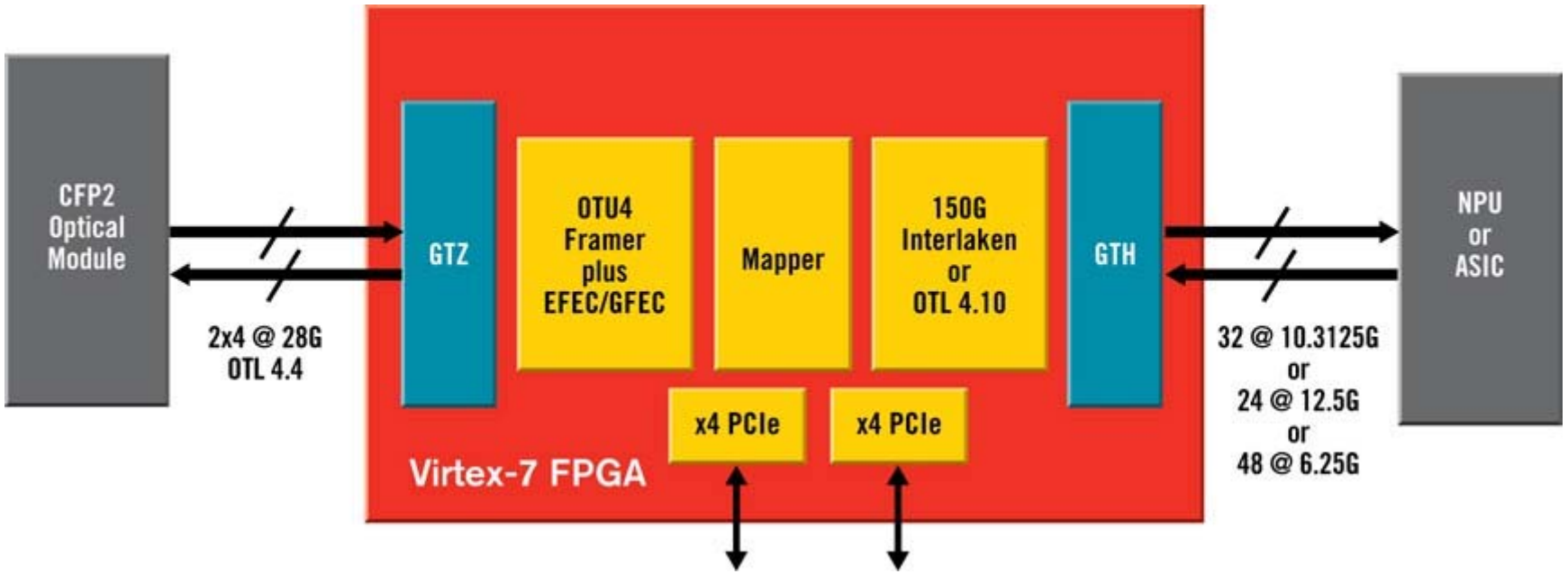
高速通信が使えるFPGAボードの例

- Stratix V GX エディション $24,995
  - SMA インタフェース用：2 チャネル
  - SFP+ インタフェース用：4 チャネル
  - QSFP インタフェース用：8 チャネル
  - CFP インタフェース用：10 チャネル
  - Interlaken インタフェース用：24 チャネル



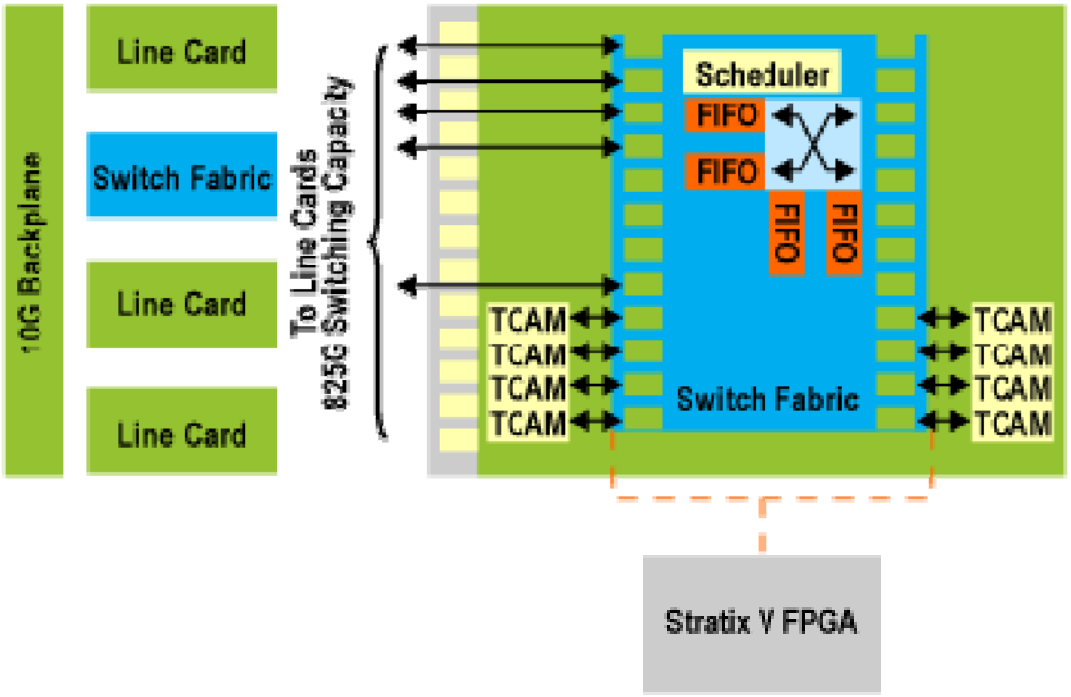
Xilinxの100G x2実装

- Virtex-7を利用
- 28.05Gbpsトランシーバ x8で片方の100Gを実装
- 12.5Gbpsトランシーバ x10でもう片方の100Gを実装



Alteraによるクロスバスイッチの実装例

- Stratix V GXを利用
- 14.1Gbpsのトランシーバ x66間の通信のスイッチング
- ルーティングのためのTCAMを併用



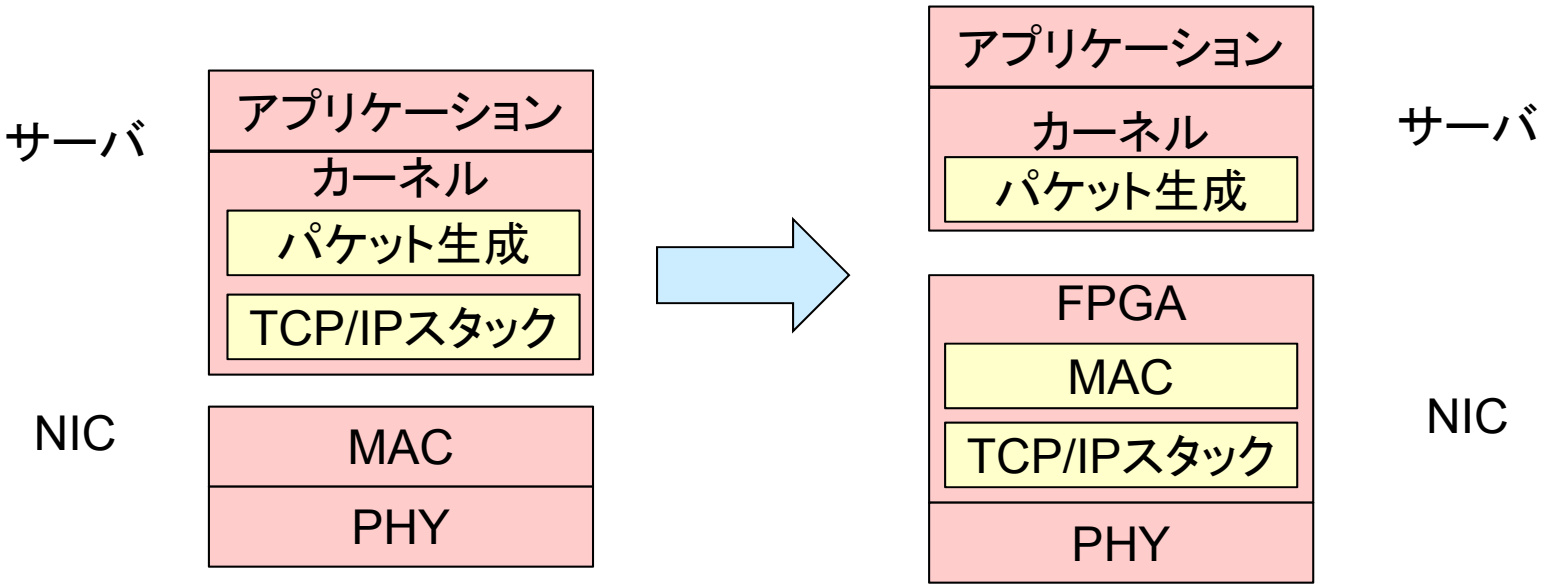
(高速)高頻度トレードにおけるFPGA利用

- HFT: High Frequeny Trading
  - アルゴリズムによる(株式)取引方法の1つ
  - 取引時のマージンを低くするが、高頻度で取引をすることで
  - ミリ秒単位の高速(株式)取引が重要になる
    - “2005円で売り”と”2010円で買い”が出そうならば、”2006円で買って2009円で売る”という
    - 最近だとマイクロ秒とかのオーダに…
- このような取引では取引依頼の少しの遅延が大きな損失に

  →FPGAによる取引依頼部ハードウェア化
  - アルゴリズムの部分は引き続きサーバ部分

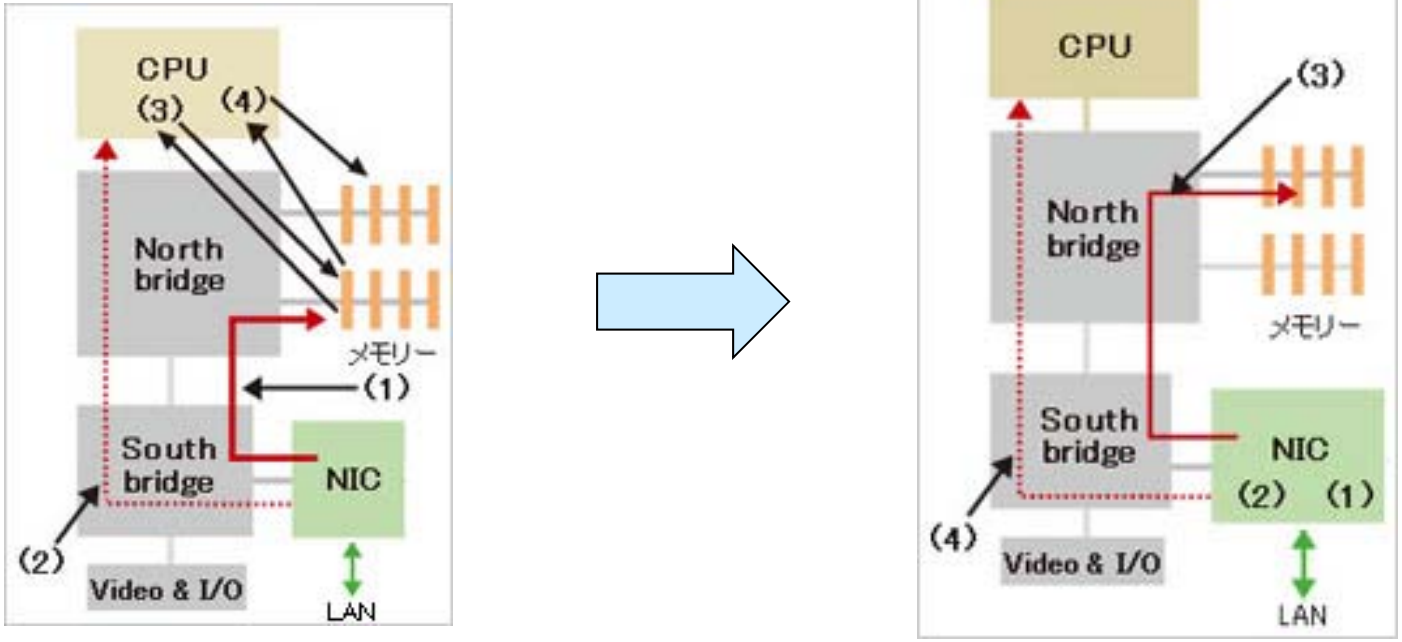
HFTのネットワークにおけるFPGA利用 (1/3)

- 初期: FPGA付きNICによるTCPオフローディング
  - TCPオフローディング: TCP/IPスタックをFPGA側で実行することでサーバ側の負荷を軽減
  - サーバで生成した取引発注の通信内容をFPGA側のTCP/IPスタックにて送信
    - FIXプロトコル:金融取引の標準プロトコル



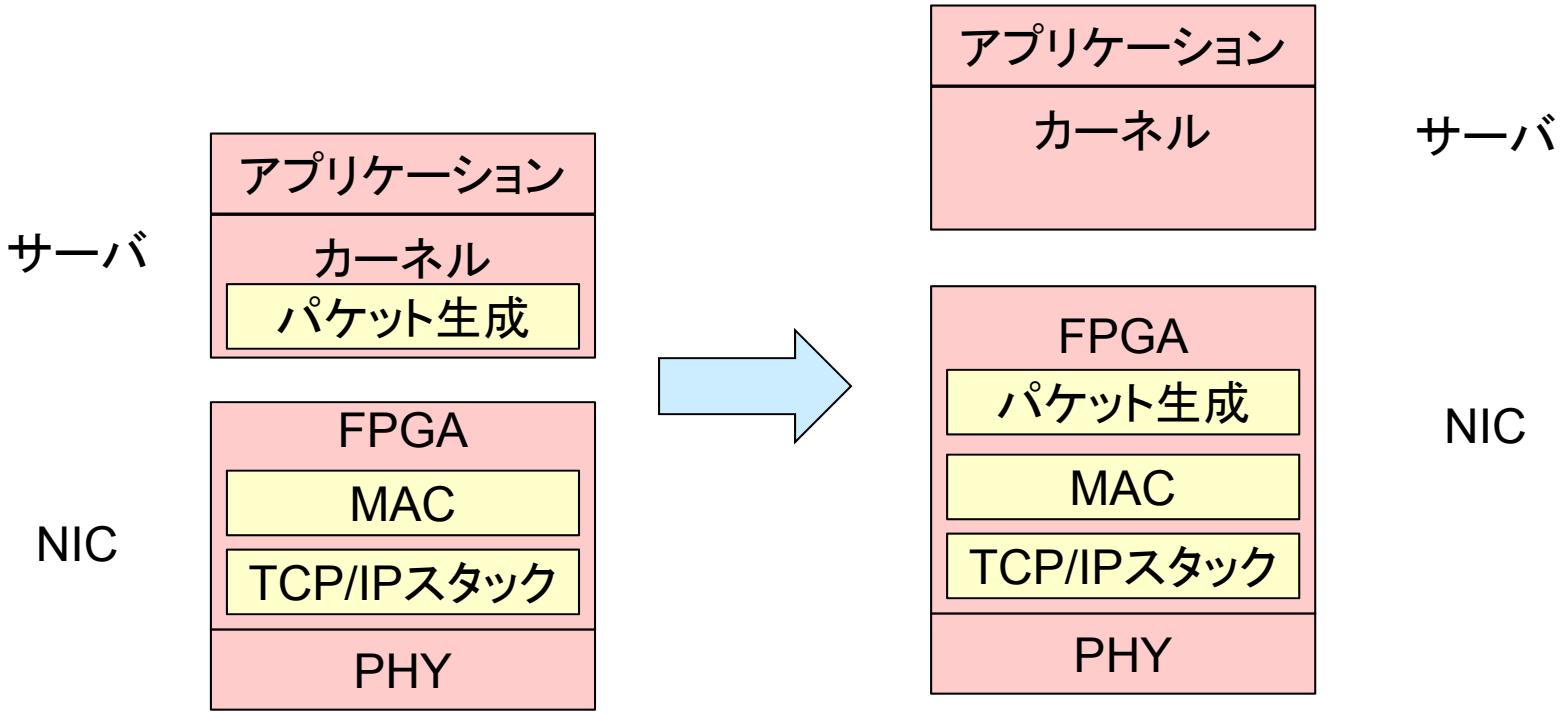
TCPオフローディング

- 従来
  - パケットデータのメモリへの読み書きにCPUが介在
- TCPオフローディング
  - パケットデータはメモリに書き込まれてから受信通知が来る
  - メモリ上のパケットデータに対して送信依頼ができる



HFTのネットワークにおけるFPGA利用 (2/3)

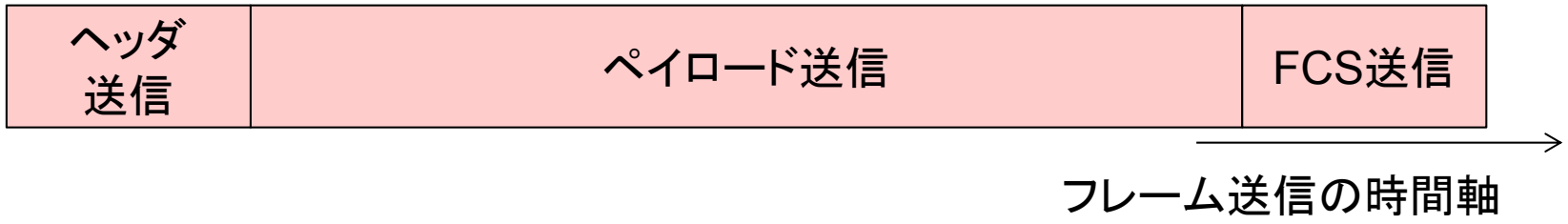
- 中期:
  - 発注の通信をFPGA内部で生成
  - サーバ側は取引発注内容自体のリクエスト処理のみ



HFTのネットワークにおけるFPGA利用 (3/3)

- 最終形: 投機的な取引リクエスト
  - 過去の値動きを元に発注すべき取引内容を予測
  - 最新の値動き結果が来る前に取引内容(のイーサネットフレーム)を送信開始
  - 予定通りの値動き: そのまま送信
  - 予定とは異なる値動き: イーサネットフレームの送信をキャンセル
    - フレーム最後のFCSに誤った値を付与
    - 非常に迷惑な行為なので、当然、証券会社側の確認は取っているはず

→あまりにもえげつないのでHFTは規制される傾向



高速トレードにおけるFPGA(小ネタ)

- J.P.MorganがポートフォリオのリスクシミュレーションにFPGAアクセラレータ利用(2011)
  - x86サーバ数千台並列で8-12時間

    →アクセラレータ付属サーバ40台で4分(120倍の高速化!)
    - 途中でGPUで14-15倍の高速化も行った
- AristaがFPGA内蔵ネットワークスイッチを出しているので、それを使ったソリューションも出てくるかも?

FPGA関連小ネタ

- IP Coreもアップグレードできる(される)
  - 例: Alteraは2013/11に10G/40G/100G Ethernet IP Coreを更新
    - 100Gは55%小型化、70%低レイテンシ
    - 40Gは40%小型化、60%低レイテンシ
    - 10Gは20%小型化、24%低レイテンシ
  - OpenCoresとかでも新しいコアが出ることはある