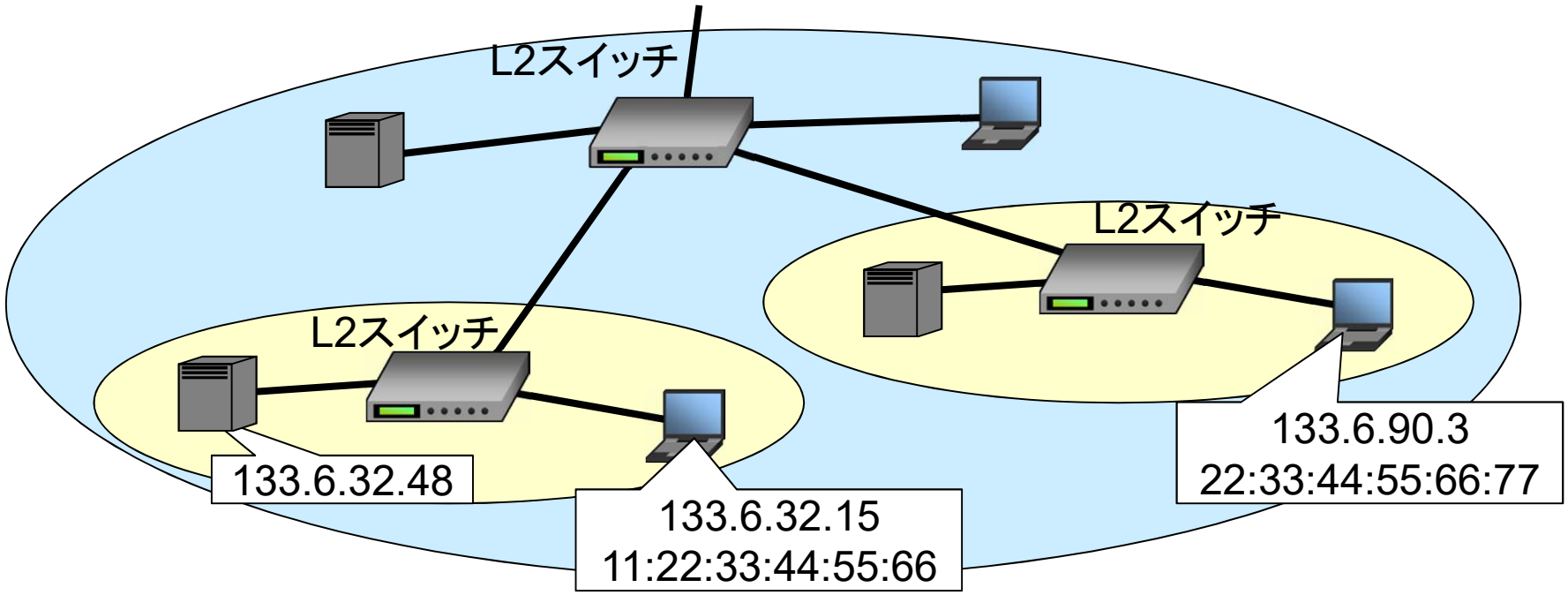
ネットワークスイッチの構成と動作

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

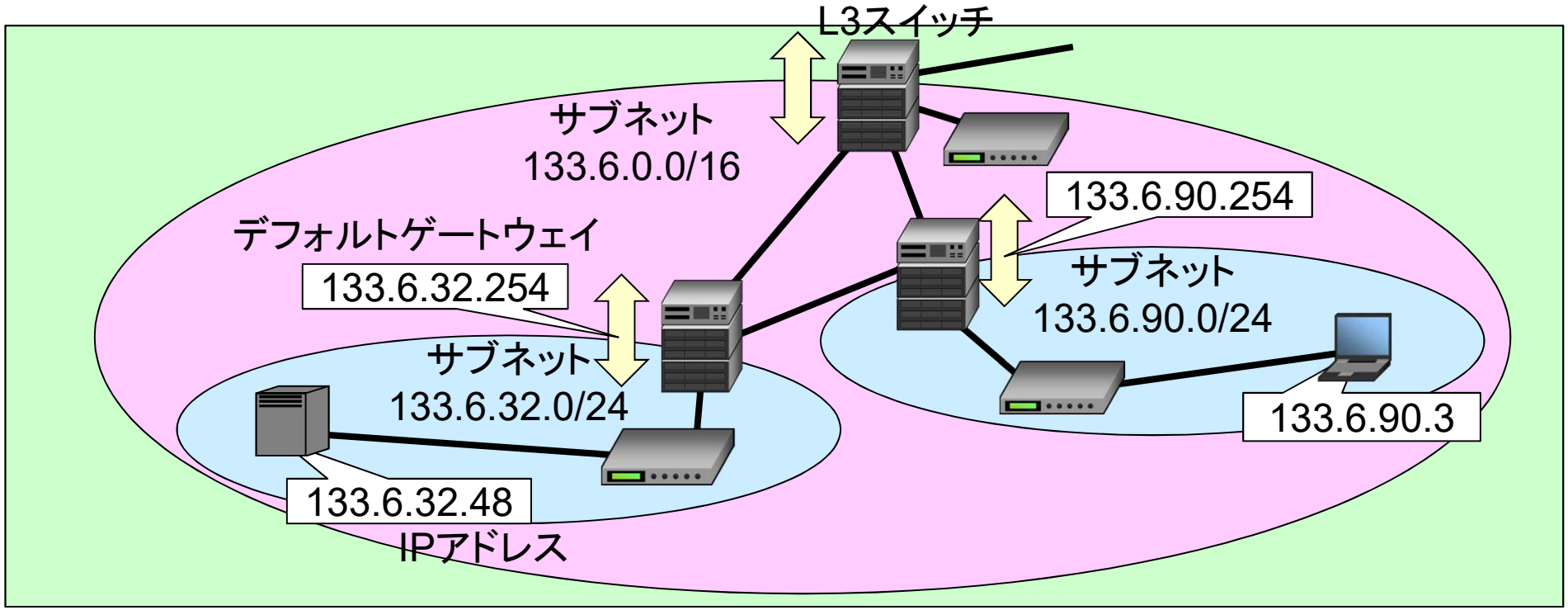
L2における通信の復習

- MACアドレスで宛先を決定
- IPアドレスからMACアドレスへの変換 →送信者がARPで解決
  - 宛先IPアドレス入りARP reqを送り、当該IPアドレスのホストがARP reply
- 宛先MACアドレスが当該L2スイッチの下に無い →ブロードキャスト



L3における通信の復習

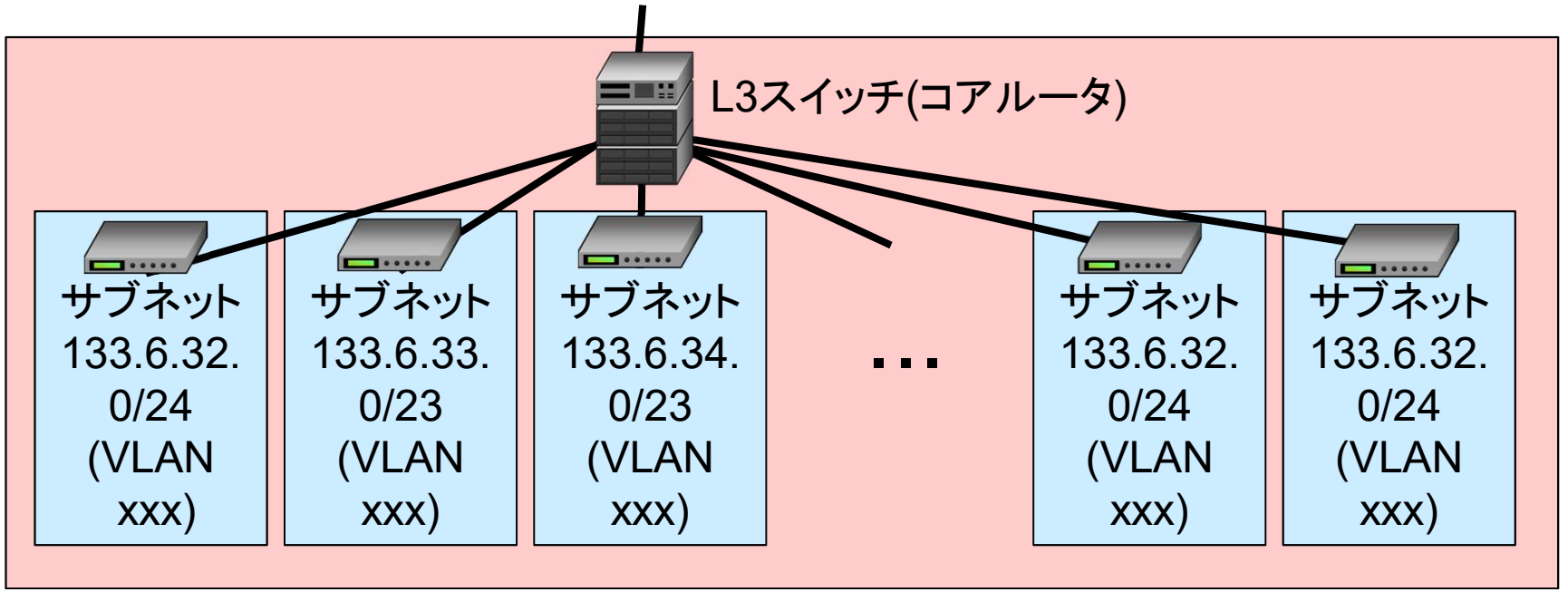
- 宛先IPアドレスを見て送信先を判別
  - L2のブロードキャストはL3スイッチで止まる
- 静的ルーティングや動的ルーティングで設定
  - 動的ルーティング: 各経路の距離や容量の情報をもとに経路選択



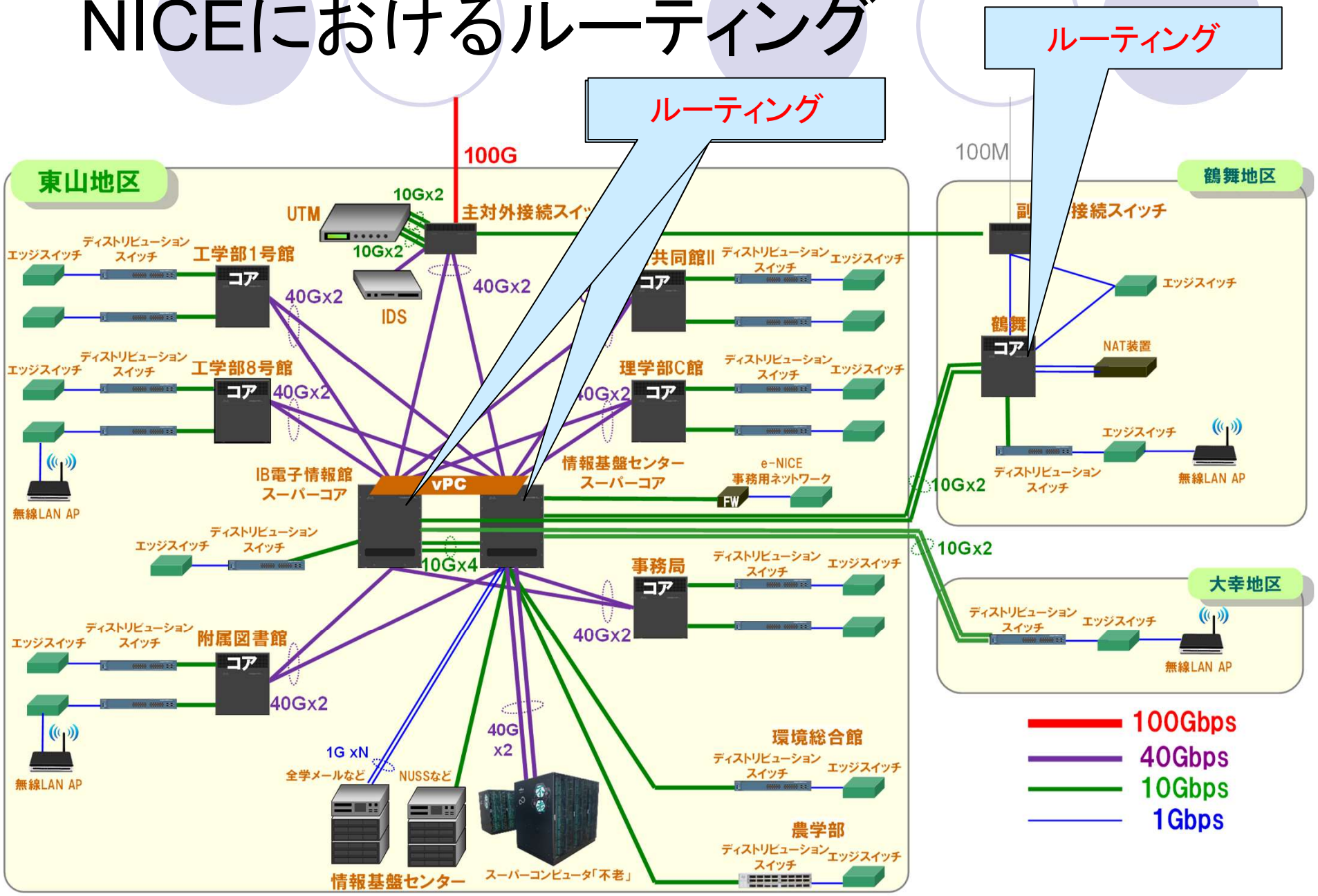
現実的なL3による通信

- L3スイッチは高価なので、各サブネット出口に1つは置けない

  →1つのL3スイッチが複数のサブネット(VLANなど)を管理
  - コア・スイッチとかコア・ルータとか呼ばれたりする

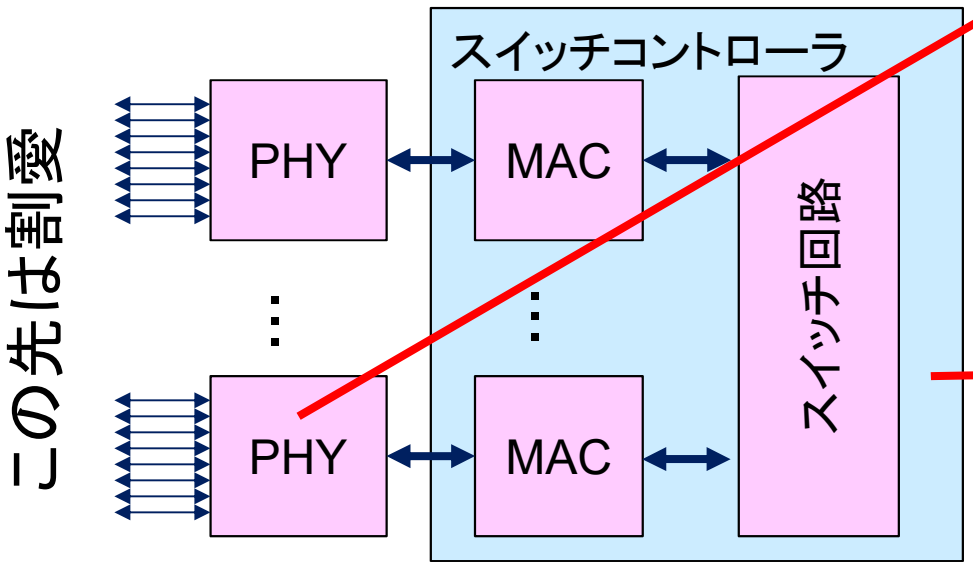


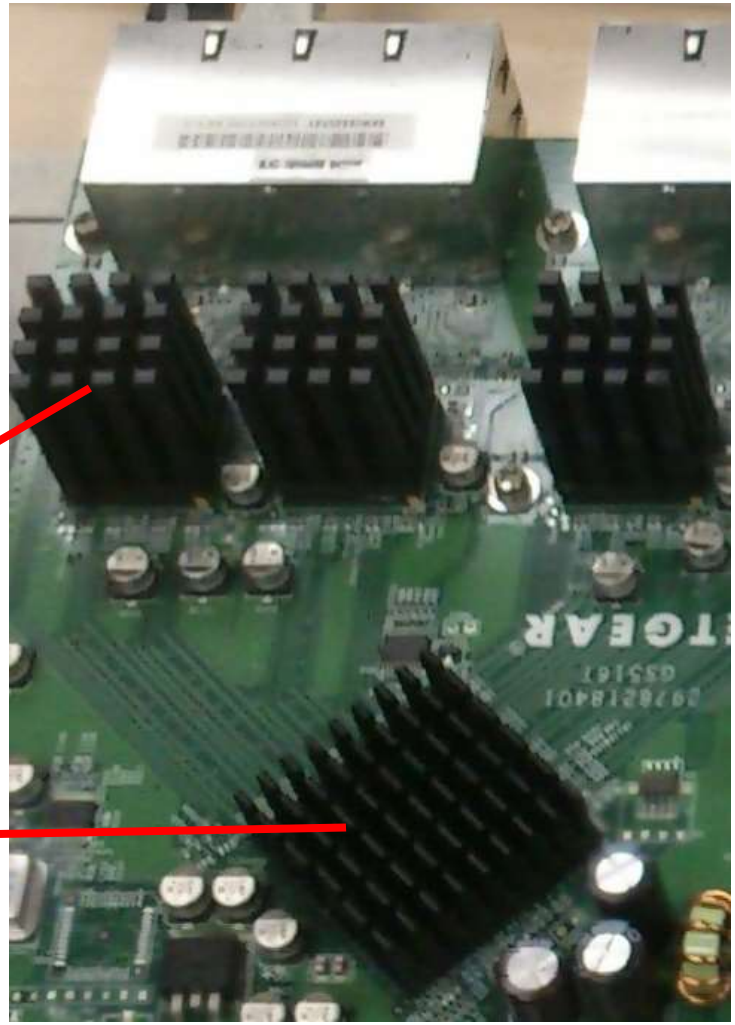
NICEにおけるルーティング



低価格L2スイッチの構成

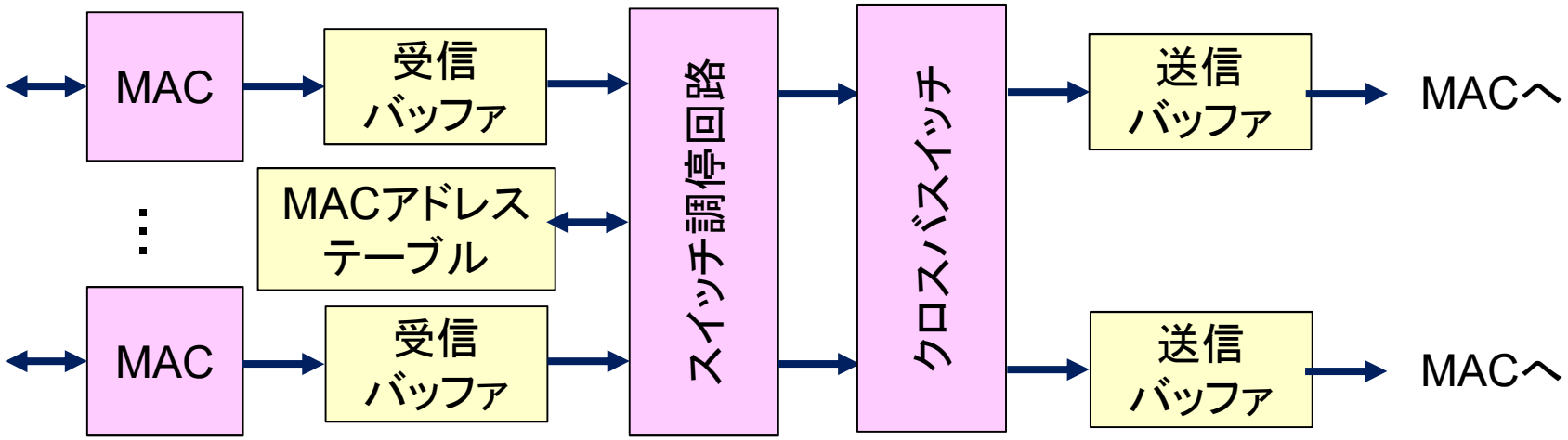
- NICのMACの先がスイッチコントローラに変わったものと考えれば良い
- 注: 要はスイッチングハブの構成です





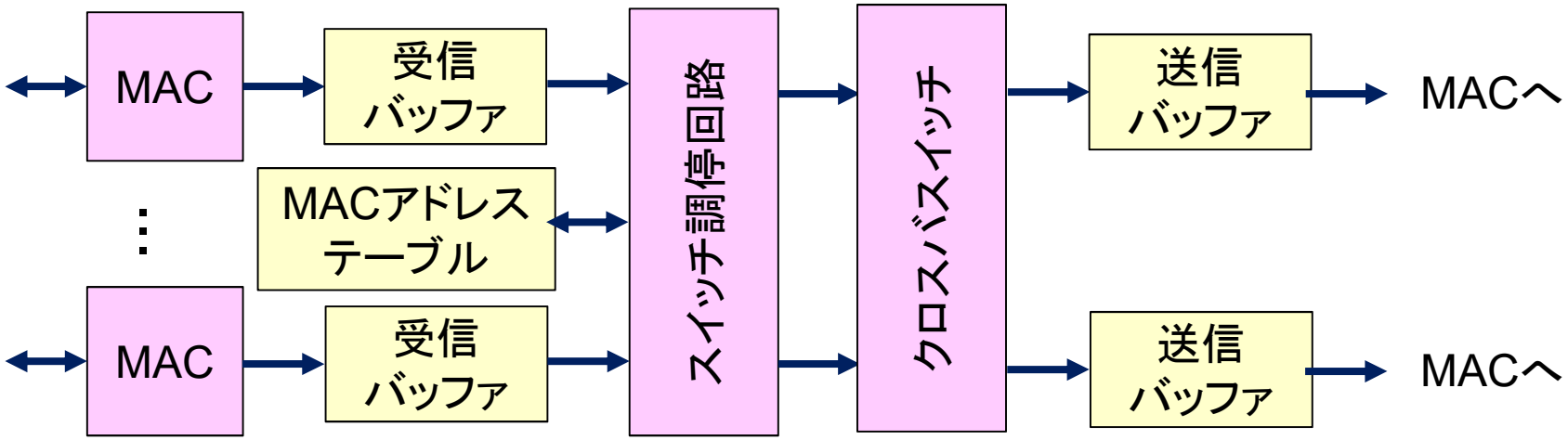
スイッチコントローラの動作(1/2)

1. MACがフレームを受信バッファに格納
2. 受信バッファ中のフレームのMACアドレスでMACアドレステーブルを検索
   - MACアドレステーブルは、どこのポートからどこのMACアドレスの通信がきたかどうかを保存
   - 一致があれば、そのポートのみに送信
   - 一致が無ければ、全ポートへ送信(ブロードキャスト)



スイッチコントローラの動作(2/2)

3. 宛先ポートの送信バッファへのクロスバスイッチの調停ができたら送信
   - 宛先ポートの利用権はFIFOなどで制御
     - もちろん、ブロードキャストも同様



予定通りに行かない時のスイッチコントローラの動作

- MACアドレステーブルが溢れた
  - サブネットがでかいとアップリンクポート側はそのうち溢れる →古いものから削除
  - MACアドレステーブルに対して端末数が多いと削除→再登録処理ばかりになって、負荷急増からスループットorスイッチが落ちる
- 受信バッファが溢れた
  - パケットロスとして、パケットが再送信されて来るのを待ちます
    →効率が悪いのでフロー制御を行う
- フロー制御: パケットバッファが溢れるのを防ぐ制御
  - バッファが溢れそうになったら、フレームの送信を一時中止するように送信元に伝える
  - IEEE 802.3xによるフロー制御
    - 送信元に対してポーズフレームを送信し、受信側は送信を一旦停止
    - バッファが空いたらポーズ解除フレームを送信して通信を再開

スイッチコントローラに関する小ネタ

- いちいちパケットはバッファに保存する?
  →保存しないやり方(カットスルー)もある
  - フレームのMACアドレスを見て、クロスバスイッチの調停を先に済ませてしまう
  - ただし、フレームの途中でエラーが見つかると、調停は無駄になる
- 一応、1000BASEでもリピータハブと同じ動作をする低価格L2スイッチは存在する
  - リピータハブ: CSMA/CDを前提に、来たパケット全てを全ポートに転送するハブ
  - サイズの小さい製品(数ポート程度)であり、出先でのパケットキャプチャのために使うのも便利

高機能な(インテリジェント)L2スイッチ

- スイッチの上でOSが入っていて色々と設定できる
- 基本的に設定できること
  - ポートごとの各種設定
  - L2でのアクセス制御(特定のMACアドレスを遮断、など)
  - 各種ログの読み出し/SNMPによる転送
- 最近ではもっと高機能なことができたりします
  - スイッチのポートやMACアドレス単位で接続時に認証をかける
    - 認証は802.1xが基本だが、独自な物も存在
  - IDS等と連携して、不審な通信のあるポートやMACアドレスを遮断
  - Power over Ethernetによる電力供給
  - Pythonスクリプトの実行

Cisco Catalyst 2960X

ホワイトボックススイッチやオープンルータ

- どちらも好きなOSを載せれるハードウェア
  - ホワイトボックススイッチ: (主に大型の)L2/L3スイッチ
  - オープンルータ: (主に小型の)L2/L3スイッチ
    - WiFiも含めた家庭用ルータと考えればOK
- 最近だと100G/400G対応のホワイトボックススイッチもいっぱいある
- 自前でファームウェアレベルでカスタマイズできる技術者を持つ超大手ITベンダが商用活用している
- オープンルータ関係では、既成品の(WiFi/ブロードバンド)ルータのファームウェアを書き換える形の実装もある
  - OpenWrtが実装としては有名

Network OS(NOS)

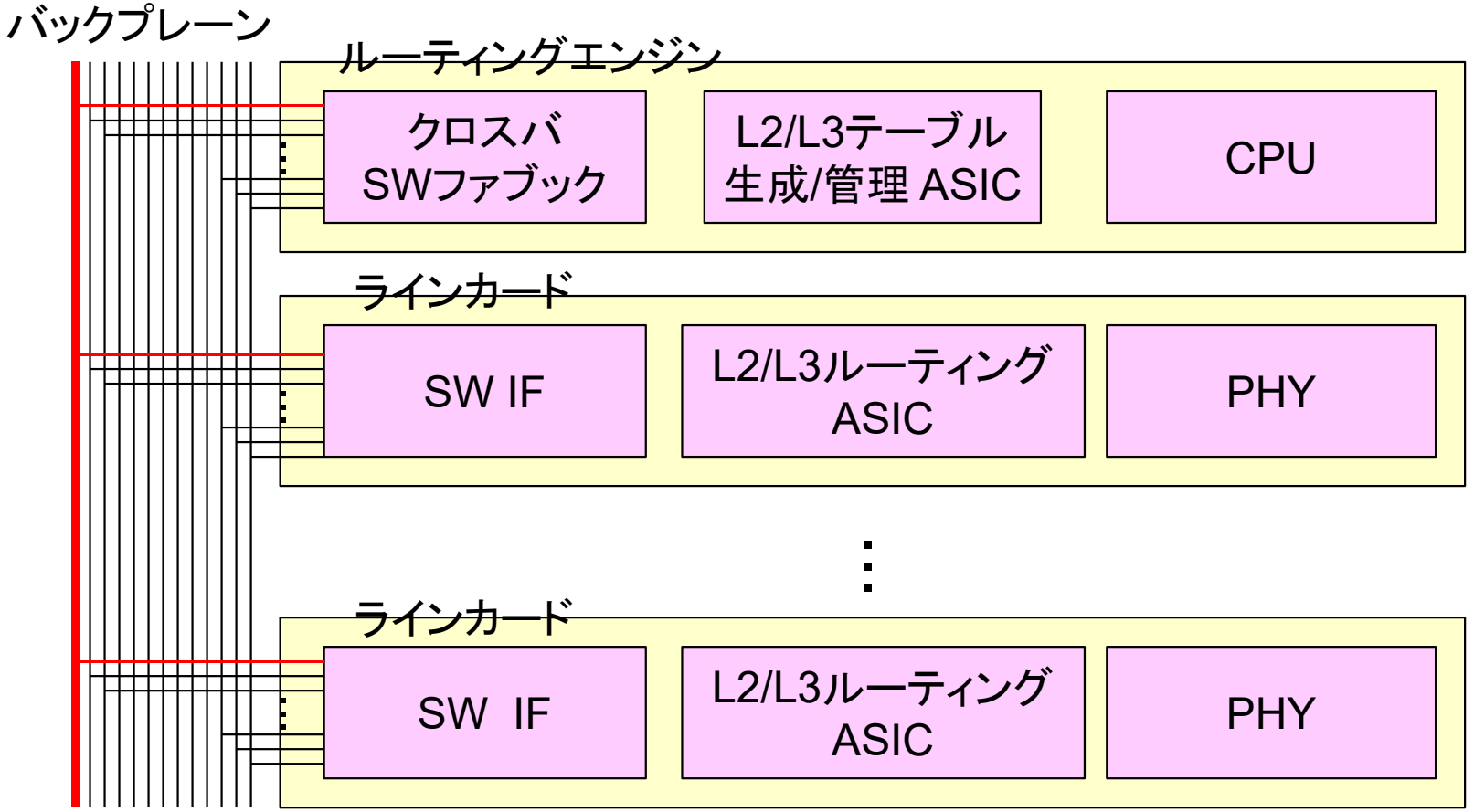
- ネットワークスイッチ(L2/L3)を構成することに特化したOS
  - Linuxベースで作られていたりする
  - 例: Cumulus Linux, Switch Light OS, VyOS, Quagga
    - 商用ではCumulus Linuxが強い(NVIDIAが2020年に買収)
- x86サーバにNICをいっぱいつけてNOS入れてL2/L3スイッチを作るのも面白い
  - 例: VyOS, Quagga
  - NICはUSB NICがお手軽
- Raspberry PiにUSB NICを複数つけてNOS入れてL2/L3スイッチを作るのも面白い

一般的な大型スイッチの構成

- 小型L3スイッチもありますが、せっかくなので大きい物を
  - 最も小型L3スイッチ(ブロードバンドルーター)はL2スイッチ+組み込みプロセッサwith組み込みLinuxによるL3処理なので
- 通常はルーティングエンジンと複数のラインカードから構成される
  - ルーティングエンジンはテーブル生成部とクロスバスイッチから構成
  - ルーティングエンジンで生成した各種テーブルは各ラインカードにも送付
- L3ルーティングテーブルを持つ
  - CAMを用いるが、一部をマスク可能なTCAM(Ternary CAM)を利用
- もちろん、L2による通信機能も持つ
  - MACアドレステーブルなどのL2スイッチの機能もある

一般的な大型スイッチの構成

- 複数1RU程度サイズのユニットをバックプレーンで接続
  - ラインカードとか呼ばれる



各部の処理(1/2)

- CPU
  - スイッチ全体の制御
  - ルーティングテーブル等の作成に必要な情報を管理
  - 管理用OSの実行
- L2/L3テーブル生成/管理ASIC
  - CPUからの指示を受けてL2/L3テーブル生成/管理
  - ルーティングエンジンで作成したテーブルのコピーを各ラインカードに保持
    - FIB(Forwarding Information Base)方式と呼ばれる
- バックプレーン
  - ルーティングエンジンやラインカード間を接続

各部の処理(2/2)

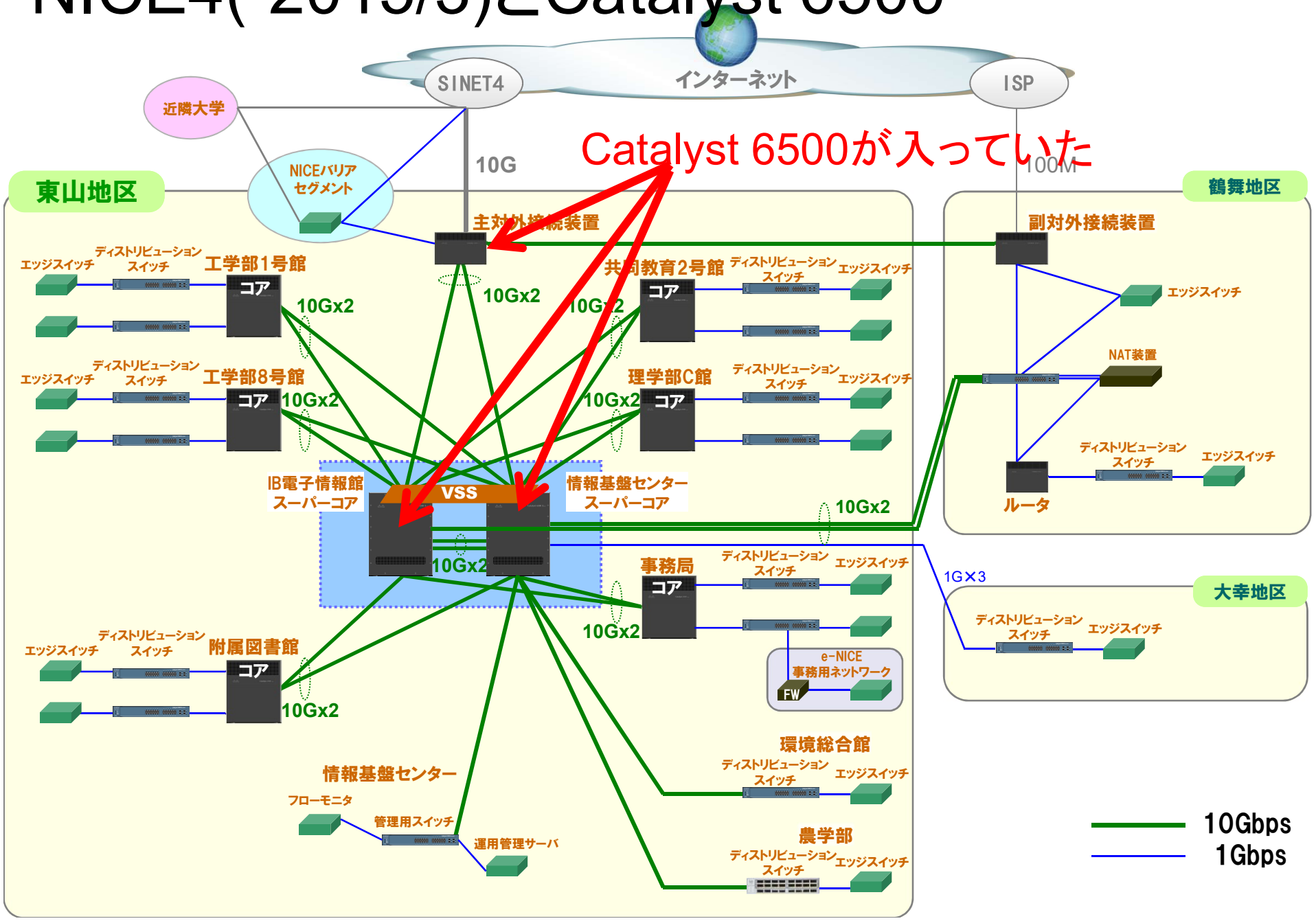
- ラインカード
  - 物理層から受け取ったフレームをバッファリング
  - テーブルにある宛先の送信制御
    - クロスバスイッチファブリックに調停要求→送信
  - テーブルにない宛先を持つフレームの送信をルーティングエンジンのCPUに依頼
  - マルチキャストフレームの複製

大型ネットワークスイッチの実例 (Cisco Catalyst 6500)

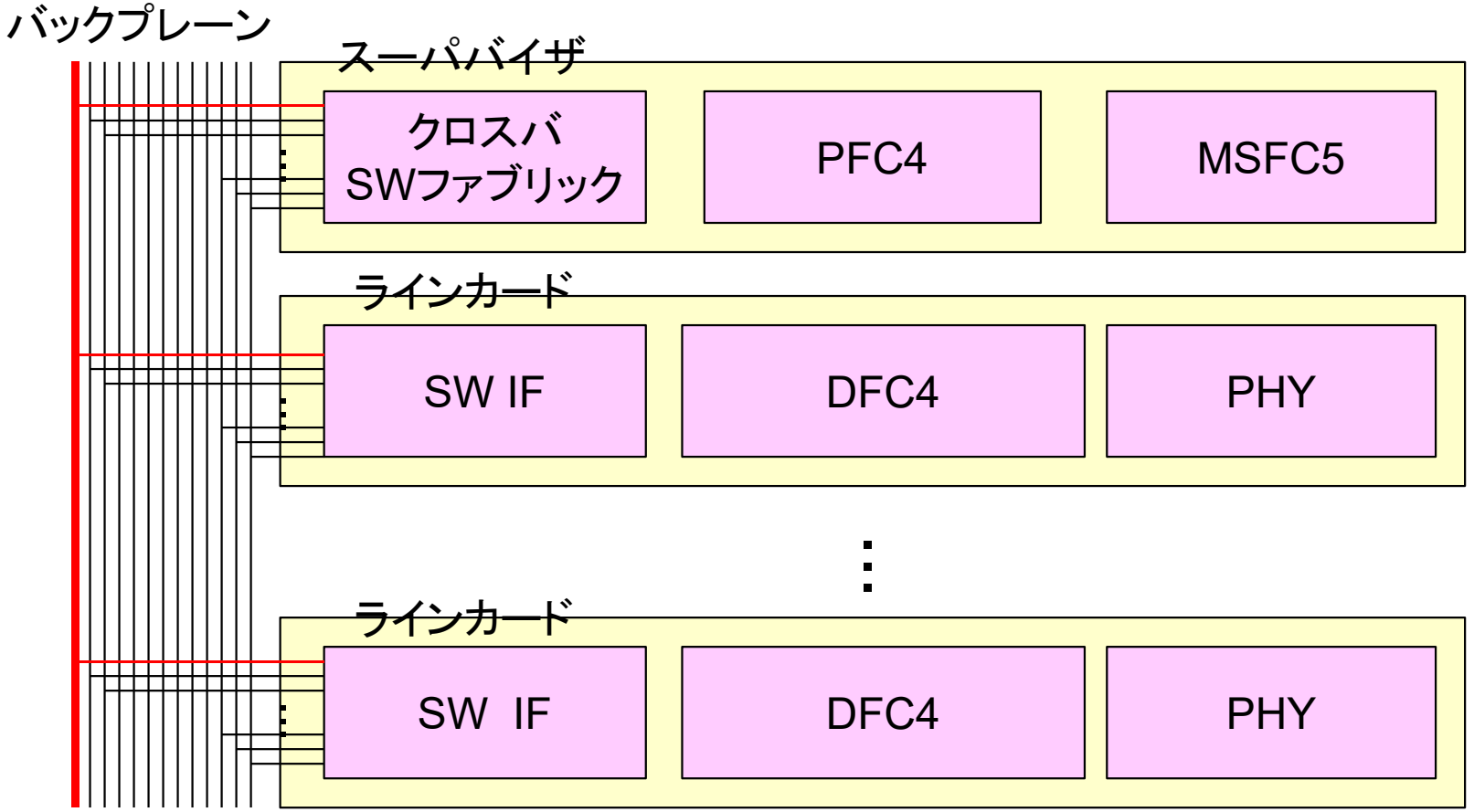
名大内有線ネットワークのNICE4で利用していた

- バックプレーン容量2Tbps
  - ルーティングエンジン(スーパバイザカード)も2Tbps対応
- 1G/10G/40Gイーサネット対応ラインカードを複数接続可能
- Virtual Switching Systemで複数のスイッチを束ねて制御可能

NICE4(-2015/3)とCatalyst 6500



Catalyst 6500の構成

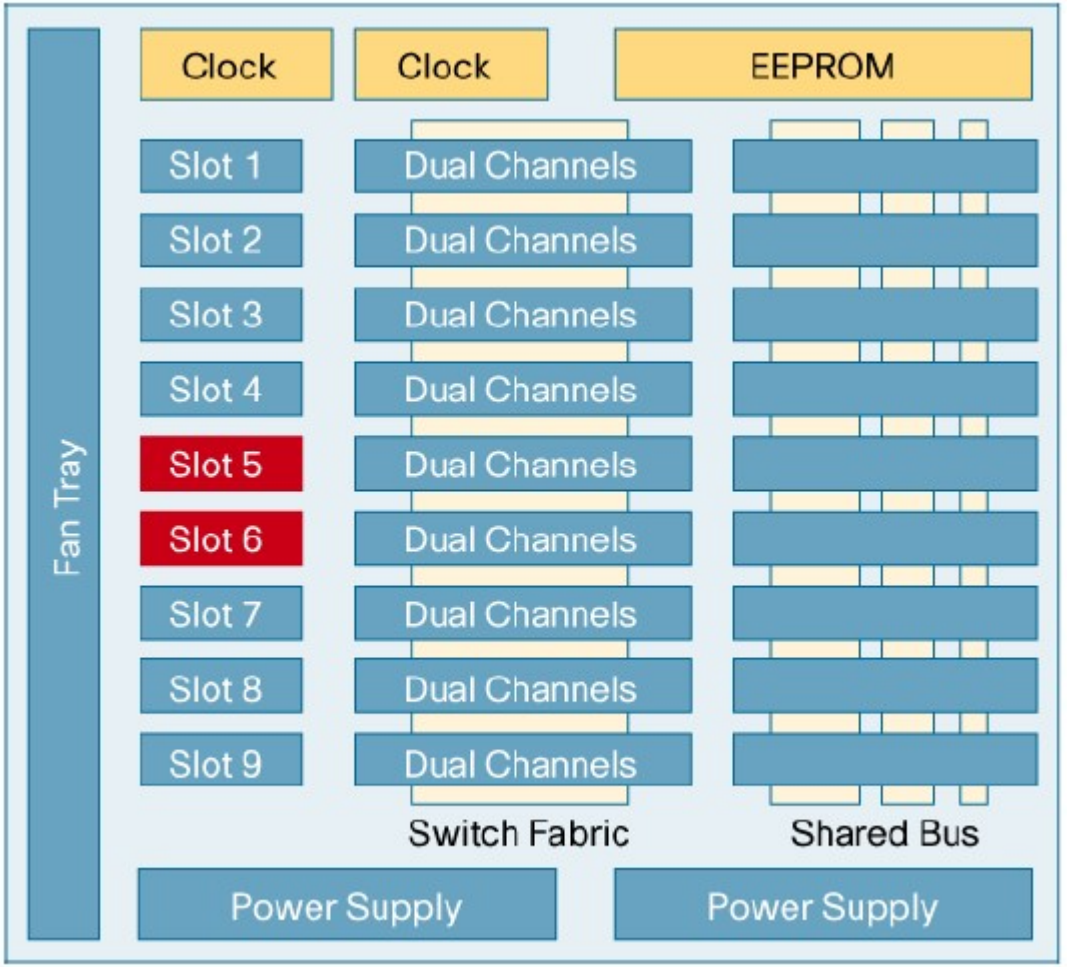


シャーシについているバックプレーン

- 共有バス接続部とクロスバスイッチ接続部に分かれている
  - おそらく、共有バスは全体にブロードキャストする時に便利なため

以下のCatalyst 6500の図は、全て[1]からの引用

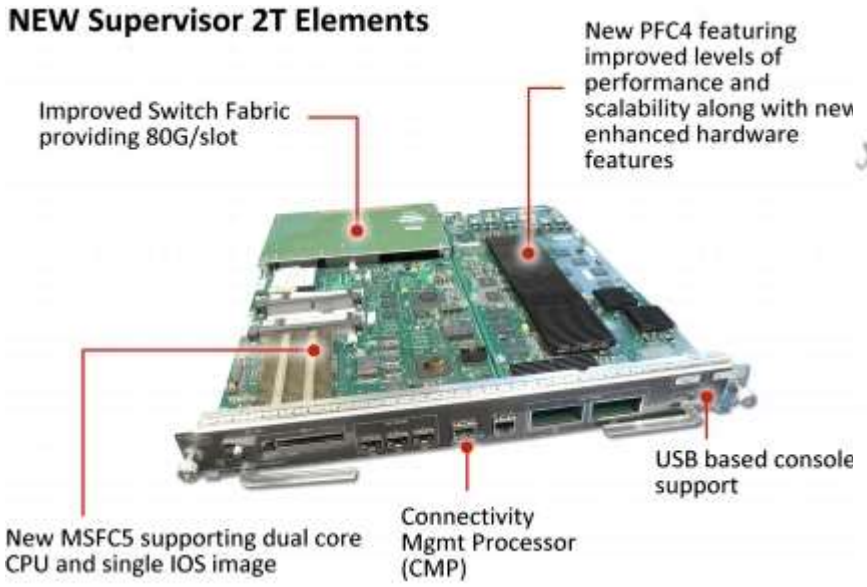
[1] Cisco, "Catalyst 6500 Architecture White Paper," Nov. 2010.

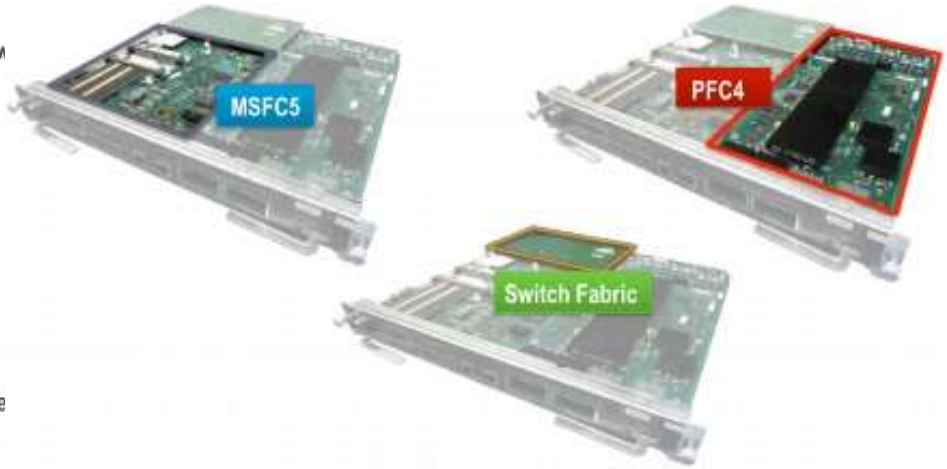


スーパバイザカード2T

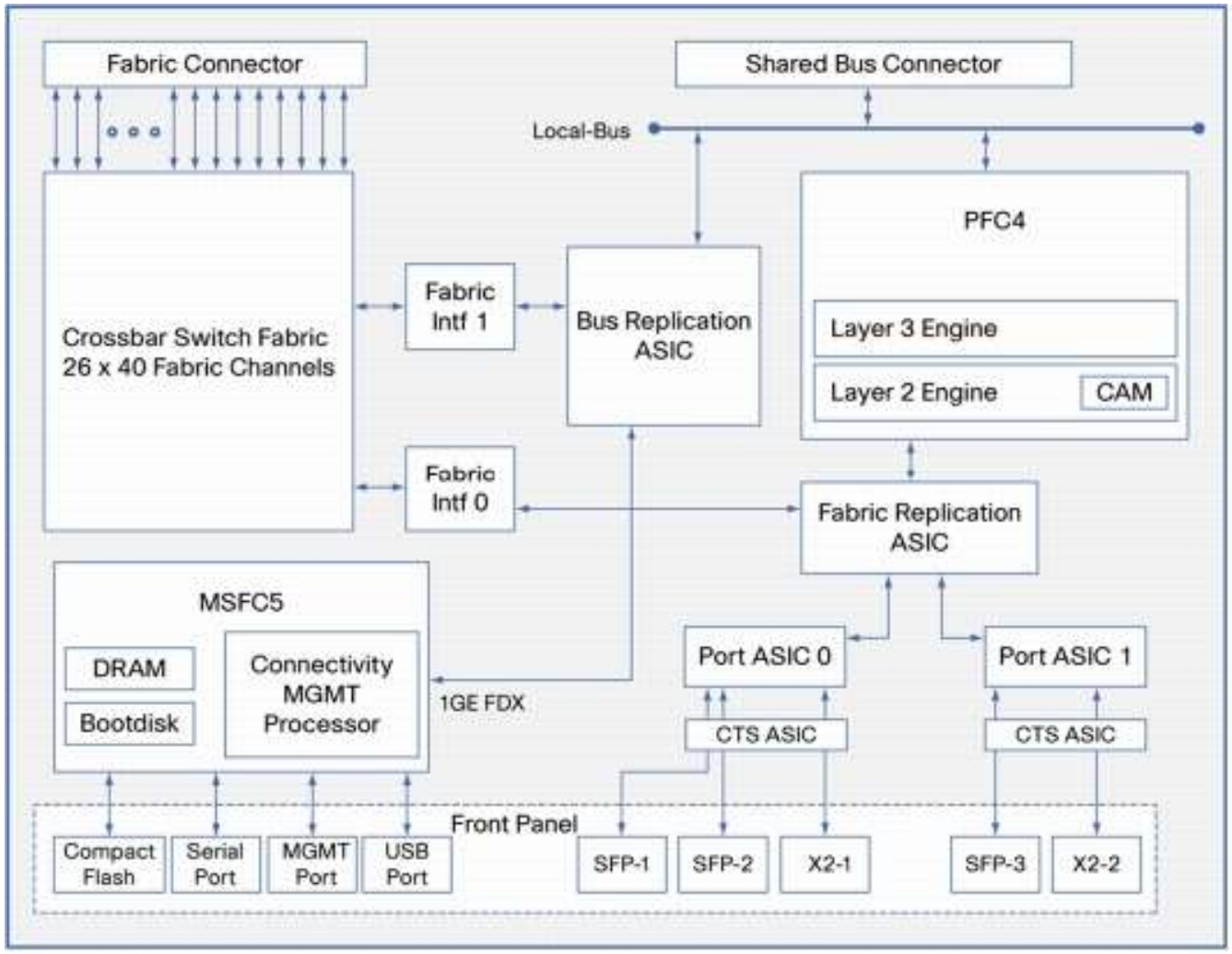
以下のブロックより構成されている

- MSFC5(Multilayer Switch Feature Card): NOS動作担当
- PFC4(Policy Feature Card): L2/L3のルーティング情報やパケット転送情報生成
- スイッチファブリック



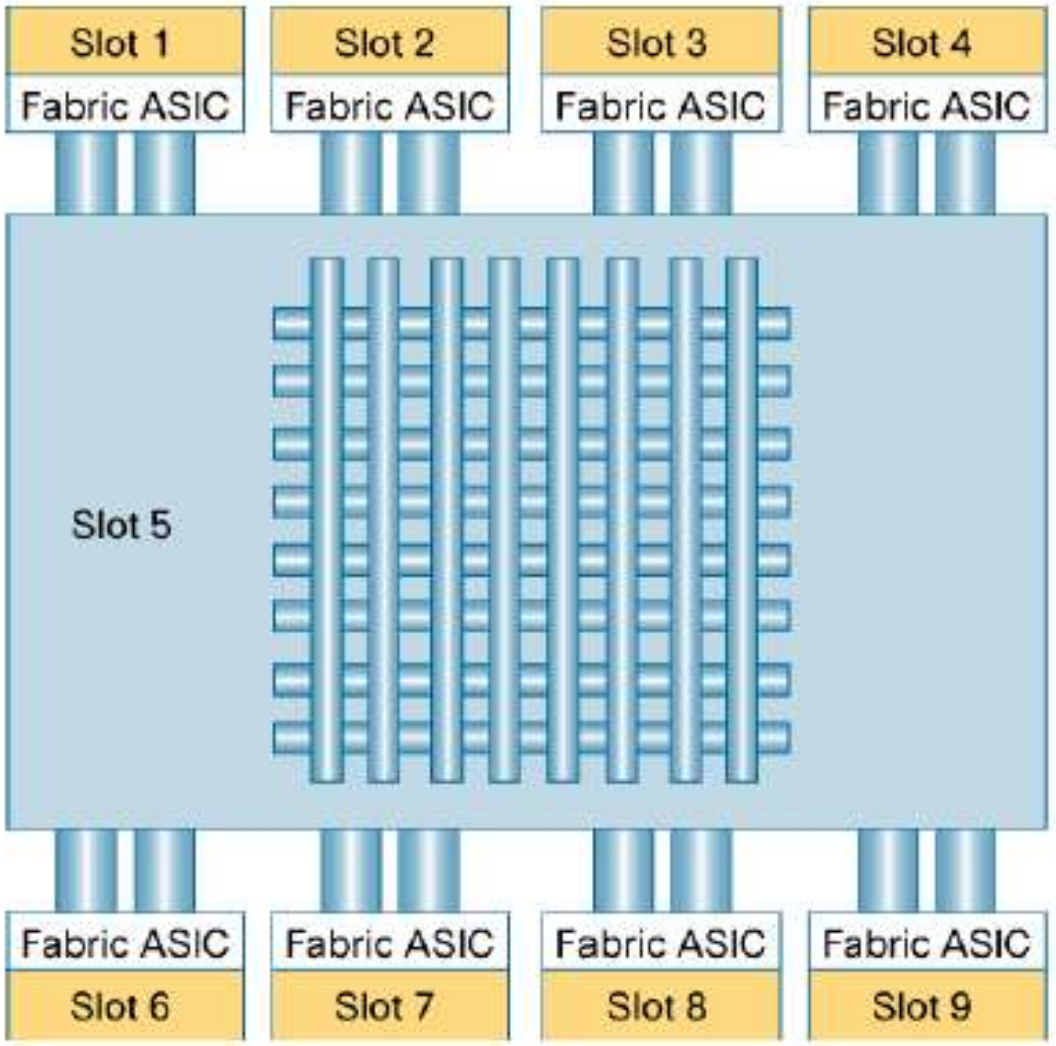


スーパバイザカードのブロック図



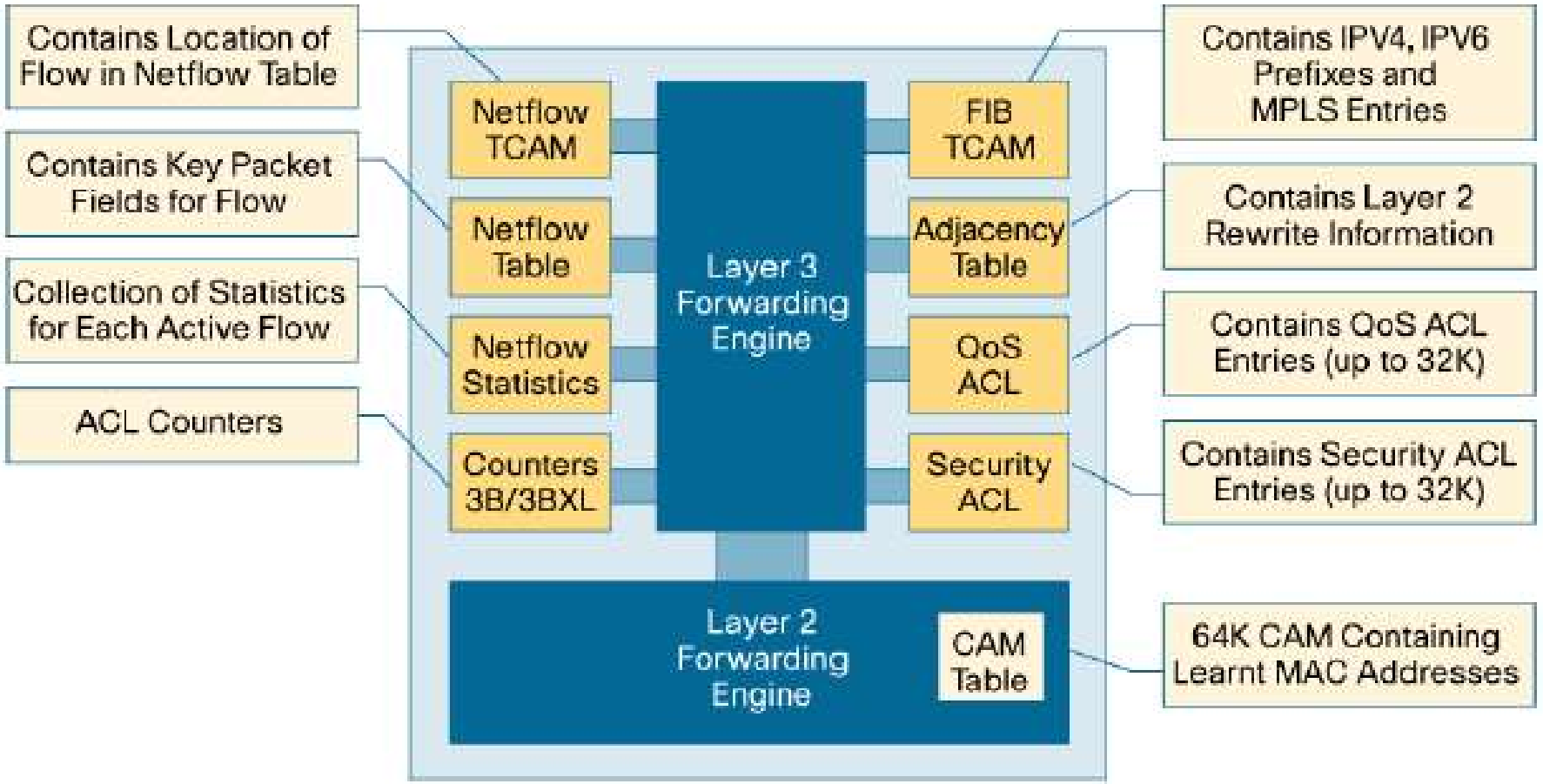
スイッチファブリック部

- 教科書通りのクロスバスイッチの構成



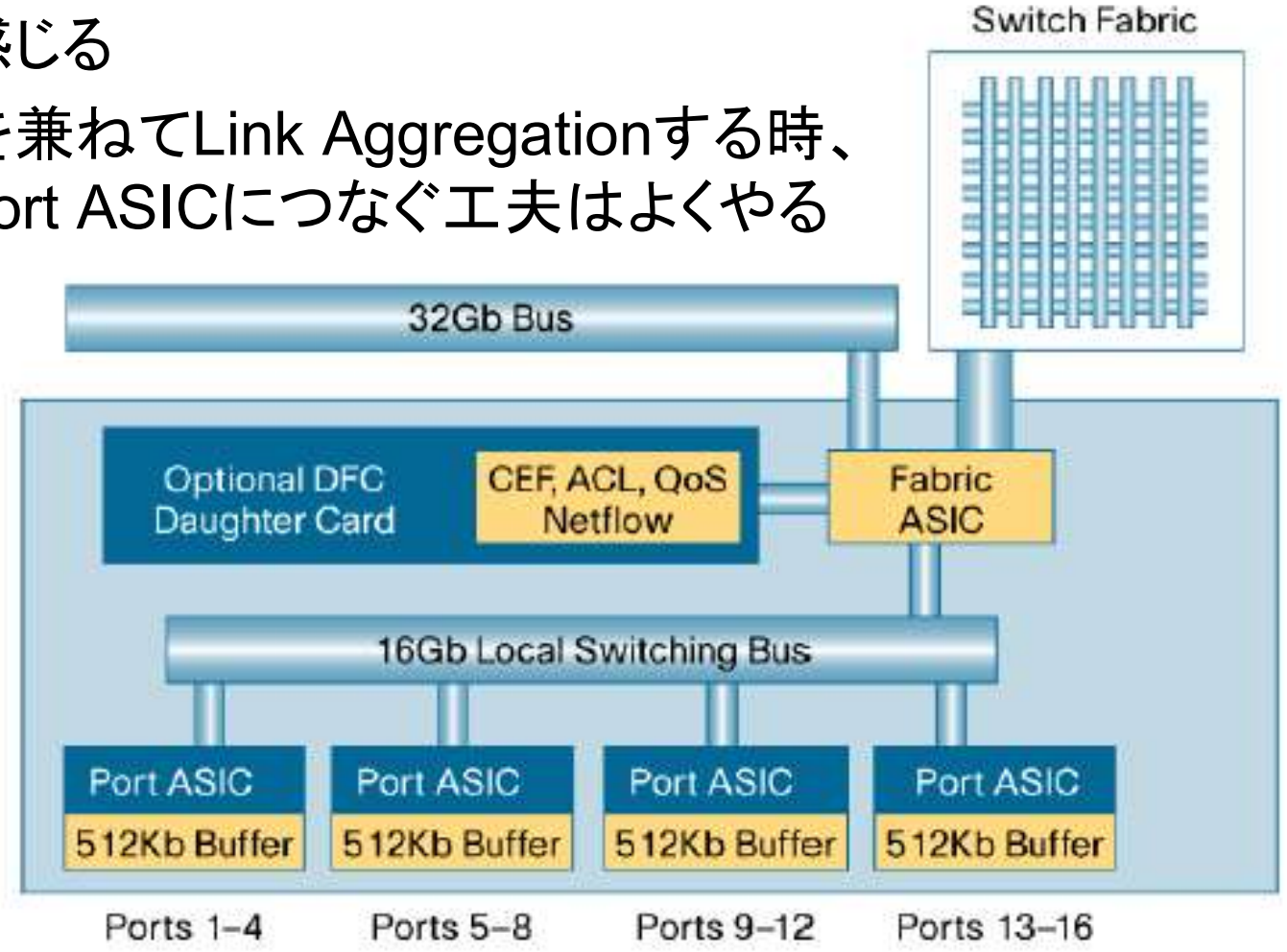
PFC4部

- TCAMで検索をハードウェア化している点が強い

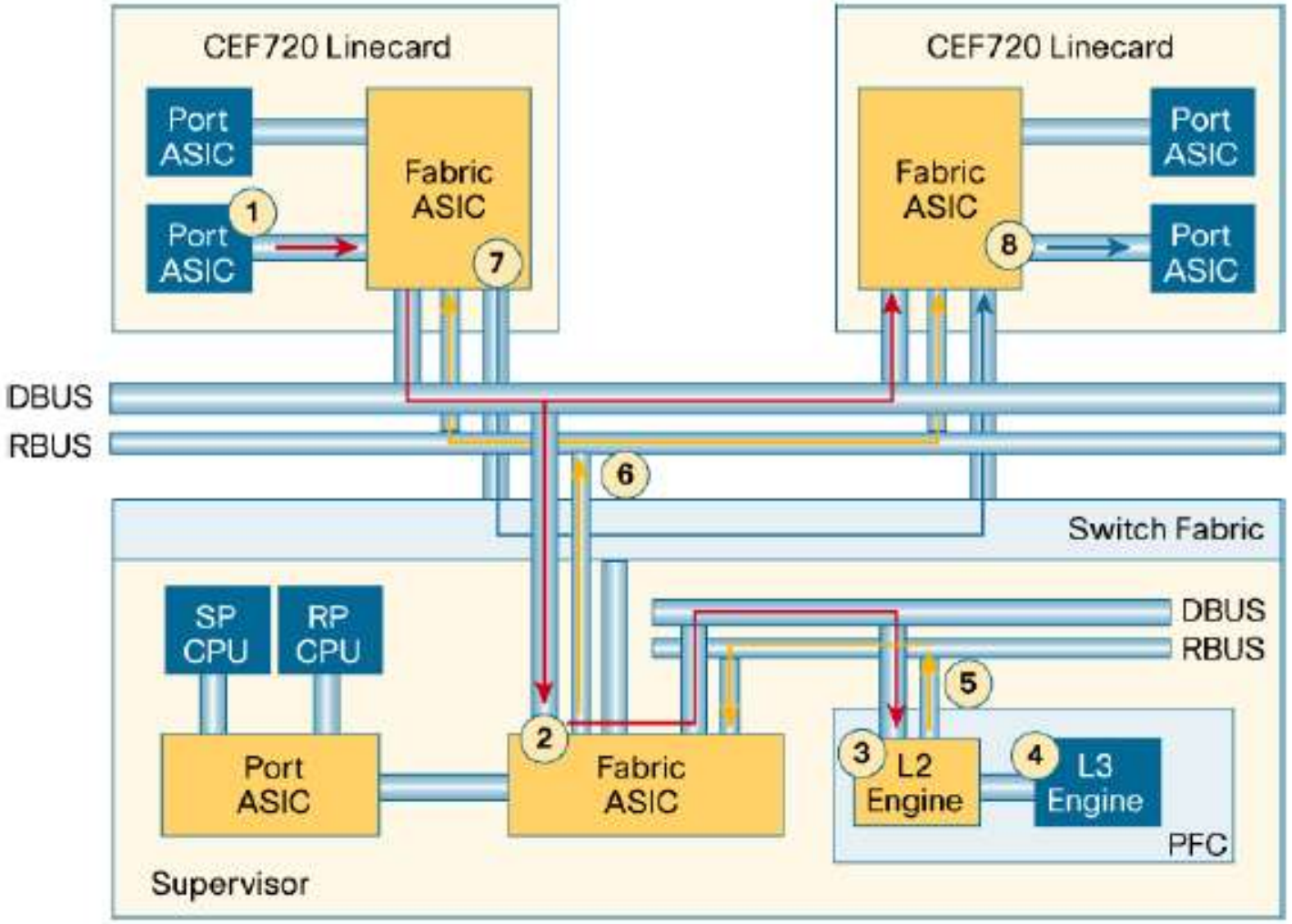


ラインカード内部の構成

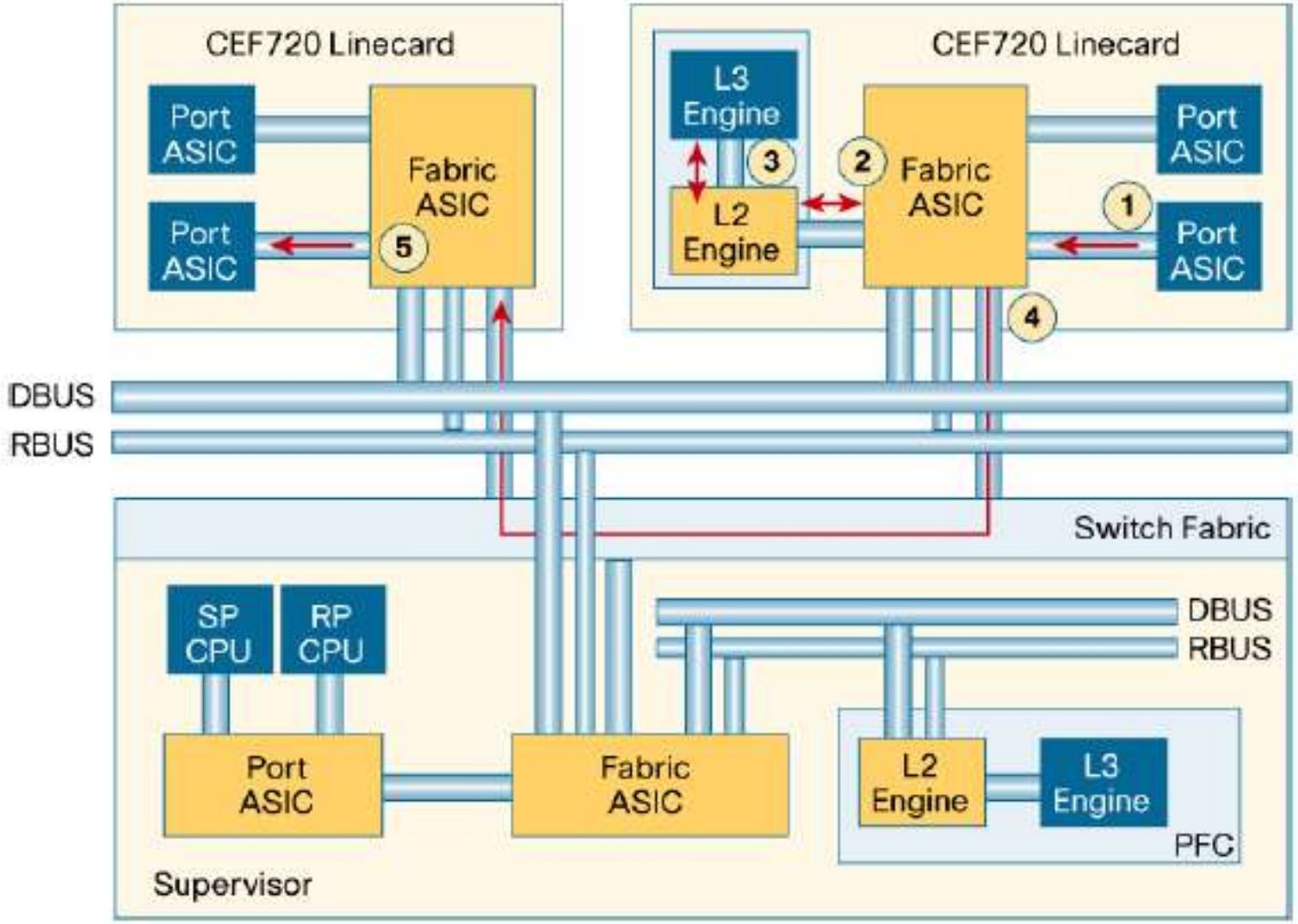
- Port ASICのバッファが大きいあたりに高級機種を感じる
- 冗長化を兼ねてLink Aggregationする時、異なるPort ASICにつなぐ工夫はよくやる



ルーティング情報が無い場合のスイッチング



ルーティング情報がある場合のスイッチング



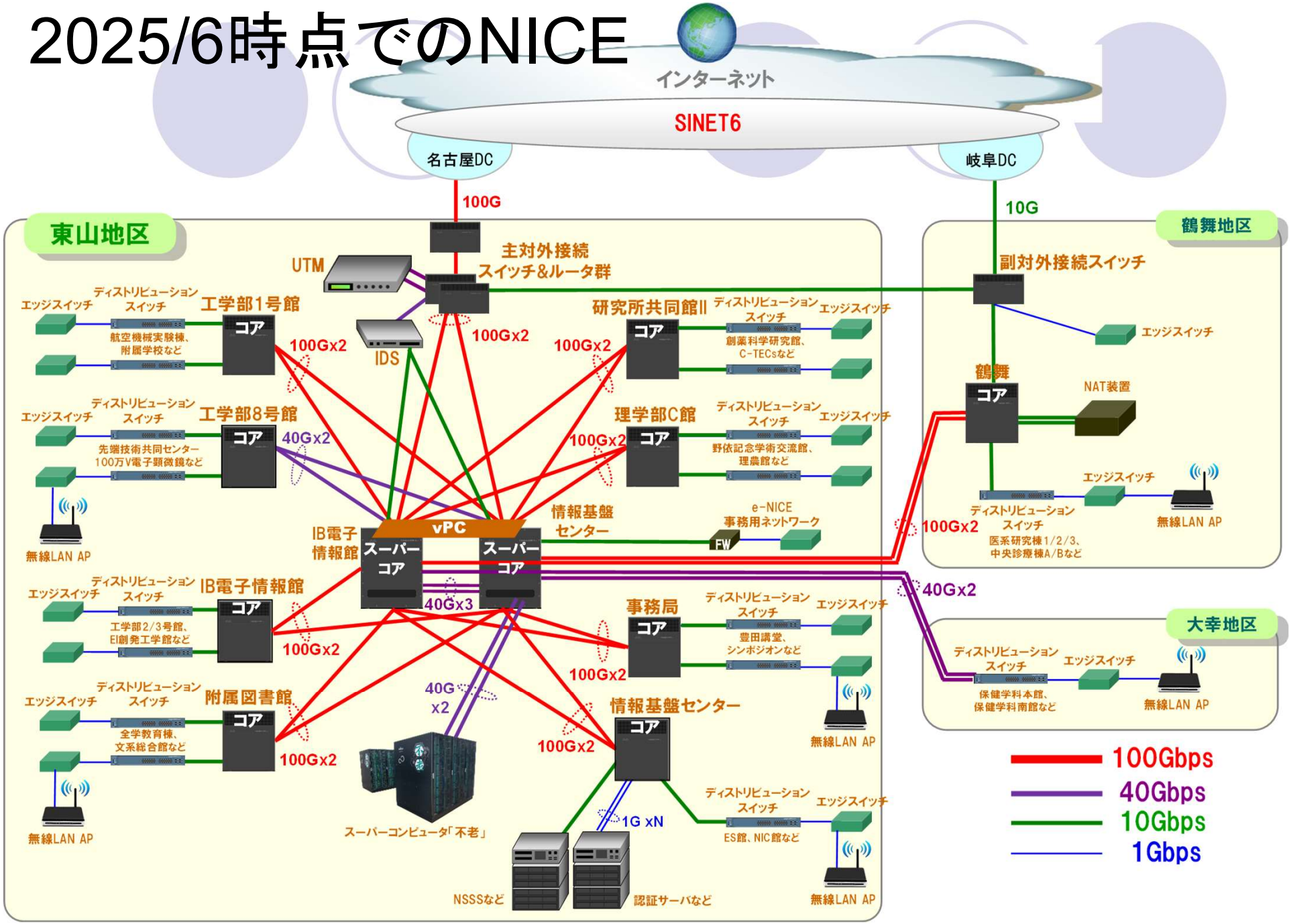
最近の大型ネットワークスイッチへの所感

- 半導体性能の向上により、シャーシ型ネットワークスイッチの需要は減っていく印象
  - 特に、ポート密度、スイッチング性能、ルーティング性能の向上によって
  - 1RUから3RUのボックス型スイッチで従来の10RU未満のシャーシ型スイッチが担っていた集線とルーティングが可能に
    - 昔の40G/100Gスイッチは1RUあたり16ポートだったりしたのが、現在の100G/400Gスイッチは1RUあたり32ポートとか48ポートとか
    - ただし、全ポート400G同時使用できるものは稀(チップあたりのスイッチング容量制限で隣接ポートの制限が入る)
- シャーシ型スイッチだけでなくボックス型スイッチの組み合わせも考慮する
  - シャーシ型でスーパバイザ冗長化対応していた所をボックス型複数台で冗長化
  - 不足する低速のポートをブレークアウトスイッチで提供

キャンパスLANの組み方の実例

- NICE(Nagoya university Integrated Communication Environment)を参考に
  - http://www.icts.nagoya-u.ac.jp/ja/services/nice/
- 現在は6世代目(NICE6)にほぼ移行
  - NICE4までは一括で更新していたが、NICEから部分的に更新
  - NICE4までは対外接続や学内コア間は10Gだったのを、通信需要増加を見越してNICE5で対外100G/学内コア間40Gへ
  - その後、NUWNETとセキュリティ機器を主に増強して一段落
    - 2020/3にBYOD講義を前提に講義室やセミナ室のNUWNETを改善
    - 2022/3にオンライン会議の参加場所を増やすために床面積を1.3倍へ
  - 老朽化でコアをNICE6にする時に学内100G化/対外400G準備
  - 次は、エッジスイッチ(情報コンセント)のmGig化を考慮中
    - さすがに全部やると投資効率悪いので、将来的に1G以上の通信が想定される講義室の無線LAN APとか研究室とか
      - 一部ポートのみmGigな低コストなスイッチもある

2025/6時点でのNICE



スイッチの階層の組み方(1/2)

- 対外接続スイッチ
  - 主対外外続は情報基盤センターからSINET名古屋DCへ100Gで
  - インターネット上の1ASとして自由に接続構成を選べるようBGPフルルートを利用可だったが、今はフルルート未対応
    - 以前はSINETとISPで経路冗長を想定していたためフルルート対応
    - どんどんIPv4もIPv6も経路数が増えてフルルートはコスト高い
    - SINETとの間でのみBGPの経路冗長運用
  - 副対外接続は鶴舞からSINET岐阜DCに10Gで接続
    - 鶴舞<->東山間が切れた時にはSINETをIPsecVPNで通る経路へ切替
    - SINETは愛知県に岡崎DCも持つが、費用差とSINET構成を見て岐阜DCを選択
    - 以前は帯域保証ありISP契約だった
  - この周りにIDS、ファイアウォール兼UTM、アンチウィルスゲートウェイ、フローモニタ、など

スイッチの階層の組み方(2/3)

- (スーパ)コアスイッチ
  - スーパコアスイッチは情報基盤センターとIB北棟
    - vPCという方式で仮想的に1台のスイッチとして運用(冗長化)
    - 学内のVLAN(L2)をルーティングするL3スイッチ
  - コアスイッチは学内に7箇所
    - NICE4までは学内のルーティングも実施
    - 鶴舞のコアスイッチは引き続きルーティングを実施中
- ディストリビューションスイッチ
  - 基本的に各建物に1台存在
    - 情報基盤センターのサーバ室にはサーバ室向けの独立したディストリビューションスイッチあり
  - 複数のエッジスイッチの通信とコアスイッチの間をとりもつ
    - 基本的に、ディストリビューションスイッチまでは光、その先はUTP
    - 「10Gで接続したい」とここまで部屋から光ファイバ引っ張る研究室も

スイッチの階層の組み方(3/3)

- エッジスイッチ
  - 基本的に、各建物の各フロアに1台
    - 1フロアが広い建物だと、情報コンセントまでのUTPが100mに近づくので、複数台設置(情報コンセントから先につながれるUTPの長さも考慮)
  - 一部の小規模な建物は、エッジスイッチだけが存在
- 無線LANアクセスポイント
  - 802.1xの認証サーバは学内3箇所に(2019あたりに増強)
    - …が、これでも最近は認証が遅いことがあるので、増強するか検討中
  - Web認証サーバは情報基盤センターに
  - 無線LAN用のUTM、無線LAN用NAT(IPv4)、IPv4/IPv6分離ルータもあって、無線通信量増加に伴うここの増強も頭が痛い
- 研究用100Gブレークアウトスイッチ
  - 安価な3rd party光モジュールでの接続OKな基幹グレードでないSW
  - 一部コアスイッチに試験的に導入して運用中

キャンパスネットワーク設計の検討点(1/2)

- 対外接続においてBGPフルルートを受けてルーティングするか?
  - BGPフルルートを受ける性能が対外接続スイッチに必要となる
    - 現状ではIPv4で90万経路ほどが必要になるが、将来のネットワーク細分化を考えると最低200万経路は欲しい
    - 最近の高級機種はIPv4/IPv6合わせて1000万経路とかも
    - 全ルートをTCAMに入れることができるようなL3スイッチは高価
  - The Internetへの出口が1つならばBGPフルルートは不要
    - 複数の出口があっても広範囲で動的に経路選択をやらない限り不要(例:どの出口も日本国内とか特定の広域ネットワークとか)
    - どの経路でも特定の広域ネットワークの下に出るならば、特定の広域ネットワークとの間のプライベートASを利用可能

→2025/3の更新でNICEはフルルートやめました(コストダウン)

- その前に「副対外線もSINETで十分」という判断もあった

キャンパスネットワーク設計の検討点(2/2)

- 学内のルーティングはどこでやる?
  - 現状ではルーティングをスーパコアスイッチに集約中
    - メーカも集約する方向を売りにしている
    - L3機能を持ったスイッチ自体が高価(保守費用も含めて)
  - ただし、鶴舞のVLANを東山でルーティングするのは無駄が多いので、鶴舞のコアスイッチでルーティング
  - 最近だと、組織内ネットワークを監査するUTMをコアスイッチにしてしまう事例も見られる
    - UTMメーカが「一体化してコストダウン」をうたっている
    - 個人的には、あまりにも集中した単一障害点を作ると、トラブル発生時に広い範囲で影響が出るので不安

スイッチ調達時に主に気にする性能(1/2)

- スイッチング容量(単位: bit per second)
  - 例: 10Gbps 32ポート、40Gbps 4ポート
    10G x 32 x 2(双方向) + 40G x 4 x 2(双方向) = 960Gbps
  - 普通は全ポートが同時にフルに通信しても問題ない性能を持つ
- パケット転送性能(単位: packet per second)
  - こちらは全ポートが同時にフルにショートパケット(64byte)で通信しても耐えれるものはまず無い
  - ショートパケットが大量に来る用途では注意
- 各種プロトコル(ルーティング、管理、QoS、など)に対応しているか?
- (最近だとネットワーク接続への認証機能とかセキュリティ機器との連携とかも要件に入ってくるかも)

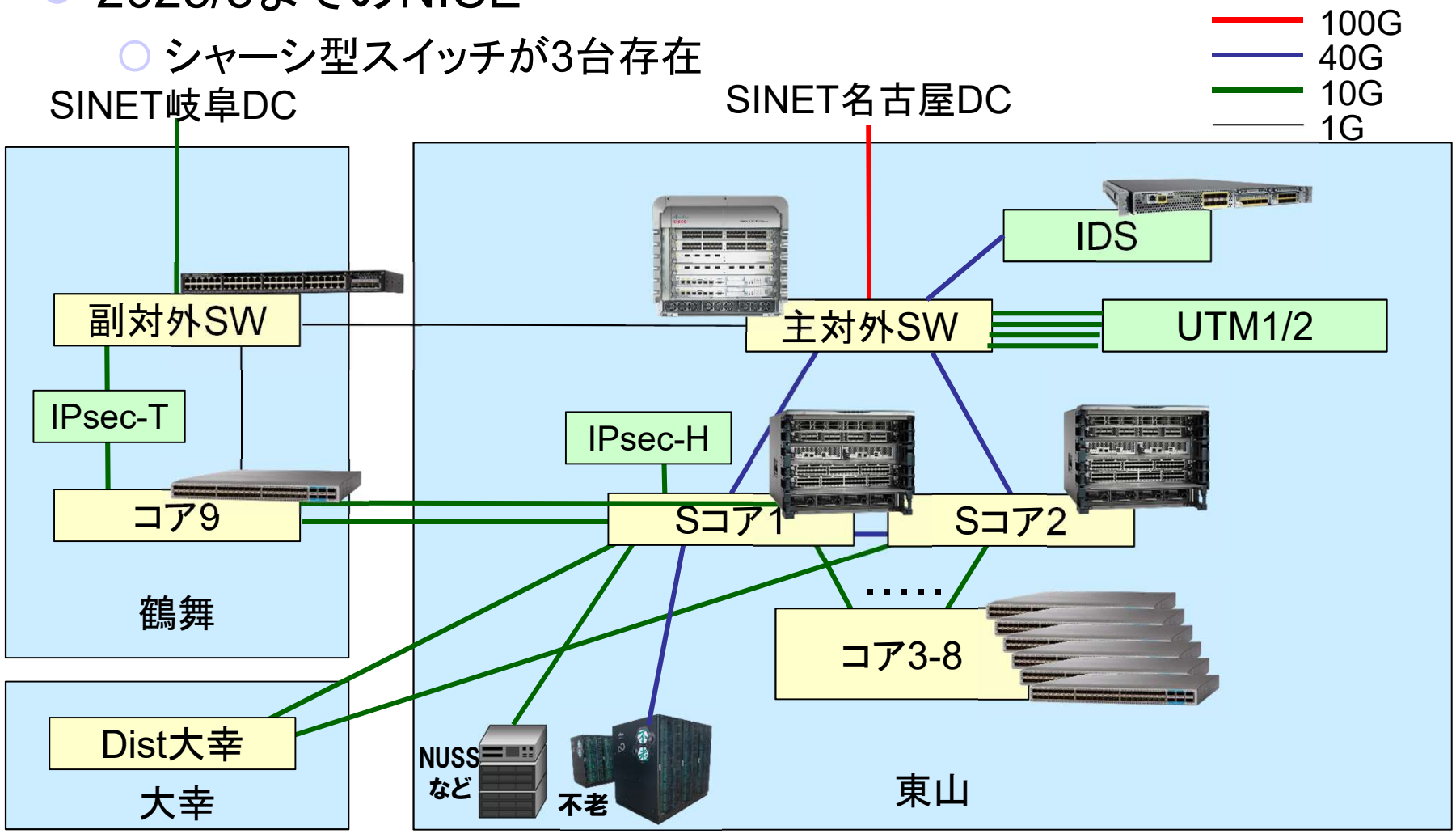
スイッチ調達時に主に気にする性能(2/2)

- 監視用の設定を色々とできるか?
  - フローモニタに出力できるか?
  - ミラーポートの切り方に自由度があるか?
  - アクセス・コントロール・リストをどれだけ設定できるか?
- メンテナンス性や冗長性は?
  - 複数のスイッチを1つとして動作させる機能は?
  - ソフトウェアアップデート時の停止時間は短いか?
- 実効的に使えるポート数は?

- なお、有名メーカでも安物モデルだと、「カタログ性能は何の設定していない全通信素通し時」みたいなことも…
  - 今どき必要なセキュリティとか監視設定とかVLAN設定すると、負荷がかかった時にあっぷあっぷする有名メーカの安物モデルを見た

NICEにおけるシャーシ型スイッチの廃止(1/3)

- 2025/3までのNICE
  - シャーシ型スイッチが3台存在



NICEにおけるシャーシ型スイッチの廃止(2/3)

将来を見据えた基幹増速に加えて検討したこと

- シャーシ型SWとボックス型SWクラスタではどちらが得か
  - 将来的な保守も考えてボックス型クラスタへ
  - 2台のボックス型SW間でeBGP/iBGPで経路冗長設定と面倒くさい所はある
- スーパコアSW直収部がスーパコアSWメンテ時に止まる問題の解消
  - スーパコアSWはL3を担当するのでファーム更新が低頻度だが発生(逆にL2のみ利用のSWでは稀)
  - スーパコアSWのボックス化でポート数が減るのでポート数確保も兼ねて、スーパコアSWに併設する新規コアSWを追加
- データサイエンスで100G等で接続したいが、基幹スイッチ接続に基幹スイッチ純正光モジュール必須で高価を何とか
  - 非常に安価な3rd party(保証は無い)光モジュールOKな安価100G SW(本体価格も基幹グレードSWの1/20とか)を追加
- 学内NW監視を強化
  - 10Gに制限した上で学内通信をIDSに入力する構成

NICEにおけるシャーシ型スイッチの廃止(3/3)

- 2025/3の更新で全てボックス型スイッチとなる
  - 対外400G ready、学内100G化、スーパコア直収廃止