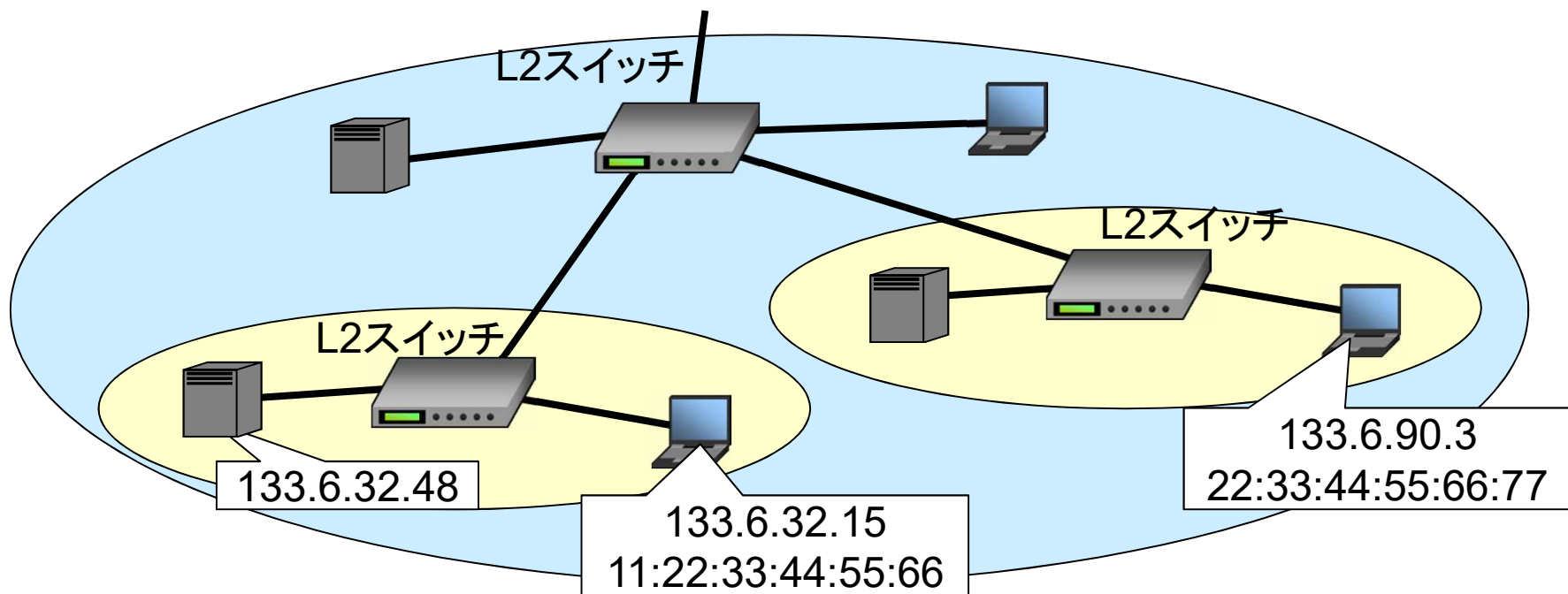


ネットワークスイッチの構成と動作

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

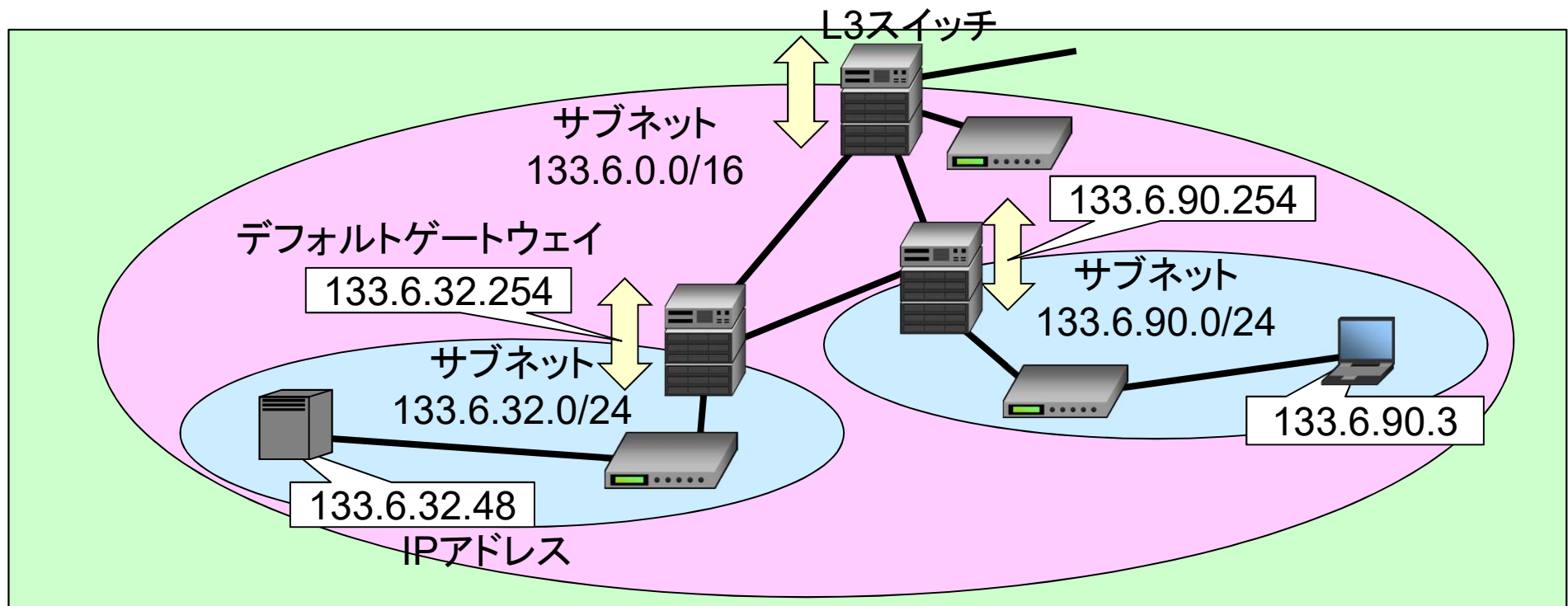
L2における通信の復習

- 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



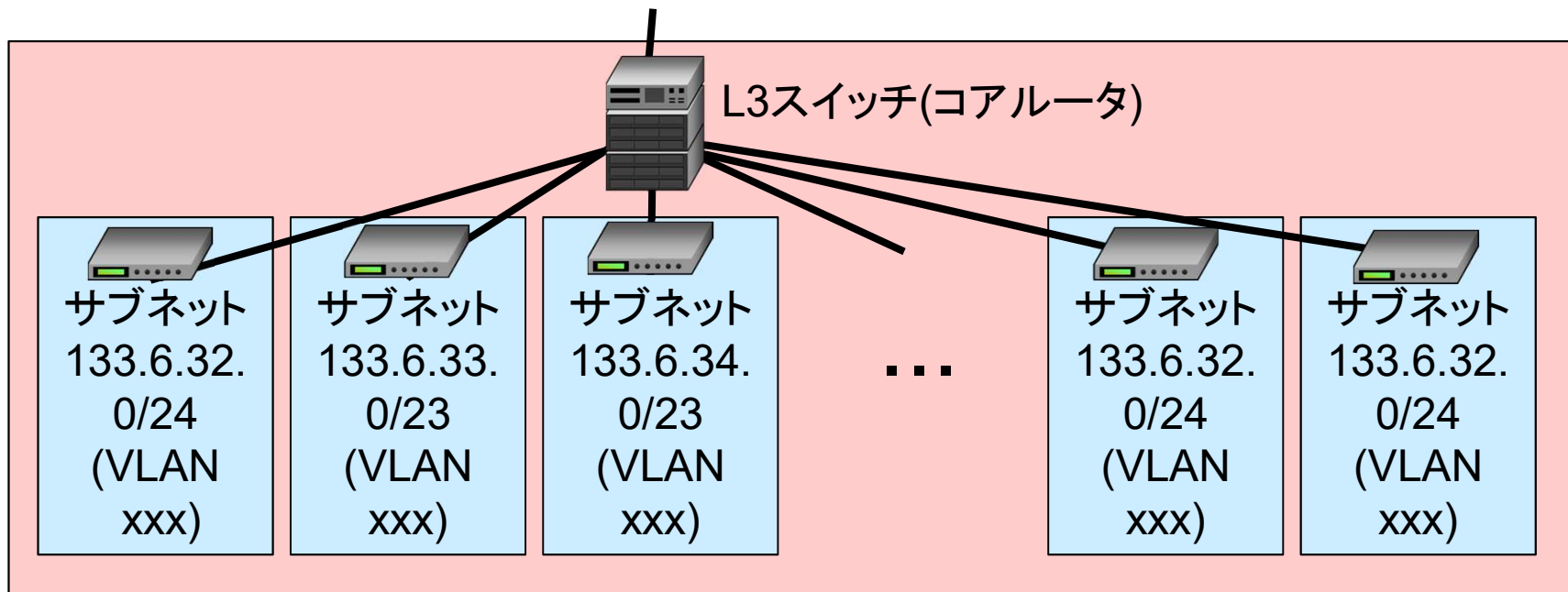
L3における通信の復習

- 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

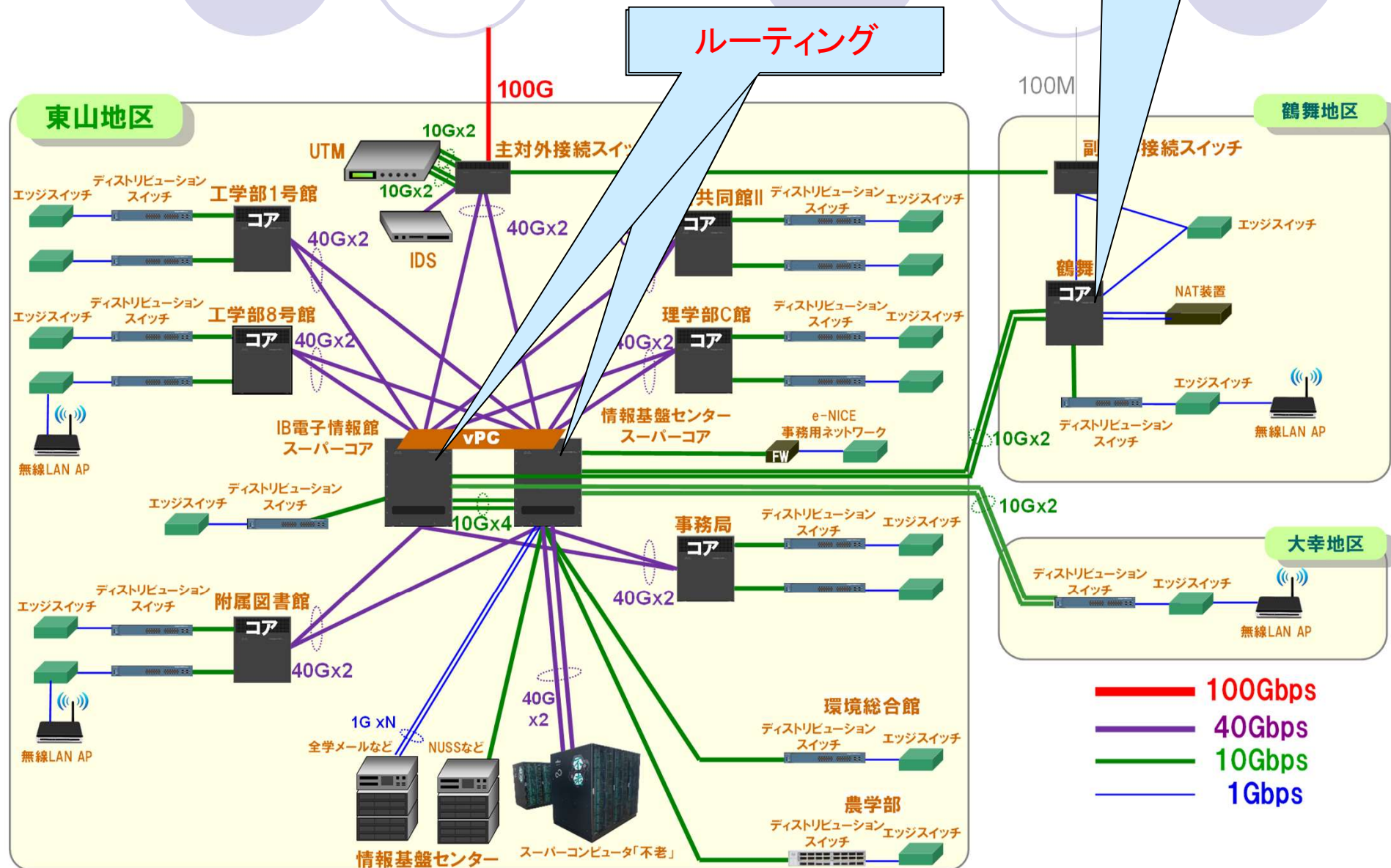


現実的なL3による通信

- 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

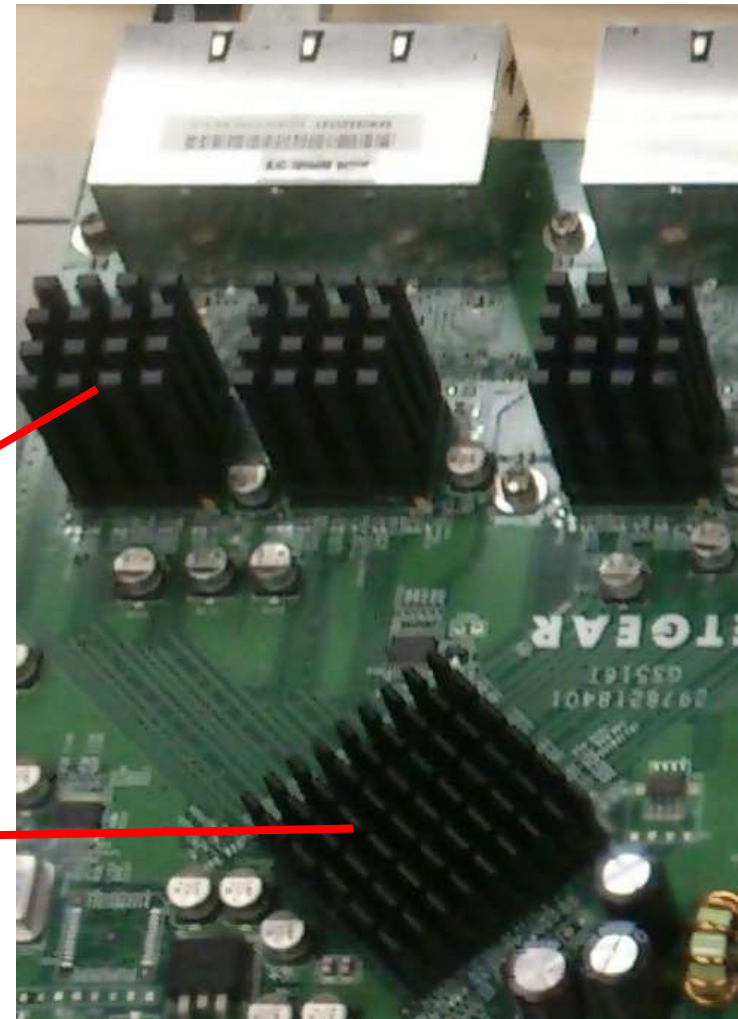
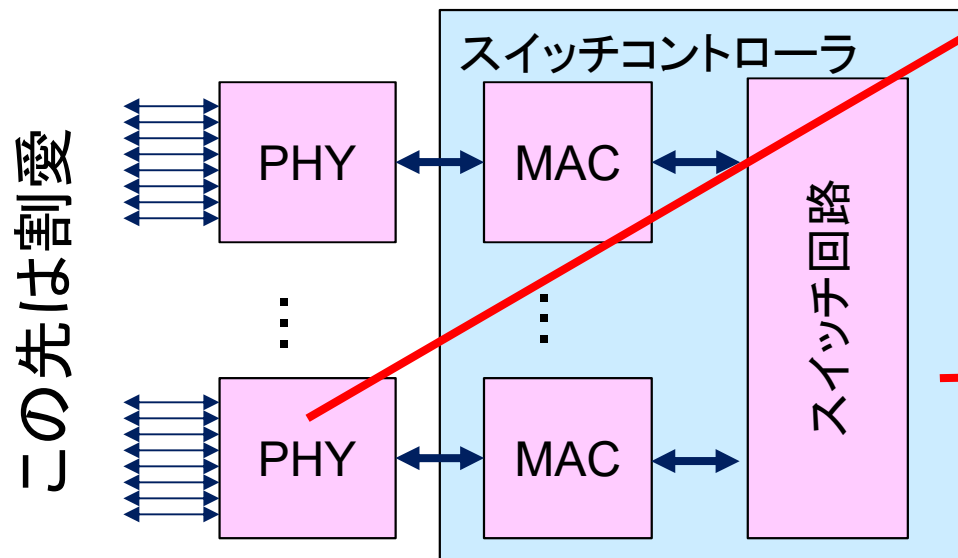


NICEにおけるルーティング



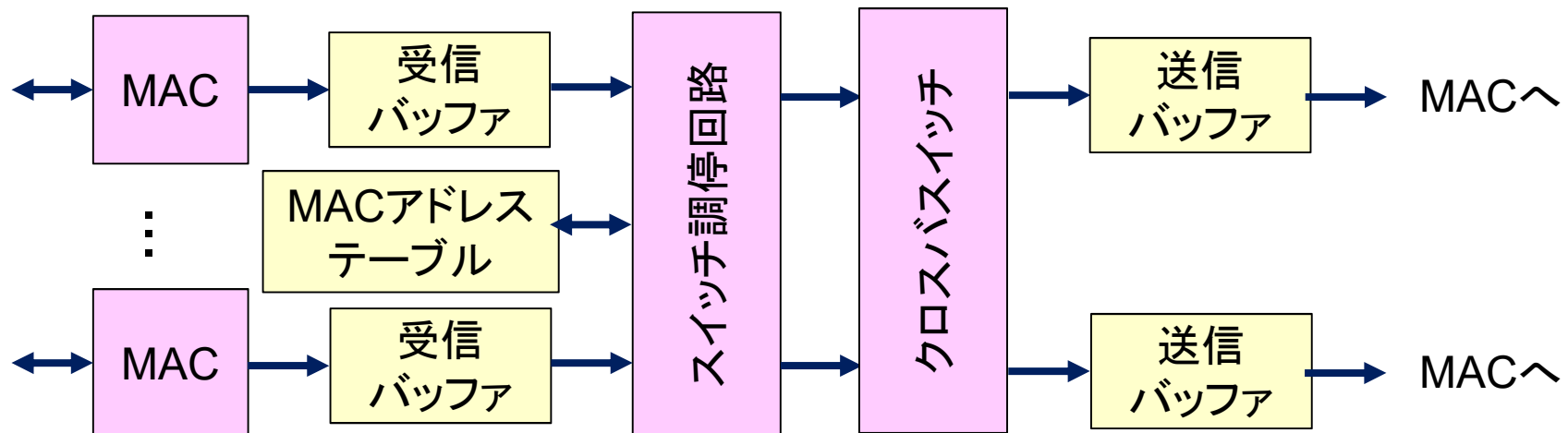
低価格L2スイッチの構成

- 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)

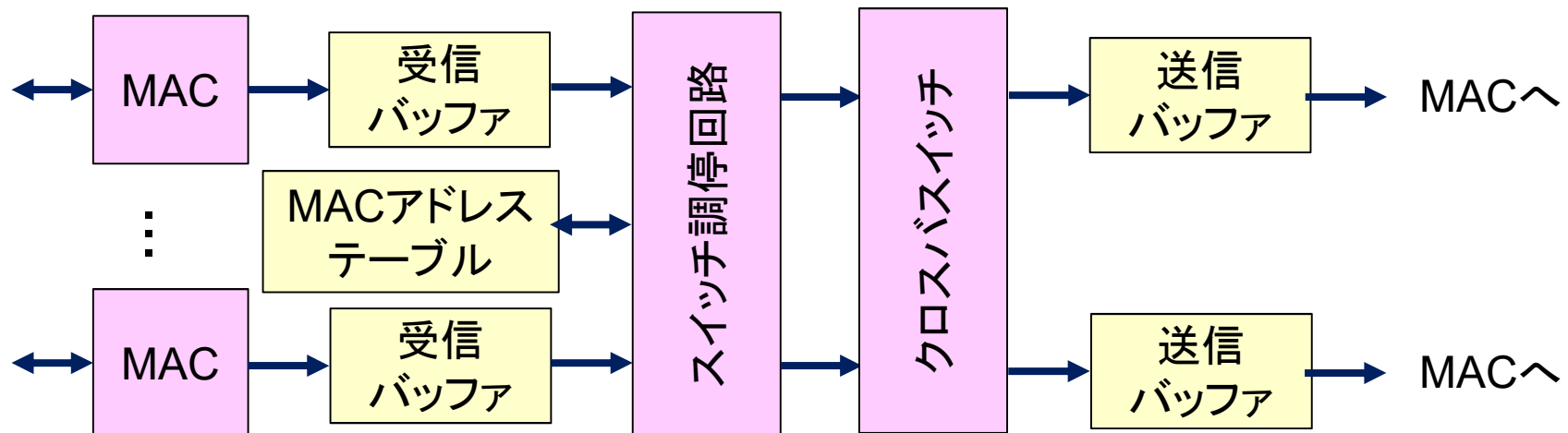
1. MACがフレームを受信バッファに格納
2. 受信バッファ中のフレームのMACアドレスでMACアドレステーブルを検索
 - MACアドレステーブルは、どこのポートからどこのMACアドレスの通信がきたかどうかを保存
 - 一致があれば、そのポートのみに送信
 - 一致が無ければ、全ポートへ送信(ブロードキャスト)



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スイッチの調停ができたなら送信

- 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ったら、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、独自の物も存在
 - 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

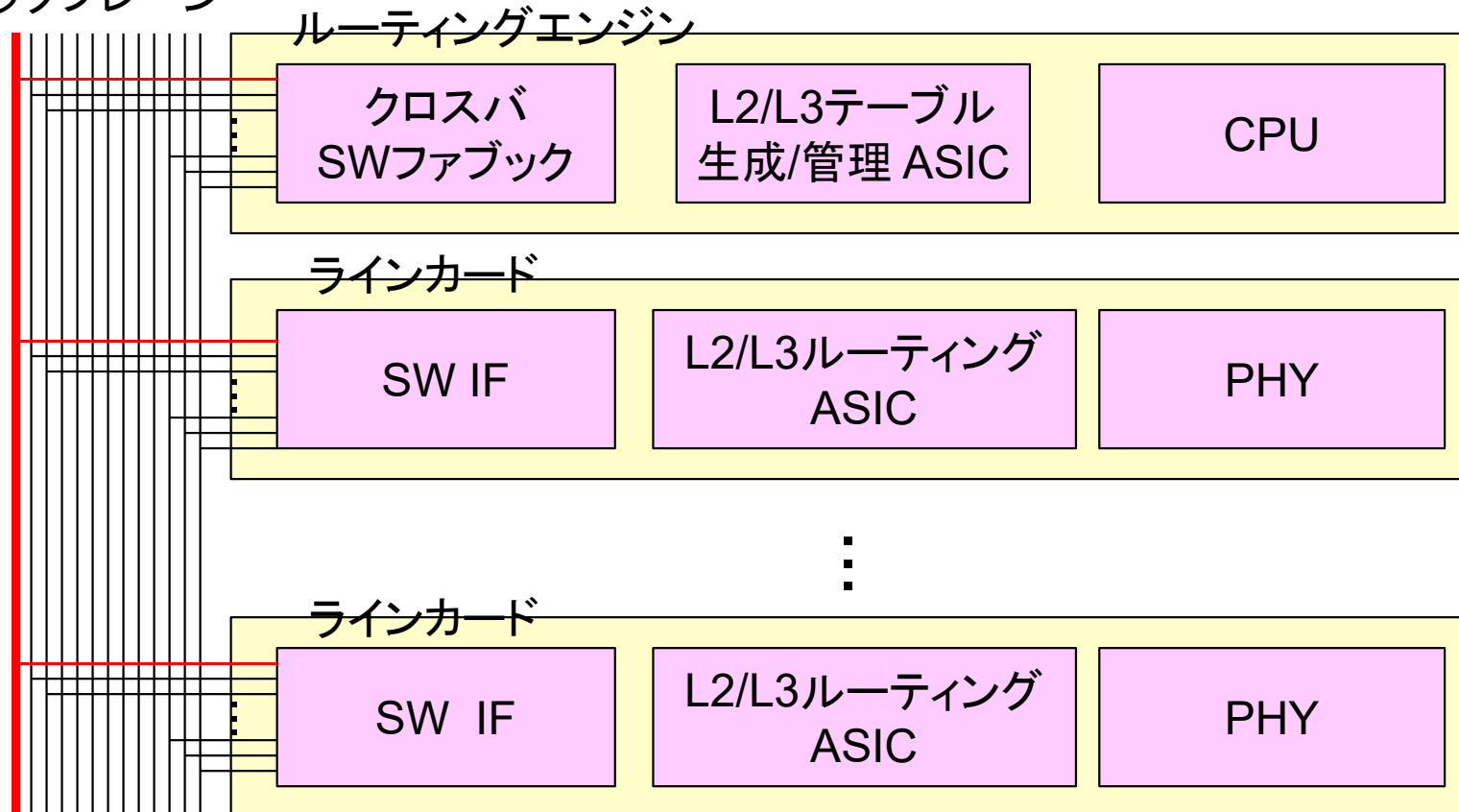
一般的な大型スイッチの構成

- 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

バックプレーン



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

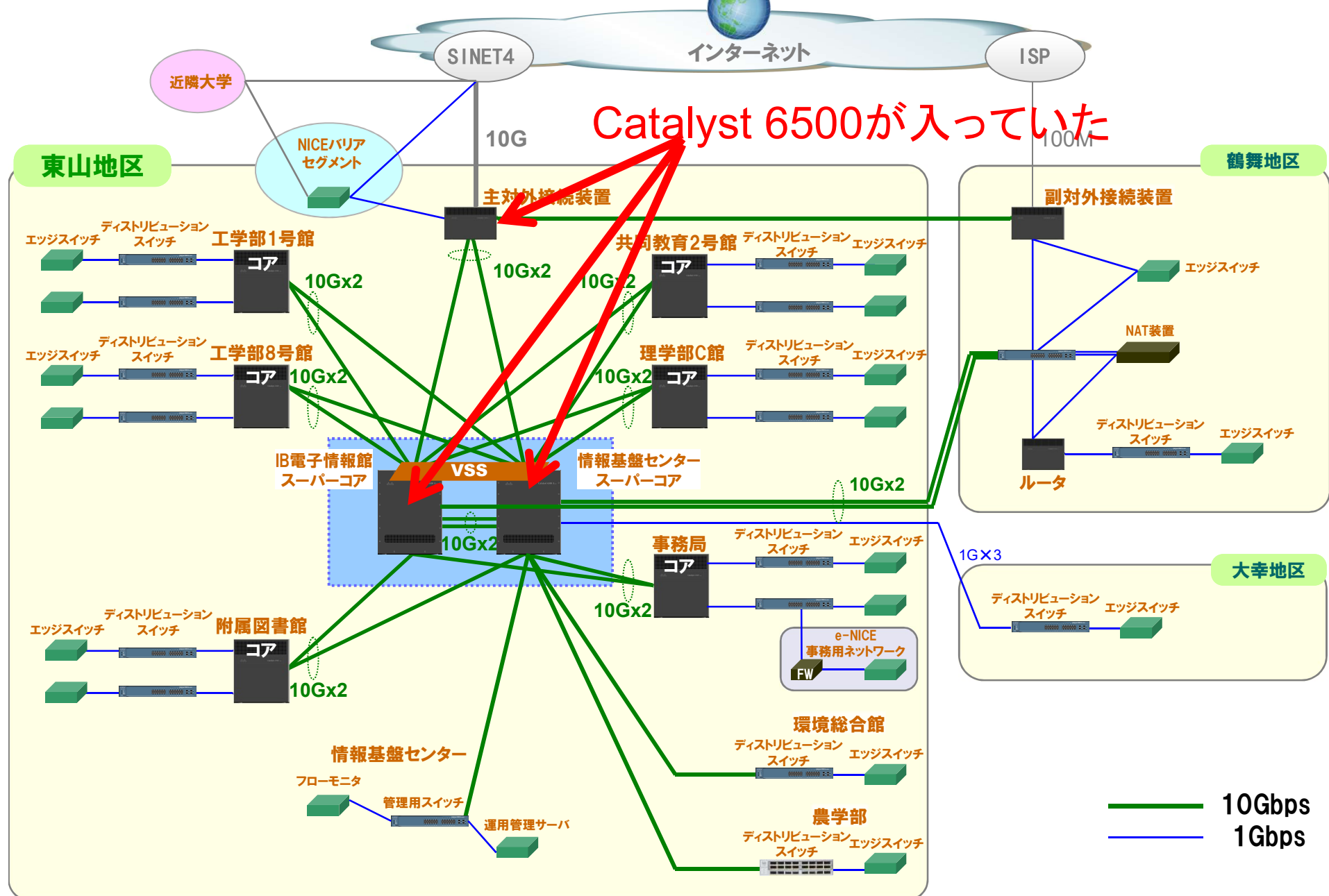
大型ネットワークスイッチの実例 (Cisco Catalyst 6500)

名大内有線ネットワークのNICE4で利用していた

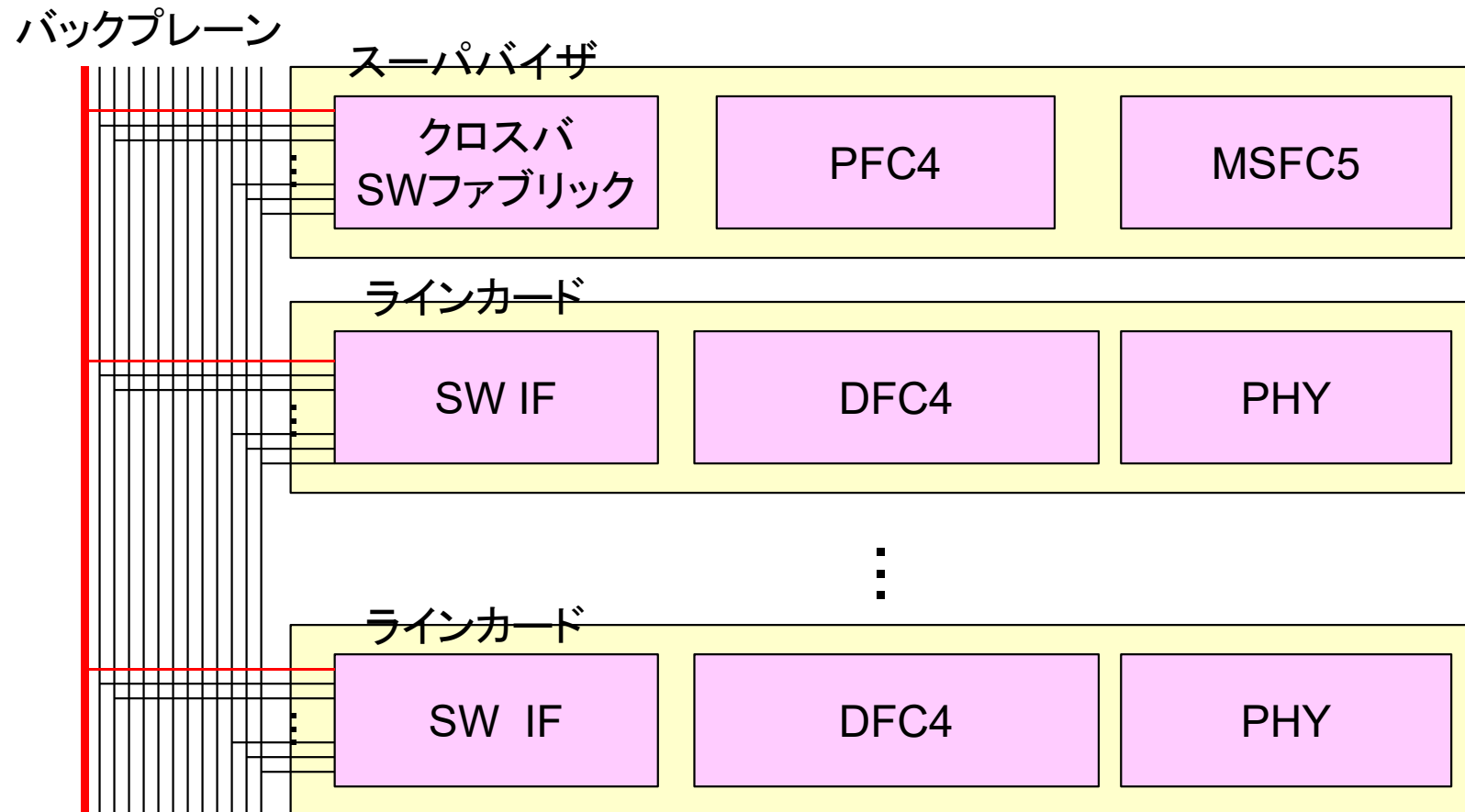
- バックプレーン容量2Tbps
 - ルーティングエンジン(スーパーバイザカード)も2Tbps対応
- 1G/10G/40Gイーサネット対応ラインカードを複数接続可能
- Virtual Switching Systemで複数のスイッチを束ねて制御可能



NICE4(-2015/3)とCatalyst 6500



Catalyst 6500の構成

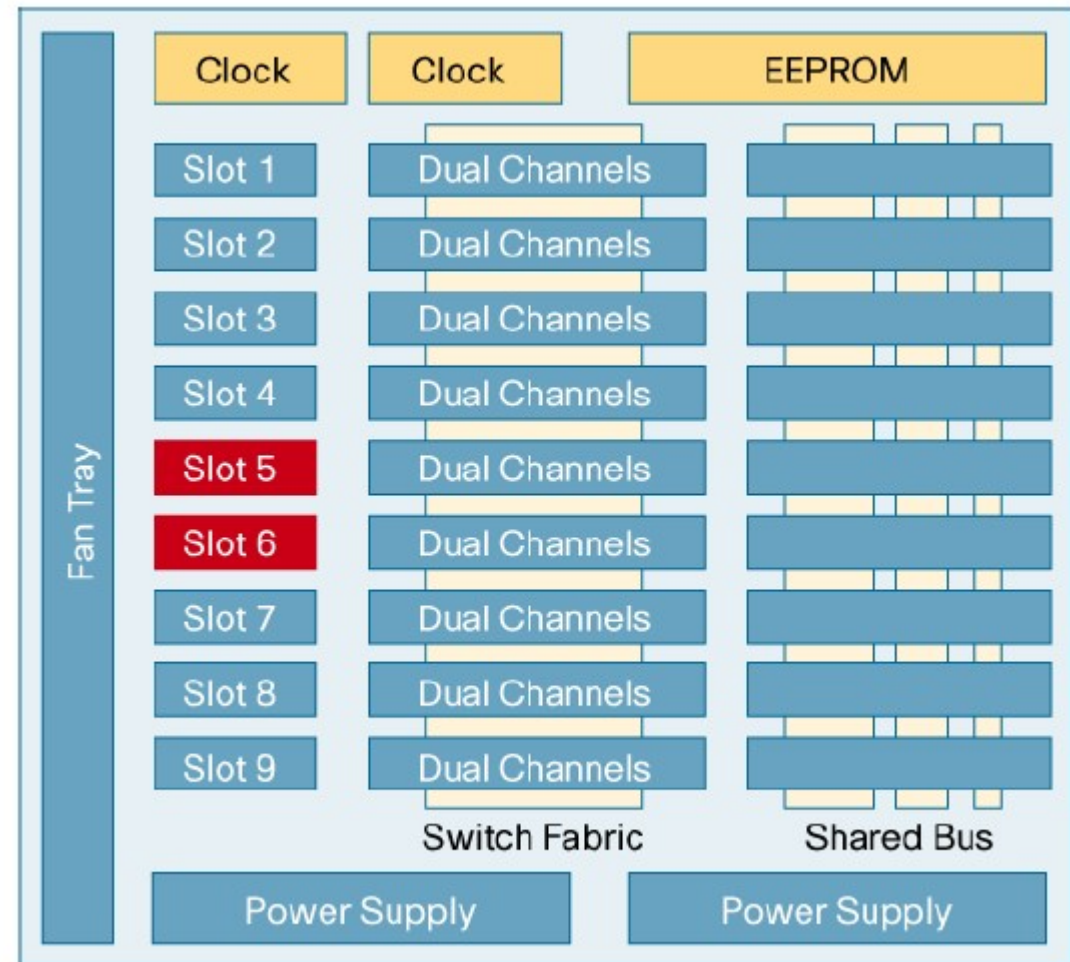


シャーシについているバックプレーン

- 共有バス接続部とクロスバススイッチ接続部に分かれています

以下のCatalyst 6500の図は、全て[1]からの引用

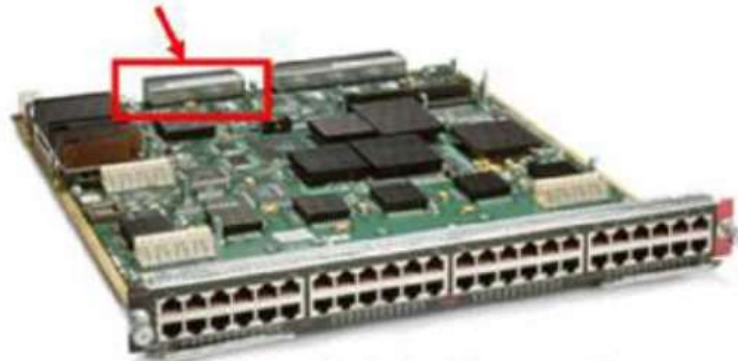
[1] Cisco, "Catalyst 6500 Architecture White Paper," Nov. 2010.



ラインカードによる接続形態の違い

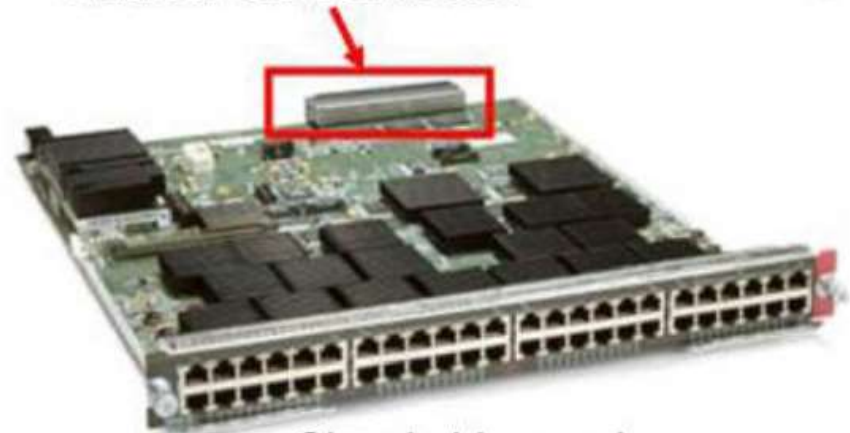
- ラインカードによっては、共有バスのみの接続を取るものがある
 - ルーティングの依頼ができなくなるので、その逆はない

Crossbar Connector



Fabric Linecard

Shared Bus Connector



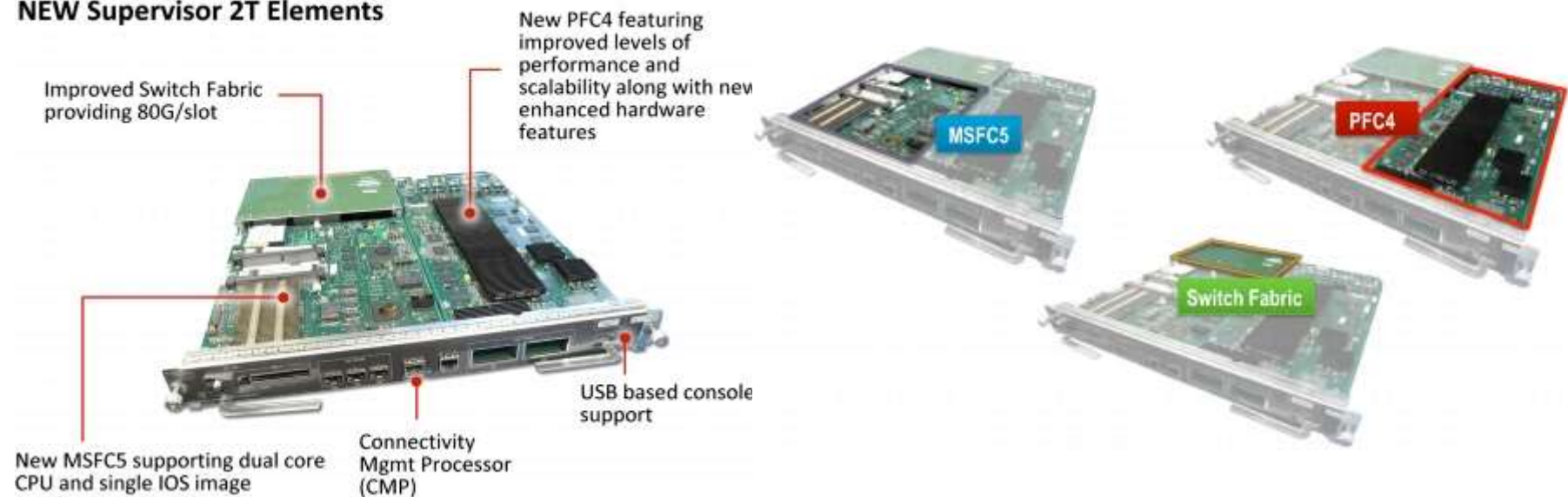
Classic Linecard

スーパーバイザカード2T

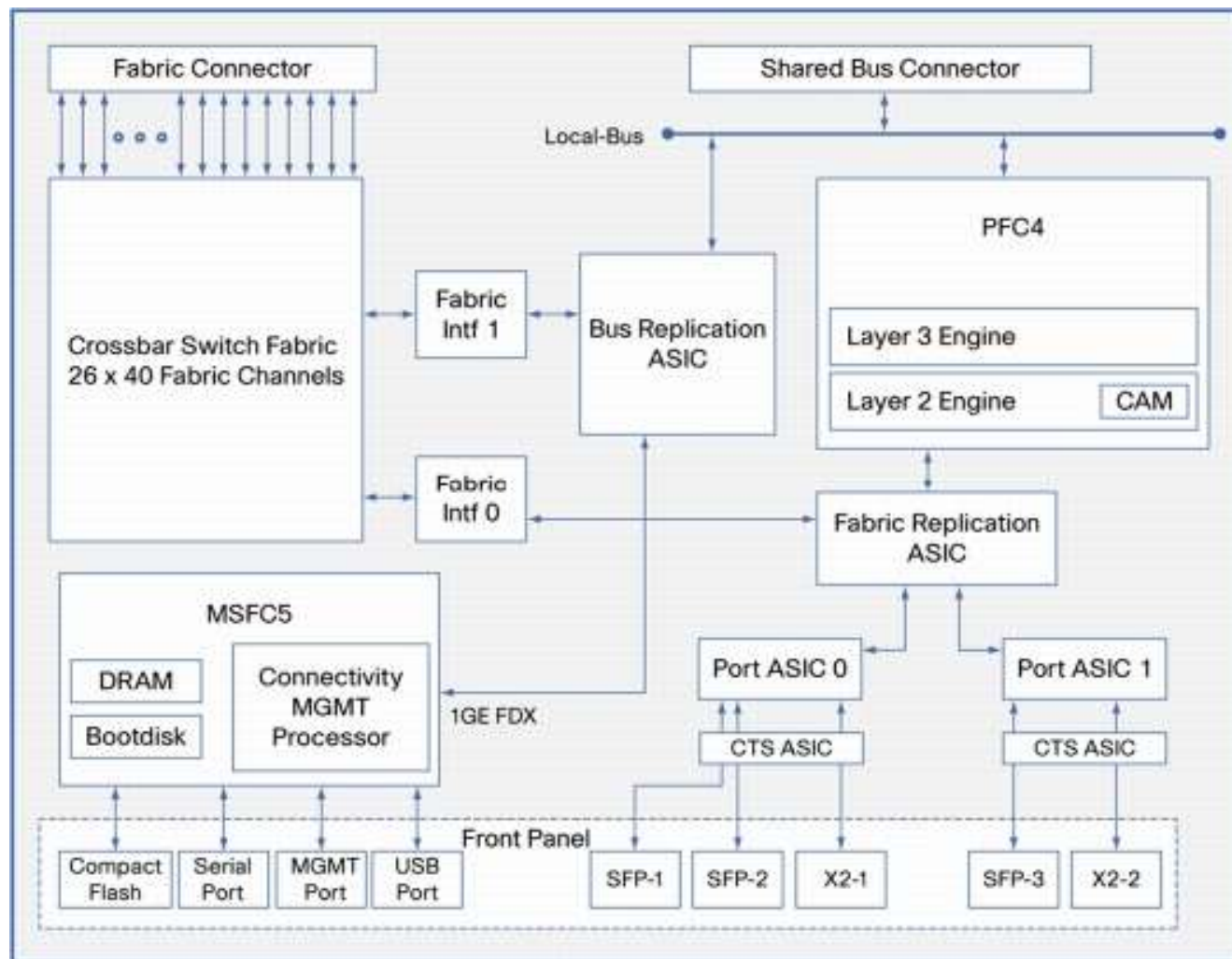
以下より構成

- MSFC5: Multilayer Switch Feature Card 5
- PFC4: Policy Feature Card 4
- スイッチファブリック

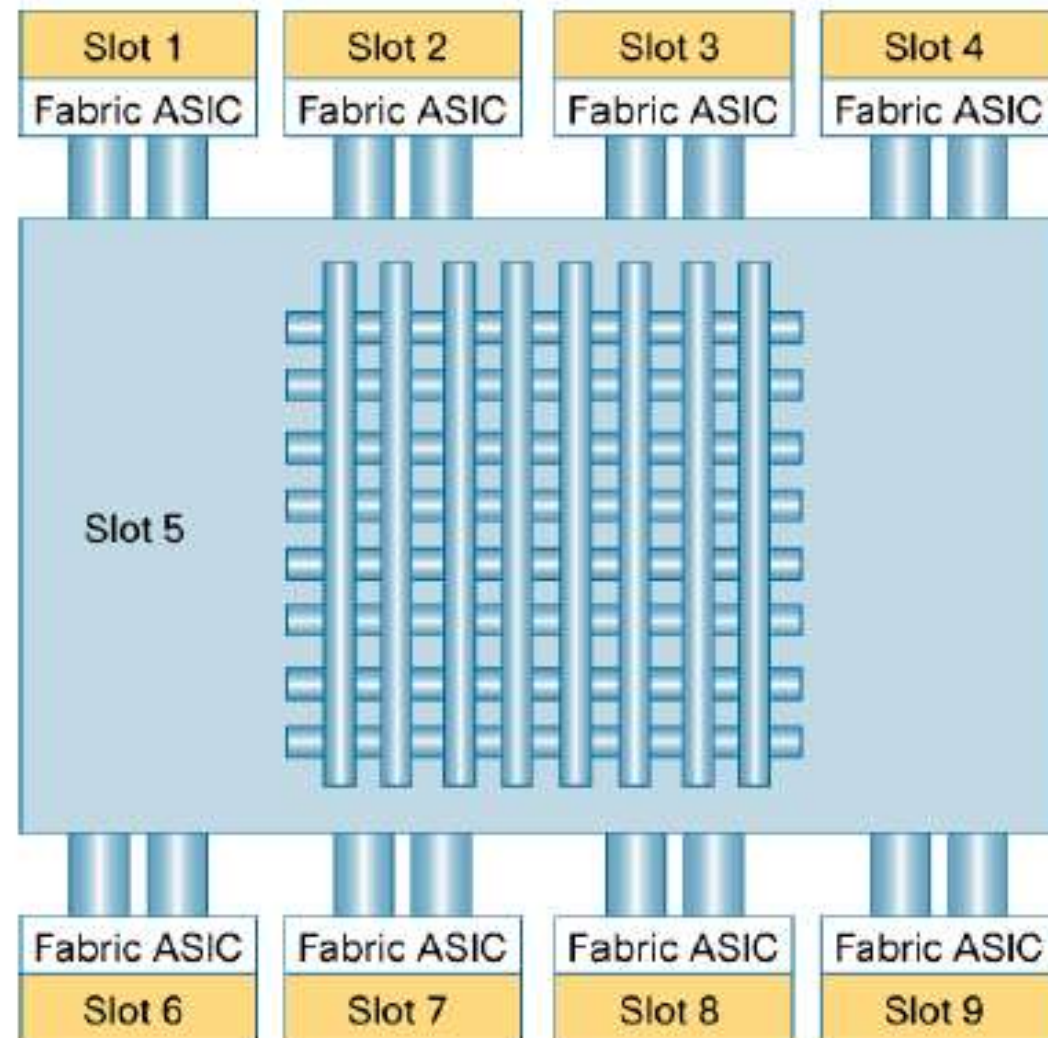
NEW Supervisor 2T Elements



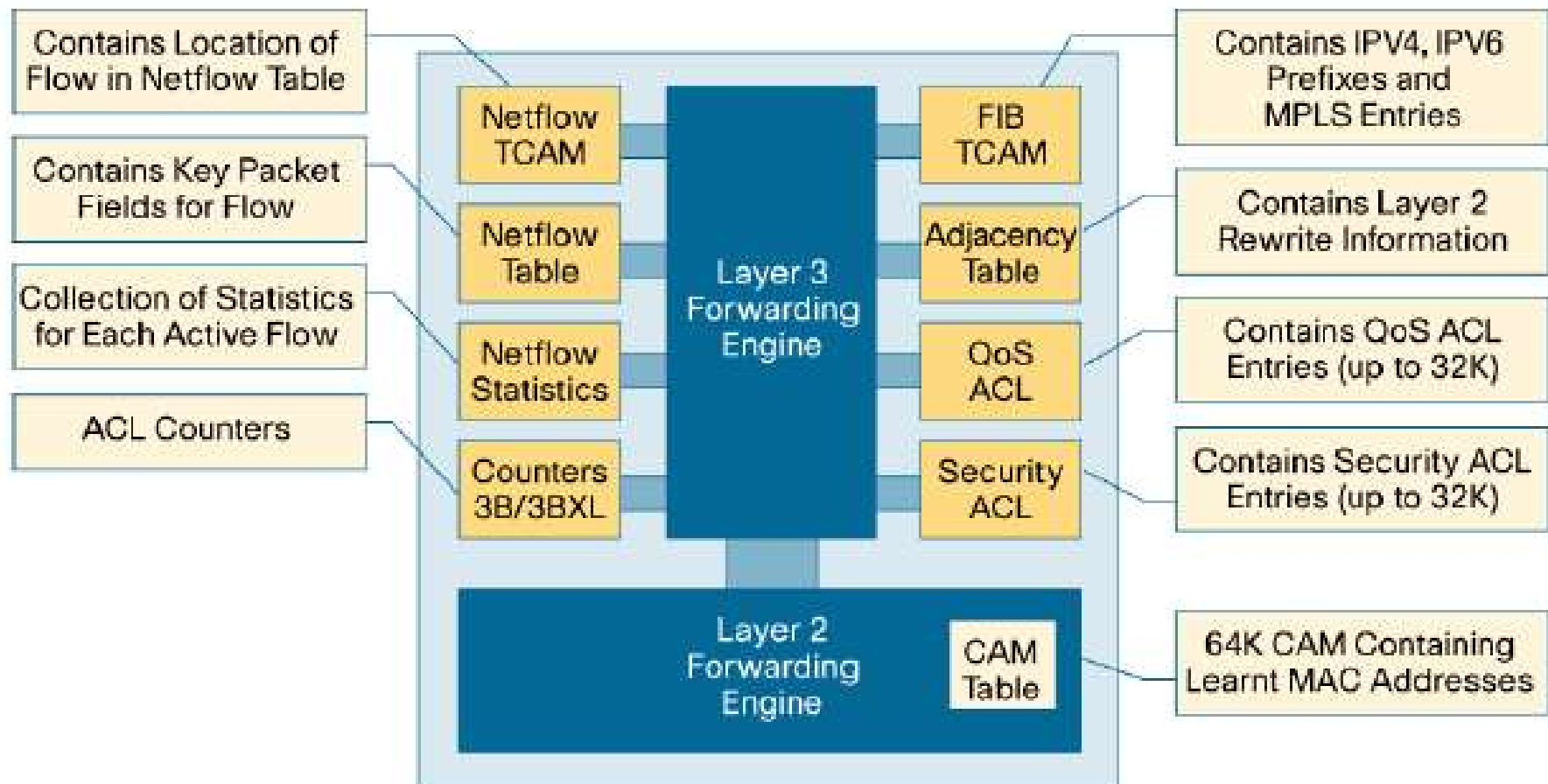
スーパーバイザカードのブロック図



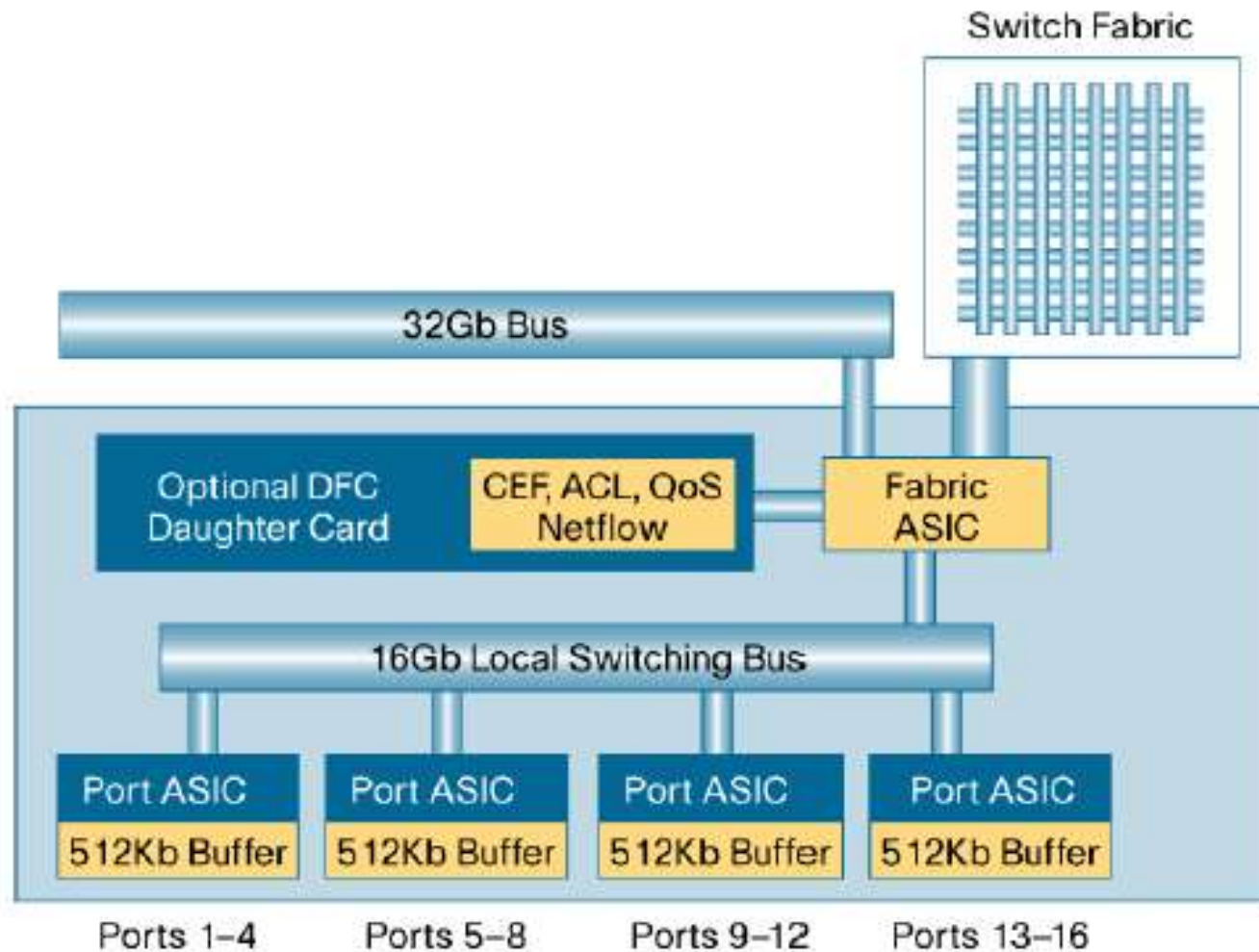
スイッチファブリック部



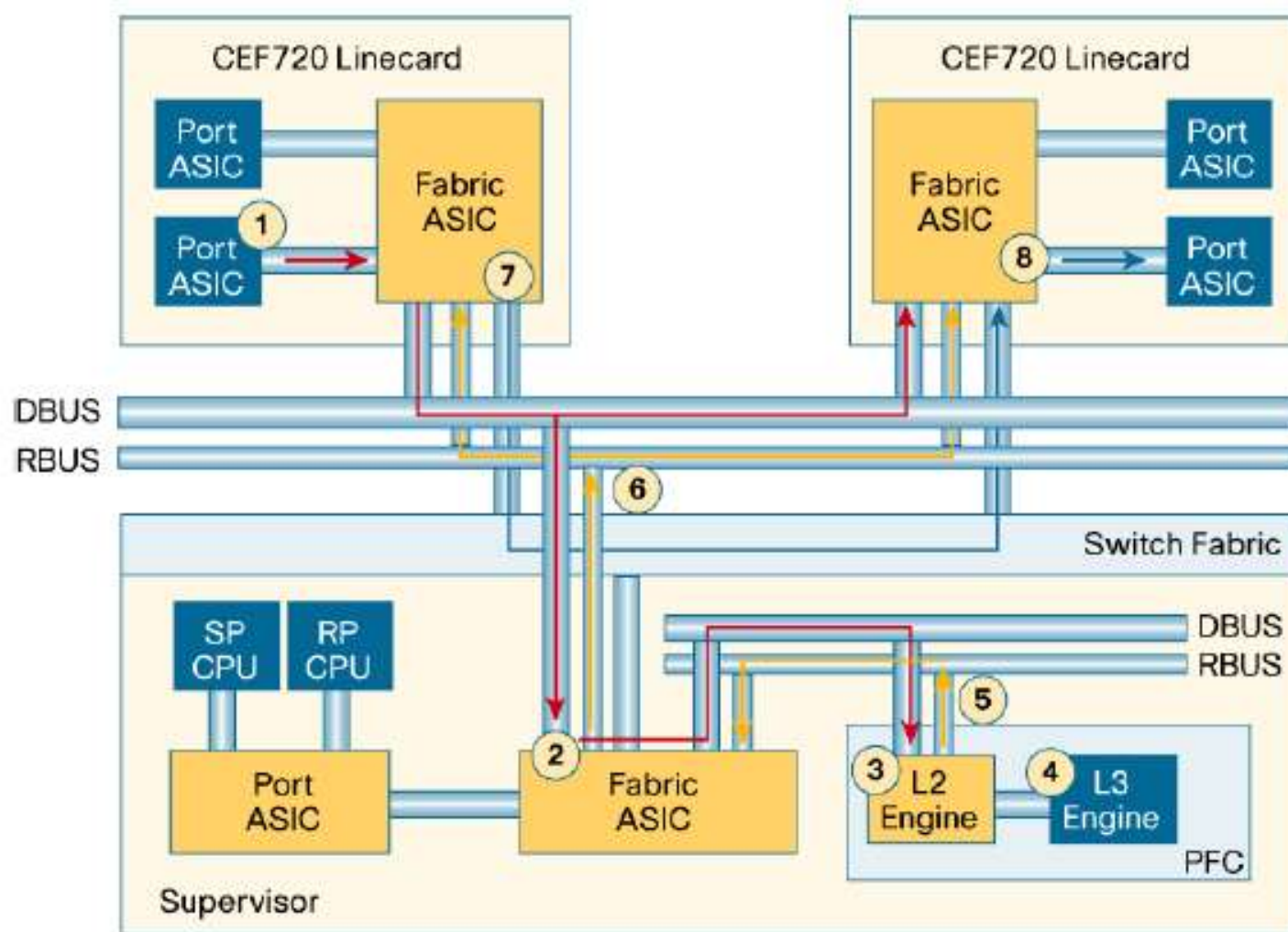
PFC4部



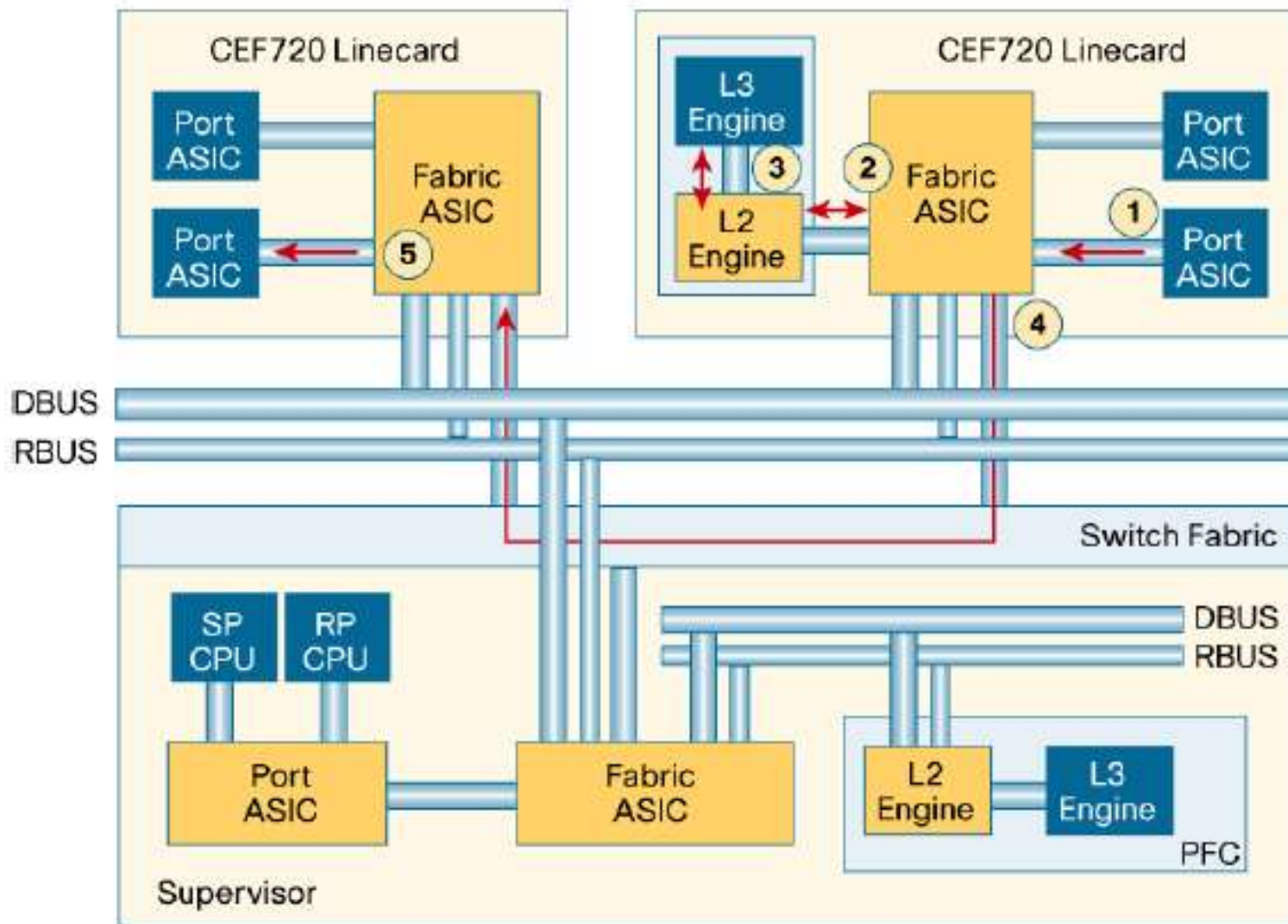
クロスバススイッチ部を 1チャンネル使うラインカード



ルーティング情報が無い場合のスイッチング



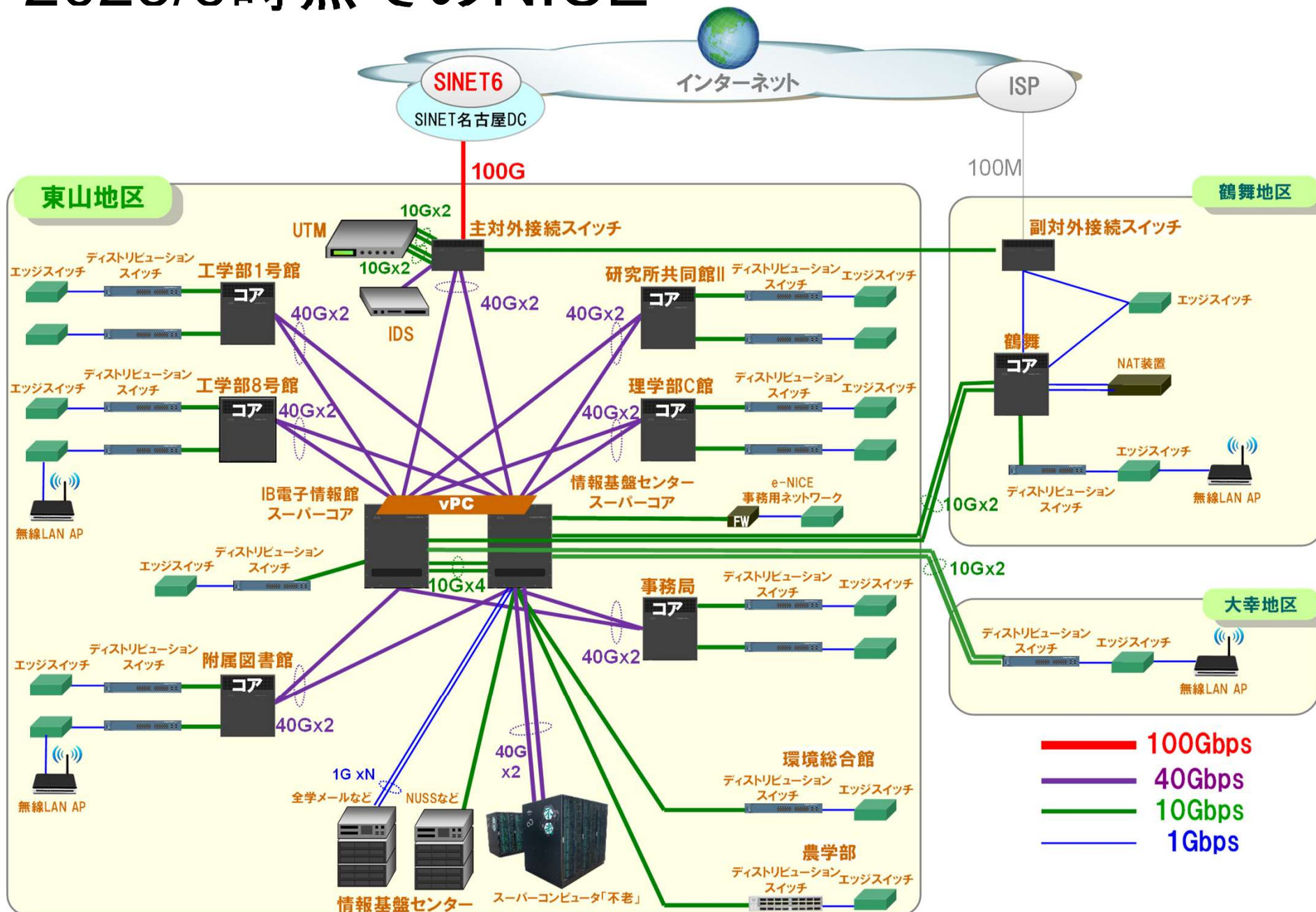
ルーティング情報がある場合のスイッチング



キャンパスLANの組み方の事例

- NICE(Nagoya university Integrated Communication Environment)を参考に
 - <http://www.icts.nagoya-u.ac.jp/ja/services/nice/>
- 現在は5世代目(NICE5)で、6世代目を検討中
 - 一括で更新したかったのだが、お金が無いので2016-7の2年で更新
 - SINET(Science Information NETwork: 学術情報ネットワーク)の方も2016/4にSINET5に更新された(2022/3にSINET6へ)
 - <http://www.sinet.ad.jp/>
 - 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 세미나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
 - 次に有線やるならば、エッジスイッチのmGig化かな
 - さすがに全部やると投資効率悪いので、大量データを使う所に
 - ...と思っていたら、世の中、情報コンセントにmGigの需要が無いらしい

2023/6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

- 主対外接続は情報基盤センターからSINETへ100Gbpsで
- インターネット上の1ASとして動けるようBGPフルルートを利用可
- 副対外接続は鶴舞から某ISP経由で100Mbpsで(ほぼ非常用)
 - こちらもSINET別DCに10Gbps接続化する調達(入札)をこの春に実施
- 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

スイッチの階層の組み方(2/2)

- 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
 - 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
- エッジスイッチ
 - 基本的に、各建物の各フロアに1台
 - 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、それでも最近は認証が遅いことがあるので、増強を計画中
 - Web認証サーバ情報基盤センターに

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

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
 - メーカー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メーカーが「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
 $10\text{G} \times 32 \times 2(\text{双方向}) + 40\text{G} \times 4 \times 2(\text{双方向}) = 960\text{Gbps}$
 - 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)で通信しても耐えら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?

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？
- ラインカードあたりのポート数は？
- なお、有名メーカー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ーの安物モデルを見た