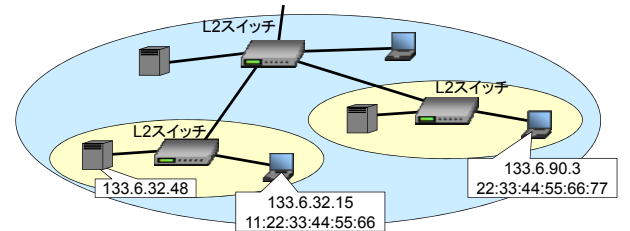


# 情報ネットワーク特論 ネットワークスイッチの構成と動作

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

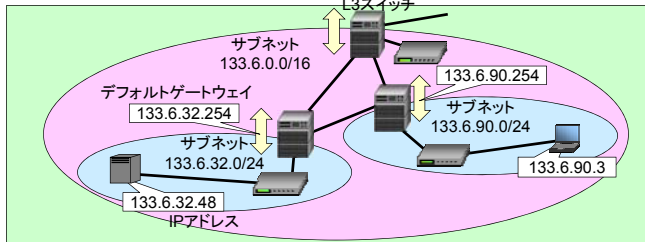
## L2における通信

- 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



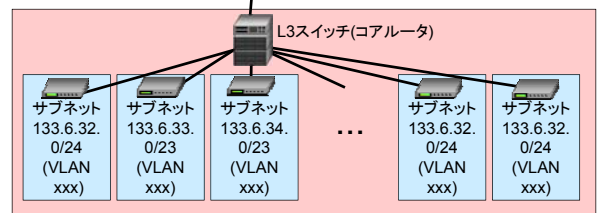
## L3における通信

- 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

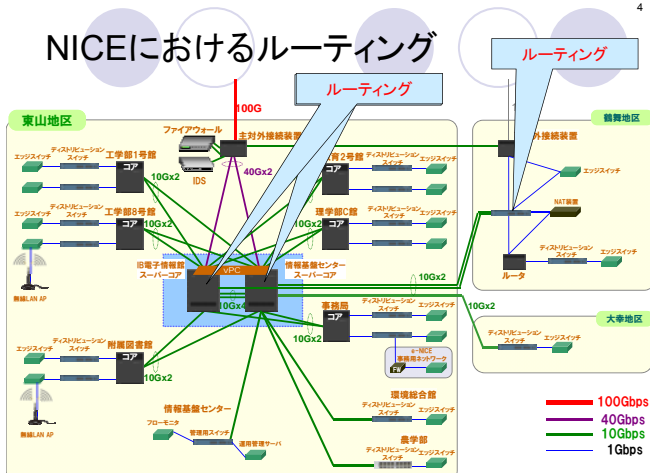


## 現実的なL3による通信

- 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

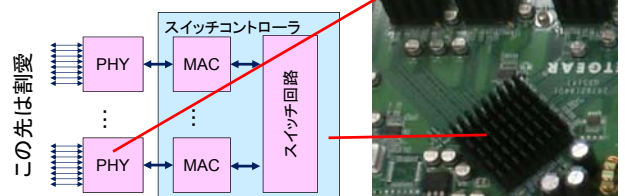


## NICEにおけるルーティング



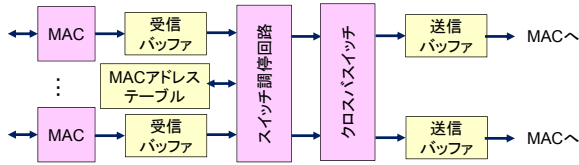
## 低価格L2スイッチの構成

- 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: スイッチングハブの構成です



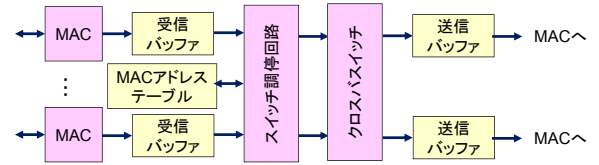
## スイッチコントローラの動作(1/2)

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



## 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
  - アップリンクポート側はそのうち溢れます
  - 古いものから削除
- 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
- 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

## 高機能なL2スイッチ

- スイッチの上でOSが入っていて色々設定できる
- 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
  - スイッチの各ポートに対して認証をかける
  - Power over Ethernetによる電力供給

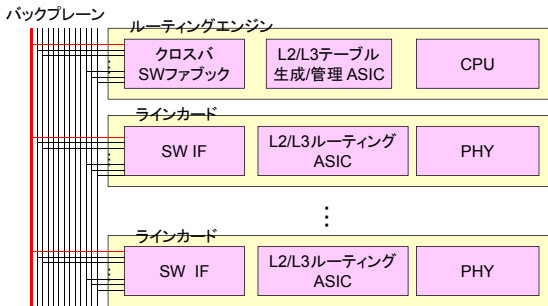


Cisco Catalyst 2960X

## 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
  - 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



# 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)

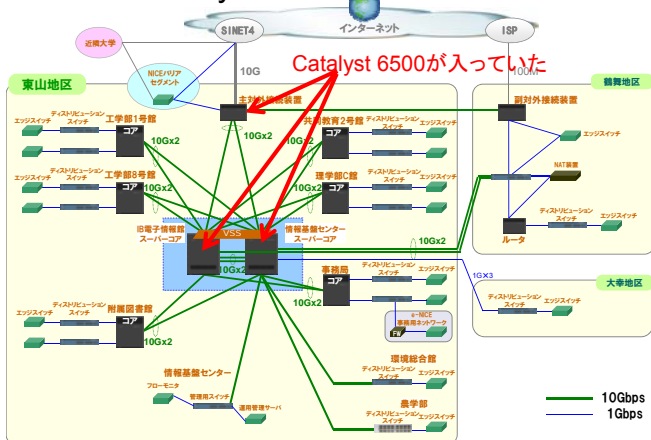
- ラインカード
  - 物理層から受け取ったフレームをバッファリング
  - テーブルにある宛先の送信制御
    - クロスバスイッチファブリックに調停要求→送信
  - テーブルにない宛先を持つフレームの送信をルーティングエンジンのCPUに依頼
  - マルチキャストフレームの複製

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

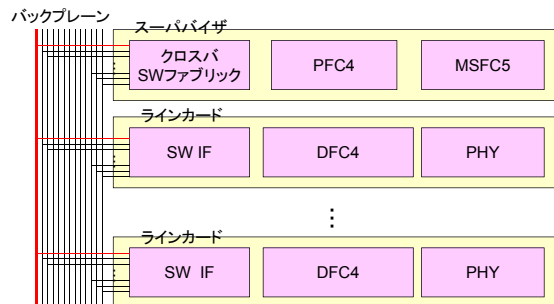
- 名大内ネットワークNICE4で利用中なので
- バックプレーン容量2Tbps
  - ルーティングエンジン(スーパーバイザカード)も2Tbps対応
- 1G/10G/40Gイーサネット対応ラインカードを複数接続可能
- Virtual Switching Systemで複数のスイッチを束ねて制御可能



# NICE4とCatalyst 6500

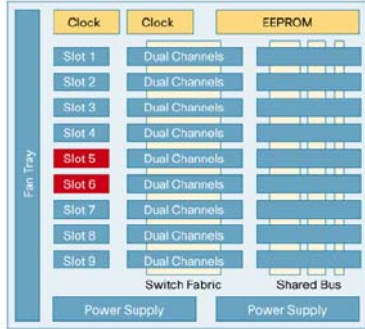


# Catalyst 6500の構成



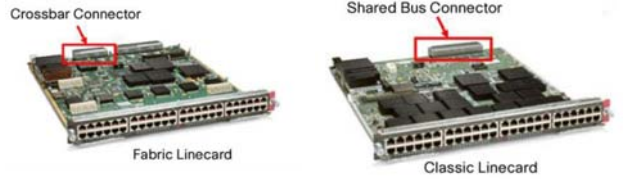
# シャーシについているバックプレーン

- 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る



# ラインカードによる接続形態の違い

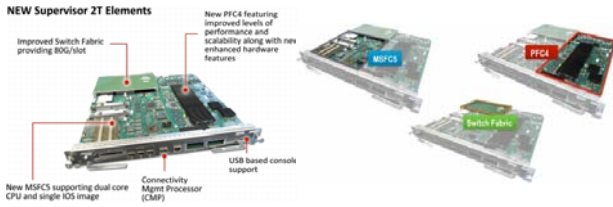
- ラインカードによっては、共有バスをみの接続を取るものがある
  - ルーティングの依頼ができなくなるので、その逆はない



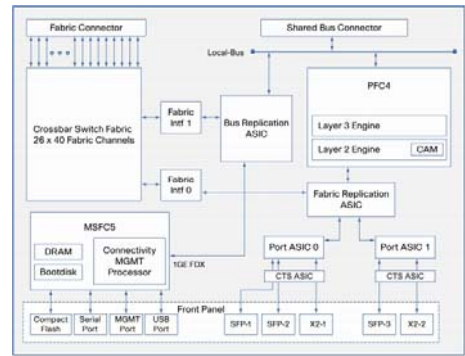
# スーパーバイザカード2T

以下より構成

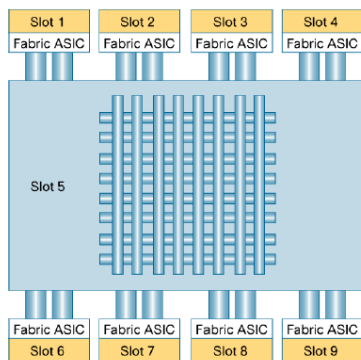
- MSFC5: Multilayer Switch Feature Card 5
- PFC4: Policy Feature Card 4
- スイッチファブリック



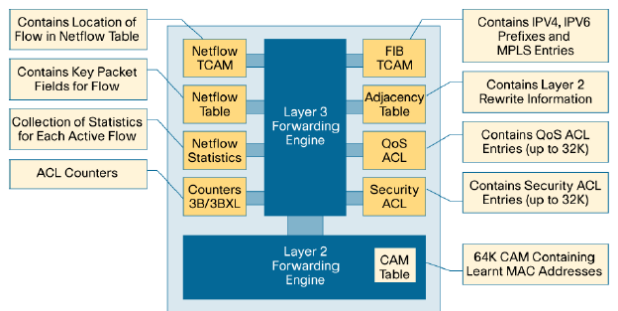
# スーパーバイザカードのブロック図



# スイッチファブリック部

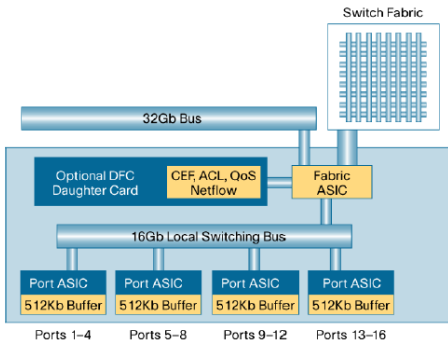


# PFC4部



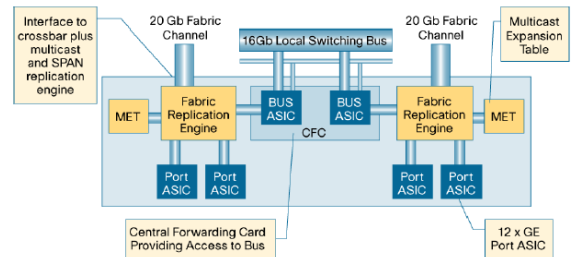
## クロスバススイッチ部を 1チャンネル使うラインカード

24



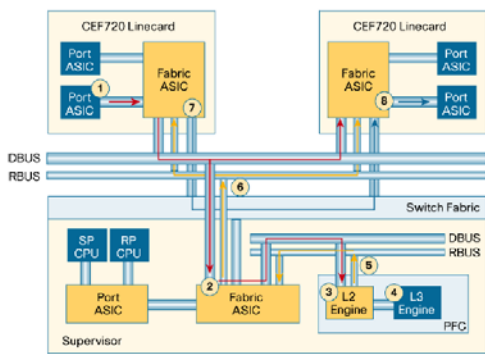
## クロスバススイッチ部を 2チャンネル使うラインカード

25



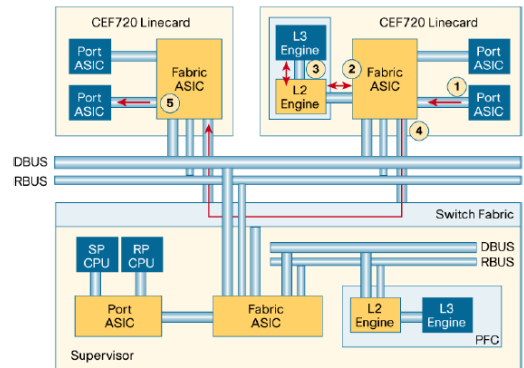
## ルーティング情報が無い場合のスイッチング

26



## ルーティング情報がある場合のスイッチング

27

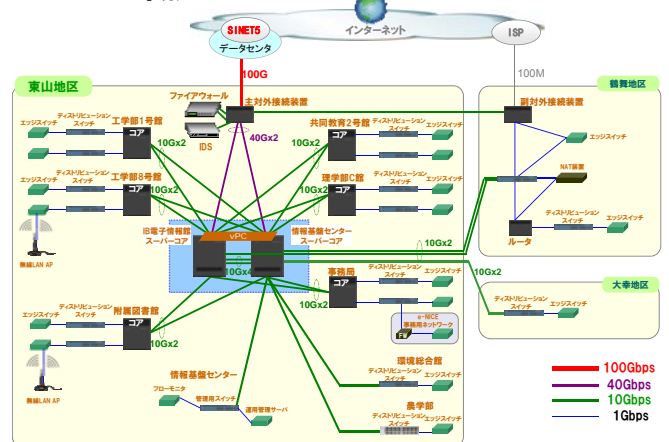


## キャンパスLANの組み方の実例

28

- NICE(Nagoya university Integrated Communication Environment)を参考に
  - <http://www.icts.nagoya-u.ac.jp/ja/services/nice/>
- 現在は4世代目(NICE4)から5世代目(NICE5)への更新途中
  - 本来は一括で更新したかったのだが、お金が無い
  - 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
    - <http://www.sinet.ad.jp/>

## 2016/4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
  - 副対外接続は鶴舞のエネルギーセンターから某プロバイダ経由で100Mbpsで(ただし、業務用)
  - この周りにIDS、ファイアウォール、アンチウイルスゲートウェイ、フローモニタ、など
- (スーパー)コアスイッチ
  - スーパー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
  - 認証系(ウェブ、802.1x)は情報基盤センターのサーバ室

## キャンパスネットワーク設計の検討点

- 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50万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100万経路は欲しい
    - 全ルートをTCAM(3値連想検索メモリ)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動的に経路選択をやらない限り不要
- 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ー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

## スイッチ調達時に主に気にする性能(1/)

- スイッチング容量(単位: bit per second)
  - 例: 10Gbps 32ポート、40Gbps 4ポート
  - $10G \times 32 \times 2(\text{双方向}) + 40G \times 4 \times 2(\text{双方向}) = 960Gbps$
  - 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/)

- 監視用の設定を色々できるか?
  - 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?



36



37