SINET5で変わる 長距離高速通信テクニック

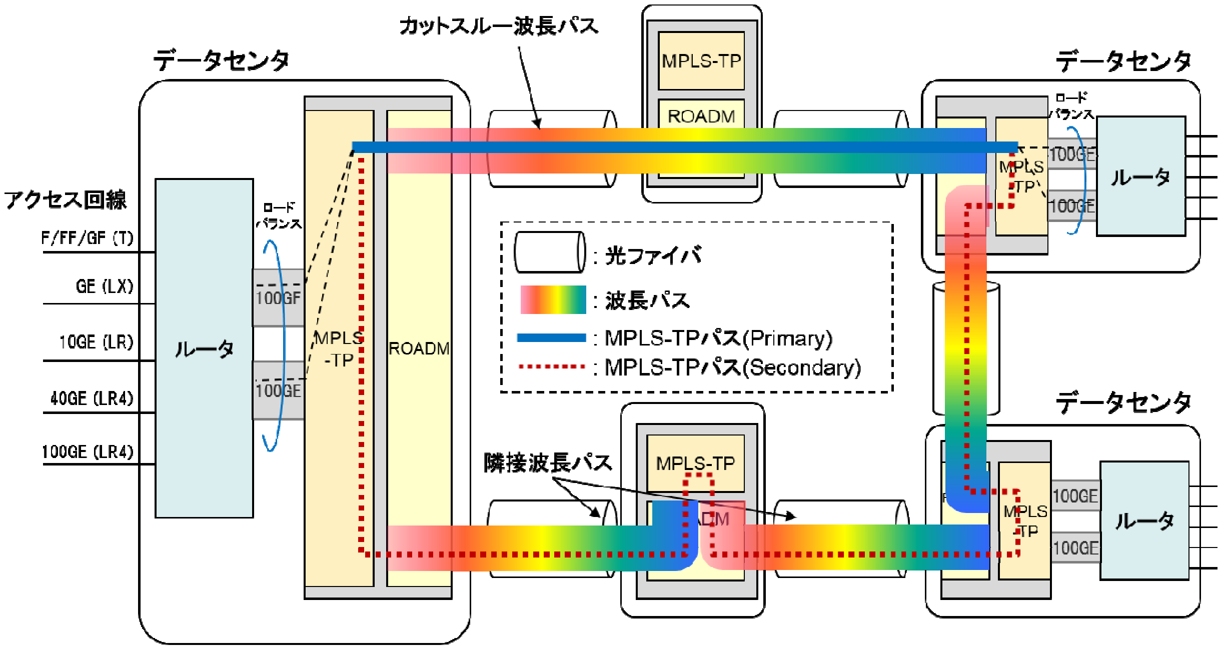
平成28年10月14日

国立情報学研究所/総合研究大学院大学
山中顕次郎, 阿部俊二, 漆谷重雄

はじめに

◆平成28年4月：新世代の学術情報ネットワークSINET5，運用開始
- SINET3，4の延長では無い，新ネットワークアーキテクチャを採用
- 全国100Gbps回線，フルメッシュ接続，…

◆本講演
- 長距離データ転送高速化の観点で，新アーキテクチャの効能を説明
- SINET5での長距離高速データ転送に，必要なテクニックを解説



SINET5の新ネットワークアーキテクチャ

目次

◆TCPで長距離高速通信を行うには：
- ・ノード間遅延(RTT)の最小化
- ・ウィンドウサイズ(Win)を大きく保つ → パケットロスの防止

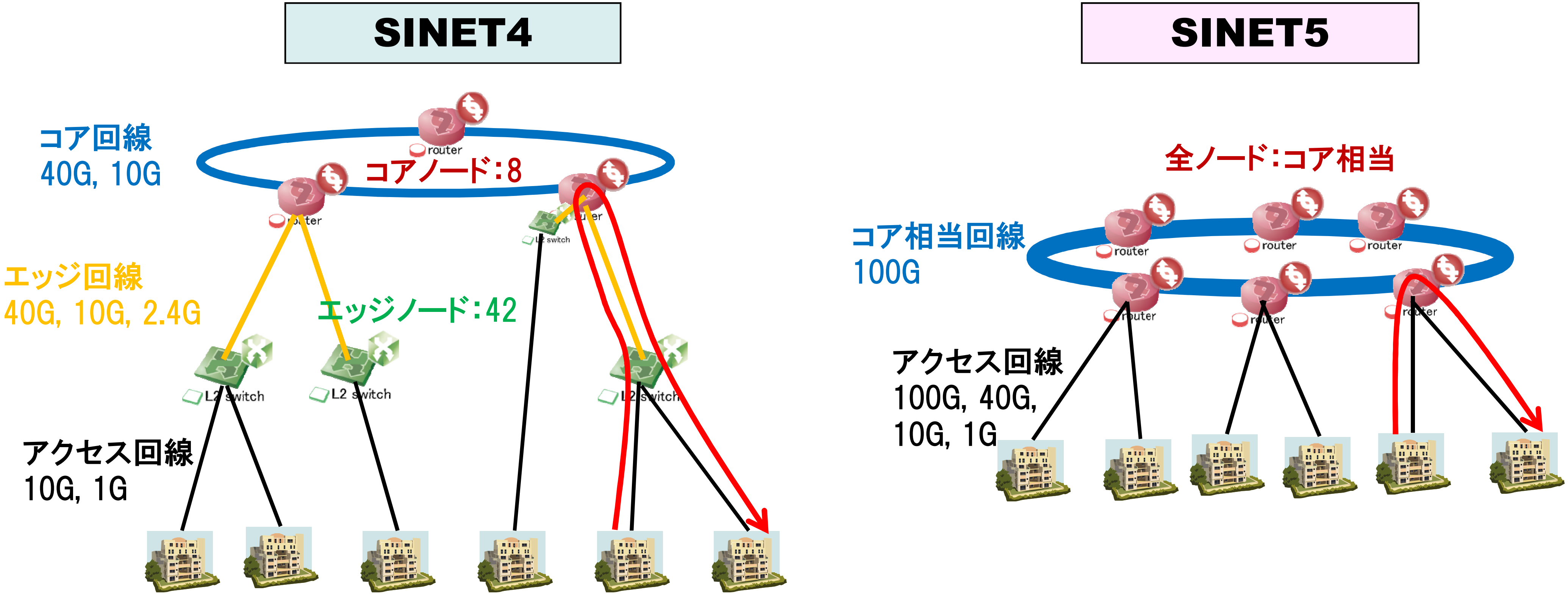
$$r = \frac{Win}{RTT}$$

1. 新アーキテクチャによるノード間遅延の最小化
   - 国内回線
   - 国際回線(欧州直結回線の新設)
2. 新アーキテクチャによるパケットロスの最小化
   - バーストトラフィック耐性の向上
3. SINET5での長距離高速通信テクニック
   - Linuxシステムパラメータチューニング
   - サイエンスDMZ

1. SINET5 国内ノード間遅延の最小化(1)

◆SINET5では、全ノードにルータを設置

◆SINET4では，エッジノードには，L2スイッチ（ルーティング機能無し）を設置
同一エッジノード収容機関同志の通信も，コアノードを経由する必要があり，遅延増大



1. SINET5 国内ノード間遅延の最小化(2)

◆SINET5では、各ノード間を2本(現用&予備)のMPLS-TP論理パスで接続. 現用パスは, ファイバルート上の最短経路に設定し, 遅延を最小化

◆IPレベルでは, 他のノードが全て隣接ノードに見える(フルメッシュ). ルータ経由数を最小化し, ルータ通過由来の遅延増を抑制

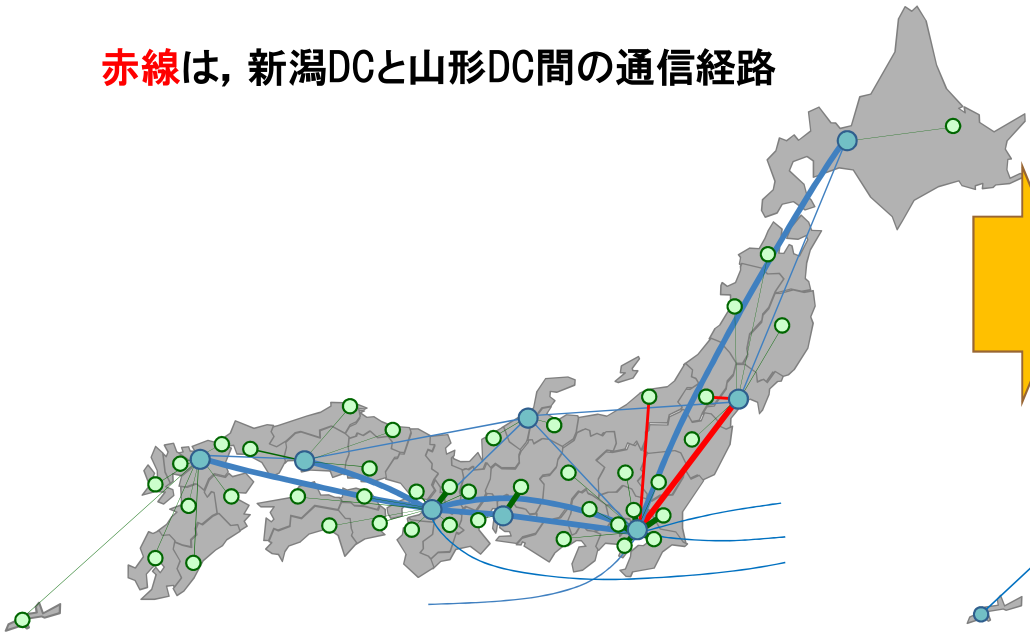
SINET4

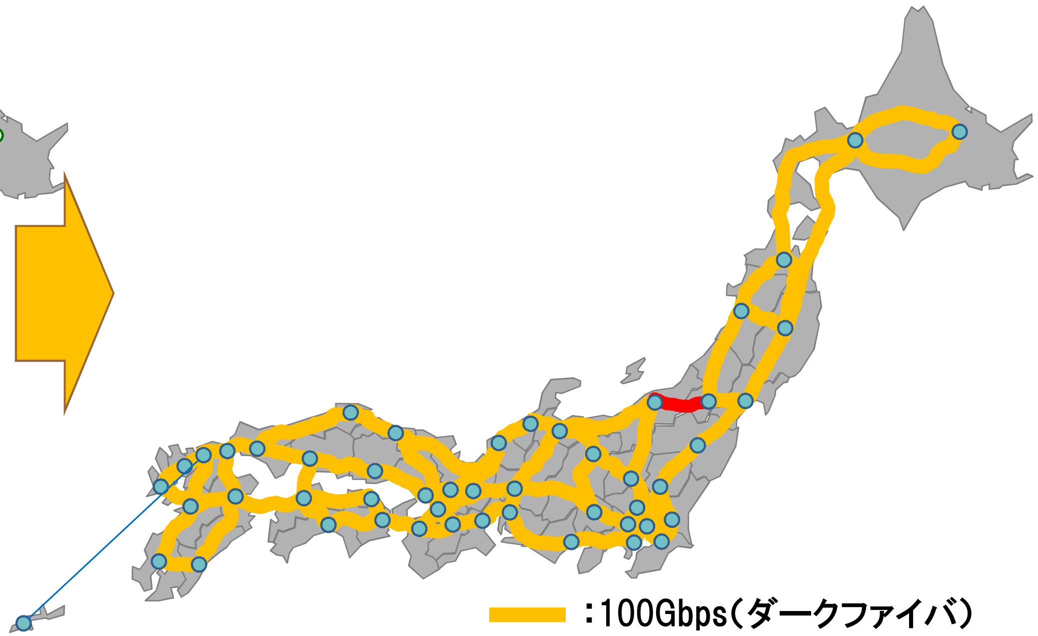
- エッジノードーコアノード間をスター接続
- ノード間は専用線で接続

SINET5

- 隣接ノード間をダークファイバ上のWDM波長パスで接続
- 各ノード間を2本(現用&予備)のMPLS-TP論理パスで接続
- 現用パスは, ファイバルート上の最短経路に設定

赤線は, 新潟DCと山形DC間の通信経路

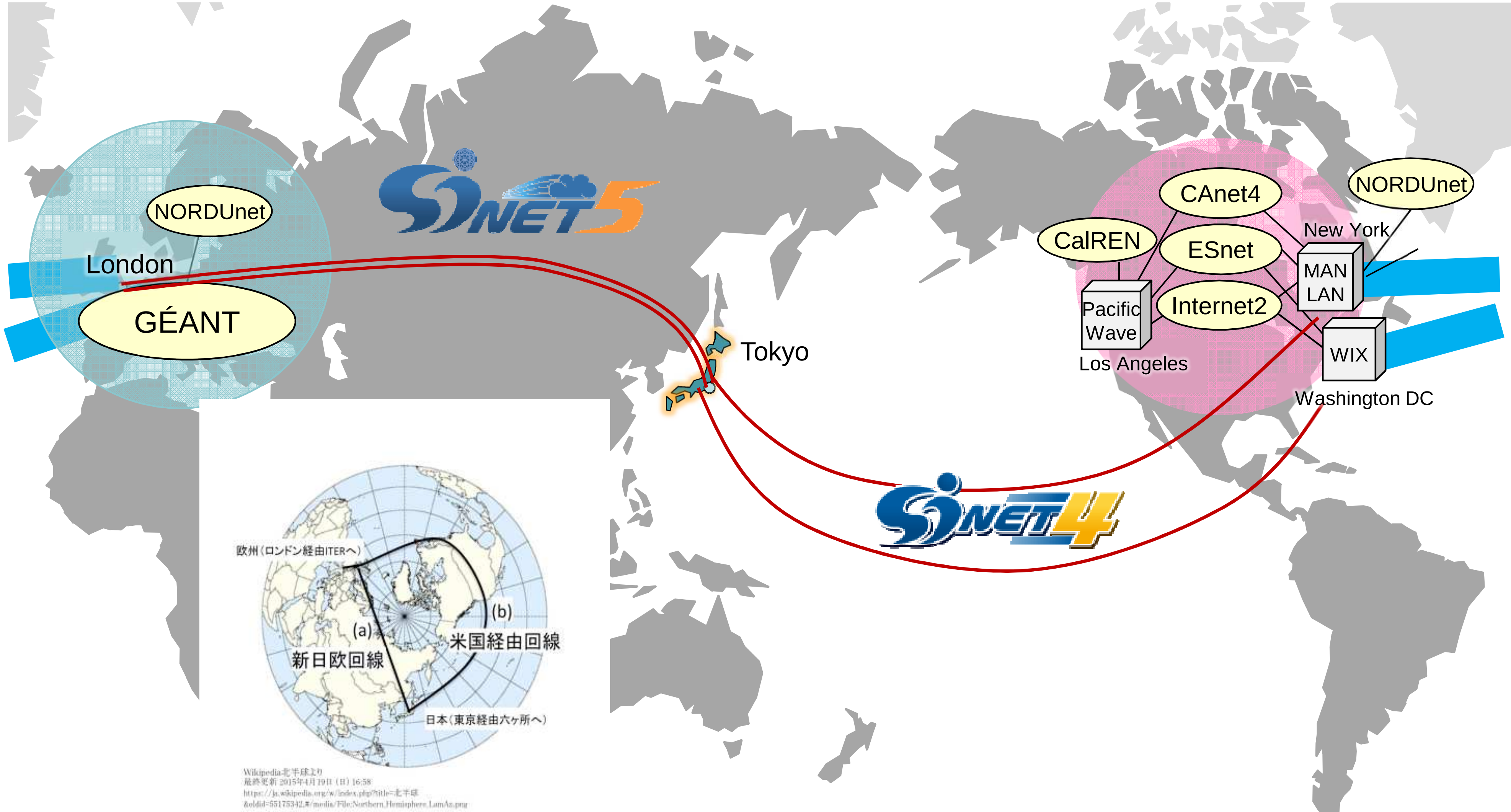




1. SINET5 国際(欧州)通信遅延の最小化

◆SINET5では、日欧直結回線(20Gbps)を新設

◆米国経由だったSINET4と比べて，遅延時間を約2/3の200ms程度に短縮

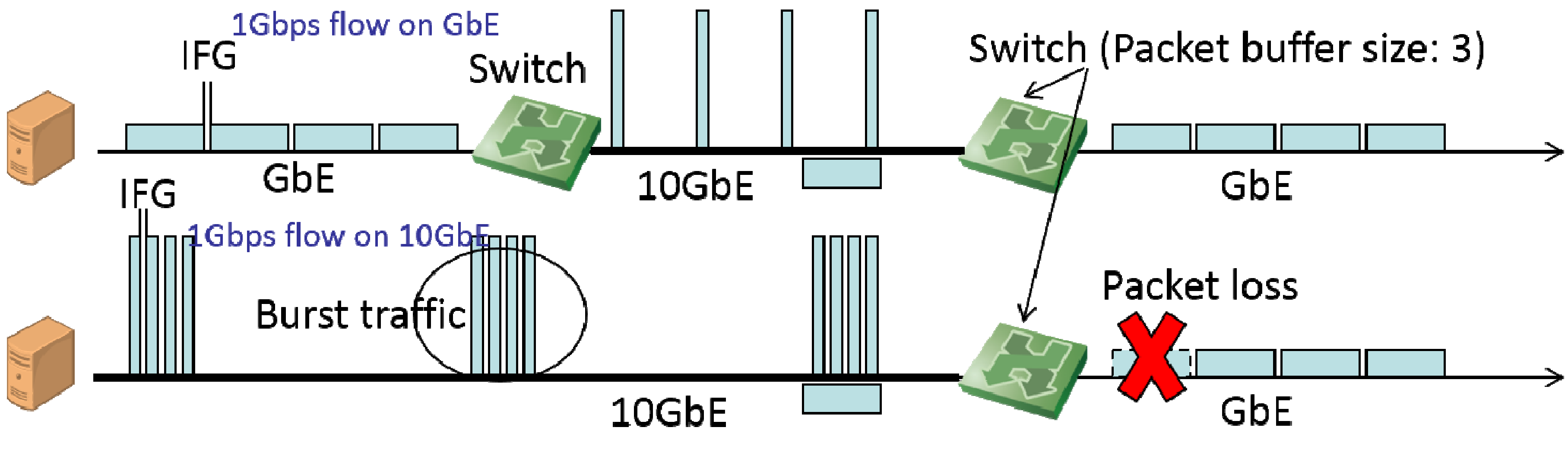


9/27量研機構他プレスリリース「国際熱核融合実験炉イーターの建設サイトから日本への大量データの高速転送を実証」より

2. SINET5 パケットロスの最小化

◆背景

- 高速NIC(10GbE以上)では，オフロード技術の副作用により，バーストトラフィックが発生しやすい．
- バーストトラフィックは，ネットワーク機器を通過する際に，パケットロスすることがある．典型的には，入インタフェース速度より出インタフェース速度が遅い場合．
- このパケットロスは，パケットバッファの不足により生じ，回線帯域に余裕があっても発生．
- 対策例：　ページング(IFGを強制的に拡大する)



IFG: Inter Frame Gap

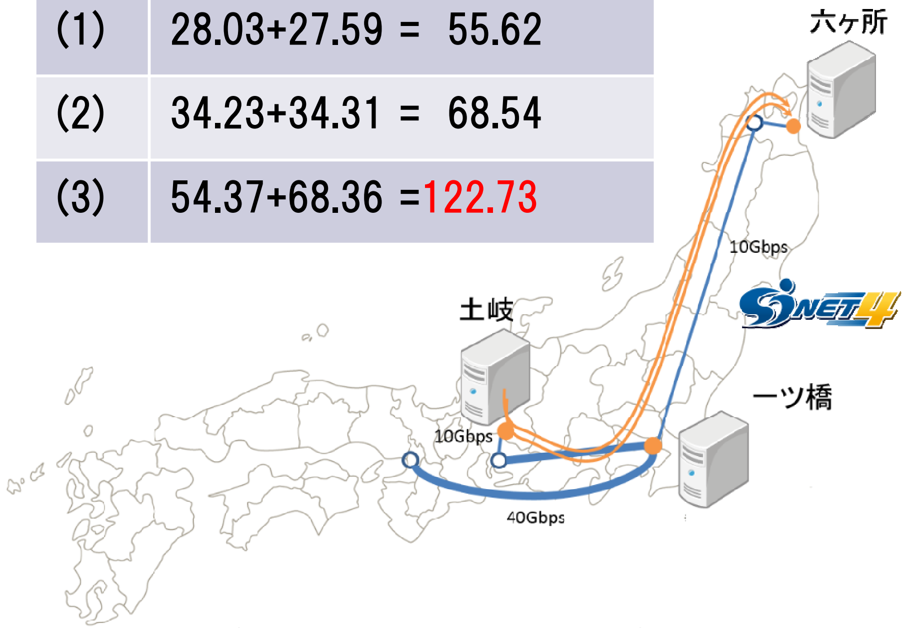
オフロード技術：　高速通信に伴うCPU負荷を軽減するため，OS機能の一部をNICが代行する技術．
　パケット分割をNICに任せる(TSO)，複数の受信パケットをNICが大きな一パケットに纏めてOSに渡す(LRO/RSC)
　一つの割込で複数パケット分の受送信完了を知らせる(IC)等

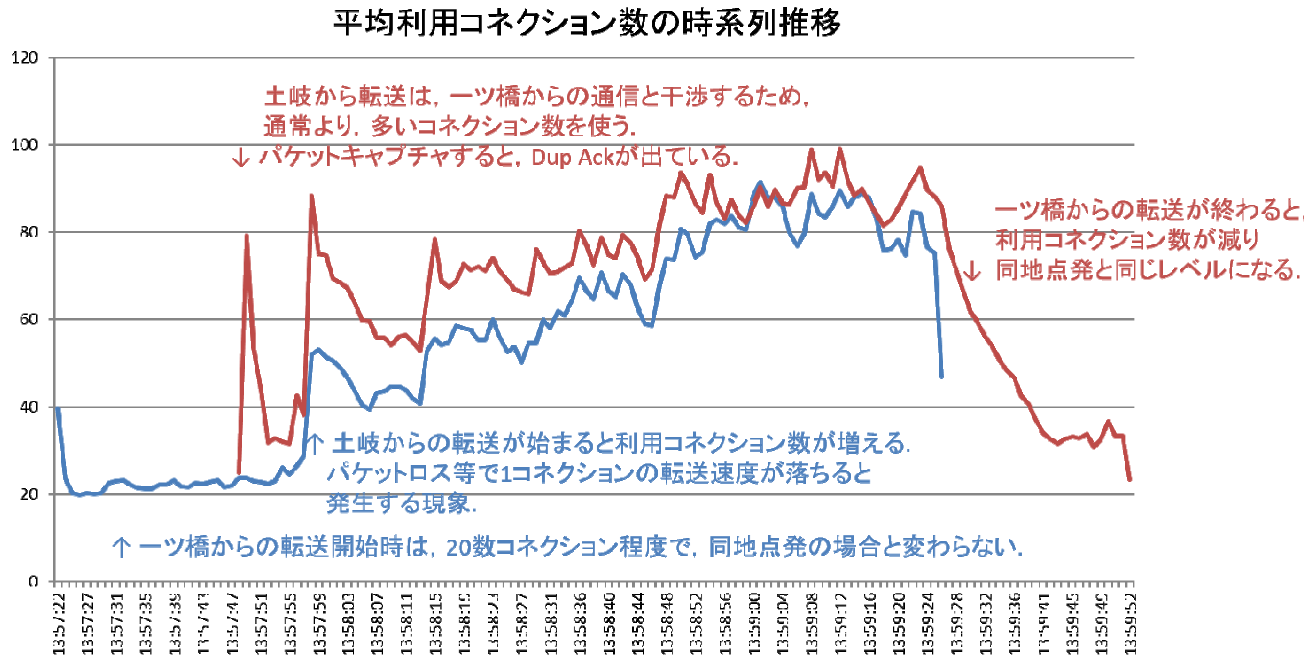
2. SINET5 パケットロスの最小化

◆SINET4でのパケットロス発生事例 (MMCFTPの転送実験)

- 目標速度1.3Gbps, 転送量20GBの転送(転送時間:約2分)を2フロー同時実行
- 一ツ橋発六ヶ所2フロー(1), 土岐発六ヶ所2フロー(2)時は, 特に問題なし
- 一ツ橋発六ヶ所と土岐発六ヶ所を同時実行 (3)時に, 前記と比べて利用コネクション数が倍に. 転送速度は, 目標速度を維持.

	平均利用コネクション数
(1)	28.03+27.59 = 55.62
(2)	34.23+34.31 = 68.54
(3)	54.37+68.36 =122.73



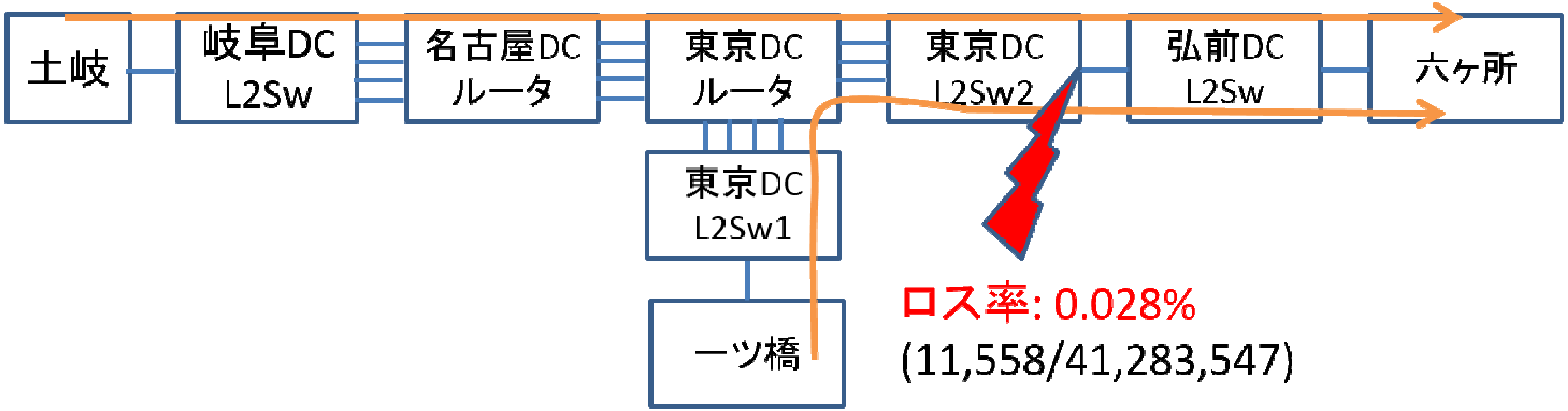


MMCFTP: Massively Multi-Connection File Transfer Protocol
利用者の指定した目標速度とネットワーク状況に応じて, 動的のTCPコネクション数を調整し,
一定速度を保つファイル転送プロトコル

2. SINET5 パケットロスの最小化

◆SINET4でのパケットロス発生事例（MMCFTPの転送実験）

- L2スイッチの10GbEx4のLAGから10GbEに速度が落ちるインタフェースで，パケットロスが発生
- 同地点発2フローの場合は，SINET入口の10GEインタフェースで，パケットがシリアライズされるのに対し，異地点発の場合は，該当インタフェースにパケットが同着する．
- 合計2.6Gbpsと回線帯域10Gbpsには余裕があるにも関わらず，同着時のパケットバッファ不足により，パケットロスが発生
- このパケットロスは，SNMPの標準MIB監視では発見できず，プライベートMIBを監視してはじめて認識できる（内部キューの構造はベンダー・製品依存のため）



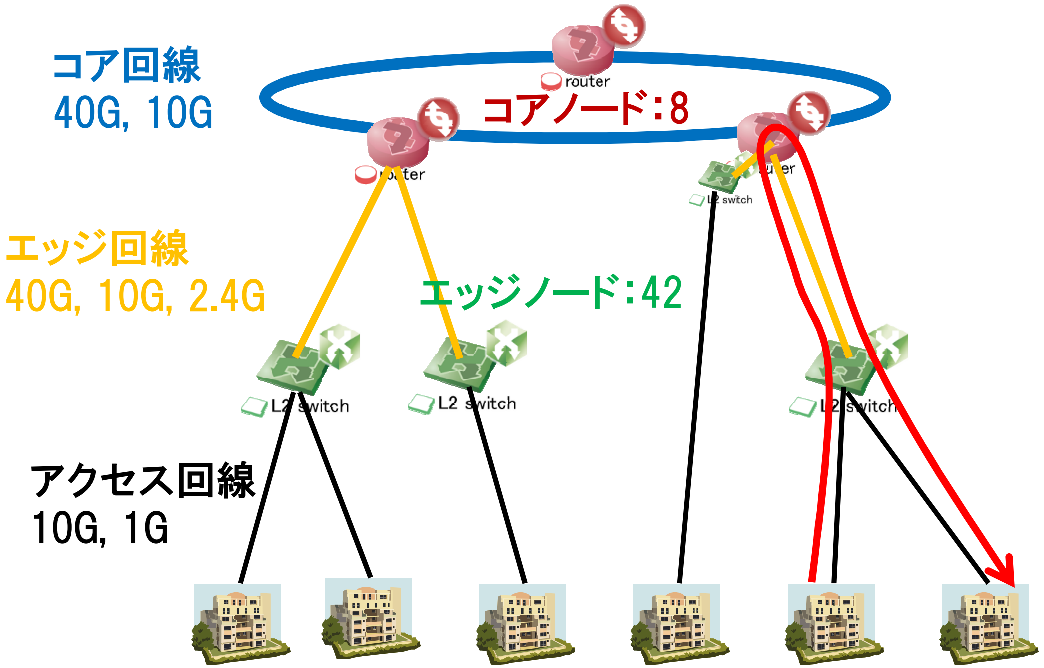
2. SINET5 パケットロスの最小化

◆SINET5では、全ノードにルータを設置し，加入機関はルータに収容

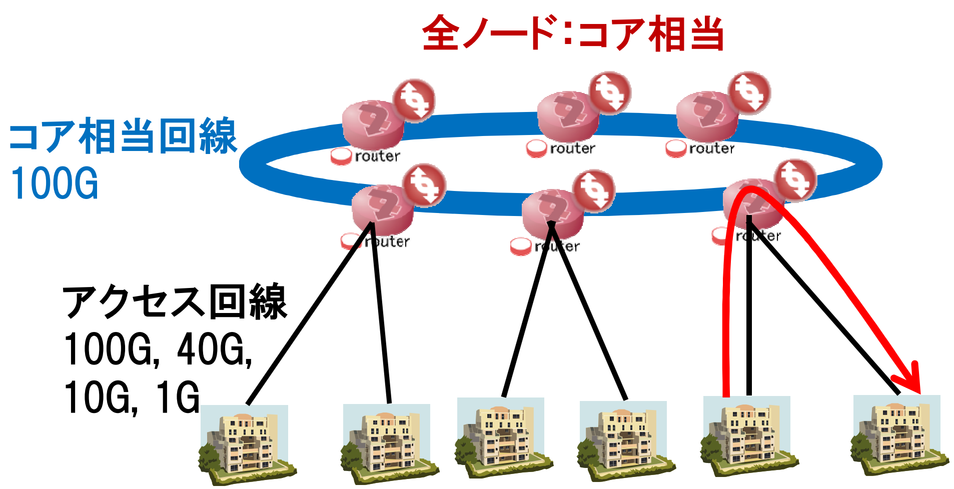
- L2スイッチは，通常，パケットバッファが少ないので，排除
- ルータの調達要件で，大容量パケットバッファを要求

2.11.5 QoS 制御を実施してもバーストトラフィックをパケットロスなく処理できるよう、インタフェースあたり 100msec 以上の遅延時間を吸収できるバッファ量を有すること。

SINET4



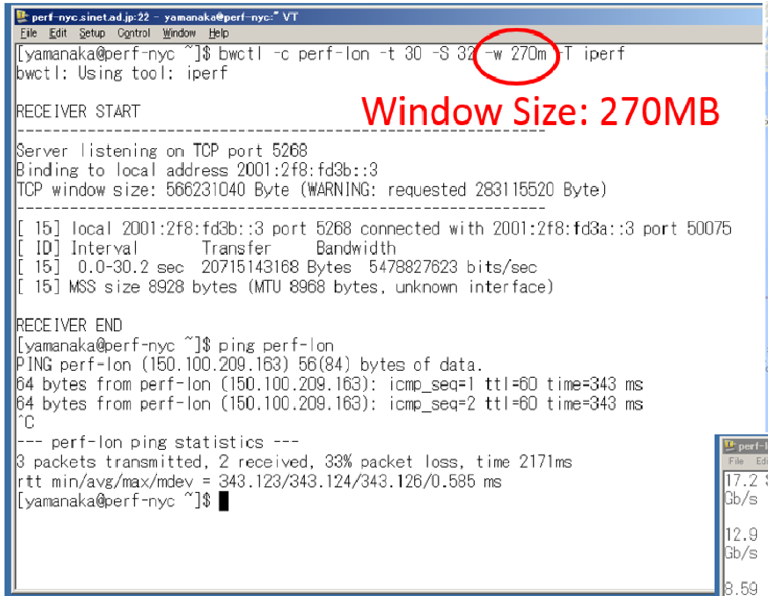
SINET5

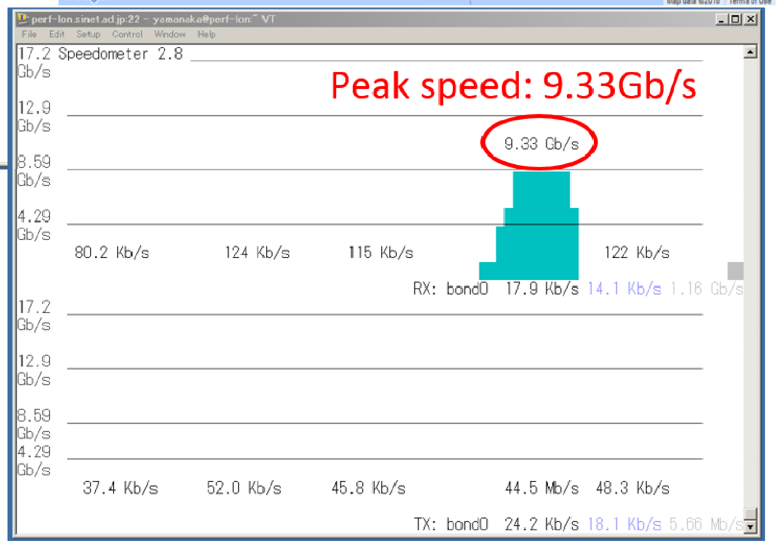


2. SINET5 パケットロスの最小化

◆SINET5での高速転送実験(iperf)

- ニューヨークとロンドンのperfSONARサーバ間で，iperfを用いて転送実験
- ウィンドウサイズを適切に指定すれば，帯域限界(10Gbps)に近い，約9Gbpsの通信が容易に可能





3. SINET5での長距離高速通信テクニック

◆SINET5では，国内通信に関し，地理的距離感と整合する，ネットワーク遅延を提供する．また，欧州との通信は直結回線により遅延が短縮された．

◆SINET5では，バーストトラフィック耐性の高いWANを提供する．ページングによるバースト性緩和は，SINET5内では必要ない．

◆以上から，SINET5は，長距離高速通信に適した環境を提供する．しかし，WANだけで高速通信が可能な訳では無い

◆自ホスト，相手ホストのマシンスペック，OS設定，LAN環境，転送アプリケーション等が通信速度に影響

◆転送アプリケーションに依存しない一般論の範囲で，長距離高速通信テクニックを紹介

- Linuxシステムパラメータチューニング
- サイエンスDMZ

3. SINET5での長距離高速通信テクニック

◆Linuxシステムパラメータチューニング

- Linuxシステムパラメータの初期値は，長距離高速通信に不適切なので，チューニングが必要

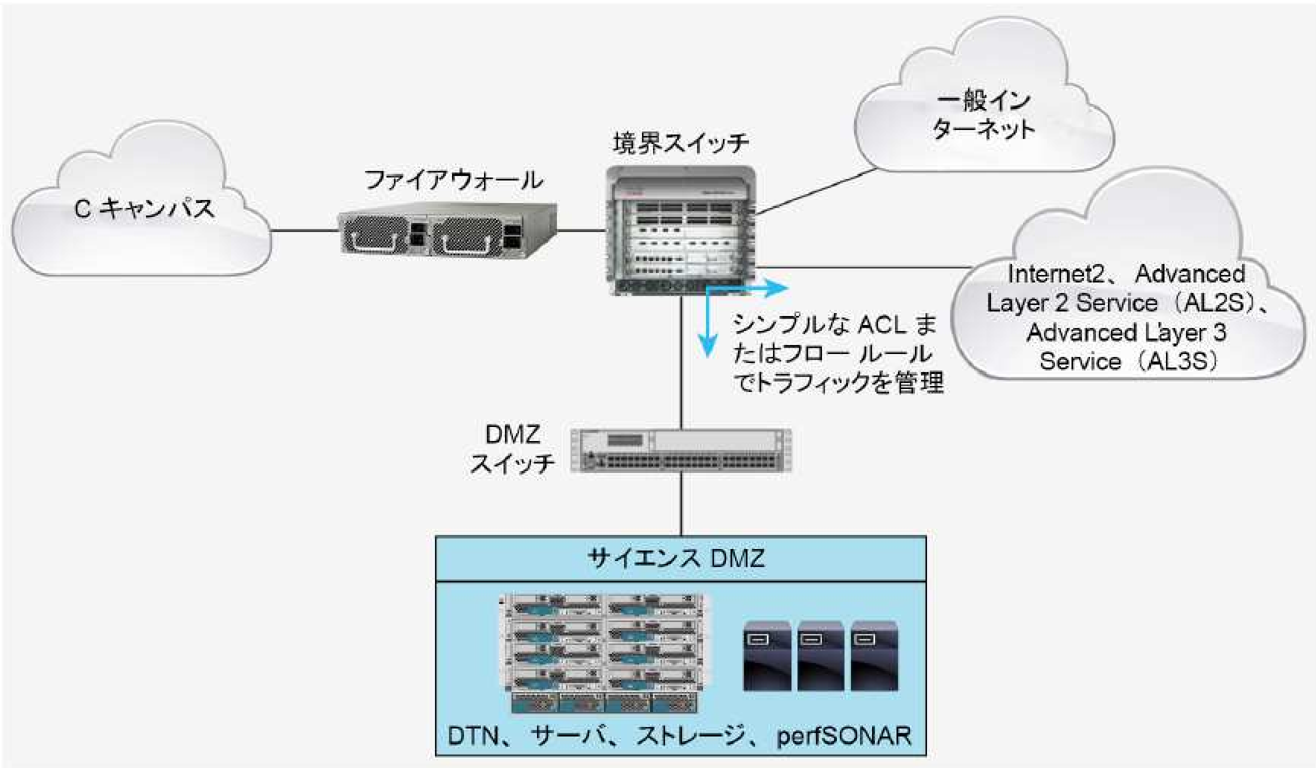
パラメータ	意味	初期値（RHEL 7.2）	初期値（RHEL 6.8）
net.core.rmem_max	受信バッファサイズとして明示的に設定された際に，許可する値の最大値	16MB	122KB
net.core.wmem_max	送信バッファサイズとして明示的に設定された際に，許可する値の最大値	16MB	122KB

- RHEL 6.8の初期値は，TCPバッファサイズの自動チューニングで規定される最大値よりも小さく，バグ．
- RHEL 7.2で，修正されたものの，16MBでは足りないこともある．
- 明示的に設定しなければ，使われない値であり，現在の64bitコンピュータでは大量のメモリを搭載できる．
- 長距離高速通信用マシンなら，TCPウィンドウサイズ最大値1GBで良いのでは？
- SINET5のperfSONARサーバは，1GB設定で運用中

3. SINET5での長距離高速通信テクニック

◆サイエンスDMZ

- インタネット経由のデータ転送では，自サイトと相手サイトで，IPS/Firewallを通過する. IPS/Firewallの性能が，データ転送速度を制約する.
- 信頼する相手と閉域網(L2VPN, L3VPN)で接続し，IPS/Firewall無しにデータ転送を行うのが，高速通信には有効.



CISCO社ホワイトペーパ「イベントベースのソフトウェア定義型ネットワーク：セキュアなサイエンスDMZ」より

ご清聴ありがとうございました