本資料は，公開用資料です．諸般の都合により，講演資料と異なる場合がございます．

パケットフォワーディングを支える技術
200Gbps
～更なる高速化への挑戦～

大江将史　< masa@fumi.org >
自然科学研究機構 国立天文台

http://fumi.org/ULTRA

自己紹介

- 大江将史　（おおえ　まさふみ）
  http://fumi.org/
- 自然科学研究機構 国立天文台
  - NAOJ: National Astronomical Observatory Of JAPAN
  - 天文データセンター　助教
- なにしてるのか？
  - 専門は、ネットワークセキュリティ、衛星通信、無線通信など
  - 天文と情報ネットワークの融合に関する研究等
  - 国立天文台のネットワーク運用や設計等

予算額70万円程度でどんなものができるでしょうか？

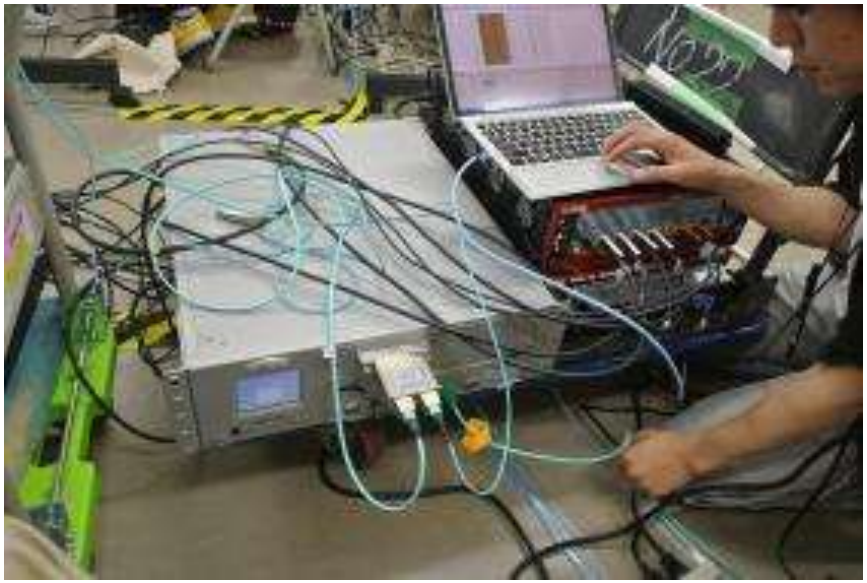
- 国立天文台では，性能限界に挑戦すべく 100Gbps級のトラフィック処理力を有するPCサーバを試作しました.
- 高機能IPルーター 「野川」と「大沢」
  - 実効L3 バックプレーン容量 75Gbps～100Gbps程度
    - 野川 I/F：QSFP+ 40GBASE-R x 1 + 10GbE x 2
    - 大沢 I/F：18x 10GBASE-R
  - 野川は，超高速SSDストレージを搭載（SSDは予算外☺）
    - 書き込み性能に特化したSSDを16台搭載し，4GByte/sec の書き込み性能

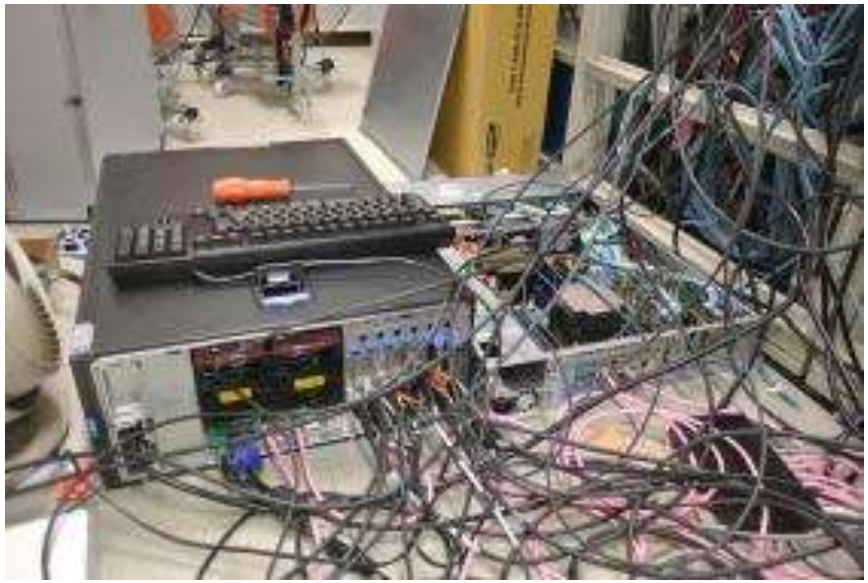
野川

大沢

性能を計ってみました

- IXIA社 とSpirent社の計測器 を接続し， L3伝送性能とトラフィック送受信能力を検証いたしました.
  - 合計で160Gbpsまで検証が可能

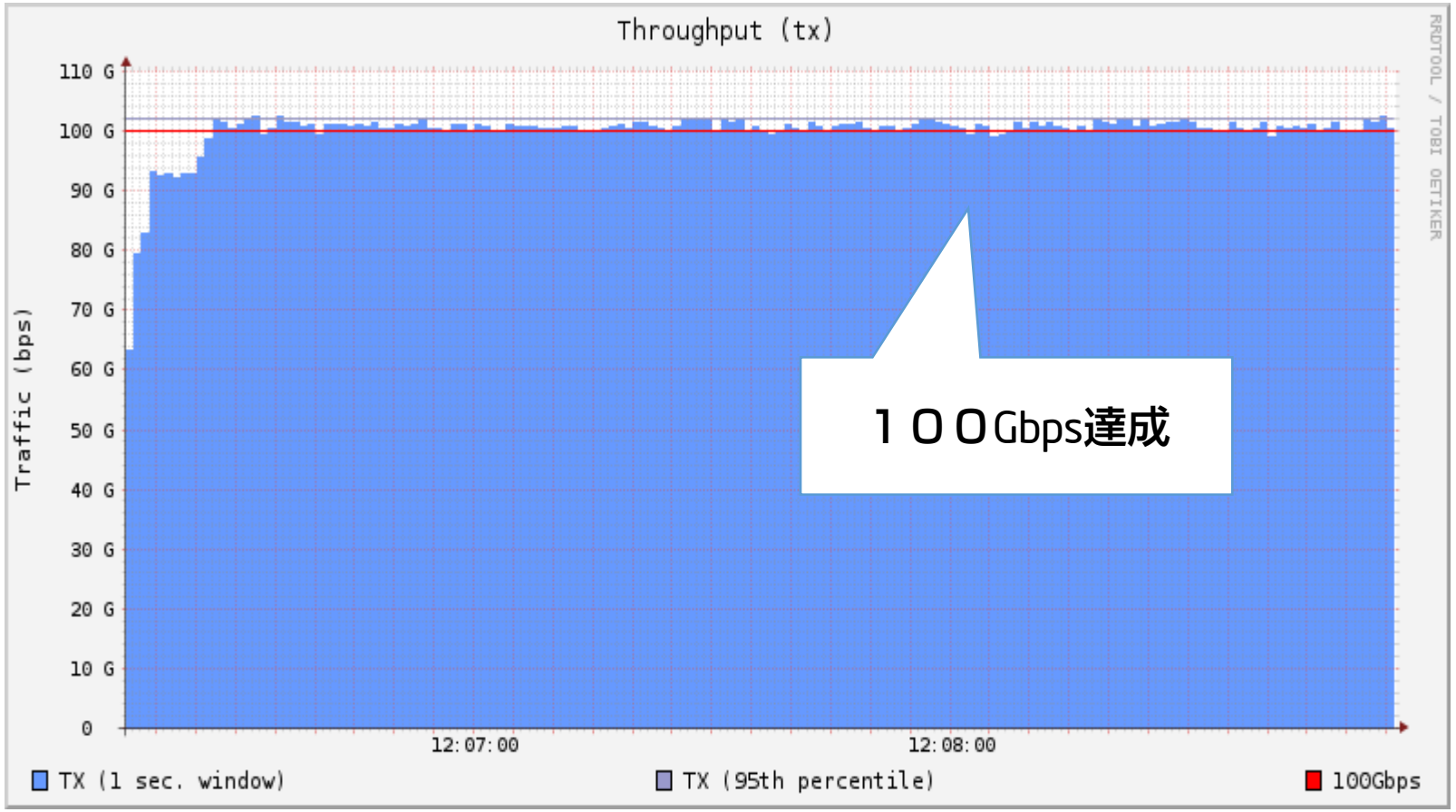




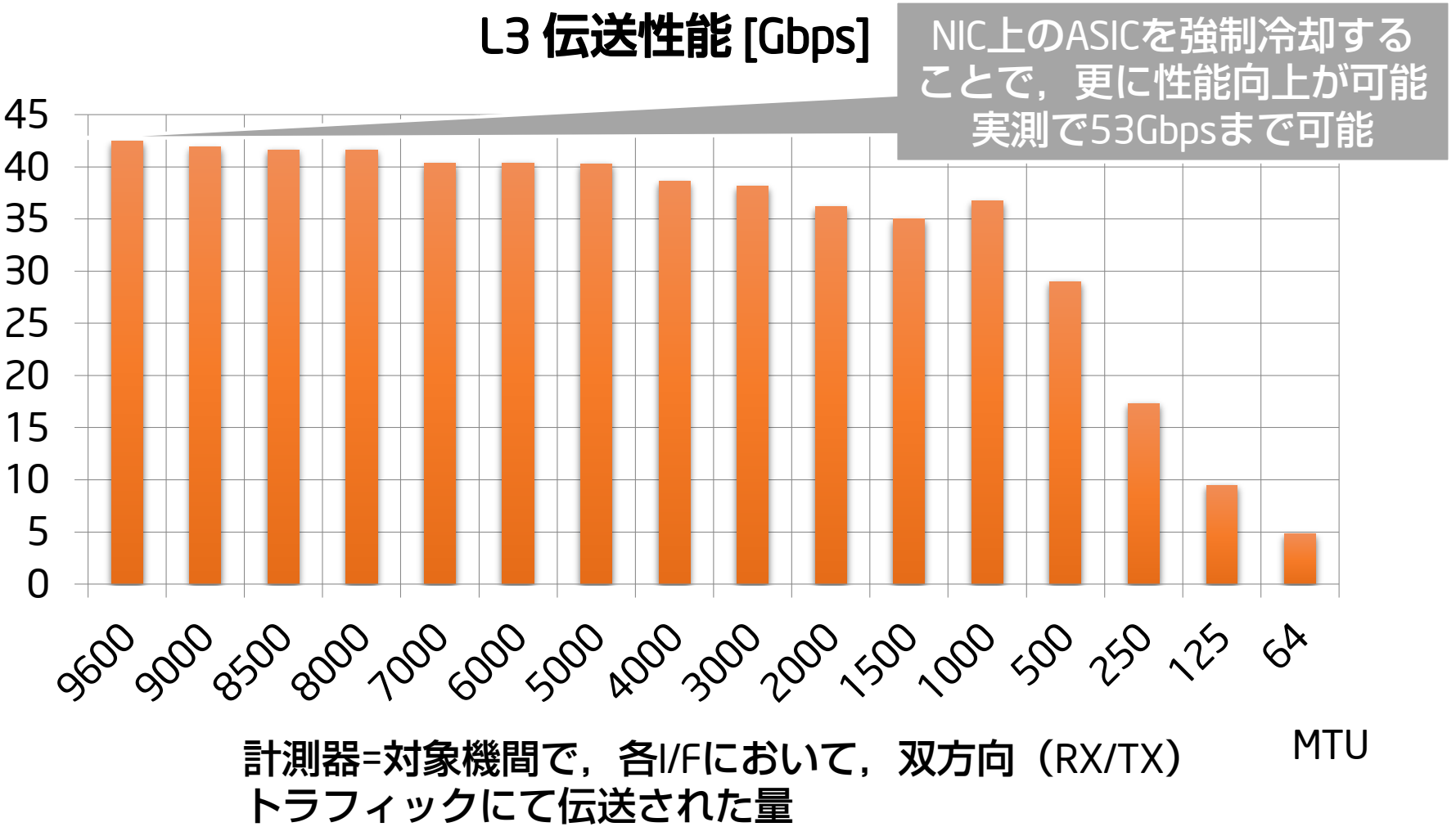
10GBASE-DA or SRにて，最大160Gbpsで接続

ぎりぎり100Gbpsのトラフィックを生成
そう！次は，200Gbpsだな.

- UDPパケットを利用したサービス妨害攻撃を想定したトラフィックを大沢から生成



IPルータとしてのフォワーディング性能



計測器=対象機間で，各I/Fにおいて，双方向（RX/TX）
トラフィックにて伝送された量

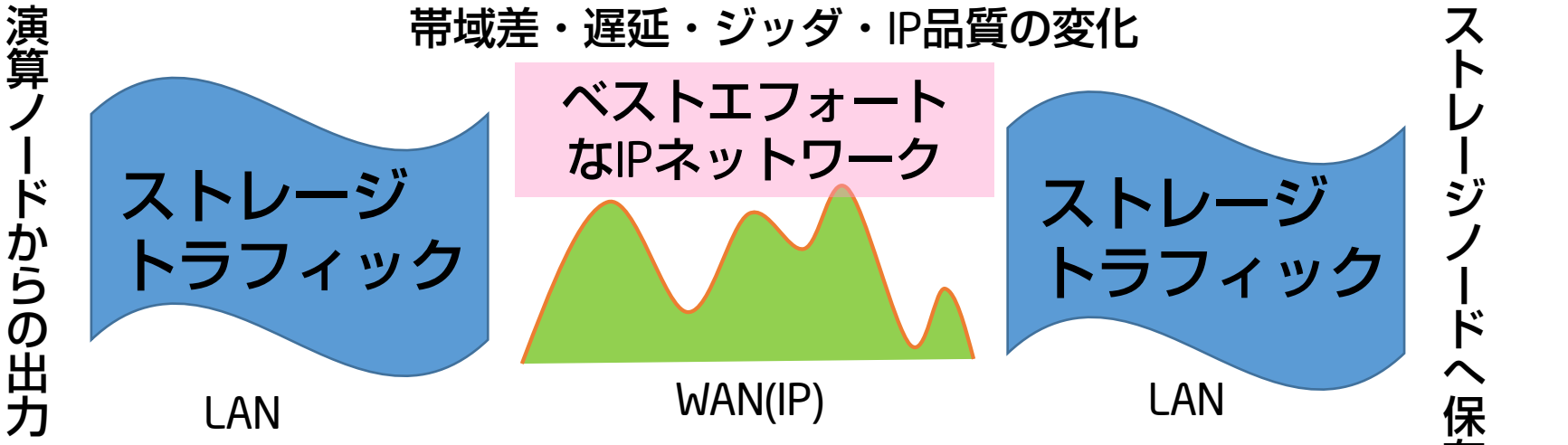
背景

登場背景

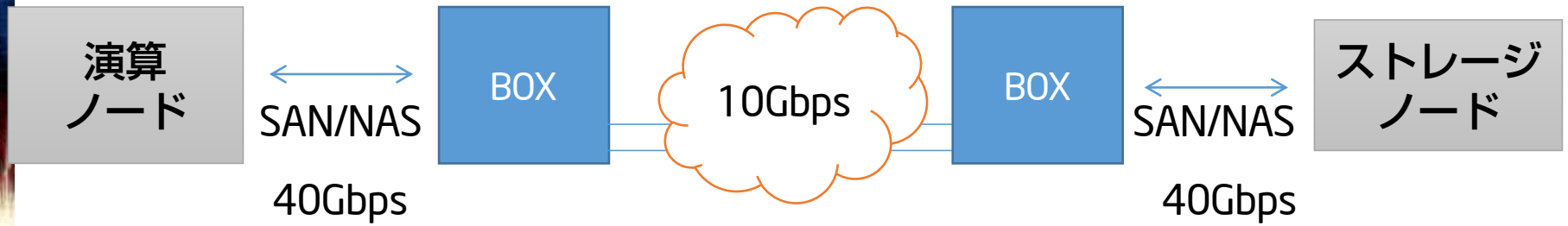
取り組みの背景

- 2013.4 CRAY社製新HPCシステム稼働
  - 岩手県奥州市・東京都三鷹市に分散した分散HPCインフラを構築
    - 演算性能：600Tflops～, 2015(?) 1Pflops
    - SAN性能：40Gbpsクラス
    - ネットワーク帯域：10Gbps(10GbE)
      - NICT JGN 奥州市水沢 AP 開設
- 国立天文台仮想化システムの構築と運用
  - 部品厳選からはじめる仮想化基盤の構築
  - 三鷹・大手町・水沢（岩手）に分散
  - インフラ，情報公開用途，実験などに利用
- 研究開発をインハウスで進める
  - 研究予算上のコストパフォーマンスや経験の伝承の観点から重要と位置づけ
- 現在，絶賛開発中
  - 各種イベントに途中経過として発表しています.

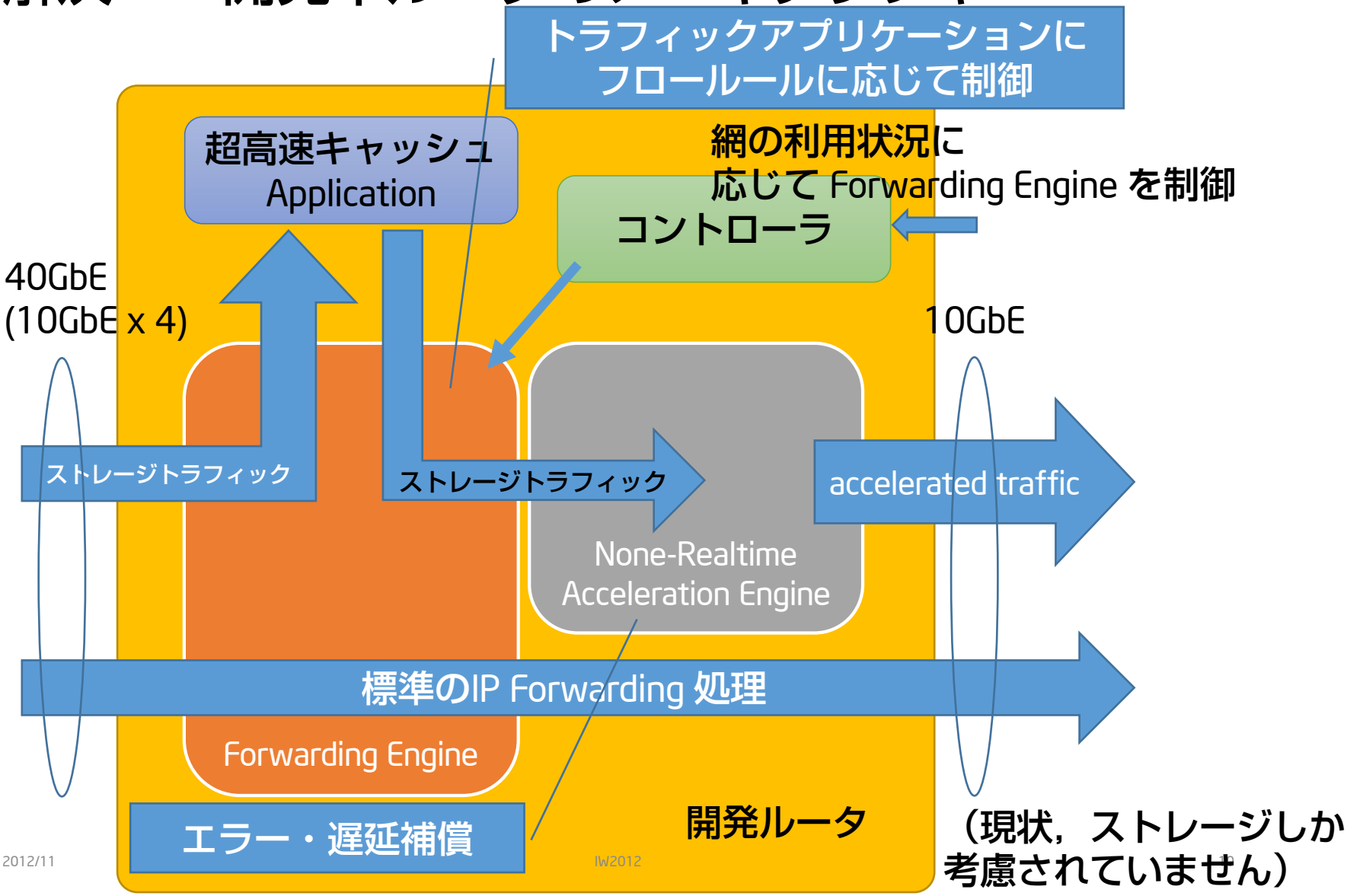
解決へ：開発中ルータのアーキテクチャ



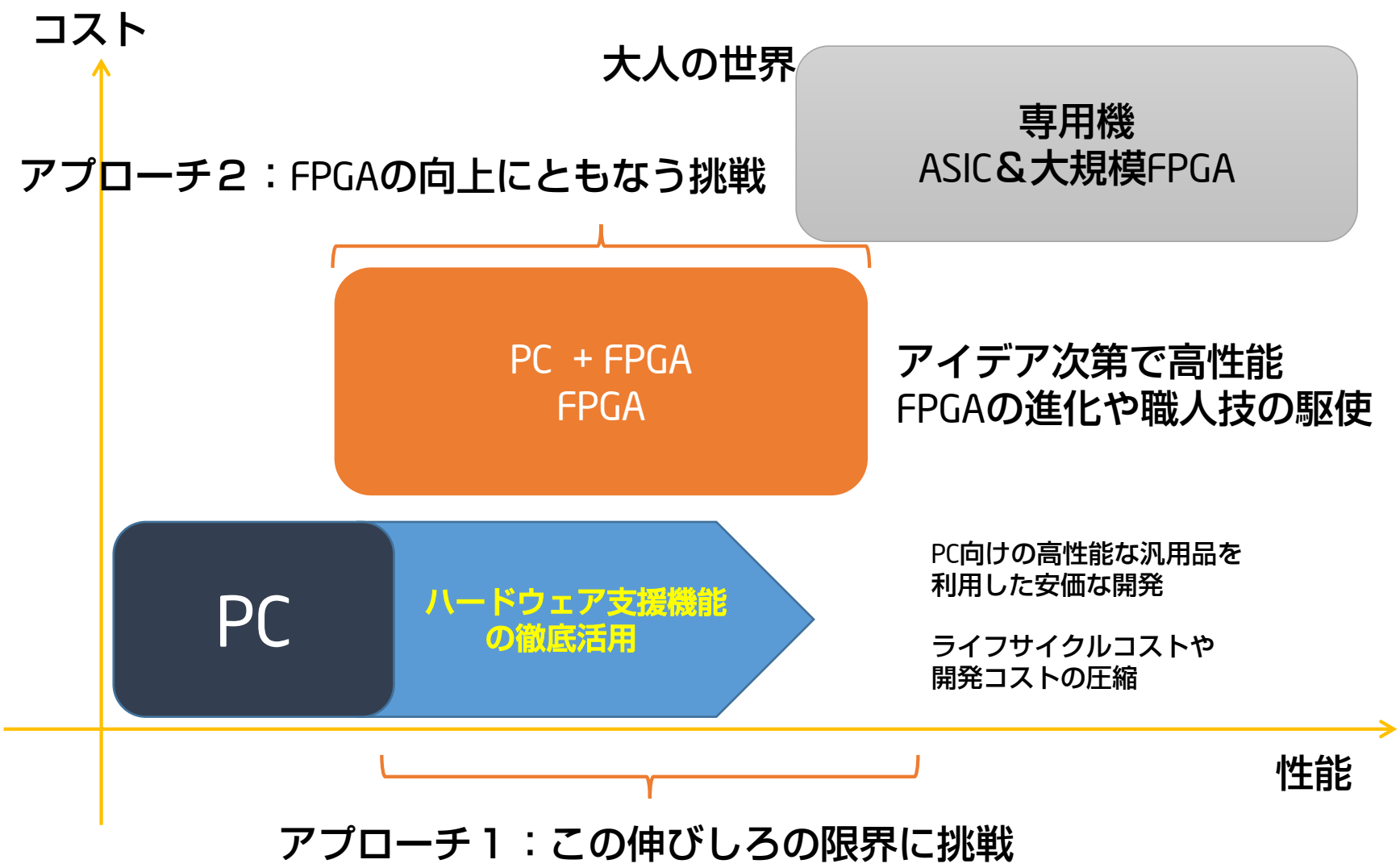
超高速ストレージキャッシュ・ストレージトラフィックの予測に基づく伝送制御・高効率伝送方式を組み合わせて解決



解決へ：開発中ルータのアーキテクチャ



解決へ：開発手法とトレードオフ



さて，どうなんでしょう？

高速イーサネット・FPGA・PC・・・私たちの今

100Gbpsと私たち
～高速イーサネット：40GbE - 802.3ba～

- 40ギガビットイーサーネット
  - データセンタ，エンタプライズでの利用が進む
  - NIC・スイッチングハブ：廉価化が進む
- 100ギガビットイーサーネット:現状，キャリア向け
- 現状
  - サーバは，10ギガ，40ギガ
  - （エンタープライズな）ネットワークでは，アクセスが10/40ギガ，バックボーンは40ギガがお買い得
    ➔来年以降なら，100Gのバックボーン，アクセス40G も視野に

Extreme BDX8
10Gx768 40Gx192

Arista 7508
10G x 384

100Gbpsと私たち
～高性能な汎用スイッチングハブ用チップの流通～

- 10G/40Gスイッチングハブ向け高性能な汎用チップにより，100万円前後の廉価な10G/40G多ポートハブが多数流通
  - Fulcrum Microsystems
  - Broadcom
  - 各社ボックススイッチは，汎用チップを利用している場合が多い.

100Gbpsと私たち
～XFPからSFP+，Ｔ，CFP から QSFP+～

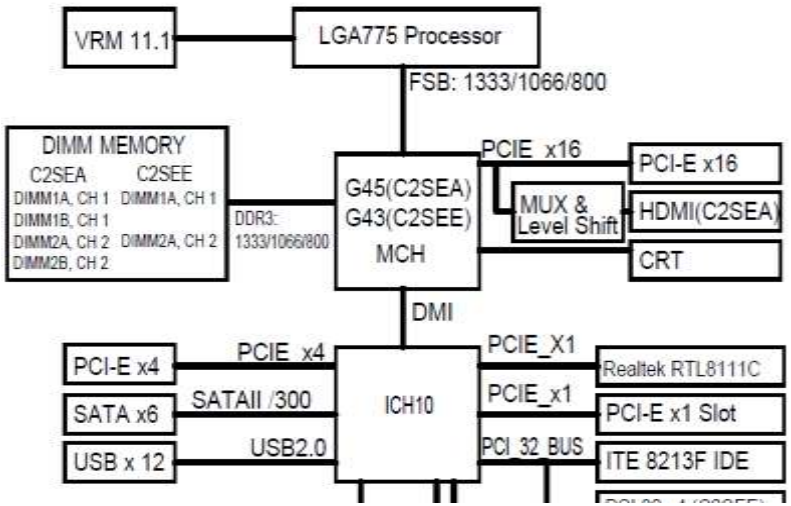
- 10GbEは，SFP+/T対応による低コスト化
  - 10Gbase-CR(DA: Direct Attach) 数十$～150$程度
  - 10Gbase-T (CR+100$: ➔0$は時間の問題)
- 40GbEは，CFPからQSFP+
  - 40Gbase-SR4 : MPO(8芯MMF)コネクタ
    - 高密度に10Gbase-SRを収容可能
  - 40Gbase-CR4(DA) (安価)
  - 40Gbase-LR4もモジュールメーカから出荷が進みつつある
    - 40Gbase-LR4 QSFP+モジュールが，約5500US$程度

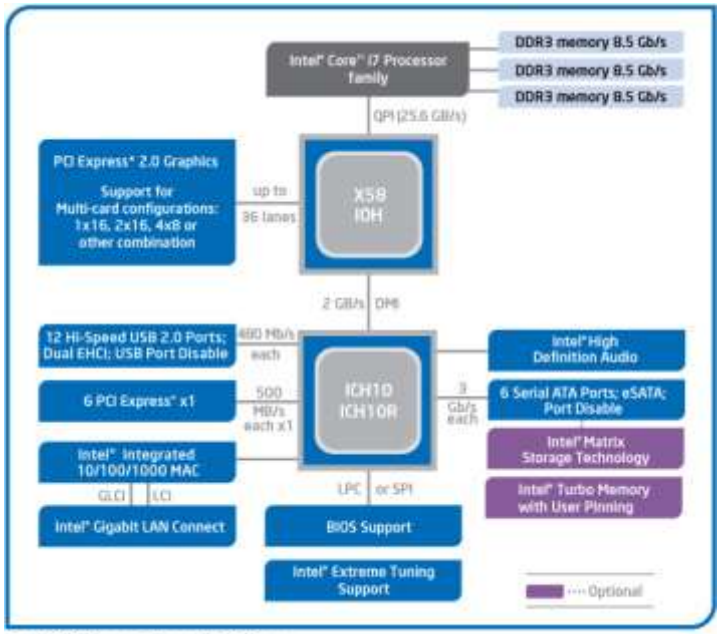
100Gbpsと私たち
～バンド幅ではなく低遅延への要求～

- Arista Networks/エクストリーム社のアプローチ
  - 伝送の低遅延化
  - キャリアコアではなく，エンタープライズコアの優位性を生かす
- 金融が低遅延化を引っ張る
  - アルゴリズムに基づくトレーディング（システムトレード）
    - 遅延＝損失リスク大
  - 自社システムの金融取引所への直結
    - 何フィートの争い
- 低遅延化も強く意識された製品が増えている
  - ＰＣの高性能化
  - 高機能ＮＩＣ（FPGA on NIC）
  - FPGA入り ハブ

100Gbpsと私たち
～PCサーバの高性能化と短寿命～

- PCアーキテクチャの進化
  - PCI-E 2.0/3.0，メモリーコントローラがCPU搭載
    - 高速・大容量メモリ
  - QPI / HT，ノースブリッジ統合，IOHハブ統合
  - 3Dゲーム，FullHDとGPU
- 専用機でしかできなかった事がコンシューマ向け汎用品でできるようになる.
  - GPU / 大容量メモリ / 高速なI/O





Intel® X58 Express Chipset Block Diagram

100Gbpsと私たち
～PCサーバは，バランス感覚が重要～

- 最高峰のCPUは，お値段も最高峰
- メモリーをたくさん載せれば，速度は低下
  - トレードオフ
- 限られた予算内でもっとも高いパフォーマンスを出すことを意識することは重要です．

例）2012年
Xeon E5-2665 2.40GHz T/B 3.1GHz
LGA2011 QPI-8GT 8Core x 2
DDR3-1600 8GB rank2 x 16　➔60万円以下

2011年:
XeonX5675 x 2 3.06GHz T/B 3.46GHz QPI6.4GT x 2 96GB ➔65万円

100Gbpsと私たち
～NICの進化・廉価化～

- 40ギガ・10ギガのネットワークカードの市場価格は？
  - ２００９年：１０ギガｘ１　２０万円
    - FCのチップを転用した10GbE NIC
  - ２０１０年：１０ギガｘ２　＜１０万円
    - ただしワイヤーレートではない.
    - ネイティブな10GbE ASIC搭載 NIC
  - ２０１１年：１０ギガｘ２　＜７万円
    - 各種Offload
    - ワイヤーレート
    - オーバーサブスクライブ（２８ギガ程度）で，１０ギガｘ４や４０ギガ登場

➔急速に高性能化と低価格化が進んでいる.

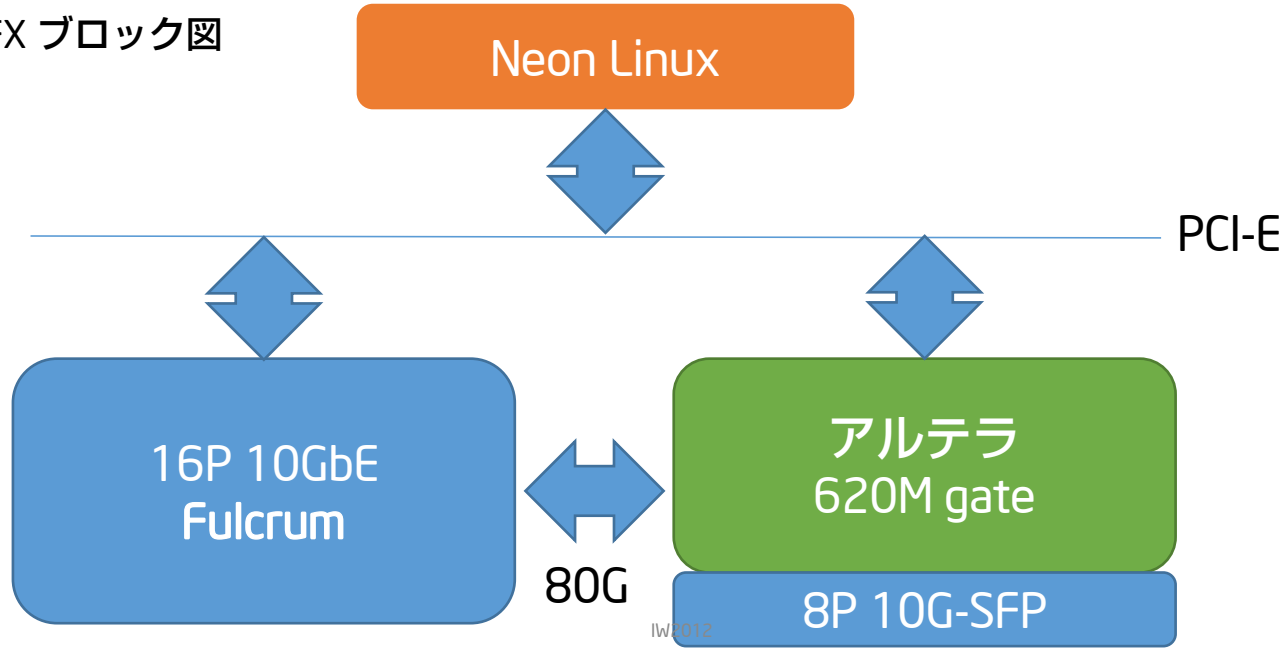
100Gbpsと私たち
～NICの進化～

- PC性能向上にあわせてNICも進化
- 帯域
  - 現在，PCI-E Gen2 NICで30Gbps級の速度，PCI-E Gen3で，40Gbps
    - 40GbE(10GbEx4の40GbaseR) や，10GbE-T NIC は，650$位
  - 2015年以降に，100Gbps登場予定
    ➔つまり，2015年位から100GbEの安価な環境が整う
- 低遅延化
  - 2.9μsec (2011)-> 1.5μsec台へ(2015?)
- 高機能化
  - オフロードエンジン(IP/TCP/UDP/ Application)
  - NIC DRAMへのダイレクトアクセス
  - 内蔵L2パケットスイッチング機能
  - プログラマブルNIC

100Gbpsと私たち
～FPGA搭載の高速スイッチングハブ～

- Arista 7124FX : FPGA programmable でありながら，Fulcrum のワイヤーレート CHIPも搭載
  - FPGA戦略で言うところの「FPGAのみ」と「PC」の中間的存在
  - 定価600万(?)

7124FX ブロック図



ここまでのまとめ

- NICもPCも手の届く範囲でワイヤーレートを達成するハードウェアが安価に整う.
- 低遅延への対応や，一部NIC搭載の機能を活用することで，特殊な応用へも対応可能
- FPGA搭載 スイッチングハブにより，ソフトウェアやNICハードウェア支援機構のみでは達成しにくいアプリケーションへの対応も可能

つまり，10Gbps/40Gbps程度をターゲットにした研究開発は，十分にできるのです.

➔野川や大沢が登場した背景

システム例：NAOJ CFCA ストレージシステム

究極：ケースなどいらないのです.

実機を見学出来ます

PCI-Eカードの
固定はネジりっこで
十分

超速で保守交換

超空冷

性能の絞り出し

- 天文台では，野川や大沢といったPCルータ以外にも，仮想化用（クラウド）システムを作成したり，大規模なストレージシステムを開発運用しています．

絞り出すには

- 最新のNIC + PC ならば，10Gbps程度を絞り出すのは簡単ですが，乾いた雑巾級に絞りきるにはノウハウが必要です．
- 最適なパーツ選定・PCI-Eのソケット選定
  - 各種M/BやNIC，RAID，SSDの性能調査やメーカーへのFirmware改善と性能向上
  - ゲーミング用M/Bは，微妙
  - SSDは，ファームで劇的に変動
    - 悪くも良くもなる．ほしいのは安定した実効性能

絞り出すには

- 割り込み処理の最適化
  - Receiver Side Scalingにより，各CPUへの割り込みを分散
    - コントールを維持するために，全部に振らないのも考えかたの一つ.
  - 割り込みの集約化と待ち時間の調整
    - 遅延と性能のバランス

絞り出すには

- Linuxカーネルでの割り込み・OS上の無駄な機能排除
  - ACPI IRQバランスの禁止・CPU speed の制御禁止
  - rx/バッファの調整・MTUの調整
    などなど

絞り出すには

- NIC搭載のハードウェアを活用
  - IP/UDP/TCP/Bondingオフロードエンジン
  - 組み込みL2フォワーディング機能
  - 2 ポート間のフォワーディングを内蔵の組み込みL2 SWを利用する
  - Userlandまでダイレクトに通過トラフィックを収集
    - CAMの関係上，100MACまで

よいこと：ノウハウをどんどん吸収して台内展開

- 国立天文台では，2009年初頭からサーバー仮想化を開始
- 仮想化＝PCハードウェアの更新頻度向上
  - コスト対性能の高いハードウェアの持続的投入
  - ハードウェア更新が仮想化与える影響小
    - ハード寿命は4年程度でも2年程度でどんどん格下げ
- 10Gbpsネットワーク機器なども安価に
  - ネットワーク機器は，40Gbpsクラスへ
- 目的達成には手段を問わないが，予算規律は厳しく
  - メーカ品には手が出ない：CPの高い構成の意識
  - 幸いにスペースはありますので，ケース無しから，ケース有りまで用途・CPに合わせて考慮

みなさんに伝えたい

作る・絞り出す・役に立つ・失敗する
➔おもろい&ノウハウも蓄積

お問い合わせ・見学歓迎

http://fumi.org/ULTRA/

masa@fumi.org