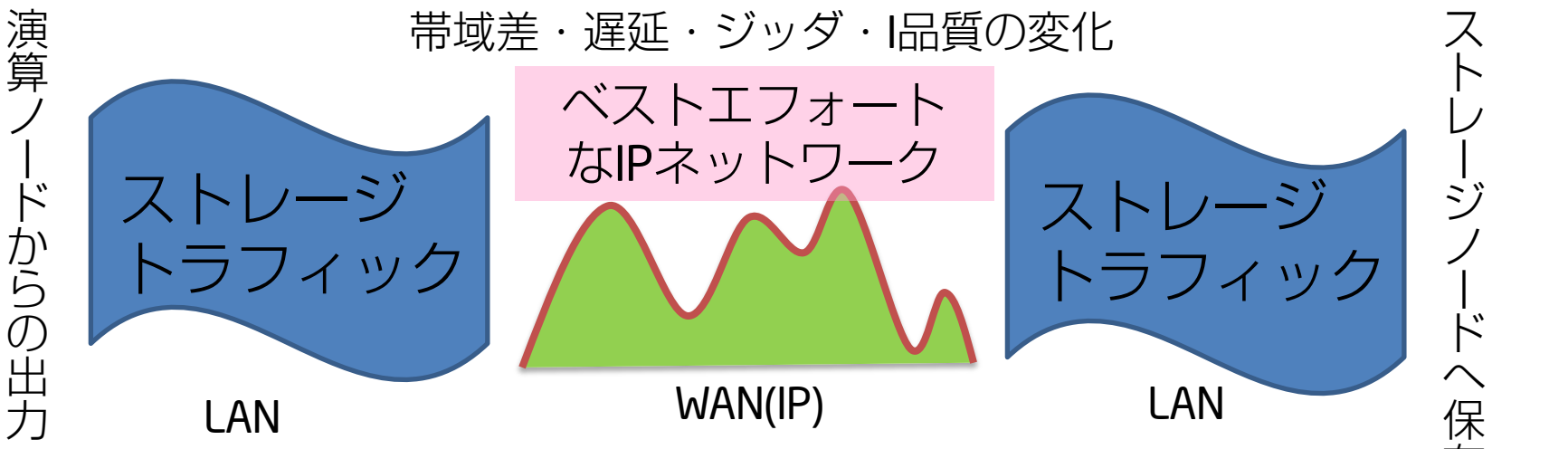
ULTRA40
～サイエンスを加速する力～

ウルトラ40プロジェクト
国立天文台
おおえ・あさい・いのうえ
National Astronomical Observatory of JAPAN

背景

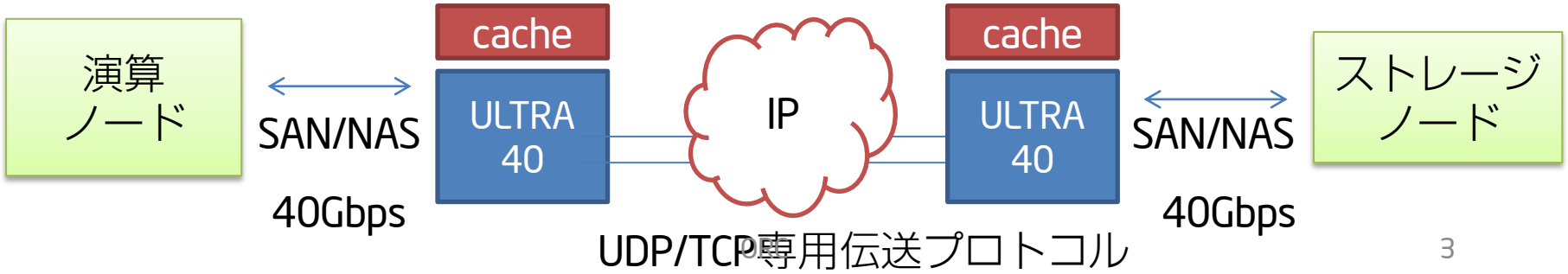
- 次期 天文台HPCシステム
  - 2013.4 岩手県奥州市・東京都三鷹市に分散した分散HPCインフラを構築
    - 演算性能：600Tflops～
    - SAN性能：40Gbpsクラス
    - IPネットワーク：10Gbps
- 演算・ストレージ・IPネットワークを効率よく連携させる仕組みが必要

課題の解決

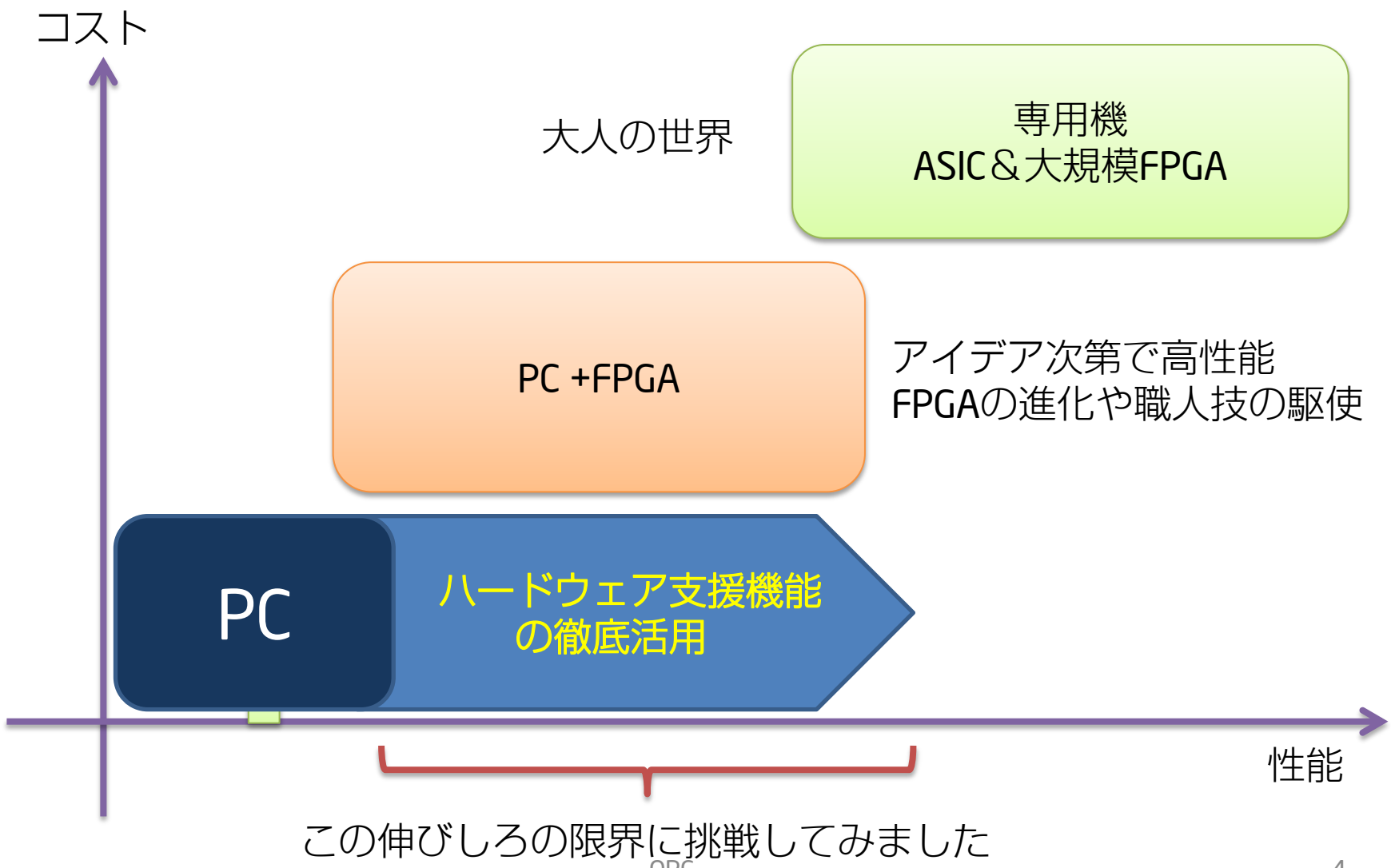


超高速ストレージキャッシュ・ストレージトラフィックの予測に基づく伝送制御・高効率伝送方式を組み合わせて解決

➔高性能なIPルータ+ 種々のアプリケーション実装が必要



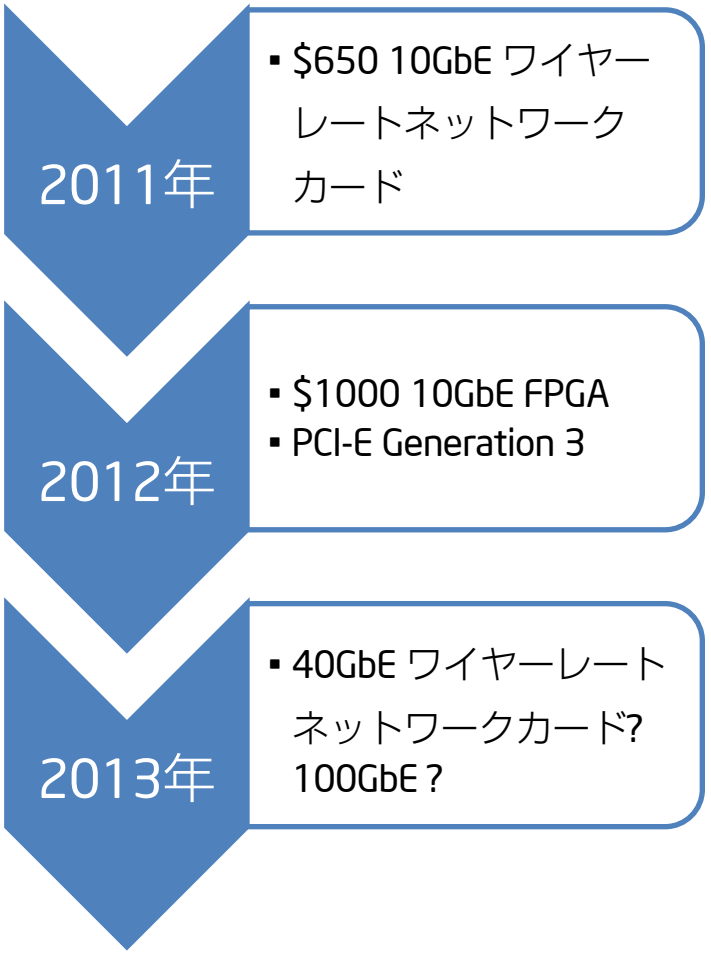
開発手法とトレードオフ



開発指針

- PC等の汎用品を利用した安価な開発(COTS)
  - 高性能な汎用品の流通とライフサイクルコストや開発コストの圧縮
- 汎用品の進化にあわせた性能強化や低価格化
- インハウスでの開発ノウハウ吸収と他への応用
  - 様々なシステムへ応用
- ノウハウの公開
  - 知の共有と技術者の育成

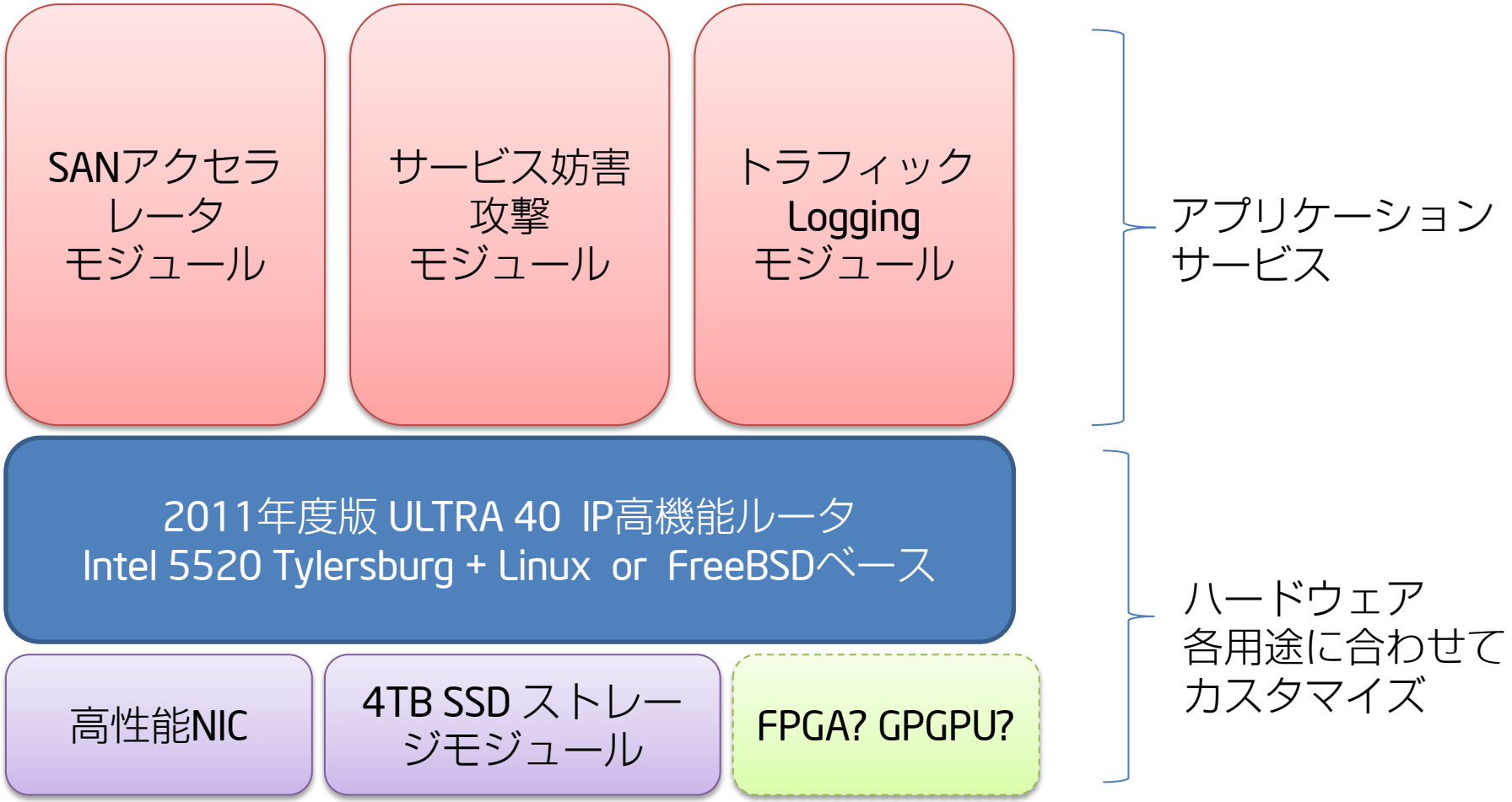
ネットワークハードウェアの動向



2つの試作高機能IPルータを出展

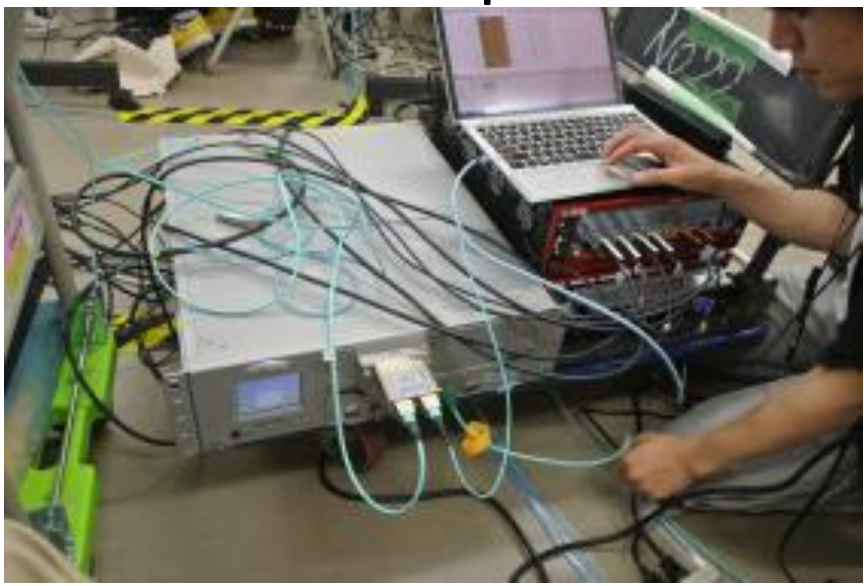
- 高機能IPルーター 「野川」
  - 実効L3 バックプレーン容量 75Gbps
    - 80Gbps (QSFP+ 40GBASE-R x 1) + 10GbE x 2
  - 超高速ストレージモジュール搭載
    - 書き込み性能に特化したSSDを16台搭載
    - トラフィックロギングモジュール
    - ストレージアクセラレータモジュール
- 高機能IPルーター 「大沢」
  - 18 x 10GBASE-R
  - 実効L3バッププレーン容量 100Gbps
  - サービス妨害攻撃モジュール搭載（評価）

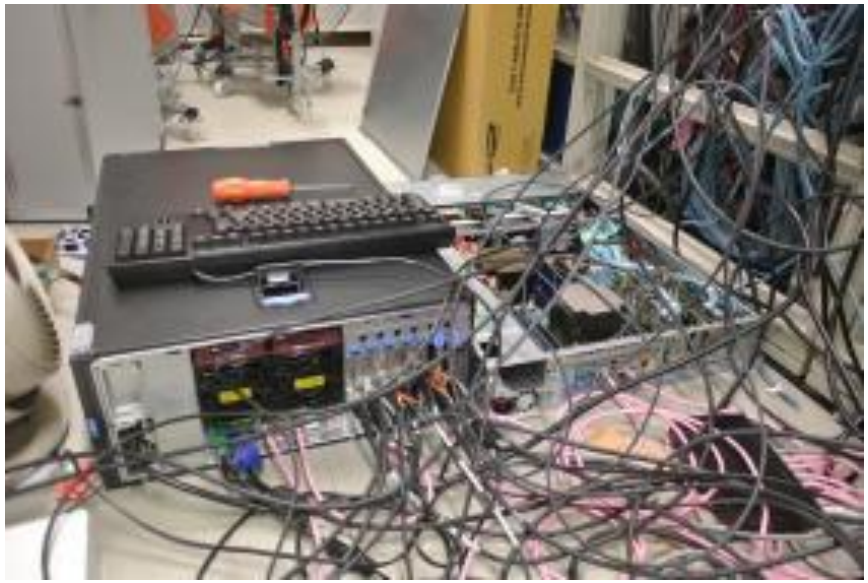
ULTRA40-アーキテクチャ



性能計測

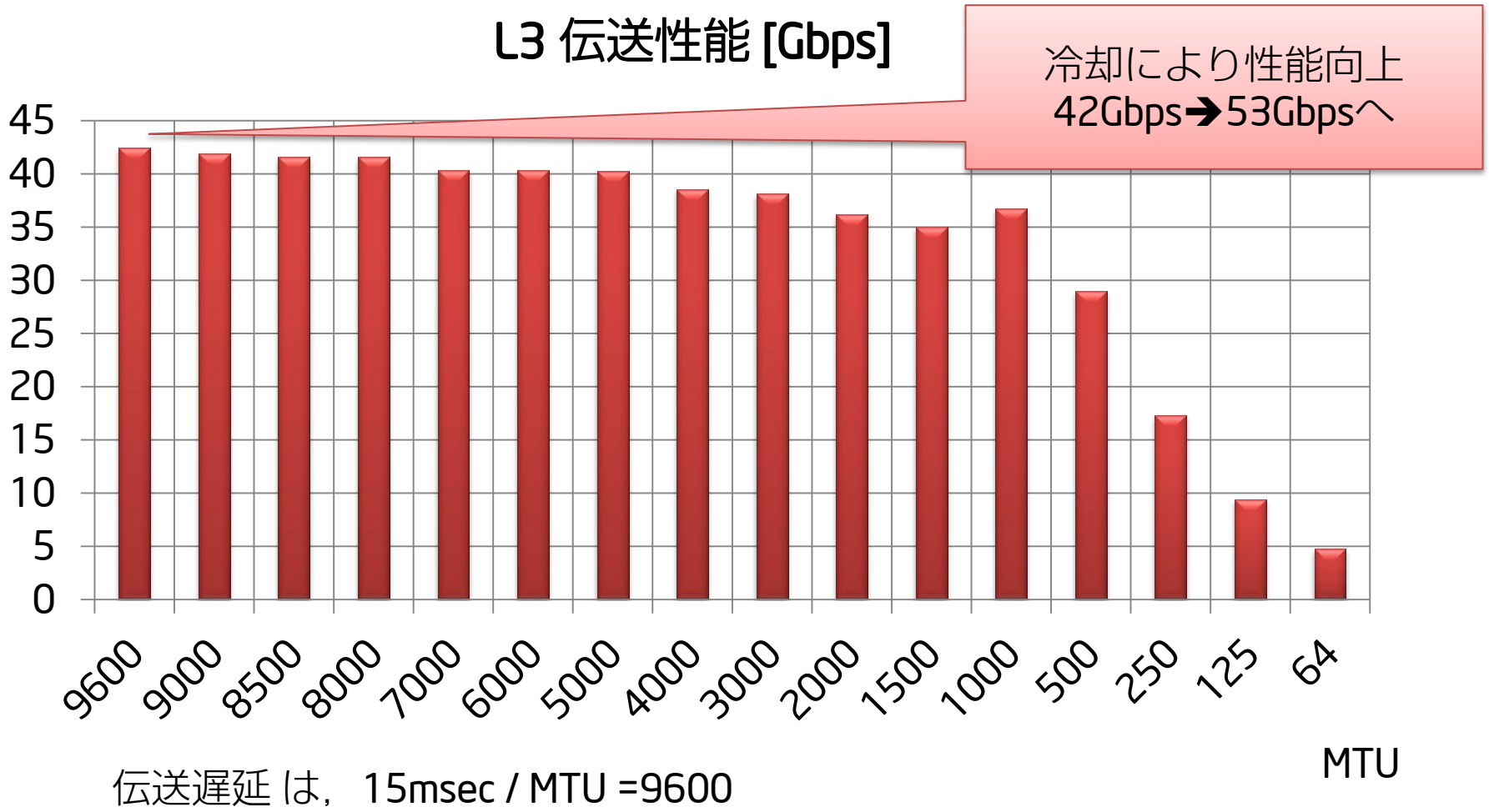
- IXIA とSpirent を接続し，160Gbps環境で性能を検証
  - IXIA様，Spirent/TOYO様に感謝いたします．





10GBASE-DA or SRにて，160Gbpsで接続

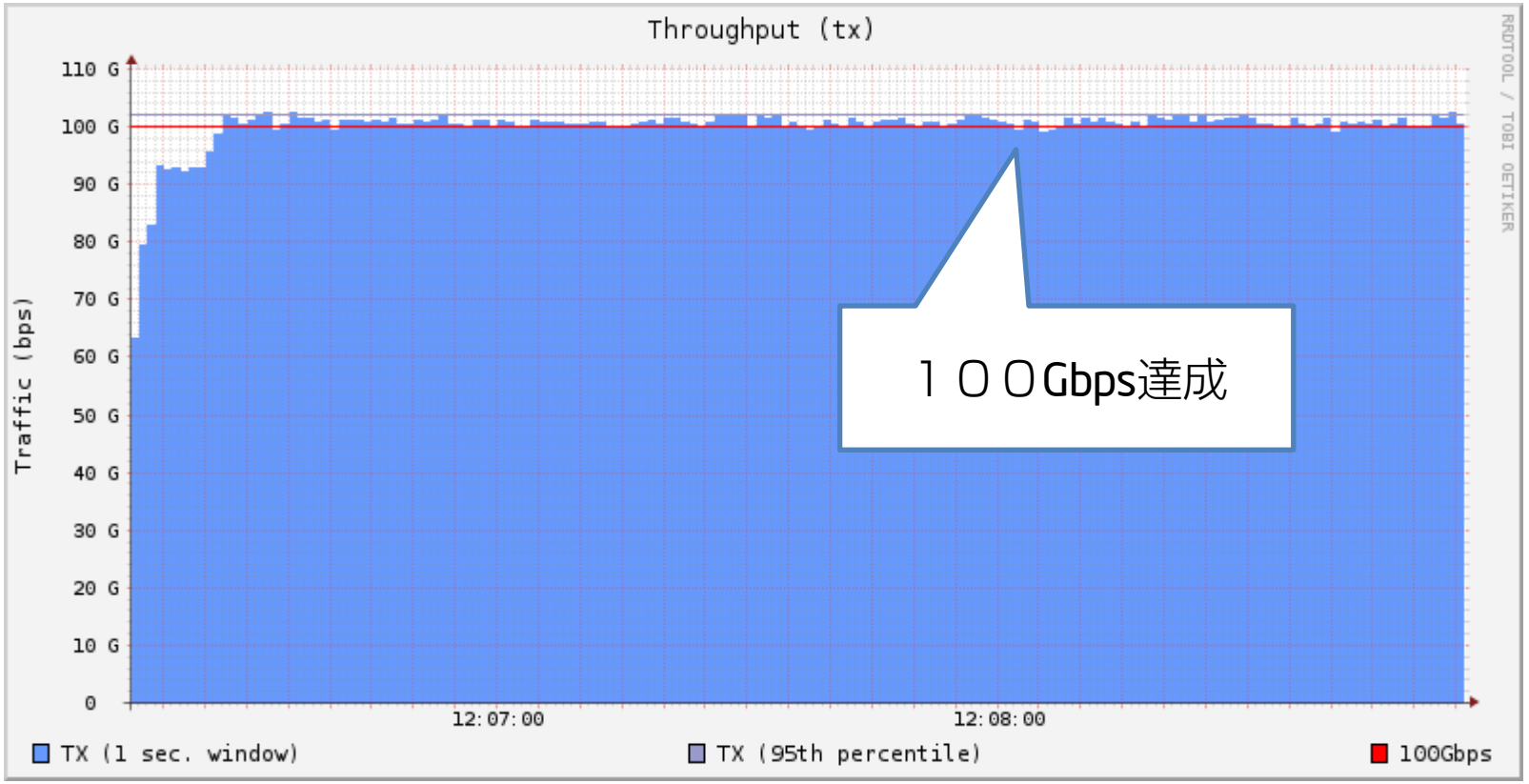
基本性能(大沢)



伝送遅延 は，15msec / MTU =9600

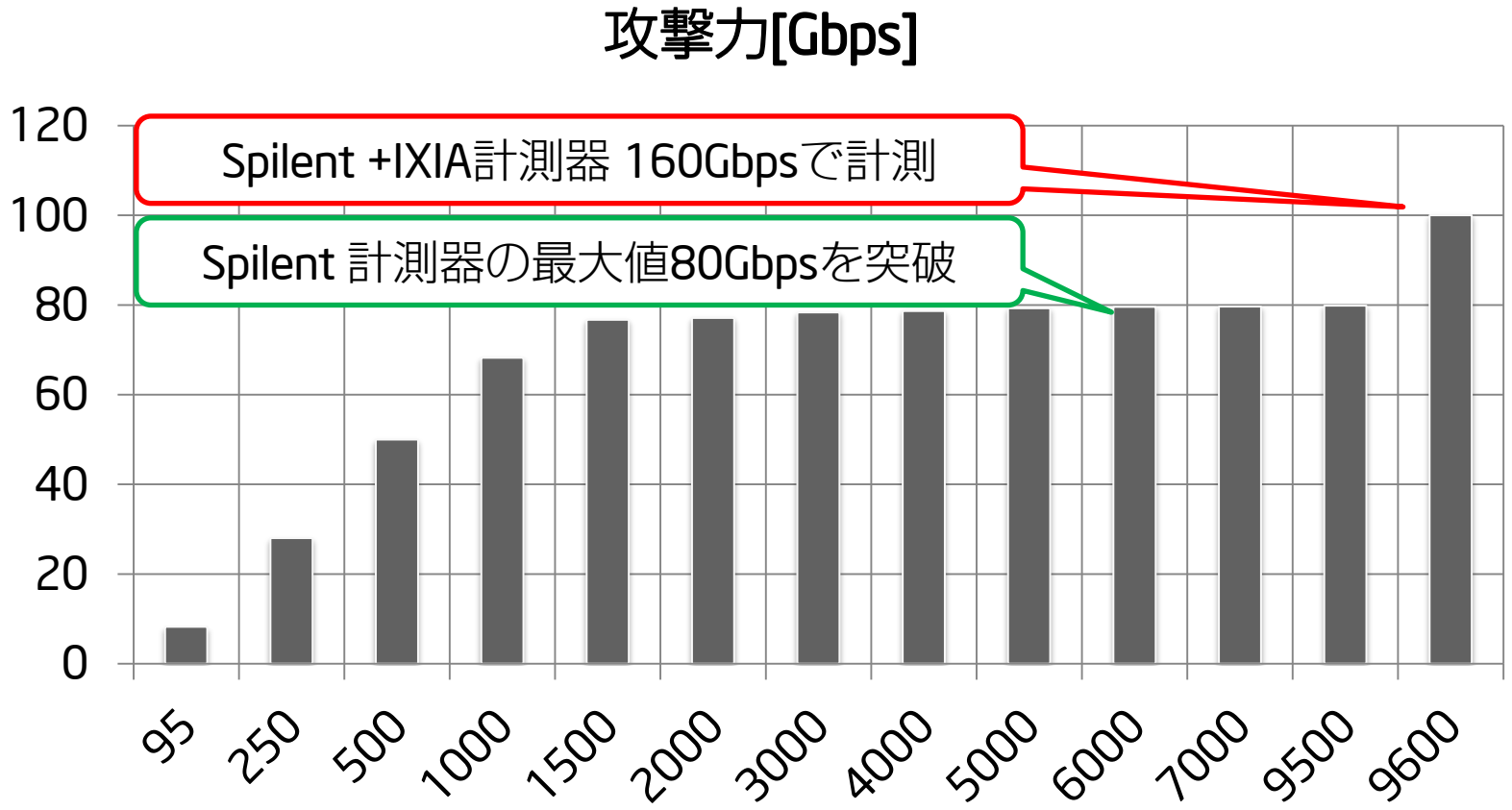
サービス妨害攻撃力（送信力）

- インターネットを破壊できるかもしれません.
  - IXIAとSpilentの組み合わせないと計測できない帯域でした.



基本性能(大沢)

- サービス妨害攻撃性能 (UDP flood)



ストレージモジュール性能

- Micron m4 SSD Firmware 0902 x 16で構成
  - Megaraid 9265 8i x 2
  - ストライピング/EXT4パラメータの最適値をSSD毎に調査
- 32Gbps/最大4TBまでの連続書き込み性能
  - 32Gbpsのトラフィックフローなら，16分間ダンプ

結論

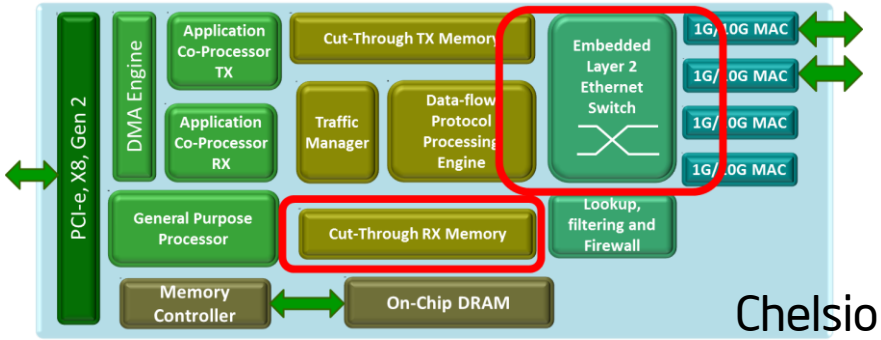
- ＰＣでも，100Gbpsは十分に扱えること
- 100万円以下で十分つかえる環境
- Competition

どのように達成したか？

- 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
- 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
  - 割り込みの集約化と待ち時間の調整
- Linuxカーネルでの割り込み・OS上の無駄な機能排
  - 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

どのように達成したか？

- 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



Chelsio T4 ASICダイアグラム

http://fumi.org/ULTRA へどうぞ

デモ

今後の課題

- さらなる道の追求と情報公開
  - SandyBridgeやゲーマーM/Bの評価
- 冷却機構の強化
  - NIC上のASICの熱により性能が低下
- 高性能化
  - 来年は，100G x n本か？
  - 受信処理性能向上，メモリのボトルネック解消

➔2013.4の天文台HPCシステム運用に向けてのブラッシュアップ