広帯域データ伝送システム ULTRAの研究開発

自然科学研究機構
国立天文台
大江将史

2014.1.27 SS研ビッグデータのためのキャンパス基盤 –「俺の部屋まで10Gを引け」と言われたら
15:50–16:40セッション

自己紹介

おおえまさふみ 検索

- 大江将史 （おおえ　まさふみ）
  http://fumi.org/
- 所属：自然科学研究機構 国立天文台
  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

[質問]
机のうえに10GbEありますか？

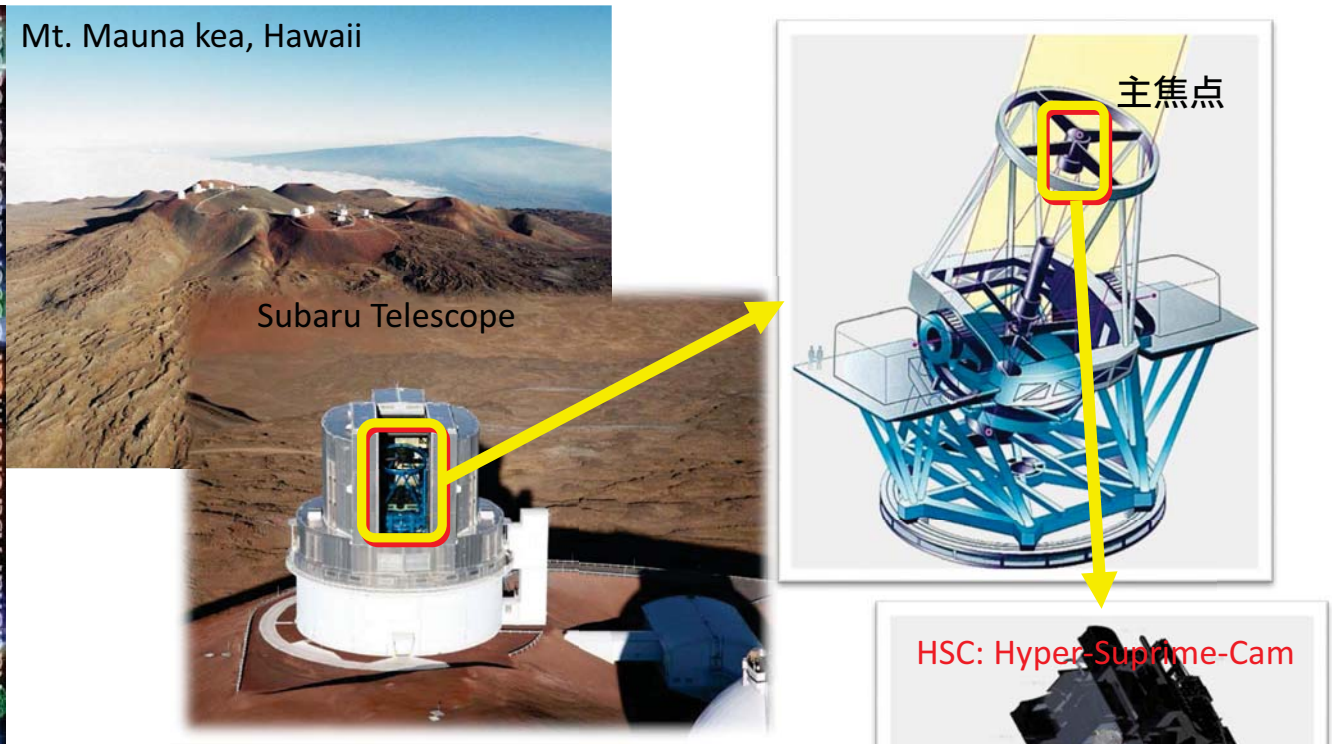
部屋まで10GbE（ユーザー向けアクセスサービスとして10GbE）きてますか？

1）きてます

2）上流のアグリゲーションスイッチなら

3）コアスイッチなら

4）ありません



アーカイブ
解析
データ公開

一晩で250GB程度の
デジタルデータを生成

天文学を支えるコンピュータネットワーク

天体望遠鏡からの観測データ
コンピュータでの
・観測データの計算機解析
・数値シミュレーション
観測装置や計算機を支えるシステム
➔ネットワークを活用
➔コンピュータ＆ネットワークによる成果

各拠点をネットワーク接続
　JGN-X/SINET-4/他
　10ギガ：水沢・大手町ＤＣ・三鷹
　1ギガ：岡山・ハワイ(2014.4-)
他もすべて接続されています.

さまざま種類のトラフィック

水沢～大手町～三鷹の場合

1）スーパーコンピュータ:アテルイ

- 特徴
  - 水沢観測所（岩手県奥州市）に設置500TFlops級のCray社のスーパーコンピューターシステム
    - 2014年度に 1PFlops級へアップグレード

1）スーパーコンピュータ:アテルイ

- 計算ジョブ（最長8時間）の間隔でデータが出力
  - ジョブ完了➔水沢からデータを東京へ取り出す
  - ジョブ継続➔再度ジョブ投入

  ➔8時間単位で，ネットワークに負荷がかかる可能性

HPC計算ノード群
(水沢)

IPネット
ワーク

ストレージ
ノード群(三鷹)

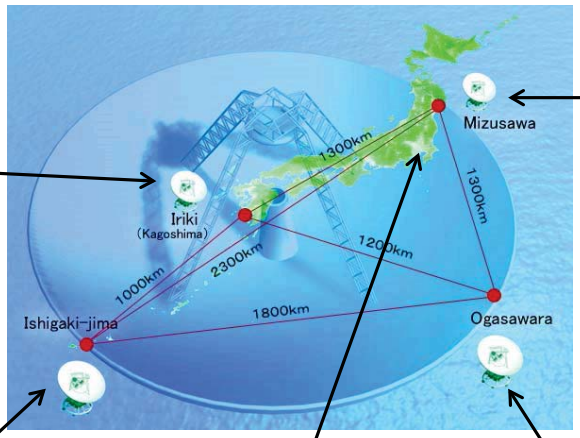
専用計算ノード群
(三鷹)

汎用計算サーバ群
(三鷹)

2）VERA: VLBI Exploration of Radio Astrometry

VERA is a VLBI array to explore the 3-D structure of the Milky Way Galaxy

IRIKI(入来), KAGOSHIMA



MIZUSAWA(水沢), IWATE

ISHIGAKIJIMA(石垣島), OKINAWA

Correlation center

MITAKA(三鷹), TOKYO

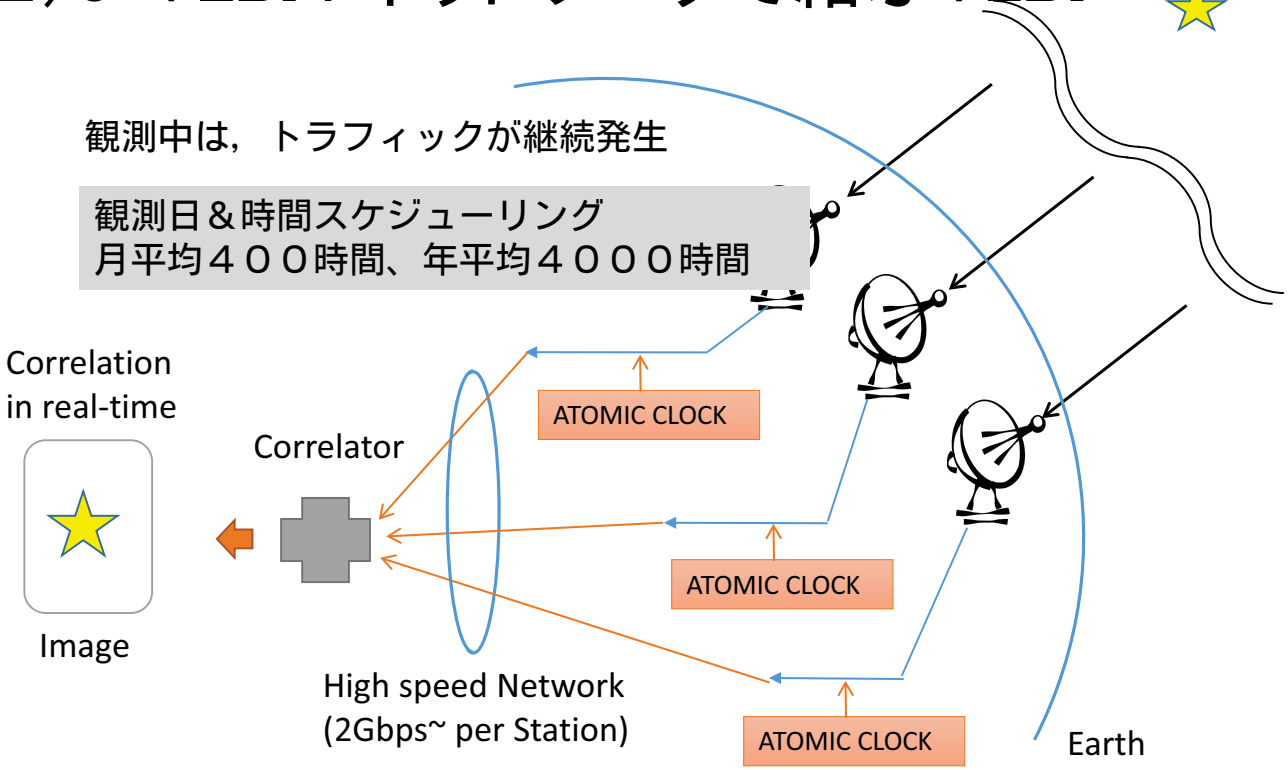
OGASAWARA(小笠原), TOKYO

望遠鏡（山口・茨城・他）

2）e-VLBI：ネットワークで結ぶVLBI

観測中は，トラフィックが継続発生

観測日＆時間スケジューリング
月平均４００時間、年平均４０００時間



そのほか

3）クラウドシステム

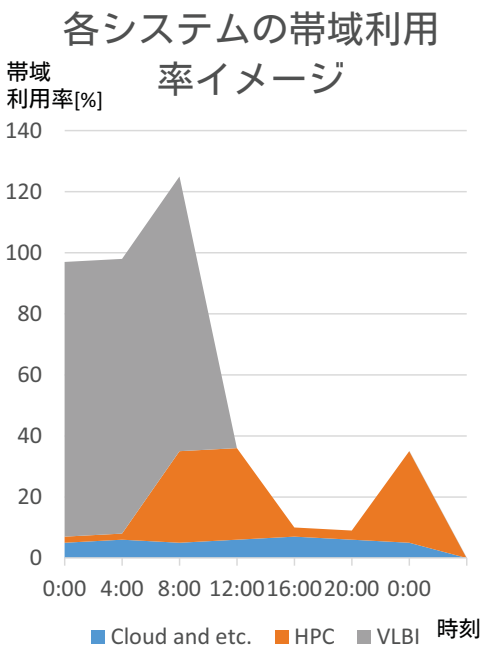
- プライベートクラウドサービスを4拠点で運用
  - 「実機より速い」が合言葉
  - 三鷹地区・大手町地区・水沢地区・岡山地区に分散したクラウドシステム
  - iSCSIネットワーク・VMノード

4）コンテンツ配信

- デジタル4次元シアター(4D2U)のコンテンツ提供
  - スパコンや観測成果に基づく科学コンテンツの配信
- アウトリーチ：観測所と学校を結んで最先端の科学にふれる
  - HDビデオ双方向遠隔授業（1から多地点）

さまざまな属性を持つトラフィックがWANを流れる

- スパコン
  - 水沢の計算ノードからの結果出力を，三鷹の恒久ストレージへ効率よく伝送
    - ノンリアルタイム・利用者の利用傾向に基づく帯域の占有予測
    - 伝送中は高効率化により帯域を占有・ロスは許容されない．
- VLBI
  - 水沢から三鷹へ観測データをバーストトラフィックで伝送
    - スケジュールされた観測時間に連動した帯域確保
    - パケットロスには寛容・通信としてのプライオリティは低い扱い
- クラウド・コンテンツ配信
  - 帯域は，クラウドのマイグレーション，ストレージトラフィック，コンテンツ配信などに強く依存
    - 帯域の変動幅が大きい
    - パケットロスに非寛容．



ULTRA計画

WAN-LAN間のギャップ解消

ULTRA計画(2012～)の背景

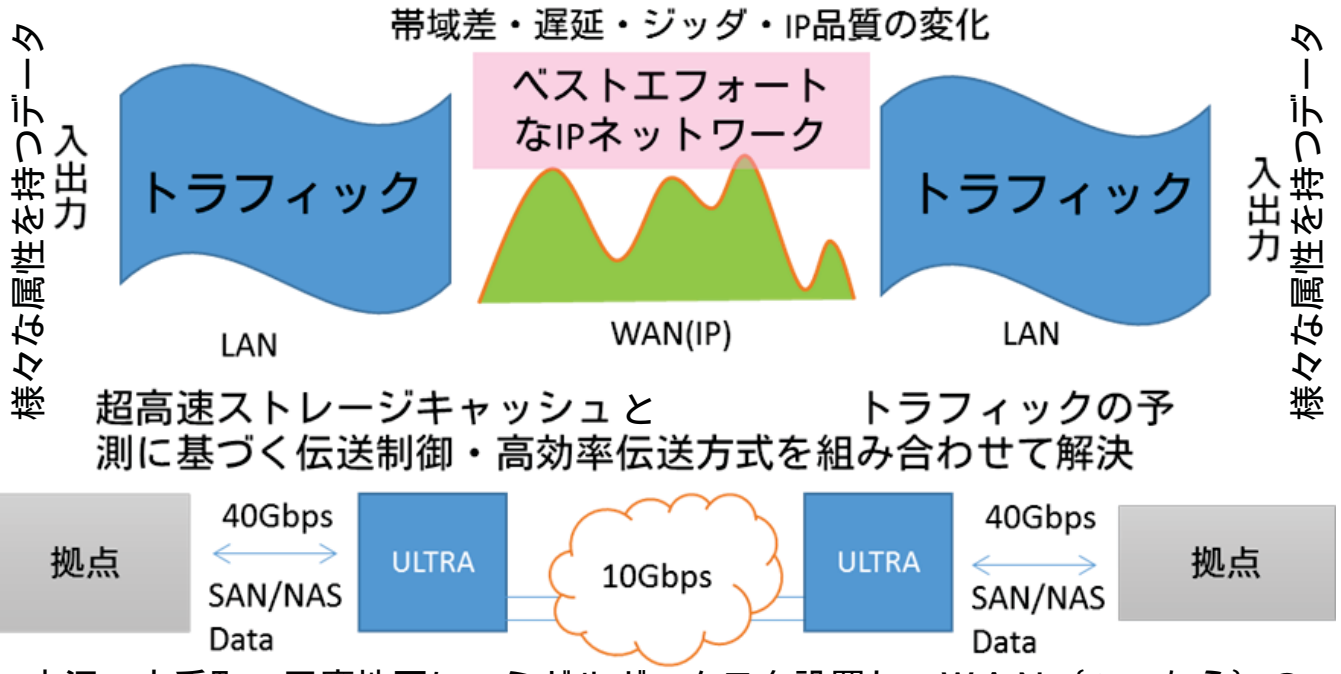
- アプリケーションが必要とするネットワークの高性能化・広帯域化・トラフィック個性
  - スパコン・VLBI・クラウド・映像中継等々
- WAN広帯域化とLANさらなる広帯域化
  - WAN-LANの帯域・性能ギャップの存在と絶対処理量の増大

➔地理的に分散する情報システムとIPネットワークを連携させ，かつ，増大する処理量にこたえられる仕組みが自然科学の発展には必要不可欠

アプリケーション要求要件からの要素技術のブレークダウン

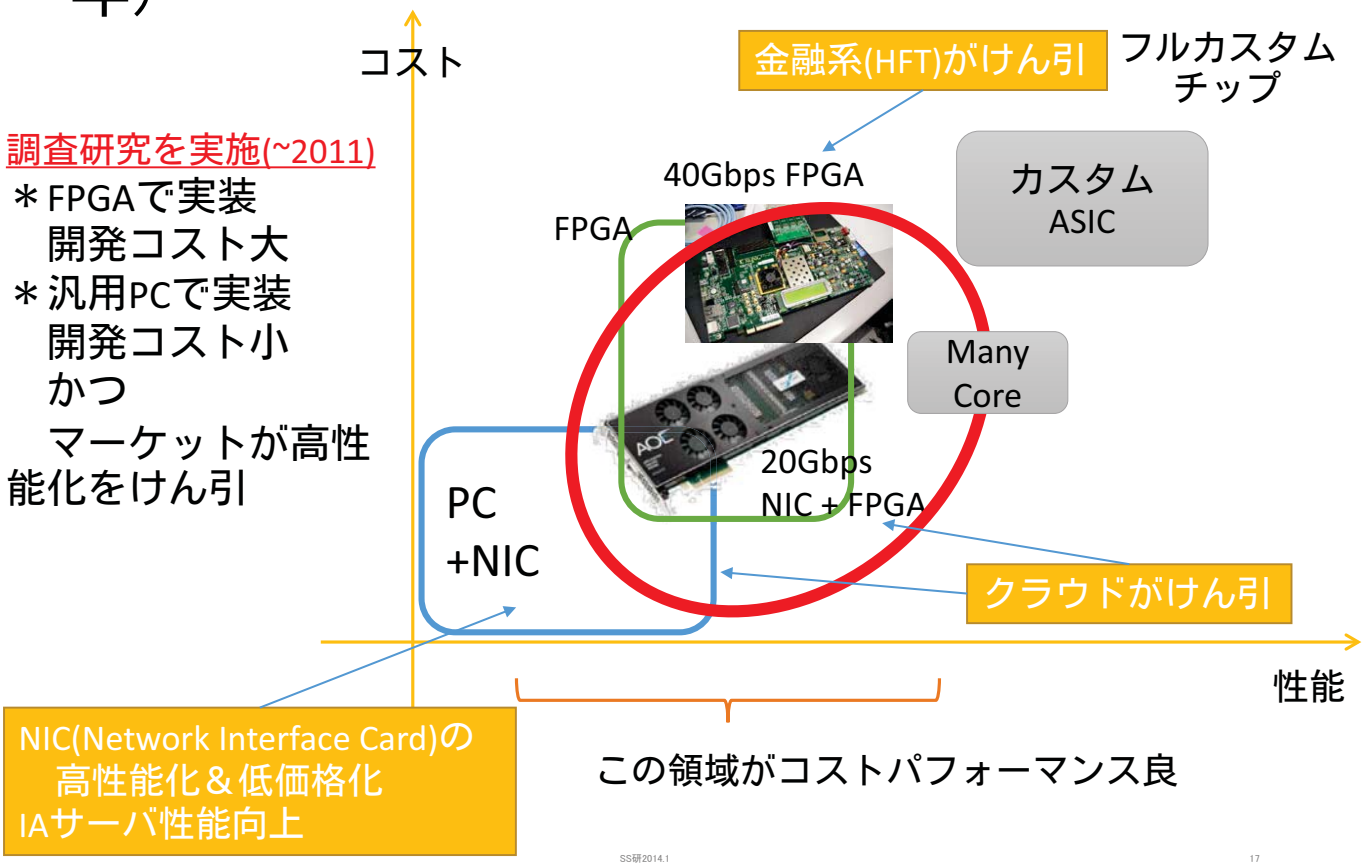
➔ULTRA計画は，高性能・高機能（＝問題解決のアイデア）を安価に実装することを目標に開始

ULTRA計画のアイデア
ミドルボックスによるデータ伝送の効率化



水沢・大手町・三鷹地区に，ミドルボックスを設置し，ＷＡＮ｛へ，から｝のトラフィックをＷＡＮのトラフィック・ウェザー情報に基づき伝送制御

コストの観点から見る実装方式の検討(2011年)



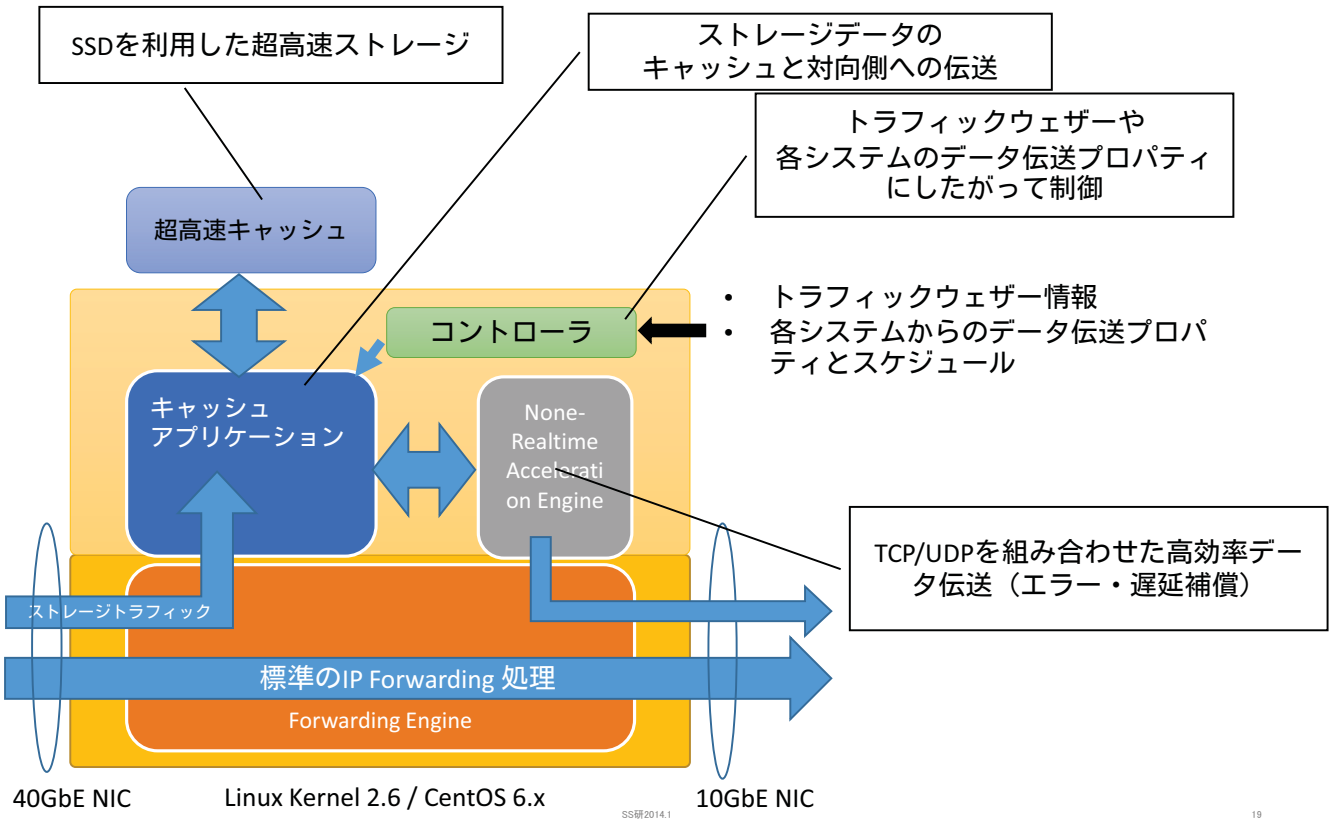
調査研究を実施(~2011)
＊FPGAで実装
　開発コスト大
＊汎用PCで実装
　開発コスト小
　かつ
　マーケットが高性能化をけん引

NIC(Network Interface Card)の
　高性能化＆低価格化
IAサーバ性能向上

開発実装：
100Gbps越えのIP通信処理

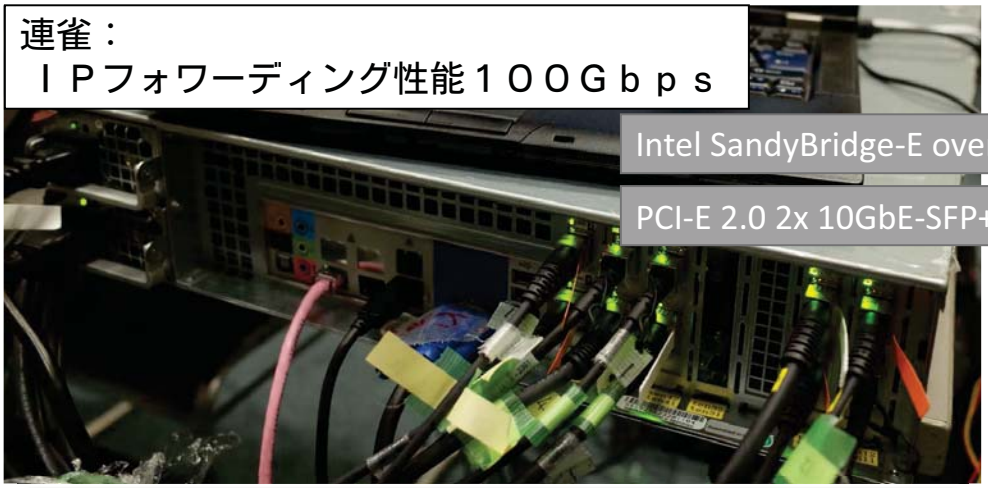
- 方針
  - 事前検証よりIAサーバ・汎用製品の組み合わせで十分な性能を叩き出せるという目算
    - コモディティ化した製品の応用によるコストパフォーマンスの追及
  - 汎用故に成果物のほかシステム・分野への転用による効率化
- 内容
  - 2つのトラックに分けて開発実装
    - キャッシュサーバ機能「野川」：SSD超高速ストレージ開発
    - IP通信機能「大沢」「連雀」：100GbE　L3ルーターの開発
    - （継続）FPGAによる方法も検討
  - 最終的には，これらを統合する

ULTRAの機能ブロック構成



「連雀(れんじゃく)」・「連雀＋」:

連雀：
ＩＰフォワーディング性能１００Ｇｂｐｓ



Intel SandyBridge-E overclock

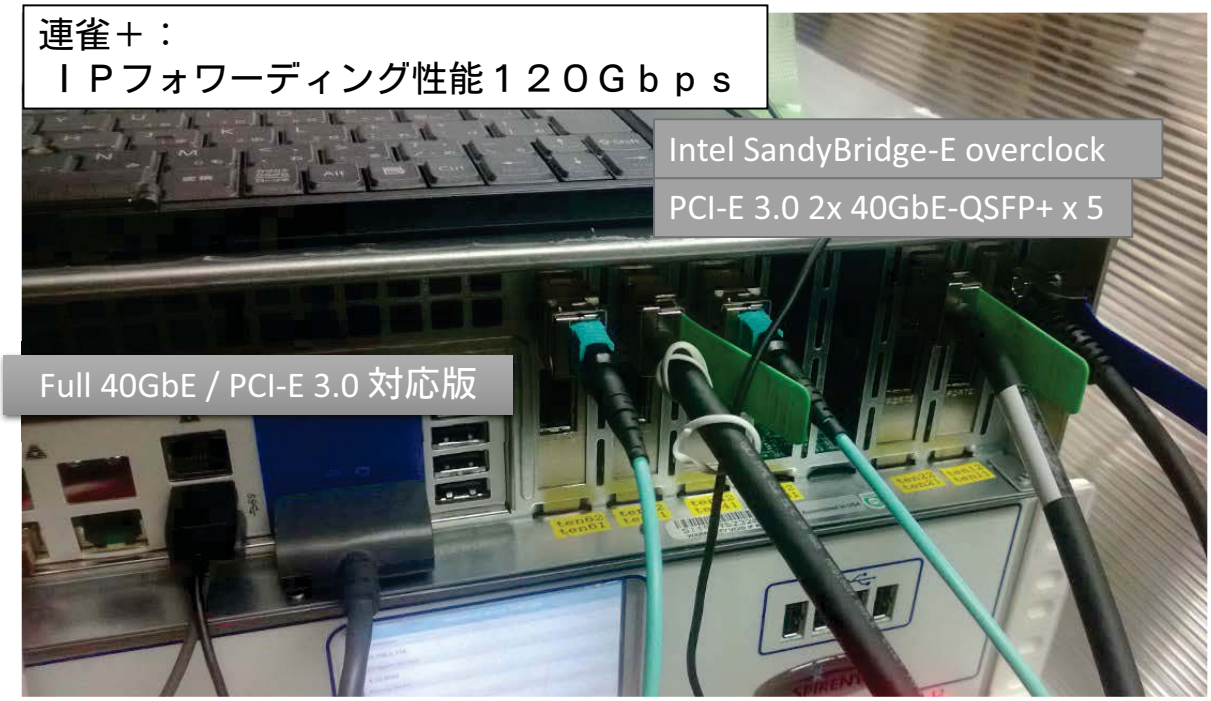
PCI-E 2.0 2x 10GbE-SFP+ x 10 (最大12port)

Interop2013 オープンルーターコンペティション(ORC)
富士通賞受賞

国立天文台が天文データ処理用のPCサーバ / ルーター プラットフォームとして開発
　Linux OSを基に低遅延・広帯域処理能力を目標に設計・開発

「連雀＋」:
40GbE対応 / 広帯域・低遅延化

連雀＋：
ＩＰフォワーディング性能１２０Ｇｂｐｓ



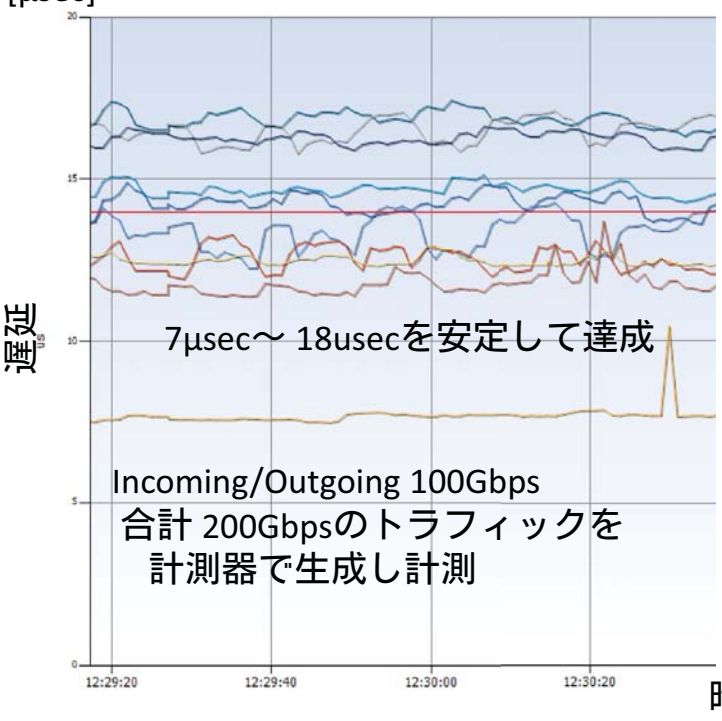
Intel SandyBridge-E overclock

PCI-E 3.0 2x 40GbE-QSFP+ x 5

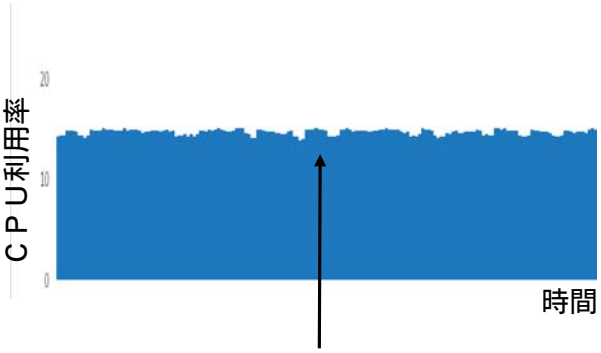
Full 40GbE / PCI-E 3.0 対応版

「連雀」の性能
低遅延＆広帯域の両立

フルメッシュＩＰフォワーディング性能を計測器で長時間検証



広帯域＆低遅延を両立



NICオフロードハードウェアを徹底活用し200Gbps相当のトラフィック処理下で，15%のCPUリソース消費

「汎用PCで 100Gbps越え」のコスト

- 連雀
  - 物品費＜80万円

    ＋10%の性能，機能，信頼性追及をしなければ，コストは，40万円以下へ．
- 野川
  - 物品費＜150万円（筐体は，40万円程度，他はSSD費用）

    （部品の選定に要した費用は含まず）

➔クラウドシステムへの応用（高性能仮想サーバと高速ストレージサーバへ）

➔汎用PCの性能向上ノウハウのインハウス蓄積

ULTRAの進化でかわる PCサーバの性能向上

低遅延性能 FGI

2013年10月 「連雀+」
220Gbps IPフォワーディング性能
遅延 < 10µsec

2013年6月 「連雀」
200Gbps IPフォワーディング性能
遅延 < 17µsec

2012年 「野川」 & 「大沢」
4GB/sec 読み書き可能な高速ストレージ
100Gbps コンテンツ送信性能
45Gbps IPフォワーディング性能
18x 10Gbase-R or 2x 40Gbase-R I/F 搭載


2013 連雀+
240Gbps / <10µsec


2013 連雀
200Gbps / <17µsec


2012 ULTRA40 大沢 45Gbps / 12msec
ULTRA40 野川 4GByte/sec SSD Storage

2011 PCルータ
20Gbps ?msec

広帯域

ULTRAの進化でかわる PCサーバの性能向上

低遅延性能 FGPAなどハードウェアの領域


2013 連雀+
240Gbps / <10µsec

汎用PCで踏み込める領域が拡大


2013 連雀
200Gbps / <17µsec


2012 ULTRA40 大沢 45Gbps / 12msec
ULTRA40 野川 4GByte/sec SSD Storage

2011 PCルータ
20Gbps ?msec

広帯域

PCサーバの性能向上は今後も続く，手段を問わず研究開発を継続

2011年 ?? Intel Core + PCI-E2.0 1x10GbE NIC

- なんとか10Gbpsを絞り出せるレベル

2012年「大沢」「野川」（第1世代）Intel Nehalem + PCI-E2.0 2x10GbE NIC + Offload

- コンテンツ送信力は，100Gbps

2013年「連雀」（第2世代）Intel SandyBridge-E + PCI-E2.0 2x10GbE NIC + Offload
「連雀＋」

- その処理力は，200Gbpsへ向上

2014年（第3世代）Intel Haswell + PCI-E3.0 NIC Full 40GbE NIC + Offload / **ULLtraDIMM**
野川系・連雀系の統合

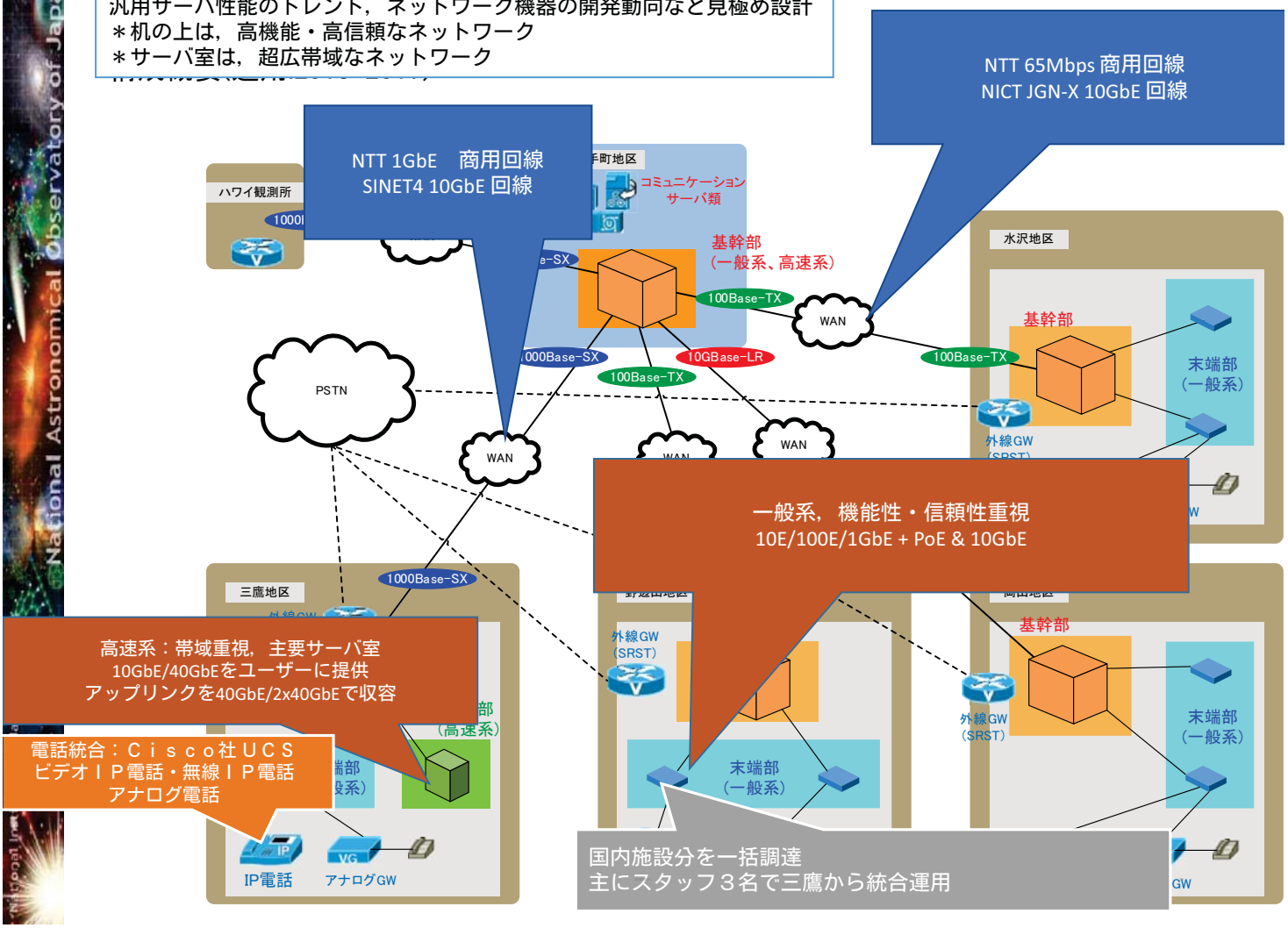
- その処理力は，400Gbpsへ？

- 誰もが100Gbps～200Gbpsを扱える時代

割り込みモデル
or
ポーリングモデル

広帯域通信を支える 国立天文台の基盤ネットワーク

汎用サーバ性能のトレンド，ネットワーク機器の開発動向など見極め設計
＊机の上は，高機能・高信頼なネットワーク
＊サーバ室は，超広帯域なネットワーク



低価格が進む要因
L3／L2スイッチング機器のトレンド

- 10ギガビットイーサネット＝1万円の世界はすぐそこ.
- L3／L2を支えるハードウェア技術
  - 専用LSI
    - 目標性能・クロックをもとに回路設計
      - コストが高く，自社専用or汎用に二極化
    - 汎用は，あらゆるスイッチング機器ベンダー向けに設計，量産効果
      - ブロードコム(Dune) ・インテル(Fulcrum)など
      - シャーシ向け，多機能，超低遅延，低価格など品種多数
    - カスタムは，自社（機能実装）向けにデザイン
      - シスコ・メラノックスなど
  - 専用ASIC
    - ASICチップに機能（論理合成）を搭載
      - LSIより安価だが，ASICのデザインクロックの制約やデザイン上の制約などから，性能を出すのがむつかしい

L3／L2スイッチング機器の構成と今後

- 事業者・製品性能に応じて，LSI／ASICを活用
  - 汎用LSIをフル活用
  - 汎用LSIを部分的に使い，独自実装のためASIC，FPGA併用
  - カスタムLSIをフル活用

  ➔Cisco ／ Mellanox ／ ARISTA ／ Extreame 等，各社ごとにデザインが異なる．

  *）LSIフル活用であっても，ASICやFPGAを支援のために使う場合もあります．
- 次世代
  - 発熱問題，リソグラフィ・プロセスの進化の必要性
  - 差別化，汎用LSI上でのプログラマブル領域実装
  - トラフィックの多段処理，メニーコア化

まとめ

- 汎用サーバ分野
  - 10ギガ級性能は超越，今は，40，100ギガです．
- ネットワーク分野
  - 広帯域の低価格が進む：10GbE＝1万円は近い．
    - LSI汎用化による低価格が進行中だが，帯域向上には，リソグラフィが課題
  - 汎用サーバを生かし切ることやその導入促進には，基盤ネットワークの整備が重要
    - 利用者は，もっとも価格競争が働き進化が速い汎用サーバ部分を用意すればよい．

まとめ:
机に10G？，まだいらいないかな

- 高性能化・広帯域化・低価格化・仮想化➔システムの集約化と集中運用のトレンド
  - 様々なシステムがハウジング内で広帯域ネットワーク上に密結合
    - 仮想サーバ，ストレージサーバ，セキュリティプラットフォーム，WAN等々
- 10・40Gが活用できるのはハウジングの中，机じゃない.
  - 組織内の計算機を垣根を越えて，横断的なリソースの運用と効率化が可能な基盤ネットワークが重要

おしらせ

＊国立天文台三鷹キャンパスでは，毎月2回公開天体望遠鏡を使った観望会を開催中！

詳しくは国立天文台ホームページをご覧ください.

ありがとうございました

口径30m次世代超大型望遠鏡(TMT) 始動
http://tmt.mtk.nao.ac.jp/

TMT 完成予想図 (c) 2010 Thirty Meter Telescope