シ ス テ ム 技 術 分 科 会 選 出
システム技術分科会 2013 年度第 1 回会合 より

# 学内情報通信基盤の耐震対策の効果

曽根 秀昭
(東北大学 サイバーサイエンスセンター)

# 学内情報通信基盤の耐震対策の効果

曽根 秀昭†

†東北大学サイバーサイエンスセンター 〒980-8578 仙台市青葉区荒巻字青葉 6-3

E-mail: †sone @ isc.tohoku.ac.jp

**あらまし** 東北地方太平洋沖地震による東北大学の施設への被害も甚大であったが，様々な耐震対策はよく効果したと考えられ，学内情報基盤への被害も比較的軽微であった。情報通信基盤の設備を維持できるための耐震補強など防災対策が，災害時のサービス継続を保つ減災対策，あるいは復旧・復興の重要な鍵になる。筆者の体験に基づく大学の情報基盤の被災状況及び耐震対策とその効果を報告する。

**キーワード** 情報通信基盤，東日本大震災，震災被害，防災，耐震対策

# 学内情報通信基盤の耐震対策の効果

東北大学
サイバーサイエンスセンター

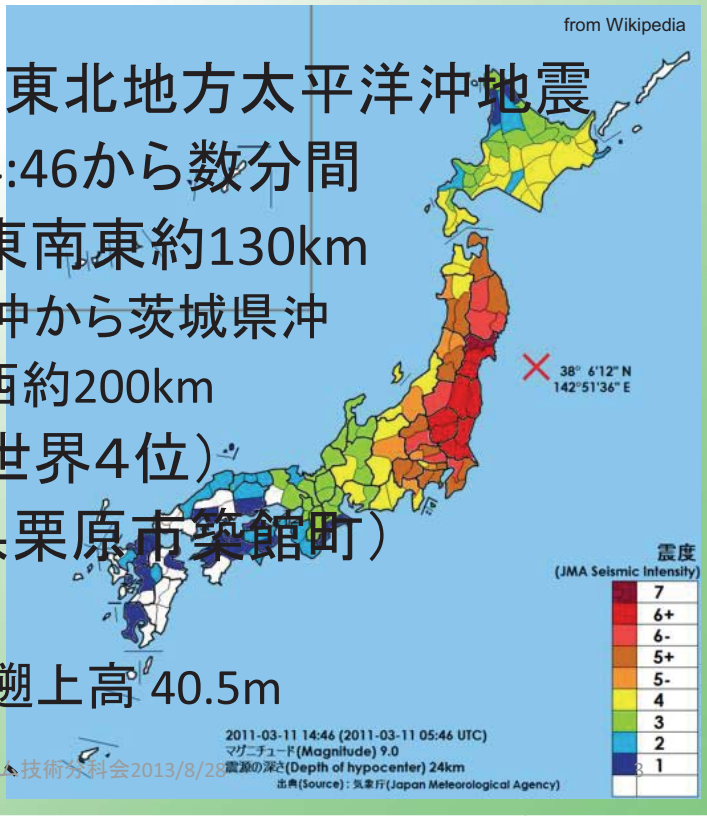
曽根秀昭



# 東北地方太平洋沖地震と東北大学の被災状況

# 地震の概要

- 平成23年（2011年）東北地方太平洋沖地震
- 発生：2011/3/11 14:46から数分間
- 震源：牡鹿半島の東南東約130km
  - 震源域は、岩手県沖から茨城県沖
  - 南北約500km、東西約200km
- Mw 9.0（観測史上世界4位）
- 最大震度7（宮城県栗原市築館町）
- 大津波が発生
  - 波高約 10m、最大遡上高 40.5m
  - 内陸 6km まで浸水






# 被災の特徴・防災の備え

- 被災がきわめて広域　ー東日本全域
  - 被災地復旧対応もきわめて広域・膨大
- 仙台市では、地震対策の効果
  - 宮城県沖地震（37年周期へあと4年）
  - 災害に強いまち：インフラ、建物の耐震補強、家具の固定、避難訓練、非常時備蓄
    - 地震による直接の死者は数十名と言われる
  - 交通，電力，水道，ガスなどインフラの迅速な復旧
    - 十分な食料・物資の供給　（物流と労働力がネック？）
- 地震よりも、津波による沿岸部の被害が甚大
  - 沿岸部では、津波による社会基盤の壊滅的喪失
  - 石油の流通の被害から大きな影響
    - 物流，交通の復旧や，発電機，復旧作業への支障


# 東北大学の被害概況

元気・前向き Powerful Positive Tohoku University 東北大学

- 建物・施設の被災
  - 安全判定・白　512棟（90%）
  - 危険判定・赤　川内2棟、雨宮3棟、青葉山8棟
  - キャンパスは内陸部・丘陵地にあり、津波浸水被害はなし
  - 沿岸部施設の流失、壊滅
    - 女川町・複合生態フィールドセンター
    - 七ヶ浜町・ヨット艇庫、名取市・ボート艇庫・合宿所
- ライフラインの復旧
  - 電気（3・13）4月4日，水道（無）4月13日，ガス（4・13）4月26日
    - 全学復旧完了日(カッコ内：サイバーサイエンスセンター)
  - バス 3月14日，地下鉄 3月13日（部分的再開）
- 学生・教職員の被害
    - 全員の安否確認は3月30日完了
  - 学生：死亡　3名、負傷14名
    - 死亡した学生：　学部学生2名、入学予定者1名：自宅等で被災
  - 教職員：死亡者・負傷者ともゼロ

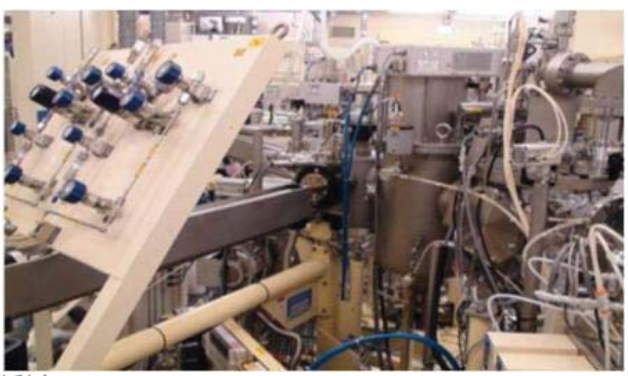


# 青葉山キャンパス

Forward 未来をともに 東北大学 Tohoku Univ

工学研究科　人間・環境系実験研究棟

電子光理学研究センター（富沢）粒子加速装置



工学研究科附属マイクロ・ナノマシニング研究教育センター

# 沿岸部施設

女川町：農学研究科附属複合生態フィールド教育研究センター


七ヶ浜ヨット艇庫

名取ボート艇庫　合宿所



# 東北大学の全学情報基盤の被災

# 全学共通情報基盤への被害

- 全学共通情報基盤関係の建物は，業務に支障なし
  - サイバーサイエンスセンター，情報推進課（本部棟）
    - 応急危険度判定：白 （外壁・内壁にクラック）
    - ラックの倒壊などなし
- 全学共通情報基盤の機器・サーバに損傷なし
  - サーバ，パソコン等情報機器及びネットワーク機器
  - キャンパス間の光ファイバ（自営）も無事
  - 一部の部局に設置した機器は建物被災のため移設
    - 教育情報基盤センターの機器はラック倒壊で大きな損害
- 停電
  - 本センターや本部棟： 46時間（3日目14時まで）
  - 病院のある星陵キャンパス： 自家発電供給
  - 各部局： その後の数日間で順次復電



# 地震時のネットワークの停止

- 3月11日(金) 地震発生（14：46から数分間）
  - 14:48 サイバーサイエンスセンター停電
    - 無停電電源装置 (UPS（4系統）) による給電に切替
  - 10分後サーバ等自動シャットダウン
    - TAINSメール、リモートアクセスサーバ等
    - ウェブサーバ、メールサーバ、DNSサーバ等が全停止
  - 28分後SINET（上流接続）停止
  - 2時間30分後TOPIC（地域ネットワーク） 停止
  - 2時間50分後TAINS 幹線停止

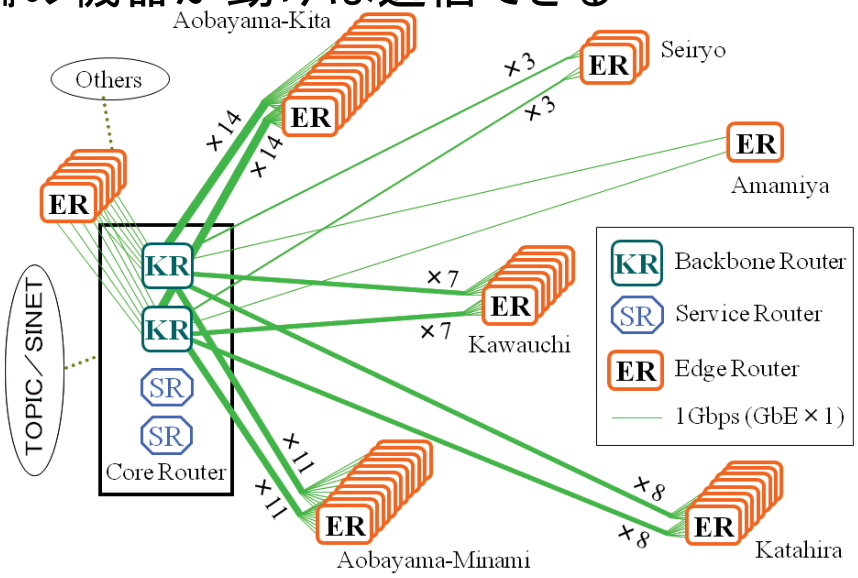
# 地震後のネットワークの運用

- 当日
  - 屋外避難後に，原則としてセンター職員は帰宅
  - ネットワーク機器室等の簡単な点検
- 3月12日(土)
  - (センター)災害対策本部で被害状況点検
  - (大学)災害対策本部へ報告
    - 建物の使用に問題なし、機器の損傷なし、復電すれば運用可能
- 3月13日(日)
  - 14:20　サイバーサイエンスセンターへの復電(午前に連絡)
  - 基幹ネットワーク、サーバ群、TOPICおよび SINET再開
- 3月14日(月)以降
  - 部局ネットワークの復旧支援
  - 被災部局(立ち入り禁止等)への支援
    - ネットワークやサーバーなどの情報基盤機能を代行など



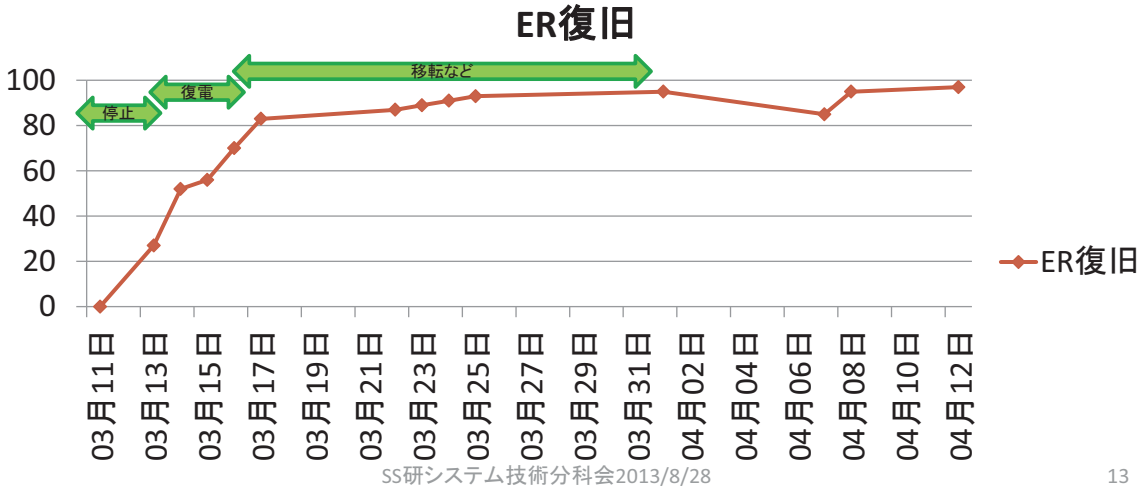
# TAINSのキャンパス間接続

- センター～各建物をファイバで直結する構成
  - 中継(キャンパス拠点)の停電の影響を回避
  - 両端の機器が動けば通信できる



# 部局ネットワークの復旧状況

- 3/11(金) 地震発生（14:46），停電
- 3/13(日) 復電，基幹ネットワーク復旧
- 3/13(日) エッジルータ 27%(13/48)復旧
- 4/1(金)  ER 95%(46/48)復旧
- 4/7(木) 余震発生　停電のためER5台が一時停止，すぐに復旧
- 4/12(火) ER 97%(47/48)復旧






# 部局ネットワークへの応急支援

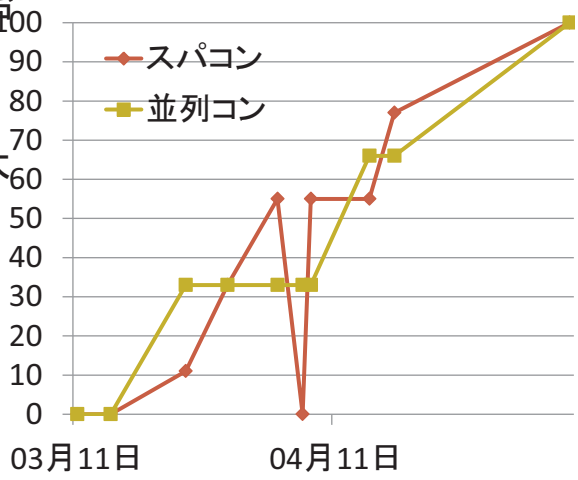
- ハウジングサービス
  - 部局サーバをサイバーサイエンスセンター内にて仮復旧
- 部局ドメイン代行サービス
  - 部局メール転送，ウェブホスティング，DNSホスティング
- ネットワークインフラ
  - ネットワーク障害復旧作業
  - 無線LANシステム と有線ネットワークを緊急増設
- 電力モニタリングシステムの調達の支援
  - 電力ピーク抑制への協力と，電力使用制限への対応
  - ウェブホスティングサービス上の構築により工期を短縮

# スパコンの停止状況

- 3/11(金)　地震の際の停電により全システム停止
- 3/13(日)　復電後のチェックで，ほぼ被害なし
  - （メモリーモジュールの不良など）
- 3/15(火)　ログインサーバ、ファイルサーバのみ運用開始

- 3/24(木)　演算用サーバ運用開始
  - スパコン11%稼働(2/18ノード),
    並列コン33%稼働(2/6ノード)
  - 利用者の回復遅れと節電要請のため，縮退運転から徐々に拡大
  - 5/9 (金) 100%稼働

- 4月の利用状況
  - 実利用者数　439人(2011.04)
    - 昨年同期比82%(92人減)
  - 利用CPU時間　昨年比約66%






# ホームページ・広報

- 2日目にウェブサーバ再開、3日目に更新開始
  - 停止（無応答）中は、さまざまな風評
    - 無停止の対策、または非常時代替対策が必要だった
- 緊急連絡ページを追加して、トップページに
  - 文字ベースで職員・学生向け情報連絡を確実に伝達
    - 対策本部のアナウンスなど、毎日更新
  - 携帯電話からの閲覧は、多くはない（が、数%）
    - シンプルなページを見て、新たな誤解も
    - 情報不足の声（リアルタイム検索）→掲載追加
- 6月から復興広報キャンペーン
  - 風評被害の払拭のため
  - スローガン、動画、説明資料
  - twitter, Youtube, facebook

元気・前向き
Powerful Positive Tohoku University
東北大学

# 近隣他大学の様子

- 建物に甚大な被害があったところ
  - ネットワーク，情報基盤も長期の停止
- 電力，通信回線が長期途絶したところ
  - 道路や橋梁の喪失のため
- 情報基盤の復旧への対応が難しかったところ
  - 平時からの運用体制が必要
- 分散キャンパス間の情報通信の確保が必須
- 災害への備えや非常時対応が良かったところ
  - 発電機，回線二重化，学外サーバなど
    - （発動発電機の騒音問題）



# 情報基盤の災害対策の効果

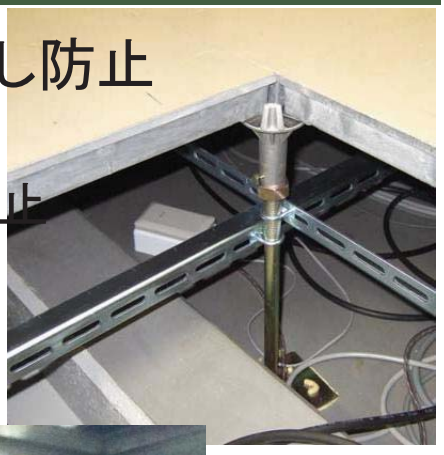
# フロアとラックの耐震補強

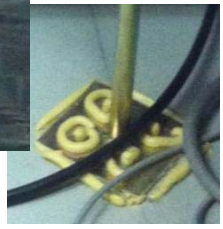
- 地震対策＝揺れても倒れないこと
  - 揺れだけなら電子機器は壊れないので
- 機器室のフリーアクセスフロア、ラックの耐震補強
  - 阪神・淡路大震災の報告を教訓とした取組み
    - フリアクの支柱が倒れると、全体がドミノ倒し
  - フリーアクセスフロアの耐震補強
  - ラックの転倒防止対策



# フロアとラックの耐震補強

- フリーアクセスフロアのドミノ倒し防止
  - 支柱の根がらみ、接着
  - 支柱の爪＝フロアパネル落下防止
- ラックの転倒防止
  - チャンネルベース（底へ固定）
  - スカート（転倒防止）





# 建物構造と耐震対策

- 頑丈な建物構造　＋　機器室を低層階に設置
  - 1，2F：　機器室：　ほぼ被害なし
  - 3F：　事務室：　棚などは耐震固定、棚から落下
  - 4F：　研究室：　歩行困難、棚や机から落下
  - 5F：　講義室：　机が転倒

- 執務室の地震対策
  - 本棚の耐震固定等対策
  - 棚からの落下防止
  - 怪我人がなく、居室の被害も少なく、業務継続



# 備蓄物資

- 備蓄物資
  - 非常食料、飲料水、懐中電灯、乾電池等
    - 2009年から、十分な備蓄
    - これにより復旧対応の執務を継続可能
    - 懐中電灯：　被災当日の帰宅と停電中の状況確認作業
    - 非常食料：　小人数職員×約2週間の昼食を供給
      - 炊き込みご飯などの種類の多いアルファ米＝精神的な支え
  - 屋上受水槽が水を確保
    - 大容量 ＋ 節水対策のため，給水再開まで断水なし
    - 飲料水ペットボトルの備蓄もあった

# 学内情報基盤の災害対策の課題など



## 事前の対策の不足

- 職員の連絡手段
  - 緊急連絡網の不通のため運用が困難
    - 一部携帯電話、IP電話回線、多機能電話、などは不通
  - 電子メールや学内情報システムが使えない状況で、
    情報伝達・共有する方法
    - 学外のメール、ウェブ、グループサービスなどの準備、
  - 移動手段の鍵 ＝ ガソリン、出勤困難、・・・

# 事前の対策の不足

- 情報サービス運用の非常時体制（BCP）
  - BCPは大学運営と密接であり、策定は困難
  - まず、各システムの復旧手順・補完手順から
    - 非常時の運用体制、部局との連絡手段確保
- 非常用電源の確保
- 情報システム・サーバの学外配置
    - 情報伝達用の、ウェブサーバ、ウェブメールサービス
    - tohoku.ac.jpドメインネームサーバ（消失・・・）
    - 安否確認システム
- ウェブサーバのコンテンツ更新の備え
- 教職員グループウェアや、学内ポータル



# 非常時の情報基盤の維持

- 非常時の情報基盤の維持に、一大学の自力でできることの限界
  - 地域の社会基盤の損壊（電力、交通、通信）
  - 情報システム最適化（コスト削減）との両立
  - 総合大学は、平常時でさえ、目が届き切らない
- 今なら、学外インフラの活用の検討も選択肢
  - 学内集約（学内DC）は、最適化とセキュリティの観点で見て、効果が中途半端になる
  - データセンター（市内，遠隔地）やクラウドサービスの活用を採り入れる検討

# 学内情報基盤の今後の課題

- 非常時に備えるための情報基盤の整備と活用
  - 基盤設備や体制の“多元化”
  - システム最適化・費用圧縮との両立または相乗効果
- 設備の防災対策
  - ラックの耐震（一部未施工）、学外設置
  - 電源確保
    - UPSの維持時間と収容機器の対応のクラス分け
    - 非常用電源の整備（UPS容量の先の運用継続）
    - 節電・計画停電、ピークシフトへの備えも兼ねる
- 運用継続
  - 情報システム業務継続計画の策定とインフラ整備
    - まず、各システムごとの障害時対応手順の整備から
  - 運用継続のための学外設置



# （続き）

- 非常時広報・連絡
  - ウェブページ，メールサービス、連絡手段の多元的確保
  - 情報不足・不正確情報への対策
  - 情報発信には，正確，丁寧，相互理解が必要
- 安否確認の方法
  - 業務・体制、学外ASP利用？、システムによらない方法？
- 学術認証連携
  - 学内・学外への避難研究室のネットワーク利用に必須
- 他大学との連携
  - 経験、ノウハウの共有
    - 地域大学間：　東北学術研究インターネットコミュニティTOPIC
    - 行政：「東日本大震災被災地自治体ICT担当連絡会（ISN）」
  - 近隣大学間、遠隔地大学との相互補助？

シ ス テ ム 技 術 分 科 会 選 出
システム技術分科会 2013 年度第 2 回会合 より

# 広帯域データ伝送システム ULTRA の研究開発

大江 将史
(国立天文台)

# 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 <masafumi.oe@nao.ac.jp>

## 1. はじめに

国立天文台(以下, 本台)での種々プロジェクトにおいて, 10 ギガビット毎秒を超える広帯域データ伝送が必要となっている. これに対して, 10 ギガビット毎秒を超える高帯域ネットワーク基盤を安価に構築することが可能となったことや, 汎用 PC(Intel Architecture: IA)サーバの処理能力が飛躍的に向上するなど, データ伝送処理を安価に実現できる環境が整うようになった.

そこで, 本台では, 低コストかつ広帯域データ伝送を実現する ULTRA の研究開発に取り組んでいる. ULTRA は, 汎用ネットワーク機器と汎用 PC サーバを基盤とした超高速 IP ルータやストレージキャッシュシステムであり, 2013 年には第 3 世代「連雀」を開発し, ピーク時に 200 ギガ級の処理能力を発揮する.

本発表では, ULTRA の紹介に加え, 科学的成果を生み出すために必要なキャンパス基盤ネットワークについても述べる.

## 2. 背景

本台の各天体観測システムや計算機システムなどは, その最適な拠点に整備され, 本台の拠点間を結ぶ WAN を通じて相互接続している. 本台の主要拠点とその WAN 構成は, 図 2 の通りである. また, 必要とする帯域や, データ伝送が必要なるタイミング, リアルタイムの必要性などの要求要件は, システムごと異なる（表 1）. 故に, 各システムがデータ伝送を無秩序に行った場合, WAN の輻輳などにより, 全体の利用効率が著しく低下する点が課題となっている.

## 3. ULTRA の開発

この課題に対して, ULTRA では, 伝送データを超高速ストレージによりキャッシュし, 各システム（アプリケーション）の稼働スケジュールや必要帯域, （プロパティ), WAN のトラフィックウェザーに基づいた WAN 伝送スケジュールに従って, TCP/UDP を組み合わせた高効率 WAN 伝送により解決をはかっている.

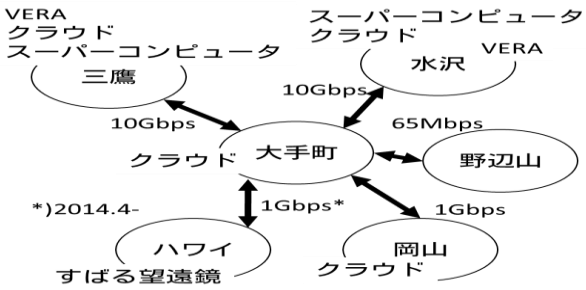



**図 1 WAN 構成図**

**表 1 システムごとに異なる要件例**

| システム名 | データ帯域 | 要求要件 |
|---|---|---|
| すばる HSC(Hyper Supreme Cam) | 一晩で 250GB～/夜（ハワイ→三鷹） | 観測計画に基づく伝送, 観測後に伝送 |
| VERA プロジェクト VLBI 観測 | 2Gbps～/局（各局→三鷹） | リアルタイム性必要, 観測計画に基づく伝送, 局数に応じた帯域が必要 |
| スーパーコンピュータ | 500M～2Gbps（水沢→三鷹） | 定常伝送, 計算バッチ終了（最大 8 時間周期）のタイミングで伝送 |

この実装のため, ULTRA では, アプリケーション要求要件を満たす要素技術の検証結果から, 汎用 PC サーバと高性能なネットワークカード(NIC)を組み合わせて実装を行っている. この理由は, 検証を行った 2011 年当時, ギガビット級レベルでは FPGA による実装が最も高性能であったが, 10Gbps 以上の帯域では, クラウドの伸長から汎用 PC と NIC の高性能化と低価格化の速度が, FPGA のそれを超えていくことが予想されたためである.

現在, ULTRA では, 超高速 SSD ストレージと 100Gbps IP ルータの 2 つのトラックに分けて, 開発を行っており, 前者は野川(2012 年), 後者は, 大沢(2012 年), 連雀/連雀+(2013 年)を発表し, 2014 年現在, その両者の統合を進めている.

野川は, Intel Nehalem マイクロアーキテクチャ上に, RAID0 接続された合計 16 台の SATA 汎用 SSD を組み合わせ, 32Gbps の連続書き込みに耐えうるストレージ性能を有している.

一方, IP ルータのである連雀+は, Spirent 社の計測器により, 120Gbps 以上の IP フォワーディング性能と 10 μsec 以下の低遅延伝送の両立を実現している. このシステム構成は, Intel 社 Sandy Bridge-E マイクロアーキテクチャ上に, PCI-e 3.0 対応の 40GbE ネットワークカードを搭載しており, OS は, CentOS6.3 をベースとしたものとなっている.

## 4. 国立天文台基盤ネットワーク

本台では, 国立天文台基盤ネットワークシステムにおいて, 各所のサーバルーム内にて, 10/40 ギガビットイーサネットをアクセスインターフェースとして提供している. この目的は, L3/L2 スイッチング機器向けの汎用 LSI の登場により, 安価に高速ネットワークの構築ができるようになった点と, 先に示した ULTRA の開発事例のようにその帯域を生かしきれるシステムを安価に構築できるようになった点にある.

また, クラウド技術の成熟により, 各システムは, これまでのシステムに紐づく計算機やストレージといった物理的な要素の組み合わせで構築されるのではなく, 要素をリソースとして, たとえば, 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, LAN, WAN などを集約化し, 各システムの目的などに応じて, リソースから仮想化技術により, 組み合わせて構築されるようになった.

この場合, リソースの結合には, 広帯域・低遅延なネットワークが必要不可欠であり, このようなトレンドにも対応できるよう本台の基盤ネットワークは設計されている.

## 5. まとめ

汎用 PC サーバや NIC の高性能化は, そのデータ処理能力を飛躍的に向上させ, だれもが, 100Gbps 以上のデータ処理が可能になった. また, その成長を支えるネットワークも汎用化により高性能低価格化が進んでいる. これらの利活用は, ノウハウを要するが, そのコスト対効果は非常に高く応用性も高いことから, 組織内の生産性を高めることができる.

本台の事例からも, 現在において, 組織内のリソースを縦横無尽に結ぶための広帯域情報ネットワークの導入と活用は, 生産性を高めるうえで, 重要であることがわかる.

# 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？

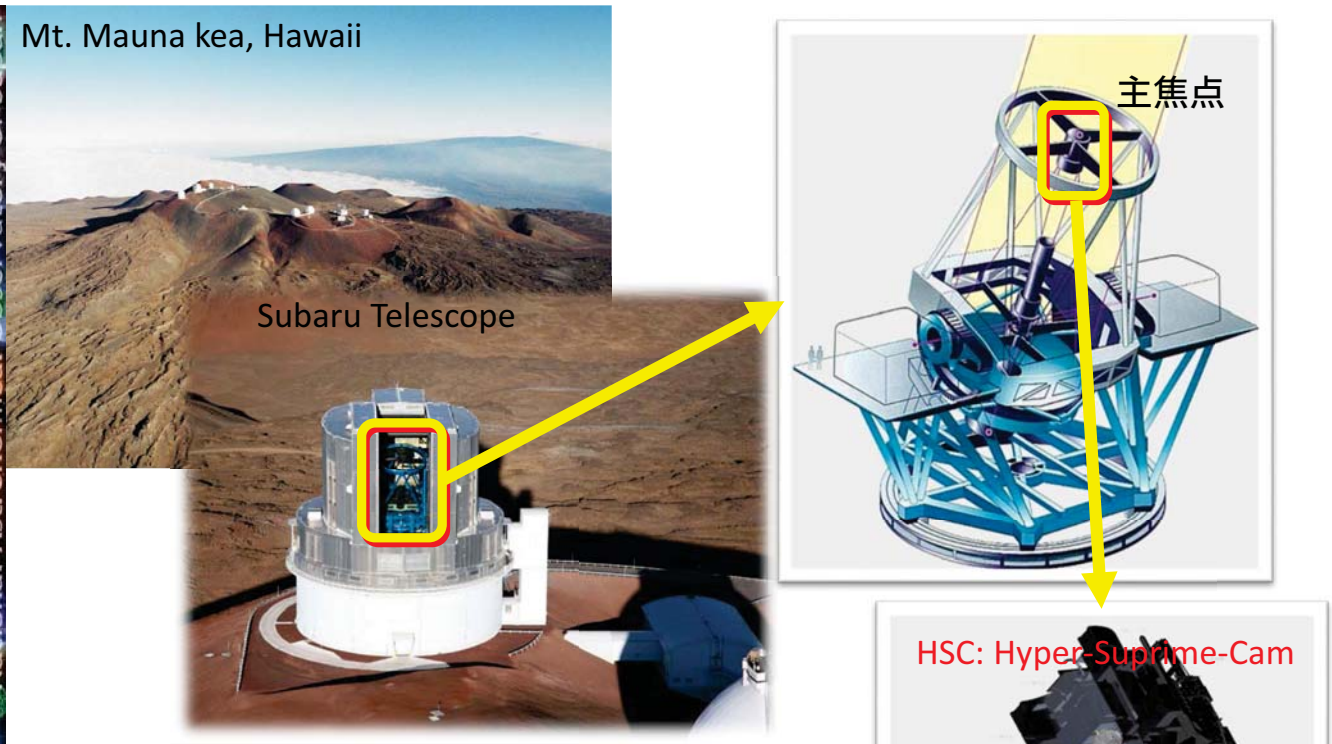
1）きてます

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)

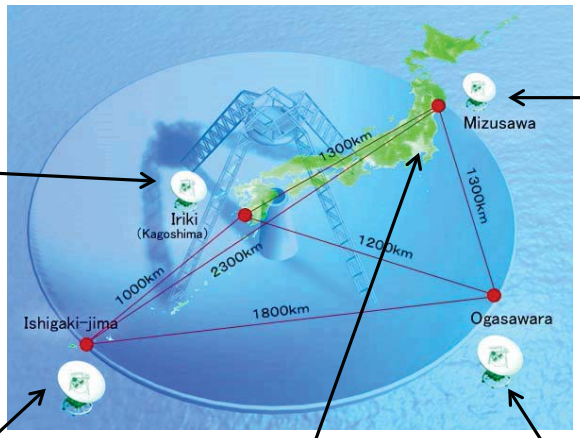
専用計算ノード群
(三鷹)

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

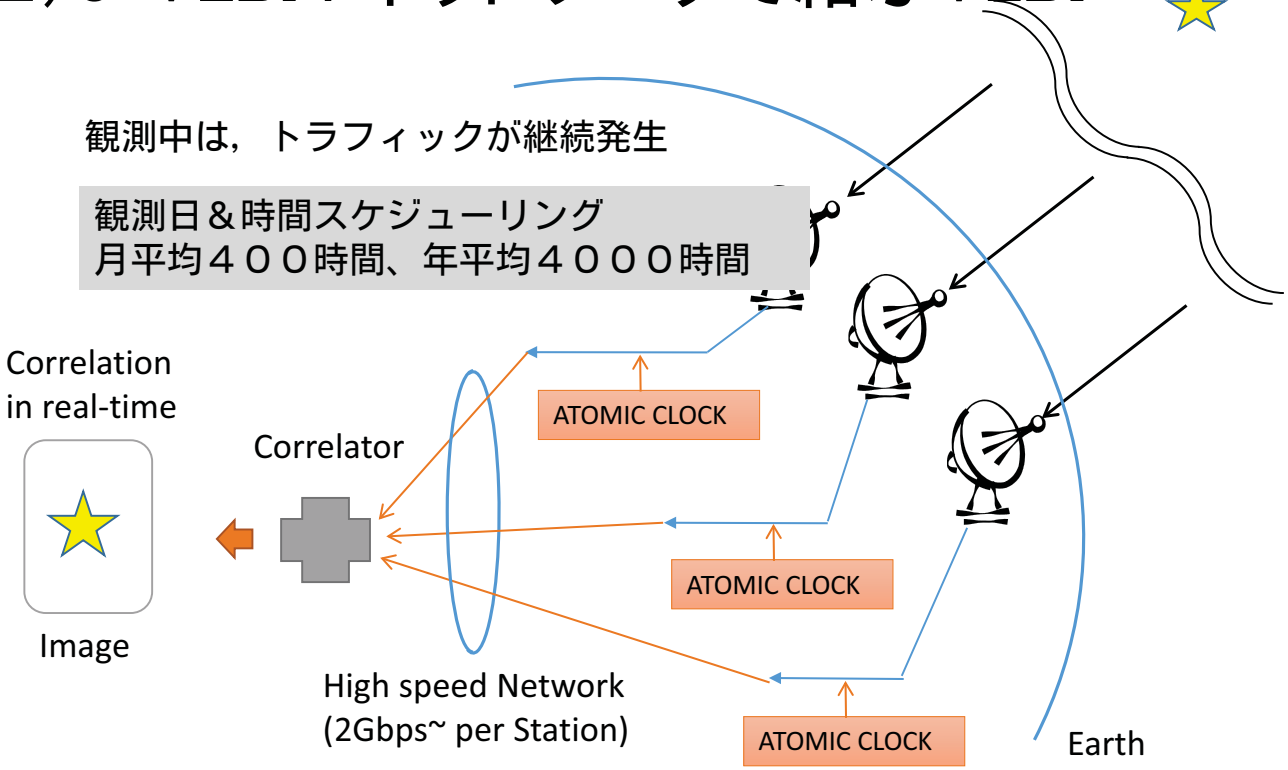
望遠鏡（山口・茨城・他）



# 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

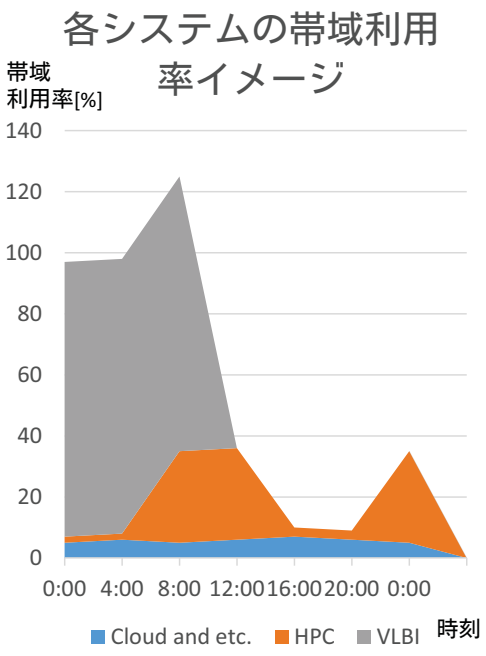
## 4）コンテンツ配信

- 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

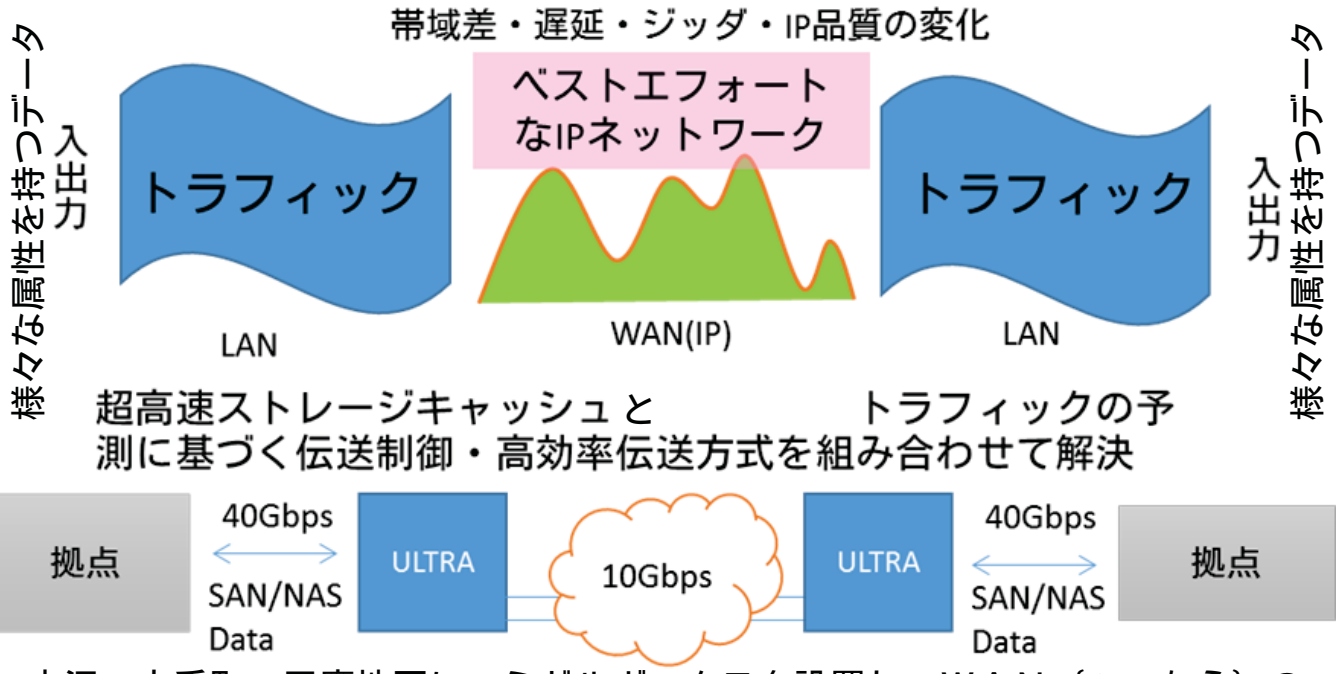
- アプリケーションが必要とするネットワークの高性能化・広帯域化・トラフィック個性
  - スパコン・VLBI・クラウド・映像中継等々
- WAN広帯域化とLANさらなる広帯域化
  - WAN-LANの帯域・性能ギャップの存在と絶対処理量の増大

➔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

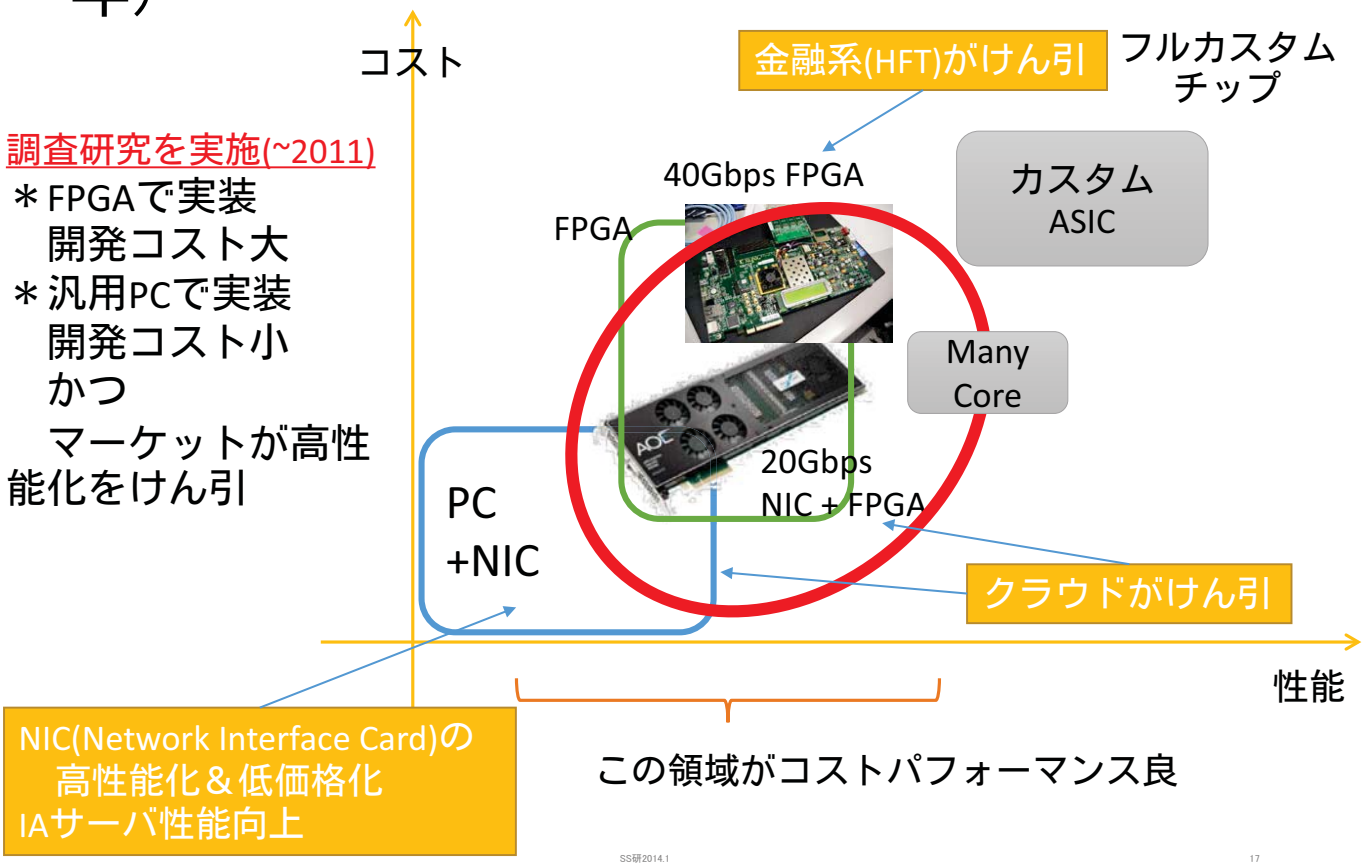
# ULTRA計画のアイデア
# ミドルボックスによるデータ伝送の効率化




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

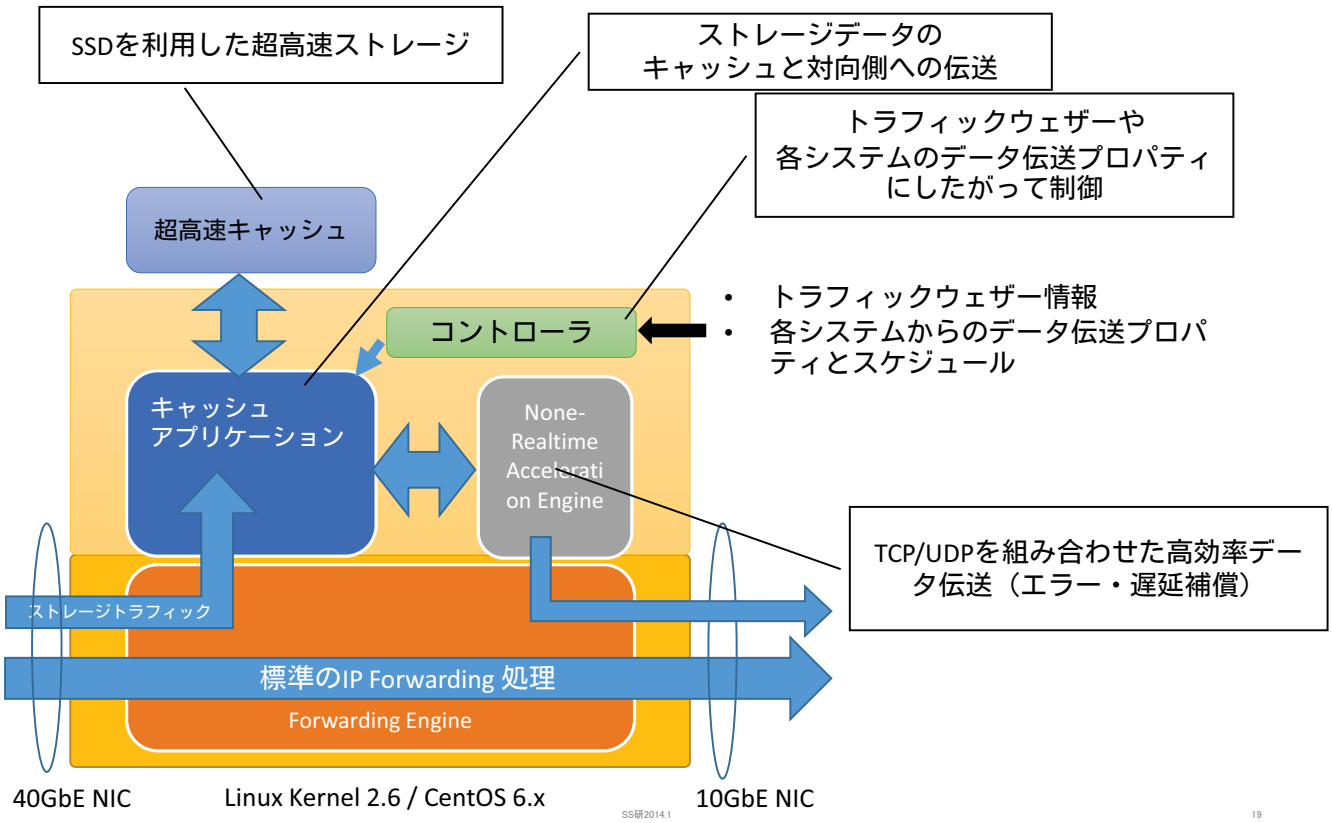
NIC(Network Interface Card)の
　高性能化＆低価格化
IAサーバ性能向上

# 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

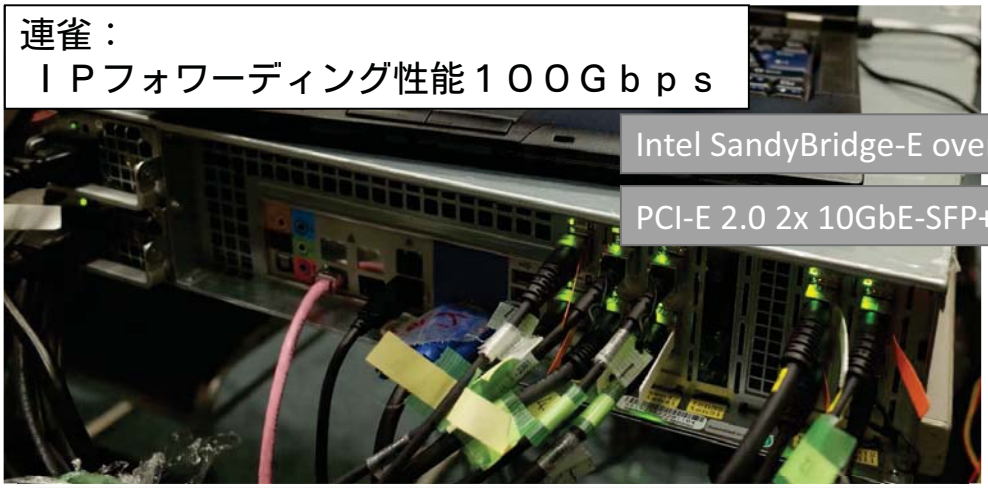


# ULTRAの機能ブロック構成



# 「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)

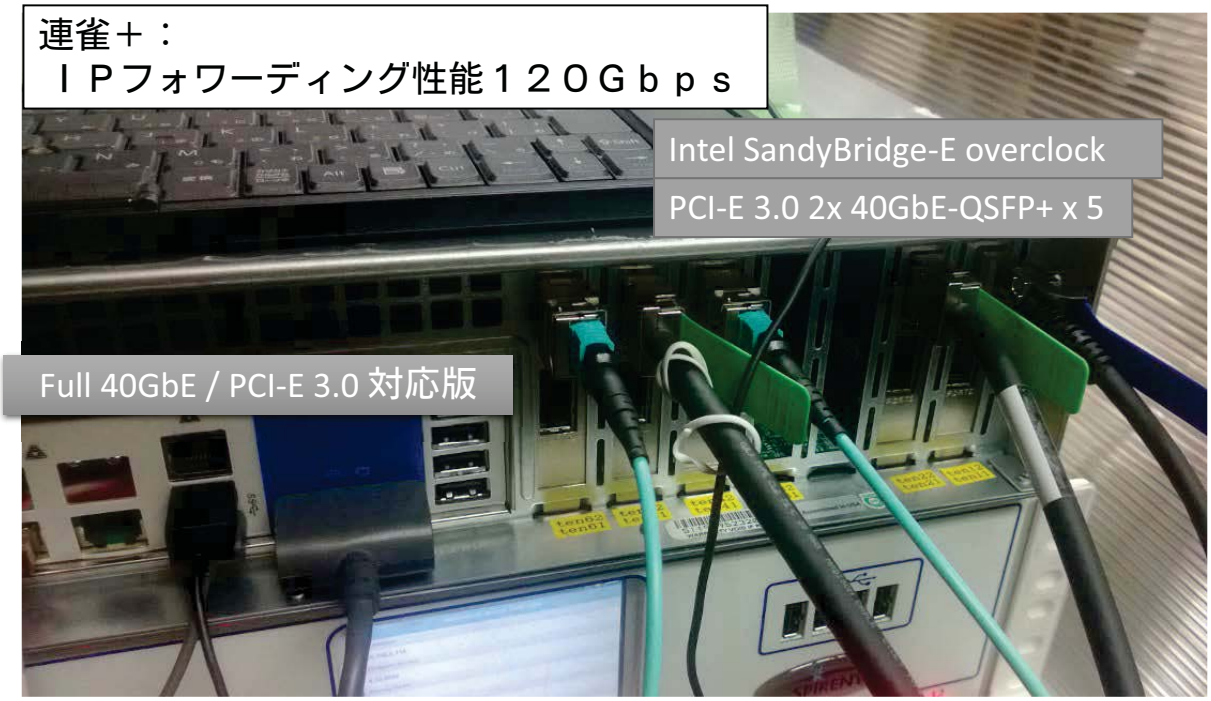
Interop2013 オープンルーターコンペティション(ORC)
富士通賞受賞

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



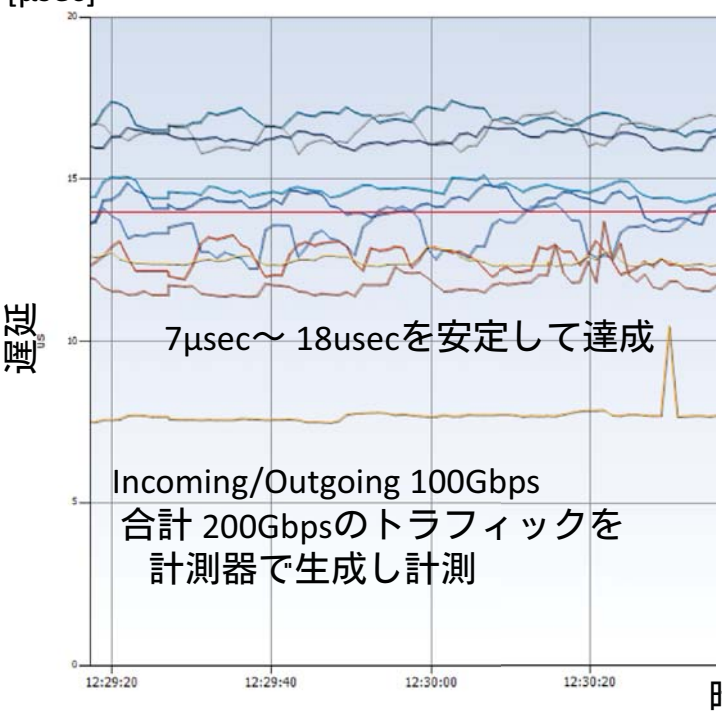
Intel SandyBridge-E overclock

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

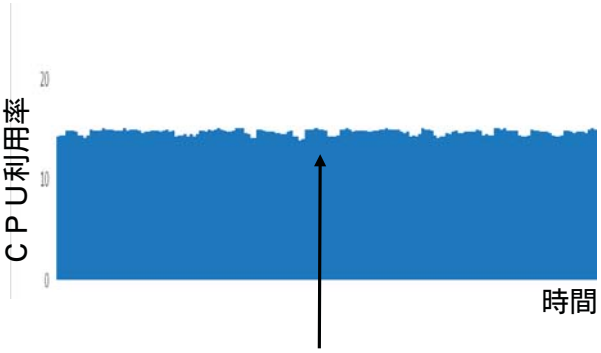
Full 40GbE / PCI-E 3.0 対応版

# 「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？

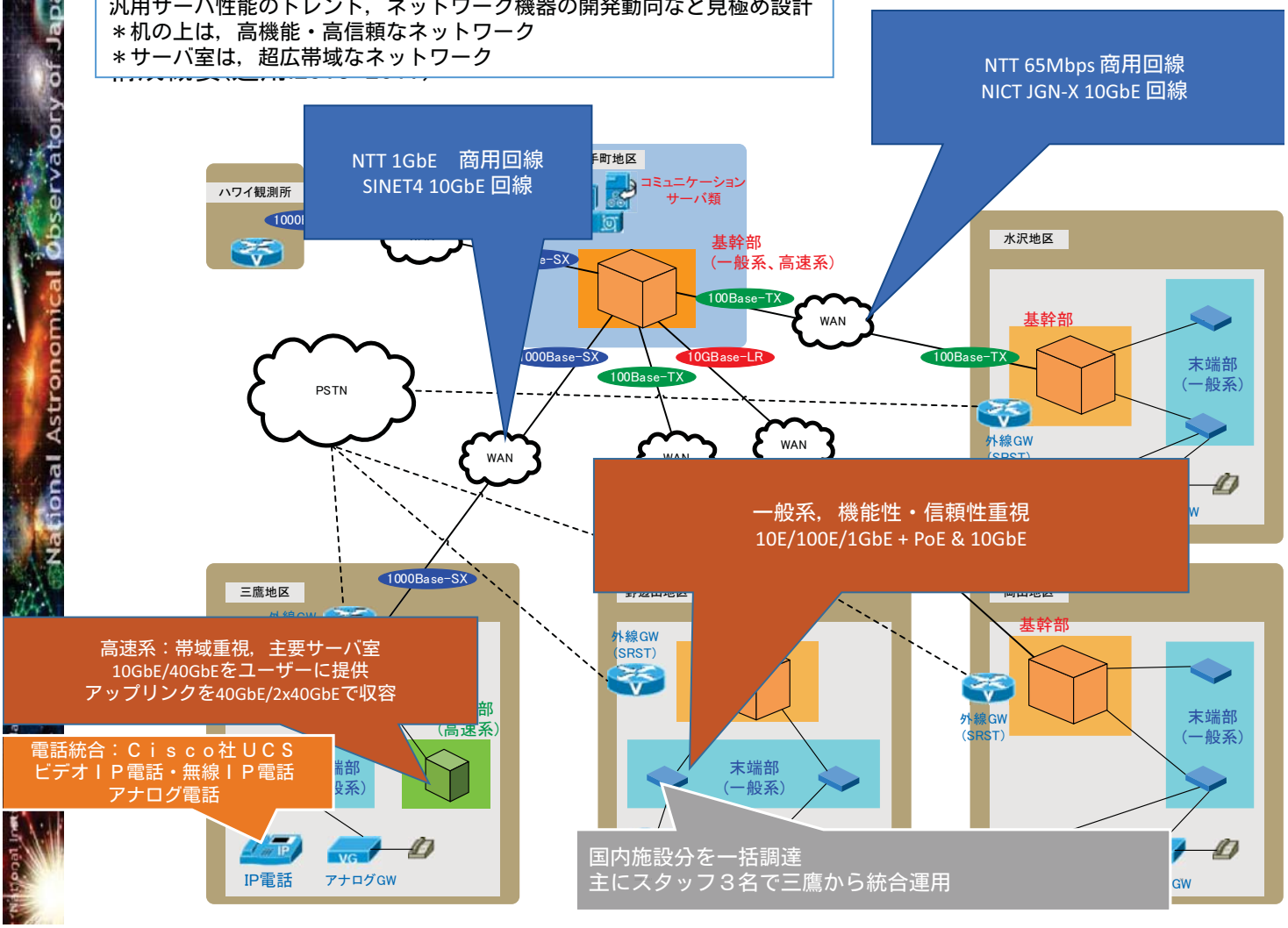
- 誰もが100Gbps～200Gbpsを扱える時代

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