超長距離超高速
データファイル転送への挑戦
Data Reservoir　プロジェクト

稲葉真理　平木敬
東京大学大学院
情報理工学系研究科

コアプロジェクトメンバ（平成12～15年）

平木敬（プロジェクトリーダ）
稲葉真理
中村誠　　東京大学（情報理工学系研究科、理学系研究科、情報基盤センタ）
亀沢寛之　　（プロジェクト統括（雑用）、ネットワークチューニング）
玉造潤史

陣崎明　　富士通研究所
下見淳一郎　　（Comet 開発グループ）
下國治

来栖竜太郎　　富士通コンピュータ技術研究所
坂元眞和　　（DRシステム開発グループ）
古川裕希
生田祐吉

中野理　　富士通コンピュータ技術研究所
鳥居健一　　（ギガアナライザ開発グループ）
吉田昇一
柳沢敏孝
水口健二

科学技術振興調整費
知的基盤整備のうち先導的なもの

平成９年　**知的基盤の整備を総合的に推進する知的基盤整備推進制度**

各省庁の国立試験研究機関，大学，民間会社等が**連携**し、 研究者の研究開発活動を安定的かつ効果的に**支える**ため、標準、試験評価方法、研究用材料及び先端的な試験装置等の知的基盤を総合的に整備する。

平成１２年

知的基盤整備 「**2010年までに米国並みの整備水準を目指す**」（閣議決定）

平成１３年　**先導的研究等の推進**

科学技術の急速な発展に先見性と機動性をもって対応するため、境界を越えた融合により新たな領域の創成が期待される先導的な研究開発を推進し、また、科学技術が**社会に与える影響**の広がりと深まりに先見性をもって対応するため、自然科学と人文・社会科学とを総合した研究開発を**先導する**ことを目的とする。

１件　３億から5億（原則3年以内）

平成１６年 「先導的研究の推進」プログラムとしては廃止。

１６年度の募集は行わない。 緊急研究については引き続き実施。

（文部科学省　科学技術・学術政策局による科学技術振興調整費説明会資料より）

平成12年　SuperSINET　開通

Super SINET を、真っ黒になるまで
（現実に役にたつことで）使ってみたい♪

平成12年　　パイロットモデル
平成13年～15年　知的基盤整備

20世紀の理学の発展

量子論、相対性理論による
自然観の変革

↓

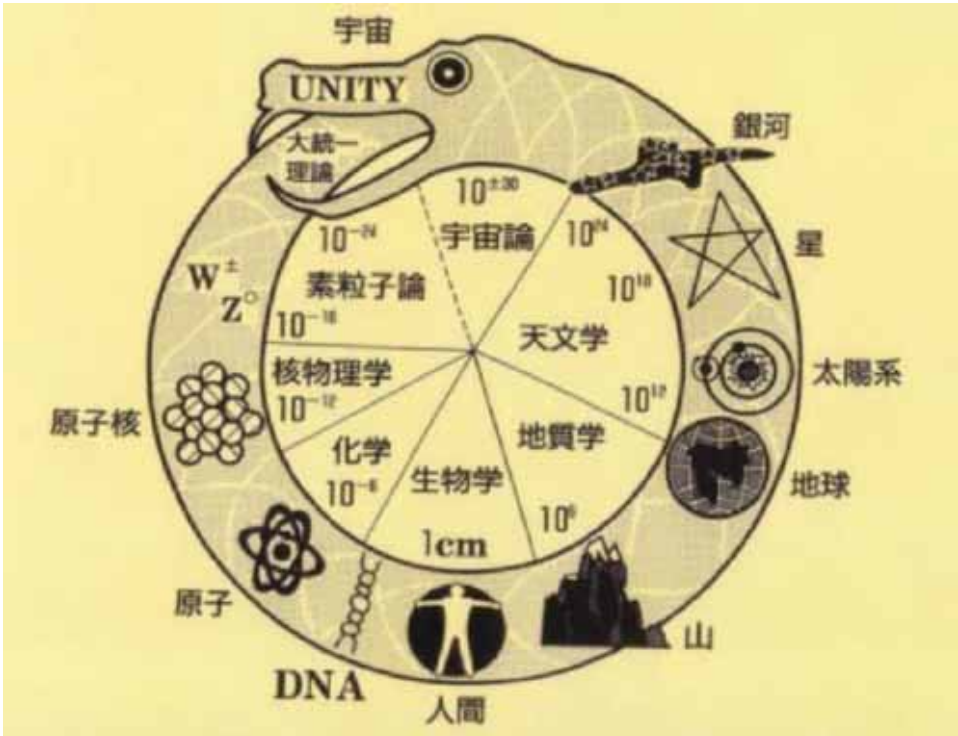
素粒子から宇宙まで

「質量とは何か」

「物質の究極構造」

「力の統一的な理解」

「ビッグバン」



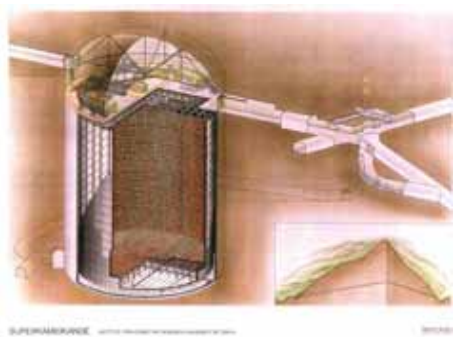
実験・観測機器の高精度化

あけぼの

加速器

すばる望遠鏡

野辺山電波天文台



スーパーカミオカンデ

データインテンシブリサーチ

- 巨大データを取り扱う新しい科学手法
  - 取得･生成データの巨大化
    実験･観測装置、シミュレータの発達
  - 計算処理能力の発達
    演算装置の高速化、メモリの高速、低価格化
- 観測装置（望遠鏡･加速器）は高価
  - 実験観測機器は順番待ちのことも
  - 実験･観察の場所と解析場所の分離。
  - 解析場所は研究グループ単位でまとまっている事が多い。

天文学専攻ネットワーク委員曰く

ネットワークは、論文やメールの交換のためにしか使わない
観測データはDLTテープを使って運ぶ。
FEDEX が、毎週、ハワイから東京へテープを運んでいる

大規模データ実験観測研究プロジェクト

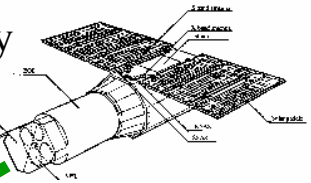
名前	所属	プロジェクト名	プロジェクト内容	国内接続先	海外接続先	現在のデータ量(平成12年当時)
酒井 英行	物理 原子核研究センター	SMART偏極実験	少数系散乱現象と三体力の検証実験 および核力の媒質効果の研究	理化学研究所 核物理研究センタ	CERN Brook Haven	CERN LEP　50±20　DAT/month CERN LHC　100 MB/sec Brook Haven　100 Mbps RCNP 加速器　50 GB/day
祖父江 義明	附属天文学教育センター	電波望遠鏡観測実験	銀河の基本構造(質量分布、ダークマター分布、銀河団の距離、ハッブル定数)の決定	国立天文台野辺山宇宙電波観測所	マックスプランク電波天文学研究所 ミリ波天文学研究所 米国電波天文学研究所	電波望遠鏡データ　200 GB
岡村 定矩	天文	スローン・デジタル・スカイサーベイ	スローン・ディジタル・スカイ・サーベイという全天の 1/4 の光学サーベイ	国立天文台 東北大学 名古屋大学	フェルミ研究所	データ総量:　10 TB 程度 フェルミ研から定期的に転送したい。
牧島 一夫	物理	初期宇宙研究観測	ASTRO-E2 HXD-II 、X線/γ線観測衛星で宇宙の構造進化や天体現象の内包する物理現象を探る。	宇宙科学研究所 広島大学 埼玉大学	NASA European Space Agency	既存衛星　1GB/day 以下 ASTRO-E2　数GB/day 次世代X線衛星　数10GB/day
山形 俊男	地球惑星科学	地球流体変動シュミレーションデータ解析	計算機シュミレーションによる地球環境の解明	地球変動研究所	なし	1シュミレーション　10 TB 現在は、50ペタのテープアーカイバを利用。
小林 富雄	素粒子物理国際研究センター	ATLAS実験	実験およびシュミレーションデータ蓄積によるヒッグス粒子、超対称性粒子等の探索	KEK 京都大学 筑波大学	CERN	CERN LHC　100 MB/sec
尾中 敬	天文	赤外線天文衛星観測実験	ASTRO-F　赤外線天文衛星のデータ解析	宇宙科学研究所 名古屋大学 国立天文台	ESA 受信局(スウェーデン・エスレンジ)	ダウンリンク一回　200MB 1分以内に関連研究機関内で交換したい
牧野 淳一郎	天文	恒星系力学天体現象シュミレーション	世界最速の多粒子系専用計算機 GRAPE-6 による大規模シミュレーションの結果を解析・可視化	国立天文台	プリンストン高等研究所 アメリカ自然史博物館	最大スループット:　100MB/s 1シミュレーション当たり　10TB
相原 博昭	物理	KEK　b－factory	物質・反物質の非対称性の研究	KEK 名古屋大学 東北大学	プリンストン大学	生データ:600 GB/day 転送データ:　10GB/day
嶋作 一大	天文	すばる望遠鏡観測	ハワイ州マウナケアすばる望遠鏡に光学カメラを取りつけた天体観測	国立天文台	国立天文台ハワイ観測所	100 GB/day。 ピーク時(リアルタイムチェック) 0.5 GB /sec　(4Gbps)

Data intensive scientific computation through SUPER-SINET

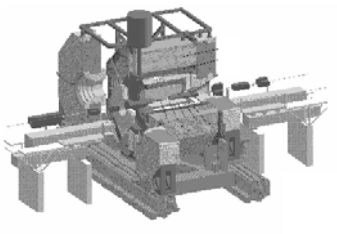
Nuclear experiment

Nobeyama Radio Observatory (VLBI)

X-ray astronomy Satellite ASUKA

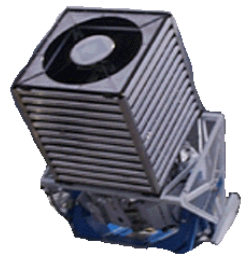


Belle Experiments

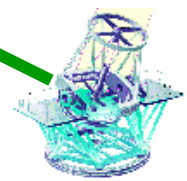


Data Reservoir

Very High-speed Network



Digital Sky Survey



SUBARU Telescope

Data Reservoir

Distributed Shared files

Data Reservoir

Local Accesses

Data analysis at University of Tokyo

CERN

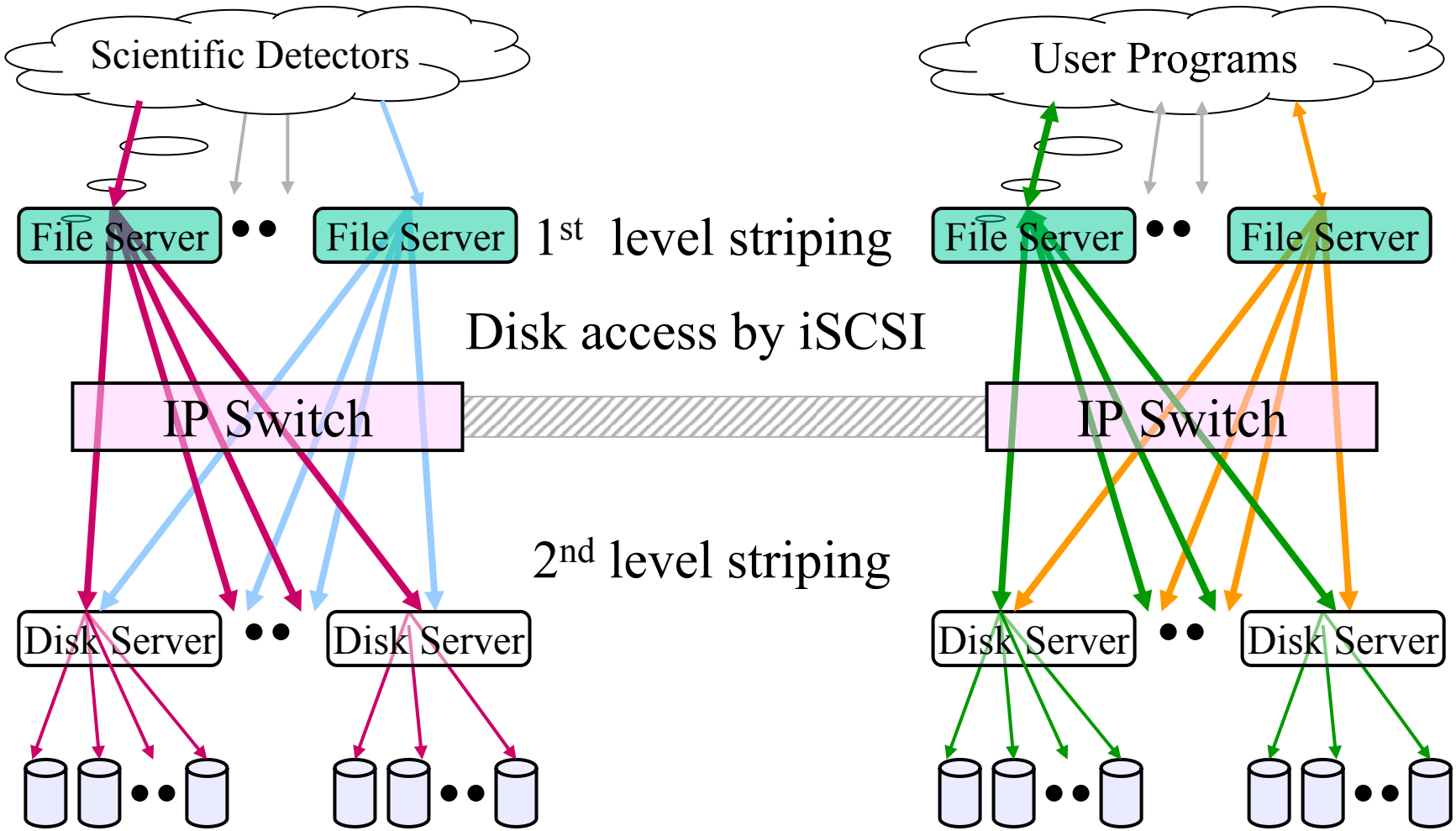
Oversea Sites

長距離高速転送への挑戦

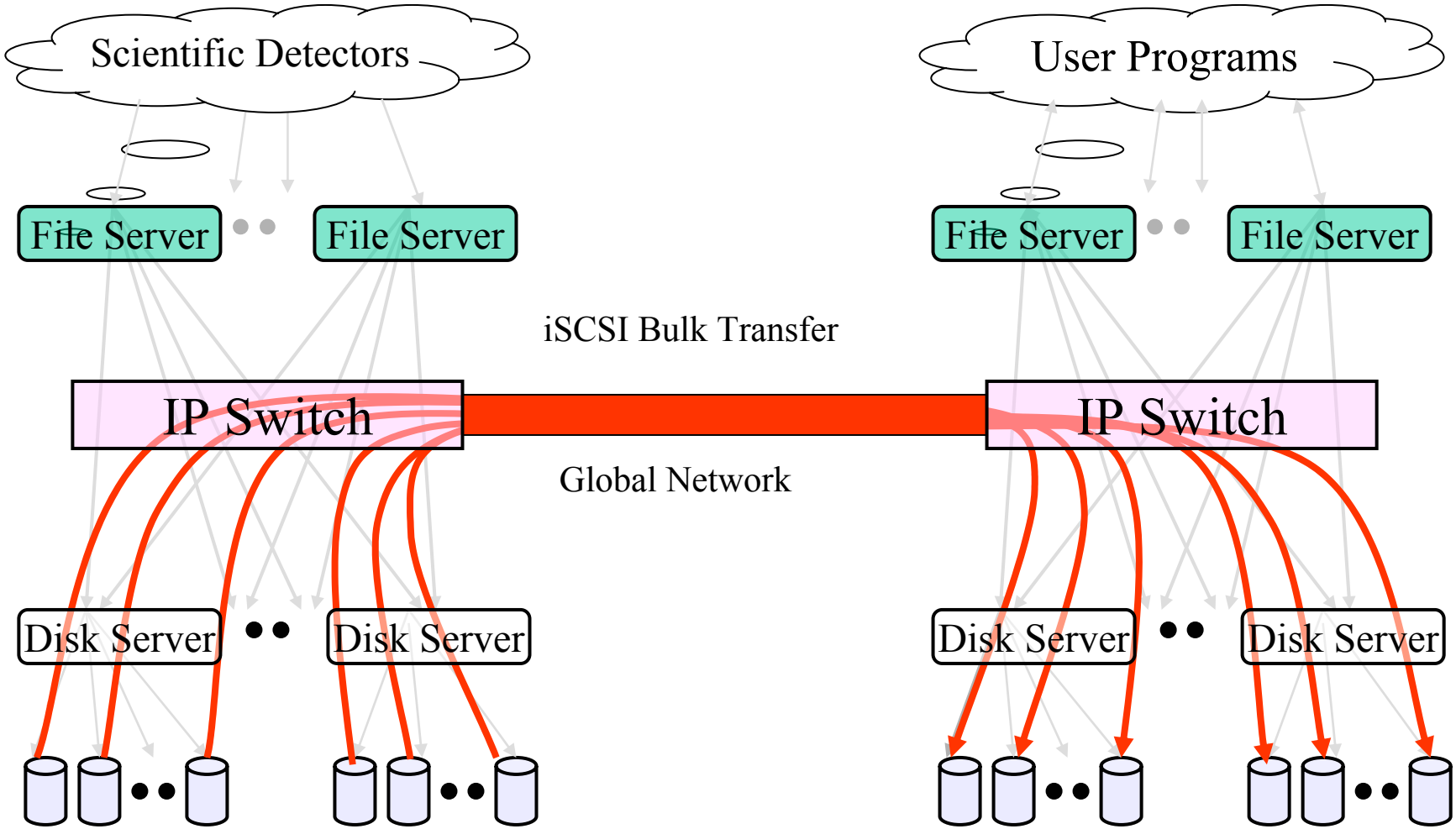
- 作りたいもの、　使ってくれる人

- システムデザイン
- 実験と改良、そして「ほしいもの」
  - 2002年5月　　国内実験
  - ツール作成　その1　　遅延箱
  - 2002年11月　U.S.実験　600Mbps
  - ツール作成　その2　ギガアナライザ
  - 2003年11月　U.S. 1往復半実験

システム概要（ローカルアクセス）



システム概要（遠距離データ転送）



ネットワーク上での転送実験

2002年春

- 室内実験　４台対向
- 40 km　（東京→相模原, 10ms, 1Gbps）　４台対向
- 1,600km　（日本半周, 26ms, 1Gbps)　　４台対向

　　東京→京都→大阪→仙台(東北）→東京

2002年秋

- 高遅延予備実験（遅延箱, 200ms, 0.6Gbps)　１台対向
- 12,000km 実験（200ms, 0.6Gbps）　４台対向

　　東京→メリーランド（富士通研究所アメリカ）

- 12,000km 実験（200ms, 0.6Gbps）26台対向 0.55Gbps

　　東京→ボルチモア（SC2002）

- 10Gbps 室内実験

2003年秋

- 15,680km 日米往復実験（171ms, 9.6Gbps) 16台対向 6.8Gbps

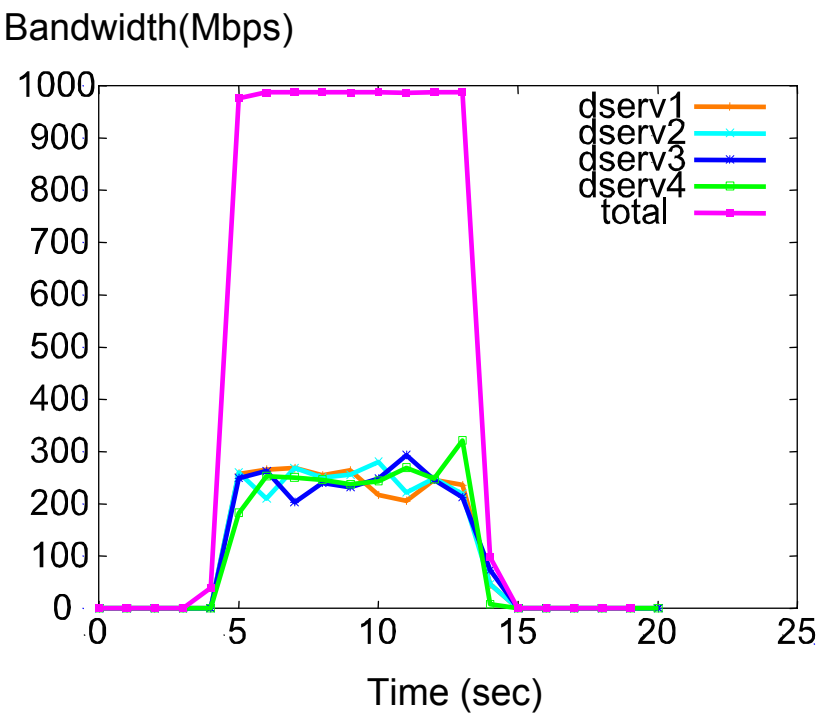
　　東京→ポートランド→東京

- 24,000km 日米１往復半実験（350ms, 8.2Gbps)　32 台対向 7.5Gbps

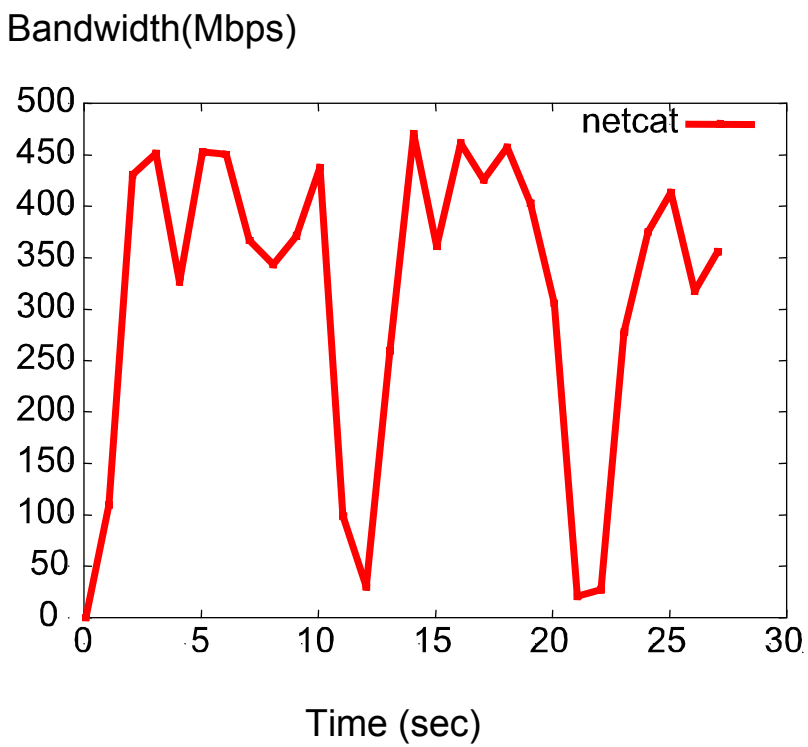
　　東京→ポートランド→東京→フェニックス（SC2003）

Basic Performance

Data Reservoir



Transfer through A file system (middleware solution)



日本半周までは幸せ。 95%の回線利用率

遠距離実験の難しさ

場所が離れてる（ブツ、配線、電気、空調）
回線の確保（独占的に使うのはとても大変）

計算機でシミュレーション （ありがち）

ほしいもの

遠くにいる親切な共同研究者♪ （重要）

実験室で、遠距離ネットワークのふりをしてくれるもの
「遅延」 と 「パケット落ち」 の実現

Comet ネットワークカード

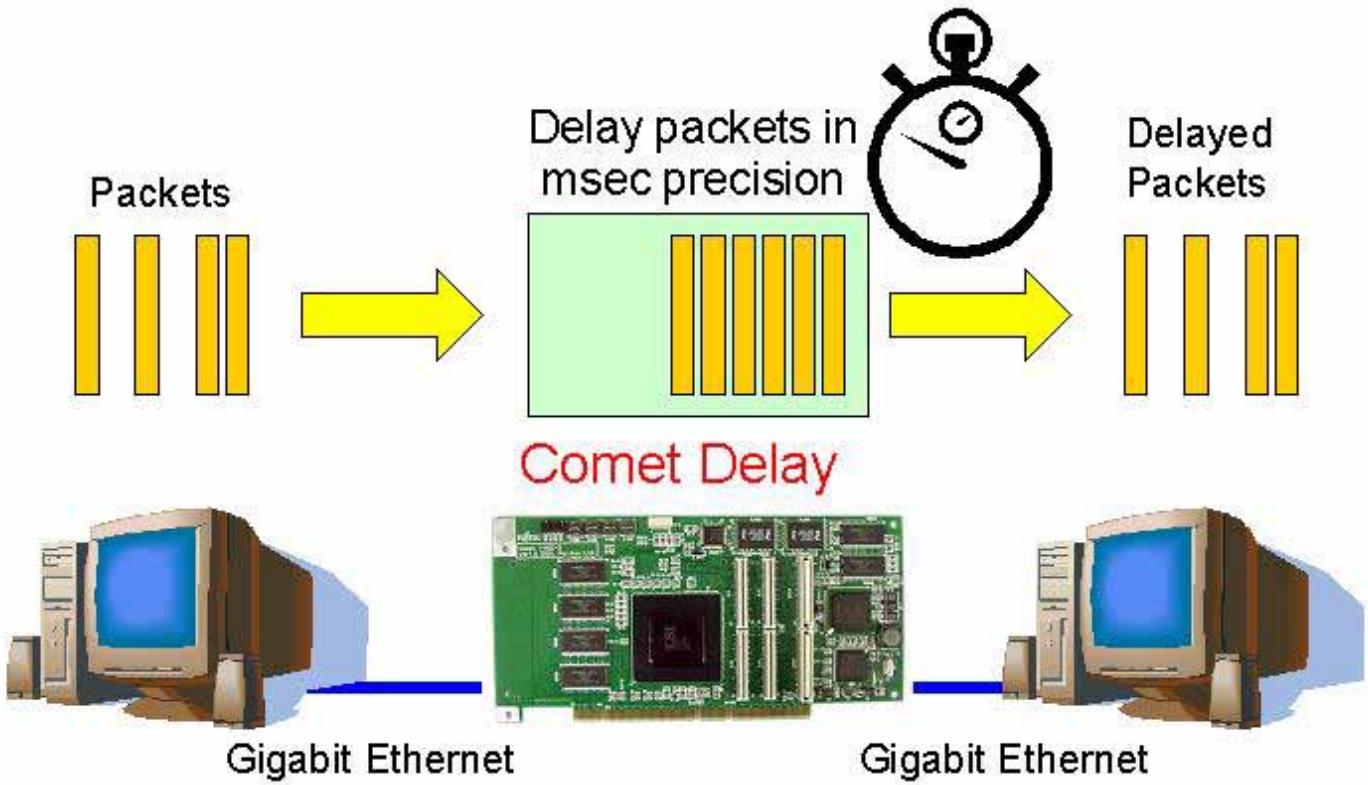
Gigabit Ethernet ２ポート
Comet NP (ネットワークプロセッサ） ＋ ARM ＋ バッファ

何が素敵か？
→ ネットワークカードでプログラムが動く
OS のオーバーヘッドの影響を受けない
CPU に負荷をかけない
制御はホストから

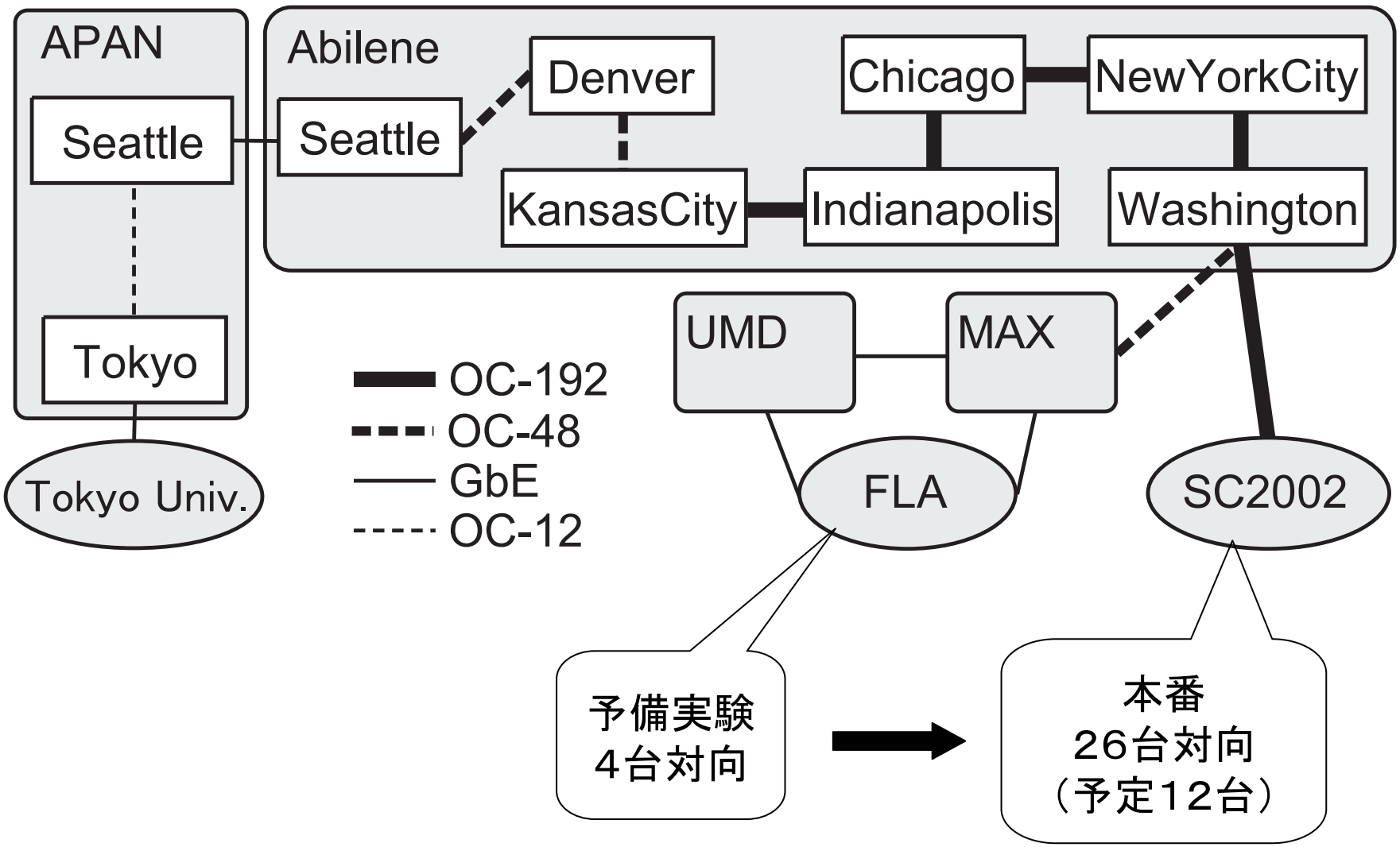
ストリーム処理に適している

遅延箱の作成

Comet Delay (バッファリング) / Comet Drop（Delay + 捨てる）



日米実験ネットワーク構成図



SC2002 （BWCの時のスライド）

BWC2002

560Mbps (200ms RTT)

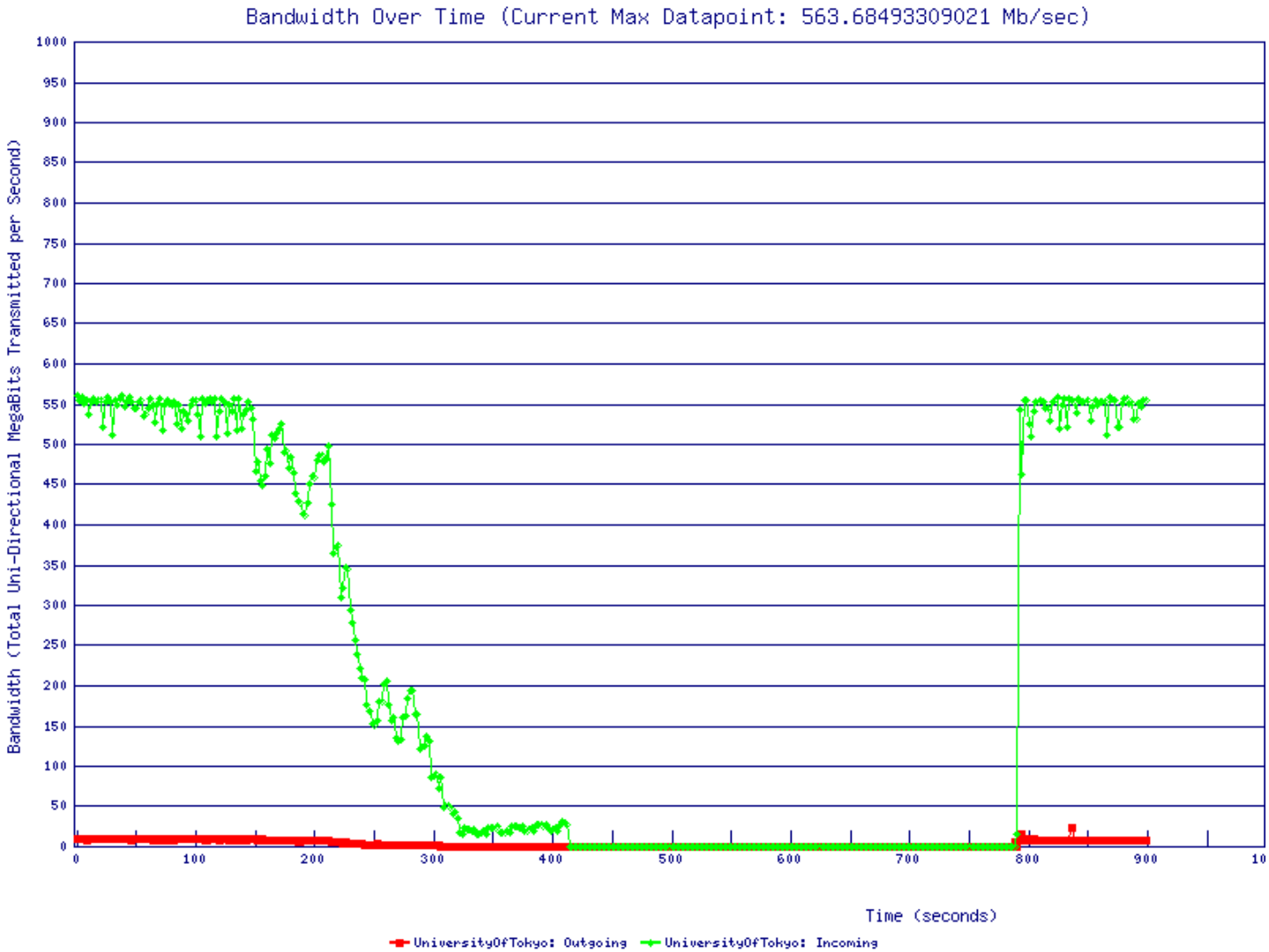
95% Utilization of available bandwidth

U. of Tokyo ⇔ Scinet (Maryland, USA)

⇒ Data Reservoir can saturate 10Gbps network when it will be available for US-JAPAN connection

と、プレゼンテーションで、約束（？）

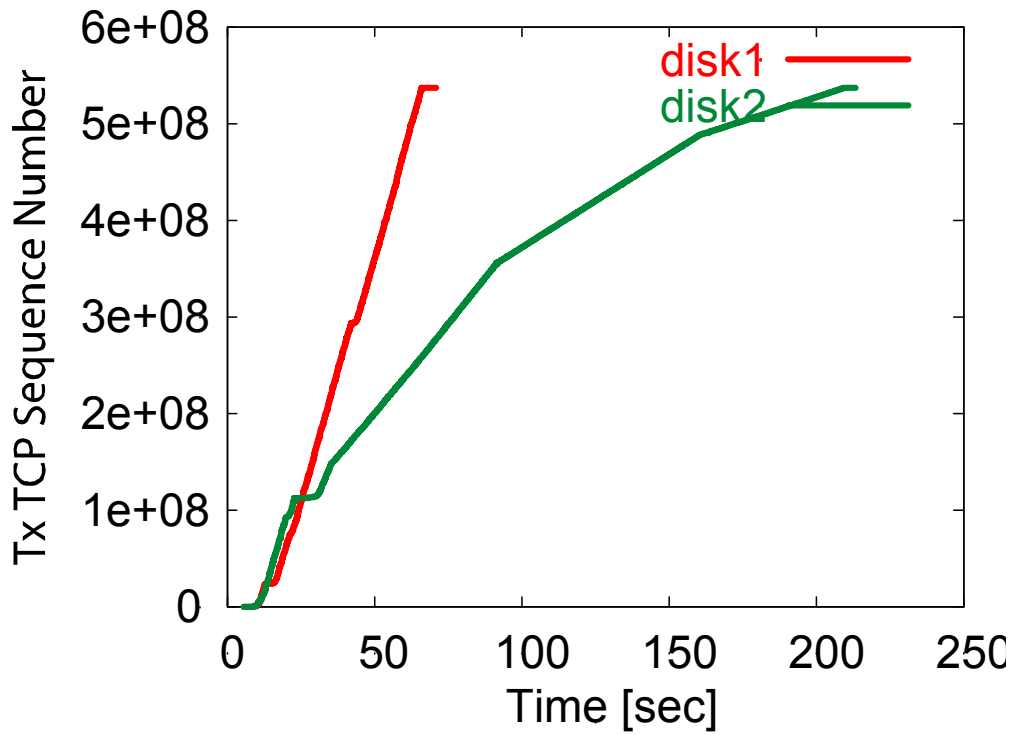
SC2002, RTT 200ms, 92 % bandwidth usage, 0.01% packet drop, exclusive use



LFN (Long Fat Pipe Network)　の難しさ

並列ストリーム

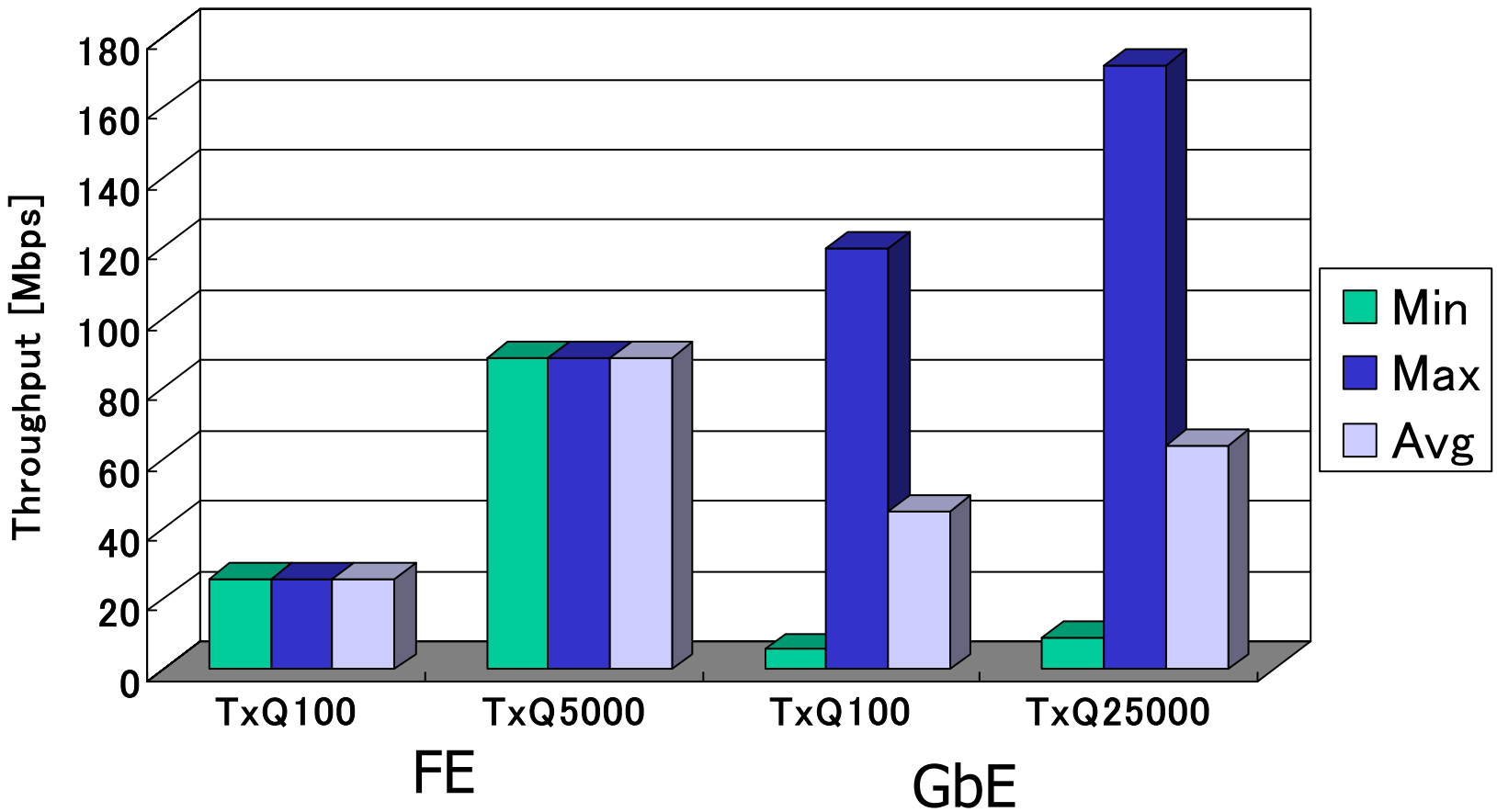
- ストリームの速度のばらつき
- 最も遅いストリームがシステムのパフォーマンスを決める
- 遅延箱実験結果とも、かなり違う



Fast Ethernet vs. GbE

Iperf　（30 sec）

最小、平均ともに: Fast Ethernet > GbE



わけのわからないことが多すぎる

理屈にあわない
実は、何もわかってない

ほしいもの

パケット解析装置

したいこと

シングルストリーム、そして並列ストリームのチューニング
日米ネットワークの両端でとったログのつきあわせ
正確なタイムスタンプつきのログ
ただし 1Gbps

1Gbps でパケットのやり取りをするということ

1パケット 1250バイトとすると、1 パケット 10000ビット
（本当は1500＋αが一般的）

1秒間に 10万パケット
1msec に 100 パケット
ということは、約10μsec にひとつ、パケットが来る。
→ タイムスタンプに、100ns くらいの精度が欲しい。
パケットロス 0．1%
→ 1秒に100個なくなる

一方一般的なUNIX のカーネルタイマーは 10msec(2002年）

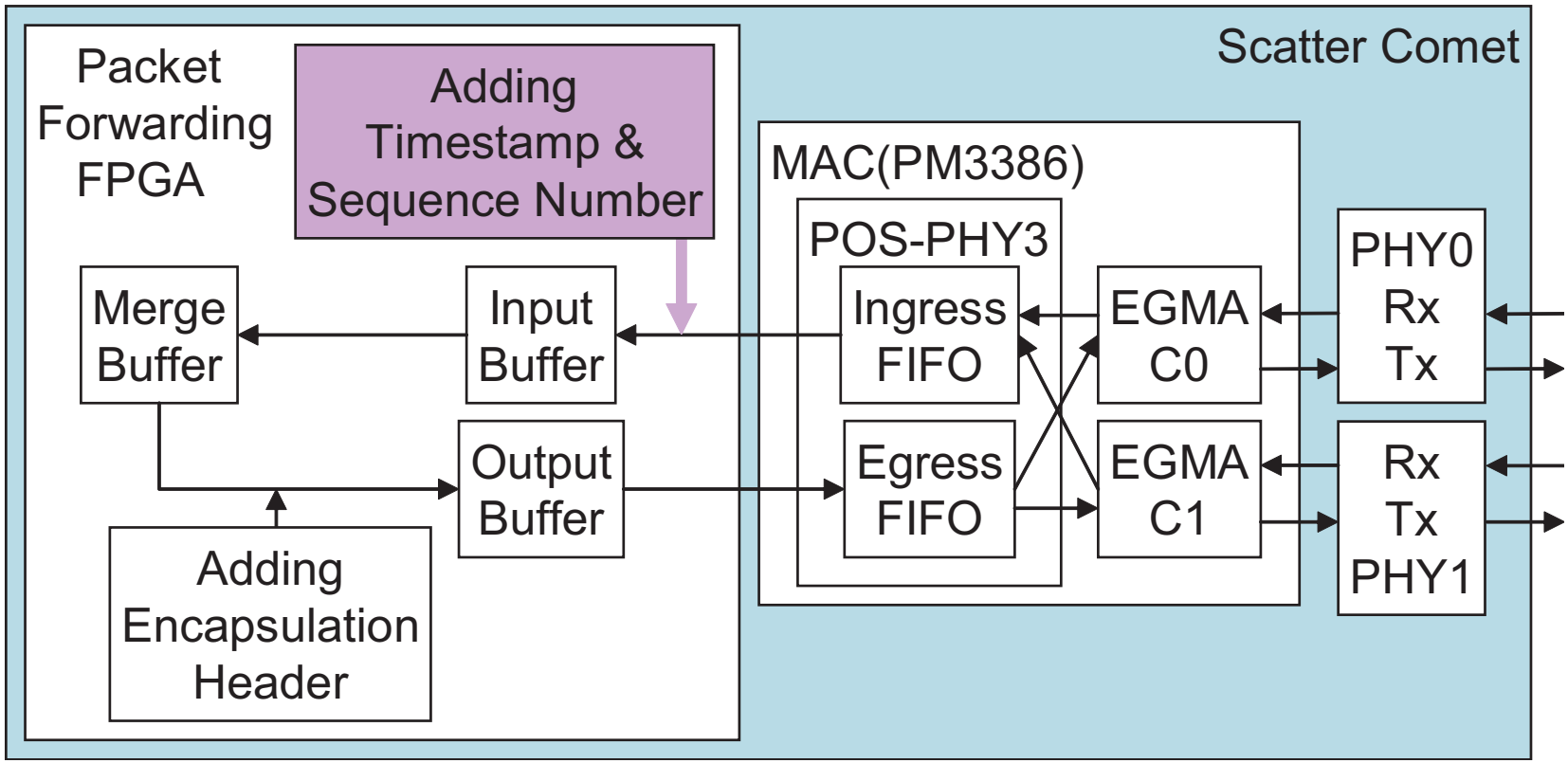
Comet ネットワークカード

Gigabit Ethernet ２ポート
Comet NP (ネットワークプロセッサ） ＋ ARM ＋ バッファ

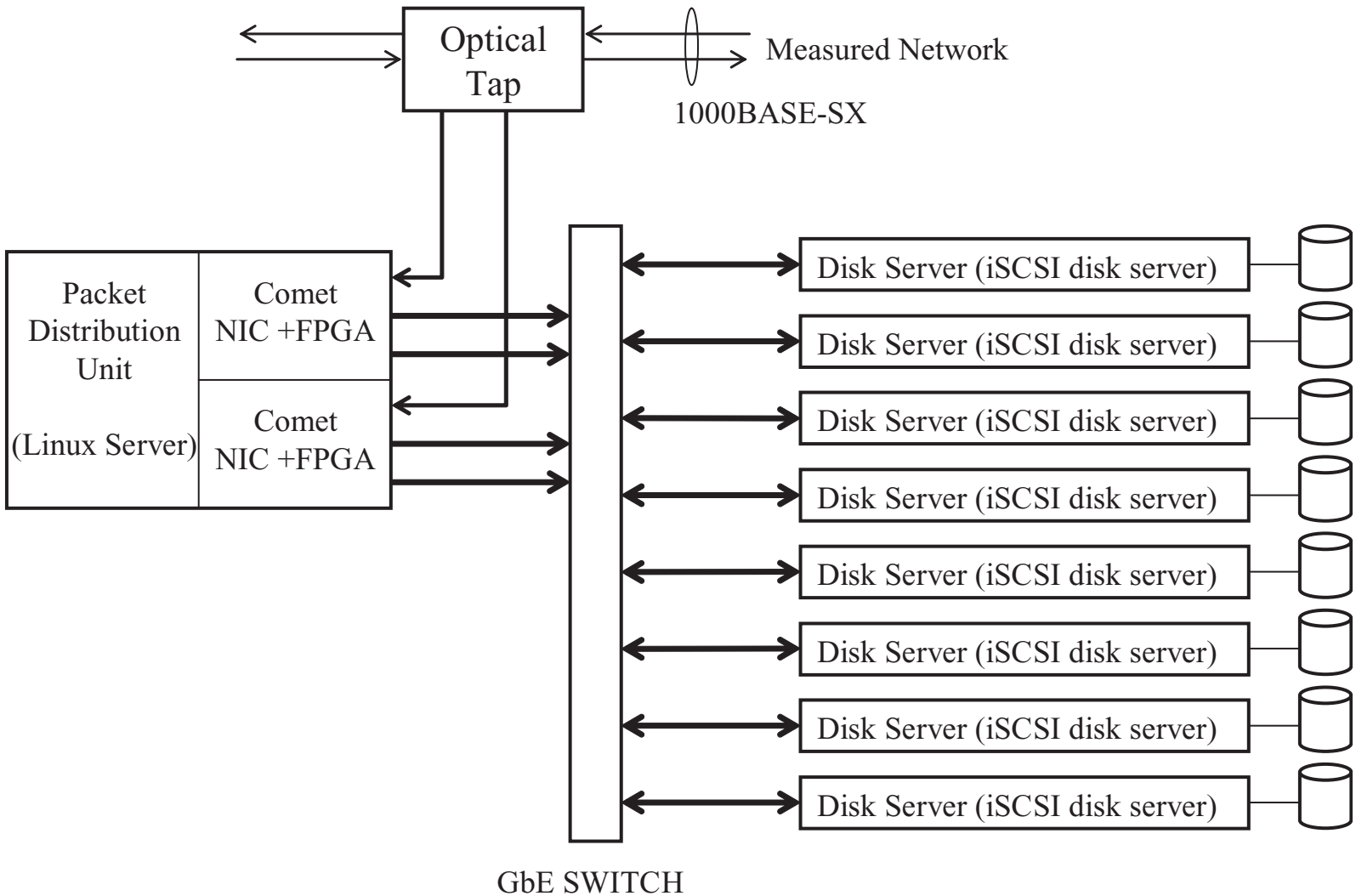
何が素敵か？
→ ネットワークカードでプログラムが動く
OS のオーバーヘッドの影響を受けない
CPU に負荷をかけない
制御はホストから

ストリーム処理に適している

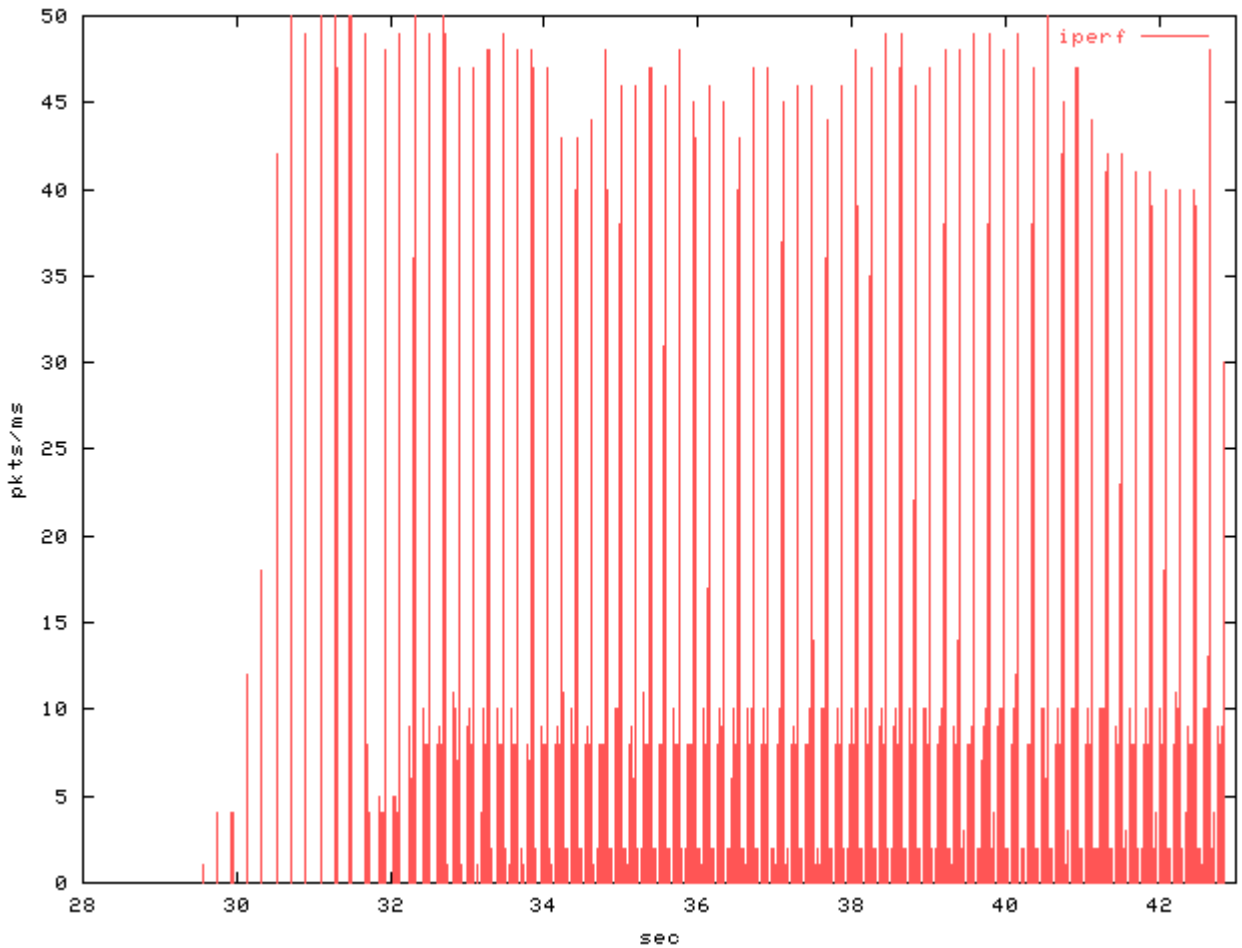
Programable NIC



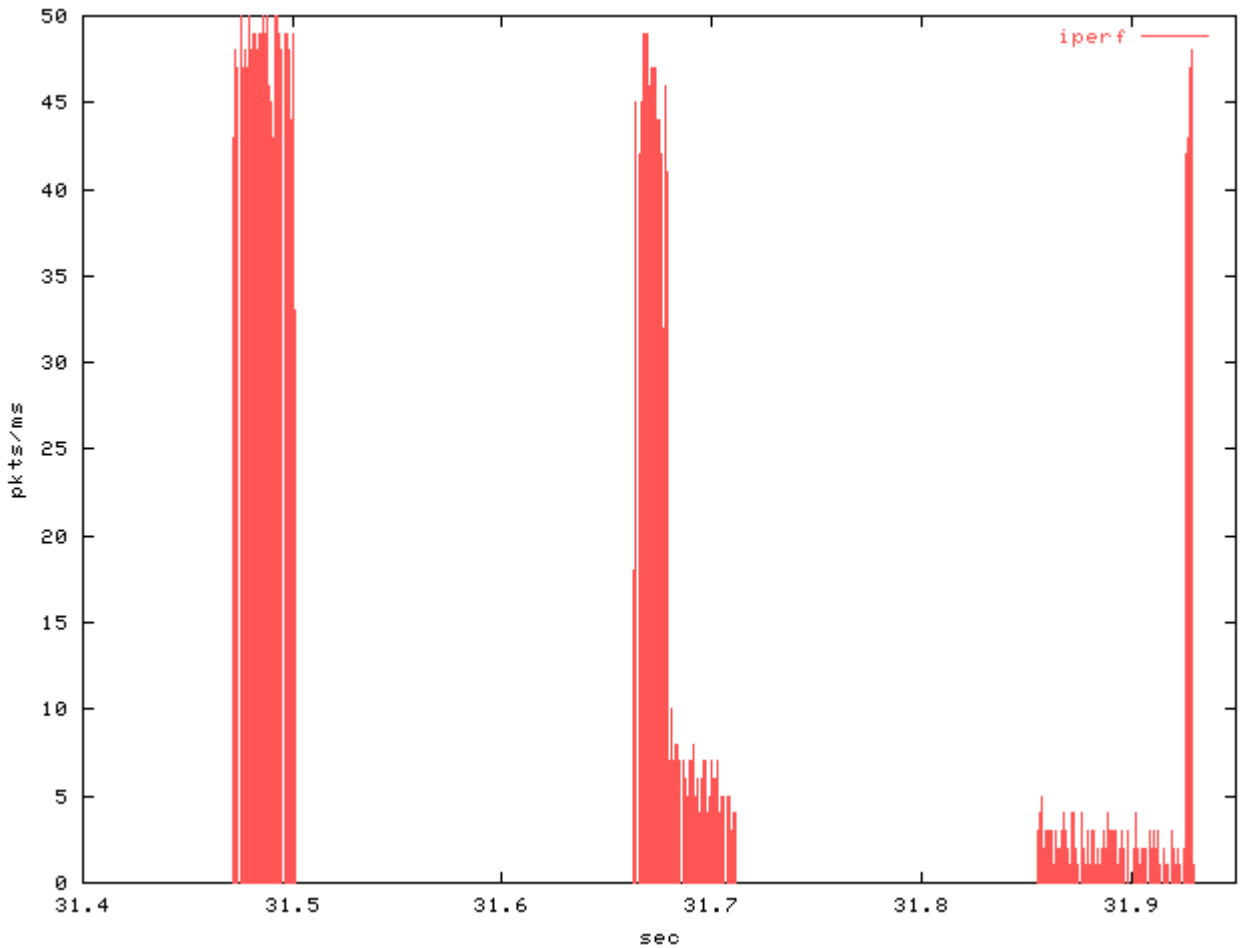
DR ギガアナライザ

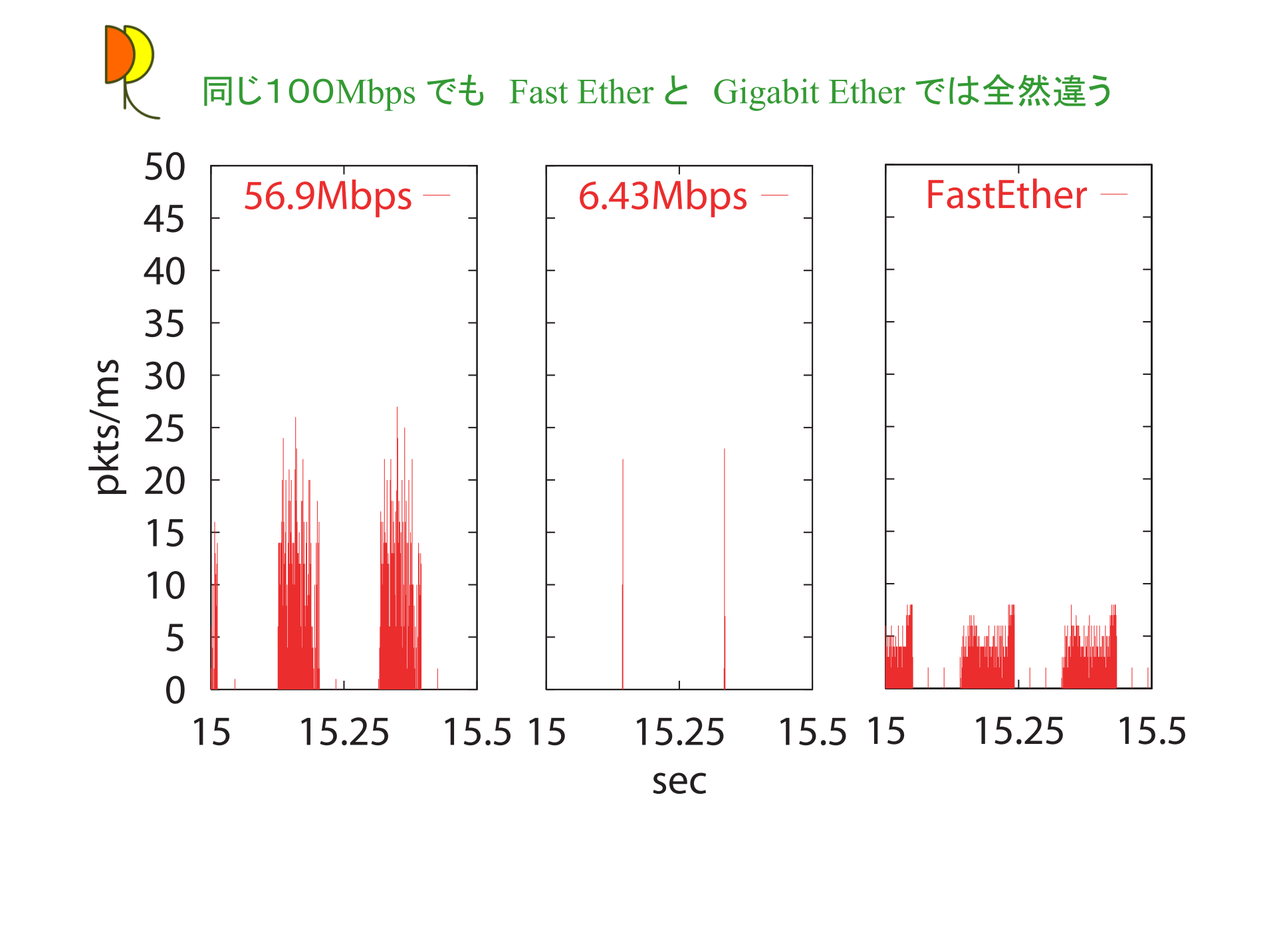

Optical Tap
Measured Network
1000BASE-SX
Packet Distribution Unit
(Linux Server)
Comet NIC +FPGA
Comet NIC +FPGA
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
Disk Server (iSCSI disk server)
GbE SWITCH

DR ギガアナライザによる
遠距離シングルストリームのパケット解析



拡大図


iperf
pkts/ms
sec
0
5
10
15
20
25
30
35
40
45
50
31.4
31.5
31.6
31.7
31.8
31.9


同じ100Mbps でも Fast Ether と Gigabit Ether では全然違う
56.9Mbps
6.43Mbps
FastEther
pkts/ms
50
45
40
35
30
25
20
15
10
5
0
15
15.25
15.5
sec

Transmission Rate Controlled TCP

バースト的なふるまいを減じることで （i.e., 遅くする）、
安定性を得て、結果的に全体を速くする。

シングルストリーム

IPG チューニング

IPG: イーサネットフレームのフレーム間ギャップ

イーサドライバ e1000 変更、8～1023バイト設定可能に

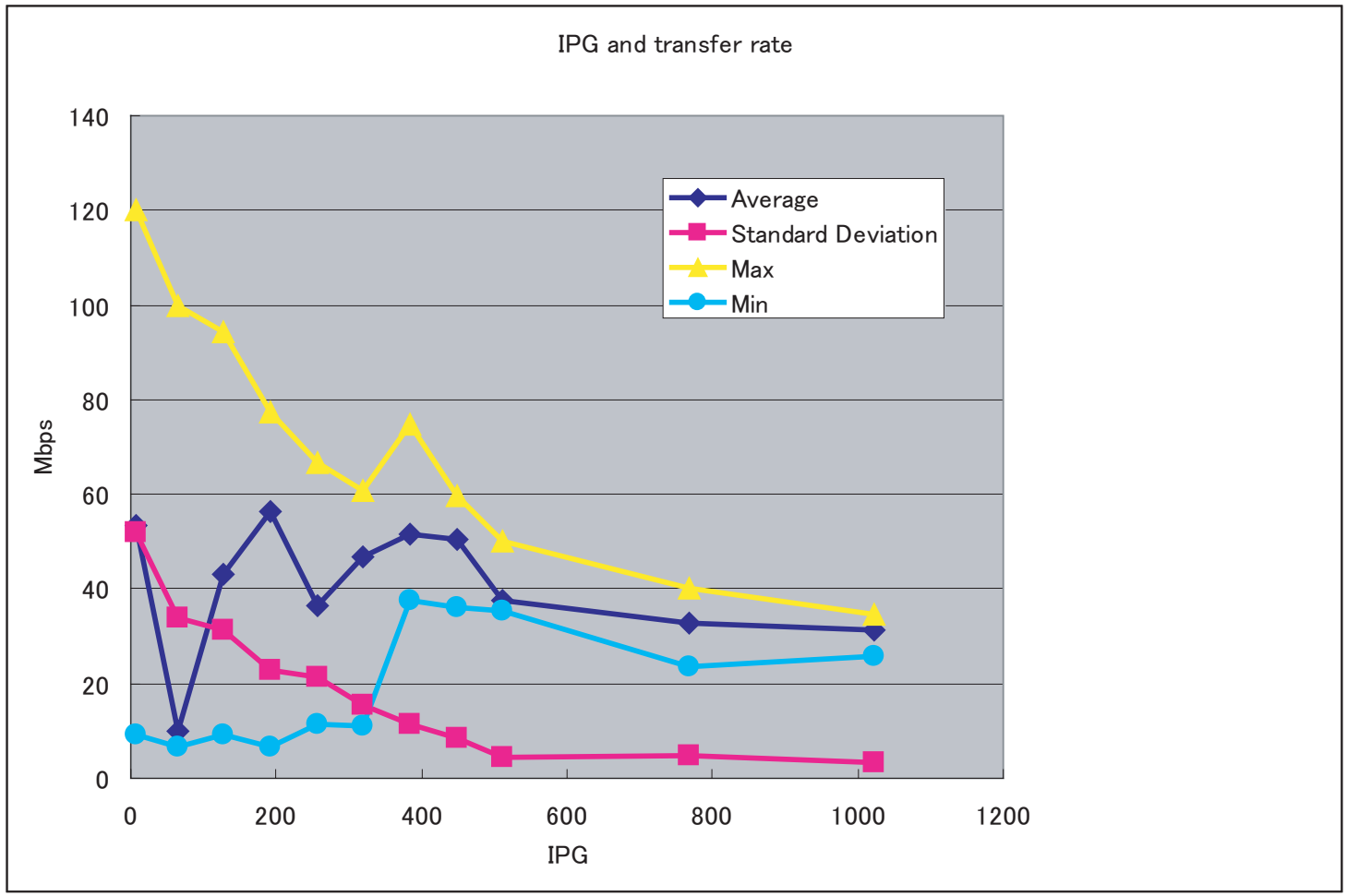
Clustered Spacing

Slow Start 時、指数的に速度が増える時に

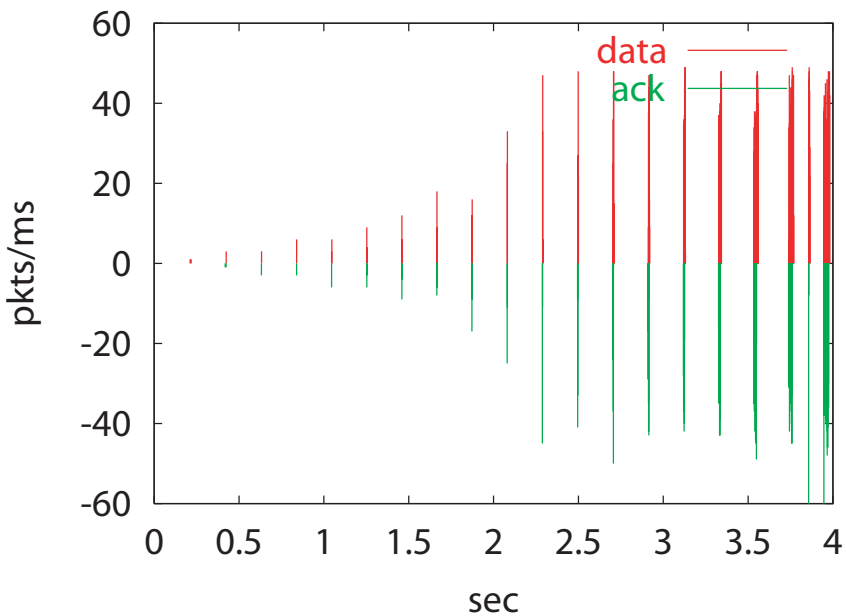
並列ストリーム

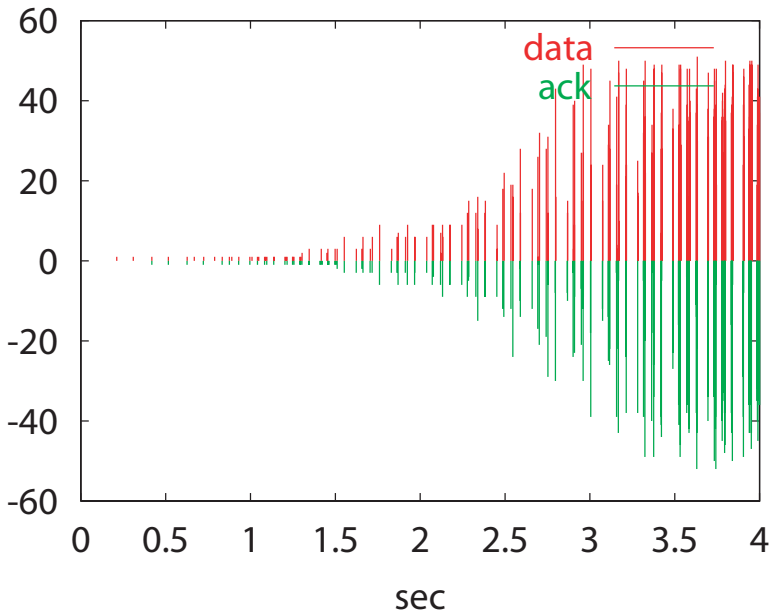
各コネクションのウィンドウ情報を収集して
ウィンドウサイズの上限を調整するインタフェイスを実装。
速いストリームを抑えることでウィンドウサイズ調整

IPG調整による　MAX,MIN,平均、標準偏差のグラフ



Clustered Spacing

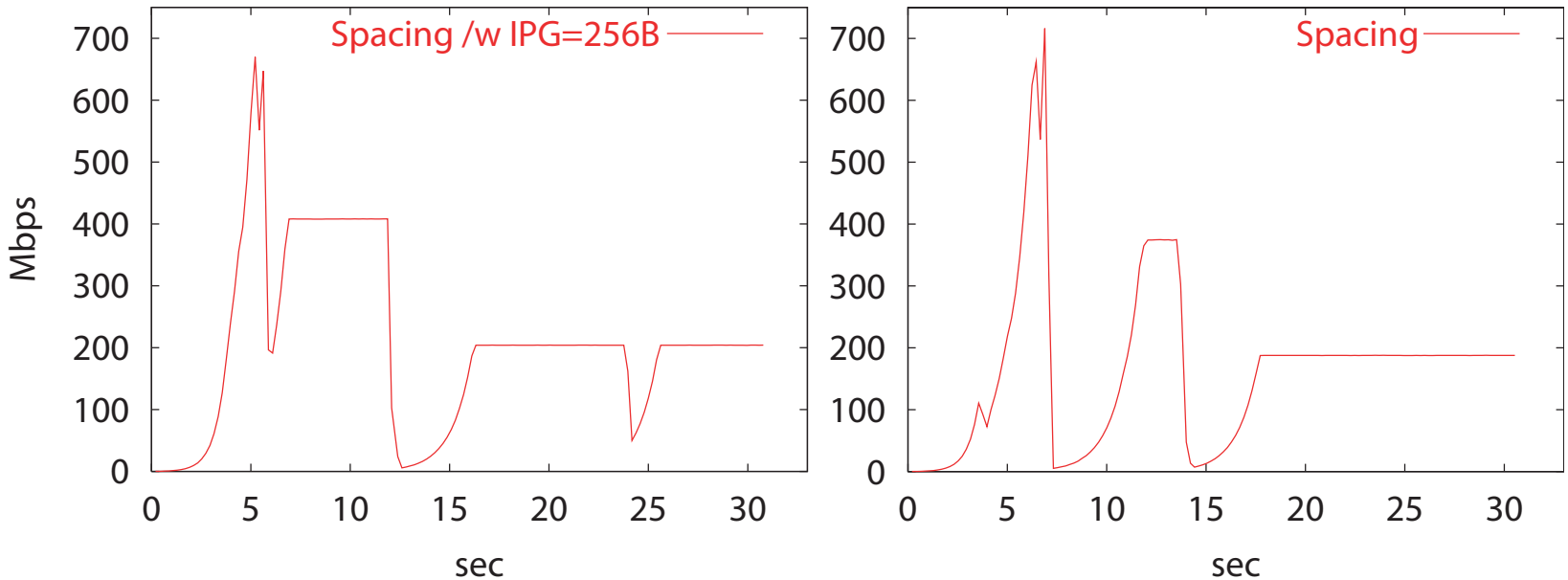




従来方式

Clustered Spacing

スループット


Spacing /w IPG=256B
Spacing
Mbps
sec
sec
0
100
200
300
400
500
600
700
0
5
10
15
20
25
30

日米１往復半実験　構成

片側　サーバ　32台

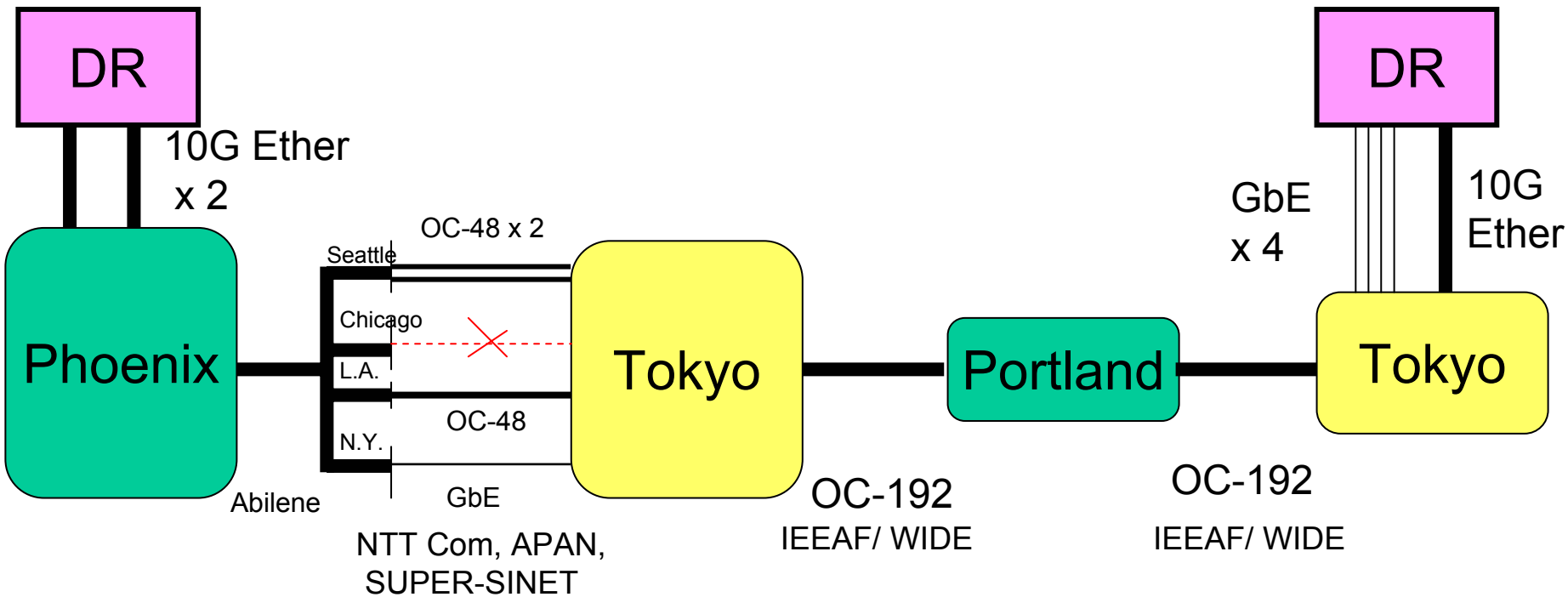
IBM x345, Dual Intel Xeon 2.40GHz, 2GBメモリ

Intel 82546EBオンボードNIC, Redhat Linux 7.3, Kernel 2.4.18

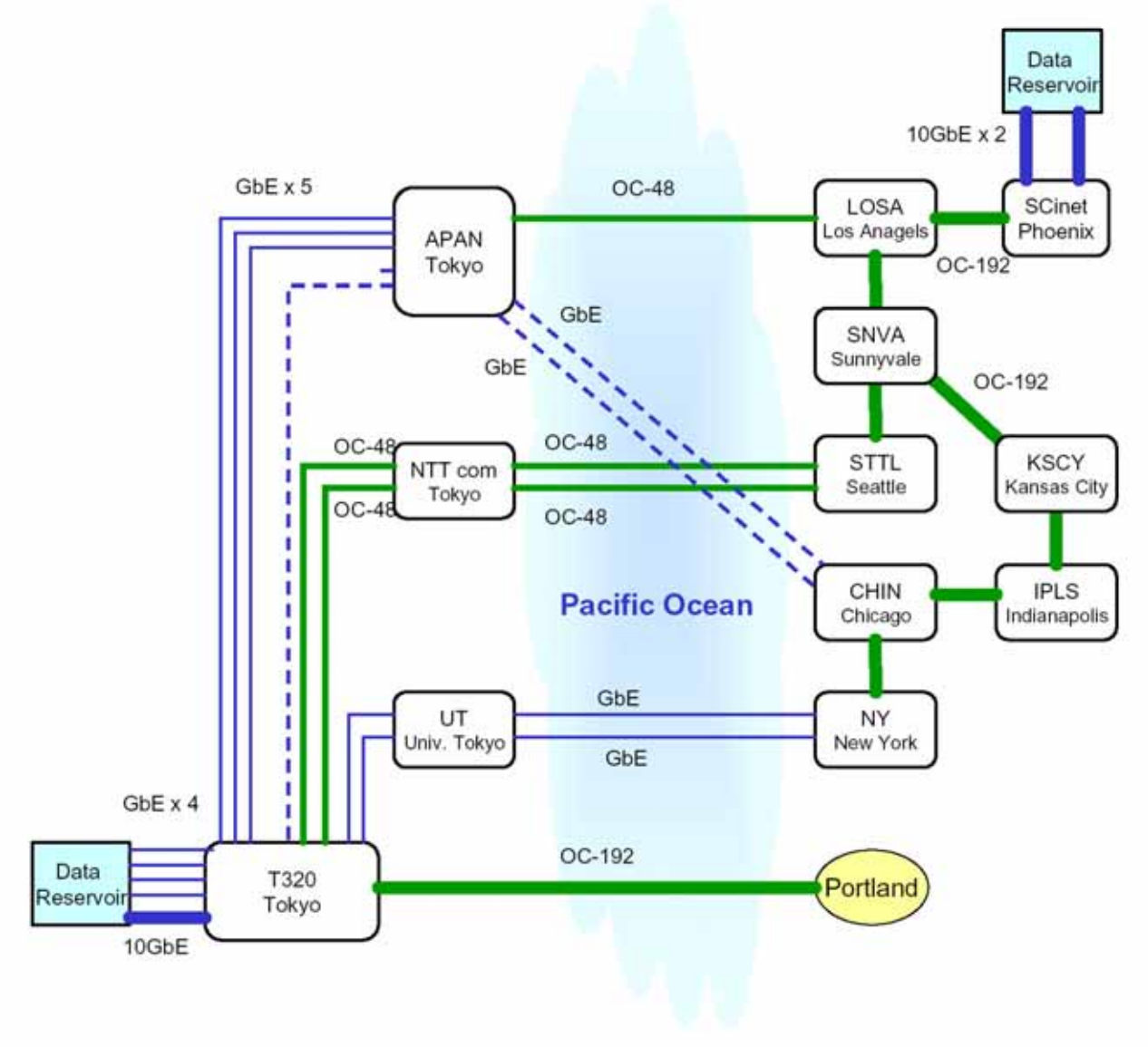
USAGI STABLE 20020408 で,

各ディスクサーバには，10,000rpm Ultra320　146GB SCSI HDD4台,

合計 18 ペタバイトのデータディスクを持つ.



ネットワーク構成


Data Reservoir
10GbE x 2
GbE x 5
APAN Tokyo
OC-48
LOSA Los Anagels
SCinet Phoenix
OC-192
GbE
GbE
SNVA Sunnyvale
OC-192
OC-48
OC-48
NTT com Tokyo
OC-48
OC-48
STTL Seattle
KSCY Kansas City
Pacific Ocean
CHIN Chicago
IPLS Indianapolis
GbE
GbE
UT Univ. Tokyo
NY New York
GbE x 4
Data Reservoir
T320 Tokyo
OC-192
Portland
10GbE

思いがけない事故

11/8　海底ファイバ切断事故

北まわりルートが利用不能に

復旧まで　２～３週間（ぎりぎり間に合わない）

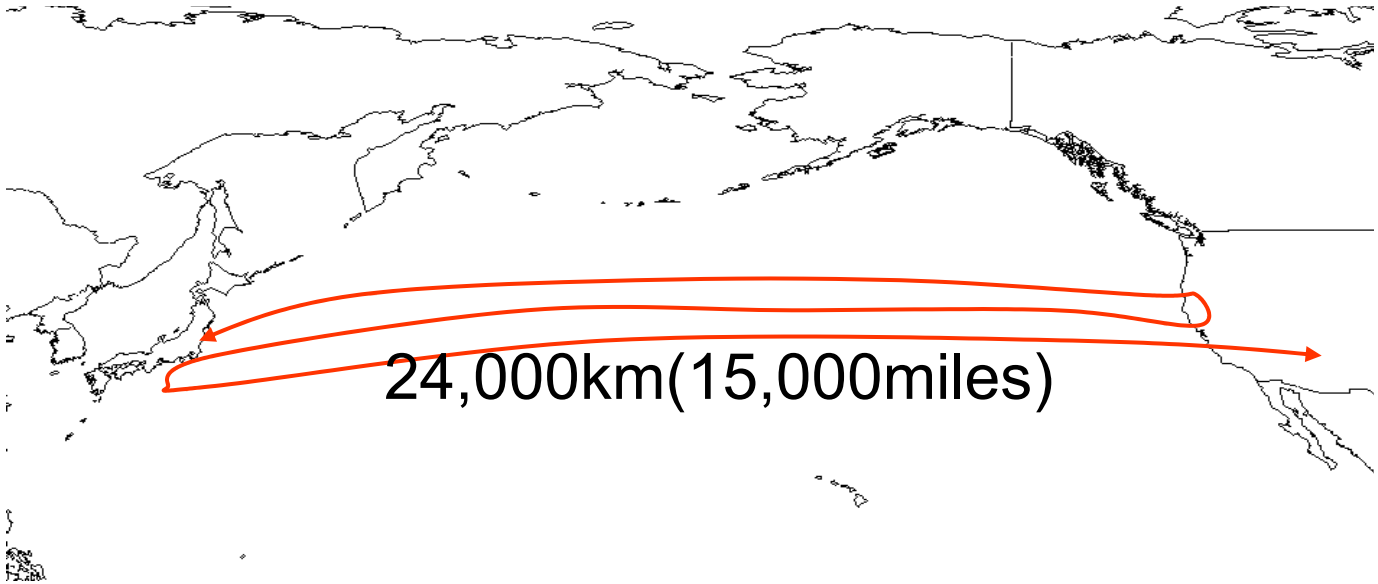
11.2 Gbps（OC48 x 3 + GbE x 4）　予定

↓

3.4 Gbps（OC48 x 1 + GbE x 1）11/8～

↓

8.2 Gbps（OC48 x 3 + GbE x 1）　11/18　本番２日前



OC-192

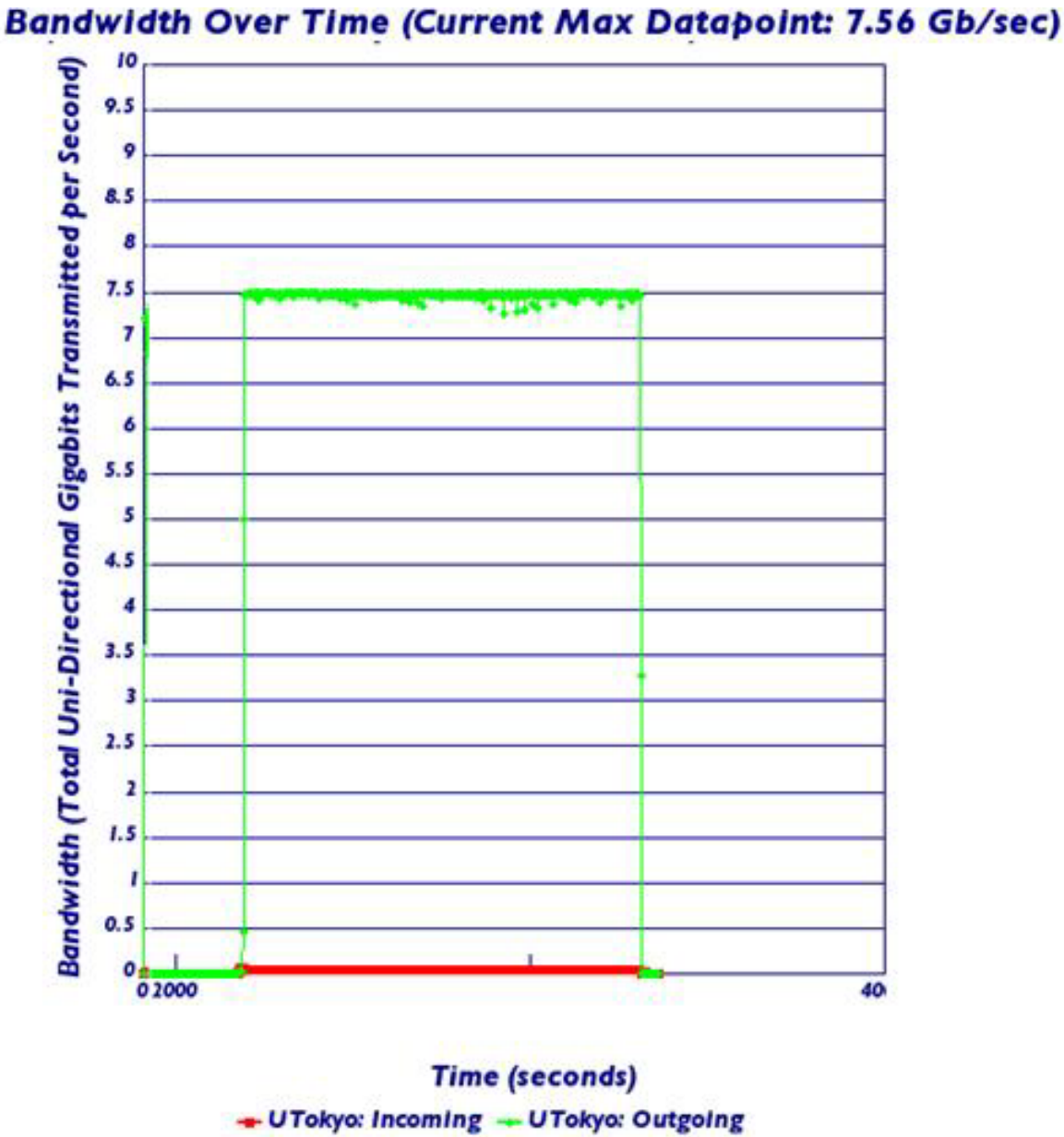
15,680km (9,800miles)

Juniper
T320

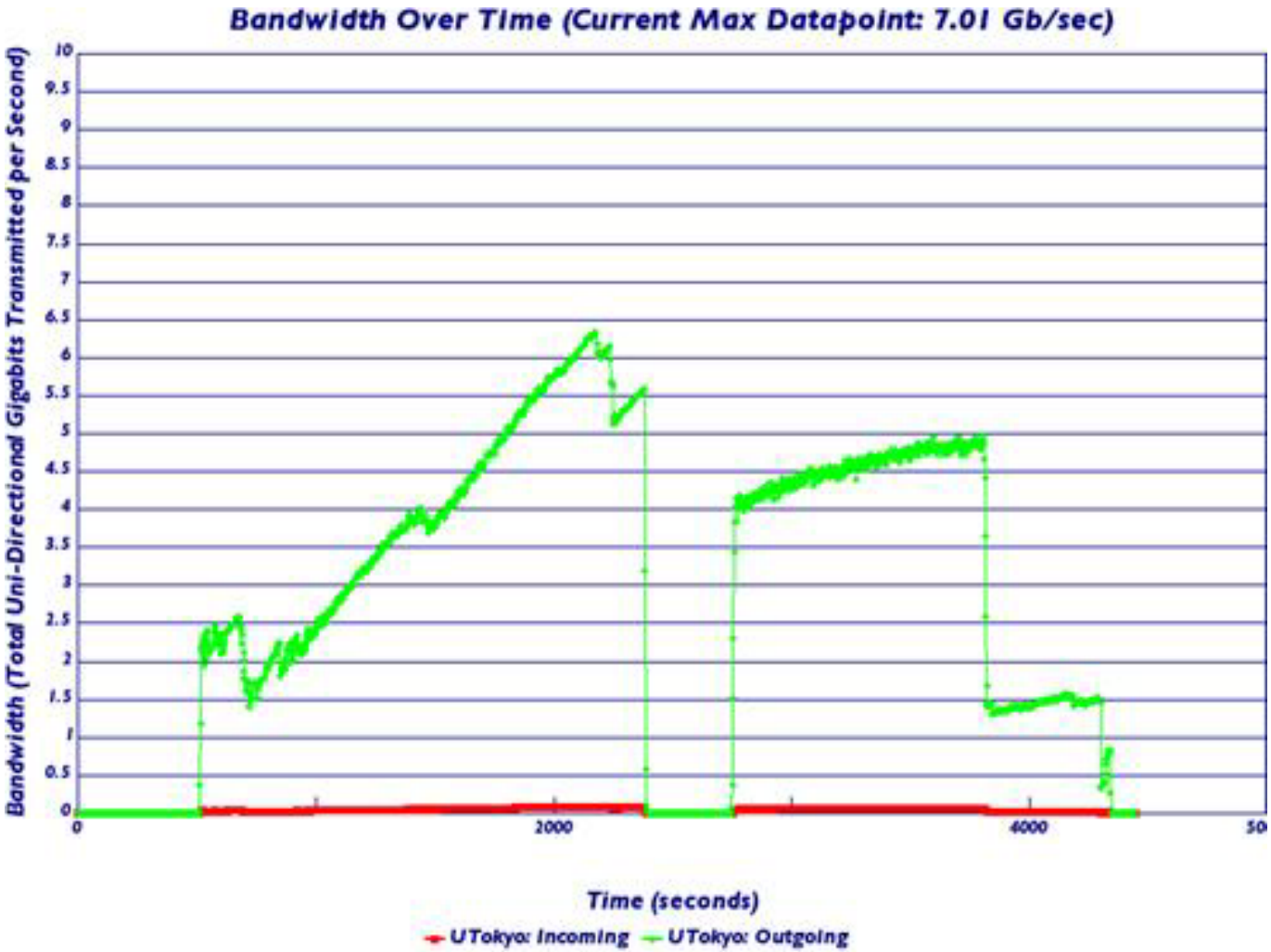
OC-48 x 3
GbE x 4

8,320km
(5,200miles)

ハードウェアによる TCP 終端方式



ソフトウェアによる Transmission Rate Controled TCP



サマリー

- Low-level protocol (iSCSI)
  - Latency hiding by Distributed Shared File architecture
  - Efficient utilization of high-speed network for single file transfer
  - File system transparency by the use of low-level protocol
- Scalability
  - Scalability to network bandwidth
  - Scalability to the size of the storage
  - Scalable computing system
- Experimental Results
  - 95% utilization of network bandwidth up to 1600km
  - 50% performance gain by the use of low-level protocol to conventional file transfer
  - Realization of 11.7 Gbps using 24 disks

Data Reservoir project is supported by Special Coordinated Fund for Science and Technology, Ministry of Education, Culture, Sports, Science and Technology, Japan.

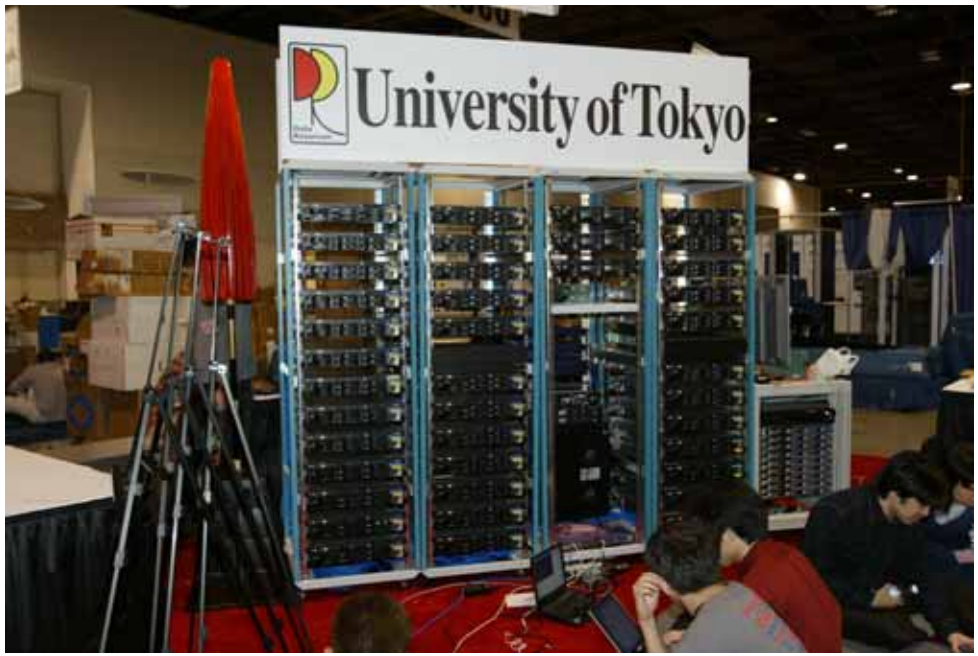
１年間で

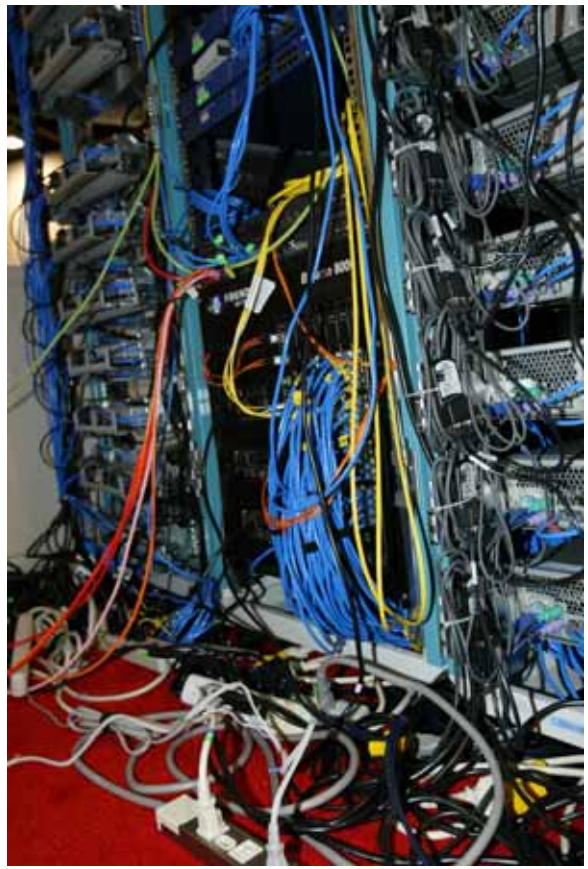
- BWC2002
  - Tokyo → Baltimore　10,800km (6,700miles)
  - Peak bandwidth (on network)　600 Mbps
  - Average file transfer bandwidth　560 Mbps
  - Bandwidth-distance products　6,048　terabit-meters/second

- BWC2003
  - Tokyo → Portland → Tokyo → Phoenix　24,000 km (15,000 miles)
  - Peak bandwidth (on network)　8.2Gbps
  - Average file transfer bandwidth　7.5Gbps
  - Bandwidth-distance products　181,440 terabit-meters/second

  - More than 25 times improvement from BWC2002 performance (bandwidth-distance products)

実験･展示の舞台裏　･･･　本当に大変





日米データ転送実験にご協力いただいた方々

東京大学情報基盤センター	加藤朗，　関谷勇司
米国富士通研究所カレッジパーク	松尾和洋，益岡竜介
東京大学大学院情報理工学系研究科	山本成一
NTT コミュニケーションズ	村上満雄, 福田健平, 長谷部克幸
IEEAF	Don Riley
KDDI	小西和憲, 北辻佳憲
国立情報学研究所	浅野正一郎, 松方純, 藤野貴之

日米24,000km 実験にご協力いただいた組織

NTT / VERIO

WIDE PROJECT

APAN

Juniper NETWORKS

IEEAF

FOUNDRY NETWORKS

SUPER SINET

Cisco Systems

tyco / Telecommunications