

Grid Datafarm for a HEP application

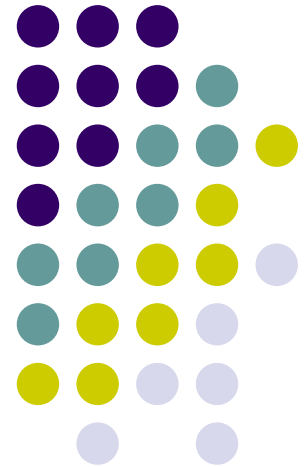
Osamu Tatebe

(Grid Technology Research Center, AIST)

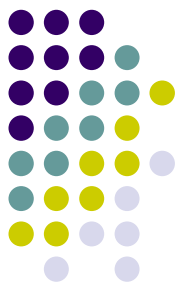
Satoshi Sekiguchi(AIST), Youhei Morita (KEK),
Satoshi Matsuoka (Titech & Nii), Kento Aida (Titech),

Donald F. (Rick) McMullen (Indiana),

Philip Papadopoulos (SDSC)



SC2002 high-performance bandwidth challenge



- demonstrate exciting applications at the maximum possible speed

アドレス http://scinet.supercomp.org/ 移動 リンク

Premier Sponsors
Force 10 **SCinet** World's Fastest Network

Information for Exhibitors

- 2002 Service Offerings
- Connection Requests
- SCinet Architecture
- Security Policy
- Wireless Service Policy
- Service Level Policy
- Sponsors & Volunteers
- Xnet
- reserved

2002 Bandwidth Challenge

- Call for Participation
- Applicants Summary
- Final Results

Connection Requests
[Manage Your Account](#)

Contact
[Registration](#)
[Bandwidth Challenge](#)
[Xnet](#)
[Helpdesk](#)
[Webmaster](#)

Previous Years
[2002 \(Baltimore\)](#)
[Photos](#)
[2001 \(Denver\)](#)

- Final Invoices for Exhibitor Connection Requests Due.
- 2002 [Bandwidth Challenge](#) Shatters Previous Records.

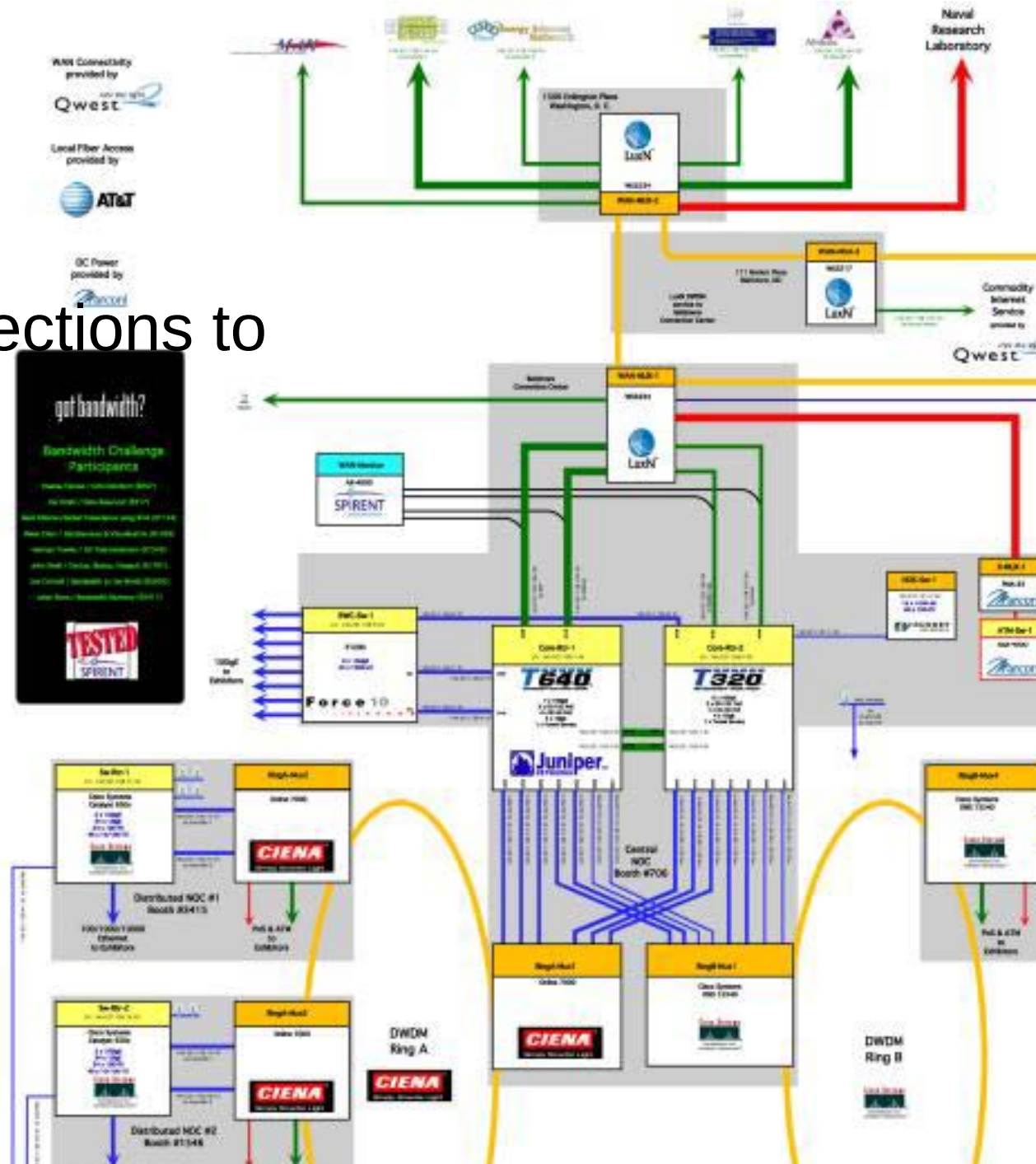


SCinet Network Operations Center, under construction, Dallas, TX, November 2000

SCinet is the high-performance network built to support the annual [International Conference for High Performance Computing and Communications](#) (SC). The SC Conference Series is co-sponsored by [ACM SIGARCH](#) and the [IEEE Computer Society](#). SCinet features both a high-performance production-quality network as well as an extremely high performance experimental network, Xnet.

SCinet

- 10 Gbps connections to Abilene, TeraGrid, . . .
- WAN-Monitor: Spirent AX-4000



Overview of Our *Challenge*

- Seven clusters in US and Japan comprise a cluster-of-cluster file system:

Gfarm file system

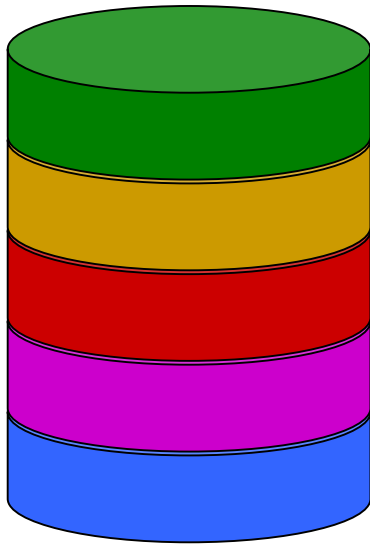
- The on demand

Bandwidth Challenge!!!!

terabyte of raw data) in the *Gfarm file system*

- The Gfarm file system replicates data across the clusters

グリッド技術 (グリッドデータファーム) を用いた テラバイト (1兆文字) クラスの大規模データ



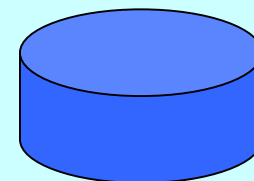
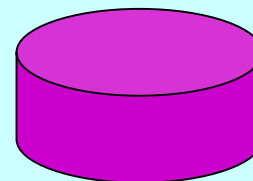
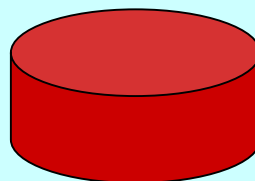
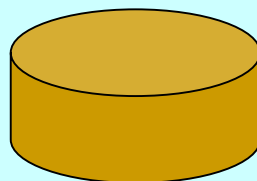
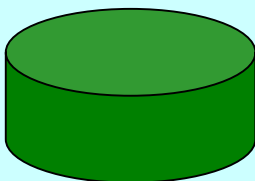
一ヶ所に保存すると

- ▶ 保存箇所以外からの書き込み・読み出しに時間がかかる
- ▶ 災害時などにデータが失われたり、アクセスできなくなったりする



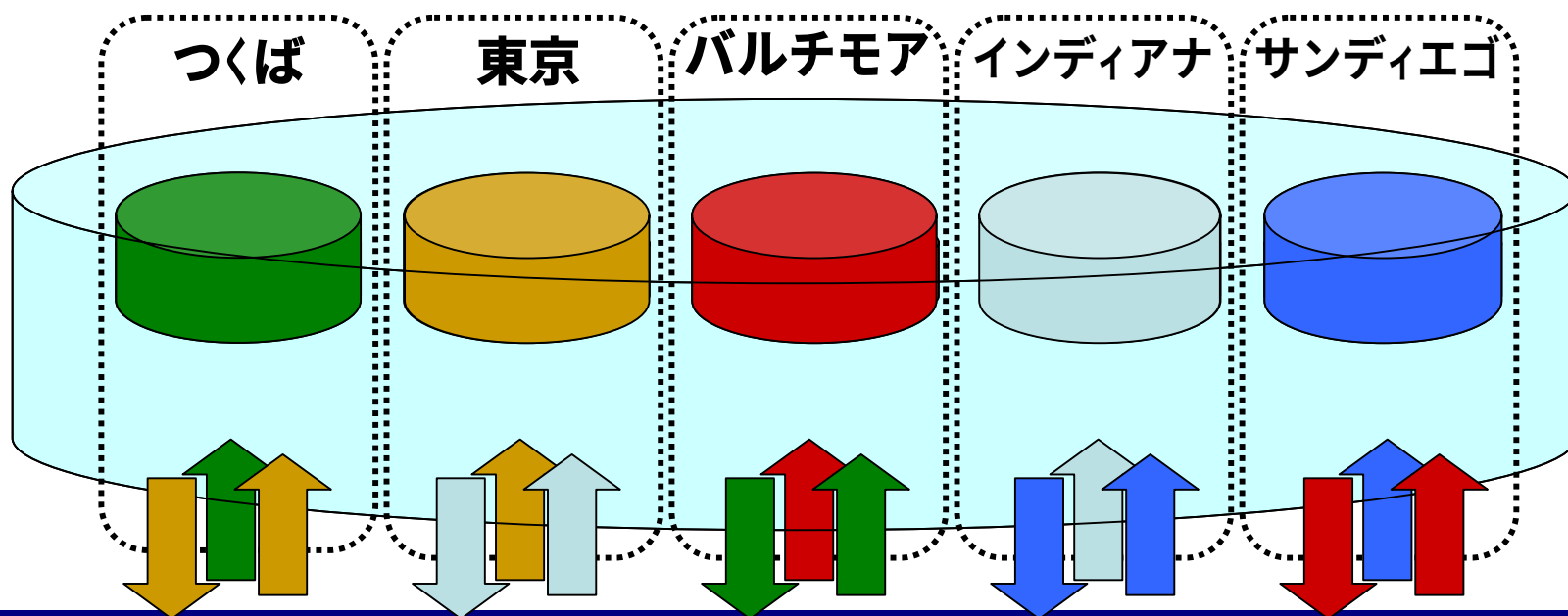
分けて別々に保存すると

- ▶ 別々の管理単位になり、管理、アクセスが大変
- ▶ 手元にないデータへの読み書きは時間がかかる
- ▶ 災害時などには部分的にデータが失われたり、アクセスできなくなったりする



AIST グリッドデータファームを使うと

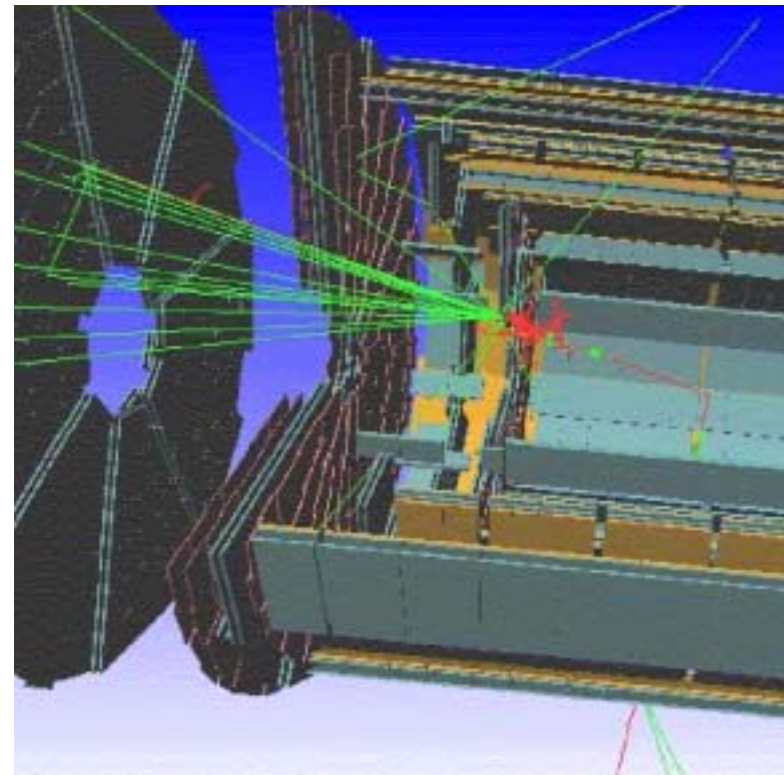
- 複数分散ディスクによりファイルシステムを構成
- 分散して書き込まれた複数ファイルを一つのファイルとして扱える
- 部分的な複製ができる
 - ▶ 複数の複製により障害に備える
- データがある場所でそのデータを高速処理する



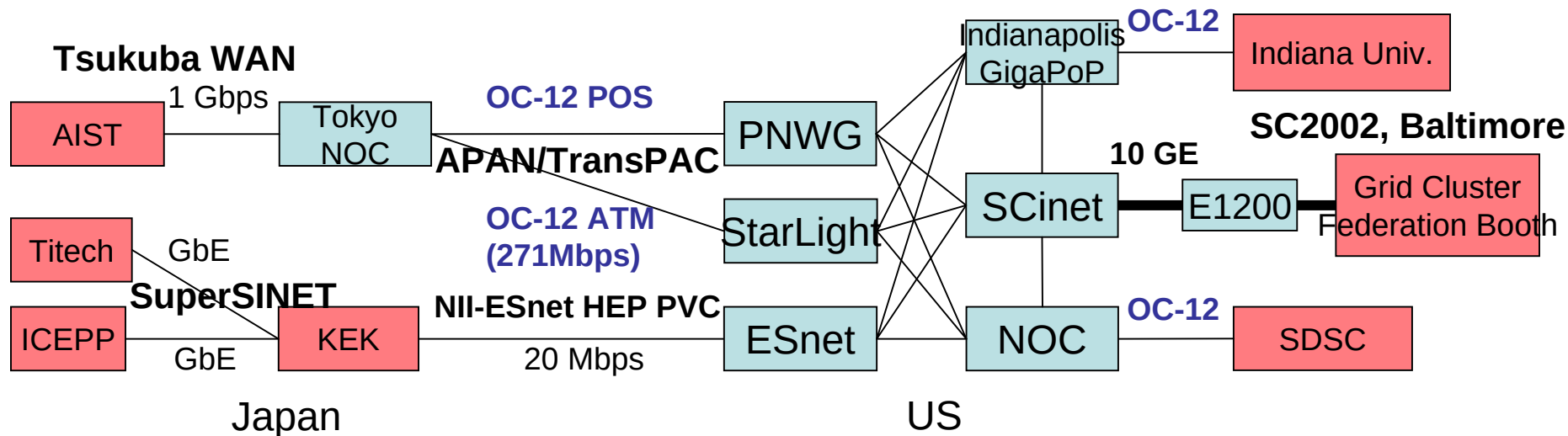
Target Application at SC2002: FADS/Goofy



- ⌘ Monte Carlo Simulation Framework with Geant4 (C++)
- ⌘ FADS/Goofy: Framework for ATLAS/Autonomous Detector Simulation / Geant4-based Object-oriented Folly
<http://atlas.web.cern.ch/Atlas/GROUPS/SOFTWARE/OO/domains/simulation/>
- ⌘ Modular I/O package selection:
Objectivity/DB and/or ROOT I/O
on top of Gfarm filesystem
with good scalability
- ⌘ CPU intensive event simulation
with high speed file replication
and/or distribution



Network and cluster configuration for SC2002 Bandwidth Challenge

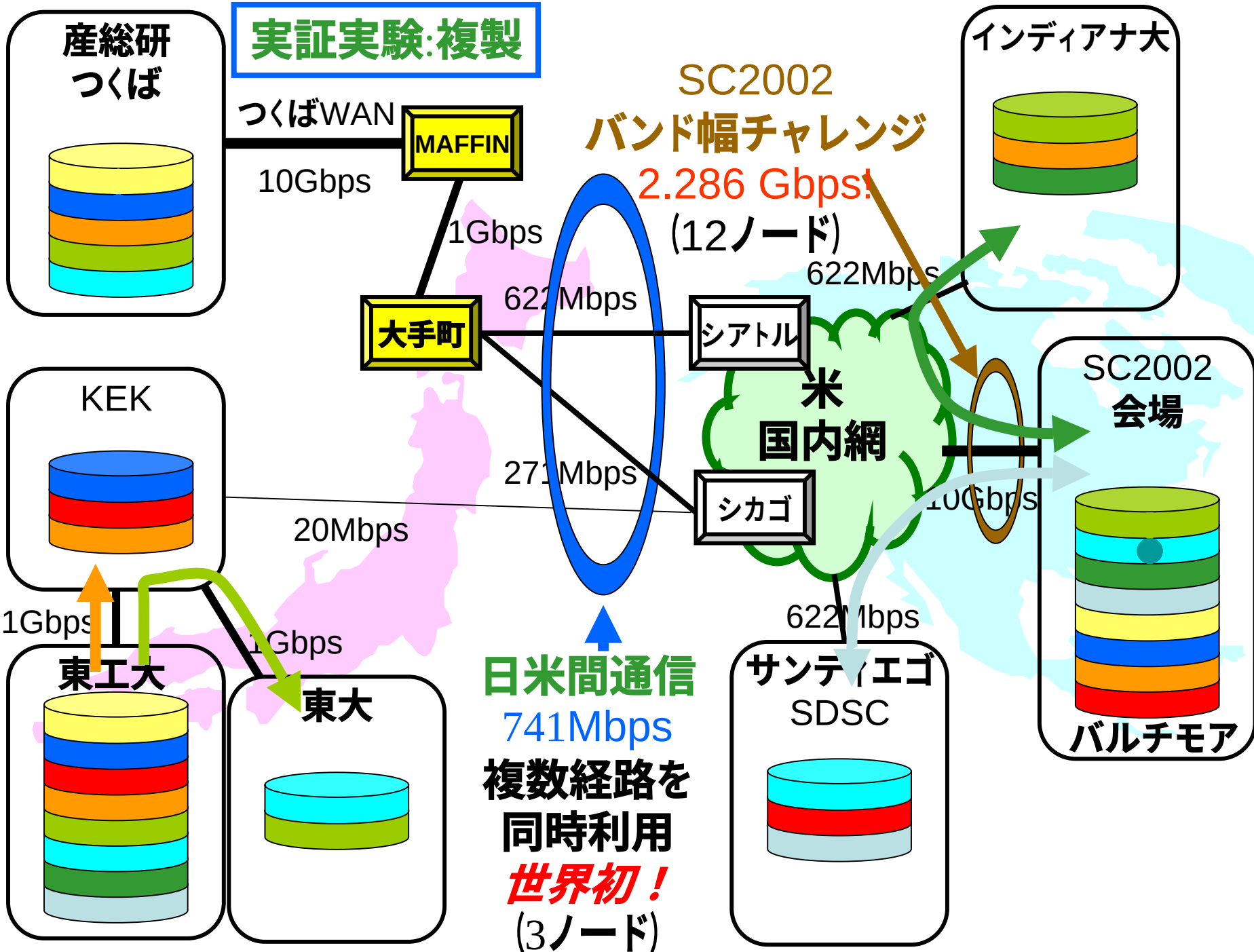


SC会場とのデータ転送の理論総バンド幅 2.137 Gbps

KEK Titech AIST ICEPP SDSC Indiana U SC2002



Total disk capacity: 18 TB, disk I/O bandwidth: 6 GB/s
Peak CPU performance: 962 GFlops



AIST ネットワーク、クラスタ構成の特徴

SC会場

- ▶ GbEで接続された12ノードのPCクラスタをForce10 E1200で10GEでSCinetに接続
- ▶ LANにおける性能
 - ④ ネットワークバンド幅は930Mbps
 - ④ ファイル転送性能は75MB/s(=629Mbps)

AIST

- ▶ 同等の7ノードのクラスタ。GbEでつくばWAN、Maffinを経て東京XPに接続

Indiana大

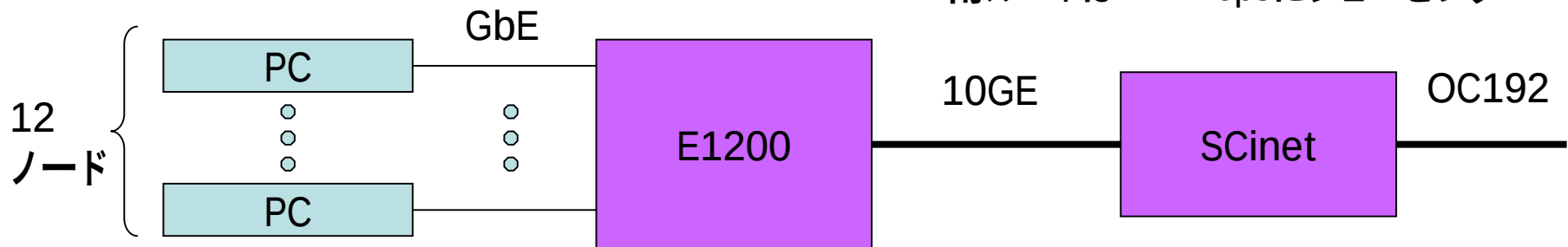
- ▶ FEで接続された15ノードのクラスタ。OC12でIndianapolis GigaPoPに接続

SDSC

- ▶ GbEで接続された8ノードのクラスタ。OC12で外部接続

TransPACの南北ルートのルーティング

- ▶ デフォルトは北ルート
- ▶ SC会場、AISTのそれぞれ3ノードについて南ルートを通るよう東京XP、Maffinで設定
- ▶ AIST、SC会場間のRTT: 北ルート 199ms、南ルート 222ms
- ▶ 南ルートは271Mbpsにシェーピング



SC2002会場のネットワーク構成

● **Large latency, high Bandwidth (aka LFN)**

- ▶ Big socket size for large congestion window
- ▶ Fast window-size recovery after packet loss
 - ⌚ High Speed TCP (internet draft by Sally Floyd)
- ▶ Network striping

● **Packet loss due to real congestion**

- ▶ Transfer control

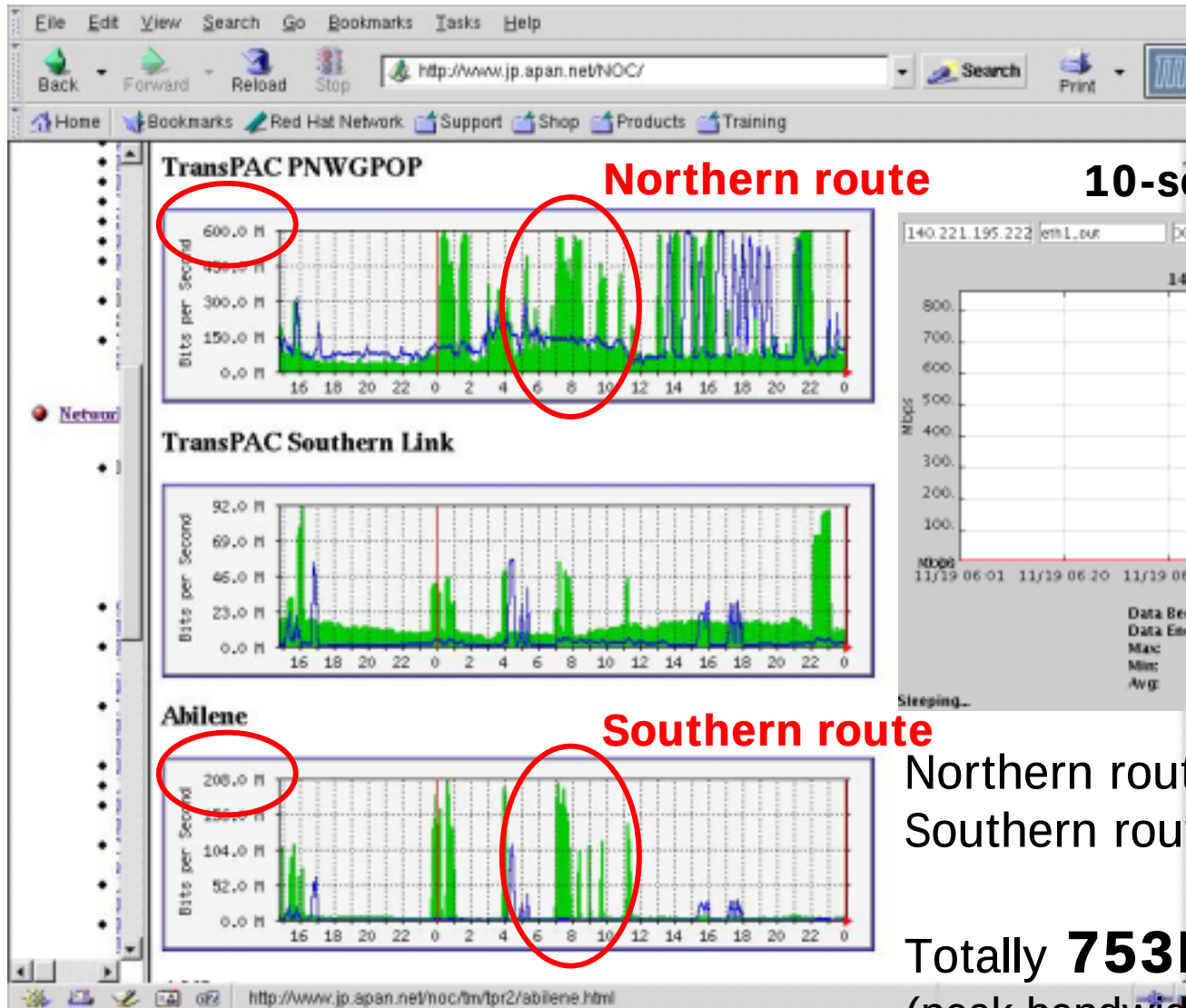
● **Poor disk I/O performance**

- ▶ 3ware RAID with 4 3.5" HDDs on each node
- ▶ Over 115 MB/s (~ 1 Gbps network bandwidth)
- ▶ Network striping vs. disk striping access
 - ⌚ # streams, stripe size

● **Limited number of nodes**

- ▶ Need to achieve maximum file transfer performance

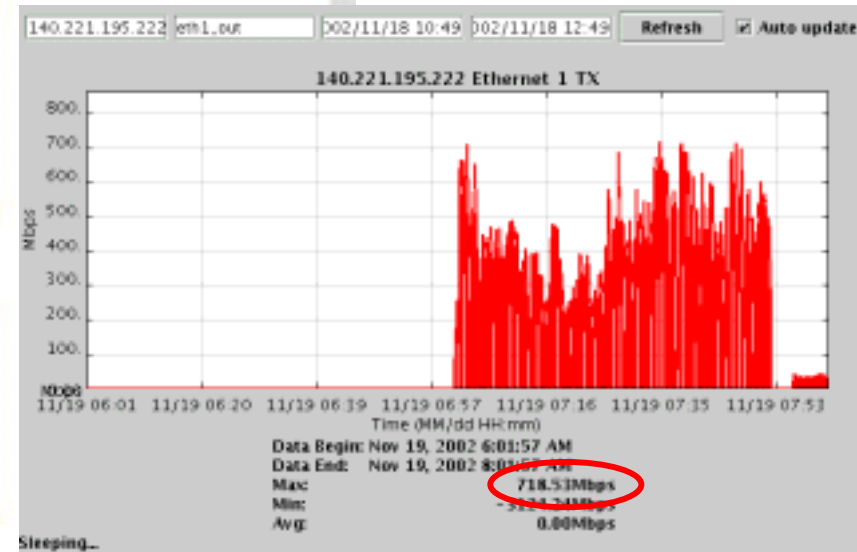
Bandwidth measurement result of TransPAC by Iperf



5-min average bandwidth

Northern route

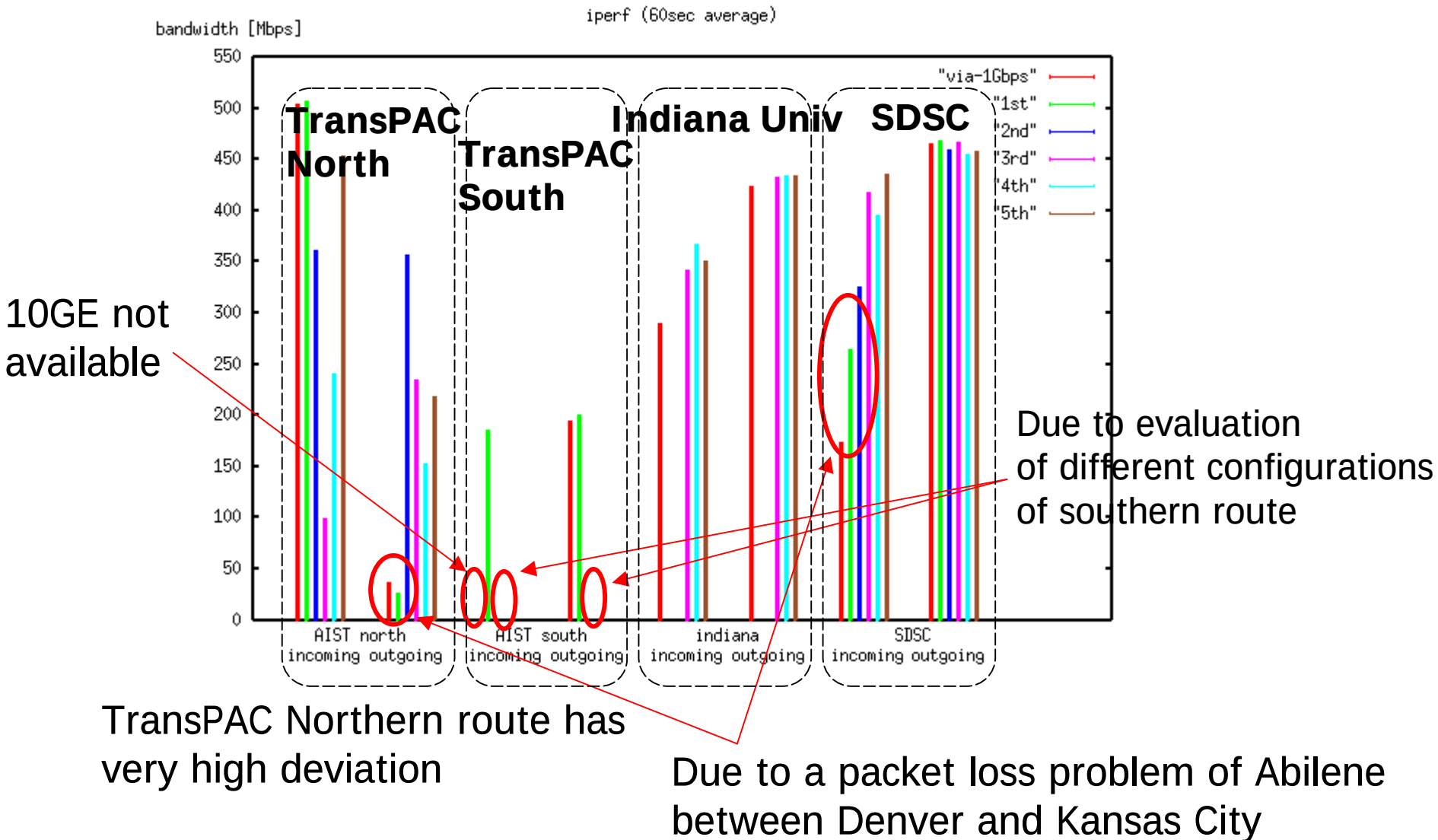
10-sec average bandwidth



Northern route: 2 node pairs
Southern route: 3 node pairs
(100Mbps streams)

Totally **753Mbps** (10-sec avg)
(peak bandwidth: 622+271=893Mbps)

Bandwidth measurement result between SC-booth and other sites by Iperf (1-min average)

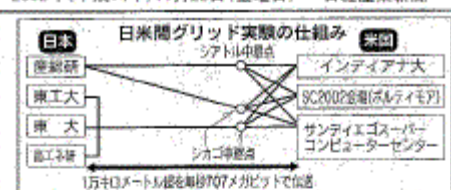


AIST 日米間のグリッド実験

産経新聞12月1日

日経産業新聞11月22日

2002年(平成14年)11月22日(金曜日) 日経産業新聞



複数結ぶ「グリッドコンピューティング」

日米間通信で最速記録

毎秒707

通信実験に成功した。多数のコンピュータをネットワークでつなぎ、スーパーコンピュータの能力を最大限に活用する。日米間の通信速度を向上させる。グリッドコンピューティングの実現に近づく。成果は、米国のスーパーコンピュータに近づく。成果は、米国のスーパーコンピュータに近づく。成果は、米国のスーパーコンピュータに近づく。



日本経済新聞(タ刊) 2002年(平成14年)11月21日(木曜日)

高速通信に成功

産総研と米インディアナ大の共同研究チームは、日米間の通信速度を向上させる。グリッドコンピューティングの実現に近づく。成果は、米国のスーパーコンピュータに近づく。成果は、米国のスーパーコンピュータに近づく。

日本経済新聞11月21日タ刊

読売新聞11月21日タ刊

読売新聞(タ刊) 2002年(平成14年)11月21日(木曜日)

◆毎秒8000万文字分のデータ転送

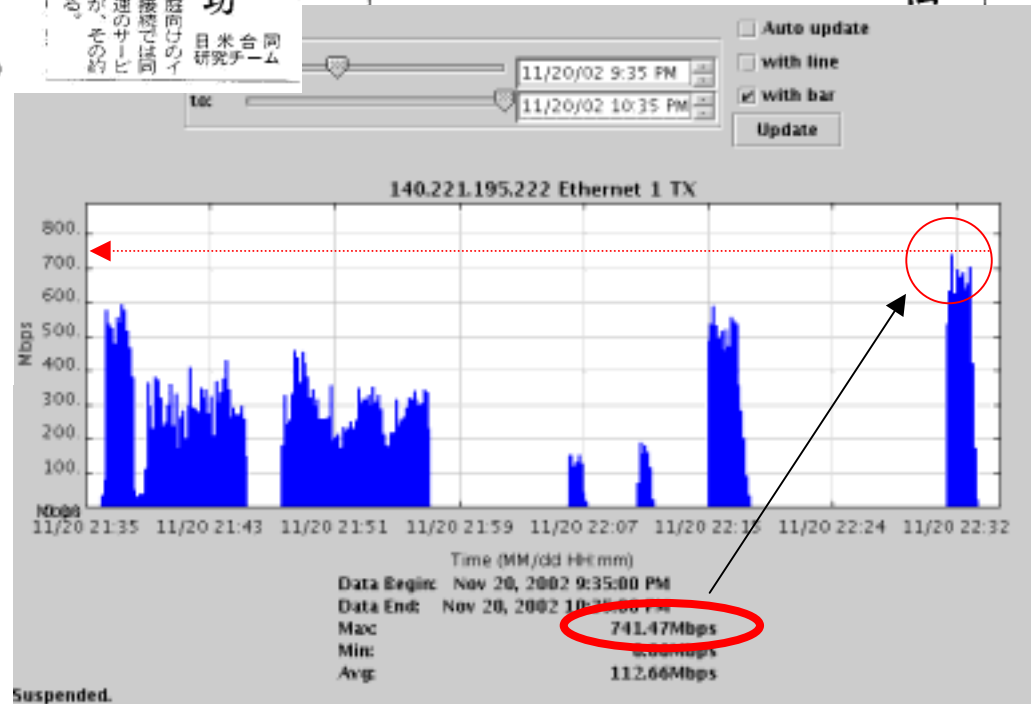
産総研と米インディアナ大の共同研究チームは、日米間の通信速度を向上させる。グリッドコンピューティングの実現に近づく。成果は、米国のスーパーコンピュータに近づく。成果は、米国のスーパーコンピュータに近づく。

産経新聞 2002年(平成14年)12月1日(日曜日)

産総研 超大規模データ転送

独立行政法人「産業技術総合研究所」(茨城県つくば市)は、日米七カ所を拠点として、日米間の通信速度を向上させる。グリッドコンピューティングの実現に近づく。成果は、米国のスーパーコンピュータに近づく。成果は、米国のスーパーコンピュータに近づく。

グリッドコンピューティング 世界初の実証



米国4ノード、日本4ノード、ファイル転送で 741 Mbps 達成!

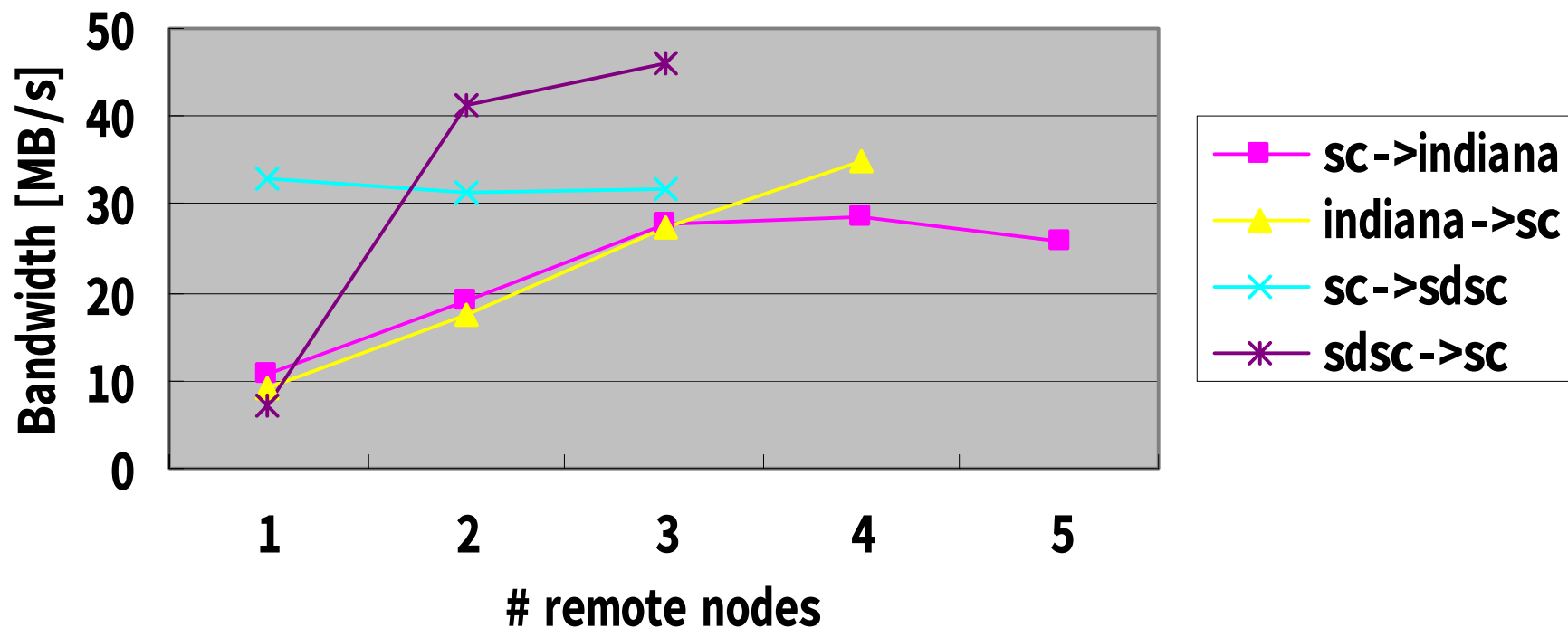
(理論ピーク性能 893 Mbps、10秒平均)

TransPACにおけるファイル転送の設定と結果

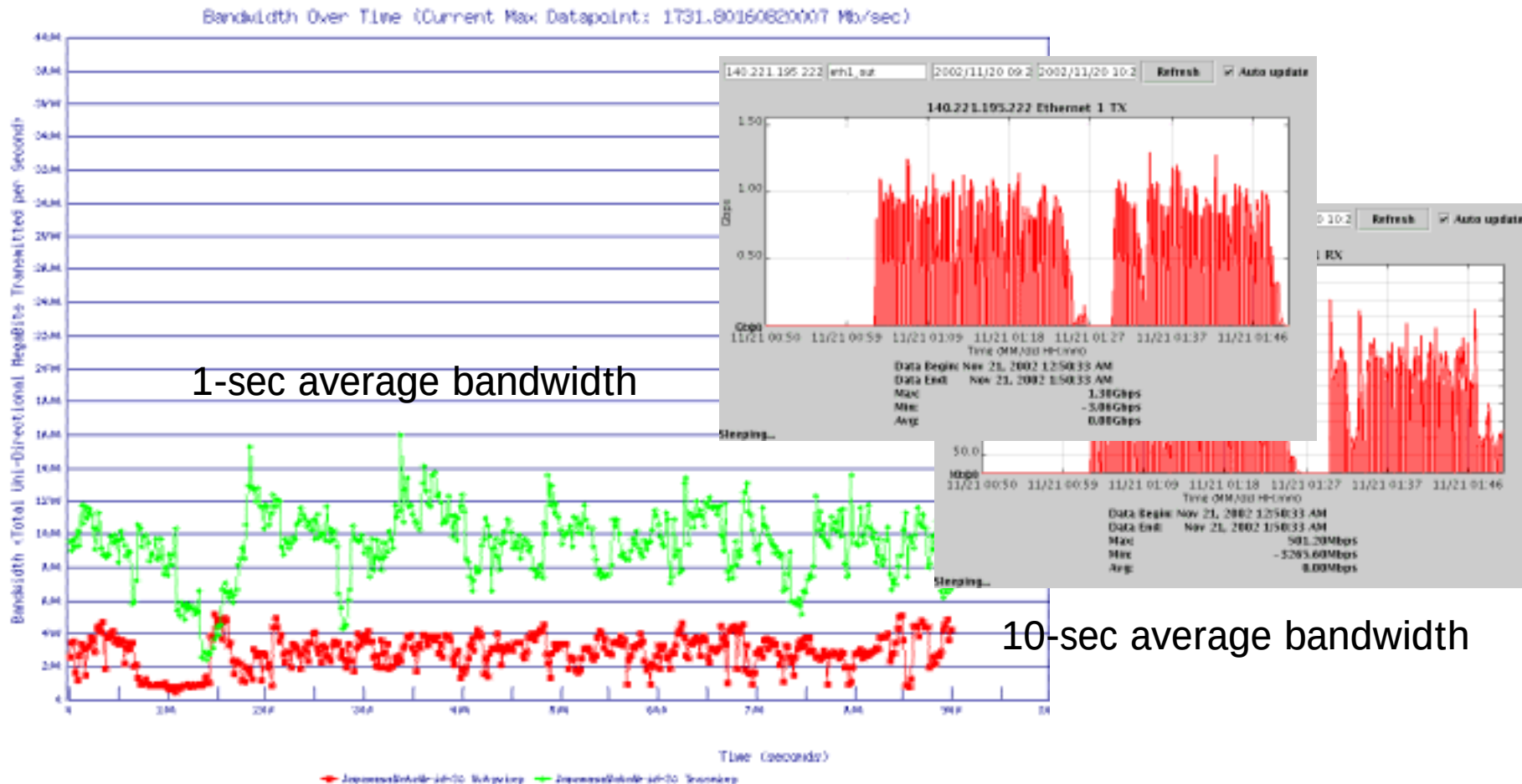
Host pair (内訳)	streams (内訳)	10秒平均最大	転送時間 (秒)	平均
1 (北1)	16 (北16x1)		113.0	152Mbps
2 (北2)	32 (北16x2)	419 Mbps	115.9	297Mbps
3 (北3)	48 (北16x3)	593 Mbps	139.6	369Mbps
4 (北3 南1)	56 (北16x3 + 南8x1)	741 Mbps	150.0	458Mbps

設定パラメータ	北回り	南回り
socket buffer size:	610 KB	250 KB
流量制限 1 ストリームあたり:	50 Mbps	28.5 Mbps
ストリーム数:	16 streams	8 streams
ホスト数:	3 hosts	1 host
stripe unit size:	128 KB	128 KB

Replication bandwidth [SC 1 node-US sites]



SC02 Bandwidth Challenge Result



We achieved 2.286 Gbps using 12 nodes!
(outgoing 1.691 Gbps, incoming 0.595 Gbps)

Special thanks to

- ❑ Rick McMullen, John Hicks (Indiana Univ, PRAGMA)
 - ❑ Phillip Papadopoulos (SDSC, PRAGMA)
 - ❑ Hisashi Eguchi (Maffin)
 - ❑ Kazunori Konishi, Yoshinori Kitatsuji, Ayumu Kubota (APAN)
 - ❑ Chris Robb (Indiana Univ, Abilene)
 - ❑ Force 10 Networks, Inc
-