

ペタバイトスケールデータインテンシブ
コンピューティングのためのGrid
Datafarmアーキテクチャ
<http://datafarm.apgrid.org/>

建部修見

産業技術総合研究所グリッド研究センター

On behalf of the Gfarm project

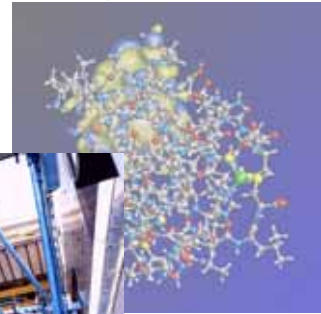


ペタスケールデータコンピューティング

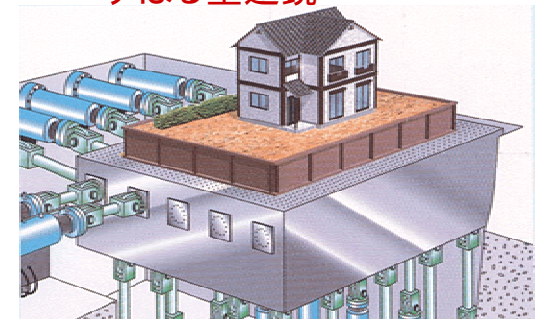
- Petascale Data Intensive Computing

- **大規模データ計算科学、データマイニング**
 - 高エネルギー物理学、粒子物理学
 - 天文台、地球惑星
 - 生命情報工学 ...

- **大規模ビジネスデータベース**
 - e-Japan、電子政府、電子商取引
 - データウェアハウス
 - 検索エンジン



すばる望遠鏡



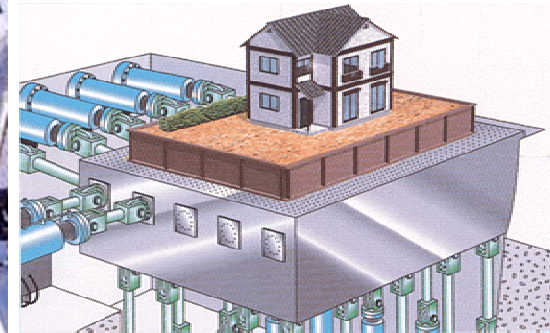
3次元地震シミュレータ

Data Grid Projects in Japan

- ATLAS/Grid Datafarm
 - AIST, KEK, Titech, UTokyo, . . .
 - New Data Grid Architecture for Petascale data-intensive computing and its reference implementation (Gfarm)
- Japanese Virtual Observatory
 - NAO, Titech, AIST, . . .
 - Distributed databases, Common access method to multiwavelength databases, Statistical analysis
- NARC, Agriculture
- RIKEN, JAIST, Genome Informatics
- Potential Projects
 - Bosai, Earthquake measurement
 - NASDA, SELENE Luna exploration



Subaru Telescope



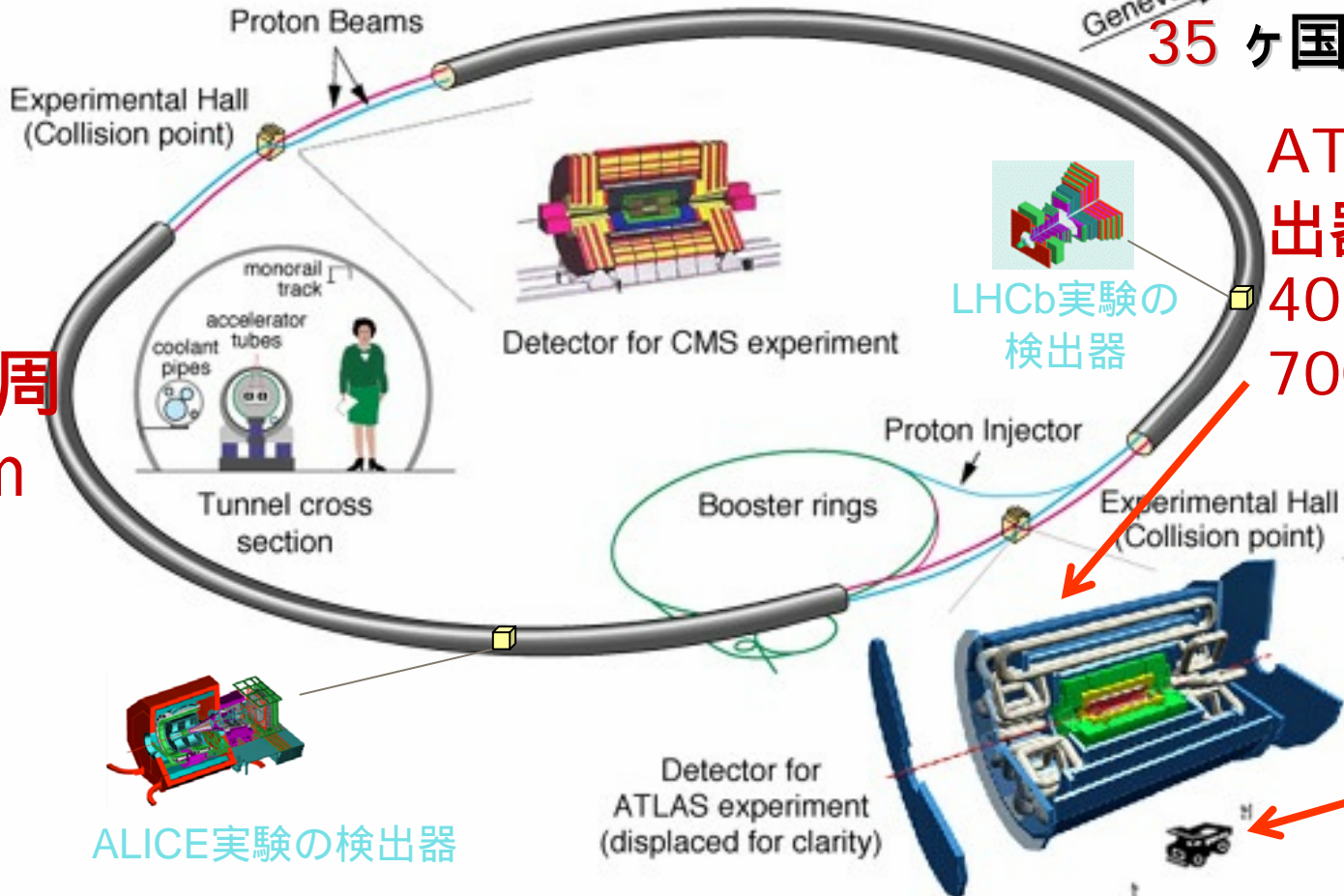
3D earthquake simulator in MIKI

例: CERN Large Hadron Collider 加速器実験

Large Hadron Collider at CERN

Circumference 26.7 km (16.6 miles)

~2000 物理学者
35 ヶ国
Geneva



ATLAS検出器
40m x 20m
7000 トン

トラック

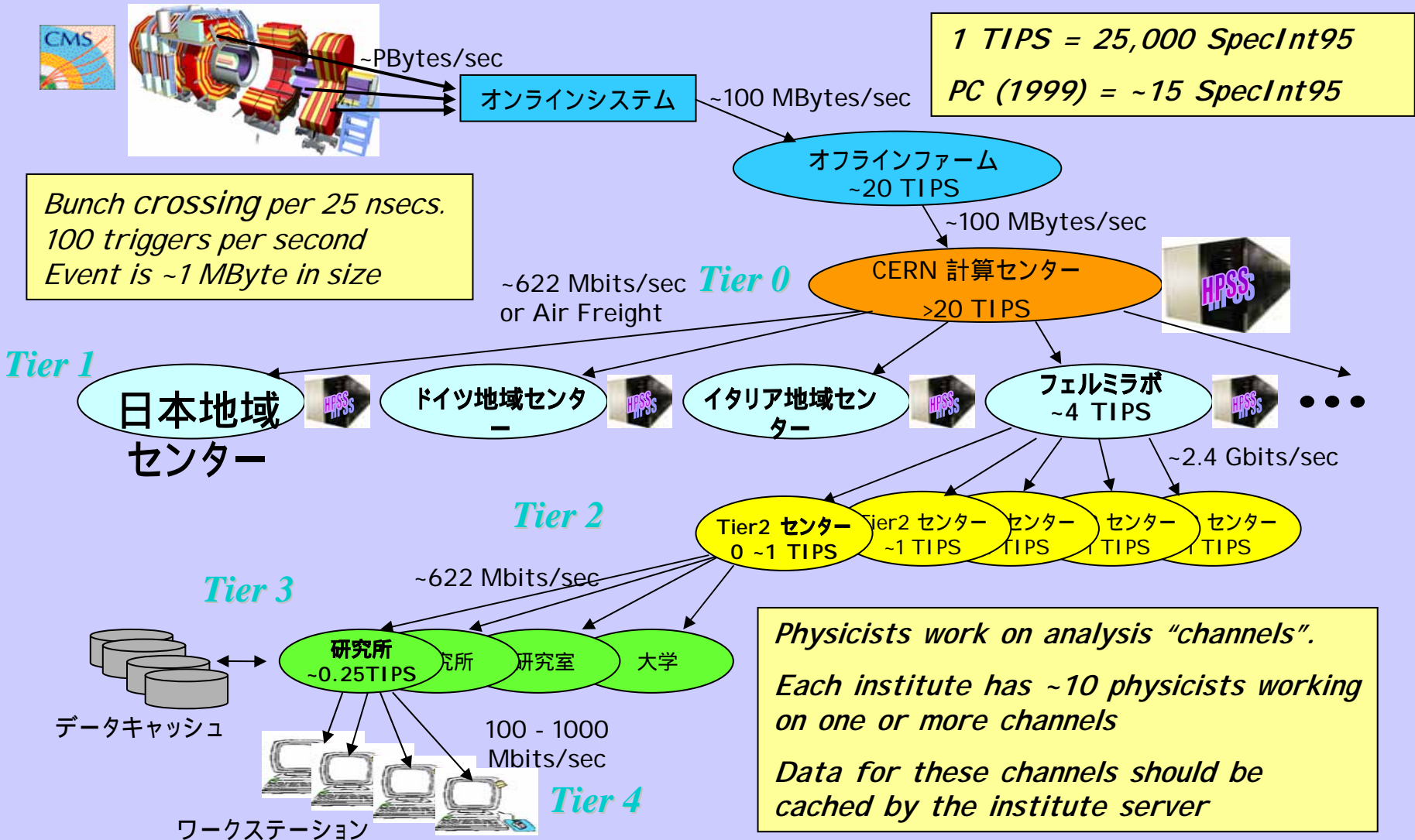
ALICE実験の検出器

LHC円周
26.7km

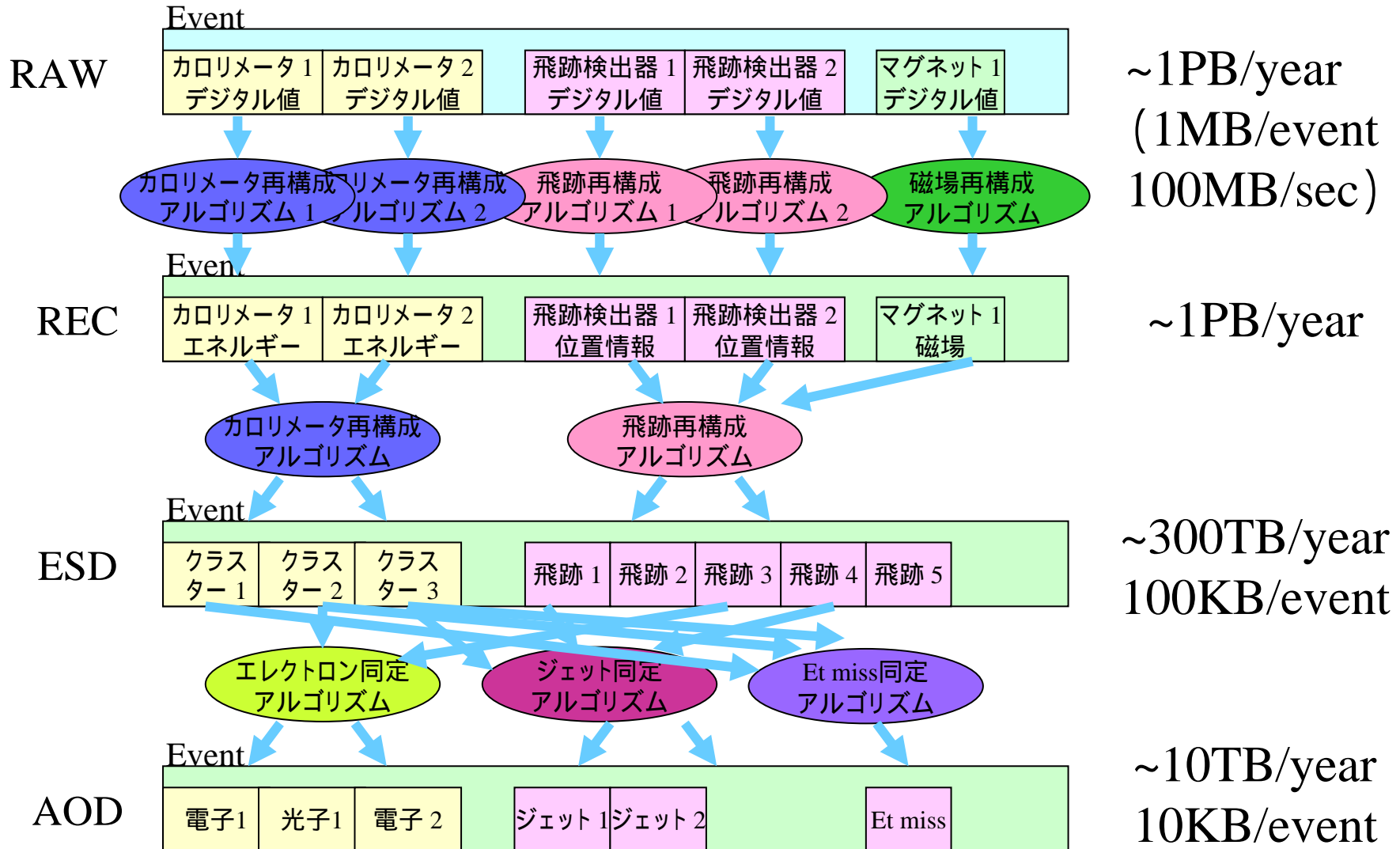


解析モデルスキーム (LHC ATLAS, CMS, ...)

● 資源階層 (Tier0, Tier1, Tier2, ...)



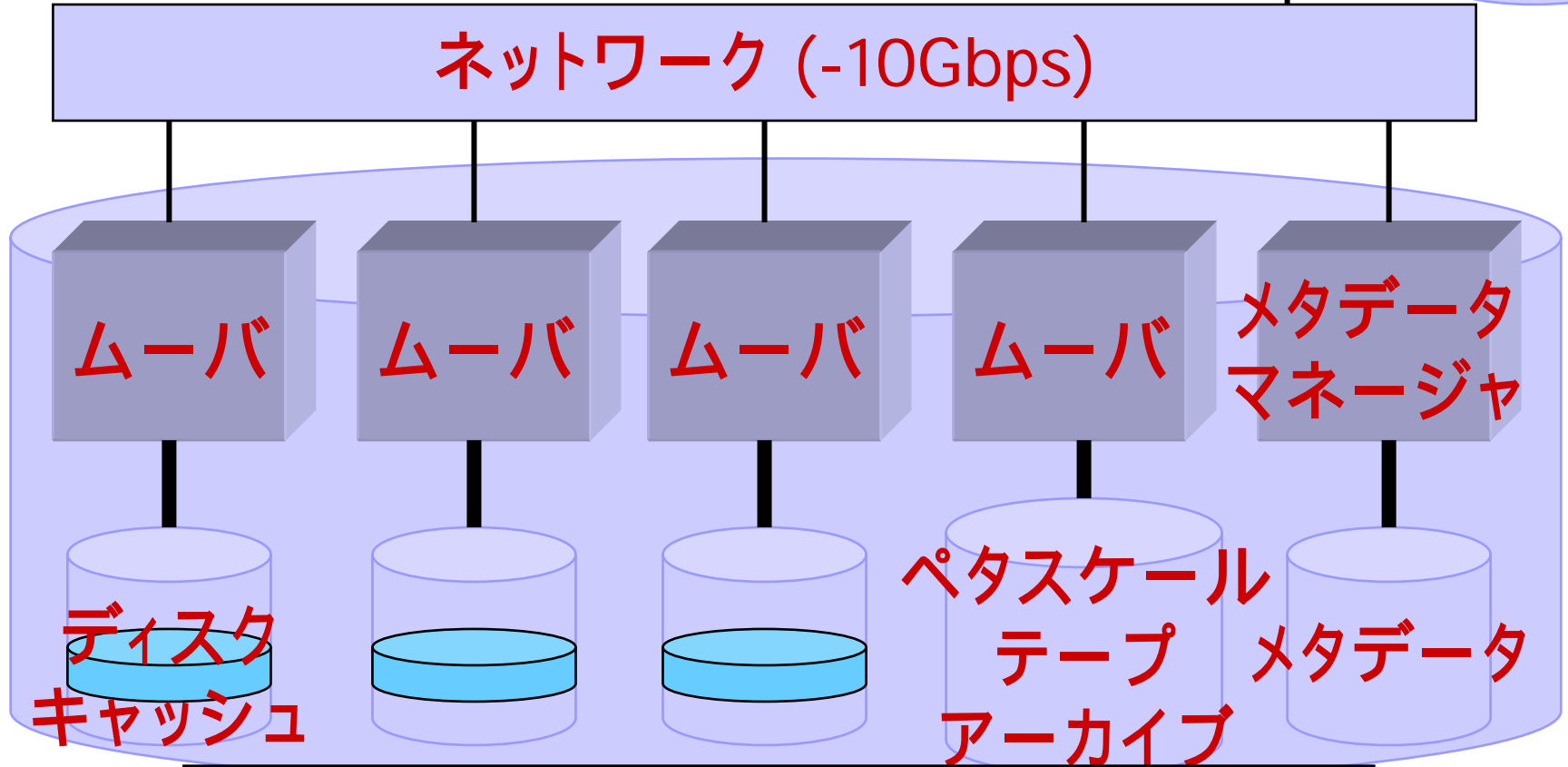
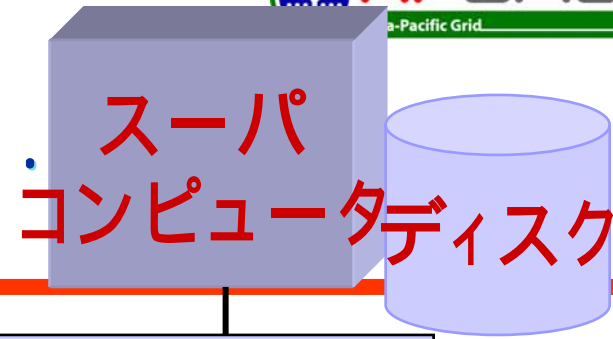
高エネルギーデータ解析の流れ



ペタスケールデータコンピューティング における要求項目

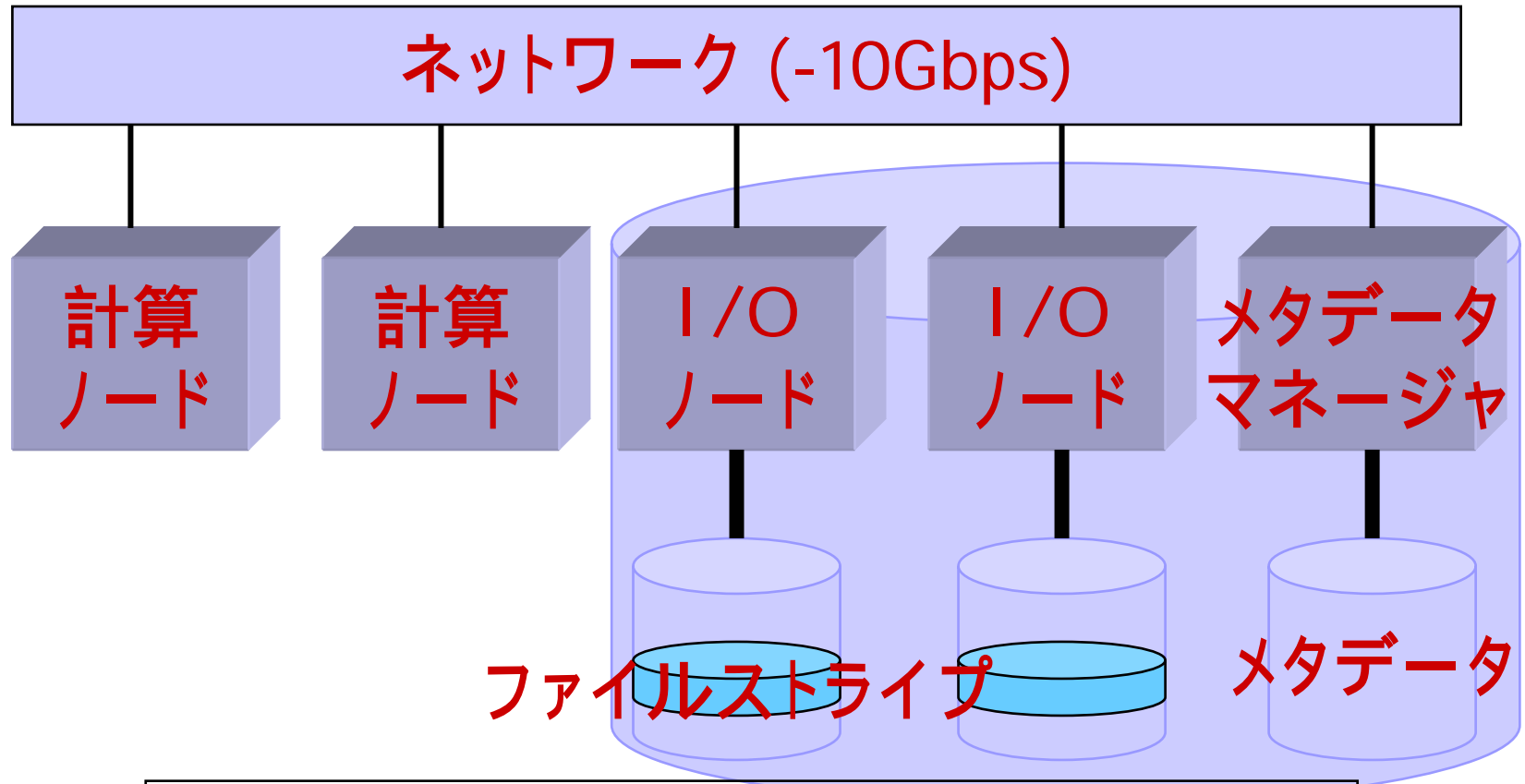
- 装置、計算機、人、可視化装置などが広域に分散するため、高速接続、効率アクセス、安全に共有する技術
 - スケーラブルな並列I/Oバンド幅
 - > 100GB/s, > 1TB/s (システム内, システム間)
 - スケーラブルな計算パワー
 - > 1TFLOPS, > 10TFLOPS
 - 安全な認証、効率的で制御されたデータ / プログラム共有、アクセス制限
 - システムモニタと管理
- 耐故障性 / 動的再配置 / データ復元、再計算

従来手法(1): HPSS/DFS, ...



単一システムイメージ、並列I/O
I/Oバンド幅はネットワークに制限される

従来手法(2): ストライピングクラスタファイルシステム - PVFS, GPFS, ...

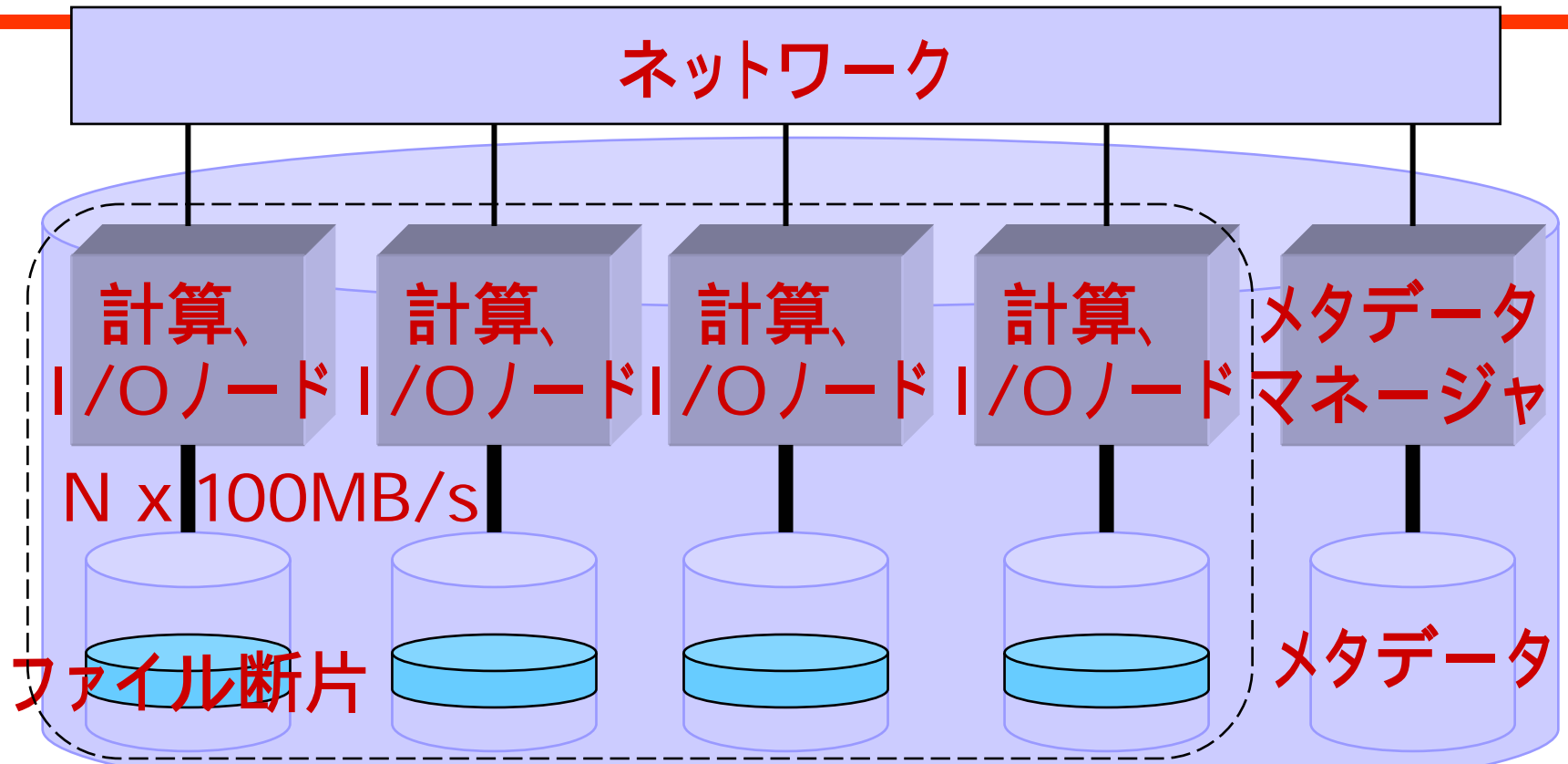


単一システムイメージ、並列I/O
I/Oバンド幅はネットワークに制限される

ペタバイトスケールコンピューティングに向けて

- 広域における効率的な共有
 - 広域高速データ転送
 - 広域データ複製管理
- TB/sを超えるスケーラブルなバンド幅のために
 - I/Oバンド幅はネットワークバンド幅に制限される
 - ローカルI/Oを積極的に利用
 - ネットワークのデータ移動を可能な限り避ける
- 耐故障性
 - 広域ネットワークの一時的不通はおこりがち
 - ノードやディスクの故障もおこりやすい
- 根本的に**新しいパラダイム**が必要

提案手法：広域データ並列ファイルシステム



単一システムイメージ、並列I/O
 ローカルファイルビュー、アフィニティスケジューリング
 主大規模ファイルに対し局所性を利用

提案手法(2): グリッド上の広域データ 並列ファイルシステム

- **グリッド上のクラスタ・オブ・クラスタファイルシステム**
 - 耐故障性と負荷分散のため、クラスタ間にファイル複製
 - クラスタファイルシステムの広域拡張
 - ファイルのブロックサイズはブロックごとに自由 - ファイル断片
 - 計算ノードとI/Oノードを統合
 - **並列I/O、並列ファイル複製**、...
- **ローカルI/Oを利用したスケーラビリティ**
 - ローカルファイルビュー - グリッド並列I/O API
 - データ分散に応じたファイルアフィニティスケジューリング
- **グリッド環境における耐故障性、負荷分散**
 - **ファイル複製**、生成履歴をファイルシステムメタデータで一貫して管理することによりデータ復元 - 複製は負荷分散にも利用

Gfarm cluster-of-cluster filesystem (1)

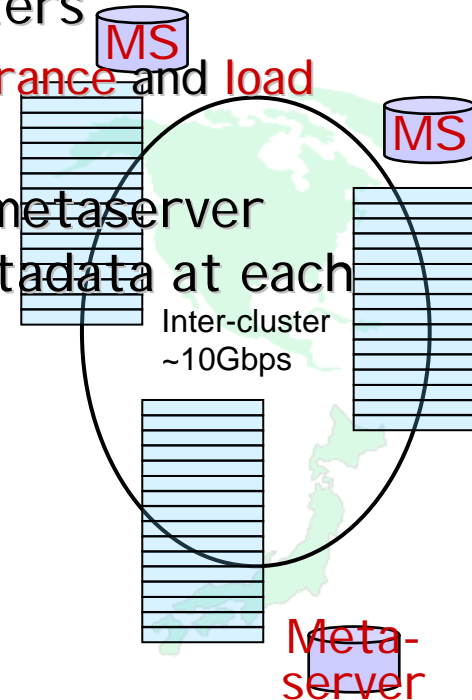
- Extension of cluster filesystem

- File is divided into file fragments
- Arbitrary length for each file fragment
- Arbitrary number of I/O nodes for each file
- Filesystem metadata is managed by metaserver
- Parallel I/O and parallel file transfer



- Cluster-of-cluster filesystem

- File replicas among (or within) clusters
 - fault tolerance and load balancing
- Filesystem metaserver manages metadata at each site



Gfarm cluster-of-cluster filesystem (2)

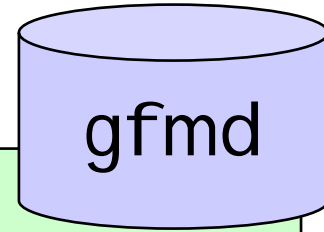
- Gfsd – I/O daemon running on each filesystem node
 - Remote file operations
 - Authentication / access control (via GSI, . . .)
 - Fast executable invocation
 - Heartbeat / load monitor
 - Process / resource monitoring, management
- Gfmd – metaserver and process manager running at each site
 - Filesystem metadata management
 - Metadata consists of
 - **Mapping** from logical filename to physical distributed fragment filenames
 - **Replica catalog**
 - **Command history** for regeneration of lost files
 - Platform information
 - File status information
 - Size, protection, . . .

Extreme I/O bandwidth (1)

- Petascale file tends to be accessed with access locality
 - Local I/O aggressively utilized for scalable I/O throughput
 - Target architecture – cluster of clusters, each node facilitating large-scale fast local disks
- File affinity process scheduling
 - **Almost Disk-owner computation**
- Gfarm parallel I/O extension - **Local file view**
 - MPI-I/O insufficient especially for irregular and dynamically distributed data
 - Each parallel process accesses only its own file fragment
 - Flexible and portable management in single system image
 - Grid-aware parallel I/O library

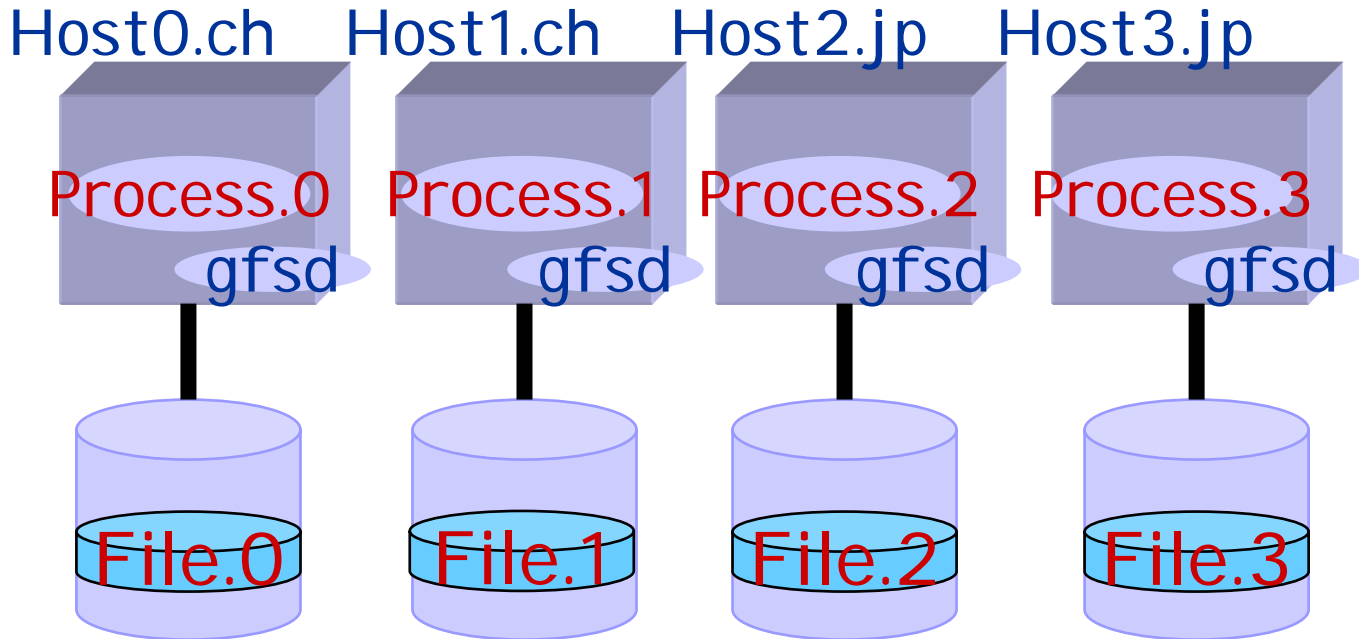
Extreme I/O bandwidth (2)

Process manager - scheduling



gfarm:File
 Host0.ch Host1.ch Host2.jp
 Host3.jp

- File affinity scheduling

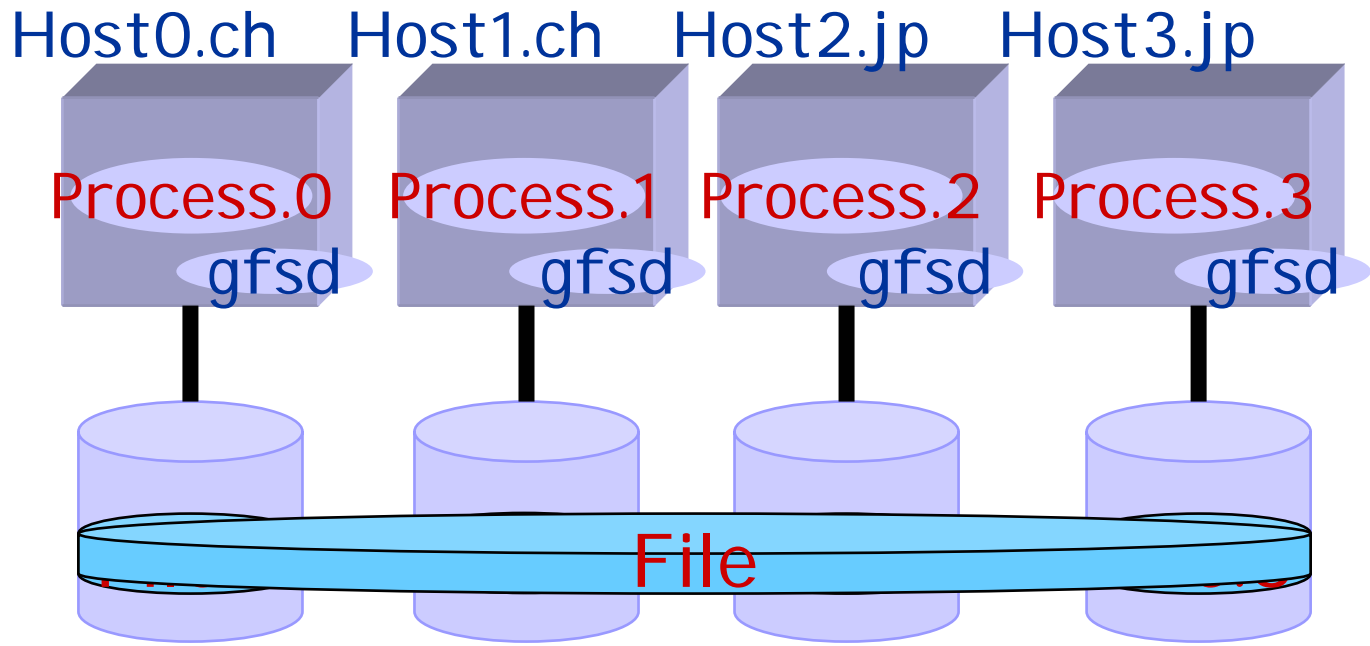
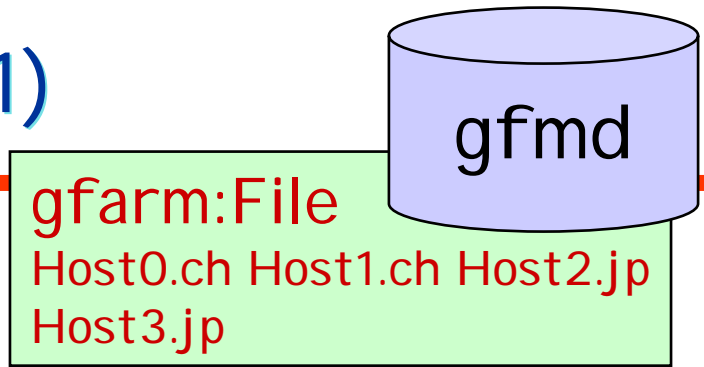


Process scheduling based on file distribution
 Ex. % gfrun -G gfarm:File Process

Extreme I/O bandwidth (3)

Gfarm I/O API - File View (1)

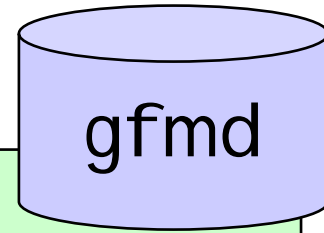
- Global file view



(I/O bandwidth limited by bisection bandwidth, ~GB/s, as an ordinal parallel filesystem)

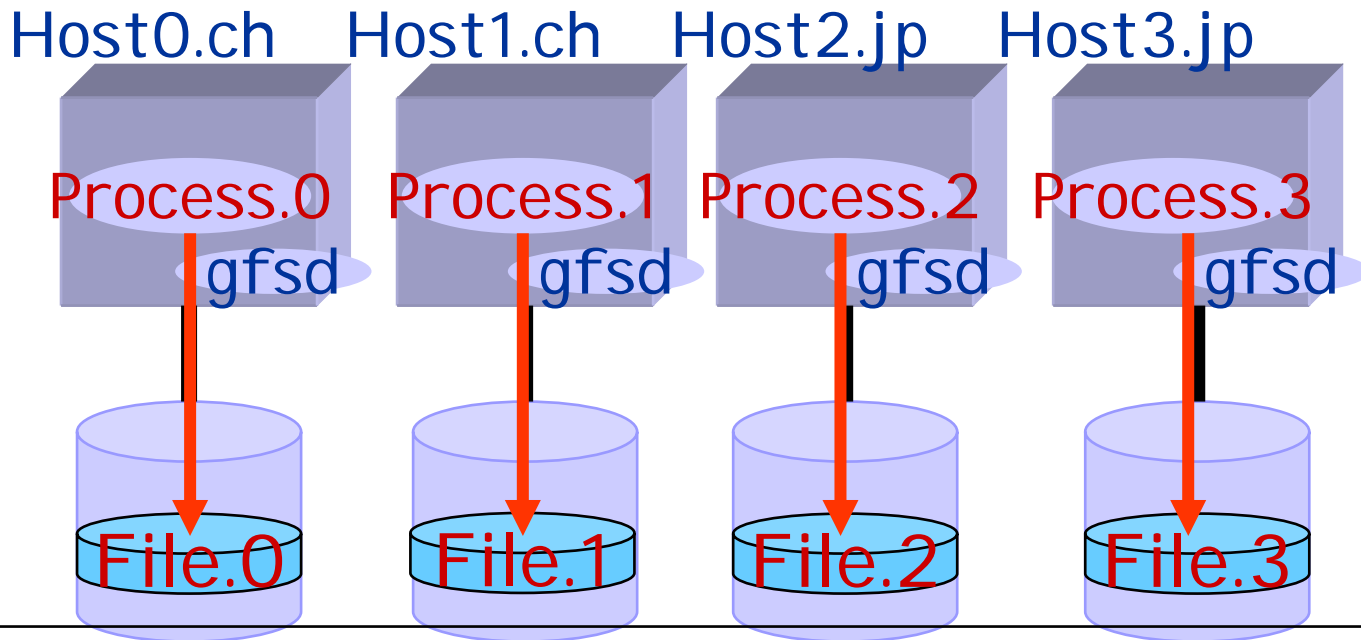
Extreme I/O bandwidth (4)

Gfarm I/O API - File View (2)



gfarm:File
 Host0.ch Host1.ch Host2.jp
 Host3.jp

Local file view



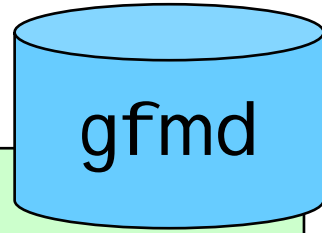
Accessible data set is restricted to a local file fragment
 Scalable disk I/O bandwidth (>TB/s)
 (Local file fragment may be stored in remote node)

Extreme I/O bandwidth support

example: gfgrep - parallel grep

```
% gfrun -G gfarm:input gfgrep
  -o gfarm:output regexp gfarm:input
```

gfarm:input
 Host1.ch Host2.ch Host3.ch
 Host4.jp Host5.jp



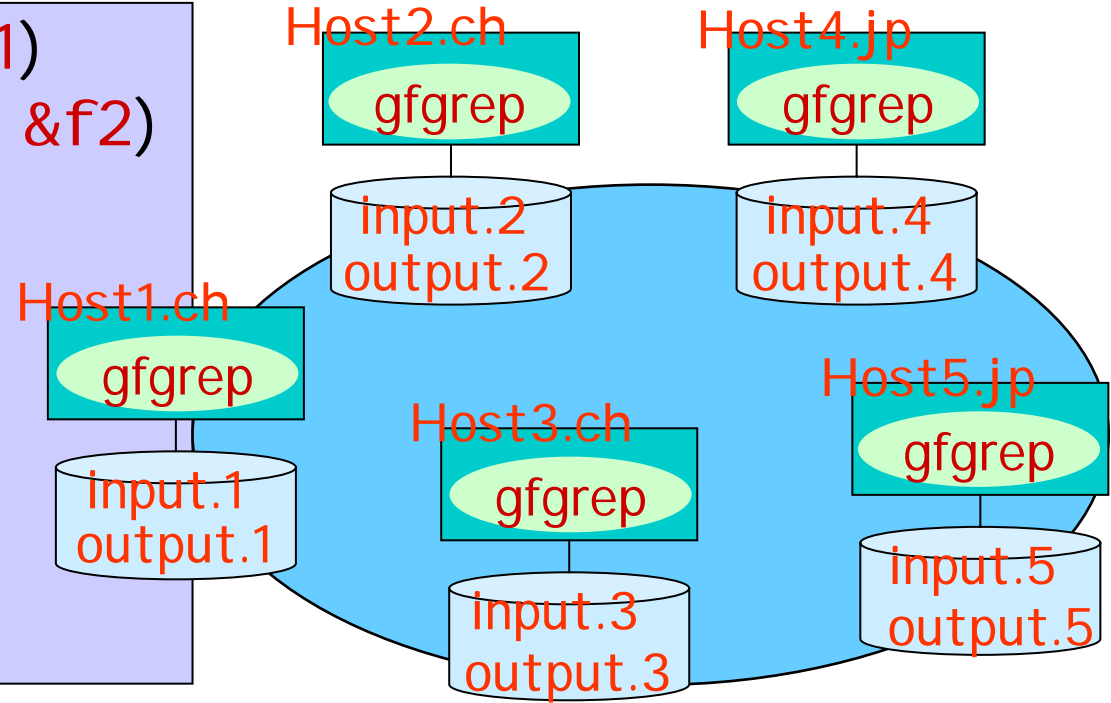
File affinity scheduling

```
open("gfarm:input", &f1)
create("gfarm:output", &f2)
set_view_local(f1)
set_view_local(f2)

↓

grep regexp

close(f1); close(f2)
```



CERN.CH

KEK.JP

耐故障性のサポート

- File replicas on an individual fragment basis
- Re-generation of lost or needed write-once files using a command history
 - Program and input files stored in fault-tolerant Gfarm filesystem
 - Program should be deterministic
 - Re-generation also supports GriPhyN virtual data concept

Gfarm API とGfarmコマンド

<http://datafarm.apgrid.org/>

Gfarm並列I/O APIs

- gfs_pio_open / create / close
- gfs_pio_set_view_local / index / global
- gfs_pio_read / write / seek / flush
- gfs_pio_getc / ungetc / putc
- gfs_mkdir / rmdir / unlink
- gfs_chdir / chown / chgrp / chmod
- gfs_stat
- gfs_opendir / readdir / closedir

主なGfarmコマンド

- gfrep
 - 並列ストリームにより
ファイル複製作成
- gfwhere
 - 複製カタログ表示
- gfls
 - ディレクトリの内容表示
- gfcp
 - 並列ストリームによる
ファイルコピー
- gfrm, gfrmdir
 - ファイル、ディレクトリ削除
- gfmkdir
 - ディレクトリ作成
- gfdf
 - ファイルシステムの空き
ブロック数の表示
- gfscck
 - ファイルシステムの検査
と修復

Porting Legacy or Commercial Applications

- Hook syscalls `open()`, `close()`, `write()`, ... to utilize Gfarm filesystem
 - Intercepted syscalls executed in local file view
 - This allows thousands of files to be **grouped automatically** and processed in parallel.
 - Quick upstart for legacy apps (but some portability problems have to be coped with)
- `gfreg` command
 - After creation of thousands of files, `gfreg` explicitly groups files into a single Gfarm file.

予備評価1 - 評価環境 Presto III Gfarm 開発クラスタ (プロトタイプ)

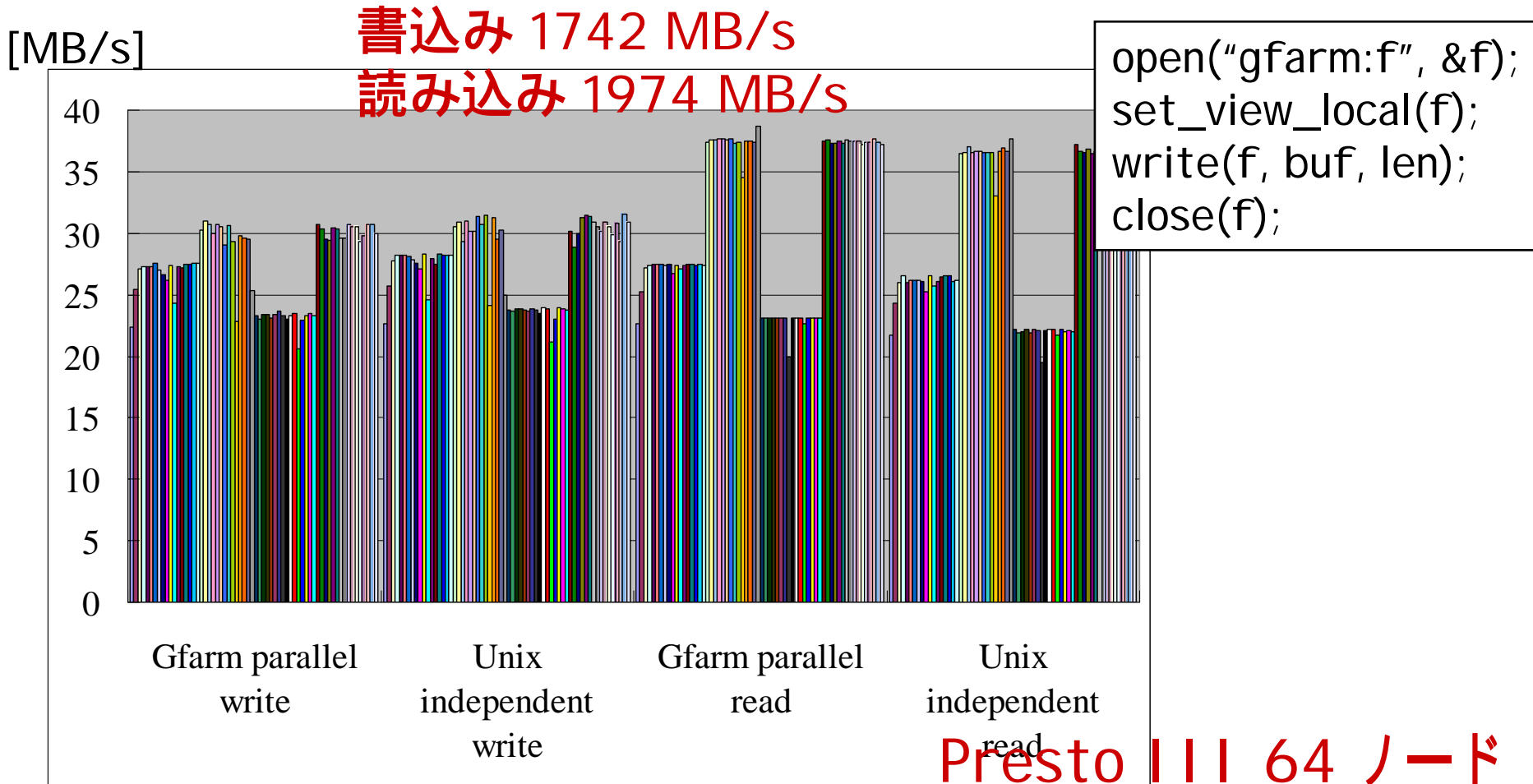
- Dual Athlon MP
1.2GHz 128ノード
- 768MB, 200GB HDD
- 総計98GBメモリ, 25TB
ディスク
- Myrinet 2K, 64bit PCI
- 614 GFLOPS (ピーク)
- 331.7GFLOPS Linpack
for Top500

2001年10月より稼動



初期性能評価(2)

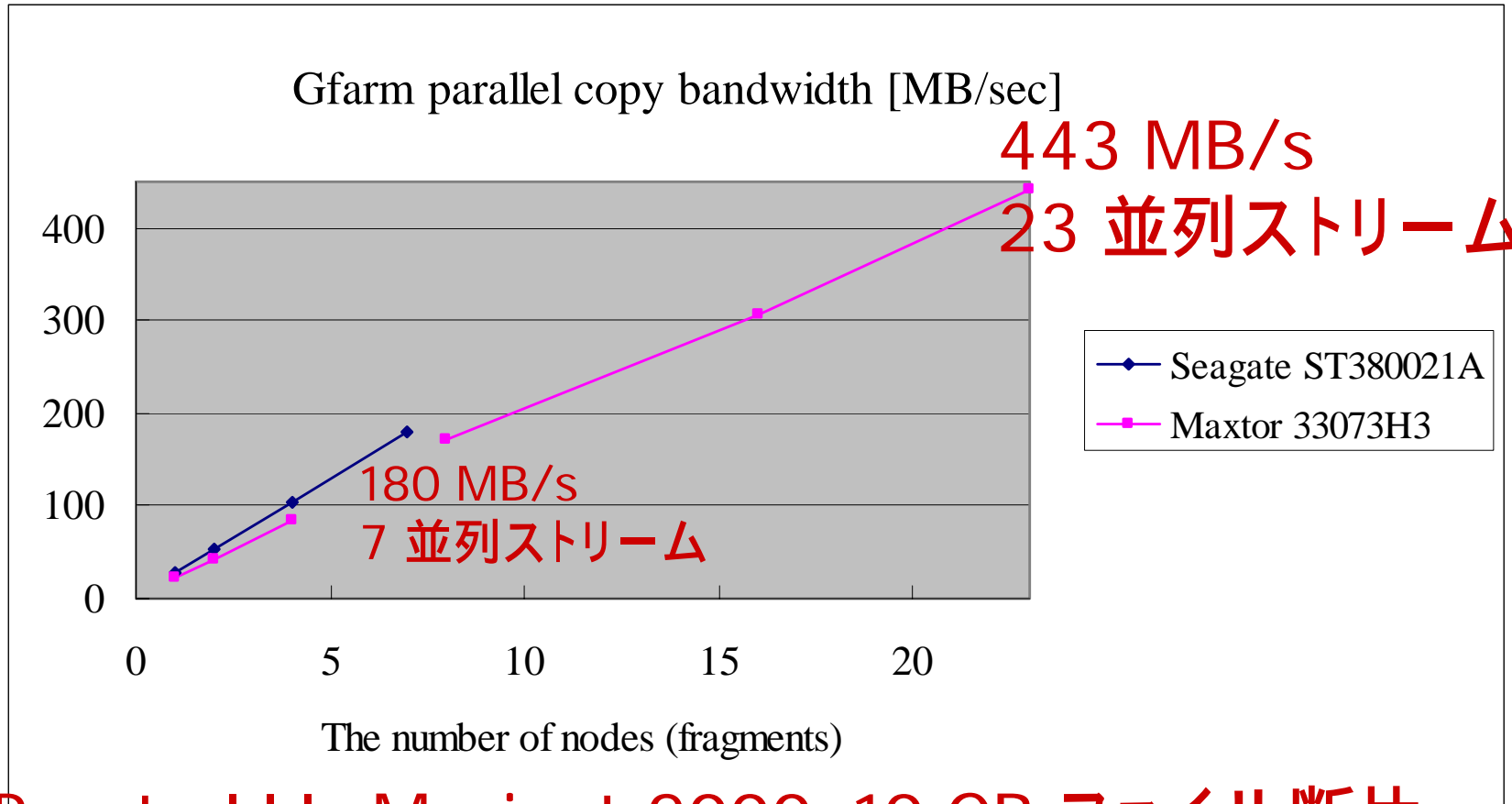
- 並列I/O (ファイルアフィニティスケジューリングと局所ファイルビュー)



Presto III 64 ノード
640 GB データ

初期性能評価 (3)

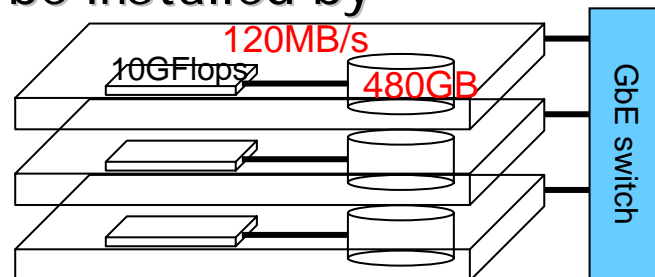
- ファイル複製 (gfrep)



Presto III, Myrinet 2000, 10 GB ファイル断片

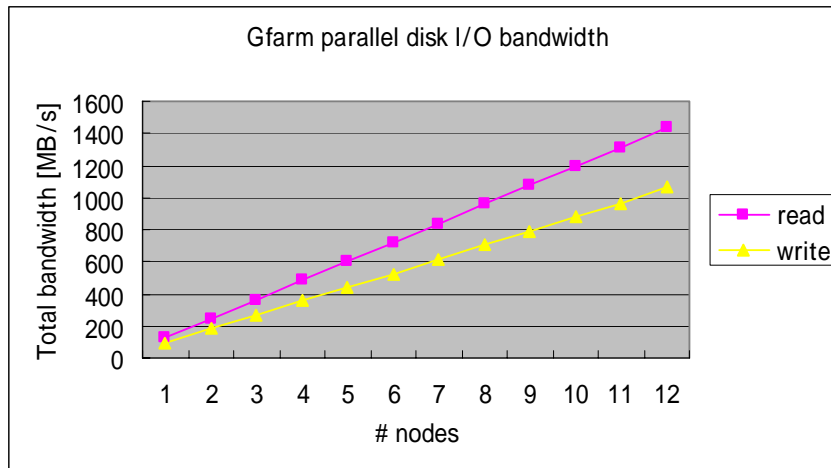
産総研Gfarmクラスタ I の設計

- クラスタノード
 - 1U, Dual 2.4GHz Xeon, GbE
 - 480GB RAID with 4 3.5" 120GB HDDs + RAID card
- 12ノードプロトタイプクラスタ (2002年10月稼動)
 - 12U + ギガビットイーサスイッチ (2U) + KVM スイッチ (2U) + キーボード
 - Totally 6TB RAID with 48 disks
 - 1063 MB/s on writes, 1437 MB/s on reads
 - 410 MB/s for file replication with 6 streams
 - Up to 4 Gbps for external network
 - WAN emulation with NistNET
- 80-node cluster will be installed by Feb 2003



産総研クラスタ初期性能評価

並列ディスクI/O性能

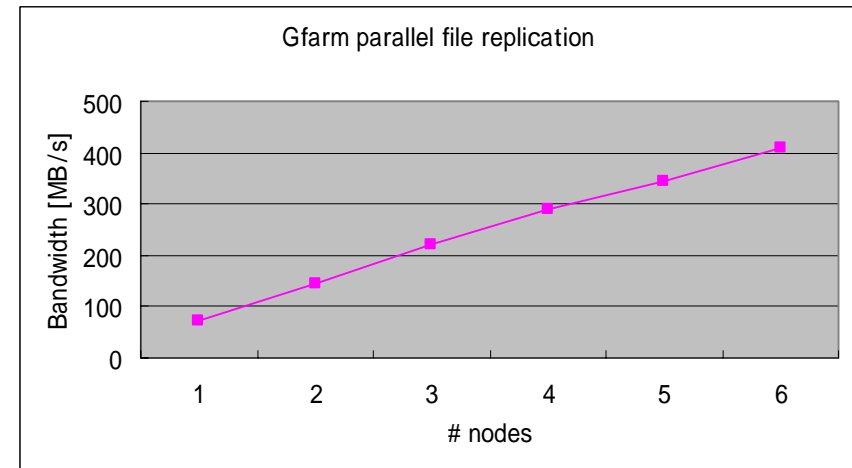


1436 MB/s for reading
1063 MB/s for writing

Per 1 node

120 MB/s for reading
89 MB/s for writing

並列ファイル複製性能

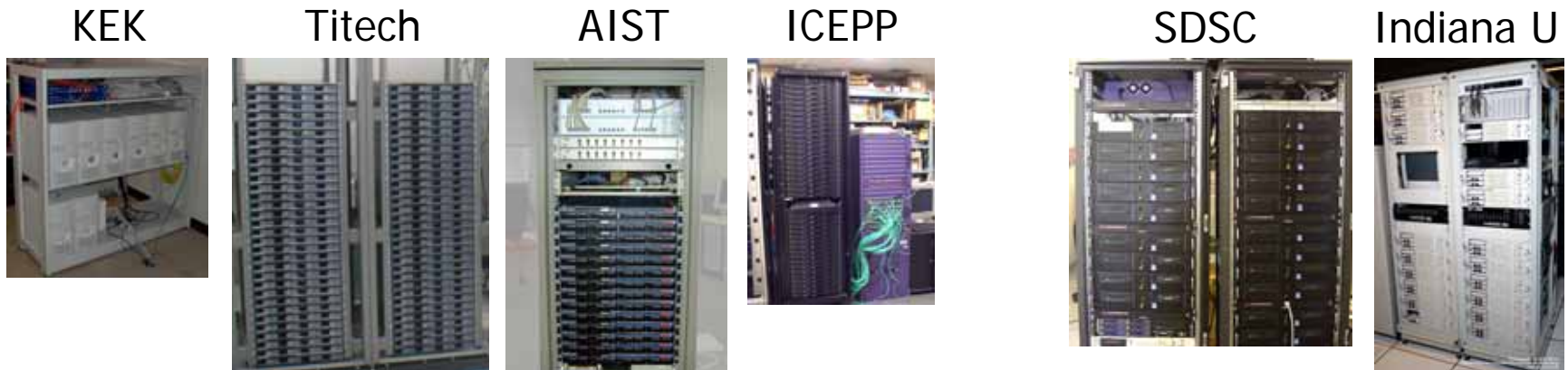
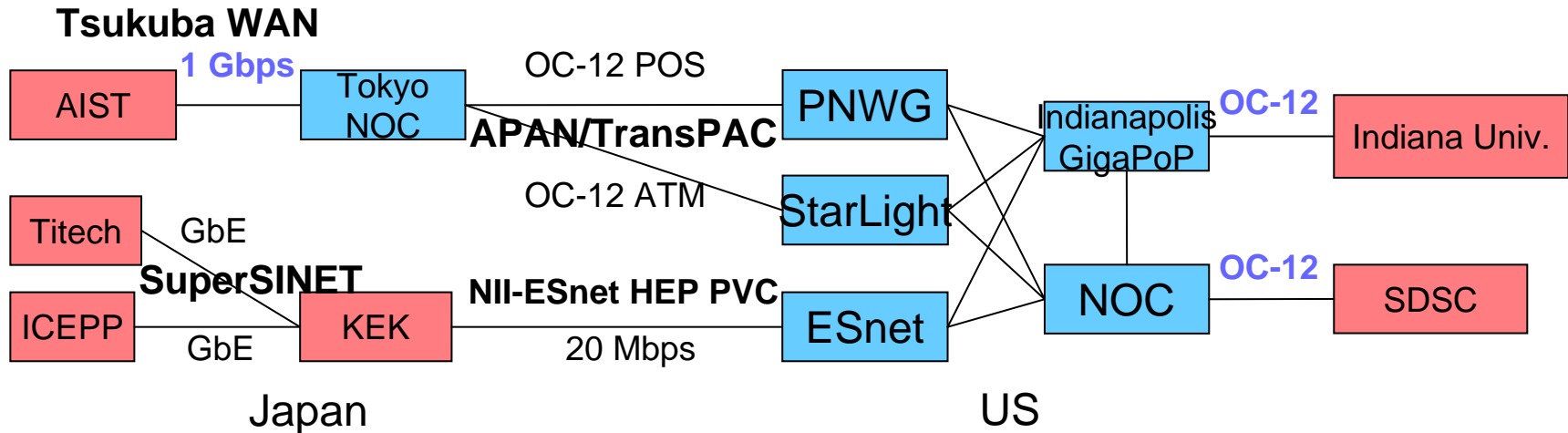


410 MB/s using 6 nodes

Per 1 node

68 MB/s = 547 Mbps

Grid Datafarm US-Japan Testbad



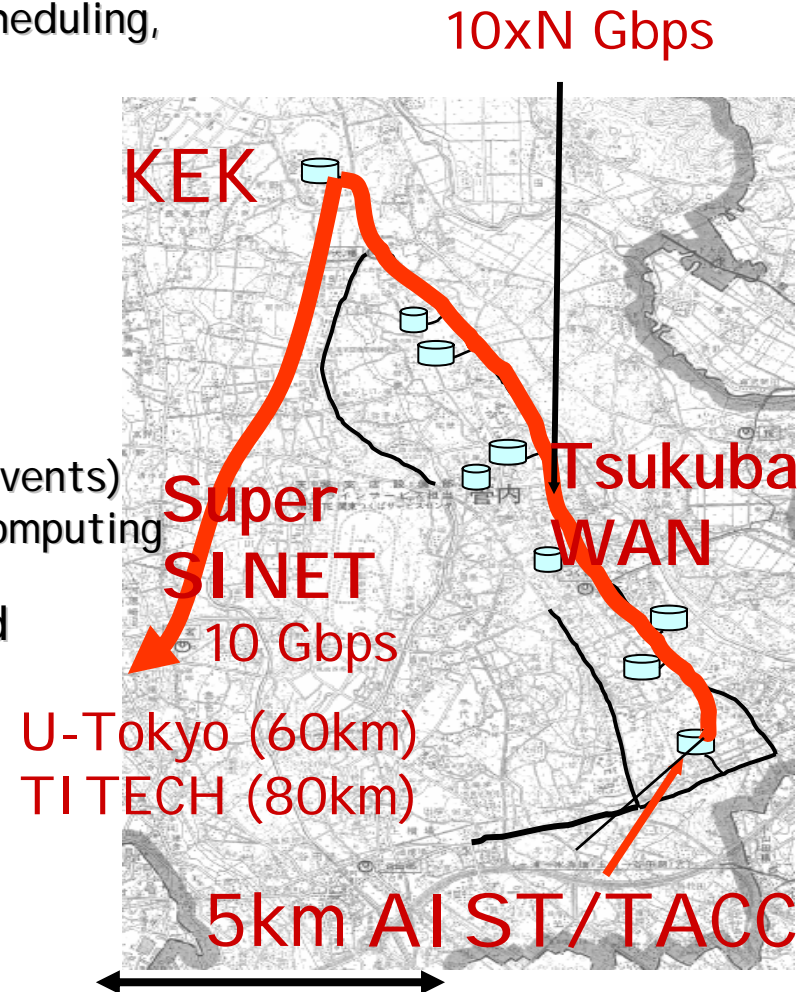
Total disk capacity: 18 TB, disk I/O bandwidth: 6 GB/s

関連研究

- MPI-I/O
 - ローカルI/Oのスケラビリティ活用の鍵となる局所ファイルビューがない
 - PVFS – ストライピングクラスタファイルシステム
 - UNIX I/O API, MPI-I/O
 - 局所性を利用しないため, ネットワークでバンド幅が制限される
 - 耐故障性??? 広域??? 数千大規模??
 - IBM PIOFS, GPFS
 - HPSS – 階層型大容量ストレージシステム
 - ネットワークバンド幅によりI/Oバンド幅が制限される
 - Distributed filesystems
 - NFS, AFS, Coda, xFS, GFS, . . .
 - 複数からの書き込みに対しバンド幅が確保できない
 - Globus – Gridツールキット
 - GridFTP – Gridセキュリティと並列ストリーム
 - 複製管理
 - 複製カタログとGridFTP
 - Kangaroo – Condor approach
 - ローカルディスクをキャッシュとして利用し, 広域における遅延を隠蔽
 - バンド幅は解決されない
- Gfarmはグリッド環境における広域クラスタ・オブ・クラスタファイルシステムの初めての試み
- ファイル複製
 - ファイルアフィニティスケジューリング、...

Grid Datafarm Development Schedule

- Initial Prototype 2000-2001
 - Gfarm filesystem, Gfarm API, file affinity scheduling, and data streaming
 - Deploy on Development Gfarm Cluster
- Second Prototype 2002(-2003)
 - Grid security infrastructure
 - Load balance, Fault Tolerance, Scalability
 - Multiple metaservers with coherent cache
 - Evaluation in cluster-of-cluster environment
 - Study of replication and scheduling policies
 - ATLAS full-geometry Geant4 simulation (1M events)
 - Accelerate by National "Advanced Network Computing initiative" (US\$10M/5y)
- Full Production Development (2004-2005 and beyond)
 - Deploy on Production GFarm cluster
 - Petascale online storage
- Synchronize with ATLAS schedule
 - ATLAS-Japan Tier-1 RC "prime customer"



Summary

<http://datafarm.apgrid.org/>
datafarm@apgrid.org

- Petascale Data Intensive Computing Wave
- Key technology: Grid and cluster
- Grid datafarm is an architecture for
 - Online >10PB storage, >TB/s I/O bandwidth
 - Efficient sharing on the Grid
 - Fault tolerance
- Initial performance evaluation shows scalable performance
 - 1742 MB/s, 1974 MB/s on writes and reads on 64 cluster nodes of Presto III
 - 443 MB/s using 23 parallel streams on Presto III
 - 1063 MB/s, 1436 MB/s on writes and reads on 12 cluster nodes of AI ST Gfarm I
 - 410 MB/s using 6 parallel streams on AI ST Gfarm I
- Metaserver overhead is negligible
- I/O bandwidth limited by not network but disk I/O (good!)

