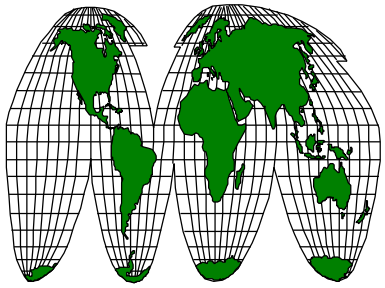
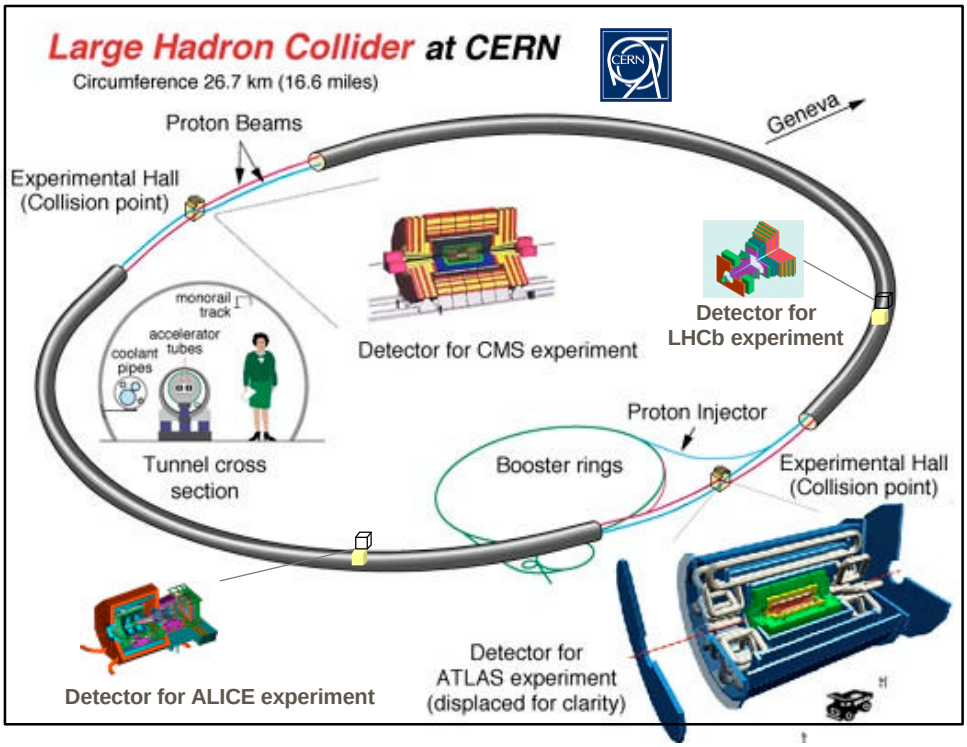
アトラス実験とグリッド・コンピューティング
～ 世界規模のデータ解析環境の構築

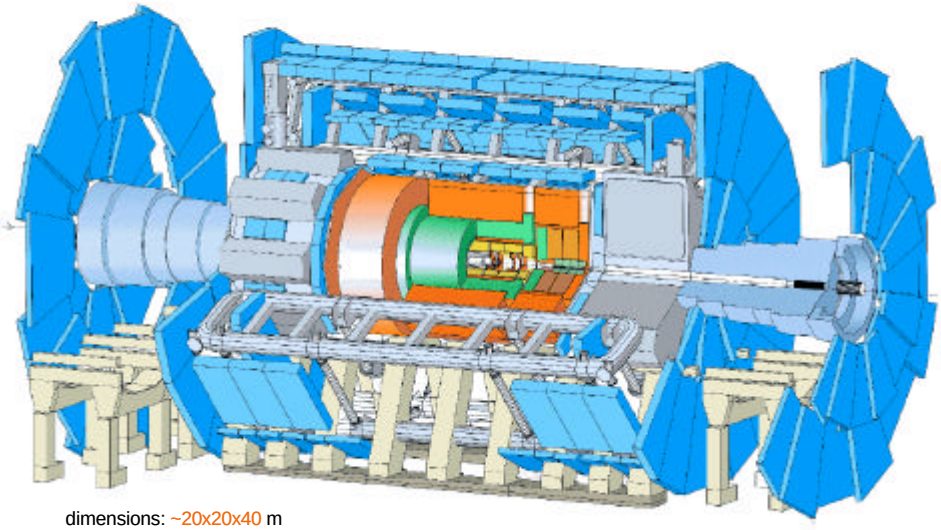


高エネルギー加速器研究機構
計算科学センター
森田洋平



ATLAS Detector

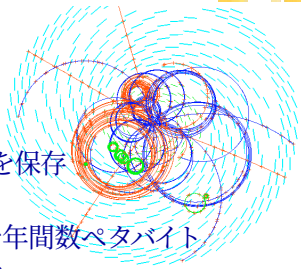
~1850 physicists
from 33 countries



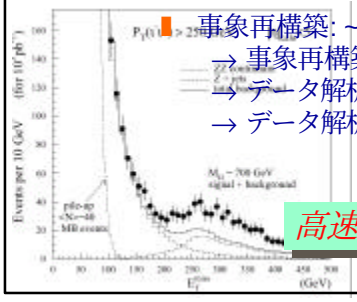
dimensions: ~20x20x40 m
weight : ~7000 ton
readout ch: ~1.5 x 10^8

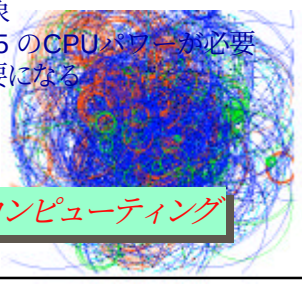
物理データ解析のチャレンジ

"干し草の山の中から針を探しだす"



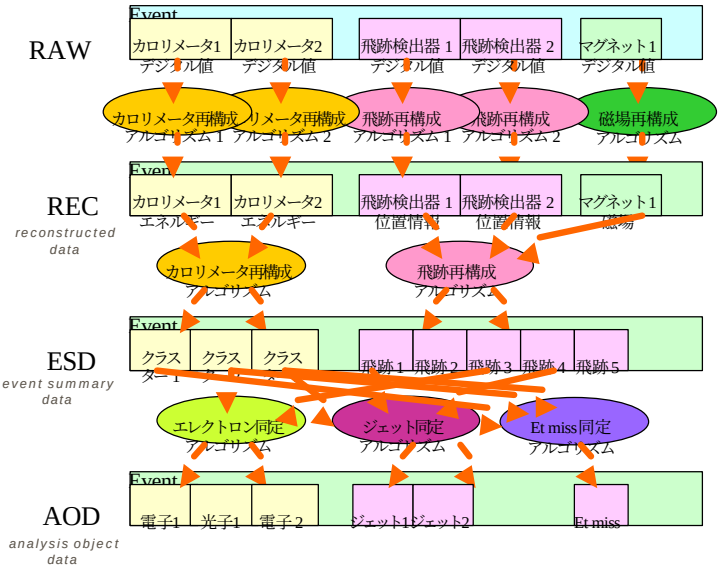
- 毎秒 10億回の衝突事象
  → オンラインで選別→ 毎秒 100 事象を保存
  → 1年あたり10億事象
- データサイズ 1 Mbyte/事象 → 4実験で年間数ペタバイト
- 事象再構築: ~ 300 SPECint95*秒/事象
  → 事象再構築だけで 20万SPECint95 のCPUパワーが必要
  → データ解析にさらにその数倍が必要になる
  → データ解析も国際協力で！





高速I/O，データ主体のコンピューティング

高エネルギー実験のデータ解析モデル



~1PB/年
1MB/事象

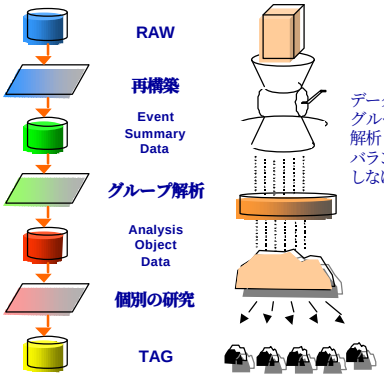
~1PB

~300TB/年
100KB/事象

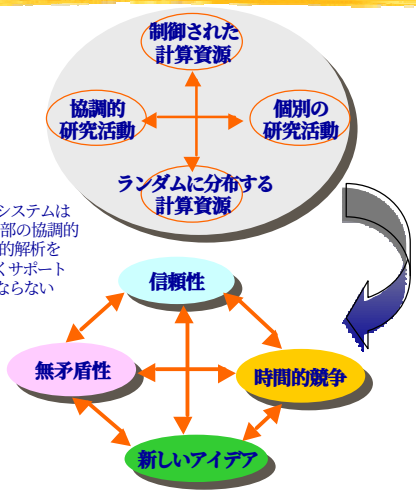
~10TB/年
10KB/事象

データ解析システム構築の考慮点

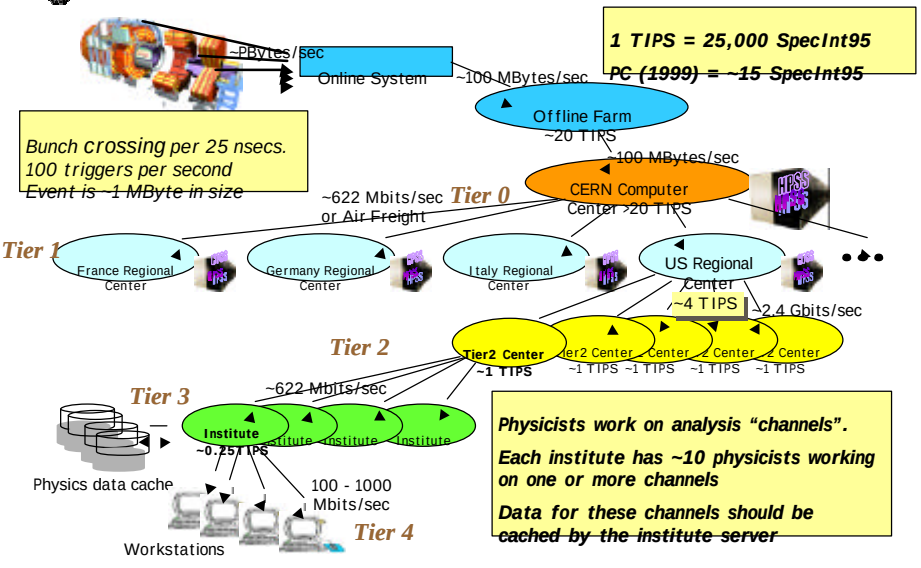
- 高エネルギー実験のデータ解析は世界中に分散した研究者のグループによる協調的かつ競争的研究活動



データ解析システムはグループ内部の協調的解析と競争的解析をバランス良くサポートしなければならない



LHCの多階層型地域解析センターモデル



24 March 2000, WW A/C Panel, P. Capiluppi

高エネルギー実験データ解析の要求事項

- 実験グループ単位の計算資源とアクセス制御
- 世界中に分散した研究者による解析作業のサポート

- 限りある計算機資源、ストレージ資源、ネットワーク資源の管理とスケジューリング
- グループ内部での実験データの共有と効率的なアクセス
- 解析プログラムの共有
- システムの運用管理と稼動状況モニタリング
- システムの可用性（フォルトトレランス、システムの動的再配置）
- その他のグローバルコンピューティング環境
- ビデオ会議システムによる多地点会議

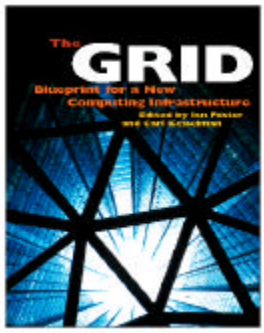
→ グリッドの各種技術が有効に利用できるという期待

アトラス実験の"データ・チャレンジ"計画

- **2001年末** **Data Challenge 0**
  アトラス解析ソフトウェアのフル稼働
- **2002年** **Data Challenge 1** **"~ 0.1％" test**
  地域解析センター試験
- **2002年末** **計算機技術デザインのまとめ**
  (Technical Design Report)
- **2003年** **計算機・ソフトウェア各国分担の覚書**
- **2003年** **Data Challenge 2** **"~10％" test**
  計算機・ソフトウェアモデルの実証的検証
- **解析ソフトウェアと解析システムを段階的に実証する**

グリッドとは

- インターネットの普及とブロードバンド化に伴う広域分散型超並列計算
- PCの低価格化、高性能化、Linux普及に伴う超並列クラスターの実現
  → 広域ユーザー認証、並列計算、データアクセス、ユーザーインターフェースなどの標準化が必要になってくる
- 仮想的な大規模並列計算機
  - Metacomputing [Smarr87]
  - “The GRID” [Foster ら98]
- 次世代のインターネットのソフトウェア基盤
  - 既存のソフトウェア基盤の上位レイヤとして
  - サービスとプロトコルの研究・提供・標準化
  - Grid Forum として活動を開始
- 電力線 "Power Grid" の計算機・ネットワーク版



www.gridforum.org

Gridサービスのアーキテクチャ

アプリケーション層	... さまざまなGridアプリケーション ...				
アプリケーション ツールキット層	リモート データ ツールキット	リモート 計算 ツールキット	**リモート 可視化 ツールキット**	非同期 コラボレーション ツールキット	.. リモート センシング ツールキット
Grid サービス層	プロトコル, 認証, 利用許可, 資源管理, 機器対応, 資源探索, その他				
Grid ファブリック層	Grid対応のアーカイブ, ネットワーク, 計算機, ディスプレイ装置, その他 および 関連するローカルなサービスなど				

世界の高エネルギー実験のGridプロジェクト

- PPDG - 米 DoE
  - 超高速ネットワーク、大規模DB実証などのR&D
- GriPhyN - 米 NSF
  - 米Atlas, 米CMS, LIGO, SDSS
    - Tier2 センター設立の為のR&D
    - 米Atlas - インディアナ大などに専属研究者
- DataGrid - 欧 IST
  - 2001年から3年計画でLHC4実験ためのミドルウェアを開発
- ApGrid-HEP - 日本
  - 高エネルギー実験データ解析システムの要求要件から出発した高エネ研、産総研、東工大、東大の共同プロジェクト
    → "Grid Data Farm" (Gfarm)

アトラス日本グループの地域解析センター

- KEKと東大 素粒子国際研究センター(ICEPP)の共同で技術開発を推進
- 2006年までに 約6万SPECint95の計算機からなるTier-1データ解析システムを国内に構築、ストレージを約1ペタバイトまで段階的に増強
- 補完的役割を担うCERN分室を設立
- 2001年末から始まるアトラスのデータ・チャレンジに参加
- NIIのSuperSINET計画にGrid/アトラスの専用回線
  - 2001年度末に KEK-ICEPP間に 1 ~ 10 Gbps

地域解析センター実現のための技術課題

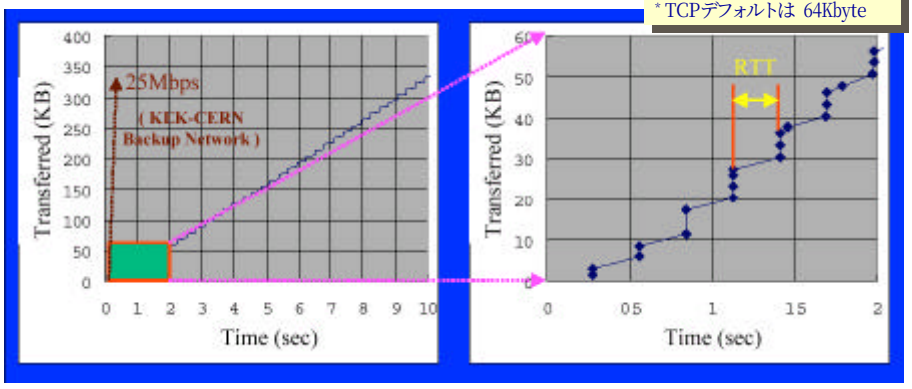
- 広域広帯域ネットワークの利用
  - TCP/IPの技術的制約と効率的ファイル転送技術の必要性
  - サイト間にまたがる研究者の認証とセキュリティの確保
  - 実験データの分配・複製機構
  - 計算資源の効率的管理
- 大規模データストレージと大規模CPUクラスター
  - スケーラブルでフォルトトレラントな大規模システム
  - 共同研究者間で透過的に利用できる広域データ共有システム

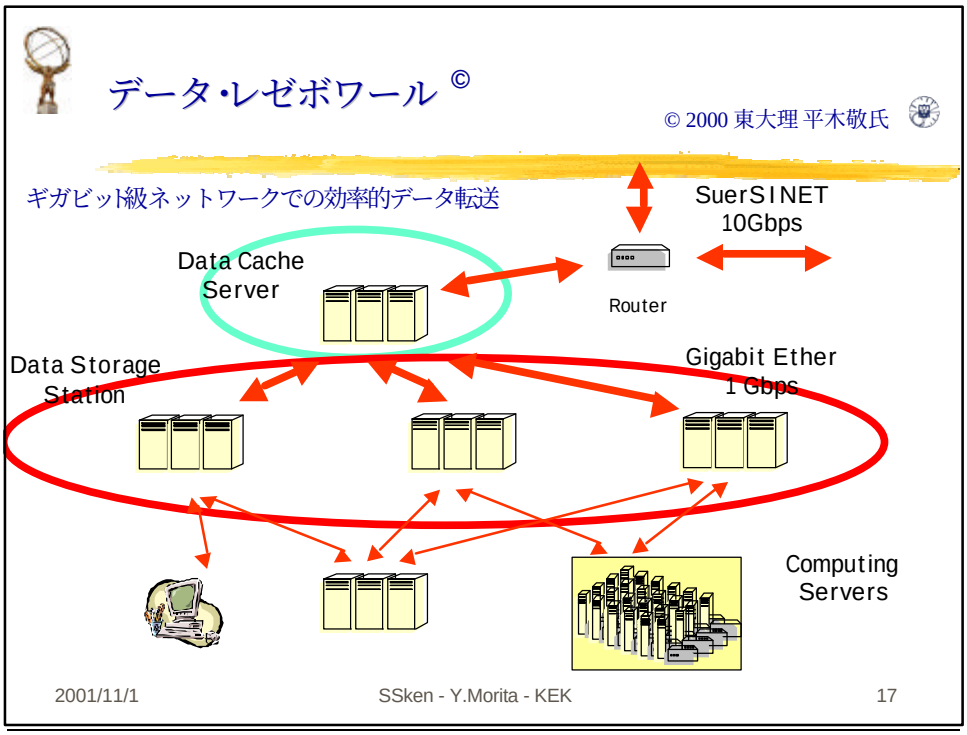
広域高速ネットワークの効率的利用

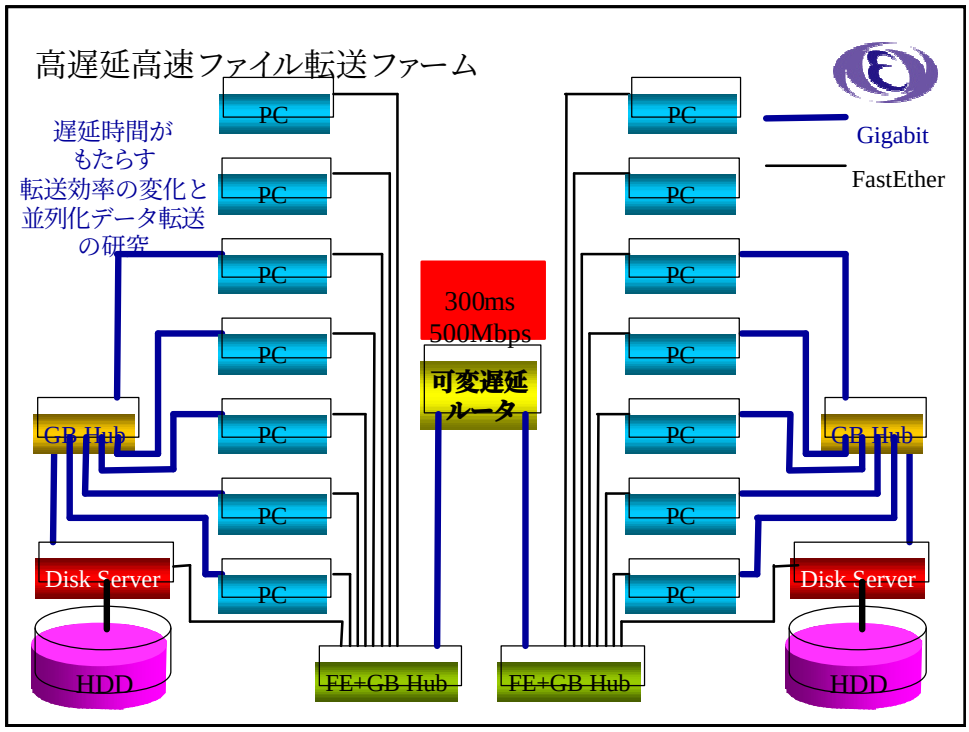
- 高速・高遅延ネットワークで転送効率を保つためには

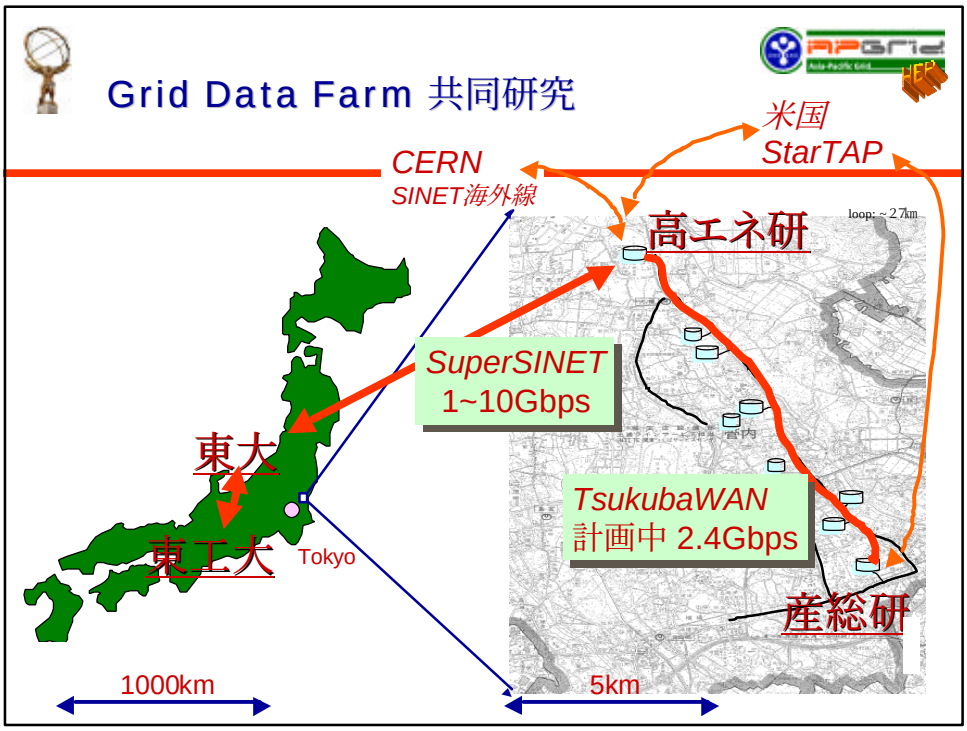
Window size ∝ Latency * Bandwidth

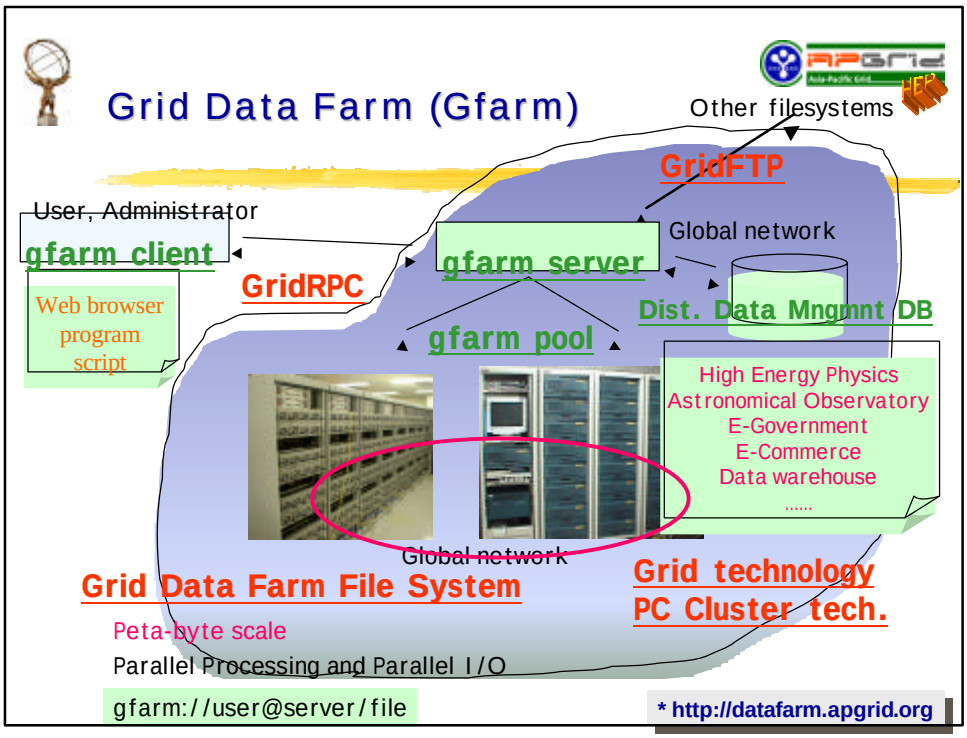
日欧回線 RTT ~ 300msec
300 msec * 1 Gbps ~ 300Mbit
→ 38 Mbyte*
* TCPデフォルトは 64Kbyte




データ・レゼボワール ©
© 2000 東大理 平木敬氏
ギガビット級ネットワークでの効率的データ転送
SuerSINET
10Gbps
Data Cache
Server
Router
Data Storage
Station
Gigabit Ether
1 Gbps
Computing
Servers
2001/11/1
SSken - Y.Morita - KEK
17

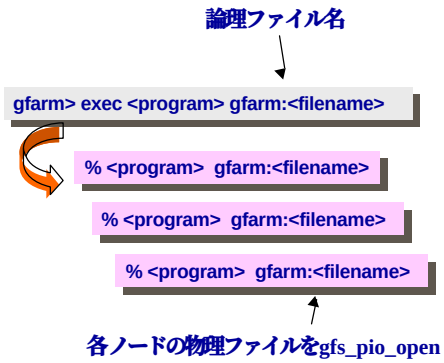

高遅延高速ファイル転送ファーム
遅延時間が
もたらす
転送効率の変化と
並列化データ転送
の研究
Gigabit
FastEther
PC
300ms
500Mbps
可変遅延
ルータ
GB Hub
Disk Server
HDD
FE+GB Hub


Grid Data Farm 共同研究
CERN
SINET海外線
米国
StarTAP
高エネ研
SuperSINET
1~10Gbps
東大
東工大
Tokyo
TsukubaWAN
計画中 2.4Gbps
産総研
1000km
5km


Grid Data Farm (Gfarm)
Other filesystems
GridFTP
User, Administrator
gfarm client
Global network
gfarm server
GridRPC
Web browser
program
script
Dist. Data Mngmnt DB
gfarm pool
High Energy Physics
Astronomical Observatory
E-Government
E-Commerce
Data warehouse
.....
Global network
Grid Data Farm File System
Peta-byte scale
Parallel Processing and Parallel I/O
Grid technology
PC Cluster tech.
gfarm://user@server/file
* http://datafarm.apgrid.org

Grid Data Farm (Gfarm) の発想

- 安価なPCを数千台規模で並べるのであれば、そのローカルディスクをシステム全体のストレージとして活用できないか？
- 解析は事象ごとに独立なのだから、データをプログラムが実行されるCPUまで送るのではなく、プログラムをデータが存在するCPUまで送ればよい
- データを細かく分けて各ノードに分散し、データのセットに論理名をつけて管理すればCPUとI/Oの双方の負荷分散になる
- データの履歴管理機能を持てばデータの動的再生成が可能になる
- ネットワーク上のファイル転送も並列に行うことができる



"もともと並列な処理なら、すべてを並列なままで扱おう"

今後の予定

- HPSSなどの大容量ストレージとの効率的かつスケーラブルな接続
- 数百~1000台規模のPCクラスターによる性能実証
- 1 ~ 10 Gbpsの高速広域ネットワークによる実証試験
- Firewall、サイト間にまたがるユーザー認証インフラの構築
- 世界規模のテストベッド構築と分散データ解析ソフトウェアの実証

まとめ

- ギガビット級の国際ネットワークで世界各地の研究所が相互接続される時代がやってきた → LANとWANの帯域幅の格差の減少
- グリッド技術は実験データの格納場所やCPUの場所を直接意識しなくすむ仮想的なデータ解析環境を提供する
- 高エネルギー実験に参加する各国が計算資源をネットワーク上に提供する、世界的な多階層型データ解析環境の構築が進みつつある
- KEKと東大素粒子国際研究センターでは2006年から始まるLHC/アトラス実験のためにTier1地域解析センター網を構築する
- 高エネルギー実験分野と計算科学分野の研究者の共同研究が世界各地で進んでいる
- ペタバイト級のストレージと数千台規模の並列処理CPU、高速・高遅延ネットワークを有効に結び付けるシステムモデルの構築と検証が急ピッチで進みつつある