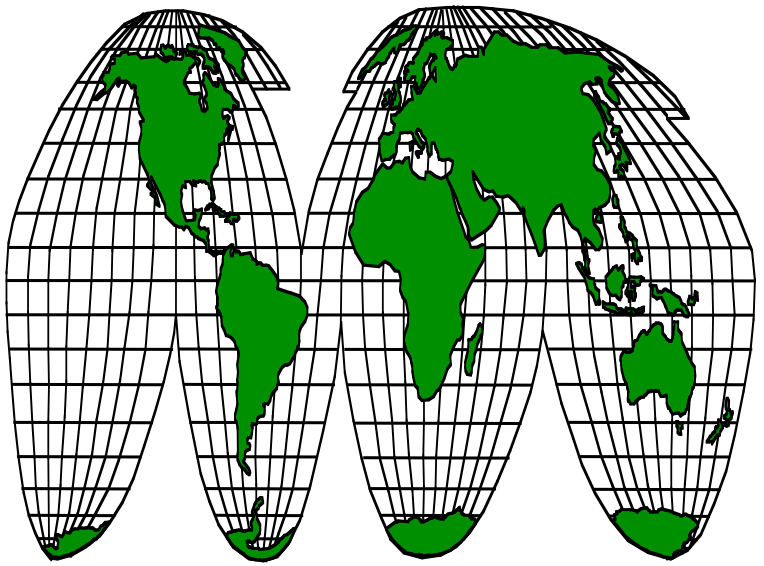
ATLAS実験とデータグリッド
～ 世界規模のデータ解析環境の構築



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

お知らせ

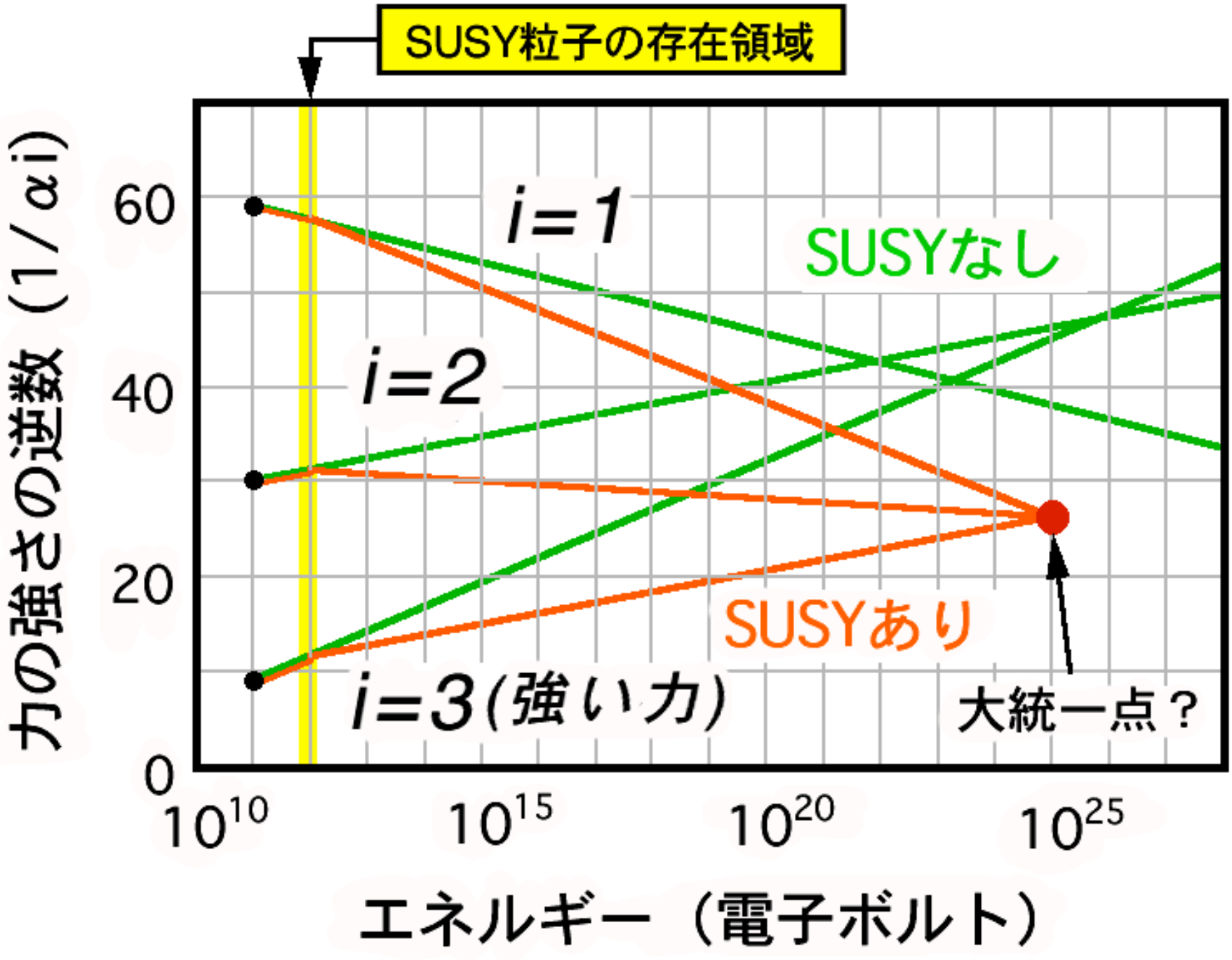
高エネルギー物理データグリッド研究会
平成15年3月10～11日

Symposium on HEP Data Grid
March 10-11, 2003

高エネルギー加速器研究機構
4号館セミナーホール

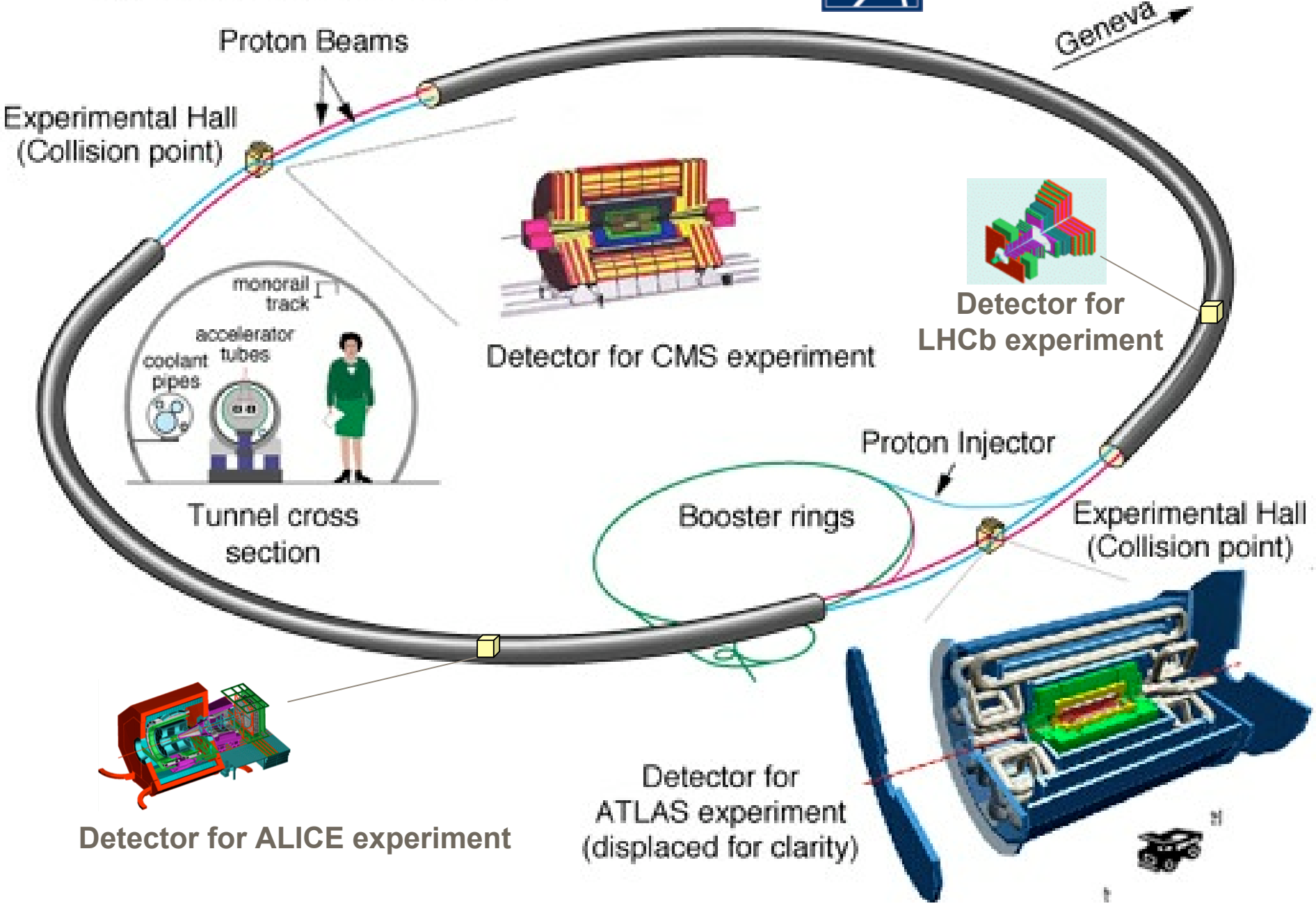
問い合わせ先：計算科学センター 渡瀬芳行
yoshiyuki.watase@kek.jp


物　質　粒　子
ゲージ粒子
第１世代
第２世代
第３世代
クォーク
u
アップ
c
チャーム
t
トップ
d
ダウン
s
ストレンジ
b
ボトム
レプトン
νe
eニュートリノ
νμ
μニュートリノ
ντ
τニュートリノ
e
電子
μ
ミューオン
τ
タウ
強い力
g
グルーオン
電磁力
γ
光子
弱い力
W+
W−
Z
W ボゾン
Z ボゾン
ヒッグス場に伴う粒子
（未発見）
H
ヒッグス粒子


SUSY粒子の存在領域
力の強さの逆数（1/αi）
60
40
20
0
i=1
i=2
i=3(強い力)
SUSYなし
SUSYあり
大統一点？
10^10
10^15
10^20
10^25
エネルギー（電子ボルト）

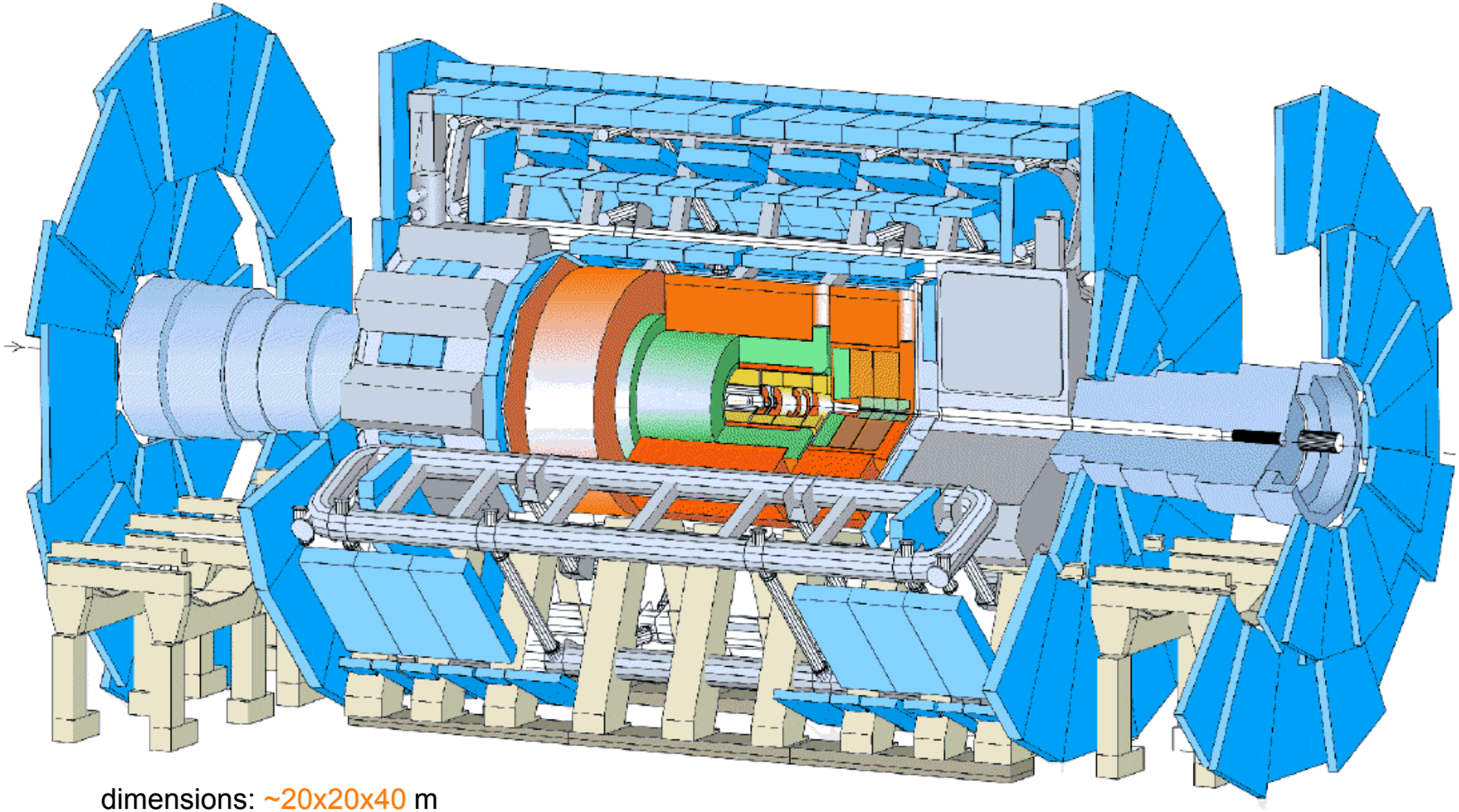
Large Hadron Collider at CERN

Circumference 26.7 km (16.6 miles)



ATLAS Detector

~1850 physicists
from 33 countries



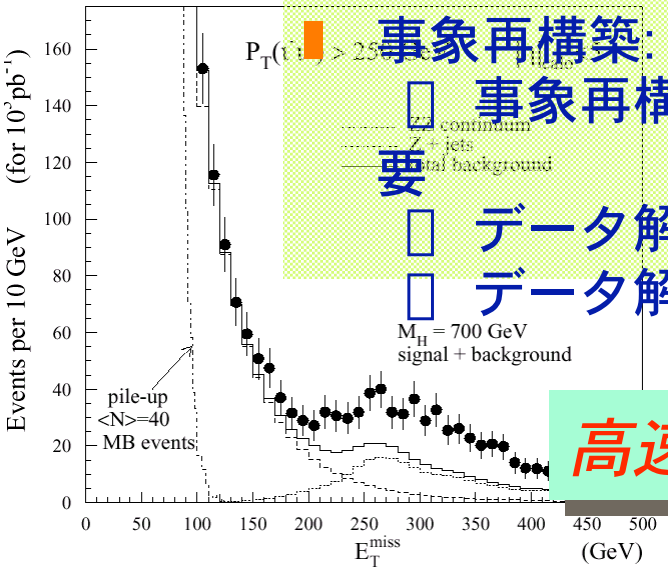
dimensions: ~20x20x40 m
weight : ~7000 ton
readout ch: ~1.5 x 10^8

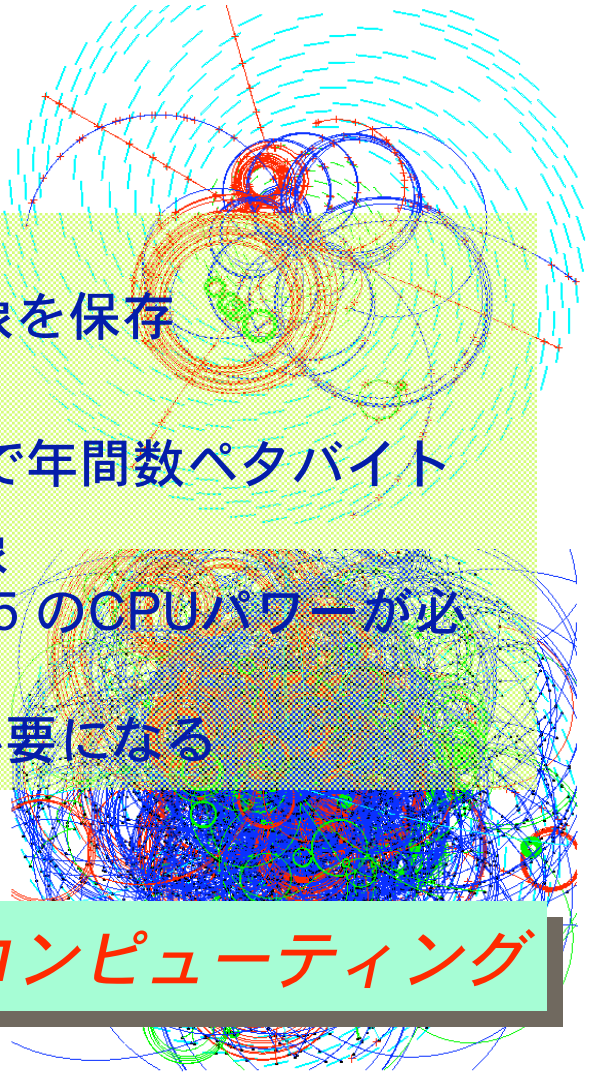

Armenia
Australia
Austria
Azerbaijan
Belarus
Brazil
Canada
China
Germany
Danmark
Spain
Finland
France
Georgia
Greece
Italy
Israel
Japan
Kazakhstan
Morocco.
Norway
Netherlands
Portugal
Poland
Romania
Russia
Sweden
Switzerland
Slovenia
Slovakia
Turkey
UK
USA
ATLAS
Collaboration

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"

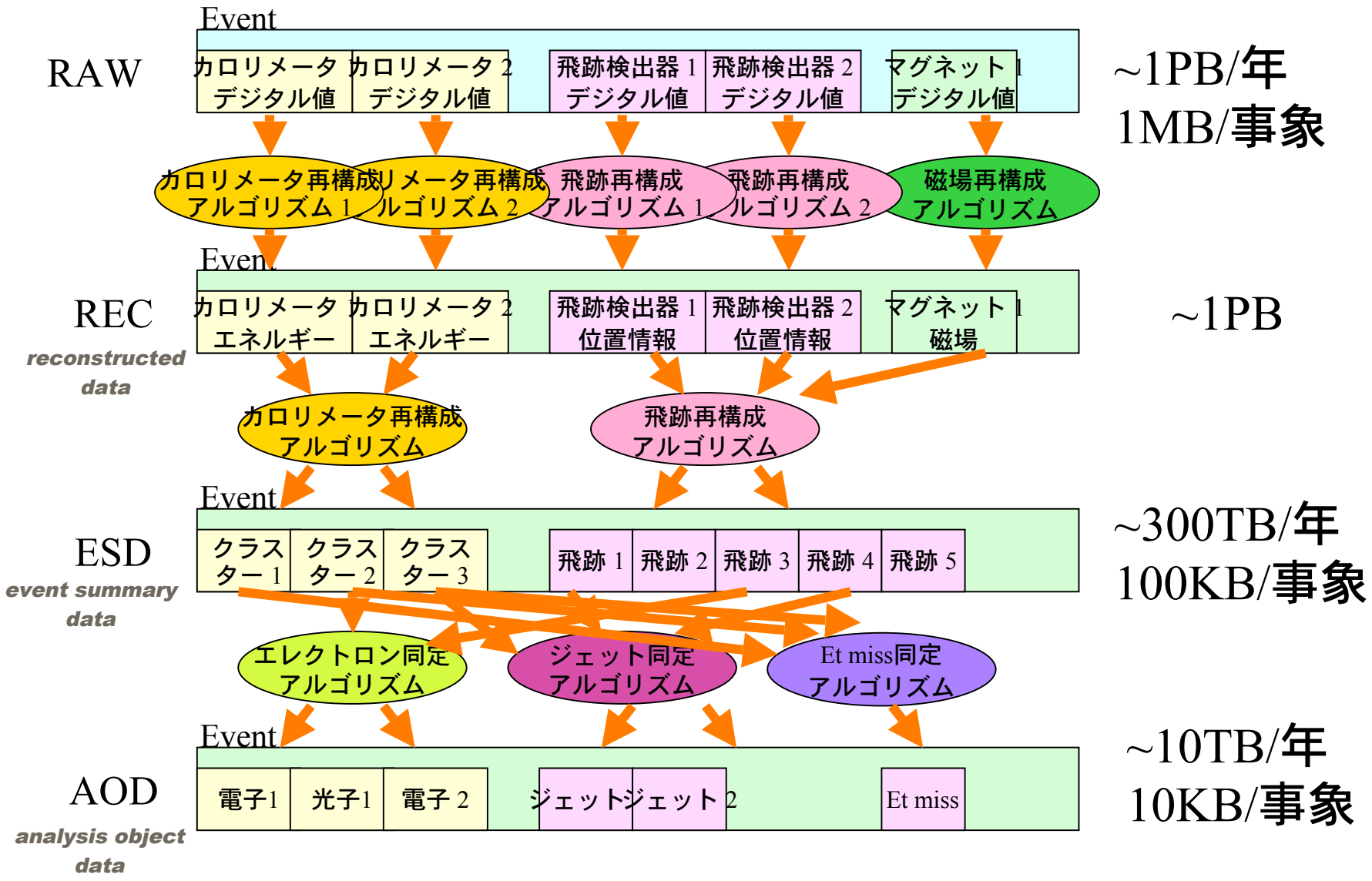
- 毎秒 10億回の衝突事象
  - オンラインで選別 毎秒 100 事象を保存
  - 1年あたり 10億事象
- データサイズ 1 Mbyte/事象 4実験で年間数ペタバイト
- 事象再構築: ~ 300 SPECint95*秒/事象
  - 事象再構築だけで 20万SPECint95 のCPUパワーが必要
  - データ解析にさらにその数倍が必要になる
  - データ解析も国際協力で！





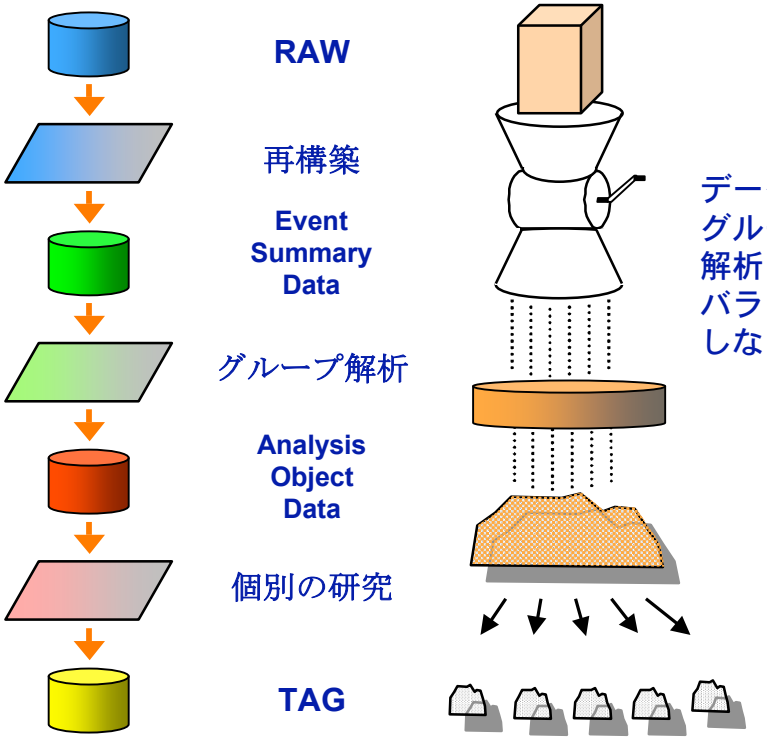
高速I/O, データ主体のコンピューティング

高エネルギー実験のデータ解析モデル

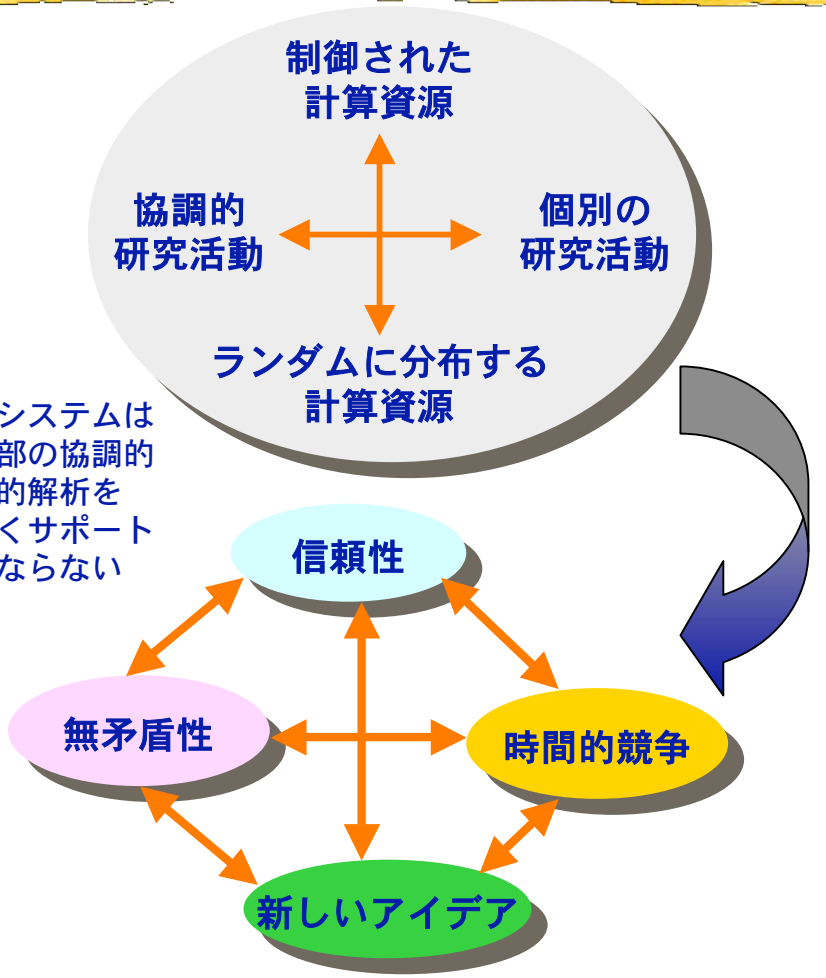


データ解析システム構築の考慮点

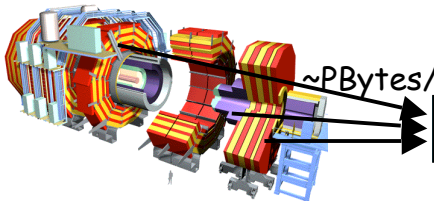
- 高エネルギー実験のデータ解析は世界中に分散した研究者のグループによる協調的かつ競争的研究活動



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



~PBytes/sec

Online System

~100 MBytes/sec

1 TIPS = 25,000 SpecInt95

PC (1999) = ~15 SpecInt95

Offline Farm
~20 TIPS

~100 MBytes/sec

Bunch crossing per 25 nsecs.
100 triggers per second
Event is ~1 MByte in size

~622 Mbits/sec
or Air Freight

Tier 0

CERN Computer
Center >20 TIPS

HPSS

Tier 1

France Regional Center

Germany Regional Center

Italy Regional Center

US Regional Center

~4 TIPS

~2.4 Gbits/sec

Tier 2

Tier2 Center
~1 TIPS

~622 Mbits/sec

Tier 3

Institute
~0.25TIPS

Physics data cache

100 - 1000
Mbits/sec

Tier 4

Workstations

Physicists work on analysis "channels".

Each institute has ~10 physicists working on one or more channels

Data for these channels should be cached by the institute server

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

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
- 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
  - LCG: LHC Computing Grid
    2002年から3年計画で LHCの計算機・ソフトウェアを実装
- Gfarm - 日本
  - 高エネルギー実験データ解析システムの要求要件をベースにした、高エネ研、産総研、東工大の共同プロジェクト
    → "Grid Data Farm" (Gfarm)

LCG Grid Deployment Board

- WG1: Choice of Security Middleware and Tools
- WG2: VO management and resources
- WG3: Registration, Authentication, Authorization and Security
- WG4: Security Operational Procedures

- LCG-1 estimates:
  Users ~ 1000
  User Registration: Peak rate ~ 25 users/day in 2003 2Q

ATLAS実験の"データ・チャレンジ"計画

- **2002年　Data Challenge 1　"~ 0.1%" test**

  地域解析センターのテスト + "HLT studies"
  4~8月　Phase1: Event Full Simulation (Fortran)
  10~1月　Phase2: Event PileUp (Fortran)
  3 x 10^6 events, ~ 25TB

- **2003~4年 Data Challenge 2　"~10%" test**

  計算機・ソフトウェアモデルの実証的検証
  (C++, Grid, ...)

- **解析ソフトウェアと解析システムを段階的に実証**

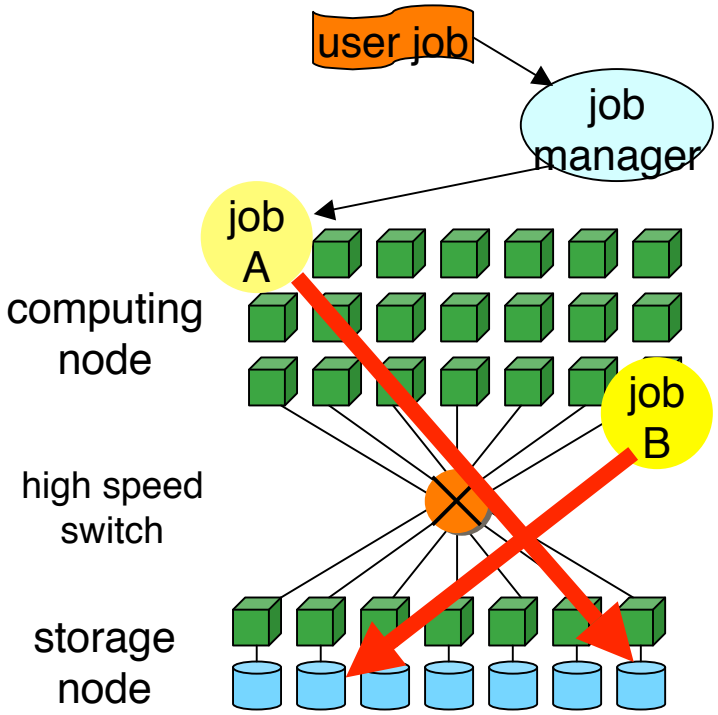
アトラス日本グループの地域解析センター

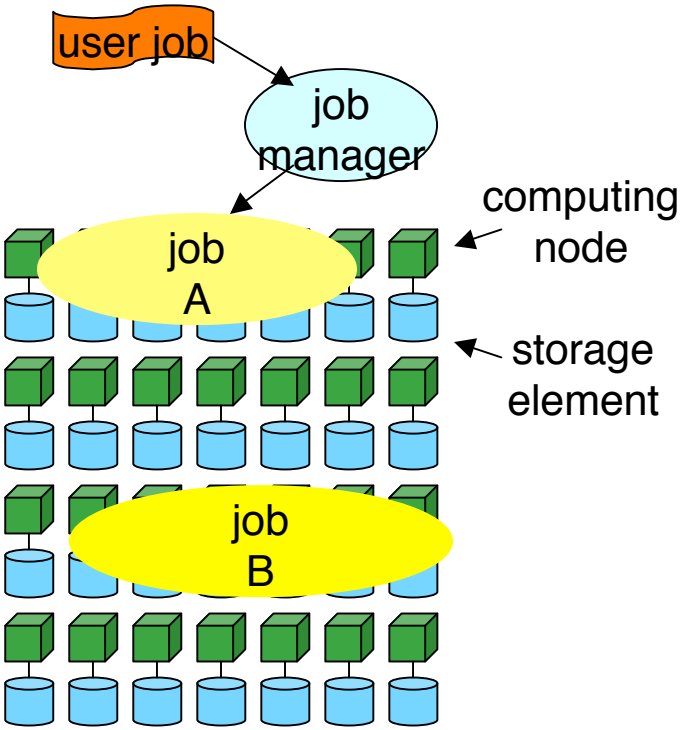
- KEKと東大・素粒子国際研究センター(ICEPP)の共同で技術開発を推進
- 2006年までに 約6万SPECint95の計算機からなるデータ解析システムを国内に構築、ストレージを約1ペタバイトまで段階的に増強
- 補完的役割を担うCERN分室を設立
- ATLAS実験のデータ・チャレンジに参加
- NIIのSuperSINET計画にGrid/アトラスの専用回線
  - KEK-ICEPP間に 1 Gbps ('02.1~) DCデータの高速転送&格納
  - KEK-東工大間に 1 Gbps ('02.10~) Geant4の大規模シミュレーション

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

CPU vs Storage



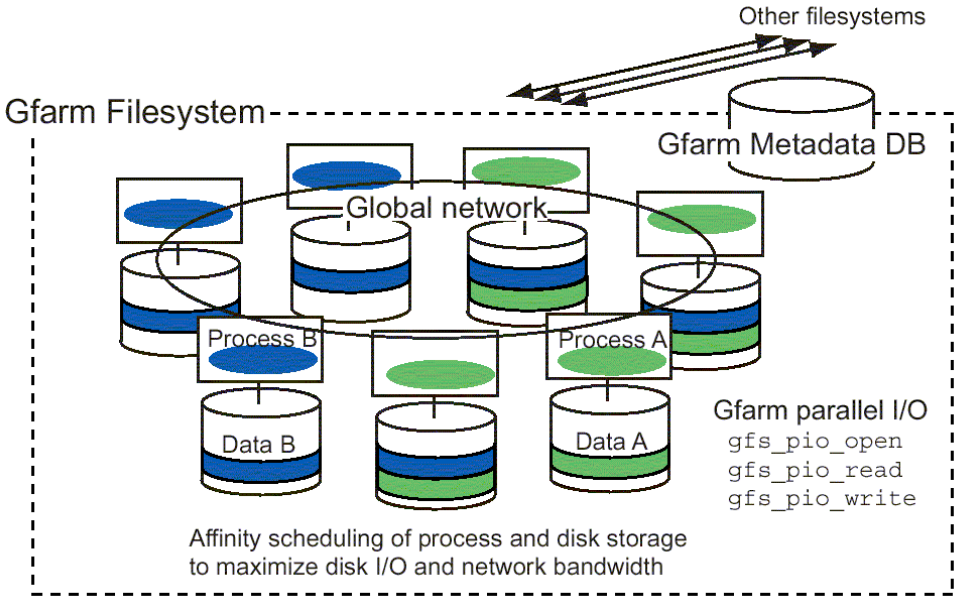


- ● システム管理・ファイル管理が容易
- ✖ 高速スイッチのスピード/チャンネル数がボトルネック
- ✖ 数百台規模以上のシステムへのスケールが困難

- ● 各ノードで独立したI/O
- ● 数千台規模への拡張可能
- ✖ システム管理・ファイル管理が複雑

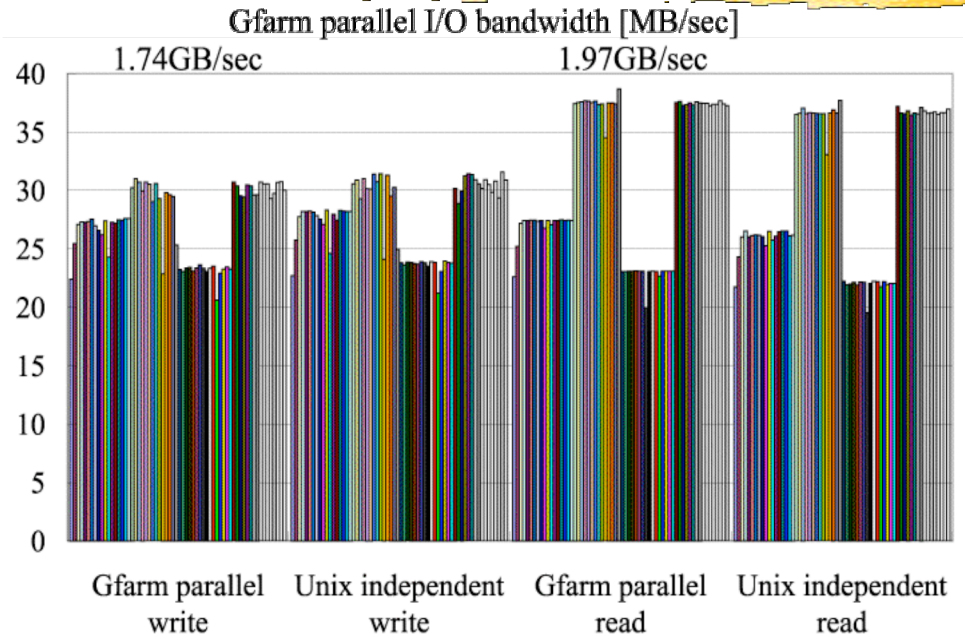
Gfarmとは

- Grid Data Farm
- KEK, 産総研, 東工大によるGridミドルウェア共同開発プロジェクト
- データ処理の独立性を利用してファイルをフラグメントに分割
- データのあるノードでジョブを実行 "owner computes"
- ジョブとフラグメントの履歴をMetadata DBで一括管理
- フラグメントの複製を作成しバックアップおよび負荷分散に
- ユーザーはシングルイメージのGfarm URLとしてファイルを管理



http://datafarm.apgrid.org/

Gfarm parallel I/O benchmark

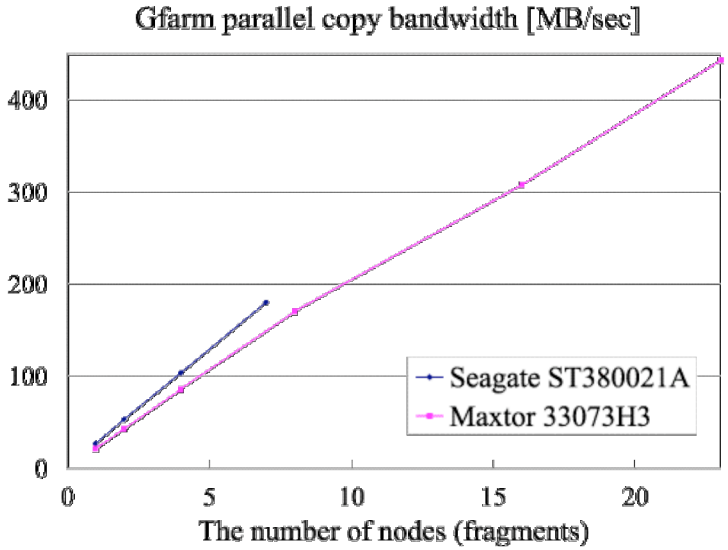



```
write_test(char *fn, void *buf, int size)
{
   GFS_File gf;
   gfs_pio_create(fn, GFS_FILE_WRONLY, mode, &gf;);
   gfs_pio_set_view_local(gf, lflag);
   gfs_pio_write(gf, buf, size, &np);
   gfs_pio_close(gf);
}
```

64 nodes, 640 GB file

File Replication of 10 GB file fragments through Myrinet 2000

443MB/s at 23 parallel streams



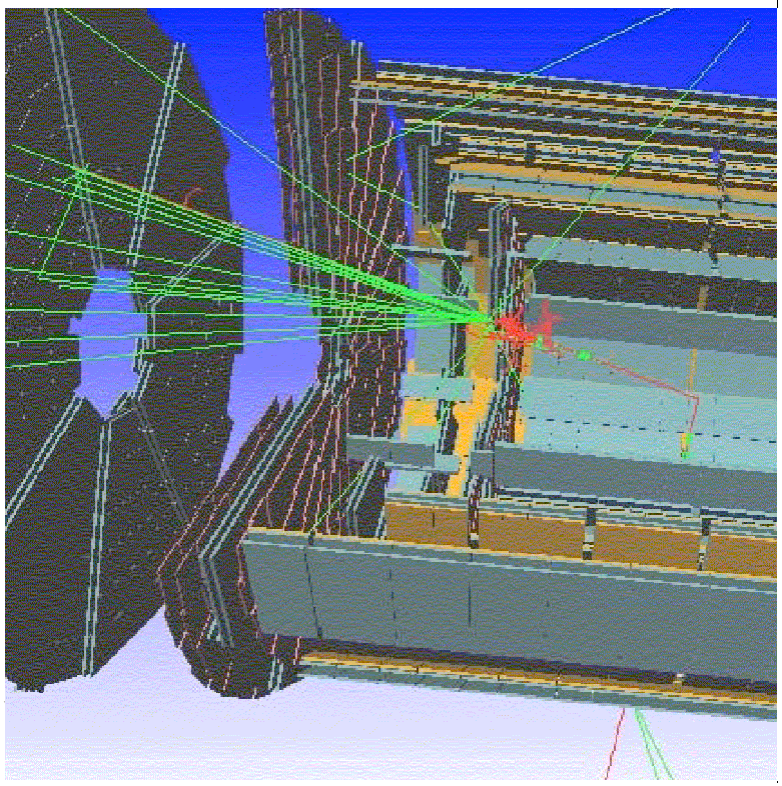
Presto-III PC Cluster @ Titech

- # of Nodes 256
- CPU AMD Athlon x 2
  (Thunderbird, AXIA core) 1.33GHz (FSB=133MHz)
- Motherboard ASUS A7V13
  (VIA KT133A Chipset)
- Memory 768MB
- HDD 40GB
- OS Debian/Lucie 2.14.7
- g++ 2.95.4
- Network Card 1 DEC 21140AF
- Network Card 2 Myricom Myrinet2000

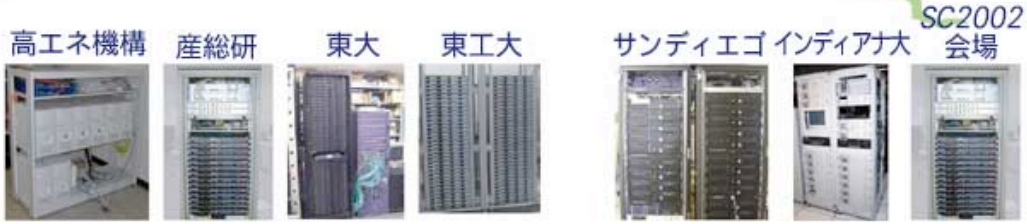
- 47-th in "TOP500" (2nd in PC cluster)

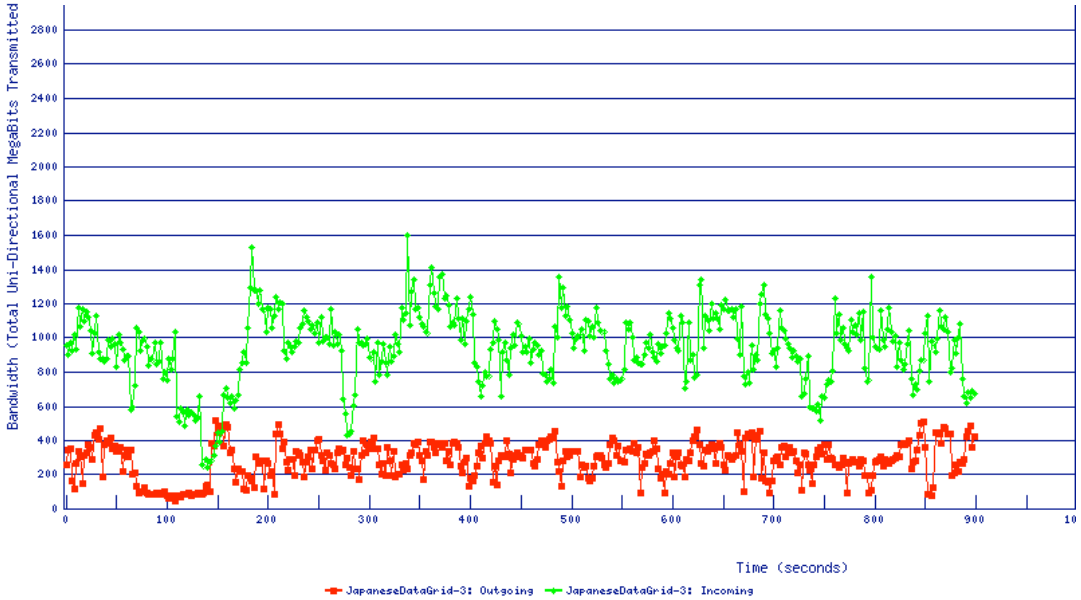
ATLAS実験の検出器シミュレーション

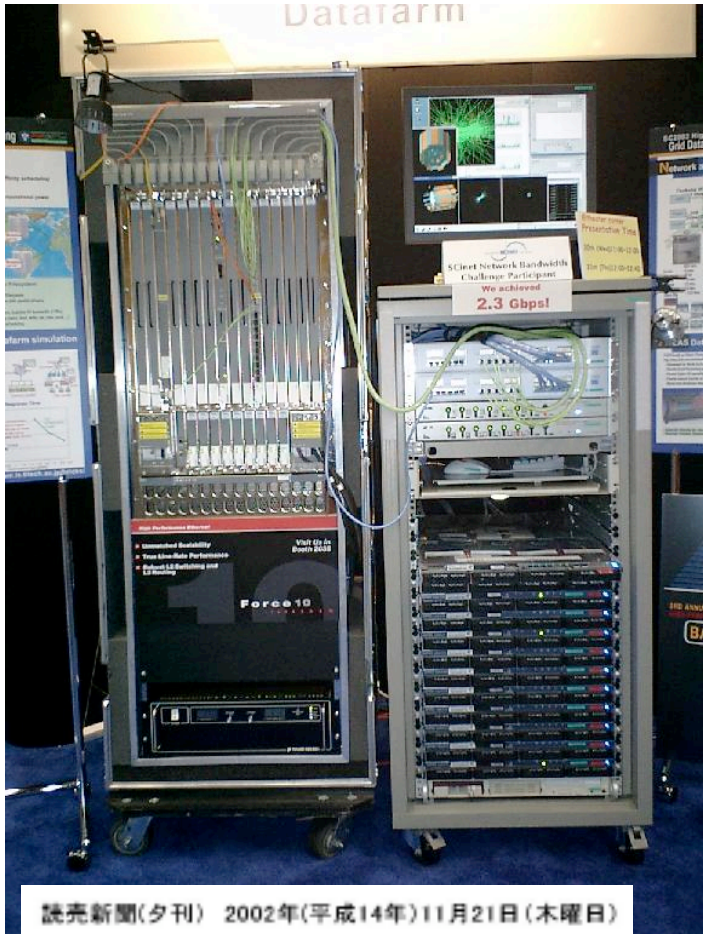
- FADS/Goofy: Framework for ATLAS Detector Simulation / Geant4-based Object-oriented Folly
- Geant4ベースのAtlas Detector Simulationのための フレームワーク
- SC2002デモ
- ATLAS Data Challenge 1に あわせ 10^6イベントを生成 400CPU　～　約2日間
- GfarmによるGigabit級WANの データ複製管理と広域分散データ解析



Cluster and Network setting for SC2002 Bandwidth Challenge (9/3)







読売新聞(夕刊) 2002年(平成14年)11月21日(木曜日)

◆毎秒8800万文字分のデータ転送

産業技術総合研究所は21日未明、複数のパソコンを接続して一台のコンピューターとして機能させ（PCクラスター）、さらにこのPCクラスターを複数統合する新技術を開発、日米間の超大規模データ転送実験に成功したと発表した。

同研究所は新開発のソフトウエアで、同研究所や高エネルギー加速器研究機構、東京大学、米インディアナ大学など日米7拠点、計190台のパソコンを統合したPCクラスターを構成、日米間で1秒当たり7.07ギガ・ビットのデータ転送速度を達成した。これは毎秒8800万文字、CDなら一枚を5・7秒で転送する速さになる。

Ganglia Cluster Toolkit

Cluster Report for Wed, 13 Nov 2002 02:41:29 +0900

Get Fresh Data

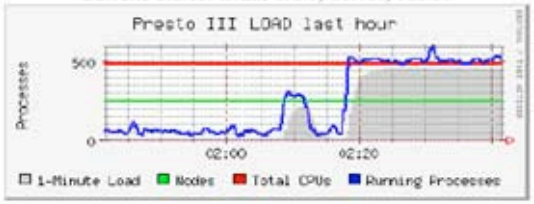
Metric load_one Last hour Sorted descending Physical View

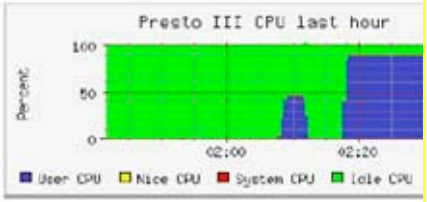
Grid > Presto III > --Choose a Node

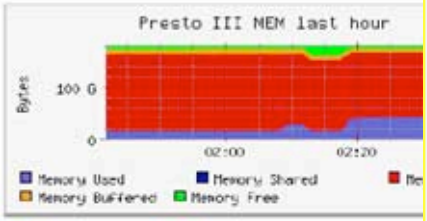
Overview of Presto III

There are **245** nodes (**490** CPUs) up and running.
There are **12** nodes down.

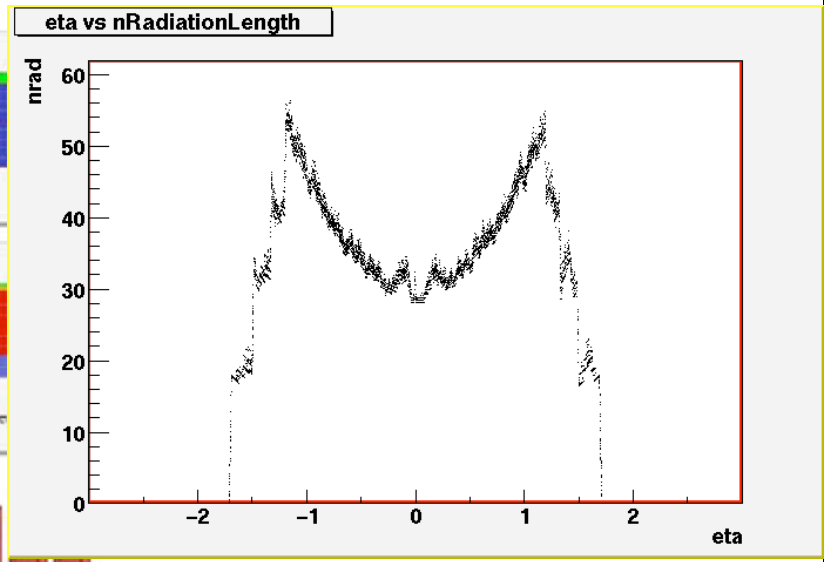
Current Cluster Load: 451.9, 440.46, 318.29

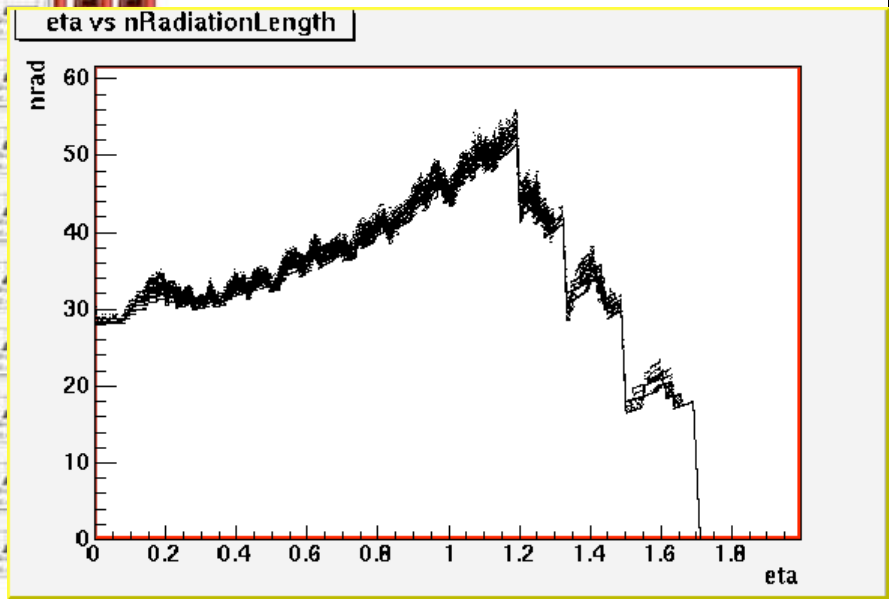






Snapshot of Presto III | Legend





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2007年から始まるLHC/ATLAS実験のための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
- SC2002で行ったバンド幅チャレンジを今後日欧間にも拡大