アトラス実験に向けた大規模計算機システムの構築とその性能

東京大学素粒子物理国際研究センター

磯部忠昭、松永浩之、上田郁夫、小林富雄、坂本宏、
真下哲郎、松井長隆

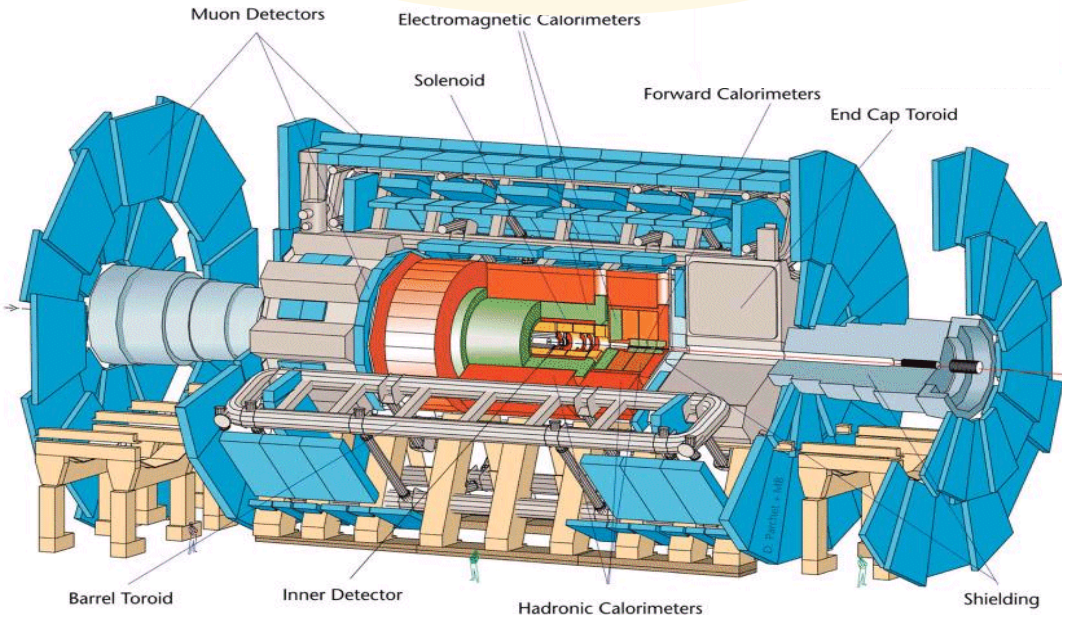
日本物理学会　２００９年秋季大会@甲南大学　岡本キャンパス

Contents

- アトラス実験
- 東京地域解析センター
  - LCG
  - ATLAS-Japanユーザー環境
- アトラス性能テスト (STEP '09)
- まとめ

アトラス(ATLAS)実験

アトラス検出器全体像
直径 22 m、長さ 44 m、重さ 7000 t
1.1億チャンネルのセンサー



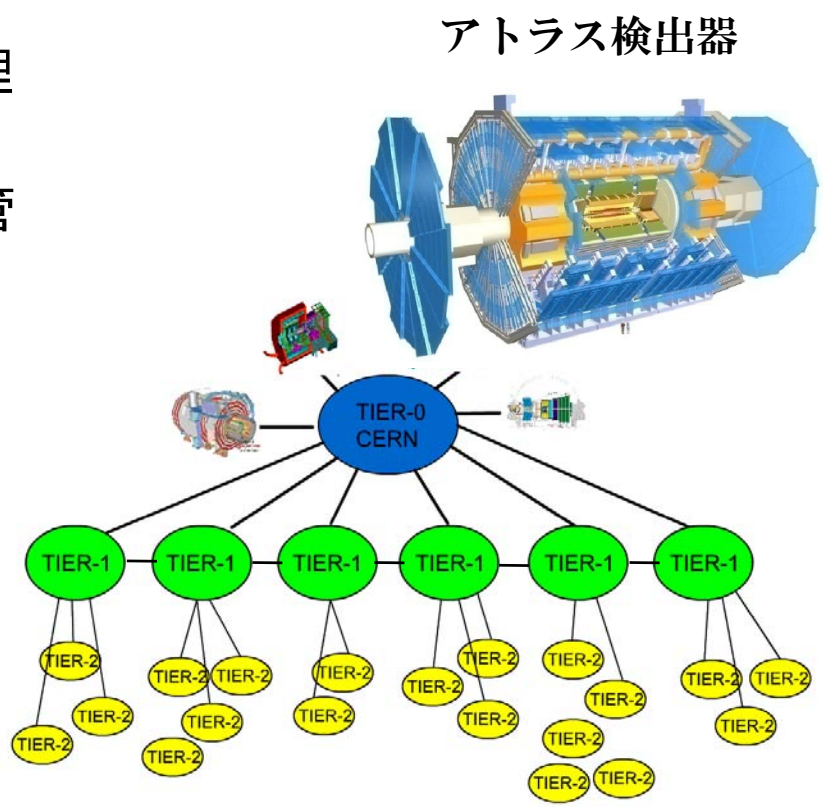
- 約2700人(37カ国)による国際共同研究
- 欧州原子核研究機構(CERN)のLHC加速器を用いた実験
  - 世界最高エネルギー14TeVでの陽子+陽子衝突
  - ヒッグス粒子やSUSY・余剰次元事象の発見が期待されている
- 25ナノ秒ごとに起きる事象を観測、200Hzで1イベントあたり約2MBytesのデータが記録される
  - 1年あたりでは数ペタバイトのデータが取得される
- この多大なデータを解析する為、WLCG(Worldwide LHC Computing Grid)と呼ばれるグリッド技術を用いたコンピューティングプロジェクトが進んでいる

アトラスで生成される主なデータ

- RAWデータ (RAW)
  - 測定器から読み出された生データ
  - Byte Stream
  - ~1.6MByte/event
- Event Summary Data (ESD)
  - データ較正、イベント再構成したもの
  - トラック、クラスター、カロリメーターセル、…
  - 実験初期の検出器の理解に必須
  - ~0.8MByte/event (MCの場合1.1MByte/event)
- Analysis Object Data (AOD)
  - 物理解析用のデータ
  - 電子、ミューオン、光子、ジェット、ミッシングエネルギー …
  - ~150kByte/event (MCの場合180kByte/event)

アトラスコンピューティングモデル

- Tier (階層)モデル
  - Tier0 (CERN):
    - RAWデータ収集、保管と一次処理
  - Tier1 (10ヶ所):
    - RAWデータ複製保管、ESDの保管とデータの再処理
    - Tier1サイトは互いに広帯域ネットワークで連携している
  - Tier2 :
    - シミュレーションデータの生成
    - ユーザによるデータ解析
- 各Tier1毎に資源がまとめられCloud(雲)が定義される。
  - 各雲にデータの保管やシミュレーションデータ生成等のデータ処理が割り当てられる
  - 各Tier2は一つのTier1雲に所属す



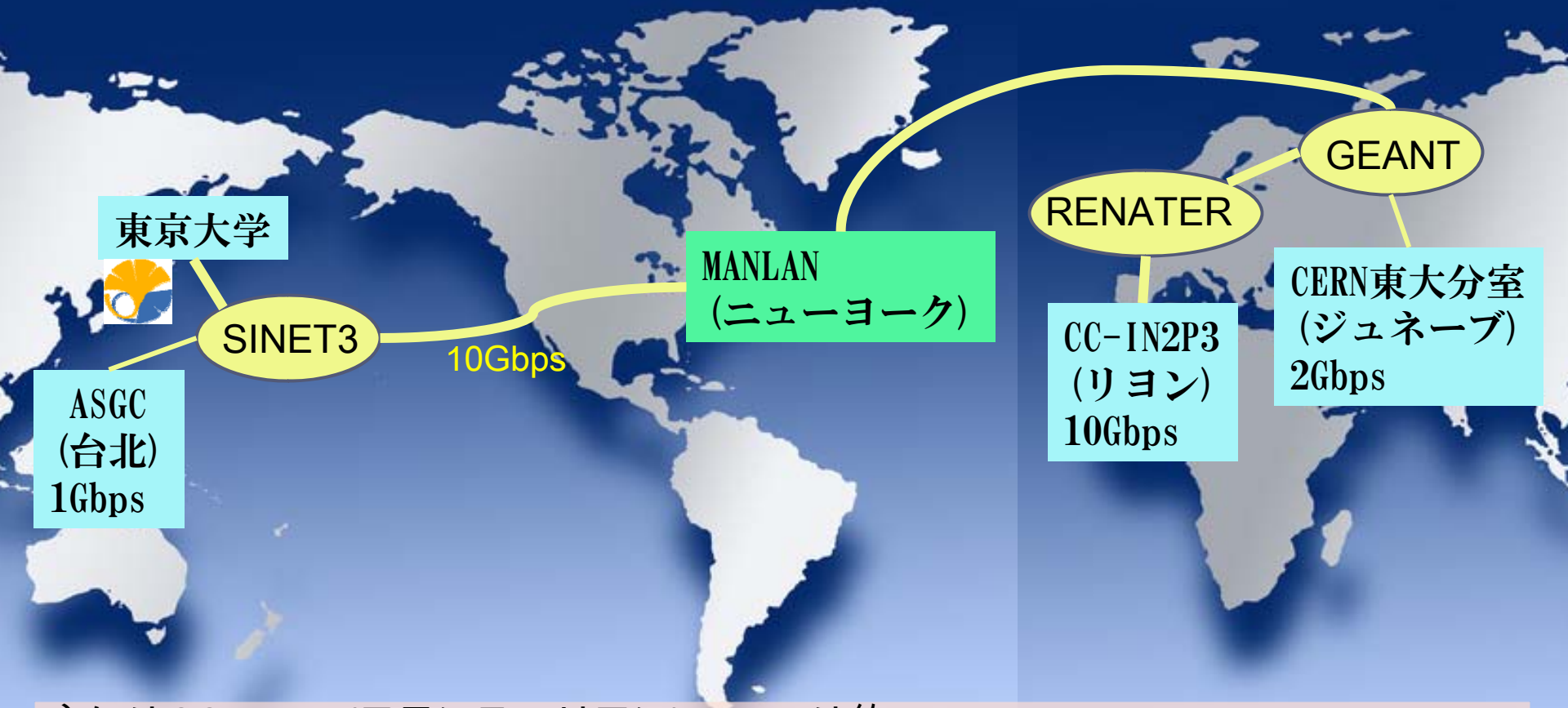
地域解析センター@東京大学

- 国内でのデータ解析を円滑に行う為、Tier2サイトとして参加
  - 東京大学本郷の素粒子物理国際研究センターにあるTier2 に対する Tier1 はフランス・リヨンのCC-IN2P3
  - gLiteグリッドミドルウェアを用いて運用されている
    - CE: torque+maui
    - SE: DPM (Disk Pool Manager)
- OSは主にScientific Linux CERNを使用
- アジア太平洋地域では、唯一100% のAODデータ保持しているTier2

東京大学地域解析センターの計算資源

- 2007年から現システム稼動
- 計算サーバー: 650 ノード x 4 コア
  - 1Gbpsでネットワークへ接続
  - 120 ノードをLCGのアトラス用に供給
  - その他のノードをLCGヘッドノードや各種サーバー、ATLAS-JAPAN専用バッチノードとして使用
- ディスクアレイ: 140台
  - 1台に 500GB x 16 HDD、RAID-6 で使用
  - 65台 (~410TB) をLCGのアトラス用に供給
  - その他のディスクは各種サーバー(AFS, CASTOR…)やATLAS-JAPAN専用のデータ置き場として使用
- ファイルサーバー28台
  - 5台のディスクアレイと4GbpsのFibre-Channelで接続
  - 10GbpsのNIC(Chelsio)にてネットワークへ接続
- テープライブラリ:8000本
  - LTO-3 (400GB/本)

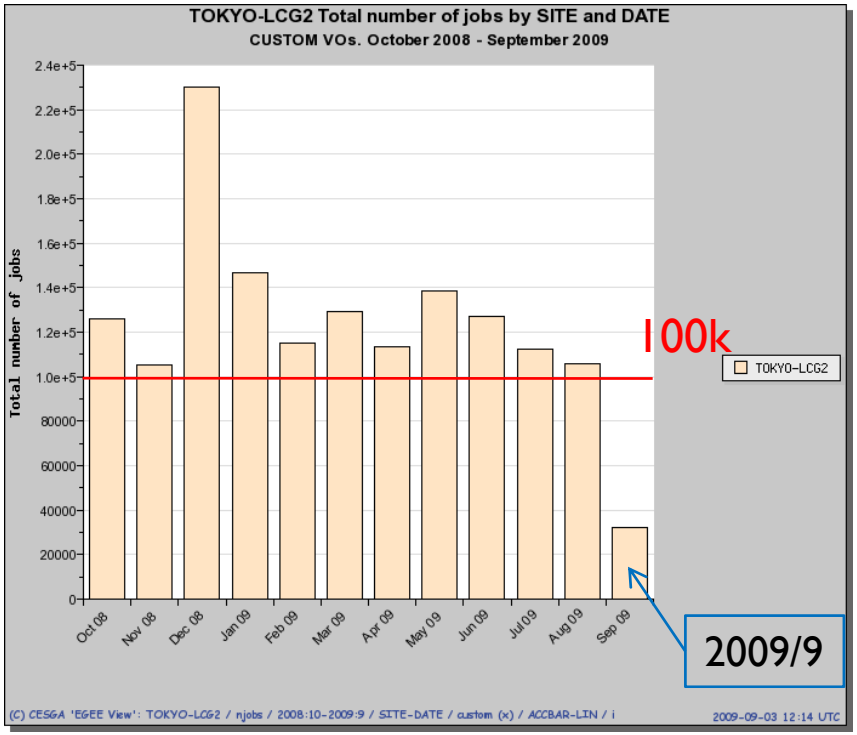
主に東京大学地域解析センターが国際回線を用いて通信を行う相手



主にはCC-IN2P3(フランス・リヨン)、RTTは約280msec
gridftpを用いてデータ転送を行う
アトラスのモデル想定外だが、CERN東大分室とのネットワークも我々には重要

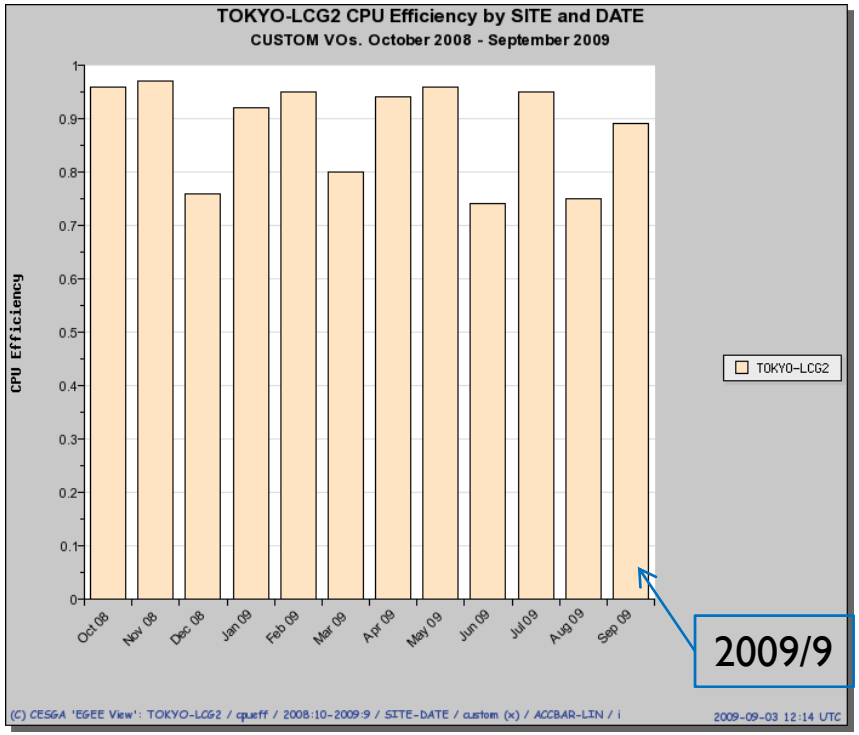
東京地域解析センターここ1年のLCG運用実績

毎月のジョブプロセス数



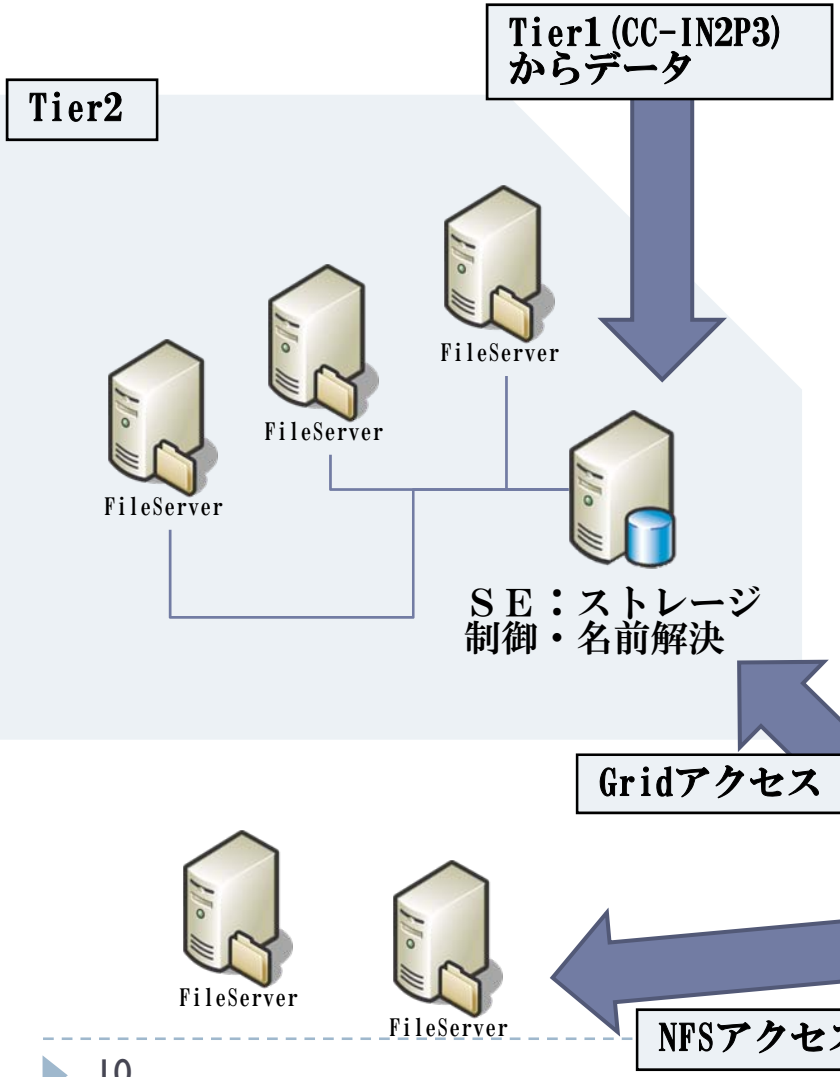
毎月100kジョブ強をプロセス(1500k/year)
多くがATLAS-Geant4のジョブと思われる

CPU効率 (CPU Time/Wall Time)



年平均90%のCPU効率

東京地域解析センターユーザーのデータアクセス



- InteractiveノードからTokyo-Tier2上のデータへ直接アクセスが可能
  - アクセス用のAPIが必要
  - rfioやgridftpを使ったstaging
- 自分でgrid上からデータをコピーしNFSディスク上へ置いてもよい
- TOKYO-Tier2へのデータアクセスはバッチノード、グリッドワーカーノード両方から可能
- LCG、非LCG領域共に同じローカルネットワーク上にある

東京地域解析センターユーザーの主な解析スキーム

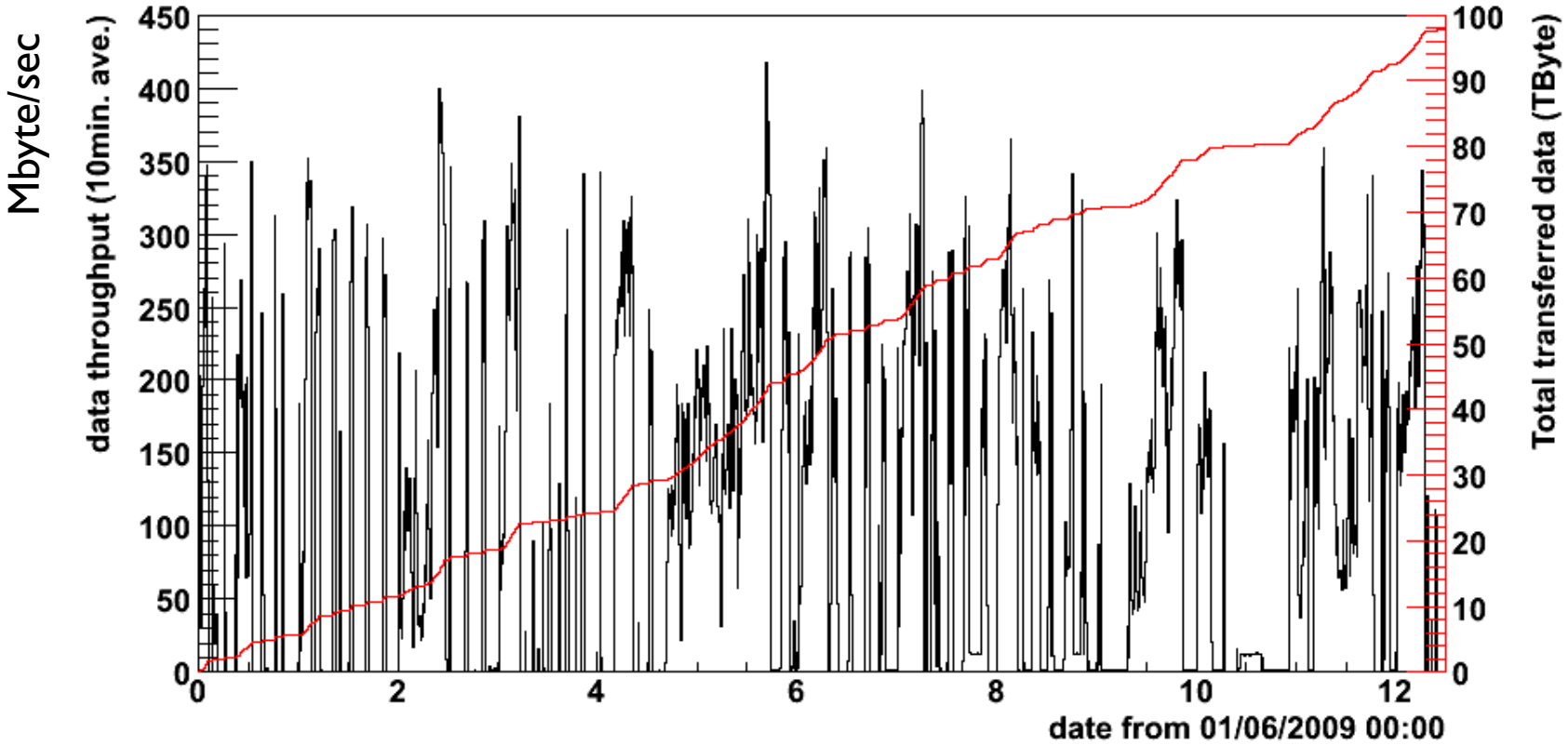
- グリッドへジョブを投げる
  - ATLASではEGEE、OSG、NDGFを統一したインターフェイスで使うため、2種類のシステム(panda, gLite-wms)を使用
  - ユーザーはATLASで開発されたユーザーに優しいコマンドを使う事でLCGサイトへジョブを投げられる
    - athena -> pathena, ganga
    - ジョブのスプリットから入力データファイルの準備、出力データファイルのカタログ登録までやってくれる
    - モニターも充実している
  - ATLASのジョブマネージメントシステムやグリッドミドルウェアといった階層化されたシステムを通す為、開始まで時間がかかる
- ATLAS-Japan専用バッチノードを使う
  - LSFで制御されている
  - グリッドへジョブを投げた場合に比べ、比較的早くジョブが始まる
  - Tier2のファイルにアクセスする場合、ファイルカタログ等の準備が必要

STEP '09

- "Scale Testing for the Experimental Programme"
- 実データが出た時を想定し、モンテカルロ生成、データreconstruction、データ配布、データ解析を同時に行い、問題点を洗い出す
  - LHC4実験合同での演習
  - 2009年6月前半（2週間弱）に行われた
- この期間で100TByte強のAODデータが生成されると想定し、データの配布が行われた
  - 200MByte/secでのAOD生成、1日あたり14時間の実験データ取得を仮定
- 東京地域解析センターには幾つかのタイプのジョブがグリッドワーカーノードへ投入された
  - Pandaによるモンテカルロ生成: 50%Fiar Shareに~8kジョブ投入
  - Panda/gLite-WNSによるユーザーの解析ジョブ: 50%Fiar Shareに~36kジョブ投入
    - ATLASでは10MByte/sec/1job程度のfile readを想定している

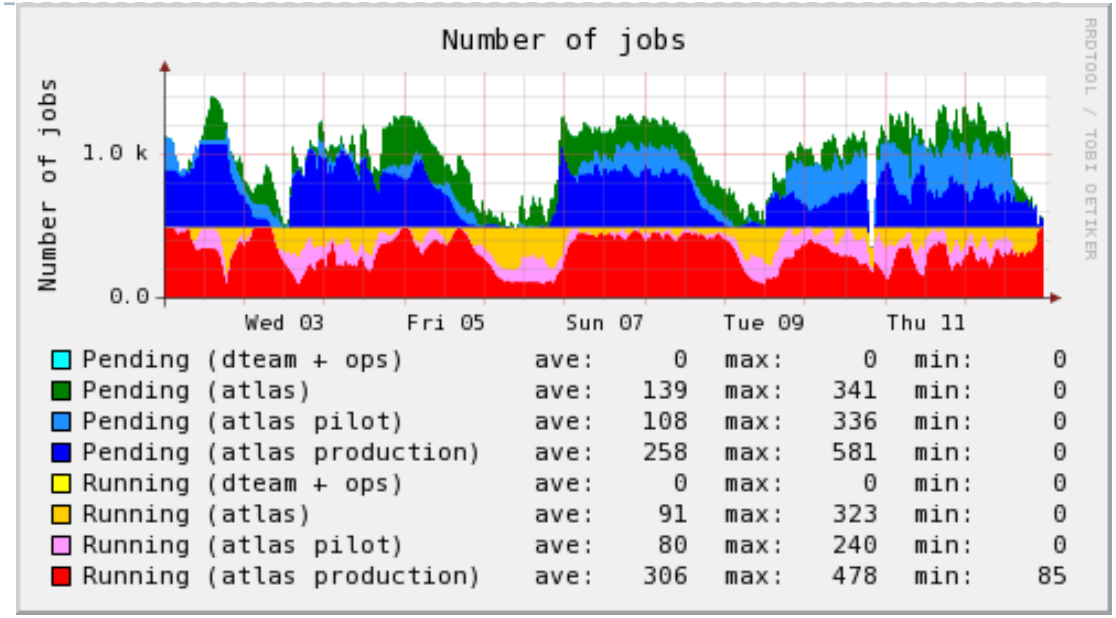
＊非グリッドノードはATLAS中央管理の範囲外で、テストの対象外

STEP '09 パフォーマンス
データ転送(CC-IN2P3→東京)

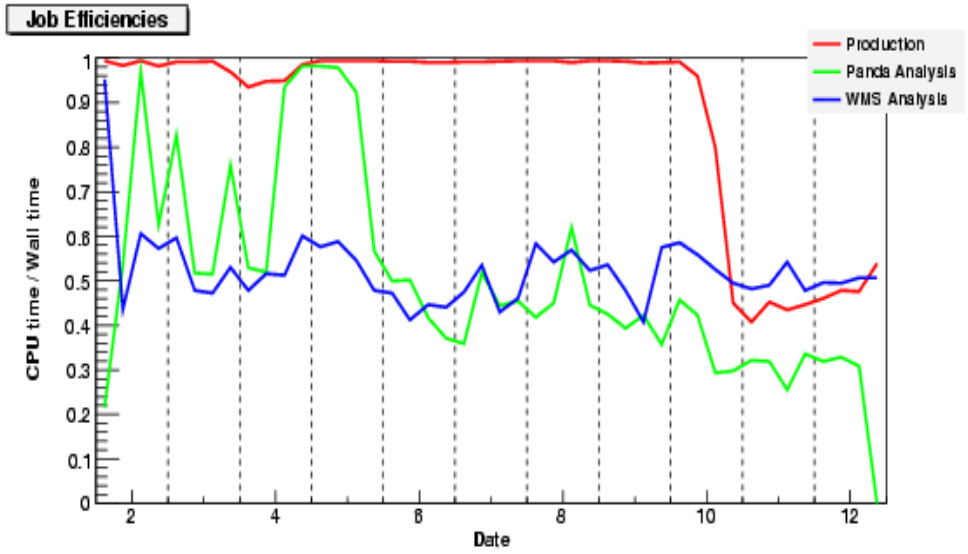


- おもにAODファイルを転送(~3GByte/file)
- 2週間弱の転送で100TByteほどのデータを転送
  - データの転送能力がある事を示す

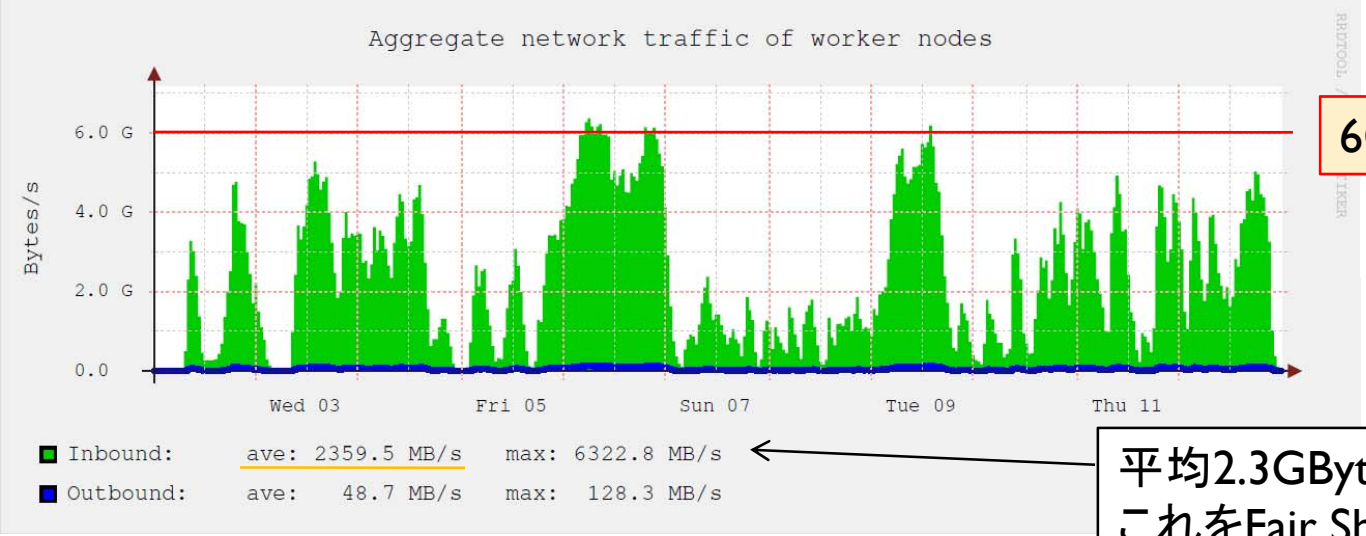
STEP '09 パフォーマンス ジョブCPU効率



- モンテカルロ生成ジョブはほぼ100%の効率
- 解析ジョブの効率 (CPU Time/Wall Time) は0.5程度
  - AODファイルのローカルディスクへのコピーに時間がかかる



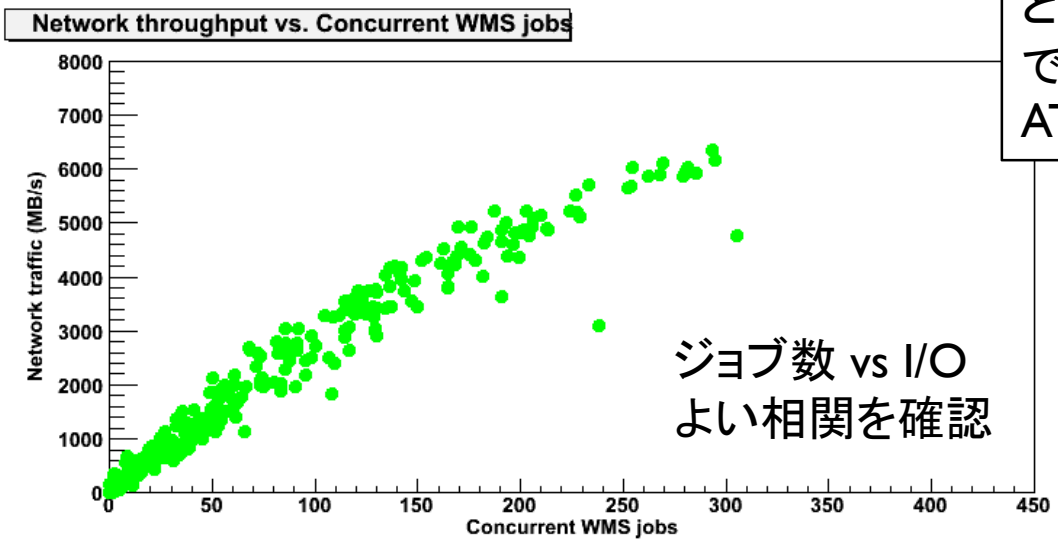
STEP '09 パフォーマンス
ファイルサーバー→LCGバッチノードのI/O



6GByte/sec

I台のファイルサーバーあたり
470MByte/sec相当

平均2.3GByte/sec:
これをFair Share50%でのプロセスと考えるとIcore辺り9.2MByte/secでのイベントプロセス相当で、ATLAS想定程度の性能



ジョブ数 vs I/O
よい相関を確認

まとめと展望

- 東京大学では2007年春より現地域解析センターが稼動しており、これまで大きな問題なく安定して稼動している
  - LCG用以外にATLAS-JAPAN専用のシステムを設置
- ATLASによる大規模なテストにて東京地域解析センターの性能を測定し、ATLASで想定される程度の性能を示した
  - 2週間弱で~100TByteのデータをTier1から転送し、生成されるAODデータの転送能力があることがわかった
  - 解析ジョブのCPU効率(CPU Time/Wall Time)は~50%程度で、~9.2MByte/sec/1coreでデータを解析、想定される解析ジョブの処理能力がある事がわかった
- 現システムは3年間のリースであり、2010年からは新しいシステムに移行する予定で、新システムへの移行の為の準備を進めている