Tier 1 at Brookhaven (US / ATLAS)

Bruce G. Gibbard

LCG Workshop

CERN

23-24 March 2004

BROOKHAVEN NATIONAL LABORATORY

Tier 1: Interface to Grid of US Facilities

- Grid of Distributed Regional Resources Includes …
  - Tier 1 Facility at Brookhaven
    - Currently operational at ~2-3% of required 2008 capacity
    - ~14% of device count complexity scale (Total facility >100% now)
  - 5 Permanent Tier 2 Facilities
    - Scheduled for selection beginning in 2004
    - Currently there are 2 Prototype Tier 2's
      - **Indiana U / U of Chicago**
      - **Boston U**
  - ~9 Currently Active Tier 3 (Institutional) Facilities
  - WAN Coordination Activity
  - Program of Grid R&D Activities
    - **Based on Grid Projects (**PPDG, GriPhyN, iVDGL, EU Data Grid, EGEE, etc.**)**
  - Grid Production & Production Support Effort

Regional Center Mission

- Contribute to the ATLAS Virtual Computing Center per agreed levels of service including …
  - capacities as a function of time
    - Compute (Production, Analysis)
    - Storage (Online, Tertiary)
    - Network (LAN, WAN)
  - levels of support
    - Up times, Response time, Etc.
- Guarantee effective participation by U.S. physicists in the ATLAS physics program & adequate support for physics topics of regional interest
  - Direct access to and analysis of physics data sets
  - Other activities to support regional usage
    - AFS mirror of ATLAS repository, including nightly builds, for example

Tier 1 Facility Proper

- Functions
  - Primary U.S. data repository for ATLAS
  - Programmatic event selection and AOD, Etc. regeneration
  - *Chaotic* high level analysis by individuals
    - Especially for large data set analyses
  - Significant source of Monte Carlo
  - Re-reconstruction as needed
  - Technical support for smaller US computing resource centers
- Co-located and operated with the RHIC Computing Facility
  - To date a very synergistic relationship
  - Some recent increased divergence (Linux RH version)
  - Substantial benefit from cross use of idle resources (2400 CPU's)

Tier 1 Facility Evolution for FY '04

- Addition of 2 FTE's and modest equipment upgrade in '04
  - 2 FTE increase expected to translate into 4 new hires distributed over year (… brings total to 6.5 FTE's for year)
    - 1 new hire now in place
    - 2 additional offers now accepted
    - 1 opening still to be filled
  - Central NFS Disk: 11 TBytes → 23 TBytes (factor of 2)
  - Dedicated Access to Tape Storage: 30 → 60 MBytes/sec (factor of 2)
    - StorageTek (4.5 PBytes, 1 GByte/sec configuration)
    - HPSS
  - CPU Farm: 30 kSPECint2k → 130 kSPECint2k (factor of 4)
    - 48 x (2 x 3.06 GHz, 1 GB, 360 GB) … so also 16 TB local IDE disk
    - Total CPU count now ~220
    - Available via both LSF and Condor (& PBS for very limited LCG-1)

Major Tier 1 Technical Activities

- Study of alternative disk technologies
- Storage Element evaluation, deployment and optimization
- Addressing increased complexity of cyber security and AAA, especially for Grid
- Grid3+ deployment & operation
- Continue LCG deployment & evolution (LCG-1 -> LCG-2)
- ATLAS Data Challenge 2 (DC2) (primarily via Grid3+)
- Wide Area Network Issues (need to pursue timely upgrade from current OC12 – 622Mb/sec)
- ATLAS divergences from RHIC

Disk Technology Evaluation

- Current primary (RCF/ACF) Technology (90% RHIC):
  - Sun/Solaris NFS servers (~32 2-4 CPU SMP's)
  - RAID 5 Disk from MTI, Data Direct, & Zyzzyx (~200 TB)
  - FibreChannel SAN connectivity via Brocade switches
- Increase Performance/Functionality and/or Decrease Cost
  - Full function access to storage at ~4 x the I/O rate
  - Storage at ~¼ x the cost
  - Probably a 2 tiered approach
- Currently Investigating
  - Panasas – Very high performance (commercial product)
  - dCache – Very cost effective (Fermilab/DESY development)
  - Lustre – Maybe both (or neither)

Analysis Model: All ESD Resident on Disk
(Preferably at Each Tier 1)

- Enables ~24 hour selection/regeneration passes (versus ~month if tape stored) – faster, better tuned, more consistent selection
- Allows navigation for individual events (to all processed, *though not Raw,* data) without recourse to tape and associated delay – faster more detailed analysis of larger consistently selected data sets
- Avoids contention between analyses over ESD disk space and the need to develop complex algorithms to optimize management of that space – better result with less effort
- **Complete set on disk at single Tier 1**
  - **Reduced sensitivity to performance of multiple Tier 1's, intervening network (transatlantic) & middleware – improved system reliability, availability, robustness and performance**

Cost Impact of All ESD on *Local* Disk

- Assumptions
  - Increase from 480 TB to 1 PB of total disk
    - Some associated increase in CPU and infrastructure
  - Simple extension of current technology
    - Using a conservative technology so cost may be over estimated
    - ***Disk distributed in Farm nodes could be a factor of 4 less expensive***
  - Personnel requirement unchanged
    - Alternative is effort spent optimizing transfer and caching schemes
- Tier 1 Facility cost impact through 2008 (33% versus 100% on disk)

	% δ
Labor	0%
MST (maint, licen, etc)	5%
Capital Equipment	29%
Total	9%

  - Since facility cost is not dominated by hardware,

    …reduction to "1/3 disk model" certainly reduces cost but not dramatically

    …and as much as a factor of 4 less in a distributed disk environment

Tier 2 Facilities

- 5 Permanent Tier 2 Facilities
  - Primary resource for simulation
  - Empower individual institutions and small groups to do autonomous analyses using more directly accessible and locally managed resources
- Currently 2 Prototype Tier 2's selected for ability to rapidly contribute to Grid development
- Permanent Tier 2 will be selected to leverage strong institutional resources
  - ***Selection of first three scheduled for spring/summer 2004***
  - Currently 9 active Tier 3's in addition to prototype Tier 2's; all candidates to be permanent Tier 2's
  - Expectation that Prototypes will become permanent but not guaranteed
- Aggregate of 5 permanent Tier 2's to be comparable to Tier 1 in CPU
  - Near term, DC2 & DC3 Tier 2's will actually exceed Tier 1 in CPU