



U.S. ATLAS Computing Facilities

Bruce G. Gibbard

DOE/NSF LHC Computing Review

Germantown, MD

8 July, 2004

Content



- * Overview
- * ATLAS Computing Model & Resource Estimate (Revision)
- * US ATLAS Computing Facilities Plan (Revisions & Status)
- * Facilities Activities
- * Data Challenge 2
- * Progress on Milestones
- * Conclusions

US ATLAS Computing Facilities



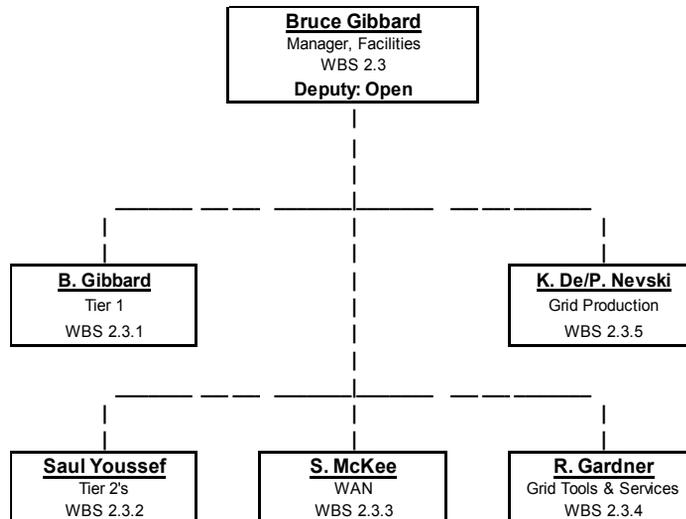
- ✳ Supply capacities to the ATLAS Distributed Virtual Offline Computing Center
 - At levels agreed to in a computing resource MoU (Yet to be written)
- ✳ Guarantee the Computing Capabilities & Capacities Required for Effective Participation by U.S. Physicists in the ATLAS Physics Program
 - Direct access to and analysis of physics data sets
 - Simulation, re-reconstruction, and reorganization of data as required to support such analyses

2.3 US ATLAS Facilities



- ✳ Coordinated Grid of Distributed Resources Including ...
 - Tier 1 Facility at Brookhaven – Bruce Gibbard
 - ✳ Currently operational at ~2.5% of required 2008 capacity
 - 5 Permanent Tier 2 Facilities – Saul Youssef
 - ✳ Selection of 3 schedule for next 4 months
 - ✳ Currently there are 2 Prototype Tier 2's
 - Indiana U – Fred Luehring / U of Chicago – Rob Gardner
 - Boston U – Saul Youssef
 - ~9 Currently Active Tier 3 (Institutional) Facilities
 - WAN Coordination Activity – Shawn McKee
 - Program of Grid R&D Activities – Rob Gardner
 - Based on Grid Projects (PPDG, GriPhyN, iVDGL, EU Data Grid, EGEE, etc.)
 - Grid Production & Production Support Effort – Kaushik De/Pavel Nevski

Corresponding WBS Organization



ATLAS Facilities Model



❄ ATLAS Virtual Offline Computing Facility

□ Distributed hierarchical set of resources – recently revised

⌘ CERN Tier 0 – Exactly 1

- Record Raw Data, Calibrate, Reconstruct, Distribute Raw & DST
- Totaling: 1.5 PB Disk, 7.5 PB Tape, 4.1 MSI2K CPU

⌘ Tier 1 – Expect ~6

- Store, serve, reprocess – 1/3 ESD, AOD, TAG's & 1/6 of Raw on Tape
- Totaling: 1.2 PB Disk (x2.5), 1.2 PB Tape, 2.1 MSI2K CPU (x.65)

⌘ Tier 2 – Expect ~4 per supporting Tier 1

- Bulk of simulation, analysis support for ~25 active users
- Store, serve – TAG's, AOD, small select ESD sample
- Totaling: 150 TB Disk, 60 TB Tape, 200 KSI2K CPU (x.3)

⌘ Institutional Facilities & Individual Users

⌘ Acceleration of ramp-up in FY '06 & '07

□ Still a work in progress

⌘ Expect further revision based on input from 17 June Workshop

Revised Tier 1 Capacity Profile (A Snapshot)



- * Extending this to US Tier 1 requirement
 - Full, rather than 1/3, ESD on local disk
 - Additional analysis CPU to exploit this enhanced data access and in particular to support projects of particular interest to US physicists

January 2004 Profile

	2001	2002	2003	2004	2005	2006	2007	2008
CPU (kSI2K)	30	30	30	125	250	750	1,500	5,000
Disk (TBytes)	0.5	12	12	25	50	143	300	1,000
Disk (MBytes/sec)	40	90	90	400	1,000	3,000	6,000	20,000
Tape (PBytes)	0.01	0.05	0.05	0.10	0.21	0.32	0.86	2.05
Tape (MBytes/sec)	10	30	30	60	60	120	240	360
WAN (Mbits/sec)	155	155	622	622	2488	2488	9952	9952

July 2004 Profile

	2001	2002	2003	2004	2005	2006	2007	2008
CPU (kSI2K)	30	30	30	135	232	772	1,737	3,860
Disk (TBytes)	0.5	12	12	24	104	346	778	1,730
Disk (MBytes/sec)	40	90	90	349	6,515	12,744	21,048	35,376
Tape (PBytes)	0.01	0.05	0.05	0.10	0.16	0.27	0.54	1.73
Tape (MBytes/sec)	10	30	30	60	60	120	180	300
WAN (Mbits/sec)	155	155	622	622	2488	2488	9952 (A)	2 x A

- * Disk already dominates cost so past year's disk evaluation work becomes relevant

Disk Technology Evaluation Status



- * **Panasas** – (Commercial)
 - RAID 5 across disk server blades directly to NFS clients (Kernel sensitive)
 - * Cost currently slightly below full cost of Sun/SAN RAID 5 NFS central disk
 - * Prototype type deployed for real use by one of RHIC experiments
 - Expect this to be high end (performance, availability, reliability) disk solution
- * **dCache** – (Fermilab / DESY)
 - Significant performance & robustness testing done but not complete
 - Enstore switch out to HPSS demonstrate but not yet *well* tested
 - SRM (Grid Storage Manager) interface under test
 - Expect this will be very large scale commodity (low price) disk solution
- * **Lustre** – (Open Source)
 - Now only being considered as a backup solution to dCache

Tier 1 Capital Equipment Cost Profiles (\$k) (Snapshot)



January 2004 Estimate (All Central Disk)

	2001	2002	2003	2004	2005	2006	2007	2008
CPU	\$ 30	\$ -	\$ -	\$ 129	\$ 118	\$ 296	\$ 293	\$ 926
Disk	\$ 100	\$ 137	\$ -	\$ 185	\$ 236	\$ 588	\$ 603	\$ 1,793
Tertiary Storage	\$ 46	\$ 25	\$ -	\$ 30	\$ 170	\$ 30	\$ 80	\$ 30
LAN	\$ 79	\$ -	\$ 20	\$ 20	\$ 90	\$ 100	\$ 250	\$ 250
Overhead	\$ 22	\$ 14	\$ 2	\$ 32	\$ 54	\$ 89	\$ 108	\$ 264
Total	\$ 277	\$ 176	\$ 22	\$ 397	\$ 668	\$ 1,104	\$ 1,334	\$ 3,263

July 2004 Estimate (All Central Disk)

	2001	2002	2003	2004	2005	2006	2007	2008
CPU	\$ 30	\$ -	\$ -	\$ 140	\$ 92	\$ 318	\$ 375	\$ 571
Disk	\$ 100	\$ 137	\$ -	\$ 170	\$ 587	\$ 1,195	\$ 1,408	\$ 2,043
Tertiary Storage	\$ 46	\$ 25	\$ -	\$ 30	\$ 170	\$ 30	\$ 80	\$ 150
LAN	\$ 79	\$ -	\$ 20	\$ 20	\$ 90	\$ 100	\$ 250	\$ 200
Overhead	\$ 22	\$ 14	\$ 2	\$ 32	\$ 83	\$ 145	\$ 186	\$ 261
Total	\$ 277	\$ 176	\$ 22	\$ 392	\$ 1,022	\$ 1,788	\$ 2,299	\$ 3,225

July 2004 Estimate (~60% Distributed Disk)

	2001	2002	2003	2004	2005	2006	2007	2008
CPU	\$ 30	\$ -	\$ -	\$ 140	\$ 92	\$ 318	\$ 375	\$ 571
Disk	\$ 100	\$ 137	\$ -	\$ 170	\$ 173	\$ 541	\$ 656	\$ 950
Tertiary Storage	\$ 46	\$ 25	\$ -	\$ 30	\$ 170	\$ 30	\$ 80	\$ 150
LAN	\$ 79	\$ -	\$ 20	\$ 20	\$ 90	\$ 100	\$ 250	\$ 200
Overhead	\$ 22	\$ 14	\$ 2	\$ 32	\$ 46	\$ 87	\$ 120	\$ 165
Total	\$ 277	\$ 176	\$ 22	\$ 392	\$ 571	\$ 1,077	\$ 1,481	\$ 2,035

Tier 1 Facility Evolution in FY '04



✳ Addition of 2 FTE's

- 2 FTE increase => 4 new hires
- ATLAS supported Tier 1 staff 4.5 last year => 8.5 (-1) now
- But have lost Rich Baker, Deputy ATLAS Facilities Manager
... hope to replace him soon

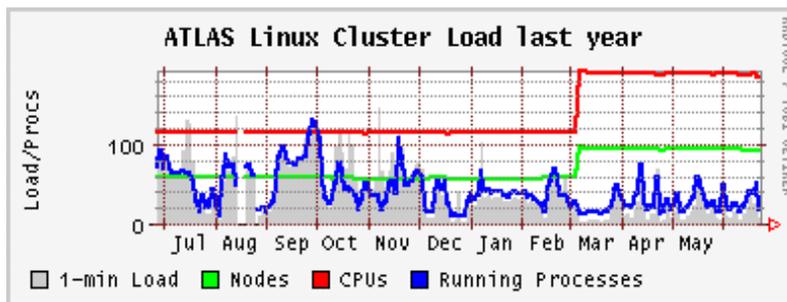
✳ Modest equipment upgrade

- Disk: 11 TBytes → 23 TBytes (factor of 2)
- CPU Farm: 30 kSPECint2000 → 130 kSPECint2000 (factor of 4)
 - ⌘ 48 x (2 x 3.06 GHz, 1 GB, 360 GB) ... so also 16 TB local IDE disk
 - ⌘ First processor farm upgrade since FY '01 (3 years)
- Robotic Tape Storage: 30 MBytes/sec → 60 MBytes/sec (factor of 2)

Tier 1 Deployment/Operation Activities



- ✳ Grid 3 (+) & ATLAS Data Challenge 2 (DC2) support
 - Major effort over past several months
- ✳ LHC Computing Grid deployment (LCG-1 -> LCG-2)
 - Very limited equipment deployed using only modest effort
- ✳ Still limited general chaotic use of facility



Combined Test Beam Support



- ✳ Combined Test Beam activities currently underway are expected to produce a ramp up in demand
 - Support calibrations
 - ✳ Several million single particle events
 - Store major samples of test beam data on local disk
 - Supply capacity to do major re-reconstruction of test beam data as new versions of software become available
 - Store and distribute test beam data for individual analyses
- ✳ Issues of support for test beam activities and other less monolithic computing in the context of DC2 production
 - Resources management policies (Queues, disk, etc.)

Tier 1 Development Activities



* Study of alternate disk technologies

- Already discussed



* Cyber security and AAA for Grid

- Continued evolution of Grid User Management System (GUMS)
- Testing / Deploying VOM/VOMS/VOX/VOMRS
- Consolidation of ATLAS VO registry with US ATLAS as a subgroup
- Privilege management project underway in collaboration with Fermilab/CMS

* BNL-Tier 1 CERN-Tier 0 data transfer optimization

* Storage Element (SRM) evaluation, testing & deployment

Data Transfer / WAN Issues



* From Last Review

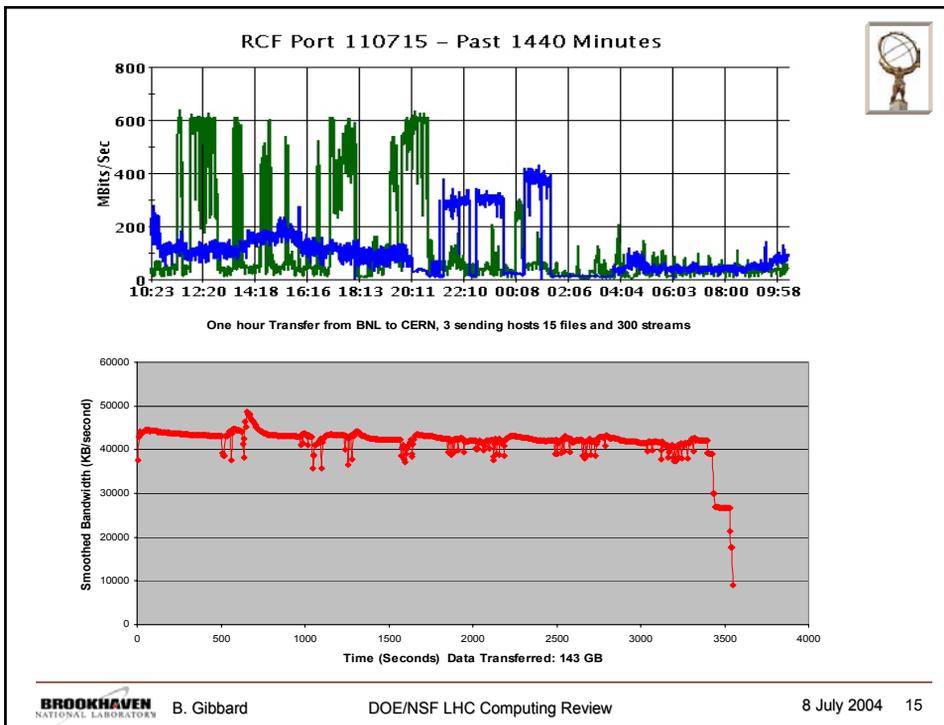
- Feb '04: Measure throughput and understand limitation in current CERN → BNL data transfers
- Apr '04: Do a first, limited effort and very limited equipment, optimization of data transfers for use in DC2
 - * Hopefully, at least 5-10 TB of data in 2 weeks
 - DC2 requires 4-8 MByte/sec average
 - WAN bandwidth limit at OC12 is ~40 MBytes/sec so should not constrain this goal

* 3 RRD GridFTP servers doing bulk disk-to-disk data transfers, ***in absence of contention***, achieved

- BNL to CERN: ~45 MBytes/sec => ~4 TB/day
- CERN to BNL: near wire speed, ~70 MBytes/sec
- Afternoon versus midnight "***contention***" effect ~15% (RHIC utilization)

* LCG Service Challenge in Networking now defined

- Sustain data transfer BNL ↔ CERN at ~45 MBytes/sec in July
- Seems well within current scope



Storage Element

- ✳ **Storage Resource Managers (SRM) under evaluation**
- ✳ **HRM/SRM developed by LBNL**
 - HPSS (BNL's HSM) capable out of the box
 - Self contained operation of associated SRM demonstrated
 - Interoperability of Web services version with other SRM's now being studied
- ✳ **dCache/SRM developed Fermilab/DESY**
 - dCache (hope to use for distributed disk management) compatible out of the box
 - With ENSTORE => HPSS demonstration becomes a full function option
 - Self contained behavior now being studied
 - Interoperability with other SRM's will be studied
- ✳ **Evaluation completion expected by September followed by deployment**

B. Gibbard
DOE/NSF LHC Computing Review
8 July 2004 16

2.3.2 Tier 2 Facilities Selection



- ✱ Now soliciting proposals for 3 of 5 planned NSF Funded US ATLAS Tier 2's computing facilities
- ✱ Tier 2 Functions
 - Primary US ATLAS resource for simulation
 - Empower individual institutions and small groups to do relatively autonomous analyses using more directly accessible and locally managed resources
- ✱ Tier 2 Scale
 - Aggregate of 5 permanent Tier 2's should be comparable to Tier 1 in CPU
 - Approximate support levels for each
 - ✱ Operating \$250K => ~2 FTE's plus MST
 - ✱ Equipment \$350k => four year refresh for ~1000 CPU's plus infrastructure
 - A primary selection criterion is ability to leverage strong institutional resources to benefit ATLAS (maximize bang for the buck)

2.3.2 Tier 2 Facilities Selection (2)



- ✱ Criteria Include
 - Qualifications and level of commitment of PI
 - Expertise and experience of staff with facility fabrics and Grids
 - Availability and quality of physical infrastructure (Space, power, HVAC, etc.)
 - Numerical metrics of expected capacity including
 - ✱ CPU and disk capacity dedicated to US ATLAS (SI2000, TBytes)
 - ✱ Integrated non-dedicated resources expected (SI2000-Years)
 - ✱ Dedicated staff supporting Tier 2 operation (FTE's)
 - ✱ Expected non-dedicated support for Tier 2 operation (FTE's)
 - ✱ Wide Area Network connectivity. (Gbits/sec)
 - Operations model (hours of attended operations, etc.)
 - Nature of Education and Outreach program
- ✱ Process
 - Proposals due Sept 30 for Selection by Oct 31 for Funding in 2nd half FY '05
 - Technical committee to produce ordered list for final decision by management

Tier 2 Facilities Capacities



✳ Requirements per Tier 2 in revised ATLAS computing model (Revision is a work in progress)

- CPU '08 => 200 kSI2K
- Disk '08 => 146 TB

✳ For Data Challenges resources at prototype Tier 2's include dedicated plus expected share of institutional while only dedicated resources are considered a the Tier 1 (No assumed share of current RHIC ~1500 kSI2K)

Tier 2/Tier 1 Resources

	For DC2		Expected for DC3	
	CPU (kSI2K)	Disk (TB)	CPU (kSI2K)	Disk (TB)
Boston U	191	4	399	138
Indiana U	144	10	144	10
U of Chicago	80	16	484	75
Total Tier 2	415	30	1027	223
Tier 1	135	24	250	100

2.3.3 Networking



✳ Responsible for:

- Specifying both the national and international WAN requirements of US ATLAS
- Communicating requirement to Network suppliers (ESnet, Internet 2, etc.)
- Monitoring the extent to which WAN requirements ...
 - ✳ ... are and will continue to be met for US ATLAS sites

✳ Small base program support effort includes:

- Interacting with ATLAS facility site managers and technical staff
- Participating in HENP networking forums
- Interacting with national & international Grid & networking standards groups
- Adopt/adapt/develop, deploy, & operate WAN monitoring tools

✳ Some progress on critical Network issues at BNL Tier 1

- Currently limited to OC12 WAN connectivity (BNL reviewing paths forward)
 - ✳ ESnet is no longer able to meet bandwidth needs in a timely fashion
 - ✳ Issue is getting to and onto National Lambda Rail
 - ✳ Investigating the formation of local consortium
 - ✳ Comparing near and long term costs of lit service versus dark fiber
 - ✳ Solutions exist but source of significant required funding unclear

Network Contention



- * Significant issue at BNL
 - BNL hosts **RHIC**, an equally bandwidth hungry program
- * For WAN connectivity, “*effectively utilization*” is a concern of equal important with “*high bandwidth*”
 - Contention between
 - * Programs: RHIC / ATLAS at BNL or ATLAS / CMS / etc. at CERN
 - * Activities: Bulk data transfer, Interactive analysis, Conferencing, etc.
- * Project to deploy “*contention management*” technology initiated
 - Multi-protocol Label Switching (MPLS) as a mechanism to achieve Quality of Service (QoS) differentiation on network paths
 - * Partition off and allocate slices of network bandwidth or otherwise prioritize traffic
 - Support from DOE, High-Performance Network Research Program, Thomas Ndousse’s office
 - * 1 FTE+
 - * MPLS capable equipment at BNL, CERN, elsewhere

DC2 Phases & Status



- * Phase 1: Distributed production ($>10^7$) events
 - In two steps:
 - * Pythia based generation
 - * Geant4 simulation & digitization
 - Produced data sent to Tier 1 centers for transfer to CERN
 - Status of Phase 1
 - * Originally scheduled to start April 1st
 - * Officially started on May 5th
 - **ATHENA software not ready, started seriously running test jobs using new production system (and latest ATHENA release 8.0.2)**
 - * Real production started June 24th
 - **Few days after ATHENA release 8.0.5 became available**
- * Phase 2: Reconstruction at CERN (Tier 0) **Delayed (1 Jun => 16 Aug)**
 - Reconstruction output will be distributed to Tier-1 centers for redistribution to Tier 2, etc.
- * Phase 3: Distributed analysis

DC2 Production System



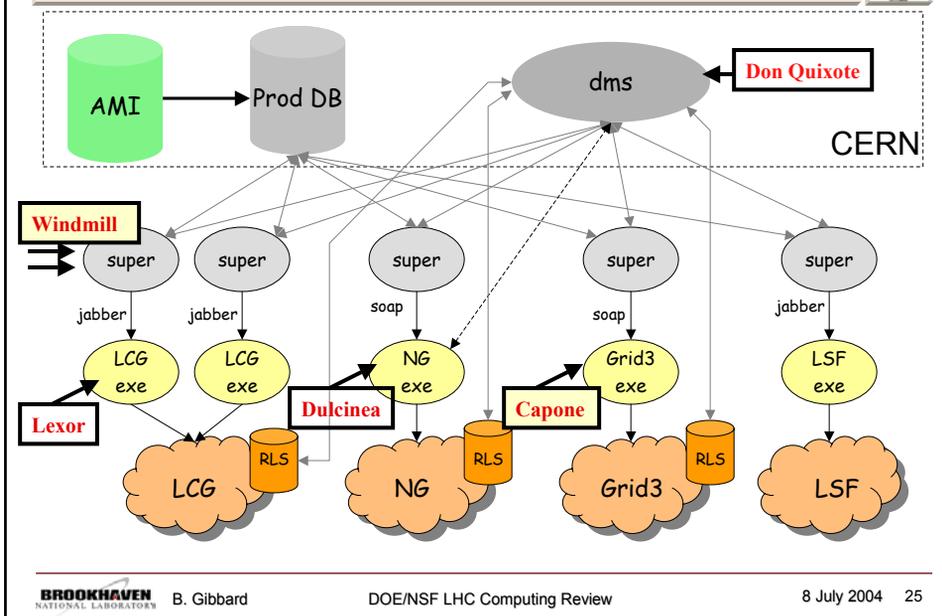
- ✱ **ATLAS Production System Designed to Integrate Use of Three Independently Configured and Operated Grids**
 - ❑ LCG
 - ❑ Grid3
 - ❑ NorduGrid
- ✱ U.S. Production Team making critical ATLAS wide contributions in design, development, deployment and testing of this **multi-Grid** production system
 - ❑ Principal component contribution, **Windmill**, delivered on time
 - ❑ Over 20,000 real jobs already executed using **Windmill**

DC2 Production System (2)



- ✱ DC2 production system consists of 4 components
 - ❑ Production Database - Oracle DB supported by CERN, developed primarily by CERN (L. Gossens) and U.S. (K. De)
 - ❑ **Windmill Supervisor** - used ATLAS wide for production, developed by U.S. production team, allows interoperability of grids (& batch)
 - ❑ Don Quixote - Data Management System, developed by CERN, allows interoperability between RLS/RC systems
 - ❑ Executors - to run jobs requested by supervisor
 - ✱ **Capone - GRID3** executor developed by U.S. GTS team
 - ✱ **Lexor** - LCG executor developed by Italy
 - ✱ **Dulcinea** - NorduGrid executor
 - ✱ **legacy** - LSF/PBS/BQS executor by Munich/Lyon groups
 - ❑ U.S. also developed the xml based messaging system between supervisor and executors

ATLAS Multi-Grid Production System



Grid Tools & Services (GTS) Activities



- ✱ While all out effort have been directed toward support for DC2 operation, there have been noteworthy intermediate benefits
 - US ATLAS GTS team has led development of new releases of Grid3+ working within the framework of grid3dev
 - All sites now upgraded to grid3v2.1 (based on VDT 1.1.14)
 - Have deployed tools allowing installation of multiple ATLAS releases at a Grid3 site
 - Instrumental in ATLAS software packaging and deployment kit via Pacman and active testing and debugging activities
- ✱ Distributed Analysis Phase of DC2
 - Will use ATLAS ADA/Dial effort lead by David Adams
 - Will serve as “fall demonstrator” for Grid3+
- ✱ Grid3+ evolution is now seen as moving toward OSG-0 in early 2005

Status of Previously Listed Major Near Term Milestones



WBS

2.3.1.3	Tier 1 Linux CPU upgrade for DC2 complete ATLAS Production on LCG-2	1/30/2004 ✓ 2/29/2004 ✓	
2.3.4	GCE 2.0: DC2 Alpha	2/1/04 ✓	1 Availability of software delayed the start of DC2 Phase 1 Start until 6/24/04
2.3.1.4	Tier 1 disk upgrade for DC2 complete	2/9/04 ✓	
2.3.3	Beta version host network diagnostics deployed	2/27/2004 ✓	
2.3.4	GCE 2.0: DC2 Delivery	3/1/2004 ✓	2 Delay of DC2 Phase 1 Start propagates into a delay for DC2 Phase 2 Start until 8/16/04
2.3.2.2	BU Tier 2 Fabric Upgrade for DC2 complete	3/5/04 ✓	
2.3.5	Deliver working Windmill supervisor	3/15/04 ✓	
2.3.1	Tier 1 Fabric upgrade fully operational for DC2	3/25/04 ✓	
2.3.2	Tier 2 Fabric upgrade operational for DC2	3/25/04 ✓	3 Delay of DC2 Phase 2 Start propagates into a delay for DC2 Analysis Phase Start until 9/15/04
2.3.1.5	Limited Optimization of CERN / BNL Transfer	4/1/2004 ✓	
2.3.5	DC2 GTS version ready for production Start ATLAS DC2 Phase 1 Combined Testbeam Start ATLAS DC2 Phase 2	4/1/2004 ✓ 4/1/04 Delayed 1 5/1/04 ✓ 6/1/04 Delayed 2	4 Uncertainty regarding funding availability delayed the call for proposals so this process is now expected to complete 10/31/04
2.3.1	Tier 2 Fabric upgrade operational for DC2 analysis Start ATLAS DC2 Analysis Phase	7/15/2004 ✓ 7/15/2004 Delayed 3	
2.3.2	Permanent Tier 2 Sites A, B, C selection complete	8/1/04 Delayed 4	5 Need to delay this milestone has not been established
2.3.5	DC2 goals achieved ATLAS Computing Model Paper Complete	10/1/2004 11/30/04 Note 5	

External Milestones



B. Gibbard

DOE/NSF LHC Computing Review

8 July 2004 27

Summary



- ✳ Ramp-up of Tier 1 technical staff (3 =>7) significantly strengthened ...
 - Authentication, Authorization, & Accounting
 - Networking
 - Data Storage & Movement
 - DC2 Operational Support

... temporary setback in management with loss of Rich Baker
- ✳ Facility fabrics at all levels adequate for DC2
 - Tier 1, prototype Tier 2's, Tier 3's (some soon to be Tier 2's)
- ✳ GTS & Production teams heavily involved in ATLAS aspects of DC2 as well as bring US ATLAS resources to bear on DC2 via Grid3
 - Major role in design & implementation of Multi-Grid architecture
 - Grid3 based resources being used to shake down DC2 operations



B. Gibbard

DOE/NSF LHC Computing Review

8 July 2004 28