U.S. ATLAS Computing Facilities
Requirements, Capabilities and Schedule

Michael Ernst

Brookhaven National Laboratory

U.S. ATLAS Tier-2 & Tier-3 Meeting

University of Michigan, Ann Arbor

27 – 28 May 2008

HEP office of high energy physics

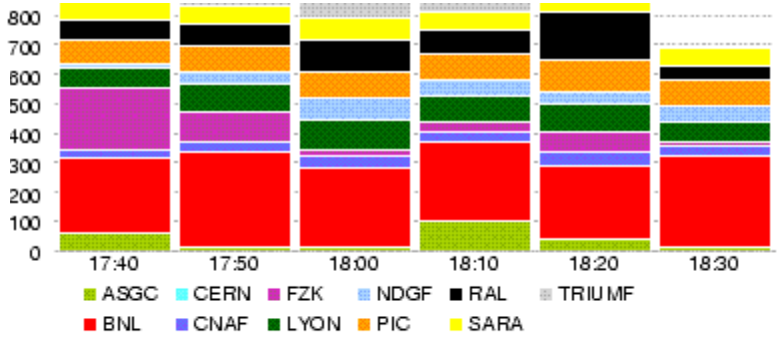
BROOKHAVEN NATIONAL LABORATORY

Planning

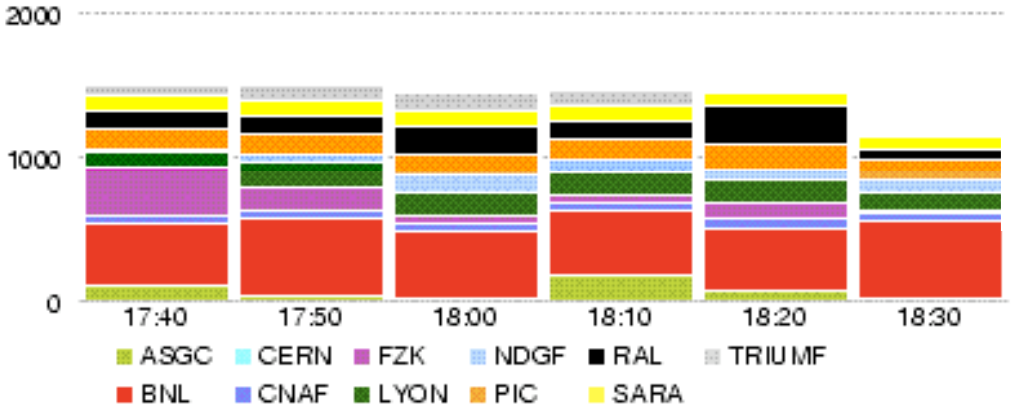

December
February
March
April
May
CC Readiness Challenge
Full DressRehersal
Cosmic Ray tests
Functional Tests
SRMv2 testing
Time for fixes and updates
S&C Week
Jamboree 2 (?)
WLCG Workshop
1
2
0
1
bis
2
bis
6
3
4
5
6
1
1
2
3

CERN to all ATLAS Tier-1's

300 GB in 10 minutes



MB/s

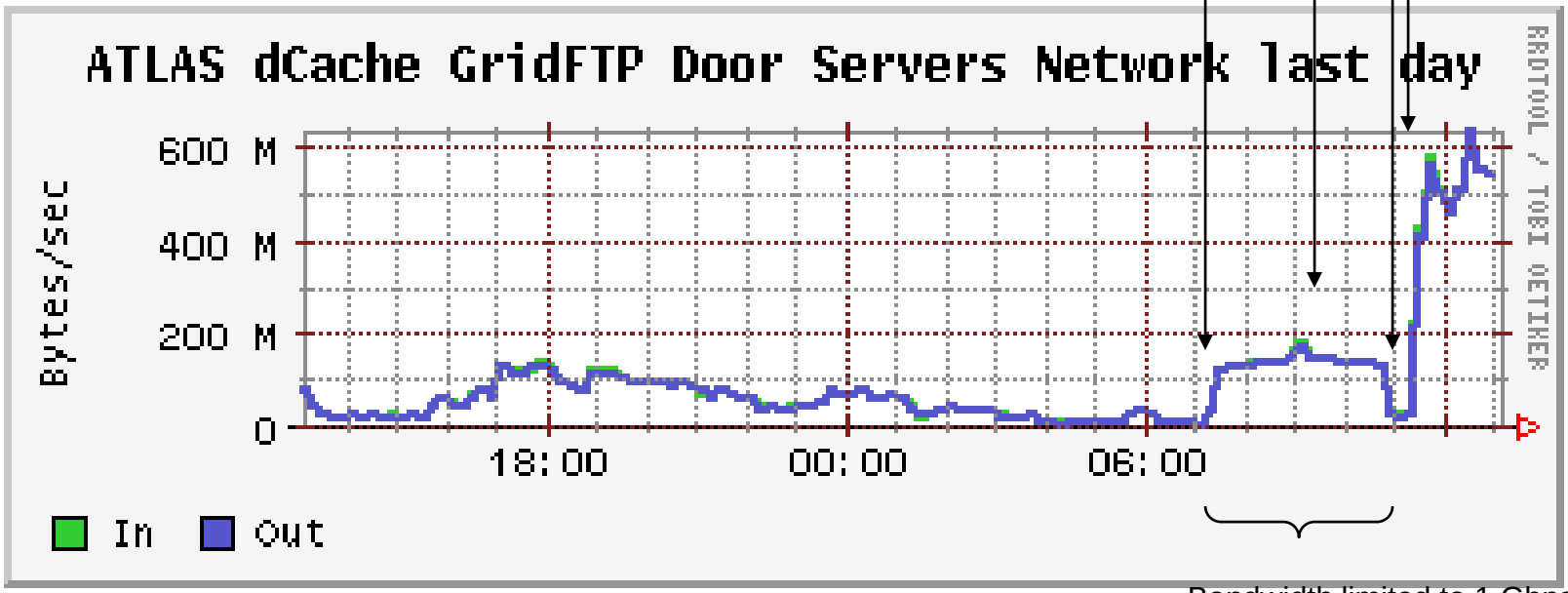


Start of ATLAS CCRC Transfers to BNL

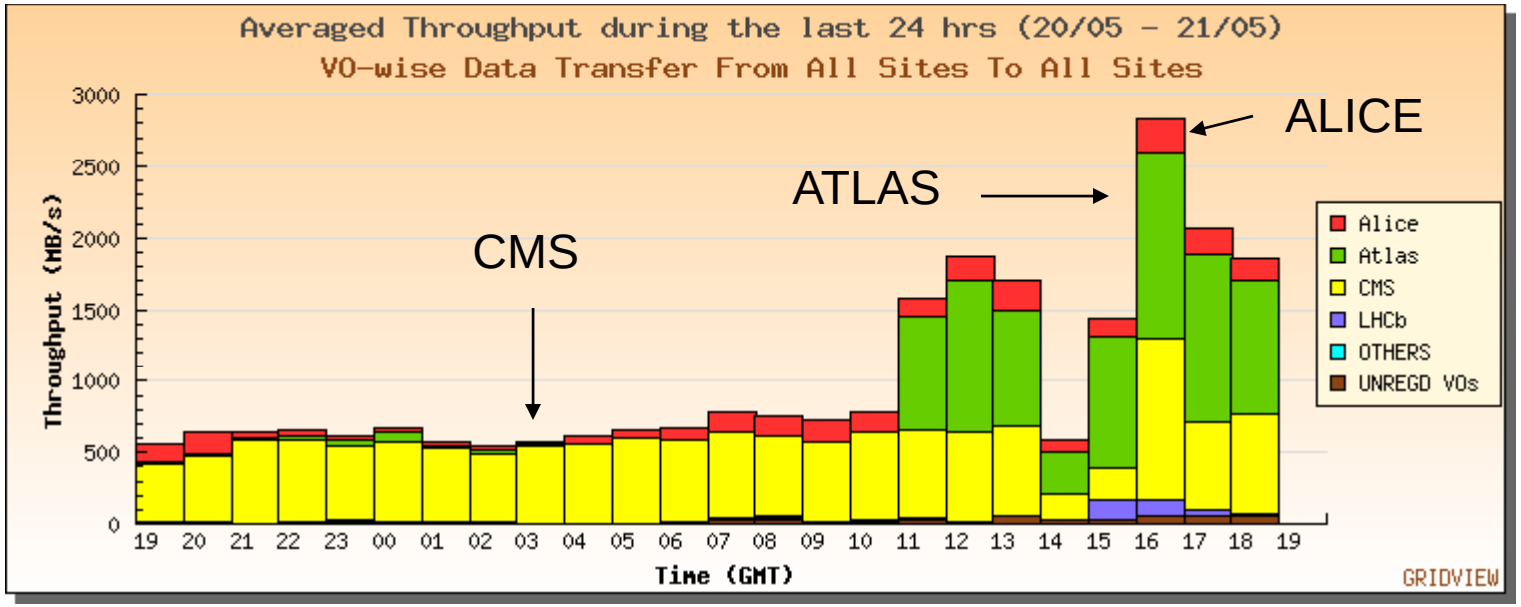
Castor outage at CERN

Problem solved by ESnet

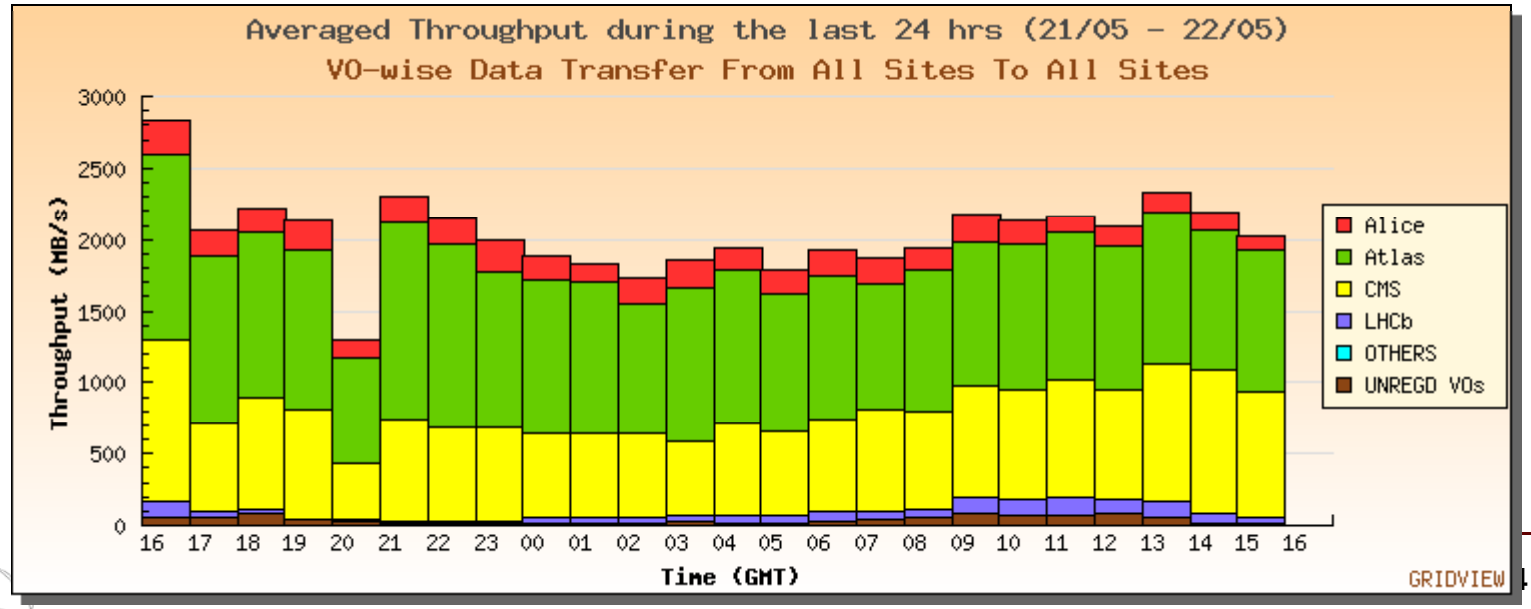
Problem reported to ESnet

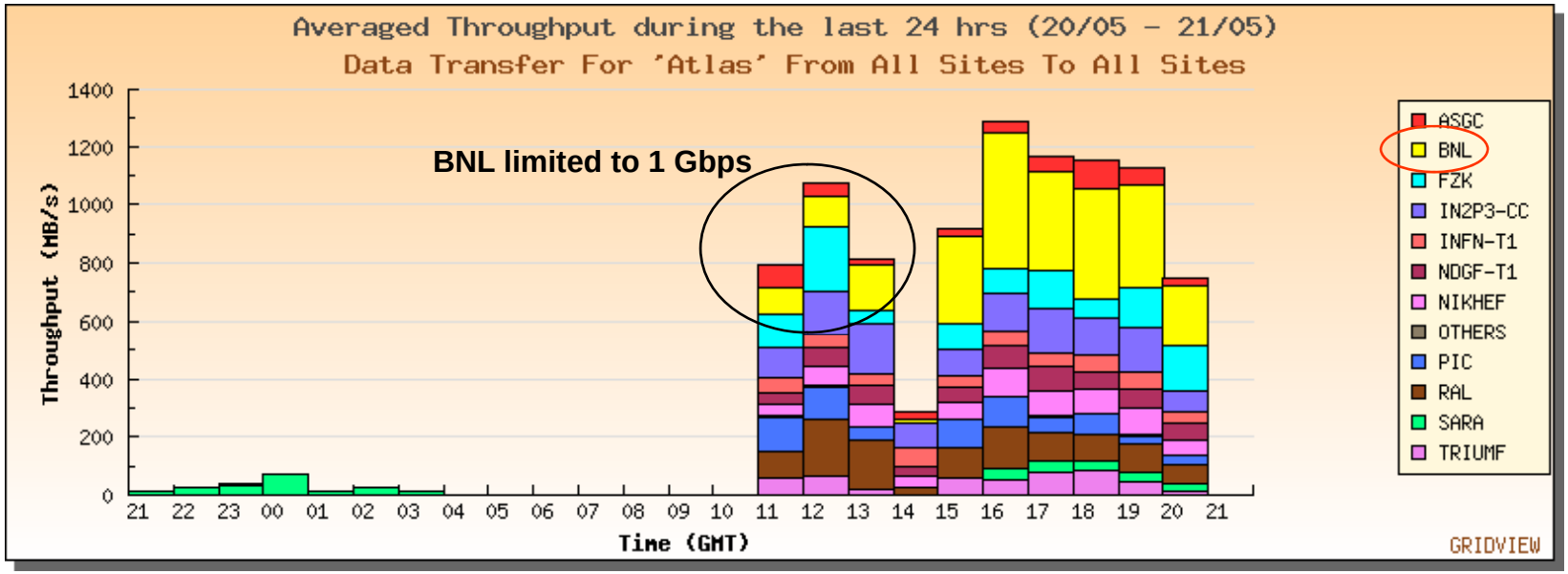


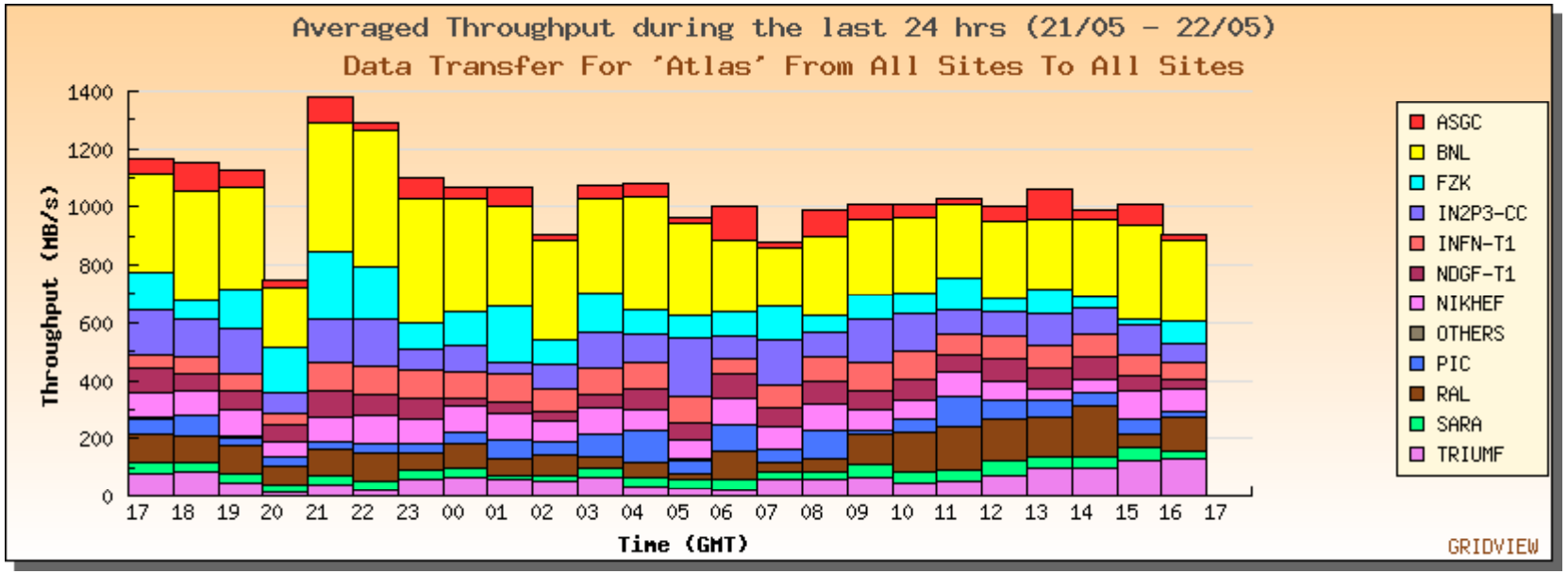
Bandwidth limited to 1 Gbps
Due to configuration problem at ESnet



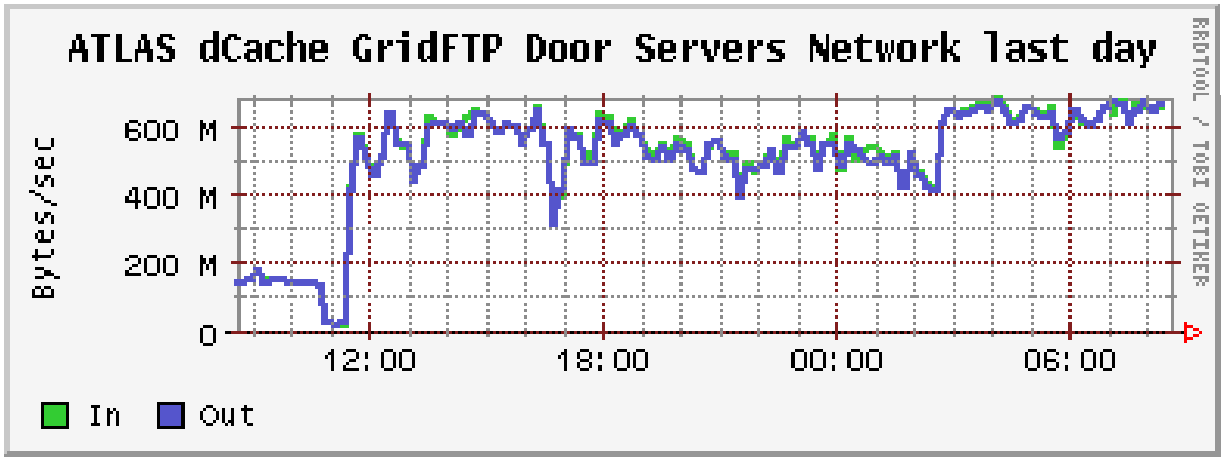
After 24 hours



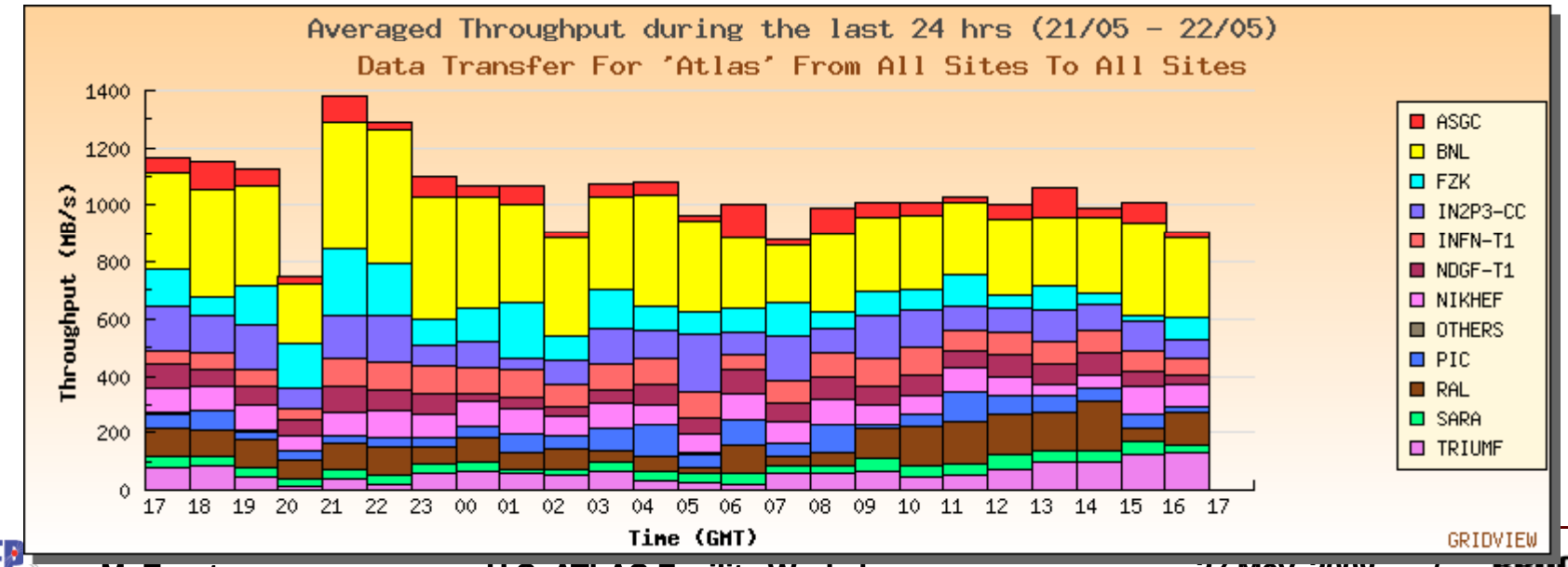

Averaged Throughput during the last 24 hrs (20/05 - 21/05)
Data Transfer For 'Atlas' From All Sites To All Sites
BNL limited to 1 Gbps
Throughput (MB/s)
Time (GMT)
ASGC
BNL
FZK
IN2P3-CC
INFN-T1
NDGF-T1
NIKHEF
OTHERS
PIC
RAL
SARA
TRIUMF
GRIDVIEW


Averaged Throughput during the last 24 hrs (21/05 - 22/05)
Data Transfer For 'Atlas' From All Sites To All Sites
Throughput (MB/s)
Time (GMT)
ASGC
BNL
FZK
IN2P3-CC
INFN-T1
NDGF-T1
NIKHEF
OTHERS
PIC
RAL
SARA
TRIUMF
GRIDVIEW

After ~24 Hours



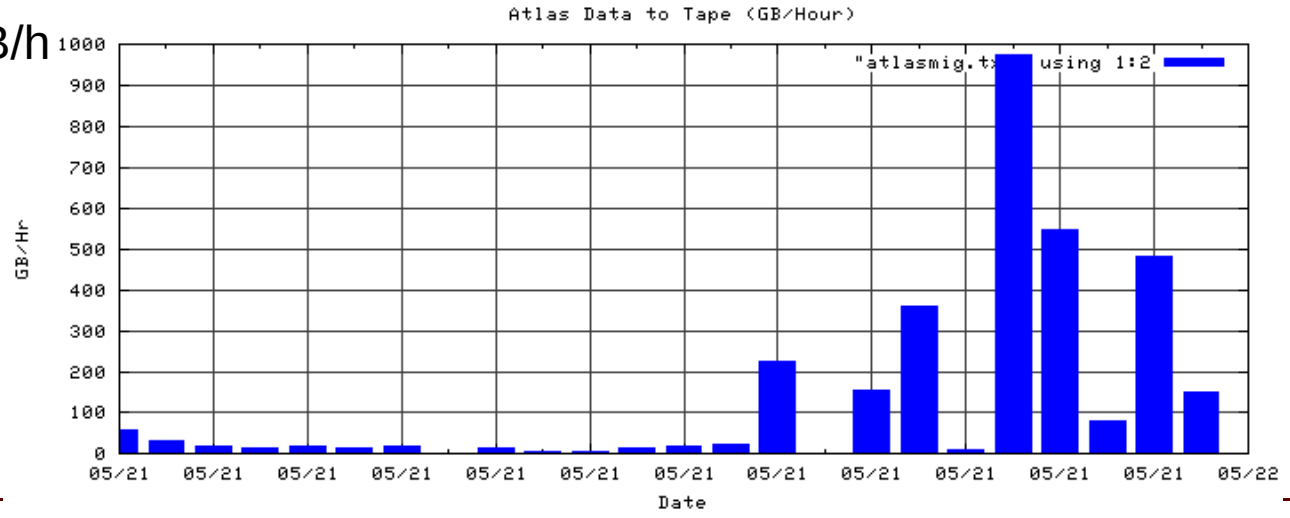
Aggregate GridFTP I/O
- CERN T0 => BNL Tier-1
- BNL-T1 => US ATLAS T2



ATLAS Distributed Data Management Dashboard

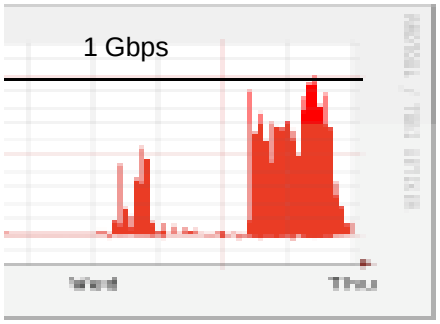
	Transfers			Registrations		Errors	
Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration
ASGC	100%	73 MB/s	81	15	81	0	0
BNL	100%	420 MB/s	542	28	542	1	0
	Click on the site name to go to the site page, '+' to see statistics for this site per source NEW						
AGLT2	0%	0 MB/s	0	0	0	0	0
AGLT2_DATADISK	0%	0 MB/s	0	0	0	0	0
AGLT2_MCDISK	0%	0 MB/s	0	0	0	0	0
AGLT2_SRM	0%	0 MB/s	0	0	0	0	0
AGLT2_UM	0%	0 MB/s	0	0	0	0	0
AGLT2_UMFS02	0%	0 MB/s	0	0	0	0	0
BNL-OSG2_DATADISK	100%	298 MB/s	423	23	423	0	0
BNL-OSG2_DATATAPE	99%	122 MB/s	119	5	119	1	0

1 TB/h

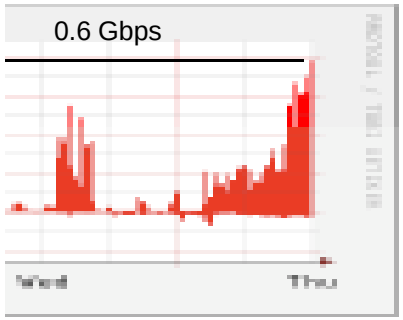


Data Migration to Tape

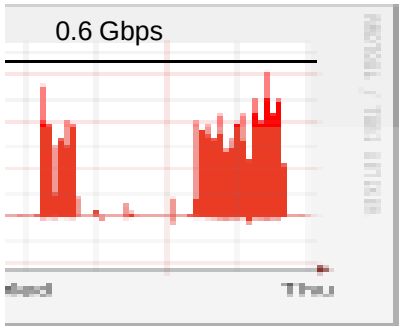
Export to the Tier-2's



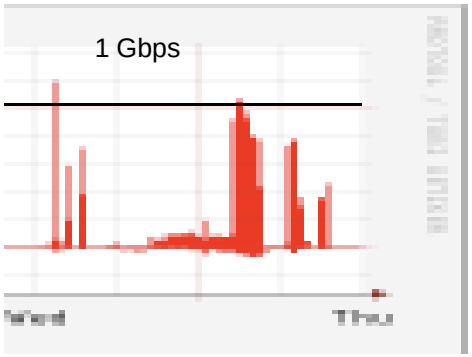
AGLT2



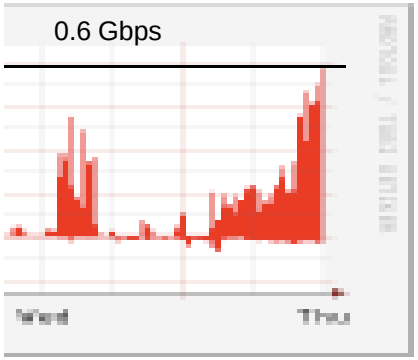
MWT2



SWT2



WT2



NET2

Data Replication Progress

Tier-1's

Datasets	Total Files in datasets	Total CpFiles on sites	Last Subscription	FC Checked	Last Transfer
8953	81466	80820	24 May 09:39	26 May 08:57	24 May 18:16

Tier1	Datasets	Total Files in datasets	Total CpFiles in datasets	Completed	Transfer	Subscribed
BNL	1476	25415	25415	1476	0	0
FZK	893	7694	7694	893	0	0
IN2P3	919	8997	8997	919	0	0
INFN	836	7058	7062	832	0	0
NDGF	774	4096	3346	749	0	25
PIC	776	4156	4156	776	0	0
RAL	811	6337	6367	807	0	0
SARA	912	9308	9208	898	13	1
TAIWAN	758	3559	3559	758	0	0
TRIUMF	798	4846	5016	750	0	0

Tier-2's

BNL

Tier2	Datasets	Total Files in datasets	Total CpFiles in datasets	Completed	Transfer	Subscribed
AGLT2_DATADISK	694	1059	1059	694	0	0
MWT2_DATADISK	694	1059	1059	694	0	0
NET2_DATADISK	694	1059	1059	694	0	0
SLACXRD_DATADISK	694	1059	1059	694	0	0
SWT2_CPB_DATADISK	694	1059	1059	694	0	0

Machine Schedule

- Two months of ramp up
  - Beam commissioning
  - Presumably we will record data during this period.
- Pilot run of 3 weeks
- Conservatively, we prepare for 3 months of data taking in 2008
  - Assume 60 days after folding in inefficiencies

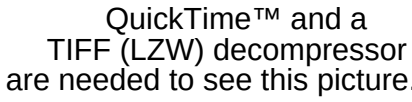
The Overall Plan

1. Tier-0 to Tier-1and Tier-1 <=> Tier-2 data distribution
2. Tier-1 data re-processing
3. Tier-2 Simulation Production
4. Tier-1/Tier-2 Physics Group Analysis
5. Tier-1/Tier-2/Tier-3 End-User Analysis

- ✓ Synchronously with Data from Detector Commissioning
- ✓ Fully rely on SRM V2 at the Tier-1 and all Tier-2's
- ✓ Move to tests at real scale (avoid data deletion)
- ✓ Move to realistic operations scenarios (Communication, shifts)

Schedule in June

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.


QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

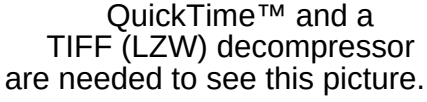
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Group Space

- We had many ATLASGRP <group> storage area's but (almost) none were used
- It seems at this stage that one for each VOMS group is over the top
- Much overhead to create too many small storage classes
- For now a catch-all ATLASGRP
- We may revert back later when we better understand the usage

Tier-2 Simulation Production

- Simulation of physics and background for FDR-2
- Have produced ~30 M events
- Simulation => HITS (4MB/ev), Digitization => RDO (2MB/ev)
- Reconstruction => ESD (1.1MB/ev), AOD (0.2MB/ev)
- Simulation was done at theTier-2's
- HITS uploaded to T1 and kept on disk
- At Tier-1: digitization => RDOs sent to BNL for mixing
- At Tier-1: Reconstruction => ESD, AOD
  - ESD, AOD archived to tape at Tier-1
  - ESD copied to one or two otherTier-1's
  - AOD copied to each other Tier-1


QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Tier-0/1/2 Group Analysis

- Done at Tier-0 & Tier-1 &Tier-2 - not at Tier-3's
- Production of primary Derived Physics Datasets (DPD's)
- DPD's are 10% of AOD's in size - but there are 10 X more
- Primary DPD's are produced form ESD and AOD at the Tier-1s
- Secondary DPD's are produced at Tier-1 and Tier-2's
- Also other file types may be produced (ntup's,hist's)
- Jobs always run by managers, data always run to/from disk
- Writable for group managers only, readable by all ATLAS

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

End-User Analysis

- Done at Tier-0 &Tier-1&Tier-2 &Tier-3's
- Users can run (CPU) anywhere where there are ATLAS resources
- But can only write where she/he has write permission (e.g. home institute)
- Each site can decide how to implement this (T1D0,T0D1)
- Data must be registered in the catalog
- Non-registered data is really Tier-3 or laptop
- The whole issue of End User Data Management is not fully understood

Summary Table for a typical Tier-2

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Data in 2008

Assumptions on Event sizes and Running Times and Trigger Rates		
RAWEvent	1.60	MB
FractionRawOnDisk	0.30	
ESDEvent	1.00	MB
ESDNmbCopies	3.00	
FractionEsdOnDisk	0.10	
AODEvent	0.20	MB
SecondsPerDay	36,000.0	seconds
DaysPerMonth	20.0	
TriggerRate	200.0	Hz

RAWRate	11.5	TB/day
ESDRate	7.2	TB/day
AODRate	1.4	TB/day

Disk Requirements for 2008 Data

Detector Data for 3 months running			3							
	RAW	to TAPE	RAW	to DISK	ESD	to DISK	AOD	to DISK	SUM	to DISK
BNL	173	TB	74	TB	463	TB	123	TB	660	TB
IN2P3	104	TB	44	TB	278	TB	123	TB	446	TB
SARA	104	TB	44	TB	278	TB	123	TB	446	TB
RAL	69	TB	30	TB	185	TB	123	TB	338	TB
FZK	69	TB	30	TB	185	TB	123	TB	338	TB
CNAF	35	TB	15	TB	93	TB	123	TB	231	TB
ASGC	35	TB	15	TB	93	TB	123	TB	231	TB
PIC	35	TB	15	TB	93	TB	123	TB	231	TB
NDGF	35	TB	15	TB	93	TB	123	TB	231	TB
Triumf	35	TB	15	TB	93	TB	123	TB	231	TB
Sum	691	TB	296	TB	1851	TB	1234	TB	3382	TB

Kors at the last TOB (05/21)

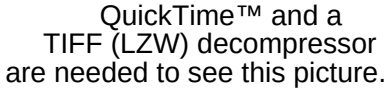
- there are uncertainties in what follows
  - numbers are within a factor of 2
  - based on current practice and very simple spreadsheet
- cannot wait to put something in place
  - T0, T1 and T2¬s ask for guidelines
  - Physics Groups want to know about their MC budget
  - We are getting more and more users
- not quite clear what our resources will be
  - Jim has assembled data for a reality check
  - it is constantly changing
  - we have used quite a bit already and it seems difficult to delete
  - there is quite a bit of hidden resource usage
- current T1 situation is discouraging
  - permanent shortage in CCRC
  - Several Tier-1's far behind (one is just installing the 2007 pledges)
- Situation in T2's still not very well known
  - Haven't done much with T2's yet other than MC production
  - … and sending them AOD's
- Talk mainly about storage, computing is much less an issue

Problems and Questions

- Is the default pool the right solution for end-users?
- From the default pool a huge number of small files are written to tape
- Need to try to clean up atlprod and atldata
- Will we delete cosmics data when the space is needed?
- Can we delete data also from tape? already using 200%
- Do we need to export more calibration streams?
- Do we need more RAW data on disk or do we make the express line available @CERN to all users?
- Do we re-distribute ESD after re-processing or only AOD?
- Do we re-distribute ESD after re-reconstruction of MC?

At the Tier-1

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.


QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Storage Classes at the Tier-1

1. DATATAPE
   - cache before RAW data is written to tape
   - needs to be a buffer in case of mal functioning tape system
   - could also be used for making RAW available
2. DATADISK
   - must contain the ESD (20%) and AOD (100%) from initial processing in the T0
   - Moreover this pool will hold version 1 and version 2 AOD's from reprocessing from other Tier-1's
3. DATADISKTAPE
   - only contains the ESD and AOD from re-processing of the RAW at this T1, so typically 10% of the full sample
   - data needs to be archived, stay on disk and be exported to all other T1's and the T0
4. MCTAPE
   - cache before HITS imported from T2's are written to tape
   - need to be big enough to have HITS on disk to be digitized and reconstructed
5. MCDISK
   - contains AOD (100%) from reconstructed MC data at other T1's
   - moreover must hold subsequent versions of digitization and reconstruction
6. MCDISKTAPE
   - only contains ESD and AOD from reconstruction at this T1, so typically 10% of the full sample
   - data needs to be archived, stay on disk and be exported to all other T1's and the T0
7. Export Pool
   - needs to be a buffer for exports, should be 2 days in case of mal functioning other T1's
8. Stage Pool
   - to stage data back from tape
9. GROUP
   - for physics group analysis, contains DPD's
10. USER
    - still needs better definition (need use-cases)

Data Pool Sizes at the U.S. ATLAS Tier-1

1. DATATAPEDATATAPE
   - cache before RAW data is written to tape => 80 TB would maintain 2 weeks of RAW
2. DATADISK
   - must contain the ESD and AOD (100%) from initial processing in the T0 => ~600 TB
   - Moreover must hold version 1 and 2 of AOD from reprocessing from other T1's => 120 TB per version
3. DATADISKTAPE
   - only contains the ESD and AOD from re-processing at this T1 => 67 TB per version
4. Export Pool
   - for data => 30 TB
5. Stage Pool
   - for data => 30 TB
6. GROUP
   - DPD's from data (same size as AOD) => 120 TB
7. USER
   - not well defined yet => 50 TB for the moment (M.E.: far too small)

Problems and Questions

- How do we adapt to changing circumstances?
  - More RAW requested than originally foreseen
  - Fewer resources available at many Tier-1's than originally pledged
- End-user behavior largely unknown
  - Need use cases to design storage solution
- How do we protect our data?
  - Need a mechanism to avoid users reading from tape
  - … and users to write into managed space
- How will we manage disk-only and tape-only space?

MC Pool Sizes needed at Tier-1

Assumption: MC is 25% of real data: 600M real events and so 150M MC events

1. MCTAPE
   - 35M HITS of 4MB/event => 140TB to have them all on disk in 2008 (and tape)
2. MCDISK
   - AOD from reconstructed MC from other T1's => 135M AOD (0.2 MB) 27 TB per version of the reconstruction
3. MCDISKTAPE
   - 15M RDO (2 MB)+ ESD (1 MB) + AOD (0.2 MB) => ~120 TB per version of the reconstruction
   - Assumption: We still need to store the RDO's ?
4. Export Pool
   - for MC => 25 TB
5. Stage Pool
   - for MC => 25 TB
6. GROUP
   - DPD's from MC (same size as AOD) => 30 TB
7. USER
   - not well defined yet => 50 TB for the moment

Service Reliability = D.I.D.O.

- The Goals - MoU Targets & Experiment Critical Services Lists
- The Techniques: Developers' golden rules
- The Techniques: The Integration Program & Deployment strategies
- **Measuring What We Deliver & Problem Resolution Targets**
- Middleware & experiment-ware surveys - how do we rack up?
- Testing Experiment Services - SAM/RSV tests et al (incl. Network)
- ATLAS Grid Services

Design, Implementation, Deployment, Operation

(Re-)implementing Services

- Cannot take existing non-HA service, add SLA and make it HA
- Several techniques available – need to be further studied
- Need to discuss options and impact early on with users
  - e.g. Replication, RAC all offer solutions in area of HA
  - Most appropriate depends on needs of service
  - How much downtime can be tolerated?
  - Pros & cons of multiple cheap boxes versus more expensive (and more complex) hardware
- Learned a lot by implementing ~24 x 7 services at Tier-1

ATLAS Critical Services (PDF)

Tier	Service	Criticality	Consequences of service interuption
0	Oracle database RAC (online, ATONR)	Very high	Possible loss of DCS, Run Control, and Luminosity Block data while running. Run start needs configuration data from the online database. Buffering possibilities being investigated.
0	DDM central services	Very high	No access to data catalogues for production or analysis. All activities stops.
0	Data transfer from Point1 to Castor	High	Short (<1 day): events buffered in SFO disks, backlog transferred as connection is resumed. Long (>1 day): loss of data.
...			
0-1	3D streaming	Moderate	No export of database data. Backlog can be transferred as [soon as] connections are resumed.
... more ...			

CHEP 2007

Planning ahead …

- Milestones and Items on the following slides are guided by Production and Analysis needs
- U.S. ATLAS Computing Integration Program will translate them into the technical steps sites have to perform
- Will only mention a few pressing items

Analysis

- Interactive Analysis
  - BNL PROOF farm available to all US ATLAS users for testing
  - BNL PROOF farm in production mode
    - Target: Complete 31 March
    - We will hear more from Ofer and Sergey
  - Tier 2 PROOF farms available
    - Target: Complete 30 June
    - Does this fit our model (User account management etc.)?
  - Support Tier-3 activities as part of the Computing Integration Meeting
    - Immediately, ongoing
    - Status: Still no regular participation from Tier-3 institutions

Storage Services

- At Tier-1
  - Evaluate Pinning with SRM v2.2
    - Target: Complete by 21 December
    - Status: Delayed
  - Propose data placement plan for data at Tier 1, including pinning, disk only partitions etc
    - Target: Complete by 31 December
    - Status: In progress
  - Develop and deploy software necessary to mange pinned files
    - Target: Complete 15 January
    - Status: In Progress
  - Disk space reconfiguration according to computing model
    - Target: Complete 31 January
    - Status: In progress
  - Develop and deploy disk-only dCache space management tools
    - Target: Complete 21 December
    - Status: Delayed
  - User space management at Tier 1, including user management, cache cleanup
    - Target: Proposal Complete 31 December
    - Target: Deployment: 31 January
    - Status: Delayed
  - LFC
    - Test system deployed: Complete 31 December
    - Test system production ready: 31 January
    - Migration to LFC completed for US, assuming successful tests: 28 February

U.S ATLAS Data

- Data Management
  - Deploy storage quota system US ATLAS wide
    - Target: Complete by 28 February
    - Status: Delayed
  - DQ2 data deletion fully operational
    - Complete by 15 December
    - Status: In Progress (ATLAS tools exist for LFC based sites)
  - Complete DQ2 lost file tagging for US
    - Complete 15 January
    - Status: In progress
- What is data flow model in Pathena?
- What if researcher produces data at Tier3?
- How is the decision to archive made?
- Are Tier2's expected to maintain precious data indefinitely?
- User Data Lifetime?
- Consistency Checks?

Summary

The U.S. ATLAS Facilities have made substantial progress toward an operational integrated computing facility for the start of the experiment

- The facilities, the Tier-1 and the Tier-2's, have performed well in ATLAS computer system commissioning and specific exercises
  - An Integration Program is in place to ensure readiness in view of the steep ramp-up
  - The Tier-2's are ready to provide resources for Analysis (still need the AODs)
  - Excellent contribution of U.S ATLAS Tier-2 Sites to high volume production in 2007
- The BNL Tier-1 serves as the hub and principal center of the US community, with scale-up for data taking underway
  - A sizable Central Analysis Facility is needed in addition to the ramp-up of Tier-3's
- Overall, progressing well towards full readiness for LHC data analysis

Questions?

If not, on with the workshop…