



# U.S. ATLAS Facility – Status and Plans

Michael Ernst

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

28 November 2007

# Overview



- Funding for U.S. ATLAS Computing Facilities
- Computing in U.S. ATLAS
- Status and Plans of the U.S. ATLAS Tier-1 Center
- Tier-2 Resource ramp-up
- Transition to Operation
- Summary

# US ATLAS Institutions

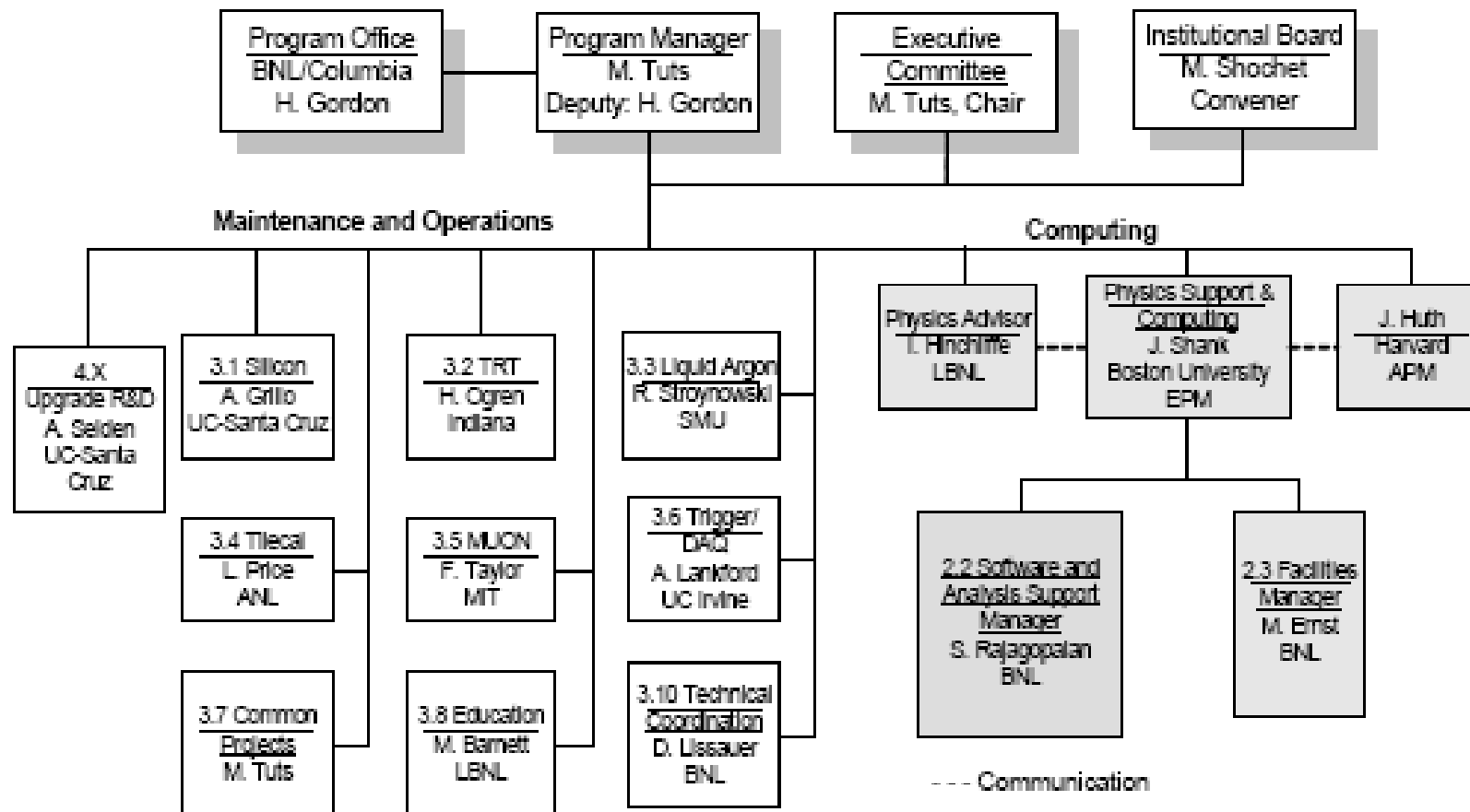


- 42 institutes (in **red** indicates new since last April) (Black = Tier 2 center)
  - ❑ Albany, ANL, Arizona, UT Arlington, Berkeley LBL and UC, Boston, Brandeis, BNL, Chicago, Columbia, UT Dallas\*, Duke, Hampton, Harvard, Indiana, U Iowa#, Iowa State, UC Irvine, Louisiana Tech\*, Massachusetts, MIT, Michigan, MSU, New Mexico, **NIU^**, NYU, Ohio State, Oklahoma, Oklahoma State, Oregon, Pennsylvania, Pittsburgh, UC Santa Cruz, SLAC, SMU, South Carolina\*, SUNY Stony Brook, Tufts, Illinois Urbana, Washington, Wisconsin, Yale
  - ❑ Corresponding to 38 voting institutions
    - \* = affiliated with BNL
    - # = affiliated with SLAC
    - ^= affiliated with ANL
- Currently in discussions with...
  - ❑ Fresno State (existing collaborator starting group there)
  - ❑ RPI (new group, coming from CLEO)
  - ❑ Overtures received from Temple University
- As of Sept 30, 2007 (used for Oct RRB)
  - ❑ 38/166 voting institutions (23%)
  - ❑ 332/1624 “current M&O authors = PhDs” (20%) – for cat A/B
  - ❑ 420/2095 including students (20%)
  - ❑ 398/1977.25 Operations tasks share (students count .75) (20%)

# US ATLAS Organization Chart



## U.S. ATLAS Research Program Organization as of February 1, 2007



# Computing Funding



# Funding Source/Management



- The Research Program (RP) is funded by the Department of Energy (DOE) and the National Science Foundation (NSF)
  - ❑ All physics groups/institutes are funded additionally for work on ATLAS by the same agencies (core funding)
- The RP funding covers all costs of the Tier 1 and 2s among other things
- Tier 3 centers funded by other sources
- The RP is managed by a Joint Oversight Group (JOG) consisting of members of NSF and DOE and BNL/FNAL Lab management.
- A U.S. ATLAS management team organizes the US participation in ATLAS
  - ❑ Keeps costs under control

# Funding Targets



- Original targets based on bottom up estimates, out years evaluated yearly
- FY08 Management reserve requests far exceed funding - prioritize

(AY M\$)	FY07	FY08	FY09	FY10	FY11
Physics & Computing	15.0	15.3	17.4	19.0	17.9
M&O	19.3	10.0	10.3	10.4	10.5
Upgrade R&D	3.1	3.4	3.2	3.2	3.2
Management Reserve	1.5	5.0	3.7	4.0	5.9
DOE guidance	22.6	24.6	25.5	27.5	28.5
NSF guidance	9.0	9.0	9.0	9.0	9.0
Unobligated/Carryover	7.3	?			
<b>Total</b>	<b>38.9</b>	<b>33.6</b>	<b>34.5</b>	<b>36.5</b>	<b>37.5</b>

# Overall Computing Needs



## US ATLAS Computing Needs Profile (AY k\$)

	FY07	FY08	FY09	FY10	FY11
Research program target	15112	15260	17406	19006	17940
<b>Current Computing Total</b>	<b>15112</b>	<b>15260</b>	<b>17406</b>	<b>19006</b>	<b>17940</b>
Difference between Target-Total	(0)	0	0	(0)	(0)
<b>sw target</b>	<b>5268</b>	<b>5179</b>	<b>5641</b>	<b>5835</b>	<b>6067</b>
sw mr	0	624	483	501	523
<b>Total sw</b>	<b>5268</b>	<b>5803</b>	<b>6124</b>	<b>6336</b>	<b>6590</b>
<b>T1 target</b>	<b>6295</b>	<b>6451</b>	<b>8416</b>	<b>9803</b>	<b>8485</b>
T1 mr	0	1762	1397	1831	1251
<b>Total T1</b>	<b>6295</b>	<b>8213</b>	<b>9813</b>	<b>11634</b>	<b>9736</b>
<b>DC/prod.</b>	<b>549</b>	<b>630</b>	<b>649</b>	<b>668</b>	<b>688</b>
Operations Coordinator (MR)		250	260	270	281
<b>T2</b>	<b>3000</b>	<b>3000</b>	<b>2700</b>	<b>2700</b>	<b>2700</b>
T2 mr		0	300	300	300
<b>Total T2</b>	<b>3000</b>	<b>3000</b>	<b>3000</b>	<b>3000</b>	<b>3000</b>
<b>Total Facilities (with MR)</b>	<b>9844</b>	<b>12093</b>	<b>13722</b>	<b>15573</b>	<b>13706</b>
<b>Total Fac. (no MR allocated)</b>	<b>9844</b>	<b>10081</b>	<b>11765</b>	<b>13171</b>	<b>11873</b>
Total with no MR allocated	15112	15260	17406	19006	17940
<b>Total with MR allocated</b>	<b>15112</b>	<b>17896</b>	<b>19846</b>	<b>21909</b>	<b>20296</b>



# Computing in U.S. ATLAS



- **Computing resources used for production and analysis**
  - ❑ BNL Tier 1
  - ❑ Five Tier 2's
  - ❑ Many Tier 3's
  - ❑ All sites are organized hierarchically
  
- **Personnel involved in production**
  - ❑ Tier 1 site support (clusters, storage, networking)
  - ❑ Service support (servers, alarms, installation, integration)
  - ❑ Tier 2 site support (also helping Tier 3's)
  - ❑ Shift team
  
- **Software systems for production**
  - ❑ Panda (including pathena) and DQ2 (including FTS)

# Personnel



## ➤ BNL T1

- ❑ 1-3 FTE for each area of work
- ❑ Storage management most critical

## ➤ Tier 2's

- ❑ Typically 1-2 FTE
- ❑ Also help with integration, tool development etc

## ➤ Tier 3's

- ❑ No dedicated ATLAS personnel – some need/request help from Tier 2's

## ➤ Production Shift team

- ❑ 3 FTE (5 people): 1 FTE BNL, 2 FTE UTA

# Service Model



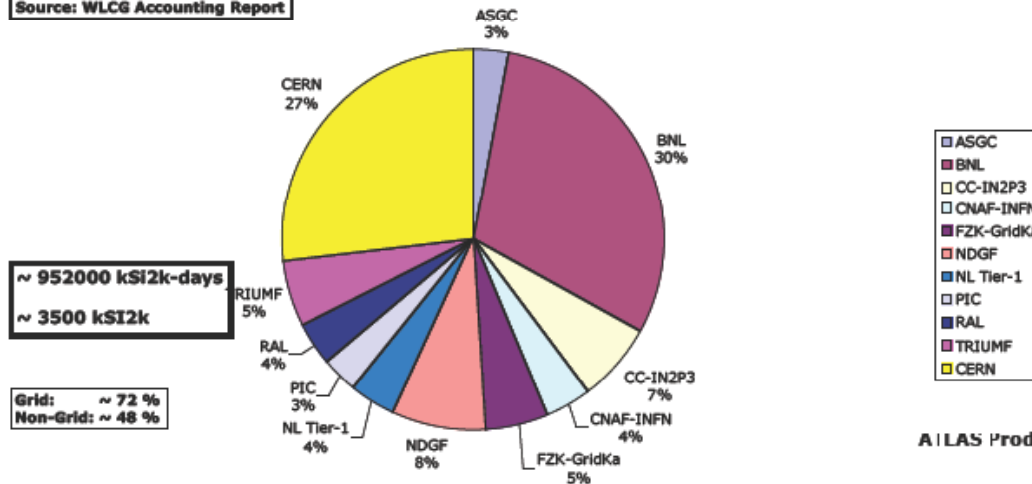
- U.S. ATLAS Computing follows a service model with dedicated personnel for common tasks + site support
- Panda production (including pathena) is monitored at Tier 1 and Tier 2 resources by Production Team, working with local site support teams
- DDM Operations team (coordinated by Alexei Klimentov) manages data distribution, data access issues
- Production team operates shifts to provide QoS
  - ❑ U.S. shift team is also supporting CA, UK, FR Panda production (new ATLAS-wide plan being developed)
- Hypernews, RT user support system, and Savannah bug reporting systems are available to users
- **Need to fill the Position of the U.S. ATLAS Operations Coordinator**

# US Contribution to Worldwide Production



ATLAS CPU at Tier-1s & Tier-0 in 2007 (Jan-Sep)

Source: WLCG Accounting Report

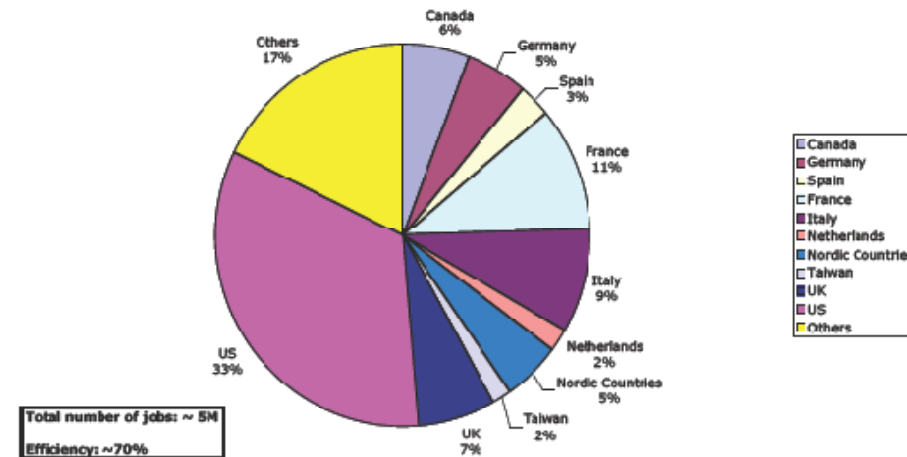


~ 952000 kSi2k-days  
~ 3500 kSi2k

Grid: ~ 72 %  
Non-Grid: ~ 48 %

ATLAS Production (per country) - Number of Jobs (Jan-Sep 2007)

Source: ATLAS Production System database



Total number of jobs: ~ 5M  
Efficiency: ~ 70%

D. Barberis at WLCG MB 11/6/07  
Details in Kaushik's Presentation

# Data Location



- Tier 1 – main repository of data (MC & Primary)
  - ❑ Store complete set of ESD, AOD, AANtuple & TAG's on disk
  - ❑ Fraction of RAW and all U.S. generated RDO data
- Tier 2 – repository of analysis data
  - ❑ Store complete set of AOD, AANtuple & TAG's on disk
  - ❑ Complete set of ESD data divided among 5 Tier 2's
- Data distribution to Tier 1 & Tier 2's is managed
- Tier 3 – unmanaged data matching local interest
  - ❑ Data through locally initiated subscriptions
  - ❑ Mostly AANtuple's, some AOD's
  - ❑ Will Tier-3's be associated with Tier-2 sites?
  - ❑ Tier 3 model is still not fully developed – evolving

# Resource Allocation



- All U.S. Tier 1/2 sites provide dedicated resources
  - ❑ For reliable storage of distributed data (previous slide)
  - ❑ CPU's for managed production (ATLAS-wide groups)
  - ❑ CPU's for regional/local production of large samples through Panda
  - ❑ CPU's for user analysis through pathena
  - ❑ CPU's for interactive Athena for testing/software development (unlikely to be available at all Tier 2's – will be available at BNL)
  - ❑ Root analysis of AANtuple's is expected to be done on personal workstations, and Tier 3 sites
  
- U.S. Resource Allocation Committee (chaired by Jim Shank) oversees fair share usage of resources
  - ❑ Set allocations between ATLAS-wide and U.S. usage
  - ❑ Set allocations between different groups
  - ❑ Set quotas for individual users

# U.S. ATLAS Tier-1 FY08 estimates



## New FY08 Funding Plan All in AY k\$

### Tier 1

Labor	2875
Space + Power	356
MST (travel, maintenanceÉ)	1220
Equipment RP \$	2000
Equipment MR \$	1762
Total Equipment	3762
Total Tier 1 RP \$	6451
Total Tier 1 MR \$	1762
Total Tier 1	8213

➤ Funding never guaranteed, but projections will give us enough to meet MOU pledges.

# From FY08 Management Reserve



## ➤ T1

- ❑ 1762 k in Equipment
  - Approx \$1M for LAN backbone upgrade
  - Approx. \$700k increase in high performance disk

## ➤ Facilities

- ❑ 250k for U.S. Operations Coordinator
- ❑ 110k for Computing Integration Coordinator (Rob Gardner)



# Tier-1 and Analysis Facility Capacity



## ➤ Revised Tier-1 and Analysis Facility Capacity Profile

YEAR	2007	2008	2009	2010	2011
CPU (kSI2k)	2,432	5,400	11,598	18,838	26,875
Disk (TB)	1,175	3,400	8,921	17,262	24,427
Tape (TB)	1,045	2,500	6,276	11,996	18,781
WAN	2 x λ	2 x λ	3 x λ	4 x λ	4 x λ

wLCG Plan to pledge US model					
CPU (kSI2k)	2,560	4,844	7,337	12,765	18,193
Disk (TB)	1,100	3,136	5,822	11,637	16,509
Tape (TB)	603	1,715	3,277	6,286	9,820

## ➤ Expect to become (almost) flat in 2012 and beyond(?)



**Director RACF**  
Michael Ernst

**Deputy Director (RCF)**

**Admin. Assistant**  
Maureen Anderson

**Processing and General Services**  
Tony Chan

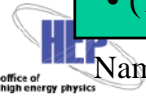
**Storage**  
Shigeki Misawa

**Grid MW and Services**  
Dantong Yu

- Linux Farms + SysAdmin
  - Tony Chan
  - Chris Hollowell
  - Richard Hogue
  - Alexander Withers
  - Tristan Ziska
- GCE Services + SysAdmin
  - Robert Petkus
  - Mizuki Karasawa
  - John McCarthy
  - Jason Smith
  - Morris Strongson
  - New Hire (offer accepted)
  - (Frank Burstein)
- Operations & Infrastructure
  - Richard Hogue
  - Kevin Casella
  - Enrique Garcia
  - (1/2 FTE from ITD)

- Mass Storage
  - Shigeki Misawa
  - Ognian Novakov
  - John Riordan
  - Grace Tsai
  - David Yu
  - New Hire
- Storage Mgmt & Data Movement
  - Gabriele Carcassi
  - Hironori Ito
  - Jane Liu
  - Ofer Rind
  - Iris Wu
  - New Hire (FTS/DDM)
- Central Storage
  - Maurice Askinazi
  - Dave Free

- Production, Data Base and User Support
  - Dantong Yu
  - John DeStefano
  - Carlos Gamboa
  - Yuri Smirnov
  - Tomasz Wlodek
- General Software Env. and Software Development
  - Dimitrios Katramatos
  - John Hover
  - Jay Packard
- Grid Middleware (OSG / WLCG)\_
  - John Hover
  - Jay Packard
  - Xin Zhao

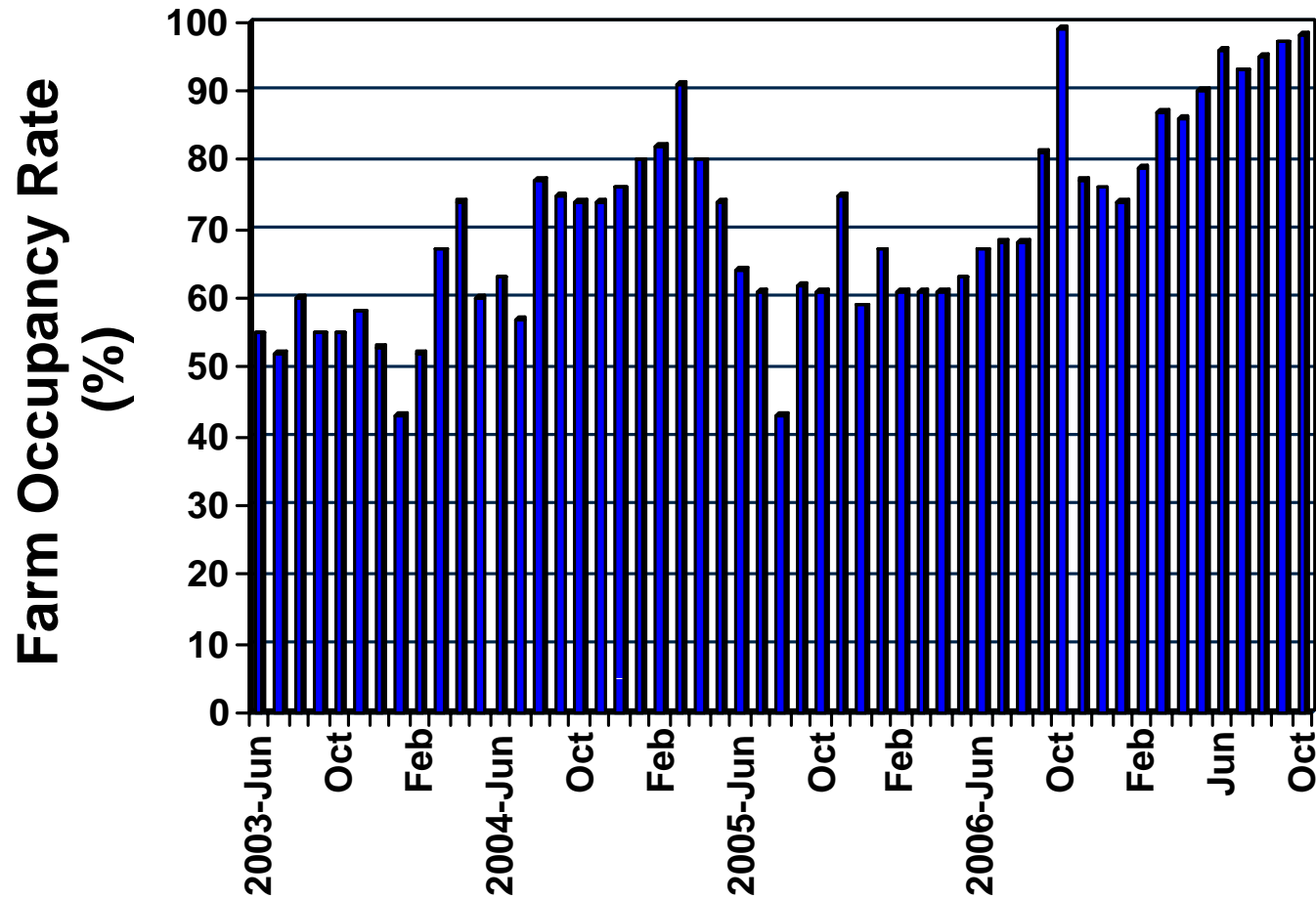


Names in alphabetical order

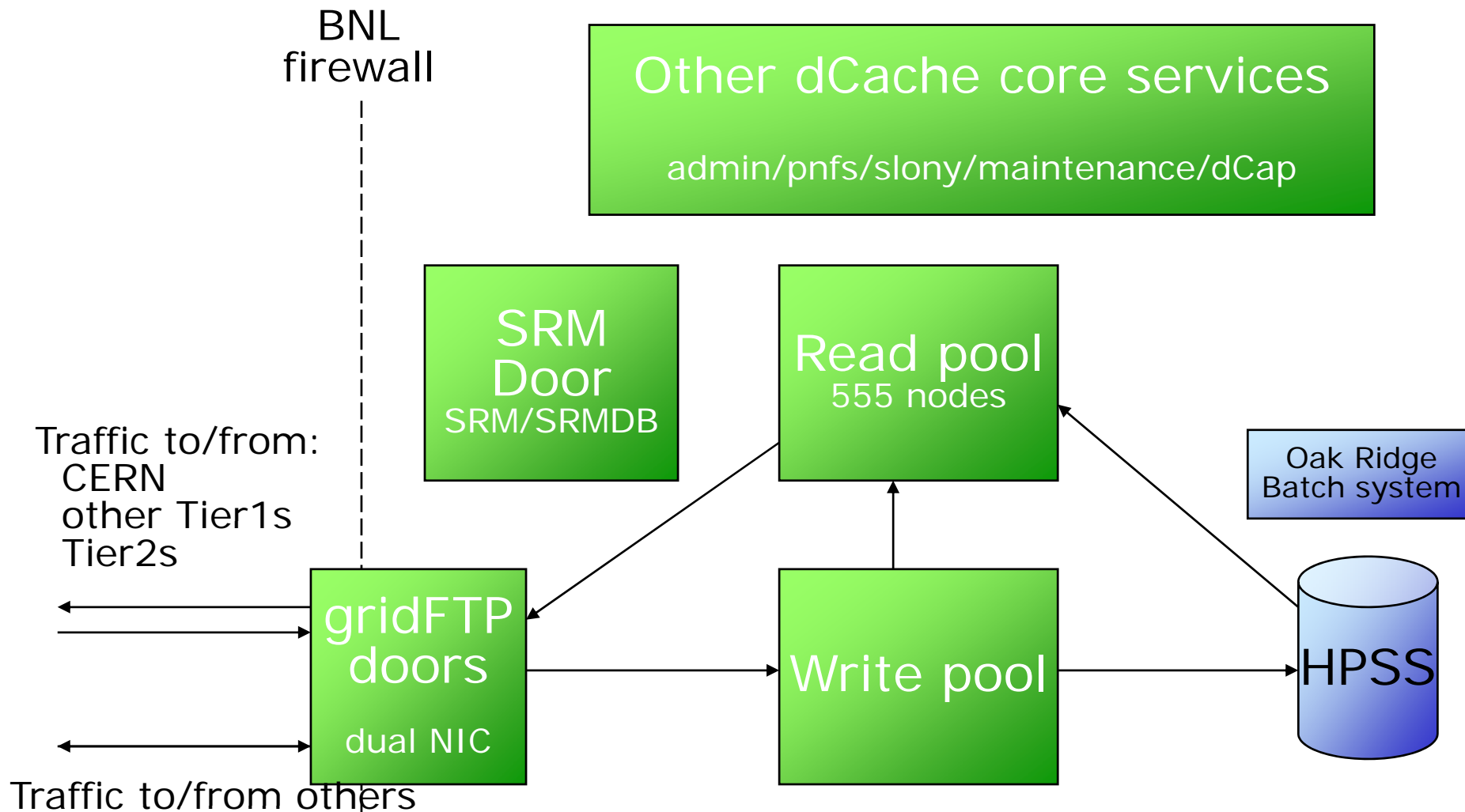
## Processing: RCF/ACF Linux Farm Occupancy



Condor general queue fully enabled in Aug. 2006



# Storage Management for ATLAS Data



Details in Gabriele's talk

Cosmic data replication within clouds (new replication pool)

Tier2	Datasets	Total Files in datasets	Total CpFiles in datasets	Completed	Transfer	Subscribed
BEIJING	50	370	30	4	10	36
CPPM	11	73	7	1	2	8
LAL	98	857	60	1	39	58
LAPP	11	184	25	0	5	6
LPC	21	159	23	3	7	11
LPNHE	98	1205	101	0	36	62
IPNE_02	11	65	16	1	5	5
IPNE_07	11	112	18	5	3	3
SACLAY	98	547	28	0	18	80
TOKYO	291	3768	156	0	65	226

	AGLT2	BU_DDM	MWT2_JU	SLACXRD	UTA_SWT2	WISC
in datasets	662	7015	7015	662	0	0
in datasets	662	7015	6955	639	13	0
in datasets	662	7015	7015	662	0	0
in datasets	651	7004	2334	261	4	386
in datasets	102	848	848	102	0	0
in datasets	662	7015	5263	462	0	190

100% of data requested by 5 T2s

BNL Cloud

Dataset	Files on Source Host	AGLT2	BU_DDM	MWT2_JU	SLACXRD	UTA_SWT2	WISC
M4.0019354.Default.L1TT-b10000000.ESD.v13002503.part0001	3						
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0001	1						
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0002	3						
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0003	2						
M4.0019536.Default.L1TT-b00000001.ESD.v13002502.part0001	49						
M4.0019544.Default.L1TT-b00000001.ESD.v13002502.part0001	93						
M4.0019562.Default.L1TT-b00000001.ESD.v13002503.part0001	39						
M4.0019580.Default.L1TT-b00000001.ESD.v13002503.part0001	4						
M4.0019585.Default.L1TT-b00000001.ESD.v13002503.part0001	16						
M4.0019587.Default.L1TT-b00000001.ESD.v13002503.part0001	8						
M4.0019619.Default.L1TT-b00000010.ESD.v13002503.part0001	1						
M4.0019626.Default.L1TT-b00000010.ESD.v13002503.part0001	25						
M4.0019638.Default.L1TT-b00000010.ESD.v13002503.part0001	8						
M4.0019654.Default.L1TT-b00000010.ESD.v13002503.part0001	1						
M4.0019656.Default.L1TT-b00000010.ESD.v13002503.part0001	10						
M4.0019659.Default.L1TT-b00000010.ESD.v13002503.part0001	3						
M4.0019660.Default.L1TT-b00000010.ESD.v13002503.part0001	3						
M4.0019673.Default.L1TT-b00000001.ESD.v13002503.part0001	0						

Tier-3 at UWM

Dataset	Files on Source Host	BEIJING	CPPM	LAL	LAPP	LPC	LPNHE	NIPNE_02	NIPNE_07	SACLAY	TOKYO
M4.0019354.Default.L1TT-b10000000.ESD.v13002503.part0001	3		1								
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0001	1										
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0002	3										
M4.0019391.Default.L1TT-b00000010.ESD.v13002503.part0003	2										
M4.0019536.Default.L1TT-b00000001.ESD.v13002502.part0001	49				8						
M4.0019544.Default.L1TT-b00000001.ESD.v13002502.part0001											
M4.0019562.Default.L1TT-b00000001.ESD.v13002503.part0001									8		
M4.0019580.Default.L1TT-b00000001.ESD.v13002503.part0001											
M4.0019585.Default.L1TT-b00000001.ESD.v13002503.part0001											
M4.0019587.Default.L1TT-b00000001.ESD.v13002503.part0001									1		
M4.0019619.Default.L1TT-b00000010.ESD.v13002503.part0001											
M4.0019626.Default.L1TT-b00000010.ESD.v13002503.part0001											
M4.0019638.Default.L1TT-b00000010.ESD.v13002503.part0001											
M4.0019654.Default.L1TT-b00000010.ESD.v13002503.part0001	1										
M4.0019656.Default.L1TT-b00000010.ESD.v13002503.part0001	10		2								
M4.0019659.Default.L1TT-b00000010.ESD.v13002503.part0001	3					2			2		

100% of data shared by T2s within the cloud

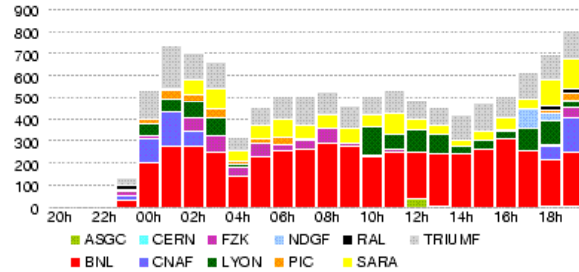
BNL T1 replicated M4 Cosmic Data to T2s and a T3

A. Klimentov at WLCG Collaboration Meeting

Victoria, 1 September 2007

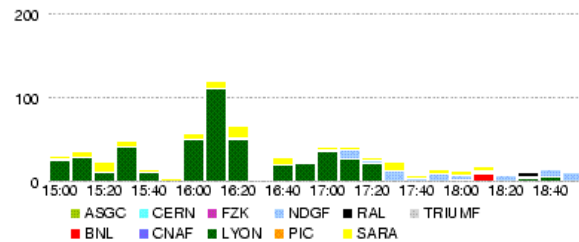
<http://indico.cern.ch/conferenceTimeTable.py?confId=3578>

# ATLAS Cosmic Run (M5) Data Replication



Data replication status 20 hours after replication started

## Data Transfer Performance



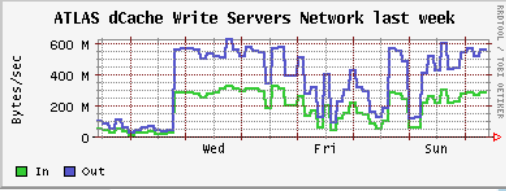
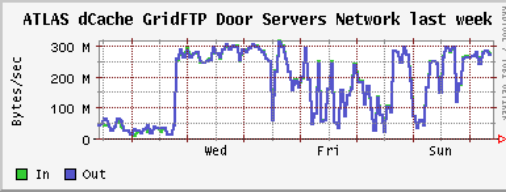
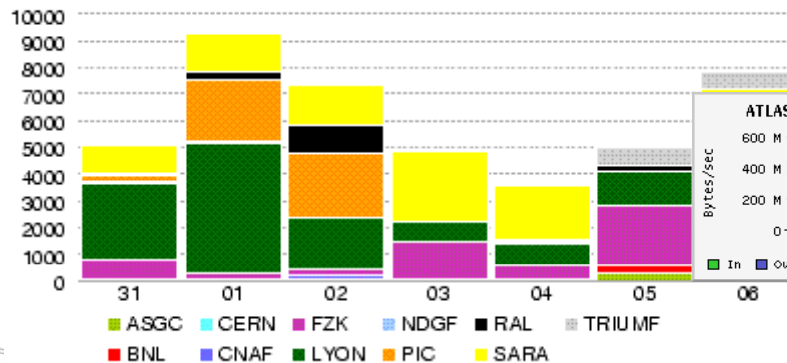
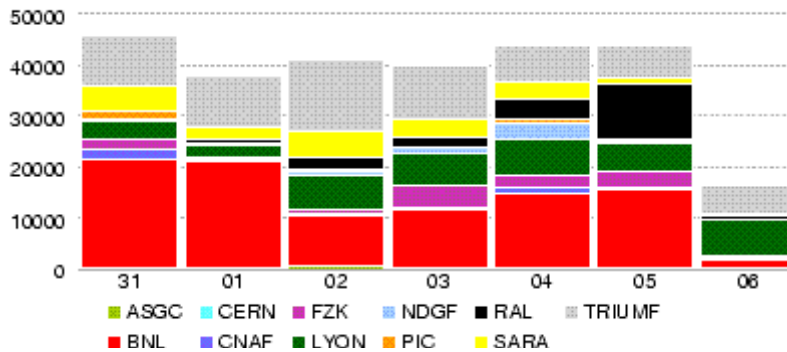
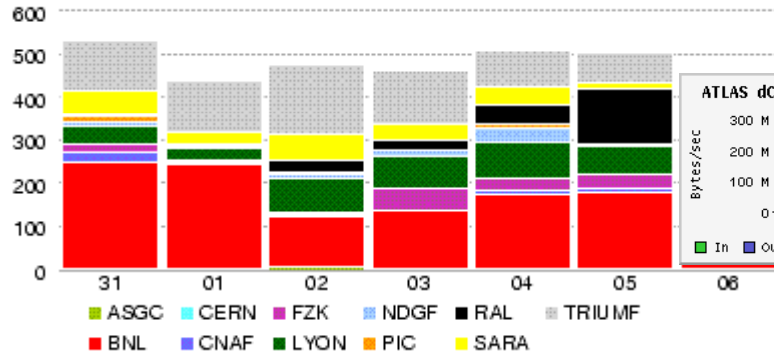
Tier1	Datasets	Total Files in datasets	Total CpFiles in datasets	Completed	Transfer	Subscribed
ASGC	33	171	169	31	2	0
BNL	398	18240	9683	128	202	68
CNAF	45	1149	796	43	2	0
FZK	46	1145	1143	45	1	0
LYON	360	17424	1973	30	103	227
NDGF	24	305	64	5	3	16
PIC	22	818	485	13	1	8
RAL	42	257	189	38	2	2
SARA	47	7872	2642	6	29	12
TRIUMF	341	18073	4433	23	153	165

## Transfer Errors

Activity Summary (Last 4 Hours)  
Click on the cloud name to view list of sites

Cloud	Transfers				Services		Errors			
	Efficiency	Throughput	Files Done	Datasets Done	DQ	Grid	Transfer	Local	Remote	Central
ASGC	99%	1 MB/s	85	24			1	0		
BNL	100%	243 MB/s	2650	136			10	0		
CERN	0%	0 MB/s	0	0			0	0		
CNAF	100%	54 MB/s	444	39			0	0		
FZK	86%	15 MB/s	170	36			27	0		
LYON	63%	57 MB/s	495	37			292	0		
NDGF	79%	25 MB/s	237	28			64	0		
PIC	86%	28 MB/s	242	13			40	0		
RAL	93%	11 MB/s	135	37			10	0		
SARA	93%	96 MB/s	1262	16			91	0		
TRIUMF	100%	123 MB/s	1396	24			4	0		

# M5 Data Replication to BNL



Datasets	Total Files in datasets	Last Subscription	LFC Checked	Last Transfer
2209	72062	Nov 06 21:12:24	Nov 07 12:25:47	Nov 07 12:16:02

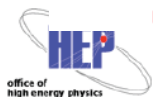
Tier1	Datasets	Total Files in datasets	Total CpFiles in datasets	Completed	Transfer	Subscribed
ASGC	246	1955	1955	246	0	0
BNL	2209	72062	72061	2208	1	0
CNAF	391	7476	7476	391	0	0
FZK	341	10501	10463	335	6	0
LYON	1149	50383	26825	467	462	220
NDGF	178	5101	5101	178	0	0
PIC	161	3540	3540	161	0	0
RAL	397	12829	7599	395	0	2
SARA	439	15484	16643	434	1	2
TRIUMF	928	47226	36508	490	394	44

## Activity Summary (Last 168 Hours)

[Click on the cloud name to view list of sites](#)

Cloud	Efficiency	Throughput	Transfers		Services		
			Files Done	Datasets Done	DQ	Grid	Transfer
ASGC	89%	3 MB/s	1858	325			231
BNL	97%	138 MB/s	66153	2609			2366
CERN	0%	0 MB/s	0	0			0
CNAF	98%	7 MB/s	6797	506			158
FZK	47%	19 MB/s	12343	448			13839
LYON	64%	62 MB/s	26511	575			14887
NDGF	97%	10 MB/s	5108	278			180
PIC	39%	3 MB/s	3096	243			4823
RAL	90%	35 MB/s	15436	509			1707
SARA	63%	30 MB/s	15258	564			8973
TRIUMF	90%	98 MB/s	33828	543			3675

CRITICAL WARNING NORMAL GOOD NO\_ACTIVITY



M. Ernst

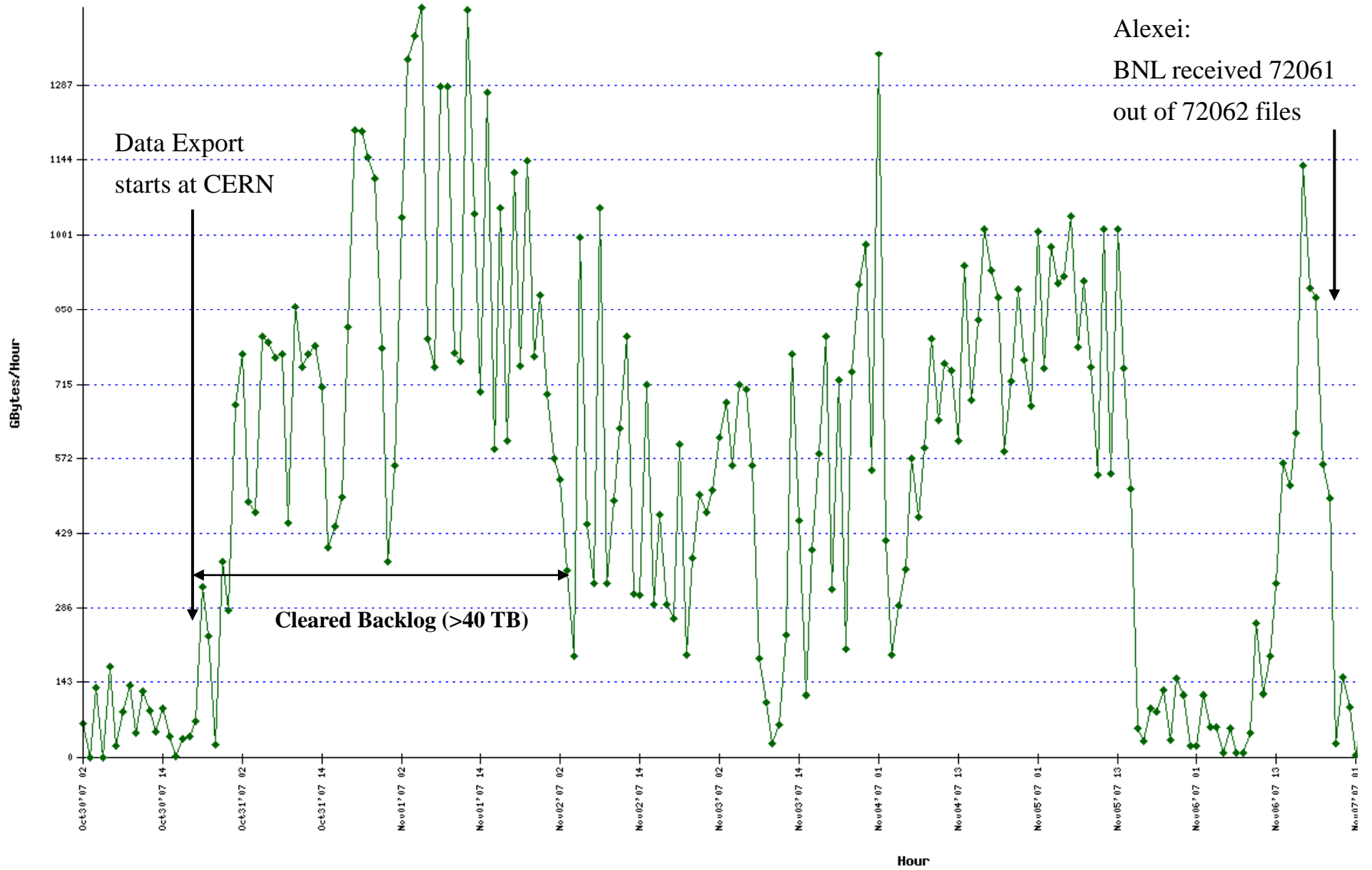
Tier-2/3 Meeting at SLAC

28 November 2007

23

BRUKHAVEN NATIONAL LABORATORY

Atlas Cosmic Ray Run: HPSS Atlas Disk Migration Rate (GBytes/Hour)  
Stared from Oct 30 6PM



4 Drives | 6 Dr. | 8 Drives →



# Roadmap

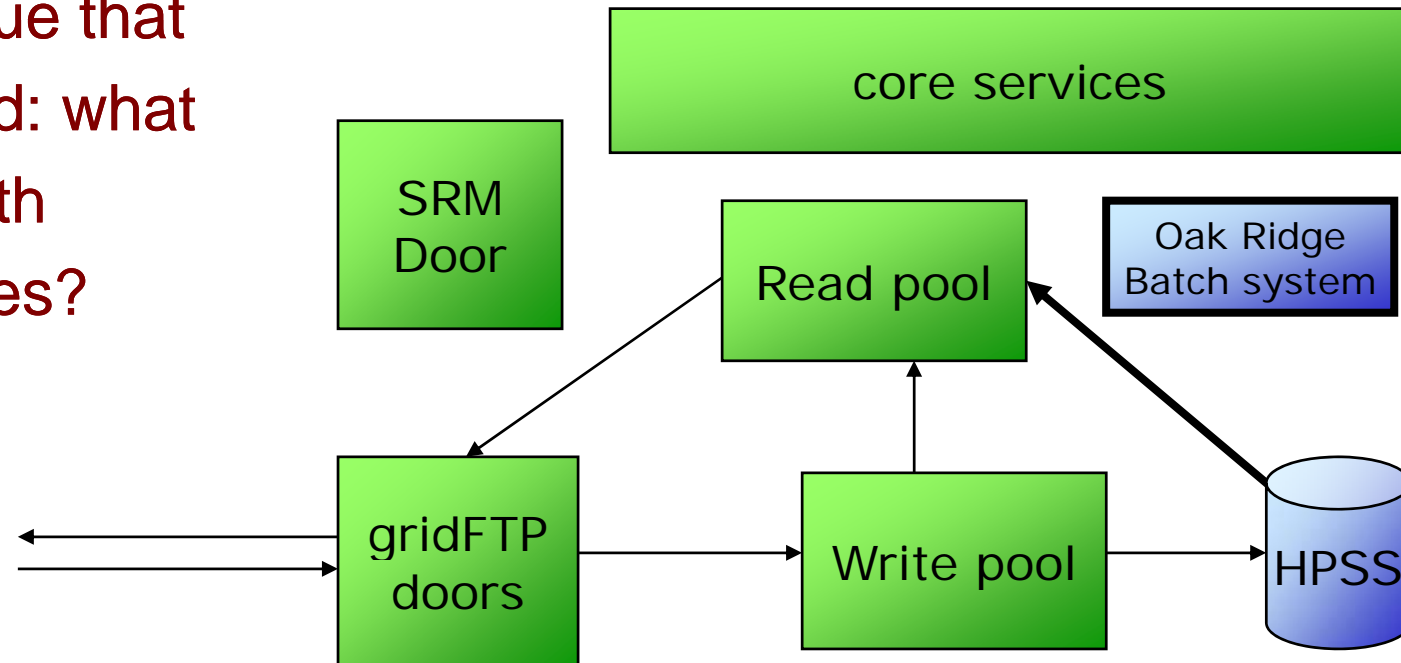


- **Increase the number of gridFTP Servers**
  - ❑ Will convert some former pool nodes to increase the external bandwidth of the system (to 20 nodes, 50 MB/s each)
  - ❑ Already started, expect completion by end December
- **Upgrade to dCache 1.8 and SRM 2.2**
  - ❑ Functionality upgrade
  - ❑ Scheduled for mid-December
- **Upgrade to Chimera**
  - ❑ Claims to solve the problems of scalability related to PNFS
  - ❑ Scheduled for mid-January

# Work in Progress: dCache/HPSS backend



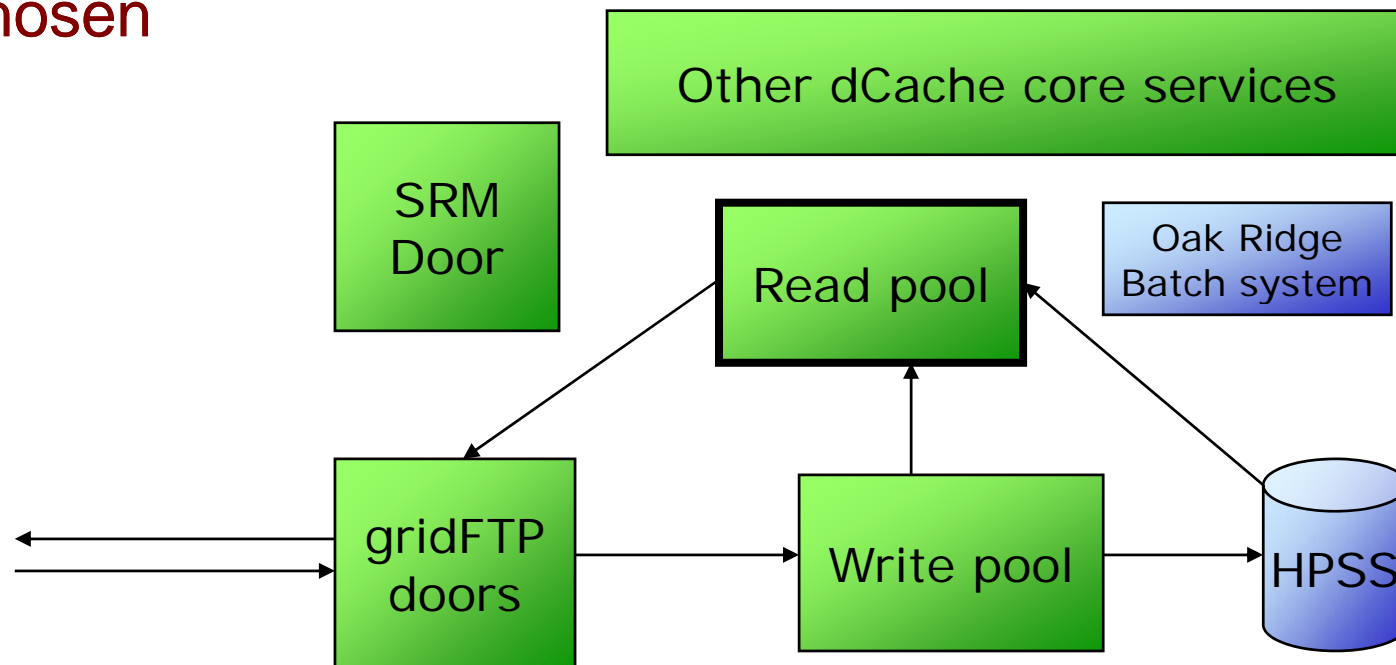
- Strengthen the communication between dCache and HPSS (already started)
- One issue that emerged: what to do with bad tapes?



# Work in Progress: Data Placement



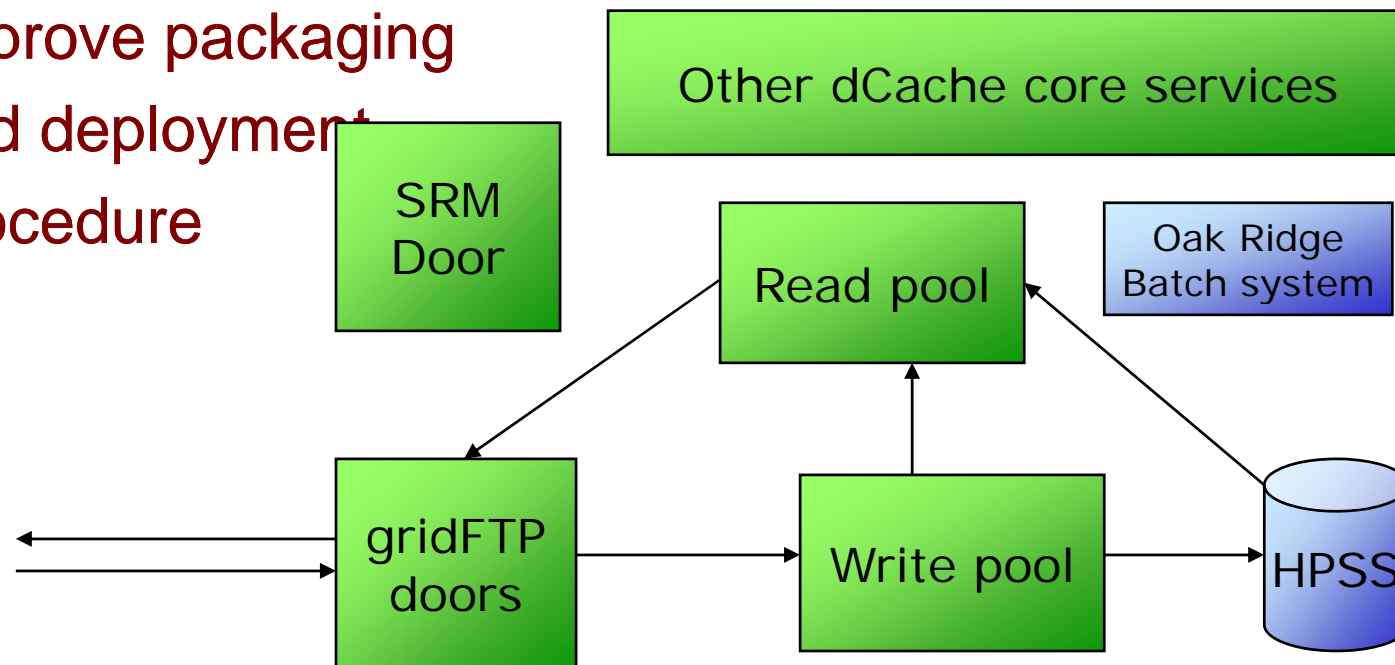
- We need better control of what is in the cache (already started)
- We need to better control how pools are chosen



# Work in Progress: Improve Operations



- Improve Monitoring and Alarming
- Improve (and create) logs that are suitable for operations, and make sure they fit well with the facility infrastructure
- Improve packaging and deployment procedure



# Issues



- Recent US production experience shows urgent need for facilities upgrades at BNL
- Too many US produced files are on tape, slowing down reprocessing
- Long Latencies for pathena users
- Rapid increase in disk storage needed

# Why Tape doesn't work well for us



## ➤ Tapes have evolved to Archive Medium

### □ Sequential versus random access

- 400/800 GB per Cartridge (LTO3/LTO4)
- Average File Access Time (LTO 3) 72 sec
- Average Rewind Time 49 sec
- Unload Time 19 sec
- Data Transfer Time (80MB/s native rate) for 100 MB File ~2 sec

### □ Adding the Robot

- Cell to Drive and vice versa 11 sec per move, w/ 8 arms per cartridge exchange 3 sec

### □ Total (assuming 1 File requested per Cartridge) 145 sec

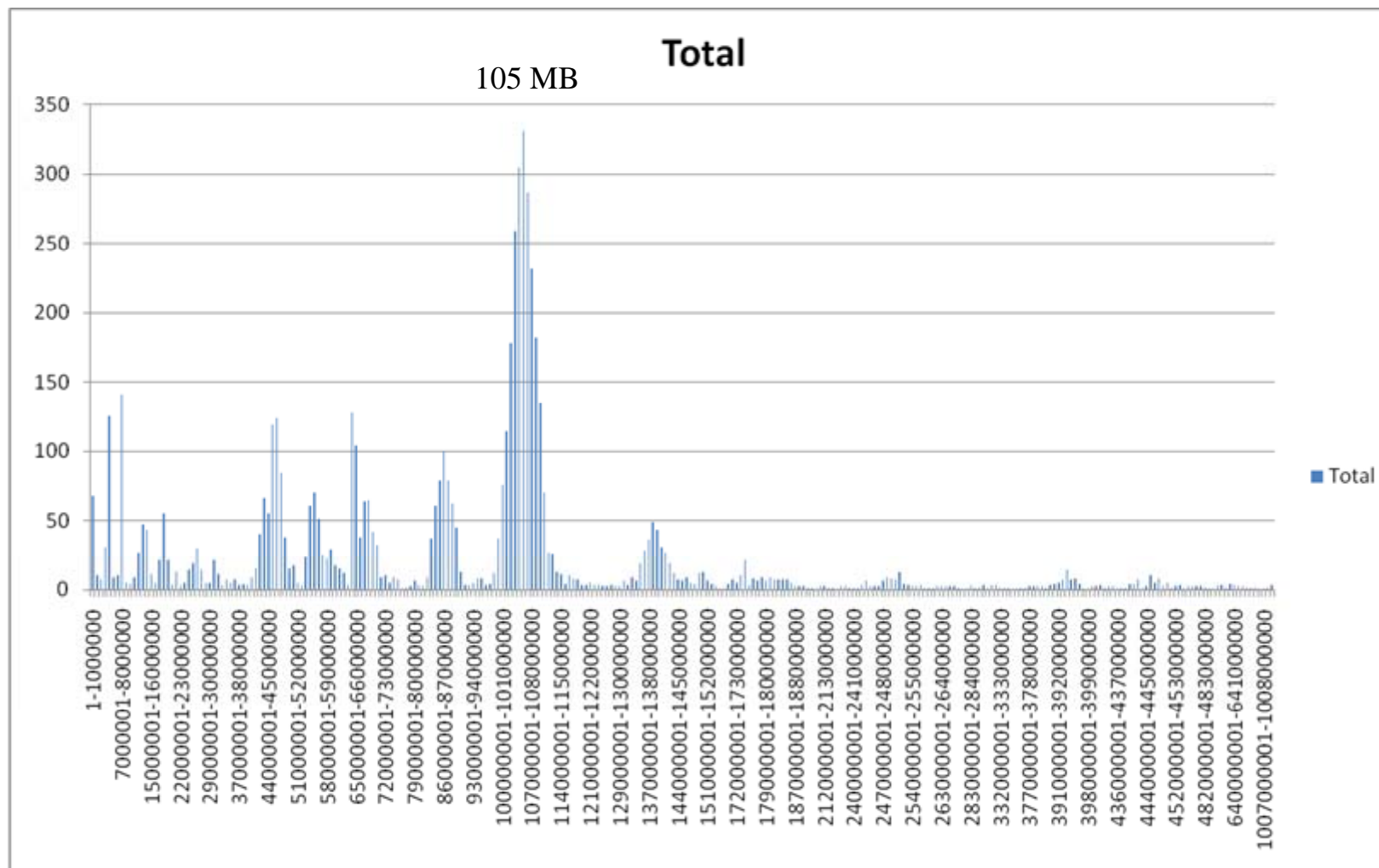
- Minimum File Staging Rate per Day w/ 10 Drives 5958
- For recent Re-Reconstruction observed 4 Files/Cartridge 13090  
Average File Staging Rate per Day

### □ Our Observation 12k – 16k Completed Requests per Day

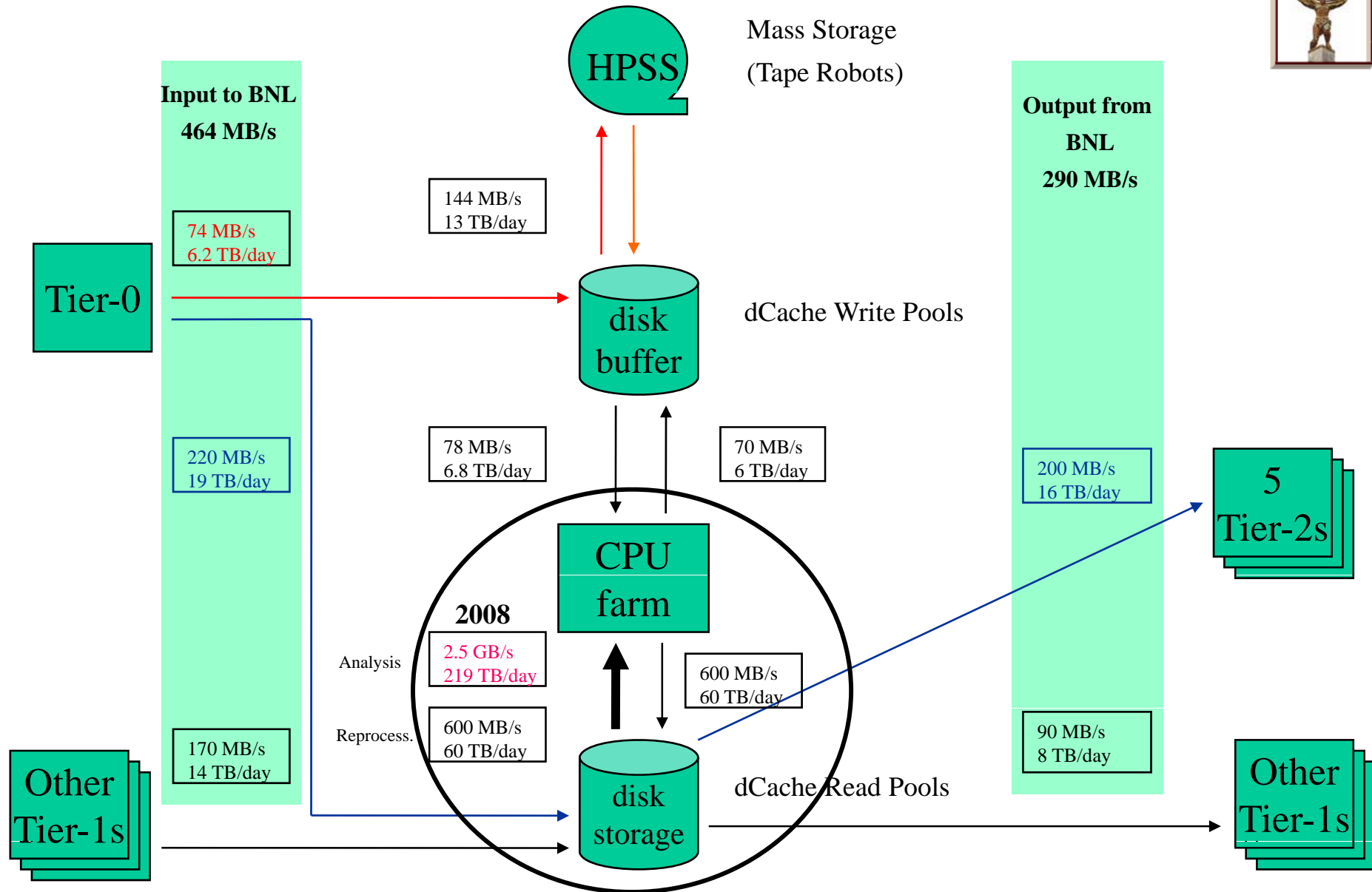
- Expect to double the rate with 10 new LTO4 Drives
  - **First results show we can do ~25k restores/day**

(LTO 3 specs at <http://www.9to5computer.com/sun/Sun%20Storage%20LTO%20ULTRIUM%203%20Tape%20Drive.htm>)

# Pre-Staging – File Size Distribution



# U.S. ATLAS Tier-1 Networking Needs





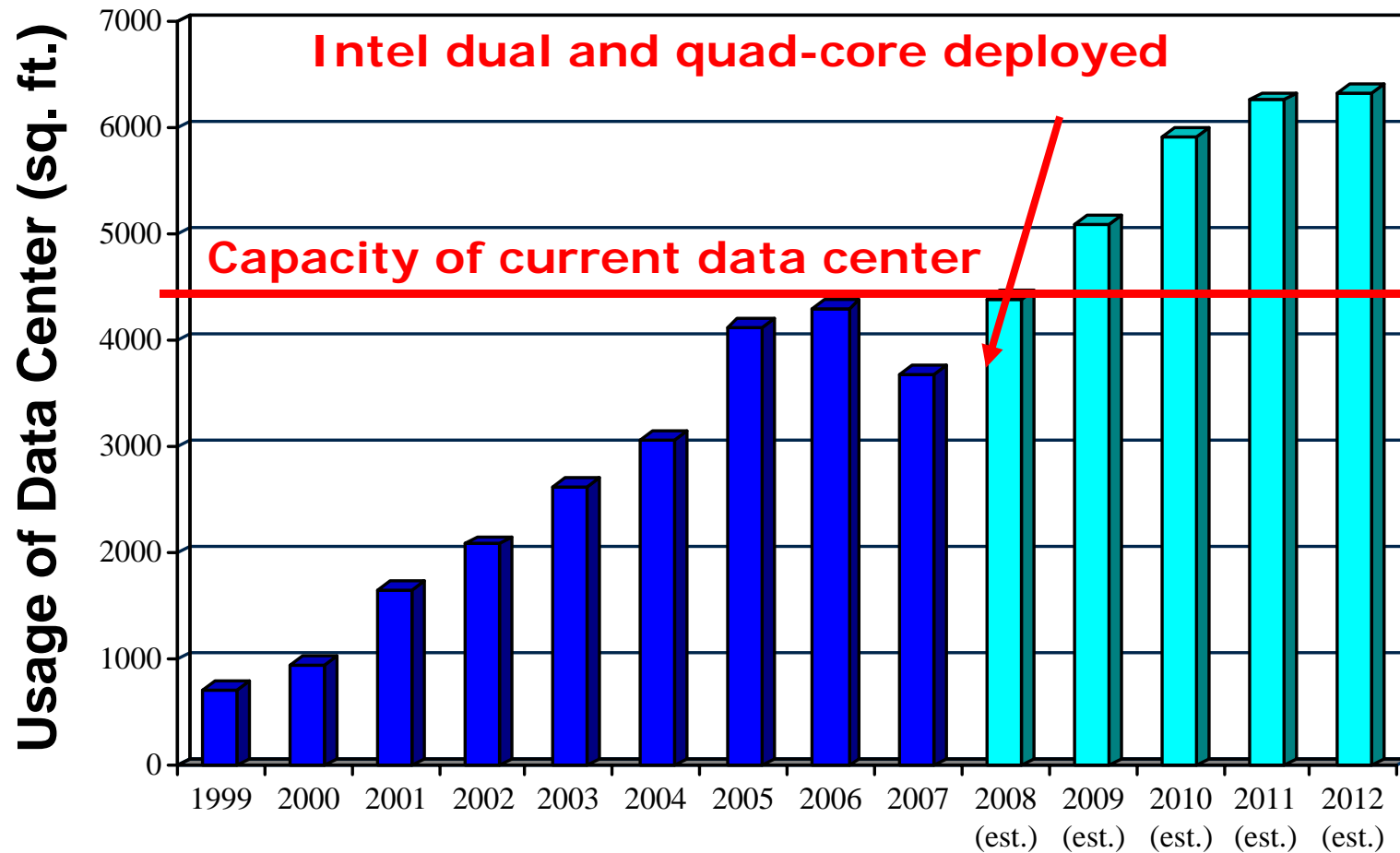
# Physical Infrastructure



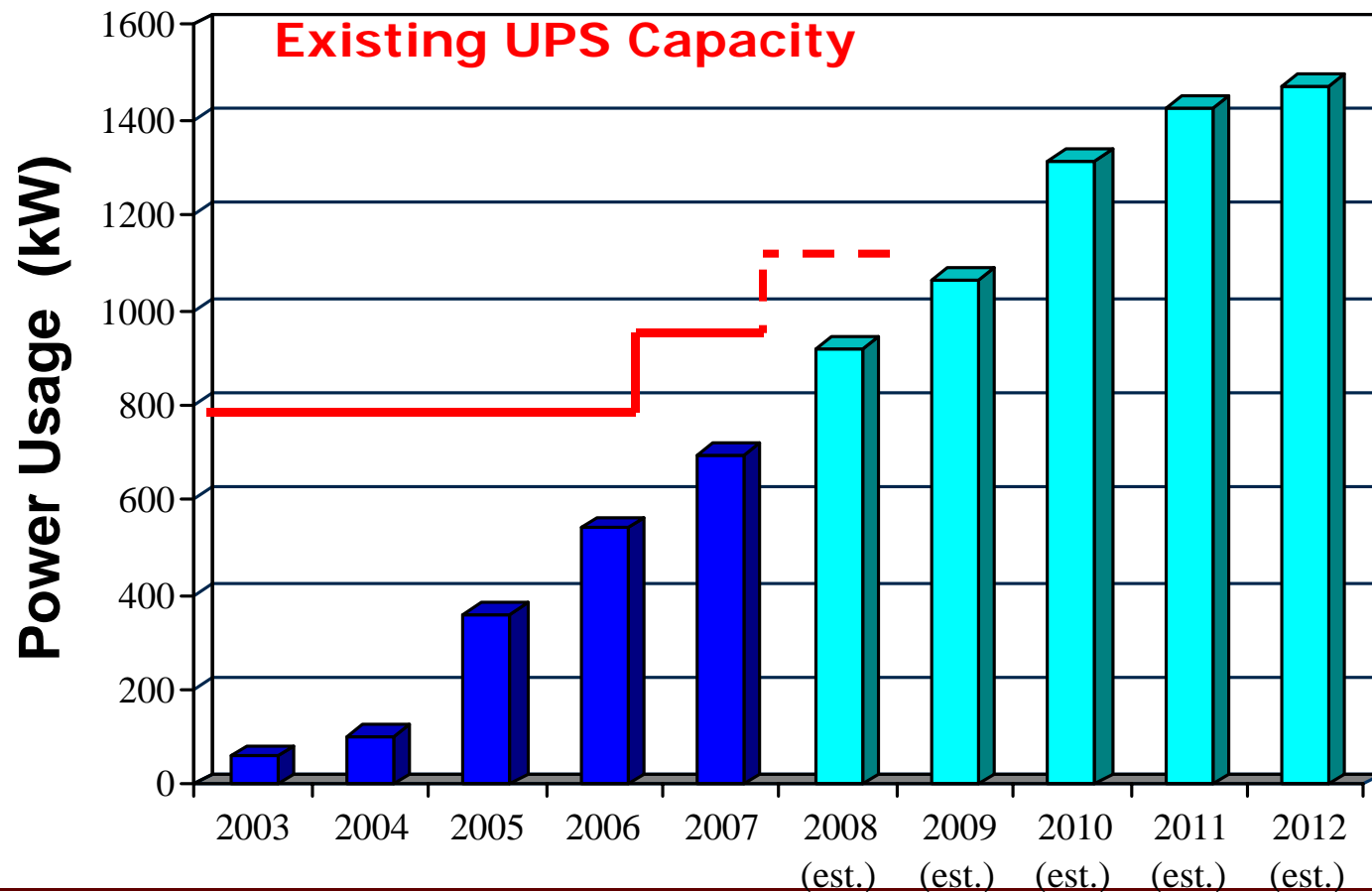
- Have reached limits in all areas
  - Reallocation of space to RCF/ACF allows 2008 expansion
    - Additional power & cooling is needed each year
  - Need expansion of space in 2008 and beyond
    - Working with ITD, BNL Plant Engineering and BNL Management on a plan
  
- This is our top concern at the moment



# Evolution of Space Usage



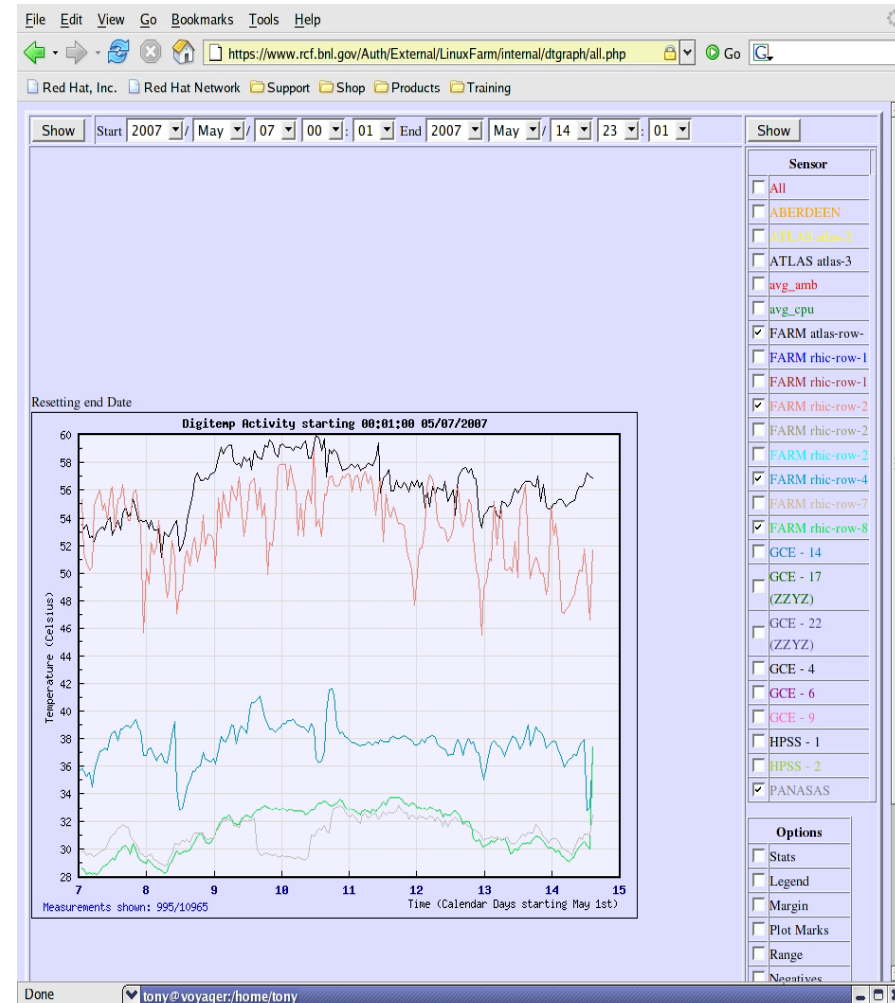
# Evolution of Power Usage



# Infrastructure



- The growth of the RACF has put considerable strain on power and cooling to the building's infrastructure
- UPS back-up power for RACF equipment
- Custom RACF-written script to monitor power and cooling issues
- Alarm escalation through RT ticketing system
- Automatic shutdown of Linux Farm during cooling or power failures



# Transition to Operations



## Stability is important, maybe more than performance

- Define milestones for uptime, success rates as measured by Site Availability Monitoring tests (building VO-specific tests on top of OSG/WLCG SAM tests) and Dataset replication exercises
- Define and put procedures in place to protect services from extended disruptions
- The Tier-1 center at BNL and the Tier-2's in the U.S. are tightly coupled
  - ❑ “BNL Cloud” according to the ATLAS Computing Model
- In preparation of letting the Tier-2's act according to the Computing Model
  - ❑ Carrying the load of MC production
  - ❑ Hosting datasets for analysis
  - ❑ Hosting the work of various analysis groups
  - ❑ Supporting “local communities”
  - ❑ The effort to produce physics results, distribution and processing of Cosmic Ray data and FDR will be important tests of our readiness

# Monitoring



- Evolution of RACF from local to globally available resource highlights the importance of a reliable, well-instrumented monitoring system
- RACF monitors service availability, system performance and facility infrastructure (power and cooling)
- Mixture of commercial, open-source and RACF-written components
  - ❑ RT
  - ❑ Ganglia
  - ❑ Nagios
  - ❑ Infrastructure
  - ❑ Condor
- Choices guided by desired features: historical logs, alarm escalation, real time information

# Facility Operations



- Facility operations is a manpower-intensive activity at the RACF
- Careful choice of technologies required for scaling of capacity and services
- Operational responsibility divided among major support groups within the facility (storage, computing, grid operations)
  - ❑ Software upgrades
  - ❑ Hardware lifecycle management
  - ❑ Integrity of facility services
  - ❑ User account lifecycle management
  - ❑ Cyber-security
- Experience of RHIC operations for the past 8 years
- Used as a starting point for U.S. ATLAS Tier 1 facility operations



# A New Operational Model for the RACF



- RHIC facility operations is a system-based approach
- ATLAS needs support for (mostly) remote users
- Service-based operational approach better suited for a distributed computing environment
- New SLA for RACF incorporates service-based approach
- Mapping of services to related systems

# Implementing the new SLA



- Further instrumentation of monitoring tools to improve diagnosis of error conditions and speed up alarm response times
- Emphasis on automation of response to common error conditions
- Service Coordinators oversee response and resolution of error conditions
- Other implementation details being discussed and refined

# A Dependency Matrix



File Edit View Insert Format Tools Data Window Help

Arial 10 75%

M13 f w Σ =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Service Class	C	C	C	C	C	C	C	C	C	C	C	C	M	M	C	C	C	M	C	C
2		Access to mass data archive - RBC	Access to mass data archive - ATLAS	Access to local NFS storage	Access to NFS storage	SMB File Serving (SMBFA)	Data Catalogs	User Databases	Grid job execution	Local job execution	PANDA Service	Central Reconstruction Service	Gateways	Web documents	Email	Printing	Data Protection	Accounting	Software Subversion Service	Support and Monitoring	
3	HPSS	X	X									X									
4	dCache		X																		
5	DG2		X																		
6	FTS		X																		
7	PANDA										X										
8	Gatekeepers								X												
9	Gateways								X	X			X								
10	Ferm								X	X			X								
11	Condor/LSF								X	X			X								
12	User Databases						X	X					X								
13	NFS			X								X									
14	AFS				X							X									
15	Backup					X						X					X				
16	Kerberos	X	X	X	X				X	X		X									
17	ModProxy		X						X	X		X									
18	VOMS		X						X	X		X									
19	GUMS		X						X	X		X									
20	DNS/LDAP/NTP	X	X	X	X	X	X	X	X	X	X	X	X	X	X						X
21	BDII		X						X	X		X		X	X						
22	Graba																	X			
23	RT																				X
24	Naples													X							X
25	Ganglia													X							
26	Web svr													X							
27	Network and firewalls	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
28	Power	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
29	A/C	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
30																					
31	Critical	C																			
32	Managed	M																			
33	Unmanaged	U																			
34																					
35																					
36																					

IME Status

Sheet 1 / 3 tony@voyager:/home/tony

# Challenges Ahead



- Growth of the facility stressing existing infrastructure -- new facility space available Fall 2008 and Summer 2009
- More efficient use of facility resources -- too many needs chasing too few resources
- Enhanced integrated operation of the facility in concert with other computing centers (T0, T1, T2, etc) -- essential in a distributed computing environment
- Changes to current operational policy required

# Securing the Facilities' Readiness



- **Towards ATLAS Milestones**
  - ❑ Computing Integration Program in place which aims at building the Integrated Virtual Computing Facility that we need to support LHC Data Analysis for the ATLAS Community in the US
  - ❑ With exercises designed to verify sites' readiness, stability and performance
  - ❑ Coordinated by Rob Gardner (UC) and the U.S. ATLAS Facility Manager
    - Organizing quarterly F-to-F meetings w/ Tier-2's, now incl. Tier-3's
  - ❑ Detail's in Rob's Presentation
- **Exploit commonality and establish (technology) baseline whenever possible**
  - ❑ Synergy allows to bundle resources (development and operations)
- **Site Certification**
  - ❑ Site admins are asked to install well defined software packages and to make needed capacities available to the Collaboration
  - ❑ We continuously run use-case oriented exercises, document and archive the results
    - Load Tests – Data Transfers on a basic level
    - Dataset replication based on high-level functionality (DDM/DQ2)
    - Processing (Analysis job profile)
      - Grid Job submission (PanDA) – distribution based on data affinity
      - Local data access (from SE)

# Projections for U.S. Tier 2's



- Totals outline capacity committed to international ATLAS
- ~ 20% Capacity on top of totals retained under US control for US physicists

		2007	2008	2009	2010	2011
Northeast T2	<i>CPU (kSI2k)</i>	394	685	1,049	1,592	1,968
	<i>Disk (TB)</i>	103	244	445	727	1,024
Great Lakes T	<i>CPU (kSI2k)</i>	581	985	1,406	1,670	2,032
	<i>Disk (TB)</i>	155	322	542	709	914
Midwest T2	<i>CPU (kSI2k)</i>	826	1,112	978	1,262	1,785
	<i>Disk (TB)</i>	213	282	358	382	512
SLAC T2	<i>CPU (kSI2k)</i>	550	820	1,202	1,191	1,685
	<i>Disk (TB)</i>	228	482	794	1,034	1,462
Southwest T2	<i>CPU (kSI2k)</i>	998	1,386	1,734	1,966	2,514
	<i>Disk (TB)</i>	143	258	328	650	1,103
<b>TOTAL US Tier 2's</b>						
	<i>CPU (kSI2k)</i>	3,348	4,947	6,367	7,681	9,982
	<i>Disk (TB)</i>	842	1,587	2,487	3,482	5,015

- Planned Pledge for 2012 (and beyond?) expected to stay flat

# Capacity at end of September 2007



## ➤ Dedicated processing cores, usable storage

- ❑ T1: 1600 cores, 1200 TB
- ❑ AGLT2: 550 cores, 297 TB
- ❑ NET2: 392 cores, 144 TB
- ❑ MWT2\_IU: 128 cores, 110 TB
- ❑ MWT2\_UC: 136 cores, 102 TB
- ❑ SWT2-UTA: 300 cores, 16 TB
- ❑ SWT2-OU: 260 cores, 16 TB
- ❑ WT2: 312 cores, 51 TB
- ❑ Total: 4060 cores / 6.5 MSI2k, >1936 TB
  - Not including new resources at MSU and UTA
  - Normalized w/ Scaling Factors as defined in OSG

# Summary



- The BNL Tier-1 serves as the hub and principal center of the US community, with scale-up for data taking underway
- US ATLAS Tier-1 facility at BNL is on track to meet the performance and capacity requirements of the ATLAS computing model augmented to supply appropriate additional support to US physicists
- The facilities, both Tier-1 and Tier-2's, have performed well in both ATLAS computer system commissioning and WLCG service challenges
  - ❑ An Integration Program is in place to ensure readiness in view of the steep ramp-up
  - ❑ 2007 – Excellent contribution of U.S ATLAS Tier-2 Sites to high volume production
- Space, Power & Cooling at the Tier-1 center on the critical path
- Overall, progressing well towards full readiness for LHC data analysis