Data Distribution and Placement Strategies

Kaushik De

University of Texas At Arlington

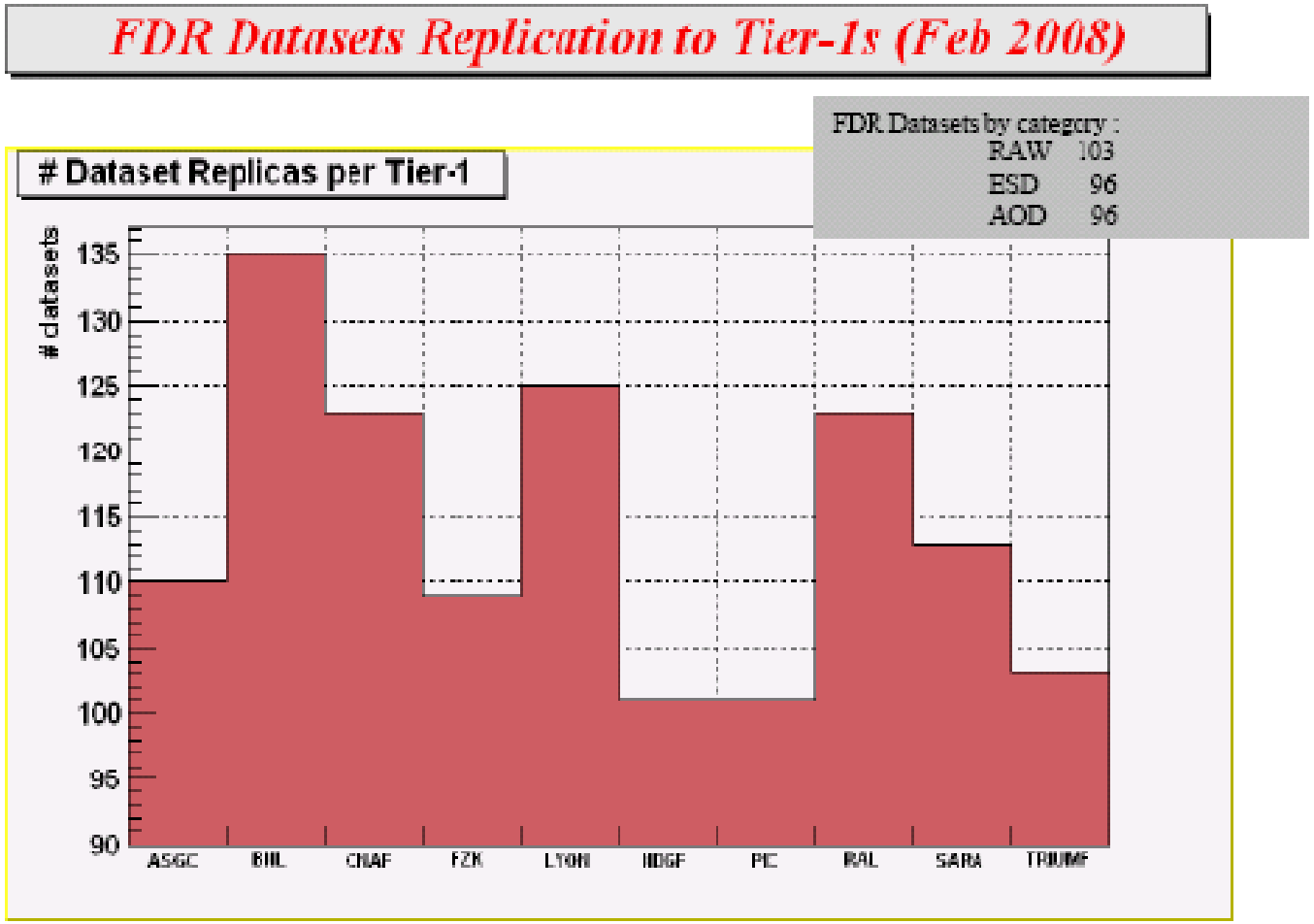
US ATLAS Facility Meeting, UNC

March 4, 2008

Introduction

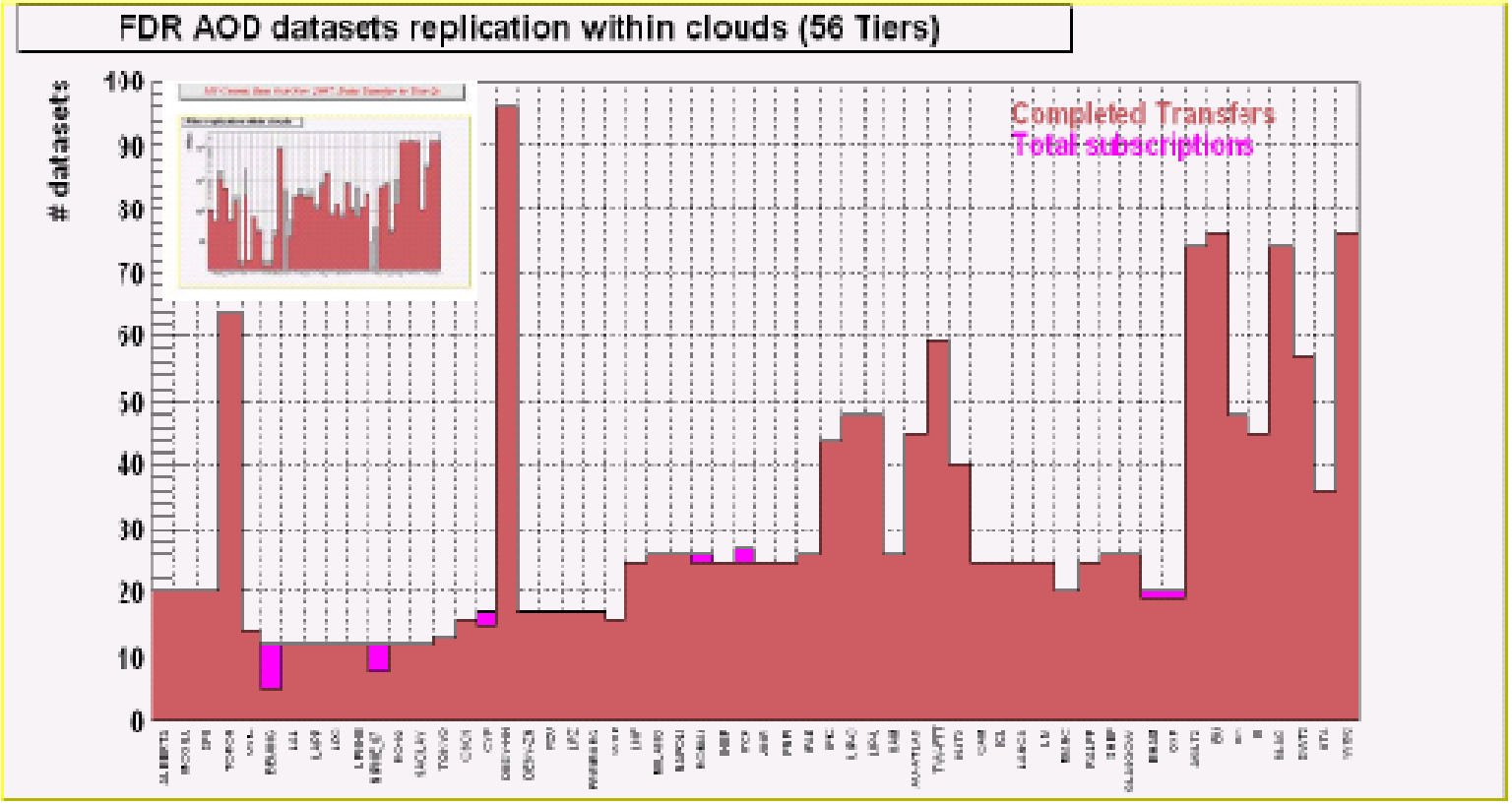
- DDM – continues to be challenging issue for ATLAS
- ADC plan for data distribution and placement is improving
  - FDR-1 and CCRC mostly positive
  - Transfer rates good – on good days
  - Dataset completion rate is improving
  - Data distributed to Tier 1's according to MoU share - works well
- Main U.S. issues
  - Central distribution to T2 sites is too rigid
  - Central deletion policy is too aggressive
  - Storage pinning works – to keep files on disk, but need strategy
  - Disk shortfall – requires storage reorganization

Data Replication to Tier 1's



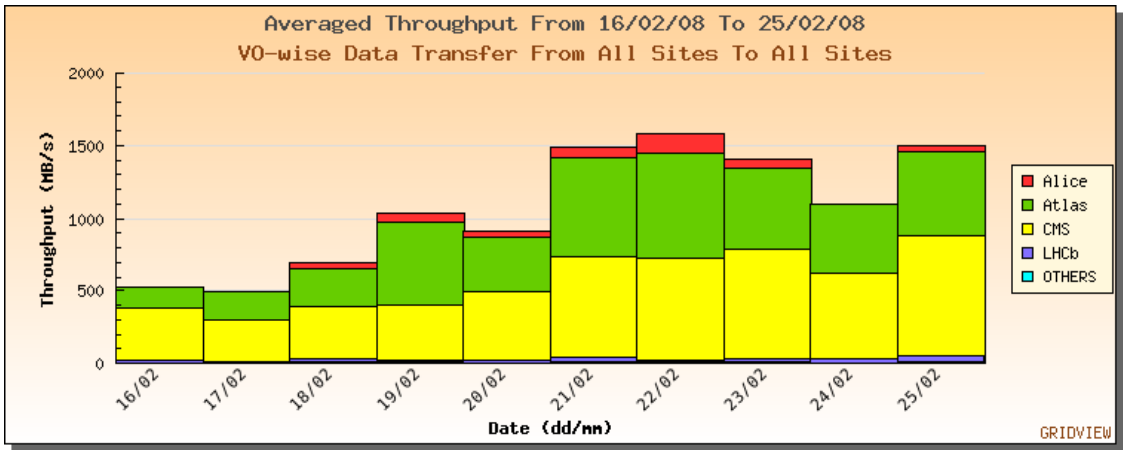
Data Replication to Tier 2's

FDR Feb 2008. Data Transfer to Tier-2s



A Klimentov ATLAS SW week Feb, 2008 10

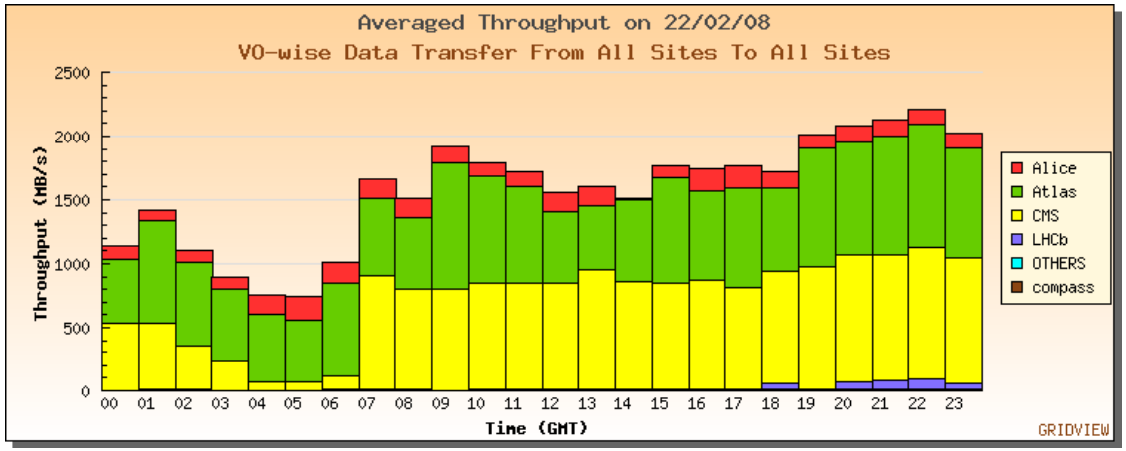
You'll never walk alone



Weekly Throughput

2.1 GB/s out of CERN

From Simone Campana



Data Location Model

- 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
  - Tier 3's will be associated with Tier 2 sites?
  - Tier 3 model is still not fully developed – evolving

Slide from SLAC meeting in Nov. 2007

Data Distribution to U.S. Sites

- We will rely on T0 export to send data to BNL T1
- We will also rely on central T1<->BNL T1 replication
- For Tier 2 and Tier 3 sites the plan is different
  - The central data distribution model follows rigid MoU shares
  - Does not have flexibility to distribute among multiple sites within T2
  - Will not distribute to Tier 3 sites
  - U.S. opting out of central plan - we are setting up machinery for data replication managed by U.S.
  - Tools are not ready yet – will need automated replication and monitoring tools within the U.S. (hopefully we can use ATLAS-wide tools instead of reinventing the wheel)
- We will not limit data according to rigid computing model – instead plan to actively manage based on physicist usage

U.S. Data Distribution Tool

- Need to be developed based on ADC toolset
  - Automatic discovery of data arriving at T1
  - Automatic subscriptions to Tier 2's
  - Ability to modify shares dynamically from web interface
  - Semi-automatic distribution to Tier 3's
  - Robust monitoring – through Panda monitoring
- Need information from Panda
  - Usage pattern of datasets - data distribution strategy should use this information dynamically
  - Space availability should be factored in
- Goal: to manage data actively – replication + deletion
- Time scale – few months

Data Deletion

- ADC DDM operations team maintains list of deletions
  - Sites clean up on their own using centrally provided scripts
  - Effort is too manual and does not scale
  - This has not worked well – new proposal to centrally delete data by ADC DDM ops group from all sites using SRM v2.2 capabilities
- U.S. deletion strategy
  - We have requested to opt out of central deletion
  - New responsibility for U.S. team – need software development
  - Need multi-level plan for Tier 1, Tier 2's (and some Tier 3's)
  - Also multi-process plan for MC production, LHC data, user data…

Multi-level Strawman Plan

- Tier 1
  - Use central deletion list automatically (Tier 1 has unique requirement – archival storage)
  - Manage data by type – sort by disk-resident, tape-resident (for example, pin all AOD's, new ESD's…)
  - Allow less used data to move out of disk cache
  - May not use space tokens rigidly – evolve based on experience
- Tier 2's
  - Need at least two partitions (space tokens – ATLASDATADISK, ATLASMCDISK or storage heirarchy?)
  - Production data – cache as much as possible (for reuse), delete periodically and automatically
  - Replicated data – manage according to physicist usage
- Tier 3 – partly automated, partly by request

SRM2.2 space tokens

- Space tokens at T1s
  - ATLASDATADISK and ATLASDATATAPE correctly confugured in all T1s
    - ATLASMCDISK and ATLASMCTAPE need testing
  - Should start thinking about T1D1
    - ATLASDATADISKTAPE and ATLASMCDISKTAPE
    - It is clear how this will be implemented at every T1?
  - Should start thinking about group analysis at T1
    - ATLASGRM<group> need to be setup for a few groups
- Space tokens at T2s
  - Many T2s are still on SRMv1
  - SRMv2 T2s set up ATLASDATADISK and ATLASMCDISK
    - Being tested right now.

Comment by KD: ATLASUSERDISK, ATLASGRM…?

From Simone Campana

Monte Carlo with SRM2.2

- Need to start using SRM2.2 for MC production
  - Panda Server and Panda job are SRM2.2 aware
    - Running in “SRM2.2 mode” in NL cloud. Encouraging results
  - Missing pieces:
    - SRM2.2 is in every T1, but what about T2s? Space tokens?
    - DPM needs a fix for ACL in directory creation.
      - Backported to DPM1.6.7 (current version). Under certification.
    - Need lcg_util 1.6.7, available only for SL4
      - All T1s are on SL4, but T2s? An ATLAS deadline (March 15) has been announced
- My feeling: FDR2 production with mixed SRMv1/v2 scenario
  - Not a problem for T2 (production “buffers”).

From Simone Campana

Space Tokens et al

- Need to flesh out a fully evolved plan for U.S. that optimizes for strengths of dcache in managing mixed storage (computing model may be too rigid in disk/tape separation of tokens, and may fragment storage too much)
- BNL is already using space tokens for FDR and CCRC
- We also need to provision for new storage needs as required by ATLAS – ex. FDR-2 mixing
- Usage model for Tier 2 space tokens is still not fully defined

FDR-2 Mixing at BNL

Job logistics: Mixing

- Note that damaged data is added in this step
- Move all RDO to BNL: estimate 50M events, 50 TB, 200K files.
- Mixing jobs run at BNL
- 1 job per lumi block per SFO: 30x5=150 jobs "per hr run": 1500 jobs total
- Average job puts out 5K events, ~15K input. 3 days CPU.
  - Jobs limited by RDOtoBS; if this were optimized.
- Assume 3 weeks for this.
- Move Bytestream to CERN for SFO loading

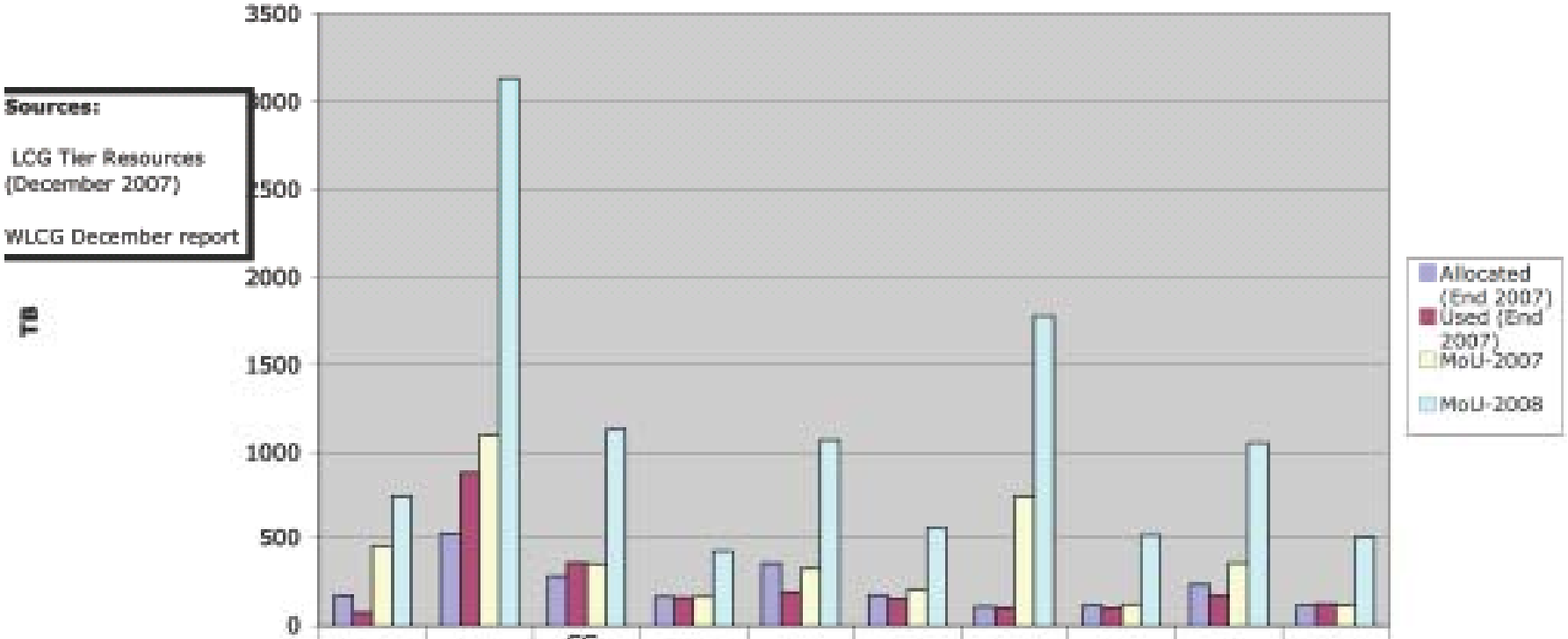
FDR Ian Hinchliffe 2/25/08 18

Storage Management

- Tier 1 storage systems
  - Disk storage projected to grow by factor of three in ~6 months
  - Additional funding also expected from management reserve
  - During past few months many new issues have emerged
- Disk/tape dcache pools – the default at BNL
  - Allows unlimited space – write pools automatically push data to tape
  - Does not work well for small files (log files), or volatile user output
  - Solution: new disk only pool was set up recently
  - Remaining issue: need tools to manage space (no longer infinite)
  - Does not work well for computing model – AOD, RDO, Evgen, DPD etc need to be on disk (large fraction of these got pushed to tape)
  - Solution: software to manage 'pinning' will be rolled out soon
- Have not tackled Tier 2 storage issues yet!

Slide from SLAC meeting in Nov. 2007

Disk resources at ATLAS Tier-1s



	ASGC	BNL	CC-IN2P3	CNAF	FZK	NDGF	SARA	PIC	RAL	TRIUMF
Allocated (End 2007)	167	520	280	160	351	168	105	111	233	110
Used (End 2007)	71	885	354	155	187	150	99	95	165	120
MoU-2007	450	1100	348	162	322	198	750	115	352	110
MoU-2008	750	3136	1133	420	1072	556	1778	512	1056	500

G. Poulard - CERN PH-ADP ATLAS Week 14 Feb. '08 12

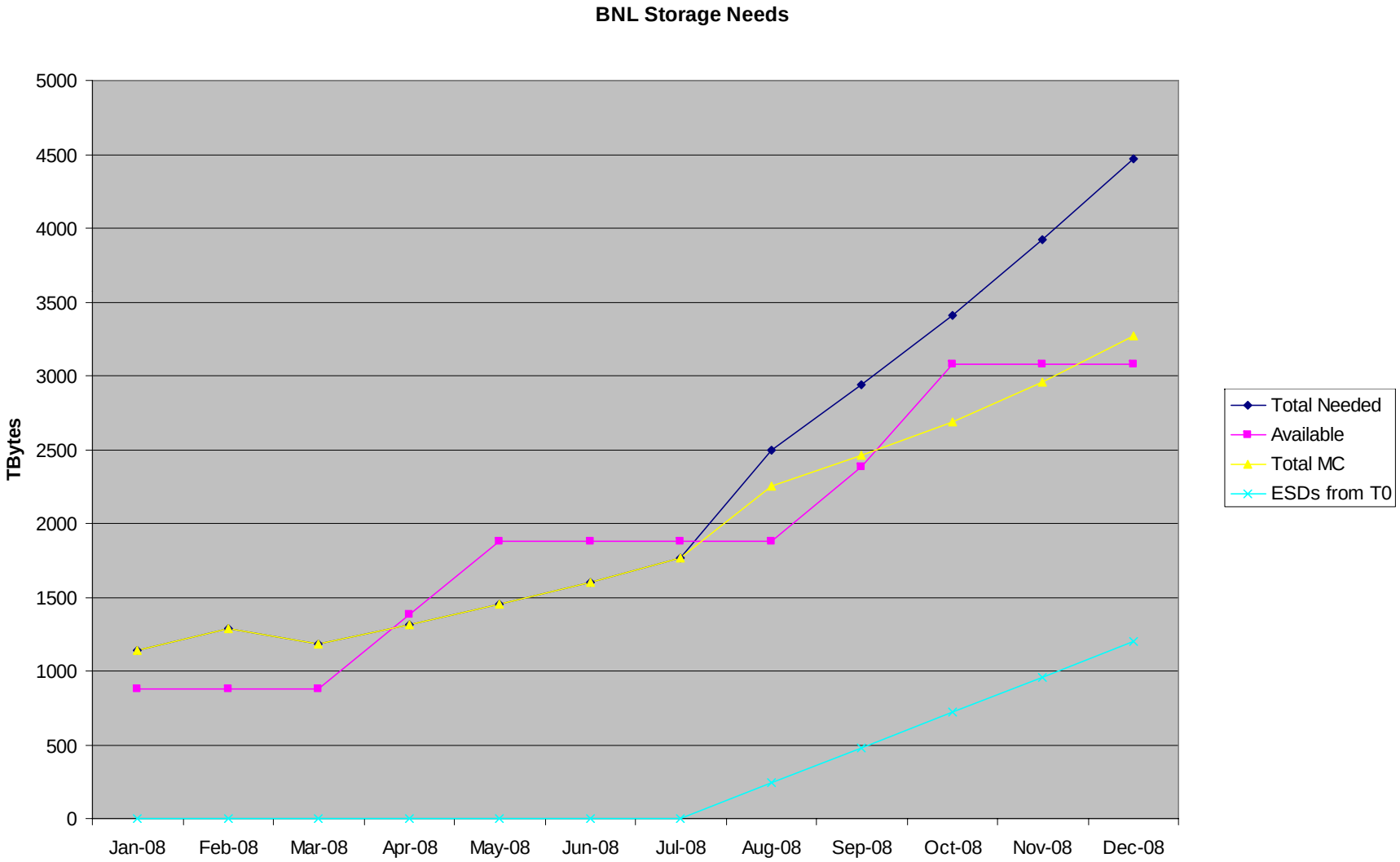
BNL Disk Usage

Type	Number of files	Size on disk (TB)
AOD	1.4 M	91
RDO	3.4 M	365
ESD	561 k	286
Hits	2.5 M	145
NTUP	912 k	25
EVNT	164 k	18
Log	4.0 M	23
misc.	816k	177
Total	13.9 M	1130

- Includes 35 TB M4 data and 90 TB of M5
- In addition to the above we have ~100TB disk space for Write Pools & Buffer and ~200TB for the FDR/CCRC.

J. Shank ATLAS Week 14 Feb., 2008 20

T1 Storage Projection for 2008



Capacity Projections

		2007	2008	2009	2010	2011
Northeast T2	*CPU (kSI2k)*	384	685	1,049	1,592	1,986
	Disk (TB)	103	244	445	727	1,024
Great Lakes T	*CPU (kSI2k)*	581	985	1,406	1,670	2,032
	Disk (TB)	155	322	542	709	914
Midwest T2	*CPU (kSI2k)*	826	1,112	978	1,262	1,785
	Disk (TB)	213	282	358	382	512
SLAC T2	*CPU (kSI2k)*	550	820	1,202	1,191	1,685
	Disk (TB)	228	462	794	1,034	1,462
Southwest T2	*CPU (kSI2k)*	998	1,386	1,734	1,986	2,514
	Disk (TB)	143	256	328	650	1,103
TOTAL US Tier 2's						
	CPU (kSI2k)	3,348	4,947	6,367	7,681	9,982
	Disk (TB)	842	1,567	2,467	3,482	5,015

BNL Tier1	**2007**	**2008**	**2009**	**2010**	**2011**
CPU (kSI2K)	2560	4844	7337	12765	18193
Disk (Tbytes)	1100	3136	5822	11637	16509
Tape (Tbytes)	603	1715	3277	6286	9820

Data Management

- DQ2 is on critical path
  - Many performance issues have been identified through operations
  - Central server load issues – still problem after Oracle migration
  - Fetcher performance issues – incomplete datasets, QoS
  - Essential features needed soon: hierarchical (container) datasets, lost file flag, tape handling…
- We expect rapid improvements via new ADC organization
- Expect higher priority for production and DA needs – since Panda was chosen for ATLAS wide use
- Panda not using DQ2 for input file transfers – PandaMover
- Need to integrate PandaMover with DQ2
- Test and implement LFC in the U.S.
  - Support problems with LRC used in U.S. – diverging from DQ2

Slide from SLAC meeting in Nov. 2007

LFC Migration

- Ongoing work
- Need to develop a set of metrics for evaluation
- Need rollout schedule
- Still using LRC at all U.S. sites

Intelligent Data Placer

- Data arrives at BNL in mostly random order
- Copied to tape by order of arrival
- When data is accessed from tape, it is done few file at a time from many tapes – expensive operation
- ATLAS has built in data organization – datasets
- We should use this to place data on tape, broadly according to datasets
  - New software development
  - Cannot be strictly according to datasets, unless they are closed
  - Use project level hierarchy (MC data, RAW data, conditions data…)

Conclusion

- DDM is a huge area of work – on critical path
- Needs some U.S. specific development of tools
- The strategy is evolving
- We need much planning/discussions in upcoming weeks/months