U.S. ATLAS Computing Facilities Overview

Michael Ernst

Brookhaven National Laboratory

U.S. ATLAS Distributed Facility Meeting

SLAC

12-13 October 2010

Introduction to the week

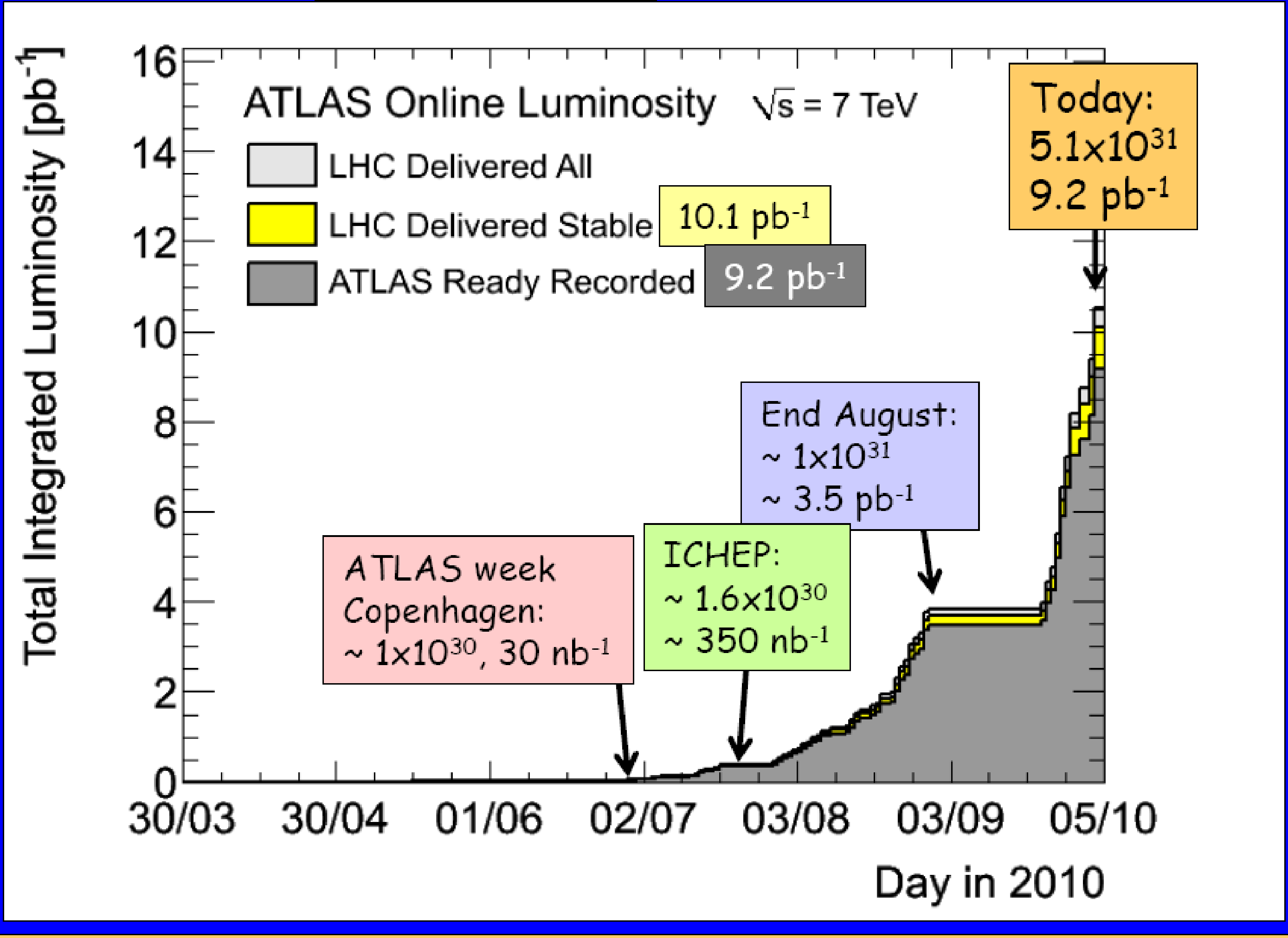
- Main achievements since last ATLAS Week and main concerns (examples ...)
- A walk through the agenda and the main goals of the week

~10 pb^{-1} !

F.Gianotti, ATLAS week, 4/10/2010

Data taking

→ see today's sessions



10.1 pb^{-1}

9.2 pb^{-1}

Today:
5.1×10^{31}
9.2 pb^{-1}

End August:
~ 1×10^{31}
~ 3.5 pb^{-1}

ICHEP:
~ 1.6×10^{30}
~ 350 nb^{-1}

ATLAS week
Copenhagen:
~ 1×10^{30}, 30 nb^{-1}

Excellent machine progression over last months (→ see J.Wenninger's talk)

❑ successful strategy: ~ 3 week-long machine commissioning (to prepare for next step in luminosity) alternated with physics periods

❑ now running with bunch trains, 8+8 bunches/train, 150 ns bunch spacing, up to 150 bunches

❑ record peak luminosity 5.1×10^{31}

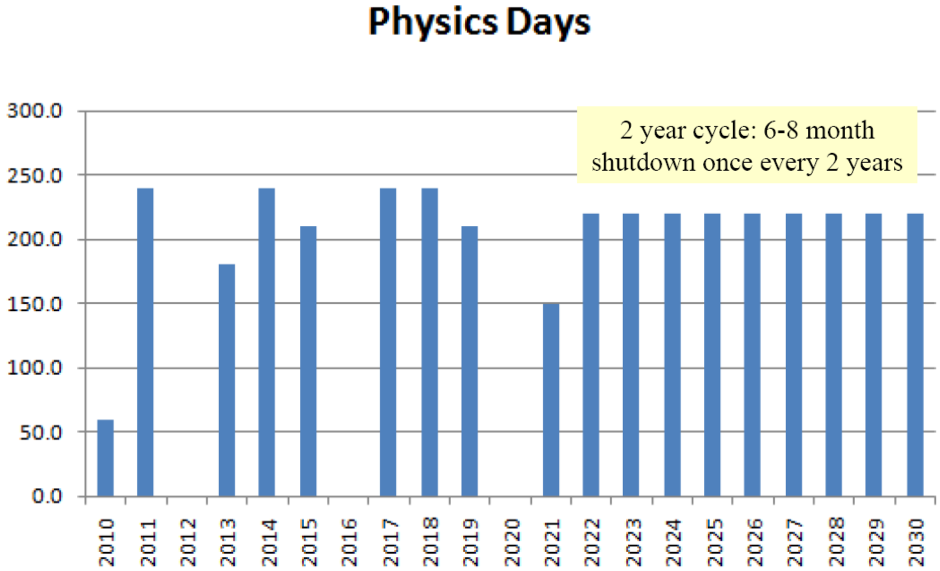
❑ stored beam energy: ~ 8.5 MJ

LHC Schedule

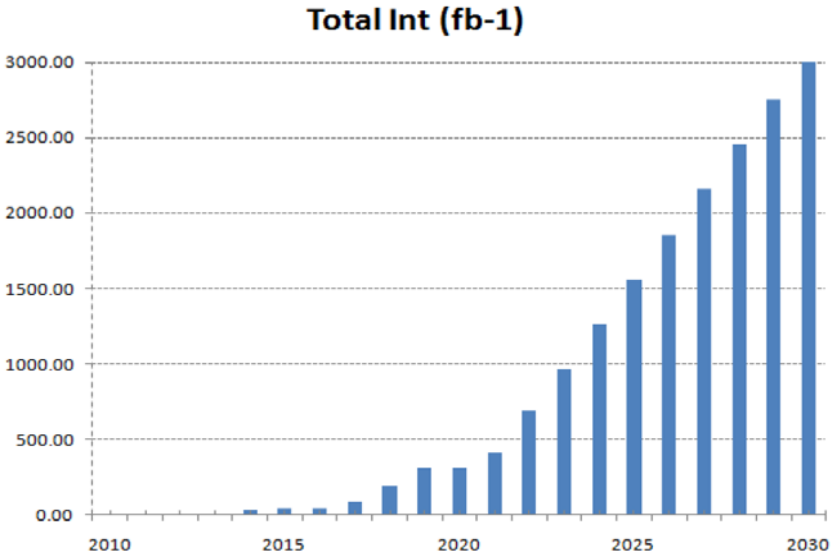
Current near term schedule

- ❑ 2010 (end): achieve 10^{32} running
- ❑ 2011 (end): goal is 1 fb^{-1} integrated
- ❑ 2012: >yearlong shutdown to carry out repairs to allow ~14 TeV
- ❑ 2013: start running at ~14 TeV
- ❑ 2016: shutdown
- ❑ 2020: shutdown

Steve Myers
ICHEP 2010



2 year cycle: 6-8 month shutdown once every 2 years

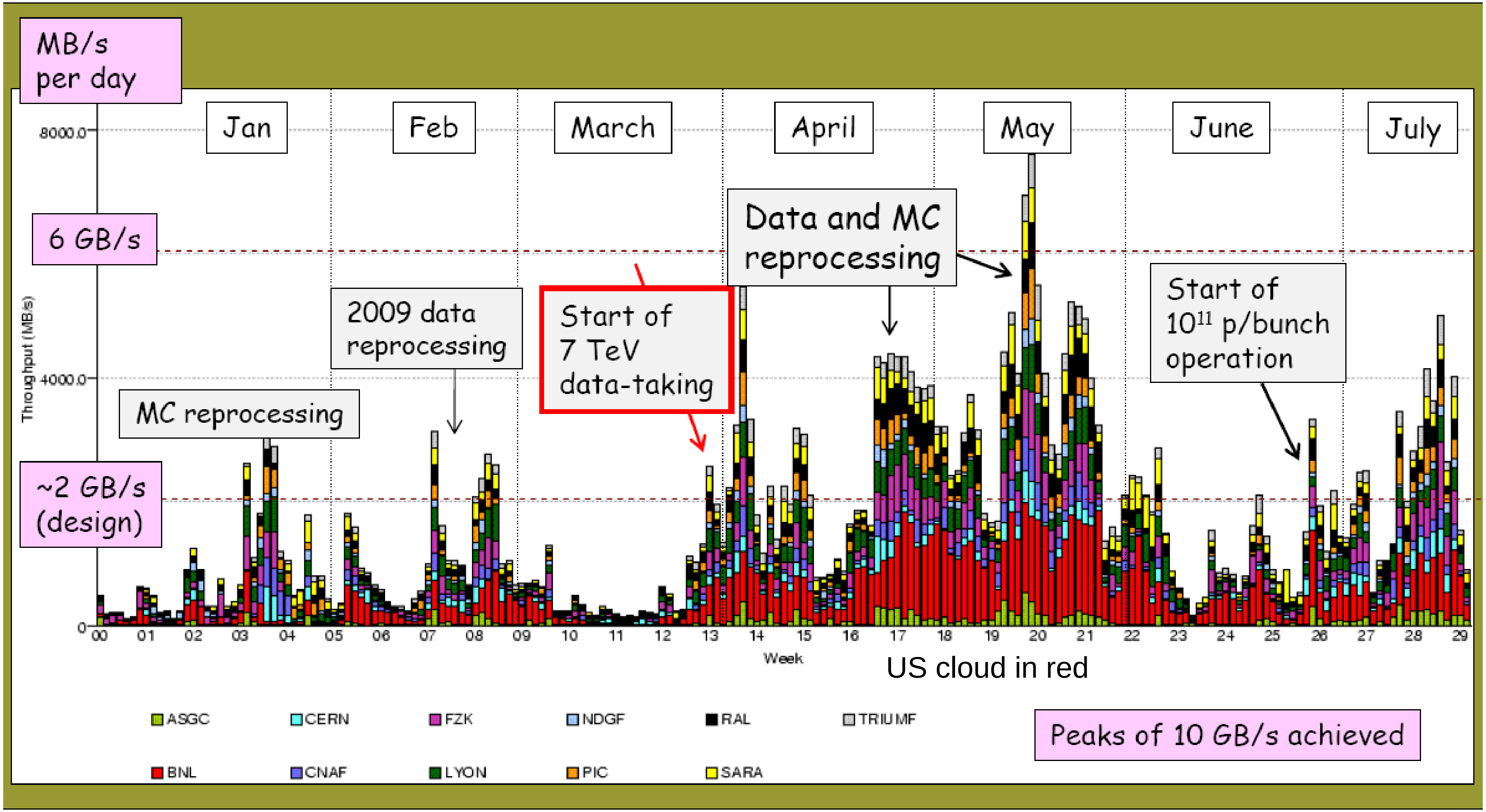


Computing Status

- T1, T2 Facilities and T3 Coordination
  - ❑ Facilities continue to perform best in ATLAS, but constrained to our MOU share
    - o High priority, must-deliver production work often preferentially sent to US
  - ❑ Utilization has far exceeded pre-startup scaling tests, but processing systems have held up very well
  - ❑ Analysis utilization consistently high; production usage more variable
    - o We flexibly shift resources from production to analysis (and back) to maximize utilization
  - ❑ Facility cost/benefit analysis and proposal for Tier 2 funding 2012-16 underway
  - ❑ Exponentially increasing space consumption brought under control with Panda usage-driven dynamic brokering; space usage now scales
  - ❑ Tier 3 on strong growth curve; new ARRA money/hardware is now arriving, purchase/setup recommendations and help are ready. ~20 new T3s coming
    - o Will hear a lot more about Tier-3s from Doug et al later this morning
- 2010 CPU, disk pledges met at T1 and T2s
- 2011, 2012 (est.) pledges delivered on time, prior to October RRB
  - ❑ Based on 23% US share

Worldwide data distribution and analysis

Total throughput of ATLAS data through the Grid: from 1st January until yesterday



6 GB/s

~2 GB/s (design)

MC reprocessing

2009 data reprocessing

Start of 7 TeV data-taking

Data and MC reprocessing

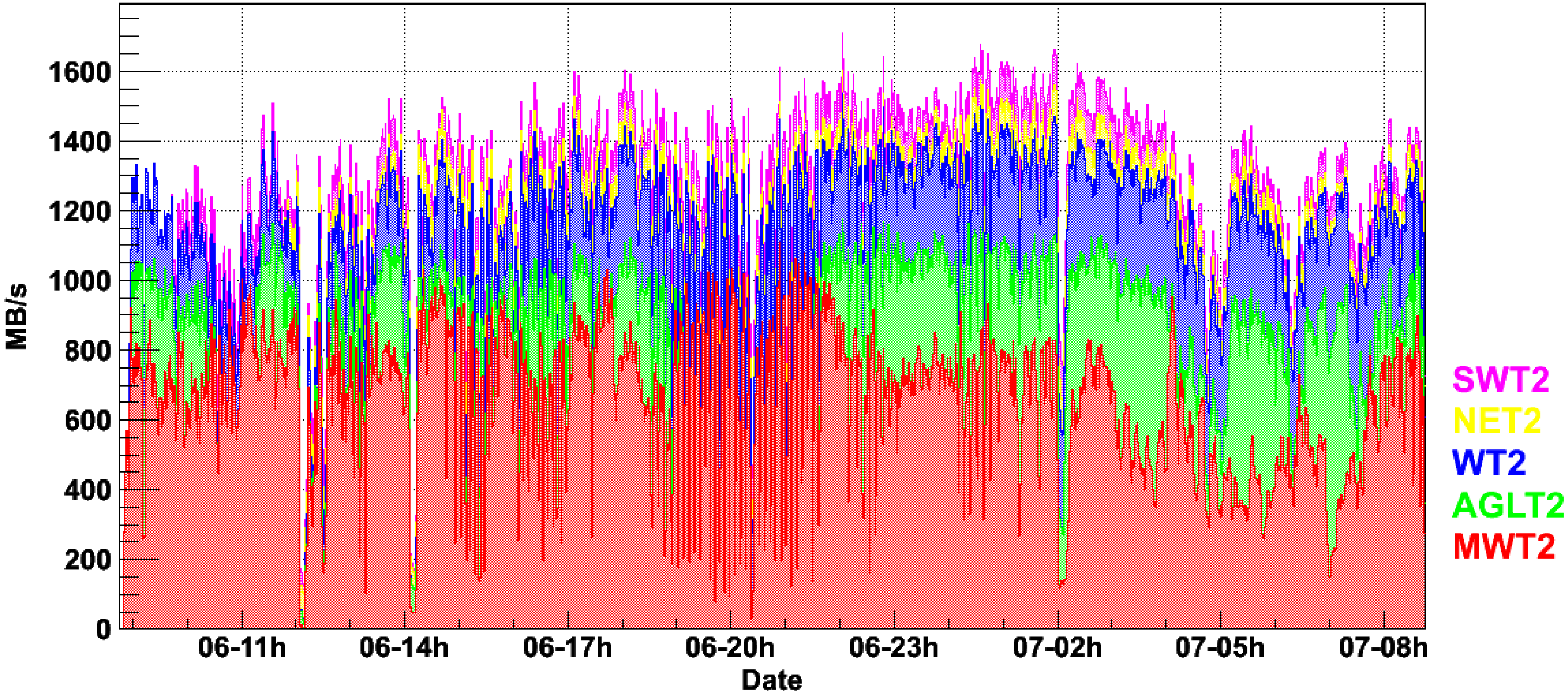
Start of 10^{11} p/bunch operation

US cloud in red

Peaks of 10 GB/s achieved

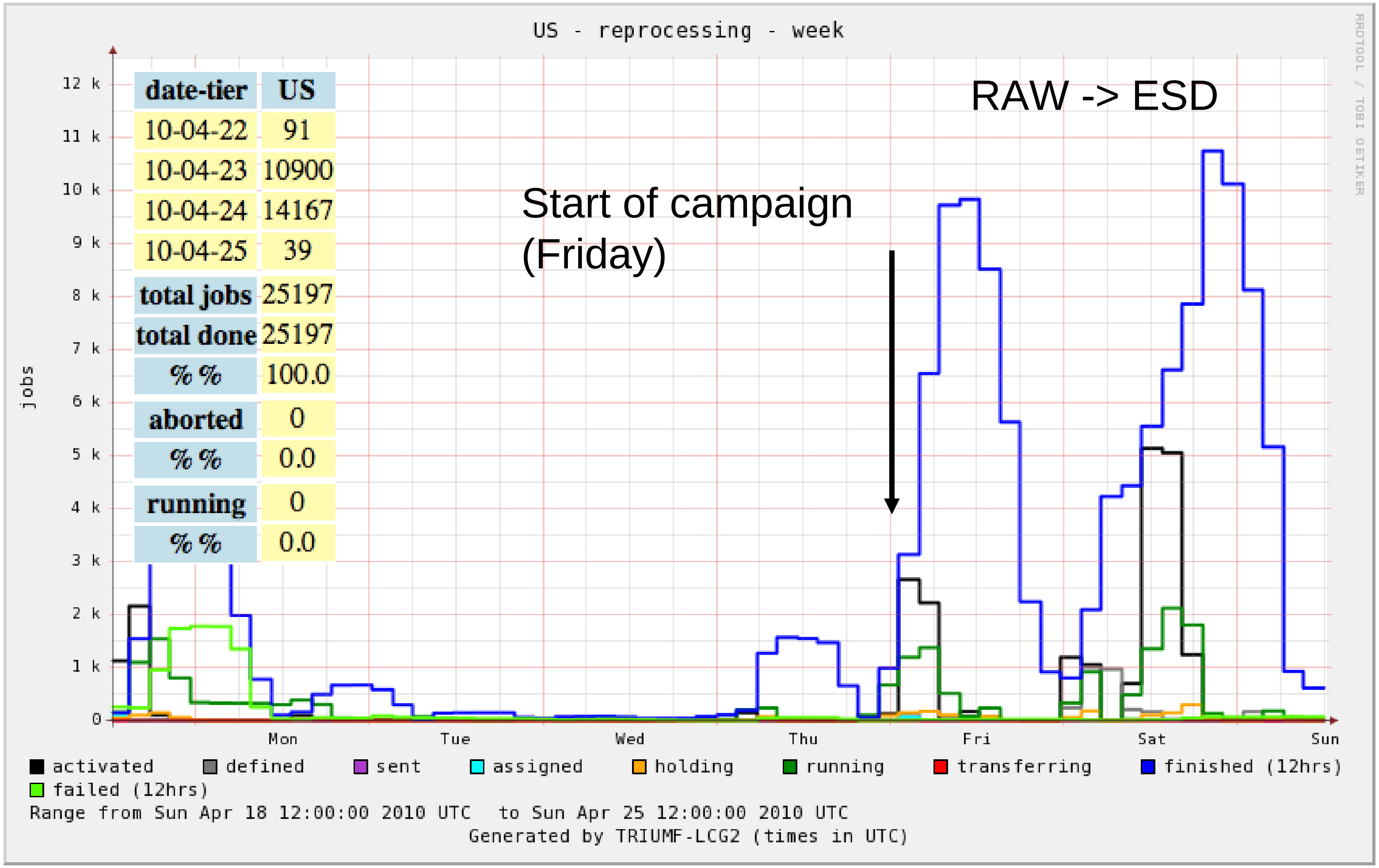
GRID-based analysis in June-July 2010:
> 1000 different users, ~ 11 million analysis jobs processed

U.S. Tier-1 to Tier-2 Replication



- Long delays not acceptable for Users and not sustainable from a technical perspective
- Observation: Only a fraction of the datasets is needed

Reprocessing (RAW->ESD->AOD-> HIST) – The easy part



date-tier	US
10-04-22	91
10-04-23	10900
10-04-24	14167
10-04-25	39
total jobs	25197
total done	25197
% %	100.0
aborted	0
% %	0.0
running	0
% %	0.0

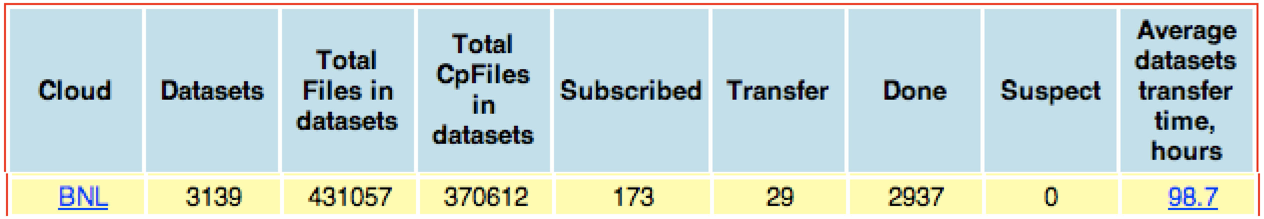
Range from Sun Apr 18 12:00:00 2010 UTC to Sun Apr 25 12:00:00 2010 UTC

Generated by TRIUMF-LCG2 (times in UTC)

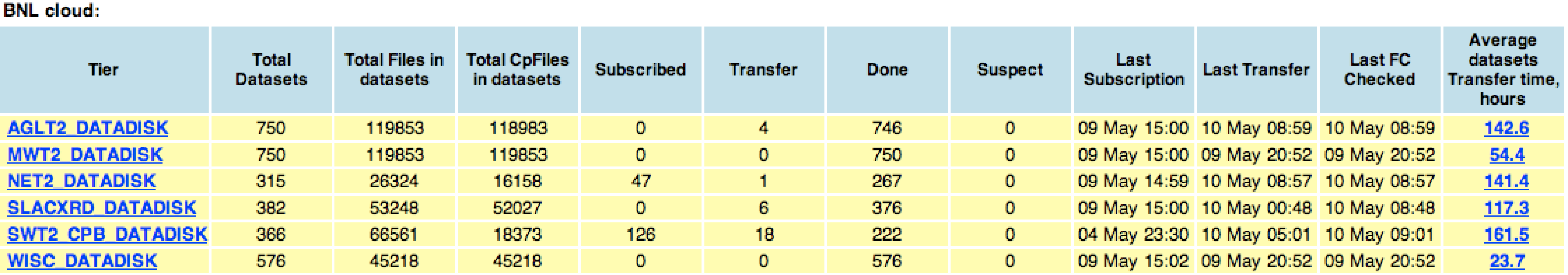
RAW -> ESD

Start of campaign (Friday)

Tier-1 -> Tier-2 – The hard part

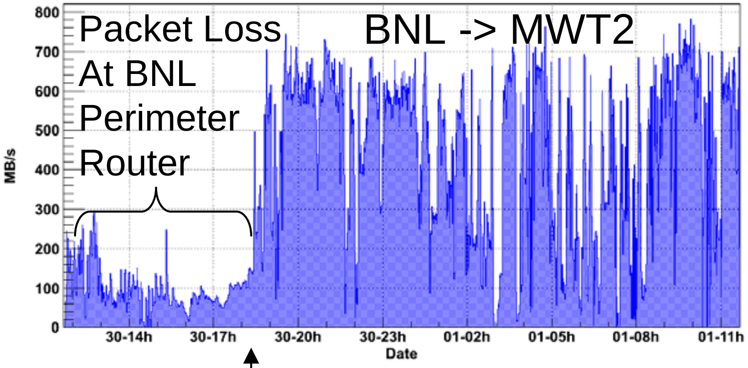


Cloud	Datasets	Total Files in datasets	Total CpFiles in datasets	Subscribed	Transfer	Done	Suspect	Average datasets transfer time, hours
BNL	3139	431057	370612	173	29	2937	0	98.7



BNL cloud:

Tier	Total Datasets	Total Files in datasets	Total CpFiles in datasets	Subscribed	Transfer	Done	Suspect	Last Subscription	Last Transfer	Last FC Checked	Average datasets Transfer time, hours
AGLT2_DATADISK	750	119853	118983	0	4	746	0	09 May 15:00	10 May 08:59	10 May 08:59	142.6
MWT2_DATADISK	750	119853	119853	0	0	750	0	09 May 15:00	09 May 20:52	09 May 20:52	54.4
NET2_DATADISK	315	26324	16158	47	1	267	0	09 May 14:59	10 May 08:57	10 May 08:57	141.4
SLACXRD_DATADISK	382	53248	52027	0	6	376	0	09 May 15:00	10 May 00:48	10 May 08:48	117.3
SWT2_CPB_DATADISK	366	66561	18373	126	18	222	0	04 May 23:30	10 May 05:01	10 May 09:01	161.5
WISC_DATADISK	576	45218	45218	0	0	576	0	09 May 15:02	09 May 20:52	09 May 20:52	23.7

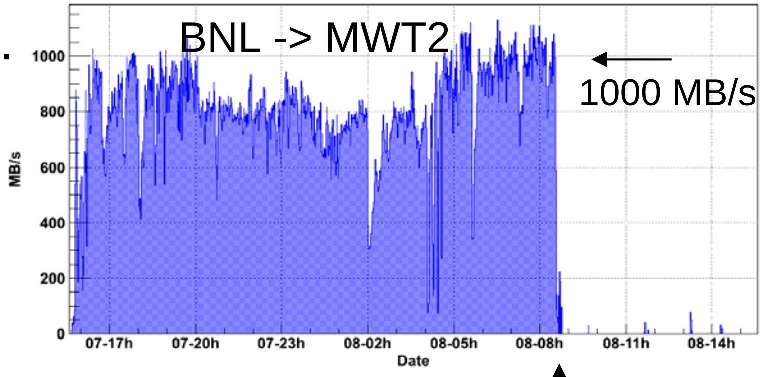


Packet Loss At BNL Perimeter Router

BNL -> MWT2

Resolved Fri, April 30

…

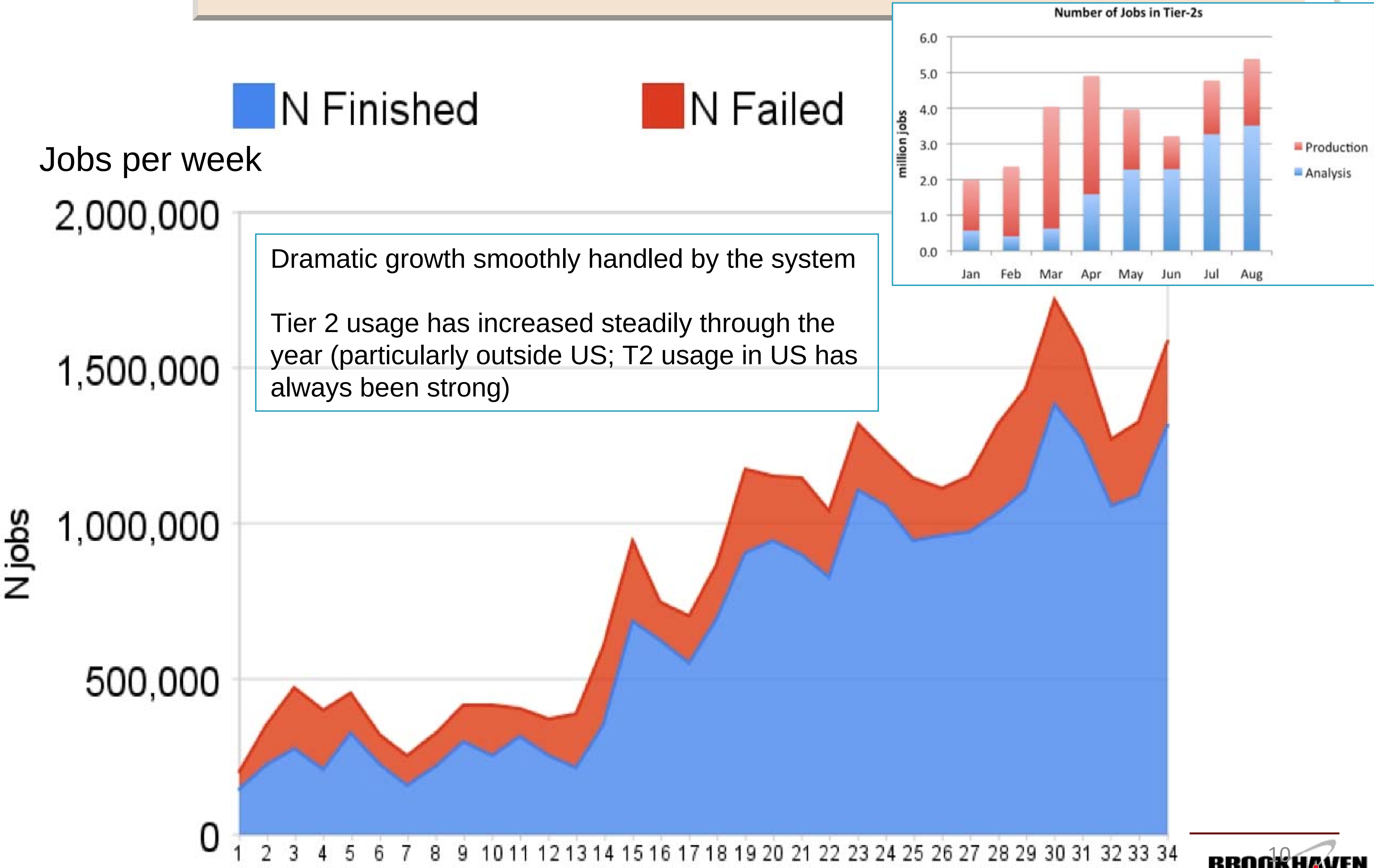


BNL -> MWT2

1000 MB/s

Sat, May 8 MWT2 received all Datasets subscribed April 27

Worldwide Panda Analysis 2010



Dramatic growth smoothly handled by the system

Tier 2 usage has increased steadily through the year (particularly outside US; T2 usage in US has always been strong)

What is PD2P

- Dynamic data placement at Tier 2's
  - Continue automatic distribution to Tier 1's – treat them as repositories
  - Reduce automatic data subscriptions to Tier 2's – instead use PD2P
- The plan
  - Panda will subscribe a dataset to a Tier 2, if no other copies are available (except at a Tier 1), as soon as any user needs the dataset
    - User jobs will still go to Tier 1 while data is being transferred – no delay
  - Panda will subscribe replicas to additional Tier 2's, if needed, based on backlog of jobs using the dataset (PanDA checks continuously)
  - Cleanup will be done by central DDM popularity based cleaning service (as described in previous talk by Stephane)
- Few caveats
  - Start with DATADISK and MCDISK
  - Exclude RAW, RDO and HITS datasets from PD2P
  - Restrict transfers within cloud for now
  - Do not add sites too small (storage mainly) or too slow

Main Goals

- User jobs should not experience delay due to data movement
- First dataset replication is 'request' based
  - Any user request to run jobs will trigger replication to a Tier 2 chosen by PanDA brokering – no matter how small or large the request
- Additional dataset replication is 'usage' based
  - Send replicas to more Tier 2's if a threshold is crossed (many jobs are waiting for the dataset)
- Types of datasets replication are 'policy' based
  - We follow Computing Model – RAW, RDO, HITS are never replicated to Tier 2's by PanDA (we may have more complex rules later, to allow for small fraction of these types to be replicated)
  - PanDA does replication only to DATADISK and MCDISK, for now
- Replication pattern is 'cloud' based
  - Even though subscription source is not specified, currently PanDA will only initiate replication if source is available within cloud (we hope to relax this in the next phase of tests)

Computing Status (2)

- We have some CPU capacity beyond pledge; mechanism deployed and under test at AGLT2 to dedicate the excess to US users
  - We have long planned (and been advised to establish) a US-specific fraction; for the first time we have the resources
  - Will be above pledge ~25% CPU, ~15% disk in FY11, roughly where we wanted to be in terms of a US-dedicated fraction
    - Thanks in large measure to local university contributions to Tier 2s
- Aspects of ATLAS computing model are being re-examined, with ATLAS moving in directions spearheaded by the US
  - More flexible utilization of the Tier 1 for analysis
  - More flexible data distribution policy to the Tier 2s (eg. ESDs)
  - Dynamically cached data based on usage, rather than predefined distribution policies
- US ATLAS a strong (and very supportive) player in “OSG prime” planning

Tier 2 Planning

- Current Tier-2 grant period ends 31 January, 2012
- US ATLAS Computing Management conducts cost/benefit analysis across US facilities, which is in progress since several weeks
- Cost breakdowns are being provided by the Tier 2s based on standard costing templates
- Analysis includes assessment of subjective factors such as the Tier-2 teams’ expertise and experience
- Input to Tier 2 planning: proposal in preparation for 2012-2016 Tier 2 funding
  - ❑ In the context of NSF cooperative agreement

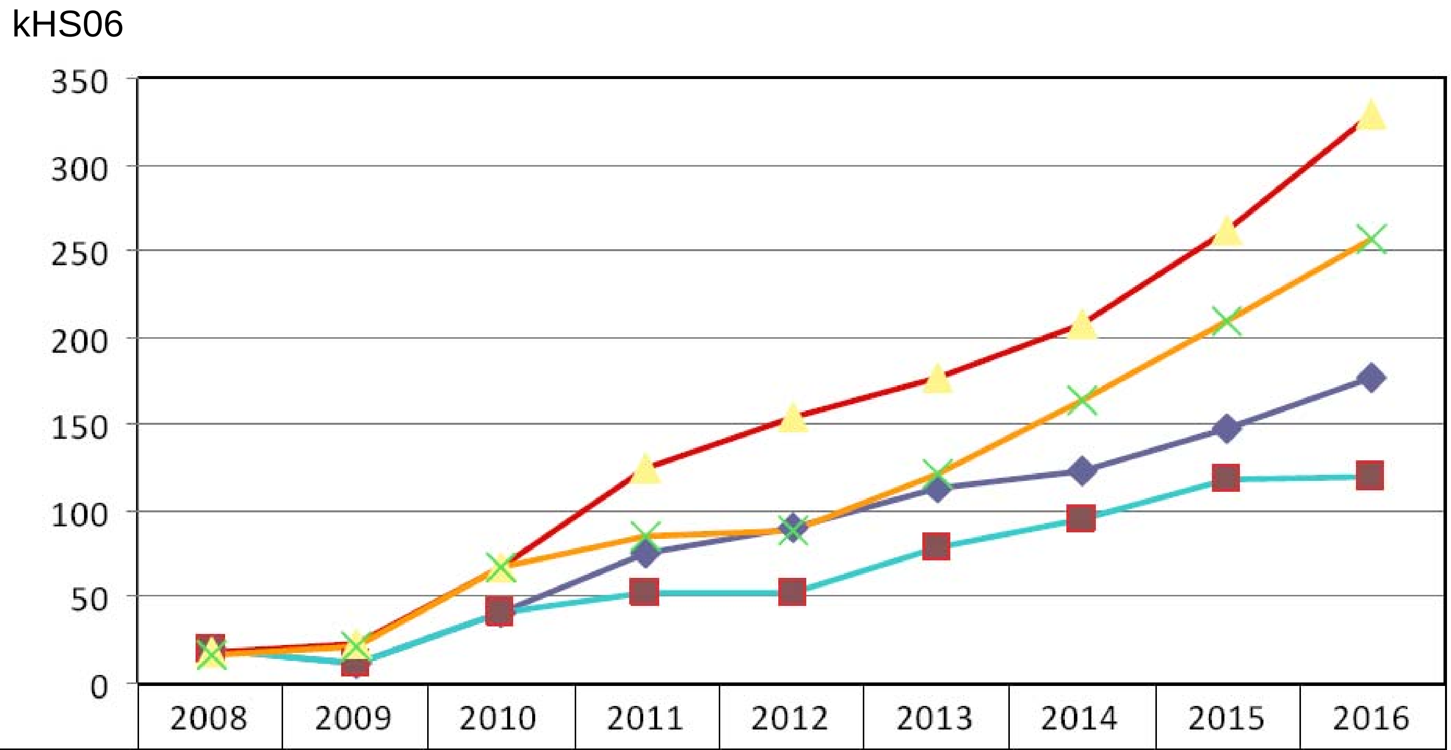
Tier 2 Institutional Involvement

- Torre has sent a letter to current Tier 2s asking for their plans for the next 5 years with respect to Tier 2 involvement
  - ❑ What will they bring to a new funding cycle, what changes can we expect
- US ATLAS solicited via the IB in 2 monthly meetings for expressions of interest from universities who may wish to become involved as Tier 2 sites
- Three heard from: Illinois Urbana/Champagne, UT Dallas, SMU
- Will integrate them into cost/benefit analysis to come to decisions on involvement
  - ❑ Factoring in what local resources and capabilities they bring
- Any new involvement will come through new consortium members in existing T2s, not new T2s
  - ❑ Avoid inflating fixed costs (eg. support manpower); share and leverage existing resources and expertise
  - ❑ Institutes expressing interest all have a natural regional T2 association (but doesn't exclude a different association)

Tier 2 Proposal 2012-2016

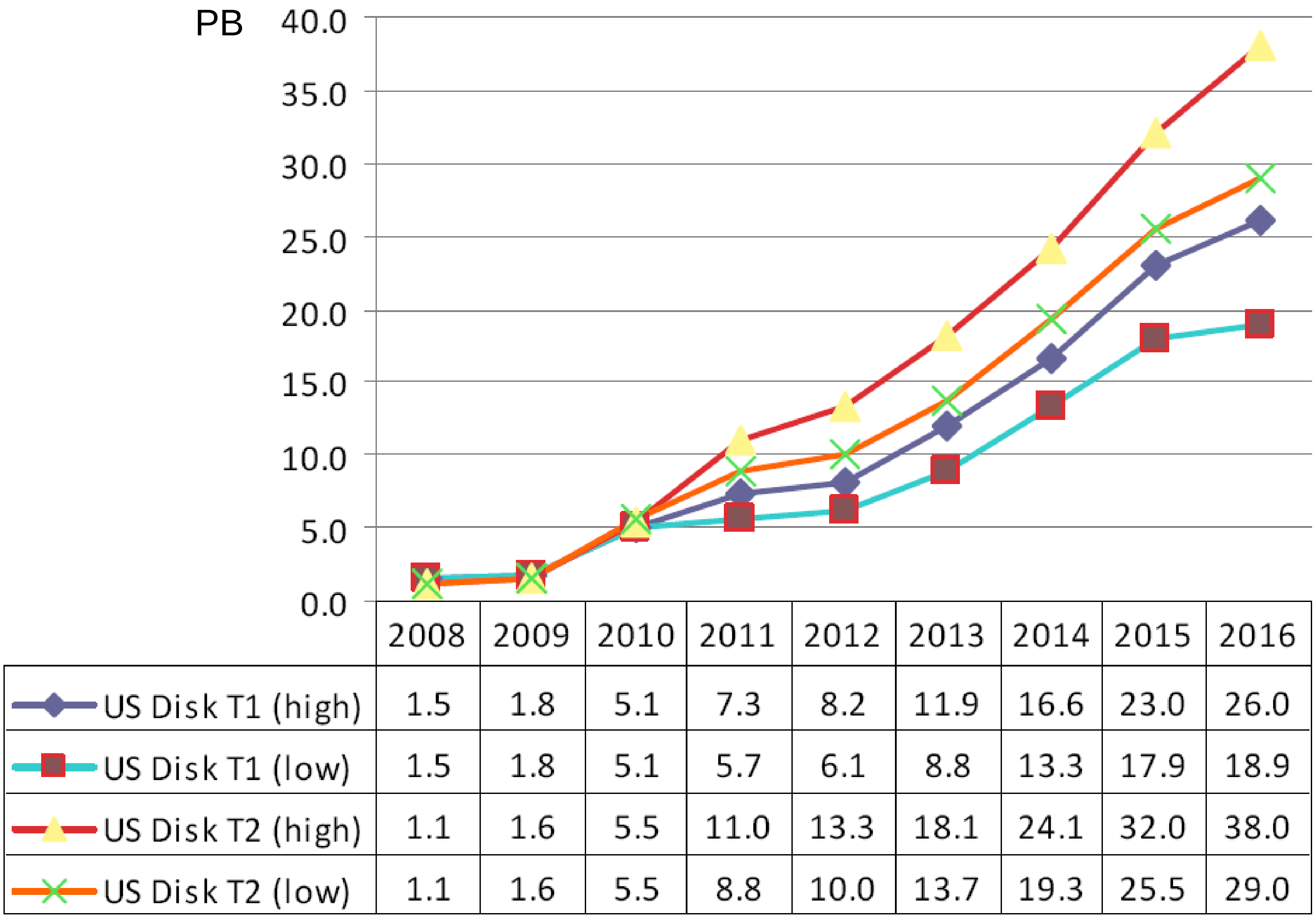
- Target date for completion of draft for US ATLAS review is Nov 1 (!)
  - ❑ Until recently we thought we had more time than that, so October is busy
- Not an impossible date with the start we've made, but there are external dependencies (the Tier 2 PIs have to interact with their universities for example)
- Objective for the proposal will not be to enumerate the detailed breakdown of resources between Tier 2s and their members, but rather
  - ❑ Describe the capabilities and resources Tier-2s anticipate making available to ATLAS
  - ❑ Describe the cost/benefit analyses that will guide the resource distribution
  - ❑ Resource distribution will be at least potentially dynamic over the 5 years as conditions evolve

CPU in the U.S. (75% MC at Tier-2s)



	2008	2009	2010	2011	2012	2013	2014	2015	2016
CPU T1 (high)	19	12	41	75	89	113	123	147	176
CPU T1 (low)	19	12	41	52	52	78	94	117	120
CPU T2 (high)	17	22	67	124	153	177	208	262	329
CPU T2 (low)	17	22	67	85	89	121	163	210	257

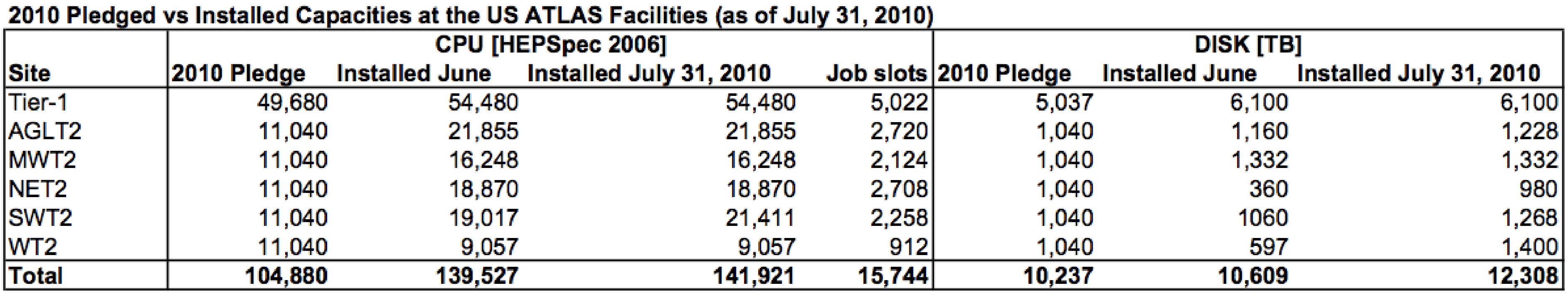
Disk in the U.S.



	2008	2009	2010	2011	2012	2013	2014	2015	2016
◆ US Disk T1 (high)	1.5	1.8	5.1	7.3	8.2	11.9	16.6	23.0	26.0
■ US Disk T1 (low)	1.5	1.8	5.1	5.7	6.1	8.8	13.3	17.9	18.9
▲ US Disk T2 (high)	1.1	1.6	5.5	11.0	13.3	18.1	24.1	32.0	38.0
✕ US Disk T2 (low)	1.1	1.6	5.5	8.8	10.0	13.7	19.3	25.5	29.0

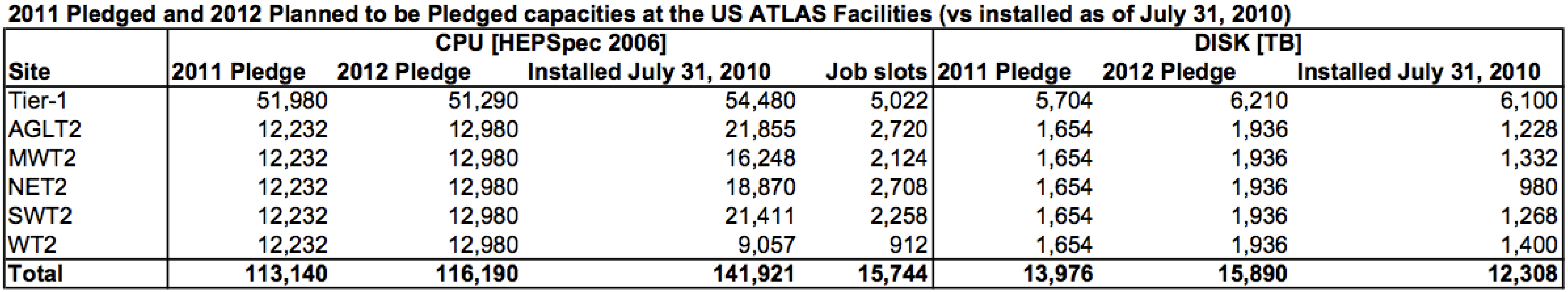
US ATLAS Computing Pledge Status

- 2010 pledges fulfilled
- 2011, 2012 pledges to be submitted as below
- More CPU in the US than pledged



2010 Pledged vs Installed Capacities at the US ATLAS Facilities (as of July 31, 2010)

	CPU [HEPSpec 2006]				DISK [TB]		
Site	**2010 Pledge**	**Installed June**	**Installed July 31, 2010**	**Job slots**	**2010 Pledge**	**Installed June**	**Installed July 31, 2010**
Tier-1	49,680	54,480	54,480	5,022	5,037	6,100	6,100
AGLT2	11,040	21,855	21,855	2,720	1,040	1,160	1,228
MWT2	11,040	16,248	16,248	2,124	1,040	1,332	1,332
NET2	11,040	18,870	18,870	2,708	1,040	360	980
SWT2	11,040	19,017	21,411	2,258	1,040	1060	1,268
WT2	11,040	9,057	9,057	912	1,040	597	1,400
Total	**104,880**	**139,527**	**141,921**	**15,744**	**10,237**	**10,609**	**12,308**



2011 Pledged and 2012 Planned to be Pledged capacities at the US ATLAS Facilities (vs installed as of July 31, 2010)

	CPU [HEPSpec 2006]				DISK [TB]		
Site	**2011 Pledge**	**2012 Pledge**	**Installed July 31, 2010**	**Job slots**	**2011 Pledge**	**2012 Pledge**	**Installed July 31, 2010**
Tier-1	51,980	51,290	54,480	5,022	5,704	6,210	6,100
AGLT2	12,232	12,980	21,855	2,720	1,654	1,936	1,228
MWT2	12,232	12,980	16,248	2,124	1,654	1,936	1,332
NET2	12,232	12,980	18,870	2,708	1,654	1,936	980
SWT2	12,232	12,980	21,411	2,258	1,654	1,936	1,268
WT2	12,232	12,980	9,057	912	1,654	1,936	1,400
Total	**113,140**	**116,190**	**141,921**	**15,744**	**13,976**	**15,890**	**12,308**

Reserving Beyond-Pledge Resources

- US ATLAS CPU resources at T1, T2 now exceed the ATLAS pledges, enabling us to reserve some resources for US use, as long planned
- Mechanism supporting this is implemented & under test
  - ❑ PanDA DB records pledged and available resource levels
  - ❑ Where available exceeds pledge level by X%, PanDA job dispatcher assigns a US job exclusively X% of the time
  - ❑ Resource thus is ‘virtually partitioned’ as an equitably shared resource of pledge-level size, plus a US piece X
  - ❑ If US jobs are insufficient to fill X, US-only constraint is dropped
  - ❑ No hard partitions, no waste

Fall Reprocessing

- 7 TeV runs, stable beam, ATLAS ready
  - Also 900 GeV data
- Software release 16.0.x.x
- Sep 24: Conditions data deadline
- Oct 1: Release physics validation completed
- Oct 5-11: Express stream repro (~50M events)
- Oct 25-Nov 15: Bulk repro campaign (~900M events by Oct 20)
- Nov 29: End of repro, data distribution done
  - Adequate analysis time before La Thuile approval deadlines
- Associated Geant4 simulation campaign already underway

HI Run

- Rate will be limited to 320 MB/sec
  - Means trigger rate of ~60 Hz (RAW=5MB)
- At T0: only express line reconstruction
  - And normal calibration loop
- RAW data exported to all T1's
  - Bulk reconstruction done at T1's
- Only ESD (and TAG) output from reconstruction
  - ESD is bigger: ~3 MB
- Distribute to 3 T2s GROUP space
  - BNL, Krakow, Israel
  - Analysis based on MinBiasD3PD

Summary

- The facilities in the U.S., the Tier-1 and the Tier-2's, have performed well in ATLAS computer system commissioning and specific exercises
  - Production and Analysis Operations Coordination provides seamless integration with ATLAS world-wide computing operations
  - The Integration Program is instrumental to ensure readiness in view of the steep ramp-up of the resources and the need to properly integrate end-user analysis facilities (Tier-3s)
  - Excellent contribution of U.S ATLAS Tier-2 Sites to high volume production (event simulation, reprocessing) and analysis
  - Steep ramp-up of in particular disk resources during LHC run needs special attention
- Overall, the Facilities are prepared for LHC data analysis …
  - … though there is still a lot to be done