U.S. ATLAS Computing Facilities Overview
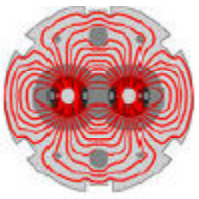
Michael Ernst

Brookhaven National Laboratory
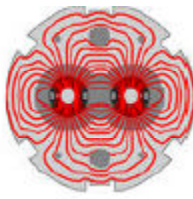
U.S. ATLAS Facility Meeting

University of Texas at Arlington
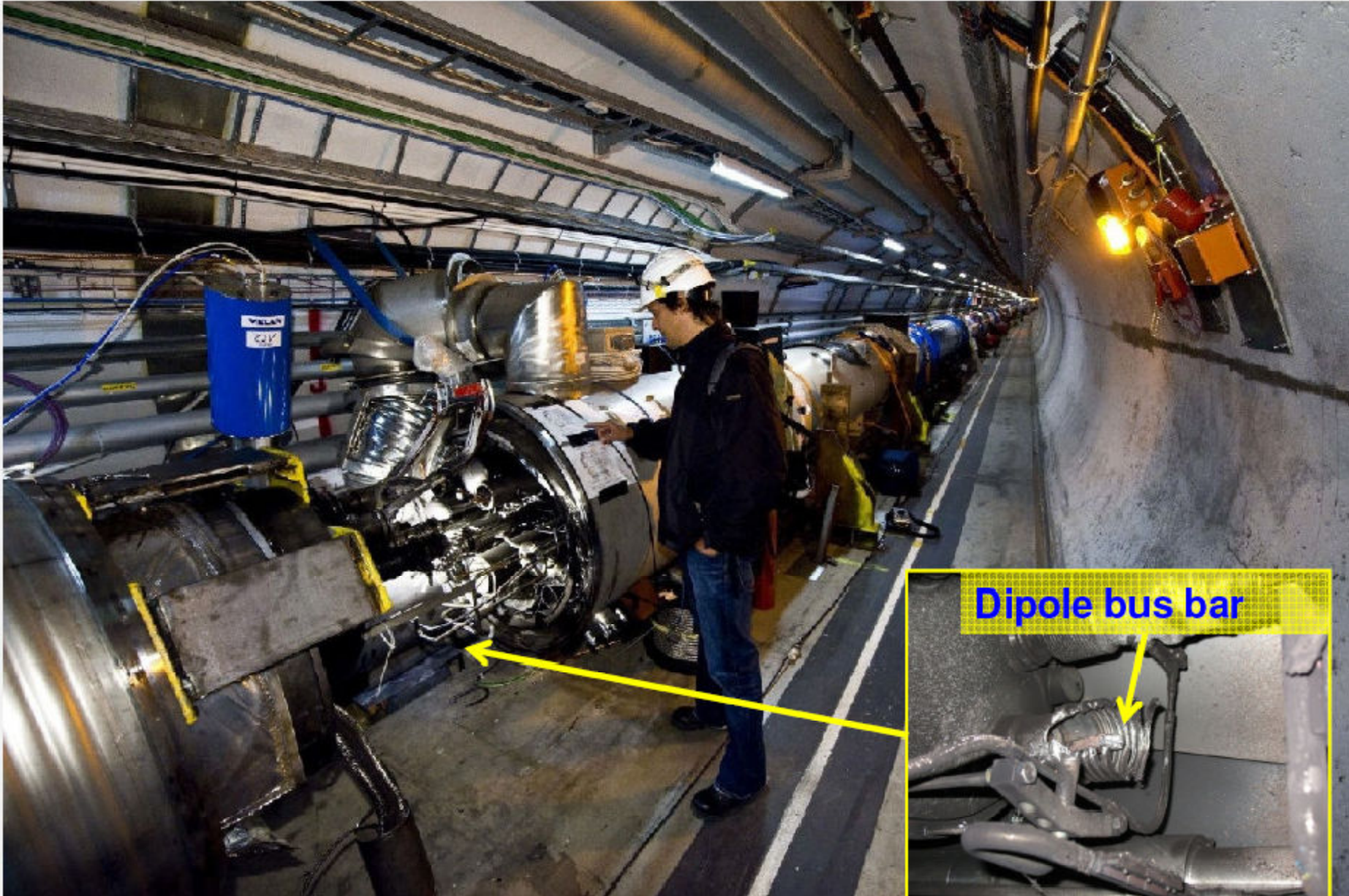
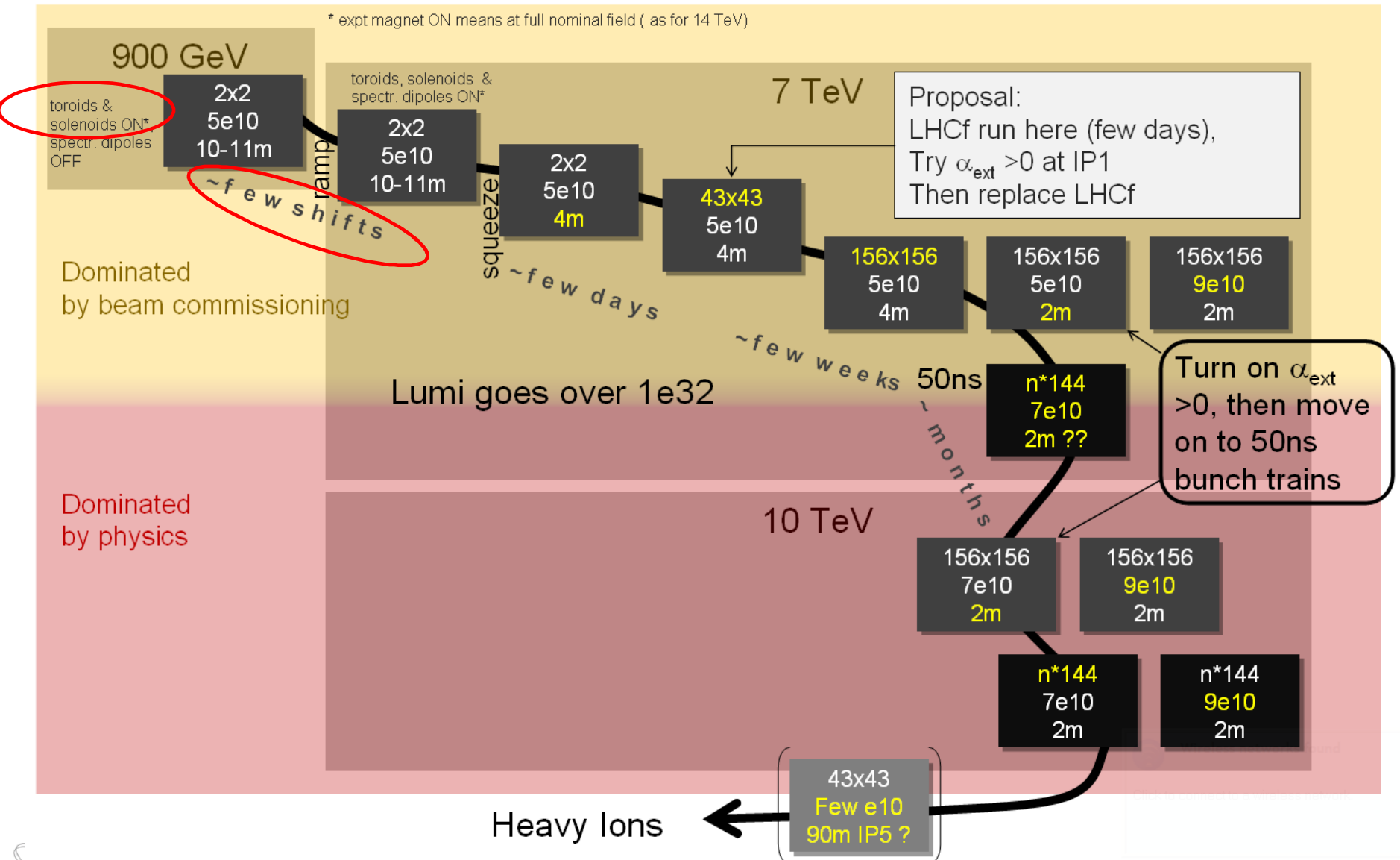10 – 12 November 2009

Status of the LHC - Evian - J. Wenninger

19th Oct. 2009

Dipole bus bar

* expt magnet ON means at full nominal field (as for 14 TeV)

900 GeV

toroids & solenoids ON* spectr. dipoles OFF

2x2
5e10
10-11m

toroids, solenoids & spectr. dipoles ON*

2x2
5e10
10-11m

7 TeV

ramp

squeeze

~few shifts

~few days

~few weeks

~months

2x2
5e10
4m

43x43
5e10
4m

Proposal:
LHCf run here (few days),
Try α_{ext} >0 at IP1
Then replace LHCf

156x156
5e10
4m

156x156
5e10
2m

156x156
9e10
2m

Dominated
by beam commissioning

Lumi goes over 1e32

50ns

n*144
7e10
2m ??

Turn on α_{ext}
>0, then move
on to 50ns
bunch trains

Dominated
by physics

10 TeV

156x156
7e10
2m

156x156
9e10
2m

n*144
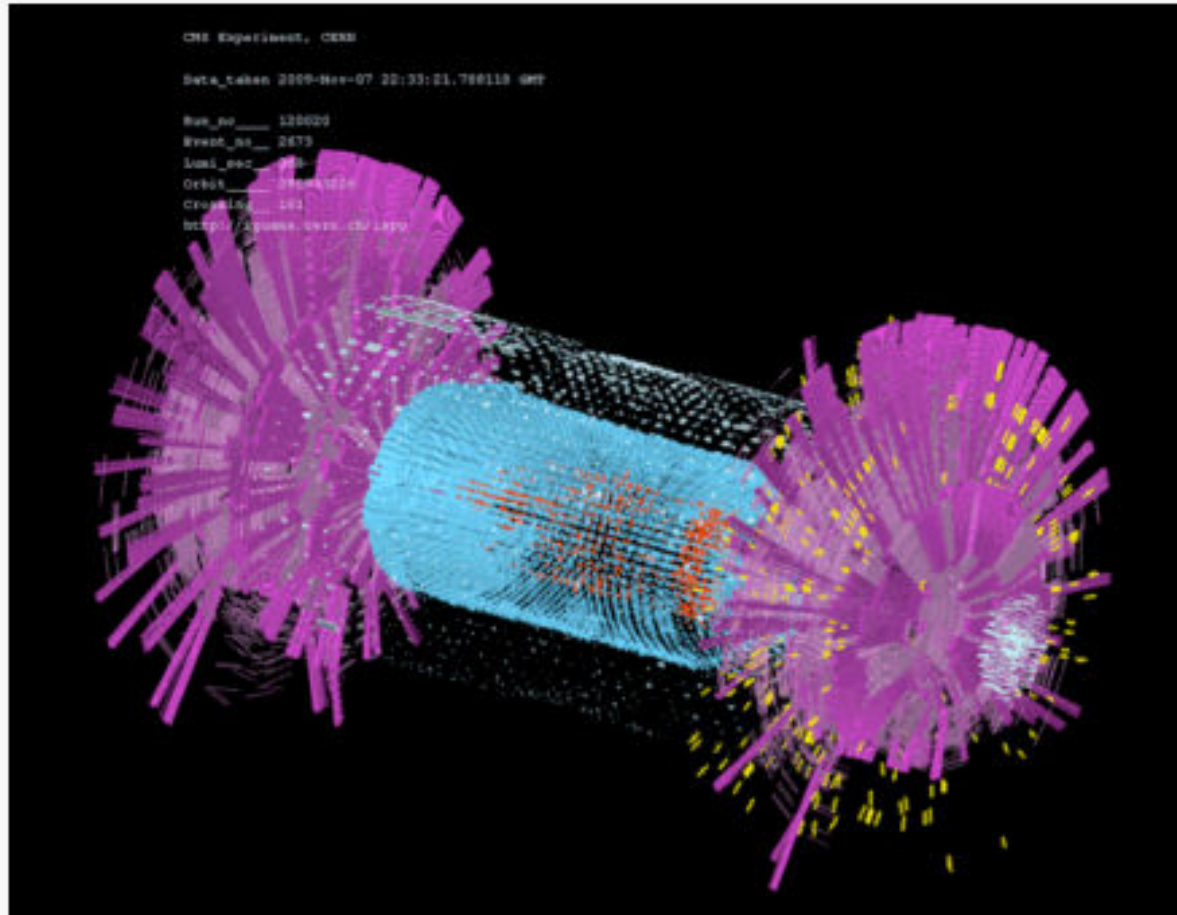7e10
2m

n*144
9e10
2m

Heavy Ions

43x43
Few e10
90m IP5 ?

❑ completing HW cmg to 6 kA (3.5 TeV) turns out to be icompatible with making collisions this year

❑ complete HW cmg to 2 kA (1.1 TeV) by ~16 Nov and make beams
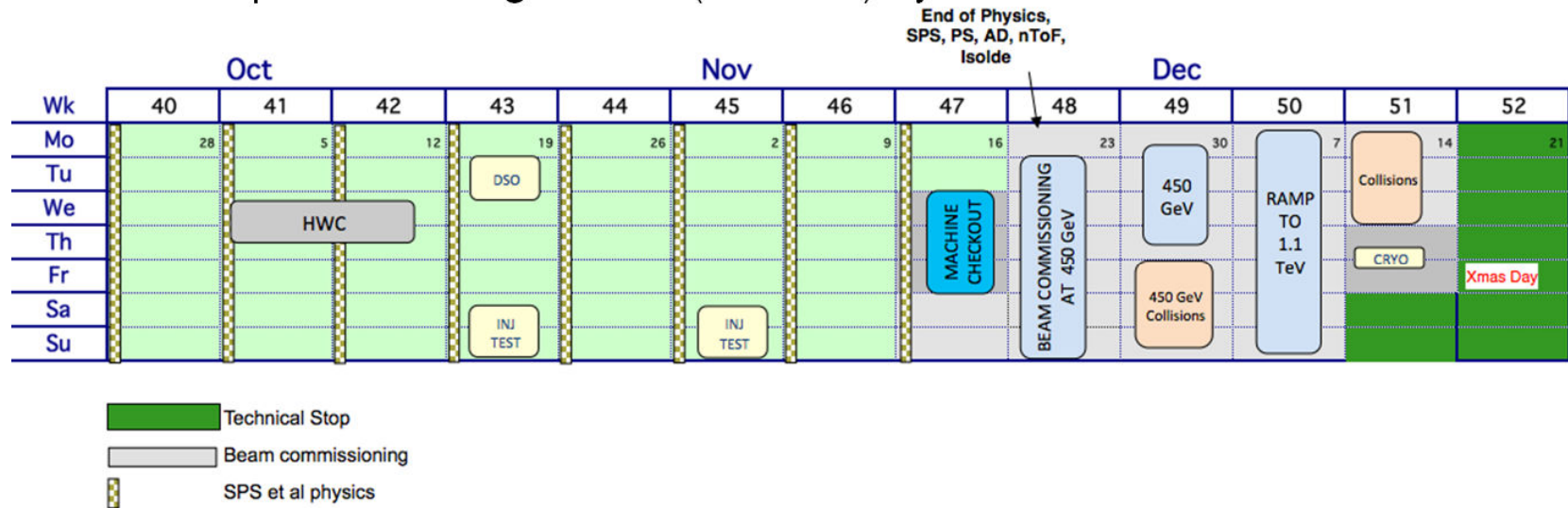
❑ how to move up to 3.5 TeV (after Xmas break) is being addressed

Particles have gone half way round the LHC

Splash event recorded by the CMS experiment on 7 November. The electromagnetic calorimeter is in red, the hadronic calorimeter in blue, the muon system is yellow and magenta. The barrel muon detector was on standby and the inner tracking detector was off.

Computing Facility: Transition to Stable Services at Data Taking Rate

- Focus now must be physics analysis performance
  - Which means integration, deployment, testing, documentation

- Stimulation will come from the physics!
  - Physicists are in the process of learning how to analyze data

But still many 'big unsolved problems':

- How can we store data more efficiently?
  - Number of complete copies vastly reduced
  - Disk capacity at Tier-1s was reduced by ~50%
  - Disk Capacity at Tier-2s has to ramp up much faster than we had planned

- Lots of ATLAS Software issues
  - Timely readiness of **validated** S/W Releases
  - Multi-core issues being addressed but when will it be production ready?
  - Inefficiencies require Facilities to provide more resources

- How should we use virtualisation – a big issue (not only) for Tier-3s?

- What analysis tasks can really be made interactive, and which are desirable?

- Is Connectivity according to the Hierarchy as defined by the Computing Model sufficient?

	kHS06 / PB / PB	OLD				NEW							
		Pledge 2008 (C-RRB Fall 2007)	Pledge 2009 (C-RRB Fall 2008)	Req. 2009 (C-RRB Fall 2008)	Req. 2010 (C-RRB Fall 2008)	Requirement 2009-Q2	Requirement 2009-Q3	Requirement 2009-Q4	Requirement 2010-Q1	Requirement 2010-Q2	Requirement 2010-Q3	Requirement 2010-Q4	Requirement 2011-Q1
Tier-0	CPU	14.1	16.2	30.4	30.4	12.7	18	30.4	30.4	30.4	30.4	30.4	30.4
	Disk	0.15	0.26	0.65	0.7	0.41	0.7	0.7	0.7	0.7	0.7	0.7	0.7
	Tape	2.8	6.2	9.7	14.6	3.8	5.1	6.4	7.8	8.9	9.9	9.9	9.9
CAF	CPU	8.6	10.2	23.2	37.4	7.6	14.2	20.8	27.4	37.4	37.4	37.4	37.4
	Disk	1.1	1.8	3.3	4.5	0.76	2	2.5	2.9	3.4	3.9	4.2	4.4
Tier-1	CPU	76.8	120.9	113.7	234.5	88.8	90	90	90	177	177	227	227
	Disk	10.9	19.9	20.9	41.3	7	19.6	21.7	23.8	27.5	32	34.3	36.7
	Tape	7.7	14.7	15.8	22.7	8.1	9.3	10.3	11.3	12.5	13.7	14.2	14.8
Tier-2	CPU	70.1	107.7	108	242.5	108	108	108	108	240	240	240	240
	Disk	6.3	10.7	13.3	24.8	5.8	9.2	11.4	13.3	17.5	21.5	23.1	24.8

Total ATLAS Resource Requirements

Tier-1	CPU	17.7	27.8	26.2	53.9	20.4	20.7	20.7	20.7	40.7	40.7	52.2	52.2
	Disk	2.5	4.6	4.8	9.5	1.6	4.5	5.0	5.0	6.3	6.3	7.9	7.9
	Tape	1.8	3.4	3.6	5.2	5.2	1.9	2.1	2.4	2.6	2.9	3.3	3.4
Tier-2	CPU	16.1	24.8	24.9	55.8	24.8	24.8	24.8	24.8	55.2	55.2	55.2	55.2
	Disk	-1.5	-2.5	-3.1	-5.7	-1.6	-2.1	-2.6	-3.1	-4.0	-4.9	-5.3	5.7

U.S. share derived from total ATLAS Resource Requirements (U.S. share is 23%)

EACH	CPU	[kHS-06]				4.96	4.96	4.96	11.04	11.04	11.04	11.04
Tier-2	Disk	[PB]				0.42	0.52	0.62	0.80	0.98	1.06	1.14

Installed Capacities vs Requested

Tier-2 Site	Resource	2010 US Pledge (what we were planning to pledge)	Installed (as of 08/2009)	2010 Request (this is what we pledged)
AGLT2	CPU (kHS-06)	6.860	13.11	11.04
	Disk (TB)	709	610	1,140
MWT2	CPU (kHS-06)	5.048	17.50	11.04
	Disk (TB)	362	576	1,140
NET2	CPU (kHS-06)	6.368	9.15	11.04
	Disk (TB)	727	377	1,140
SWT2	CPU (kHS-06)	7.864	9.14	11.04
	Disk (TB)	650	263	1,140
WT2	CPU (kHS-06)	4.764	5.00	11.04
	Disk (TB)	619	275	1,140
Total	CPU (kHS-06)	30.904	53.90	55.20
	Disk (TB)	3,067	2,101	5,700

Data Replicas at Tier-1s and Tier-2s

	Tier-1s	Tier-2s
Real RAW	1	0.1
Real ESD (current)	2	—
Real ESD (previous)	1	—
Real AOD (current)	2	10
Real AOD (previous)	1	10
Real DPD (current)	2	10
Sim HITS/RDO	0.1	0.1
Sim ESD (current)	2	—
Sim ESD (previous)	1	—
Sim AOD (current)	2	10
Sim AOD (previous)	1	4

Event size		
Real RAW size	MB	1.6
Real ESD size	MB	0.8
Real AOD size	MB	0.15
Sim RAW size	MB	2
Sim ESD size	MB	1.1
Sim AOD size	MB	0.18

Reality Check - 2010 Tier-1 Resources

➢ **Pledges presented at October C-RRB**

	CPU [HEP-SPEC06]	Disk [Tbytes]	Tape [Tbyte]
CA	10800	1095	710
DE	21600	2190	1420
ES	10288	1043	681
FR	27000	2738	1775
IT	16000	1800	1400
NDGF	7640	860	650
NL	33496	3001	2122
TW	14000	1750	1750
UK	27000	2738	1775
US	49680	5037	3266
SUM	217504	22252	15549
Needs	192000	21900	14200
Delta	13%	2%	10%

➢ **OK, little excess of CPU**

❑ Pledges submitted before change of ATLAS Tier-1 CPU requirement

2010 T1 – Disk [Tbytes]

Relative disk distribution per T1
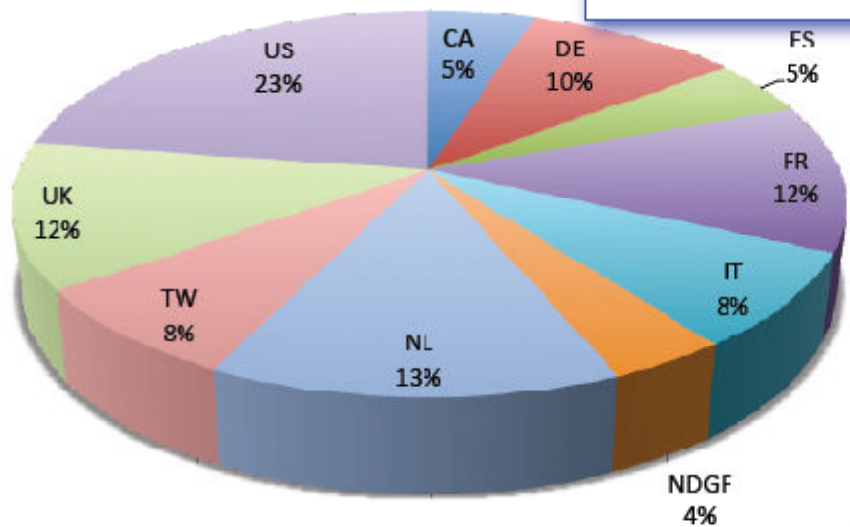
Disk/CPU (small dispersion)

2010 T1 – Disk/CPU
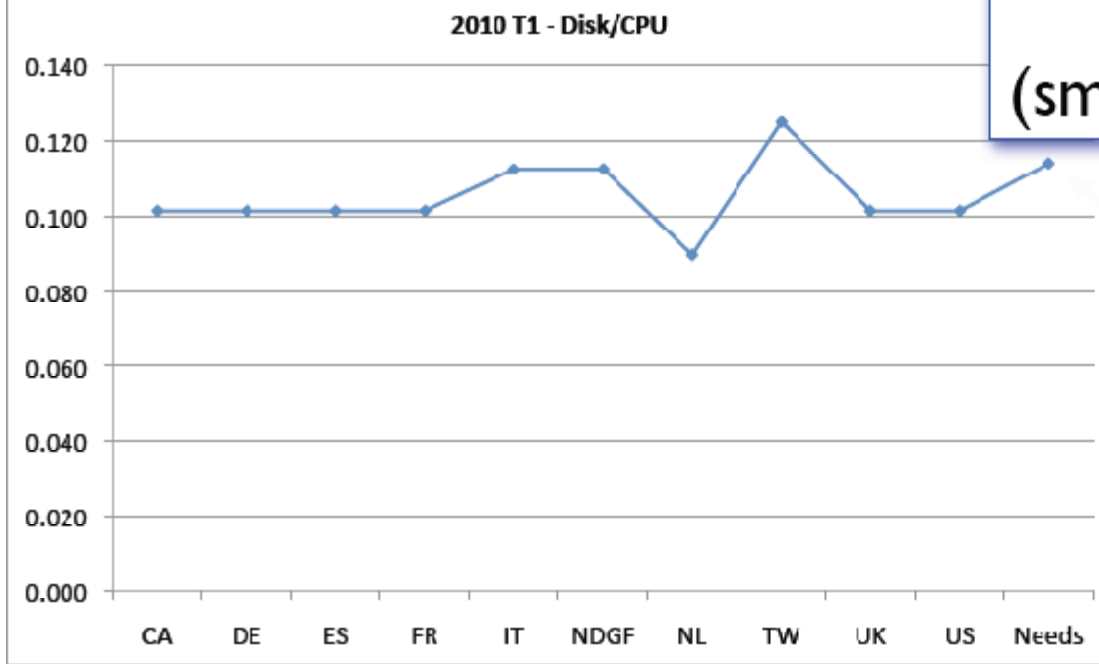
Pledges compared to M&O Share
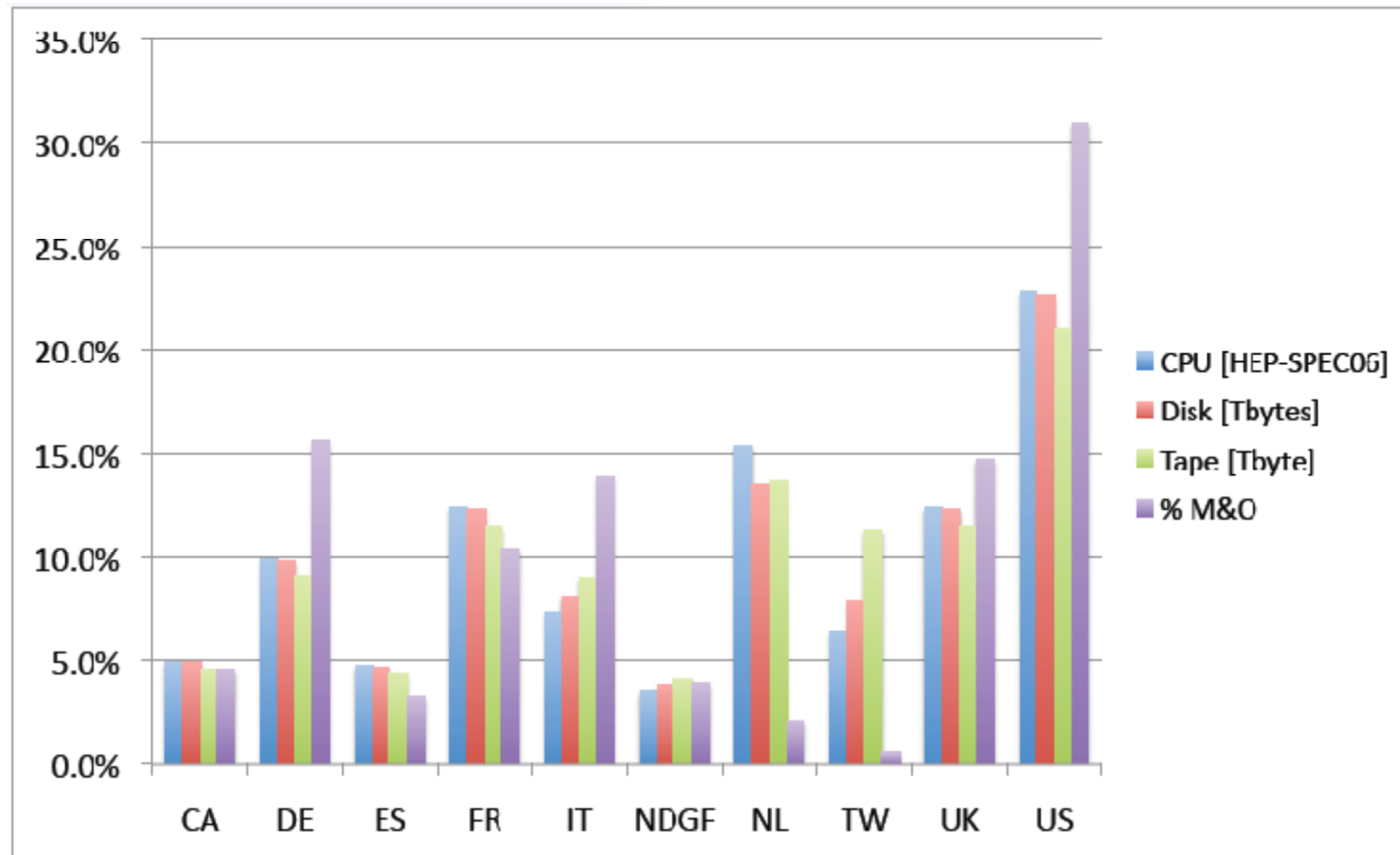
	CPU [HEP-SPEC06]	Disk [Tbytes]
CA	11520	1190
DE	30708	3361
ES	13508	1307
FR	41444	3136
IT	12000	1000
ND	11530	1000
NL	8300	925
TW	4100	425
UK	18695	2006
US	55200	5705
Needs	240000	24800
sum	207005	20055
Delta	-14%	-19%

2010 resources in T2s per cloud
15-20% deficit

2010 T2 - CPU [HEP-SPEC06]

US 27%, CA 5%, DE 15%, ES 6%, FR 20%, IT 6%, ND 6%, NL 4%, TW 2%, UK 9%

2010 T2 - Disk [Tbytes]

US 28%, CA 6%, DE 17%, ES 6%, FR 16%, IT 5%, ND 5%, NL 5%, TW 2%, UK 10%

2010 T2 - Disk/CPU

CM
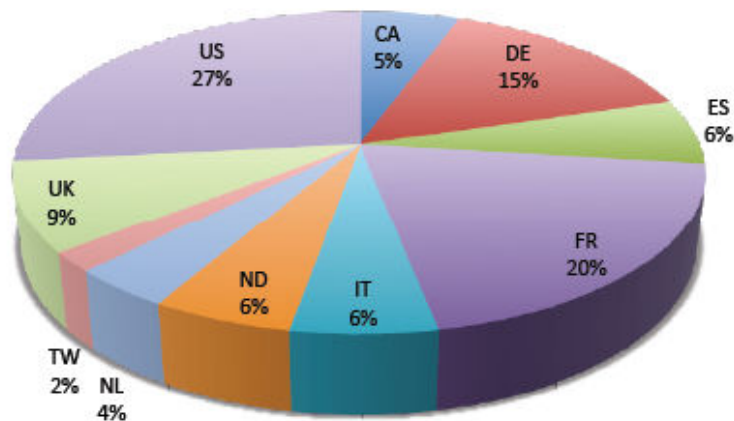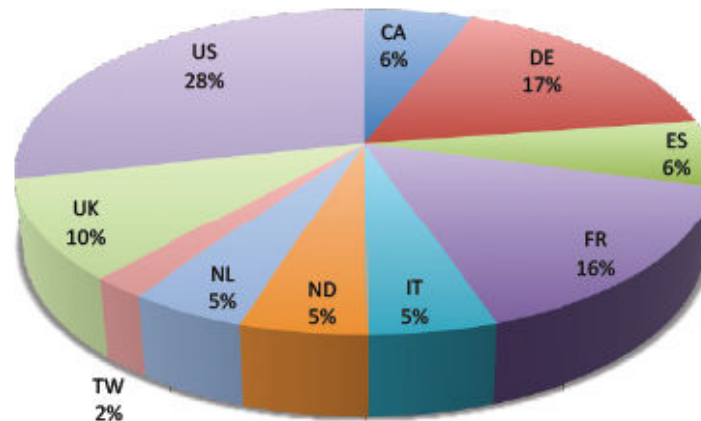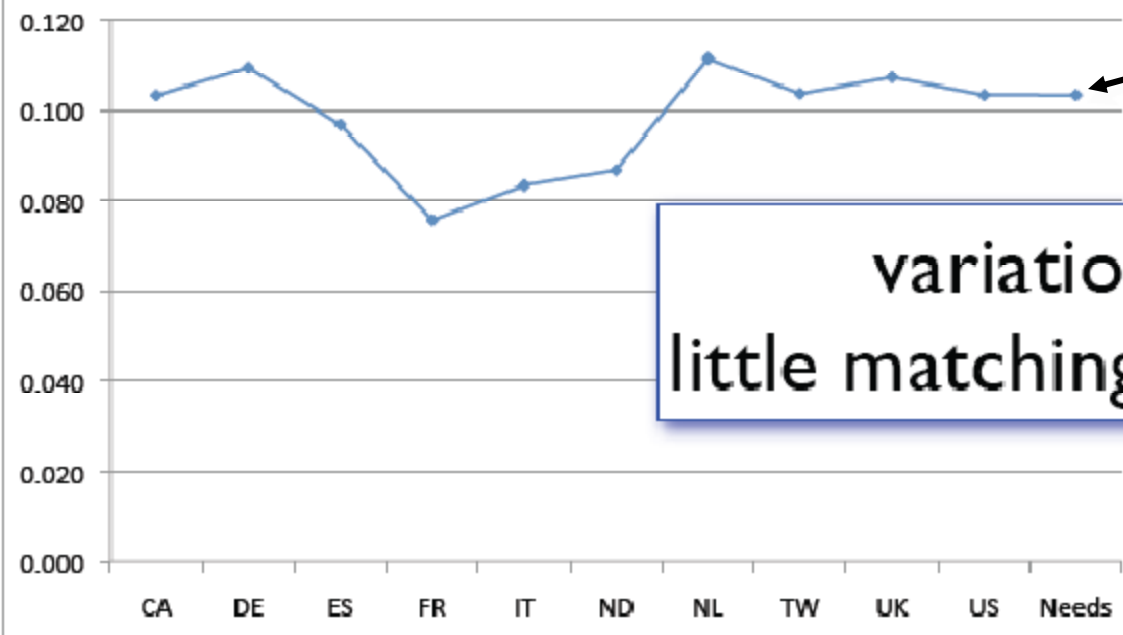
variations of Disk/CPU
little matching to computing model

2010 - CPU-T2/Disk-T1
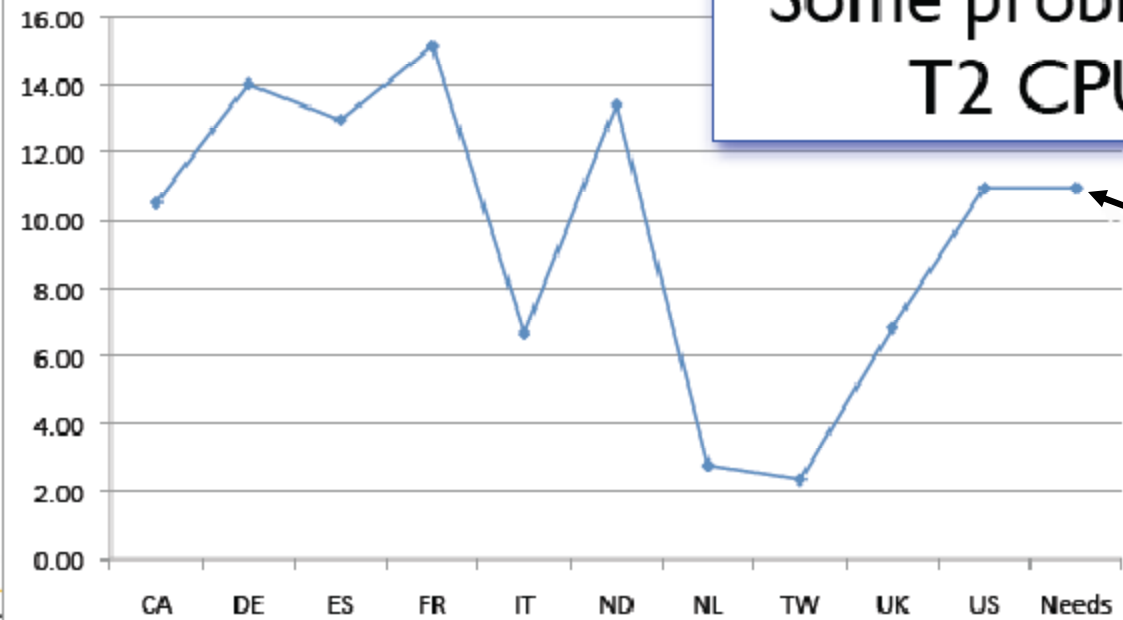
Some problems of matching between
T2 CPUS and T1 Disk Space

CM

Possible Data Distribution Patterns

1. AODs/DPDs are distributed to Tier-2s according to each Tier-2 cloud size
   - ❑ i.e. a 5% cloud gets 0.5 copies and a 25% cloud gets 2.5 copies
     - o Allocated statically
   - ❑ Needs planning in advance and leaves most disks empty till the end of data-taking

2. AODs/DPDs are distributed to (almost) fill disk space
   - ❑ Needs an automatic policy to reduce the number of replicas and make room for newer data
     - o Bare minimum is 2 replicas on disk in different clouds, better 3 or more
     - o Replicas that are not used for N weeks will go first
       - ▪ **Automatic warning but no negotiation**
   - ❑ Needs definition of custodial sites
   - ❑ Smaller clouds will have a rapid cycling of data and need to be able to keep up with deletions

➢ In both cases ATLAS will need to rank sites according to reliability
   - ❑ In case (1) to store data there
   - ❑ In case (2) to define custodial sites
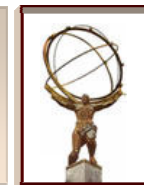
Data Management for 2009/2010 Run

- ➢ ATLAS Computing Model (ACM)
  - ❑ hIps://twiki.cern.ch/twiki/bin/view/AtIas/ComputingModel

- ➢ Evolution of the ATLAS Computing Model and Resource Needs for the First Year of Data-taking
  - ❑ hIps://twiki.cern.ch/twiki/pub/AtIas/Compu6ngModel/Evolution_of_ATLAS_Computing_Model_and_Resources-v05b.pdf

- ➢ Analysis Model for the First Year (AMFY)

- ➢ Pledged Resources
  - ❑ http://lcg.web.cern.ch/LCG/resources.htm

Analysis Data sample for DATADISK

- ➢ 200 Hz, 6M sec, AOD=0.18 MB => 216 TB

- ➢ If dESD have the same volume => 432 TB

- ➢ 2010 pledges are due in June (about half way into the run)

- ➢ Need to count on a 200 TB data sample initially

- ➢ Current grid-wide Tier-2 disk resources are ~3 PB

- ➢ ATLAS-wide we can store 10 copies of most recent version and 5 of the one but last

- ➢ Operational constraints may reduce this even further

Computing Model for Analysis Data

➢ Tier-1: 2 Copies of AOD & dESD of current version and 1 Copy of the previous version

➢ Tier-2: 10 Copies of AOD & dESD of the current version and 10 Copies of the previous version

- ❑ This requires 4 PB until June but only 3 PB deployed so far
- ❑ ATLAS will remove unused data
- ❑ Guarantee a minimum of "custodial" data
  - o Tier-1: 1 custodial copy of current version only
  - o Tier-2: 7 custodial copies of current version and 2 copies of the previous version

Tier-2 Resources from ATLAS' Perspective

- ➤ **Tier-2 sites even more unbalanced than Tier-1s**
  - ❑ From sites with 1 TB to sites with 1 PB storage

- ➤ **~25% of disk space is already used**

- ➤ **Tier-2s are for user analysis but should only be considered if**
  - ❑ They have at least 100 TB of storage space
  - ❑ They have passed the HC validation test at 90% efficiency, 150M events/day throughput

- ➤ **Tier-2s that don't pass criteria should/will not be considered to host any data**
  - ❑ Further reduces resources available to the collaboration

BROOKHAVEN
NATIONAL LABORATORY

office of
high energy physics

Tier-2 Data Placement & Deletion

➢ Put 10 copies of the latest version in Tier-2s

❑ 7 copies declared "custodial"

➢ When writing version n reduce

❑ To 2 the number of custodial copies of version n-1

❑ To 0 the number of custodial copies of version n-2

➢ When a site is >80% full

❑ Select least used and old data and

❑ Mark and publish non-custodial data to be deleted

❑ Delete 1 week later until site <60% full

What's New?

- ➤ ATLAS relies on data deletion rather than data placement
  - ❑ Less dependant on exact space provided
  - ❑ Sites will be kept always more or less full
  - ❑ Requires improvements of deletion mechanism

- ➤ Method guarantees a minimal number of custodial copies
  - ❑ But they may be scattered over sites and clouds

- ➤ Data may not be (any more) where users expected it
  - ❑ E.g. at their own/favorite site

- ➤ Tier-3 users need to copy data they want to keep locally
  - ❑ LOCALGROUP disk or find space somewhere on the grid

- ➤ The system automatically evolves to a selection of data at a site which has been frequently used

- ➤ Tier-2 Selection
  - ❑ E.g. according to performance as observed w/ HC tests, user feedback, etc

The Distributed U.S. ATLAS Computing Facility
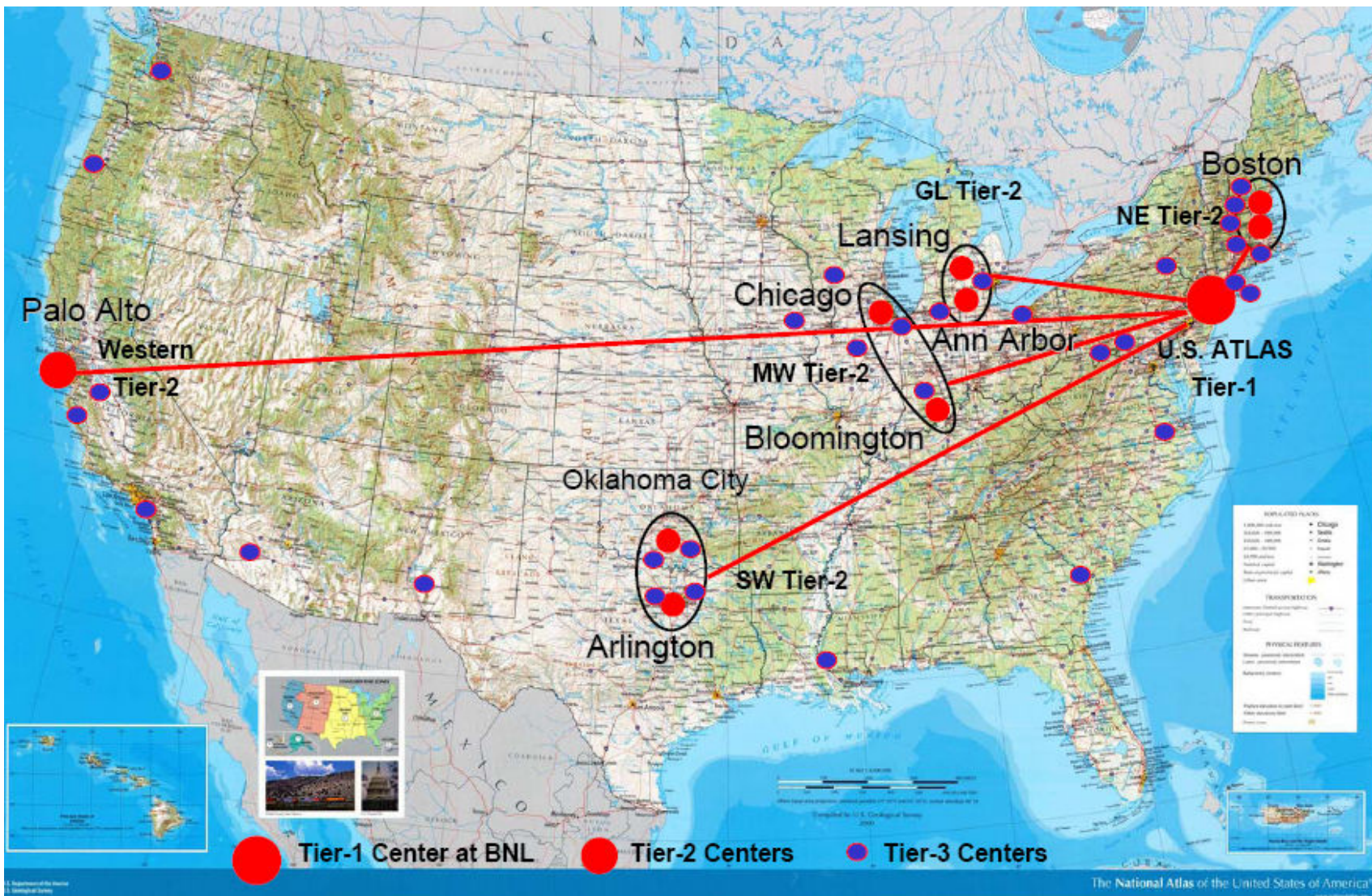
Computing Facilities- Recent Developments

➢Stability and Performance of all facility instances greatly improved over the course of the last 6 months

- ❑ ~7,500 jobs running continuously at the Tier-1 and the Tier-2 centers
- ❑ ~400M event sample produced for physics analysis and to exercise the facility components relevant to analysis computing

➢Progress at the U.S. ATLAS Tier-1

- ❑ Deployed 1.7 PB of Disk space in July
- ❑ New building (+ 6,400 sq.ft.) ready for occupancy
- ❑ 1 MW of UPS added (Diesel Generator backed Flywheel)
- ❑ Pledges were adjusted according to ATLAS' latest resource request

➢Scalable access to Conditions Data with Frontier and Squid

➢Progress on Networking

- ❑ Dedicated Circuits for Tier-1/Tier-2 Transfers
  - o Using ESnet's DCN Infrastructure
  - o Demonstrated up to 1 GigaByte/s between BNL and Universities
- ❑ CERN/Tier-1s ⇔ BNL from 10 Gbps => 20 Gbps by the end of September, 2009
- ❑ PerfSONAR monitoring implemented at all sites in collaboration with ESnet and Internet2

World Wide Production Jobs (last 12 months)

10k → US
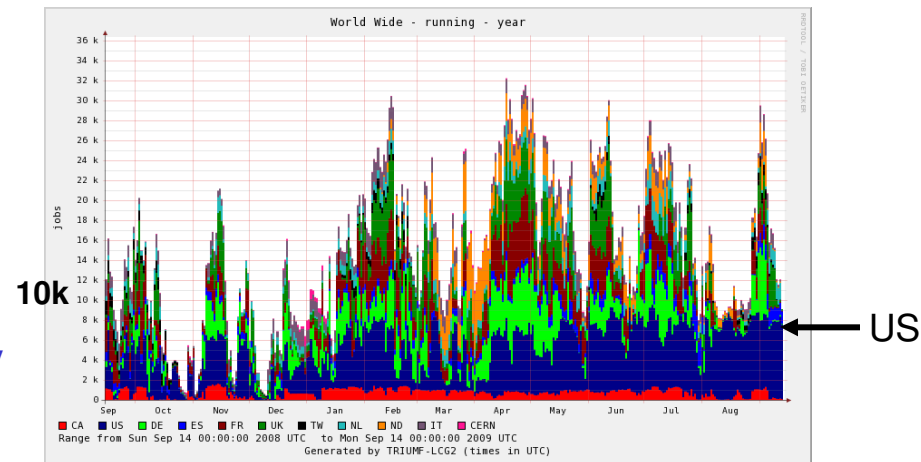
➢Added 2 areas to Facility Coordination

- ❑ Tier-3 Integration, Rik Yoshida (ANL)
- ❑ Facility related Physics Analysis Support , Nurcan Ozturk (UTA)
  - o Organizing analysis support shifts monitoring readiness of analysis resources at sites, responding to facility related problems and questions raised by users

➢Started to work on a coherent Grid Middleware Architecture

- ❑ In collaboration w/ OSG, USCMS and LIGO

US ATLAS Tier-3

➢ The aim of Tier-3 coordination by Rik Yoshida (ANL) and Doug Benjamin (Duke) is to maximize the effectiveness of US ATLAS Tier-3's for physics analysis. Rik and Doug are charged with coordinating the efforts of the institutes to bring up Tier-3 centers to support end-user analysis, operate them and integrate them into the US ATLAS computing system.

➢ Tentative plan of action to bring Tier-3's up and working early next year

   ❑ Tier-3g design (not complete) tentatively based on ANL and Duke Tier-3

   ❑ Discussion, cooperation with OSG, VDT, Condor experts

   ❑ A lot of work on T3 Grid Storage Element (SE) underway

➢ Are we on the right track?

   ❑ A lot of fruitful discussion driven by analysis needs (as we understand them today) to verify the key features of the design and its implementation

➢ Community Support

   ❑ Technical experts volunteered to contribute to designing and implementing Tier-3 centers

➢ Moving ahead very well, lots of progress in short time

Facilities Performance – a selection

➢ Production

➢ Data Replication

➢ Reprocessing from ESDs

❑ Using Tier-1 and Tier-2 Resources (unique in ATLAS)

➢ Analysis

Production
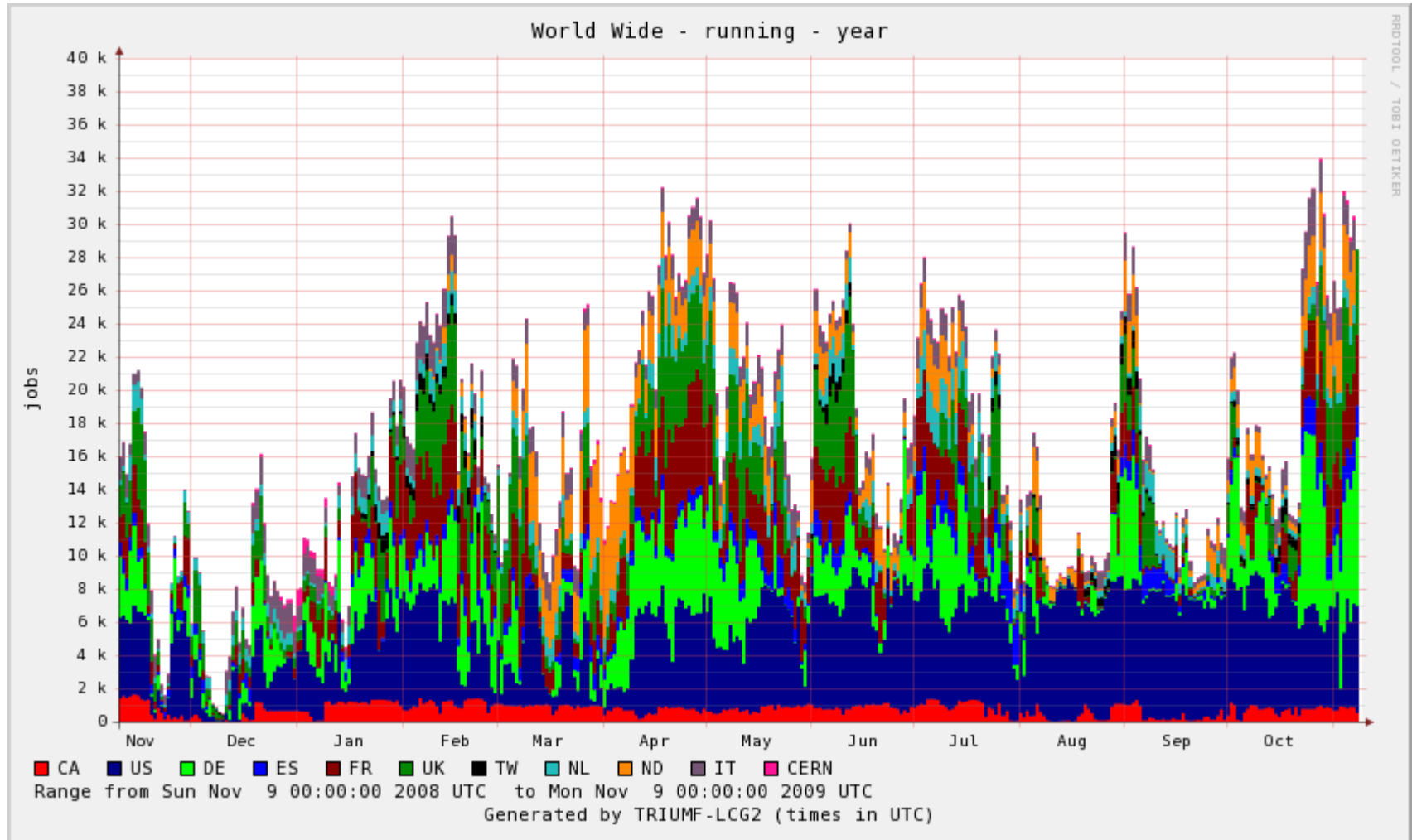
Data Replication

Throughput Test Result (From BNL to T2s)

Legend:
- WT2
- AGLT2
- MWT2_IU
- MWT2_UC

Source: Hiro

ATLAS dCache GridFTP Door Servers Cluster Network last hour

1.5 GB/s

In (Min:115.8M Avg: 1.0G Max: 1.8G)
Out (Min:116.8M Avg: 1.0G Max: 1.8G)
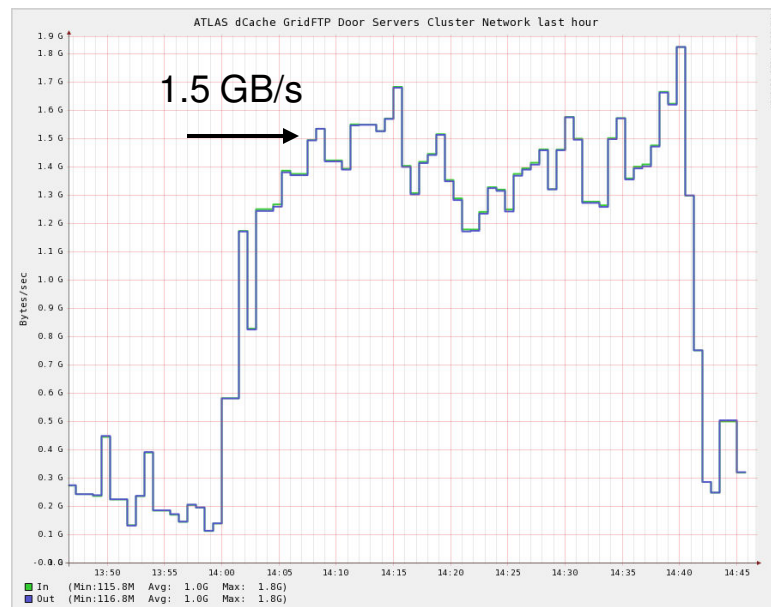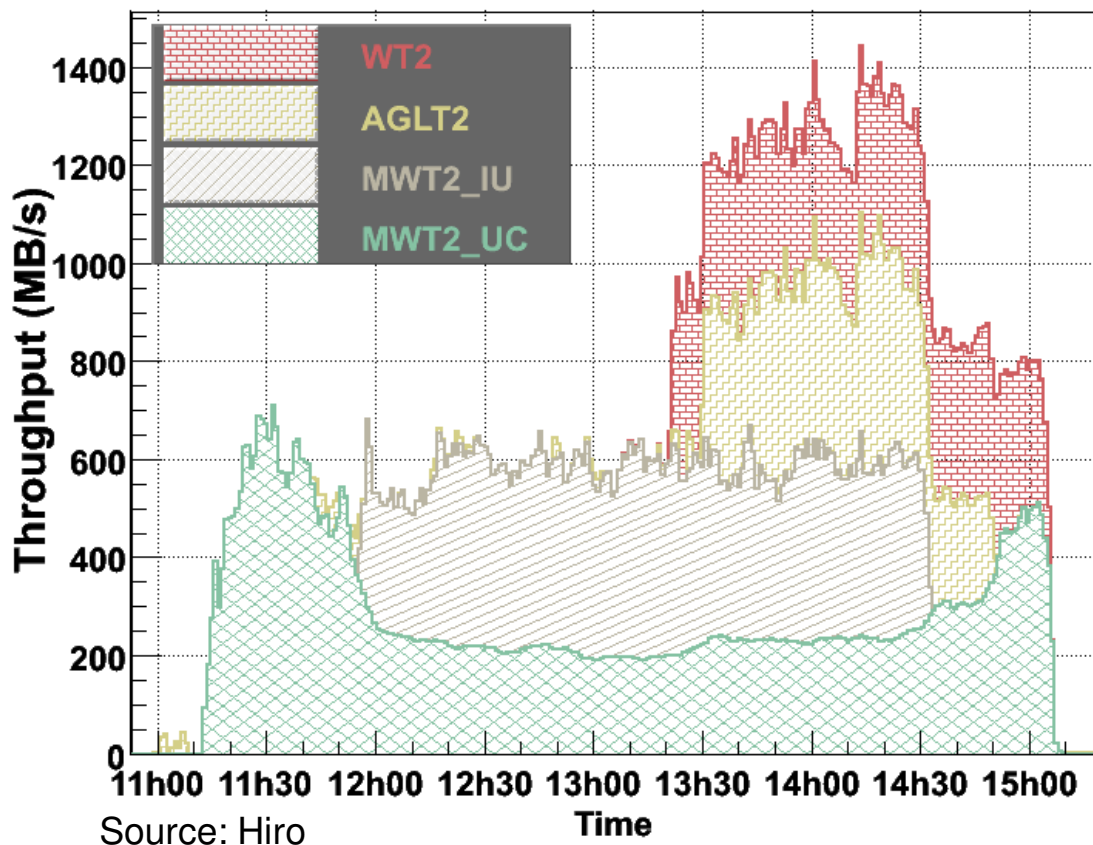
BNL Output Rate

Reprocessing

> ## Results from recent reprocessing Exercise

- ❑ 300 TB of Cosmic Ray Data taken in July

- ❑ Following a Fast Reprocessing exercise ATLAS wanted to reprocess the data from ESDs
  - o 138 runs, 951 containers, 298,378 files, 83.3 TB
  - o Less I/O than reprocessing from RAW
  - o 30% less CPU (~7 sec/event)
  - o Reprocessing to run at Tier-1 centers
    - · **8 out of 10 Tier-1 centers participated**
    - · **In the U.S. we included the Tier-2s**
  - o BNL had a complete replica (2.6 replicas ATLAS-wide)
  - o Task assignment to clouds by PanDA brokerage (taking data availability into account)

- ❑ Smooth operation over the course of the exercise
  - o 76% of 380k jobs done within 4 days
  - o 24% of jobs needed another 4 days caused by well understood/fixed issues

- ❑ Observed code validation issues
  - o The whole procedure relies on the groups thoroughly testing tags before they go into a cache, and then thoroughly checking the results of limited processing before signing off
  - o Due to known software issues we had to tolerate an error rate of 1.6%

BROOKHAVEN
NATIONAL LABORATORY

Reprocessing from ESDs

Reprocessing Jobs in all Clouds

date-tier	CA	DE	ES	IT	ND	NL	UK	US	sum
09-09-19	0	0	0	0	0	0	0	6370	6370
09-09-20	0	0	0	0	0	0	0	18618	18618
09-09-21	0	0	0	0	0	0	0	14184	14184
09-09-22	0	0	0	0	0	0	0	17766	17766
09-09-23	0	0	0	0	0	0	0	3	3
09-09-24	7793	3957	3365	5091	7008	3826	1163	24130	56333
09-09-25	7103	6644	3045	1578	1635	12417	5551	28244	66217
09-09-26	2442	9392	3049	6247	3494	13756	10965	41963	91308
09-09-27	1788	2087	1096	1614	1317	3495	10729	50364	72490
09-09-28	0	12	0	67	131	78	3216	15708	19212
09-09-29	0	0	0	4	2	27	175	6494	6702
09-09-30	3	0	0	0	0	6	35	274	318
09-10-01	0	0	0	0	0	0	0	125	125
09-10-02	0	0	0	0	0	0	0	3	3
total jobs	19527	22793	11028	15191	13969	34134	32358	228435	377435
total done	19129	22092	10555	14601	13587	33605	31834	224246	369649
%%	98.0	97.0	95.7	96.1	97.3	98.5	98.4	98.2	98.0
aborted	398	701	473	590	382	529	524	4189	7786
%%	2.1	3.1	4.3	3.9	2.8	1.6	1.6	1.9	2.1

ESD Reprocessing - Finished Jobs
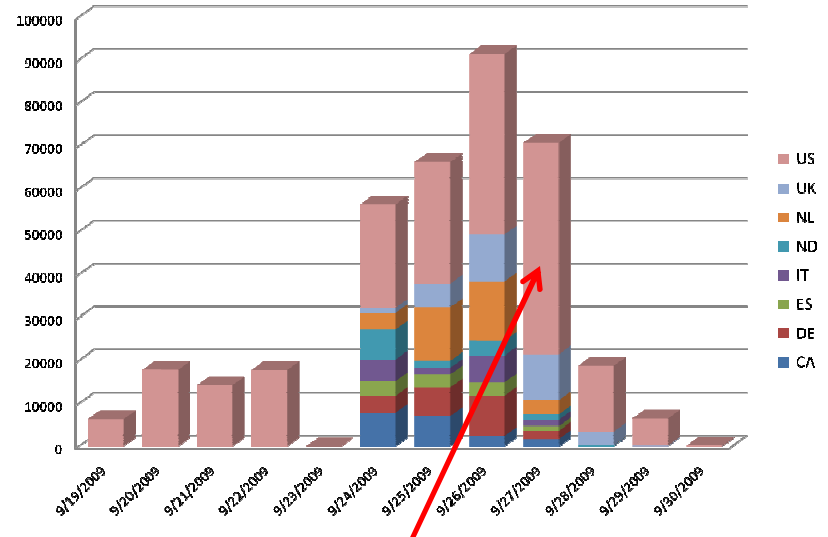
US demonstrated ability to process ~50k jobs/day

ESD Reprocessing by Cloud
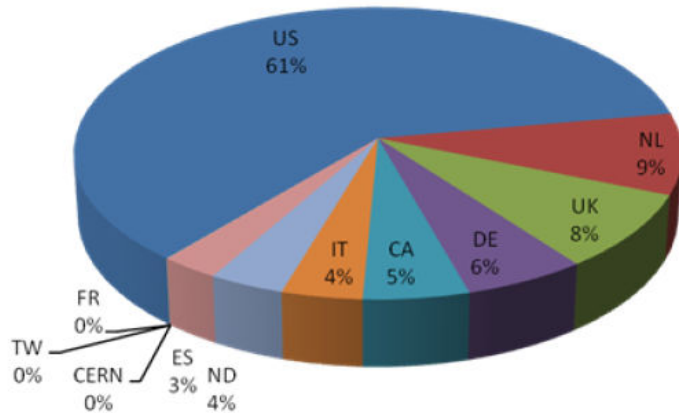
ESD
Reprocessing
U.S. Sites
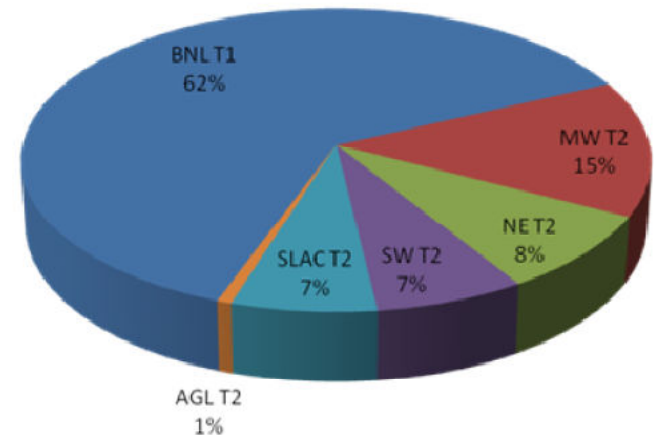
Source: Kaushik

AS Facility Meeting

Analysis Test - Oct. 28 @ ~2 pm

Source: Nurcan

Source: Nurcan

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

Additional Material

US-ATLAS Tier1	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	2560	4844	7337	12765	18193	21132	32200
Disk (Tbytes)	1100	3138	5822	11837	18509	18985	27800
Tape (Tbytes)	603	1715	3277	6286	9820	15085	18630
Nominal WAN (Mbits/sec)	9952	19904	29856	39808	39808	39808	39808
USA, Northeast ATLAS T2	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	394	685	1049	1592	1988	3990	5474
Disk (Tbytes)	103	244	445	727	1024	1851	2578
USA, Southwest ATLAS T2	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	898	1386	1734	1966	2514	3990	5474
Disk (Tbytes)	143	256	328	650	1103	1851	2578
USA, Midwest ATLAS T2	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	826	1112	978	1262	1785	3990	5474
Disk (Tbytes)	213	282	358	362	512	1851	2578
USA, Great Lakes ATLAS T2	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	581	965	1406	1670	2032	3990	5474
Disk (Tbytes)	155	322	542	709	914	1851	2578
USA, SLAC ATLAS T2	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	550	820	1202	1191	1685	3990	5474
Disk (Tbytes)	228	482	794	1034	1482	1851	2578
US ATLAS FACILITY	2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K)	5809	9792	13706	20446	28175	41080	59570
Disk (Tbytes)	1942	4702	8289	15119	21524	26220	40488

M. E

Facility Organization – Recent Development

Facility divided into two principal lines

❑ Production and Analysis Operation Coordination
- ❑ Nurcan Ozturk has agreed to coordinate Analysis Support

❑ Computing Deployment, Integration and Operation
- ❑ Rik Yoshida has agreed to coordinate Tier-3 development within Facilities
- ❑ Doug Benjamin coordinating Tier-3 Integration into Analysis Support (WBS 2.4) making important contributions to facility related Tier-3 development

❖ Many thanks to Nurcan, Doug and Rik !!!
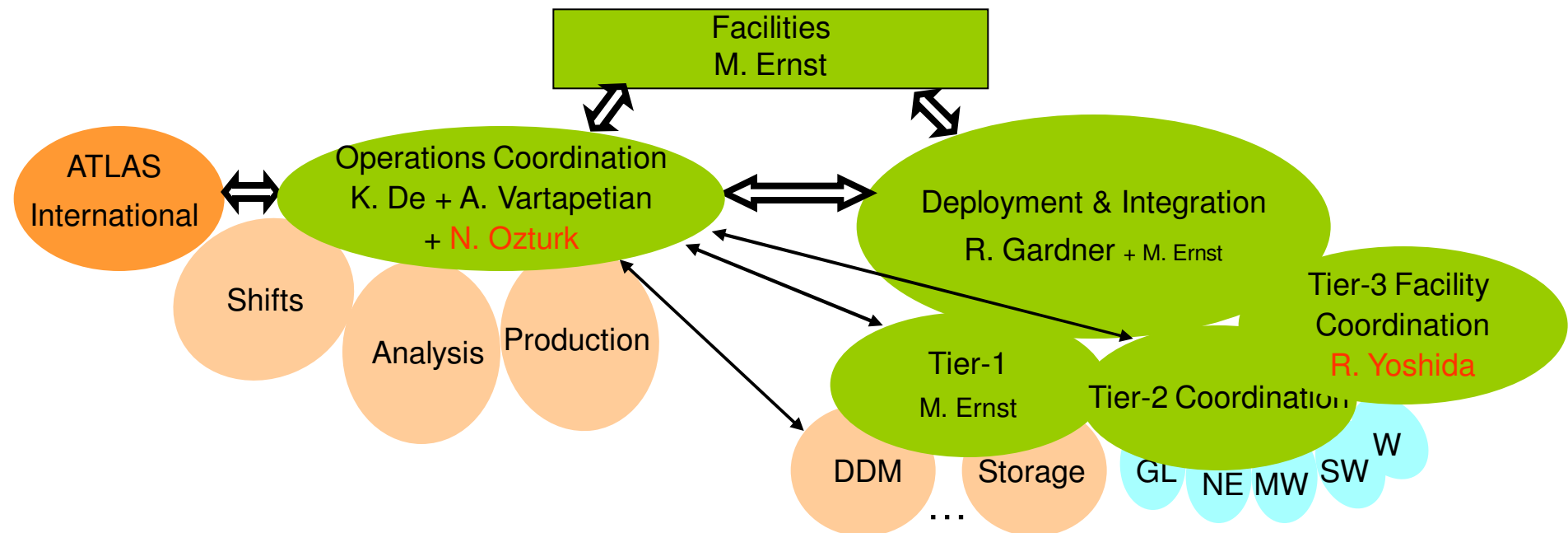
Roles for Sites

The roles in the ATLAS model are remarkably stable

- **CERN Tier-0:**
  - Prompt first pass processing on express/calibration & physics streams with old calibrations - calibration, monitoring
  - Calibration tasks on prompt data
  - 24-48 hours later, process full physics data streams with reasonable calibrations

- **CERN Analysis Facility**
  - Access to ESD and RAW/calibration data on demand
  - Essential for early calibration / Detector optimization / algorithmic development

- **U.S. ATLAS Tier-1 Center at BNL - 10 Tier-1s Centers worldwide**
  - Reprocess 1-2 months after arrival with better calibrations
  - Reprocess all resident RAW at run end with improved calibration and software
  - Centrally managed (Physics) Group Analysis / D1PD Production

- **5 distributed Tier-2s in the U.S. - 30+ Tier-2 Facilities worldwide**
  - On demand / grid-based user analysis of shared datasets
  - Limited access to ESD and RAW data sets
  - Simulation (some at Tier-1s until start of LHC data taking)

- **~30 Tier-3s expected in U.S.**
  - End-User on demand physics analysis
  - Data private and local - summary datasets (D3PDs, ntuples)
  - Not a centrally managed resource,
  - Not funded through U.S. ATLAS Operations Program, some funding from DOE and NSF