Текущее состояние и ближайшие перспективы компьютинга для АТЛАСа в России

А.Минаенко

Совещание по физике и компьютингу,

27 января 2011 г. МИФИ, Москва

ATLAS RuTier-2 tasks

- **Russian Tier-2 (RuTier-2) computing facility is planned to supply with computing resources all 4 LHC experiments including ATLAS. It is a *distributed* computing center including computing farms of 6 institutions: ITEP, RRC-KI, SINP (all Moscow), IHEP (Protvino), JINR (Dubna), PNPI (St.Petersburg). Two smaller sites MEPhI and FIAN are now present in the TiersOfAtlas list but they have smaller resources and really can be used as Tier-3 sites only**
- **The main RuTier-2 task is providing facilities for physics analysis of collected data using mainly AOD, DESD and group/user derived data formats**
- **Now group *atlas/ru* exists in the framework of ATLAS VO. It includes physicists intending to carry analysis in RuTier-2 and the group list contains 49(81) names at the moment. The group will have privilege of write access to local RuTier-2 disk resources (space token LOCALGROUPDISK)**
- **All the data used for analysis should be stored on disks**
- **The second important task is production and storage of MC simulated data**
- **The full size of data and CPU needed for their analysis are proportional to the collected statistics. The resources needed should constantly grow with the increase of the number of collected events.**

Current RuTier-2 resources for ATLAS

	CPU	Disc, TB	ATLAS Disk, TB
RRC-KI	1024	1000	316
JINR	916	732	240
IHEP	400	358	226
PNPI	208	184	126
ITEP	268	150	10
SINP	184	153	3
MEPhI	192	60	34
FIAN	52	49	28
Total	3244	2686	983

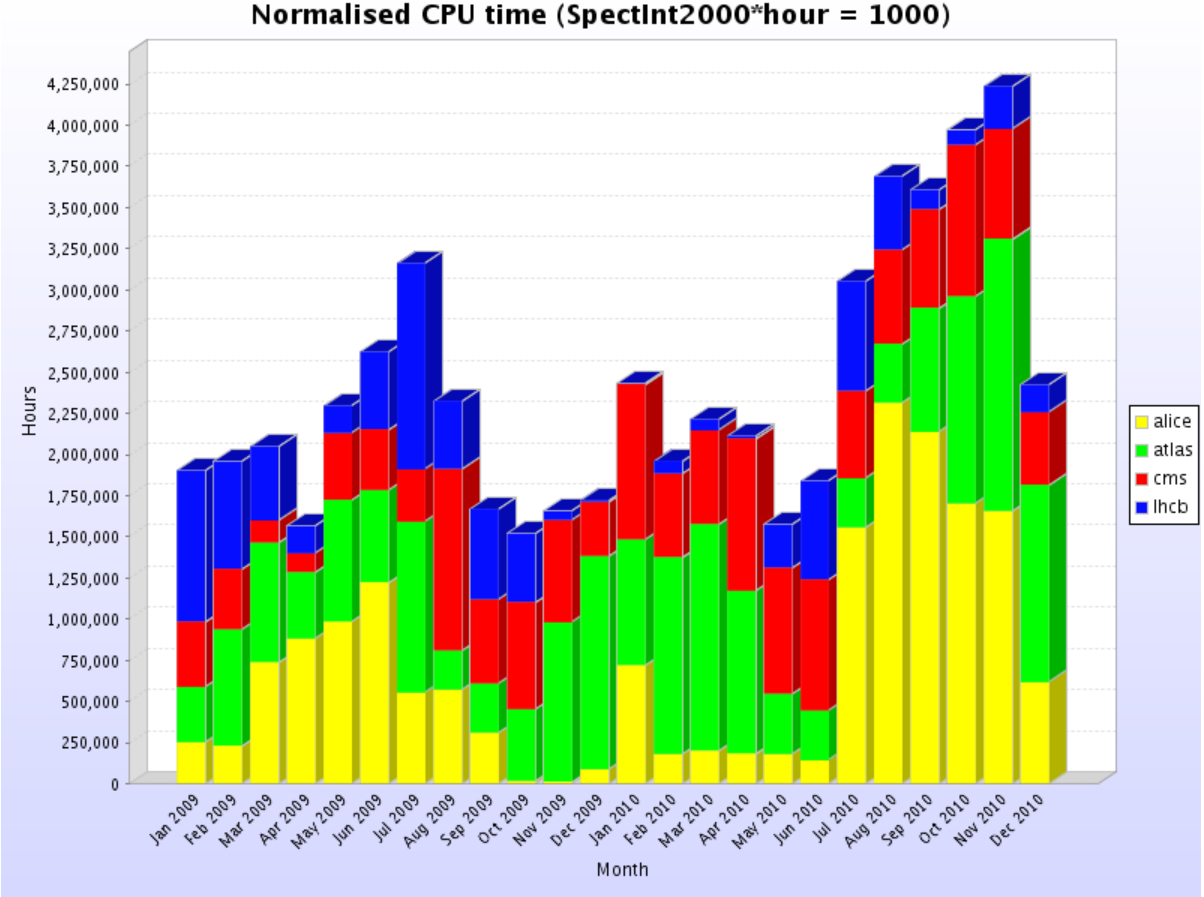
- Red – sites for user analysis of ATLAS data, the other for simulation only
- The total number of CPU cores in 2010 (2009) is about 3200 (2500), increase by 40%
- ATLAS disk resource in 2010 (2009) is about 980 (560) TB), increase by 75%. This disk space is exactly sufficient to keep ATLAS statistics 2009-2010 (AOD+DESD+group data)
- Now the main type of LHC grid jobs is official production jobs and CPU resources are at the moment dynamically shared by all 4 LHC VO
- For 2011 the resource increase has NOT taken place

Space tokens at RuTier-2

Tokens (TB)	RRC-KI	JINR	IHEP	PNPI	MEPhI	ITEP	SINP	FIAN	Sum
DATADISK	134	105	100	60	6	3	1	7	416
MCDISK	90	60	55	50	6	2	0	7	270
GROUPDISK	60	50	50	0	16	0	0	7	183
SCRATCHDISK	12	10	8	6	1	1	0	1	39
LOCALGROUPDISK	12	9	8	6	2	1	0	1	39
PRODDISK	5,5	4	3	2	1	2	1	1	19,5
HOTDISK	2	2	2	2	2	1	1	1	13
Total	315,5	240	226	126	34	10	3	25	979,5
Planned	315,5	235	202	118	34,5	10	3	26,5	944,5

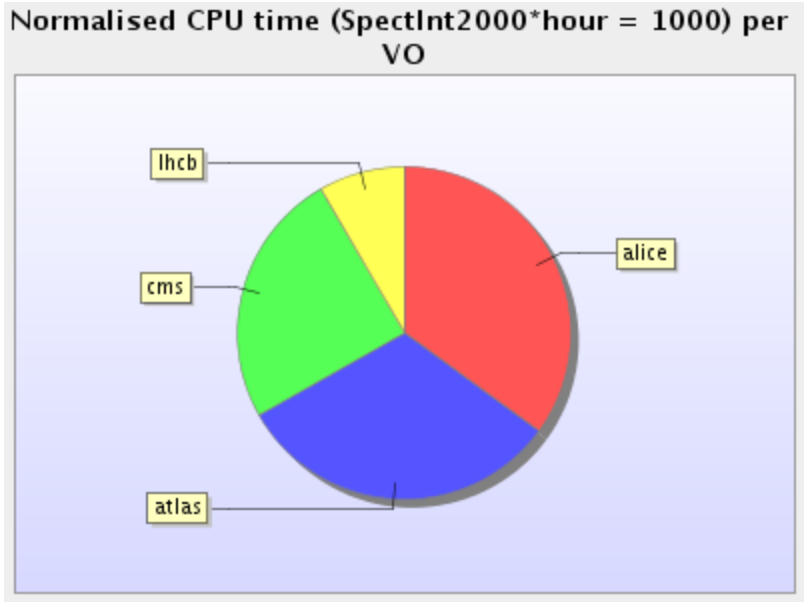
- Minimal ATLAS requests for minimal main token sizes at Tier-2:
  - DATADISK – 50 TB
  - MCDISK – 50 TB
  - GROUPDISK – 50 TB
- DATADISK – 399 TB, MCDISK – 255 TB
- 3 group disks are assigned to RuTier-2
  - RRC-KI – exotic
  - JINR – SM
  - IHEP - JetEtmiss

RuTier-2 CPU resources usage in 2009 and 2010 (January-December)

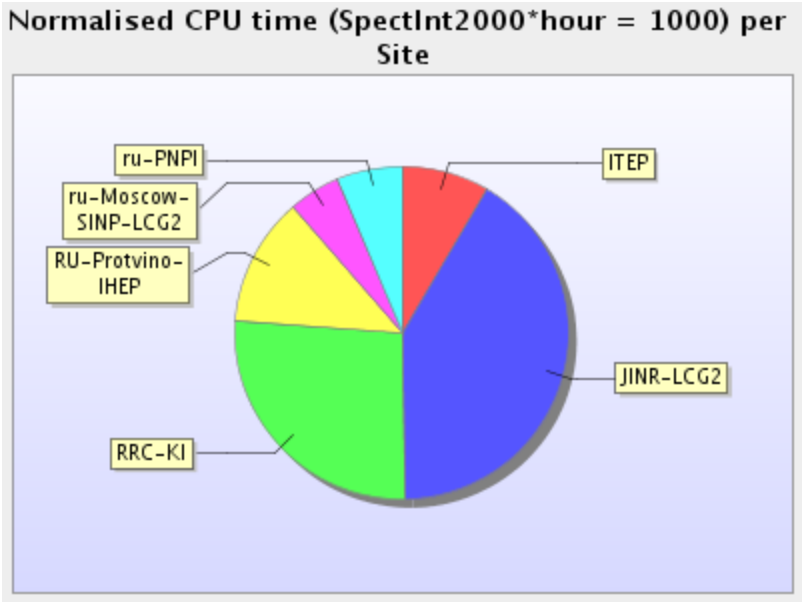


2009: 26.8 M*kSI2k*hour **2010: 33.1 M*kSI2k*hour**

RuTier-2 CPU resources usage in January-December 2010 (all 4 LHC exp.)

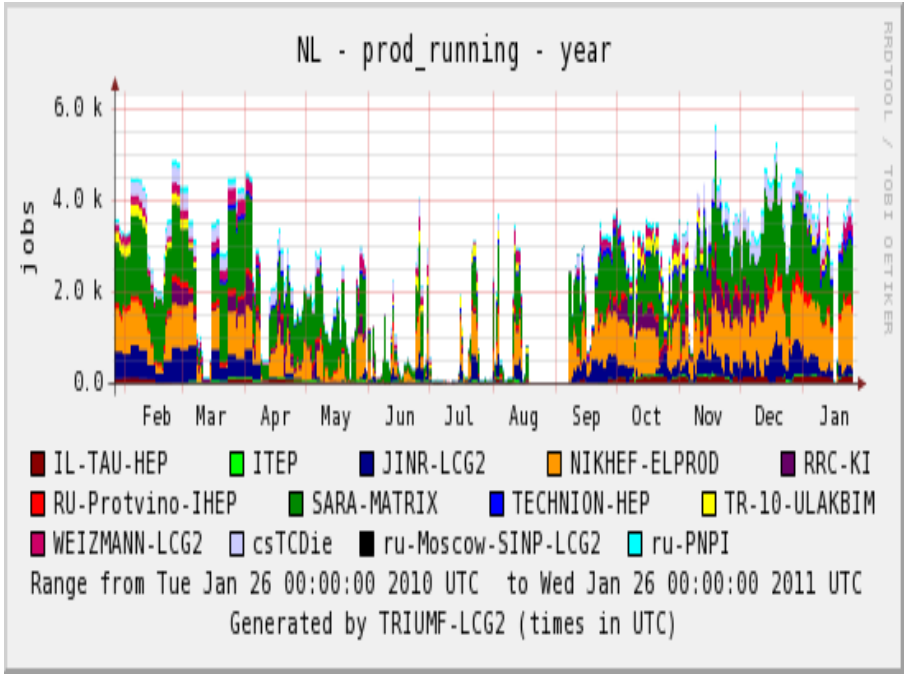


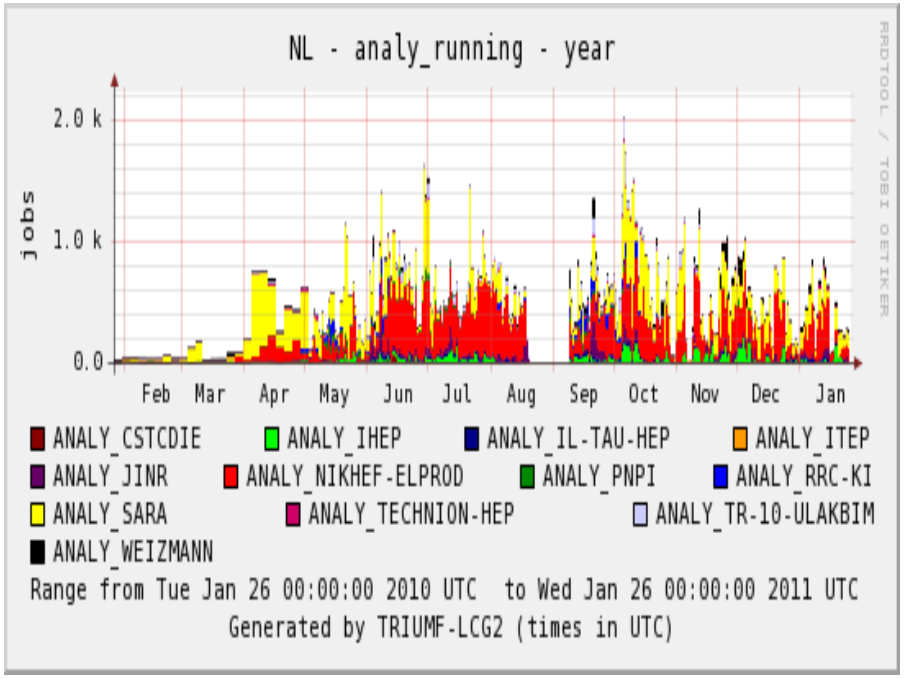
- ALICE – 30%
- ATLAS – 32%
- CMS – 24%
- LHCb – 14%



- JINR – 41%
- RRC-KI – 26%
- IHEP – 12%
- ITEP – 9%
- PNPI – 6%
- SINP – 5%

Total number of production (left) and analysis (right) jobs in the NL cloud during 2010





- **Number of production jobs is fluctuating but at the same mean level**
- **Now the production jobs takes the most part of consumed CPU time (larger number of jobs, longer in time and using CPU time more efficiently)**

ATLAS RuTier-2 and data distribution in the fist part of 2010

- The sites of RuTier-2 are associated with ATLAS Tier-1 SARA
- Now 8 sites IHEP, ITEP, JINR, RRC-KI, SINP, PNPI, MEPhI, FIAN are included in TiersOfAtlas list and FTS channels are tuned for data transfers to/from the sites
- 4 sites of them (IHEP, JINR, RRC-KI, PNPI) will be used by ATLAS data analysis and all physics data need for analysis will be kept at these sites. The other 4 sites will be used for MC simulations only/mostly
- RuTier-2 is was subscribed to get all simulated AOD as well as real data AOD, DESD, Tags . The data transfer was done automatically under steering and control of central ATLAS DDM (Distributed Data Management) group
- Real data ESDs were replicated also to RuTier-2 sites. But when all available disk space will be filled, ESD data will be gradually deleted and replaced with AOD, DESD
- The used shares correspond to disk resources available for ATLAS at the sites:
  - ✓ RRC-KI – 35%
  - ✓ JINR – 25%
  - ✓ IHEP – 25%
  - ✓ PNPI – 15%

Space token current status

DATADISK	Total TB	Used TB	Used %
RRC-KI	121	98.2	81
JINR	105	93.6	89
IHEP	100	82.6	82
PNPI	60	52.8	88

LOCALGROUPDISK	Total TB	Used TB	Used %
RRC-KI	12	0	0
JINR	9	2.7	29
IHEP	8	0.14	1
PNPI	6	0	0

GROUPDISK	Total TB	Used TB	Used %
RRC-KI	50	2.3	4
JINR	50	11.6	23
IHEP	30	24.1	83

Estimate of resources needed to fulfil RuTier-2 tasks in 2011

Resource type	ALICE	ATLAS	CMS	LHCb	Sum/Average
VO authors	568	1835	1367	351	4121
Ru+JINR authors	46	88	79	32	245
Ru+JINR authors (%)	8,10	4,80	5,78	9,12	5,95
CPU all T2 (kSH06), 2010	95,0	239,4	199,0	31,7	565,1
CPU all T2 (kSH06), 2011	111,0	277,5	305,0	36,0	729,5
Disks all T2 (PB), 2010	10,0	20,1	9,0	0,0	39,1
Disks all T2 (PB), 2011	6,0	34,2	18,1	0,0	58,3
Ru-Tier2 CPU (kSH06), 2010	7,7	11,5	11,5	2,9	33,6
Ru-Tier2 CPU (%), 2010	22,9	34,2	34,3	8,6	100,0
Ru-Tier2 CPU (kSH06), 2011	9,0	13,3	17,6	3,3	43,2
Ru-Tier2 CPU (%), 2011	20,8	30,8	40,8	7,6	100,0
Ru-Tier2 Disks (PB), 2010	0,81	0,96	0,52	0,14	2,4
Ru-Tier2 Disks (%), 2010	33,3	39,6	21,4	5,8	100,0
Ru-Tier2 Disks (PB), 2011	0,49	1,64	1,05	0,200	3,4
Ru-Tier2 Disks (%), 2011	14,4	48,6	31,0	5,9	100,0
Ru-Tier2 CPU (kSH06), 2010 fact	6,9	10,3	10,3	2,6	30,0
Ru-Tier2 Disks (PB), 2010 fact	0,81	0,96	0,96	0,14	2,9

- **CPU for ATLAS in 2011 (2010) – 1500 (1300)**
- **Discs for ATLAS in 2011 (2010) – 1640 (960) TB**
- **$13.8*10^6$ sec * 200 ev * 150 kB = 414 TB (AOD)**

2011: “reasonable” numbers

- 4 TeV (to be discussed at Chamonix)
- 936 bunches (75 ns)
- 3 micron emittance
- 1.2 x 10^{11} protons/bunch
- beta* = 2.5 m, nominal crossing angle

Peak luminosity	6.4 x 10^{32}
Integrated per day	11 pb^{-1}
200 days	2.2 fb^{-1}
Stored energy	72 MJ

Usual warnings apply – see problems, problems above

Slides by Roger Bailey
LHCC, November 17

Ultimate reach

- 4 TeV
- 1400 bunches (50 ns)
- 2.5 micron emittance
- 1.5 x 10^{11} protons/bunch
- beta* = 2.0 m, nominal crossing angle

Peak luminosity	2.2 x 10^{33}
Integrated per day	38 pb^{-1}
200 days	7.6 fb^{-1}
Stored energy	134 MJ

Usual warnings particularly apply – see problems, problems above

Slides by Roger Bailey
LHCC, November 17

Kors Bos slide 1

Changes to the Computing Model

- Our current resources are nearly full with 2010 data
- We don't know how much increase we get for 2011
- More beam and luminosity in 2011
  - 200 days of uninterrupted running
- Likely even more so in 2012
  - To be decided before the April RRB
- New energy (8 (?) TeV) in 2011
  - To be decided in Chamonix in January
- Request for higher trigger rates
  - 400 Hz and 600 Hz need to be considered
- We need revolution, evolution is not enough

Kors Bos slide 2

New parameters to work with

- 200 days of pp and 28 days of HI during 2011
- Most likely the same for 2012
- 50% of the time stable beams
- We will take data at 400 Hz
- Energy will be 4+4 TeV
- 7 TeV MC will still be needed for some time
- HI MC will be needed
- 30% increase of resources on April 1st 2011 and 2012

Kors Bos slide 3

Revolutionary ideas needed

- Reduce event size from Calo, Indet, Trigg, ..
  - Factor 3 (?) in RAW size, gives also smaller ESD
- Use PD2P everywhere (maybe slower, but ..)
  - And make it more intelligent
- Run bulk processing on the grid
- Get more T1s (stable site with tape) (CERN, ..)
- Custodial copies more distributed:
  - RAW on tape: 1 @CERN and 1 @T1s
  - HITS on tape: 1 @T1s , @T2s
  - ESD on tape: 1 @T1s
- Fewer primary copies on disk:
  - ESD 1 copy @T1s
  - AOD, DESD, etc. on disk: 2 (3?) copies @T1s
  - No a priori data @T2s (rely on PD2P)
- Do away with space tokens (at least in T2s)
  - Just one big cache
  - Will dramatically improve life for Groups

What is PD2P

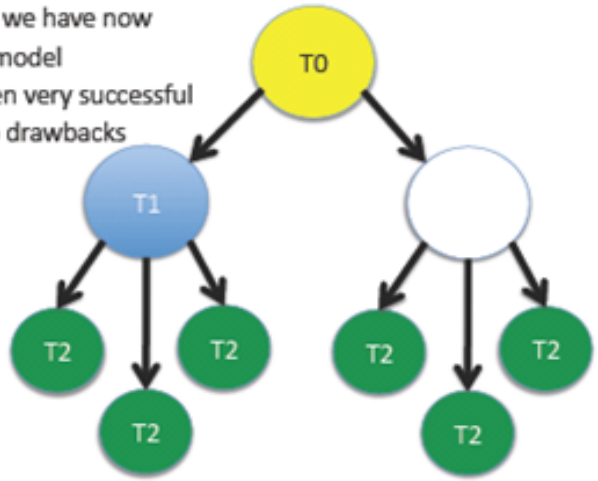
- Dynamic data placement at Tier 2's
  - Keep automatic distribution to Tier 1's – treat them as repositories
  - Reduce automatic data subscriptions to Tier 2's – instead use PD2P
- The plan
  - Panda will subscribe a dataset to a Tier 2, if no other copies are available (except at a Tier 1), as soon as any user needs the dataset
    - User jobs will still go to Tier 1 while data is being transferred – no delay
  - Panda will subscribe a replica 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
- 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 which are too small (storage mainly) or too slow
  - Provenance metadata is not GP, hidden metadata is not True

Data Distribution is Very Important

- Most user analysis jobs run at Tier 2 sites
  - Jobs are sent to data
  - We rely on pushing data out to Tier 2 sites promptly
  - Difficult since many data formats and many sites
  - We adjusted frequently the number of copies and data types in April & May
  - But Tier 2 sites were filling up too rapidly, and user pattern was unpredictable
  - Most datasets copied to Tier 2's were never used

Data placement model
The"Monarch Model"

- This is what we have now
- It is a push model
- And has been very successful
- But has also drawbacks

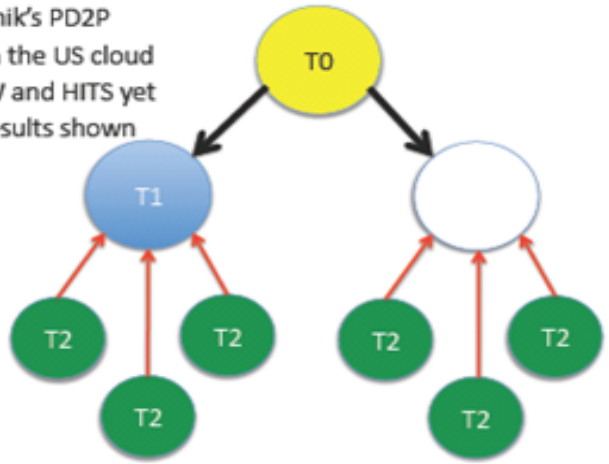


We Changed Data Distribution Model

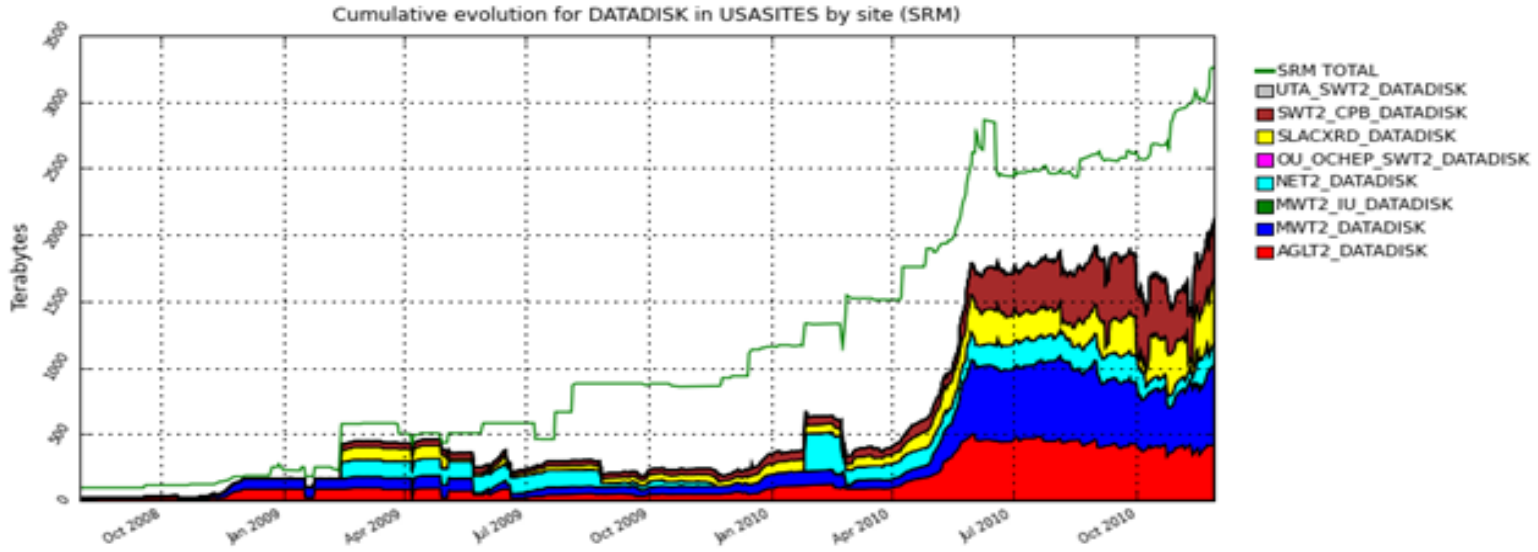
- Reduce pushed data copies to Tier 2's
  - Only send small fraction of AOD's automatically
  - Pull all other data types, when needed by users
  - Note: for production we have always pulled data as needed
- But users were insulated from this change
  - Did not affect the many critical ongoing analyses
  - No delays in running jobs
  - No change in user workflow

Data pull model I

- This is Kaushik's PD2P
- Runs now in the US cloud
- Not for RAW and HITS yet
- Intersting results shown



Data Flow to Tier 2's



- Example above is from US Tier 2 sites
  - Exponential rise in April and May, after LHC start
  - We changed data distribution model end of June – PD2P
  - Much slower rise since July, even as luminosity grows rapidly

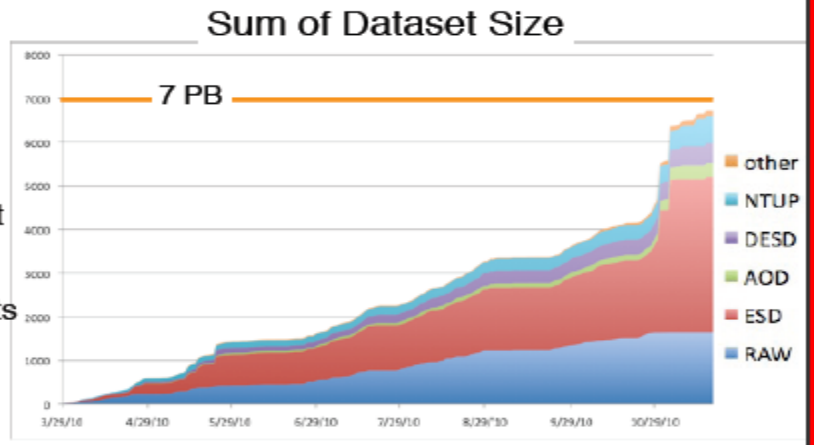
Computing model for 2011-2012

Data Volume

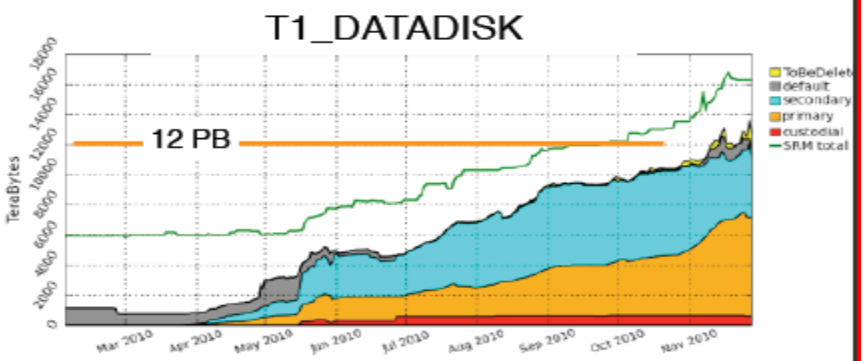
data10_7TeV total
- RAW=1.6 PB, ESD=3.5 PB

data10 Oct average event size
- RAW: 1.40 MB/event, ESD: 1.48 MB/event

data10_hi (as of Dec 2.)
- RAW: 1.48 MB/event, 300 TB, 202M events
- ESD: 2.01 MB/event, (400 TB)



2011 prospects (a naive calculation)
- 400 Hz x 200 days x 50% = 3.5 Gevts
- RAW (1.4 MB) : 4.8 PB
- ESD (1.48 MB) : 5.1 PB
- **2 repro (merged) = 10 PB ESD**
- **1 repro (recon+merge) = 10 PB ESD**



Note:
- data size depends on the stream
- change of trigger rate could affect average event sizes

Computing model for 2011/12 – I. Ueda (ATLAS S&C Workshop 2010.12.03.)

3

Computing model for 2011-2012 (I.Ueda)

Data Volume

Pledges 2010

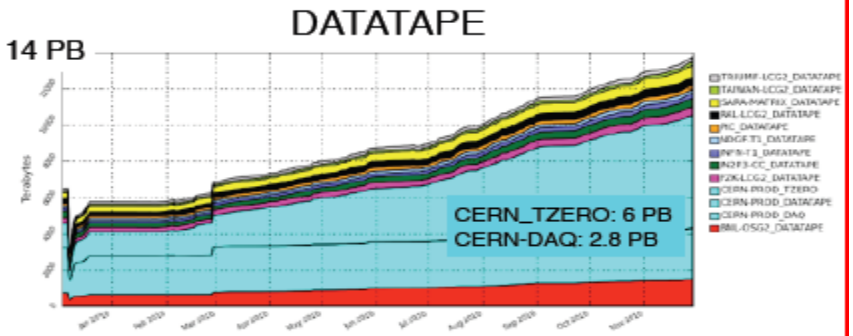
- T1_Disk=22 PB, T1_Tape=15 PB, T2_Disk=21 PB

Pledges 2011 (tentative)

- T1_Disk=26 PB, T1_Tape=32 PB, T2_Disk=34 PB
- eg. datadisk=15 PB, mcdisk=5 PB

Do not keep ESD on disk (put them all onto tape)

- we have not used T1_TAPE much



Concequences

- Analysis jobs would not be able to run on ESD because they are on tape.
  - No PD2P of ESD to T2s.
- Group productions run on ESD and provide dESD/D3PD to user analysis

Computing model for 2011-2012 (I.Ueda)

Tier2 Disks

Having both pre-defined distribution + PD2P is problematic for many Tier-2s.

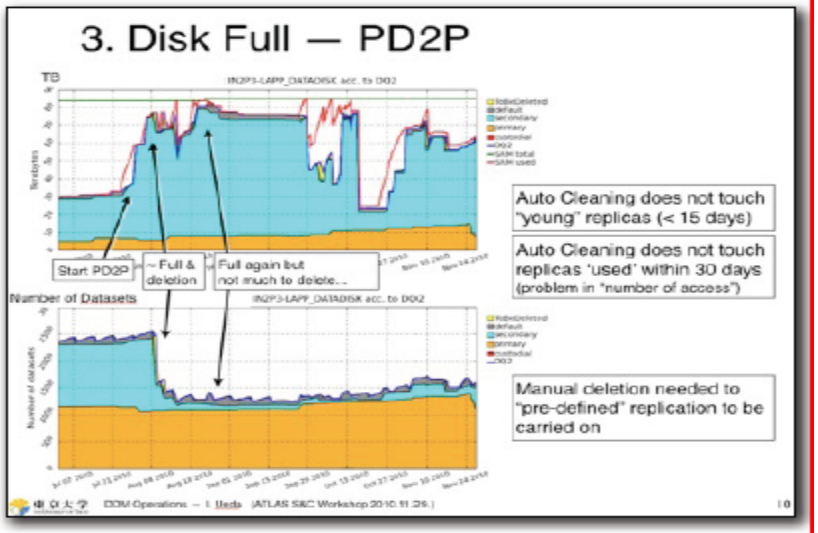
questions posed.

- "Why do we need 'primary' at Tier-2s?"
- "What are 'primaries' at Tier-2s?"

It was to ensure for the users to have access to the data

- but now we have PD2P and its extension soon

No more pre-defined distribution to Tier-2s (and Tier-3s)



- All the T2_DATADISK and T2_MCDISK spaces can be used for PD2P and on-demand replication (DaTRI)

Computing model for 2011/12 – I. Ueda (ATLAS S&C Workshop 2010.12.03.) 5

Computing model for 2011-2012 (I.Ueda)

Space Tokens

Space token (reserved space) has been used as a substitute for "quota"

- now we have some accounting tools, and better accounting system is coming soon (see DDM session)
- No more T2_datadisk,T2_mcdisk and T2_groupdisk but a single PD2P cache
- "Global quota" on group data on T1s would require flexibility in per-site spaces
  - Changing the quota centrally is better than asking sites to adjust the shares among the spaces

Merge datadisk, mcdisk and groupdisk into one

- probably would be called as 'datadisk'
- even proddisk?

Computing model for 2011/12 – I. Ueda (ATLAS S&C Workshop 2010.12.03.) 8

Computing model for 2011-2012 (I.Ueda)

Summary

Do not keep ESD on disk (put them all onto tape)

- start immediately once decided, or before start of run next year

Merge datadisk, mcdisk and groupdisk into one

- starting with the new pledges (2011)

T1_DISK

- data/mc/group shares controlled centrally
  - with the new accounting system (Jan-Feb 2011)
- Tier-1s host group data (persistent store) with 'global quota'
  - can start with quota per T1 with auto-distribution. need some development. need the new merged space.

T2_DISK

- No more pre-defined distribution to Tier-2s (and Tier-3s)
  - can stop even now. wait for extended PD2P (and renamed)
- PD2P and on-demand requests to fill the space (incl. group data)
  - PD2P extension on-going. (Jan-Feb 2011)

Computing model for 2011/12 – I. Ueda (ATLAS S&C Workshop 2010.12.03.) 9