Reliability Engineering for ATLAS Petascale Data Processing on the Grid

Dmitry Golubkov, Andrei Minaenko and Alexandre Vaniachine

ATLAS Computing Workshop held in conjunction with the Grid2012 Conference

July 17, 2012

JINR, Dubna, Russia

Open Areas of Research

End-to-End Data Solutions for Distributed Petascale Science

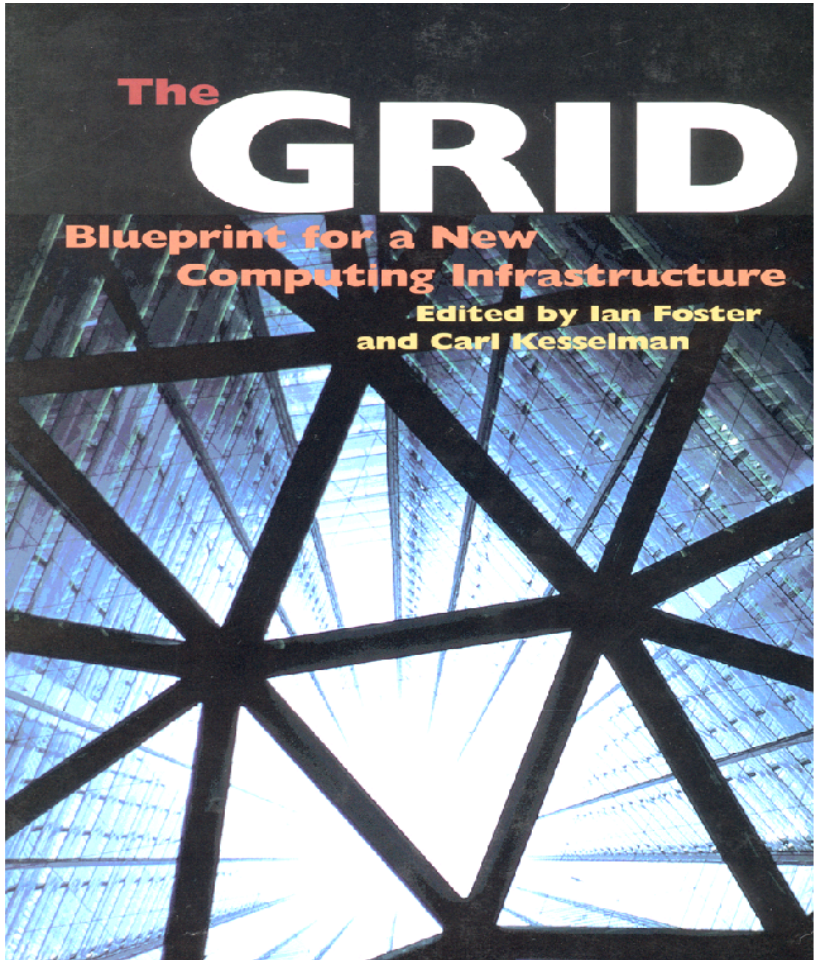
Jennifer M. Schopf[1,2], Ann Chervenak[3], Ian Foster[1,2,4], Dan Fraser[1,2], Dan Gunter[5], Nick LeRoy[6], Brian Tierney[5]

[1] Computation Institute, University of Chicago and Argonne National Laboratory
[2] Mathematics and Computer Science Division, Argonne National Laboratory
[3] Information Sciences Institute, University of Southern California
[4] Department of Computer Science, University of Chicago
[5] Lawrence Berkeley National Laboratory
[6] Department of Computer Science, University of Wisconsin

- Failure detection techniques including online monitoring and operation retry to detect and recover from multiple failure modalities
- A need for predictability and coordinated scheduling in spite of variations in load and competing use of storage space, bandwidth to the storage system, and network bandwidth
  - Failure detection and performance prediction are considered open areas of research by many

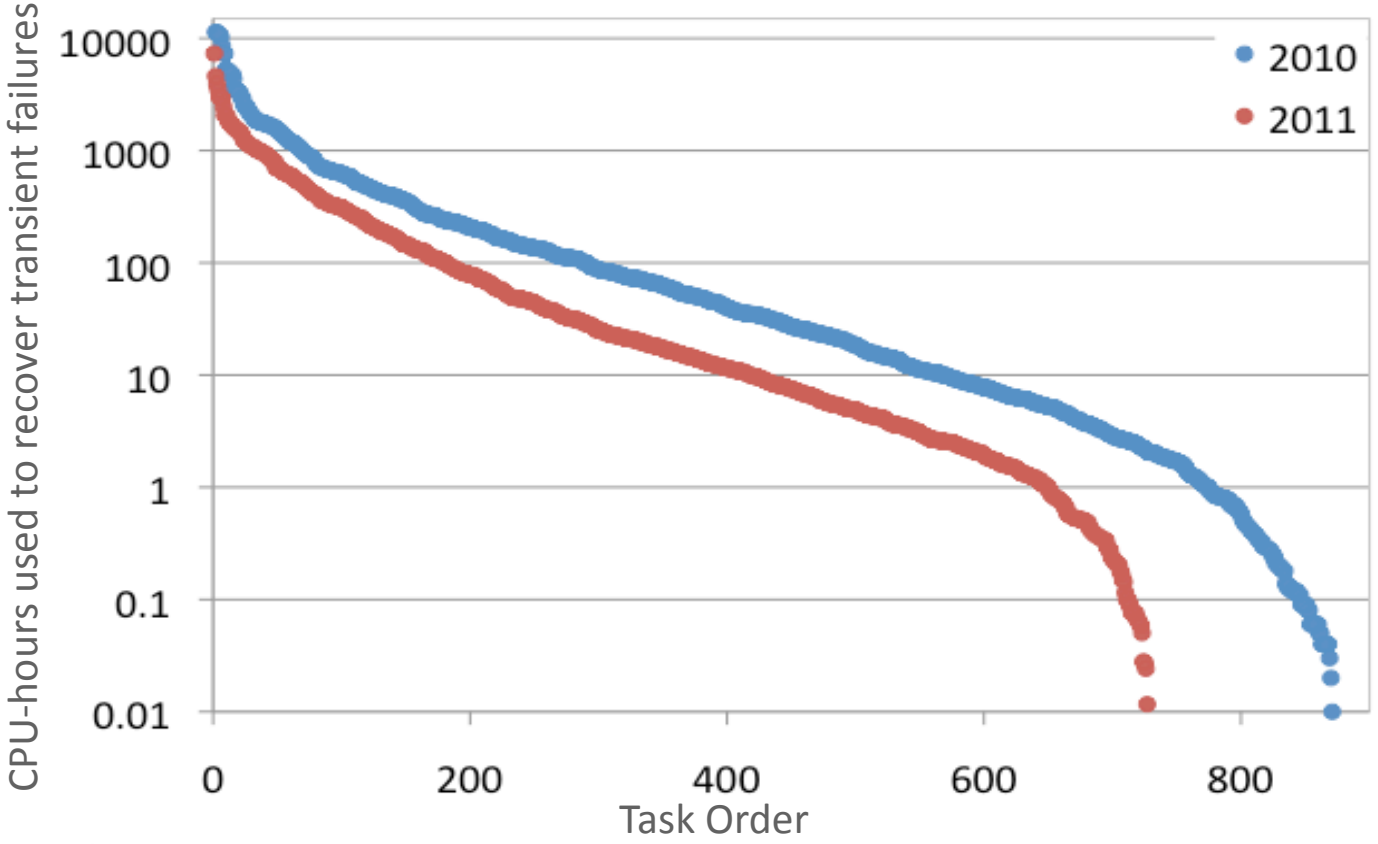
Reliability Engineering for the Grid

- ATLAS experience has shown that Grid failures can occur for a variety of reasons
  - Grid heterogeneity makes failures hard to diagnose and repair quickly
    - ATLAS data processing on the Grid must tolerate a continuous stream of failures errors and faults
- While many fault-tolerance mechanisms improve the reliability of ATLAS data processing in the Grid, their benefits come at costs
- Reliability Engineering provides a framework for fundamental understanding of the ATLAS petascale data processing on the Grid, which is not a desirable enhancement but a necessary requirement



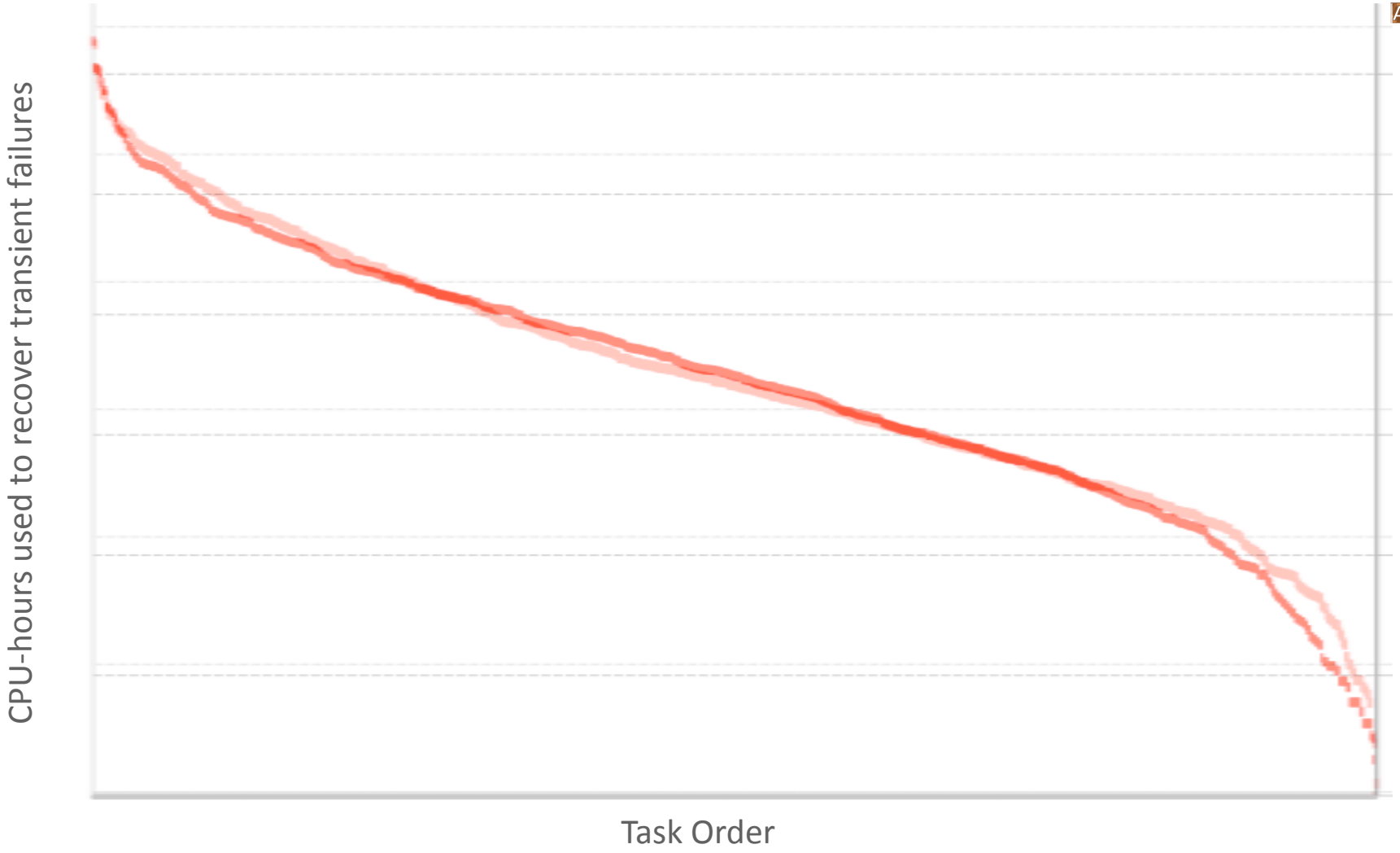
ATLAS Failure Recovery Cost

- Job resubmission avoids data loss at the expense of CPU time used by the failed jobs
  - Distribution of tasks ordered by CPU time used to recover transient failures is not uniform:
    - Most of CPU time required for recovery was used in a small fraction of tasks



- In 2010 reprocessing, the CPU time used to recover transient failures was 6% of the CPU time used for reconstruction
- In 2011 reprocessing, the CPU time used to recover transient failures was reduced to 4% of the CPU time used for the reconstruction

2010 vs. 2011: Universal Behavior



Waloddi Weibull

1939 Weibull published his paper on Weibull distribution in probability theory and statistics

1941 Weibull received a personal research professorship in Technical Physics at the Royal Institute of Technology in Stockholm from the arms producer Bofors

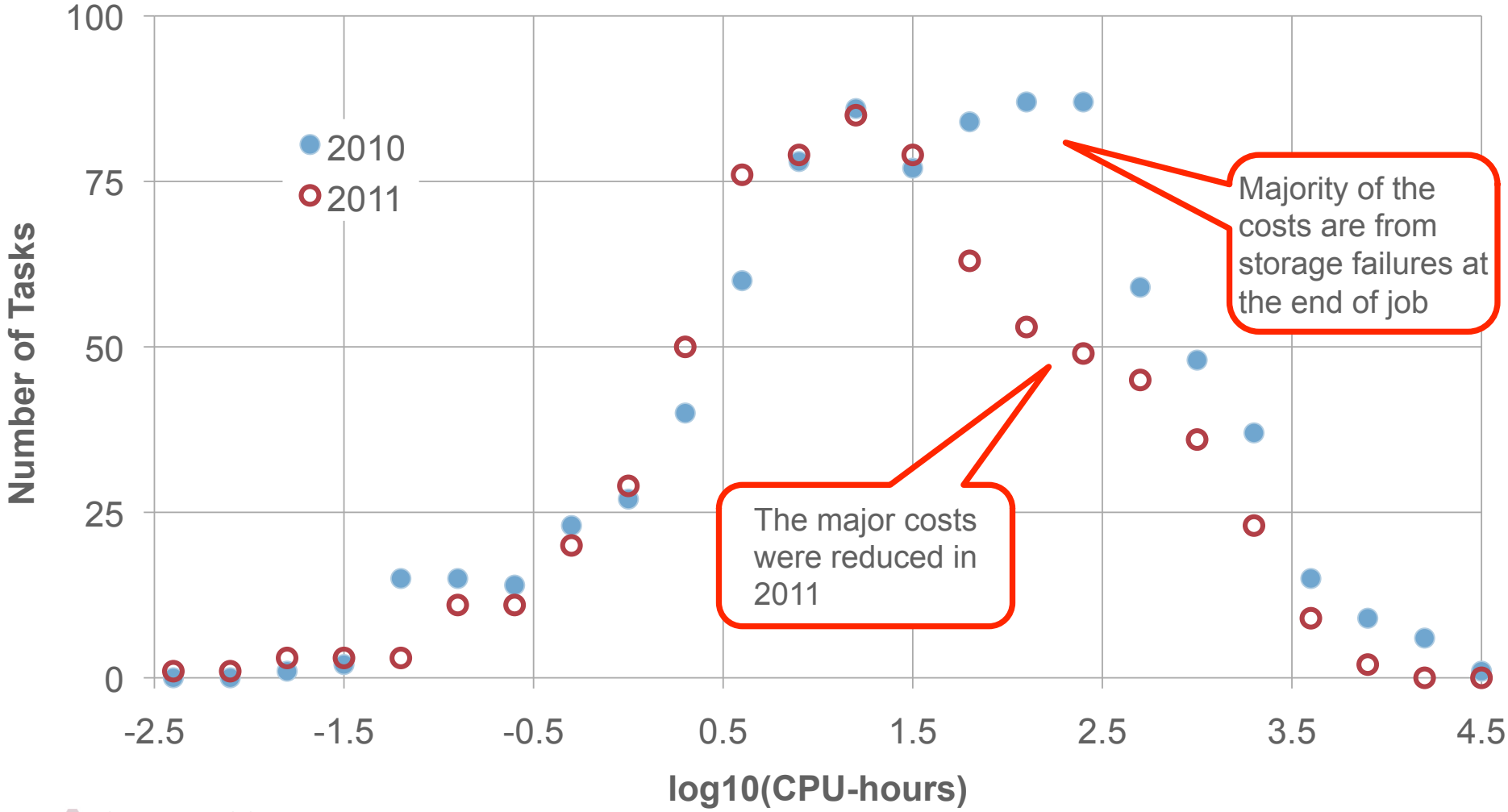
1951 Weibull presented his most famous paper to the American Society of Mechanical Engineers (ASME) on Weibull distribution, using seven case studies

1972 American Society of Mechanical Engineers awarded Weibull the Gold Medal

1978 Royal Swedish Academy of Engineering Sciences awarded Weibull the Great Gold Medal

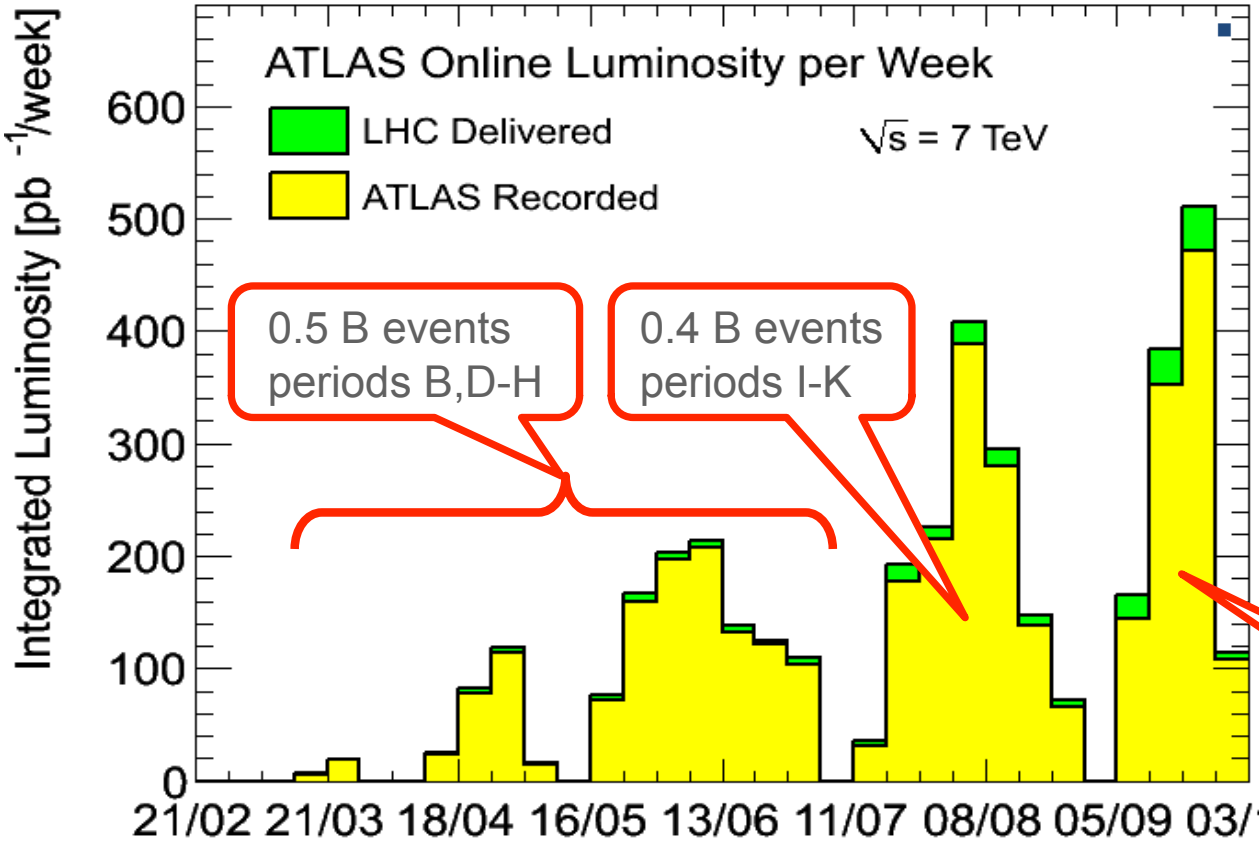
- The Weibull distribution is by far the world's most popular statistical model for production data

Bi-modal Weibull Distribution: CPU-time Used to Recover Job Failures in ATLAS Data Reprocessing



Two Parts in ATLAS 2011 Reprocessing

- 2011 reprocessing campaign of more than 0.9 B pp events was split in two parts
  - Part I reprocessed the dataset taken 3/21-6/29, Part II - data taken 7/13-8/22
    - Each part has about 1.25 fb^{-1} of integrated luminosity



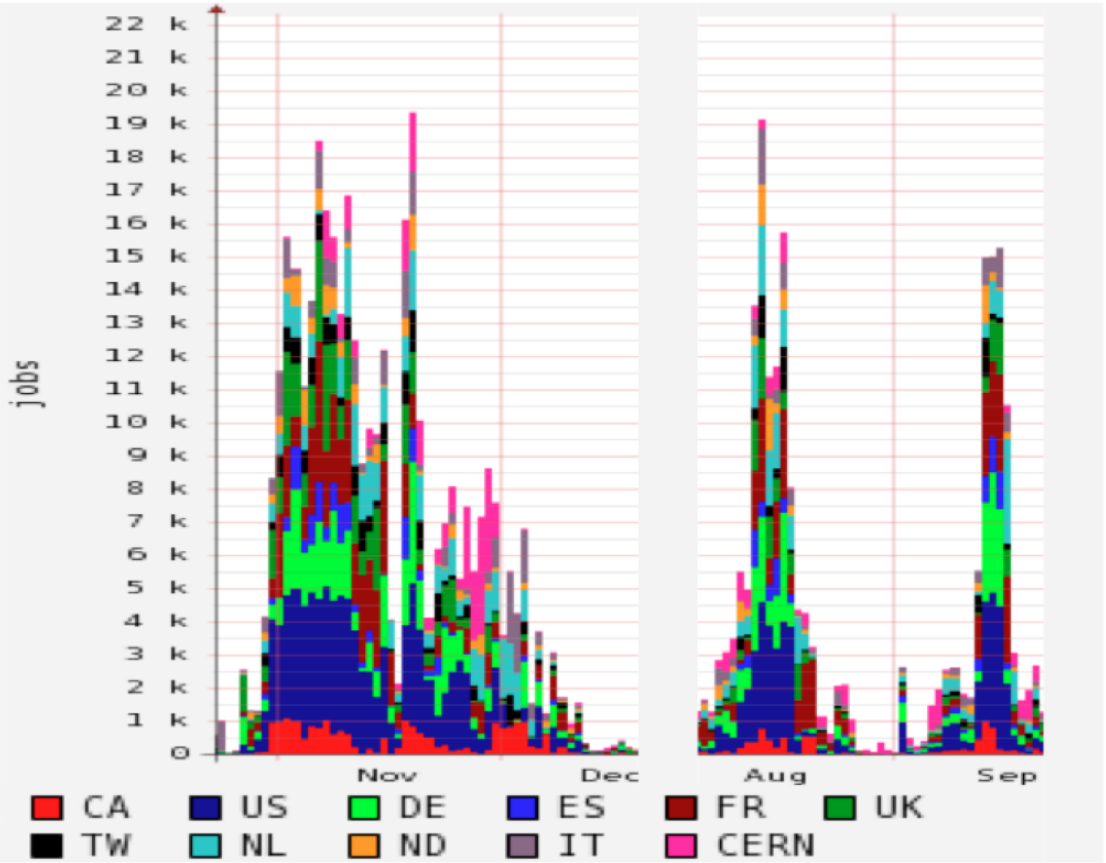
Reprocessing provided large scale validation of the new software release assuring its smooth deployment at Tier-0, so that the end of 2011 pp datataking was in sync with the reprocessed dataset to be used for physics analysis for winter conferences and papers

Detailed Look at 2011 Reprocessing

Representative Task Shows that Most CPU Costs Caused by Storage Failures at the End of the Job

PandaID, Owner, Working group	Job	Status	Created	Time to start	Duration	Ended/ Modified	Cloud/Site, Type	Priority
1312824903 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000179.job 1	failed	2011-09-14 19:51	9:29:57	5:54:40	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: LFC setup and mkdir failed: LFC_HOST=atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/log/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.log.r2713_tid512594_00: Name server not activelLog put error: LFC setup and mkdir failed:							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312824904 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000180.job 1	failed	2011-09-14 19:51	9:29:59	5:54:39	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: LFC setup and mkdir failed: LFC_HOST=atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/log/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.log.r2713_tid512594_00: Name server not activelLog put error: LFC setup and mkdir failed:							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312824986 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000262.job 1	failed	2011-09-14 19:51	9:32:00	5:52:35	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: /opt/lcg/bin/lcg-cr lcg_util-1.7.6-2 GFAL-client-1.11.8-3 Using grid catalog type: lfc Using grid catalog : atlas-lfc-fzk.gridka.de Checksum type: None SE type: SRMv2 Destination SURL : srm://atlassrm-fzk.gridka.de:8443/srm/managerv2?SFN=/pnfs/gridka							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312825012 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000288.job 1	failed	2011-09-14 19:51	9:35:55	5:49:33	09-15 11:16	DE/DE.FZK-LCG2, production	810
	Error details: pilot: Put error: ysics_JetTauEtmiss.recon.ESD.r2713_tid512594_00/ESD.512594._000288.pool.root.1: Registration failed, please register it by hand, when the problem will be solved lcg_cr: Communication error on send guid:0EF44EF4-5CDF-E011-A432-001A6478950C							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312825126 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000399.job 1	failed	2011-09-14 19:51	9:40:00	5:44:31	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: LFC setup and mkdir failed: LFC_HOST=atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/log/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.log.r2713_tid512594_00: Name server not activelLog put error: LFC setup and mkdir failed:							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312829438 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._000824.job 1	failed	2011-09-14 20:00	9:48:33	5:26:57	09-15 11:16	DE/DE.FZK-LCG2, production	810
	Error details: pilot: Put error: LFC setup and mkdir failed: LFC_HOST atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/ESD/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.ESD.r2713_tid512594_00: Name server not active							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312833405 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._001048.job 1	failed	2011-09-14 20:10	9:49:17	5:15:53	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: LFC setup and mkdir failed: LFC_HOST=atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/log/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.log.r2713_tid512594_00: Name server not activelLog put error: LFC setup and mkdir failed:							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							
1312833444 lucotte@lpsc.in2p3.fr **Reprocessing**	data11_7TeV.00186923.physics_JetTauEtmiss.recon.r2713_tid512594._001087.job 1	failed	2011-09-14 20:10	9:51:17	5:13:48	09-15 11:15	DE/DE.FZK-LCG2, production	810
	Error details: pilot: LFC setup and mkdir failed: LFC_HOST=atlas-lfc-fzk.gridka.de cannot create /grid/atlas/dq2/data11_7TeV/log/r2713 /data11_7TeV.00186923.physics_JetTauEtmiss.recon.log.r2713_tid512594_00: Name server not activelLog put error: LFC setup and mkdir failed:							
	In: data11_7TeV.00186923.physics_JetTauEtmiss.merge.RAW **Out:** data11_7TeV.00186923.physics_JetTauEtmiss.recon.HIST.r2713_tid512594_00							

Time Overhead



- It takes about three million core-hours to processes one petabyte of ATLAS data
- Transient job failures and retries delay the reprocessing duration
- Optimization of ATLAS Grid Data Processing workflow and other improvements cut the “tails” and halved the duration of the petabyte-scale reprocessing on the Grid from almost two months in 2010 to less than four weeks in 2011

- Optimization of fault-tolerance techniques vs. time overhead (“tails”) induced in task completion is an active area of research in ATLAS Grid Data Processing

Conclusions

- In ATLAS Grid Data Processing physicists deal with datasets, not individual files
  - A task (comprised of many jobs) became a unit of ATLAS Grid Data Processing
- Reliability Engineering provides a framework for fundamental understanding of ATLAS Grid Data Processing , which is not a desirable enhancement but a necessary requirement
  - Improvements in ATLAS Grid Data Processing reduced the cost of the automatic failed jobs re-tries to the level below 5% of total CPU-hours used
    - It is no longer a problem
  - Fault-tolerance achieved through automatic re-tries of the failed jobs induces a time overhead in the task completion, which is difficult to predict
    - Reduction in the duration of Grid Data Processing tasks is an active area of research in ATLAS