Reliability Engineering Analysis of ATLAS Data Reprocessing Campaigns

A. Vaniachine, D. Golubkov and D. Karpenko
on behalf of the ATLAS Collaboration

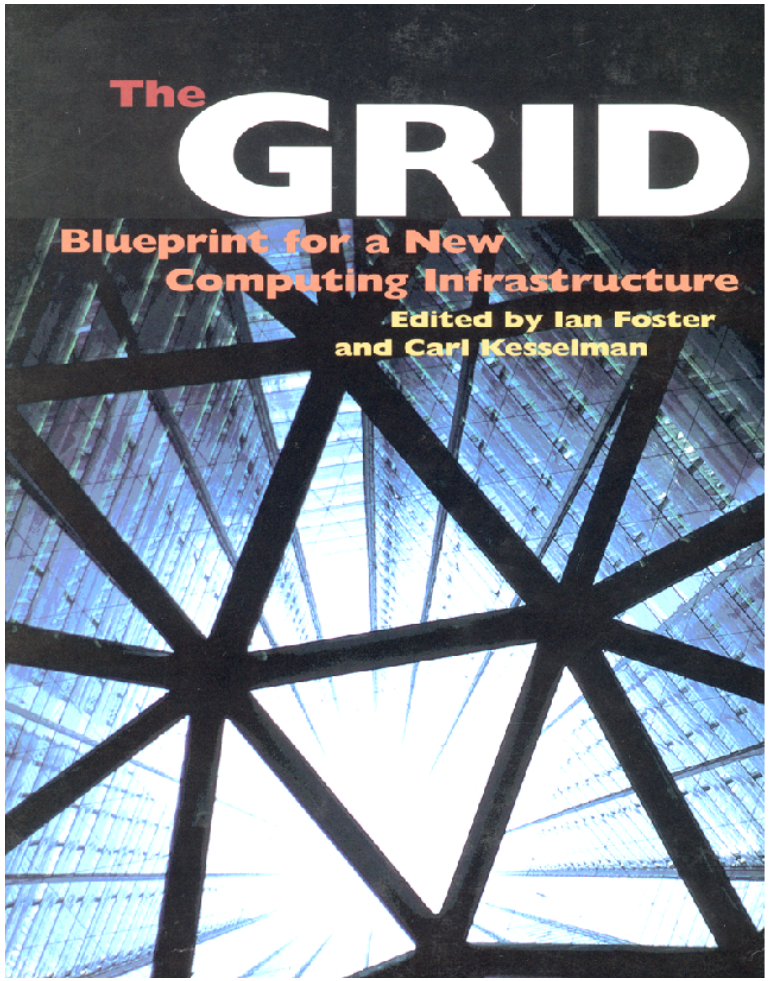
XX International Conference on Computing in High Energy and Nuclear Physics
October 14-18, 2013
Amsterdam, The Netherlands

Introduction

- Scheduled LHC upgrades will increase the data taking rates tenfold, which increases demands on computing resources
  - However, a tenfold increase in WLCG resources for LHC upgrade needs is not an option
- The ATLAS experiment needs to exercise due diligence in evolving its Computing Model to optimally use the required resources
  - A comprehensive end-to-end solution for the composition and execution of the reprocessing workflow within given CPU and storage constraints is necessary
- During three years of LHC data taking, the ATLAS collaboration completed three petascale data reprocessing campaigns on the Grid, with up to 2 PB of “raw” data being reprocessed every year
  - We present the Reliability Engineering analysis of ATLAS data reprocessing campaigns providing the foundation needed to scale up data reprocessing beyond petascale

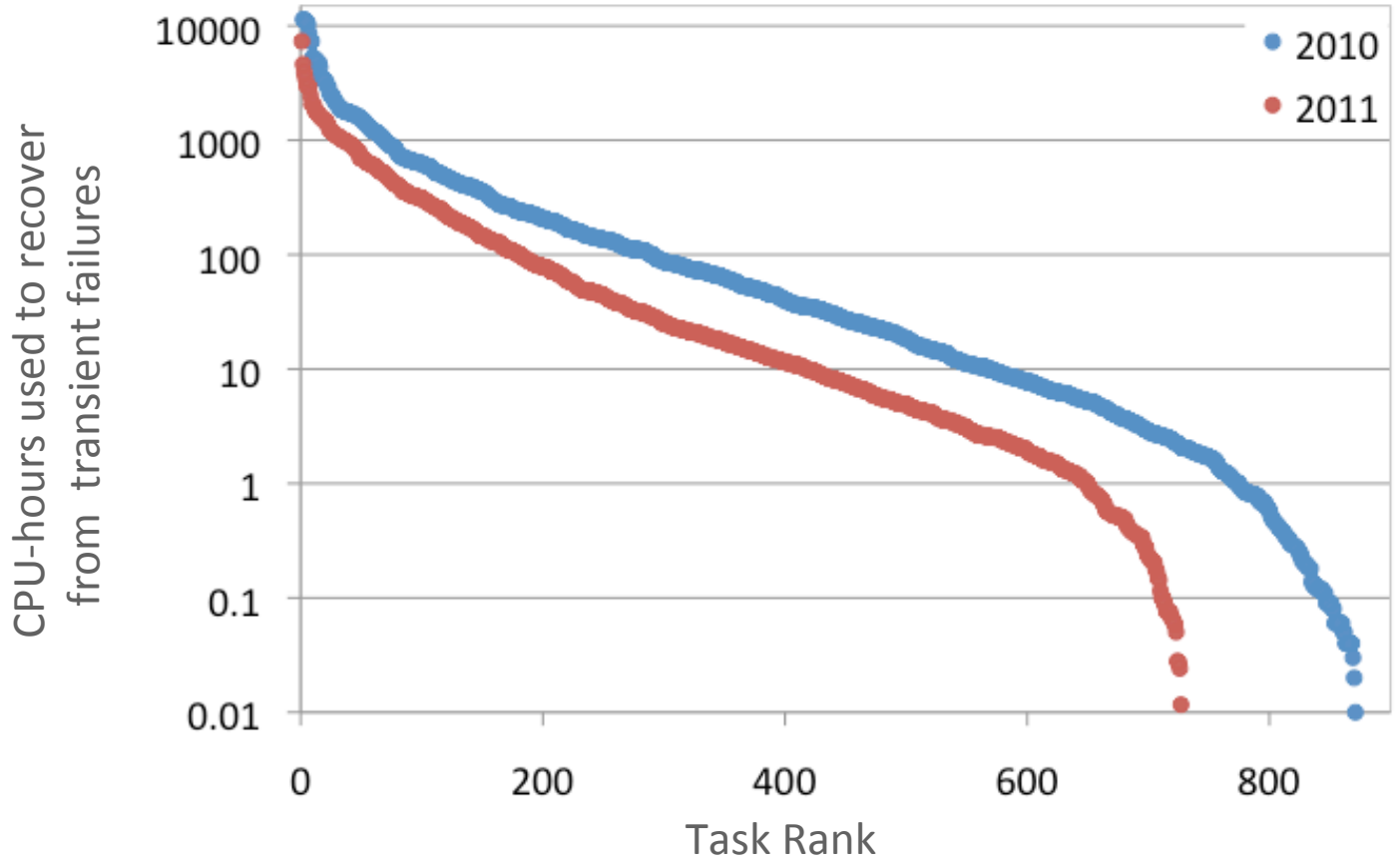
Reliability Engineering on the Grid

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



CHEP2012: Costs of Recovery from Failures

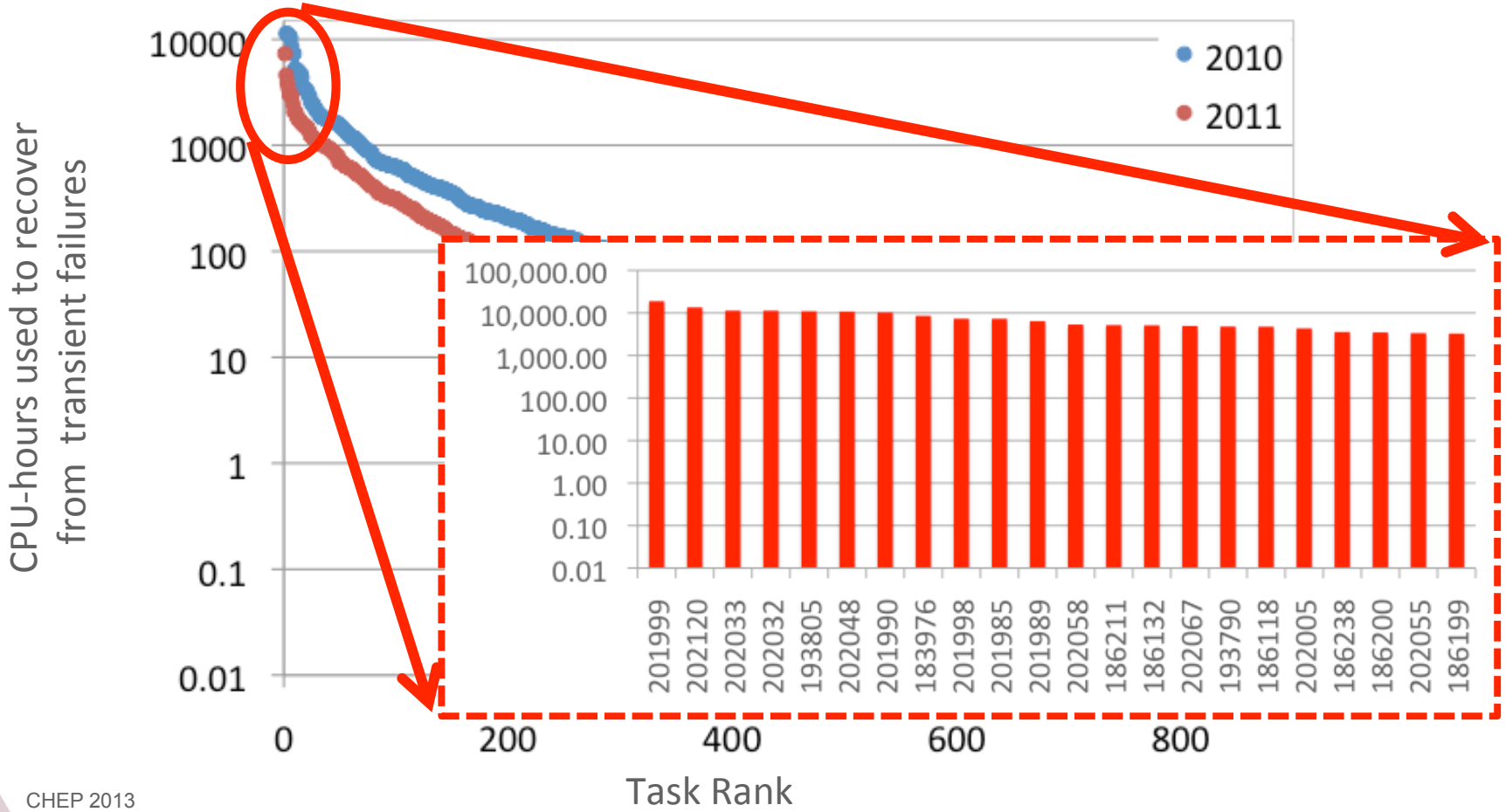
- Job re-tries avoids data loss at the expense of CPU time used by the failed jobs
  - Distribution of tasks[1] ranked by CPU time used to recover from transient failures is not uniform:
    - Most of CPU time required for recovery was used in a small fraction of tasks



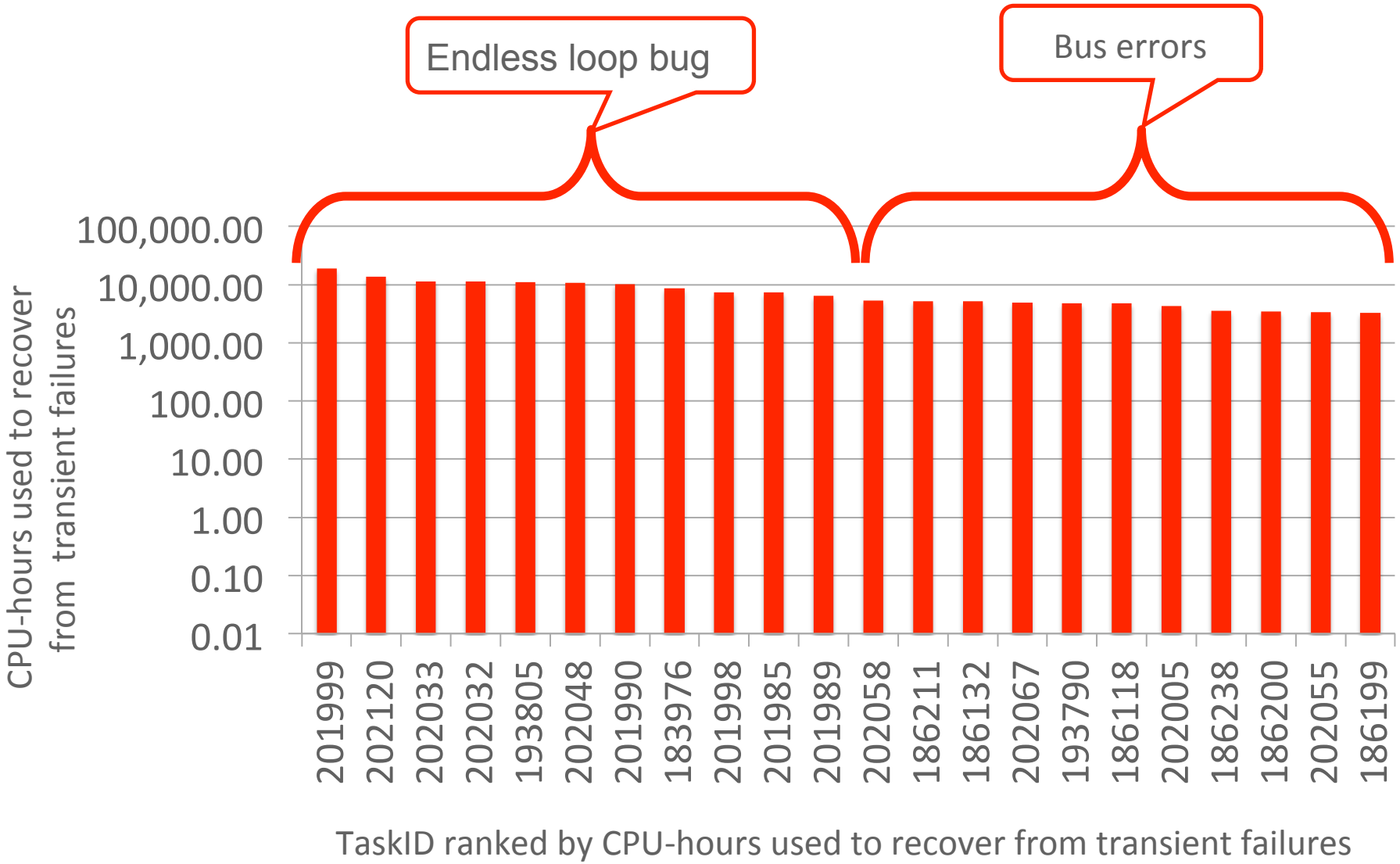
[1]In ATLAS data reprocessing jobs from the same run are processed in the same task

Zoom at Tasks with Most CPU-hours Used

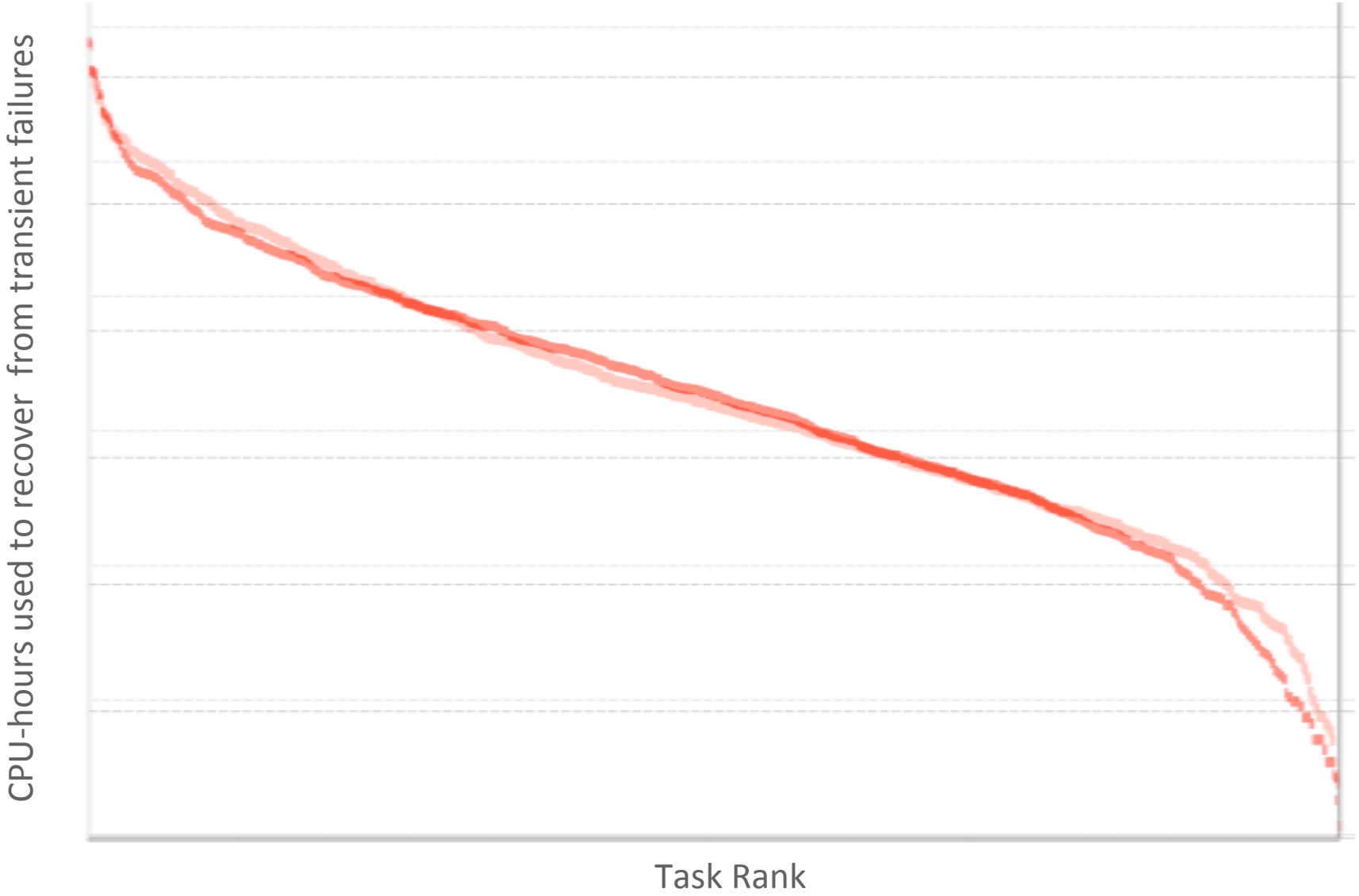
- Job re-tries avoids data loss at the expense of CPU time used by the failed jobs
  - Distribution of tasks ranked by CPU time used to recover from transient failures is not uniform:
    - Most of CPU time required for recovery was used in a small fraction of tasks



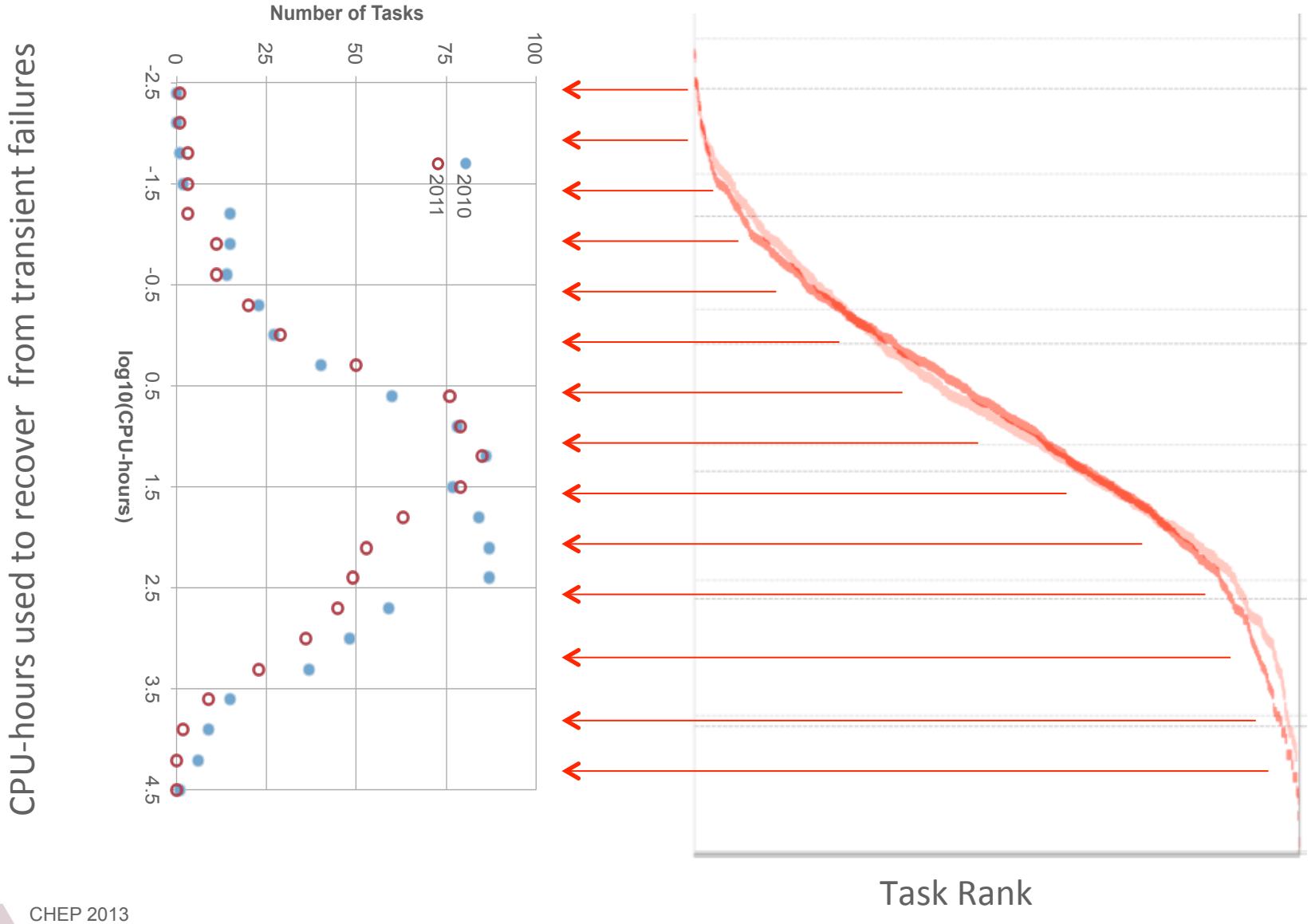
2010: Highest Costs of Recovery from Failures



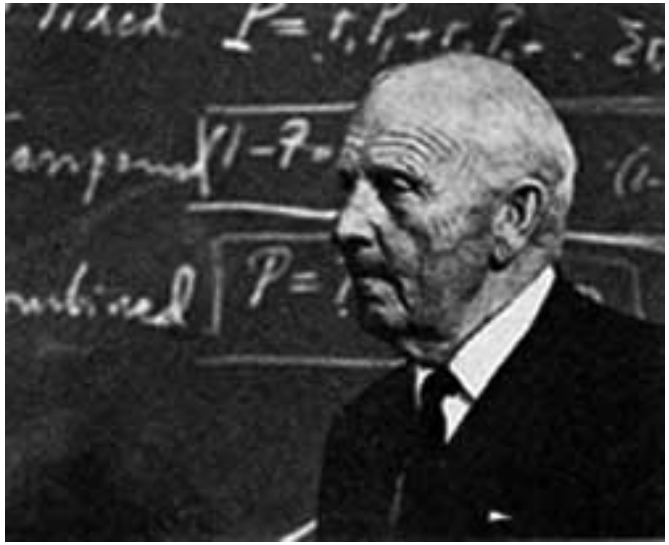
2010 vs. 2011: Universal Behavior



Universal Behavior Indicates Weibull Distribution



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

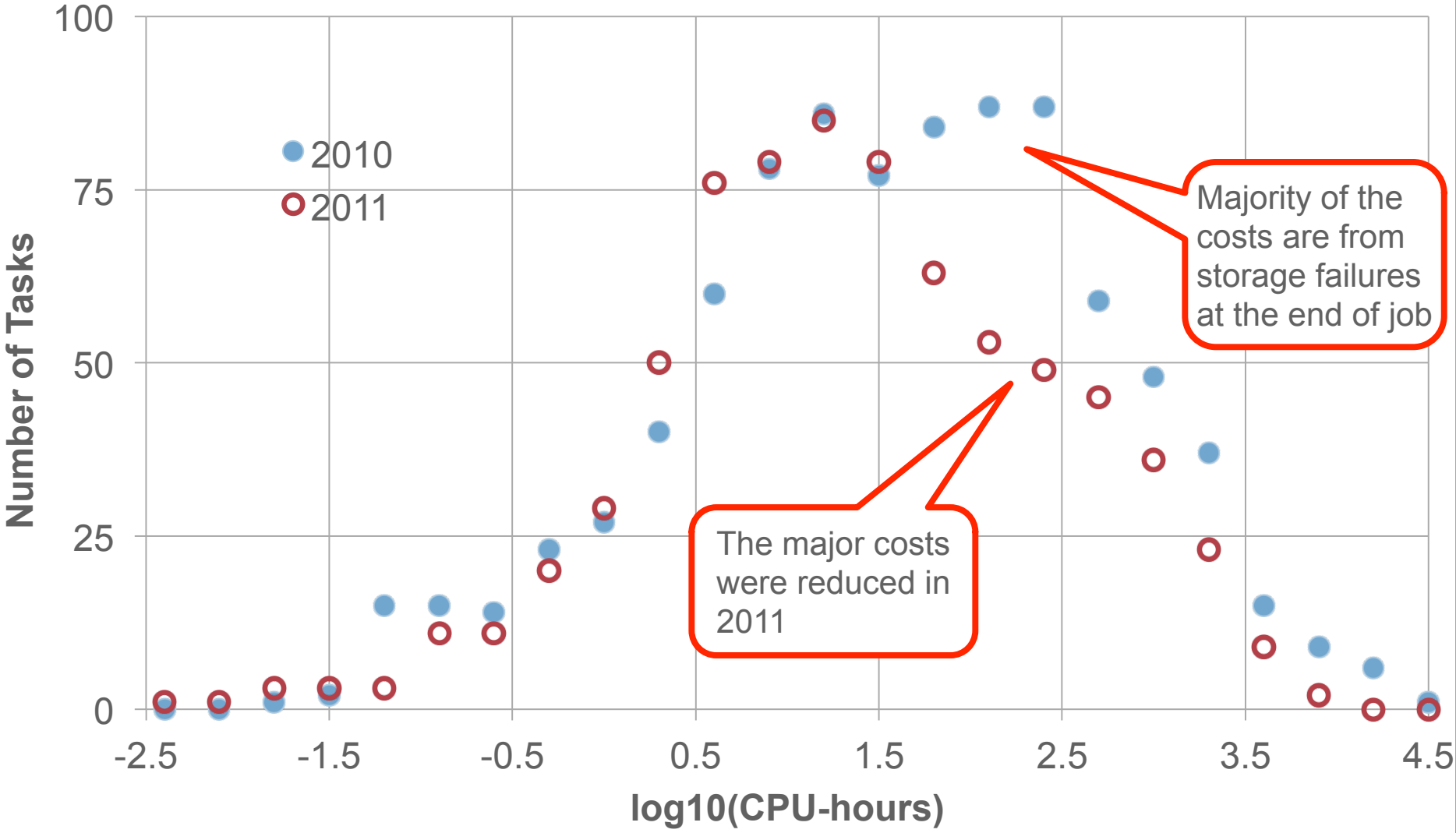
1972 American Society of Mechanical Engineers awarded Weibull the Gold Medal

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

$$f(x) = \frac{k}{\lambda}\left(x/\lambda\right)^{k-1}\exp\left(-\left(x/\lambda\right)^{k}\right)$$

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

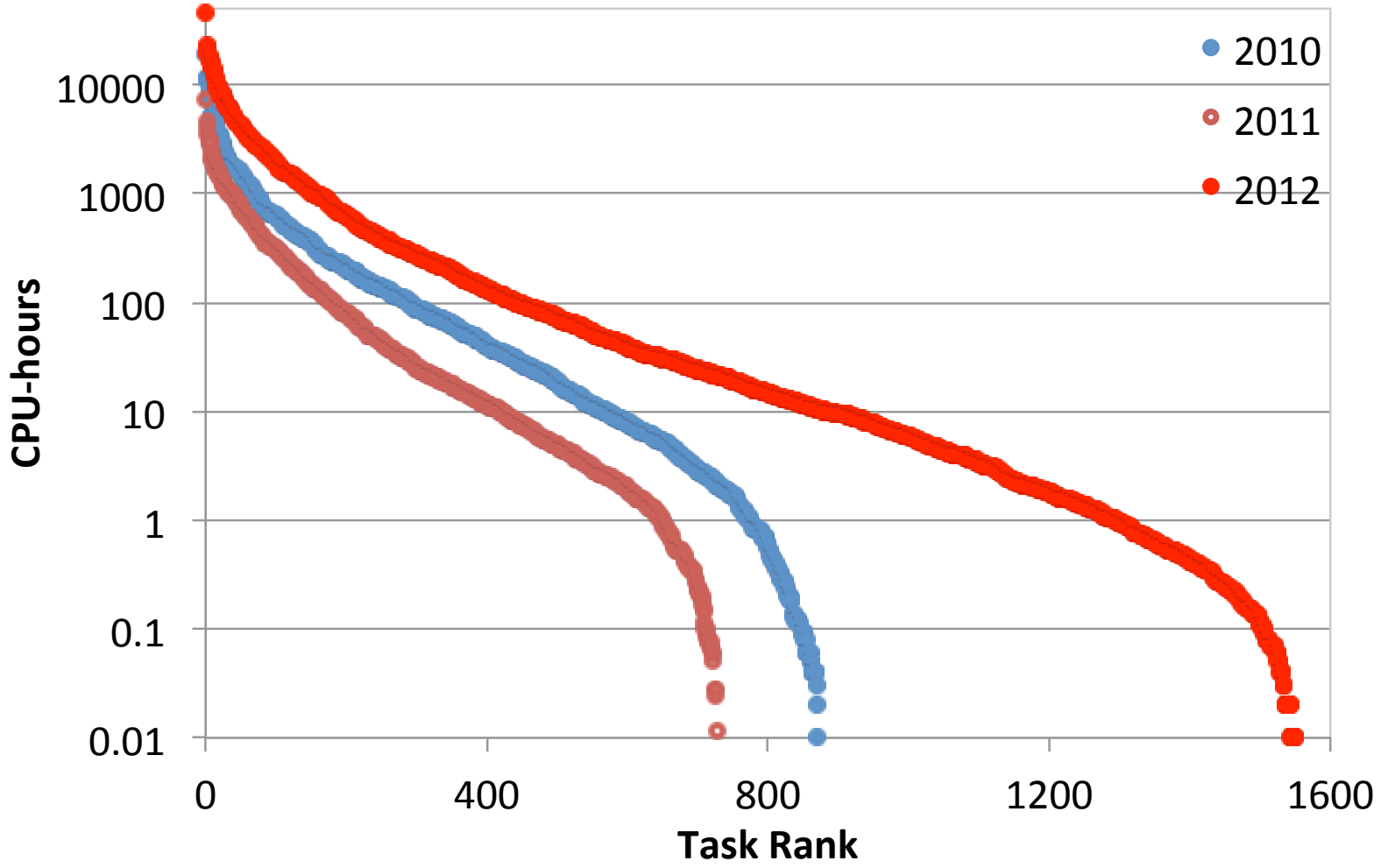
CPU-time Used to Recover from Job Failures in Data Reprocessing: Bi-modal Weibull Distribution



Cost of Recovery from Failures in Reprocessing

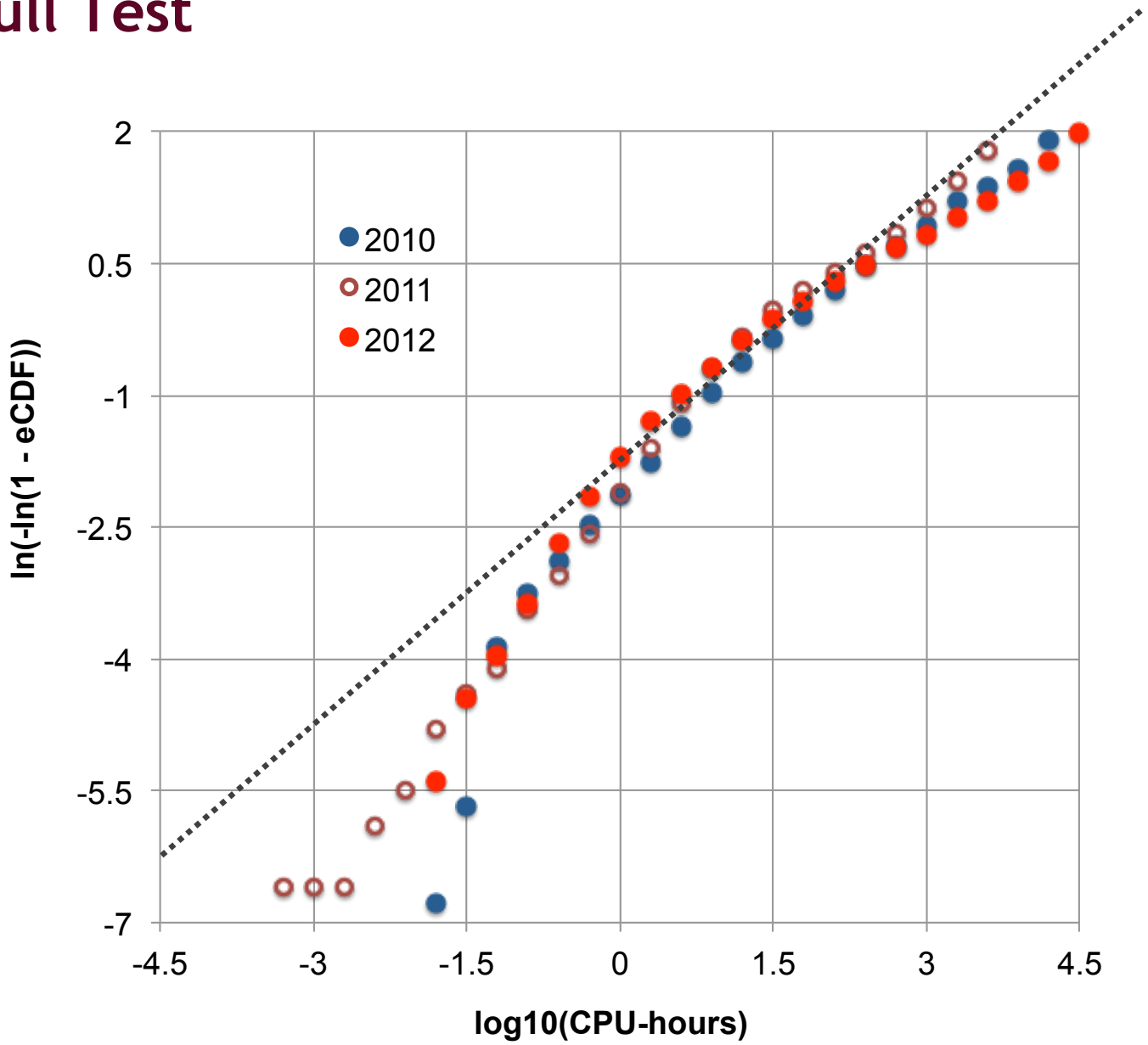
Reprocessing campaign	Input Data Volume (PB)	CPU Time Used for Reconstruction ($10^6 h$)	Fraction of CPU Time Used for Recovery (%)
2010	1	2.6	6.0
2011	1	3.1	4.2
2012	2	14.6	5.6

Same Behavior Reproduced in 2012 Reprocessing

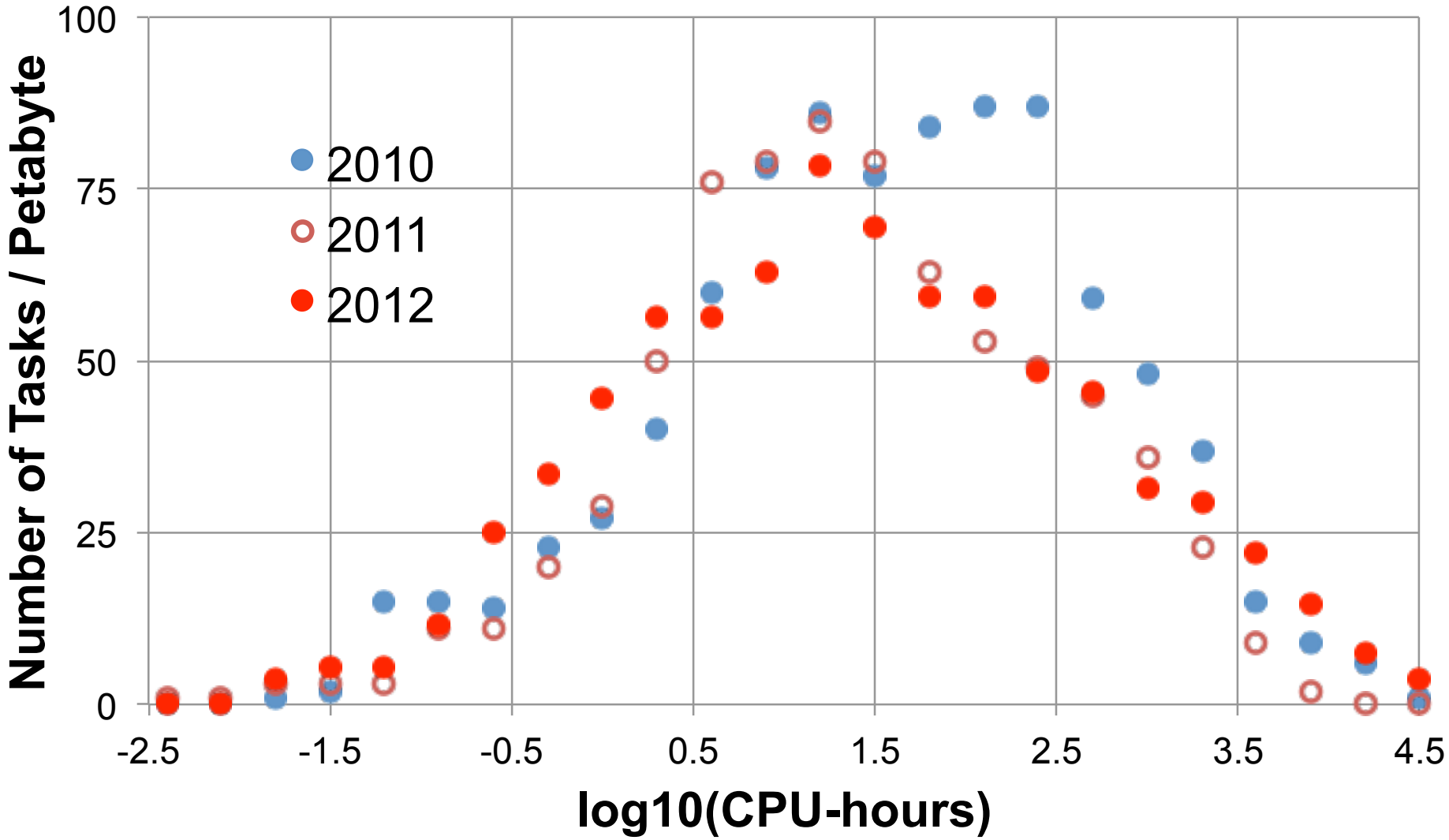


– There were more tasks in 2012 reprocessing of 2 PB of 2012 p-p data

Weibull Test



CPU-time Used to Recover from Job Failures



Log-Normal Behavior

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)$$

- The distribution symmetry indicates the log-normal behaviour rather than a skewed Weibull distribution
  - However, the overall log-normal behaviour of the distribution of tasks vs. CPU-hours used for failure recovery is unexpected
    - Although, the log-normal distributions are often observed in Reliability Engineering analysis of failures, such as mean-time-to-failure of the Grid hardware, these are expected when some multiplicative process is present, such as multiplicative degradation (Kolmogorov, 1941)
- We were unable to identify the multiplicative process describing our case where the CPU-time used for recovery is added upon each job retry
  - Thus, we compared our data with two necessary conditions under which the additive process may result in the log-normal distribution:
    - Mouri H 2013 *Preprint* arXiv:1309.5709

Conditions Check

$$\left|\frac{\langle(\ln z_* - \langle \ln z_* \rangle)^3\rangle}{\langle(\ln z_* - \langle \ln z_* \rangle)^2\rangle^{1.5}}\right| < \left|\frac{\langle(z_* - \langle z_* \rangle)^3\rangle}{\langle(z_* - \langle z_* \rangle)^2\rangle^{1.5}}\right| \quad (1)$$

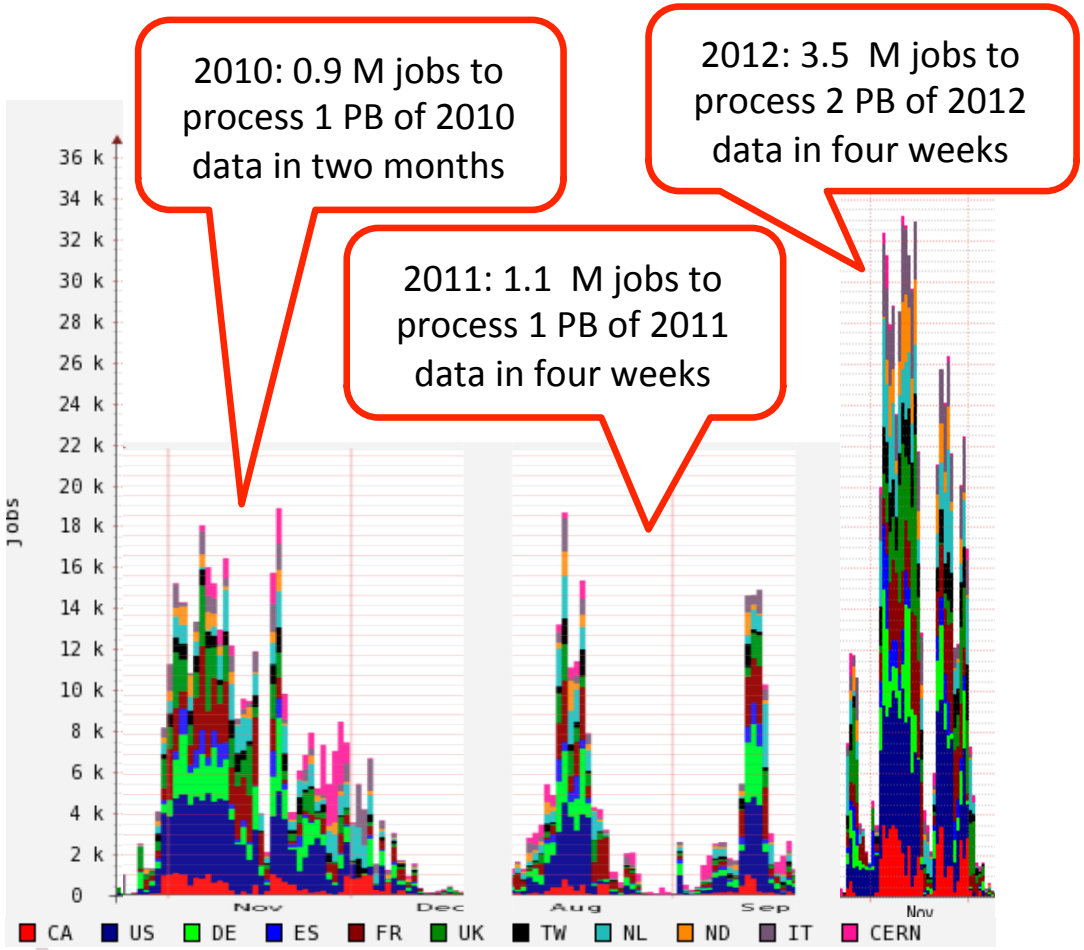
$$\left|\frac{\langle(\ln z_* - \langle \ln z_* \rangle)^4\rangle - 3\langle(\ln z_* - \langle \ln z_* \rangle)^2\rangle^2}{\langle(\ln z_* - \langle \ln z_* \rangle)^2\rangle^2}\right| < \left|\frac{\langle(z_* - \langle z_* \rangle)^4\rangle - 3\langle(z_* - \langle z_* \rangle)^2\rangle^2}{\langle(z_* - \langle z_* \rangle)^2\rangle^2}\right| \quad (2)$$

- Here z_* is the random variable
  - in our case it is the CPU-hours used to recover from transient job failures
- Table below shows that conditions (1) and (2) are met in our case:
  - for all reprocessing campaigns the right-sides are much greater than the left-sides

Reprocessing campaign	Condition (1)		Condition (2)	
	Left-side	Right-side	Left-side	Right-side
2010	0.27	7.90	0.23	79.6
2011	0.26	7.17	0.31	74.0
2012	0.06	10.1	0.46	153.5

Duration of Reprocessing Campaigns

- Transient job failures and retries delay the reprocessing duration
  - Speeding up the completion of reprocessing is an active area of research



- Optimization of ATLAS Grid Data Processing workflow and other improvements cut the delays 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
- To assure timely results for the 2012 Moriond Conference we reprocessed twice more data within the same time period as in 2011 reprocessing, with each 2012 event taking twice longer to reconstruct than the 2011 event

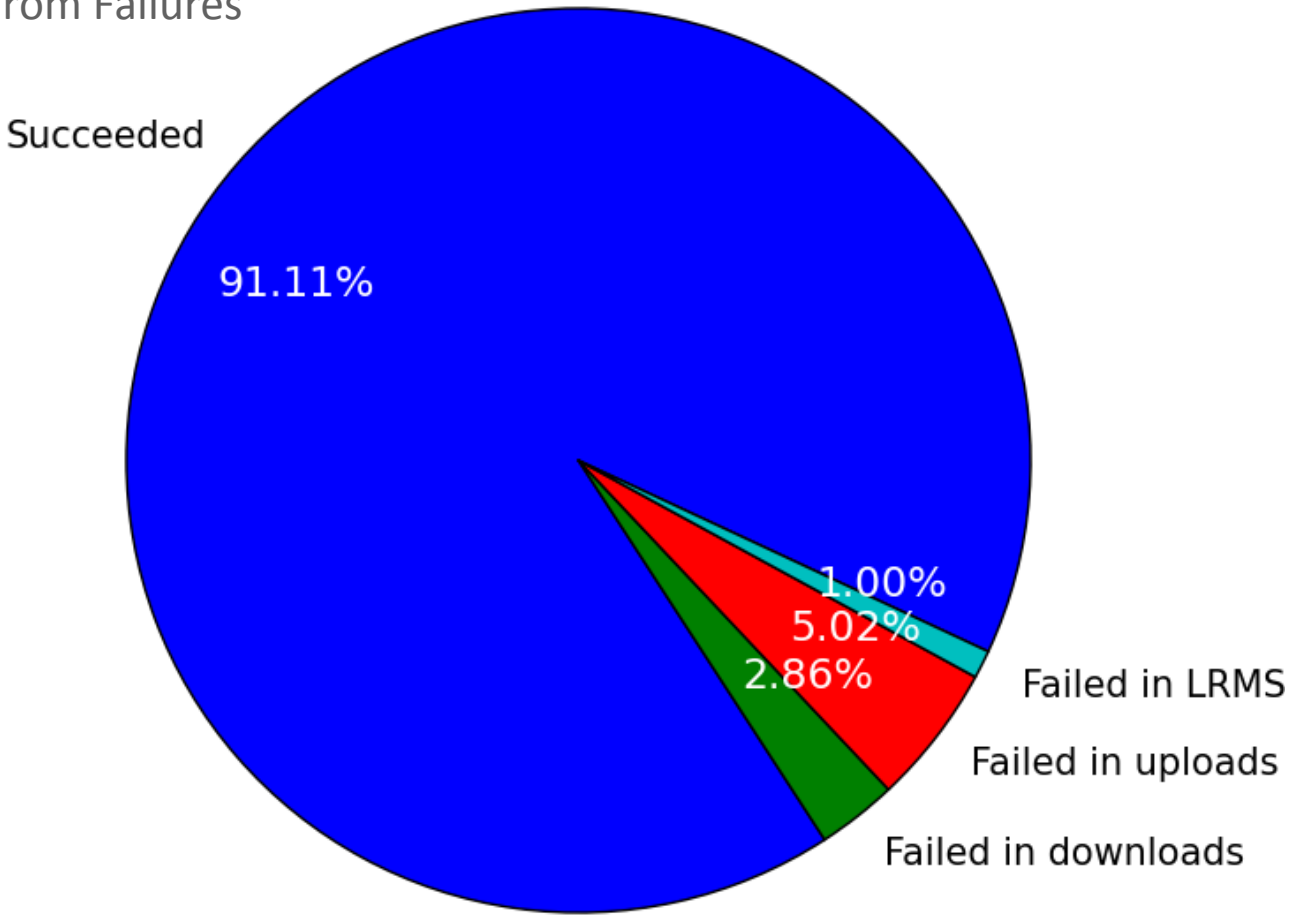
Conclusions

- Reliability Engineering provides a framework for fundamental understanding of data reprocessing workflow on the Grid
  - ATLAS data reprocessing keeps the cost of automatic re-tries of the failed jobs at the level of 4-6% of total CPU-hours used for data reconstruction
- Fault-tolerance achieved through automatic re-tries of the failed jobs induces a time overhead in the task completion, which is difficult to predict
  - Optimization of data reprocessing duration is an active area of research
- The ATLAS experiment needs to exercise due diligence in evolving its Computing Model to optimally use the required resources
  - A comprehensive end-to-end solution for the composition and execution of the reprocessing within given CPU and storage constraints is necessary

Extra Materials

Representative Causes for Job Re-tries

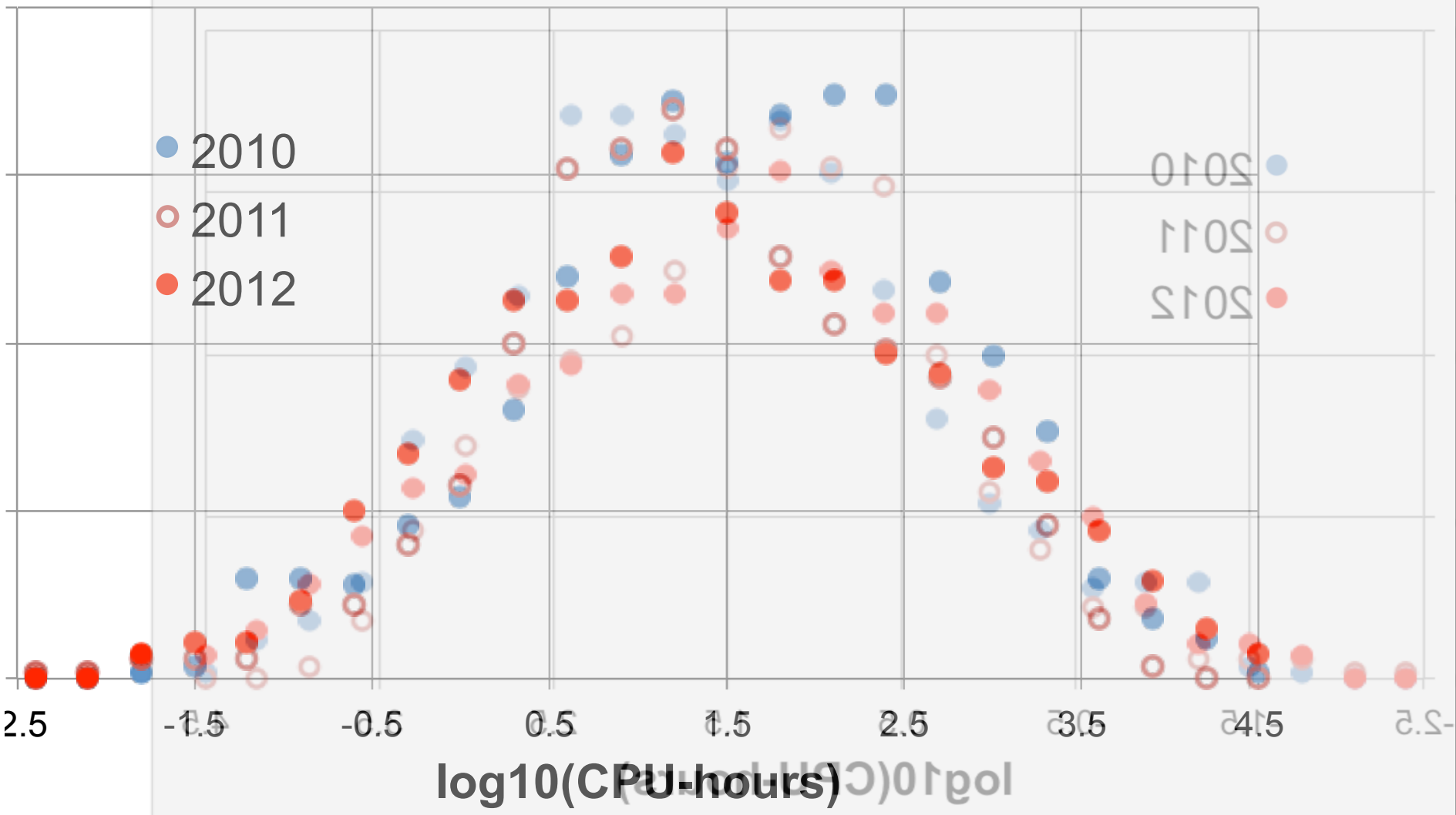
- Storage Failures at the End of the Job (Uploads) Increase the CPU Cost of Job Recovery from Failures



Representative Task with CPU Costs from Job Re-tries Caused by Storage Failures at the End of 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							

Symmetry Indicates Log-Normal Behaviour



Kolmogorov-Smirnov Test Prefers Log-Normal

Log-normal distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)$$

Reprocessing campaign	μ	σ	K-S test
2010	3.29±0.09	2.63±0.06	0.07
2011	2.77±0.09	2.50±0.07	0.05
2012	2.92±0.07	2.90±0.05	0.04

Weibull distribution:

$$f(x) = \frac{k}{\lambda}(x/\lambda)^{k-1} \exp\left(-(x/\lambda)^k\right)$$

Reprocessing campaign	k	λ	K-S test
2010	0.41±0.01	96.6±8.3	0.05
2011	0.43±0.01	54.5±4.9	0.10
2012	0.35±0.01	79.4±6.0	0.08