Characterizing a High Throughput Computing Workload: The Compact Muon Solenoid (CMS) Experiment at LHC

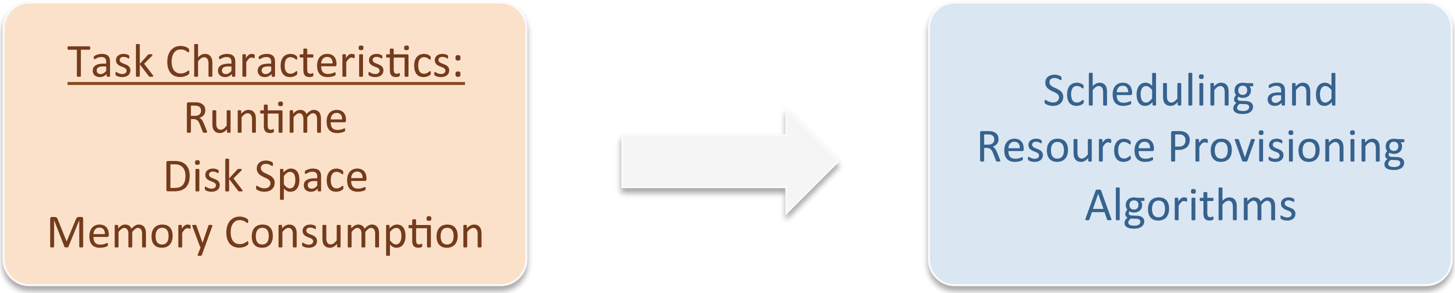
Rafael Ferreira da Silva[1], Mats Rynge[1], Gideon Juve[1], Igor Sfiligoi[2,]
Ewa Deelman[1], James Letts[1], Frank Würthwein[2], Miron Livny[3]

International Conference on Computational Science
June 1-3, 2015, Reykjavik, Iceland

[1] University of Southern California, Information Sciences Institute, Marina Del Rey, CA, USA
[2] University of California at San Diego, Department of Physics, La Jolla, CA, USA
[3] University of Wisconsin Madison, Madison, WI, USA

USC Viterbi
School of Engineering
Information Sciences Institute

Introduction



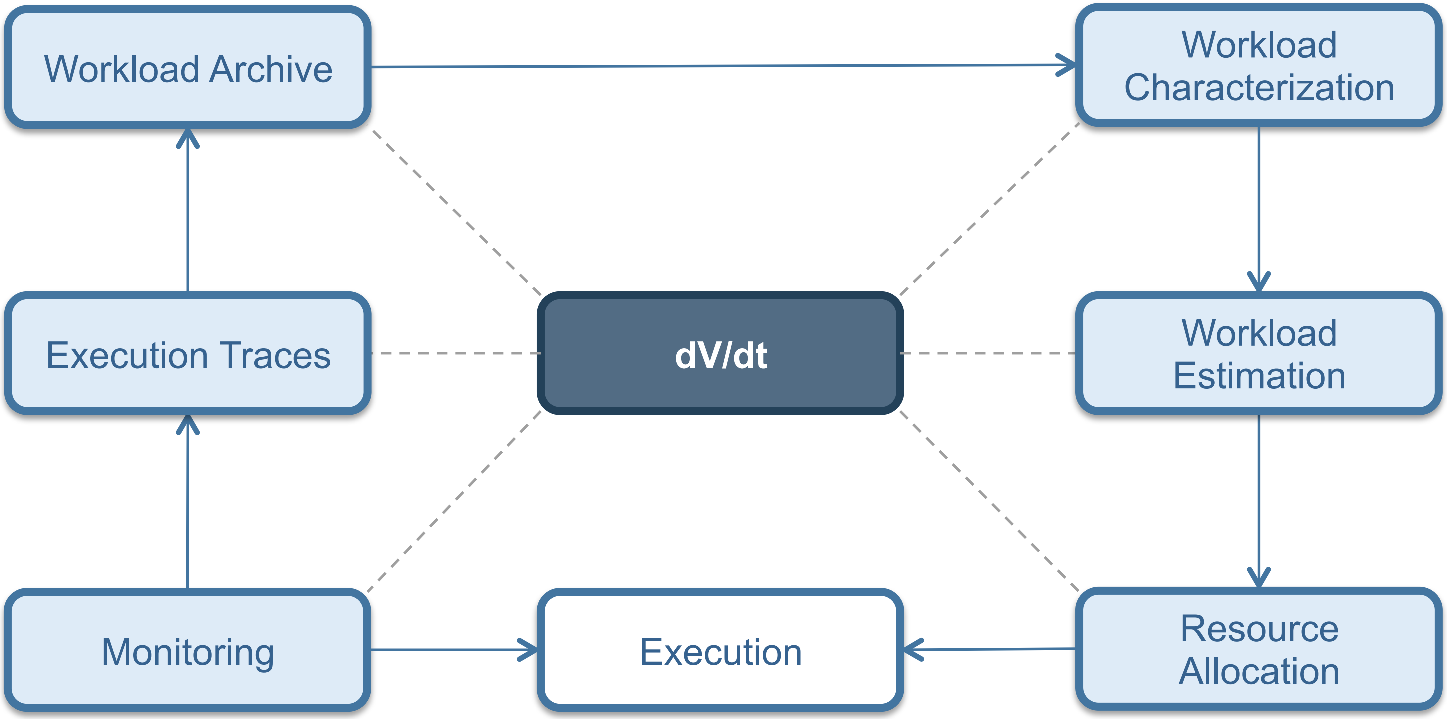
Task Characteristics:
Runtime
Disk Space
Memory Consumption

Scheduling and Resource Provisioning Algorithms

- Methods assume that accurate estimates are available
  - It is hard to compute accurate estimates in production systems

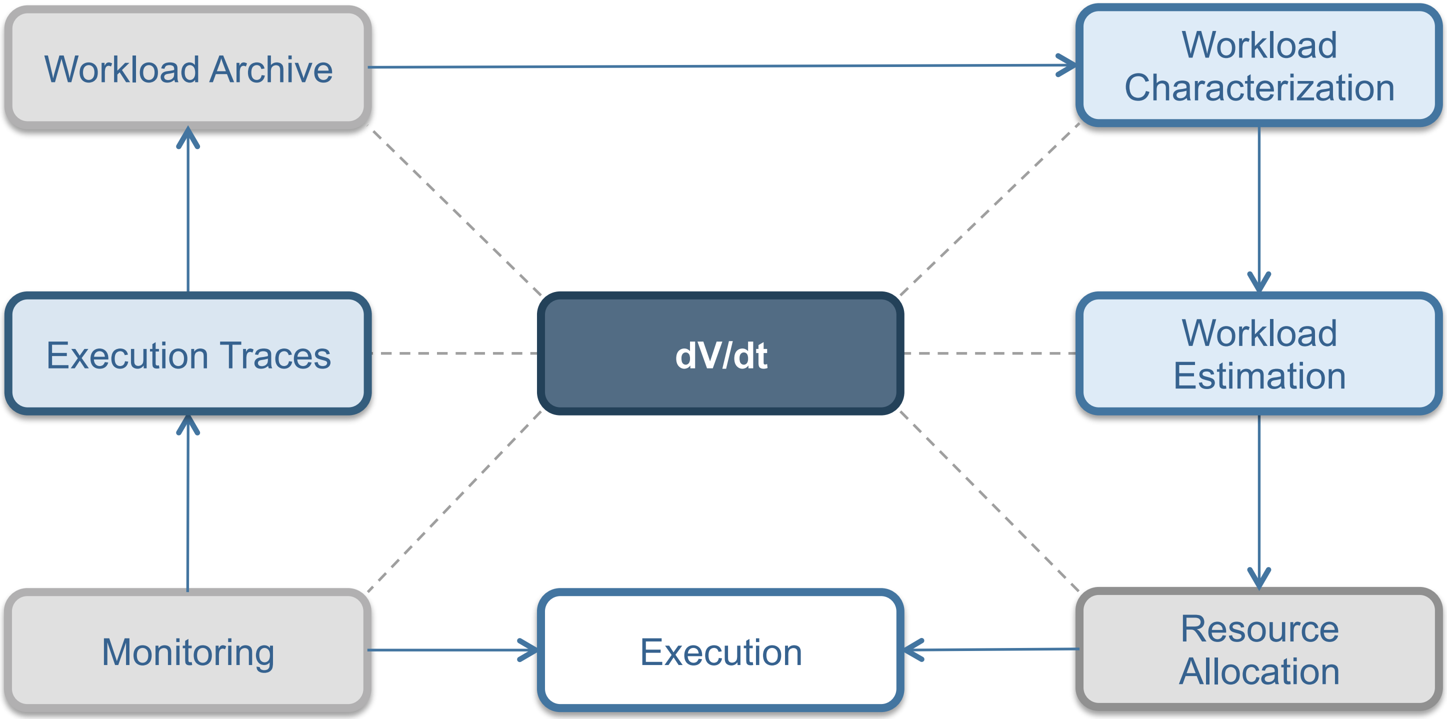
- Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC)
  - Process millions of jobs submitted by hundreds of users
  - The efficiency of the workload execution and resource utilization depends on how these jobs are scheduled and resources are provisioned

Overview of the Resource Provisioning Loop



USC Viterbi
School of Engineering
Information Sciences Institute

What is covered in this work?



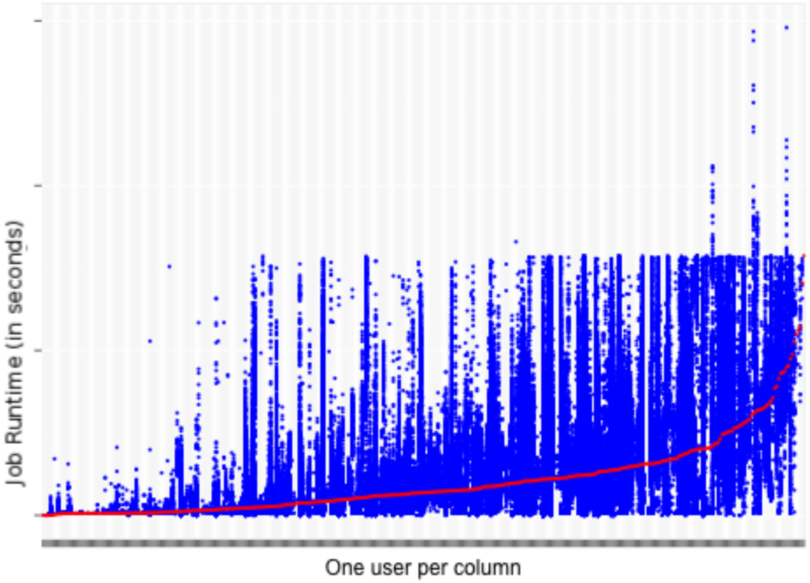
Workload Characteristics

Characteristics of the CMS workload for a period of a month (**Aug 2014**)

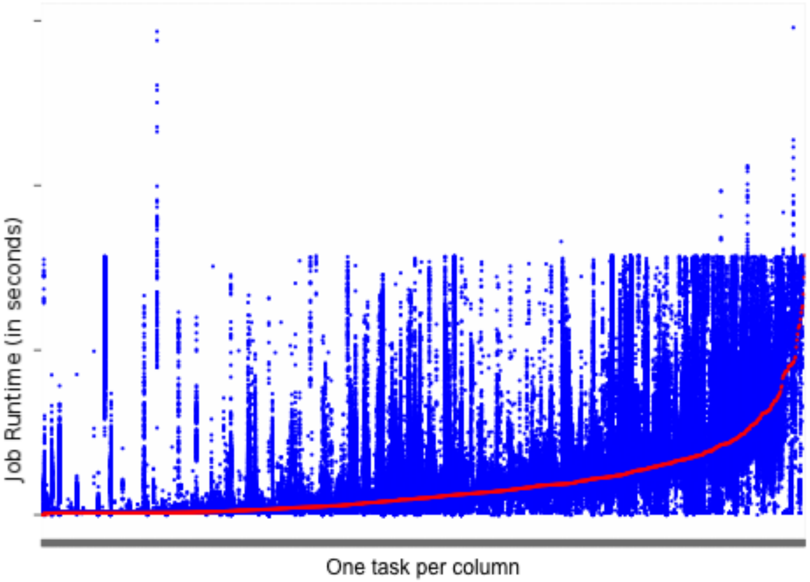
Characteristic	Data
General Workload	
Total number of jobs	1,435,280
Total number of users	392
Total number of execution sites	75
Total number of execution nodes	15,484
Jobs statistics	
Completed jobs	792,603
Preempted jobs	257,230
Exit code (!= 0)	385,447
Average job runtime (in seconds)	9,444.6
Standard deviation of job runtime (in seconds)	14,988.8
Average disk usage (in MB)	55.3
Standard deviation of disk usage (in MB)	219.1
Average memory usage (in MB)	217.1
Standard deviation of memory usage (in MB)	659.6

Workload Execution Profiling

- The workload shows similar behavior to the workload analysis conducted in [Sfiligoi 2013]
  - The magnitude of the job runtimes varies among users and tasks

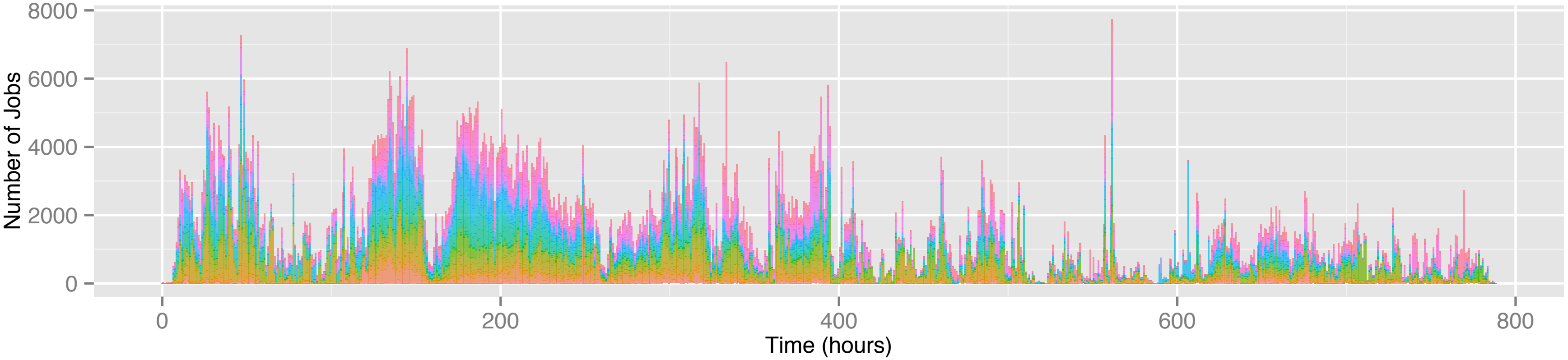


Job runtimes by user
sorted by per-user mean job runtime



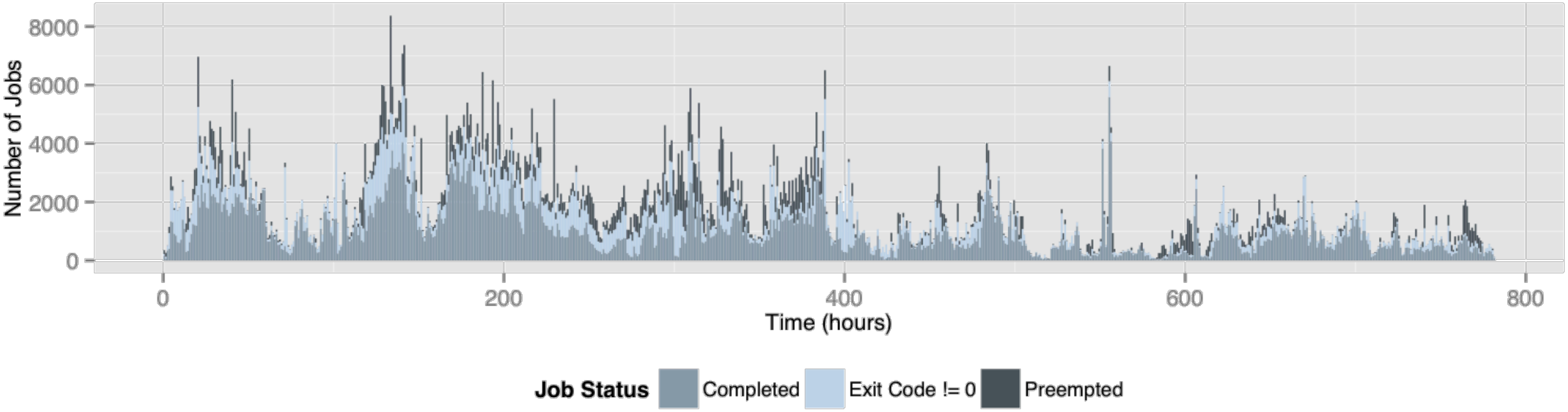
Job runtimes by task
sorted by per-task mean job runtime

Workload Execution Profiling (2)



Job start time rate

Colors represent different execution sites – job distribution is relatively balanced among sites

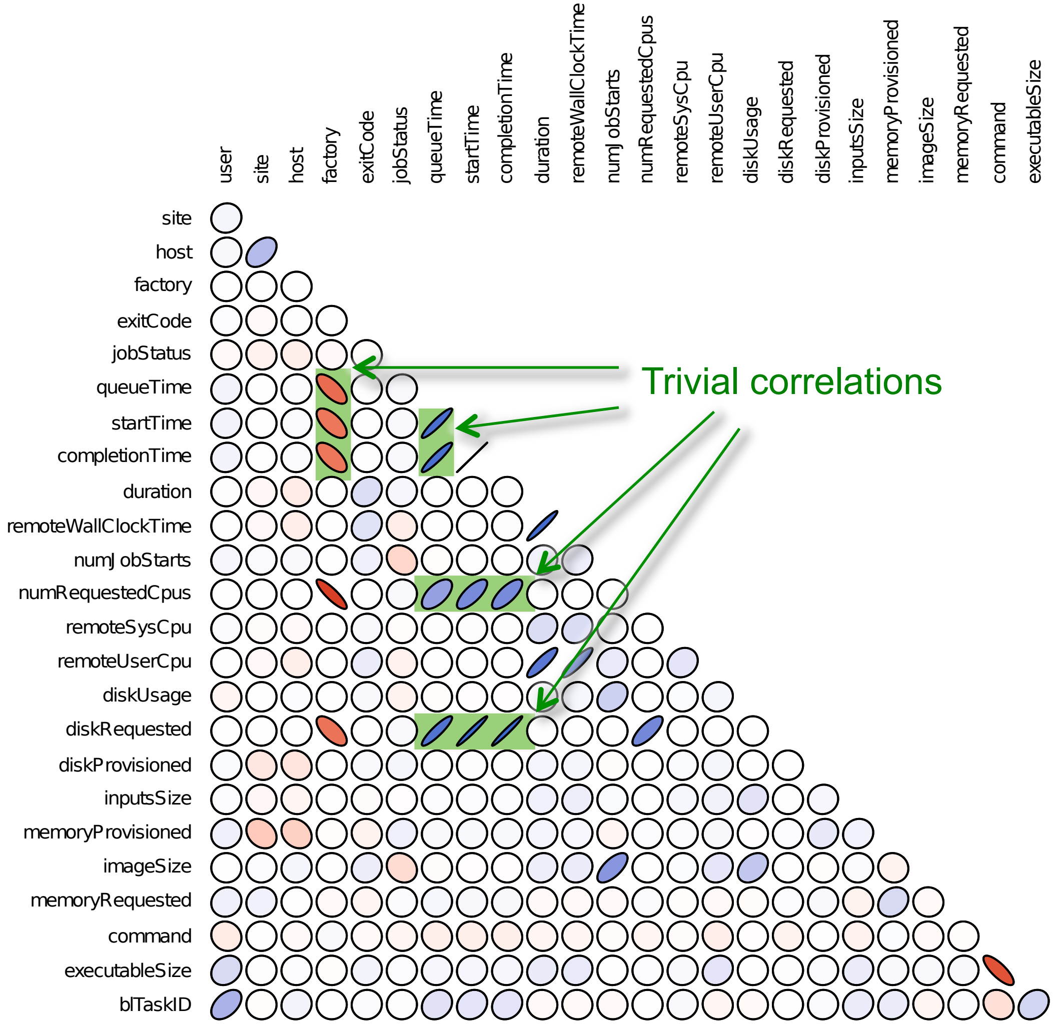


Job completion time rate

Colors represent different job status

Workload Characterization

- Correlation Statistics
  - Weak correlations suggest that none of the properties can be directly used to predict future workload behaviors
  - Two variables are correlated if the ellipse is too narrow as a line



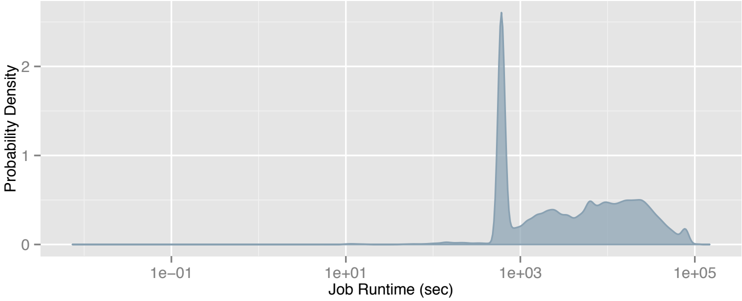
Trivial correlations

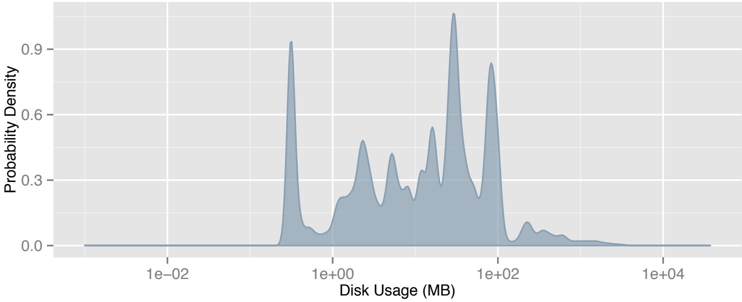
Workload Characterization (2)

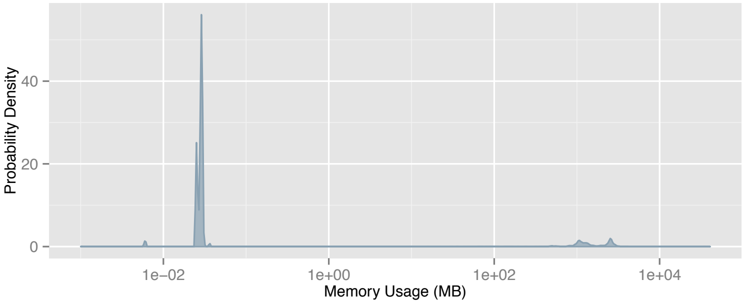
- Correlation measures are sensitive to the data distribution

- Probability Density Functions
  - Do not fit any of the most common families of density families (e.g. Normal or Gamma)

- Our approach
  - Statistical recursive partitioning method to combine properties from the workload to build Regression Trees

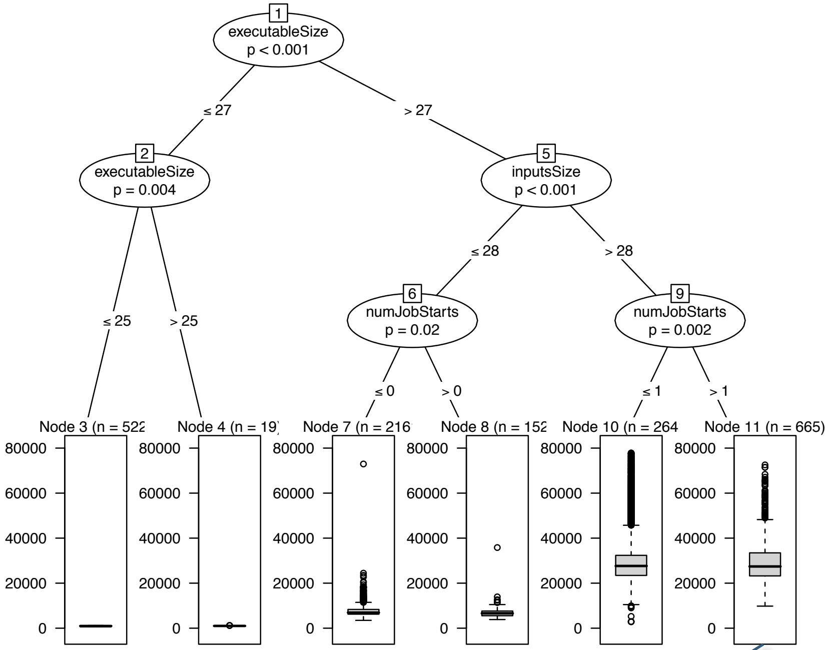






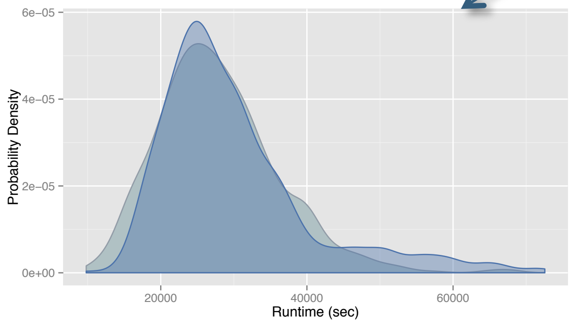
Regression Trees

- The recursive algorithm looks for PDFs that fit a family of density
  - In this work, we consider the Normal and Gamma distributions
  - Measured with the Kolmogorov-Smirnov test (K-S test)



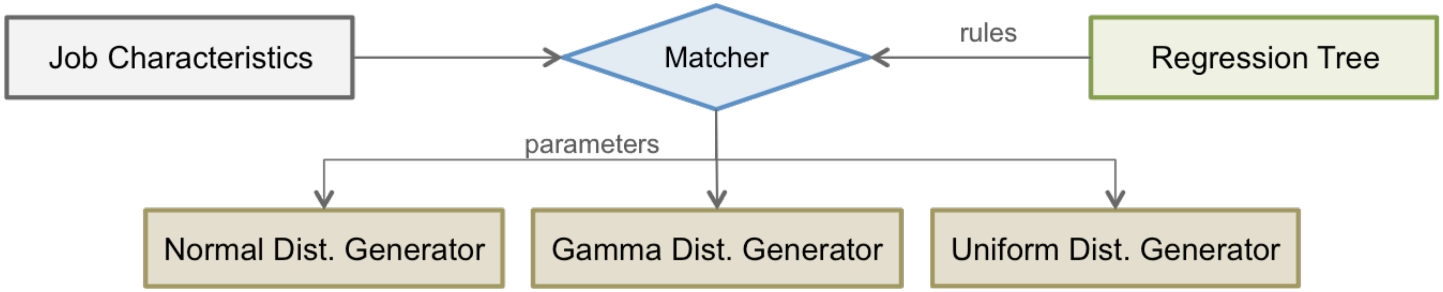
The PDF for the tree node (in blue) fits a Gamma distribution (in grey) with the following parameters:

Shape parameter = 12
Rate parameter = 5x10^{-4}
Mean = 27414.8
p-value = 0.17

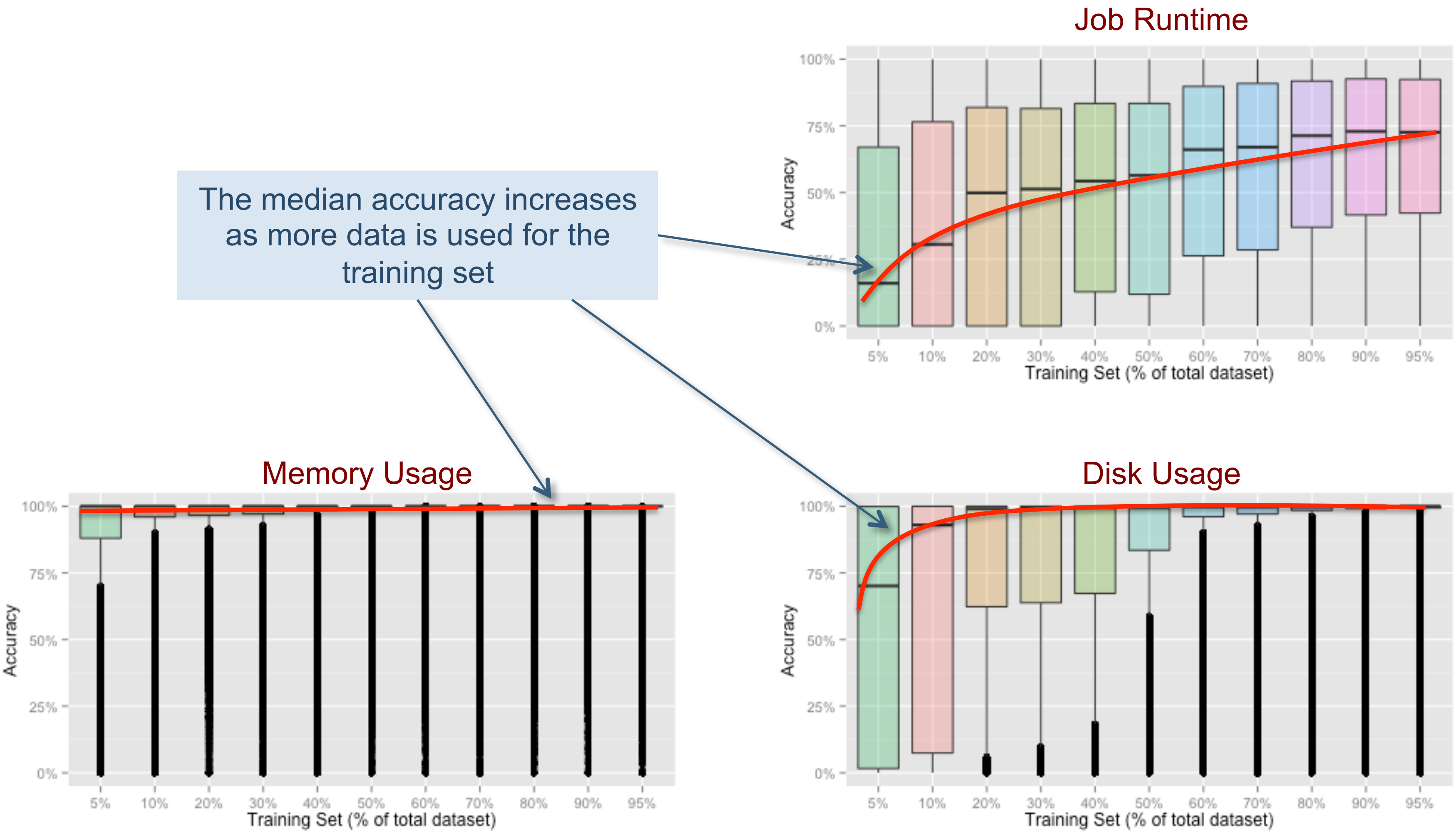


Job Estimation Process

- Based on the regression trees
  - We built a regression tree per user
  - Estimates are generated according to a distribution
    (Normal, Gamma, or Uniform)



Experimental Results



Job Runtime

The median accuracy increases as more data is used for the training set

Memory Usage

Disk Usage

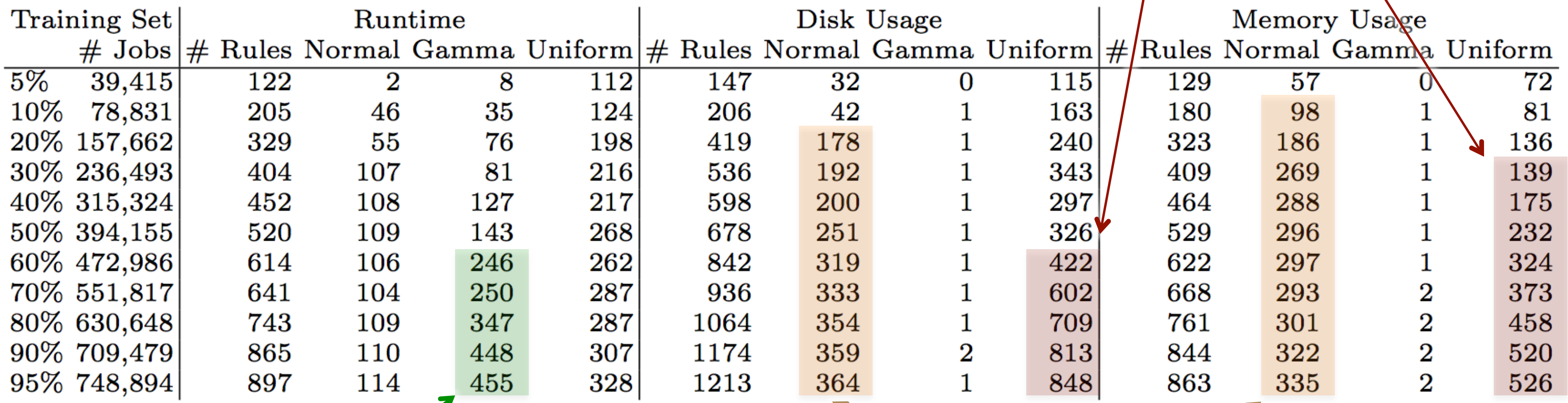
Average accuracy of the workload dataset
The training set is defined as a portion of the entire workload dataset

USC Viterbi
School of Engineering
Information Sciences Institute

Experimental Results (2)

- Number of Rules per Distribution
  - **Runtime**: better fits Gamma distributions
  - **Disk**: better fits Normal distributions
  - **Memory**: better fits Normal distributions

Specialization



Training Set		Runtime				Disk Usage				Memory Usage			
	# Jobs	# Rules	Normal	Gamma	Uniform	# Rules	Normal	Gamma	Uniform	# Rules	Normal	Gamma	Uniform
5%	39,415	122	2	8	112	147	32	0	115	129	57	0	72
10%	78,831	205	46	35	124	206	42	1	163	180	98	1	81
20%	157,662	329	55	76	198	419	178	1	240	323	186	1	136
30%	236,493	404	107	81	216	536	192	1	343	409	269	1	139
40%	315,324	452	108	127	217	598	200	1	297	464	288	1	175
50%	394,155	520	109	143	268	678	251	1	326	529	296	1	232
60%	472,986	614	106	246	262	842	319	1	422	622	297	1	324
70%	551,817	641	104	250	287	936	333	1	602	668	293	2	373
80%	630,648	743	109	347	287	1064	354	1	709	761	301	2	458
90%	709,479	865	110	448	307	1174	359	2	813	844	322	2	520
95%	748,894	897	114	455	328	1213	364	1	848	863	335	2	526

accuracy above 60%

Fits mostly Normal distributions

Prediction of Future Workloads

- Experiment Conditions
  - Used the workload from Aug 2014 to predict job requirements for October 2014
- Experiment Results
  - Median estimation accuracy

    Runtime: 82% (50% 1st quartile, 94% 3rd quartile)

    Disk and Memory consumption: over 98%

Characteristics of the CMS workload for a period of a month (**October 2014**)

Characteristic	Data
General Workload	
Total number of jobs	1,638,803
Total number of users	408
Jobs statistics	
Completed jobs	810,567

Conclusion

- Contributions
  - Workload characterization of 1,435,280 jobs
  - Use of a statistical recursive partitioning algorithm and conditional inference trees to identify patterns
  - Estimation process to predict job characteristics
- Experimental Results
  - Adequate estimates can be attained for job runtime
  - Nearly optimal estimates are obtained for disk and memory consumption
- Remarks
  - Data collection process should be refined to gather finer information
  - Applications should provide mechanisms to distinguish custom user codes from the standard executable

Characterizing a High Throughput Computing Workload: The Compact Muon Solenoid (CMS) Experiment at LHC

Thank you.

rafsilva@isi.edu

http://pegasus.isi.edu

This work was funded by DOE under the contract number ER26110, “dV/dt - Accelerating the Rate of Progress Towards Extreme Scale Collaborative Science”