ANALYSIS OF USER SUBMISSION BEHAVIOR ON HPC AND HTC

Rafael Ferreira da Silva

USC Information Sciences Institute

CS&T Directors' Meeting

USC Viterbi
School of Engineering
Information Sciences Institute

OUTLINE

Introduction

HPC and HTC
Job Schedulers
Problem Statement

Workload Characterization

Mira (ALCF)
Compact Muon Solenoid (CMS)

User Behavior in HPC

Think Time
Runtime and Waiting Time
Job Notifications

User Behavior in HTC

Think Time
Batches of Jobs
Batch-Wise Submission

Summary

Conclusions
Future Research Directions

REFERENCES

Consecutive Job Submission Behavior at Mira Supercomputer

S. Schlagkamp, R. Ferreira da Silva, W. Allcock, E. Deelman, and U. Schwiegelshohn
25th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC), 2016

Understanding User Behavior: from HPC to HTC

S. Schlagkamp, R. Ferreira da Silva, E. Deelman, and U. Schwiegelshohn
International Conference on Computational Science (ICCS), 2016

INTRODUCTION

Feedback-Aware Performance Evaluation of Job Schedulers

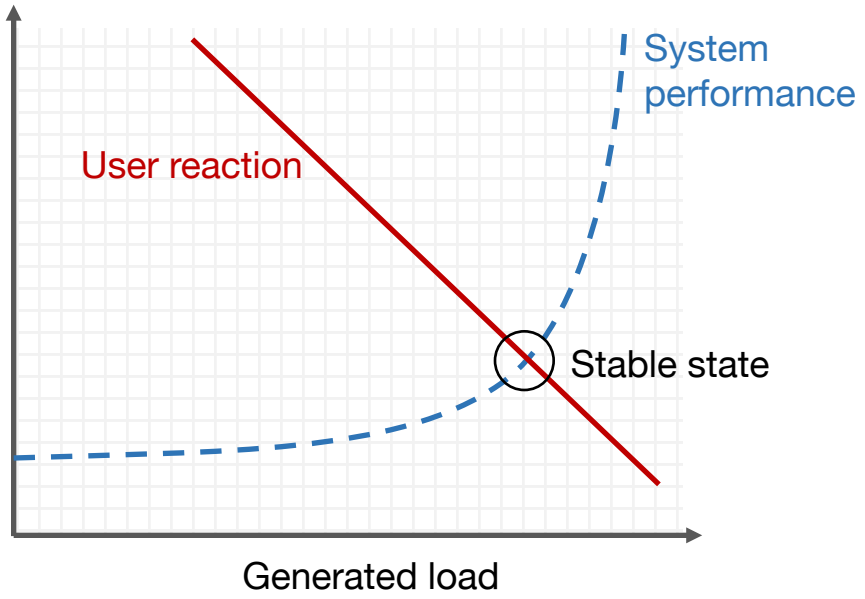
Evaluation with previously recorded workload traces

One instantiation of a dynamic process
User reaction is a mystery

D. G. Feitelson, 2015

Lack between theory and practice
Further understanding of reactions to system performance

U. Schwiegelshohn, 2014



THINK TIME

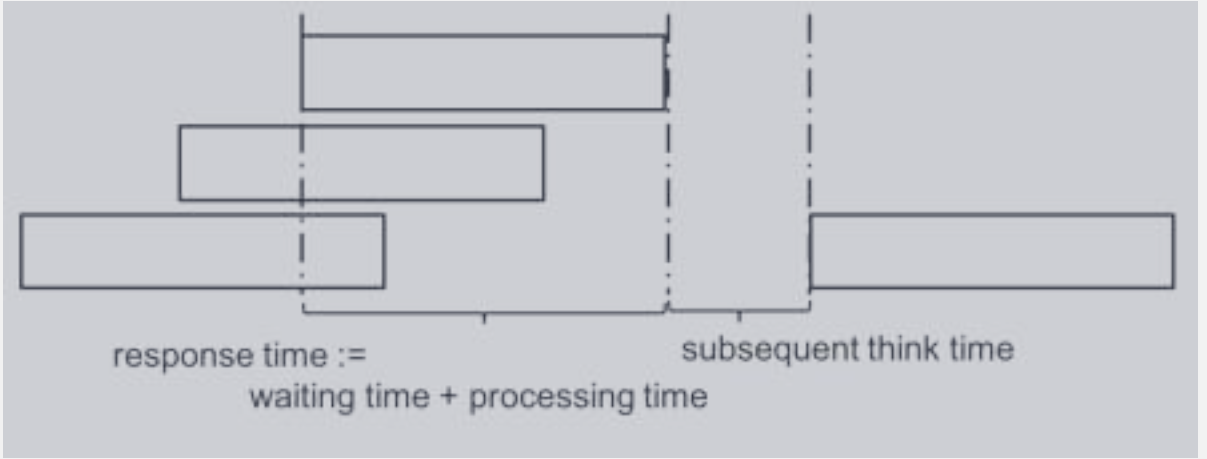
ID	Submit Time	Wait Time	Run Time	...	Req. Resources	User ID
248	727216	9120	9585	...	128	12
249	727280	10830	16273	...	256	12
250	727531	2560	1720	...	32	12
251	735783	1204	3440	...	128	12

timings + resources

subsequent submission behavior predictions

How do users react to system performance?
data-driven analysis

Think Time
Time between job completion and consecutive job submission



OUTLINE

Introduction

HPC and HTC
Job Schedulers
Problem Statement

Workload Characterization

Mira (ALCF)
Compact Muon Solenoid (CMS)

User Behavior in HPC

Think Time
Runtime and Waiting Time
Job Notifications

User Behavior in HTC

Think Time
Batches of Jobs
Batch-Wise Submission

Summary

Conclusions
Future Research Directions

WORKLOAD CHARACTERISTICS

MIRA (ALCF) 2014

49,152
Total Number of Nodes

786,432
Total Number of Cores

487
Total Number of Users

78,782
Total Number of Jobs

5,698
CPU hours (millions)

6,093
Avg. Runtime (seconds)

Physics

73 Users
24,429 Jobs
2,256 CPU hours (millions)
7,147 Avg. Runtime (sec)

Materials Science

77 Users
12,546 Jobs
895 CPU hours (millions)
5,820 Avg. Runtime (sec)

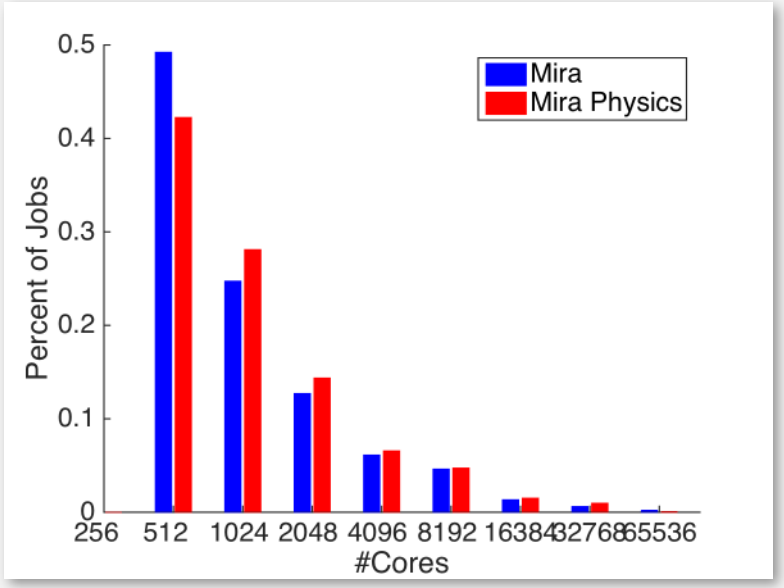
Chemistry

51 Users
10,286 Jobs
810 CPU hours (millions)
6,131 Avg. Runtime (sec)

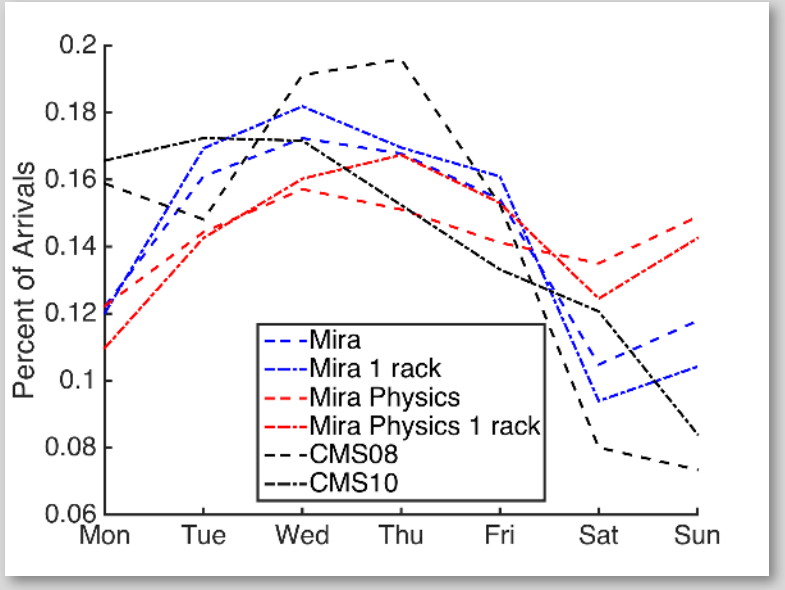
WORKLOAD CHARACTERISTICS

AUGUST 2014		COMPACT MUON SOLENOID (CMS)		OCTOBER 2014
1,435,280 Total Number of Jobs	**392** Total Number of Users		**408** Total Number of Users	**1,638,803** Total Number of Jobs
75 Execution Sites	**15,484** Execution Nodes		**15,034** Execution Nodes	**72** Execution Sites
792,603 Completed Jobs	**385,447** Exit Code (!= 0)		**476,391** Exit Code (!= 0/)	**816,678** Completed Jobs
9,444.6 Avg. Runtime (sec)	**55.3** Avg. Disk Usage (MB)		**32.9** Avg. Disk Usage (MB)	**9967.1** Avg. Runtime (sec)

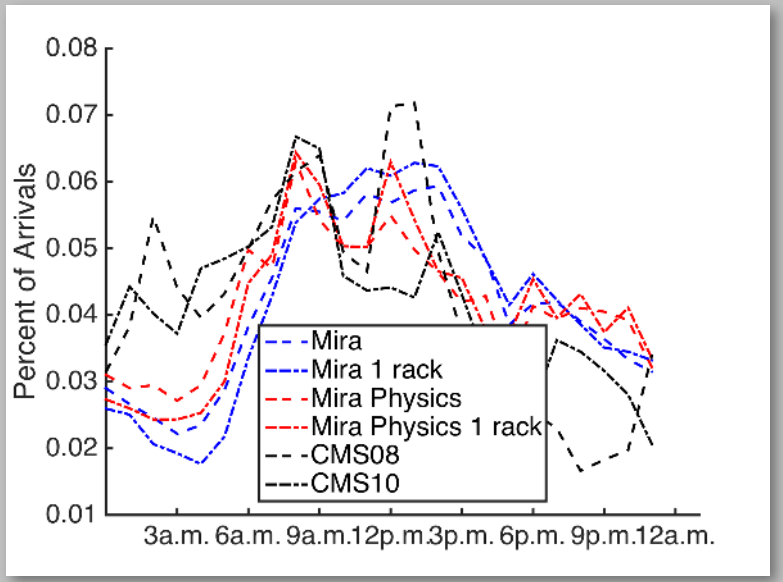
WORKLOAD CHARACTERIZATION



Jobs' resource requirements at Mira



Job submission interarrival times per day



Job submission interarrival times per hour

OUTLINE

Introduction

HPC and HTC
Job Schedulers
Problem Statement

Workload Characterization

Mira (ALCF)
Compact Muon Solenoid (CMS)

User Behavior in HPC

Think Time
Runtime and Waiting Time
Job Notifications

User Behavior in HTC

Think Time
Batches of Jobs
Batch-Wise Submission

Summary

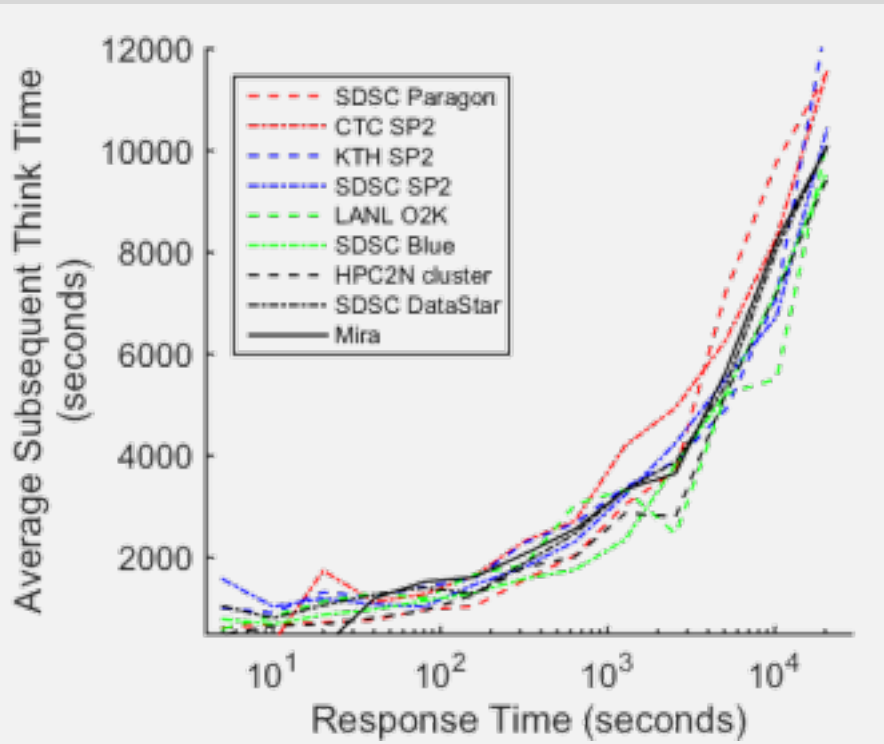
Conclusions
Future Research Directions

USER THINK TIME

The *user's think time* quantifies the timespan between a job completion and the submission of the next job (by the same user)

We only account for positive times (and less than 8 hours) between subsequent job submissions

No change in the past 20 years



Average think times in several traces from the Parallel Workloads Archive and Mira

CORRELATIONS

Response Time

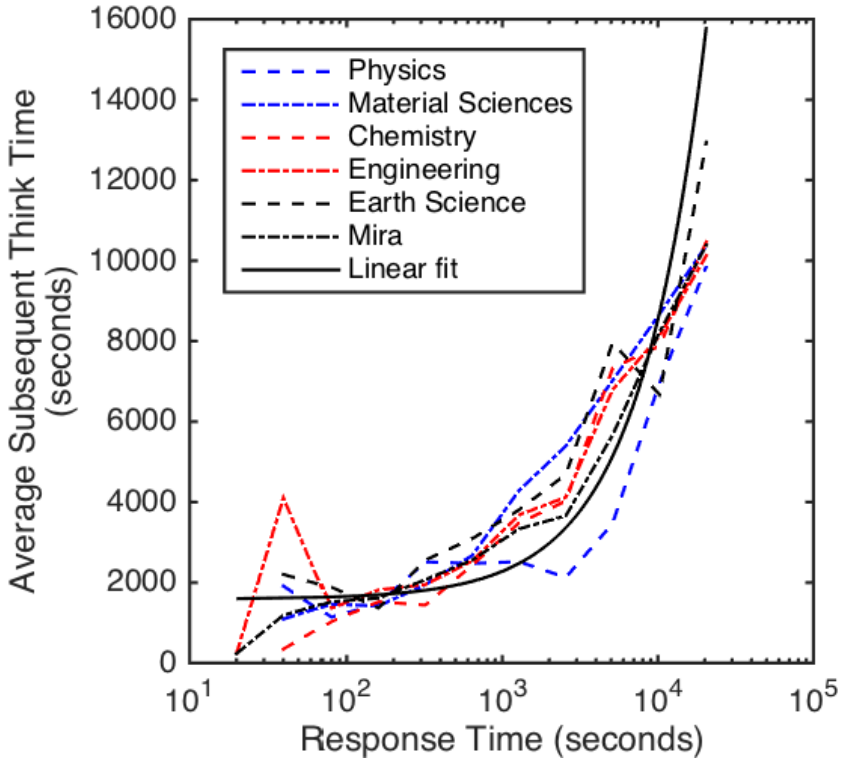
Waiting Time + Runtime

General Behavior

Subsequent behavior independent of the science field/application

Low values (nearly instantaneous submissions) is typically due to the user of automated scripts

Peaks (e.g., Engineering) is due to outliers (about 8h)

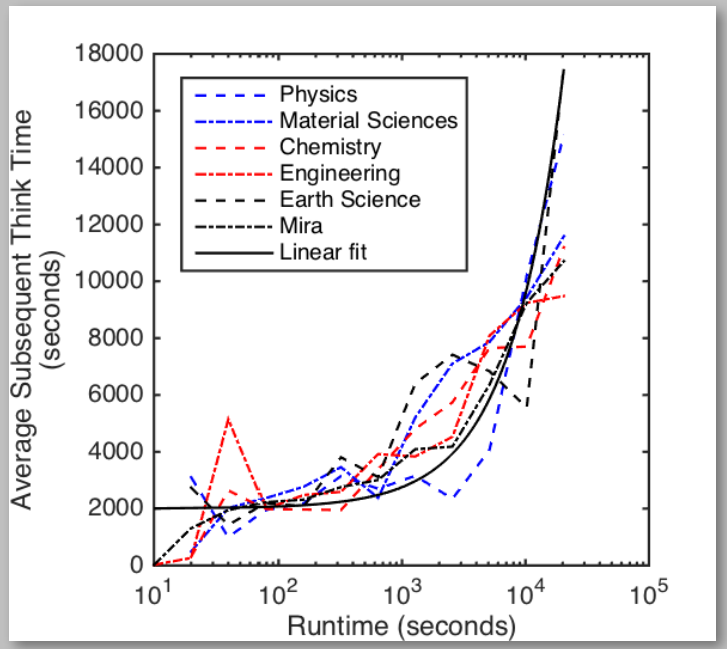


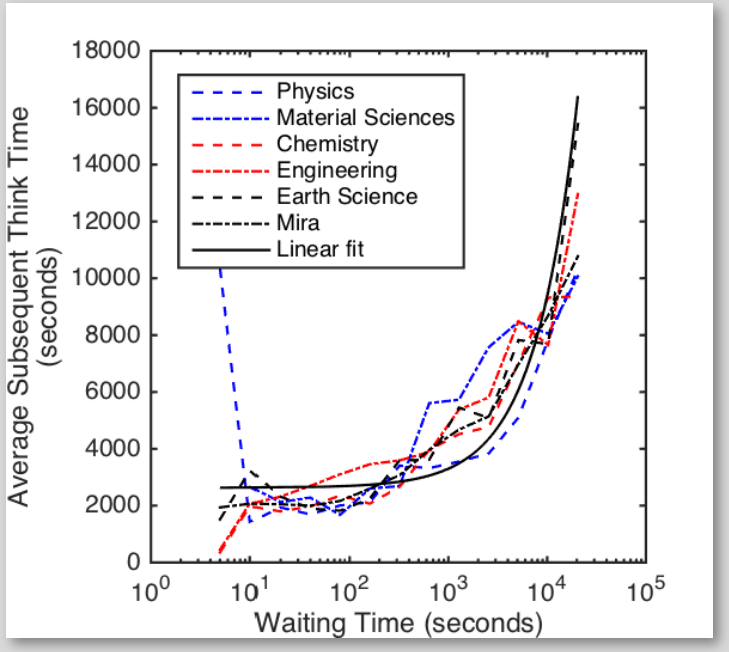
CORRELATIONS

Both runtime and waiting time have ***equal influence*** *on the user behavior*

Reducing queuing times would not significantly improve think times for long running jobs

RUNTIME

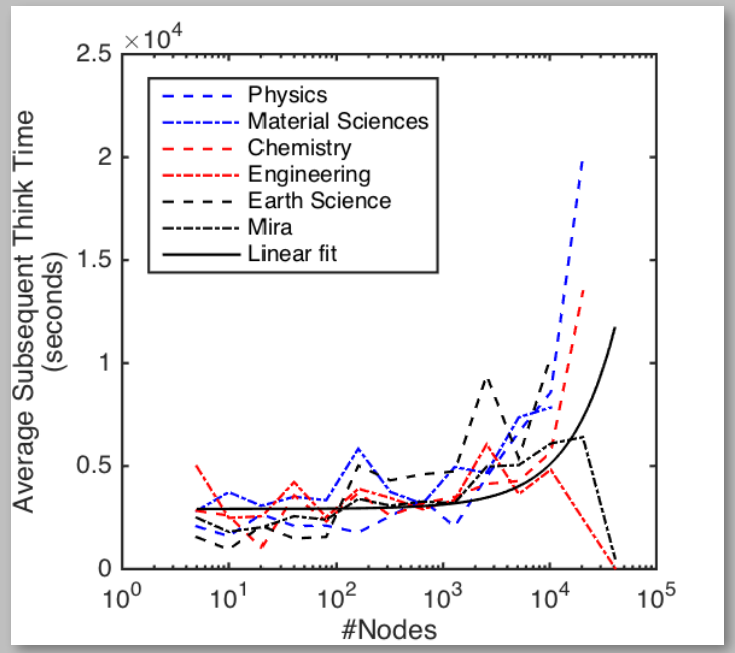




WAITING TIME

CORRELATIONS

NODES



For small jobs (~10^3 nodes), average think times are relatively small (<1.5h)

For larger jobs, it substantially increases:

- Users do not fully understand the behavior of their applications as the number of cores increase
- Resource allocation for larger jobs is delayed
- Larger jobs require additional settings and refinements (increased job complexity)

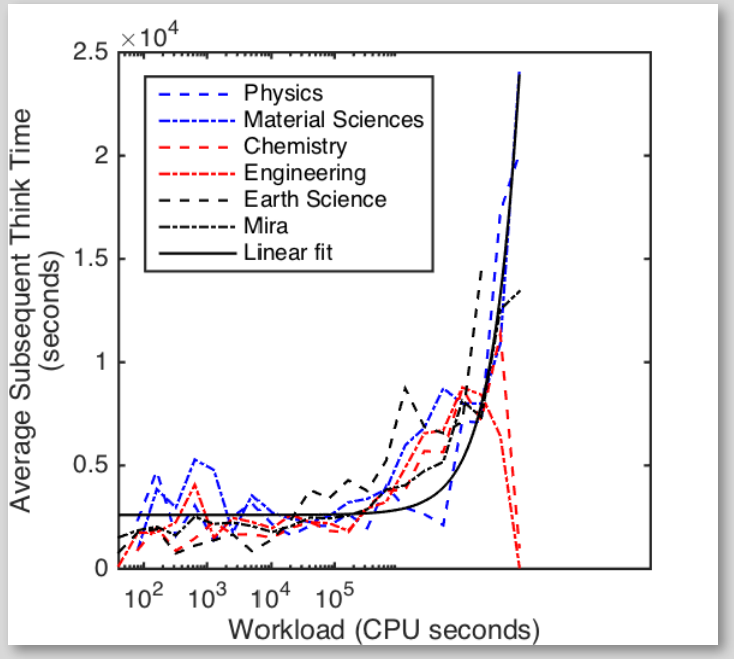
CORRELATIONS

*Think time is **heavily** correlated to the workload*

Workload has more impact as it also considers runtime

`w(j) = processing time x number of nodes`

Similar conclusions to the number of nodes analysis



WORKLOAD

ANALYSIS OF JOB CHARACTERISTICS

Think times are small when *runtime* prevails

User behavior is not impacted by the job size

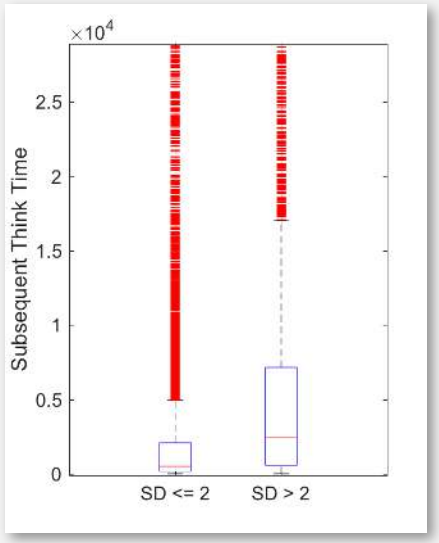
More complex jobs do yield **higher** think times, however there is a similar behavior when runtime or waiting time prevail

- *Complex jobs requires more think time*
- *Lack of accurate runtime estimates*

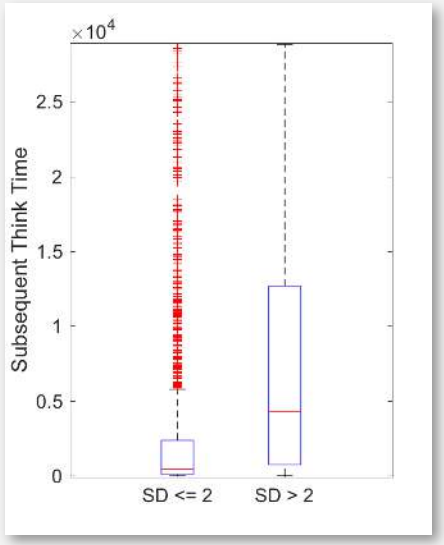
small jobs
Represent **49.2%** of total jobs

less complex jobs
Consume less than 277 CPU hours

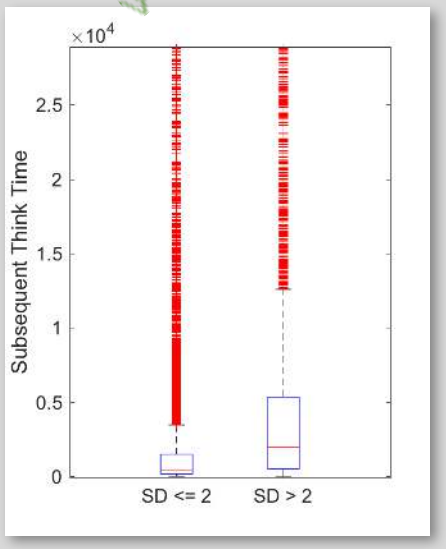
NODES



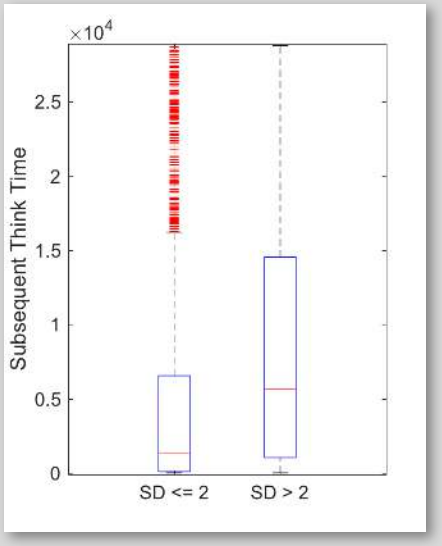
<= 512



> 512



<= 10^6



> 10^6

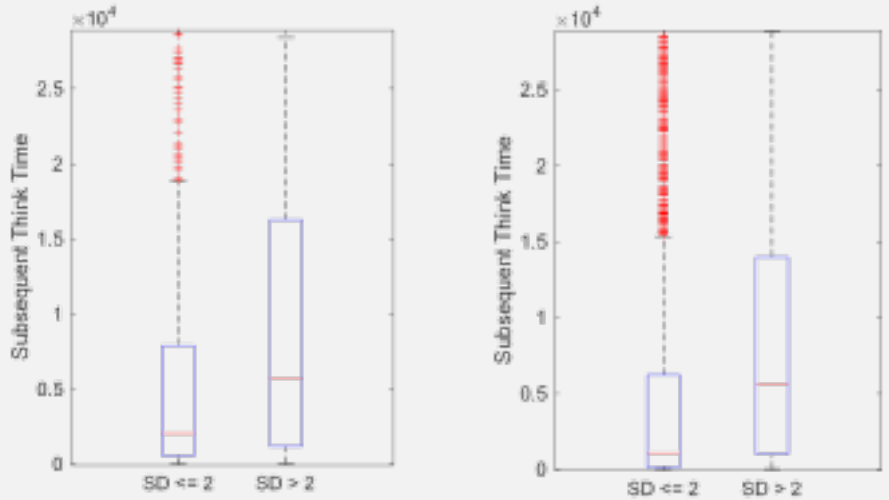
WORKLOAD

ANALYSIS OF JOB NOTIFICATIONS

The overall user behavior is nearly ***identical*** regardless of whether the user is notified

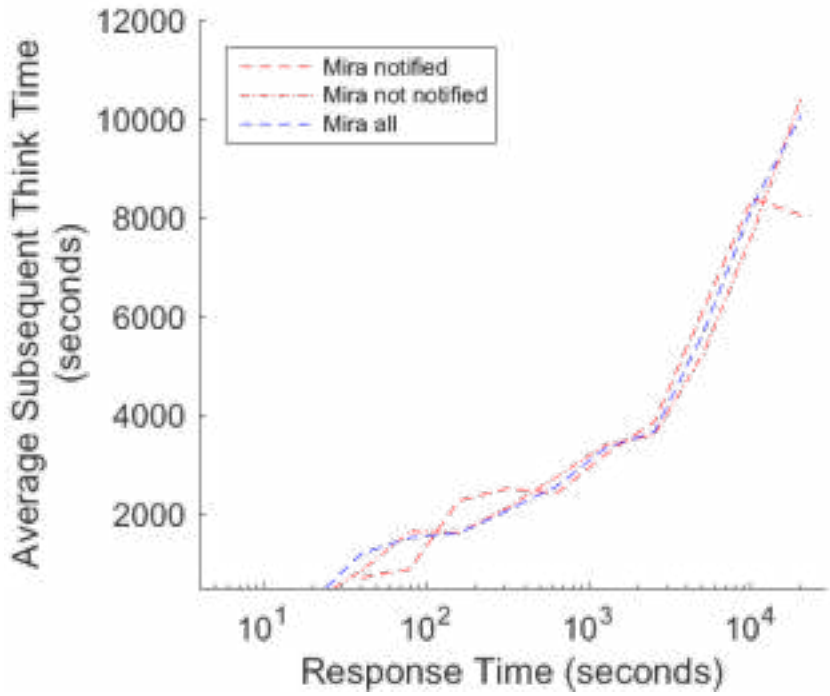
17,736 out of 78,782 jobs used the email notification mechanism

LARGE JOBS



With notification

Without notification



Average think time as a function of response time for jobs with and without notification upon job completion

SUMMARY

User Behavior in HPC

Think Time
Runtime and Waiting Time
Job Notifications

>

Discussion

There is no shift on the think time behavior during the past 20 years. This similar behavior is due to the current restrictive definition to model think time

Simulating submission behavior has to consider other job characteristics and system performance components

A notification mechanism has no influence on the subsequent user behavior. Thus, there is no urging to model user *(un)awareness* of job completion in performance evaluation simulations

OUTLINE

Introduction

HPC and HTC
Job Schedulers
Problem Statement

Workload Characterization

Mira (ALCF)
Compact Muon Solenoid (CMS)

User Behavior in HPC

Think Time
Runtime and Waiting Time
Job Notifications

User Behavior in HTC

Think Time
Batches of Jobs
Batch-Wise Submission

Summary

Conclusions
Future Research Directions

CHARACTERIZING THINK TIME

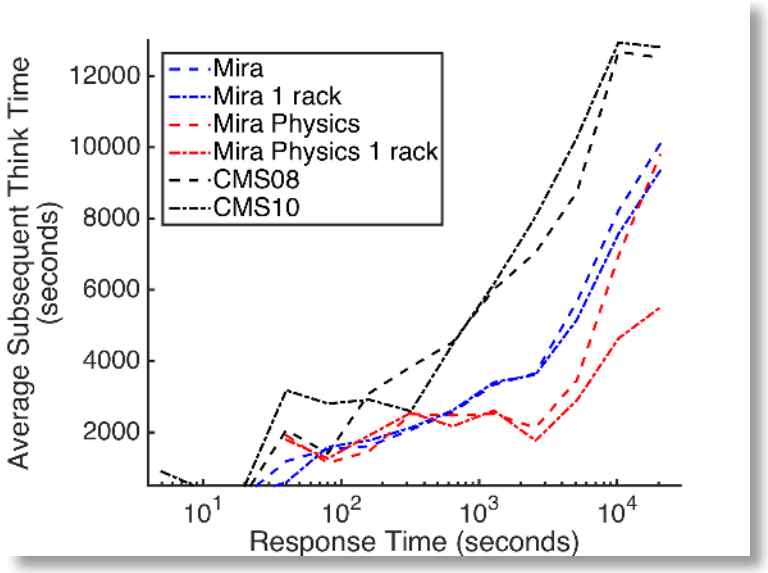
HPC

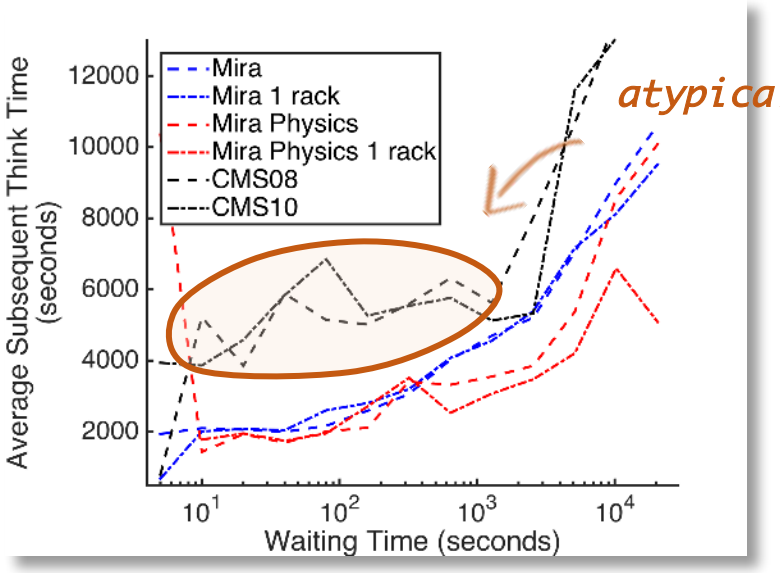
Tightly coupled applications
Methods: Think time

> > >

HTC

Embarrassingly parallel applications





BATCHES OF JOBS

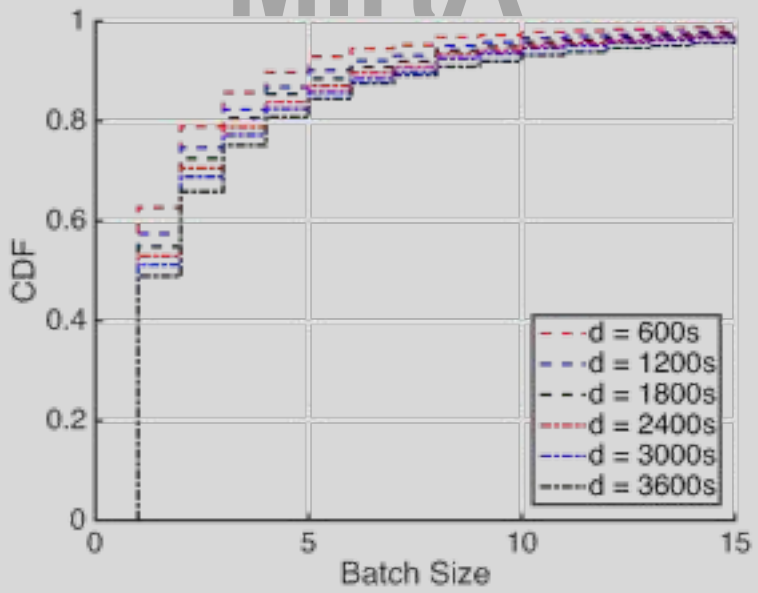
User-triggered job submissions are often clustered and denoted as ***batches***

Large interarrival thresholds may not capture the actual job submission behavior in HTC systems

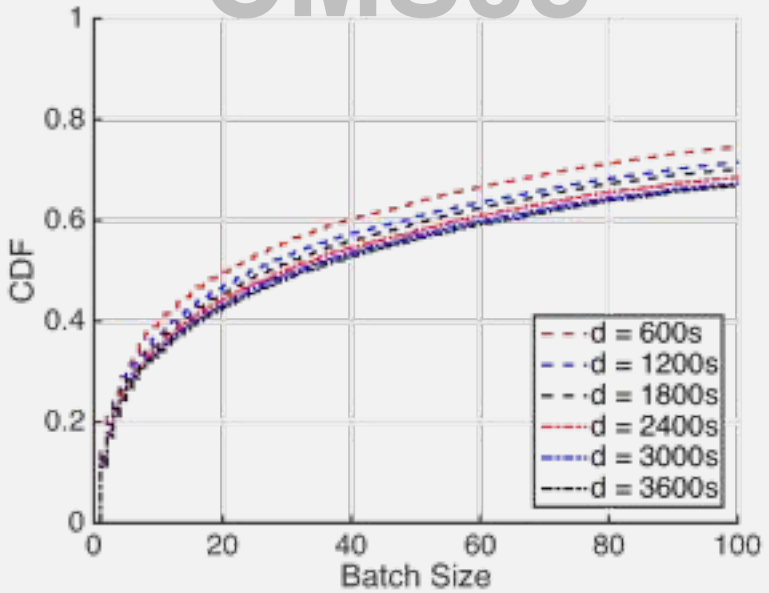
Two jobs successively submitted by a user belong to the same batch if the interarrival time between their submissions is within a threshold:

$$i_{j,j'} := s'_j - s_j \leq \Delta$$

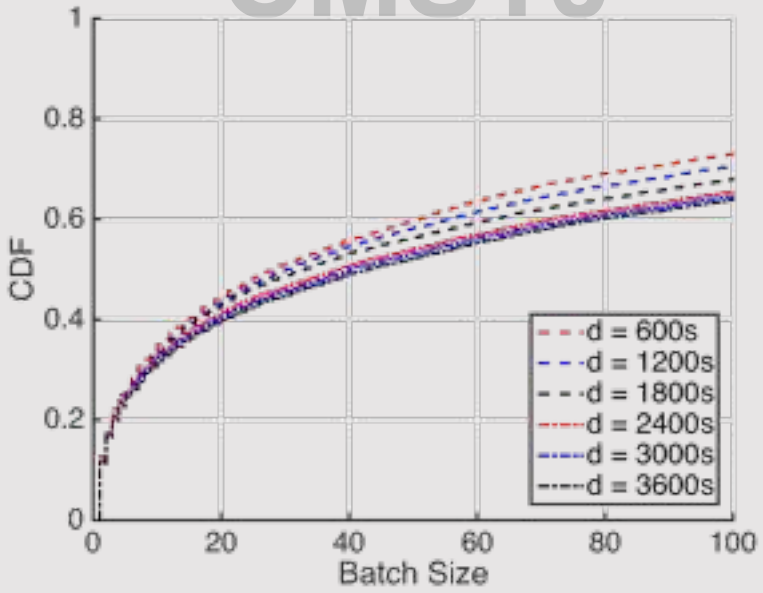
MIRA



CMS08



CMS10



REDEFINING THINK TIME

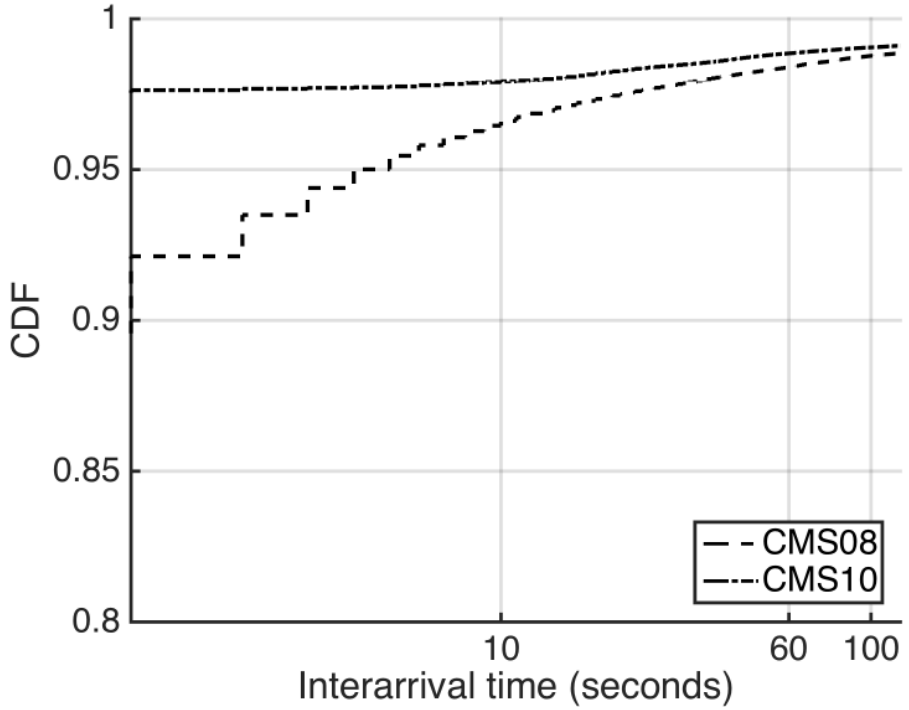
Think Time for HTC

Quantifies the timespan between two subsequent submissions of **bags of tasks**

General Behavior

Most of the jobs belonging to the same experiment and user (97%) are submitted within one minute

We use the threshold of 60s to distinguish between automated bag of tasks submissions and human-triggered submissions (batches)



Distribution of interarrival times (CDF)

THINK TIME IN HTC

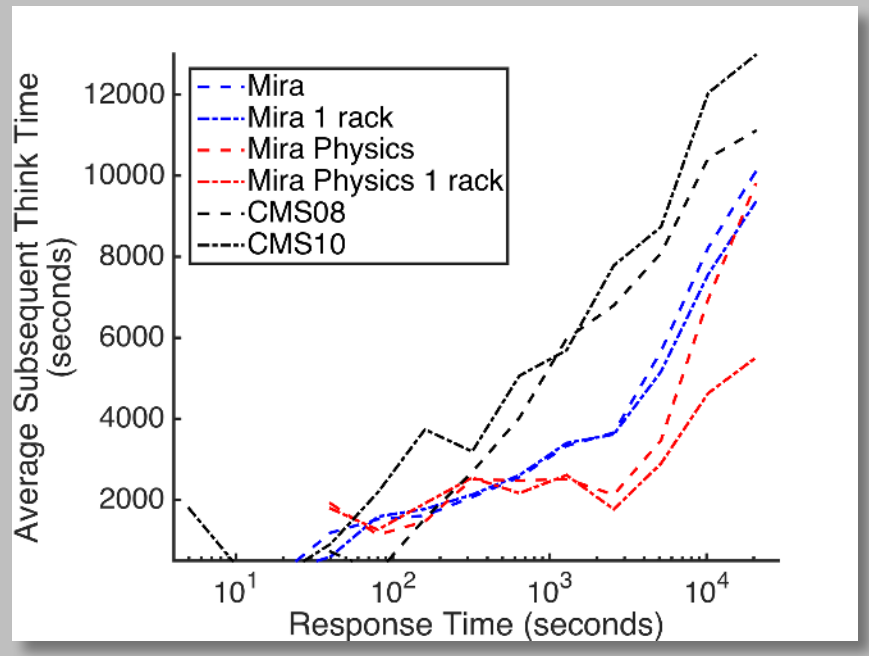
Both HTC workloads follow the same linear trend

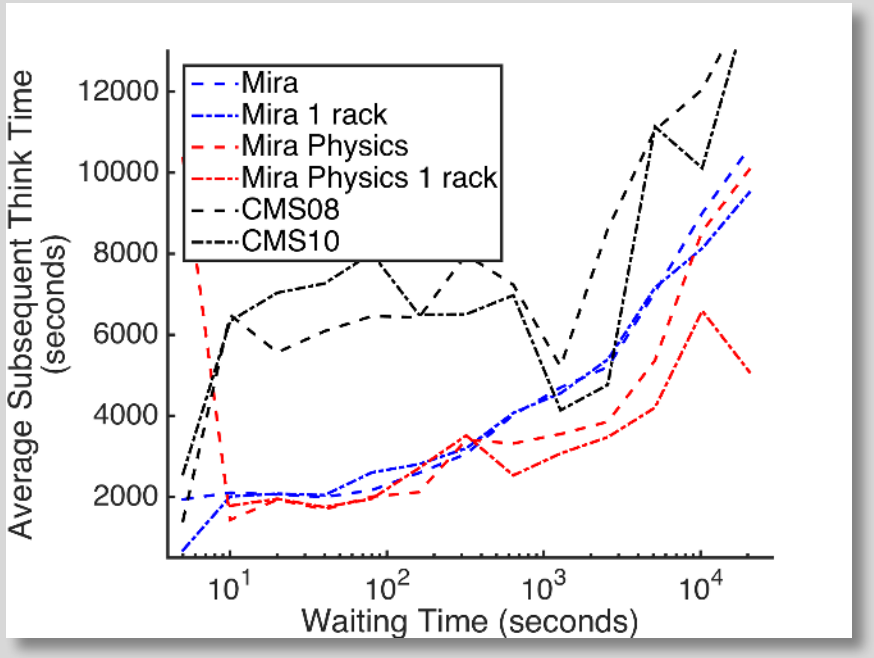
Batch-Wise Analysis: **Lower** think times when compared to standard analysis based on individual jobs

HTC BoTs are comparable to HPC jobs

User behavior in CMS is not strictly related to waiting time

RESPONSE





WAITING TIME

ALTERNATIVE THINK TIME DEFINITIONS

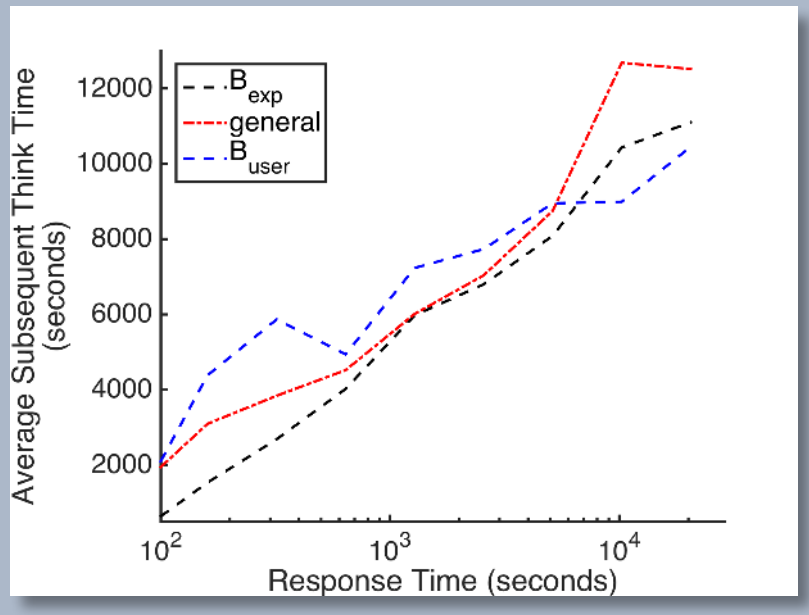
B_{exp} the ground truth knowledge (from CMS traces)

B_{user} bag of tasks based on jobs submitted by the same user (most common approach)

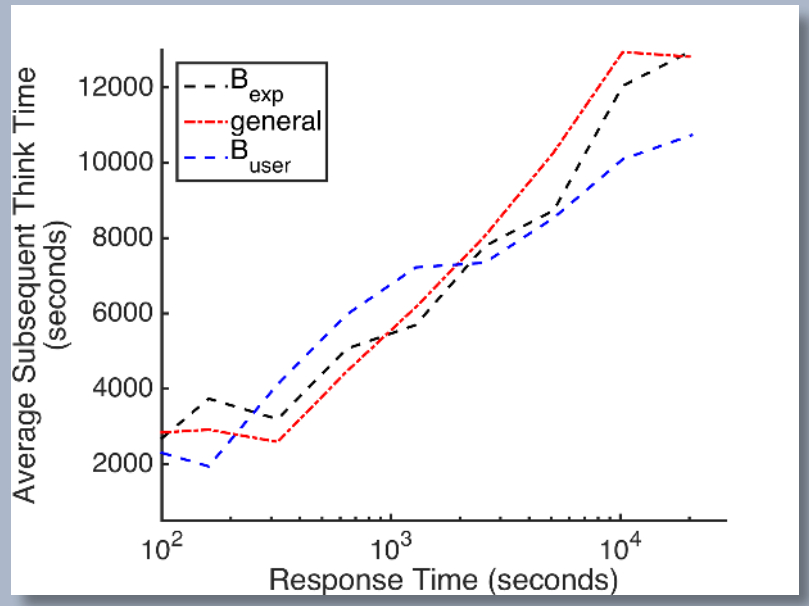
general jobs are treated individually

> *The subsequent think time behavior for the **general** behavior is closer to B_{exp} than the B_{user}*

CMS 08



Comparison of different data interpretations for think time



CMS 10

SUMMARY

User Behavior in HTC

Think Time
Batches of Jobs
Batch-Wise Submission

>

Discussion

Although HTC jobs are composed of thousands of embarrassingly parallel jobs, the general human submission behavior is comparable to HPC

Additional information is required to properly identify HTC batches

Subsequent behavior in HPC is sensitive to the job complexity, while BoTs drives the HTC behavior

There is no strong correlation between waiting and think times in the CMS experiments due to the dynamic behavior of queuing times within BoTs

FUTURE WORK

Summary

Conclusion
Future Research Directions

Future Research Directions

Extend and explore different think time definitions (e.g., based on concurrent activities)

Model think time as a function of job complexity from past job submissions

Cognitive studies: understand user reactions based on waiting times and satisfaction

In-depth characterization of waiting times in bags of tasks to improve correlation analysis between queuing time and think time

ANALYSIS OF USER SUBMISSION BEHAVIOR ON HPC AND HTC

Thank You

Questions?

Rafael Ferreira da Silva, Ph.D.

Research Assistant Professor
Department of Computer Science
University of Southern California

rafsilva@isi.edu – http://rafaelsilva.com