Trace-Based Evaluation of Job Runtime and Queue Wait Time Predictions in Grids

Ozan Sonmez, Nezih Yigitbasi,
Alexandru Iosup, Dick Epema

Parallel and Distributed Systems Group
(PDS)
Department of Software Technology
Faculty EEMCS, Delft,
the Netherlands

Introduction

- **Grids**
  - Multi-site and heterogeneous resource structure
  - Dynamic and heterogeneous workloads

  → Highly variable **job runtimes** and **queue wait times** limit the efficient use of the resources by users

Introduction (cont.)

- Remedy: **Prediction-based methods**
  - Extensive body of research for space-shared Parallel Production Environments (**PPEs**)
  - **Grids** differ from traditional PPEs in both structure and typical use (e.g., heterogeneous resources, more bursty job arrivals)

  - **Goal:**
  - A systematic evaluation of job runtime and queue wait time predictions in grids using **real traces**

What to predict?

- **Job Runtime**
- **Queue Wait Time**
- CPU Load
- Resource Availability
- Resource Failure Rates

What to predict?

- **Job runtime predictions for**
  - Improving the performance of backfilling in batch queueing systems*
  - Predicting queue wait times
- **Queue wait time predictions for**
  - Guiding the decisions of a user/grid scheduler

*D. Tsafrir, Y. Etsion, and D. G. Feitelson. *Backfilling Using System-Generated Predictions Rather than User Runtime Estimates*. IEEE TPDS, 18(6):789–803, 2007

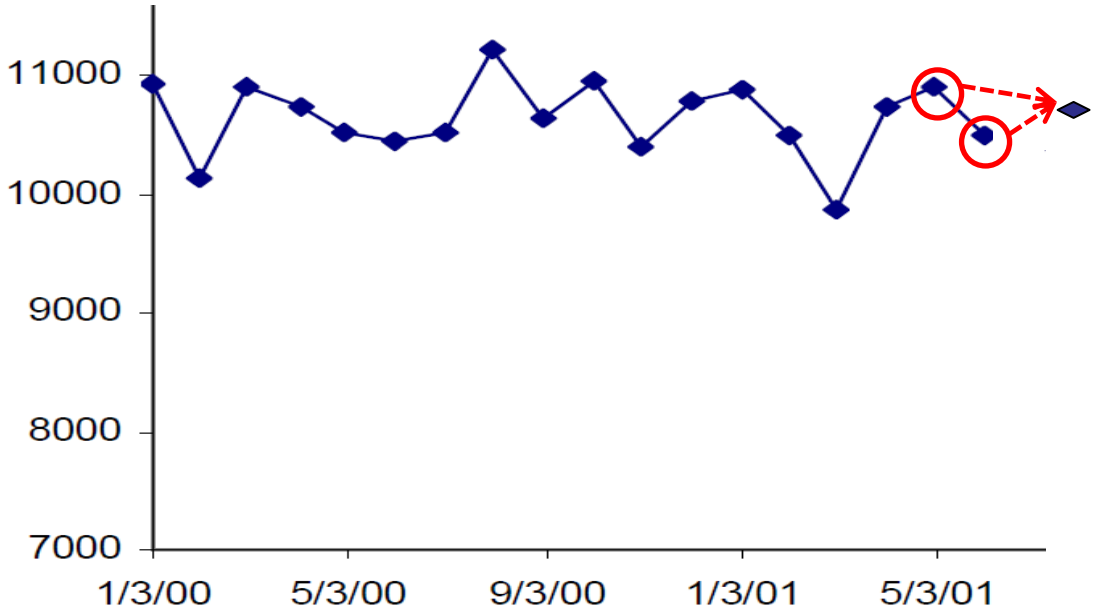
Prediction Methods

- **Time Series-based** → **Easy to implement**
  Fast delivery of predictions
- Analytical Benchmarking
- Code Profiling
- Genetic Algorithms
- Instance-based Learning

Time Series Prediction

- Based on historical (classified) data
  - Time ordered set of past observations

- **Example: Last2**



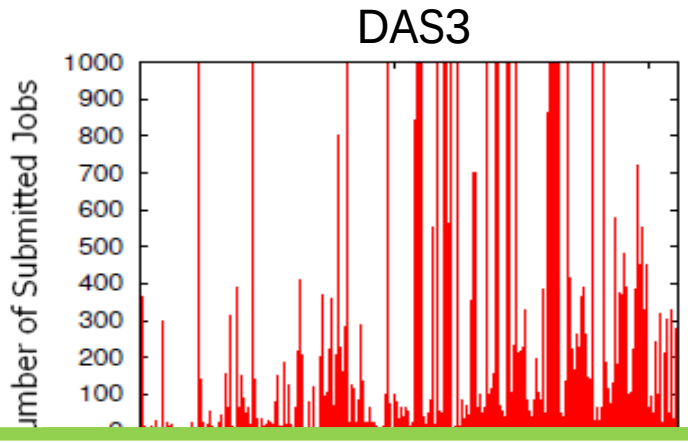
TU Delft

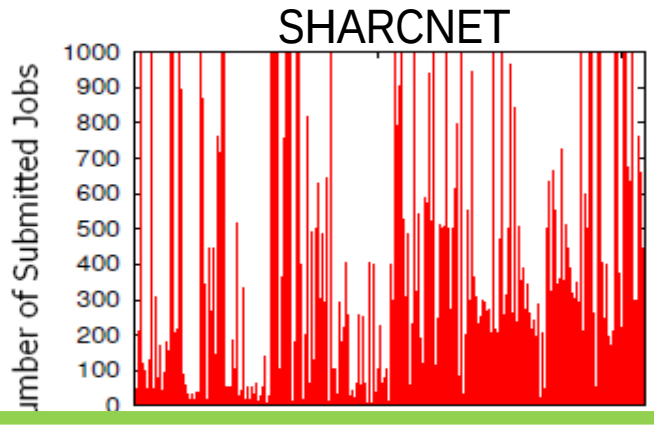
Grid Workload Traces*

Traces	Type	# CPUs	Duration (Months)	# Tasks	Parallel Jobs
DAS2	Research	400	18	1.1 M	66%
GRID5000	Research	2500	27	1.0 M	45%
DAS3	Research	544	18	2 M	15%
SHARCNET	Research	6828	12	1.2 M	10%
AUVER	Production	475	12	0.4 M	0%
NORDU	Production	2000	24	0.8 M	0%
LCG	Production	24515	4	0.2 M	0%
NGS	Production	-	6	0.6 M	0%
GRID3	Production	3500	18	1.3 M	0%

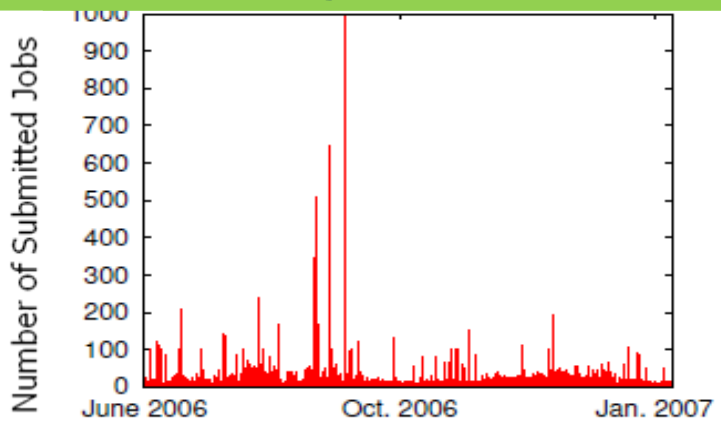
*The Grid Workloads Archive: http://gwa.ewi.tudelft.nl/pmwiki/

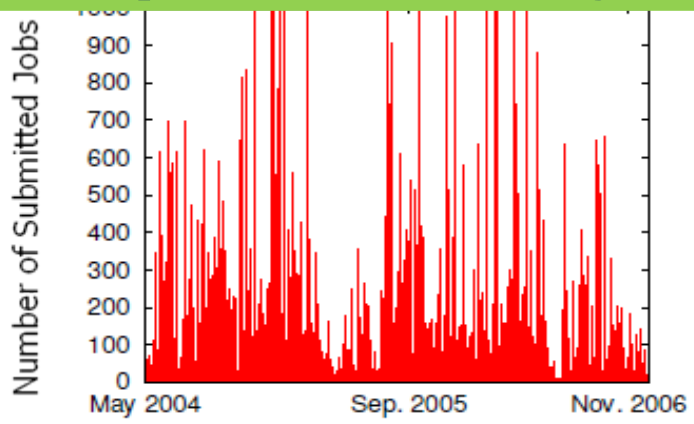
Grid Workload Traces:
Bursty Job Arrivals (5 minute intervals)





Bursty arrivals reduce predictability!





TU Delft

Research Questions

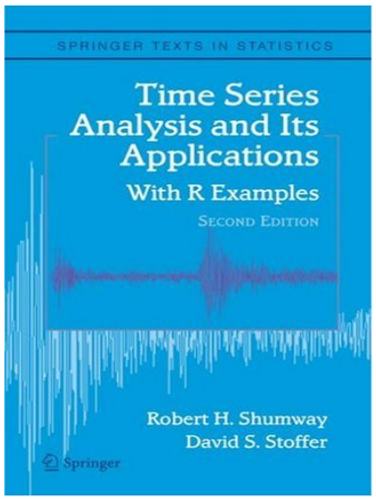
1. What is the performance of **job runtime** predictors in grids?

2. What is the performance of **queue wait time** predictors in grids?

3. Can **prediction-based grid scheduling** policies perform better than traditional policies?

Job Runtime Predictions

- We have evaluated the accuracy of five time series methods under four job classifications



- **Time series methods**
  - Last
  - Last2
  - Running Mean (RM)
  - Sliding Median (SM)
  - Exponential Smoothing (ES)

Job Runtime Predictions

- **Job Classification Methods**
  - Create classes according to job attributes
  - Site, User, User on Site,
    (User + Application Name + Job Size) on Site

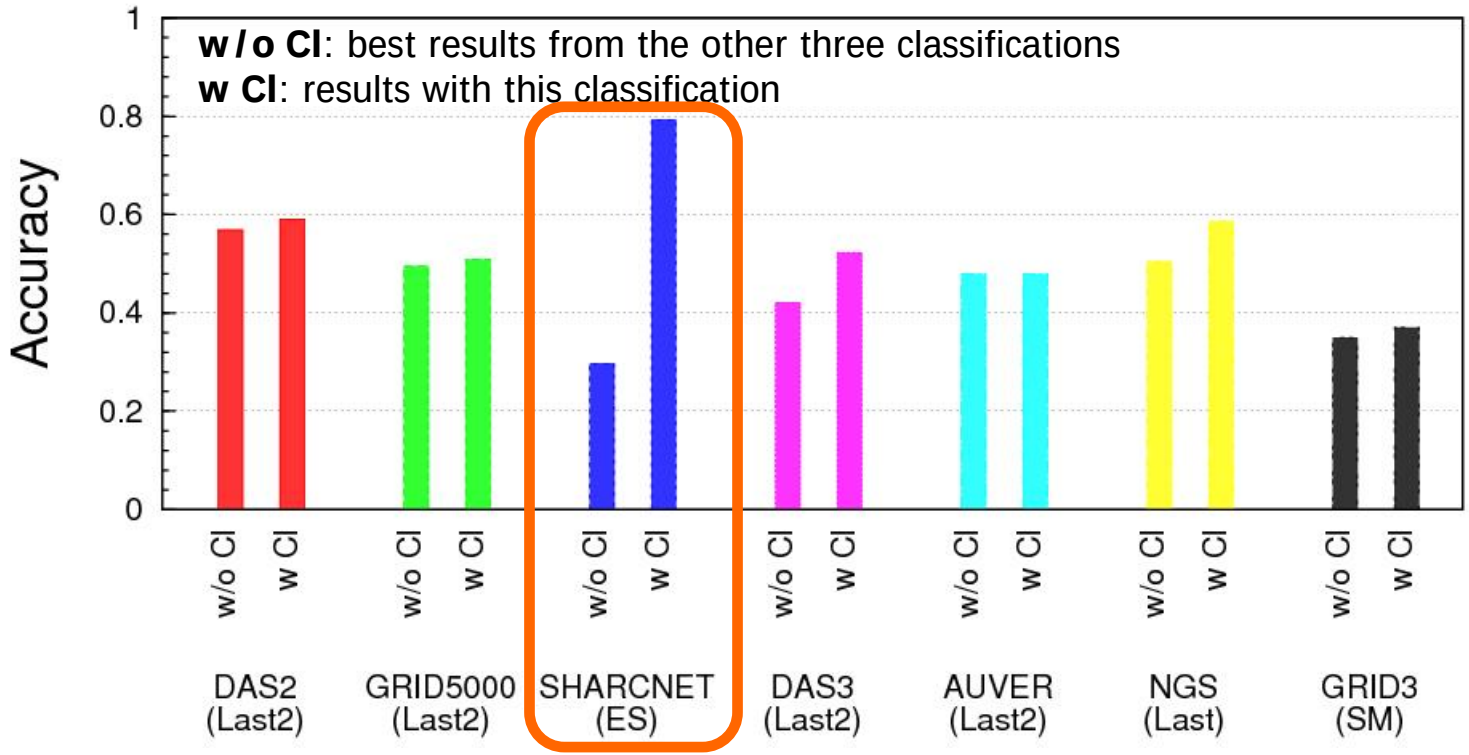
- **Performance Metric**

$$accuracy = \begin{cases} 1 & \text{if } P = T_r, \\ T_r/P & \text{if } P > T_r, \\ P/T_r & \text{if } P < T_r, \end{cases}$$

P : Predicted runtime
Tr : Actual runtime

Job Runtime Predictions

Classification: (User + Application Name + Job Size) on Site

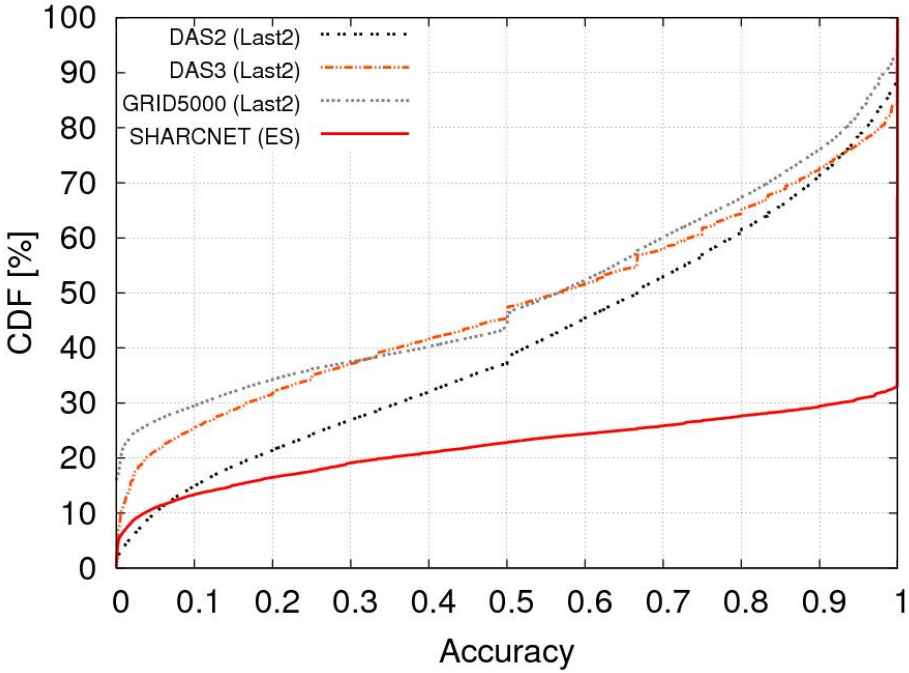


More specific classification improves the accuracy
No dominant prediction method

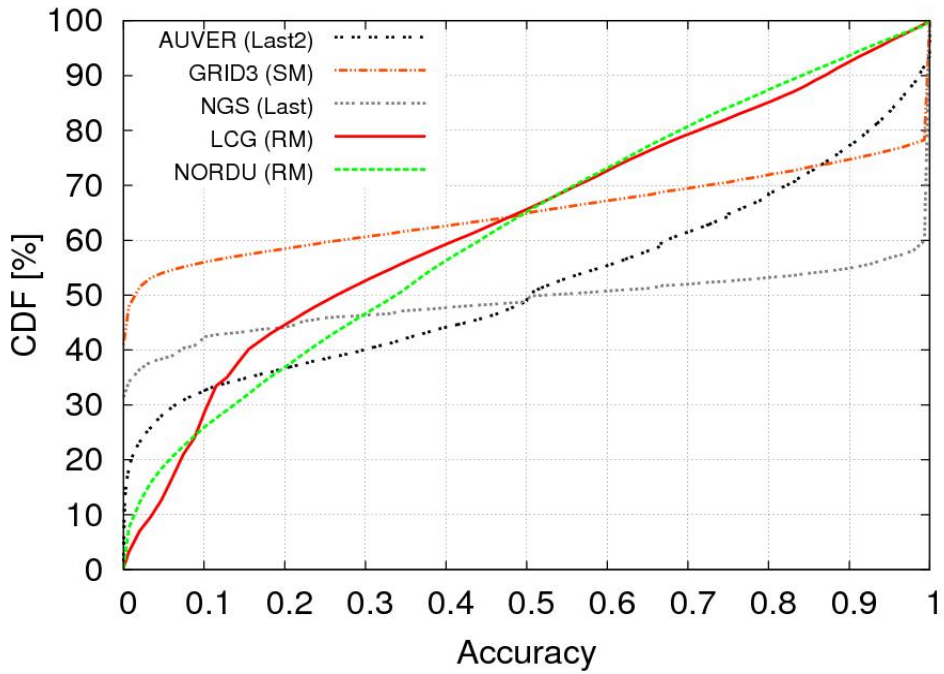
TU Delft

Job Runtime Predictions

Research Grids



Production Grids



Lower curves have higher accuracy

Job runtimes are predicted more accurately in research grids

TU Delft

Job Runtime Predictions: Summary of the results

- More specific classification improves job runtime prediction performance

- Job runtime prediction accuracy is low across all grids (except SHARCNET)
  - **Bursty Arrivals:** Same prediction error is made for all the jobs submitted together
  - Lack of **Stationarity**

    (no constant long-term mean and variance)

Queue Wait Time Predictions

- **Point-value predictions**
  - Simulate the local scheduling policy with predicted job runtimes to predict job queue wait times

- **Upper-bound predictions**
  - Predict upper bounds for queue wait times with a specified confidence level
  - Obviate the need to know the internal operation of local scheduling policies

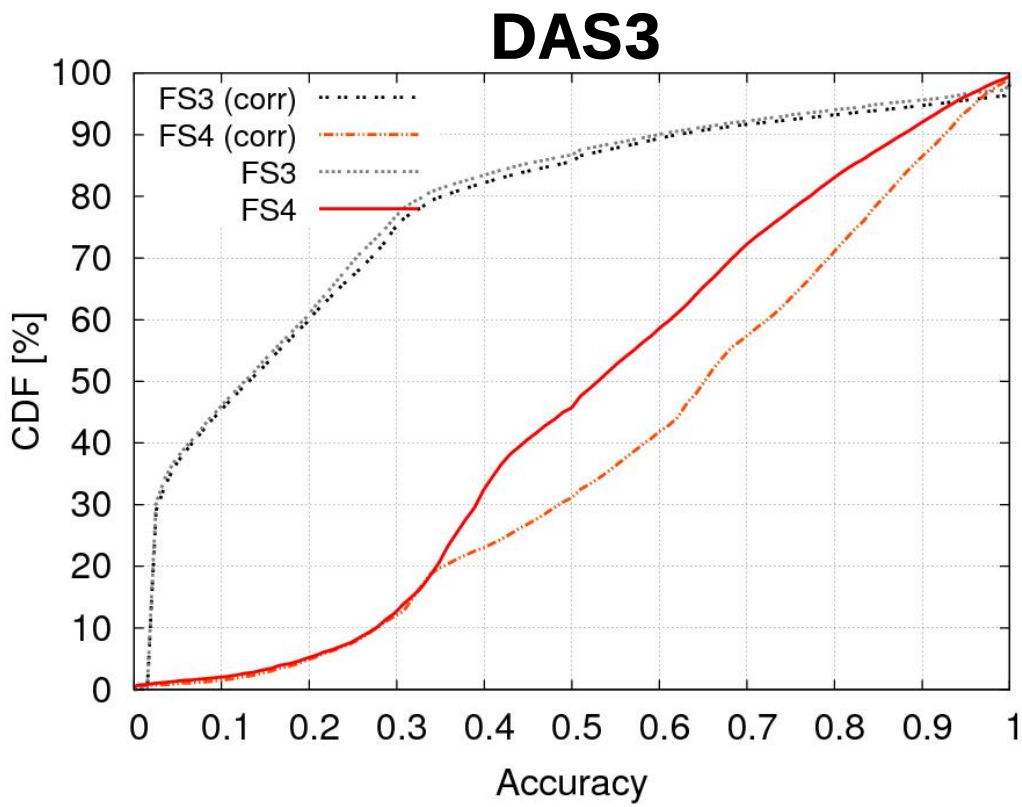
Point-Value Predictions

- **Simulation Model**
  - **FCFS** as the local scheduling policy
  - Jobs assigned to their original execution sites
  - A point-value predictor runs on each site
    - Job runtimes are predicted with **Last2**

- **Prediction Correction Mechanism**
  - On departure, update the predicted runtimes of both the queued and the running jobs accordingly

  - **Traces:** DAS2, DAS3, GRID5000, and AUVER

Point-Value Predictions



Accuracy of the point-value predictor is low
Correction mechanism improves the prediction accuracy (1% to 10%)

TU Delft

Upper-Bound Predictions

- Binomial Method Batch Predictor **(BMBP)***
  - Predicts the specified quantile of the wait time distribution with a specified confidence level

- A predictor based on **Chebyshev's Inequality**
  - No more than $1/k^2$ of the values are more than k standard deviations away from the mean

- We consider a quantile (for BMBP) and a confidence level of 95%

- **Traces:** DAS2, DAS3, GRID5000, and AUVER

* J. Brevik, D. Nurmi, and R. Wolski. *Predicting bounds on queuing delay for batch-scheduled parallel machines*. In PPoPP, pages 110–118, 2006.

Upper-Bound Predictions

BMBP

Grid-Site	Avg. Accuracy	Under-predictions	Perfect-predictions	Over-predictions
DAS2-FS1	0.50	8%	9%	83%
DAS3-FS4	0.41	15%	4%	81%
Auver-clr01	0.20	12%	1%	87%
GRID5K-G1	0.72	20%	0%	80%

Chebyshev

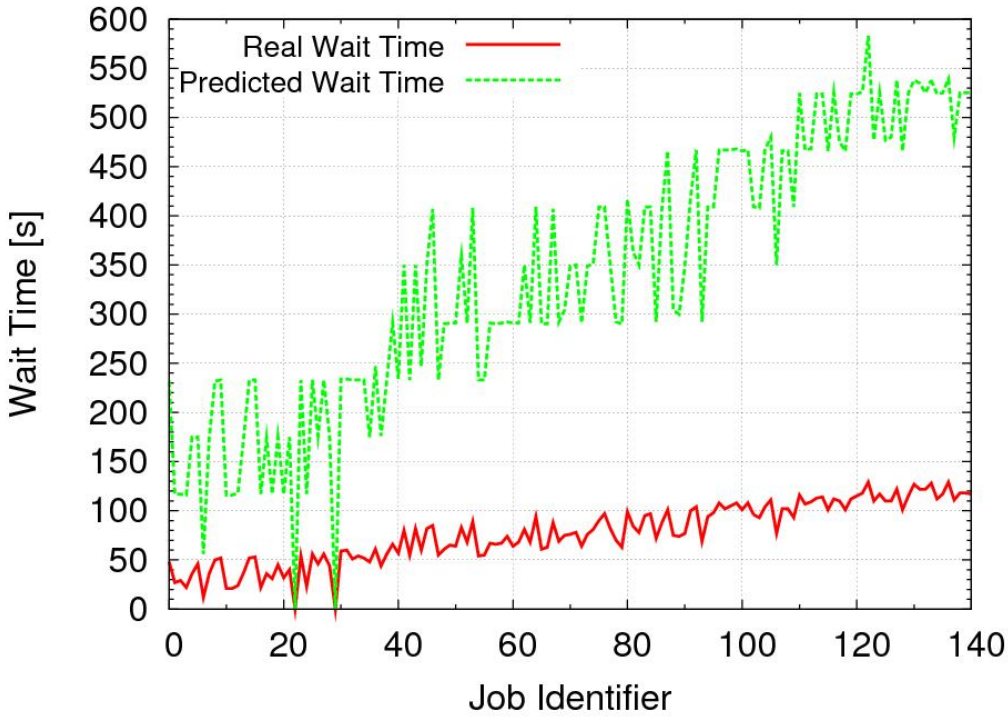
DAS2-FS1	0.21	8%	0%	92%
DAS3-FS4	0.23	7%	1%	82%
Auver-clr01	0.10	7%	0%	93%
GRID5K-G1	0.24	16%	0%	84%

Trade-off between accuracy and tightness of the upper bounds

Upper-Bound Predictions

- Both BMBP and Chebyshev fail when jobs arrive in bursts
- **User runtime estimates**, if available, can also be used in predicting upper bounds

A burst period of DAS3-FS4



TU Delft

Performance of Prediction-Based Grid Scheduling

- **Global Scheduling Policies**
  - Earliest Completion Time (ECT)-Perfect
  - ECT-Last2

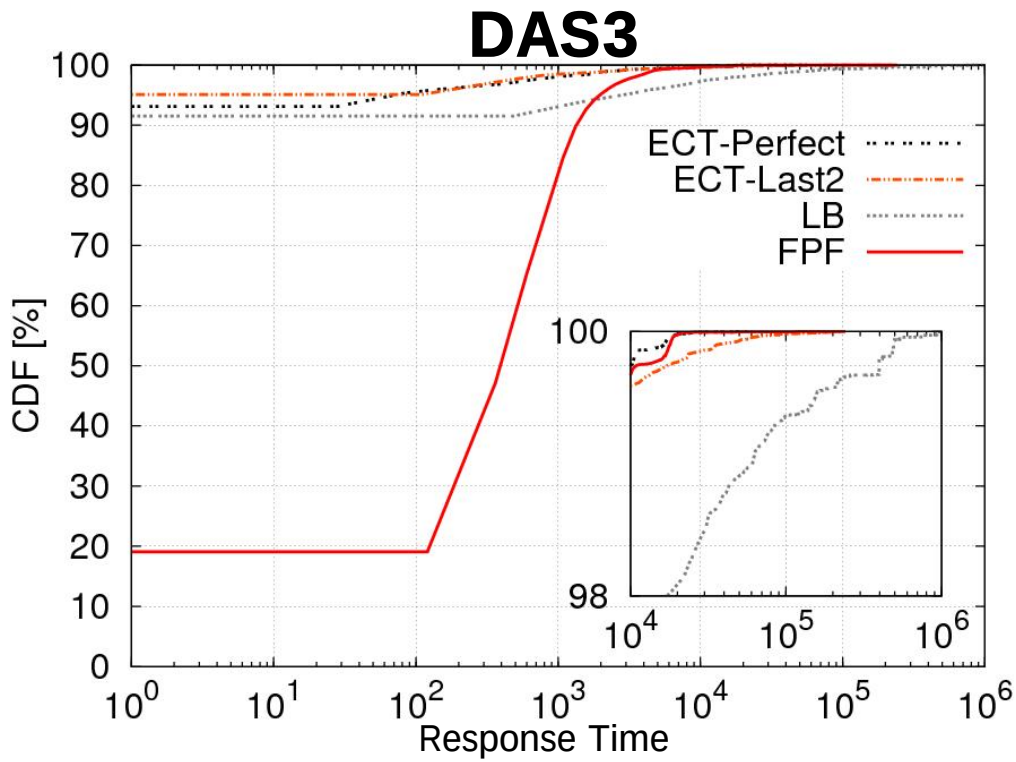
  Prediction-based
  - Load Balancer
  - Fastest Processor First (FPF)

  Traditional

- **Simulation Model**
  - DAS3 and AUVER
  - Jobs arrive to a global scheduler
  - A point-value predictor runs on each cluster (Last2+Correction)

Trace	Period	Number of Jobs	Avg. Util.
DAS3	July-Oct. 2008	~220,000	~30%
AUVER	Aug.-Nov. 2006	~90,000	~70%

Performance of Prediction-Based Grid Scheduling

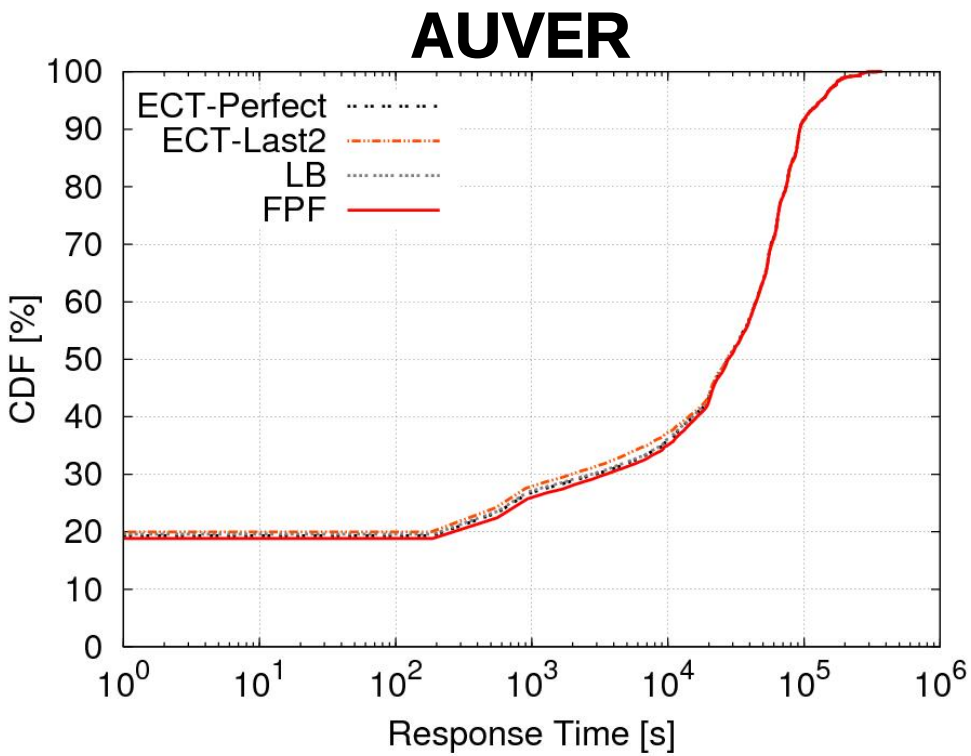


DAS3	ECT-Perfect	ECT-Last2	LB	FPF
Avg. Response Time [s]	1320	1400	4318	1911
Avg. Wait Time [s]	105	186	3061	681

Prediction-based policies perform better

TU Delft

Performance of Prediction-Based Grid Scheduling



AUVER	ECT-Perfect	ECT-Last2	LB	FPF
Avg. Response Time [s]	40951	41003	40959	41334
Avg. Wait Time [s]	6515	6574	6534	6898

All policies have similar performance

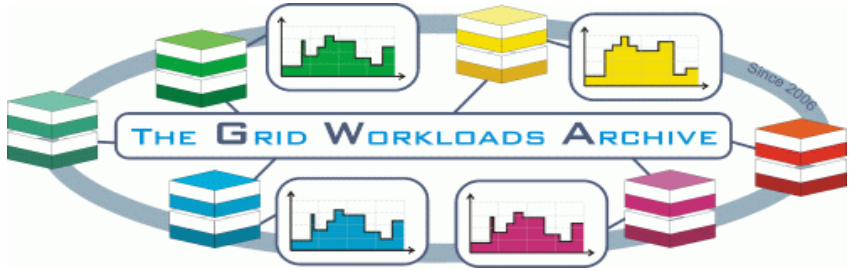
TU Delft

Conclusion

- We presented a systematic evaluation of job runtime and queue wait time predictions in grids using **real traces**
  - Simple time-series methods revealed low accuracy
  - Current predictors cannot handle bursty arrivals
  - More accurate predictions do not imply a better performance of grid scheduling

- **Future Work**
  - Simple vs. Complex (AI-based) prediction methods

Questions?



More Information:

- The Grid Workloads Archive: http://gwa.ewi.tudelft.nl/pmwiki/
- DGSim: www.pds.ewi.tudelft.nl/~iosup/dgsim.php
- see PDS publication database at: www.pds.twi.tudelft.nl/

email: o.o.sonmez@tudelft.nl

This work was carried out in the context of the Virtual Laboratory for e-Science project (www.vl-e.nl). Part of this work is also carried out under the FP6 Network of Excellence CoreGRID funded by European Commision.

PDS Group

TU Delft