Measuring the Robustness of Resource Allocations for Distributed Computer Systems in a Stochastic Dynamic Environment

Jay Smith[1], Luis D. Briceño[2], Anthony A. Maciejewski[2], Howard Jay Siegel[2,3], Timothy Renner[2], Vladimir Shestak[2], Joshua Ladd[4], Andrew Sutton[3], David Janovy[2], Sudha Govindasamy[2], Amin Alqudah[2], Rinku Dewri[2], and Puneet Prakash[2]

[1]IBM, Colorado State University: [2]Electrical and Computer Engineering Department, [3]Computer Science Department, [4]Mathematics Department

Introduction

Events that attract world wide audiences such as the World Cup football tournament (soccer for you westerners), implement distributed computing systems to support the aggregation and dissemination of the generated data. During the summer of 1998, the web site for the World Cup processed more than 1.3 billion HTTP requests. Robust allocation of available computing resources so that web requests for such large events are handled in a responsive manner is the focus of this research.

Problem Statement

Allocate computing resources such that incoming HTTP requests have a high probability of completing before the user's browser times out.

Performance Metric

The evaluation of a resource allocation heuristic is based on minimizing the number of tasks that miss their deadline.

Definitions

- Incoming HTTP request is a computing task (*i*) that has an associated deadline (β_i^{max}).
- Received tasks belong to set of known tasks.
- Tasks are grouped into predefined classes (*C*) that represent relative complexity of executing the task.
- Each class contains a task execution time probability distribution for each machine in the set of machines (*M*).
- Task arrival times follow the observed traffic patterns of the 1998 World Cup.
- Resource mapping heuristics allocate tasks to machines.
- Mapping events occur when a task arrives and when a task completes.
- Mapping time is limited.
- Machine completion time is equal to the time when all tasks that have been mapped to it have completed.

Simulation Setup

Eight machines, 1024 tasks, 10 simulation runs used to predict DSRM value, 100 runs for performance metric evaluation.

Stochastic Robustness Metric (SRM)

Convolution is used to combine a task's execution time distribution with a machine's completion time distribution to obtain a task's completion time distribution.

SRM at a given time is the product of the probabilities of the task completion time distributions.

Dynamic SRM (DSRM) is defined as the average of the SRM values. Determining the DSRM over a limited number of SRM values is sufficient to indicate the performance of the heuristic.

Heuristics

Three heuristics were used to evaluate the effectiveness of the dynamic SRM value. During selection, ties are resolved arbitrarily.

Two Phase Greedy (TPG)

- While not all tasks mapped:
- Phase 1: pair each task with a machine where the task can complete before its deadline; otherwise pick any machine.
- Phase 2: from the pairs in Phase 1, map the pair that minimizes the performance metric and update the machine completion time.
- Repeat.

Segmented Two Phase Greedy (STPG)

- Tasks are assigned a weight based on their probability to complete before their deadline (smaller probability = smaller weight).
- Tasks are sorted in ascending order by their average weighted expected completion time over all machines.
- Sorted tasks divided into *n* segments.
- TPG applied to each segment.

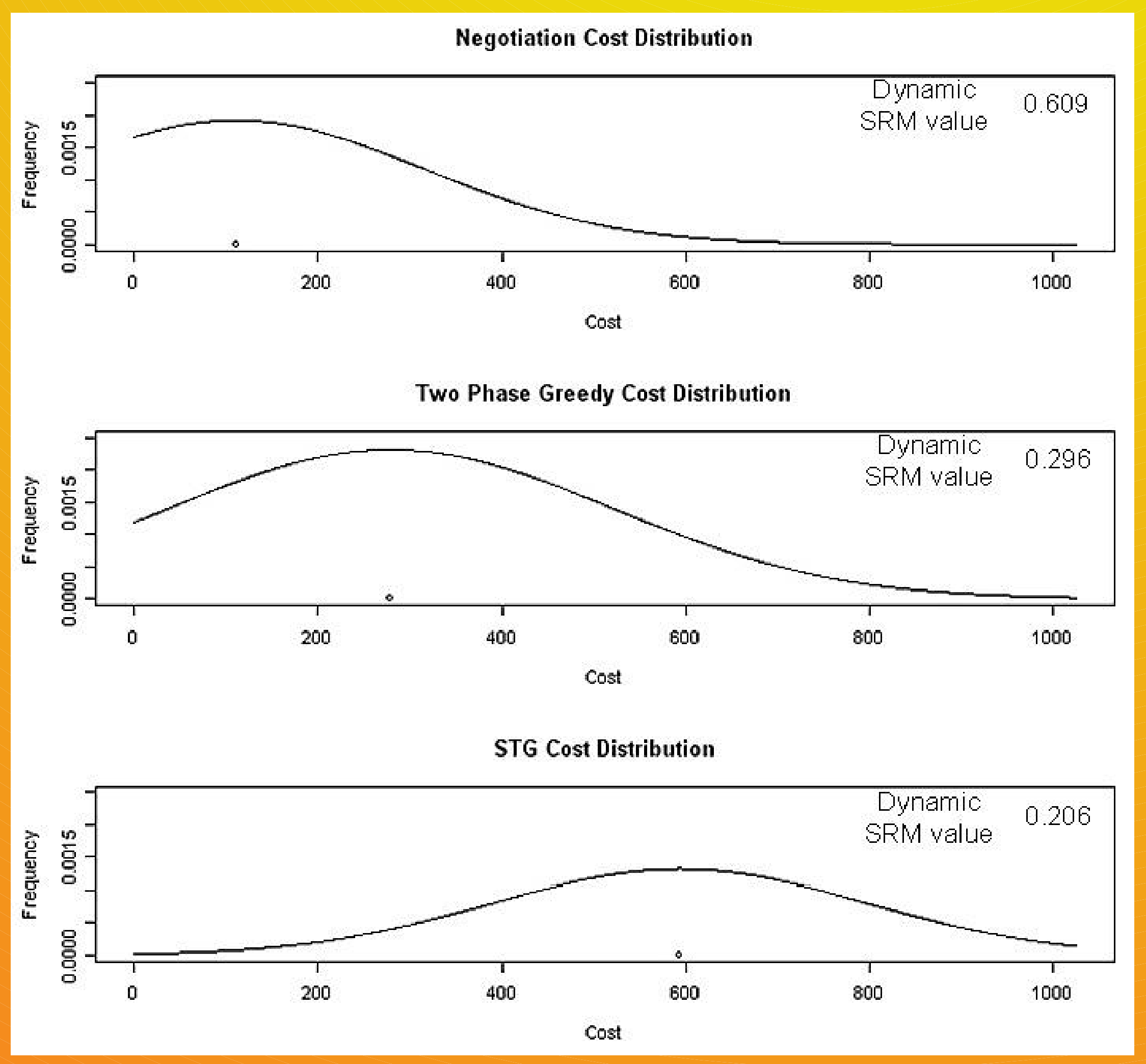
Negotiation

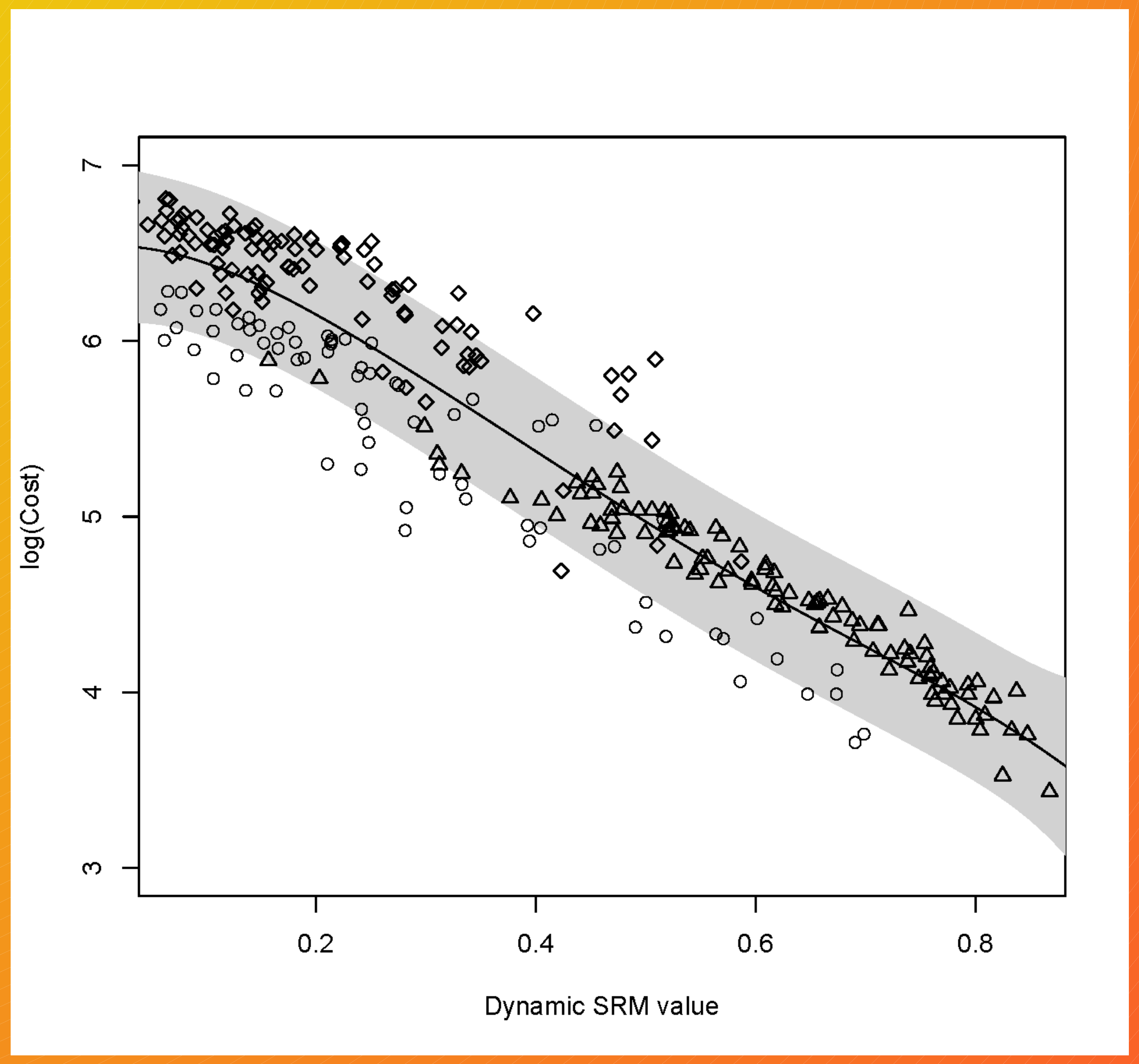
- Similar to hill-climbing search where the best solution is always kept.
- Total ordering of tasks is iteratively permutated.
- Task local earliness metric (*LEM*) = deadline – expected completion time.
- Global earliness metric (*GEM*) for an ordering is the sum of the task *LEMs*.
- Ordering with best GEM is kept.
- Two tasks are swapped and process is repeated.

Results

The dynamic stochastic robustness metric (DSRM) has an inverse relationship to the performance metric. The figure (below, left) shows the cost distributions for the three heuristics. The small circle indicates the mean of each distribution. The graphs were generated using a kernel density estimator with a Gaussian distribution.

The plot on the right shows the DSRM value versus the logarithm of the costs for all three heuristics. A Bayesian regression model has been used to create the curve shown in the plot and the shaded area represents one standard deviation from the curve.





Reference: Jay Smith, Luis D. Briceño, Anthony A. Maciejewski, Howard Jay Siegel, Timothy Renner, Vladimir Shestak, Joshua Ladd, Andrew Sutton, David Janovy, Sudha Govindasamy, Amin Alqudah, Rinku Dewri, and Puneet Prakash, "Measuring the Robustness of Resource Allocations in a Stochastic Dynamic Environment," submitted.

This research was supported by the NSF under Contract No: CNS-0615170, by the Colorado State University Center for Robustness in Computer Systems (funded by the Colorado Commission on Higher Education Technology Advancement Group through the Colorado Institute of Technology), and by the Colorado State University George T. Abell Endowment.