Event Aware Elasticity Control

Matthew Sladescu
Supervisors: Alan Fekete, Anna Liu, Kevin Lee
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

INTRODUCTION

One of the most attractive features offered by cloud computing is elasticity, where resource provisioning policies can acquire and release compute resources on an as-needed basis.

The popularity of online events like product announcements, sale events, and auction *events have often drawn flash crowds that can render systems with inadequate resource provisioning control policies unavailable.*

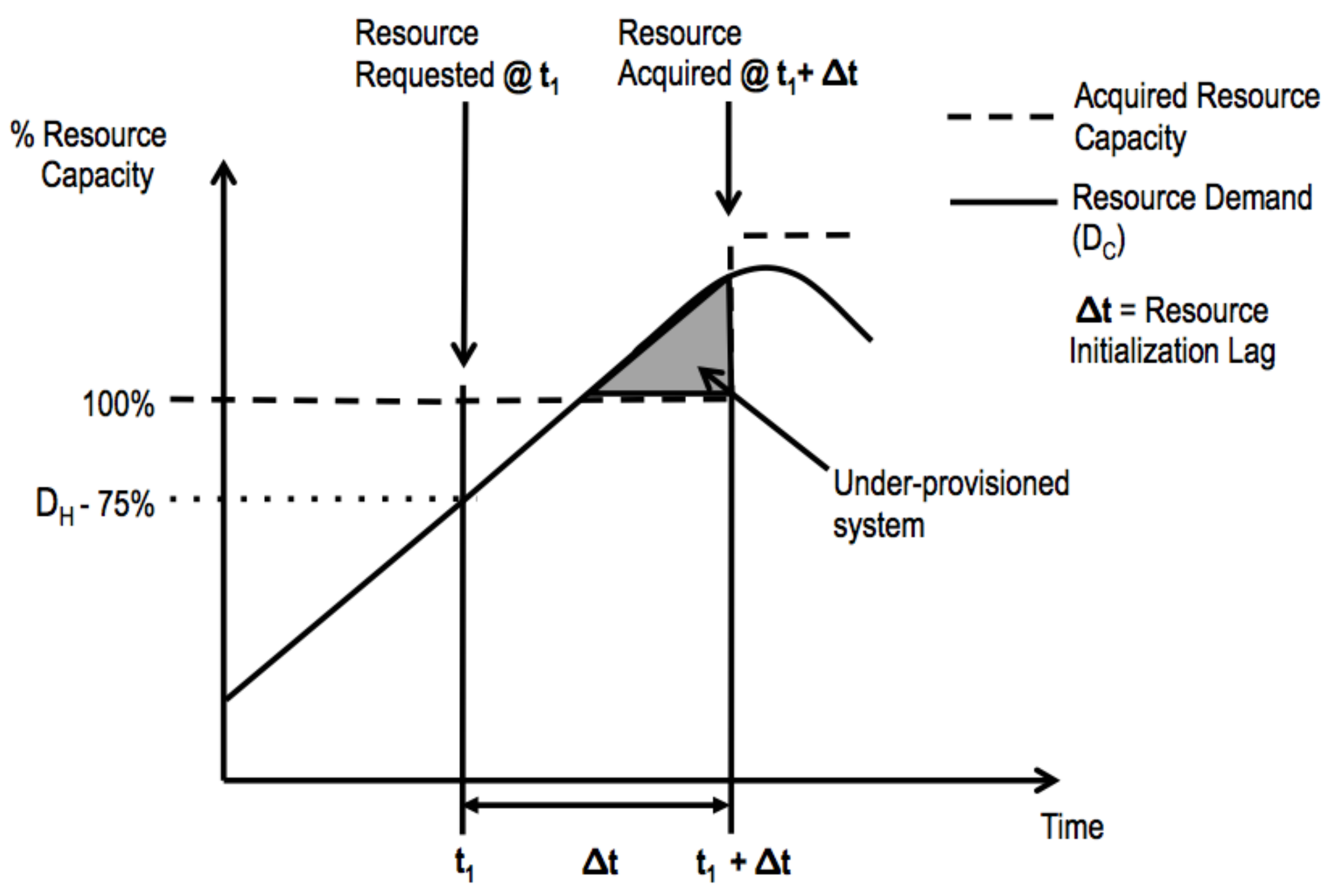
We propose a novel event-aware prediction approach, for use in a resource provisioning control policy, which can more effectively provision resources on-time to serve event-associated bursts in a cloud computing environment.

EXISTING RESOURCE PROVISIONING APPROACHES

The current resource provisioning control approaches are often classified as either being reactive or predictive (proactive):

REACTIVE CONTROL APPROACHES

- Command issued to acquire more resource capacity when resource demand exceeds a high level threshold (D_H)


% Resource Capacity
Resource Requested @ t_1
Resource Acquired @ $t_1 + \Delta t$
Acquired Resource Capacity
Resource Demand (D_c)
Δt = Resource Initialization Lag
100%
D_H - 75%
Under-provisioned system
Time
t_1
Δt
$t_1 + \Delta t$

- Resource Demand (D_C) can exceed acquired capacity while acquired resources are initializing (Δt), leaving a system under-provisioned.

PREDICTIVE CONTROL APPROACHES

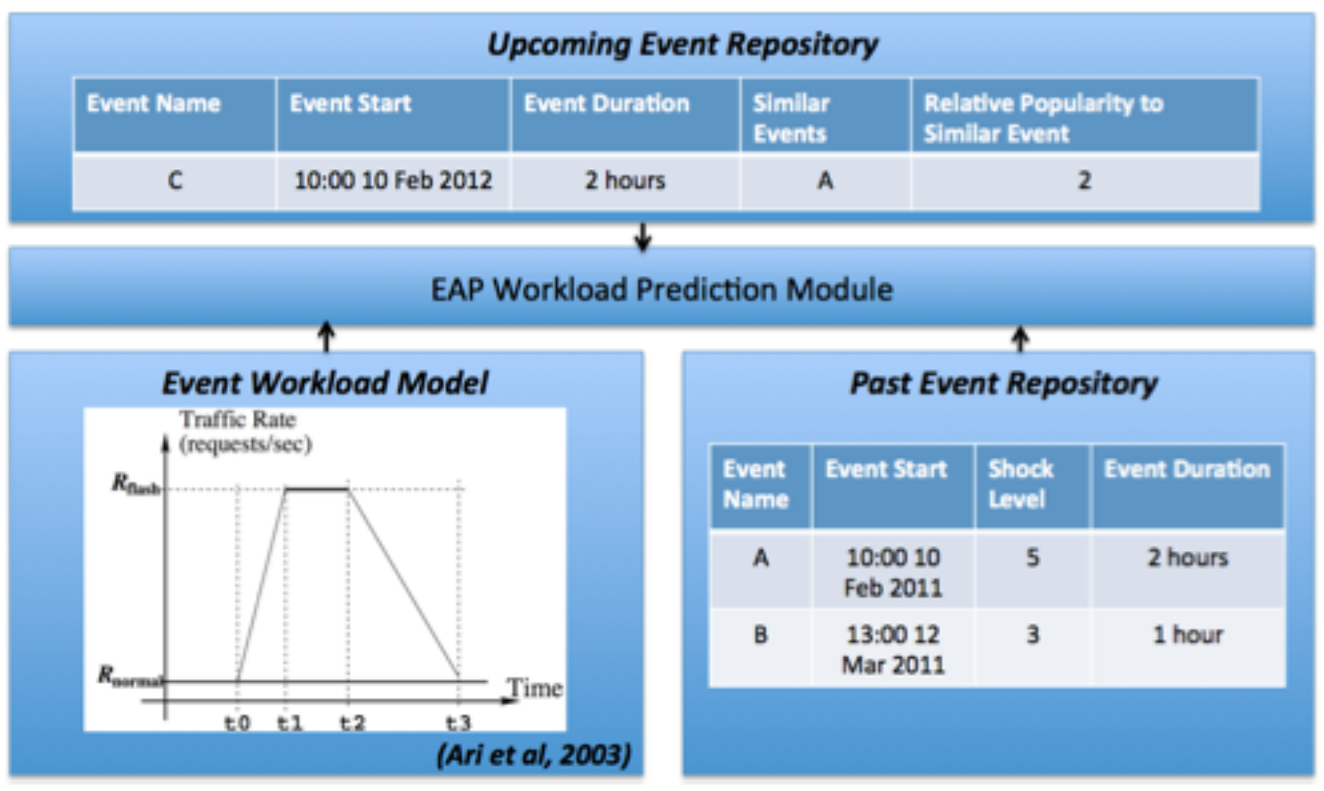
Prediction of future resource demand can assist with:

- Overcoming initialization lag issues by pre-provisioning resources in anticipation of increased demand.
- Cost effective resource acquisition (eg: allows informed decisions about when to "reserve" instances for increased demand).
- Devising smarter event scheduling policies (eg: scheduling several smaller events, at separate times, that can be handled by fewer resources instead of scheduling one large event).

The most commonly and recently used methods for load prediction use only past load history, in combination with methods like artificial neural networks and support vector regression to predict load. *These methods have been found to be ineffective [1] for predicting workload bursts.*

EVENT AWARE PREDICTION

In contrast to existing approaches, we propose an *Event Aware Prediction approach (EAP), which recognizes and makes use of the inherent link often found between events and workload bursts.* EAP makes use of information about how workload was influenced by events in the past, to predict how events will influence workload in the future, using the two phases described below:



Upcoming Event Repository

Event Name	Event Start	Event Duration	Similar Events	Relative Popularity to Similar Event
C	10:00 10 Feb 2012	2 hours	A	2

EAP Workload Prediction Module

Event Workload Model

Past Event Repository

Event Name	Event Start	Shock Level	Event Duration
A	10:00 10 Feb 2011	5	2 hours
B	13:00 12 Mar 2011	3	1 hour

EAP Framework Elements [2]

Training Phase

Information from a repository of past events, (where any details about events like announcement time, product announced, or teams playing can be stored, together with load history for each past event), is used to build a workload model that describes how workload fluctuated during the events included in the repository of past events.

Prediction Phase

Workload is predicted for a given time, $t_{i,}$ by finding all events in the upcoming event repository that are active at the time t_i, and using the workload model derived in the training phase to predict the load for each event at time $t_{i.}$

EVALUATING EAP

EAP's burst location prediction accuracy has so far been assessed in comparison to:

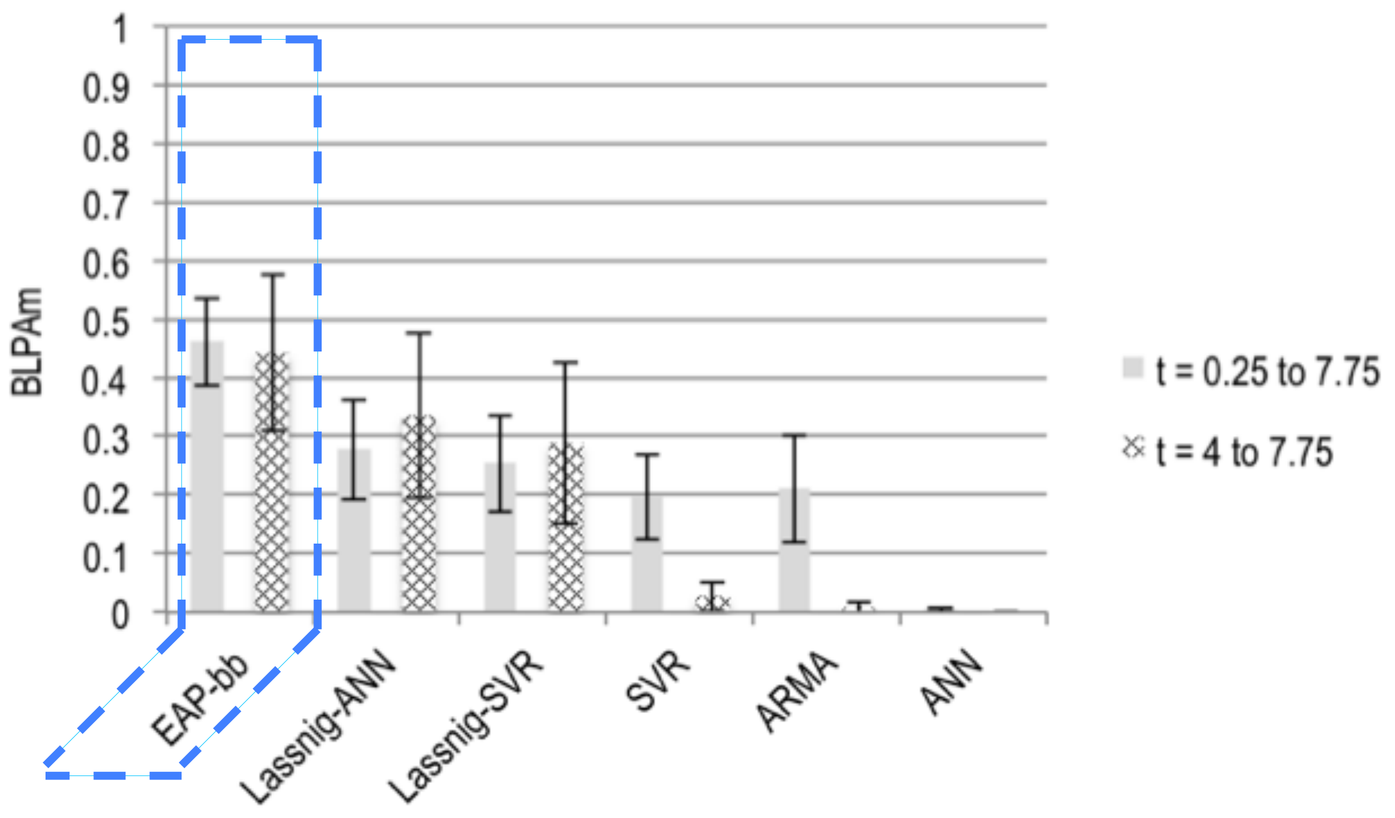
- Artificial Neural Networks (ANN)
- ARMA
- Support Vector Regression (SVR)
- A recently developed approach dedicated to predicting burst location [1].

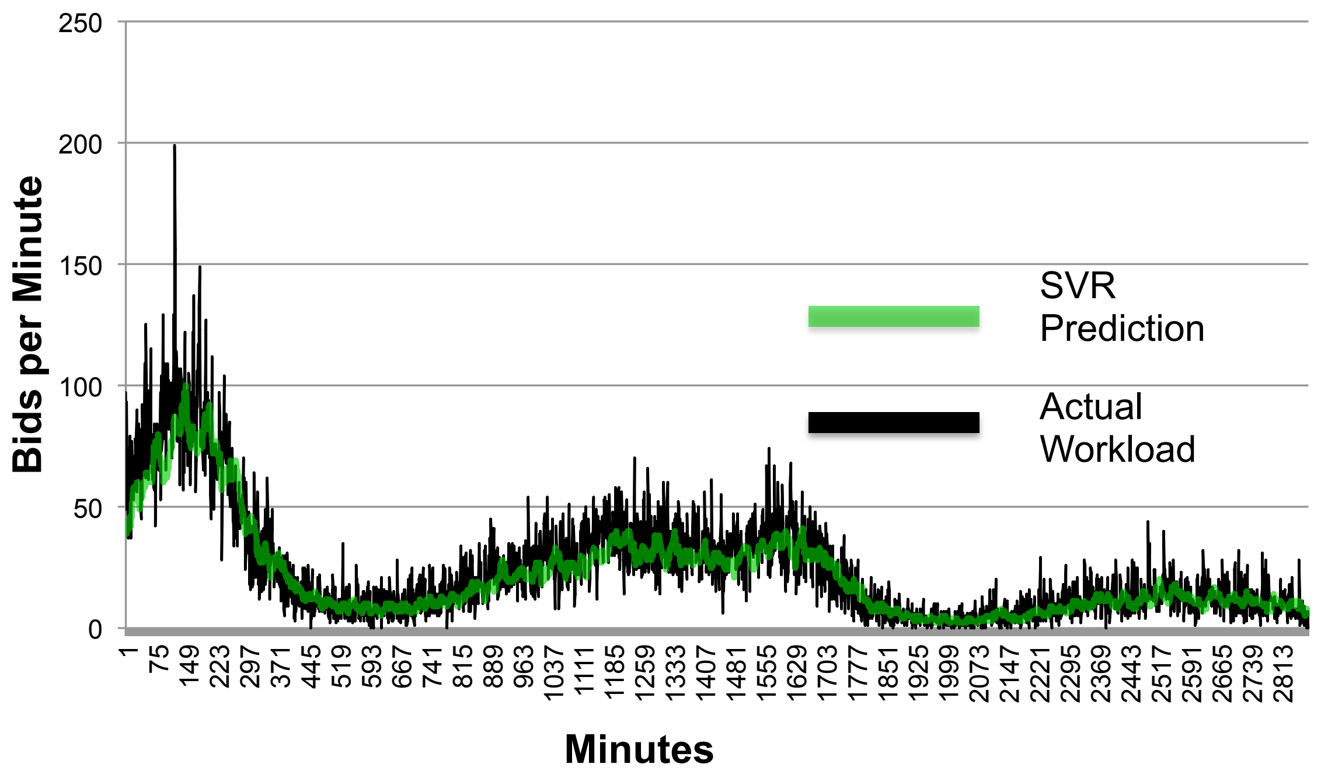
Real world data-sets are used to assess prediction accuracy performance:

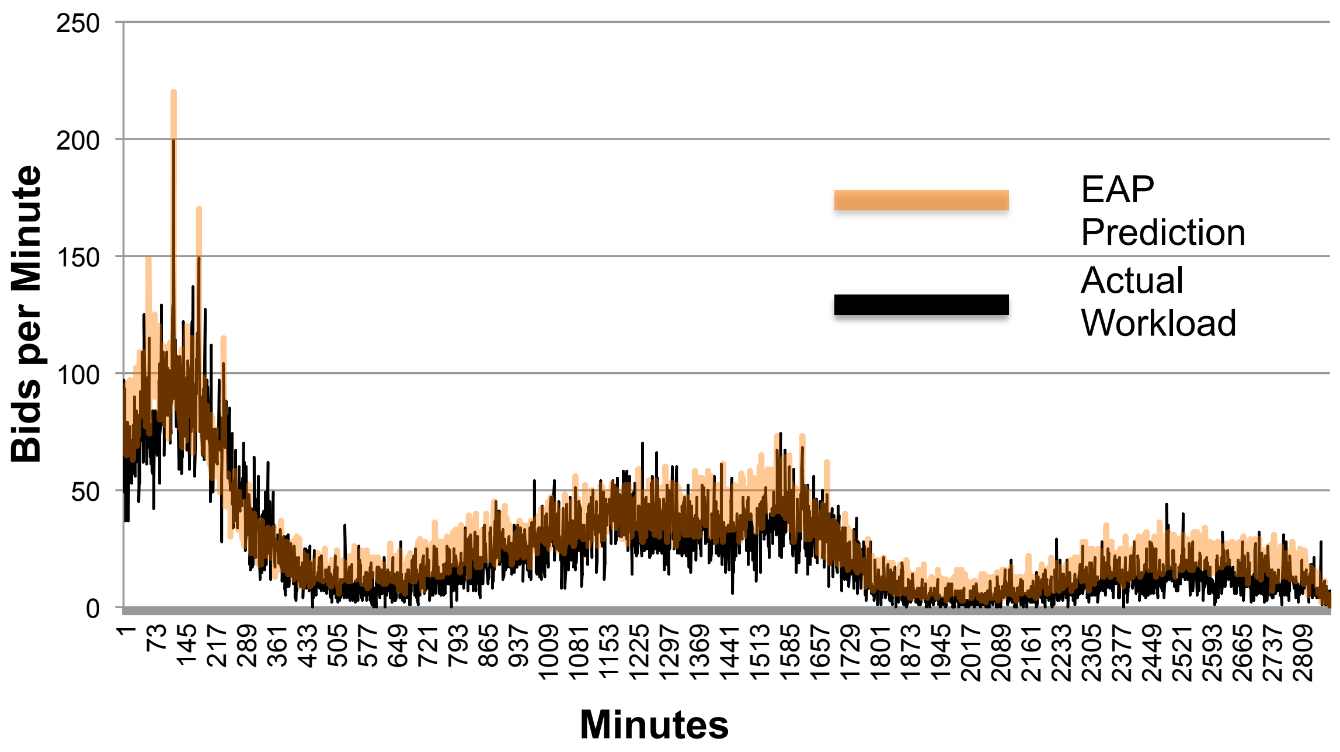
- Online auction workload traces
- Wikipedia load traces
- World Cup Soccer website load traces

RESULTS

Burst location prediction accuracy (BLPAm) results [2] for a 15 minute prediction horizon on real bidding workload are shown below. Higher "t" values correspond to more severe bursts.







Experiments are currently underway to assess prediction performance using Wikipedia load traces, and further experiments to evaluate performance using world cup soccer load traces are scheduled.

CONCLUSION

The results shown above indicate the superior burst prediction accuracy of EAP relative to all other methods tested, and concur with [1] that traditional methods like ARMA, ANN, and SVR are ineffective in predicting workload bursts when only using knowledge about the past. The advantage of EAP is in its use of often readily available information about future events to predict workload.

REFERENCES

[1] Lassnig, M. et. al.: Identification, Modelling and prediction of Non-periodic Bursts in Workloads. In: Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. pp. 485–494. IEEE Computer Society (2010)

[2] Sladescu, M. et. al: Event Aware Workload Prediction: A Study Using Auction Events. In: Proceedings of the 2012 13th Web Information System Engineering

THIS RESEARCH IS SPONSORED BY