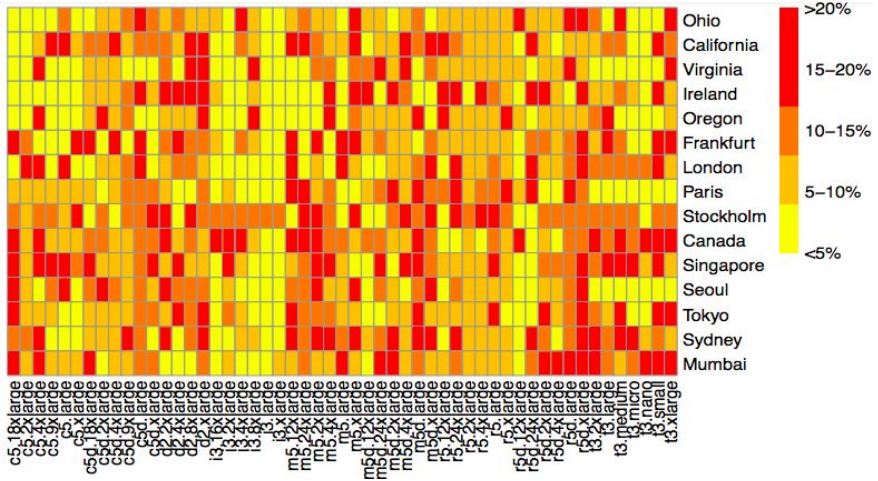
An Approximate Bribe Queueing Model for Bid Advising in Cloud Spot Markets

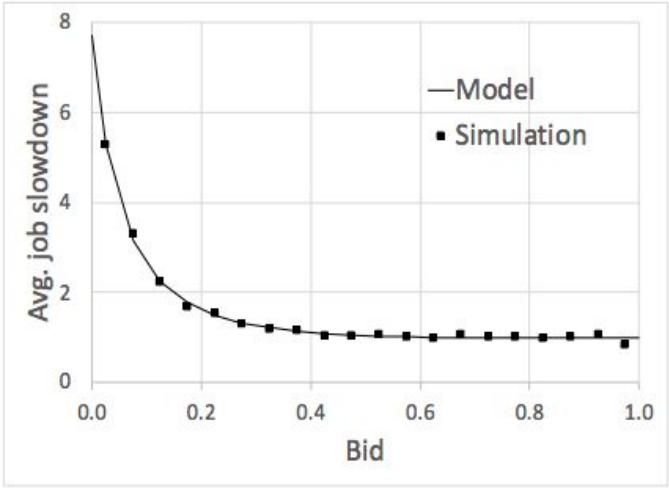
Bogdan Ghit and Asser Tantawi

QEST 2021

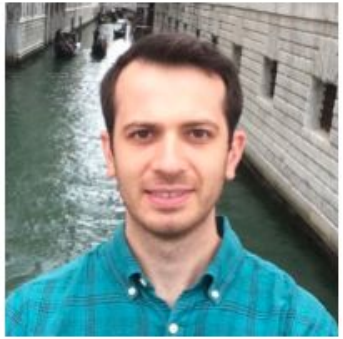
An Approximate Bribe Queueing Model for Bid Advising in Cloud Spot Markets

Bogdan Ghit and Asser Tantawi





- ➔ Simple, approximate job slowdown expression as a function of bid
- ➔ Validated prediction accuracy using simulations



Tech Lead | Sr. Software Engineer @ Databricks
- BI tools performance

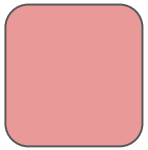
Research Intern @ IBM T.J. Watson
- Spot market analysis

PhD @ TU Delft
- Scheduling in datacenters

https://bogdanghit.github.io

Cloud Spot Markets

On-demand instance



Spot instance
Added when bid > spot price

Announcing low-priority VMs on scale sets now in public preview

Posted on May 3, 2018

Meagan McCrory, Senior Program Manager, Azure Compute

Microsoft Azure

Google Cloud Platform Blog

Product updates, customer stories, and tips and tricks on Google Cloud Platform

Introducing Preemptible VMs, a new class of compute available at 70% off standard pricing

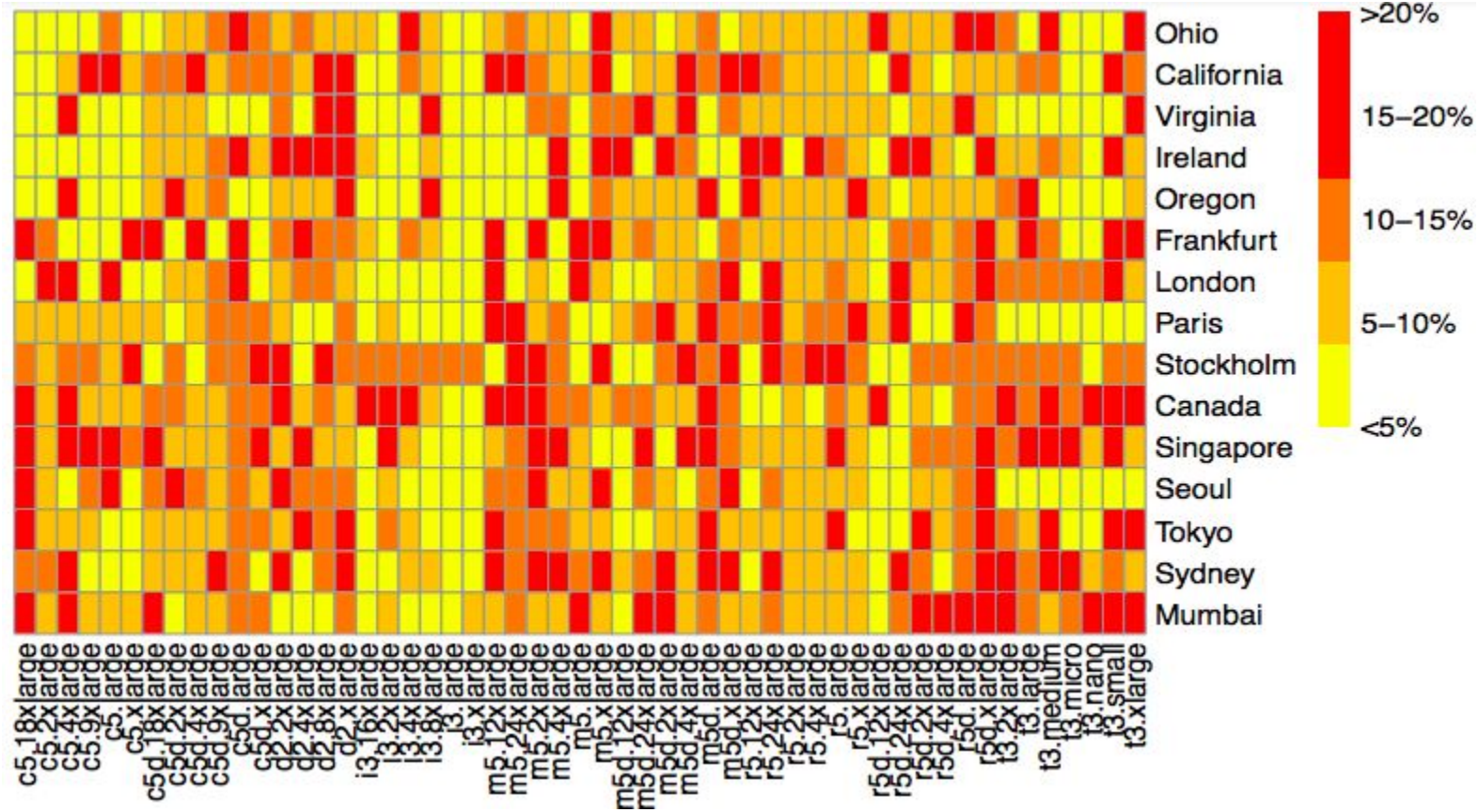
Monday, May 18, 2015

Announcing Amazon EC2 Spot Instances

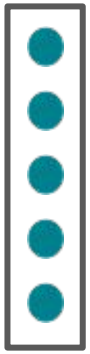
Posted On: Dec 14, 2009

We are excited to announce the introduction of Amazon EC2 Spot Instances, a new way to purchase and consume Amazon EC2 Instances. Spot Instances allow customers to bid on unused Amazon EC2 capacity and run those instances for as long as their bid exceeds the current Spot Price. The Spot Price changes periodically based on supply and demand, and customers whose bids meet or exceed it gain access to the available Spot Instances. Spot Instances are complementary to On-Demand Instances and Reserved Instances, providing another option for obtaining compute capacity. If you have flexibility in when your applications can run, Spot Instances can significantly lower your Amazon EC2 costs. Additionally, Spot Instances can provide access to large amounts of additional capacity for applications with urgent needs. To learn more, please visit the Amazon EC2 Spot Instances detail page.

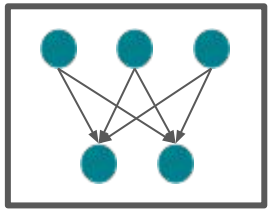
Interruption Rates for AWS Spot Instances



Cloud Analytics Applications

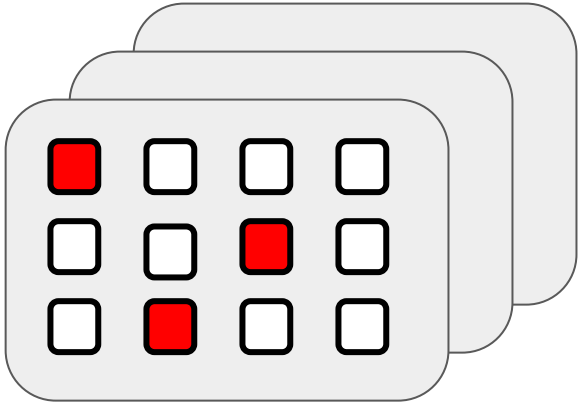


Bag of tasks



Workflow

Flexible execution model with allocation that can grow and shrink over time

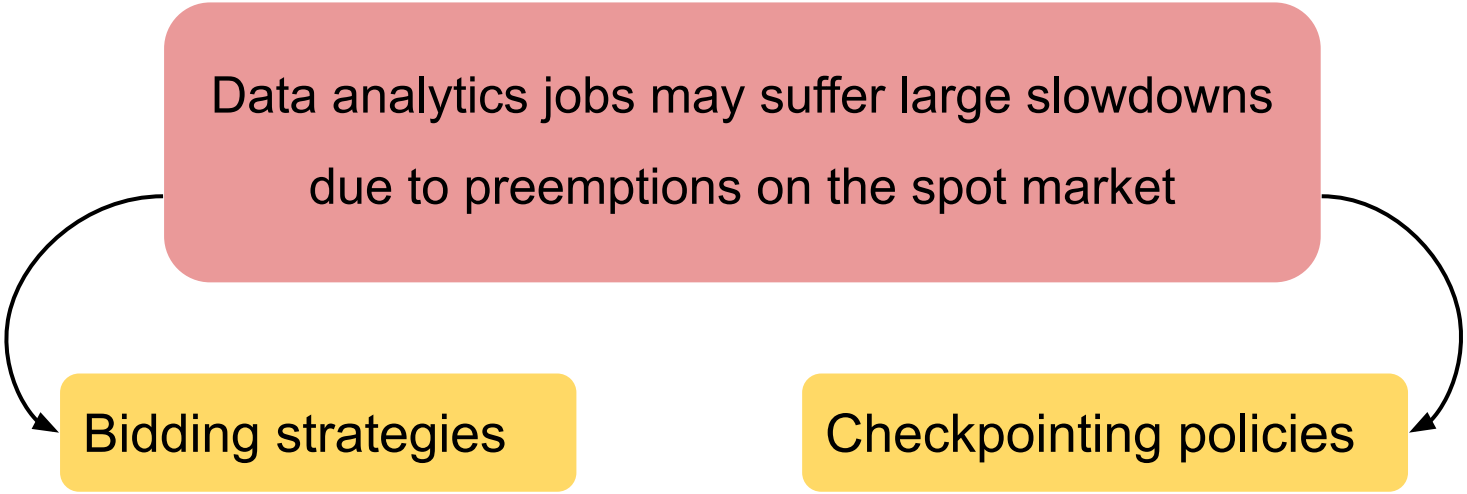


Data dependencies lead to recomputations when tasks are preempted due to instance revocation

Preemption prolongs the job runtime

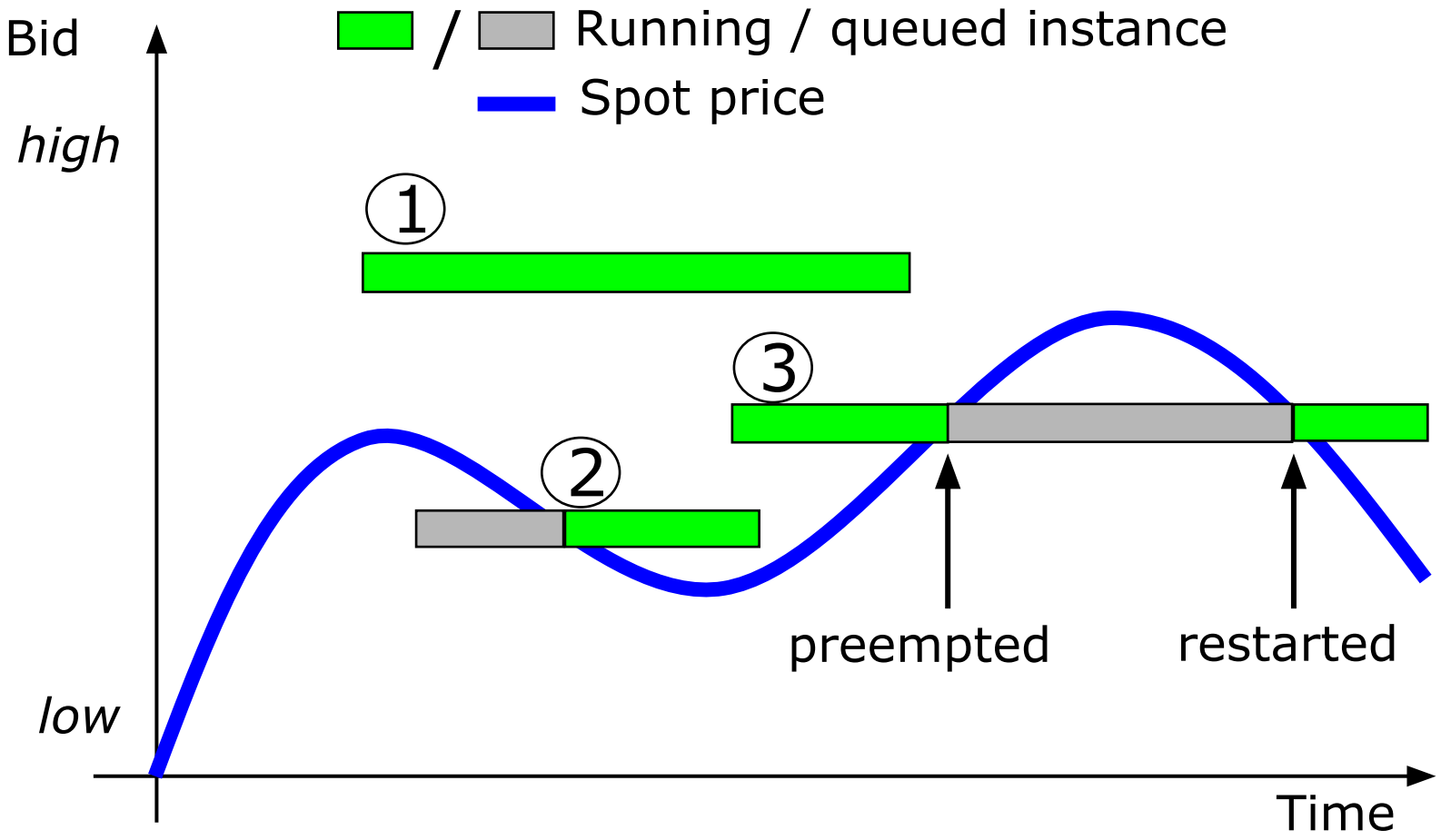
- HotCloud'10
- HPDC'12
- HPDC'17

This Work

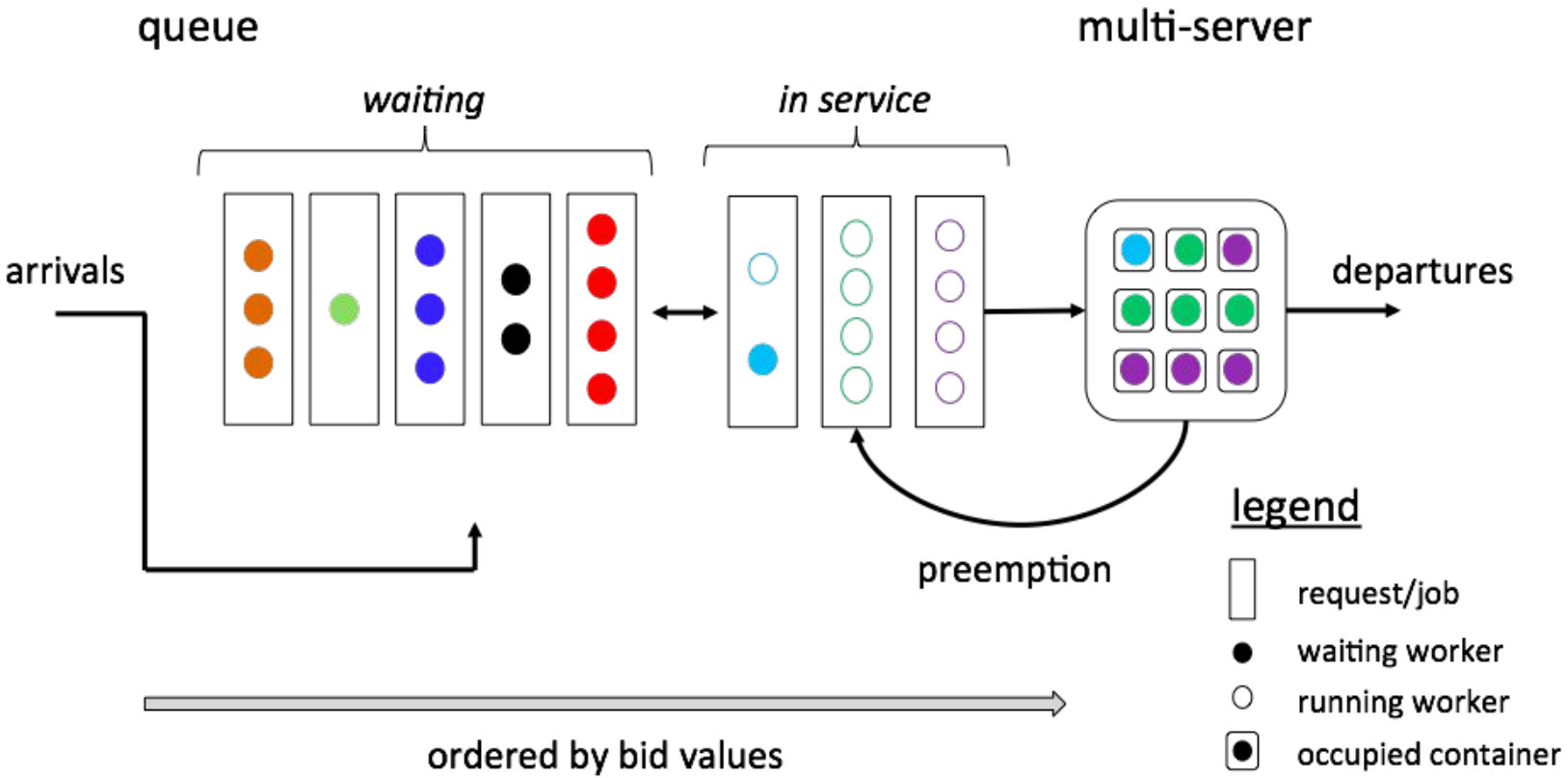


Our goal: provide an expectation of the job slowdown as a function of the bid value

Elastic Applications

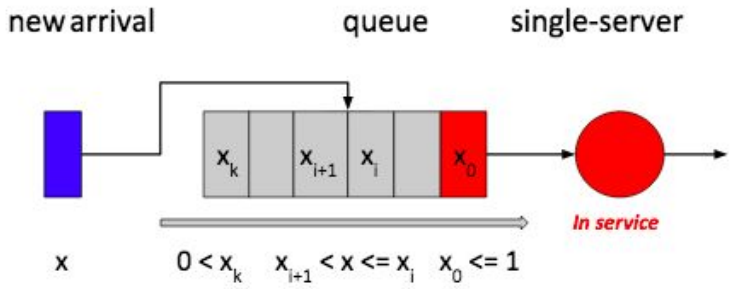


Multi-worker, multi-server bribing queue



Bribery Queueing Model

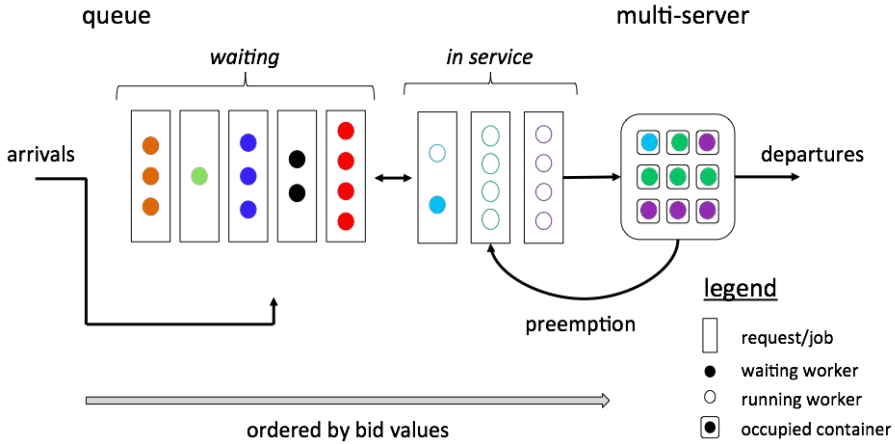
M/M/1 bribing queue



$$S(x) = \frac{1}{(1 - \rho(1 - B(x)))^2}$$

Kleinrock

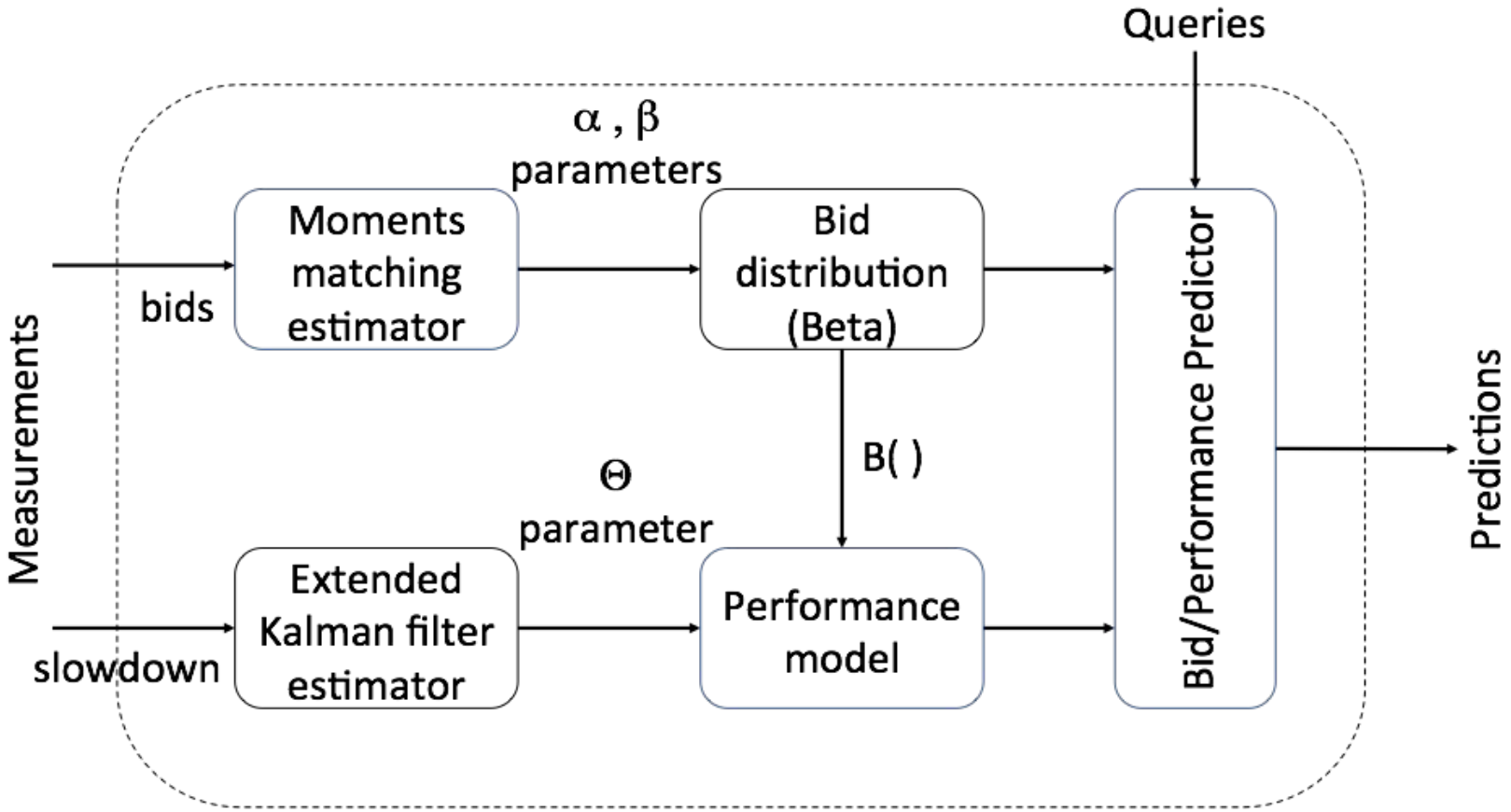
Multi-worker, multi-server bribing queue



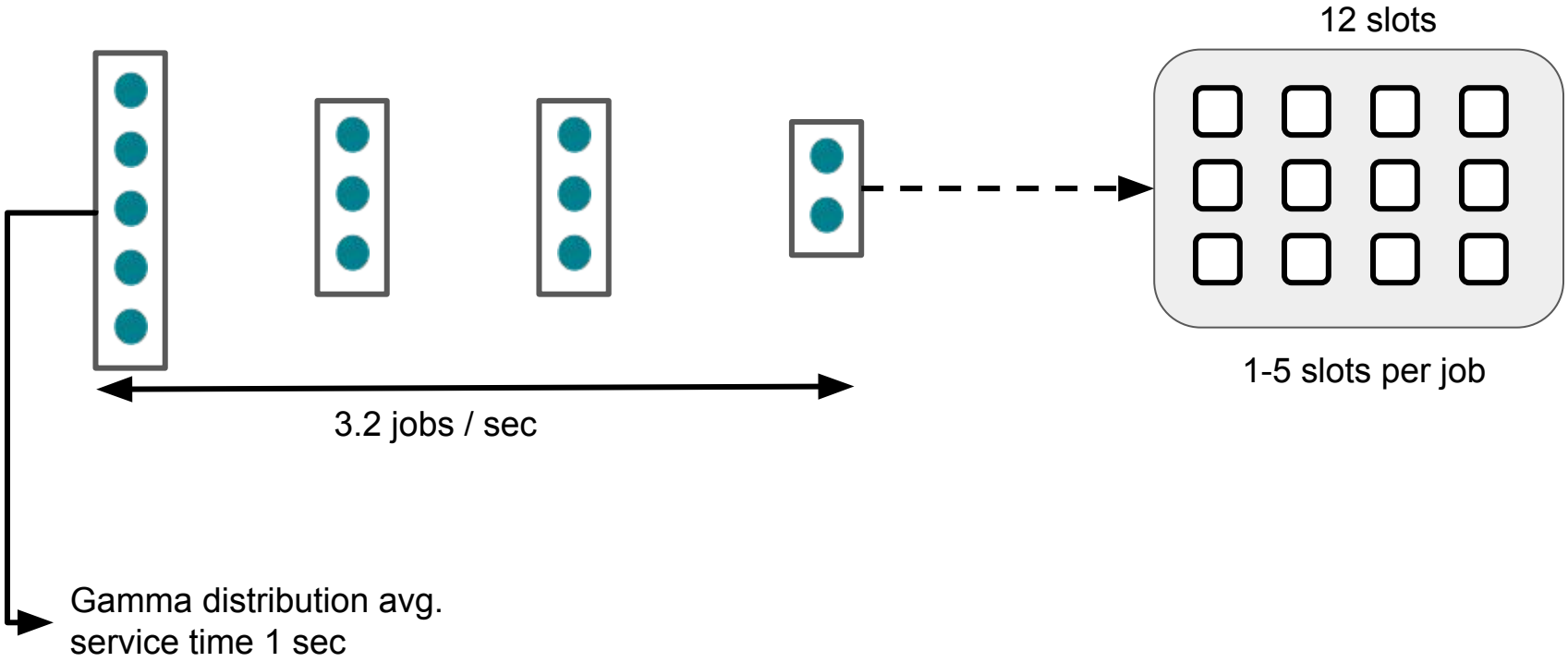
$$S(x; B, \Theta) = \frac{1}{[1 - \theta_0(1 - B(x))^{\theta_1}]^2}$$

Our extension

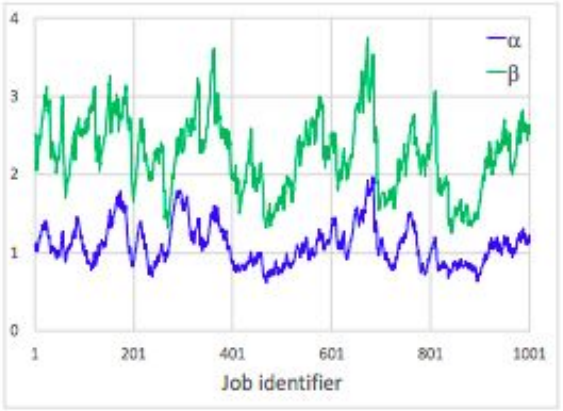
Prediction Framework



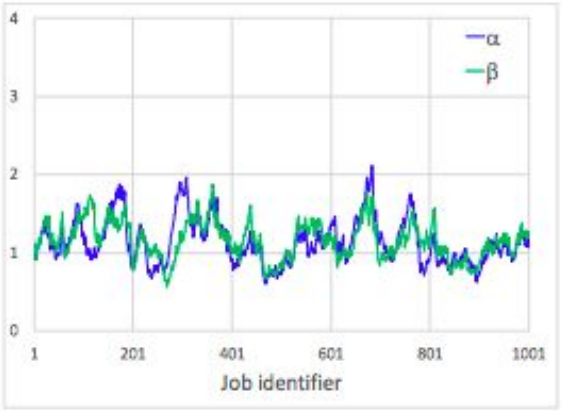
Simulation Setup



Bid Distribution Parameters



(a) Linearly decreasing



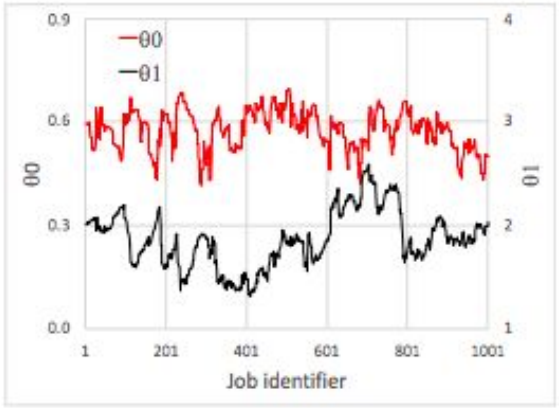
(b) Uniform

(c) Linearly increasing

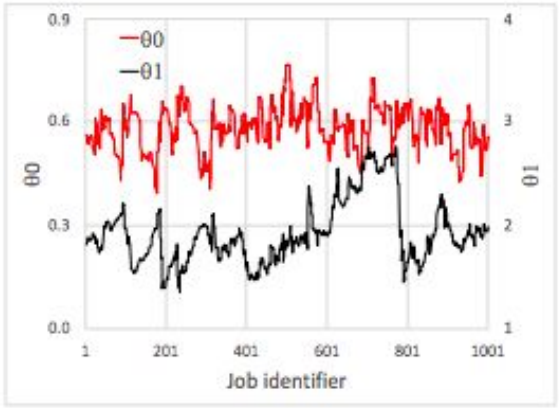
b(x) = 2(1-x) **b(x) = 1** **b(x) = 2x**

- **Parameters fluctuate around their values governed by α/β = 0.5, 1, 2**

Model Parameters



(a) Linearly decreasing

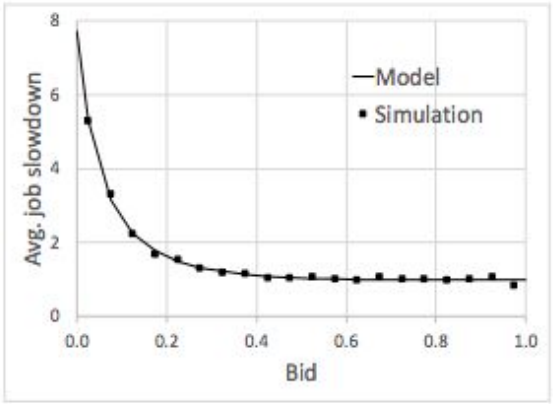


(b) Uniform

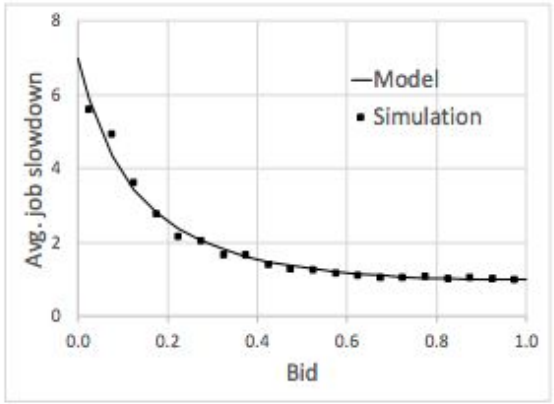
(c) Linearly increasing

θ_1 has a catalyzing effect when θ_0 is overestimated

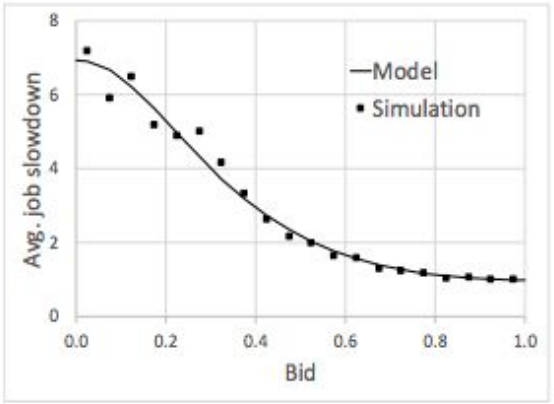
Prediction Accuracy



(a) Linearly decreasing



(b) Uniform



(c) Linearly increasing

High accuracy for the entire range of bid values, irrespective of the shape of the bid distribution.

Conclusion

➔ Simple, approximate job slowdown expression as a function of bid

➔ Methodology for dynamically estimating the model parameters

➔ Validated prediction accuracy using simulations