Empirical Evaluation of Workload Forecasting Techniques for Predictive Cloud Resource Scaling

In Kee Kim, Wei Wang, Yanjun (Jane) Qi, and Marty Humphrey

Computer Science @ University of Virginia

Motivation - Cloud Resource Scaling Approach

Reactive Auto Scaling

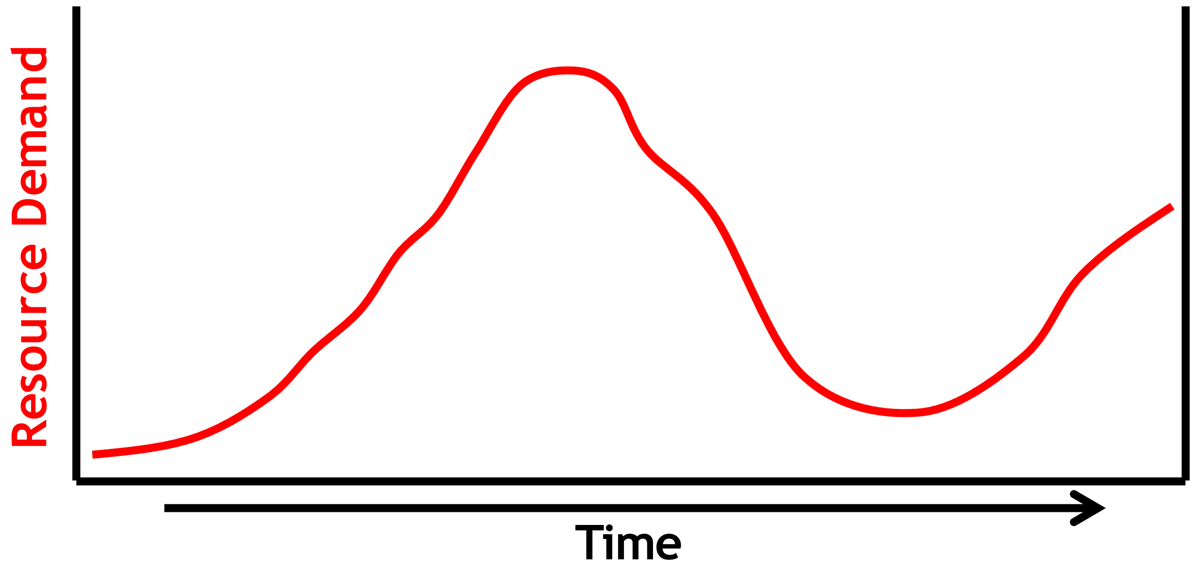
[AWS, Google, Azure, etc.]

Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O…

Motivation - Cloud Resource Scaling Approach

<u>Reactive Auto Scaling</u>
[AWS, Google, Azure, etc.]

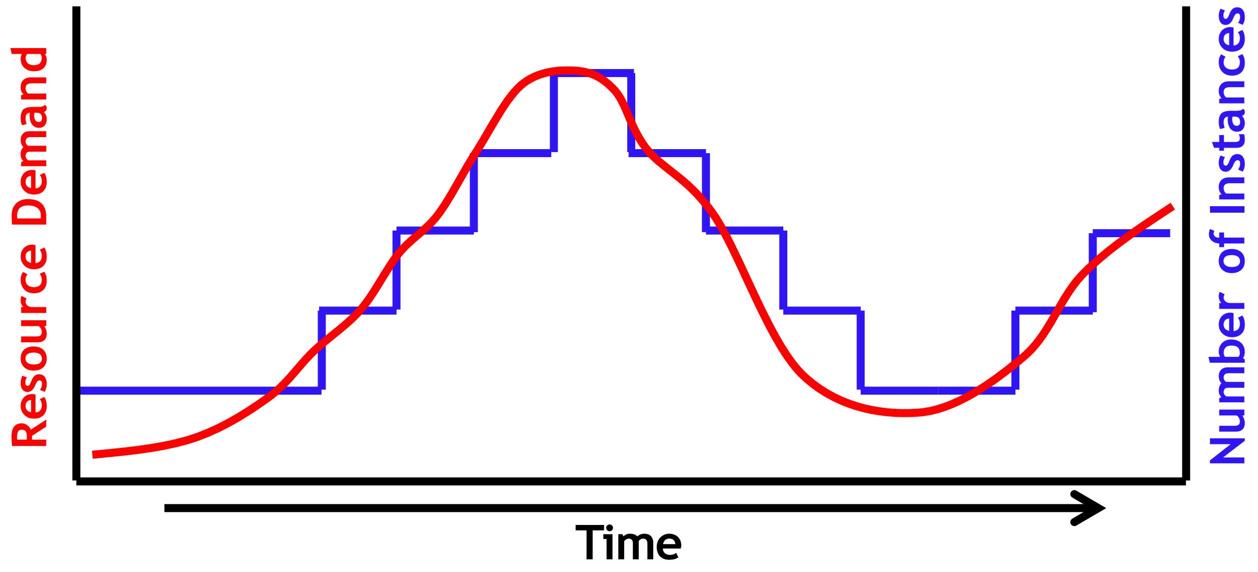
Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O...



Motivation - Cloud Resource Scaling Approach

<u>Reactive Auto Scaling</u>
[AWS, Google, Azure, etc.]

Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O…

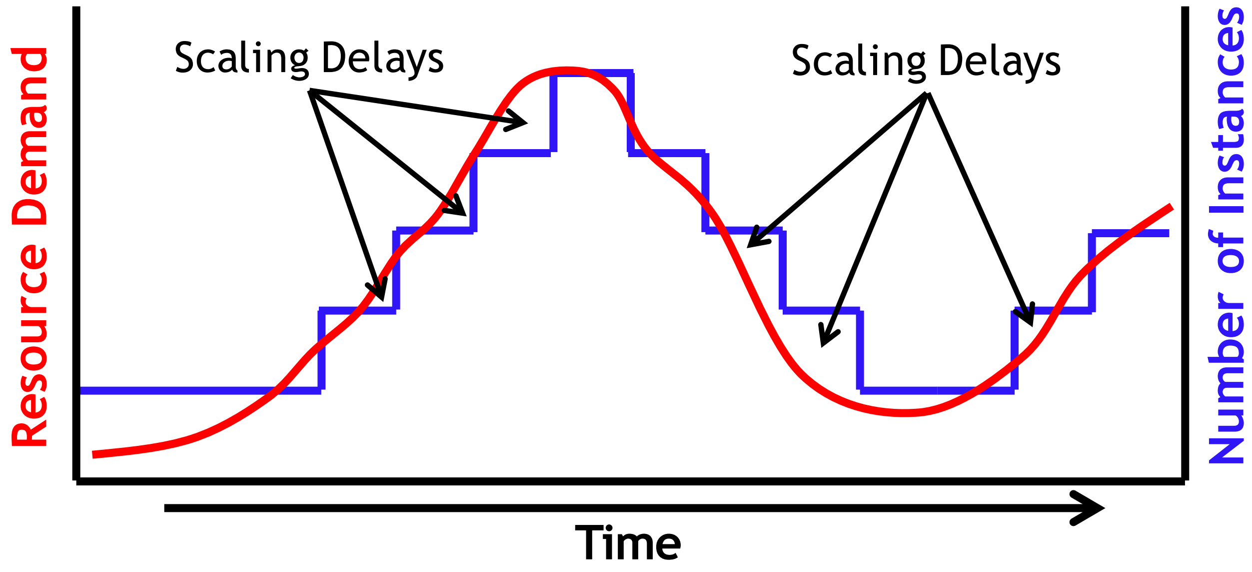


Motivation - Cloud Resource Scaling Approach

<u>Reactive Auto Scaling</u>

[AWS, Google, Azure, etc.]

Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O…



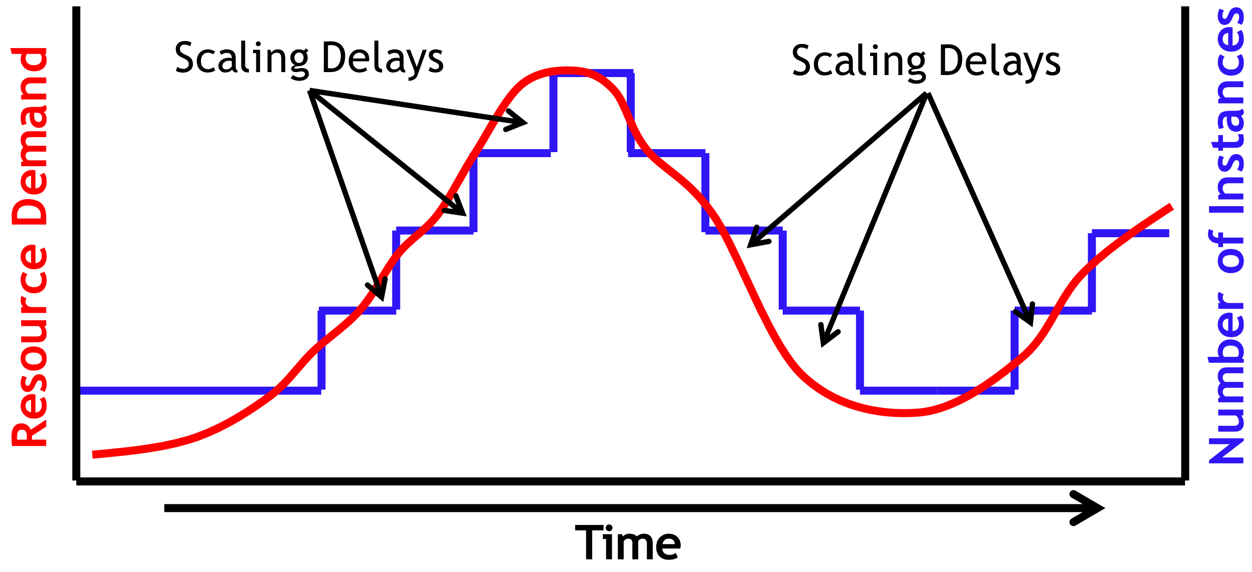
Motivation - Cloud Resource Scaling Approach

Reactive Auto Scaling
[AWS, Google, Azure, etc.]

Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O…

Predictive Resource Scaling

Resource Scaling based on forecasting:
1. Future Resource Usage
2. Workload Arrival Pattern



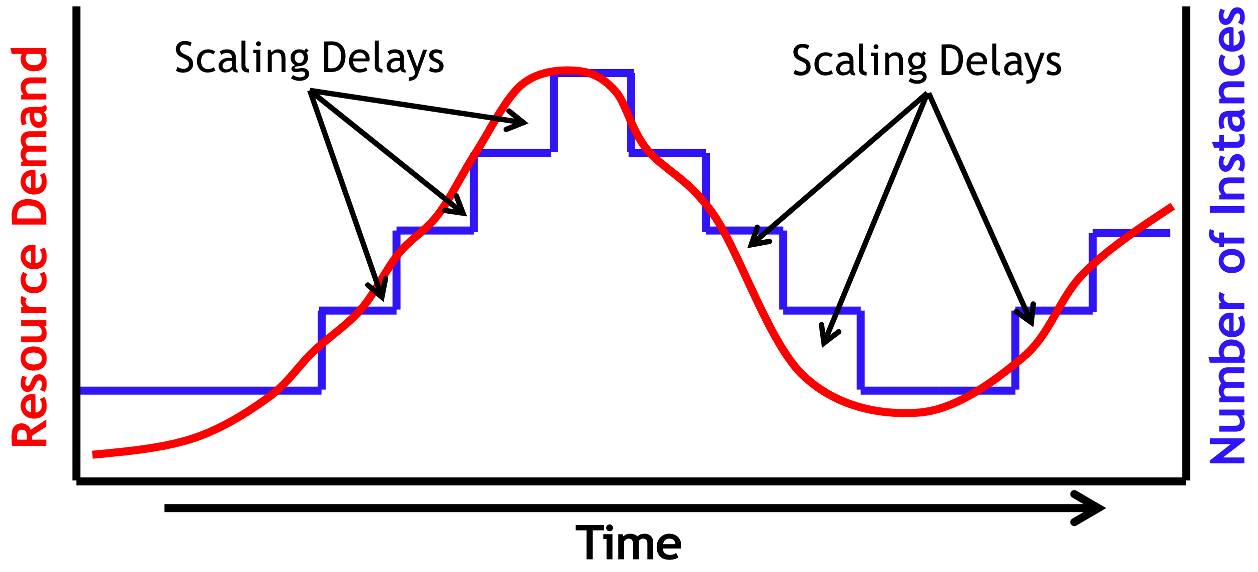
Motivation - Cloud Resource Scaling Approach

Reactive Auto Scaling
[AWS, Google, Azure, etc.]

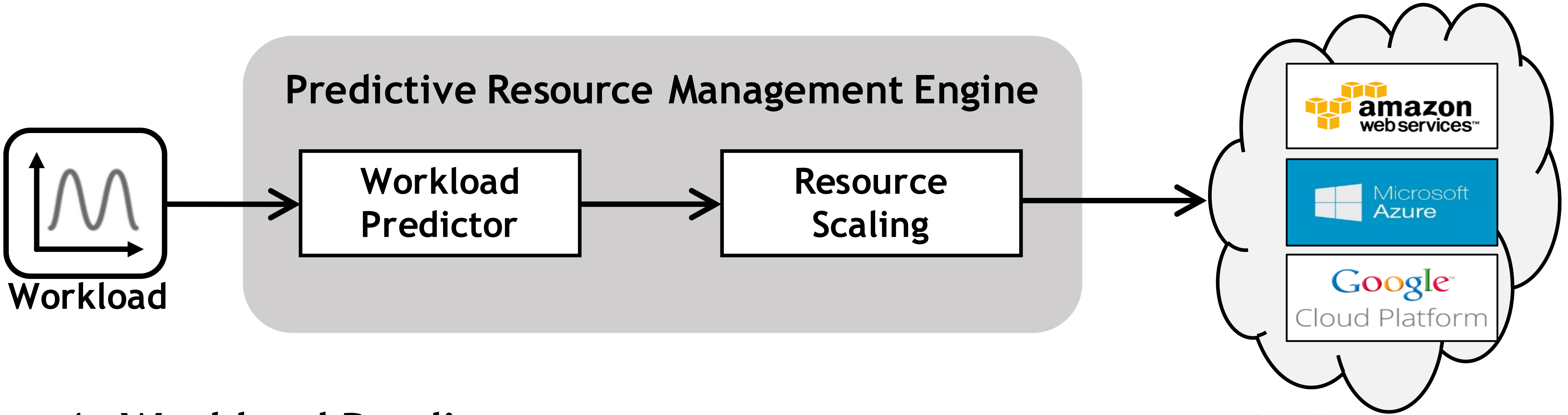
Autoscaling based on Resource Utilization:
CPU, Memory, Network-I/O…

Predictive Resource Scaling

Resource Scaling based on forecasting:
1. Future Resource Usage
2. Workload Arrival Pattern



Predictive Resource Scaling



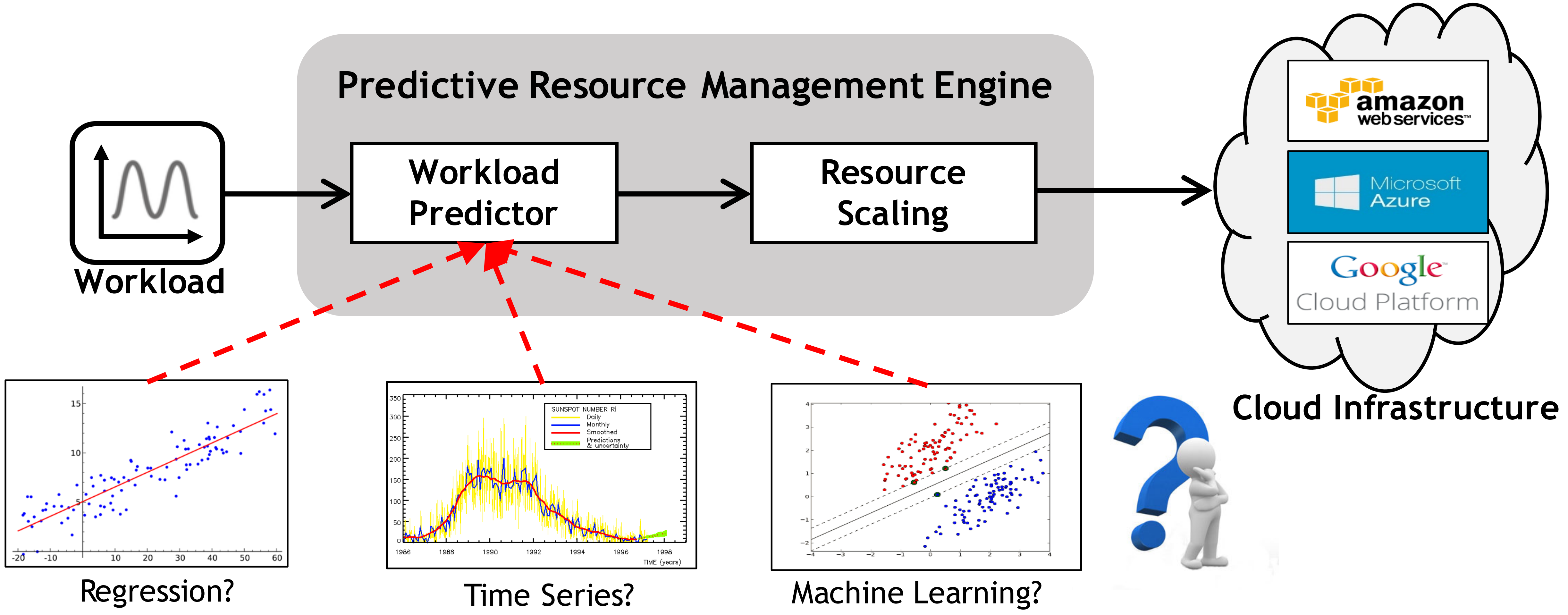
1. Workload Predictor

- Detects cloud workload pattern.
- Predicts job arrival pattern in near future.

2. Resource Scaling

- allocates/deallocates cloud resources based on the prediction.

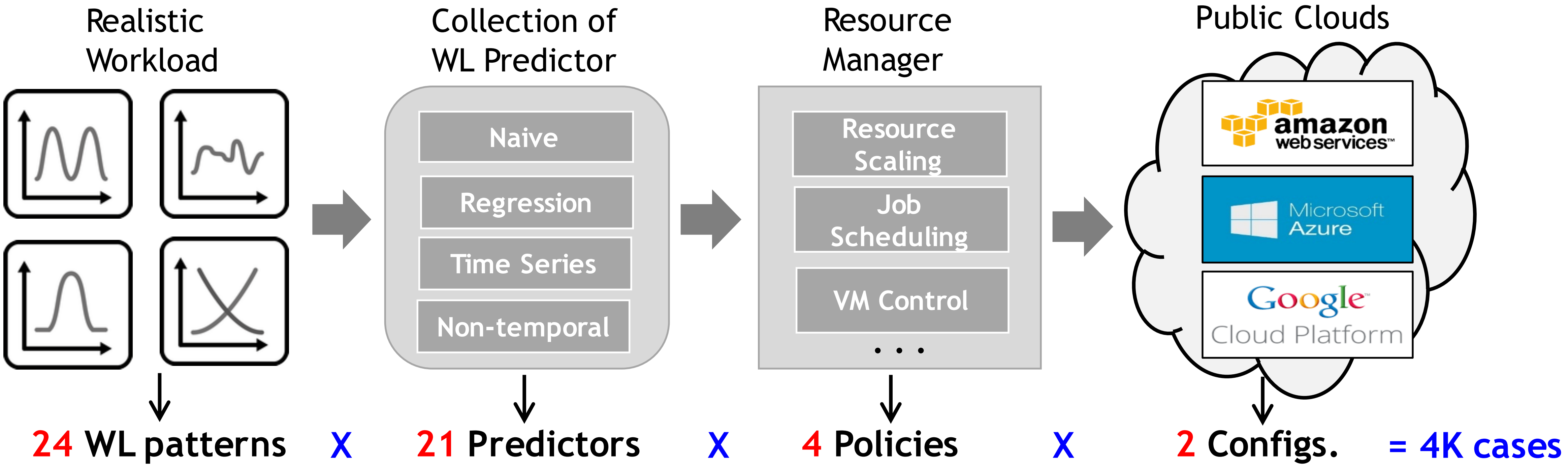
Predictive Resource Scaling



Research Questions

- **Question #1**: Which workload predictor has the highest accuracy for job arrival time prediction?
- **Question #2**: Which exiting workload predictor has the best cost efficiency and performance benefits?
- **Question #3**: Which styles of predictive scaling achieves the best cost efficiency and performance benefits?

Research Big Picture



- 4K cases are very challenging via actual deployment on IaaS clouds.
  - Use PICS (Public IaaS Cloud Simulator) – KWH – CLOUD'15

Experiment Design

- Collection of Workload Predictors.
- Simulation Workloads.
- Design of Resource Management System.
- Implementation and Performance Tuning.

Collection of (Existing) Workload Predictors

- We collect all 21 workload predictors:

1) Naïve Models

- Mean-based
- Recent-mean (*k*NN)

2) Regression Models

- Global Model (Linear, Quad, Cubic)
- Local Model (Linear, Quad, Cubic)

3) Time Series Models

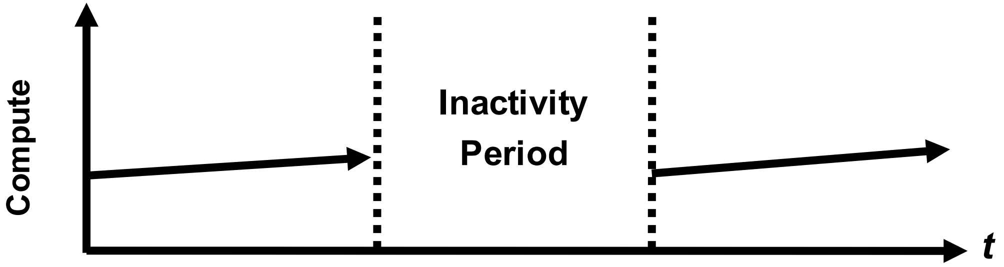
- Smoothing (WMA, EMA, DES)
- Box-Jenkins (AR, ARMA, ARIMA)

4) Non-Temporal (ML) Models

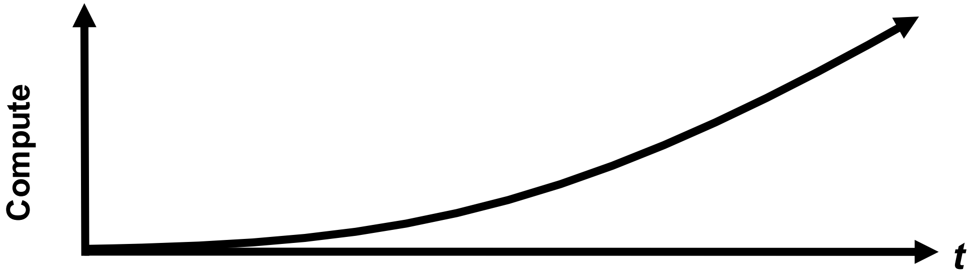
- SVMs (Linear, Gaussian)
- Decision Tree
- Ensemble (RF, GBM, Exts)

Simulation Workload Patterns

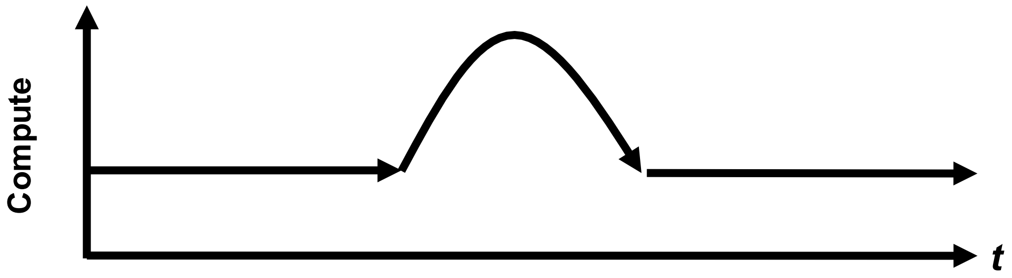
- We generate 24 workload patterns based on:



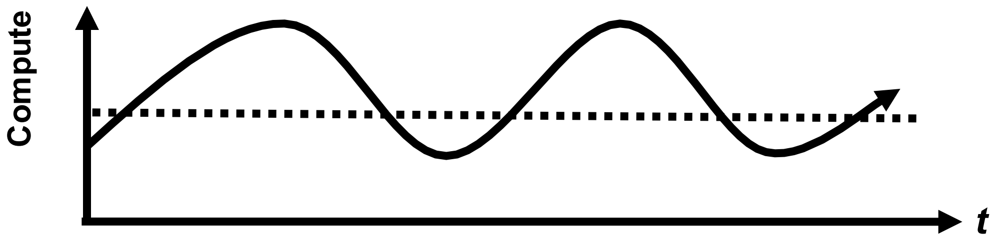
On and Off (Batch/Scientific)



Growing (Emerging Service)

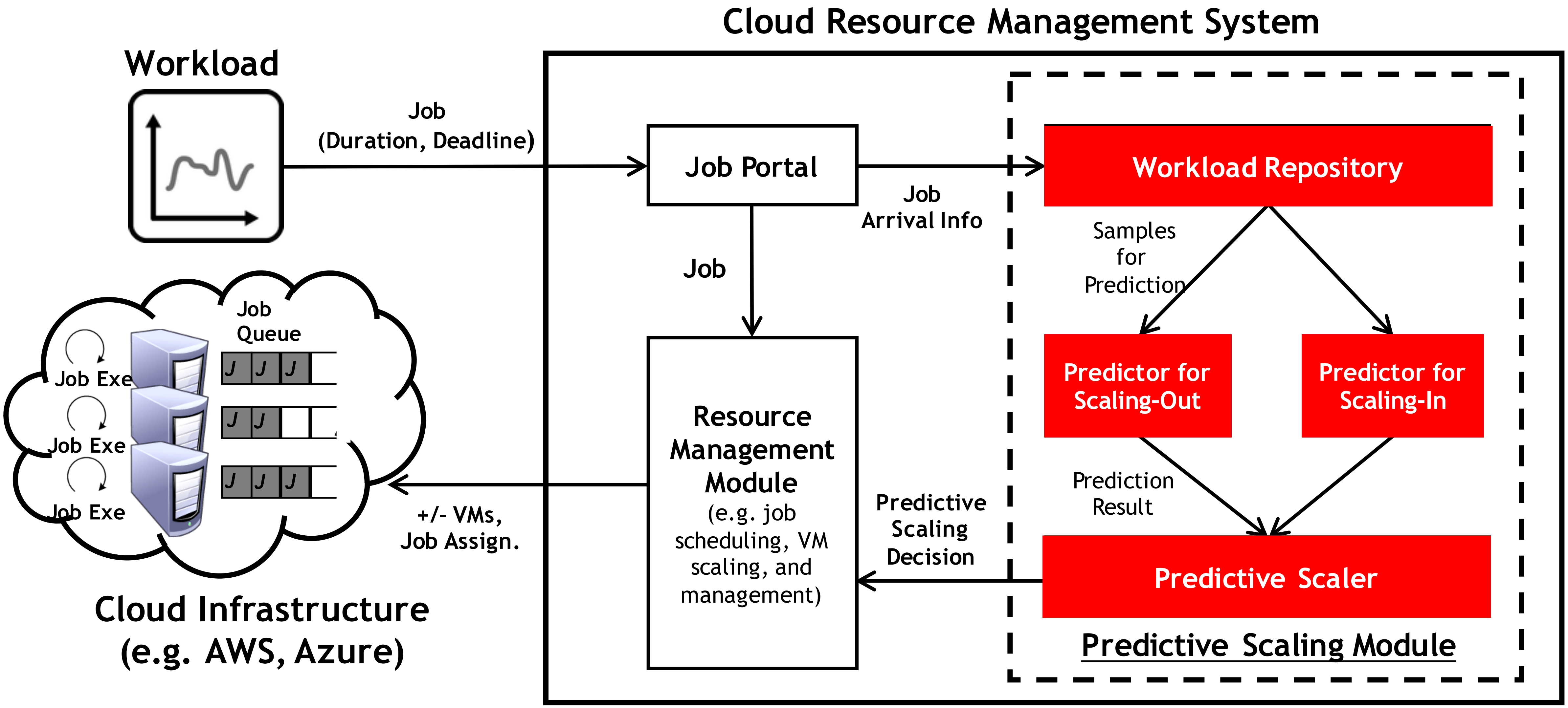


Random/Unpredictable (Media)



Cyclic Bursting (E-Commerce)

Design of Resource Management System



Implementations and Performance Tuning

- Workload Predictor Implementation.
  - All predictors are written in `Python`.
  - `numpy` and `Pandas`.
  - `statsmodels` for time-series model implementation.
  - `scikit-learn` machine learning lib for non temporal models.

- Predictor Performance Tuning.
  - (Training) Sample Size Decision:
    - a tradeoff between prediction performance and overhead.
    - Most predictors use 50 -- 100 of most recent job arrival samples.
  - Parameter Selection:
    - a grid search algorithm with prediction accuracy.

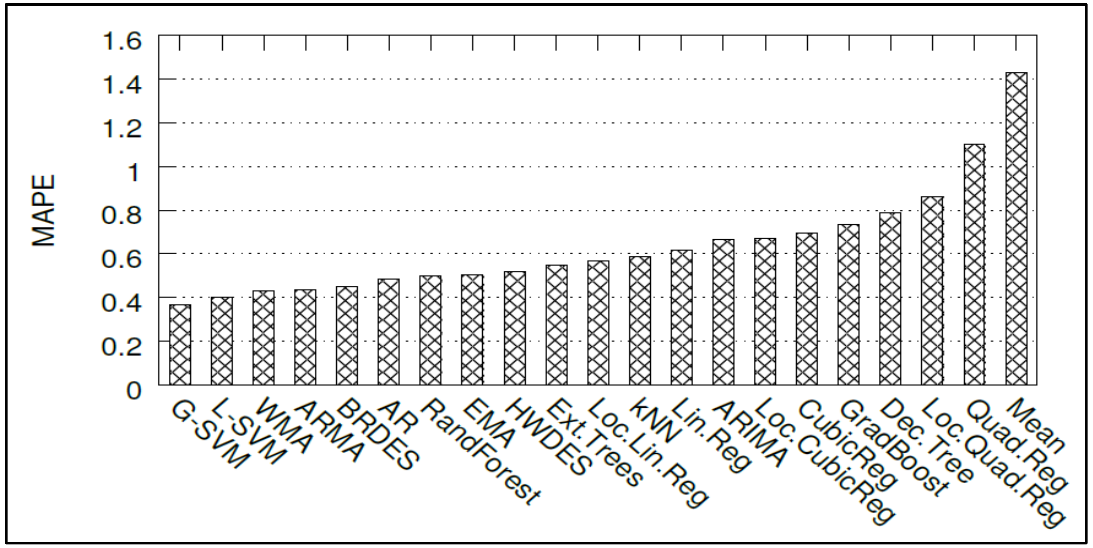
Performance Evaluation

- Experiment #1 – Statistical Predictor Performance.
- Experiment #2 – Predictive Scaling Performance.

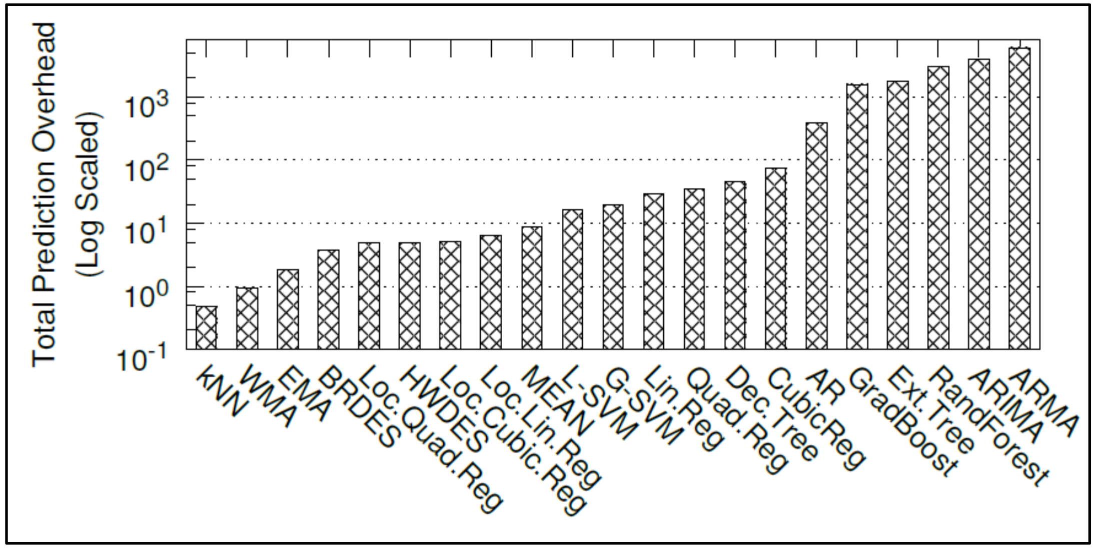
Experiment #1 – (Statistical) Predictor Performance

- Purpose: Measuring Statistical Predictor Accuracy and Overhead.
  - Accuracy: MAPE – Mean Absolute Percentage Error.
  - Overhead: Sum of All Prediction Time.

- Overall Results:



Overall Prediction Accuracy

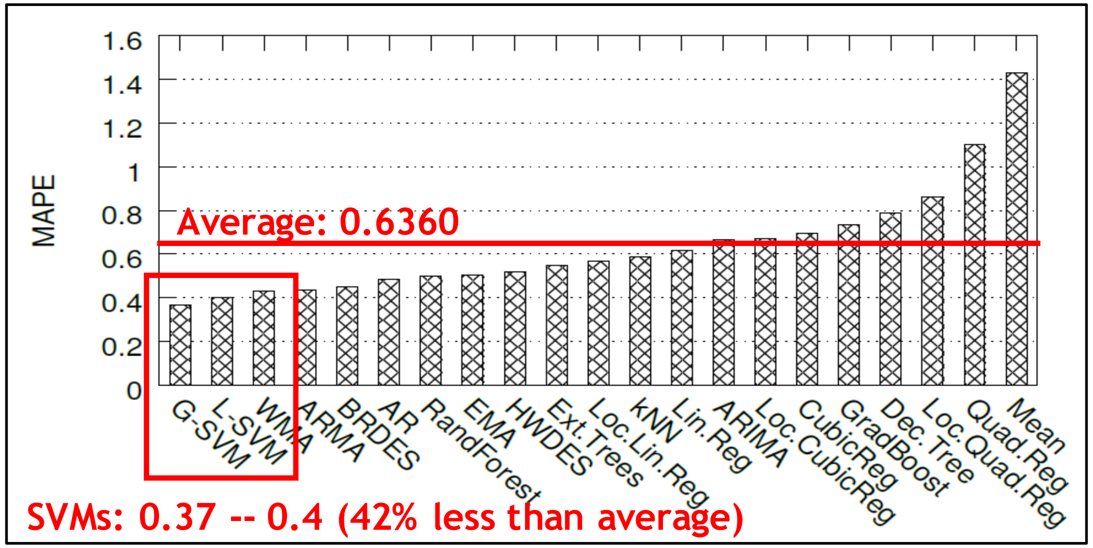


Overall Prediction Overhead (10K Jobs)

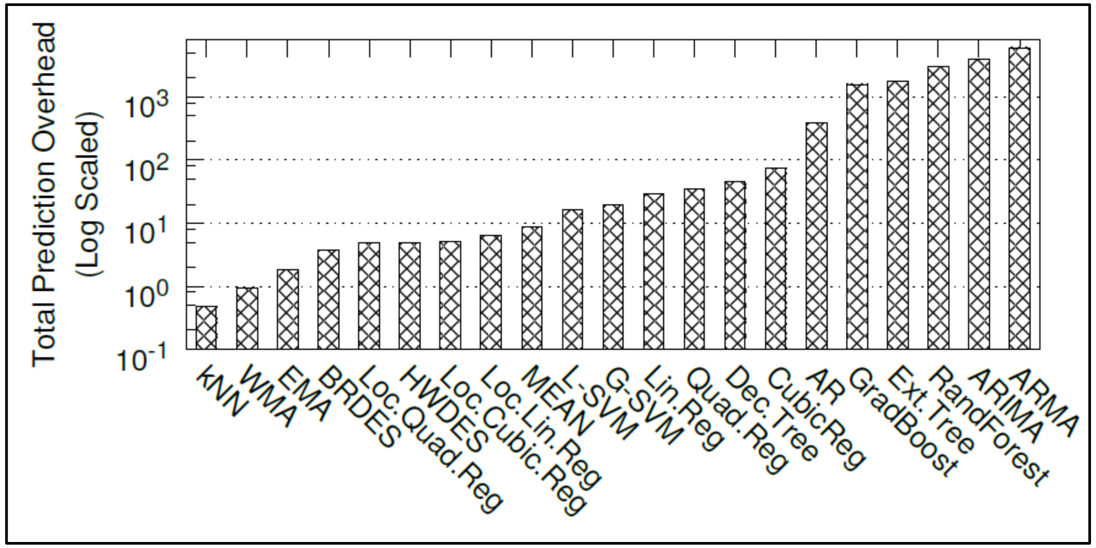
Experiment #1 - (Statistical) Predictor Performance

- Purpose: Measuring Statistical Predictor Accuracy and Overhead.
  - Accuracy: MAPE – Mean Absolute Percentage Error.
  - Overhead: Sum of All Prediction Time.

- Overall Results:



Overall Prediction Accuracy

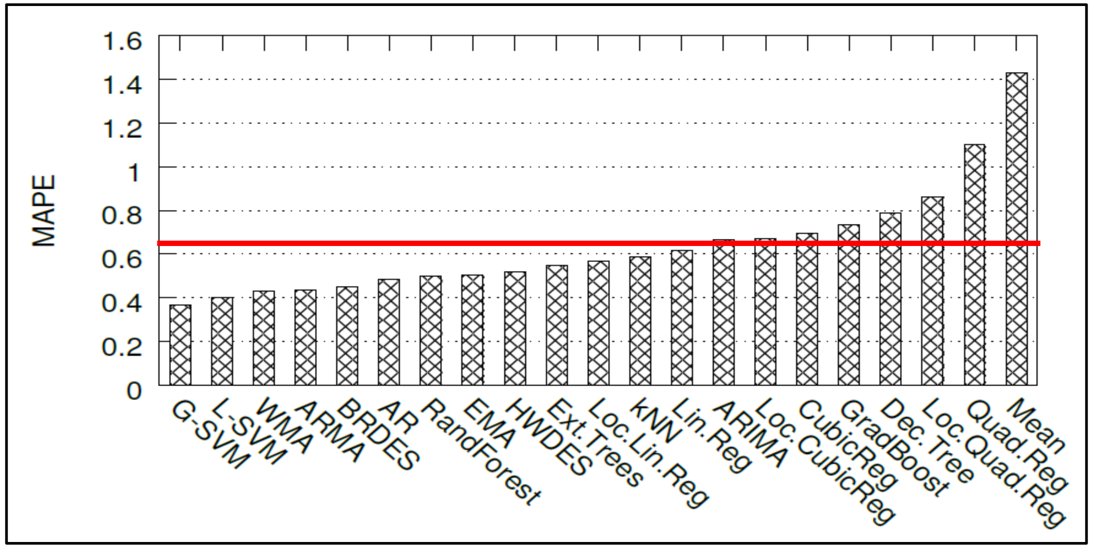


Overall Prediction Overhead (10K Jobs)

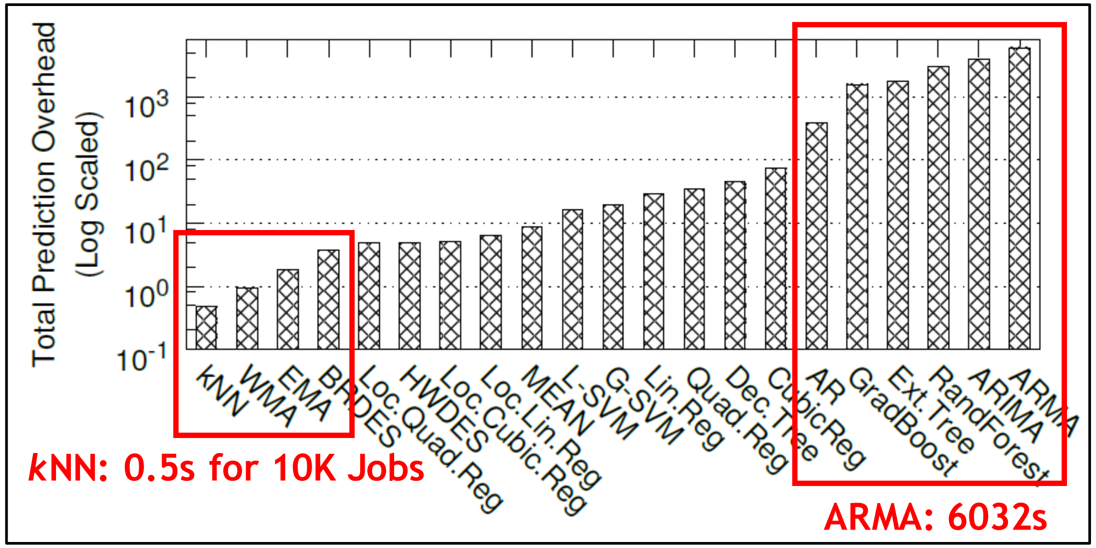
Experiment #1 - (Statistical) Predictor Performance

- Purpose: Measuring **Statistical** **Predictor Accuracy** and **Overhead**.
  - Accuracy: MAPE – Mean Absolute Percentage Error.
  - Overhead: Sum of All Prediction Time.

- Overall Results:



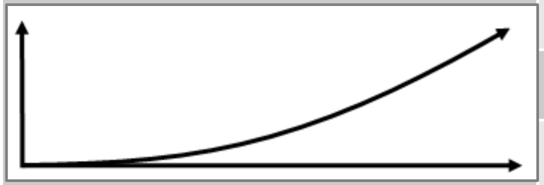
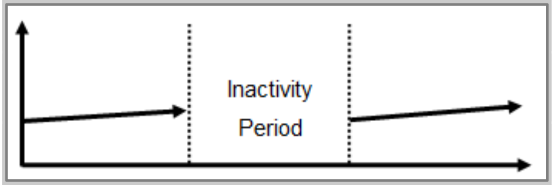
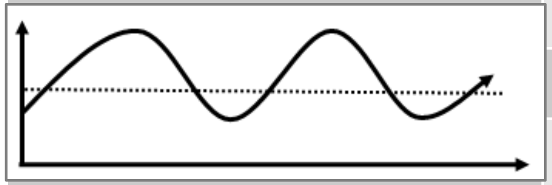
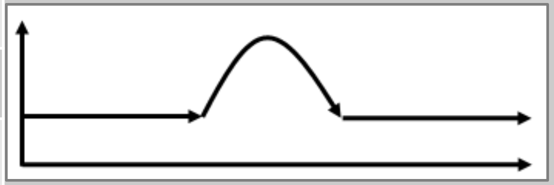
Overall Prediction Accuracy



Overall Prediction Overhead (10K Jobs)

Experiment #1 - (Statistical) Predictor Performance

- Accuracy of Workload Predictor per Pattern.

Workload	Rank	Predictor	MAPE	Workload	Rank	Predictor	MAPE
Growing	1	Lin. SVM	0.28	On/Off	1	Gau. SVM	0.22
	2	AR	0.29		2	ARMA	0.30
	3	ARMA	0.30		3	Lin. SVM	0.44
	Avg.	--	0.51		Avg.	--	0.69
Bursty	1	ARIMA	0.38	Random	1	Gau. SVM	0.45
	2	Brown's DES	0.41		2	Lin. Reg.	0.46
	3	Lin. SVM	0.43		3	Lin. SVM	0.46
	Avg.	--	0.75		Avg.	--	0.52

Experiment #2 - Predictive Scaling Performance

- Purpose: **How much benefits** RM can achieve by applying
  1. "Good Predictor"
  2. Different Styles of Predictive Scaling.

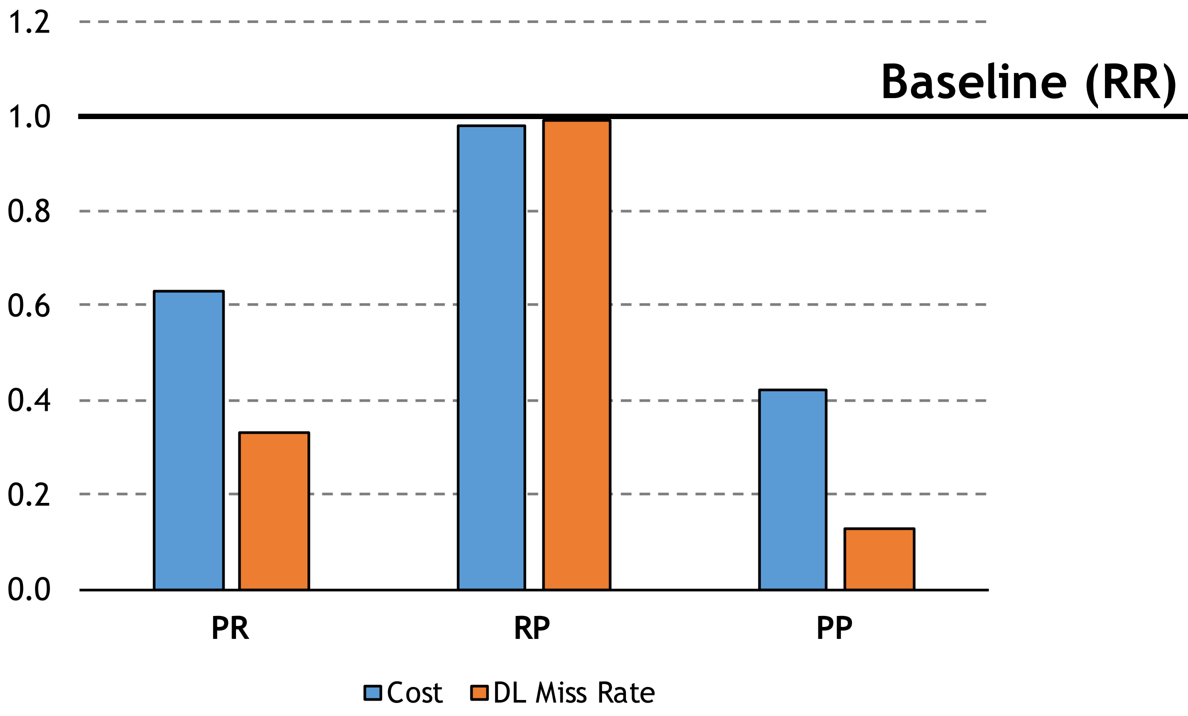
- List of Predictors: **8 best predictors** from evaluation #1.
  - Linear Regression, WMA, BRDES, AR, ARMA, ARIMA, Linear SVM, Gaussian SVM

- Four Different Styles of Resource Scaling.
  - **RR** (Reactive Scaling-Out + Reactive Scaling-In) -- **Baseline**
  - **PR** (Predictive Scaling-Out + Reactive Scaling-In)
  - **RP** (Reactive Scaling-Out + Predictive Scaling-In)
  - **PP** (Predictive Scaling-Out + Predictive Scaling-In)

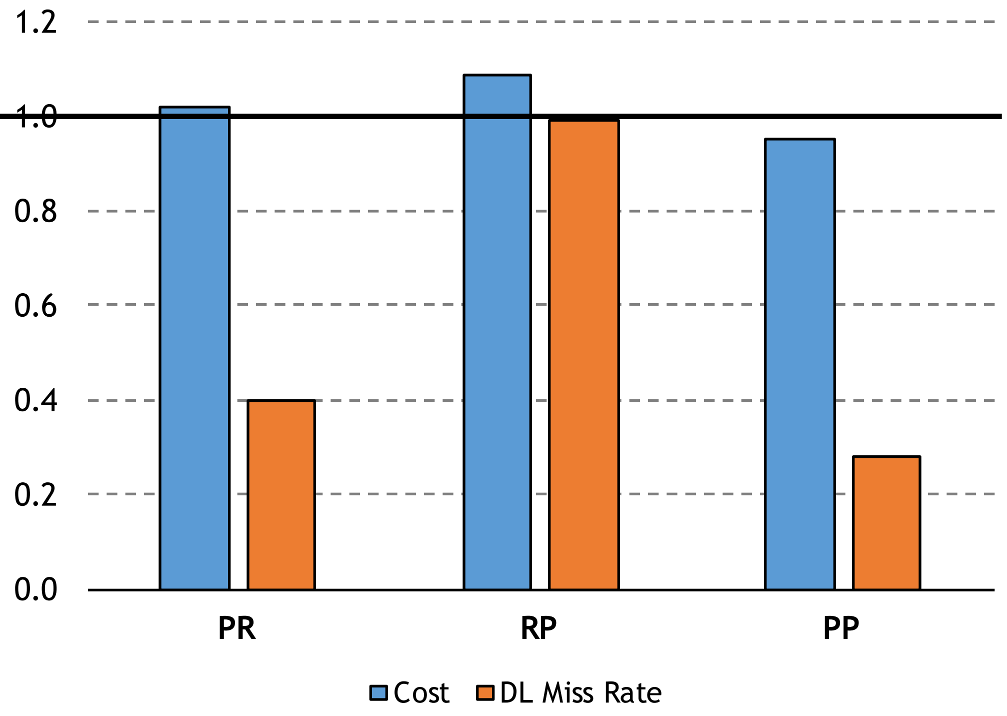
- Cloud Configurations: Two Pricing Models -- Hourly and Minutely.
- Metrics: Cloud Cost and Job Deadline Miss Rate.

Experiment #2 - Predictive Scaling Performance

- Overall Results:



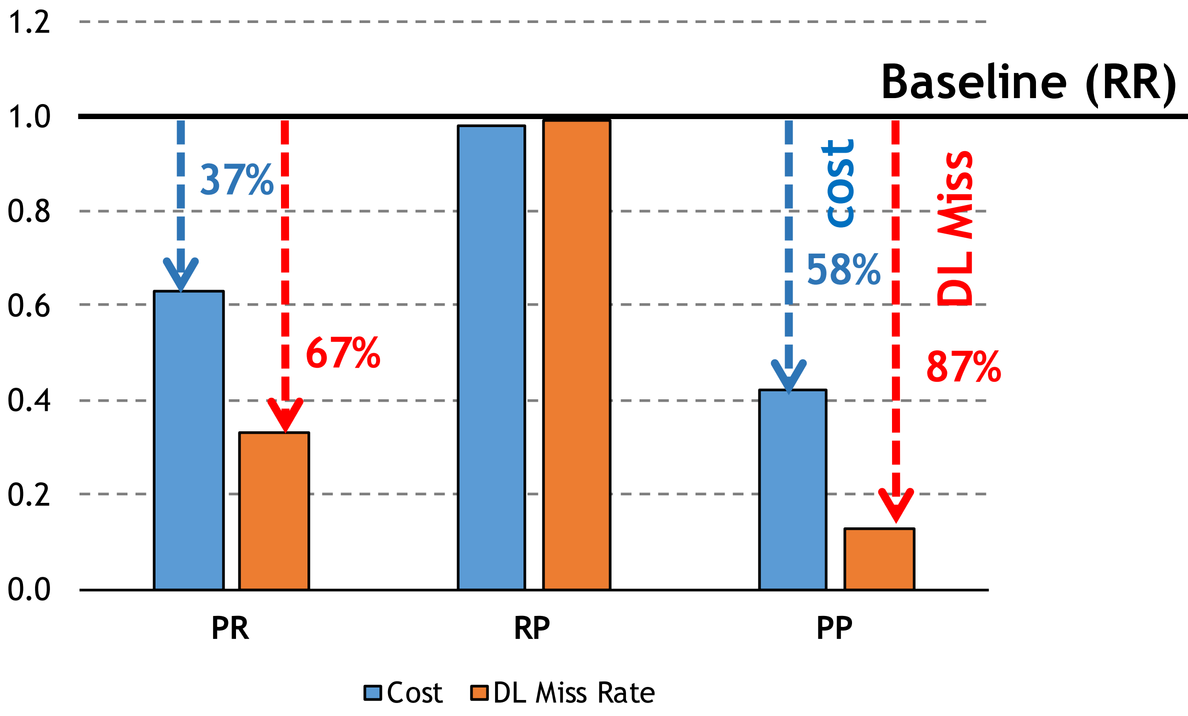
(a) **Hourly Pricing Model**



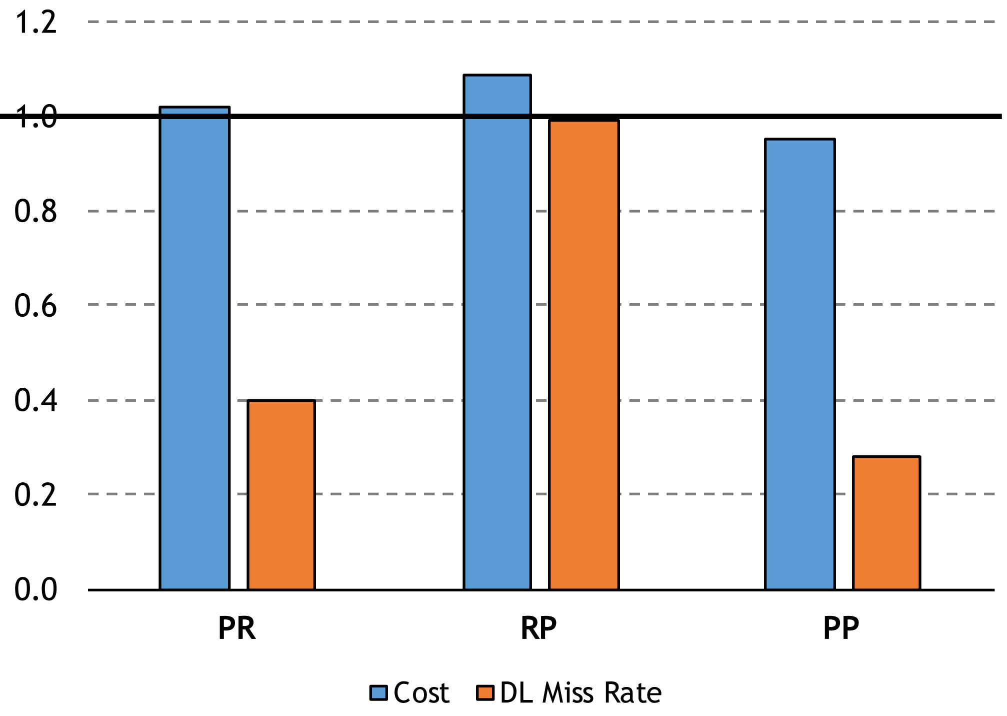
(b) **Minutely Pricing Model**

Experiment #2 - Predictive Scaling Performance

- Overall Results:



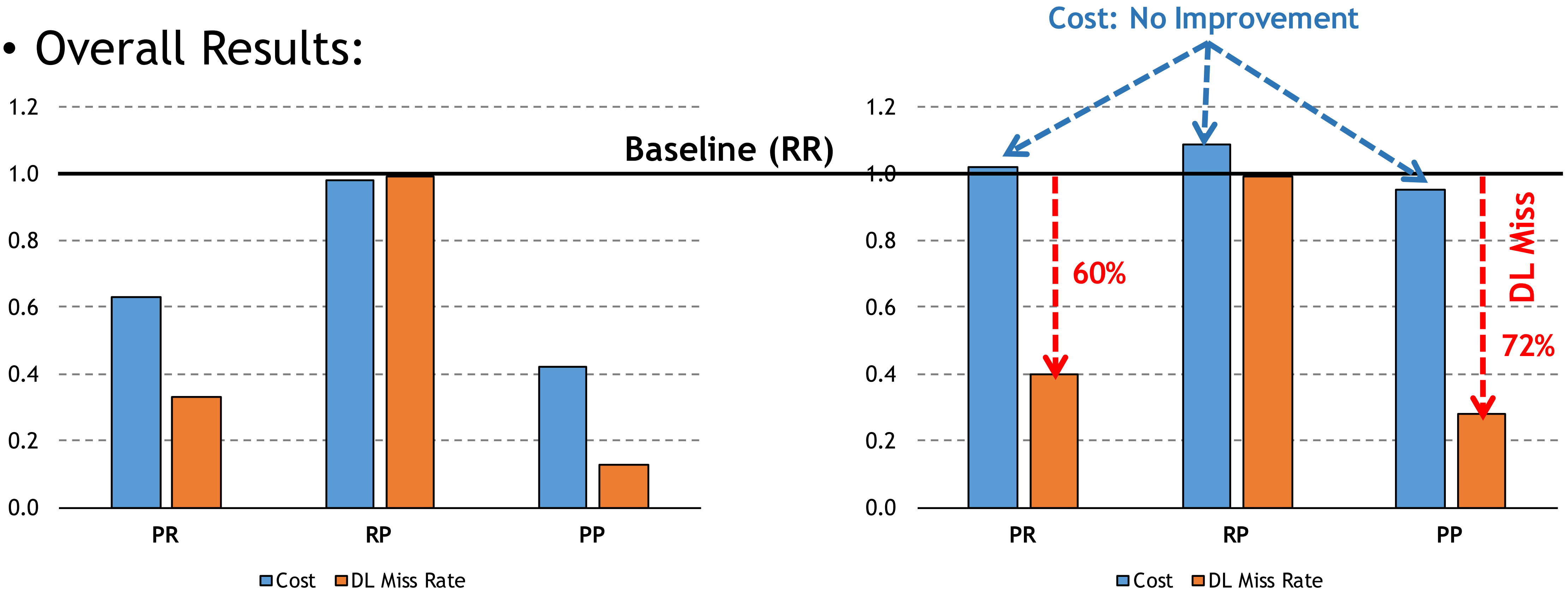
(a) **Hourly Pricing Model**



(b) **Minutely Pricing Model**

Experiment #2 - Predictive Scaling Performance

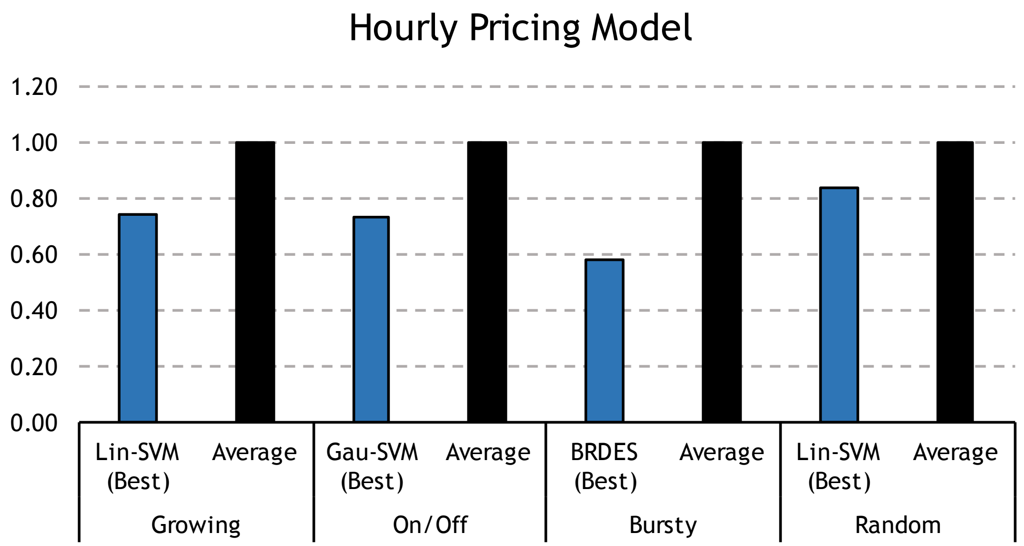
- Overall Results:

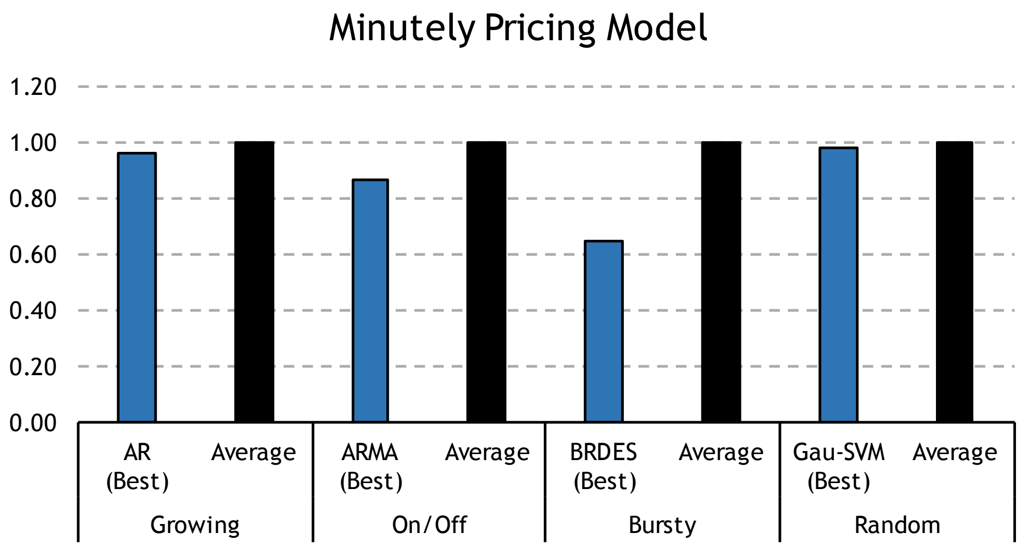


(a) **Hourly Pricing Model** (b) **Minutely Pricing Model**

Experiment #2 - Predictive Scaling Performance

- **PP** (**P**redictive Scaling-Out – **P**redictive Scaling-In) Details -- Deadline Miss Rate

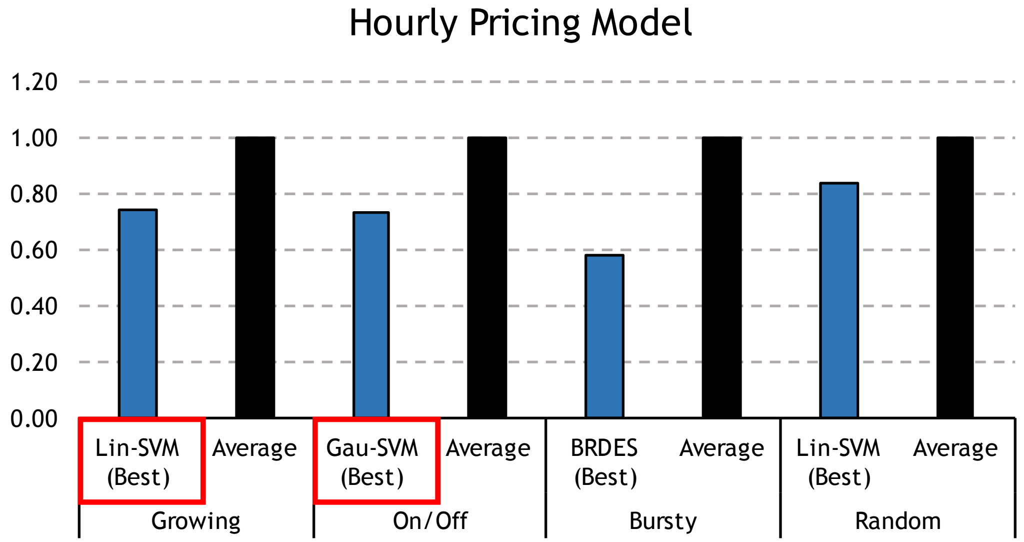


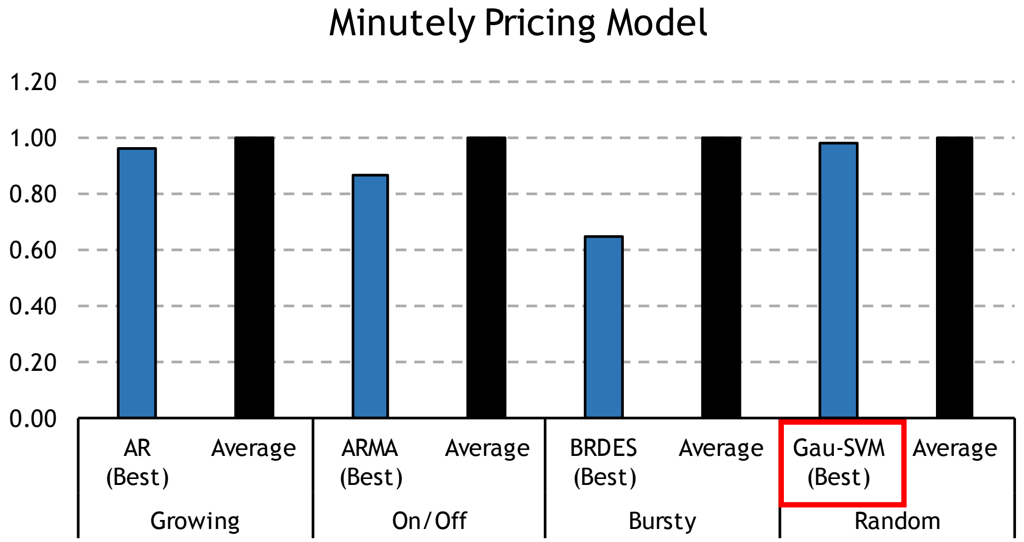


Workloads	Top 1	Top 2	Top 3
Growing	Linear SVM	AR	ARMA
On/Off	Gaussian SVM	ARMA	Linear SVM
Bursty	ARIMA	Brown’s DES	Linear SVM
Random	Gaussian SVM	Linear Regression	Linear SVM

Experiment #2 - Predictive Scaling Performance

- **PP** (**P**redictive Scaling-Out – **P**redictive Scaling-In) Details -- Deadline Miss Rate

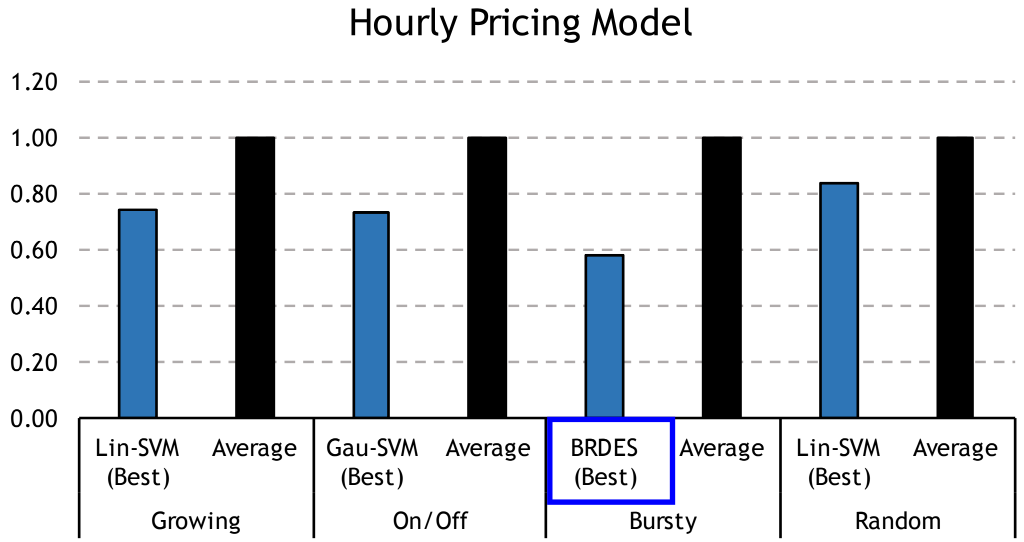


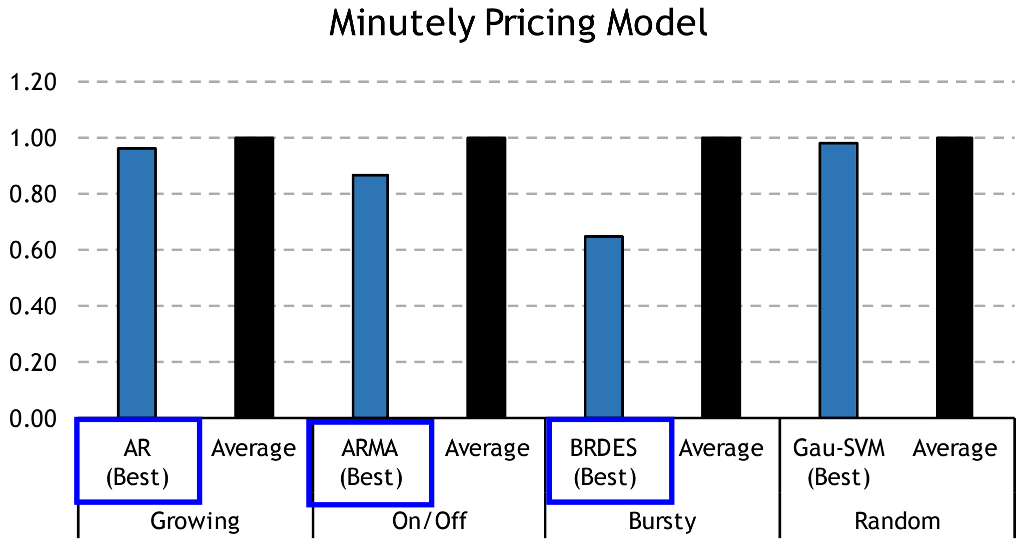


Workloads	Top 1	Top 2	Top 3
Growing	Linear SVM	AR	ARMA
On/Off	Gaussian SVM	ARMA	Linear SVM
Bursty	ARIMA	Brown’s DES	Linear SVM
Random	Gaussian SVM	Linear Regression	Linear SVM

Experiment #2 - Predictive Scaling Performance

- **PP** (**P**redictive Scaling-Out – **P**redictive Scaling-In) Details -- Deadline Miss Rate

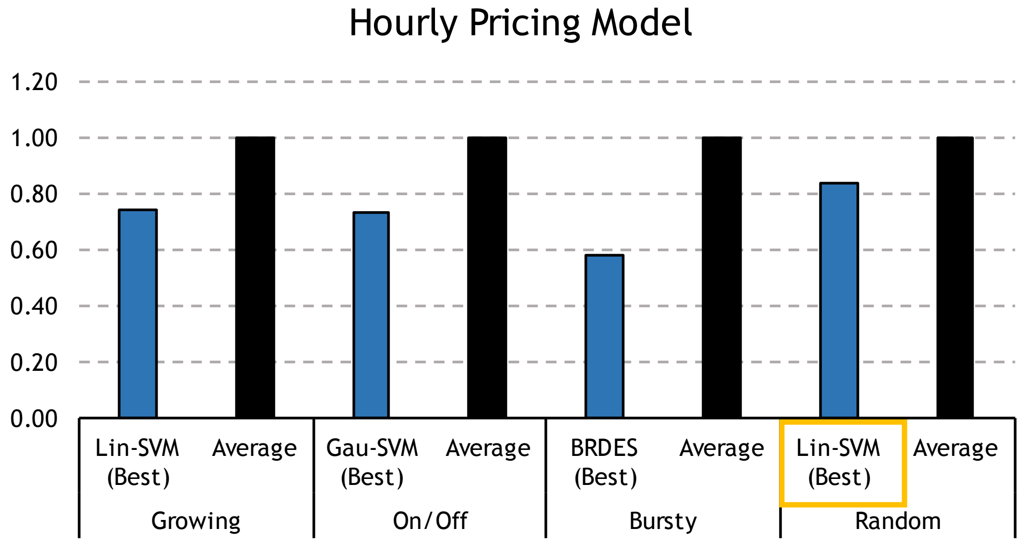


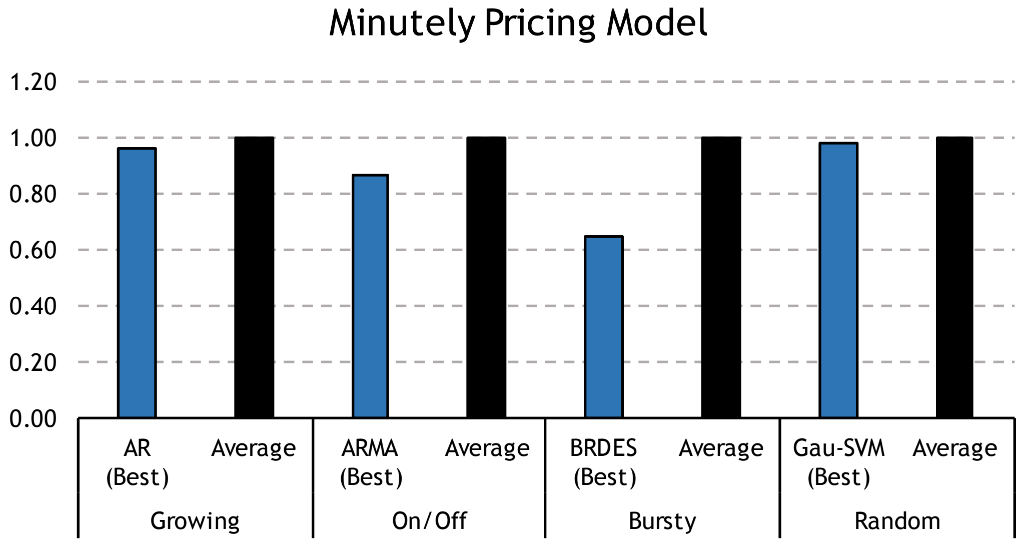


Workloads	Top 1	Top 2	Top 3
Growing	Linear SVM	AR	ARMA
On/Off	Gaussian SVM	ARMA	Linear SVM
Bursty	ARIMA	Brown's DES	Linear SVM
Random	Gaussian SVM	Linear Regression	Linear SVM

Experiment #2 - Predictive Scaling Performance

- **PP** (**P**redictive Scaling-Out – **P**redictive Scaling-In) Details -- Deadline Miss Rate





Workloads	Top 1	Top 2	Top 3
Growing	Linear SVM	AR	ARMA
On/Off	Gaussian SVM	ARMA	Linear SVM
Bursty	ARIMA	Brown's DES	Linear SVM
Random	Gaussian SVM	Linear Regression	Linear SVM

Summary -- Revisit 3 Research Questions

- **Q1**: Which WL predictor has the highest accuracy?
  - **No one predictor fits all workload patterns.**
- **Q2**: Which WL predictor provides the best performance benefits?
  - **Similar with Q1 – no universally best workload predictor exits.**
    - **Depends on workload patterns and cloud configurations (e.g. billing model)**
  - **In general, best workload predictor (cloud metric) is one of top three most (statistically) accurate predictors.**
- **Q3**: Which styles of predictive scaling provides the best performance benefits?
  - **PP (Predictive Scaling-Out/In) is the best style of predictive scaling.**
    - **"Predictive Scaling-Out" can improve cloud metrics.**

Questions?

Thank you!

ik2sb@virgina.edu