VADARA

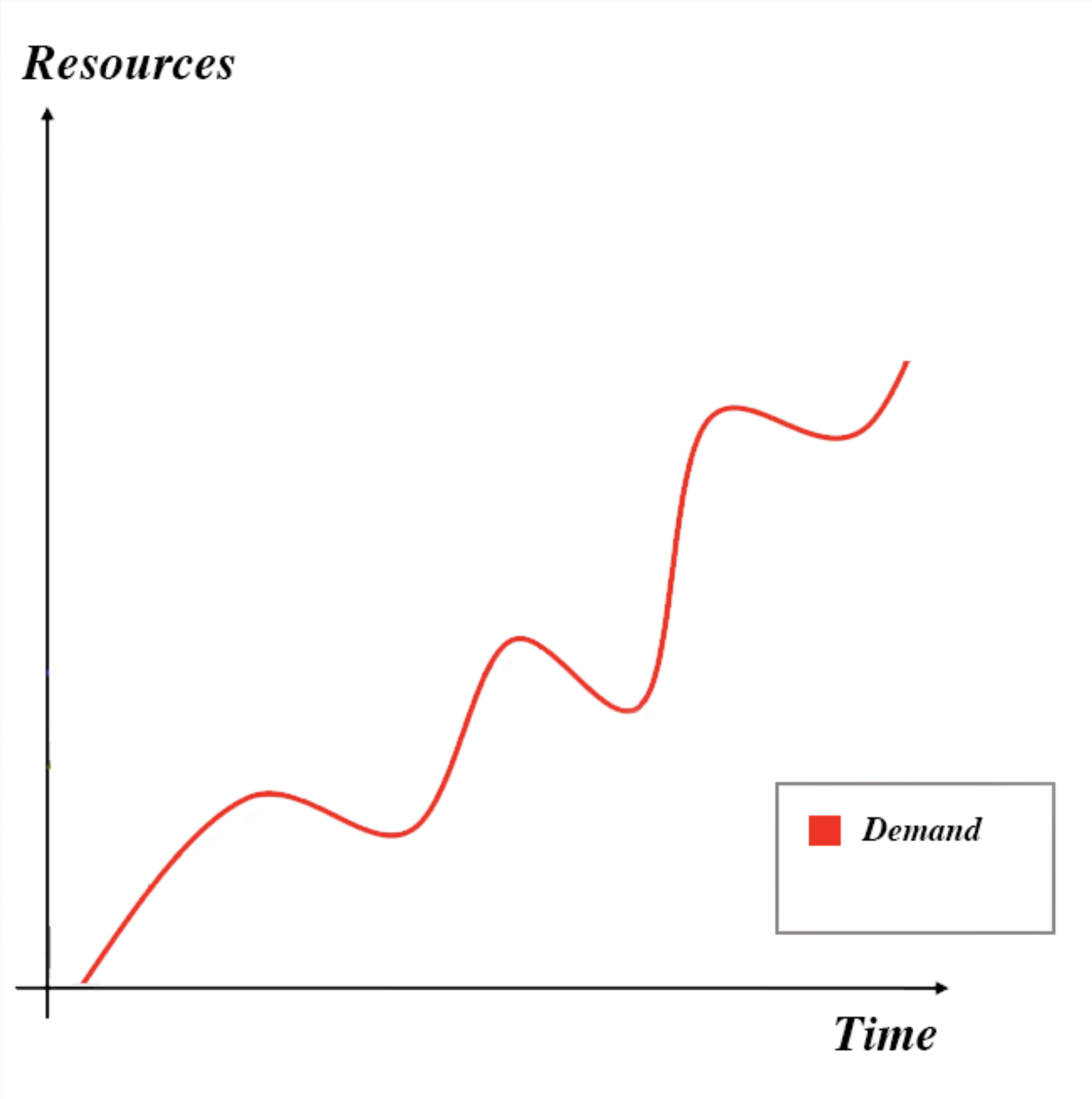
PREDICTIVE ELASTICITY FOR CLOUD APPLICATIONS

JOÃO LOFF & JOÃO GARCIA

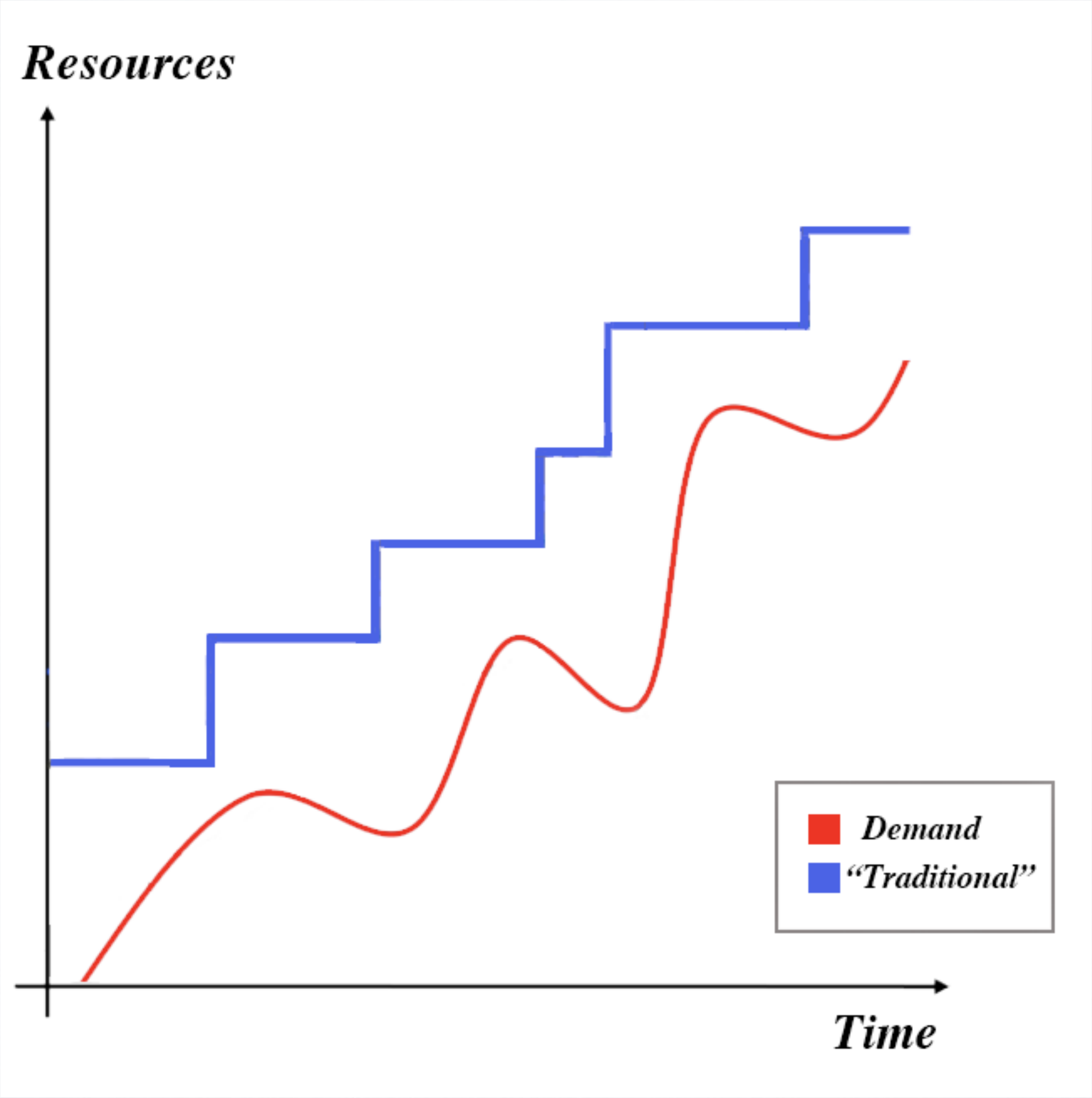
INSTITUTO SUPERIOR TÉCNICO - UNIVERSIDADE DE LISBOA
INESC-ID LISBOA

IEEE CLOUDCOM 2014

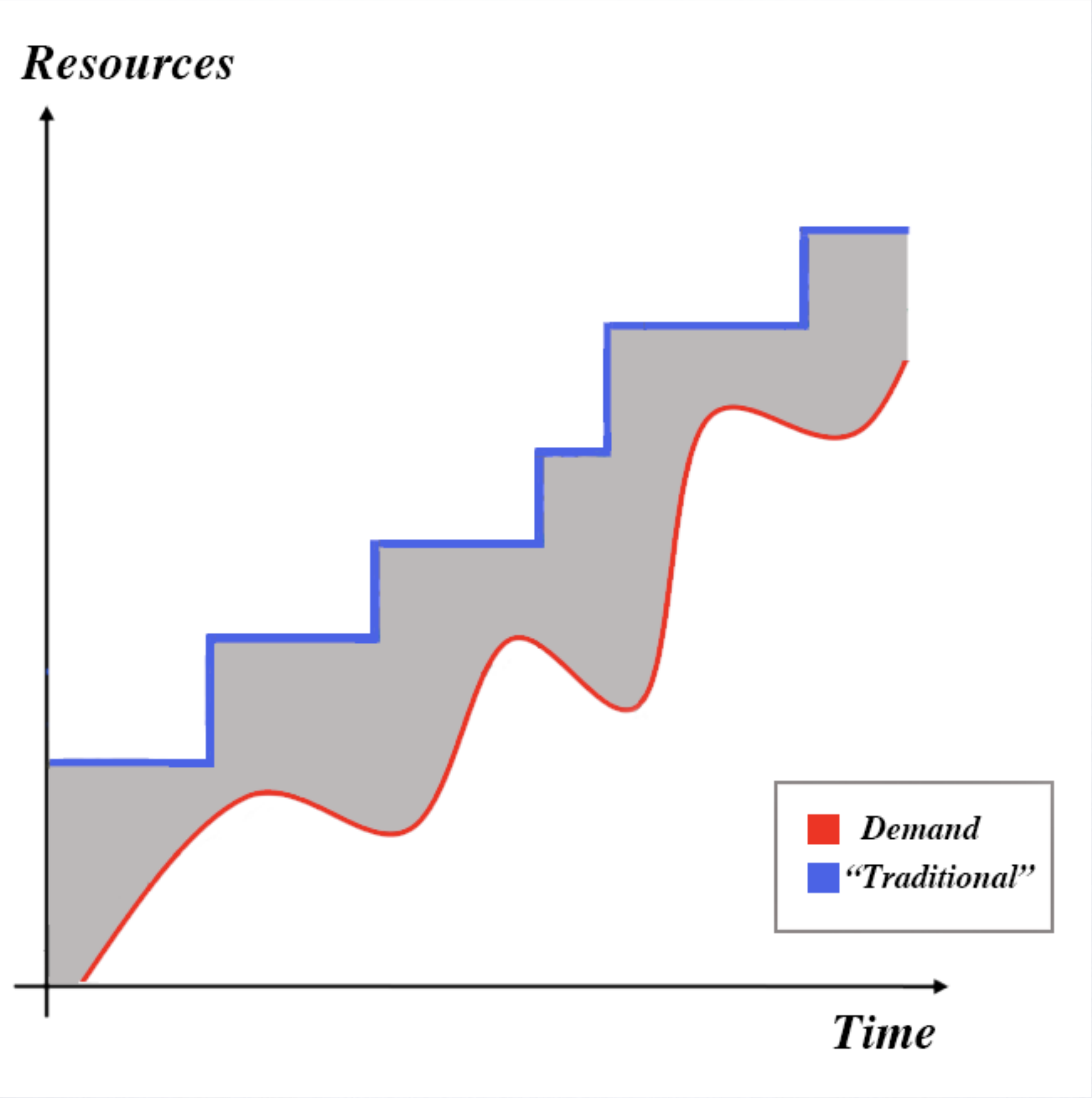
TYPICAL IAAS CLOUD



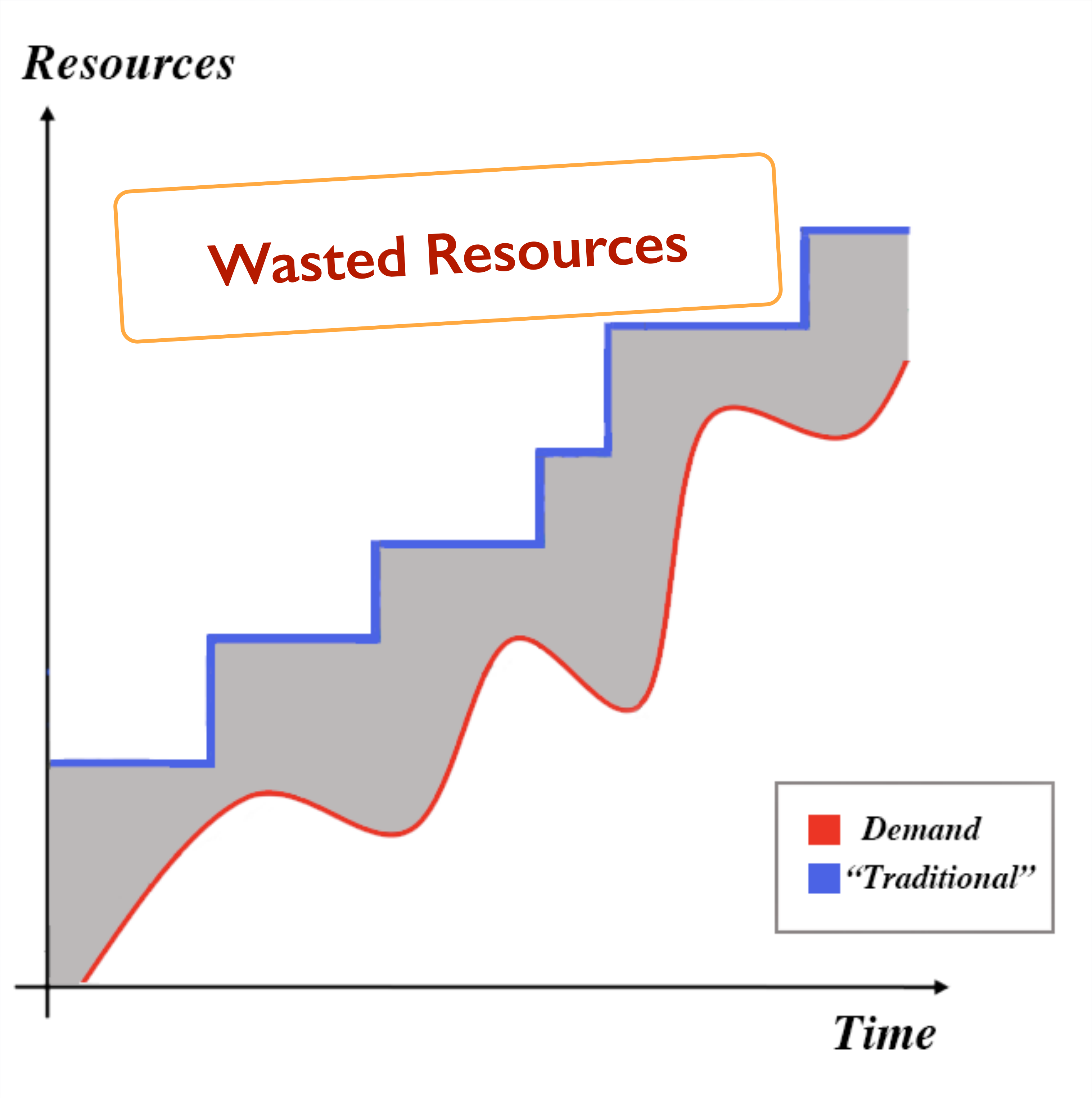
TYPICAL IAAS CLOUD



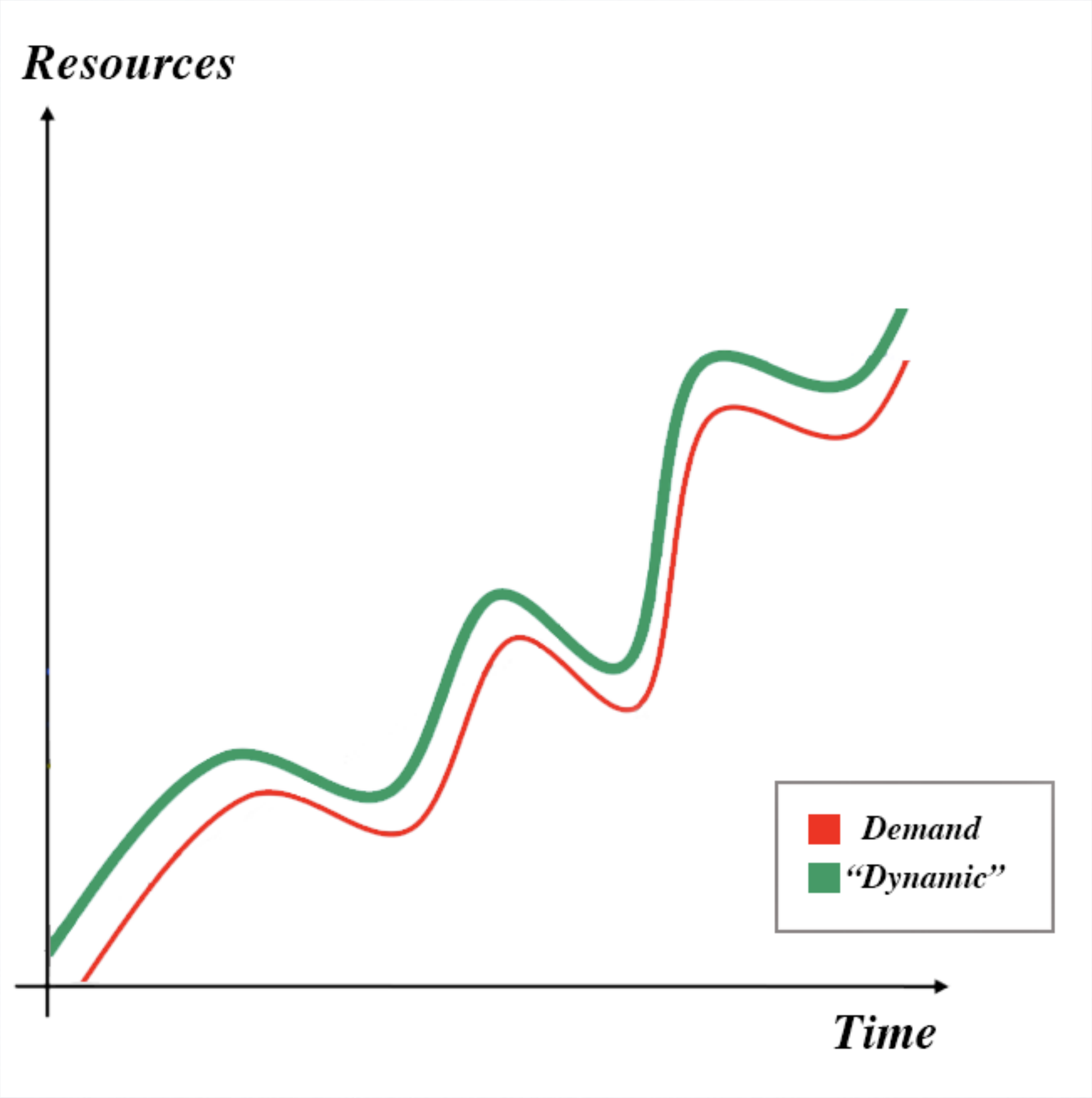
TYPICAL IAAS CLOUD



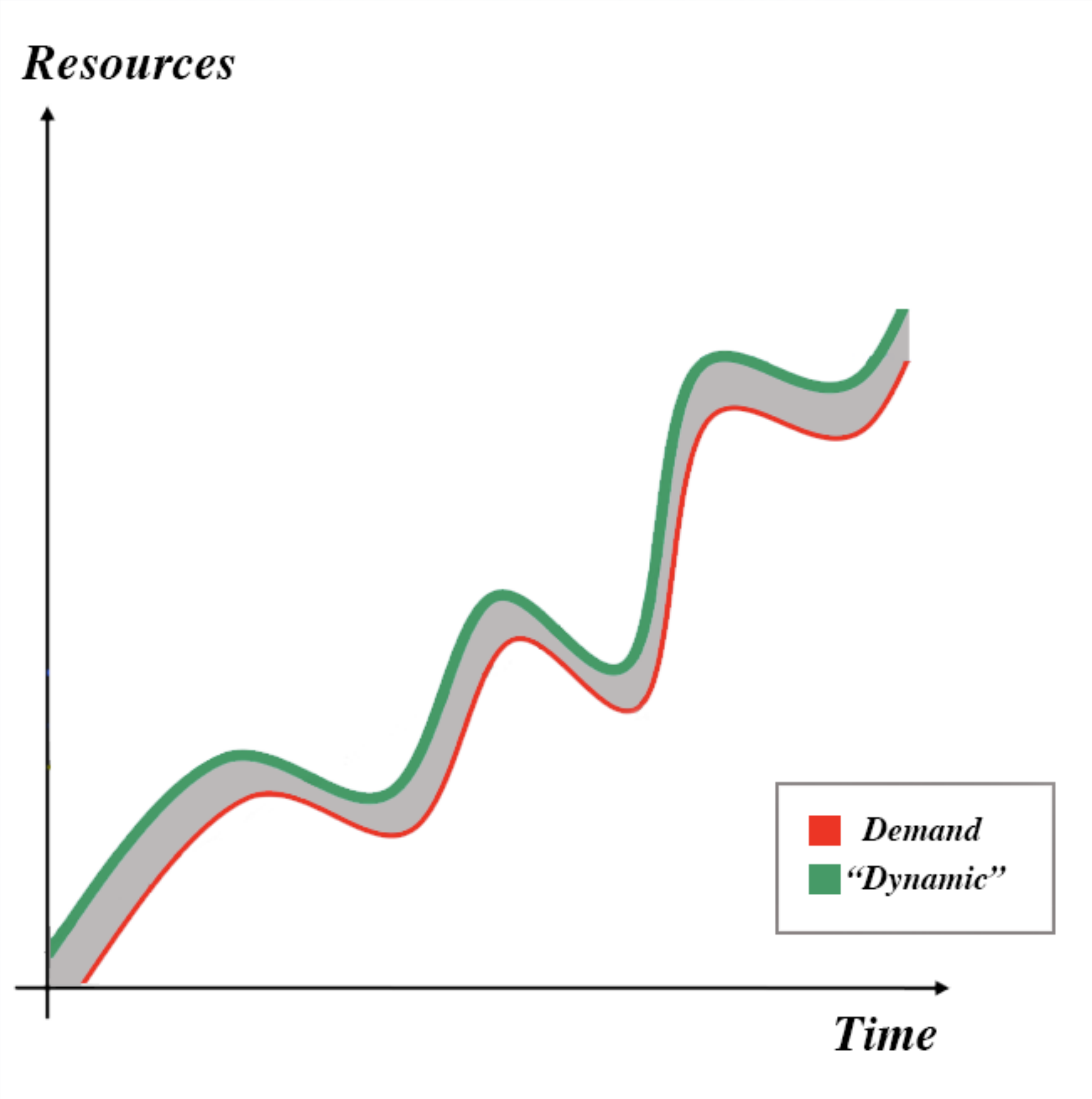
TYPICAL IAAS CLOUD



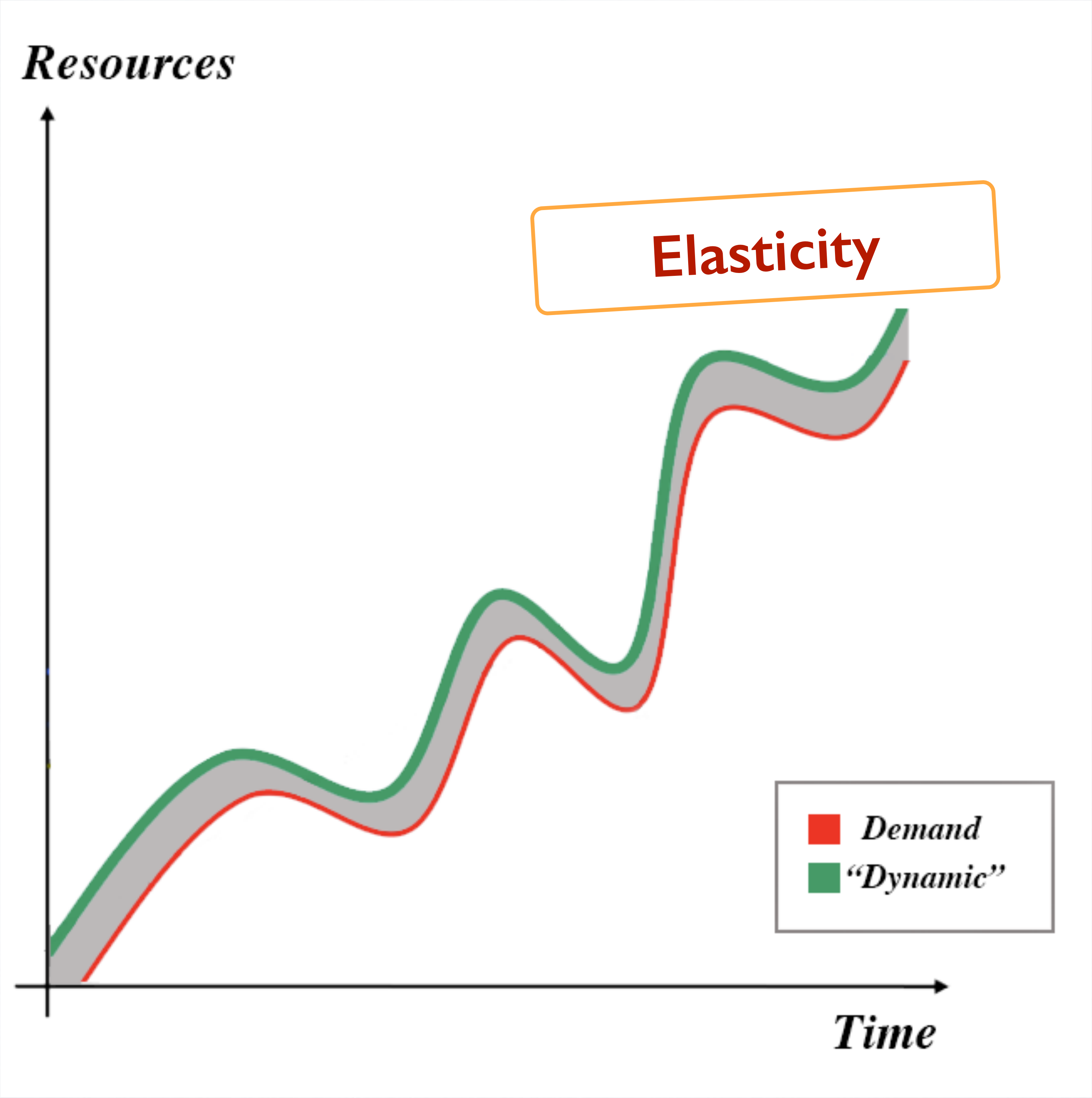
TYPICAL IAAS CLOUD



TYPICAL IAAS CLOUD



TYPICAL IAAS CLOUD



INTRODUCTION

INTRODUCTION

Current CP elasticity mechanisms:

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***
- Mostly **reactive-based** solutions

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***
- Mostly **reactive-based** solutions

Hinders developers from:

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***
- Mostly **reactive-based** solutions

Hinders developers from:

- **Re-using knowledge** between platforms

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***
- Mostly **reactive-based** solutions

Hinders developers from:

- **Re-using knowledge** between platforms
- Developing **reusable** elasticity algorithms

INTRODUCTION

Current CP elasticity mechanisms:

- No **standardised CP model**
- **CP elasticity *lock-in***
- Mostly **reactive-based** solutions

Hinders developers from:

- **Re-using knowledge** between platforms
- Developing **reusable** elasticity algorithms
- Develop innovative solutions: **predictive**

GOAL

GOAL

We aim for elasticity strategies independent from:

GOAL

We aim for elasticity strategies independent from:

CP

GOAL

We aim for elasticity strategies independent from:

CP

Application

GOAL

We aim for elasticity strategies independent from:

CP

Application

Workload Pattern

SOLUTION

SOLUTION

A innovative framework, **Vadara**:

SOLUTION

A innovative framework, **Vadara**:

- **unique** set of features

SOLUTION

A innovative framework, **Vadara**:

- **unique** set of features
- **decoupled** from the CP

SOLUTION

A innovative framework, **Vadara**:

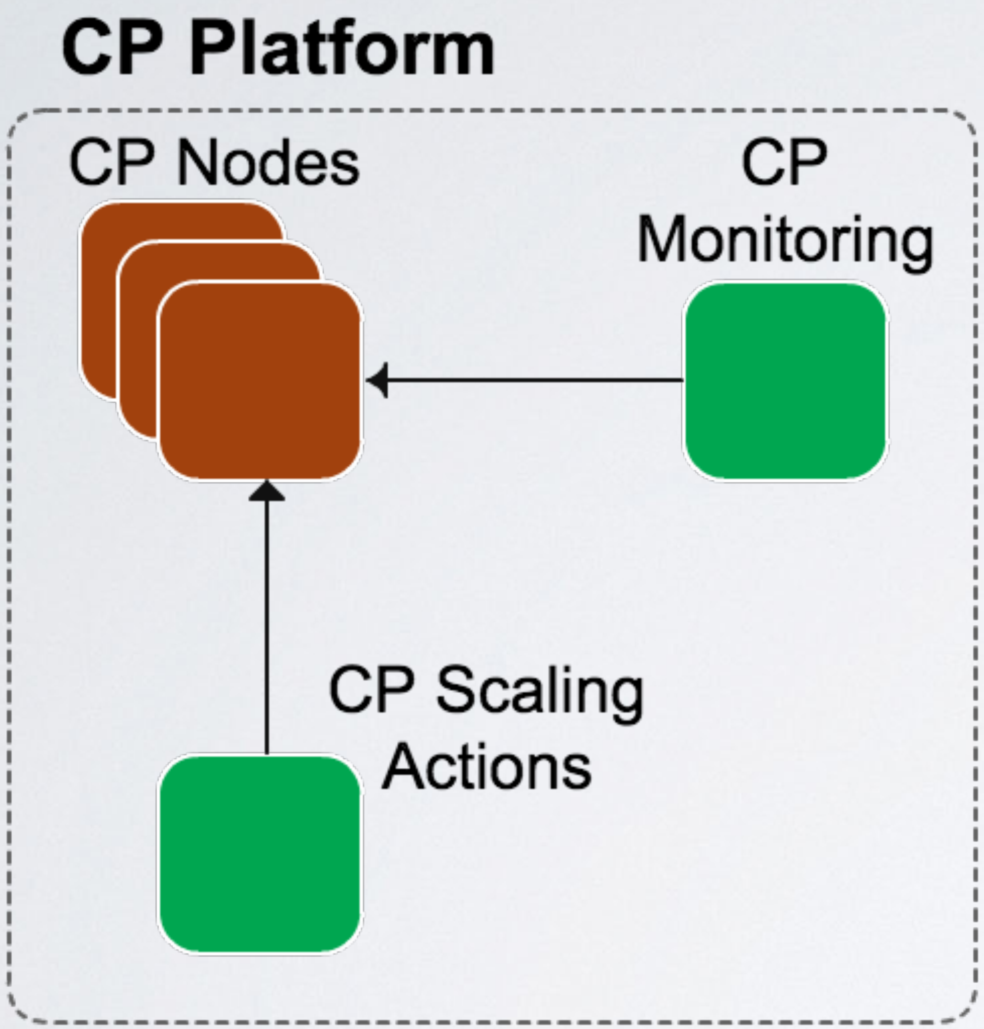
- **unique** set of features
- **decoupled** from the CP
- **generic** regarding the employed elasticity strategy

SOLUTION

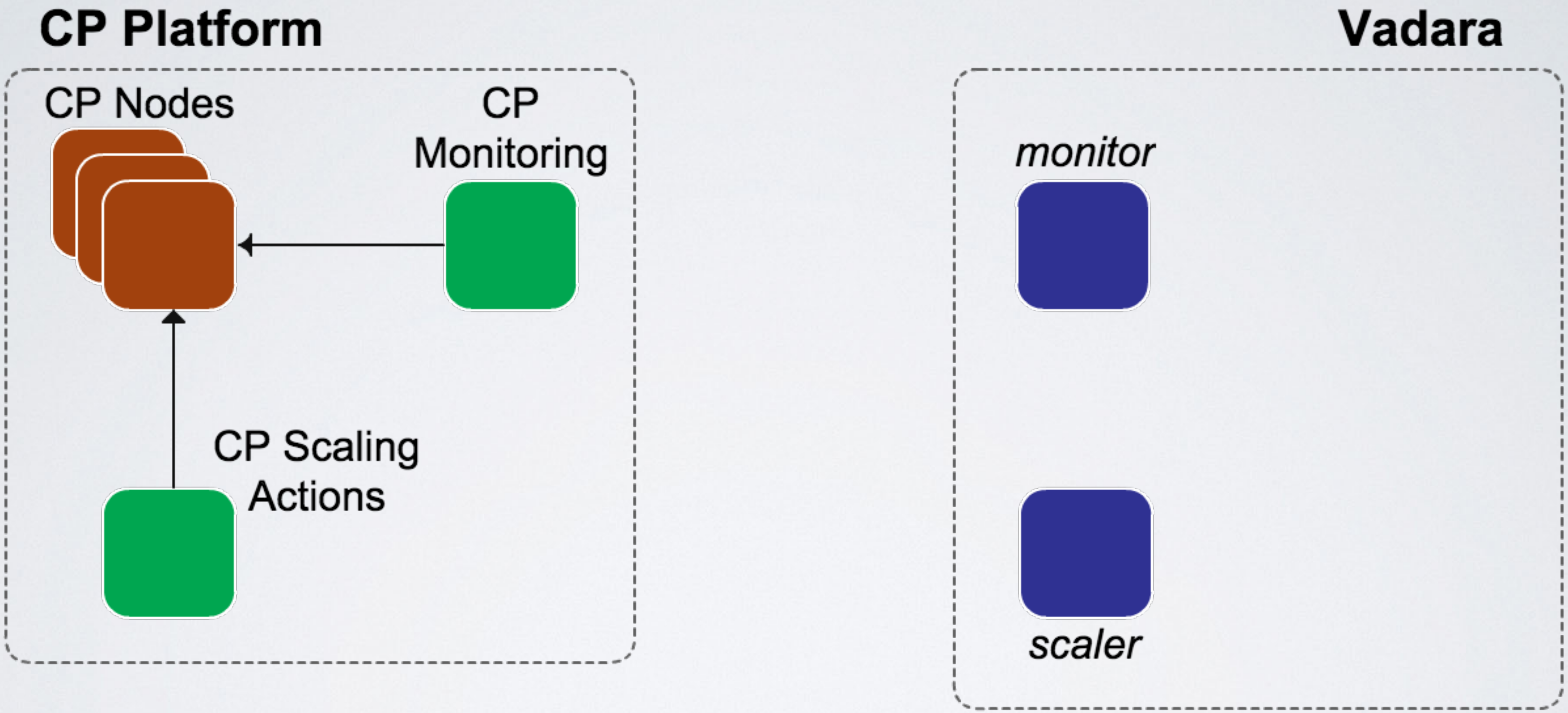
A innovative framework, **Vadara**:

- **unique** set of features
- **decoupled** from the CP
- **generic** regarding the employed elasticity strategy
- **bypasses** CP elasticity lock-in

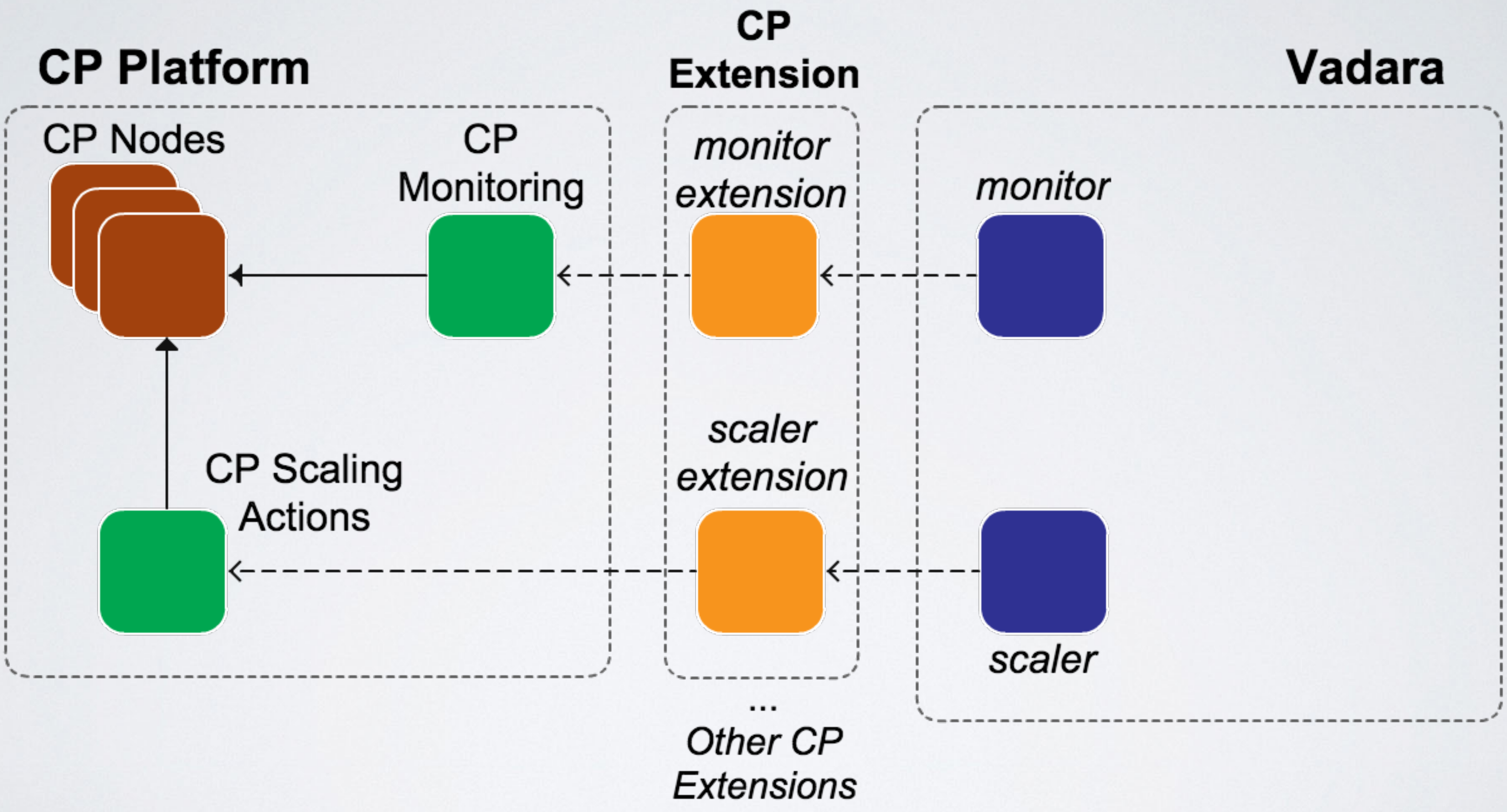
ARCHITECTURE



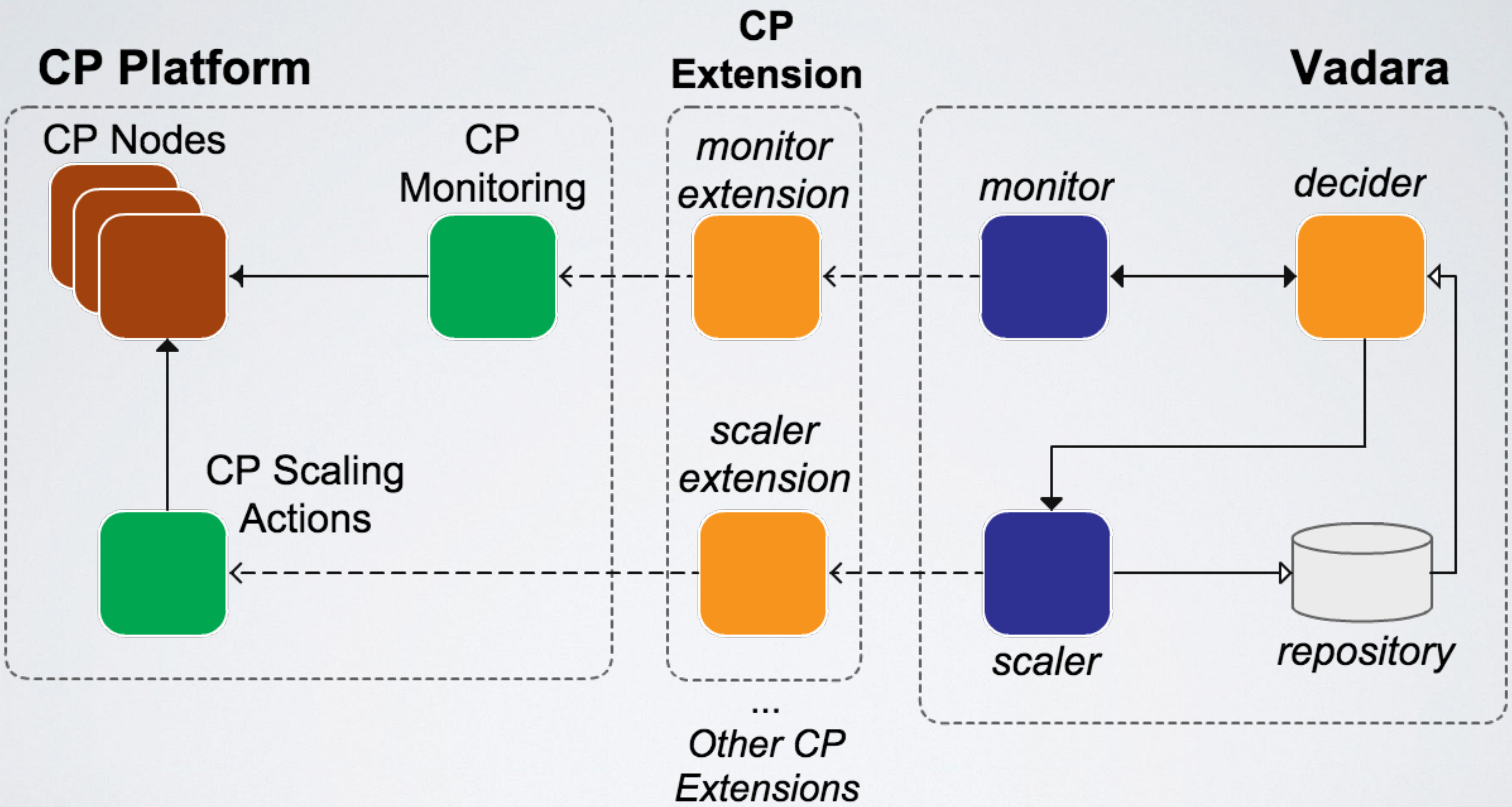
ARCHITECTURE



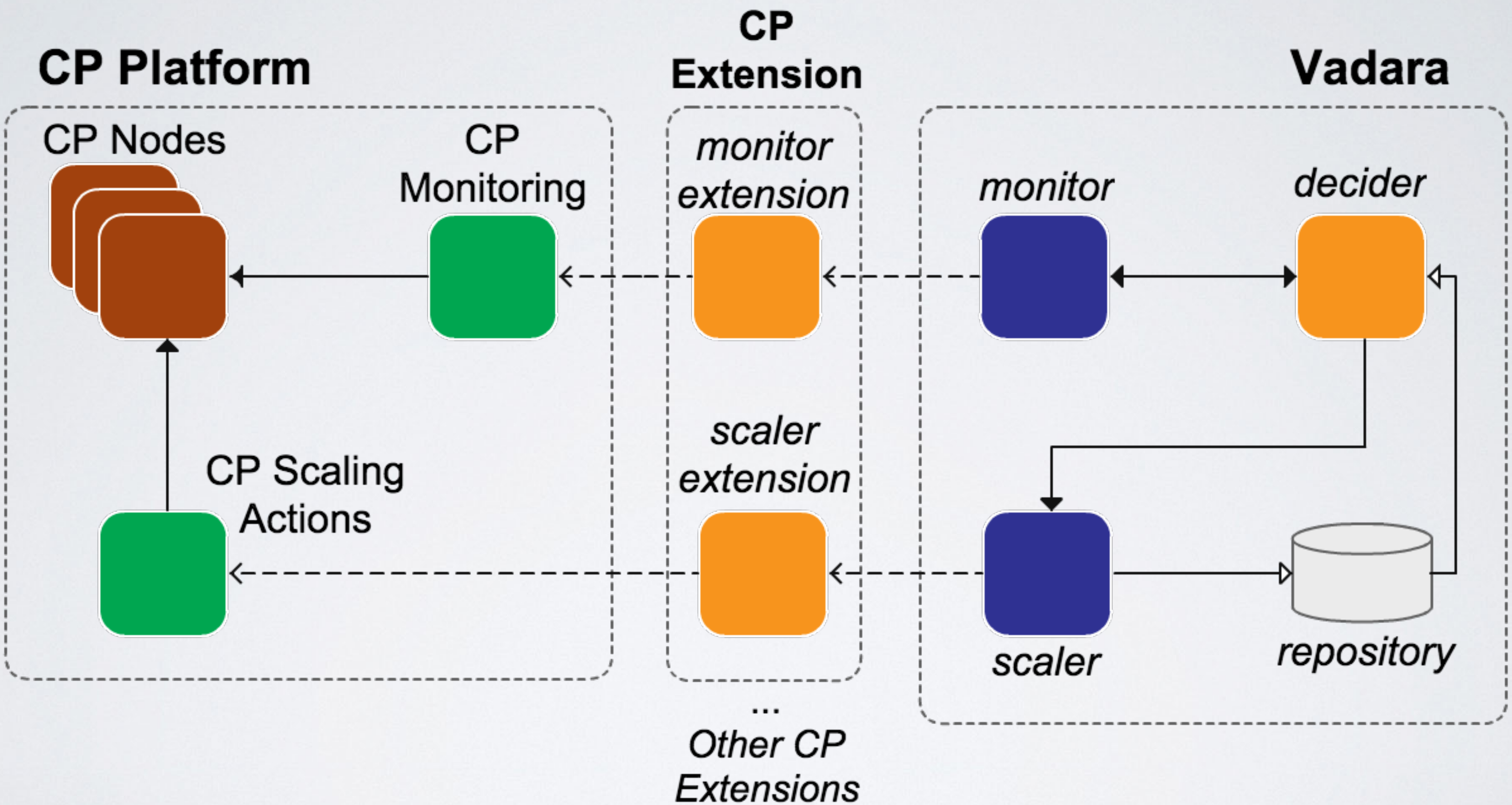
ARCHITECTURE



ARCHITECTURE



ARCHITECTURE



https://github.com/jfloff/vadara

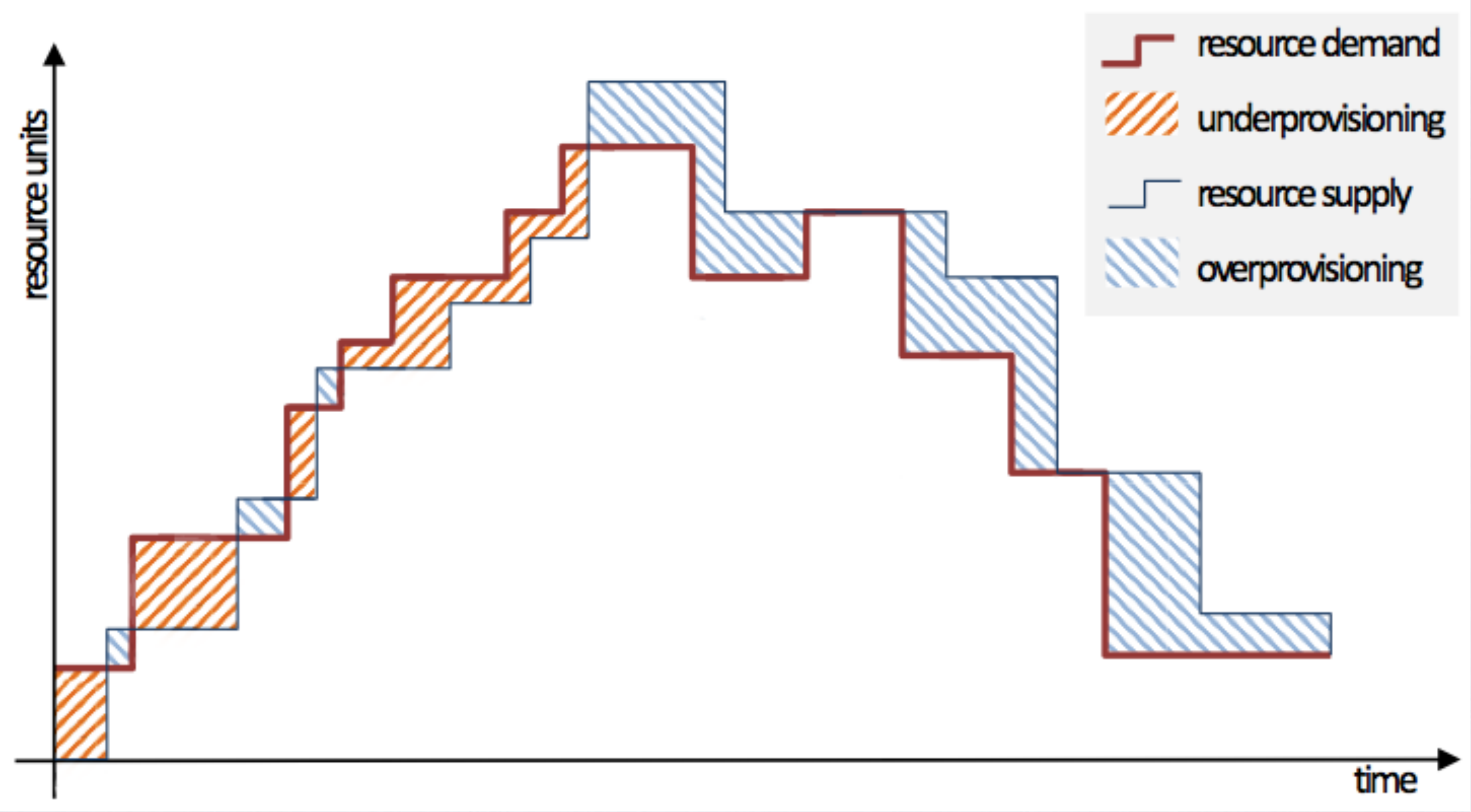
FORECASTING

FORECASTING

Known forecasting methods are only concerned with how close they are to the real value

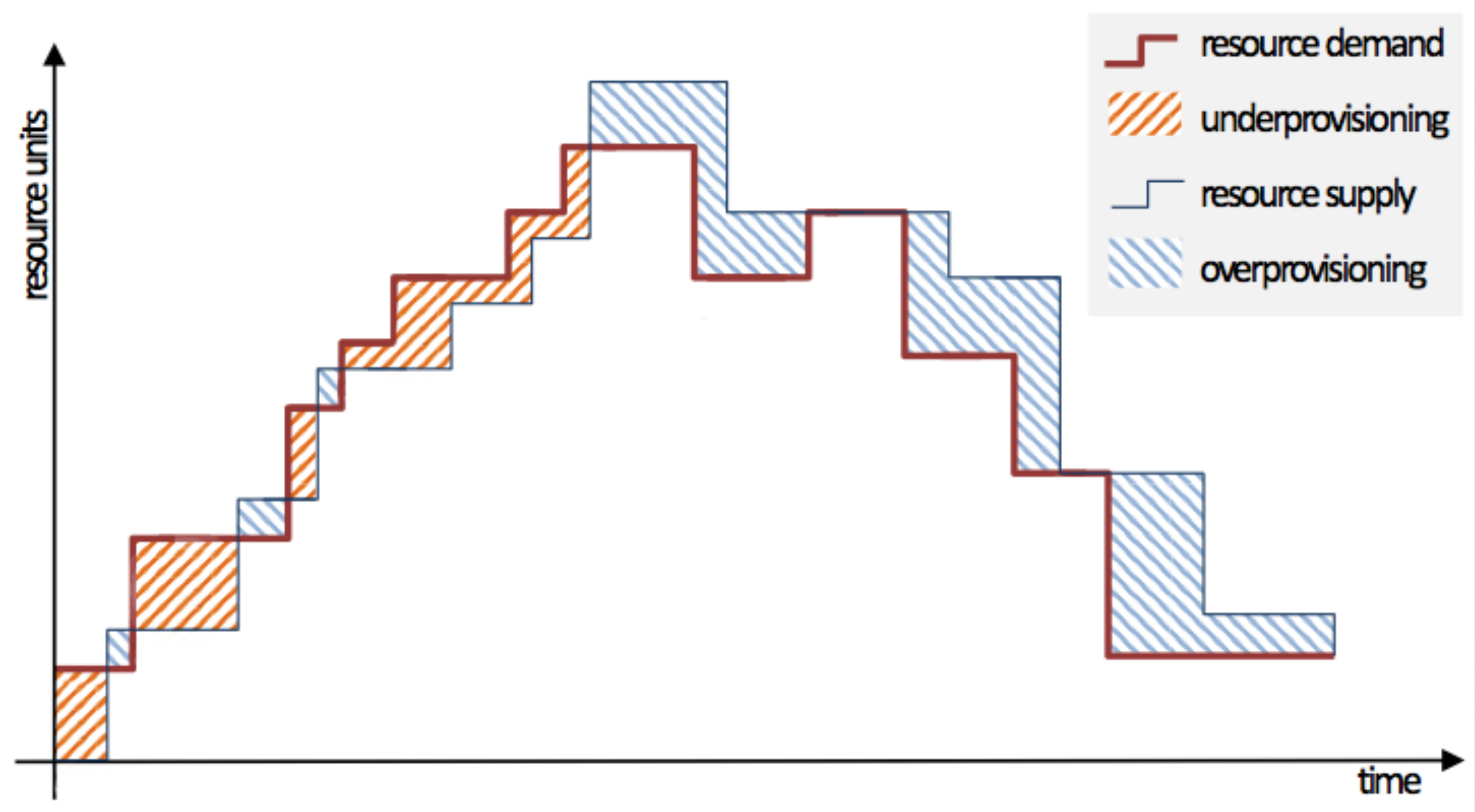
FORECASTING

Known forecasting methods are only concerned with how close they are to the real value



FORECASTING

Known forecasting methods are only concerned with how close they are to the real value



Goal: take a known forecasting method and dynamically pad its value, fixing under and over-provisioning occurrences

DYNAMIC PADDING

DYNAMIC PADDING

INDIVIDUAL METHODS: dynamically pad a methods' original forecast based on the most recent prediction errors of under and over-provisioning.

$$pad_t =$$

DYNAMIC PADDING

INDIVIDUAL METHODS: dynamically pad a methods' original forecast based on the most recent prediction errors of under and over-provisioning.

1. Calculate **EME** = weighted average of error observations where most recent observations have more weight.

$$pad_t =$$

DYNAMIC PADDING

INDIVIDUAL METHODS: dynamically pad a methods' original forecast based on the most recent prediction errors of under and over-provisioning.

1. Calculate **EME** = weighted average of error observations where most recent observations have more weight.

$$pad_t = \quad EME_t(O_t) + \quad EME_t(U_t)$$

DYNAMIC PADDING

INDIVIDUAL METHODS: dynamically pad a methods' original forecast based on the most recent prediction errors of under and over-provisioning.

1. Calculate **EME** = weighted average of error observations where most recent observations have more weight.

2. Count the number of errors for both occurrences.

$$pad_t = \frac{n_O}{n} EME_t(O_t) + \frac{n_U}{n} EME_t(U_t)$$

DYNAMIC PADDING

INDIVIDUAL METHODS: dynamically pad a methods' original forecast based on the most recent prediction errors of under and over-provisioning.

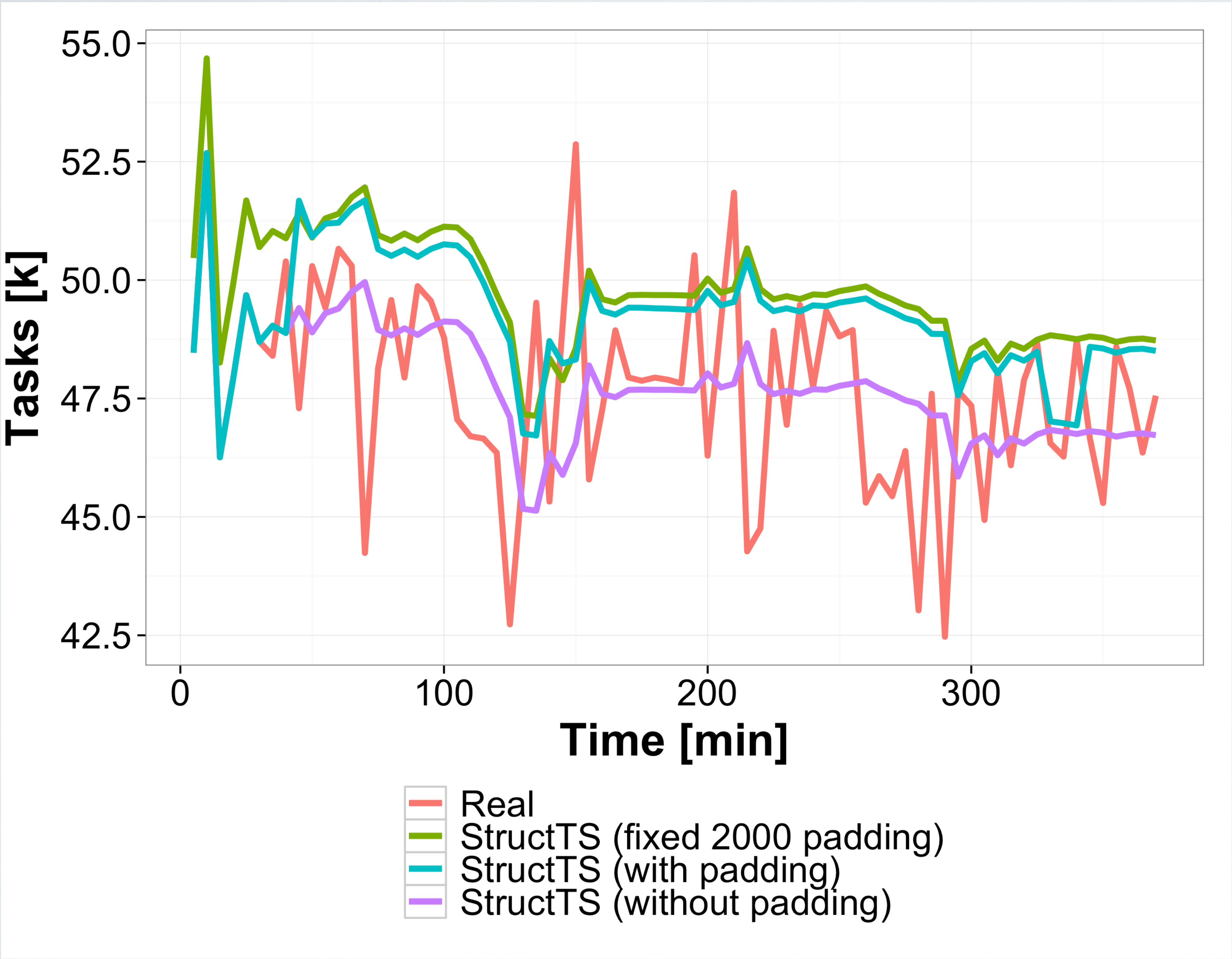
1. Calculate **EME** = weighted average of error observations where most recent observations have more weight.

2. Count the number of errors for both occurrences.

3. Padding value is a weighted average of both EMEs, where the weights are the ratios of over- and under-provisioning occurrences

$$pad_t = \frac{n_O}{n} EME_t(O_t) + \frac{n_U}{n} EME_t(U_t)$$

PADDING



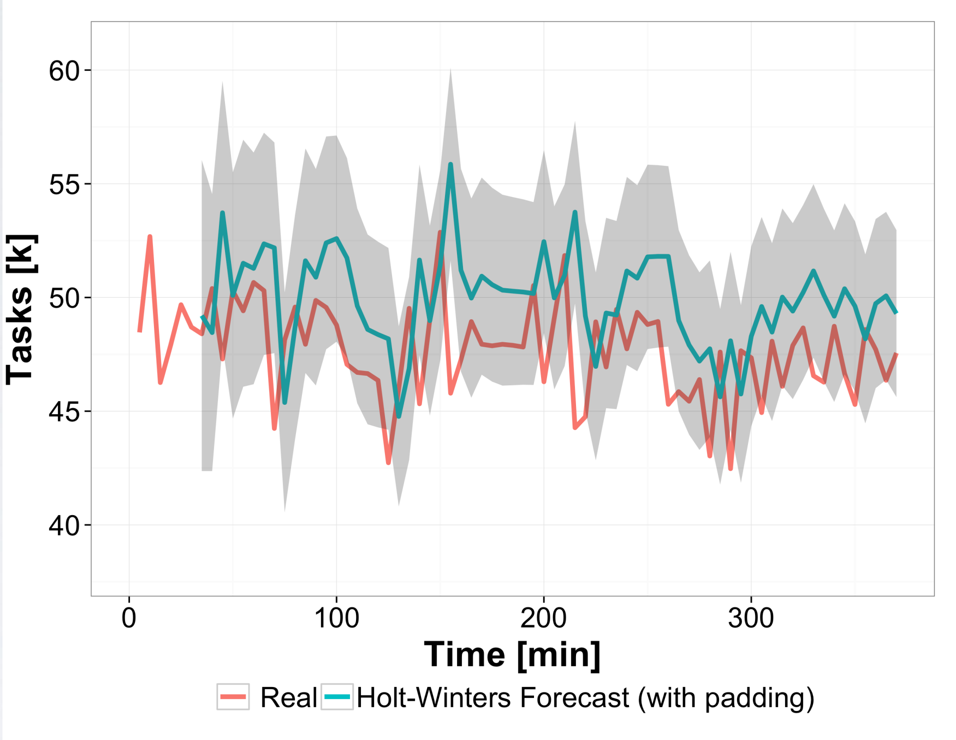
Mean Average Percentage Error (MAPE):

- **No padding: 3.2%**
- **Fixed padding: 5.1%**
- **Dynamic padding: 4.2%**

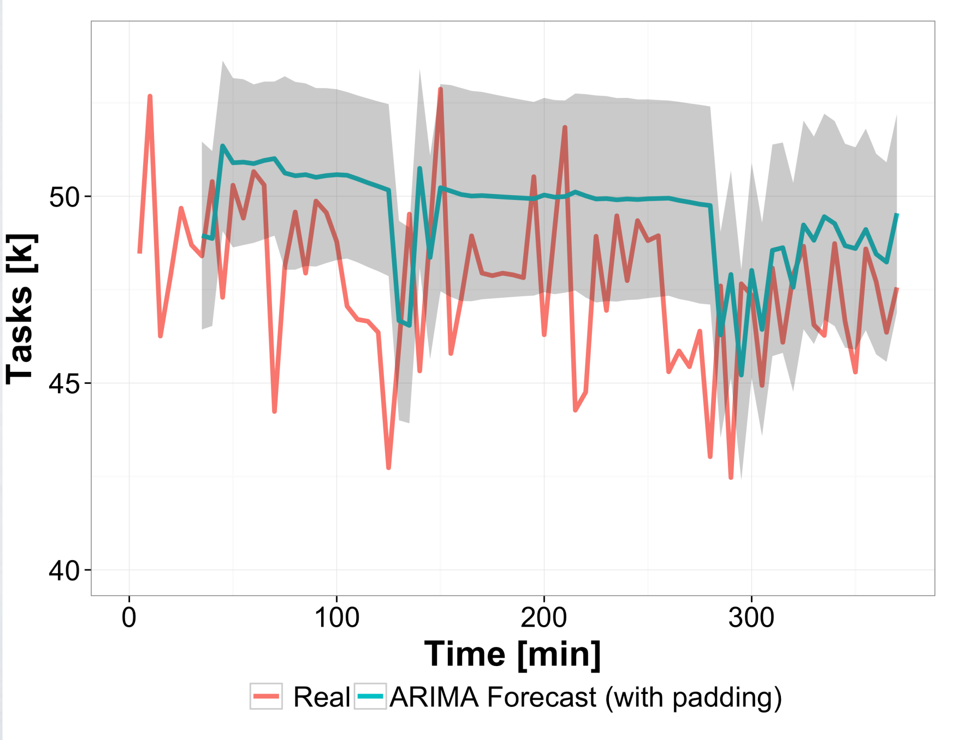
BEHAVIOUR

BEHAVIOUR

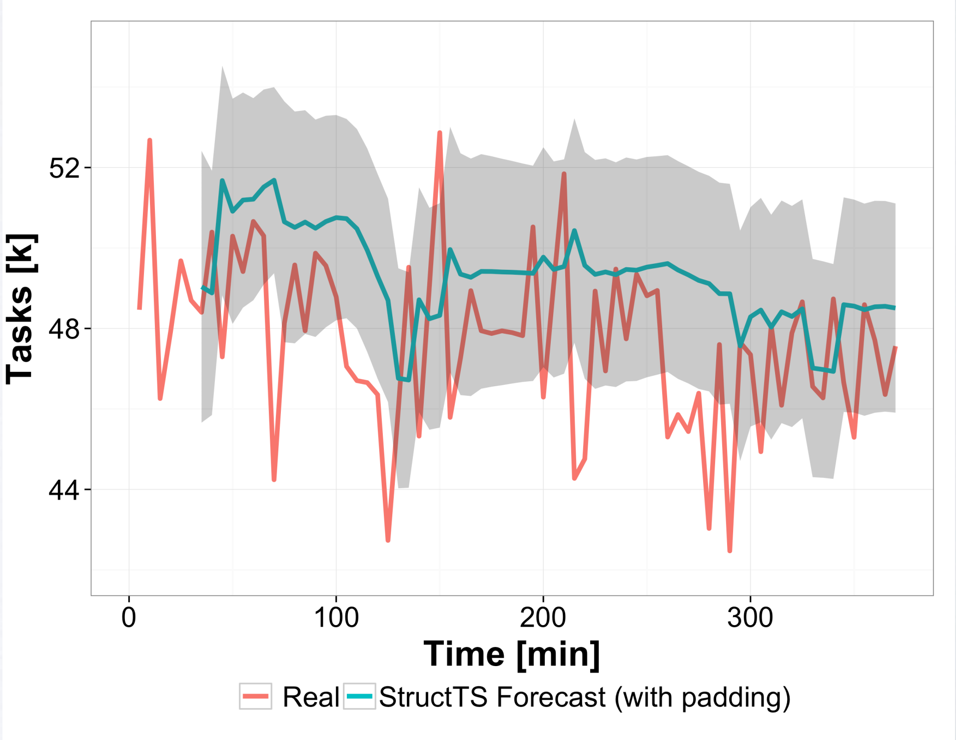
Holt-Winters



ARIMA

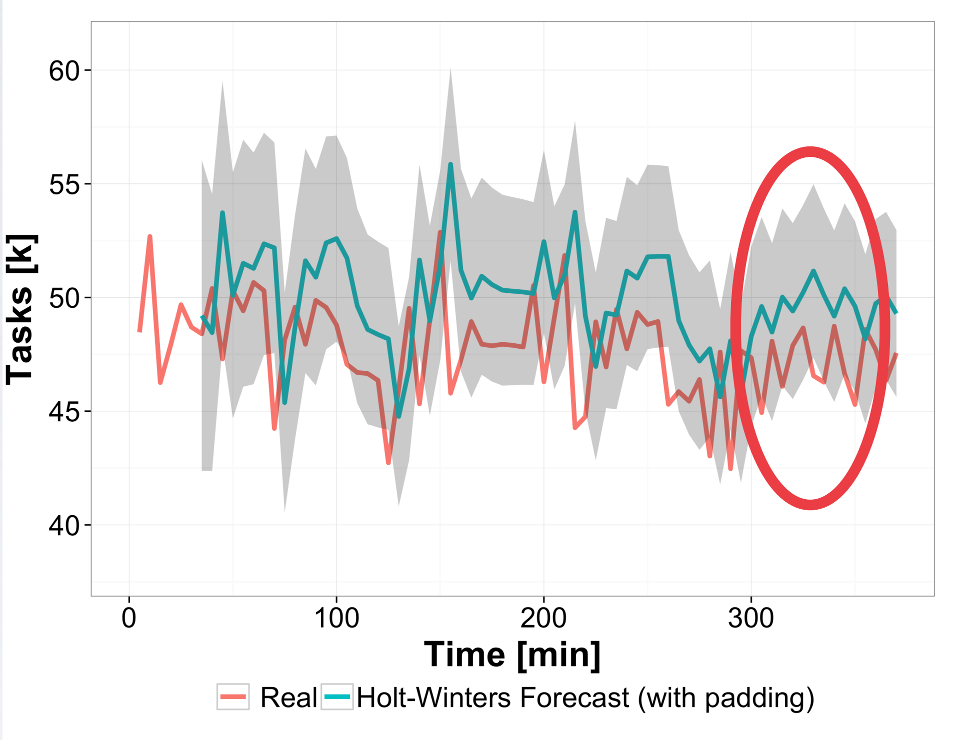


StructTS

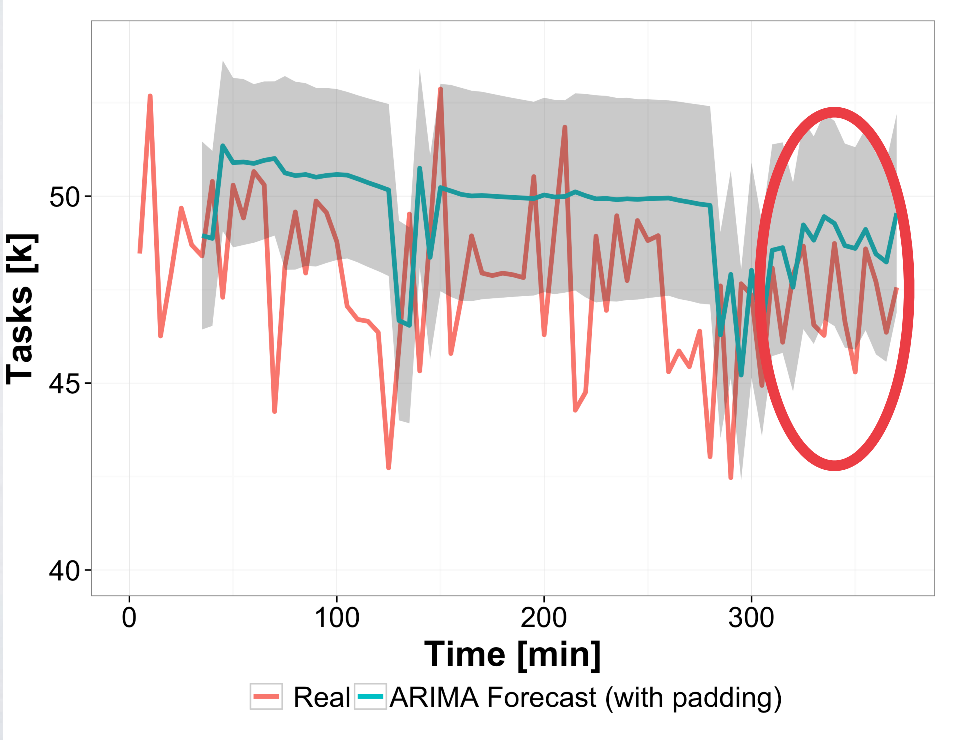


BEHAVIOUR

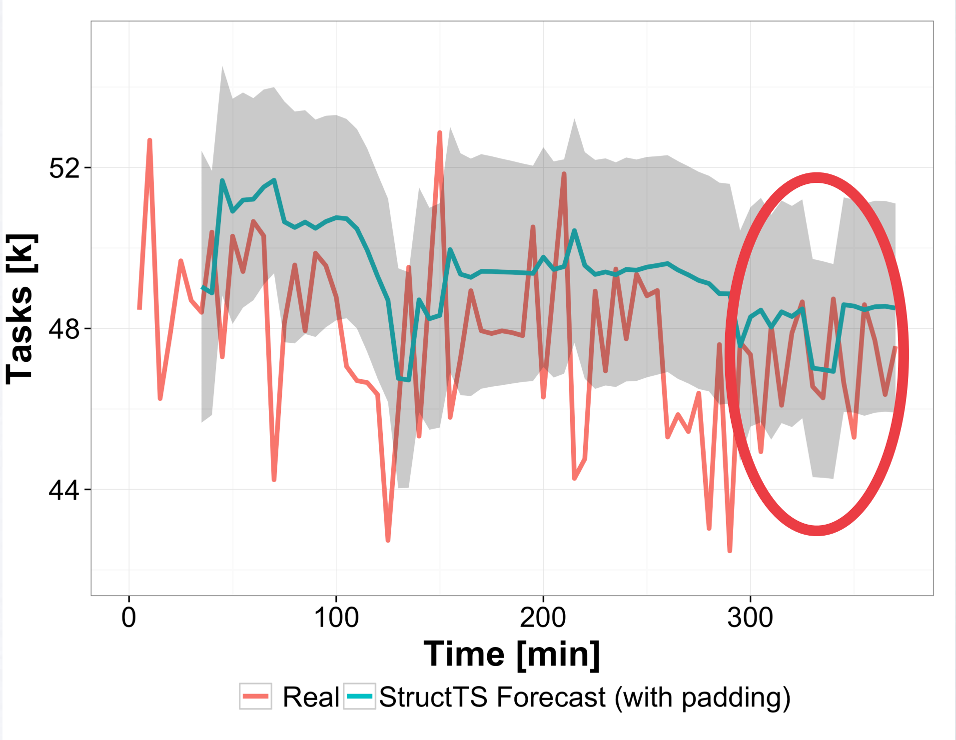
Holt-Winters



ARIMA



StructTS



ENSEMBLE APPROACH

ENSEMBLE APPROACH

ENSEMBLE METHOD: a weighted kNN-like algorithm by most recent forecast performance:

ENSEMBLE APPROACH

ENSEMBLE METHOD: a weighted kNN-like algorithm by most recent forecast performance:

1. Compute *p* individual forecasting methods (previously padded)

ENSEMBLE APPROACH

ENSEMBLE METHOD: a weighted kNN-like algorithm by most recent forecast performance:

1. Compute *p* individual forecasting methods (previously padded)

2. For each method *p* compute the *EME* of its accuracy (MAPE)

ENSEMBLE APPROACH

ENSEMBLE METHOD: a weighted kNN-like algorithm by most recent forecast performance:

1. Compute *p* individual forecasting methods (previously padded)
2. For each method *p* compute the *EME* of its accuracy (MAPE)
3. Choose the *k* individual methods that have recently been closer to the real workload value

ENSEMBLE APPROACH

ENSEMBLE METHOD: a weighted kNN-like algorithm by most recent forecast performance:

1. Compute *p* individual forecasting methods (previously padded)
2. For each method *p* compute the *EME* of its accuracy (MAPE)
3. Choose the *k* individual methods that have recently been closer to the real workload value
4. Calculate the final forecast value:

$$\hat{Y}_t = \sum_{i=1}^{k} w_i Y_{k\,t}, \text{with } w_i = {}^{1}/_{EME_t(A_{t_k})}$$

EVALUATION

EVALUATION

- Does Vadara correctly handles cloud application's behaviour?

EVALUATION

- Does Vadara correctly handles cloud application's behaviour?
- Can it handle more than one CP?

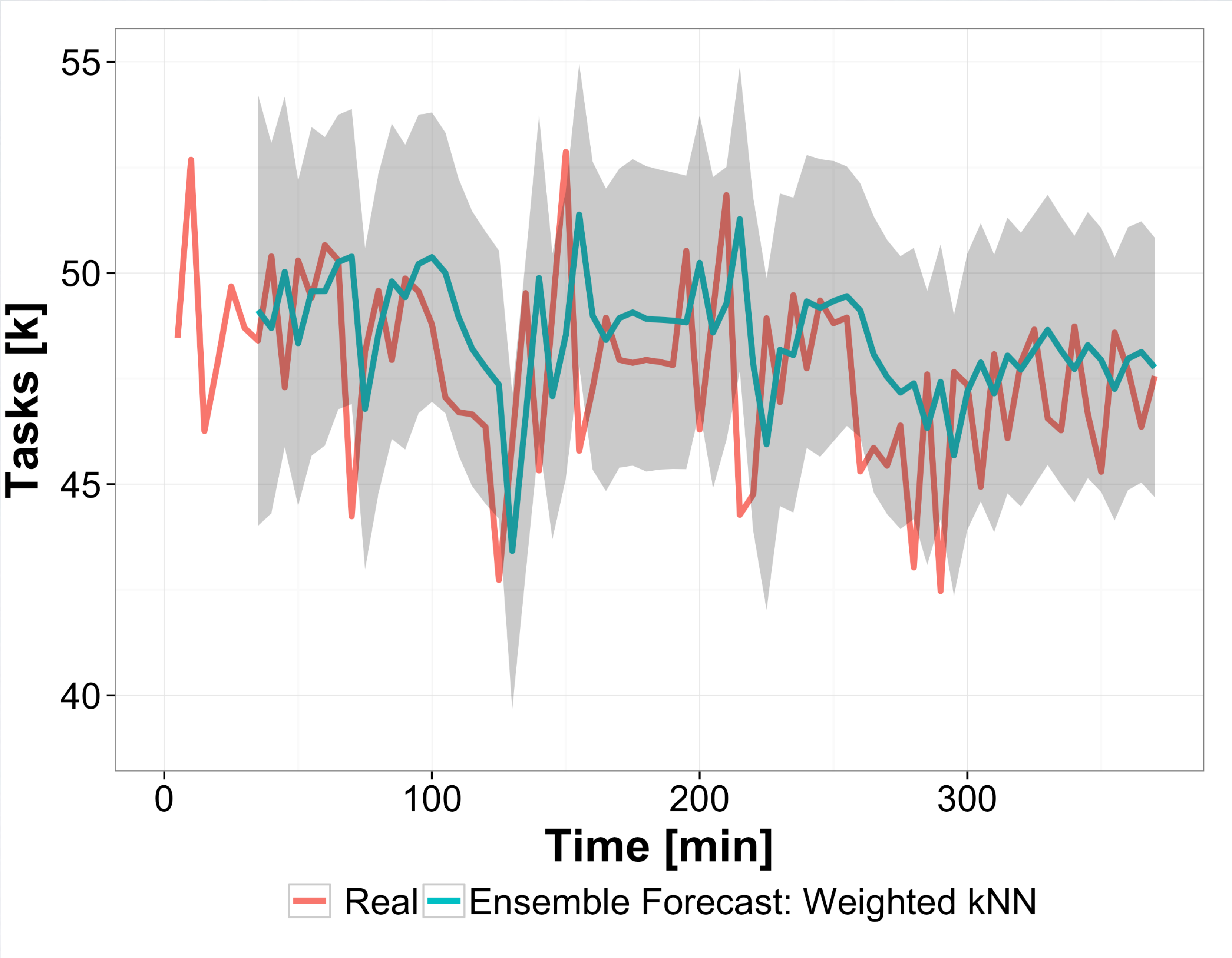
EVALUATION

- Does Vadara correctly handles cloud application's behaviour?
- Can it handle more than one CP?
- Does our ensemble approach correctly forecasts cloud application's demand?

EVALUATION

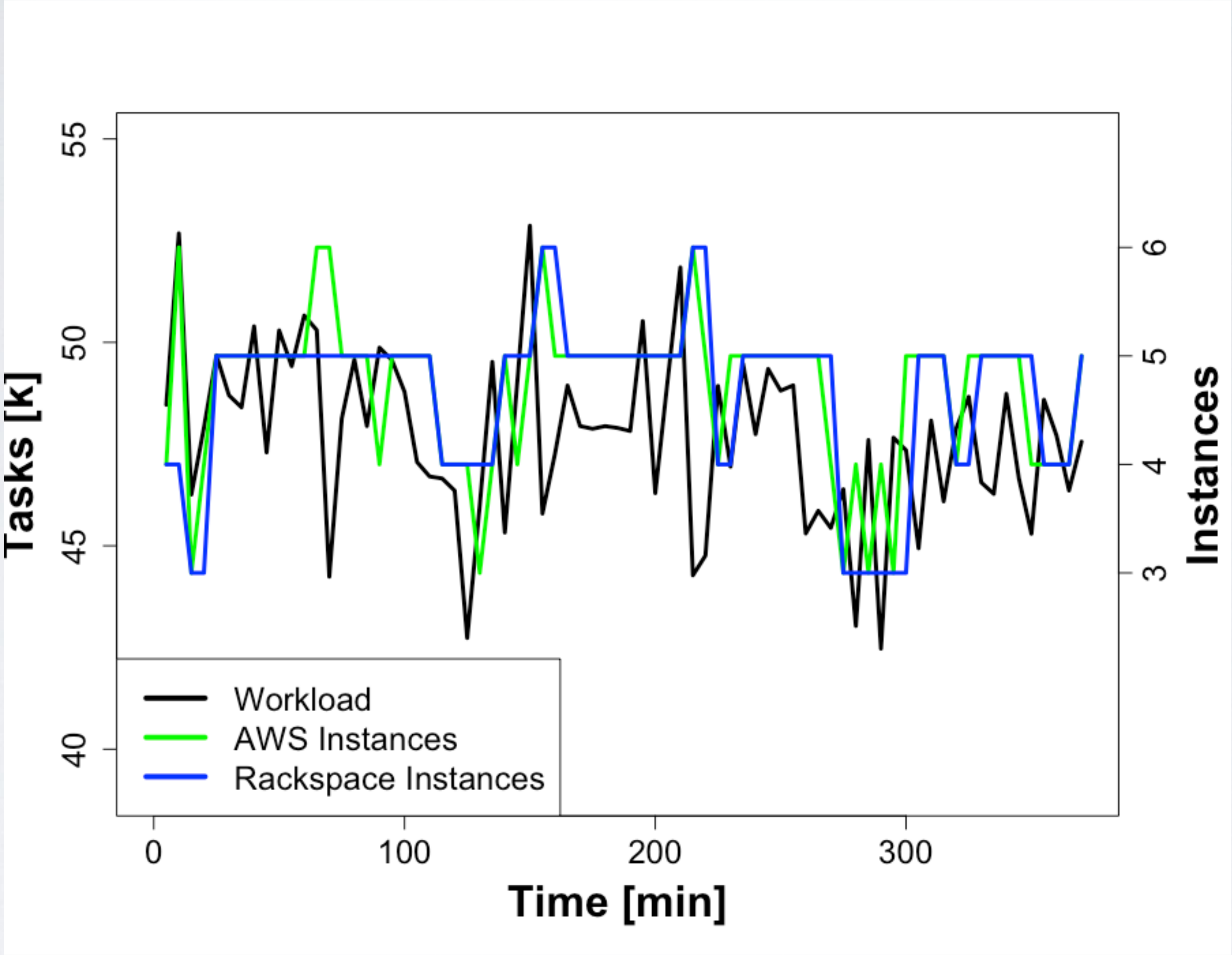
- Does Vadara correctly handles cloud application's behaviour?
- Can it handle more than one CP?
- Does our ensemble approach correctly forecasts cloud application's demand?
- How does it compare to individual methods?

ENSEMBLE APPROACH



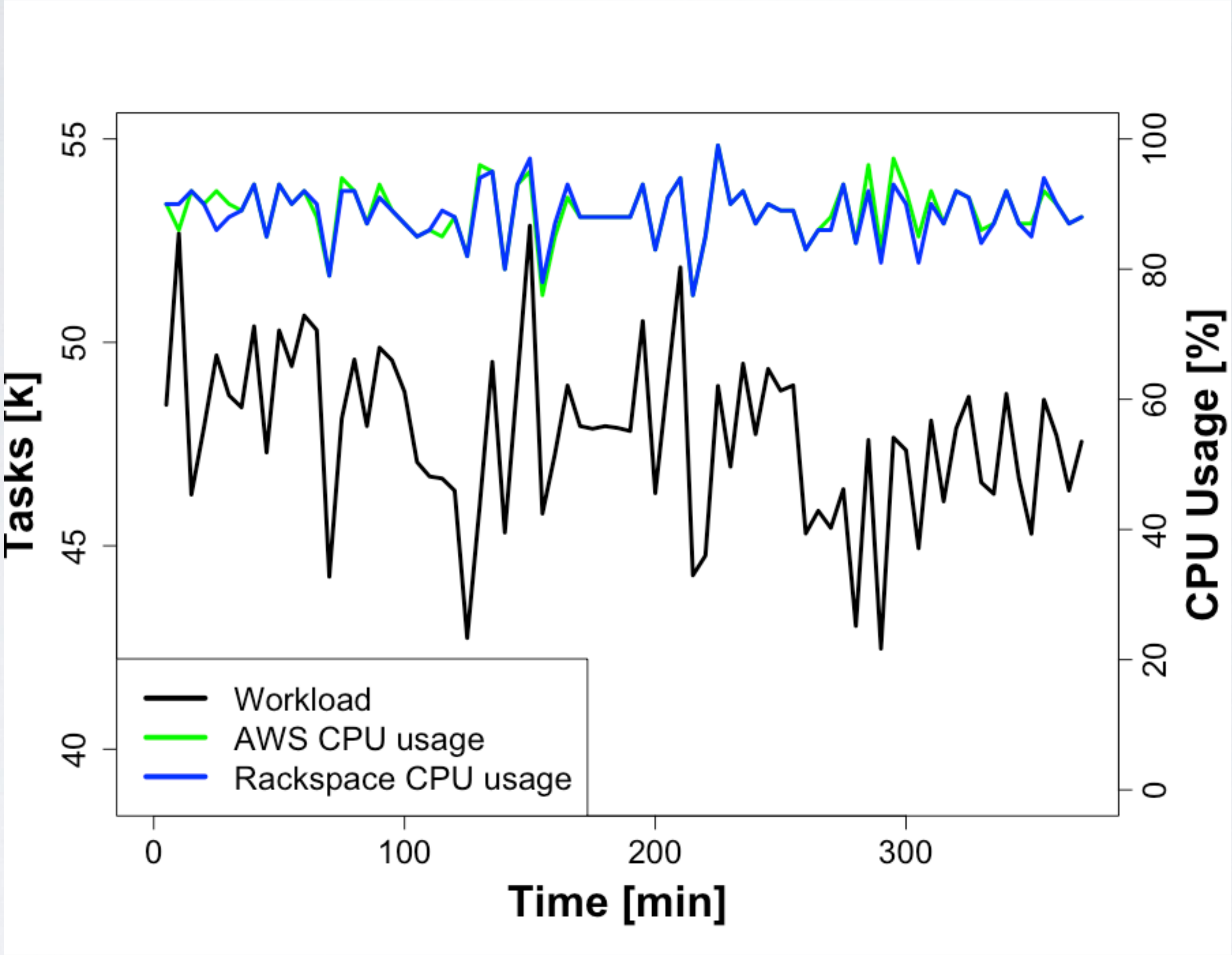
2.5% MAPE — 55% Improvement

ENSEMBLE APPROACH



CPU Bound application

ENSEMBLE APPROACH



Stays maximized!

RESULTS

RESULTS

For **individual methods**, using **padding**:

RESULTS

For **individual methods**, using **padding**:

- Reduction in observed under-provisioning occurrences (30% on average)

RESULTS

For **individual methods**, using **padding**:

- Reduction in observed under-provisioning occurrences (30% on average)

Ensemble method:

RESULTS

For **individual methods**, using **padding**:

- Reduction in observed under-provisioning occurrences (30% on average)

Ensemble method:

- Less than 22% of under-provisioning occurrences

RESULTS

For **individual methods**, using **padding**:

- Reduction in observed under-provisioning occurrences (30% on average)

Ensemble method:

- Less than 22% of under-provisioning occurrences
- Near 65% of over-provisioning occurrences

RESULTS

For **individual methods**, using **padding**:

- Reduction in observed under-provisioning occurrences (30% on average)

Ensemble method:

- Less than 22% of under-provisioning occurrences
- Near 65% of over-provisioning occurrences
- Near 13% of near 'perfect' forecasts

CONTRIBUTIONS

CONTRIBUTIONS

- **Vadara:** generic framework that allows the development of CP agnostic strategies.

CONTRIBUTIONS

- **Vadara:** generic framework that allows the development of CP agnostic strategies.
- **Padding system:** for demand forecasts based on most recent under and over-provisioning observations

CONTRIBUTIONS

- **Vadara:** generic framework that allows the development of CP agnostic strategies.
- **Padding system:** for demand forecasts based on most recent under and over-provisioning observations
- **Ensemble forecasting algorithm:** a weighted kNN-like algorithm by most recent forecast performance

CONTRIBUTIONS

- **Vadara:** generic framework that allows the development of CP agnostic strategies.
- **Padding system:** for demand forecasts based on most recent under and over-provisioning observations
- **Ensemble forecasting algorithm:** a weighted kNN-like algorithm by most recent forecast performance
    1. Reduction in under-provisioning observations in over 15%

CONTRIBUTIONS

- **Vadara:** generic framework that allows the development of CP agnostic strategies.
- **Padding system:** for demand forecasts based on most recent under and over-provisioning observations
- **Ensemble forecasting algorithm:** a weighted kNN-like algorithm by most recent forecast performance

1. Reduction in under-provisioning observations in over 15%
2. MAPE reduction in more than half

THANK YOU!

QUESTIONS?

RELATED WORK

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch

Doesn't allow the development of predictive strategies

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch

It's another form of lock-in

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**

It's another form of lock-in

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**
  - Yang et al., Mao et al., Kranas et al. and Morais et al.

It's another form of lock-in

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**
  - Yang et al., Mao et al., Kranas et al.

It's another form of lock-in

Doesn't offer the same set of features as Vadara

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**
  - Yang et al., Mao et al., Kranas et al.
- **Demand forecasting:**

It's another form of lock-in

Doesn't offer the same set of features as Vadara

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**
  - Yang et al., Mao et al., Kranas et al.
- **Demand forecasting:**
  - Shen et al. , Jiang et al. , Gandhi et al. , Roy et al.

It's another form of lock-in

Doesn't offer the same set of features as Vadara

RELATED WORK

- State of the art: **Galante et al.** and **Lorido-Botrán et al.**
- **CPs:** AWS, Rackspace, Azure
- **CMPs:** RightScale, Sclar, Enstratius, AzureWatch
- **Frameworks:**
  - Yang et al., Mao et al., Kranas et al.
- **Demand forecasting:**
  - Shen et al. , Jiang et al. , Gandhi et al.

It's another form of lock-in

Doesn't offer the same set of features as Vadara

High number of under-provisioning