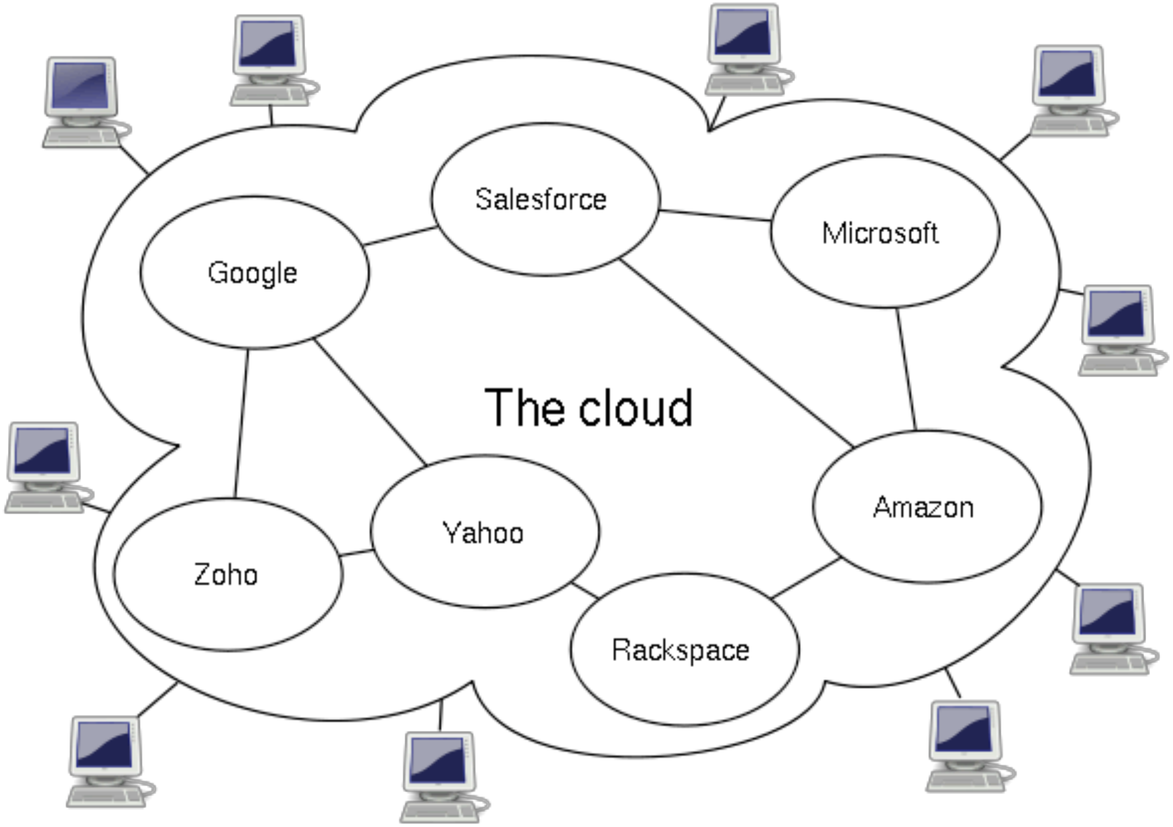
Cloud Computing Overview

CSC 501

Cloud Computing

- Paradigm shift in computing model
  - Provision of dynamically *scalable* and *virtualized* resources as service over Internet
  - Computing and storage details abstracted from users



History

- 1960s: concept of utility computing
  - Computing resources offered as public utility
- Late 1990s: SaaS
  - Software as a Service
  - Application licensed for use as service on demand
  - Extended by Microsoft through web services
- 2005: Amazon
  - Modernized datacenters
  - Started AWS (Amazon Web Services)
    - EC2: Amazon Elastic Compute Cloud
- 2007: Google, IBM, and universities: cloud computing initiative

What is Cloud Computing?

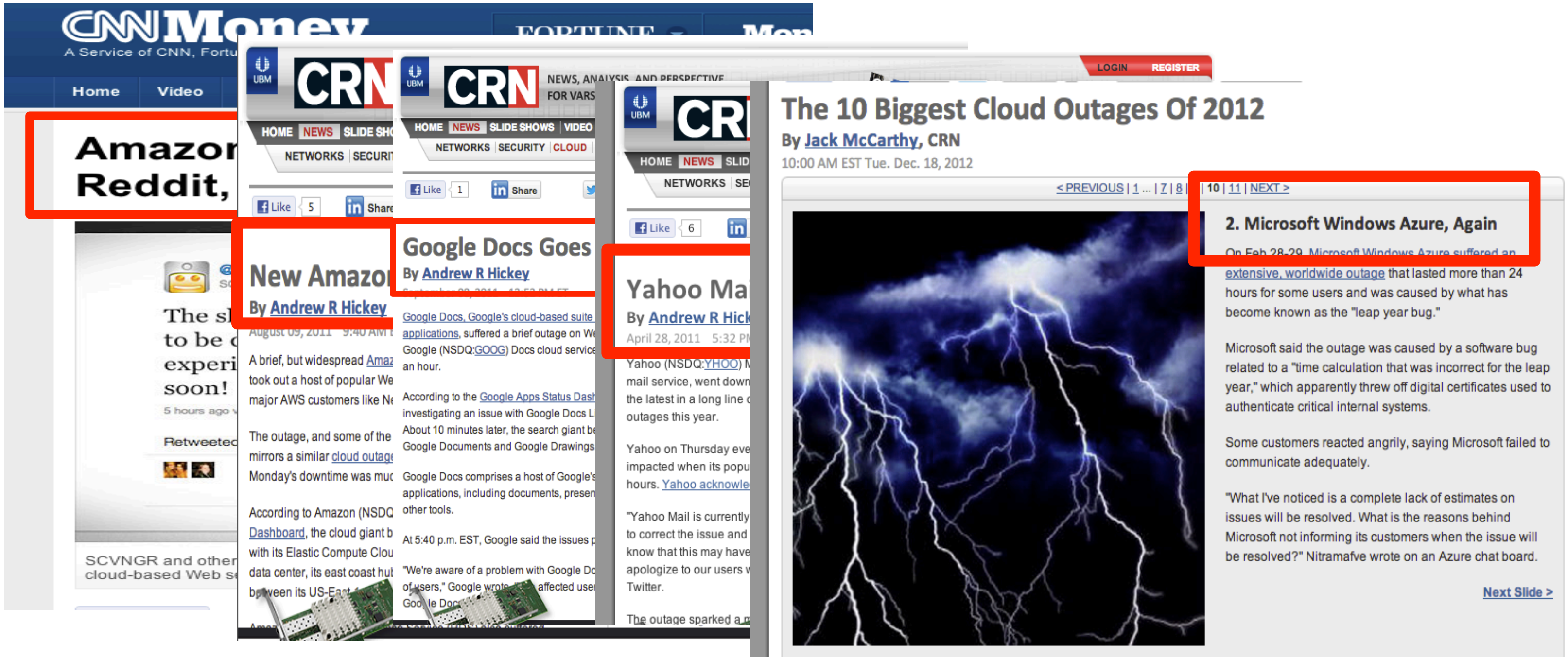
- Computing resource leasing infrastructure
  - Rent computing resources (hardware, infrastructure, virtual machines, software platforms, etc.) for a period of time rather than own resources permanently
  - Pay-as-you-go service model
  - Illusion of infinite resources
  - Resources on-demand
  - No up-front cost
  - Fine-grained billing (e.g. hourly)

Why Cloud Computing?

- Pay-as-you-go resource leasing
  - Datacenter ownership is expensive
  - Avoid infrastructure maintenance hassle
- Elastic scaling
  - Instant resource on-demand
  - Illusion of infinite resources
  - No up-front cost
- Ubiquitous accessibility
  - Data aggregation

Cloud Computing Challenges

- Robustness



Amazor
Reddit,

The sl
to be c
experi
soon!

5 hours ago

Retweetec

SCVNGR and other
cloud-based Web se

New Amazor
By Andrew R Hickey
August 09, 2011 9:40 AM

A brief, but widespread Amaz
took out a host of popular We
major AWS customers like Ne

The outage, and some of the
mirrors a similar cloud outage
Monday's downtime was muc

According to Amazon (NSDQ
Dashboard, the cloud giant b
with its Elastic Compute Clou
data center, its east coast hul
between its US-East

Google Docs Goes
By Andrew R Hickey
September 08, 2011 12:52 PM ET

Google Docs, Google's cloud-based suite
applications, suffered a brief outage on We
Google (NSDQ:GOOG) Docs cloud service
an hour.

According to the Google Apps Status Dash
investigating an issue with Google Docs L
About 10 minutes later, the search giant be
Google Documents and Google Drawings

Google Docs comprises a host of Google's
applications, including documents, presen
other tools.

At 5:40 p.m. EST, Google said the issues p

"We're aware of a problem with Google Do
of users," Google wrote affected use
Google Doc

Yahoo Ma
By Andrew R Hick
April 28, 2011 5:32 PM

Yahoo (NSDQ:YHOO) M
mail service, went down
the latest in a long line o
outages this year.

Yahoo on Thursday eve
impacted when its popu
hours. Yahoo acknowle

"Yahoo Mail is currently
to correct the issue and
know that this may have
apologize to our users w
Twitter.

The outage sparked a m

The 10 Biggest Cloud Outages Of 2012
By Jack McCarthy, CRN
10:00 AM EST Tue. Dec. 18, 2012

2. Microsoft Windows Azure, Again

On Feb 28-29, Microsoft Windows Azure suffered an extensive, worldwide outage that lasted more than 24 hours for some users and was caused by what has become known as the "leap year bug."

Microsoft said the outage was caused by a software bug related to a "time calculation that was incorrect for the leap year," which apparently threw off digital certificates used to authenticate critical internal systems.

Some customers reacted angrily, saying Microsoft failed to communicate adequately.

"What I've noticed is a complete lack of estimates on issues will be resolved. What is the reasons behind Microsoft not informing its customers when the issue will be resolved?" Nitramafve wrote on an Azure chat board.

- Resource and energy efficiency
  - Over-provision resources for performance
- Accountability
  - How do I trust the results from the cloud?

Efficient Cloud Resource Management

- Elastic resource scaling
  - Reduce resource waste
  - Minimize SLO violations
  - Reduce energy consumption by slowing down CPU frequencies
- Server consolidation
  - Reduce the number of physical hosts
  - Reduce energy consumption by shutting down machines

Motivation

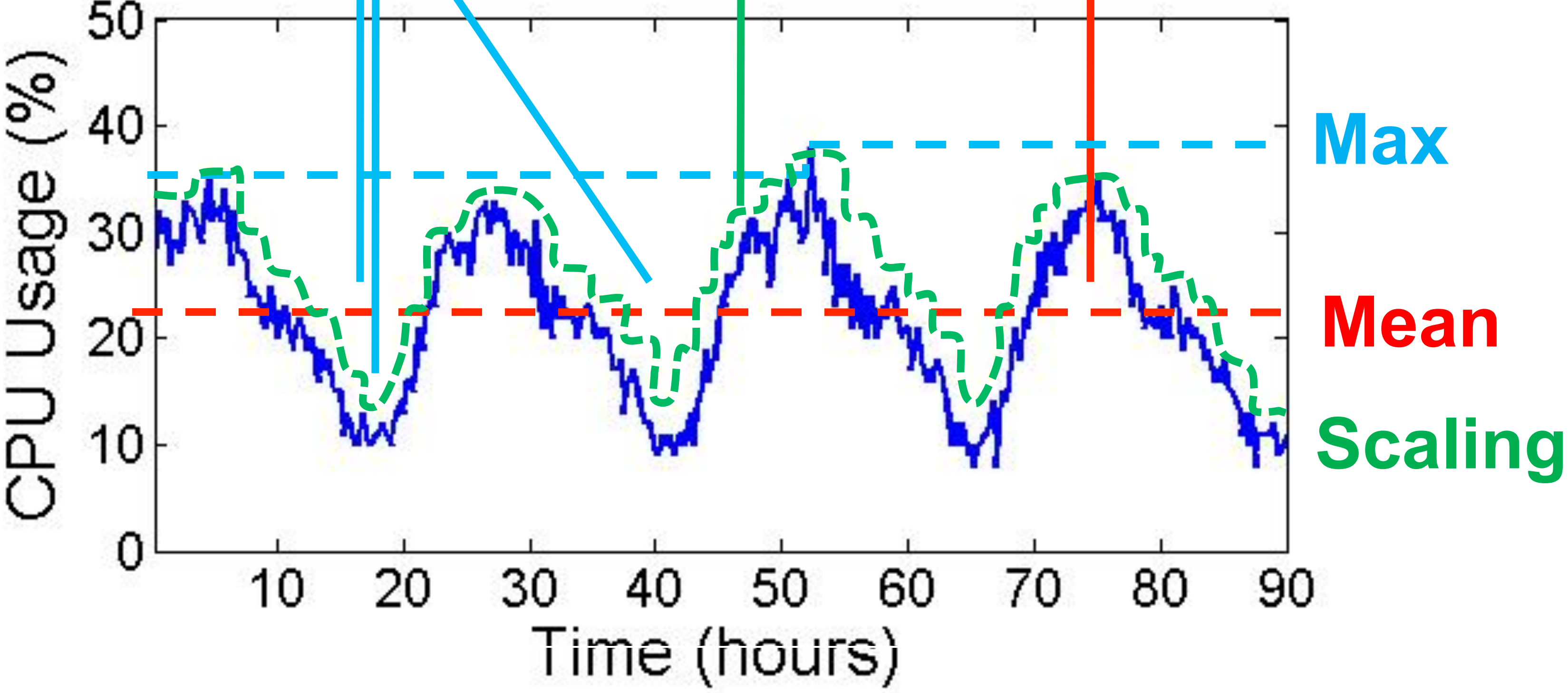
- RUBiS Web server driven by real workload trace:

Over-provisioning (waste)

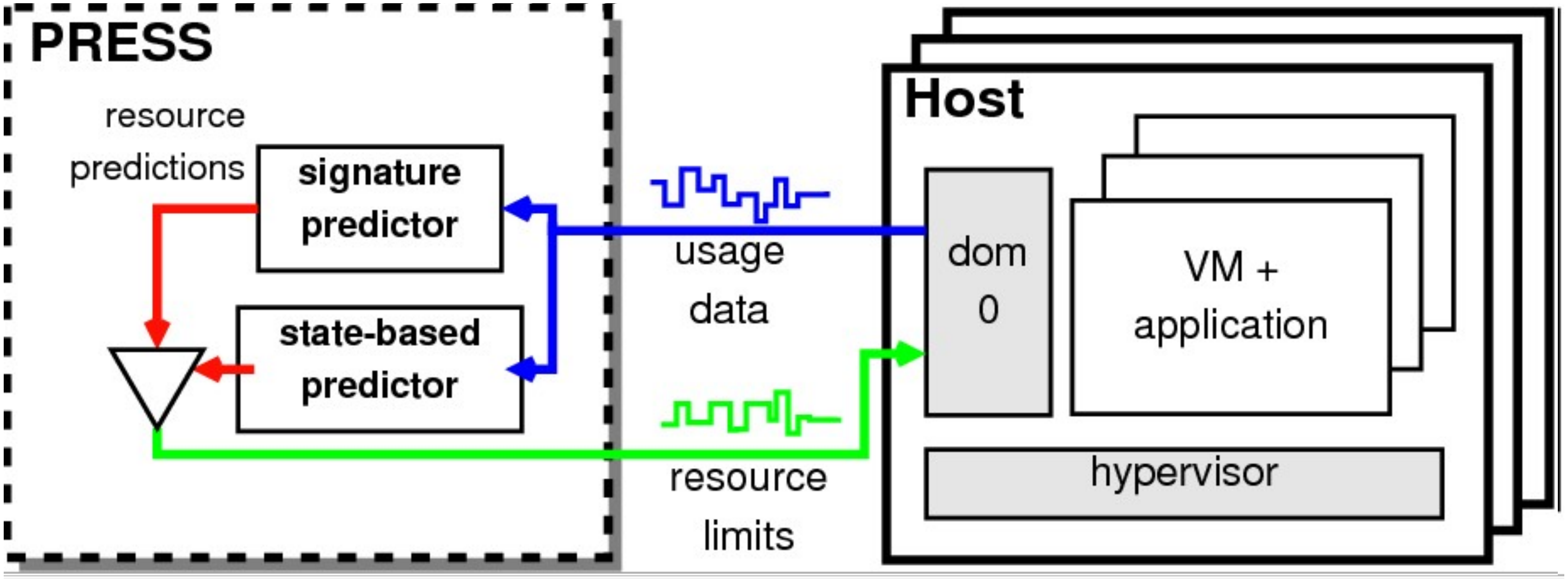
Over-provisioning (waste)

Under-provisioning (starve)

Always just enough resource



System Architecture

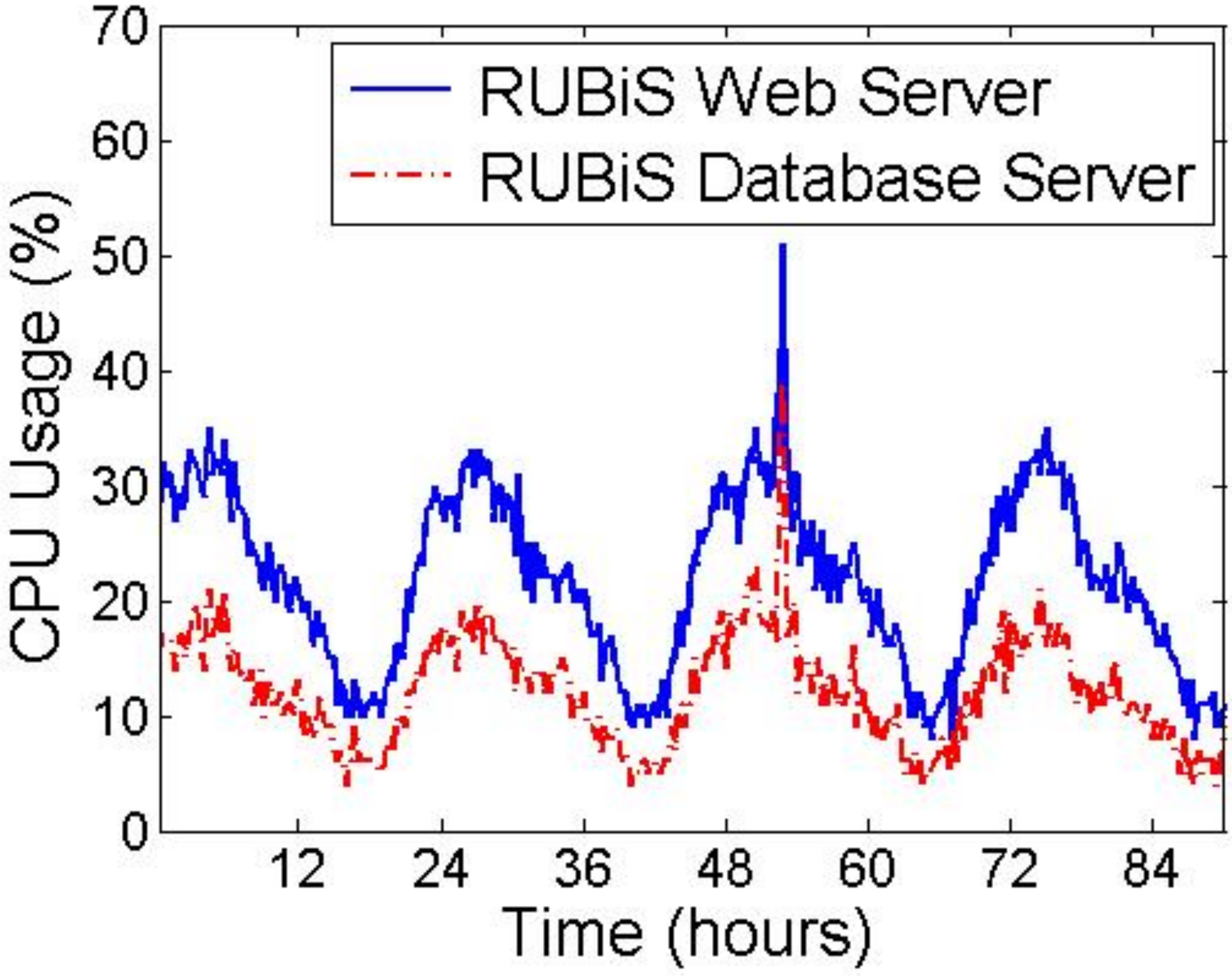


- Collect historical resource usage
- Online prediction for future usage
- Scaling: dynamically adjust resource caps based on prediction
- Black-box: no need for application knowledge

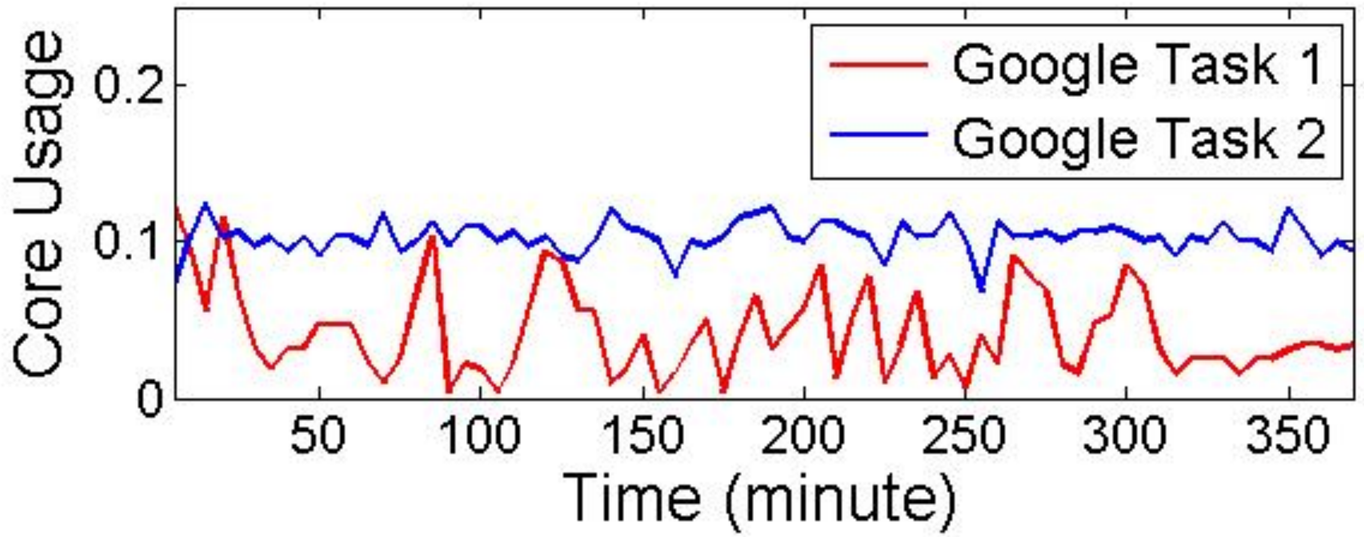
Research Challenges

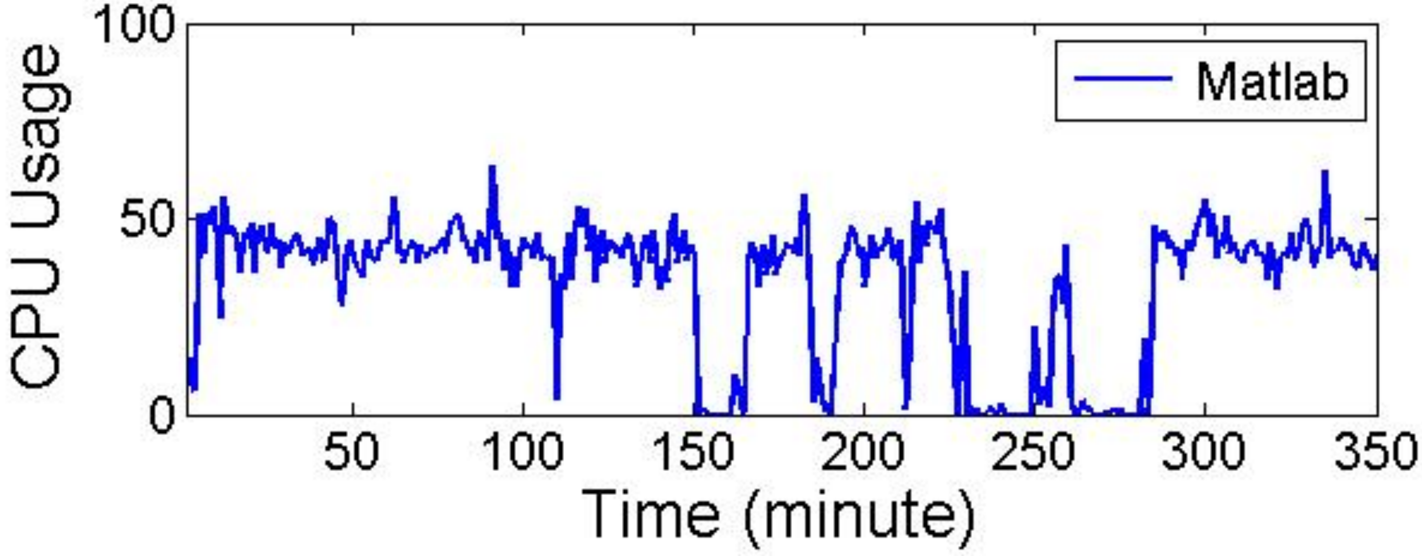
- We need prediction to do scaling
  - Fast reaction to dynamic workload changes
- A good prediction algorithm is hard
  - Resource demands are dynamic
  - Need to avoid both over- and under-estimation
  - Accommodate different workloads
  - Low overhead to manage large cloud system

Repeating Workload



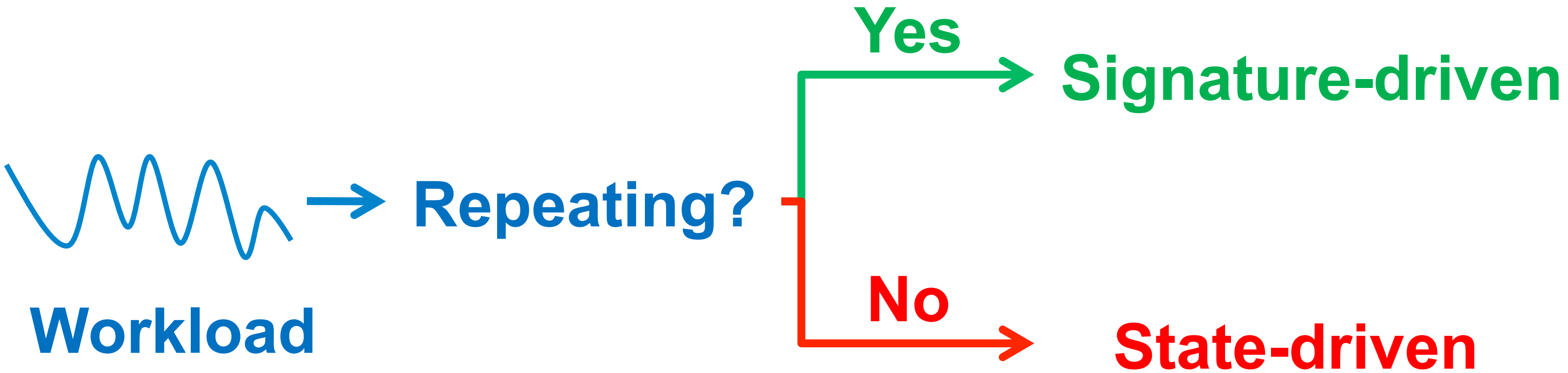
Non-Repeating Workload





Our approach: Hybrid Prediction Model

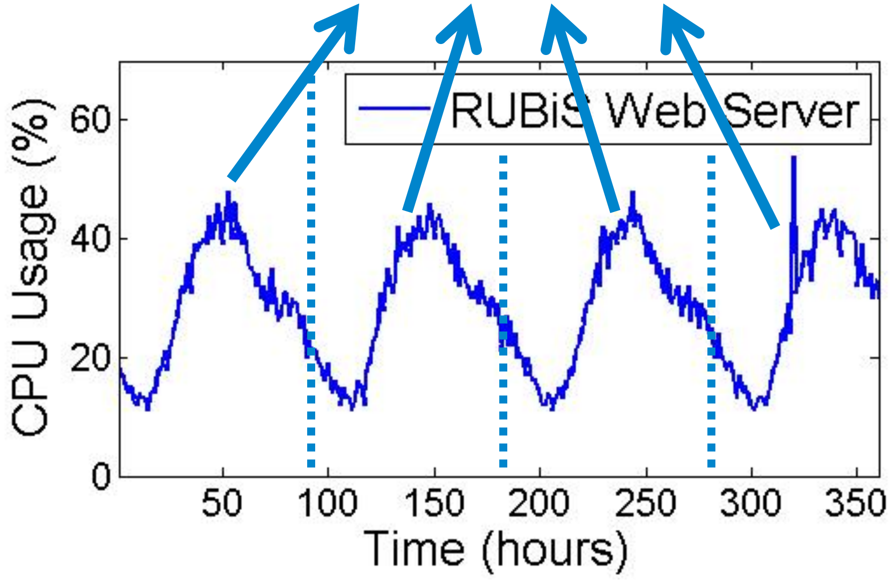
- For repeating workload: signature-driven
  - Use fast Fourier transform (FFT) to find repeating patterns for prediction
  - Lower overhead
- For non-repeating workload: state-driven
  - Use Markov chain model for prediction
  - Higher overhead, accommodate different workloads



Signature-Driven Approach

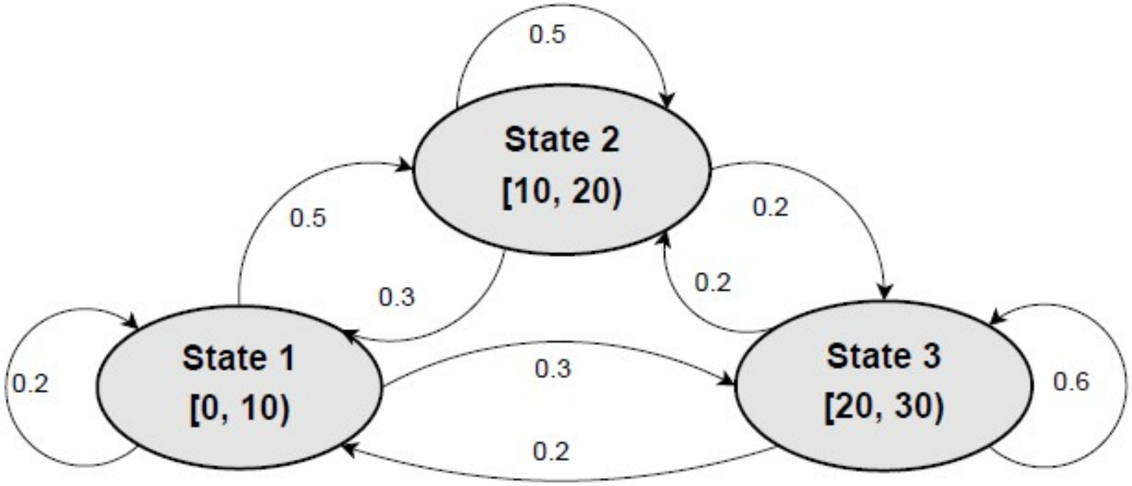
- Signature Extraction
  - Use fast Fourier transform (FFT) to discover dominant frequency
  - Use comprehensive difference check to validate repeating pattern
  - Signature alignment to predict next value

Similar? (mean value, correlation)



State-Driven Approach

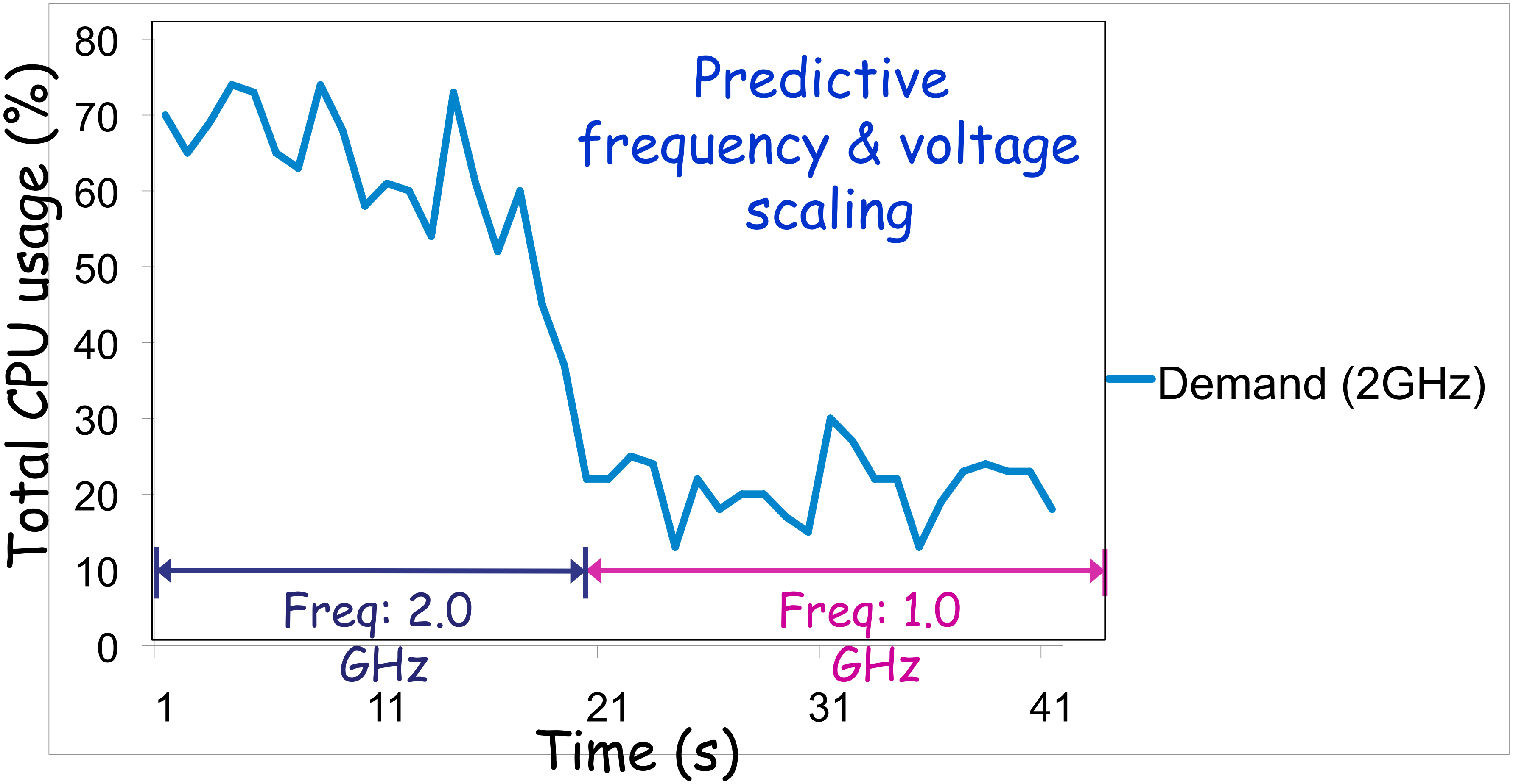
- Repeating patterns sometimes do not exist
- Uses discrete-time Markov chain to do short-term prediction
  - Dynamically learn transition probability matrix to predict future values



state transition matrix:

$$\begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$$

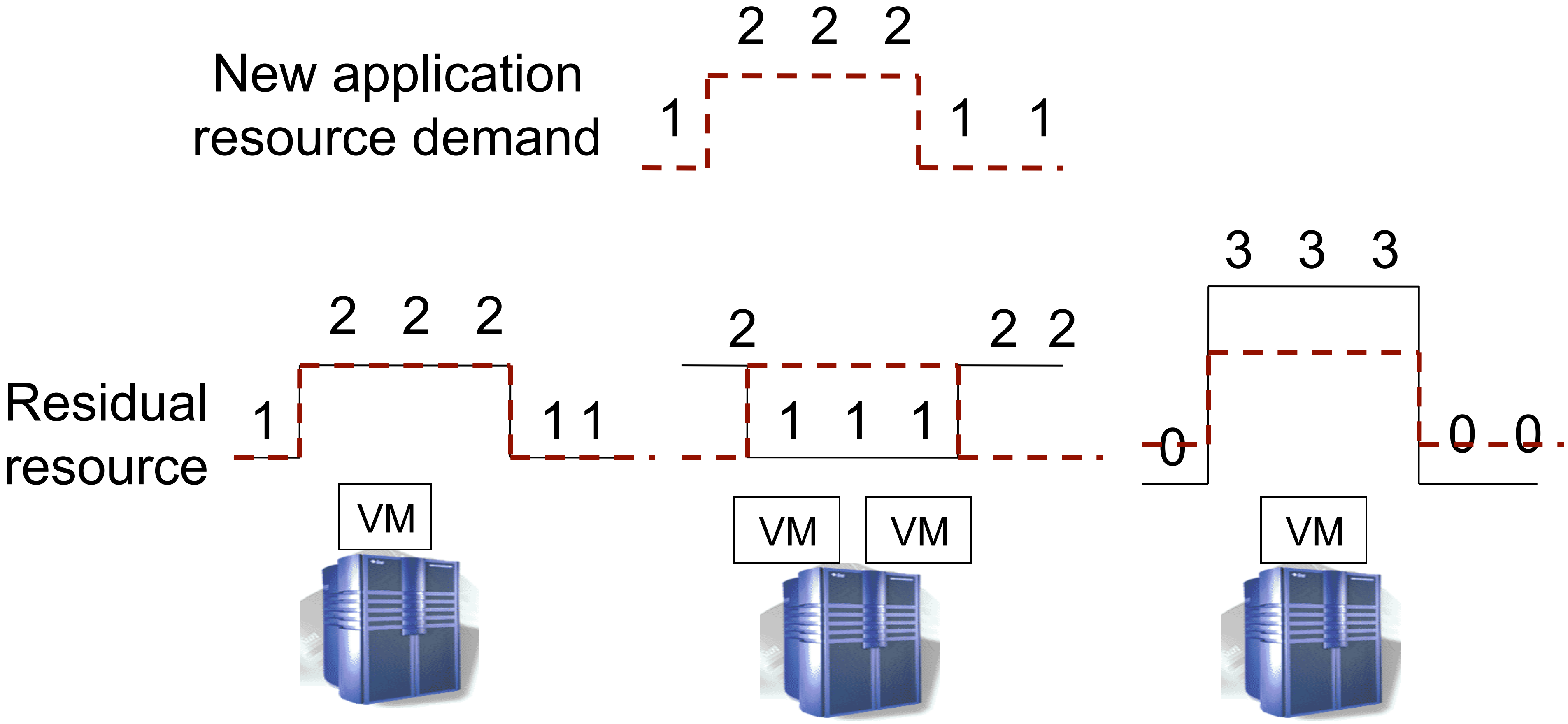
Energy Saving via Resource Scaling



Problem: Application Consolidation

- Application consolidation is needed for
  - Use fewest hosts to run all applications
  - Reduce resource provisioning cost
  - Reduce energy consumption
- Application consolidation needs to achieve
  - Meet dynamic resource requirements of different applications
  - Alleviate negative sharing impact
  - Balance resource utilization

A Case for Fine-Grained Patten Matching



New application resource demand

Residual resource

PAC: Pattern-driven Consolidation

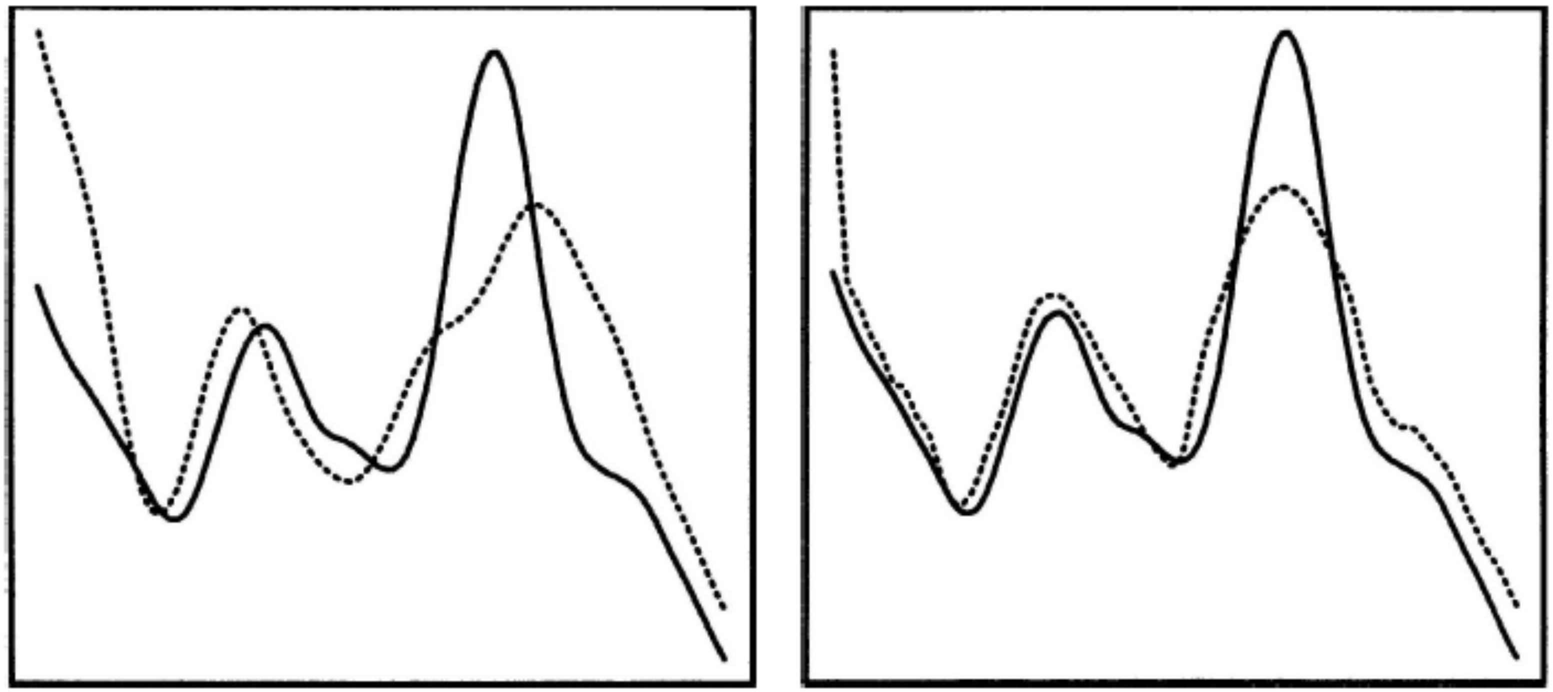
- Use time series to capture fine-grained signatures of dynamic applications
  - Dynamically discover repeating patterns using signal processing techniques

- Perform dynamic VM consolidation using signatures
  - Use time series indexing and dynamic time warping (DTW) to achieve fast and robust signature matching

PAC: Pattern-driven Consolidation

- Challenges
  - Fine-grained pattern matching in distributed computing environments
  - Scalable application consolidation
  - Pattern heterogeneity
- Solutions
  - Use dynamic time warping (DTW) to achieve robust matching
  - Use time series indexing to achieve fast signature matching
  - Pattern alignment

Dynamic Time Warping (DTW)



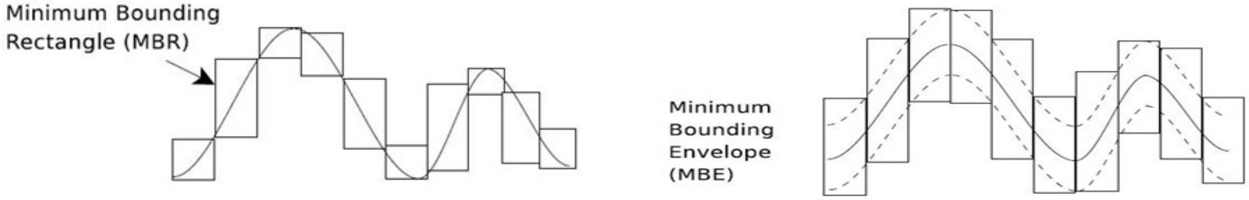
(a) Two different time-series that are out of phase by little

(b) After alignment of the time-series by dynamic time warping

DTW Phase alignment example

Signature Indexing for Pre-Filtering

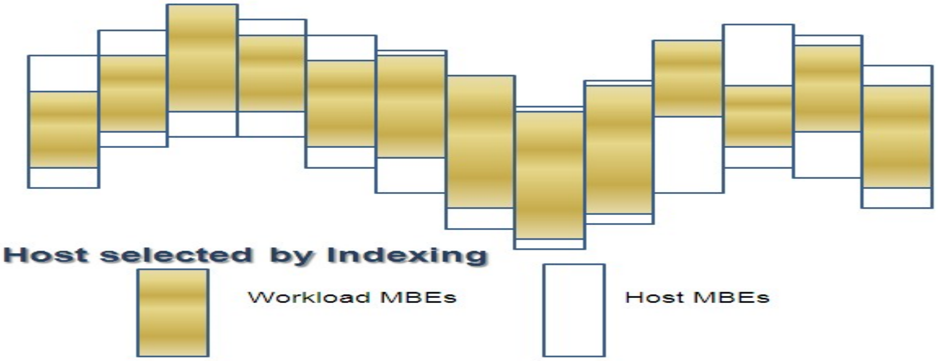
- One-dimensional case for simplicity:
  - Split the time series to MBRs expand to be MBE



(a) Signature indexing using MBR.

(b) Signature Pre-filtering.

  - Matching lower-bound of the load MBE is lower than the upper-bound of the host resource MBE



  - Admission Control: Pass if more then 65% matching in all MBEs (adjustable threshold)

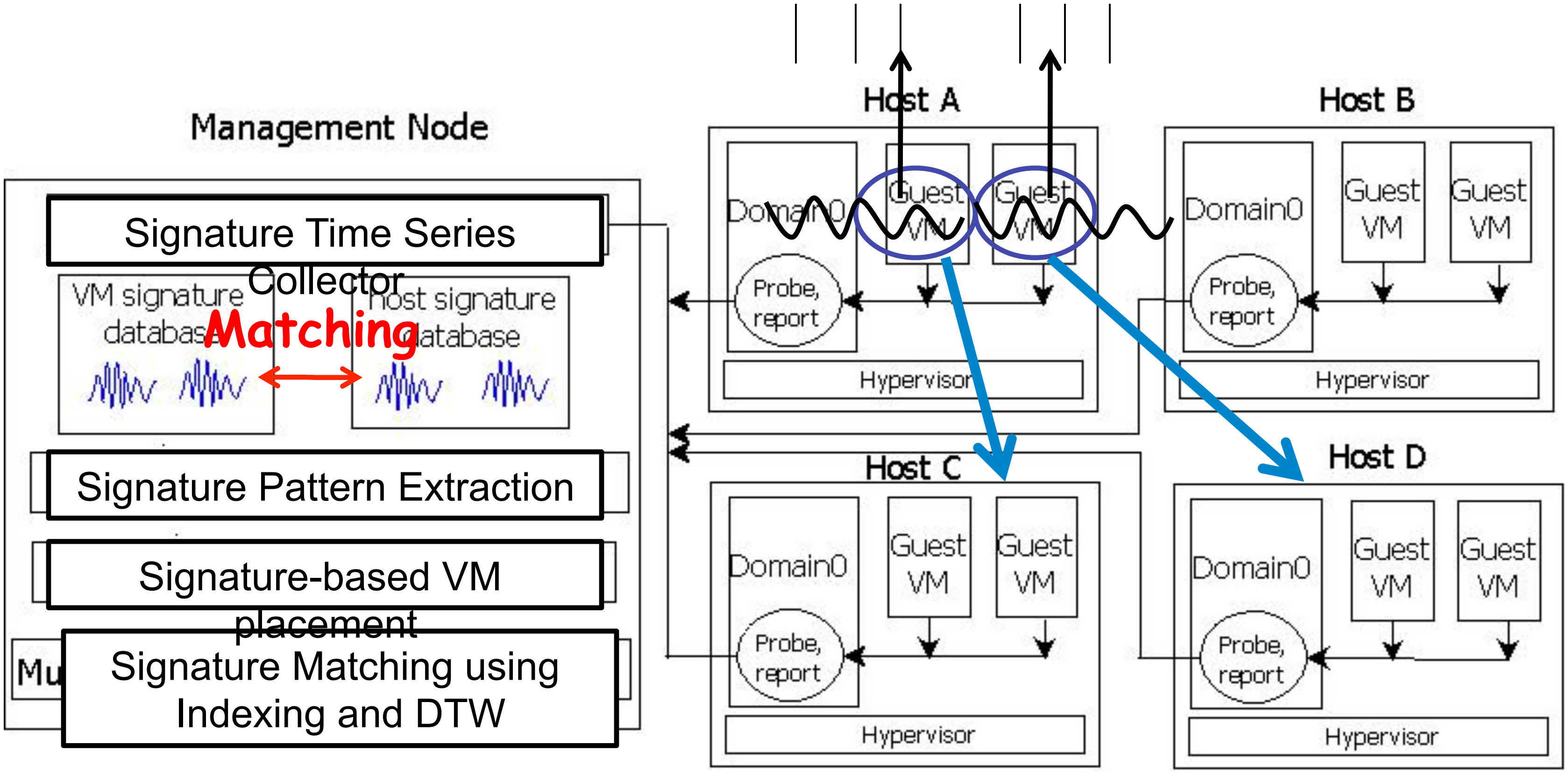
Pattern-Driven Consolidation

Collect VM resource usage in the recent past

Extract signature patterns and predict for future

Use indexing and DTW to match VMs and hosts

Perform VM consolidation based on matching result



Signature Time Series Collector

Signature Pattern Extraction

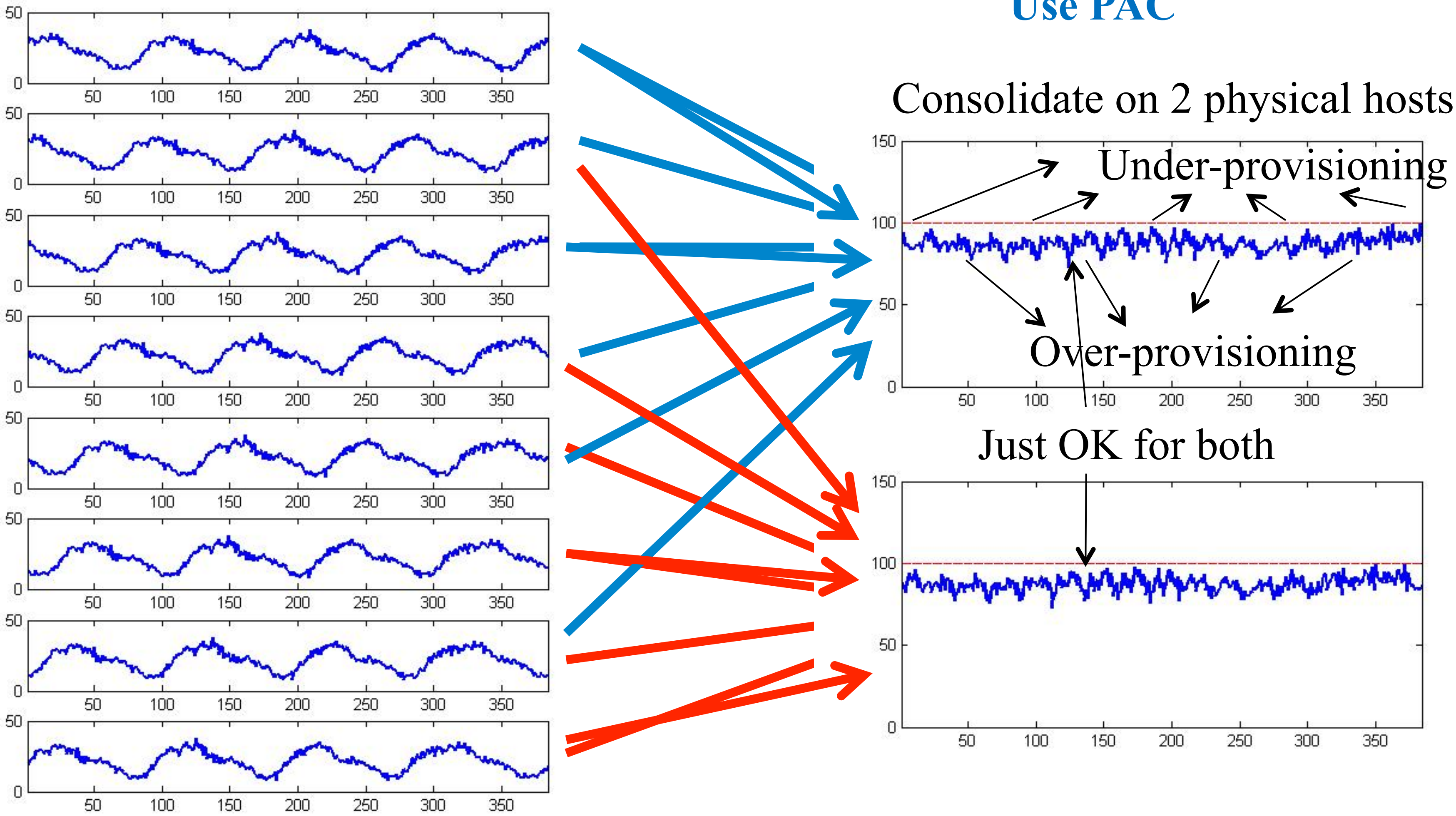
Signature-based VM placement

Signature Matching using Indexing and DTW

Consolidation example

Resource usage for 8 VMs

Use coarse-grained approach
Use PAC



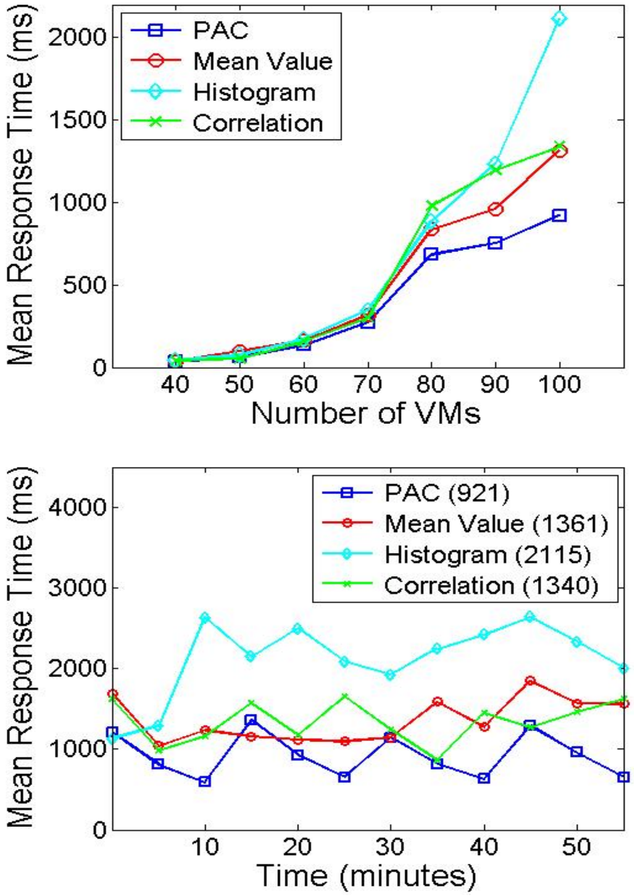
Consolidate on 2 physical hosts

Under-provisioning

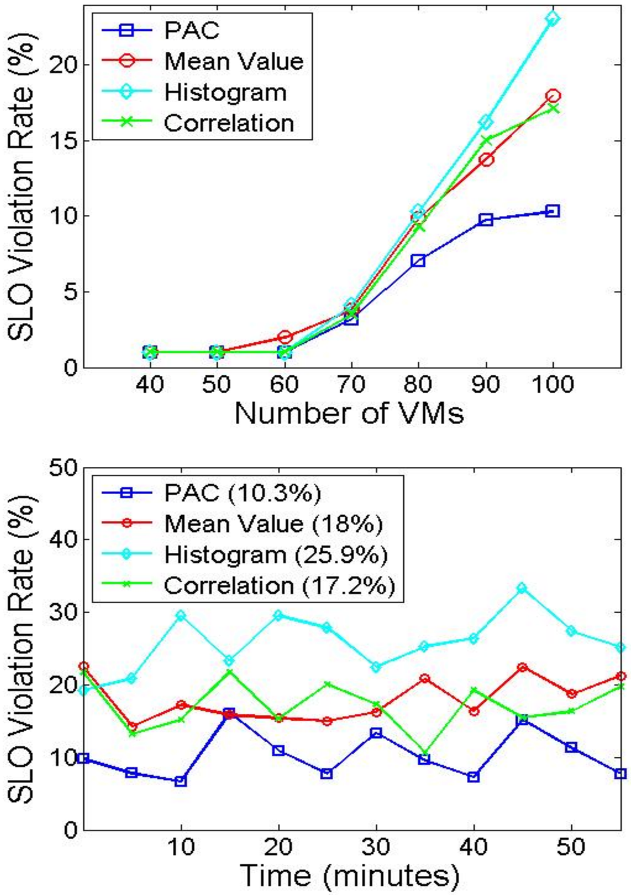
Over-provisioning

Just OK for both

System Consolidation Performance



Response time comparison



SLO violation comparison