institute imdea software

Is Machine Learning Necessary for Cloud Resource Usage Forecasting?

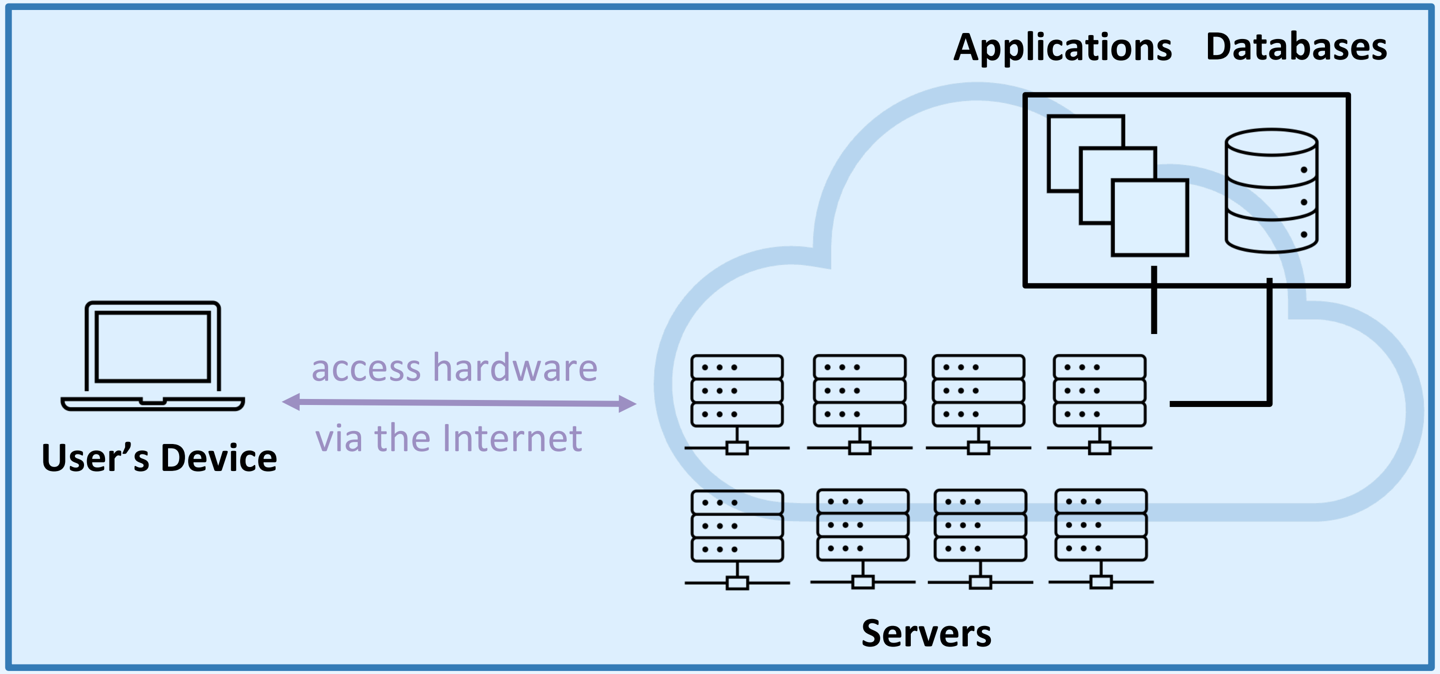
Georgia Christofidi

Paper Presented at the Symposium of Cloud Computing (SoCC ‘23)

@ S3, November 14th

What is Cloud Computing?

Cloud providers.



The cloud.

Cloud servers are located in **data centres.**

Computing and **storage** takes place on **servers.**

Businesses and single users → cloud computing →

1. Reduce IT costs

↓ Buy and manage physical servers.

2. Easier to operate internationally

✓ Files and Applications can be accessed from different devices.

Basic Concepts of Cloud Computing

1. Workload

An **Application** performing a specific task

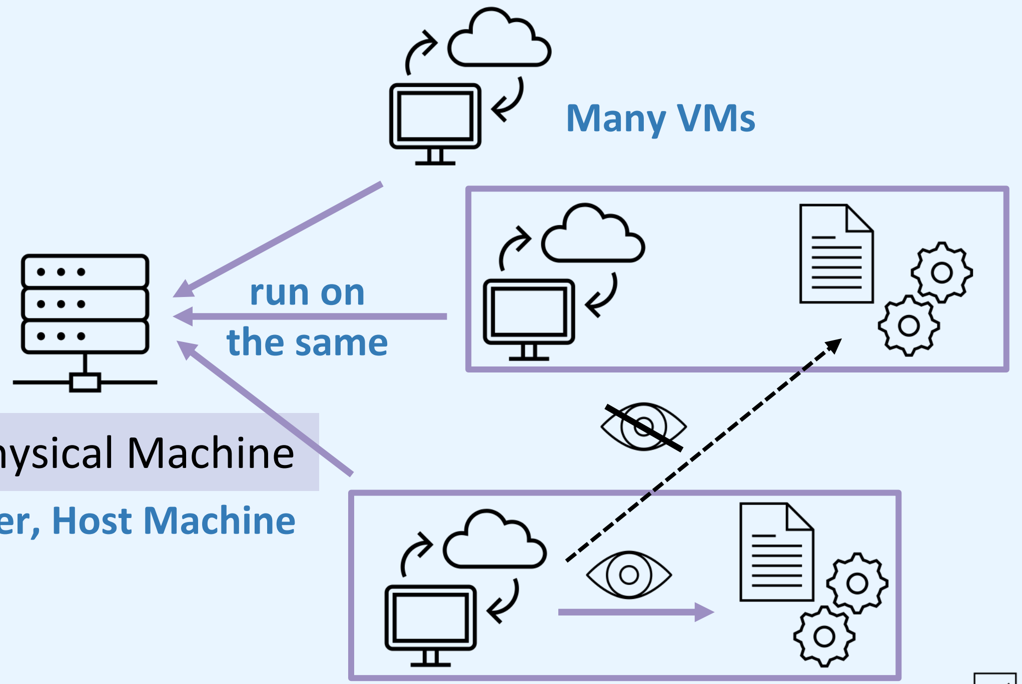
that uses

a specific amount of resources (processing, storage, network etc)

2. Virtual Machine (VM)

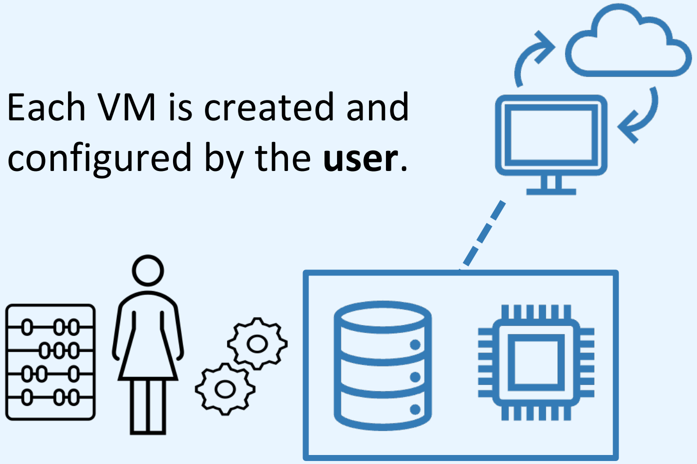
digital-only computer

behaves as a **physical computer** with its own hardware



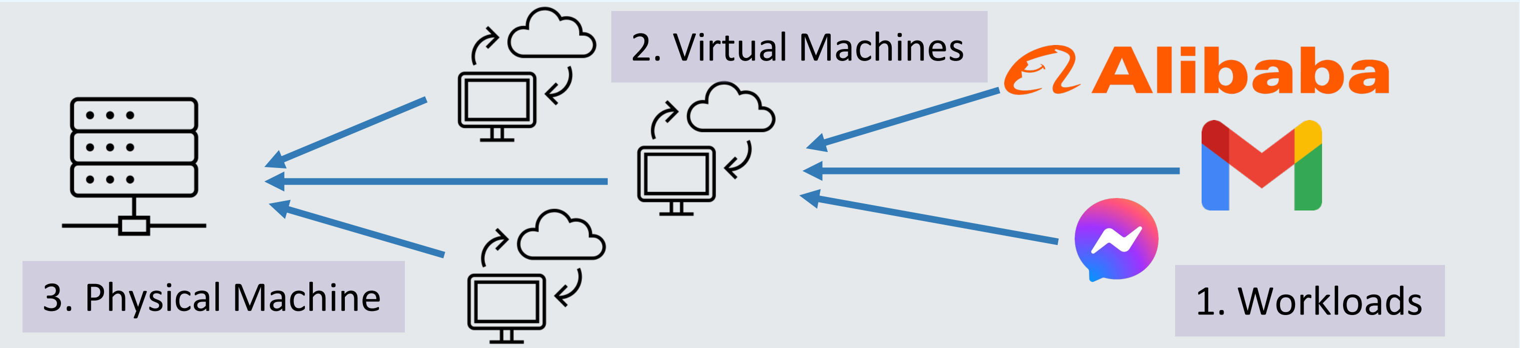
VMs do **not interact** with each other.

Each VM is created and configured by the **user**.



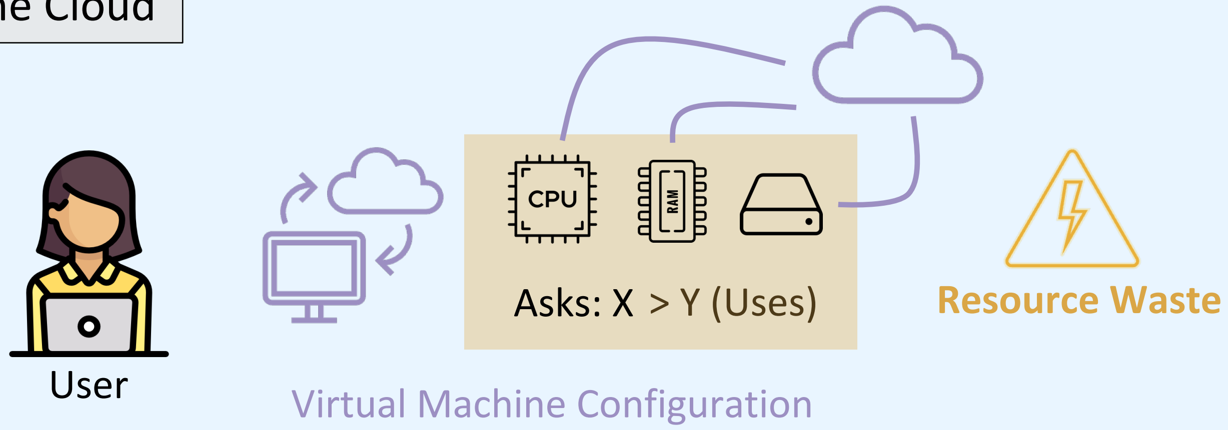
- ☑ Serve more customers at once
- ☑ Low cost – High hardware efficiency

Levels at which we can observe Resource Usage:



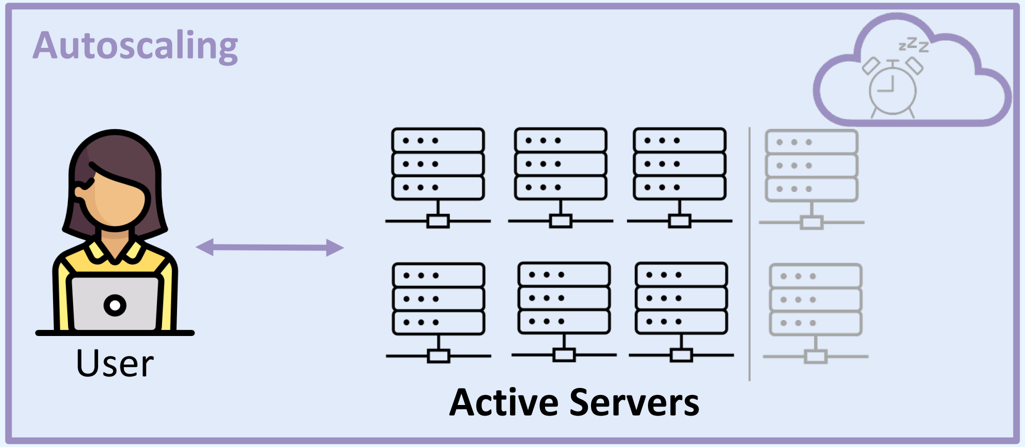
The Problem of Cloud Resource Usage Forecasting

Low resource efficiency in the Cloud



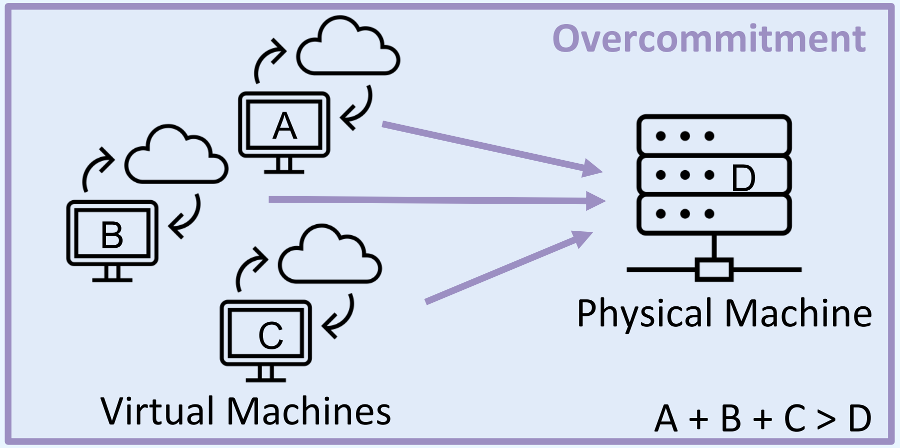
Cloud resource management systems

Depending on the **load** and the **user needs**



Dynamic adjustment of the number of computational resources e.g., active servers

Resource shared between VMs



Sum of resources **allocated** on all the VMs **>** resources **available** on physical servers

The Problem of Cloud Resource Usage Forecasting

Approach: Future Resource Usage Forecasting

Input: Past Resource Usage
$x_1, x_2, \ldots, x_n$

Forecasting Models
(ML, Statistical, Heuristic, Hybrid)

Output: Future Resource Usage
$X_{n+1}, x_{n+2}, \ldots, x_{n+k}$

Challenge: Achieving High Accuracy in Forecasting

1. ↑ Resource Efficiency

2. ↓ Costs

3. ↑ Energy Efficiency

4. ↑ Application Performance

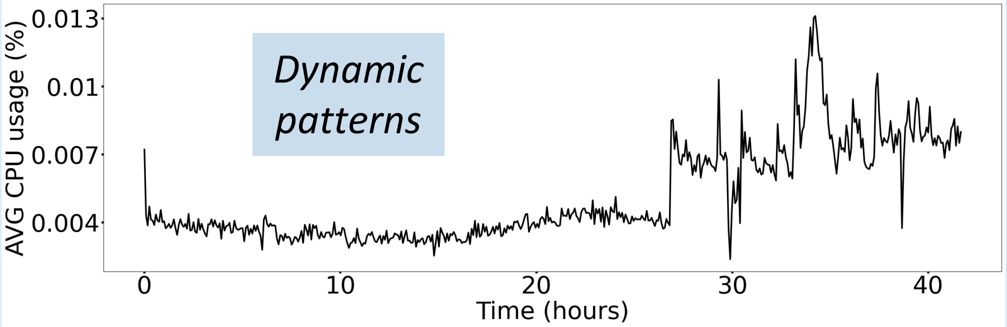
↑ Meeting Service Level Agreements
↑ User Experience

↓ Service Interruptions
↓ Response time

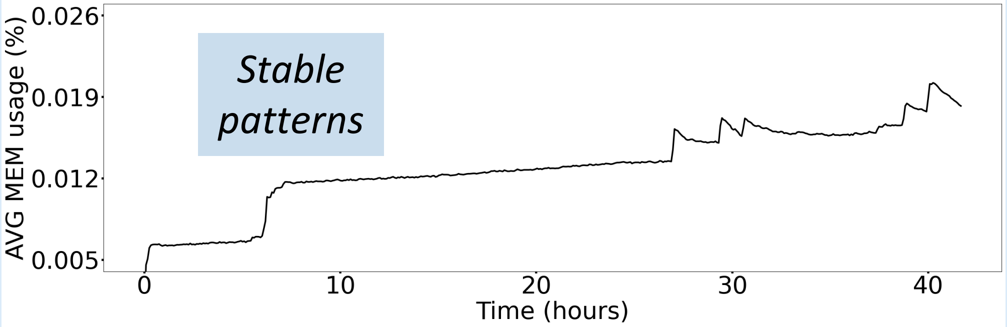
The Patterns of Cloud Resource Usage

Workload level

Average CPU usage

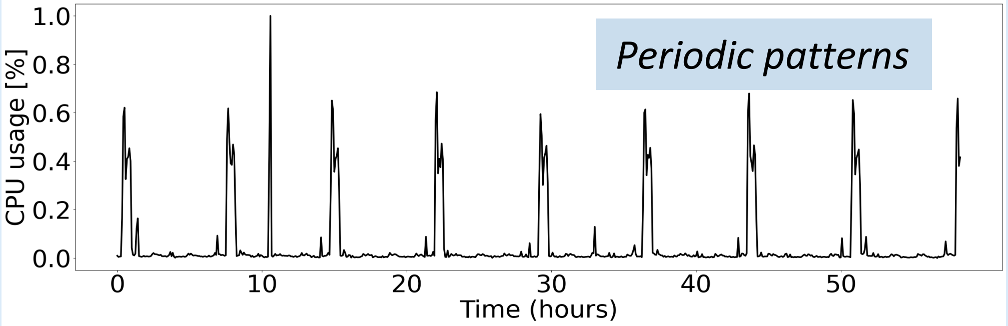


Average memory usage

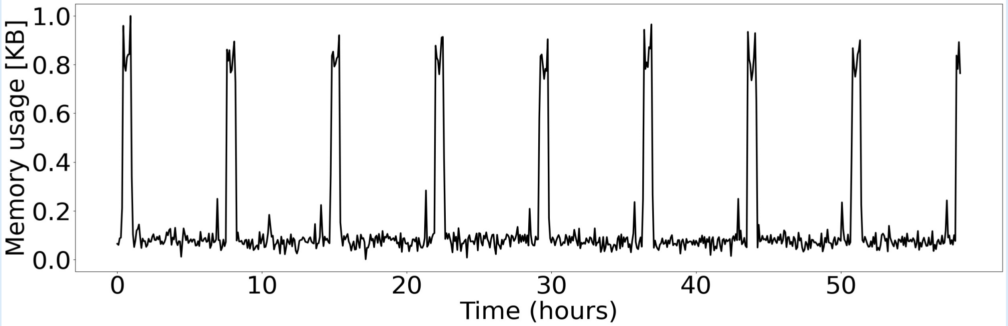


Virtual Machine level

CPU usage



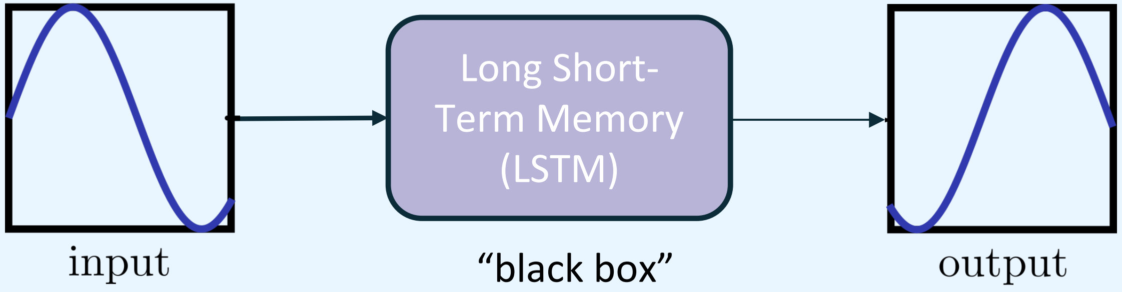
Average memory usage



Takeaway: Patterns differ across different types of resources and levels of use (Workload vs VM).

Do we need ML to **accurately predict all** of the different patterns?

Forecasting with Machine Learning



High accuracy when predicting:

Weather

Stock Market Prices

Power Consumption

Traffic Conditions

LSTMs for **Cloud** Resource Usage Forecasting

"BHyPreC: A Novel Bi-**LSTM** Based Hybrid Recurrent Neural Network Model to Predict the CPU Workload of Cloud Virtual Machine"
IEEE Access, 2021

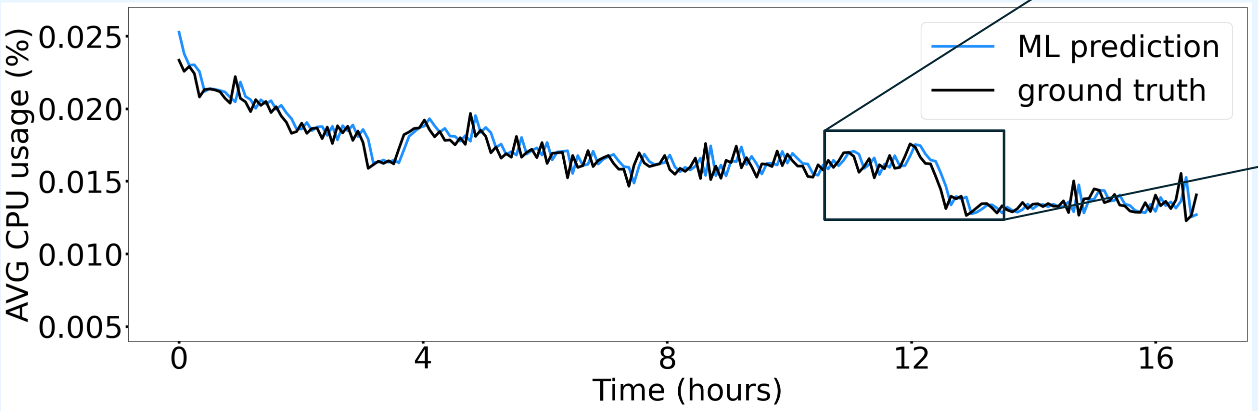
Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems
"We used **LSTM** for time series forecasting."
EuroSys, 2023

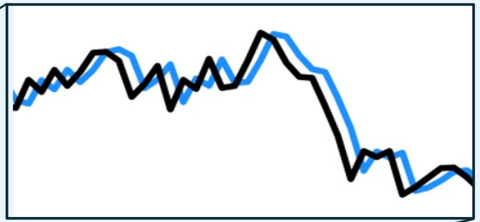
Seer: Leveraging Big Data to Navigate the Complexity of Performance Debugging in Cloud Microservices
"The **LSTM** is especially effective at capturing load patterns over time."
ASPLOS, 2019

"Large-scale computing systems workload prediction using parallel improved **LSTM** neural network"
IEEE Access, 2021

Debunking the High Accuracy of LSTMs

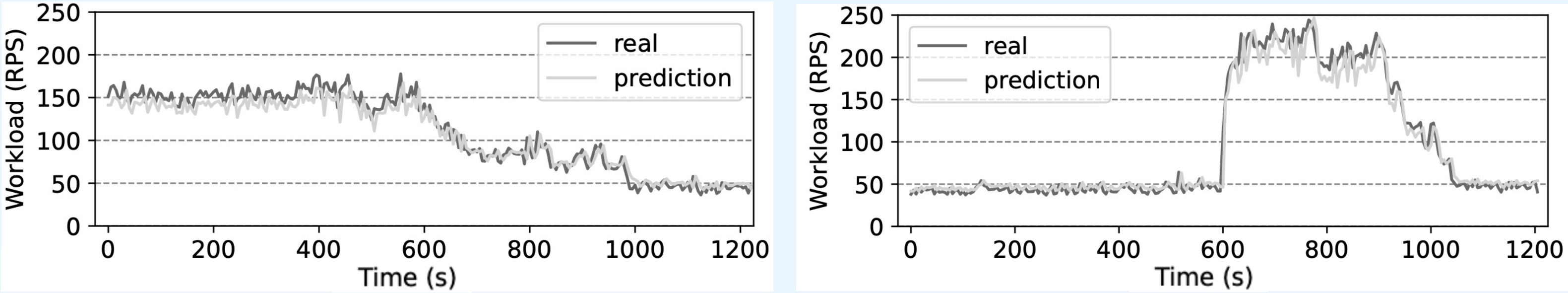
Usecase: Cloud Workloads.





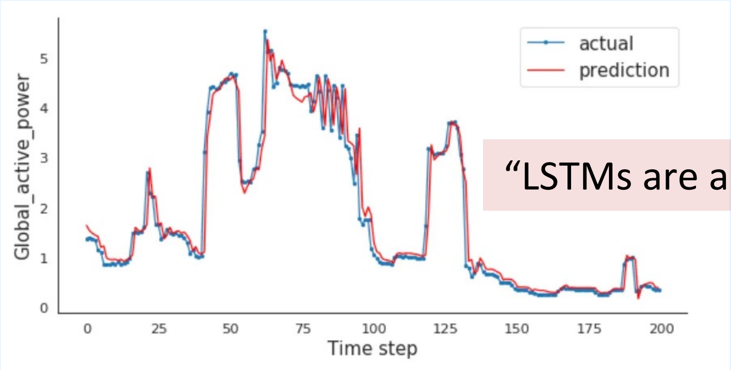
Our Insight: LSTM predictions resemble the **previous** timestep of the timeseries.

Usecase: ML Inference Services.



Source: Figures 5 & 8 from paper “Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems” published at EuroMLSys 2023. Twitter trace workload.

Usecase: Global Active Power Consumption



“LSTMs are amazing!”

Source: Figure 12 from blog post “Time Series Analysis, Visualization & Forecasting with LSTM” on https://towardsdatascience.com

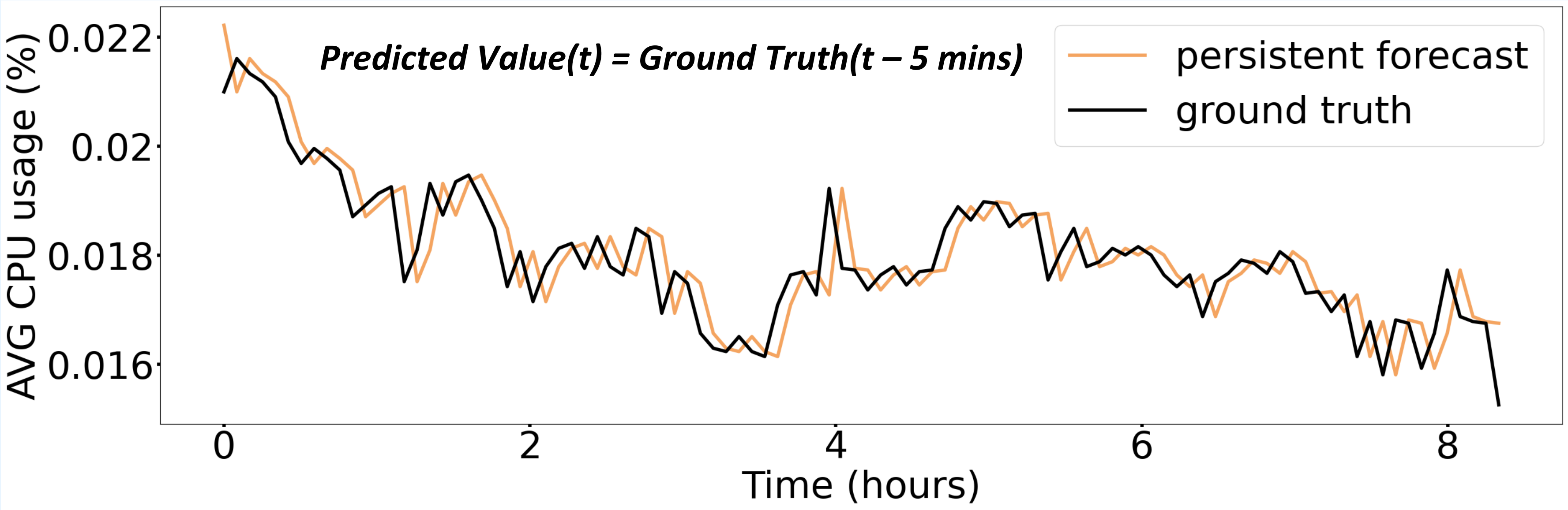
Do we need ML to produce such “shifted” predictions?

Our Approach: Persistent Forecast

Let's do something **simple**!

For each timestep t in the timeseries, the prediction is the value at the **previous** timestep.

We call this the **Persistent Forecast**.



The prediction (Persistent Forecast) is a shifted version of the ground truth.

Simple, Lightweight
Application agnostic
No overheads

Prediction Accuracy

Experimental Methodology

Extensive experimental evaluation with cloud resource usage data.

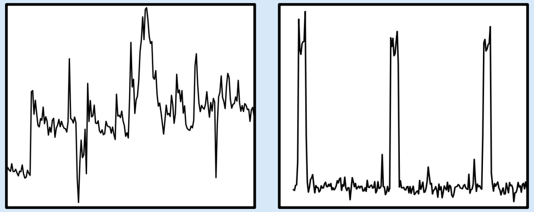
Public open-source datasets across different:

Cloud providers
Alibaba Cloud
Google Cloud
Microsoft Azure
Bitbrain

Resource Types
CPU
RAM

Resource Levels
Physical Machine
Virtual Machine
Workload

Usage patterns



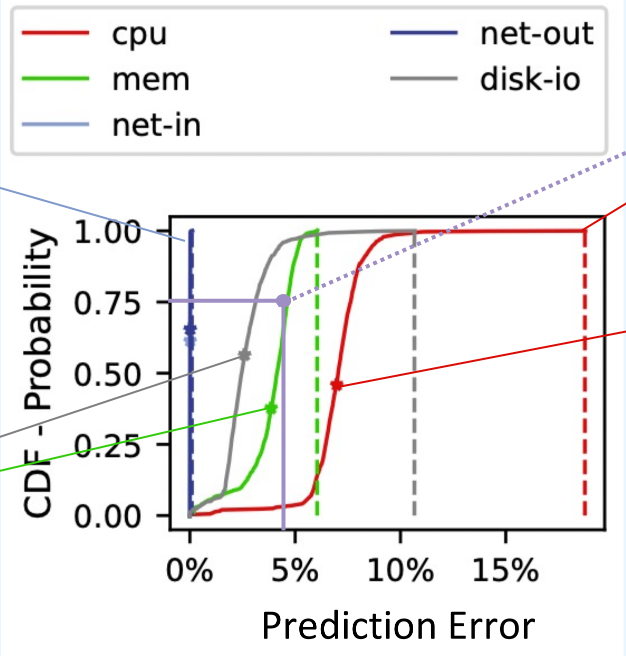
Frequency
Hourly/ Daily/
Weekly Windows

We calculate the **prediction error** of the persistent forecast.

Experimental Results – Physical Machine Level

Physical Machine

Alibaba Dataset
Physical Machine Level



higher is better

lower is better

NET-IN & NET-OUT: Negligible Average and Maximum Error Values

DISK-IO & MEM: Average Error < 4%

The probability of the error being equal or less than **4%** is **75%**.

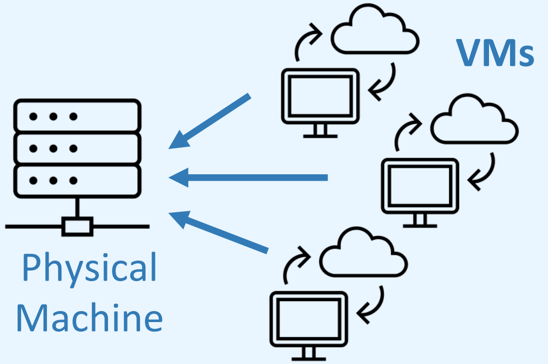
CPU: has the largest tail

CPU: 6.97% on average (more dynamic patterns)

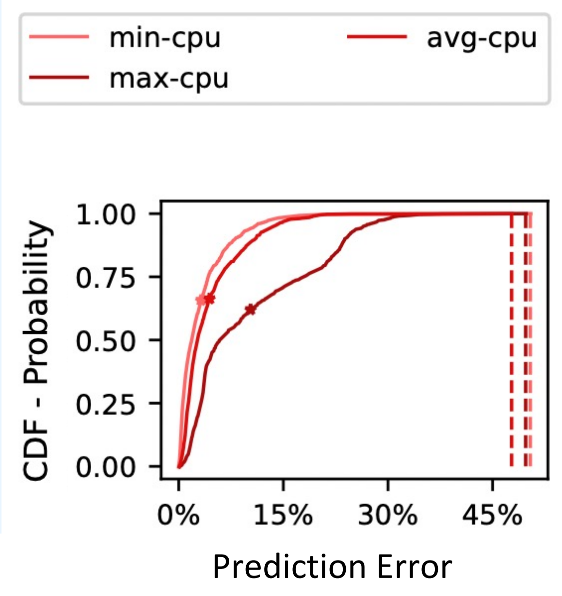
We want **high** probability of **low** errors.

Takeaways: The Persistent Forecast is **highly accurate**, across resource types, levels of use and measurements, *because* cloud resource usage values **persist** over time.

Experimental Results -Virtual Machine Level

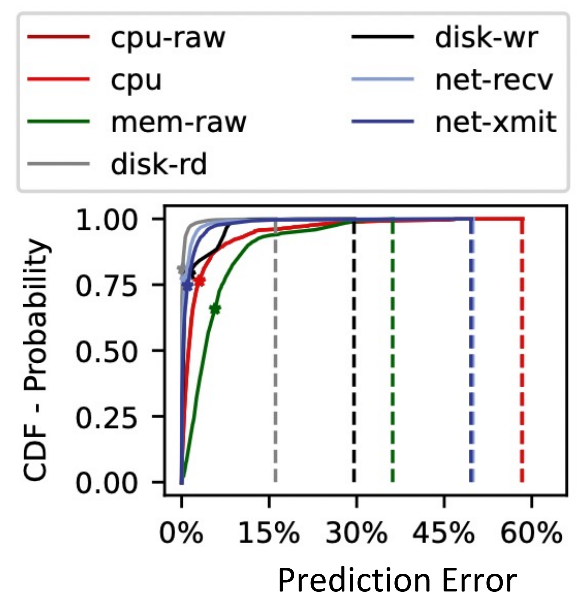


Azure Dataset
Virtual Machine Level



Average Prediction Error < 10%

Bitbrains Dataset
Virtual Machine Level



Average Prediction Error < 6 %

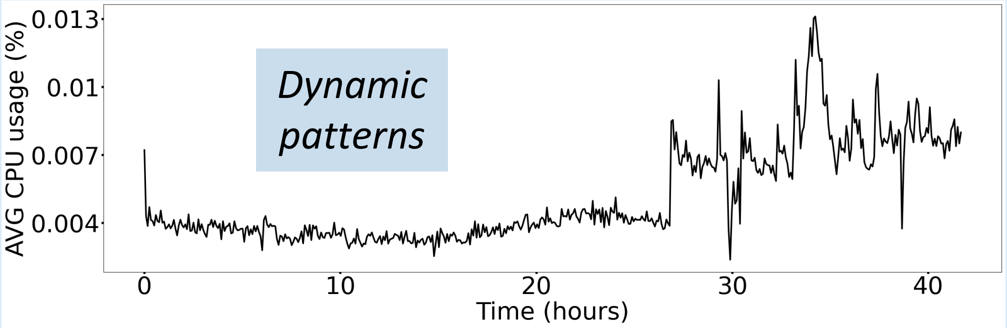
Resource Type	Average Error	Raw Error
CPU usage MHZ	3.05%	83.64 MHZ
Memory Usage KB	5.73%	129.63 MB
Disk reads KB/sec	0.21%	57.60 KB/sec
Disk writes KB/sec	1.41%	36.78 KB/sec
Network in KB/sec	0.66%	29.76 KB/sec
Network out KB/sec	1.03%	26.62 KB/sec

Takeaways: The Persistent Forecast gives **very low average error values** on the virtual machine level, less than 10%. The tail gets larger, because **patterns** become more **dynamic**, as we mesure resource usage on a deeper level.

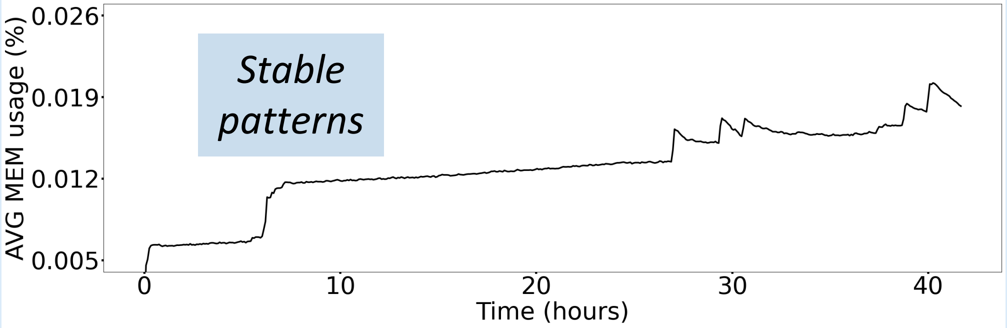
Experimental Results –Workload Level

Workload level

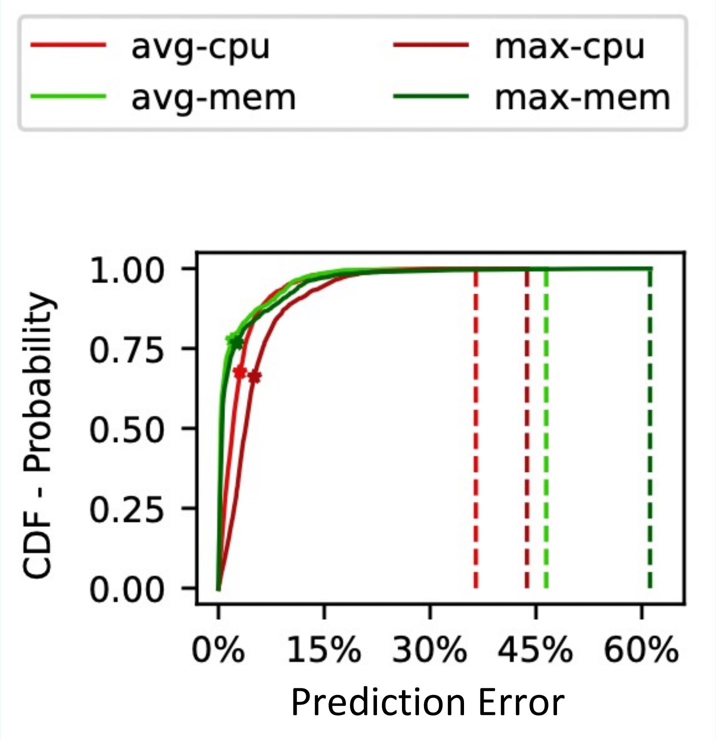
Average CPU usage



Average memory usage



Google Dataset
Workload Level



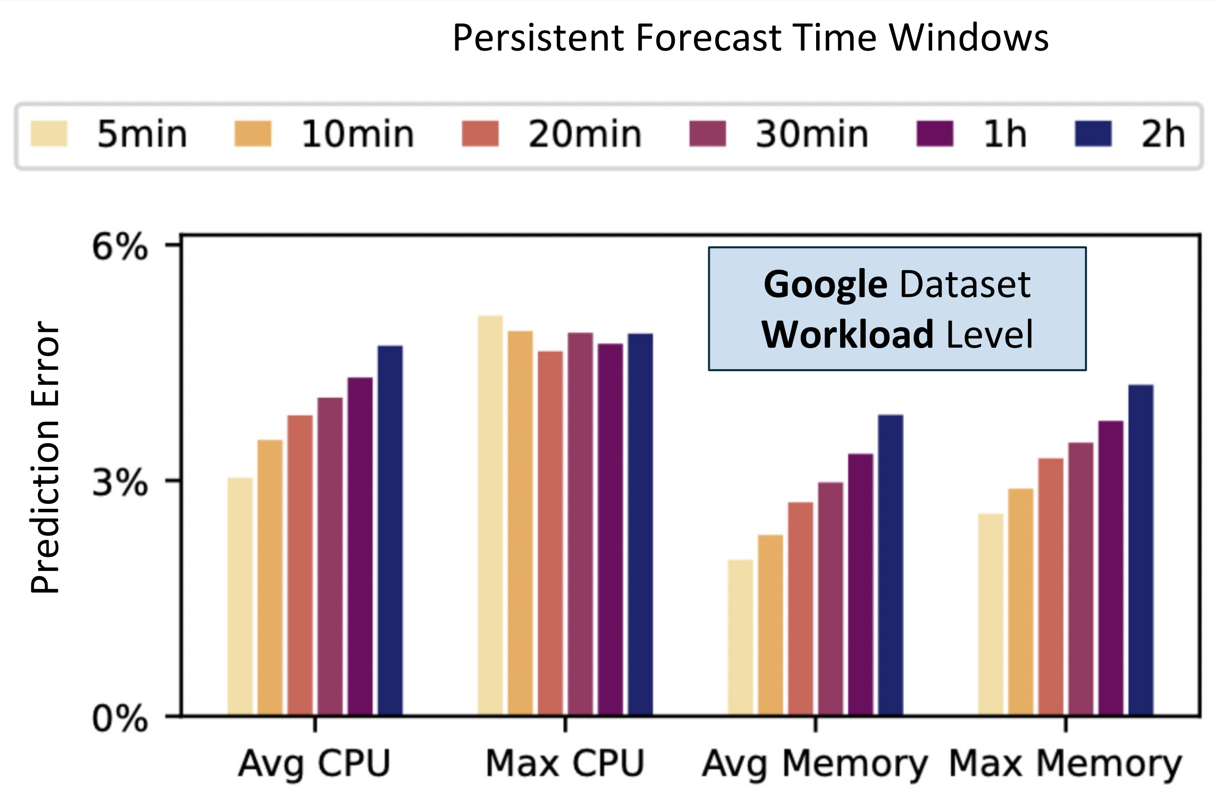
Average Prediction
Error Values < 6%

Takeaways: At the workload level, patterns become even more dynamic. CPU usage has larger prediction error values than memory usage.

Sensitivity on the length of the time window

Persistent forecast time window = 5 minutes
Predicted Value(t) = Ground Truth(t – 5 mins)

What happens when we increase the time window?
Predicted Value(t) = Ground Truth(t – time_window)



Takeways: Low sensitivity to length of the time window.
This validates that the values **persist** over time and reveals potential **repeating patterns** in the data.
This unlocks an **opportunity** for lower prediction error values, if the time window matches the data periodicity.

Lessons Learned

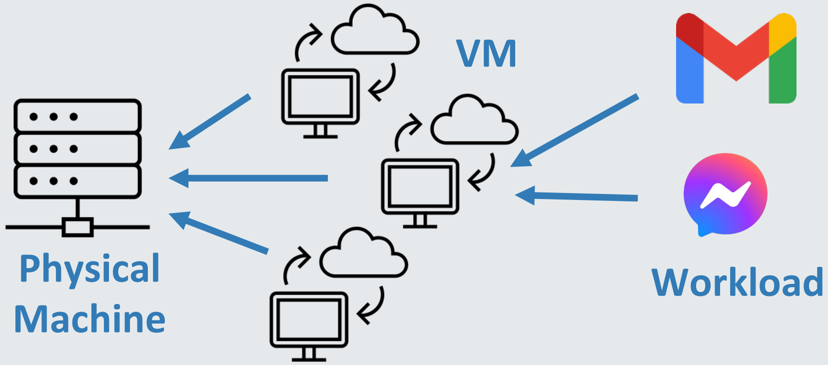
Prediction Error
of Persistent Forecast

Resource Measurement

Prediction Error of
MAX > AVG > MIN

depends on

Resource Level



More stable and periodic utilization levels

More dynamic resource usage

Resource Type

- Average CPU usage shows periodic and daily patterns.
- Memory levels more stable over time.

Is Machine Learning Necessary for Cloud Resource Usage Forecasting?

Scan for code & paper:

No.

(for the most part)

Open questions

1. When to use ML?

- exact use case
- data pattern

predictions ⇅ system's performance and decision-making

2. Which ML method to use, *when necessary*?

Probably not LSTMs

Other state-of-the-art ML methods for timeseries forecasting

Suggestions

1. Revisit existing systems and study the **data patterns**.

Values persist over time?

Try the **Persistent Forecast**

2. Insightful and **judicious** use of ML, simple mechanisms to the extent possible.