Is Machine Learning Necessary for Cloud Resource Usage Forecasting?

Vision Paper

Georgia Christofidi, Konstantinos Papaioannou, Thaleia Dimitra Doudali

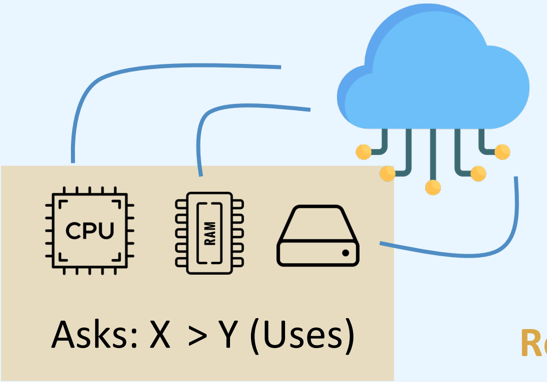
@ SoCC23, October 30th

institute imdea software

The Problem of Cloud Resource Usage Forecasting

Challenge: Low resource efficiency in the Cloud

User



Resource Waste

Solution: Future Resource Usage Forecasting

Input: Past Resource Usage
$x_1, x_2, ..., x_n$

Forecasting Models
(ML, Statistical, Heuristic, Hybrid)

Output: Future Resource Usage
$x_{n+1}, x_{n+2}, ..., x_{n+k}$

Problem: Achieving High Accuracy in Forecasting

1. ↑ Resource Efficiency

2. ↓ Costs

3. ↑ Energy Efficiency

4. ↑ Application Performance

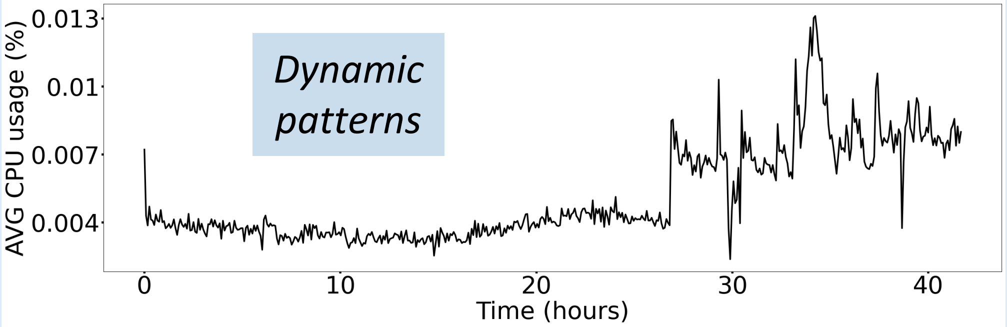
↑ Meeting Service Level Agreements
↑ User Experience

↓ Service Interruptions
↓ Response time

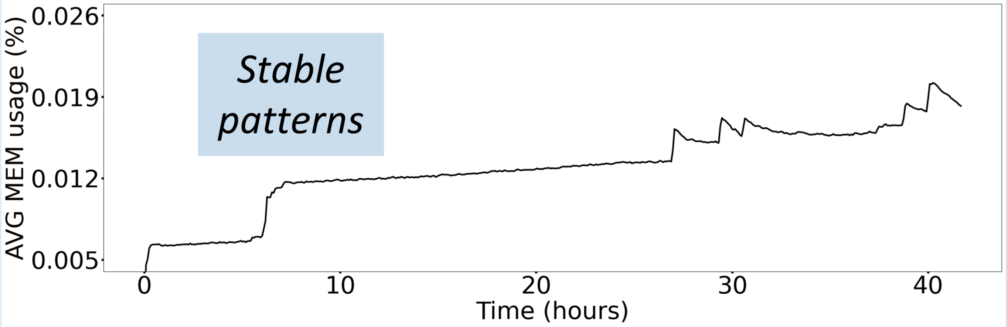
The Patterns of Cloud Resource Usage

Workload level

Average CPU usage

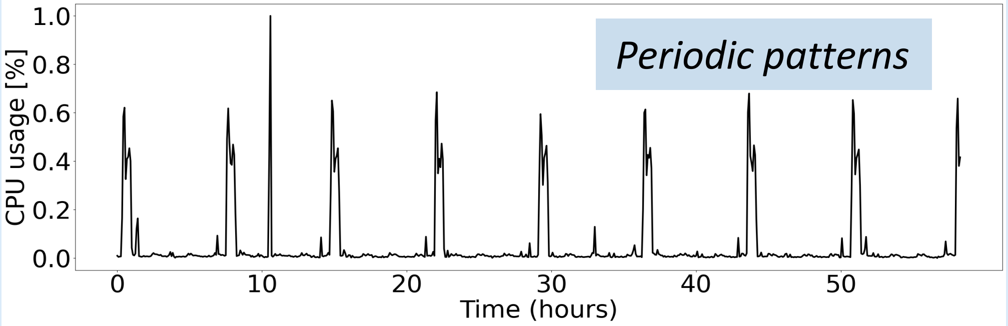


Average memory usage

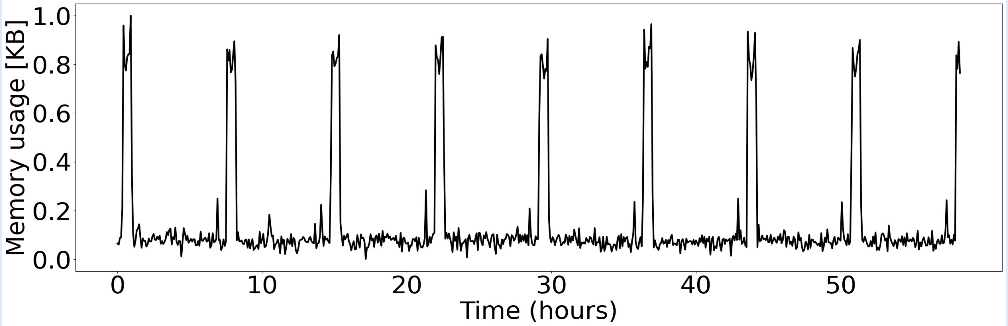


Virtual Machine level

CPU usage



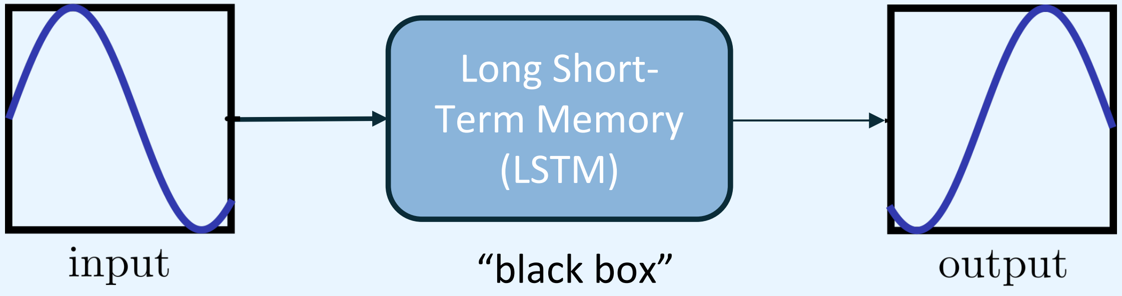
Average memory usage



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

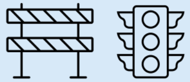


High accuracy when predicting:


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

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

Reconciling High Accuracy, Cost-Efficiency, and Low Latency of Inference Serving Systems

EuroSys, 2023

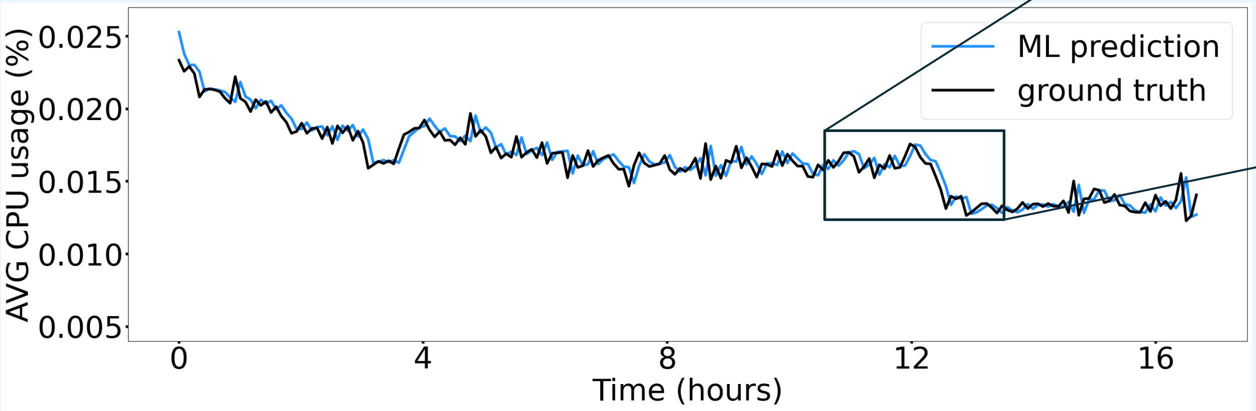
"We used **LSTM** for time series forecasting."

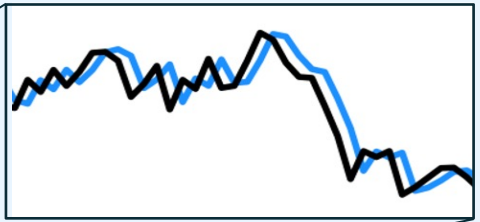
Seer: Leveraging Big Data to Navigate the Complexity of Performance Debugging in Cloud Microservices

"The **LSTM** is especially effective at capturing load patterns over time."
ASPLOS, 2019

Debunking the High Accuracy of LSTMs

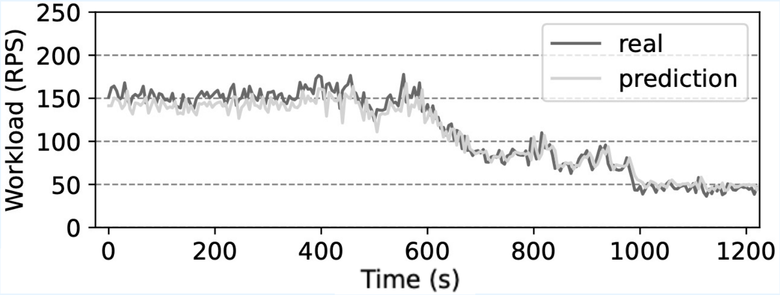
Usecase: Cloud Workloads.

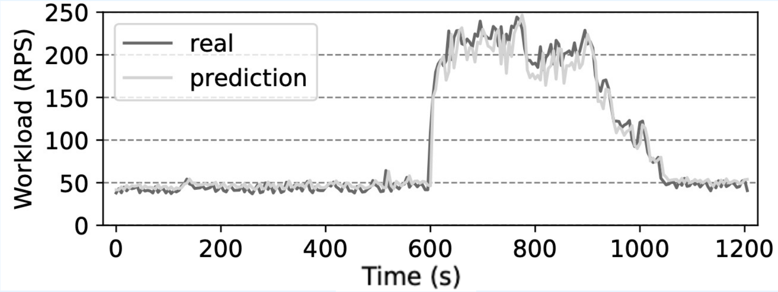




Our Insight: LSTM predictions resemble the **previous** timestep of the timeseries.

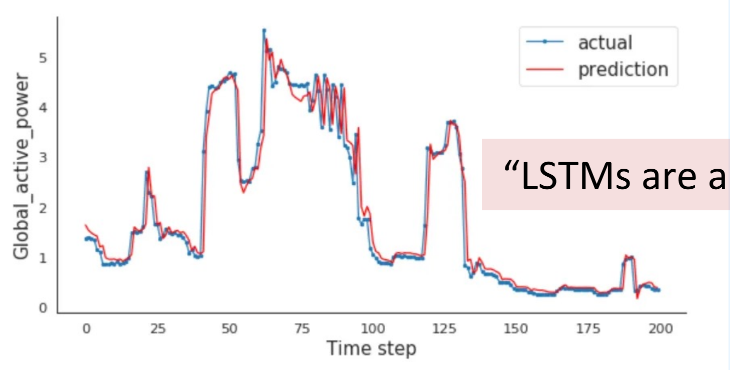
Usecase: ML Inference Services.





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

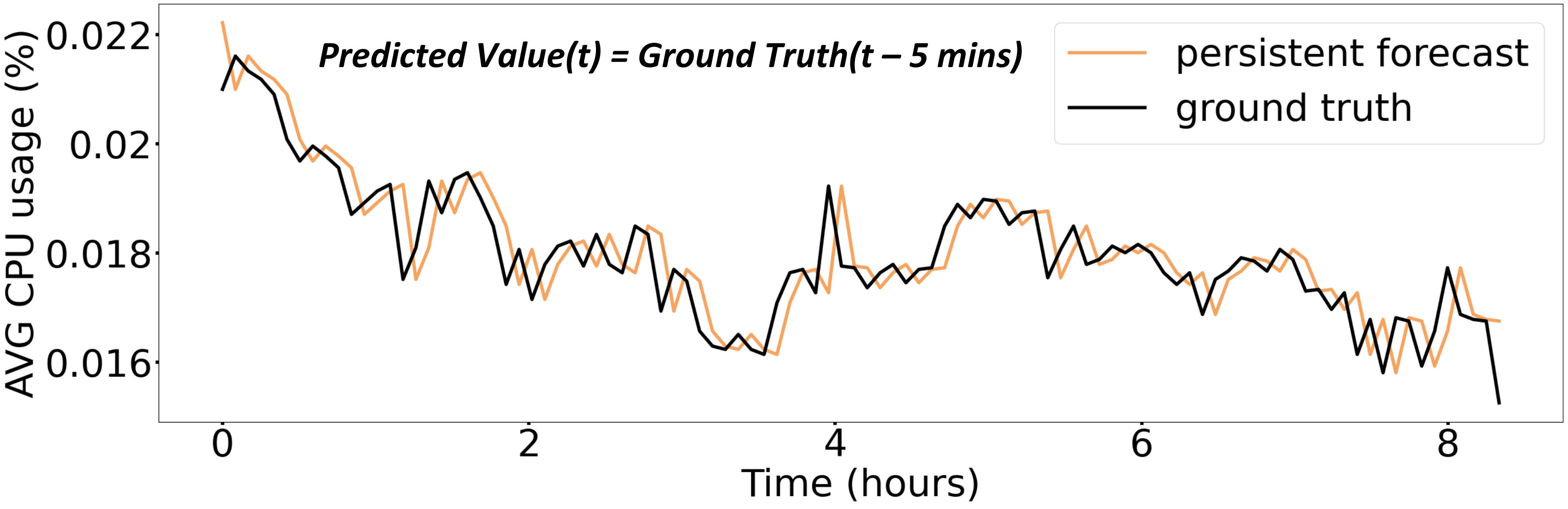
Do we need ML to produce such "shifted" predictions?

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

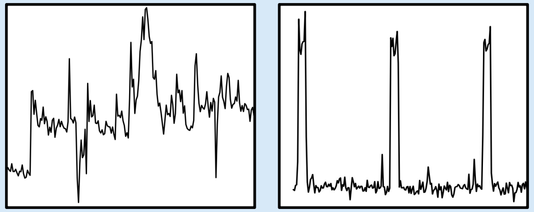
Resource Types

CPU
RAM

Resource Levels

Physical Machine
Virtual Machine
Workload

Usage patterns



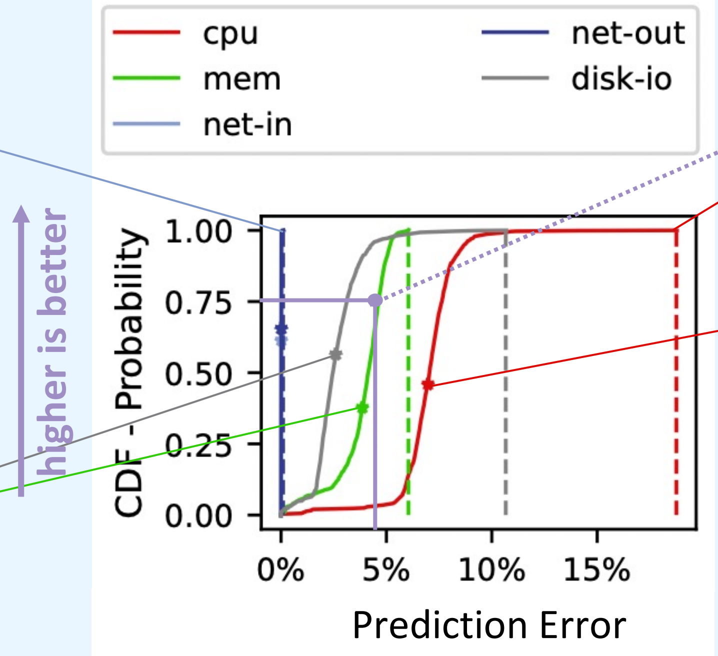
Frequency

Hourly/ Daily/ Weekly Windows

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

Experimental Results

Alibaba Dataset
Physical Machine Level



lower is better

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

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

CPU: has the largest tail

CPU: 6.97% on average (more dynamic patterns)

DISK-IO & MEM: Average Error < 4%

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

More experiments and graphs in the paper!

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