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 servers and data centres are located in computing and storage takes place on servers.

Businesses and single users benefit from cloud computing in two main ways:

1. Reduce IT costs
2. Easier to operate internationally

Files and Applications can be accessed from different devices.

Buy and manage physical servers.

Cloud providers:
Basic Concepts of Cloud Computing

2. Virtual Machine (VM)

- **digital-only** computer
- behaves as a **physical computer** with its own hardware

3. Physical Machine

Server, Host Machine

- run on the same

Many VMs

- VMs do not interact with each other

Levels at which we can observe Resource Usage:

1. **Workloads**
   - An **Application** performing a specific task that uses a specific amount of resources (processing, storage, network etc)
   - Serve more customers at once
   - Low cost – High hardware efficiency

2. **Virtual Machines**
   - Each VM is created and configured by the user

3. **Physical Machines**
The Problem of Cloud Resource Usage Forecasting

Low resource efficiency in the Cloud

Cloud resource management systems

Autoscaling

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

Active Servers

-取决于负载和用户需求

Virtual Machine Configuration

Asks: \( X > Y \) (Uses)

Resource Waste

Resource shared between VMs

- 资源共享

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

Overcommitment

- 物理机

\[ A + B + C > D \]
The Problem of Cloud Resource Usage Forecasting

**Approach:** Future Resource Usage Forecasting

- **Input:** Past Resource Usage $x_1, x_2, ..., x_n$
- **Forecasting Models:** (ML, Statistical, Heuristic, Hybrid)
- **Output:** Future Resource Usage $x_{n+1}, x_{n+2}, ..., x_{n+k}$

**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**

**Dynamic patterns**

**Average CPU usage**

**Stable patterns**

**Average memory usage**

**Virtual Machine level**

**Periodic patterns**

**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

LSTMs for **Cloud** Resource Usage Forecasting

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

*EuroSys, 2023*

“We used LSTM for time series forecasting.”

**“BHyPreC: A Novel Bi-LSTM Based Hybrid Recurrent Neural Network Model to Predict the CPU Workload of Cloud Virtual Machine”**

*IEEE Access, 2021*

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*

High accuracy when predicting:

- Weather
- Stock Market Prices
- Power Consumption
- Traffic Conditions
Debunking the High Accuracy of LSTMs

Use case: Cloud Workloads.

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

Use case: ML Inference Services.

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


Use case: Global Active Power Consumption

“LSTMs are amazing!”

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

$\text{Predicted Value}(t) = \text{Ground Truth}(t - 5 \text{ mins})$
Experimental Methodology

Extensive experimental evaluation with cloud resource usage data.

Public open-source datasets across different:

Cloud providers: Alibaba Cloud, Google Cloud, Bitbrain

Resource Types:
- Physical Machine
- Virtual Machine
- Workload

Resource Levels:

Usage patterns:

Frequency:
- Hourly/Daily/Weekly Windows

We calculate the prediction error of the persistent forecast.
Experimental Results – Physical Machine Level

**Alibaba Dataset Physical Machine Level**

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

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

**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 measure resource usage on a deeper level.
Experimental Results – Workload Level

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

\[ \text{Predicted Value}(t) = \text{Ground Truth}(t - 5 \text{ mins}) \]

What happens when we increase the time window?

\[ \text{Predicted Value}(t) = \text{Ground Truth}(t - \text{time\_window}) \]

**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

Resource Level

- Physical Machine
- VM
- Workload

More stable and periodic utilization levels
More dynamic resource usage

Resource Type

- CPU
- Memory

Resource Measurement

Prediction Error of MAX > AVG > MIN

Lessons Learned

- Average CPU usage shows periodic and daily patterns.
- Memory levels more stable over time.
Is Machine Learning Necessary for Cloud Resource Usage Forecasting?

No. (for the most part)

Open questions

1. **When** to use ML?
   - exact use case
   - data pattern

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.