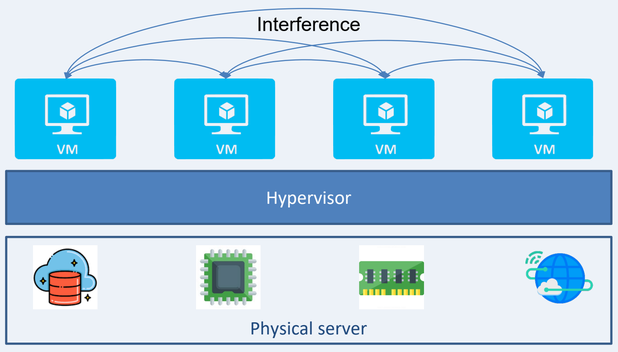
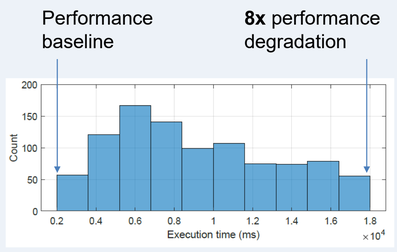
CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds

Darong Huang, Luis Costero, Ali Pahlevan, Marina Zapater, David Atienza
Embedded Systems Laboratory

Introduction

- Infrastructure as a service (IaaS):
  - IaaS contributes to 25% market share in all cloud services
  - IaaS is forecasted to experience the highest end-user spending growth in 2023 at 30.9%
- Performance problem of IaaS:



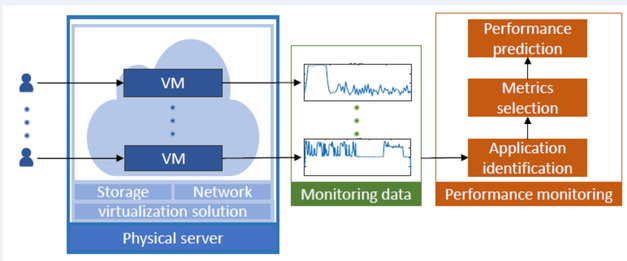


Performance of Redis benchmark

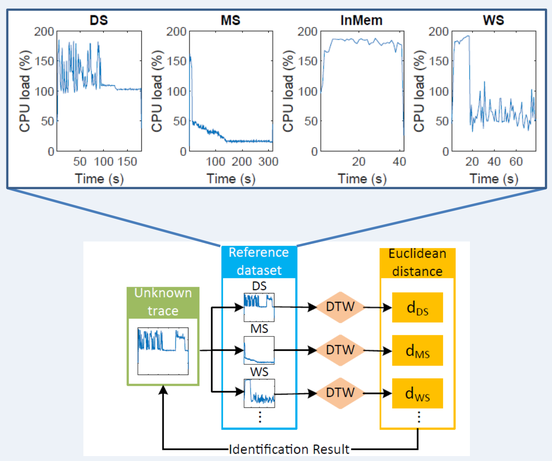
  - Collocated VMs can suffer from severe performance degradation!

Motivation and challenges

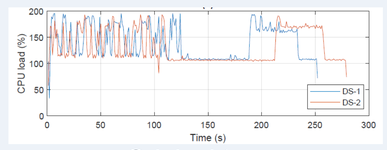
- Motivation:
  - Performance prediction of VMs
- Challenges:
  - VM is a black-box
  - Resource contention induced performance variation
  - Workload induced performance variation



Application Identification



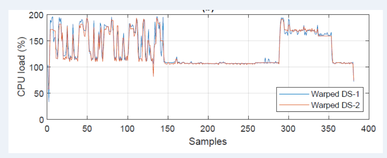
- Offline: Create the reference dataset
- Online: Dynamic time warping (DTW) - based identification



Original traces

$$d(m,n) = \sqrt{\sum_{i \in t} (m_i - n_i)^2}$$

- Low prediction accuracy on original traces
- Misprediction due to large distance



Warped traces

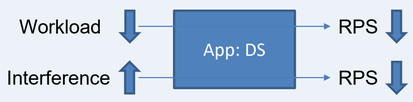
$$d_{min}(m,n) = \min_{\pi \in A(m,n)} \left(\sum_{(i,j) \in \pi} \sqrt{(m_i - n_j)^2} \right)$$

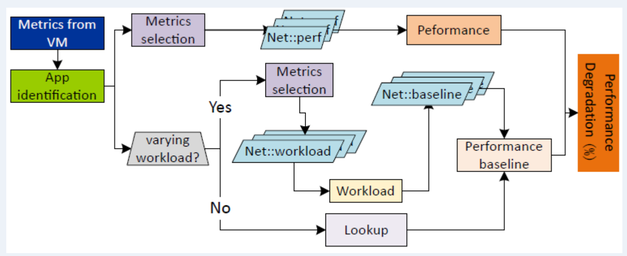
Dynamic time warping

- Warped traces are minimized in the distance
- For better identification accuracy

Performance degradation prediction

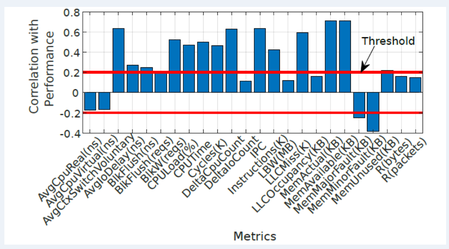
- **Problem:** Both workload and interference influence the performance level of the application!
- **Solution:** Alongside the **performance prediction**, two additional NNs are used for **workload prediction** and **performance baseline prediction**

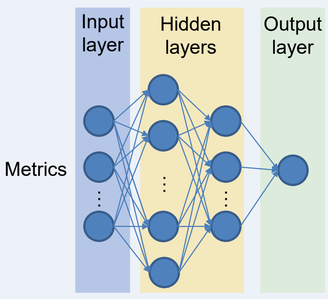


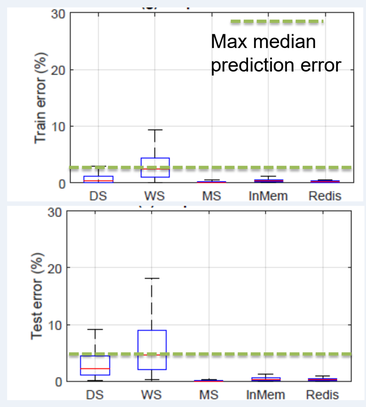


I. Performance prediction

- Highly correlated metrics are selected for the prediction
- Proposed method achieves a mean error below 5% for all applications!



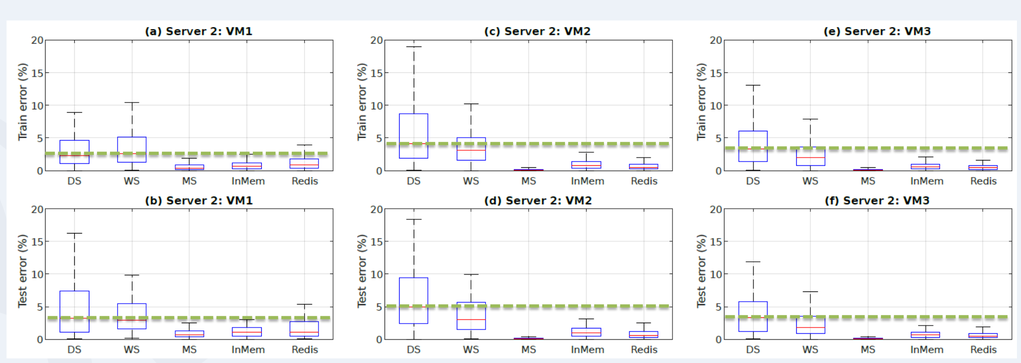




II. Performance degradation prediction

- NN-based performance degradation prediction method achieves a mean error below 7% for all applications
- Moreover, the proposed method can work with different server and VM configurations

App	Train error (%)			Test error (%)		
	mean	max	std	mean	max	std
DS	4.5	44.7	5.6	5.4	26.0	5.2
WS	4.6	27.0	4.3	6.8	32.1	6.1
MS	0.5	7.6	0.9	1.2	16.4	2.9
InMem	0.4	3.3	0.4	0.6	7.4	0.8
Redis	0.2	1.3	0.2	0.3	1.7	0.3



Server	CPU	Memory	OS
S1	2x Intel Xeon E5-2620 v3 @ 2.4GHz	128GB ECC	EulerOS 2.0
S2	2x Intel Xeon Gold 6242R @ 3.1GHz	384GB ECC	CentOS 7.9

VM name	CPU	Memory	OS
VM1	2x vCPU	4GB	Ubuntu 20.04
VM2	4x vCPU	4GB	Ubuntu 20.04
VM3	4x vCPU	8GB	Ubuntu 20.04

Acknowledgments

This work has been partially supported by the EC H2020 RECIPE FET-HPC project (No. 801137), the ERC Consolidator Grant COMPUSAPIEN (No. 725657), an Industrial Grant from Huawei Cloud to ESL-EPFL, and Grant PID2021-126576NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

D. Huang, L. Costero, A. Pahlevan, et al. CloudProphet: A Machine Learning-Based Performance Prediction for Public Clouds arXiv preprint arXiv: 2309.16333

Summary

- Black box VM assumption
- DTW-based app. identification
- Accurate performance prediction (7% mean error)
- Support different server and VM configurations