Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

AGH University of Krakow

Towards Sustainable Cloud Environments by Leveraging Time Series Forecasting for Enhanced Resource Utilization

Mateusz Smendowski

☁ Cloud Computing in a Nutshell ☁

on-demand – *pay-as-you-go* – scalable – flexible

90% of IT enterprises will shift to cloud solutions shortly.

In 2020, data centers consumed **200 TWh** of energy.

While clouds are essential, they also entail inherent **risks**.

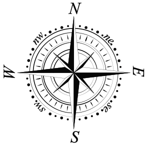
Pessimistic resource allocation

Costly on-demand resources

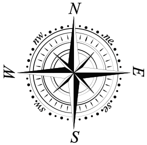
Regional resource shortages

Limited strategic foresight

energy inefficiency
resource wastage
escalated costs
SLAs violations
low QoS & QoE



GCC – Green Cloud Computing



FinOps – Financial Operations

Sustainable Cloud Environments

What? – Objective

Towards Sustainable Cloud Environments

How? – Method

by Leveraging Time Series Forecasting

Why? – Motivation

for Enhanced Resource Utilization

What? – Objective

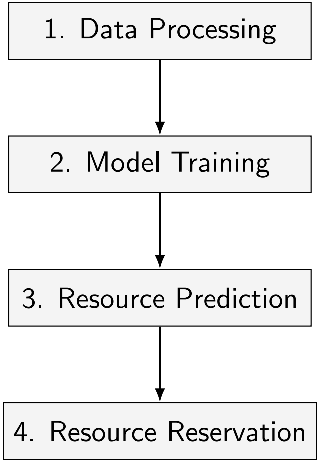
Towards Sustainable Cloud Environments

How? – Method

by Leveraging Time Series Forecasting

Why? – Motivation

for Enhanced Resource Utilization



High-level Resource Usage Optimization Scheme

Process multivariate historical resource consumption metrics from real-life production cloud environment – CPU, RAM, Disk, Network.

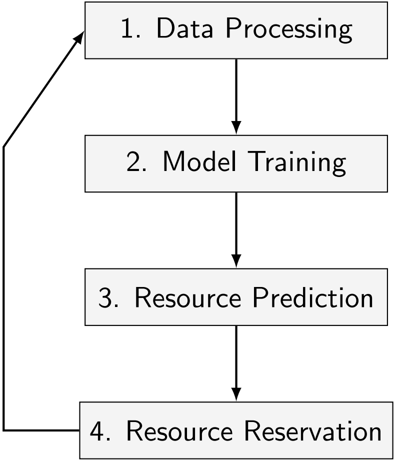
– variable length and dimensionality in time series –

Train long-term time series forecasting model(s).

– Gated Recurrent Unit (GRU) Neural Network(s) –

Forecast CPU and RAM usage with a weekly horizon for each of the 8 virtual machines.

Create weekly resource reservation plans based on predicted demand and adjust resources accordingly.



High-level Resource Usage Optimization ~~Scheme~~ Loop

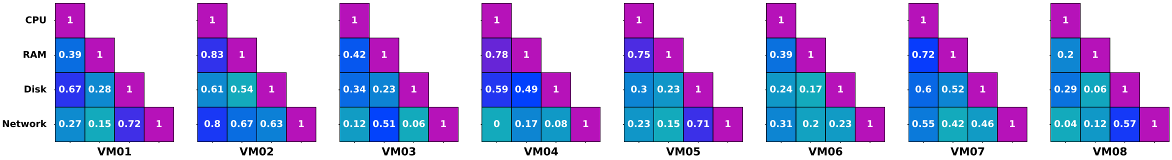
General Purpose Virtual Machines

High-Performance Computing Virtual Machines

Cloud environments are evolving **rapidly** *and* **dynamically.**

Optimization solutions themselves should be **scalable** *and* **cost-effective.**

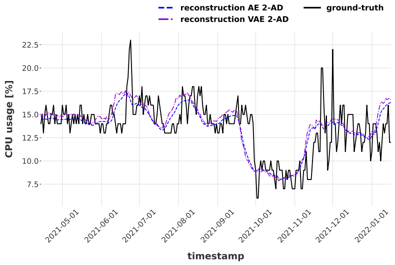
GFM	GSFM	LFM
Global Forecasting Model	Group-specific Forecasting Model	Local Forecasting Model



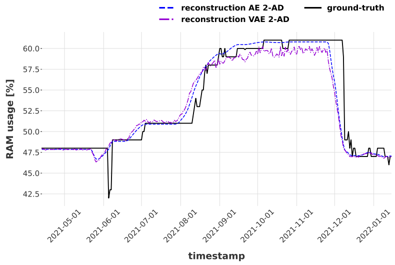
Relations between resource usage metrics, even within the same virtual machine, **are not obvious** – based on an 8-month training set.

GFM	**GSFM**	**LFM**
Global Forecasting Model	Group-specific Forecasting Model	Local Forecasting Model

Similarity-based Time Series Grouping (STG) – use a recurrent autoencoder for semantic representation of weekly time series sequences, apply K-means clustering with the DTW distance, and group time series based on majority membership of embedded sequences to distinguished clusters.



VM06 – CPU reconstruction



VM03 – RAM reconstruction

Resource	**AE 1**	**AE 1 -AD**	**AE 2**	**AE 2 -AD**	**VAE 1**	**VAE 1 -AD**	**VAE 2**	**VAE 2 -AD**
CPU	0.150	0.130	0.144	**0.126**	0.135	0.130	0.135	**0.129**
RAM	0.180	0.165	0.179	**0.164**	0.172	0.178	**0.170**	0.175
Disk	0.168	0.147	0.183	**0.142**	0.184	0.169	0.183	**0.159**
Network	0.130	0.118	0.120	**0.117**	0.127	**0.121**	0.122	0.124

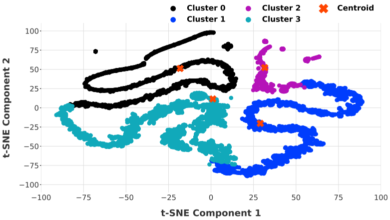
Based on 1st level metric – RMSE, **AE 2-AD is superior** to VAE 2-AD.

Cluster assessment metrics using **AE 2-AD** as the sequence embedding model.

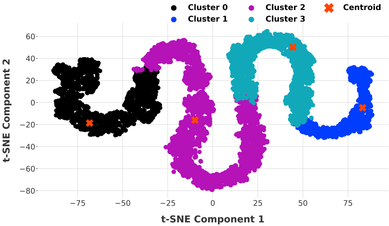
No of clusters	Silhouette score	Davies-Bouldin index	Calinski-Harabasz index	Global purity	Global Gini index	No of idle global clusters	No of time series per global cluster
2	0.542	0.746	11,490	0.788	0.063	0	18; 14
3	0.561	0.578	17,847	0.739	0.313	0	15; 16; 1
4	**0.557**	**0.542**	**22,597**	**0.629**	**0.281**	**0**	**8; 12; 1; 11**
5	0.568	0.509	28,546	0.599	0.338	0	9; 6; 1; 12; 4
6	0.572	0.489	34,813	0.544	0.396	0	10; 9; 2; 9; 1; 1
7	0.556	0.495	41,869	0.491	0.402	0	9; 9; 1; 5; 1; 1; 6

Cluster assessment metrics using **VAE 2-AD** as the sequence embedding model.

No of clusters	Silhouette score	Davies-Bouldin index	Calinski-Harabasz index	Global purity	Global Gini index	No of idle global clusters	No of time series per global cluster
2	0.548	0.541	14,604	0.782	0.343	0	27; 5
3	0.524	0.544	22,873	0.670	0.208	0	8; 7; 17
4	**0.545**	**0.528**	**31,881**	**0.629**	**0.188**	**0**	**9; 5; 12; 6**
5	0.498	0.562	32,744	0.564	0.250	1	3; 5; 12; 7; 5
6	0.481	0.585	38,416	0.487	0.292	1	3; 5; 12; 5; 5; 2
7	0.473	0.600	40,632	0.450	0.303	2	4; 5; 11; 4; 4; 3; 1

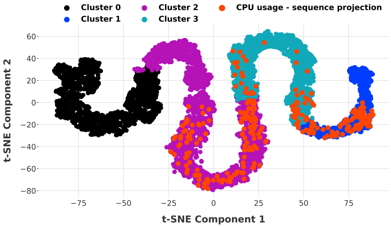


Clustering post AE 2-AD

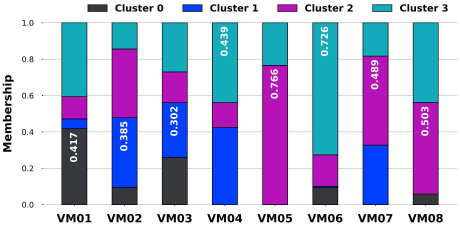


Clustering post VAE 2-AD

Based on 2nd level metrics, **VAE 2-AD is superior** to AE 2-AD.

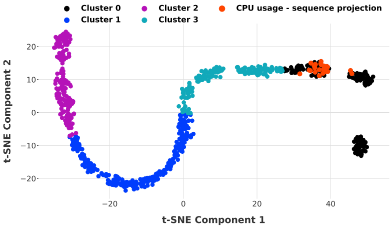


VM07 – CPU sequences projection

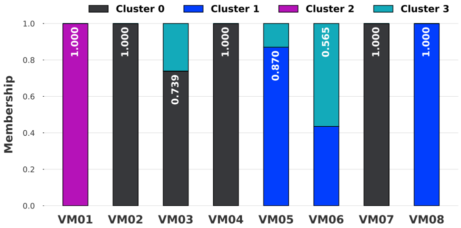


CPU Membership

Assigning time series to a group is **ambiguous** when STG is performed on both the training (8 months) and validation (1 month) sets.

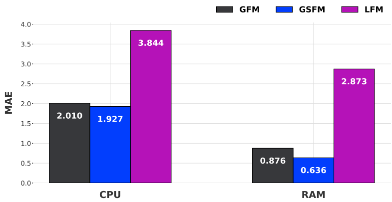


VM07 – CPU sequences projection

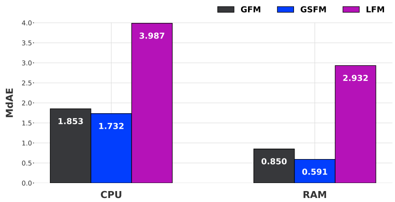


CPU Membership

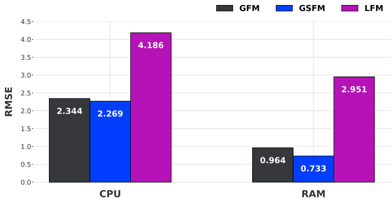
Assigning time series to a group is **unambiguous** when STG is performed on the validation set only – earliest available resource usage metrics for analysis.



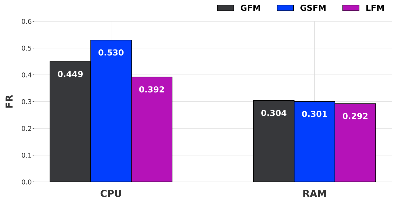
MAE



MdAE



RMSE



FR

Aggregated evaluation metrics for predictions on test (3 months) set.

Initial resource reservation plans – aggregated domain-specific metrics.

Reference type	Percentage cost reduction (with GSFM)	Daily USD cost (without GSFM)	Daily USD cost (with GSFM)	Percentage CPU usage (with GSFM)	Percentage RAM usage (with GSFM)	Scaling events	Violation events	Percentage availability
e2-std-2	0.00	2.07	2.07	17.69	22.13	0.00	0.00	100.00
e2-std-4	39.40	4.14	2.51	35.39	38.17	0.63	0.13	99.81
e2-std-8	48.24	8.29	4.29	48.88	44.40	1.13	0.63	99.04
e2-std-16	50.74	16.58	8.17	55.41	56.86	2.25	2.00	96.93
e2-std-32	57.58	33.15	14.06	60.60	57.52	2.25	2.13	96.75

Initial Plan → **Administrative Rules** → Adjusted Plan

Recommend a more robust machine type within the e2-std flavor hierarchy if the forecasted demand surpasses 80% of the initially suggested machine's computational capacity.

Adjusted resource reservation plans – aggregated domain-specific metrics.

Reference type	Percentage cost reduction (with GSFM)	Daily USD cost (without GSFM)	Daily USD cost (with GSFM)	Percentage CPU usage (with GSFM)	Percentage RAM usage (with GSFM)	Scaling events	Violation events	Percentage availability
e2-std-2	0.00	2.07	2.07	17.69	22.13	0.00	0.00	100.00
e2-std-4	20.87	4.14	3.28	29.52	36.42	0.75	0.00	100.00
e2-std-8	34.02	8.29	5.47	35.33	40.23	1.50	0.00	100.00
e2-std-16	36.98	16.58	10.45	35.06	41.12	4.75	0.25	99.62
e2-std-32	44.71	33.15	18.33	36.07	39.75	4.75	0.25	99.62

Optimization **Towards Sustainable Cloud Environments** is indeed a trade-off.

References:

- M.Smendowski, P.Nawrocki;
  Optimizing multi-time series forecasting for enhanced cloud resource utilization based on machine learning;
  Knowledge-Based Systems, 2024;
  DOI: https://doi.org/10.1016/j.knosys.2024.112489
- P.Nawrocki, M.Smendowski;
  FinOps-driven optimization of cloud resource usage for high-performance computing using machine learning;
  Journal of Computational Science, 2024;
  DOI: https://doi.org/10.1016/j.jocs.2024.102292

Thank you for your attention.

Let's connect on