

Do Predictors for Resource Overcommitment Even Predict?

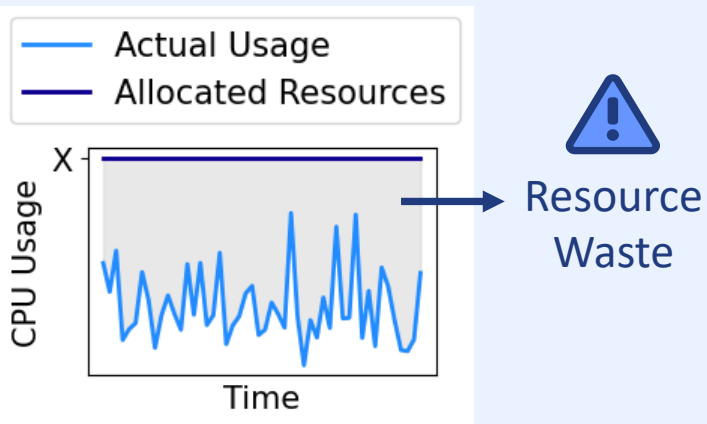
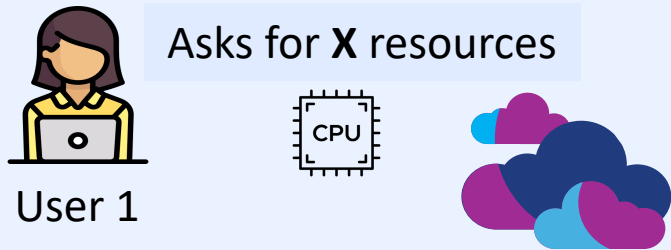
Georgia Christofidi, Thaleia Dimitra Doudali
@EuroMLSys '24

IMDEA Software Institute, Madrid, Spain



Problem Space – Resource Overcommitment

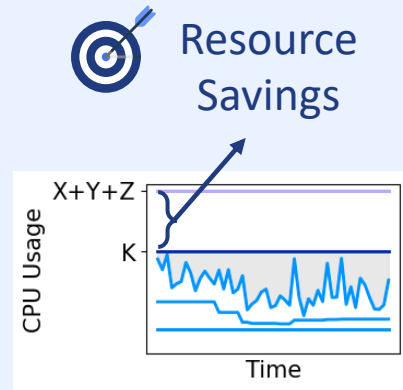
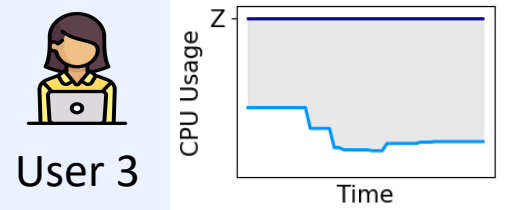
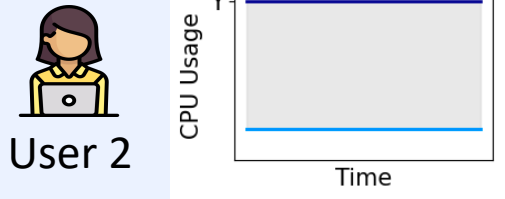
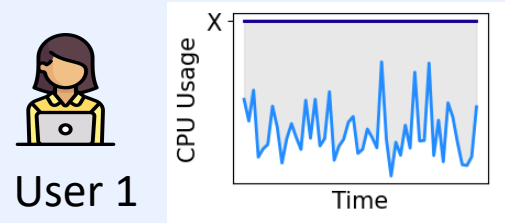
1. Traditional Resource Allocation



Application allocated to a machine with capacity $K \geq X$



2. Resource Allocation with **Overcommitment**



Applications allocated to a machine with capacity $K < X+Y+Z$

? How?
Predict resource usage!

✓ **Benefit:** Improve resource and cost efficiency

✗ **Risk:** Resource usage underestimations may cause competition for resources

Accurate predictions will increase benefit and reduce risk.

Do Current Predictors Even Predict?

Scan for code and paper:



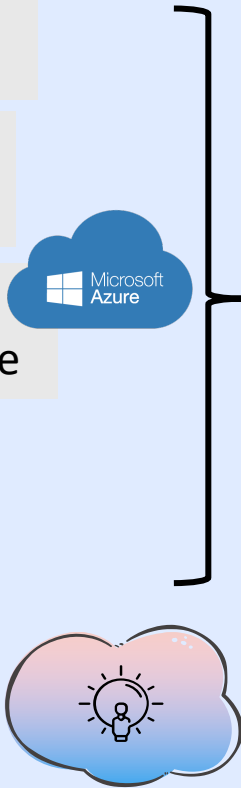
Existing Predictors

Future Usage =

1. Borg
 $90\% * U$ Google Cloud

2. Resource Central
sum of the k-th %-ile

3. N-Sigma
 $U + N * std(U)$



Google Cloud
4. TITTL
Max (1, 2, 3)



Our Insight: TITTL has an average prediction error of 568%.
Prediction > Resource Limit **94%** of the times
(No Overcommitment!)

Usable resource savings (2%)

Maximum feasible resource savings (61%) with an accurate predictor

Resource waste: can never be used

Future Work: build a lightweight, practical predictor that **accurately predicts** and closes the existing gap in attainable resource savings.

