CP395 Weekly Report 2

Sufiya Rahemtulla – 169018559

1. Systematic Search Strategy & Search Log

- Identify state-of-the-art literature (2020–2025) on cloud autoscaling, specifically focusing on the comparison between reactive (threshold-based) and predictive (machine learning/reinforcement learning) approaches.
- Used IEEE Xplore, ACM Digital Library, Google Scholar
- Inclusion/Exclusion Criteria:
    - **Timeframe:** 2020–2026 (to ensure relevance to modern container/cloud-native environments).
    - **Required Content:** Papers must include a quantitative evaluation using system-level metrics (SLA violation rates, resource utilization, or cost).
    - **Exclusion:** Purely theoretical proposals without experimental validation, papers focusing solely on hardware energy consumption without performance trade-offs, and non-English results.
- **Search log**:

Search Log					
DataBase	**Query String**	**Date**	**Results**	**Papers Retained**	**Notes**
IEEE Xplore	("cloud autoscaling" OR "workload prediction") AND "SLA"	18-Jan-26	27	4	*open access only
Springer	"predictive autoscaling"	18-Jan-26	3	0	*open access only
Springer	"reactive AND "autoscaling"	18-Jan-26	22	4	*open access only
Google Scholar LABS (AI)	similar to Deep Q-learning" vs "Q-learning" autoscaling cloud workload that have been published between 2020-2026 AND are open access	18-Jan-26	10	1	*open access only
Google Scholar	"Google cluster trace v3" workload characterization autoscaling	18-Jan-26	10	1	*open access only

2. Taxonomy Overview

I have organized the literature into a taxonomy based on the decision mechanism (Reactive vs. Predictive) and the learning method (Heuristic vs. Machine Learning). The table below compares the representative approaches identified.

Source	Title	Predictive vs reactive?	Heuristic vs learning-based?	Offline vs online control	RL based vs supervised learning
IEEE Xplore	**Predictive Hybrid Autoscaling for Containerized Applications**	Predictive (Hybrid)	Learning-based	Online Control (Offline Training)	Supervised (Time-series)
IEEE Xplore	**Online Workload Burst Detection for Efficient Predictive Autoscaling...**	Predictive	Learning-based	Online Control	Supervised (Classification)
IEEE Xplore	**Proactive Random-Forest Autoscaler for Microservice Resource Allocation**	Predictive	Learning-based	Online Control (Offline Training)	Supervised (Random Forest)
IEEE Xplore	**On the Stability of the Kubernetes Horizontal Autoscaler Control Loop**	Reactive	Heuristic (Threshold/PID)	Online Control	N/A (Rule-based)
Springer	**Effective priority-based... using hybrid tree-enhanced vector machine model**	Predictive	Learning-based	Online Control (Offline Training)	Supervised (Vector Machine)
Springer	**Enhancing Machine Learning-Based Autoscaling for Cloud Resource Orchestration**	Predictive	Learning-based	Online Control (Offline Training)	Supervised
Springer	**Signature-based Adaptive Cloud Resource Usage Prediction Using Machine Learning...**	Predictive	Learning-based	Online Control	Supervised (Anomaly Detection)
Springer	**Application Optimisation: Workload Prediction and Autonomous Autoscaling...**	Predictive	Learning-based	Online Control (Offline Training)	Supervised
Google Scholar	**A Deep Q-learning Scaling Policy for Elastic Application Deployment**	Predictive (Implicit)	Learning-based	Online Training & Control	RL Based (Deep Q-Network)
Google Scholar	**Autopilot: workload autoscaling at Google**	Reactive (Window-based)	Heuristic (Sliding Window)	Online Control	N/A (Statistical Rule-based)

***Taxonomy Summary/Comparison DRAFT: Key Papers (Reactive vs Predictive vs RL)**

- *Has been submitted in the Dropbox as an Excel file (spacing errors will not let me paste here). Summary is below:*
- **Reactive Baseline (e.g., Kubernetes HPA):** While simple to implement and relying only on standard inputs like CPU utilization, these threshold-based methods suffer from significant "provisioning lag" and oscillation when facing bursty workloads.

- **Predictive ML (e.g., Random Forest):** Supervised learning approaches address the lag problem by forecasting demand based on historical traces. However, they rely heavily on the assumption that past patterns (seasonality) will repeat and are vulnerable to "concept drift" if the workload character changes unexpectedly.

Reinforcement Learning (e.g., Deep Q-Network): RL offers the most flexibility by learning optimal policies through trial-and-error without explicit thresholds. However, the table identifies critical risks for a 12-week project, specifically the high training overhead and initial instability during the "exploration" phase

Based on this taxonomy, my project will likely proceed with a **Supervised Learning (Predictive)** approach. It offers the proactive benefits required to minimize SLA violations while avoiding the high computational cost and stability risks associated with training an RL agent from scratch. The Reactive (Kubernetes) model will serve as the control group for evaluation

3. Preliminary Research Gaps

- **Lack of Realistic Constraints in RL Evaluations:** Many RL-based autoscaling papers assume instant VM startup times. My project will introduce realistic provisioning delays to test if predictive models fail under 'laggy' conditions.
- **Over-Reliance on Prediction Accuracy:** Existing studies often optimize for low prediction error (MSE) but fail to measure if that accuracy translates to fewer SLA violations in a noisy system.
- **Absence of Failover Mechanisms:** Few predictive approaches discuss safety strategies. I intend to explore a hybrid mechanism where reactive rules override the predictive model during extreme prediction errors."

4. Status of Literature Review Draft

- Completed: systematic research, paper selection (10 papers, see autobiography below), taxonomy construction
- In Progress: Writing the comparative discussion and synthesizing the “Background” section

5. Artifact Links:

- **Annotated Bibliography**: https://lauriercloud-my.sharepoint.com/:w:/r/personal/rahe8559_mylaurier_ca/Documents/Annotated%20Bibliography.docx?d=w74f1ef6037c94e4e96625280a5d6c9a4&csf=1&web=1&e=aqS1JC
- **Overleaf Literature Review DRAFT:** https://www.overleaf.com/read/jfgyhqcmfhtq#647fa3
- **Git Repository:** https://github.com/scr200/CP395-Research-Project
- **Search Log & Taxonomy Tables:** https://lauriercloud-my.sharepoint.com/:x:/r/personal/rahe8559_mylaurier_ca/Documents/Search%20Log%20CP395.xlsx?d=w1971d6047491461eadfe3ad43ecf2dfa&csf=1&web=1&e=LZc08n