ITMO

SpacSec 2026
International Conference on Cyber Warfare, Security & Space Computing

Alexandra A. Chudinova

WORKLOAD ANALYSIS: UNCERTAINTY AND MULTI-CRITERIA

Scientific advisor:
Doctor of Technical Sciences, Professor **Bogatyrev Vladimir Anatolyevich**

2026

Purpose

To establish that the trade-off between service quality and resource expenditure in Q-learning and threshold-based control depends on the load level, and to determine the conditions (load and SLO) under which an adaptive strategy offers a practical advantage.

Compare two autoscaling scenarios *in a simulation cluster (40 nodes, 320 cores, load ranging from 1,000 to 12,000 RPS): a deterministic threshold-based controller and an adaptive controller based on tabular Q-learning.*

Uncertainty and Multi-criteria

Multi-objective optimization defines what we want to achieve simultaneously (speed and low cost).

Uncertainty makes this difficult, as **the workload is not known in advance**.

Adaptive reinforcement learning (**RL**) is necessary precisely because the cost of the trade-off shifts as the workload changes, a factor that threshold-based methods fail to account for.

Scientific novelty

The scientific novelty lies in a quantitative analysis of how the trade-off between **QoS** and **resource costs** depends on the **workload level**, specifically for threshold-based and reinforcement learning (RL) approaches to autoscaling, as well as in framing the comparison as an optimization problem subject to service-level objective (SLO) constraints, which makes it possible to determine the conditions under which an adaptive strategy is practically advantageous.

Research Questions

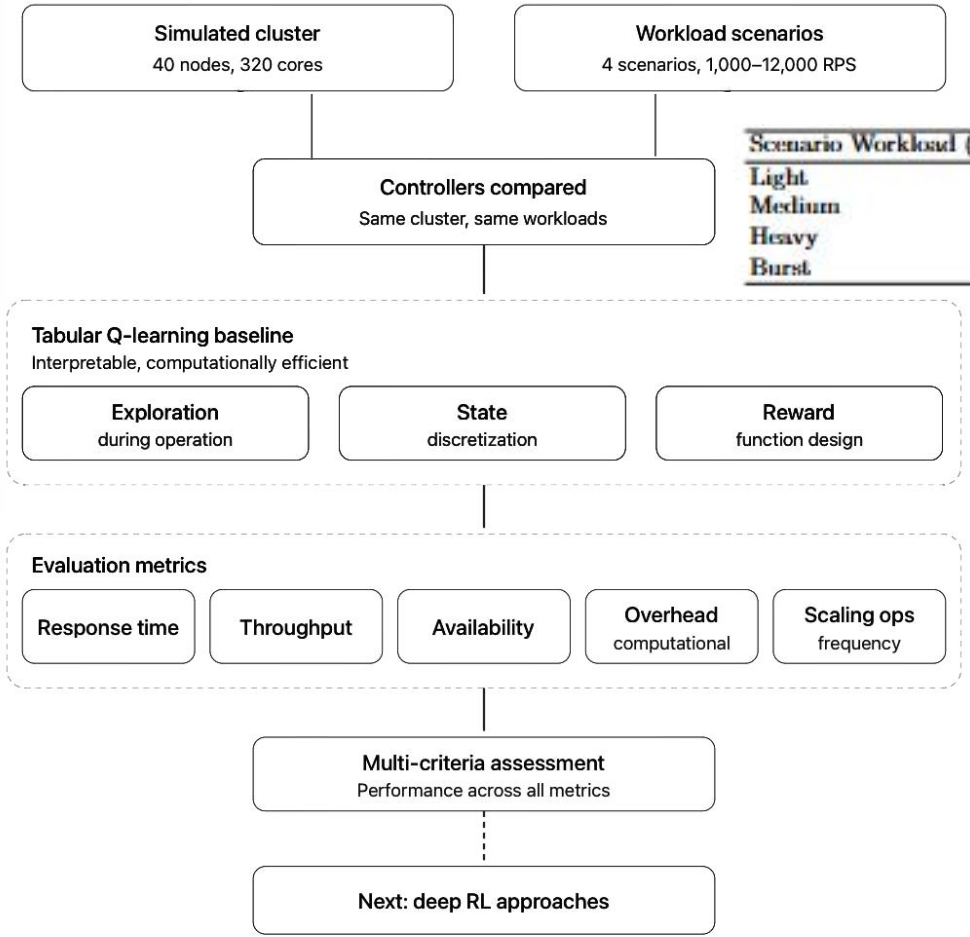
1. How do latency, throughput, and resource cost change for threshold-based and **Q-learning** policies under different workload levels?

2. Can an RL policy reduce allocated resources and **normalized computational cost** while maintaining acceptable service performance?

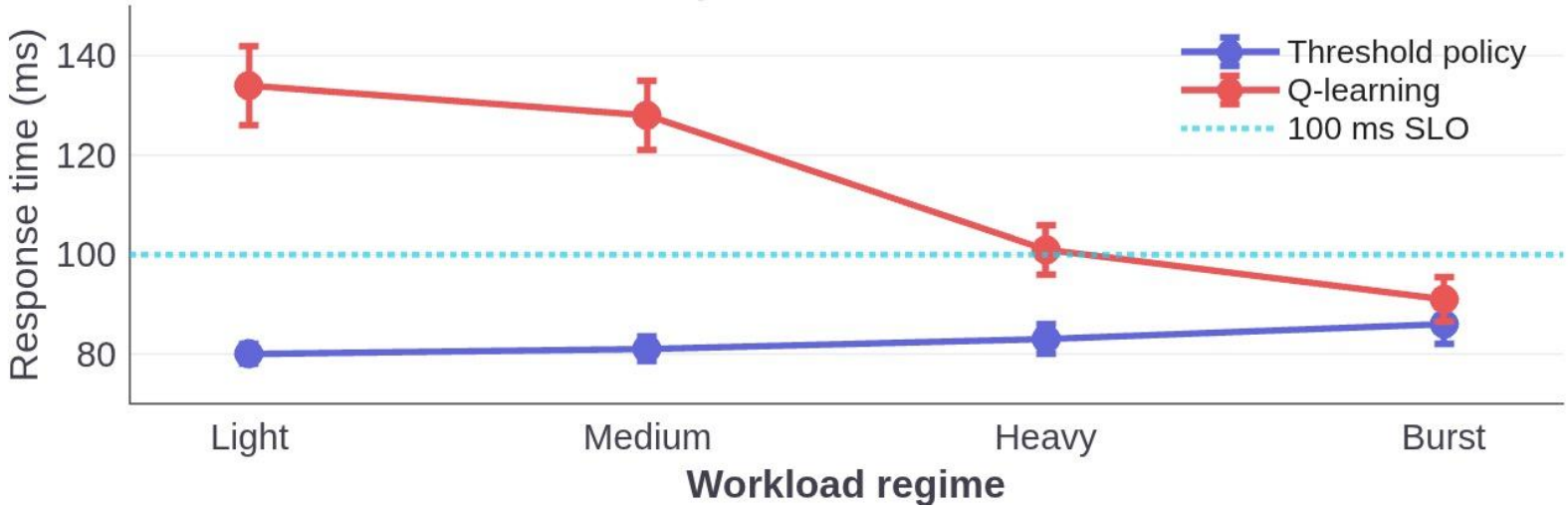
3. How does **the latency** and **cost trade-off** change from light and medium workloads to heavy and burst workloads?

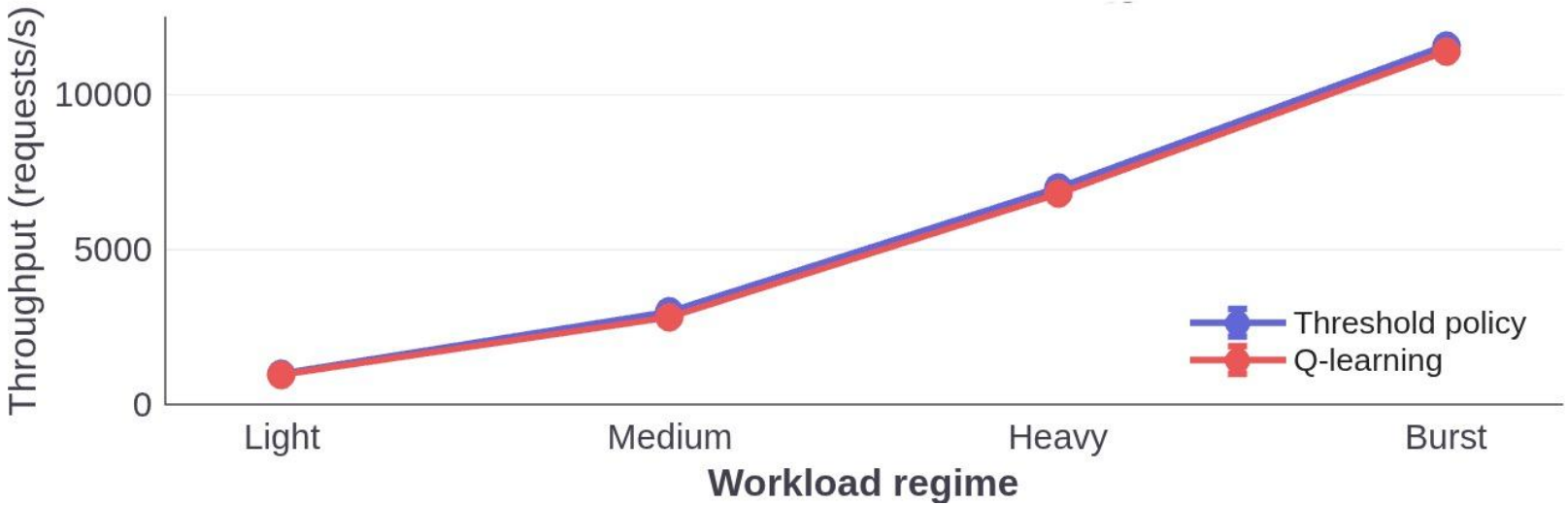
4. Under what **workload conditions** can an adaptive **RL policy** provide practical advantages over a deterministic threshold policy?

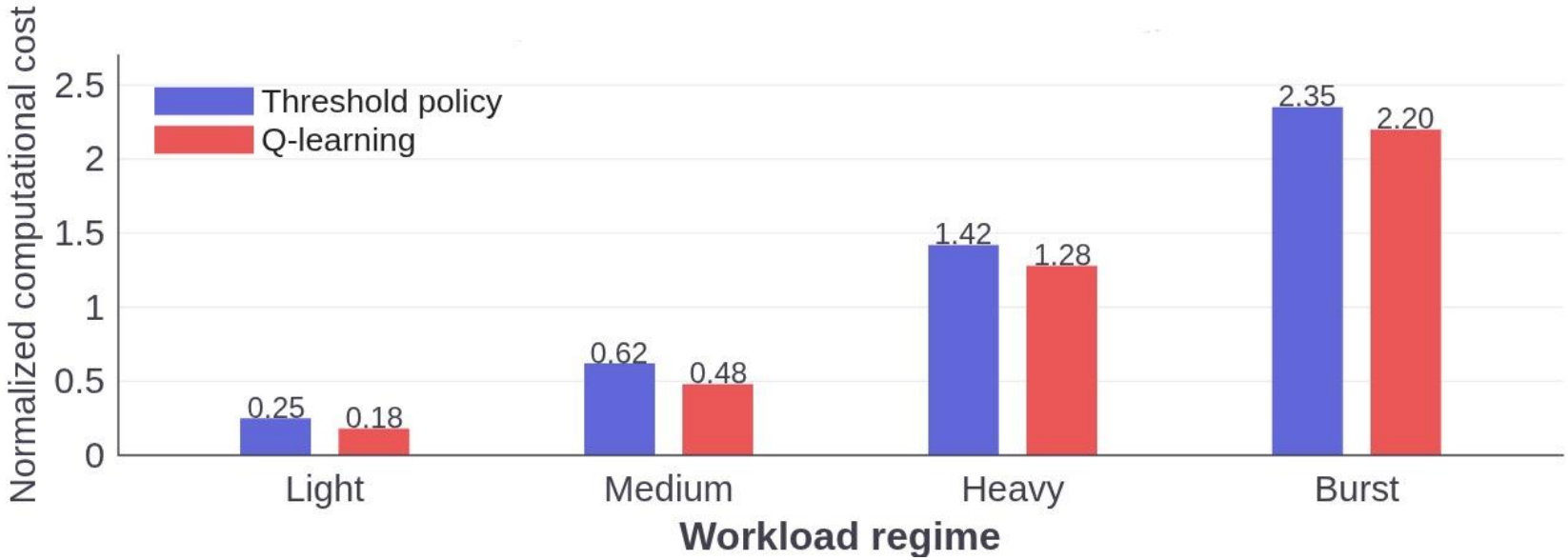
Parameter	Value
Number of nodes	40
CPU cores per node	8
Total CPU capacity	320 units
Replica capacity	50 requests/s
Base response time	80 ms
Latency coefficient	0.8
Node failure probability	0.002 per step
Repair probability	0.1 per step
Episodes per run	30
Steps per episode	60
CPU scale-up threshold	0.70
CPU scale-down threshold	0.30
RT scale-up threshold	200 ms
RT scale-down threshold	100 ms
CPU bins	6
Response-time bins	6
Replica levels	8
Learning rate η	0.10
Discount factor γ	0.95
Exploration probability ϵ	0.15

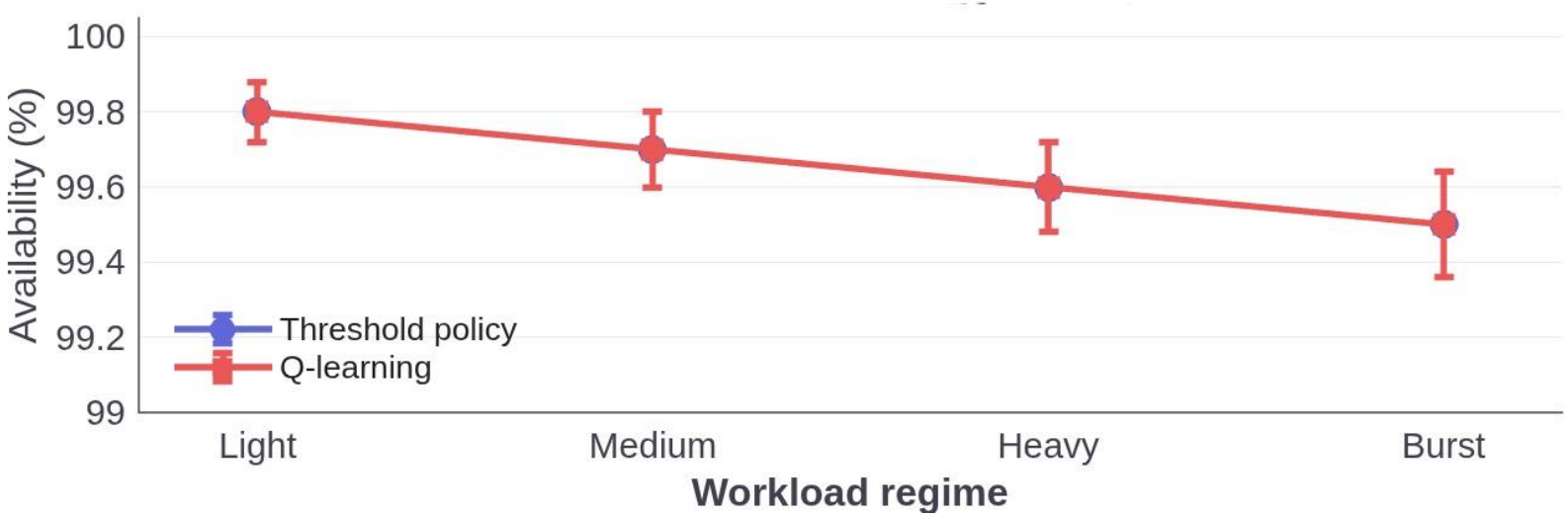


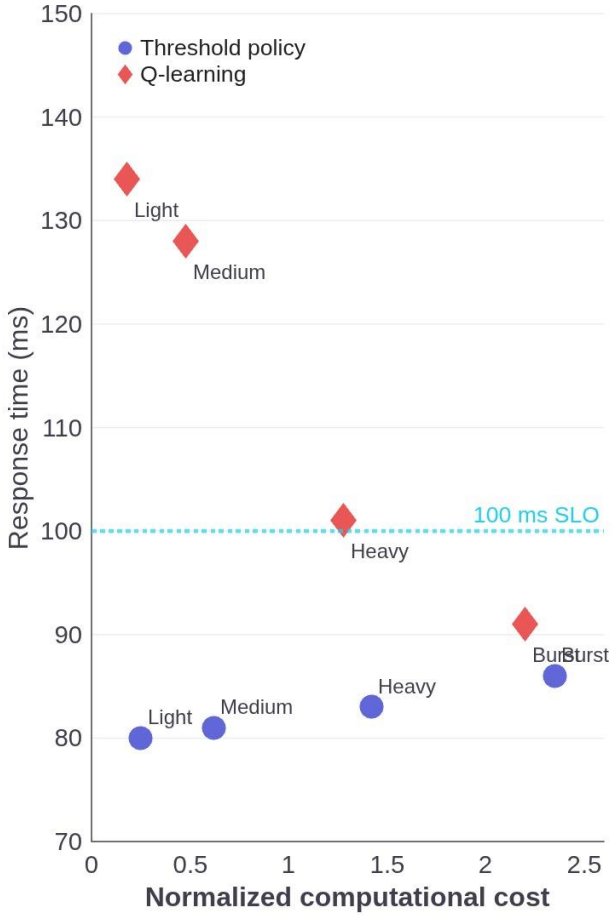
Scenario	Workload (requests/s)	Interpretation
Light	1,000	Low demand
Medium	3,000	Moderate demand
Heavy	7,000	High sustained demand
Burst	12,000	High-intensity demand


Response time (ms)
140
120
100
80
Threshold policy
Q-learning
100 ms SLO
Light
Medium
Heavy
Burst
Workload regime


Throughput (requests/s)
10000
5000
0
Light
Medium
Heavy
Burst
Workload regime
Threshold policy
Q-learning


Normalized computational cost
2.5
2
1.5
1
0.5
0
Threshold policy
Q-learning
0.25
0.18
0.62
0.48
1.42
1.28
2.35
2.20
Light
Medium
Heavy
Burst
Workload regime


100
99.8
99.6
99.4
99.2
99
Availability (%)
Threshold policy
Q-learning
Light
Medium
Heavy
Burst
Workload regime


Threshold policy
Q-learning
Light
Medium
Heavy
100 ms SLO
Burst
Heavy
Medium
Light
Response time (ms)
Normalized computational cost
150
140
130
120
110
100
90
80
70
0
0.5
1
1.5
2
2.5

Results

- The compromise depends on the **load**
- **Q-learning** reduces the number of replicas and costs (low/medium regime)
- The cost of resource scarcity is rising (heavy/burst)
- $\min C \quad \text{subject to} \quad RT \leq RT_{\text{SLO}}, \qquad T \geq T_{\text{SLO}}.$
- RL for balancing performance and cost
- The weakness of Q-Learning under low load is explained by the properties of the baseline solution

THANK YOU!



alexandra.a.chudinova@mail.ru