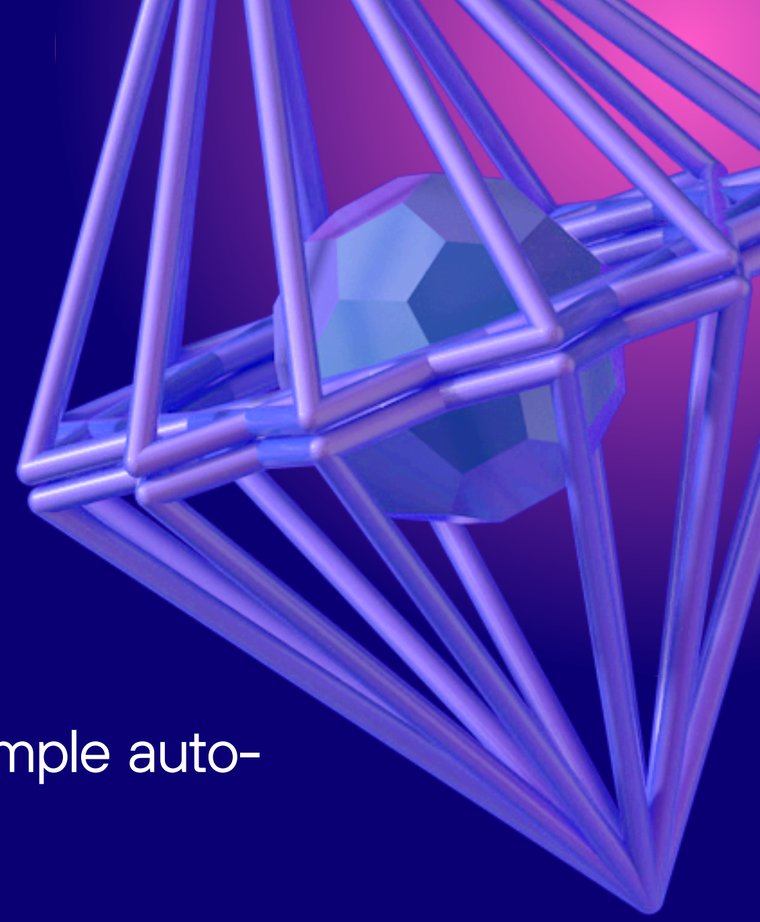
CLOUD EFFICIENCY AGENT

Aiswarya K B22CS028
Viswanadhapalli Sujay B22CS063
Yesha Shah B22CS067

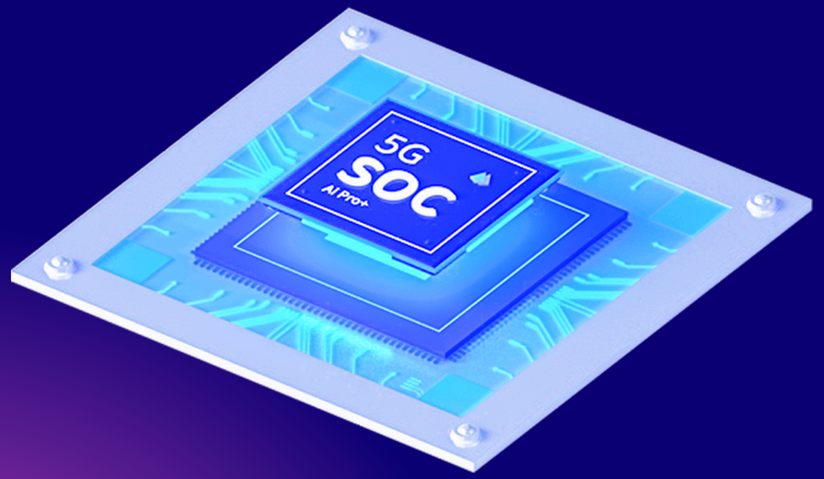
PROBLEM STATEMENT



Cloud computing environments face highly dynamic workloads.
Traditional resource allocation methods (like static thresholds or simple auto-scaling)
often lead to resource underutilization or performance bottlenecks.

We address this with an AI-based agent that dynamically allocates resources to Virtual Machines (VMs) to optimize efficiency and reduce operational cost.

OBJECTIVE



To design and simulate a reinforcement learning agent
that learns to adjust VM resource allocation in real time
based on workload fluctuations.

GOAL

- Maximize resource utilization
- Minimize over-provisioning and cost
- Improve overall system performance

LITERATURE REVIEW

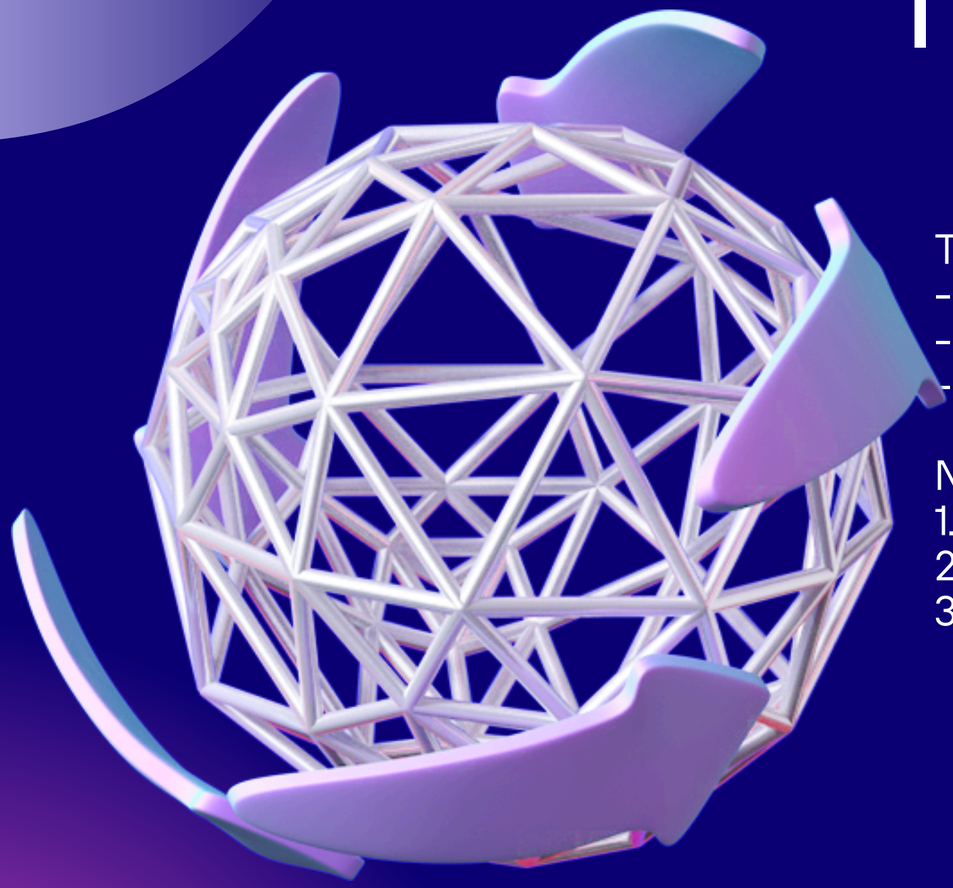
1. AI-Powered Dynamic Optimization of Cloud Resource Allocation
   [ResearchGate]

(https://www.researchgate.net/publication/387724349_AI-Powered_Dynamic_Optimization_of_Cloud_Resource_Allocation)
   → RL-based cloud scaling, cost efficiency, dynamic behavior modeling

2. Node Failure Prediction in Cloud Service Systems
   [Paper Link](https://hongyujohn.github.io/NodeFailures.pdf)
   → Predictive models for system reliability and resource redundancy

Existing approaches:
- Static scaling → Poor adaptation
- Manual tuning → Inefficient at scale
Our RL agent addresses both limitations

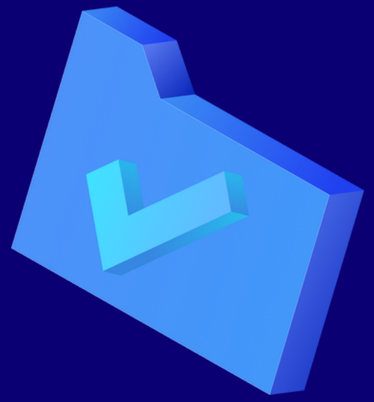
IMPLEMENTATION



Tools & Platforms:
- Python, TensorFlow, NumPy
- PPO (Proximal Policy Optimization)
- Custom simulation environment (similar to OpenAI Gym)

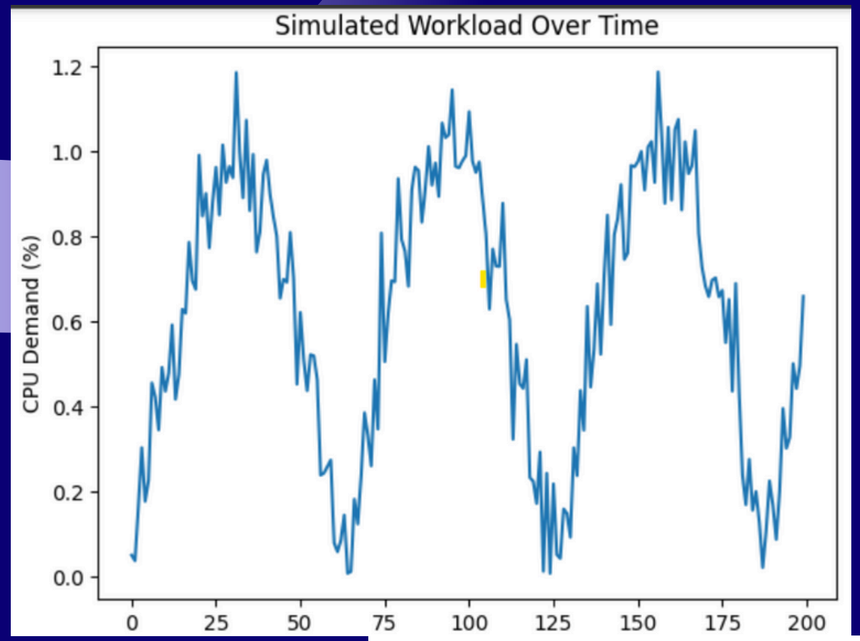
Notebooks:
1. PPO.ipynb – Core reinforcement learning model
2. VCC_PROJECT.ipynb – Cloud resource simulation
3. Node_Failure_Prediction.ipynb – Adds fault-awareness to the agent

RESULTS

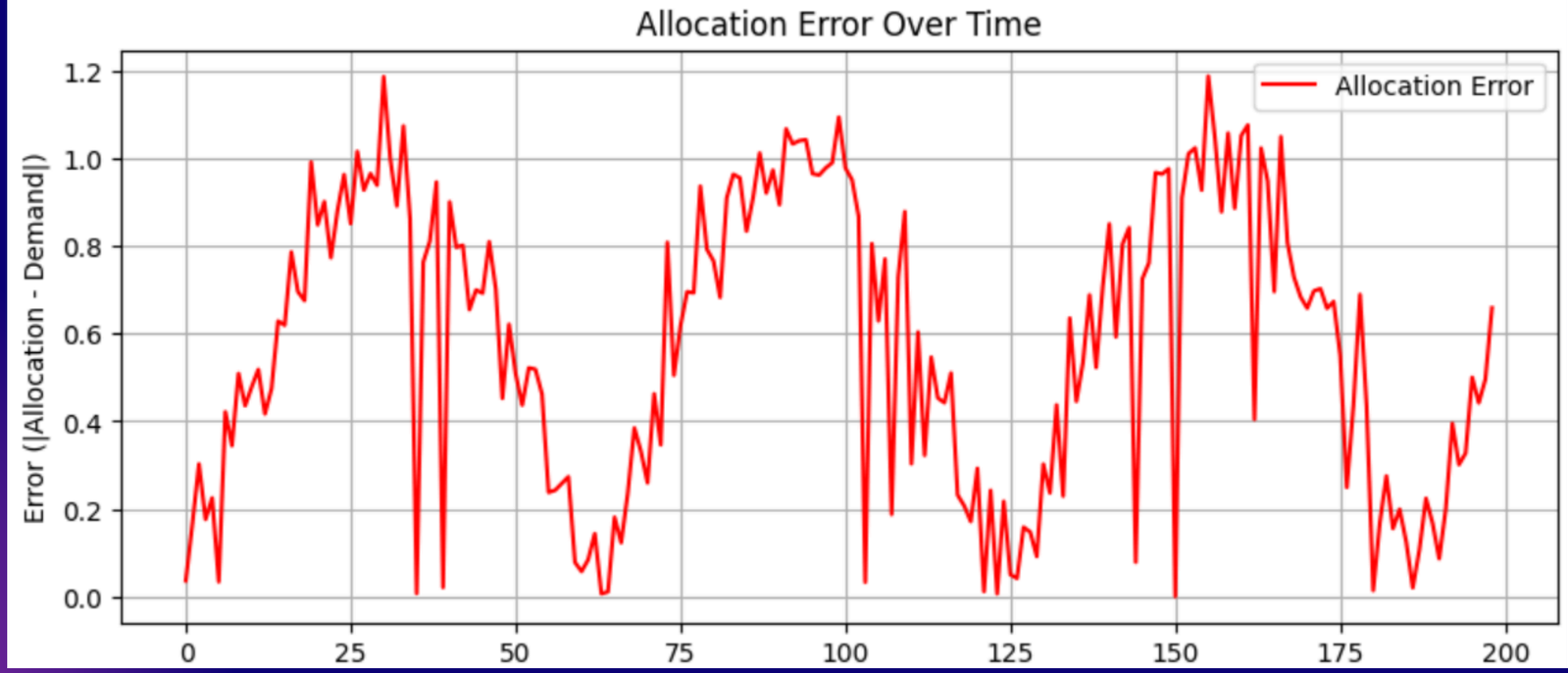


Key Observations:
- Resource usage optimized (↑ CPU utilization from ~45% to ~75%)
- Cost reduced (↓ over-provisioning incidents by ~30%)
- Agent stabilized actions over time

Metrics:
- Avg Reward: ↑ with training
- Cost Deviation: ↓ over episodes
- Uptime stability: ↑ in fault-aware simulations


Simulated Workload Over Time
CPU Demand (%)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200


Agent Reward Over Time
Reward
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
−1.50
0
25
50
75
100
125
150
175
200
Timestep
Reward per Step


Allocation Error Over Time
Error (|Allocation - Demand|)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200
Allocation Error

COMPARATIVE ANALYSIS

Feature	MING	Wang & Yang	Kaipu
Primary Goal	Node failure prediction	Workload-aware scheduling	Self-adaptive resource control
Forecasting Method	LSTM (time) + RF (space)	LSTM (workload demand)	Feedforward NN
Action Mechanism	Ranking & migration	DQN scheduling	PPO scaling agent
Type of Learning	Supervised + Ranking	Supervised + RL	Reinforcement Learning
SLA Optimization	Indirect	Direct	Emergent via training
Environment Complexity	Medium	High (RL + LSTM)	Medium (RL + Predictor)

CONCLUSION & CONTRIBUTION

- Implemented and trained a reinforcement learning agent that optimizes VM resource allocation in cloud environments.
- Incorporated fault prediction for reliability.
- Achieved better utilization, reduced cost, and improved performance.
- Project demonstrates practical potential of RL in cloud management.

THANK YOU