Control Plane

5) Action

4) Flatten Observation

7) Reward

Reward

3) Predicted CPU of Nodes

3) Cluster State

3) YOLO Profiled Data

3) Cluster State

6) Container Placement and Model Switch Action in GKE/ Simulated Testing

6) Container Placement and Model Switch Action in Training

Kubernetes Cluster Test

Node 1

Node 2

Node 3

Node 4

Flask

8) Public Traffic

IoT Layer

Gymnasium Environment

Kubernetes Cluster Simulation

PatchTST

1) Nodes CPU

Positional Encoding

2a) Patching

2a) Vectors

2a) Transformers

YOLO

1) YOLO Data Retrieval

Time Series Patching

Vectors

Transformers Core

Kubernetes Cluster

Gymnasium Environment

DRL Agent Reward

DRL Action to be implemented

IoT Users

YOLO Features

Cluster State Features

Flask API

Heterogeneous Pods

YOLO Hosted Versions

Gymnasium Flow