universität
innsbruck

HyPA: A Hybrid Pod Autoscaler

Autoscaling of container-orchestration environments for enhanced VoIP performance

Dominik Gratz, René Hueber

Supervisors: Zahra Najafabadi Samani, PhD
Juan Aznar Poveda, PhD

Cooperation with World-Direct

- Subsidiary company of A1
- Manages over 90.000 VoIP ports
- Transitions its telephone infrastructure to **Kubernetes**
  - Microservices
  - API-first approach
  - **Zero-Downtime** architecture

Cooperation with World-Direct

- Subsidiary company of A1
- Manages over 90.000 VoIP ports
- Transitions its telephone infrastructure to **Kubernetes**
  - Microservices
  - API-first approach
  - **Zero-Downtime** architecture

Pros

- Less complexity on the client side
- Central maintenance
- High flexibility and **scalability**

Cooperation with World-Direct

- Subsidiary company of A1
- Manages over 90.000 VoIP ports
- Transitions its telephone infrastructure to **Kubernetes**
  - Microservices
  - API-first approach
  - **Zero-Downtime** architecture

Pros

- Less complexity on the client side
- Central maintenance
- High flexibility and **scalability**

Cons

- Complex infrastructure
- Efficient architectures necessary
- **Timely scaling**

Problem formulation: Scaling realtime applications

- Challenges in VoIP:
  - **Stateful** protocols
  - Time-sensitive signaling
  - Sessions over a long period of time

Problem formulation: Scaling realtime applications

- Challenges in VoIP:
  - **Stateful** protocols
  - Time-sensitive signaling
  - Sessions over a long period of time
- High call volume traffic:
  - Increased load on the telephone system core
  - Increased latency
  - Call failures

Problem formulation: Scaling realtime applications

- Challenges in VoIP:
  - **Stateful** protocols
  - Time-sensitive signaling
  - Sessions over a long period of time
- High call volume traffic:
  - Increased load on the telephone system core
  - Increased latency
  - Call failures
- **Default Kubernetes** scaling is not enough:
  - Static thresholds
  - Limited option for custom parameters
  - No hybrid scaling approach

Thesis goal

HyPA

- **Hy**brid scaling:
  - Horizontal $\rightarrow$ variable number of replicas (pods)
  - Vertical $\rightarrow$ variable CPU/memory assignment of a pod
- **P**od **A**utoscaling:
  - Automatically scale service pods at runtime

Thesis goal

HyPA

- **Hy**brid scaling:
  - Horizontal $\rightarrow$ variable number of replicas (pods)
  - Vertical $\rightarrow$ variable CPU/memory assignment of a pod
- **P**od **A**utoscaling:
  - Automatically scale service pods at runtime

Challenges

- Maintain high call throughput with small latency
- Resource conservation
- Ensure no service downtime

Proposed Model

Overview

- RL learning approach
- Deployed in customer namespace
- Baseline model
- Focuses on CPU scaling

Proposed Model

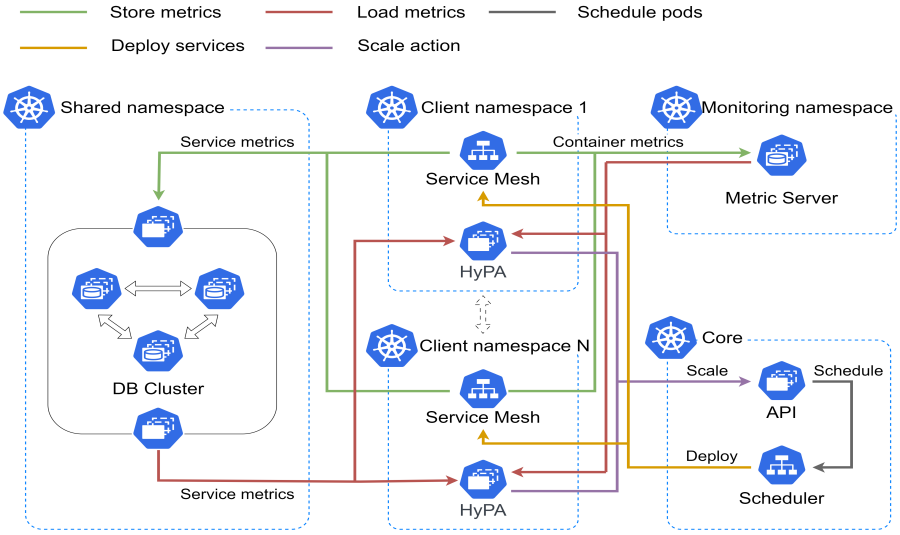
Overview

- RL learning approach
- Deployed in customer namespace
- Baseline model
- Focuses on CPU scaling

Model complexity reductions

- No vertical memory scaling
- Discrete finite action space

Infrastructure Model



Model Training (1)

Call Data

- No existing datasets
- Analyzed historic call data
- Cover **all ranges** of clients
- Train a baseline model

Model Training (1)

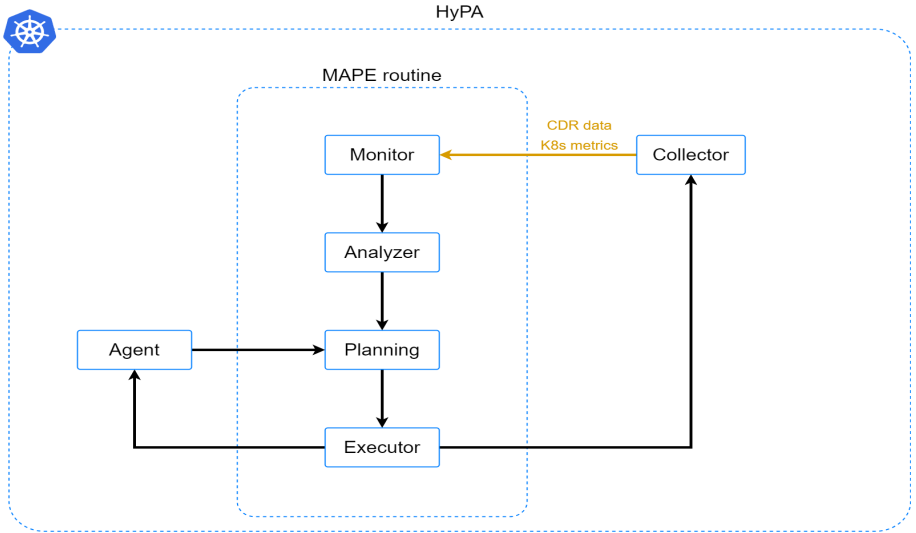
Call Data

- No existing datasets
- Analyzed historic call data
- Cover **all ranges** of clients
- Train a baseline model

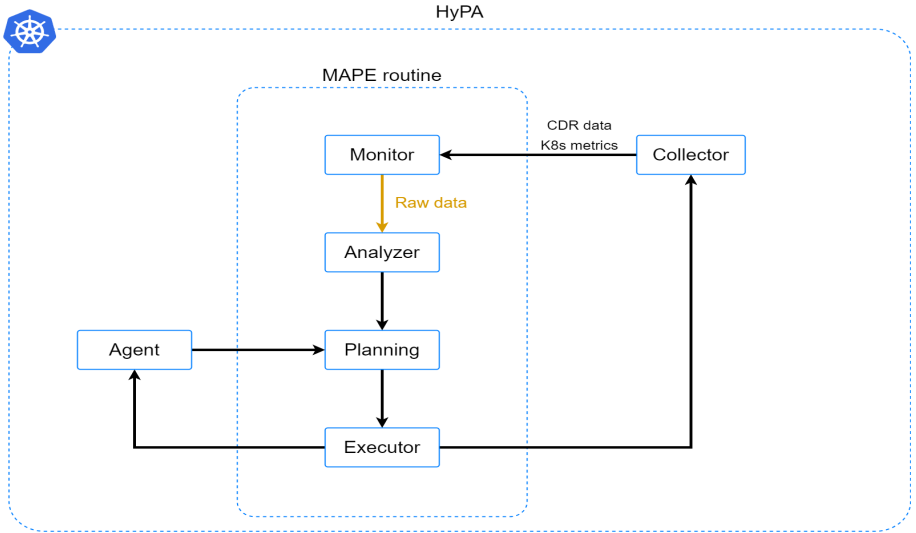
Call Generation

- Based on real call patterns
- Custom scenarios utilizing **SIPp**

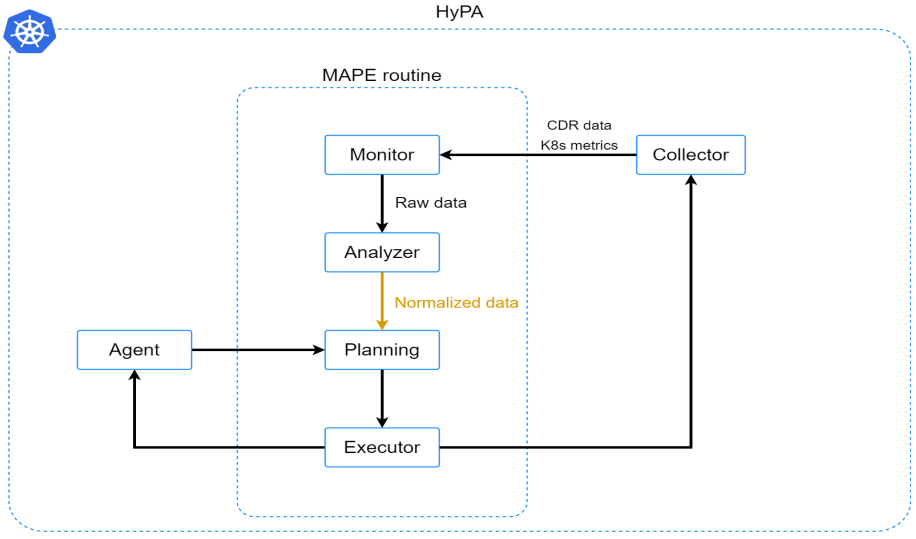
Model Training (2)



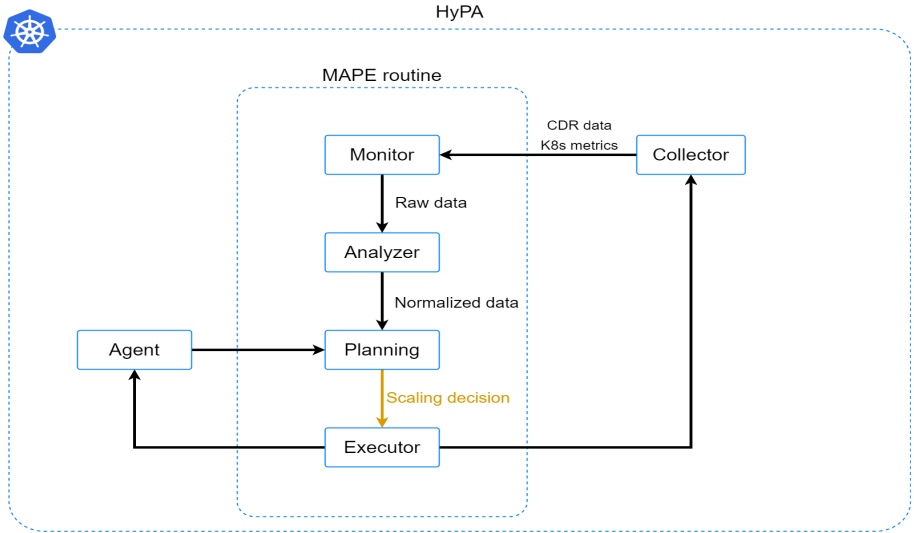
Model Training (3)



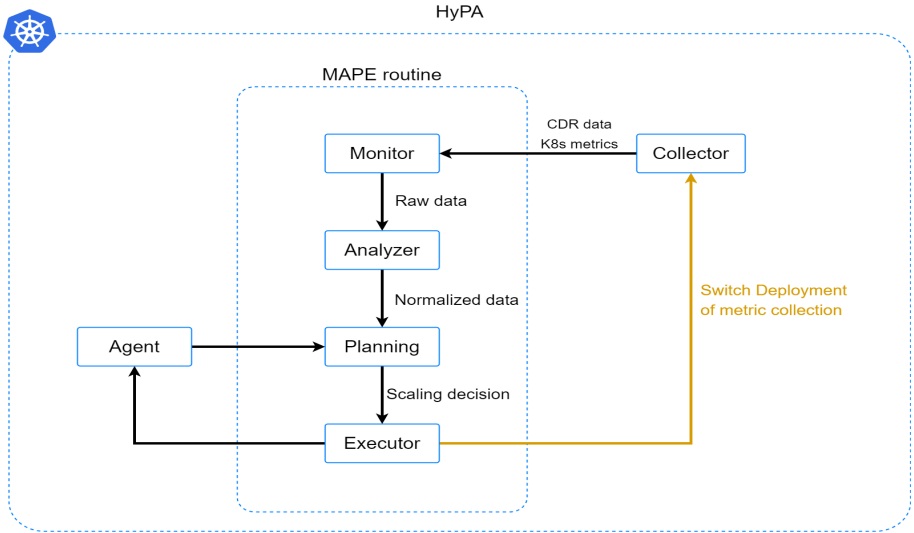
Model Training (4)



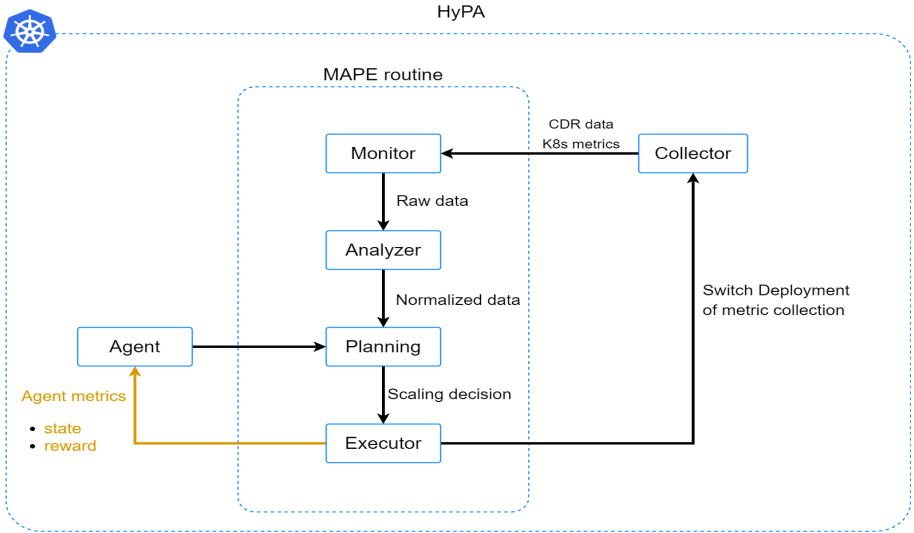
Model Training (5)



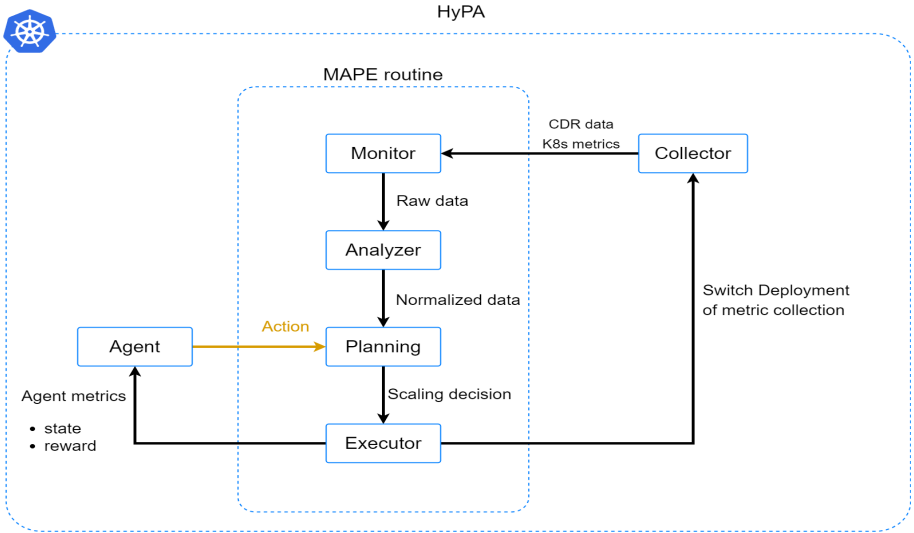
Model Training (6)



Model Training (7)



Model Training (8)



RL Training Disadvantage

Training duration

- Time-consuming tasks:
  - Scaling operation
  - Environment response
  - Data collection
- Significant delay after every agent decision (step)
- Limits training rate of model

RL Training Disadvantage

Training duration

- Time-consuming tasks:
  - Scaling operation
  - Environment response
  - Data collection
- Significant delay after every agent decision (step)
- Limits training rate of model

Problem

- Development and retraining time consuming
- Significant training duration reduction needed

RL Training Optimization

Analytical Model

- Finite scaling options covering WD’s cases
- Deploying all scaling options
- Same workload everywhere
- Only metric collection switched

RL Training Optimization

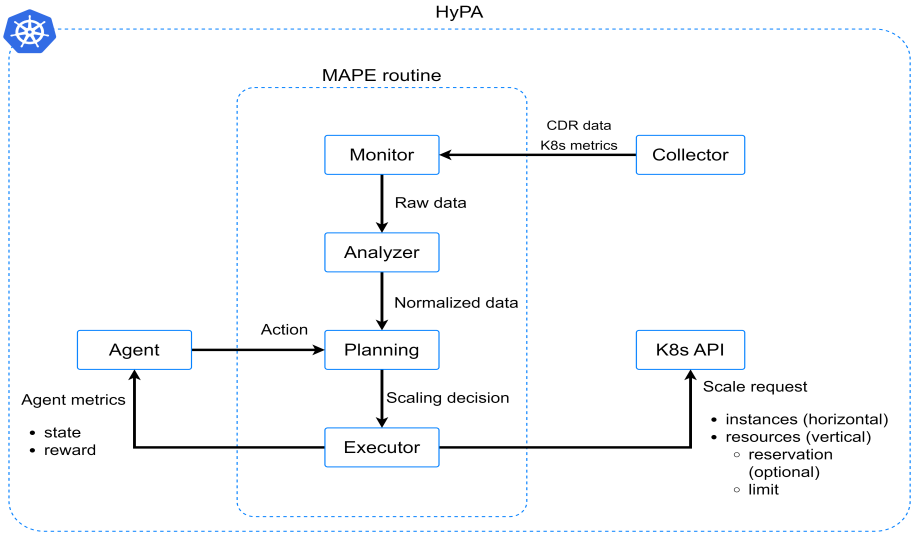
Analytical Model

- Finite scaling options covering WD’s cases
- Deploying all scaling options
- Same workload everywhere
- Only metric collection switched

Improvement

- Faster environment response
- Broader agent decision exploration
- Training three times faster

Evaluation HyPA



Evaluation Competitors

Horizontal Pod Autoscaler (HPA)

- Kubernetes default
- Threshold based
- Reactive autoscaler
- No expert knowledge for setup

Evaluation Competitors

Horizontal Pod Autoscaler (HPA)

- Kubernetes default
- Threshold based
- Reactive autoscaler
- No expert knowledge for setup

Multi-Objective-Hybrid-Autoscaling (MOHA)

- Machine Learning based (NN, SVM, LR)
- Hybrid scaling
- Code modifications to support usecase

Evaluation Scenarios

Call Volume Groups

Group	Call volume	Client share
light	$< 10^4$	31.99 %
medium	$10^4 - 10^5$	53.02 %
heavy	$10^5 - 10^6$	14.34 %
very heavy	$\geq 10^6$	0.65 %

Evaluation Scenarios

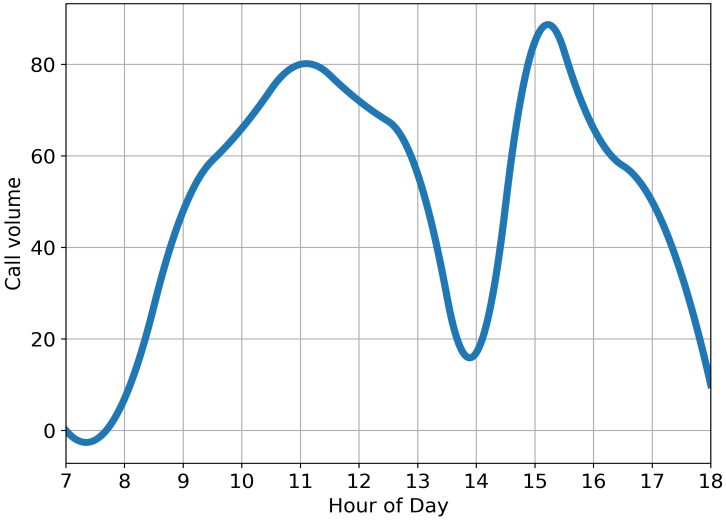
Call Volume Groups

Group	Call volume	Client share
light	$< 10^4$	31.99 %
medium	$10^4 - 10^5$	53.02 %
heavy	$10^5 - 10^6$	14.34 %
very heavy	$\geq 10^6$	0.65 %

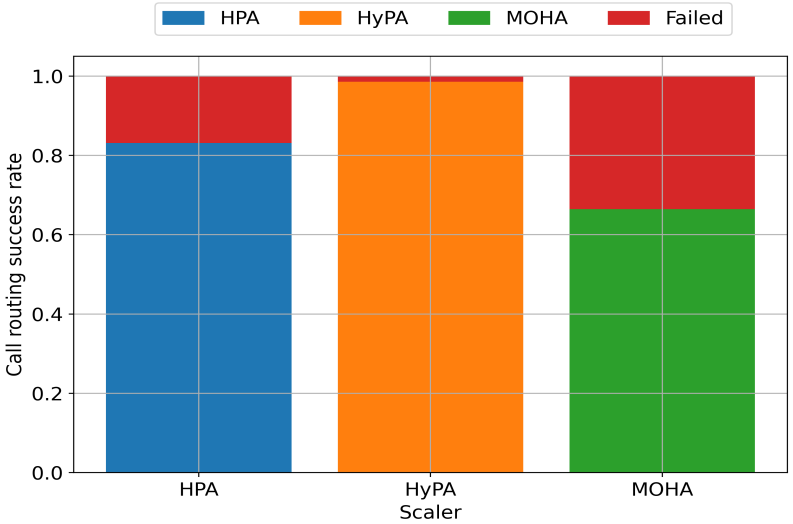
Scenarios

1. Random common client
2. Averaged call volume per group

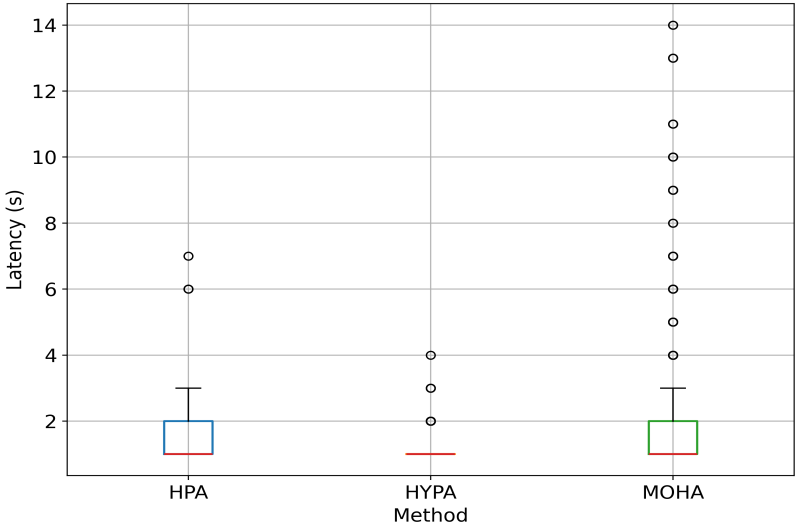
Scenario 1 Visualization



Scenario 1 Result Calls



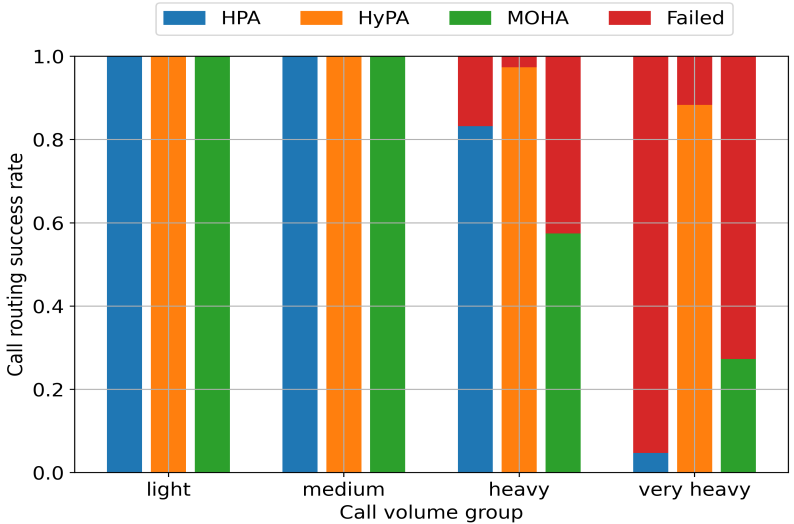
Scenario 1 Result Latency



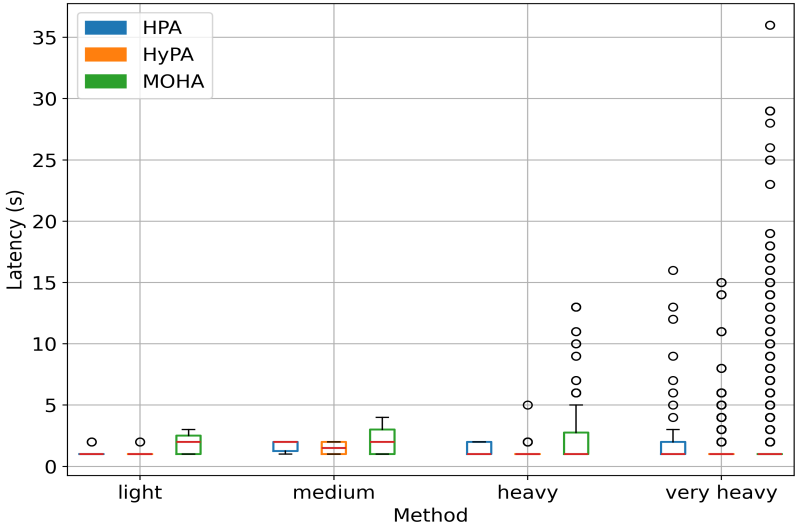
Scenario 2 Visualization

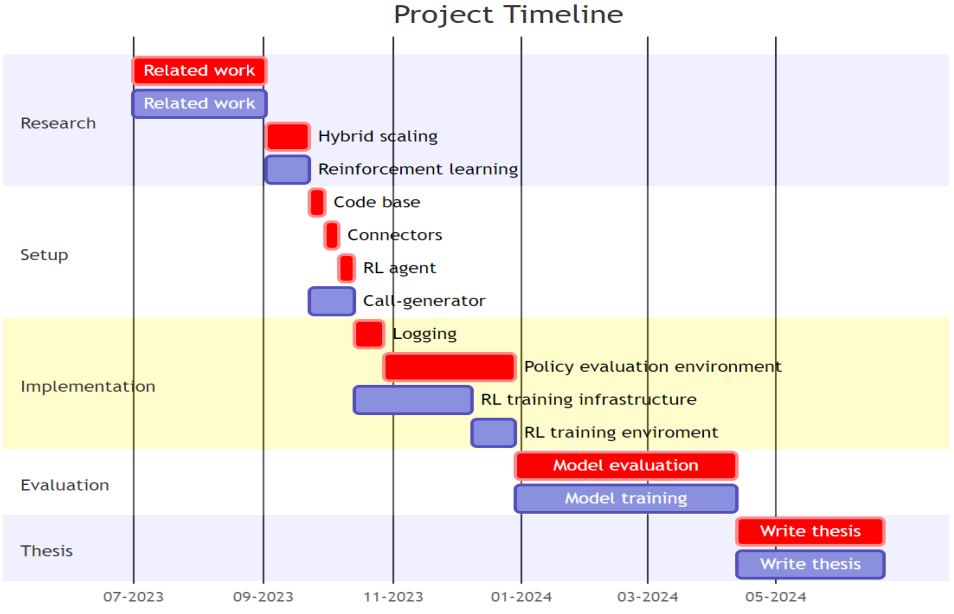
Hour	light	medium	heavy	very heavy
7	1	3	51	984
8	2	7	159	3069
9	2	9	193	4187
10	2	9	191	7196
11	2	8	175	3901
12	1	4	91	2468
13	1	6	128	3058
14	1	6	127	3025
15	1	5	109	2901
16	1	3	75	1803
17	1	2	36	726
Client share	31.99%	53.02%	14.34%	0.65%

Scenario 2 Result Calls



Scenario 2 Result Latency




Project Timeline
Research
Related work
Related work
Hybrid scaling
Reinforcement learning
Setup
Code base
Connectors
RL agent
Call-generator
Implementation
Logging
Policy evaluation environment
RL training infrastructure
RL training enviroment
Evaluation
Model evaluation
Model training
Thesis
Write thesis
Write thesis
07-2023
09-2023
11-2023
01-2024
03-2024
05-2024

References I

Achim Zeileis Reto Stauffer.
UIBK Latex Beamer Theme.