Into the Wild: Real-World Testing for ML-Based ABR

Benjamin Hoffman, Alexander Dietmüller, Ayush Mishra, Laurent Vanbever

PACMI at SOSP 2025

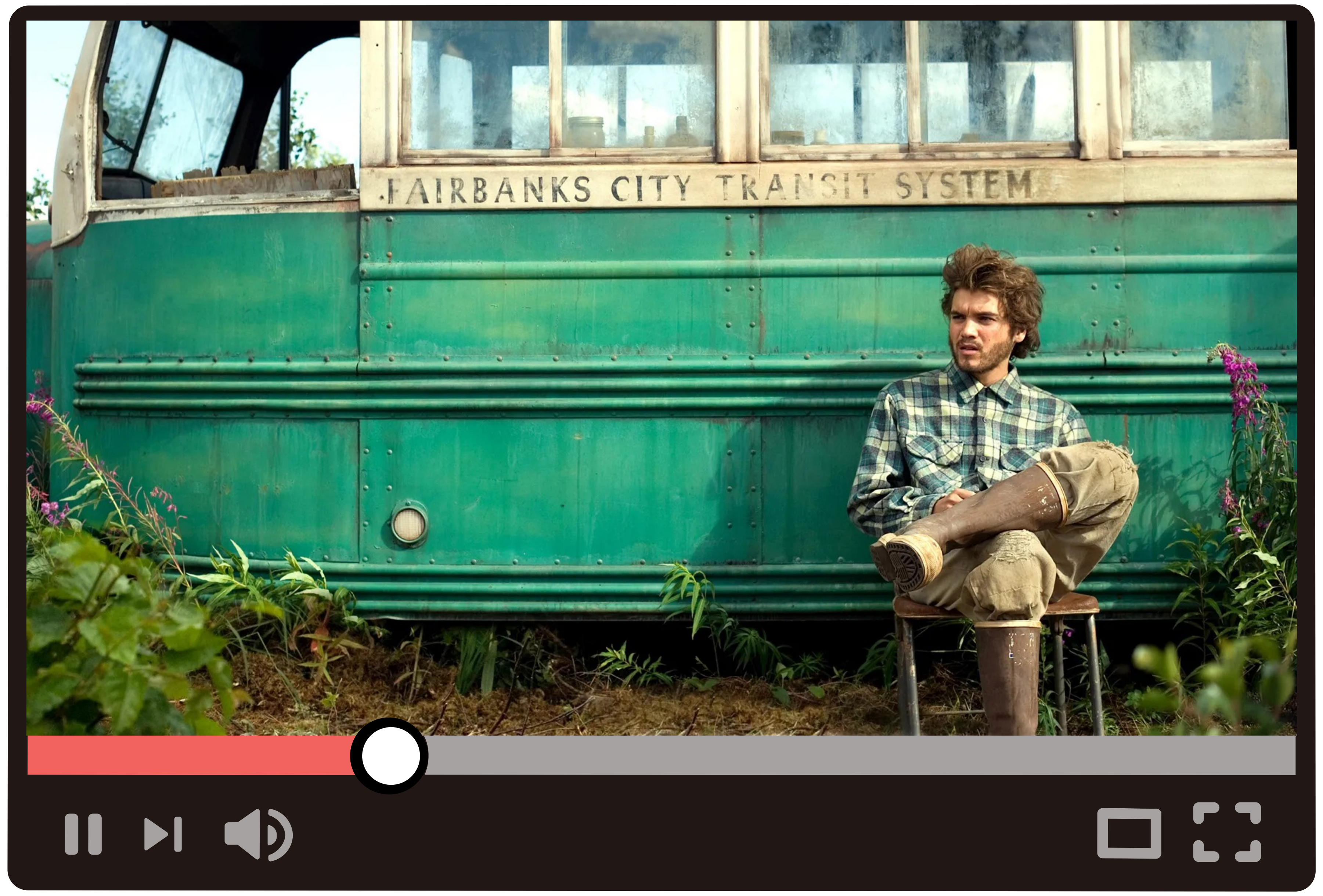
Networked Systems
ETH Zürich — seit 2015

ETH zürich

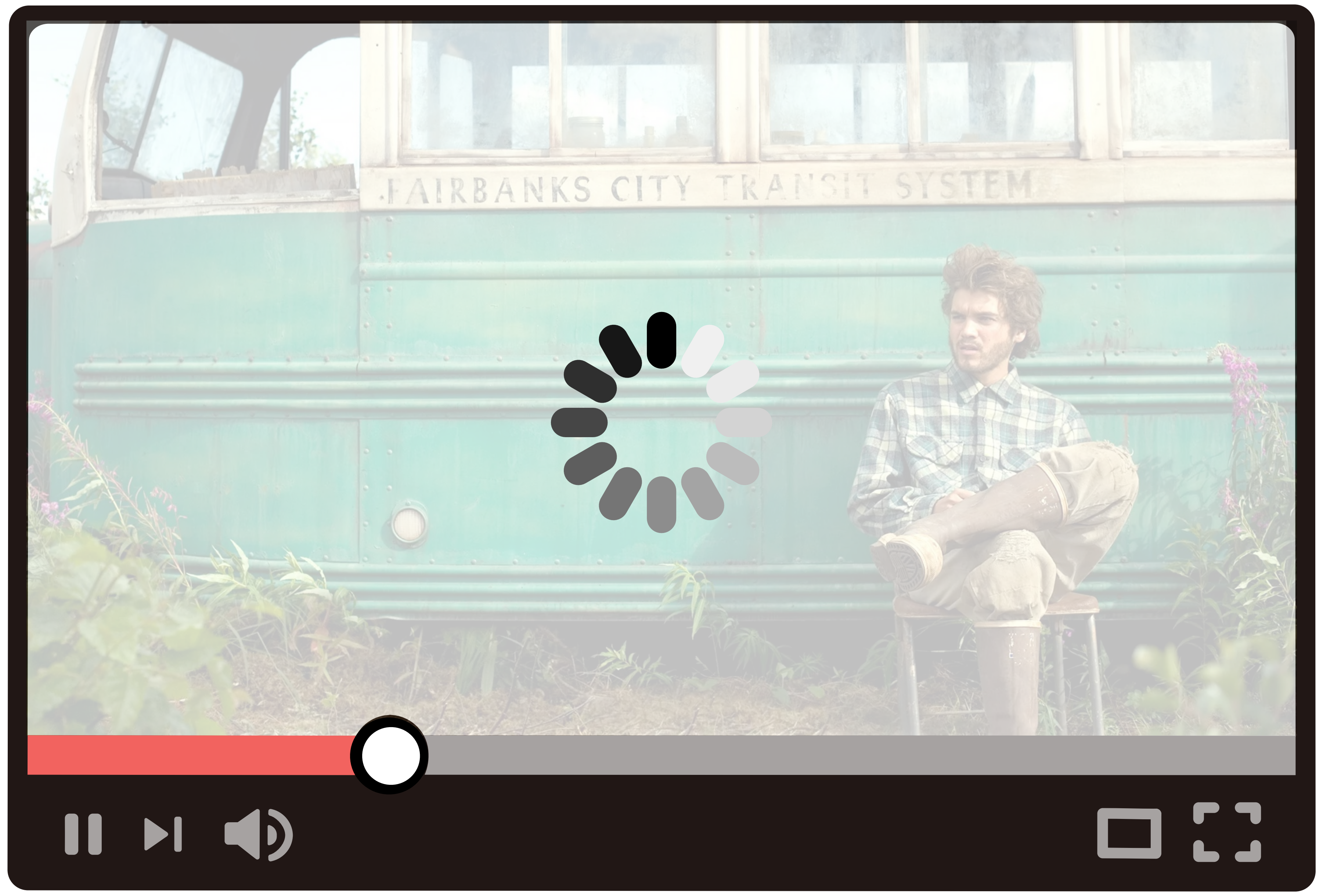
Bad video streaming…

Bad video streaming...

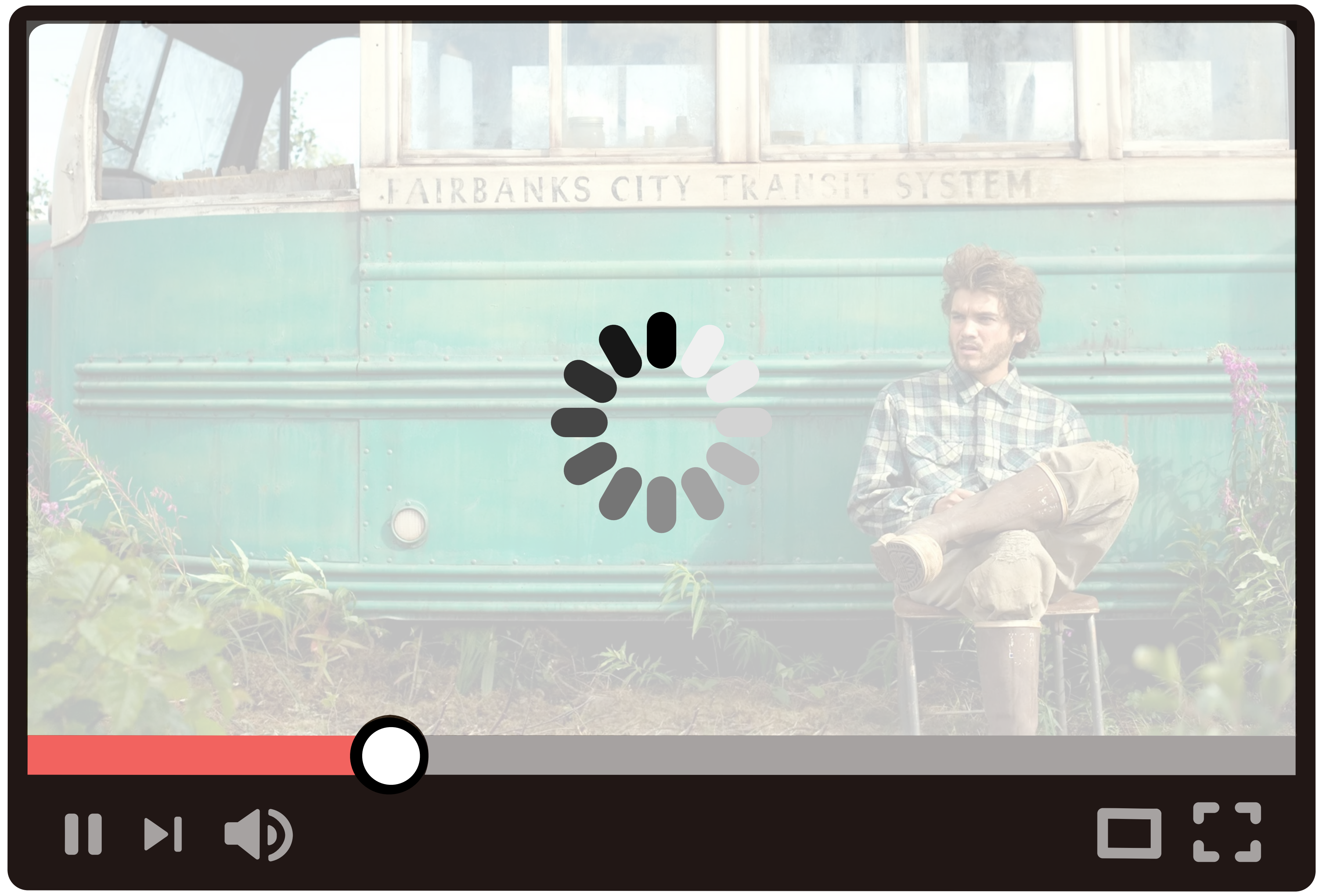
Bad video streaming...



Bad video streaming...



Bad video streaming... sucks



Why does it matter?

Why does it matter?

- Rebuffering is detrimental to user experience[1]

[1] A. Dietmüller, et al., 2024

Why does it matter?

- Rebuffering is detrimental to user experience[1]
- Streaming is the Internet's most prominent workload (> 65% of downstream traffic[2])

[1] A. Dietmüller, et al., 2024 | [2] Sandvine Corporation, 2023

Why does it matter?

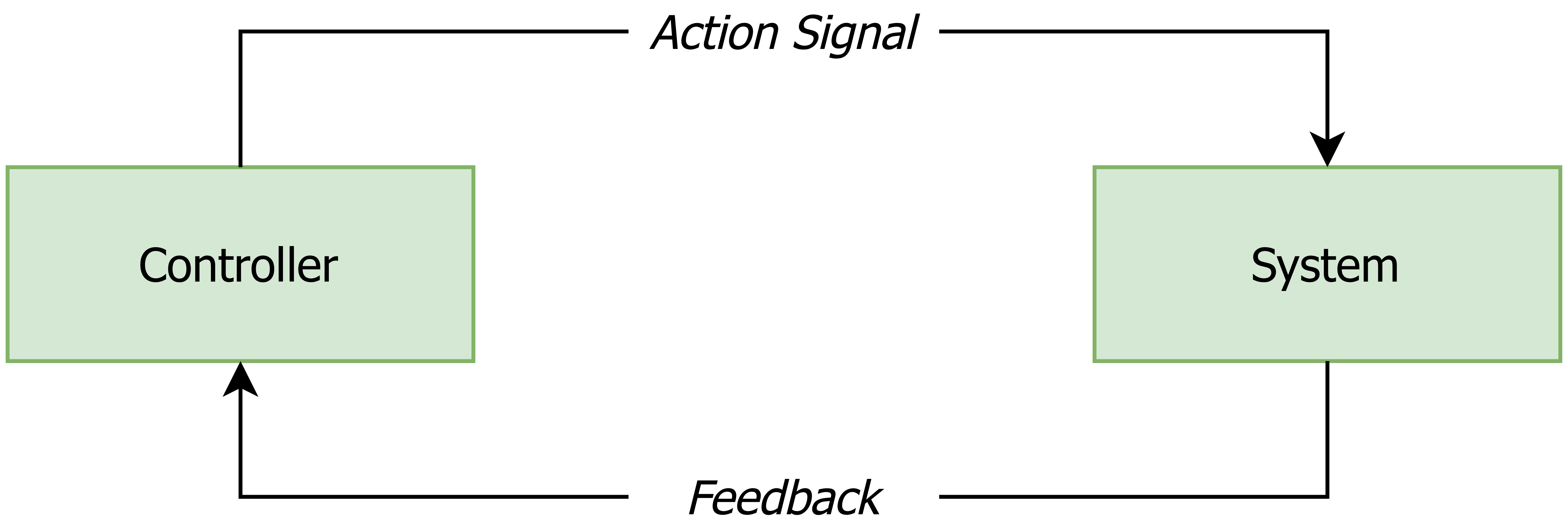
- Rebuffering is detrimental to user experience[1]
- Streaming is the Internet's most prominent workload (> 65% of downstream traffic[2])
- Providing high Quality of Experience (QoE) to users is key!

[1] A. Dietmüller, et al., 2024 | [2] Sandvine Corporation, 2023

Providing high QoE

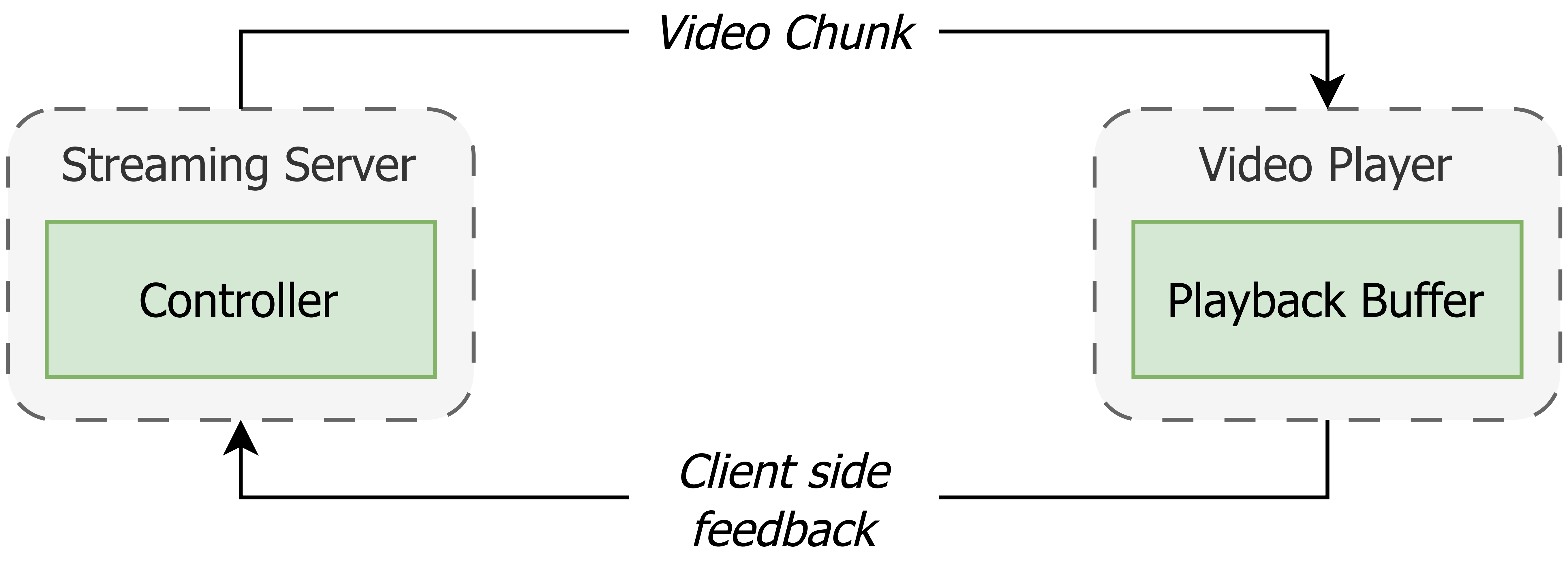
Providing high QoE

Control problem:



Providing high QoE

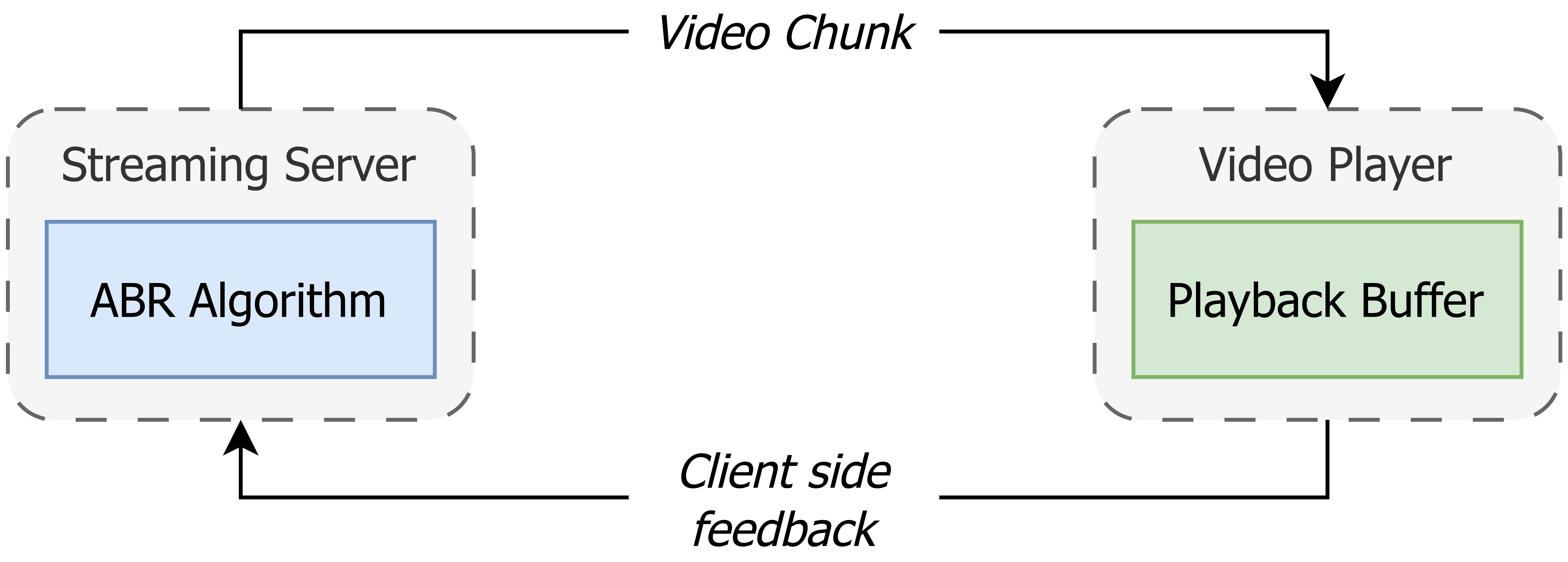
Control problem: adapt sending behavior based on changing conditions



Providing high QoE

Control problem: adapt sending behavior based on changing conditions

- Adaptive Bitrate (ABR) Algorithms



Providing high QoE

Control problem: adapt sending behavior based on changing conditions

- Adaptive Bitrate (ABR) Algorithms
- Minimize stall time while maximizing video quality

Providing high QoE

Control problem: adapt sending behavior based on changing conditions

- Adaptive Bitrate (ABR) Algorithms
- Minimize stall time while maximizing video quality
- Select appropriate bitrate after sensing or predicting network state

Providing high QoE

Control problem: adapt sending behavior based on changing conditions

- Adaptive Bitrate (ABR) Algorithms
- Minimize stall time while maximizing video quality
- Select appropriate bitrate after sensing or predicting network state
- Traditionally via buffer- and rate-based control, or control theoretic approaches

Optimizing QoE is hard

Optimizing QoE is hard

Optimizing QoE over the Internet is challenging:

Optimizing QoE is hard

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space

Optimizing QoE is hard

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space with increasing complexity over time

ML to the rescue

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space with increasing complexity over time

→ *Research turns to learning-based methods that enable:*

ML to the rescue

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space with increasing complexity over time

→ *Research turns to learning-based methods that enable:*

- Handling high-dimensional modeling spaces

ML to the rescue

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space with increasing complexity over time

→ *Research turns to learning-based methods that enable:*

- Handling high-dimensional modeling spaces and rapidly processing data

ML to the rescue

Optimizing QoE over the Internet is challenging:

- high-dimensional modeling space with increasing complexity over time

→ *Research turns to learning-based methods that enable:*

- Handling high-dimensional modeling spaces and rapidly processing data
- Learning from past experience instead of tuning heuristically

ML to the rescue

And using learning-based methods shows promise!

CS2P: Improving Video Bitrate Selection and Adaptation with Data-Driven Throughput Prediction

Yi Sun[⊗], Xiaoqi Yin[†], Junchen Jiang[†], Vyas Sekar[†]
Fuyuan Lin[⊗], Nanshu Wang[⊗], Tao Liu[⋄], Bruno Sinopoli[†]
[⊗] ICT/CAS, [†] CMU, [⋄] iQIYI
{sunyi, linfuyuan, wangnanshu}@ict.ac.cn, yinxiaoqi522@gmail.com,
junchenj@cs.cmu.edu, vsekar@andrew.cmu.edu, liutao@qiyi.com,
brunos@ece.cmu.edu

ABSTRACT

Bitrate adaptation is critical to ensure good quality-of-experience (QoE) for Internet video. Several efforts have argued that accurate throughput prediction can dramatically improve the efficiency of (1) *initial* bitrate selection to lower startup delay and offer high initial resolution and (2) *mid-stream* bitrate adaptation for high QoE. However, prior efforts did not systematically quantify real-world throughput predictability or develop good prediction algorithms. To bridge this gap, this paper makes three contributions. First, we analyze the throughput characteristics in a dataset with 20M+ sessions. We find: (a) Sessions sharing similar key features (e.g., ISP, region) present similar initial throughput values and dynamic patterns; (b) There is a natural "stateful" behavior in throughput variability within a given session. Second, building on these insights, we develop CS2P, a throughput prediction system which uses a data-driven approach to learn (a) clusters of similar sessions, (b) an ini-

Keywords

Internet Video; TCP; Throughput Prediction; Bitrate Adaptation; Dynamic Adaptive Streaming over HTTP (DASH)

1 Introduction

There has been a dramatic rise in the volume of HTTP-based adaptive video streaming traffic in recent years [1]. Delivering good application-level video quality-of-experience (QoE) entails new metrics such as low buffering or smooth bitrate delivery [5, 22]. To meet these new application-level QoE goals, video players need intelligent bitrate selection and adaptation algorithms [27, 30].

Recent work has shown that accurate throughput prediction can significantly improve the QoE for adaptive video streaming (e.g., [47, 48, 50]). Specifically, accurate prediction can help in two aspects:

- *Initial bitrate selection:* Throughput prediction can help

Y. Sun, et al., 2016 (SIGCOMM)

Neural Adaptive Video Streaming with Pensieve

Hongzi Mao, Ravi Netravali, Mohammad Alizadeh
MIT Computer Science and Artificial Intelligence Laboratory
{hongzi,ravinet,alizadeh}@mit.edu

ABSTRACT

Client-side video players employ adaptive bitrate (ABR) algorithms to optimize user quality of experience (QoE). Despite the abundance of recently proposed schemes, state-of-the-art ABR algorithms suffer from a key limitation: they use fixed control rules based on simplified or inaccurate models of the deployment environment. As a result, existing schemes inevitably fail to achieve optimal performance across a broad set of network conditions and QoE objectives.

We propose Pensieve, a system that generates ABR algorithms using reinforcement learning (RL). Pensieve trains a neural network model that selects bitrates for future video chunks based on observations collected by client video players. Pensieve does not rely on pre-programmed models or assumptions about the environment. Instead, it learns to make ABR decisions solely through observations of the resulting performance of past decisions. As a result, Pensieve automatically learns ABR algorithms that adapt to a wide range of environments and QoE metrics. We compare Pensieve to state-of-the-art ABR algorithms using trace-driven and real world experiments spanning a wide variety of network conditions, QoE metrics, and video properties. In all considered scenarios, Pensieve outperforms the best state-of-the-art scheme, with improvements in average QoE of 12%–25%. Pensieve also generalizes well, outperforming existing schemes even on networks for which it was not explicitly trained.

content providers [12, 25]. Nevertheless, content providers continue to struggle with delivering high-quality video to their viewers.

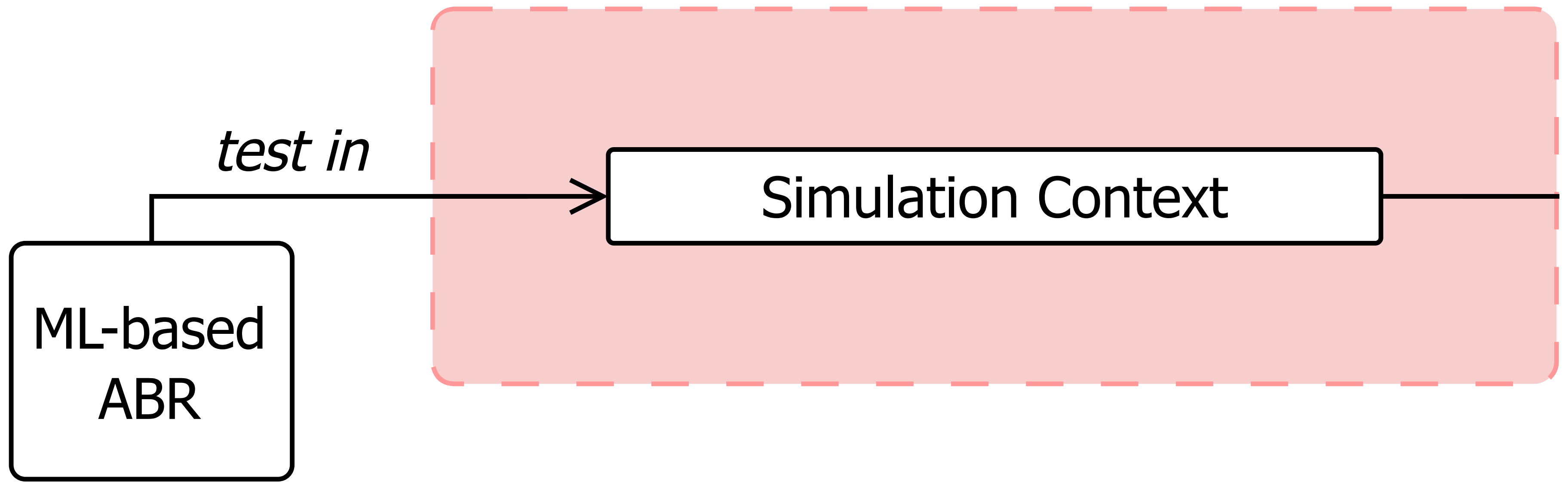
Adaptive bitrate (ABR) algorithms are the primary tool that content providers use to optimize video quality. These algorithms run on client-side video players and dynamically choose a bitrate for each video *chunk* (e.g., 4-second block). ABR algorithms make bitrate decisions based on various observations such as the estimated network throughput and playback buffer occupancy. Their goal is to maximize the user's QoE by adapting the video bitrate to the underlying network conditions. However, selecting the right bitrate can be very challenging due to (1) the variability of network throughput [18, 42, 49, 52, 53]; (2) the conflicting video QoE requirements (high bitrate, minimal rebuffering, smoothness, etc.); (3) the cascading effects of bitrate decisions (e.g., selecting a high bitrate may drain the playback buffer to a dangerous level and cause rebuffering in the future); and (4) the coarse-grained nature of ABR decisions. We elaborate on these challenges in §2.

The majority of existing ABR algorithms (§7) develop fixed control rules for making bitrate decisions based on estimated network throughput ("rate-based" algorithms [21, 42]), playback buffer size ("buffer-based" schemes [19, 41]), or a combination of the two signals [26]. These schemes require significant tuning and do not generalize to different network conditions and QoE objectives. The state-of-the-art approach, MPC [51], makes bitrate decisions by solv-

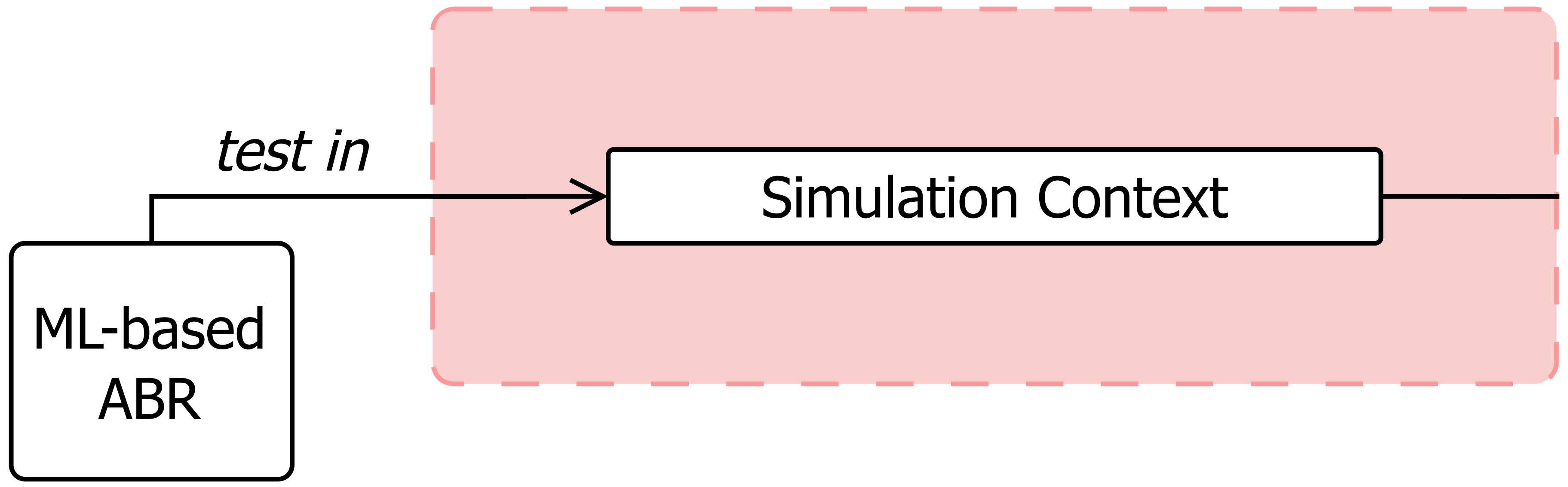
H. Mao, et al., 2017 (SIGCOMM)

ML to the rescue...?

ML to the rescue...?



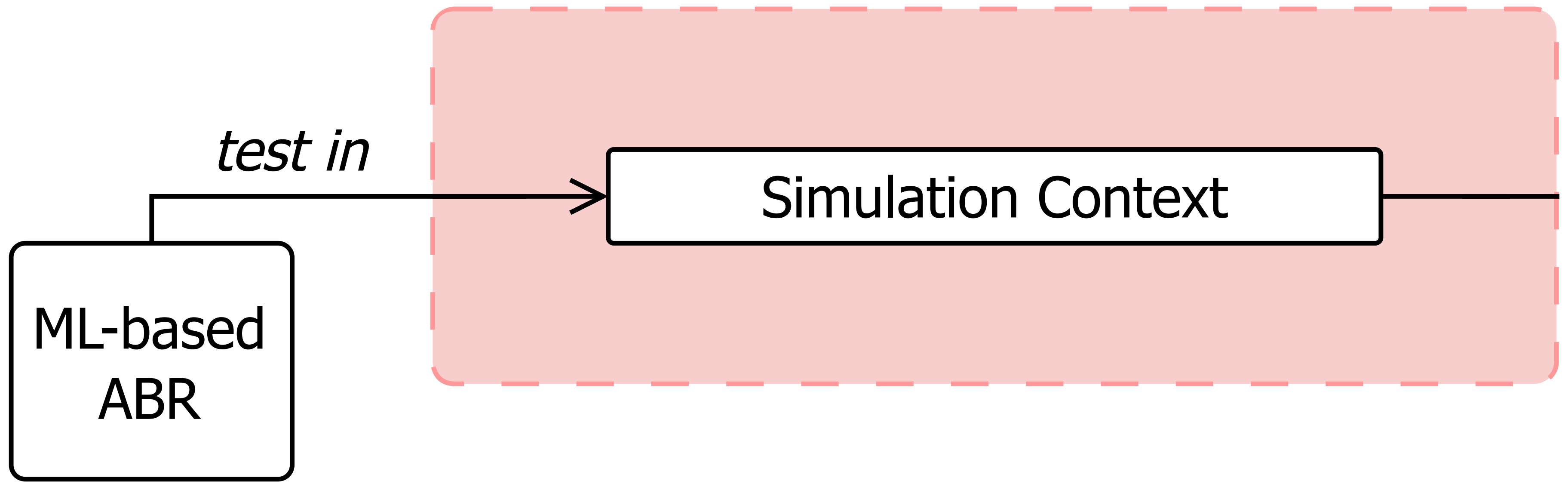
ML to the rescue...?



ML-based algorithms are failing to perform in practice[1,3]:

[1] A. Dietmüller, et al., 2024 | [3] F. Yan, et al., 2020

ML to the rescue...?

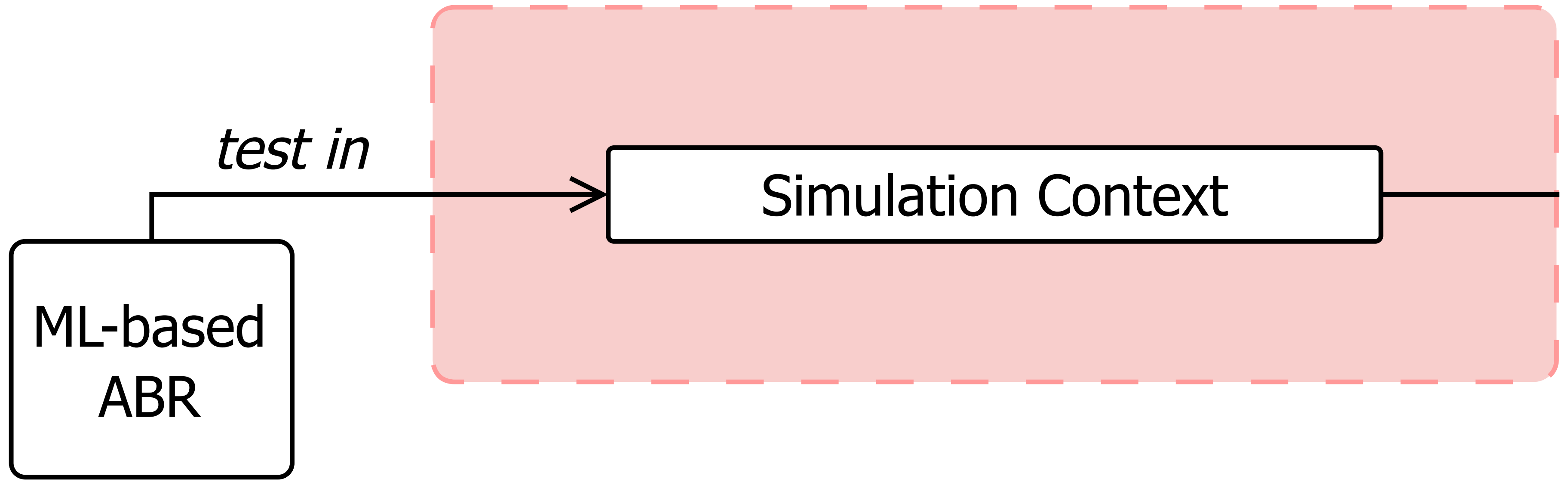


ML-based algorithms are failing to perform in practice[1,3]:

- Network conditions vary across the Internet

[1] A. Dietmüller, et al., 2024 | [3] F. Yan, et al., 2020

ML to the rescue...?

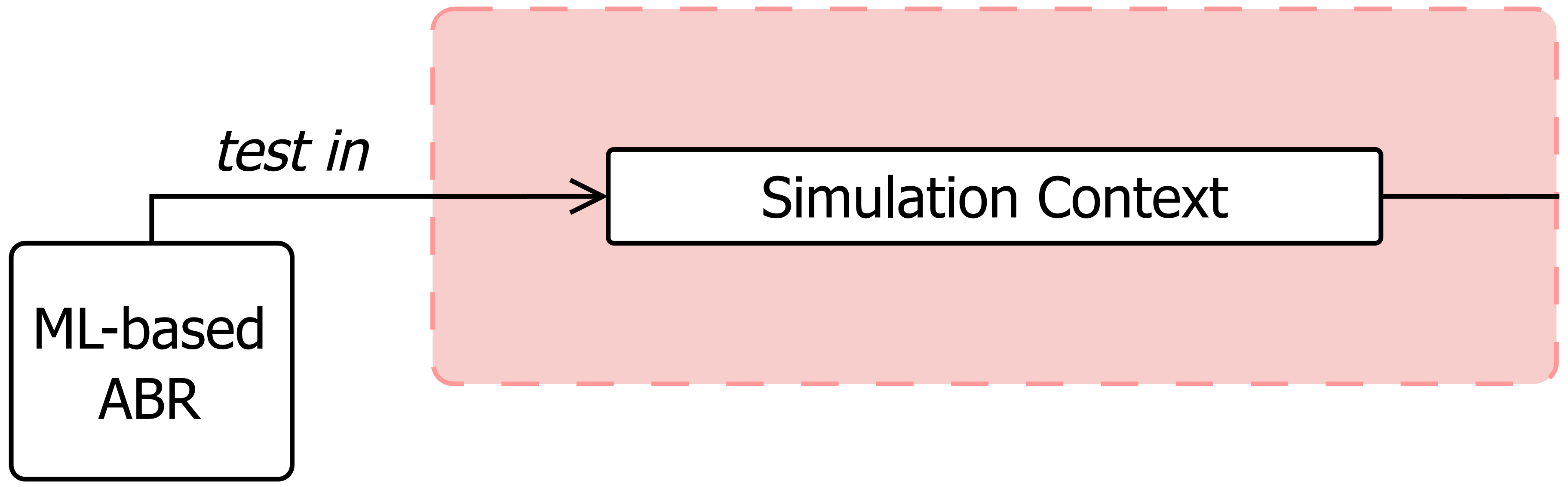


ML-based algorithms are failing to perform in practice[1,3]:

- Network conditions vary across the Internet
- Traffic and user behavior is heavy-tailed

[1] A. Dietmüller, et al., 2024 | [3] F. Yan, et al., 2020

ML to the rescue...?



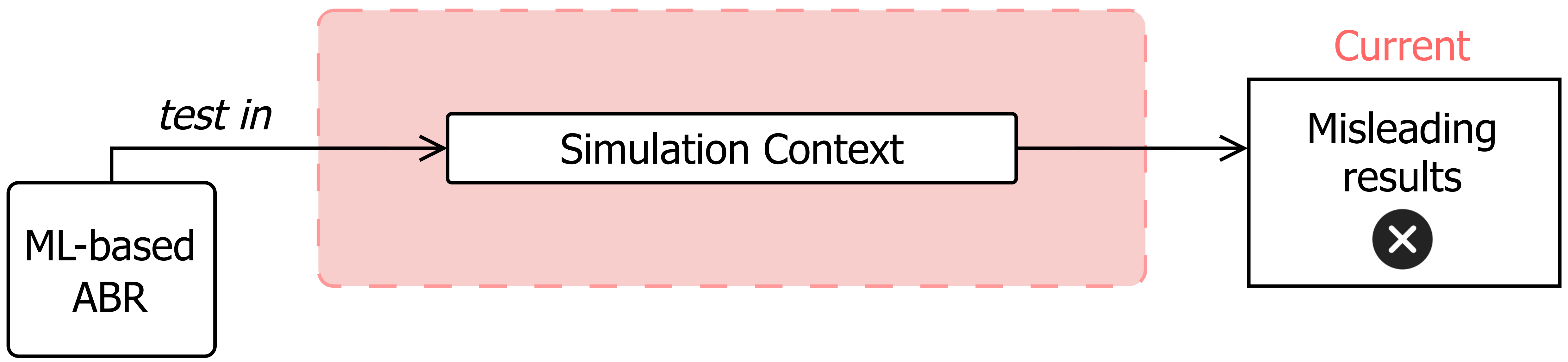
ML-based algorithms are failing to perform in practice[1,3]:

- Network conditions vary across the Internet
- Traffic and user behavior is heavy-tailed

Modeling networks is hard!

[1] A. Dietmüller, et al., 2024 | [3] F. Yan, et al., 2020

ML to the rescue...?



ML-based algorithms are failing to perform in practice[1,3]:

- Network conditions vary across the Internet
- Traffic and user behavior is heavy-tailed

Modeling networks is hard!

[1] A. Dietmüller, et al., 2024 | [3] F. Yan, et al., 2020

Testing on the Internet

Testing on the Internet

The Puffer project:

Puffer: F. Yan, et al., 2020

Testing on the Internet

The Puffer project:

- YouTube like streaming platform

Puffer: F. Yan, et al., 2020

Testing on the Internet

The Puffer project:

- YouTube like streaming platform
- Streaming live TV to users across the US

Puffer: F. Yan, et al., 2020

Testing on the Internet

The Puffer project:

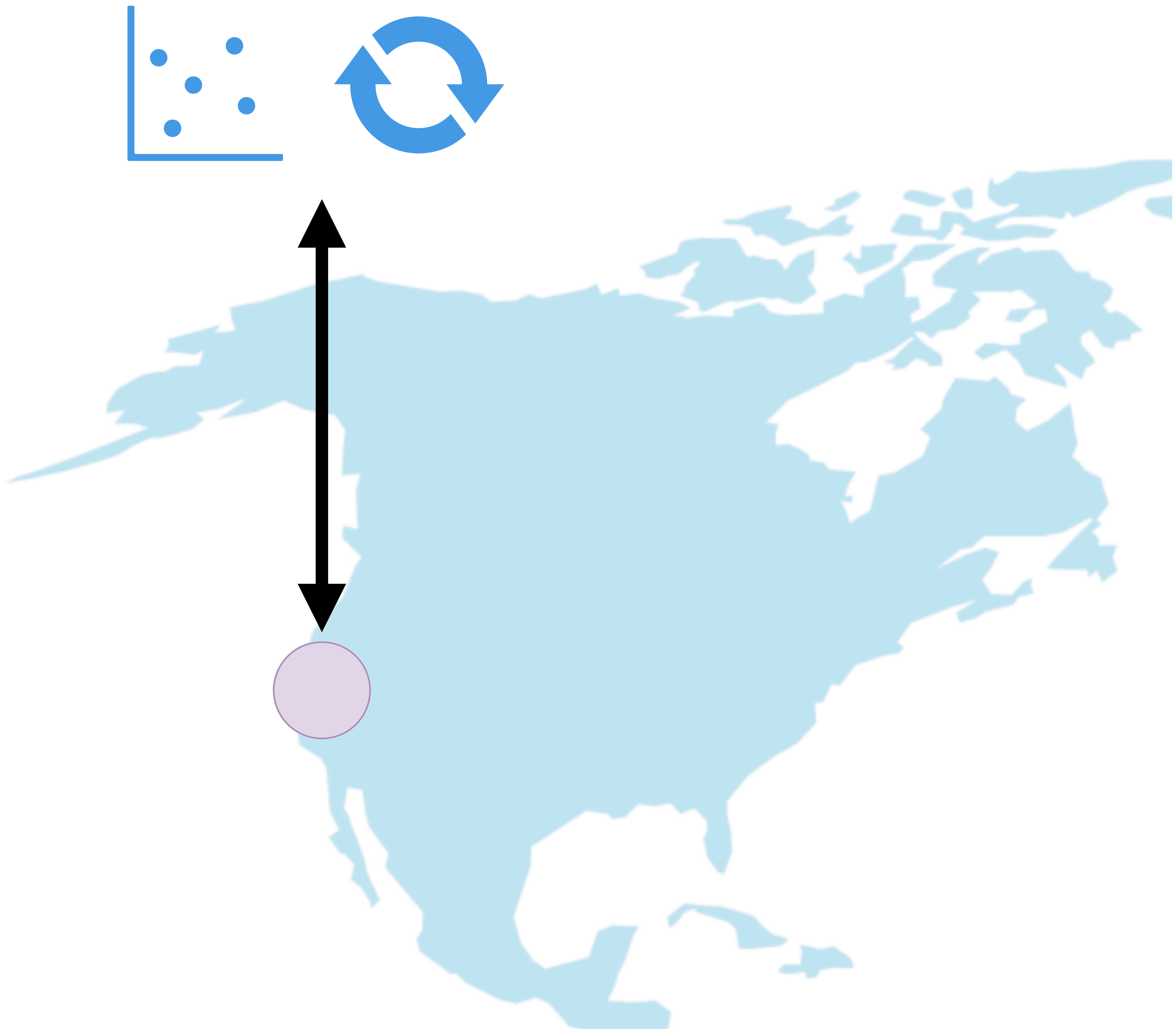
- YouTube like streaming platform
- Streaming live TV to users across the US
- Hosts various state-of-the-art ABR schemes

Puffer: F. Yan, et al., 2020

Testing on the Internet

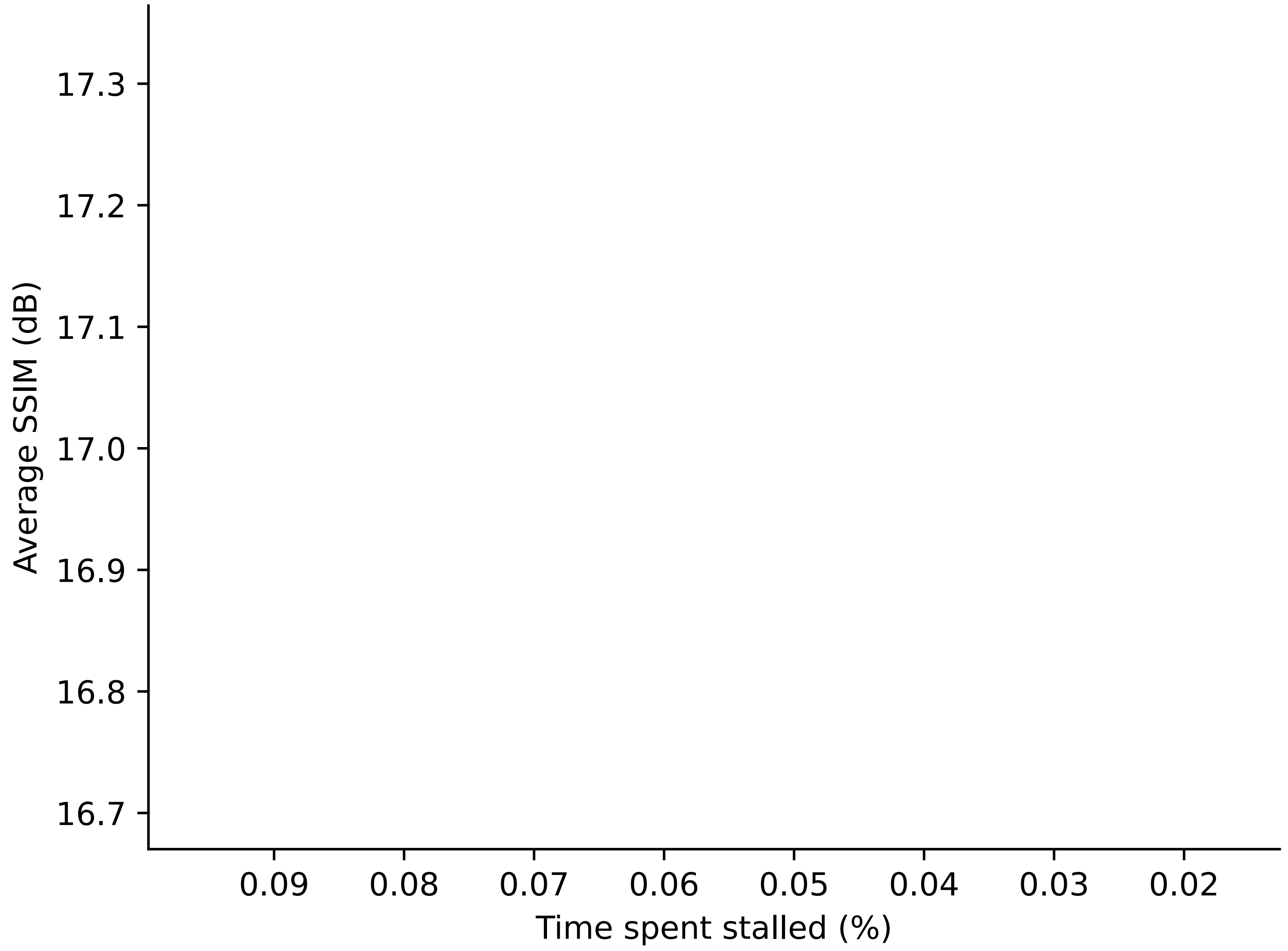
The Puffer project:

- YouTube like streaming platform
- Streaming live TV to users across the US
- Hosts various state-of-the-art ABR schemes
- Data collection for evaluation and training *in situ*



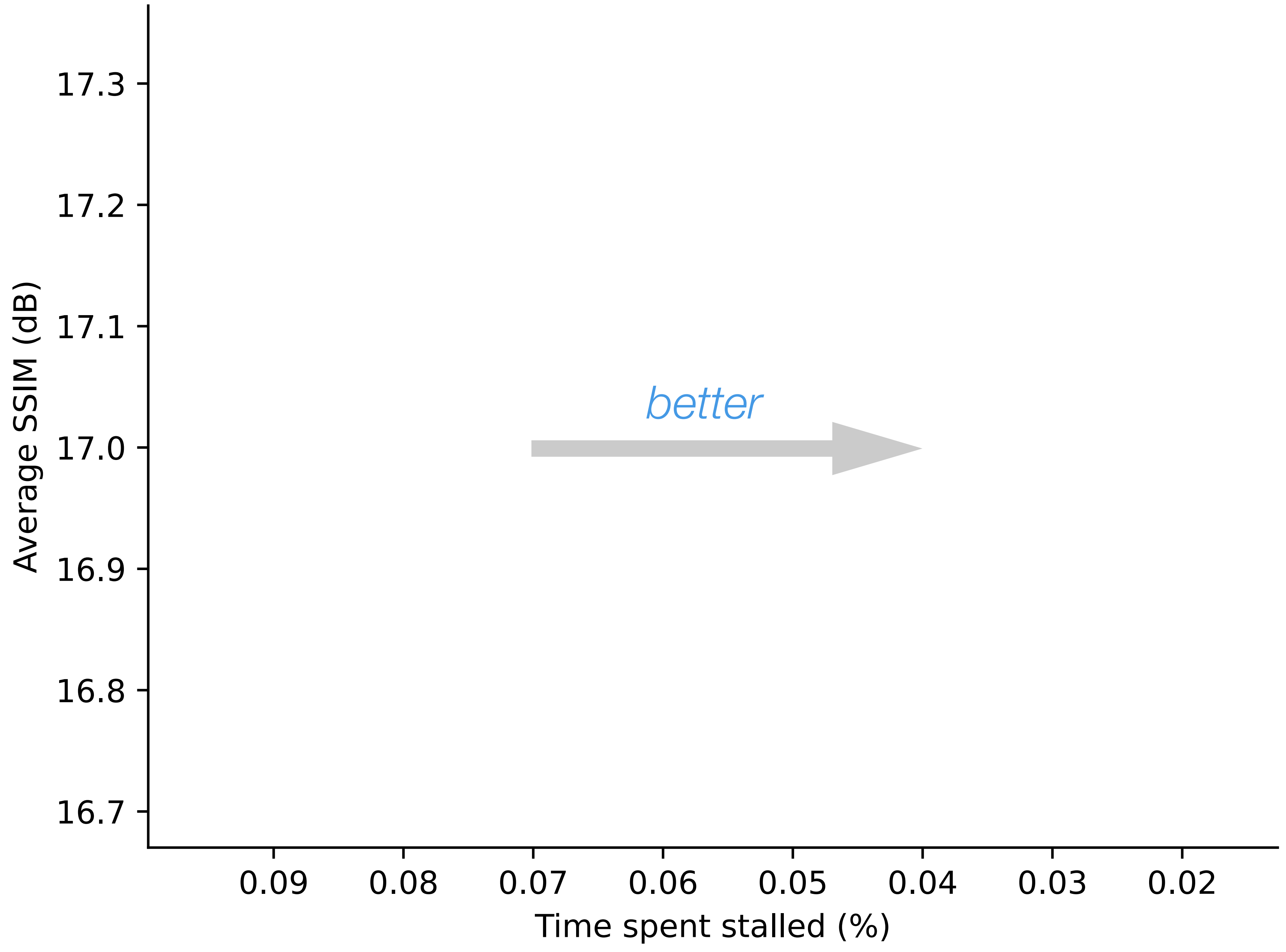
Puffer: F. Yan, et al., 2020

Testing on the Internet



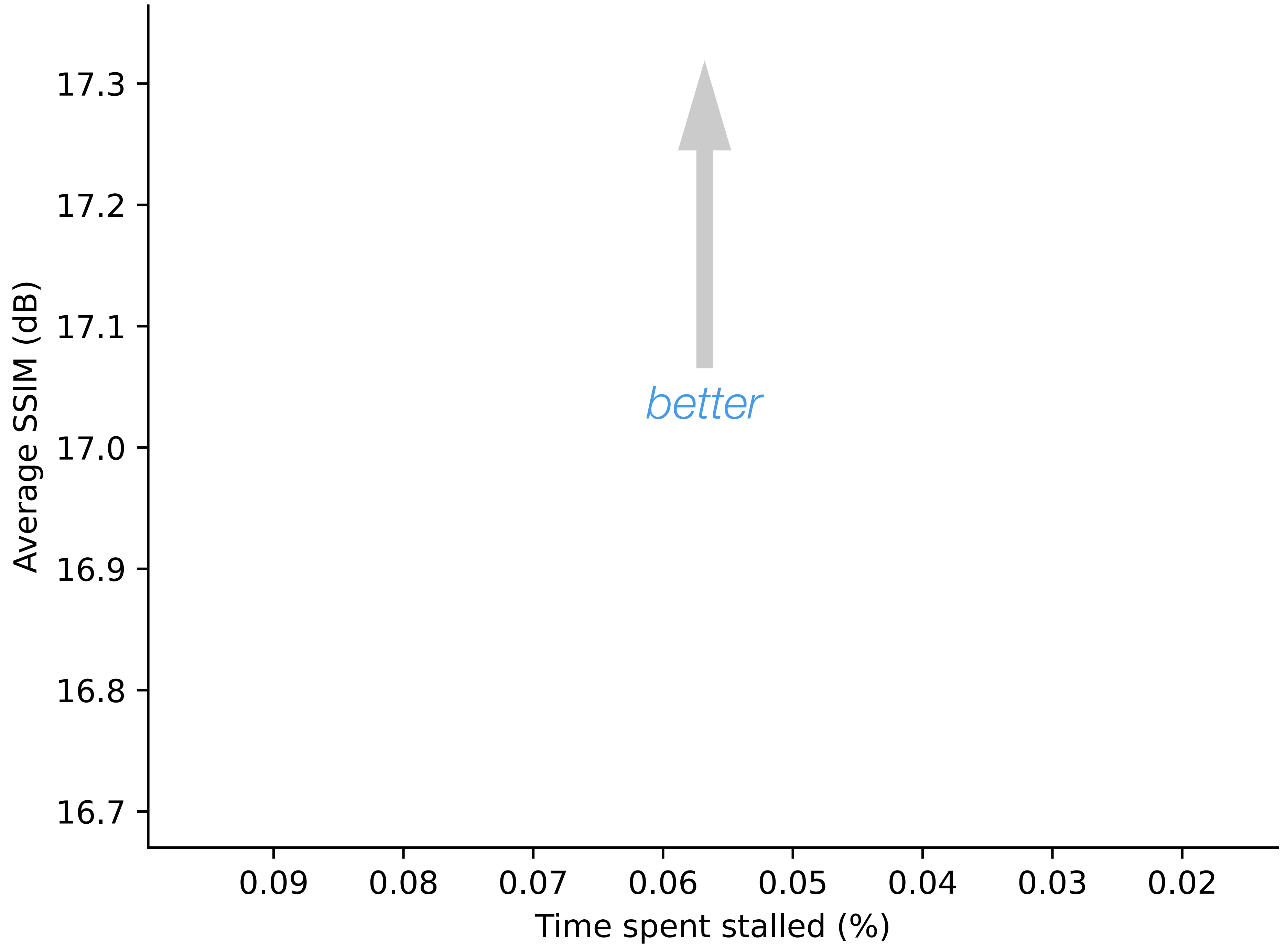
Data from Puffer: F. Yan, et al., 2020

Testing on the Internet



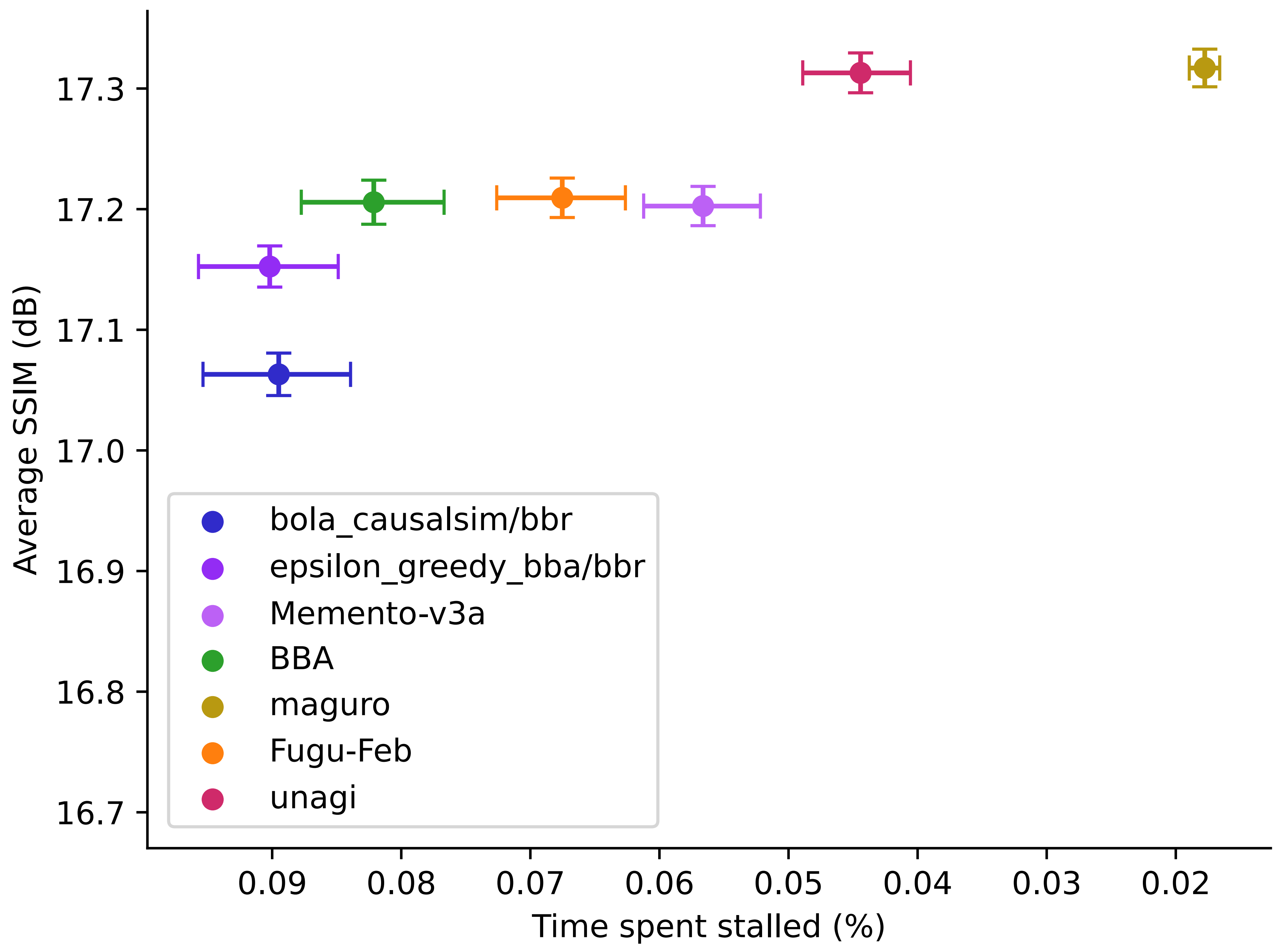
Data from Puffer: F. Yan, et al., 2020

Testing on the Internet



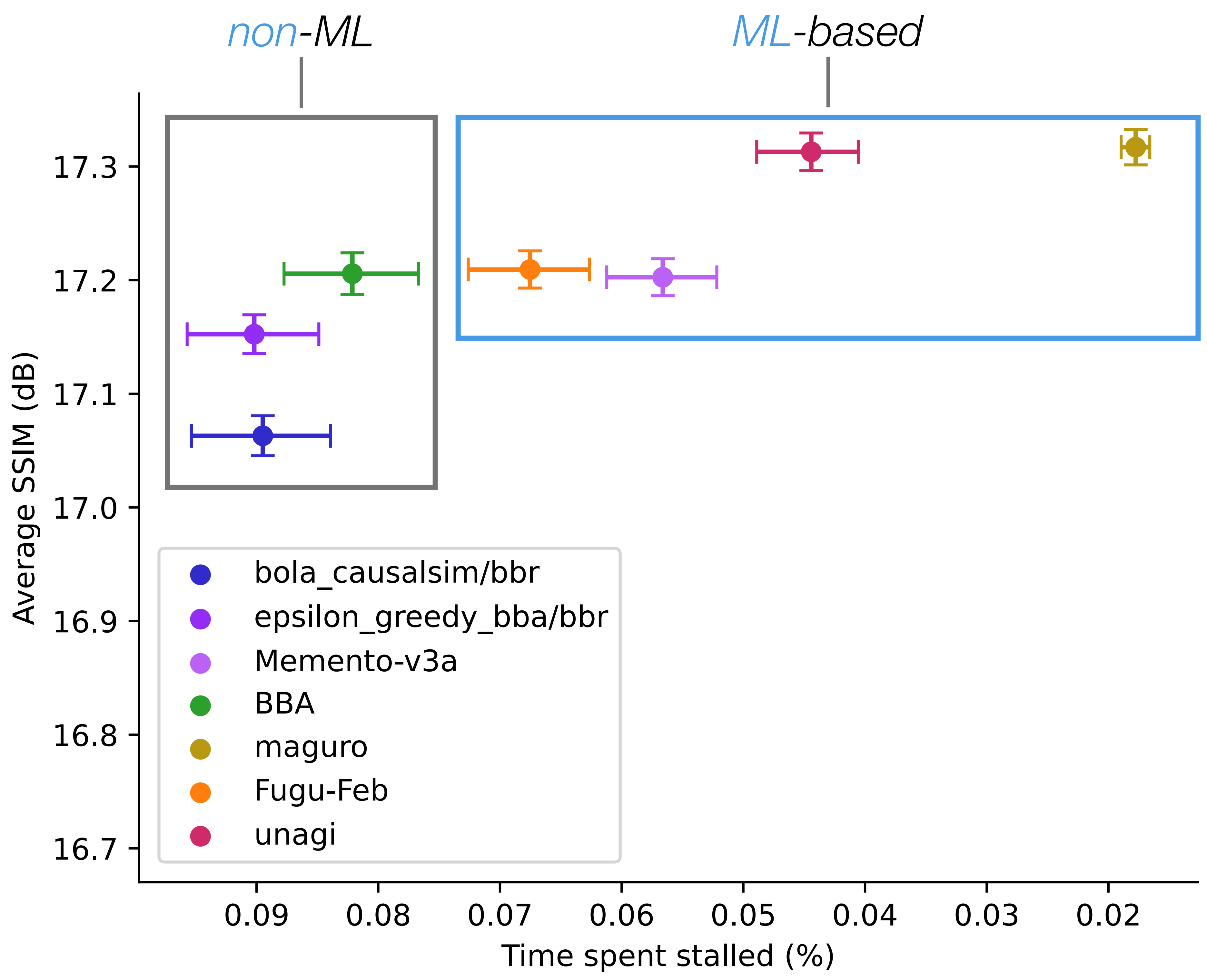
Data from Puffer: F. Yan, et al., 2020

Testing on the Internet



Data from Puffer: F. Yan, et al., 2020

Testing on the Internet



Data from Puffer: F. Yan, et al., 2020

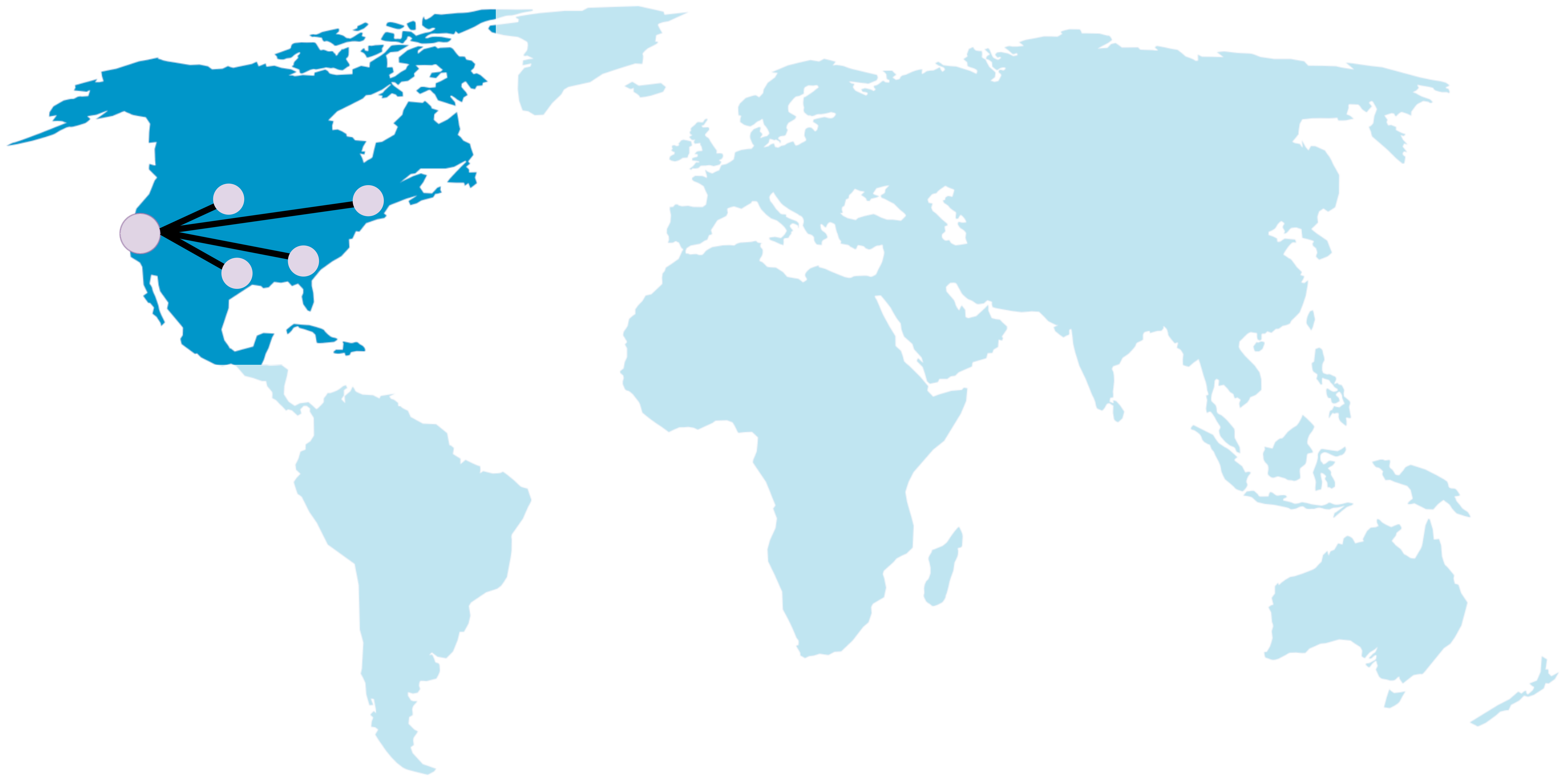
Puffer's limitation

The Puffer project … is limited:

Puffer's limitation

The Puffer project … is limited:

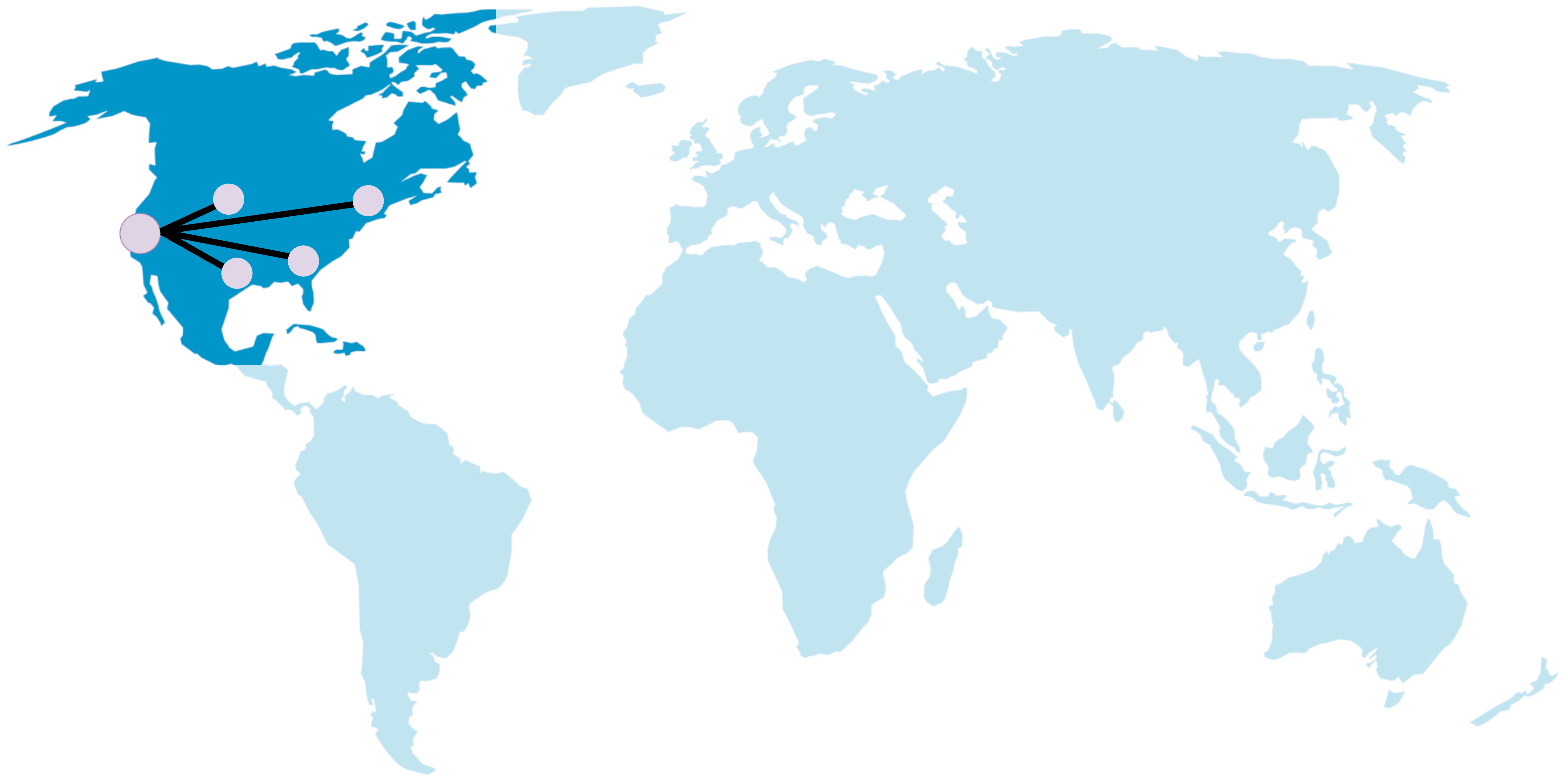
- Server in Stanford and US clients



Puffer's limitation

The Puffer project … is limited:

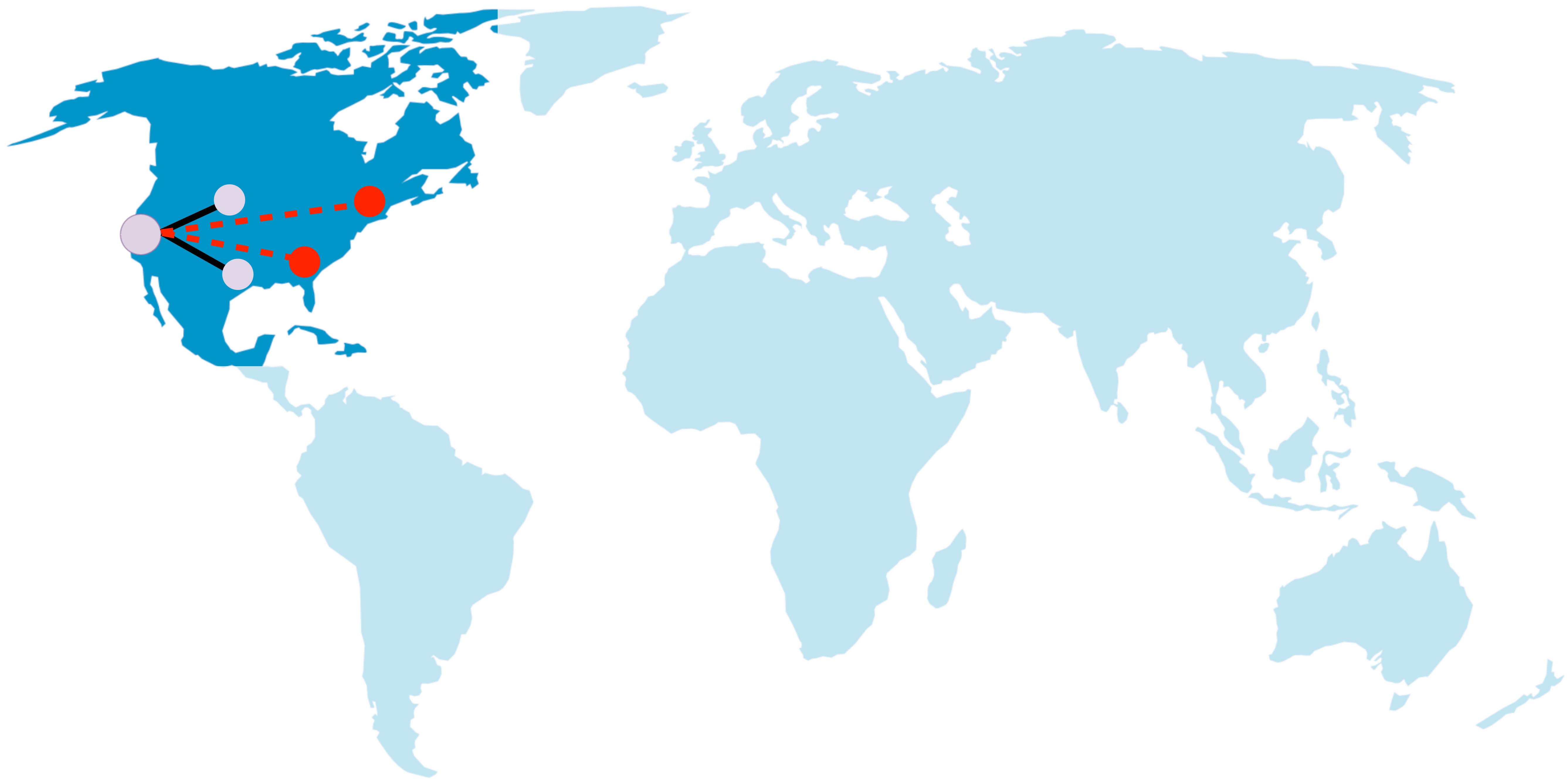
- Server in Stanford and US clients
- Results indicate survivorship bias



Puffer's limitation

The Puffer project … is limited:

- Server in Stanford and US clients
- Results indicate survivorship bias

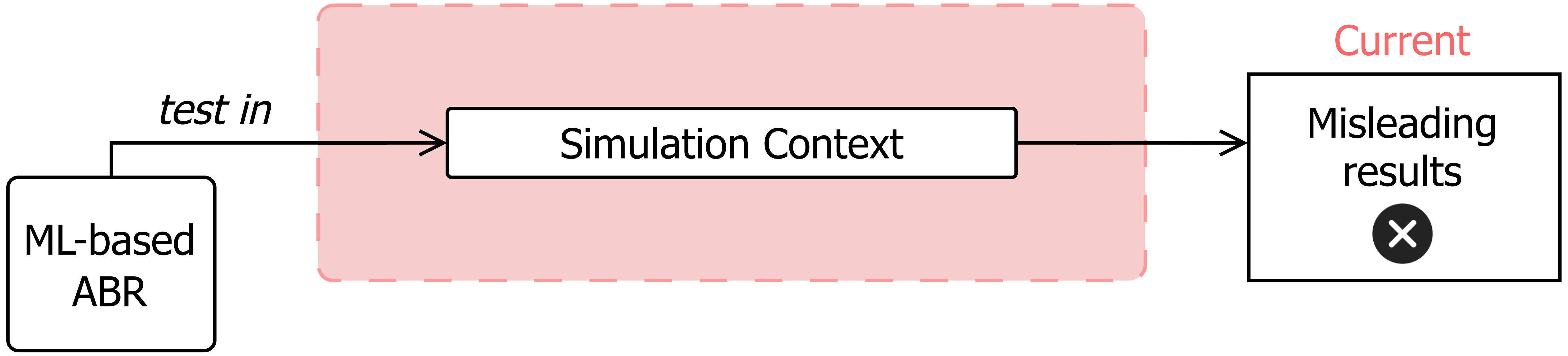


Puffer's limitation

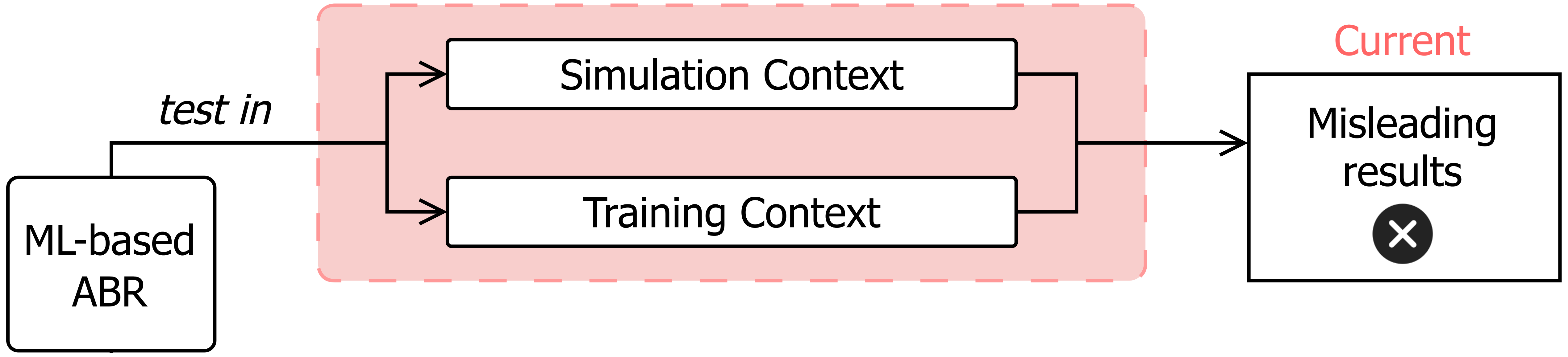
The Puffer project … is limited:

- Server in Stanford and US clients
- Results indicate survivorship bias
- Hard to build, deploy and scale

Puffer's limitation

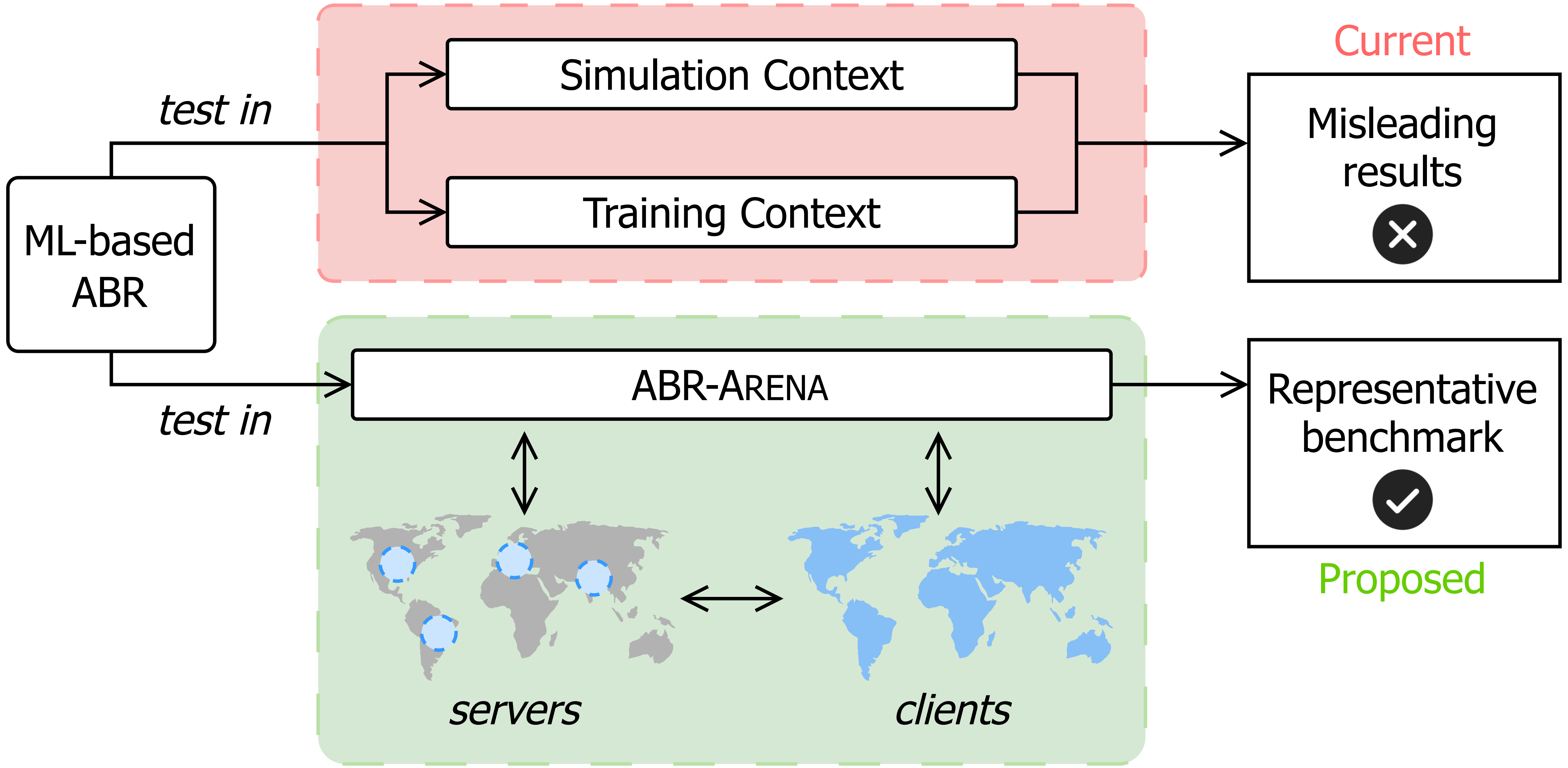


Puffer's limitation



→ *Do Puffer's results suffer from similar limitations as results from synthetic environments?*

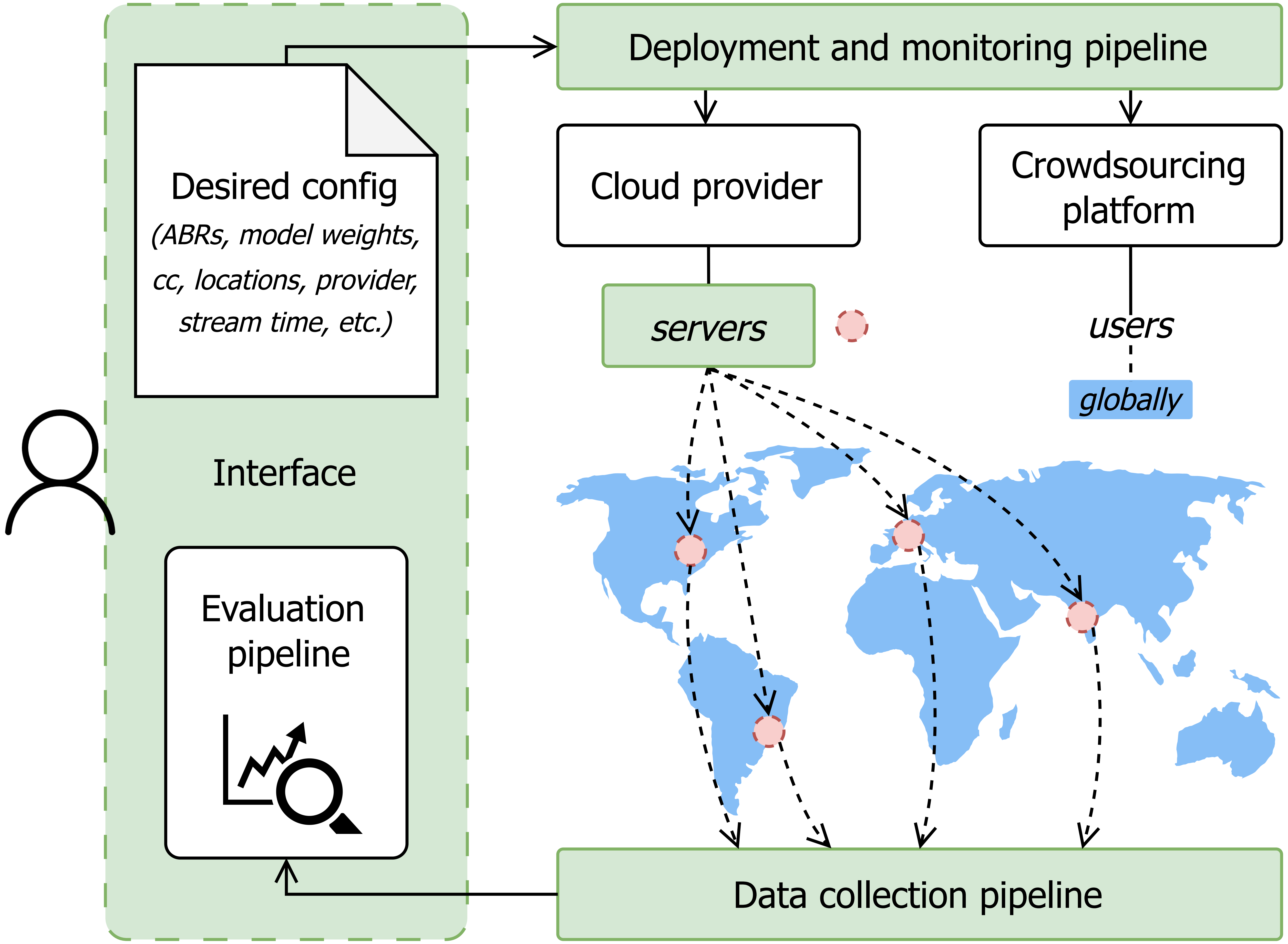
Introducing ABR-Arena



System design

Addressing:

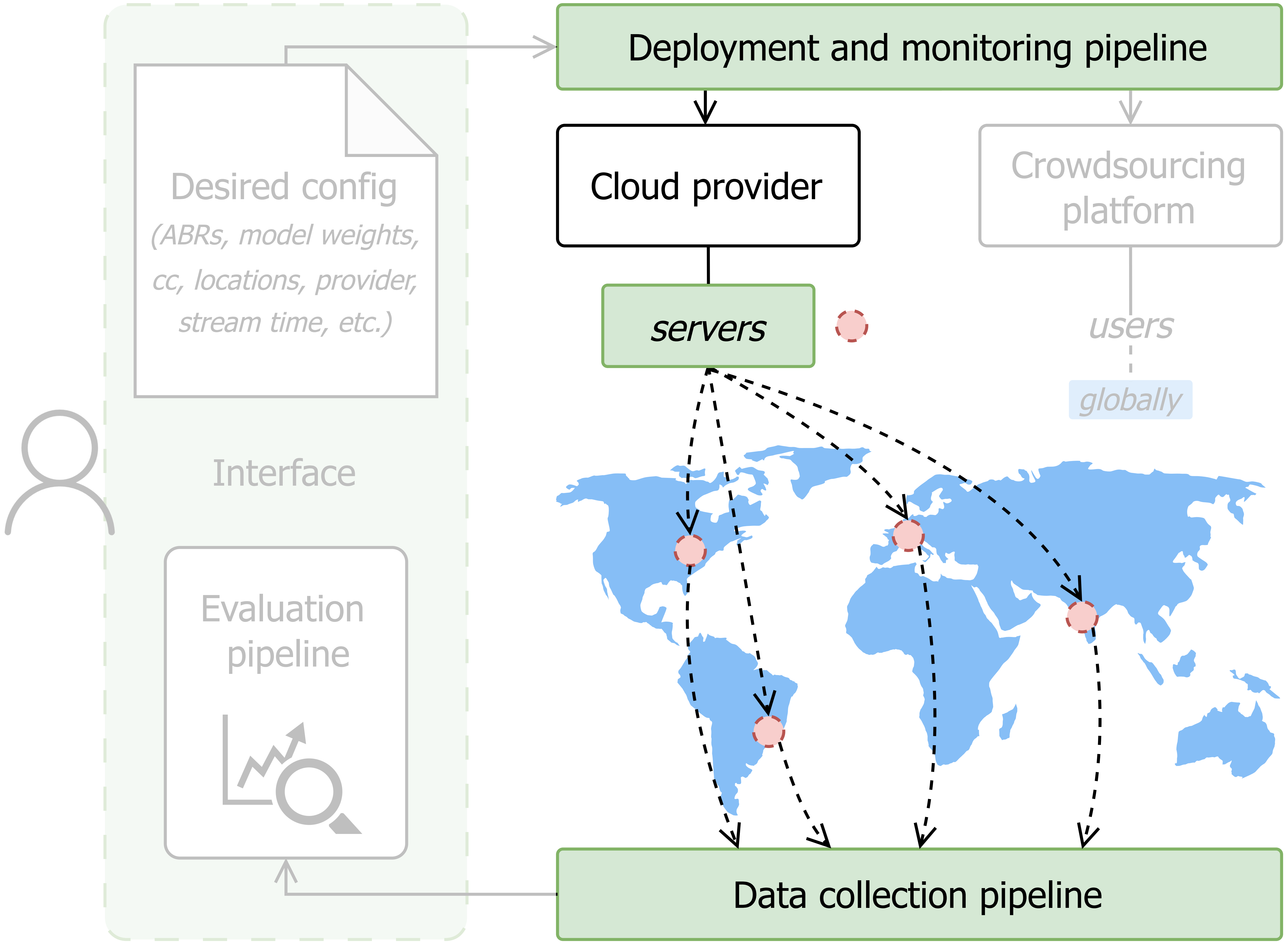
- Regional diversity
- Survivorship bias
- Scalability



System **design**

Addressing:

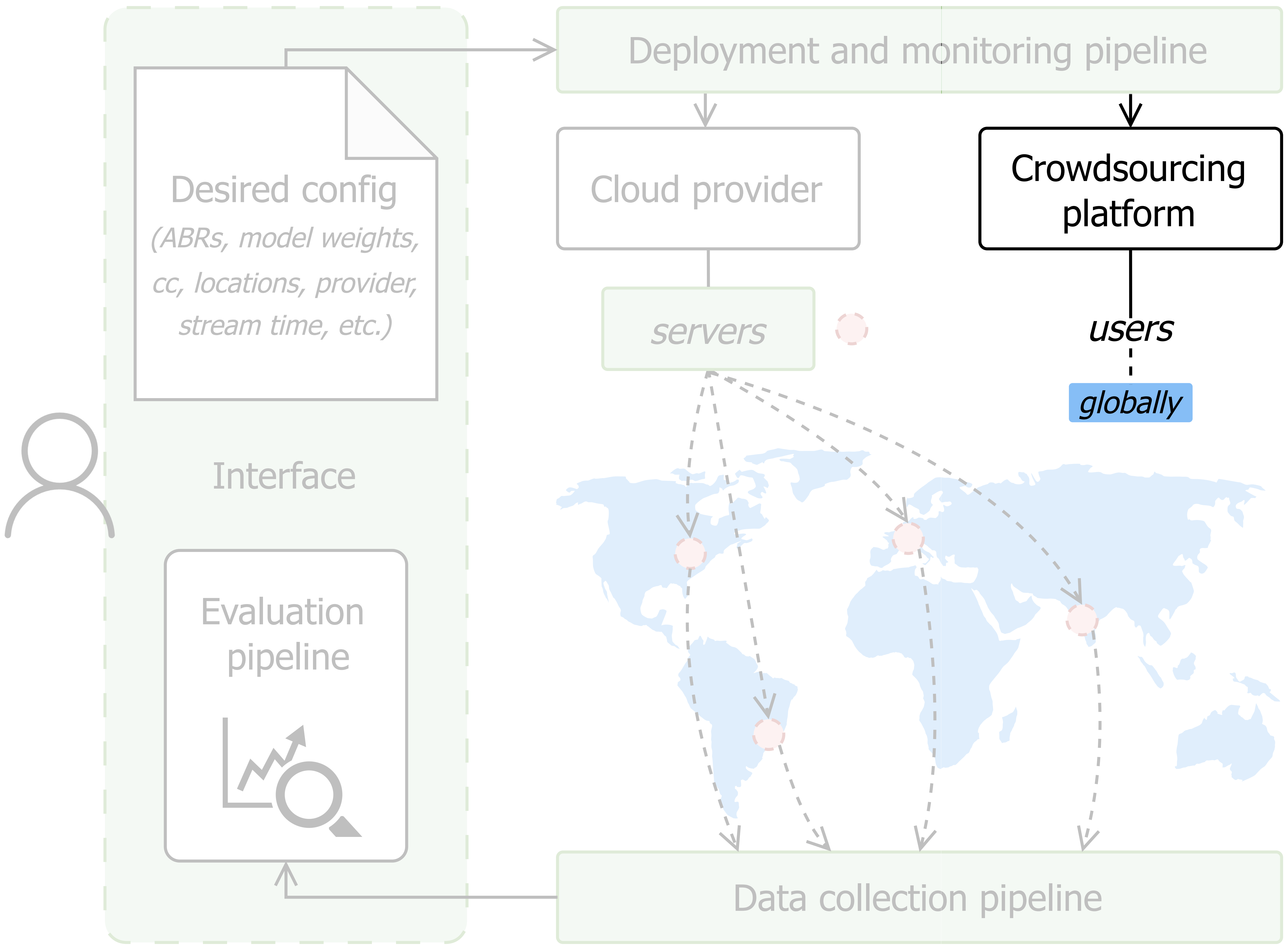
- *Regional diversity*
- Survivorship bias
- Scalability



System *design*

Addressing:

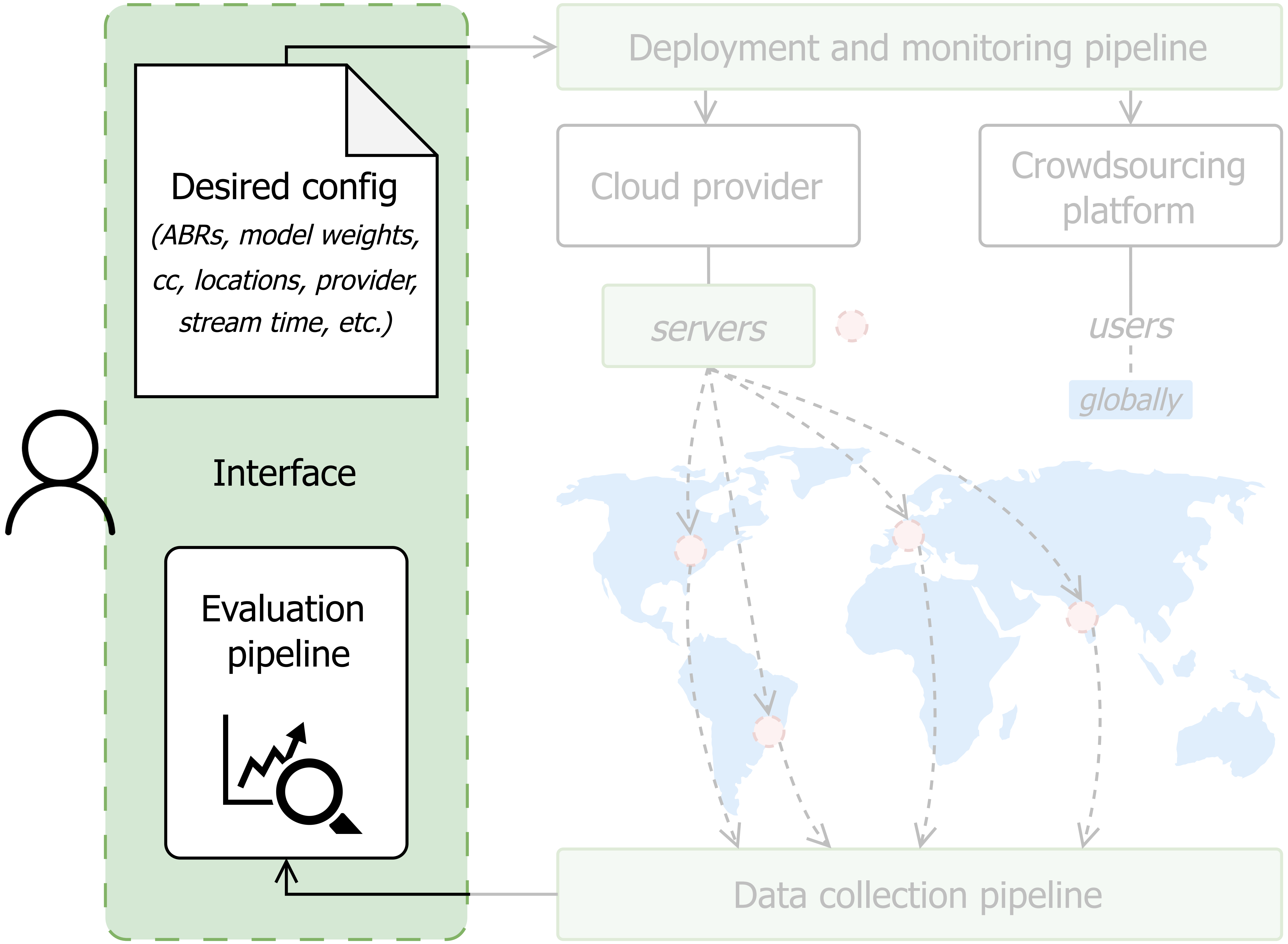
- Regional diversity
- *Survivorship bias*
- Scalability



System design

Addressing:

- Regional diversity
- Survivorship bias
- *Scalability*



Experiments and results

Experiments and results

Is there a gap between results on Puffer and deployments in practice?

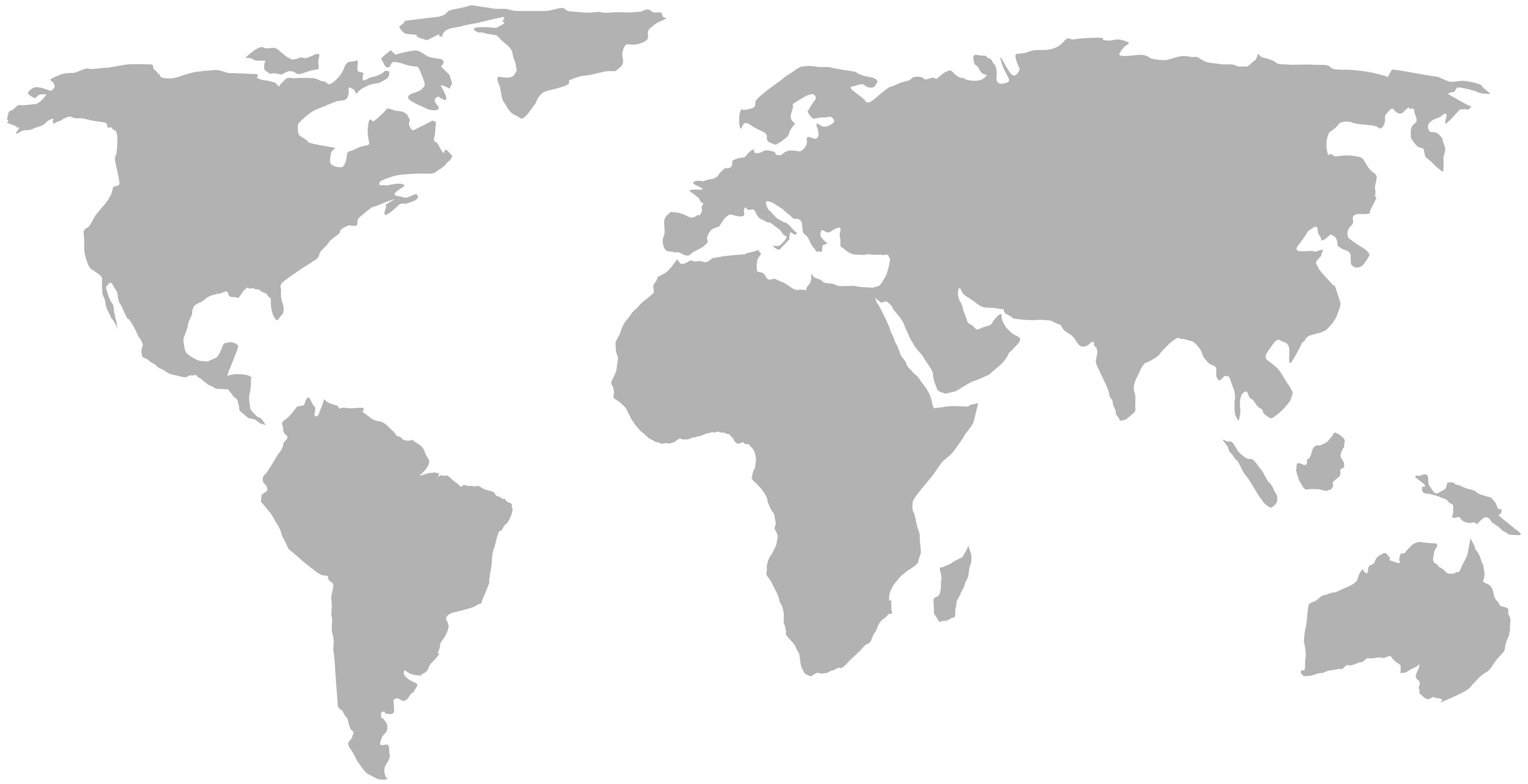
Experiments and results

Experiment setup:

Experiments and results

Experiment setup:

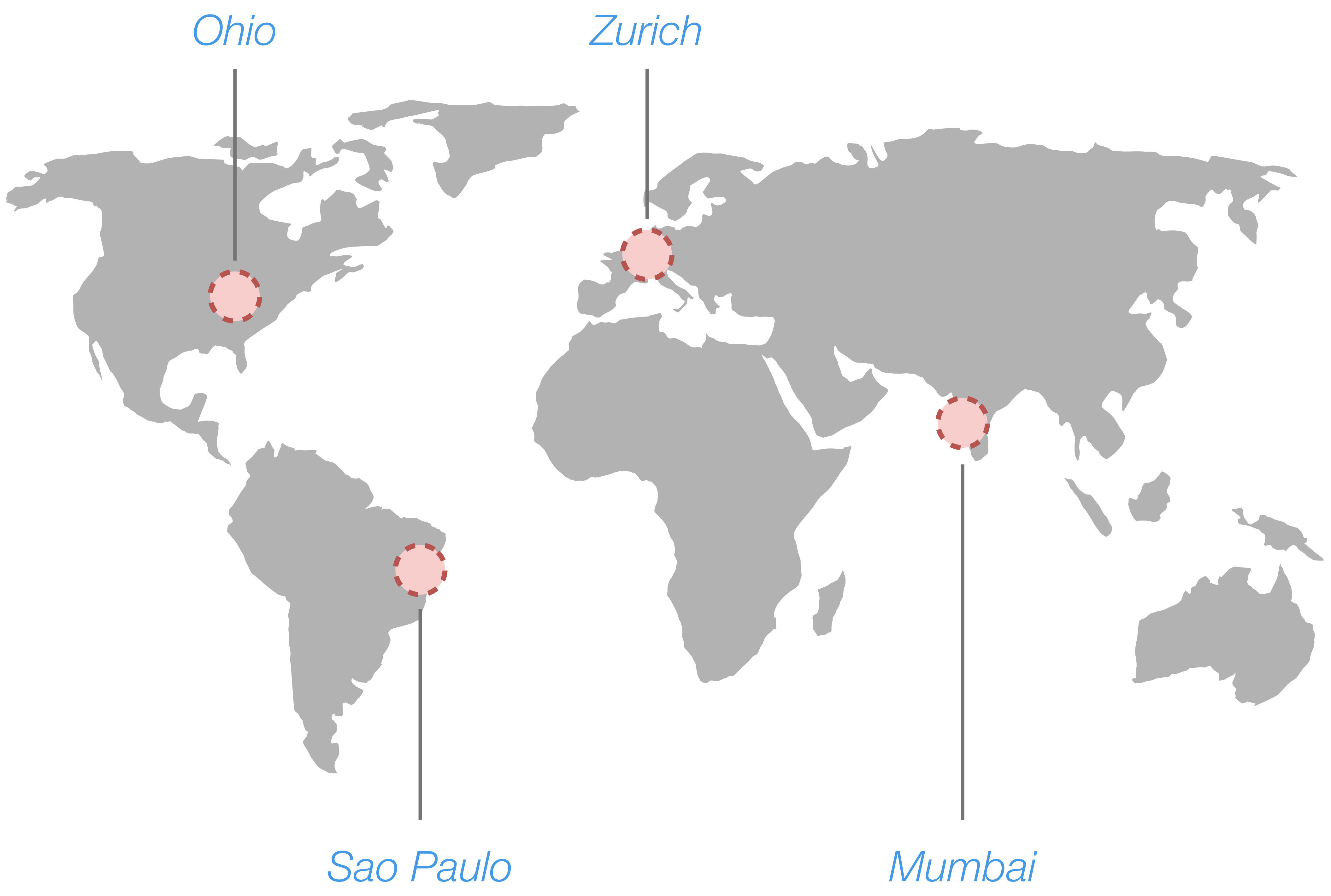
- Source users globally



Experiments and results

Experiment setup:

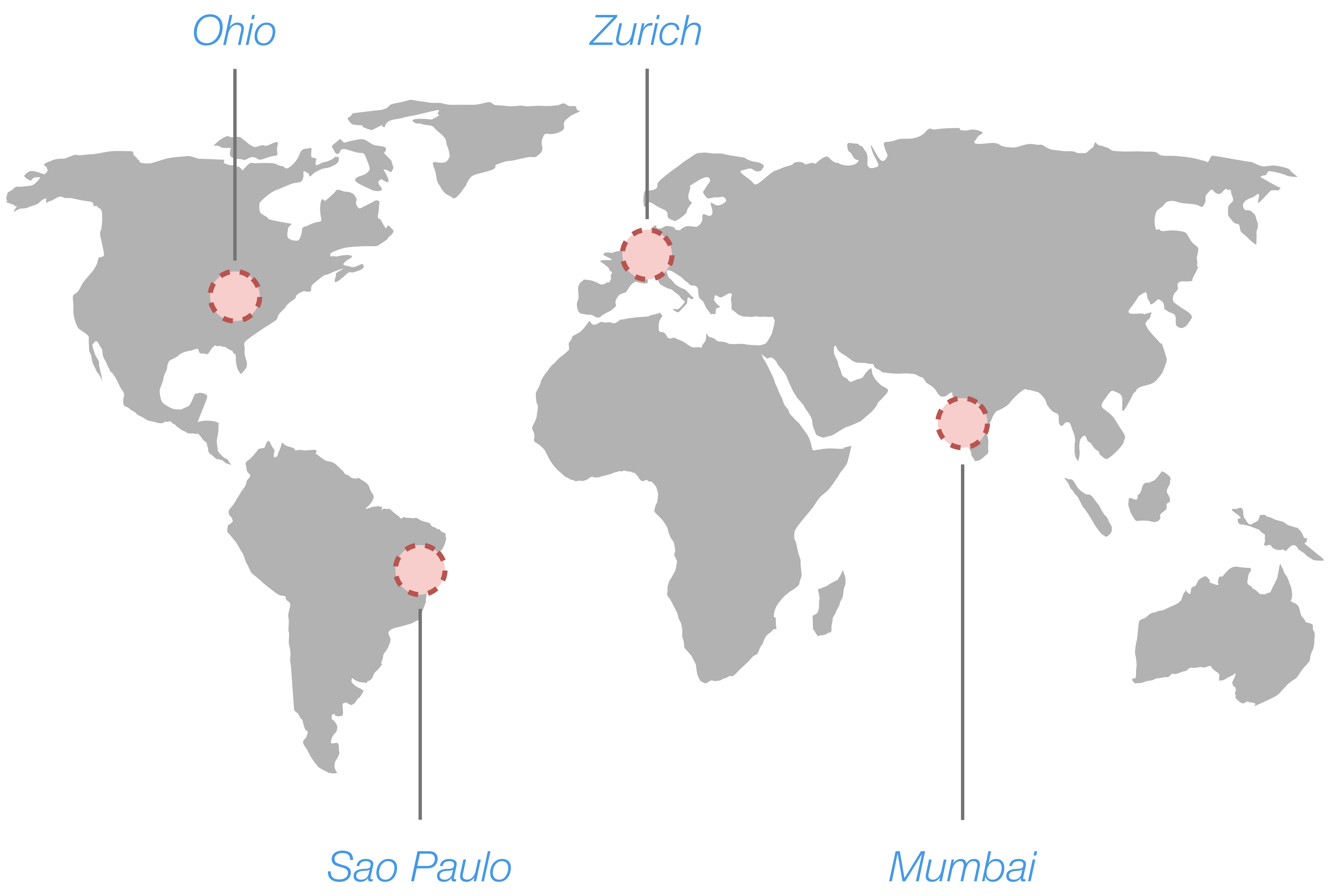
- Source users globally
- Four server locations



Experiments and results

Experiment setup:

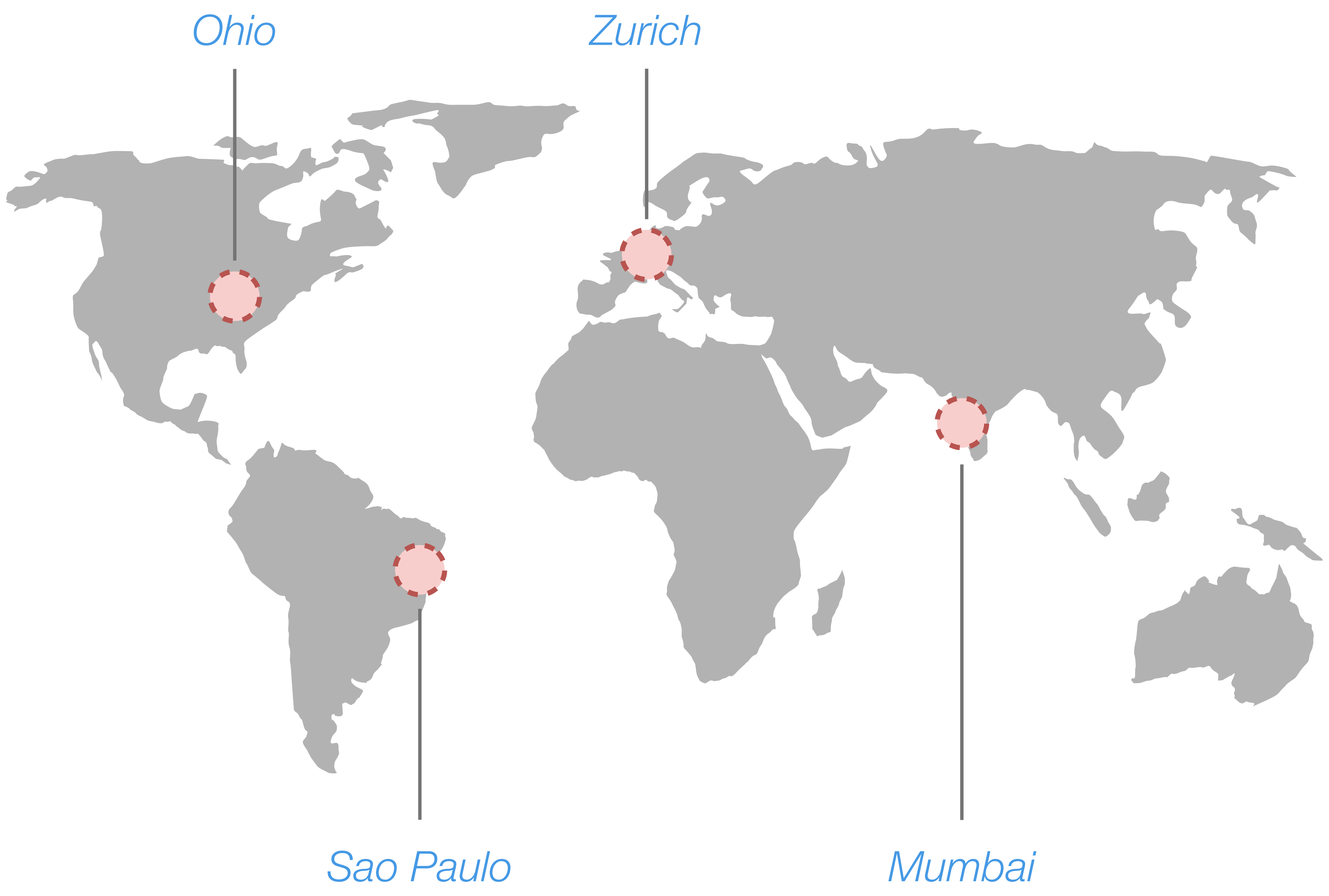
- Source users globally
- Four server locations
- Stream video for 160 seconds



Experiments and results

Experiment setup:

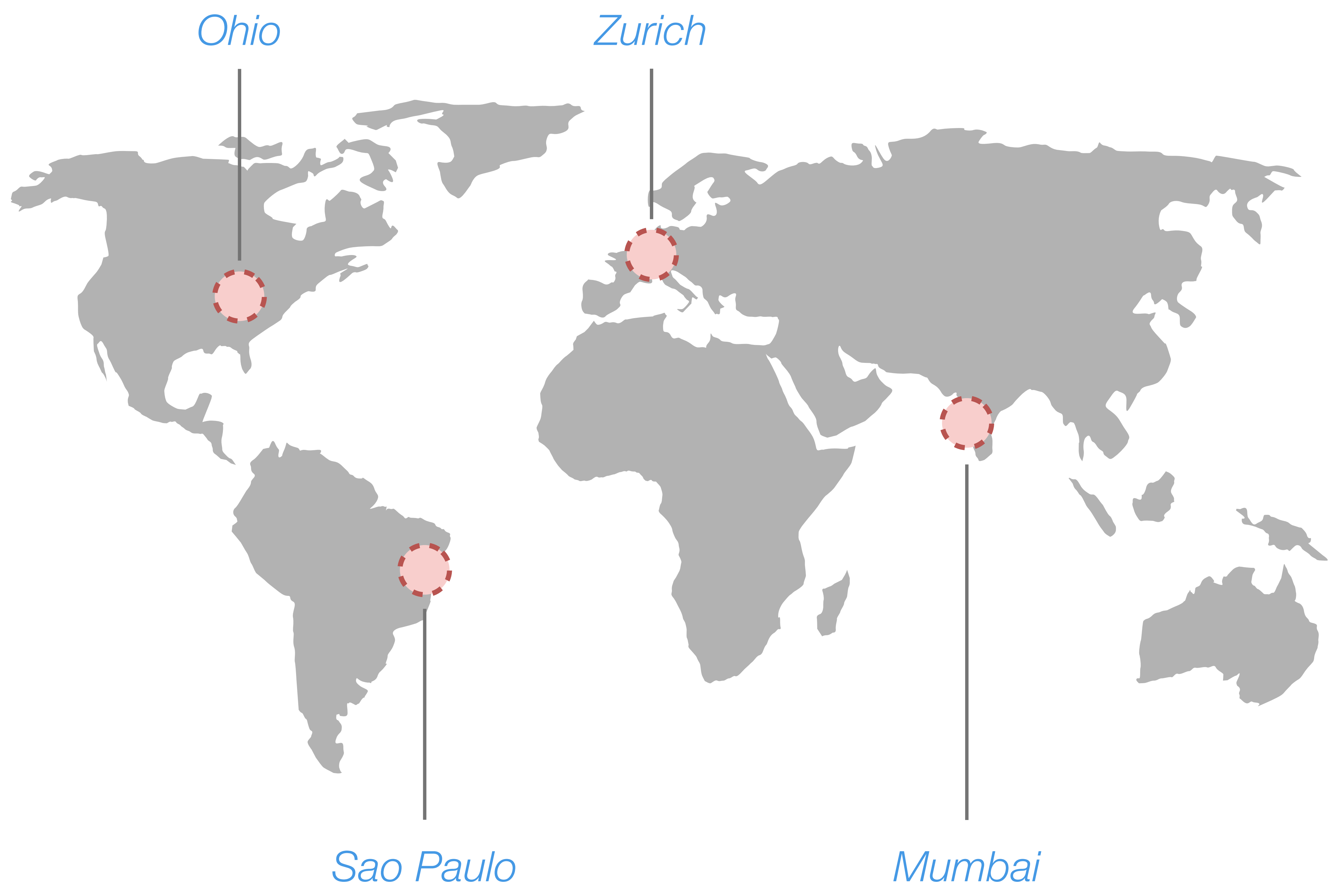
- Source users globally
- Four server locations
- Stream video for 160 seconds
- 11’156 users and 510 hours total



Experiments and results

Experiment setup:

- Source users globally
- Four server locations
- Stream video for 160 seconds
- 11’156 users and 510 hours total
- Four ABR algorithms



Experiments and results

Experiment setup:

- Source users globally
- Four server locations
- Stream video for 160 seconds
- 11'156 users and 510 hours total
- Four ABR algorithms

ML: Unagi[4], Maguro[4] and Fugu[3]

[3] F. Yan, et al., 2020 | [4] S. Patel, et al., 2020

Experiments and results

Experiment setup:

- Source users globally
- Four server locations
- Stream video for 160 seconds
- 11'156 users and 510 hours total
- Four ABR algorithms

ML: Unagi[4], Maguro[4] and Fugu[3] – all trained on Puffer

Experiments and results

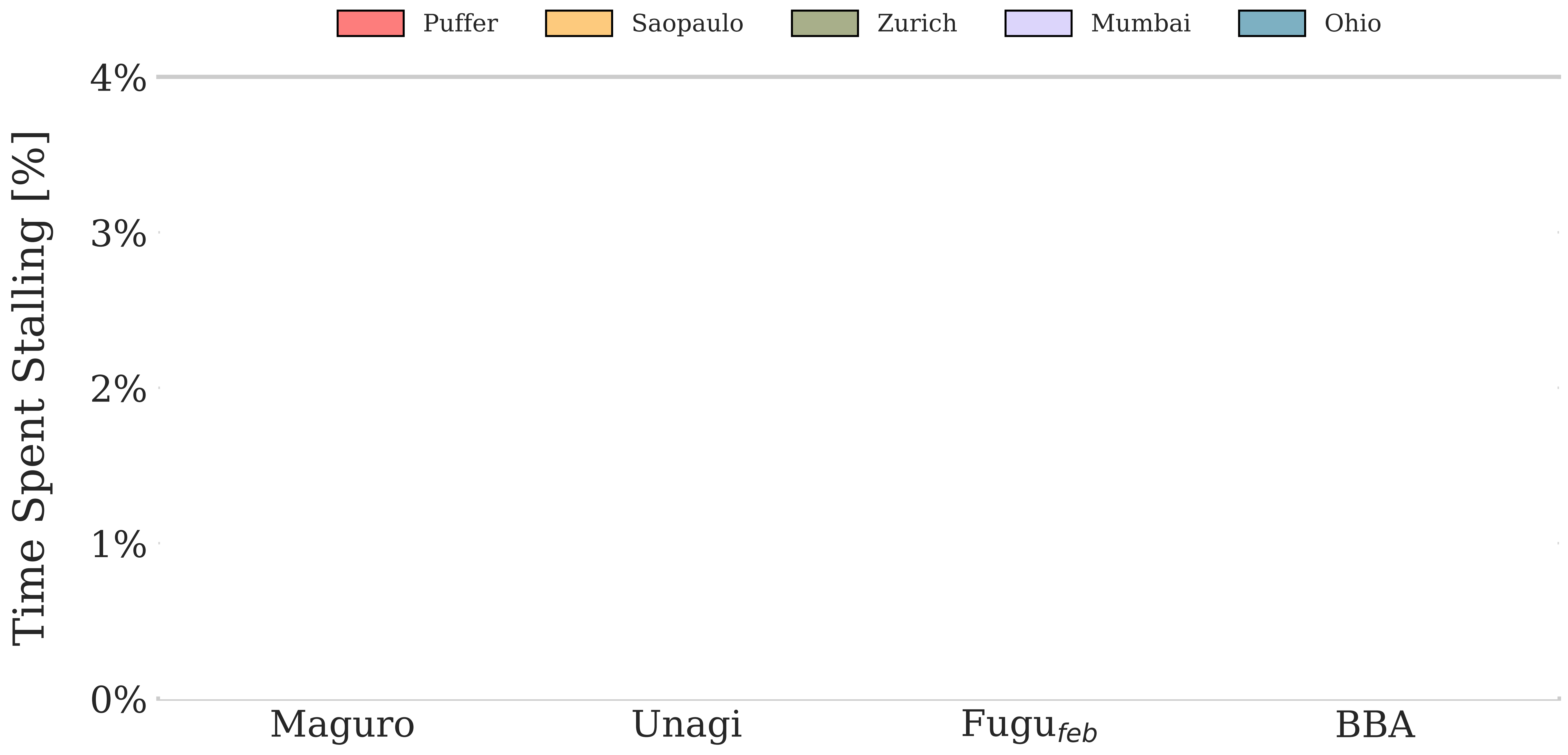
Experiment setup:

- Source users globally
- Four server locations
- Stream video for 160 seconds
- 11’156 users and 510 hours total
- Four ABR algorithms
  - ML: Unagi[4], Maguro[4] and Fugu[3] – all trained on Puffer
  - Non-ML: BBA[5]

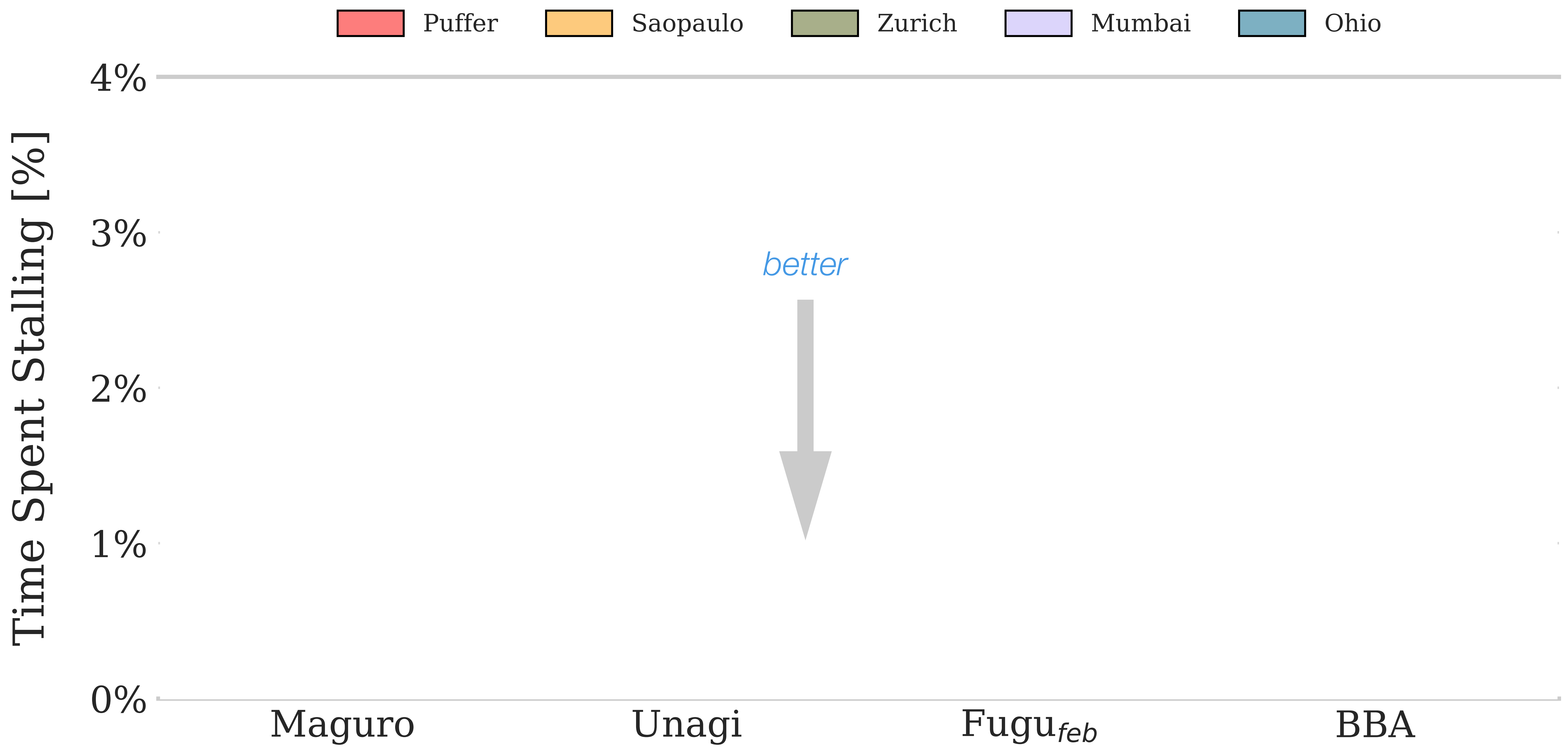
[3] F. Yan, et al., 2020 | [4] S. Patel, et al., 2020 | [5] T. Huan, et al., 2014

Experiments and results

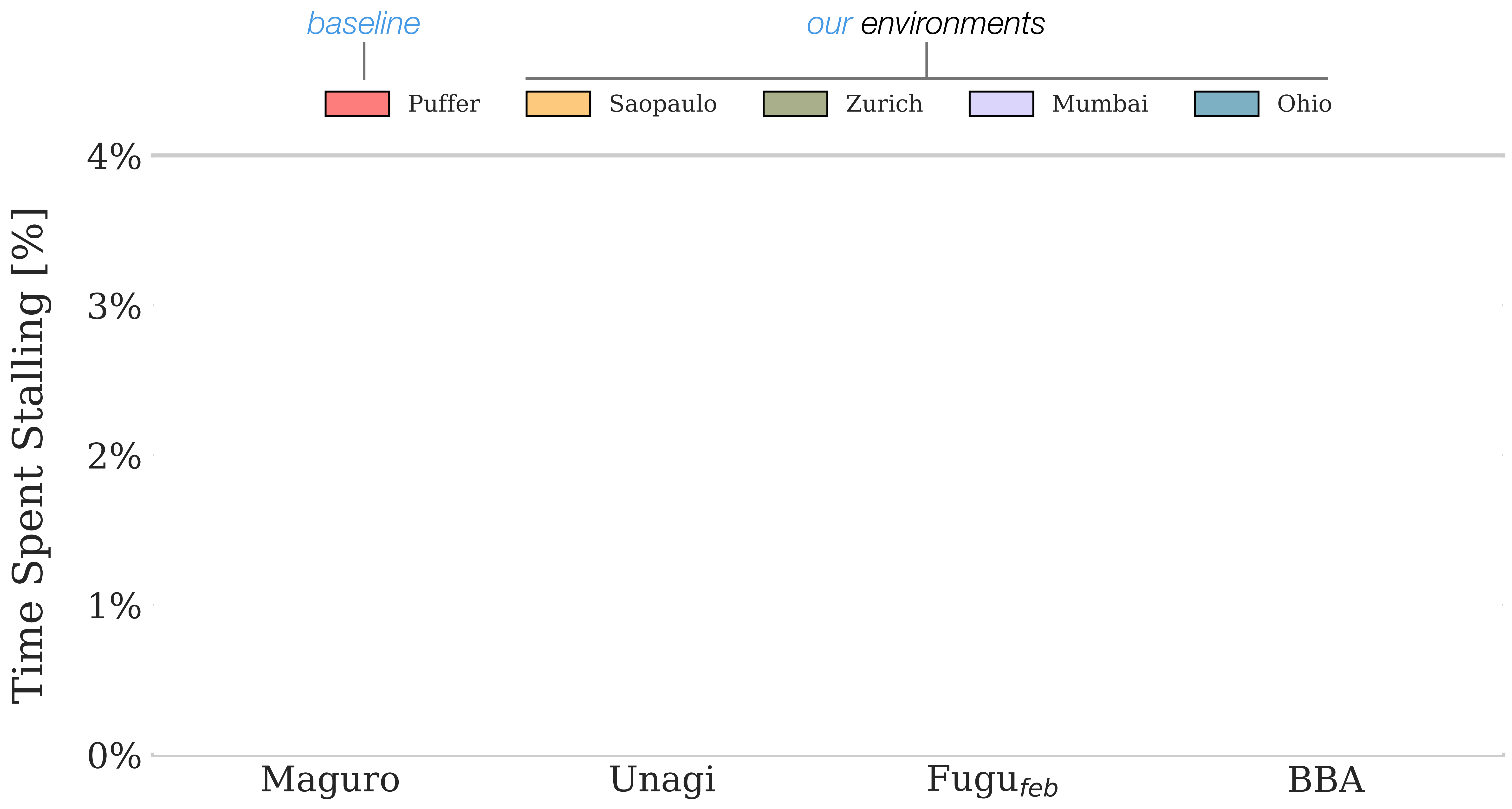
Testing outside of training context



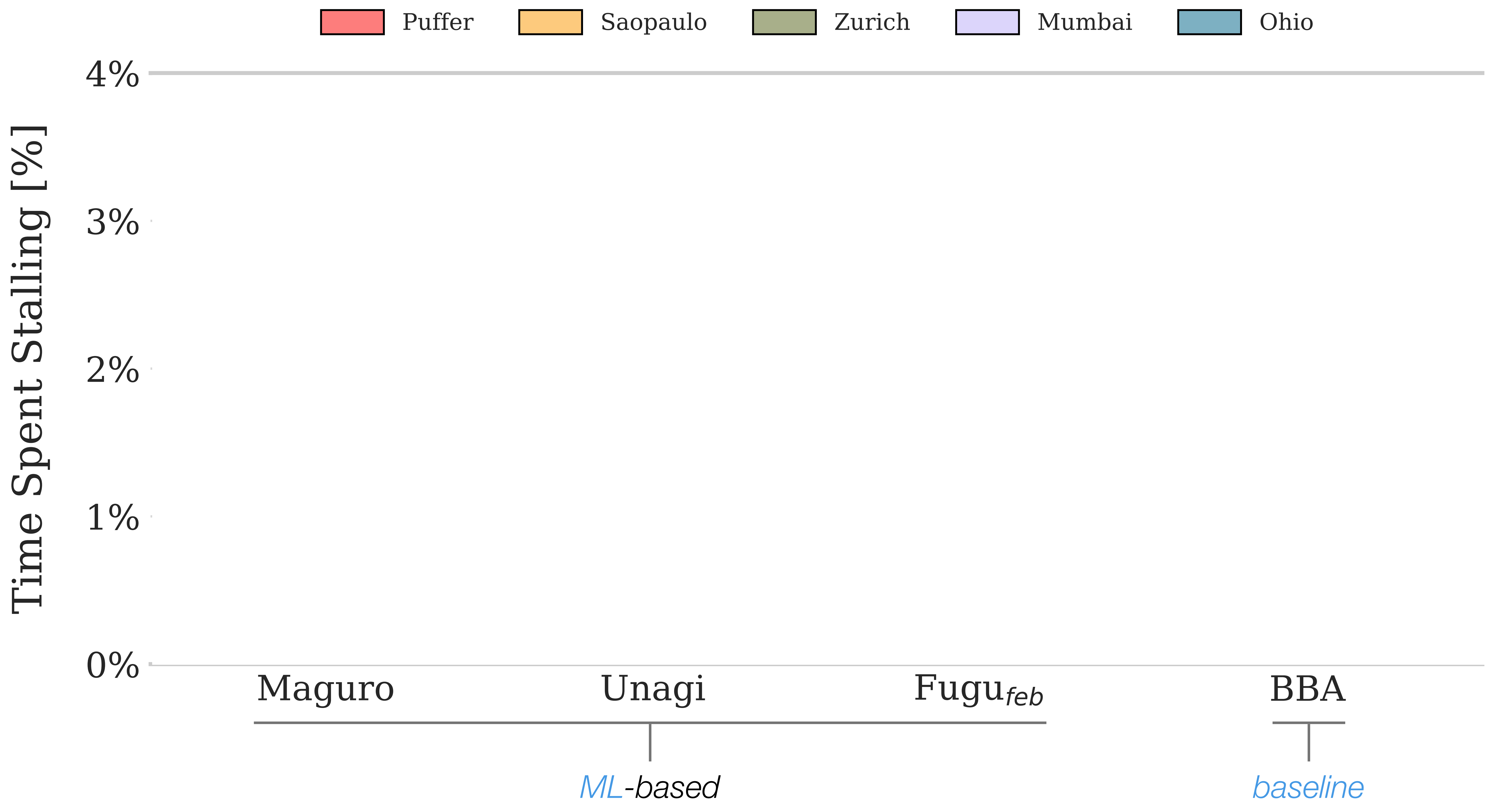
Testing outside of training context



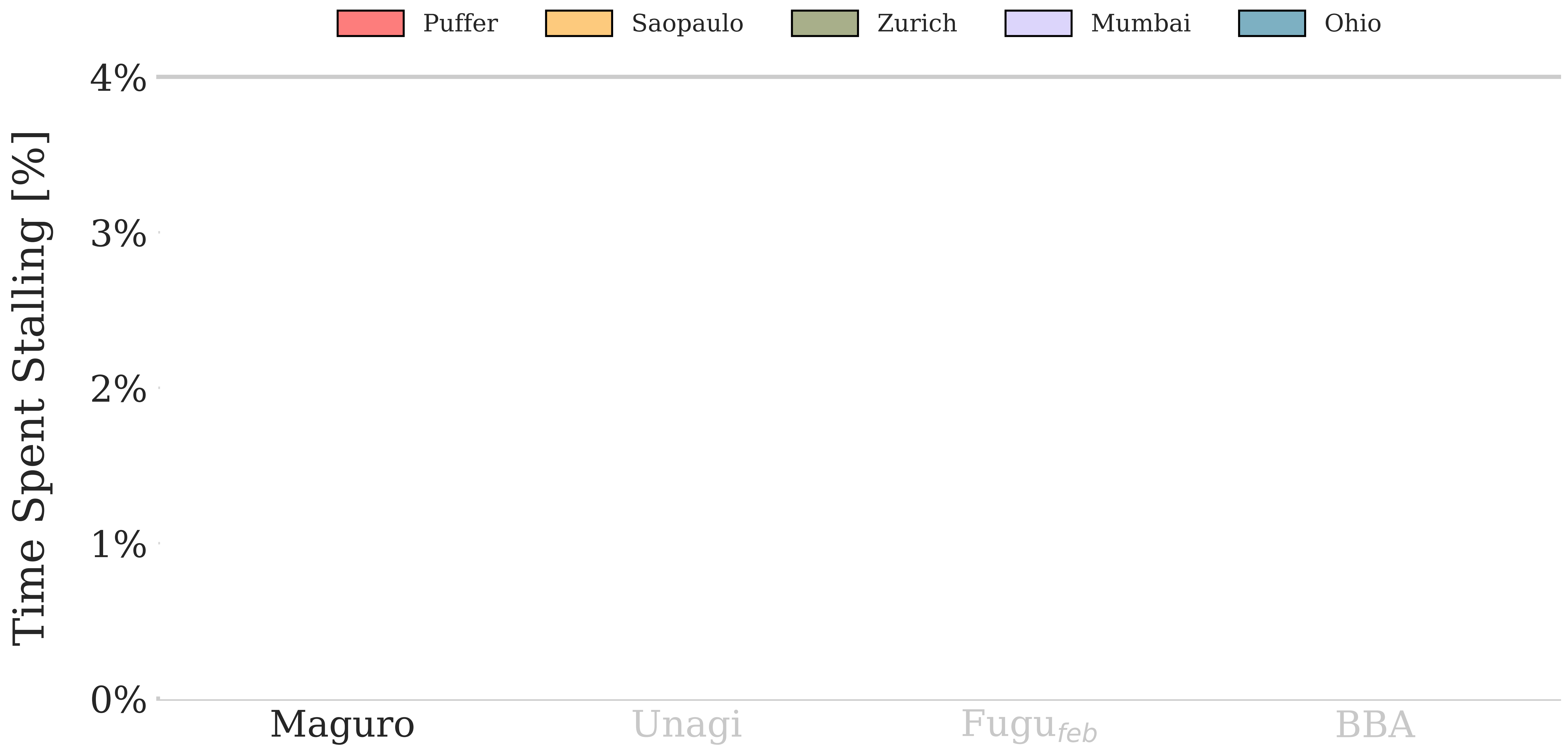
Testing outside of training context



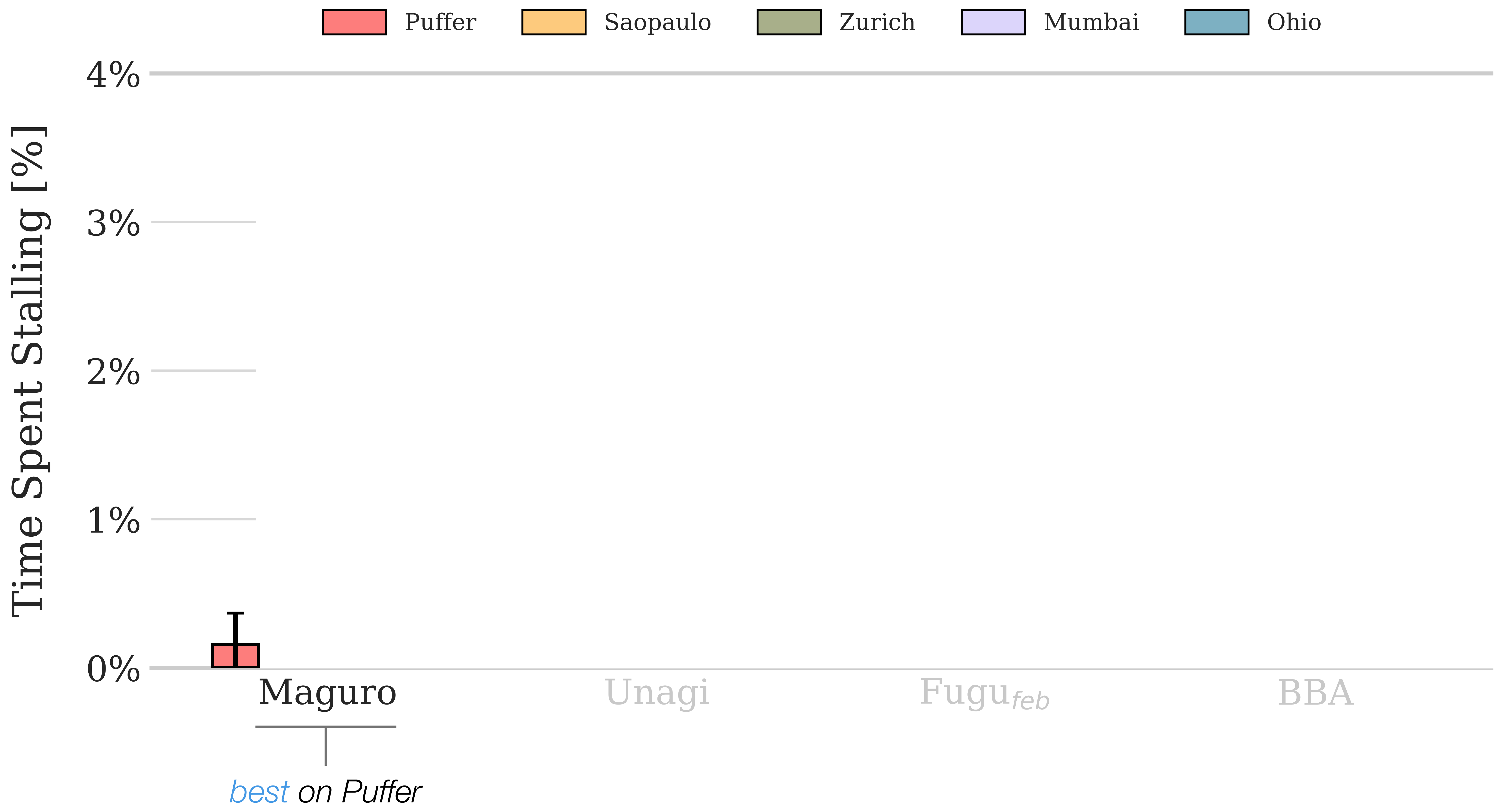
Testing outside of training context



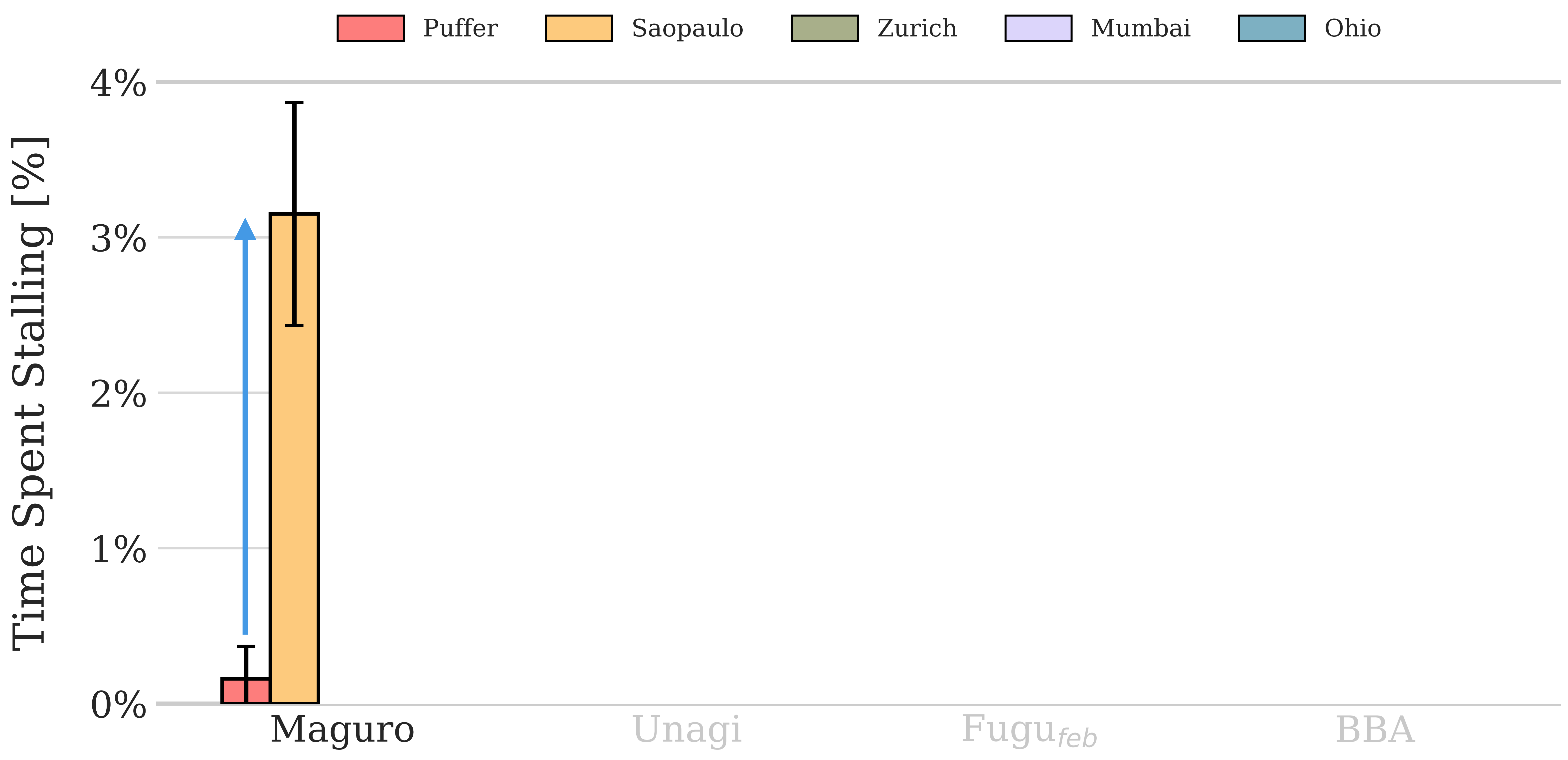
Testing outside of training context



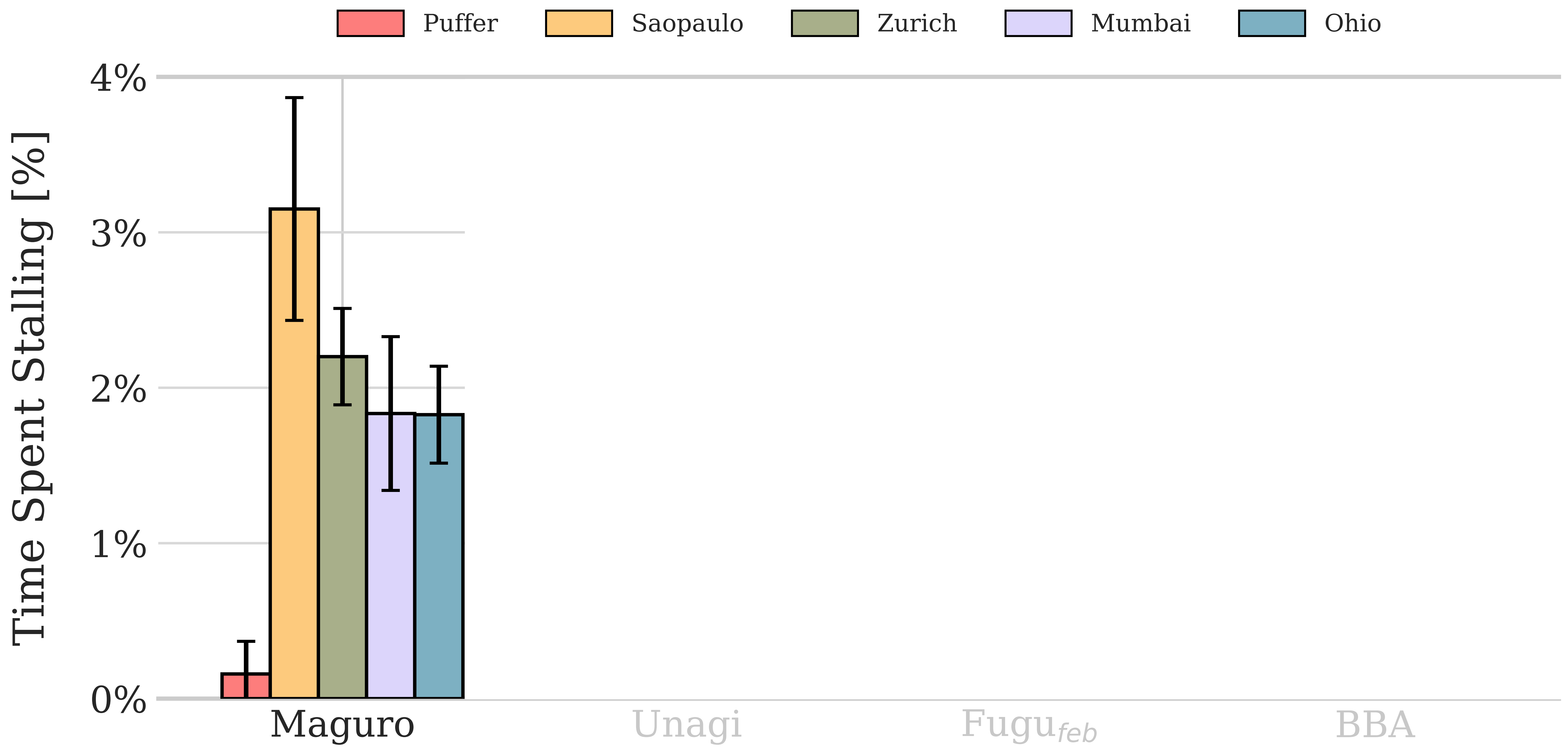
Testing outside of training context



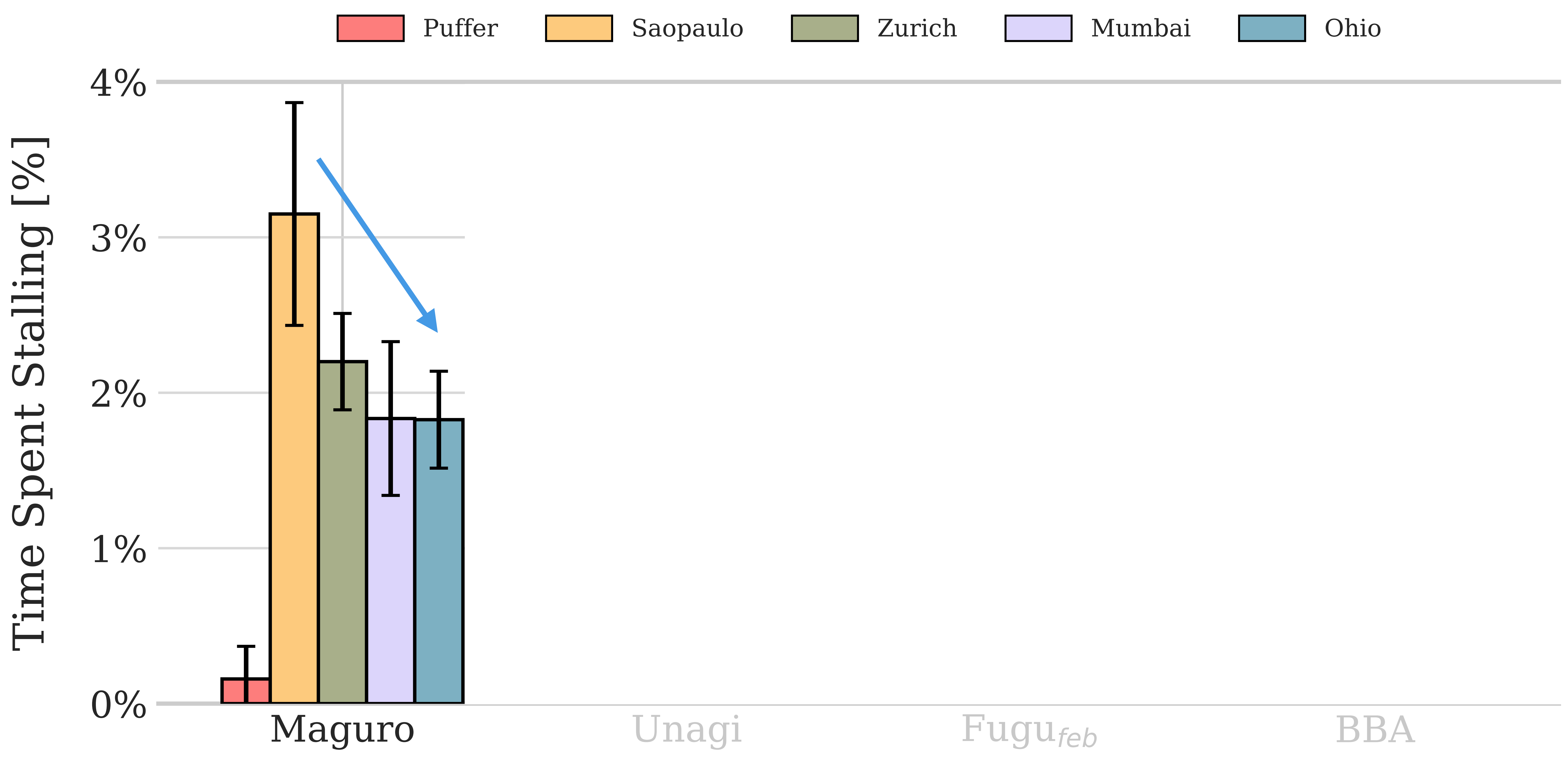
Testing outside of training context



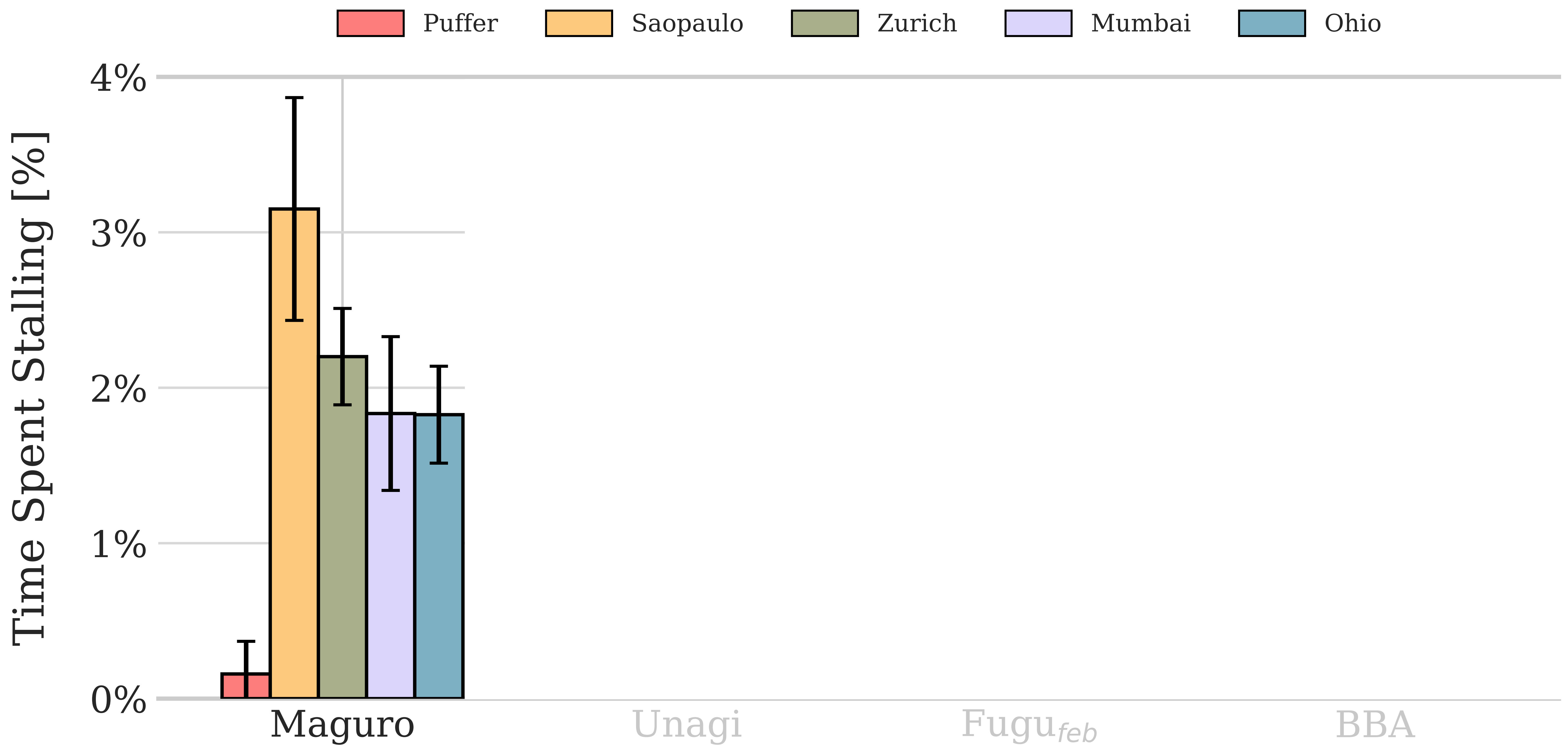
Testing outside of training context



Testing outside of training context

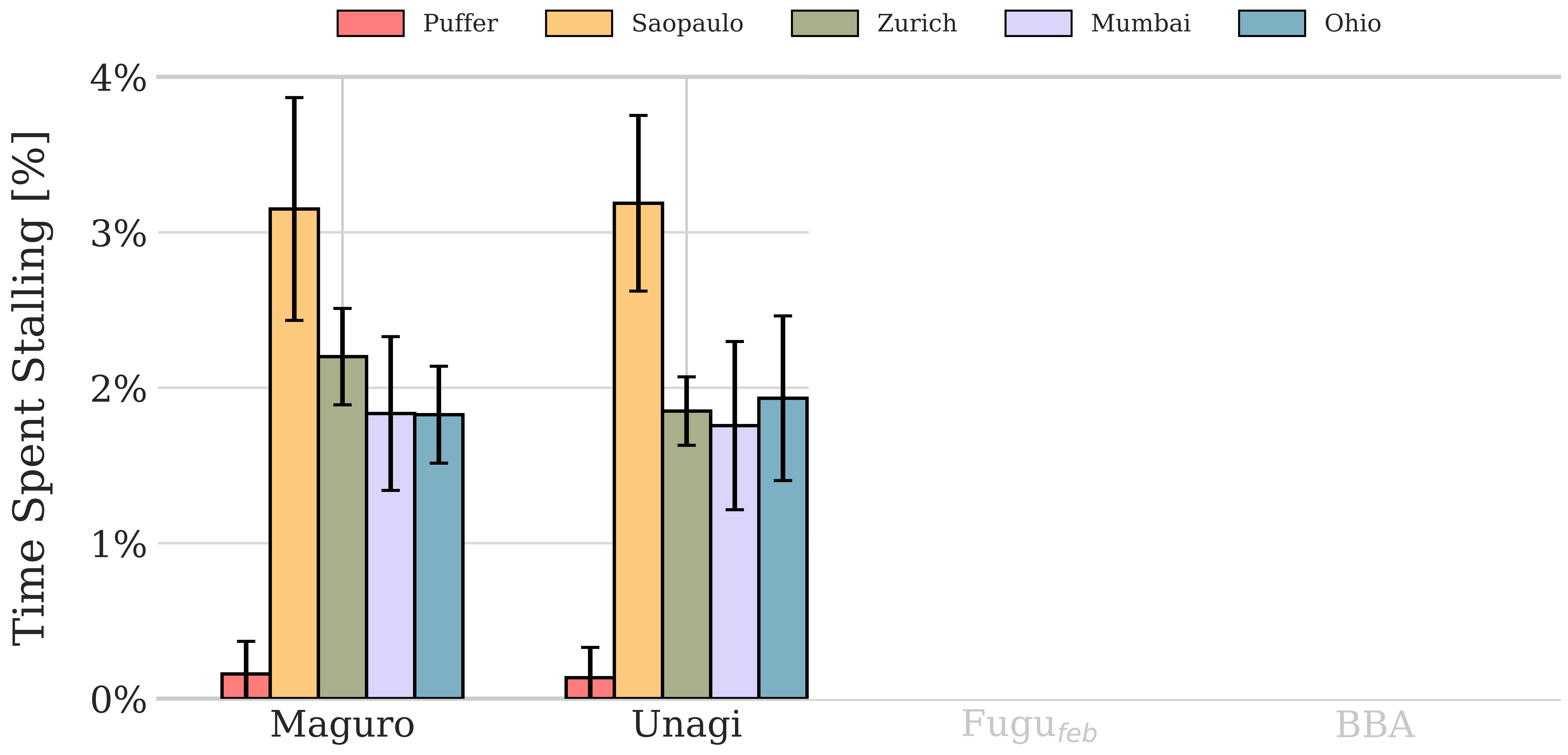


Testing outside of training context

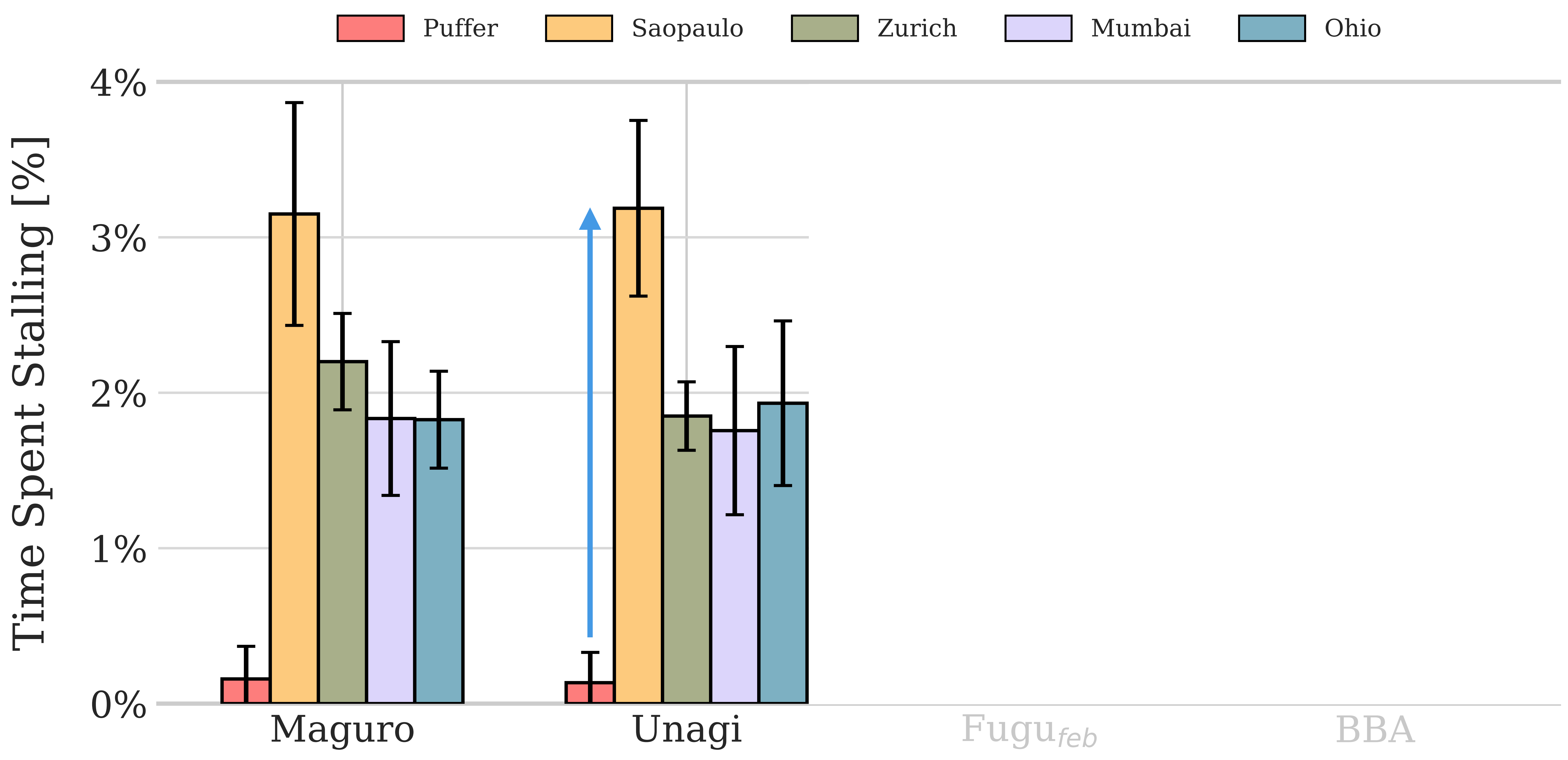


→ *There is likely a gap between results on Puffer and deployments in practice!*

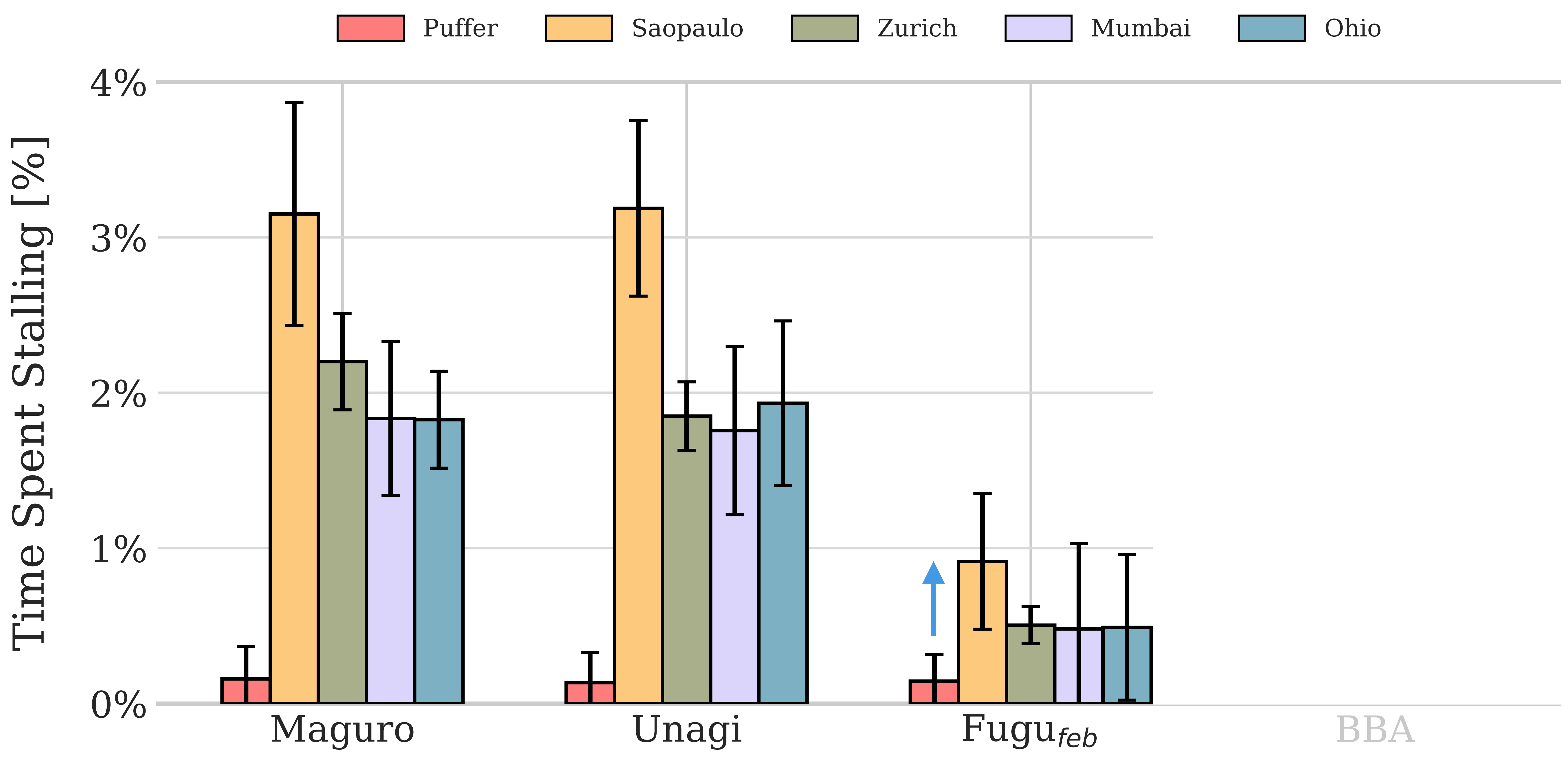
Testing outside of training context



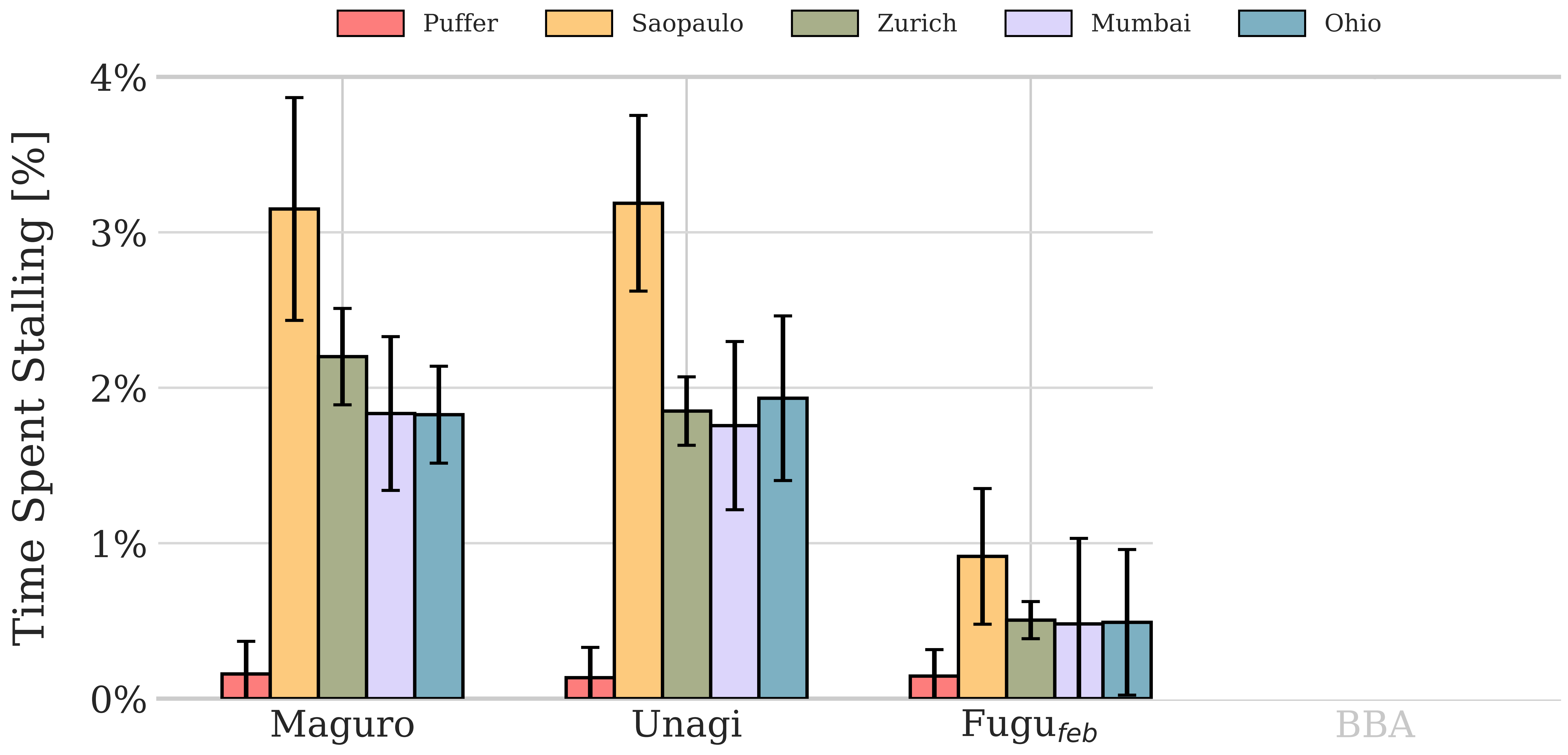
Testing outside of training context



Testing outside of training context

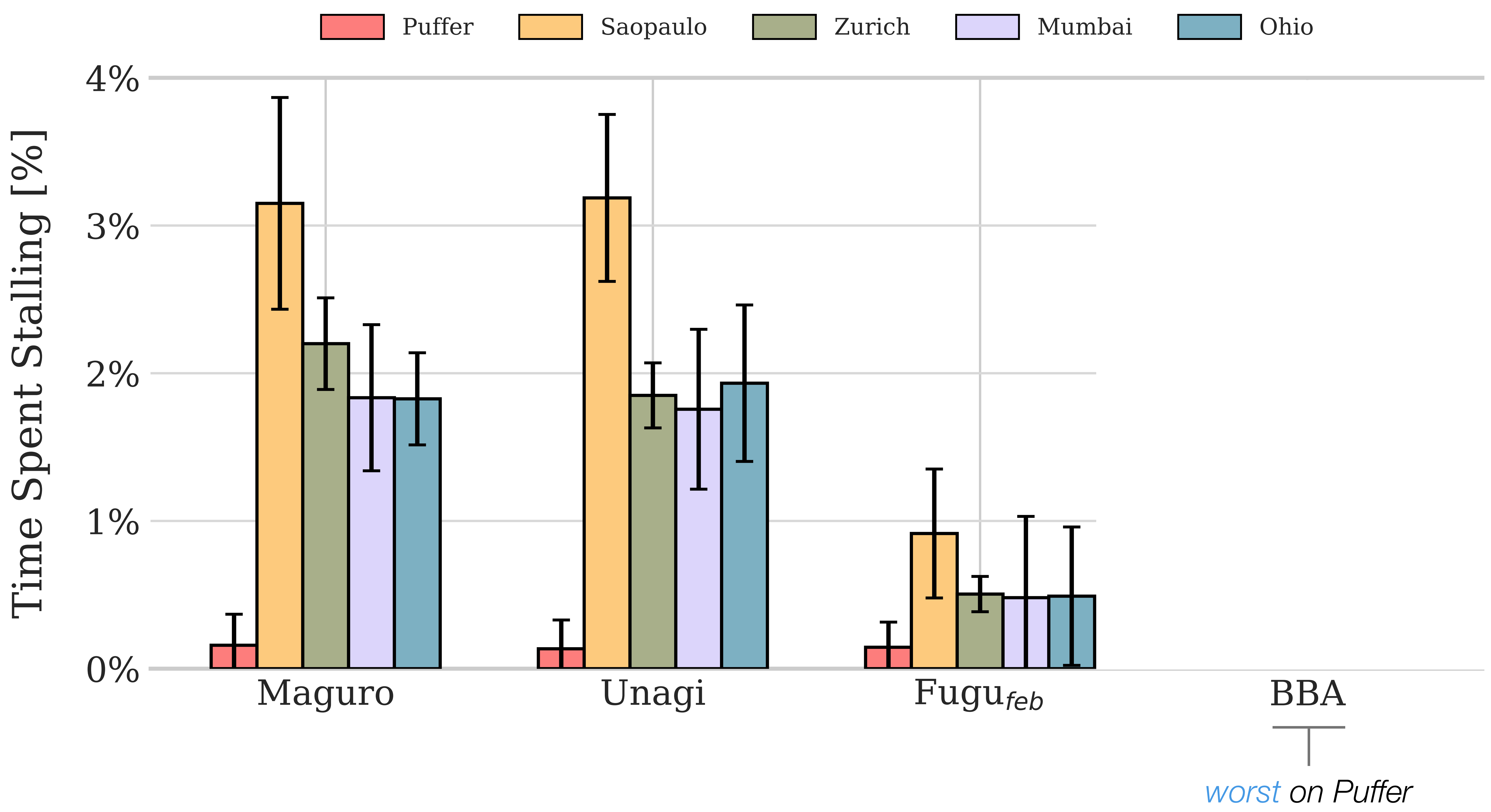


Testing outside of training context

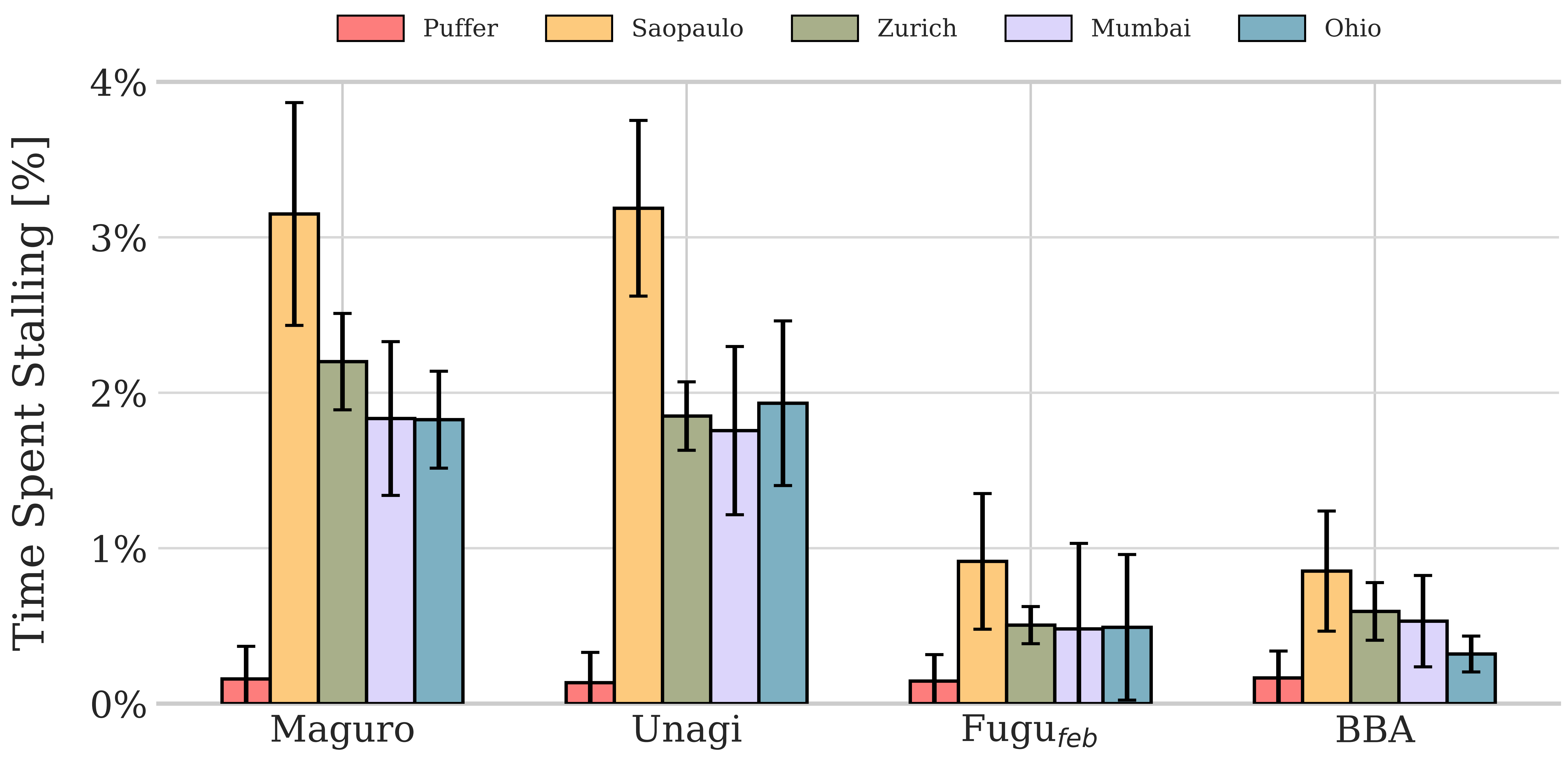


→ *There is ~~likely~~ a gap between results on Puffer and deployments in practice!*

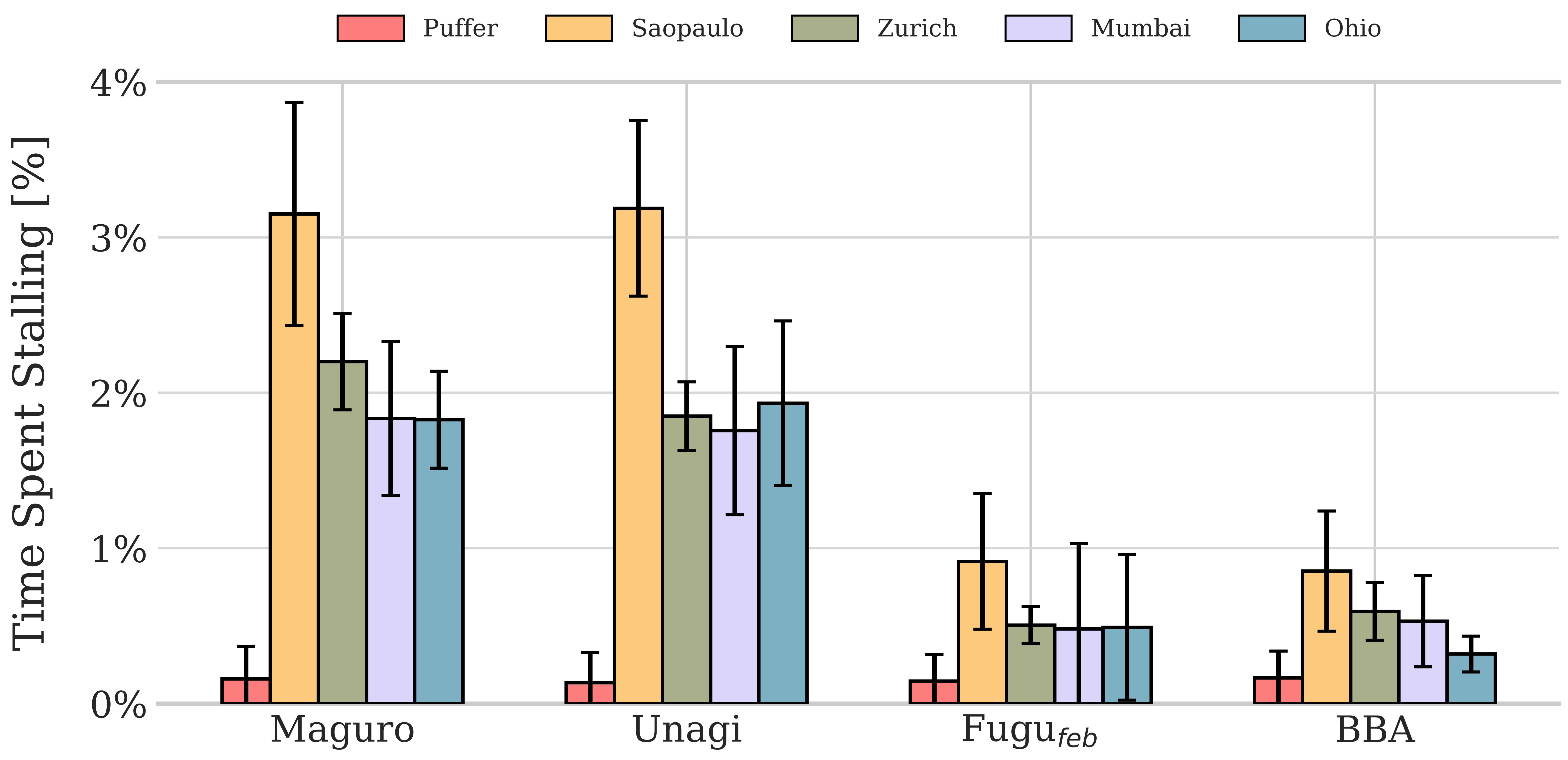
Testing outside of training context



Testing outside of training context



Testing outside of training context



→ *ABR-Arena allows for establishing a more robust leaderboard!*

Experiments and results

Experiments and results

- Performance differs significantly between training context and deployment

Experiments and results

- Performance differs significantly between training context and deployment
- Performance varies across our ABR-Arena deployments

Experiments and results

- Performance differs significantly between training context and deployment
- Performance varies across our ABR-Arena deployments
- Relative ranking between ABR algorithms varies

What’s the verdict?

What's the verdict?

ABR-Arena allows for:

What’s the verdict?

ABR-Arena allows for:

- Identifying the gap between results in training context versus in practice

What’s the verdict?

ABR-Arena allows for:

- Identifying the gap between results in training context versus in practice
- Establishing a robust comparison or leaderboard of ABR algorithms

What’s the verdict?

ABR-Arena allows for:

- Identifying the gap between results in training context versus in practice
- Establishing a robust comparison or leaderboard of ABR algorithms
- *Closing the performance gap by collecting diverse training data*

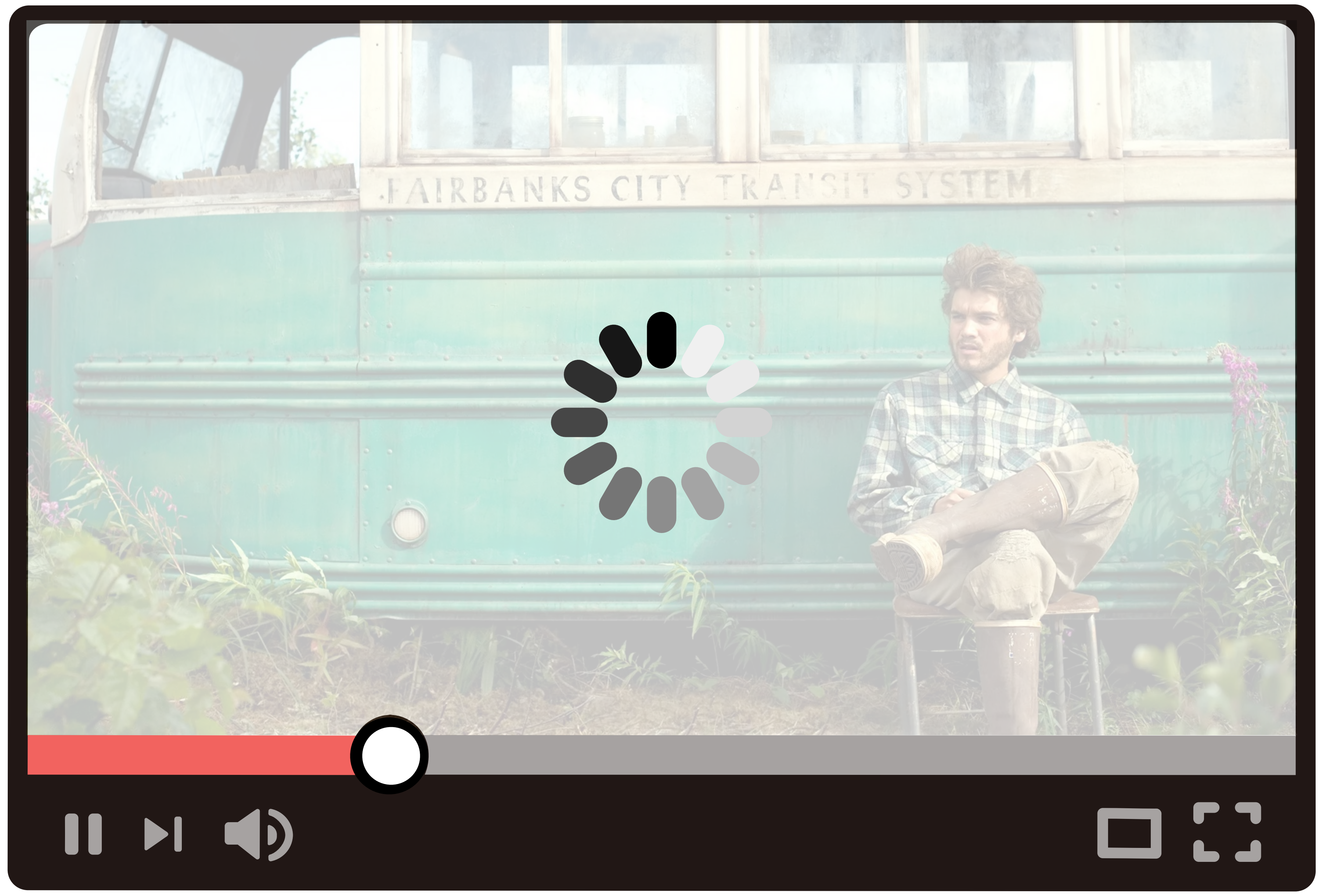
What's the verdict?

ABR-Arena allows for:

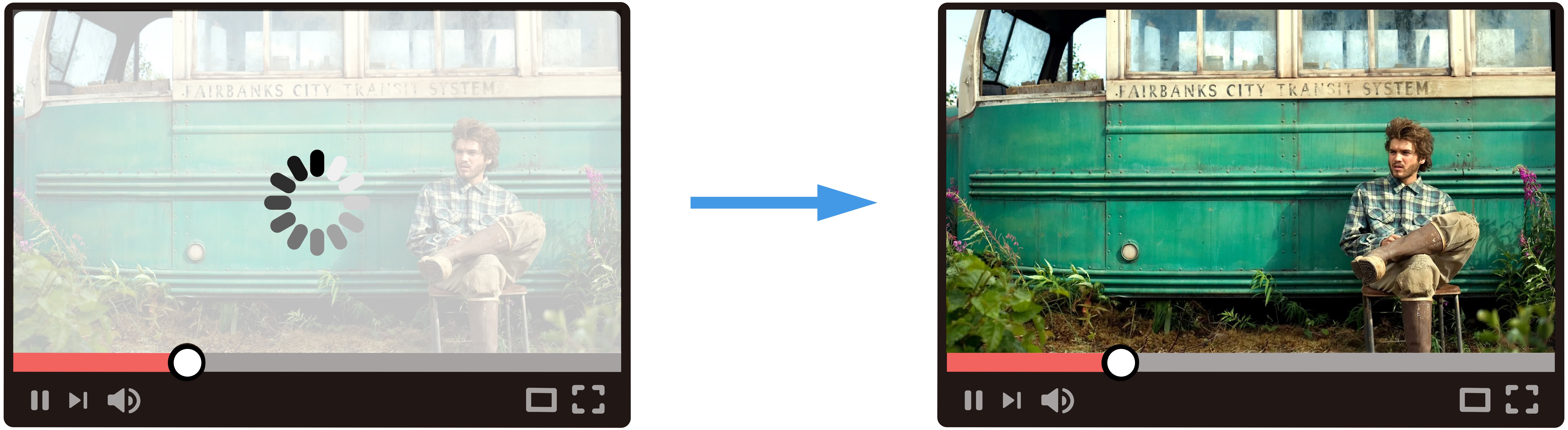
- Identifying the gap between results in training context versus in practice
- Establishing a robust comparison or leaderboard of ABR algorithms
- *Closing the performance gap by collecting diverse training data*

→ *We will open-source ABR-Arena to support the community in their research*

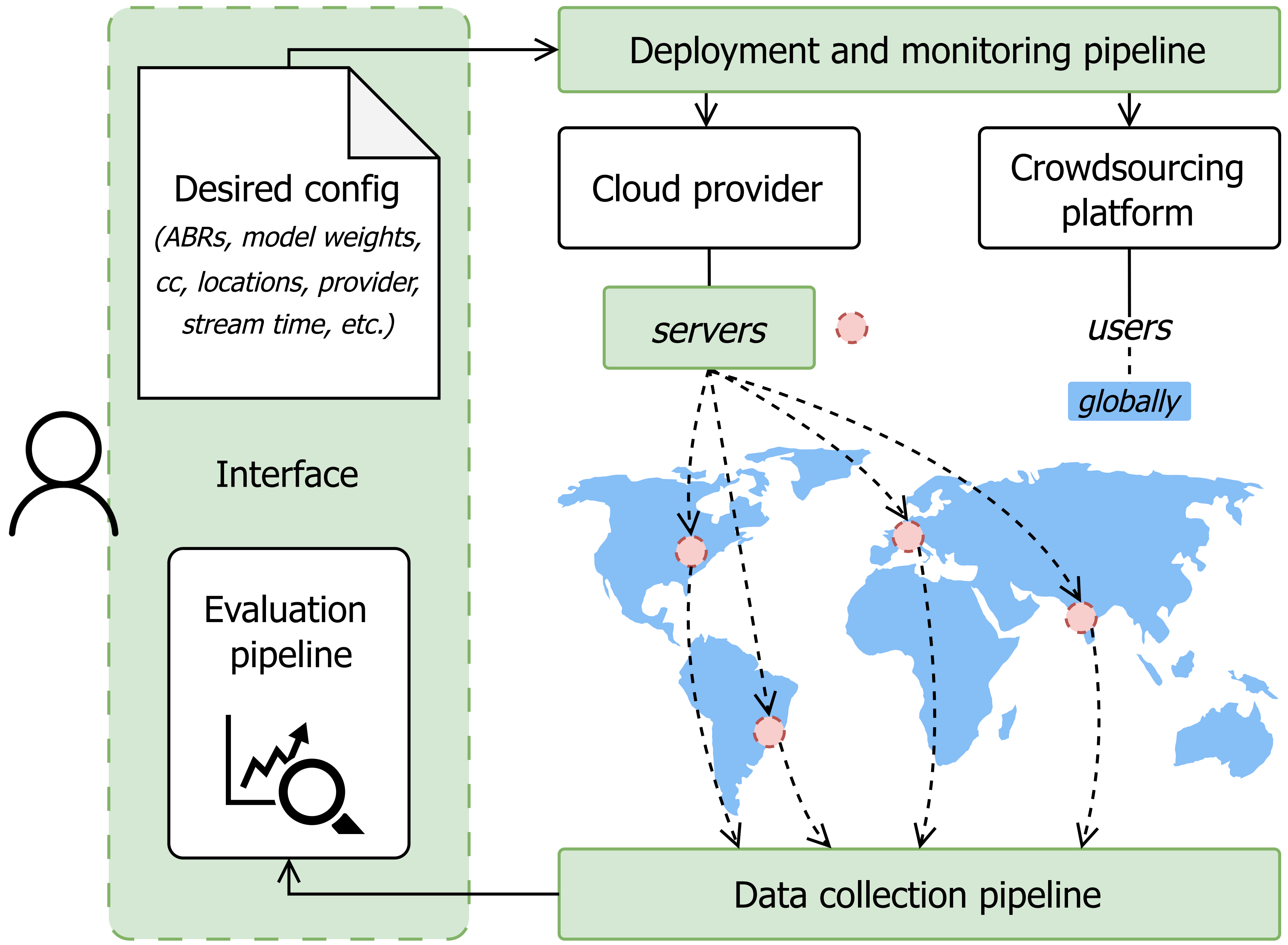
What’s the verdict?



What's the verdict?



Thank you for your attention!



Find ABR-Arena:

- ***Website: abrarena.com***
- *Code: github.com/nsg-ethz/ABR-Arena*

Get in touch:

- *Linkedin: linkedin.com/in/benjaminhoff*
- *Email: bhoffman@ethz.ch*

References

[1] A. Dietmüller, et al., “On Sample Selection for Continual Learning: a Video Streaming Case Study,” 2024

[2] Sandvine Corporation, “Video Permeates, Streaming Dominates,” 2023

[3] F. Yan, et al., “Learning in situ: a randomized experiment in video streaming,” 2020

[4] S. Patel, et al., “Practically High Performant Neural Adaptive Video Streaming,” 2024

[5] T. Huang, et al., “A buffer-based approach to rate adaptation: evidence from a large video streaming service,” 2014

[1] A. Dietmüller, et al., 2024 | [2] Sandvine Corporation, 2023 | [3] F. Yan, et al., 2020 | [4] S. Patel, et al., 2020 | [5] T. Huan, et al., 2014