Distributed On-Demand Deployment for Transparent Access to 5G Edge Computing Services

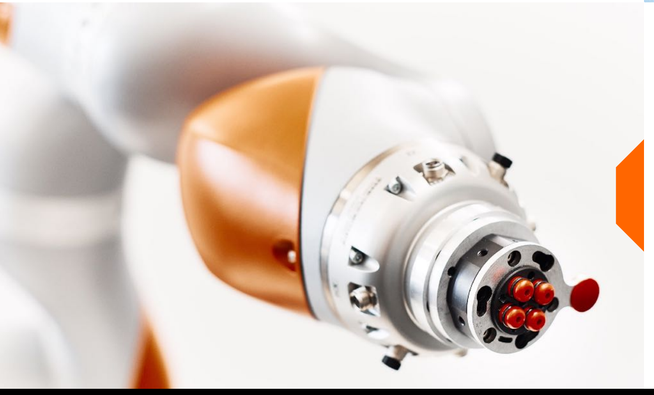
Josef Hammer
Hermann Hellwagner

UNIVERSITÄT KLAGENFURT

5GPlayground.at Research Projects

Virtual Realities

Communication in Swarms



Wireless Industrial Robotics

Smart City

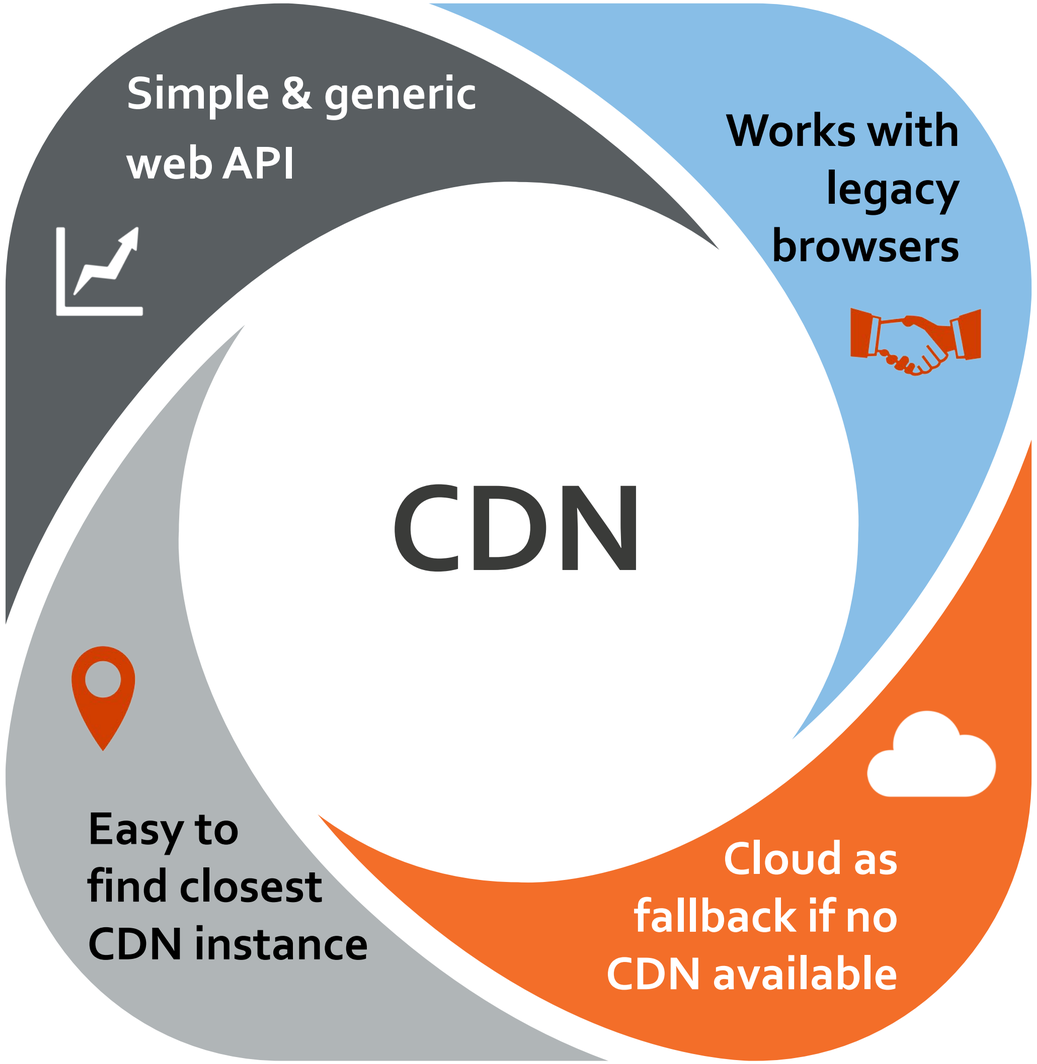
Q

How to access edge services?

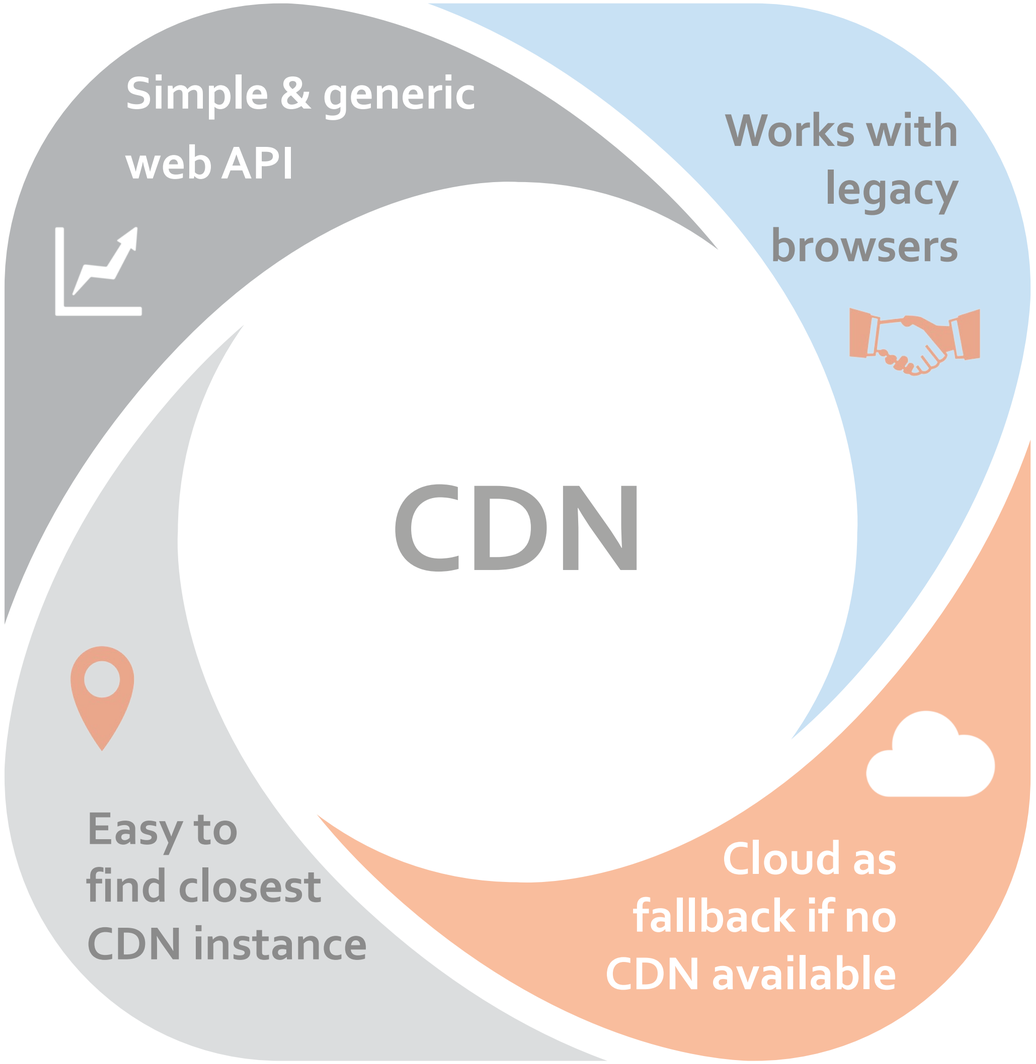
Transparent Access to Edge Clouds

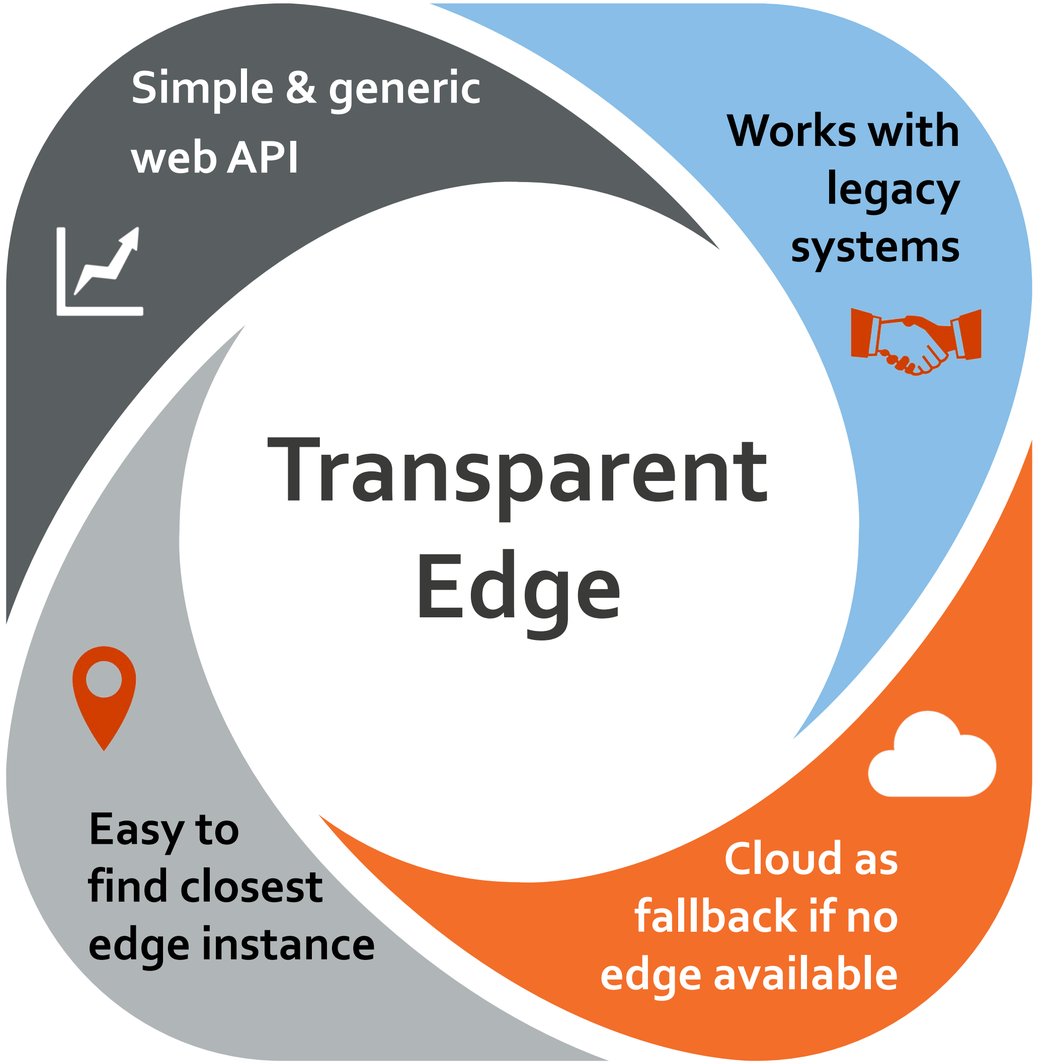
… similar to Content Delivery Networks (CDNs)

Why Transparent Access?



Why Transparent Access?

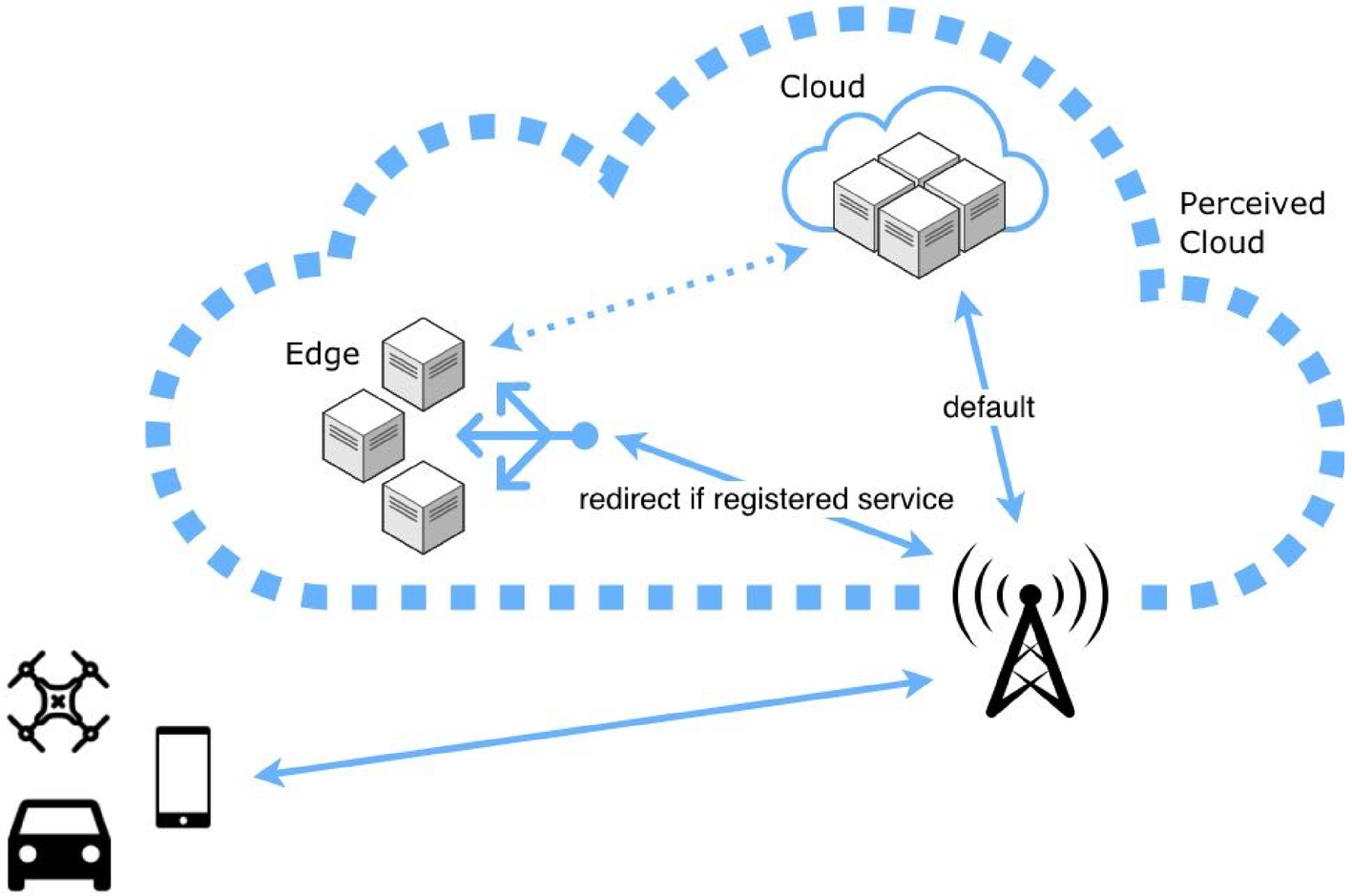




A

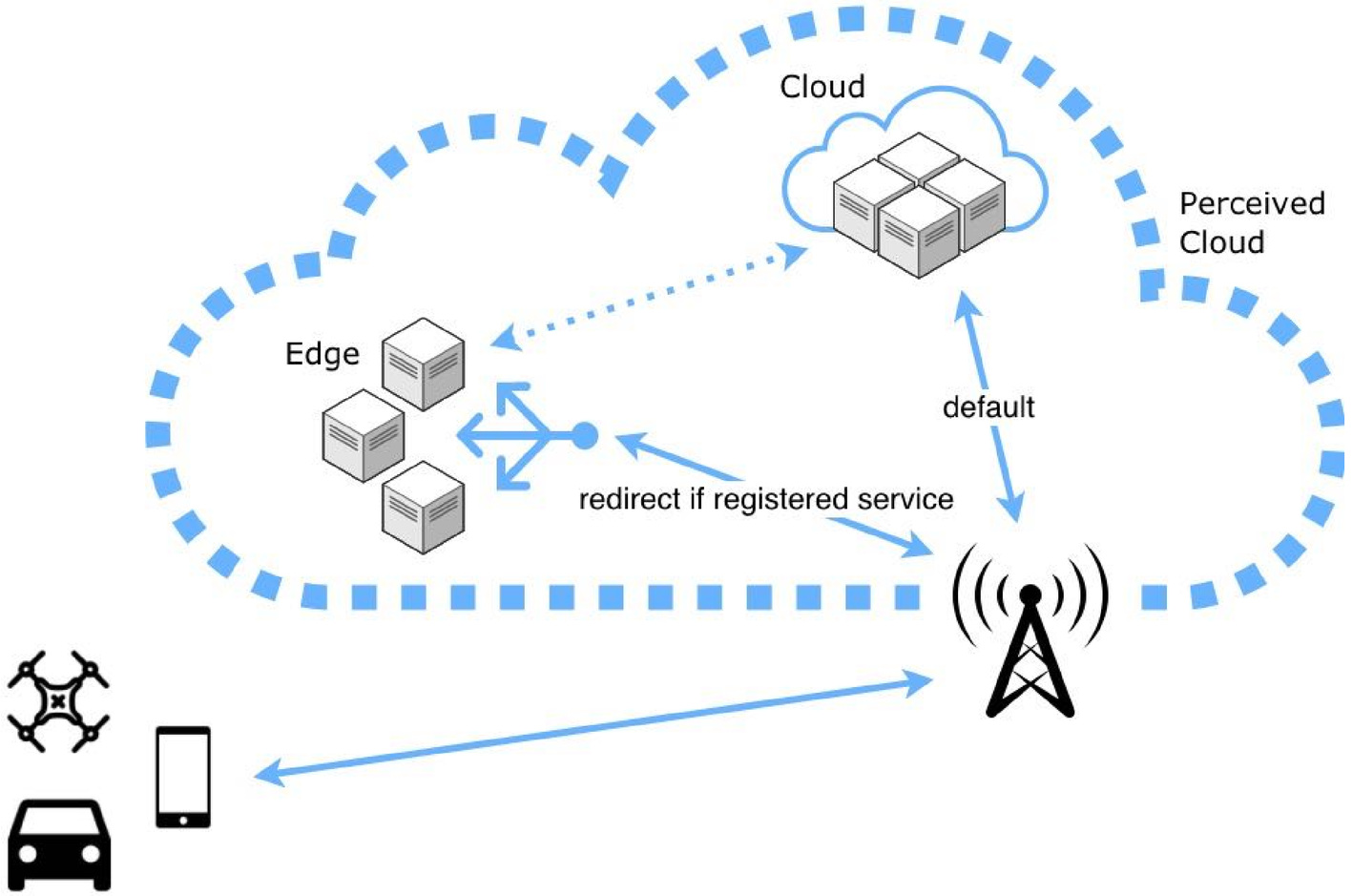
Approach

Use (virtual) cloud service IP addresses to access the local edge instance using SDN



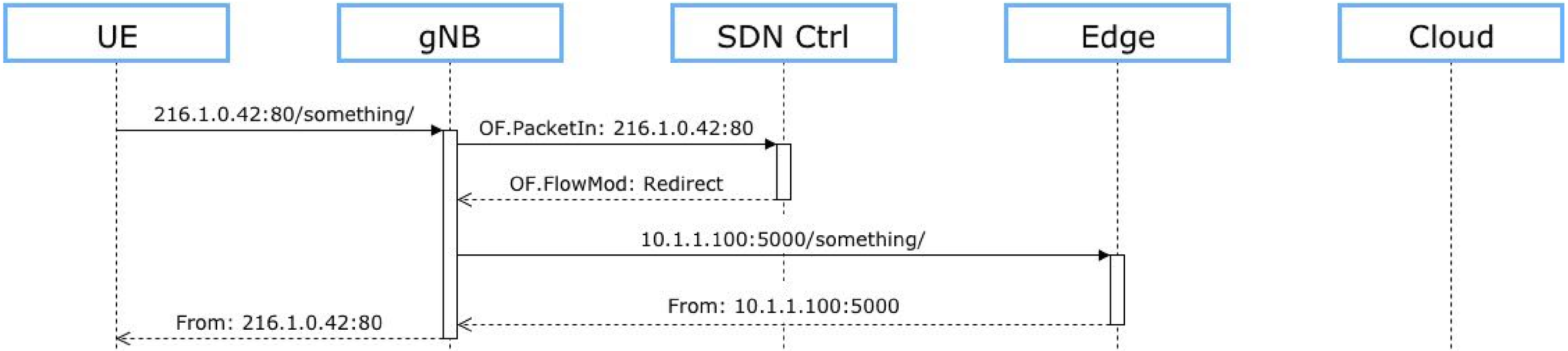
Use (virtual) cloud service IP addresses to access the local edge instance using SDN

The user (UE) seems to communicate with a remote cloud – it never sees the IP of the edge service



Routing with a registered service IP

Redirected to the closest/… edge server

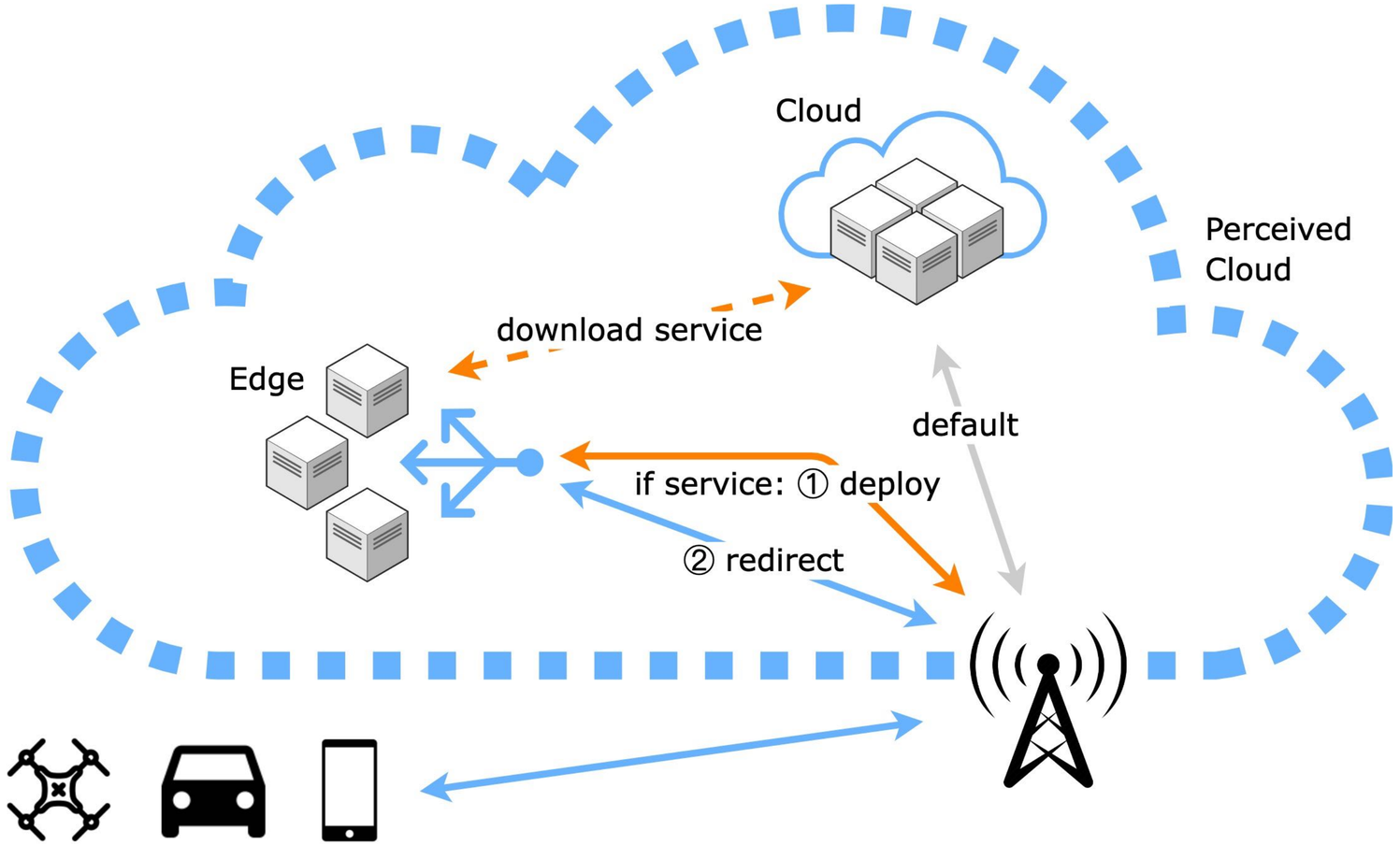


What if the service instance is not yet running in a nearby edge cluster?

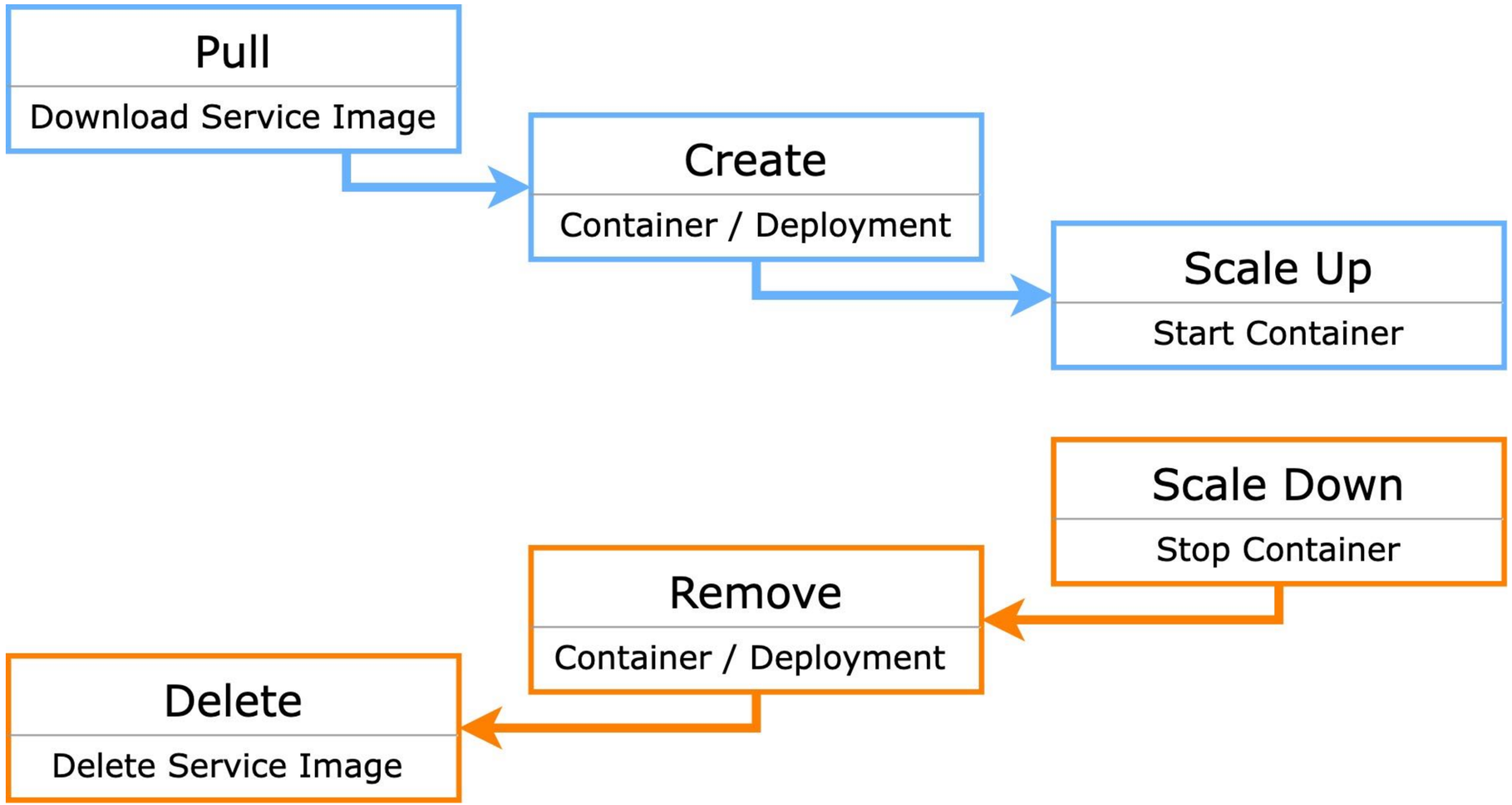
S

Solution:
On-Demand Deployment

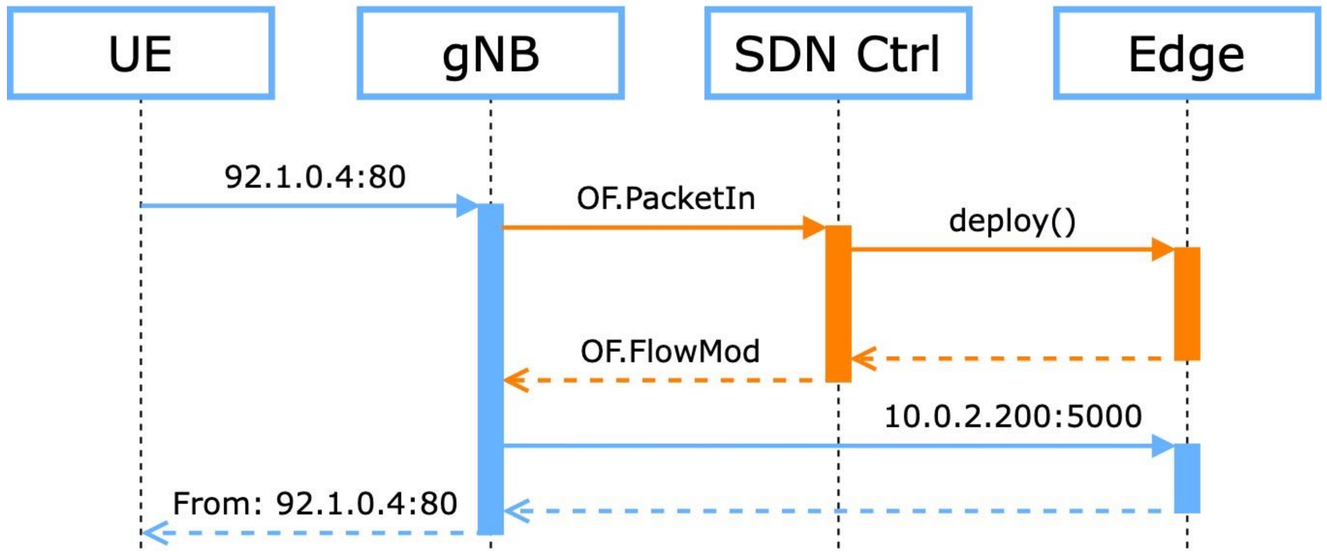
On-Demand Deployment

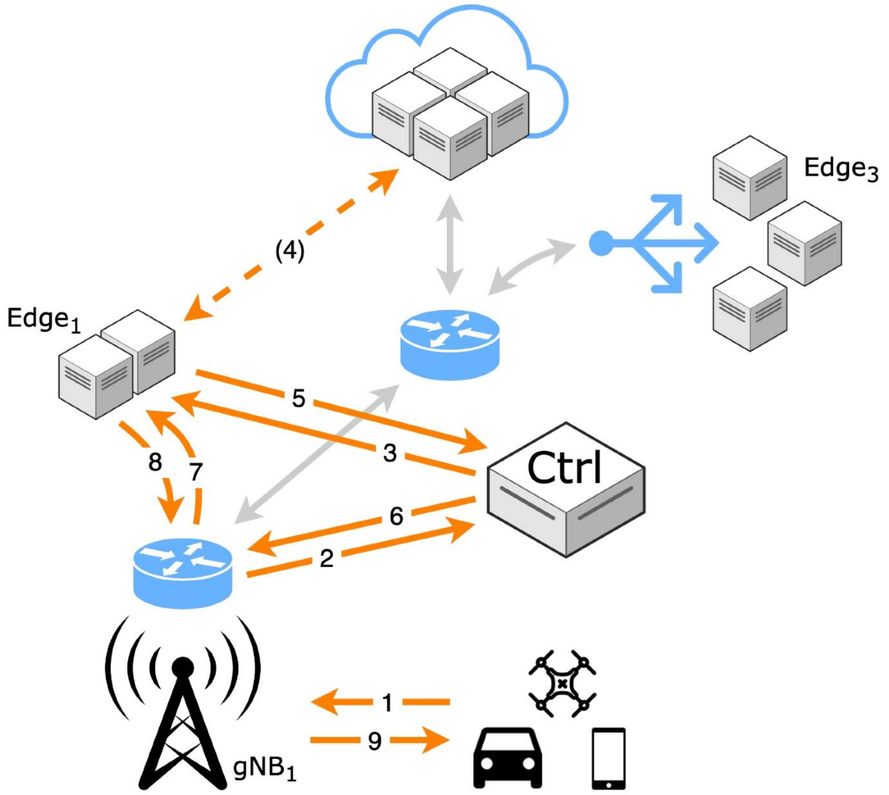


Deployment Phases



On-Demand Deployment with Waiting



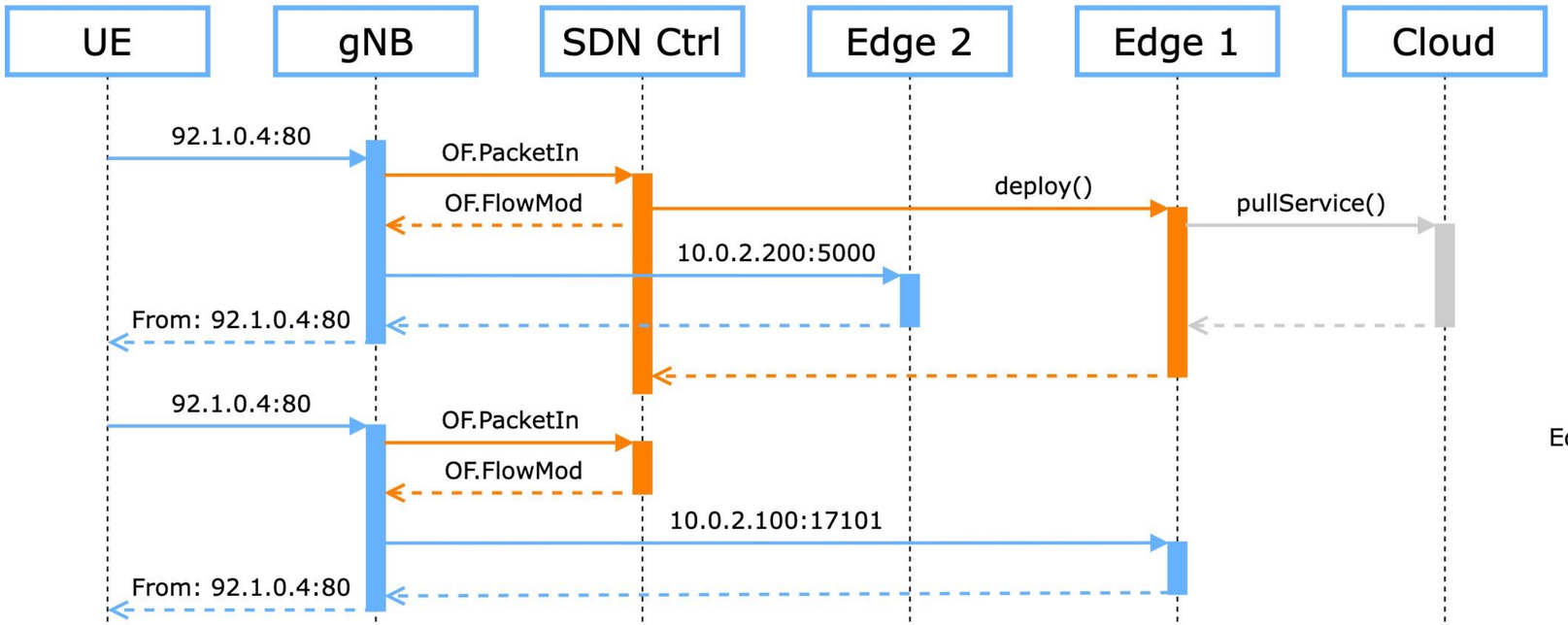


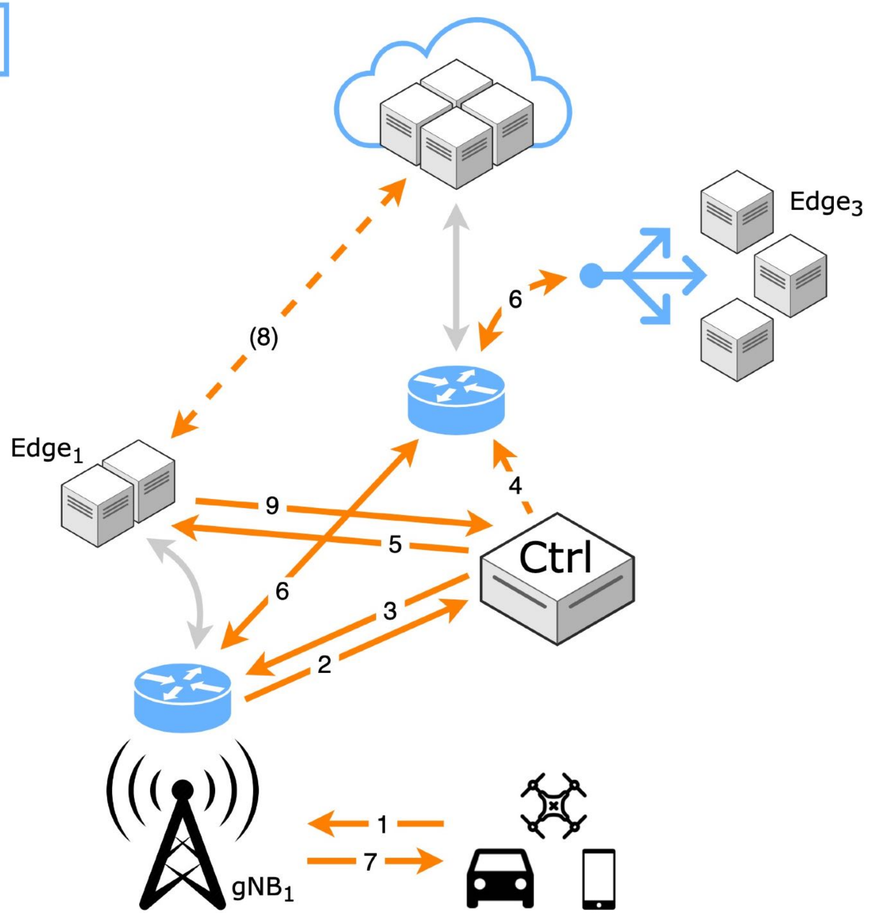
What if we need a lower latency?

S

Solution:
Two Choices:
FAST + BEST

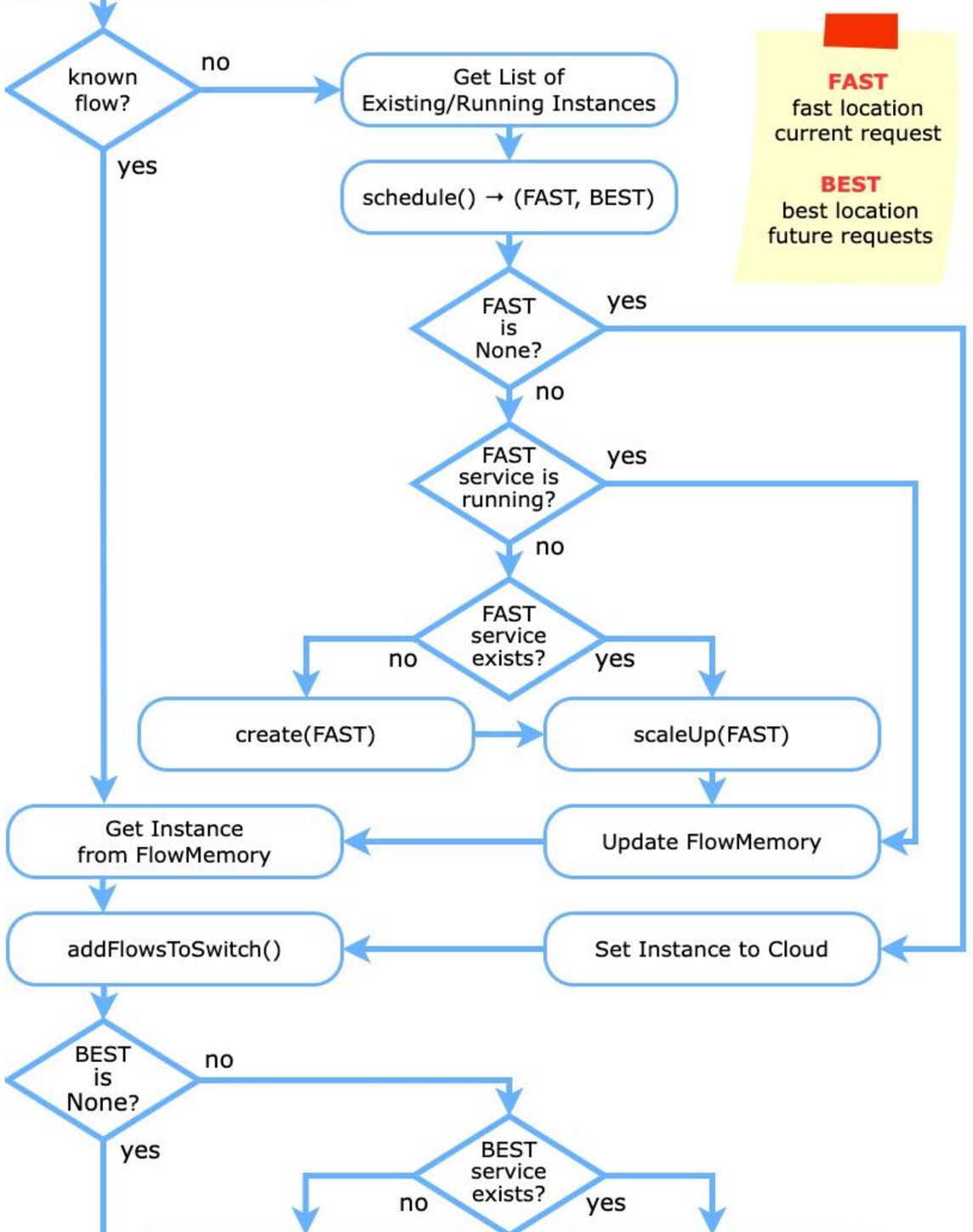
On-Demand Deployment without Waiting





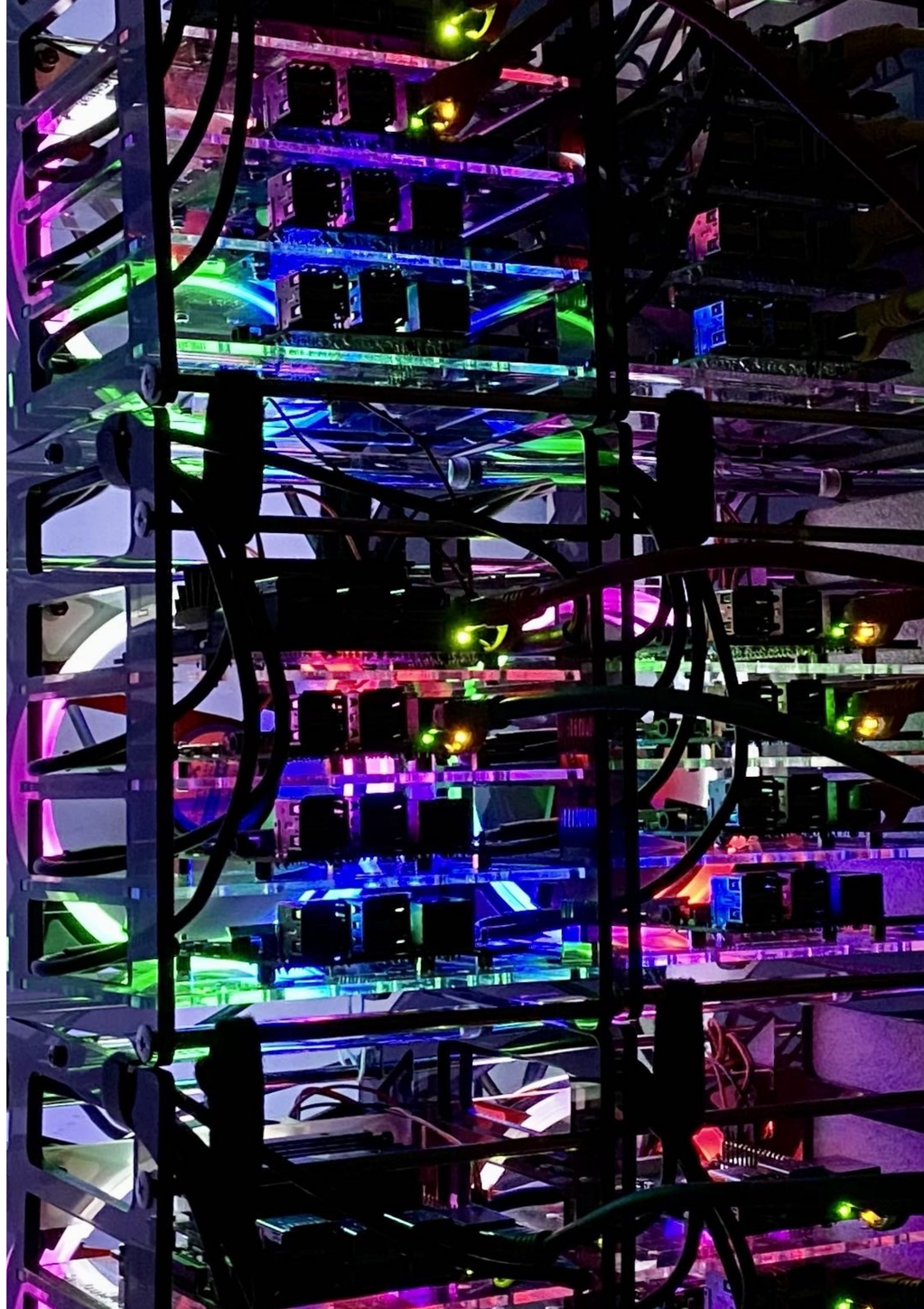
Dispatching Algorithm

- Flow memorized in **FlowMemory**?
- Dispatcher gathers a list of available and running instances and passes it to the Scheduler
- Scheduler returns **two choices: FAST + BEST**



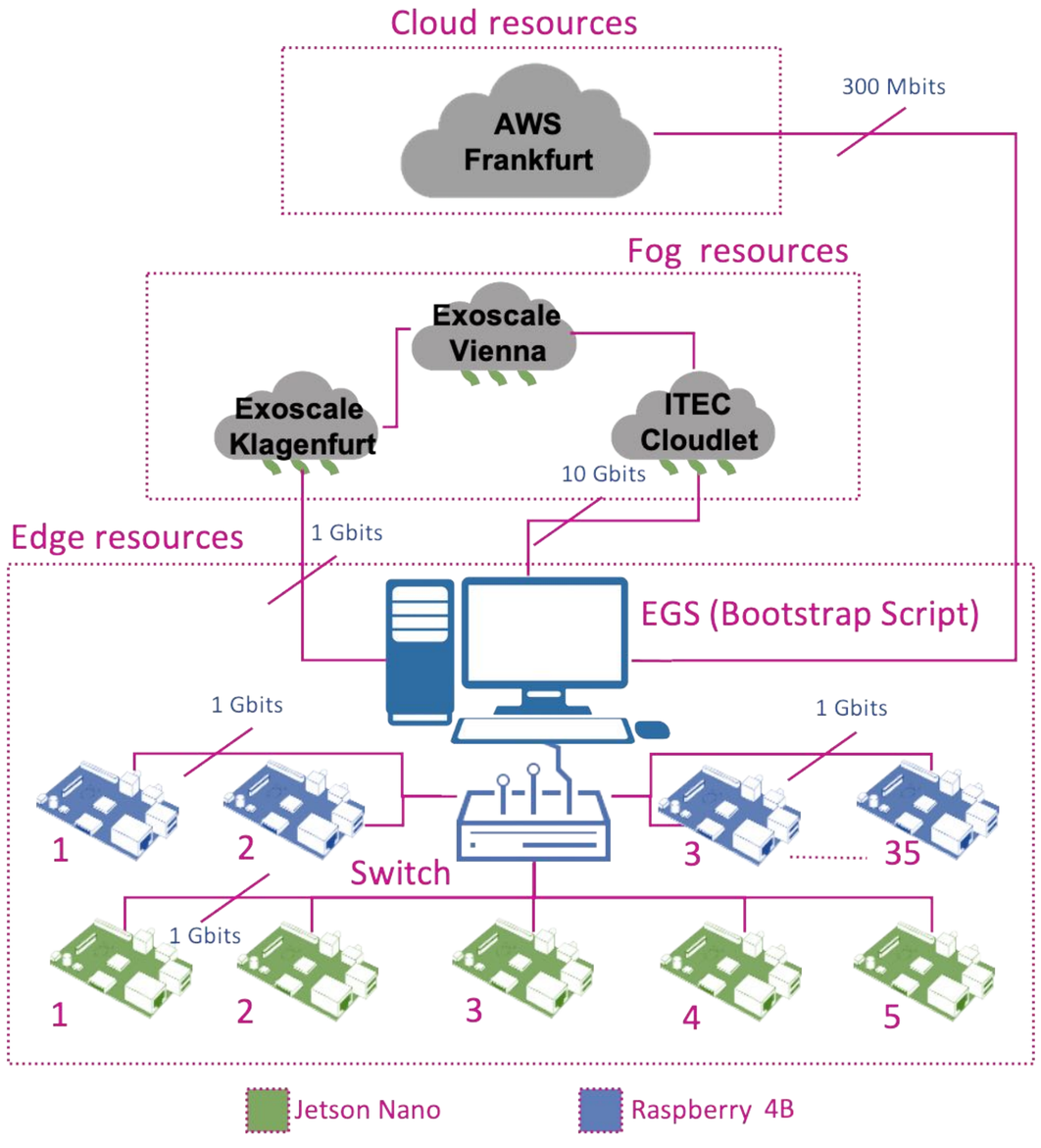
E

Evaluation



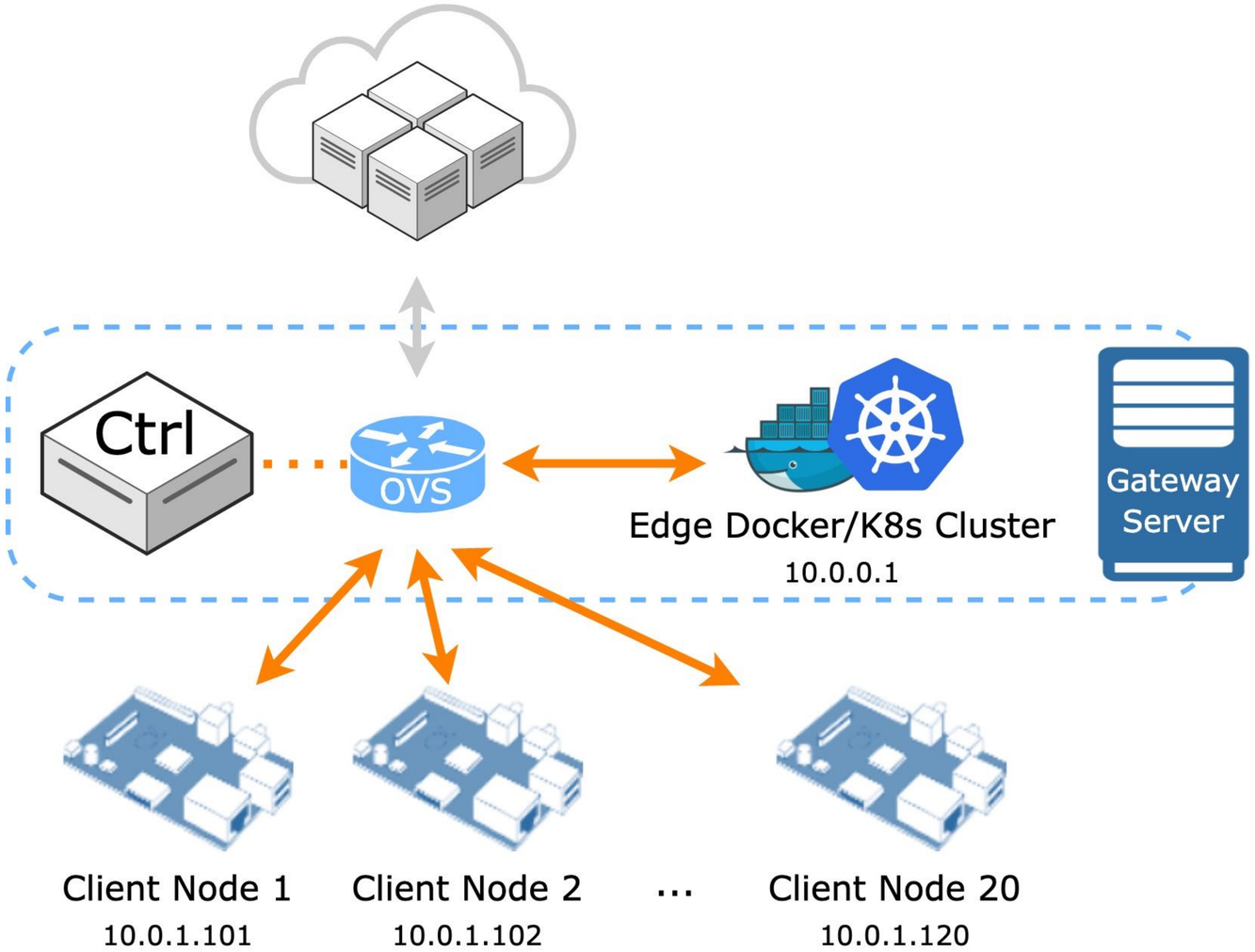
Carinthian Computing Continuum (C³) Testbed

c3.itec.aau.at



Evaluation Setup

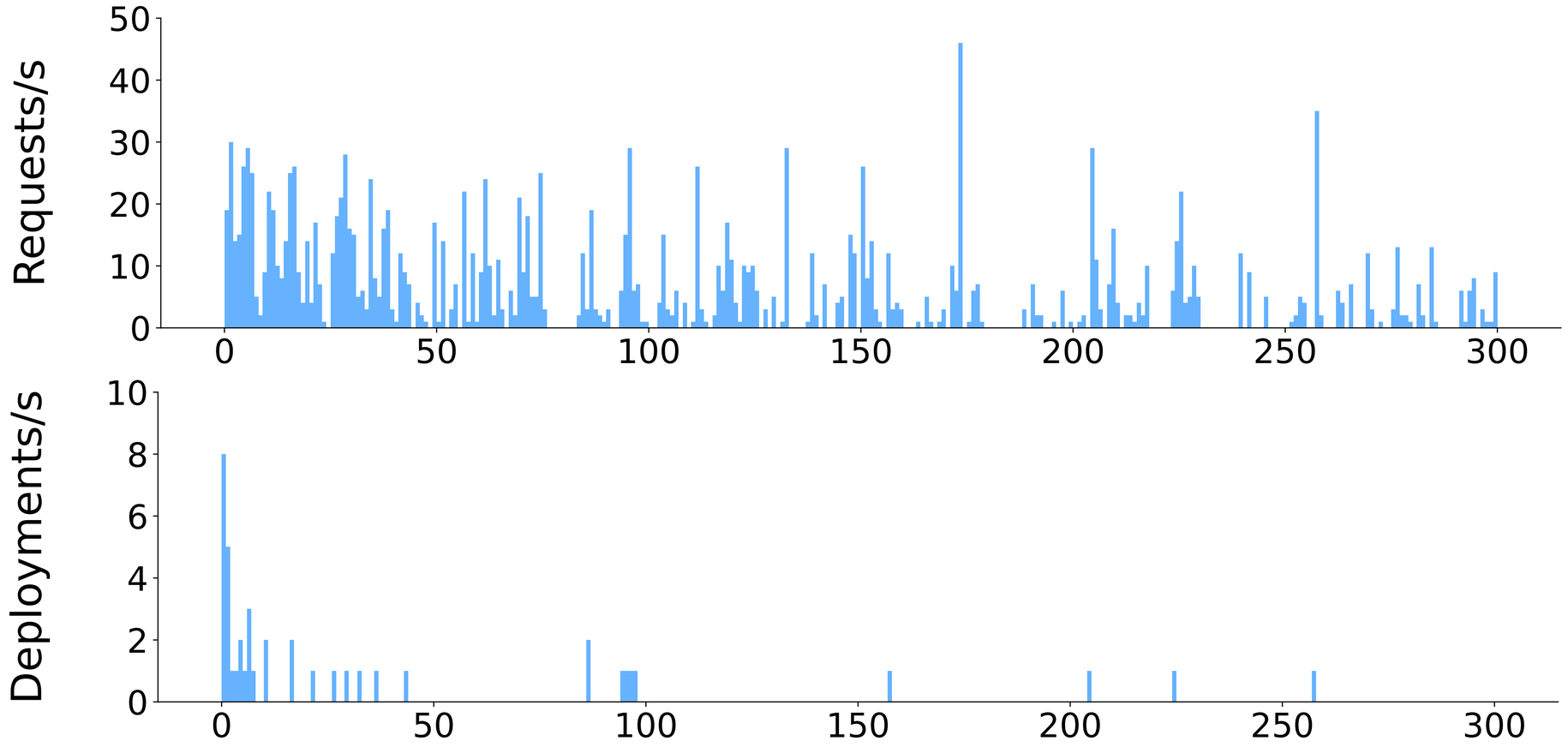
edge.itec.aau.at



Real Network Traffic Dataset

– **1708 requests** to **42 different edge services** over **five minutes**

– deployment if service is not running yet

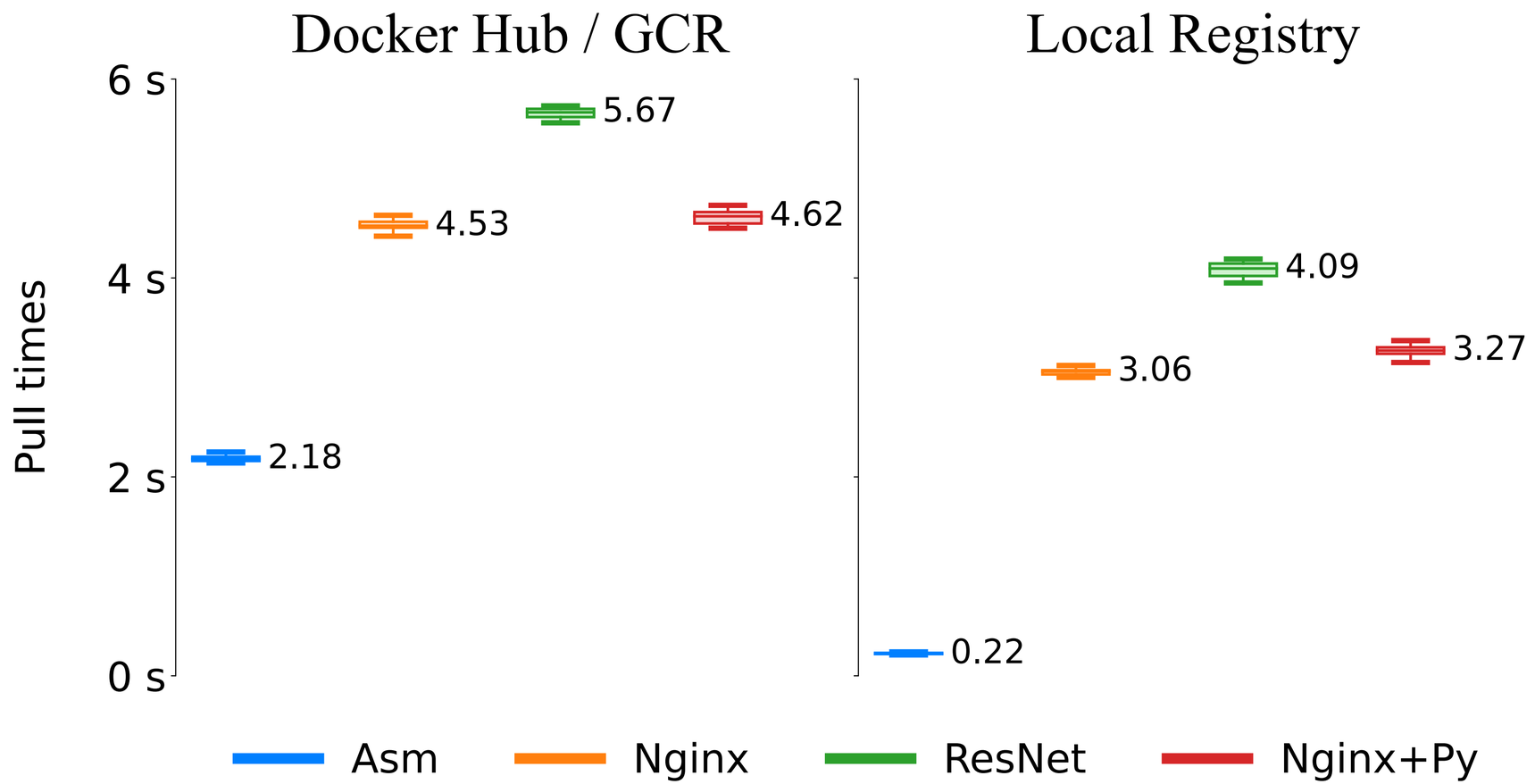


Services

	Service	Image(s)	Size / Layers	Containers	HTTP
`Asm`	Assembler Web Server (asmttpd [28])	josefhammer/web-asm:amd64	6.18 KiB / 1	1	GET
`Nginx`	Nginx Web Server	nginx:1.23.2	135 MiB / 6	1	GET
`ResNet`	TensorFlow Serving with pre-trained ResNet50 model	gcr.io/tensorflow-serving/resnet	308 MiB / 9	1	POST
`Nginx+Py`	Nginx Web Server + Python Application	nginx:1.23.2 + josefhammer/env-writer-py	181 MiB / 7	2	GET

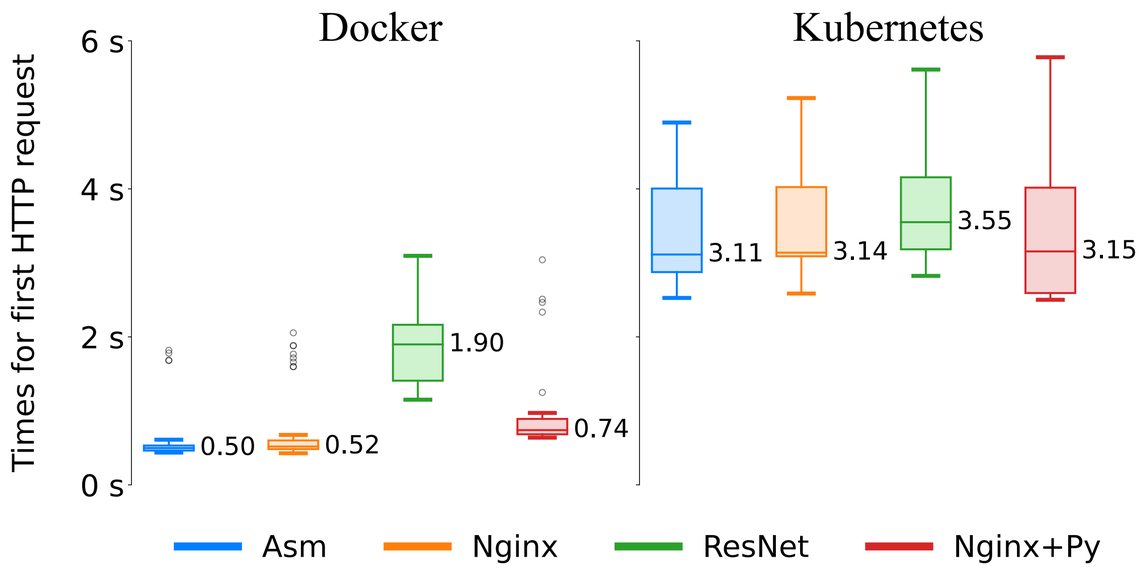
Pull Times

	Size / Layers	Containers
Asm	6.18 KiB / 1	1
Nginx	135 MiB / 6	1
ResNet	308 MiB / 9	1
Nginx+Py	181 MiB / 7	2

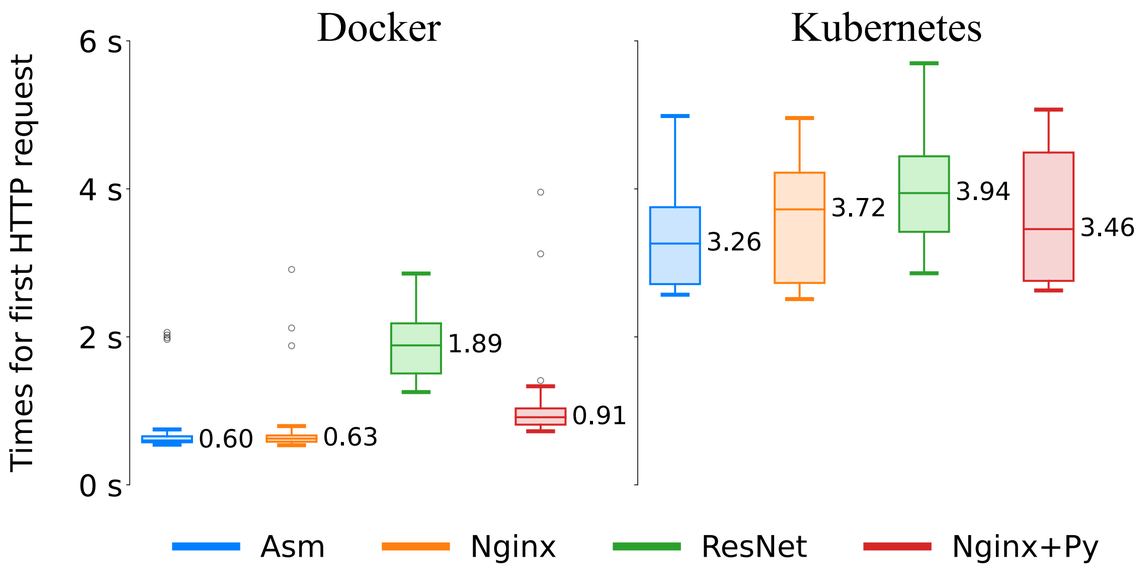


Create + Scale Up Times

Total time (median) to deploy **four different services** on **two different clusters** (42 instances per test).



Scale Up

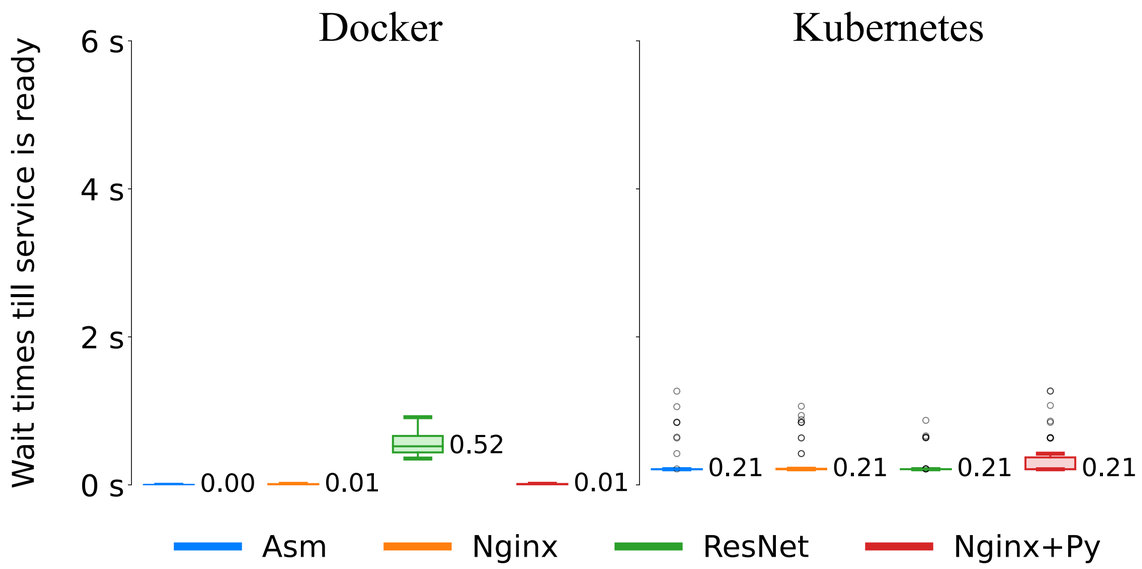


Create + Scale Up

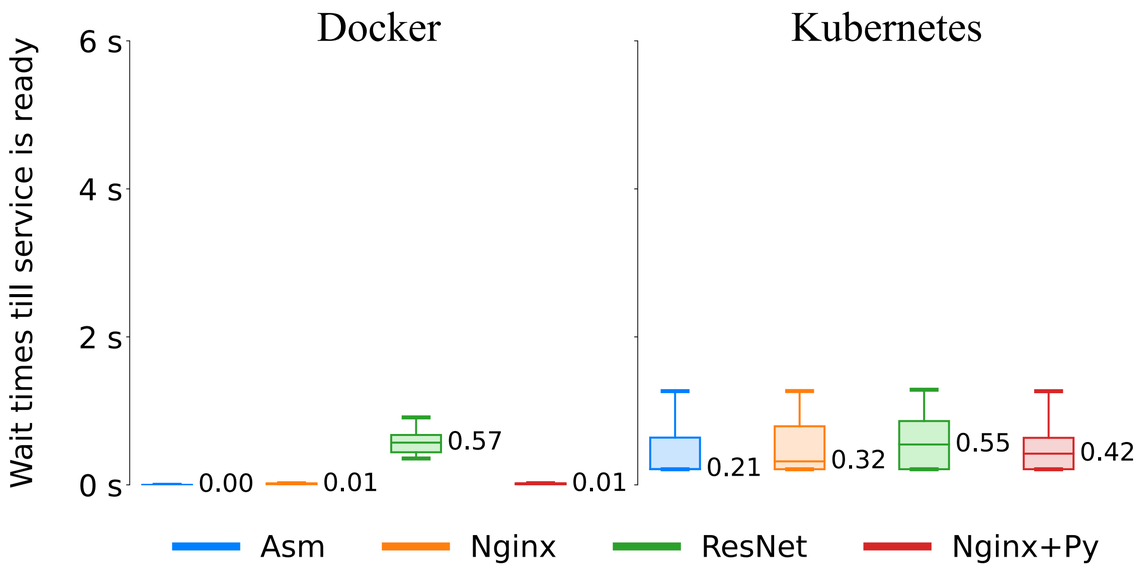
Wait Times for Open Port

Wait time (median) until the services are ready after being deployed on two different clusters.

SDN controller continuously tests whether the respective **port is open** before setting up the flows.



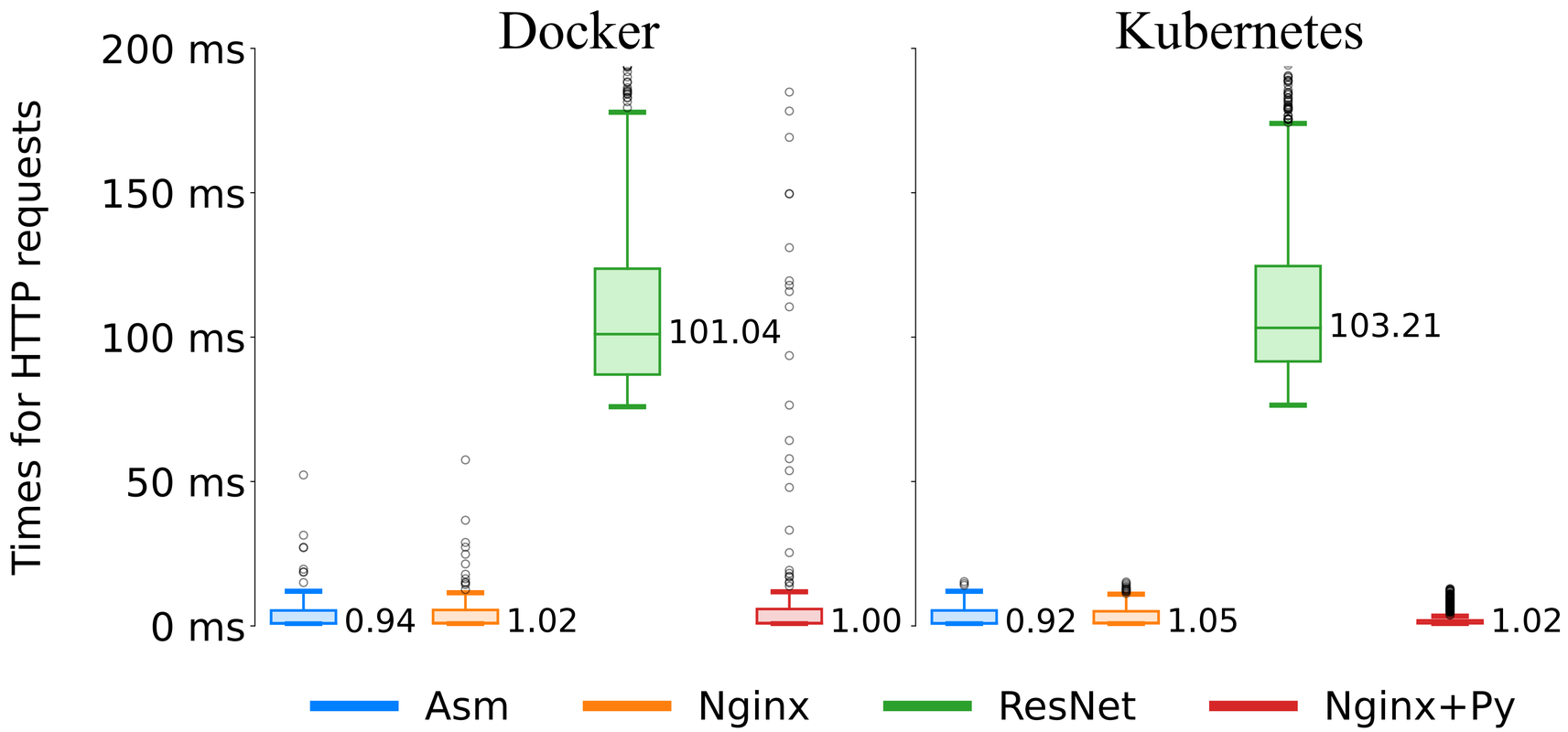
Scale Up



Create + Scale Up

Further Requests

Total time (median) for client requests to the edge services when the instance is already running on the cluster.



On-Demand Deployment For Transparent Edge Services:

1. FAST vs. BEST Choice
2. Less than a Second on Docker
   for Create + Scale Up

edge.josefhammer.com

josef.hammer@aau.at

www.aau.at