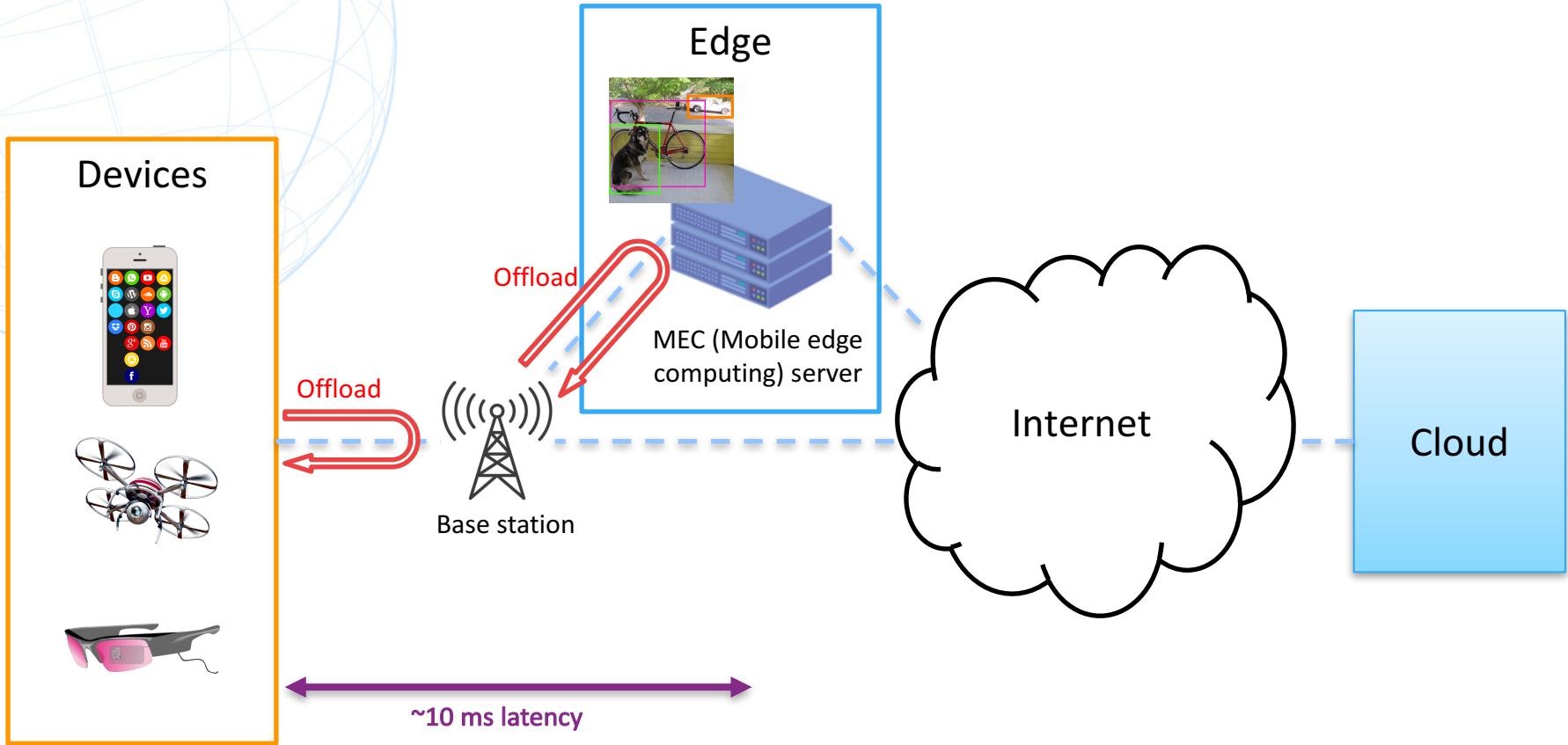


TVM for edge computing platform

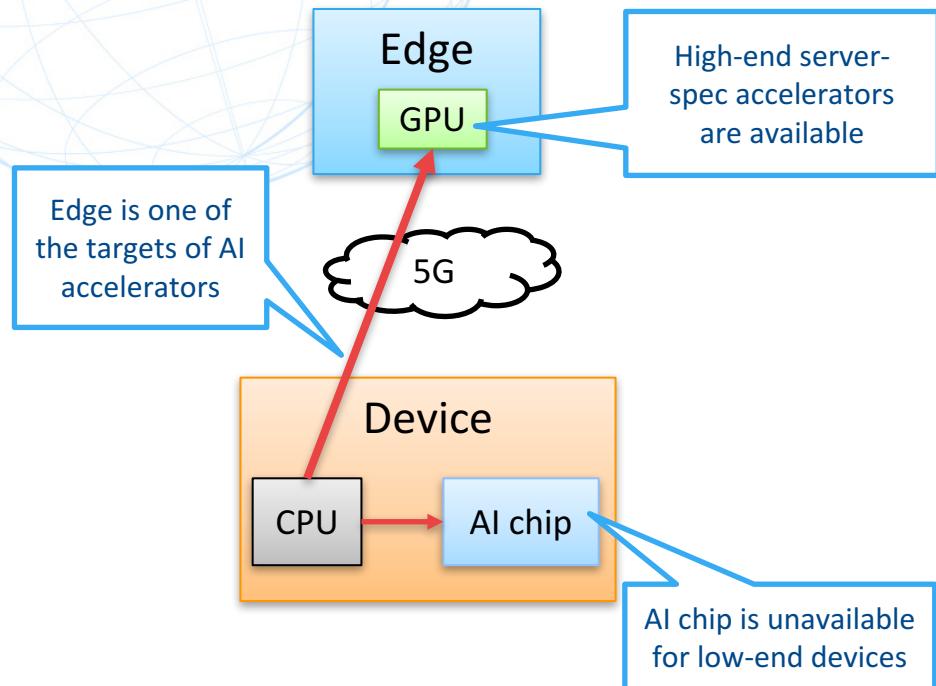
NTT Software Innovation Center
Kazutaka Morita

Inference in 5G era

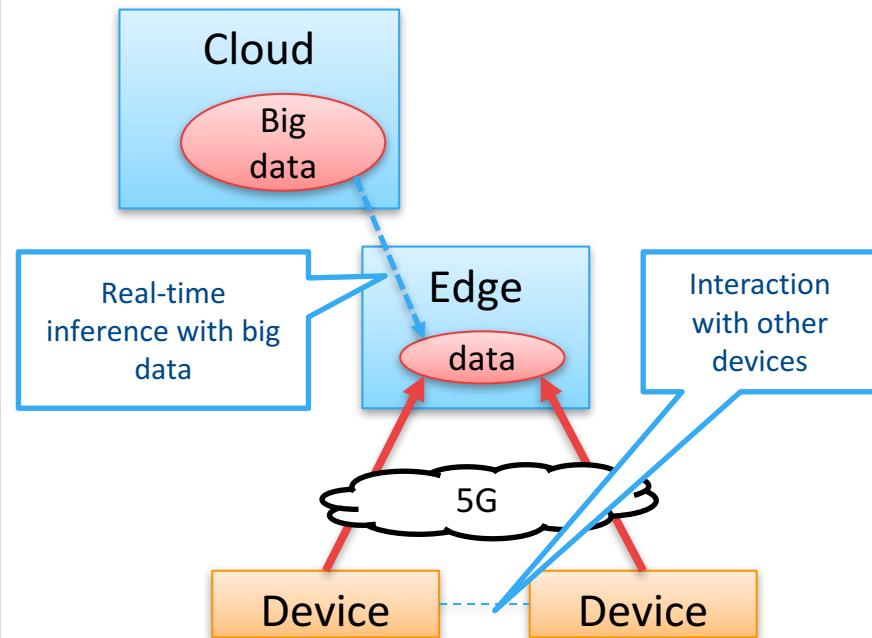


Benefits of offloading inference

Computing resource



Inference with data



Example – Augmented Reality

Occlusion



Object segmentation inference



object will not collide

Plane detection



Object detection inference
can also provide collider
from moving real world objects

bouncing object

Many Inference tasks

Point cloud



Point
cloud
data

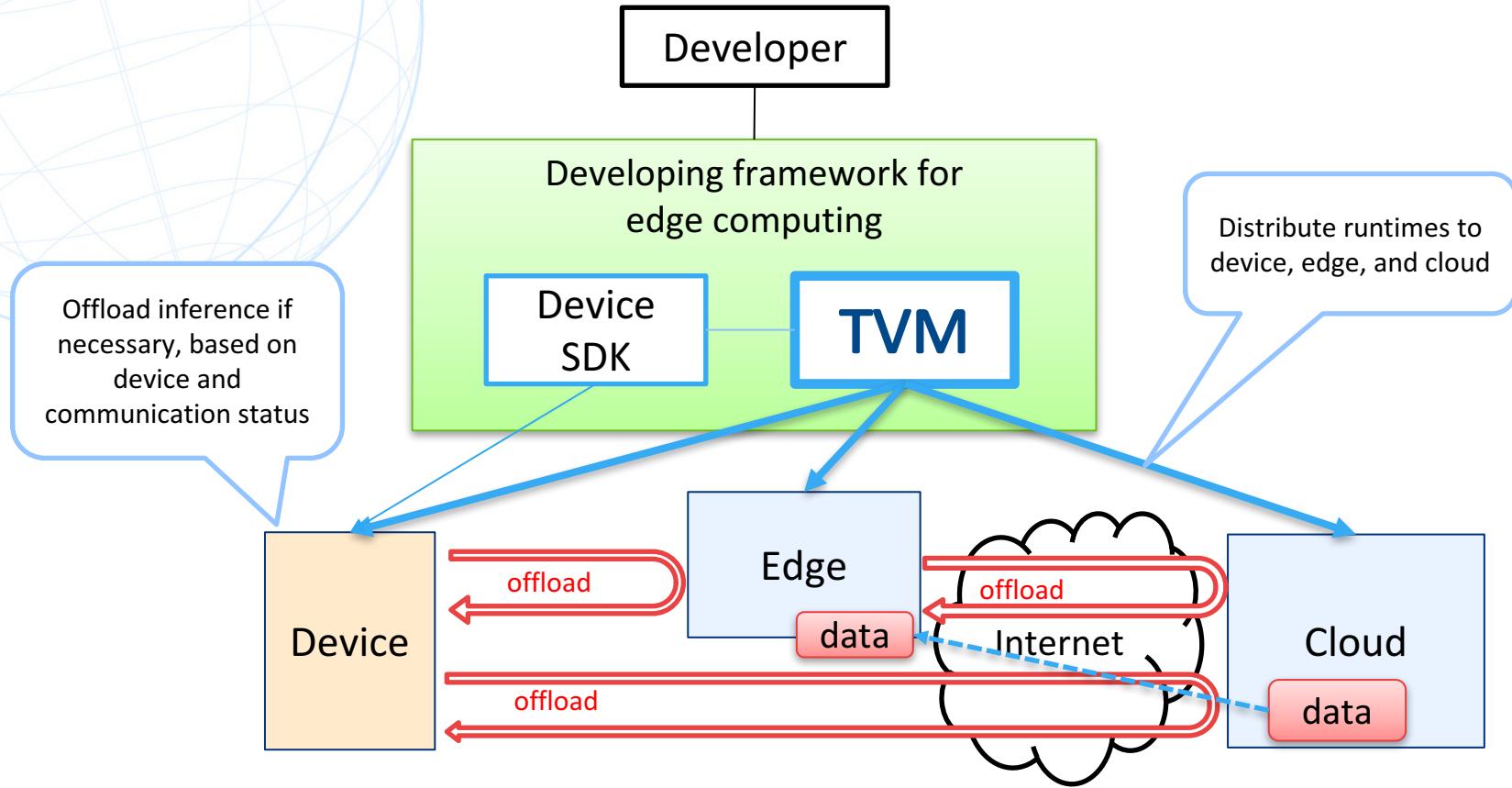
Captured
images



HYPER-REALITY: <https://vimeo.com/166807261>

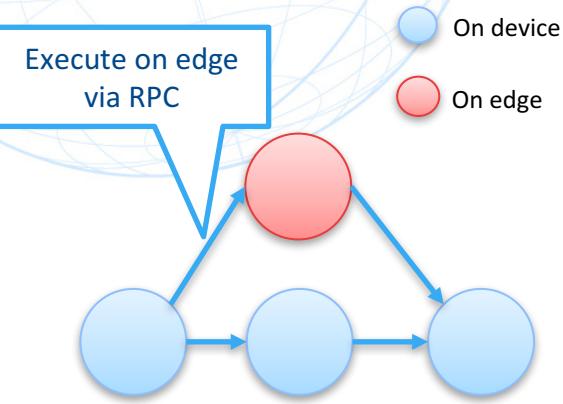
Inference with big data in the cloud

Edge computing platform with TVM



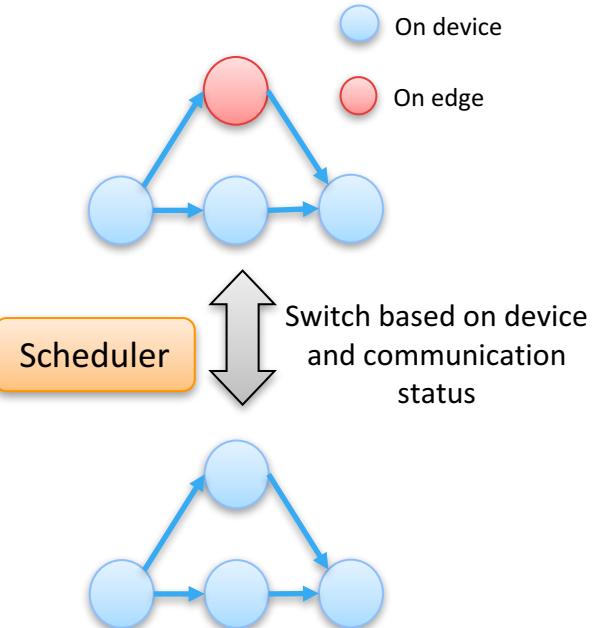
What are required for TVM?

Heterogeneous runtime with offloading support

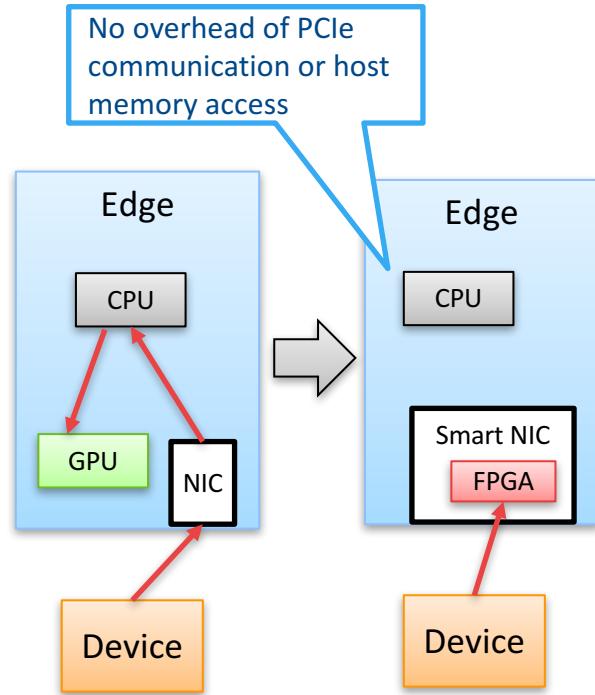


Auto tuning support would be also nice

Dynamic runtime



Smart NIC support



Thank You!

Email: kazutaka.morita.fp@hco.ntt.co.jp