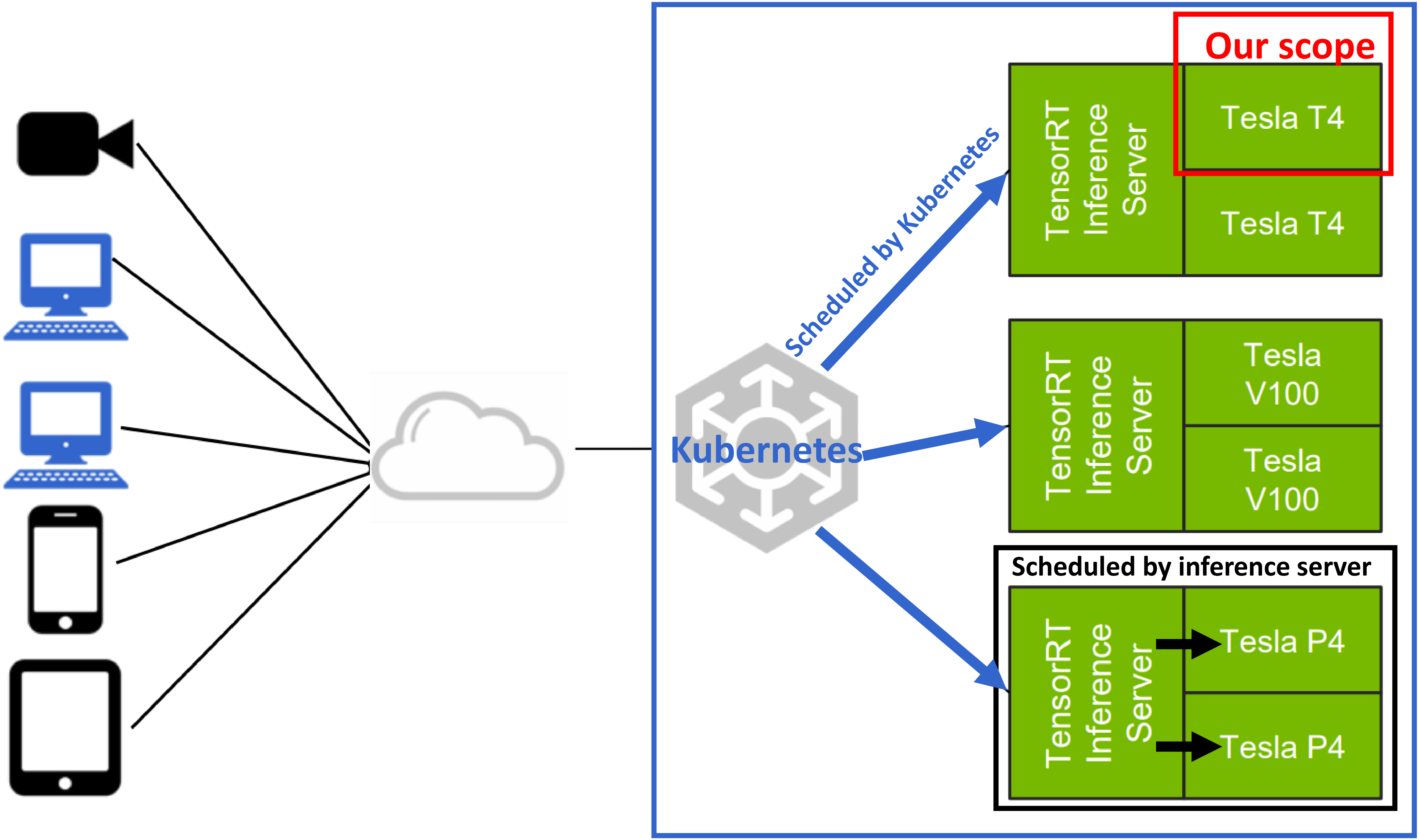
Our scope
Scheduled by Kubernetes
Kubernetes
TensorRT Inference Server
Tesla T4
Tesla T4
TensorRT Inference Server
Tesla V100
Tesla V100
Scheduled by inference server
TensorRT Inference Server
Tesla P4
Tesla P4

* Tripti Singhal, "Maximizing GPU Utilization for Data Center Inference with NVIDIA TensorRT Inference Server", NVIDIA Webinar, 2019