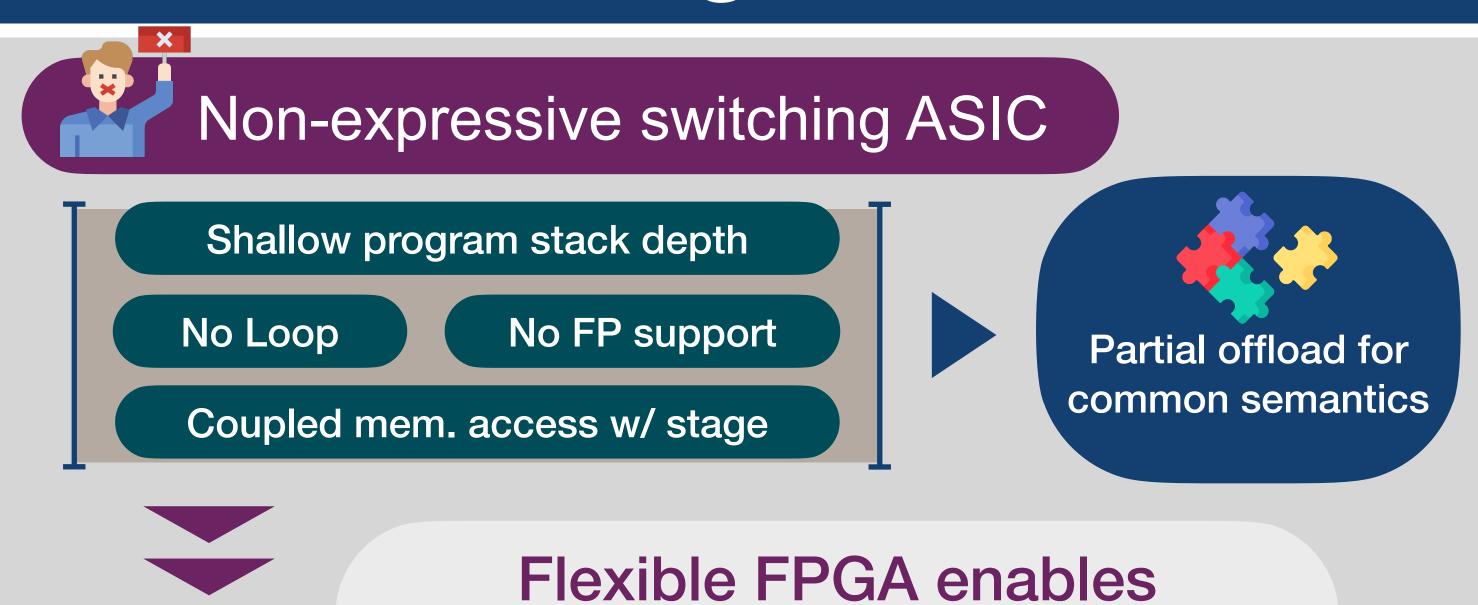
# Bridging Software and Hardware: Stateless Remote Accelerator Calls



Ziyi Yang, Krishnan B. Iyer, Yixi Chen, Ran Shu^, Zsolt István\*, Marco Canini, Suhaib A. Fahmy

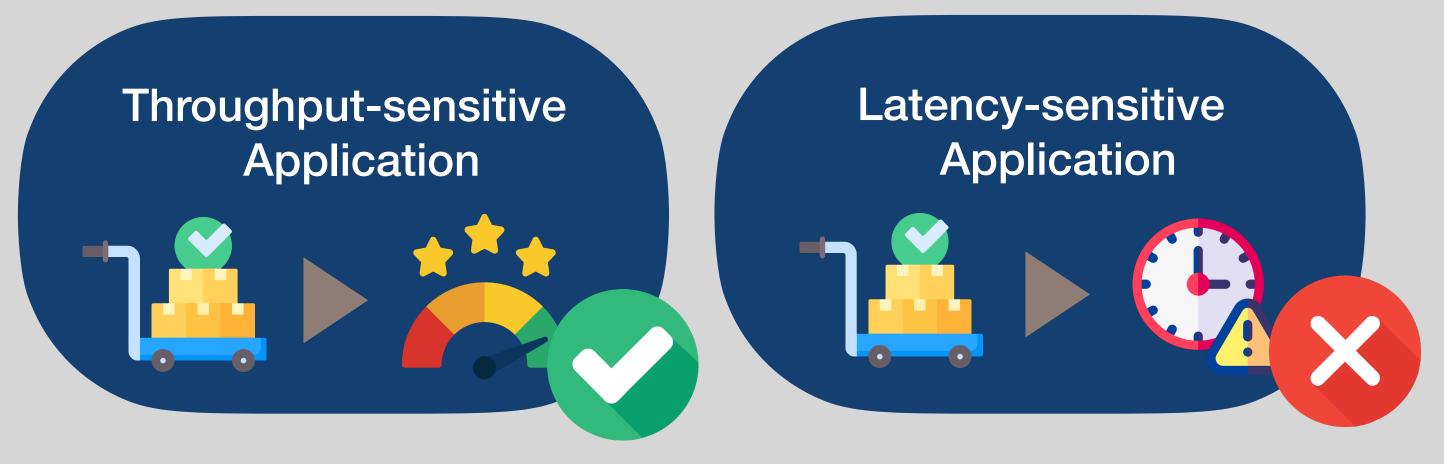
KAUST, 'Microsoft Research, \*TU Darmstadt

### Background



complex varied accelerators

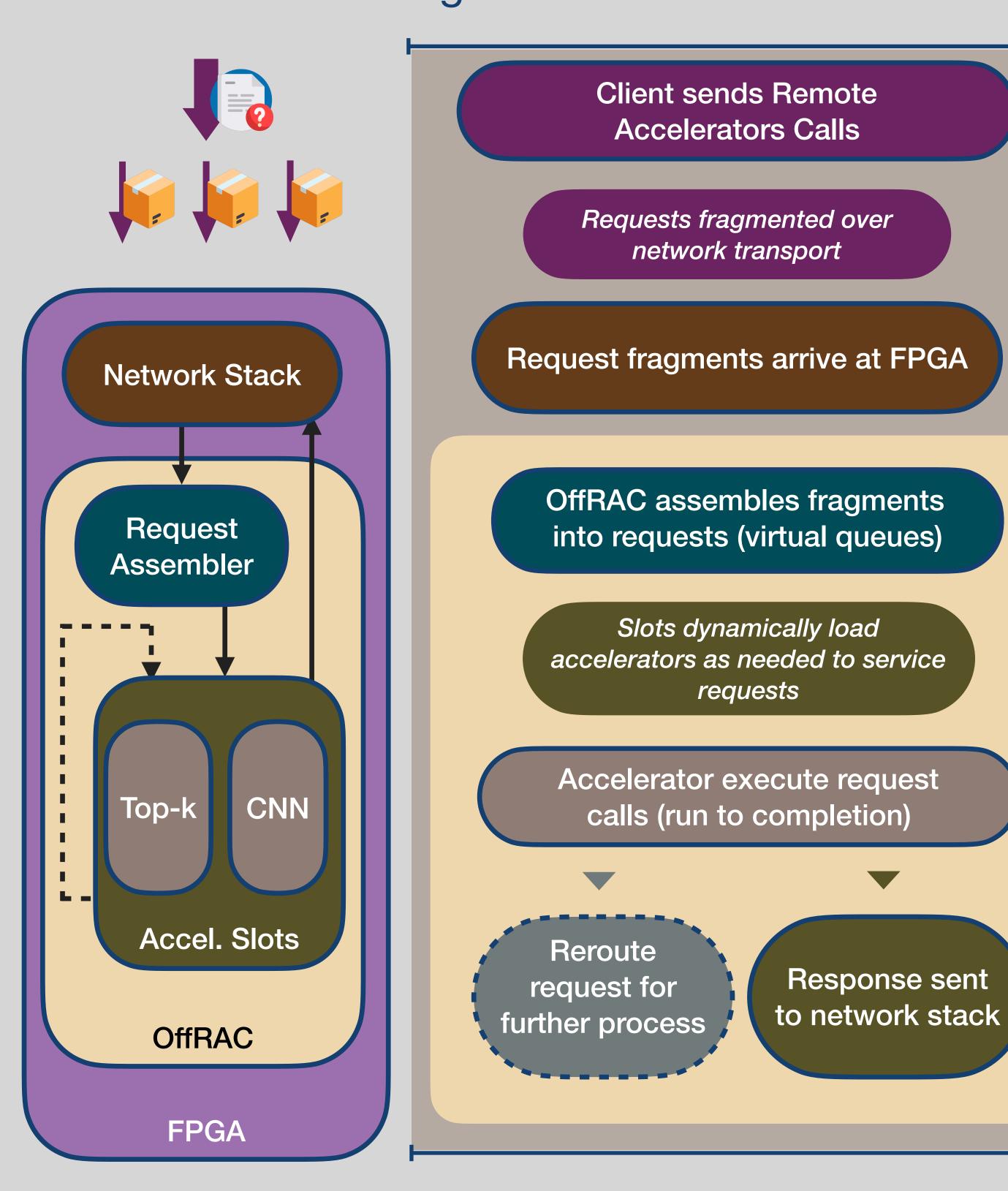




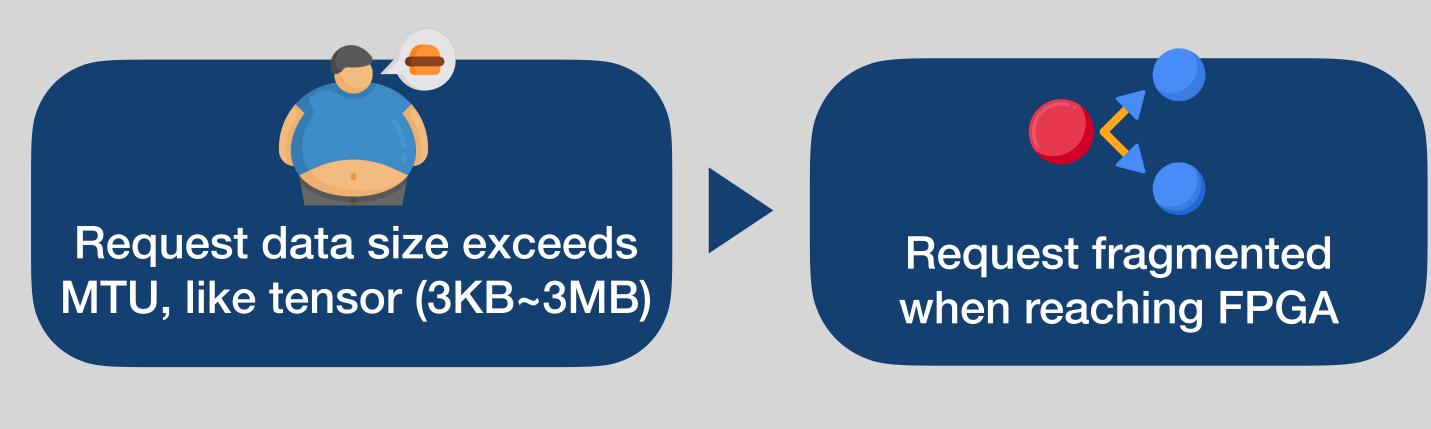
Batching hides data movement costs only works for throughput-sensitive applications!

# System Design

OffRAC hosts multiple accelerator slots on FPGA, accessible using remote accelerator calls

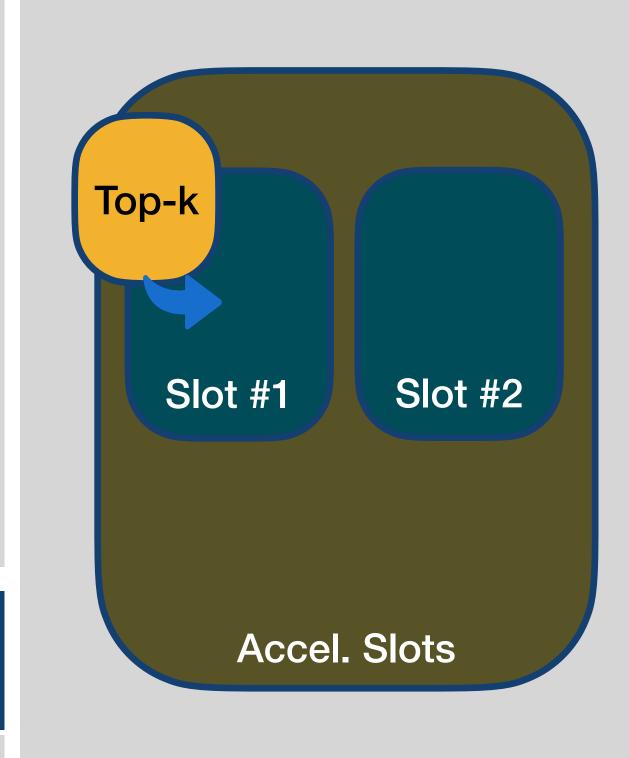


# Request Reassembly



Request reassembly crucial for virtualizing accelerator service across multiple clients for our run-to-completion model

#### Accelerator Slot Abstraction







## Preliminary Results





CPU-based DPDK (LibTPA) consumes 104-110W, while OffRAC on FPGA only requires 28-31W







