Designing Hierarchical Edge Cloud for Mobile Computing

Wei Gao

The University of Tennessee, Knoxville

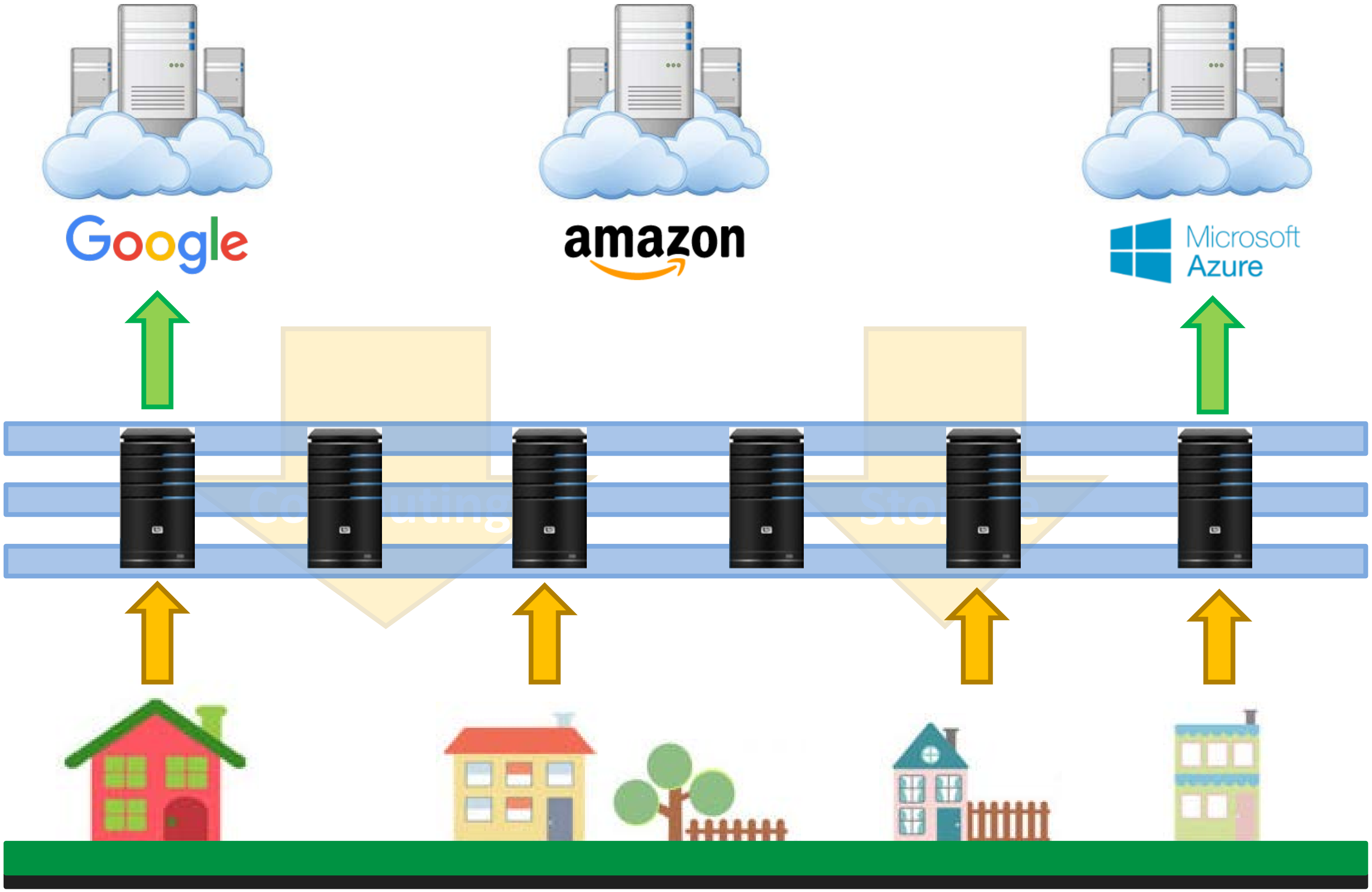
Supported by NSF Award No. CNS-1527612

Edge Cloud

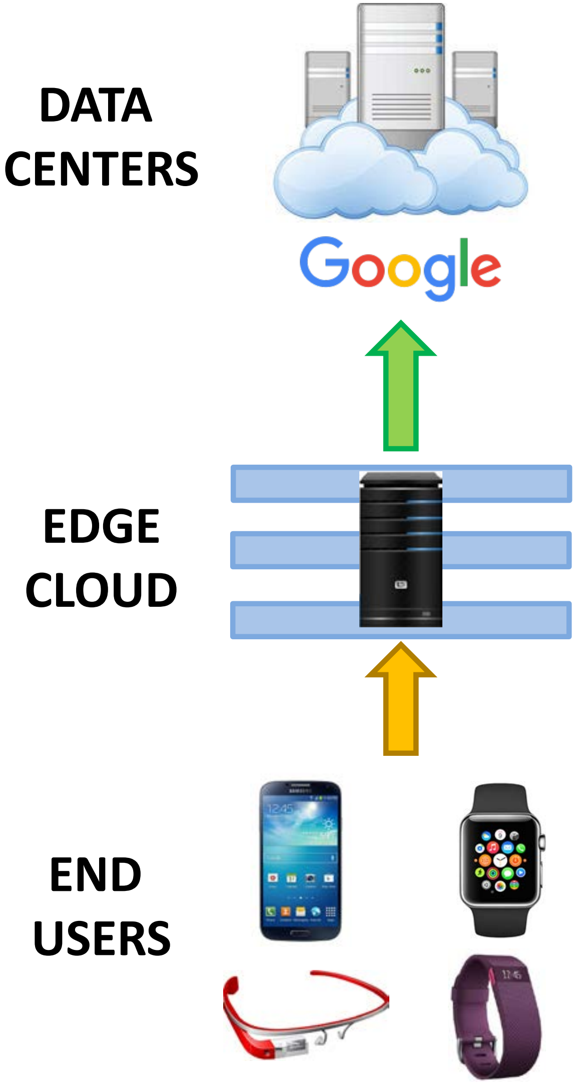
DATA CENTERS

EDGE CLOUD

END USERS



Edge Cloud for Mobile Computing



- Reduced response latency
  - Delay-sensitive mobile applications

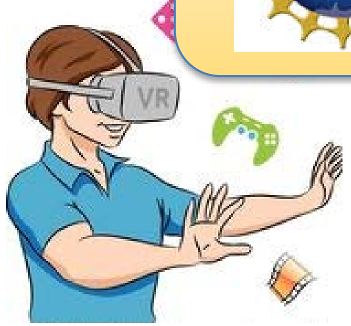
- Higher efficiency of resource utilization
  - Distributed processing

Applications of Edge Cloud

Cloud

Things

NSF CSR
Hightlighted Area



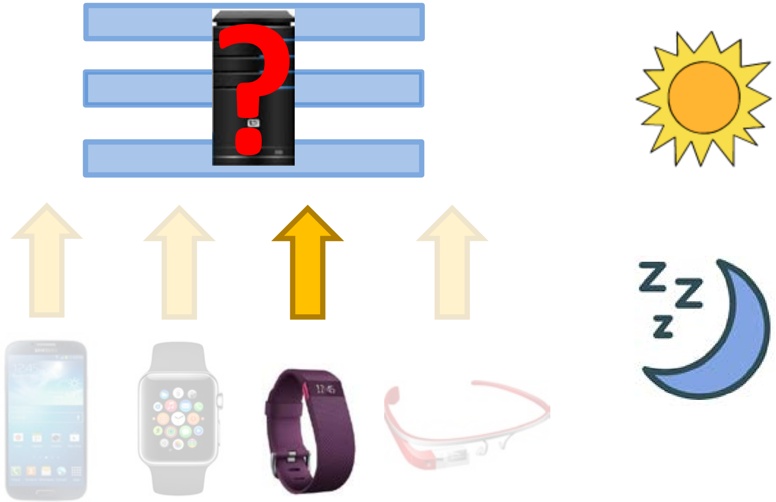
Virtual Reality

Smart Cities and Communities

Challenges

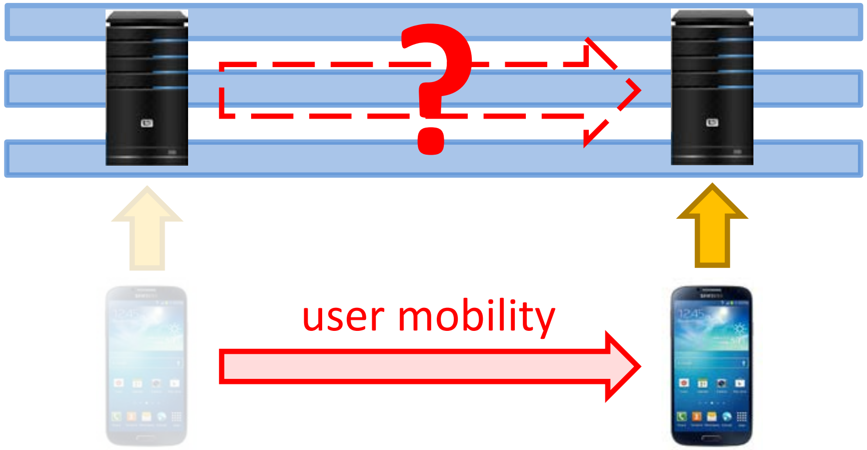
- Adaptability
  - Optimized performance?
  - Minimized cost?

~~Provis ng for the p load~~



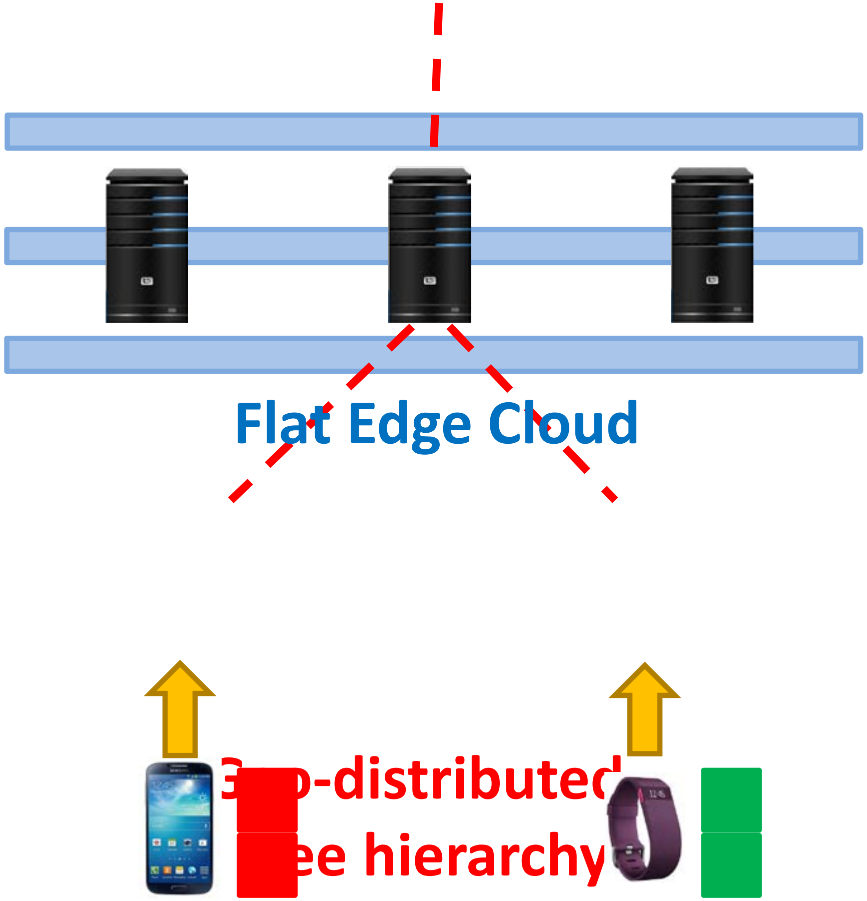
- User mobility
  - Minimized cost?

~~Comple nove of data ar gram~~



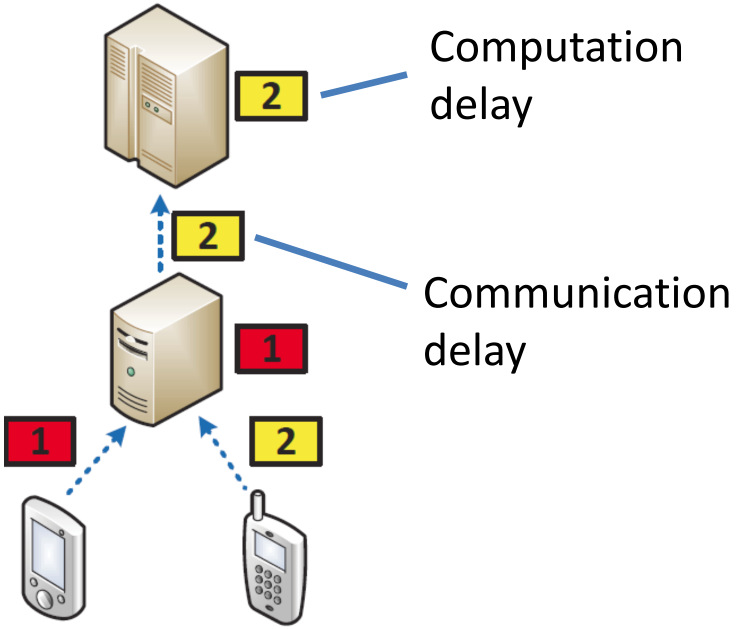
Our Solution: Hierarchical Edge Cloud

- Adaptability
  - Aggregation of peak load
- User mobility
  - Partial migration of data and program



Task 1: Optimal Workload Placement

- Our focus: minimized response latency
  - Where to place a workload
  - How much capacity for a workload
- Challenge
  - Delay tradeoff



Response latency = Computation + Communication

High tiers 🟥 🟩

Low tiers

A Hierarchical Edge Cloud Architecture for Mobile Computing, *in IEEE INFOCOM'16*.

Task 1: Optimal Workload Placement

- Distributed optimization

$$\min f = \sum_{i=1}^{m}\left(\overbrace{\frac{w_i}{\lambda_{i,\gamma_i}c_{\gamma_i}}}^{\text{Computation delay}} + \overbrace{\left(L(\boxed{\gamma_i}) - 1\right)}^{\text{Communication delay}}\frac{s_i}{B_{\gamma_i}}\right),$$

$$\text{s.t.} \sum_{j \in O_j} \boxed{\lambda_{i,j}} = 1, j = 1,2, \ldots, n$$

Capacity allocation of server *j* to workload *i*

Placement of workload *i*

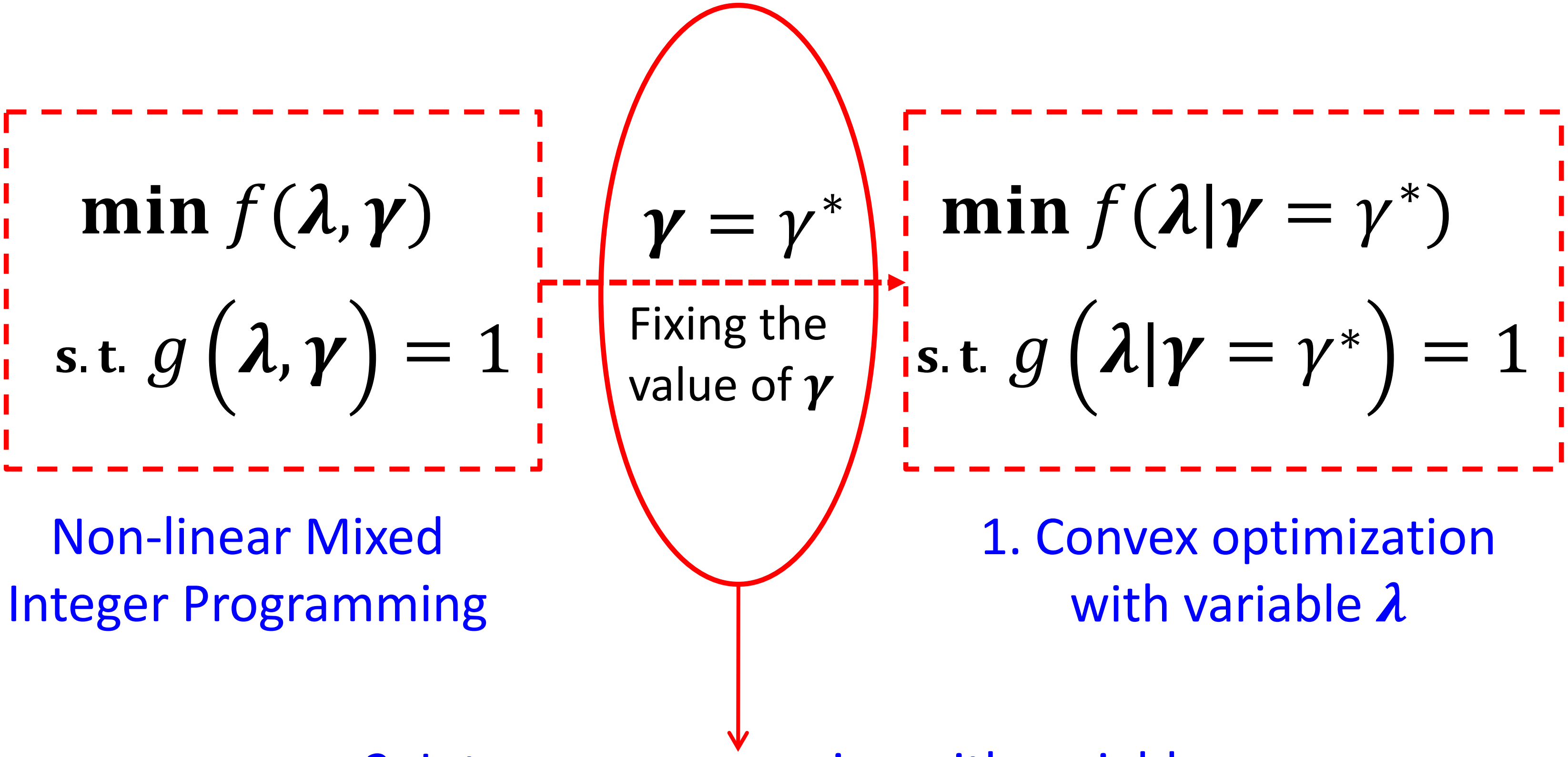
Non-linear mixed integer programming

?

A Hierarchical Edge Cloud Architecture for Mobile Computing, *in IEEE INFOCOM'16.*

Task 1: Optimal Workload Placement

- Problem transformation



$$\min f(\boldsymbol{\lambda},\boldsymbol{\gamma})$$

$$\text{s.t. } g\left(\boldsymbol{\lambda},\boldsymbol{\gamma}\right)=1$$

Non-linear Mixed Integer Programming

$$\boldsymbol{\gamma}=\gamma^*$$

Fixing the value of $\boldsymbol{\gamma}$

$$\min f(\boldsymbol{\lambda}|\boldsymbol{\gamma}=\gamma^*)$$

$$\text{s.t. } g\left(\boldsymbol{\lambda}|\boldsymbol{\gamma}=\gamma^*\right)=1$$

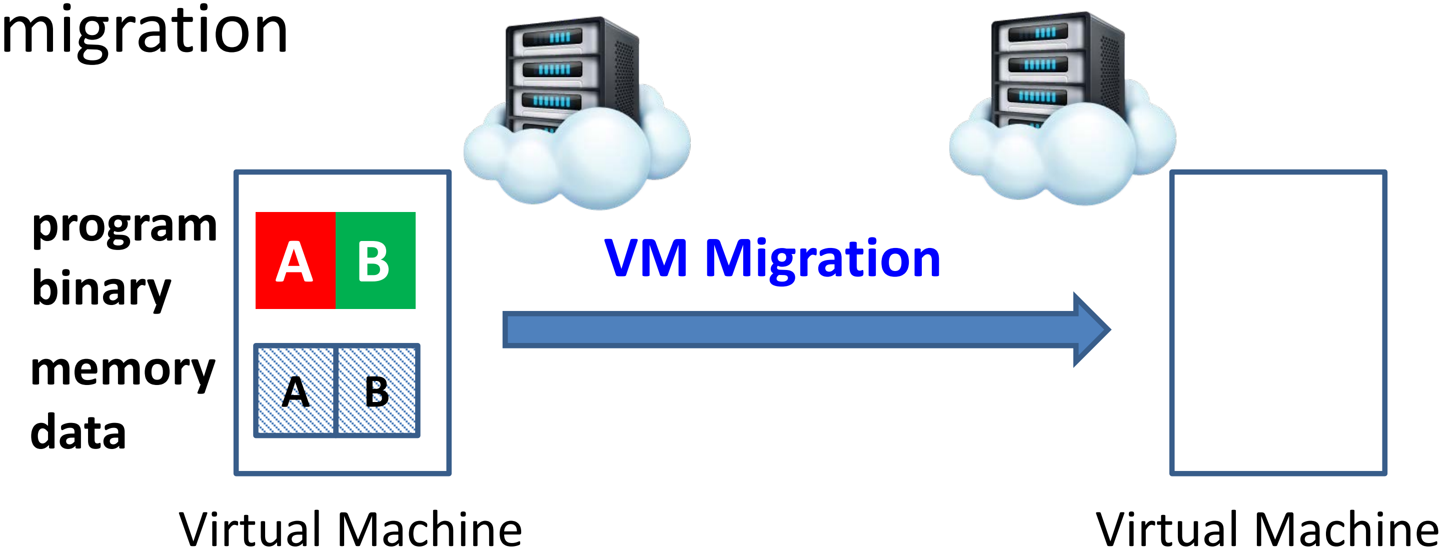
1. Convex optimization with variable $\boldsymbol{\lambda}$

2. Integer programming with variable $\boldsymbol{\gamma}$

A Hierarchical Edge Cloud Architecture for Mobile Computing, *in IEEE INFOCOM'16*.

Task 2: Supporting User Mobility

- Remote program execution with least context migration

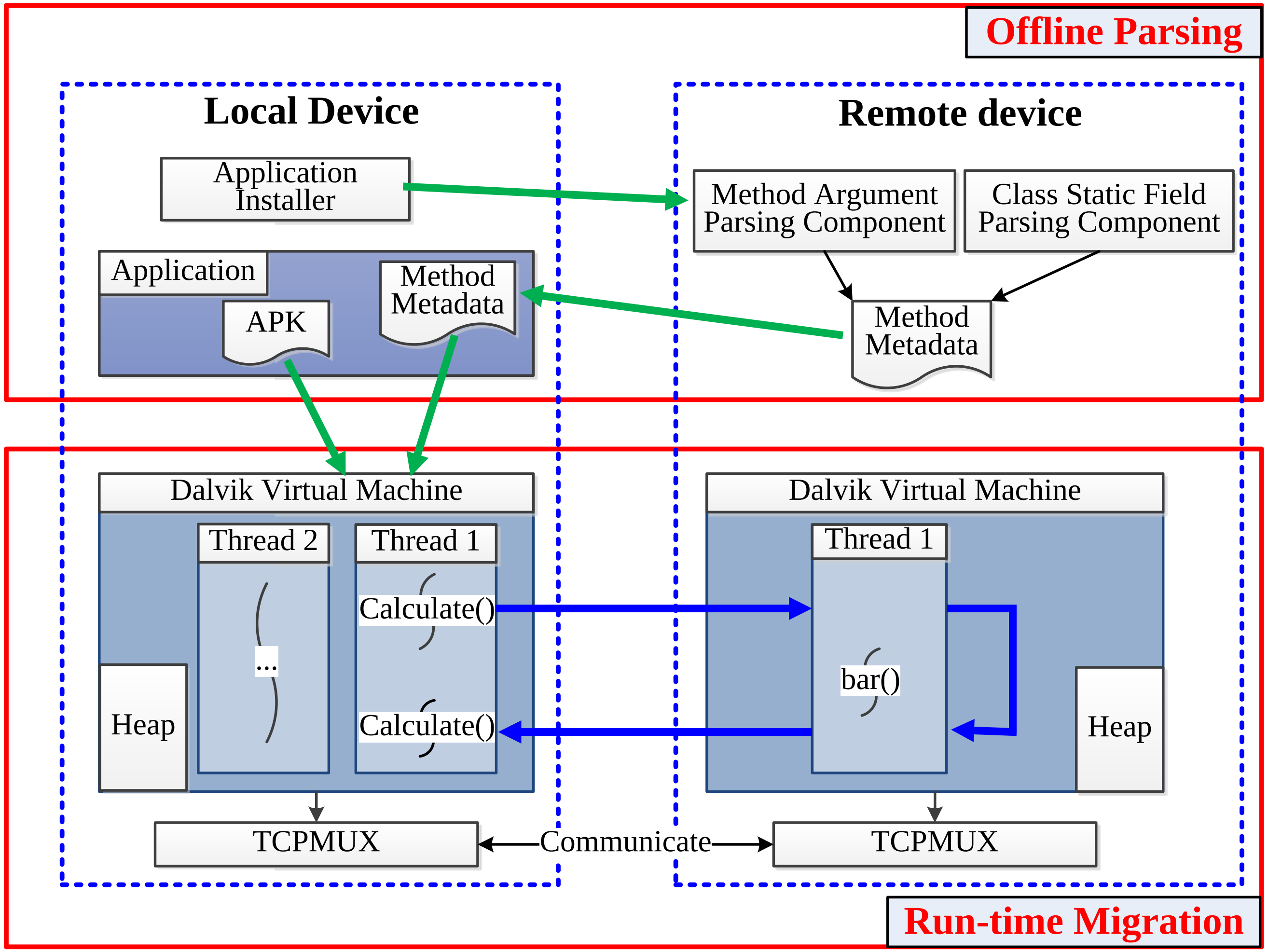


Offline program parsing

Run-time migration

Minimizing Context Migration in Mobile Code Offload, *in IEEE Transactions on Mobile Computing, 2017.*

Task 2: Supporting User Mobility



Minimizing Context Migration in Mobile Code Offload, *in IEEE Transactions on Mobile Computing, 2017.*

Implementation

- Heterogeneous mobile and wearable platforms

Samsung Galaxy S4

LG Watch Urbane

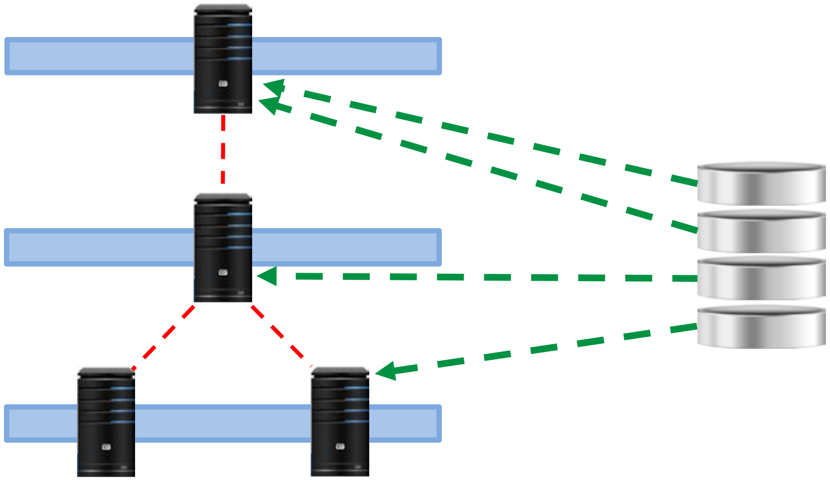
Samsung Nexus 10 Tablet

  - 1,500 LoC in Java and 1,000 LoC in C++ over Android v5.1.1 OS kernel

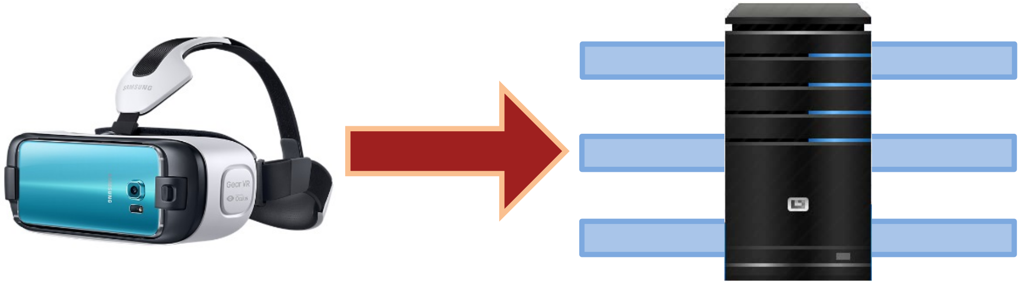
- Edge cloud servers
  - x86-based instances of Dalvik VM
  - Executing ARM-based native programs

Ongoing Work and Future Directions

- Optimal resource provisioning
  - Minimizing both CapEx and OpEx



- Virtual reality over mobile platforms



Thank you

- Questions?