Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand

Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, Onur Mutlu

SAFARI  Carnegie Mellon  Google  SAMSUNG  ETH Zürich
Consumer Devices
Consumer Devices
Consumer Devices

Consumer devices are everywhere!
Consumer Devices

Consumer devices are everywhere!
Consumer Devices

Consumer devices are everywhere!

Energy consumption is a first-class concern in consumer devices.
Popular Google Consumer Workloads
Popular Google Consumer Workloads

Chrome
Google’s web browser
Popular Google Consumer Workloads

**Chrome**
Google’s web browser

**TensorFlow Mobile**
Google’s machine learning framework
Popular Google Consumer Workloads

Chrome
Google’s web browser

TensorFlow Mobile
Google’s machine learning framework

VP9
Video Playback

VP9
Video Capture
Popular Google Consumer Workloads

Chrome
Google’s web browser

TensorFlow Mobile
Google’s machine learning framework

VP9
Video Playback
Google’s video codec

VP9
Video Capture
Google’s video codec
Energy Cost of Data Movement
Energy Cost of Data Movement

SoC

CPU

L1

L2

DRAM

Data Movement
Energy Cost of Data Movement

1st key observation: 62.7% of the total system energy is spent on data movement
Energy Cost of Data Movement

1st key observation: 62.7% of the total system energy is spent on data movement

Potential solution: move computation close to data
Energy Cost of Data Movement

1st key observation: 62.7% of the total system energy is spent on data movement.

Potential solution: move computation close to data
Energy Cost of Data Movement

1st key observation: 62.7% of the total system energy is spent on data movement

Potential solution: move computation close to data

Challenge: limited area and energy budget
Using PIM to Reduce Data Movement
Using PIM to Reduce Data Movement

2nd key observation: a significant fraction of the data movement often comes from simple functions.
Using PIM to Reduce Data Movement

2nd key observation: a significant fraction of the data movement often comes from simple functions.

We can design lightweight logic to implement these simple functions in memory.
2nd key observation: a significant fraction of the data movement often comes from simple functions.

We can design lightweight logic to implement these simple functions in memory.

Small embedded low-power core

PIM Core
Using PIM to Reduce Data Movement

2nd key observation: a significant fraction of the data movement often comes from simple functions

We can design lightweight logic to implement these simple functions in memory

- Small embedded low-power core
  - PIM Core
- Small fixed-function accelerators
  - PIM Accelerator
Using PIM to Reduce Data Movement

2nd key observation: a significant fraction of the data movement often comes from simple functions

We can design lightweight logic to implement these simple functions in memory

Small embedded low-power core
Small fixed-function accelerators

Offloading to PIM logic reduces energy and improves performance, on average, by 55.4% and 54.2%
Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand
Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, Onur Mutlu

Location: Virgina EF
Time: 11:10 AM