# Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

SAFARI

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

Carnegie Mellon · Google · SAMSUNG · ETH zürich

## Google Consumer Workloads

We look at widely-used Google workloads to identify major sources of energy consumption:

**Chrome**

Google's web browser

**TensorFlow Mobile**

Google's machine learning framework

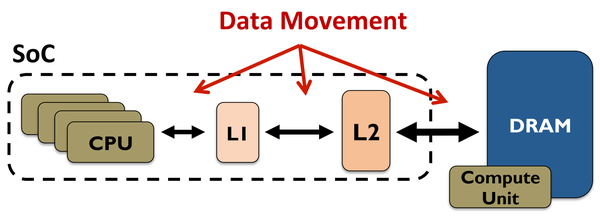
**VP9**

Video Playback and Video Capture

Google's video codec

## Data Movement Cost

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




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

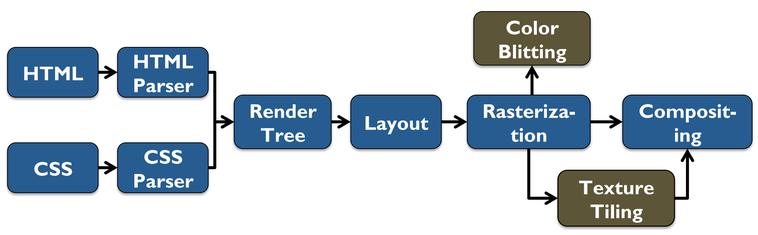
## Goals

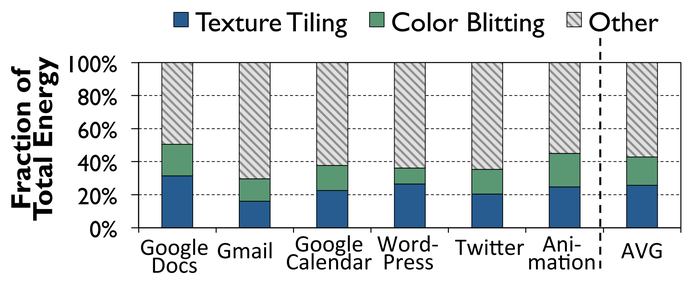
1. Understand the data movement related bottlenecks in modern consumer workloads
2. Analyze the benefits of processing-in-memory (PIM) to mitigate data movement cost
3. Investigate the PIM logic that can maximize energy efficiency given the limited area and energy budget in consumer devices

## Browser

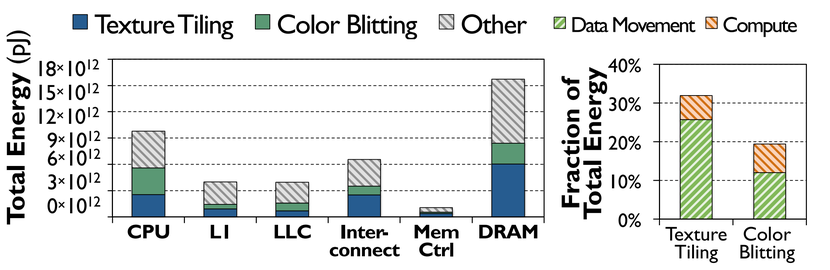
### Scrolling

#### Chrome Rendering Pipeline



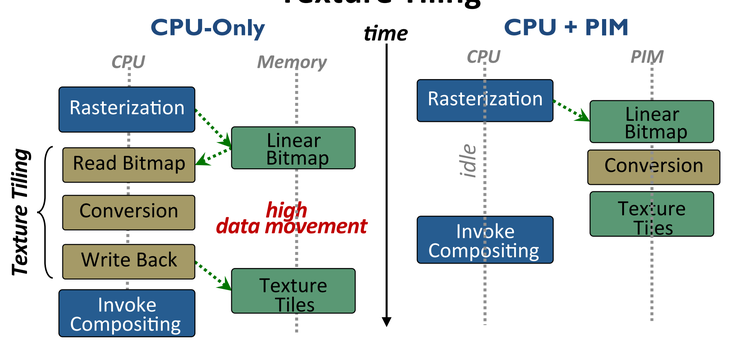

41.9% of page scrolling energy is spent on texture tiling and color blitting




77% of total energy consumption goes to data movement

49.1% of total data movement comes from texture tiling and color blitting

#### Texture Tiling




Texture tiling is a good candidate for PIM execution


Requires simple primitives: memcopy, bitwise operations, and simple arithmetic operations

**PIM Core**

9.4% of the area available for PIM logic

**PIM Accelerator**

7.1% of the area available for PIM logic

PIM core and PIM accelerator are feasible to implement in-memory texture tiling

#### Color Blitting

Generates a large amount of data movement

Accounts for 19.1% of the total system energy during scrolling

Color blitting is a good candidate for PIM execution

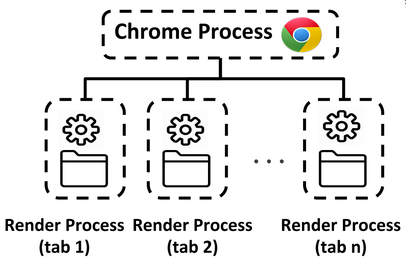
Requires low-cost operations:

Memset, simple arithmetic, and shift operations

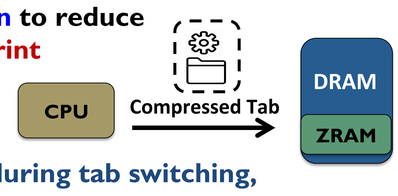
Color blitting is feasible to implement in PIM logic

### Tab Switching

- Chrome employs a multi-process architecture
- Main operations during tab switching:
  - Context switch
  - Load the new page

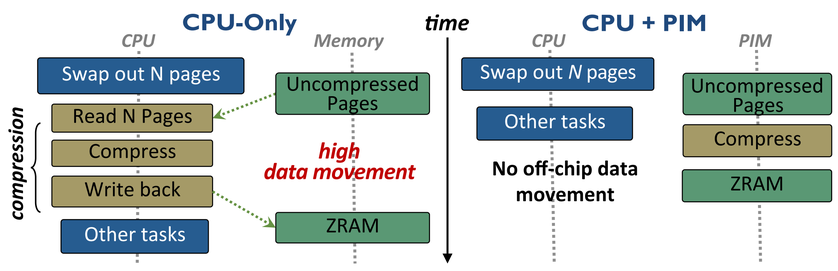



- Chrome uses compression to reduce each tab's memory footprint



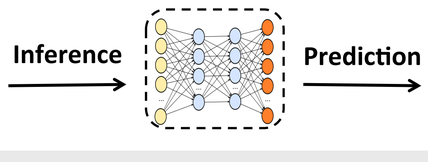

- To study data movement during tab switching, we perform an experiment:
  - A user opens 50 tabs (most-accessed websites)
  - Scrolls through each for a few second
  - Switches to the next tab

1. Compression and decompression contribute to 18.1% of the total system energy
2. 19.6 GB of data move between CPU and ZRAM




Both functions can benefit from PIM execution and can be implemented at PIM logic

## TensorFlow Mobile




57.3% of the inference energy is spent on data movement

54.4% of the data movement energy comes from packing/unpacking and quantization


Reorders elements of matrices to minimize cache misses

Up to 40% of the inference energy and 31% of inference execution time

Packing's data movement accounts for up to 35.3% of the inference energy


Converts 32-bit floating point to 8-bit integers

Up to 16.8% of the inference energy and 16.1% of inference execution time

Majority of quantization energy comes from data movement

A simple data reorganization process that requires simple arithmetic

A simple data conversion operation that requires shift, addition, and multiplication

## Video Playback and Capture


63.5% of the system energy is spent on data movement

80.4% of the data movement energy comes from sub-pixel interpolation and deblocking filter

Sub-pixel interpolation: interpolates the value of pixels at non-integer location

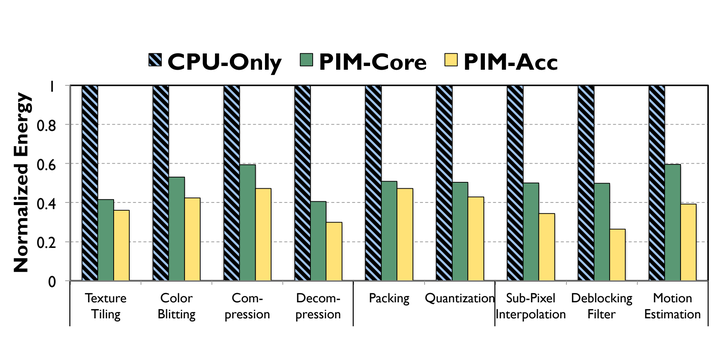
Deblocking filter: a simple low-pass filter that attempts to remove discontinuity in pixels


59.1% of the system energy is spent on data movement

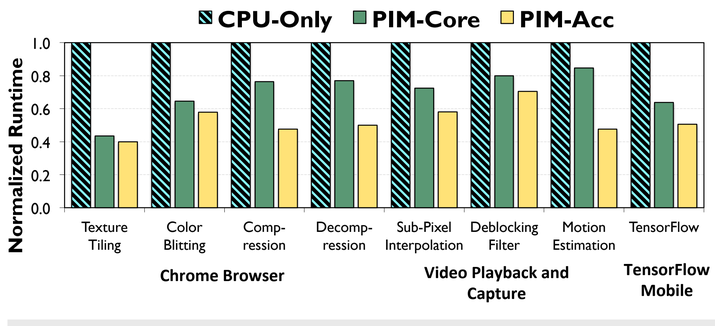
Majority of the data movement energy comes from motion estimation

Motion estimation: compresses the frames using temporal redundancy between them

## Evaluation




On average, energy consumption reduces by 49.1% using PIM core and 55.4% using PIM accelerator




On average, energy consumption reduces by 44.6% using PIM core and 54.2% using PIM accelerator