A Case for Core-Assisted Bottleneck Acceleration in GPUs

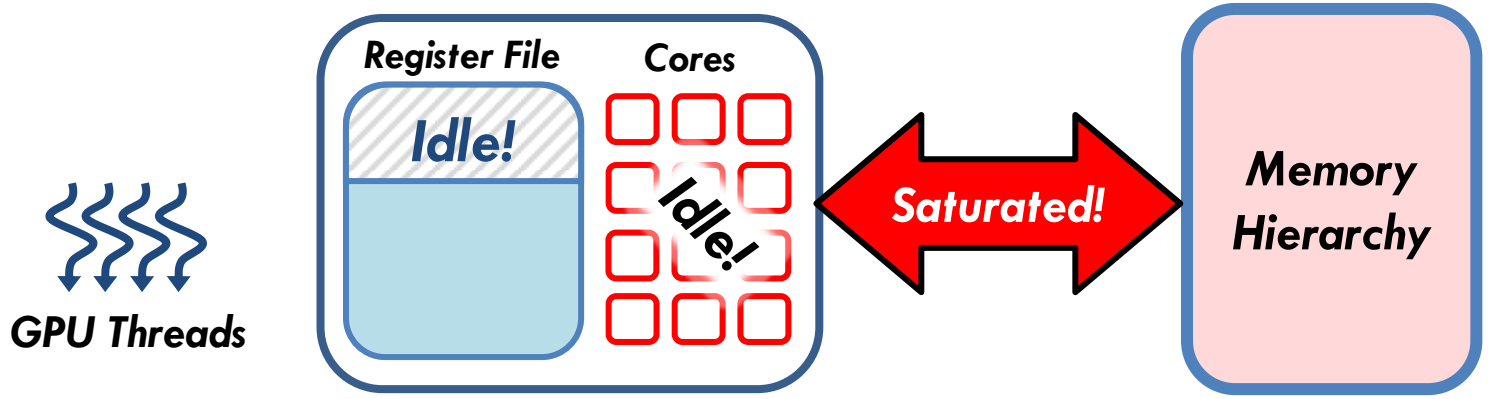
Enabling Flexible Data Compression with Assist Warps

Nandita Vijaykumar

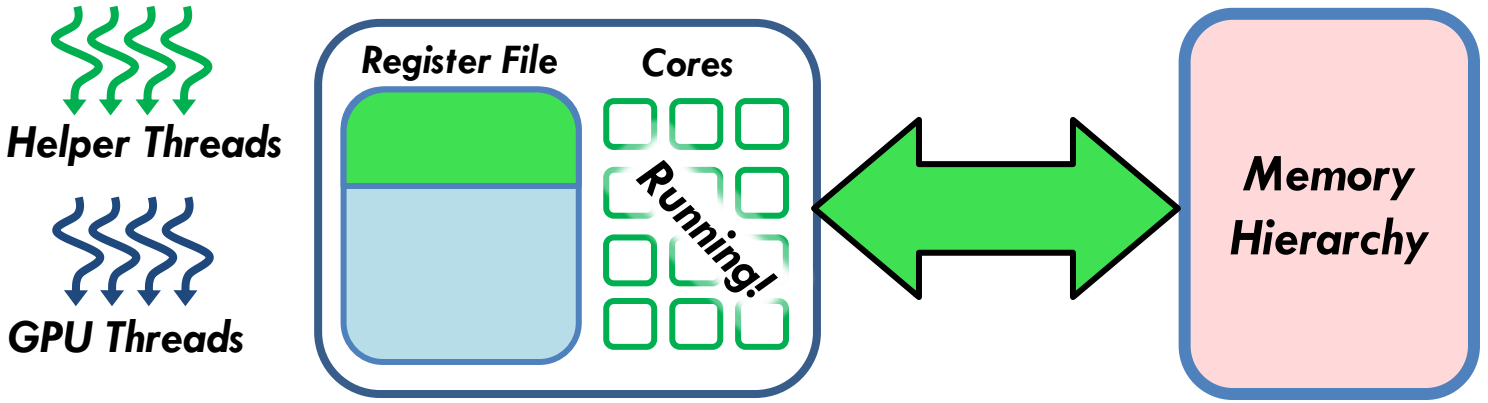
Gennady Pekhimenko, Adwait Jog, Abhishek Bhowmick, Rachata Ausavarangnirun, Chita Das, Mahmut Kandemir, Todd C. Mowry, Onur Mutlu

SAFARI Carnegie Mellon PENNSTATE 1855

Observation: Imbalances in execution leave GPU resources underutilized



Our Goal: Employ idle resources to do something useful: **accelerate the bottleneck – using helper threads**



Challenge: How do you **manage and use** helper threads in a **throughput-oriented** architecture?

Our Solution: CABA

- **A new framework** to enable helper threading in GPUs
  - **CABA (Core-Assisted Bottleneck Acceleration)**
- **Wide set of use cases**
  - Compression, prefetching, memoization, …
- **Flexible data compression** using CABA alleviates the memory bandwidth bottleneck
  - 41.7% performance improvement