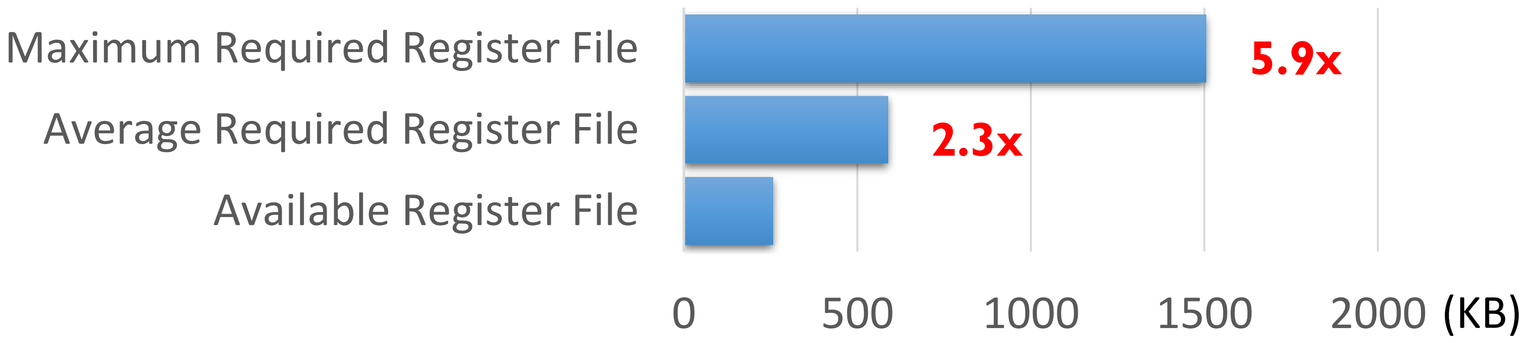
LTRF: Enabling High-Capacity Register Files for GPUs via Hardware/Software Cooperative Register Prefetching

Mohammad Sadrosadati
Amirhossein Mirhosseini
Seyed Borna Ehsani
Hamid Sarbazi-Azad
Mario Drumond
Babak Falsafi
Rachata Ausavarungnirun
Onur Mutlu

ETH Zürich
MICHIGAN
Sharif University of Technology
EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Carnegie
Mellon
University

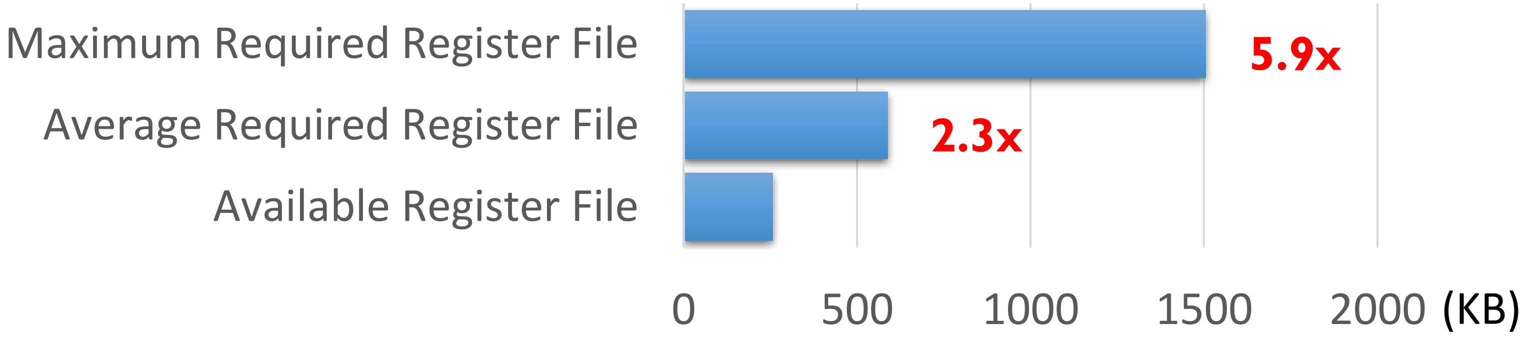
Register file size limits GPU scalability

- Register file already accounts for 60% of on-chip storage
- But, there is still demand for more registers to achieve maximum performance and concurrency



Register file size limits GPU scalability

- Register file already accounts for 60% of on-chip storage
- But, there is still demand for more registers to achieve maximum performance and concurrency

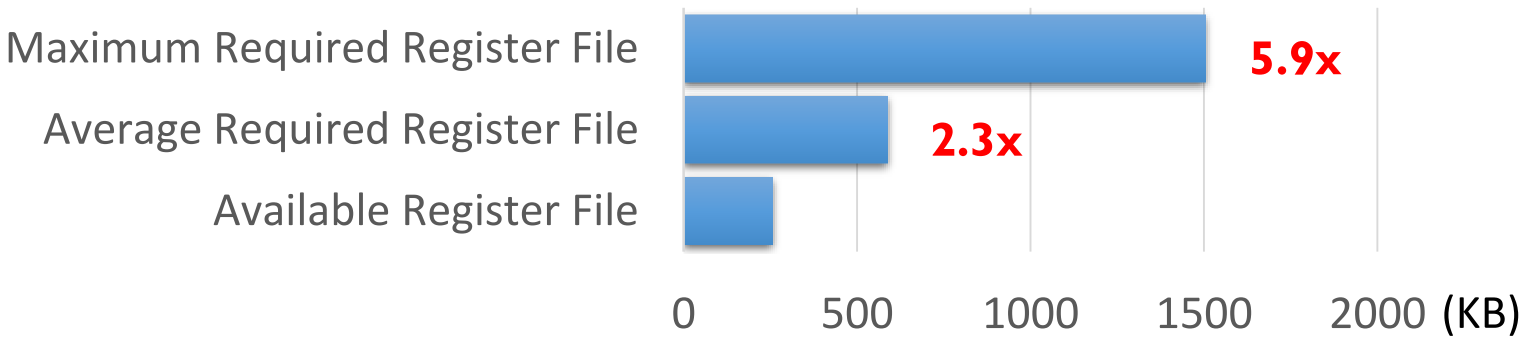


•Unfortunately, all mechanisms to expand RF capacity without large area/power overheads significantly increase access latency

- Emerging technologies, register file compression/virtualization, etc.

Register file size limits GPU scalability

- Register file already accounts for 60% of on-chip storage
- But, there is still demand for more registers to achieve maximum performance and concurrency



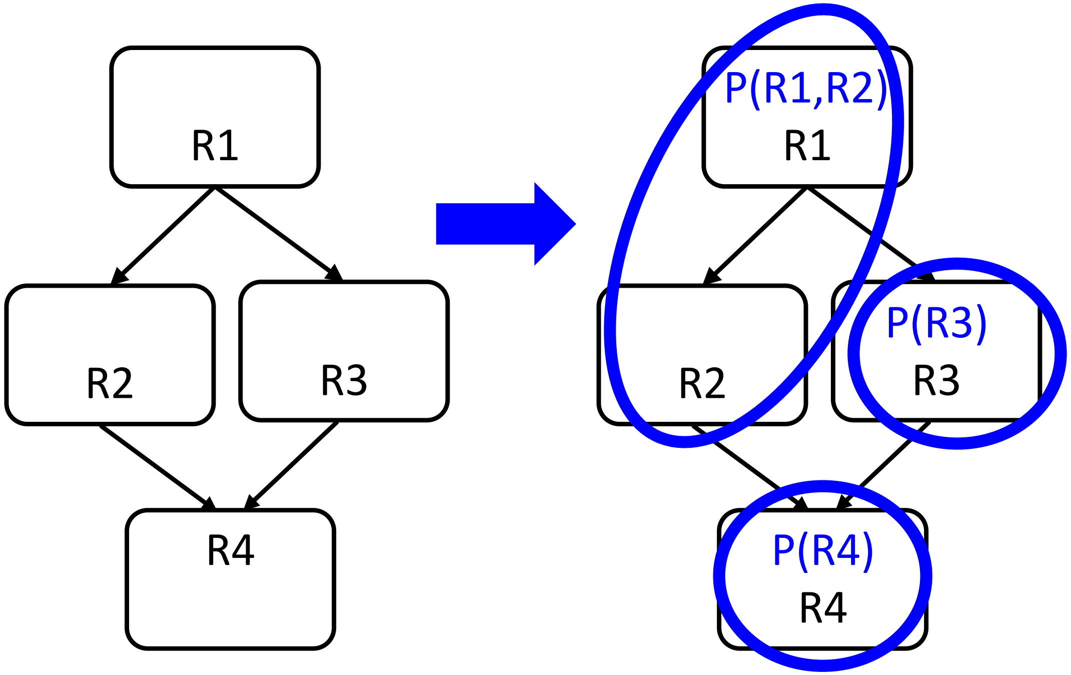
•Unfortunately, all mechanisms to expand RF capacity without large area/power overheads significantly increase access latency

- Emerging technologies, register file compression/virtualization, etc.

Goal: Tolerate register file latencies

Contributions (1)

- **Compiler-driven Register Prefetching**
  - Break control flow graph into "prefetch subgraphs"
  - Prefetch registers at the beginning of each subgraph
  - Interval analysis to identify prefetch subgraphs

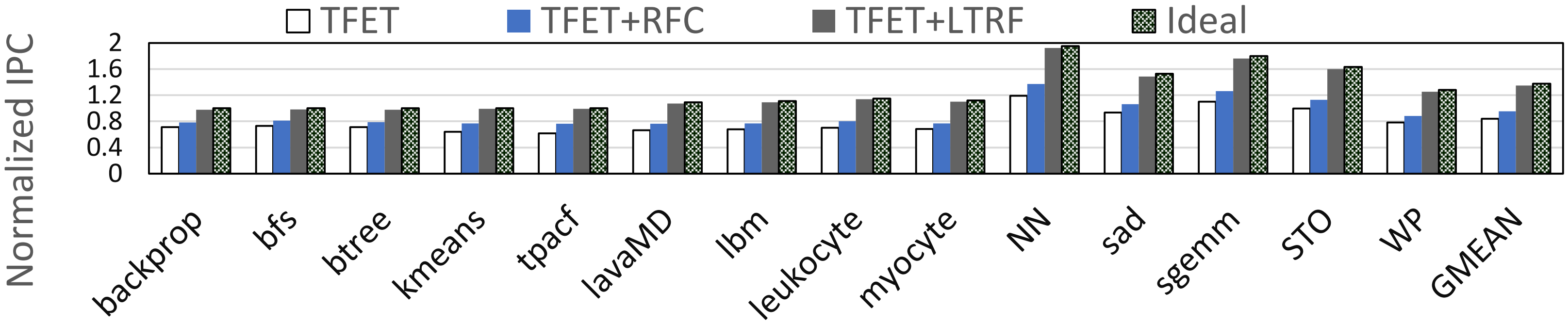


Contributions (2)

- **Latency Tolerant Register File (LTRF)**
  - “2-level” main register file + register cache
  - Performs prefetch ops while executing other warps
  - Tolerates up to 6x higher RF access latencies
  - Paves the way for various area/power optimizations

Contributions (2)

- **Latency Tolerant Register File (LTRF)**
  - “2-level” main register file + register cache
  - Performs prefetch ops while executing other warps
  - Tolerates up to 6x higher RF access latencies
  - Paves the way for various area/power optimizations



Example LTRF deployment enables 8× larger register file and 32% higher performance

LTRF: Enabling High-Capacity Register Files for GPUs via Hardware/Software Cooperative Register Prefetching

Mohammad Sadrosadati
Amirhossein Mirhosseini
Seyed Borna Ehsani
Hamid Sarbazi-Azad
Mario Drumond
Babak Falsafi
Rachata Ausavarungnirun
Onur Mutlu

ETH Zürich
MICHIGAN
Sharif University of Technology
EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Carnegie Mellon University

LTRF: Enabling High-Capacity Register Files for GPUs via Hardware/Software Cooperative Register Prefetching

Mohammad Sadrosadati
Amirhossein Mirhosseini
Seyed Borna Ehsani
Hamid Sarbazi-Azad
Mario Drumond
Babak Falsafi
Rachata Ausavarungnirun
Onur Mutlu

ETH Zürich
MICHIGAN
Sharif University of Technology
EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Carnegie Mellon University

Please attend our talk at 2:00 in Virginia EF