USC University of Southern California

USC Viterbi School of Engineering

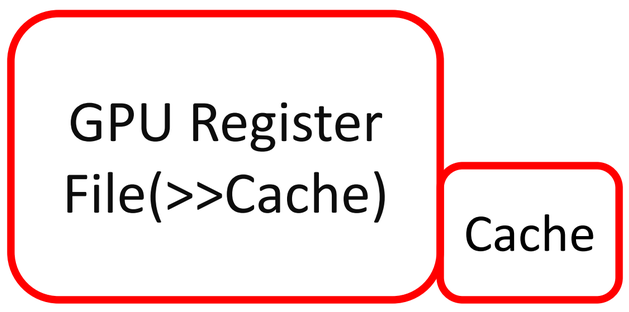
Ming Hsieh Department of Electrical Engineering

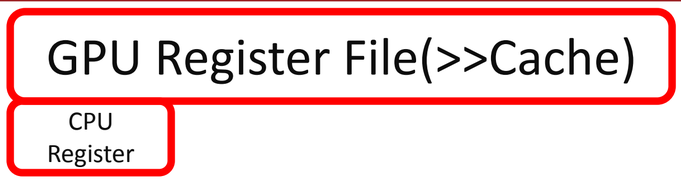
Warped Register File: A Power Efficient Register File for GPGPUs

Mohammad Abdel-Majeed and Murali Annavaram

Introduction

- **GPGPU Register Files**
  - Larger than GPU Cache
  - Extra-wide (128 bytes) : Needed to support 32 parallel threads

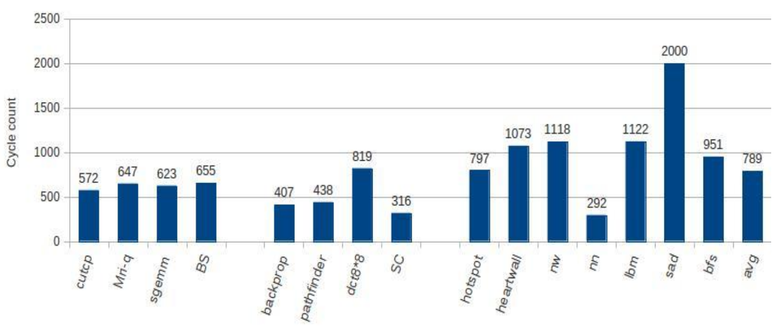




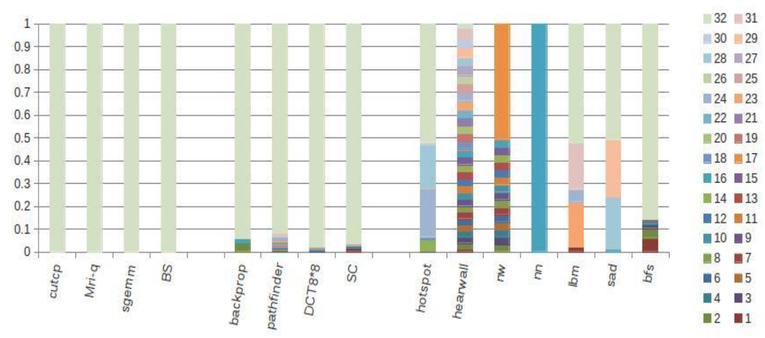
Register files burn massive leakage and dynamic power

Motivation

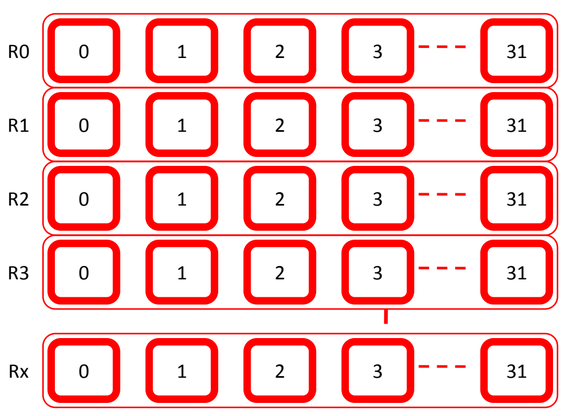
- **Register Access Inefficiency**
  - Unused registers = 46%!
  - Time between two accesses = 789 cycles!



- **Varied Thread Parallelism**
  - 32 threaded hardware, but only 24 or less are active (code divergence)
  - All 128 bytes are activated per access



Baseline : Leave all registers ON

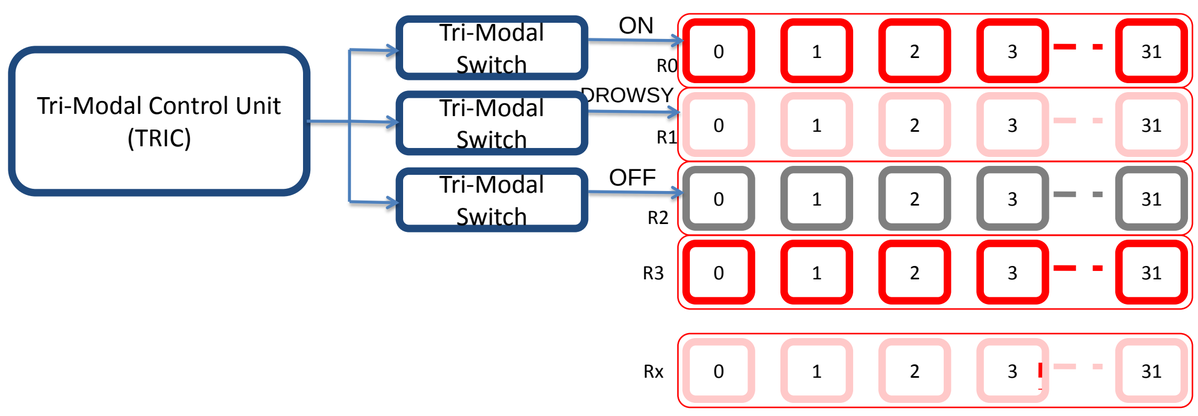


Baseline

Leakage Power

- **Save Leakage Power**
  - OFF unallocated registers
  - Drowsy Registers
  - Turn ON only during access
- **Architectural Support**
  - Use Tri-Modal Switch
  - Tri-Modal Control Unit(TRIC)
    - Activates a register only on access

Instruction: ADD R0,R0,R3
Active Mask: No role in saving power

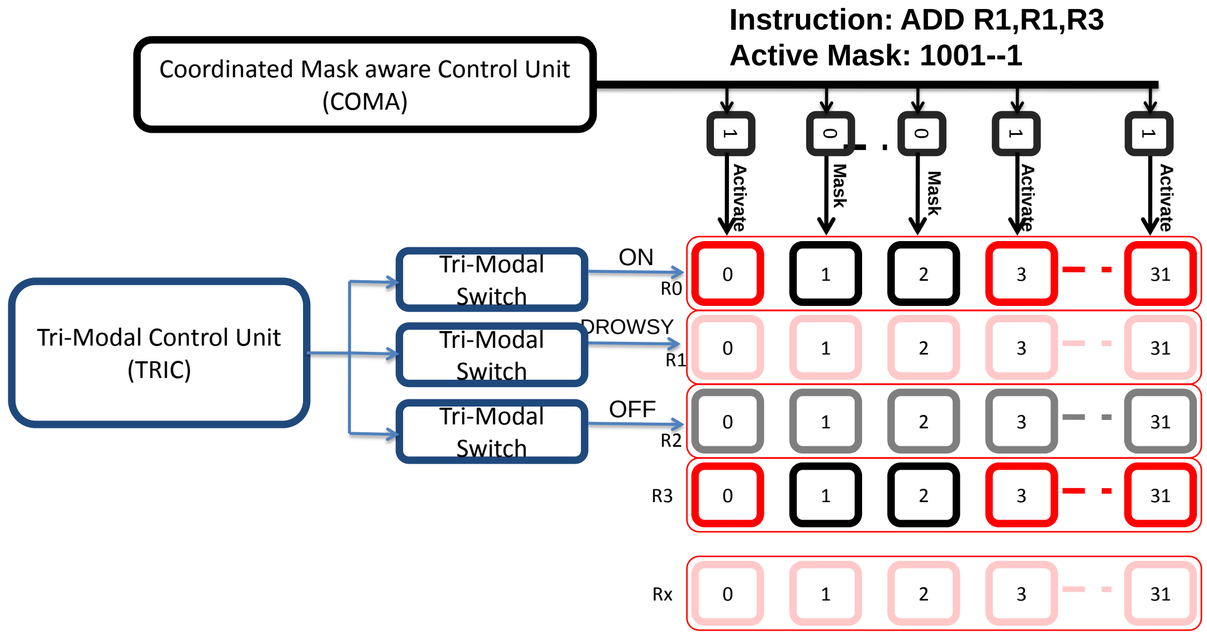


TRIC only

Dynamic Power

- **Save Dynamic Power**
  - Enable partial width access
- **Architectural Support**
  - Divide wordline into segments
  - Use the warp active mask to activate segments
  - Coordinated Mask Aware Control Unit (COMA)

Instruction: ADD R1,R1,R3
Active Mask: 1001--1



Combined

Results

15	14%	90%	69%	1%
Benchmarks GPGPU-sim simulator Cadence for circuits Simulations	Dynamic Power savings 0% - 59%	Leakage power reduction 85%-98%	Total Power savings 55%-90%	Performance overhead 3 cycles wakeup latency -.07%-3.16%

abdelmaj@usc.edu, annavara@usc.edu

NSF DARPA

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering