**TABS: Temperature Aware Thread Block Scheduling in GPGPUs**

Rajib Nath‡, Raid Ayoub‡, Tajana Šimunić Rosing‡

‡ Intel Lab, ‡ UCSD

---

**Thermal Problems in GPGPUs**

- Single GPGPU
  - There is no thermally aware scheduling in current GPGPUs
  - Recent GPGPUs allow multiple kernels to run concurrently
  - High GPGPU temperature
  - Computation slowdown due to throttling
  - High leakage power
  - High cooling costs
  - High energy consumption

- Multiple GPGPUs
  - Kernels are scheduled to dedicated GPGPUs
  - GPGPUs share a single fan
  - Computation slowdown
  - High leakage power and cooling costs
  - High energy consumption

- Setup
  - Modified hotspot simulator for GPGPUs
  - Developed thermal model for GTX280
  - Ambient temperature set to 45°C
  - Used merge benchmark suite

---

**Thermal Management via Thread Block Scheduling**

**Motivation and Novelty**

- Workload Characteristics
  - TBs have a short life time (μs to ms) and a stable power profile
  - Abundance of TBs
  - Thermal heterogeneity between kernels

- Our Contribution
  - Intermixing TBs from heterogeneous kernels to minimize thermal power
  - Intermix TBs without any thread migration
  - Extended the solution for multi-GPGPU
  - Provided a H/W+S/W approach

- **Heterogeneous Mixing**
  - Alternate
  - Mixed Alternate
  - Default
  - Mixed Uniform
  - Mixed Non-Uniform

- **TABS Architecture**
  - OS Scheduler is aware of power density index & lifetime
  - Scheduling Policy
  - New policy set by the OS scheduler
  - Used merge benchmark suite

- **Results**
  - Our Baseline policy clock gates GPGPU SMs until SM temperature falls below 85°C
  - TABS is implemented with four different intermixing policies: A, MU, MNU, MA
  - In the absence of heterogeneous kernels, TABS employs DVFS
  - Improvements indicates reduction in computation slowdown due to throttling

---

**Single GPGPU**

- **Improvements over Baseline**
  - TABS improves over Baseline by 57%-60% on average
  - TABS improves over DVFS by 40%-45% on average
  - Improvement is higher when heterogeneity is higher– up to 97%
  - Improvement is 82%-86% on average for the target cases (WL1-WL6)

- **Energy Savings over Baseline**
  - Average energy savings with TABS is 6.75%
  - DVFS saves 1.8% energy on average
  - Higher heterogeneity leads to higher energy savings (e.g., 15.75%)
  - Energy savings are 9.48% on average for the target cases (WL1-WL6)

---

**Multiple GPGPU**

- **Improvements over Baseline**
  - Baseline and DVFS have similar performance and energy consumption
  - TABS improves over Baseline by 44%-48% on average
  - Benefit exists when memory overhead is less than the gain through thermally aware scheduling

- **Effect of Memory Technology**
  - As the memory and the interconnect technology improves the gain through TABS gets larger

---

**Interactions**

- Intermixes kernels whenever possible
- Looks for a cold kernel during thermal emergencies
- In the absence of a thermal emergency, it looks for a hot kernel to spread the heat over time
- Intermixes kernels in a given intermixing window
- Works proactively

---

**Setup**

- GTX 280 Graphics Card
- Intel Xeon X5650 (2.67GHz)
- NVIDIA GTX 690
- Target workloads
  - Single GPGPU: Intel TBB benchmark suite
  - Multiple GPGPU: Mixed Uniform (MU) and Mixed Alternate (MA) policies