Idle Resources in GPUs

- A lot of stalls and idle time in the shaders
- Often there are insufficient warps to hide the memory/computational latencies

Observations

- A significant amount of time is spent waiting for data from memory
- Stalls are also due to long latency ALU operations and resource contention
- Compute units are idle during this time
  - Idle cores can be used to perform useful computation

Shader Assisted Data Decompression

Idle shaders can be used to perform computationally intensive decompression algorithms

- Mechanism
  - Convert decompression algorithm into multiple simple instructions
  - Retrieve compressed data into registers
  - Spontaneously generate a helper warp to perform decompression

Performance Impact

Accelerating Bottlenecks

- Idle pipelines in the cores can be used to accelerate bottlenecks
  - Memory limited
    - Data Compression
    - Redundant/Speculative Execution
    - Scheduling
  - Computationally limited
    - Memoization
    - Prefetching