Parallel Computation Patterns
Atomic Operation Performance
Objective

- To learn the main performance considerations of atomic operations
  - Latency and throughput of atomic operations
  - Atomic operations on global memory
  - Atomic operations on shared L2 cache
  - Atomic operation on shared memory
Atomic Operations on DRAM

- An atomic operation starts with a read, with a latency of a few hundred cycles.
- The atomic operation ends with a write, with a latency of a few hundred cycles.
- During this whole time, no one else can access the location.
Atomic Operations on DRAM

- Each Load-Modify-Store has two full memory access delays
- All atomic operations on the same variable (DRAM location) are serialized
Latency determines throughput

• Throughput of an atomic operation is the rate at which the application can execute an atomic operation.

• The rate for atomic operation on a particular location is limited by the total latency of the read-modify-write sequence, typically more than 1000 cycles for global memory (DRAM) locations.

• This means that if many threads attempt to do atomic operation on the same location (contention), the memory bandwidth is reduced to <
You may have a similar experience in supermarket checkout

• Some customers realize that they missed an item after they started to check out
• They run to the isle and get the item while the line waits
  • The rate of checkout is reduced due to the long latency of running to the isle and back.
• Imagine a store where every customer starts the check out before they even fetch any of the items
  • The rate of the checkout will be $1 / (\text{entire shopping time of each customer})$
Hardware Improvements

- Atomic operations on Fermi L2 cache
- Medium latency, but still serialized
- Shared among all blocks
- “Free improvement” on Global Memory atomics

![Diagram showing time, internal routing, data transfer, atomic operation N, and atomic operation N+1]
Hardware Improvements

- Atomic operations on Shared Memory
  - Very short latency, but still serialized
  - Private to each thread block
  - Need algorithm work by programmers (more later)
No Recommended Reading