CUDA memory model

- gld_efficiency

global memory load efficiency

- gld_throughput

global memory load throughput

- gld_transactions

global memory transactions

- gld_transactions_per_request

how many memory transactions needed for one memory request.

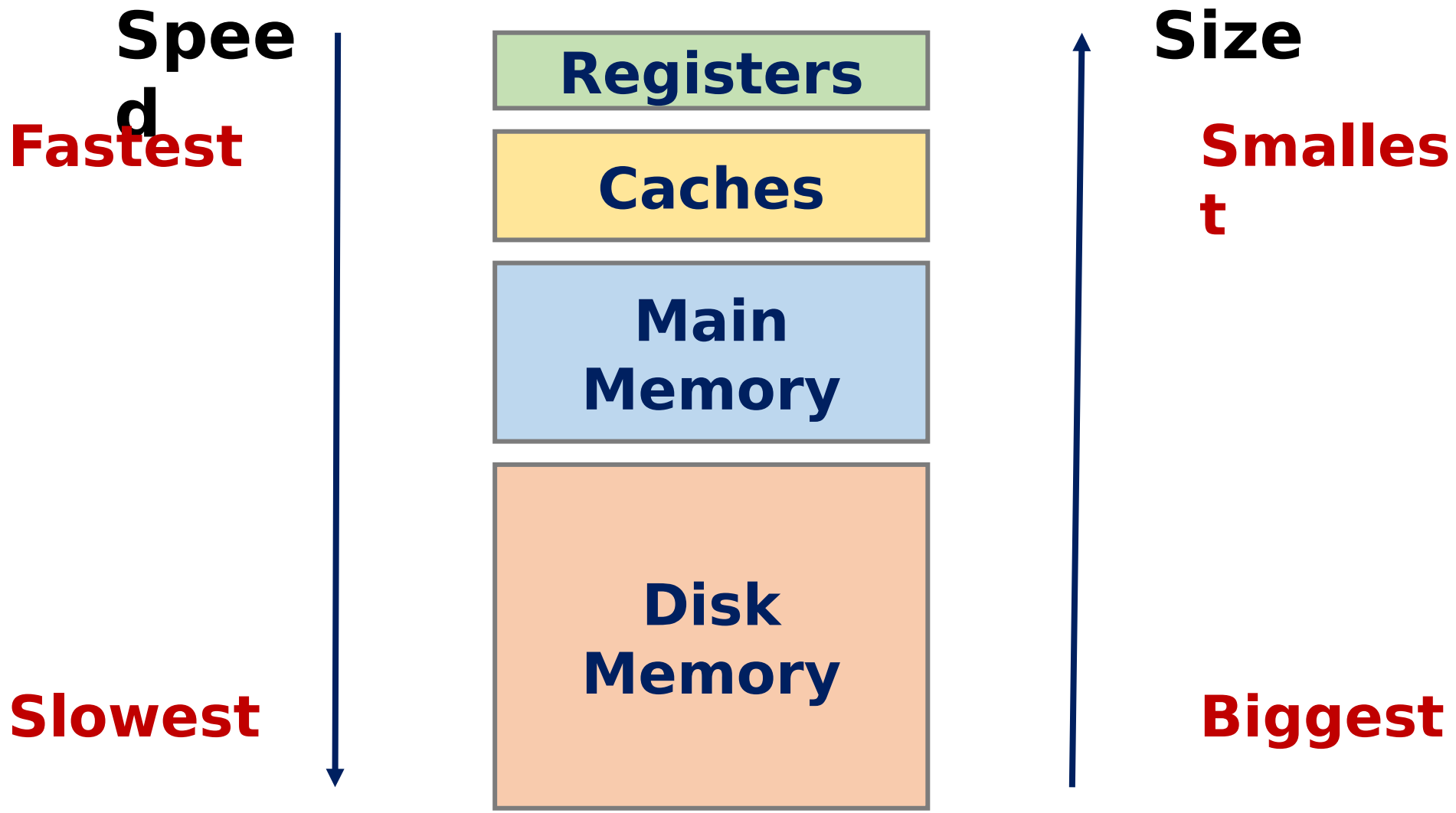
CUDA memory model

Locality

- Applications access a relatively small and localized portion of their address space at any point-in-time
  - Temporal locality
  - Spatial locality

Locality in time

Locality in space


Speed
Fastest
Slowest
Registers
Caches
Main Memory
Disk Memory
Size
Smallest
Biggest