Johns Hopkins
Engineering for Professionals

605.767 Applied Computer Graphics

Brian Russin

Module 11H
Memory and Bandwidth

Memory Architecture

- Dedicated memory
  - Called video memory
  - Not accessible by CPU
  - Common uses
    - Buffers (vertex, uniform, framebuffers)
    - Texture
- **Unified Memory Architecture (UMA)**
  - Memory shared between graphics and host
  - Caches are different
  - Typical in gaming consoles (Xbox, Playstation) and embedded devices (iOS devices)
  - Intel's Gen 9 UMA Architecture

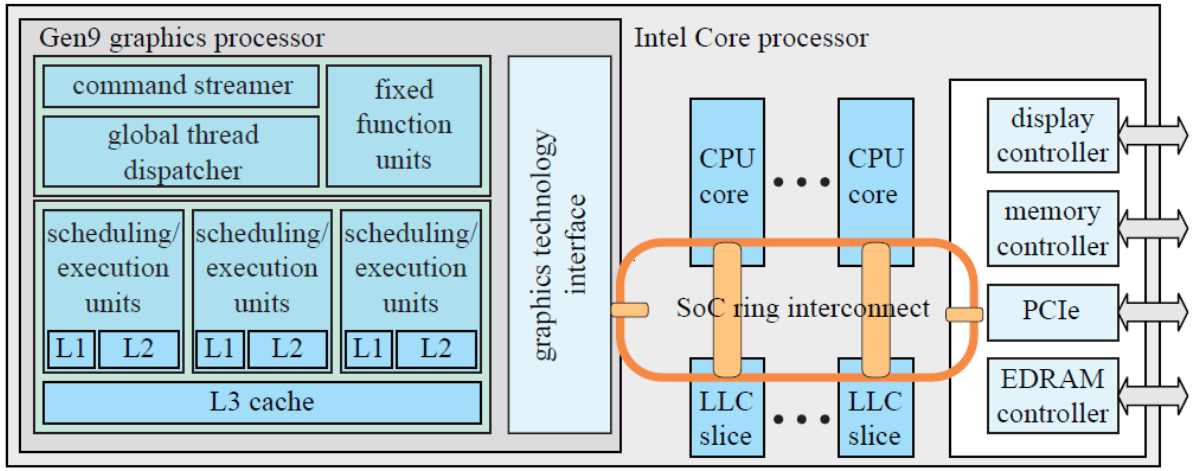

Figure 23.9 (UMA): Intel's Gen9

Texture Caching and Prefetching

- Trend has been to use more textures per primitive
- Reading from texture memory is main use of bandwidth
  - Steps must be taken to reduce bandwidth and latency
  - To reduce bandwidth, caches are used at various places in pipeline
  - To reduce latency, prefetching is used
- Small on-chip memory caches store result of recent texture reads
  - Neighboring pixels often access same or nearby texels
    - Likely to find in the cache
  - Reading texels into the cache takes time
    - So caching is often combined with prefetching
    - Extremely useful processing many fragments simultaneously

Port and Bus Bandwidth

- Port – channel for sending data between 2 devices
  - Accelerated Graphics Port (AGP) connects CPU and graphics accelerator
- Bus – shared channel for sending data among more than 2 devices
- Bandwidth – describes throughput of data over port or bus in bytes per second
- Cases
  - PCIe v3: 15.75 GB/s
  - PCIe v4: 31.51 GB/s
  - Pascal (dedicated video memory) for GTX 1080: 320 GB/s

Latency

- Latency - time between making a query and getting the result
  - In pipelined system with n stages it takes at least n clock cycles to get through entire pipeline
  - Example: 700 pipeline stage average, 200 MHz = 5ns per clock cycle
    - 5ns*700pipeline stages=3.5 μs (microseconds)
    - 20 ms per frame = 50 frames per second
      - Can pass through entire pipeline many times per frame
- Latency can hurt performance if it is necessary to read back from the pipeline
  - Read back a 256x256 square of the depth buffer can take approx. 1 ms
  - For real-time performance one should avoid any reads back from graphics accelerator
- Also, pipeline often has to be flushed before reading back data from graphics accelerator
  - CPU is idle during this time

Caching and Compression

- Memory hierarchy typically used
- Cached data in tile tables
  - Exploits the locality fetching often seen in graphics applications
    - Sequential memory reads often nearby
- Compression
  - Different data uses different compression algorithms
    - Lossless
    - i.e. Depth data compressed differently than color
      - Color data compression can be lossy; this is an active area or research
  - Pre-cache compression: data compressed before placed in cache
    - Increases the effective capacity of the tile tables but increase complexity
  - Post-cache compression: data compressed after placed in cache, but before placed in memory hierarchy