Johns Hopkins
Engineering for Professionals

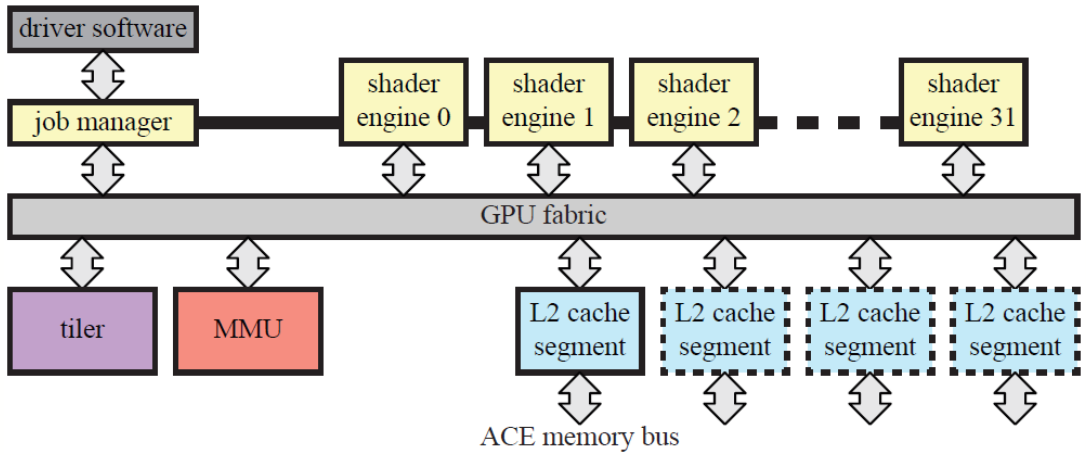
605.767 Applied Computer Graphics

Brian Russin

Module 11I
Case Studies

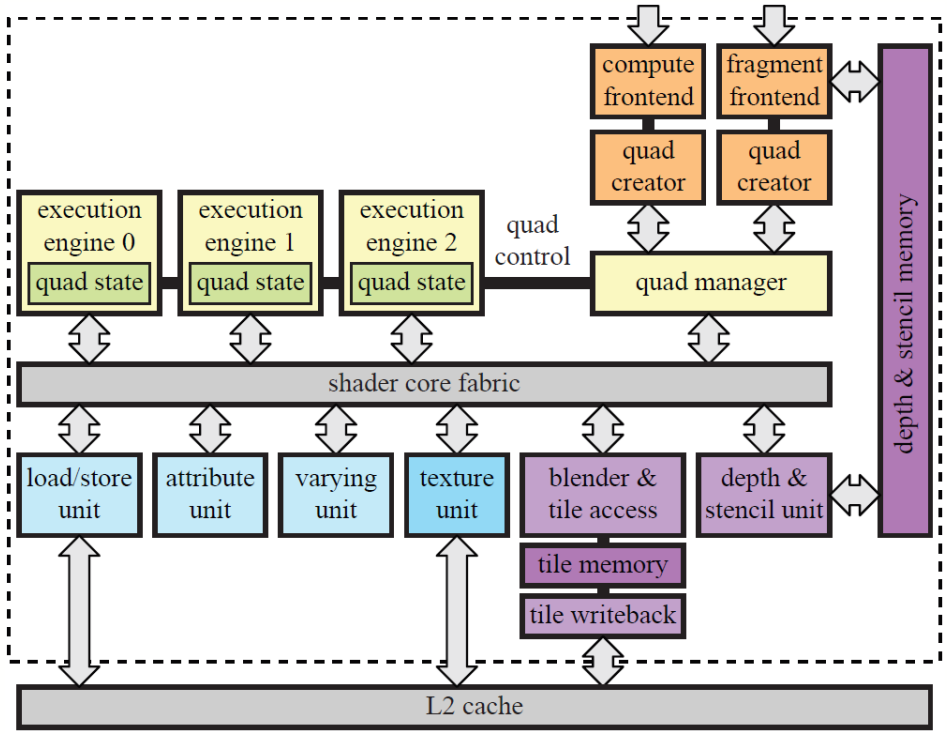
ARM Mali G71 Bifrost

- Targets mobile and embedded devices
  - Energy efficiency is a primary concern
- Sort-middle design
- 12 ALUs (supports 12 simultaneous threads)
- 32 cores (shader engines)
  - G71 has 32, the design scales much higher
  - cores communicate with each other via the GPU fabric
- Job manager schedules jobs to the shader engines
- All memory accessed through the Memory Management Unit (MMU)



ARM Mali G71 Bifrost - Shader Engine

- 3 execution engines
  - With 4 FMA (fused-multiply-and-add) units for SIMD of 4
    - Each core has 12 SIMD lanes
- **Unified** design
  - Can perform vertex, fragment, and compute shading
- Dedicated tile memory for framebuffers



ARM Mali G71 Bifrost (cont.)

- Vertices executed all at once
- Triangles results sorted into bins
  - Bins are executed in parallel
- Bin execution
  - Rasterization
  - Pixel Processing
- Other features
  - **Pixel Locale Storage (PLS)**
    - Fragment shader can read from the color buffer
    - Custom blending and deferred shading
  - Transaction elimination
    - Each tile has unique signature
      - Subsequent tiles (next frame) with same signature not output
  - Smart composition
    - Memory cached from frame to frame if no changes have occurred

NVIDIA Pascal

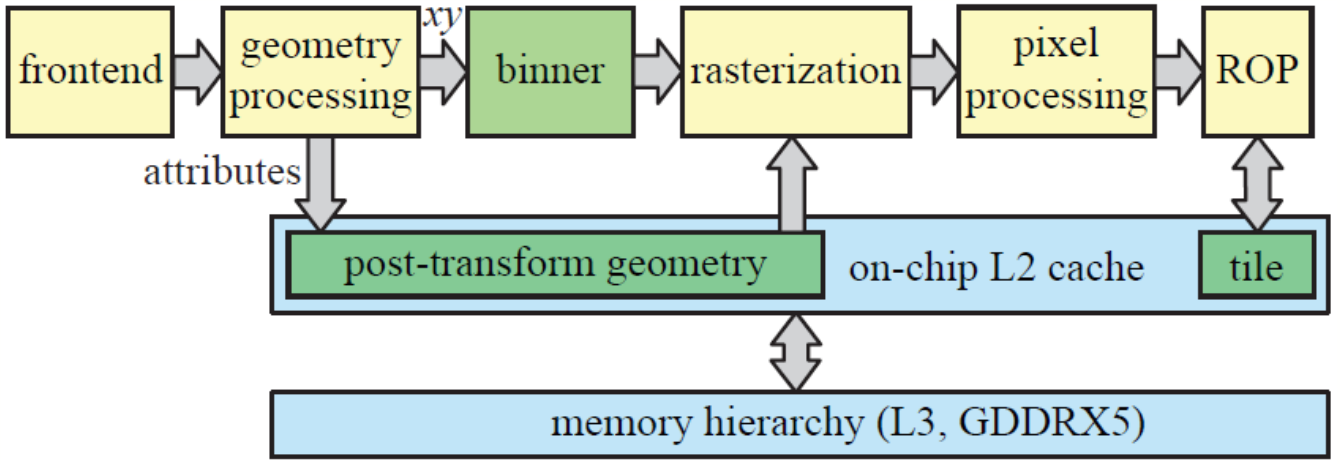
- Used in GeForce GTX 1000 series (i.e. GeForce GTX 1080)
- Separate architecture for graphics and compute (general-purpose)
- Unified ALUs
  - Has both floating-point unit and integer unit
  - Also known as CUDA Cores
- Arranged in blocks of 32 called a **streaming multiprocessor (SM)**
  - Can execute 4 warps of 32 threads simultaneously
  - 256 kB of instructions per SM
  - 48 kB L1 cache memory per SM
- Shading performed in 2 x 2 pixel blocks
  - Scheduler can group vertices, pixels, primitives (tessellation), or compute shader work

NVIDIA Pascal (cont.)

- **Polymorph Engine** unit performs geometry-related tasks
  - i.e. Vertex fetch, tessellation, attribute setup, etc.
- **Texture Processing Cluster (TPC)**
  - A polymorph engine unit coupled with a streaming multiprocessor
- **Graphics Processing Cluster (GPC)**
  - 5 TPCs
  - Can be thought of as a mini-GPU
- Pascal variations
  - GTX 1080 has 4 GPCs
  - GTX 1080 Ti has 6 GPCs
    - 2 GPCs are smaller than those in the 1080
  - GTX 1070 has 3 GPCs
  - GTX 1050 has 2 GPCs

NVIDIA Pascal (cont.)

- **Tiled caching**
  - Hybrid sort-middle and sort-last fragment methods
  - Exploits locality and the L2 cache
  - Sorts geometry into bins, but also keeps the framebuffer in the L2 cache until geometry has finished processing



Other Architectures

- See text for description of AMD GCN Vega
  - Used in Xbox One and Playstation 4
- Real-time Ray Tracing
  - Text was written before NVIDIA's RTX
  - Focus on graph traversal and intersections
  - Uses BV Hierarchy and AABBs
  - Ray Tracing Unit
    - Intersection processor
    - Coherency engine
      - Gathers rays with similar properties to exploit locality