Realtime Computer Graphics on GPUs

Speedup Techniques, Other APIs

Jan Kolomazník

Department of Software and Computer Science Education
Faculty of Mathematics and Physics
Charles University in Prague

Optimizations Intro

Performance Bottlenecks I

- Most of the applications require steady framerate – What can slow down rendering?
- Too much geometry rendered
  - CPU/GPU can process only limited amount of data per second
  - Render only what is visible and with adequate details
- Lighting/shading computation
  - Use simpler material model
  - Limit number of generated fragments
- Data transfers between CPU/GPU
  - Try to reuse/cache data on GPU
  - Use async transfers

Performance Bottlenecks II

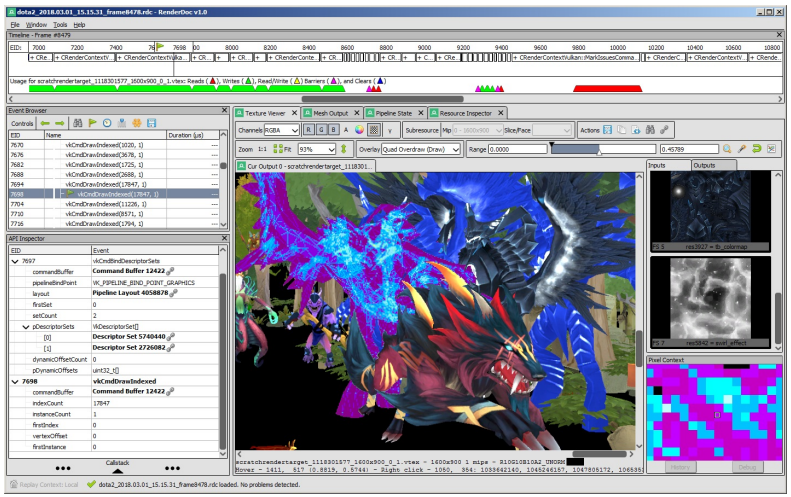
- State changes
  - Bundle object by materials
  - Use UBOs (uniform buffer objects)
- GPU idling – cannot generate work fast enough
  - Multithreaded task generation
  - Not everything must be done in every frame – reuse information (temporal consistency)
- CPU/Driver hotspots
  - Bindless textures
  - Instanced rendering
  - Indirect rendering

Different Needs

- Large open world
  - Rendering lots of objects
  - Not so many details needed for distant sections
- Indoors scenes
  - Often lots of same objects – instaced rendering
  - Only small portion of the scene visible – occluded by walls, ...
- CAD, molecular biology, ...
  - All geometry visible at once
  - Lots of self-occlusions
  - Switching levels of detail

PROFILING

- Three rules of profiling:
  - Measure!
  - Measure!
  - Measure!
- How to measure?
  - Code instrumentation
  - Sampling profilers
  - HW event counters
- Profiling influences performance
- Tools for GPU:
  - NVIDIA Nsight
  - Radeon GPU Profiler
  - RenderDoc
  - apitrace
  - CodeXL



Scene Graph

- Boundary volume hierarchies (BVH)
- Spatial partitioning
  - Binary space partitioning
  - Quadtrees, octrees
  - Grids
- Additional meta-objects:
  - Occluders
  - Collision geometry

Standard Rendering Procedure

1. Analyze world view
2. Transfer data on GPU
   - Upload/reuse geometry in VBOs
   - Upload/reuse textures
3. Issue draw commands

Optimize Rendering

Query Objects

- How to use:

```
glGenQueries(GLsizei n, GLuint * ids);
glBeginQuery(GLenum target, GLuint id);
   // ...
glEndQuery(GLenum target);
```

- Example queries:

  `GL_SAMPLES_PASSED`

  `GL_ANY_SAMPLES_PASSED`

  `GL_PRIMITIVES_GENERATED`

  ...

- Usualy used with one frame lag

Occlusion Culling

- Do not render objects hidden behind others
- Helper objects – occluders
- CPU processing
  - Analyze scene graph + occluders to filter rendered geometry
- GPU processing
  - Z-buffer pre-render
  - Render occluders to Z-buffer
  - Occlusion queries
  - Temporal consistency – Z-buffer reprojection

Conditional Rendering

- GL version is 3.0 or greater
- Render cheap object
- Use occlusion query to see if any of it is visible
- If it isn't, skip rendering of expensive object

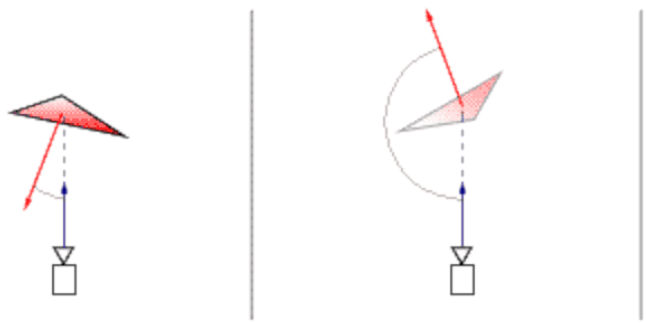
```
glBeginConditionalRender(GLuint id, GLenum mode);
glEndConditionalRender();
```

Z-buffer Optimizations

- Early Z-test:
  - Fragment shader cannot modify depth value
  - Front-to-back rendering (rough sort)
  - Double speed Z-only:
    - First pass depth (stencil) only, write z-buffer – faster when no color buffer present
    - Second pass color writes, z-buffer read-only
  - Optimizations:
    - Z-pass for major occluders
- Deffered shading

Backface Culling

- Good closed meshes
  - No need for triangle back-side rasterization
- GPU can filter (face cull) according to vertex order:
  - glEnable(GL_CULL_FACE);
  - glFrontFace(GL_CCW);
  - glCullFace(GL_BACK); // draw front faces only



Clipping Planes

- Objects outside of the view frustum are not rendered
- Set far-plane reasonable close (also increases z-buffer precision)
- Harsh end of all geometry
  - Render background geometry
  - Hide the far plane in fog (glFog())
  - More sophisticated effect in fragment shader/deffered shading

Level of Detail (LoD)

- Optimal rendering efficiency:
  - details of distant objects (pixel size) need not be drawn
  - the closest objects (and/or user focus) need the best available visual quality
- dynamic level of detail
  - rendering system adjusts individual rendering accuracy
  - global parameter definition (e.g. total approximate number of drawn triangles)
  - advance data preparation: discrete LoD levels / continuous LoD pre-processing..

Discrete LoD

- fixed LoD levels are prepared in advance
  - shape approximations with different accuracy
  - can come from the finest model – generalization (can be time consuming)
- rendering – choosing appropriate LoD level:
  - according to object-camera distance
  - according to bounding object projection size or even exact object projection – errors perpendicular to viewing direction are most noticeable
  - object importance, “player focus”, ..
  - global balancing (declared number of triangles to draw)

Continuous LoD

- Progressive meshes (Hoppe)
  - Hierarchy of simplifying operations
  - Smooth transition between differetn LoDs
  - High memory consuption
  - Problematic to implement on parallel architectures (GPUs)
- LoD streaming:
  - Analyze required LoD levels
  - Stream missing geometry and required operations to GPU
  - Temporal coherence

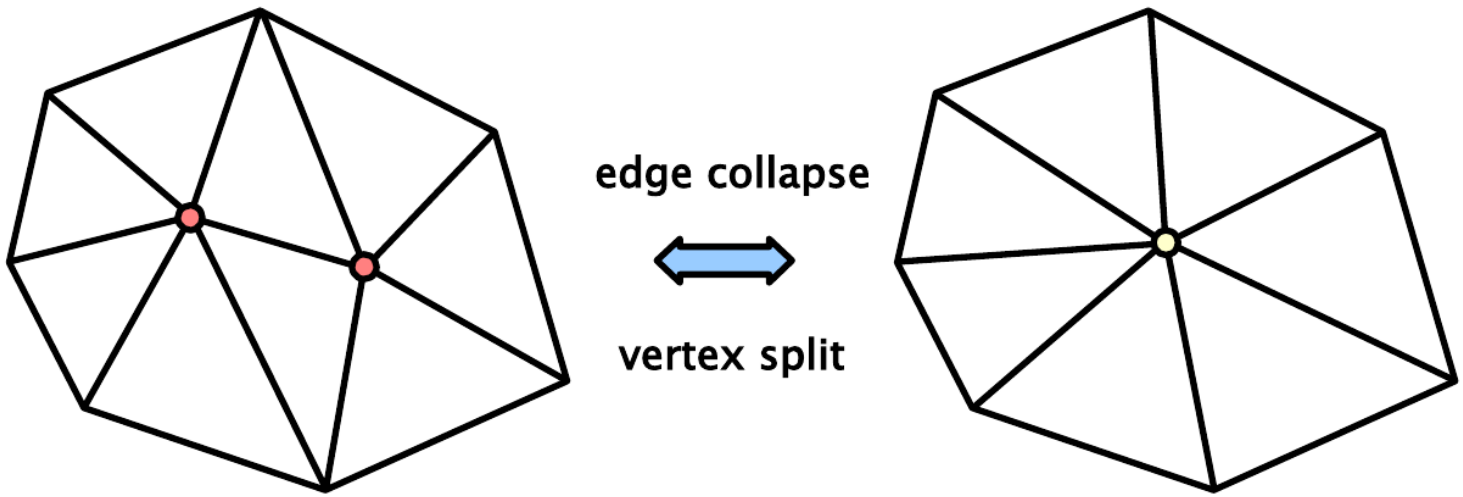
LoD Transitions

- simple switching
  - hysteresis must be introduced (against unwanted ”popping”)
- blending neighboring LoD levels
  - drawing both neighboring levels using transparency
  - linear combination (blending, “transition”)
  - another option
    - current LoD level is opaque
    - additional semitransparent new LoD (“over” a operation)
    - “z-writing” enabled only for the current LoD level

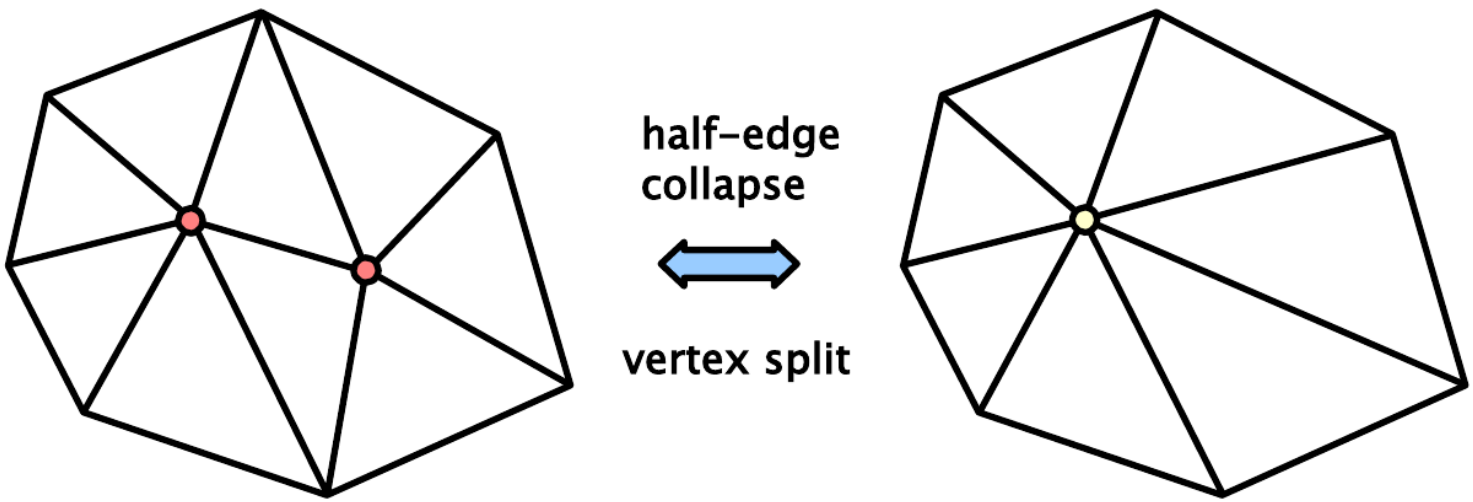
Generating LoDs

- top-down approach
  - based on the simplest shape representation (low-poly), additional details are introduced
  - less frequent (subdivision surfaces..)
- bottom-up
  - starts with detailed (most accurate) model
  - gradual simplification / generalization (data reduction)

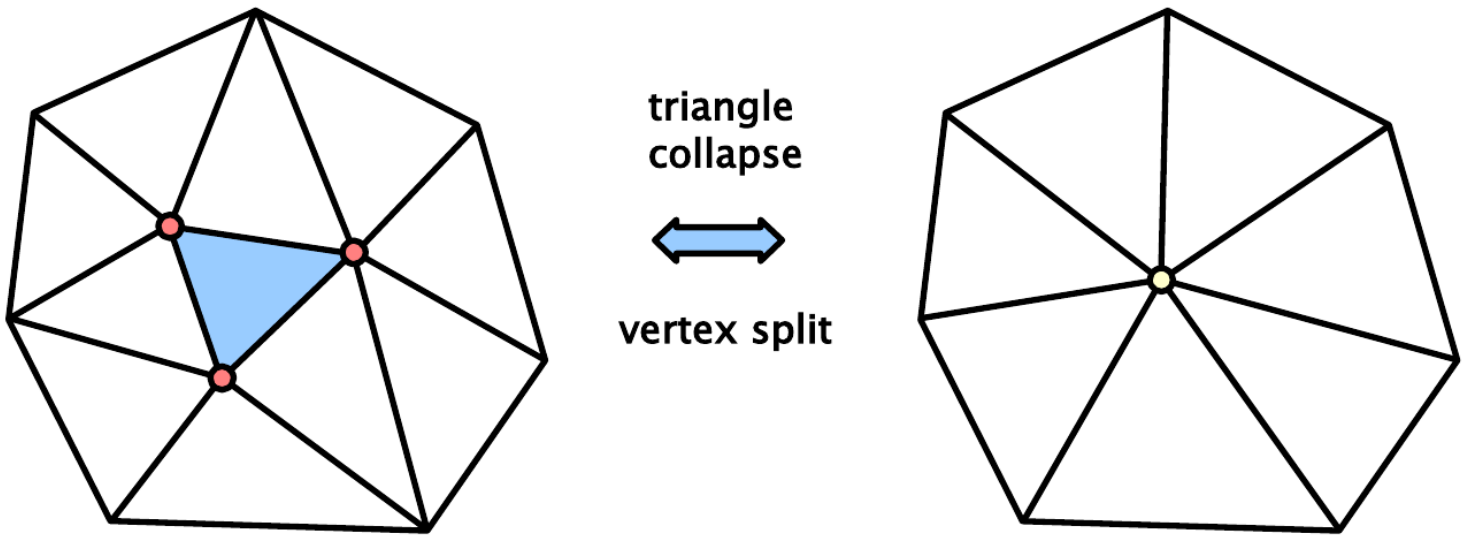
EDGE COLLAPSE VS. VERTEX SPLIT



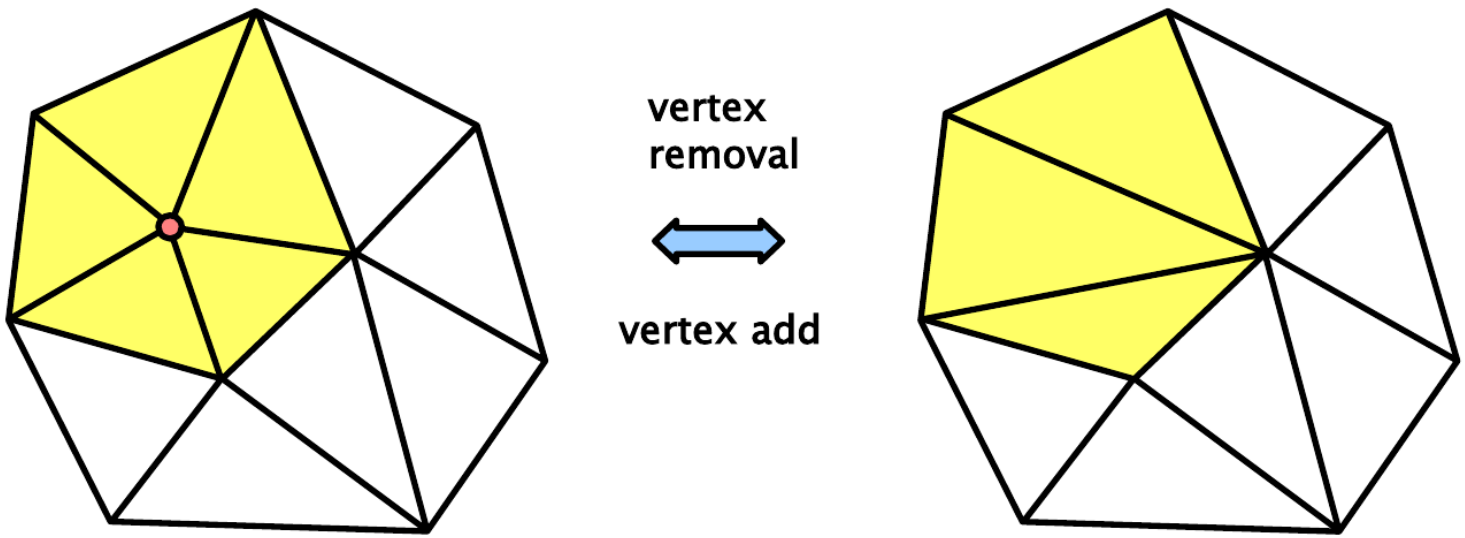
HALF-EDGE COLLAPSE VS. VERTEX SPLIT



TRIANGLE COLLAPSE COLLAPSE VS. VERTEX SPLIT



VERTEX REMOVAL VS. VERTEX ADD



TRIANGLE STRIPS/FANS

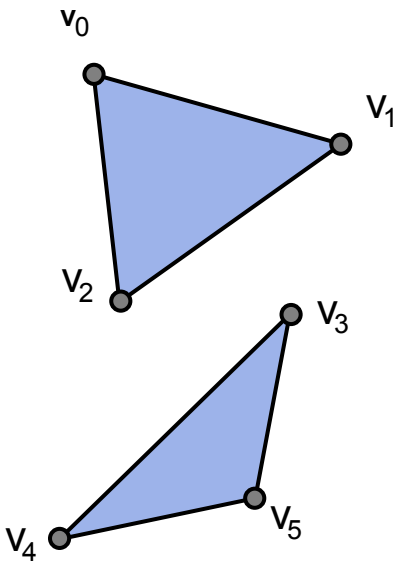


Figure:
GL_TRIANGLES

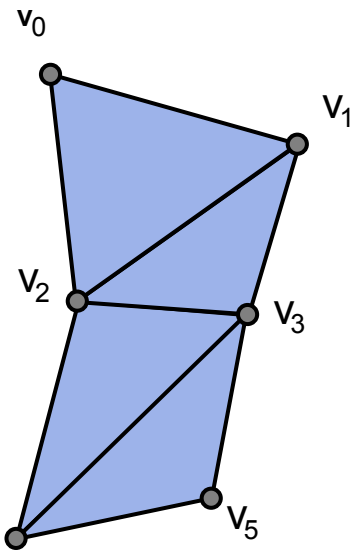


Figure:
GL_TRIANGLE_STRIP

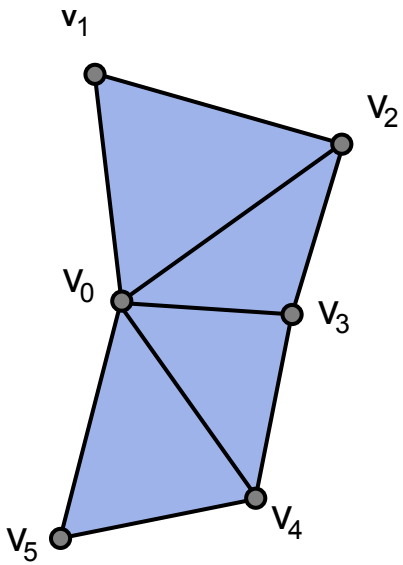


Figure:
GL_TRIANGLE_FAN

Limit Draw Calls

- Instanced rendering
- Multidraw commands – bundle more indexed render calls together

```
void glMultiDrawElements(GLenum mode, const GLsiz
```

- Indirect rendering
  - GL_DRAW_INDIRECT_BUFFER
  - Fill on GPU

```
void glDrawElementsIndirect()
void glDrawArraysIndirect()
void glMultiDrawElementsIndirect()
```

MULTITHREADING

- Contexts are not thread-safe
- Context can be current only in one thread
  - It can be used in multiple threads, but not in parallel
- Multiple contexts can be used concurrently
  - Enable resource sharing – textures, VBOs, . . .
- Split preparation work between multiple threads (task systems)

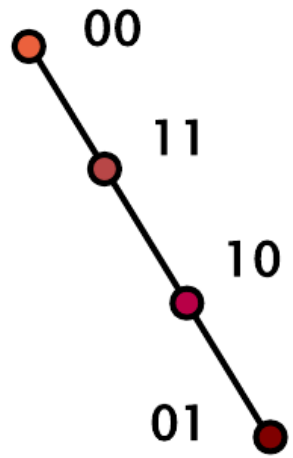
Textures

Texture Access

- Texture accesses expensive
  - Limit number of tex reads per fragment
- Use caching efficiently
  - Textures have spatial caching
  - Neighboring fragments should access neighboring texels
  - Use reasonable sized textures
  - Use mip-mapping

Texture Compression I

- Invention: S3 in DirectX 6 (1998)
  - DXTC, in OpenGL: S3TC, DXT1
- Fixed compression ratio
  - necessary for memory management
  - 4:1 to 6:1 lossy compression
- Decomposition into rectangle tiles (4×4 px)
  - each tile: two 16-bit colors and sixteen 2-bit indices (together – 4 bpp)
    - two extreme colors (R5G6B5), two more in-between colors (or one in-between and black)
    - each pixel is represented by a reference to one color
  - Extensions add alpha compression (128-bit): DXT3, DXT5



Texture Compression II

- NVIDIA VTC (Volume Texture Compression)
  - 3D variant of S3TC
  - data blocks $4 \times 4 \times 1$, $4 \times 4 \times 2$, $4 \times 4 \times 4$
- BPTC Texture Compression:
  - Byte stream
  - Unsigned + float
  - Multiple gradients

Other APIs

OpenGL ES

- OpenGL for Embedded Systems (smartphones, tablet computers, video game consoles)
- Most widely deployed 3D graphics API in history
- Subset of the OpenGL API
- Latest version of OpenGL ES 3.2
  - Adds tesselation and geometry shaders

WebGL

- JavaScript API for rendering interactive 2D and 3D graphics within any compatible web browser without the use of plug-ins
- Control code written in JavaScript and shader code that is written in OpenGL ES Shading Language
- WebGL 1.0 is based on OpenGL ES 2.0, HTML5 canvas element
- WebGL 2.0 is based on OpenGL ES 3.0
- WebAssembly
  - Compile code for virtual architecture
  - Executable in browser + OpenGL ES
- WebGPU – future API based on Vulcan and Metal

DirectX (Direct3D) I

- Available only for Microsoft platforms (Windows, XBox)
- Wraps same hardware – similar principles to OpenGL
- API organization different
  - State machine vs. parameter objects
  - Resources managed by developer
- Component Object Model (COM)
  - MS binary-interface standard for software components
  - Support for interprocess communication, multiple languages
  - Entities (textures, render targets, shaders, . . .) are COM objects
- Shading language HLSL
  - Similar to GLSL (almost 1:1 mapping)
  - Cross compilers available

DirectX (Direct3D) II

- Not a standard – more flexible to introduce changes (backward incompatible)
- DirectX 9 $\longleftrightarrow$ OpenGL 2
- DirectX 10 $\longleftrightarrow$ OpenGL 3
- DirectX 11 $\longleftrightarrow$ OpenGL 4
- DirectX 12 $\longleftrightarrow$ Vulkan, Metal

Direct 3D

Basic interfaces:

```
IDXGISwapChain *swapchain;              // the pointer to the swap chain interface
ID3D11Device *dev;                      // the pointer to our Direct3D device interface
ID3D11DeviceContext *devcon;            // the pointer to our Direct3D device context
```

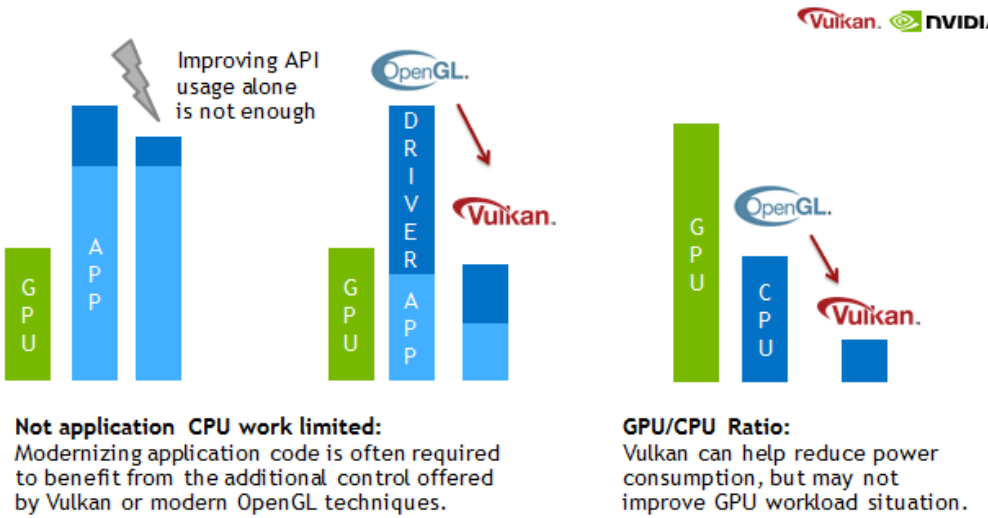
- ID3D11Device thread-safe
  - Can prepare resources in multiple threads
- ID3D11DeviceContext must use some synchronization
  - Draw commands cannot be issued in parallel

New APIs

- Opengl, DirectX (11 and older)
  - Old APIs – designed fo different ecosystems
  - Hotfixed for demands of modern hardware and software architectures
- CPU time in driver – hot spot in lots of modern applications
  - Multicore CPUs cannot efficiently feed command to GPUs
- AMD Mantle $\rightarrow$ Vulkan, Metal (Apple)
  - Explicit command buffer control
  - Multithreaded parallel rendering
  - Lower level API – developers have more control (also more code is needed)
- DirectX 12
  - Added similar features
  - Command queues, feeding from multiple threads

VULKAN

- Managed by Khronos
- Same API for desktop and embedded systems
- Low CPU usage, efficient usage of multi-core systems (thread-safe)
- No global state
- GLSL compiled to SPIR-V (Standard Portable Intermediate Representation)
- OpenCL convergence to Vulkan



Vulcan Usage I

- Create a VkInstance
  - Specify required features, extensions
- Select graphics card (VkPhysicalDevice)
  - Listing cards supporting requested features
- Create a logical device (VkDevice)
- Create queues (VkQueue) from queue families
  - Graphics
  - Compute
  - Memory transfers
- Create a window, window surface and swap chain

Vulcan Usage II

- Wrap the swap chain images into VkImageView
- Create a render pass
  - Specifies the render targets and usage
- Create framebuffers for the render pass
- Set up the graphics pipeline (VkPipeline)
  - Configured in advance
- Allocate and record a command buffer with the draw commands
- Draw frames
  - Acquire images
  - Submit draw commands
  - Return images to swap chain
  - Synchronization required