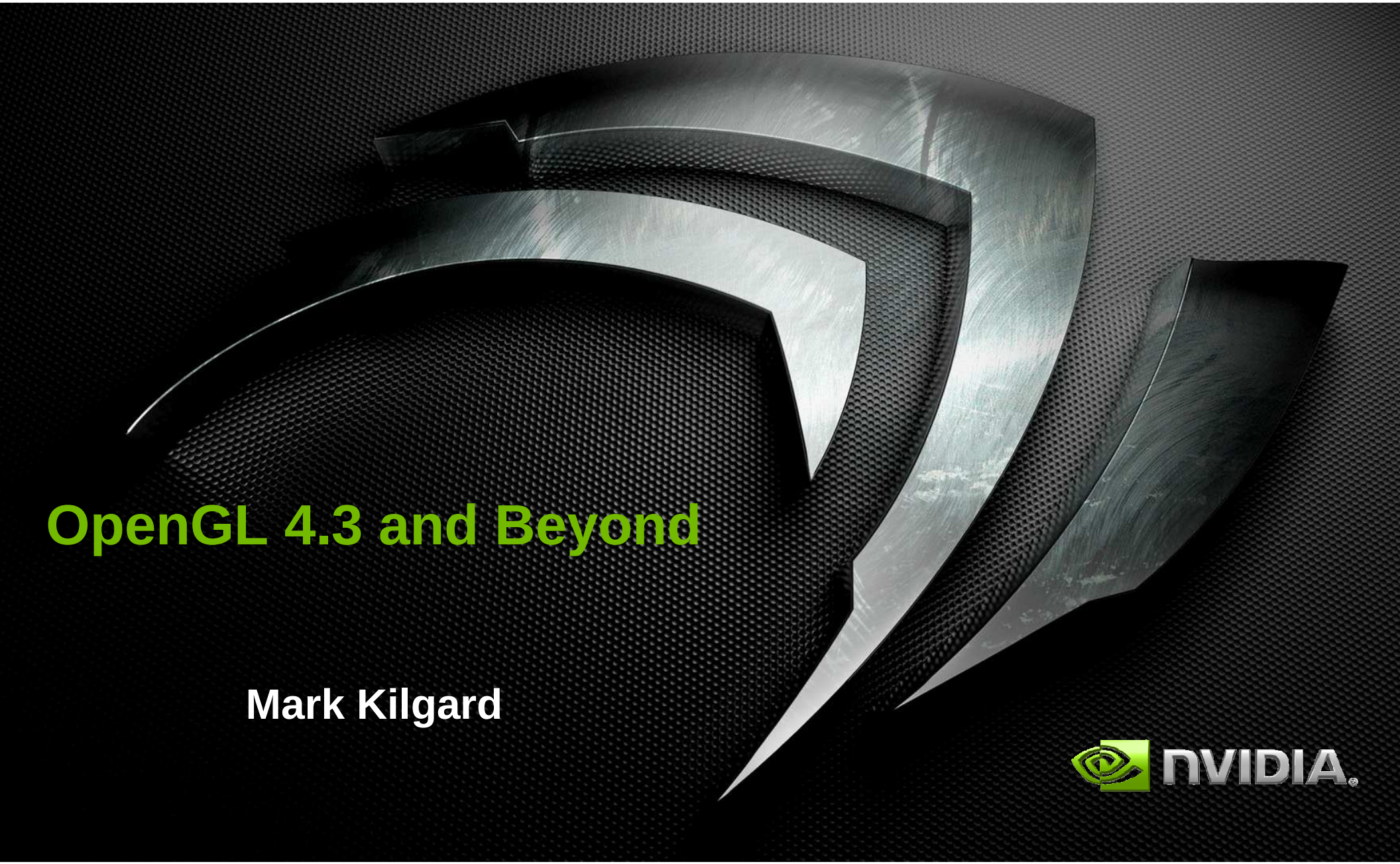

OpenGL 4.3 and Beyond
Mark Kilgard
NVIDIA

Talk Details

Location: Conference Hall K, Singapore EXPO
Date: Thursday, November 29, 2012
Time: 11:00 AM - 11:50 PM

Mark Kilgard (Principal Software Engineer, **NVIDIA**, Austin, Texas)

Abstract: Attend this session to get the most out of OpenGL on NVIDIA Quadro and GeForce GPUs. Learn about the new features in OpenGL 4.3, particularly Compute Shaders. Other topics include bindless graphics; Linux improvements; and how to best use the modern OpenGL graphics pipeline. Learn how your application can benefit from NVIDIA's leadership driving OpenGL as a cross-platform, open industry standard.

Topic Areas: Computer Graphics; Development Tools & Libraries; Visualization; Image and Video Processing

Level: Intermediate

Mark Kilgard

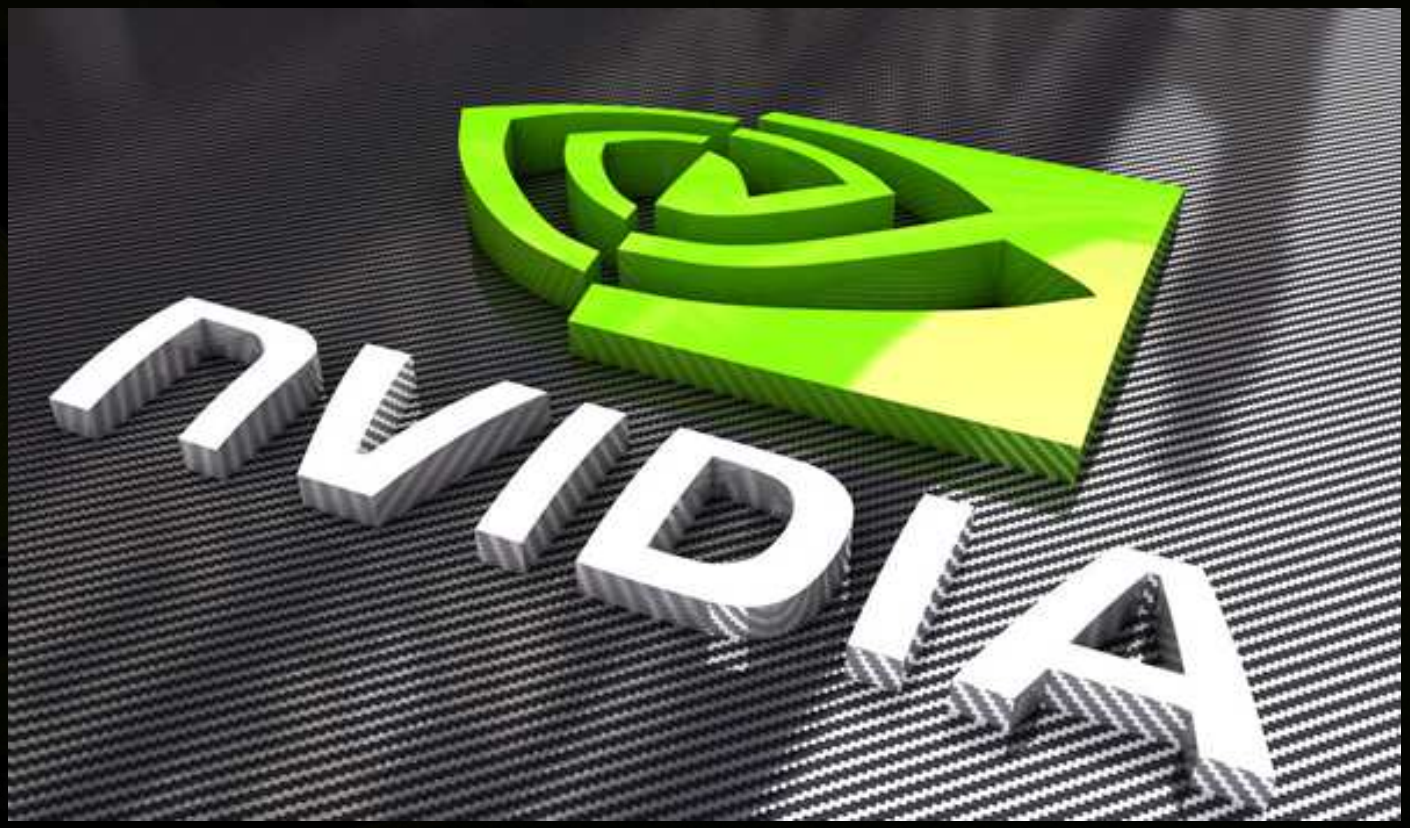
- Principal System Software Engineer
  - OpenGL driver and API evolution
  - Cg (“C for graphics”) shading language
  - GPU-accelerated path rendering
- OpenGL Utility Toolkit (GLUT) implementer
- Author of *OpenGL for the X Window System*
- Co-author of *Cg Tutorial*

- *Worked on OpenGL for 20+ years*

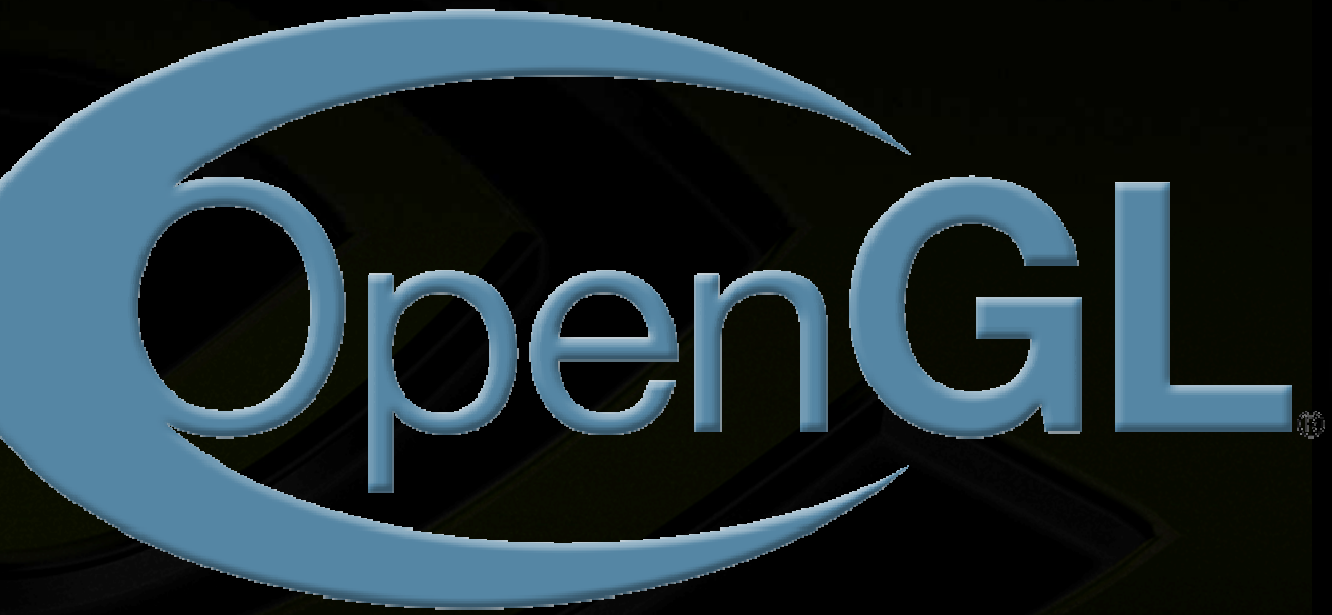
Outline

- State of OpenGL & OpenGL’s importance to NVIDIA
- Compute Shaders explored
- Other stuff in OpenGL 4.3
- Further NVIDIA OpenGL Work
- How to exploit OpenGL’s modern graphics pipeline

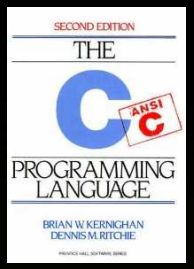
State of OpenGL &
OpenGL's importance to NVIDIA



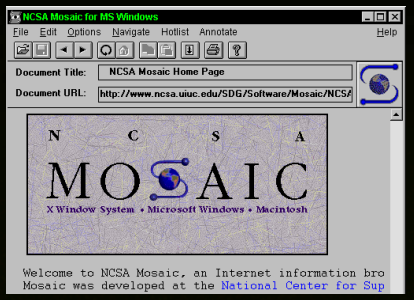
OpenGL Standard is 20 Years and Strong

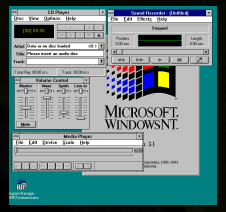


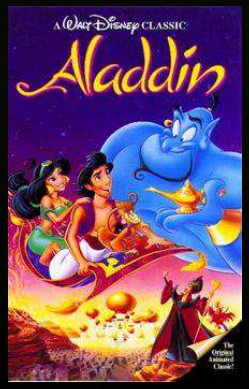
Think back to Computing in 1992



- Programming Languages
  - ANSI C (C 89) was just 3 years old
  - C++ still implemented as a front-end to C
  - OpenGL in 1992 provided FORTRAN and Pascal bindings
- One year before NCSA Mosaic web browser first written
  - Now WebGL standard in almost every browser
- Windows version
  - Windows 3.1 ships!
  - NT 3.1 still a year away
- Entertainment
  - Great video game graphics? Mortal Kombat?
  - Top grossing movie (Aladdin) was animated
    - Back when animated movies were still hand-drawn

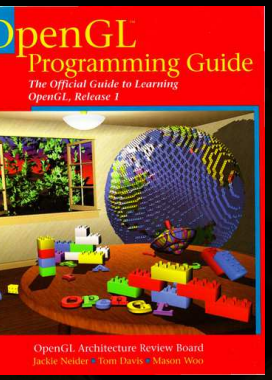


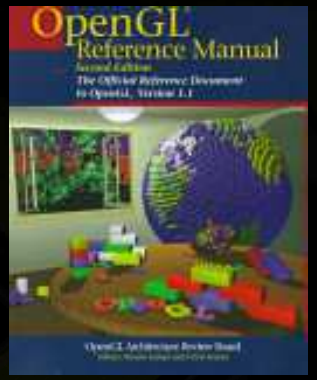


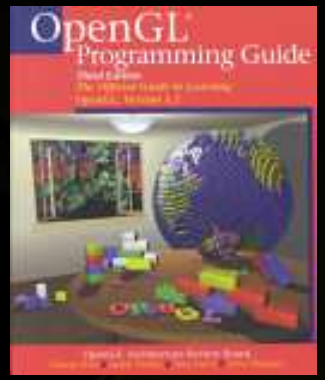


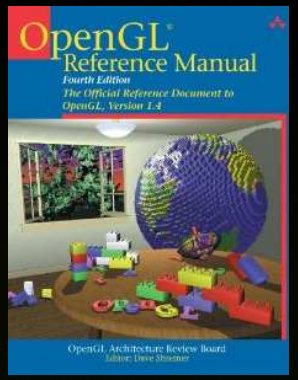
20 Years Ago: Enter OpenGL

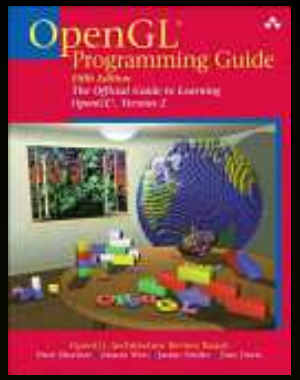
20 Years in Print

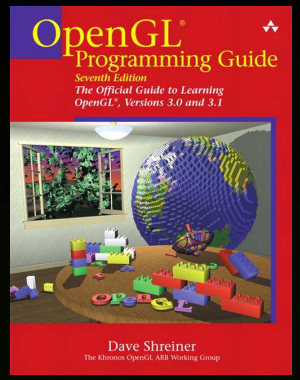


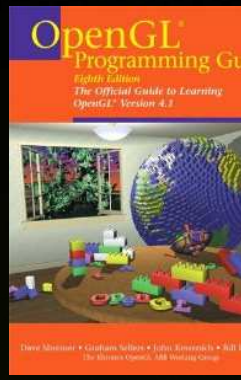


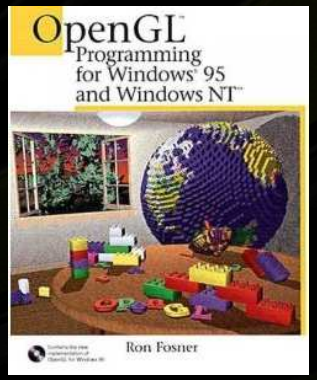


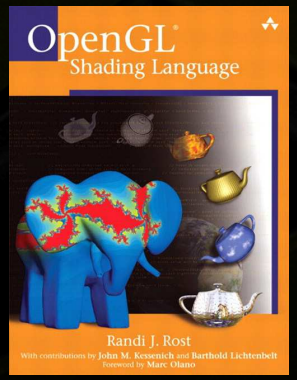


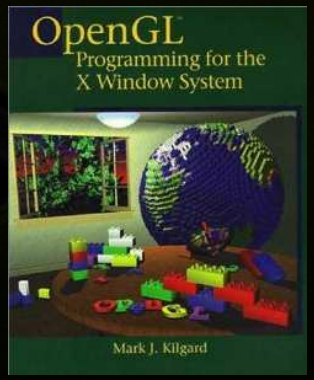


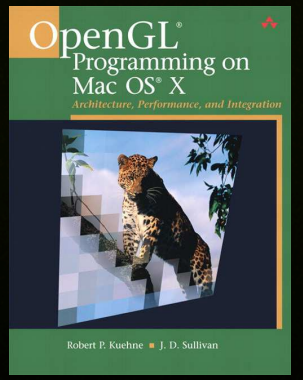












Then and Now

OpenGL 1.0: Per-vertex lighting

1992

2012

OpenGL 4.3: Real-time Global Illumination

[Cras

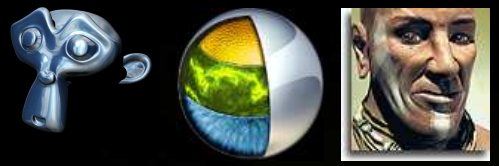
Big News

OpenGL 4.3

- OpenGL 4.3 announced at SIGGRAPH Los Angeles
  - August 6, 2012
  - *Moments later…* NVIDIA beta OpenGL 4.3 driver on the web
    http://www.nvidia.com/content/devzone/opengl-driver-4.3.html
  - Now fully WHQL'ed production driver with OpenGL 4.3 now shipping
- OpenGL 4.3 brings substantial new features
  - Compute Shaders!
  - OpenGL Shading Language (GLSL) updates (multi-dimensional arrays, etc.)
  - New texture functionality (stencil texturing, more queries)
  - New buffer functionality (clear buffers, invalidate buffers, etc.)
  - More Direct3D-isms (texture views, parity with DirectX compute shaders)
  - OpenGL ES 3.0 compatibility

quee
ture

NVIDIA's OpenGL Leverage



Programmable
Graphics (GLSL, Cg)

GeForce

Debugging w
Parallel Nsigh

OpenGL

Tegra

OptiX

Quadro

Single 3D API for Every Platform

OS X

Windows

Linux

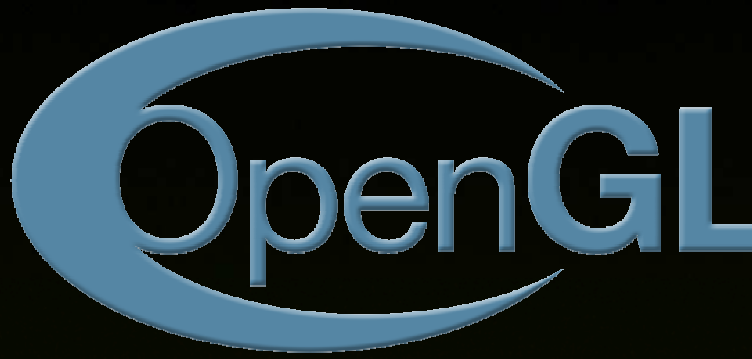
FreeBSD

Solaris

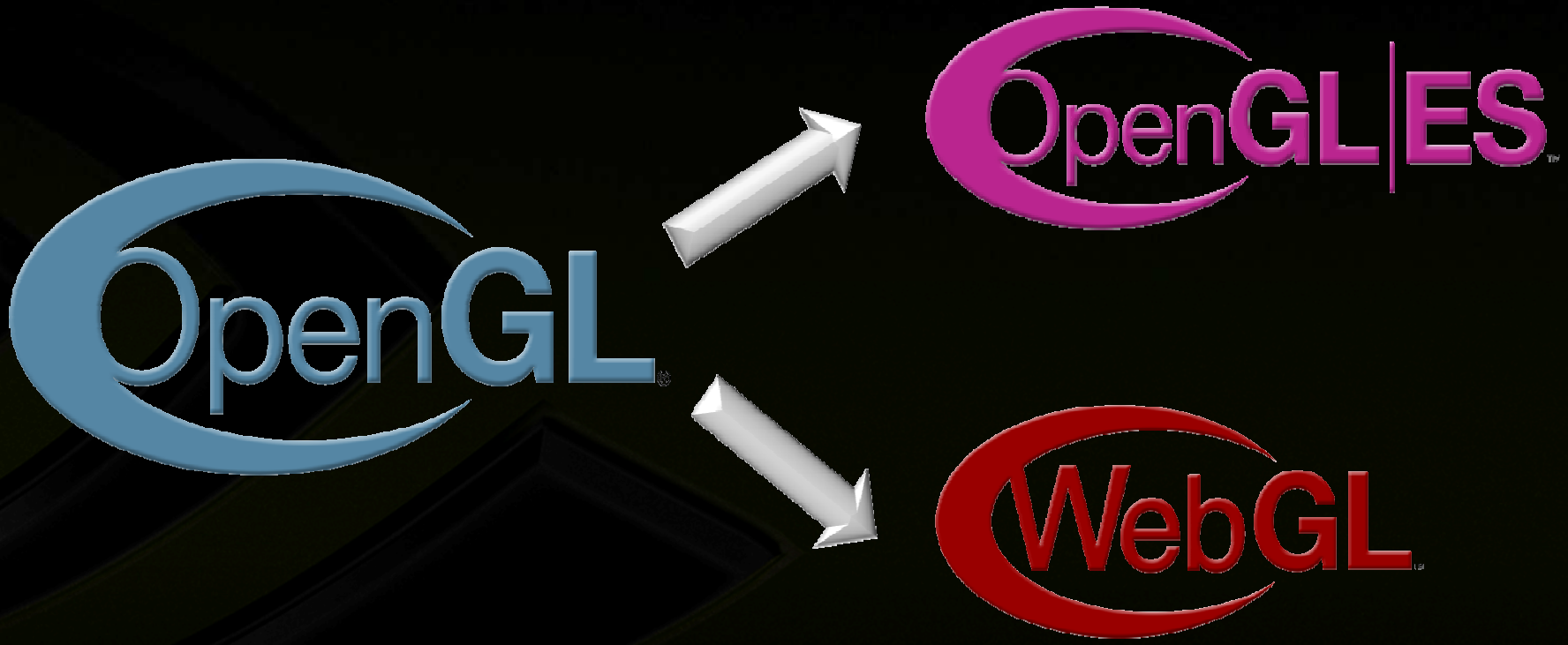
Android

OpenGL 3D Graphics API

- cross-platform
- most functional
- peak performance
- open standard
- inter-operable
- well specified & documented
- 20 years of compatibility



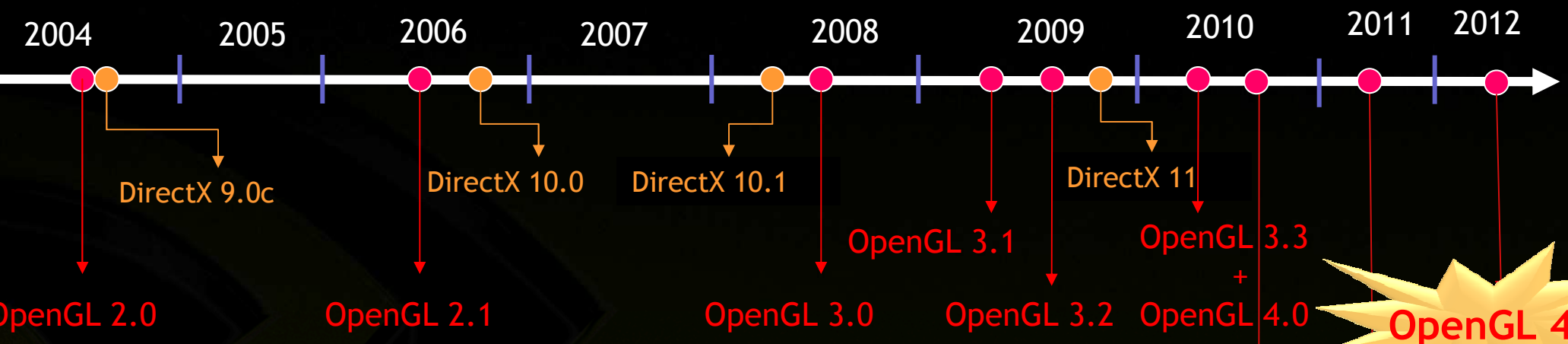
OpenGL Spawns Closely Related Standards



Congratulations: WebGL officially approved, February 201

"The web is now 3D enabled"

Accelerating OpenGL Innovation



- **OpenGL has fast innovation + standardization**
  - Pace is 7 new spec versions in four years
  - Actual implementations following specifications closely
- **OpenGL 4.3 is a superset of DirectX 11 functionality**
  - While retaining backwards compatibility

OpenGL Today – DirectX 11 Superset

- First-class graphics + compute solution
  - OpenGL 4.3 = graphics + compute shaders
  - NVIDIA still has existing inter-op with CUDA / OpenCL

Buffer *and* Event Interop

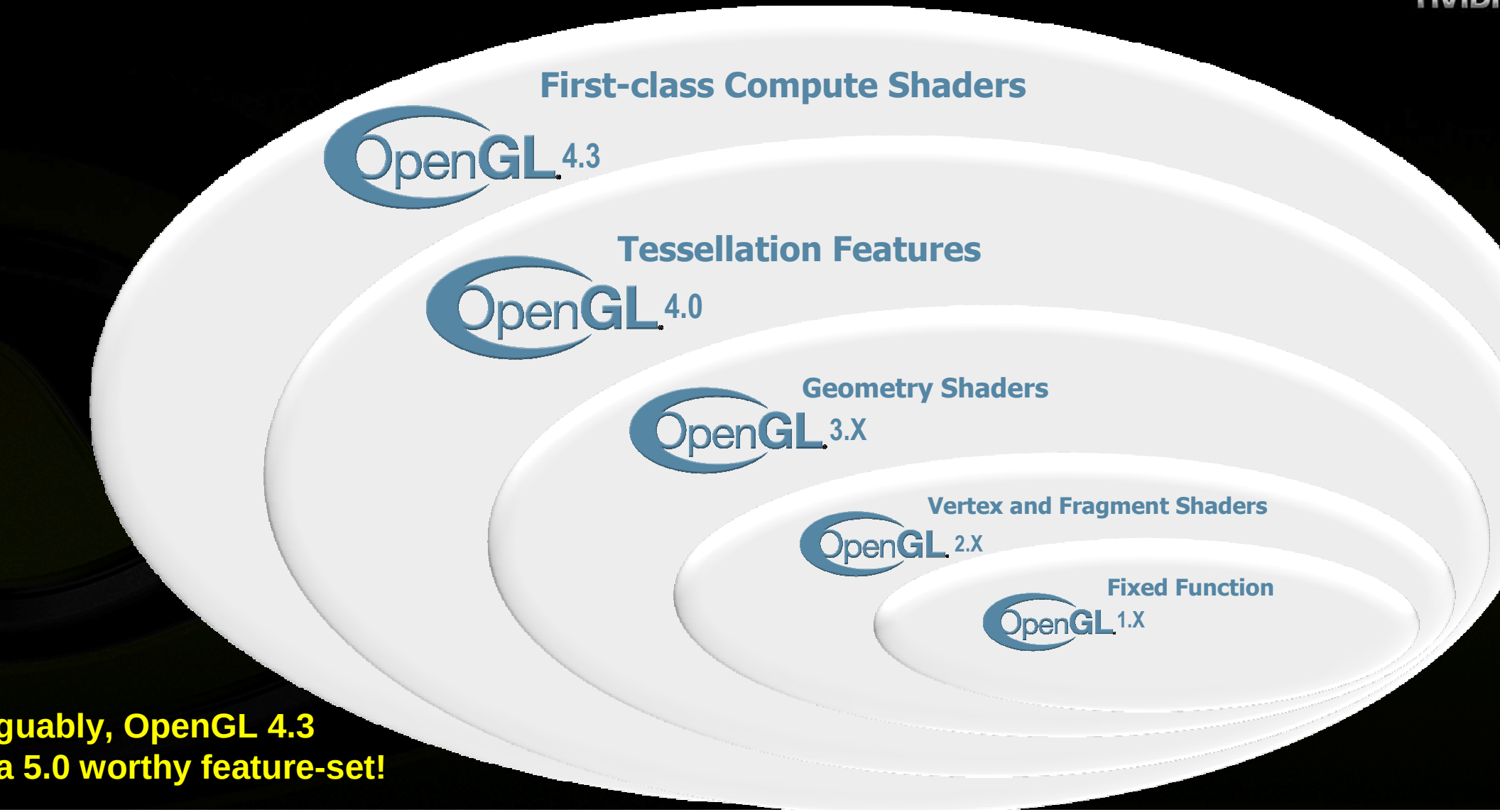
NVID
CUD

- Shaders can be saved to and loaded from binary blobs
  - Ability to query a binary shader, and save it for reuse later

- Flow of content between desktop and mobile
  - Brings ES 2.0 and 3.0 API and capabilities to desktop
  - WebGL bridging desktop and mobile

- Cross platform
  - Mac, Windows, Linux, Android, Solaris, FreeBSD
  - Result of being an open standard

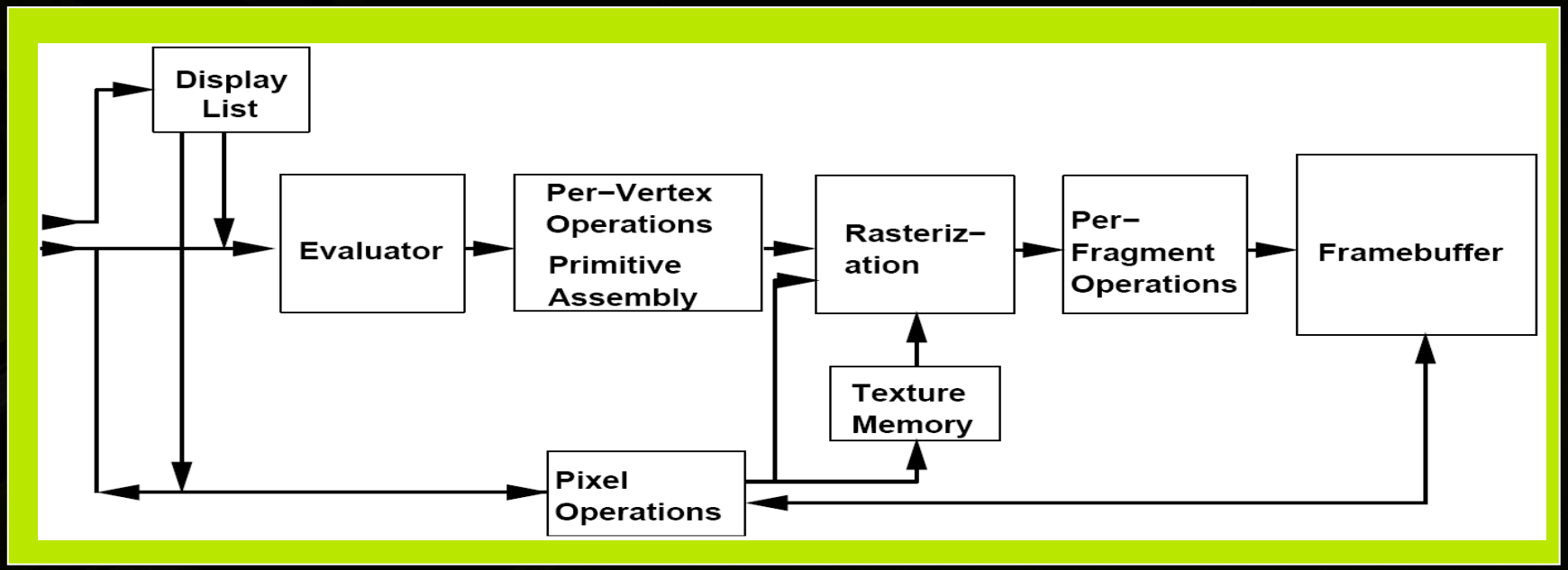
Increasing Functional Scope of OpenGL



guably, OpenGL 4.3
a 5.0 worthy feature-set!

Classic OpenGL State Machine

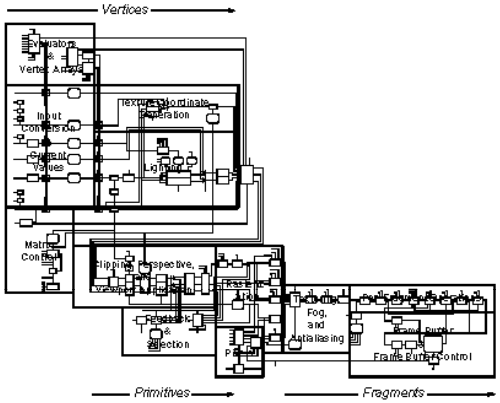
- From 1991-2007
  - * vertex & fragment processing got programmable 2001 & 2003



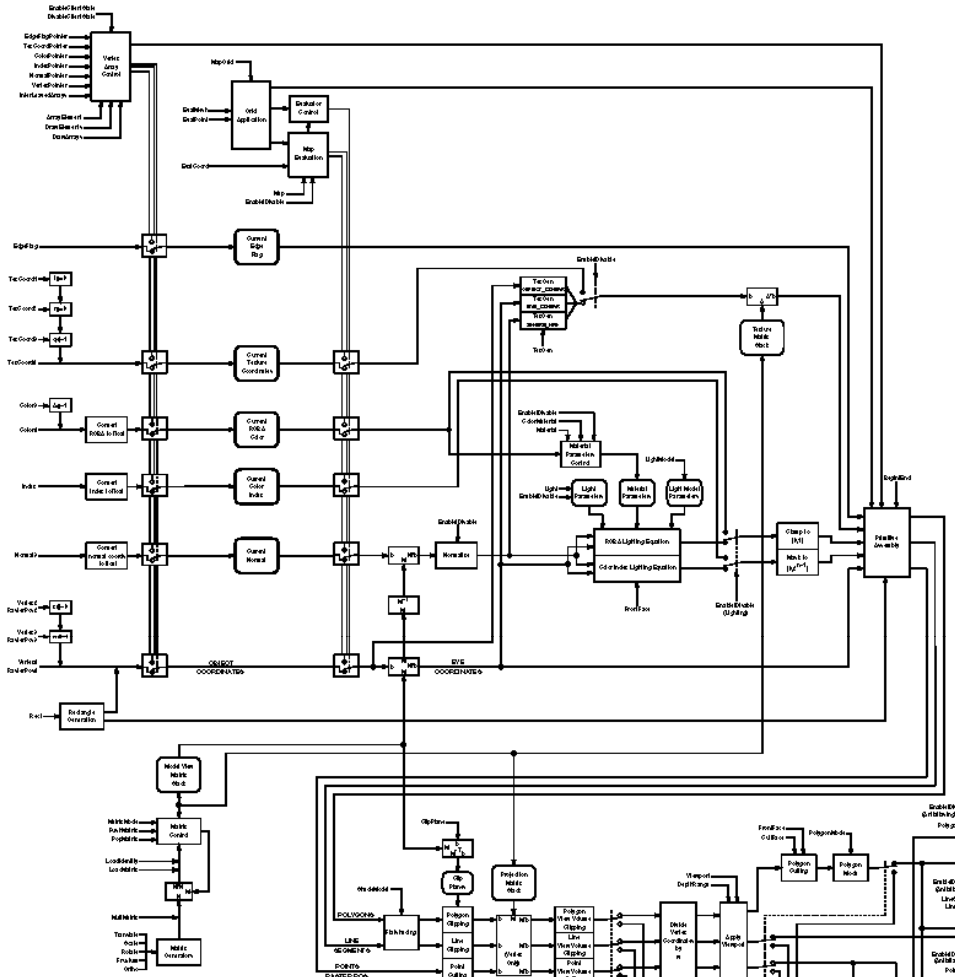
[source: GL 1.0 specification, 1992]

The OpenGL® Machine

Complicated from inception



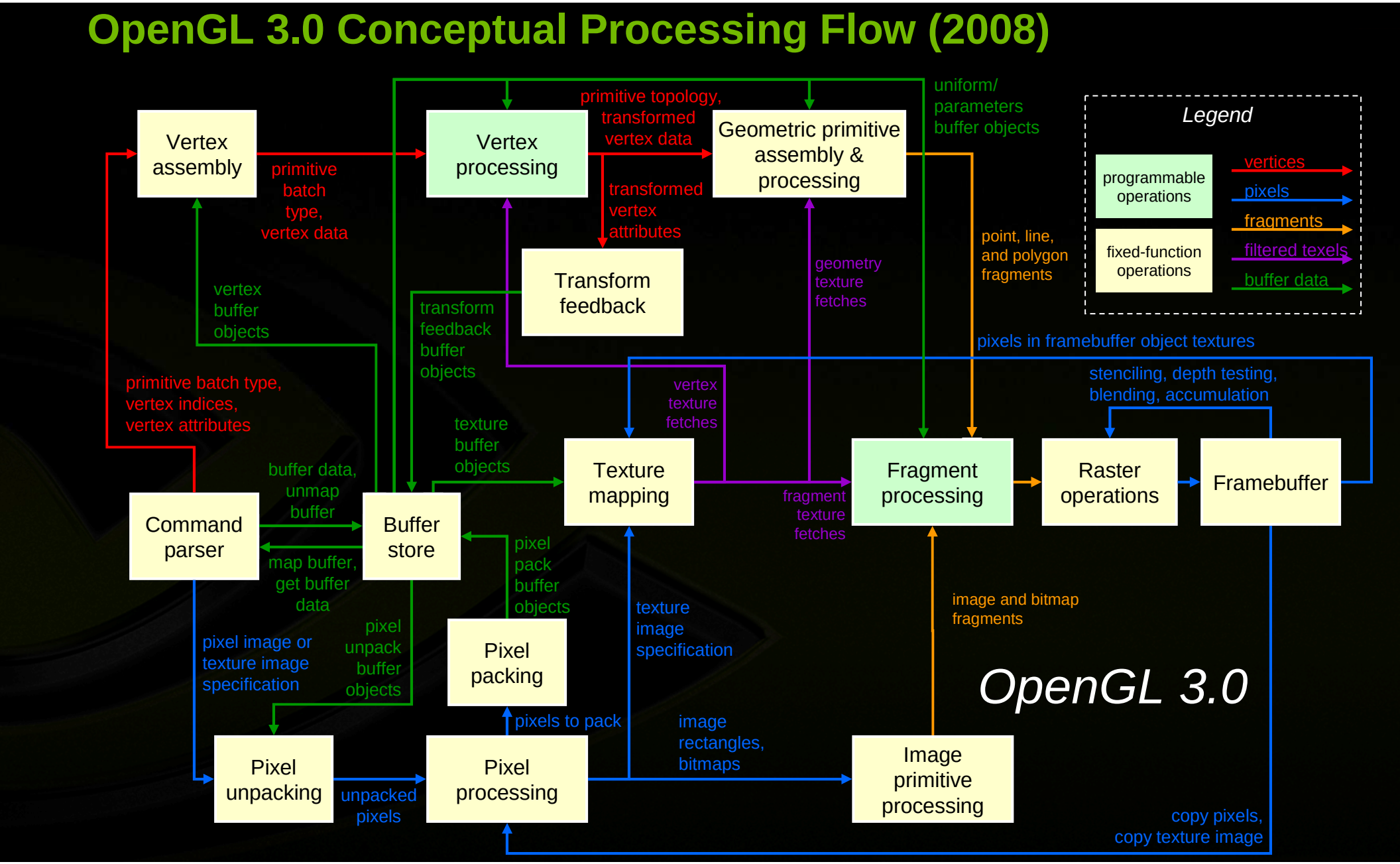
Key to OpenGL Operations

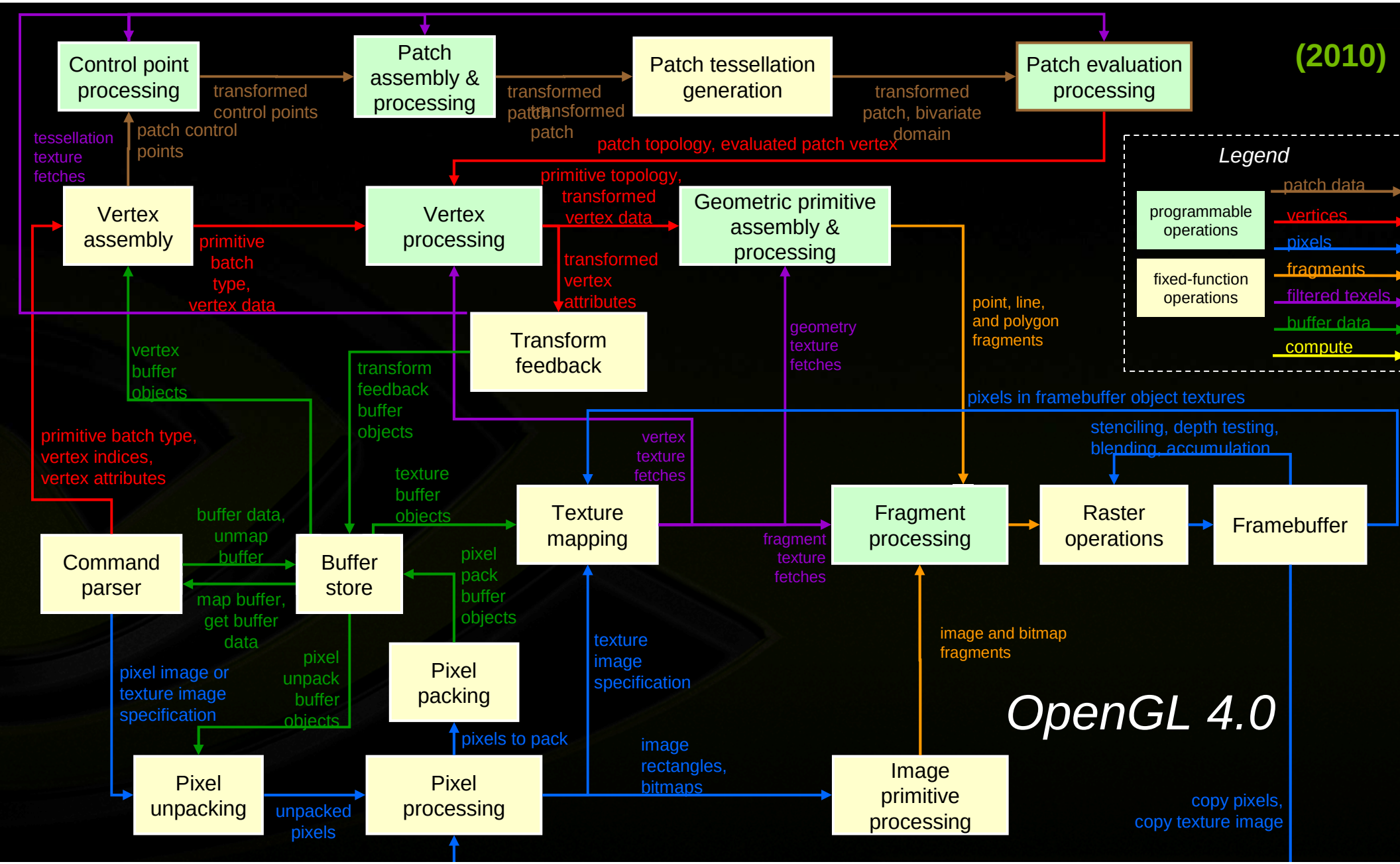


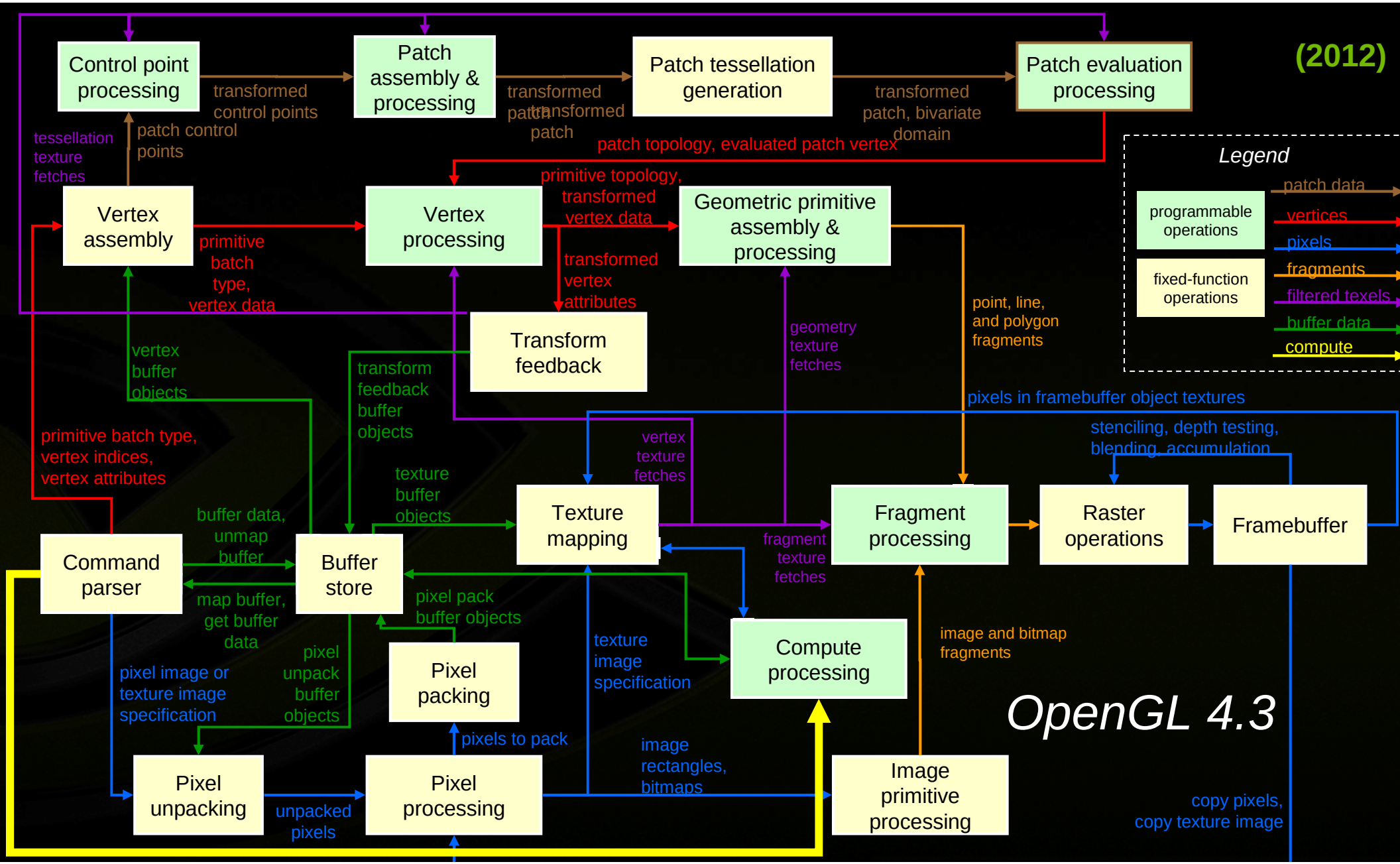
Notes:

1. Commands (and constants) are shown without the gl (or GL_) prefix.
2. The following commands do not appear in this diagram: glAccum, glClearAccum, glHint, display list commands, texture object commands, commands for obtaining OpenGL state (glGet commands and glIsEnabled), and glPushAttrib and glPopAttrib. Utility library routines are not shown.
3. After their execution, glDrawArrays and glDrawElements leave affected current values indeterminate.
4. This diagram is schematic; it may not directly correspond to any actual OpenGL implementation.

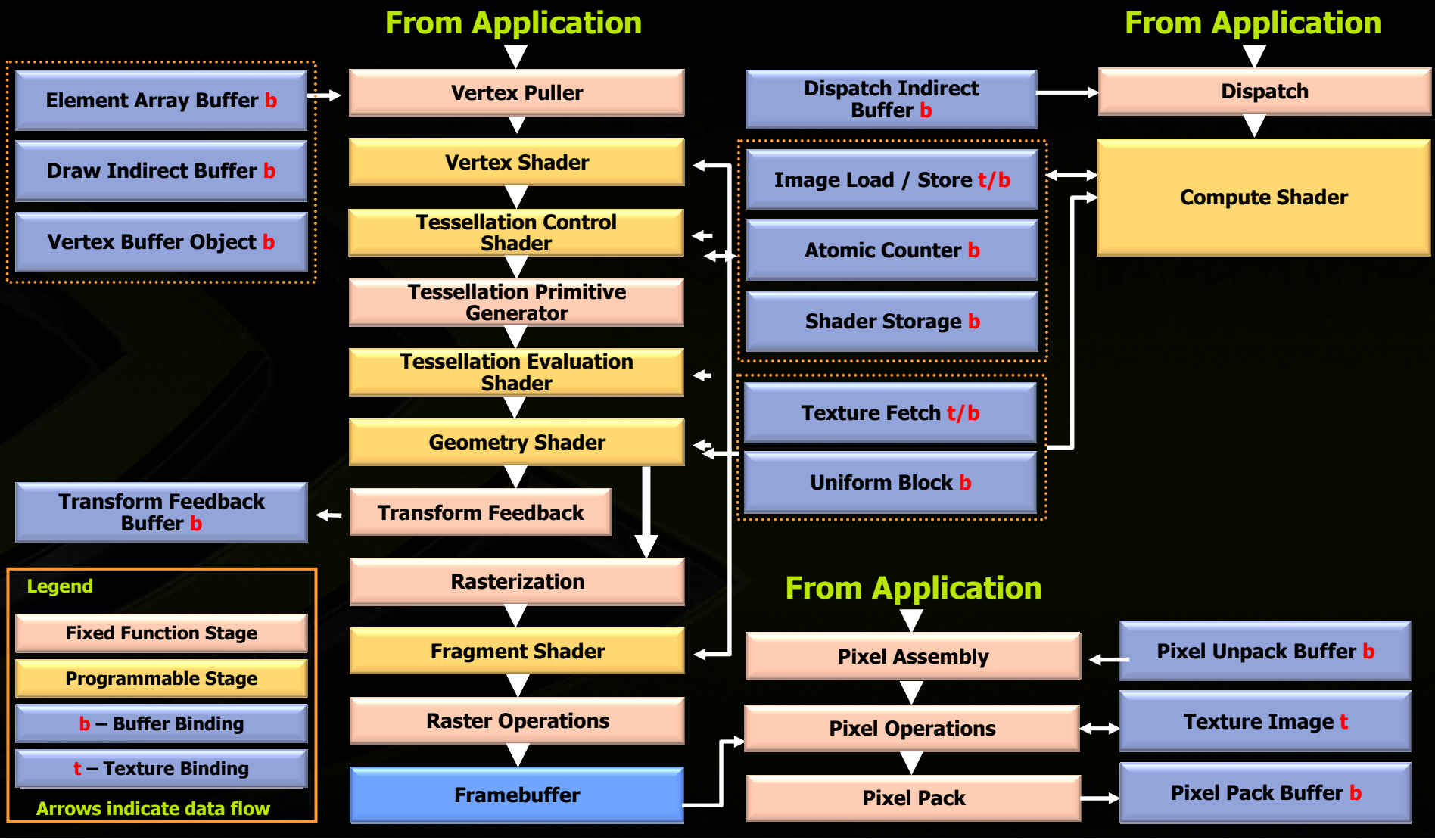
OpenGL 3.0 Conceptual Processing Flow (2008)



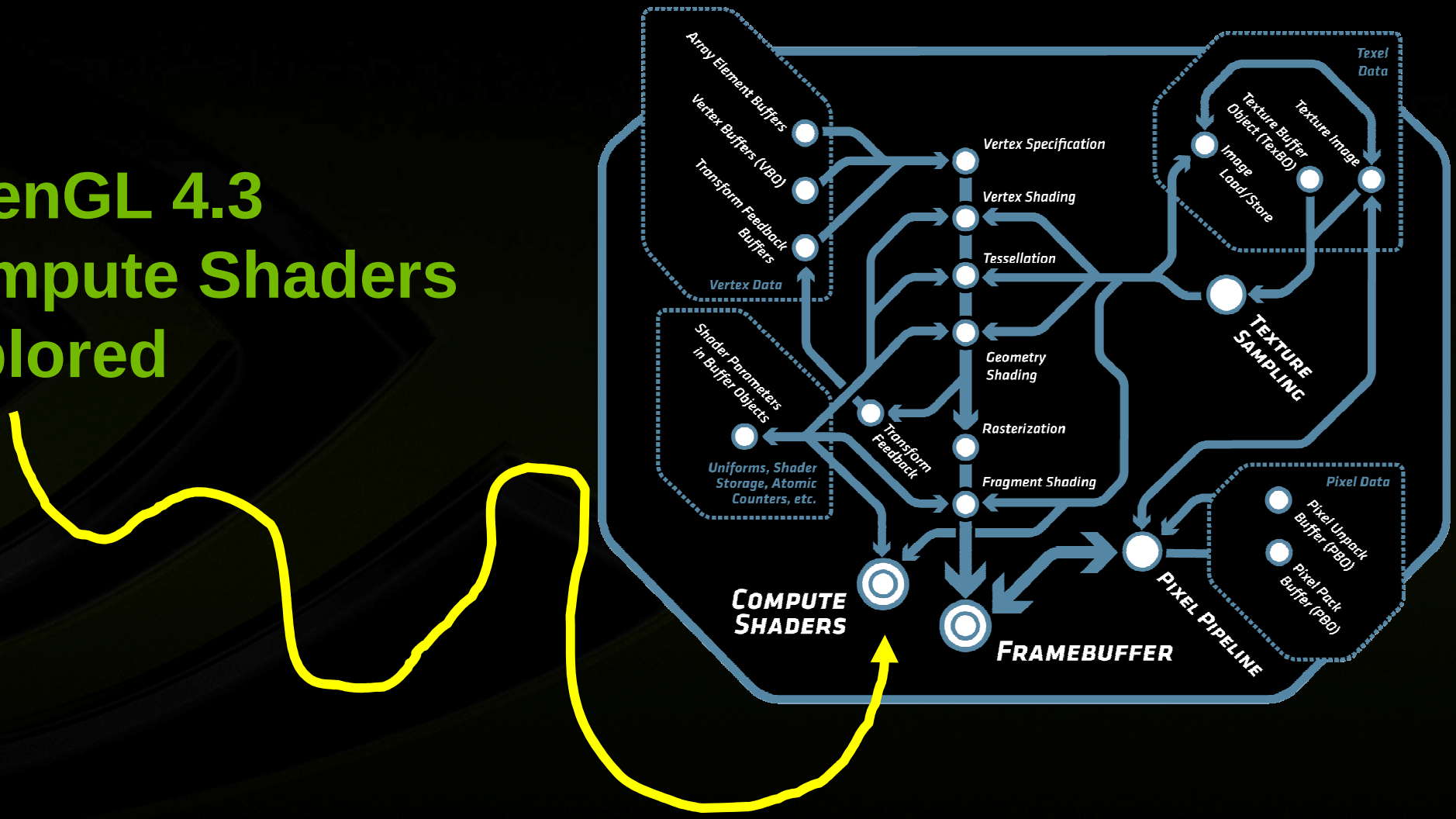

(2010)
Control point processing
transformed control points
Patch assembly & processing
transformed patch
transformed patch
Patch tessellation generation
transformed patch, bivariate domain
Patch evaluation processing
tessellation texture fetches
patch control points
patch topology, evaluated patch vertex
Legend
patch data
vertices
pixels
fragments
filtered texels
buffer data
compute
programmable operations
fixed-function operations
Vertex assembly
primitive batch type, vertex data
Vertex processing
primitive topology, transformed vertex data
transformed vertex attributes
Geometric primitive assembly & processing
point, line, and polygon fragments
Transform feedback
geometry texture fetches
vertex buffer objects
transform feedback buffer objects
pixels in framebuffer object textures
primitive batch type, vertex indices, vertex attributes
vertex texture fetches
stenciling, depth testing, blending, accumulation
texture buffer objects
Texture mapping
Fragment processing
Raster operations
Framebuffer
buffer data, unmap buffer
map buffer, get buffer data
Command parser
Buffer store
pixel pack buffer objects
fragment texture fetches
Pixel packing
texture image specification
image and bitmap fragments
pixel image or texture image specification
pixel unpack buffer objects
OpenGL 4.0
pixels to pack
image rectangles, bitmaps
Pixel unpacking
unpacked pixels
Pixel processing
Image primitive processing
copy pixels, copy texture image


(2012)
Control point processing
transformed control points
Patch assembly & processing
transformed patch
transformed patch
Patch tessellation generation
transformed patch, bivariate domain
Patch evaluation processing
tessellation texture fetches
patch control points
patch topology, evaluated patch vertex
Legend
patch data
vertices
pixels
fragments
filtered texels
buffer data
compute
programmable operations
fixed-function operations
Vertex assembly
primitive batch type, vertex data
Vertex processing
primitive topology, transformed vertex data
transformed vertex attributes
Geometric primitive assembly & processing
point, line, and polygon fragments
Transform feedback
geometry texture fetches
vertex buffer objects
transform feedback buffer objects
pixels in framebuffer object textures
stenciling, depth testing, blending, accumulation
primitive batch type, vertex indices, vertex attributes
vertex texture fetches
texture buffer objects
Texture mapping
Fragment processing
Raster operations
Framebuffer
buffer data, unmap buffer
map buffer, get buffer data
Command parser
Buffer store
fragment texture fetches
pixel pack buffer objects
texture image specification
Compute processing
image and bitmap fragments
pixel image or texture image specification
pixel unpack buffer objects
Pixel packing
pixels to pack
image rectangles, bitmaps
Pixel unpacking
unpacked pixels
Pixel processing
Image primitive processing
copy pixels, copy texture image
OpenGL 4.3

OpenGL 4.3 Processing Pipelines

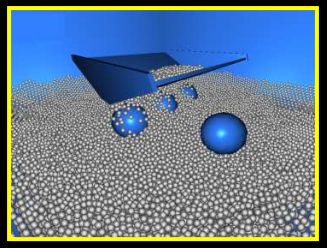


OpenGL 4.3 Compute Shaders explored



Why Compute Shaders?

- Execute algorithmically general-purpose GLSL shaders
  - Read and write uniforms and images
  - Grid-oriented Single Program, Multiple Data (SPMD) execution model with communication via shared variables
- Process graphics data in context of the graphics pipeline
  - Easier than interoperating with a compute API when processing ‘close to the pixel’
  - Avoids involved “inter-op” APIs to connect OpenGL objects to CUDA or OpenCL
- Complementary to OpenGL
  - Gives full access to OpenGL objects (multisample buffers, etc.)
  - Same GLSL language used for graphic shaders
  - Not a full heterogonous (CPU/GPU) programming framework using full ANSI C
    - In contrast to CUDA C/C++,
- Standard part of all OpenGL 4.3 implementations
  - Matches DirectX 11 functionality



particle physics

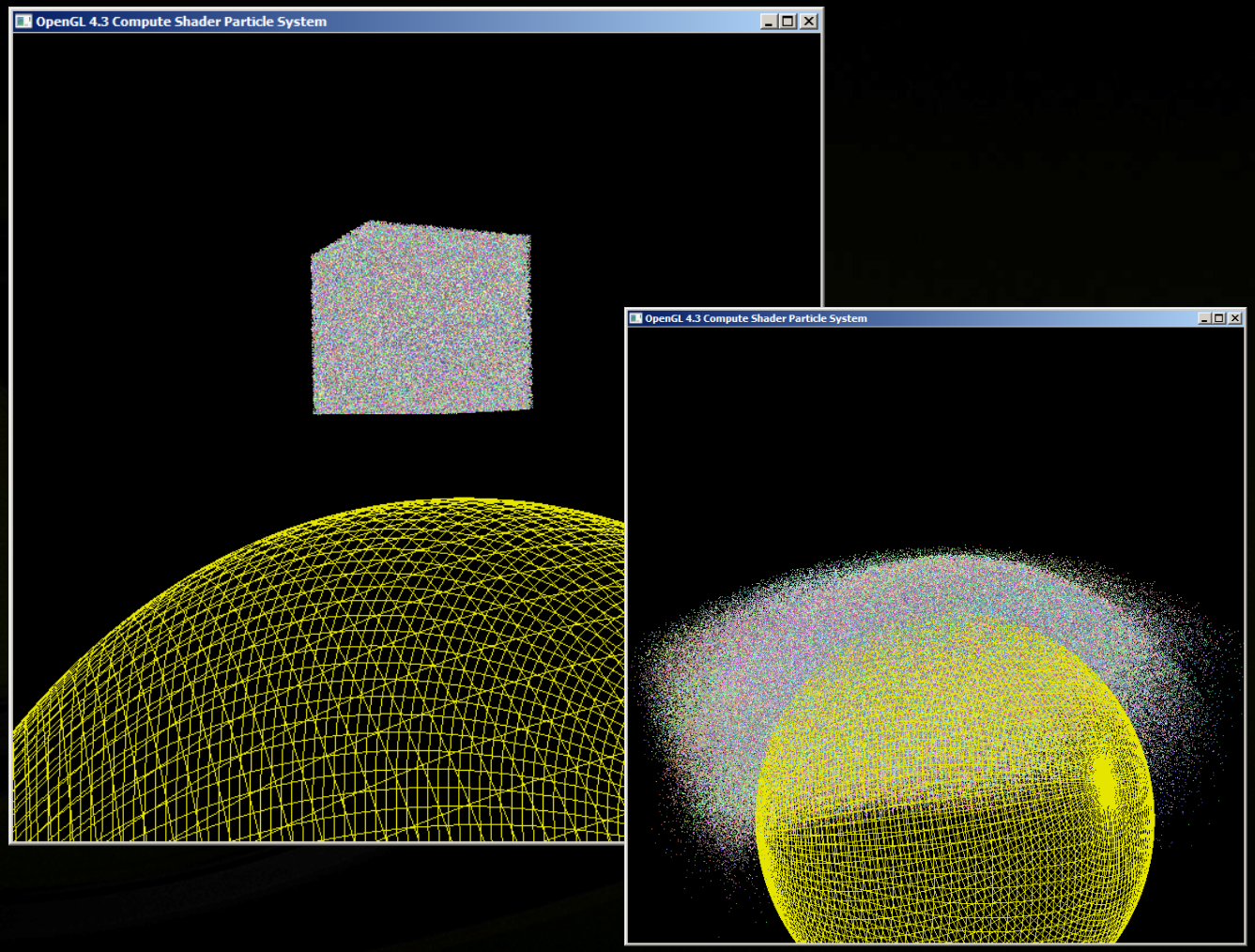
fluid behavio

crowd simulati

ray tracing

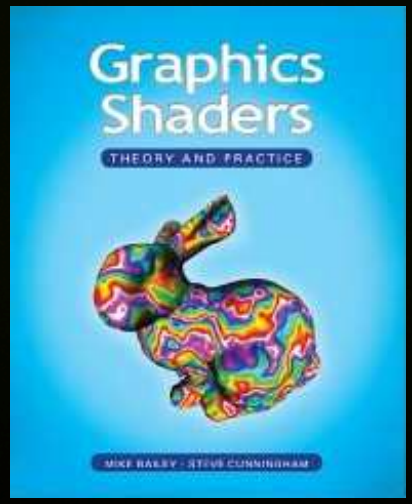
global illumina

Compute Shader Particle System Demo



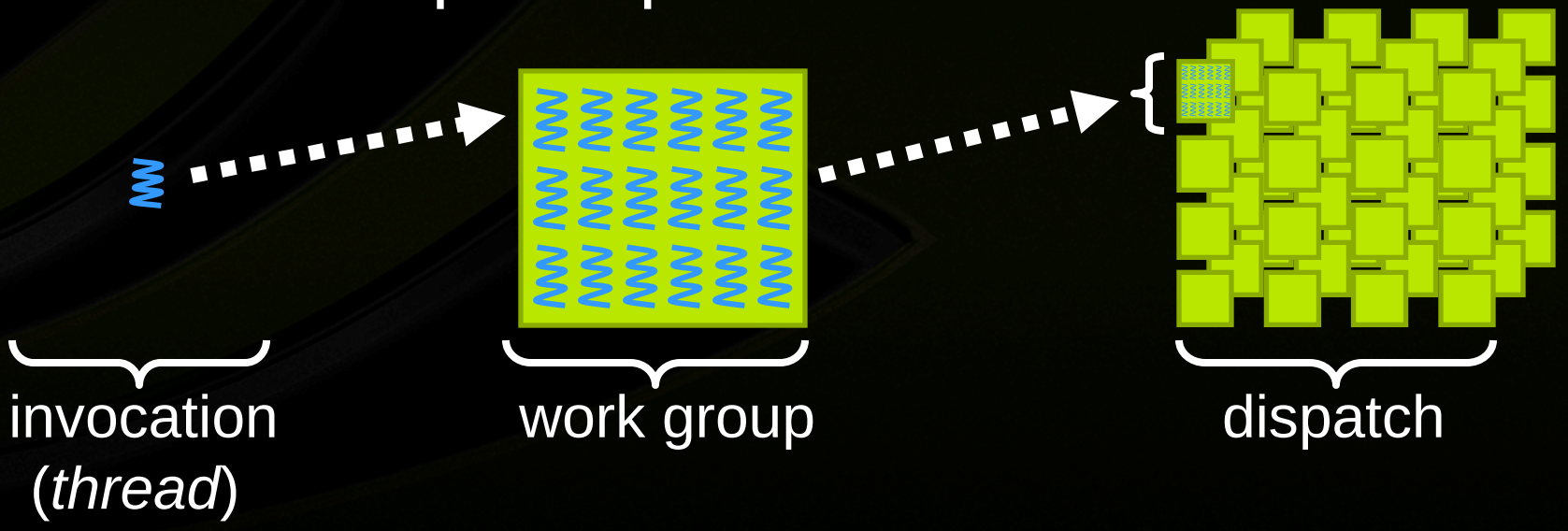
written by
Mike Bailey
@ Oregon
State Univer

also
author
of



OpenGL 4.3 Compute Shaders

- **Single Program, Multiple Data (SPMD) Execution Model**
  - **Mental model: "scores of threads jump into same function at once"**
- **Hierarchical thread structure**
  - **Threads in Groups in Dispatches**



Single Program, Multiple Data Example

Standard C Code, running single-threaded

```
void SAXPY_CPU(int n, float alpha, float x[256], float y[256])
{
    if (n > 256) n = 256;
    for (int i = 0; i < n; i++)  // loop over each element explicitly
        y[i] = alpha*x[i] + y[i];
}
```

```
#version 430
layout(local_size_x=256) in;  // spawn groups of 256 threads!
buffer xBuffer { float x[]; };  buffer yBuffer { float y[]; };
uniform float alpha;
void main()
{
    int i = int(gl_GlobalInvocationID.x);
    if (i < x.length())      // derive size from buffer bound
        y[i] = alpha*x[i] + y[i];
}
```

OpenGL Compute Shader, running SPMD

SAXPY = BLAS library's
single-precision alpha times x plus

Examples of Single-threaded Execution vs. SPMD Programming Systems

Single-threaded

C/C++

FORTRAN

Pascal

CPU-centric,
hard to make multi-threaded & parallel

Single Program, Multiple Data

CUDA C/C++

DirectCompute

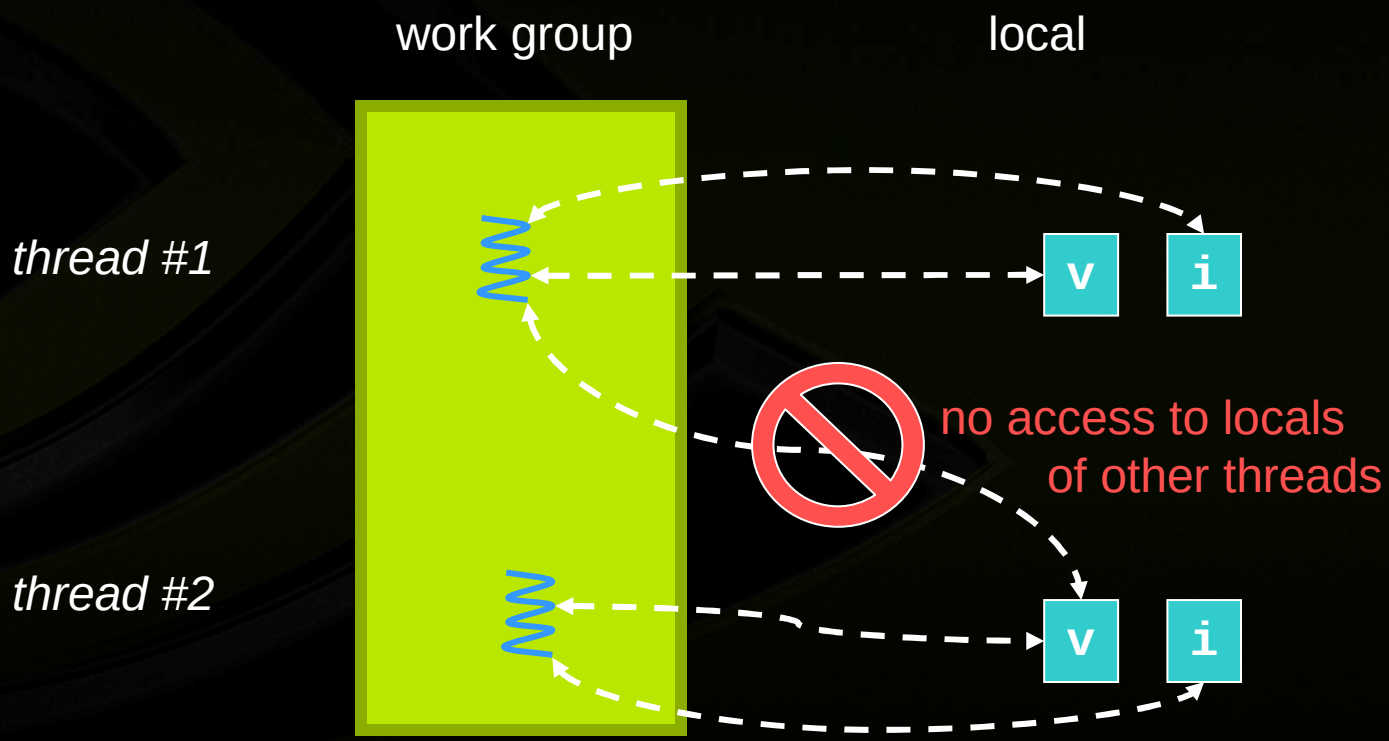
OpenCL

OpenGL Compute Shaders

GPU-centric,
naturally expresses parallelism

Per-Thread Local Variables

- **Each thread can read/write variables “private” to its execution**
  - **Each thread gets its own unique storage for each local variable**



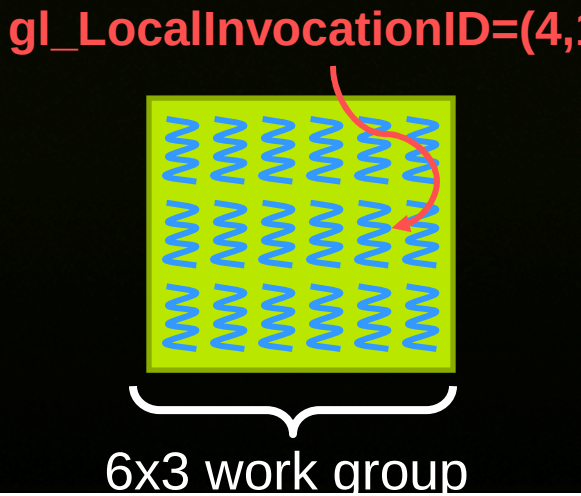
Compute Shader
source code

```
int i;
float v;

i++;
v = 2*v + i;
```

Special Per-thread Variables

- Work group can have a 1D, 2D or 3D “shape”
  - Specified via Compute Shader input declarations
  - Compute Shader syntax examples
    - 1D, 256 threads: `layout(local_size_x=256) in;`
      2D, 8x10 thread shape: `layout(local_size_x=8,local_size_y=10) in;`
    - 3D, 4x4x4 thread shape:
      `layout(local_size_x=4,local_size_y=4,local_size_z=4) in;`

- Every thread in work group has its own invocation #
  - Accessed through built-in variable
    `in uvec3 gl_LocalInvocationID;`
  - Think of every thread having a “who am I?” variable
  - Using these variables, threads are expected to
    - Index arrays
    - Determine their flow control
    - Compute thread-dependent computations



Per-Work Group Shared Variables

- Any thread in a work group can read/write shared variables
  - Typical idiom is to index by each thread's invocation #

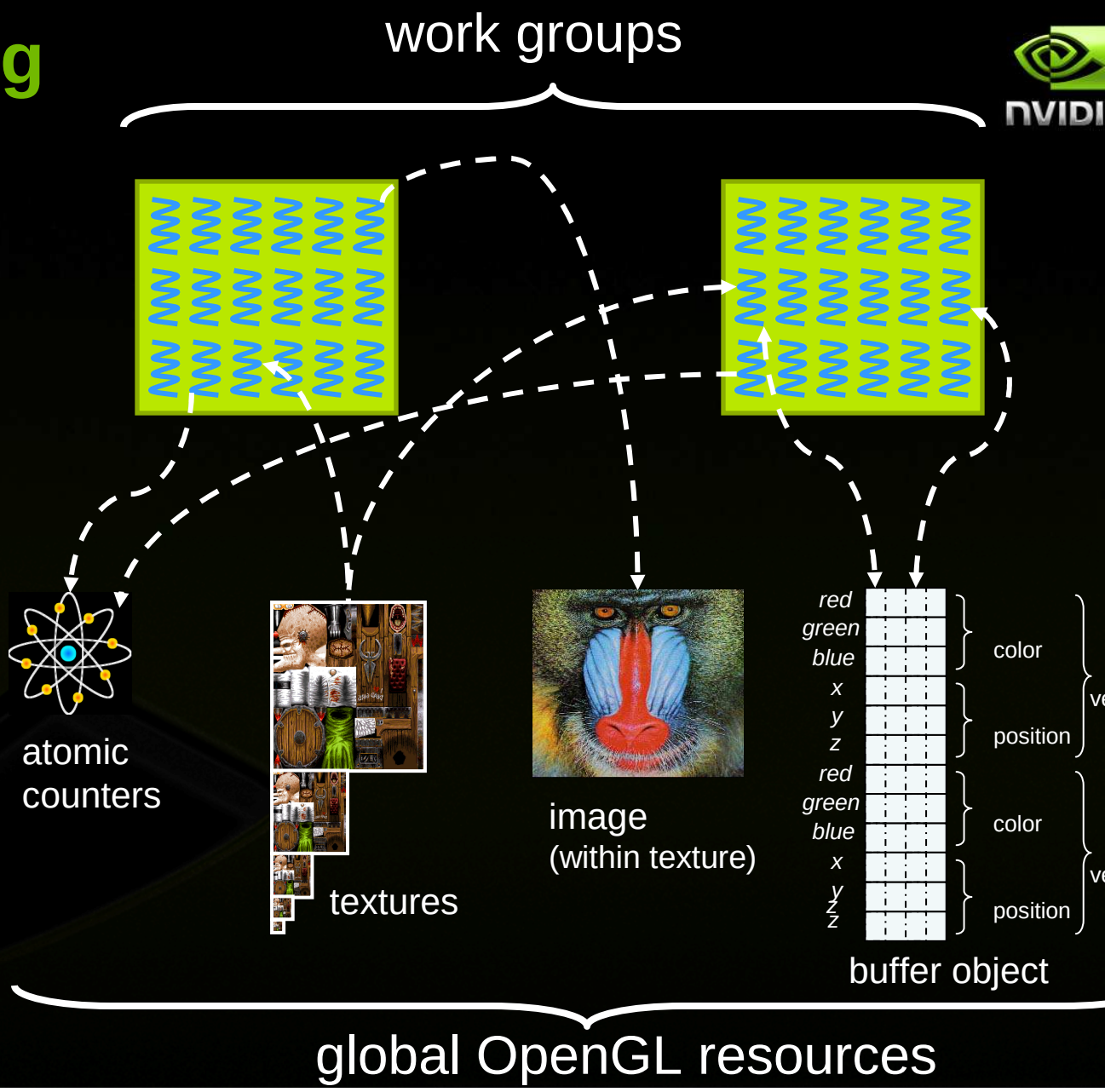
Compute Shader
source code

```
shared float a[3][4];
unsigned int x =
  gl_LocalInvocationID.
unsigned int y =
  gl_LocalInvocationID.

a[y][x] = 2*ndx;
a[y][x^1] += a[y][x];
memoryBarrierShared();
a[y][x^2] += a[y][x];
```

use shared memory barrie
to synchronize access to
shared variables

Reading and Writing Global Resources

- *In addition to local and shared variables…*
- **Compute Shaders can also access global resources**
  - **Read-only**
    - **Textures**
    - **Uniform buffer objects**
  - **Read-write**
    - **Texture images**
    - **Uniform buffers**
    - **Shader storage buffers**
    - **Atomic counters**
    - **Bindless buffers**
    - **Take care updating shared read-write resources**



Simple Compute Shader

- Let's just copy from one 2D texture image to another…

Pseudo-code:

for each pixel in source image
*** copy pixel to destination image***

✓ pixels could be copied fully in parallel

How would we write this as a compute shader...

Simple Compute Shader

- Let's just copy from one 2D texture image to another…

```
#version 430  // use OpenGL 4.3's GLSL with Compute Shaders
#define TILE_WIDTH 16
#define TILE_HEIGHT 16
const ivec2 tileSize = ivec2(TILE_WIDTH,TILE_HEIGHT);

layout(binding=0,rgba8) uniform image2D input_image;
layout(binding=1,rgba8) uniform image2D output_image;

layout(local_size_x=TILE_WIDTH,local_size_y=TILE_HEIGHT) in;

void main() {
  const ivec2 tile_xy = ivec2(gl_WorkGroupID);
  const ivec2 thread_xy = ivec2(gl_LocalInvocationID);
  const ivec2 pixel_xy = tile_xy*tileSize + thread_xy;

  vec4 pixel = imageLoad(input_image, pixel_xy);
  imageStore(output_image, pixel_xy, pixel);
}
```

Compiles into NV_compute_program5 Assembly

```
!!NVcp5.0 # NV_compute_program5 assembly
GROUP_SIZE 16 16; # work group is 16x16 so 256 threads
PARAM c[2] = { program.local[0..1] }; # internal constants
TEMP R0, R1; # temporaries
IMAGE images[] = { image[0..7] }; # input & output images
MAD.S R1.xy,invocation.groupid,{16,16,0,0}.x,invocation.localid;
MOV.S R0.x, c[0];
LOADIM.U32 R0.x, R1, images[R0.x], 2D; # load from input image
MOV.S R1.z, c[1].x;
UP4UB.F R0, R0.x;  # unpack RGBA pixel into float4 vector
STOREIM.F images[R1.z], R0, R1, 2D; # store to output image
END
```

What is NV_compute_program5?

- NVIDIA has always provided assembly-level interfaces to GPU programmability in OpenGL
  - NV_gpu_program5 is Shader Model 5.0 assembly
    - And NV_gpu_program4 was for Shader Model 4.0
  - NV_tessellation_program5 is programmable tessellation extension
  - NV_compute_program5 is further extension for Compute Shaders
- Advantages of assembly extensions
  - Faster load-time for shaders
  - Easier target for dynamic shader generation
    - Allows other languages/tools, such as Cg, to target the underlying hardware
  - Provides concrete underlying execution model
    - You don’t have to guess if your GLSL compiles well or not

Launching a Compute Shader

- First write your compute shader
  - Request GLSL 4.30 in your source code: `#version 430`
  - The text of our “copy” kernel is an example
- Second compile your compute shader
  - Same compilation process as standard GLSL graphics shaders…
  - glCreateShader/glShaderSource with Compute Shader token

```
GLuint compute_shader = glCreateShader(GL_COMPUTE_SHADER);
```

  - glCreateProgram/glAttachShader/glLinkProgram
    - (compute and graphics shaders cannot mix in the same program)

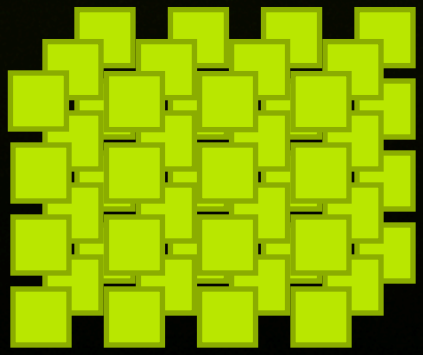
- Bind to your program object

```
glUseProgram(compute_shader);
```

- Dispatch a grid of work groups

```
glDispatchCompute(4, 4, 3);
```

dispatches a
4x4x3 grid
of work groups



Launching the Copy Compute Shader

- Setup for copying from source to destination texture
  - Create an input (*source*) texture object

```
glTextureStorage2DEXT(input_texobj, GL_TEXTURE_2D,
    1, GL_RGBA8, width, height);
glTextureSubImage2DEXT(input_texobj, GL_TEXTURE_2D,
    /*level*/0, /*x,y*/0,0, width, height,
    GL_RGBA, GL_UNSIGNED_BYTE, image);
```

  - Create an empty output (*destination*) texture object

```
glTextureStorage2DEXT(output_texobj, GL_TEXTURE_2D,
    1, GL_RGBA8, width, height);
```

OpenGL 4.2 or ARB_texture-_storage plus EXT_direct_state_ac

  - Bind level zero of both textures to texture images 0 and 1

```
GLboolean is_not_layered = GL_FALSE;
glBindImageTexture(/*image*/0, input_texobj, /*level*/0,
    is_not_layered, /*layer*/0, GL_READ_ONLY, GL_RGBA8);
glBindImageTexture( /*image*/1, output_texobj, /*level*/0,
    is_not_layered, /*layer*/0, GL_READ_WRITE, GL_RGBA8);
```

OpenGL 4 ARB_sha _image- _load_sto

  - Use the copy compute shader

```
glUseProgram(compute_shader);
```

- Dispatch sufficient work group instances of the copy compute shader

```
glDispatchCompute((width + 15) / 16, (height + 15) / 16), 1);
```

OpenG

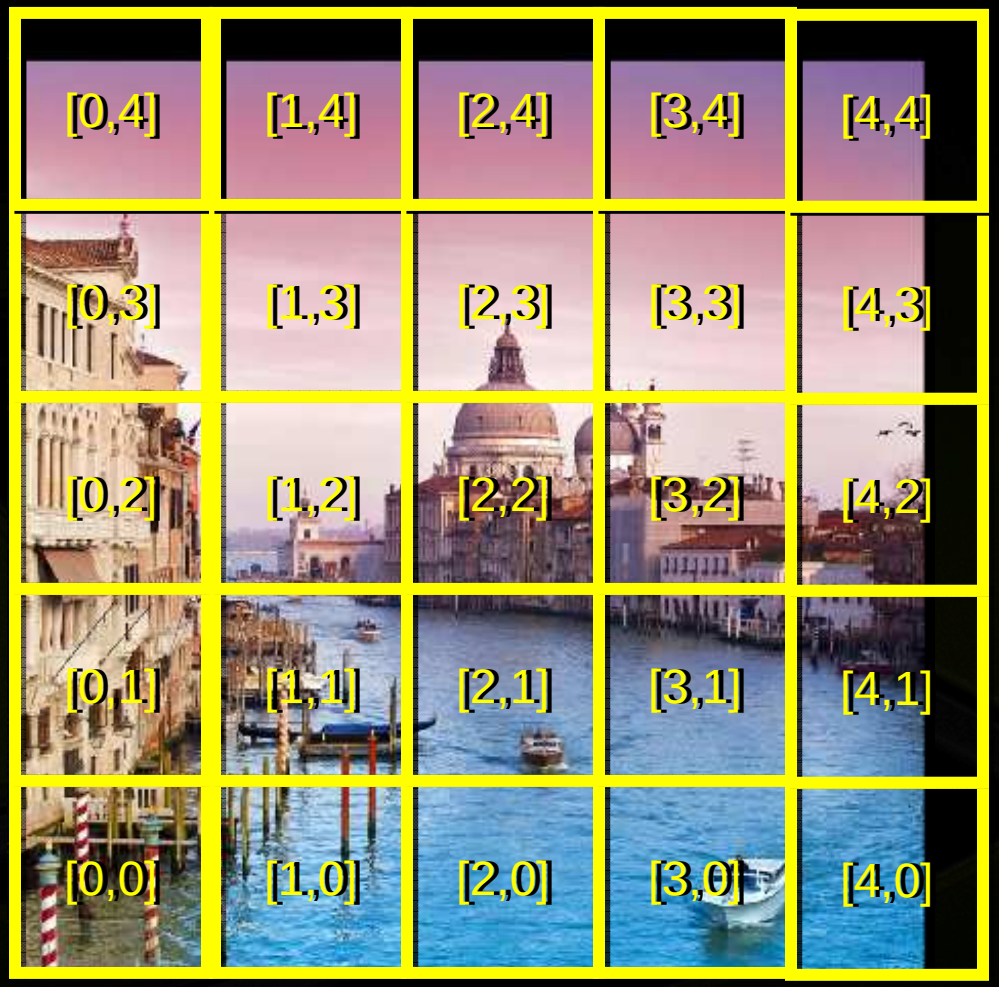
Copy Compute Shader Execution

Input (*source*) image

Output (*destination*) image

Copy Compute Shader Tiling

gl_WorkGroupID=[x,y]



Input (*source*) image 76x76

Output (*destination*) image 76x76

Next Example: General Convolution

- Discrete convolution: common image processing operation
  - Building block for blurs, sharpening, edge detection, etc.
- Example: 5x5 convolution (N=5) of source (input) image s
  - Generates destination (output) image d, given NxN matrix of weights w

$$d_{x,y} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} \, s_{x+i-\left\lfloor \frac{N}{2} \right\rfloor, y+j-\left\lfloor \frac{N}{2} \right\rfloor}$$

Input Image

Output Image after 5x5 Gaussian Blur

Gaussi
sigma=

Implementing a General Convolution

- Basic algorithm
  - Tile-oriented: generate MxM pixel tiles
  - So operating on a (M+2n)x(M+2n) region of the image
    - For 5x5 convolution, n=2
- Phase 1: Read all the pixels for a region from input image
- Phase 2: Perform weighted sum of pixels in [-n,n]x[-n,n] region around each output pixel
- Phase 3: Output the result pixel to output image

General Convolution: Preliminaries

```
#version 430  // use OpenGL 4.3's GLSL with Compute Shaders
// Various kernel-wide constants
const int tileWidth = 16,
          tileHeight = 16;
const int filterWidth = 5,
          filterHeight = 5;
const ivec2 tileSize = ivec2(tileWidth,tileHeight);
const ivec2 filterOffset = ivec2(filterWidth/2,filterHeight/2);
const ivec2 neighborhoodSize = tileSize + 2*filterOffset;

// Declare the input and output images.
layout(binding=0,rgba8) uniform image2D input_image;
layout(binding=1,rgba8) uniform image2D output_image;

uniform vec4 weight[filterHeight][filterWidth];

uniform ivec4 imageBounds;  // Bounds of the input image for pixel coordinate clamping.

void retirePhase() { memoryBarrierShared(); barrier(); }

ivec2 clampLocation(ivec2 xy) {
    // Clamp the image pixel location to the image boundary.
    return clamp(xy, imageBounds.xy, imageBounds.zw);
}
```

General Convolution: Phase 1

```
layout(local_size_x=TILE_WIDTH,local_size_y=TILE_HEIGHT) in;

shared vec4 pixel[NEIGHBORHOOD_HEIGHT][NEIGHBORHOOD_WIDTH];

void main() {
    const ivec2 tile_xy = ivec2(gl_WorkGroupID);
    const ivec2 thread_xy = ivec2(gl_LocalInvocationID);
    const ivec2 pixel_xy = tile_xy*tileSize + thread_xy;
    const uint x = thread_xy.x;
    const uint y = thread_xy.y;

    // Phase 1: Read the image's neighborhood into shared pixel arrays.
    for (int j=0; j<neighborhoodSize.y; j += tileHeight) {
        for (int i=0; i<neighborhoodSize.x; i += tileWidth) {
            if (x+i < neighborhoodSize.x && y+j < neighborhoodSize.y) {
                const ivec2 read_at = clampLocation(pixel_xy+ivec2(i,j)-filterOffset);
                pixel[y+j][x+i] = imageLoad(input_image, read_at);
            }
        }
    }
    retirePhase();
```

General Convolution: Phases 2 & 3

```
  // Phase 2: Compute general convolution.
    vec4 result = vec4(0);
    for (int j=0; j<filterHeight; j++) {
        for (int i=0; i<filterWidth; i++) {
            result += pixel[y+j][x+i] * weight[j][i];
        }
    }

    // Phase 3: Store result to output image.
    imageStore(output_image, pixel_xy, result);
}
```

Separable Convolution

- Many important convolutions expressible in “separable” form
  - More efficient to evaluate
  - Allows two step process: 1) blur rows, then 2) blur columns
  - Two sets of weights: column vector weights **c** and row vector weights **r**

$$d_{x,y} = \sum_{i=1}^{N} \sum_{j=1}^{N} \underbrace{c_i \; r_j} \; S_{x+i-\left\lfloor \frac{N}{2} \right\rfloor, y+j-\left\lfloor \frac{N}{2} \right\rfloor}$$

weight is product of separable row & column weights

- Practical example for demonstrating Compute Shader shared variables…

Example Separable Convolutions

Original

Original

Original

Example Separable Convolutions

Gaussian filter, sigma=2.25

Sobel filter, horizontal

Sobel filter, vertical

GLSL Separable Filter Implementation

<< assume preliminaries from earlier general convolution example>>

```
layout(local_size_x=TILE_WIDTH,local_size_y=NEIGHBORHOOD_HEIGHT) in;

shared vec4 pixel[NEIGHBORHOOD_HEIGHT][NEIGHBORHOOD_WIDTH]; // values read from input image
shared vec4 row[NEIGHBORHOOD_HEIGHT][TILE_WIDTH];           // weighted row sums

void main()  // separable convolution
{
  const ivec2 tile_xy = ivec2(gl_WorkGroupID);
  const ivec2 thread_xy = ivec2(gl_LocalInvocationID);
  const ivec2 pixel_xy = tile_xy*tileSize + (thread_xy-ivec2(0,filterOffset.y));
  const uint x = thread_xy.x;
  const uint y = thread_xy.y;

  // Phase 1: Read the image's neighborhood into shared pixel arrays.
  for (int i=0; i<NEIGHBORHOOD_WIDTH; i += TILE_WIDTH) {
    if (x+i < NEIGHBORHOOD_WIDTH) {
      const ivec2 read_at = clampLocation(pixel_xy+ivec2(i-filterOffset.x,0));
      pixel[y][x+i] = imageLoad(input_image, read_at);
    }
  }
  retirePhase();
```

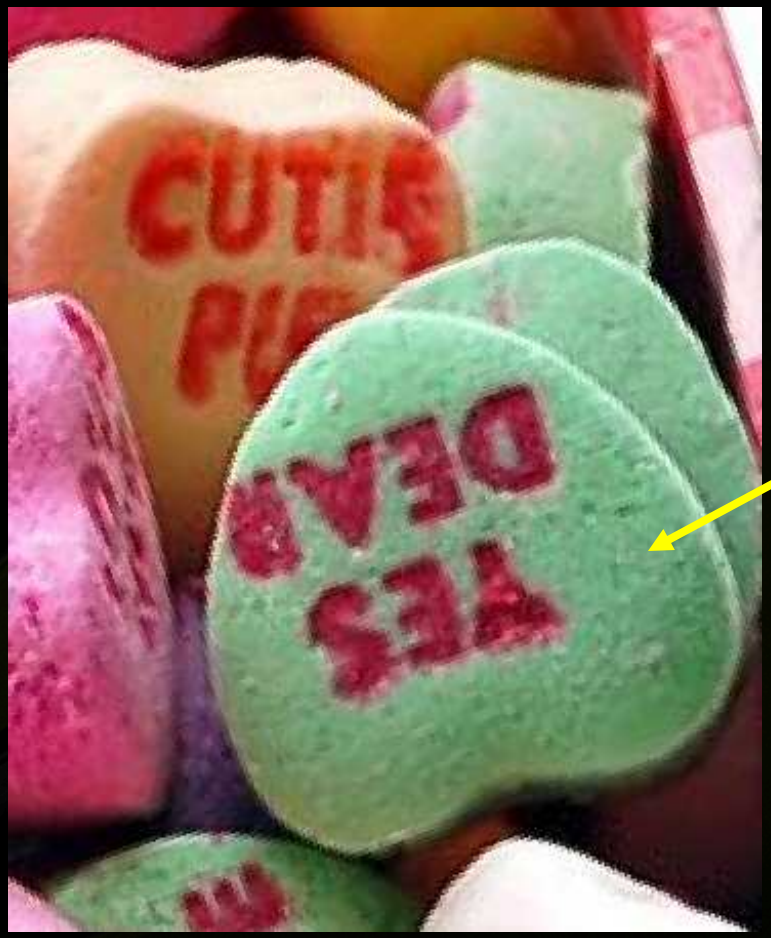
GLSL Separable Filter Implementation

```
// Phase 2: Weighted sum the rows horizontally.
row[y][x] = vec4(0);
for (int i=0; i<filterWidth; i++) {
  row[y][x] += pixel[y][x+i] * rowWeight[i];
}
retirePhase();

// Phase 3: Weighted sum the row sums vertically and write result to output image.
// Does this thread correspond to a tile pixel?
// Recall: There are more threads in the Y direction than tileHeight.
if (y < tileHeight) {
  vec4 result = vec4(0);
  for (int i=0; i<filterHeight; i++) {
    result += row[y+i][x] * columnWeight[i];
  }

  // Phase 4: Store result to output image.
  const ivec2 pixel_xy = tile_xy*tileSize + thread_xy;
  imageStore(output_image, pixel_xy, result);
}
```

Compute Shader Median Filter

- Simple idea
  - “For each pixel, replace it with the median-valued pixel in its NxN neighborhood”
  - Non-linear, good for image enhancement through noise reduction
  - Expensive: naively, requires lots sorting to find median
    - Very expensive when the neighborhood is large
- Reasonably efficient with Compute Shaders

Median Filter Example

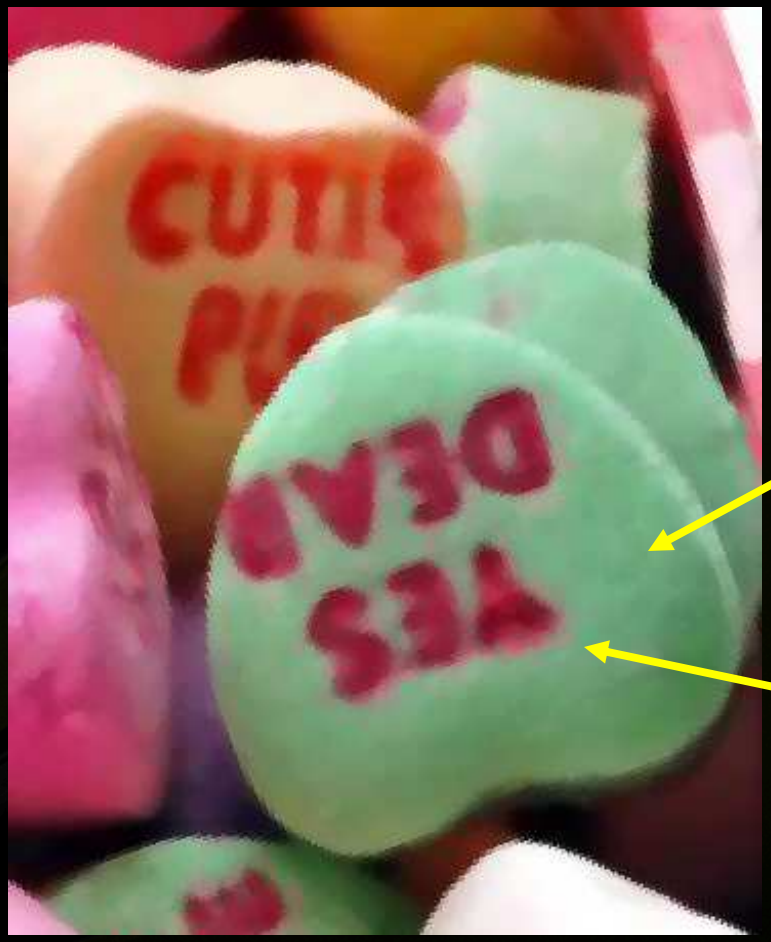


Original

Median Filter Example

Gaussian 5x5 blur

Median Filter Example



Median filter 5x5

Large Median Filters for Impressionistic Effect

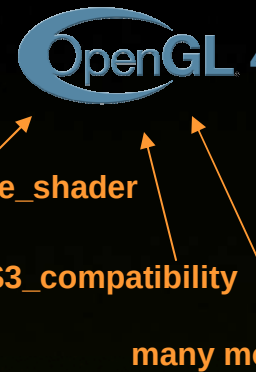
Original

7x7 Estimated Median Filter

Other stuff in OpenGL 4.3

OpenGL Evolves Modularly

- Each core revision is specified as a set of extensions
  - Example: ARB_compute_shader
    - Puts together all the functionality for compute shaders
    - Describe in its own text file
  - May have dependencies on other extensions
    - Dependencies are stated explicitly
- A core OpenGL revision (such as OpenGL 4.3) “bundles” a set of agreed extensions—and mandates their mutual support
  - Note: implementations can also “unbundle” ARB extensions for hardware unable to support the latest core revision
- So easiest to describe OpenGL 4.3 based on its bundled extensions…



OpenGL 4.3 debugging support

- **ARB_debug_output**
  - OpenGL can present debug information back to developer
- **ARB_debug_output2**
  - Easier enabling of debug output
- **ARB_debug_group**
  - Hierarchical grouping of debug tagging
- **ARB_debug_label**
  - Label OpenGL objects for debugging

OpenGL 4.3 new texture functionality

- ARB_texture_view
  - Provide different ways to interpret texture data without duplicating the texture
  - Match DX11 functionality
- ARB_internalformat_query2
  - Find out actual supported limits for most texture parameters
- ARB_copy_image
  - Direct copy of pixels between textures and render buffers
- ARB_texture_buffer_range
  - Create texture buffer object corresponding to a sub-range of a buffer's data store
- ARB_stencil_texturing
  - Read stencil bits of a packed depth-stencil texture
- ARB_texture_storage_multisample
  - Immutable storage objects for multisampled textures

OpenGL 4.3 new buffer functionality

- ARB_shader_storage_buffer_object
  - Enables shader stages to read & write to very large buffers
    - NVIDIA hardware allows every shader stage to read & write
  - structs, arrays, scalars, etc.
- ARB_invalidate_subdata
  - Invalidate all or some of the contents of textures and buffers
- ARB_clear_buffer_object
  - Clear a buffer object with a constant value
- ARB_vertex_attrib_binding
  - Separate vertex attribute state from the data stores of each array
- ARB_robust_buffer_access_behavior
  - Shader read/write to an object only allowed to data owned by the applicatio
  - Applies to out of bounds accesses

OpenGL 4.3 new pipeline functionality

- ARB_compute_shader
  - Introduces new shader stage
  - Enables advanced processing algorithms that harness the parallelism of GPUs
- ARB_multi_draw_indirect
  - Draw many GPU generated objects with one call
- ARB_program_interface_query
  - Generic API to enumerate active variables and interface blocks for each stage
  - Enumerate active variables in interfaces between separable program objects
- ARB_ES3_compatibility
  - features not previously present in OpenGL
  - Brings EAC and ETC2 texture compression formats
- ARB_framebuffer_no_attachments
  - Render to an arbitrary sized framebuffer without actual populated pixels

GLSL 4.3 new functionality

- ARB_arrays_of_arrays
  - Allows multi-dimensional arrays in GLSL. `float f[4][3];`
- ARB_shader_image_size
  - Query size of an image in a shader
- ARB_explicit_uniform_location
  - Set location of a default-block uniform in the shader
- ARB_texture_query_levels
  - Query number of mipmap levels accessible through a sampler uniform
- ARB_fragment_layer_viewport
  - `gl_Layer` and `gl_ViewportIndex` now available to fragment shader

New KHR and ARB extensions

- Not part of core but important and standardized at same time as OpenGL 4.3…

- KHR_texture_compression_astc_ldr
  - Adaptive Scalable Texture Compression (ASTC)
  - 1-4 component, low bit rate < 1 bit/pixel – 8 bit/pixel
- ARB_robustness_isolation
  - If application causes GPU reset, no other application will be affected
  - For WebGL and other un-trusted 3D content sources

Getting at OpenGL 4.3

- Easiest approach…
- Use OpenGL Extension Wrangler (GLEW)
  - Release 1.9.0 already has OpenGL 4.3 support
  - http://glew.sourceforge.net

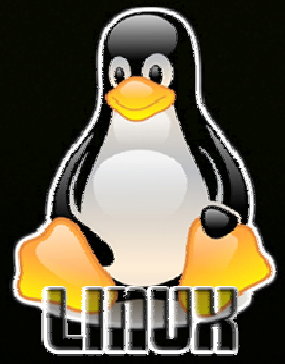
Further NVIDIA OpenGL Work

Further NVIDIA OpenGL Work

- Linux enhancements
- Path Rendering for Resolution-independent 2D graphics
- Bindless Graphics
- Commitment to API Compatibility

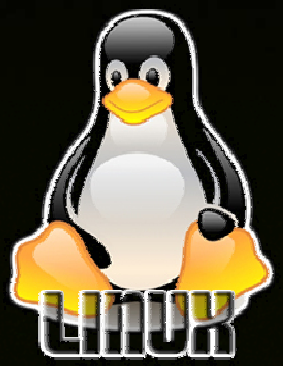
OpenGL-related Linux Improvements

- Support for X Resize, Rotate, and Reflect Extension
  - Also known as RandR
  - Version 1.2 and 1.3
- OpenGL enables, by default, “Sync to Vertical Blank”
  - Locks your glXSwapBuffers to the monitor refresh rates
  - Matches Windows default now
  - Previously disabled by default



OpenGL-related Linux Improvements

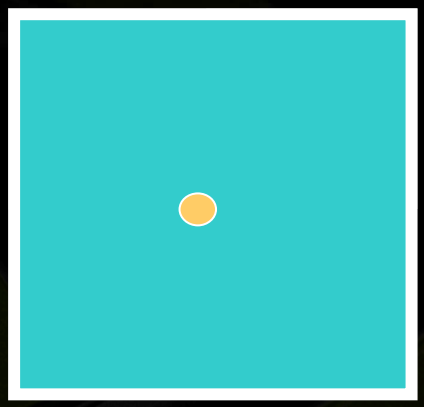
- Expose additional full-scene antialiasing (FSAA) modes
  - 16x multisample FSAA on all GeForce GPUs
    - 2x2 supersampling of 4x multisampling
  - Ultra high-quality FSAA modes for Quadro GPUs
    - 32x multisample FSAA
      - 2x2 supersampling of 8x multisampling
    - 64x multisample FSAA
      - 4x4 supersampling of 4x multisampling



- Coverage sample FSAA on GeForce 8 series and better
  - 4 color/depth samples + 12 depth samples

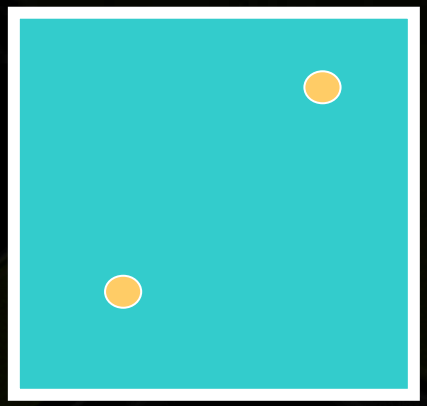
Multisampling FSAA Patterns

aliased
1 sample/pixel



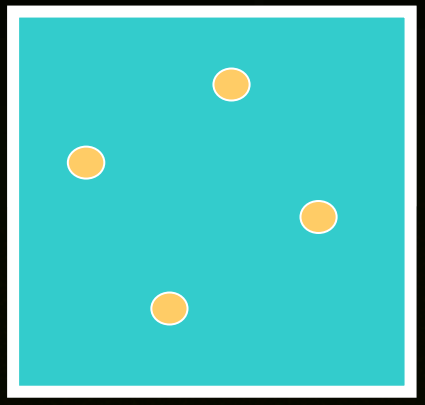
64 bits/pixel

2x multisampling
2 samples/pixel



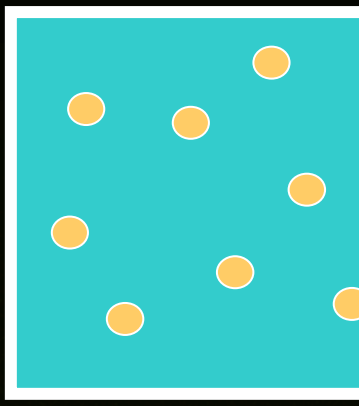
128 bits/pixel

4x multisampling
4 samples/pixel



256 bits/pixel

8x multisamp
8 samples/pi

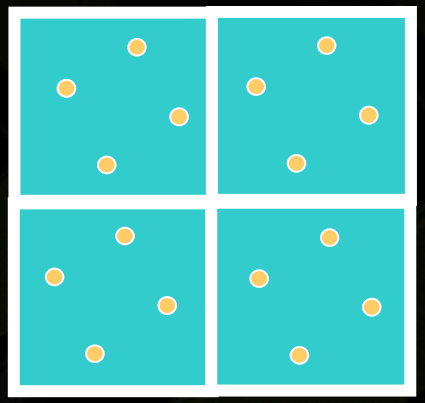


512 bits/pixe

Assume: 32-bit RGBA + 24-bit Z + 8-bit Stencil = 64 bits/sample

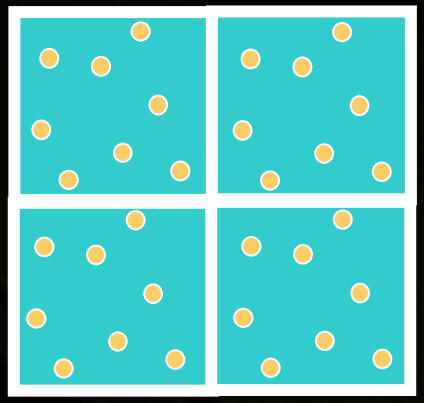
Supersampling FSAA Patterns

2x2 supersampling
of 4x multisampling
16 samples/pixel



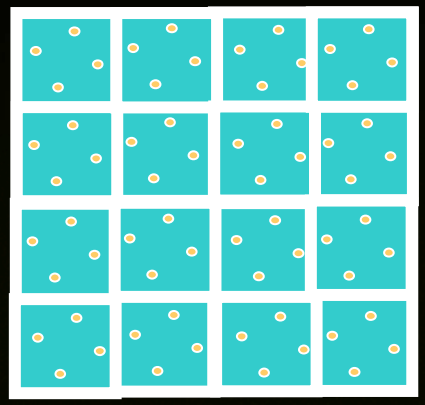
1024 bits/pixel

2x2 supersampling
of 8x multisampling
32 samples/pixel



2048 bits/pixel

4x4 supersampling
of 16x multisampling
64 samples/pixel



4096 bits/pixel

Quadro GPUs

Assume: 32-bit RGBA + 24-bit Z + 8-bit Stencil = 64 bits/sample

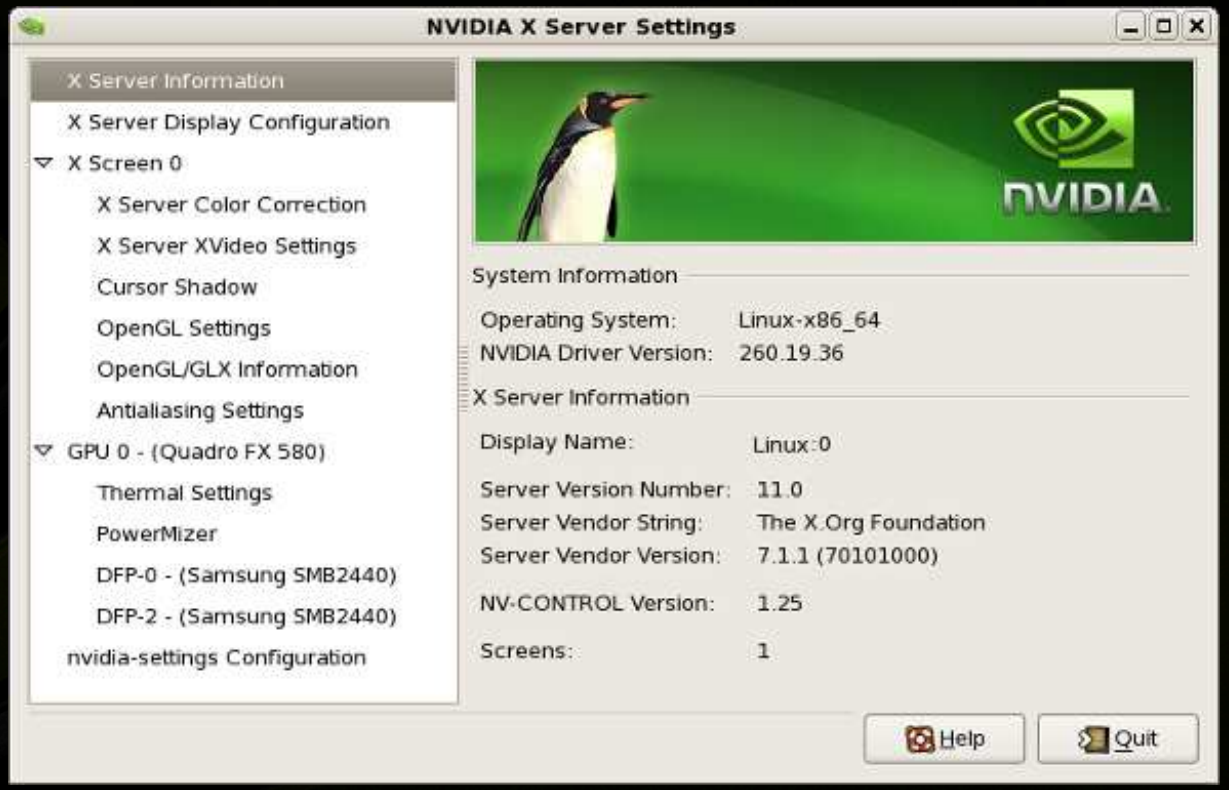
Image Quality Evolved

NVIDIA Fast Approximated Anti-Alias (FXAA)

Supported on Windows for seve driver releases…

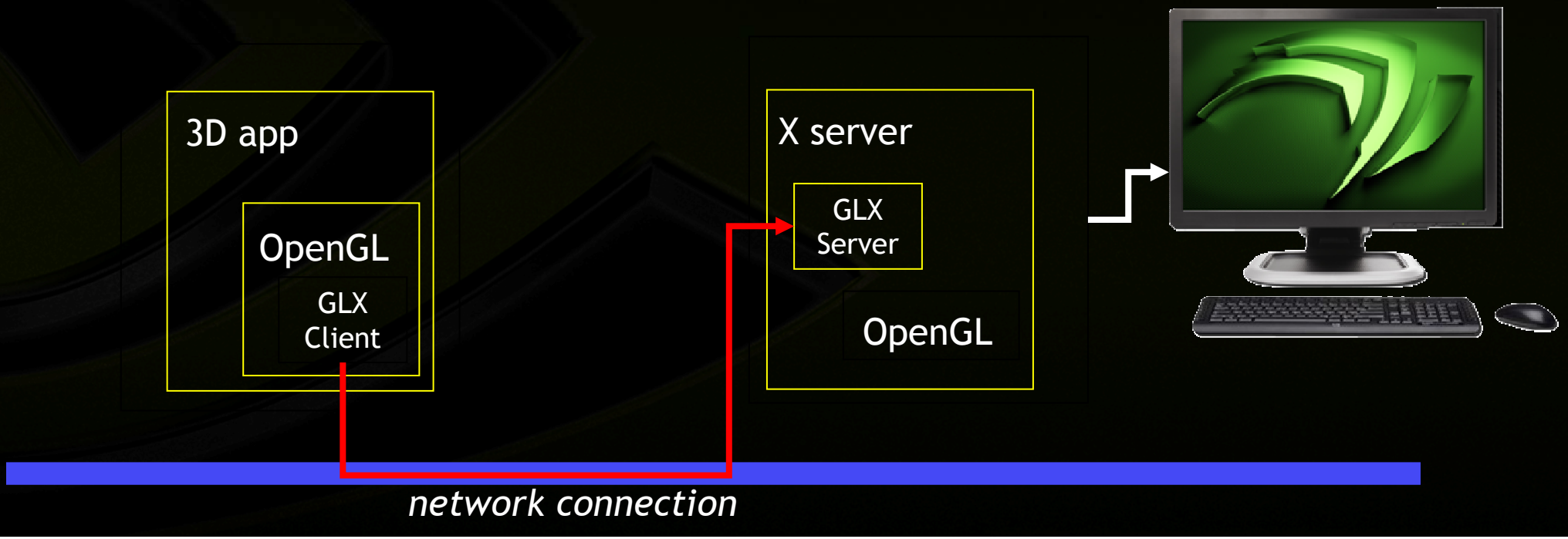
Now enabled fo Linux in 304.xx drivers

NVIDIA X Server Settings for Linux Control Panel



GLX Protocol

- Network transparent OpenGL
  - Run OpenGL app on one machine, display the X and 3D on a different machine



OpenGL-related Linux Improvements

Official GLX Protocol support for OpenGL extensions

- ARB_half_float_pixel
- ARB_transpose_matrix
- EXT_blend_equation_separate
- EXT_depth_bounds_test
- EXT_framebuffer_blit
- EXT_framebuffer_multisample
- EXT_packed_depth_stencil
- EXT_point_parameters
- EXT_stencil_two_side
- NV_copy_image
- NV_depth_buffer_float
- NV_half_float
- NV_occlusion_query
- NV_point_sprite
- NV_register_combiners2
- NV_texture_barrier

OpenGL-related Linux Improvements

Tentative GLX Protocol support for OpenGL extensions

- ARB_map_buffer_range
- ARB_shader_subroutine
- ARB_stencil_two_side
- EXT_transform_feedback2
- EXT_vertex_attrib_64bit
- NV_conditional_render
- NV_framebuffer_multisample_coverage
- NV_texture_barrier
- NV_transform_feedback2

Synchronizing X11-based OpenGL Streams

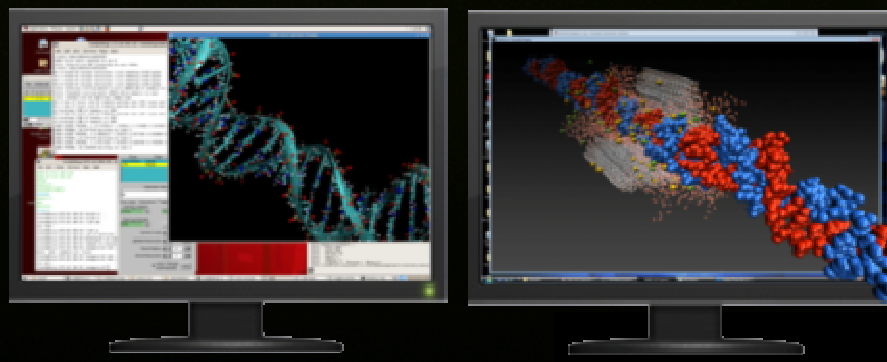
- New extension
  - GL_EXT_x11_sync_object
- Bridges the X Synchronization Extension with OpenGL 3.2 “sync” objects (ARB_sync)
- Introduces new OpenGL command
  - GLintptr sync_handle;
  - GLsync glImportSyncEXT (GLenum *external_sync_type*, GLintptr *external_sync*, GLbitfield *flags*);
    - *external_sync_type* must be GL_SYNC_X11_FENCE_EXT
    - *flags* must be zero

Other Linux Updates

- GL_CLAMP behaves in conformant way now
  - Long-standing work-around for original Quake 3
- Enabled 10-bit per component X desktop support
  - GeForce 8 and better GPUs
- Support for 3D Vision Pro stereo now

What is 3D Vision Pro?

- For Professionals
- All of 3D Vision support, plus
  - Radio frequency (RF) glasses, Bidirectional
  - Query compass, accelerometer, battery
  - Many RF channels – no collision
  - Up to ~120 feet
  - No line of sight needed to emitter
  - NVAPI to control



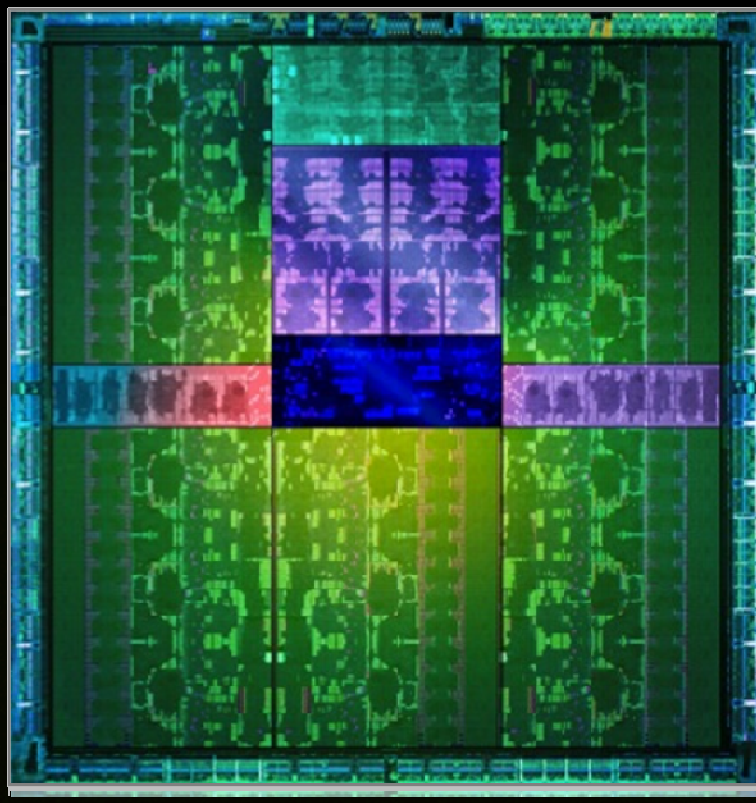
NV_path_rendering

- An NVIDIA OpenGL extension
  - GPU-accelerates resolution-independent 2D graphics
    - Very fast!
  - Supports PostScript-style rendering
- Come to my afternoon paper presentation to learn more
  - “GPU-Accelerated Path Rendering”
  - Garnet 217
  - Tomorrow, Friday, November 30
  - 14:15 - 16:00

Teaser Scene: 2D and 3D mix!

NVIDIA's Vision of Bindless Graphics

- **Problem:** Binding to different objects (textures, buffers) takes a lot of validation time in driver
  - And applications are limited to a small palette of bound buffers and textures
  - Approach of OpenGL, but also Direct3D
- **Solution:** Exposes GPU virtual addresses
  - Let shaders and vertex puller access buffer and texture memory by its virtual address!



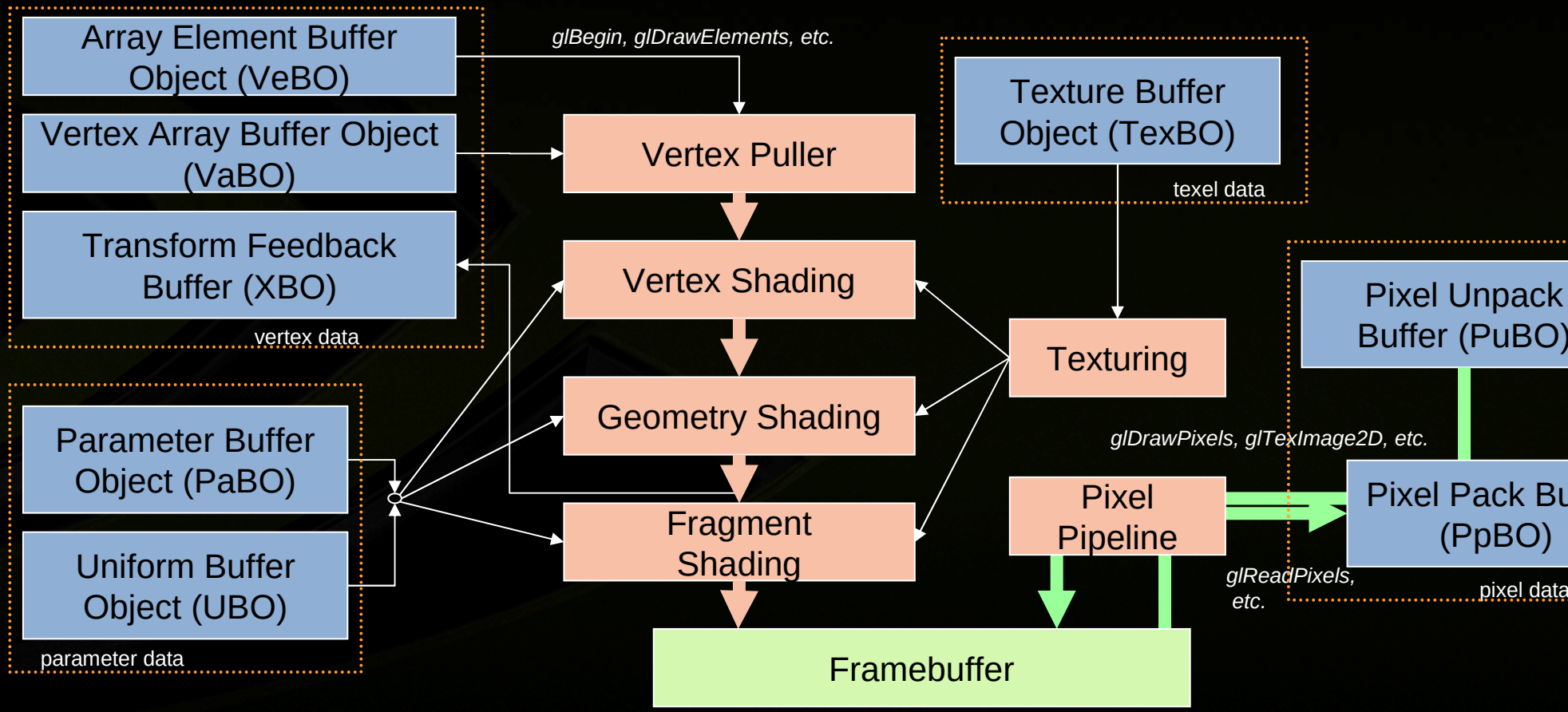
Kepler GPUs support bindless texture

Prior to Bindless Graphics

- Traditional OpenGL
  - GPU memory reads are “indirected” through bindings
    - Limited number of texture units and vertex array attributes
  - glBindTexture—for texture images and buffers
  - glBindBuffer—for vertex arrays

Buffer-centric Evolution

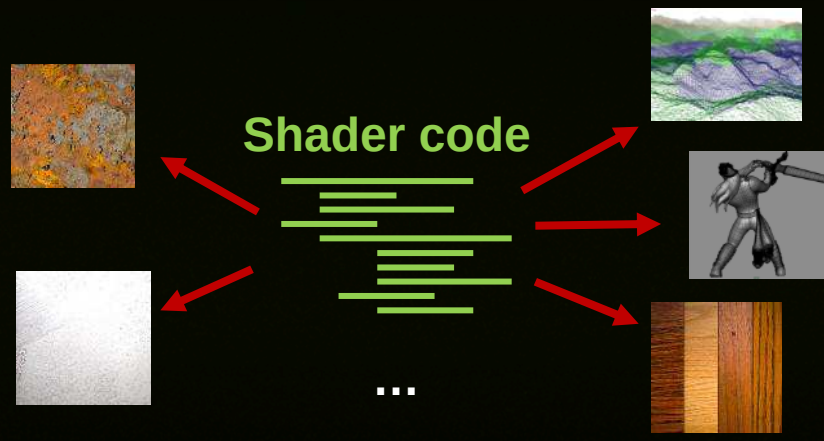
- Data moves onto GPU, away from CPU
  - Apps on CPUs just too slow at moving data otherwise



Kepler – Bindless Textures

- Enormous increase in the number of unique textures available to shaders at run-time
- More different materials and richer texture detail in a scene

Pre-Kepler texture binding model



Kepler bindless textures
over 1 million unique textures

Kepler – Bindless Textures

Pre-Kepler texture binding model

CPU

Load texture A
Load texture B
Load texture C
Bind texture A to slot I
Bind texture B to slot J
Draw()

GPU

Read from texture at slot I
Read from texture at slot J

CPU

Bind texture C to slot K
Draw()

GPU

Read from texture at slot K

Kepler bindless textures

CPU

Load textures A, B, C
Draw()

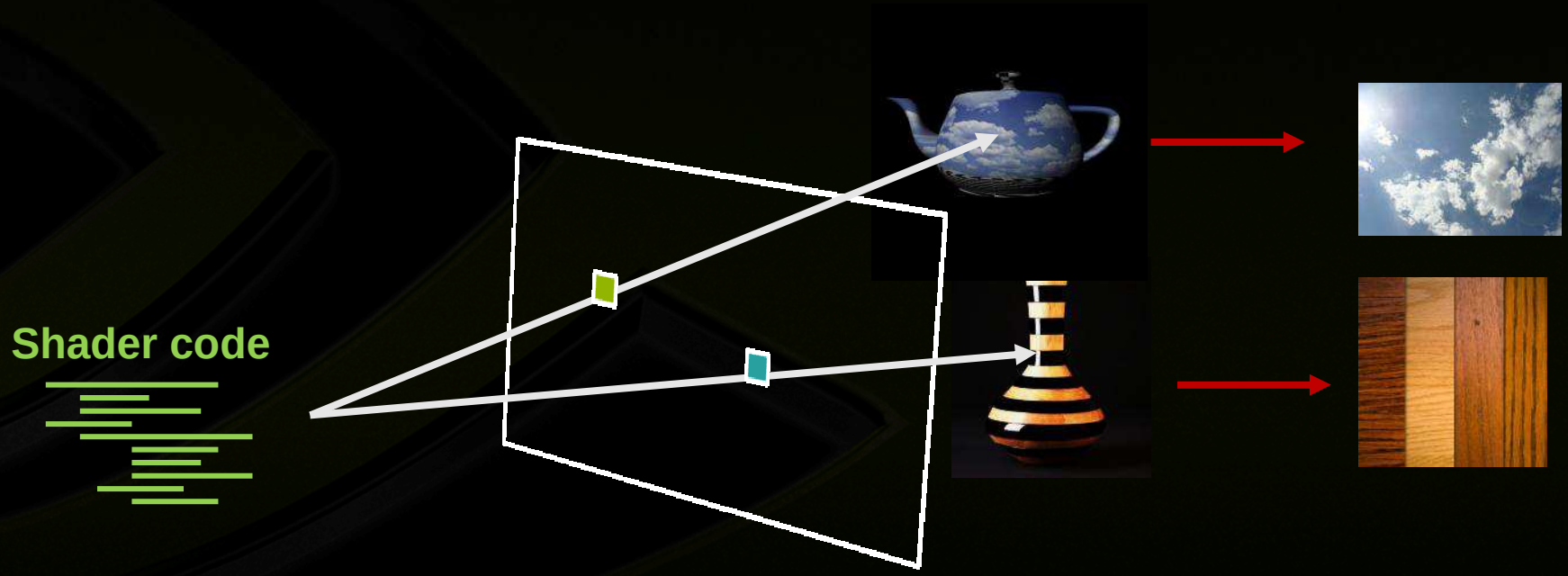
GPU

Read from texture A
Read from texture B
Read from texture C

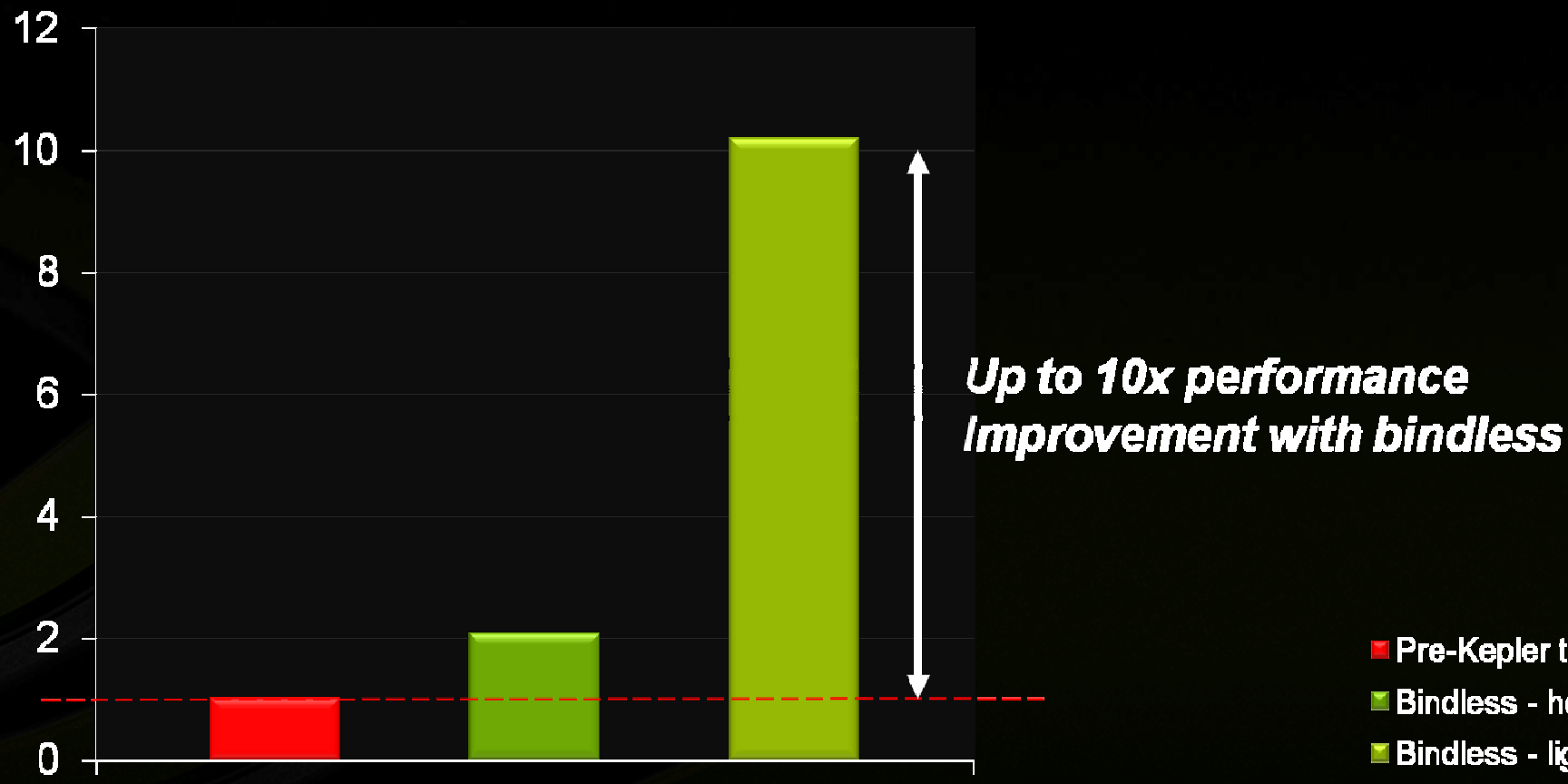
Bindless model reduces CPU overhead and improves GPU acces efficiency

Bindless Textures

- Apropos for ray-tracing and advanced rendering where textures cannot be “bound” in advance



Bindless performance benefit



Numbers obtained with a directed test

More Information on Bindless Texture

- Kepler has new NV_bindless_texture extension
  - Texture companion to
    - NV_vertex_buffer_unified_memory for bindless vertex arrays
    - NV_shader_buffer_load for bindless shader buffer reads
    - NV_shader_buffer_store (also NEW) for bindless shader buffer writes
- API specifications are publically available
  - http://developer.download.nvidia.com/opengl/specs/GL_NV_bindless_texture.txt

NVIDIA's Position on OpenGL Deprecation: Core vs. Compatibility Profiles

- OpenGL 3.1 introduced notion of "core" profile
- Idea was remove stuff from core to make OpenGL "good-er"
  - Well-intentioned perhaps but…
  - Throws API backward compatibility out the window
- Lots of useful functionality got removed that is in fast hardware
  - Examples: Polygon mode, line width, GL_QUADS
- Lots of easy-to-use, effective API got labeled deprecated
  - Immediate mode
  - Display lists
- Best advice for real developers
  - Simply use the "compatibility" profile
  - Easiest course of action
    - Requesting the core profile requires special context creation gymnastics
  - Avoids the frustration of "they decided to remove what??"
  - Allows you to use existing OpenGL libraries and code as-is
- No, your program won't go faster for using the "core" profile
  - It may go slower because of extra "is this allowed to work?" checks

Nothing changes with OpenGL 4.3
NVIDIA still committed to compatibility without compromise

Questions?

Other OpenGL-related NVIDIA Sessions at SIGGRAPH Asia

- Scaling OpenGL Applications Across Multiple GPU's
  - Thursday, 29 November 15:00 - 15:45, Conference Hall K (here)
  - **Shalini Venkataraman** / Senior Applied Engineer, NVIDIA
- Developing an Optimized Maya Plugin Using CUDA and OpenGL
  - Friday, 30 November 11:00 - 11:45, Conference Hall K
  - **Wil Braithwaite** / Senior Applied Engineer, NVIDIA
- OpenGL ES Performance Tools and Optimization for Tegra Mobile Devices
  - Friday, 30 November 13:00 - 13:45, Conference Hall K
  - **Xueqing Yang** / Developer Technology Engineer, NVIDIA
- GPU-accelerated Path Rendering
  - Friday, 30 November 14:15 – 16:00, Peridot 206
  - Third paper in “Points and Vectors” technical papers session
  - **Mark Kilgard** / Principal Software Engineer, NVIDIA