Candidate features for   
OpenGL 5 and Direct3D 12 hardware

6 April 2014, [Christophe Riccio](mailto:mail@g-truc.net)



[G-Truc Creation](http://www.g-truc.net)

# Table of contents

Table of contents 2

Introduction 4

1. Draw submission 5

1.1. GL\_ARB\_multi\_draw\_indirect 5

1.2. GL\_ARB\_shader\_draw\_parameters 6

1.3. GL\_ARB\_indirect\_parameters 7

1.4. GL\_NV\_bindless\_multi\_draw\_indirect 7

1.5. GL\_AMD\_interleaved\_elements 7

1.6. WGL\_AMD\_gpu\_association and WGL\_NV\_gpu\_affinity 7

2. Resources 9

2.1. GL\_ARB\_bindless\_texture 9

2.2. GL\_NV\_shader\_buffer\_load and GL\_NV\_shader\_buffer\_store 9

2.3. GL\_ARB\_sparse\_texture 9

2.4. GL\_AMD\_sparse\_texture 10

2.5. GL\_AMD\_sparse\_texture\_pool 10

2.6. Texture stitching 10

2.7. GL\_KHR\_texture\_compression\_astc 10

2.8. GL\_INTEL\_map\_texture 10

2.9. GL\_ARB\_seamless\_cubemap\_per\_texture 11

2.10. DMA engines 11

3. Shader operations 12

3.1. GL\_ARB\_shader\_group\_vote 12

3.2. GL\_NV\_shader\_thread\_group 12

3.3. GL\_NV\_shader\_thread\_shuffle 12

3.4. GL\_NV\_shader\_atomic\_float 13

3.5. GL\_AMD\_shader\_atomic\_counter\_ops 13

3.6. GL\_ARB\_compute\_variable\_group\_size 13

3.7. GL\_NV\_gpu\_shader5 13

3.8. GL\_AMD\_gpu\_shader\_int64 13

3.9. GL\_NV\_vertex\_attrib\_integer\_64bit 14

3.10. GL\_AMD\_ shader\_trinary\_minmax 14

4. Framebuffer 15

4.1. GL\_AMD\_sample\_positions 15

4.2. GL\_EXT\_framebuffer\_multisample\_blit\_scaled 15

4.3. GL\_NV\_multisample\_coverage and GL\_NV\_framebuffer\_multisample\_coverage 15

4.4. GL\_AMD\_depth\_clamp\_separate 15

5. Blending 16

5.1. GL\_NV\_texture\_barrier 16

5.2. GL\_EXT\_pixel\_local\_storage (ES) 16

5.3. GL\_EXT\_shader\_framebuffer\_fetch (ES) 16

5.4. GL\_ARM\_shader\_framebuffer\_fetch\_depth\_stencil (ES) 16

5.5. GL\_INTEL\_fragment\_shader\_ordering 16

5.6. GL\_INTEL\_conservative\_rasterization 16

5.7. GL\_KHR\_blend\_equation\_advanced 16

5.8. GL\_AMD\_blend\_minmax\_factor 16

6. Stencil 17

6.1. GL\_AMD\_stencil\_operation\_extended 17

6.2. GL\_AMD\_shader\_stencil\_export 17

7. Rendering pipeline 18

7.1. GL\_AMD\_vertex\_shader\_layer 18

7.2. GL\_AMD\_vertex\_shader\_viewport\_index 18

7.3. GL\_AMD\_transform\_feedback3\_lines\_triangles 18

7.4. GL\_AMD\_transform\_feedback4 18

8. Profiling 19

Conclusions 20

Is there room for an OpenGL 4.5 specification? Always! 20

Deprecation for OpenGL 5 hardware? 20

Features for OpenGL 5 hardware level? 21

Last words 21

# Introduction

The announcement of Mantle few months ago has triggered a lot of discussions about graphics API design. I think there are technical issues in OpenGL but those are precise problems that needs to be solve individually and following the hardware designs. These redesigning an entire API might be fun, it is keeping marketing people busy and [it makes OpenGL people communicating about how good is OpenGL](https://www.youtube.com/watch?v=-bCeNzgiJ8I&list=PLckFgM6dUP2hc4iy-IdKFtqR9TeZWMPjm).

What I like with Direct3D is that Microsoft was important enough to drive IHVs to standardize hardware features. The Khronos Group is certainly getting better at it, the [ASTC texture format](http://www.opengl.org/registry/specs/KHR/texture_compression_astc_hdr.txt) is a good example as I expect this format will be supported on all mobile and desktop GPUs in the future. How strong is Microsoft these days? This is something we will be able to judge at GDC.

What's an OpenGL 5 hardware feature? Following the conversions for OpenGL 3 and OpenGL 4, it's any hardware feature that can't be implemented on all OpenGL 4 hardware but would be implementable on newer hardware by all IHVs.

In this article I would like to point at hardware features available through OpenGL extensions and ideas that may or may not be interesting to standardize. As we will see, there are a lot of great features that could build this OpenGL 5 and Direct3D 12 hardware generation.

# 1. Draw submission

## 1.1. [GL\_ARB\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt)

With OpenGL 4.2 the application is responsible to dispatch the draws to the GPU. With OpenGL 1.0 is was done using [glBegin](http://www.opengl.org/sdk/docs/man2/xhtml/glBegin.xml)/[glEnd](http://www.opengl.org/sdk/docs/man2/xhtml/glEnd.xml), the immediate mode but quickly it appears that such approach building and sending vertex one by one was way too slow to be interesting because the GPUs were faster to consume the primitives than the CPU was able to submit them.

OpenGL has evolved many times to ensure that this submission could be done quickly enough:

* [Vertex Array](http://www.opengl.org/registry/specs/EXT/vertex_array.txt) (GL1.1);
* [Vertex Buffer Object](http://www.opengl.org/registry/specs/ARB/vertex_buffer_object.txt) (GL1.5);
* [Vertex Array Object](http://www.opengl.org/registry/specs/ARB/vertex_array_object.txt) (GL3.0);
* [Base Vertex](http://www.opengl.org/registry/specs/ARB/draw_elements_base_vertex.txt) (GL3.2);
* [Instancing](http://www.opengl.org/registry/specs/ARB/draw_instanced.txt) (GL3.2);
* [Instanced Arrays](http://www.opengl.org/registry/specs/ARB/instanced_arrays.txt) (GL3.3);
* [Base Instance](http://www.opengl.org/registry/specs/ARB/base_instance.txt) (GL4.2) ; and
* [Vertex Attrib Binding](http://www.opengl.org/registry/specs/ARB/vertex_attrib_binding.txt) (GL4.3)

By batching multiple VAOs into a single VAO by copying multiple sets of buffers into a single set of buffers and then calling [glDrawElementsInstancedBaseVertexBaseInstance](http://www.opengl.org/sdk/docs/man4/xhtml/glDrawElementsInstancedBaseVertexBaseInstance.xml) or [glDrawArraysInstancedBaseInstance](http://www.opengl.org/sdk/docs/man4/xhtml/glDrawArraysInstancedBaseInstance.xml) multiple times on that single VAO to draw the needed meshes we can already archive extremely good performances and the CPU overhead taken by this approach will be very low. However, the CPU remains in change of figuring out which meshes needs to be drawn for a specific frame which for many applications can occupy an entire CPU core…

Thanks to [ARB\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt) and [ARB\_compute\_shader](http://www.opengl.org/registry/specs/ARB/compute_shader.txt) it is possible to move this processing on the GPU, the compute shader becoming responsible to create a draw indirect buffer, which will store the parameters for multiple draw calls. This buffer will then be consumed by [glMultiDrawElementsIndirect](http://www.opengl.org/sdk/docs/man4/xhtml/glMultiDrawElementsIndirect.xml) or [glMultiDrawArraysIndirect](http://www.opengl.org/sdk/docs/man4/xhtml/glMultiDrawArraysIndirect.xml).

The glMultiDraw\*Indirect functions are nothing more than evolutions of the glDraw\*Indirect introduced with OpenGL 4.0. Instead of processing a single draw per draw call, the new functions can submit many draws per calls.

|  |
| --- |
| glBindBuffer(GL\_DRAW\_INDIRECT\_BUFFER, BufferName);  for(std::size\_t DrawIndex = 0; DrawIndex < DrawCount; ++DrawIndex)  glDrawElementsIndirect(GL\_TRIANGLES, GL\_UNSIGNED\_SHORT, BUFFER\_OFFSET(Offset)); |

Listing 1.1.1: CPU draw dispatching.

|  |
| --- |
| glBindBuffer(GL\_DRAW\_INDIRECT\_BUFFER, BufferName);  glMultiDrawElementsIndirect(GL\_TRIANGLES, GL\_UNSIGNED\_SHORT, nullptr, GLsizei(DrawCount), 0); |

Listing 1.1.2: GPU draw dispatching.

Vertex construction

A first usage of the shader storage buffer is when it is brought together with rendering without vertex attribute and multi draw indirect ([ARB\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt) promoted from [AMD\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/AMD/multi_draw_indirect.txt)). We can imagine that a compute shader can select in advance which parts of a mesh actually need to be rendered. This detection generated the multi draw indirect buffer, which will be used to pull specific draws from shader storage buffers that store a larger mesh. This method will be demonstrated in my GPU Pro article to be released.

Multi draw indirect is part of OpenGL 4.3 core specification but it's arguably an OpenGL 5 hardware feature allowing the GPU to submit itself draws to be executed. This feature can be implemented through software emulation quite easily using the CPU to push each individual draw but this is really slow. Currently, all the Intel GPU and AMD Evergreen support multi draw in software. Hardware implementation gives another magnitude of performance.

With 800000 draws per frame at 60Hz on Kepler and 300000 draws per frame on Southern Islands with a synthetic test rendering 2 triangles per draw on 4 pixels. That huge amount of draws provides an amazing control over the rendering.

Current support: NVIDIA Fermi; AMD Evergreen (emulated), Southern Islands; Intel Haswell (emulated)

Expected support: All OpenGL 5 hardware

## 1.2. [GL\_ARB\_shader\_draw\_parameters](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt)

This extension exposes three new built-in inputs to the vertex shader stage: gl\_BaseIntanceARB, gl\_BaseVertexARB that exposes the values passed in the draw commands but also gl\_DrawIDARB that behaves for multi-draws just like gl\_InstanceID for draw instancing. A massive difference between these new vertex shader inputs and gl\_InstancesID is that there are dynamically uniform variables so that we can use them to address arrays of resources.

|  |
| --- |
| layout(binding = INDIRECTION) uniform indirection {  int Transform[MAX\_DRAW];  } Indirection;  layout(binding = TRANSFORM0) uniform transform {  mat4 MVP[MAX\_DRAW];  } Transform;  layout(location = POSITION) in vec3 Position;  layout(location = TEXCOORD) in vec3 Texcoord;  out gl\_PerVertex {  vec4 gl\_Position;  };  out block {  vec2 Texcoord;  } Out;  void main(){  Out.Texcoord = Texcoord.st;  gl\_Position = Transform.MVP[Indirection.Transform[gl\_DrawIDARB]] \* vec4(Position, 1.0);  } |

Listing 2.1.1: Use sample of gl\_DrawIDARB to use a different matrix per draw in a multi draw call.

With the perspective of programmable vertex pulling, we could imagine removing the Position and Texcoord input variables and store them into a shader storage buffer and using gl\_BaseVertexARB and gl\_VertexID to fetch ourselves the vertex data.

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware, AMD Southern Islands

## 1.3. [GL\_ARB\_indirect\_parameters](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt)

An issue of [ARB\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/ARB/multi_draw_indirect.txt) is that it requires that we submit the number of draws from the CPU side, as <drawcount> is [glMultiDrawElementsIndirect](http://www.opengl.org/sdk/docs/man/xhtml/glMultiDrawElementsIndirect.xml) parameter. What if we use a compute shader to build the indirect multi-draw buffer? We need to query on the CPU side the number of draws stored in that buffer to feed the <drawcount> parameter: Very inefficient because we stall the CPU waiting on the query. Another workaround is to reserve a large buffer of <drawcount> elements and set to zero the primitive counts of each draw we want to skip. Unfortunately, according to my measurements, that solution is inefficient because GPUs are barely faster at skipping a draw than executing it.

The proposed solution is to add a parameter called <maxdrawcount> which value is sourced from an indirect parameter buffer. The maximum of executed draws becomes min(drawcount, maxdrawcount). Why not only source drawcount from a buffer? Because some command processors needs to know from the CPU the drawcount.

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware, AMD Southern Islands

## 1.4. [GL\_NV\_bindless\_multi\_draw\_indirect](http://www.opengl.org/registry/specs/NV/bindless_multi_draw_indirect.txt)

This is the last piece of NVIDIA bindless API allowing draw submission without binding vertex arrays or indirect draw buffer for lower CPU overhead.

Current hardware support: NVIDIA Fermi

Not needed with OpenGL 5 hardware

## 1.5. [GL\_AMD\_interleaved\_elements](http://www.opengl.org/registry/specs/AMD/interleaved_elements.txt)

This extension is really ugly but the functionality is really interesting. Instead of having a single element array, thanks to this extension we can have up to 4 element arrays and we can index each vertex attribute with the element array of our choice. There is quite a few software actually generating meshes using multiple element array so this functionality sounds extremely useful. Taking advantage of this feature can save bandwidth avoiding duplicating vertex attributes.

Current hardware support: AMD Southern Islands

## 1.6. [WGL\_AMD\_gpu\_association](http://www.opengl.org/registry/specs/AMD/wgl_gpu_association.txt) and [WGL\_NV\_gpu\_affinity](http://www.opengl.org/registry/specs/NV/gpu_affinity.txt)

I have never really explored either [AMD\_gpu\_association](http://www.opengl.org/registry/specs/AMD/wgl_gpu_association.txt) or [NV\_gpu\_affinity](http://www.opengl.org/registry/specs/NV/gpu_affinity.txt) by lack of interest. These extensions enable CrossFire and SLI on AMD and NVIDIA GPUs. Such feature could probably be standardized into an OpenGL ARB extension.

Expected hardware support: OpenGL 3 hardware

# 2. Resources

## 2.1. [GL\_ARB\_bindless\_texture](http://www.opengl.org/registry/specs/ARB/bindless_texture.txt)

[ARB\_bindless\_texture](http://www.opengl.org/registry/specs/ARB/bindless_texture.txt) is a very similar extension to [NV\_bindless\_texture](http://www.opengl.org/registry/specs/NV/bindless_texture.txt).

If there is still something wrong with the OpenGL API it's the texture binding API. Bindless texture API doesn’t fixed the current design but it provides an alternative that allows using about an infinite number of different textures simultaneously. I would be particularly interested to see some work on [Ptex](http://ptex.us/) using this extension.

Current hardware support: NVIDIA Kepler

Expected hardware support: All OpenGL 5 hardware, AMD Southern Islands

## 2.2. [GL\_NV\_shader\_buffer\_load](http://www.opengl.org/registry/specs/NV/shader_buffer_load.txt) and [GL\_NV\_shader\_buffer\_store](http://www.opengl.org/registry/specs/NV/shader_buffer_store.txt)

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware, AMD Southern Islands

## 2.3. [GL\_ARB\_sparse\_texture](http://www.opengl.org/registry/specs/ARB/sparse_texture.txt)

[ARB\_sparse\_texture](http://www.opengl.org/registry/specs/ARB/sparse_texture.txt) is a subset of [AMD\_sparse\_texture](http://www.opengl.org/registry/specs/AMD/sparse_texture.txt) enabling virtual texturing with seamless texture filtering. Thanks to this extension we can create 16K by 16K texel textures that memory isn’t fully resident. In practice when we create a sparse texture, a large table of pointers to memory pages is allocated. The allocation of these memory pages is potentially an independent task performed with glTexPageCommitmentARB on a subsection of that texture.

The API has changed a bit from the original extension. It's worth noticing that all the shader queries are gone. Allocation and copy have been decoupled so that we are no longer required to pass data for a sparse texture that we are actually going to fill afterward.

Sparse textures can be used for sampling and rendering. Supported formats are not specified. Multisample textures are explicitly not supported. For others formats, including compressed and depth stencil formats, it’s a matter of querying GL\_NUM\_VIRTUAL\_PAGE\_SIZES\_ARB. Supporting depth formats is a serious advantage to do high-resolution shadow map generation.

This extension is a really nice new feature for OpenGL that Direct3D 11.2 has the good idea to inspire itself. We can regret that the ARB didn't take time to create an ARB\_sparse\_buffer extension at the same time.

This feature remains too limited as it only relies on the virtual addressing of GPUs but it is bound by the current texture filtering capabilities. A 16K x 16K x 2K (arrays) texture is nice but 1M x 1M x 1M sparse textures would be so much better. AMD supports texture 3D sparse textures however the tiles are only 2D planes on Southern Islands. This is inefficient and it doesn’t make sparse texture on this platform a great candidate for sparse voxel storage. For OpenGL 5 hardware? Let's wish it!

Current hardware support: AMD Southern Islands, NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware

## 2.4. [GL\_AMD\_sparse\_texture](http://www.opengl.org/registry/specs/AMD/sparse_texture.txt)

Current hardware support: AMD Southern Islands

Expected hardware support: Future hardware

## 2.5. GL\_AMD\_sparse\_texture\_pool

Current hardware support: AMD Volcanic Islands

Expected hardware support: Future hardware

## 2.6. Texture stitching

Expected hardware support: Future hardware

## 2.7. [GL\_KHR\_texture\_compression\_astc](http://www.opengl.org/registry/specs/KHR/texture_compression_astc_hdr.txt)

The Khronos Group has standardized a new texture format called ASTC that provides very low bit rate and HDR support. Because, it's a KHR extension, it means that both the OpenGL ES group and the OpenGL ARB group voted to support that feature which gives me good hope that we will "soon" have support for this format on all platforms.

|  |  |  |
| --- | --- | --- |
| Block size | Bits per pixels | Compression ratio |
| 4x4 | 8.00 | 4:1 |
| 5x4 | 6.40 | 5:1 |
| 5x5 | 5.12 | 6.25:1 |
| 6x5 | 4.27 | 7.5:1 |
| 6x6 | 3.56 | 9:1 |
| 8x5 | 3.20 | 10:1 |
| 8x6 | 2.67 | 12:1 |
| 8x8 | 2.00 | 16:1 |
| 10x5 | 2.56 | 12.5:1 |
| 10x6 | 2.13 | 15:1 |
| 10x8 | 1.60 | 20:1 |
| 10x10 | 1.28 | 25:1 |
| 12x10 | 1.07 | 30:1 |
| 12x12 | 0.89 | 36:1 |

Current hardware support: Imagination Technologies PowerVR6XT, ARM Mali T700, NVIDIA Maxwell

Expected hardware support: Future hardware

## 2.8. [GL\_INTEL\_map\_texture](http://www.opengl.org/registry/specs/INTEL/map_texture.txt)

This feature allows to map the storage of a texture 2D just like we can do it with a buffer. The extension supports linear and [**tiled formats**](https://duckduckgo.com/Z-order_(curve)) however if this feature had to be standardize only the linear storage could be supported across current GPUs. However, linear storage is texture cache inefficient. Each GPU tile format is vendor specific so if we had to have that feature, the texture layout would have to be standardized.

Current hardware support: Intel Sandy Bridge

Expected hardware support: All OpenGL 5 hardware, AMD Evergreen, NVIDIA Fermi

## 2.9. [GL\_ARB\_seamless\_cubemap\_per\_texture](http://www.opengl.org/registry/specs/ARB/seamless_cubemap_per_texture.txt)

OpenGL 3.2 and [ARB\_seamless\_cube\_map](http://www.opengl.org/registry/specs/ARB/seamless_cube_map.txt) provide a state for sampling a cube map accessing multiple faces to avoid seams. This functionality is embodied by a global state that affects every cubemaps. If we want to use seamless cube map filtering for one cube map we need to call glEnable(GL\_TEXTURE\_CUBE\_MAP\_SEAMLESS). If we want to don’t want to use it on another texture, we need to call glDisable(GL\_TEXTURE\_CUBE\_MAP\_SEAMLESS). If we would to apply these two textures on a single mesh, then we need to do two rendering passes. [ARB\_seamless\_cubemap\_per\_texture](http://www.opengl.org/registry/specs/ARB/seamless_cubemap_per_texture.txt) changes this behavior giving to each cube map texture and sampler a state to enable or not the seamless cubemap filtering so that we can in a single pass sample a single cube map both ways.

Current hardware support: AMD RV700, NVIDIA Kepler

Expected hardware support: All OpenGL 5 hardware

## 2.10. DMA engines

NVIDIA Fermi and AMD Northern Islands have dedicated DMA engines that can live their lives on their own. Hence a dedicated thread can could be in charge of streaming resources because at some point the application figure out that they might become useful. During these transfers, the graphics engine can continue his life independently without any required synchronization. Obviously, the transfers would have to be completed before using the resources but with enough anticipation we could need a synchronization object only for the purpose of guarantying correctness on all possible hardware but without actually hitting that fence.

Currently NVIDIA supports this behavior but only by creating a separated context on a dedicated thread. This is workable but cumbersome and it costs thread safety penalty for the entire OpenGL implementation.

An explicit use of the DMA engine for fully asynchronous transfers and performing transfer outside of the rendering code would be really nice to have.

Current hardware support: AMD Northern Islands, NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware

# 3. Shader operations

## 3.1. [GL\_ARB\_shader\_group\_vote](http://www.opengl.org/registry/specs/ARB/shader_group_vote.txt)

[ARB\_shader\_group\_vote](http://www.opengl.org/registry/specs/ARB/shader_group_vote.txt) is a small subset of [NV\_gpu\_shader5](http://www.opengl.org/registry/specs/NV/gpu_shader5.txt).

Branching is a very interesting topic with GPUs. I guess GPU design is essentially based on how we access memory and how multiple shader invocations diverge. After that, ALUs are just the cherry on the cacke. As a result, branching is a very important (if not the most important) topic when it comes to performance. For example, in [AMD Southern Islands architecture](http://developer.amd.com/wordpress/media/2012/12/AMD_Southern_Islands_Instruction_Set_Architecture.pdf), I found five different methods to handle branching.

Nevertheless, the OpenGL ARB has tackled this issue by considering how could we help the compiler to produce more efficient branching code. The proposed solution is exposed by [GL\_ARB\_shader\_group\_vote](http://www.opengl.org/registry/specs/ARB/shader_group_vote.txt), which provides the GLSL functions anyInvocationARB, allInvocationsARB, allInvocationsEqualARB to compare values across shader invocations.

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware, AMD Southern Islands

## 3.2. [GL\_NV\_shader\_thread\_group](http://www.opengl.org/registry/specs/NV/shader_thread_group.txt)

This extension goes into the "super resolution" range of idea where we no long want to think at a fixed pixel resolutions but instead we want to think at higher or lower resolution than the native resolution. GPU doesn't actually execute anything on a per pixel or per vertex base but in many different kind of grouping. The warp/wavefront is the grouping for shader invocation and another famous one is the quadpixel, a set of 4 fragments. The texture LOD calculation is computed per quadpixel because it is very complex to compute anatically the derivatives required for the texture LOD computation but it is really easy to compute within a quadpixel: It's only the different between the values across quadpixels.

This extension gives access to quadpixels allowing to swizzle the intermediate results accross each fragments. Let's say that fragment shader requires 4 texture sampling. In some areas, we could consider that it is not that useful to sample per fragment and we can deal will sampling per quadpixels. This feature should interact pretty well with [GL\_ARB\_shader\_group\_vote](http://www.opengl.org/registry/specs/ARB/shader_group_vote.txt).

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware, AMD Evergreen

## 3.3. [GL\_NV\_shader\_thread\_shuffle](http://www.opengl.org/registry/specs/NV/shader_thread_shuffle.txt)

This extension extends [GL\_NV\_shader\_thread\_group](http://www.opengl.org/registry/specs/NV/shader_thread_group.txt) to any of the shader invocations of a wrap/wavefront. It seems very likely that we could use [GL\_NV\_shader\_thread\_group](http://www.opengl.org/registry/specs/NV/shader_thread_group.txt) on any GPU because all GPUs use quadpixels however, the warp/wavefront size is different for each GPU vendors. 32 for NVIDIA, 64 for AMD and variable for Intel, between 4 to 16 according to the cases. This feature sounds particularly useful for post processed antialiazing and maybe things like soft shadows.

Current hardware support: NVIDIA Kepler

Expected hardware support: Future hardware

## 3.4. [GL\_NV\_shader\_atomic\_float](http://www.opengl.org/registry/specs/NV/shader_atomic_float.txt)

This extension is simply extending add and exchange to float atomic operations.

Current hardware support: NVIDIA Fermi

Expected hardware support: Future hardware

## 3.5. [GL\_AMD\_shader\_atomic\_counter\_ops](http://www.opengl.org/registry/specs/AMD/shader_atomic_counter_ops.txt)

[ARB\_shader\_atomic\_counters](http://www.opengl.org/registry/specs/ARB/shader_atomic_counters.txt) and OpenGL 4.2 introduced the concept of atomic counter operations be those where limited to increment, decrement and query. AMD GPUs support these atomic operations in GDS memory which is faster than image and buffer atomic operations. However, AMD GPUs support more atomic operations from GDS: Increment and decrement with wrap ; addition and subtraction ; minimum and maximum ; bitwise operators (AND, OR, XOR, etc.) ; masked OR operator ; exchange, and compare and exchange operators. [GL\_AMD\_shader\_atomic\_counter\_ops](http://www.opengl.org/registry/specs/AMD/shader_atomic_counter_ops.txt) exposes all these operations.

Current hardware support: AMD Southern Islands

## 3.6. [GL\_ARB\_compute\_variable\_group\_size](http://www.opengl.org/registry/specs/ARB/compute_variable_group_size.txt)

The purpose of this extension is simply to specify the sizes of a workgroup at dispatch time instead of compile time. This is an OpenCL 1.2 features that is not effectively implemented by AMD implementation. Also it sounds like an interesting feature, the implementation needs to be capable to set this size without recompiling the shaders.

Current hardware support: NVIDIA Fermi

Expected hardware support: All OpenGL 5 hardware

## 3.7. [GL\_NV\_gpu\_shader5](http://www.opengl.org/registry/specs/NV/gpu_shader5.txt)

This extension what release with Fermi GPUs. It offers many features that went to core specifications. However, one of the aspects of that specification is the large range of types it provides: 8, 16, 32 bits integers and half-precision floating point types.

Current hardware support: NVIDIA Fermi

## 3.8. [GL\_AMD\_gpu\_shader\_int64](http://www.opengl.org/registry/specs/AMD/gpu_shader_int64.txt)

[AMD\_gpu\_shader\_int64](http://www.opengl.org/registry/specs/AMD/gpu_shader_int64.txt) is a superset of the 64-bit support exposed by [NV\_gpu\_shader5](http://www.opengl.org/registry/specs/NV/gpu_shader5.txt) and supported by AMD Southern Islands. This extension introduces the following features:

* support for 64-bit scalar (int64\_t, uint64\_t) and vector integer data types (i64vec\*, u64vec\*), including uniform API, uniform buffer object, transform feedback, and shader input and output support;
* new built-in functions to pack and unpack 64-bit integer types into a two-component 32-bit integer vector (int64BitsToDouble, uint64BitsToDouble);
* new built-in functions to convert double-precision floating-point values to or from their 64-bit integer bit encodings (doubleBitsToInt64, doubleBitsToUint64);
* vector relational functions supporting comparisons of vectors of 64-bit integer types; and
* common functions abs, sign, min, max, clamp, and mix supporting arguments of 64-bit integer types.

[AMD\_gpu\_shader\_int64](http://www.opengl.org/registry/specs/AMD/gpu_shader_int64.txt) seems to be an obvious candidate to become an ARB extension and OpenGL 5 hardware feature.

Current hardware support: AMD Southern Islands

Expected hardware support: All OpenGL 5 hardware, NVIDIA Fermi

## 3.9. [GL\_NV\_vertex\_attrib\_integer\_64bit](http://www.opengl.org/registry/specs/NV/vertex_attrib_integer_64bit.txt)

[NV\_vertex\_attrib\_integer\_64bit](http://www.opengl.org/registry/specs/AMD/gpu_shader_int64.txt) requires [NV\_gpu\_shader5](http://www.opengl.org/registry/specs/NV/gpu_shader5.txt) to provide 64-bit integer and unsigned integer support for vertex attributes.

Current hardware support: NVIDIA Fermi

Expected hardware support: Not needed in OpenGL 5 hardware

## 3.10. [GL\_AMD\_ shader\_trinary\_minmax](http://www.opengl.org/registry/specs/AMD/shader_trinary_minmax.txt)

This extension adds functions to find the minimum, maximum and median of three float or integer scalar or vector inputs.

|  |  |
| --- | --- |
| Syntax | Description |
| genType min3(genType x, genType y, genType z) genIType min3(genIType x, genIType y, genIType z) genUType min3(genUType x, genUType y, genUType z) | Returns the per-component minimum value of x, y, and z |
| genType max3(genType x, genType y, genType z) genIType max3(genIType x, genIType y, genIType z) genUType max3(genUType x, genUType y, genUType z) | Returns the per-component maximum value of x, y, and z |
| genType mid3(genType x, genType y, genType z) genIType mid3(genIType x, genIType y, genIType z) genUType mid3(genUType x, genUType y, genUType z) | Returns the per-component median value of x, y, and z |

Current hardware support: AMD Southern Islands

Expected hardware support: Future Hardware

# 4. Framebuffer

## 4.1. [GL\_AMD\_sample\_positions](http://www.opengl.org/registry/specs/AMD/sample_positions.txt)

Setting the sample positions is to me a very useful feature for post processing antialising but also for very high multisampled rendering with could be nice for text rendering for example.

Current hardware support: AMD Evergreen

Expected hardware support: All OpenGL 5 hardware

## 4.2. [GL\_EXT\_framebuffer\_multisample\_blit\_scaled](http://www.opengl.org/registry/specs/EXT/framebuffer_multisample_blit_scaled.txt)

This extension is a collaboration between Apple and NVIDIA. It is obviously design to handle the high DPI screens. When calling glBlitFramebuffer to resolve a multisampled framebuffer, the OpenGL implementation can resolve and scale the framebuffer.

Current hardware support: NVIDIA G80

Expected hardware support: All OpenGL 5 hardware

## 4.3. [GL\_NV\_multisample\_coverage](http://www.opengl.org/registry/specs/NV/multisample_coverage.txt) and [GL\_NV\_framebuffer\_multisample\_coverage](http://www.opengl.org/registry/specs/NV/framebuffer_multisample_coverage.txt)

NVIDIA uses something call multisample coverage which allows to have coverage samples than color samplers. Hence, the implementation can adapted the multisampling according to the number of coverage samples covering the color sample.

In the second party of this article, we will discuss the possibilities for blending, stencil, profiling, rendering pipeline and misc features. We will discuss the announcements at GDC if there is enough public information given away that I could discuss about relevant things. We will conclude by my personal wish list for OpenGL 5 and Direct3D 12 hardware class.

Current hardware support: NVIDIA G80

## 4.4. [GL\_AMD\_depth\_clamp\_separate](http://www.opengl.org/registry/specs/AMD/depth_clamp_separate.txt)

Current hardware support: AMD RV670

# 5. Blending

Programmable blending has been on the wish list of many graphics programmers There are three approaches: Through a per tile shader stage, through a new per pixel shader stage or by modifying the fragment shader stage.

## 5.1. [GL\_NV\_texture\_barrier](http://www.opengl.org/registry/specs/NV/texture_barrier.txt)

Texture barrier is an NVIDIA extension but that extension has been implemented by all vendors (even Apple!). This extension was the very first step toward programmable blending allowing to read onces and write onces at the same pixel location within a fragment shader invocation.

Current hardware support: AMD R600, NVIDIA G80

Expected hardware support: Intel Sandy Bridge

## 5.2. GL\_EXT\_pixel\_local\_storage (ES)

Current hardware support: Imagination Technologies PowerVR Rogue

Expected hardware support: Future hardware

## 5.3. GL\_EXT\_shader\_framebuffer\_fetch (ES)

## 5.4. GL\_ARM\_shader\_framebuffer\_fetch\_depth\_stencil (ES)

## 5.5. [GL\_INTEL\_fragment\_shader\_ordering](http://www.opengl.org/registry/specs/INTEL/fragment_shader_ordering.txt)

Current hardware support: Intel Haswell, AMD Southern Islands

Expected hardware support: Future hardware

## 5.6. GL\_INTEL\_conservative\_rasterization

Current hardware support: Intel Haswell

Expected hardware support: Future hardware

## 5.7. [GL\_KHR\_blend\_equation\_advanced](http://www.opengl.org/registry/specs/NV/blend_equation_advanced.txt)

Current hardware support: NVIDIA Fermi

## 5.8. [GL\_AMD\_blend\_minmax\_factor](http://www.opengl.org/registry/specs/AMD/blend_minmax_factor.txt)

Current hardware support: AMD Northern Islands

# 6. Stencil

## 6.1. GL\_[AMD\_stencil\_operation\_extended](http://www.opengl.org/registry/specs/AMD/stencil_operation_extended.txt)

Current hardware support: AMD Southern Islands

Expected hardware support: Future hardware

## 6.2. GL\_[AMD\_shader\_stencil\_export](http://www.opengl.org/registry/specs/AMD/shader_stencil_export.txt)

Current hardware support: AMD RV670

Expected hardware support: Future hardware

# 7. Rendering pipeline

## 7.1. [GL\_AMD\_vertex\_shader\_layer](http://www.opengl.org/registry/specs/AMD/vertex_shader_layer.txt)

OpenGL 3 hardware introduced layered rendering allowing to render each primitive to a different framebuffer attachment. However, to leverage this functionality, we need to use a geometry shader to specify gl\_LayerID per generated primitive. Using a geometry shader is not free on contrary of setting gl\_LayerID. Following this reasoning, AMD published [AMD\_vertex\_shader\_layer](http://www.opengl.org/registry/specs/AMD/vertex_shader_layer.txt) which allows to set gl\_LayerID in the vertex shader. Considering that mobile GPUs don't have a geometry shader, it would be particularly useful to use layered rendering.

## 7.2. [GL\_AMD\_vertex\_shader\_viewport\_index](http://www.opengl.org/registry/specs/AMD/vertex_shader_viewport_index.txt)

This extension follows the same reasoning than [AMD\_vertex\_shader\_layer](http://www.opengl.org/registry/specs/AMD/vertex_shader_layer.txt), enabling to choose the rendering view port gl\_ViewportIndex from the vertex shader stage.

## 7.3. [GL\_AMD\_transform\_feedback3\_lines\_triangles](http://www.opengl.org/registry/specs/AMD/transform_feedback3_lines_triangles.txt)

With unextended OpenGL, the application can use multiple transform feedback streams but only the first stream can output primitives that are not points. [AMD\_transform\_feedback3\_lines\_triangles](http://www.opengl.org/registry/specs/AMD/transform_feedback3_lines_triangles.txt) removes this restriction; any primitive can be generated to in any stream.

## 7.4. [GL\_AMD\_transform\_feedback4](http://www.opengl.org/registry/specs/AMD/transform_feedback4.txt)

[AMD\_transform\_feedback4](http://www.opengl.org/registry/specs/AMD/transform_feedback4.txt) extends [AMD\_transform\_feedback3\_lines\_triangles](http://www.opengl.org/registry/specs/AMD/transform_feedback3_lines_triangles.txt) so that each stream can be rendered in a single draw.

# 8. Profiling

# Conclusions

## Is there room for an OpenGL 4.5 specification? Always!

Continuing to rant about not having a better support of direct state access sounds like a desperate cause 5 years after the release of [EXT\_direct\_state\_access](http://www.opengl.org/registry/specs/EXT/direct_state_access.txt). Maybe, but this vision for the API remains as legitimate as even. That issue has to be tackle even if it will be particularly disrupting for the OpenGL API.

I guess a massive work that needs to be undertaken is separating the sampler and the texture in the shader. Direct3D 11 requires 128 texture units but only 32 samplers but with OpenGL we are stock at 32 texture units because the sampler and the texture is a monolithic unit.

I am also still missing the concept of sizeof and locationof in GLSL to be able to set my location depending on the definition of structures and variable that I don’t necessarily know or that might constantly change during the development process. We, OpenGL programmers, are used to use sizeof in C++ programs, so we will feel confortable working with sizeof in the shaders.

Subroutines could be improved. They are such a weak API that we still need to setup on the CPU side. With the multi draw indirect capabilities we have a great approach to indexing. With sparse textures and bindless textures we have resources to use for a large number of draws. However, we need have bad API to choose the shader code we want to execute in the shader. Subroutine could leverage that if we get a way to avoid the CPU interventions.

Some developers would like something like Direct3D 11 deferred context. That approach doesn’t work so we shouldn’t expect it. However, we could consider the idea of exposing the command buffer.

Finally, OpenCL continues moving forward with [SPIR](http://www.khronos.org/registry/cl/specs/spir_spec-1.2-provisional.pdf), an IR for OpenCL, we could picture many developers happy to have an equivalent for shader code robustness, obfuscation and compile time performances.

## Deprecation for OpenGL 5 hardware?

Considering the idea of an OpenGL 5 hardware level we could imagine a new pass of deprecations:

* Deprecate vertex array object and the array buffer: Use shader storage buffer instead.
* Deprecate texture buffer object. Use shader storage buffer instead.
* Deprecate transform feedback buffer. Use shader storage buffer instead.
* Deprecate atomic counter buffer. Use shader storage buffer instead.
* Deprecate MIRROR wrapping modes.
* Deprecate the concept of locations for input and output variables.
* Deprecate compressed formats that have better alternatives.
* Deprecate mutable textures and buffers.
* Deprecate all conversions not directly supported by the DMA controllers.
* Deprecate renderbuffers.
* And probably many more!

## Features for OpenGL 5 hardware level?

On the domain of OpenGL 5 hardware level features, the OpenGL ARB has strike in the recent past leaving Direct3D just following what OpenGL has been doing. This is the case for example of the “tiled resources” which is nothing more than [AMD\_sparse\_texture](http://www.opengl.org/registry/specs/ARB/sparse_texture.txt). On other aspects the Direct3D 11 API doesn’t even reach parity with OpenGL 4.4.

The new ARB extensions for OpenGL 5 hardware are great but we can expect more from Kepler and Southern Islands architectures. Could we have ARB sparse buffers and ARB bindless buffers?

Then there is feature that probably no available hardware can do: Could we have a new pixel shader stage following the fragment shader stage to do programmable blending. This is more a tile based rendering view of the GPU world but I quite believe that convergence between desktop and mobile will happen and that it will turn out to be a programmable tile based rendering architecture. There are actually three approaches to imagine programmable blending:

* 1/ In the fragment shader where output would be array where we accumulate processed fragment.
* 2/ In a dedicated pixel shader stage where the input would be a list of fragments.
* 3/ In a dedicated tile shader stage where we could access all the fragments within a tile.

All three methods require on-chip memory to be efficient but when we consider the 64 KB of LDS in AMD Southern Islands CU, it sounds like that we are no very far from enough on-chip memory.

Hardware vendors also need to do some works on their command processors. Target for OpenGL 5 hardware: Being capable to process 1 million draws per seconds. Draws need to become smaller to allow finer granularity culling and to reduce partially used large batches from brut-force processing.

[ARB\_indirect\_parameters](http://www.opengl.org/registry/specs/ARB/indirect_parameters.txt) opens interesting perspectives but the feature might be too limited. We could imagine an approach where per-draw user-defined parameters could be defined to index resources in the shaders. Technically, we can only already do this by using vertex attributes with the divisor set to 1 and base instance set to 0. However in practice each vertex attributes are fetched per vertex, which make it quite wasteful. It’s hard to imagine whether this approach could be efficiently implemented on any current hardware but probably not for a target of 1 million draws per seconds.

We need a good support for sparse resources in OpenGL 5 hardware. This means textures and buffers with shader queries to check the residence of memory pages and capable to address a terabyte of data, 1 million layers textures, 1 million by 1 million texel textures and a 3D tile order for texture 3D to allow more efficient storage of sparse voxels.

## Last words

We are reaching the end of the evolution of the OpenGL 4/Direct3D generation with a pretty mature OpenGL 4.4 API. While that generation will remains for years a base line for productions, innovations through a new base line of GPU features are coming as the new ARB extensions rise to our attentions.

The medium term future of real time rendering is already written: The renderer will submit less than 10 draw calls per frames but generating thousands of draws for each of them, each indexing only the resources it needs thanks to sparse and bindless resources. Each draw will also dynamically select the code it needs to execute with-in über-shaders reducing the total number of shaders to a few. Bandwidth remains more critical as ever which will reduce the size of the batch and multiply the draws for a finer granularity GPU culling.

Ultimately, I think that mobile and desktop GPU architectures will converge to a form of programmable tile based GPUs where the triangle lists would be created by a compute shader and assigned to a multi draw indirect buffer. Going for tiles could also open more opportunities for programmable blending to do for example on-chip [order independent transparency](https://www.google.co.uk/search?q=order+independent+transparency&spell=1&sa=X&ei=vZf1UeLjGMXEtAbim4CwCA&ved=0CCoQvwUoAA) (OIT). [NV\_blend\_equation\_advanced](http://www.opengl.org/registry/specs/NV/blend_equation_advanced.txt) or [AMD\_blend\_minmax\_factor](http://www.opengl.org/registry/specs/AMD/blend_minmax_factor.txt) show recent interests in more blending programmability. My bet? All that could happen with Maxwell and I will keep my fingers crossed for it. 2014 should be another pretty interesting year!

* [OpenGL 4.4 core specification](http://www.opengl.org/registry/doc/glspec44.core.withchanges.pdf)
* [GLSL 4.4 specification](http://www.opengl.org/registry/doc/GLSLangSpec.4.40.diff.pdf)
* [OpenGL 4.4 review](http://www.g-truc.net/doc/OpenGL%204.4%20review.pdf)
* [OpenGL 4.3 review](http://www.g-truc.net/doc/OpenGL%204.3%20review.pdf)
* [OpenGL 4.2 review](http://www.g-truc.net/doc/OpenGL%204.2%20review.pdf)
* [OpenGL 4.1 review](http://www.g-truc.net/doc/OpenGL%204.1%20review.pdf)
* [OpenGL 4.0 review](http://www.g-truc.net/doc/OpenGL%204.0%20review.pdf)
* [OpenGL 3.3 review](http://www.g-truc.net/doc/OpenGL%203.3%20review.pdf)
* [OpenGL 4.4 Pipeline Map](http://www.g-truc.net/doc/OpenGL%204.4%20Pipeline%20Map.pdf)
* [OpenGL ES 3.0 Pipeline Map](http://www.g-truc.net/doc/OpenGL%20ES%203.0%20Pipeline%20Map.pdf)

A big thanks to [Vladimir Hrincar](https://twitter.com/LokimanSK) for the detailed review of the review and [Daniel Rákos](https://twitter.com/aqnuep) for sending his comments on the new extensions.