OpenCL / OpenGL Vertex Buffer Interoperability:
A Particle System Case Study

See the video at:
http://cs.oregonstate.edu/~mjb/cs575/Projects/particles.mp4

Mike Bailey
mjb@cs.oregonstate.edu
Oregon State University

Your C++ program writes initial values into the buffer on the GPU
OpenCL acquires the buffer
Each OpenCL kernel reads an (x,y,z) value from the buffer
Each OpenCL kernel updates its (x,y,z) value
OpenCL releases the buffer
OpenGL draws using the (x,y,z) values in the buffer on the GPU

In the Beginning of OpenGL …

You listed the vertices with separate function calls:

```c
glBegin( GL_TRIANGLES );
glVertex3f( x0, y0, z0 );
glVertex3f( x1, y1, z1 );
glVertex3f( x2, y2, z2 );
glVertex3f( x0, y0, z0 );
glVertex3f( x3, y3, z3 );
glVertex3f( x4, y4, z4 );
glEnd( );
```

Then someone noticed how inefficient that was, for three reasons:
1. Sending large amounts of small pieces of information is less efficient than sending small amounts of large pieces of information
2. The vertex coordinates were being listed in the CPU and were being transferred to the GPU every drawing pass
3. Some vertices were listed twice

Here’s What OpenGL Has Been Moving To: Vertex Buffer Objects

A Little Background -- the OpenGL Rendering Context

The OpenGL Rendering Context contains all the characteristic information necessary to produce an image from geometry. This includes transformations, colors, lighting, textures, where to send the display, etc.

If we were implementing the OpenGL state as a C++ structure, we might do something like this:

```c
struct context {
    float [4][4] Transformation;
    struct Texture * Texture;
    struct DataArrayBuffer * ArrayBuffer;
    ... Context;
}
```

More Background --
How do you create a special OpenGL Array Buffer called a Vertex Buffer Object?

In C++, objects are pointed to by their address.
In OpenGL, objects are pointed to by an unsigned integer handle. You can assign a value for this handle yourself (not recommended), or have OpenGL generate one for you that is guaranteed to be unique. For example:

```c
GLuint buf;
glGenBuffers( 1, &buf );
```

This doesn’t actually allocate memory for the buffer object yet, it just acquires a unique handle. To allocate memory, you need to bind this handle to the Context.
More Background – What is an OpenGL “Object”?

An OpenGL Object is pretty much the same as a C++ object: it encapsulates a group of data items and allows you to treat them as a unified whole. For example, a Data Array Buffer Object could be defined in C++ by:

```c
struct DataArrayBuffer {
    enum dataType;
    void * memStart;
    int memSize;
};
```

Then, you could create any number of Buffer Object instances, each with its own characteristics encapsulated within it. When you want to make that combination current, you just need to point the ArrayBuffer element of the Context to that entire struct (“bind”). When you bind an object, all of its information comes with it.

A Little Background -- the OpenGL Rendering Context

It’s very fast to re-bind a different vertex buffer. It amounts to just changing a pointer.

```c
glBindBuffer( GL_ARRAY_BUFFER, buf );
```

More Background -- “Binding” to the Context

The OpenGL term “binding” refers to “attaching” or “docking” (a metaphor which I find to be more visually pleasing) an OpenGL object to the Context. You can then assign characteristics, and they will “flow” through the Context into the object.

When you want to use that Vertex Buffer Object, just bind it again. All of the characteristics will then be active, just as if you had specified them again.

```c
glBindBuffer( GL_ARRAY_BUFFER, buf );
```

Vertex Buffers: Putting Data in the Buffer Object

```c
glBufferData( type, numBytes, data, usage );
```

*type* is the type of buffer object this is:

- GL_ARRAY_BUFFER to store floating point vertices, normals, colors, and texture coordinates

*numBytes* is the number of bytes to store in all. Not the number of numbers, but the number of bytes!

*data* is the memory address of (i.e., pointer to) the data to be transferred to the graphics card. This can be NULL, and the data can be transferred later.

Preview: We are going to use a Particle System as a Case Study
Vertex Buffers: Putting Data in the Buffer Object

```c
glBufferData( type, numbytes, data, usage );
```

`usage` is a hint as to how the data will be used: `GL_xxx_yyy`

where `xxx` can be:
- `STREAM`: this buffer will be written lots
- `STATIC`: this buffer will be written seldom and read seldom
- `DYNAMIC`: this buffer will be written often and used often

and `yyy` can be:
- `DRAW`: this buffer will be used for drawing
- `READ`: this buffer will be copied into
- `COPY`: not a real need for now, but someday...

`GL_STATIC_DRAW` is the most common usage.

Vertex Buffers: Step #1 – Fill the Arrays

```c
GLfloat Vertices[ ][3] = {
    { 1., 2., 3. },
    { 4., 5., 6. },
    /*...
```

```c
int numVertices = sizeof(Vertices) / (3*sizeof(GLfloat));
```

Vertex Buffers: Step #2 – Create the Buffers and Fill Them

```c
glGenBuffers( 1, &buf );
```

```c
glBindBuffer( GL_ARRAY_BUFFER, buf );
```

```c
glBufferData( GL_ARRAY_BUFFER, 3*sizeof(GLfloat)*numVertices, Vertices, GL_STATIC_DRAW );
```

Vertex Buffers: Step #3 – Activate the Array Types That You Will Use

```c
glEnableClientState( type )
```

where `type` can be any of:
- `GL_VERTEX_ARRAY`
- `GL_COLOR_ARRAY`
- `GL_NORMAL_ARRAY`
- `GL_TEXTURE_COORD_ARRAY`

- Call this as many times as you need to enable all the arrays that you will need.
- There are other types, too.
- To deactivate a type, call:
  ```c
glDisableClientState( type )
```

Vertex Buffers: Step #4 – To Draw, First Bind the Buffers

```c
glBindBuffer( GL_ARRAY_BUFFER, buf );
```

Vertex Buffers: Step #5 – Specify the Data

```c
glVertexPointer( size, type, stride, rel_address );
```

```c
glColorPointer( size, type, stride, rel_address );
```

```c
glNormalPointer( type, stride, rel_address );
```

```c
glTexCoordPointer( size, type, stride, rel_address );
```

`size` is the spatial dimension, and can be: 2, 3, or 4

type can be:
- `GL_SHORT`
- `GL_INT`
- `GL_FLOAT`
- `GL_DOUBLE`

`stride` is the byte offset between consecutive entries in the array (0 means tightly packed)

rel_address, the 4th argument, is the relative byte address from the start of the buffer where the first element of this part of the data lives.
The Data Types in a vertex buffer object can be stored either as “packed” or “interleaved”.

**Packed:**
- glVertexPointer(3, GL_FLOAT, 3*sizeof(GLfloat), 0);
- glColorPointer(3, GL_FLOAT, 3*sizeof(GLfloat), 3*numVertices*sizeof(GLfloat));

**Interleaved:**
- glVertexPointer(3, GL_FLOAT, 6*sizeof(GLfloat), 0);
- glColorPointer(3, GL_FLOAT, 6*sizeof(GLfloat), 3*sizeof(GLfloat));

The relative byte address, the 4th argument, is the relative byte address from the start of the buffer where the first element of this part of the data lives.

### Vertex Buffers: Step #6 – Specify the Connections

```gl
GLfloat Vertices[ ] = {
    { x0, y0, z0 },
    { x1, y1, z1 },
    { x2, y2, z2 },
    { x3, y3, z3 },
    { x4, y4, z4 },
    { x5, y5, z5 }
};
```

```gl
int numVertices = sizeof(Vertices) / (3*sizeof(GLfloat));
```

```gl
glDrawArrays( GL_TRIANGLES, 0, numVertices );
```

### Vertex Buffers: Writing Data Directly into a Vertex Buffer

Map the buffer from GPU memory into the memory space of the application:

```gl
glBindBuffer( buf, GL_ARRAY_BUFFER );
```

```gl
glBufferData( GL_ARRAY_BUFFER, 3*sizeof(float)*numVertices, NULL, GL_STATIC_DRAW );
```

```gl
float * vertexArray = glMapBuffer( GL_ARRAY_BUFFER, usage );
```

`usage` is an indication of how the data will be used:
- GL_READ_ONLY the vertex data will be read from, but not written to
- GL_WRITE_ONLY the vertex data will be written to, but not read from
- GL_READ_WRITE the vertex data will be read from and written to

You can now use `vertexArray` like any other floating-point array.

When you are done, be sure to call:

```gl
glUnMapBuffer( GL_ARRAY_BUFFER );
```

### Either OpenGL or OpenCL can use the Vertex Buffer at a time, but not both.

All of this happens on the GPU

OpenGL draws using the (x,y,z) values in the buffer on the GPU

Your C++ program writes initial values into the buffer on the GPU

OpenCL acquires the buffer

Each OpenCL kernel reads an (x,y,z) value from the buffer

Each OpenCL kernel writes its (x,y,z) value back to the buffer

OpenCL releases the buffer

Map the buffer from GPU memory into the memory space of the application:

### Structures We Will Use to Fill the Vertex Buffers

```c
#include <stdio.h>
#define _USE_MATH_DEFINES
#include <math.h>
#include <string.h>
#include <stdlib.h>
#include <ctype.h>
#include <omp.h>
#ifdef WIN32
    #include <windows.h>
#endif
#ifdef WIN32
    #include "glew.h"
#endif
#include <GL/gl.h>
#include <GL/glu.h>
#include "glut.h"
#include "glui.h"
#include "CL/cl.h"
#include "CL/cl_gl.h"
```

```c
// structs we will need later:
struct xyzw {
    float x, y, z, w;
};
struct rgba {
    float r, g, b, a;
};
```

```c
// structs we will need later:
struct xyzw {
    float x, y, z, w;
};
struct rgba {
    float r, g, b, a;
};
```
OpenCL Global Variables

```c
``` 

A Deceptively-Simple Main Program

```c
int main( int argc, char *argv[ ] )
{
    glutInit( &argc, argv );
    InitGraphics( );
    InitLists( );
    InitCL( );
    Reset( );
    InitGlui( );
    glutMainLoop( );
    return 0;
}
``` 

GLEW – the GL Extension Wrangler

```c
#ifdef WIN32
GLenum err = glewInit( );
if( err  !=  GLEW_OK )
{
    fprintf( stderr, "glewInit Error
" );
}
#endif
``` 

Setting up OpenCL: Querying the Existence of an OpenCL Extension

```c
void IsCLExtensionSupported( const char *extension )
{
    // see if the extension is bogus:
    if( extension == NULL  ||  extension[0] == '\0' )
        return false;
    char * where = (char *) strchr( extension, ' ' );
    if( where != NULL )
        return false;
    // get the full list of extensions:
    size_t extensionSize;
    clGetDeviceInfo( Device, CL_DEVICE_EXTENSIONS, 0, NULL, &extensionSize );
    char *extensions = new char [ extensionSize ];
    clGetDeviceInfo( Device, CL_DEVICE_EXTENSIONS, extensionSize, extensions, NULL );
    for( char * start = extensions ; ; )
    {
        where = (char *) strstr( (const char *) start, extension );
        if( where == 0 )
        {
            delete [] extensions;
            return false;
        }
        char * terminator = where + strlen(extension);  // points to what should be the separator
        if( *terminator == ' '  ||  *terminator == '\0'  ||  *terminator == '\r'  ||  *terminator == '\n' )
        {
            delete [] extensions;
            return true;
        }
        start = terminator;
    }
}
``` 

Setting up OpenCL: The Interoperability Context

```c
void CreateInterContext( )
{
    // 1. get the platform id:
    status = clGetPlatformIDs( 1, &Platform, NULL );
    PrintCLError( status, "clGetPlatformIDs: ");
    // 2. get the device id:
    status = clGetDeviceIDs( Platform, CL_DEVICE_TYPE_GPU, 1, &Device, NULL );
    PrintCLError( status, "clGetDeviceIDs: ");
    // 3. create a special opencl context based on the opengl context:
    cl_context_properties props[ ] =
    { 
        CL_GL_CONTEXT_KHR,  // default
        CL_WGL_HDC_KHR,  // use the wgl context
        CL_CONTEXT_PLATFORM, Platform,
        0
    };
    cl_context Context = clCreateContext( props, 1, &Device, NULL, NULL, &status );
    PrintCLError( status, "clCreateContext: ");
``` 

Querying the Existence of an OpenCL Extension
The Interoperability Context is Different for each OS (oh, good…)

For Windows:
```c
cl_context_properties props[] = {
    (cl_context_properties) NULL,
    0,
    CL_GL_CONTEXT_KHR, (cl_context_properties) NULL,
    CL_CONTEXT_PLATFORM, (cl_context_properties) NULL,
};
```

For Linux:
```c
cl_context_properties props[] = {
    0,
    CL_GLX_DISPLAY_KHR, (cl_context_properties) glXGetCurrentDisplay(),
    CL_GL_CONTEXT_KHR, (cl_context_properties) glXGetCurrentContext(),
    CL_CONTEXT_PLATFORM, (cl_context_properties) NULL,
};
```

For Apple:
```c
cl_context_properties props[] = {
    0,
    CL_CONTEXT_PROPERTY_USE_CGL_SHAREGROUP_APPLE,
    0,
    CL_GLX_DISPLAY_KHR, (cl_context_properties) NULL,
    CL_GL_CONTEXT_KHR, (cl_context_properties) NULL,
    CL_CONTEXT_PLATFORM, (cl_context_properties) NULL,
};
```

```
for (int i = 0; i < NUM_PARTICLES; i++)
    colors[i].a = 1.;
    colors[i].b = Ranf(&Seed, 0., 1.);
    colors[i].g = Ranf(&Seed, 0., 1.);
    colors[i].r = Ranf(&Seed, 0., 1.);
```

```
for (int i = 0; i < NUM_PARTICLES; i++)
    points[i].w = 1.;
    points[i].z = Ranf(&Seed, ZMIN, ZMAX);
    points[i].y = Ranf(&Seed, YMIN, YMAX);
    points[i].x = Ranf(&Seed, XMIN, XMAX);
```

Setting the Initial Particle Parameters

void ResetParticles()
```c
    glMapBuffer(GL_ARRAY_BUFFER, GL_WRITE_ONLY);
    glBindBuffer(GL_ARRAY_BUFFER, hPobj);
    glBufferData(GL_ARRAY_BUFFER, 4 * NUM_PARTICLES * sizeof(float), NULL, GL_STATIC_DRAW);
    glUnmapBuffer(GL_ARRAY_BUFFER);
```

void ResetColors()
```c
    glMapBuffer(GL_ARRAY_BUFFER, GL_WRITE_ONLY);
    glBindBuffer(GL_ARRAY_BUFFER, hCobj);
    glBufferData(GL_ARRAY_BUFFER, 4 * NUM_PARTICLES * sizeof(float), NULL, GL_STATIC_DRAW);
    glUnmapBuffer(GL_ARRAY_BUFFER);
```

Setting-up the Device-Side Buffers

void InitCL()
```c
    // 1. create the opengl version of the velocity array:
    hVel = new struct xyzw[NUM_PARTICLES];
    delete[] hVel;
```

```
    // 2. write the data from the host buffers to the device buffers:
    status = clEnqueueWriteBuffer(CmdQueue, dVel, CL_FALSE, 0, 4*sizeof(float)*NUM_PARTICLES, hVel, 0, NULL, NULL, &status);
```

```
    // 3. create the velocity array and the opengl vertex array buffer and color array buffer:
    clCreateBuffer(Context, CL_MEM_READ_WRITE, dVel, &status);
    clCreateFromGLBuffer(Context, GL_ARRAY_BUFFER, hVel, &status);
    clCreateFromGLBuffer(Context, GL_ARRAY_BUFFER, hCobj, &status);
```

Note: you don't need an OpenGL-accessible buffer for the velocities. Velocities aren't needed for drawing.

... to Match the Kernel's Parameter List

```
    // 10. setup the arguments to the Kernel object:
    status = clSetKernelArg(Kernel, 1, sizeof(cl_mem), &hPobj);
    status = clSetKernelArg(Kernel, 2, sizeof(cl_mem), &hVel);
    status = clSetKernelArg(Kernel, 3, sizeof(cl_mem), &hCobj);
```

```
    // fill those arrays and buffers:
    ResetParticles();
    ResetColors();
```

```
    // 11. create the velocity array and the opengl vertex array buffer and color array buffer:
    status = clCreateBuffer(Context, CL_MEM_READ_WRITE, dVel, &status);
    clCreateFromGLBuffer(Context, GL_ARRAY_BUFFER, hVel, &status);
```

```
    // 12. create the kernel:
    Kernel = clCreateKernel(Context, "kernel", &status);
```

```
    // 13. setup the arguments to the Kernel object:
    status = clSetKernelArg(Kernel, 1, sizeof(cl_mem), &hPobj);
    status = clSetKernelArg(Kernel, 2, sizeof(cl_mem), &hVel);
    status = clSetKernelArg(Kernel, 3, sizeof(cl_mem), &hCobj);
```

```
    // call the kernel:
    clEnqueueNDRangeKernel(CmdQueue, Kernel, 1, NULL, NULL, NULL, 1, NULL, NULL, &status);
```
The OpenGL "Idle Function" Tells OpenCL to Do Its Computing

```cpp
void Animate()
{
    // acquire the vertex buffers from opengl:
    glutSetWindow( MainWindow );
    glFinish(  );
    cl_int status =
        clEnqueueAcquireGLObjects( CmdQueue, 1, &dPobj, 0, NULL, NULL );
    PrintCLError( status, "clEnqueueAcquireGLObjects (1) : " );
    status =
        clEnqueueAcquireGLObjects( CmdQueue, 1, &dCobj, 0, NULL, NULL );
    PrintCLError( status, "clEnqueueAcquireGLObjects (2) : " );
    Wait();
    double time0 = omp_get_wtime( );
    // 11. enqueue the Kernel object for execution:
    cl_event wait;
    status = clEnqueueNDRangeKernel( CmdQueue, Kernel, 1, NULL, GlobalWorkSize, LocalWorkSize, 0, NULL, &wait );
    PrintCLError( status, "clEnqueueNDRangeKernel: " );
    Wait();
    double time1 = omp_get_wtime( );
    ElapsedTime =  time1 - time0;
    clFinish( CmdQueue );
    clEnqueueReleaseGLObjects( CmdQueue, 1, &dCobj, 0, NULL, NULL ) ;
    PrintCLError( status, "clEnqueueReleaseGLObjects (1): " );
    clEnqueueReleaseGLObjects( CmdQueue, 1, &dPobj, 0, NULL, NULL );
    PrintCLError( status, "clEnqueueReleaseGLObject (2): " );
    glutSetWindow( MainWindow );
    glutPostRedisplay(  );
}
```

Redrawing the Scene: The Particles

```cpp
void Display()
{
    ...
    if( ShowPerformance )
    {
        char str[128];
        sprintf( str, "%6.1f GigaParticles/Sec", (float)NUM_PARTICLES/ElapsedTime/1000000000. );
        glutSetWindow( MainWindow );
        glutPostRedisplay(  );
    }
    glDisable( GL_DEPTH_TEST );
    glMatrixMode( GL_PROJECTION );
    glLoadIdentity( );
    gluOrtho2D( 0., 100.,     0., 100. );
    glMatrixMode( GL_MODELVIEW );
    glLoadIdentity( );
    glColor3f( 1., 1., 1. );
    DoRasterString( 5., 5., 0., str );
    ...
}
```

13. Clean-up

```cpp
void Quit()
{
    Glui->close( );
    glutSetWindow( MainWindow );
    glFinish( );
    glutDestroyWindow( MainWindow );
    // 13. clean everything up:
    clReleaseKernel(                   Kernel   );
    clReleaseProgram(                Program  );
    clReleaseCommandQueue(  CmdQueue );
    clReleaseMemObject(           dPobj );
    clReleaseMemObject(           dCobj );
    exit( 0 );
}
```

Redraw the Scene: The Performance

```cpp
typedef float4 point;
typedef float4 vector;
typedef float4 color;
typedef float4 sphere;
constant float4 G               = (float4) ( 0., -9.8, 0., 0. );
constant float   DT             = 0.1;
constant sphere Sphere1 = (sphere)( -100., -800., 0.,  600. );

bool IsInsideSphere( point p, sphere s )
{
    float r = fast_length( p.xyz - s.xyz );
    return  ( r < s.w );
}

void Particle(  global point * dPobj,  global vector * dVel,  global color * dCobj )
{
    int gid = get_global_id( 0 ); // particle #
    point p   = dPobj[gid];
    vector v = dVel[gid];
    point pp   = p + v*DT + G * (point)(.5*DT*DT); // p'
    vector vp = v + G*DT; // v'
    pp.w = 1.;
    vp.w = 0.;
    if(   IsInsideSphere( pp, Sphere1 )   )
    {
        vp = BounceSphere( p, v, Sphere1 );
        pp = p + vp*DT + G * (point)( .5*DT*DT );
    }
    dPobj[gid] = pp;
    dVel[gid]   = vp;
}
```

particles.cl
vector Bounce( vector in, vector n )
{
    n.w = 0.;
    n = normalize( n );
    vector out = in - n * (vector)( 2. * dot( in.xyz, n.xyz ) );
    out.w = 0.;
    return out;
}

vector BounceSphere( point p, vector v, sphere s )
{
    vector n;
    n.xyz = fast_normalize( p.xyz - s.xyz );
    n.w = 0;
    return Bounce( in, n );
}