The Open Computing Language (OpenCL)

OpenCL consists of two parts: a C/C++-callable API and a C-ish programming language.

The OpenCL programming language can run on NVIDIA GPUs, AMD GPUs, Intel CPUs, Intel GPUs, mobile devices, and (supposedly) FPAGAs (Field-Programmable Gate Arrays).

But, OpenCL is at its best on compute devices with large amounts of data parallelism, which usually implies GPU usage.

You break your computational problem up into lots and lots of small pieces. Each piece gets farmed out to threads on the GPU.

Each thread wakes up and is able to ask questions about where it lives in the entire collection of (thousands of) threads. From that, it can tell what it is supposed to be working on.

OpenCL can share data, and interoperate, with OpenGL.

There is a JavaScript implementation of OpenCL, called WebCL.

There is a JavaScript implementation of OpenGL, called WebGL.

WebCL can share data, and interoperate, with WebGL.

The GPU does not have a stack, and so the OpenCL C-ish programming language cannot do recursion and cannot make function calls. It also can’t use pointers.

The Khronos Group

http://www.khronos.org/opencl/
http://en.wikipedia.org/wiki/OpenCL

Who is Part of the Khronos Group?
Example of using OpenCL in a System-on-a-Chip: Qualcomm Node – Full Linux and OpenCL

The OpenCL Paradigm

The OpenCL Language also supports Vector Parallelism

OpenCL code can be vector-oriented, meaning that it can perform a single instruction on multiple data values at the same time (SIMD).

Vector data types are: char\( _n \), int\( _n \), float\( _n \), where \( n = 2, 4, 8, \) or 16.

```
float4 f, g;
f = (float4)(1.f, 2.f, 3.f, 4.f);
float16 a16, x16, y16, z16;
f.x = 0.;
f.xy = g.zw;
x16.s89ab = f;
float16 a16 = x16 * y16 + z16;
```

(Note: just because the language supports it, doesn’t mean the hardware does.)
Computer Graphics

Device

CU  CU  CU
CU  CU  CU

Compute Units and Processing Elements are Arranged in Grids

A GPU Device is organized as a grid of Compute Units.

Each Compute Unit is organized as a grid of Processing Elements.

So in NVIDIA terms, their Turing GPU has 68 Compute Units, each of which has 64 Processing Elements, for a grand total of 4,352 Processing Elements.

OpenCL Software Terminology:

Work-Groups and Work-Items are Arranged in Grids

An OpenCL program is organized as a grid of Work-Groups.

Each Work-Group is organized as a grid of Work-Items.

In terms of hardware, a Work-Group runs on a Compute Unit and a Work-Item runs on a Processing Element (PE).

One thread is assigned to each Work-Item.

Threads are swapped on and off the PEs.
Rules

- Threads can share memory with the other Threads in the same Work-Group
- Threads can synchronize with other Threads in the same Work-Group
- Global and Constant memory is accessible by all Threads in all Work-Groups
- Global and Constant memory is often cached inside a Work-Group
- Each Thread has registers and private memory
- Each Work-Group has a maximum number of registers it can use. These are divided equally among all its Threads

OpenCL Error Codes

This one is #define’d as zero. All the others are negative.

<table>
<thead>
<tr>
<th>Error Code</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL_SUCCESS</td>
<td></td>
</tr>
<tr>
<td>CL_DEVICE_NOT_FOUND</td>
<td>Device Not Found</td>
</tr>
<tr>
<td>CL_DEVICE_NOT_AVAILABLE</td>
<td>Device Not Available</td>
</tr>
<tr>
<td>CL_COMPILER_NOT_AVAILABLE</td>
<td></td>
</tr>
<tr>
<td>CL_MEM_OBJECT_ALLOCATION_FAILURE</td>
<td></td>
</tr>
<tr>
<td>CL_OUT_OF_RESOURCES</td>
<td></td>
</tr>
<tr>
<td>CL_OUT_OF_HOST_MEMORY</td>
<td></td>
</tr>
<tr>
<td>CL_PROFILING_INFO_NOT_AVAILABLE</td>
<td></td>
</tr>
<tr>
<td>CL_MEM_COPY_OVERLAP</td>
<td></td>
</tr>
<tr>
<td>CL_IMAGE_FORMAT_MISMATCH</td>
<td></td>
</tr>
<tr>
<td>CL_IMAGE_FORMAT_NOT_SUPPORTED</td>
<td></td>
</tr>
<tr>
<td>CL_BUILD_PROGRAM_FAILURE</td>
<td></td>
</tr>
<tr>
<td>CL_MAP_FAILURE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_VALUE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_DEVICE_TYPE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_PLATFORM</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_CONTEXT</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_QUEUE_PROPERTIES</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_COMMAND_QUEUE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_HOST_PTR</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_MEM_OBJECT</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_IMAGE_FORMAT_DESCRIPTOR</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_IMAGE_SIZE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_SAMPLER</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_BINARY</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_BUILD_OPTIONS</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_PROGRAM</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_PROGRAM_EXECUTABLE</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL_NAME</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL_DEFINITION</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL_INDEX</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL_ARGS</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_KERNEL_ARGS</td>
<td></td>
</tr>
<tr>
<td>CL_INVALID_WORK_DIMENSION</td>
<td></td>
</tr>
</tbody>
</table>

A Way to Print OpenCL Error Codes – get this from our Reference Page

```c
struct errorcode
{
    cl_int statusCode;
    char * meaning;
} ErrorCodes[] = {
    { CL_SUCCESS,                                        "\n"                                                },
    { CL_DEVICE_NOT_FOUND,                    "Device Not Found"                    },
    { CL_DEVICE_NOT_AVAILABLE,              "Device Not Available"                },
    { CL_COMPILER_NOT_AVAILABLE,            "Device Not Available"                },
    { CL_MEM_OBJECT_ALLOCATION_FAILURE | "Invalid MIP Level"                     },
    { CL_INVALID_GLOBAL_WORK_SIZE,      "Invalid Global Work Size"         },
};

void PrintCLError( cl_int errorCode,  char * prefix,  FILE *fp )
{
    const int numErrorCodes = sizeof( ErrorCodes ) / sizeof( struct errorcode );
    char * meaning = "\n";
    for( int i = 0; i < numErrorCodes; i++ )
    {
        if( errorCode == ErrorCodes[i].statusCode )
        {
            meaning = ErrorCodes[i].meaning;
            break;
        }
    }
    fprintf( fp, "%s %s\n", prefix, meaning );
}
```
// find out how many devices are attached to each platform and get their ids:
status = clGetDeviceIDs( platform, CL_DEVICE_TYPE_ALL, 0, NULL, &numDevices );
devices = new cl_device_id[ numDevices ];

// find out how many platforms are attached here and get their ids:
cl_uint numPlatforms;
status = clGetPlatformIDs( 0, NULL, &numPlatforms );
if( status != CL_SUCCESS )
fprintf( stderr, "clGetPlatformIDs failed (1)
" );
fprintf( OUTPUT, "Number of Platforms = %d
", numPlatforms );
cl_platform_id *platforms = new cl_platform_id[ numPlatforms ];
status = clGetPlatformIDs( numPlatforms, platforms, NULL );
if( status != CL_SUCCESS )
fprintf( stderr, "clGetPlatformIDs failed (2)
" );

cl_uint numDevices;
cl_device_id *devices;
for( int i = 0; i < (int)numPlatforms; i++ )
{
fprintf( OUTPUT, "Platform #%d:
", i );
size_t size;
char *str;
clGetPlatformInfo( platforms[i], CL_PLATFORM_NAME, 0, NULL, &size );
str = new char[ size ];
clGetPlatformInfo( platforms[i], CL_PLATFORM_NAME, size, str, NULL );
fprintf( OUTPUT, "Name = '%s'
", str );
delete[] str;
clGetPlatformInfo( platforms[i], CL_PLATFORM_VENDOR, 0, NULL, &size );
str = new char[ size ];
clGetPlatformInfo( platforms[i], CL_PLATFORM_VENDOR, size, str, NULL );
fprintf( OUTPUT, "Vendor = '%s'
", str );
delete[] str;
clGetPlatformInfo( platforms[i], CL_PLATFORM_VERSION, 0, NULL, &size );
str = new char[ size ];
clGetPlatformInfo( platforms[i], CL_PLATFORM_VERSION, size, str, NULL );
fprintf( OUTPUT, "Version = '%s'
", str );
delete[] str;
clGetPlatformInfo( platforms[i], CL_PLATFORM_PROFILE, 0, NULL, &size );
str = new char[ size ];
clGetPlatformInfo( platforms[i], CL_PLATFORM_PROFILE, size, str, NULL );
fprintf( OUTPUT, "Profile = '%s'
", str );
delete[] str;

status = clGetDeviceIDs( platforms[i], CL_DEVICE_TYPE_ALL, 0, NULL, &numDevices );
if( status != CL_SUCCESS )
fprintf( stderr, "clGetDeviceIDs failed (2)
" );
devices = new cl_device_id[ numDevices ];
status = clGetDeviceIDs( platforms[i], CL_DEVICE_TYPE_ALL, numDevices, devices, NULL );
if( status != CL_SUCCESS )
fprintf( stderr, "clGetDeviceIDs failed (2)
" );
for( int j = 0; j < (int)numDevices; j++ )
{
fprintf( OUTPUT, "Device #%d:
", j );
cl_device_type type;
cl_uint ui;
switch( type )
{
    case CL_DEVICE_TYPE_CPU:
fprintf( OUTPUT, "CL_DEVICE_TYPE_CPU
" );
break;
    case CL_DEVICE_TYPE_GPU:
fprintf( OUTPUT, "CL_DEVICE_TYPE_GPU
" );
break;
    case CL_DEVICE_TYPE_ACCELERATOR:
fprintf( OUTPUT, "CL_DEVICE_TYPE_ACCELERATOR
" );
break;
    default:
fprintf( OUTPUT, "Other...
" );
break;
}
switch( type )
{
    case CL_DEVICE_TYPE_CPU:
    fprintf( OUTPUT, "CL_DEVICE_TYPE_CPU
" );
break;
    case CL_DEVICE_TYPE_GPU:
    fprintf( OUTPUT, "CL_DEVICE_TYPE_GPU
" );
break;
    case CL_DEVICE_TYPE_ACCELERATOR:
    fprintf( OUTPUT, "CL_DEVICE_TYPE_ACCELERATOR
" );
break;
    default:
    fprintf( OUTPUT, "Other...
" );
break;
}
}
Typical Values from Querying the Device

Number of Platforms = 1
Platform #0:
  Name = 'NVIDIA CUDA'
  Vendor = 'NVIDIA Corporation'
  Version = 'OpenCL 1.1 CUDA 4.1.1'
  Profile = 'FULL_PROFILE'
Device #0:
  Type = 0x0004 = CL_DEVICE_TYPE_GPU
  Device Vendor ID = 0x10de
  Device Maximum Compute Units = 15
  Device Maximum Work Item Dimensions = 3
  Device Maximum Work Item Sizes = 1024 x 1024 x 64
  Device Maximum Work Group Size = 1024
  Device Maximum Clock Frequency = 1401 MHz
  Kernel Maximum Work Group Size = 1024
  Kernel Local Memory Size = 0

Querying to see what extensions are supported on this device

size_t extensionSize;
clGetDeviceInfo ( device, CL_DEVICE_EXTENSIONS, 0, NULL, &extensionSize );
char *extensions = new char [extensionSize];
clGetDeviceInfo ( devices, CL_DEVICE_EXTENSIONS, extensionSize, extensions, NULL );
fprintf ( stderr, "Device Extensions:
" );
for ( int i = 0; i < (int)strlen ( extensions ); i++ ) {
    if ( extensions[ i ] == ' ' )
        extensions[ i ] = '
';
}
fprintf ( stderr, "%s
", extensions );
delete [] extensions;

Steps in Creating and Running an OpenCL program

1. Program header
2. Allocate the host memory buffers
3. Create an OpenCL context
4. Create an OpenCL command queue
5. Allocate the device memory buffers
6. Write the data from the host buffers to the device buffers
7. Read the kernel code from a file
8. Compile and link the kernel code
9. Create the kernel object
10. Setup the arguments to the kernel object
11. Enqueue the kernel object for execution
12. Read the results buffer back from the device to the host
13. Clean everything up
#include <stdio.h>
#include <math.h>
#include <string.h>
#include <stdlib.h>
#include <omp.h>  // for timing

#include "cl.h"

1. .cpp Program Header

2. Allocate the Host Memory Buffers

```c
// allocate the host memory buffers:
float * hA = new float [ NUM_ELEMENTS ];
float * hB = new float [ NUM_ELEMENTS ];
float * hC = new float [ NUM_ELEMENTS ];

// fill the host memory buffers:
for( int i = 0; i < NUM_ELEMENTS; i++ )
{
  hA[ i ] = hB[ i ] = sqrtf( (float) i );
}

// array size in bytes (will need this later):
size_t dataSize = NUM_ELEMENTS * sizeof( float );

// opencl function return status:
cl_int status; // test against CL_SUCCESS
```

This could have also been done like this:

```c
float hA[ NUM_ELEMENTS ];
```

Global memory and the heap typically have lots more space than the stack does. So, typically, you do not want to allocate a large array like this as a local variable.

(Here, it's being done on the heap. It could also have been done in global memory.)

3. Create an OpenCL Context

```c
cl_context context = clCreateContext( NULL, 1, &device, NULL, NULL, &status );
```

4. Create an OpenCL Command Queue

```c
cl_command_queue cmdQueue = clCreateCommandQueue( context, device, 0, &status );
```
5. Allocate the Device Memory Buffers

```
// allocate memory buffers on the device:
cl_mem dA = clCreateBuffer( context, CL_MEM_READ_ONLY, dataSize, NULL, &status );
cl_mem dB = clCreateBuffer( context, CL_MEM_READ_ONLY, dataSize, NULL, &status );
cl_mem dC = clCreateBuffer( context, CL_MEM_WRITE_ONLY, dataSize, NULL, &status );
```

The read and write terminology is with respect to the OpenCL device.
So, CL_MEM_READ_ONLY means that the OpenCL device can only get this data – it
can’t send it back to the host CPU. Other options are CL_MEM_WRITE_ONLY and
CL_MEM_READ_WRITE.

6. Write the Data from the Host Buffers to the Device Buffers

```
// enqueue the 2 commands to write data into the device buffers:
status = clEnqueueWriteBuffer( cmdQueue, dA, CL_FALSE, 0, dataSize, hA, 0, NULL, NULL );
status = clEnqueueWriteBuffer( cmdQueue, dB, CL_FALSE, 0, dataSize, hB, 0, NULL, NULL );
```

Enqueuing Works Like a Conveyor Belt

```
kernel void ArrayMult(  global const float *dA,  global const float *dB,  global float *dC )
{
    int gid = get_global_id( 0 );
}
```

The .cl File

```
# Which dimension's index are we fetching?
0 = X, 1 = Y, 2 = Z
Since this is a 1D problem, X is the only index we need to get.
```
OpenCL code is compiled in the Driver . . .

7. Read the Kernel Code from a File into a Character Array

A Warning about Editing on Windows and Running on Linux

Some of you will end up having strange, unexplainable problems with your csh scripts, .cpp programs, or .cl programs. This could be because you are typing your code in on Windows (using Notepad or Wordpad or Word) and then running it on Linux. Windows likes to insert an extra carriage return (‘\r’) at the end of each line, which Linux interprets as a garbage character.

You can test this by typing the Linux command:

```
od -c loop.csh
```

which will show you all the characters, even the ‘\r’ (which you don’t want) and the ‘\n’ (newlines, which you do want).

To get rid of the carriage returns, enter the Linux command:

```
tr -d '\r' < loop.csh > loop1.csh
```

Then run loop1.csh

Or, on some systems, there is a utility called dos2unix which does this for you:

```
dos2unix < loop.csh > loop1.csh
```

Sorry about this. Unfortunately, this is a fact of life when you mix Windows and Linux.
**Something new: Intermediate Compilation**

- You pre-compile your OpenCL code with an external compiler
- Your OpenCL code gets turned into an intermediate form known as SPIR-V
- SPIR-V gets turned into fully-compiled code at runtime

**Advantages:**
1. Software vendors don’t need to ship their OpenCL source
2. Syntax errors appear during the SPIR-V step, not during runtime
3. Software can launch faster because half of the compilation has already taken place
4. This guarantees a common front-end syntax
5. This allows for other language front-ends

---

**How does that array-of-strings thing actually work?**

```c
char *ArrayOfStrings[3];
ArrayOfStrings[0] = "..one commonly-used function...
ArrayOfStrings[1] = "..another commonly-used function...
ArrayOfStrings[2] = "..the real OpenCL code...

cl_program program = clCreateProgramWithSource( context, 1, (const char**)ArrayOfStrings, NULL, &status);
```

These are two ways to provide a single character buffer:

```c
char *buffer[] = "..the entire OpenCL code...
cl_program program = clCreateProgramWithSource( context, 1, (const char**)buffer, NULL, &status);
```

```c
char *buffer = "..the entire OpenCL code...
cl_program program = clCreateProgramWithSource( context, 1, (const char**)buffer, NULL, &status);
```

---

**Why use an array of strings to hold the OpenCL program, instead of just a single string?**

1. You can use the same OpenCL source and insert the appropriate “#defines” at the beginning
2. You can insert a common header file (≈ a .h file)
3. You can simulate a “#include” to re-use common pieces of code
9. Create the Kernel Object

```c
cl_kernel kernel = clCreateKernel( program, "ArrayMult", &status );
```

10. Setup the Arguments to the Kernel Object

```c
status = clSetKernelArg( kernel, 0, sizeof(cl_mem), &dA );
status = clSetKernelArg( kernel, 1, sizeof(cl_mem), &dB );
status = clSetKernelArg( kernel, 2, sizeof(cl_mem), &dC );
```

```c
void ArrayMult(  global const float *dA,  global const float *dB,  global float *dC )
```

11. Enqueue the Kernel Object for Execution

```c
size_t globalWorkSize[ 3 ] = { NUM_ELEMENT, 1, 1 }; 
size_t localWorkSize[ 3 ] = { LOCAL_SIZE, 1, 1 }; 
clWait( cmdQueue ); // will be covered in the OpenCL event notes 
double time0 = omp_get_wtime( ); 
status = clEnqueueNDRangeKernel( cmdQueue, kernel, 1, NULL, globalWorkSize, localWorkSize, 0, NULL, NULL ); 
clWait( cmdQueue ); // will be covered in the OpenCL event notes 
double time1 = omp_get_wtime( );
```

**Work-Groups, Local IDs, and Global IDs**

- **NDRange Index Space** can be 1D, 2D, or 3D. This one is 1D.

\[
\begin{align*}
G_x &= 20 \\
W_x &= 5 \\
L_x &= 4
\end{align*}
\]

\[
#WorkGroups = \frac{GlobalIndexSpaceSize}{WorkGroupSize} = \frac{20}{4} = 5
\]
Work-Groups, Local IDs, and Global IDs

“NDRange Index Space” can be 1D, 2D, or 3D. This one is 2D.

\[
G_x = 20 \quad G_y = 12
\]

\[
L_x = 4 \quad L_y = 3
\]

\[
W_x = 5 \quad W_y = 4
\]

\[
5 \times 4 = \frac{20 \times 12}{4 \times 3}
\]

\[
L_z = 3
\]

\[
L_z = 4
\]

Work-Groups, Local IDs, and Global IDs

“NDRange Index Space” can be 1D, 2D, or 3D. This one is 3D.

Figuring Out What Thread You Are and What Your Thread Environment is Like

- `uint get_work_dim();`
- `size_t get_global_size(uint dimindx);`
- `size_t get_global_id(uint dimindx);`
- `size_t get_local_size(uint dimindx);`
- `size_t get_local_id(uint dimindx);`
- `size_t get_num_groups(uint dimindx);`
- `size_t get_group_id(uint dimindx);`
- `size_t get_global_offset(uint dimindx);`

0 ≤ dimindx ≤ 2

12. Read the Results Buffer Back from the Device to the Host

status = clEnqueueReadBuffer( cmdQueue, dC, CL_TRUE, 0, dataSize, hC, 0, NULL, NULL );
13. Clean Everything Up

```c
// clean everything up:
clReleaseKernel( kernel );
clReleaseProgram( program );
cReleaseCommandQueue( cmdQueue );
cReleaseMemObject( dA );
cReleaseMemObject( dB );
cReleaseMemObject( dC );
delete[] hA;
delete[] hB;
delete[] hC;
```

13. Clean Everything Up

Do this because we created these arrays with new

---

Array Multiplication Performance: What is a Good Work-Group Size?

<table>
<thead>
<tr>
<th>Array Size (K)</th>
<th>Work-Group Size</th>
<th>GigaMultiplications/Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>1024</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>2048</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>4096</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>8192</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

---

Writing out the .cl Program’s Binary Code

```c
size_t binary_sizes;
status = clGetProgramInfo( Program, CL_PROGRAM_BINARY_SIZES, 0, NULL, &binary_sizes );
size_t size;
status = clGetProgramInfo( Program, CL_PROGRAM_BINARY_SIZES, sizeof(size_t), &size, NULL );
unsigned char *binary = new unsigned char[size];
status = clGetProgramInfo( Program, CL_PROGRAM_BINARIES, size, &binary, NULL );
FILE *fpbin = fopen( "particles.nv", "wb" );
if( fpbin == NULL ) {
    fprintf(stderr, "Cannot create 'particles.bin'
    } else {
        fwrite( binary, 1, size, fpbin );
        fclose( fpbin );
    }
delete[] binary;
```
Importing that Binary Code back In:
8. Compile and Link the Kernel Code

Instead of doing this:
```c
char *strings[1] = 
cl_program program = clCreateProgramWithSource( context, 1, (const char**)strings, NULL, &status );
delete [] clProgramText;
```

You would do this:
```c
unsigned char byteArray[ numBytes ];
cl_program program = clCreateProgramWithBinary( context, 1, &device, &numBytes, &byteArray, &binaryStatus, &status );
delete [] byteArray;
```

And you still have to do this:
```c
char *options = 
size_t size;
clGetProgramBuildInfo( program, device, CL_PROGRAM_BUILD_LOG, 0, NULL, &size );
cl_char *log = new cl_char[ size ];
clGetProgramBuildInfo( program, device, CL_PROGRAM_BUILD_LOG, size, log, NULL );
delete [] log;
```