How OpenCL enables easy access to FPGA performance?

Suleyman Demirsoy
Agenda

- Introduction
- OpenCL Overview
  - S/W Flow
  - H/W Architecture
- Product Information & design flow
- Applications
- Additional Collateral
Introduction
The Quest for Performance

Programmability

CPU

- Single-Core C/C++
- Multi-Core AVX/OpenMP

GPGPU

- Stream CUDA
- Driver API

Heterogeneous Programming

- Architecture PCIe Accelerator
- Programming Language OpenCL

Performance

© 2013 Altera Corporation—Public
✓ Massive Parallelism
- Millions of logic elements
- Thousands of 20Kb memory blocks
- Thousands of Variable Precision DSP blocks
- Dozens of High-speed transceivers

✗ Hardware-centric
- VHDL/Verilog
- Synthesis
- Place&Route
OpenCL Overview
OpenCL (Open Computing Language) Overview

- **Software programming model:**
  - C/C++ API for host program
  - OpenCL C for acceleration device

- **Provides increased performance with hardware acceleration**
  - CPU offload to appropriate accelerator
    - Local Memory
    - Explicit Parallelism
      - Task (SMT)
      - Data (SPMD)

- **Open, royalty-free, standard**
  - Managed by Khronos Group
  - Altera active member
  - Conformance requirements
    - V1.0 is current reference
    - V2.0 is current release
  - [http://www.khronos.org](http://www.khronos.org)
Altera OpenCL Program Overview

- **2010 research project**
  - Toronto Technology Center

- **2011 Development started**
  - Proof of concept
  - 9 customer evaluations

- **2012 Early Access Program**
  - Demo’s at Supercomputing ‘12
  - Over 60 customer evaluations

- **2013 First public release**
  - Public announcement with release
  - Passed v1.0 conformance

- **Public release v13.1**
  - **Installation** image accessible from ACDS download infrastructure
  - **Documentation** available online
  - **Boards** available from vendor web site and ACDS installation
  - **Support** flow in place
  - **Optimization** improvements
  - **SoC** support
  - **Design Examples** on Altera.com
Passed OpenCL Conformance!

- First FPGA to pass OpenCL conformance
  - OpenCL v1.0 specification
- >8500 Programs tested

Conformant Products

<table>
<thead>
<tr>
<th>Company</th>
<th>Date</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altera Corporation</td>
<td>2013-08-03</td>
<td>OpenCL_1.0</td>
</tr>
<tr>
<td>Nallatech PCIe-385N (Altera Stratix V A7)</td>
<td>2013-07-23</td>
<td>OpenCL_1.2</td>
</tr>
</tbody>
</table>

http://www.khronos.org/conformance/adopters/conformant-companies
http://www.khronos.org/conformance/adopters/conformant-products
Heterogeneous Platform Model

OpenCL Platform Model

Host

Global Memory

Host Memory

(Compute) Device

Compute Unit

Processing Element

Example Platform

x86

PCIe
Heterogeneous Platform Model

OpenCL Platform Model

- Host
- Global Memory

Device

- CU (Compute Unit)

Example Platform

- x86

PCIe
main() {
    read_data( ... );
    manipulate( ... );
    clEnqueueWriteBuffer( ... );
    clEnqueueNDRange(..., sum, ...);
    clEnqueueReadBuffer( ... );
    display_result( ... );
}

__kernel void sum
    (__global float *a,
     __global float *b,
     __global float *y)
{
    int gid = get_global_id(0);
    y[gid] = a[gid] + b[gid];
}
Use Model: clCreateProgramWithBinary

```c
fp = fopen("file.aocx","rb");
fclose(fp);
free((void*)fp);
```

```
const char** program(const char** binaries, long lengths[], int length);
```

```
clCreateProgramWithBinary(const char** program, const char** binaries, long lengths[], int length);
```

```
clBuildProgram(cl_program, Program (exe));
```

```
clGetPlatforms(
    cl_platform
);
```

```
clGetDevices(cl_platform, cl_device);
```

```
clCreateContext(cl_context, cl_platform);
```

```
clCreateCommandQueue(cl_context, cl_device, cl_command_queue);
```

```
clEnqueueNDRangeKernel(cl_command_queue, Kernel (src), cl_program, exe);
```

```c
typedef struct {
    void* program;
    const char** binaries;
    long lengths[];
    int length;
} Program;
```

```
clCreateKernel(cl_program, Kernel (src), exe);
```

```
clEnqueueNDRangeKernel(cl_command_queue, Kernel (src), cl_program);
```

```
clGetPlatforms(cl_platform);
```

```
clGetDevices(cl_platform, cl_device);
```

```
clCreateContext(cl_context, cl_platform);
```

```
clCreateCommandQueue(cl_context, cl_device, cl_command_queue);
```

```
clEnqueueNDRangeKernel(cl_command_queue, Kernel (src), cl_program, exe);
```

```
clCreateKernel(cl_program, Kernel (src), exe);
```

```
clEnqueueNDRangeKernel(cl_command_queue, Kernel (src), cl_program);
```

```
CL File ≈ OpenCL “Program” ≈ Bitstream
```
# Reference Platforms

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Network Enabled</th>
<th>High Performance Computing (HPC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>Half-Size</td>
<td>Full-Size</td>
</tr>
<tr>
<td>Component</td>
<td>Single</td>
<td>Dual</td>
</tr>
<tr>
<td>Global Memory</td>
<td>DDR3-1600 and QDRII+ 550MHz</td>
<td>DDR3-1333/FPGA</td>
</tr>
<tr>
<td>IO Channels</td>
<td>2x10GbE (MAC/UOE)</td>
<td>None (Minimize IP overhead)</td>
</tr>
<tr>
<td>Reference Design</td>
<td>• OPRA (Streaming) • Trading (with global memory access)</td>
<td>• Option Pricing</td>
</tr>
</tbody>
</table>
Altera HPC Reference Platform for OpenCL

C/C++ API

OpenCL C

host.c
device.cl

Compiler

Software Layer

OpenCL.h
AOCL install
AOCL flash
AOCL program
AOCL diagnose

HAL
MMD

Hardware Layer

Interconnect

SV: CvP Update

SV D8

PCIe gen2x8

Reference Platform

Reference Design

Reference Board

Host

Device

64-bit
- RHEL 6.4
- Windows 7

s5_hpc (S5PE-DS)

18
Altera Network Enabled Reference Platform for OpenCL

C/C++ API

OpenCL C

host.c
device.cl

Compiler

Software Layer

Hardware Layer

OpenCL.h
AOCL install
AOCL flash
AOCL program
AOCL diagnose

HAL
MMD

SV: CvP Update

Interconnect

10Gbe
MAC/UEF
10Gbe
MAC/UEF

s5_hft (S5PH-Q)

SV D8

Reference Design

Reference Platform

Reference Board

Host

Device

64-bit
- RHEL 6.4
- Windows 7
Altera Network Enabled Reference Platform for OpenCL

- **Device C/C++ API**

  - Software Layer
  - Hardware Layer
  - Compiler
  - OpenCL

  **Reference Platform**
  - **Reference Board**
  - **Reference Design**

**Host**

- **10G Network**

- **PCIe gen2x8 Host Interface**

**Interconnect**

- **DDR3 Memory Interface**
- **QDRII Memory Interface**
- **10Gb MAC/UEO Data Interface**

**OpenCL C**

- **CvP Update**

- **OpenCL Kernels**

**Host**

- 64-bit
  - RHEL 6.4
  - Windows 7

- **s5_hft (S5PH-Q)**
OpenCL Modular Hardware Architecture

- DDR
- DDR
- QDR
- QDR
- QDR
- QDR
- QDR
- DDR
- DDR
- QDR
- QDR
- 10G Network
- Host

Interconnect:
- DDR3 Memory Interface
- DDR3 Memory Interface
- QDR II Memory Interface
- QDR II Memory Interface
- QDR II Memory Interface
- QDR II Memory Interface
- 10Gb MAC/UOE Data Interface
- 10Gb MAC/UOE Data Interface
- PCIe gen2x8 Host Interface

CvP Update

OpenCL Kernels

OpenCL Kernels

Prebuilt BSP with standard HDL tools

Built with Altera OpenCL Compiler

© 2013 Altera Corporation—Public
OpenCL on SoC Platform – Single Chip Solution

- **Lightweight bridge**
  - Starting/stopping kernel, reconfiguring PLL, etc…

- **FPGA to SDRAM bridge**
  - Default way to move data between HPS and FPGA
  - 256bits wide @ 100Mhz ~ 2.6GB/s

- **FPGA external memory**
  - Scratch ram for FPGA kernels
    - Store intermediate data before passing to next kernel
  - FPGA to HPS and HPS to FPGA bridges are connected to it as a secondary connection
    - Very slow: 32bits @50Mhz w/out DMA
The Key to Performance

Maximize Throughput
Minimize Latency

More Operations Per Second

Parallelism
- Pipelining
  - Instructions
  - Processes
- Loop unrolling
- Duplication (SPMD)
- Multi-threading (SMT)

Quick Data Access

Memory Access
- Avoid transfer/copy
- Work in local memory instead of shared memory
- Coalesce accesses
Altera Platform
- Multiple Devices/Board
- Multiple Boards/Host
Memories with different characteristics
- DDR
  - Sequential Access
- QDR
  - Random Access

Attribute-based
Standard OpenCL

Vendor Extension Channels
Interfaces to compiler

- Host CPU Interface: *Avalon Memory Map*
- Global Memory: *Avalon Memory Map*
- Option IO: *Avalon Streaming*
OpenCL + FPGA Key Benefits

- **Higher performance/watt vs. CPU/GPGPU**
  - Offload performance-intensive functions from the host processor to an FPGA
    - Implement exactly what you need
    - Pipeline parallel structures
    - Custom interconnect converging with data processing cores

- **Lower power vs. CPU/GPGPU**
  - Core frequency lower: 200-250MHz vs 1GHz
  - Turn off unused logic
  - Up to 1/5 the power

- **Faster development vs. traditional FPGA design flow**
  - Higher level of design abstraction

- **Higher department productivity**
  - Leverage your software development team
  - Familiar C-based design entry

- **Portability & Obsolescence free**
  - Code can transfer between different HW accelerators (CPU, GPGPU, FPGA, etc)
  - FPGA life cycle considerably longer than CPUs or GPGPUs
Product Information & Design Flow
How to Get OpenCL

- Part of ACDS v13.1 installation
- Requires licensed Quartus software
- Supported on Windows and Linux
- Still need standard GCC compiler for host side code
  - Visual Studio, Eclipse... etc.
What is included with the Altera SDK for OpenCL?

- **Offline compiler (aoc)**
  - GCC based model

- **Altera OpenCL utility (aocl)**
  - Diagnostics for board installation
  - Flash or program FPGA image
  - Install board drivers (typically PCIe)

- **Host libraries**
  - Required by host code and provided by the vendor (Altera)

- **APB Board driver**

- **Design examples**
  - FFT, vectoradd, matrixmult, moving average
Altera SDK for OpenCL Product Details

- **Altera SDK for OpenCL Licensing**
  - Purchase a 1 year perpetual license
  - Fixed & float available
  - 60-day evaluation license available on request
  - Requires Quartus II v13.1 Subscription Edition or Development Kit Edition

- **OS**
  - Microsoft 64-bit Windows 7
  - Red Hat Enterprise 64-bit Linux (RHEL) 6.x

- **Memory requirements**
  - SDK: Computer equipped with at least 16 GB RAM
  - Quartus II: Refer to memory requirements for target FPGA
ALTERA®

Public

Altera Preferred Board Partner Program for OpenCL

■ Provide customers with a portfolio of COTS boards to evaluate, develop and go-to-production with

■ Customers can develop code and target that preferred board
  – Altera SDK for OpenCL flow has been verified on the board
  – Ensures an exceptional out-of-box customer experience for the customer

■ Customer purchase directly from partners

■ Altera’s Preferred Board include:
  – Includes Quartus II Development Kit Edition Software (one year license)
  – Includes an Altera SDK for OpenCL License (one year, perpetual license)
### APBs Available as of 13.1 Release

<table>
<thead>
<tr>
<th>Partner</th>
<th>Board</th>
<th>Altera Device</th>
<th>Where to Get</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altera</td>
<td><strong>DK-DEV-5CSXC6NES</strong></td>
<td>Cyclone V SX SoC</td>
<td>Part of ACDS 13.1</td>
</tr>
<tr>
<td>BittWare</td>
<td><strong>S5PH-Q</strong></td>
<td>Stratix V D5</td>
<td>Part of ACDS 13.1</td>
</tr>
<tr>
<td></td>
<td><strong>S5PH-Q</strong></td>
<td>Stratix V D8</td>
<td>Part of ACDS 13.1</td>
</tr>
<tr>
<td></td>
<td><strong>S5PH-Q</strong></td>
<td>Stratix V A7</td>
<td>Contact BittWare</td>
</tr>
<tr>
<td></td>
<td><strong>S5PH-Q</strong></td>
<td>Stratix V AB</td>
<td>Contact BittWare</td>
</tr>
<tr>
<td></td>
<td><strong>S5PH-DS</strong></td>
<td>Dual Stratix V AB</td>
<td>Contact BittWare</td>
</tr>
<tr>
<td>Terabox</td>
<td>8, <strong>S5PH-DS Boards</strong></td>
<td></td>
<td>Contact BittWare</td>
</tr>
<tr>
<td>Nallatech</td>
<td><strong>385-A7 Accelerator Card</strong></td>
<td>Stratix V A7</td>
<td>Part of ACDS 13.1</td>
</tr>
<tr>
<td></td>
<td><strong>385-D5 Accelerator Card</strong></td>
<td>Stratix V D5</td>
<td>Part of ACDS 13.1</td>
</tr>
<tr>
<td>PLDA</td>
<td><strong>XP5S620LP-40G</strong></td>
<td>Stratix V A7</td>
<td>Contact PLDA</td>
</tr>
<tr>
<td>Terasic</td>
<td><strong>DE5</strong></td>
<td>Stratix V A7</td>
<td>Contact Terasic</td>
</tr>
</tbody>
</table>

© 2013 Altera Corporation—Public

**ALTERA**
MEASURABLE ADVANTAGE™
Altera SDK for OpenCL Design Flow

**Getting Started Guide** (document)
- Install Quartus II v13.1 with Altera SDK for OpenCL
- Install C Compiler or Development Environment
- Obtain and setup license from the Self Service Licensing Center
- Install the FPGA (OpenCL) board

```
aocl install
```

**Programming Guide** (document)
- Develop kernel code and compile on CPU/GPU for functional correctness
- Build, compile & link the host application (Visual Studio/GCC)
- Compile the OpenCL kernel with Altera offline Compiler (aoc)
- Run the application

**Optimization Guide** (document)
- Optimize kernel for FPGA hardware

Set Up

Design

Optimize
Applications
AES Encryption

- **Encryption/decryption**
  - 256bit key
  - Counter (CTR) method

- **Advantage FPGA**
  - Integer arithmetic
  - Coarse grain bit operations
  - Complex decision making

- **Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (GB/s)</th>
<th>Efficiency (GB/s/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5503 Xeon Processor (single core)</td>
<td>est 80</td>
<td>0.01</td>
<td>1.25e-4</td>
</tr>
<tr>
<td>AMD Radeon HD 7970</td>
<td>est 100</td>
<td>0.33</td>
<td>3.30e-3</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>25</td>
<td>5.20</td>
<td>2.08e-1</td>
</tr>
</tbody>
</table>

© 2013 Altera Corporation—Public
Multi-Asset Barrier Option Pricing

- **Monte-Carlo simulation**
  - No closed form solution possible
  - High quality random number generator required
  - Billions of simulations required

- **Used GPU vendors example code**

- **Advantage FPGA**
  - Complex Control Flow

- **Optimizations**
  - Channels, loop pipelining

- **Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (Bsims/s)</th>
<th>Efficiency (Msims/s/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3690 Xeon Processor</td>
<td>130</td>
<td>.032</td>
<td>0.0025</td>
</tr>
<tr>
<td>nVidia Kepler20</td>
<td>212</td>
<td>10.1</td>
<td>48</td>
</tr>
<tr>
<td>Bittware S5-PCle-HQ</td>
<td>45</td>
<td>12.0</td>
<td>266</td>
</tr>
</tbody>
</table>

© 2013 Altera Corporation—Public
Document Filtering

- **Unstructured data analytics**
  - Bloom Filter

- **Advantage FPGA**
  - Integer Arithmetic
  - Flexible Memory Configuration

**Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (MTs)</th>
<th>Efficiency (MTs/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3690 Xeon Processor</td>
<td>130</td>
<td>2070</td>
<td>15.92</td>
</tr>
<tr>
<td>nVidia Tesla C2075</td>
<td>215</td>
<td>3240</td>
<td>15.07</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>25</td>
<td>3602</td>
<td>144.08</td>
</tr>
</tbody>
</table>

© 2013 Altera Corporation—Public
Consumer (Japan)

- **Image Processing**
  - Adaptive weighted images

\[
p_{xy} = \sum \frac{c_1 d_1 + c_2 d_2 + c_2 d_2}{W}
\]

- **Advantage FPGA**
  - Integer Arithmetic

- **Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (FPS)</th>
<th>Efficiency (FPS/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3565 Xeon Processor</td>
<td>est 130</td>
<td>0.05</td>
<td>.0004</td>
</tr>
<tr>
<td>nVidia Quadro 4000</td>
<td>est 150</td>
<td>2.94</td>
<td>.0200</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>21</td>
<td>4.29</td>
<td>.2040</td>
</tr>
</tbody>
</table>

© 2013 Altera Corporation—Public
Smith-Waterman

- **Sequence Alignment**
  - Scoring Matrix

\[
H(i, 0) = 0, \ 0 \leq i \leq m \\
H(0, j) = 0, \ 0 \leq j \leq n \\
\text{if } a_i = b_j \text{ then } w(a_i, b_j) = w(\text{match}) \text{ or if } a_i \neq b_j \text{ then } w(a_i, b_j) = w(\text{mismatch}) \\
H(i, j) = \max \left\{ \begin{array}{l}
H(i-1, j-1) + w(a_i, b_j) \\
H(i-1, j) + w(a_i, \_)
\end{array} \right\}, \ 1 \leq i \leq m, 1 \leq j \leq n
\]

- **Advantage FPGA**
  - Integer Arithmetic
  - SMT Streaming

- **Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (MCUPS)</th>
<th>Efficiency (MCUPS/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3565 Xeon Processor</td>
<td>140</td>
<td>40</td>
<td>.29</td>
</tr>
<tr>
<td>nVidia K20</td>
<td>225</td>
<td>704</td>
<td>3.13</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>25</td>
<td>32596</td>
<td>1303.00</td>
</tr>
</tbody>
</table>
**Multi Function Printer**

- **Image Processing**
  - RGB output of raster scanner converted to CMYK colorants for printing

- **Advantage FPGA**
  - SoC Solution
  - IO and Kernel Channels
  - Heterogeneous memory accesses

- **Goal 50PPM at A4/letter size**

- **Results**
  - >40X improvement over C based algorithm on ARM only
    - No NEON coprocessor used
  - C6 speed grade part improved 20% to 128PPM
Additional Resources
Additional Altera Collateral

- White papers on OpenCL
- OpenCL online demos
- OpenCL design examples
- Instructor-Led training
  - OpenCL for Altera FPGAs Training by Acceleware – (4 Day)
  - Parallel Computing with OpenCL Workshop by Altera – (1 Day)
  - Optimization of OpenCL for Altera FPGAs Training by Altera – (1 Day)
- Online training
  - Introduction to Parallel Computing with OpenCL
  - Writing OpenCL Programs for Altera FPGAs
  - Running OpenCL on Altera FPGAs
- OpenCL board partners page
Summary

- **Productivity**
  - Unified software programmer friendly design environment for a variety of devices, now including FPGA, in a heterogeneous platform

- **Performance**
  - Excellent throughput and latency for algorithms pushing SMT limits, and SPMD with large local memory demands

- **Efficiency**
  - Dedicated custom processors for the parallel tasks make for the most compelling performance/Watt results

- **Cost**
  - SoC solution with host and accelerator in a single device creates a simpler system and can lowers system costs for real-time performance acceleration