CSCI-GA.3033-004

Graphics Processing Units (GPUs): Architecture and Programming

More on GPU Architecture

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com
On the motherboard
PCI-e

- Developed by Intel
- Peripheral Component Interconnect
- PCI Express architecture is a high performance, IO interconnect for peripherals.
- A serial point-to-point interconnect between two devices
- PCI-Express slots also accepts older PCI cards
- Data sent in packets
- Synchronous
- No shared bus but a shared switch
PCI-e

- Data is sent serially in packets through pairs of transmit and receive signals called lanes.
- Each lane enables 250 MBytes/s bandwidth per direction.
- Multiple lanes can be grouped together into x1 (“by-one”), x2, x4, x8, x12, x16, and x32 lane widths to increase bandwidth to the slot.
Bus Example: PCI-e

<table>
<thead>
<tr>
<th>Link Width</th>
<th>x1</th>
<th>x2</th>
<th>x4</th>
<th>x8</th>
<th>x12</th>
<th>x16</th>
<th>x32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregate BW (GBytes/s)</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>
Source: National Instruments
PCIe Card

• Is any device connected to PCIe bus

Graphics card PCIe (x16)
Source: National Instrument
Some Hardware Advances

- NVLINK
- Stacked Memory
NVLINK

- By NVIDIA
- High-bandwidth, energy-efficient interconnect
- enables ultra-fast communication between:
  - CPU and GPU
  - GPUs
- Allows data sharing at rates 5 to 12 times faster than the traditional PCIe.
- Complementing PCIe, not replacing it
NVLINK

Stacked Memory

3D instead of 2D

Stacked Memory

• Higher chip capacity
• Increases bandwidth not latency, in general
• Different flavors:
  – Hybrid memory cubes
  – High bandwidth memory
Future from NVIDIA

• Pascal  ~2016
• Volta  ~2018
PASCAL  $\Rightarrow$ 10x MAXWELL performance

- Equipped with NVLINK
  - $\sim 200$GB/s per NVLINK
- stacked memory:
  - 16 GB and up to 32 GB of HBM2 memory
  - a massive 4096bit memory interface
  - 4 stacks
  - so, getting away from GDDR5
- TSMC 16nm technology
- NVIDIA has designed a module to house Pascal at one-third the size of the standard boards used today.
  - They’ll put the power of GPUs into more compact form factors than ever before.
NVIDIA Volta

- Expected to double the maximum memory capacity to 64GB
- ~2018 for mass market
- ~2017 for first supercomputer
  - the Summit from Oak Ridge National Laboratory
  - Sierra from Lawrence Livermore National Laboratory.
  - Both of these supercomputers will feature several next generation IBM POWER9 CPUs and also several NVIDIA Volta GPUs.
How About AMD?
AMD

• heterogeneous compute compiler (HCC) for C++ programming

• for applications already developed in CUDA, they can now be ported into C++.  
  – This is achieved using the new Heterogeneous-computing Interface for Programmers (HIP) tool that ports CUDA runtime APIs into C++ code.
Accelerated Processing Units (APU)

• combine a CPU and GPU onto the same chip
• often intended for smaller and more energy-efficient machines
• Examples: Sempron, Athlon, A-series
• The architecture of AMD GPU is called Graphics-core next (GCN)
  – Used in all product lines from low-end to high-end
### AMD Radeon R9 Fury X

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>28nm</td>
</tr>
<tr>
<td>Stream Processors</td>
<td>4096</td>
</tr>
<tr>
<td>Compute Units$^{14}$</td>
<td>64</td>
</tr>
<tr>
<td>Engine Clock</td>
<td>Up to 1050MHz</td>
</tr>
<tr>
<td>Compute Performance</td>
<td>8.6 TFLOPS</td>
</tr>
<tr>
<td>Texture Units</td>
<td>256</td>
</tr>
<tr>
<td>Texture Fill-Rate</td>
<td>268.8 GT/s</td>
</tr>
<tr>
<td>ROPs</td>
<td>64</td>
</tr>
<tr>
<td>Pixel Fill-Rate</td>
<td>67.2 GP/s</td>
</tr>
<tr>
<td>Z/Stencil</td>
<td>256</td>
</tr>
<tr>
<td>Memory Configuration</td>
<td>4GB HBM</td>
</tr>
<tr>
<td>Memory Interface</td>
<td>4096-bit</td>
</tr>
<tr>
<td>Memory Speed / Data Rate</td>
<td>500MHz / 1.0Gbps</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>Up to 512 GB/s</td>
</tr>
<tr>
<td>Power Connectors</td>
<td>2 x 8-pin</td>
</tr>
<tr>
<td>Typical Board Power</td>
<td>275W</td>
</tr>
<tr>
<td>PCIe® Standard</td>
<td>PCIe® 3.0</td>
</tr>
</tbody>
</table>
AMD Radeon R9 Fury X

Fiji GPU Architecture (AMD's Graphics Core Next (GCN) microarchitecture.)

CU: Compute Unit
An now ... Intel
<table>
<thead>
<tr>
<th>Microarchitecture</th>
<th>Process Node</th>
<th>Tick or Tock</th>
<th>Release Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conroe/Merom</td>
<td>65nm</td>
<td>Tock</td>
<td>2006</td>
</tr>
<tr>
<td>Penryn</td>
<td>45nm</td>
<td>Tick</td>
<td>2007</td>
</tr>
<tr>
<td>Nehalem</td>
<td>45nm</td>
<td>Tock</td>
<td>2008</td>
</tr>
<tr>
<td>Westmere</td>
<td>32nm</td>
<td>Tick</td>
<td>2010</td>
</tr>
<tr>
<td>Sandy Bridge</td>
<td>32nm</td>
<td>Tock</td>
<td>2011</td>
</tr>
<tr>
<td>Ivy Bridge</td>
<td>22nm</td>
<td>Tick</td>
<td>2012</td>
</tr>
<tr>
<td>Haswell</td>
<td>22nm</td>
<td>Tock</td>
<td>2013</td>
</tr>
<tr>
<td>Broadwell</td>
<td>14nm</td>
<td>Tick</td>
<td>2014</td>
</tr>
<tr>
<td>Skylake</td>
<td>14nm</td>
<td>Tock</td>
<td>2015</td>
</tr>
<tr>
<td>Kaby Lake</td>
<td>14nm</td>
<td>Tock</td>
<td>2016</td>
</tr>
</tbody>
</table>
Source: The Compute Architecture of Intel® Processor Graphics Gen9 Whitepaper, August 2015
Intel’s Skylake Microarchitecture

- Increased chipset I/O throughput, Tablet I/Os, Audio DSP Upgrade, Sensor Hub
- Higher resolution display
- Bigger/wider core, better instruction per clock, improved power efficiency
- Enhanced ring/LLC for improved throughput
- Integrated camera ISP
- Extended overclocking capabilities
- Faster DDR Memory
- Advanced Processor Graphics GT3 + eDRAM, GT4 + eDRAM; OpenCL™ 2.0 API, DirectX® 12, OpenGL® 4.4

Intel Next Generation Microarchitecture Code Name Skylake
Source: The Compute Architecture of Intel® Processor Graphics Gen9 Whitepaper, August 2015
Intel Gen9 Embedded GPU

- Compute components called execution units (EU).
- Execution units are clustered into subslices.
- Subslices are gathered into slices.

**EU: Execution Unit**

- **Instruction Fetch**
  - 28KB GRF: 7 thrs x 128x SIMD8 x 32b

- **Thread Arbiter**
  - Send
  - Branch

- **GRF**: general purpose register file
- **ARF**: Architecture register file
- **FPU**: Floating point unit
  - But supports both FP and integer ops.
EU

• SIMD ALUs
  – pipelined across multiple threads
• A combination of:
  – simultaneous multi-threading (SMT)
  – fine-grained interleaved multi-threading (IMT).
• For gen9-based products there are 7 threads
• Each EU thread:
  – has 128 general purpose registers
  – Each register stores 32 bytes
  – Accessible as a SIMD 8-element vector of 32-bit data elements.
  – Thus each gen9 thread has: 4 Kbytes of general purpose register file (GRF).
EU

- Depending on the software workload, the hardware threads within an EU may
  - all be executing the same compute kernel code
  - each EU thread could be executing code from a completely different compute kernel

- On every cycle:
  - an EU can co-issue up to four different instructions
  - must be sourced from four different threads.
  - The EU’s thread arbiter dispatches these instructions to one of four functional units for execution

- Each FPU can SIMD:
  - execute up to four 32-bit floating-point (or integer) operations
  - or execute up to eight 16-bit integer or 16-bit floating-point operations.
Subslice

- Each subslice contains its own local-thread dispatcher
- The **sampler** is a read-only memory fetch unit.
Slice

- Shared local memory:
  - a structure within the L3 complex
  - supports programmer-managed data
  - for sharing among EU hardware threads within the same subslice
Conclusions

- GPU moves to more general purpose
- GPUs are becoming the main part in many supercomputer
- Players are now tackling the main issues in GPUs:
  - communication: NVLINK
  - memory: 3D stacking
- We are still waiting for a bit friendlier programming model!