Data Acceleration: Challenges & Opportunities

Scott Schweitzer
Achronix Semiconductors
Agenda

• **Challenges**
  - Hardware
    - Bandwidth
    - Queueing
    - Available On Chip B/W
  - Software
    - Programmability
    - Security

• **Opportunities**
  - Hardware
    - Computational Scaling
    - Memory I/O
    - Network on Chip Bandwidth
  - Software
    - Programmability
    - Security
Challenge is Bandwidth

Context: SmartNICs, DPUs & IPUs

<table>
<thead>
<tr>
<th>NIC Interface</th>
<th>Year*</th>
<th>Speed (GB/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10GbE x 2</td>
<td>2005</td>
<td>2.5</td>
</tr>
<tr>
<td>PCIe Gen 1x8</td>
<td>2005</td>
<td>2.0</td>
</tr>
<tr>
<td>PCIe Gen 2x8</td>
<td>2006</td>
<td>4.0</td>
</tr>
<tr>
<td>40GbE x 2</td>
<td>2010</td>
<td>10</td>
</tr>
<tr>
<td>PCIe Gen 3x8</td>
<td>2010</td>
<td>8</td>
</tr>
<tr>
<td>25GbE x 2</td>
<td>2015</td>
<td>6.25</td>
</tr>
<tr>
<td>PCIe Gen 3x16</td>
<td>2016</td>
<td>16</td>
</tr>
<tr>
<td>100GbE x 2</td>
<td>2017</td>
<td>25</td>
</tr>
<tr>
<td>PCIe Gen 4x16</td>
<td>2020</td>
<td>32</td>
</tr>
<tr>
<td>400GbE x 2</td>
<td>2022</td>
<td>100</td>
</tr>
<tr>
<td>PCIe Gen 5x16</td>
<td>2022</td>
<td>64</td>
</tr>
</tbody>
</table>
OPPORTUNITY – COMPUTATIONAL SCALING

**GPU**
- NVIDIA GeForce RTX 3080
- 8,704 CUDA Cores @ 1.44-1.71GHz
- 272 Tensor Cores
- 10GB GDDR6X RAM
- 320 bit Internal Bus Width
- PCIe Gen 4x16

**FPGA**
- Bittware VectorPath
- 692K 6-input lookup tables (LUTs)
- 2,560 Machine Learning Processors
- 16B GDDR6 RAM
- 256 bit Internal Bus Width
- PCIe Gen 4x16

**DPU**
- NVIDIA BlueField-2 DPU
- 16 Arm Cores, 256 Threads @ 3GHz
- 16GB DDR5 RAM
- PCIe Gen 4x32

ODSA Workshop 2021
Challenges - Queueing

- Classification
- Encryption
- Encoding
- Inspection
- Compression
- NVMe
- Action
Opportunity – Memory I/O

DDR5

GDDR6

HBM2
Challenge – On-Chip BW

Classification ➔ Encryption ➔ Encoding ➔ Inspection ➔ Action ➔ NVMe ➔ Compression

ODSA Workshop 2021
Opportunity – High Speed Network on Chip
Programmability

- **Challenges**
  - Tools
  - Interfaces
  - Timing
  - Queueing

- **Opportunities**
  - Open Source
  - CXL
  - Timing Flat
  - >>Memory
Security – Challenges & Opportunities

• Where and how to store secrets
  • Trusted Platform Module (TPM)
  • Onboard Secure Root of Trust
  • Physically Uncloneable Functions (PUF)

• Which accelerated libraries to include and at what speeds

• Side Channel Attacks (SCA)
  • SCA Counter Measures
The Exciting Future of Chiplets & Accelerators

• Standardization
  • Packaging issues: dimensions, thermal, connection and pin placement
  • Interface issues: signaling, common control plane or security context
  • Protocols between chiplets and to the host
  • Documentation
  • Tools

• Ecosystem, market place ➔ tools and flows, simulation, analysis
  • Imagine issues stacking several complex accelerators in the same package
  • Thermal, magnetic, signal interference