Cloud and FPGAs in Physics and HPC
from Data Acquisition to the Cloud

Andrew Putnam
July 22, 2022
FPGAs in Cosmology

EOR Science can be done with a paperclip and a supercomputer

-- Don C. Backer

Cosmologists often refer to their telescopes as “software telescopes”
FPGAs in Physics Applications

Nhan Tran (FermiLab), Philip Harris (MIT), Javier Duarte (UCSD)
Teaching Old Technology New Tricks

• Physicists are familiar for data acquisition and near-sensor processing
• You’re going to have an FPGA developer on the project...
• But what else can you do with FPGAs?
Catapult: Long, Fruitful FPGA Investment

**Catapult v1: Mt Granite**
- Distributed solution
- Integrated with WCS (OCP) 1.0

**v0: Research POC**
- Built v0 board w/ 6 Xilinx FPGAs
- 30k lines of Bing code on FPGA

**v1: Scale Pilot**
- 1632 servers deployed
- Bing IndexServe accelerated

**v2: Pikes Peak**
- Integrated Bing + Azure design
- Bump-in-the-wire introduced

**v2 Production and ramp**
- FPGAs reach production
- Deployed in all new servers

**Azure AccelNet Unveiled**
- Azure production launch
- AI Supercomputer demo

**Project BrainWave / Storm Peak**
- Real-Time AI
- First 3rd Party FPGA Service

**Catapult v3: Longs Peak**
- DNN Platform for Bing
- 50Gb w/ integrated NIC

**Overlake: Celestial Peak**
- 100G w/ SoC
- Networking + Storage

Pre-History:
- May 2009: Bing Launched
- Feb 2010: Azure Launched
- Dec 2010: Catapult concept
Catapult: Long, Fruitful FPGA Investment

Pre-History:
May 2009: Bing Launched
Feb 2010: Azure Launched
Dec 2010: Catapult concept

v0: Research POC
Built v0 board w/6 Xilinx FPGAs
30k lines of Bing code on FPGA

v1: Scale Pilot
1632 servers deployed
Bing IndexServe accelerated

v2 Production and ramp
FPGAs reach production
Deployed in all new servers

v2: Scale Pilot
1632 servers deployed
Bing IndexServe accelerated

v3: Pikes Peak
Integrated with WCS (OCP) 1.0

Catapult v3: Longs Peak
DNN Platform for Bing
50Gb w/ integrated NIC

Azure AccelNet Unveiled
Azure production launch demo

Project BrainWave / Storm Peak
Real-Time AI
First 3rd Party FPGA Service

Cloud DNNs

SDN Offload

Hypervisor Offload

AI at the Edge

Systolic Arrays and Feature Extraction

Designed for Decision Tree Scoring

2013
2014
2015
2016
2017

Azure Databox Edge
On-Site Inference

Overlake: Celestial Peak
100G w/ SoC
Networking + Storage

Designed for Decision Tree Scoring

Microsoft Supercharges
Bing Search with Programmable Chips

AI at the Edge
Cloud DNNs
Dominant state-of-the-art models evolving rapidly

Classic ML → Deep CNNs → LSTMs/GRUs → Transformers

2012 → 2019+

Sparse transformers?
Sparse mixture of experts?
Dynamic, recursive networks?
Graph neural networks?

Figure sources:
1. Han et al., Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification
2. Vaswani et al., "Attention is all you need"
Rapid Iteration and Deployment

<table>
<thead>
<tr>
<th>Year</th>
<th>Feature</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Low latency LSTM inference</td>
<td>Brainwave v1</td>
</tr>
<tr>
<td>2017</td>
<td>Narrow Precision Breakthrough</td>
<td>Brainwave v2</td>
</tr>
<tr>
<td>2018</td>
<td>Convolution Optimizations</td>
<td>Brainwave v3</td>
</tr>
<tr>
<td>2019</td>
<td>Generalized ISA, Transformers</td>
<td>Brainwave v4</td>
</tr>
</tbody>
</table>
HyperScale -- Bing’s 500 Petaflops Inference Supercomputer

Bing Compute Server

Bing FPGA Appliances
Slick new hardware
Useful hardware
What makes a public cloud company successful?

Hardware Companies

- HUAWEI
- IBM
- Intel
- NVIDIA
- SK hynix
- AMD
- Samsung
- Broadcom

Software Companies

- Amazon
- Baidu 百度
- Microsoft
- Tencent 腾讯
- Google
- Alibaba.com

© Microsoft Corporation
Innovation in Software vs. Hardware

- SW is flexible, but is also SO much bigger
- You can’t lead from the bottom
  - Just look at AMD GPUs vs. nVidia
  - x86 and Windows aren’t the leaders because they’ve always been the best
- Nobody wants to do throw-away work
  - Work needs to (plausibly) span multiple generations

Focus on enabling your customers / developers, not on HW
Why is the FPGA a good choice as an accelerator?

- Greater Performance and Efficiency than CPU, more general purpose than ASIC
- Many applications aren’t about throughput or double-precision floating point
  - AI/ML, Bioinformatics, text processing, financial services...
- Exploits different forms of parallelism than other accelerators

![Diagram of Pipeline Parallelism and Instruction Streams]

- **Pipeline Parallelism**: Shows the flow of data through different stages.
- **Instruction Streams**:
  - Single: No Parallelism (SISD)
  - Multiple: Different instructions on same data (MISD)
- **Data Streams**:
  - Single: CPU
  - Multiple: GPUs (FP) FPGAs (Int)

© Microsoft Corporation
Why is the FPGA a good choice as an accelerator?

- Greater Performance and Efficiency than CPU, more general purpose than ASIC
- Many applications aren’t about throughput or double-precision floating point
  - AI/ML, Bioinformatics, text processing, financial services...
- Exploits different forms of parallelism than other accelerators

### Multiple instruction streams, single data stream (MISD)

<table>
<thead>
<tr>
<th># Instruction Streams</th>
<th># Data streams</th>
<th>Main article: MISD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single SISD</td>
<td>Single</td>
<td>CPU</td>
</tr>
<tr>
<td>Multiple MISD</td>
<td>Multiple</td>
<td>FPGAs</td>
</tr>
</tbody>
</table>

- SISD: No Parallelism
- MISD: Different ops to same data
- SIMD: Same thing to lots of data
- MIMD: Embarrassingly Parallel

Multiple instructions operate on one data stream. This is an uncommon architecture which is generally used for fault tolerance. Heterogeneous systems operate on the same data stream and must agree on the result. Examples include the Space Shuttle flight control computer.[5]
Query: “FPGA Configuration”

NumberOfOccurrences_0 = 7  NumberOfOccurrences_1 = 4  NumberOfTuples_0_1 = 1
Feature Extraction Accelerator

SM1

SM2 AllSpans

SM3 NumberOfOccurrences

SM4

SM5

SM6

SM7

SM8

SM9 NumberOfPerfectMatches

SM10
ASIC vs. FPGA

Performance Range

ASIC

FPGA
**ASIC vs. FPGA**

Performance Range

- **ASIC**
- **FPGA**

**First Use**

- **Light change**
- **Heavy change**

- **ASIC Superiority**
  - Under heavy change

- **FPGA Superiority**
ASIC vs. FPGA

Performance Range

ASIC

FPGA

Time

Performance

First Use

ASIC Superiority

FPGA Superiority

Under heavy change

Light change

Heavy change

Planning

Development

Deployment

Superiority Depends

2-3 Years

6+ Years

Requirement lock to decommission is over a decade (as long as Azure itself has existed)

A lot changes over a decade
Resource Functionality Over Time for 40Gbps Generation

<table>
<thead>
<tr>
<th>Date</th>
<th>Pkts/sec</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/9/2017</td>
<td>22.5M PPS</td>
<td></td>
</tr>
<tr>
<td>7/27/2018</td>
<td>22.5M PPS</td>
<td>PFC Added</td>
</tr>
<tr>
<td>12/15/2018</td>
<td>22.5M PPS</td>
<td>Fast Offload, new Lookup</td>
</tr>
<tr>
<td>6/11/2020</td>
<td>22.5M PPS</td>
<td>PdParser, multi-tenancy, Flow Scaling to 4Million+</td>
</tr>
<tr>
<td>9/24/2020</td>
<td>100M PPS</td>
<td>GFT-V2, 100MPPS, Shell Update</td>
</tr>
<tr>
<td>10/1/2020</td>
<td>100M PPS</td>
<td>PCAP-V3, Filtering</td>
</tr>
</tbody>
</table>
HPC with the Cloud?

- The idea *sounds* great
- Pay for compute only when you use it
- When it breaks, it’s someone else’s problem
- No need to call the realtor / utility company when you want a bigger machine
- New hardware just shows up. No retrofits needed.
Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but...

- Top 500 often include specialized accelerators (especially GPUs)
- Networks are highly specialized, tuned for low-latency, high bandwidth
- Won’t running virtual machines kill performance?
<table>
<thead>
<tr>
<th>Rank</th>
<th>System Information</th>
<th>Rank (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Ppeak (MW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Voyager-EUS2 - ND96amsr_A100_v4, AMD EPYC 7V12 48C 2.45GHz, NVIDIA A100 80GB, Mellano HDR Infiniband, Microsoft Azure Azure East US 2 United States</td>
<td>253,440</td>
<td>30,050.0</td>
<td>39,531.2</td>
</tr>
</tbody>
</table>
Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but...

- Top 500 often include specialized accelerators (especially GPUs)
- Networks are highly specialized, tuned for low-latency, high bandwidth
- Won’t running virtual machines kill performance?
Accelerator Integration

Traditional Accelerator Integration

Bump in the Wire -- In-Network Acceleration
Bump-in-the-wire Architecture

Azure ML
BrainWave

Hardware as a Service
Network Acceleration
Compute Acceleration

Azure AccelNet

NIC

40G Ethernet

CPU

PCIe Gen 3

PCIe Gen 3
Global-Scale FPGA

7 FPGAs Involved

2 FPGAs Involved
Network Latencies

- Extremely low latency (Similar to Infiniband)
- Global-scale FPGA
Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but...

- Top 500 often include specialized accelerators (especially GPUs)
- Networks are highly specialized, tuned for low-latency, high bandwidth
- Won’t running virtual machines kill performance?
Virtualization Overhead – Standard Virtual Machines

SDN Stack Functions
SLB: Software Load Balancers
ACL: Access Control Lists
NAT: Network Address Translation
VNET: Virtual Networks
Metering

10.1.1.2
10.1.1.1
157.2.21.4
157.2.21.4

VM
(10.1.1.1)

Guest OS

Hypervisor

SDN Stack

NIC
Virtualization Overhead – SmartNICs & Bump-in-the-Wire
Why hasn’t Supercomputing moved to the Cloud?

CPUs look largely the same, but...

- Top 500 often include specialized accelerators (especially GPUs)
- Networks are highly specialized, tuned for low-latency, high bandwidth
- Won’t running virtual machines kill performance?

What will *really* make HPC developers adopt the cloud?
Developer Experience

• **Focus on the Customer**
• In Supercomputing, developers are often the customer
• Traditional HPC machines require long, in-advance reservations
• Cloud allows for gradual scaling, 24/7/365 availability

• *Enabling physicists / chemists / biologists / etc.. to experiment* is far more *important to impact* than peak performance
Conclusion

- Software is more important than hardware when you want to make an impact on the world
- Think of FPGAs as a *complement* to GPGPUs, not just a competitor
- FPGAs play a role in all parts of the HPC stack
- The Cloud will replace dedicated supercomputers
  - In large part due to developer experience
- High Flexibility enables a much longer lifetime, especially in new areas