The Configurable Cloud --
Accelerating Hyperscale Datacenter Services with FPGAs

Andrew Putnam – Microsoft
What is a Cloud / Big Data Application?
Technology Scaling

- Moore’s Law (transistors) is still alive
- Dennard Scaling (keeping energy under control) is dead
- 2x users requires 83% more servers
- Need increased efficiency

Datacenter Scaling

~100%+ Growth for the past 4 years
Modern HyperScale Datacenters

- Microsoft > 1,000,000 servers
- 100s of MegaWatts
- $100M+ power bill
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax [TFlop/s]</th>
<th>Rpeak [TFlop/s]</th>
<th>Power [kW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National Supercomputing Center in Wuxi, China</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,435.9</td>
<td>15,371</td>
</tr>
<tr>
<td>2</td>
<td>National Super Computer Center in Guangzhou, China</td>
<td>Tianhe-2 (MilkyWay-2) - TH-IWB-FEP Cluster, Intel Xeon E5-2692 12C 2.20GHz, TH Express-2, Intel Xeon Phi 3151P NUDT</td>
<td>3,120,000</td>
<td>33,862.7</td>
<td>54,902.4</td>
<td>17,808</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>Titan - Cray XK7, Opteron 6274 16C 2.20GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.</td>
<td>560,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,209</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL United States</td>
<td>Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM</td>
<td>1,572,864</td>
<td>17,173.2</td>
<td>20,132.7</td>
<td>7,890</td>
</tr>
</tbody>
</table>

Datacenter: 
~100,000 dual-socket servers

40,960 single-socket servers

16,000 dual-socket servers
3 Xeon Phi / server

18,688 single-socket servers

98,304 single-socket servers
Efficiency via Specialization

Source: Bob Broderson, Berkeley Wireless group
## Silicon Technologies for Computing

<table>
<thead>
<tr>
<th>Technology</th>
<th>Performance/Watt</th>
<th>Efficiency</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs</td>
<td>1X</td>
<td></td>
<td>Today’s standard, most programmable, good for services changing rapidly.</td>
</tr>
<tr>
<td>Manycore CPUs</td>
<td>3X</td>
<td></td>
<td>Many simple cores (10s to 100s per chip), useful if software can be fine-grain parallel, difficult to maintain.</td>
</tr>
<tr>
<td>GPUs</td>
<td>5-30X</td>
<td></td>
<td>Good for data parallelism by merged threads (SIMD), high memory bandwidth, power hungry.</td>
</tr>
<tr>
<td>FPGAs</td>
<td>5-30X</td>
<td></td>
<td>Most radical fully programmable option. Good for streaming/irregular parallelism. Power efficient but currently need to program in H/W languages.</td>
</tr>
<tr>
<td>Structured ASICS</td>
<td>20-100X</td>
<td></td>
<td>Lower-NRE ASICs with lower performance/efficiency. Includes domain-specific (programmable) accelerators.</td>
</tr>
<tr>
<td>Custom ASICS</td>
<td>&gt; 100X</td>
<td></td>
<td>Highest efficiency. Highest NRE costs. Requires high volume. Good for functions in very widespread use that are stable for many years.</td>
</tr>
</tbody>
</table>

**Programming Languages:**

- Conventional programming
- Alternative programming
- Can’t change functionality

**Languages:**

- C/C++
- CUDA
- Verilog
What are FPGAs?

- **Field Programmable Gate Array**
- FPGAs are a sea of generic logic and interconnect
  - “Silicon Legos” – build them into exactly the right circuit for each task
- Special-purpose hardware (FPGAs) is faster and more efficient than general-purpose hardware (CPUs)
- **Change the hardware anytime!**
  - 100 ms to 1 second reconfiguration time
Customize both the processing logic and the I/O
Efficiency via Specialization

Source: Bob Broderson, Berkeley Wireless group
Why not use ASICs?

![Graph showing the comparison between Cloud Applications and time workload stability for using ASICs vs FPGA vs Software.](image)

- **Software**: Running on a virtualized server.
- **FPGA**: Customizable and reusable for similar workloads.
- **ASIC**: Efficient for stable workloads over long periods.

**Ideal for ASICs:** Workloads that are stable for more than 5 years and account for more than 10% of servers.
Why not use GPUs?

- SIMD is best for batch jobs
- Customer-facing (interactive) workloads are small batches, need low latency
- Limited floating point for most workloads
  - Scientific computing and ML are exceptions
- Optimize for the whole fleet, not for one application

Also.... power
Bing: How it all Began

- Launched June 1, 2009
  - MSN Search
  - Windows Live Search
  - Live Search
- October 2010: Approached Microsoft Research for help optimizing performance
- December 2010: Designed an FPGA accelerator for one critical stage
  - First designed at Starbucks – in PowerPoint
User Demands

- Cloud users want interactive applications
- User platforms cannot do all the processing
- Users want look/feel/predictability of local processing
99%+ Responsive

- Cloud services are typically thousands of servers working together
- Overall response time is determined by slowest node
- 1% of nodes can make all the others look bad
99%+ Reliable

- When the Cloud is unavailable, apps stop working
- No backup – Clients can’t run all functionality
- Goal of 99.999% availability is typical
- Redundancy helps, but adds to cost and complexity

Customer Reviews

<table>
<thead>
<tr>
<th>Rating</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>12</td>
</tr>
<tr>
<td>4 star</td>
<td>0</td>
</tr>
<tr>
<td>3 star</td>
<td>3</td>
</tr>
<tr>
<td>2 star</td>
<td>48</td>
</tr>
<tr>
<td>1 star</td>
<td>526</td>
</tr>
</tbody>
</table>

Average Customer Review

⭐⭐⭐⭐⭐ (589 customer reviews)

"Service disruptions are temporary. 1-star reviews are forever."
Datacenter Environment

- Machines last 3-5 years, purchased ~quarterly
- Machines usually repurposed multiple times
- Little/no HW maintenance, no accessibility
- Most datacenter problems are caused by misconfigurations and missed test coverage
- Load balancing, rapid scale up/down are key
- Homogeneity is highly desirable
Datacenter Servers

- Microsoft Open Compute Server
- 1U, 1/2 wide servers
- Enough space & power for 1/2 height, 1/2 length PCIe card
- Squeeze in a single small ~25W card

BFB-1

- Created a custom board with 6 big FPGAs
  - First designed at Starbucks (& Duke’s Chowder House) – in PowerPoint
And Datacenter Operations says...
Centralized Model Complicates Deployment

- Single point of failure
- Server is different from the surrounding servers
- Complicates rack design, thermals, maintainability
- CPU network needed for FPGA communication
  - Definition of the Network In-cast problem
  - Precludes many latency-sensitive workloads
- Limited elasticity
  - What if you need more than six FPGAs?
Fitting FPGAs in the Datacenter

- All servers should be the same
- How about just 1 FPGA in each server?
- Area must be small. Temperatures high. Power low.
Catapult v1 FPGA Accelerator Card

- Altera Stratix V GS D5
  - 172k ALMs, 2,014 M20Ks, 1,590 DSPs
- 8GB DDR3-1333
- 32 MB Configuration Flash

- PCIe Gen 3 x8
- 8 lanes to Mini-SAS SFF-8088 connectors
- Powered by PCIe slot

[Image of Stratix V FPGA module with 8GB DDR3 and Configuration Flash]

[Image of PCIe Gen3 x8 with 4x 20 Gbps Torus Network]
Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per 1/2 Rack
- Network among FPGAs:
  - 6x8 Torus at 20 Gb/link

Data Center Server (1U, 1/2 width)
1,632 Server Pilot Deployed in a Production Datacenter
1,632 Servers with FPGAs Running Bing Page Ranking Service (~30,000 lines of C++)

- **2x Increase in Throughput**
- **29% Latency Reduction**
- **Reduced # of servers**
- **More compute time for improving relevance**
- **< 30% Cost**
- **< 25 W Power**
- **0 HW Failures**
Was Catapult v1 deployed into production?
2.4+ million emails per day

5.8+ billion worldwide queries each month

8.6+ trillion objects in Microsoft Azure storage

48+ million users in 41 markets

50+ million active users

1 in 4 enterprise customers

400+ million active accounts

250+ million active users

50+ billion minutes of connections handled each month

50+ million active users

400+ million active accounts

250+ million active users

400+ million active accounts

200+ Cloud Services: Diversity

1+ billion customers · 20+ million businesses · 90+ markets worldwide
Workload Diversity

- Bing was the big dog, but Azure grew much faster.
- Compute offload for Bing isn’t enough to justify hyperscale deployment.

- Could go Bing-specific
  - Misconfiguration is the leading cause of problems in the datacenter
  - Increased hardware diversity means increased chances of misconfigurations

- Could try to do offload for all the other services
  - But is there enough time to learn each new application?
Compute vs. Infrastructure

- **Compute** acceleration is application-specific
  - Bing Ranking
  - Machine Learning / DNNs
  - Bing ≈ 1-2 years

- **Infrastructure** acceleration benefits common software and services
  - Network offload and processing
  - Encryption / Compression
  - Security

- BOTH are critical when dealing with diverse cloud workloads
  - ≈ 12 years?
Catapult v2 – Bump in the Wire

- 40G NIC and TOR
- FPGA
- 4GB DDR
- 40 Gb NIC
- Catapult v2 FPGA Card
- CPU
- PCIe Gen 3 x8
- ToR
- Pikes Peak
- Storey Peak
Catapult v1 Architecture

NIC

Compute Acceleration

PCIe Gen 3

CPU

PCIe Gen 3

PCIe Gen 3
Bump-in-the-wire Architecture

Hardware as a Service
Network Acceleration
Compute Acceleration

NIC
40G Ethernet
PCIe Gen 3
CPU
PCIe Gen 3
Network Latencies

- **6x8 Torus Latency**
- **LTL Average Latency**
- **LTL 99.9th Percentile**

**Examples of L2 latency histograms for different pairs of FPGAs**

- **LTL L0 (same TOR)**
- **LTL L1**
- **LTL L2**

**Examples of L0 latency histogram**

**Example L1 latency histogram**

**Catapult Gen1 Torus** (can reach up to 48 FPGAs)

**LTL Average Latency**

**LTL 99.9th Percentile**

**Number of Reachable Hosts/FPGAs**

- LTL L2
Inter-FPGA Communication

- **LTL** -- Lightweight Transport Layer
- A Low-latency, Reliable, Connection-based communication channel for FPGA-to-FPGA messaging over standard Ethernet network
- Send-side buffering and retransmit until recv. ACK
- Built in send-side queueing mechanism to handle serialization of messages during contention
- FPGAs can communicate without any CPU intervention
Benefits of Bump-in-the-Wire

- Compute & Infrastructure Acceleration with one board
- A global hyperscale FPGA fabric – 100k+ FPGAs
- Improved Robustness & Fault Tolerance
- Fewer hops between FPGAs
- Independent of physical location
- Customized network hardware & protocols
- Allows sharing of underutilized FPGA resources

All while retaining hardware homogeneity!
Was *this* Catapult deployed into production?

**Yes**
• FPGAs Included in every new server for all major services
• Deployed across 15 countries and 5 continents
• Already in large scale production in both Bing and Azure
What workloads run well?

- Microsoft Azure
- Bing
- Compression
- Encryption
- DNNs
Compute Acceleration -- Bing Ranking

- 2x Faster at 2x higher load
- Much lower variance
Data Compression

**Xpress8 L2 (5.6GB/s)**
- 30x throughput
- 20% compression loss
- In-line compression

**Xpress9 L6 (300MB/s)**
- 4x throughput
- No compression throughput
- Short/mid-term data

**LZMA (5MB/s)**
- 5x throughput
- 5% compression loss
- Long-term storage

*Measured on Canterbury dataset*
Homomorphic Encryption

- Allows computation on data without revealing the data itself to the entity performing the computation
- Premium example: cloud computing
  - Clients encrypt data and store encrypted data in the cloud
  - The cloud server performs computation but an attacker (malicious admin, hacker, intelligence agency) cannot see the actual data
- Only user can interpret the result

<table>
<thead>
<tr>
<th>Author</th>
<th>Architecture</th>
<th>Parameter</th>
<th>Mult</th>
<th>Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catapult</td>
<td>FPGA (Stratix 5)</td>
<td>YASHE, n=16384</td>
<td>49 ms</td>
<td>1 ms</td>
</tr>
<tr>
<td>[LN’14]</td>
<td>3.5 GHz Intel i7</td>
<td>YASHE, n=4096</td>
<td>49 ms</td>
<td>0.7 ms</td>
</tr>
<tr>
<td>[RJVDV’15]</td>
<td>FPGA (Virtex-7)</td>
<td>YASHE, n=32768</td>
<td>112 ms</td>
<td>0.17 ms</td>
</tr>
</tbody>
</table>
Infrastructure Acceleration

**SmartNIC: SDN and Crypto offload**

- Generic Flow Table (GFT) rule based packet rewriting
- Enhanced network security
- 10x latency reduction vs software, CPU load now < 1 core
- 25Gb/s throughput at 25μs latency – the fastest cloud network
Deep Learning -- Image Classification via CNN

2x 8-core 2.10 GHz Xeon (95W TDP)  One Stratix V D5 FPGA (25 W)
the first AI super computer in action.
- >10x the AI capacity of the world’s largest supercomputers with Catapult v2
- Next generation is more than 3x more powerful
Microsoft Claims Fastest Network in the Cloud

Microsoft CEO Satya Nadella demonstrated some of the AI supercomputing capabilities in its Azure public cloud that make it now among the fastest cloud services available. The secret sauce is the use of field-programmable gate arrays.

By Jeffrey Schwartz  09/30/2016

Microsoft CEO Satya Nadella demonstrated some of the AI supercomputing capabilities of the company's Azure platform. He said, in his keynote this week at Microsoft Ignite in Atlanta, that the company two years ago started upgrading every node in its Azure public cloud with software-defined network (SDN) infrastructure, developed using field-programmable gate arrays (FPGAs).

The result is that Microsoft's Azure public cloud fabric is now built on a 25 gigabit-per-second
Microsoft's Project Catapult is why Intel bought FPGA-maker Altera for $16.7 billion last year

By Shawn Knight on September 26, 2016, 4:45 PM

Intel last year acquired FPGA-maker Altera for $16.7 billion in cash, the chipmaker's largest purchase in history. As it turns out, Microsoft played a key role in Intel's decision to make the purchase.
WHAT COULD POSSIBLY GO WRONG?
How do you program FPGAs?

- 10:1 or larger ratio of SW to HW programmers
- Verilog is the predominant programming language
- High-level synthesis (SDAccel, OpenCL, ROCCC, LegUp...) make it a little easier
  - Debugging still a challenge
  - Often requires platform-specific pragmas for good performance, which requires detailed knowledge of the FPGA architecture
- So far, most successful model is a custom contract between HW/SW, and still programming in Verilog
99.9% software latency
99.9% FPGA latency

Day 1  Day 2  Day 3  Day 4  Day 5

1.0  2.0  3.0  4.0  5.0  6.0  7.0
Normalized Load & Latency

And ruiner of accelerator dreams
Usage Models

- Could allow users to have raw access to the FPGAs
  - Amazon EC2 F1
- High-Level Synthesis tools make this easier

- More common is a library of services built on top of FPGAs
  - DNN computation service, linear algebra, web search (Bing)...

- Infrastructure acceleration enables gradual migration to FPGAs
What am I worried about?

- I don’t think the biggest problem is software engineers being able to program FPGAs
- I think our biggest problem is that we’re going to make software engineers fight old battles
  - Libraries, linkers, backwards compatibility
Catapult Academic Program

- Donated 3 full racks of machines at TACC for research (96 machines per rack)
- Individual boards are available for in-house development use
- Very little effort to move from the academic cluster to the production machines

- See: http://research.microsoft.com/catapult for details
Conclusion

- Specialization with FPGAs is critical to the future Cloud
- FPGAs are harder to program, and improving that will greatly improve efficiency for Big Data / Cloud apps
- FPGAs allow optimization of both compute and I/O operations, so think beyond the core application
- What can you build with your global pile of Legos?