Doing FPGA in a Former Software Company

Feng-hsiung Hsu (fhh@microsoft.com)
Hardware Computing Group
Microsoft Research Asia (MSRA)
Outline

• Catapult at Hot Chips 2014
• Video Stabilization, 2005-2007
• Machine Learning, 2006-
• Index Serve, 2009-
• Stereo Matching, 2012-2013
• Lessons Learned
• FPGA in **ALL** Microsoft Datacenter Machines?
• Conclusions
Flint Center, Cupertino
Site of Hot Chips 2014
(Picture taken from Apple iPhone 6 event a month later)
Catapult FPGA Accelerator Card

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

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

---

**Stratix V**

**Config Flash**

**8GB DDR3**

**PCle Gen3 x8**

**4x 20 Gbps Torus Network**
Board Details

- 16 Layer, FR408
- 9.5cm x 8.8cm x 115.8 mil
- 35mm x 35mm FPGA
- 14.2mm high heatsink
Microsoft Open Compute Server

- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow

200 LFM

68 °C Inlet
Economy Case for Catapult

- Less than 1/10 cost of compute servers
- For Bing RaaS (Ranking as a Service), 2x performance expected
- However, for Bing index servers, RaaS is only 20% of work load
  - SaaS (Selection as a Service) is about 60%
- Azure and other Microsoft cloud services?
- Given multiple use scenarios, to make full economic sense, just RaaS is not enough
In the Beginning
Video Stabilization

FPGA (EP2S60)

USB2.0 Controller

1394 Controller

Composite Controller

SRAM

DDR SDRAM Socket

Configuration Device (EPCS64)

CF/SD Card Socket

PCI Chip

Development Board

PCI Interface
Functional Description

- StratixII device with 60k LEs, 2.5Mbits memory and 36 DSP blocks*
- Up to 2 GB DDR SDRAM
- 4MX32bits SRAM
- PCI 33MHz 32bit universal interface
- Configure/debug FPGA through PCI

*The device can be replaced by a larger FPGA with 90k LEs, 4.5Mbit memory and 48 DSP blocks without any change.
Note 1: Select PCI or JTAG to configure FPGA and FPGA configuration device through jumpers
Video Stabilization Architecture
Video Stabilization, 2005-2007

• Built own board due to difficulty in getting suitable FPGA board from US at the time
• Used as a project to train personnel
• Spare pixels on the rim used to do virtual image sensor movement
  • Translation
  • Rotation (can't be corrected with lens based stabilization)
• Optical flow based
• No relevant product group adoption
  • Not on Windows Mobile team's agenda
Machine Learning for Web Search (2006-)

• Began as part of a hardware acceleration project for web search
• First step: RankBoost
  • Local expertise
  • Comparable NDCG to RankNet/Lambda Rank
  • Combination with RankNet/Lambda Rank could lead to even better NDCG
FPGA Accelerated RankBoost
Result - Speed

- On the same training set (3.4GB)

<table>
<thead>
<tr>
<th>Implementations</th>
<th>Time (hour)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>RankBoost (Old algorithm)</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>Distributed RankBoost (New algorithm, 10 threads)</td>
<td>1.5</td>
<td>3 per thread</td>
</tr>
<tr>
<td>Accelerated RankBoost</td>
<td>0.256</td>
<td>176</td>
</tr>
</tbody>
</table>
Subsequent Developments

• RankBoost Used as a feature selection mechanism
• FPGA based solution used by one team initially
• But distributed software version became the practice due to easy software integration
Beyond RankBoost

• Lambda Rank also accelerated by 20-50x on FPGA and GPU, compared to distributed version
  • Same fate as the FPGA version of RankBoost
• A few other algorithms also accelerated
• Some interest in Deep Neural Networks, but no active FPGA work
Opportunities in Bing (Index Serve, 2009-)

- Crawler
- Index Builder
- Feature Extraction
- Feature Selection
- NN Training
- NN Evaluation
- Static Rank

Offline:
- Inverted Index
- Static Ranks
- RankBoost

Online:
- Front End
- Caching
- Indexing Serving Rows
  - Ranking Models
    - LambdaRank
    - Ranking

Query
Index Serve

• Query Processing for Bing
• Largest capital and operational cost item
  • Hundreds of million to over a billion in terms of annual cost
• Matching and L0/L1 ranking (SaaS) and L2 ranking (RaaS) represents over 80% of the workload
  • SaaS represented ~70% of the workload in 2009
• Work on SaaS began in 2009, fully realizing the FPGA board available then may not be adequate
• A quick and dirty version completed for TechFest 2010
• Things were looking good, but then...
Devil is in the Details

• The task was harder than expected
  • Intrinsic complexity
  • Software constantly changed
  • Discovered a gross inefficiency in Bing document format at the time
    • Conveyed to Bing team, but Bing team busy with something else despite an easy 2x gain
    • Worked out a new format for hardware implementation
    • Concept picked up by Tiger team at MSRA, but with yet another format...
  • Loss of major personnel

• Product team (and management) indifference
  • Cost reduction was not a major concern for Bing despite annual billion dollar operational loss (the NEXT syndrome)
Marched On

• First “real” implementation in 2012
  • All the major pieces for SaaS implemented, but not integrated
    • Won’t fit on the old board
  • Numbers estimated
• RaaS efforts began in Redmond and went through their own difficulties
  • Eventually migrated to the Catapult design
  • Millions of dollars of investment
• Ongoing port of SaaS to Catapult
Craig Mundie’s 7-Holes (Stereo Matching)

- Craig Mundie, Microsoft's former CTO, believed a low cost Kinect-like capability should be available on all sort of devices, including phones, tablets, PCs, and TV sets.
AD Census, FPGA Implementation
Stereo Matching Excursion, 2012-2013

• Real time AD Census completed in 2012
• “Oculus”/7-Hole project in Redmond
  • Potential successor to Kinect 2
  • GPU version working, but no FPGA and nowhere near ASIC
  • MSRA team agreed to help
    • Implemented pieces, while Redmond refined GPU version due to new requirements
    • Implemented two alternative complete algorithms on FPGA with lower estimated ASIC cost
• Oculus project killed off
Lesson Learned

• Despite the transition to a “Devices and Services” company, and now “Mobile First, Cloud First”, Microsoft still a software company at the bone, with its strengths and limitations
• Working with product teams requires knowing what they are measured by
• Know who are the influencers
• Go with the trends
• And pray
Datacenter Deployment?
Microsoft Datacenter Environment

• Software services change monthly
• Machines last 3 years, purchased on a rolling basis
• Machines repurposed ~½ way into lifecycle
• Little/no HW maintenance, no accessibility

• Homogeneity is **highly** desirable

**The paradox:** Specialization *and* homogeneity
Efficiency via Specialization

Source: Bob Broderson, Berkeley Wireless group
Design Requirements

Don’t Cost Too Much
<30% Cost of Current Servers

1. Specialize HW with an FPGA Fabric
2. Keep Servers Homogeneous

Don’t Burn Too Much Power
<10% Power Draw (25W max, all from PCIe)

Don’t Break Anything
Work in existing servers
No Network Modifications
Do not increase hardware failure rate
Datacenter Servers

- Microsoft Open Compute Server
- 1U, ½ wide servers
- Enough space & power for ½ height, ½ length PCIe card
- Squeeze in a single FPGA
- Won’t fit (or power) GPU

Catapult FPGA Accelerator Card

- Altera Stratix V D5
- 172,600 ALMs, 2,014 M20Ks, 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network
Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per ½ Rack
- 6x8 Torus Network among FPGAs
  - 20 Gb over SAS SFF-8088 cables

Data Center Server (1U, ½ width)
Selection as a Service (SaaS)
- Find all docs that contain query terms,
- Filter and select candidate documents for ranking

Ranking as a Service (RaaS)
- Compute scores for how relevant each selected document is for the search query
- Sort the scores and return the results

Query → Selected Documents → 10 blue links
FPGA Accelerator for RaaS

Document
- FE: Feature Extraction
- FFE: Free-Form Expressions
- MLS: Machine Learning Scoring

8-Stage Pipeline
- FPGA 0
- FPGA 1
- FPGA 2
- FPGA 3
- FPGA 4
- FPGA 5
- FPGA 6
- FPGA 7

Routes:
- Route to Head
- Document Scoring Request
- Return Score

RaaS Servers
- Server
- Server
- Server
- Server
- Server
- Server
- Server
Scalable Deployment Challenges

Issues with Spanning Multiple FPGAs
- Health monitor to detect stalled pipelines
- Reconfiguration protocol to remove lockups
- Re-mapper shifts images on machine failure

General Issues with an FPGA Fabric
- PCIe driver tuning for FPGA configuration
- SEU scrubbing of the FPGA
- Wiring and board check at integration
1,632 Server Pilot Deployed in a Production Datacenter
Expected RaaS Deployment

• Pilot run in first half of 2015
• 2x Speedup expected
• Up to 30% latency reduction
• 20% of total workload for index serve
• But potential for relevance improvement
SaaS FPGA under test

CPU
- Query compiler
- Index manager

Host Memory System
- Term LUT
- Main index
- Per doc index

PCle Gen3 x8

FPGA
- L1 ranker
- Feature Extraction
- Matcher
- Pruner
- Index Readers
- Decoders
- MMU

On-board DDR3 (4GB)
- Main index cache
- High freq. Per doc index
SaaS vs. RaaS

- 60% of the workload vs. 20%
- No communication between FPGAs
- Memory bandwidth and PCI-e bandwidth critical
- Bigger chunk of the system latency
Expected SaaS Deployment

• Second half of 2015 for pilot run
• Target at least 50% speedup (up to 3x plausible)
• 2-10x reduction in latency
Beyond Bing

• Other Microsoft services are also looking into FPGA deployment
• Product groups are forming FPGA teams
• Longer term, programming tools, such as Open CL and so on, will be critical
Conclusions

• Microsoft is doing FPGAs
  • Lots of execution risks ahead
• Being a “former” software company means that it is no longer just doing software
  • Mobile First, Cloud First
• But it is still a software company
  • Having great tools that allow knowledgeable software programmers to do FPGAs will be important
• And it is a company
  • FPGA deployment would need to make business sense