# Adaptable Computing The Future of FPGA Acceleration



Dan Gibbons, VP Software Development June 6, 2018

# Adaptable Accelerated Computing



# Three Big Trends



#### 02 Dawn of Al

> Adoption across all industries

> Injecting new intelligence into apps

> From endpoints to edge to cloud





Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

#### 03

#### Computing After Moore's Law

> Heterogeneous computing with accelerators
 > Breadth of apps require different architectures

> Speed of innovation outpacing silicon design cycles



#### © Copyright 2018 Xilinx



### The Evolution of Computing

#### Trend to Heterogeneous Architectures with Acceleration of New Workloads

Mainframe	PC	Mobile	Pervasive Intelligence	
Era	Era	Era	Era	

**E** XILINX

### The Need for Adaptable Intelligence

The intelligent connected world needs adaptable accelerated computing.

Everything Intelligent & Connected

Deployed at Global Scale Dynamic Needs & Rapid Innovation



## Why it Matters – Personalized Medicine Example

- Whole genome diagnosis to treat critically ill newborns
- Analysis reduced from 1 day to 20 minutes
- Patient-specific genomics dynamically optimized
- Medical data and research needs to be securely accessed across the globe



![](_page_5_Picture_6.jpeg)

![](_page_5_Picture_7.jpeg)

edico

# The FPGA Advantage

![](_page_6_Picture_1.jpeg)

# The FPGA Advantage for Machine Learning Inference

#### Adaptive Architecture

> Customer dataflow, precision, optimizations

#### **Custom Memory Hierarchy**

> Keeps data inside vs. external memory bottleneck

#### Workload + ML Inference

> Unleashes the power of on-chip system dataflow

![](_page_7_Picture_7.jpeg)

![](_page_7_Figure_8.jpeg)

![](_page_7_Picture_9.jpeg)

### **Powerful FPGA Optimizations: Precision**

![](_page_8_Figure_1.jpeg)

© Copyright 2018 Xilinx

# Powerful FPGA Optimizations: Compression

![](_page_9_Figure_1.jpeg)

![](_page_9_Figure_2.jpeg)

30x to 50x compression rate without impacting accuracy (AlexNet)

![](_page_9_Picture_4.jpeg)

#### A Survey of Model Compression and Acceleration for Deep Neural Networks

Yu Cheng, Duo Wang, Pan Zhou, Member, IEEE, and Tao Zhang, Senior Member, IEEE

![](_page_9_Picture_7.jpeg)

With the world's leading research in neural network model compression, DeePhi develops DECENT (DEep ComprEssioN Tool). It firstly introduces pruning, quantization, weight-sharing and Huffman encoding to reduce model size from 5x to 50x without loss of accuracy.

![](_page_9_Picture_9.jpeg)

100%

### FPGA Advantage: Deterministic Latency

#### "Batch" Inference

- > Parallel batch of data to feed SIMD
- > High batch => low latency, higher throughput
- > Lower compute efficiency at low batch

![](_page_10_Figure_5.jpeg)

- > Low and deterministic latency
- > High throughput regardless of batch size
- > Consistent compute efficiency

![](_page_10_Figure_9.jpeg)

![](_page_10_Figure_10.jpeg)

ML Inference Integrated with Other Workloads

#### Live video summary using CNN & RNN

![](_page_11_Figure_2.jpeg)

![](_page_11_Picture_3.jpeg)

#### Adaptable Compute Use Cases Across the Datacenter

![](_page_12_Picture_1.jpeg)

![](_page_12_Picture_2.jpeg)

![](_page_12_Picture_3.jpeg)

- ✓ ML Inference
- ✓ Database / Big Data Analytics
- ✓ Video Transcoding
- ✓ Financial Services Analytics
- ✓ Genomics

- ✓ Compression
- ✓ Encryption
- ✓ Key-Value Store
- ✓ ML Inference
- ✓ Database / Big Data Analytics

- ✓ IPSec/SSL
- ✓ OVS Offload
- ✓ Bare Metal Services
- ✓ Security
- ✓ Monitoring

### Zynq SoCs: Adaptable Computing on the Edge

![](_page_13_Picture_1.jpeg)

![](_page_13_Picture_2.jpeg)

![](_page_13_Picture_3.jpeg)

![](_page_13_Picture_4.jpeg)

Face Detect

Traffic SSD

✓ 4 CNN Models
✓ 3 Live Inputs + File IO

✓ Under 10 Watts!

#### Xilinx Enables Adaptable Accelerated Computing

![](_page_14_Picture_1.jpeg)

#### XILINX 'FPGA as a Service' goes wide

![](_page_15_Picture_1.jpeg)

![](_page_15_Picture_2.jpeg)

Launched Nov 2016

![](_page_15_Picture_4.jpeg)

Launched Jul 2017

![](_page_15_Picture_6.jpeg)

Tencent Cloud Launched Aug 2017

![](_page_15_Picture_8.jpeg)

Launched Sep 2017

Aliyun

Alibaba Cloud Computing

Launched Oct 2017

![](_page_15_Picture_13.jpeg)

### Towards Software as a Service (SaaS)

![](_page_16_Figure_1.jpeg)

![](_page_17_Picture_0.jpeg)

#### Optimal acceleration results requires platform performance, compiler efficiency and programming proficiency

![](_page_17_Picture_2.jpeg)

High Performance Platform

![](_page_17_Picture_4.jpeg)

Advanced Compiler

	and a future for the state		
T+ II & 0+ 0+ 4+ # 1	EN-LA C BOIS-D-NO-N-	0.4	7344
I Falage Lighter 11 5 TO	The second secon	If Outer II * D	<b>3</b> H =
······································	<ul> <li>arches the advance of comment provides and provides a families of producted class Contemporate Inglements (Dynational/Provider )</li> </ul>	A N V + V *	0 # ·
* Bing rolps a condex je 1 2 janukring ombare	pretected final static toring 600015 "inv_ingenets"; //000.001 pretected TheiriseSphere (Personalphares are befaitDectorphet pretected int Restores are frequent(0);	* @ jard sectored * @ jard sectored along and \$ @'Segment	3+++ 2.61
1.3. jackwarpstreep 1.3. jackwarpstreep	- presented weld paras(Discussion document) (	1 @ CamereProvides - Report Object	The joint Edmar examples
<ul> <li>If just the period</li> <li>If its of the complete</li> </ul>	<pre>DBt spears accesses, greatered(lines(); DBt increments With aux/Wath.read(lines / 10), 10); det increments with aux/Wath.read(lines / 10), 10);</pre>	Facture Factor	Annunchain The coandison Thrattanes
<ul> <li>II Jacobio Meccapic p</li> <li>II Jacobio Meccapic p</li> <li>II Jacobio Meccapic p</li> <li>II Jacobio Meccapic Coll</li> </ul>	104 Long The Secondary, 104 (Long + Recement, - Lines) 114 (Long + Recement, - Lines)	R - count and count R - solution Charged Sc R - solution Charged Sc R - solution Charged Sc	contrain tea contrain tea adhana, 8 adhan showed
<ol> <li>Jacobarton pro</li> <li>Provension Action pro</li> <li>Insul de Managemente</li> </ol>	Terr 1	+ updated	mpoor at adaptive a Se adapti
E toggingeneratiation p? 5 © organizans as anaroles ja 2 © organizans as a segmeter, pe	<pre>ist each document grtLindfTurt[line + Imagth); isagine each = dfur() Real(isagine each Real(isagine); Real(is); document _ADFur(); real(SECRE); b);</pre>		part and he to define a custom
1. Biogiaclipia in scamples pa 1. Program Latin Transfe	Content, old/ner Septent/Decapiferent/Jornald/tork		provide for upe he Hat
1 Billing in Dependencies	) enteb (BallocationTeception a) (	-	comple in
<ul> <li>B. (RE Sovies Links (RE E.S.)</li> <li>D. may Anni</li> </ul>	Pattern Lauris Delaute & Canala II	411 - 0+ C+ TO	denotes that
P Drawn	148		inter adding
1 champions		1	support is
G about herei			the Ecliptor Jave Table
S bythe hubble (			Here the

![](_page_17_Picture_7.jpeg)

Productive IDE & Optimized Libraries

**User Onboarding** 

![](_page_17_Picture_10.jpeg)

### Rich Stack Integrated with Frameworks

![](_page_18_Figure_1.jpeg)

# **Transformation Through Innovation**

![](_page_19_Figure_1.jpeg)

### The Era of Heterogeneous Computing Architectures is Here

FPGA's are uniquely suited for adaptable accelerated computing

Xilinx is leading the way with platforms, tools, applications and FaaS

Now is the opportunity for application development and deployment

![](_page_20_Picture_4.jpeg)

![](_page_20_Picture_5.jpeg)

![](_page_20_Picture_6.jpeg)

![](_page_20_Picture_7.jpeg)