Adaptable Computing
The Future of FPGA Acceleration

Dan Gibbons, VP Software Development
June 6, 2018
Adaptable Accelerated Computing
Three Big Trends

01 Explosion of Data
- > 90% unstructured
- > Video & image content
- > Needs higher throughput & real-time computing

02 Dawn of AI
- > Adoption across all industries
- > Injecting new intelligence into apps
- > From endpoints to edge to cloud

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 Era  PC Era  Mobile Era  Pervasive Intelligence Era
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
The FPGA Advantage
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
Powerful FPGA Optimizations: Precision

Impact of Precision on Performance

Similar accuracy 10+ years of research

Active research area (binary, variable, bit serial...)

© Copyright 2018 Xilinx
Powerful FPGA Optimizations: Compression

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

“Batch-less” Inference
> Low and deterministic latency
> High throughput regardless of batch size
> Consistent compute efficiency
Live video summary using CNN & RNN
Adaptable Compute Use Cases Across the Datacenter

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

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

Networking
- IPSec/SSL
- OVS Offload
- Bare Metal Services
- Security
- Monitoring
Zynq SoCs: Adaptable Computing on the Edge

- 4 CNN Models
- 3 Live Inputs + File IO
- Under 10 Watts!
Xilinx Enables Adaptable Accelerated Computing
XILINX ‘FPGA as a Service’ goes wide

- **aws**: Launched Nov 2016
- **Nimbix**: Launched Nov 2016
- **Baidu**: Launched Jul 2017
- **Tencent Cloud**: Launched Aug 2017
- **Huawei**: Launched Sep 2017
- **Aliyun**: Alibaba Cloud Computing, Launched Oct 2017
Towards Software as a Service (SaaS)
Optimal acceleration results requires platform performance, compiler efficiency and programming proficiency
Rich Stack Integrated with Frameworks

- Open Frameworks
- Accelerated Libraries
- Development Environment
- Platforms

- Machine Learning
  - TensorFlow
  - mxnet
  - Caffe
- Video Transcoding
  - FFmpeg
- Data Analytics
  - PostgreSQL
  - Database Analytics

- SDAccel Environment

- FPGA as a Service (FaaS)
  - AWS
  - Aliyun
  - Alibaba Cloud Computing
  - Huawei
  - Nimbix

- On Premise Boards
Transformation Through Innovation

- **World’s First FPGA**
- **First Virtex FPGA**
- **First 3D FPGA & HW/SW Programmable SoC**
- **First MPSoC & RFSoC**
- **Virtex-2 Pro**
- **ACAP**

Timeline:
- **1980**
- **1990**
- **2000**
- **2010**
- **2020**
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