

Hardware acceleration with FPGAs on AWS

Julien Simon, Principal Evangelist AI/ML, AWS

Ramine Roane, Senior Director, Product Management, Xilinx

Sébastien Delerse, Co-founder, Mipsology



XILINX
ALL PROGRAMMABLE™

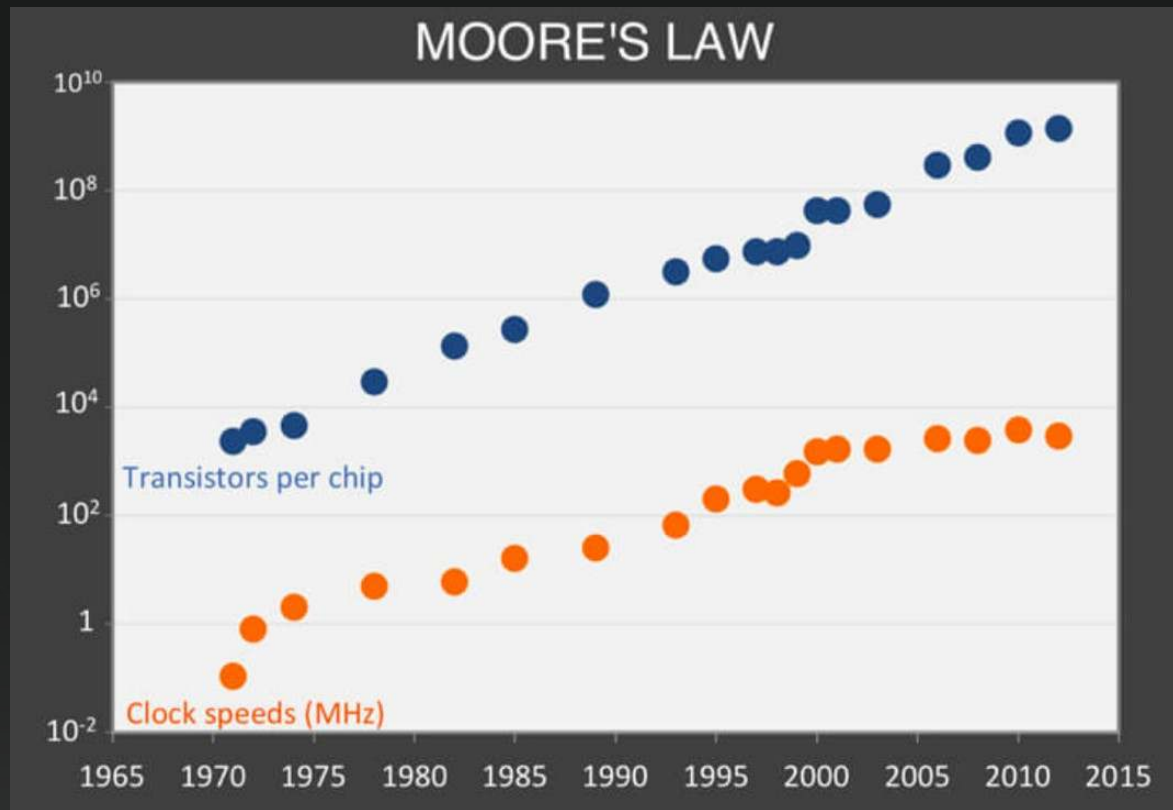
Mipsology

Agenda

- The case for accelerating computing
- What is an FPGA?
- Using FPGAs on AWS
- Customer case studies
- Resources



The case for accelerated computing



Source:
Intel

Moore's winter is (probably) coming

- « *I guess I see Moore's Law dying here in the next decade or so, but that's not surprising* », Gordon Moore, 2015
- **Technology limits:** a Skylake transistor is around 100 atoms across.
- New workloads require **higher parallelism** to achieve good performance.

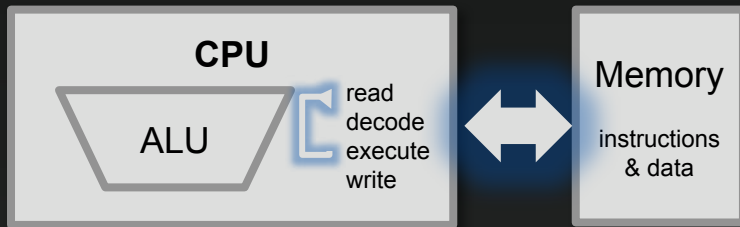
GPUs are not optimal for some applications

- Power **consumption** and **efficiency** (TOPS/Watt)
- Strict **latency** requirements
- Other **requirements**
 - Custom data types, irregular parallelism, divergence
- Building your own **ASIC** may solve this, but:
 - It's a huge, costly and risky effort
 - ASICs can't be reconfigured
- What about **FPGA**?

Why FPGA: Application Specific HW & Memory

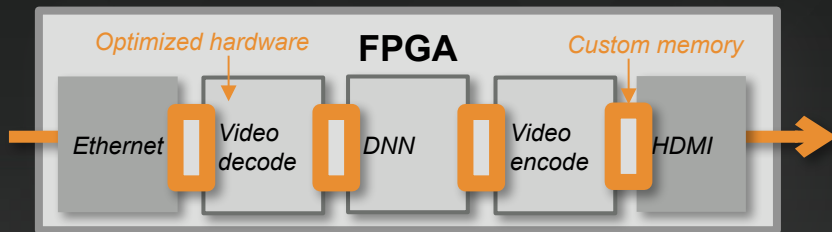


➤ Given an algorithm to implement...



➤ CPU/GPU implementation: Von Neumann

- Rigid sequential execution (SIMD for GPU)
- Memory access bottleneck
- Not optimal for custom width or decision handling



➤ FPGA implementation: WYSIWYG

- Custom dataflow, width, decision handling
- Custom memory hierarchy: keeps data inside
- Custom IOs: high throughput & low latency

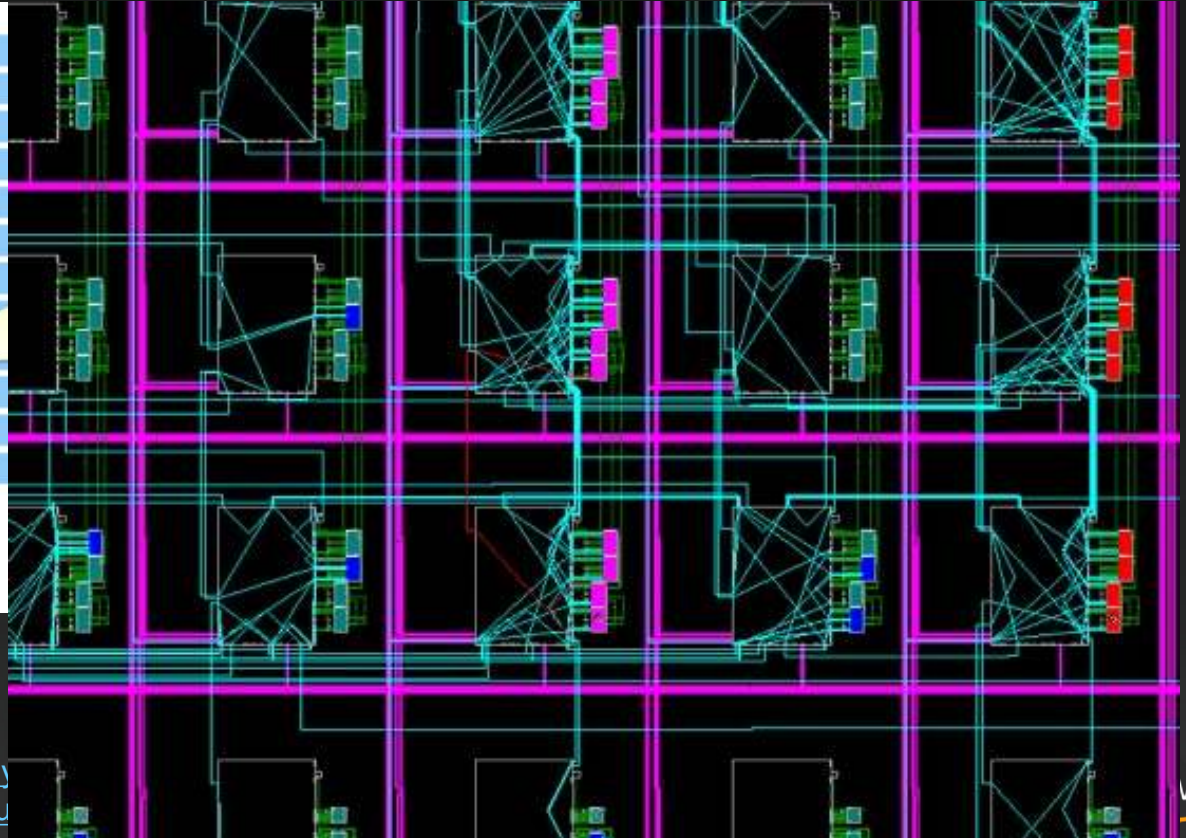
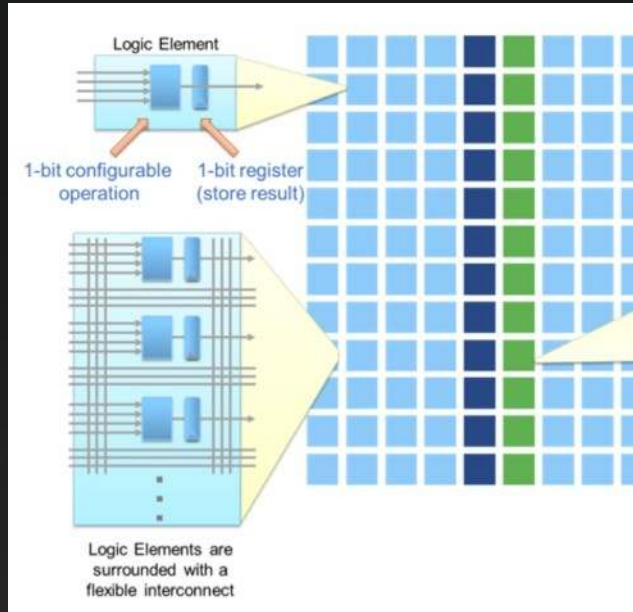


What's an FPGA?

The FPGA

- First commercial product by Xilinx in 1985
- **Field Programmable Gate Array**
- Not a CPU (although you could build one with it)
- « **Lego** » **hardware**: logic cells, lookup tables, DSP, I/O
- Small amount of very fast on-chip memory
- Build **custom logic** to accelerate your SW application

FPGA architecture



Sources:

<https://www.embedded-vision.com/industry-analy>

<http://www.bober-optosensorik.de/fpga-entwicklu>

Where are FPGAs Traditionally Used?



Communications

- Wired networking
- Wireless infrastructure

Automotive

- Infotainment
- Driver assistance



Datacenter

- High performance computing
- Solid state drives

Aerospace and Defense

- Avionics, Communications
- Space



Test & Measurement

- Communications instruments
- Semiconductor test equipment

Audio, Video, Broadcast

- 3D cameras
- Video transport



Industrial, Scientific, Medical

- Ultrasound systems
- Motor controllers

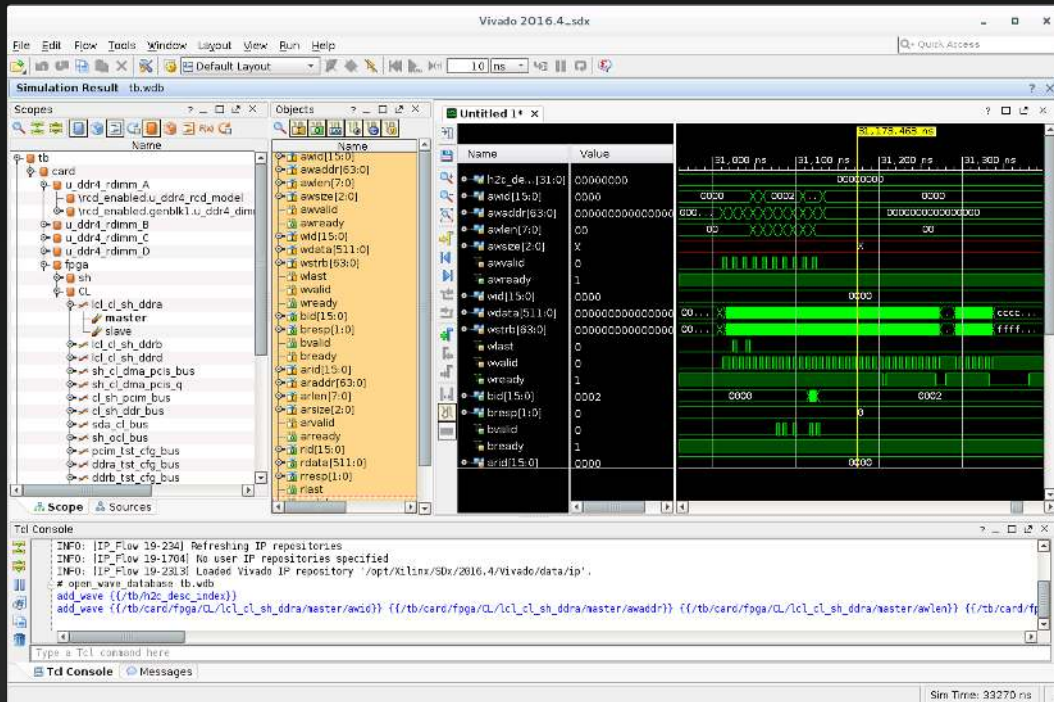
Consumer

- 3D television
- eReaders



Developing FPGA applications

- Languages
 - VHDL, Verilog
 - OpenCL (C++)
- Software tools
 - Design
 - Simulation
 - Synthesis
 - Routing
- Hardware tools
 - Evaluation boards
 - Prototypes



Expensive and hard to scale



Using FPGAs on AWS

Amazon EC2 F1 Instances

- Up to 8 **Xilinx UltraScale Plus VU9P** FPGAs
- Each FPGA includes
 - Local **64 GB DDR4** ECC protected memory
 - Dedicated **PCIe x16** connections
 - Up to 400Gbps bidirectional ring connection for high-speed streaming
 - Approximately **2.5 million logic elements**, and approximately **6,800 DSP engines**

Model	FPGAs	vCPU	Mem (GiB)	SSD Storage (GB)	Networking Performance
f1.2xlarge	1	8	122	470	Up to 10 Gigabit
f1.16xlarge	8	64	976	4 x 940	20 Gigabit

Deploy faster wherever you like

18 Regions – **54** Availability Zones



The FPGA Developer Amazon Machine Image (AMI)

- AWS FPGA **SDK**
 - Amazon FPGA Image (AFI) Management Tools
 - Linux drivers
 - Command line
- AWS FPGA **HDK**
 - Xilinx SDAccel 2017.1
 - **Free license** for F1 FPGA development
 - Supports VHDL, Verilog, OpenCL
 - Design files and scripts required to build an AFI
 - Shell: platform logic to handle external peripherals, PCIe, DRAM, and interrupts
- Run simulation and design on a c4 to save money!

Benefits of Amazon EC2 F1



➤ Business model

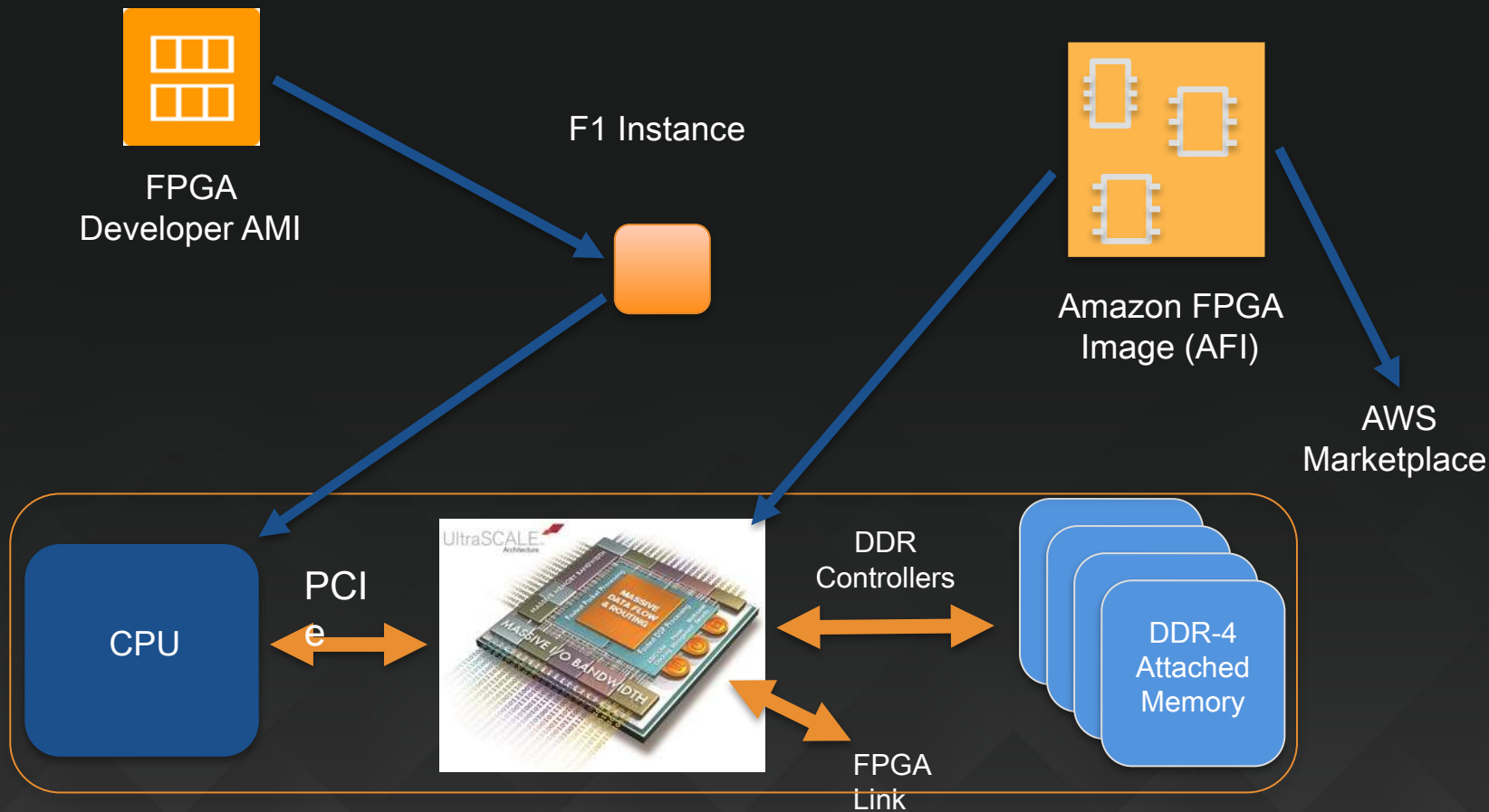
- Turns expensive HW appliance to cost effective, pay-per-use SaaS / API
- Takes sales cycles / evaluations from months to hours
- Scales to millions of AWS users



➤ Development model

- Access to Cloud-based SW development tools
- Access to fastest / best FPGA from anywhere
- Applications can elastically scale to as many accelerators as needed

FPGA Acceleration Using F1 instances





aws marketplace



FPGA Developer AMI

★★★★★ (0) | Version 1.3.3 | Sold by Amazon Web Services

The FPGA (field programmable gate array) AMI is a supported and maintained CentOS Linux image provided by Amazon Web Services. The AMI is pre-built with FPGA development...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)

Mipsology

ZEBRA on 1 FPGA (image classification)

★★★★★ (0) | Version 2017.04.1 | Sold by Mipsology

Zebra offers users FPGA-based class-leading acceleration for neural network inference. The user-defined neural network works on Zebra just as it would on GPU or CPU. Zebra...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)

Mipsology

Free Trial

Zebra Deep-Learning engine for Caffe (1 FPGA)

★★★★★ (0) | Version 2017.10.1 | Sold by Mipsology SAS

Zebra accelerates neural network inference using FPGA. User-defined neural networks are computed by Zebra just as they would be by a GPU or a CPU. Zebra is fully integrated...

Linux/Unix, Ubuntu 14.04 - 64-bit Amazon Machine Image (AMI)



FPGA-Accelerated Deep-Learning Inference with Binarized Neural Networks

★★★★★ (0) | Version 1.2 | Sold by Missing Link Electronics, Inc.

Starting from \$5.00 to \$5.00/hr for software + AWS usage fees

Image classification of the CIFAR10 dataset using the CNV neural network. Based on Xilinx public proof-of-concept implementation of a reduce-precision, Binarized Neural Network...

Linux/Unix, Ubuntu 16.04 - 64-bit Amazon Machine Image (AMI)



Visual System Integrator for FPGA and Embedded Development

★★★★★ (0) | Version 2017.1_Autoupdate | Sold by System View

Starting from \$0.50/hr or from \$2,500.00/yr (43% savings) for software + AWS usage fees

Visual System Integrator is the one-of-a-kind tool for embedded development which, for the first time, makes it possible to develop a full functioning system. Visual System...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)



Hyperion F1 10G RegEx File Scan

★★★★★ (0) | Version 5.5.2 | Sold by TITAN-IC SYSTEMS LTD

Starting from \$50.00 to \$50.00/hr for software + AWS usage fees

The Hyperion F1 10G RegEx File Scan instance provides a preloaded IP image of the 10Gbps Regular Expression Processor (RXP) on the FPGA and a pre-installed SDK allowing the...

Linux/Unix, Amazon Linux 2017.08.0 - 64-bit Amazon Machine Image (AMI)



FireSim Demo v1.0

★★★★★ (0) | Version 1.0 | Sold by Berkeley Architecture Research

FireSim is an FPGA-accelerated hardware simulation tool that cycle-accurately simulates RISCV RocketChip-based clusters, with peripherals like disks and network interface...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)



Merlin Compiler AMI

★★★★★ (0) | Version 1.0.1a | Sold by Falcon Computing Solutions, Inc.

14 Day Free Trial Available - The Merlin Compiler AMI is provided by Falcon Computing Solutions, Inc. The AMI is pre-built with Merlin Compiler that provides push-button C/C++...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)

PLUNIFY

InTime

★★★★★ (0) | Version 2.5.0 | Sold by Plunify

InTime is an automated optimization software for FPGA design by Plunify. It optimizes timing and design performance using machine learning to find the best combination of...

Linux/Unix, CentOS 6.9 - 64-bit Amazon Machine Image (AMI)



Free Trial

NGCodec HEVC/H.265 Encoder D01

★★★★★ (0) | Version 5.01 | Sold by NGCodec

Starting from \$3.25 to \$10.00/hr for software + AWS usage fees

Using an F1 instance, offload HEVC encoding to an FPGA. This version of the NGCodec Encoder features 18-P68 frame encoding at up to 1080p60 resolution/frame rate. The performance...

Linux/Unix, CentOS 7.2 - 64-bit Amazon Machine Image (AMI)



Machine Learning Development Stack from Xilinx, Preview Edition

★★★★★ (0) | Version 12.12.15 | Sold by Xilinx

In this Machine Learning Development Stack, Preview Edition AMI, users easily integrate machine learning into their current applications and deploy them quickly. Users can...

Linux/Unix, CentOS 7.3 - 64-bit Amazon Machine Image (AMI)



Toolkit Powered by RYFT Heterogeneous Computing

★★★★★ (0) | Version v2.1.1 | Sold by Ryft

Starting from \$5.00 to \$5.00/hr for software + AWS usage fees

Ryft's Toolkit is a pre-configured, ready to run image for instantly integrating smarter, more sophisticated FPGA-accelerated search & analysis capabilities into existing...

Linux/Unix, Ubuntu 16.04 - 64-bit Amazon Machine Image (AMI)



Elasticsearch Powered By RYFT Heterogeneous Computing

★★★★★ (0) | Version v2.1.1 | Sold by Ryft

Starting from \$8.00 to \$8.00/hr for software + AWS usage fees

Ryft's ELK is a pre-configured, ready to run image for deploying the powerful open source, distributed real-time search and analytics engine, Elasticsearch, on Amazon's FPGA-accelerated...

Linux/Unix, Ubuntu 16.04 - 64-bit Amazon Machine Image (AMI)



Free Trial

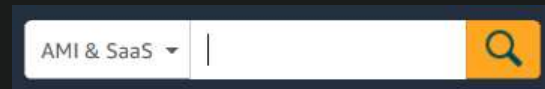
Torch: A scientific computing framework for LuaJIT by Miri Infotech

★★★★★ (0) | Version 1 | Sold by Miri Infotech

Starting from \$0.03 to \$0.03/hr for software + AWS usage fees

Torch is a suite of business tools that uses data mining, machine learning and artificial intelligence to automate work, save money and helps the business grow. It is easy...

Linux/Unix, Amazon Linux 2016.03.03 - 64-bit Amazon Machine Image (AMI)



DeePhi Descartes Efficient Speech Recognition Engine

★★★★★ (0) | Version 2018.02.2a | Sold by Beijing DeepPhi Technology Co., Ltd.

This is an end-to-end ASR (Automatic Speech Recognition) system with FPGA acceleration on AWS F1 by DeePhi. We modify the Baidu DeepSpeech2 framework

(https://github.com/SeanNaren/deepspeech.pytorch)...

Linux/Unix, CentOS 8.1.0-693.2.2.el7 x86_64 - 64-bit Amazon Machine Image (AMI)



Free Trial

Accelerated Machine Learning

★★★★★ (0) | Version AML_v1.0 | Sold by InAccel

Starting from \$2.00 to \$2.00/hr for software + AWS usage fees

AML is InAccel's accelerated machine learning library. It aims to maintain the practical and easy to use interface of other open source frameworks, i.e. of Apache Spark, and...

Linux/Unix, Ubuntu 16.04 - 64-bit Amazon Machine Image (AMI)



Free Trial

DRAGEN Complete Suite - Exome (approx. \$2 per Exome)

★★★★★ (0) | Version 2.2 | Sold by Edico Genomics

Starting from \$9.25 to \$22.80/hr for software + AWS usage fees

The DRAGEN Complete Suite (Exome) enables ultra-rapid analysis of Next Generation Sequencing (NGS) data for small data sets, such as whole exomes and targeted panels. This...

Linux/Unix, CentOS 7.2 - 64-bit Amazon Machine Image (AMI)



Free Trial

DRAGEN Complete Suite - Genome (approx. \$15 per Genome)

★★★★★ (0) | Version 2.2 | Sold by Edico Genomics

Starting from \$10.35 to \$22.10/hr for software + AWS usage fees

The DRAGEN Complete Suite (Genome) enables ultra-rapid analysis of Next Generation Sequencing (NGS) data for large data sets, such as whole genomes. This application uses...

Linux/Unix, CentOS 7.2 - 64-bit Amazon Machine Image (AMI)



Customer use cases

Compelling Use Applications on Amazon EC2 F1



Machine Learning Inference

Speech recognition

40x



Video Streaming

Frame rate for HEVC encoding

10x



Genomics

20 min vs. 33 hours for whole genome analysis

100x



Big Data Analytics

40 min vs. 60 hours for logfile query

90x



Case study: Edico Genome



- Genome diagnosis to treat critically ill newborns
- Analytics reduced from 1+ day to 20 minutes
- Dynamically reconfigures for patient-specific genomics acceleration



Case study: NGCodec



- Provider of **UHD video compression** technology
- Up to 50x faster vs. software H.265
- **Higher quality** video than x265 'veryslow' preset
 - Same bit rate
 - 60+ frames per second
- **Lower latency** between live stream and end viewing
- **Optimized cost**

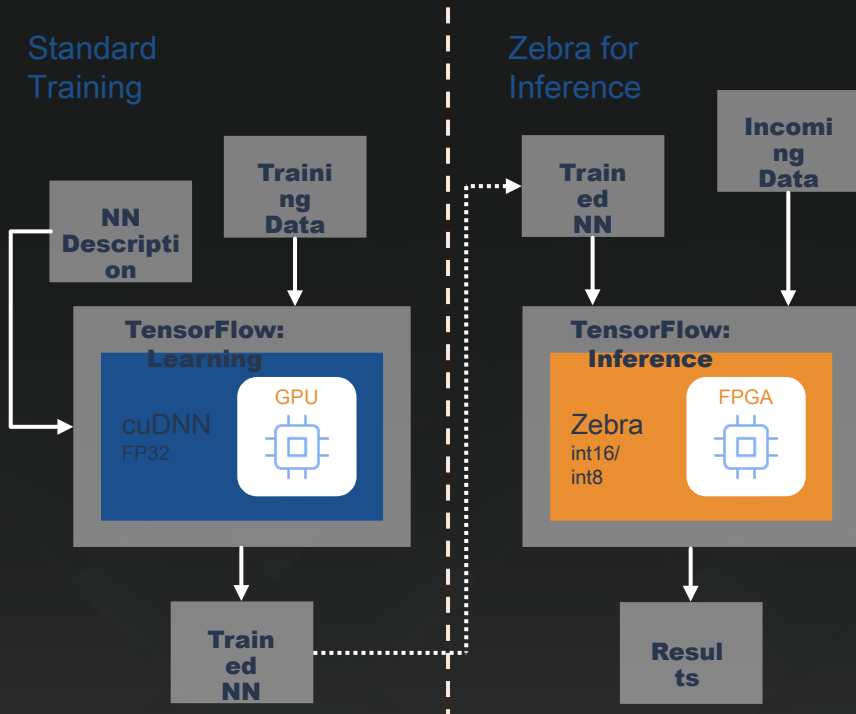
Zebra from Mipsology



- Zebra is a Deep Learning Computing Engine



- Zebra allows to replace GPU/CPU by FPGA seamlessly so users can compute their neural networks faster, with lower power, at lower cost.



Integration of Zebra in AWS EC2 F1 FPGA

- Zebra was designed to require only a PCIe and DDRs
- So, it was easy to port on AWS F1

R&D work

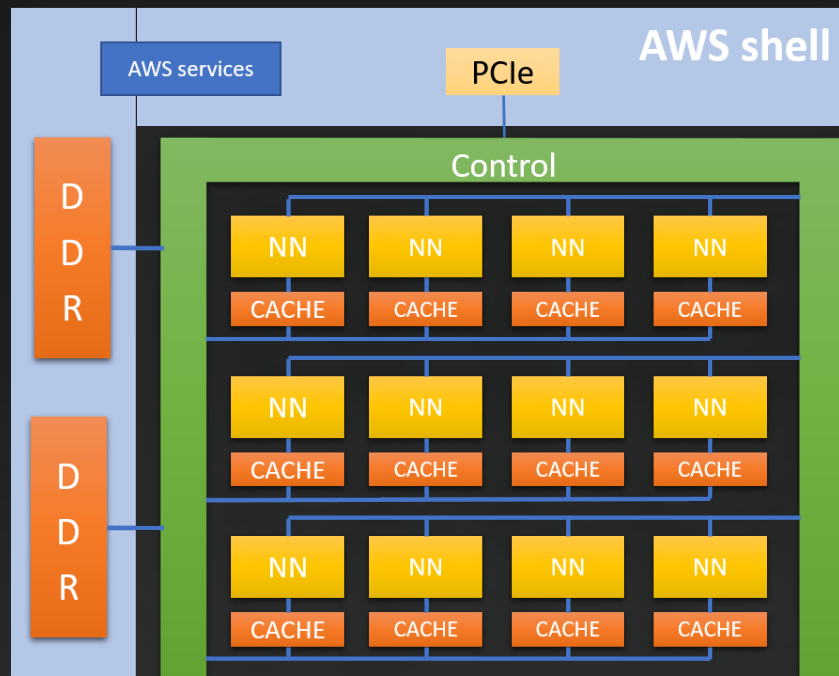
- Created a PCI-AXI bridge
- Created a AXI-2-DDR bridge
- Input clocks are provided by the shell

Challenge

- Delivering the same performances level as with a shell-free FPGA

Verification

- Used HDK BFM for the shell and the external DDR memories in simulation



Integration of Zebra in AWS EC2 F1 API

- Zebra is fully integrated in AI Frameworks, like TensorFlow, MXNet or Caffe
- From a user point of view, the Zebra AMI just makes the neural network runs on EC2 F1

R&D Work

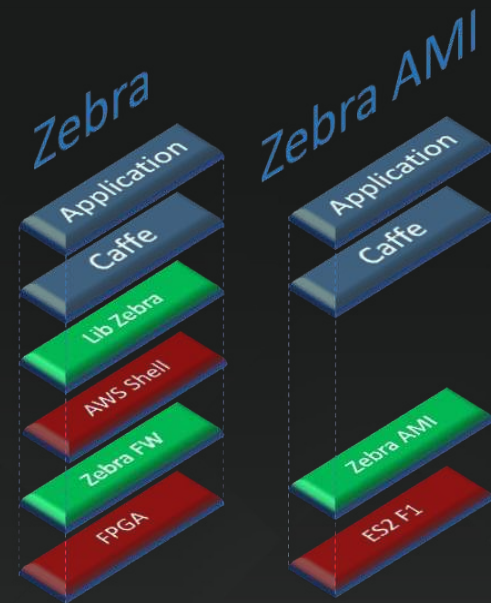
- Used Deep Learning AMI from AWS as base
- Integrated SDK library to manage FPGA loading

Validation

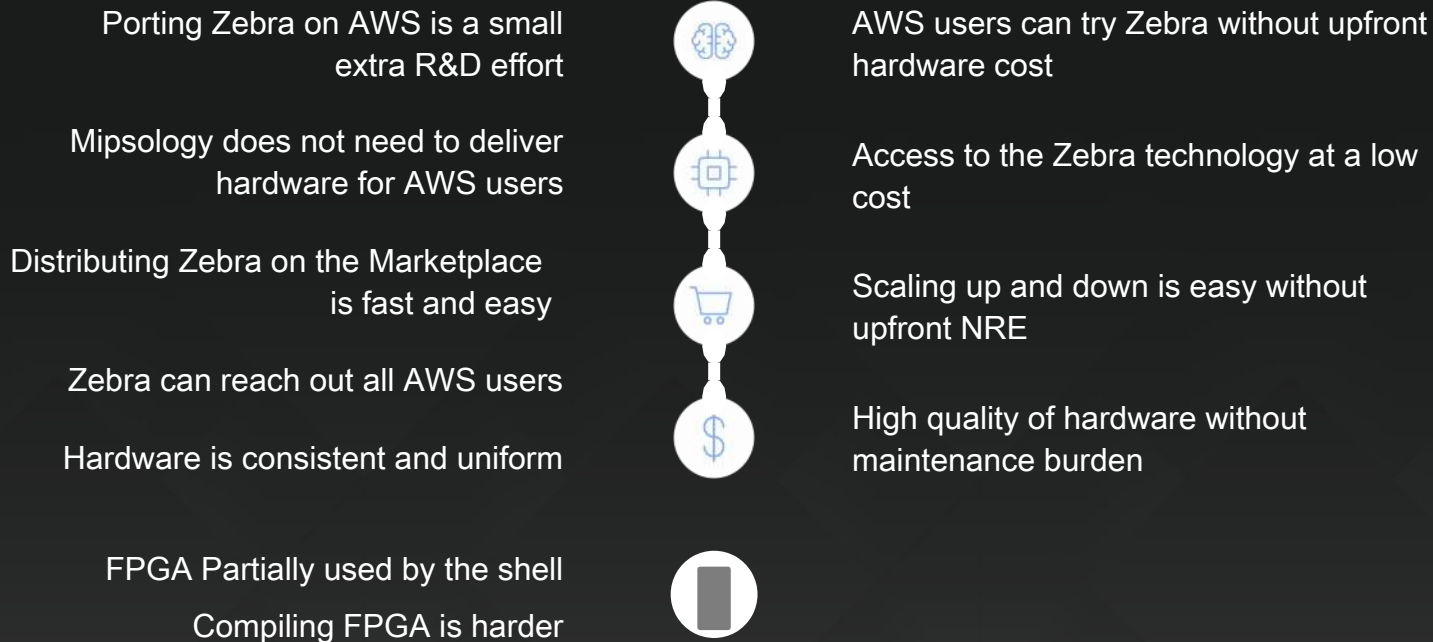
- Scripts to download the ILSVRC images data base and run a CNN quality check (i.e. top1/top5)

Delivery on AWS

- ZEBRA on 1 FPGA (image classification)
- Zebra Deep Learning engine for Caffe (1 FPGA)



Why Did We Provide Zebra on AWS EC2 F1



Get Started on AWS F1 in 5 Simple Steps

- Onboarding program with step-by-step instructions, online training and Github examples
- Setup and test your AWS environment
- Run your first SDAccel hello world!
- Build your SDAccel knowledge
- Practice and Experiment
- Install and run SDAccel on your own machine

Getting Started on AWS F1 with SDAccel

Thomas Sollaert edited this page 7 days ago · 2 revisions

This getting started guide is intended for hardware designers looking to create SDAccel applications on AWS F1 leveraging new or existing RTL code.

1. Setup and test your environment on AWS F1

This step will show you how to:

- Start an AWS F1 instance
- Download SDAccel examples from Github
- Configure and test your SDAccel environment on AWS F1

You will need the following:

- A Xilinx account - create one [here](#)

The screenshot shows the AWS website's main landing page. At the top, there's a navigation bar with links for 'Menu', 'Products', 'Solutions', 'Pricing', 'Software', 'More', 'English', 'My Account', and a 'Create an AWS Account' button. The main banner features the text 'Start Building on AWS Today' with a sub-headline: 'Whether you're looking for compute power, database storage, content delivery or other functionality, AWS has the services to help you build sophisticated applications with increased flexibility, scalability and reliability.' Below the banner, there are four columns of content: 'Broad & Deep Platform', 'Customer Success', 'Pace of Innovation', and 'Global Infrastructure'. At the bottom, there are three numbered steps: 1. 'Sign up for an AWS account', 2. 'Learn with 10-Minute Tutorials', and 3. 'Start Building with AWS'.

Resources

<https://aws.amazon.com/ec2/instance-types/f1>

<https://aws.amazon.com/ec2/instance-types/f1/partners/>

<https://github.com/aws/aws-fpga>

<https://github.com/aws/aws-fpga/blob/master/SDAccel/README.md>

<https://www.xilinx.com/>

https://github.com/Xilinx/SDAccel_Examples/wiki/Getting-Started-on-AWS-F1-with-SDAccel-and-RTL-Kernels

<http://www.mipsology.com/>

<https://aws.amazon.com/marketplace/seller-profile?id=904a5b3b-2c57-476f-95c2-3c4aeaa3ab8c>



Thank you!

Julien Simon, Principal Evangelist AI/ML, AWS

Ramine Roane, Senior Director, Product Management, Xilinx

Sébastien Delerse, Co-founder, Mipsology



XILINX
ALL PROGRAMMABLE™

Mipsology