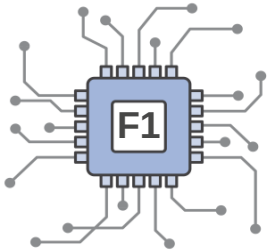
Accelerated Computing on AWS

Applications for GPUs and FPGAs



David Pellerin, Amazon Web Services

ASAP 2017, Seattle

July 11, 2017

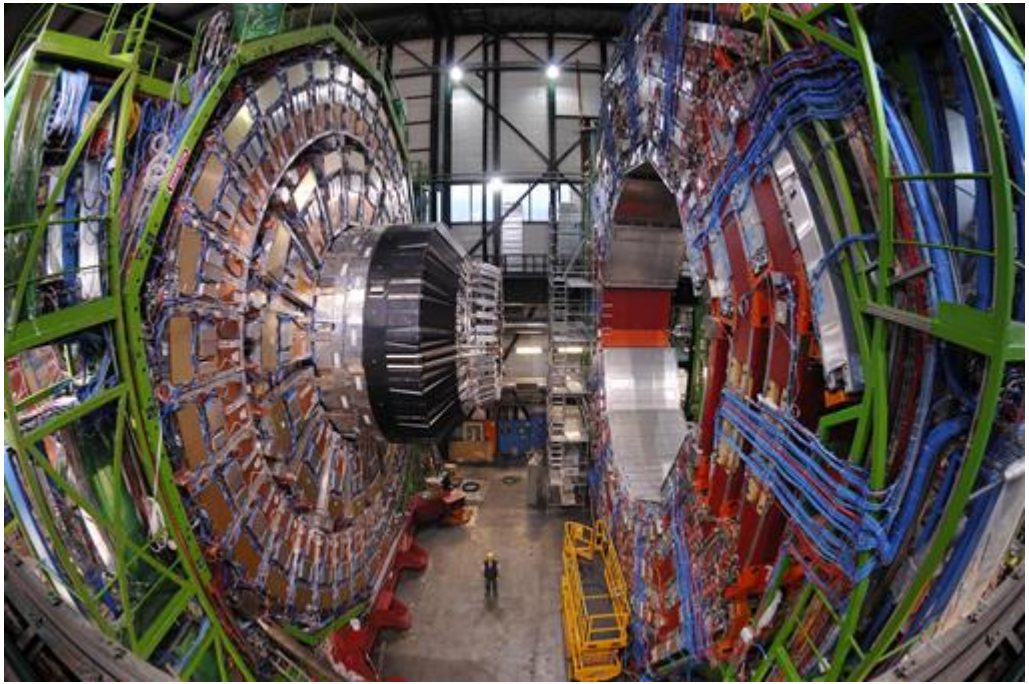
amazon web services

Data Drives Computing at Scale

Why Accelerate?

Big Compute in Big Science

Acceleration at the Edge…



50PB of data expected in 2017

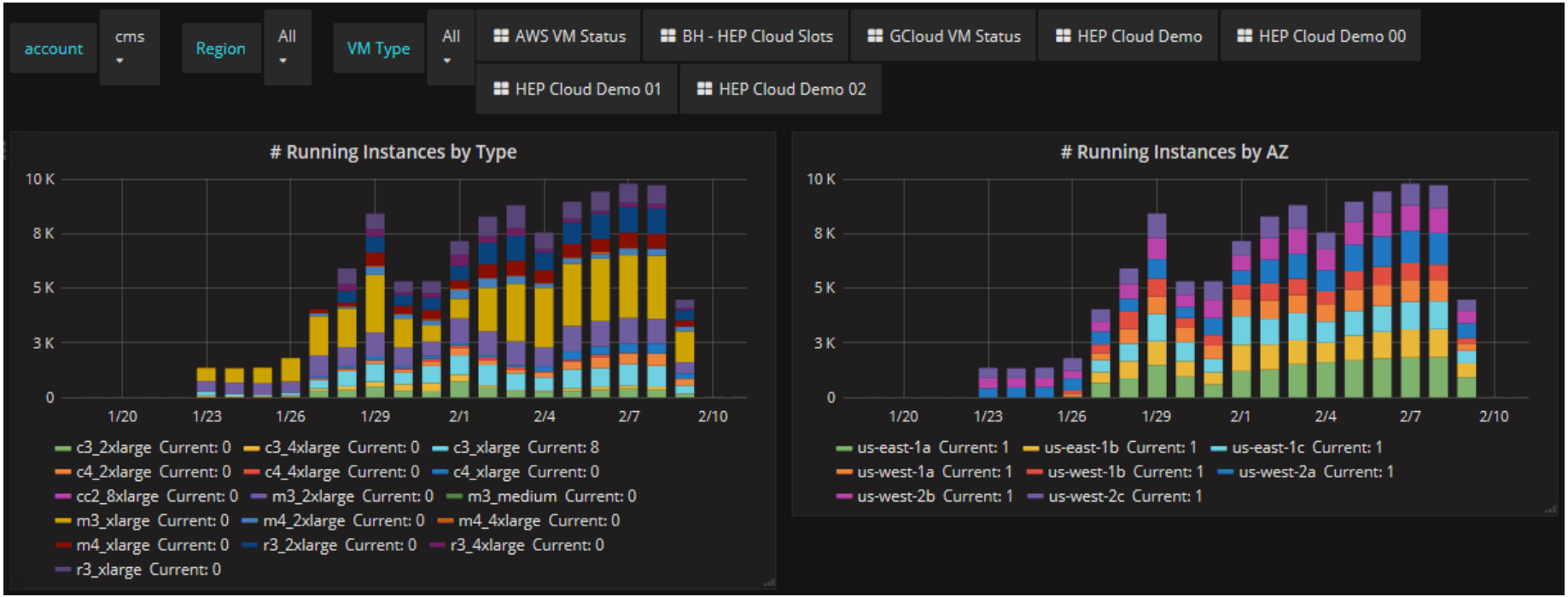
"The overall processing architecture is well matched to FPGA processing. FPGAs operate optimally using highly parallel streams with pipelined steps running at data link speed…"

A time-multiplexed track-trigger for the CMS HL-LHC upgrade, G.Hall, CMS, https://doi.org/10.1016/j.nima.2015.09.075

High Throughput Using CPUs

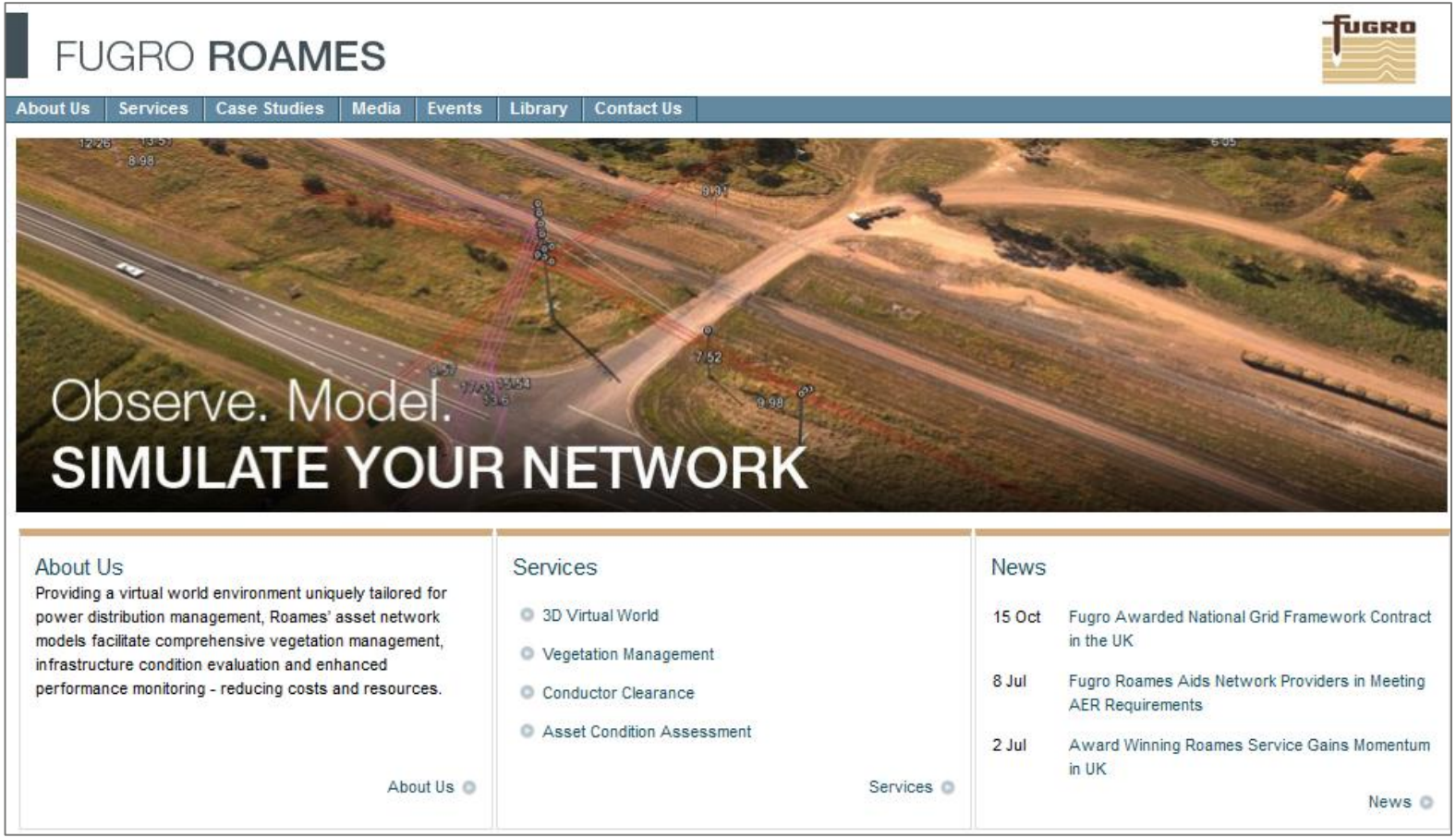
BROOKHAVEN NATIONAL LABORATORY

High Throughput Using CPUs



amazon web services

Data Drives Scale in Energy Management

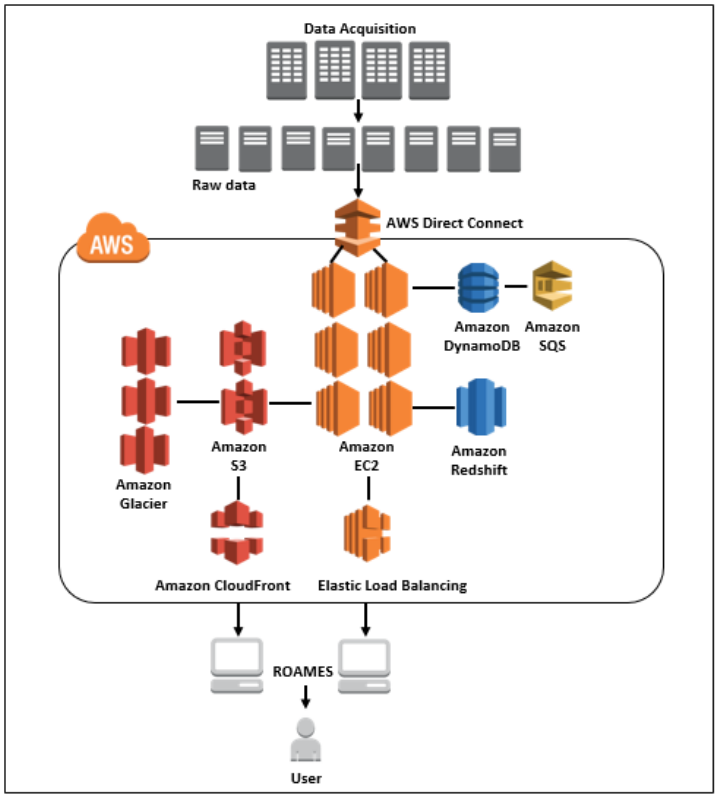


Big Data Meets Big Compute

- Aircraft equipped with cameras, laser sensors
- Repeated overflights of power networks
- Captured data is used to render detailed 3D models of the power lines, and the environment
- Analytics and simulations are run to generate actionable reports for directing post-disaster repair and prioritizing ongoing maintenance

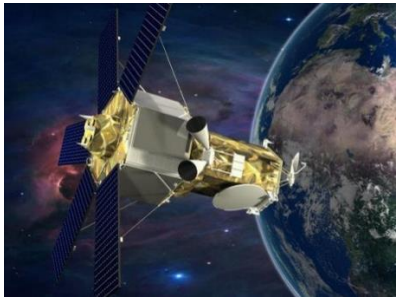
"Fugro Roames has enabled Ergon Energy to reduce the cost of vegetation management from AU$100 million to AU$60 million per year."

- Josh Passenger, Technical Architect, Fugro Roames

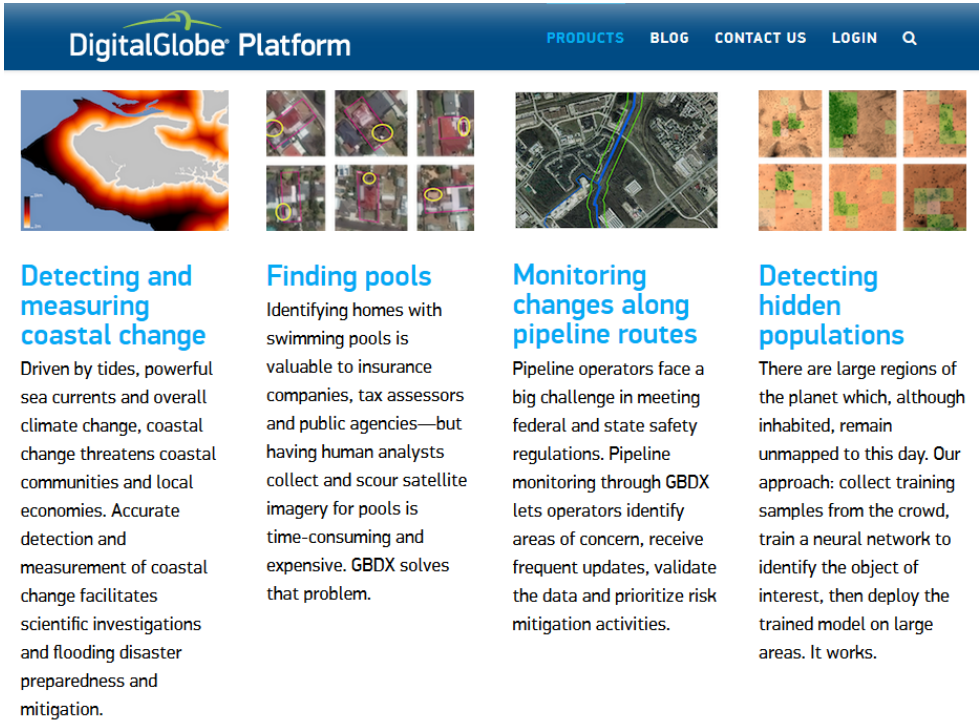


Data Should Not Be The Bottleneck

AWS Snowmobile, transporting 100PB of data



"In the past 17 years, we've collected more than 7 billion square kilometers of imagery. These are big, heavy images from a processing standpoint and add up to a lot of storage; an image from a satellite like WorldView-3 can be 30 GB. Our archive now consumes 100 petabytes of storage and increases by 10 PB per year."



Compute Should Not Be a Bottleneck Either

HGST applications for engineering:

- Molecular dynamics, CAD, CFD, EDA
  - Using CPUs and GPUs
- Collaboration tools for engineering
- Big data for manufacturing yield analysis

Running drive-head simulations at scale:

Millions of parallel parameter sweeps, running months of simulations in just hours

Over 85,000 Intel cores running at peak, using Spot Instances

amazon web services

Why Accelerated Computing in the Cloud?

Parallelism increases throughout…

CPU: High speed, low efficiency

GPU/FPGA: High throughput, high efficiency

GPUs and FPGAs can provide massive parallelism and higher efficiency than CPUs for certain categories of applications

Sample Use-Cases for Acceleration

Deep Learning

Engineering Simulations

Financial Computing

Molecular Dynamics

VR Content Rendering

Accelerated Search and Databases

Scale Drives Optimization

AWS Compute Instance Types

General purpose

T2

M4

M3

Compute optimized

Announced

C5

C4

C3

CC2

Storage and IO optimized

I3

D2

I2

HS 1

Memory optimized

X1

R4

R3

GPU and FPGA accelerated

F1 — 2017

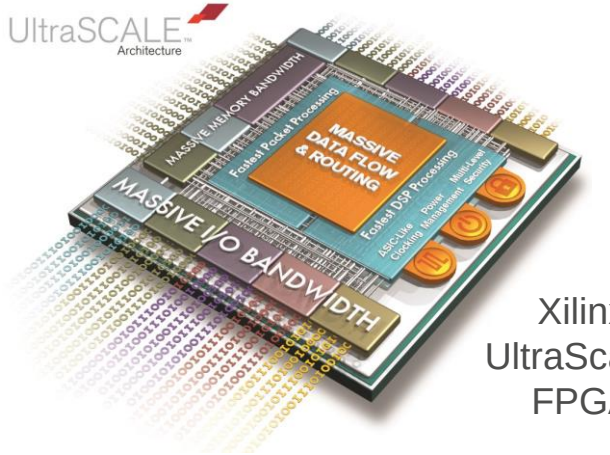
P2 — 2016

G2 — 2013

CG1 — 2011

GPU and FPGA for Accelerated Computing

NVIDIA GPU



Xilinx
UltraScale+
FPGA

P2: GPU-accelerated computing

- Enabling a high degree of parallelism – each GPU has thousands of cores
- Consistent, well documented set of APIs (CUDA, OpenACC, OpenCL)
- Supported by a wide variety of ISVs and open source frameworks

F1: FPGA-accelerated computing

- Massively parallel – each FPGA includes millions of parallel system logic cells
- Flexible – no fixed instruction set, can implement wide or narrow datapaths
- Programmable using available, cloud-based FPGA development tools

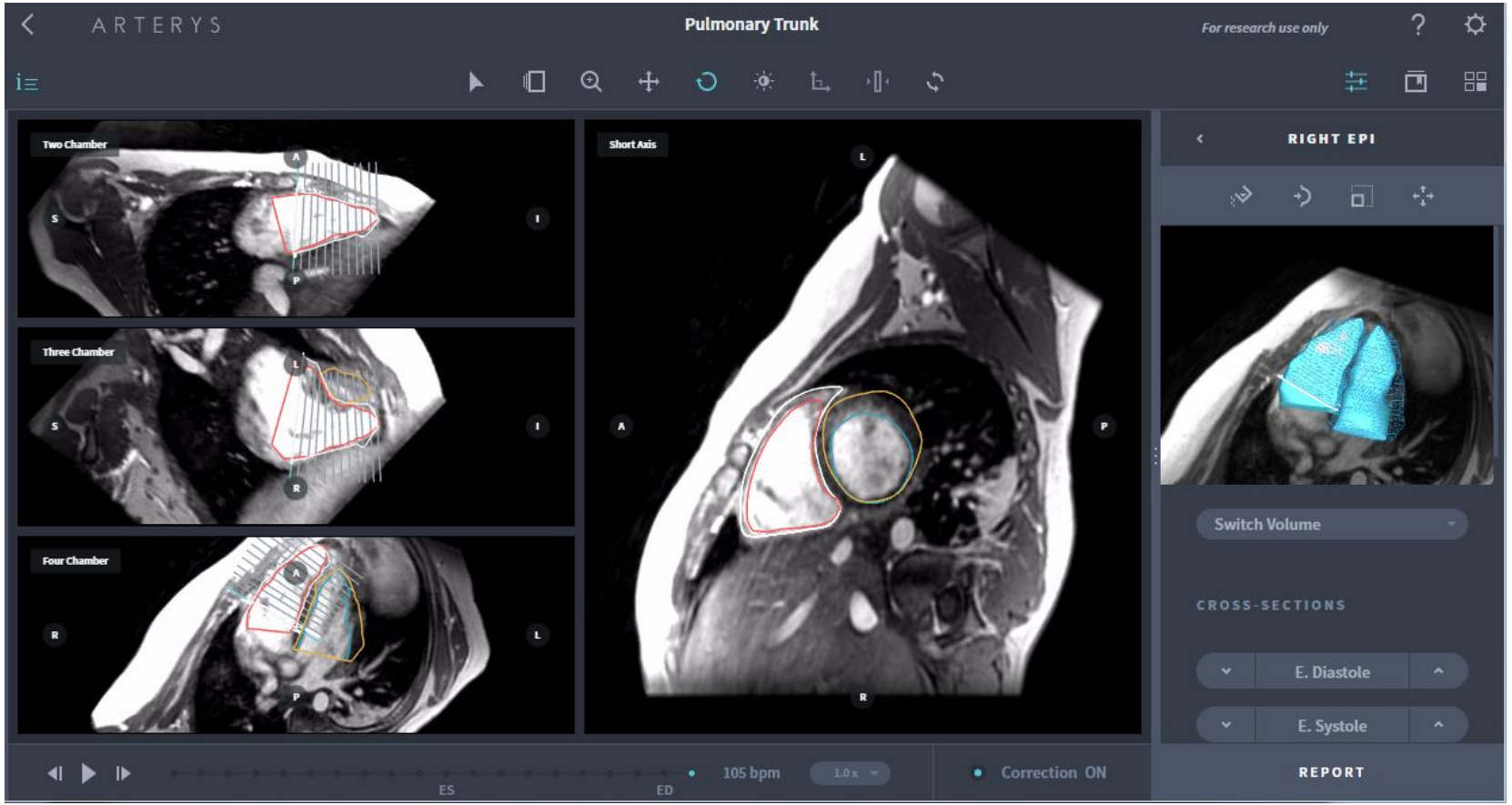
GPU Computing

P2 GPU Instance Types on AWS

- Up to 16 NVIDIA GK210 GPUs (8 K80 cards) in a single instance
- The **16xlarge** size provides:
  - A combined 192 GB of GPU memory, 40 thousand CUDA cores
  - 70 teraflops of single precision floating point performance
  - Over 23 teraflops of double precision floating point performance
- Example workloads include:
  - Deep learning training, engineering simulations, computational finance, seismic analysis, molecular modeling, VR rendering, accelerated databases

amazon web services

Accelerating Medical Imaging



GPUs for Risk Modeling & Hedging

"Using AWS helps us **reduce a 10-day process to 10 minutes**. That's transformative: it broadens our ability to discover."

Peter Phillips
Managing Director, Aon Benfield Securities

UK-based Aon plc, the ultimate parent company of Aon Benfield Securities, is a leading global provider of risk management, insurance and reinsurance brokerage

The Challenge
Spinning up up large numbers of GPUs quickly and inexpensively to meet ABSI's customers financial modeling & reporting needs

ABSI uses proprietary Monte Carlo algorithms running millions of iterations

The solution
ABSI moved its infrastructure to AWS and deprecated its co-located data center

ABSI built a front-end on AWS for its processing solution, automatically running GPU instances on Amazon EC2 using EBS in an Amazon VPC for security.

The Result
Can be as much as 500 times more efficient in terms of performance per dollar for some clients

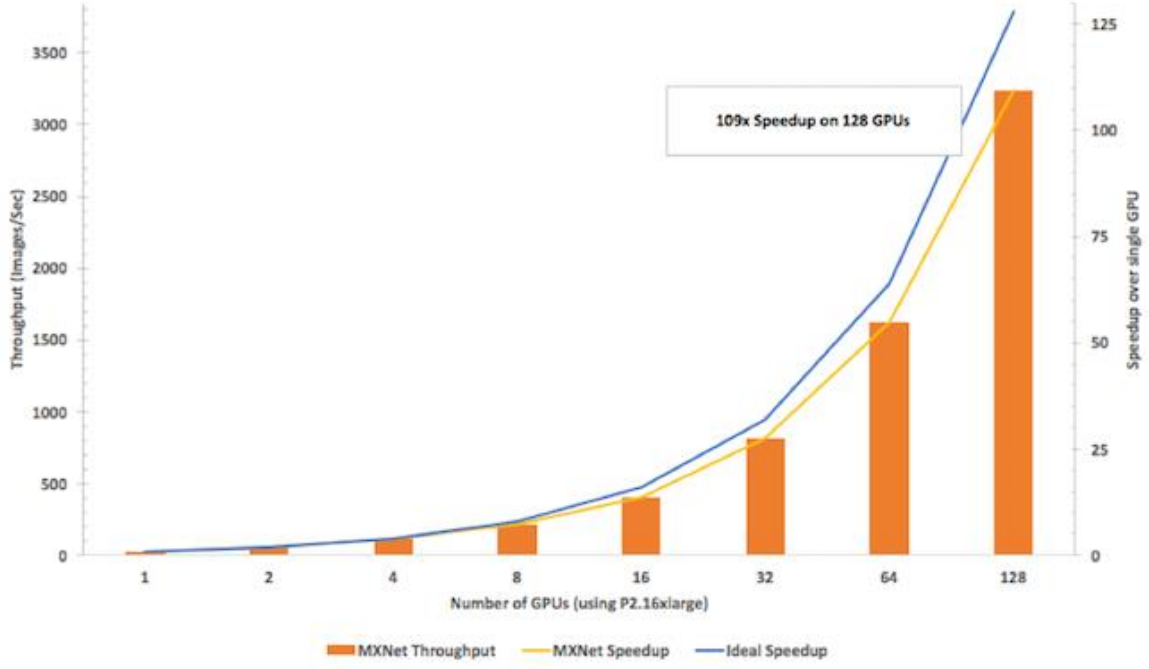
Deep Learning on AWS

P2 GPU instances for high performance DL training and inference

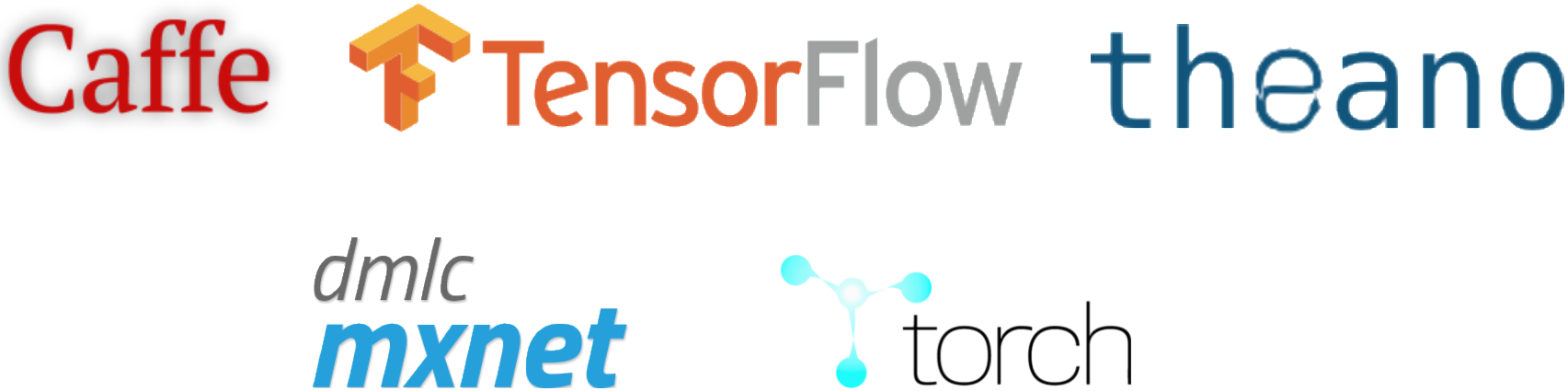
MXNet training on EC2 P2 instances:

We trained a popular image analysis algorithm, Inception v3, using MXNet and running on P2 instances

MXNet had the fastest throughput of any library we evaluated (as measured by the number of images trained per second), and the throughput rose by almost the same rate as the number of GPUs used for training, with a scaling efficiency of 85%.

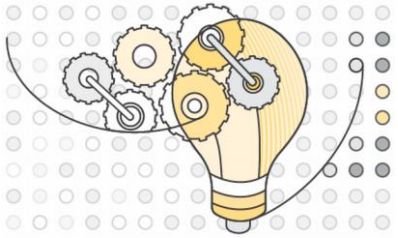


Deep Learning Frameworks



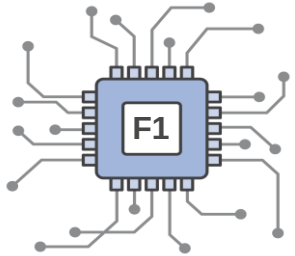
Sample Deep Learning Use Cases

- Recommendation engines
- Voice recognition / chat bots
- Fraud and intrusion detection
- Picture recognition and tagging
- Document tagging and classification
- Autonomous driving and robotics



FPGA Computing

FPGAs on AWS

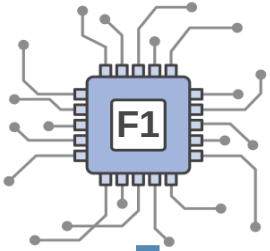


- Not an experiment!
- Driven by customer demand
- Compelling, at-scale use-cases:
  - Image and video processing
  - Analytics and machine learning
  - Genomics and proteomics
  - Financial computing
  - Engineering simulations
  - And more

FPGA Acceleration in the AWS Cloud: Goals

- **Make FPGAs available as standard AWS instances** to a large community of developers, and to millions of potential end-customers
- **Simplify the development process** by providing cloud-based FPGA development tools
- **Allow developers to focus on algorithm design**, by abstracting FPGA I/O using well-defined interfaces
- **Provide a Marketplace for FPGA applications**, providing more choice and easy access for all AWS customers

Guiding principle: allow FPGAs to be included in a customer's deployment as easily as any other AWS instance type or service

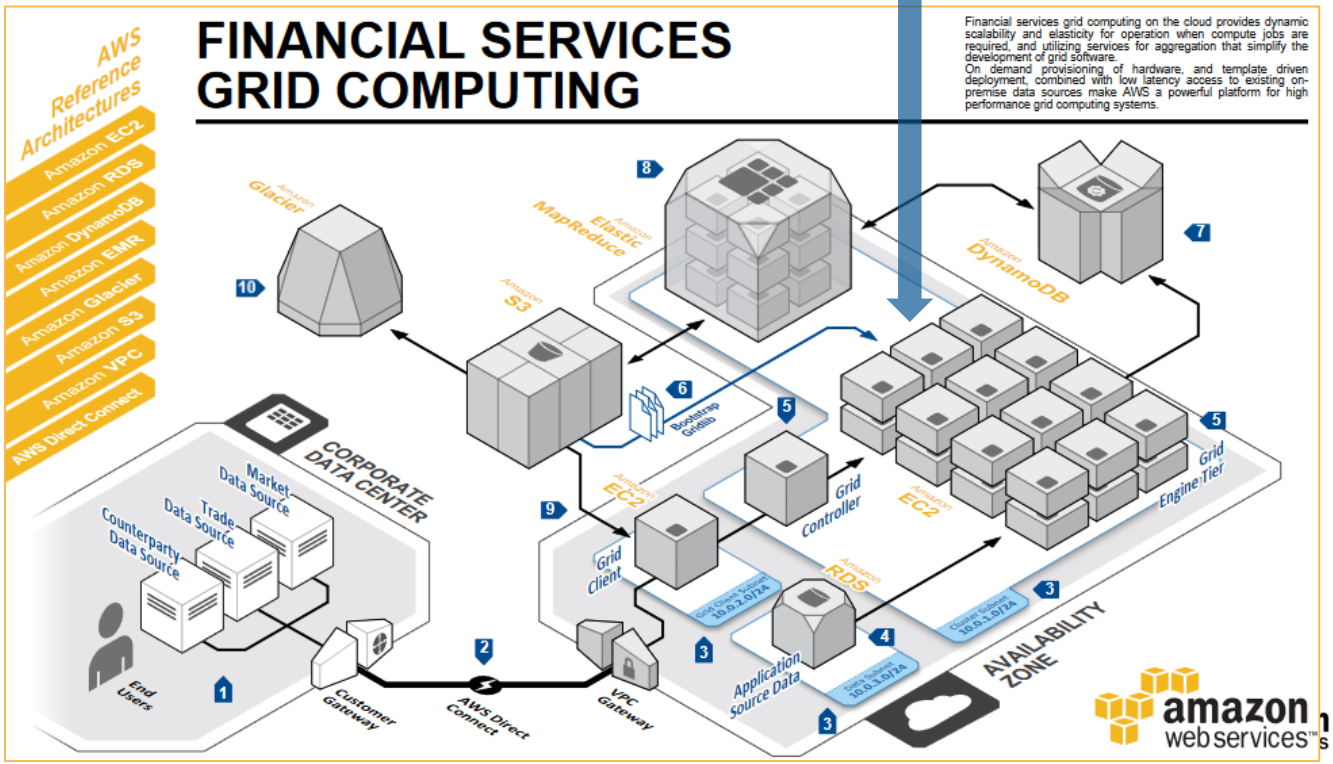


Important concepts:

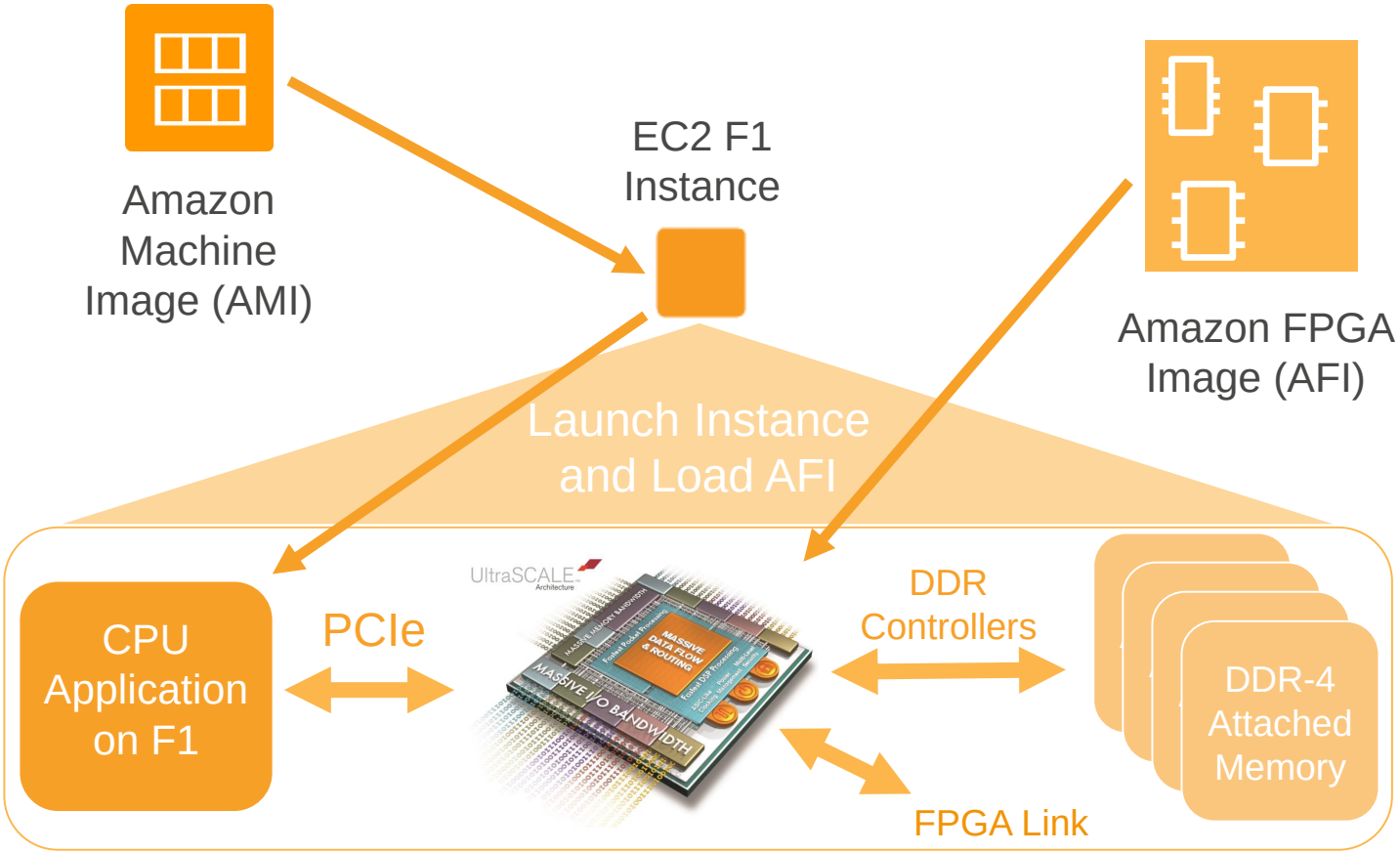
Region
Availability Zone (AZ)
Virtual Private Cloud (VPC)
Elastic Compute Cloud (EC2)
Amazon Machine Image (AMI)
EC2 Instance
AWS Marketplace

Additional for F1:

FPGA Developer AMI
Amazon FPGA Image (AFI)

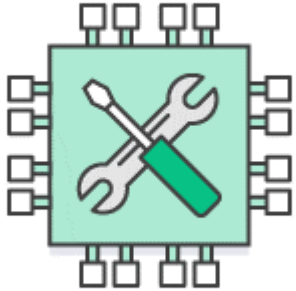


FPGA Acceleration Using F1



An F1 instance can have any number of AFIs

An AFI can be loaded into the FPGA in less than 1 second

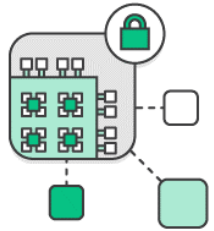


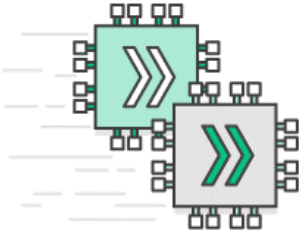
FPGA Programming Flow

Developing Applications for F1

Development steps

1. Launch the AWS-provided FPGA Developer AMI, which includes all needed FPGA design and programming software, as well as the AWS FPGA Hardware Development Kit (HDK)
2. Use Xilinx Vivado or SDAccel software and a hardware description language (Verilog, VHDL, or OpenCL) with the HDK to describe and simulate your custom FPGA logic
3. After successful simulation, use Vivado or SCAccel to synthesize and place/route the FPGA logic to create an FPGA Design Check Point (DCP), encrypt, and generate an Amazon FPGA Image (AFI)
4. Launch an F1 instance and load the AFI to the FPGA, using AFI management tools provided by AWS





FPGA Developer AMI

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 tools and run time tools required to develop and use custom FPGAs for hardware acceleration. The FPGA developer AMI includes a prepackaged tool development environment, with scripts and tools for simulating your FPGA design, compiling code, building and registering your AFI (Amazon FPGA Image). Developers can deploy the FPGA developer AMI on an Amazon EC2 instance and quickly provision the resources they need to write... Read more

Customer Rating	★★★★★ (0 Customer Reviews)
Latest Version	1.2.1
Operating System	Linux/Unix, CentOS 7.3
Delivery Method	64-bit Amazon Machine Image (AMI) (Read more)
Support	See details below
AWS Services Required	Amazon EC2, Amazon EBS
Highlights	▪ Xilinx Vivado 2017.1 and 2016.4 SDx - Free license for F1 FPGA development ▪ AWS Integration - includes packages and configurations that provide tight integration with Amazon Web Services

Product Description

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 tools and run time tools required to develop and use custom FPGAs for hardware acceleration. The FPGA developer AMI includes a prepackaged tool development environment, with scripts and tools for simulating your FPGA design, compiling code, building and registering your AFI (Amazon FPGA Image). Developers can deploy the FPGA developer AMI on an Amazon EC2 instance and quickly provision the resources they need to write and debug FPGA designs in the cloud. The AMI is designed to provide a stable, secure, and high performance development environment. The FPGA AMI is provided at no additional charge to Amazon EC2 users.

Continue You will have an opportunity to review your order before launching or being charged.

Pricing Information

Use the Region dropdown selector to see software and infrastructure pricing information for the chosen AWS region.

For Region

US East (N. Virginia)

Pricing Details

Software pricing is based on your chosen options, such as subscription term and AWS region. Infrastructure prices are estimates only. Final prices will be calculated according to actual usage and reflected on your monthly report.

1 Software Pricing

The data below shows pricing per instance for services hosted in US East (N. Virginia).

FPGA Developer AMI - Hourly

EC2 Instance Type	Software /hr	EC2 /hr	Total /hr
c4.4xlarge	$0.00	$0.796	**$0.796**
c4.8xlarge	$0.00	$1.591	**$1.591**

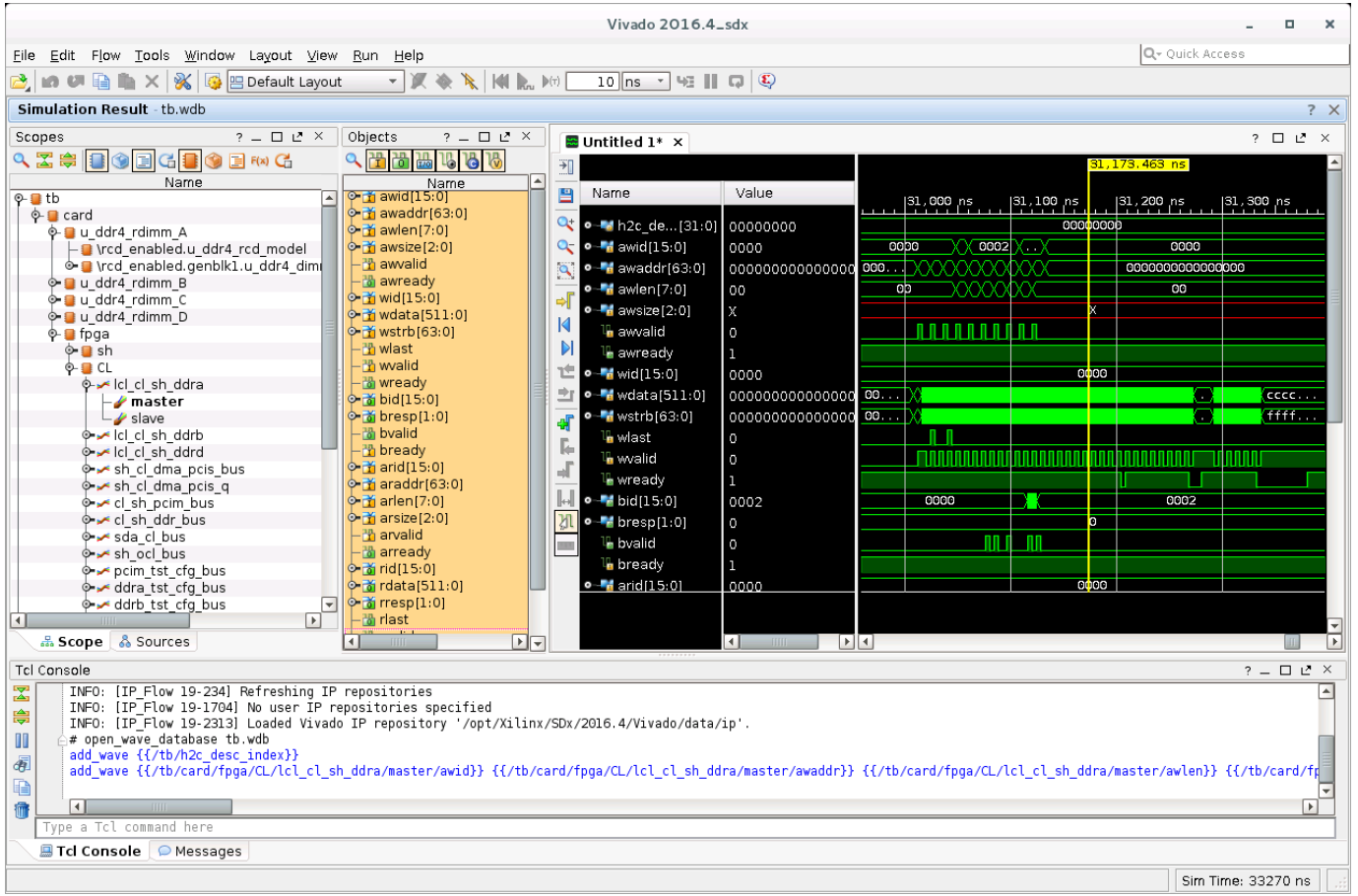
1 Software Pricing

The data below shows pricing per instance for services hosted in US East (N. Virginia).

FPGA Developer AMI - Hourly

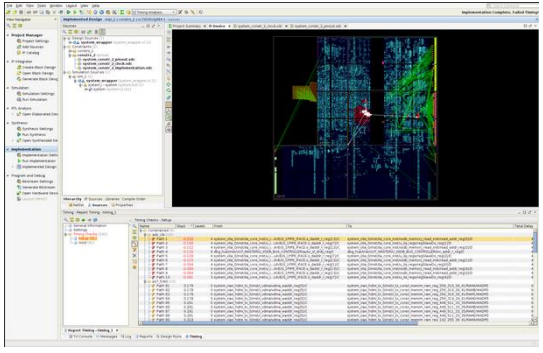
EC2 Instance Type	Software /hr	EC2 /hr	Total /hr
c4.4xlarge	$0.00	$0.796	**$0.796**
c4.8xlarge	$0.00	$1.591	**$1.591**
m4.2xlarge	$0.00	$0.431	**$0.431**
m4.4xlarge	$0.00	$0.862	**$0.862**
m4.10xlarge	$0.00	$2.155	**$2.155**
m4.16xlarge	$0.00	$3.447	**$3.447**
t2.2xlarge	$0.00	$0.376	**$0.376**
f1.2xlarge	$0.00	$1.65	**$1.65**
f1.16xlarge	$0.00	$13.20	**$13.20**
r4.xlarge	$0.00	$0.266	**$0.266**
r4.2xlarge	$0.00	$0.532	**$0.532**
r4.4xlarge	$0.00	$1.064	**$1.064**
r4.8xlarge	$0.00	$2.128	**$2.128**
r4.16xlarge	$0.00	$4.256	**$4.256**

Xilinx Vivado on AWS

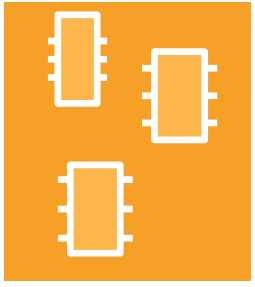


Developing Applications for F1 – AFI Creation

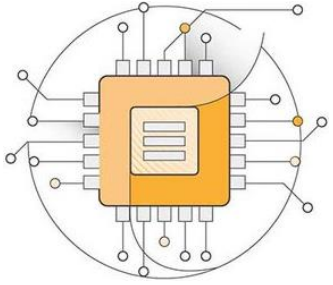
DCP



FPGA Development using Xilinx Vivado on C4 or M4 instance



Generate an Amazon FPGA Image (AFI)

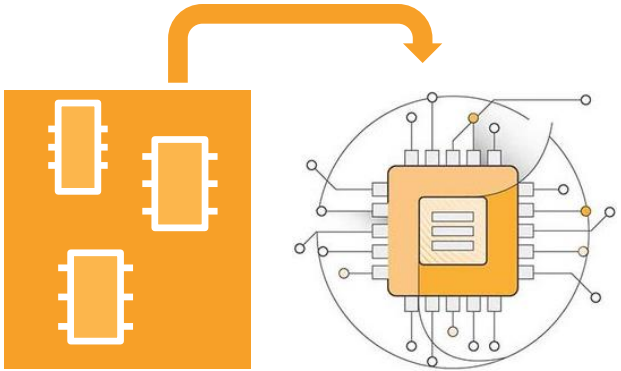


Securely deploy AFI on one or more F1 instances

```
$ aws ec2 create-fpga-image \
    --name <afi-name> \
    --description <afi-description> \
    --input-storage-location Bucket=<dcp-bucket-name>,Key=<path-to-tarball> \
    --logs-storage-location Bucket=<logs-bucket-name>,Key=<path-to-logs> \
    [ --client-token <value> ] \
    [ --dry-run | --no-dry-run ]
```

Developing Applications for F1

Launch an F1 instance and download the AFI to the FPGA, using AFI management tools provided by AWS



Generate an Amazon FPGA Image (AFI)

Deploy AFI on one or more F1 instances

SDK

Management options:

[A] Shell FPGA Management Tools

[B] C-library FPGA Management

[C] OpenCL runtime library

Runtime code for I/O:

[D] FPGA PCIe Lib

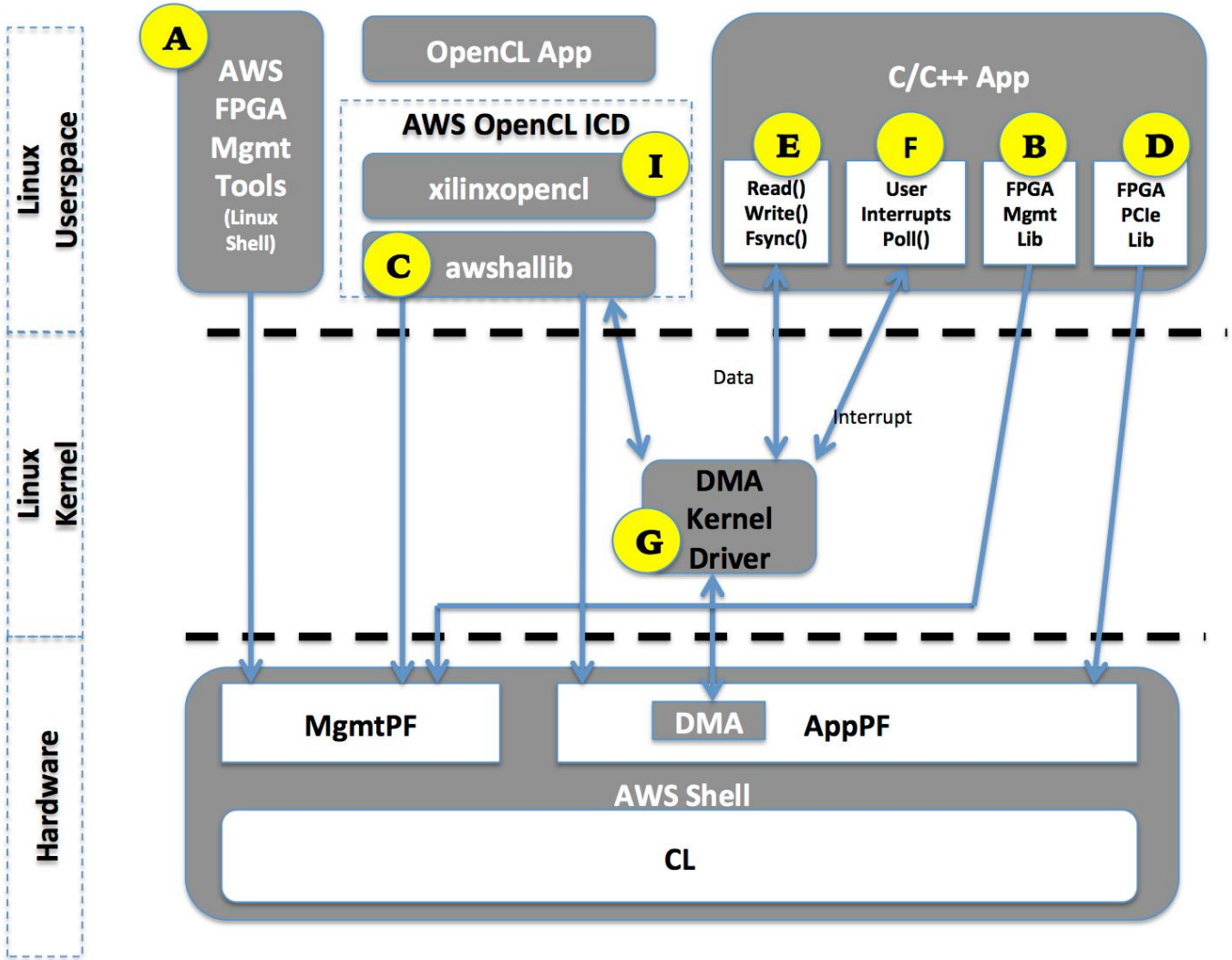
[E] DMA Interface

[F] Interrupt/Event notification

[I] OpenCL Installable Client Driver

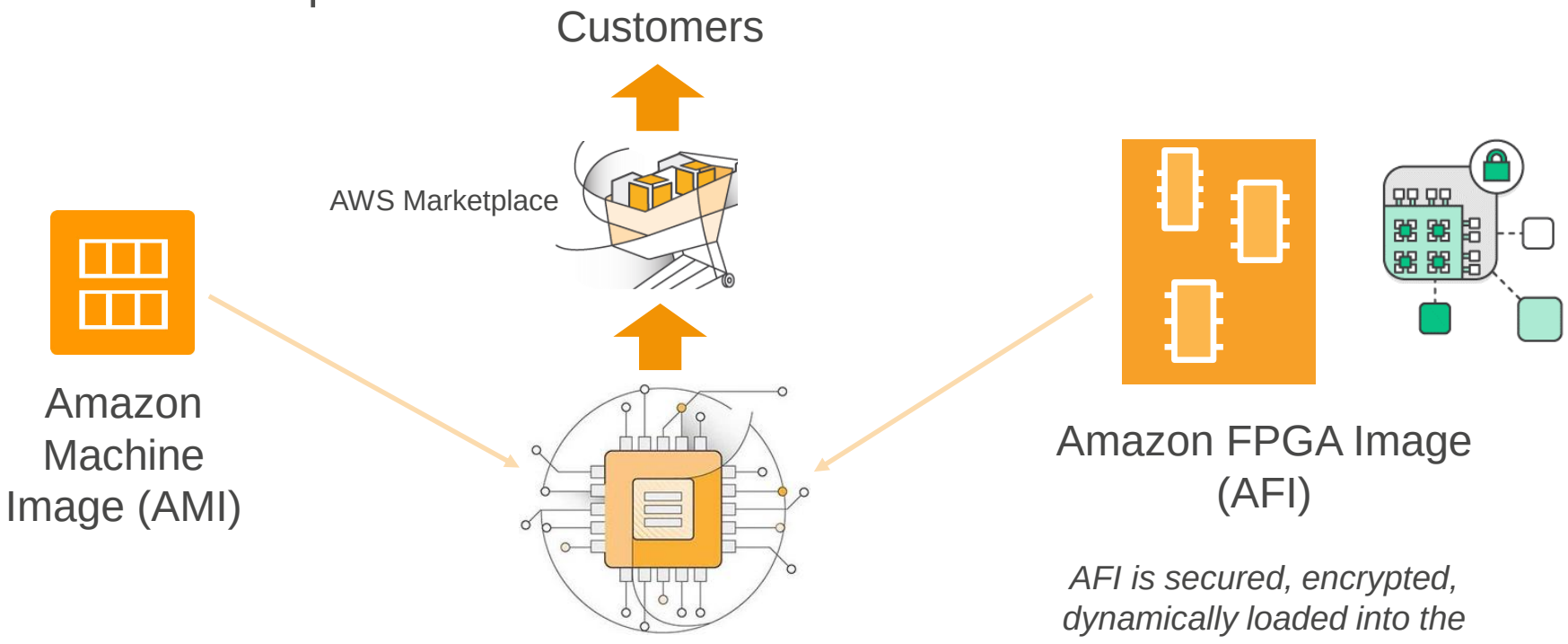
Linux Kernel Driver:

[G] DMA Kernel Driver



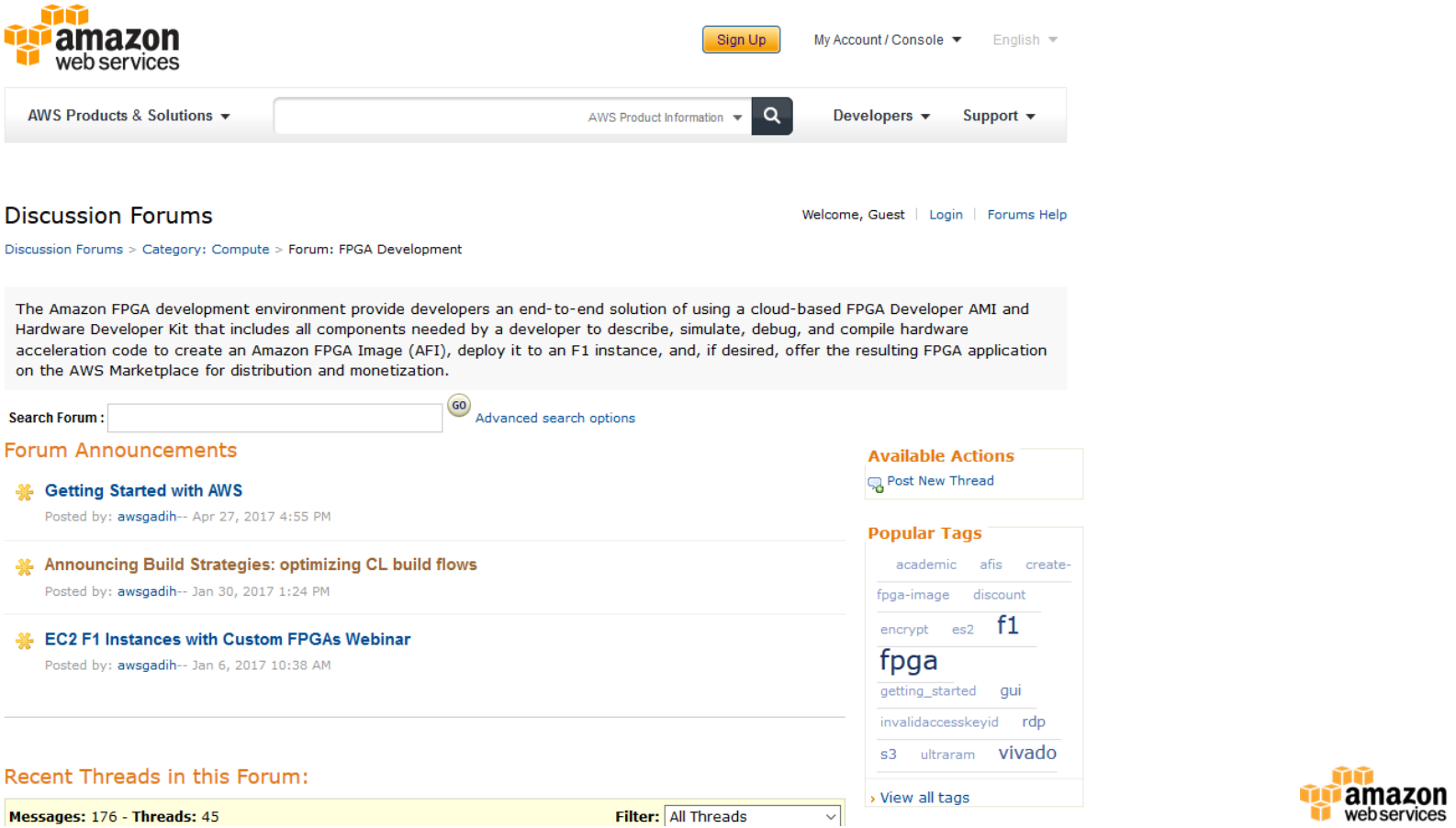
Delivering FPGA Partner Solutions

via AWS Marketplace



AFI is secured, encrypted, dynamically loaded into the FPGA - can't be copied or downloaded

F1 Discussion Forum at forums.aws.amazon.com/



amazon web services

aws educate

aws.amazon.com/education/awseducate/
https://aws.amazon.com/grants/

Institutions

Provide educators and students with resources for cloud-related learning. Those at member institutions receive twice as many AWS credits, demos and special on-campus programs.

Apply for AWS Educate for Institutions »

Educators

Professors, teaching assistants, and educators receive access to AWS technology, open source content for their courses, training resources, and a community of cloud evangelists.

Apply for AWS Educate for Educators »

Students

Students receive credits for hands-on experience with AWS technology, training, content, career pathways, and job board.

Apply for AWS Educate for Students »

amazon web services

Thank you!

David Pellerin dpelleri@amazon.com