Developing Cloud Scale FPGA Accelerations Using AWS F1

Gadi Hutt, Amazon Web Services

What is the Cloud ?

The AWS Platform

Account Support

- Support
- Managed Services
- Professional Services
- Partner Ecosystem
- Training & Certification
- Solution Architects
- Account Management
- Security & Pricing Reports
- Technical Acct. Management

Marketplace

- Business Applications
- DevOps Tools
- Business Intelligence
- Security
- Networking
- Database & Storage
- SaaS Subscriptions
- Operating Systems

Mgmt. Tools

- Monitoring
- Auditing
- Service Catalog
- Server Management
- Configuration Tracking
- Optimization
- Resource Templates
- Automation

Analytics

- Query Large Data Sets
- Elasticsearch
- Business Analytics
- Hadoop/Spark
- Real-time Data Streaming
- Orchestration Workflows
- Managed Search
- Managed ETL

Dev Tools

- Private Git Repositories
- Continuous Delivery
- Build, Test, and Debug
- Deployment

AI

- Deep L. Camera
- Transcribe Translate
- Managed Models
- Video
- Voice & Text Chatbots
- Machine Learning
- Text-to-Speech. NLP
- Image Analysis

IoT

- Defender
- Management
- Analytics
- Rules Engine
- Local Compute and Sync
- Device Shadows
- Device Gateway
- Registry

Mobile

- Build, Test, Monitor Apps
- Push Notifications
- Build, Deploy, Manage APIs
- Device Testing
- Identity

Enterprise Applications

- Document Sharing
- Email & Calendaring
- Hosted Desktops
- Application Streaming
- Backup

Media

- Store/Convert
- Live

VR

- VR/AR

Game Development

- 3D Game Engine
- Multi-player Backends

Migration	Application Discovery	Application Migration	Data Migration	Database Migration	Server Migration				
Hybrid	Data Integration	Integrated Networking	Identity Federation	Resource Management	VMware on AWS	Devices & Edge Systems			
Application Services	Transcoding	Step Functions	Messaging						
Security	Identity & Access	Key Storage & Management	Active Directory	DDoS Protection	Application Analysis	Certificate Management	Web App. Firewall	Threat detection	
Database	Aurora	MySQL	PostgreSQL	Oracle	SQL Server	MariaDB	Data Warehousing	NoSQL	Graph
Storage	Object Storage	Archive	Exabyte-scale Data Transport	Block Storage	Managed File Storage	Select			
Compute	Virtual Machines	Simple Servers	Web Applications	Auto Scaling	Batch	Containers	Event-driven Computing	Bare-Metal	
Networking	Isolated Resources	Dedicated Connections	Global CDN	Load Balancing	Scalable DNS				
Infrastructure	Regions	Availability Zones	Points of Presence						

The benefits of the AWS Cloud

SPEED

agility and speed of innovation

EFFICIENCY

cost savings

ELASTICITY

scale quickly as needed

BREADTH

range of functionality

GLOBAL

go global in minutes

Focus on the core mission

Lower the time spent on infrastructure

Concentrate on new business initiatives

Dedicate more resources to innovation

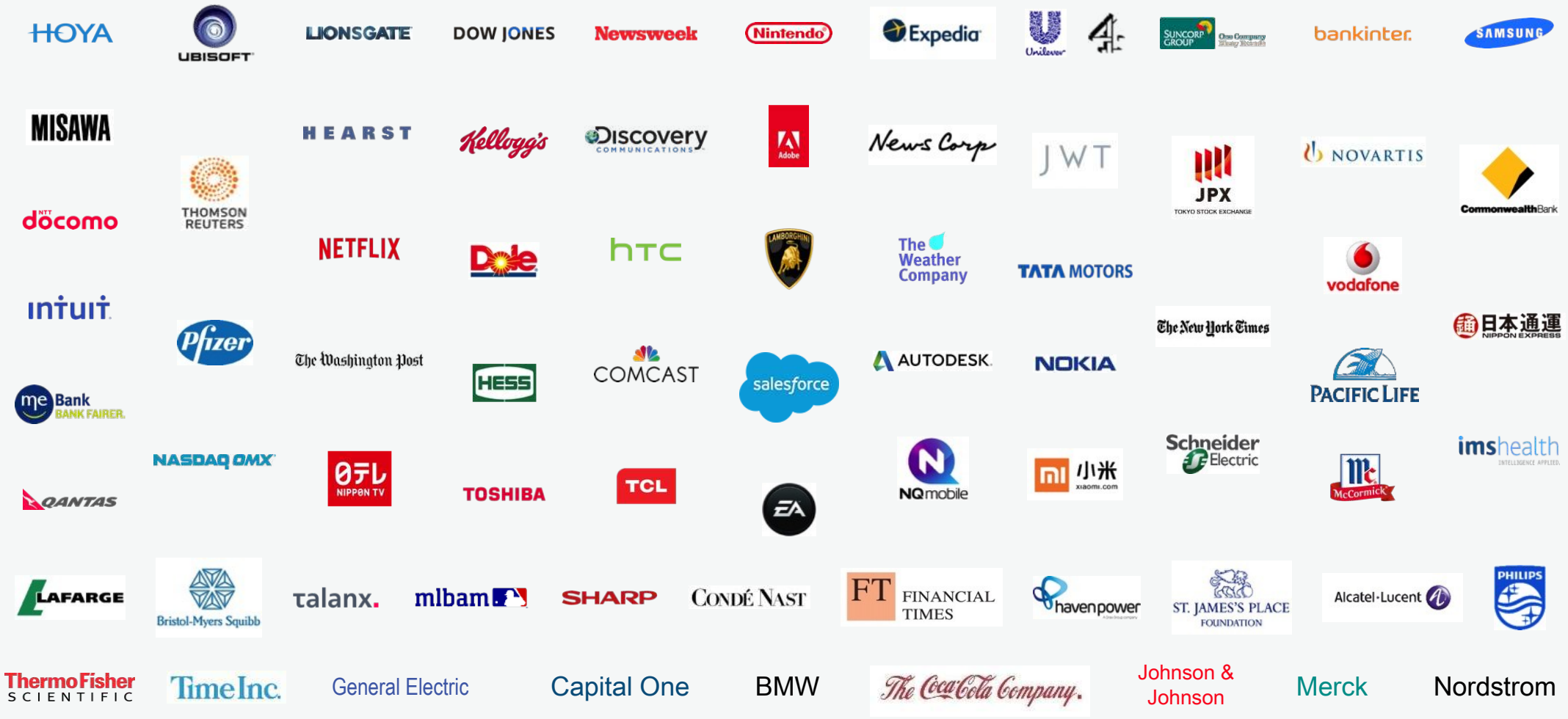
"3M HIS is not in the IT business. Rather, we are a healthcare-analytics company. We are getting out of IT operations, and by going to AWS, we can focus our R&D team on the science of healthcare. For us that means analytics rather than IT, enabling us to attack the healthcare industry's cost and quality challenges"

David Frazee

Chief Technology Officer

3M Health Information Systems

Global Enterprise Customers



AWS EC2 compute instance types

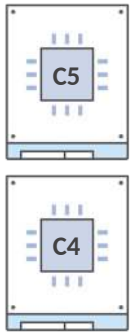
General Purpose

T2

M4

M3

Compute Optimized



Storage and I/O Optimized

I3

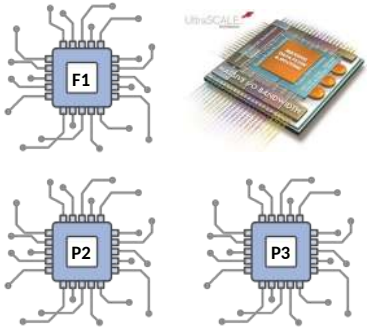
D2

Memory Optimized

X1

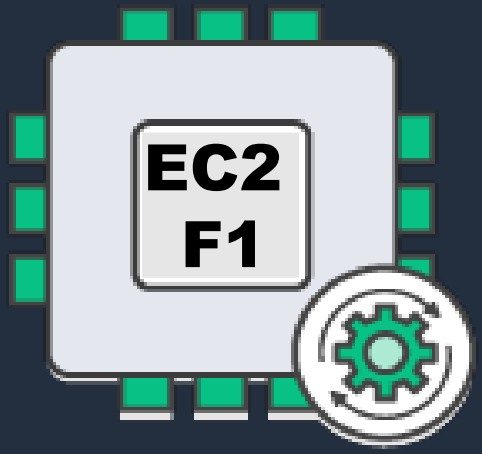
R4

GPU, FPGA Accelerated



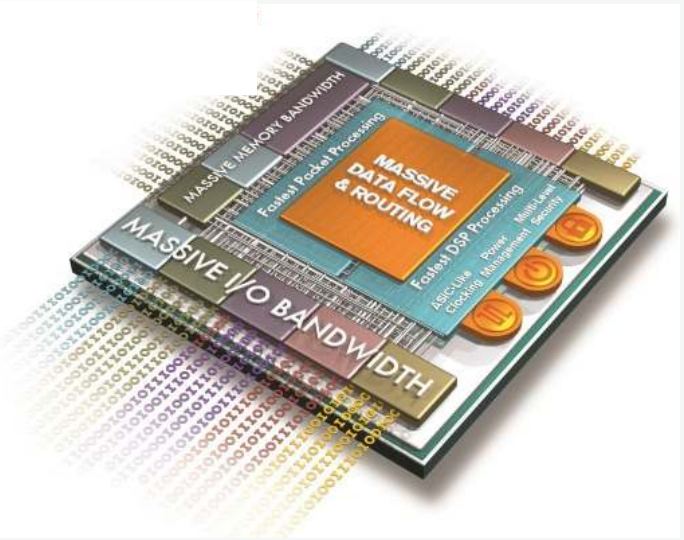
Democratizing Hardware Accelerations

- **Make FPGAs available as standard AWS instances**
- **Simplify the development process**
- **Allow developers to focus on algorithm design**
- **Provide a Marketplace for FPGA applications**



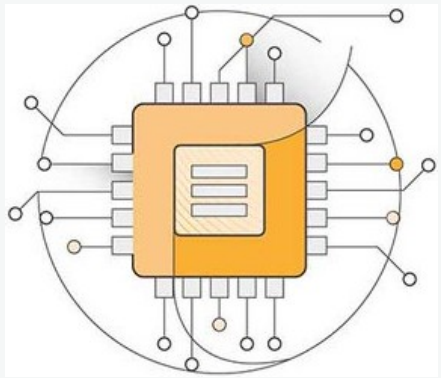
How Application Acceleration Works

FPGA handles compute-intensive, deeply pipelined operations



application

```
module filter1 (clock, rst, strm_in, strm_out)
   integer i,j; //index for loops
   always@(posedge clock)
   for (i=0; i<NUMUNITS; i=i+1)
      tmp_kernel[j] = k[i*OFFSETX];
```



CPU handles the rest

F1 FPGA instance types on AWS

- Up to 8 Xilinx UltraScale+ 16nm VU9P FPGA devices in a single instance
- The **f1.16xlarge** size provides:
  - 8 FPGAs, each with over 2 million customer-accessible FPGA programmable logic cells and over 5000 programmable DSP blocks
  - Each of the 8 FPGAs has 4 DDR-4 interfaces, with each interface accessing a 16GiB, 72-bit wide, ECC-protected memory

Instance Size	FPGAs	DDR-4 (GiB)	vCPUs	Instance Memory (GiB)	NVMe Instance Storage (GB)	Network Bandwidth
f1.2xlarge	1	4 x 16	8	122	1 x 470	Up to 10 Gbps
f1.16xlarge	8	32 x 16	64	976	4 x 940	25 Gbps

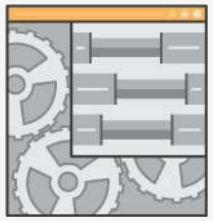
FPGA acceleration using F1

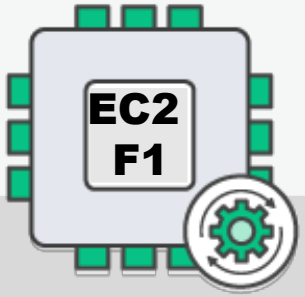
Amazon FPGA Image (AFI)

An F1 instance can have any number of AFIs

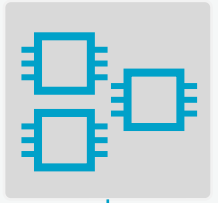
An AFI can be loaded into the FPGA in seconds

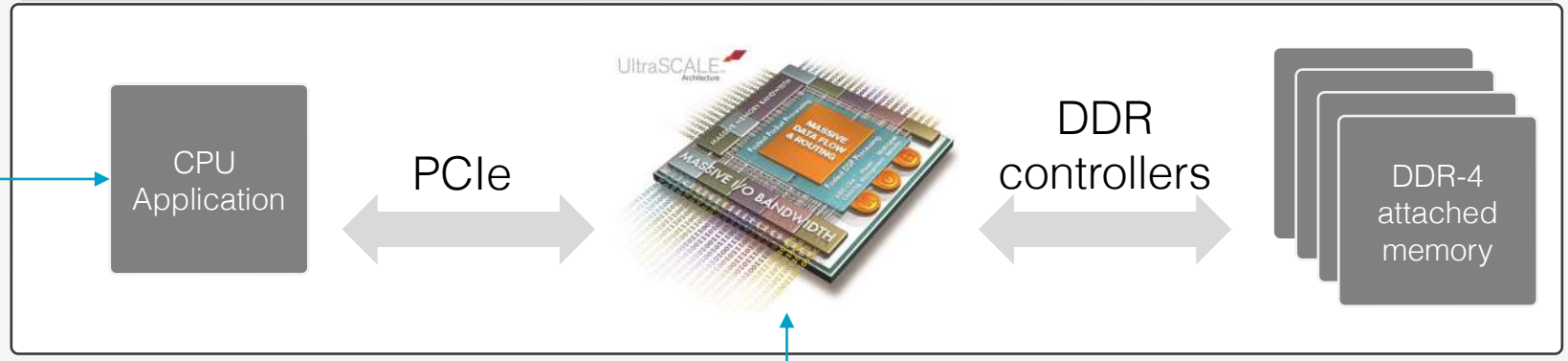
Amazon Machine Image (AMI)





Launch instance and load AFI



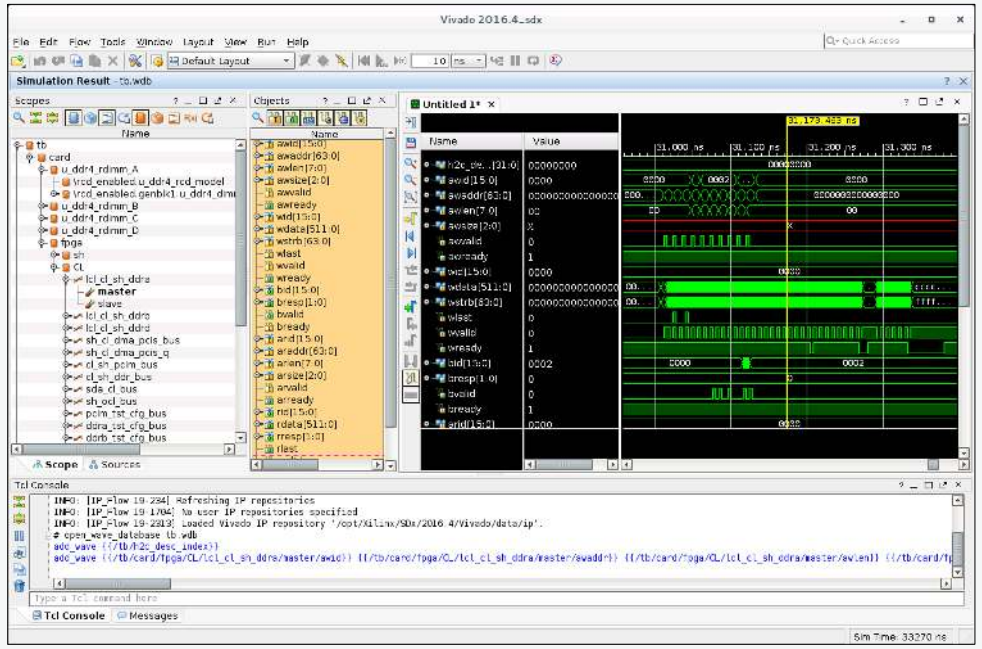


Developing applications for F1

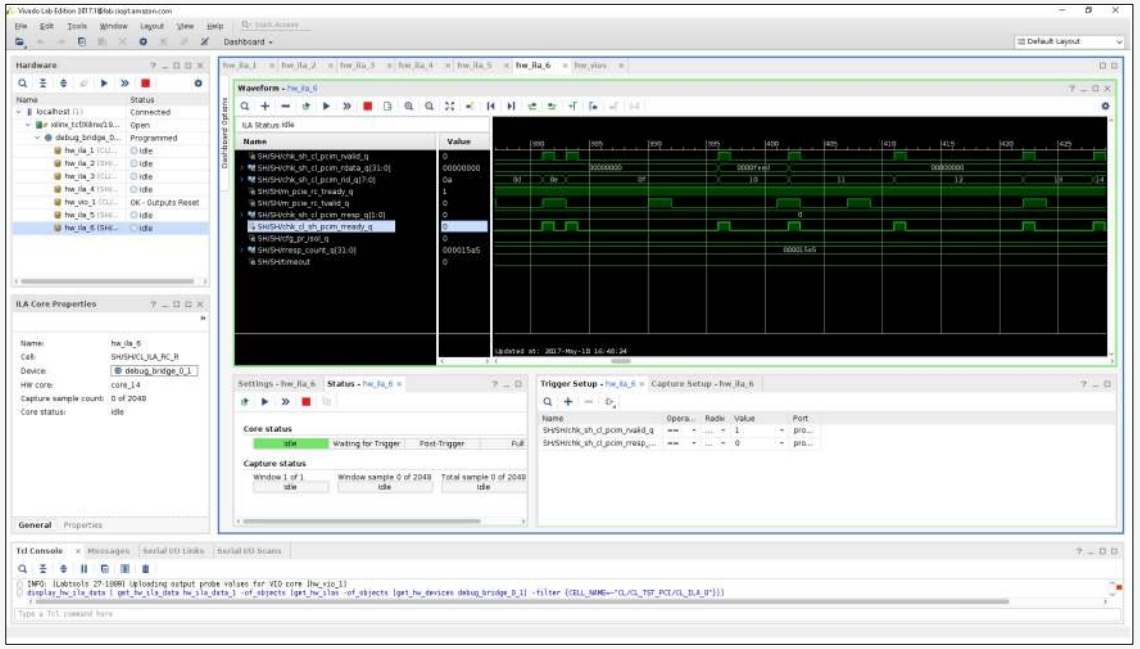
The FPGA Developer AMI

Use Xilinx Vivado and a hardware description language (Verilog or VHDL for RTL) with the HDK to describe and simulate your FPGA logic

Xilinx Vivado for custom logic development



Virtual JTAG for interactive debugging



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	1.3.3
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 SDx 2017.1 - 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: Asia Pacific (Mumbai)

Free Tier Eligible EC2 charges for Micro instances are free for up to **750 hours** a month if you qualify for the AWS Free Tier.

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 Asia Pacific (Mumbai).

1 Software Pricing

The data below shows pricing per instance for services hosted in Asia Pacific (Mumbai).

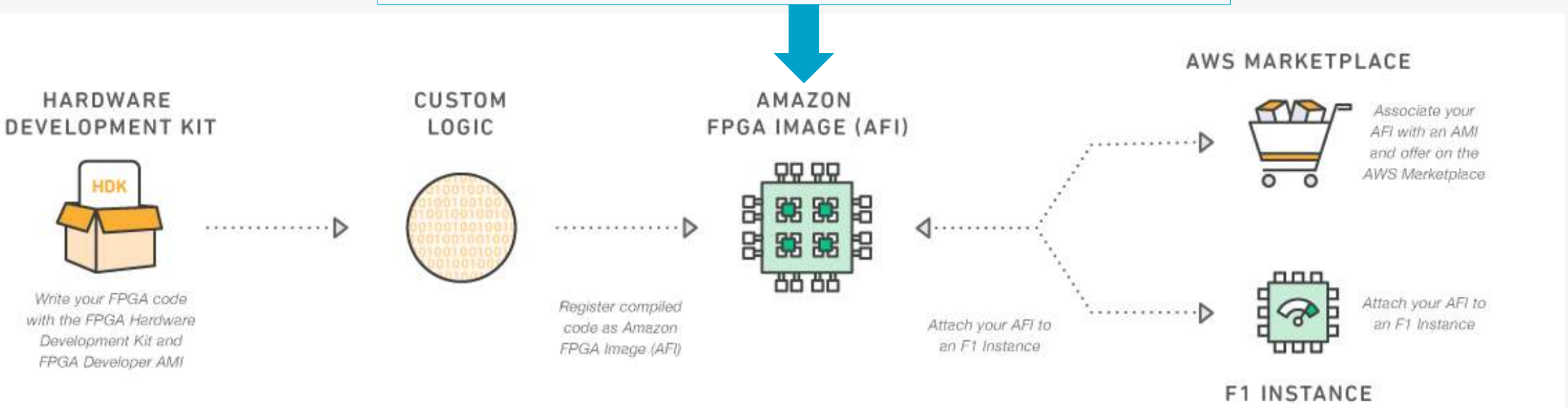
FPGA Developer AMI - Hourly

EC2 Instance Type	Software /hr	EC2 /hr	Total /hr
m4.10xlarge	$0.00	$2.46	$2.46
r4.16xlarge	$0.00	$4.864	$4.864
m4.xlarge	$0.00	$0.246	$0.246
m4.2xlarge	$0.00	$0.492	$0.492
t2.large	$0.00	$0.119	$0.119
c4.2xlarge	$0.00	$0.439	$0.439
t2.micro	$0.00	$0.015	$0.015
r4.4xlarge	$0.00	$1.216	$1.216
t2.2xlarge	$0.00	$0.476	$0.476
t2.xlarge	$0.00	$0.238	$0.238
m4.large	$0.00	$0.123	$0.123
c4.8xlarge	$0.00	$1.756	$1.756
m4.4xlarge	$0.00	$0.984	$0.984
t2.medium	$0.00	$0.059	$0.059
c4.large	$0.00	$0.11	$0.11
t2.small	$0.00	$0.03	$0.03
r4.xlarge	$0.00	$0.304	$0.304
c4.4xlarge	$0.00	$0.878	$0.878
c4.xlarge	$0.00	$0.22	$0.22
r4.8xlarge	$0.00	$2.432	$2.432
r4.2xlarge	$0.00	$0.608	$0.608
r4.large	$0.00	$0.152	$0.152
m4.16xlarge	$0.00	$3.936	$3.936

Create and Use the Amazon FPGA image (AFI)

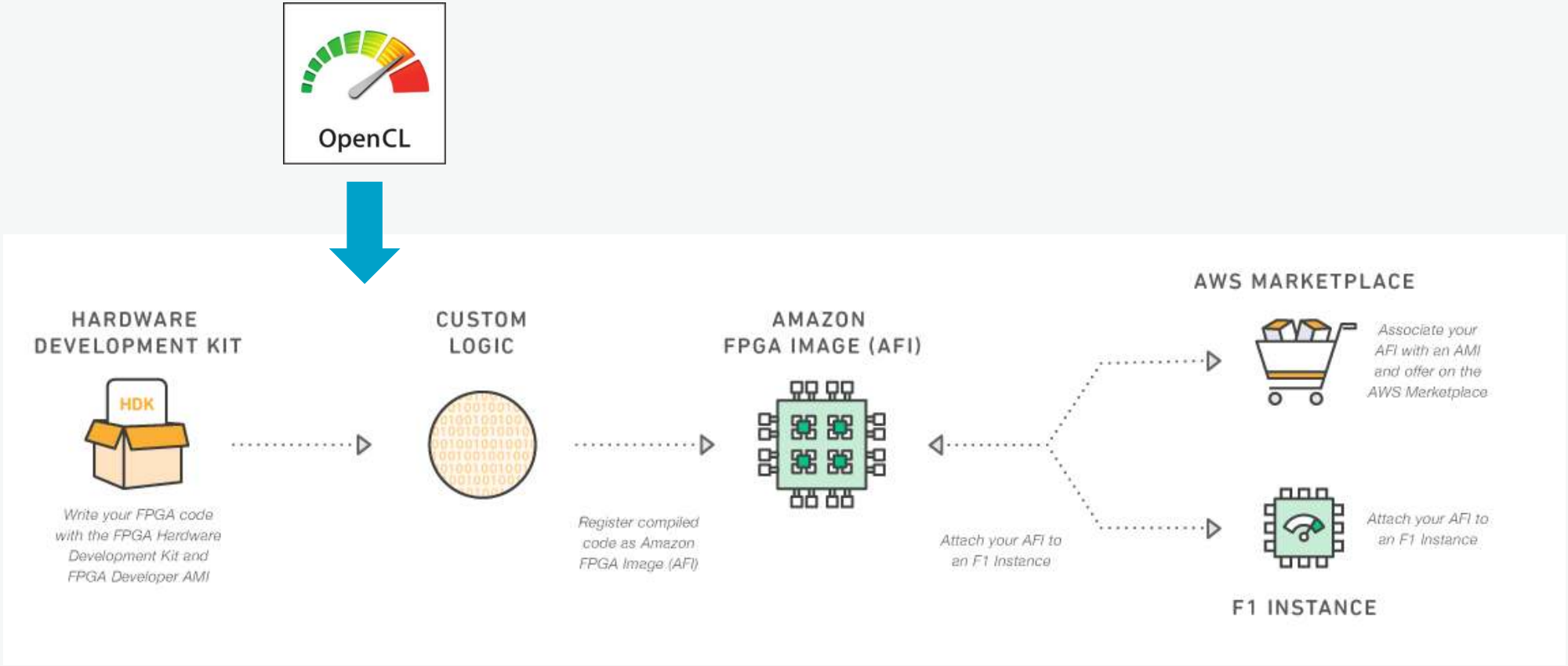
Generate an encrypted AFI and deploy on F1 / publish on Marketplace

```
$ 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 ]
```



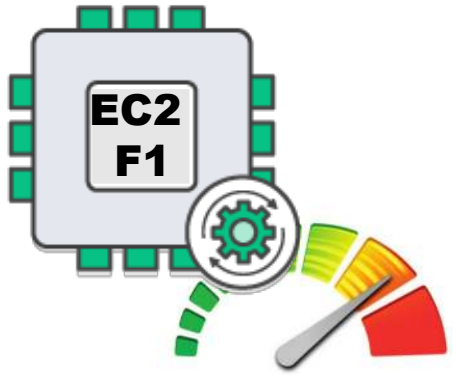
Create the Amazon FPGA image (AFI)

OpenCL provides a parallel, higher level design flow



OpenCL generally available for F1

- Familiar development experience to accelerate C/C++ applications
- 80+ F1 code examples available on github: security, image processing and accelerated algorithms
- Already supported on the FPGA Developer AMI, no need to upgrade/install



OpenCL

GETTING STARTED WITH F1 C/C++ PROGRAMMING

https://github.com/awslabs/aws-fpga-app-notes/tree/master/reInvent17_Developer_Workshop

- Gain hands-on experience with AWS F1 C/C++
- Learn how to develop FPGA-accelerated applications
- Learn the OpenCL flow with Xilinx SDAccel development environment

re:Invent 2017 Developer Workshop

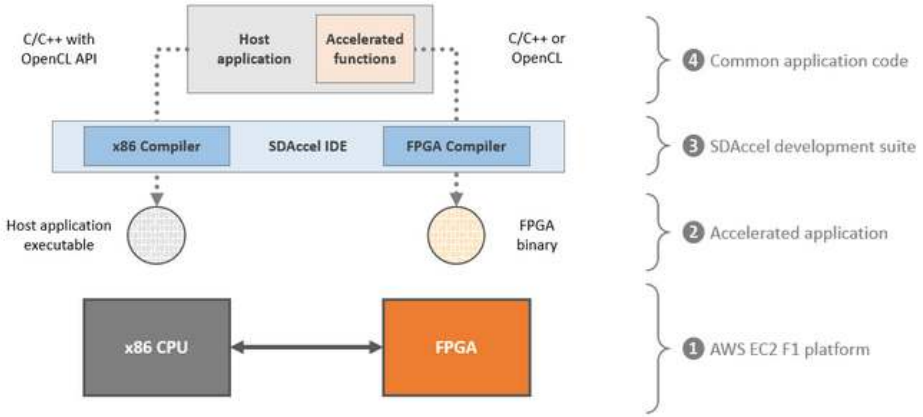
Introduction	1. Connecting to your F1 instance	2. Experiencing F1 acceleration	3. Developing F1 applications	4. Wrapping-up

Introduction

Welcome to the re:Invent 2017 Developer Workshop. During this session you will gain hands-on experience with AWS F1 and learn how to develop accelerated applications using the AWS F1 OpenCL flow and the Xilinx SDAccel development environment.

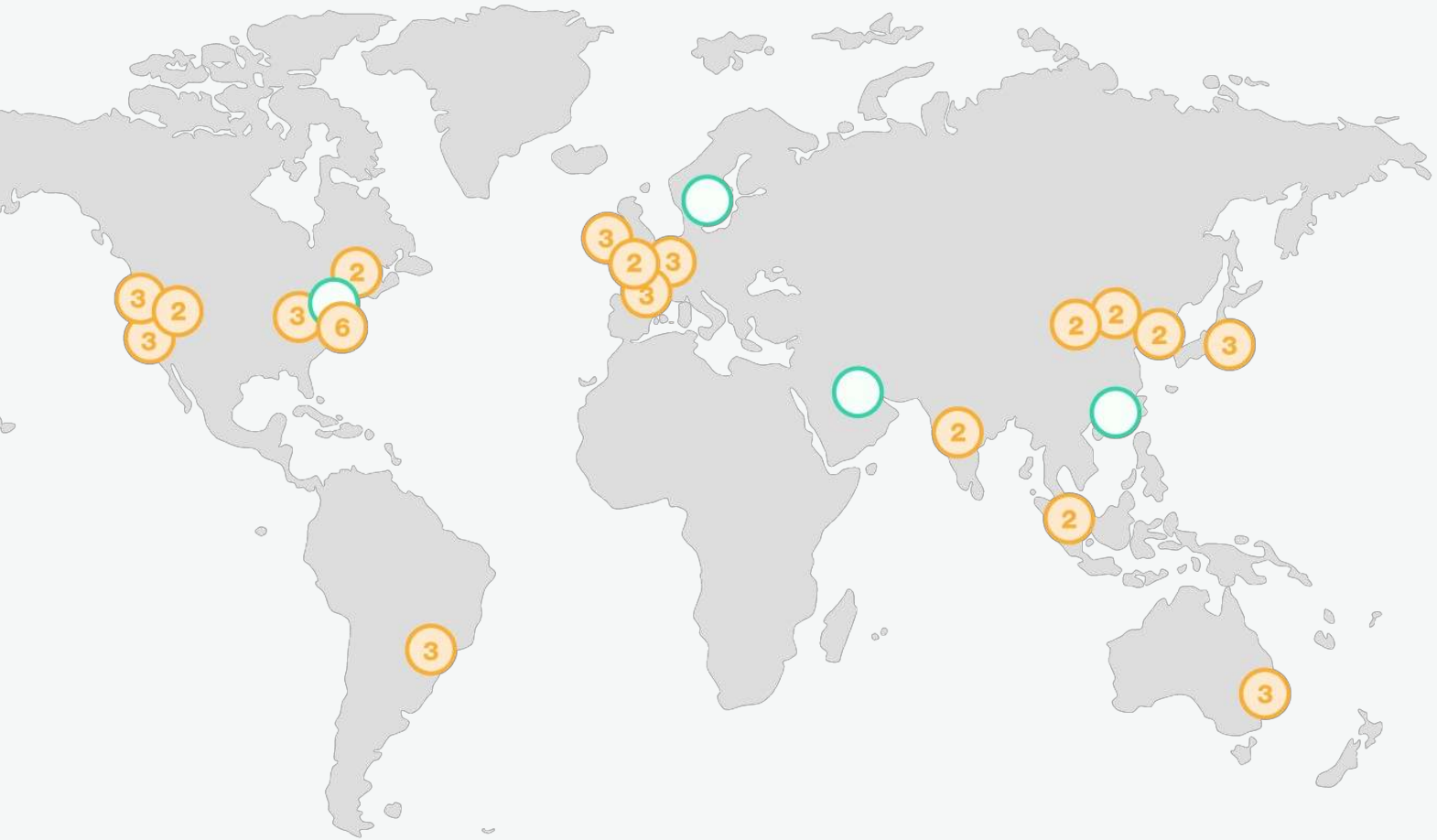
Overview of the AWS F1 platform and SDAccel flow

The architecture of the AWS F1 platform and the SDAccel development flow are pictured below:



Deploy faster wherever you like

18 Regions – **49** Availability Zones

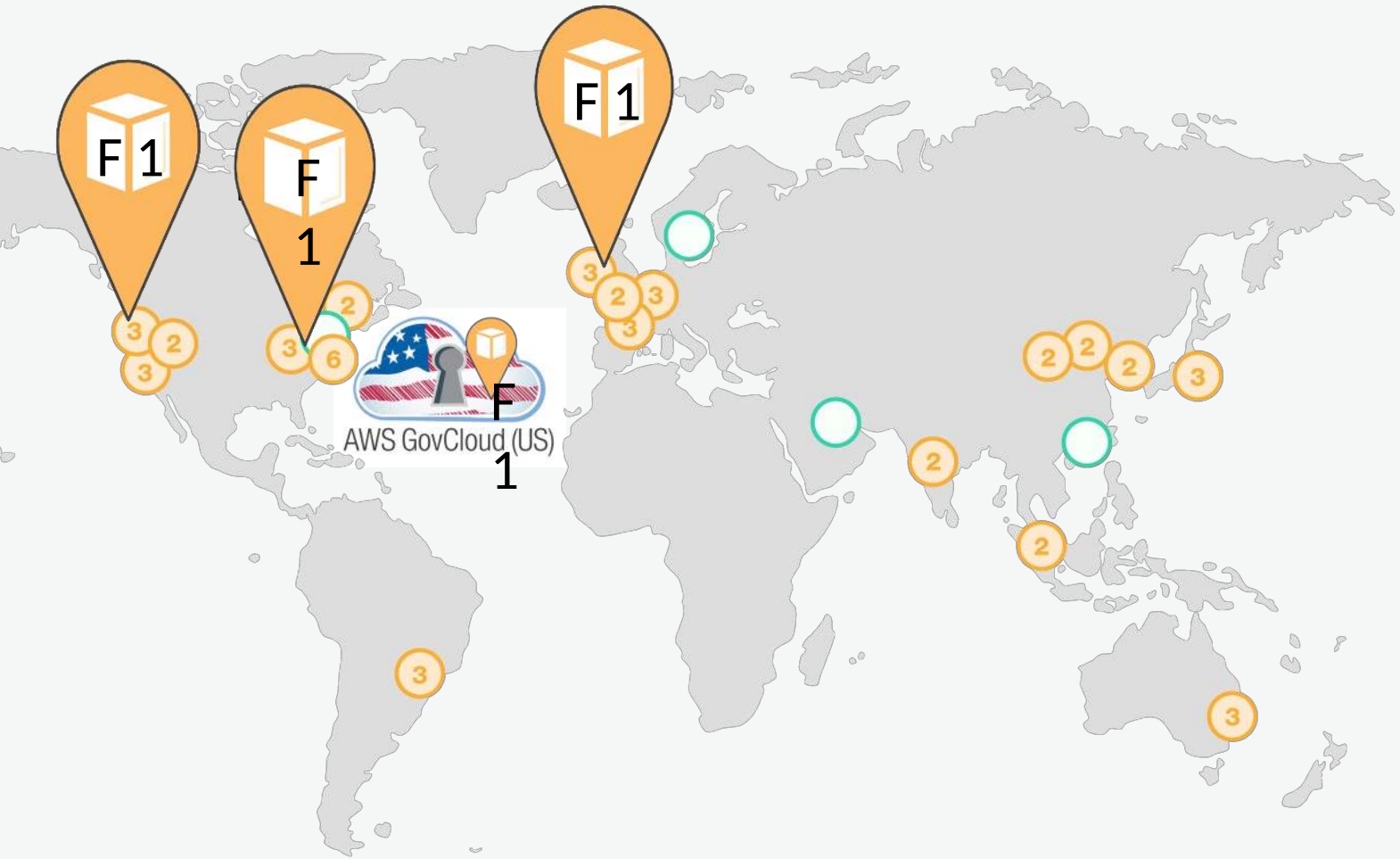


US East
N. Virginia (6), Ohio (3)
US West
N. California (3), Oregon (3)
Asia Pacific
Mumbai (2), Seoul (2), Singapore (2), Sydney (3), Tokyo (3)
Canada
Central (2)
China
Beijing (2), Ningxia (2)
Europe
Frankfurt (3), Ireland (3), London (2), Paris (3)
South America
São Paulo (3)
AWS GovCloud (US-West) (2)

aws

Deploy faster wherever you like

18 Regions – **49** Availability Zones



US East
N. Virginia (6), Ohio (3)
US West
N. California (3), Oregon (3)
Asia Pacific
Mumbai (2), Seoul (2), Singapore (2), Sydney (3), Tokyo (3)
Canada
Central (2)
China
Beijing (2), Ningxia (2)
Europe
Frankfurt (3), Ireland (3), London (2), Paris (3)
South America
São Paulo (3)
AWS GovCloud (US-West) (2)

aws

FPGA Partner and user eco-system

F1 use cases and partners

- Financial computing
- Genomics Sequencing
- Engineering simulations
- Image and video processing
- Big data and machine learning
- Security, Compression
- …and more

Connecting FPGA partners with AWS users

Accelerating Precision Medicine at Scale

Children's Hospital of Philadelphia and Edico Genome Achieve Fastest-Ever Analysis of 1,000 Genomes

Orlando, Fla., Oct 19, 2017 – The Children's Hospital of Philadelphia (CHOP) and Edico Genome today set a new scientific world standard in rapidly processing whole human genomes into data files usable for researchers aiming to bring precision medicine into mainstream clinical practice. Utilizing Edico Genome's DRAGEN™ Genome Pipeline, deployed on 1,000 Amazon EC2 F1 instances on the Amazon Web Services (AWS) Cloud, **1,000 pediatric genomes were processed in two hours and 25 minutes.**

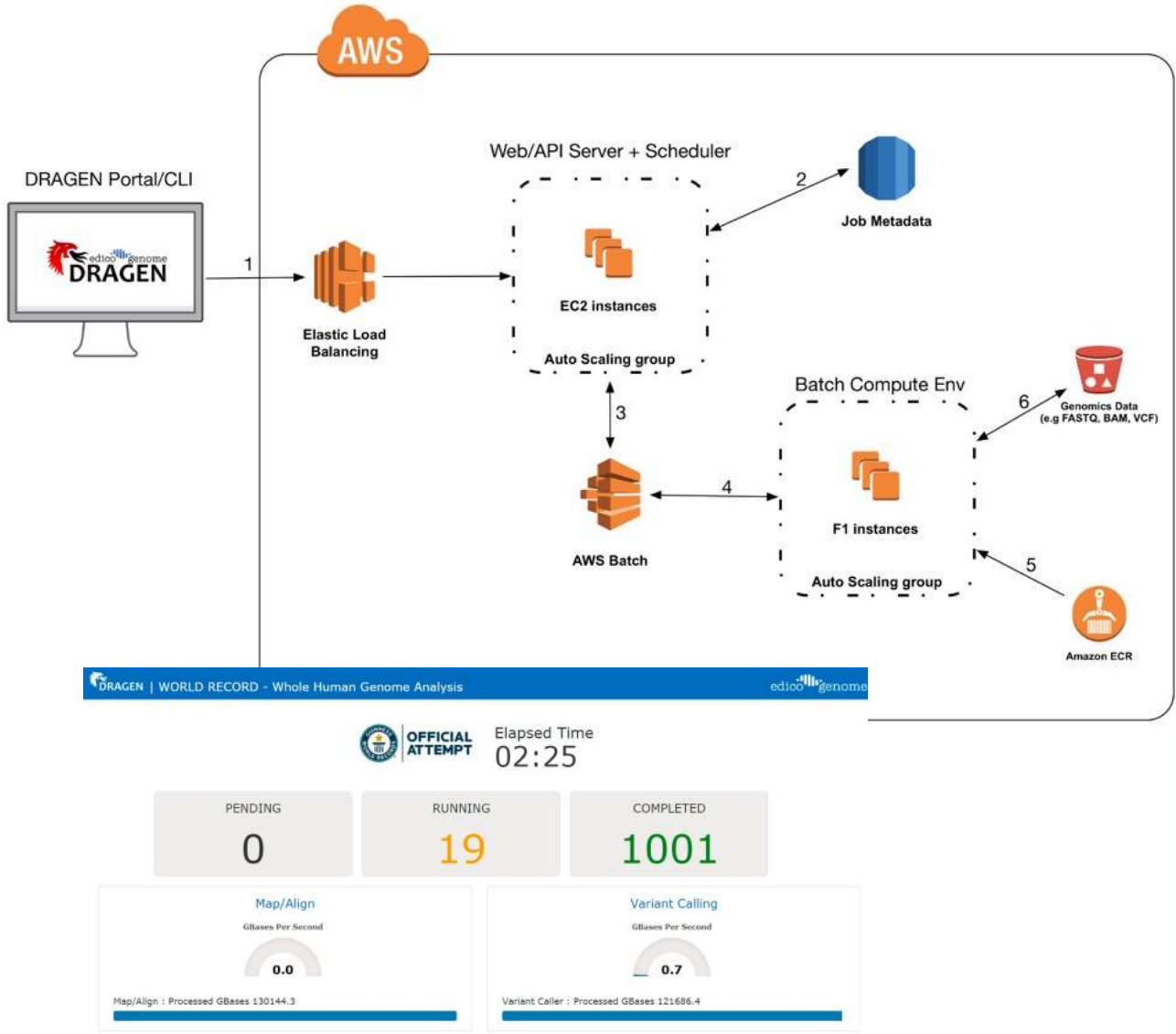
GUINNESS WORLD RECORDS™
RECORD HOLDER

edico genome

aws

Acceleration Solutions require more than an FPGA

- **AWS Batch** to schedule instances with auto scaling groups
- **Amazon Aurora** used to store jobs metadata
- **Amazon ECR**: docker container service to load needed binaries efficiently
- **Amazon S3** storing the genomic data
- **Elastic load balancer**
- **EC2 F1** instances (1020 of them)



Source: https://aws.amazon.com/blogs/compute/accelerating-precision-medicine-at-scale/

Connecting FPGA partners with AWS users

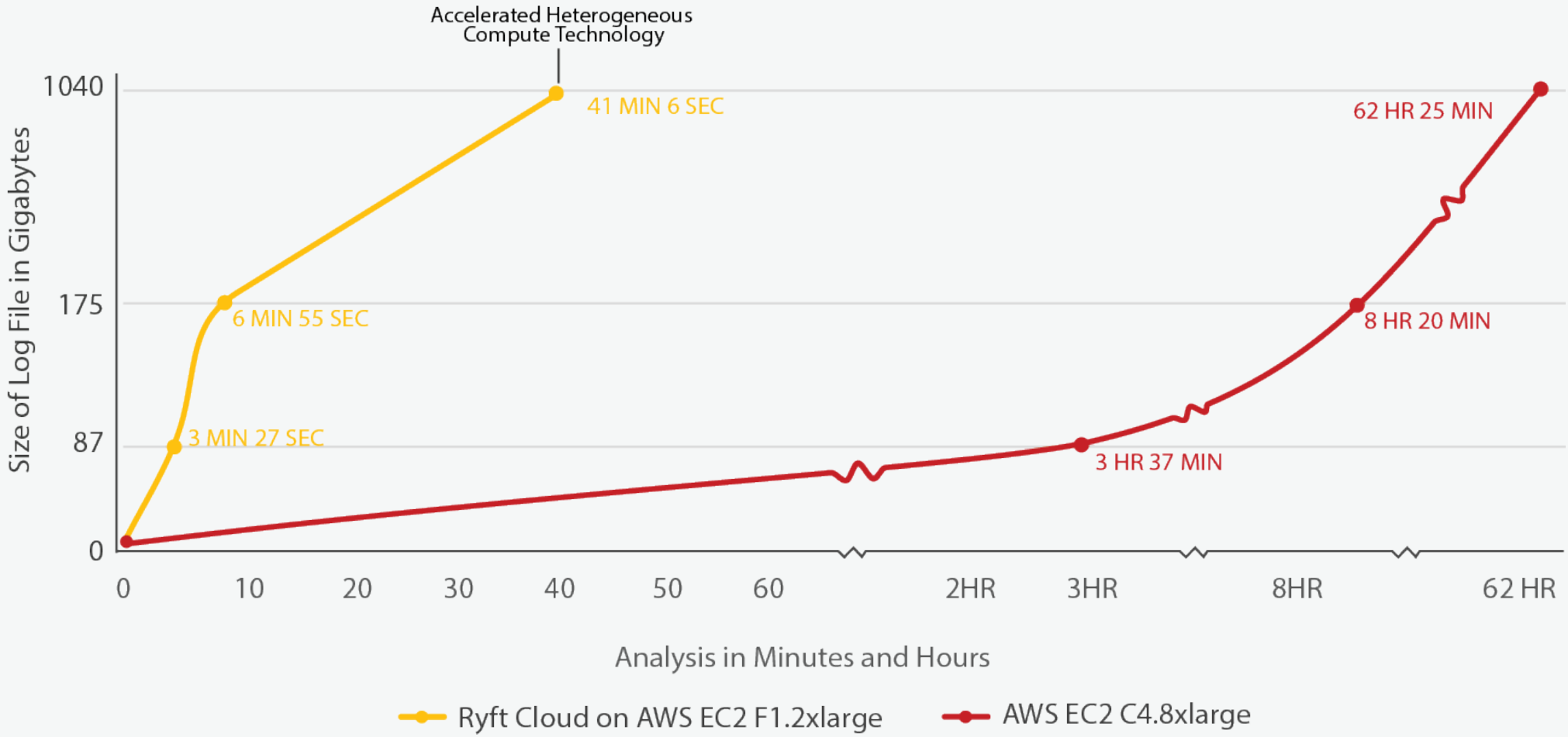
Connecting FPGA technology partners with end customers

- **World's first implementation** of cloud hardware accelerated low latency HEVC VR360 video streaming and Live Broadcast H.265/HEVC Cloud Video Encoding
- Amazon EC2 FPGA instances leveraged to **reduce encoding costs by 10X compared** to current software solutions
- At lower bitrates than legacy VR360 video streaming systems, delivers **significantly better quality with lower latency**

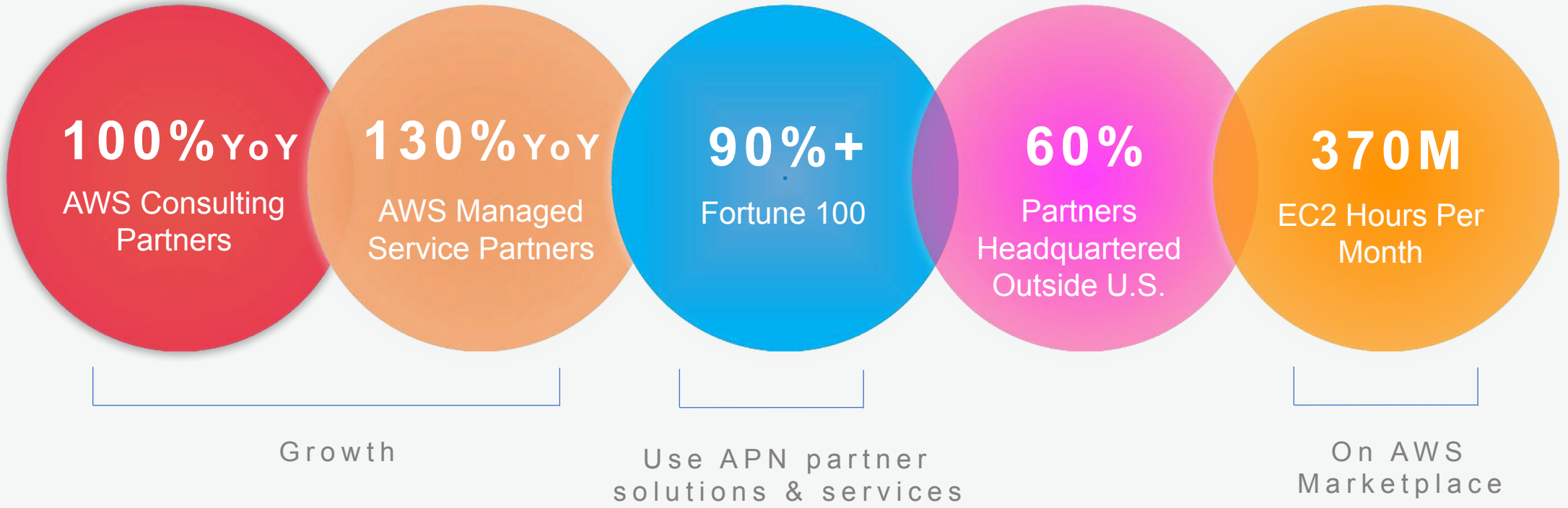
"NGCodec and Tiledmedia Show Cloud Hardware-Accelerated Low-Latency HEVC VR360 Streaming Solution"

Ryft Supercharges Analysis by 91X to Render Big Data

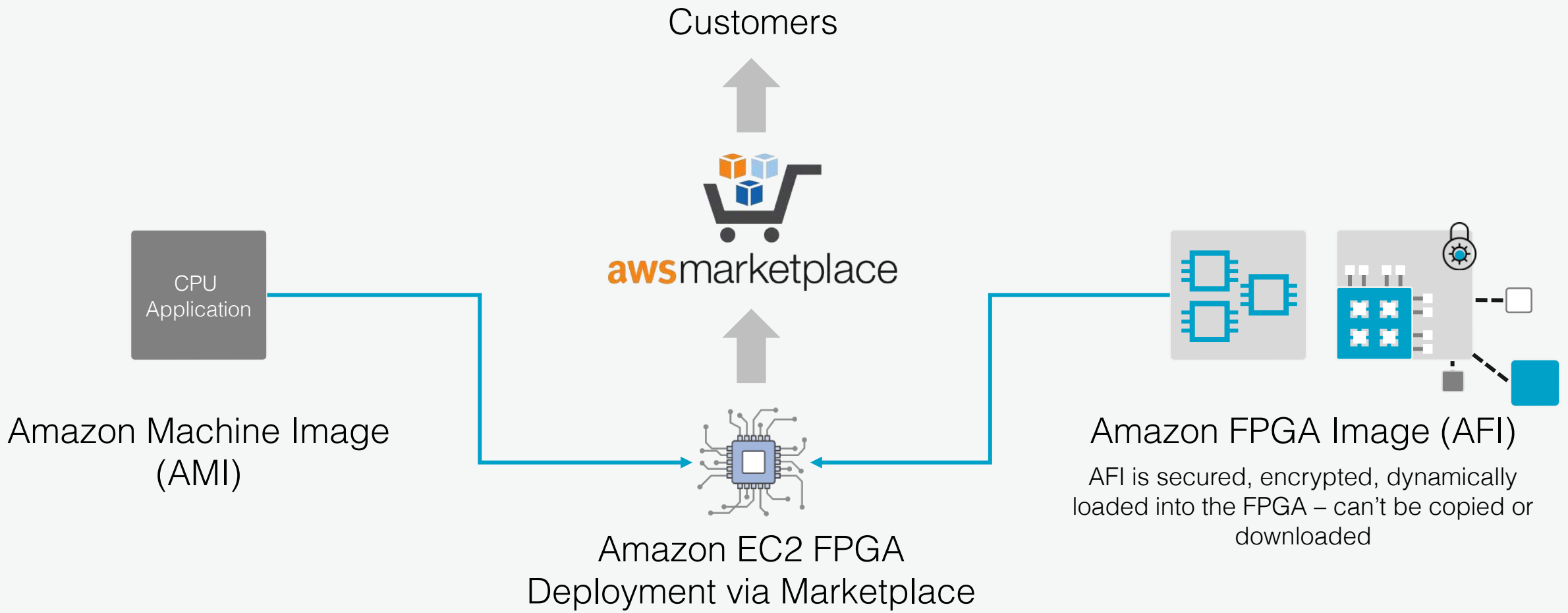
Benchmark comparison of Elasticsearch on Ryft Cloud with FPGA-acceleration vs. on CPU



AWS Partner Network

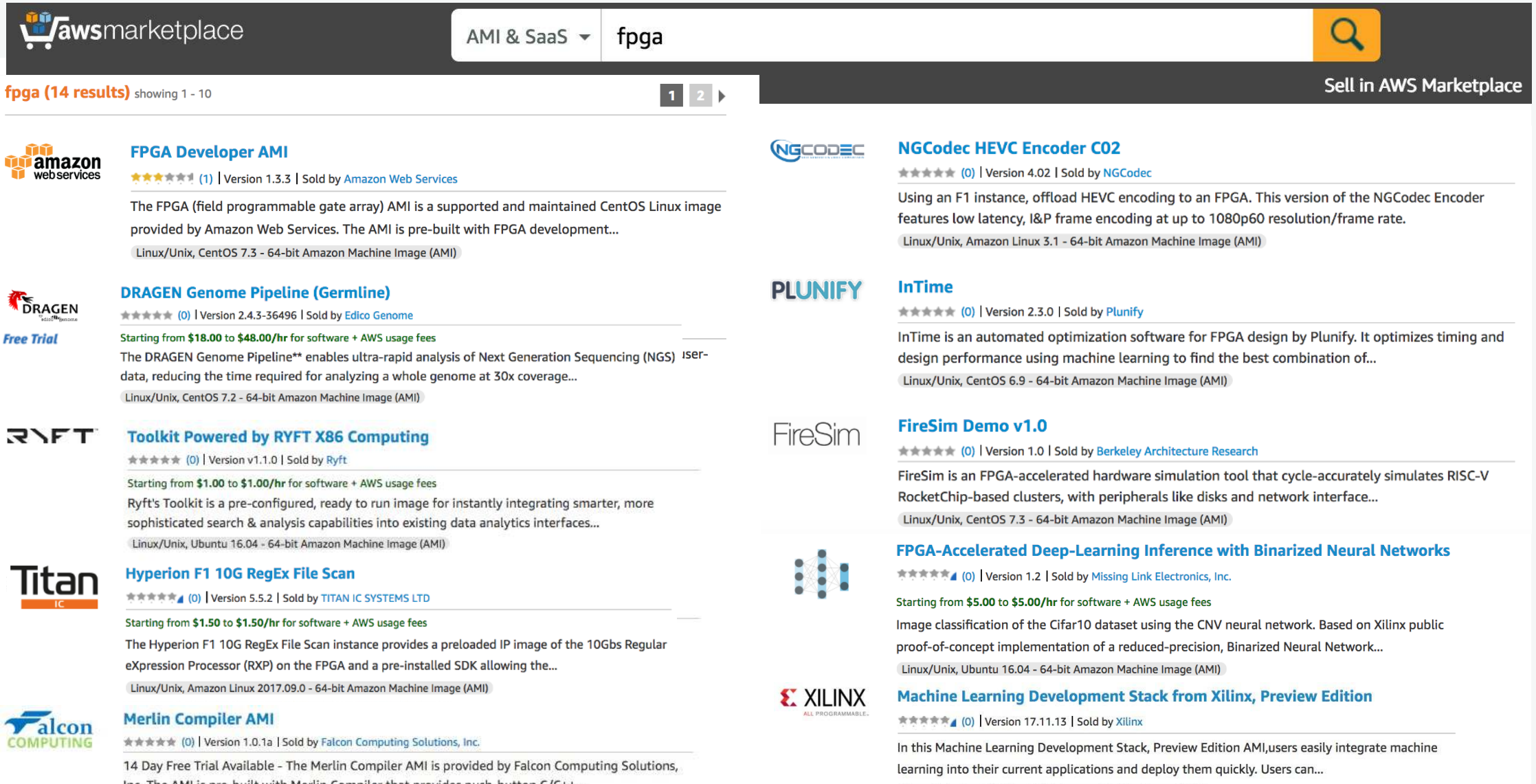


Delivering FPGA partner solutions



AWS Marketplace

Discover, Procure, Deploy, and Manage Software in the Cloud



Interested in becoming an AWS EC2 F1 Instance Partner?

AWS EC2 F12 Instance participates in the AWS Service Delivery Program. The Service Delivery Program recognizes APN Partners with a verified track record of delivering specific AWS services and workloads to AWS customers, including AWS EC2 F1 Instance.

To apply to become a AWS Service Delivery Partner, apply online through the APN Portal today.

https://aws.amazon.com/partners/

Thank you!

gadi@amazon.com


@gadi_hutt