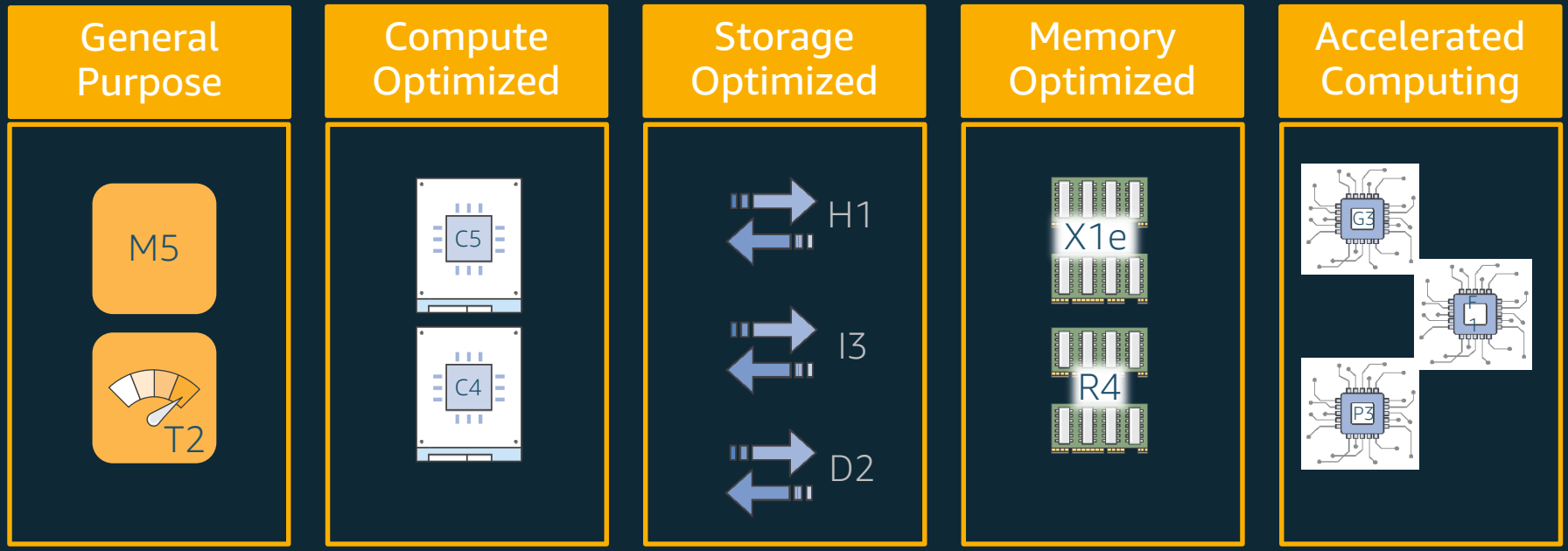
Deep Dive on Amazon EC2 Accelerated Computing Instances

Chetan Kapoor, Senior Product Manager – AWS EC2

May 2nd, 2018

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Amazon EC2 Instance Types



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

EC2 Accelerated Computing Instances

P3 GPU Compute Instance

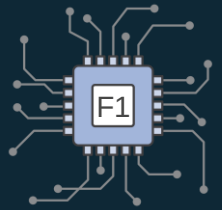
- Up to 8 NVIDIA V100 GPUs in a single instance, with NVLink for peer-to-peer GPU communication
- Supporting a wide variety of use cases including deep learning, HPC, financial computing, and batch rendering

G3: GPU Graphics Instance

- Up to 4 NVIDIA M60 GPUs, with GRID Virtual Workstation features and licenses
- Designed for workloads such as 3D rendering, 3D visualizations, graphics-intensive remote workstations, video encoding, and virtual reality applications

F1: FPGA instance

- Up to 8 Xilinx Virtex® UltraScale+™ VU9P FPGAs in a single instance. Programmable via VHDL, Verilog, or OpenCL. Growing marketplace of pre-built application accelerations.
- Designed for hardware-accelerated applications including financial computing, genomics, accelerated search, and image processing



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

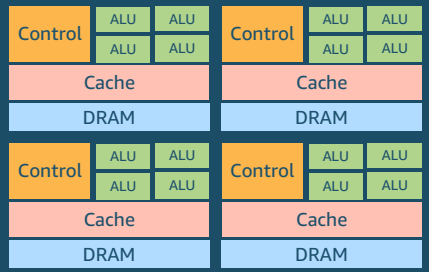
aws

AWS EC2 P3 Instances for Compute Acceleration

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

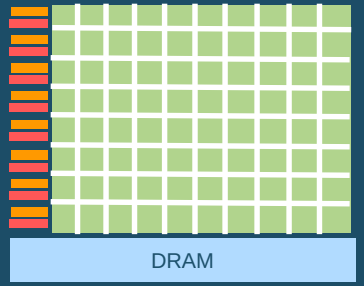
CPUs vs GPUs vs FPGA for Compute

CPU



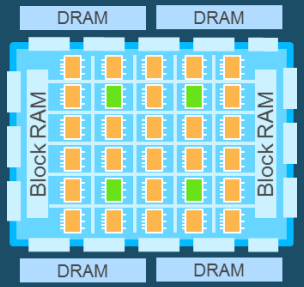
- 10s-100s of processing cores
- Pre-defined instruction set & datapath widths
- Optimized for general-purpose computing

GPU



- 1,000s of processing cores
- Pre-defined instruction set and datapath widths
- Highly effective at parallel execution

FPGA



- Millions of programmable digital logic cells
- No predefined instruction set or datapath widths
- Hardware timed execution

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

Amazon EC2 P3 Instances (October 2017)

One of the fastest, most powerful GPU instances in the cloud

- Up to eight NVIDIA Tesla V100 GPUs
- 1 PetaFLOPs of computational performance
  – Up to 14x better than P2
- 300 GB/s GPU-to-GPU communication (NVLink) *– 9X better than P2*
- 16GB GPU memory with 900 GB/sec peak GPU memory bandwidth

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

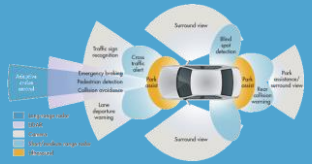
Use-Cases for P3 Instances

Machine Learning/AI

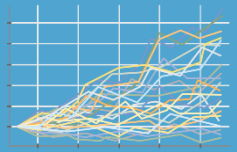
Natural Language Processing

Image and Video recognition

Autonomous vehicle systems

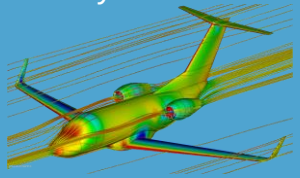


Recommendation Systems



High Performance Computing

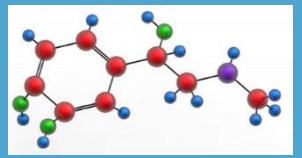
Computational Fluid Dynamics



Financial and Data Analytics

Weather Simulation

Computational Chemistry



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

P3 Instances Details

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*	1-yr RI Effective Hourly*	3-yr RI Effective Hourly*
P3.2xlarge	1	No	8	61	Up to 10Gbps	1.7Gbps	$3.06	$1.99 (35% Disc.)	$1.23 (60% Disc.)
P3.8xlarge	4	NVLink	32	244	10Gbps	7Gbps	$12.24	$7.96 (35% Disc.)	$4.93 (60% Disc.)
P3.16xlarge	8	NVLink	64	488	25Gbps	14Gbps	$24.48	$15.91 (35% Disc.)	$9.87 (60% Disc.)

Regional Availability

P3 instances are generally available in AWS US East (Northern Virginia), US East (Ohio), US West (Oregon), EU (Ireland), Asia Pacific (Seoul), Asia Pacific (Tokyo), AWS GovCloud (US) and China (Beijing) Regions.

Framework Support

P3 instances and their V100 GPUs supported across all major frameworks (such as TensorFlow, MXNet, PyTorch, Caffe2 and CNTK)

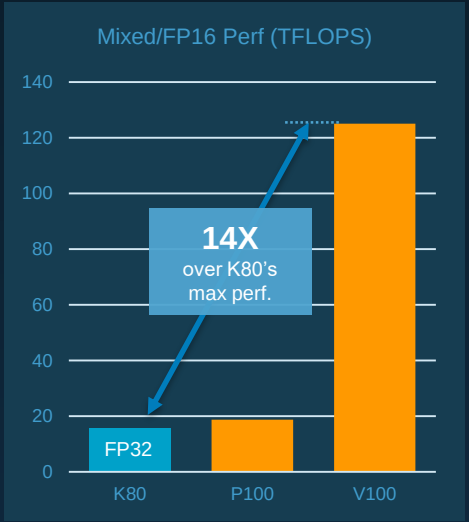
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

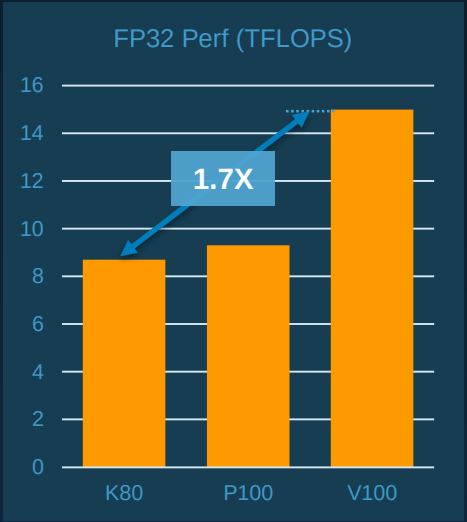
aws

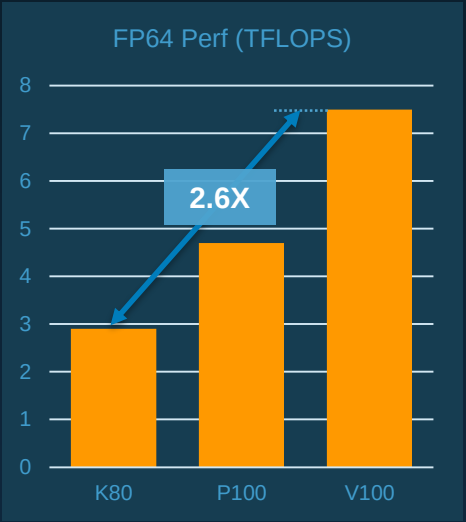
AWS P3 vs P2 Instance

GPU Performance Comparison

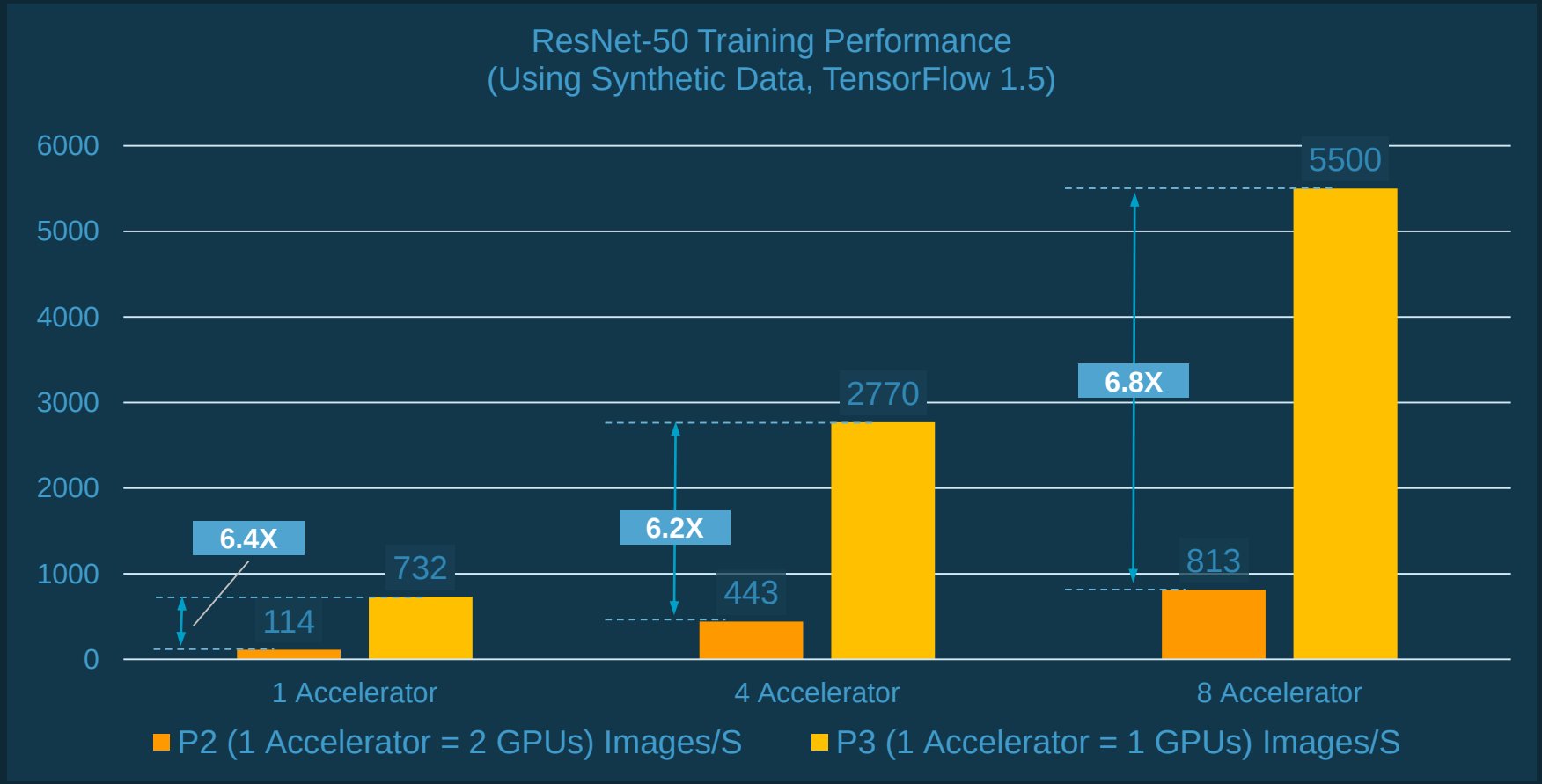
- P2 Instances use K80 Accelerator (Kepler Architecture)
- P3 Instances use V100 Accelerator (Volta Architecture)







© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

P3 Instances Details

Instance Size	GPUs	GPU Peer to Peer	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth	On-Demand Price/hr*	1-yr RI Effective Hourly*	3-yr RI Effective Hourly*
P3.2xlarge	1	No	8	61	Up to 10Gbps	1.7Gbps	$3.06	$1.99 (35% Disc.)	$1.23 (60% Disc.)
P3.8xlarge	4	NVLink	32	244	10Gbps	7Gbps	$12.24	$7.96 (35% Disc.)	$4.93 (60% Disc.)
P3.16xlarge	8	NVLink	64	488	25Gbps	14Gbps	$24.48	$15.91 (35% Disc.)	$9.87 (60% Disc.)

- P3 instances provide GPU-to-GPU data transfer over **NVLink**
- P2 instanced provided GPU-to-GPU data transfer over **PCI Express**

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

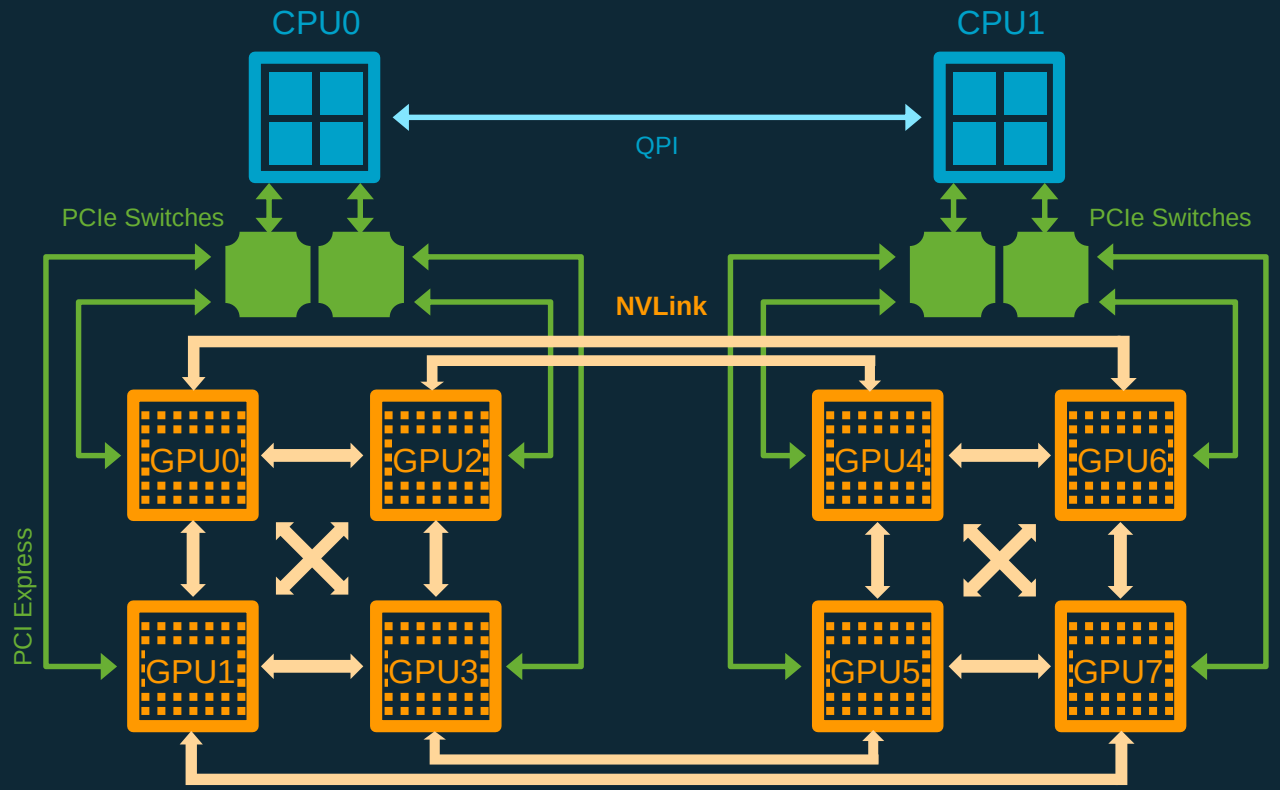
P3 vs P2 Peer-to-Peer Configurations

Description	P3.16xlarge	P2.16xlarge	P3 GPU Performance Improvement
Number of GPUs	8	16	-
Number of Accelerators	8 (V100)	8 (K80)	
GPU – Peer to Peer	NVLink – 300 GB/s	PCI-Express - 32 GB/s	9.4X
CPU to GPU Throughput PCIe throughput per GPUs	8 GB/s	1 GB/s	8X
CPU to GPU Throughput Total instance PCIe throughput	64 GB/s (Four x16 Gen3)	16 GB/s (One x16 Gen3)	4X

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

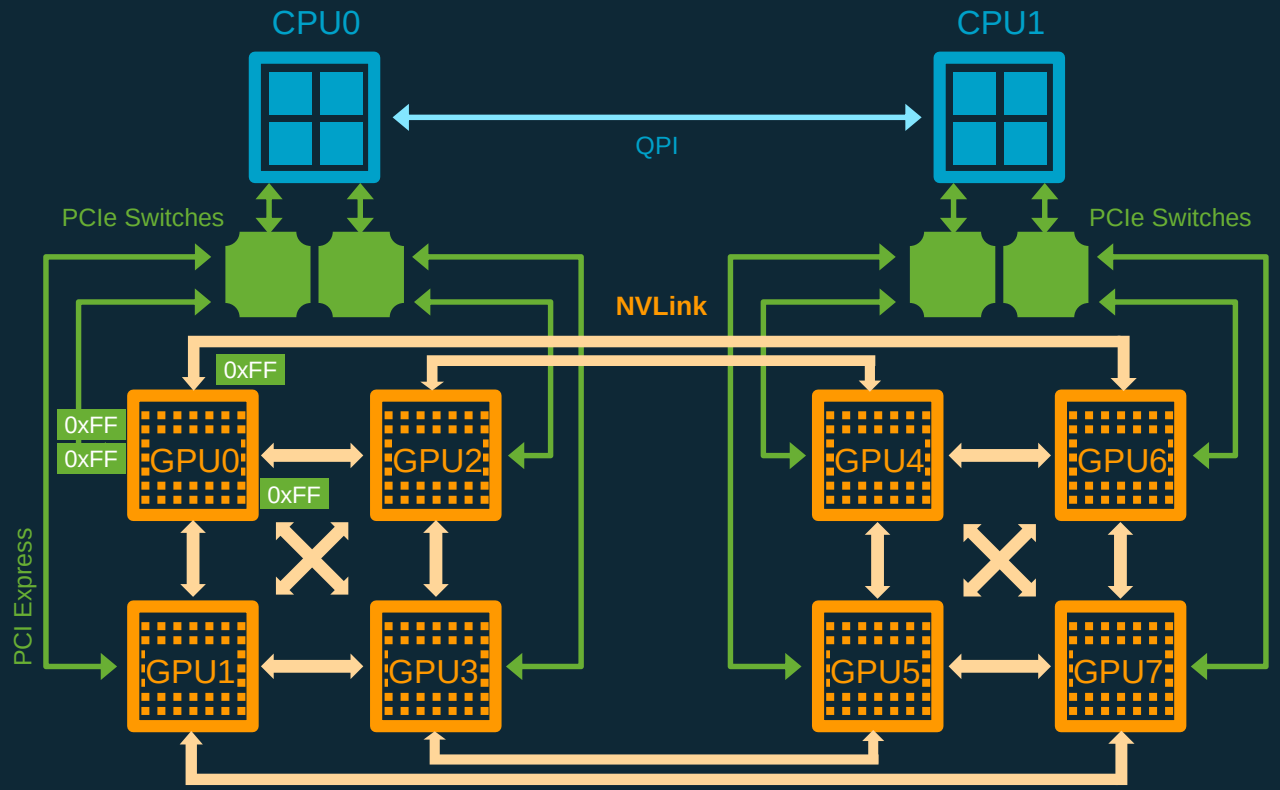
aws

P3 PCIe and NVLink Configurations



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

P3 PCIe and NVLink Configurations



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS Storage Options

EFS	EC2+EBS	Amazon S3	Amazon Glacier
Highly available, multi-AZ, fully managed network-attached elastic file system.	Create a single-AZ shared file system using EC2 and EBS, with third-party or open source software (e.g., ZFS, Intel Lustre, etc).	Secure, durable, highly-scalable object storage. Fast access, low cost.	Secure, durable, long term, highly cost-effective object storage.
For near-line, highly-available storage of files in a traditional NFS format (NFSv4).	For near-line storage of files optimized for high I/O performance.	For long-term durable storage of data, in a readily accessible get/put access format.	For long-term storage and archival of data that is infrequently accessed.
Use for read-often, temporary working storage	**Use for high-IOPs, temporary working storage**	**Primary durable and scalable storage for data**	**Use for long-term, lower-cost archival of data**

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

Data Ingestion Options

- Within a P3 instance, we have high data throughput in to GPUs (PCI Express to/from host CPUs) and between GPUs (NVLink)
- To maintain high utilization of GPUs, need high throughput data stream coming in to P3 instances
- **Option 1:** Use Multiple EBS Volumes
  - Each Provisioned IOPS SSD (io1) EBS volume and provide about **500 MB/s** of read or write throughput (need to be provisioned with 20,000 IOPS)
  - Customer can use independent EBS volume or combine multiple volumes via RAID to create a single logical volume (5 io1 volumes can support **1.65 GB/s)**
  - http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/raid-config.html
- **Option 2**: Amazon S3 -> EC2
  - We have increased data transfer from Amazon S3 directly in to EC2 from 5 Gbps to 25Gbps
  - Need to parallelize connections to Amazon S3 by using the TransferManager available in Amazon S3's Java SDK
  - https://docs.aws.amazon.com/sdk-for-java/v1/developer-guide/examples-s3-transfermanager.html

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Software Support for P3

Required Drivers & Libraries

- Hardware Driver version 384.81 or newer
- CUDA 9 or newer
- CuDNN 7 or newer & NCCL 2.0 or newer
    - Generally packaged with CUDA

Machine Learning Frameworks

- For customers to take advantage of the new Tensor Cores in V100 GPUs, they will need to use latest distros of ML framework
- All major frameworks have formally released support for V100 GPUs (ex - TensorFlow, MXNet, Pytorch, Caffe)
- http://docs.nvidia.com/deeplearning/sdk/pdf/Training-Mixed-Precision-User-Guide.pdf

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

AWS Deep Learning AMI

- Get started quickly with easy-to-launch tutorials
- Hassle-free setup and configuration
- Pay only for what you use – no additional charge for the AMI
- Accelerate your model training and deployment
- Support for popular deep learning frameworks

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

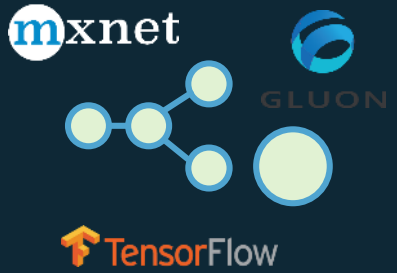
aws

Amazon SageMaker

Build, train, and deploy machine learning models at scale

End-to-End Machine Learning Platform

Zero setup



Flexible Model Training

Pay by the second

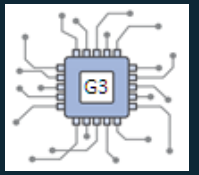
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

AWS EC2 G3 Instances for Graphics Acceleration

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS G3 GPU instances



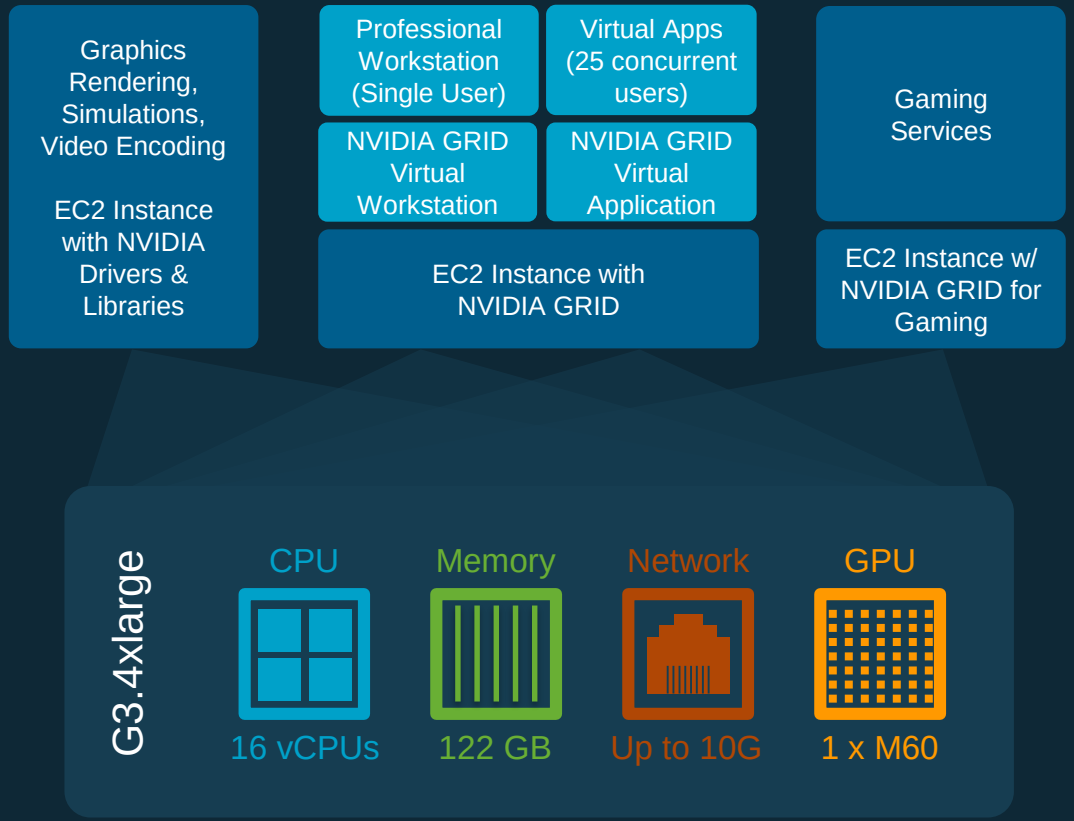
- Up to four NVIDIA M60 GPUs
- Includes GRID Virtual Workstation features and licenses, supports up to four monitors with 4096x2160 (4K) resolution
- Includes NVIDIA GRID Virtual Application capabilities for application virtualization software like Citrix XenApp Essentials and VMWare Horizon, supporting up to 25 concurrent users per GPU
- Hardware encoding to support up to 10 H.265 (HEVC) 1080p30 streams, and up to 18 H.264 1080p30 streams per GPU
- Designed for workloads such as 3D rendering, 3D visualizations, graphics-intensive remote workstations, video encoding, and virtual reality applications

Instance Size	GPUs	vCPUs	Memory (GiB)	Linux price per hour (N. Virginia)	Windows price per hour (N. Virginia)
g3.4xlarge	1	16	122	$1.14	$1.88
g3.8xlarge	2	32	244	$2.28	$3.75
g3.16xlarge	4	64	488	$4.56	$7.50

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

4 Modes of using G3 instances



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

G3 GRID Workstation vs. Virtual Application Modes

Feature	Workstation	Virtual Applications
	For professional 3D graphics applications at full performance	For PC-level applications, server-hosted RDSH desktops, XenApp
Concurrent users per GPU	1	25
NVIDIA Quadro feature	Yes	No
Desktop virtualization	Yes	No
Display & Resolution	4 monitors with 4096 x 2160 resolution	N/A
CUDA, OpenGL, DirectX and OpenCL	Yes	Yes

How to switch between the modes:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/activate_grid.html

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

AWS AMIs for G3 Instances

Available in AWS Marketplace

Microsoft Windows Server 2016 with NVIDIA GRID Driver

- No Additional Charge

Microsoft Windows Server 2012 R2 with NVIDIA GRID Driver

- No Additional Charge

Microsoft Windows Server 2016 with NVIDIA GRID Driver (Gaming Services)

- $0.023/hr to $0.092/hr additional software charge

Microsoft Windows Server 2012 with NVIDIA GRID Driver (Gaming Services)

- $0.023/hr to $0.092/hr additional software charge

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

G3 Use Cases

- Seismic Analysis, Energy E&P, Cloud GPU rendering & visualization, such as high end car configurators, AR/VR
- Desktop and Application Virtualization
  - Productivity and consumer apps
  - Design and engineering
  - Media and entertainment post-production
- Media and entertainment: video playout/broadcast, encoding/transcoding
- Cloud Gaming

HALLIBURTON
Landmark

E&P - Analytics

ZL ZeroLight

Auto – Car Configurators

BEBOP TECHNOLOGY

M&E – Content Creation

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

AWS G3 GPU instances – Data Analysis Platform

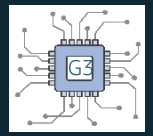
- Retrieving data from the S3 storage service to quickly hydrate for large data analysis (25 Gb/s) workloads
- Jupyter plugins with Bokeh, using WebGL backends allows for large scale data visualizations

NYC Taxi 1.3Billion Datapoints

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

AWS G3 GPU instances – FX Rendering



\+ NVIDIA GRID vDWS Workstation License +

- NVIDIA Grid Quadro features with NICE DCV remote visualization technology allows scalable delivery for advanced VFX rendering
- No additional cost on AWS using the G3 instance type
- Support for Direct X, OpenGL APIs
- Scale your desktops automatically based on number of rendering users

NVIDIA Faceworks – "Digital Ira" Demo

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

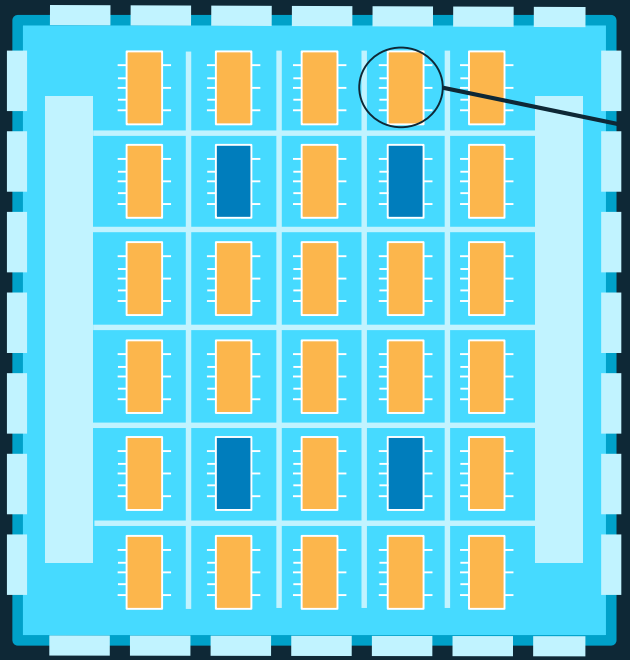
AWS EC2 F1 Instances for Custom Hardware Acceleration

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Parallel Processing in FPGAs

An FPGA is effective at processing data of many types in parallel, for example creating a complex pipeline of parallel, multistage operations on a video stream, or performing massive numbers of dependent or independent calculations for a complex financial model…

- An FPGA does not have an instruction-set!
- Data can be any bit-width (9-bit integer? No problem!)
- Complex control logic (such as a state machine) is easy to implement in an FPGA



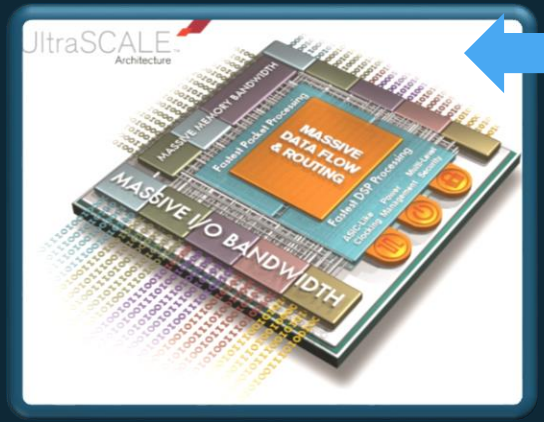
Each FPGA in F1 has more than 2M of these cells

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

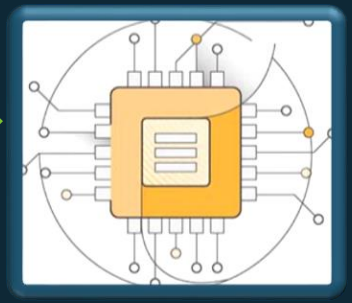
How FPGA Acceleration Works

FPGA handles compute-intensive, deeply pipelined, hardware-accelerated 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

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

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	FPGA Memory (GB)	vCPUs	Instance Memory (GB)	NVMe Instance Storage (GB)	Network Bandwidth
f1.2xlarge	1	64	8	122	1 x 470	Up to 10 Gbps
f1.16xlarge	8	512	64	976	4 x 940	25 Gbps

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

3 methods to use F1 Instance

1 Hardware Engineers/Developers

- Developers who are comfortable programming FPGA
- Use F1 Hardware Development Kit (HDK) to develop and deploying custom FPGA accelerations using Verilog and VHDL

2 Software Engineers/Developers

- Developers who are not proficient in FPGA design
- Use OpenCL to create custom accelerations

3 Software Engineers/Developers

- Developers who are not proficient in FPGA design
- Use pre-build and ready to use accelerations available in AWS Marketplace

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

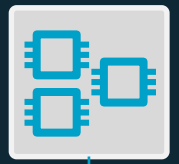
FPGA acceleration Development

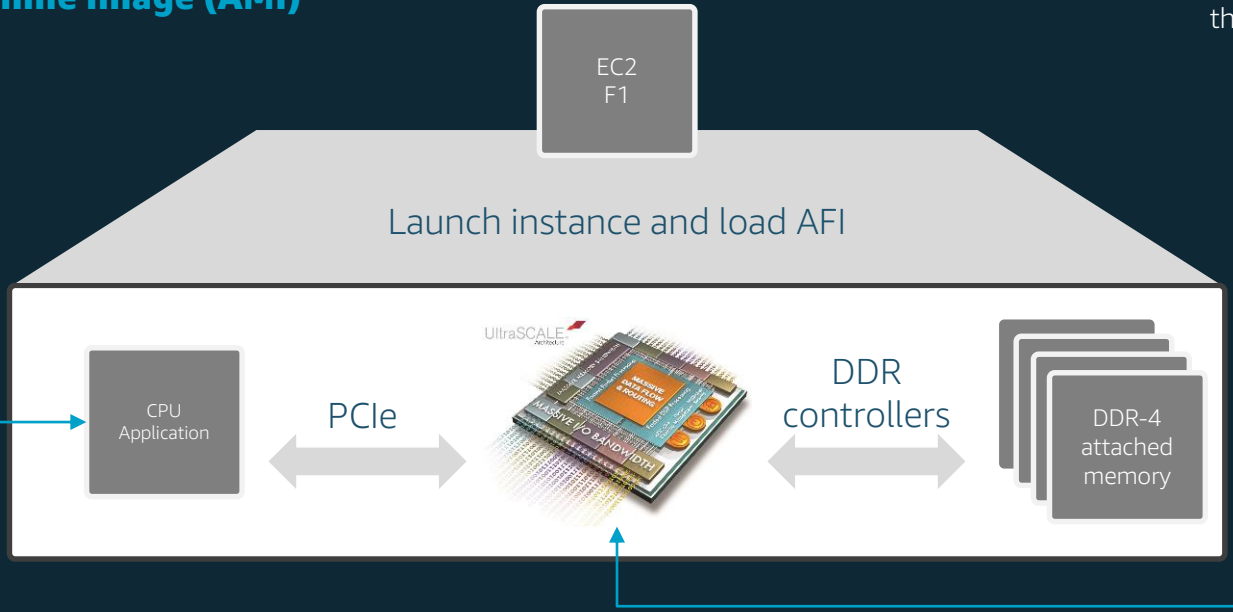
Amazon FPGA Image (AFI)

An F1 instance can have any number of AFIs

An AFI can be loaded into the FPGA in seconds

Amazon Machine Image (AMI)





© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

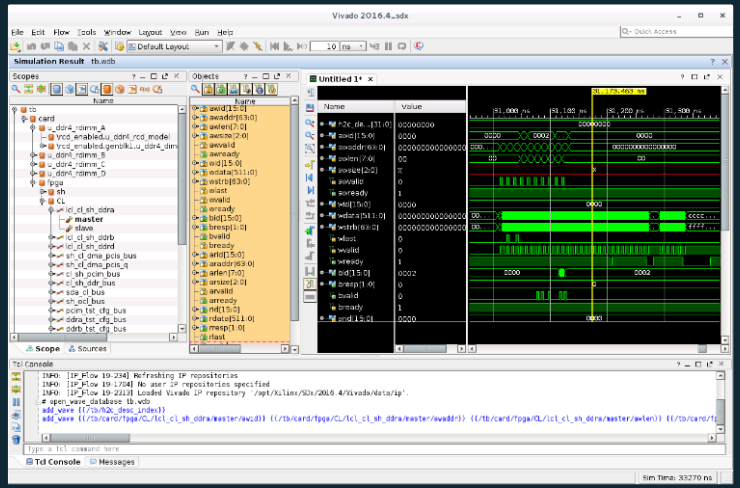
aws

Developing Custom Accelerations

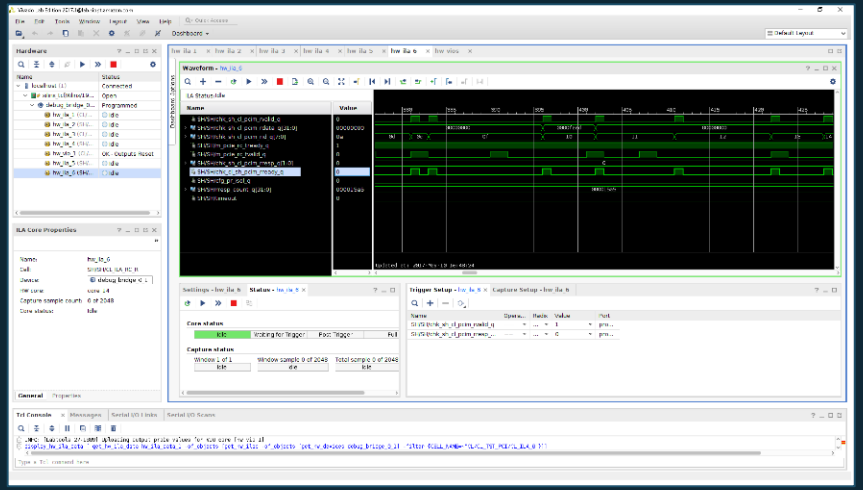
The FPGA Developer AMI

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



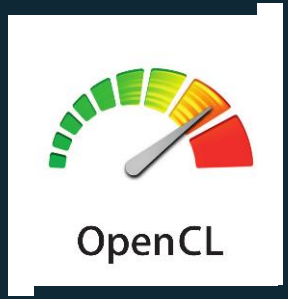
Virtual JTAG for interactive debugging



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

OpenCL generally available for F1

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

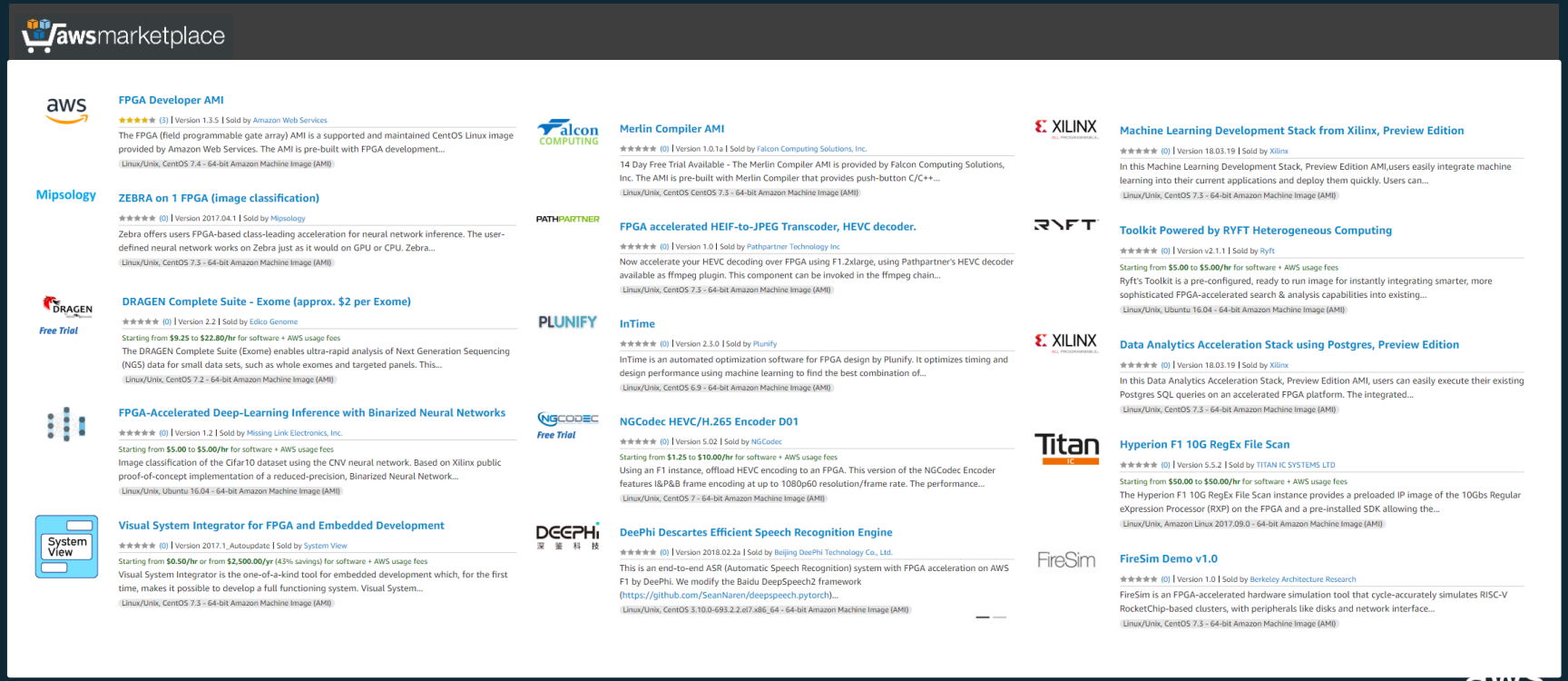


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

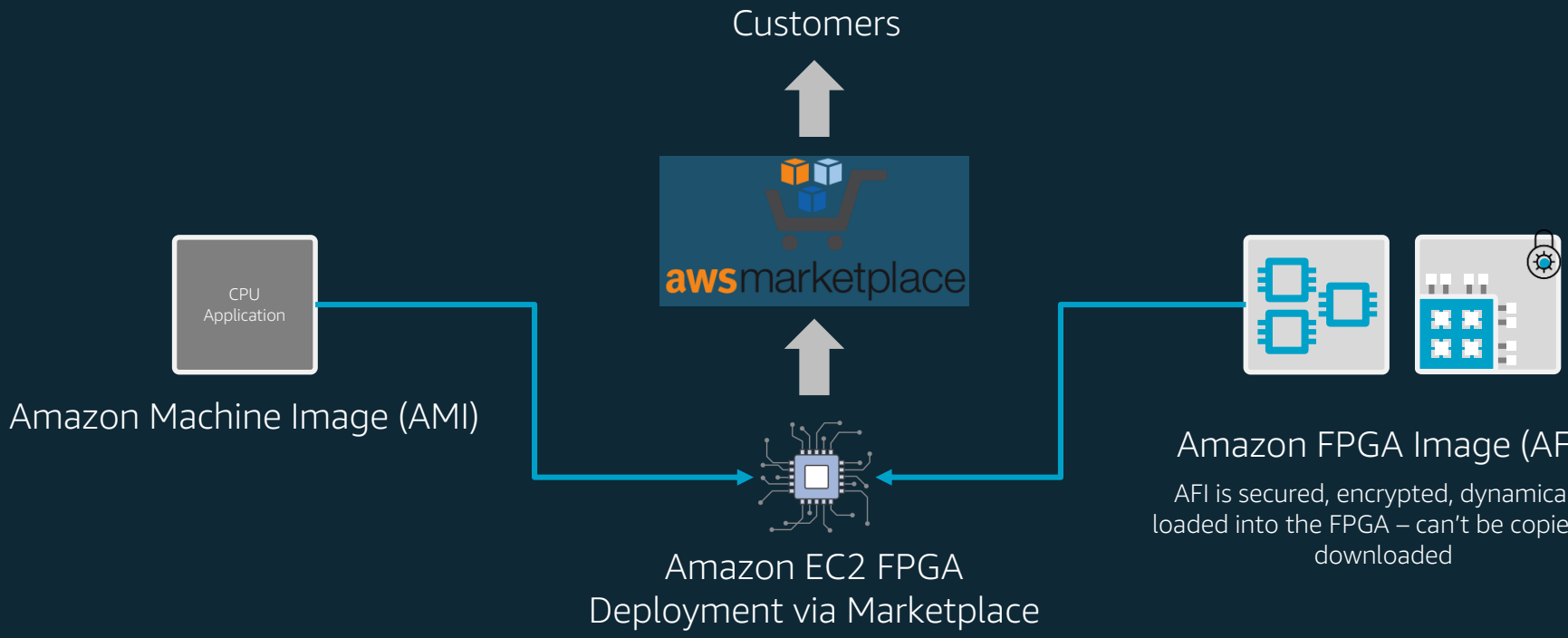
AWS Marketplace

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



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Delivering FPGA partner solutions



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

EC2 Accelerated Computing Instances

P3 GPU Compute Instance

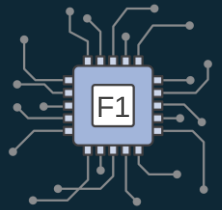
- Up to 8 NVIDIA V100 GPUs in a single instance, with NVLink for peer-to-peer GPU communication
- Supporting a wide variety of use cases including deep learning, HPC simulations, financial computing, and batch rendering

G3: GPU Graphics Instance

- Up to 4 NVIDIA M60 GPUs, with GRID Virtual Workstation features and licenses
- Designed for workloads such as 3D rendering, 3D visualizations, graphics-intensive remote workstations, video encoding, and virtual reality applications

F1: FPGA instance

- Up to 8 Xilinx Virtex® UltraScale+™ VU9P FPGAs in a single instance. Programmable via VHDL, Verilog, or OpenCL. Growing marketplace of pre-built application accelerations.
- Designed for hardware-accelerated applications including financial computing, genomics, accelerated search, and image processing



© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

Thank You!

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.