Develop & deploy accelerated applications at global scale on Amazon EC2 F1 Instances

Michał Skiba

March 24, 2021

AWS IS BUILT TO SUPPORT VIRTUALLY EVERY WORKLOAD

World-class performance, security, and innovation

Broadest and deepest platform choice

Tools for managing cost and complexity

Enabling enterprise applications

Bringing the cloud closer to you

Amazon EC2: Broadest and deepest platform choice

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

+

CAPABILITIES

Choice of processor
(AWS, Intel, AMD)

Fast processors
(up to 4.0 GHz)

High memory footprint
(up to 12 TiB)

Instance storage
(HDD and SSD)

Accelerated computing
(GPUs and FPGA)

Networking
(up to 100 Gbps)

Bare Metal

Size
(Nano to 32xlarge)

+

OPTIONS

Amazon EBS

Amazon Elastic Inference

=

MORE THAN

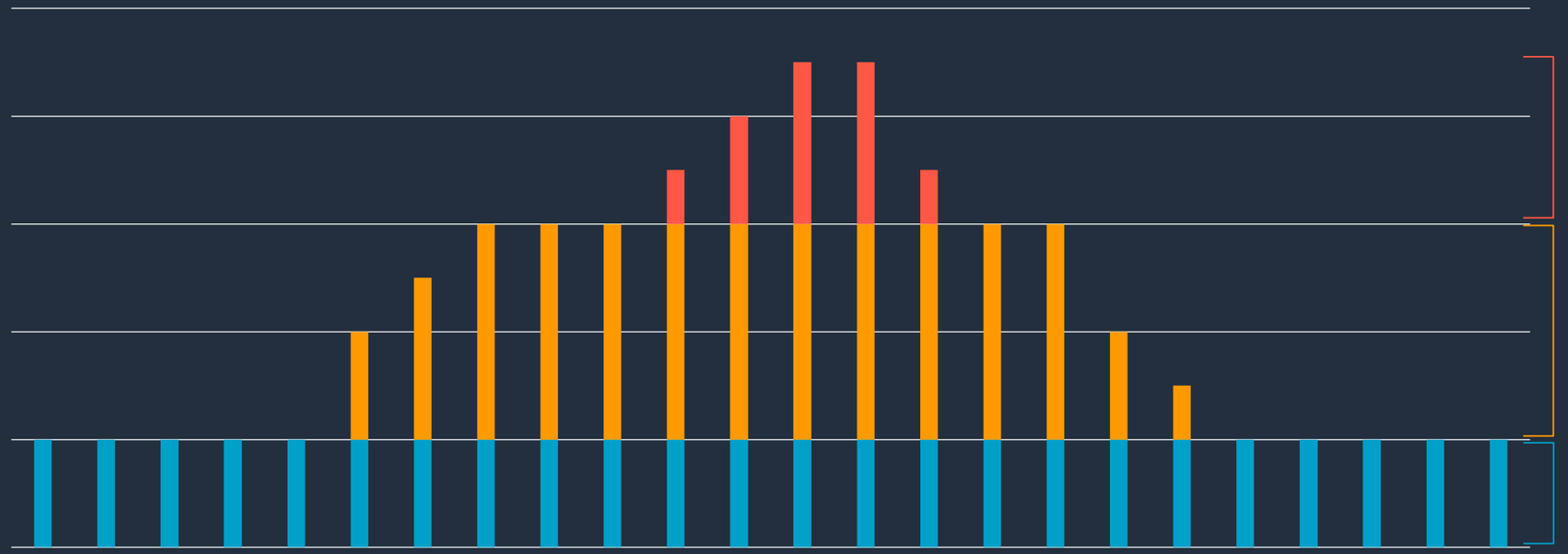
300+

INSTANCE TYPES

for virtually every workload and business need

aws

Simplifying capacity and cost optimization



SCALE USING **SPOT** FOR FLEXIBLE, FAULT-TOLERANT WORKLOADS

SCALE USING **ON-DEMAND** FOR NEW OR STATEFUL SPIKY WORKLOADS

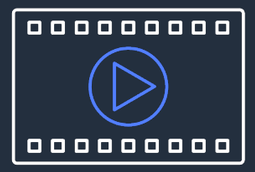
USE **RIs** AND **SAVINGS PLANS** FOR KNOWN/ STEADY-STATE WORKLOADS

AWS SERVICES MAKE THIS EASY AND EFFICIENT

- Amazon EC2 Auto Scaling
- EC2 Fleet
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Thinkbox
- Amazon EMR
- AWS CloudFormation
- AWS Batch

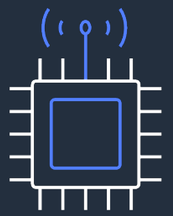
AWS FPGA customer use cases

HPC



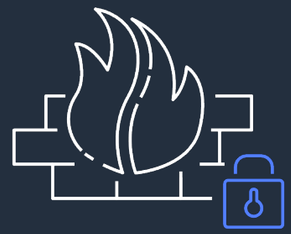
Video and Imaging

Database and Analytics



ASIC Prototyping

Fintech



Security

“We have been able to use the FPGA-based F1 instances to greatly speed up the process of whole genome sequencing. As a result, what used to take us 20 hours of compute time can now be achieved in only 3 hours.”

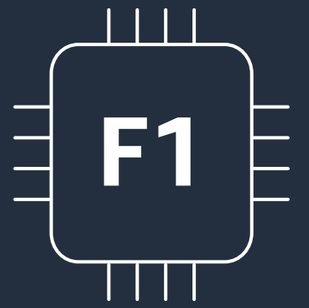
–Professor Dr. Torsten Haferlach, Chief Executive Officer, Munich Leukemia Lab

“Using Amazon EC2 F1 instances, Valtix can provide high performance and low latency SSL encryption/decryption and deep packet inspection (DPI) for threat detection.“

– Vijay Chander, CTO and Co-founder, Valtix

aws

Amazon EC2 F1 Instances



- Xilinx 16nm UltraScale+ VU9P FPGA – Up to 8 FPGAs per instance
  - Each FPGA includes 64 GiB DDR4 + ECC, 512bit wide channel
  - Each FPGA with PCIe Gen3 x16 connectivity
  - 2.5 million logic elements and 6,800 Digital Signal Processing engines
- Three different instance sizes with up to 64 vCPUs per instance
- Up to 1 TB of RAM
- Local NVMe SSD storage, data automatically encrypted at-rest

Model	FPGA	vCPU	Memory (GiB)	Instance storage (GiB)	Networking performance	EBS bandwidth
f1.2xlarge	1	8	122	1 x 470 NVMe SSD	Up to 10,000 Mbps	1,700 Mbps
f1.4xlarge	2	16	244	1 x 940 NVMe SSD	Up to 10,000 Mbps	3,500 Mbps
f1.16xlarge	8	64	976	4 x 940 NVMe SSD	25,000 Mbps	14,000 Mbps

aws

Amazon EC2 F1 Instance global footprint

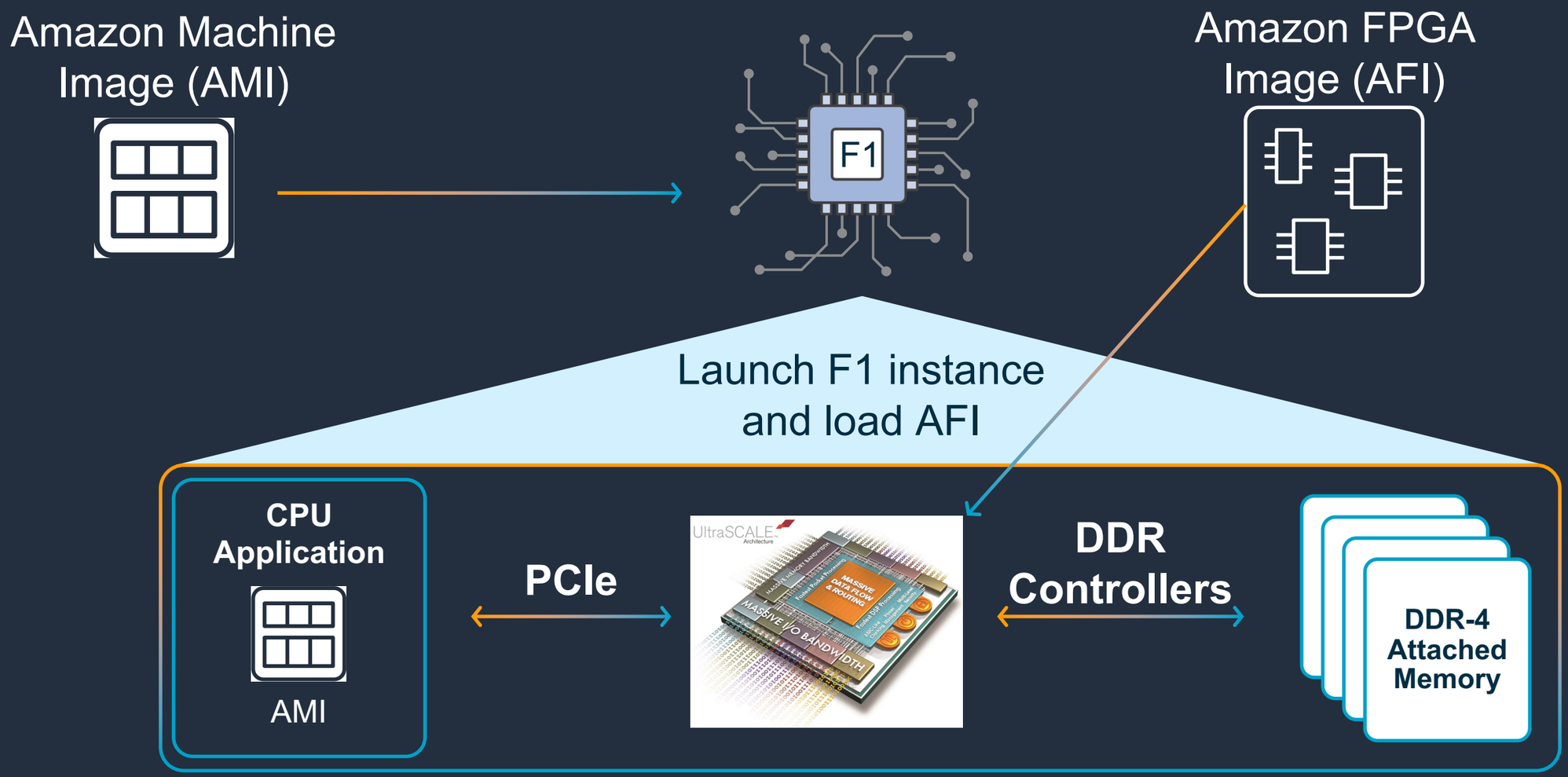
F1 instances available around the World to meet data gravity & sovereignty needs

North America	Europe/Middle East/Africa	Asia Pacific
US East (Northern Virginia)	Europe (Ireland)	Asia Pacific (Sydney)
US West (Oregon)	Europe (Frankfurt)	Asia Pacific (Singapore)
GovCloud (US-West)	Europe (London)	Mainland China (Beijing)
US East (Ohio)		
Canada (Central)		

Orange denotes preview availability

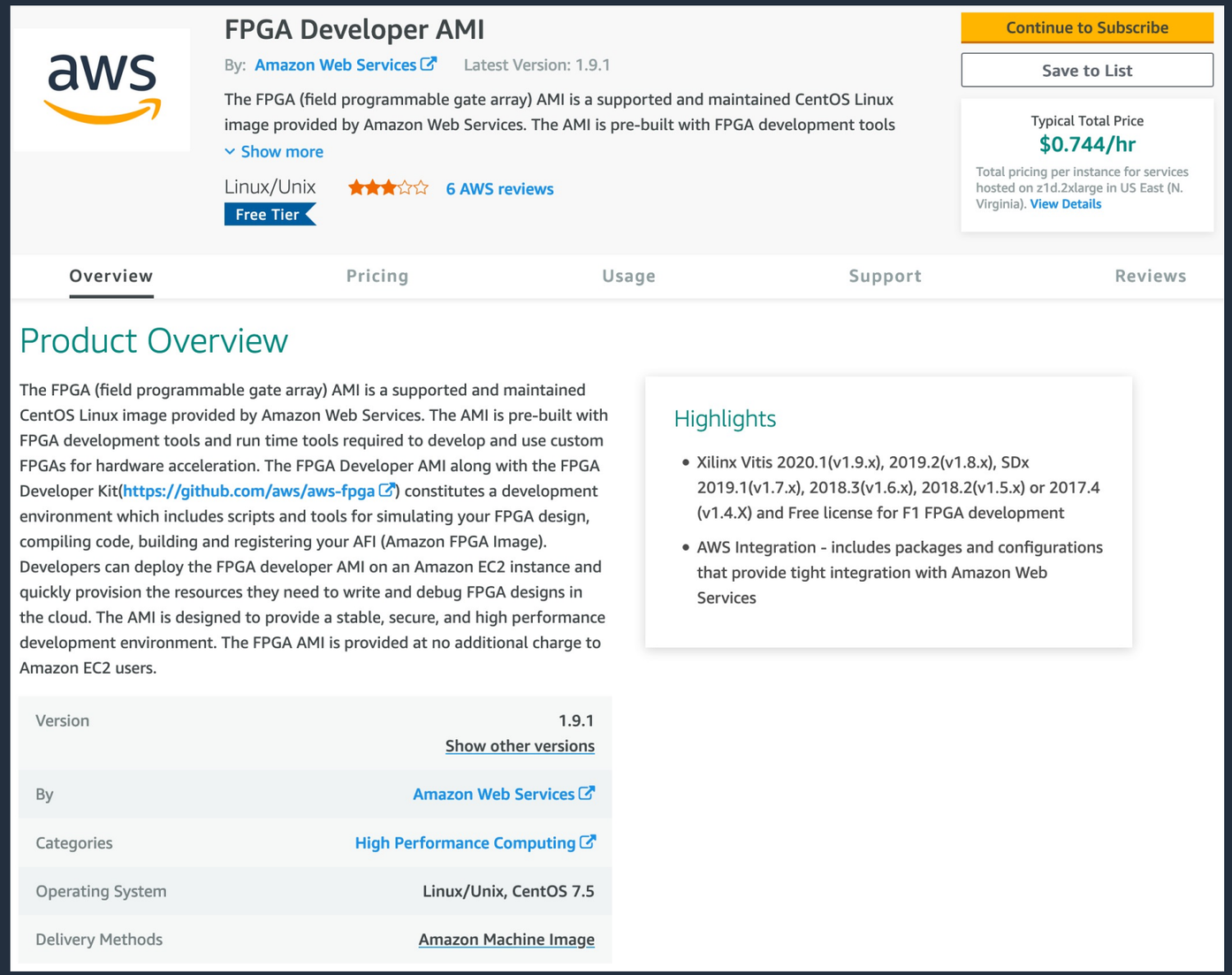
aws

Programming Amazon EC2 F1 Instances



FPGA Developer Amazon Machine Image (AMI)

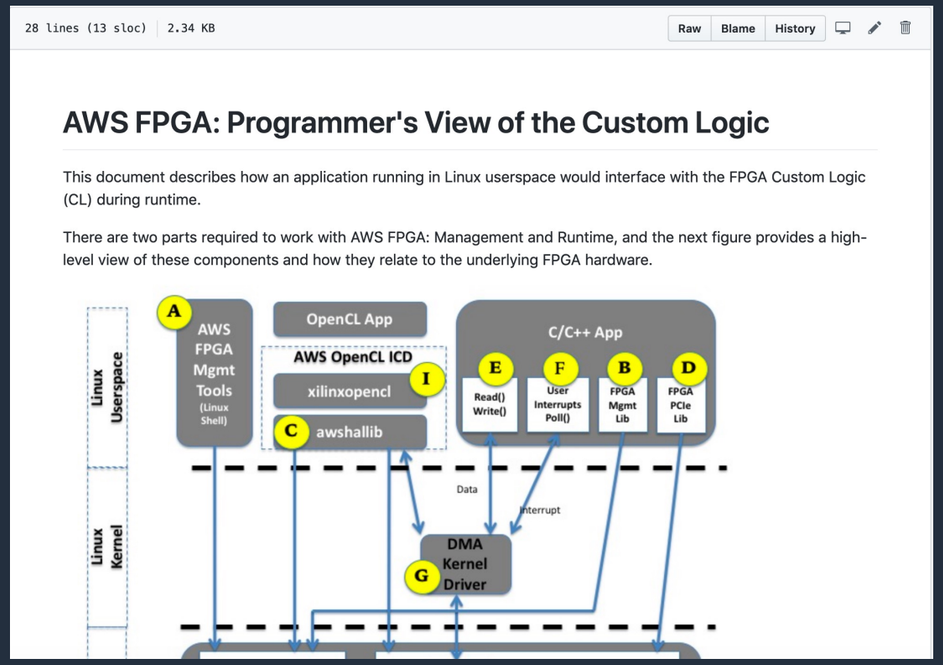
- Amazon Machine Image (AMI) launches virtual machine in minutes on any compute or FPGA instance
- New release supports Xilinx Toolset v2020.2 (Vivado and Vitis), and includes licenses
- Regularly updated with new developer resources
- Available for CentOS & Amazon Linux

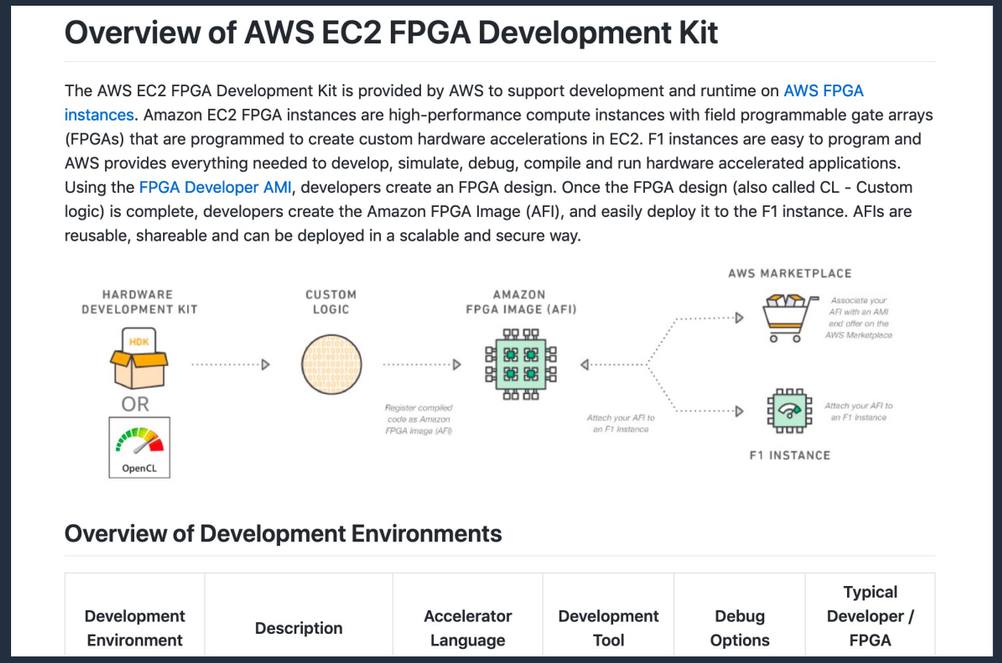


AWS FPGA developer resources

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

AWS FPGA Github contains all the drivers, code, examples and tutorials needed to develop a hardware acceleration for the AWS FPGAs





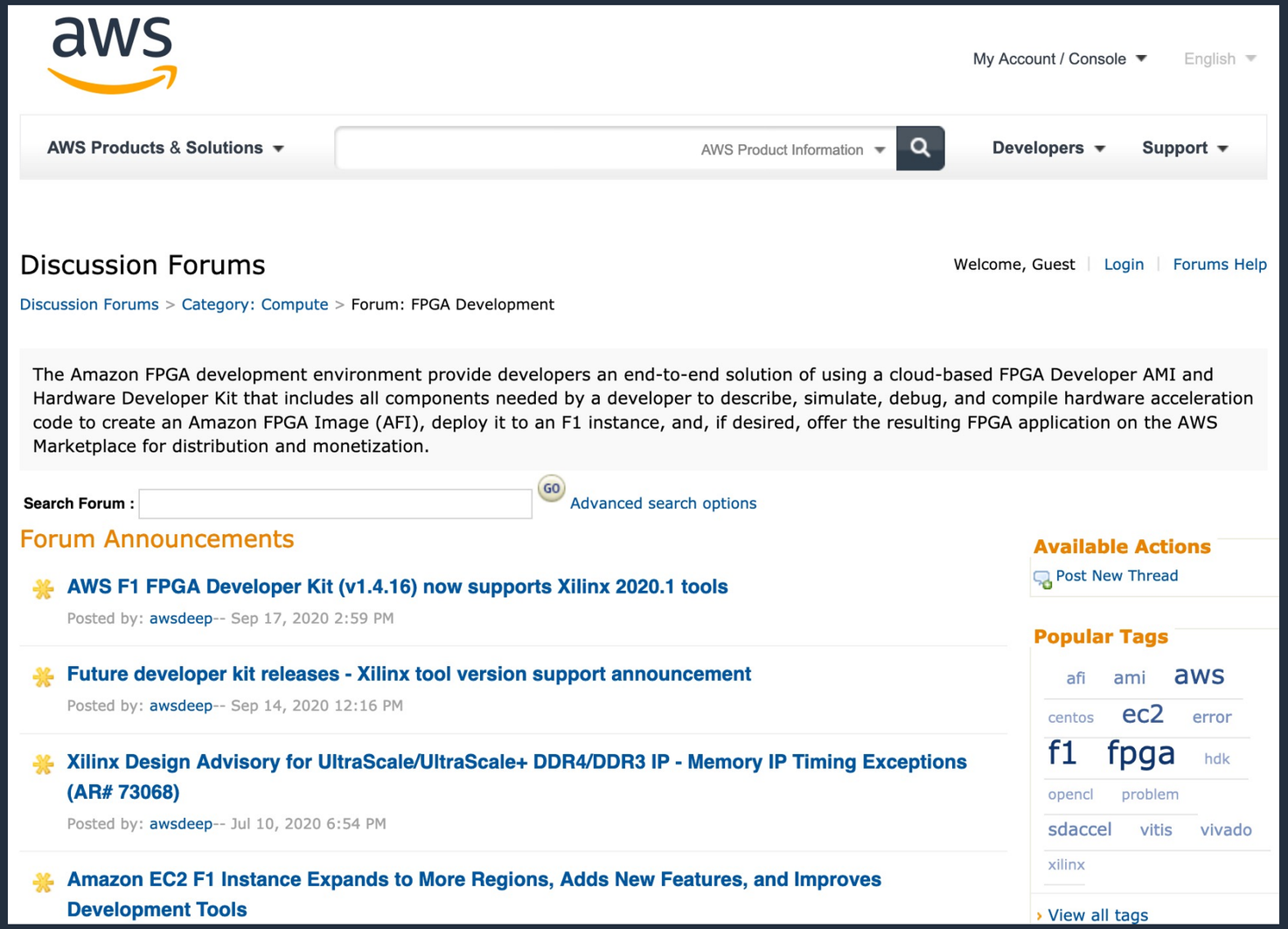
AWS FPGA support forum

Built in to your AWS Console

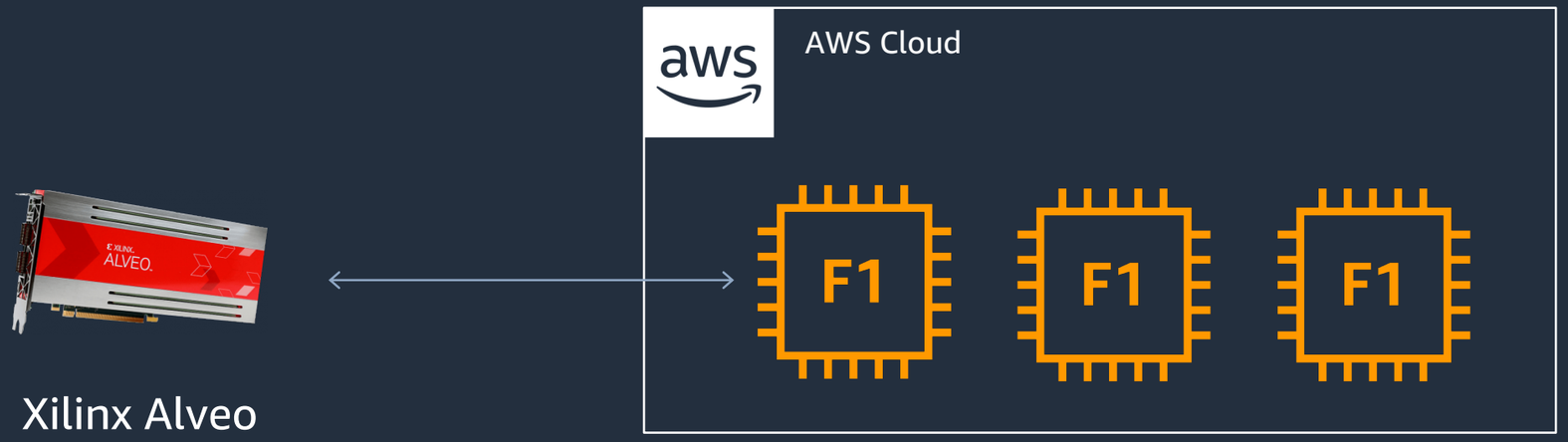
Deep knowledge repository

Hundreds of questions and answers related to the AWS FPGA compute platforms

Monitored by the community, AWS and Xilinx

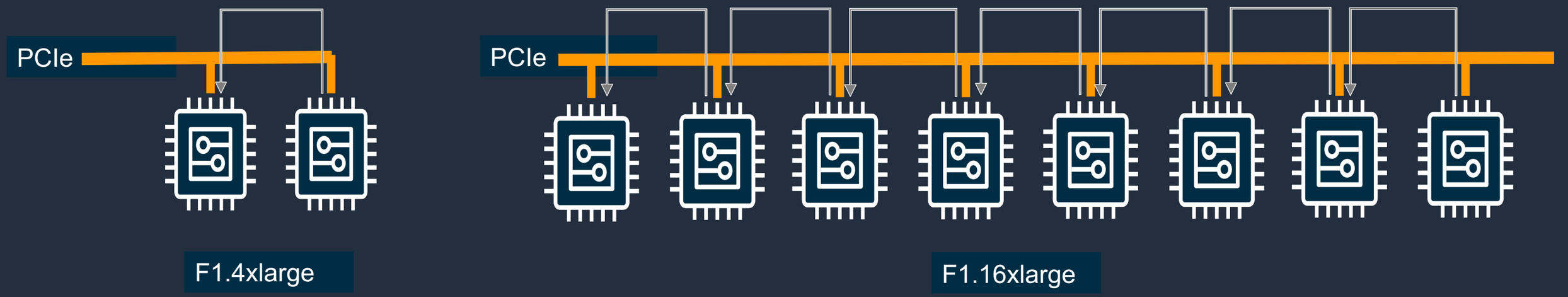


Scale on-prem applications to Amazon EC2 F1 Instances



- Re-use on-prem investment to scale to AWS cloud
- Same Vitis design with no/minimal changes
- https://github.com/aws/aws-fpga/blob/master/Vitis/docs/Alveo_to_AWS_F1_Migration.md

Improving latency with FPGA peer-to-peer (P2P)



P2P enables data transfers between FPGAs (F1.4xlarge and F1.16xlarge)

- Bypass host memory: improving latency and scalability
- Using F1 XDMA – directly write data to/from FPGA memory

Best use cases are:

- AFI Pipeline (as shown above)
- Other topologies: Producer/Consumer

https://github.com/awslabs/aws-fpga-app-notes/blob/master/Using-PCIe-Peer2Peer/README.md

aws

Simplified networking interface

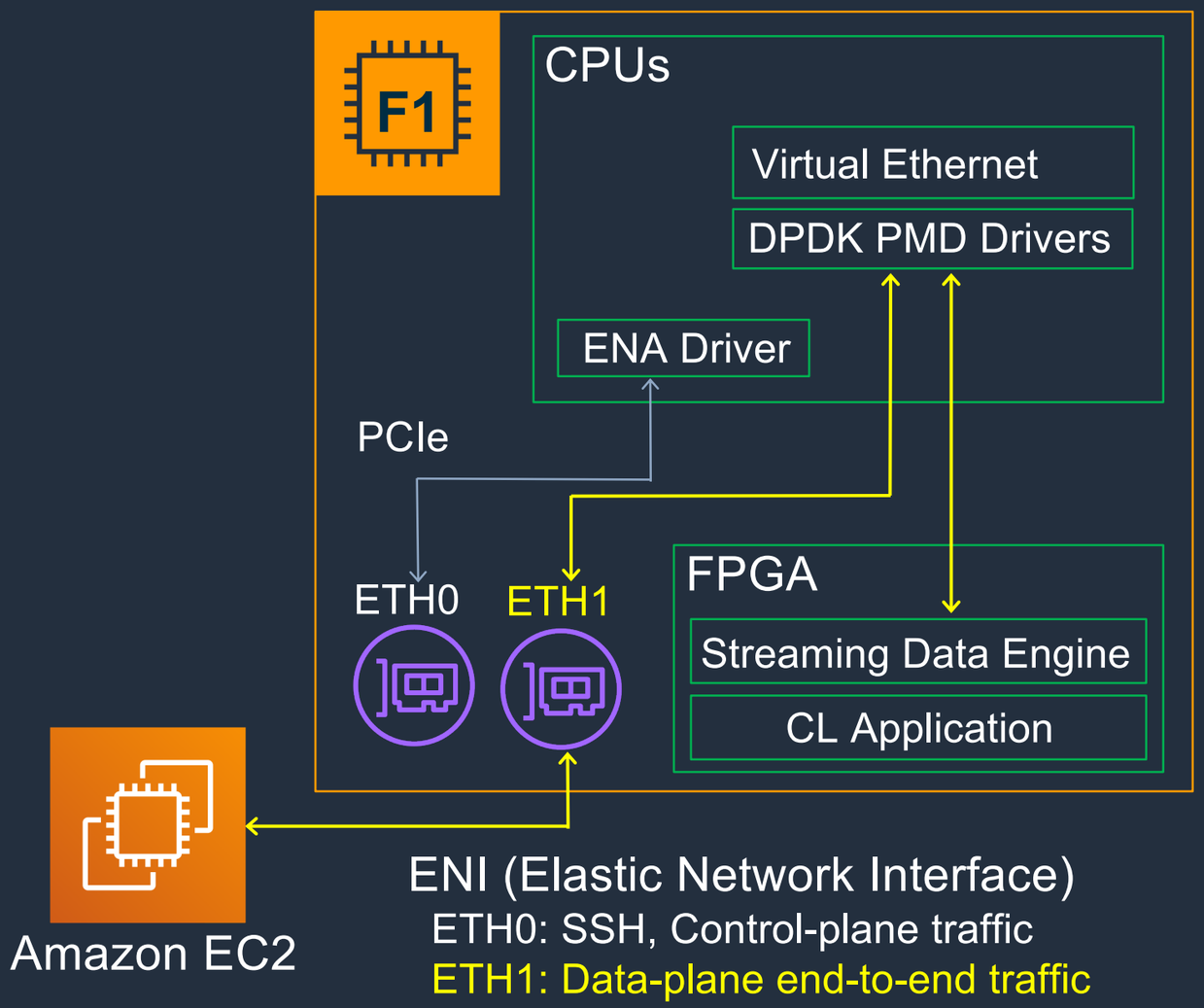
Integrated virtual Ethernet interface

Open source example design

Leveraging **DPDK** for high performance

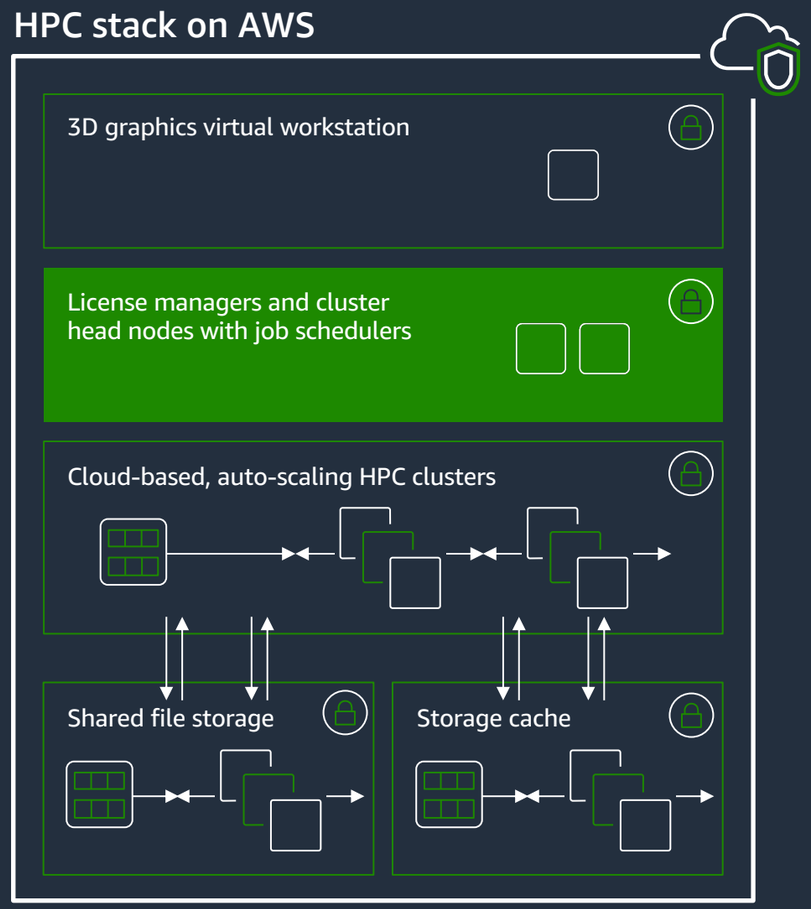
Use cases:

- Routing, Switching
- Packet processing
- SDN (Software Defined Networking)
- DPI (Deep Packet Inspection)
- Stream encryption or compression



https://github.com/aws/aws-fpga/tree/master/sdk/apps/virtual-ethernet

Simplified workload scaling with AWS ParallelCluster



Enables customers to quickly build an HPC compute environment in AWS and run large-scale workloads

Offers a variety of batch schedulers such as AWS Batch, SGE, Torque, and Slurm

Customers can build higher level workflows, such as a Genomics portal that automates the entire DNA sequencing workflow

F1 Instances available in AWS Educate Classrooms

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

F1Educator@amazon.com

Get started today!

Amazon EC2 F1 Instances

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

Github Repository with Design Guidance & Examples

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

Leverage our partners' solutions via the AWS Marketplace

https://aws.amazon.com/marketplace/

aws

Thank you