Alveo Overview for Hyperscale and HPC Applications

Viraj Paropkari
Senior Manager, Global Data Center Marketing

Arm HPC User Group (AHUG) 2019
20th June 2019, Frankfurt, Germany

XILINX

Agenda

- Data Center Focus & Strategy
- Alveo Data Center Boards & Ecosystem Overview
- Use Cases- Compute, Storage, Networking
- Future of Computing
- Getting Started

XILINX

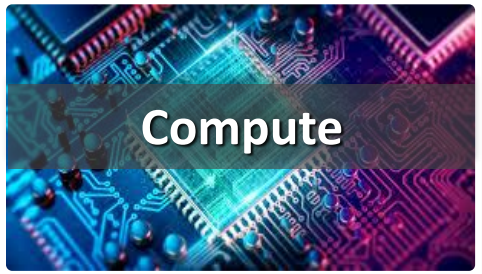
- Headquarters
- Research and Development
- Sales and Support

$3.0B	~4,400	20K+	60+	4,000+
Revenue	**Employees Worldwide**	**Customers**	**Industry Firsts**	**Patents**

Datacenter Focus & Strategy

XILINX

Data Center Opportunity



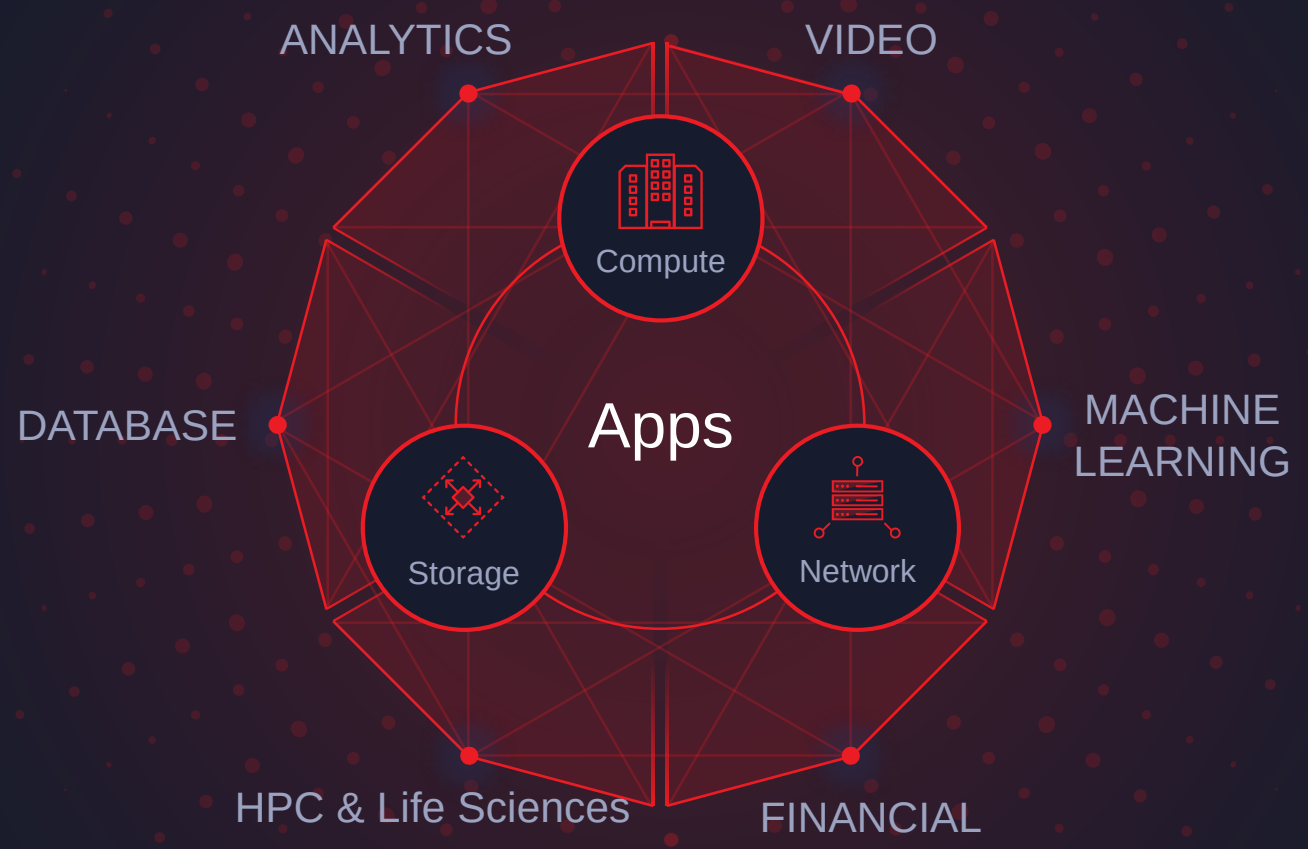
> Heterogeneous Computing post Moore's Law

> Exponential Data Growth

> Dawn of AI

XILINX

Era of Reconfigurable Accelerators

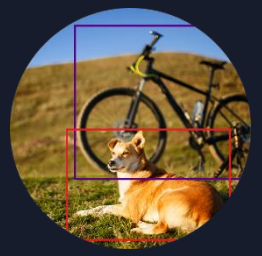


Algorithm Diversity and Fast Evolution

APPLICATIONS

Classification

Object Detection



Speech Recognition

Data Analytics

Recommendation Engine

Anomaly Detection

CNN

RNN, LSTM

RF, LR

MLP

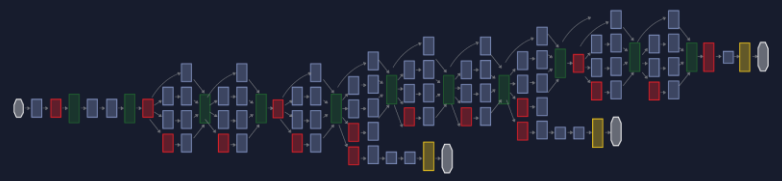
Diverse AI models and Neural Networks (NN's) for a broad range of applications

XILINX

New Algorithms Need New Architectures

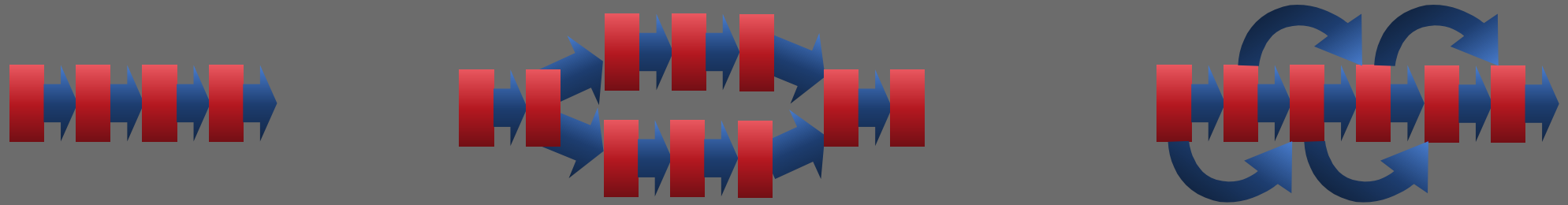
AlexNet

GoogLeNet



DenseNet

Highest throughput, latency, and efficiency requires different HW architecture



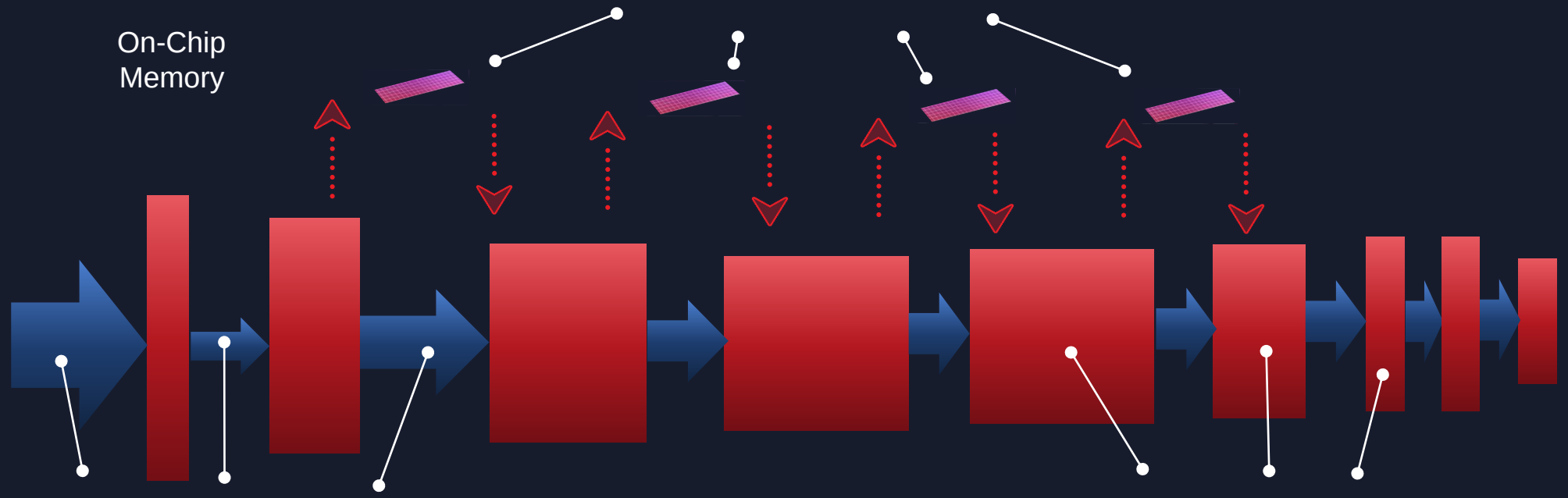
XILINX

Optimized Performance

Requires Custom Memory and Datapath

Off-Chip DDR

On-Chip Memory



XILINX

Reconfigurable Data Center Accelerators

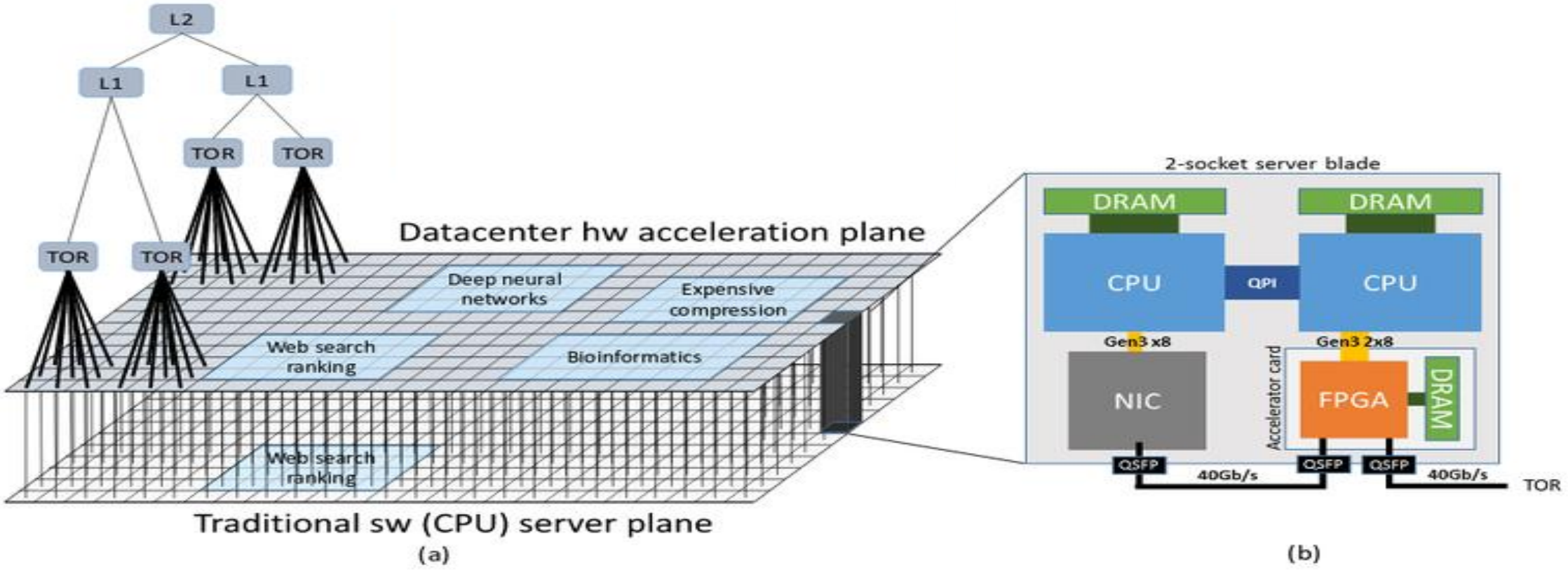


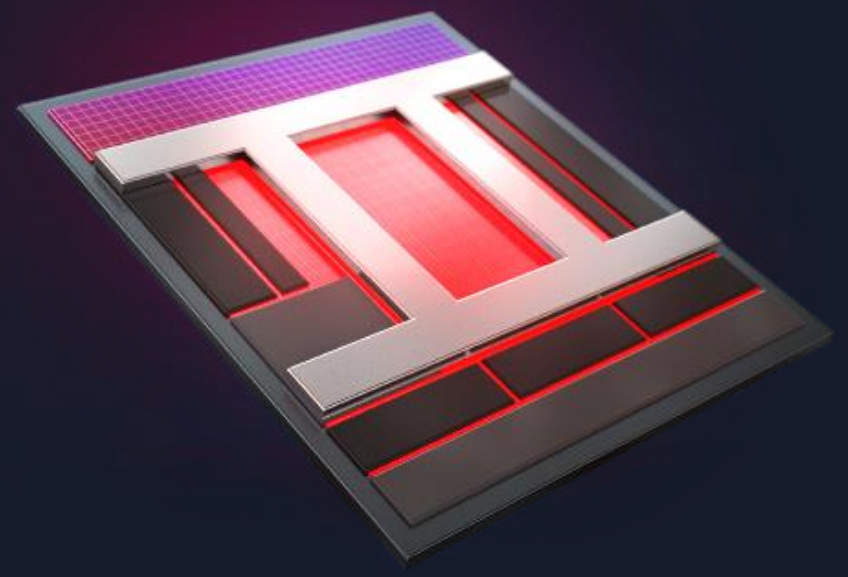
Fig. 1. (a) Decoupled Programmable Hardware Plane, (b) Server + FPGA schematic.

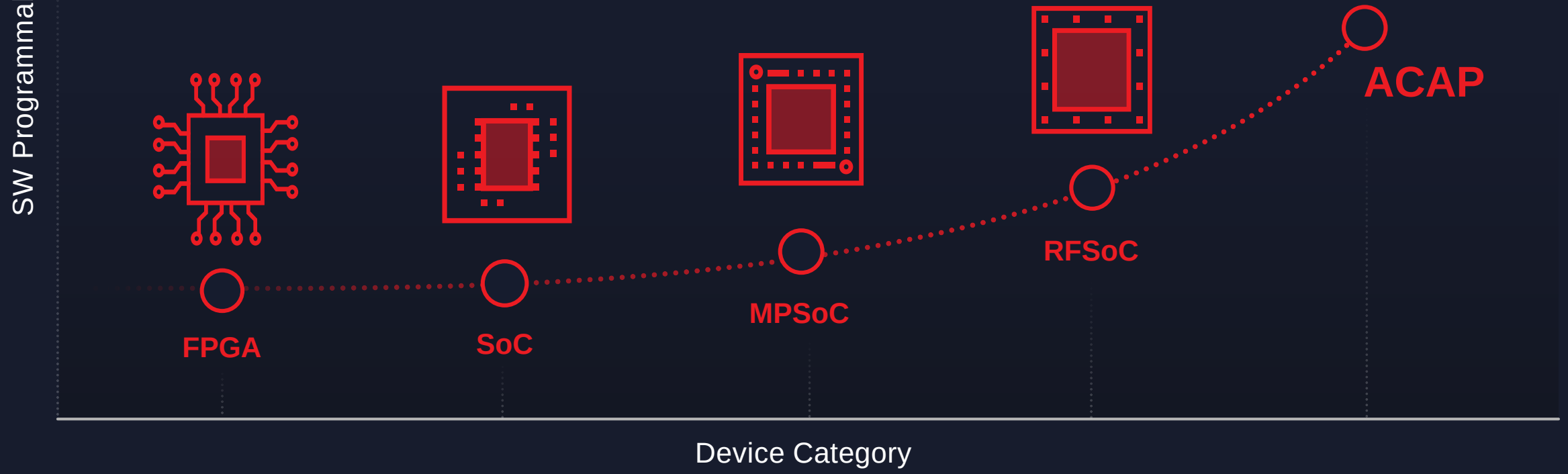
Tier1 HS adopting FPGA for general purpose acceleration across multiple in-house and external workloads

XILINX

Xilinx Transformation

From Devices to Platforms





XILINX

Alveo Product & Ecosystem Overview

XILINX

ALVEO™

FAST

Built for high throughput, ultra-low latency

Accelerate compute, networking, storage

ADAPTABLE

Deploy optimized domain-specific architectures

Adapt to changing algorithms

ACCESSIBLE

Deploy in the cloud or on-premises

Rich set of accelerated Applications

XILINX

U200

892K LUTs	35MB Internal SRAM Capacity	31TB/s Internal SRAM Bandwidth	3100img/s CNN Throughput*

U250

1,341K LUTs	54MB Internal SRAM Capacity	38TB/s Internal SRAM Bandwidth	4100img/s CNN Throughput*

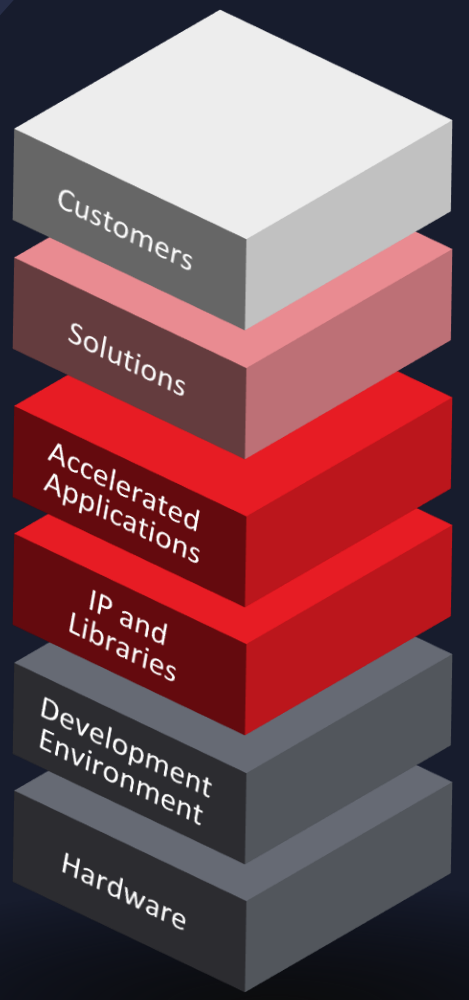
U280

1,079K LUTs	41MB Internal SRAM Capacity	30TB/s Internal SRAM Bandwidth	460GB/s HBM2 Memory Bandwidth

*Low-latency GoogLeNet v1

XILINX

ALVEO Solution Stack



End Customers

Solution Providers

App & IP Developers

Channel Partners

Data Analytics	Video & Image Processing	Machine Learning	Life Science & HPC	Financial Computing

Titan IC, NGCODEC, edico genome, MAXELER Technologies, DEEPHI, PLUNIFY, CTACCEL, DEEPOLY, falcon COMPUTING, Mipsology, byte LAKE, XELERA, BLACKLYNX, VITESSE DATA, skreens, swarm64, ALGO-LOGIC, MEGH COMPUTING, boon logic, V-NOVA, Nextera Video, bigstream, rENIAC, inaccel, BigZetta Systems, LegUp, mle

aws, Tencent Cloud, NIMBIX, SUPERMICRO, inspur, DELL, Alibaba Cloud, Baidu 百度, HUAWEI

CLOUD **ON-PREMISE**

XILINX

Growing Ecosystem

Data Analytics

bigstream
BLACKLYNX
swarm64
XELERA
inaccel
VITESSE DATA
BigZetta Systems
LegUp
rENIAC

Life Sciences & HPC

illumina
byte LAKE
inaccel
Falcon COMPUTING
MAXELER Technologies
deep Genetics

Video Processing

NGCODEC
V-NOVA
MEGH COMPUTING
Nextera Video
skreens
VYUSYNC

Machine Learning

Mipsology
DEEPHi 深鉴科技
XELERA
SumUp

Financial Computing

ALGO-LOGIC
MAXELER Technologies
SciComp

Image Processing

DEEPOLY
CTACCEL

XILINX

Growing Acceleration Ecosystem

>2x
Applications

27 May 2018 » 65 Today

Published Apps

>6x
Partners

84 May 2018 » 555 Today

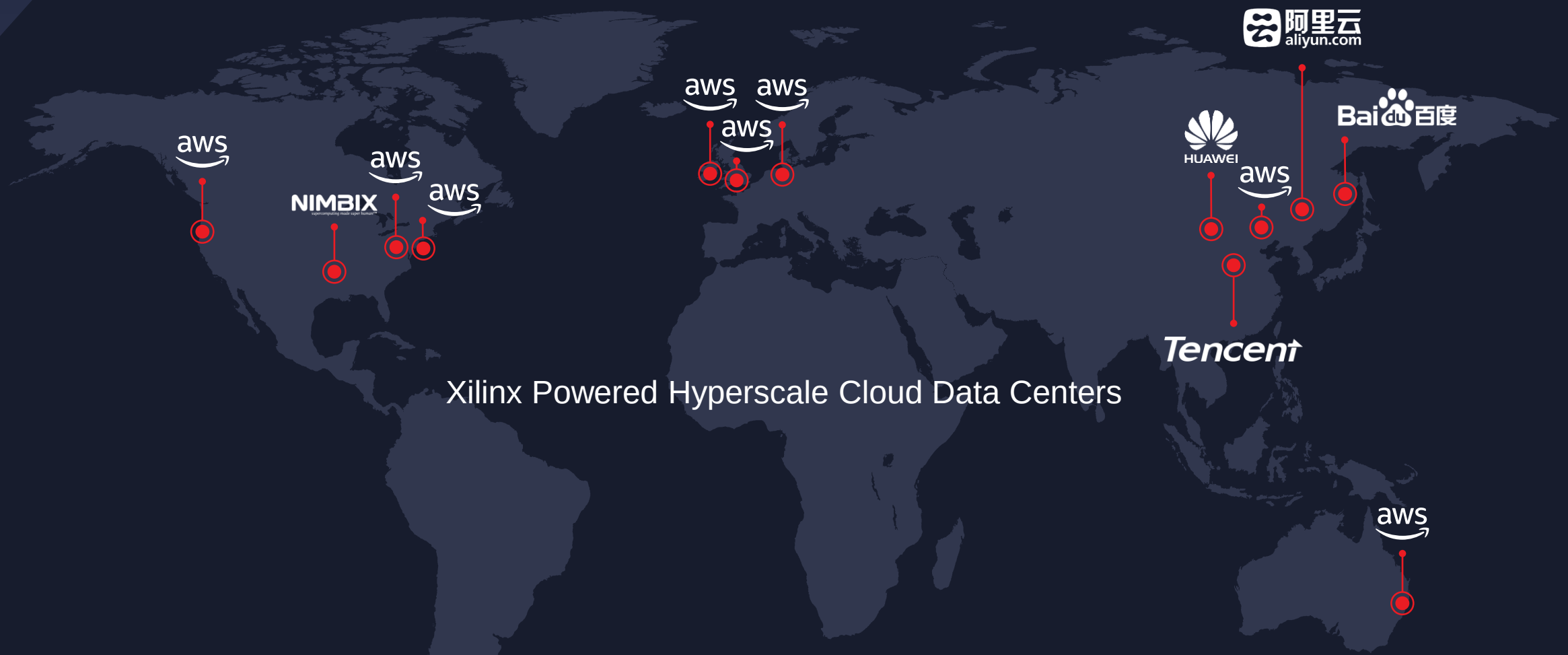
Accelerator Program

>4x
Developers Trained

716 May 2018 » 3089 Today

Companies + Academia

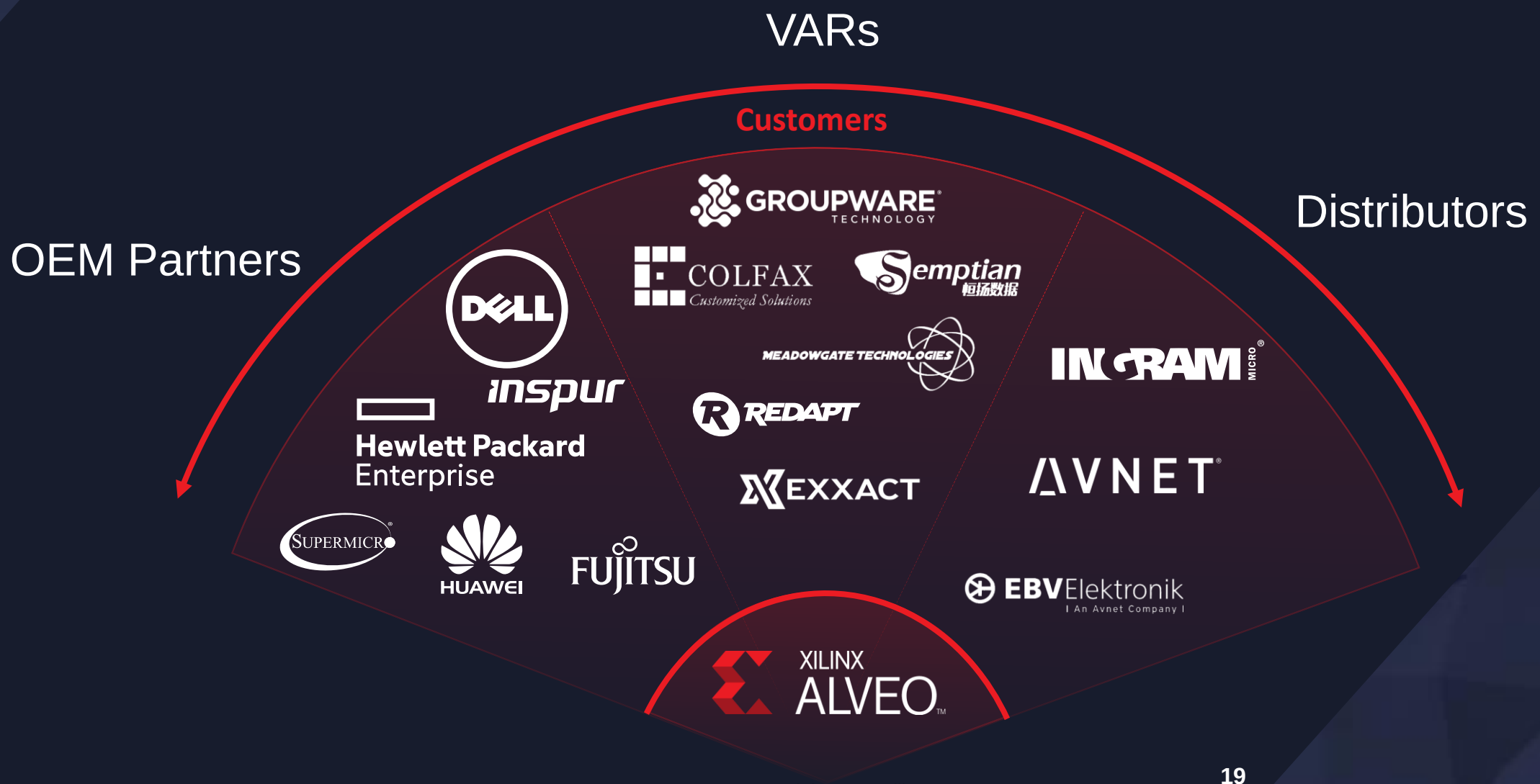
Growing Cloud Availability



Xilinx Powered Hyperscale Cloud Data Centers

XILINX

Growing Solution Partner Network



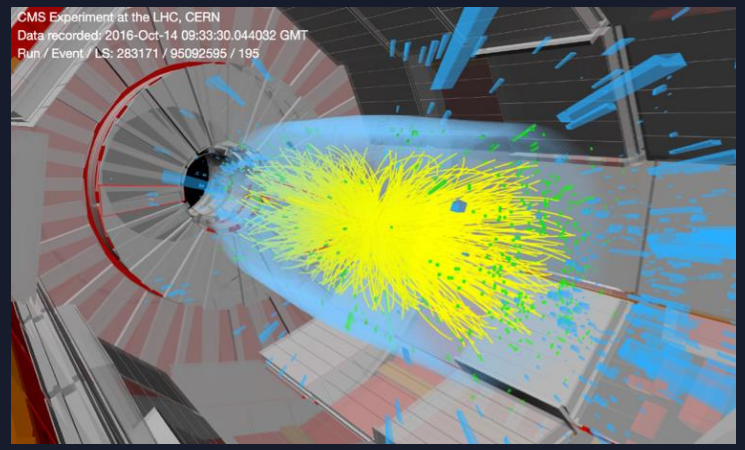
Use Cases

XILINX

AI Accelerated Dark Matter Search (CERN)

Real-time ML Inference + Sensor pre-processing

Achieving 100ns Inference Latency on 150 Terabytes/Second Data Rates Unachievable by CPUs & GPUs



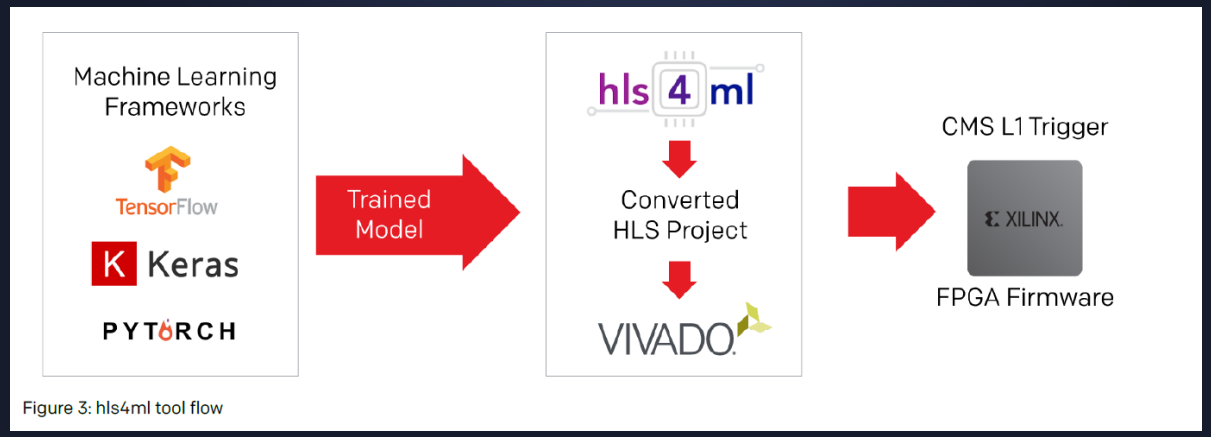
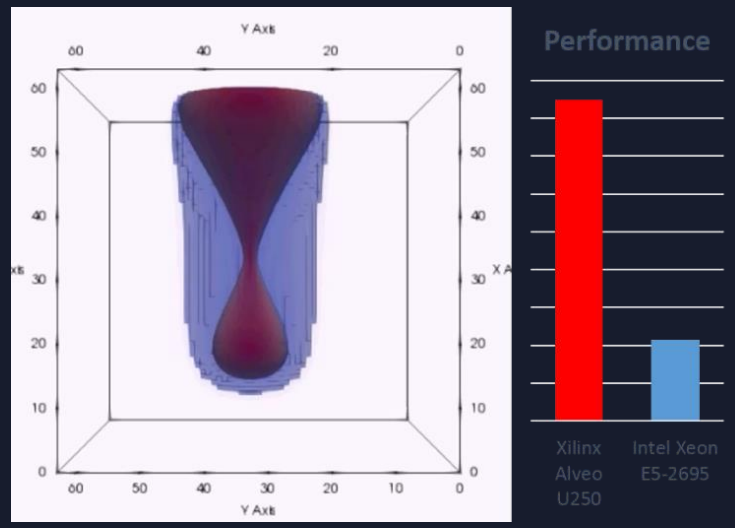


Figure 3: hls4ml tool flow

https://www.xilinx.com/content/dam/xilinx/publications/powered-by-xilinx/cern-case-study-final.pdf

Computational Fluid Dynamics

ALVEO Accelerated CFD Kernels



byte LAKE

Faster Time to insight, Fewer Nodes

- 4x Faster simulation time
- 80% lower energy consumption
- 6x better performance per Watt

XILINX

Precision Medicine

Genomic Data Analytics

Accelerates Sequencing by 90x

illumina

Source: Xilinx Analysis

XILINX

Live Video Streaming

Video Transcoding for VP9 Live Stream

30x higher performance

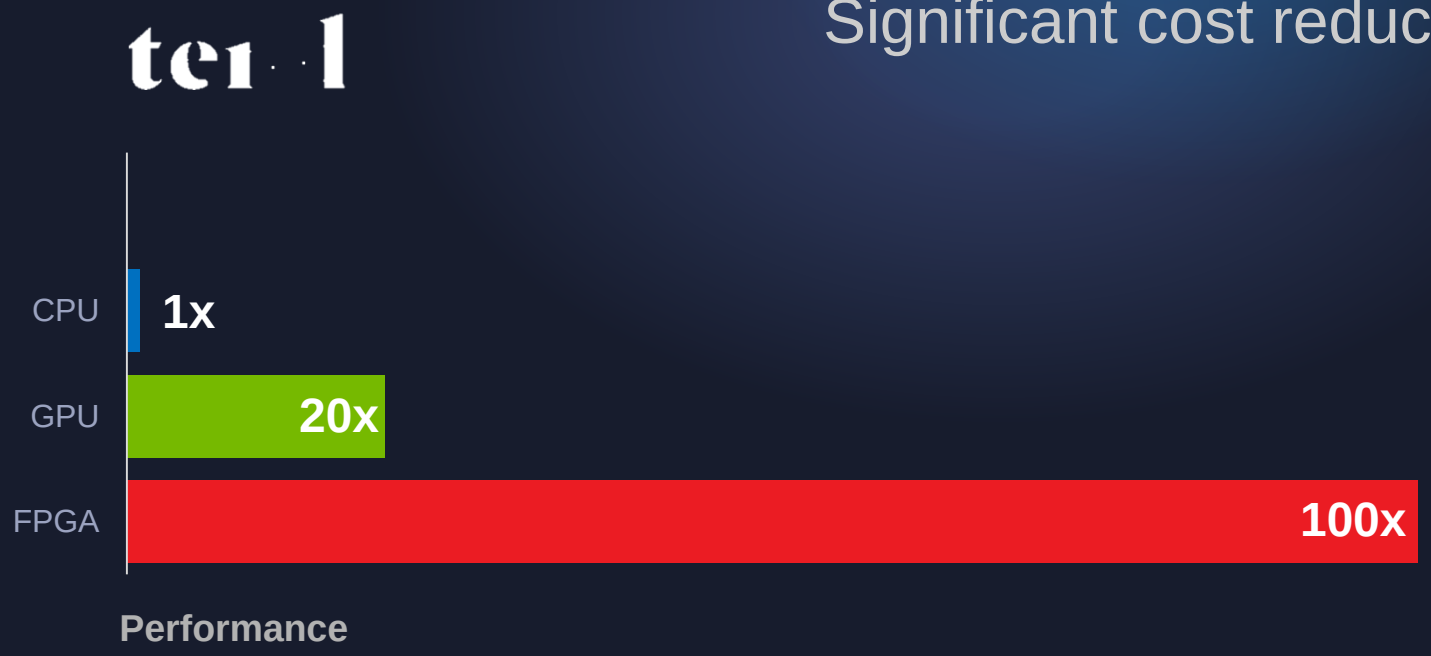
25% cost reduction

Source: Xilinx Analysis

Video Processing + AI Inference

100x higher performance

Significant cost reduction



Source: Xilinx Analysis

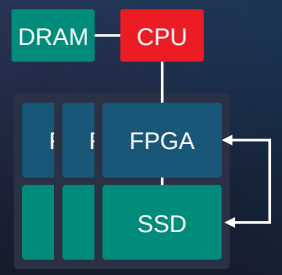
Computational Storage

5x Speedup for Data-Intensive Workloads

Compression
Encryption
Database query offloads

Reduces TCO >2x

Increased performance
Frees up Host CPU cycles



NGD systems
SmartIOPS
ScaleFlux™

Smart Networking Acceleration

SolarFlare SmartNIC
Processing >100 million packets/sec
Dual 100G QSFP
Under 75 watts

More Efficient Infrastructure
Adaptable to support new protocols
Extensible with programming in P4 & C/C++

Data Center Platforms

Software Applications

Python/Java/C++ Programmability

Compute Platform | Storage Platform | Network Platform

Software Stack

Software Frameworks

IDE

Libraries, Compilers, Middleware

Firmware

Hardware

XILINX ALVEO

Hardware Devices and Boards

Future of Computing

XILINX

Expanding Xilinx Value with ACAP

Designed for SW programmability

Integrates acceleration functions

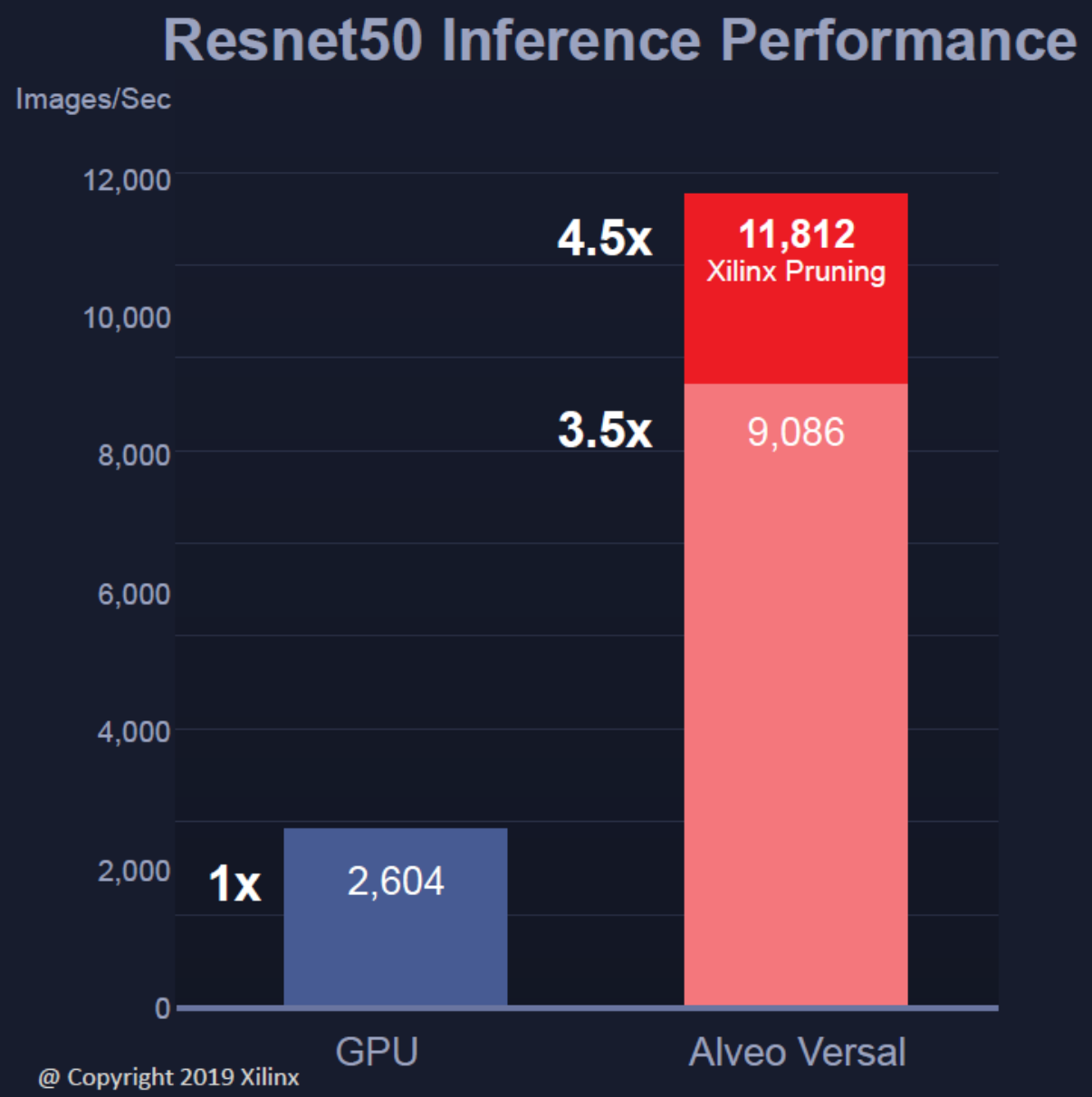
Higher performance

More power efficient

XILINX. VERSAL™

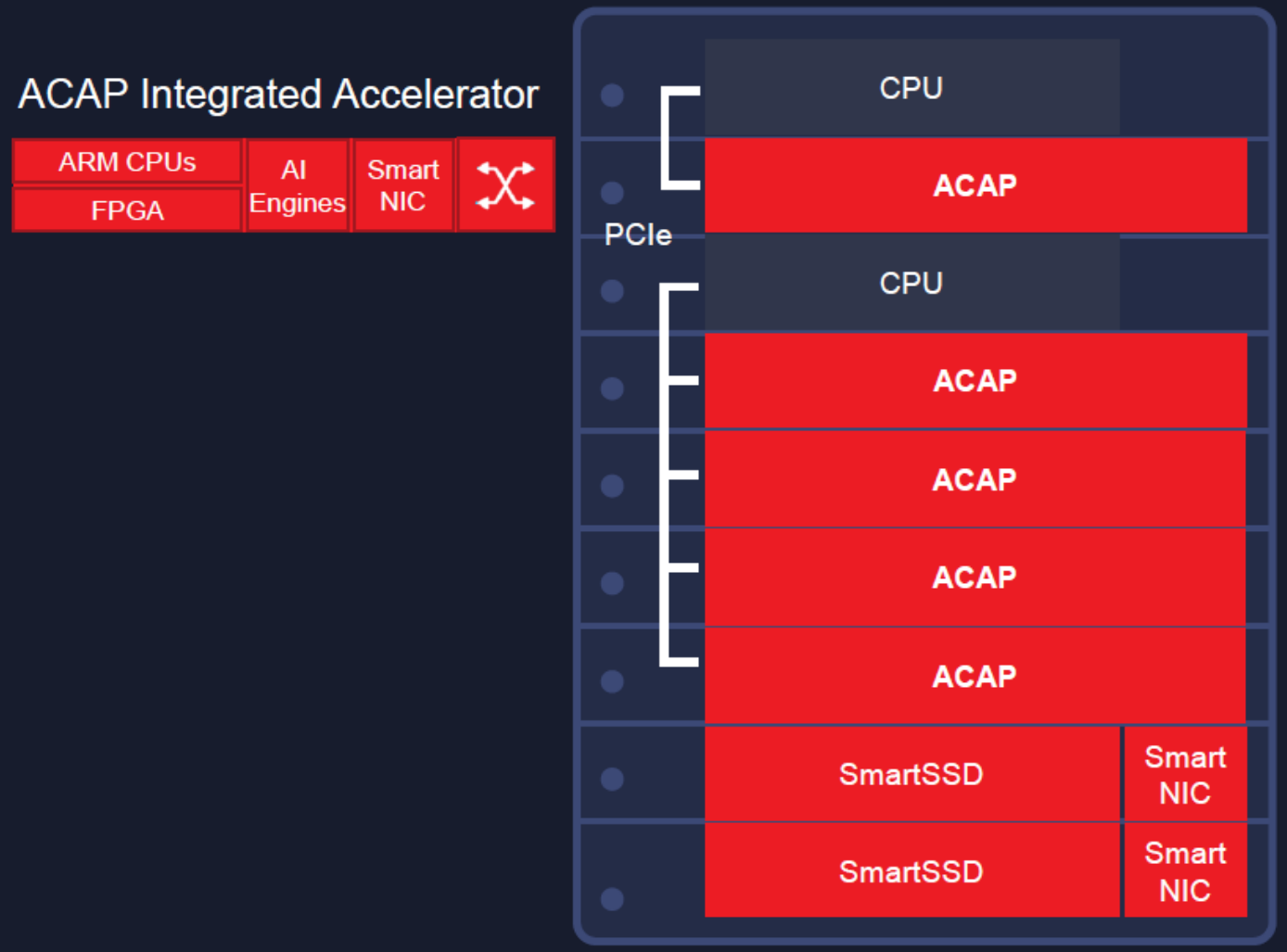
Now Shipping to multiple tier-1 customers.
GA 2H 2019

Versal AI Inference Performance

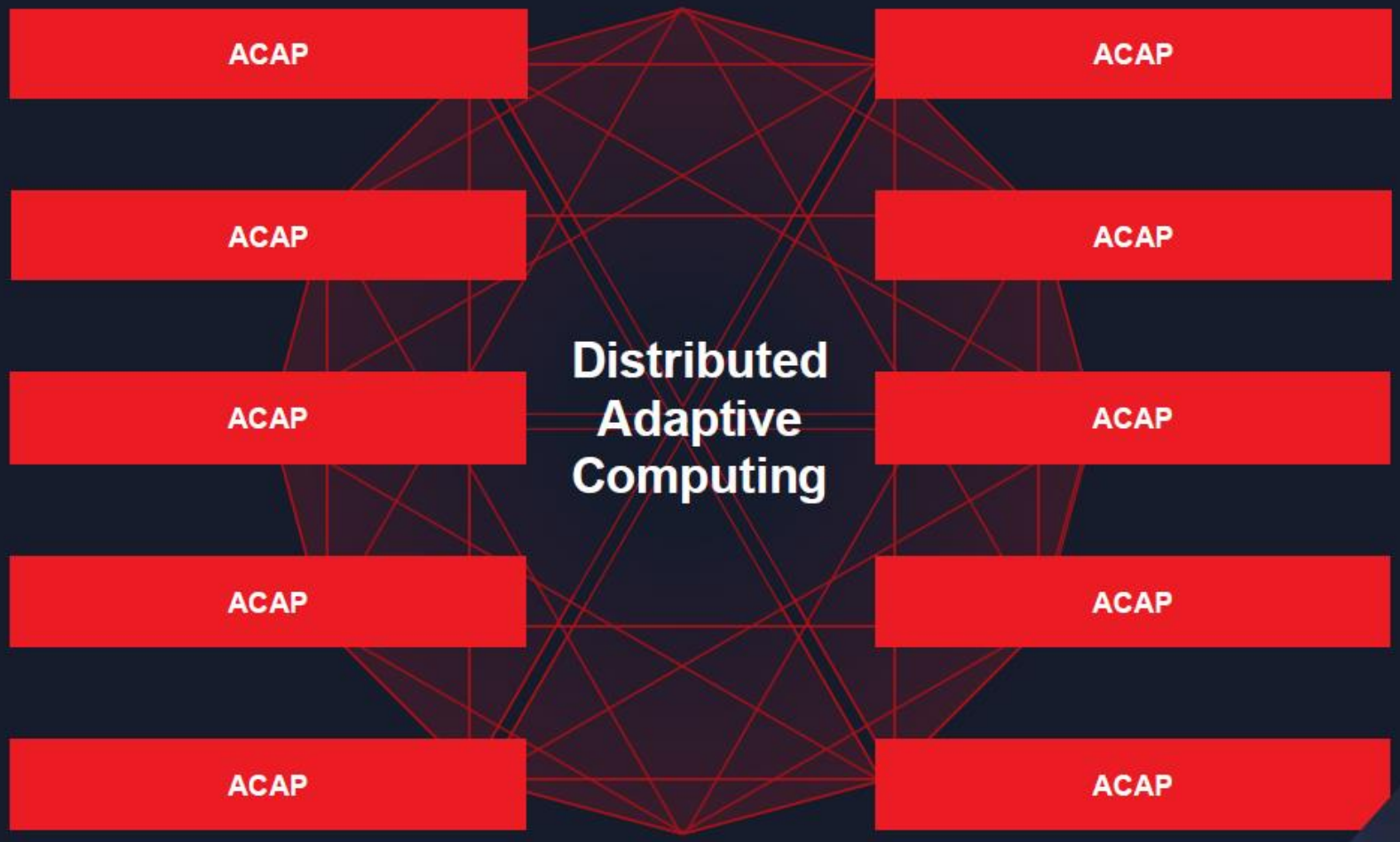


Sources:
GPU: Nvidia T4 TensorRT 5, Published March 2019
(INT8, Batch=4, 1.5ms Latency)
Alveo Versal Card, Projected (INT8, Batch=8, 1.5ms Latency)

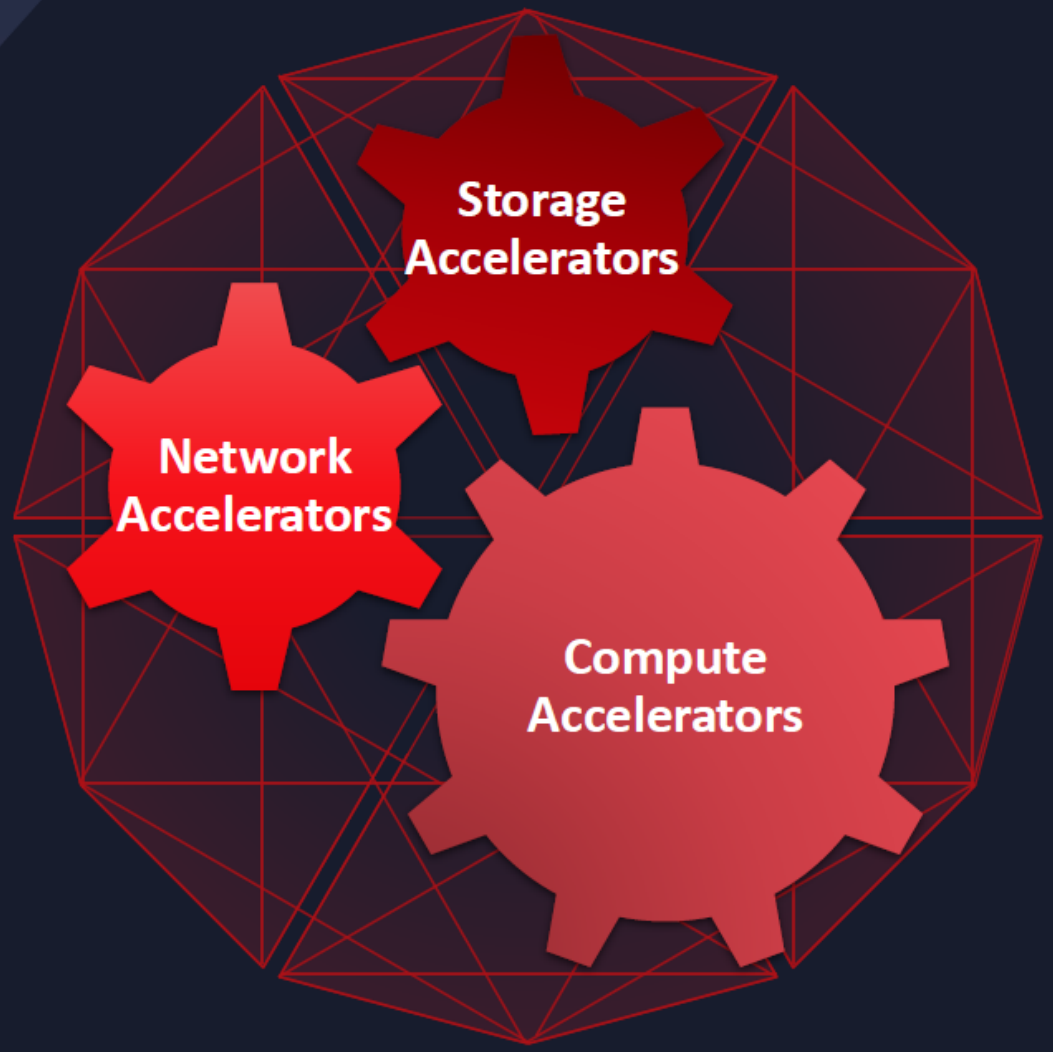
Vision: ACAP for Distributed Adaptive Computing



Harnessing the Power and Flexibility of Multiple Accelerators



Value of Distributed Adaptive Computing



Whole application acceleration across compute, networking and storage

Virtualization and independent scale-out of resources

Best compute optimization and utilization within the lowest power envelope

Significant TCO savings

Getting Started

XILINX

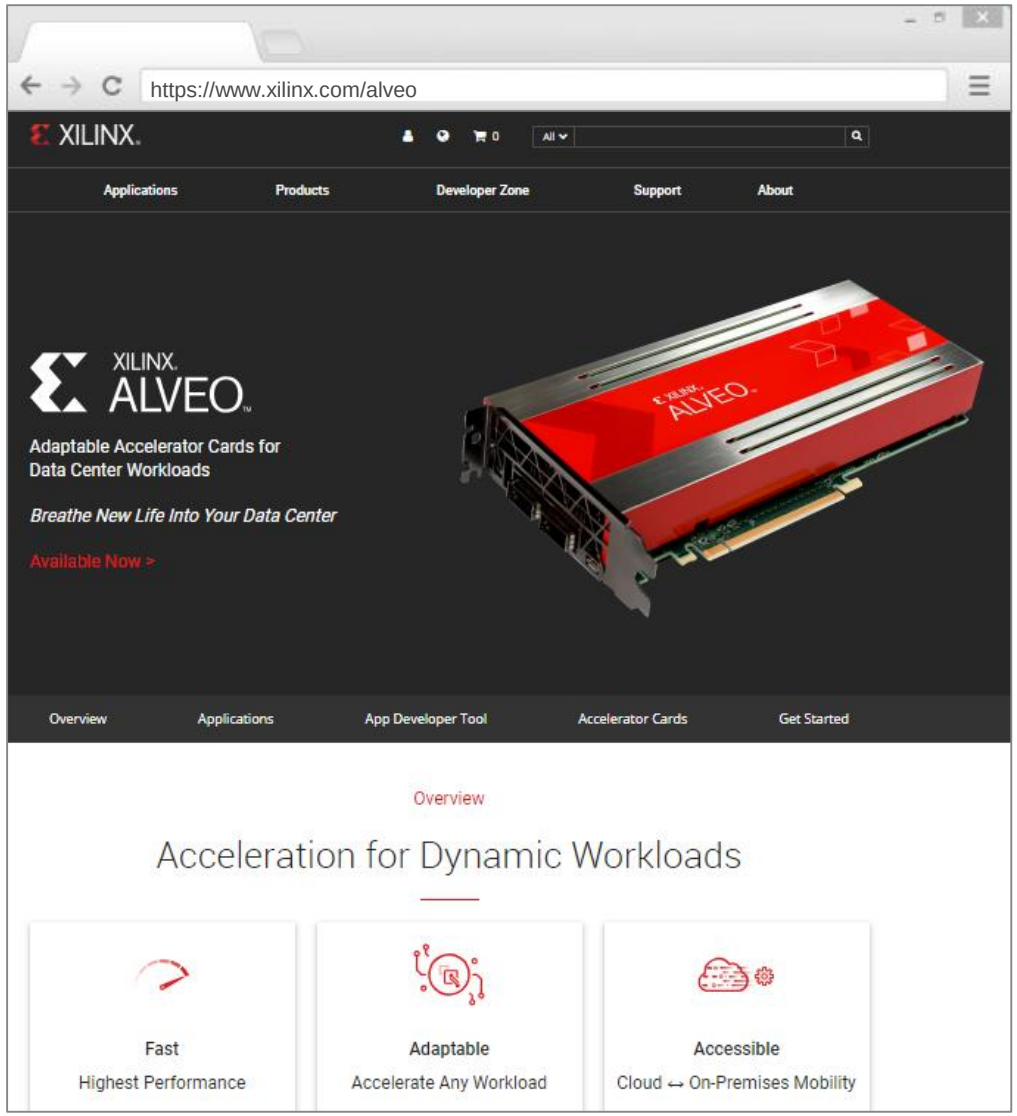
Accelerator Program

A quick start program to enable companies to accelerate products and services with Alveo™ Data Center accelerator cards and FPGAs in the cloud. Target workloads include data analytics, genomics, video processing, machine learning, financial technology, security, and storage.

Apply Now

WWW.XILINX.COM/ACCELERATOR-PROGRAM

More Information Available on Xilinx.com



Xilinx.com
Product Brief
Product Selection Guide
Getting Started Guide
Data Sheet
ML Solution Brief
SDAccel Solution Brief
ABR Transcoding Solution Brief
Accelerating DNNs with Alveo White Paper
Applications Directory

Thank you!
Contact email : virajp@xilinx.com

XILINX

Adaptable.
Intelligent.

XILINX