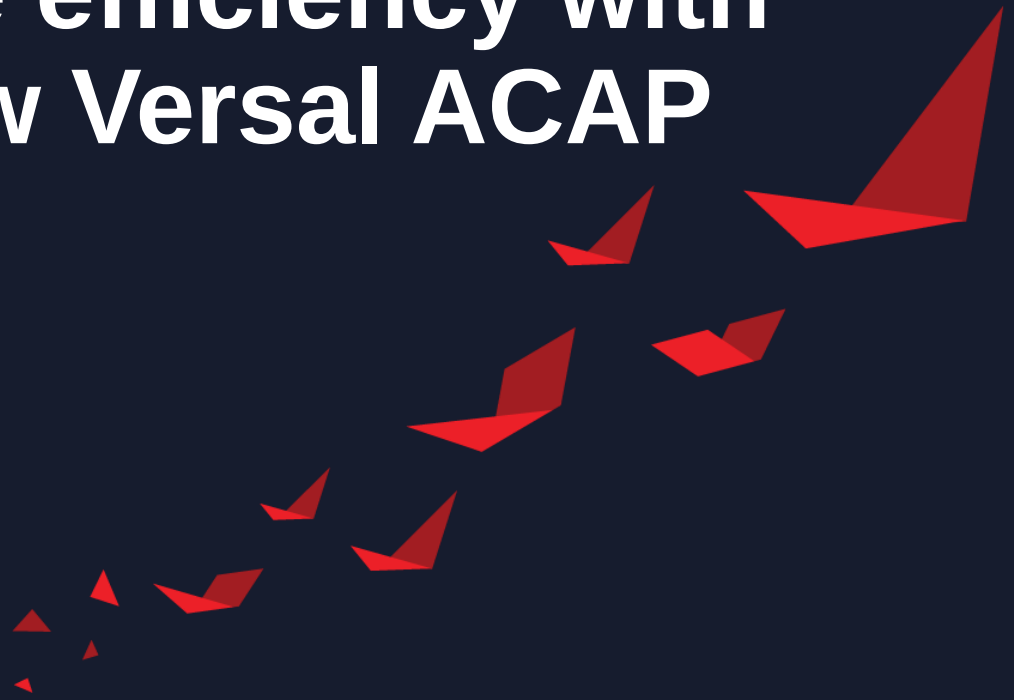
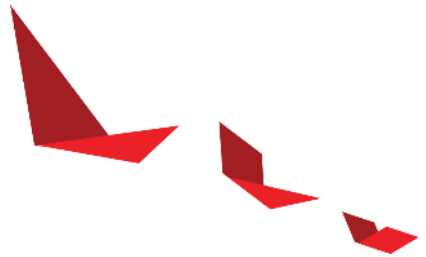
Next generation compute efficiency with Xilinx FPGAs and the new Versal ACAP

Cathal McCabe
Xilinx University Program, Xilinx Ireland
18 February 2020



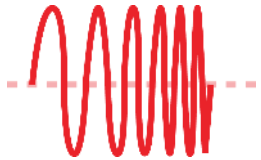
Overview



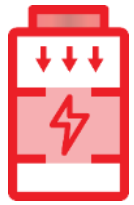
- Requirements for next generation compute systems
- The technology conundrum
- FPGA technology evolution
- Current and next generation Xilinx technologies

What are your requirements for next generation HEP systems?

Performance



Data rates



Power

Cost

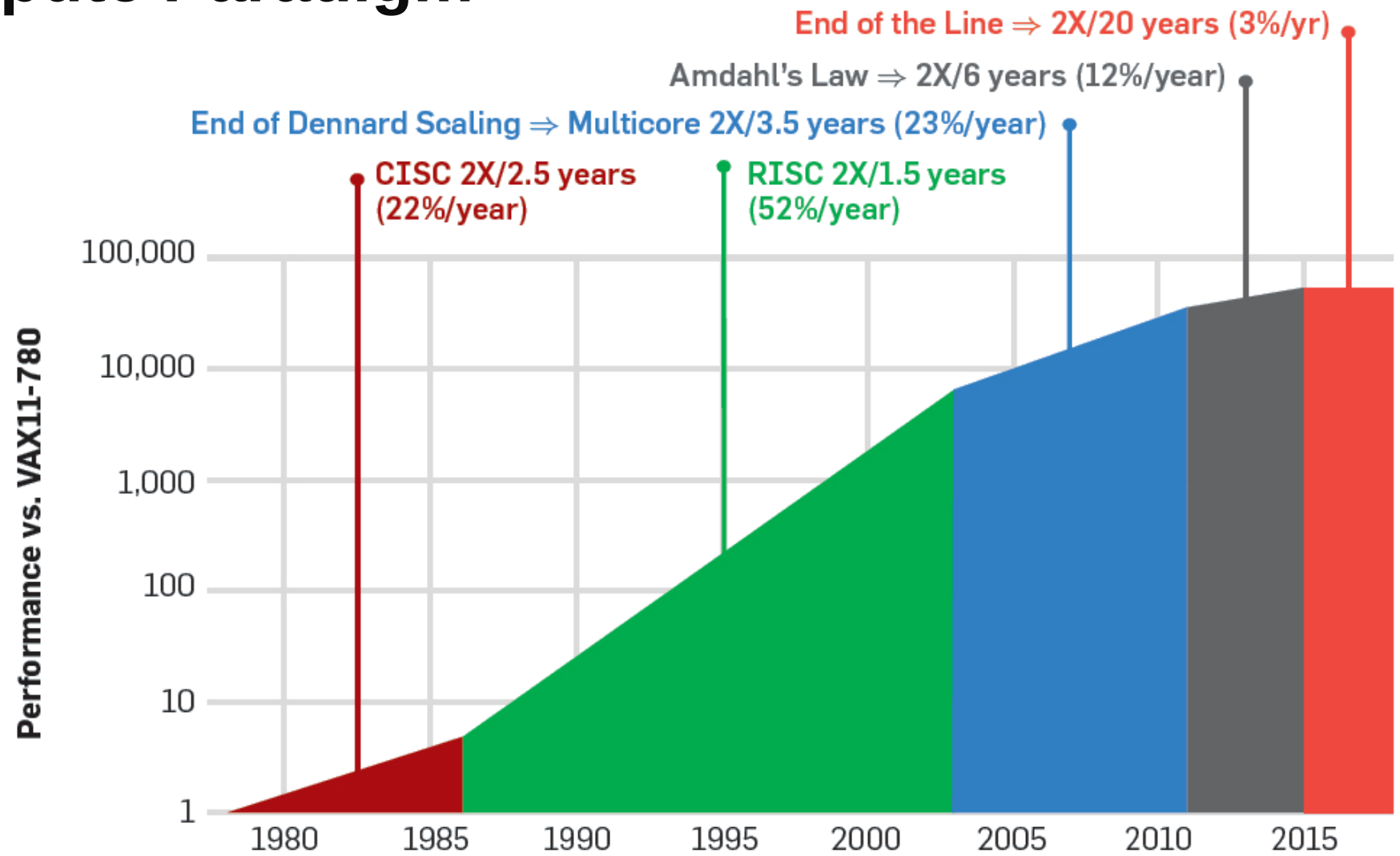
Compute density

Machine Learning

Adaptability

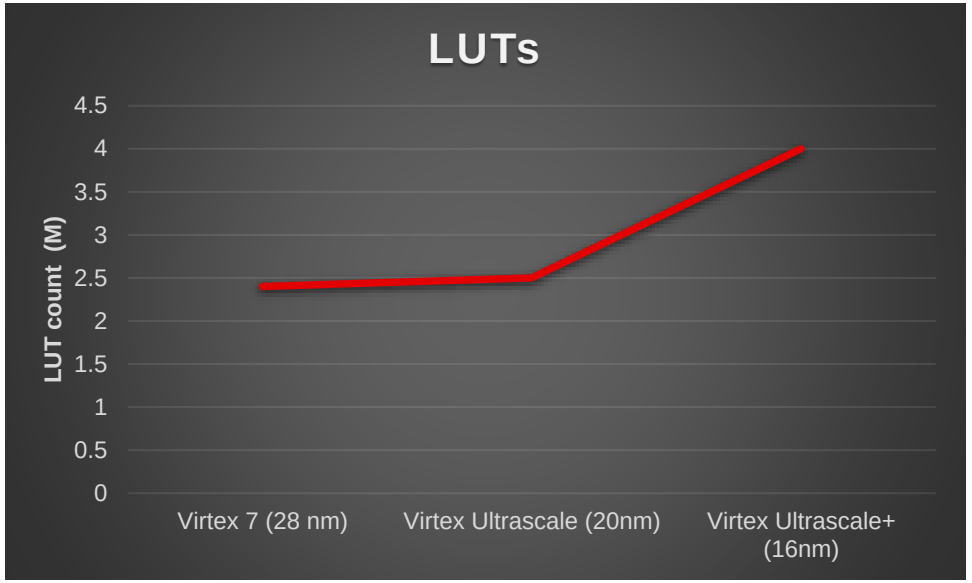
Cloud scalability

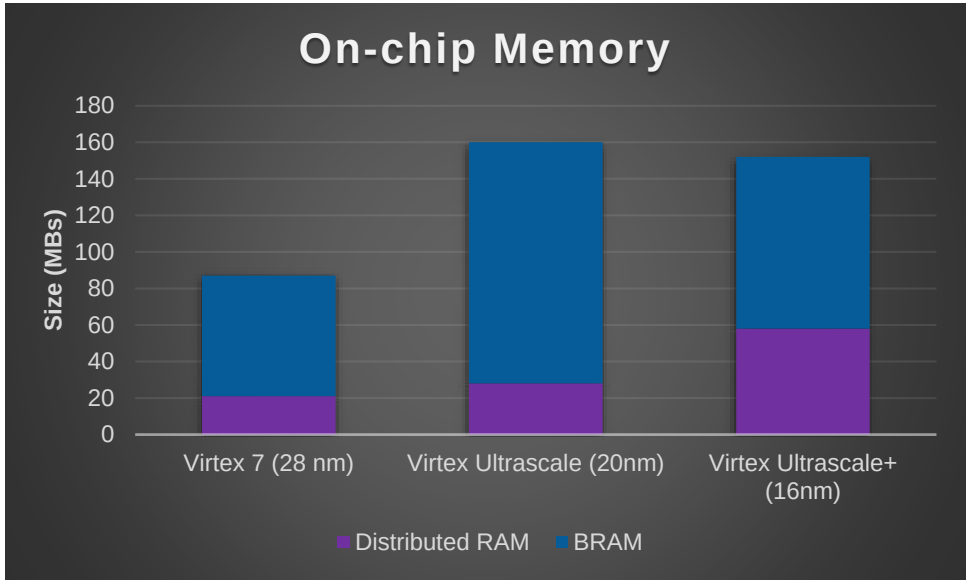
The Technology Conundrum .. And the Need for a New Compute Paradigm

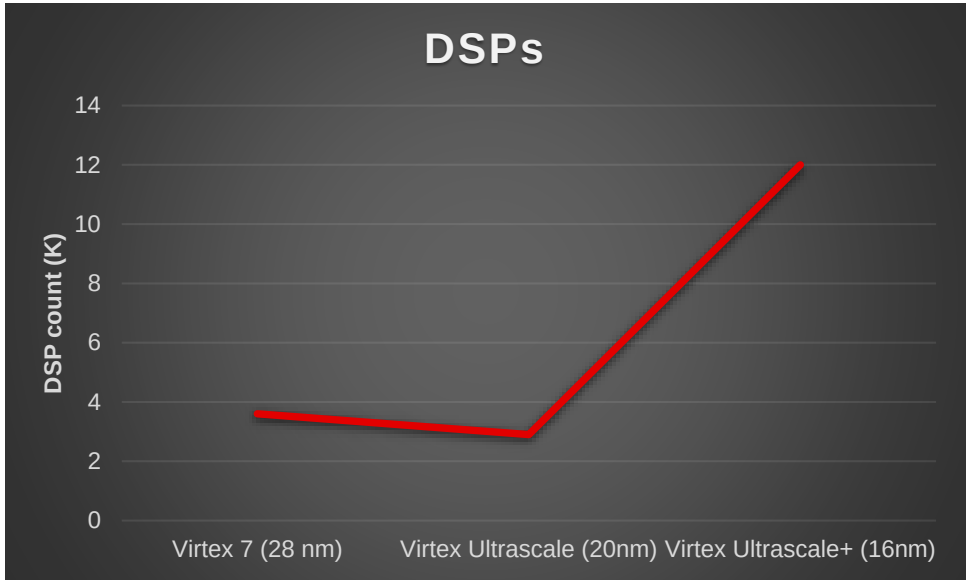


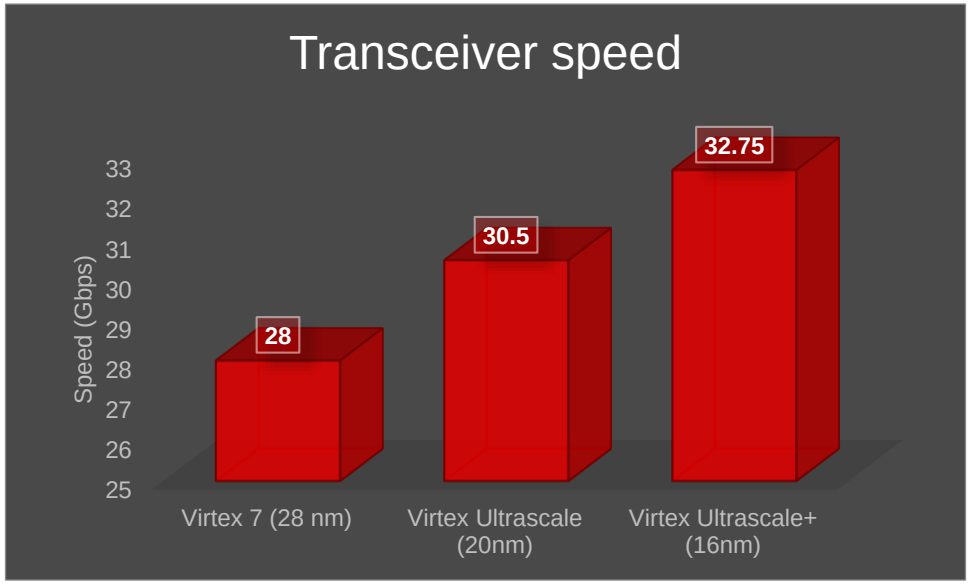
*John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

FPGA scaling

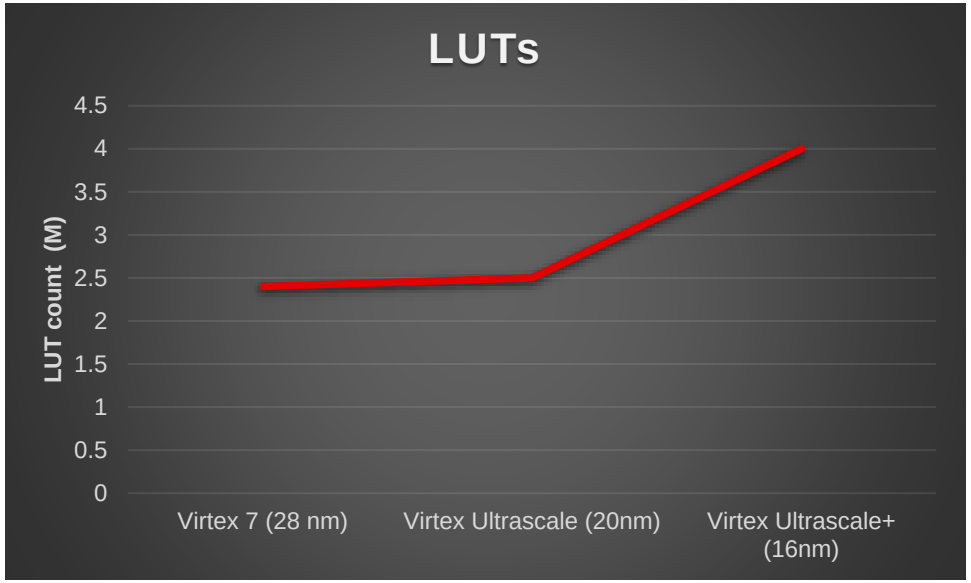


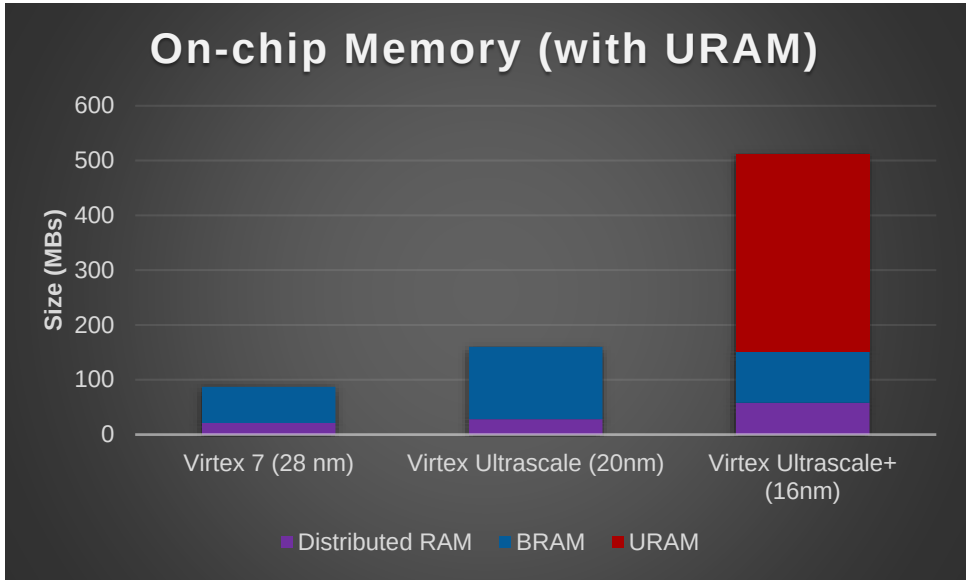


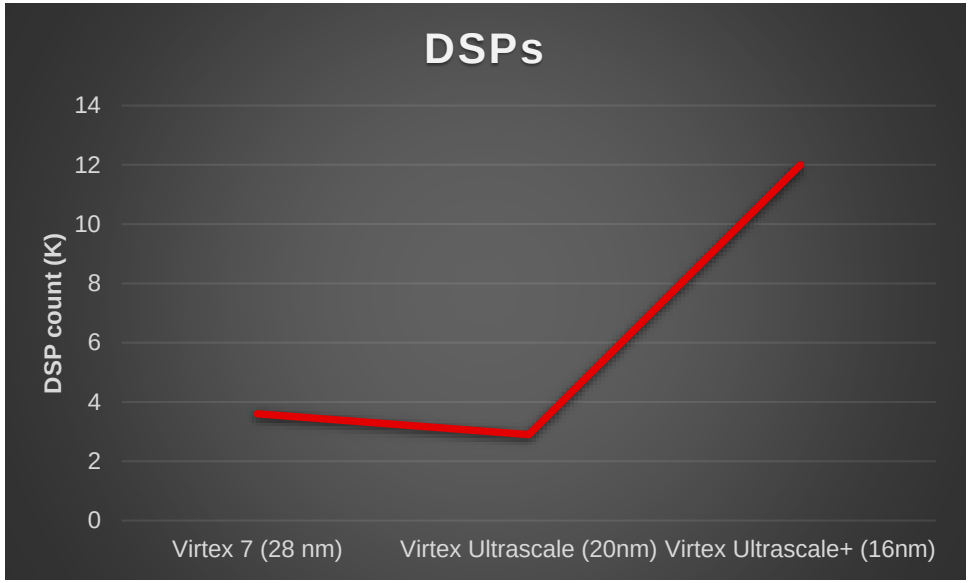


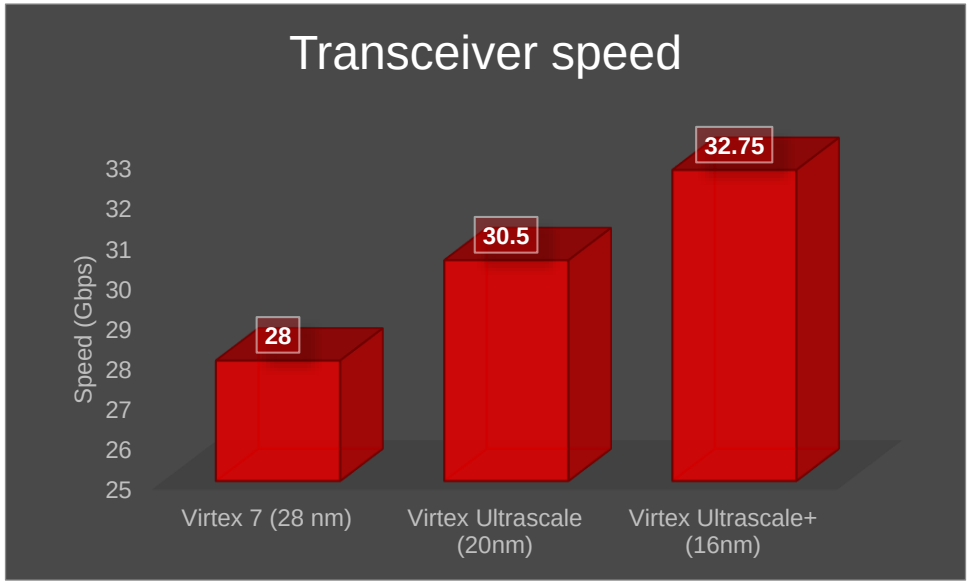


FPGA scaling - URAM example

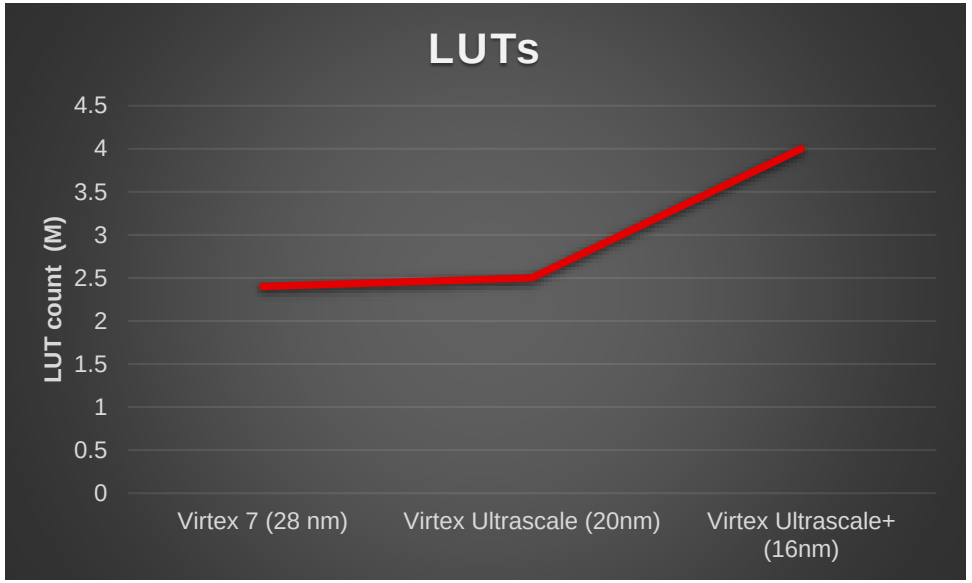


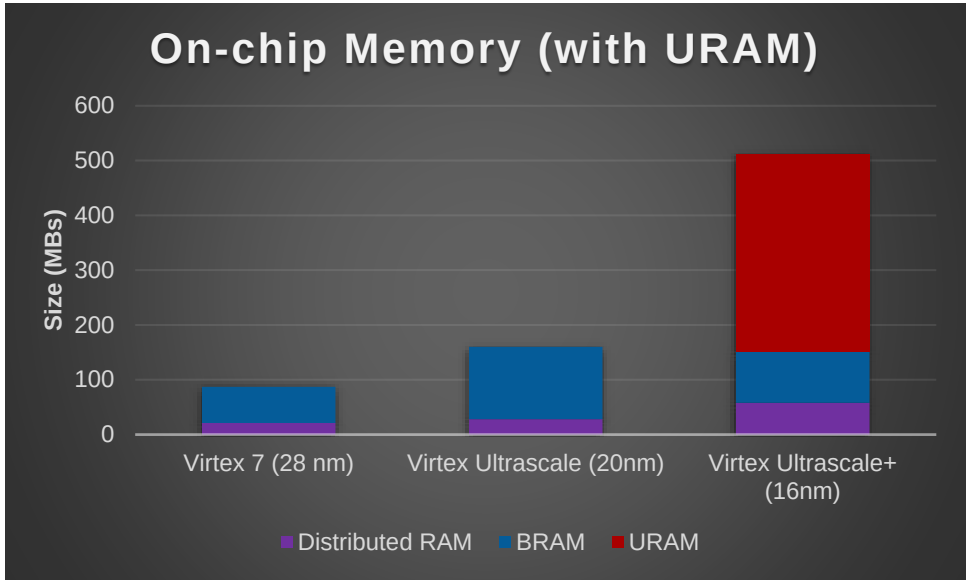


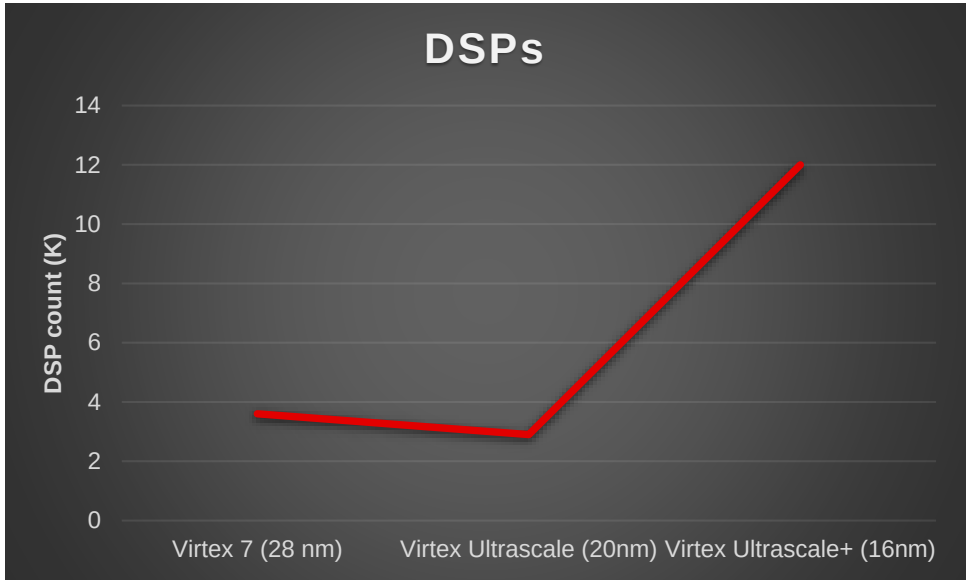


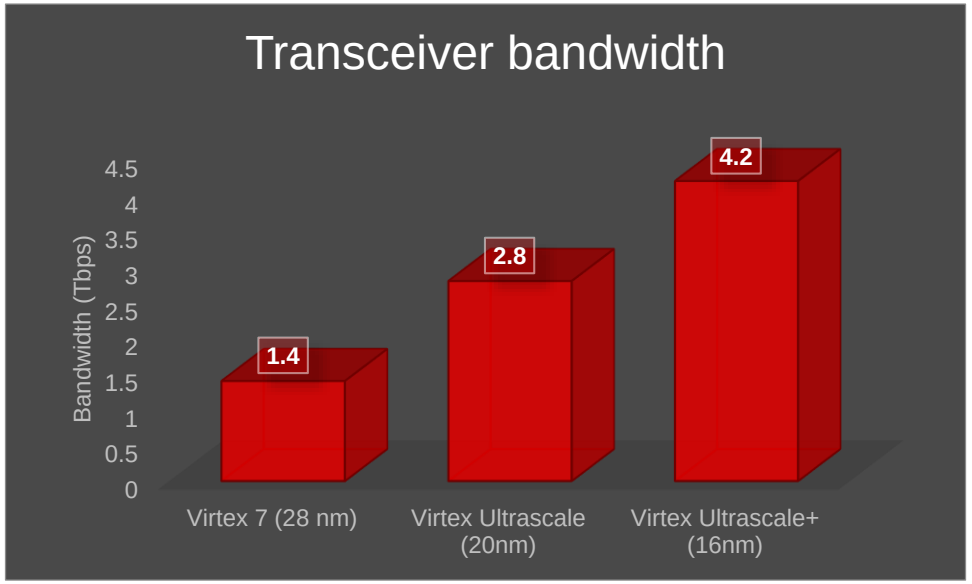


FPGA scaling – transceivers example



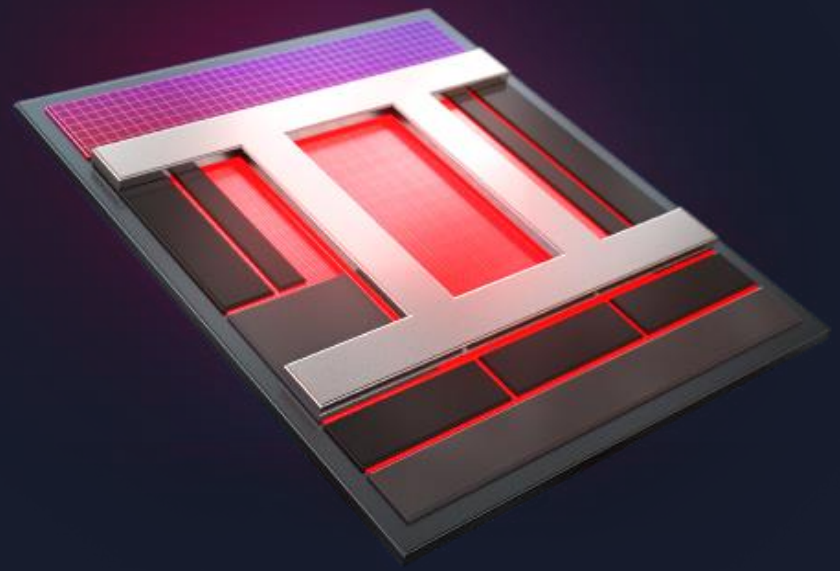


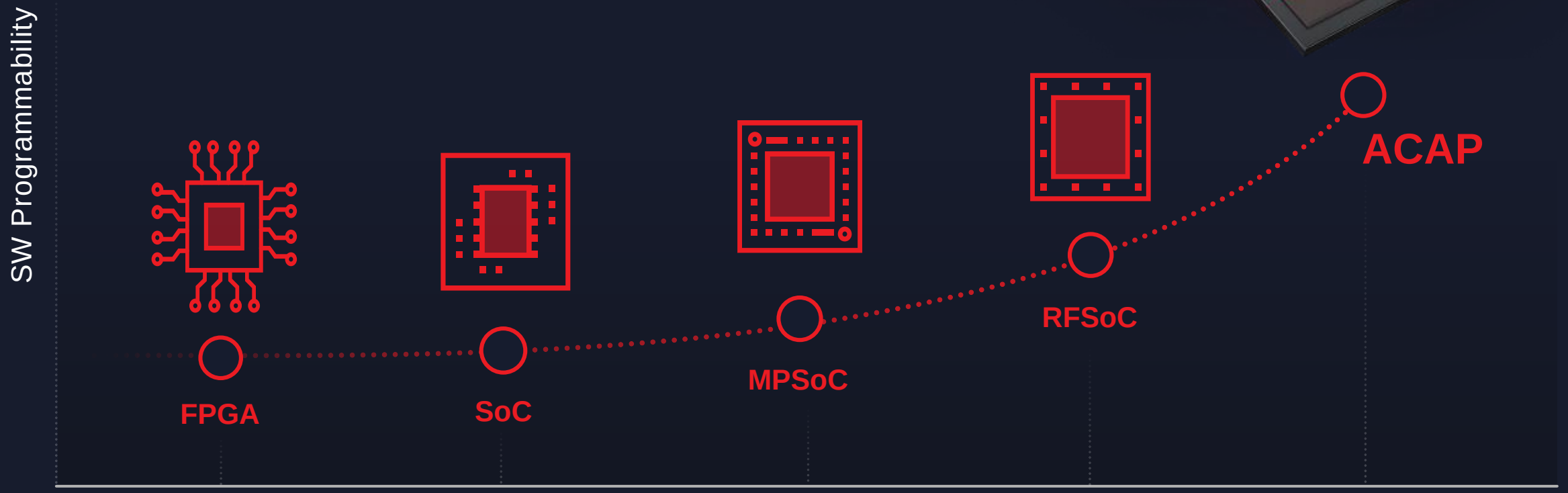




Xilinx Transformation

From Devices to Platforms





XILINX.

Xilinx technologies

XILINX ALVEO™

XILINX VITIS™

XILINX VERSAL™

XILINX

ALVEO™

FAST

Built for high throughput, ultra-low latency

Accelerate compute, networking, storage

ADAPTABLE

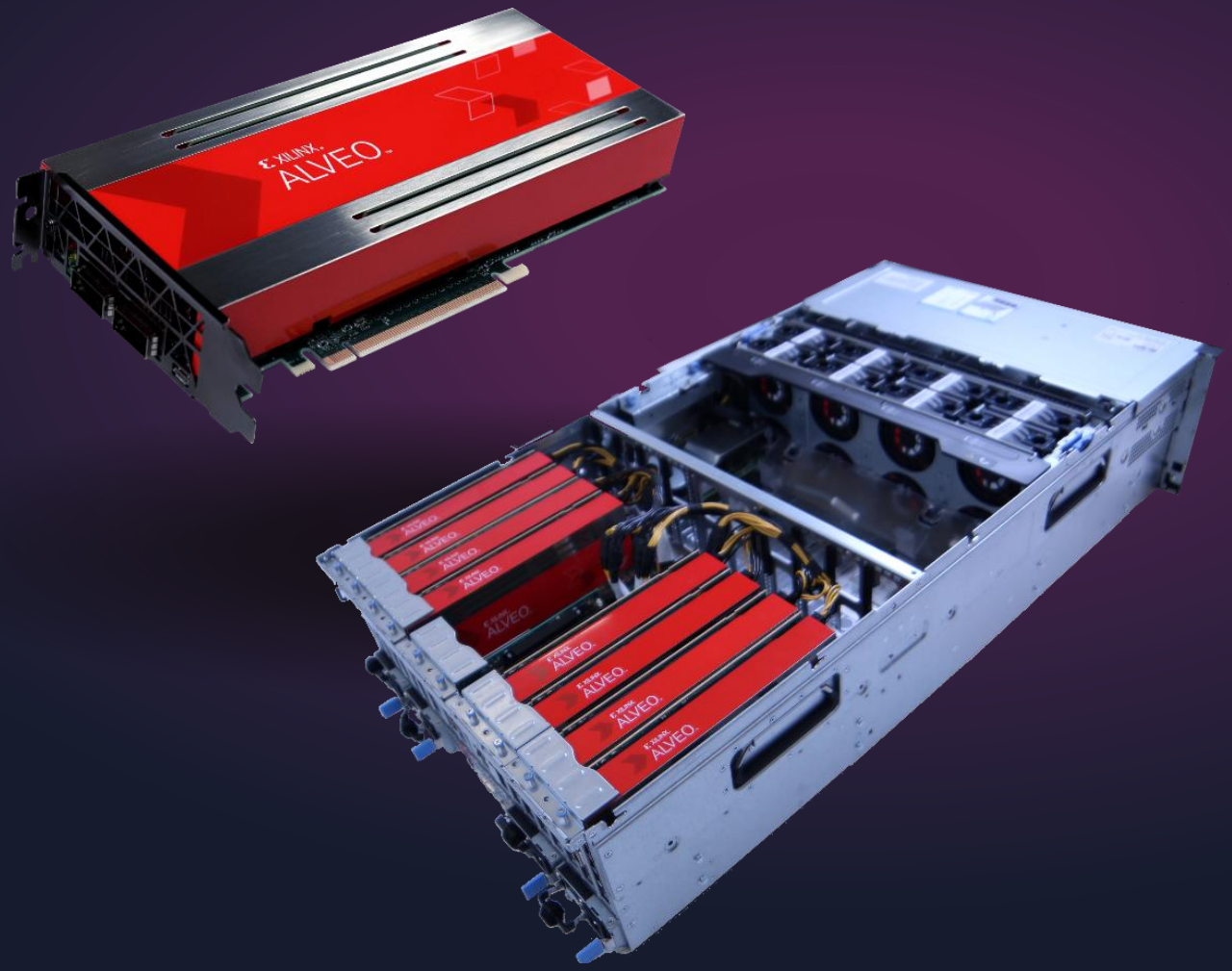
Deploy optimized domain-specific architectures

Adapt to changing algorithms

ACCESSIBLE

Deploy in the cloud or on-premises

Rich set of accelerated Applications



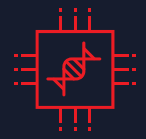
Data Center and AI Accelerator Cards

90x
Database Search & Analytics

89x
Financial Computing

20x
Machine Learning

12x
Video Processing



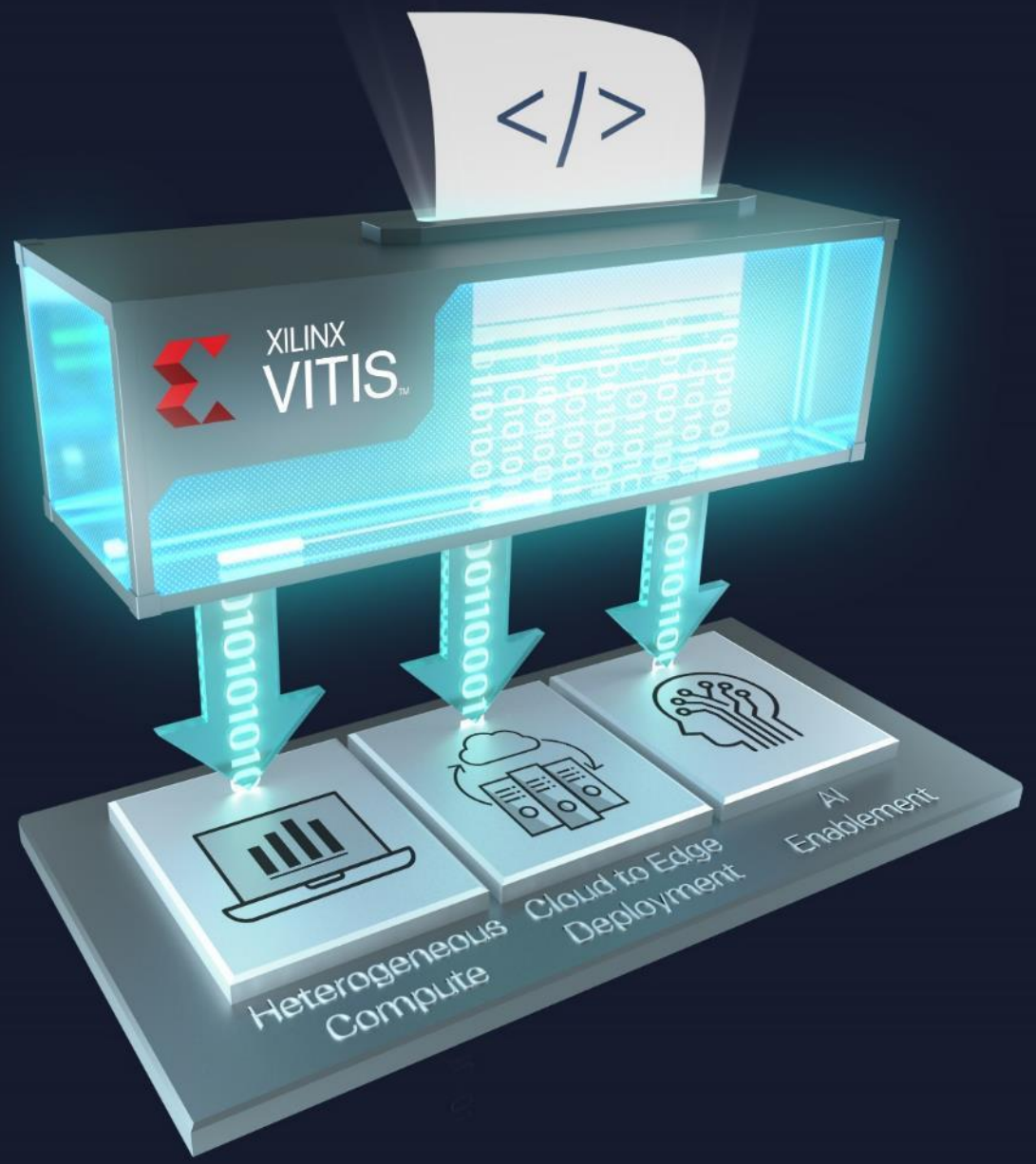
10x
HPC & Life Sciences

XILINX.

XILINX VITIS™

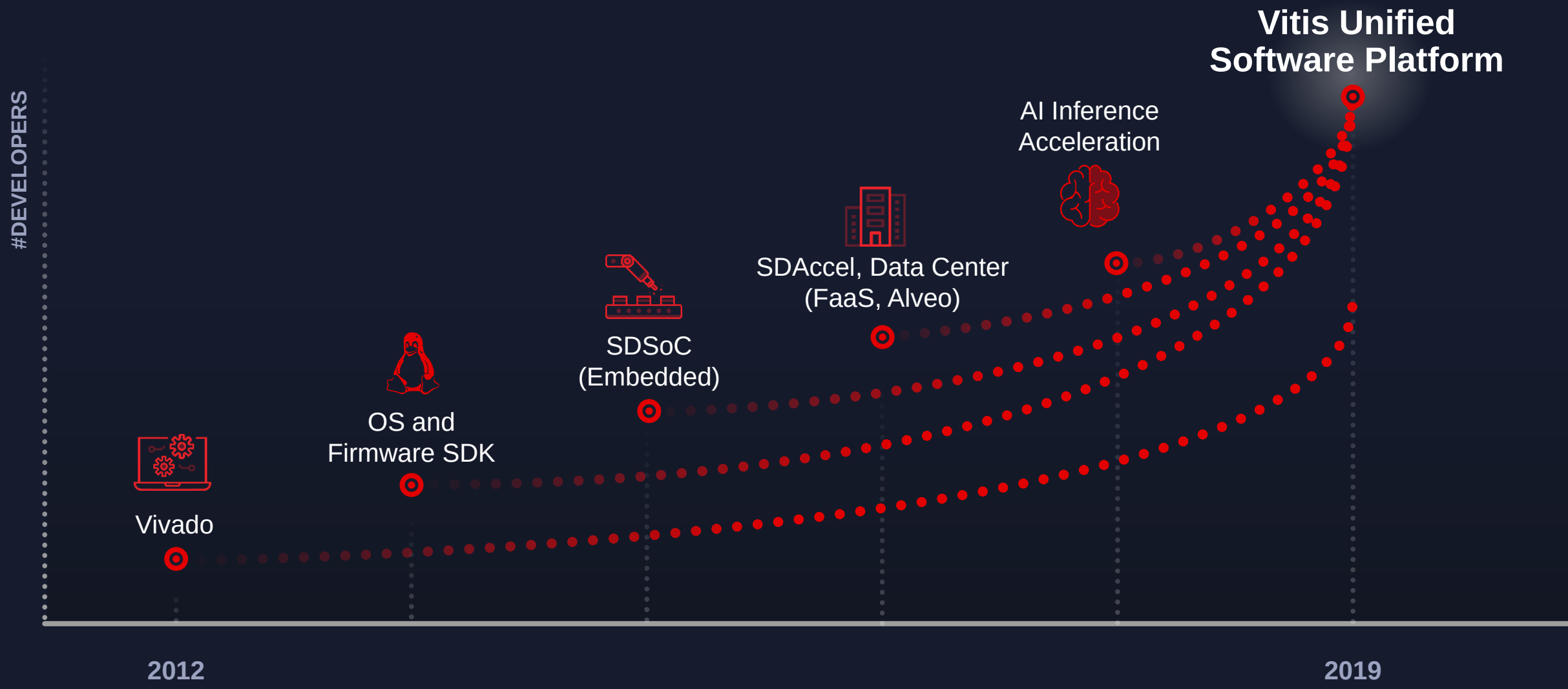
Unified Software Platform

- Unified methodology edge to cloud
- Focus on platform and acceleration
- Available for free



*Open source Xilinx Runtime library (XRT), Accelerated libraries, AI Models

Platform Transformation



XILINX

Vitis: Unified Software Platform

Domain-specific development environment

Vitis accelerated libraries

Vitis core development kit

TensorFlow

Vitis AI

FFmpeg

Vitis Video

Partners
Genomics,
Data Analytics,
And more

OpenCV Library

BLAS Library

Finance Library

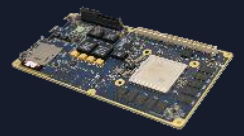
Compilers

Analyzers

Debuggers

Xilinx runtime libraries (XRT)

Vitis target platform



XILINX

Build: Extensive, Open Source Libraries

GitHub

Domain-Specific Libraries

Vision & Image

Finance

Data Analytics & Database

Data Compression

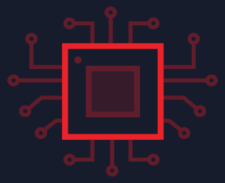
Data Security

Common Libraries

Math

Linear Algebra

Statistics



DSP

Data Management

400+ functions across multiple libraries for performance-optimized out-of-the-box acceleration

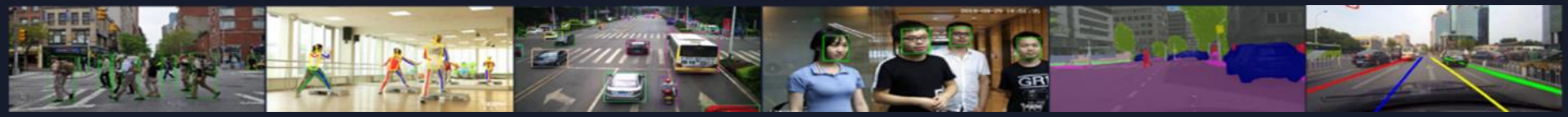
XILINX

Vitis AI: Deep Learning Acceleration

Frameworks

TensorFlow Caffe PyTorch

Vitis AI models



Vitis AI development kit

AI Optimizer | AI Quantizer | AI Compiler | AI Profiler | AI Library

Xilinx runtime library (XRT)

Deep Learning Processing Unit

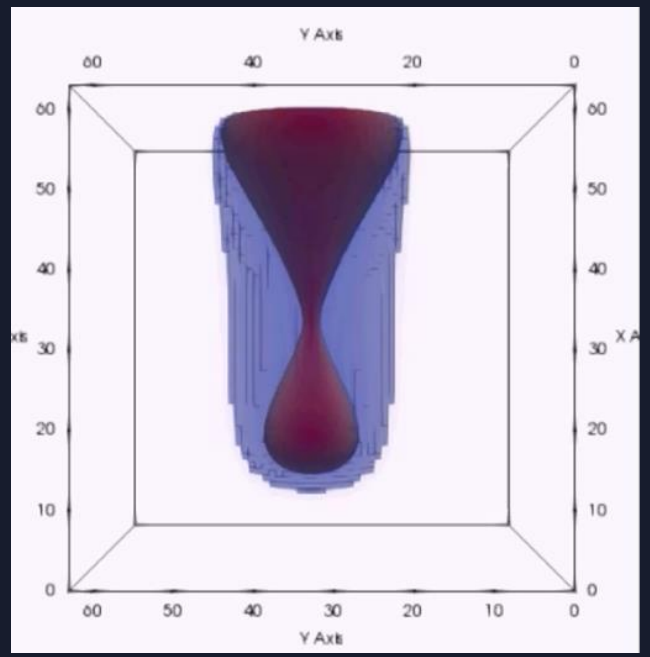
CNN DPU | LSTM DPU | MLP DPU

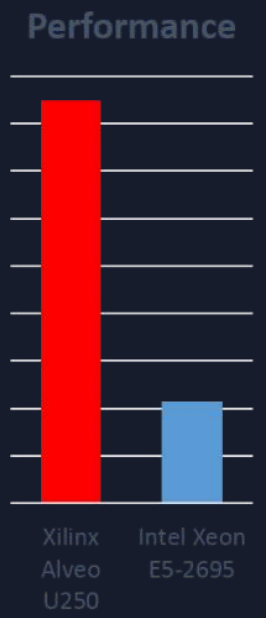
XILINX

Example performance

Computational Fluid Dynamics

ALVEO Accelerated CFD Kernels





byte LAKE

Faster Time to insight, Fewer Nodes

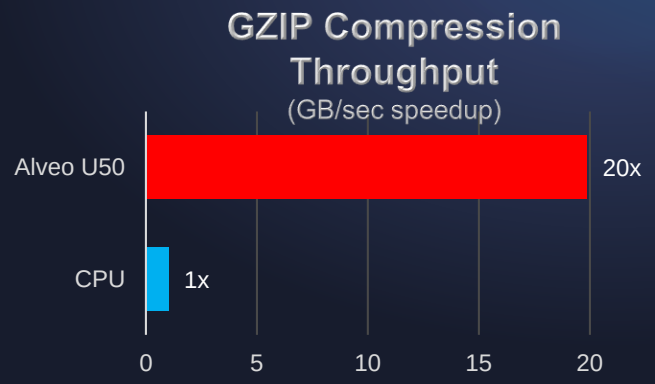
- 4x Faster simulation time
- 80% lower energy consumption
- 6x better performance per Watt

XILINX

Computational Storage

Line-rate Data Compression Acceleration

Compression, decompression, erasure coding, encryption all accelerated on one platform



Intel Skylake-SP 6152 @2.10GHz 22-core CPU (Ubuntu 16.04), GB/s compression per CPU core = .0229. Alveo U50 = 10GB/s

Source: Xilinx Analysis

Computational Storage

Line-rate Data Compression Acceleration

20x Throughput Per Node

2x Less Nodes

40% Lower Total Cost

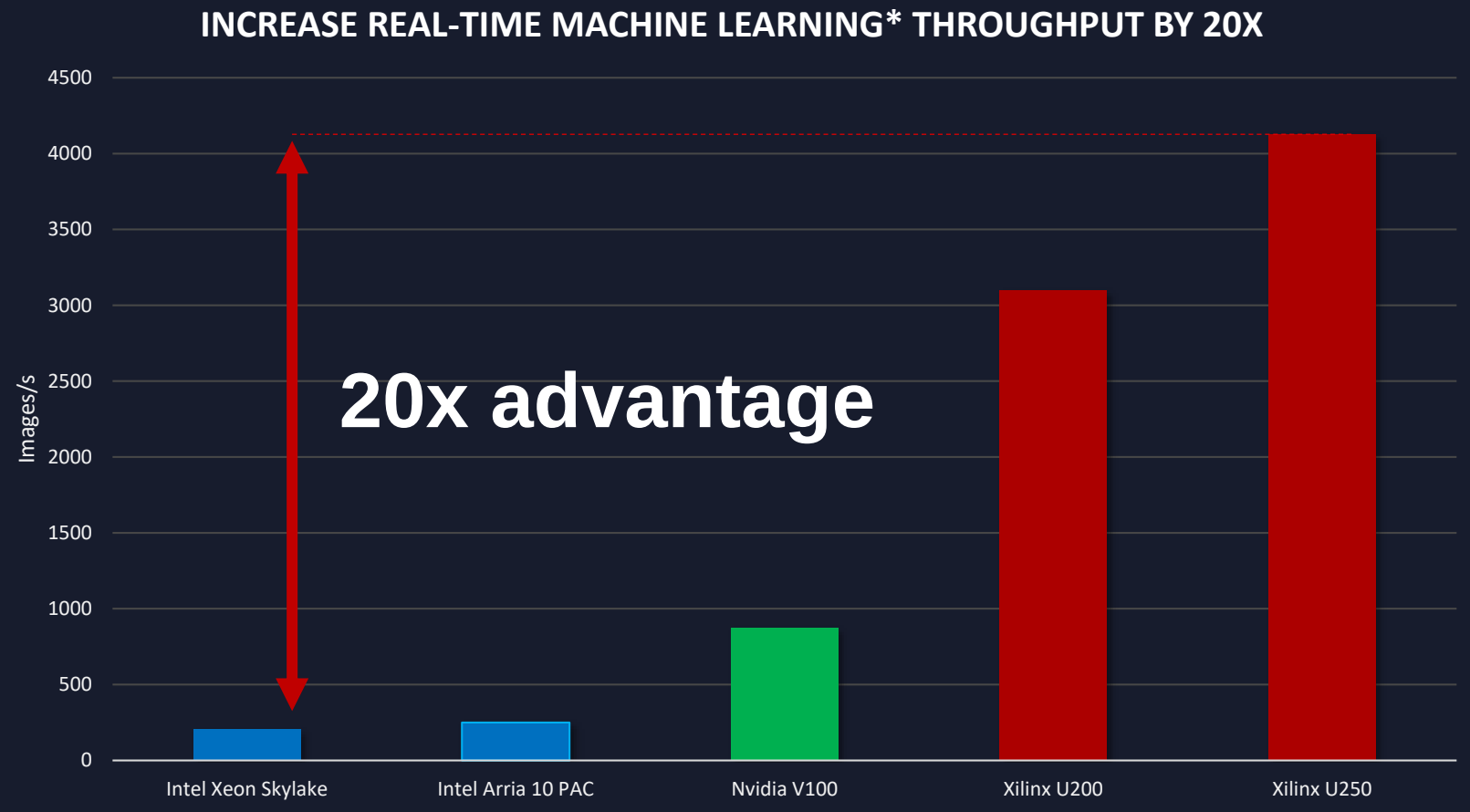
Alveo U50 Acceleration

2x Dual CPU Servers
192TB SSDs, 1GB/sec Per Node Compression Throughput

Alveo Server with 2x Alveo U50
96TB SSDs (192TB effective), 20GB/sec Per Node Compression Throughput

Intel Skylake-SP 6152 @2.10GHz CPU (Ubuntu 16.04), GB/s compression per CPU core = .0229. Alveo U50 = 10GB/s, Assume 2:1 compression

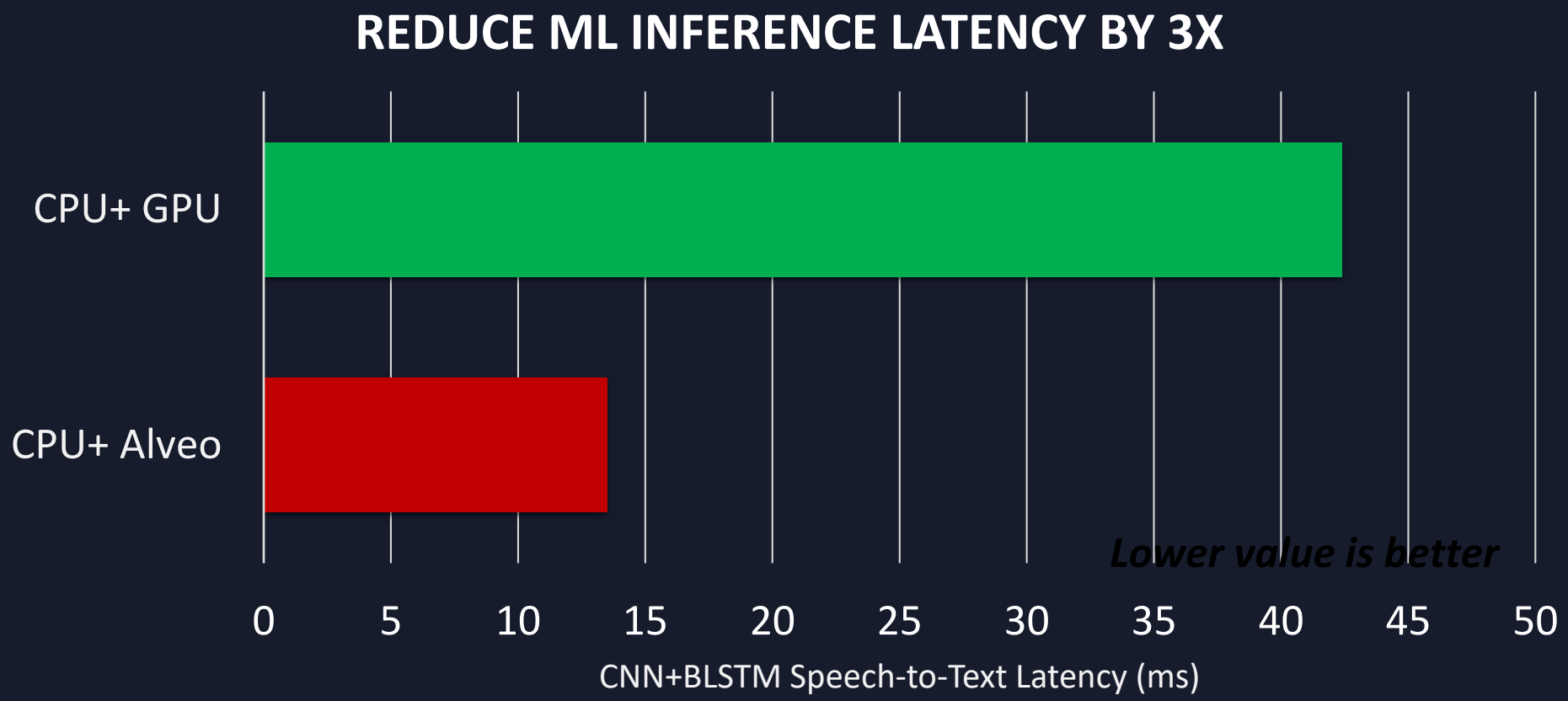
Advantages in Machine Learning Inference



* Source Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper

XILINX

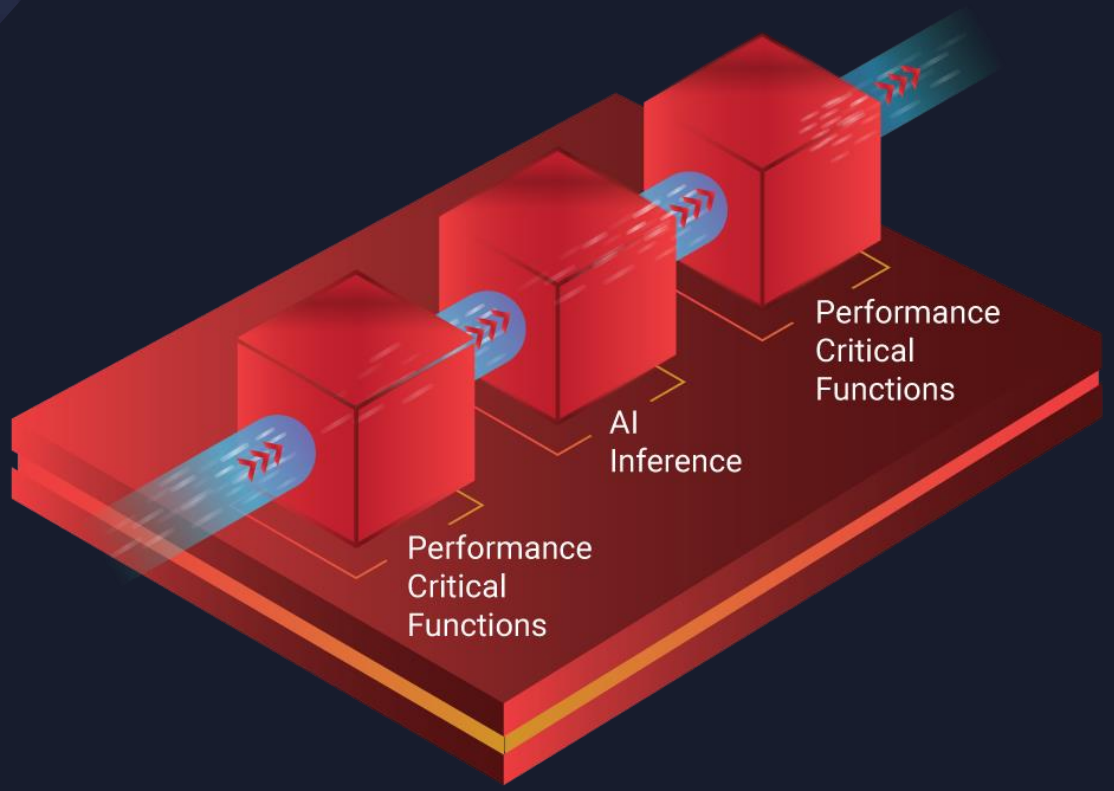
Advantages in Latency



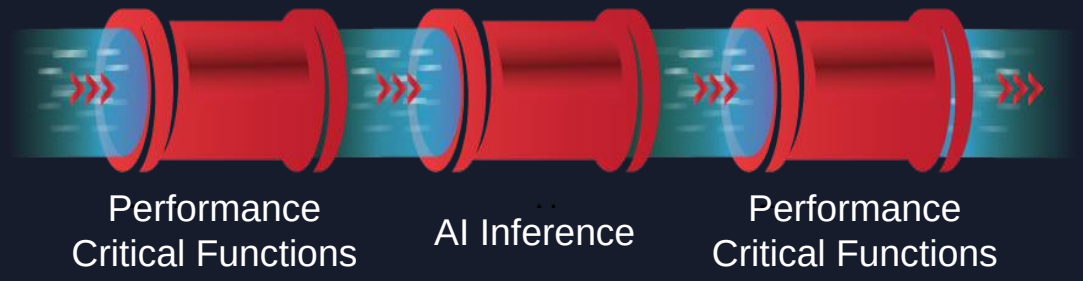
Alveo Provides Massive Parallel Compute with Lowest Latency vs GPUs

XILINX

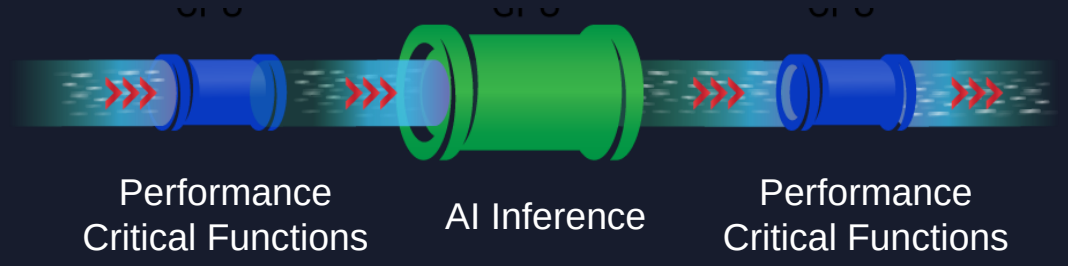
Whole Application Acceleration



Xilinx – Matched Throughput



Other solutions – Mismatched Throughput

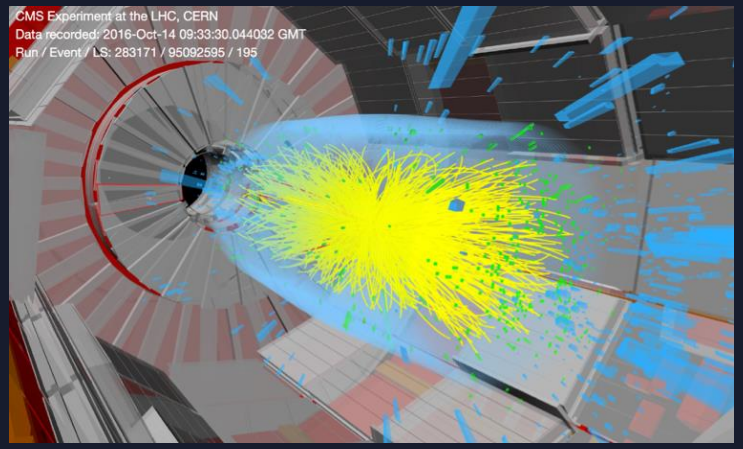


XILINX

AI Accelerated Dark Matter Search (CERN)

Real-time ML Inference + Sensor pre-processing

100ns Inference Latency on 150 Terabytes/Second Data Rates



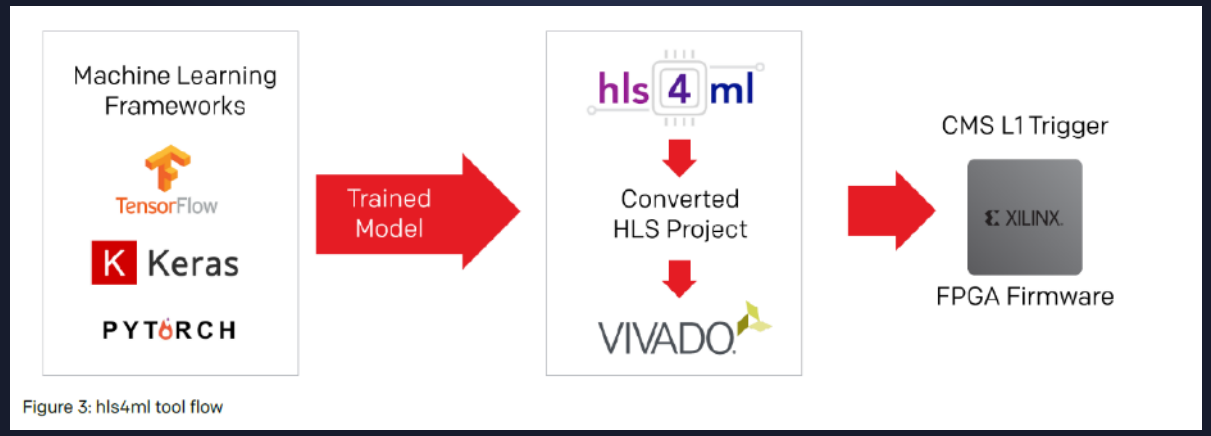


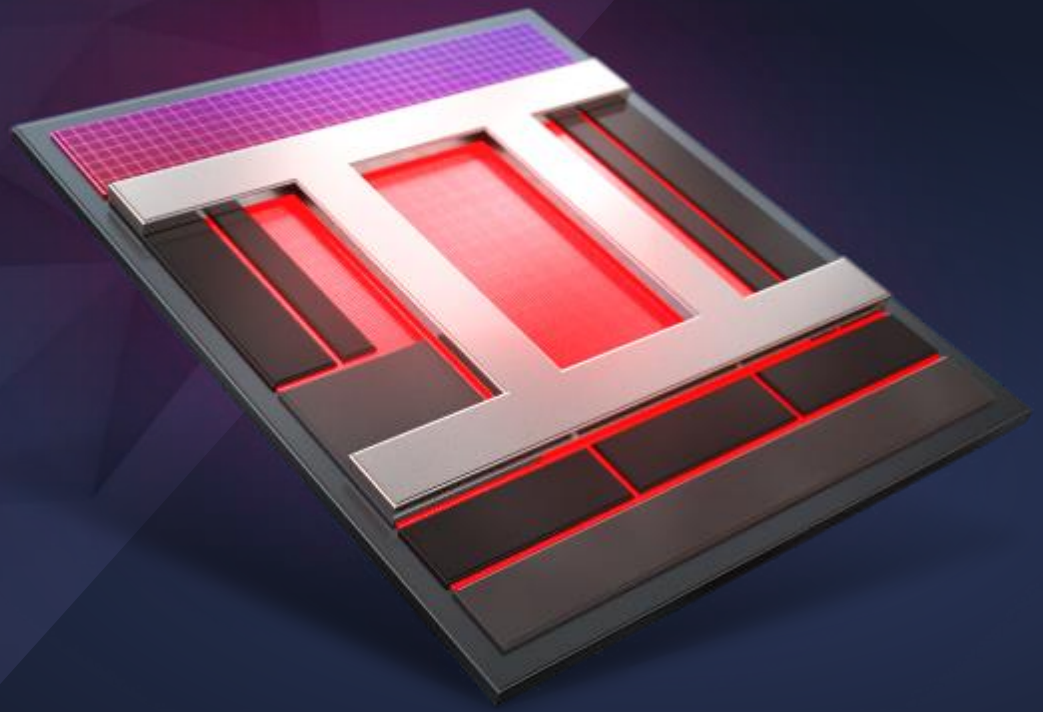
Figure 3: hls4ml tool flow

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

XILINX

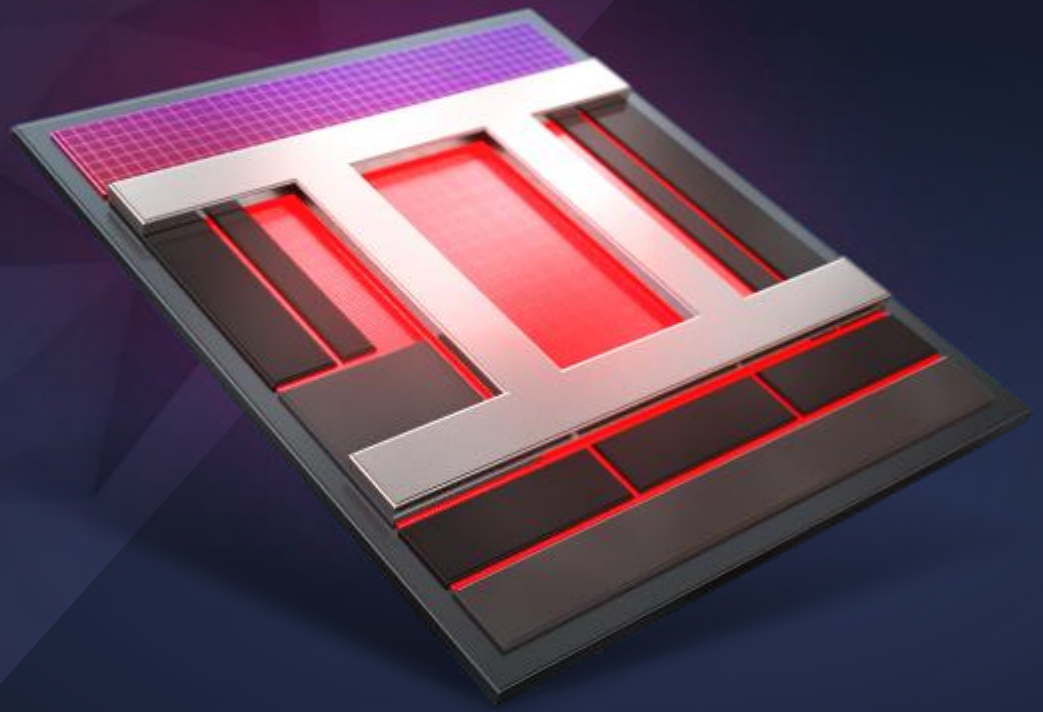
Introducing the Versal ACAP

XILINX.



ACAP

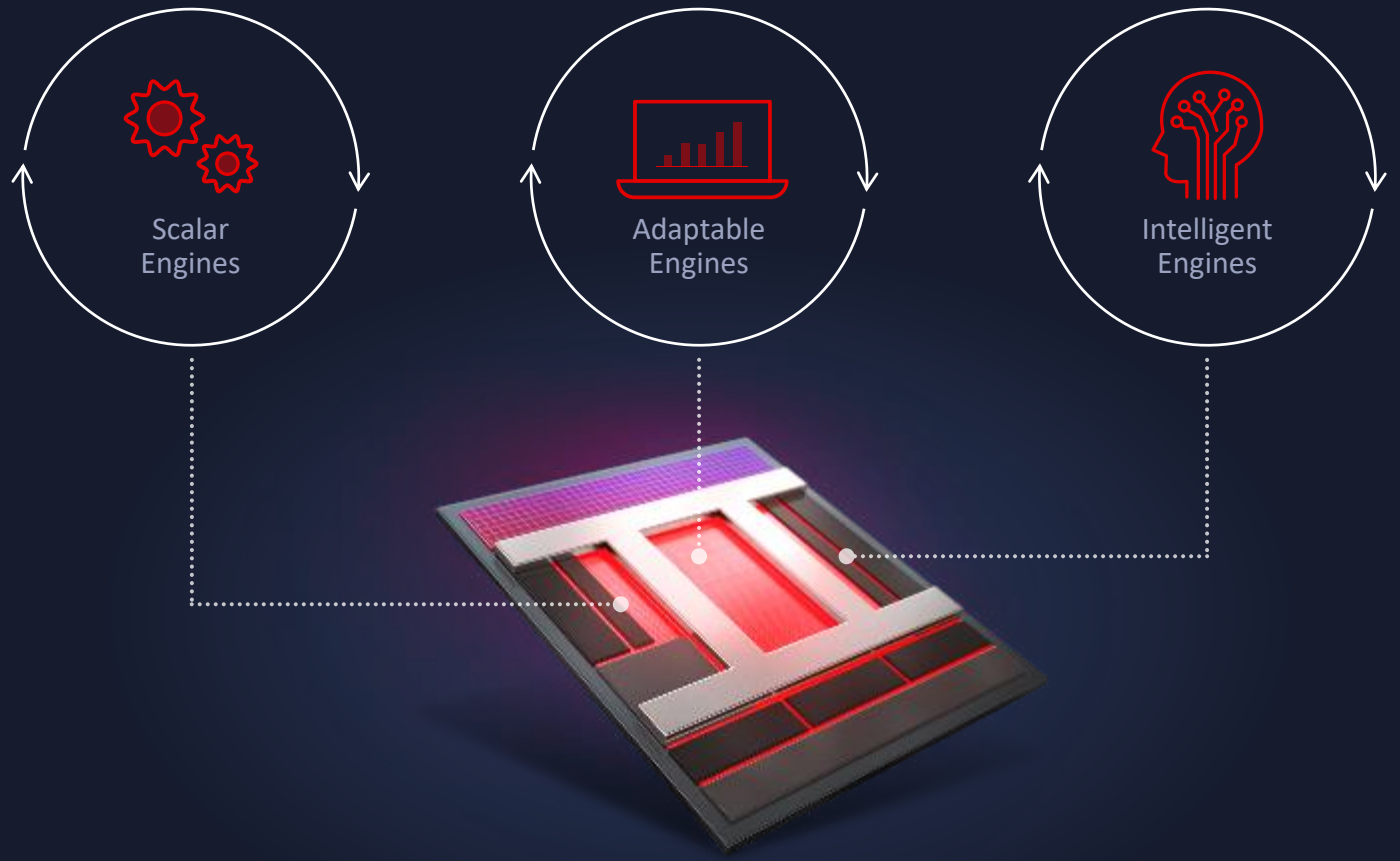
XILINX.



Adaptive Compute Acceleration Platform

XILINX.

Compute Acceleration



XILINX.

XILINX®

VERSAL™

The Industry's First ACAP

Heterogeneous Acceleration

For Any Application

For Any Developer

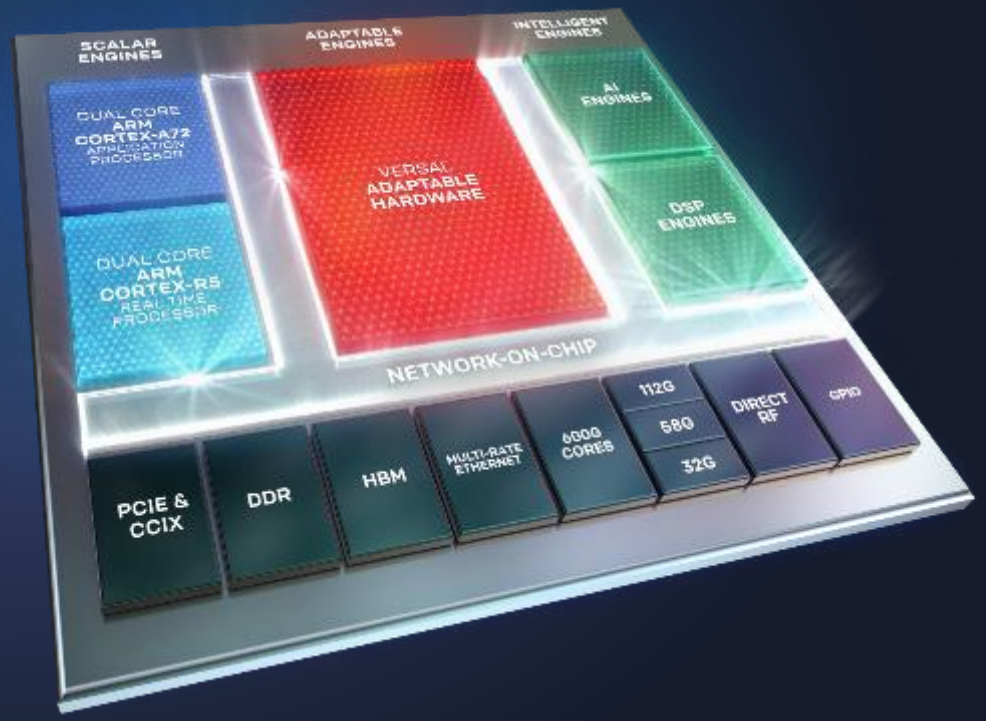
7nm
FinFET

tsmc

XILINX

Versal ACAP Technology Tour

- Scalar Processing Engines
- Adaptable Hardware Engines
- Intelligent Engines
  SW Programmable, HW Adaptable
- Breakout Integration of Advanced Protocol Engines

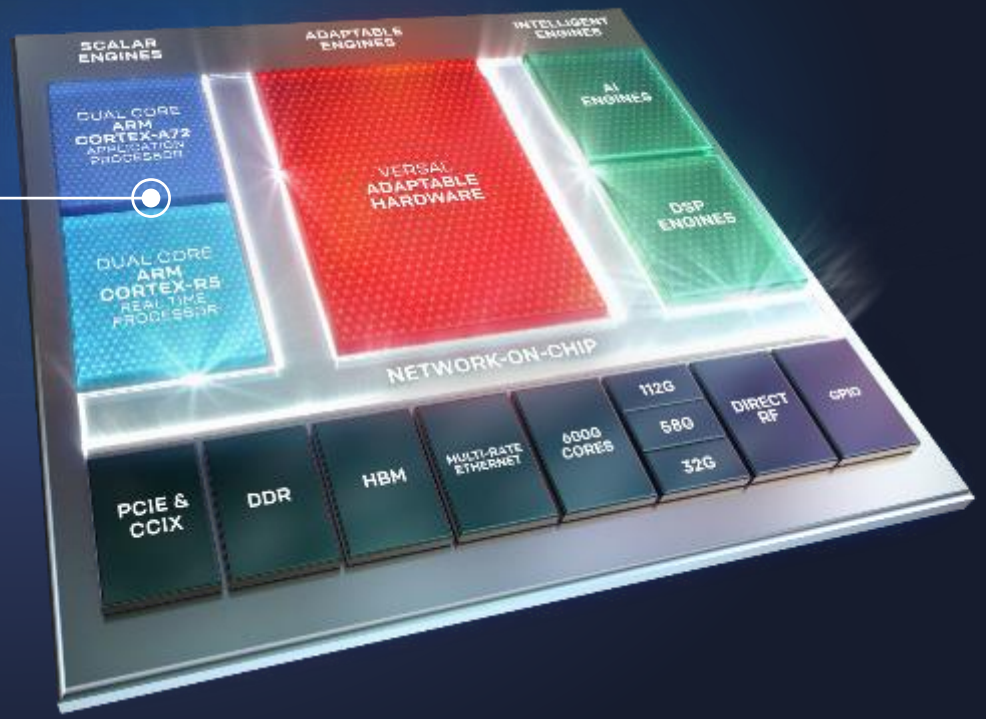


Scalar Processing Engines

Arm Cortex-A72
Application Processor

Arm Cortex-R5
Real-Time Processor

Platform Management Controller



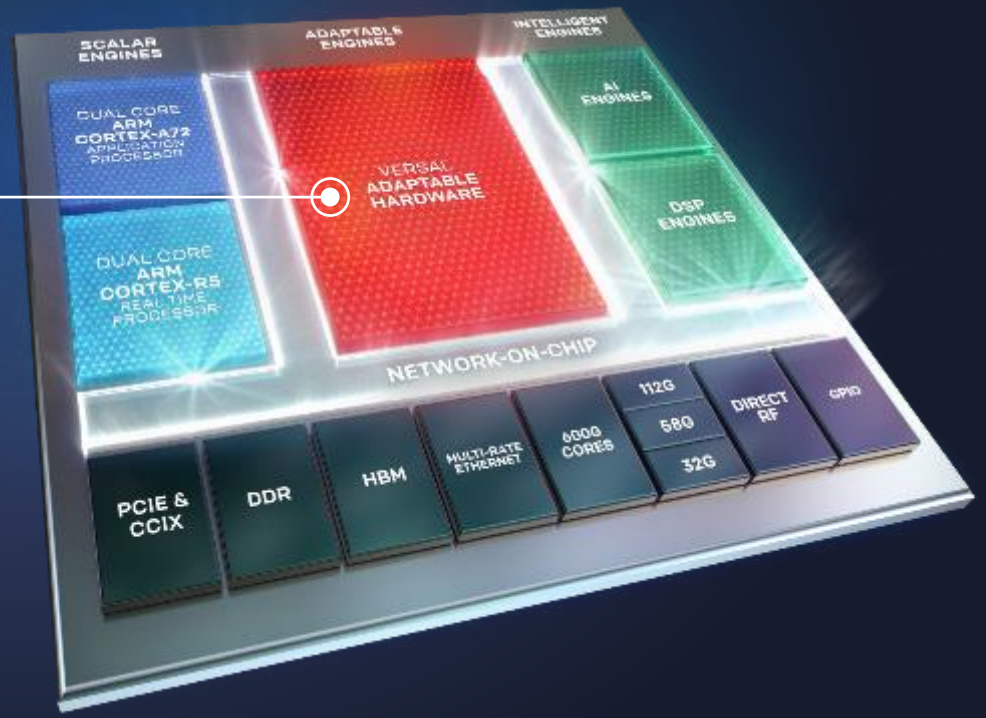
XILINX.

Adaptable Hardware Engines

Re-architected foundational HW fabric for greater compute density

Enables custom memory hierarchy

8X Faster Dynamic Reconfiguration (“on-the-fly”)



XILINX.

Intelligent Engines

DSP Engines

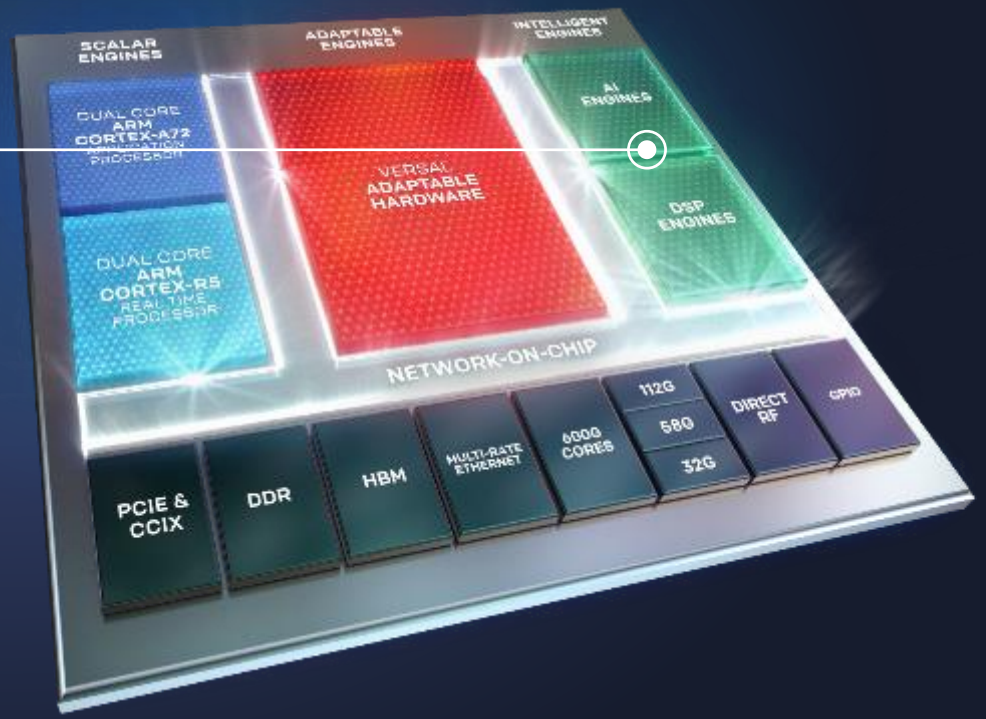
High-precision floating point & low latency

Granular control for customized datapaths

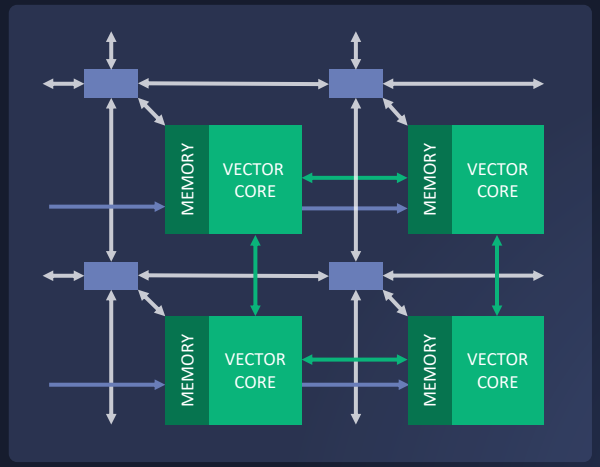
AI Engines

High throughput, low latency, and power efficient

Ideal for AI inference and advanced signal processing

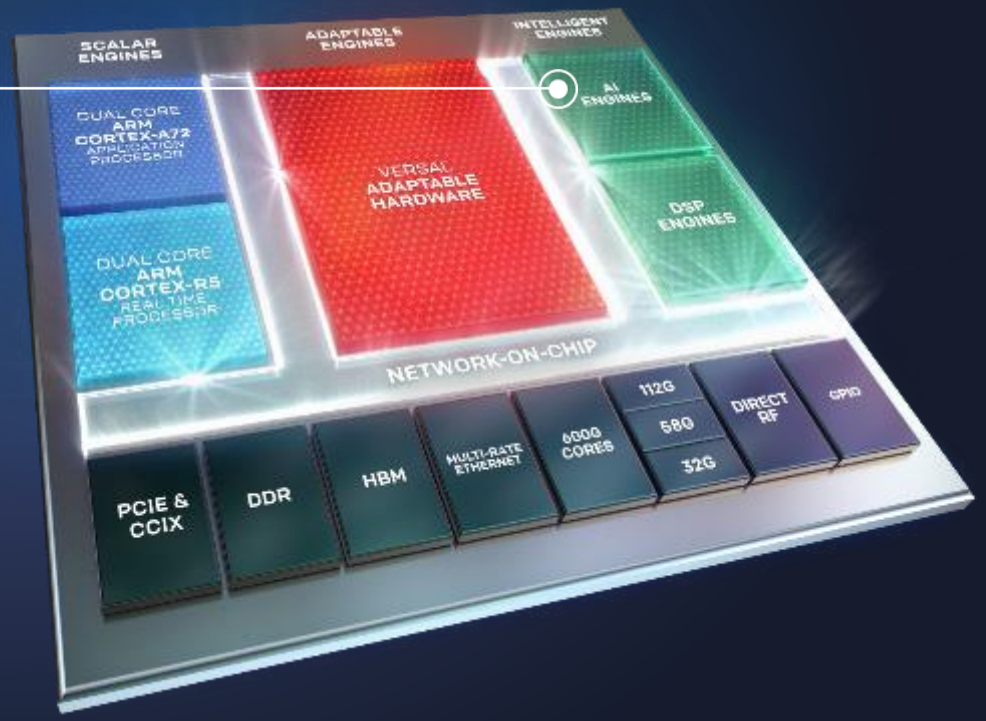


XILINX.

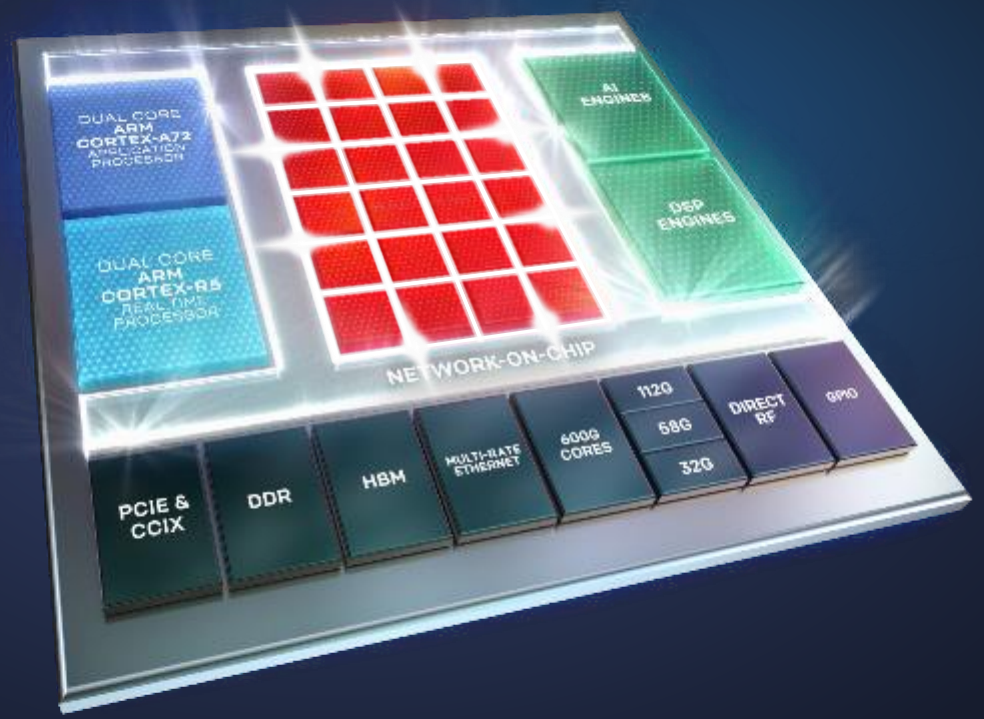


AI Engines

Optimized for AI Inference and
Advanced Signal Processing Workloads



XILINX.



Network-on-Chip (NoC)

Ease of Use

Inherently software programmable
Available at boot, no place-and-route required

High Bandwidth and Low Latency

Multi-terabit/sec throughput
Guaranteed QoS

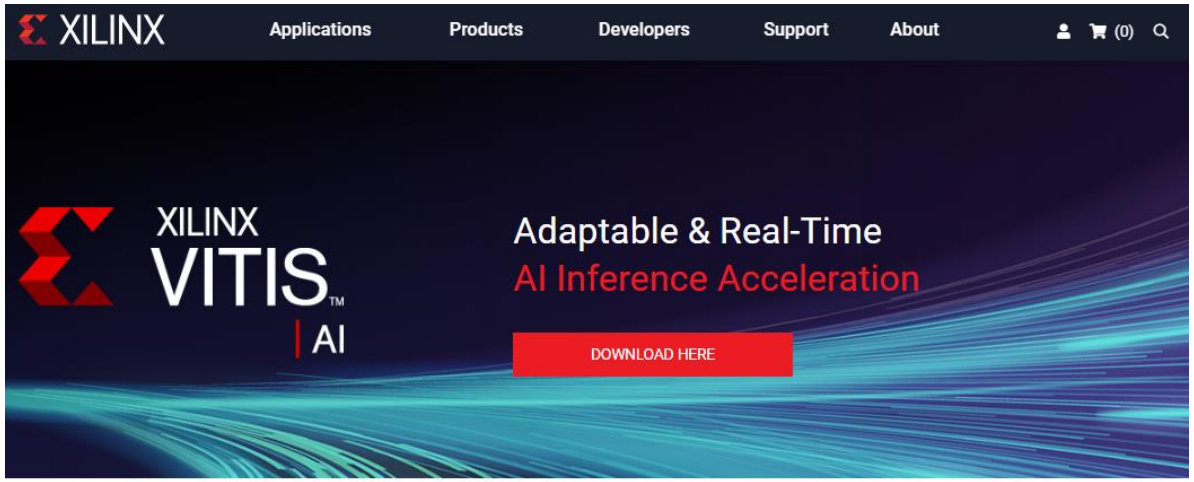
Power Efficiency

8X power efficiency vs. soft implementations
Arbitration across heterogeneous engines

XILINX.

Xilinx University Program

- Donation program
- Training materials
- Request tutorials

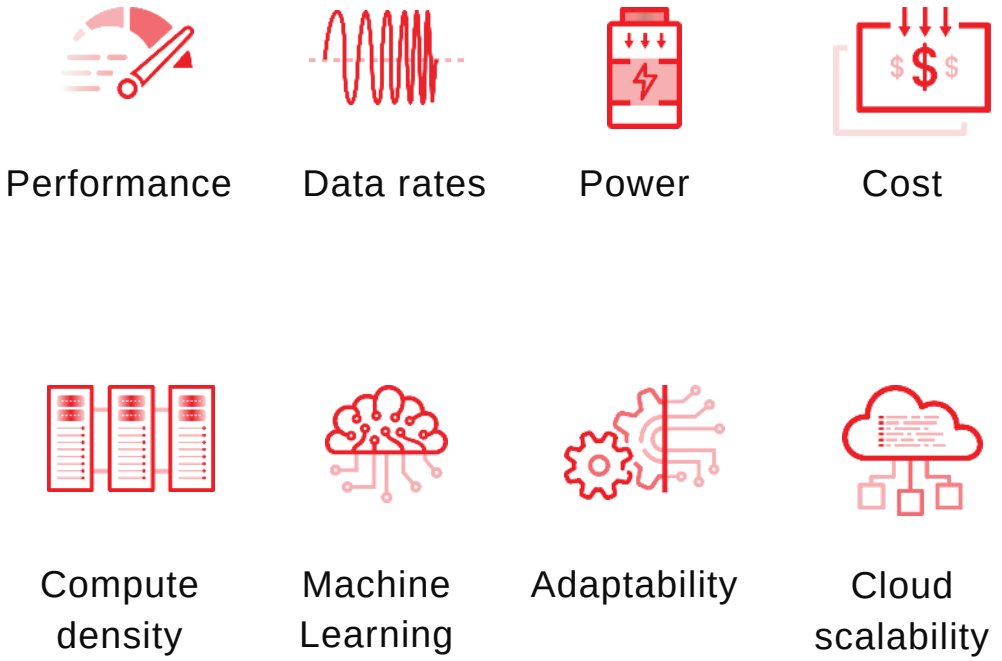


cathal.mccabe@xilinx.com

www.Xilinx.com/university

Next generation compute efficiency with Xilinx FPGAs and the new Versal ACAP

- Try the new Vitis software for platform design free
- Test drive Alveo – production ready accelerator cards
- Next generation Versal ACAP



XILINX

Thank You

Xilinx Mission

Building the Adaptable, Intelligent World