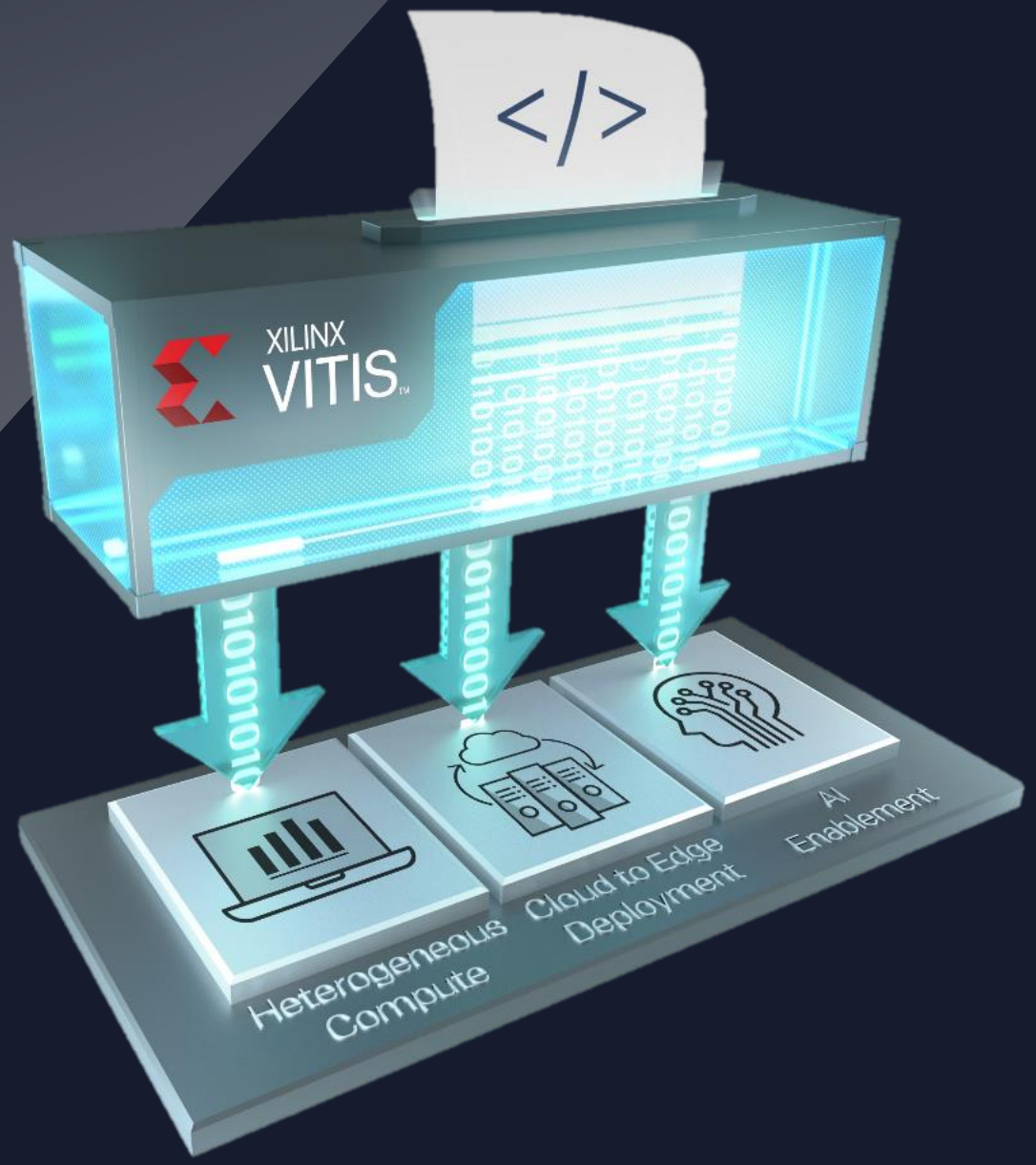
Xilinx – Enabling FPGAs in High Performance Computing

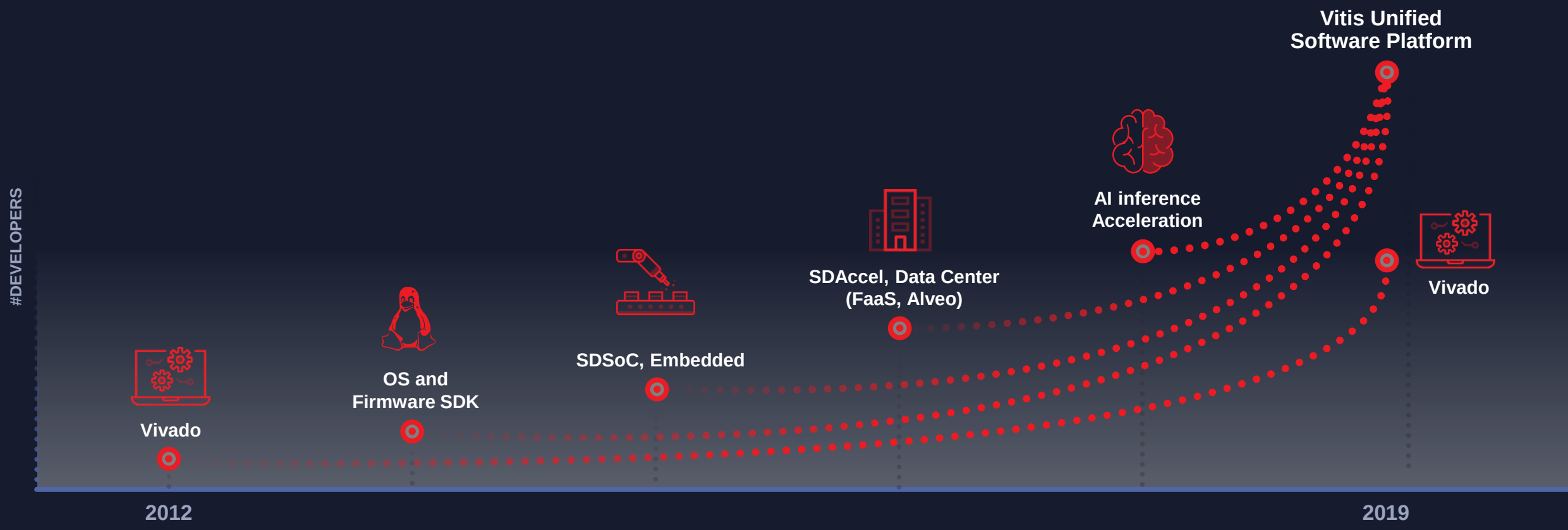
Sergei Storojev
Strategic Application Engineer
22 Sept 2020

XILINX

Vitis Unified Software Platform



Evolution of Xilinx's Software Platform



Unified: All Developers can Build and Deploy to All Platforms

Build

Embedded Developers

Enterprise Application Developers

Enterprise Infrastructure Developers

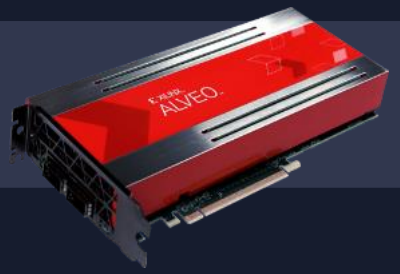
Data & AI Scientists

XILINX VITIS™

Deploy

Zynq

Ultrascale



Alveo

Data Center Rack

Vitis: Unified Software Platform

Coming soon…

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

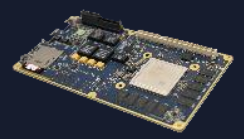
Compilers
ARM, HLS, AI Engines

Analyzers

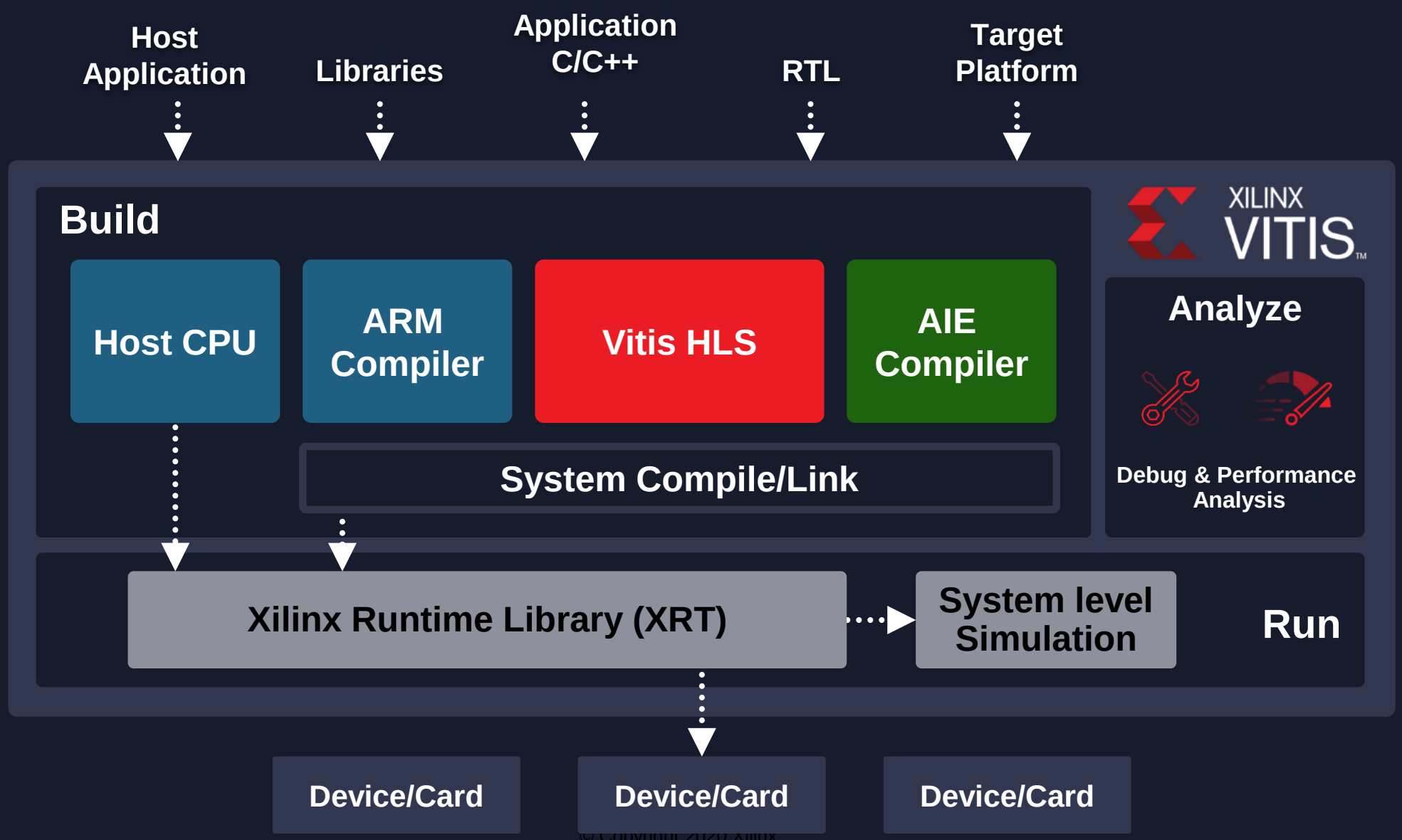
Debuggers

Xilinx runtime library (XRT)

Vitis target platform



Comprehensive Development Tool Suite



Open Source, Standards Based Libraries

Domain-Specific Libraries

Vision & Image

Finance

Data Analytics & Database

Data Management

Data Security

Common Libraries

Math

Matrix Decomposition (Cholesky, LU, etc.)
Linear Solvers
Eigenvalue Solvers
Others

Linear Algebra

amax, asum, copy, gbmv, scal, swap, trmv, others
GEMM

Statistics

Random Num Gen
Brownian Bridge Trans
Heston Model
Black-Scholes
Interpolations
Others

DSP

Monte-Carlo
Box-Meuller Trans
Probability Density
Binomial Tree
Markov Chain
Others

Data Compression

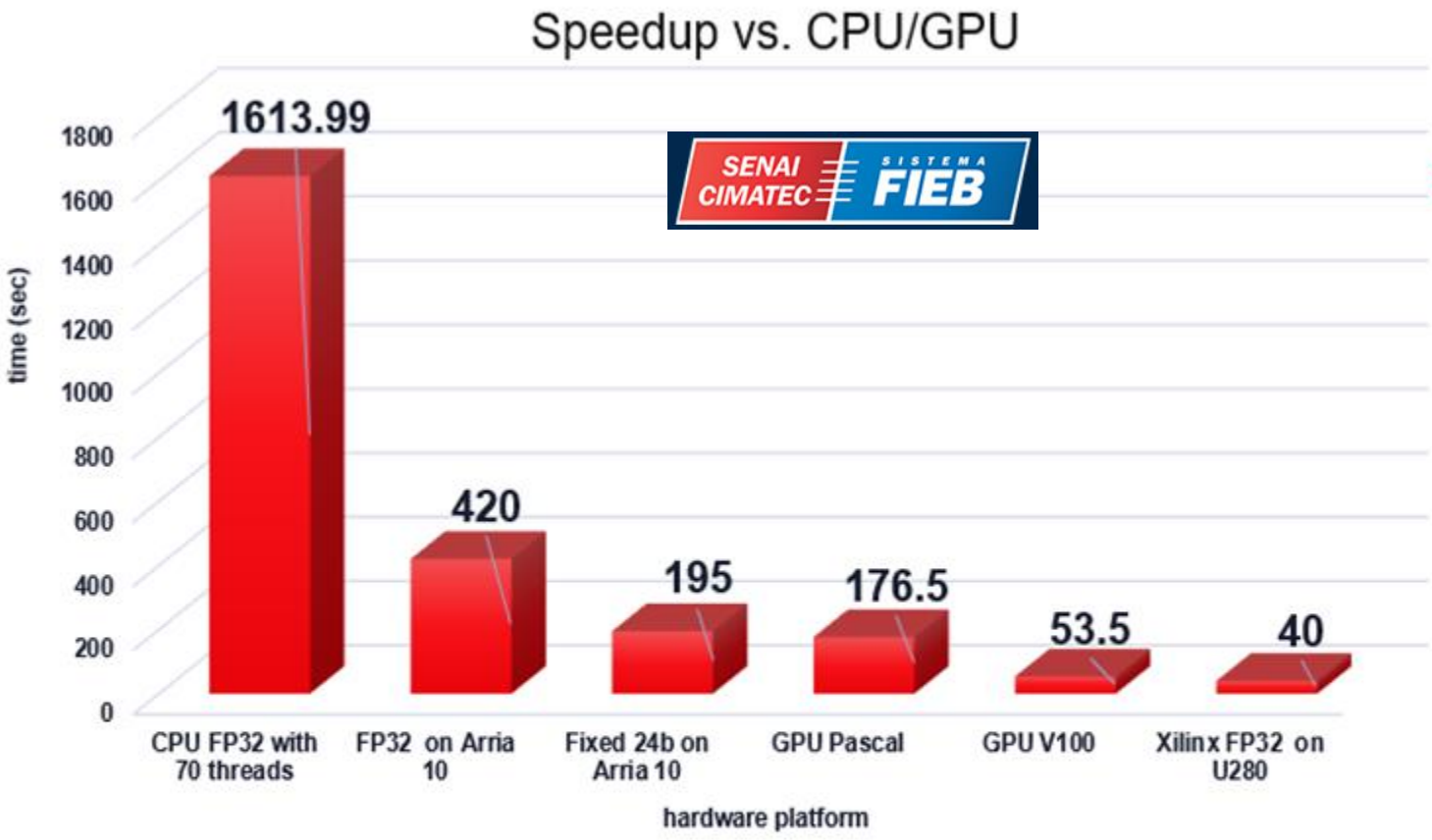
lz4 Comp/Decomp
Huffman Enc/Dec
Snappy Comp/Decomp
Others

GitHub

- Library
  - Docs
    - Source
    - Tests
    - Examples
    - Benchmarks

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

SW Programming Model - RTM Benchmarking



- **40X faster than XEON Gold CPU, 8x faster than Intel FPGA (with hardened floating point), 34% faster than Volta V100 at <1/4th the power**
  - Xilinx implementation currently based on 2x forward + backward kernels
  - Currently working on shot-independent/kernel independent optimization to improve beyond V100 performance
  - Power Measurements:
    - V100 = 182W avg., peak 225W
    - U280 = 40W max.

Not the traditional programming model for FPGAs:

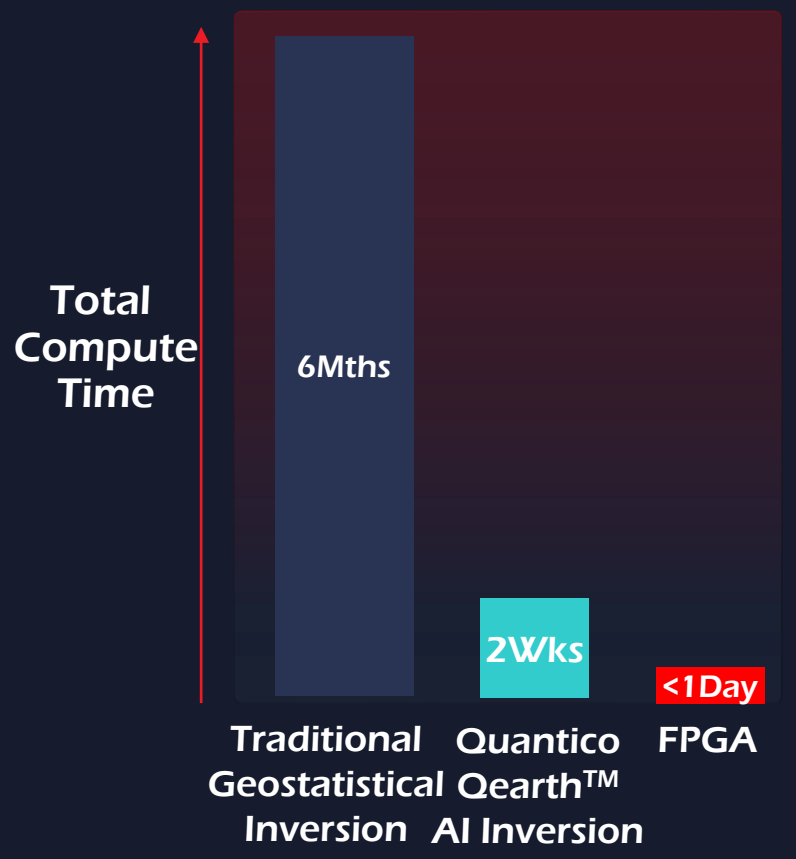
- One engineer/one month to describe & implement entire 2D RTM Algo in C++ from scratch
- Standard language, open source tools and libraries, excellent performance

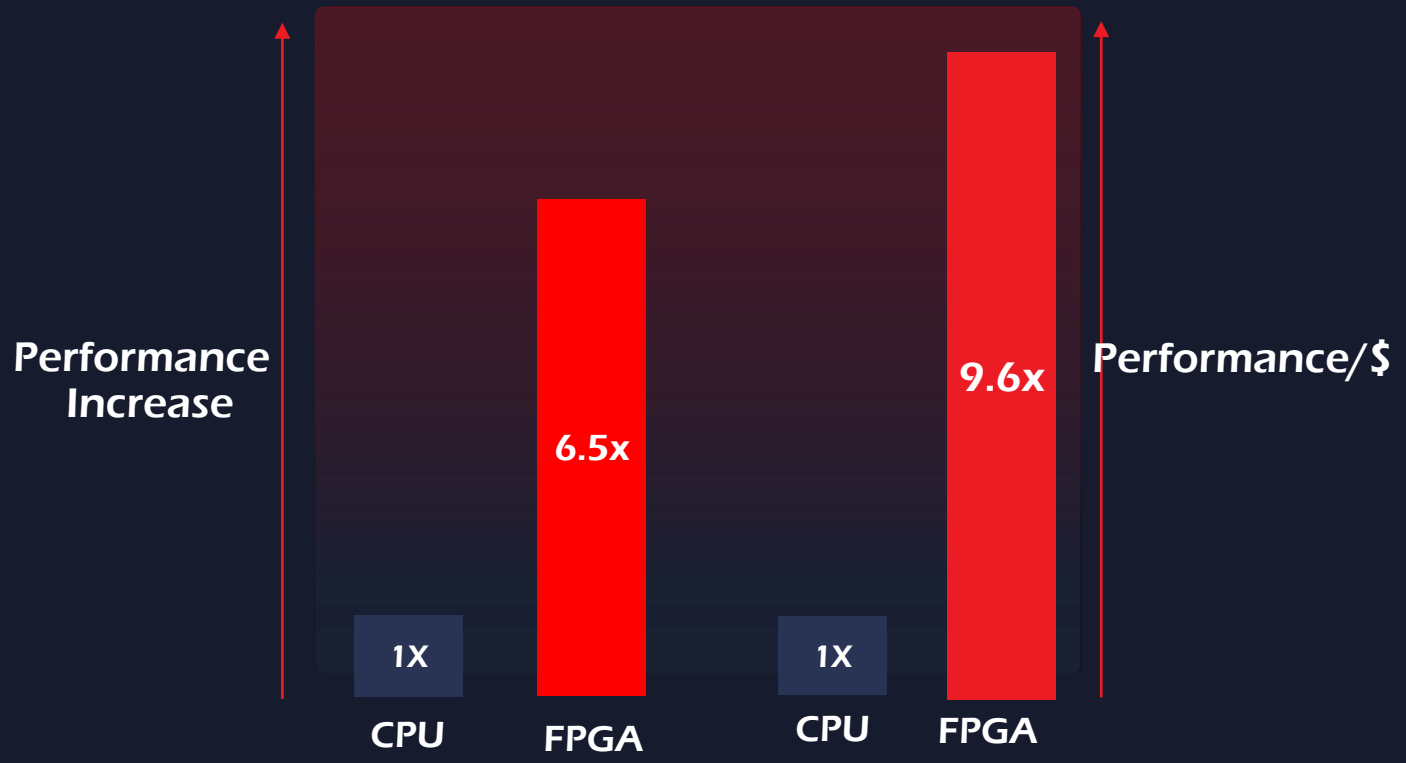
HPC Use Cases

XILINX

Oil & Gas - Realtime Subsurface Imaging using AI/ML

QUANTICO
ENERGY SOLUTIONS



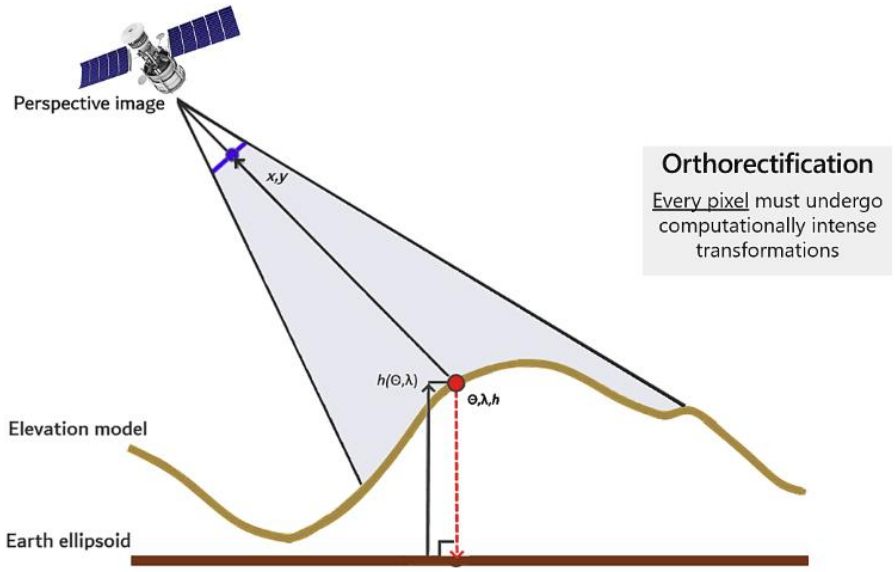


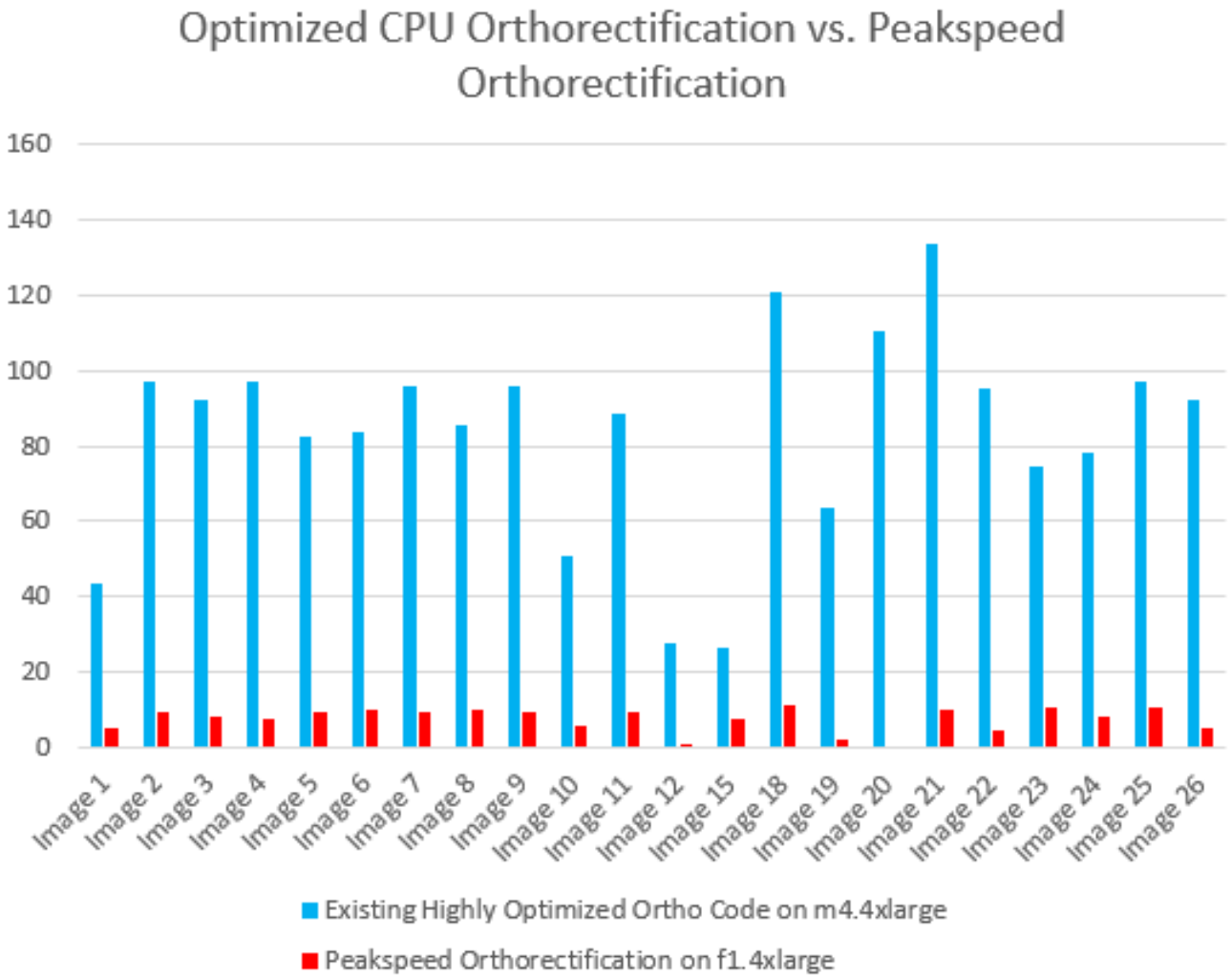
Xilinx/Quantico Analysis, QEarth running 200K-1.7M traces

XILINX

Satellite/Aerial Imaging

PEAKSPEED





Results of a pilot in AWS

Peakspeed's TrueView Orthorectification

vs.

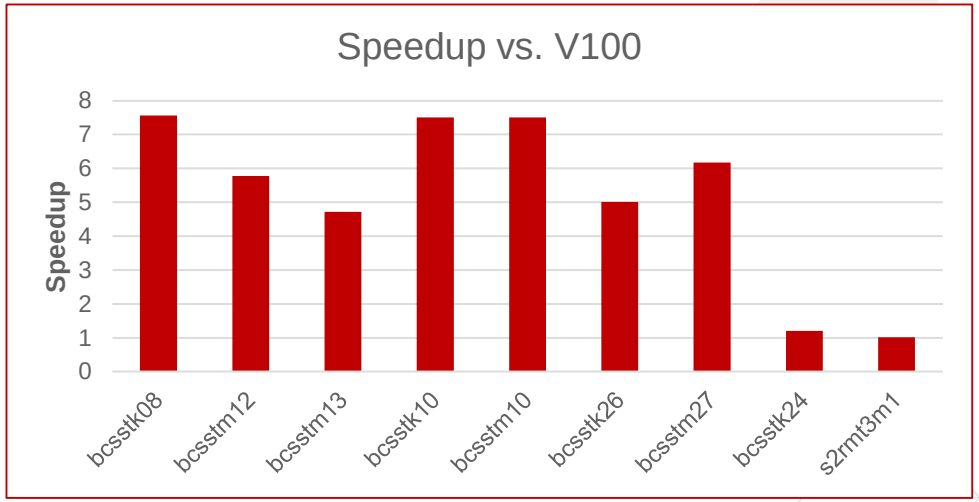
Optimized Customer solution on production images

- Reduced time to insight
- Reduced IT total cost of ownership (TCO).

PEAKSPEED XILINX

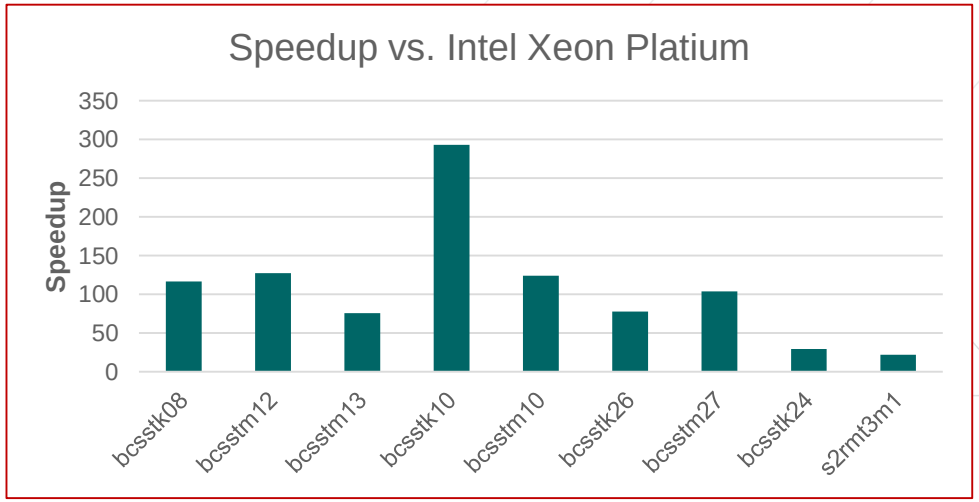
Sparse Matrix Multiplication Vitis Library

- Based on SuiteSparse Matrix Collection (https://sparse.tamu.edu/)
  - Real application problems spanning structural engineering, computational fluid dynamics, thermodynamics, quantum chemistry, financial modeling, etc.
- Performance results for 24 channel design on Alveo U280
  - Medium size (NNZs < 100K) matrices in “structural” category in the benchmark
    - Up to 7.5x better than V100
    - Up to ~292x better than Intel Xeon Platinum 8268 CPU
  - Big size matrices
    - Up to 39x better than Xeon Platinum 8268 CPU
    - Up to 20% better than V100 when NNzs < 300K



For each device, the mesured time is the time for executing the sparse kernel only
GPU: Nvidia V100 SXM2 16GB, CUSPARSE, CUDA 10.2

~13K Increasing number of NNZs ~219K

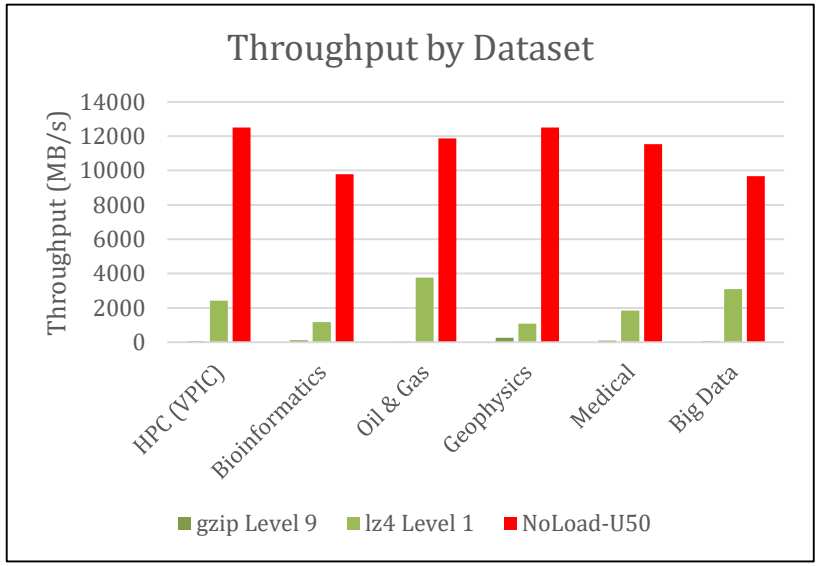


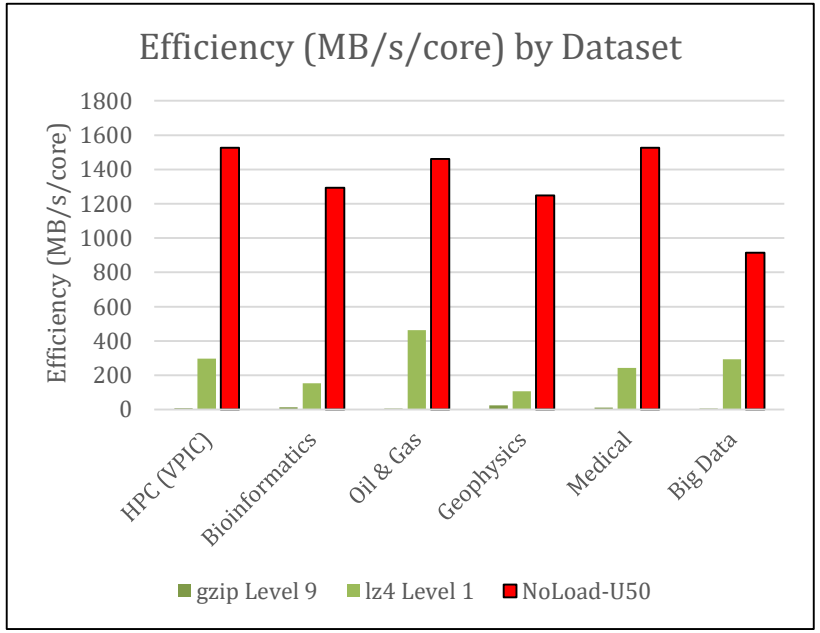
CPU: Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz, 2 sockets, 96 cores (using all 96 cores).
Intel MKL mkl_2020.1.217

XILINX

Transparent/Line-Rate Compression

EIDETICOM





NoLoad® provides gzip levels of compression with better throughput and efficiency than lz4!

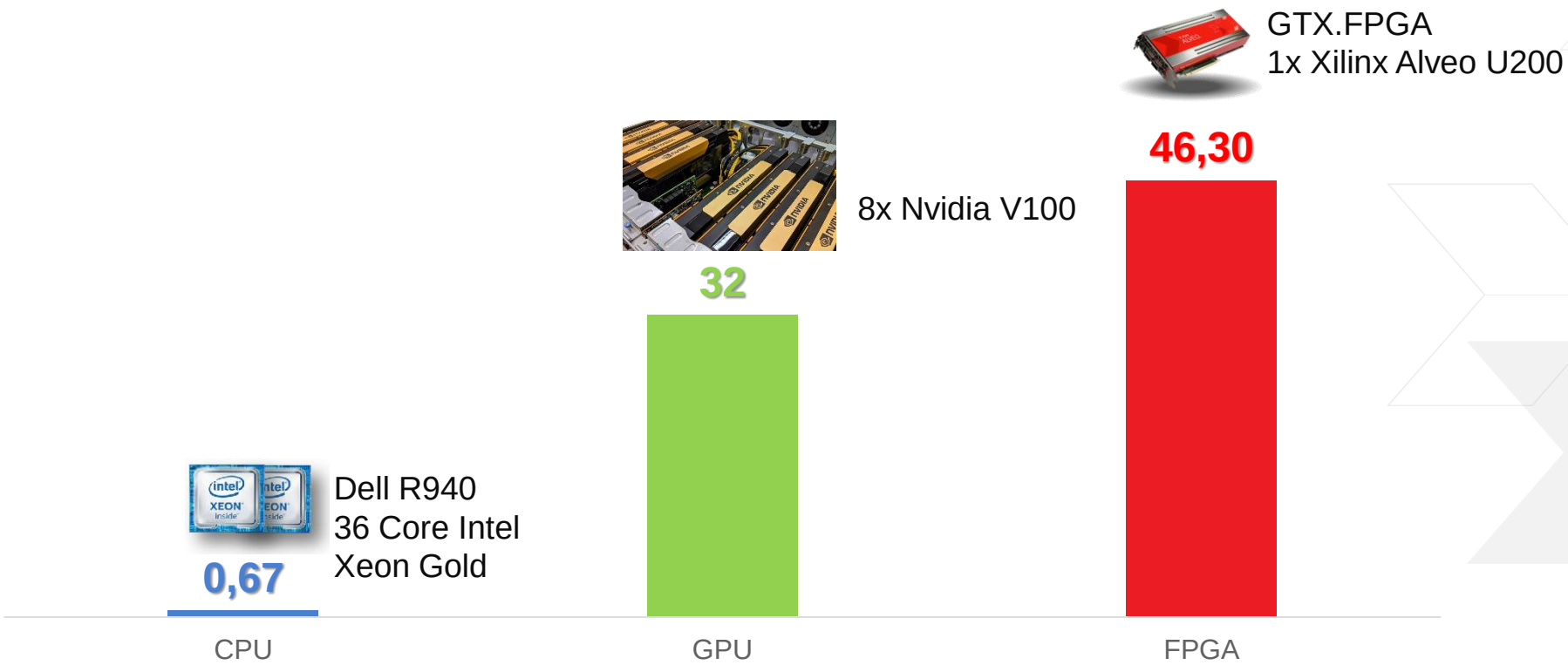
Dataset	NoLoad		gzip (level 9)		lz4 (level 1)	
	CR	MB/s/core	CR	MB/s/core	CR	MB/s/core
petroleum	2.11	1462	2.2	5	1.97	473
seismic	1.42	1320	1.43	25	1.28	363
medical	2.24	1410	2.35	22	1.57	401
video	1.02	1020	1.02	36	1.02	484
genomics	2.01	1293	2.07	13	1.42	154
big data	2.91	914	3.53	5	2.43	292
HPC (VPIC)	1.23	1526	1.23	7.1	1.01	296

NoLoad® on Alveo U50 achieves **over 12GB/s of compression input** per card and scales linearly as you add cards!

XILINX

Genomic Sequencing Analysis

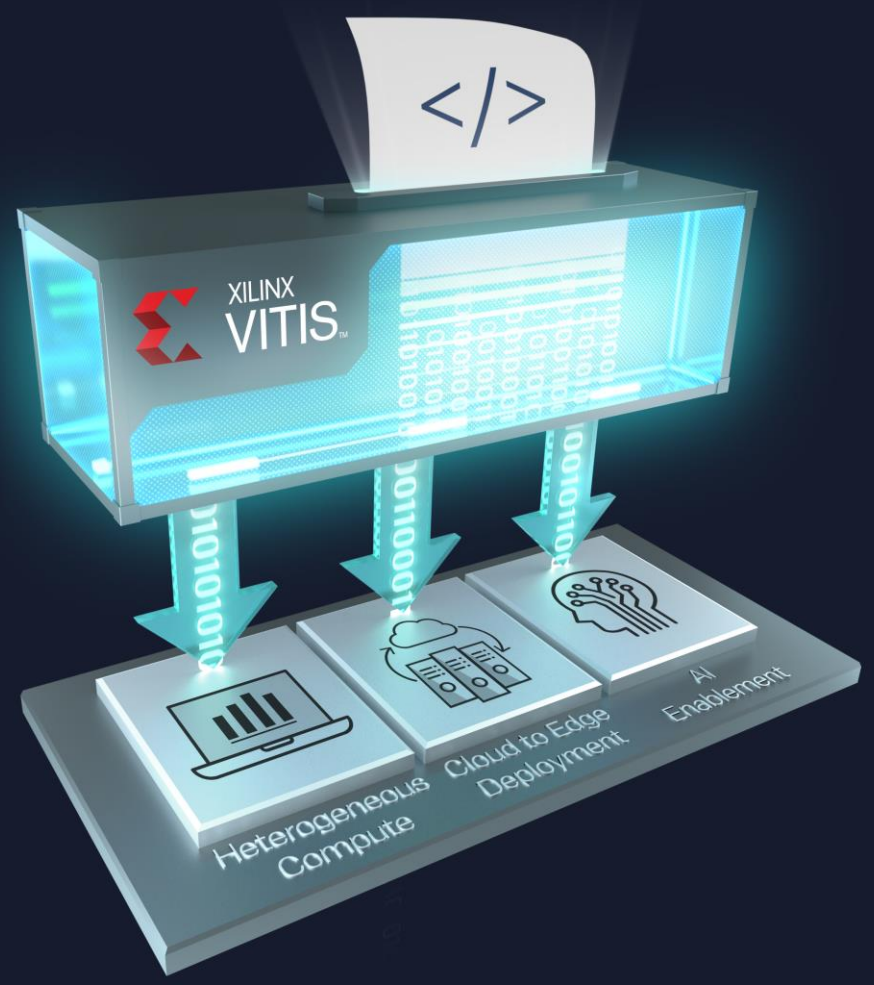
Number of WGS* Samples Processed in 24 hours



Analysis Pipeline: GATK Best Practice Pipeline for 30x Human WGS Variant Calling

**WGS: Whole Genome Sequencing*

Summary



Unified Software Platform

Cloud to edge, software and AI

Comprehensive tools, runtime, libraries and models

Standards, Open Source, Free

Embracing & participating in open source

Use of standard environments & APIs

How to get started

Download Vitis for Free

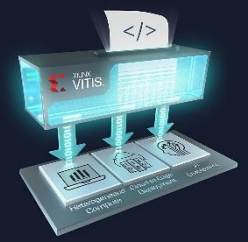
www.xilinx.com/products/design-tools/vitis.html

Access our Vitis Open Source Libraries
www.xilinx.com/products/design-tools/vitis/vitis-libraries.html

Getting Started Tutorials

github.com/Xilinx/Vitis-Tutorials/tree/master/docs/vitis-getting-started

XILINX

Vitis™		Vitis™ Unified Software Platform Overview Adaptive Computing Challenge I Developer Contest Vitis™ Unified Software Platform Accelerated Libraries
Alveo™ cards		ALVEO accelerator cards Getting Started with the Alveo U200 & U250 Vitis Acceleration Development Flow on Alveo
Versal™ ACAP		Versal ACAP Introducing the Versal Premium ACAP Versal ACAP: AI Engine
Boards		Evaluation Boards System on Modules (SoM) Boards Board and Kit Accessories

Adaptable.
Intelligent.

XILINX