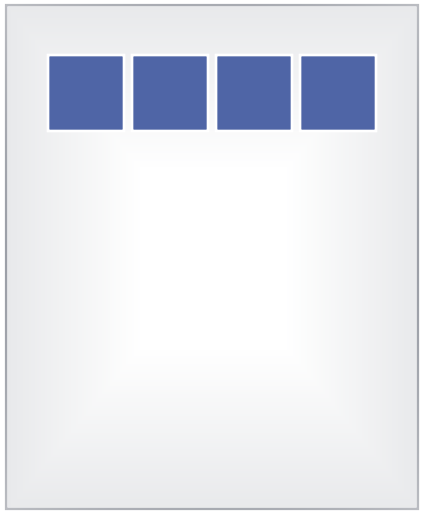
XILINX

Real-life Benchmarks vs CPUs and GPUs

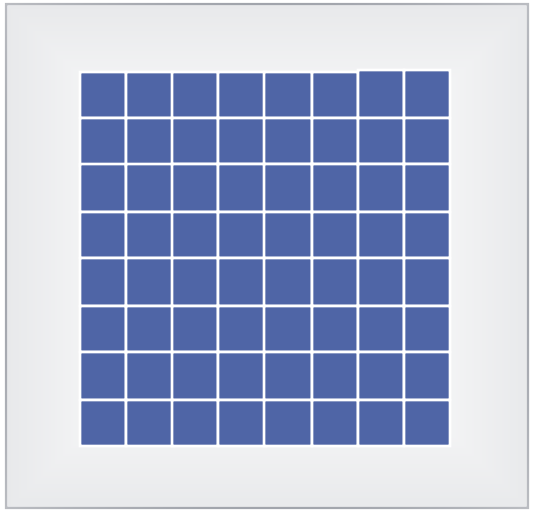
Nick Ni, Director of Product Marketing, AI – Software - Ecosystem

Era of Domain-Specific Architectures (DSAs)



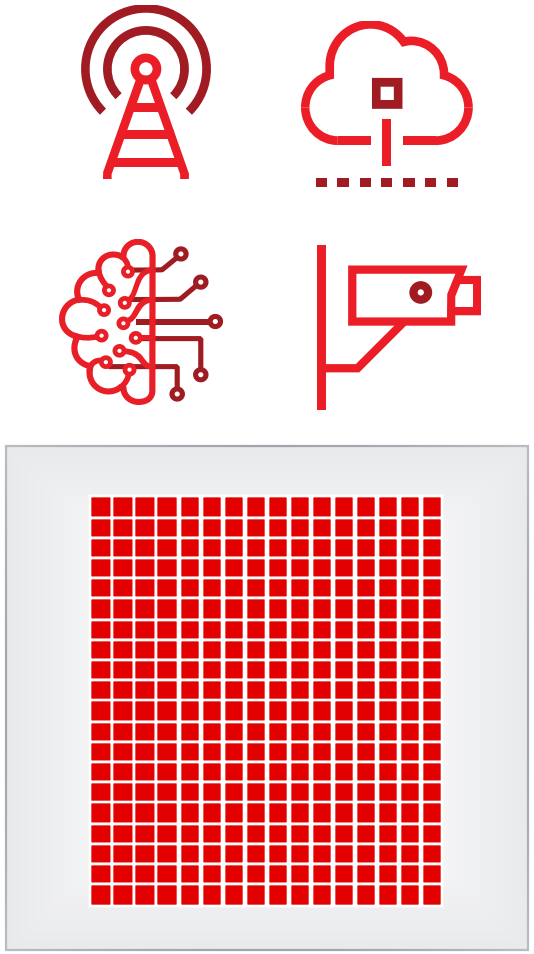
CPU

CISC → RISC → Multi-Core



Fixed HW Accelerators

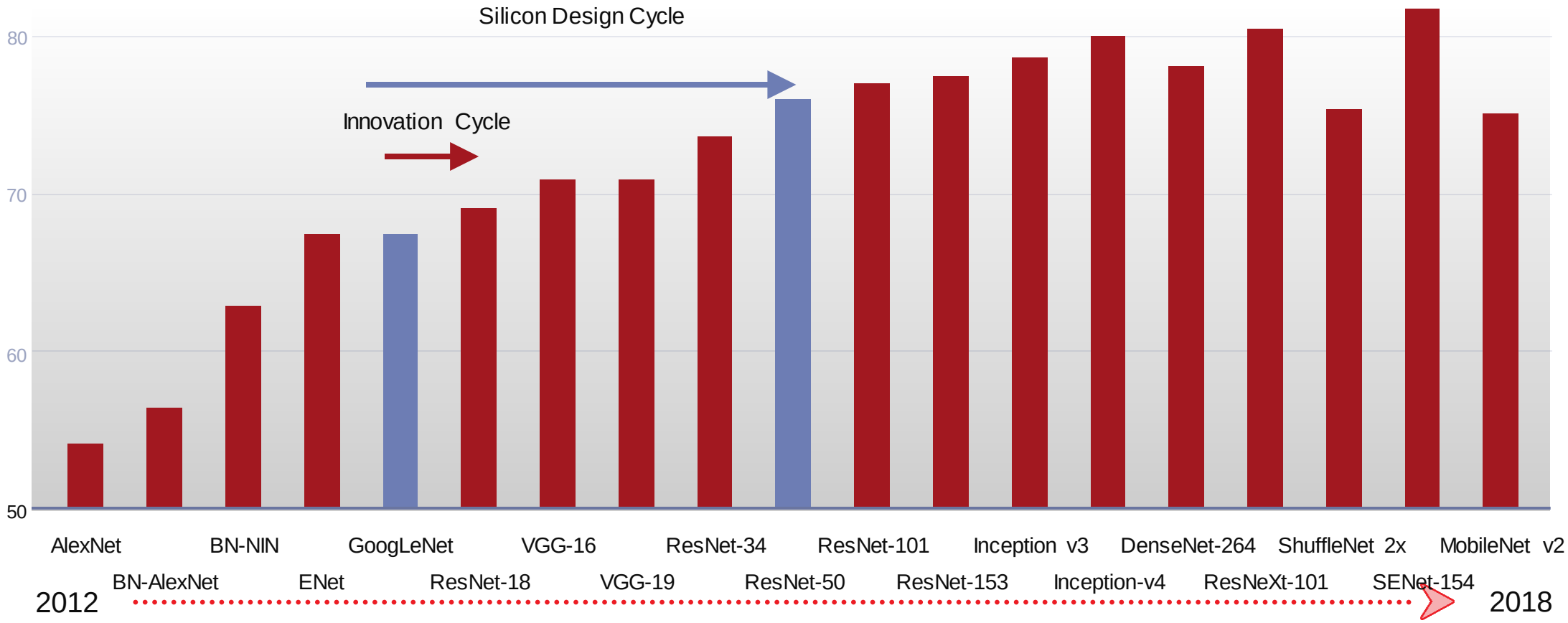
GPU, ASSP, ASIC



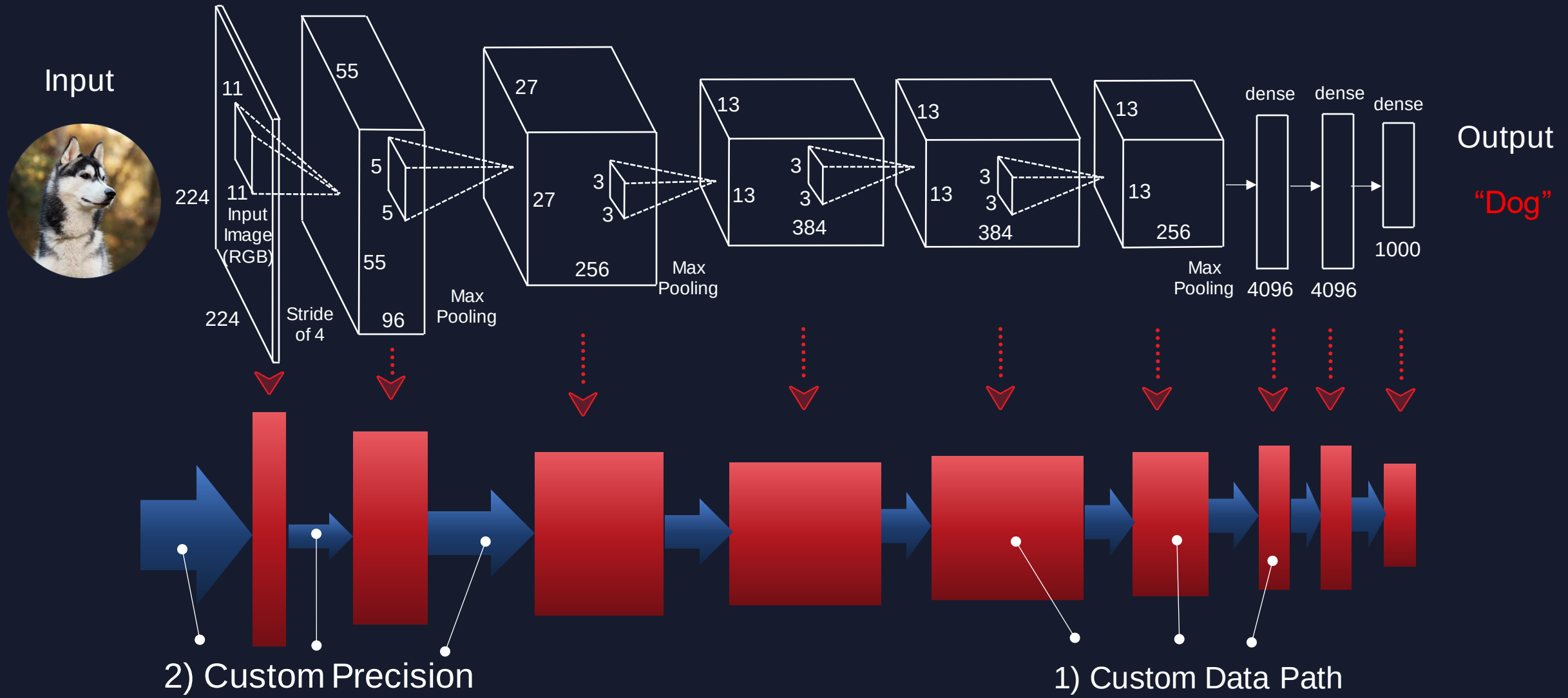
DSA

TPU, DLA
FPGA, Zynq, ACAP

Speed of Innovation Outpaced Silicon Design Cycles

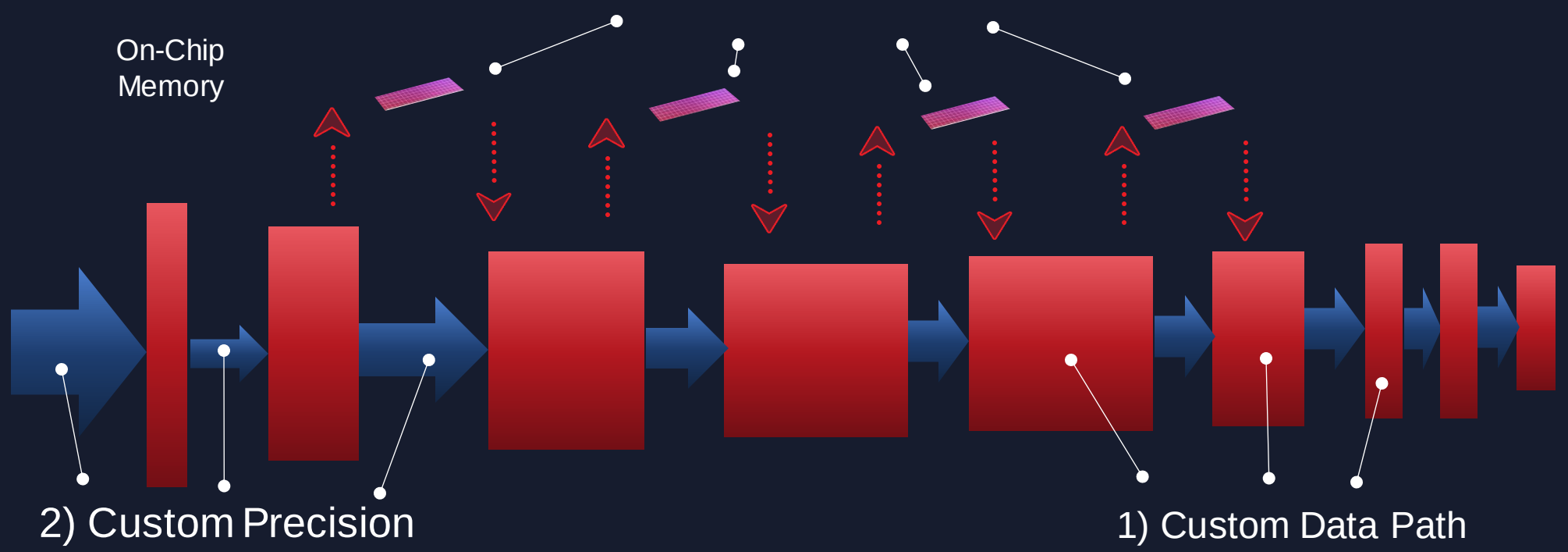


What is a Domain-Specific Architecture?



XILINX

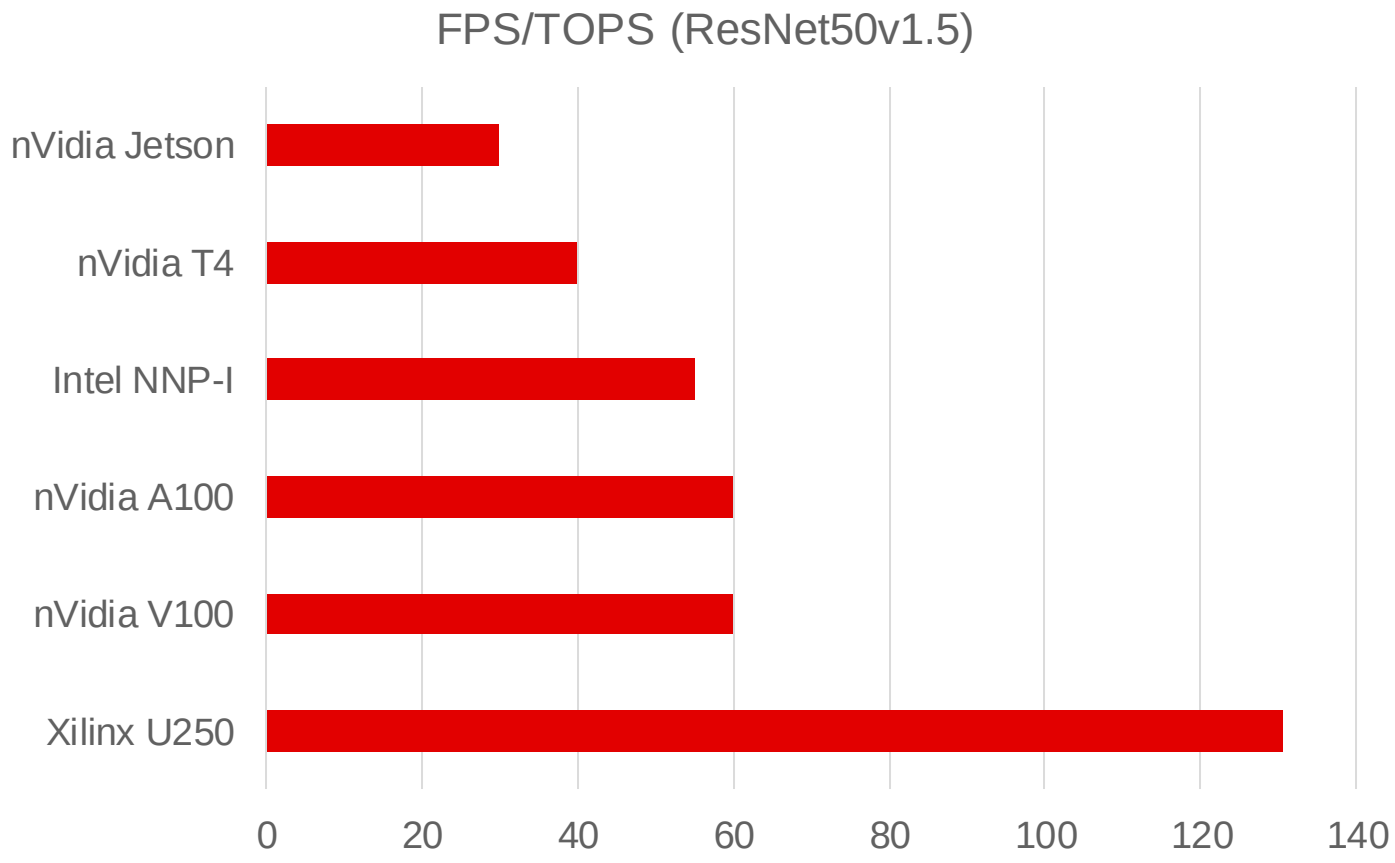
What is a Domain-Specific Architecture?



Xilinx Achieves the Highest Efficiency in Computation

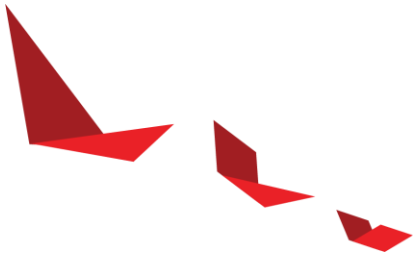
On a standard benchmark in MLPerf v0.7

MLPerf

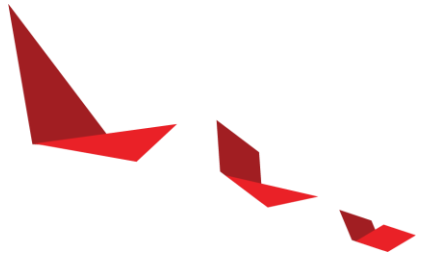


XILINX.

Agenda



- AI
- Computer Vision
- Database
- HPC



AI

Super Resolution
Sentiment Analysis
Visual Search
Point Cloud LIDAR

XILINX.

Super Resolution (Image upscaling with EDSR)

Mipsology

- Model: EDSR, Pytorch paper

FHD 2K

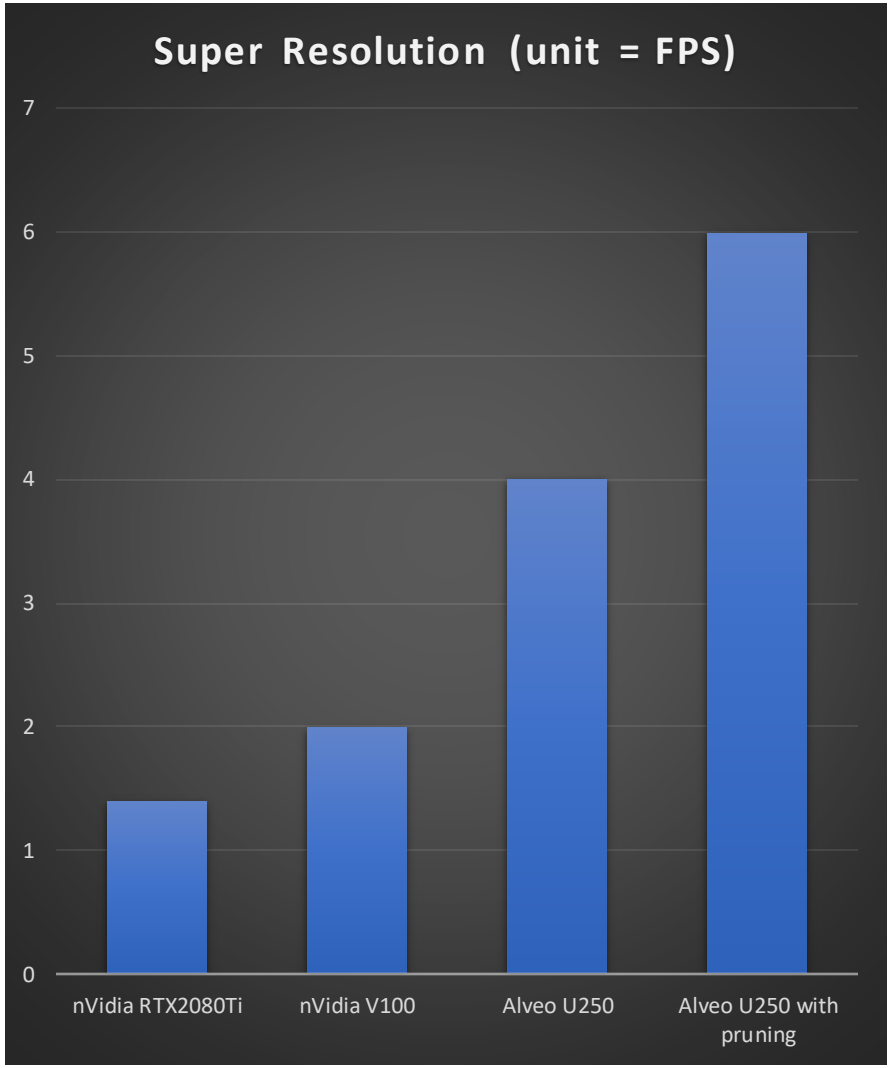
2x

UHD 4K

3x performance

4x lower COGS

8x lower TCO

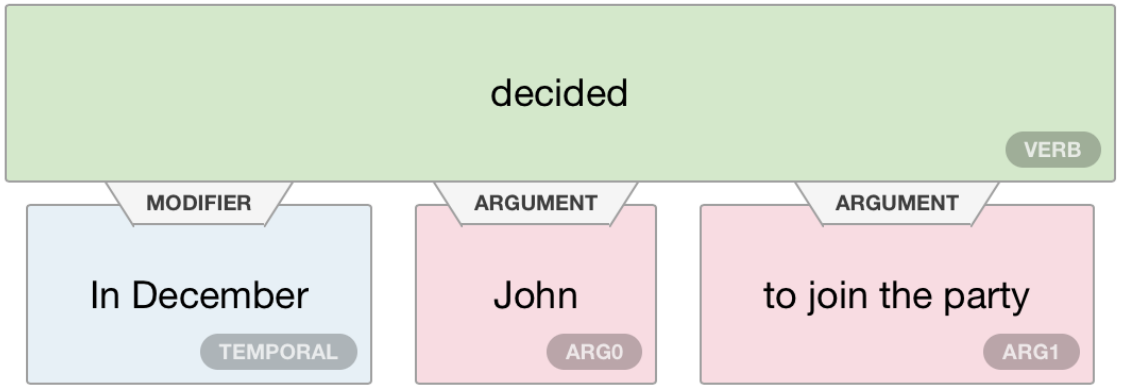


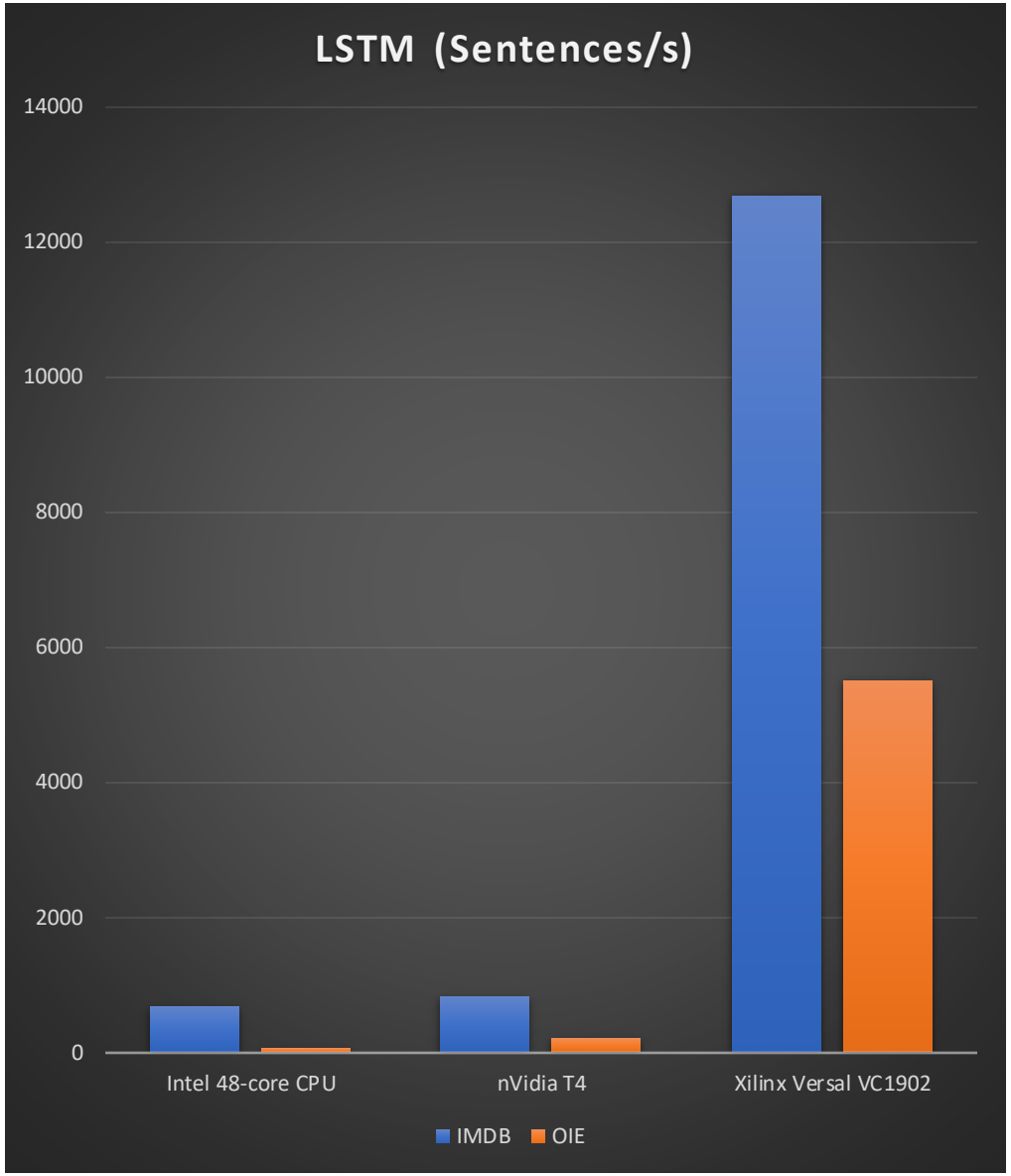
XILINX.

LSTM - Sentiment Analysis

- IMDB Sentiment Analysis - link
- Open Information Exchange - link

In December , John decided to join the party .

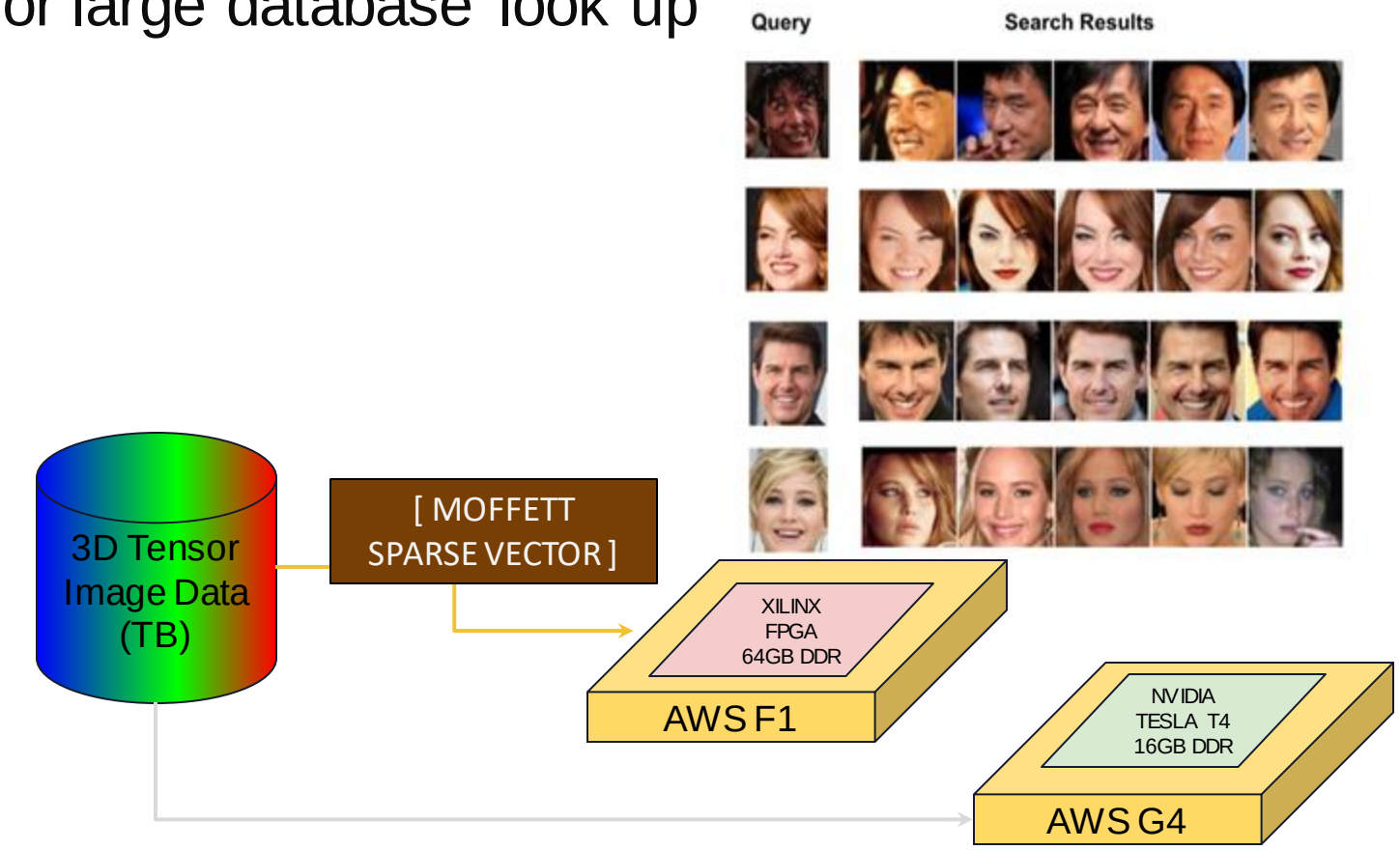


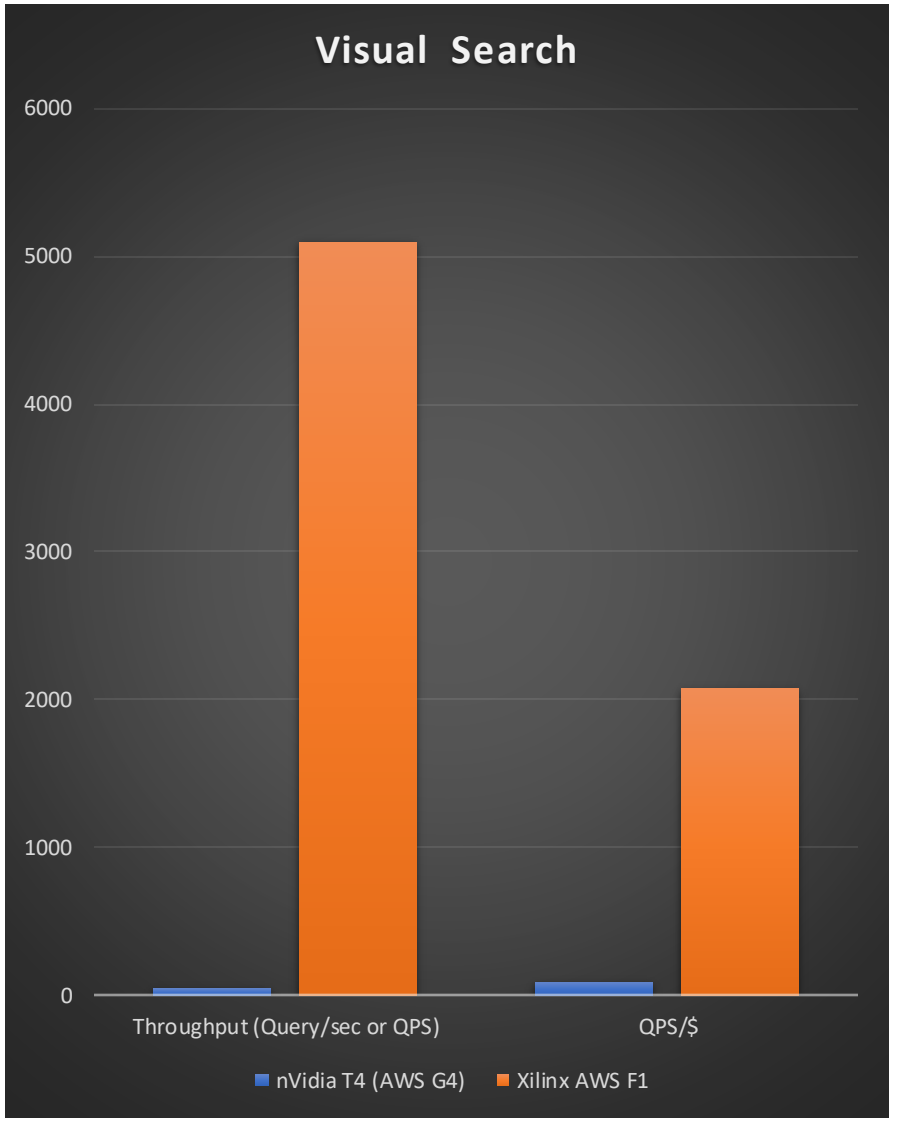


CPU: 48 cores, Intel® Xeon(R) Gold 6136 CPU 3.00GHz | **float32** model
GPU-T4: | **float32** model
Xilinx Versal VC1902: **320**Cores Design on Versal | INT8/INT16 model

Visual Search

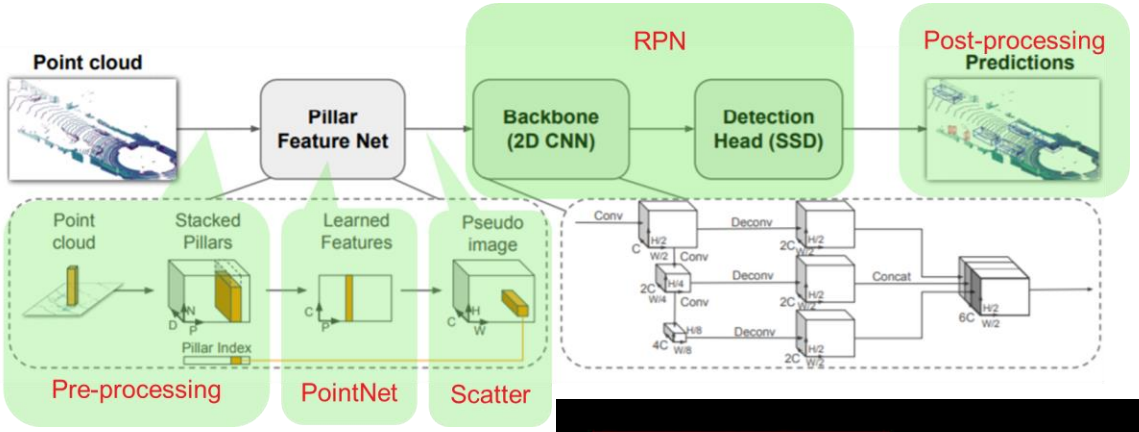
- 1st stage: CNN model for feature extraction
- 2nd stage: Highly sparse Fully Connected models for large database look up

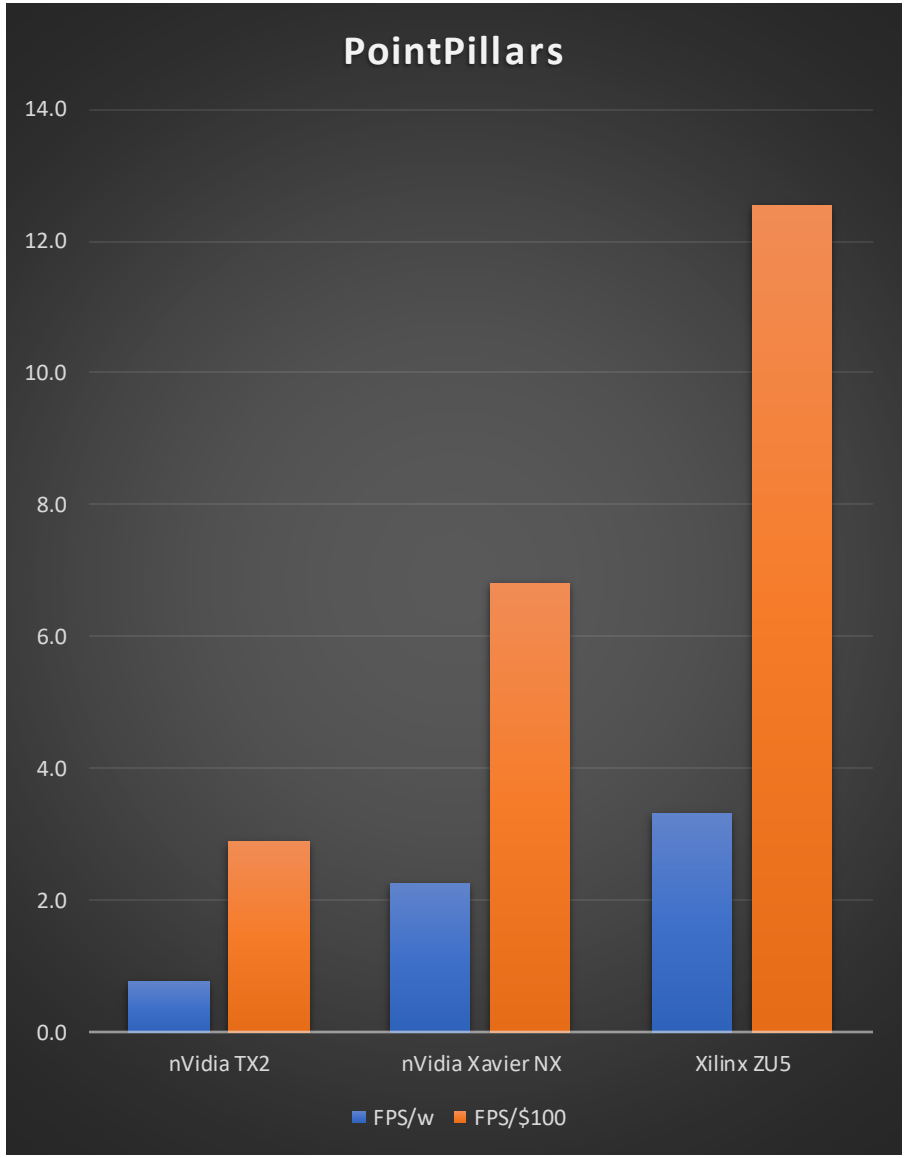




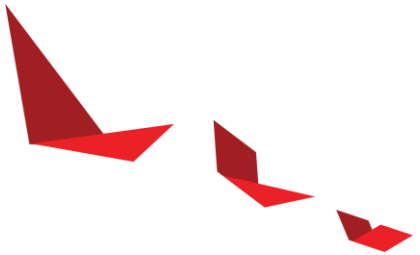
LIDAR Perception based on Point Cloud input

- Model: PointPillars with Pytorch (paper)





XILINX.



Computer Vision

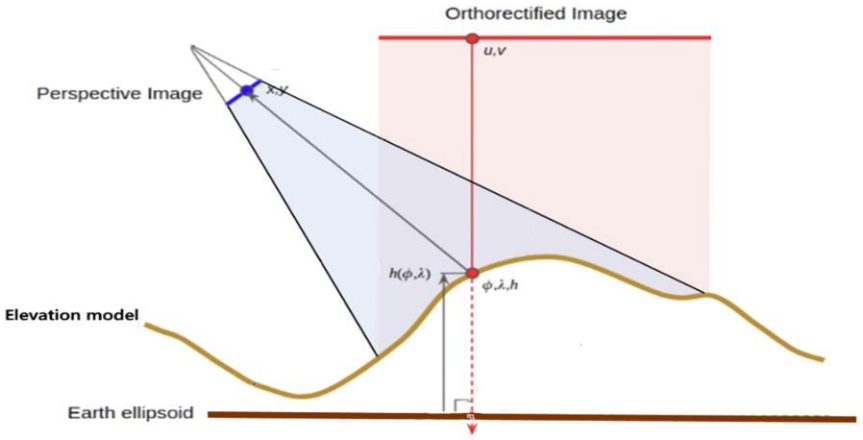
Satellite imaging
Depth estimation

XILINX.

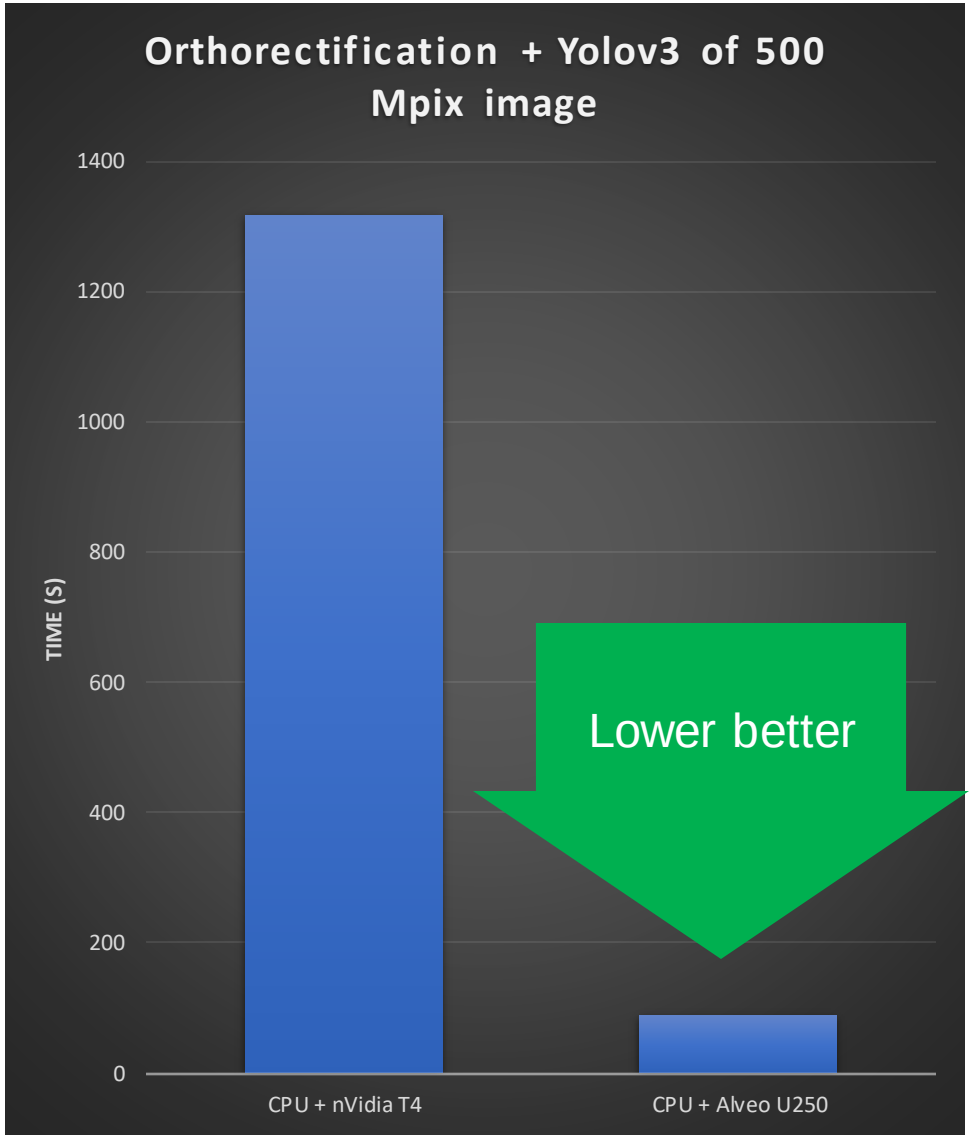
Satellite Imaging

PEAKSPEED

- 1st stage: Orthorectification
- 2nd stage: Yolov3 object detection



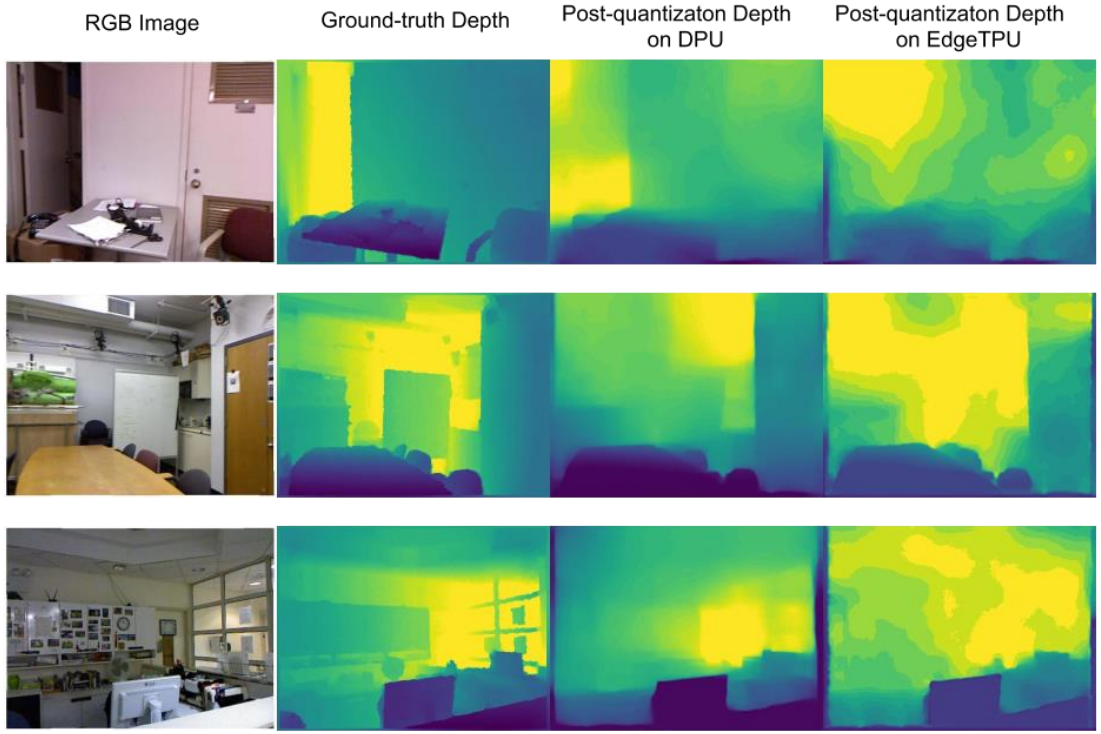
Credit: https://www.mdpi.com/2072-4292/12/1/44/htm

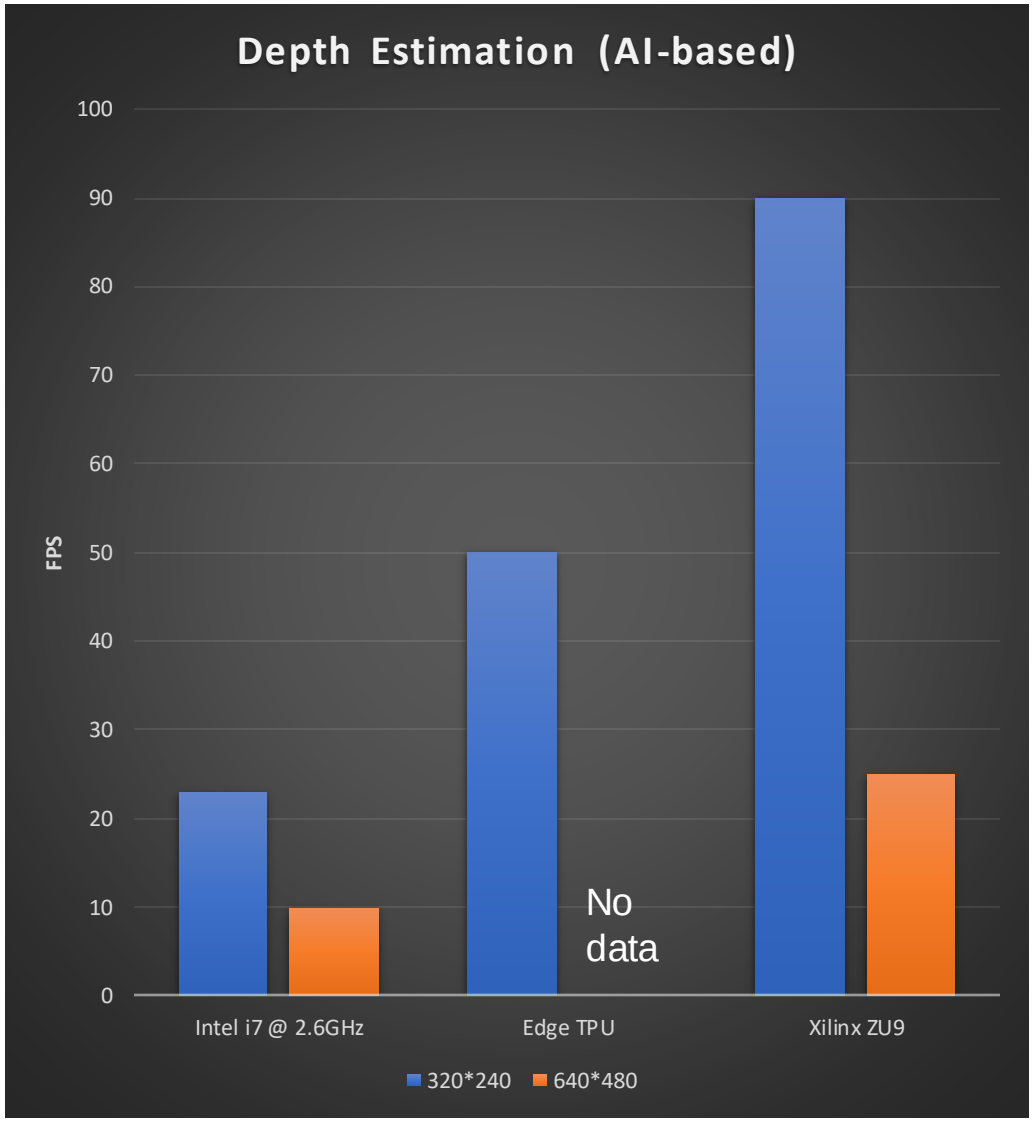


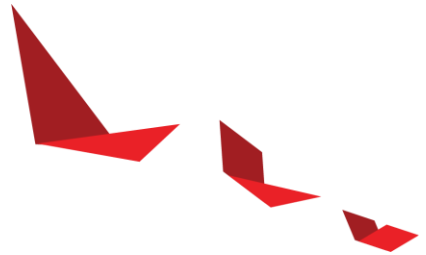
XILINX.

Depth Estimation with CV and Deep Learning

- 1st stage: Encoder-Decoder depth network
- 2nd stage: Fast bilinear filter







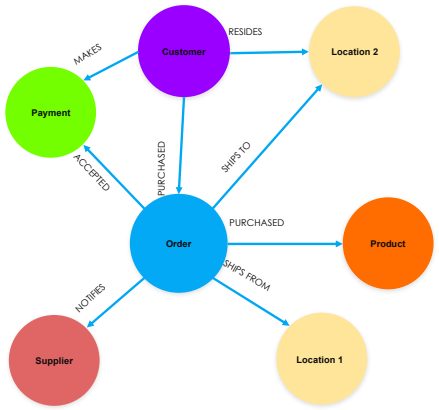
Database

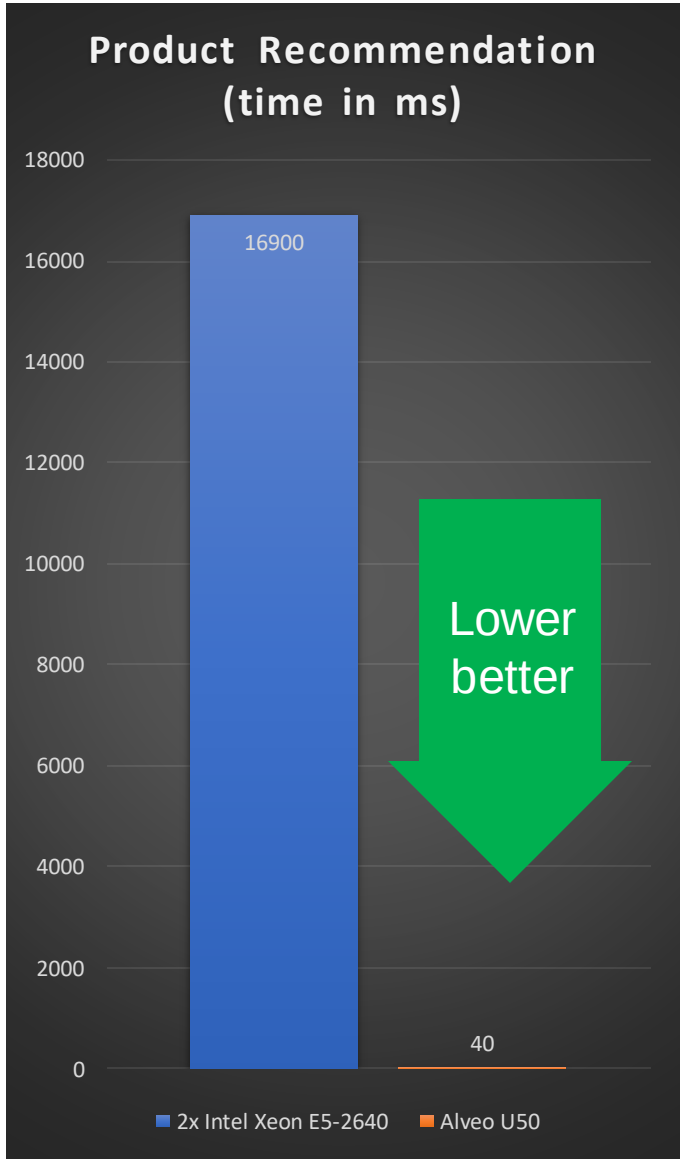
Graph Database
Big Data Queries
Transaction Processing (Decision Tree)
Regular Expression

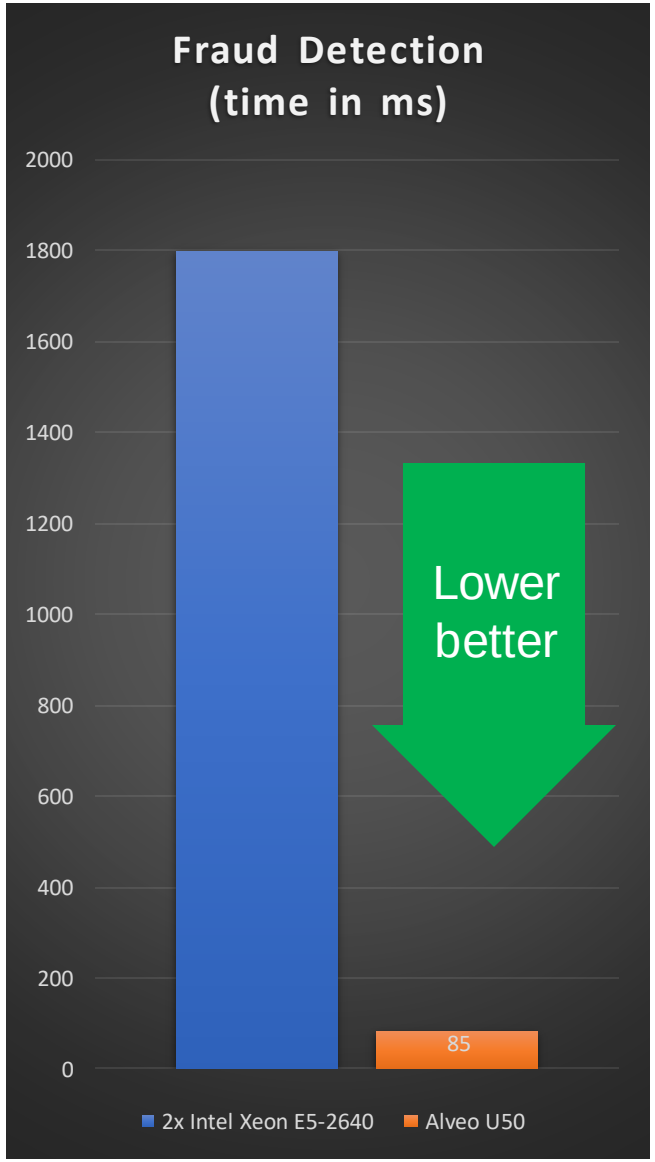
Graph Database: Recommendation, Fraud Detection

TigerGraph

- Graph DB: faster deeper and wider insights on connected data
- App 1: Recommendation with Cosine Similarity for 1.5 million patients
- App 2: Fraud Detection with Louvain Modularity for 50M nodes



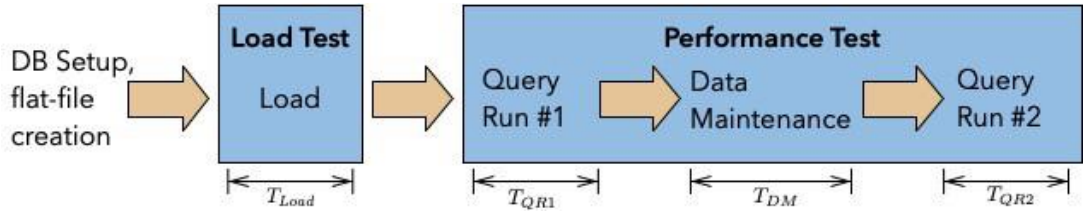




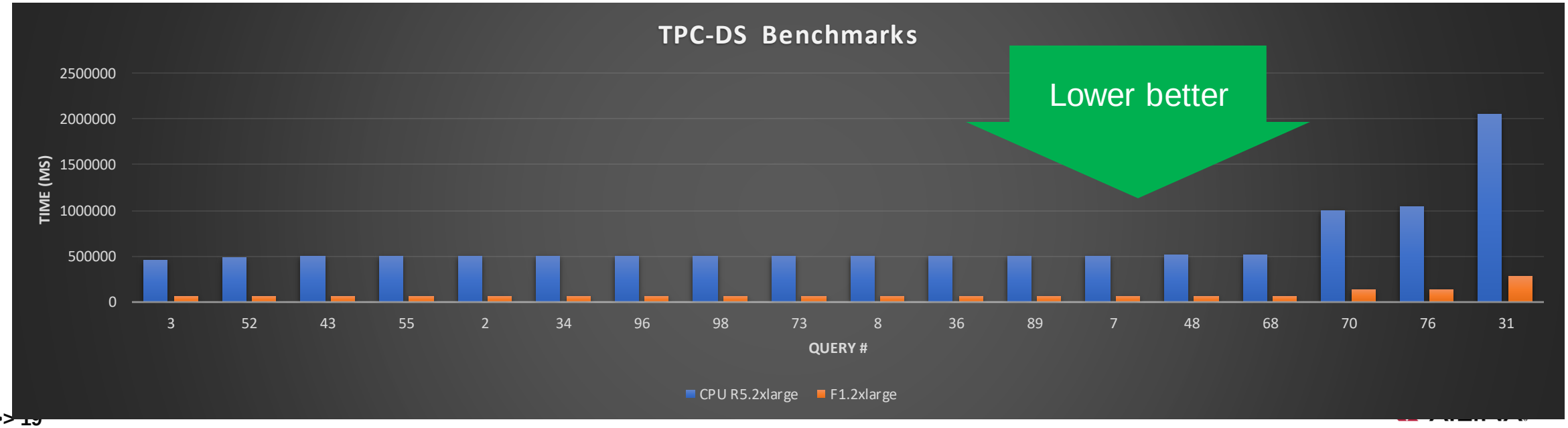
XILINX

Big Data Queries

bigstream

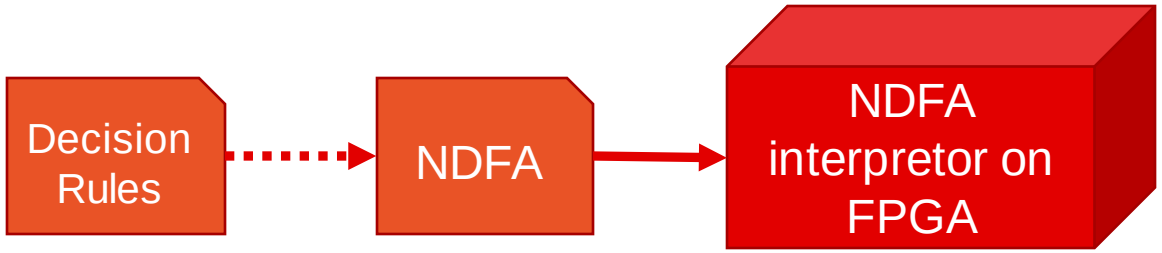


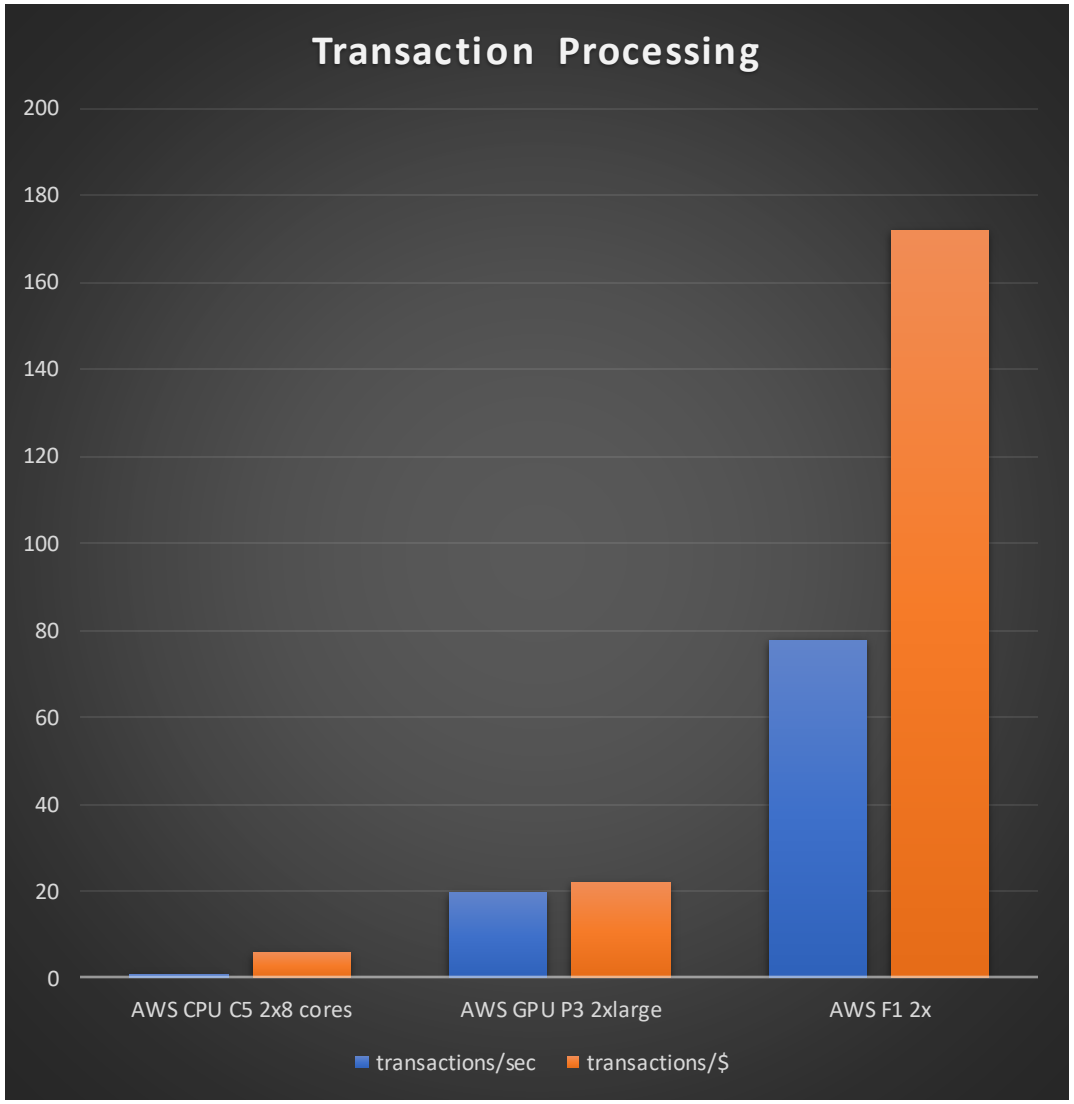
- Apache Spark – TPC-DS Benchmarks
- Using Spark on AWS R5.2xl and F1.2xl cluster with S3 storage
- 250SF JSON Gzip compressed data (~130GB)



Transaction Processing

- Non-Deterministic finite automaton





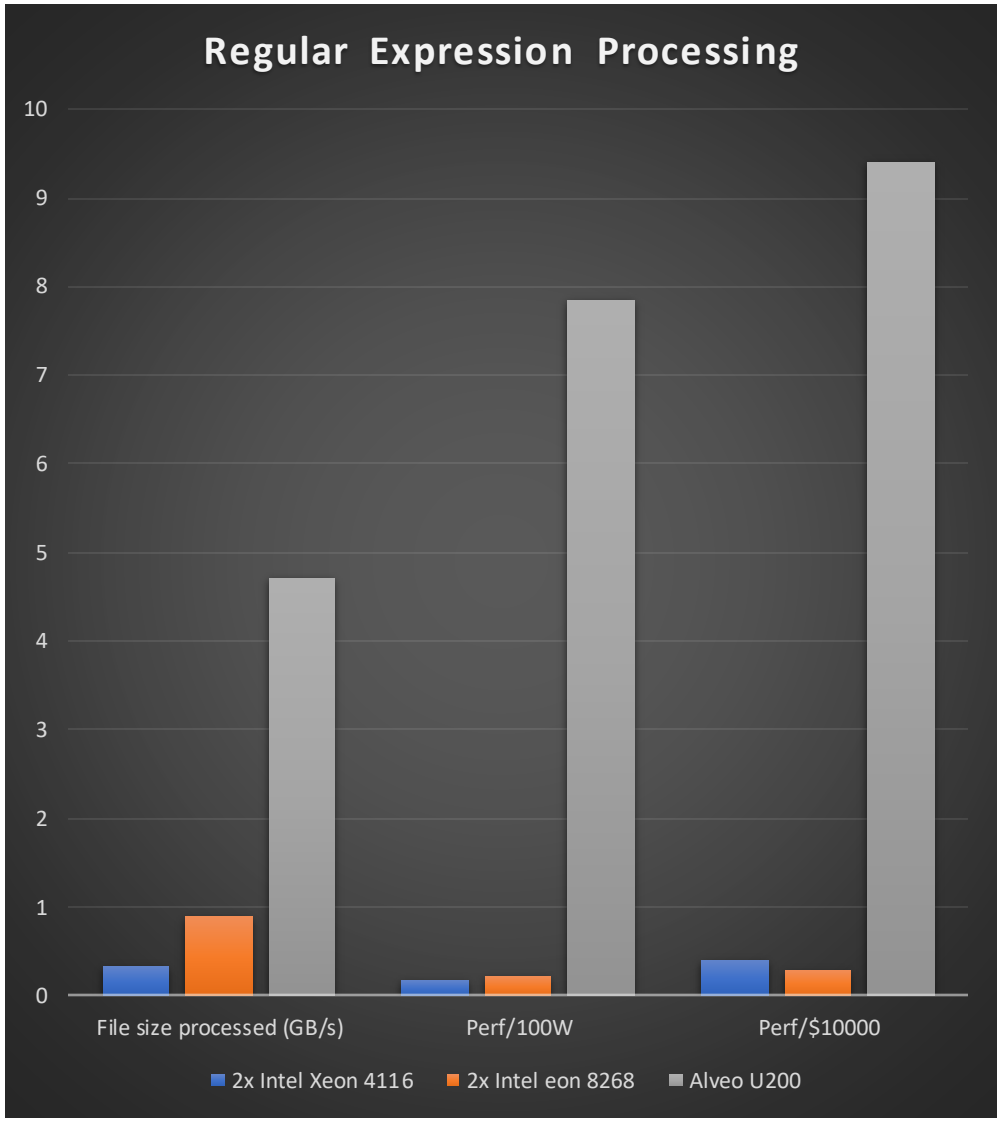
Picture credit: https://www.ttgasia.com/2020/06/04/amadeus-troovo-unite-to-automate-b2b-travel-payments/

GDPR Compliance Check – Regular Expression

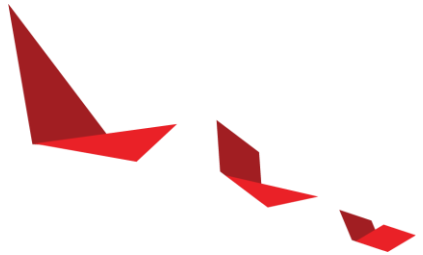
- Regular expression with Capture Group (identify kind: SSN, DOB, CC and Email, and offsets of match)

Operator	Description
.	Any character
*	0 or more, greedy
+	1 or more, greedy
?	0 or 1, greedy
*?	0 or more, lazy
+?	1 or more, lazy
??	0 or 1, lazy

Operator	Description
x|y	x or y, prefer x
[…]	"one of" character class
[^…]	"none of" character class
(…)	capture group
(?:…)	non capture group
\	escape special char



XILINX.



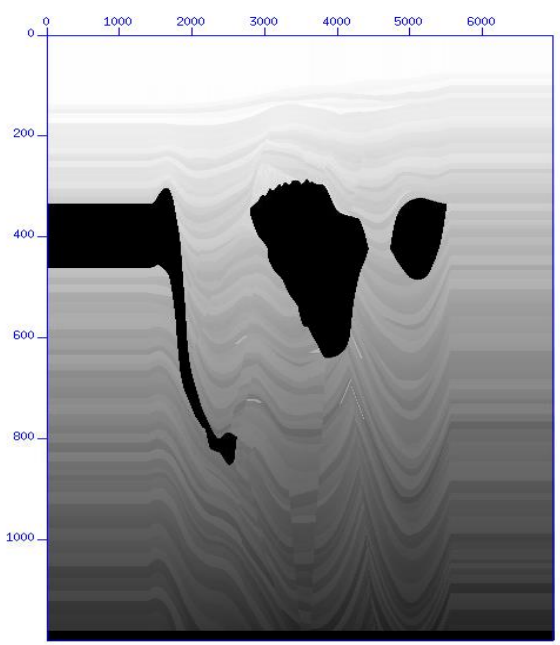
HPC

Seismic Imaging (RTM)
Genomic Sequencing

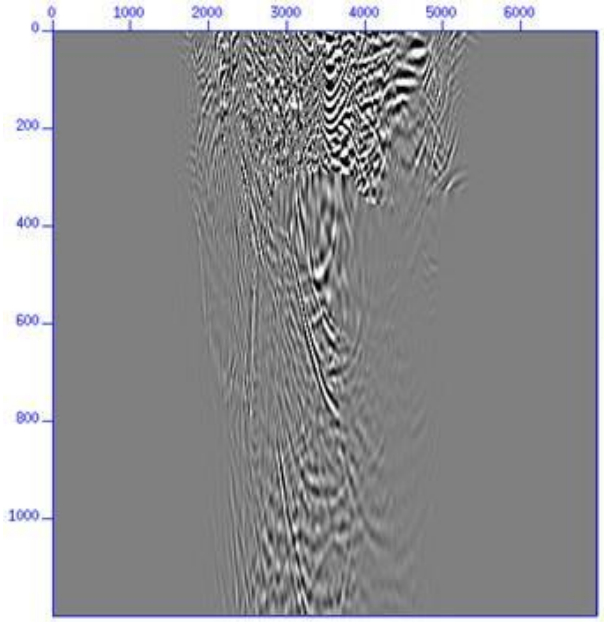
XILINX.

Seismic Imaging

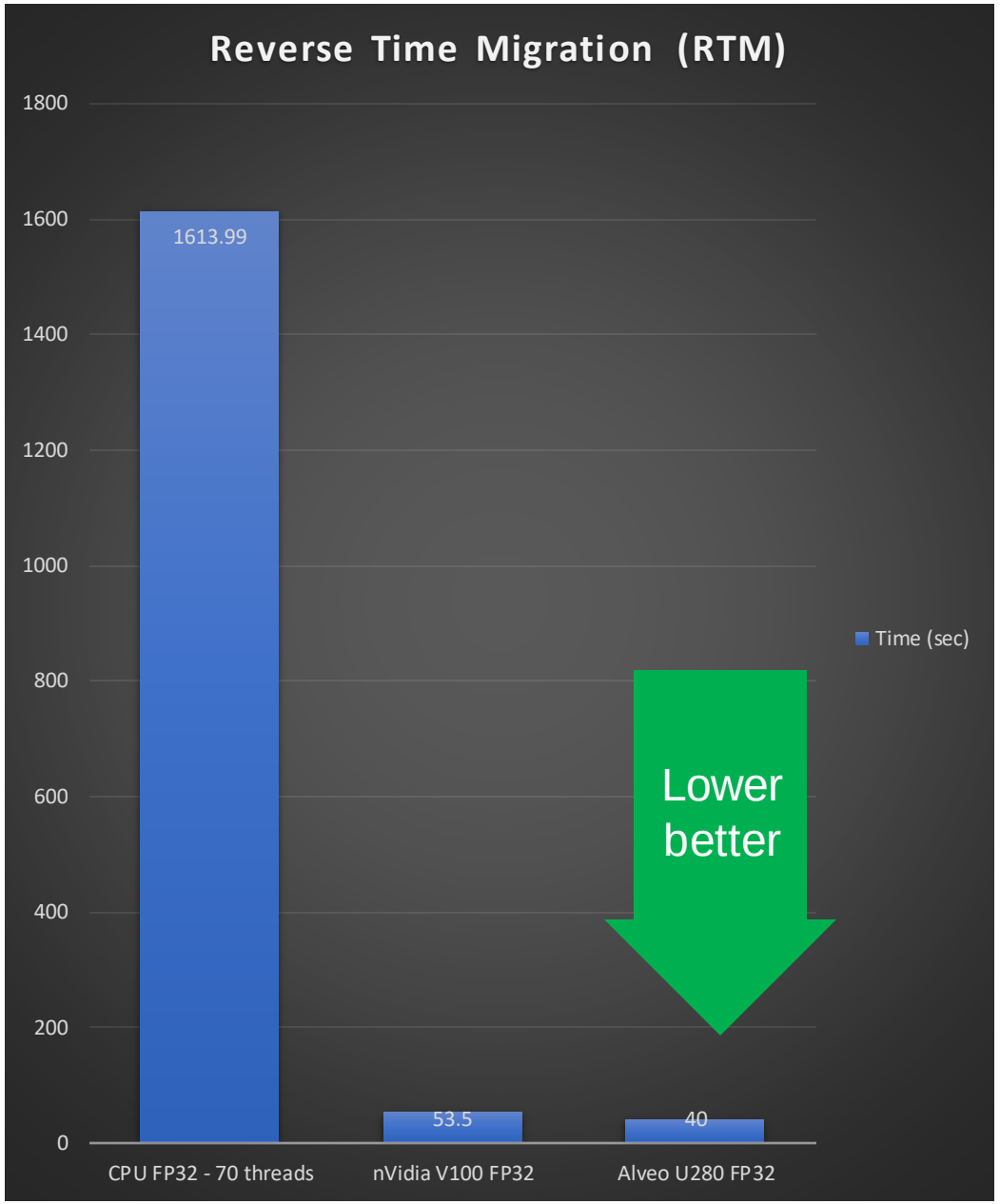
- Reverse Time Migration (RTM)
- Benchmarked with Pluto velocity model



Original input



FPGA 1-shot

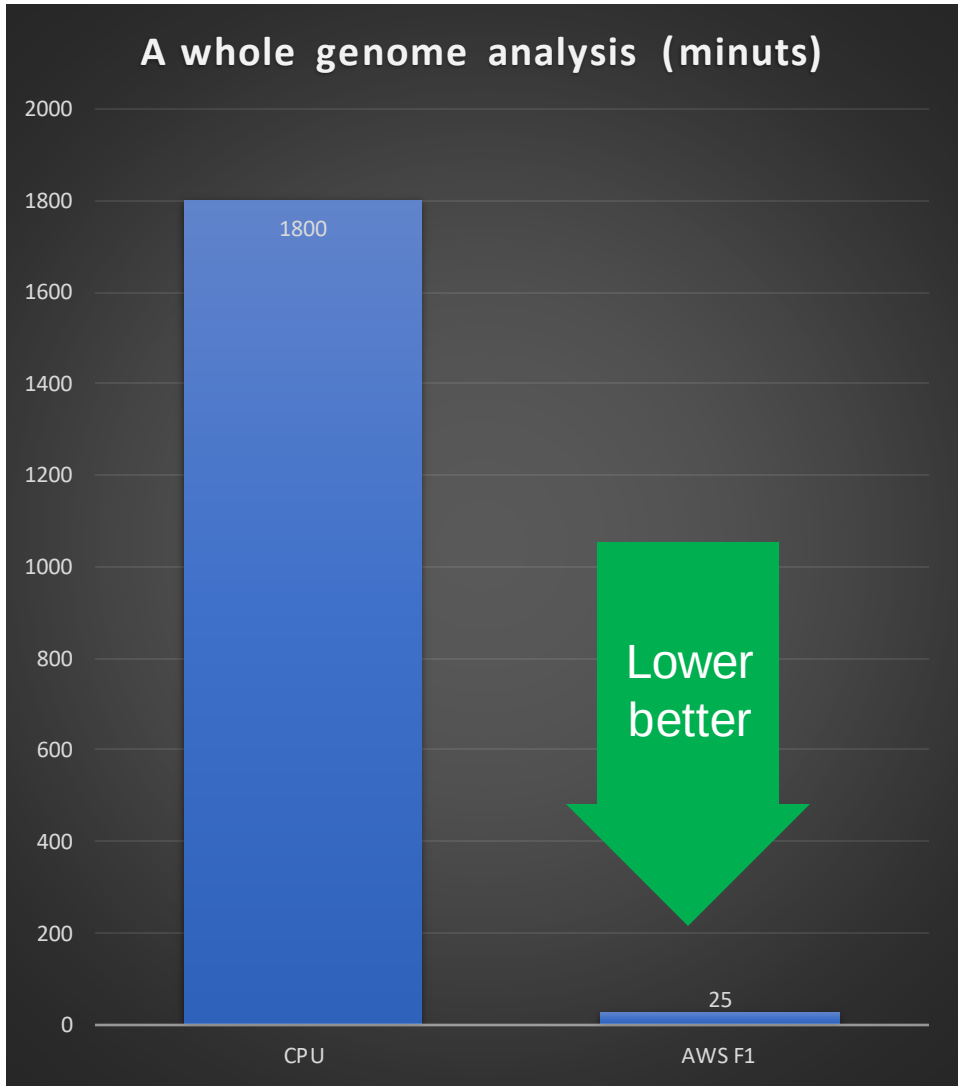


XILINX.

Genomic Sequencing

illumina®

- Reduced a whole genome sequencing time from 30 hours to 25 min
- Framework: BWA-GATK



XILINX.

Conclusion

- Workload-optimized Domain Specific Architecture (DSA) is critical to achieve high performance in today's demanding compute needs
- Innovation is outpacing silicon pace in many domains such as AI, big data and 5G
- Xilinx platform allows DSA to be developed in months on an existing silicon, not years to develop a new chip
- Wide range of benchmarks prove major advantages over CPU and GPUs

Thank You