AMD
XILINX

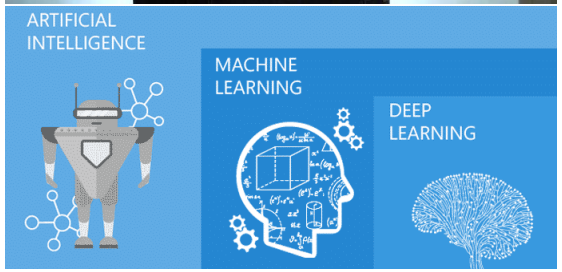
# AMD new ACAP Architecture enables close to Zero Dark Silicon

Jens Stapelfeldt – Sr. DBM
Data Center AI & Compute Markets

Jens.Stapelfeldt@amd.com
LinkedIn: www.linkedin.com/in/JensStapelfeldt

# Jens Stapelfeldt TSL / AI BDM – AMD / Xilinx

Jens.Stapelfeldt@amd.com
LinkedIn: www.linkedin.com/in/JensStapelfeldt


## Jens technical background:

- 1996 /1998 - 3 Y ASIC designer
- 1998 - 12 Y Business Manager CE and Trainer for Doulos in CE and ARM ATC WW
  - Design methodology (Verification, SystemC,..)
  - (V)HDL, FPGA, ARM Architecture (ARM7 - Cortex-M to Cortex-A)
- 2010 - 5 Y Sr. Embedded FAE at TI
  - Sitara, DaVinci, OMAP, Keystone II
  - Industrial App. (IoT, Industrial Ethernet (EtherCAT, ProfiNet,..), Camera Vison ..)
- **Since Oct. 2016** Tech Sales Lead for **AMD/Xilinx in EMEA**
  - Supporting DACH, EE, Russia, Israel, India
  - 2019 BDM Data Center
  - TSL DACH, Nordic UK & Ireland
  - 2022 Sr. BDM Data Center AI & Compute Markets
- **2016-2018 Part time MBA** in Bremen with weeks in Hong Kong, Cambridge, Dublin Business school!
- **Since Oct. 2018 guest lecturer** for “Digital Signal Processing (VHDL / FPGA design”).

MBA: In May 2018 I finished my MBA with Master Thesis in International Marketing looking into about 600 AI/ML Start-ups in EMEA!
LinkedIn Article: https://www.linkedin.com/pulse/analysing-europes-aiml-start-up-landscape-using-jens-stapelfeldt/

# High Performance and Adaptive Computing

Cloud, Network, Hyperscale & Supercomputing

5G & Comms Infrastructure

AI & Analytics Everywhere

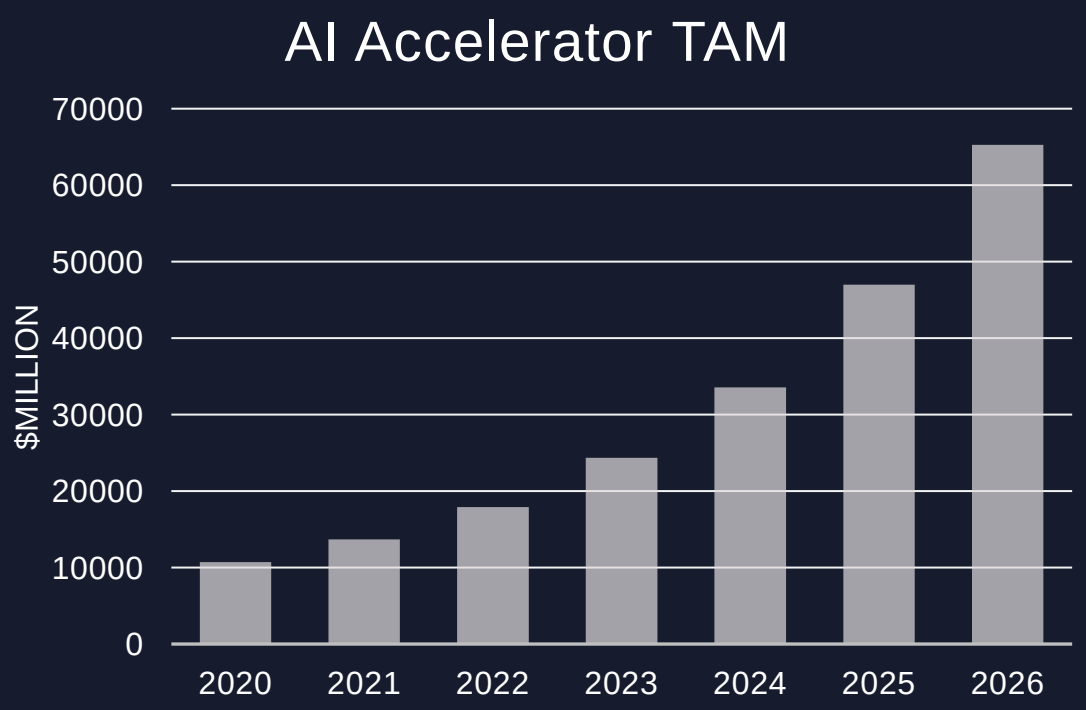
Adaptive Intelligent Systems

Gaming, Simulation and Visualization

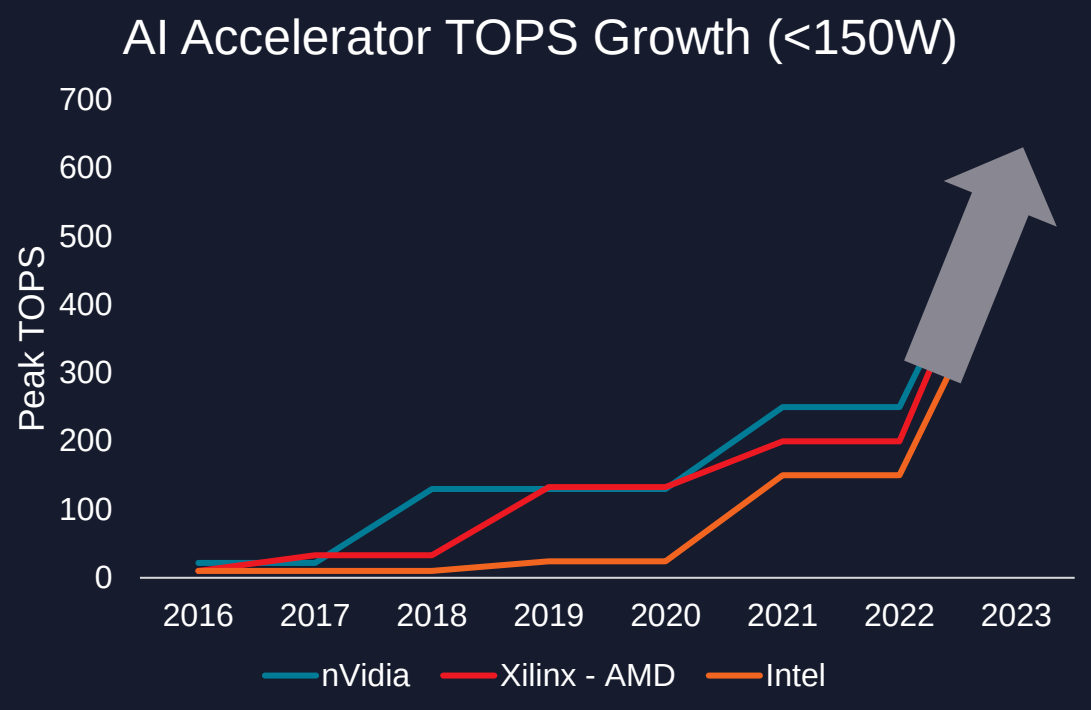
Smarter Client Devices & Edge

## At The Center of Today's Intelligent World

# AI Accelerator Demands Major Innovation




DC AI Capex Growing at 36.7% CAGR, projected to $65B TAM by 2026




AI Accelerator Peak TOPS Growing Exponentially to Keep Up with the Model Innovation



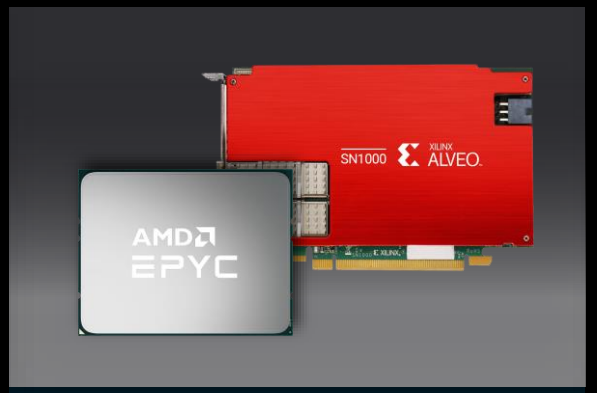
*Source: Markets and Markets (Data Center Accelerator Market Global Forecast to 2026)*

**2022-2023 is projected based on history*
*** Only including AI cards with 150W or less power consumption*

AMD
XILINX

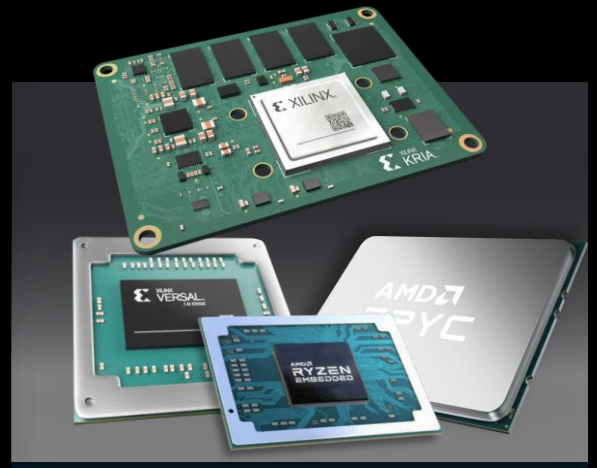
# AMD + XILINX AI OPPORTUNITIES


AI Inference and Training




Data Center and Communications


Automotive




Embedded

## - AI OPPORTUNITIES in all areas

# Adaptive Compute Acceleration Platform
## A New Device Category

### ADAPTIVE
- Adaptable to diverse workloads
- Future-proof algorithms

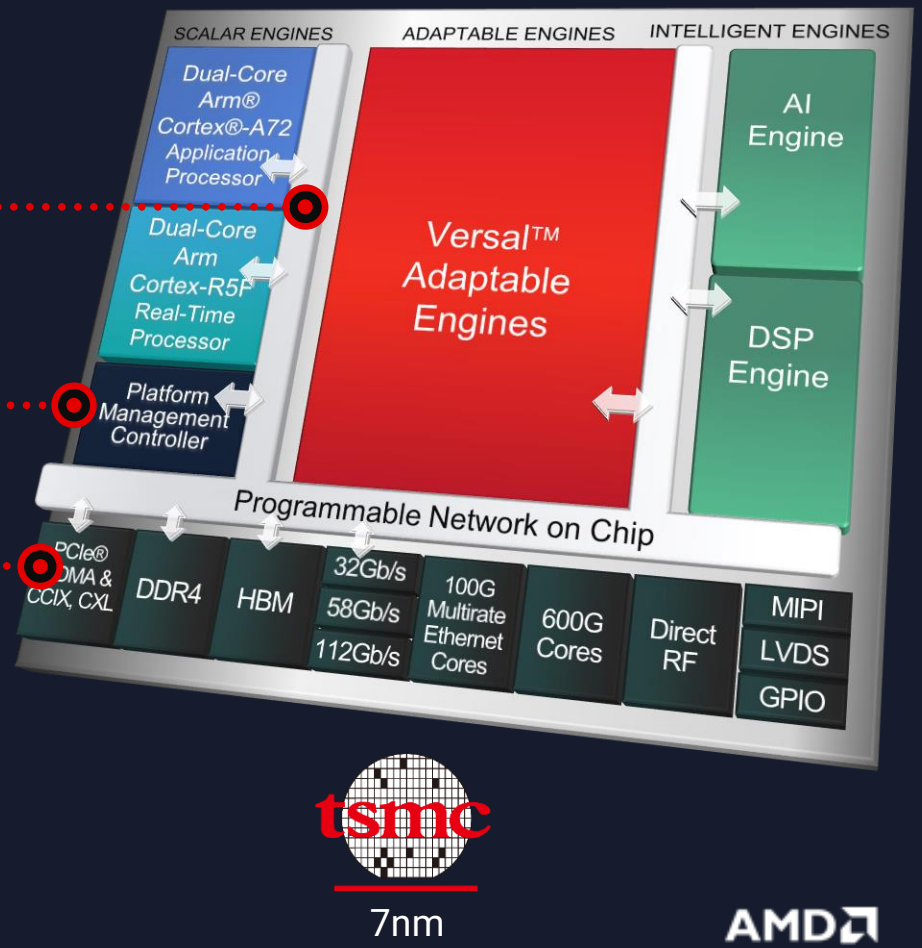
### COMPUTE ACCELERATION
- Scalar Engines
- Adaptable Engines
- Intelligent Engines

### PLATFORM
- SW programmable silicon infrastructure
- Pre-engineered connectivity
- Platform available at boot

Programmable Network on Chip

SW-Controlled Platform Management

Dedicated Interfaces for Compute (PCIe®, DDR4)




tsmc 7nm

AMD XILINX

# Intelligent Engines

*Massive AI Inference Throughput and Wireless Compute*

## 1.3GHz VLIW / SIMD vector processors

- Versatile core for ML and other advanced DSP workloads

## Massive array of interconnected cores

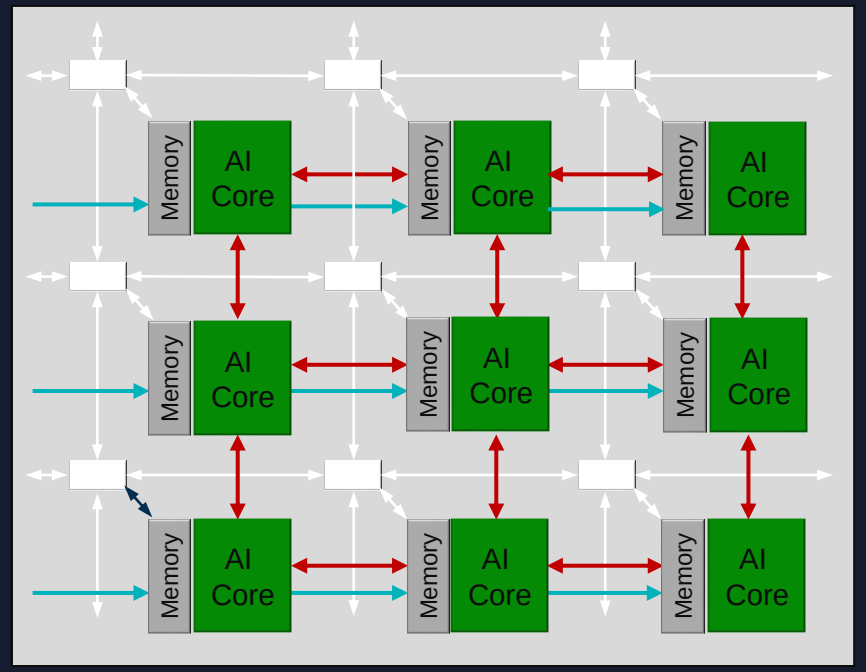
- Instantiate multiple tiles (10s to 100s) for scalable compute

## Terabytes/sec of interface bandwidth to other engines

- Direct, massive throughput to adaptable HW engines
- Implement core application with AI for “Whole App Acceleration”

## SW programmable for any developer

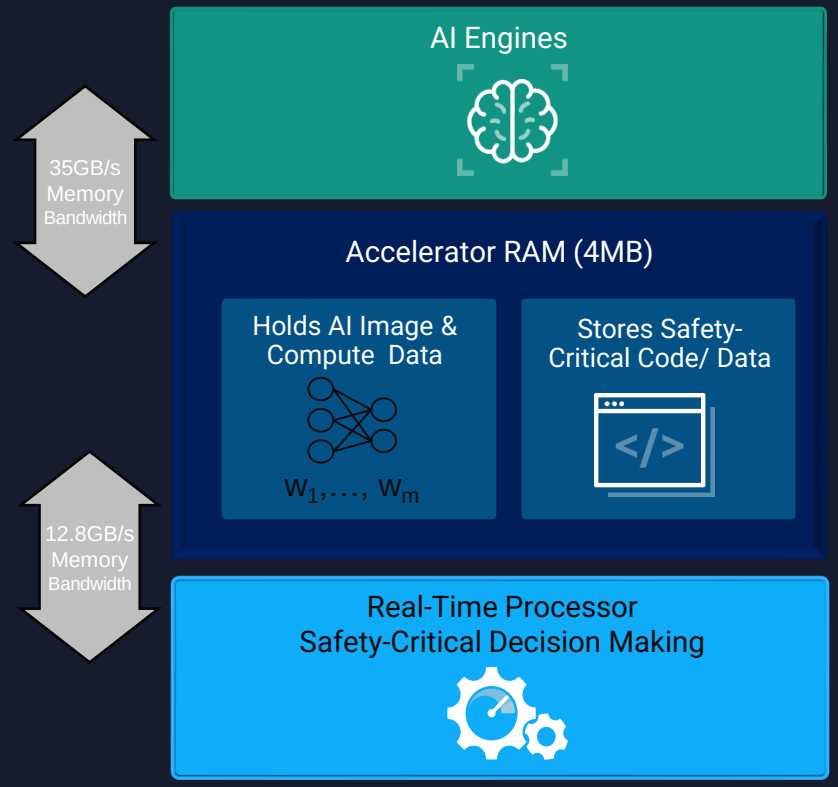
- C programmable, compile in minutes
- Framework-based design for ML developers



# Adaptable Memory Architecture - Accelerator RAM

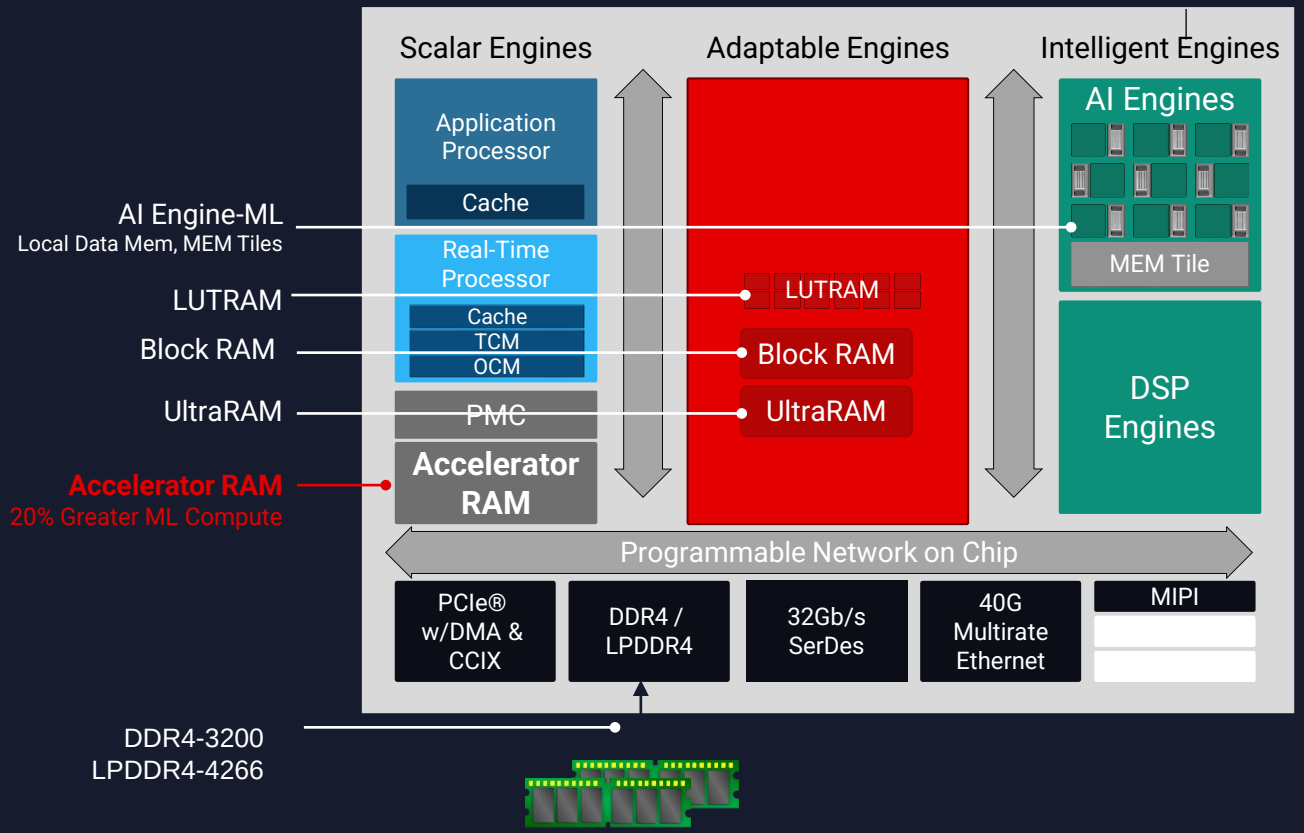
## For ML Compute and Safety-Critical Operation

Avoid DDR to store compute data, safety-critical code/data




## Part of the Adaptable Memory Hierarchy

Select the right memory for bandwidth requirement

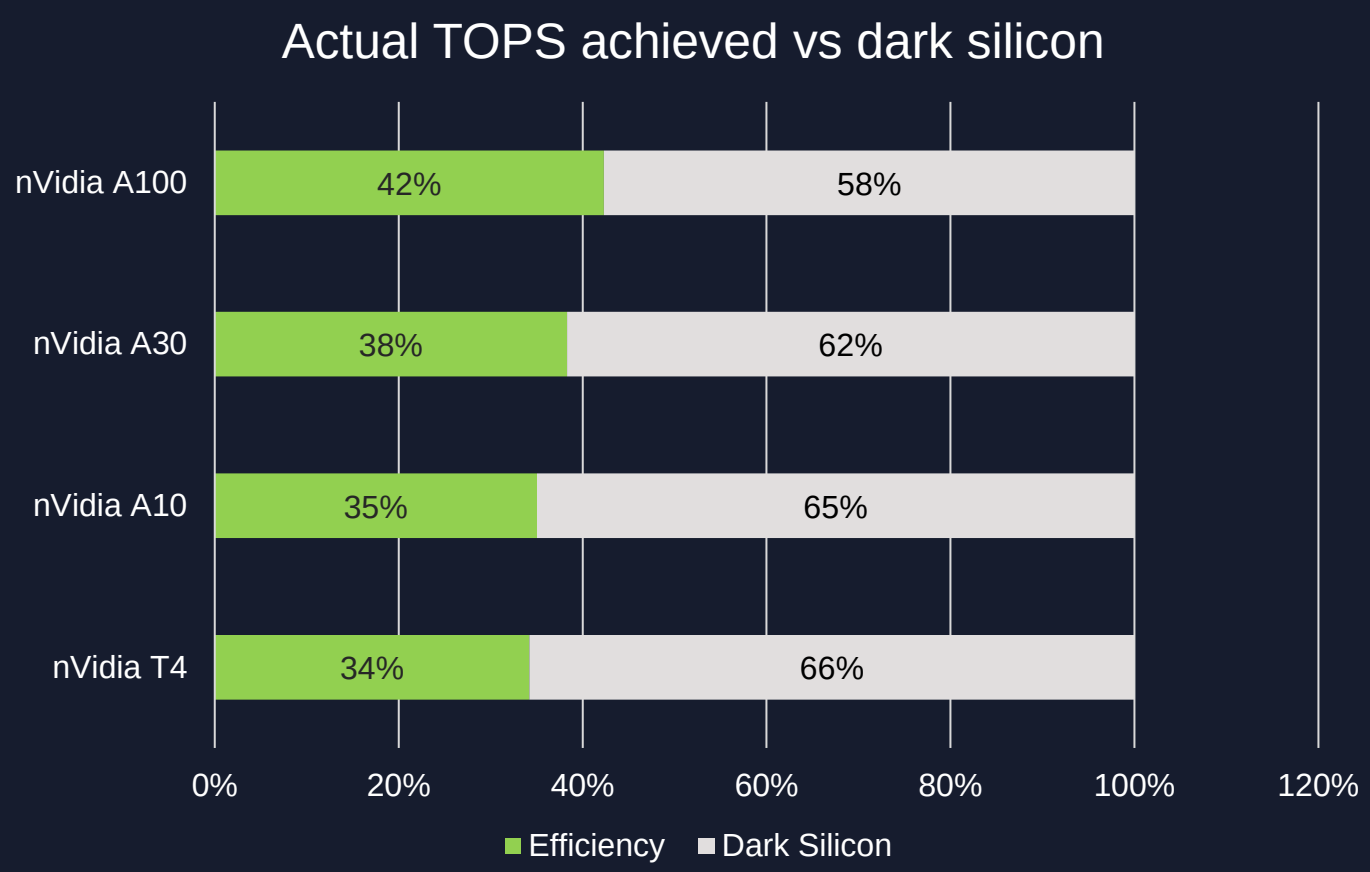







XILINX.

# AI Chips Claim large TOPS, but with Two-Thirds “Dark Silicon”




| | ResNet-50 (img/s) | Peak TOPS | Actual TOPS |
|---|---|---|---|
| A100 | 32,204 | 624 | 264 |
| A30 | 15,411 | 330 | 126 |
| A10 | 10,676 | 250 | 88 |
| T4 | 5,423 | 130 | 44 |

The market flagship barely achieves 40% in the most basic AI benchmark

*Source: https://developer.Nvidia.com/deep-learning-performance-training-inference*

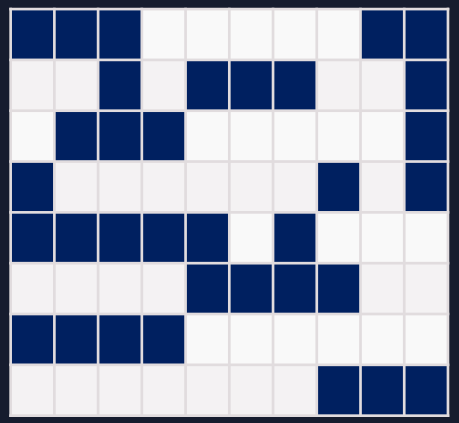


AMD XILINX

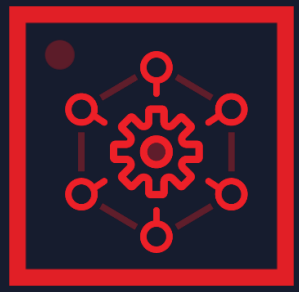
# Fixed Hardware Cannot Pump Data Fast Enough

Memory

”Data Bubbles”



**Fixed** Data Mover



**Fixed** AI Processor

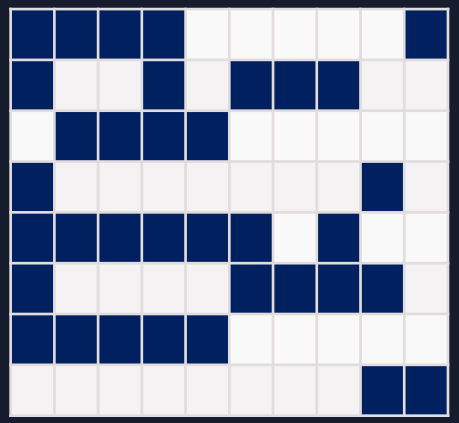


AMD XILINX

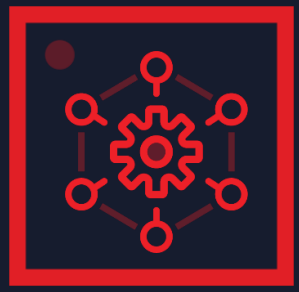
# Fixed Hardware Cannot Pump Data Fast Enough

Memory

”Data Bubbles”



**Fixed** Data Mover

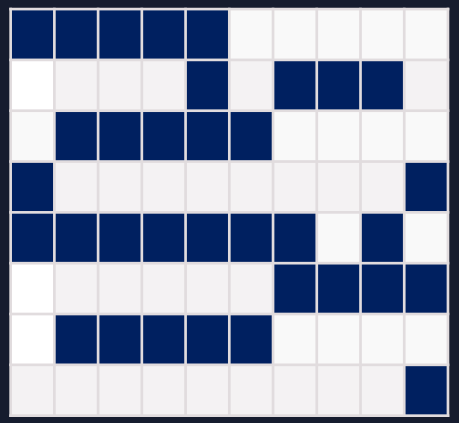


**Fixed** AI Processor

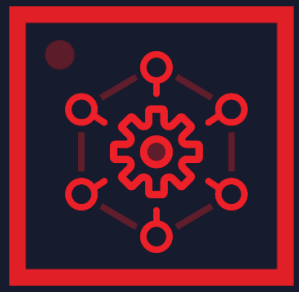
# Fixed Hardware Cannot Pump Data Fast Enough

Memory

”Data Bubbles”

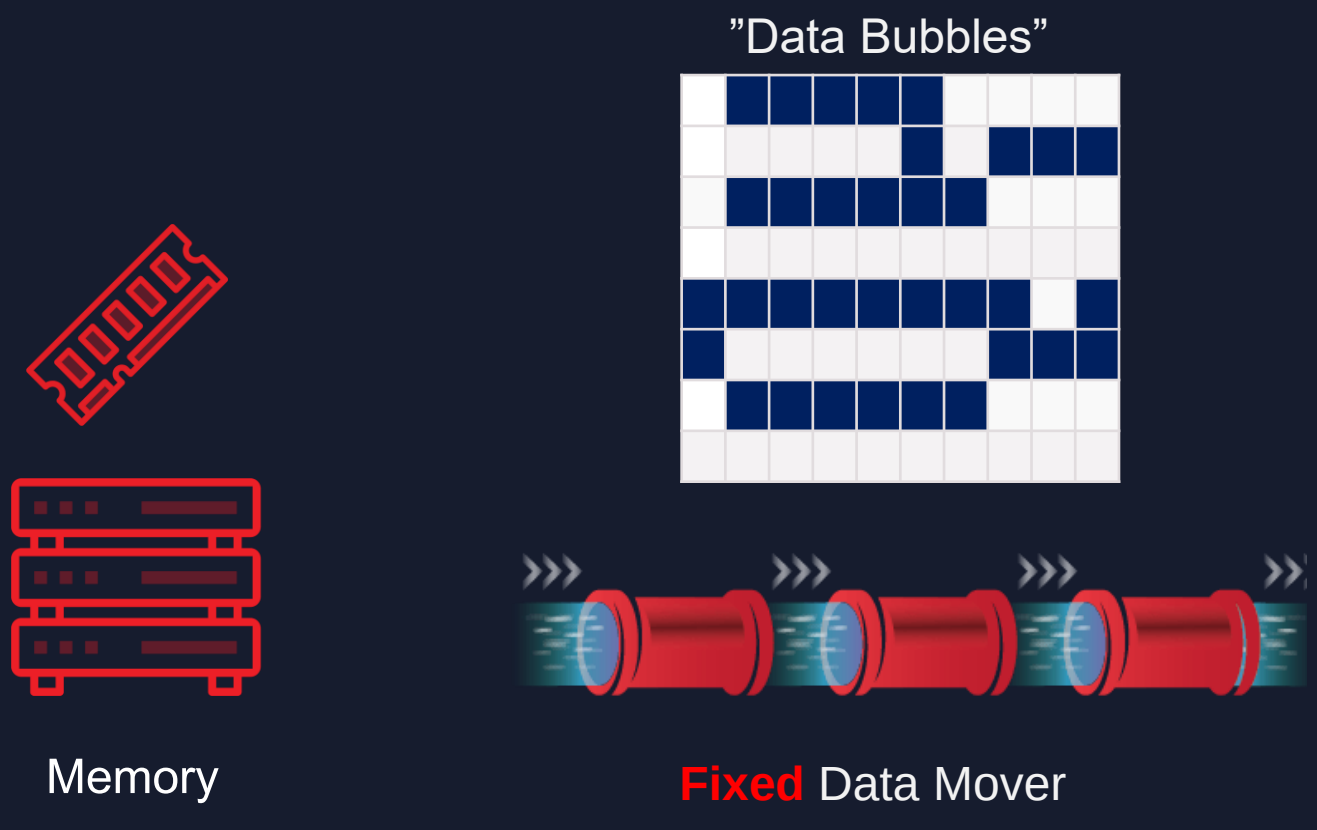


**Fixed** Data Mover

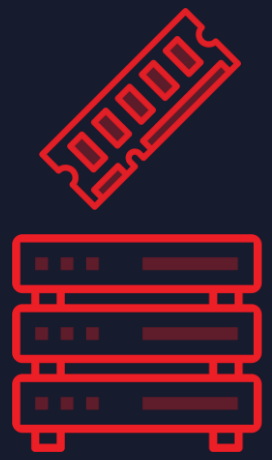


**Fixed** AI Processor

# Fixed Hardware Cannot Pump Data Fast Enough

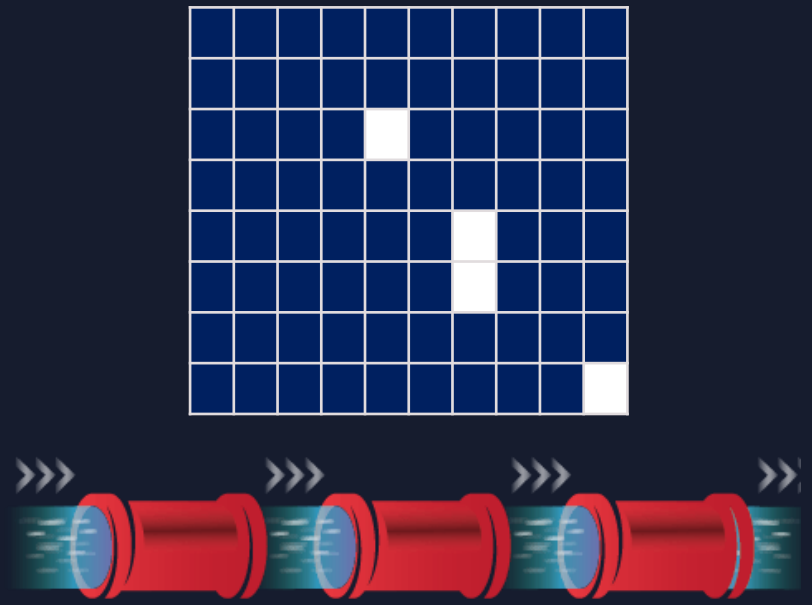


# Xilinx + AMD: Adaptable Hardware to Remove “Data Bubble”

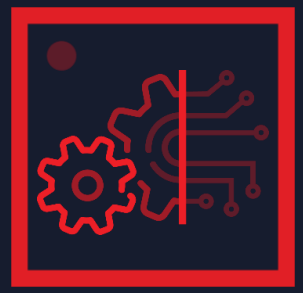


Memory

”Data Bubbles”

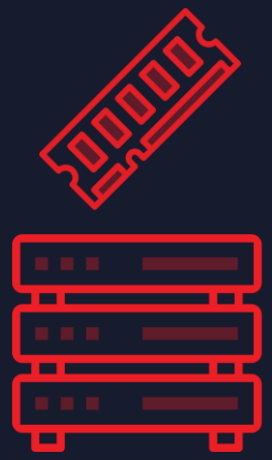


**Adaptable** Data Mover



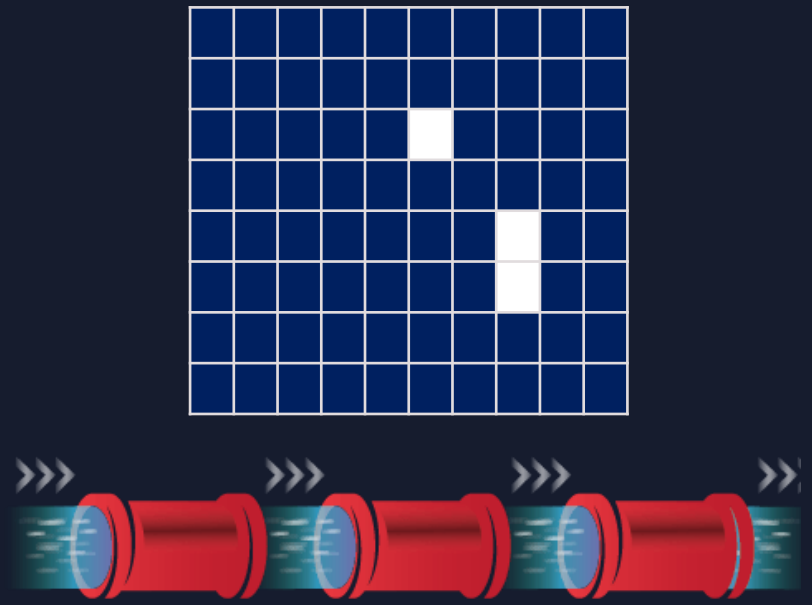
**Adaptable** AI Engine

# Xilinx + AMD: Adaptable Hardware to Remove “Data Bubble”

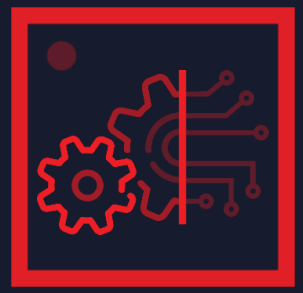


Memory

”Data Bubbles”



**Adaptable** Data Mover

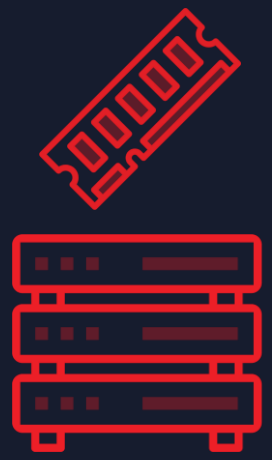


**Adaptable** AI Engine



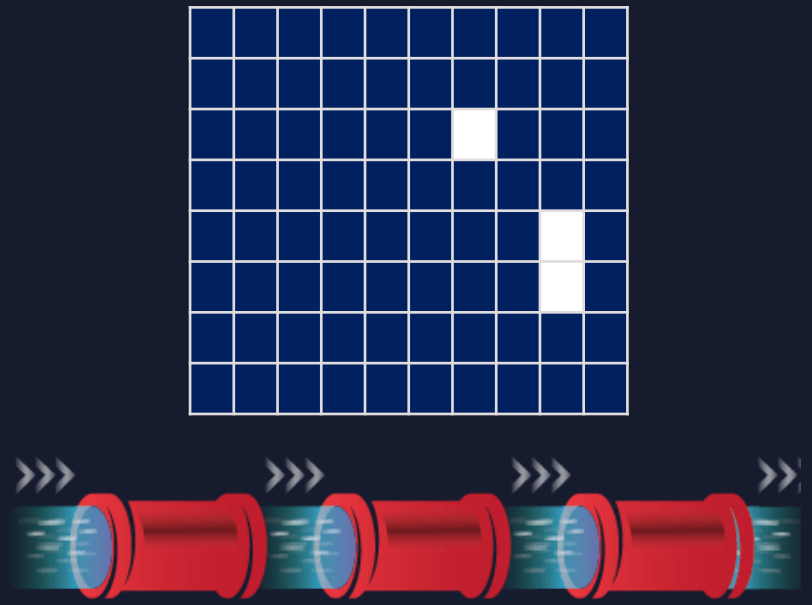
AMD XILINX

# Xilinx + AMD: Adaptable Hardware to Remove “Data Bubble”

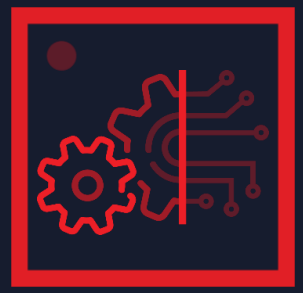


Memory

”Data Bubbles”

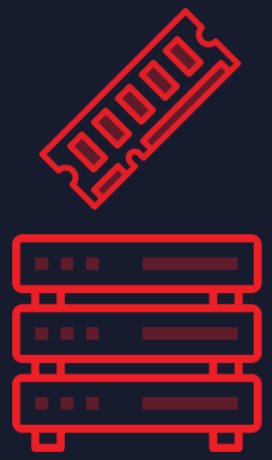


**Adaptable** Data Mover



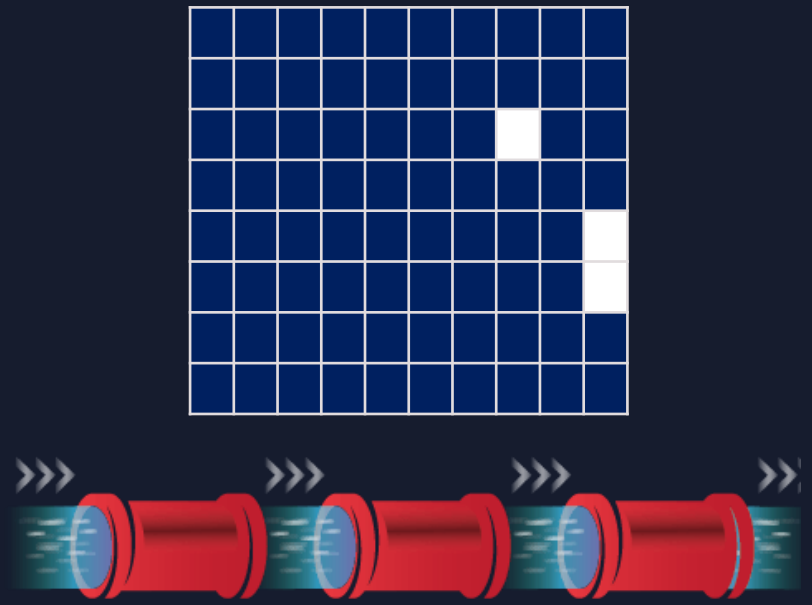
**Adaptable** AI Engine

# Xilinx + AMD: Adaptable Hardware to Remove “Data Bubble”

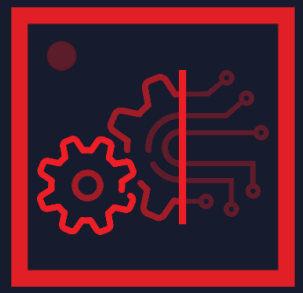


Memory

”Data Bubbles”

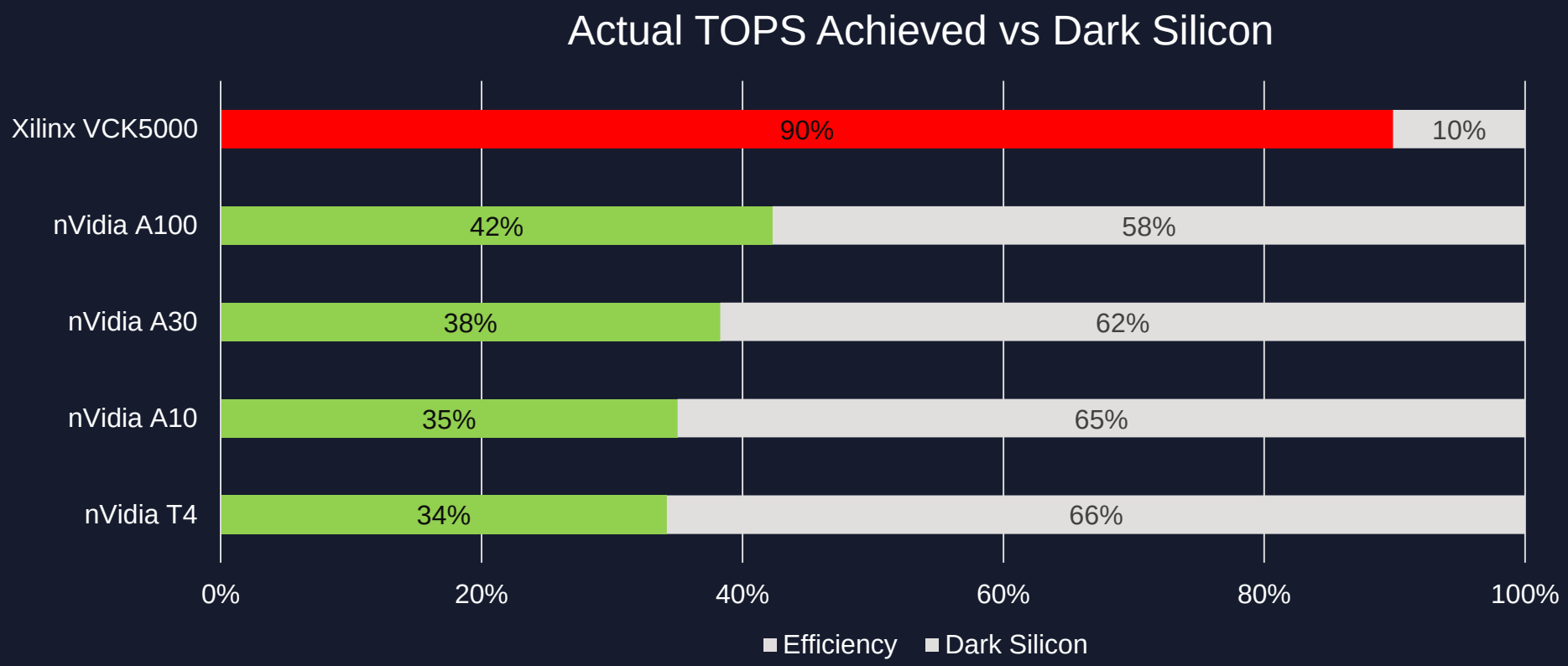


**Adaptable** Data Mover



**Adaptable** AI Engine

# The World's First "0 Dark Silicon" AI Accelerator

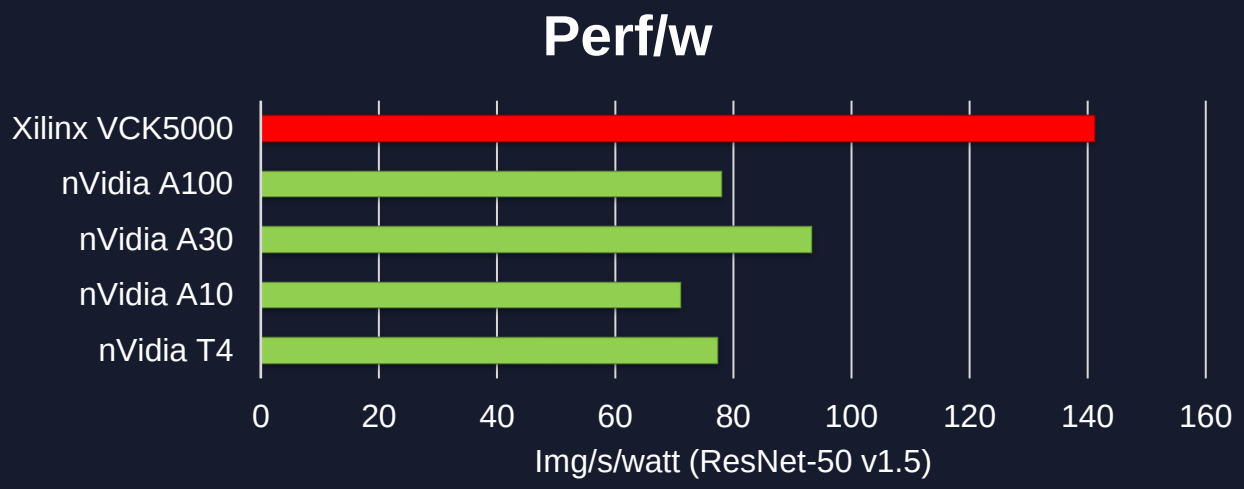


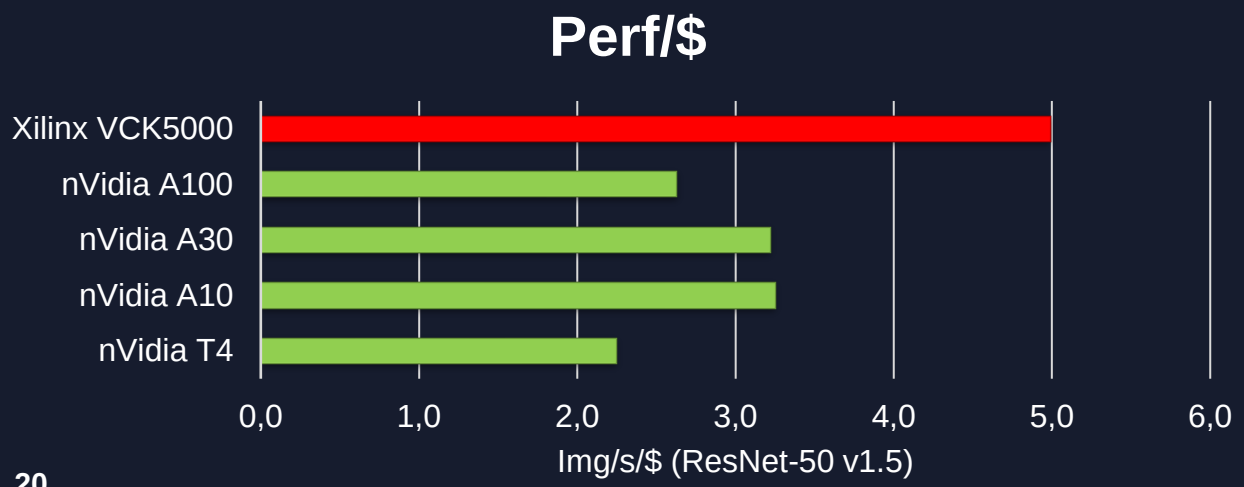

Near 100% efficiency: Achieving True Peak TOPS at Real AI Model Workloads



AMD
XILINX

# Double Performance / Watt / $ vs Nvidia Flagship AI Cards




| | ResNet-50 (img/s) | Power | SRP** |
|---|---|---|---|
| VCK5000 | 13,700 | 97W | $2,745 |
| A100 SXM | 32,204 | 413W | $12,235* |
| A30 | 15,411 | 165W | $4,787 |
| A10 | 10,676 | 150W | $3,283 |
| T4 | 5,423 | 75W | $2,410 |

* A100 SXM pricing not available, using A100 PCIe 80GB pricing instead. SXM price is typically more expensive than PCIe
** SRP captured from acmemicro.com as of Feb 22, 2022



AMD
XILINX

# Real AI Deployment > Just AI

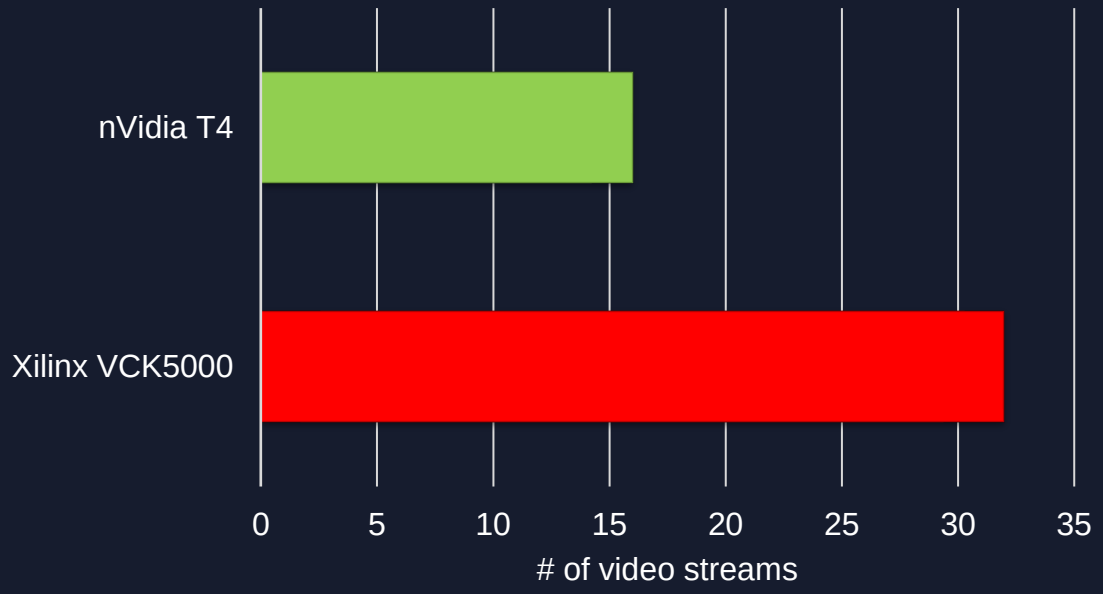
*100s of AI, Image Manipulators, Video Decode Simultaneously*




AMD
XILINX

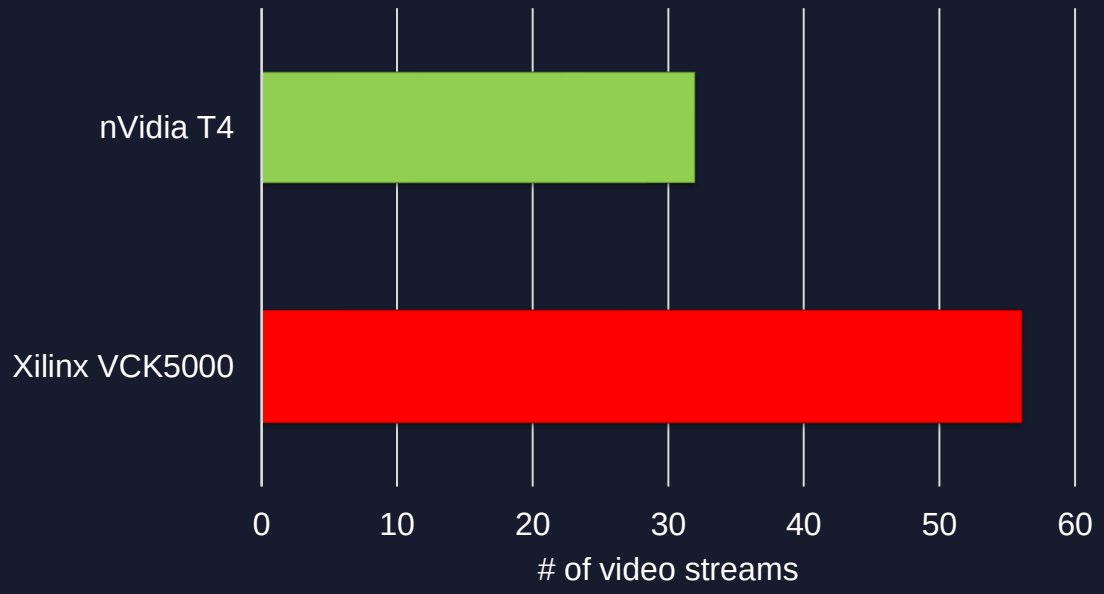
# AMD Epyc + Xilinx VCK5000 Deliver 2x TCO over Nvidia T4

## ML-heavy pipeline




H.264 Decode + Yolov3 + 3x ResNet-18

## Video-heavy pipeline




H.264 Decode + tinyYolov3 + 3x ResNet-50



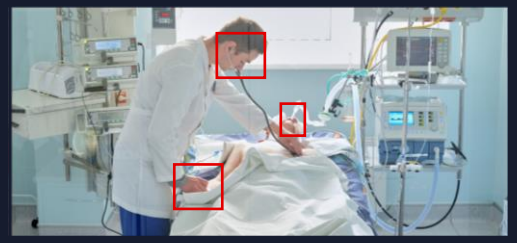
AMD
XILINX

# Target Applications

## SMART CITY

- Accident/ Emergency detection
- Safety- mask, social distancing
- Facial recognition
- Crowd detection and people counting
- Parking monitoring
- Vehicle detection and classification
- Number plate recognition

## SAFETY CRITICAL INFRASTRUCTURE

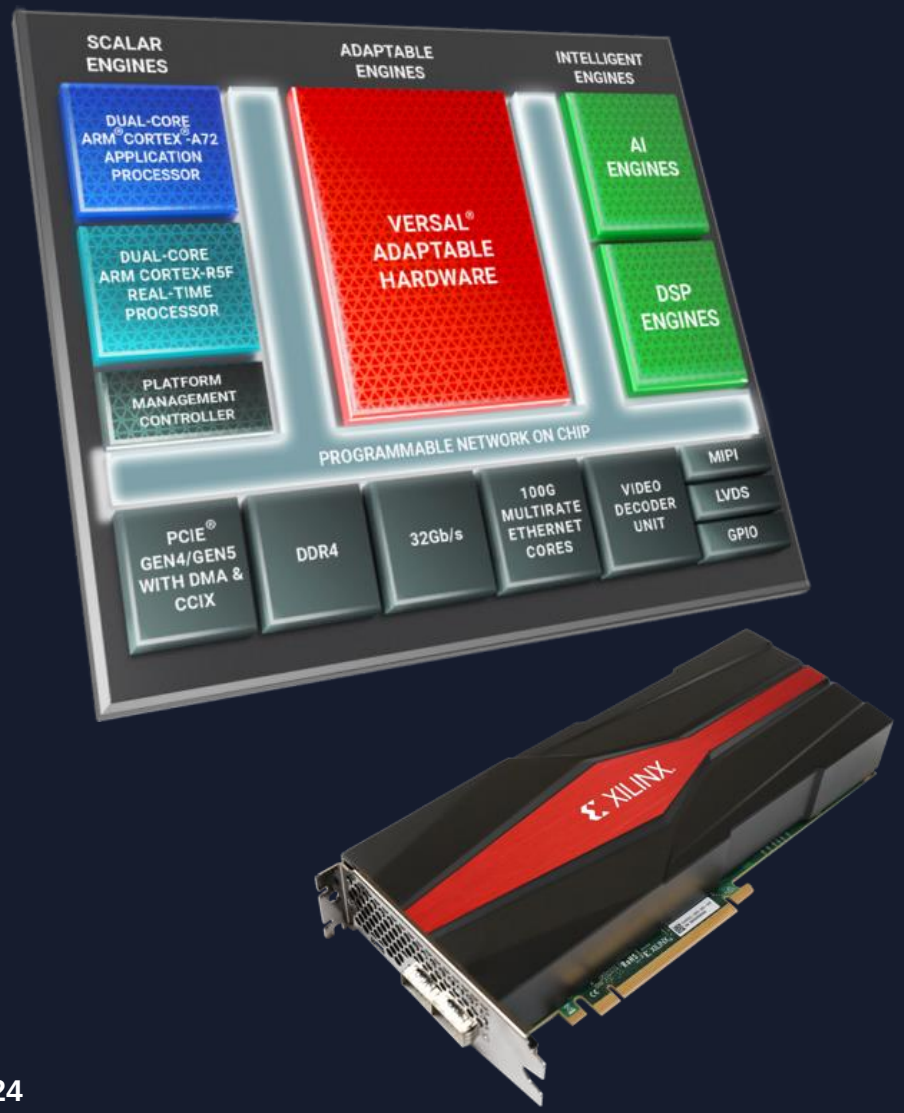


- Patient Monitoring
- Fall/ Emergency detection
- People tracking
- Weapon/ Explosive detection
- Fire detection
- Video anomaly detection
- Virtual fencing
- Worker/ Industrial safety

## SMART RETAIL

- Shrinkage reduction
- Retail demographics
- Object detection and classification
- Emotion detection
- Planograms

# VCK5000 Product Overview




| Silicon | 7nm Versal ACAP |
|---|---|
| Peak TOPS (INT8) | 125* |
| SRP | $2,745 |
| Form Factor | FH3/4L Dual |
| PCI Express | Gen3 x16 / Gen4 x 8 |
| TDP | 75, 150, 225W |
| Off-chip Memory | DDR4 16 GB |
| Internal SRAM | 36.4 MB |
| FPGA Logic (LUTs) | 900K |

*AI Engines running at 1.25GHz



AMD XILINX

# Data Center AI Success Stories

**Alibaba**
E-commerce Business


**SK Telecom**
Home Security


**Kuaishou**
Automatic Speech Recognition

# Availability



VCK5000 on Xilinx.com
**$2,745**

Try TensorFlow direct inference
Docker by Mipsology

Try Video Analytics pipeline
Docker by Aupera

# VCK5000 on the Cloud Now Live – VMACCEL

AMD XILINX


**5 Free Evaluation Hours in the Cloud**

Experience the ease of use and flexibility of best-in-class compute acceleration for deploying real-world AI applications. Access the Zebra® software in a few fast and simple steps and run on the VCK5000 Versal® development card in the cloud.
**>> Try direct inference from TensorFlow / PyTorch**

Rapidly build, configure, and deploy computer vision pipelines using a graphical user interface (GUI); with no coding. Run the Aupera Video Machine Learning Streaming Server Solution on VCK5000 hosted on VMAccel® FPGA cloud.
**>> Try a full AI video analytics pipeline**

| | ResNet-50 (img/s) | Video Pipeline (# of streams) |
|---|---|---|
| VCK5000 | 8,400 | 55 |
| Nv A2 | 2,499 | 16 |
| Nv T4 | 5,423 | 32 |

Try Dockers for AI software evaluation (VCK5000)

- CNN inference: https://www.vmaccel.com/zebrademo
- Video analytics pipeline: https://www.vmaccel.com/vmssdemo

XILINX

# Executive Summary – What's New





First Versal Data Center Card: VCK5000

Improved 3x Performance

World's First 0 Dark Silicon in AI Inference

2x TCO vs Nvidia Ampere

Video Analytics SDK available on Versal

2x Nvidia T4 in video stream throughput

Seamless TensorFlow / Pytorch inference

Video analytics plugin



AMD XILINX


# Thank You

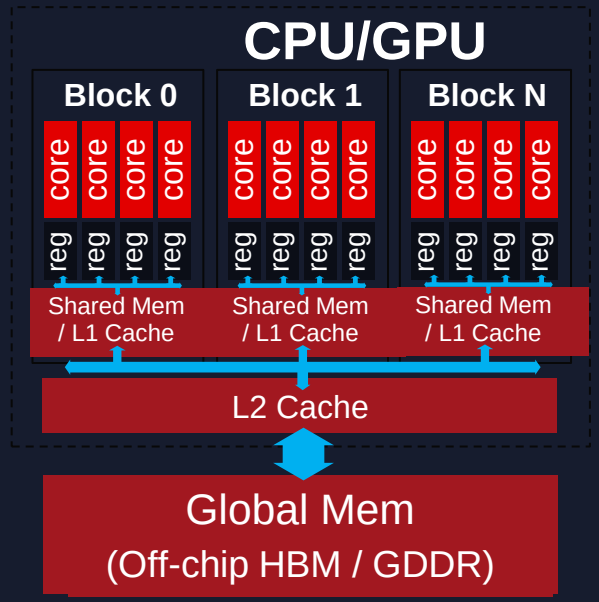
# Xilinx - Flexible and Adaptable Memory Hierarchy is the Key!

Latency : Power

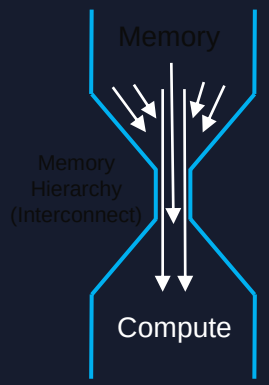
1X 1X

2X 10X

80X 100X

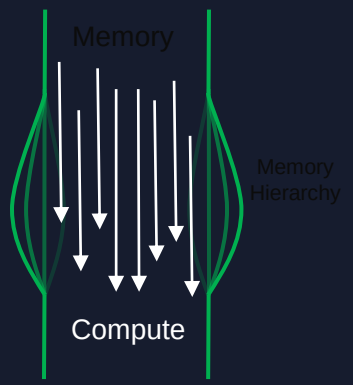


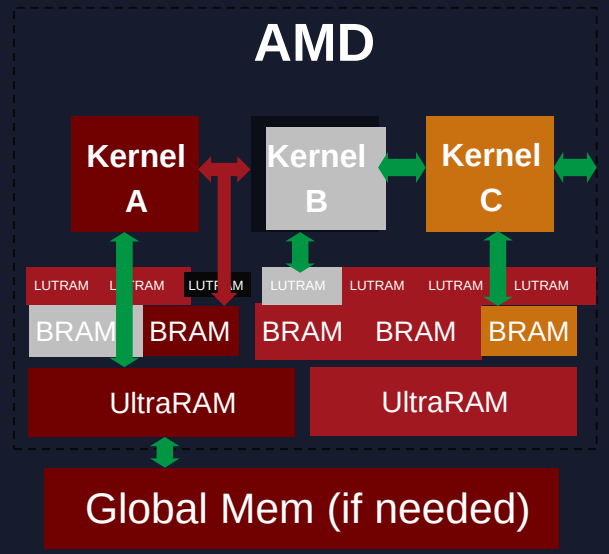

Fixed




- Rigid memory hierarchy & SW defined dataflow
- High “data locality” required for workload efficiency
- “Batching” improves efficiency at expense of latency

Adaptable



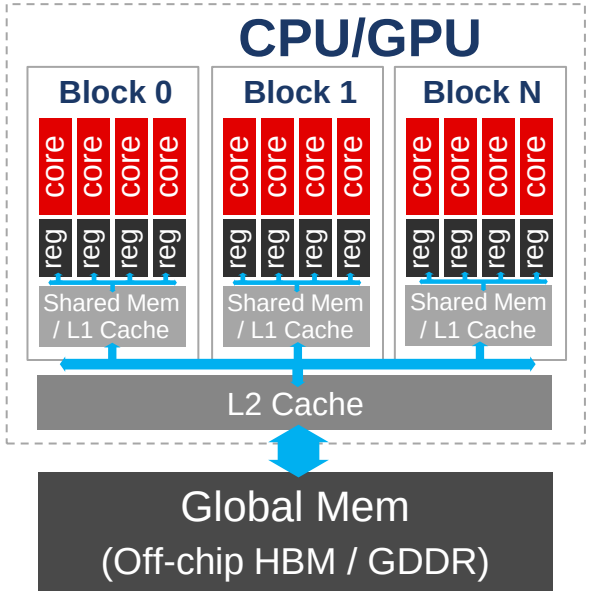

- **Adaptable memory hierarchy & datapath**
- **~5X more on-chip memory**
- **Max throughput, min latency – no batching required**

## Resultant Workload Speed-Up vs. CPU / GPU

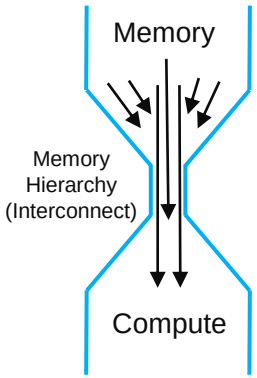
| DB - SQL | DB - RegEx | ML Infer | Genomics | Video |
|---|---|---|---|---|
| 4X | 3X | 3X | 6X | 6X |

# Xilinx - Flexible and Adaptable Memory Hierarchy is the Key!

| Latency | Power |
|---|---|
| 1X | 1X |
| 2X | 10X |
| 80X | 100X |

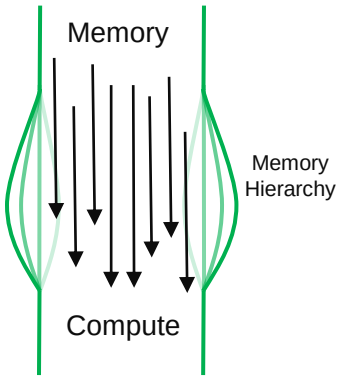


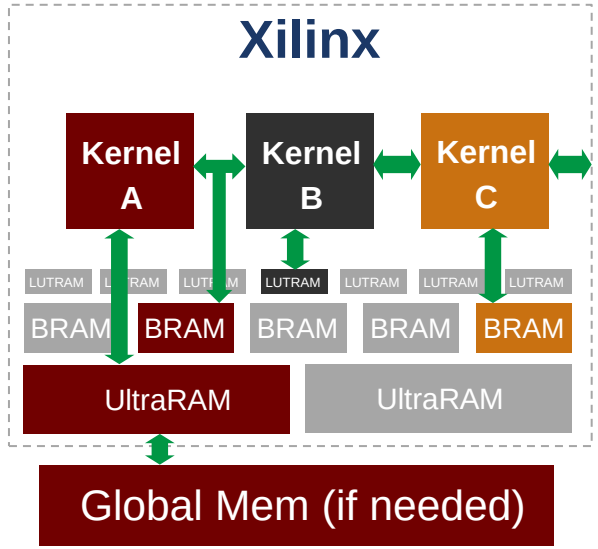

## Fixed




- Rigid memory hierarchy & SW defined dataflow
- High “data locality” required for workload efficiency
- “Batching” improves efficiency at expense of latency

## Adaptable



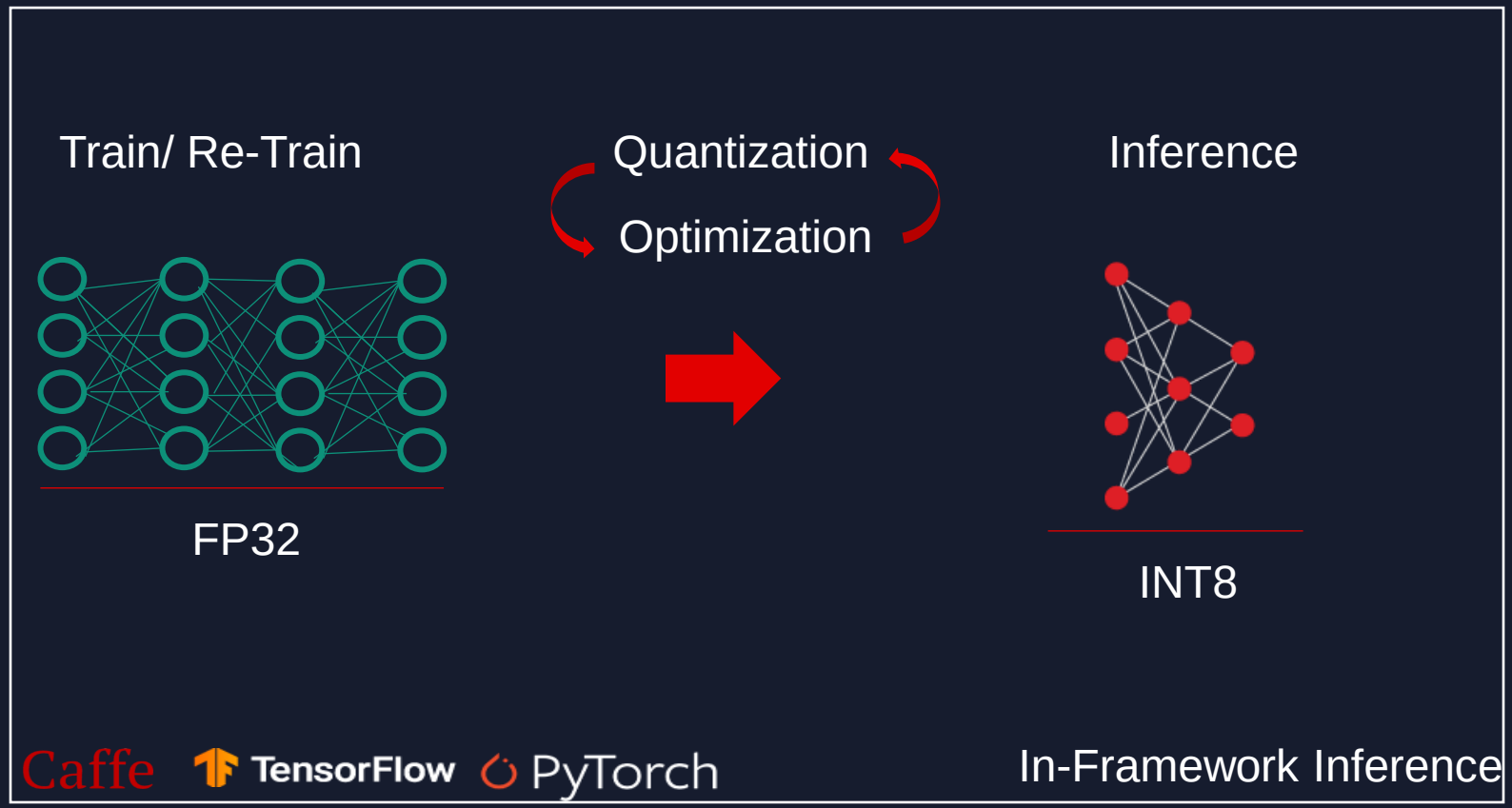

- **Adaptable memory hierarchy & datapath**
- **~5X more on-chip memory**
- **Max throughput, min latency – no batching required**

## Resultant Workload Speed-Up vs. CPU / GPU

| DB - SQL | DB - RegEx | ML Infer | Genomics | Video |
|---|---|---|---|---|
| 4X | 3X | 3X | 6X | 6X |

# Full Solution Stack- 3 Steps To Deployment

1 Import Models



# Full Solution Stack- 3 Steps To Deployment

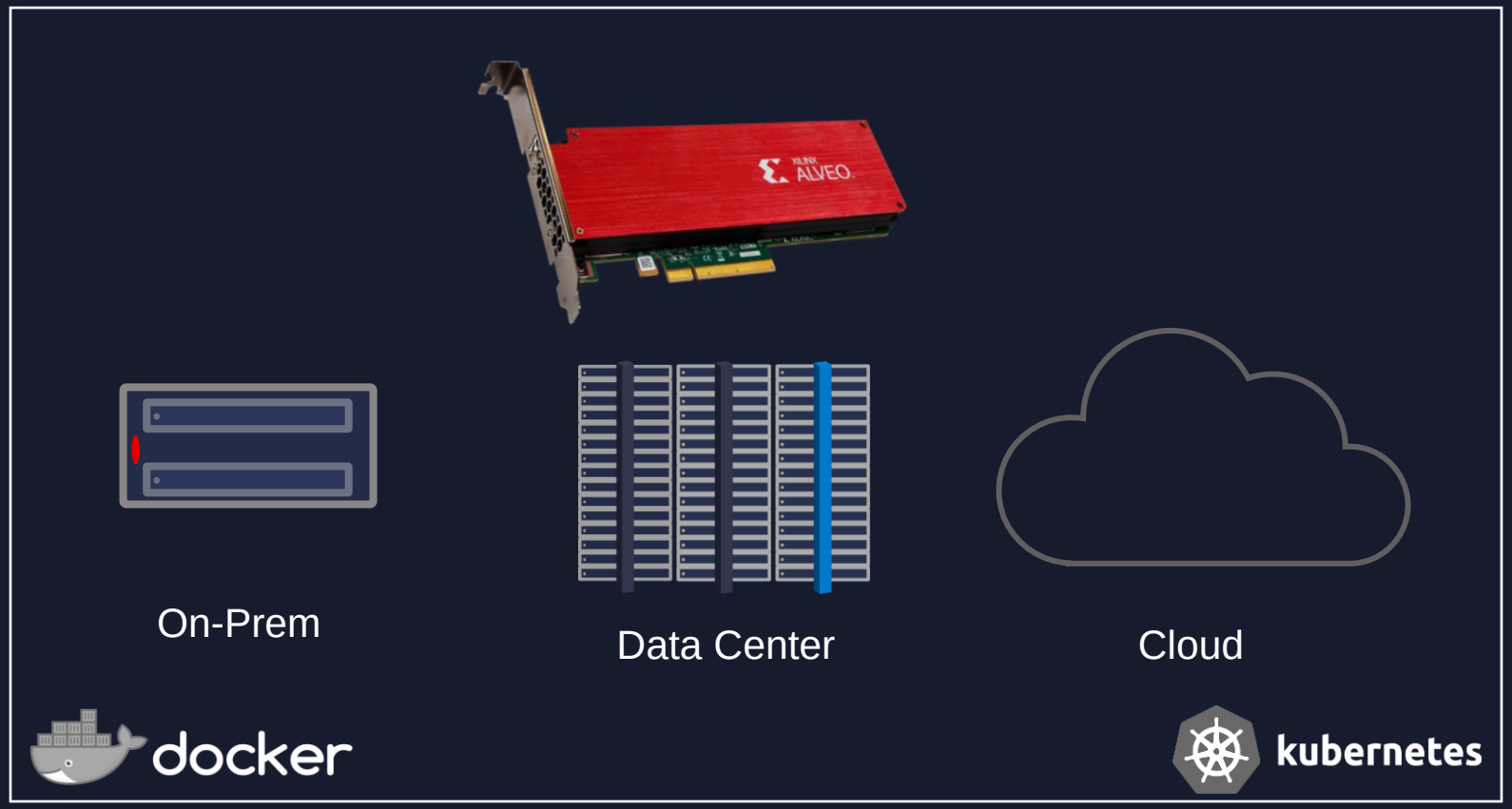
2 Configure Whole Application Pipeline




XILINX.

# Full Solution Stack- 3 Steps To Deployment

3 Deploy in production






XILINX