Alveo™ Accelerator Card

Accelerated AI and High Performance Computing Workloads with Adaptive Computing Solutions

Alveo™ Accelerator Card

Designed to meet the constantly changing needs of the modern Data Center, providing up to 90X performance increase over CPUs.

Applications

- **Computational Storage**
- **Database and Data Analytics**
- **Financial Technology**
- **High Performance Computing**
- **Network Acceleration**
- **Video and Imaging**
- **Machine Learning**
- **Tools and Services**

Figure 1: Adaptable Acceleration for Dynamic Workloads

Unified Development Platform

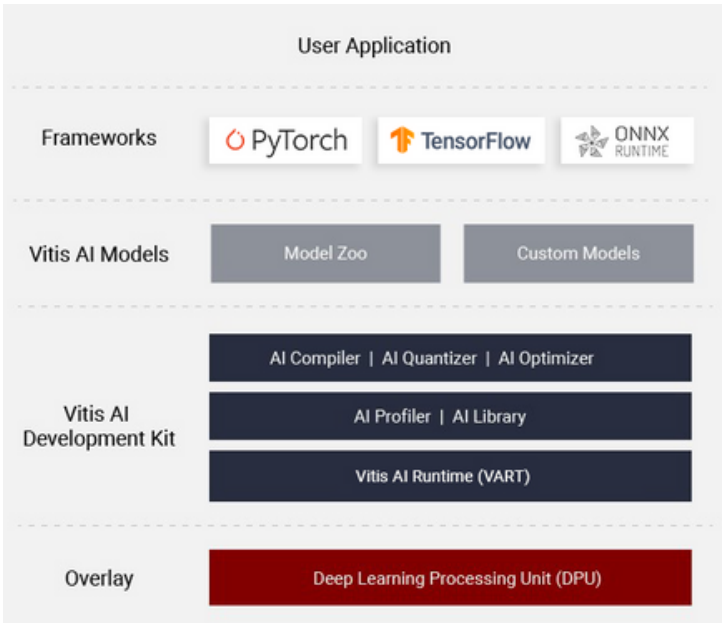


Figure 2: Vitis AI Development Platform

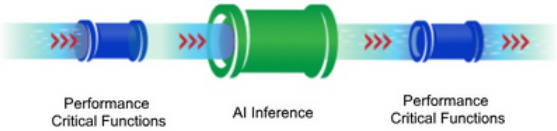
Key Advantages

Fast & Efficient
Highest Performance

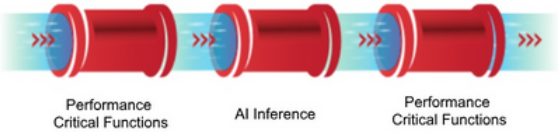
Adaptable
Accelerate Any Workload

Accessible
Cloud ↔ On-Premises Mobility

CPU/GPU Bottleneck



Alveo Breakthrough Performance



89/13 Amornpan 205 Tower 1, 11th floor, Ratchadapisek 7 (Nathong) Alley, Ratchadapisek Road, Din Daeng, Bangkok, 10400 THAILAND
Contact us hpc-sales@design-gateway.com www.design-gateway.com (+66)2-103-4327
DesignGateway DGIPcore @design_gateway www.dgway.com/blog_E/

ALVEO™ V70 AI Inference Accelerator Card

High Performance & High Efficient AI Accelerator Card

OVERVIEW

The Alveo™ V70 accelerator card is the first AMD Alveo production card leveraging AMD XDNA™ architecture with AI Engines providing a tightly integrated heterogeneous compute platform for CNN, RNN, and NLP acceleration targeting cloud and edge applications. V70 is designed to be the most energy efficient AI Inference card in the AMD portfolio tuned for video analytics and natural language processing workloads and offers industry standard framework support, directly compiling models trained in TensorFlow and PyTorch. The card is a PCIe®-based half height, half length, single slot card that supports passive cooling for closed-loop thermal control in the server PCIe expansion slot. The card is equipped with a 7nm Versal® ACAP device which has an integrated AI Engine core to complement adaptable and scalar engines and 16 GB of DDR4 memory. Providing low power and a low-profile form factor, the V70 helps reduce cost per AI channel and provides high channel density for video applications.

AMD ALVEO™ V70

AI inference accelerator

AMD XDNA AI Engine

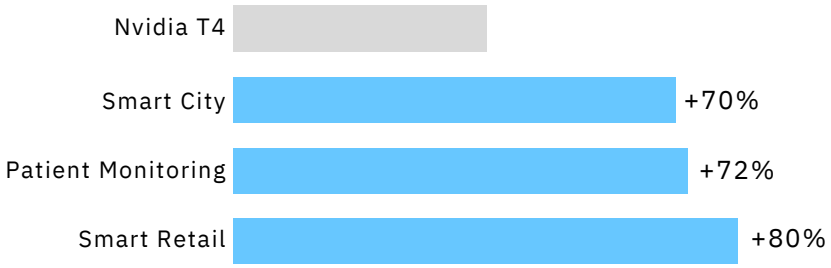
400 TOPS of AI compute

PCIe® 5.0 | **75W TDP**

Cloud-to-Client symmetry for AI developers

AMD ALVEO™ V70 Accelerator Card

Leadership AI inference performance at 75Watts!



Comparison	AMD Alveo V70	NVIDIA A30	NVIDIA A100	NVIDIA H100
TOPS*(INT8)	404	330	624	3,958
TOPS*(BF16)	202	165	312	1,979
Memory size	DDR4: 16GB FPGA: BRAM 21Mbit / URAM 74Mbit	24GB HBM2	80GB HBM2e	80GB HBM3
Memory bandwidth	76.8 GB/s (DDR4) 47.6 TB/s (FPGA)	933GB/s	1,935GB/s	3.35TB/s
Form Factor	Single slot(half size)	Double slot	Double slot	Double slot
Power(TDP) Watts	75W	165W	300W	700W
TOPS(INT8)/Watt Performance	5.38	2.00	2.08	5.65

89/13 Amornpan 205 Tower 1, 11th floor, Ratchadapisek 7 (Nathong) Alley, Ratchadapisek Road, Din Daeng, Bangkok, 10400 THAILAND
Contact us hpc-sales@design-gateway.com www.design-gateway.com (+66)2-103-4327
DesignGateway DGIPcore @design_gateway www.dgway.com/blog_E/