# ALVEO™ V70 AI INFERENCE ACCELERATOR CARD

## OVERVIEW

With intelligent connected devices in our homes, cars, offices, factories, cities, and in the cloud, the cost for this proliferation of AI enabled applications is an exponential increase in the data-processing and energy efficiency requirements placed on the chips powering these devices. The challenge is not just how to deploy the AI model, but how to deploy the AI application most efficiently. The best implementation of an AI application doesn't need to be the fastest, it needs to be the most efficient, yet remain flexible. AMD XDNA architecture built on adaptive computing platforms provide the best of both worlds for AI inference workloads – from cloud, edge or endpoint.

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.


## TARGET APPLICATIONS

### KEY VALUE PROPS

- Helps reduced cost per AI channel
- Entire pipeline execution on accelerator
- High channel density for video analytics applications

### TARGET USERS

- Data Scientist
- AI Developers
- Software Developers
- Data Center Infrastructure Engineers

### USE CASES

- Video Content Analysis
- Natural Language Processing
- Smart City/ Smart Retail
- Recommendation Ranking
- Medical Imaging

# SPECIFICATIONS

| FEATURE | DETAILS |
|---|---|
| **Application** | • AI Inference |
| **Architecture** | • AMD XDNA – Versal AI Core |
| **AI Engine** | • 2nd-gen AIE-ML tiles |
| **TOPS* (INT8 )** | • 404 |
| **TOPS* (BF16)** | • 202 |
| **Memory Bandwidth** | • 47600 GB/s for internal memory<br>• 76.8 GB/s for external DDR |
| **High Density Video Decoder**** | • 96 channels of 1920x1080p<br>• H.264/H.265 |
| **PCIe interface** | • Gen 4/5 x 8 |
| **Form Factor** | • Half Height, Half Length |
| **Cooling** | • Passive |
| **Power (TDP)** | • 75W |

* Using 50% weight sparsity
** @10 fps, H.264/H.265

# NEXT STEPS

- Order Alveo V70 Development Card
- For additional questions, contact V70_solutions@amd.com