Unmatched Global Market Coverage

aiworld

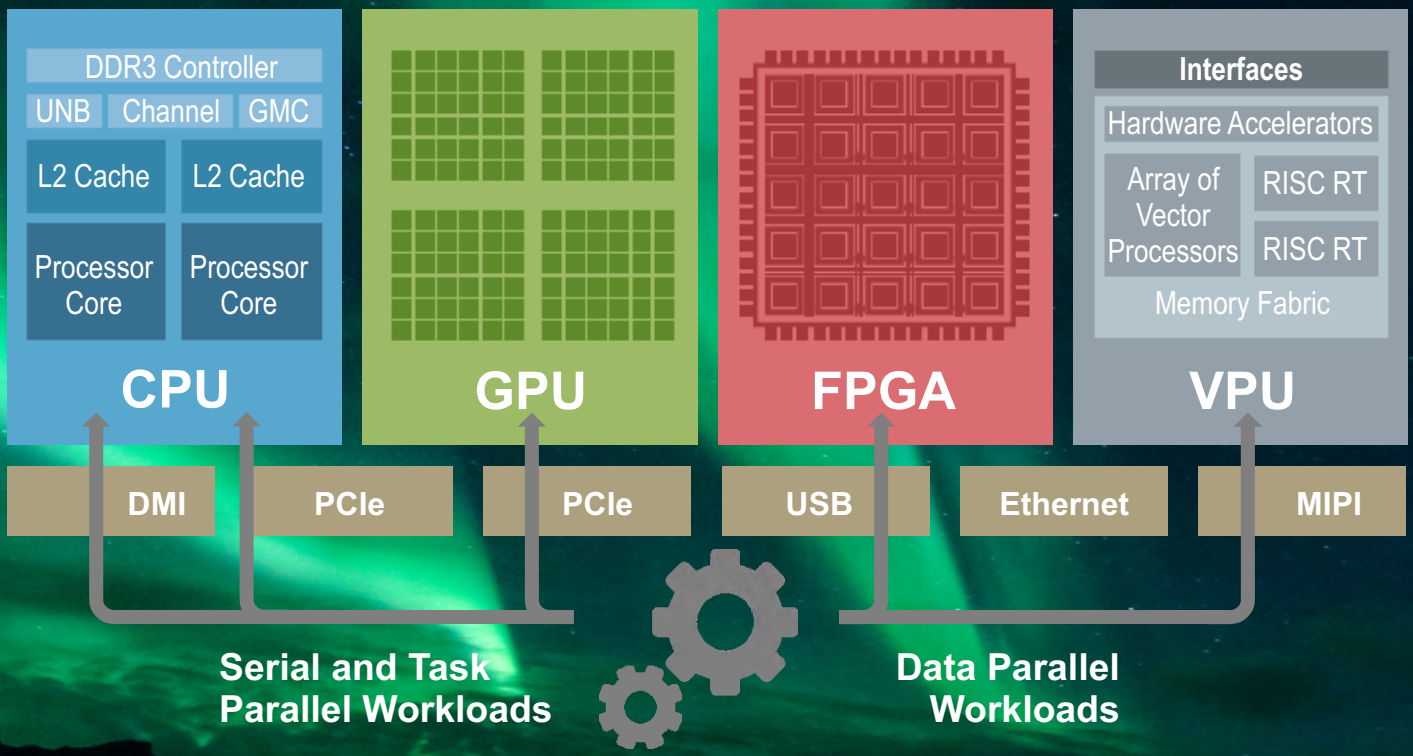
Artificial Intelligence UPDATE

Machine Learning - Deep Learning - Software Research
Products - Technology - Applications - Industry News

AI World is a Global Publication of e2mos

Jan-Feb 2019

Why AI at the Edge



AI Training Platform

Validated with NVIDIA® Tesla®
P100/V100 accelerators
21-billion-transistor
Crazy performance

Smart Camera

Breaking the bounderies of smart camera and embedded vision systems

Sponsered by ADLINK

Products & Technology from Silicon to Application-ready Platforms

SPECIAL Edition ADLINK
Artificial Intelligence

Topics include:

- Why AI at the Edge
  Heterogeneous Computing for Artificial Intelligence at the Edge

- Vortex Edge: Connect the Unconnected, Stream Anywhere, Control the Edge

- 4U Network Appliance for Cloud, Edge, Security, DPI, …

- AI Training Platform Validated with NVIDIA® Tesla® P100/V100 accelerators

- Breaking the boundaries of smart camera and embedded vision systems

- Smart Camera - Starter Kit

- ADLINK to Showcase Plug-and-Play AI and Edge Computing Solutions at NVIDIA 2019 GPU Technology Conference

Daniel Dierickx
CEO & co-Founder
at e2mos
Acting Chief Editor

Over 3 Decades Semiconductors & Computer Systems Market Expertise

Dear Reader,

Here is your free copy of **AI World** one of our six e-magazines

Your benefits:
- Selected subjects
- Easy reading
- Many direct links for more
- Efficient and time saving
- FREE Worldwide

FREE Subscription
Click on the logos below

Semi Update World
aiworld
IoT World
Embedded Systems World
Telecom COTS World
Broadband Broadcast IoT AI Convergence
ATCA World

Editor/Publisher: e2mos
WEB: www.e2mos.com
Contact: mgt@e2mos.com

ADLINK
Leading EDGE COMPUTING

Solution Brief 2019

Heterogeneous Computing for Artificial Intelligence at the Edge

ADLINK supplies flexible heterogeneous computing platforms and helps users optimize their system architectures to fulfill their application and ROI objectives.

Optimized AI Solution

Many industries are pursuing artificial intelligence (AI) with the hope of transforming their business through higher levels of automation and machine learning. There are countless examples, including manufacturers experimenting with AI-enabled machine vision for defect classification and using AI-enabled optical character recognition to extract data from legacy machines. However, AI is still in its infancy, and the complexity and diversity of hardware and software solutions can be overwhelming.
In order to reach an optimized solution, system architects need to first decide whether to run the bulk of their AI algorithms near the sensors (i.e., at the edge) or in the cloud. This decision will then impact what they choose for hardware solutions with respect to performance, size, weight, and power (SWaP) requirements. To maximize AI performance at the edge, an optimized solution will often employ a heterogeneous computing platform, meaning it has two or more different types of computing cores, such as:

- General-purpose CPU
- Field programmable gate array (FPGA)
- Graphics processing unit (GPU)
- Application-specific integrated circuit (ASIC)

This brief discusses the tradeoffs for these core types when implementing AI "at the edge." In addition, it covers the techniques ADLINK uses to help its customers optimize their AI solutions.

Why AI at the Edge

The Internet of Things (IoT) is progressing from simple devices feeding data to the cloud for analysis to smart devices performing sophisticated inferencing and pattern-matching themselves. Processing AI algorithms locally on a smart device in the field provides many benefits, including:

- Faster response: Minimize delay by eliminating the need to send data to the cloud for AI processing.
- Enhanced security: Decrease the risk of data tampering by sending less data across networks.
- Improved mobility: Reduce reliance on inconsistent wireless networks (i.e., dead zones, service outages) by performing AI functions locally on the mobile system.
- Lower communications cost: Spend less on network services by transmitting less data.

AI Design Challenges

The field of AI is incredibly diverse. System architects are applying AI workloads to a wide range of inputs, like video, text, voice, images, and sensor data, with the goal of improving a system's decision making. They must choose from a range of decision making processes that implement various deep learning frameworks (e.g., TensorFlow, Torch, and Caffe) and neural networks (e.g., recurrent and convolutional) with different numbers of layers. Particular combinations of neural networks and frameworks, running on specialized computing cores, are ideal for specific tasks, like image processing, character recognition, and object classification.

Many AI workloads require large amounts of memory, parallel computing, and low-precision computation.1 The challenge for system architects is to define an optimized AI platform that cost-effectively delivers these computing resources in ways that satisfy their speed and accuracy requirements. For platforms deployed at the edge, system architects must address additional requirements, such as environmental hardening and stringent SWaP constraints.

... to next page

Solution Brief 2019
Heterogeneous Computing for AI at the Edge

... from previous page

AI Design Solutions

When designing an AI platform, system architects should consider using a heterogeneous computing architecture, containing multiple core types, including CPU, GPU, FPGA, and ASIC. The goal is to run AI workloads on the best-suited core, resulting in faster computation and less power consumed for a particular function, compared to a homogeneous platform.

Although developing a heterogeneous platform will be more complex than a homogeneous platform, ADLINK simplifies the design process by offering heterogeneous platforms that provide a mix of core types, as shown in Figure 1. System architects can configure ADLINK platforms according to their AI computing needs, reduce their development effort, and benefit from a scalable solution.

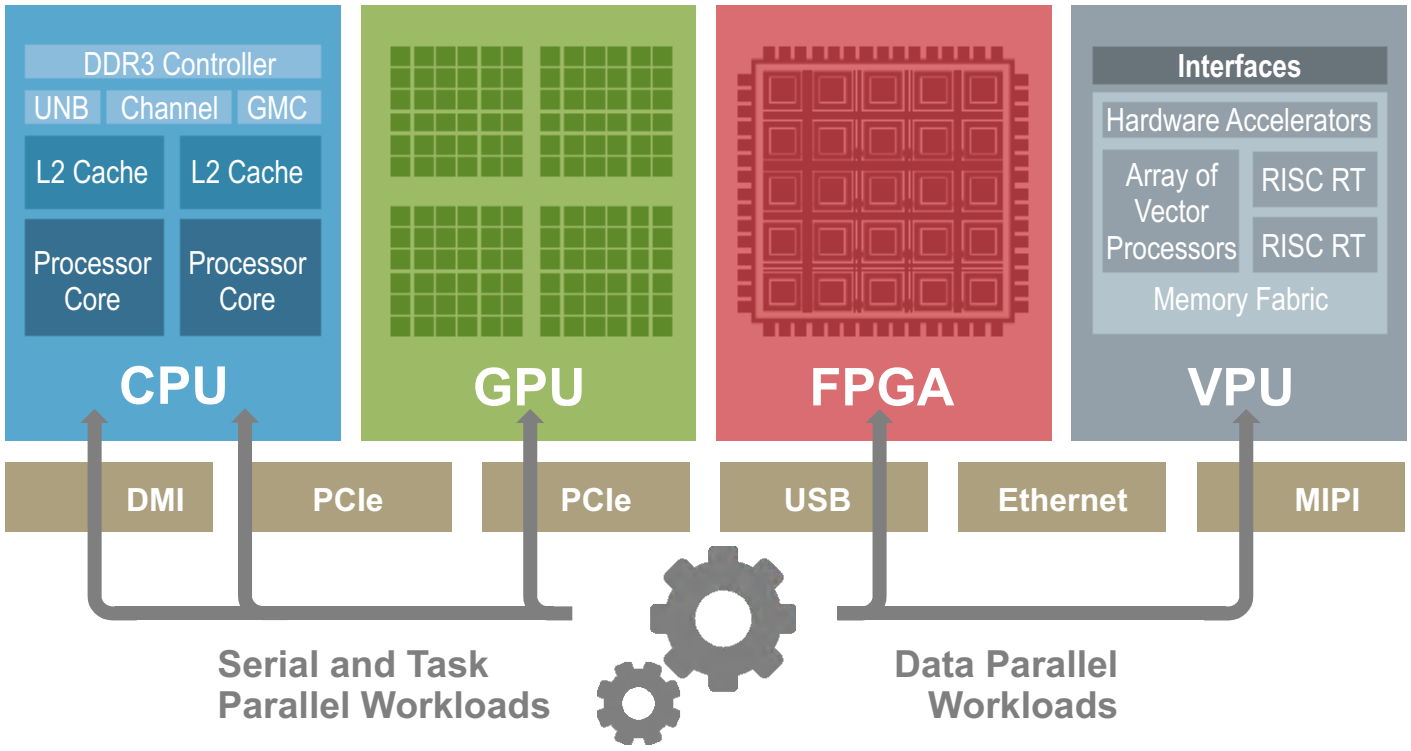


Figure 1. ADLINK heterogeneous architecture options for artificial intelligence applications

Core Type Comparison

The following provides a brief overview of some of the strengths and constraints for different core types used to process AI workloads. Additional information is presented in Table 1.

General-purpose CPU
Generally, every AI platform will have a CPU for running platform management, feature-rich applications, and, possibly, a user interface. In addition, CPUs work well with mixed data input (e.g., audio, text, image, etc.), and extract, transform, and load (ETL) processes.

Graphics processing unit (GPU)
A GPU is a highly-task-parallel, specialized core used for graphics processing, and its architecture is well-suited for AI workloads. With hundreds or thousands of small cores used to execute sophisticated mathematical and statistical computations, GPUs can perform both training deep neural networks (DNNs) and inferencing; however, GPUs can have a large footprint and high power consumption.

... to next page

Solution Brief 2019
Heterogeneous Computing for AI at the Edge

... from previous page

Field programmable gate array (FPGA)
FPGAs have configurable logic gates that can be programmed for a custom application and later reprogrammed in the field, if necessary, offering a high level of flexibility.

Application-specific integrated circuit (ASIC)
ASICs are custom logic chips designed using a manufacturer's circuit libraries. These cores can quickly perform complex, repetitive computation, but they are expensive (high non-recurring engineering costs) and time consuming (one to two years) to design.

- **Vision processing unit (VPU)** VPUs are low-power, small-footprint, customized ASICs used for computer vision and image processing. They are suitable for trained models and less so for training workloads, like machine learning.

- **Tensor processing unit (TPU)** Google developed the first TPU for the computational workloads (e.g., inferencing) of neural networks in edge cores. This custom ASIC is optimized for Google's machine learning framework, called TensorFlow.

Core Type	Custom ASIC	Typical Power Consumption	Description	Strengths	Constraints
CPU		High	Flexible, general purpose processing units	• Complex instructions and tasks • System management	• Possible memory access bottlenecks • Few cores (4-16)
GPU		High	Parallel cores for high quality graphics rendering	• High performance AI processing • Highly parallel core with 100's or 1,000's of cores	• High power consumption • Large footprint
FPGA		Medium	Configurable logic gates	• Flexible • In-field reprogrammability	• High power consumption • Programming complexity
ASIC		Low	Custom logic designed with libraries	• Fast and low power consumption • Small footprint	• Fixed function • Expensive custom design
	VPU	Ultra-low	Image and vision processor/co-processor	• Low power & small footprint • Dedicated to image and vision acceleration	• Limited dataset and batch size • Limited network support
	TPU	Low to medium	Custom ASIC developed by Google	• Specialized tool support • Optimized for TensorFlow	• Proprietary design • Very limited framework support

AI Application Examples

ADLINK is committed to helping system architects bring AI running on a heterogeneous computing platform to the edge, as shown in Figure 2. Here are some computer vision examples:

Automated Optical Inspection
Automated optical inspection (AOI) is being used to spot product defects during the manufacturing process, helping factory personnel quickly fix product yield and quality issues. AOI machines based on ADLINK high-performance, edge computing platforms deliver near-real-time defect detection and identification, and run AI workloads to develop domain knowledge used to better classify defects.

... to next page

Solution Brief 2019
Heterogeneous Computing for AI at the Edge

... from previous page

Optical Character Recognition
Another computer vision application is optical character recognition (OCR), used to read data from the graphics displays of unconnected legacy machines. Embedded GPUs running AI algorithms on ADLINK heterogeneous computing hardware greatly increase the speed and accuracy of OCR workloads.

Autonomous Mobile Robots
A new generation of autonomous mobile robots (AMR) is using VPU-accelerated AI computation for vision-based guidance and collision avoidance. These capabilities allow them to adjust to changes in a facility's floorplan or processes through a straightforward software update that allows them to navigate properly and carry out new tasks. Future mobile robots will be controlled by fleet software that assigns tasks to robots based on their availability and location, thus increasing their efficiency, productivity, and ability to work collaboratively with other robots and humans.

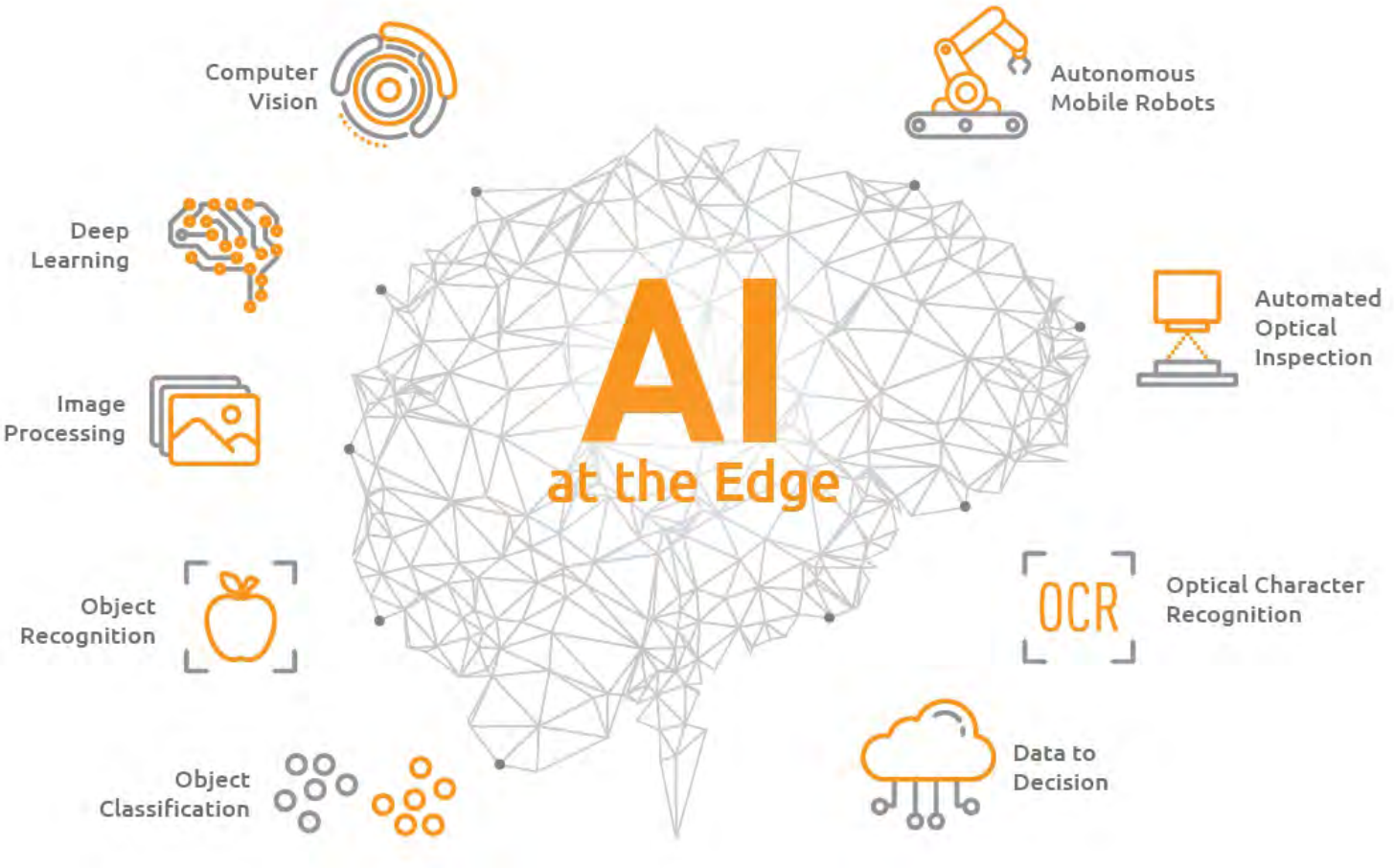


Figure 2. ADLINK is committed to helping system architects bring AI running on a heterogeneous computing platform to the edge.

AI-Powered Edge Devices

ADLINK Technology is enabling the IoT with innovative embedded computing solutions for edge computing. Now, ADLINK is taking embedded computing to the next level with heterogeneous computing platforms optimized for AI. ADLINK heterogeneous computing platforms consist of GPU- and VPU-accelerated board-, system-, and server-level products, enabling system architects to construct and optimize system architecture for both AI inferencing and training applications, as shown in Figure 3. In addition to power efficiency and longevity support, ADLINK's hardware offers the high performance required to quickly process data for deep learning inferencing, pattern-matching, and autonomous machine learning. With intelligence moving to the edge, ADLINK heterogeneous computing platforms perform real-time streaming of data between edge devices and systems, ultimately leading to better decision making.

... to next page

Solution Brief 2019
Heterogeneous Computing for AI at the Edge

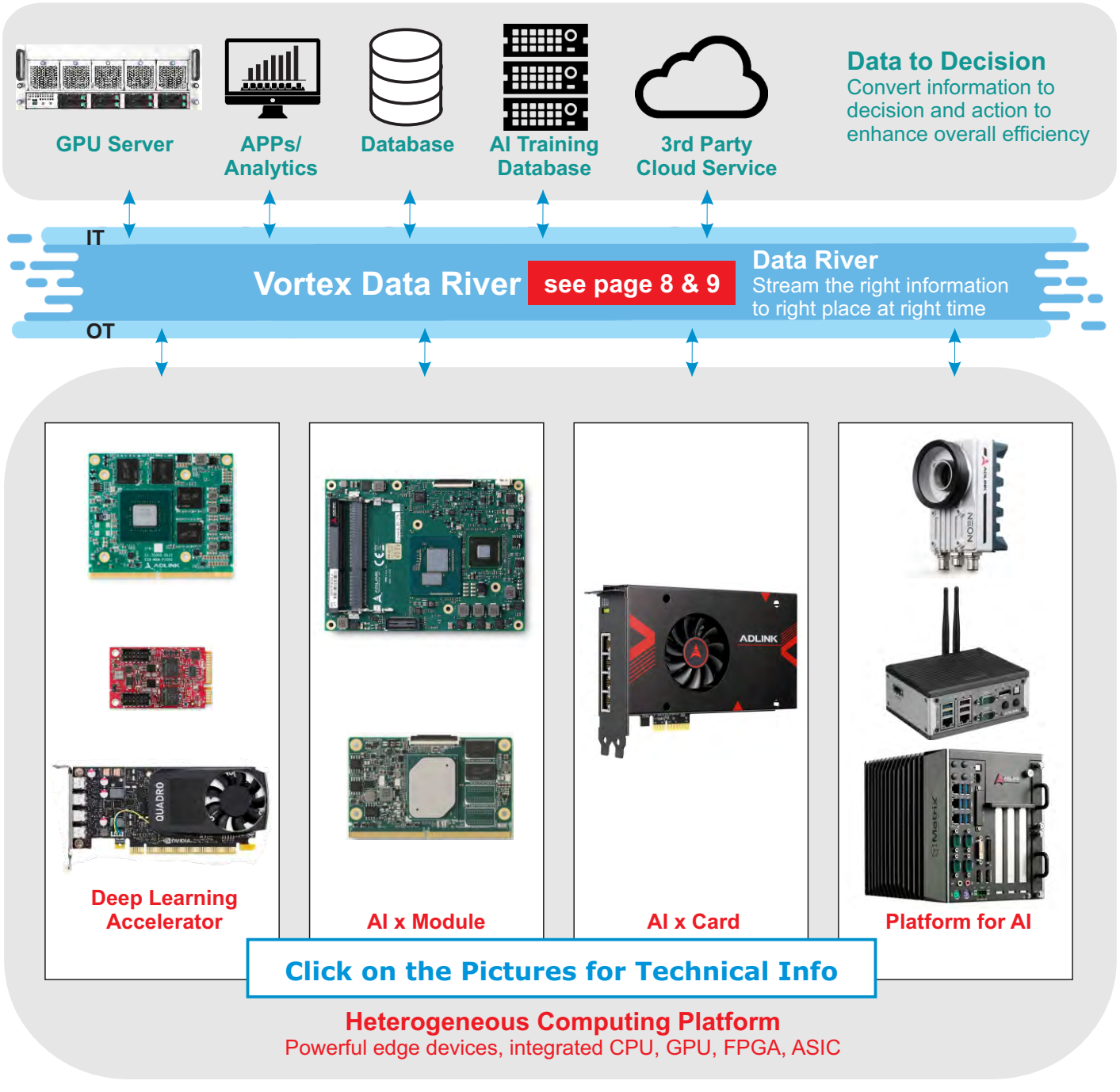
... from previous page

ADLINK System Optimization Services

In addition to its large variety of heterogeneous computing products, ADLINK offers 'consultancy services' via 'deep learning profiling' to help users determine the right platform to cost-effectively satisfy their applications needs. ADLINK is able to make hardware recommendations on how to optimize performance/watt and performance/cost for AI applications in smart manufacturing, smart city, and defense.

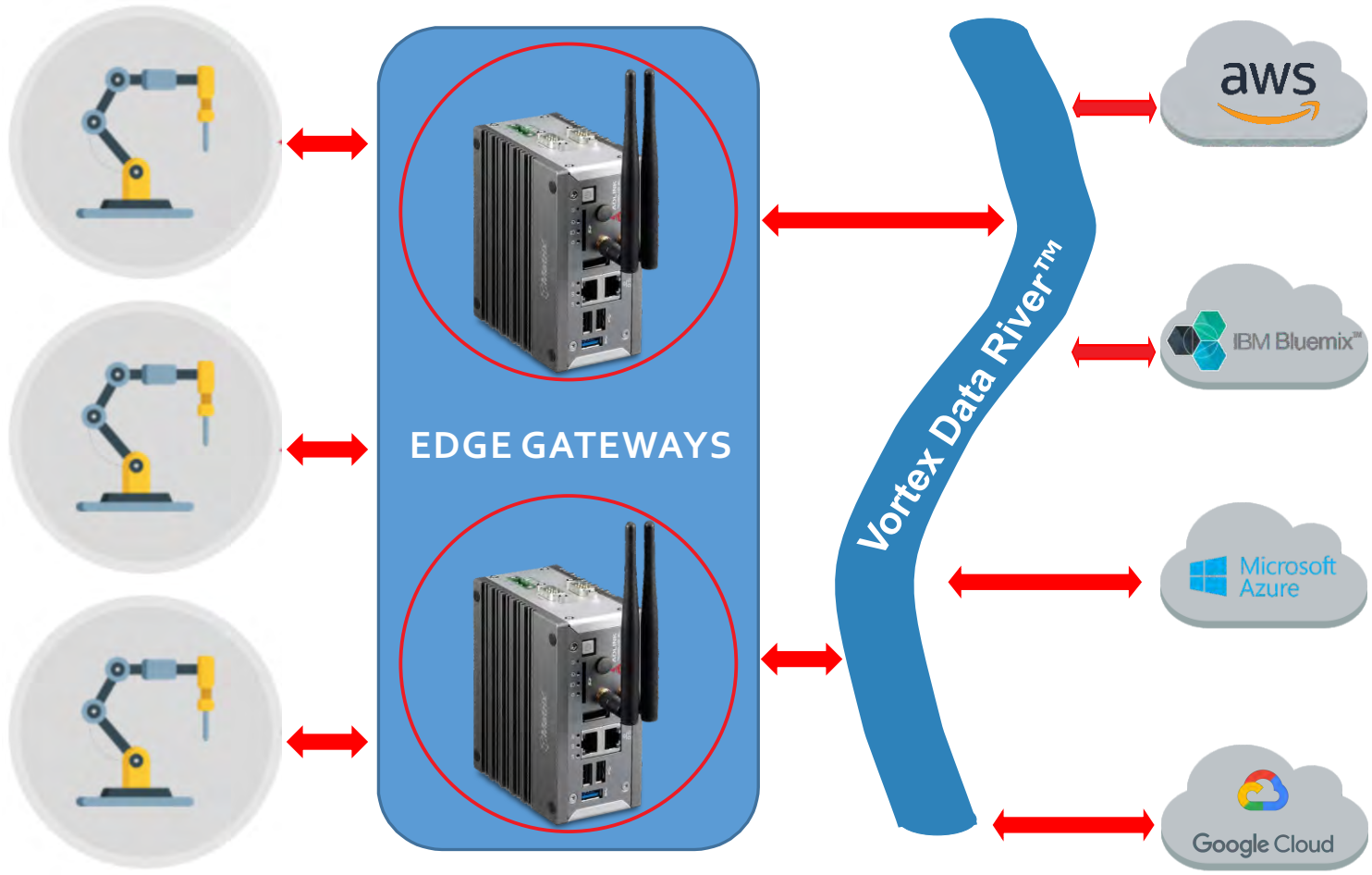
ADLINK is also working with research bodies and academic institutions to find bottlenecks on AI platforms using an analyzer to profile system performance. For example, it is possible to determine if the system is making too many memory copies or if increasing resources (e.g., memory size) will boost performance.

Take advantage of ADLINK's embedded computing solutions and deep learning profiling to optimize the performance of AI-enabled edge devices.



Vortex Edge

Product Summary:
Connect the Unconnected, Stream Anywhere, Control the Edge



At ADLINK we believe in vendor neutral solutions that are secure, scalable and can evolve with your needs in a world where disruption can change the rules of engagement, almost overnight. Vortex Edge is the flexible platform that brings together the best hardware, software and services to create holistic solutions to the challenges of the Internet of Things (IoT).

The available Vortex Edge microservices are:

- Vortex Edge OT Connect
- Vortex Edge Cloud Connect
- Vortex Edge Data Store Connect
- Vortex Edge Persistence
- Vortex Edge Visualization

These can be combined in any combination along with smart gateways, certified sensors and expert DAQ remote I/O modules, as required, to provide a fully featured Vortex Edge installation.

Get your IoT project up and running quickly with a Vortex Edge Digital Experiment
Contact Us for more details.

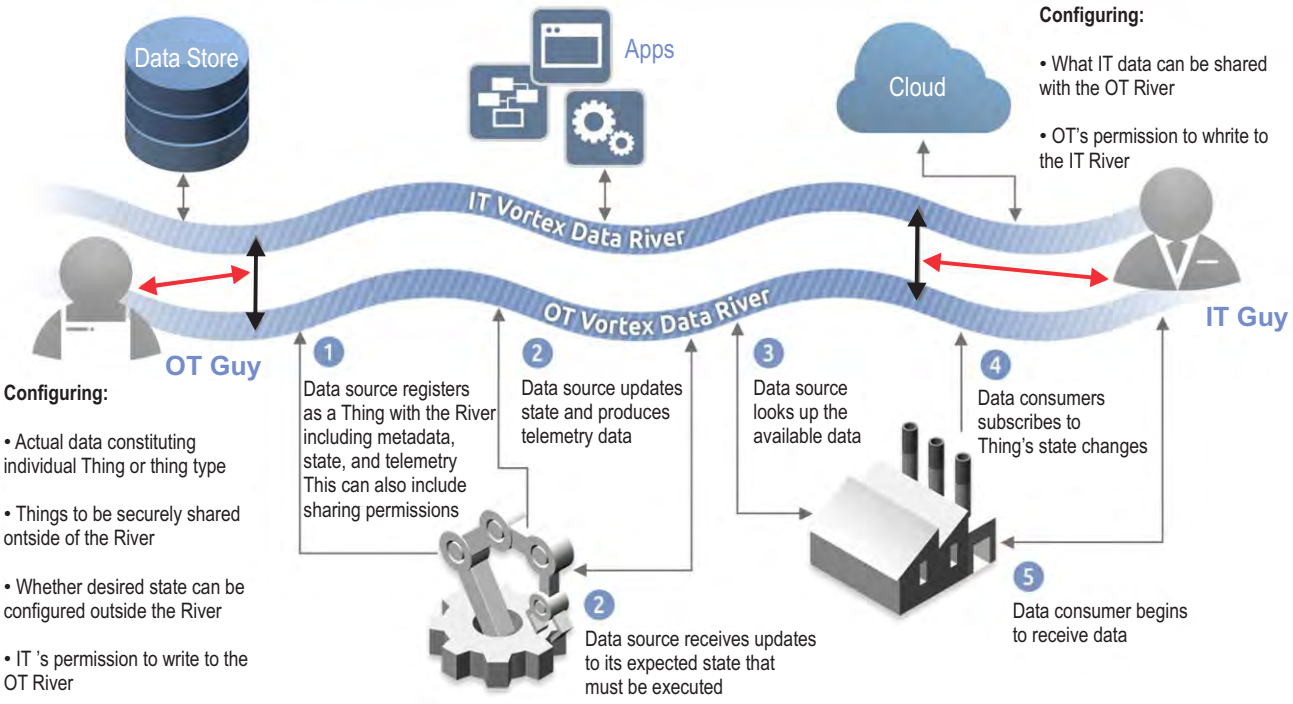
Useful Resources
Vortex Edge Overview Brochure PDF
Introduction to Vortex Edge

Vortex Edge™ Makes IoT Simple

A unique mix of edge hardware, data connectivity software, and services to get your data moving in the correct format to who needs it, where it needs to be, precisely when it needs to be there.

See the Video | **Request a DEMO** | **Read the Success Story** | **Learn How to Get Started**

Converting IT and OT



Edge Hardware
+ Can Include:
Data extraction devices
IoT gateways
Industrial servers
Embedded systems & modules
Sensor packs
Smart vision cameras
Machine condition monitoring hardware

Data Connectivity Software
+ Can Include:
Universal communication protocols
Cloud connectivity
Real-time data streaming connectivity
Peer to peer networking
Edge connectivity

Services = Vortex Edge®
Can Include:
Device management
Analytics
Visualization
Storage
Data subscription management
Professional services

4U Network Appliance 19" - CSA-7400

Cloud - Edge - Security

APPLICATIONS

- High-end Network Security
- Telecom DPI, IDS/IPS, DDoS, NGFW, vBRAS/vCPE
- Hyper-Converged Platform (HC)
- Cloud Edge, Cloud Interconnection, Cloud Security

4U 19'' Network Appliance
CSA-7400 Click Here

KEY FEATURES

- Based on **Open Compute Carrier-grade Edge Reference Architecture** (OCCERA)
- 4U high density platform powered by four dual Intel® Xeon® E5 or Scalable processors
- Up to 8 CPU with 160 Cores in one CSA-7400 system
- DDR4 2666/2400 MHz memory ECC, 48 DIMMs, up to 1,536 GB
- Flexible IO combinations via choice of switch sleds (MXN-3610, MXN-4100) and Network Interface Modules (NIM-1610, NIM-0440)
- Advanced chassis management
- Redundant AC/DC PSUs (N+1)
- Optional integration of **Wind River Titanium Server software to provide carrier grade NFV service** for **5G**
- Support for hardware acceleration for Open vSwitch and OpenFlow protocol processing, accelerating **SDN services**

AI Training Platform 19" - ALPS-4800

Artificial Intelligence - HPC

APPLICATIONS

- Machine Learning (ML)
- Deep Learning (DL)
- High Performance Computing (HPC)

AI Training Platform
ALPS-4800 Click Here

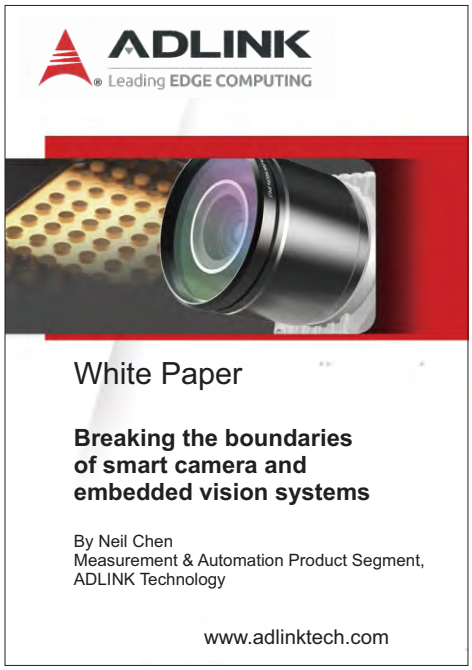
KEY FEATURES

- 8x PCIe x16 Gen3 **GPGPU** slots (300W/slot)
- Validated with **NVIDIA® Tesla® P100/V100** accelerators
- Dual **Intel® Xeon®** Scalable processors
- 24x DDR4-2666 RDIMM
- 8x SATA 6Gb/s hot-swappable 2.5" drives
- 1x FHFL and 2x Low Profile PCIe x16 Gen3 add-on slots
- 1600W AC/DC Platinum PSU, 3+1 redundancy
- Separate airflow for CPU and GPU

Contact ADLINK

World Wide Offices

Breaking the boundaries of smart camera and embedded vision systems



DOWNLOAD WHITE PAPER

Simple, Easy Development — All-in-one solutions with compact size, moderate computing power, and limited resolution

Complex, Multiple Inspection — Multiple channel, flexible, high performance solutions featuring open architecture

Smart Camera - Starter Kit

The Smart Camera Starter Kit contains everything you need to start inspecting right away

Key Features

- Intel® Atom™ Quad-Core Processor E3845 1.91GHz
- 4MP, 60fps, monochrome global shutter CMOS sensor
- IP67-rated housing and M12 connectors thoroughly protect against dust & moisture
- Advanced image processing support
- Additional GigE Vision 1 slave camera support reduces TCO
- Built-in PWM lighting control
- Flexible software support for STEMMER Common Vision Blox, MVTec HALCON, and many others
- GeniCam , GenTL, Open CV and Open CL compatible with image acquisition
- Windows Embedded Standard 7

REQUEST QUOTE

ADLINK Offices Worlwide CONTACT

ADLINK to Showcase Plug-and-Play AI and Edge Computing Solutions at NVIDIA 2019 GPU Technology Conference

ADLINK Technology, Inc., a global provider of advanced edge computing products, will showcase a series of plug-and-play AI and heterogeneous computing platforms at the NVIDIA GPU Technology Conference (GTC), March 18-21, at the San Jose McEnery Convention Center in Booth #317 and Booth #1749 in the Jetson Pavilion.

ADLINK's heterogeneous computing platforms consist of GPU board, system, and server-level products, enabling system architects to construct and optimize applications for both AI inference and training applications. With intelligence moving to the edge, ADLINK's heterogeneous computing platforms with plug-and-play AI can perform real-time streaming of data between edge devices and systems, enabling faster, better decision-making.

"ADLINK is taking embedded computing to the next level with GPUs. As an NVIDIA Quadro Embedded Partner with extensive experience in embedded applications, ADLINK is uniquely qualified to bring NVIDIA Quadro GPUs with GPUDirect, NVIDIA Video Codec SDK, custom firmware and longevity support into embedded applications, enabling them to tap the power of embedded graphics and AI," said Elizabeth Campbell, GM for ADLINK Americas.

ADLINK will showcase its plug-and-play AI and heterogeneous computing platforms for edge applications at including:
• A portfolio of heterogeneous computing platforms and configurable industrial PCs with NVIDIA Quadro embedded solutions optimized for AI at the edge.
• An automated optical inspection (AOI) application, based on ADLINK's industry leading heterogeneous computing platform with NVIDIA Quadro embedded GPU.
• AI-accelerated PCIe frame grabbers for industrial vision applications, fast implementation with no component changes required.
• A highly configurable, scalable and serviceable AI training platform with workload optimization capability, designed to enable AI/machine learning/deep learning applications with varying processing requirements.
• ADLINK's latest multi-access edge computing server, designed to meet the requirements of ultra-low latency, high bandwidth, real-time access to 5G radio networks. ADLINK and partner, Charles Industries, a leader in telecom approved outdoor enclosures, wireless cabinets and concealment solutions, will demonstrate the Charles Micro Edge Enclosure (CMEE), the industry's first outdoor pole mounted 5G compatible AI MEC solution that is GR-487 certified with integrated power and cooling.

AI at the edge is a shared vision. To provide a glimpse of how AI transforms edge applications, ADLINK, with extensive experience in embedded markets across industries, will showcase its latest innovation based on NVIDIA's Jetson family in a separate location in the Jetson Pavilion at GTC in Booth #1749, including: • A rugged deep learning inference platform based on NVIDIA Jetson TX2, demonstrating complex traffic monitoring and analytics at the edge
An AI-enabled smart camera powered by Jetson TX2, performing object detection and classification in outdoor or factory settings • A fanless robot controller taking advantage of server-class performance of Jetson AGX Xavier to enable autonomous navigation in automated mobile robots (AMR). **MORE:** https://www.adlinktech.com/en/index.aspx