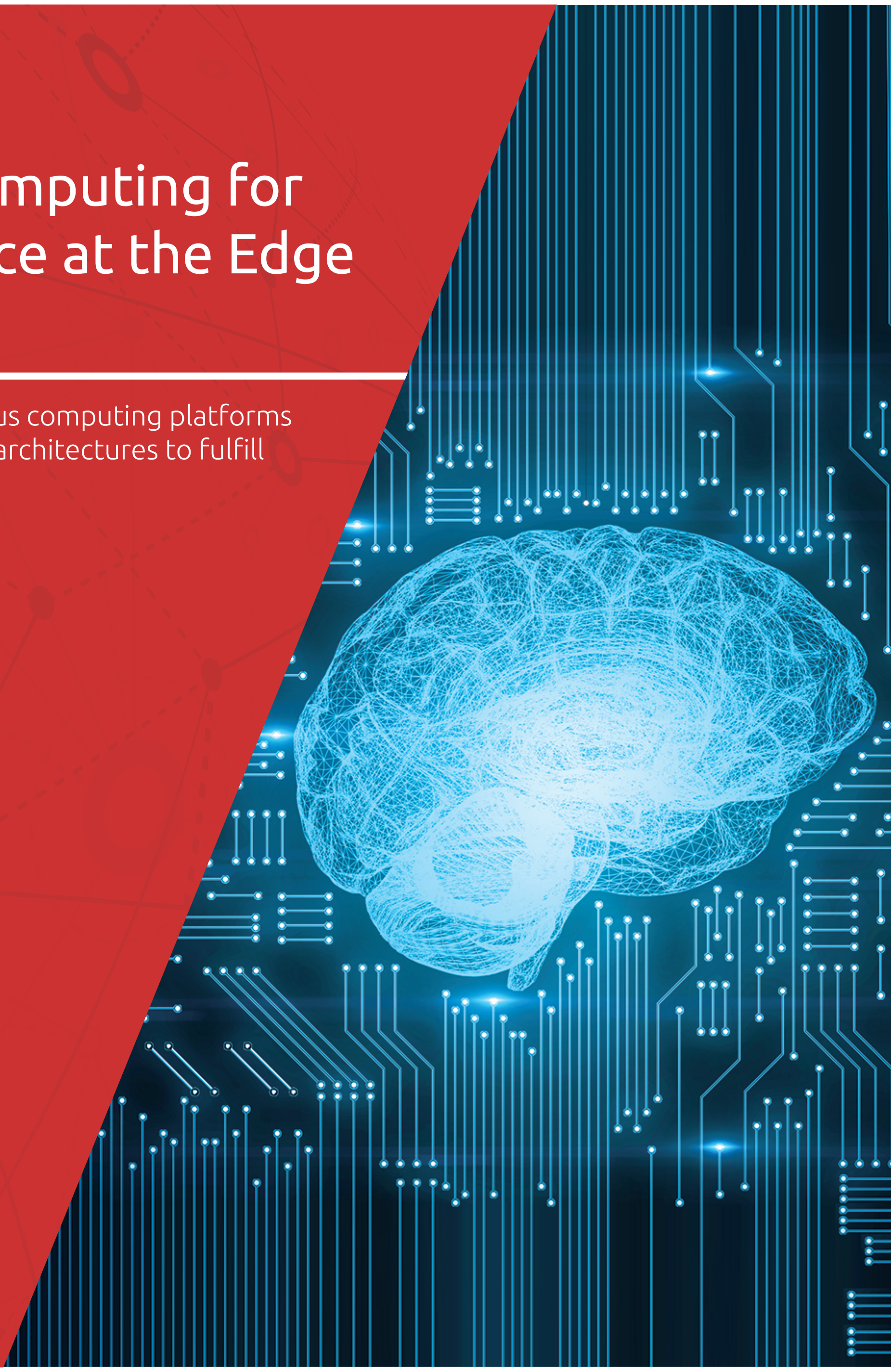
ADLINK
Leading EDGE COMPUTING

Solution Brief

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.



www.adlinktech.com
2019

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.

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.

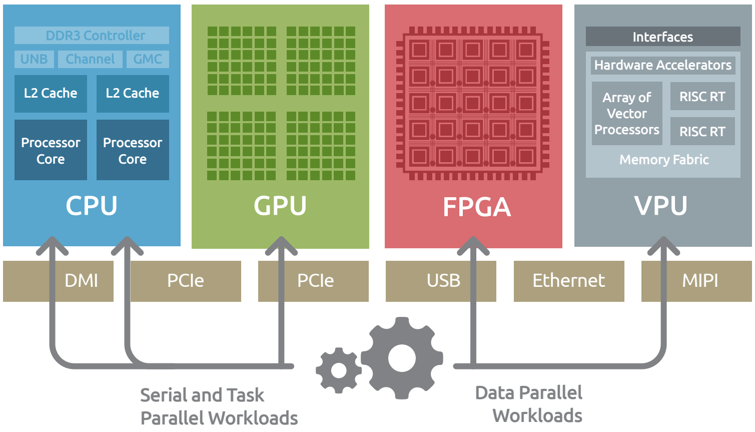


Figure 1. ADLINK heterogeneous architecture options for artificial intelligence applications

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.

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 and 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

Table 1. Comparison of core types used in artificial intelligence applications

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.

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.

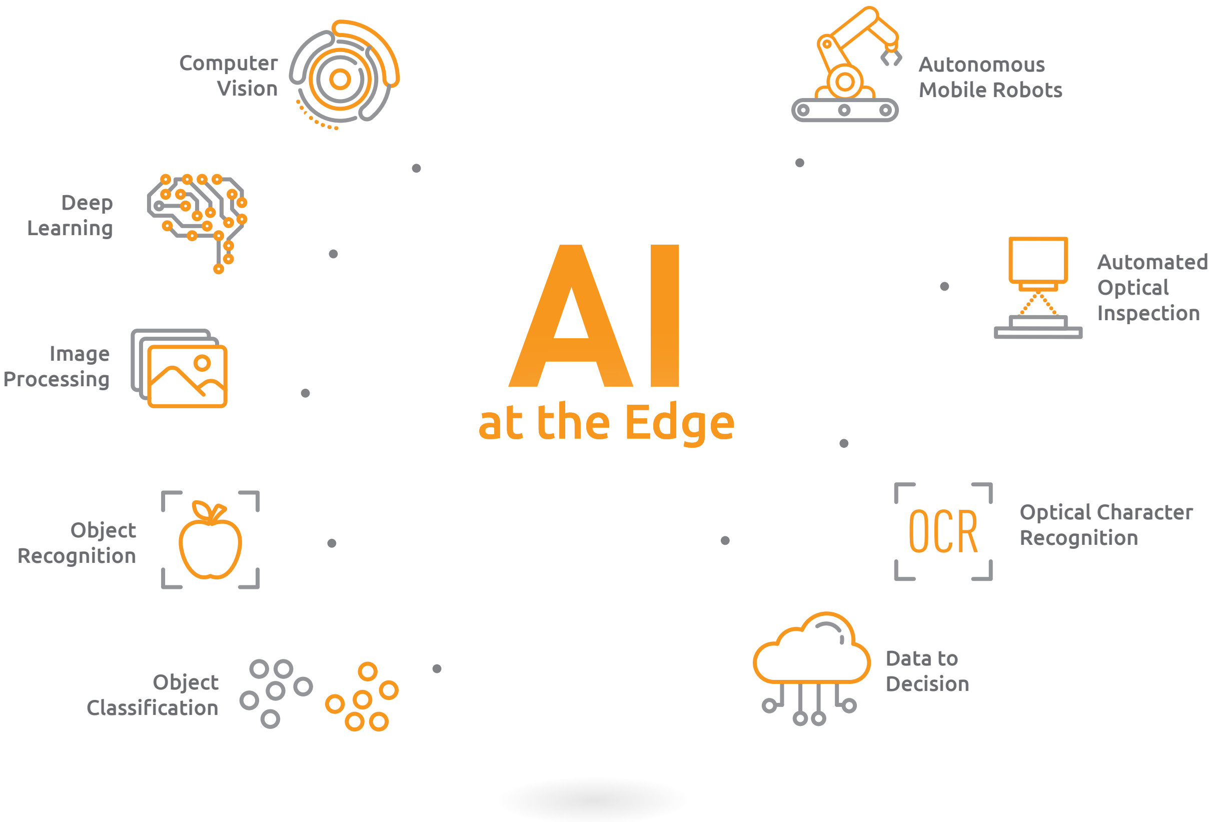


Figure 2. ADLINK is committed to helping system architects bring AI running on a heterogeneous computing platform to the edge.

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.

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.

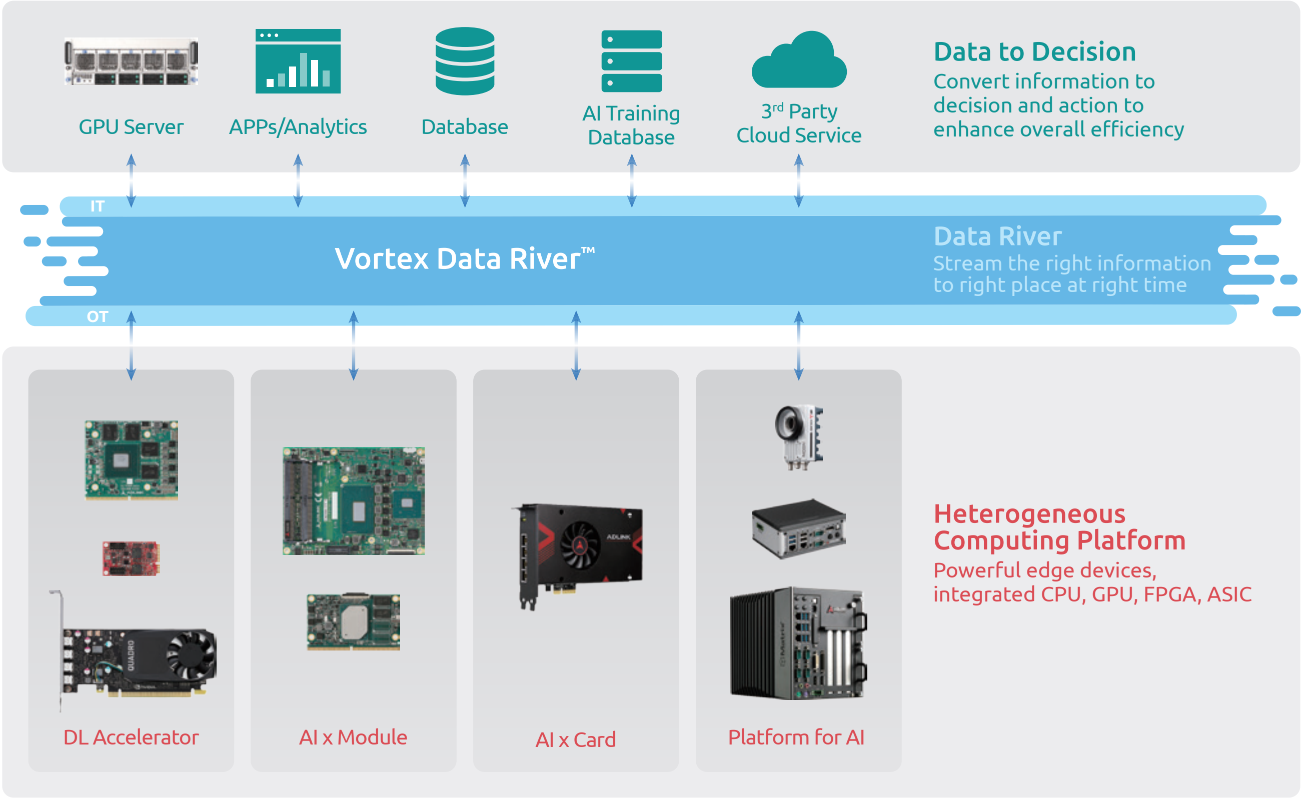

GPU Server
APPs/Analytics
Database
AI Training Database
3rd Party Cloud Service
Data to Decision
Convert information to decision and action to enhance overall efficiency
IT
Vortex Data River™
Data River
Stream the right information to right place at right time
OT
DL Accelerator
AI x Module
AI x Card
Platform for AI
Heterogeneous Computing Platform
Powerful edge devices, integrated CPU, GPU, FPGA, ASIC

Figure 3. ADLINK heterogeneous computing platforms consist of GPU- and VPU-accelerated board-, system-, and server-level products.

[1] Sundeep Bajikar, "Why AI Workloads Require New Computing Architectures – Part 1," June 20, 2018, http://blog.appliedmaterials.com/ai-workloads-computing-architectures-part-1.

WORLDWIDE OFFICES

ADLINK Technology, Inc.

9F, No.166 Jian Yi Road, Zhonghe District
New Taipei City 235, Taiwan
新北市中和區建一路166號9樓
Tel: +886-2-8226-5877
Fax: +886-2-8226-5717
Email: service@adlinktech.com

Ampro ADLINK Technology, Inc.

5215 Hellyer Avenue, #110 San Jose, CA 95138, USA
Tel: +1-408-360-0200
Toll Free: +1-800-966-5200 (USA only)
Fax: +1-408-360-0222
Email: info@adlinktech.com

ADLINK Technology Singapore Pte, Ltd.

84 Genting Lane #07-02A, Axxel Innovation Centre,
Singapore 349584
Tel: +65-6844-2261
Fax: +65-6844-2263
Email: singapore@adlinktech.com

ADLINK Technology Singapore Pte. Ltd. (Indian Liaison Office)

#50-56, First Floor, Spearhead Towers
Margosa Main Road (between 16th/17th Cross)
Malleswaram, Bangalore - 560 055, India
Tel: +91-80-65605817, +91-80-42246107
Fax: +91-80-23464606
Email: india@adlinktech.com

ADLINK Technology Japan Corporation

〒101-0045 東京都千代田区神田鍛冶町3-7-4
ユニゾ神田鍛冶町三丁目ビル4F
Unizo Kanda Kaji-cho 3 Chome Bldg. 4F,
3-7-4 Kanda Kajicho, Chiyoda-ku, Tokyo 101-0045, Japan
Tel: +81-3-4455-3722
Fax: +81-3-5209-6013
Email: japan@adlinktech.com

ADLINK Technology, Inc. (Korean Liaison Office)

경기도 용인시 수지구 신수로 767
A동 1008호 (동천동, 분당수지유타워) (우) 16827
A-1008, U-TOWER, 767 Sinsu-ro, Suji-gu, Yongin-si,
Gyeonggi-do, Republic of Korea, 16827
Toll Free:+82-80-800-0585
Tel: +82-31-786-0585
Fax: +82-31-786-0583
Email: korea@adlinktech.com

ADLINK Technology (China) Co., Ltd.

上海市浦东新区张江高科技园区芳春路300号 (201203)
300 Fang Chun Rd., Zhangjiang Hi-Tech Park
Pudong New Area, Shanghai, 201203 China
Tel: +86-21-5132-8988
Fax: +86-21-5192-3588
Email: market@adlinktech.com

ADLINK Technology Beijing

北京市海淀区上地东路1号盈创动力大厦E座801室(100085)
Rm. 801, Power Creative E, No. 1 Shang Di East Rd.
Beijing, 100085 China
Tel: +86-10-5885-8666
Fax: +86-10-5885-8626
Email: market@adlinktech.com

ADLINK Technology Shenzhen

深圳市南山区科技园南区高新南七道数字技术园
A1栋2楼C区 (518057)
2F, C Block, Bldg. A1, Cyber-Tech Zone, Gao Xin Ave. Sec. 7
High-Tech Industrial Park S., Shenzhen, 518054 China
Tel: +86-755-2643-4858
Fax: +86-755-2664-6353
Email: market@adlinktech.com

ADLINK Technology GmbH

Hans-Thoma-Strasse 11, D-68163
Mannheim, Germany
Tel: +49 621 43214-0
Fax: +49 621 43214-30

(Deggendorf) Ulrichsbergerstrasse 17, 94469
Deggendorf, Germany
Tel: +49 (0) 991 290 94-10
Tel: +49 (0) 991 290 94-29
Email: emea@adlinktech.com

ADLINK Technology, Inc. (French Liaison Office)

6 allée de Londres, Immeuble Ceylan 91940
Les Ulis, France
Tel: +33 (0) 1 60 12 35 66
Fax: +33 (0) 1 60 12 35 66
Email: france@adlinktech.com

ADLINK Technology, Inc. (UK Liaison Office)

First Floor West Exeter House, Chichester Fields
Business Park Tangmere, West Sussex,
PO20 2FU, United Kingdom
Tel: +44-1243-859677
Email: UK@adlinktech.com

ADLINK Technology, Inc. (Israel Liaison Office)

SPACES OXYGEN, 62 Medinat, Ha-yehudim st
4673300, Herzliya, Israel, P.O.Box – 12960
Tel: +972-54-632-5251
Fax: +972-77-208-0230
Email: israel@adlinktech.com

ADLINK®
Leading EDGE COMPUTING