ADLINK
Leading EDGE COMPUTING

Solution Brief

# Optimizing Heterogeneous Computing

ADLINK's consultation services and Data River™ help lower cost, increase performance, and simplify the deployment of deep learning and artificial intelligence (AI) solutions



www.adlinktech.com

2019

Distribuída no Brasil pela LRI | vendas@lri.com.br | 51 2129-3000

## Heterogeneous Computing for AI

Companies seeking to maximize the innovation and productivity gains from deep learning and artificial intelligence (AI) should consider using a computing platform optimized for the associated algorithms.

Deep learning and AI algorithms tend to work on a lot of data and, as a result, require fast memory transfer, parallel computing, and mixed-precision computation (e.g., 8-bit integer or 8/16/32-bit floating point)1 to improve the latency and responsiveness of AI models. In order to handle these diverse requirements, an optimized computing platform will typically employ two or more different types of computing cores, as shown in Figure 1. This design approach is called heterogeneous computing.

This solution brief describes several ways ADLINK can help system developers, OEMs, and systems integrators improve the hardware and software performance of their heterogeneous computing platforms.

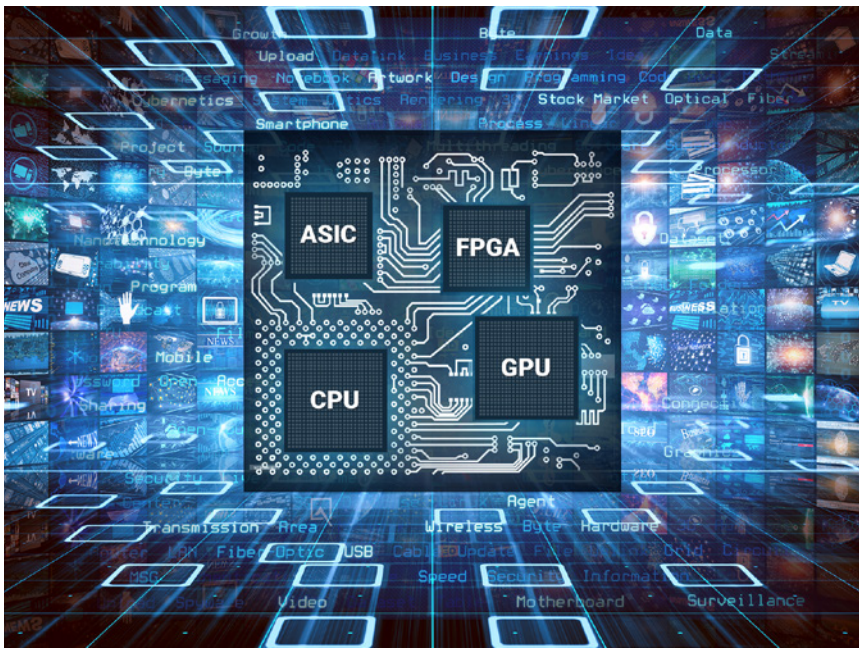



Figure 1. ADLINK heterogeneous computing options for artificial intelligence applications.

## AI Workload Optimization

In order to attain an optimized computing platform for deep learning and AI solutions, developers may need to evaluate many areas (Figure 2), including:

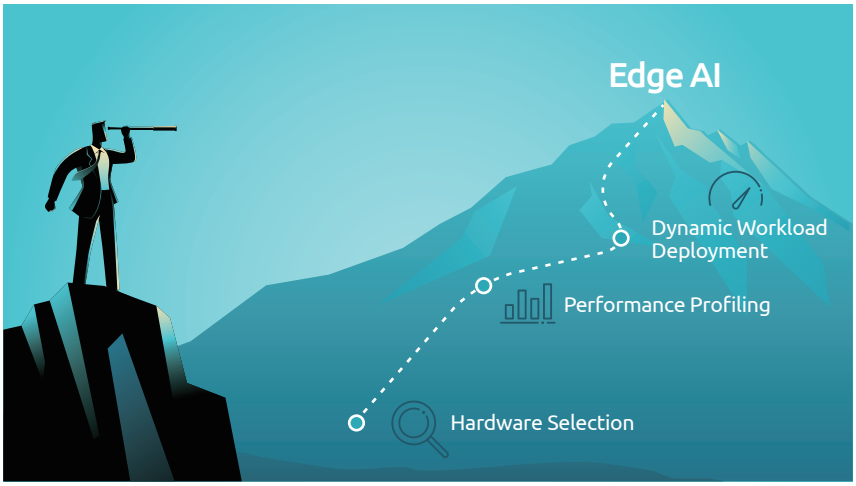



Figure 2. Developers need to evaluate many areas to attain an optimized computing platform for AI solutions.

### Selecting Hardware

Determine which computing cores are best suited to run the required AI algorithms and how much computing power and I/O bandwidth are needed. Hardware selection should also consider size, weight, power consumption (e.g., SWaP), and cost constraints, particularly when deploying AI at the edge of the network.

### Tuning AI Performance

Identify bottlenecks in the software or hardware due to insufficient platform resources (e.g., memory, I/O, computing cores, and cache), inefficient scheduling of software threads, or contention between various running processes. This analysis typically requires software profilers and other types of performance tuning tools.

### Deploying Dynamic Workloads across the Network

Develop the ability to deploy and provision dynamic workloads across the network in order to improve inference results with refined AI models and address new challenges. Dynamic workloads require a flexible and adaptable computing architecture that enables a scalable, real-time, and reliable deployment environment.



Distribuída no Brasil pela LRI | vendas@lri.com.br | 51 2129-3000

## Hardware Selection

ADLINK offers consultancy services via deep learning profiling to help users determine the right hardware platform to cost-effectively fulfill their application needs. ADLINK developed a profiling tool that models computing platform performance based on the elements of AI algorithms, such as types of neural networks, like AlexNet, MobileNet, ResNet (Figure 3).

The profiling tool runs this and other inputs against a large database of AI and deep learning accelerators and generates statistics, such as inferences per second; performance per watt; and performance per dollar, for popular neural networks. Additionally, the tool helps developers determine whether they are using the best accelerators for the neural networks used by their application and suggests which accelerator may offer the best performance.

## Performance Profiling Service

For in-depth performance tuning, developers can have ADLINK run x86 and GPU performance analyzers on their software. These analyzers help identify hardware and software bottlenecks that, when remedied, can greatly increase throughput and inference per second, performance per dollar, and SWaP.

ADLINK is also working with research bodies and academic institutions to find bottlenecks on AI platforms to profile system issues that can be addressed for performance improvement. 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.

## Dynamic Workload Deployment

ADLINK's Data River™ is an underlying technology for deploying dynamic workloads and enabling data to move freely. It solves a key challenge when developing AI solutions, which is allowing data exchange and workload provisioning across the entire network. The solution helps communicate data among computing nodes and devices.

ADLINK's solution is scalable acceleration engines over a distributed data service (DDS) that is a type of message-oriented middleware supporting a data-centric publish and subscribe style of communications. This distributed architecture allows a massive number of computing nodes to be attached to the network. In other words, users can add additional hardware accelerators that best fit their application needs based on profiling results, even after AI is deployed.




Figure 3. ADLINK offers consultancy services to help users determine the right hardware platform.

## Reduce AI Design Effort

System developers, OEMs, and systems integrators can more easily develop deep learning and AI solutions with help from ADLINK's consultation services and Data River™. By working with ADLINK, it is possible to more easily:

- **Lower solution cost:** Select the right hardware platform for the target workload.
- **Increase system performance:** Eliminate platform bottlenecks that slow down AI algorithms.
- **Simplify AI deployment:** Leverage ADLINK's Data River™ to more easily share data and distribute AI processing across the network.

## Deploy AI at the Edge

Optimized for AI, ADLINK heterogeneous computing platforms include a mix of GPUs, vision processing units (VPUs), FPGAs, and ASICs that accelerate AI workloads running on board-, system-, and server-level products. With these platforms, system developers, OEMs, and systems integrators can construct and optimize system architecture for both AI inferencing and training applications, as shown in Figure 4.

### ADLINK GPU- and VPU-Powered Solutions

ADLINK's portfolio of GPU-based and VPU-based solutions offer different performance level, power efficiency, and form factors for AI applications.

- ADLINK's families of embedded graphics products provide hundreds or thousands of parallel cores for AI inferencing and training that require sophisticated image processing, and mathematical and statistical computations. Mobile PCI Express Module (MXM) modules are well-suited for SWaP-constrained applications. PCI Express graphic (PEG) cards connect via a common interface, are easy to integrate, and are used in many embedded segments. For applications that must meet stringent size constraints and other requirements, onboard GPU integration is a customized solution to accommodate size and performance needs.
- For image-based inferencing, embedded VPU modules dedicated to image processing feature low power and a small footprint, and excel in computer vision applications.

### System Components for AI Workload Integration

For SWaP-constrained applications, ADLINK's Matrix compact fanless embedded computers offer the best performance-per-watt and high availability with expandability options, including the previously discussed MXM cards. Actively-cooled DLAP platforms support applications in less thermally-challenged settings.

For mission-critical applications, ADLINK's configurable embedded computers provide an expandable building block to host multiple accelerators, enabling the consolidation of workloads, including highly-parallel graphics computing; motion control; and data acquisition, onto one system. For applications demanding even greater scalability, ADLINK offers highly configurable motherboards and rackmount industrial chassis that provide an enclosure for ADLINK's large family of industrial ATX motherboards. The boards feature multiple PCIe/PCI/LAN/USB 3.0, enable immediate multi-tasking deployment, and balance performance and expandability.

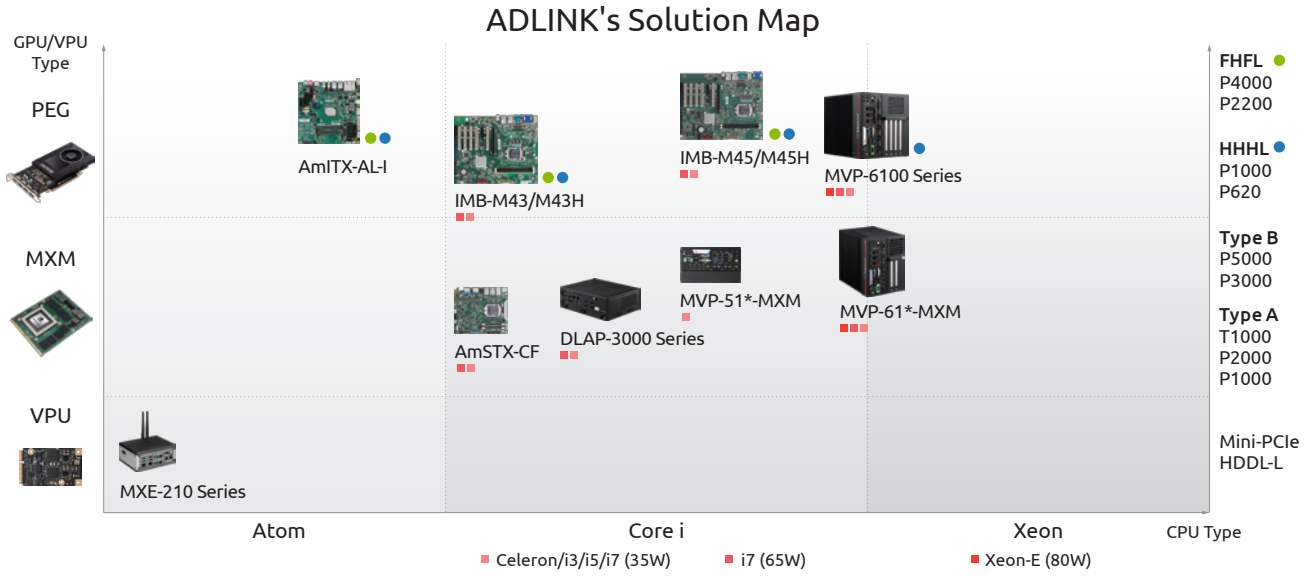



Figure 4. System developers, OEMs, and systems integrators can construct and optimize system architecture for AI applications.

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



Distribuída no Brasil pela LRI | vendas@lri.com.br | 51 2129-3000

# 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: +91-80-42246107, +91-80-23464606
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
Email: germany@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

NVIDIA QUADRO CERTIFIED
EMBEDDED


All products and company names listed are trademarks or trade names of their respective companies.
All specifications are subject to change without further notice.



LRI Distribuída no Brasil pela LRI | vendas@lri.com.br | 51 2129-3000