White Paper

Author: John Bernard

AAEON Technology Inc.

The AAEON UP AI Toolkit

Bridging the Gap Between AI Model Development and Deployment

June, 2025

Contents

Executive Summary

AAEON's UP – Bridge the Gap brand was founded in 2015 with the strategic objective of providing the professional developer community with innovative and reliable single-board computers (SBCs) equipped with the latest Intel® technology. Given the brand's close ties with Intel® and the development-focused nature of its key customer base, the UP brand has traditionally provided a greater breadth of resources with which its customers can build, troubleshoot, and collaborate to bring projects to fruition.

While UP offers a selection of development kits, software preinstallation and customization services, and universal support for Intel® development frameworks such as the Intel® Distribution of OpenVINO™ Toolkit, the brand has now released the UP Edge AI Toolkit, a comprehensive software solution designed to accelerate AI development and deployment on AAEON edge computing platforms. This white paper explores the toolkit's innovative features, addressing common pain points in AI development and deployment, and outlines the benefits of its integrated solutions, which can be summarized as:

- One-click environment setup
- Cross-platform model conversion (Intel, NVIDIA, Hailo)
- Built-in performance benchmarking tools
- Out-of-the-box compatibility with UP hardware

Understanding Key AI Development Pain Points

Complex Environment Setup

Creating development environments for AI projects comes with substantial challenges due to the unique requirements of various platforms with respect to drivers, software development kits (SDKs), and dependencies. For example, the Intel® Distribution of OpenVINO™ Toolkit, NVIDIA® TensorRT™, and Hailo TAPPAS each have their own installation procedures, and so dependency conflicts between libraries like CUDA, Python, and Linux distributions are common.

Each platform requires a different set of drivers, SDKs, and dependencies, often with strict version compatibility. For developers without strong knowledge of multiple platforms, setting up the environment from scratch is time-consuming and error-prone, often making the whole process slow and nonlinear.

Therefore, building a development environment often involves ensuring all cross-vendor components are compatible, as well as aligned to the correct version of components within a vendor's software stack.

The primary challenge for developers in this regard is ensuring all of the tools used can interoperate without causing runtime errors or suboptimal performance, with tight coupling between the SDK version, the compiler, and model schemas necessary.

Complicated Model Conversion

AI models need to be converted into formats that are compatible with the target hardware. This process is often complex because each platform supports specific model formats and has its own conversion pipeline. Developers must spend significant time learning different tools and configuring conversion parameters to successfully convert original models (e.g., PyTorch, TensorFlow) into deployable formats.

Lack of Standardization in Model Performance Evaluation

Before deploying AI models, developers must spend time evaluating their performance across different platforms. However, there is currently no standardized method for doing this, making it difficult to compare performance. As such, the current modus operandi is for developers to evaluate model performance individually across several platforms.

Unsurprisingly, the time, resources, and effort needed to do this make finding the optimal platform very inefficient, as a model that performs well on one platform may require major adjustments to run on another.

The Key Features & Architecture of the UP AI Toolkit

The UP AI Toolkit addresses the three key challenges outlined above by segmenting the execution of tasks needed to solve the challenges and bundling them into three key layers:

1. Foundation Layer – One-Click Runtime Environment Setup
2. Middleware Layer – Unified, Cross-Platform Model Conversion
3. Application Layer – Embedded Benchmarking Tools for Performance Simulation

Foundation Layer – One-Click Runtime Environment Setup

At the foundational level, the UP AI Toolkit consolidates all necessary components such SDKs, drivers, and libraries across Intel, NVIDIA, and Hailo platforms into a unified bundle. As a result, users are able to access a platform-agnostic runtime environment with just one click and immediately utilize the tools available to build models tailored to their required application.

This layer can be broken down into two key subsections.

- **Hardware**
  This layer includes the components either integrated with or embedded in the user's UP hardware, such as a board or system's Intel CPU, iGPU, NPU, Hailo Acceleration module, or NVIDIA GPU.

- **OS/Driver**
  This section of the toolkit houses the operating system on which the user's application will run, as well as drivers for peripheral device integration and the UP Framework/API/SDK.

Middleware Layer – Unified, Cross-Platform Model Conversion

The UP AI Toolkit also includes a unified, cross-platform model conversion engine, allowing users easily convert pre-trained models into formats optimized for supported UP hardware. In theory, this means that models trained with popular frameworks such as machine learning models, datasets, and tools obtained from open-source platforms such as Hugging Face can be successfully converted and deployed on UP hardware.

This layer can be broken down as follows:

- **AI Inference Runtime**
  This is a self-contained ONNX runtime environment that operates within the framework to deploy machine learning models trained on TensorFlow, ONNX, and PyTorch across UP products, acting as a unified AI inference container with platform-specific optimizers to provide enhanced performance without the need for platform-specific adjustments.

- **Platform-Specific Optimizers**
  - **Intel® Distribution of OpenVINO™ Toolkit**
    Models are converted into the Intermediate Representation format, which can be run by the inference engine across Intel hardware such as CPUs, GPUs, and NPUs.
  - **HailoRT**
    A production-grade, light, scalable runtime software used to convert AI models for deployment on Hailo AI Accelerator devices, which focuses on power-efficiency.
  - **The NVIDIA® CUDA® Toolkit**
    A development environment purpose-built to assist in the creation of high-performance, NVIDIA GPU-accelerated applications. This serves the purpose of leveraging the parallel processing offered by NVIDIA GPUs to allow models to run faster with lower memory consumption.

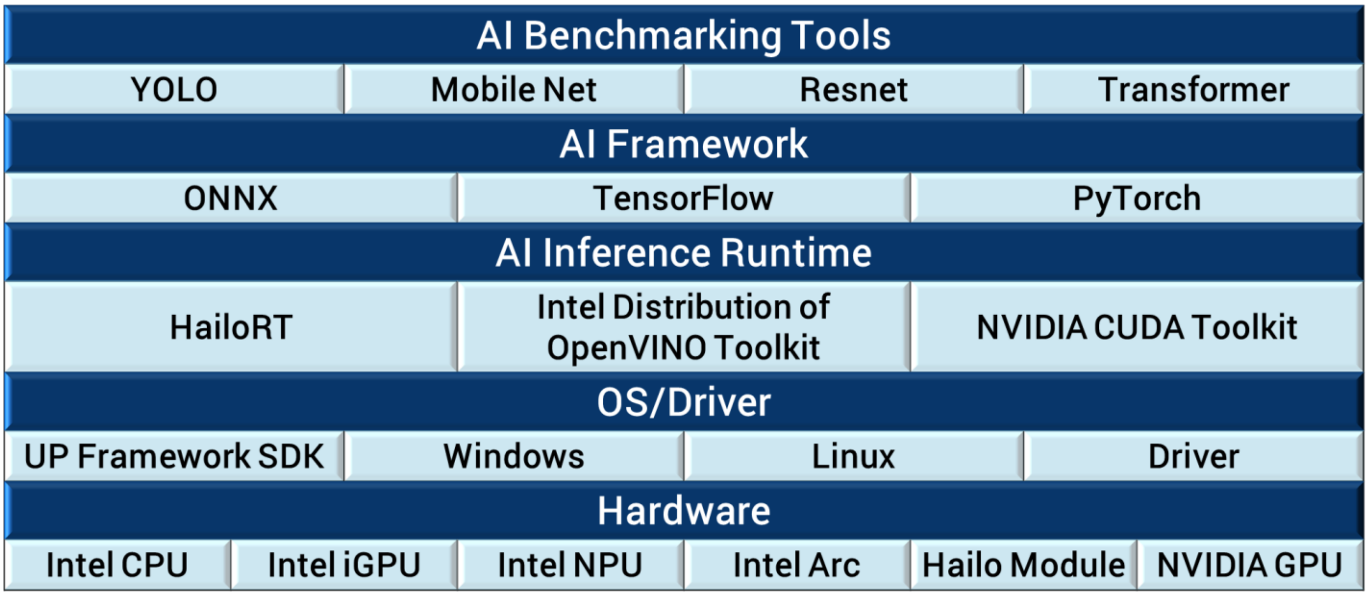
Application Layer – Embedded Benchmarking Tools

Once a model has been converted and deployed onto the user's chosen UP hardware, the UP AI Toolkit provides a built-in benchmarking tool to reduce the time needed between deployment and proof of concept testing.

In short, this tool allows users to evaluate multiple different models on the same hardware platform, simplifying the process by allowing for straightforward comparisons between model architectures on the platform, rather than testing individually across several platforms.

The kit's benchmarking tool is compatible with a variety of AI model architectures, such as ResNet, YOLO, MobileNet, and Transformer. These are commonly employed for inference tasks, and this layer allows users to compare performance metrics such as latency, throughput, power consumption, and thermal efficiency across Intel, NVIDIA, and Hailo platforms for image classification, object detection, and Natural Language Processing (NLP) tasks.

Competitive Advantages of the UP AI Toolkit



AI Benchmarking Tools

YOLO | Mobile Net | Resnet | Transformer

AI Framework

ONNX | TensorFlow | PyTorch

AI Inference Runtime

HailoRT | Intel Distribution of OpenVINO Toolkit | NVIDIA CUDA Toolkit

OS/Driver

UP Framework SDK | Windows | Linux | Driver

Hardware

Intel CPU | Intel iGPU | Intel NPU | Intel Arc | Hailo Module | NVIDIA GPU

One-Click Runtime Environment Setup

The UP AI Toolkit's one-click runtime library setup removes common issues regarding version compatibility across the entire development process. By integrating the necessary drivers, libraries, and SDKs for Intel, NVIDIA, and Hailo software platforms, the toolkit eliminates the need to manage version conflicts between Python, CUDA, and other platform-specific dependencies, thus reducing the time and resources needed to set up a development environment for AI applications.

In addition to the advantages of the UP AI Toolkit's software agnosticism, the same principle applies to the hardware platforms chosen for AI projects. Because the entire software stack is isolated from the host system, this architecture makes deploying and testing converted AI models easier across different vendor hardware, such as those commonly embedded or installed on UP platforms.

In effect, this setup removes the need for developers to manually resolve cross-vendor incompatibilities, enabling streamlined build/test/deploy cycles and enhancing development efficiency.

Out-of-the-Box Compatibility

The UP AI Toolkit's range of pre-configured runtime images are natively compatible with a large section of UP's hardware portfolio. As such, the toolkit makes it so pre-trained models can be executed across different platforms powered by Intel, NVIDIA, or Hailo hardware without the common issues associated with low-level integration.

UP 4000

UP 7000

UP Squared

UP Squared V2

UP Squared 6000

UP Squared 7100

UP Squared i12

UP Squared Pro

UP Squared Pro 7000

UP Xtreme i11

UP Xtreme i12

UP Xtreme i14

A key point to note in this is that the compatibility of the UP AI Toolkit allows for the adoption of AI functionality on platforms that prioritize different qualities, such as power efficiency. The toolkit's compatibility with the UP 710S, UP Squared 7100, and UP Xtreme 7100, all powered by Intel® Processor N-series processors, which are designed for applications requiring a low power footprint, provides a particular benefit as it allows users to integrate similarly efficient AI accelerators without the need for CPUs with higher power needs.

Conversely, the toolkit is also supported across platforms powered by the latest Intel® Core™ and Intel® Core™ Ultra processors for projects requiring higher levels of compute performance.

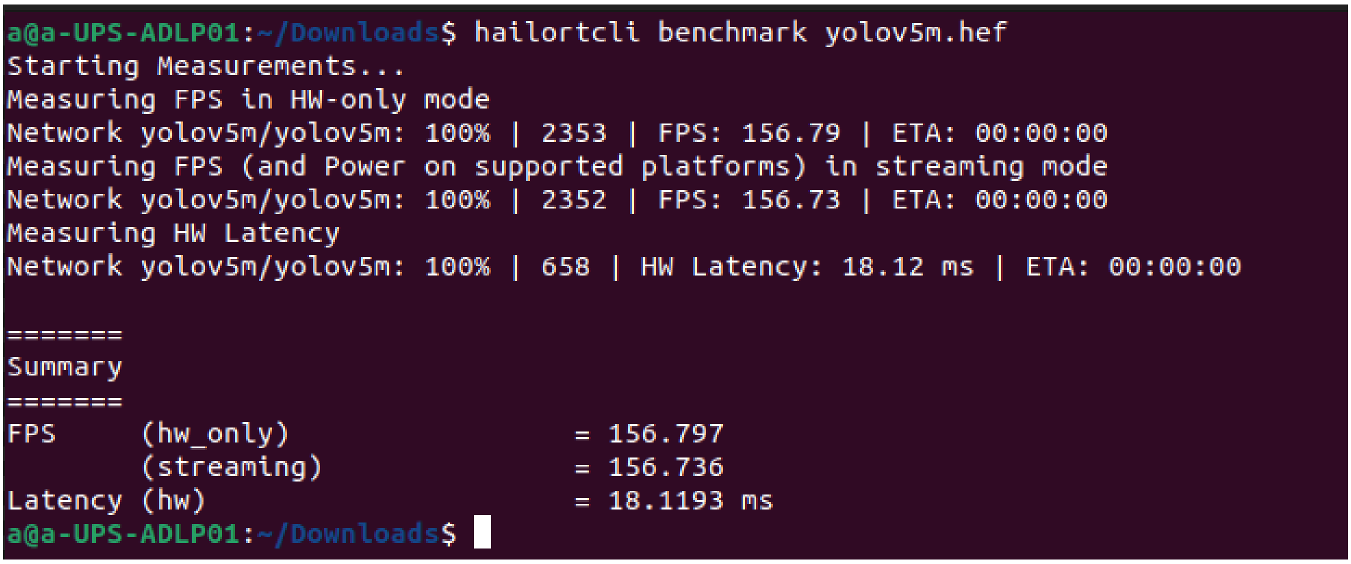
This is also true for operating system compatibility, with the toolkit being supported on multiple Windows and Ubuntu versions according to AI acceleration hardware needs, as shown below:

AI Accelerator	Windows 10 21H2 LTSC	Windows 11 24H2 LTSC	Ubuntu 22.04	Ubuntu 24.04
Hailo 8	✓	✓	✓	✗
Intel NPU	✗	✓	✓	✓
Intel GPU	✓	✓	✓	✓
Nvidia GPU	✓	✓	✓	✓

The key takeaway from this is that the benefits offered by the UP AI Toolkit's shortening of the environment setup and model conversion processes can be leveraged by users integrating models built on frameworks for a variety of different performance and efficiency needs

Reliable Benchmarking

Before deployment, developers need to evaluate AI model performance across different platforms. However, due to the complexity of simulation tools or workflows, performance evaluation often presents as a roadblock, slowing down the route the proof of concept due to the effort expended on research to find the most suitable hardware platform.



```
a@a-UPS-ADLP01:~/Downloads$ hailortcli benchmark yolov5m.hef
Starting Measurements...
Measuring FPS in HW-only mode
Network yolov5m/yolov5m: 100% | 2353 | FPS: 156.79 | ETA: 00:00:00
Measuring FPS (and Power on supported platforms) in streaming mode
Network yolov5m/yolov5m: 100% | 2352 | FPS: 156.73 | ETA: 00:00:00
Measuring HW Latency
Network yolov5m/yolov5m: 100% | 658 | HW Latency: 18.12 ms | ETA: 00:00:00

=======
Summary
=======
FPS     (hw_only)                 = 156.797
        (streaming)               = 156.736
Latency (hw)                      = 18.1193 ms
a@a-UPS-ADLP01:~/Downloads$ 
```

Combating this, the UP AI Toolkit's standardized benchmarking suite collects metrics across platforms and aggregates results in a structured format (JSON/CSV), regardless of the platform in question.

This effectively solves the problems that commonly arise due to there being no standardized way of benchmarking the performance of models across different hardware platforms.

With the introduction of the UP AI Toolkit, users not only benefit from hardware-agnostic model performance evaluation, but a vastly reduced time from model conversion to application deployment. This impact reflects the broader aim of the product, which is to accelerate how quickly and accurately users can develop and deploy AI models on AAEON edge computing platforms

Conclusion

With the introduction of the UP AI Toolkit, AAEON’s UP brand will allow users to streamline AI development and deployment for edge computing applications in a unique way. Given the brand’s position within the developer community, the UP AI Toolkit is a product offering that embodies its value proposition of helping its customers “Bridge the Gap from Idea to Success”.

By addressing the three pain points of the roadblocks caused by complex environment setups, the configuration required to successfully convert AI models, and the lack of standardized performance benchmarking, the toolkit is tailored to increase the speed with which AI applications can be built and deployed.

As a result of this toolkit, AAEON’s UP brand positions itself as a unique provider of a comprehensive AI ecosystem that will act as a conduit to cross-vendor model development and deployment. This development will crucially not be limited to specific UP hardware, with boards suitable to both efficient, low power AI tasks and high-performance, heavy duty AI inference supported.

The UP AI Toolkit Quick Start Guide

Download the latest version v1.0.0 and unzip it to up-ai folder.

Windows Installation

1. Navigate to installation directory: up-ai
2. Run `prepare.bat`
   - System reboot required after installation
   - Internet connection required
3. Launch application with `Start_app.bat`
   - Follow prompts to select demo type and hardware

Linux Installation

1. Navigate to installation directory: `cd up-ai`
2. Give `prepare.sh` and `start_app.sh` execution permissions

   ```
   chmod +x prepare.sh start_app.sh
   ```
3. Run `prepare.sh`
   - Select option "2" for automatic installation
   - System reboot required after installation
   - Internet connection required
4. Launch application with `start_app.sh`
   - Follow prompts to select demo type and hardware

     ```
     ./start_app.sh
     ```

Getting Started with the UP AI Toolkit's Pre-Built Models

To give users an introduction to the UP AI Toolkit and its capabilities, its GitHub Wiki contains two pre-built models to showcase inference workflows from common model architectures and UP hardware platform configurations.

These include:

- A chatbot that utilizes advanced Natural Language Processing (NLP) technology to:
  - Process and understand user text input
  - Generate contextually relevant responses
  - Enable natural conversational flow
  - Handle complex semantic and grammatical structures
- A real-time object detection model with:
  - Live video and camera feed processing
  - Multi-object detection and tracking
  - Support for various object classes (people, vehicles, animals, etc.)
  - Real-time performance optimization

Accessing the UP AI Toolkit's Pre-Built Models

1. Visit the AI Examples Section
   Navigate to: up-division/up-ai Wiki
   → Look for **Demo Library: Chatbot & Object Detection**
2. Install the Runtime Environment
   Follow the Quick Start Guide to download and set up the toolkit.
3. Select and Run an Example
   Choose a use case (e.g., object detection), then follow the step-by-step SOP to:
   - Load a pre-converted model.
   - Deploy it to a UP device (e.g., UP Xtreme, UP Squared).
   - Start inference using a connected camera or terminal interface.
4. Customize and Extend
   Developers can replace sample models with their own converted ones using the Model Converter Tool, or modify the inference logic to better align with specific application needs.

About UP

UP Bridge the Gap is a brand founded by AAEON Technology Europe in 2015, which since its inception has strived to produce developer platforms for all, becoming one of the developer community's most trustworthy and innovative brands. UP is committed to providing professional developer platforms to help its customers accelerate and bridge the gap between initial concept and mass production for professional developers.

About AAEON

Established in 1992, AAEON is one of the leading designers and manufacturers of industrial IoT and AI Edge solutions. With continual innovation as a core value, AAEON provides reliable, high-quality computing platforms including industrial motherboards and systems, rugged tablets, embedded AI Edge systems, uCPE network appliances, and LoRaWAN/WWAN solutions. AAEON also provides industry-leading experience and knowledge to provide OEM/ODM services worldwide. AAEON works closely with premier chip designers to deliver stable, reliable platforms. For an introduction to AAEON's expansive line of products and services, visit www.aaeon.com.

AAEON Always Agile, Always Ahead.

Follow Us

Facebook YouTube LinkedIn X