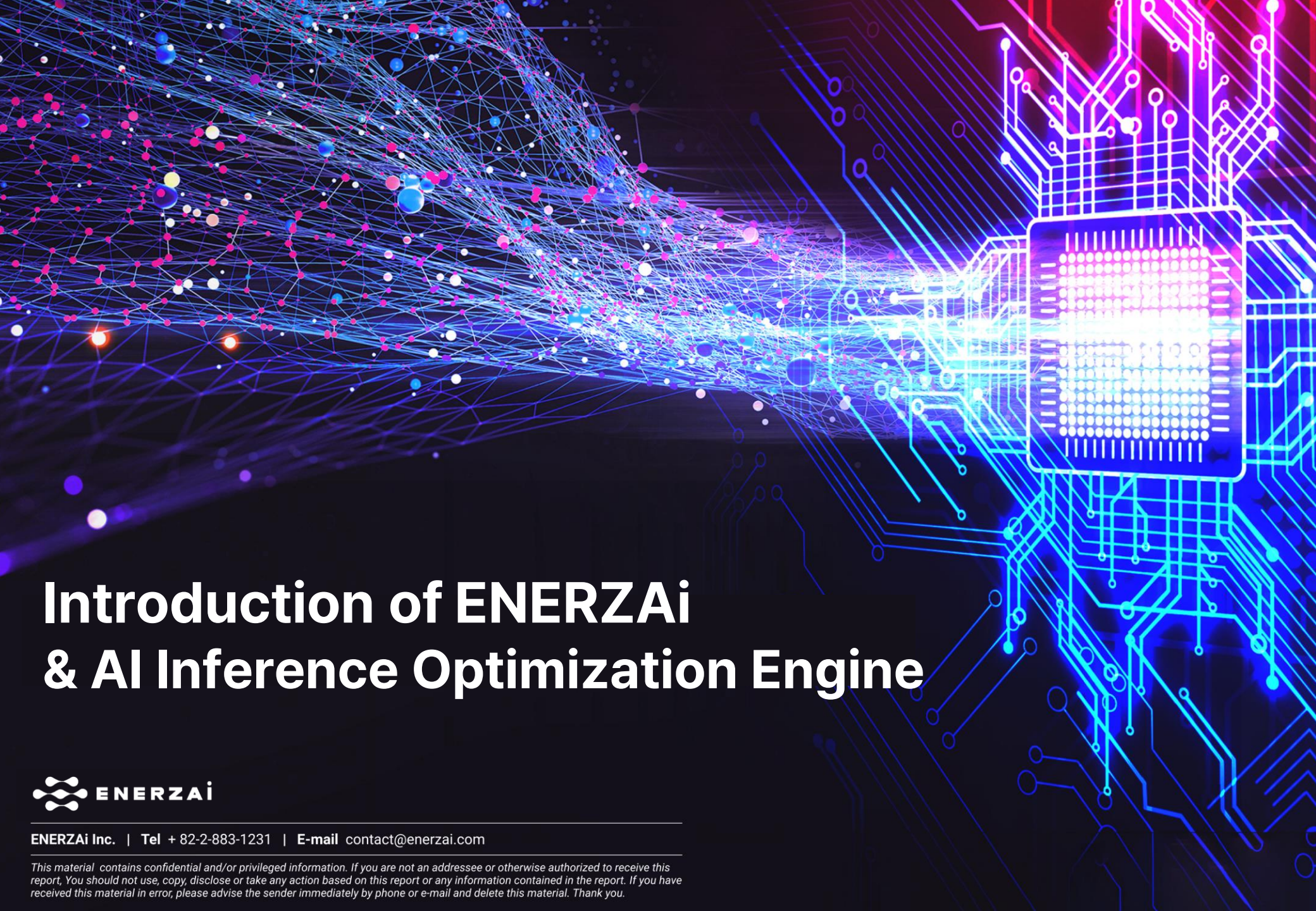


Introduction of ENERZAi & AI Inference Optimization Engine

ENERZAi

ENERZAi Inc. | **Tel** + 82-2-883-1231 | **E-mail** contact@enerzai.com

Company Overview

Team

- ENERZAi Inc. founded in Jan. 2019
- Members from top universities(SNU & KAIST), as well as major companies (Samsung & SK)
- Seed funding from NAVER affiliate VC(SpringCamp) and selected as beneficiary of TIPS by Korean govn't
- Raised $3.0M in Series A funding

Partners

intel partner Gold
Microsoft
arm
Accelerated by SAMSUNG
SK telecom
NVIDIA
Qualcomm

Technology

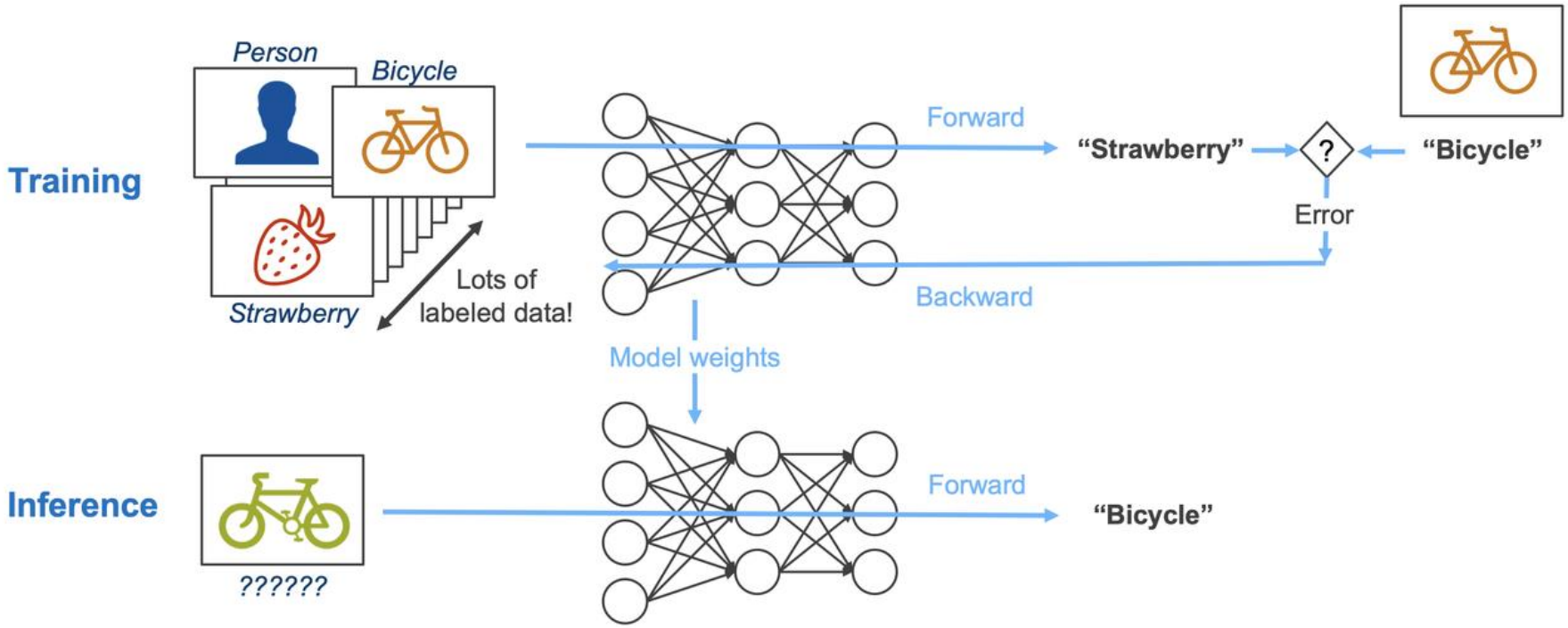
- AI model compression that makes AI model smaller while maintaining accuracy
- Low-level optimization that maximizes AI model performance for target H/W

Reference

- Selected as Intel Partner Alliance Gold & Samsung C-Lab Outside
- Ranked 1st and runner-up in 2 tracks at Mobile AI & AIM 2022 Challenge
- Ranked 3rd in 2 tracks at CVPR 2021 - Mobile AI Workshop

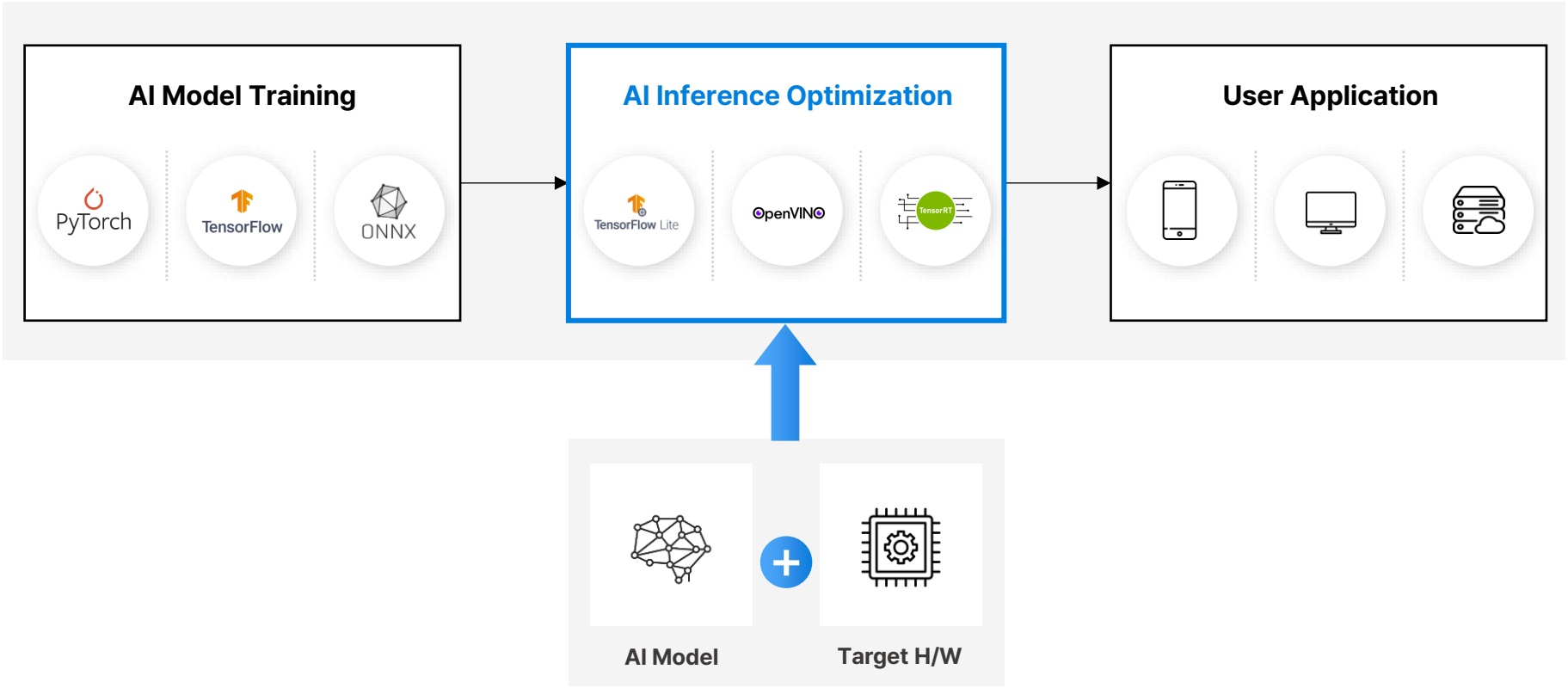
Background (1/2) – What's AI Inference?

Utilize a trained AI model to make predictions or decisions



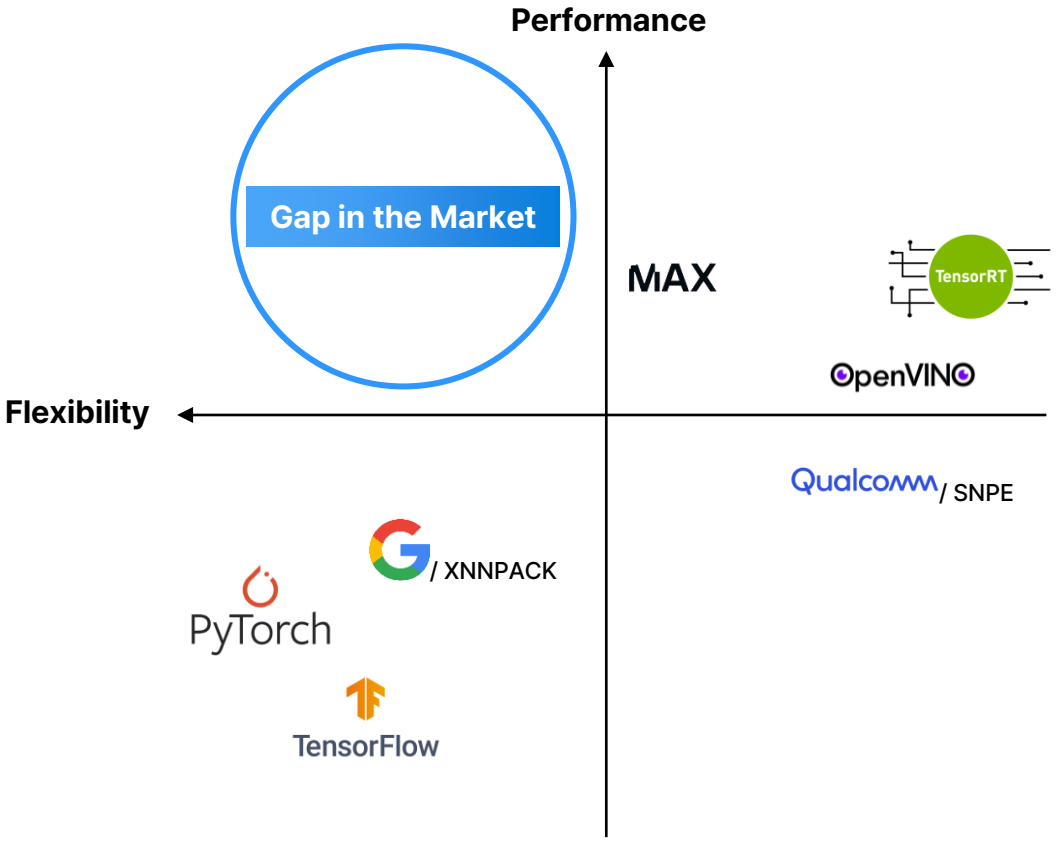
Background (2/2) – What's AI Inference Optimization?

Ensure that a trained AI model achieve high inference performance



Problem (1/3) - Overview

No existing tools capable of proper AI inference optimization



1 Low Performance

> *"Inference speed dropped more than expected on-device, costing us too much time for extra work to enable real-time processing.*
>
> [VR Device Manufacturer A Interview]

2 Limited Flexibility

> *"As more hardware platforms are incorporated into cars, AI model deployment becomes increasingly challenging due to the use of different inference optimization engines.*
>
> [Tier-2 Company B Interview]

Problem (2/3) – Low Performance

Slow inference speed reduces competitiveness and increases costs

1 Low Performance

- Using existing AI inference engines results in slower inference speed than expected
- Slow inference speed reduces product/service competitiveness and increases costs

Decline in product & service competitiveness

Critical to real-time applications including autonomous driving

Increase in costs

The Next Platform
https://www.nextplatform.com › AI

The Battle Begins For AI Inference Compute In The ...

2024. 9. 10. — **The high cost of AI inference** in the datacenter is, in fact, a major gating factor for the rollout of GenAI in the enterprise to enhance ...

Data Center Dynamics
https://www.datacenterdynamics.com › news

OpenAI training and inference costs could reach $7bn for ...

2024. 7. 24. — ... , as of March, **the company was set to spend nearly $4 billion this year on using Microsoft's servers to run inference workloads** for ChatGPT.

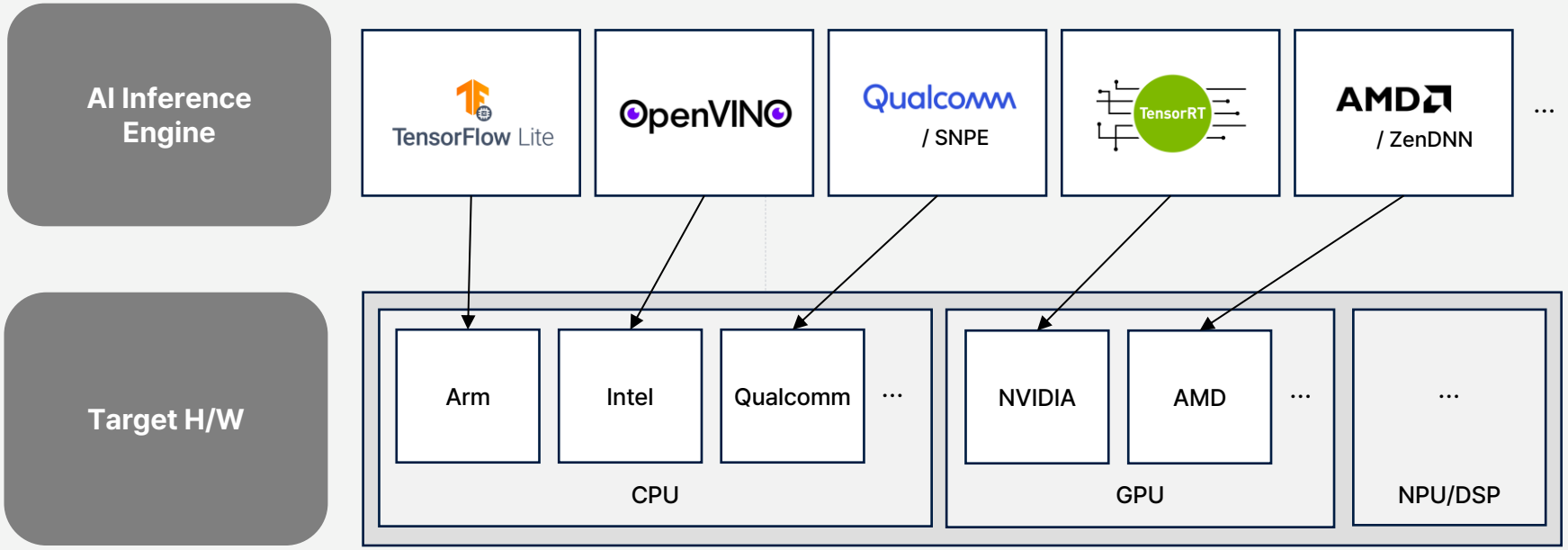
Critical to applications operating large AI models, such as generative AI models

Problem (3/3) – Limited Flexibility

Limited flexibility of existing tools hinders AI model deployment

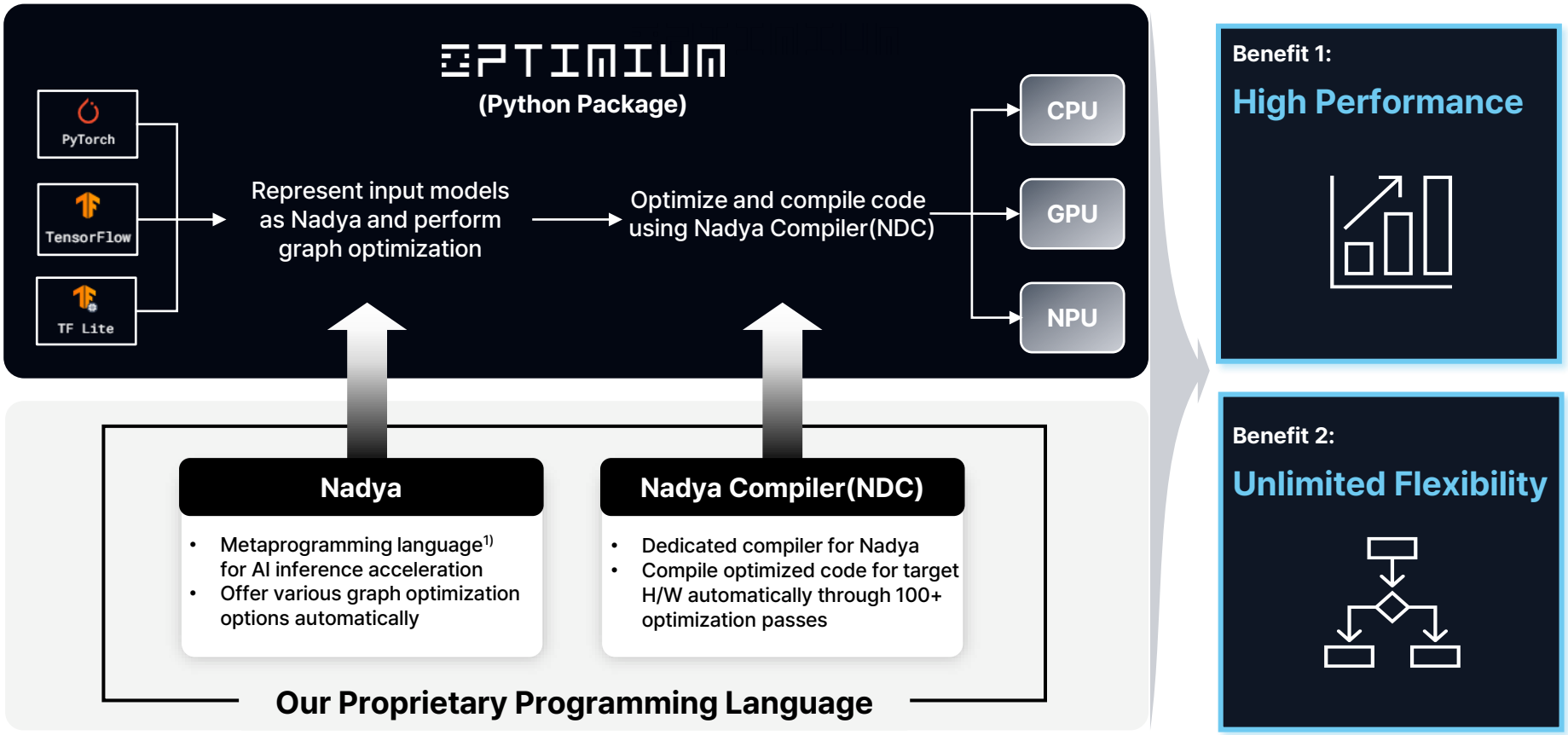
2 Limited Flexibility

- Different AI inference engines should be used for each hardware
- Deploying a single AI model across various hardware adds complexity



Solution (1/3) - Overview

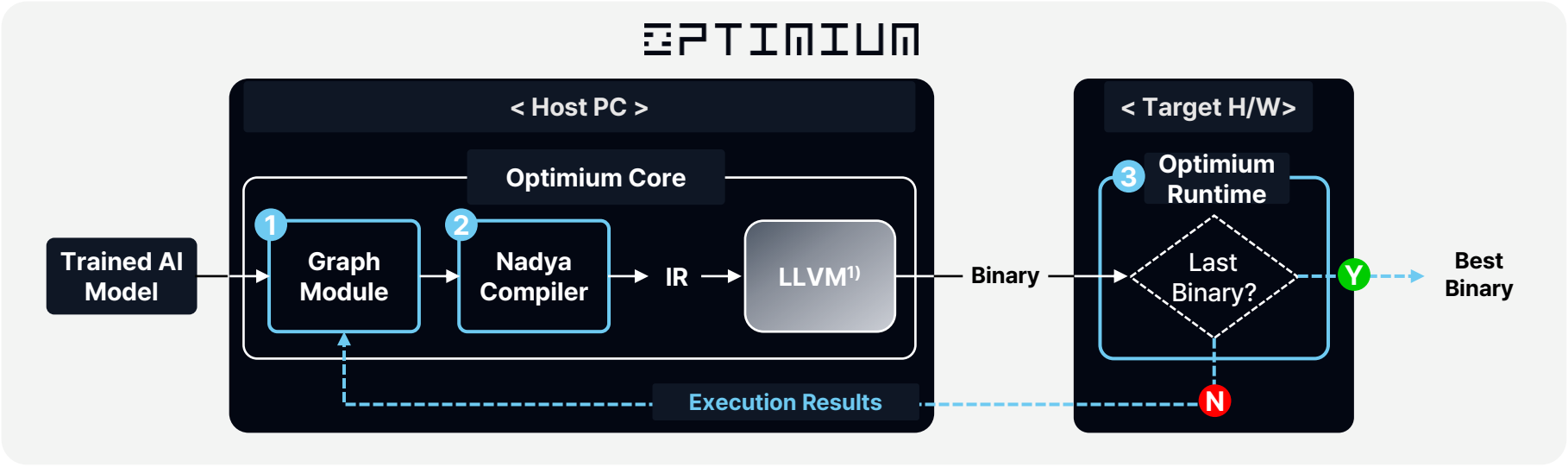
AI inference optimization engine based on our proprietary language



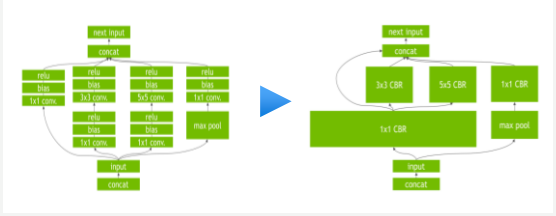
Note
1) A programming language where code(program) can automatically generate and modify new code(program).

Solution (2/3) - Workflow

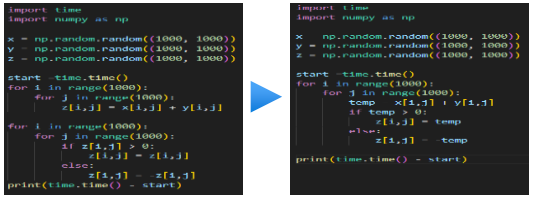
Each module performs tailored optimizations for target H/W



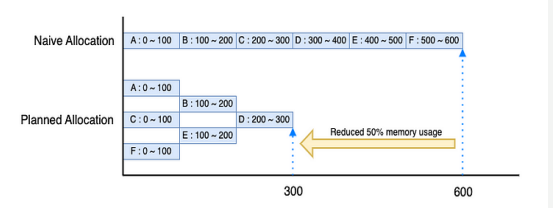
Module 1: Graph Module



Module 2: Nadya Compiler



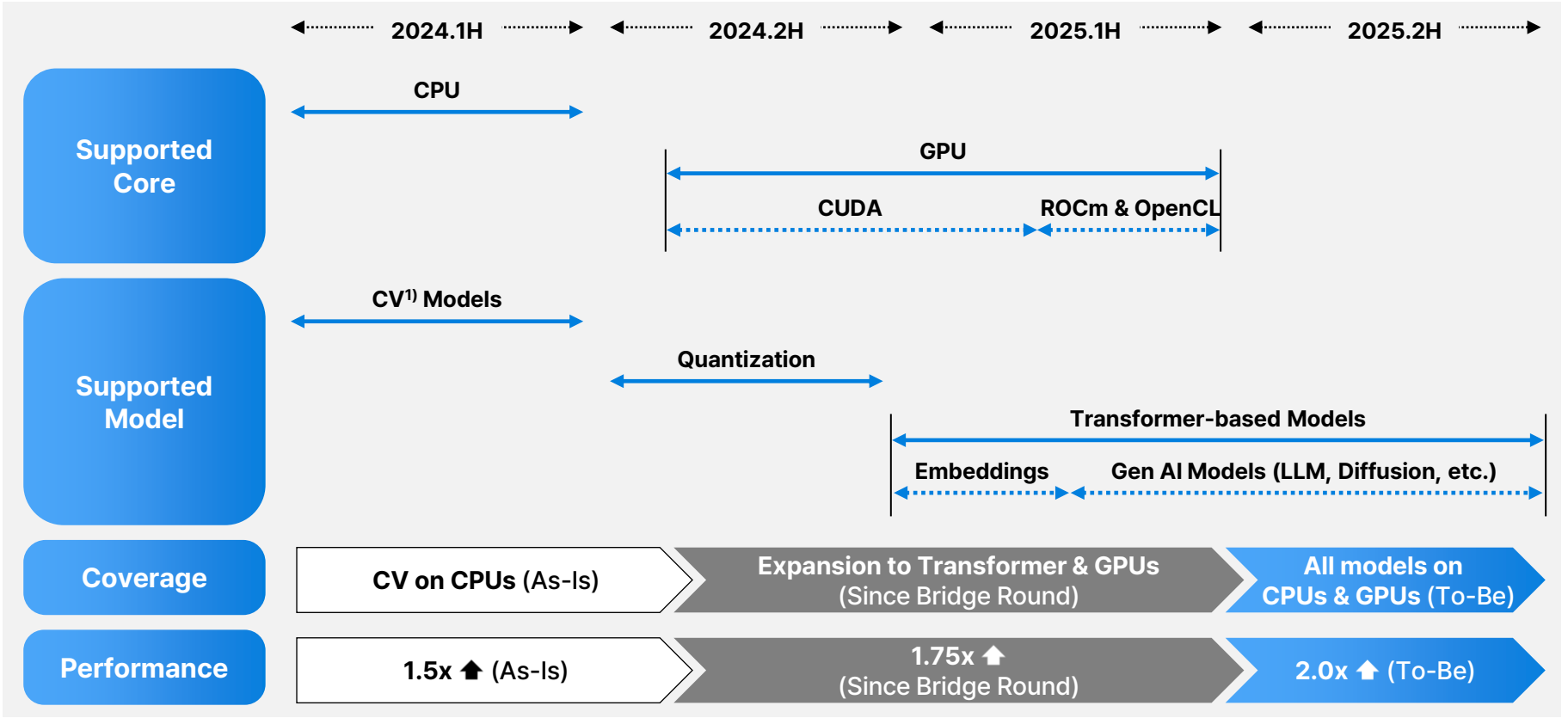
Module 3: Runtime Module



Note
1) An open-source compiler project primarily focused on middle-end/back-end compiler toolchain for generating machine code for modern programming languages

Solution (3/3) – Product Roadmap

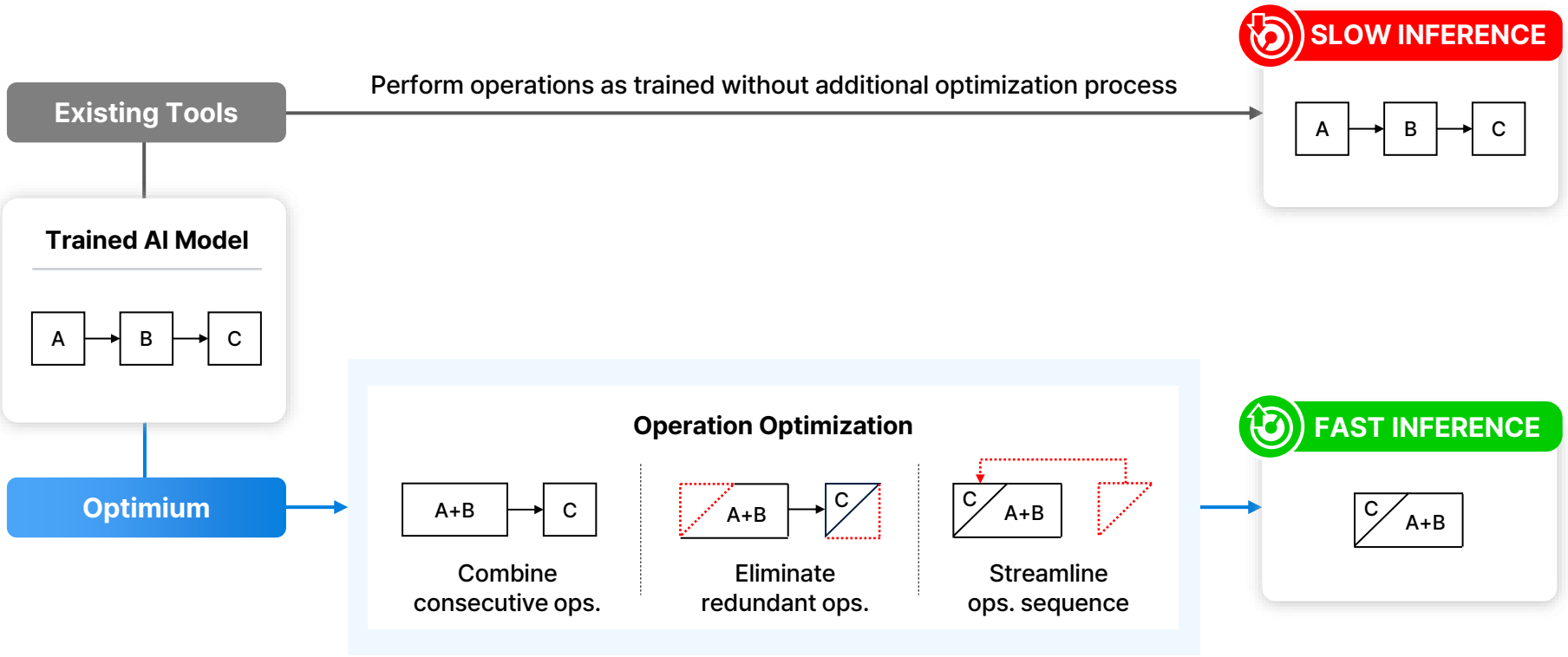
Plan to support optimization of Transformer models on GPUs by 2025



Note
1) Computer Vision

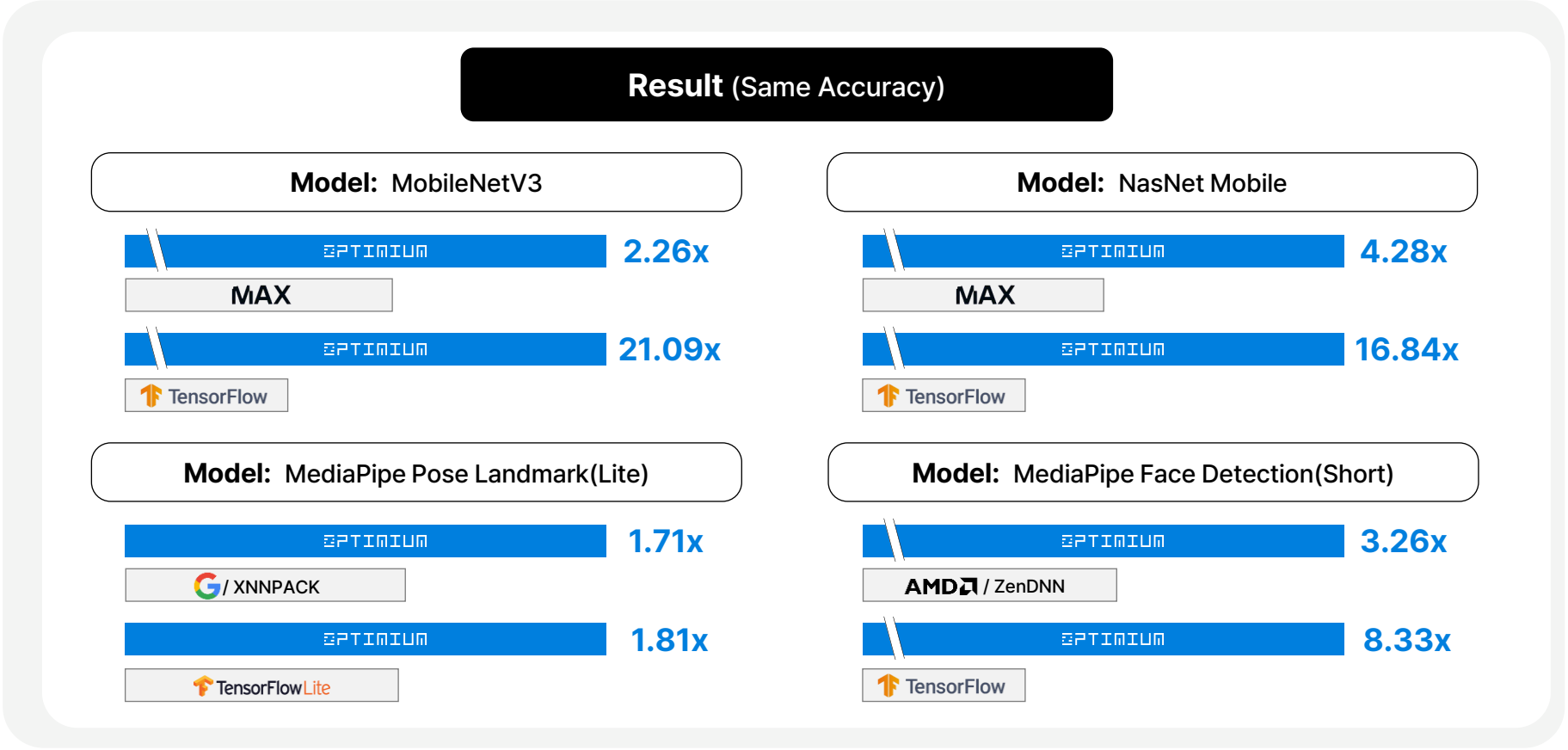
Competitiveness (1/2) – Performance

Significantly faster inference speed compared to existing tools



Competitiveness (1/2) – Performance (Cont'd)

Demonstrate superior performance over existing tools on various models

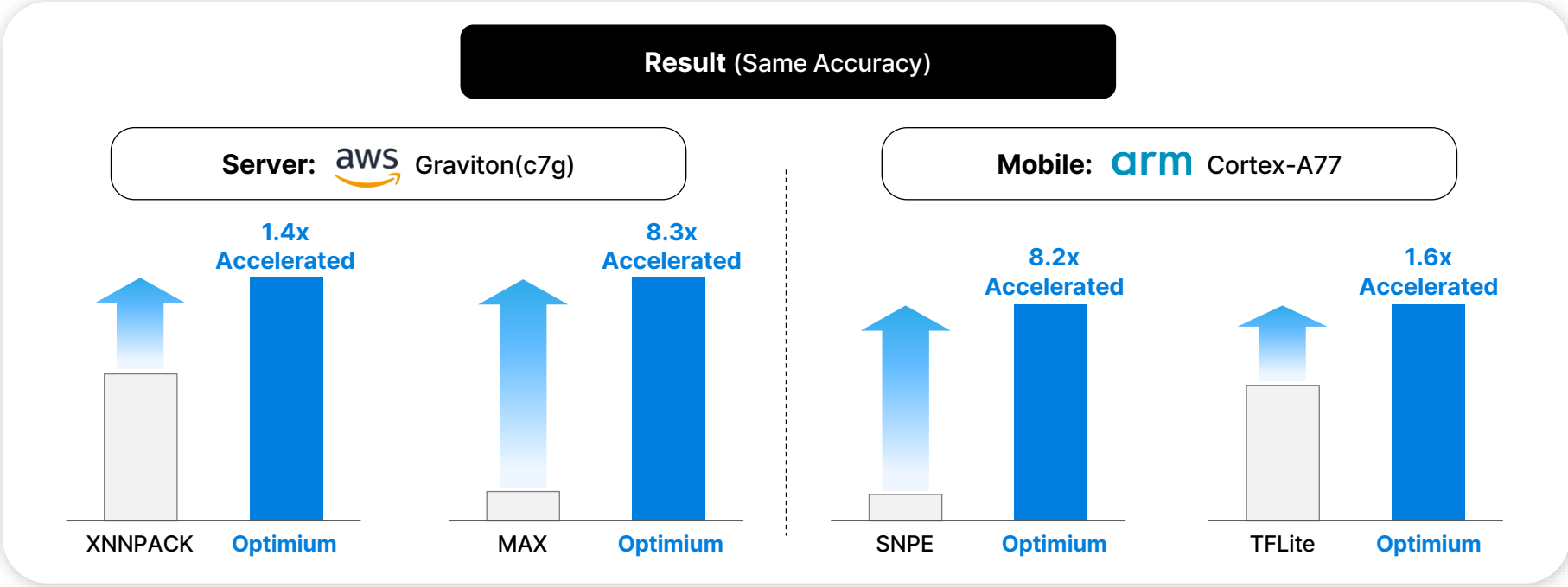


Competitiveness (1/2) – Performance (Cont'd)

Improved inference speed in diverse target H/W including mobile & server

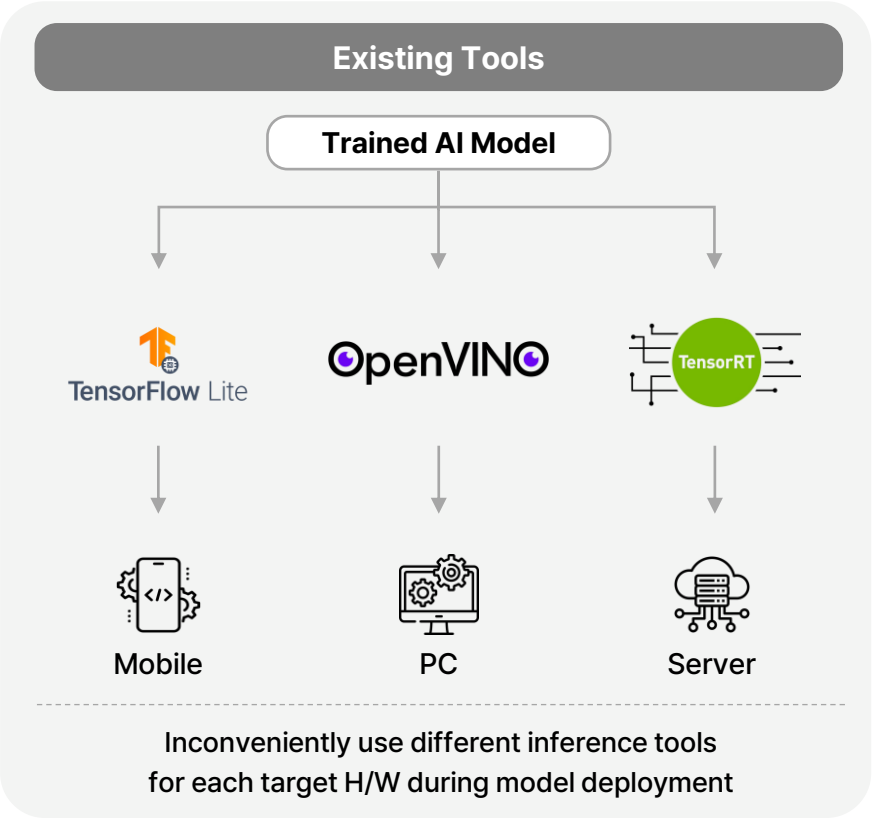
Access control solution company

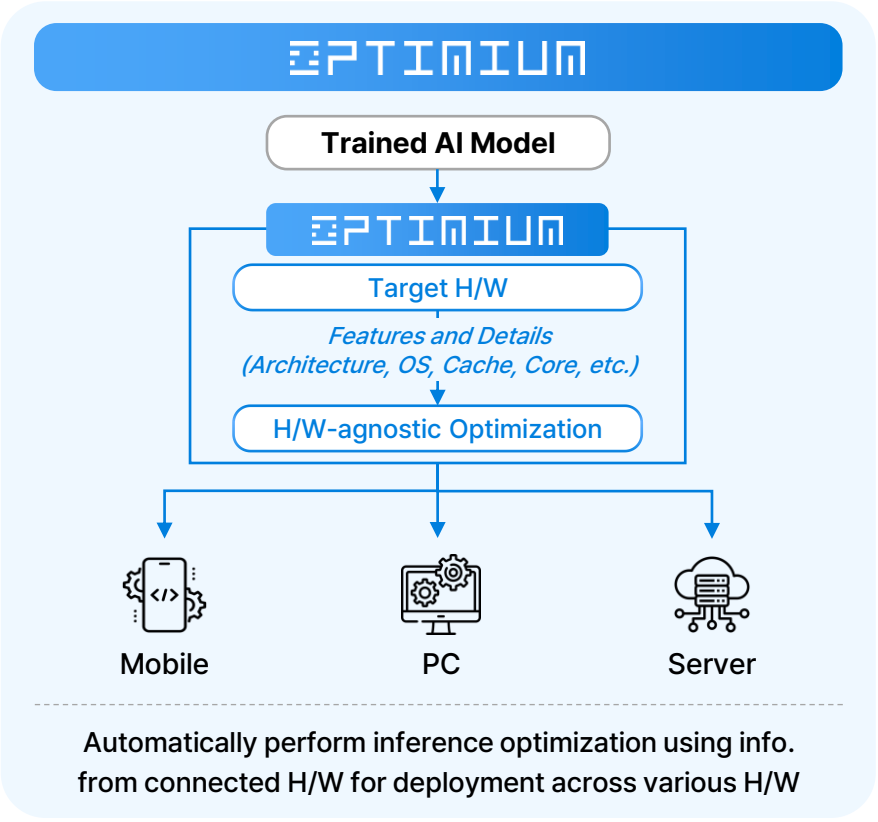
- Experienced entry/exit delays due to the low inference speed of the face recognition model
- With Optimium, the client accelerated facial recognition by enhancing inference speed across various H/W



Competitiveness (2/2) – Flexibility

Automatically deploy AI model across various H/W with a single tool



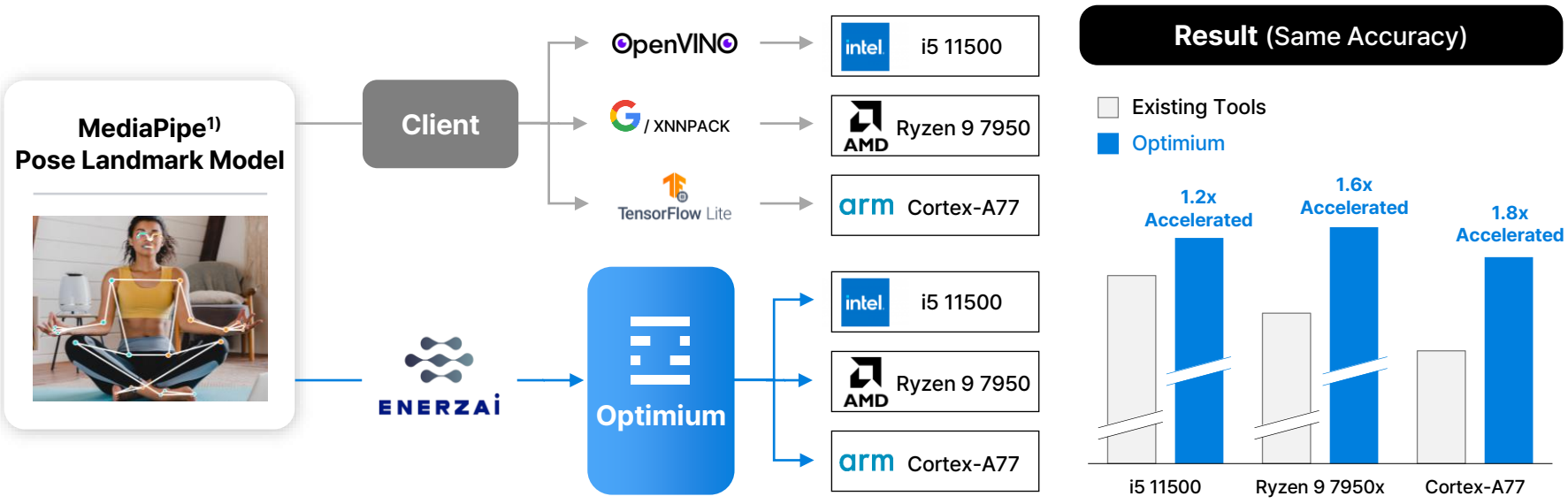


Competitiveness (2/2) – Flexibility (Cont'd)

Successfully deployed customized AI model on various H/W

Healthcare solution company

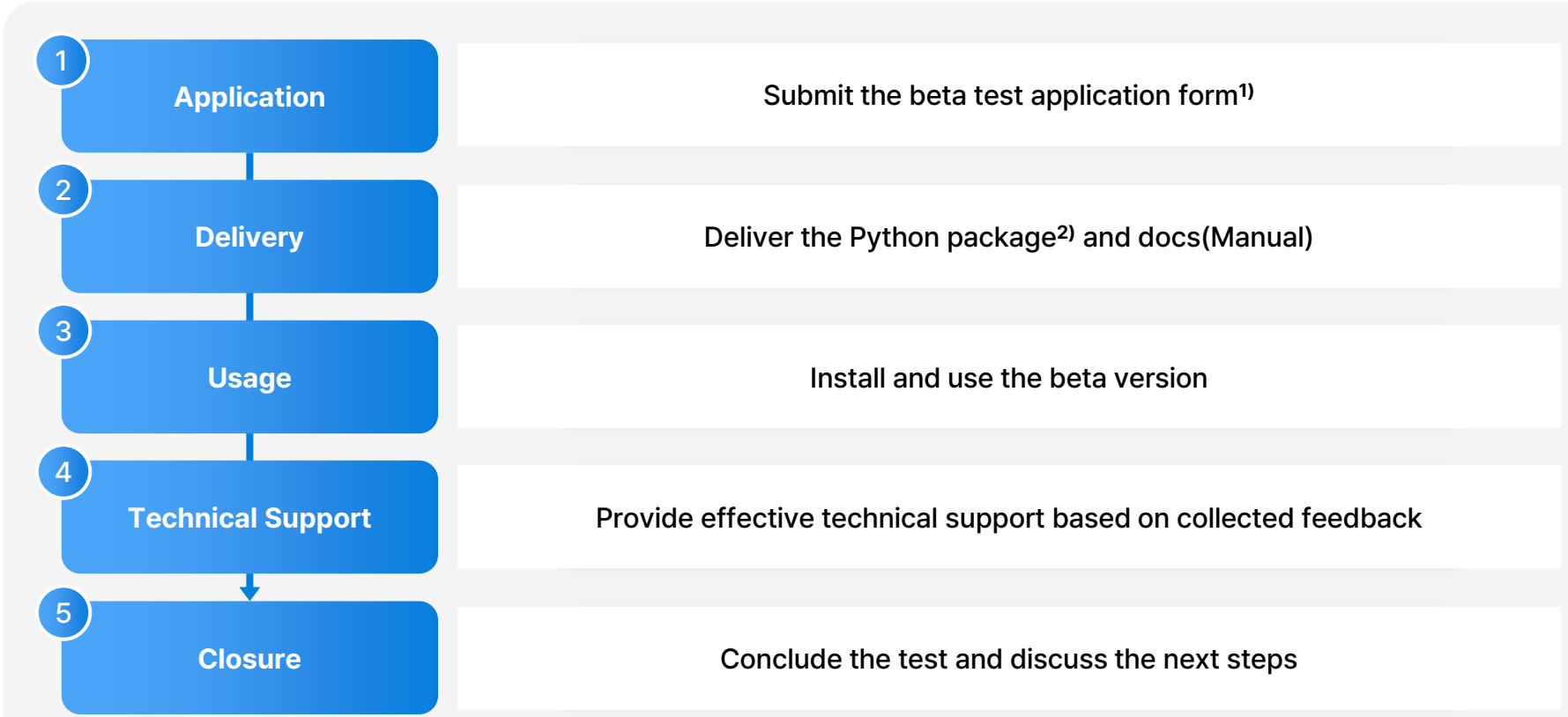
- Experienced problems with costly service updates due to the use of different inference tools for each H/W
- Optimium allows swift AI model deployment and service update for various H/W using a single tool



Note
1) Google's open-source framework for real-time ML learning applications such as face recognition, pose estimation, and palm recognition
2) Used OpenVINO Benchmark Tool and utilized measurements with initial 30 runs as warm-up and averaged latency for the next 970 runs

Beta Test Process

Beta test is conducted through the following process



Note
1) Beta test application link: https://wft8y29gq1z.typeform.com/to/fp059MY5
2) Including 30-day license keys for each company

End of Document

ENERZAi

ENERZAi Inc. | **Tel** + 82-2-883-1231 | **E-mail** contact@enerzai.com