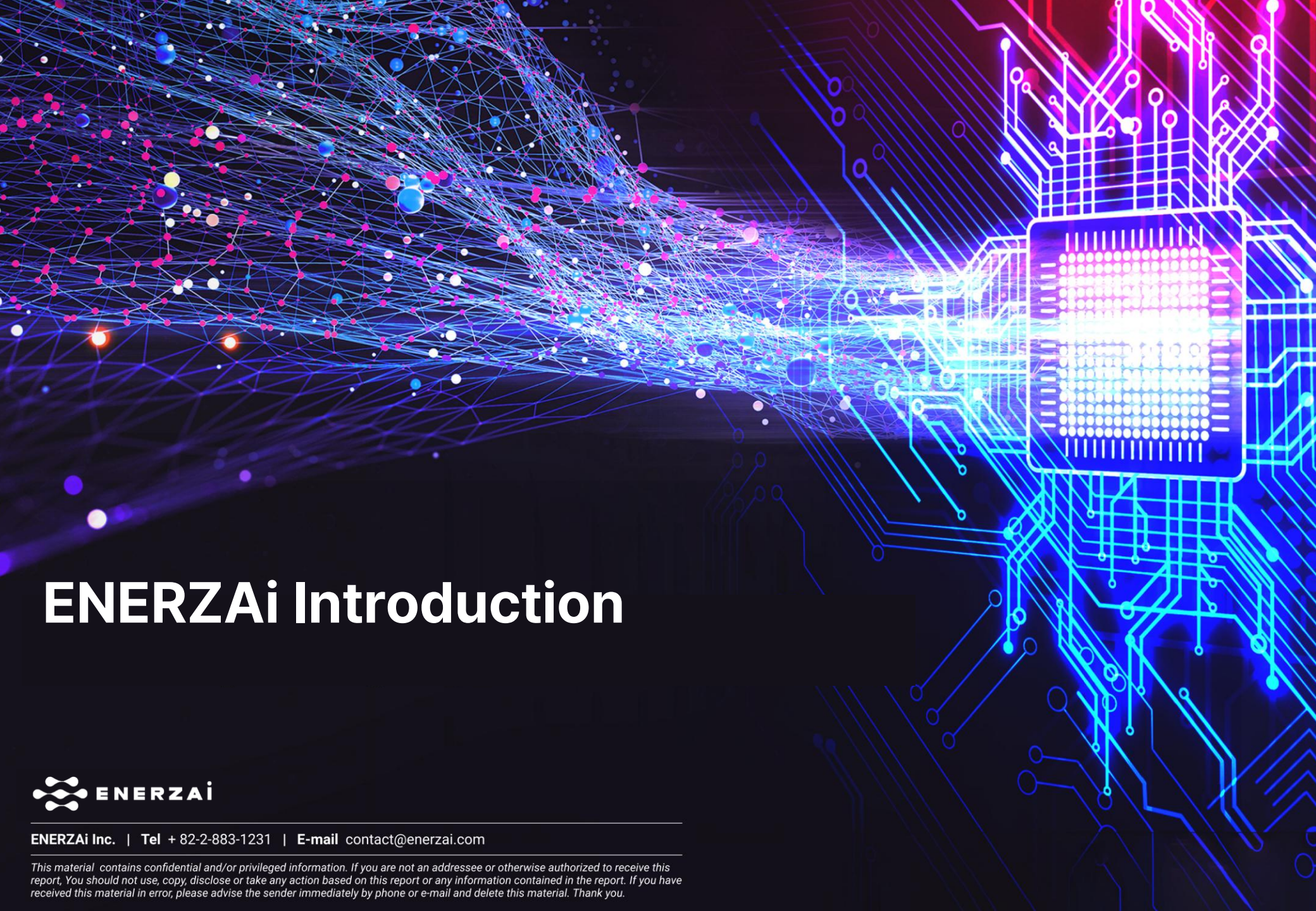


ENERZAi Introduction

ENERZAi

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

This material contains confidential and/or privileged information. If you are not an addressee or otherwise authorized to receive this report, You should not use, copy, disclose or take any action based on this report or any information contained in the report. If you have received this material in error, please advise the sender immediately by phone or e-mail and delete this material. Thank you.

Company Overview

Team

- ENERZAi Inc. founded in Jan. 2019
- 26 members from top universities, as well as major companies
- Raised a total of +$9M in funding
- Selected as beneficiary of TIPS & Post-TIPS by Korean government
- Patent: 28 registered, and 23 applied

Product

- On-device AI model: Low-bit, high-performance model that can run on-device
- Optimium: Next-generation AI inference engine with superior inference performance and hardware flexibility

Partners

Accelerated by SAMSUNG

INCEPTION PROGRAM

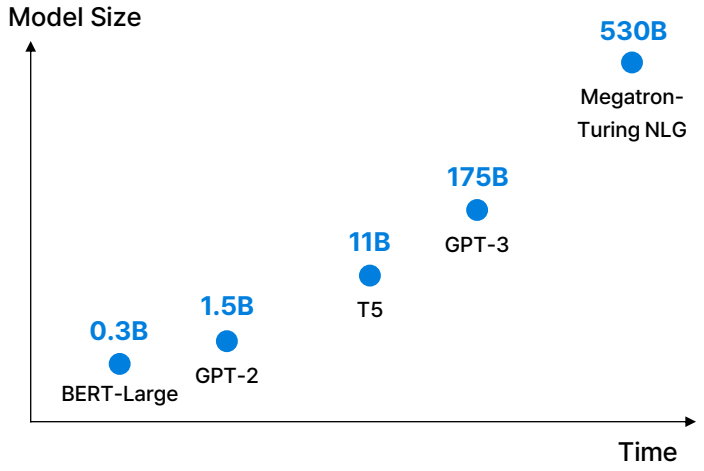
Reference

- Ranked 1st and runner-up in 2 tracks at Mobile AI & AIM 2022 Challenge
- Selected for 'Google for Startups Accelerator' & 'Microsoft Startup Program'
- Selected as the winner of the 2025 Edge AI and Vision Product of the Year Award

Problem

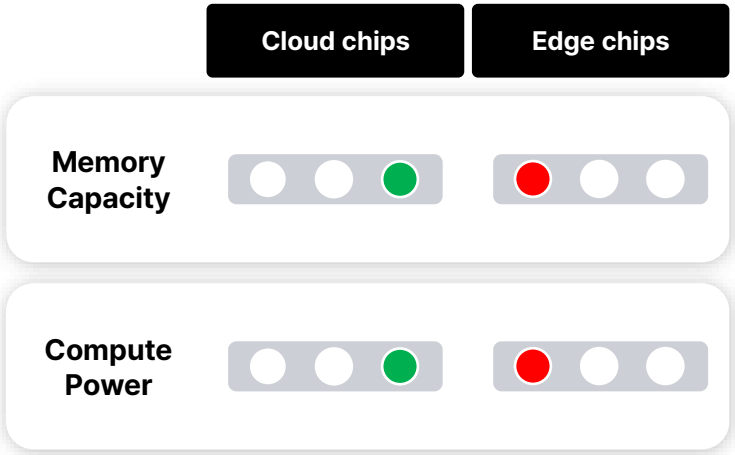
The gap between AI model sizes and edge hardware specs is widening

Resource-Intensive AI



AI models become larger and more complex,
requiring much higher hardware resources

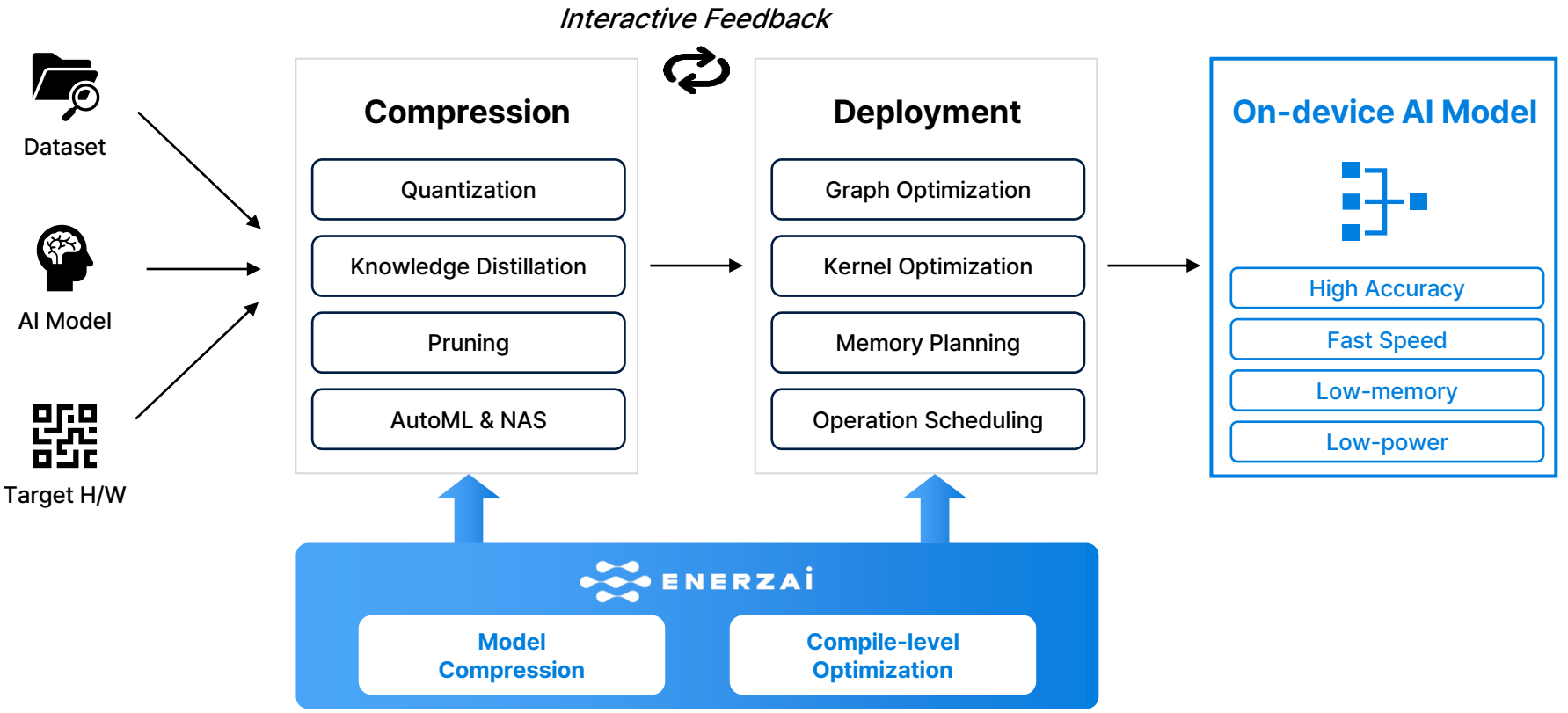
Limited Hardware Resources



Most edge devices use low-spec chips
with limited memory and compute power

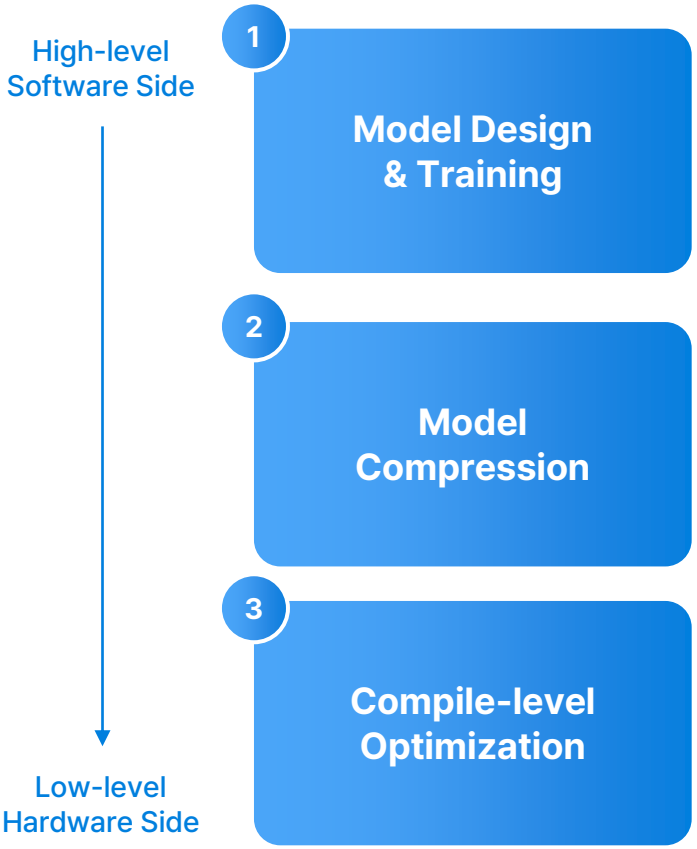
Technology (1/4) – Overview

We are the only company with full-stack Edge AI software technology



Technology (1/4) – Overview (Cont'd)

We are the only company with full-stack Edge AI software technology



- Wide expertise across models — from image and video enhancement and computer vision models to transformers
- Mainly focusing on voice & language models

- Expertise in quantization — particularly in enabling and commercializing extreme low-bit models, including 1.58-bit implementations for transformer architectures

- Building Optimium — the most powerful MLIR-based AI compiler when it comes to deploying AI models on edge, especially on Arm-based SoCs

Technology (2/4) – Competition

Awarded at worldwide AI competitions, surpassing big tech firms

Real Image Denoising Challenge

Result

Target H/W: SAMSUNG Exynos 2100

Rank	Team	Model Size (KB)	Accuracy (PSNR)	Runtime (ms)
1	Huawei	209	37.52	39
2	Megvii	14,276	37.83	84
3	ENERZAi	81	36.33	11
4	Xiaomi	1,404	37.37	54
5	ENERZAi	118	36.22	23

Learned Smartphone ISP[1)] Challenge

Result

Target H/W: MEDIATEK Dimensity 1000+

Rank	Team	Model Size (KB)	Accuracy (PSNR)	Runtime (ms)
1	Dahua	21	23.2	61
2	ITRI	123	23.73	90.9
3	ENERZAi	9	22.97	65
4	Dahua	175	22.78	77
5	Huawei	244	23.08	94.5

Note
1) ISP(Image Signal Processor) is a processor that converts Bayer-patterned raw image into RGB coming from an image sensor, and the goal of ISP challenge is to replace classical ISP with D/L based ISP pipeline

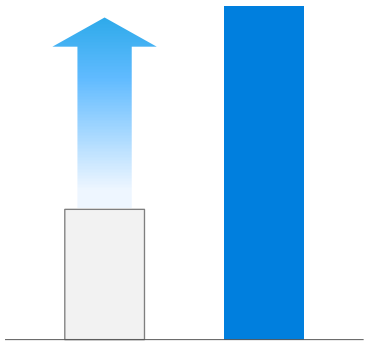
Technology (3/4) – Project

Achieved faster AI Inference speed in projects with global corporations

Project with a Global Smartphone Company

Result (Same Accuracy)

Target H/W: Qualcomm Snapdragon 888



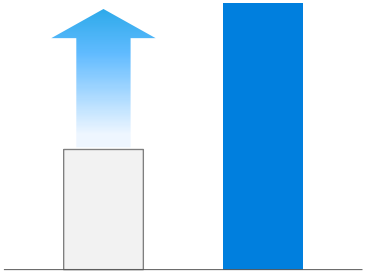
Inference Speed

2.5x

Project with an Automotive Tier 1 Company

Result (Same Accuracy)

Target H/W: Texas Instruments IWR 6843



Inference Speed

2.2x

Technology (4/4) – Patent & Recognition

Driving innovation with IP strength and global recognition

IP Strength

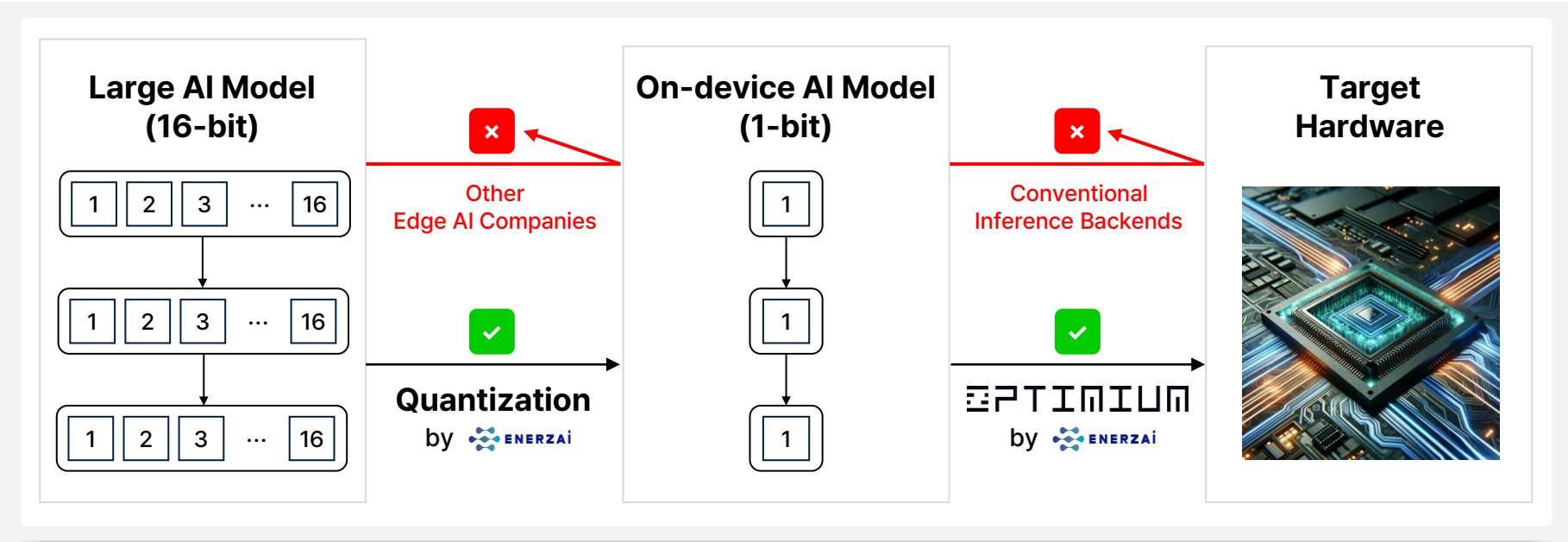
# of Granted Patents	28
# of Applications Filed	23
# of Applications to Be Filed	20+

Global Recognition

2025 EDGE AI AND VISION PRODUCT OF THE YEAR	*Awarded '2025 Edge AI and Vision Product of the Year Award'*
Google for Startups Accelerator	*Selected for 'AI First - Google for Startups Accelerator'*
Microsoft	*Selected for 'Microsoft Startup Program'*
aws	*Selected for 'AWS Gen AI US Delegation Program'*
NextRise 2025.Seoul	*Awarded 'Next Innovator Prize', NextRise Awards 2025*

On-device AI Model (1/4) – Solution

Provide the world's smallest AI models that run directly on edge devices



High Accuracy

Fast Speed

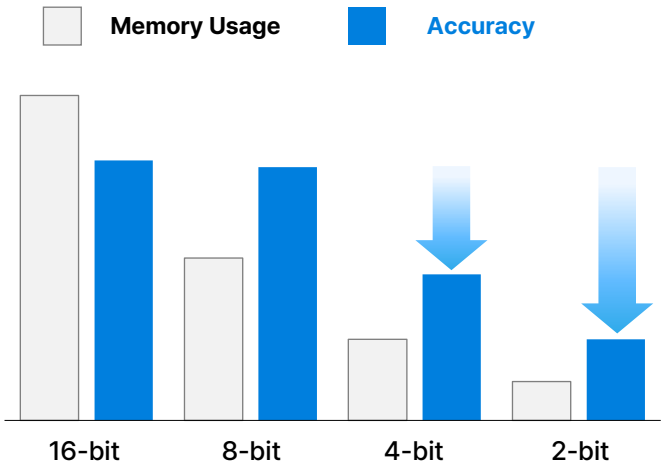
Minimal Memory Usage

Minimal Power Usage

On-device AI Model (1/4) – Solution (Cont'd)

Broke the technical barrier through our technology

Challenges in Low-bit Quantization



Although quantization can be effective,
it results in a sharp drop in accuracy

Inference Backend Gap

	16-bit	8-bit	Sub-4-bit
TensorFlow Lite	✓	✓	✗
ONNX RUNTIME	✓	✓	✗
PyTorch	✓	✓	✗
TensorRT	✓	✓	✗
OpenVINO	✓	✓	⚠

Even when quantization is successfully applied,
here is a lack of tools to flexibly run such models

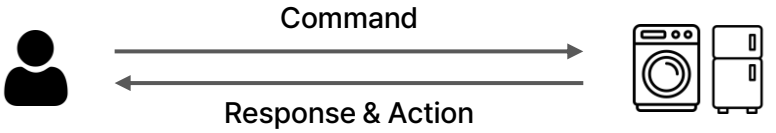
On-device AI Model (1/4) – Solution (Cont'd)

Overcome the limitations of cloud AI through on-device AI

Cloud AI

Continuous cost proportional to usage	🔴 ⚪ ⚪
Dependency on network environment	🔴 ⚪ ⚪
Risk of personal data leakage	🔴 ⚪ ⚪
Long latency during transfer	⚪ 🟡 ⚪

On-device AI



No additional cost after deployment	⚪ ⚪ 🟢
Applicable in offline environments	⚪ ⚪ 🟢
No risk of personal data leakage	⚪ ⚪ 🟢
No latency via on-device processing	⚪ ⚪ 🟢

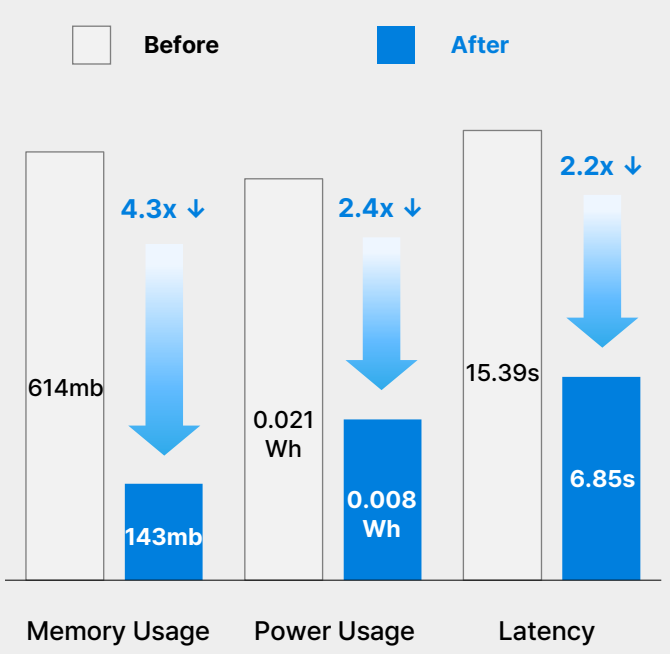
On-device AI Model (2/4) - Performance

Achieved dramatic memory & power efficiency, along with fast speed

1.58-bit Whisper[1] Demonstration (w/Synaptics)

Target H/W: synaptics SL1680

Result



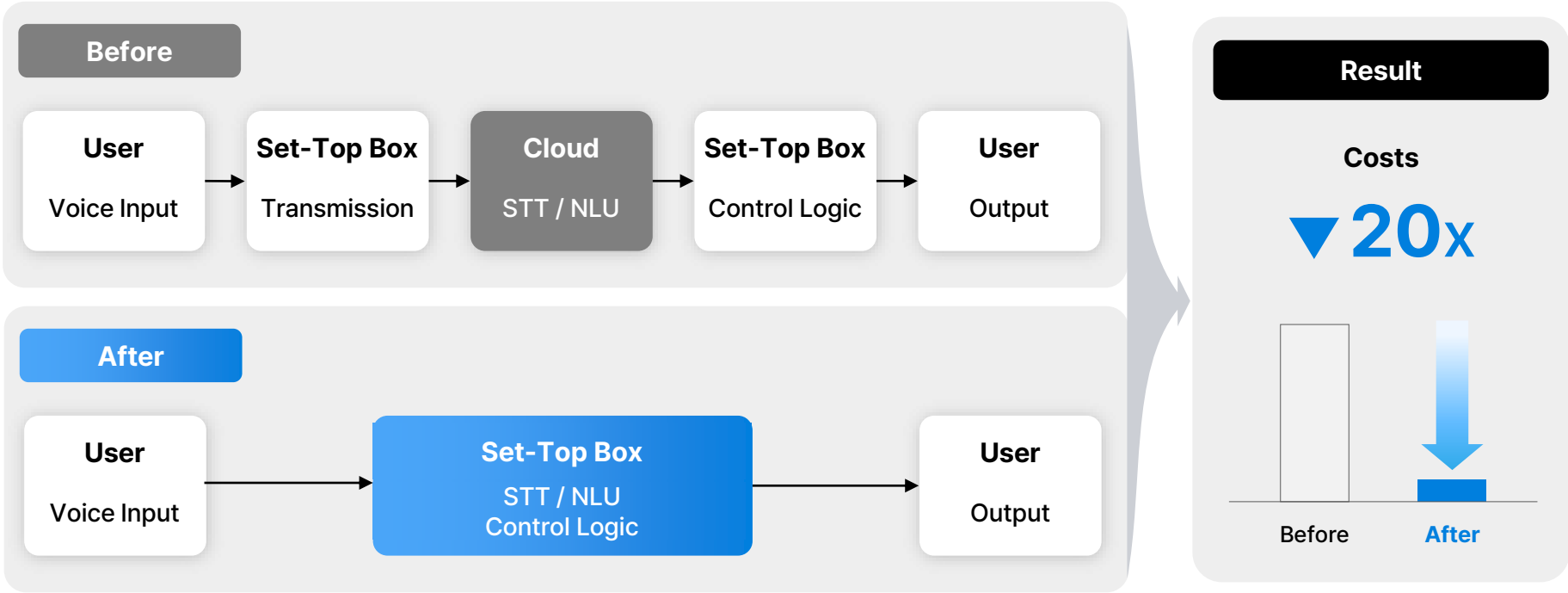
Note
1) Whisper is an automatic speech-recognition (ASR) model released as open source by OpenAI.
2) Demo video: https://www.youtube.com/watch?v=9Fa6JRqz4Hs

On-device AI Model (3/4) - Reference

Deployed on-device voice AI models on +2 million set-top boxes

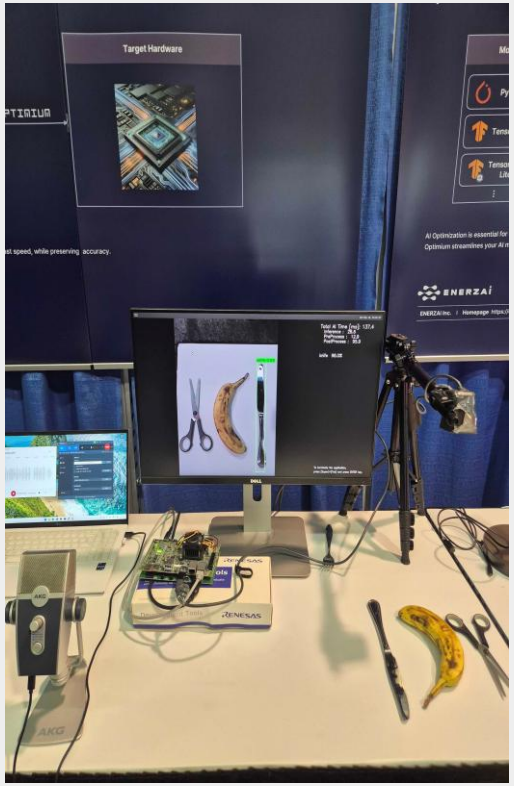
Telecom Company

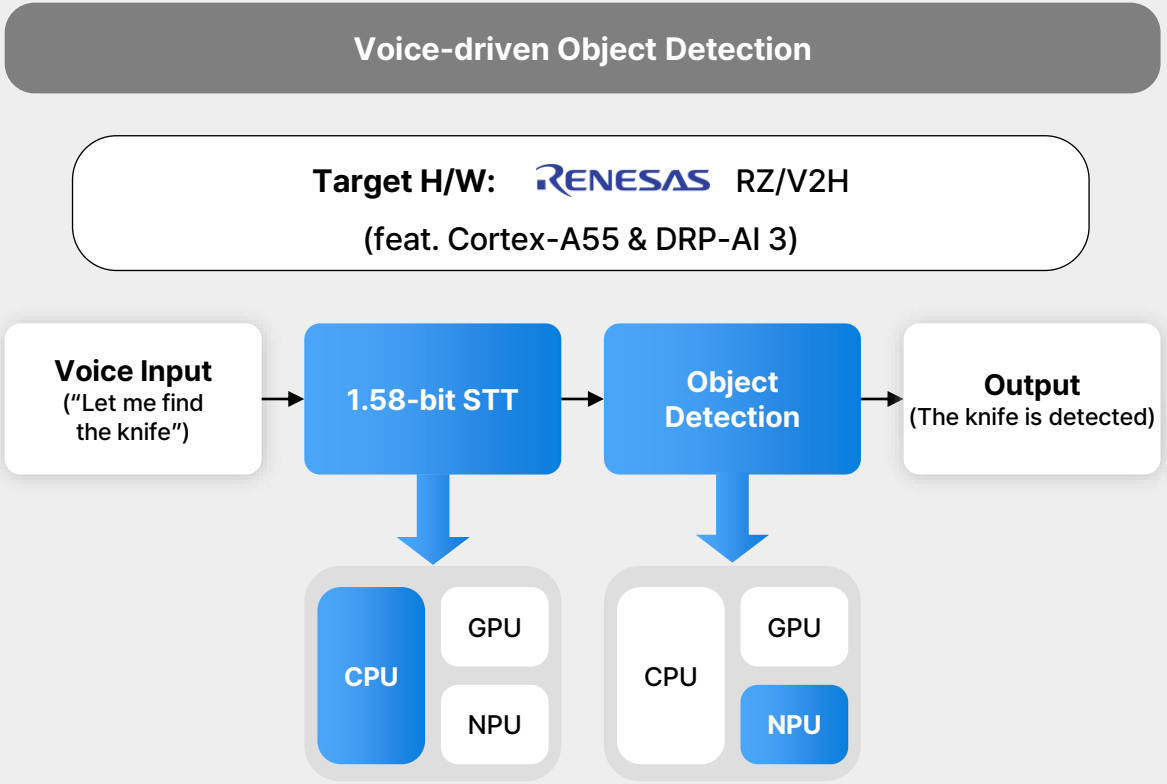
- The company had relied on a cloud-based voice AI model due to limited hardware resources
- We implemented the same functionality to run entirely on-device on a low-spec chip inside a Set-Top Box, using only 100MB of memory, which enabled significant cost reduction



On-device AI Model (4/4) – Key Demos

Live voice-driven object detection pipeline

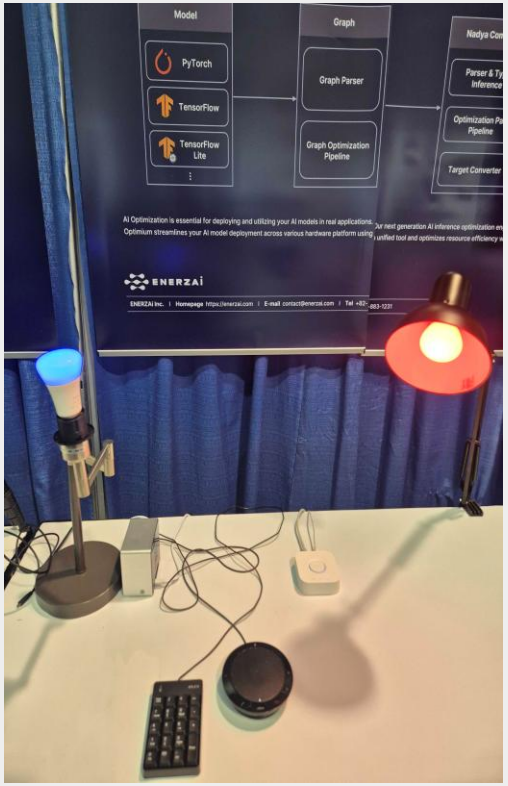


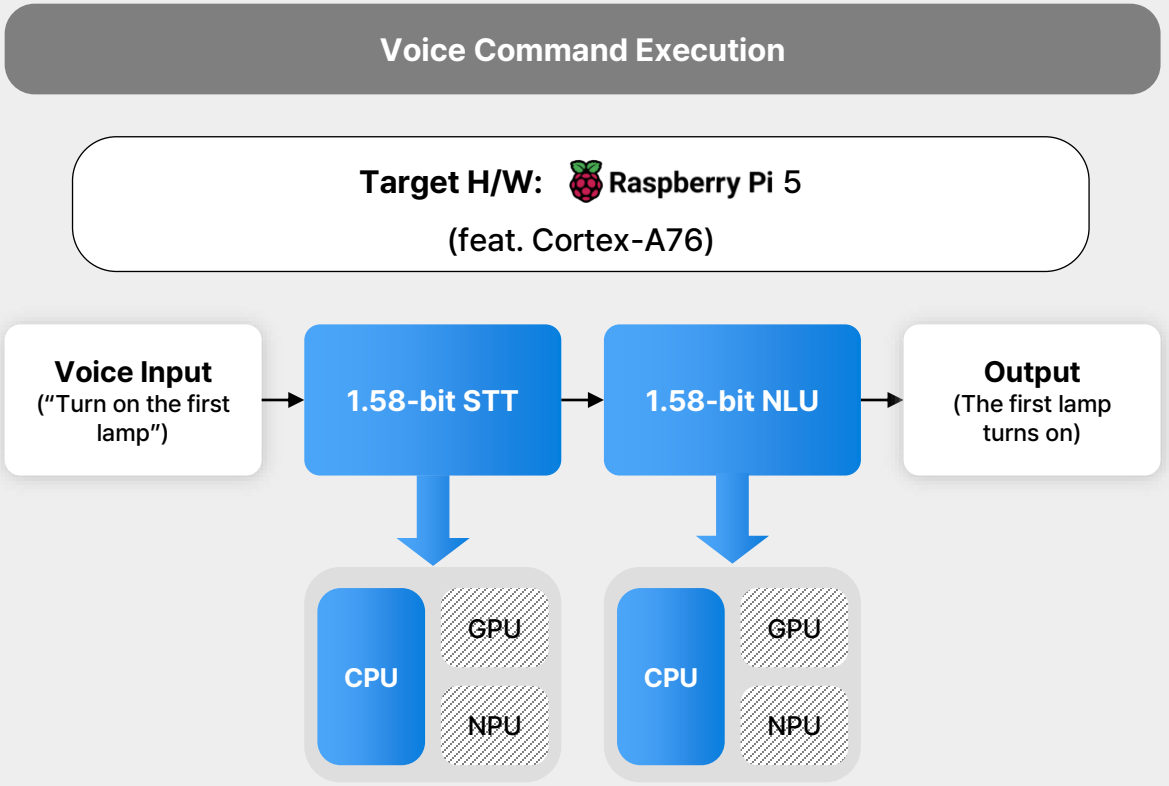


Note
1) Demo Video: https://www.youtube.com/watch?v=9Fa6JRqz4Hs&t=102s

On-device AI Model (4/4) – Key Demos (Cont'd)

Live voice control pipeline

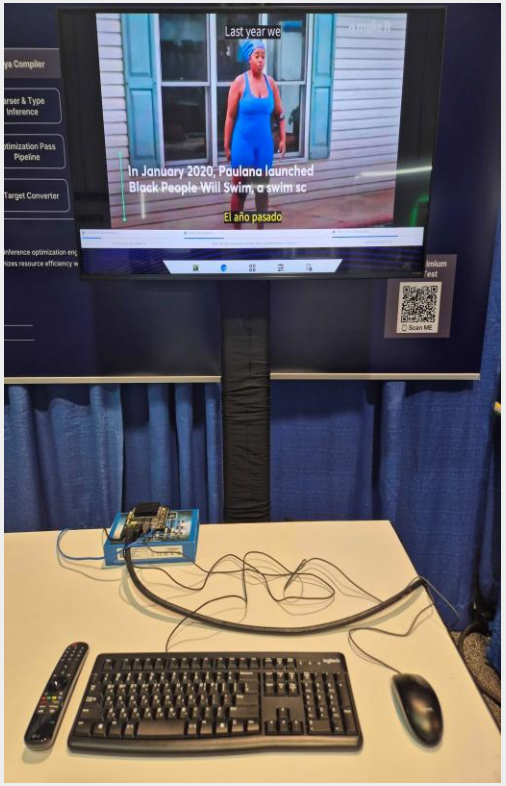


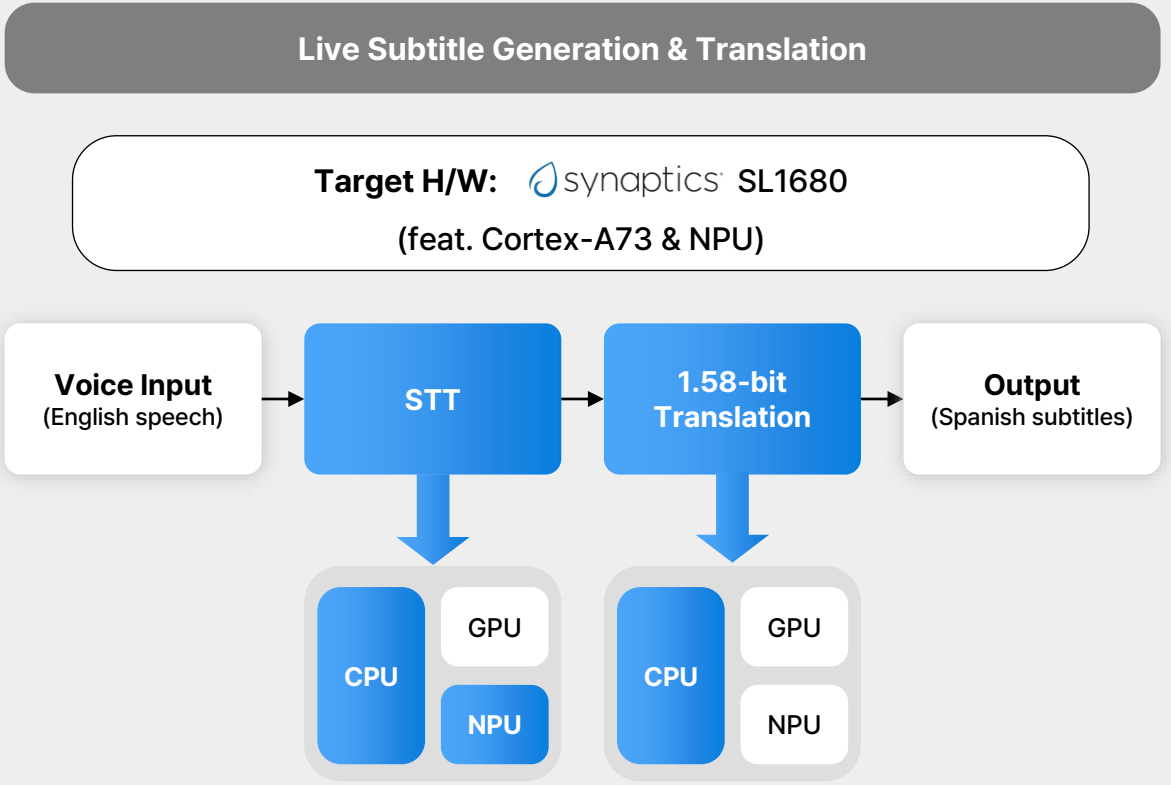


Note
1) Demo Video: https://www.youtube.com/watch?v=yCqNNh4PC4k

On-device AI Model (4/4) – Key Demos (Cont'd)

Live subtitle generation & translation pipeline

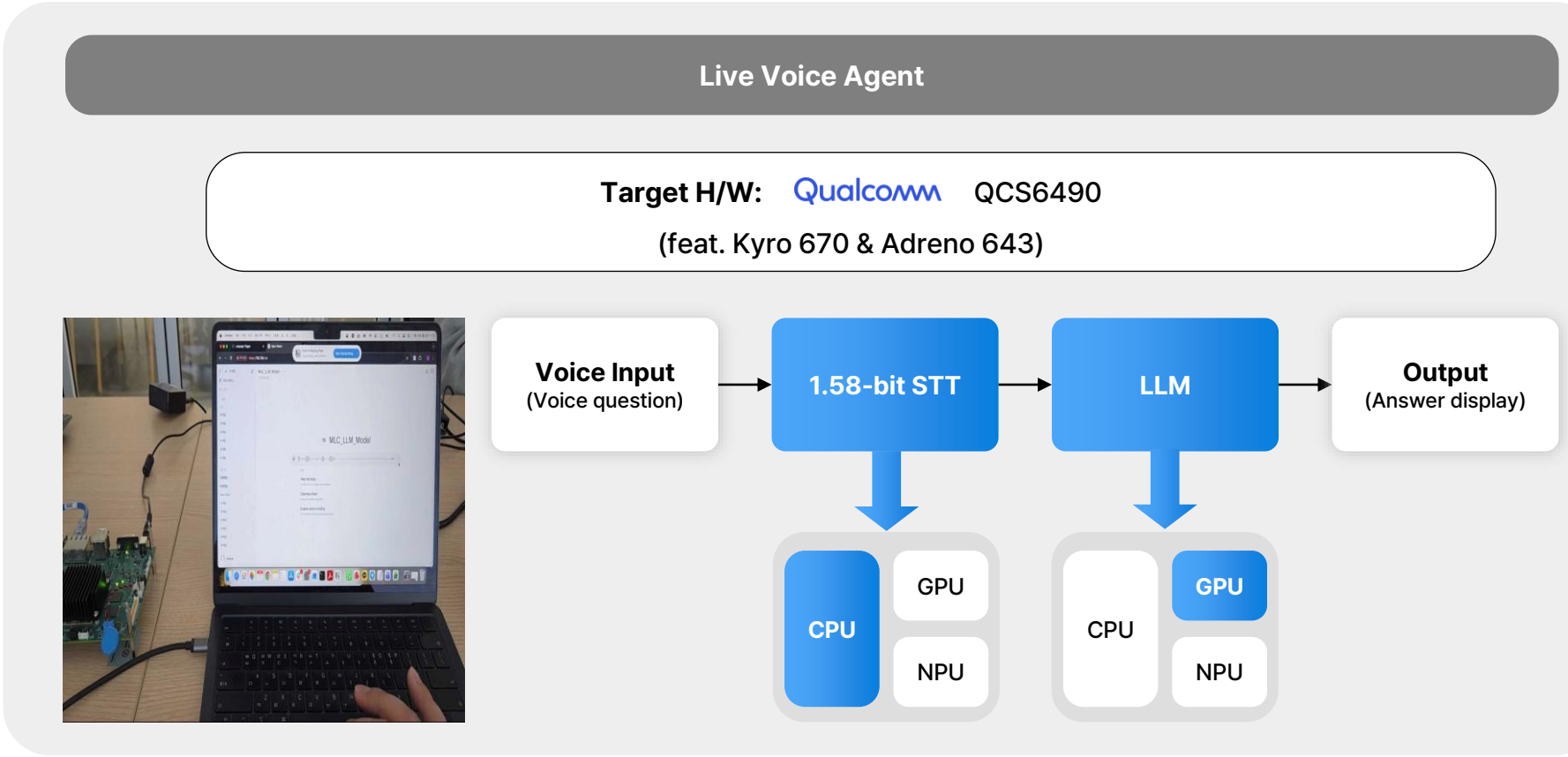




Note
1) Demo Video: https://www.youtube.com/watch?v=0u327q5aTMY

On-device AI Model (4/4) – Key Demos (Cont'd)

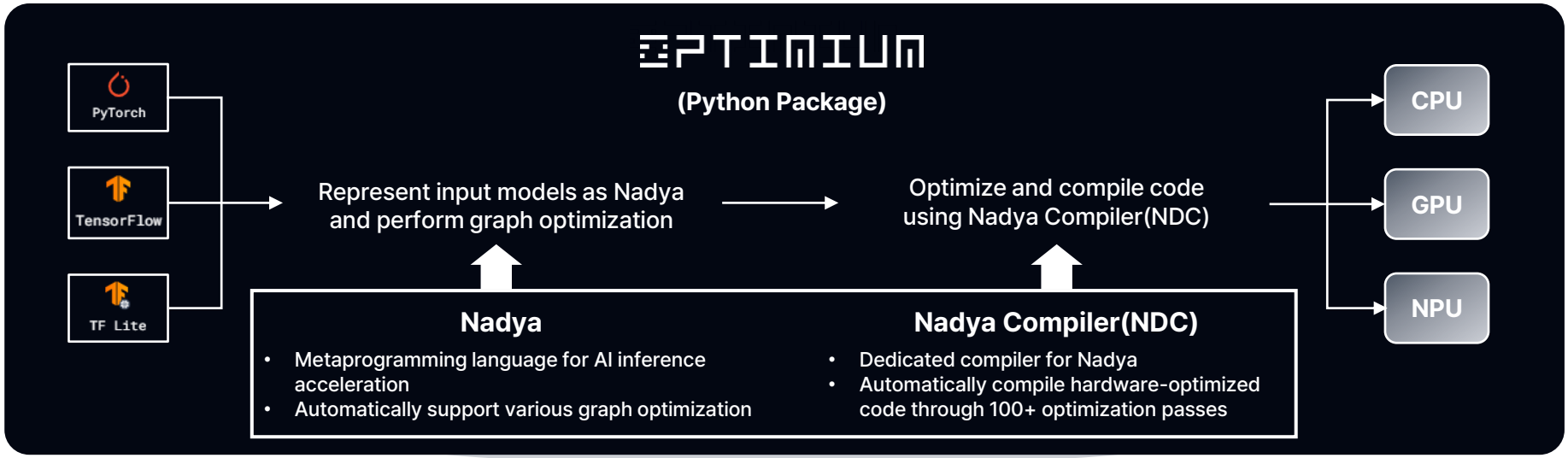
Live voice agent pipeline

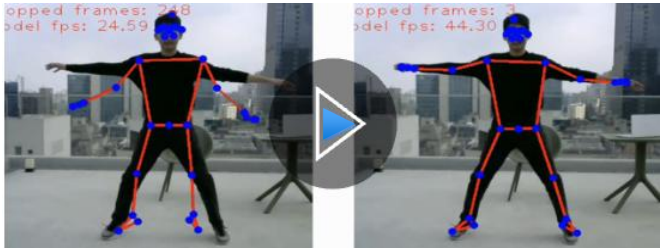


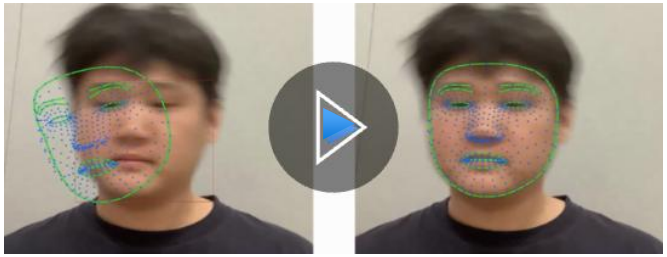
Note
1) Demo Video: https://www.youtube.com/watch?v=B0Ox5PNU5pQ

Optimium (1/3) – Overview

AI inference optimization engine based on our proprietary language

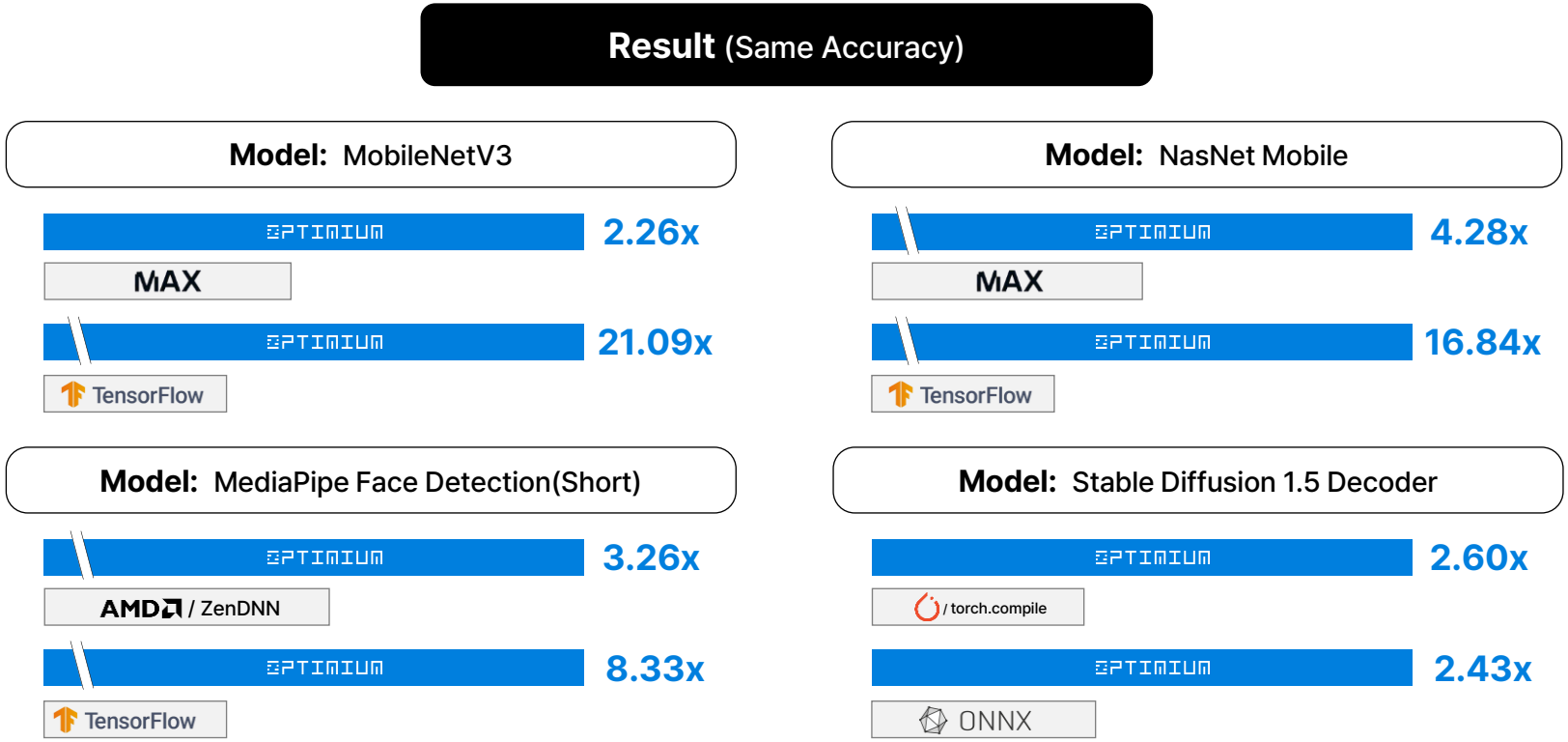






Optimium (2/3) - Performance

Demonstrate superior performance over existing tools on various models

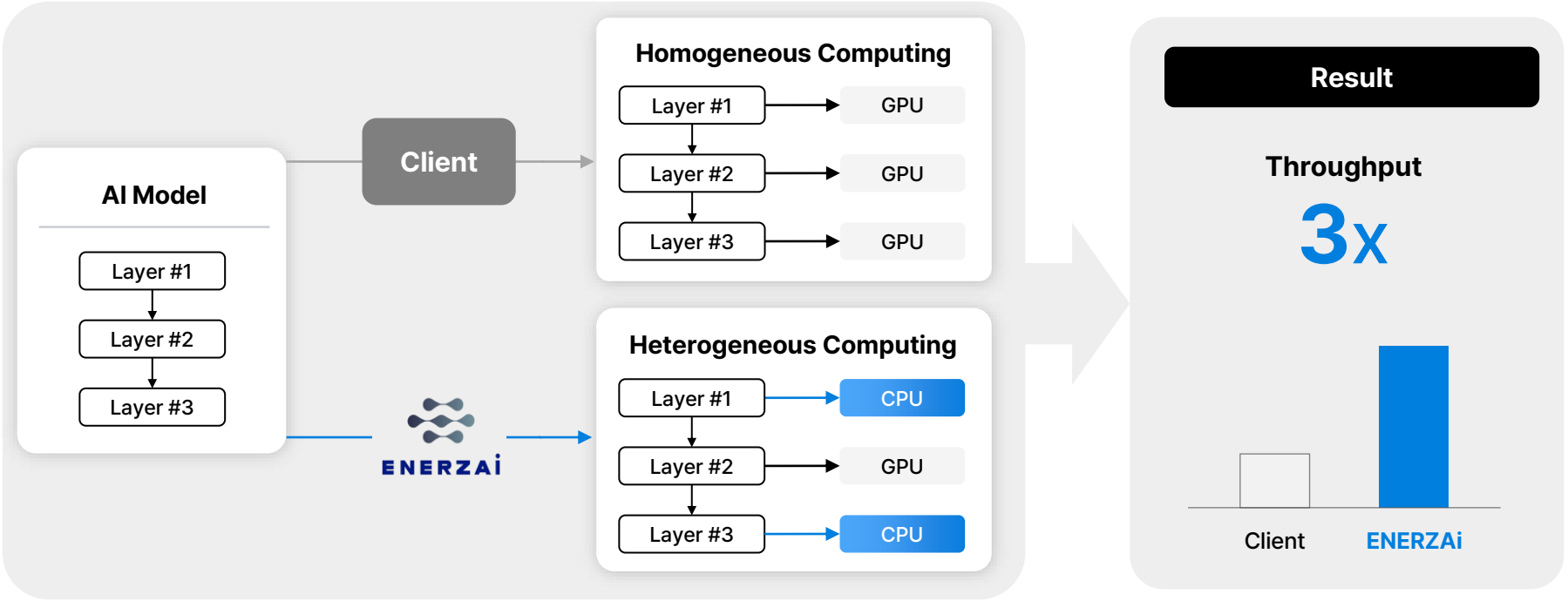


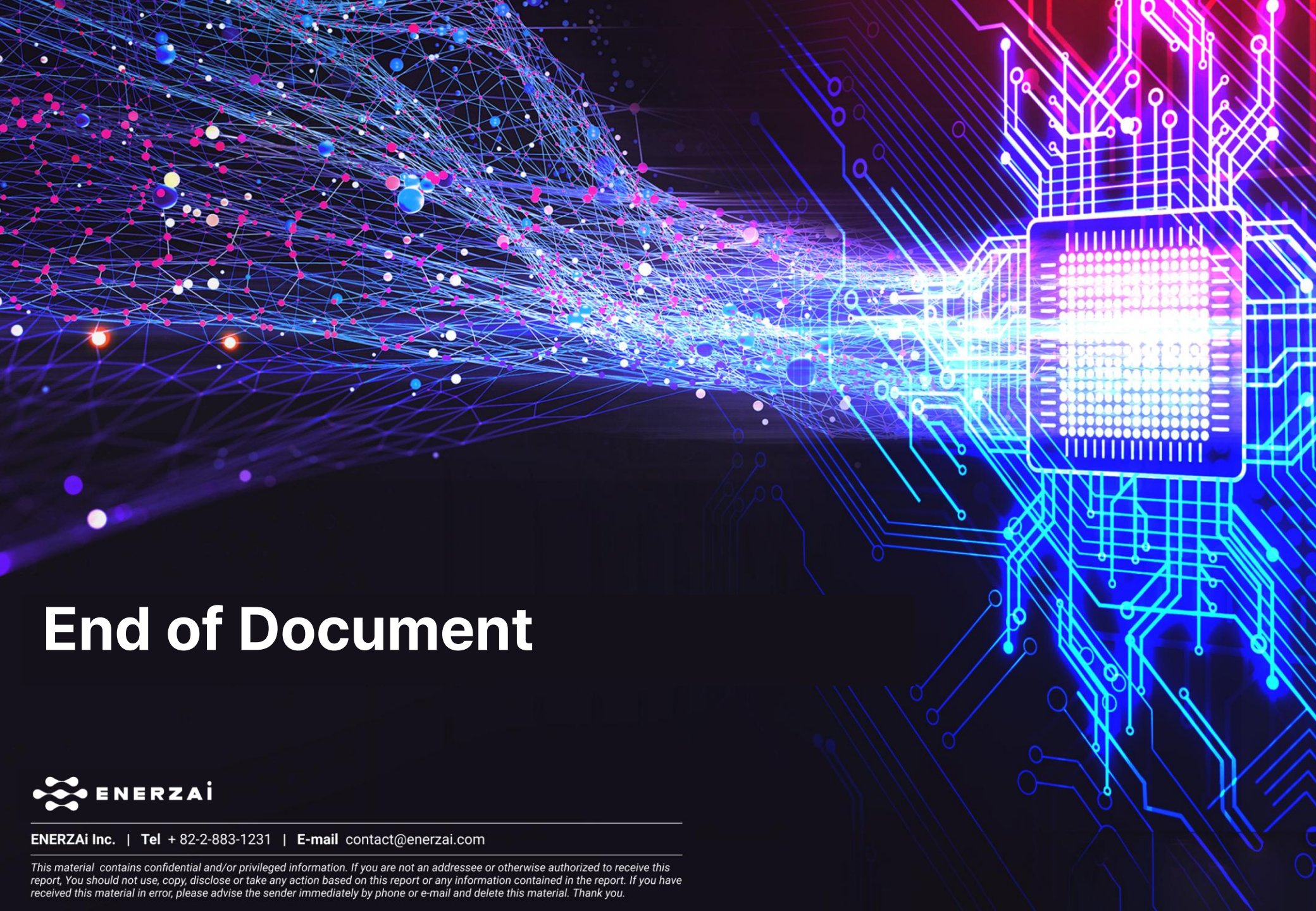
Optimium (3/3) - Reference

Maximize AI inference performance by optimally utilizing Qualcomm SoC

Electronics company

- Experienced issues with AI inference performance on a new device equipped with a Qualcomm SoC
- Our team efficiently utilized GPU & CPU in the SoC through heterogeneous computing, improving throughput by 3x – without any hardware modifications





End of Document

ENERZAi

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

This material contains confidential and/or privileged information. If you are not an addressee or otherwise authorized to receive this report, You should not use, copy, disclose or take any action based on this report or any information contained in the report. If you have received this material in error, please advise the sender immediately by phone or e-mail and delete this material. Thank you.