Scaling GPT:
The Future is on the Edge

WWW.KNERON.COM

Our Mission

Revolutionize AI with state-of-the-art, energy-efficient chips designed for the edge

Democratize AI by delivering powerful, secure, and decentralized solutions to every device across all industries

Introducing Kneron

Founded: 2015
Headquarters: San Diego, California, USA
Industry: AI, Semiconductors, Edge Computing

WHAT WE DO

Develop unique reconfigurable NPUs and high performance ISPs for edge AI application to democratize AI across industries

EMPLOYEES

- 210+ employees
- 20+ PHD experts
- >80% R&D department

APPLICATIONS

- Automobile
- Edge Server
- Security
- AIoT

FUNDING

- ~$200 million funding from Global VCs and strategic investors

Management Team

Albert Liu

Founder/CEO

30+ International Patents
70+ Papers in Major Journals

Frank Chang

Co-Founder

UCLA EE Wintek Chair
Former President of NCTU

Jimmy Lai

CFO

VP of SMIC, CFO of Dago New Energy & China Online Education

Hsiang Tsun Li

Chief Scientist

AVP of Qualcomm, Spreadtrum, and Huawei

Roger Liu

COO

COO/CTO of MilkyWay Silicon Technology

Kneron

NVIDIA

QUALCOMM

SAMSUNG

vivo

energy consumption

DATA CENTER VS. A SMALL NATION

=

Large data center equivalent to the power consumption of Singapore

energy consumption

COST OF AI ADOPTION

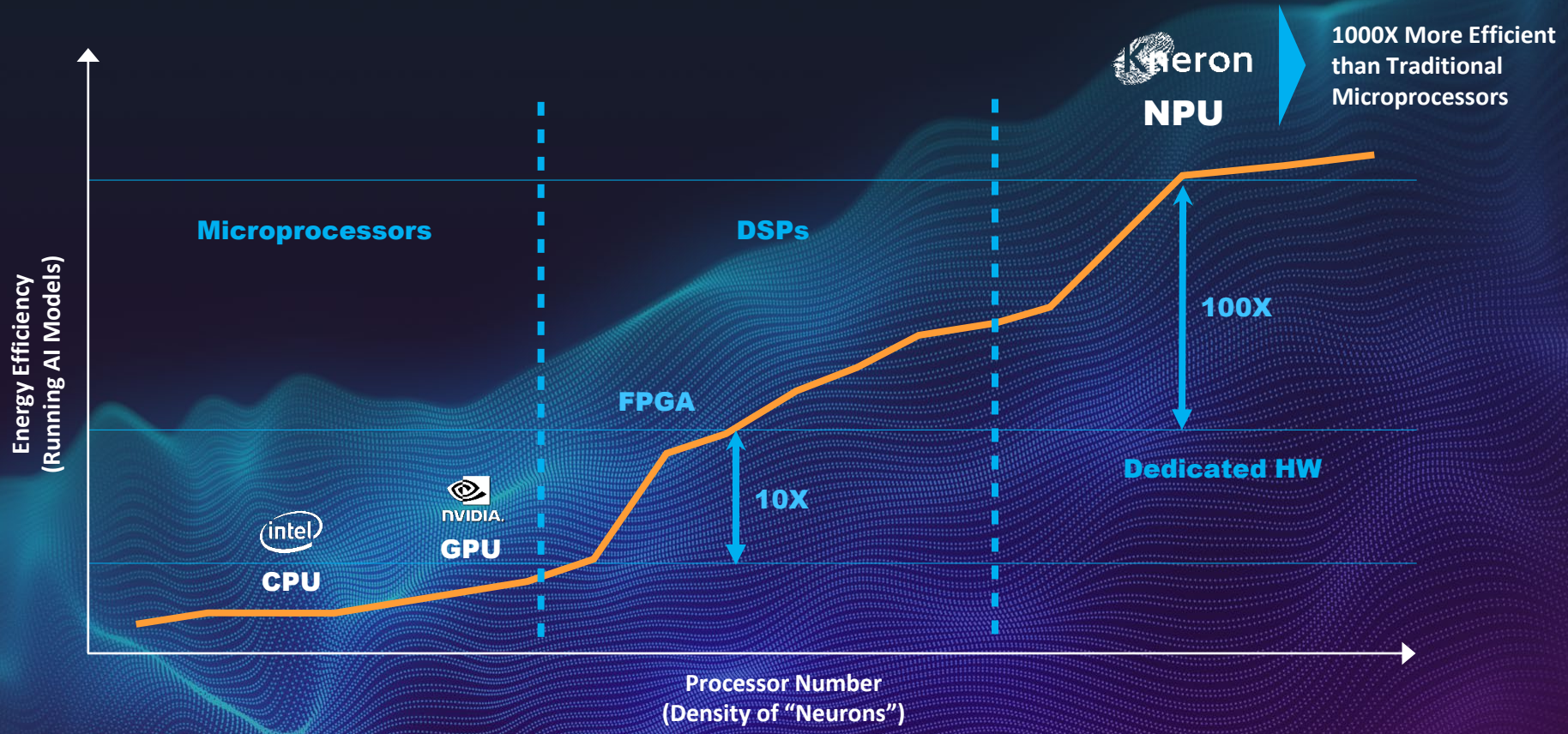
To serve 100 million GPT users

$5 billion

Training Google Switch Transformer

179 MW **59 Tons**

Kneron NPU Solves Energy Efficiency Issue



Source: ISSCC Proceedings

The Problem with Cloud AI Computing

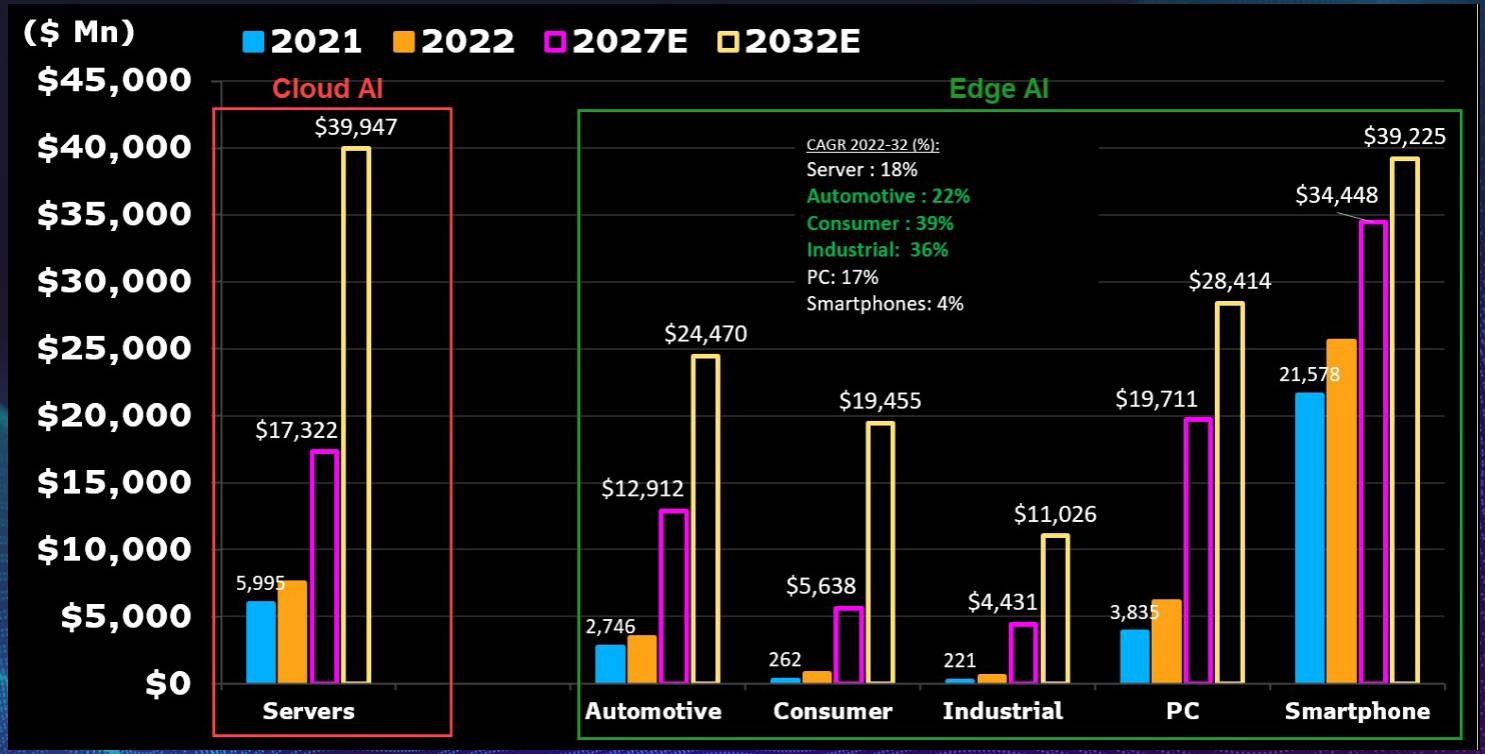
- Unstainable energy demand
- Recurring latency issues
- Ongoing security concerns given centralized processing and storage

- Efficient energy utilization
- Low to no latency issues
- Enhanced security given localized processing and storage

Computation for many future AI applications must be performed on the edge

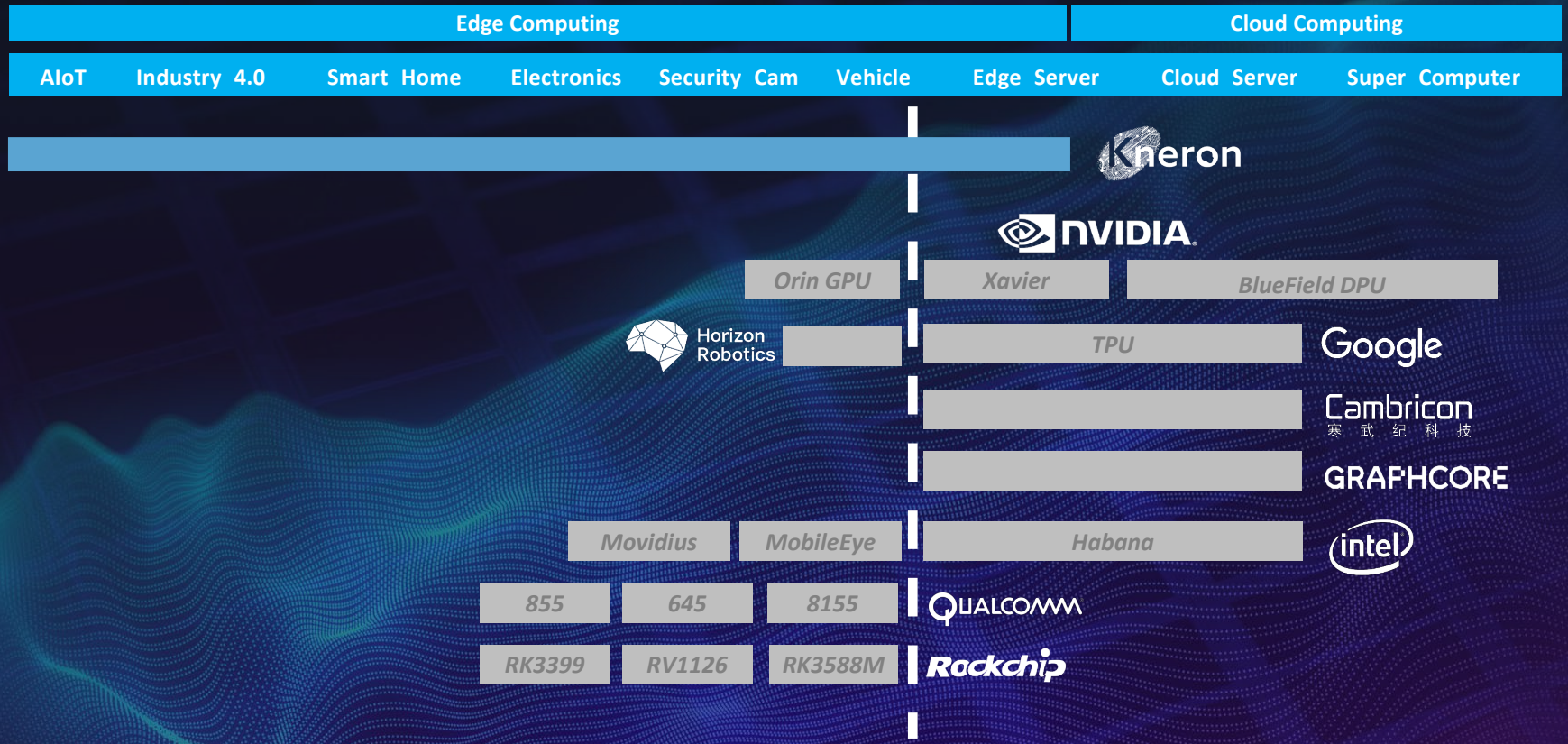
Edge AI Semiconductor Will Dominate the Market

"Edge AI semiconductor market could be as much as 3.37X the size of the cloud-based AI market by end of 2032" **Bloomberg Intelligence**



Source: Bloomberg Intelligence, Gartner

Kneron Goal is to Provide Solutions Across All Edge Use Cases



Source: Public company product information and Kneron analysis

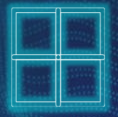
Kneron's Strategy Enabled by Patented Reconfigurable NPU

Analogous Layers

Pooling Units

Rectified Linear Units

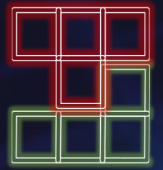
Convolution Units



Reconfigured via Kneron's Software Programming

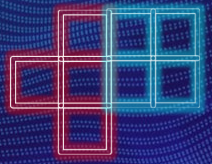
Illustrative CNN Models

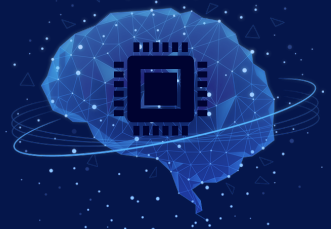
Resnet



YOLO

MobileNet





- Highly flexible architecture allows reconfiguration to run different AI models based on customers' needs
- Efficient, close-fitting operations greatly reduce power consumption – 1000X more efficient than traditional microprocessors (including GPUs)

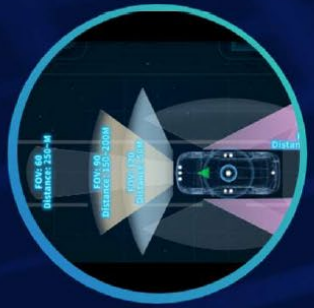
Kneron Has Already Secured Wide-Range of Customers

AIOT

GREE
PHILIPS
DESMAN 德施曼 · 智能锁
GARMIN.
QUALCOMM
Panasonic
ADVANTECH
DELTA 台达
msi
AAEON an ASUS company
合盈光電 H.P.B. Optoelectronics
OFILM 欧菲光
BARCO
PRIMAX

SECURITY

Hanwha
SKYBELL
VIVOTEK
TUTK
LITEON
spark
3S
MOXA
kbro 凯擘 大寬頻
ATN



AUTO-MOTIVES

TUS
REC
oToBrite
MIH
JVC KENWOOD
FUJITSU
Yupiteru
Polaroid
CHIMEI MOTOR
PX
LAN
wistron

EDGE SERVERS

中華電信 Chunghwa Telecom
Quanta
EDIMAX NETWORKING PEOPLE TOGETHER
AAEON an ASUS company
ENTERTAINMENT
spark

LLM&GPT STATION

Quanta
FOXCONN
COMPAL
Tongtai
Unimicron 欣興電子
THE UNIVERSITY OF CALIFORNIA UCLA
LELAND STANFORD JUNIOR UNIVERSITY 1891

Mobilizing Communities to Create with AI

Core AI curriculum textbook used at many leading universities (e.g. Princeton, UCLA, UCSD, UVA, NTU, NTHU, etc)

The Future of Semi?

CPU

GPU

NPU

GPT ai servers

ENERGY SAVING

DISTRIBUTED DATA

OFFLINE DEPLOYMENT

SERVICE RELIABILITY

KNEO 300

- 330Tops AI computing power
- Edge GPT Inference

KNEO 330

- 548Tops AI computing power
- Edge GPT Inference

KNEO 1000

- 240Tops AI computing power
- Cloud GPT inference

2023

2024

2025

AI PC applications

Video Conferencing

- Image enhancement (HDR and low light)
- Background segmentation and blurring
- Eye gaze and mouth detection
- Meeting summary

Human-Machine Interaction

- Natural language control
- Gesture control

RAG (Retrieval-augmented generation)

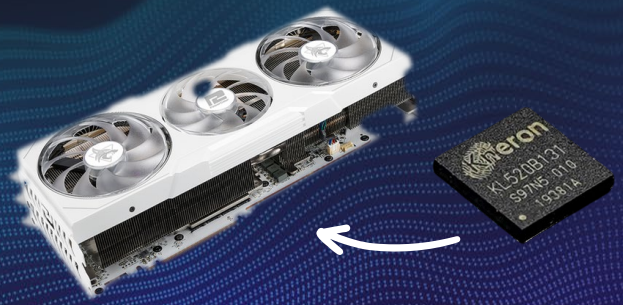
- Local knowledge base

Virtual Assistant

- Local AI Q&A, translation, inferencing, summary
- Virtual Avatar

GPU + NPU integration

OPTIMIZED AI & PROCESSING



10-25% FASTER PERFORMANCE

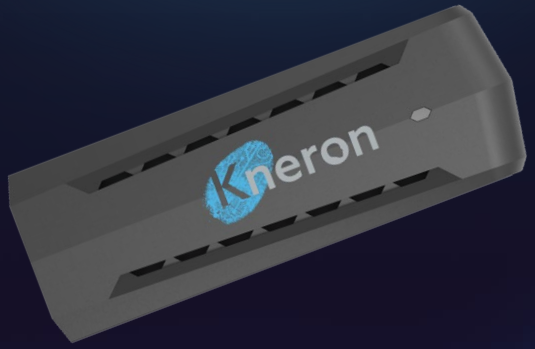
Optimize AI workloads for superior speed.

30% POWER SAVINGS

Reduce energy consumption to extend product life.

AI MODEL TUNING

Optimize computational efficiency and reduce power load.



adaptable ai solutions

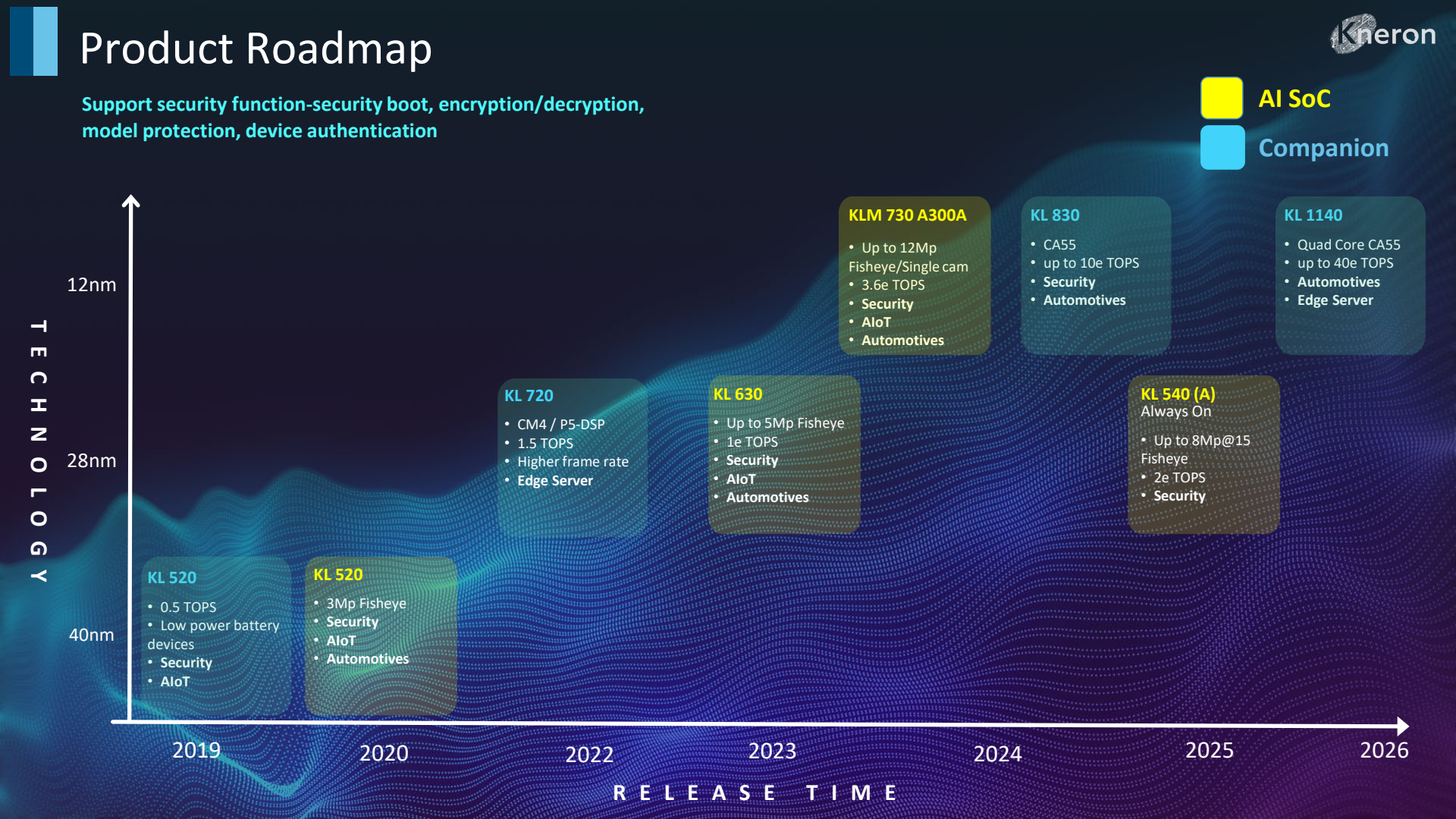
EDGE GPT DONGLE

- KL830 Inside
- 10eTOPS@8bit
- Tiny-Llama
- Plug&Play
- USB interface

for any device with diverse product options

Product Roadmap

Support security function-security boot, encryption/decryption, model protection, device authentication



thank you

www.kneron.com