Kneron KL-530 introduction
-How we define the next generation of Edge AI Chip

David Yang
Senior Director

ABOUT KNERON

GLOBAL R&D SITES

210+ No. of employees

75% R&D

100M USD funding

EDGE AI CHIP AND ALGORTHM PROVIDER

Albert
Founder / CEO

30+ International Patents
70+ Papers in Major Journals

SAMSUNG

QUALCOMM®

Frank
Co-Founder

Former president of NCTU
UCLA EE Wintek Chair

UCLA

國立交通大學
National Chiao Tung University

OUR INVESTORS

100 Million Raised

Horizons Ventures 维港投资

QUALCOMM VENTURES

SEQUOIA

中華開發資本國際 CDIB CAPITAL INTERNATIONAL

Alibaba Entrepreneurs Fund 阿里巴巴創業者基金

FOXCONN

winbond

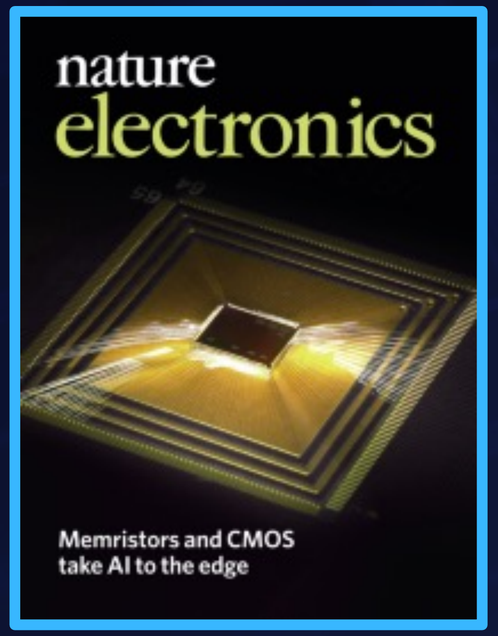
DELTA

Himax

MXIC 旺宏電子

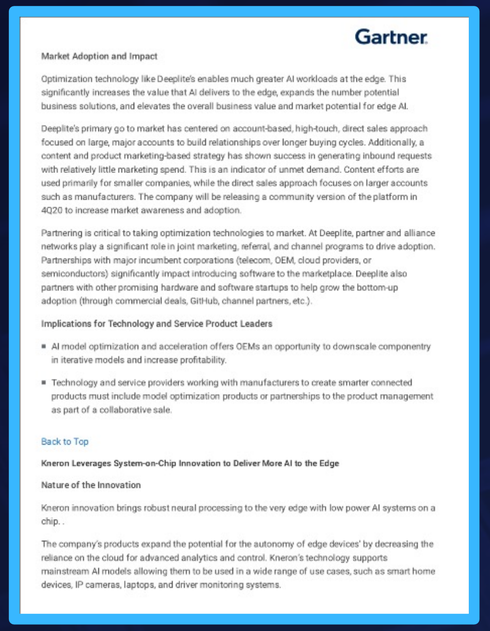
RECOGNITION AND HONORS

nature



Nature Electronics
Volume 2 Issue 9

Gartner



Gartner.

Market Adoption and Impact

Optimization technology like Deeplite's enables much greater AI workloads at the edge. This significantly increases the value that AI delivers to the edge, expands the number potential business solutions, and elevates the overall business value and market potential for edge AI.

Deeplite's primary go to market has centered on account-based, high-touch, direct sales approach focused on large, major accounts to build relationships over longer buying cycles. Additionally, a content and product marketing-based strategy has shown success in generating inbound requests with relatively little marketing spend. This is an indicator of unmet demand. Content efforts are used primarily for smaller companies, while the direct sales approach focuses on larger accounts such as manufacturers. The company will be releasing a community version of the platform in 4Q20 to increase market awareness and adoption.

Partnering is critical to taking optimization technologies to market. At Deeplite, partner and alliance networks play a significant role in joint marketing, referral, and channel programs to drive adoption. Partnerships with major incumbent corporations (telecom, OEM, cloud providers, or semiconductors) significantly impact introducing software to the marketplace. Deeplite also partners with other promising hardware and software startups to help grow the bottom-up adoption (through commercial deals, GitHub, channel partners, etc.).

Implications for Technology and Service Product Leaders

- AI model optimization and acceleration offers OEMs an opportunity to downscale componentry in iterative models and increase profitability.
- Technology and service providers working with manufacturers to create smarter connected products must include model optimization products or partnerships to the product management as part of a collaborative sale.

Back to Top

Kneron Leverages System-on-Chip Innovation to Deliver More AI to the Edge

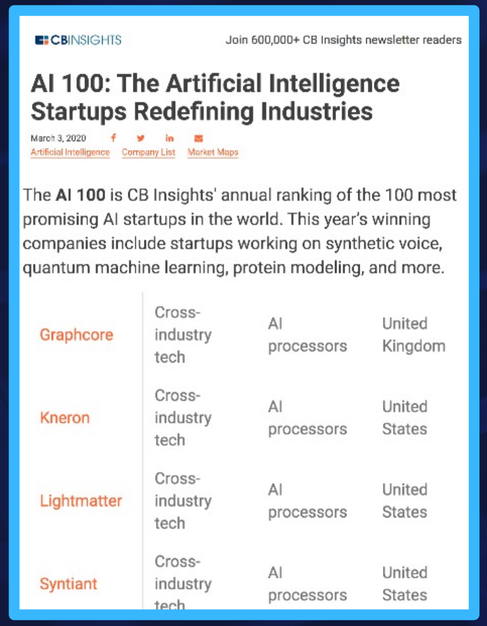
Nature of the Innovation

Kneron innovation brings robust neural processing to the very edge with low power AI systems on a chip. .

The company's products expand the potential for the autonomy of edge devices' by decreasing the reliance on the cloud for advanced analytics and control. Kneron's technology supports mainstream AI models allowing them to be used in a wide range of use cases, such as smart home devices, IP cameras, laptops, and driver monitoring systems.

Gartner
Emerging Technologies:
Tech Innovators in Edge AI

CBINSIGHTS



CBINSIGHTS Join 600,000+ CB Insights newsletter readers

AI 100: The Artificial Intelligence Startups Redefining Industries

March 3, 2020
Artificial Intelligence | Company List | Market Maps

The **AI 100** is CB Insights' annual ranking of the 100 most promising AI startups in the world. This year's winning companies include startups working on synthetic voice, quantum machine learning, protein modeling, and more.

Graphcore	Cross-industry tech	AI processors	United Kingdom
Kneron	Cross-industry tech	AI processors	United States
Lightmatter	Cross-industry tech	AI processors	United States
Syntiant	Cross-industry tech	AI processors	United States

CB Insights
AI 100

IEEE

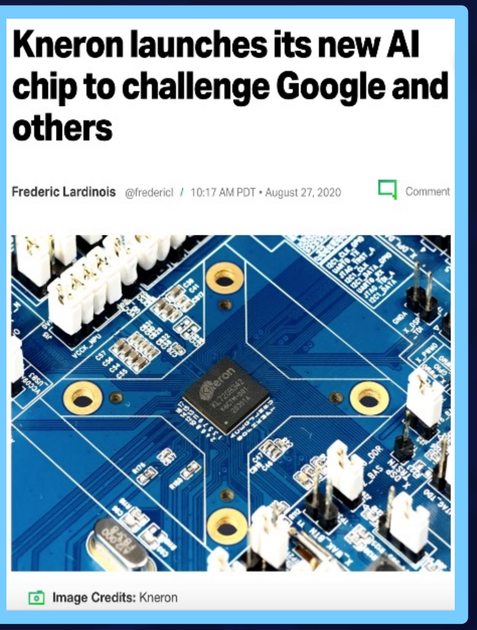
Darlington Award Recipients

Darlington Award Recipients

Year Awarded	Awardee(s)	Paper Title	Publication Title
2021	Li Du, Yuan Du, Yilei Li, Junjie Su, Yen-Cheng Kuan, Chun-Chen Liu, and MauChung Frank Chang	A Reconfigurable Streaming Deep Convolutional Neural Network Accelerator for Internet of Things	IEEE Transactions on Circuits and Systems I: Regular Papers, Volume 65, Issue 1, Jan. 2018
2020	Francesco Conti, Robert Schilling, Pasquale Davide Schiavone, Antoni Pullini, Davide Rossi, Frank Kağan Gürkaynak, Michael Muehlberghuber, Michael Gautschi, Igor Loi, Germain Haugou, Stefan Mangard, Luca Benini	An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics	IEEE Transactions on Circuits and Systems I: Regular Papers, Volume: 64, Issue: 9, Sept. 2017 Page(s): 2481 - 2494
2019	Not Awarded		
2018	Jia Zhou; Tong Ge, Joseph S. Chang	"Printed Electronics: Effects of Bending and a Self-Compensation Means"	IEEE Transactions on Circuits and Systems I: Regular Papers, Volume: 64, Issue: 3, March 2017 Page(s): 505 - 515
2017	Bo Zhao, Nai-Chung Kuo, and Ali M. Niknejad	An Inductive-Coupling Blocker Rejection Technique for Miniature RFID Tag	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 63, no. 8, pp. 1305-1315, August 2016
2016	Dong Yang; Andrews, C.; Molnar, A	Optimized Design of N-Phase Passive Mixer-First Receivers in Wideband Operation	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 11, pp. 2759–2770, November 2015
2015	Chengwu Tao, Ayman Fayed	A Low-Noise PFM-Controlled Buck Converter for Low-Power Applications	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 59, no. 12, pp. 3071-3080, December 2012
2014	Katharina Hausmair, Shuli Chi, Peter Singerl, and Christian Vogel	Aliasing-free Digital Pulse-width Modulation for Burst-mode RF Transmitters	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 2, pp. 415-427, February 2013
2012	Alan N. Willson, Jr.	Desensitized Half-Band Filters	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 57, no. 1, pp. 152-165, January 2010
2011	Zhen Cao, Brian Foo, Lei He, Michaela van der Schaar-Mitrea	Optimality and Improvement of Dynamic Voltage Scaling Algorithms for Multimedia Applications	IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 57, no. 3, pp. 681-690, March 2010
2010	Sudip Shekhar, Mozhgan Mansuri, Frank O'Mahony, Ganesh Balamurugan, James E.	Strong Injection Locking in Low-Q LC Oscillators: Modeling and	IEEE Transactions on Circuits and Systems I: Regular Papers,

IEEE
Darlington Award
2021

TechCrunch



Kneron launches its new AI chip to challenge Google and others

Frederic Lardinois @fredericl / 10:17 AM PDT • August 27, 2020 Comment

Image Credits: Kneron

TechCrunch
Kneron Challenges
Google and Intel

LEADING AI CHIP STRUCTURE DESIGN BRINGS COST ADVANTAGES

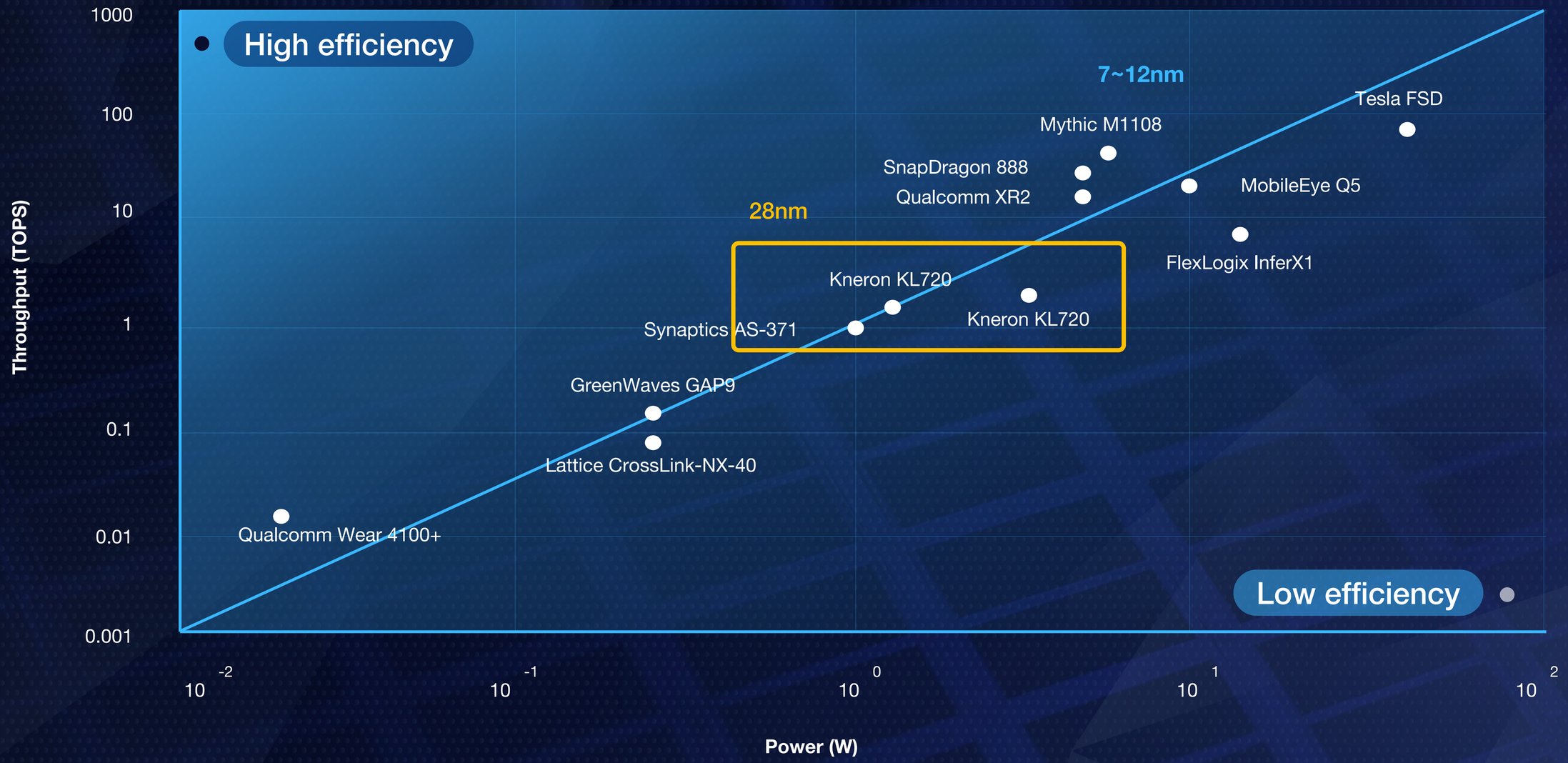


Figure 9: Throughput comparison of different commercial edge processors for NN inference at the edge.

Published by UC Berckly, A Survey of Quantization Methods for Efficient Neural Network Inference, Mar 2021

KNERON LEADING THE EDGE AI INDUSTRY

2018

50nm
0.1 Tops

Smart home

2019

40nm
0.3Tops

AIoT

2020

28nm
1.5Tops

ADAS/Edge server

GREE 格力

TUTK

ADVANTECH

OTUS IMAGING

QCT™

TOYOTA

NEXT GENERATION CHIP

Intelligent ISP

-AI based tuning
-Integrated AI AEC/AWB control

RISC-V CPU

-open source
-pre/post processing for NPU

Optimized Tool Chains

-Auto graph partition
-Multiple AI companion

Ultra low power

-500mW average

KL530

Security mechanism

-Efficient encryption on models and data

Reconfigurable NPU

-CNN/RNN/LSTM/Transformer
-searches optimal bit width

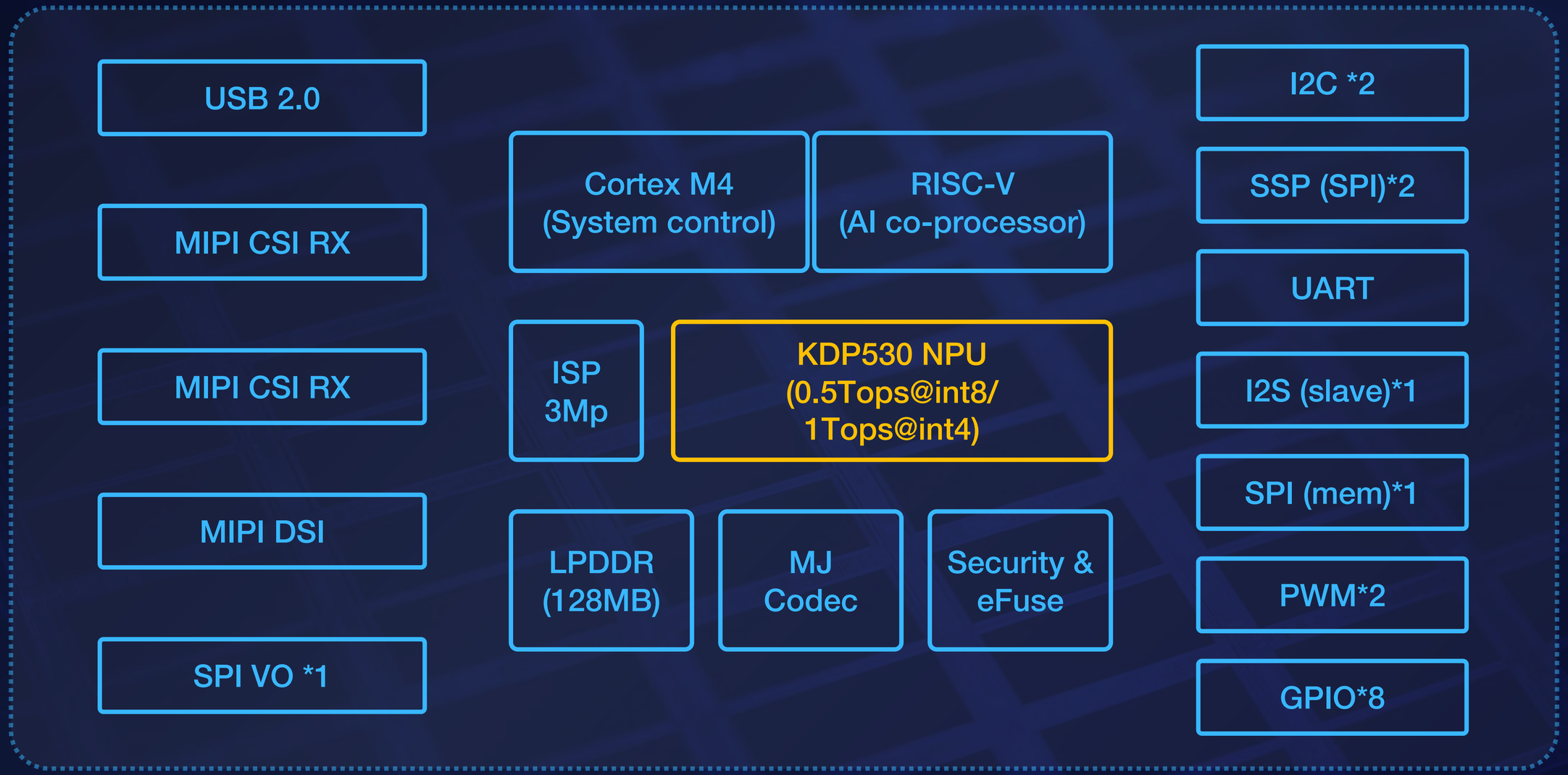
Support Int 4

- Higher computing efficiency
- Mixed precision Int8/float

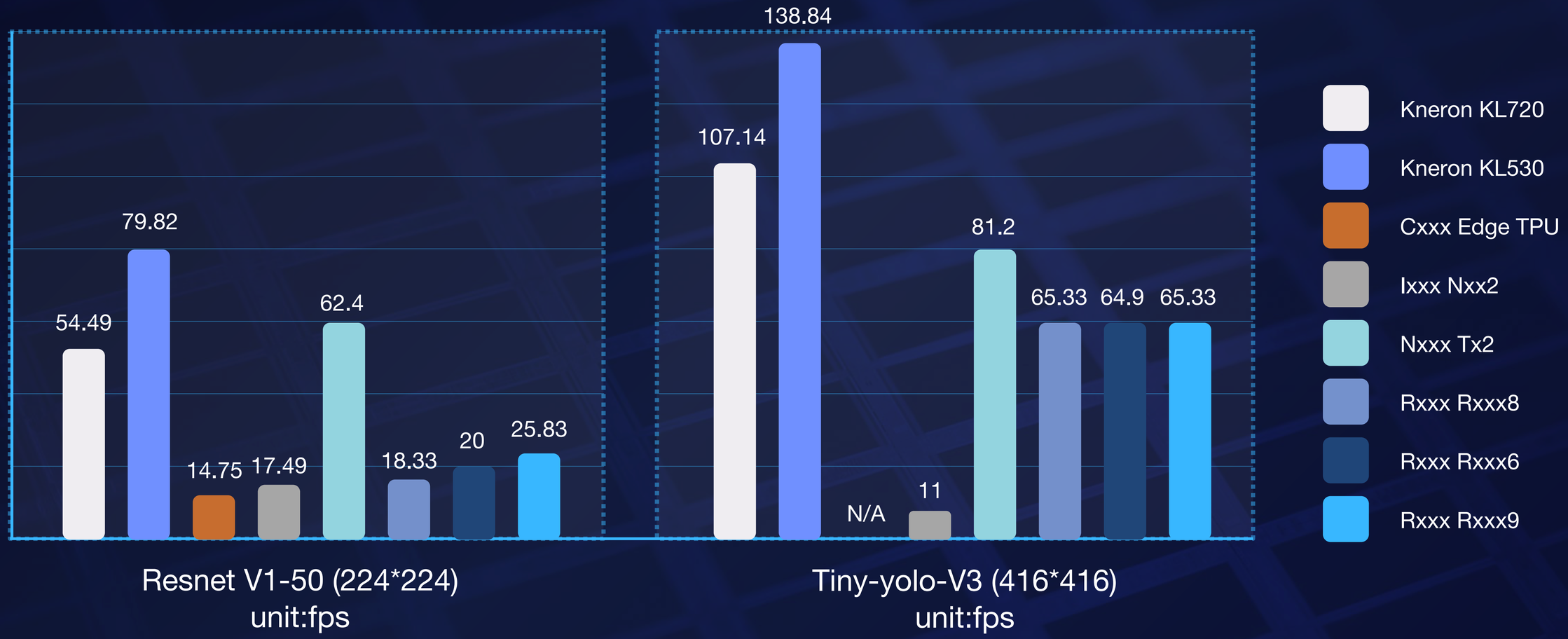
Transformer network

- Higher computing accuracy
- Has largest model capacity

KL530 DIAGRAM

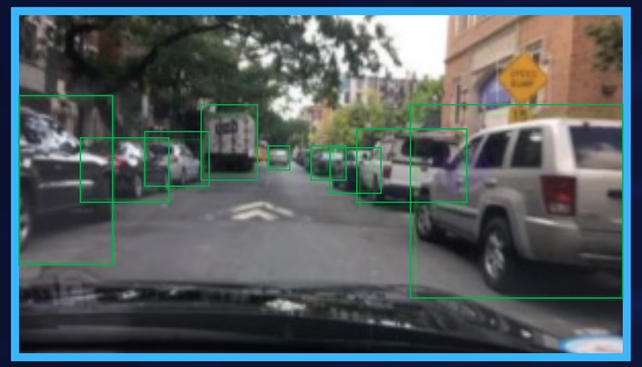


COMPUTING EFFICIENCY UP TO 10X HIGHER

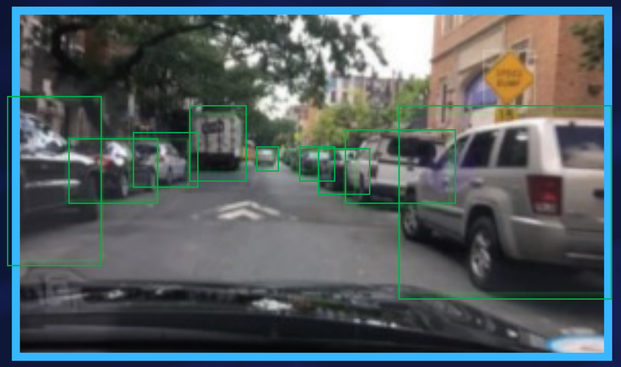


LEADING NPU PERFORMANCE BY INT 4

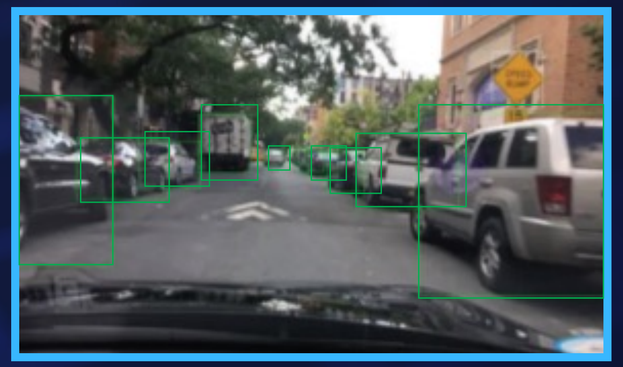
Latency save up to 70% accuracy remains the same



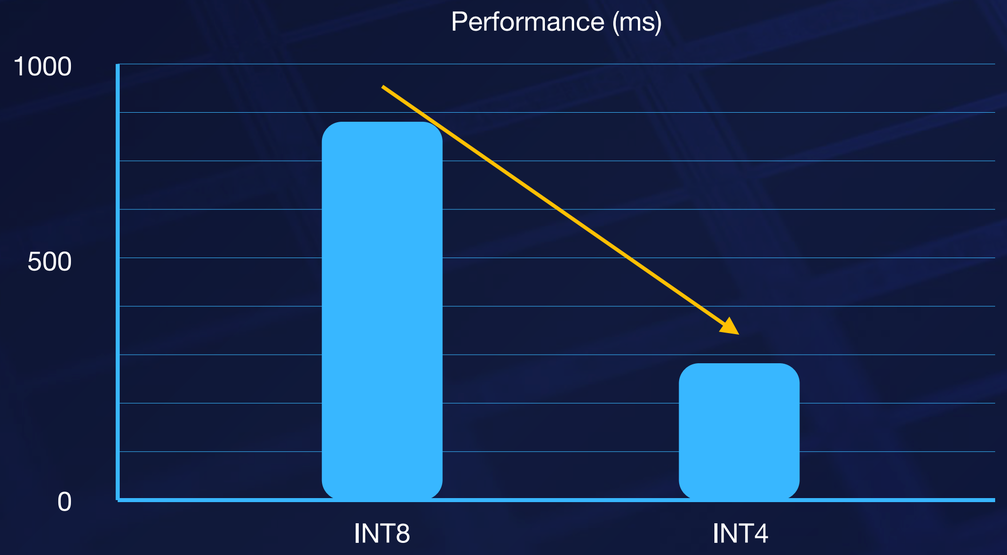
(a) GroundTruth



(b) 8-Bit



(c) 4-Bit



Model	Float	INT8	Dynamic INT4/8 (>90% INT4)
mobilenetv2_100	73	72.7	72
resnet50	79	79	78.9
mobilenetv2_120d	77.3	76.8	76.3
tf_efficientnet_lite4	81.3	81	81

HIGHER RECOGNITION ACCURACY WITH TRANSFORMER

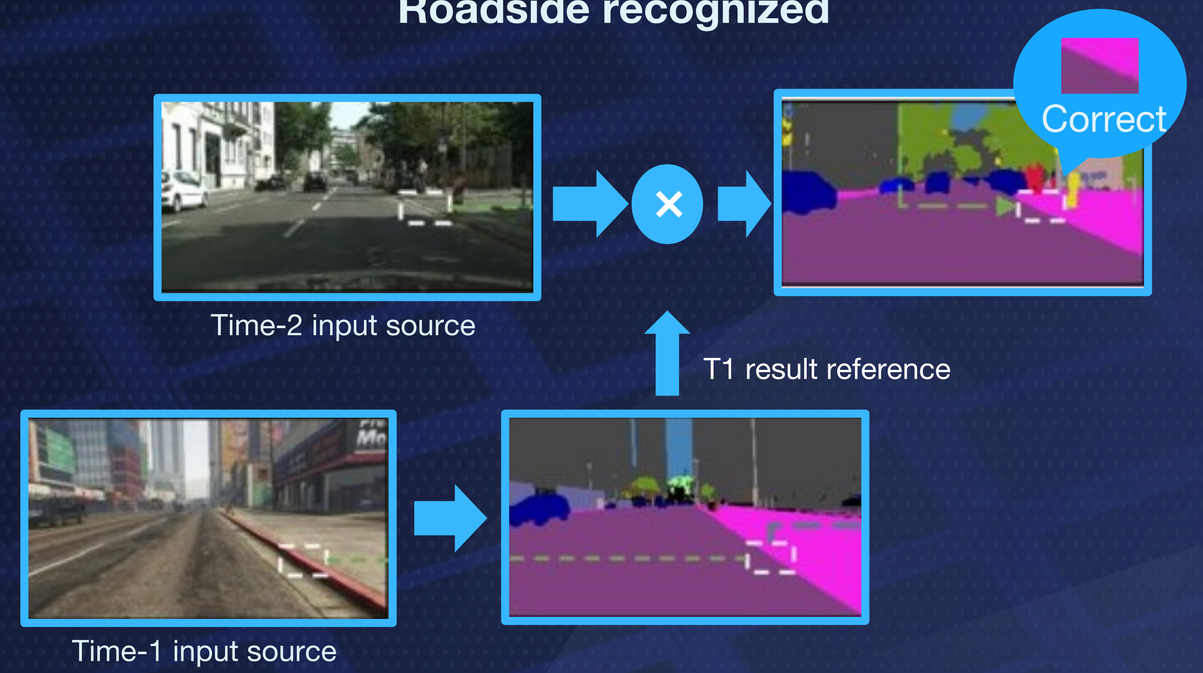
cross reference of semantic segmentation recognition results

CNN Only

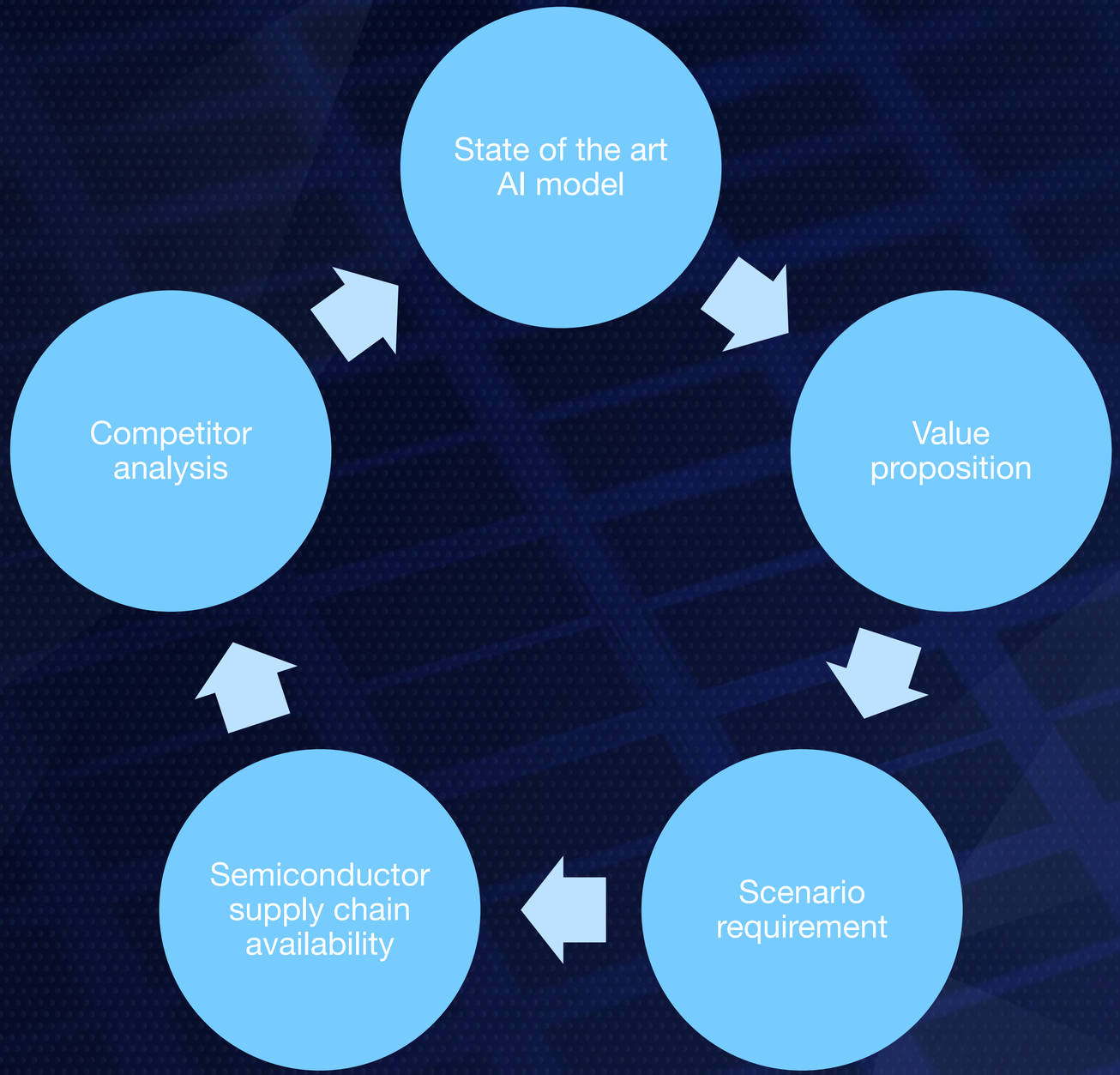
Roadside not recognized

Transformer & CNN Hybrid

Roadside recognized



SW DEFINE HW


State of the art AI model
Value proposition
Scenario requirement
Semiconductor supply chain availability
Competitor analysis

THANK YOU

www.kneron.com