Energy-efficient Semiconductors for AI in the age of Large Language Models

Dr. Sakyasingha Dasgupta

CEO & Founder
EdgeCortix

EDGECORTIX™

EDGECORTIX®

Our Mission

"To deliver cloud-level performance at the edge, with orders of magnitude better energy efficiency and processing speed, drastically reducing customer operating costs."

We are pioneering the future of the connected intelligent edge with a Software-driven Edge Artificial Intelligence (AI) Platform

Backed By

SBI

RENESAS
BIG IDEAS FOR EVERY SPACE

Strategic Partnership with Renesas in AI Acceleration

RENESAS + EDGECORTIX

Across Microcontrollers & Microprocessors

July 2022
Collaboration on DRP-AI TVM compiler announced.
DRP-AI Press Release

Sept 2022
Renesas announces RZ/V will use MERA in DRP-AI.
RZ/V Press Release

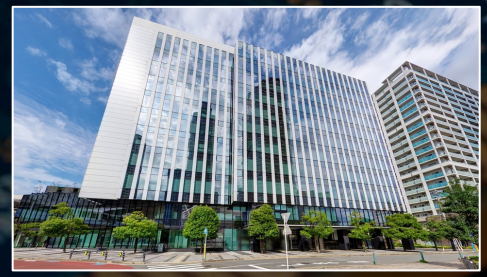
Oct 2023
Renesas invests in EdgeCortix as part of $20M in additional funding.
Funding PR Press Release

2023 & beyond ..
Collaboration to streamline AI/ML Development.
AI/ML Press Release

Since 2021

EdgeCortix is headquartered in Japan, with offices in USA, Singapore and India

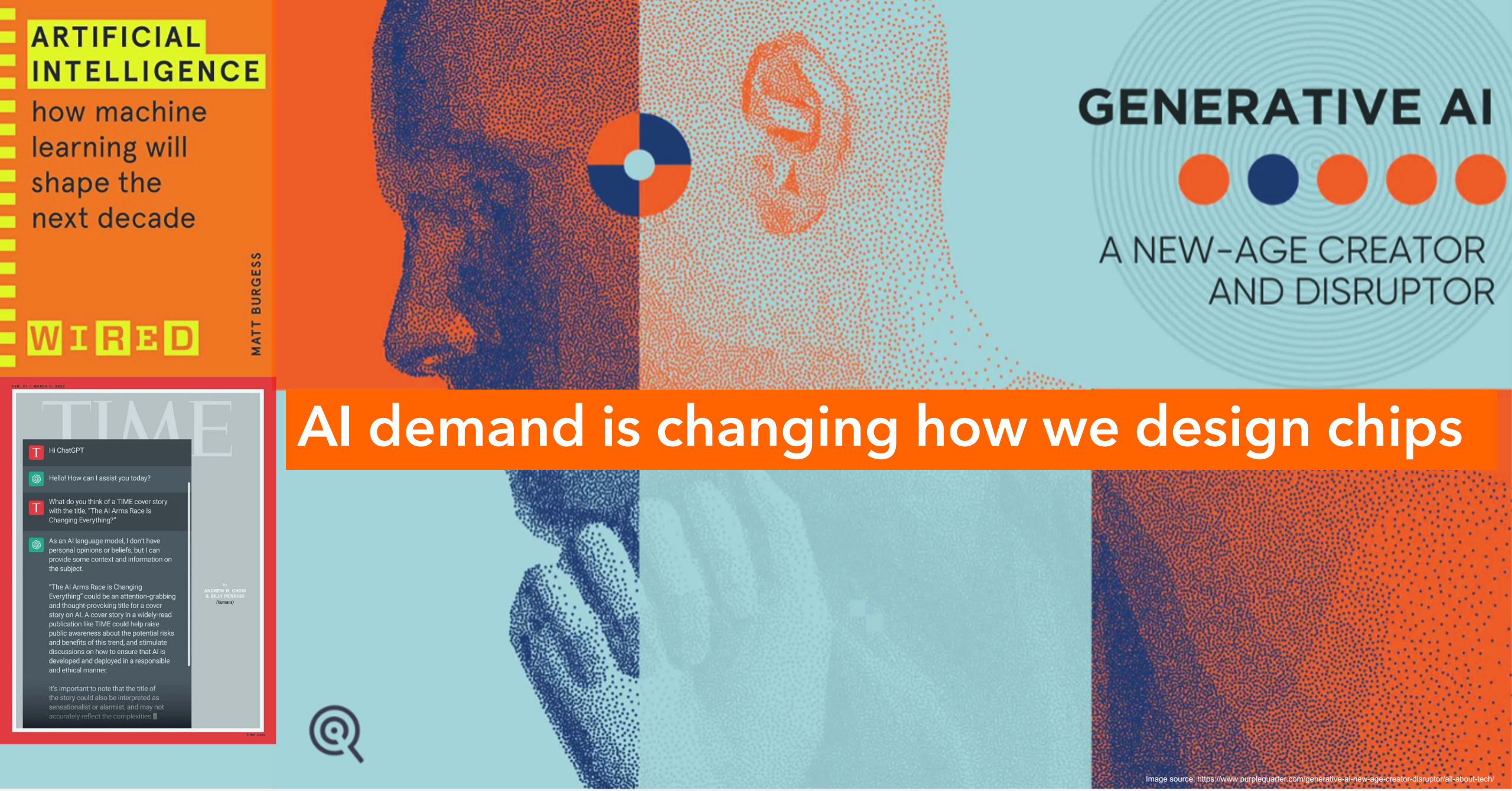
Founded July 2019



EdgeCortix Engineerring Center Mushashi-Kosugi, Kawasaki-Shi

Among Global 20* (Pitchbook)

* Pitchbook – Leading Edge Computing Semiconductor Companies

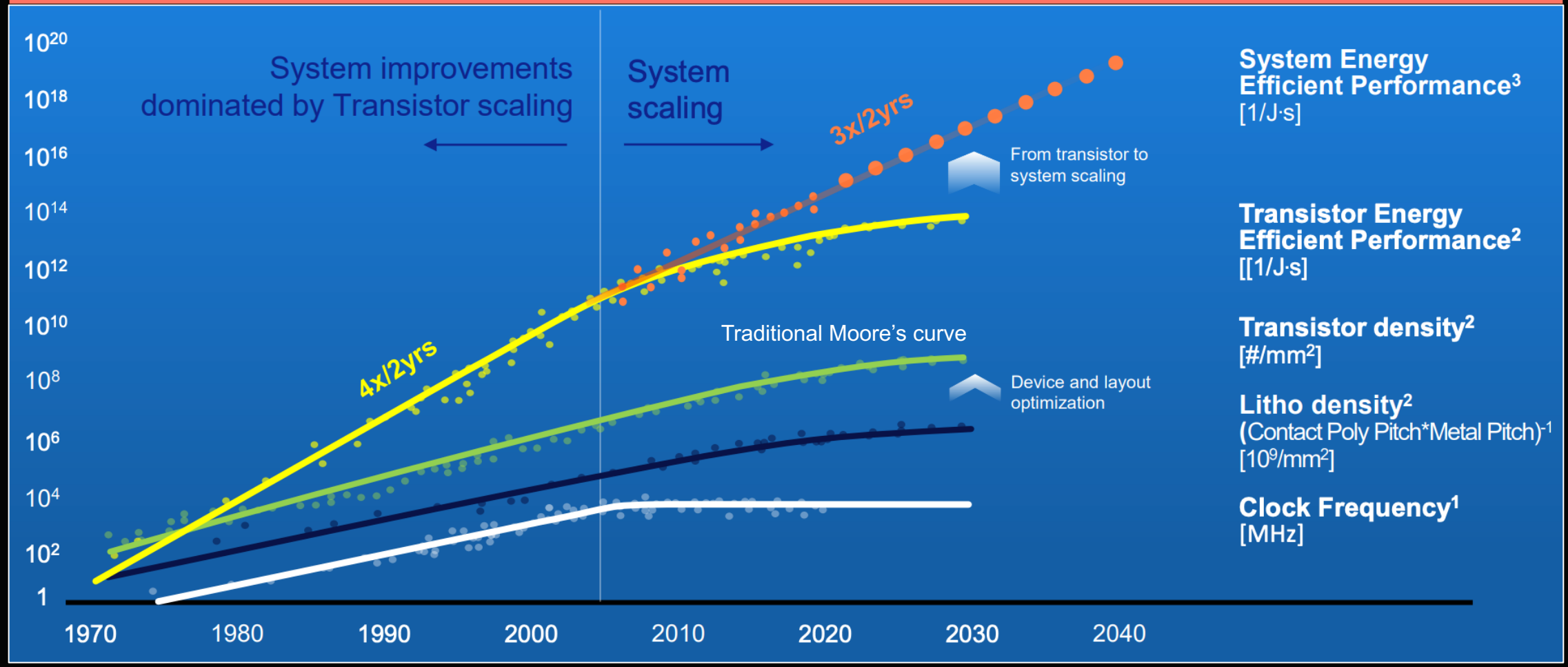


AI demand is changing how we design chips

Image source: https://www.purplequarter.com/generative-ai-new-age-creator-disruptor/all-about-tech/

Revisiting Moore's Law Evolution – An optimistic view

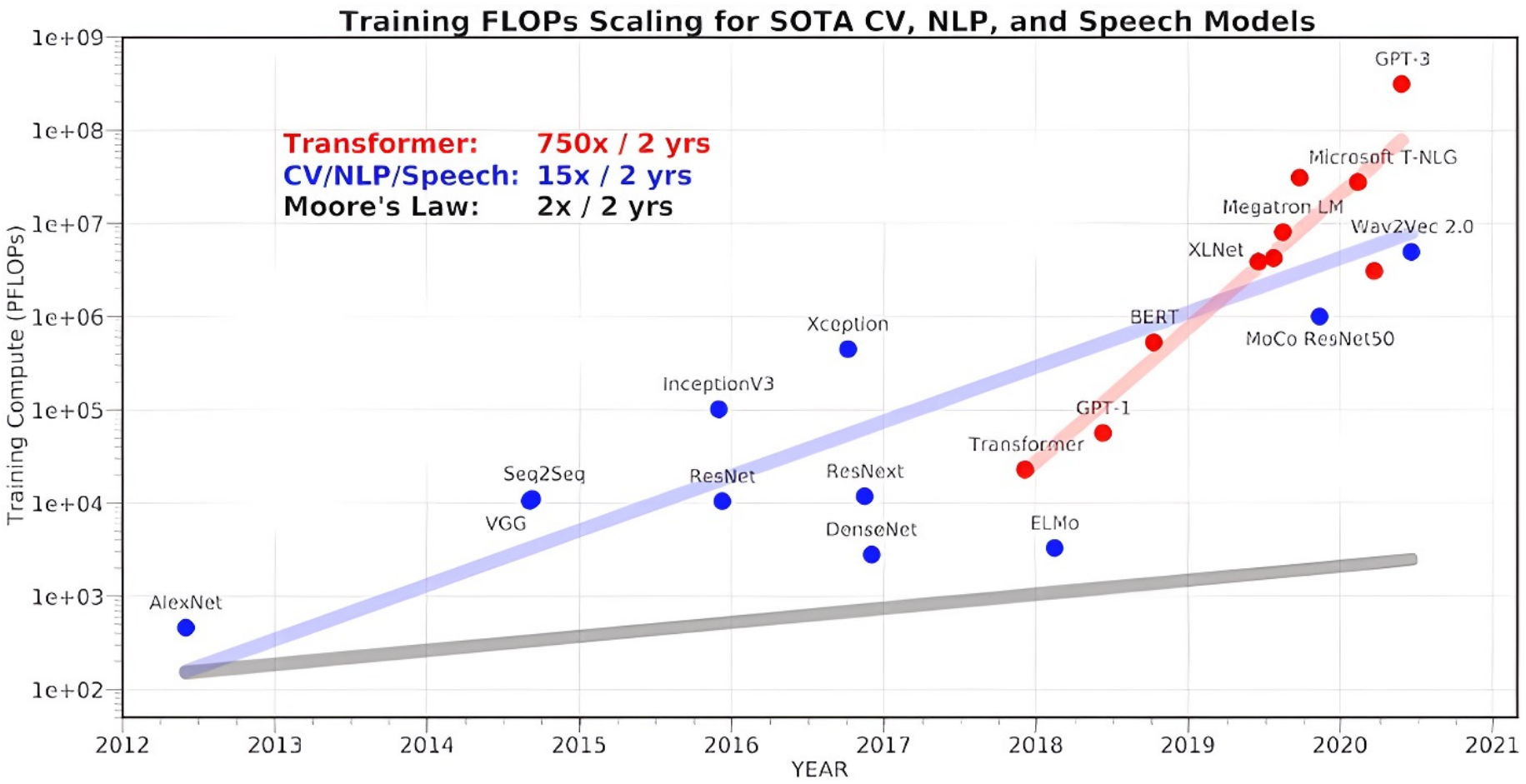
Moving into a System Scaling Era of Beyond Moore. Is this enough ?



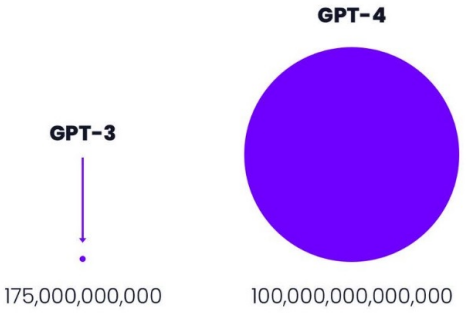
Sources: [1]Karl Rupp [2]ASML Presentation, [3]Mark Liu, TSMC, normalized to transistor EEP in 2005.

Artificial Intelligence Computing Demand vs Moore's Law

Generative AI Driving a Hyper-exponential Demand in Computing: > 100x vs Moore's

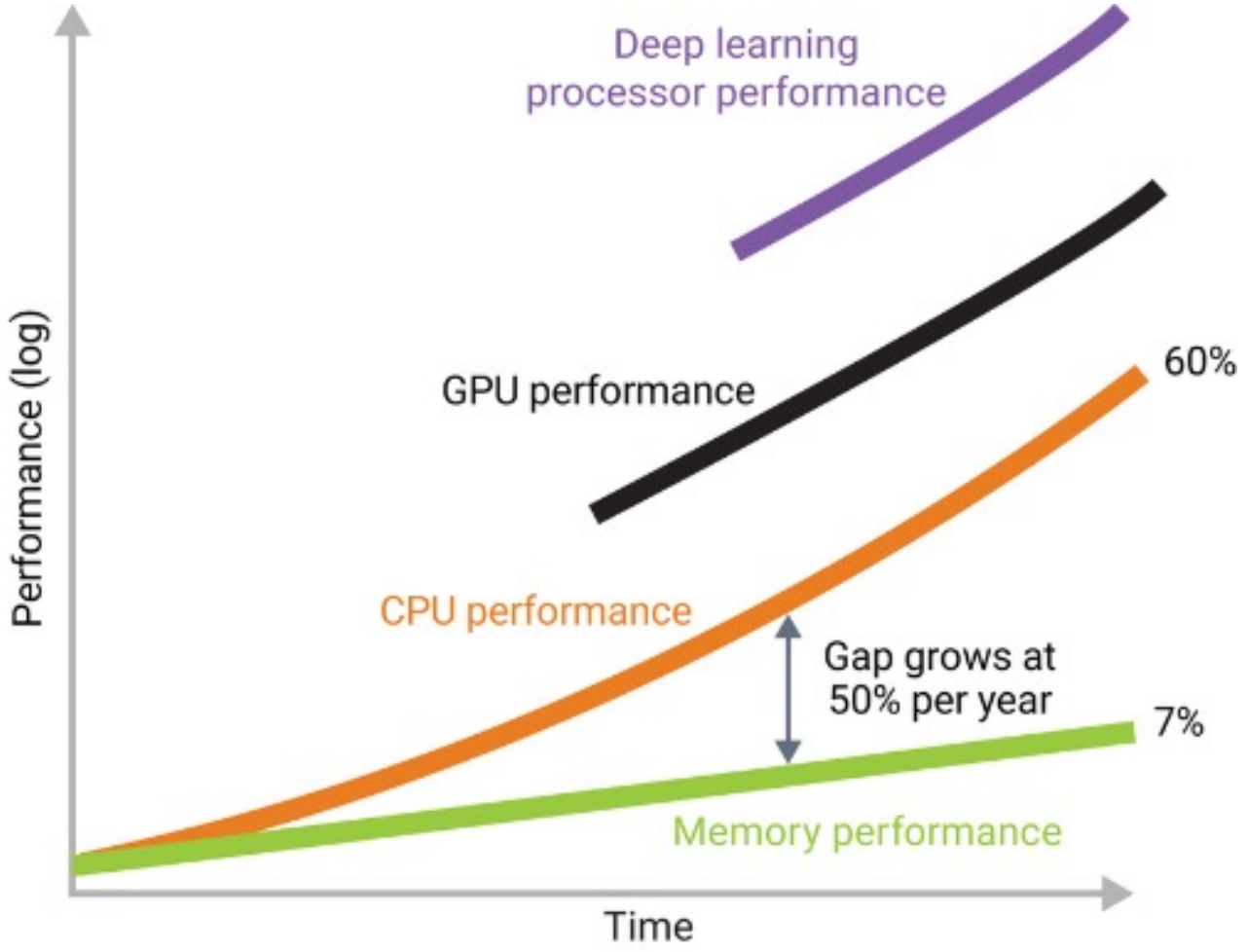


Recent Models like GPT-4 uses 1 Trillion Paramters or more ..



Overcoming the Memory Wall of AI Processing

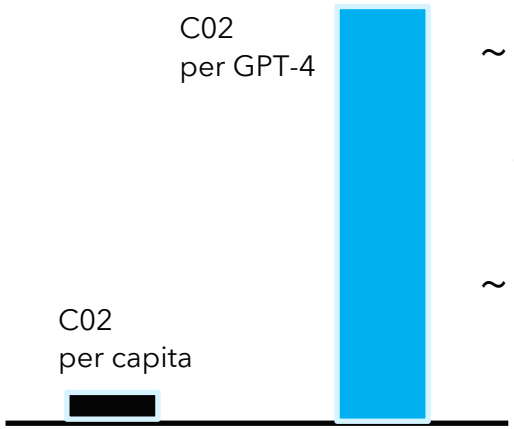
Memory is now a fundamental performance and energy bottleneck for AI Processing



Not Just Performance, Need for Higher Energy-Efficiency

Need orders of magnitude improvements in AI processor performance/watt

GPT-4 Carbon Footprint



~ 6912 Metric Tons $C0_2$ emissions for GPT-4

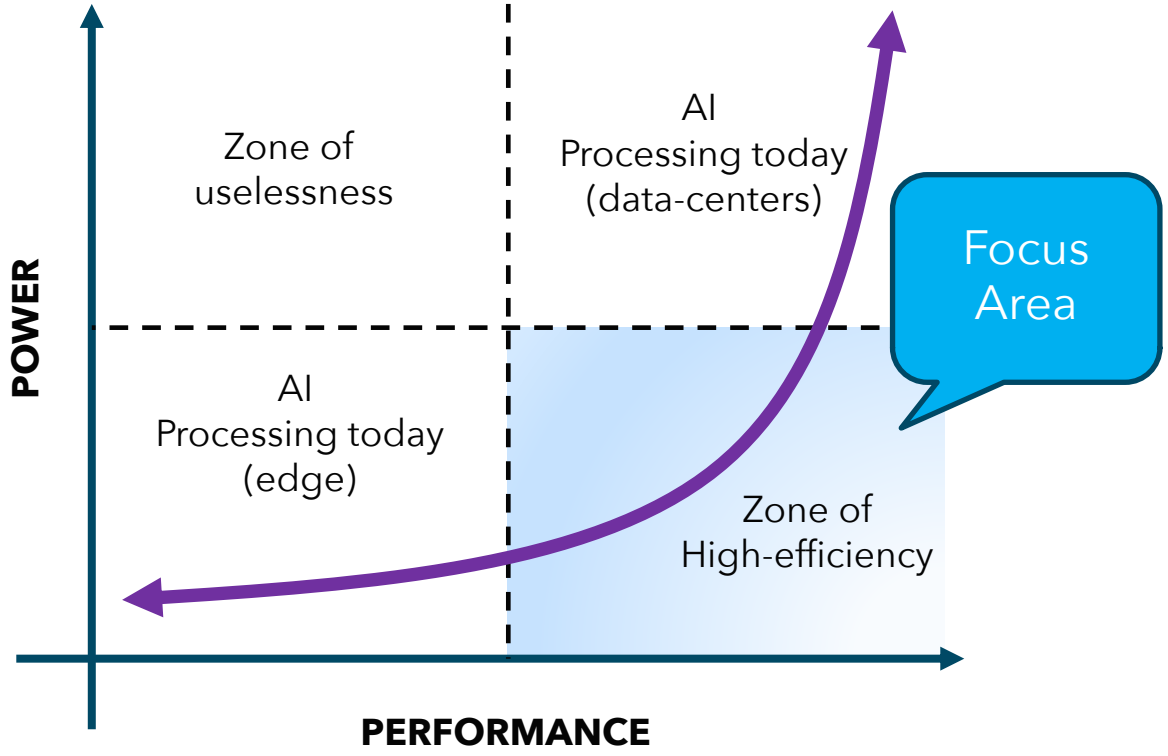
~ 1300 homes for one year

Environmental Impact of Select Machine Learning Models, 2022

Source: Luccioni et al., 2022 | Table: 2023 AI Index Report

Model	Number of Parameters	Datacenter PUE	Grid Carbon Intensity	Power Consumption
Gopher	280B	1.08	330 gC02eq/kWh	1,066 MWh
BLOOM	176B	1.20	57 gC02eq/kWh	433 MWh
GPT-3	175B	1.10	429 gC02eq/kWh	1,287 MWh
OPT	175B	1.09	231 gC02eq/kWh	324 MWh

Semiconductor Design Goal



Enable High Performance, Low Power Processing at the Edge

Reduce Expensive Data Movement by Moving more AI Processing to the Site of Data Creation

74 ZB

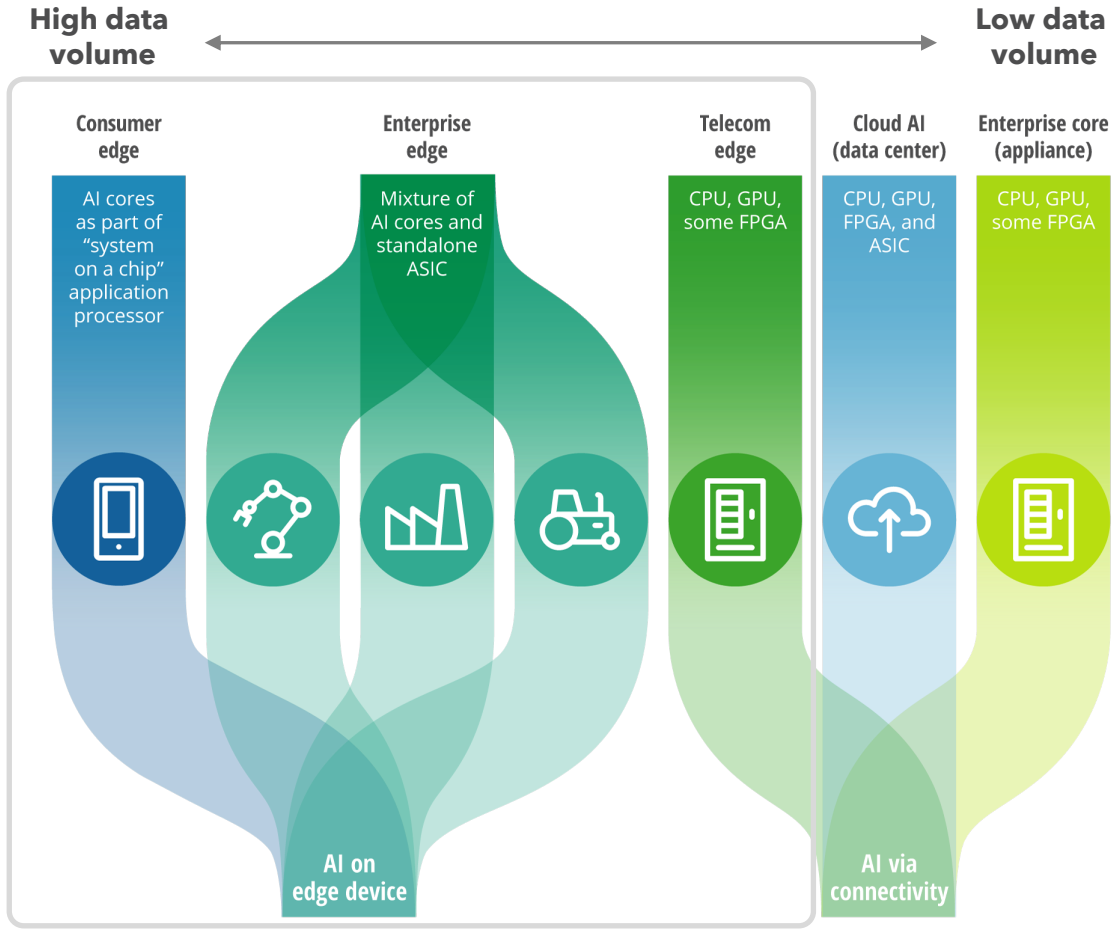
Data generated at the edge[1]

75%

Of enterprise-generated data created and processed at the edge by 2025[2]

+$80B

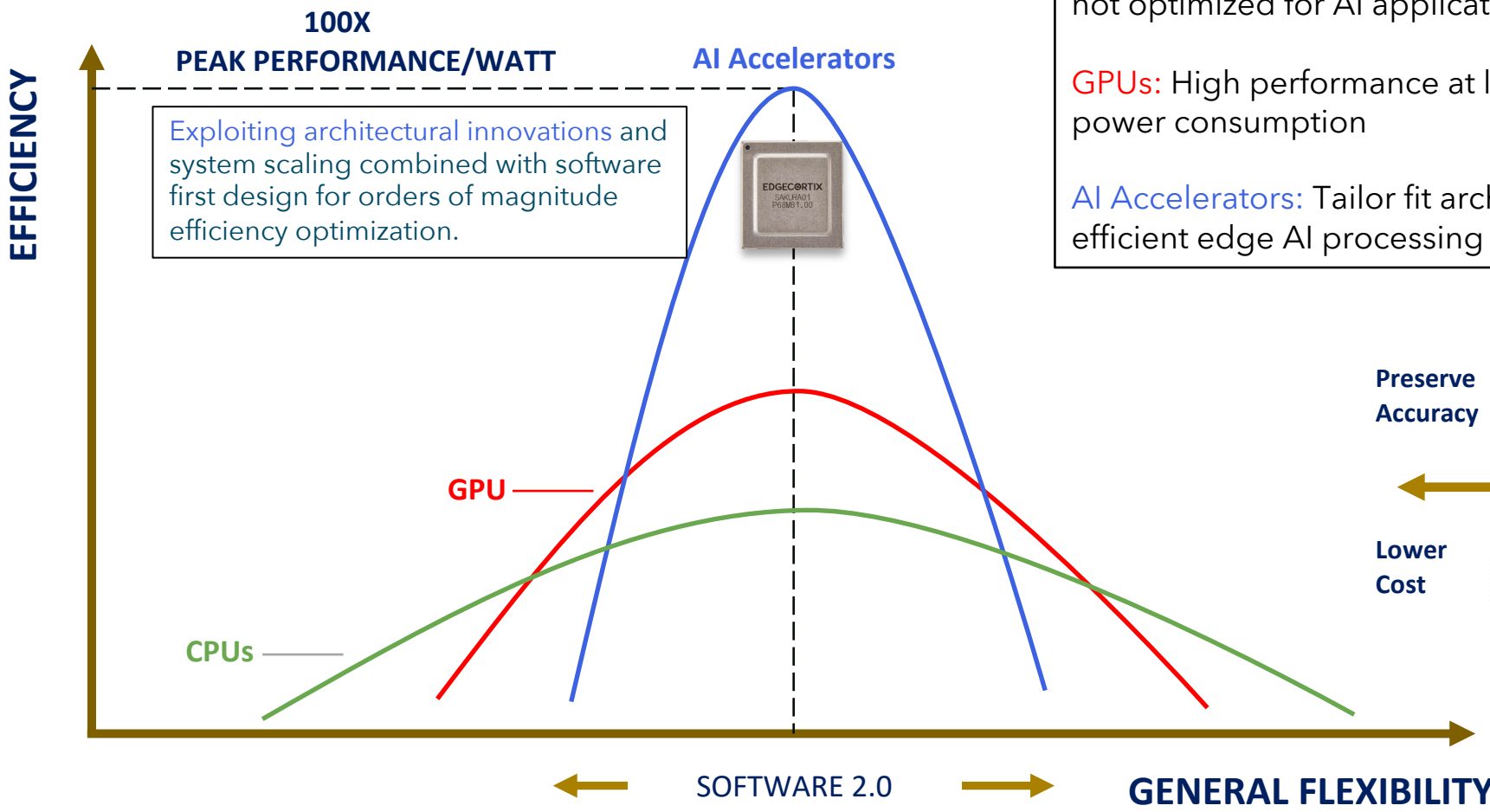
Edge AI market[3]



Source: Deloitte analysis.

Sources: [1]IDC annual estimate for 2025 [2]Gartner [3]IDC

Breakthrough Efficiency with AI Domain Specific Accelerators



CPUs: Lots of design flexibility, but highly inefficient and not optimized for AI applications

GPUs: High performance at low efficiency and higher power consumption

AI Accelerators: Tailor fit architecture for highly energy-efficient edge AI processing

Software Driven Unified Platform Delivering Highest Efficiency

Combining our AI Accelerator with Flexible Software to Deploy Power Efficient Solutions

EdgeCortix MERA

Industry first software platform enabling **AI inference** across **Heterogeneous systems**

EdgeCortix SAKURA

High Performance
Low-Power

Generative AI Ready
Effectively handles multi-billion parameter models

Unique Software

Efficient Hardware

Proprietary Architecture

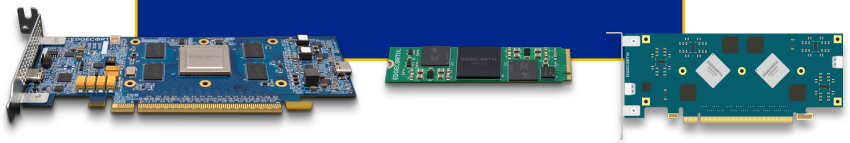
Deployable System Solutions

Dynamic Neural Accelerator Technology

Domain-specific
Highly-configurable
Architecture

Accelerator Modules & Boards

Up to 240 TOPS on board
Powered by the latest SAKURA AI Accelerators

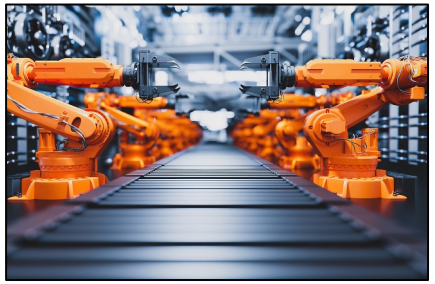


Enabling Low-power Generative Edge AI Across Markets

Smart City

Smart Retail

Smart Appliances



Smart Manufacturing

Smart Agriculture

Security

Autonomous Vehicles

Robotics

AI-RAN & Multi-Access Edge Computing (MEC)

Efficient Edge AI Processing

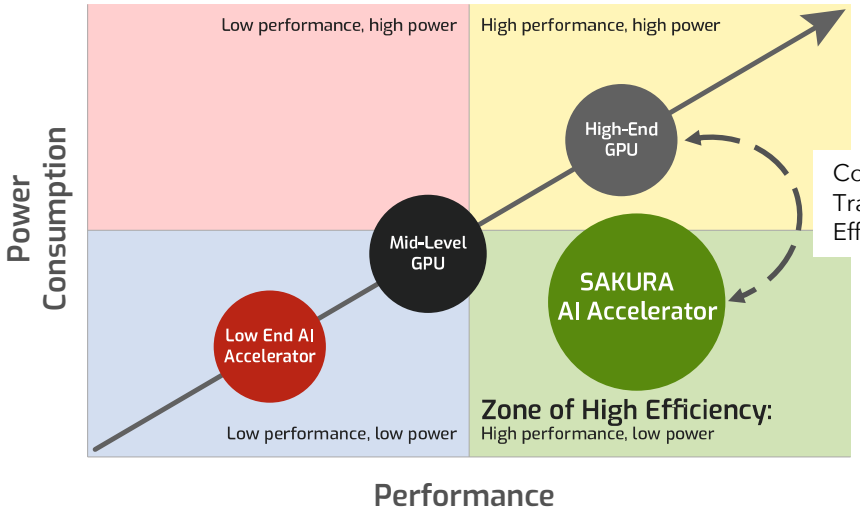
- Natural Language Processing
- Object Recognition
- Person Recognition
- AI enabled RAN
- Segmentation
- Defect Identification
- Obstacle Avoidance
- Signal Processing /MEC

Aerospace & Defense

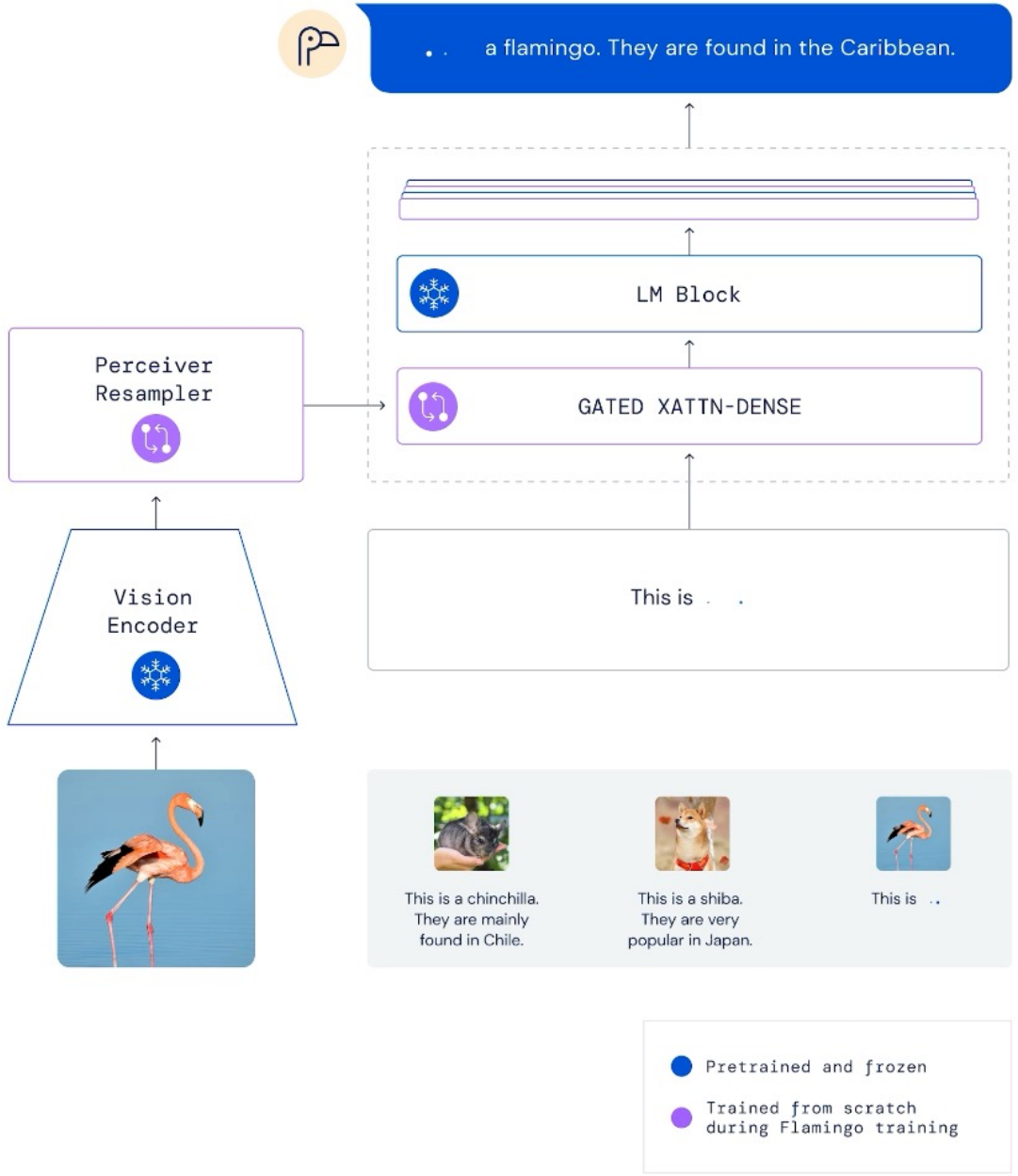
With Next-gen SAKURA:

Shaping the future of Energy-efficient Generative AI At the Edge

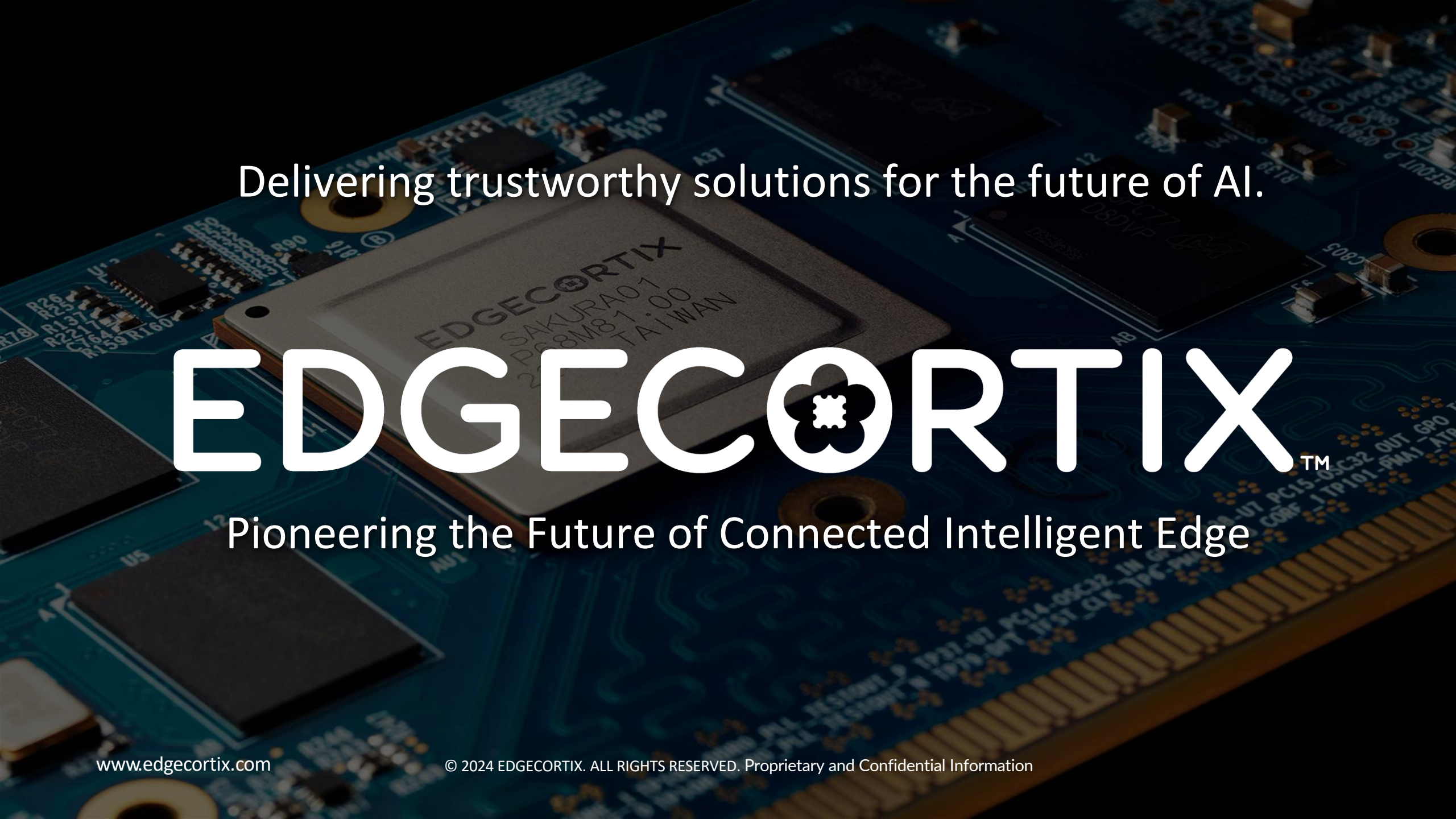
Best-in-class Performance Efficiency



Combination enables Training to Inference Efficiency



credit: deepmind.google



Delivering trustworthy solutions for the future of AI.

EDGECORTIX™

Pioneering the Future of Connected Intelligent Edge

www.edgecortix.com