socionext™
The Solution SoC Company

Beyond 100TOPS/W :
The next generation of Edge AI SoC for Smart Devices

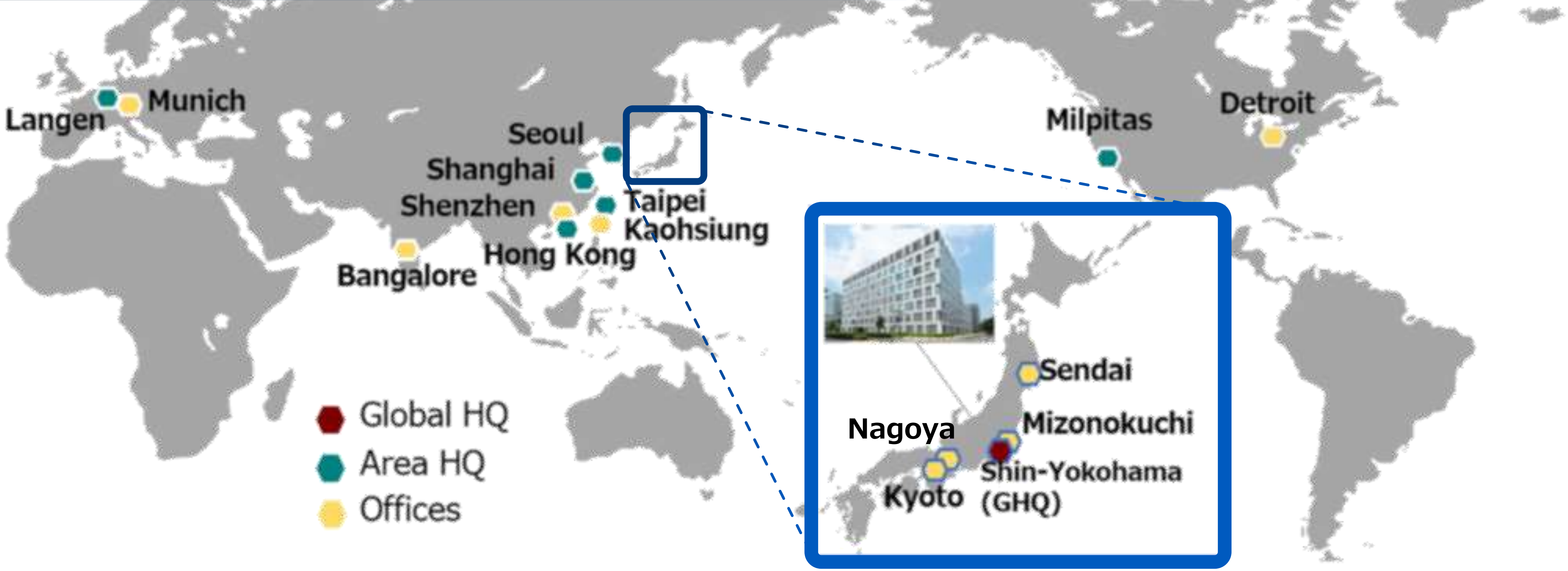
MPSoC'25

Bell Stewart (Stewart.Bell@eu.socionext.com)
Kotaro Esaki (Esaki.kotaro@socionext.com)
Socionext Inc.

about us

Socionext Inc., A Fabless Semiconductor Company

Headquarters	2-10-23 Shin-Yokohama, Kohoku-ku, Yokohama, Kanagawa, 222-0033, Japan
Established	March 1, 2015 (Merger of former Fujitsu & Panasonic System LSI divisions)
Business description	Design, development, and sale of SoC's and solutions, including services
Employees	2,600+ (1900+ Engineers, many of 20+ years experts)
Office	JP / US / EU / CN / TW / IN / KR
IPO	October 2022



smart devices

Data Center & Networking

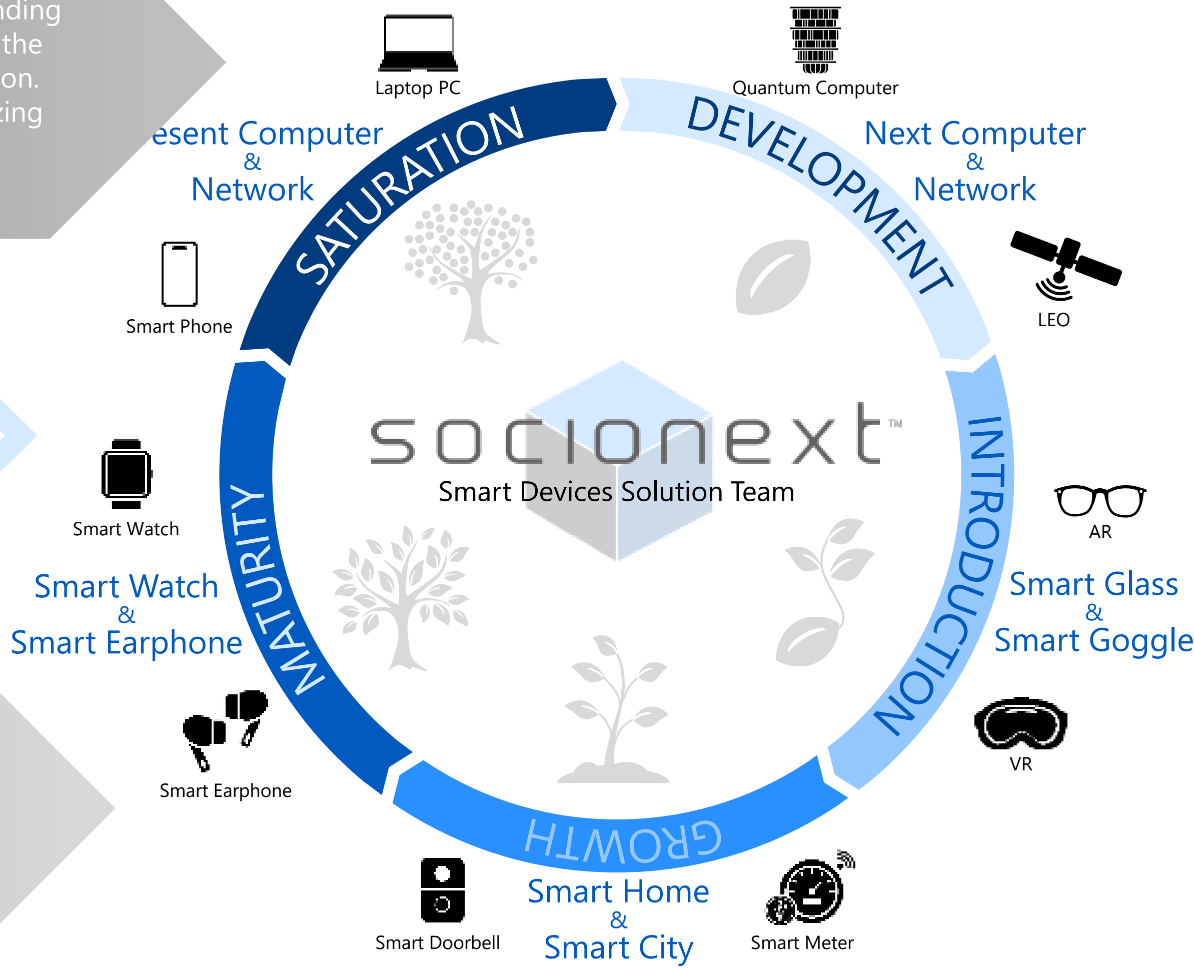
In the data center field, the increasing volume of data and application processing and communication traffic along with low latency represent major challenges. In networking, demand for large-volume, high-speed wireless communications is expanding globally, combined with the launch of full-scale services for the fifth generation (5G) ultra-high-speed wireless communication. Socionext develops high-performance custom SoCs by utilizing its multi-core design capability and low-power AI engine and accelerator to meet the customers' exact needs for feature and performance.

Smart Devices

The smart device market is expected to grow rapidly as AR/VR products and smart glasses are becoming smaller and more feature-rich.
Socionext will continue to further advance its technologies for high performance SoC designs to meet the needs for smaller form factors and reduced power consumption.

Automotive

The next generation of automotive development features a shift to autonomous driving.
Dedicated SoCs with the most advanced technologies are needed to enable these highly demanding systems.
We collaborate with major car OEMs and Tier-1 manufacturers that lead the global market, delivering custom SoCs required for advanced applications, as well as contributing to a safe, secure, environmentally friendly, and comfortable mobility.

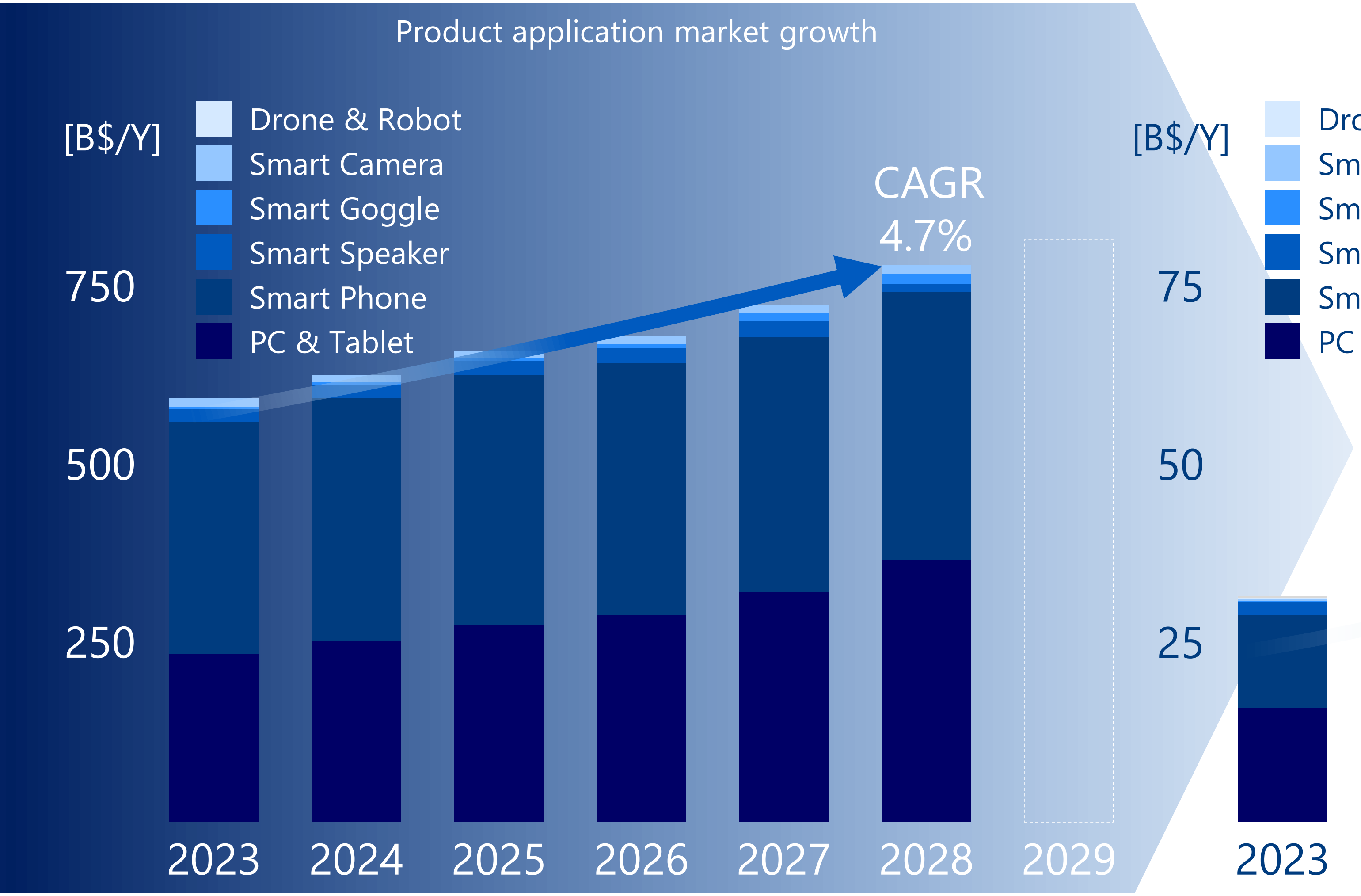


Socionext serves the diverse needs of the world's leading companies by leveraging its system expertise and core technologies garnered over the years through ASIC and ASSP businesses.
By contributing to the following focus areas, we take on challenges to create a better society and ensure a better quality of experiences for people around the world.

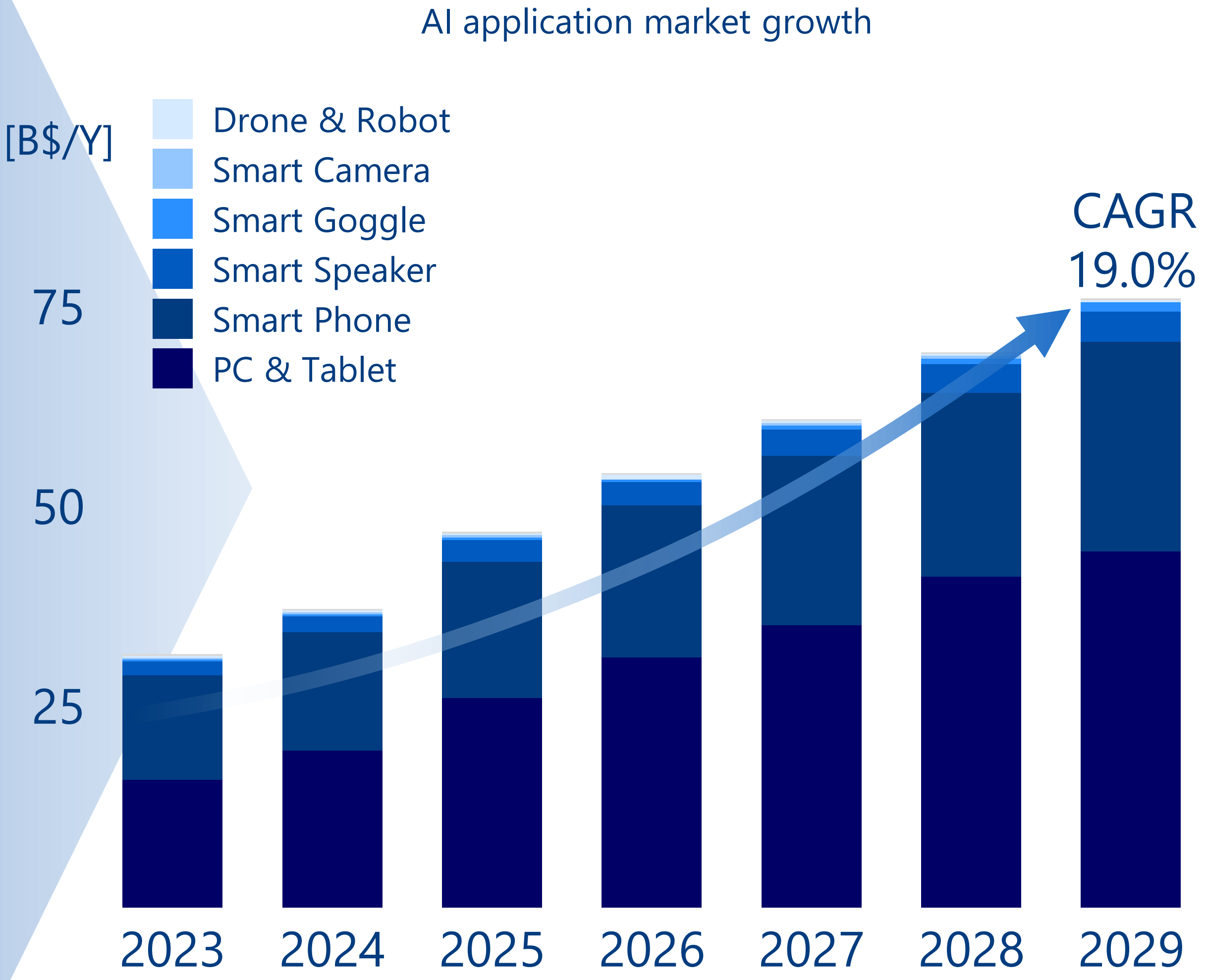
Smart devices market growth is accelerated by AI built in.

market trend

The smart device market, centered on consumer products, is expected to grow moderately, while AI in smart devices is expected to grow rapidly in the future. In particular, while the market for PCs and smartphones is almost saturated and is not expected to grow significantly, the market for AI-equipped PCs and smartphones is expected to grow significantly.



Created by SOCIONEXT based on OMDIA AMFT Shipment: World & Regions Application Market Forecast Tool –2Q/2024

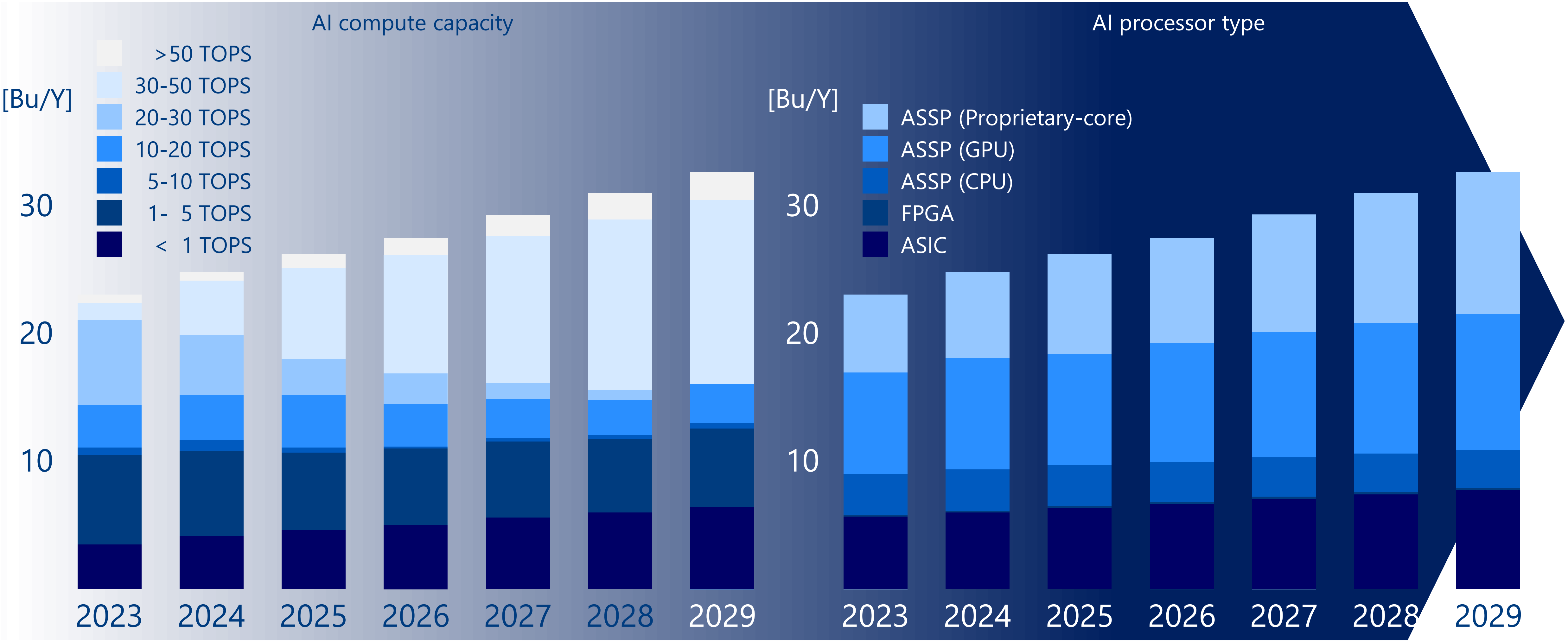


Created by SOCIONEXT based on OMDIA AI Processors for the Edge Forecast Report - 2024

AI will be divided into highs and lows, and half will be custom or proprietary

ai trend

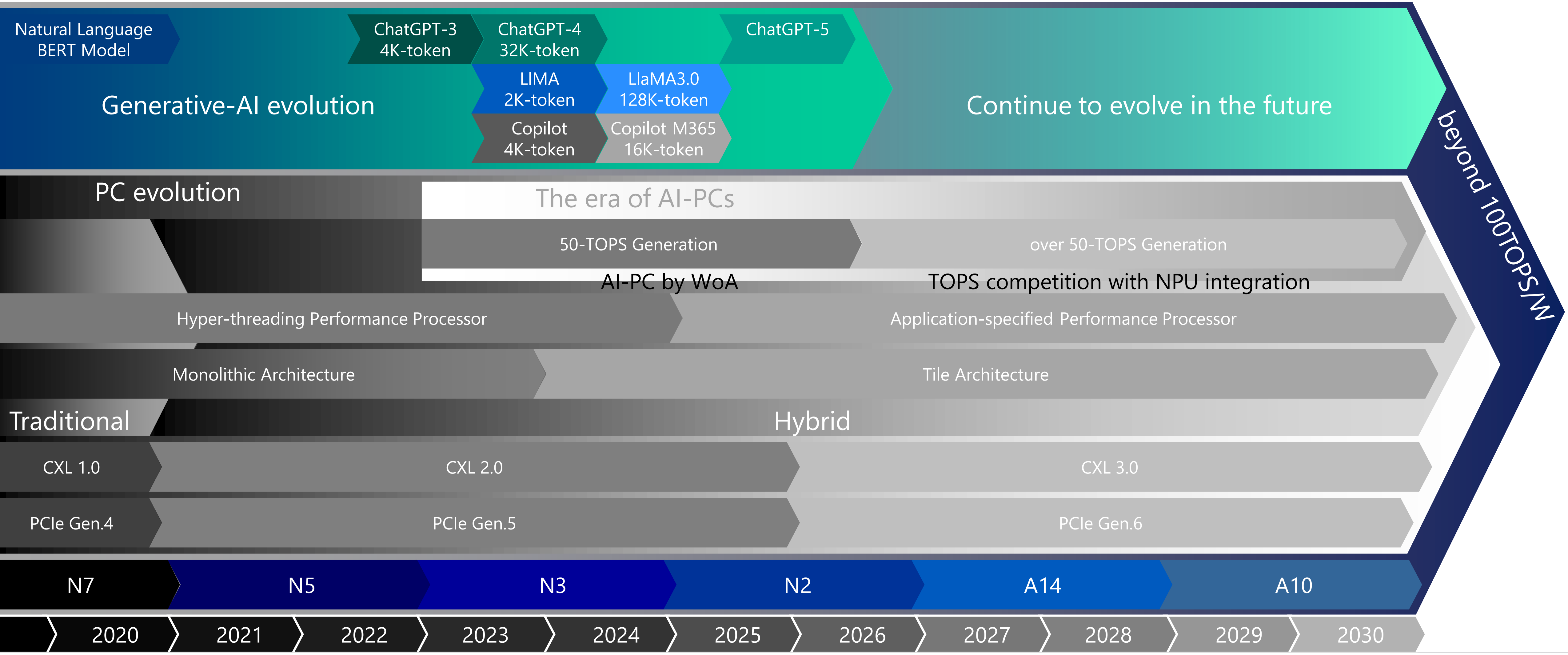
In the future, AI installed in smart devices will become polarized between 10TOPS or less and 30TOPS or more, and AI devices with medium performance are expected to disappear due to lack of use. About 1/4 of SoCs for realizing edge AI are ASICs. If SoCs that implement their proprietary AI cores are included, about half are expected to be for devices specialized in applications.



Created by SOCIONEXT based on OMDIA AI Processors for the Edge Forecast Report - 2024

Changes in Computer Architecture Caused by the Evolution of AI

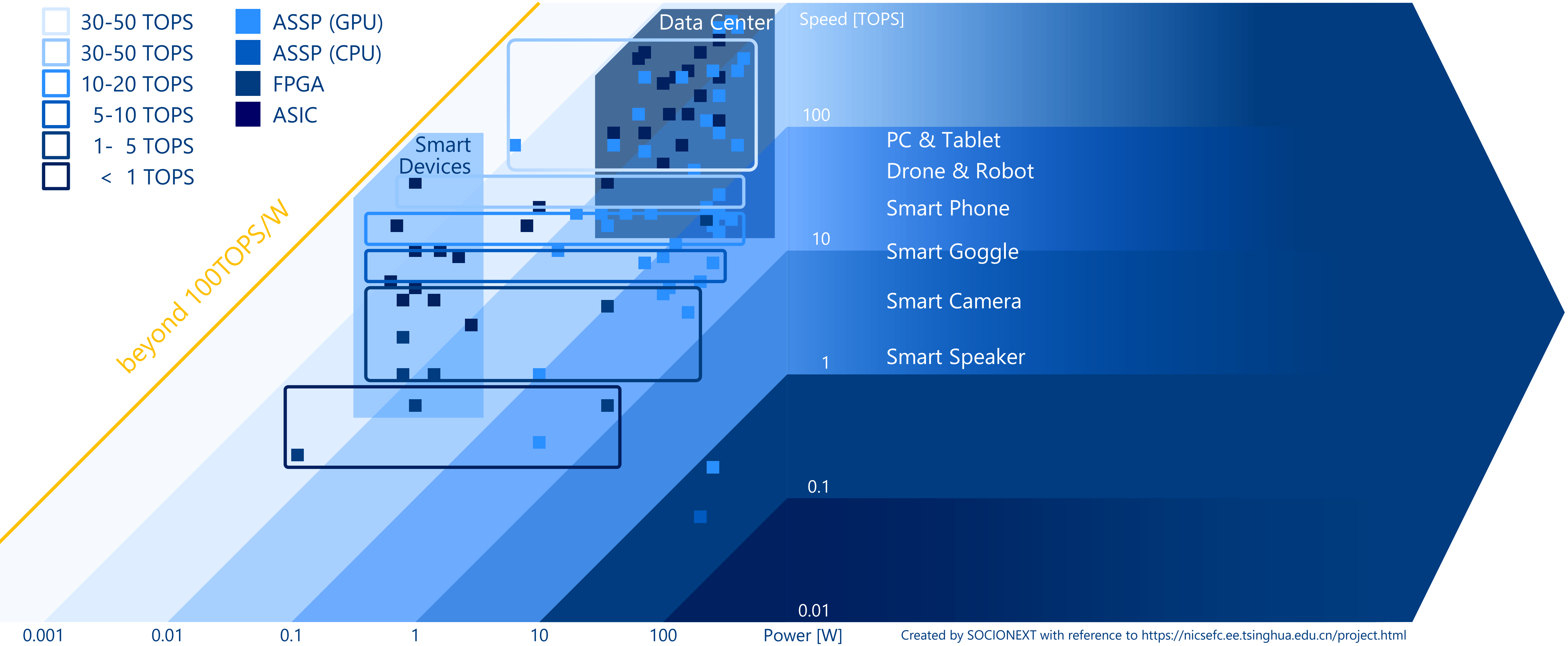
With the miniaturization of semiconductor manufacturing technology, computer architecture is evolving, shifting from traditional monolithic hyperthreading designs to application-specific types utilizing tile structures enabled by faster interfaces. AI-focused components like NPUs are noteworthy, particularly with the evolution of LLMs. Various AI processors have been developed, surpassing 50TOPS in performance, with a goal set at 100TOPS while emphasizing power efficiency.



Motivation to achieve 100TOPS/W

ai target

When the current AI processors are plotted by type on the graph of performance versus power, and arranged by AI compute capacity, ASIC is almost adopted to realize SoC with high power efficiency for smart devices. ASIC has also been widely adopted for high-performance AI processors for Server, but it has not yet broken through 100TOPS/W, including for smart devices, and this breakthrough will be a future proposition.

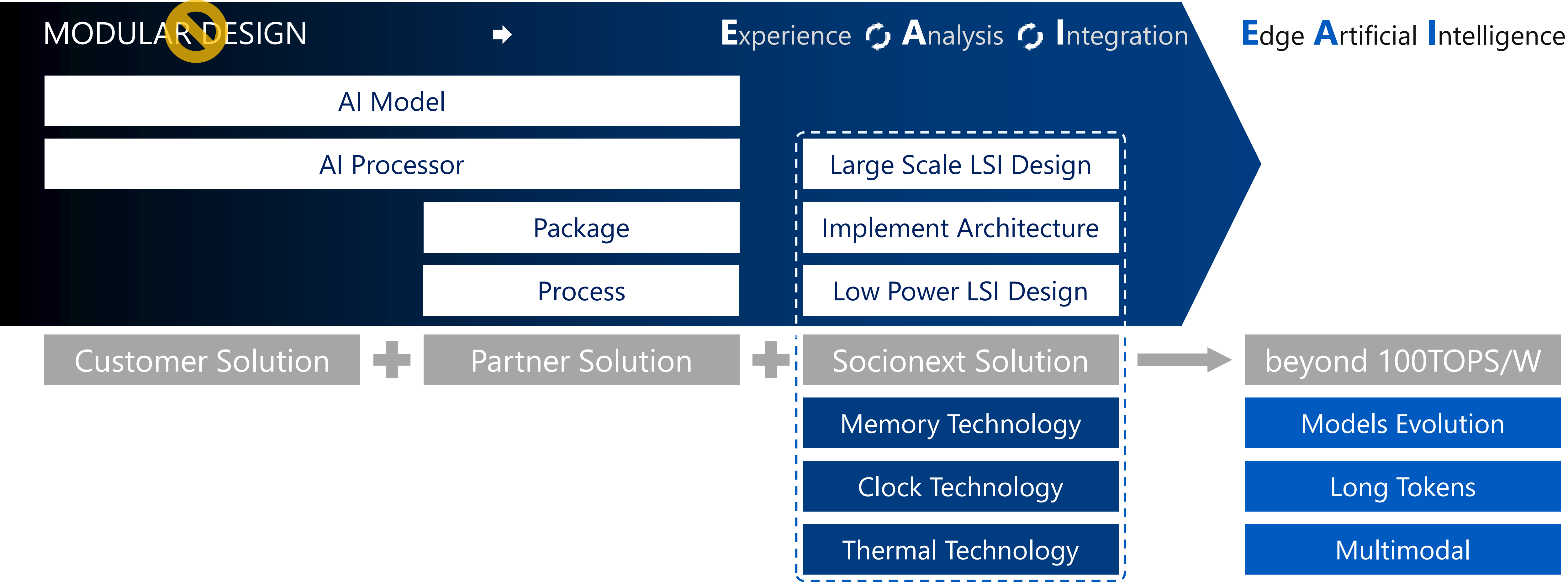


Created by SOCIONEXT with reference to https://nicsefc.ee.tsinghua.edu.cn/project.html

Experienced Analytic Integration for Edge Artificial Intelligence

socionext solution

In order to achieve the ambitious target of 100 TOPS/W, conventional modular design methods have reached their limits. Socionext focuses on analyzing requirements and leveraging our extensive experience to find the best solution. We work with our customers and partners to save power and reduce cost, especially through advances in memory technologies, clock strategies, and thermal studies.



MODULAR DESIGN ➡ Experience ↻ Analysis ↻ Integration

Edge Artificial Intelligence

AI Model

AI Processor

Package

Process

Large Scale LSI Design

Implement Architecture

Low Power LSI Design

Customer Solution ➕ Partner Solution ➕ Socionext Solution → beyond 100TOPS/W

Memory Technology

Clock Technology

Thermal Technology

Models Evolution

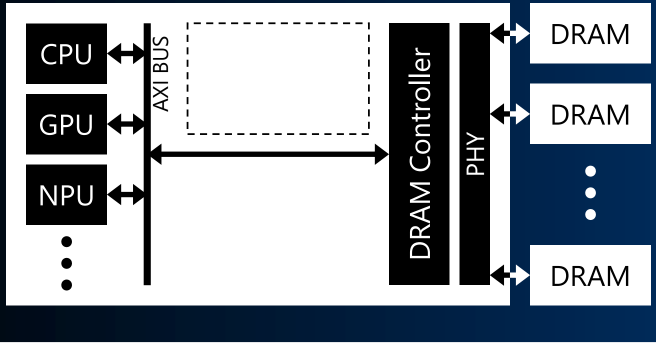
Long Tokens

Multimodal

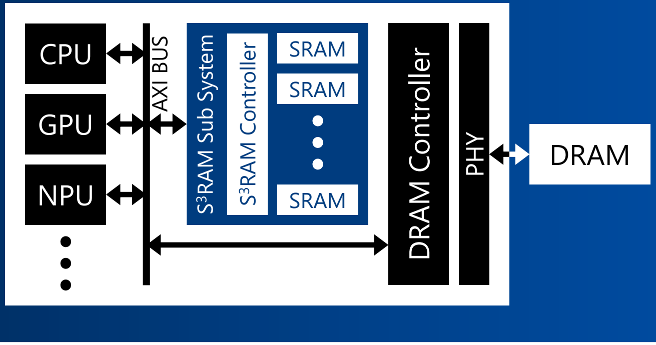
Socionext Smart SRAM for realization of DRAM-LESS and DRAM-LITE systems

In smart-devices and the like, ultra-low power consumption is required due to physical limitations of mountable batteries and long operation times (from half a day to several days). In AI processing, DRAM has been used as a memory to store learning models, but DRAM-LESS and DRAM-LITE (DRAM Access Reduction) systems are required to satisfy the requirements for edge AI installed in smart devices.

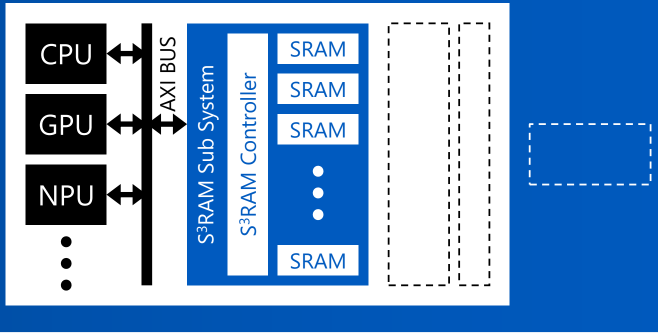
■ Conventional System Configuration



■ DRAM-LITE System Configuration



■ DRAM-LESS System Configuration



■ S3RAM Feature

■ Intelligent Layout

- HLB SRAM with control circuit in a few decade Mbit units
- Large capacity SRAM with low layout difficulty

■ Power Management

- Unit active/sleep automatic mode control
- Power minimization control using local memory access

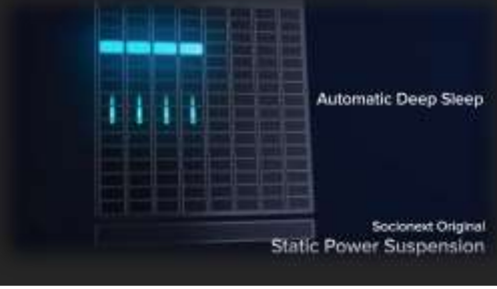
■ Memory Interface

- General purpose bus connection such as AXI
- Realization of DRAM-LESS/LITE with minor modification

■ Large Memory Implementation Technic

Silicon proven

■ Static Power Reduction Technology



Already developed

■ Dynamic Power Reduction Technology

Under development

Socionext Clock Driver for optimal clock structure for AI processor implementations

clock technology

For design convergence, power reduction, and performance improvement, three different clock structures are used together to minimize latency, skew and power consumption to realize SoC. In particular, since the spine tree method fits the implementation of AI processors, we are also working on the development of a super clock driver that is optimal for this architecture.

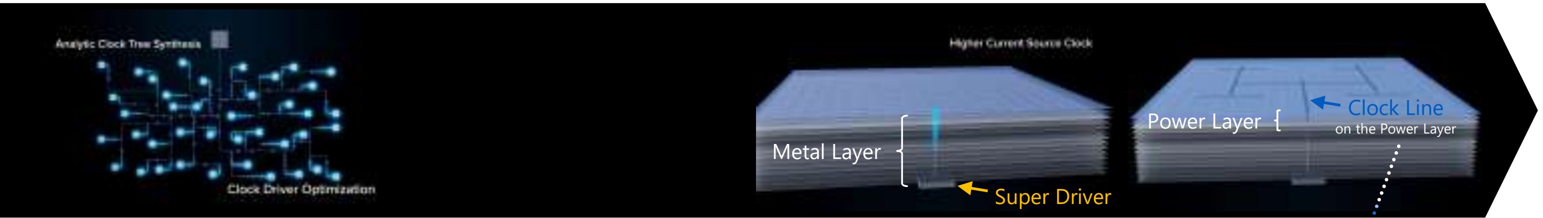
■ Block Level in SoC

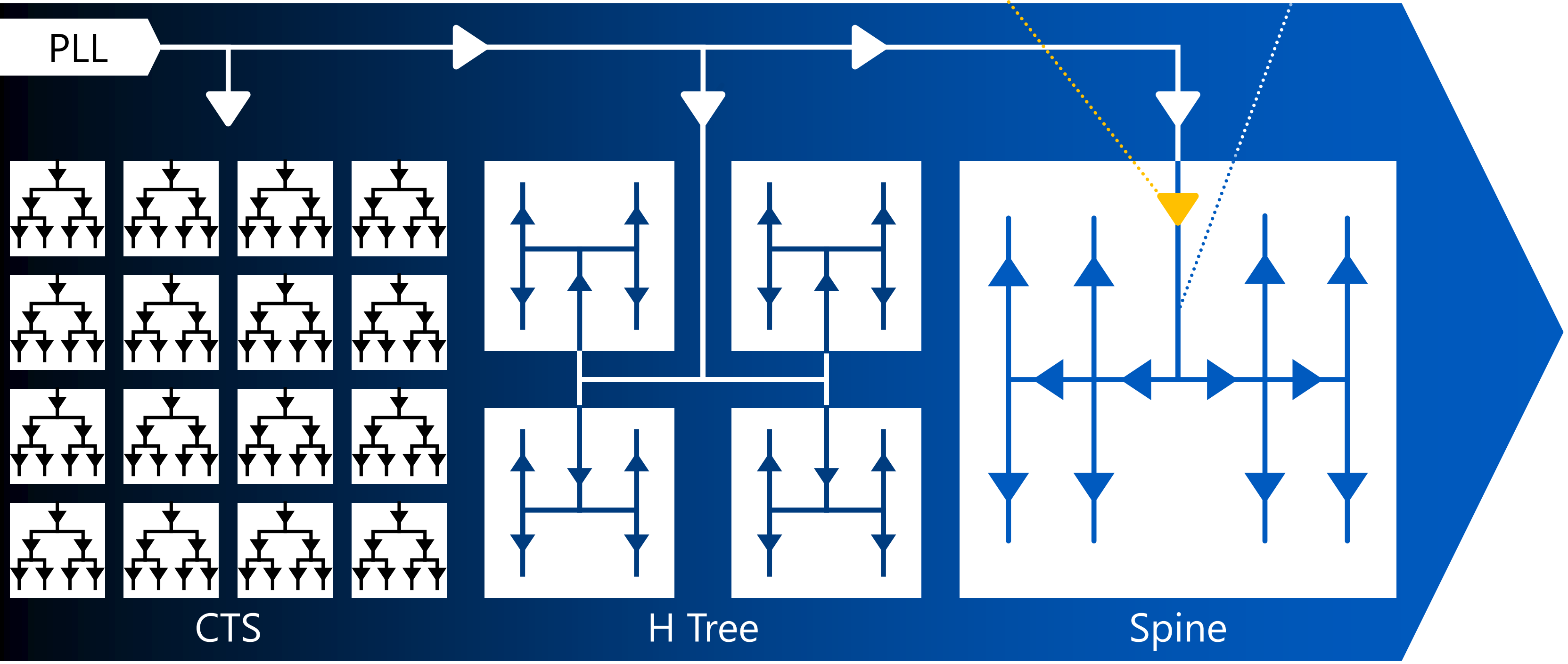
- ■ Analyze clock skew and latency
- ■ Determine clock structure

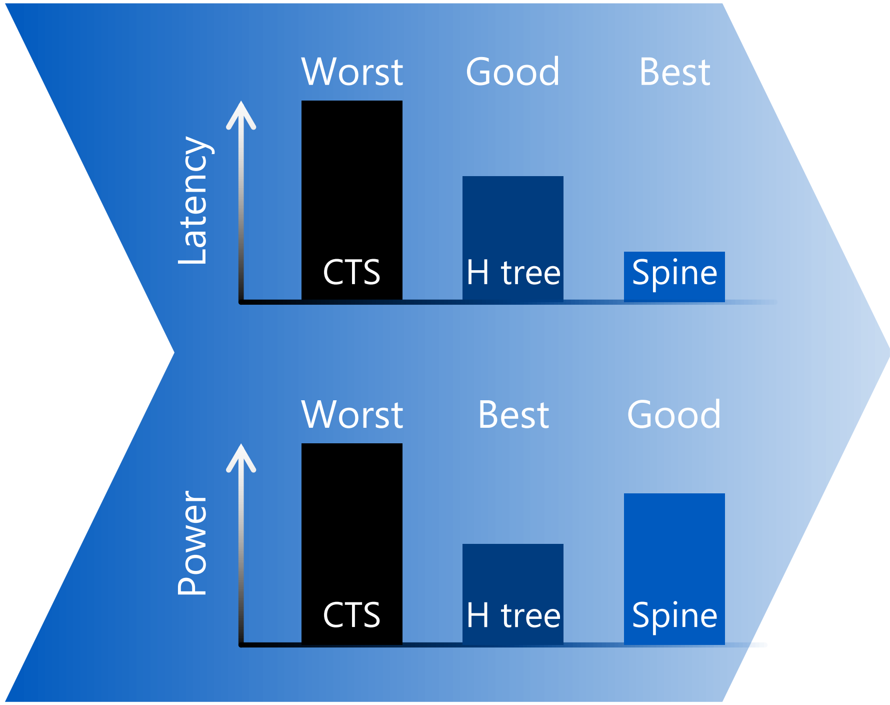
Explore the best clock structure

■ Top Level in SoC

- ■ Search for clock distribution path
- ■ Evaluate on chip variation


Analytic Clock Tree Synthesis
Clock Driver Optimization
Higher Current Source Clock
Metal Layer
Super Driver
Power Layer
Clock Line
on the Power Layer



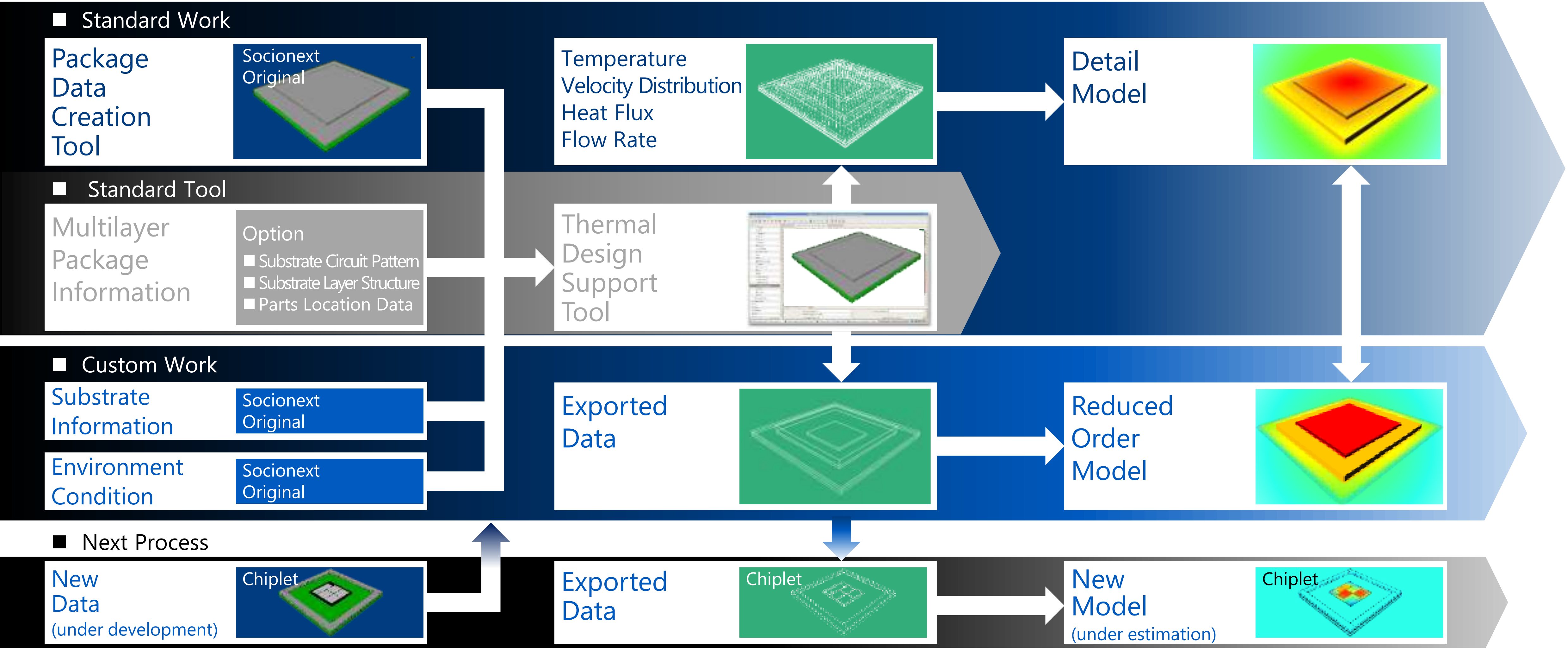


Socionext Thermal Solution to help customers solve their thermal problems

thermal technology

As market demands for thermal analysis evolve, it is essential to enhance our expertise and gather insights, such as system-level evaluations and hotspot behavior. We leverage standard analysis tools alongside decades of packaging experience to perform precise thermal simulations. This expertise enables us to develop reduced-order models that can be effectively shared with customers. We are also preparing for more advanced packages.

■ Legacy Process



■ Standard Work

Package Data Creation Tool — Socionext Original

Temperature
Velocity Distribution
Heat Flux
Flow Rate

Detail Model

■ Standard Tool

Multilayer Package Information

Option
- ☐ Substrate Circuit Pattern
- ☐ Substrate Layer Structure
- ☐ Parts Location Data

Thermal Design Support Tool

■ Custom Work

Substrate Information — Socionext Original

Environment Condition — Socionext Original

Exported Data

Reduced Order Model

■ Next Process

New Data (under development) — Chiplet

Exported Data — Chiplet

New Model (under estimation) — Chiplet

Shift modular design to Suriawase design as a method

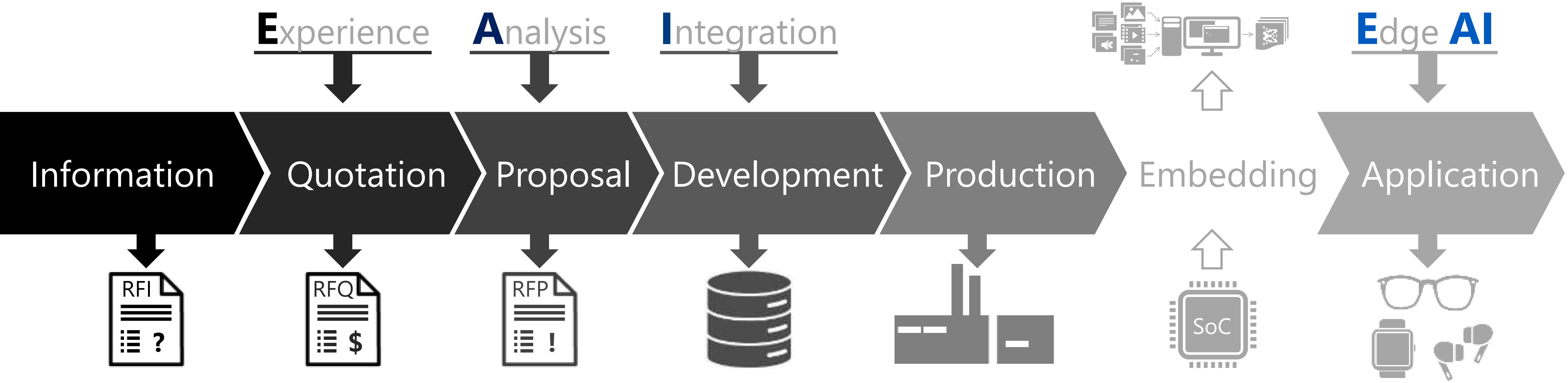
summary

In order to achieve the ambitious target of 100 TOPS/W, conventional modular design methods have reached their limits. Socionext focuses on analyzing requirements and leveraging our extensive experience to find the best solution. Instead of the traditional modular design approach, we intend to breakthrough an unprecedented goal with a design approach called SURIAWASE in Japanese.

■ SURIAWASE

- ■ Coordinate : To coordinate with multiple stakeholders
- ■ Align : To align our goals and strategies
- ■ Harmonize : To harmonize opinions and interests
- ■ Reconcile : To reconcile differences and interfaces
- ■ Fine-tune : To fine-tune designs and tests

beyond 100 TOPS/W



Experience
Analysis
Integration
Edge AI

Information
Quotation
Proposal
Development
Production
Embedding
Application

socionext™

The Solution SoC Company