Enabling Intelligence

Through Sustainable Computing

SIEMENS

"

sus·tain·a·bil·i·ty

noun

- The ability to be maintained at a certain rate or level:

"**the sustainability of economic growth**" · "**the long-term sustainability of the project**“

- Avoidance of the depletion of natural resources to maintain an ecological balance:

"the **pursuit of global environmental sustainability**" · "the ecological sustainability of the planet"

Three pillars of sustainability

Technological

Ensuring the product can be built and is fit for purpose

- Power
- Performance
- etc

Economic

Can it be built in a way that is profitable

- Cost of manufacture
- Cost of ownership
- etc

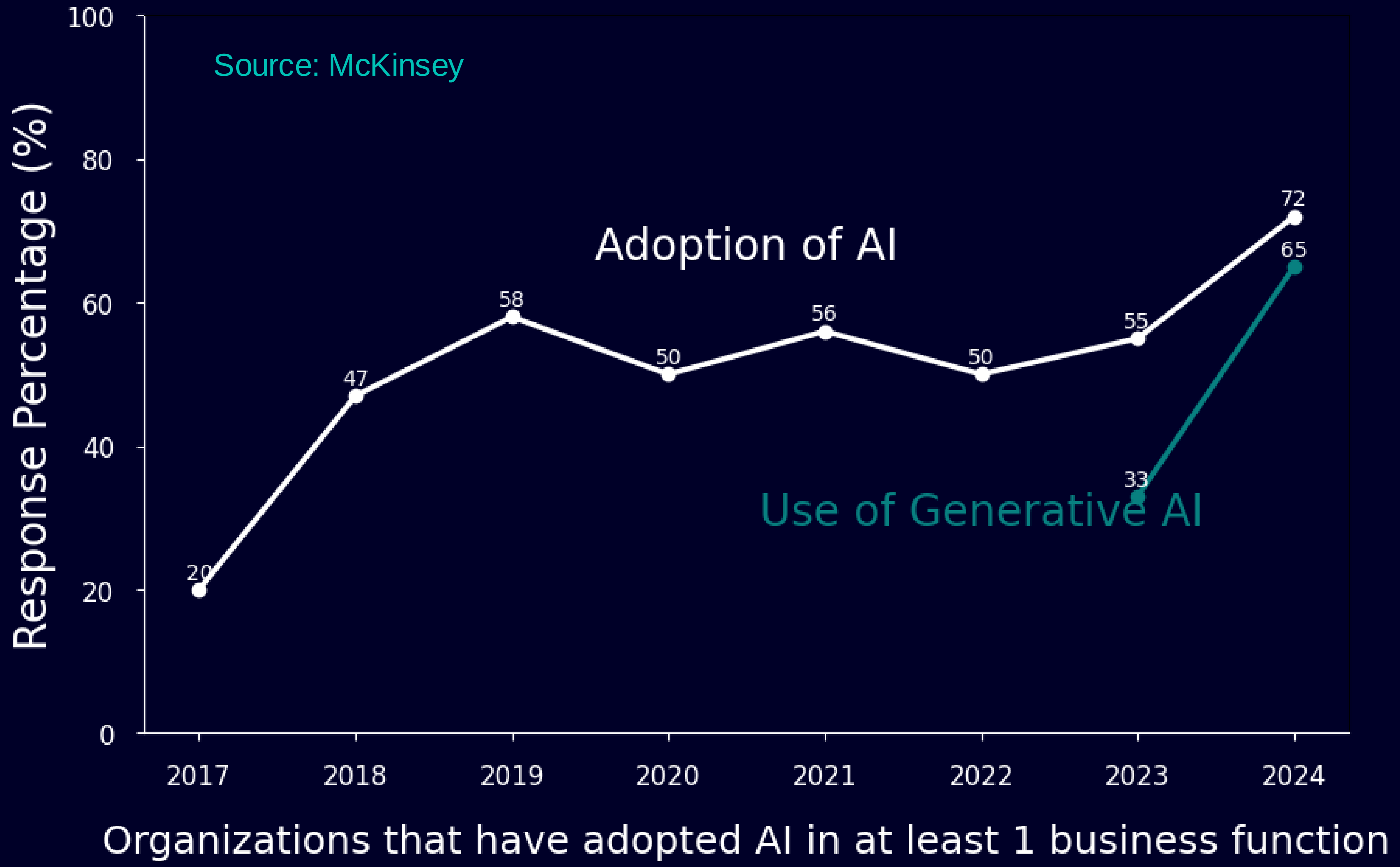
Environmental

Minimizing the negative impact on the environment

- Energy to manufacture
- Materials
- etc

Need to keep all three in balance to build successful businesses, it is bigger than just environmental sustainability

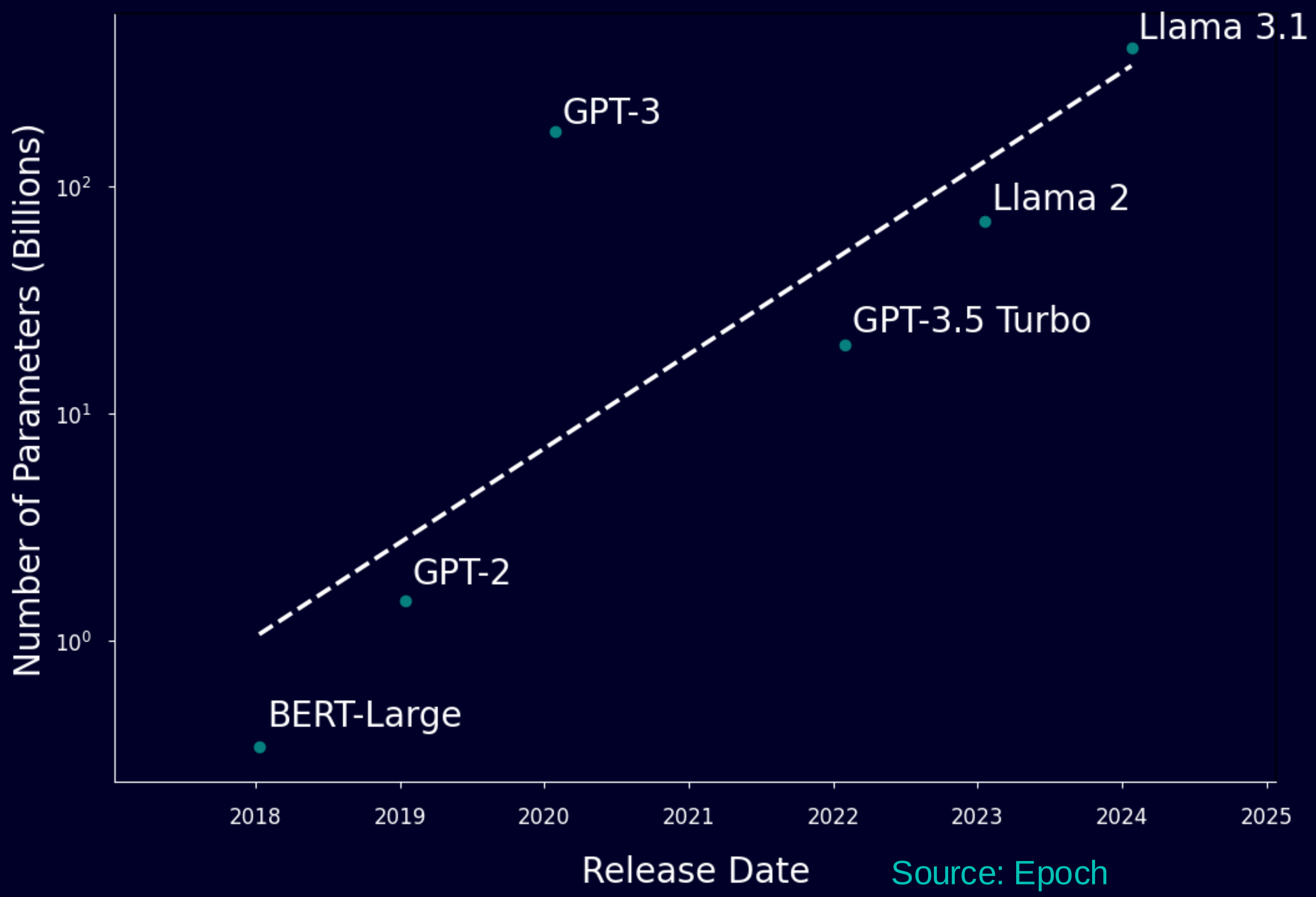
AI adoption trending upwards, with GenAI usage doubling since 2023



Growth will continue to accelerate as more use-cases come online and larger models expand applicability

SIEMENS

AI model size increase continues to rapidly increase

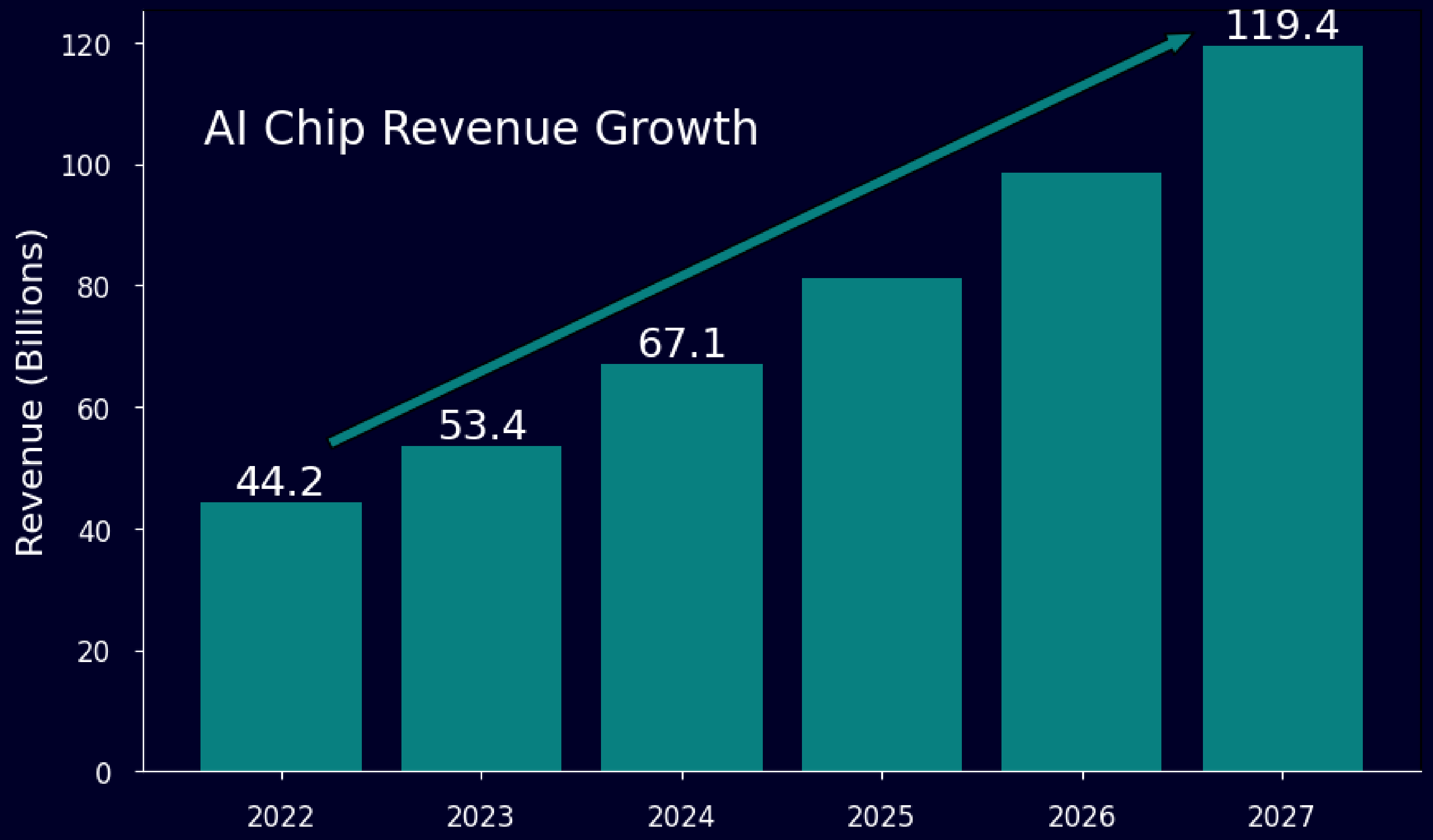


Source: Epoch

As AI model size grows so does its capability and demands on sustainability

- More complex technology needs
- Higher power demands
- Greater resource utilization

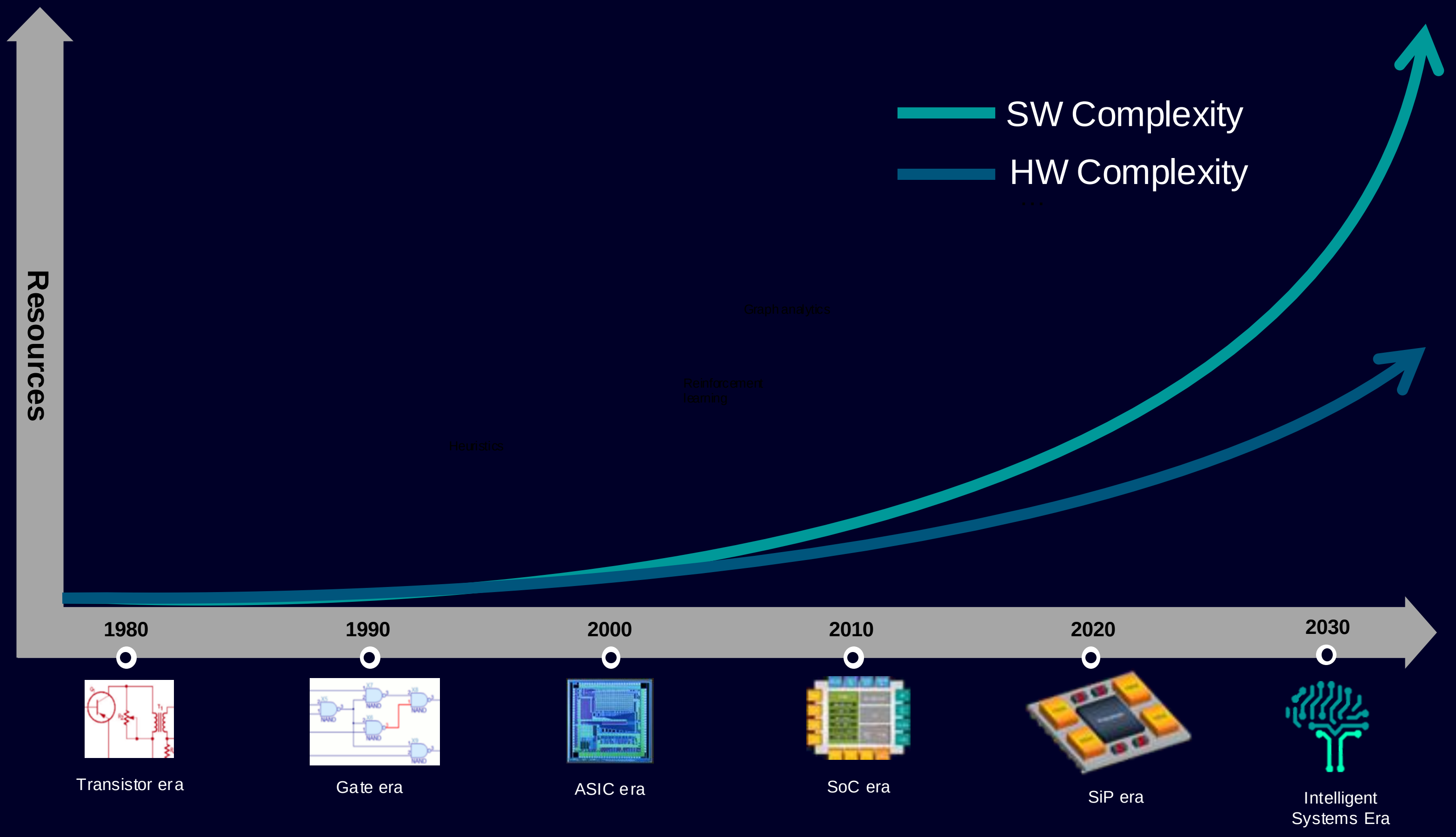
The AI growth drives compute demand, boosting the semiconductor industry



Source: Gartner

SIEMENS

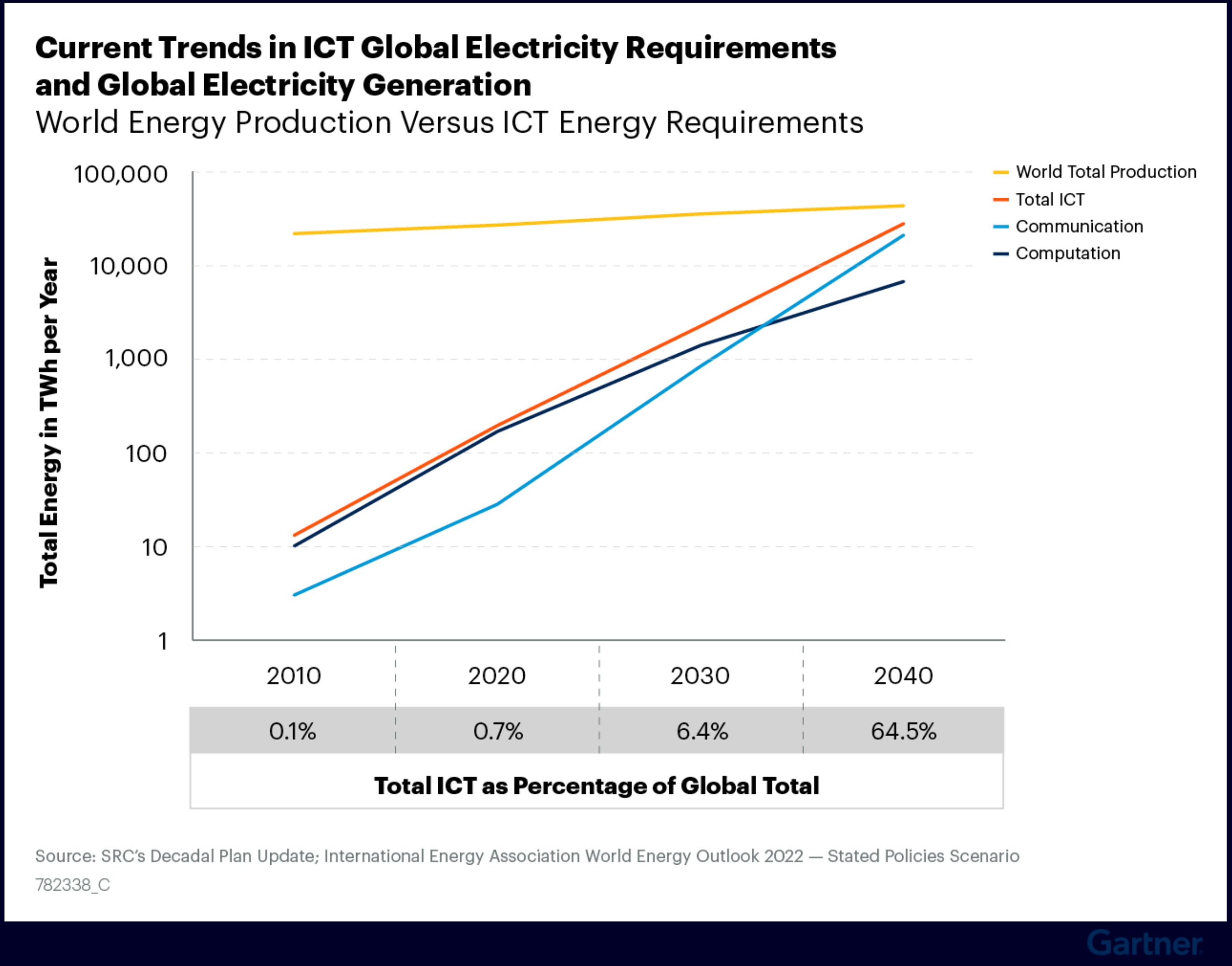
Complexity is compounding



Hardware and software **complexity is linked** and growing **at an exponential rate**

SIEMENS

AI and sustainability are on a collision course



- Energy demand growth **significantly outpaced** supply growth
- Hitting an **energy wall** by 2040

SIEMENS

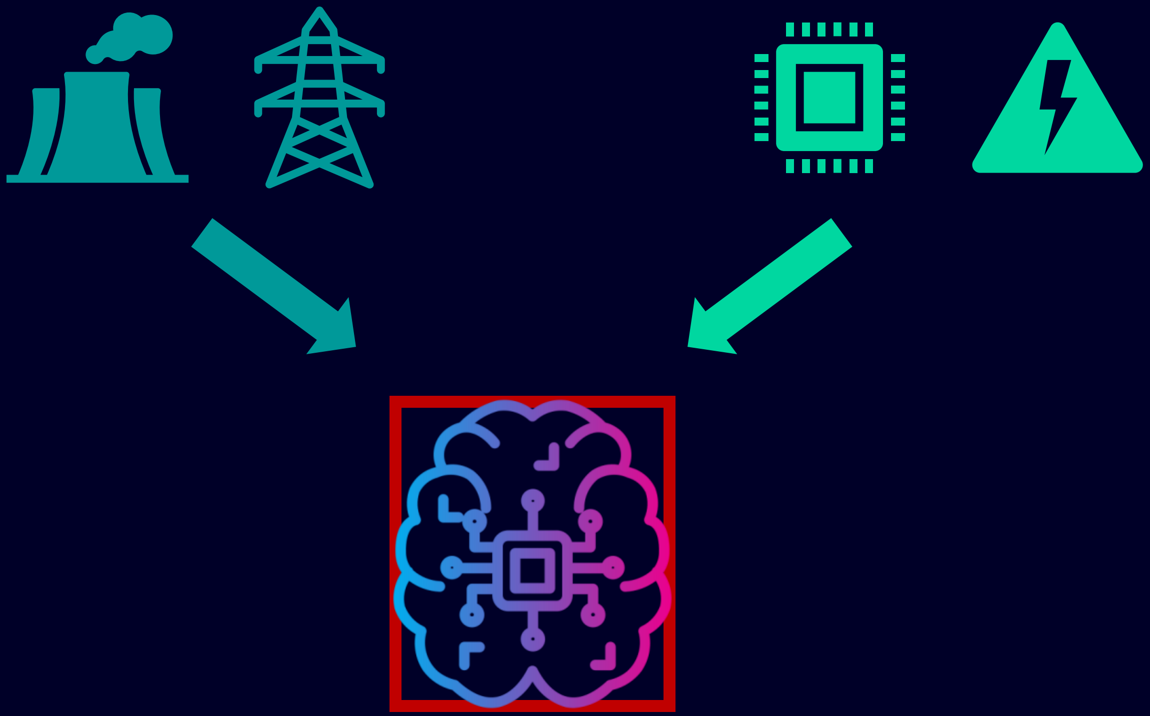
"

Powerful new AI models requires more energy consumption, there is no way to get there without energy breakthrough.

- Sam Altman, OpenAI

On the current path AI applications will be “Energy Bound”

Power Infrastructure **Device physics**



> ChatGPT consumes a lot of energy in the process, up to 25 times more than a Google search.
>
> Additionally, a lot of water is also used in cooling for the servers that run all that software. Per conversation of about 20 to 50 queries, half a liter of water evaporates – a small bottle, in other words
>
> \- Brussels Times

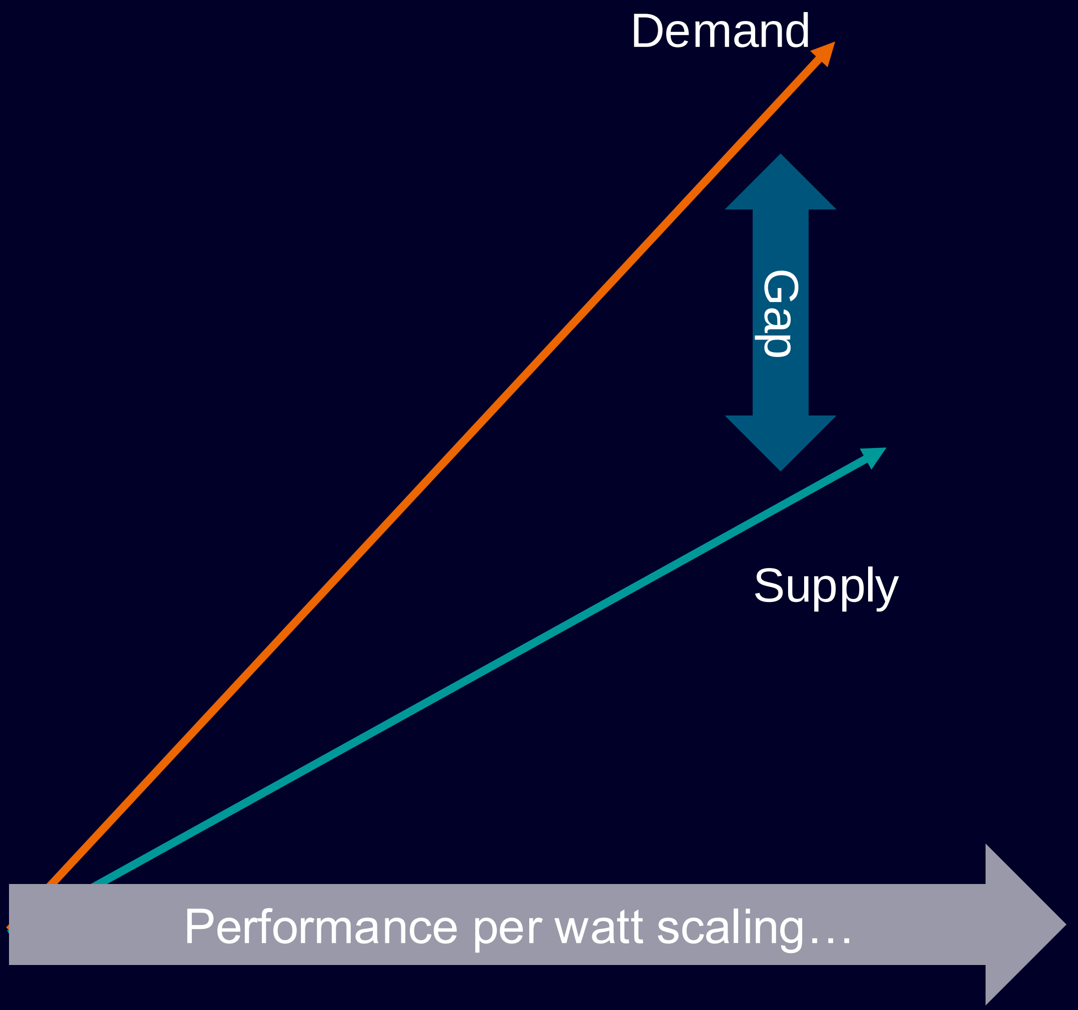
Power limits technology adoption, limits industry growth, and impacts corporate sustainability goals

Performance per Watt is now the limiting factor

Compute **performance is growing slower than power demand** due to slowing node scaling

Double impact as **increased compute power requires increased cooling power** and resources

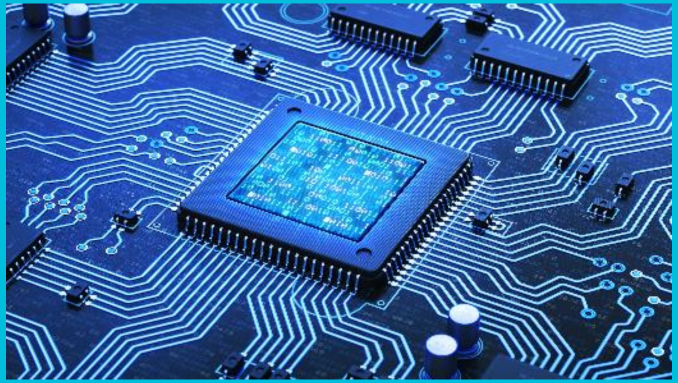
> Nvidia's upcoming Blackwell generation boosts power consumption even further, with the B200 consuming up to 1,200W, and the GB200 (which combines two B200 GPUs and one Grace CPU) expected to consume 2,700W. **This represents up to a 300% increase in power consumption across one generation of GPUs with AI systems increasing power consumption at a higher rate.**
>
> *-Forbes*



SIEMENS

Scope of potential solutions

Device



- Deliver more efficient semiconductor devices
- Better utilization of existing technology solutions

Product

- Reduce buffer and waste in design architecture
- Improved functional allocation between HW/SW, Electrical/Mechanical

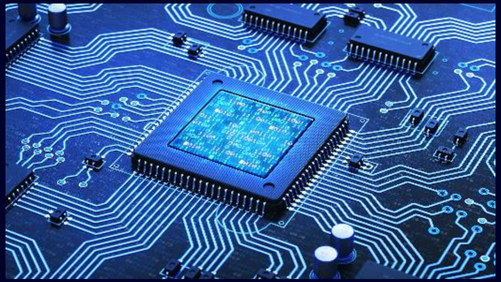
Lifecycle

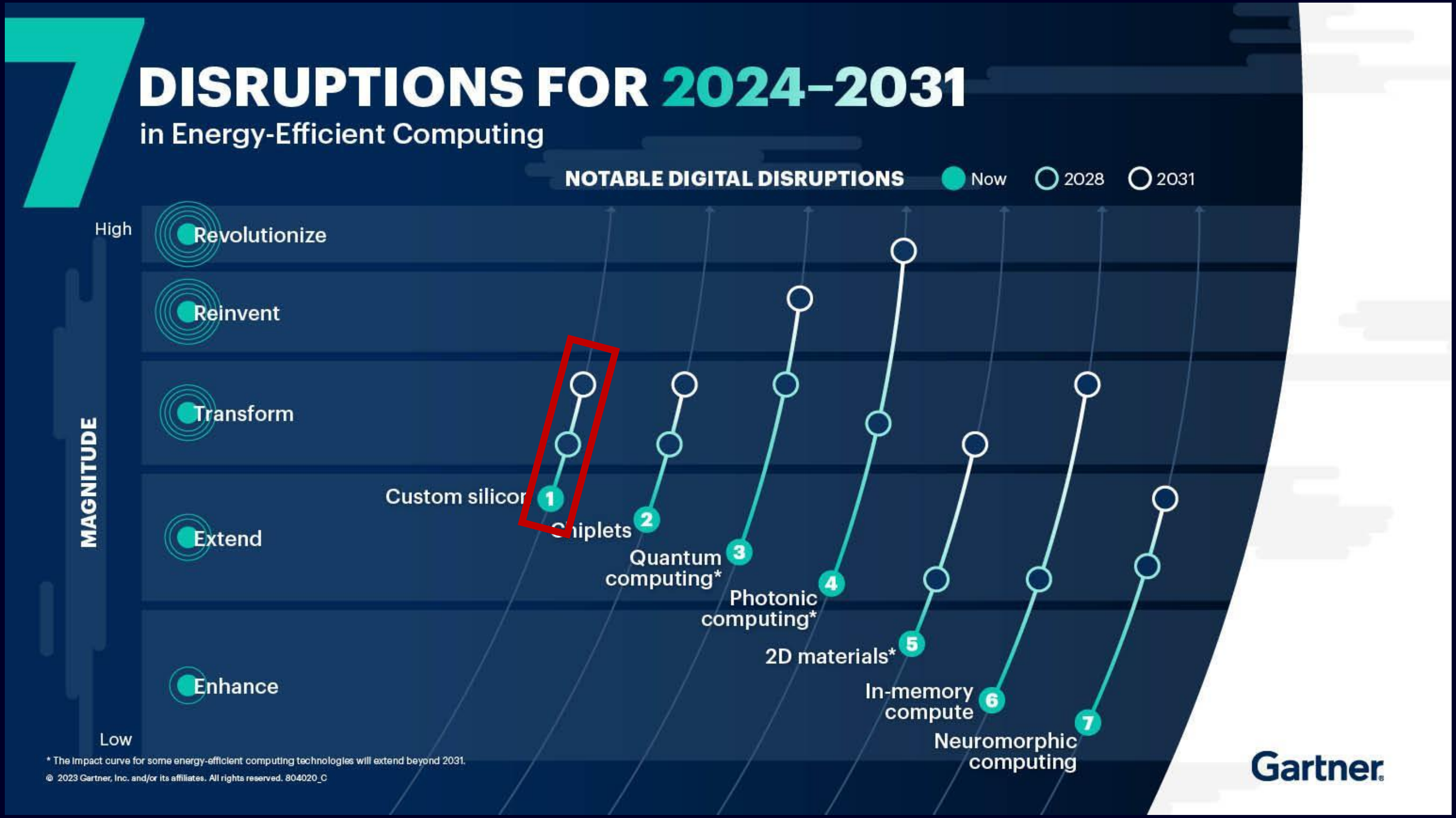
(concept – end-of-life)

- Impact of supply chain members
- Manufacturing resource utilization
- Recyclability and longevity

SIEMENS

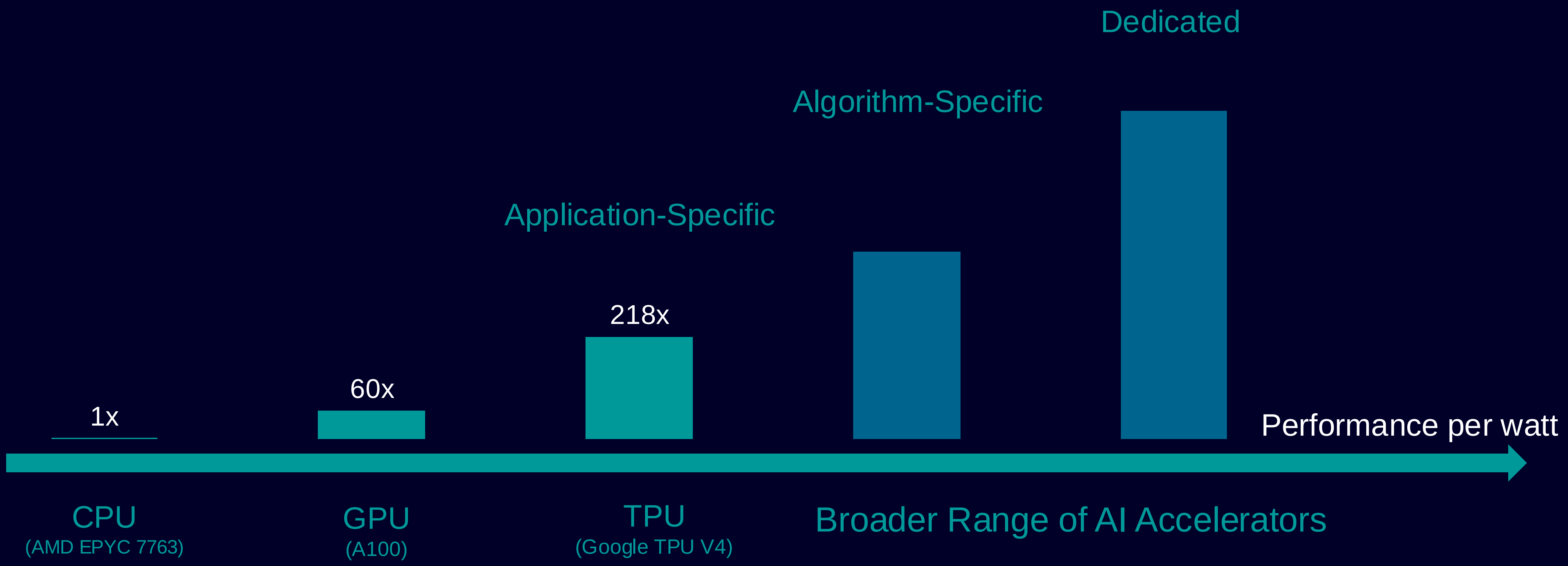
Addressing energy challenges while continuing growth of AI





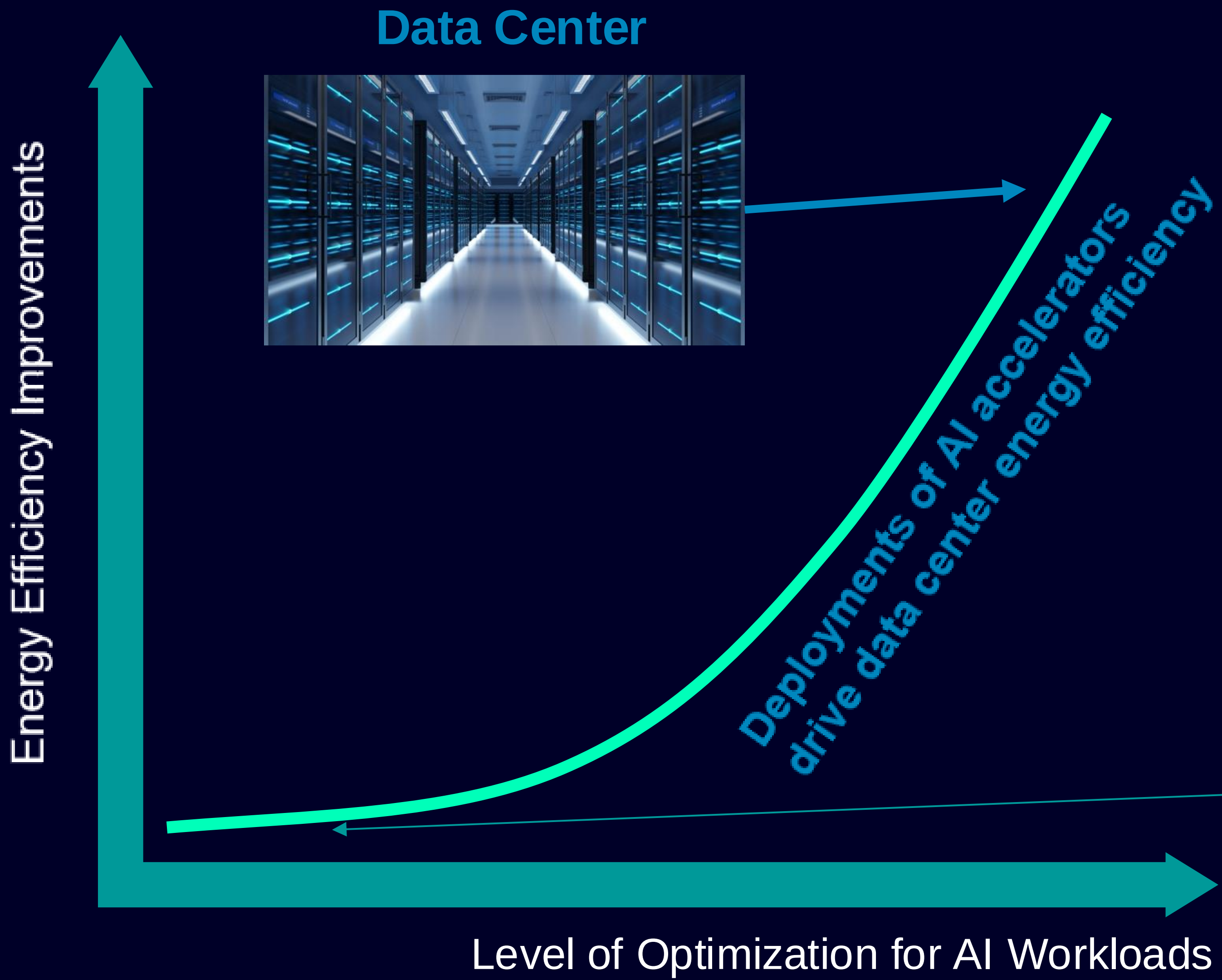
Workload Optimized Silicon

Workload-optimization leads to greater energy efficiency



SIEMENS

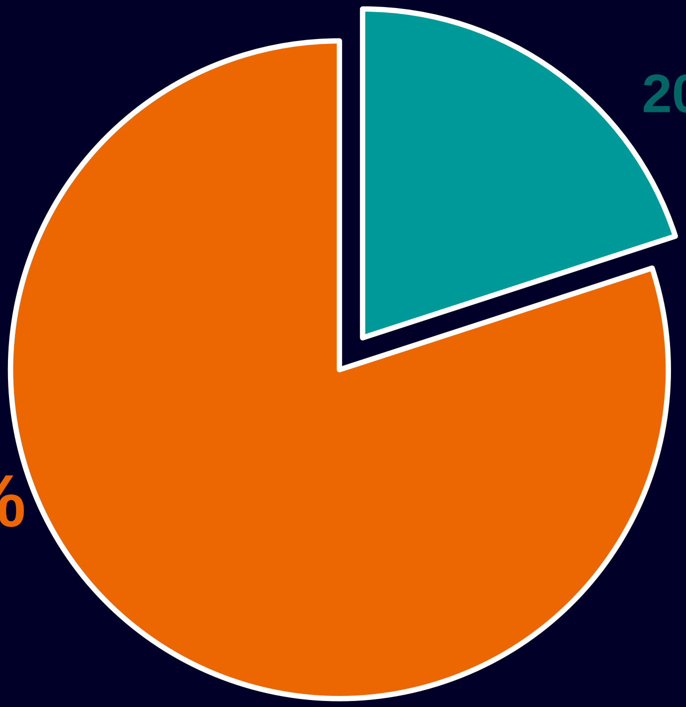
Evolving compute infrastructure: data center to edge for AI applications



Edge

SIEMENS

Transformation is needed on the edge



Environmental Footprint
(Source: World Economic Forum)

Training:
Centralized computing in the data centers

Inference:
Decentralized processing on the edge and KEY to a sustainable AI compute infrastructure

SIEMENS

However, designing workload-optimized hardware has challenges

Unique design requirements

Complex & lengthy processes

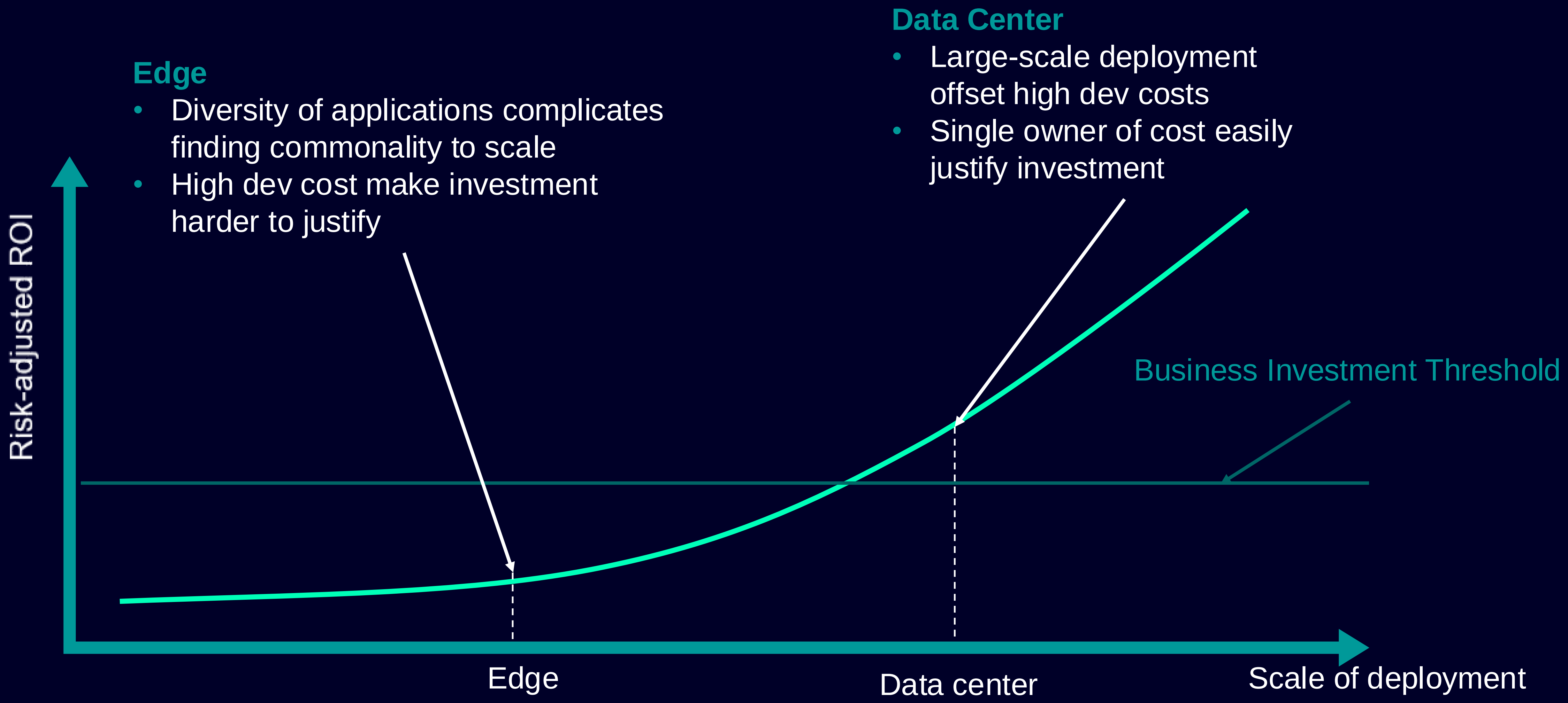
Limited reusability

Long development cycle

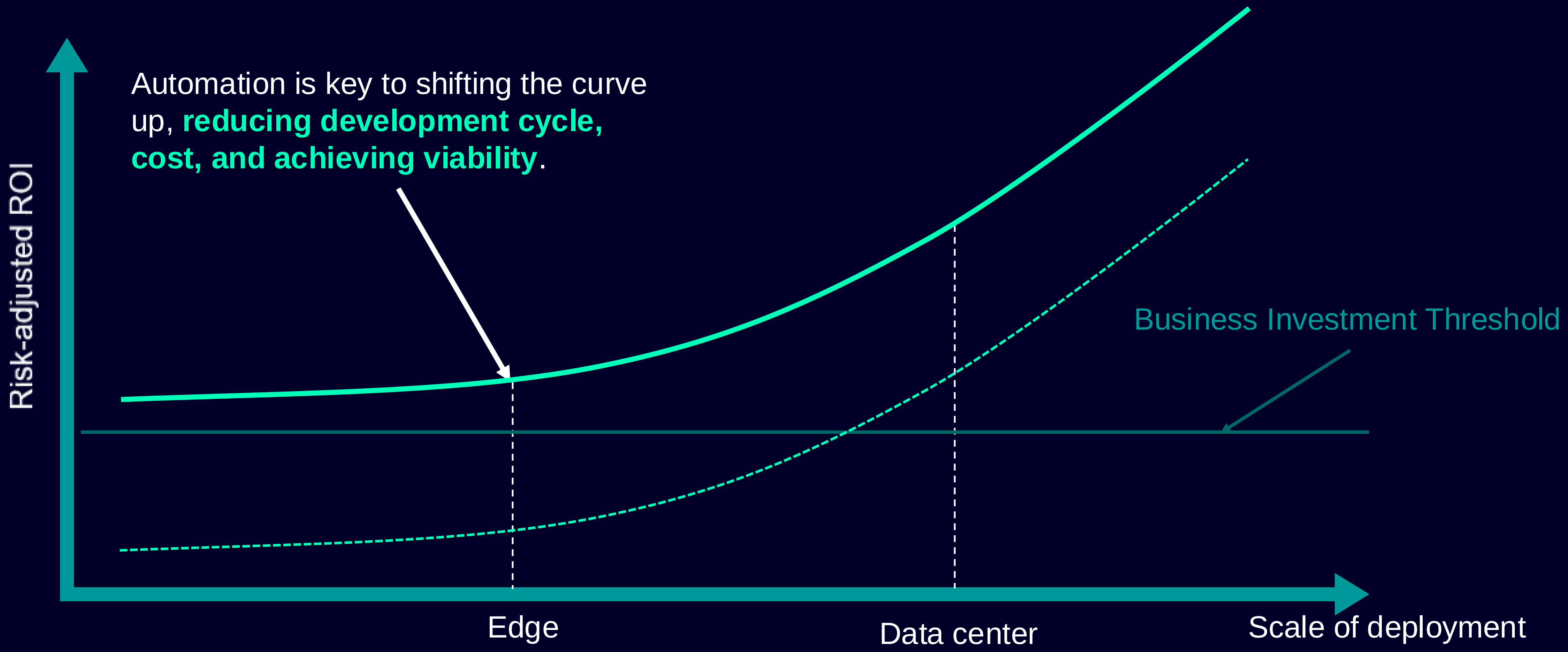
High development cost

Using traditional approaches optimizing devices for their unique workloads has increased overall business risk

Optimization traditionally favors centralized resources



Making workload-optimized hardware on the edge viable



Higher abstraction accelerates product development

Level of Abstraction

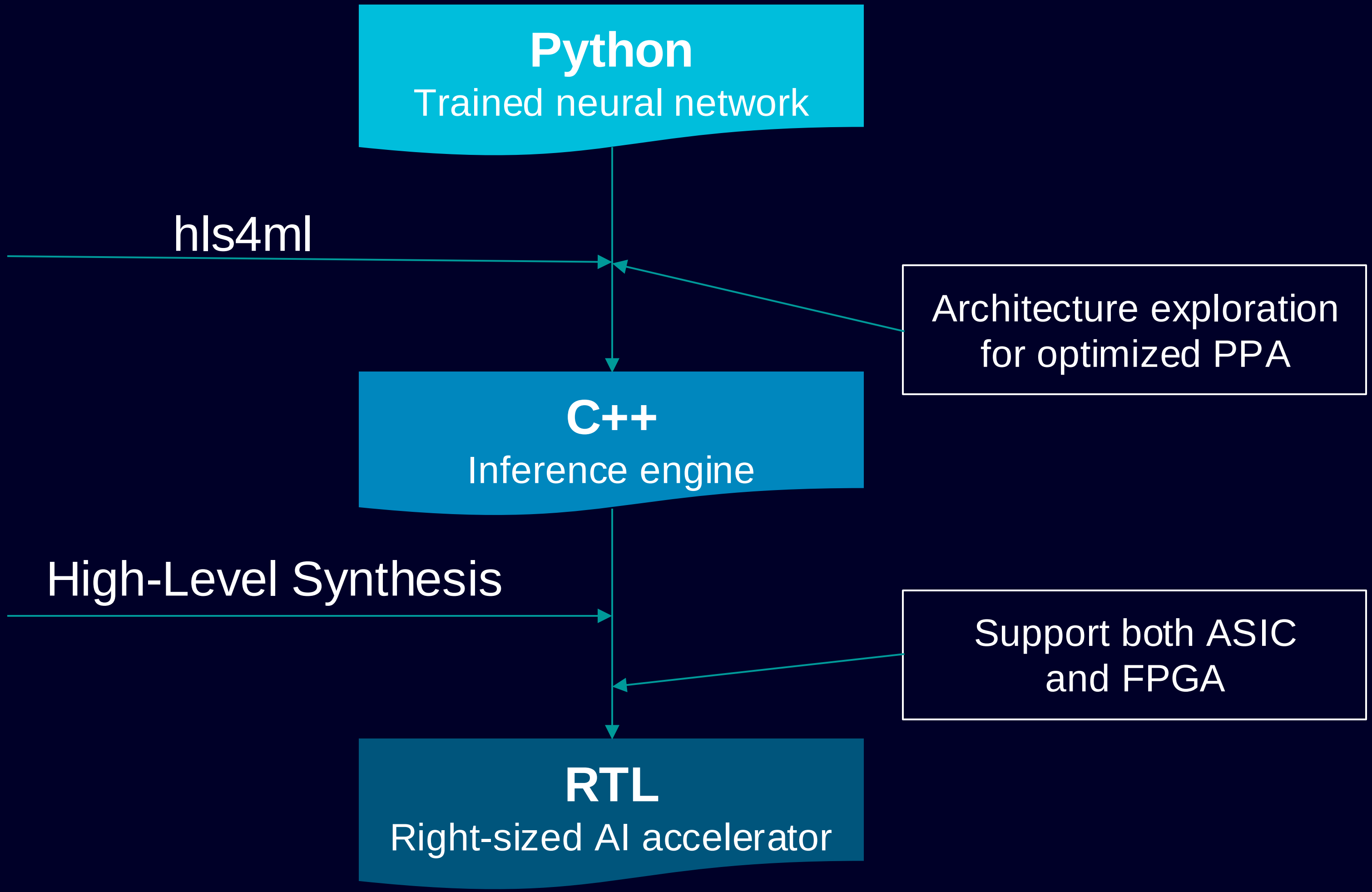
Prompt Level Design requirements in text	High-Level Synthesis with GenAI	Days
Application Level AI frameworks in Python	High-Level Synthesis for AI	Weeks
Algorithm Level Algorithm in SystemC/C	High-Level Synthesis	Months
Register Transfer Level Hardware representation in HDL	RTL Synthesis	Years

Accessibility

Catapult AI NN: automation from AI network to semiconductor implementation

Automated Translation from High-level AI Framework

Optimized Design Workflow on the Application Level



Wakeword inference: a new use case unlocked with Catapult AI NN

The Design

Audio processing system with trained models for inference

The Challenge

High latency unable to meet real-time performance

High energy consumption due to AI algorithm in design

The Solution

Create accelerators with HLS to execute AI algorithm

Evaluated design configurations to balance PPA and Accuracy

90%

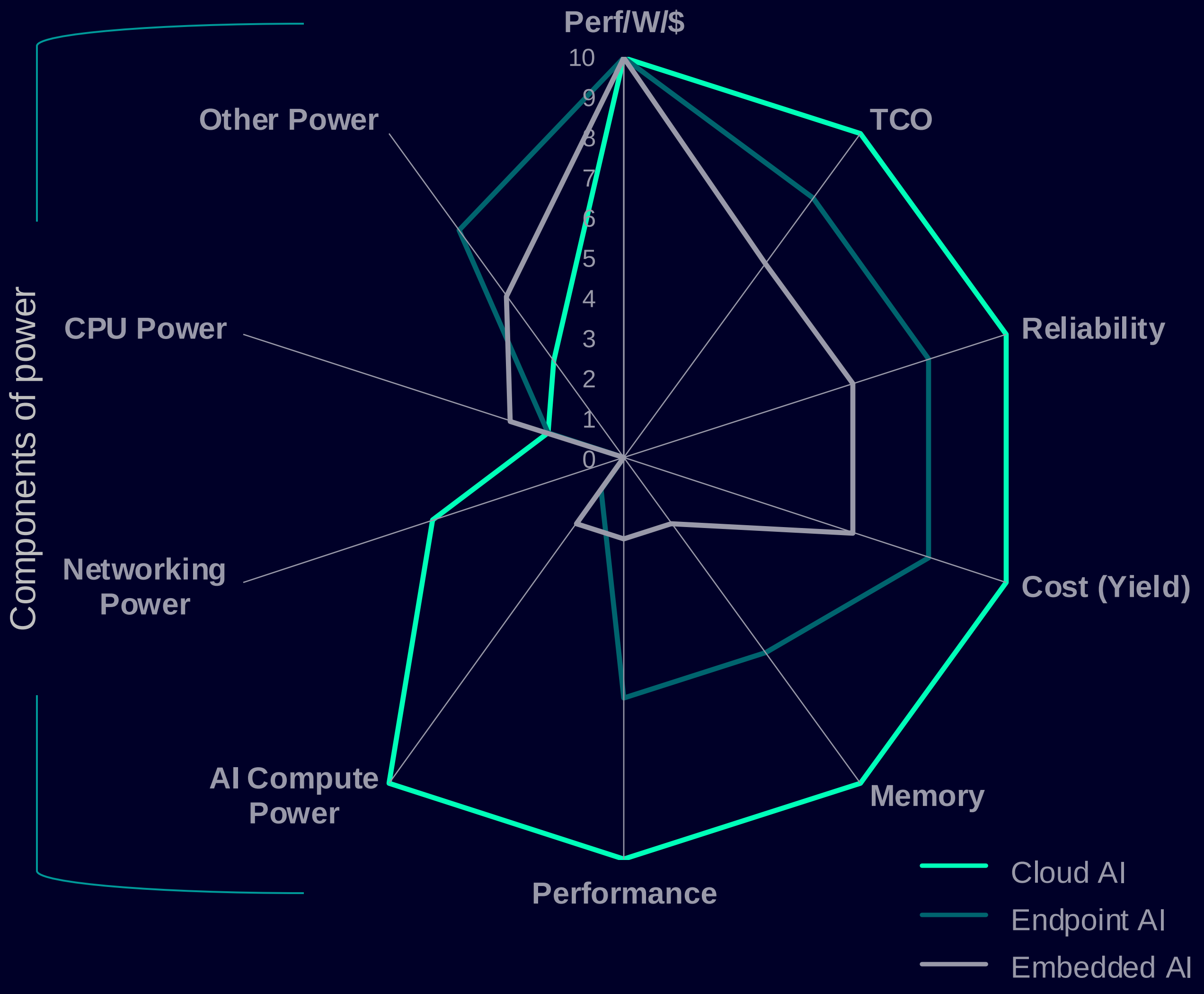
Reduction in energy per inference

15X

Faster execution time

Configuration	Speed msecs	Power mW	Area Sq micron	Energy/Inf uJ
RISC-V Software	285	189	393,521	33,577
Software + Accelerators	18	203	722,521	3,683

Recognizing your top careabouts…



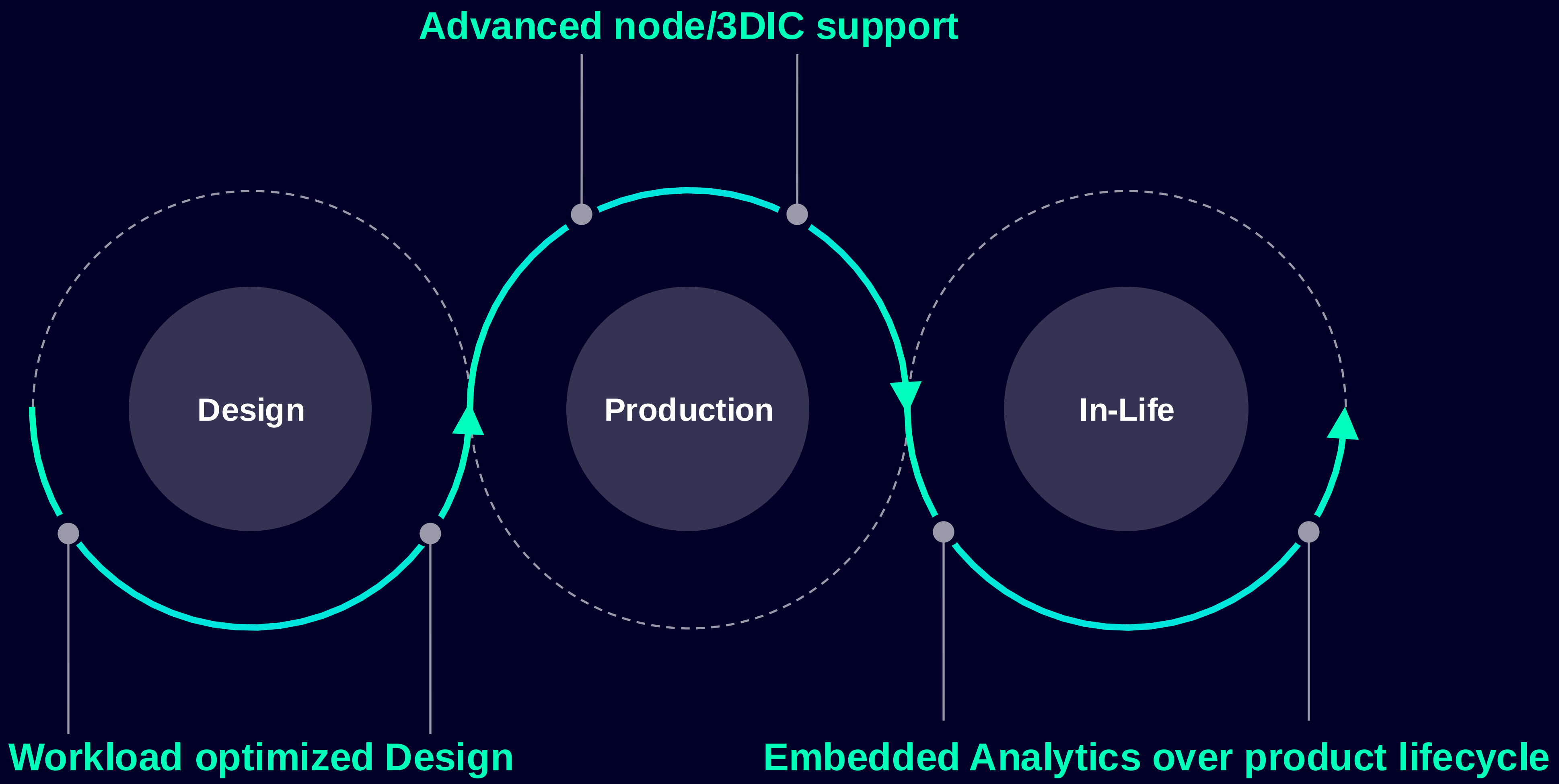
Cloud AI (large-scale training, inference)
- Total datacenter power limits model size
- Network power contribution is significant

Endpoint AI (personal assistants, games, copilot)
- Trade-off between user experience and battery life

Embedded/Edge AI (automotive, robotics, medical)
- Total power envelope highly constrained

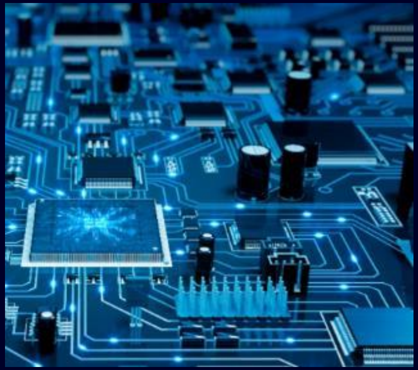
PPA optimized C-to-GDS solution
Lifecycle Management to deliver best perf/TCO

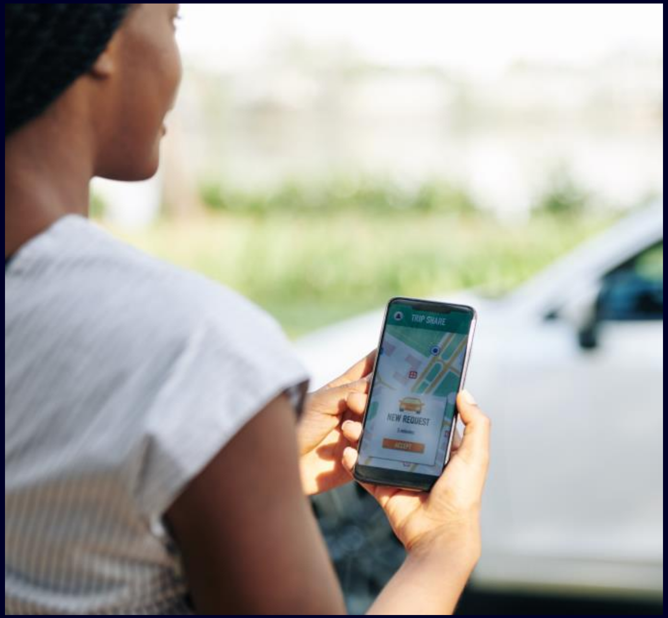
… and addressing challenges throughout the lifecycle



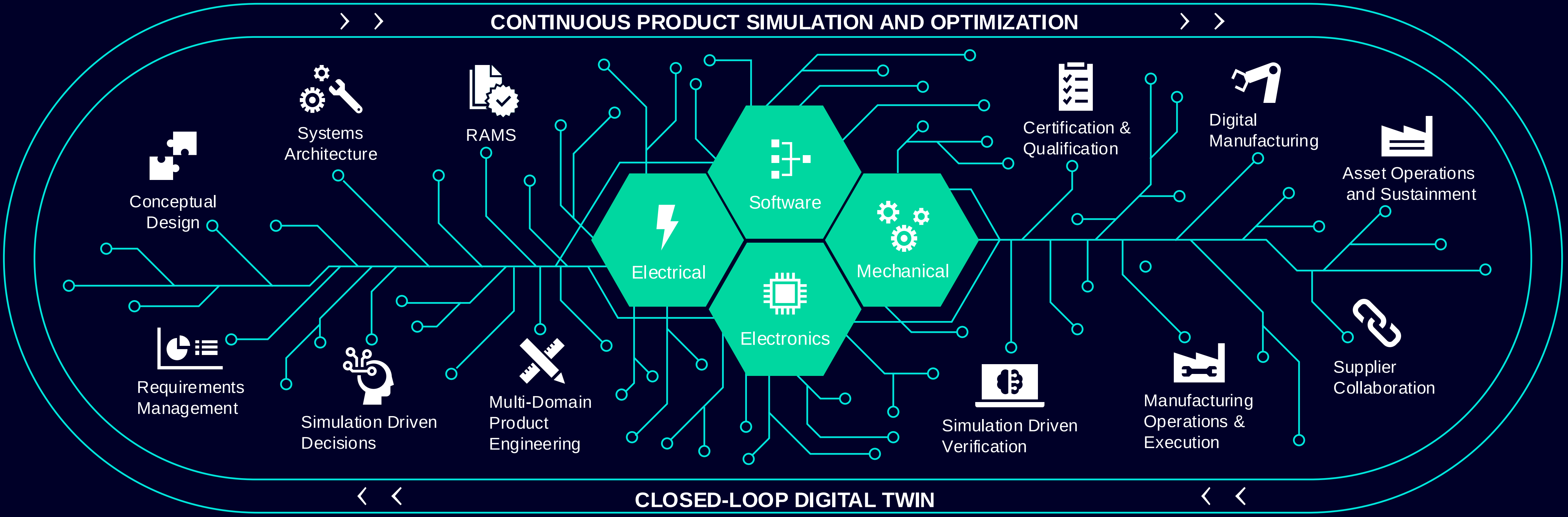
But Semiconductors are just part of a more complex system





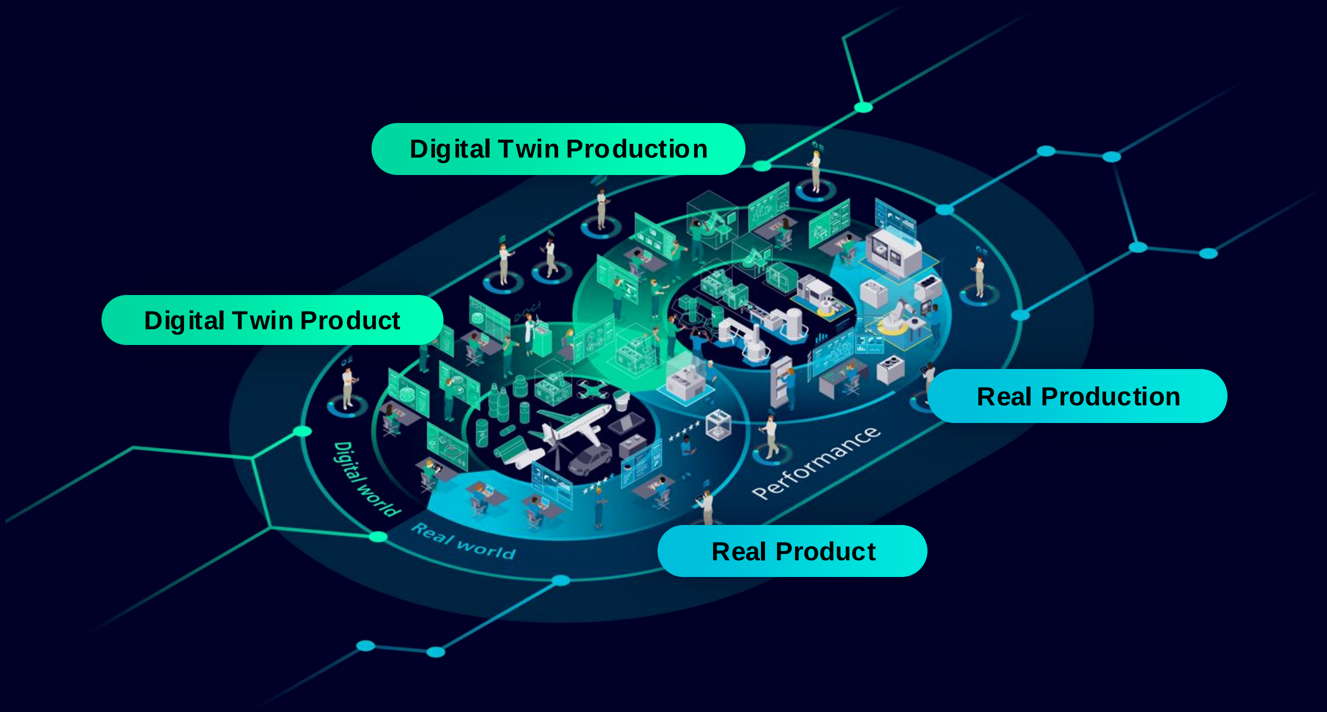


Solving the challenges for Software Defined Products



SIEMENS

Which requires a Comprehensive Digital Twin



Understand design and manufacturing considerations early in the design process

Perform design trade-offs with visibility into the implications of any change

Make decisions with a more holistic view of the product lifecycle

SIEMENS

Optimizing edge AI Infrastructure drives growth and scales sustainability

Energy Efficiency

Scalability of systems

Resource Efficiency

Sustainable manufacturing

Infrastructure Efficiency

Lifecycle solutions

Siemens offers solutions enabling a sustainable computing future

Contact

Published by Siemens 2024

Ankur Gupta
SVP and GM Digital Design Creation Platform (DDCP)
Siemens EDA

ankurgupta@siemens.com