

Passive Components Challenges in AI Systems

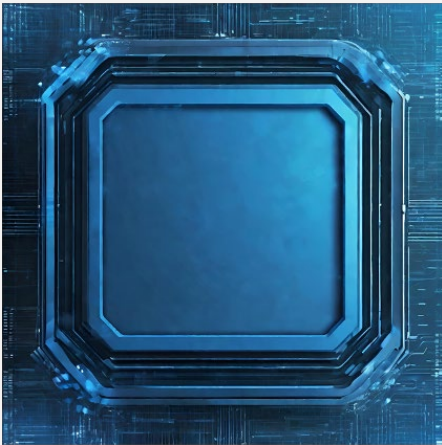


European
Passive Components
Institute

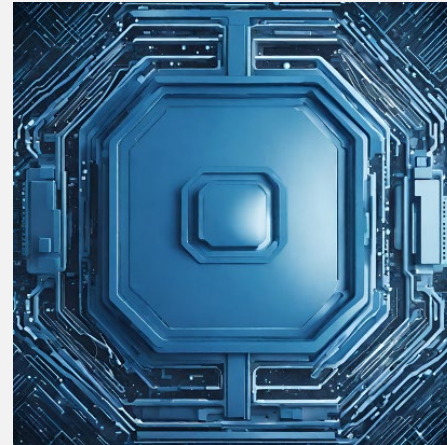
www.passive-components.eu

Tomáš Zedníček Ph.D.

Outline



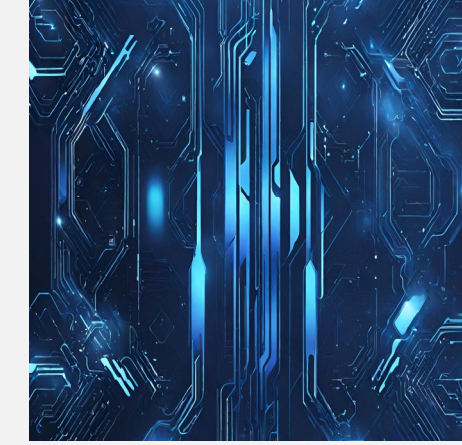
INTRODUCTION



AI POWER
MANAGEMENT



PASSIVE COMPONENTS
CONSEQUENCES

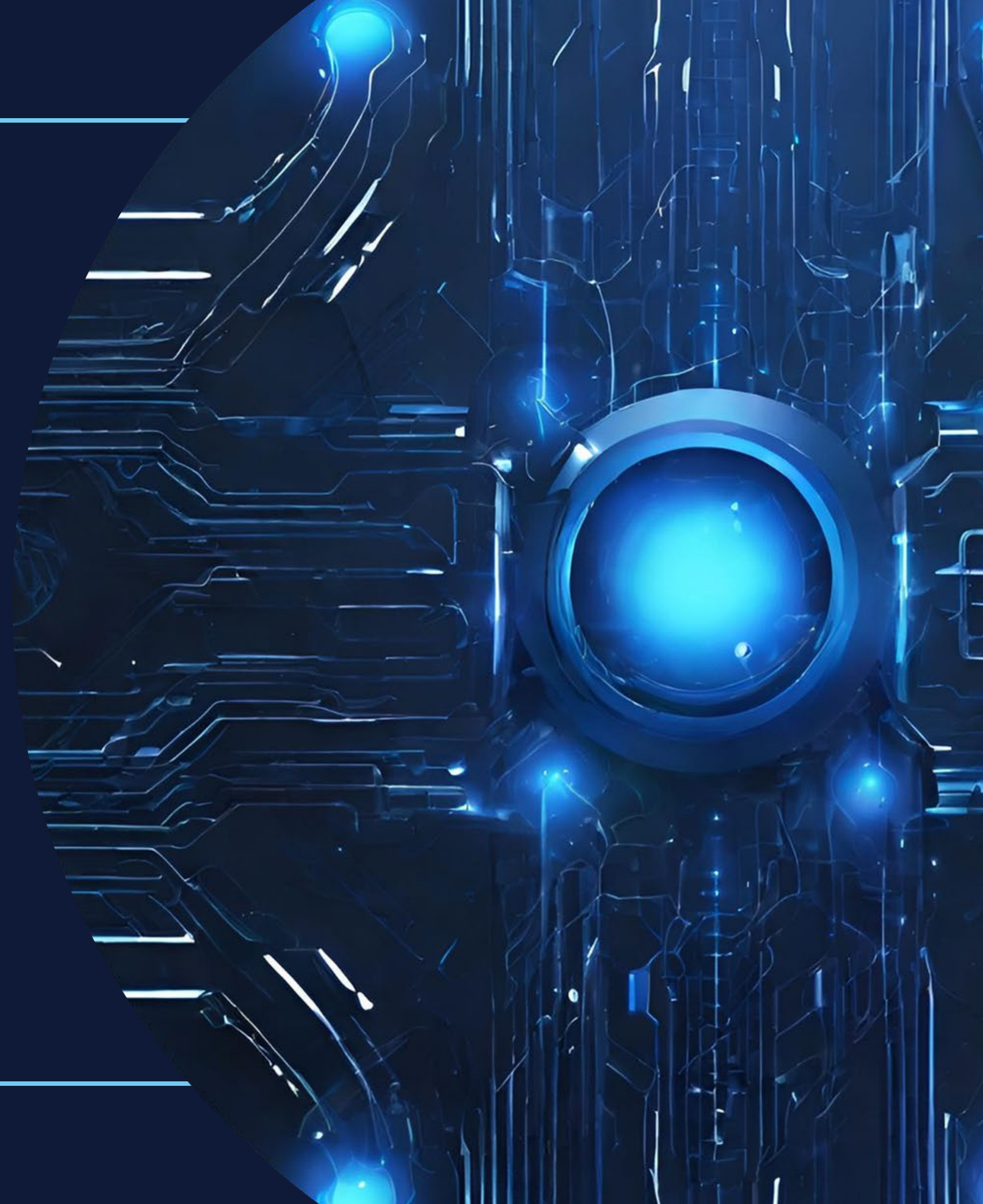


BEHIND
HORIZON

1

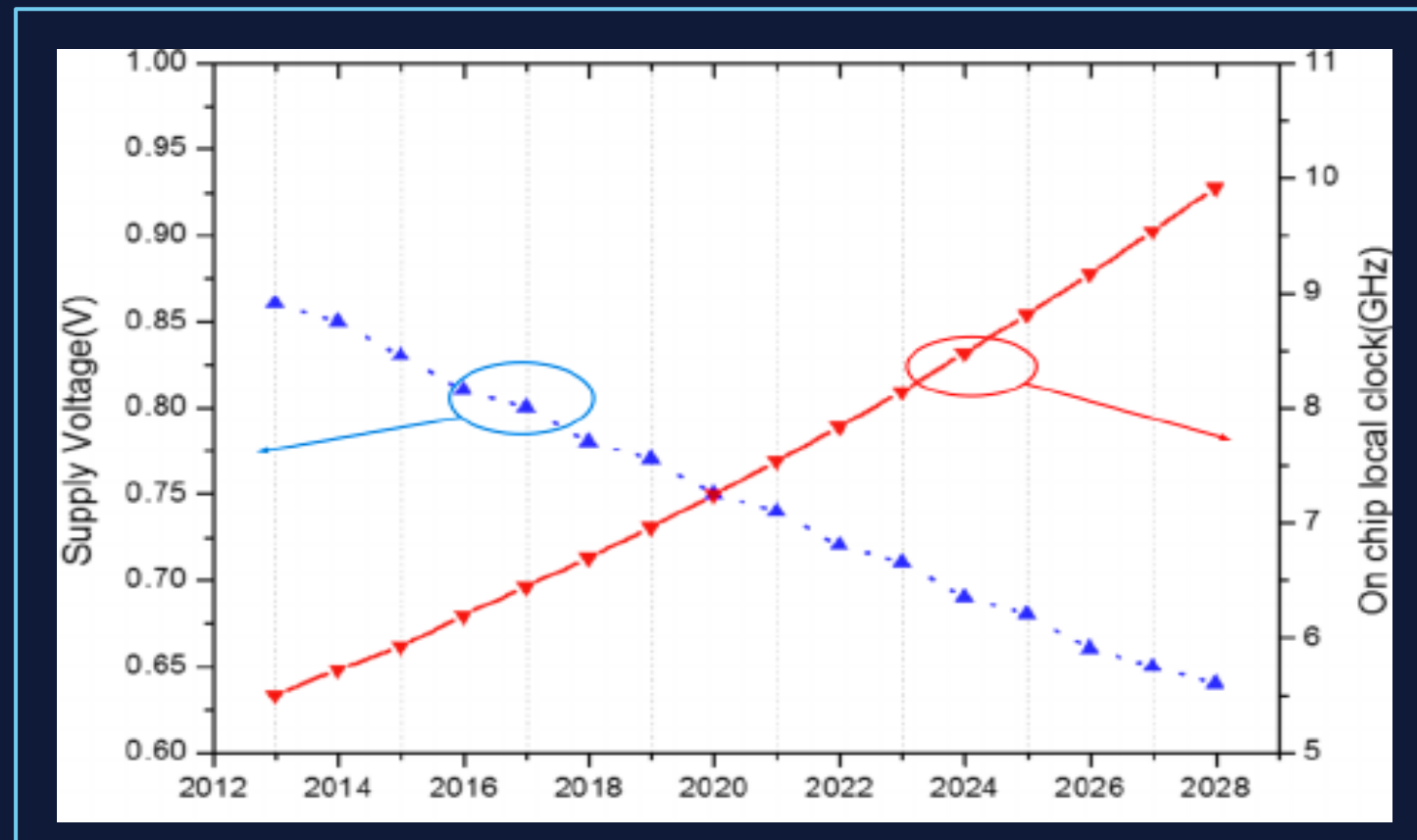
Introduction

IC Technology Background



Semiconductor IC Development – Processors

DIE SCALING HAS DROPPED IC SUPPLY VOLTAGE



Best Fit Mass Volume Capacitor Technology:

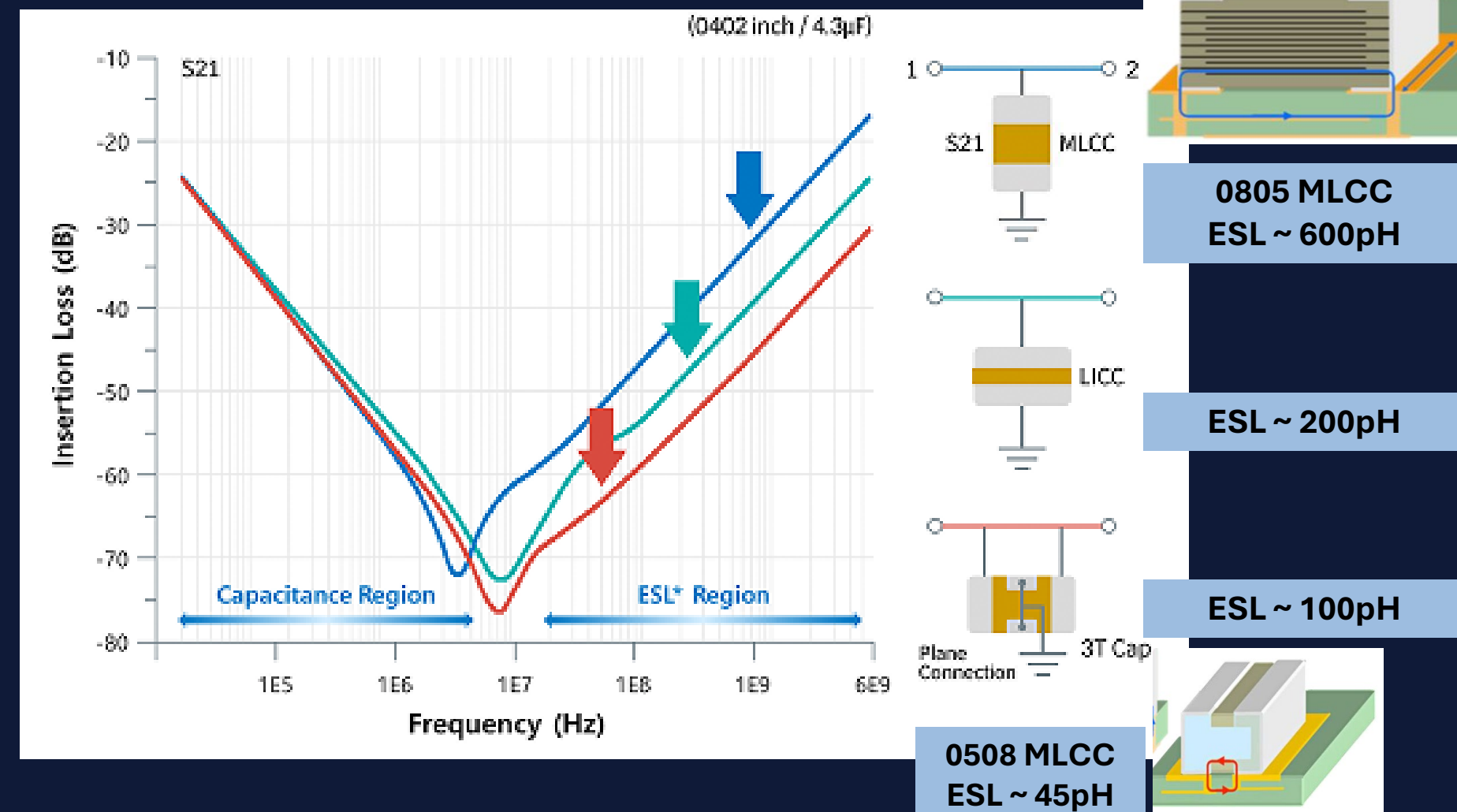
Past: Electrolytics + MLCC
Current: MLCC / Silicon Capacitors
Future: Integrated on Chip

Source: ITSC, KYOCERA AVX, Samsung

- *Capacitors job decoupling more critical*
- *Clock & data speeds making di/dt drawn larger*

Coupling Capacitor Consequences

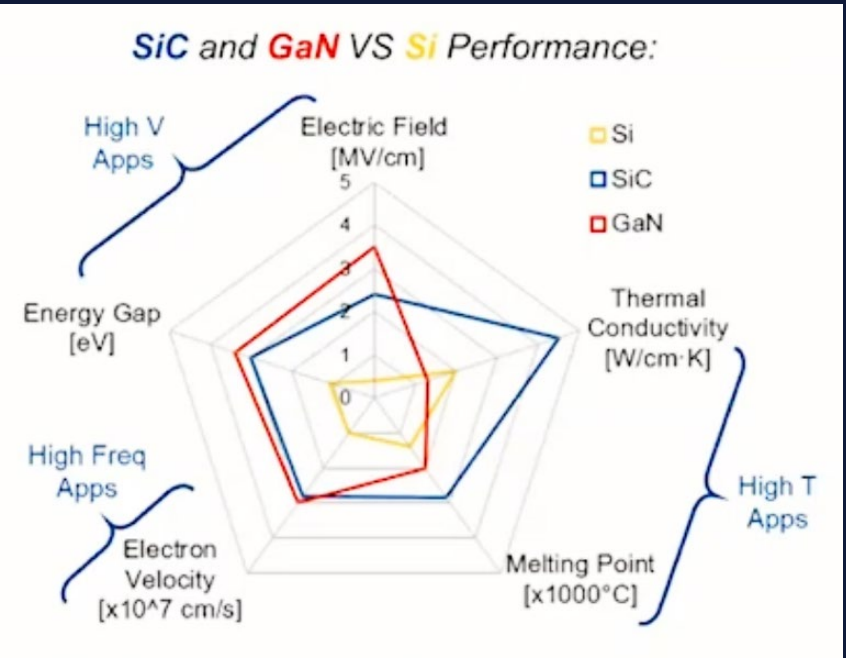
- Lower ESL
- Lower ESR
- Higher power
- Smaller Size
- Lower Profile
- Higher Precision
- EMI/noise issues



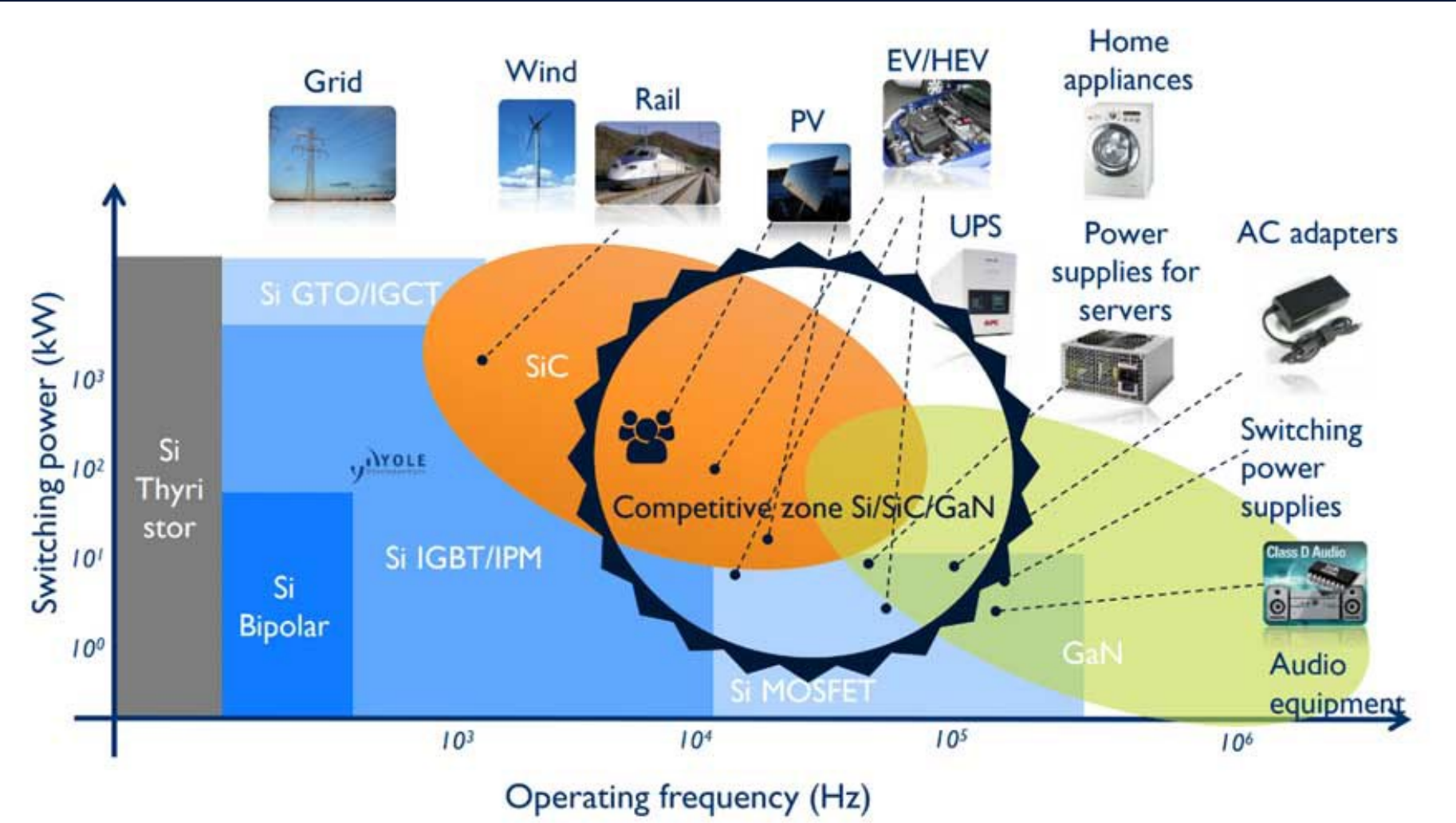
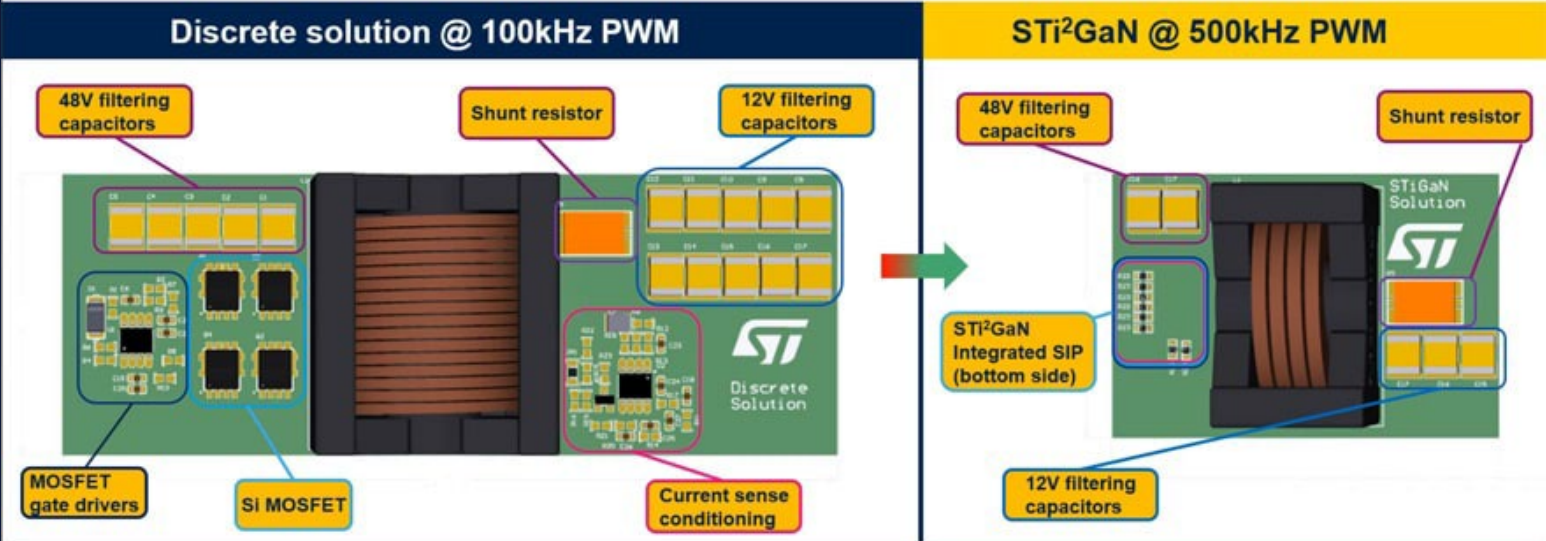
Key Growth Areas

Semiconductor IC Development Background

Semiconductor IC Development – Wide Gap GaN/SiC „Revolution“

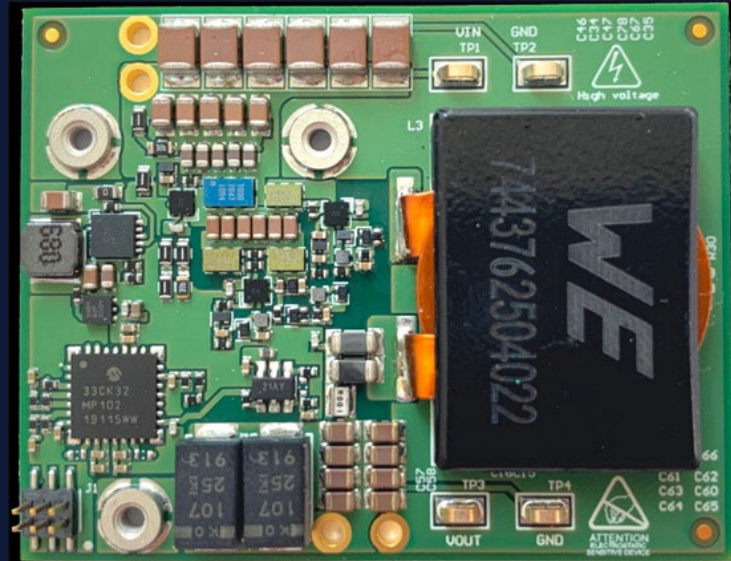


Size matters
example 12V/48V DCDC STi2GaN vs Si discrete



Source: Yole

Passive Component Changes:
Selection Changes
High Frequency,
Higher Temp
High Power
Low Loss



48 V three-stage synchronous buck converter with GaN technology

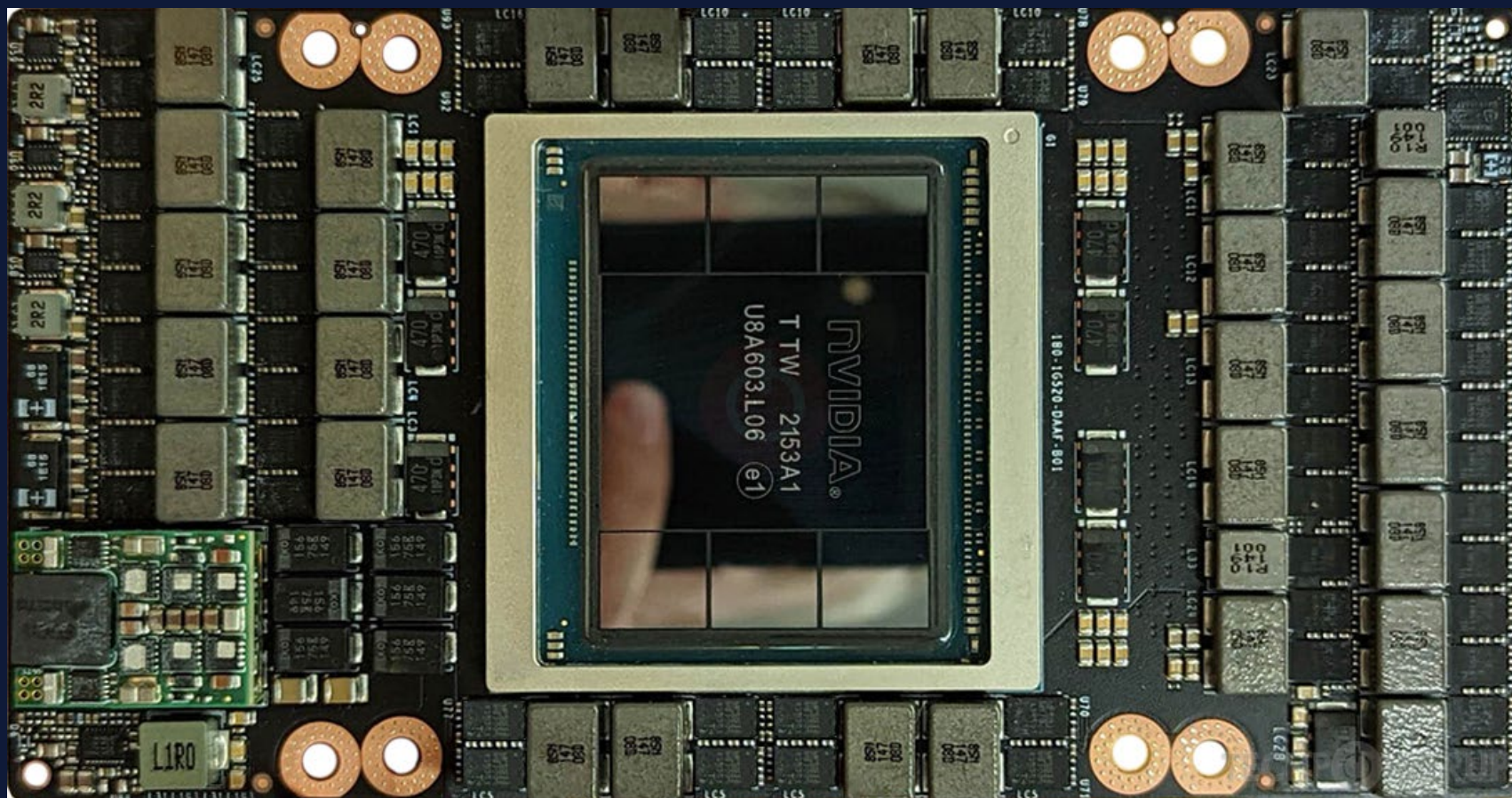
2

AI Power

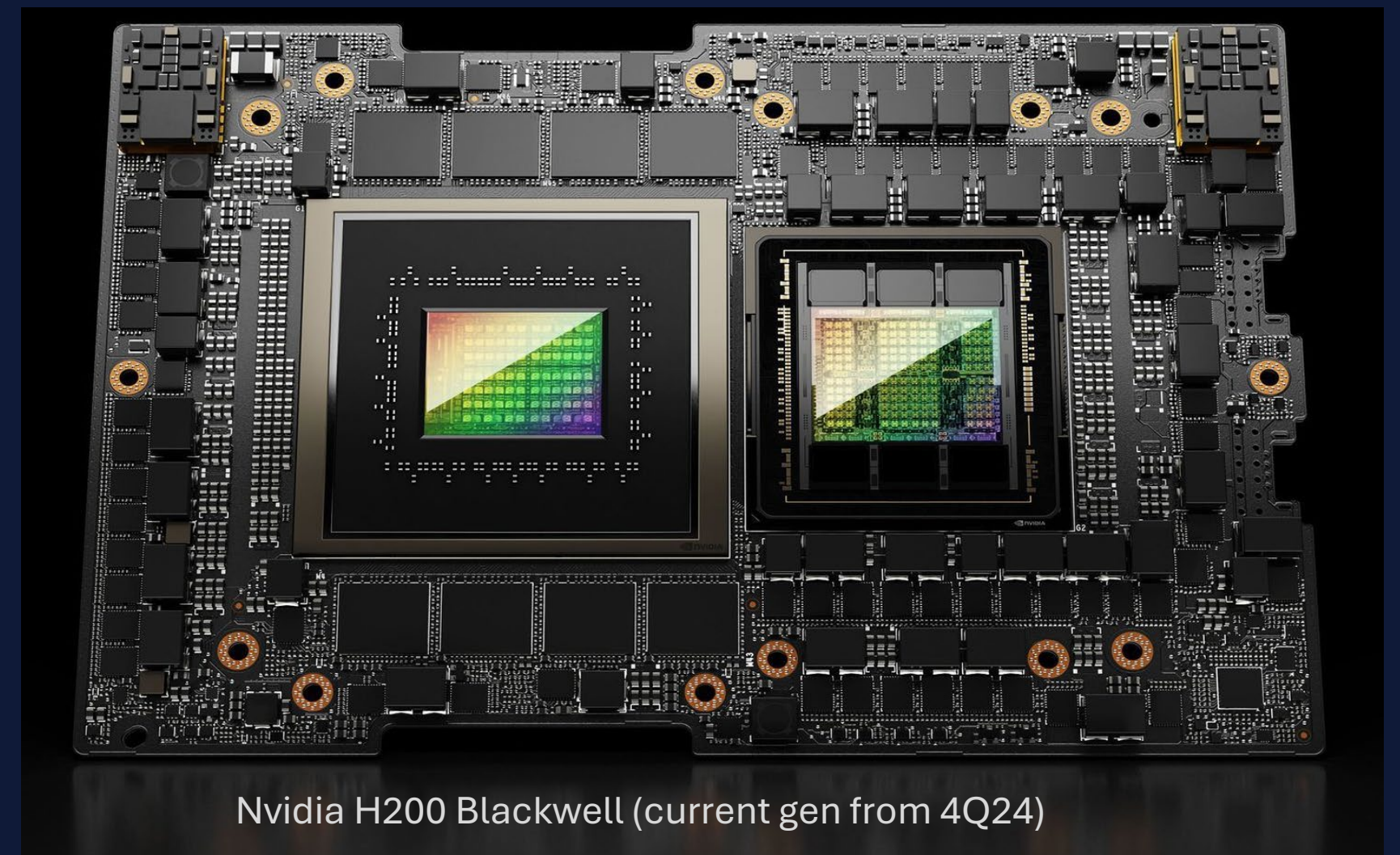
AI Energy & Power
Management Challenges



- AI demand stays very strong through CY25; bottlenecks have shifted from GPUs to datacenter power capacity and floor space.
- Nvidia H100 lead times ~10 weeks; H200 Blackwell (B200) shipping since C3Q24, tight supply into mid-2025 due to advanced packaging limits.
- Blackwell Ultra (B300): ramps in C2H25, ~50 % more compute power.
- Vera Rubin (VR200): launches C2H26 with ~3.3× B300 compute performance.
- Datacenter constraints: rack densities > 140 kW drive adoption of liquid/immersion cooling; some hyperscalers adding dedicated generation (incl. SMRs) to bypass grid limits.
- Cloud vs. on-device AI: cloud growth remains robust; the long-term compute split is still evolving.



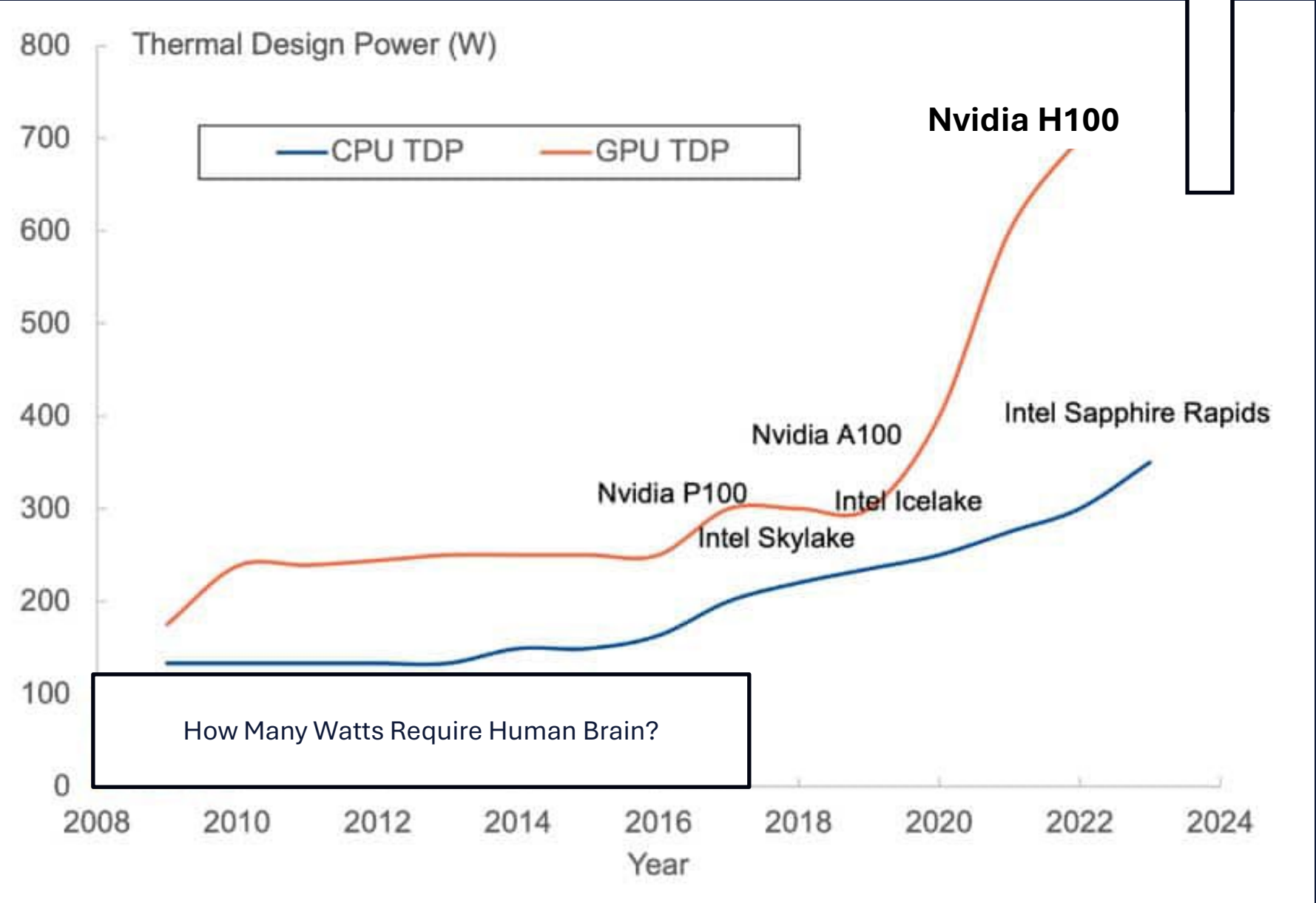
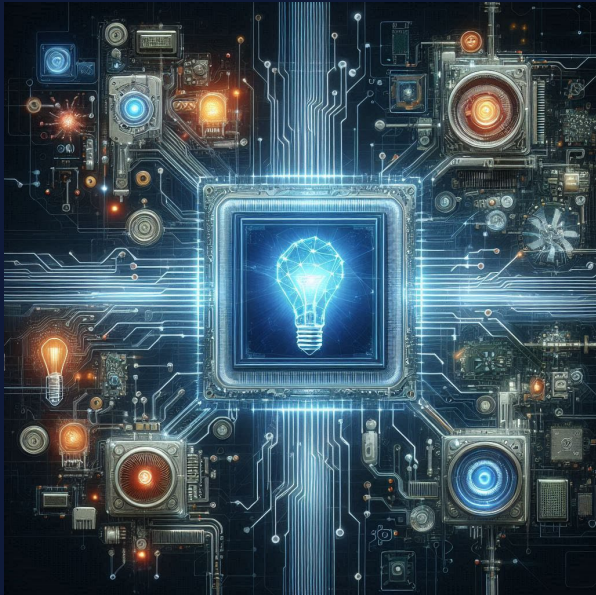
Nvidia AI Cloud GPU H100 (first gen)



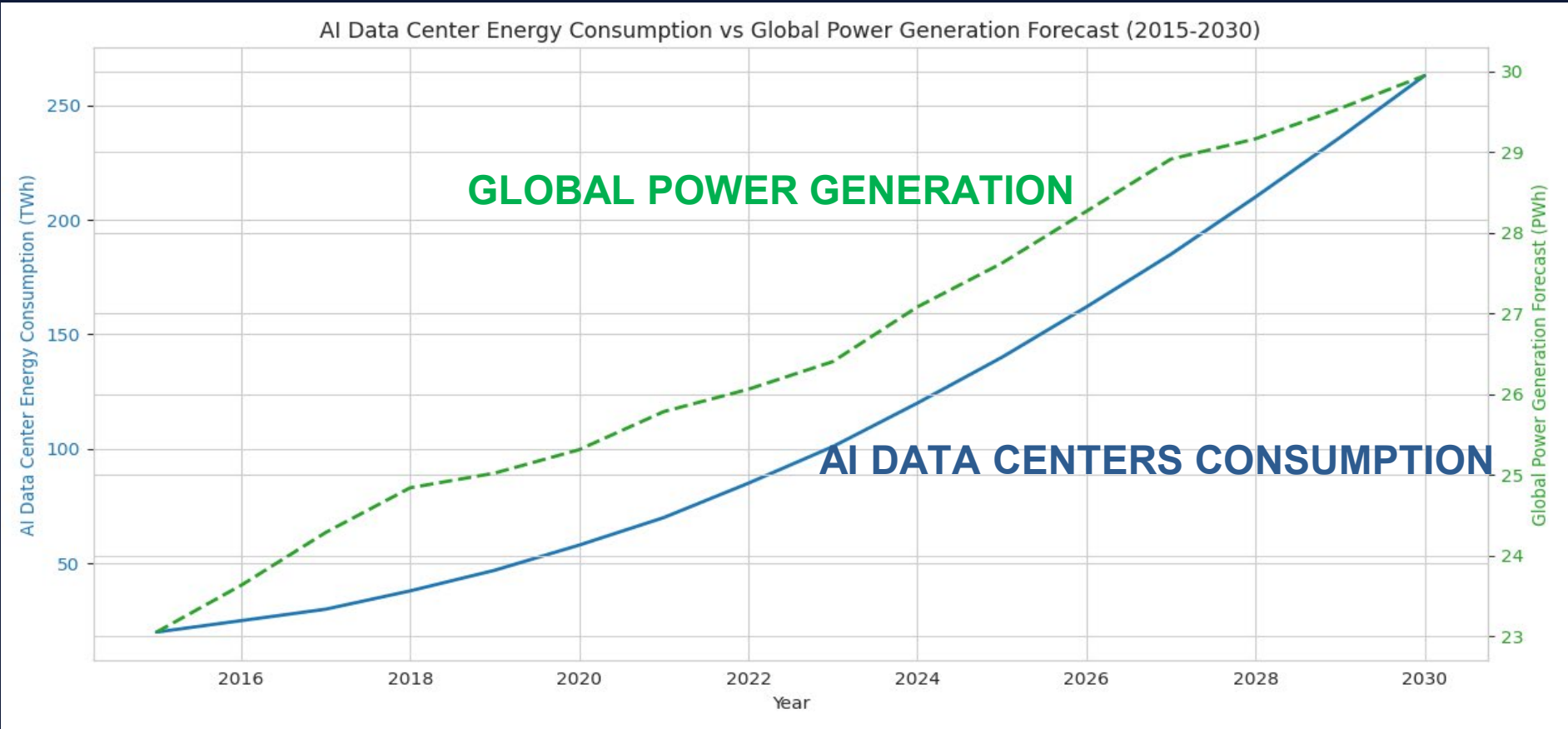
Nvidia H200 Blackwell (current gen from 4Q24)

1200W Nvidia H200

since Q3CY24



The Energy Challenge of AI in Data Centers



source: statista.com

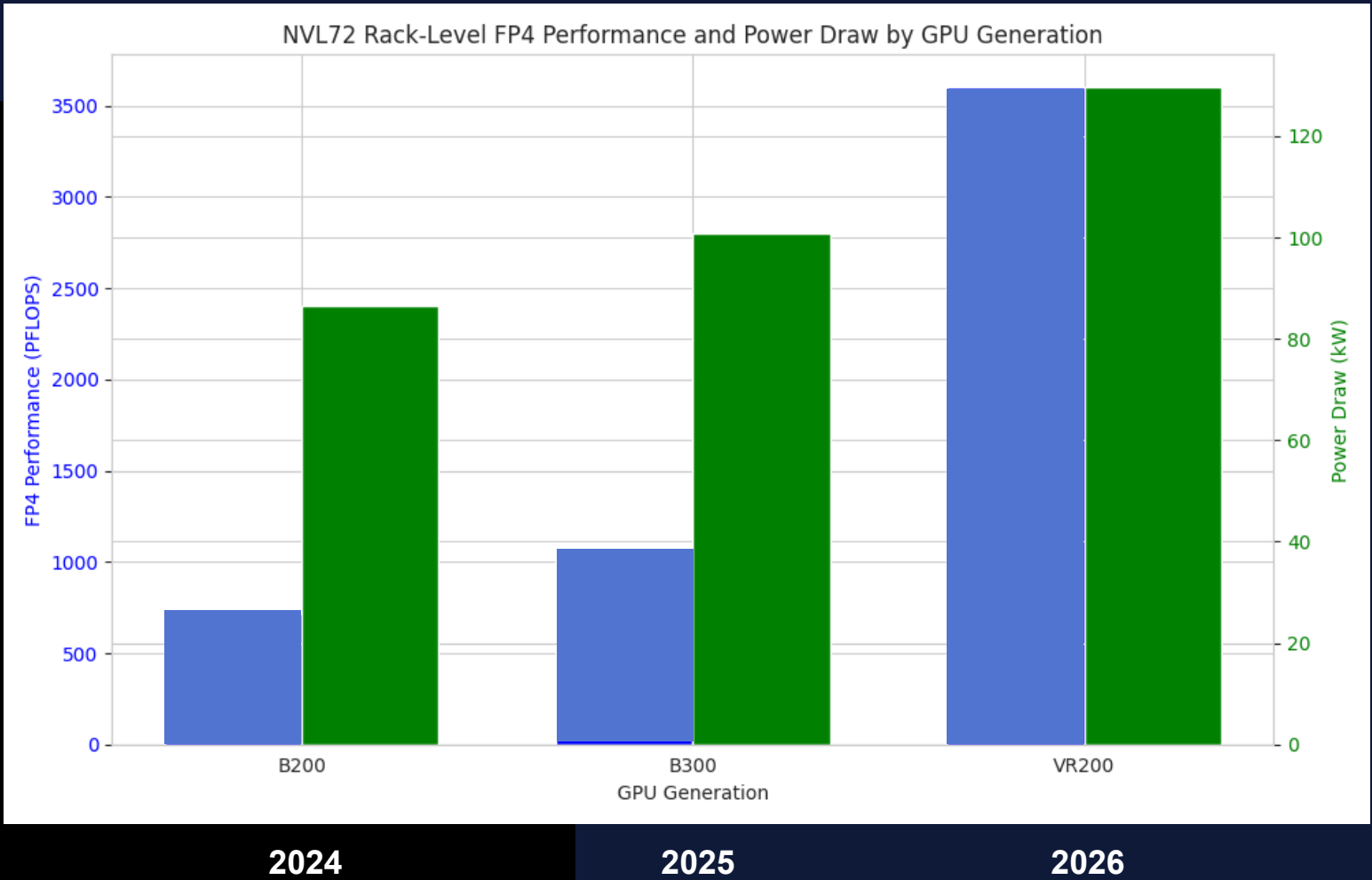
AI Server Performance vs Power Consumption 2024 - 2026

Notes:

- **PFLOPS** - PetaFLOPS = *peta* (10^{15}) floating-point operations per second.
A GPU rated at 15 PFLOPS **FP4** can perform 15 quadrillion FP4 operations every second under ideal conditions.
- **FP4** - Floating Point 4-bit, AI models often don't need full 32-bit or 16-bit precision to get accurate results. Using fewer bits speeds up computation, reduces memory use, and lowers power draw.

5x

1.5x

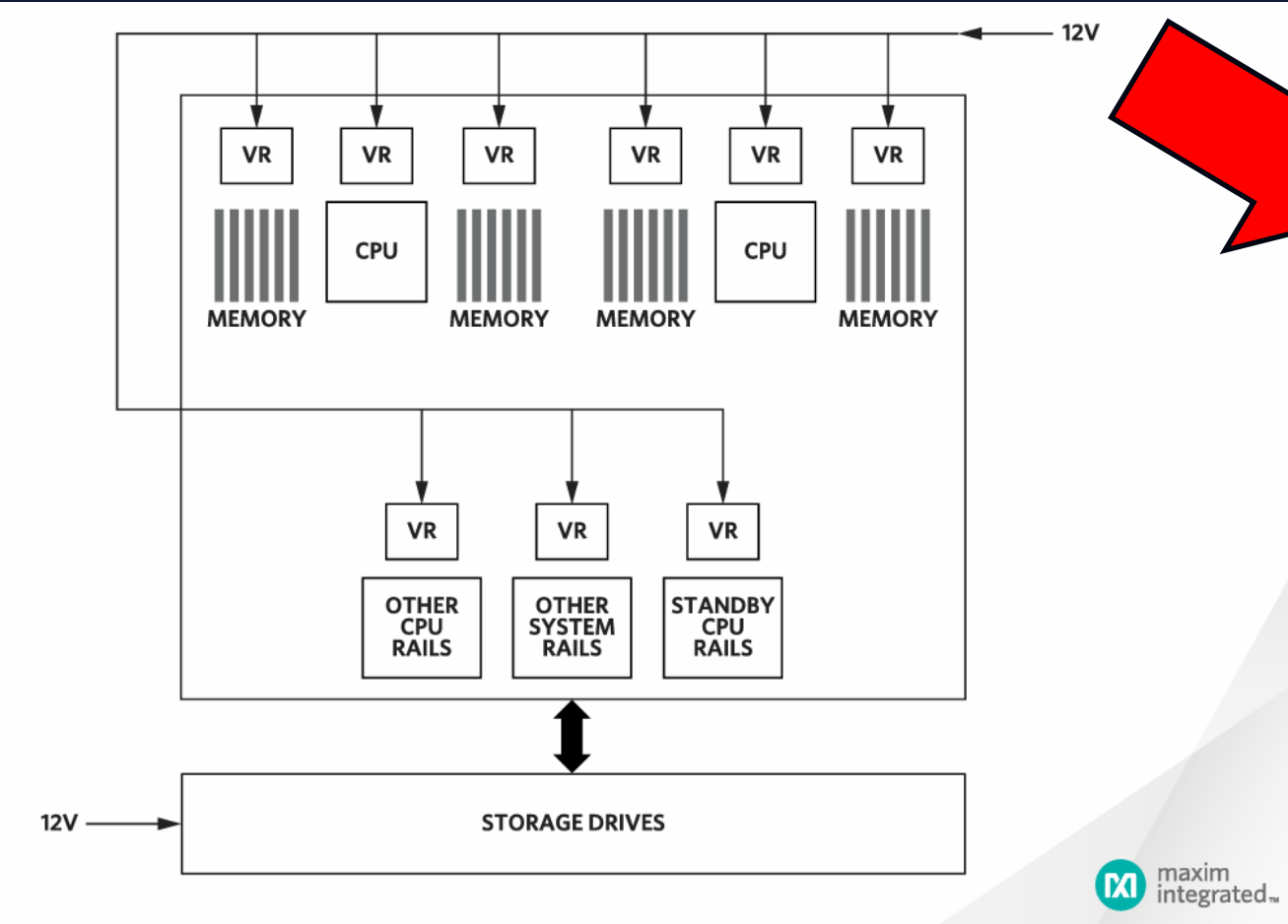


- GPU Power to reach **1800W** in 2026 (Vera Rubin)
- Rack level Power Consumption to reach **130kW** in 2026
- CPU and Memory consumes ~80% of power energy
- In two years from now

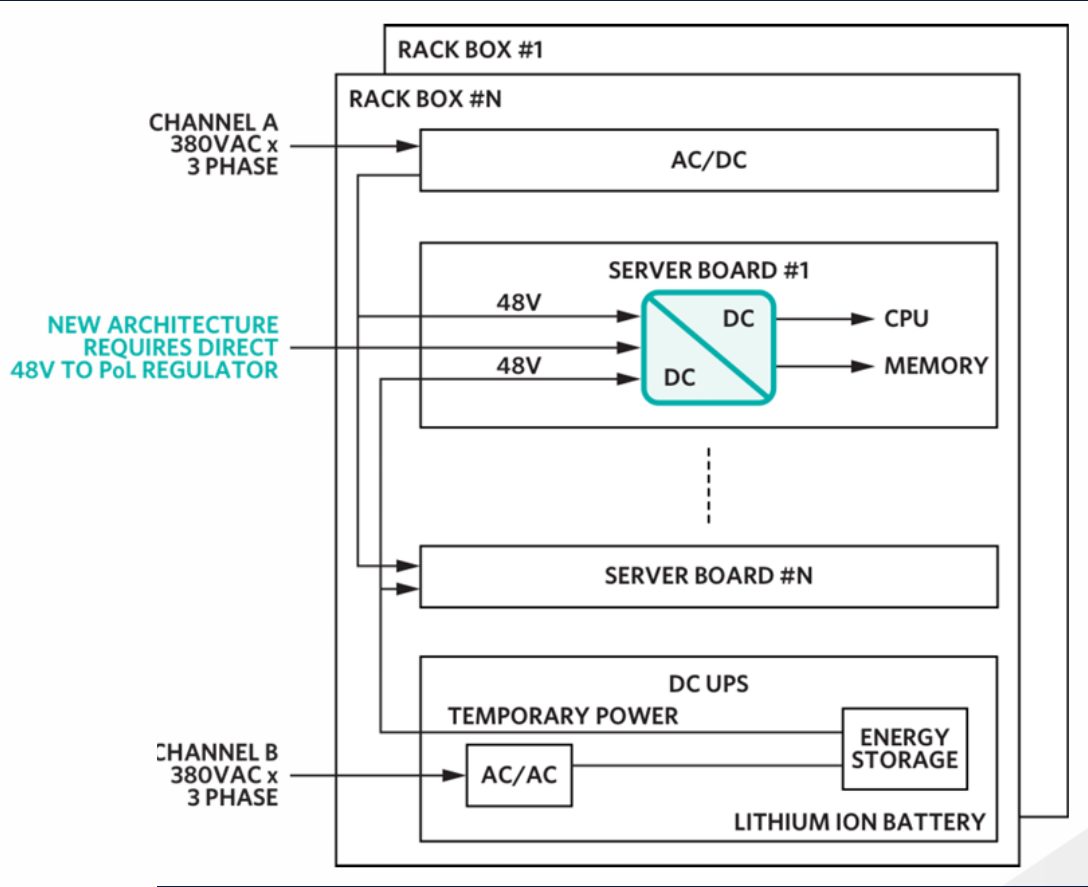
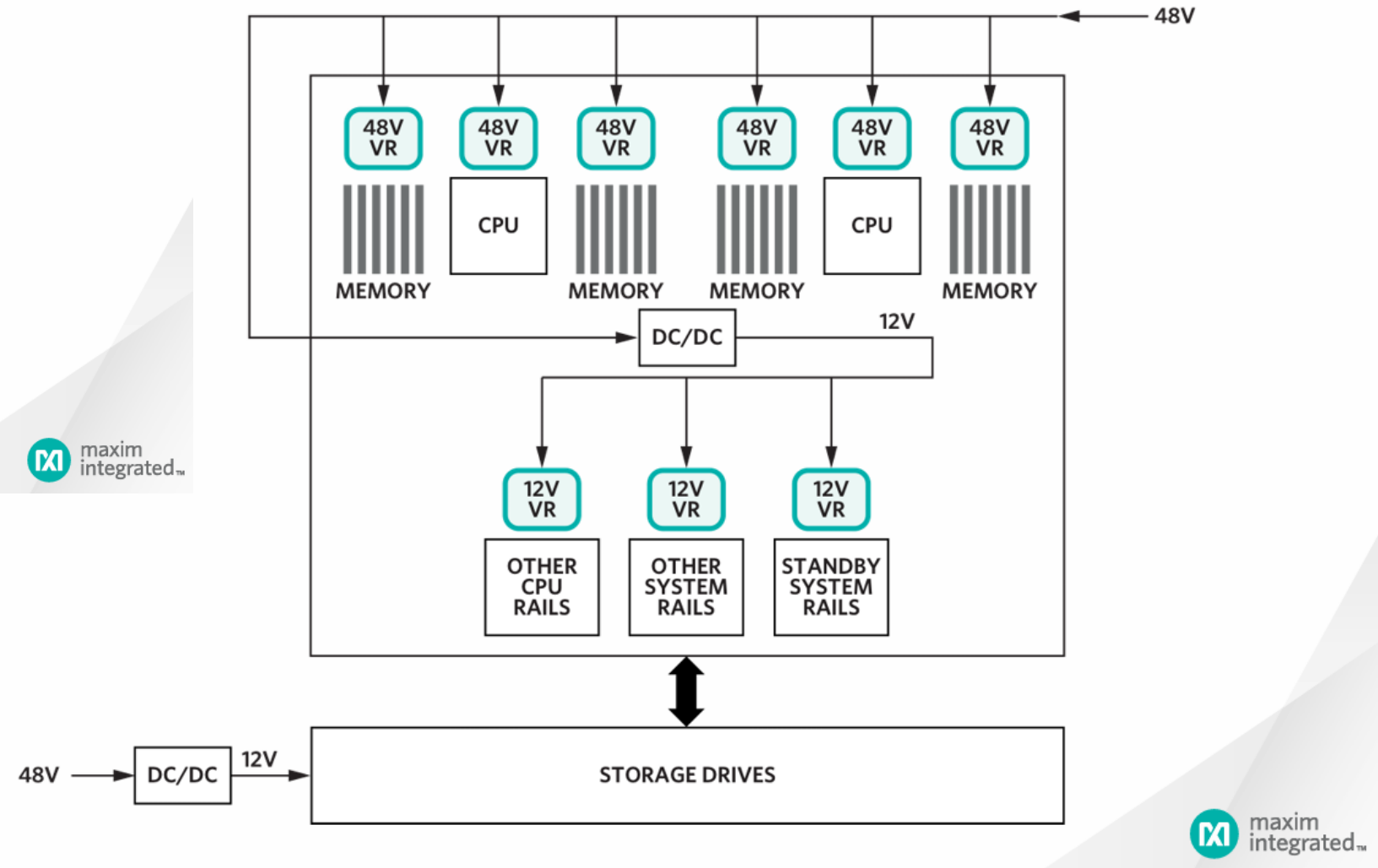
Performance increase 5x
Consumption increase 1.5x

source: NVIDIA

12V Server Motherboard Power Architecture



48V/54V Server Motherboard Power Architecture



48V Power Rack Energy Savings

- > Up to 30% less conversion losses
- > 16x less power distribution losses

Open Rack V3 Power Shelf voltage is 51 V nominal, with a range of 46 V to 52 V, or 54 V nominal, with a range of 52 V to 56 V. The power is connected through four connectors, each rated at 100 A, for a total power of about 20 kW. The current trend is increasing from 20 kW shelf to 30 kW or more

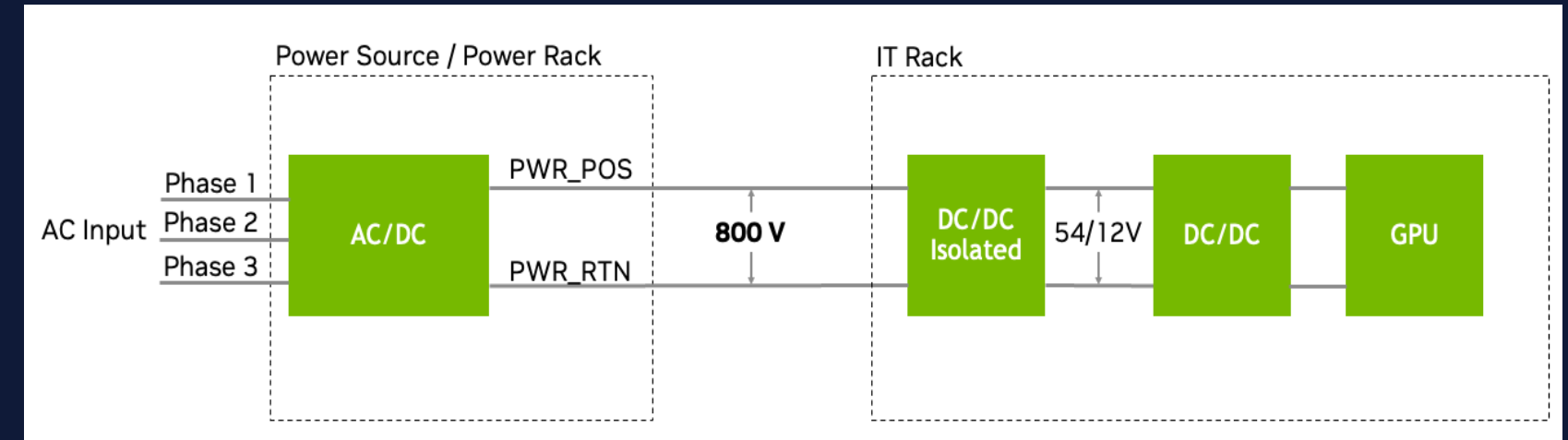
AI Energy Supply Challenge

800V Power Source Architecture

800 VDC Architecture - NVIDIA 2027 initiative

“Traditional 48/54 V in-rack power distribution, designed for kilowatt (KW)-scale racks, isn’t designed to support the megawatt (MW)-scale racks coming soon to modern AI factories. ”

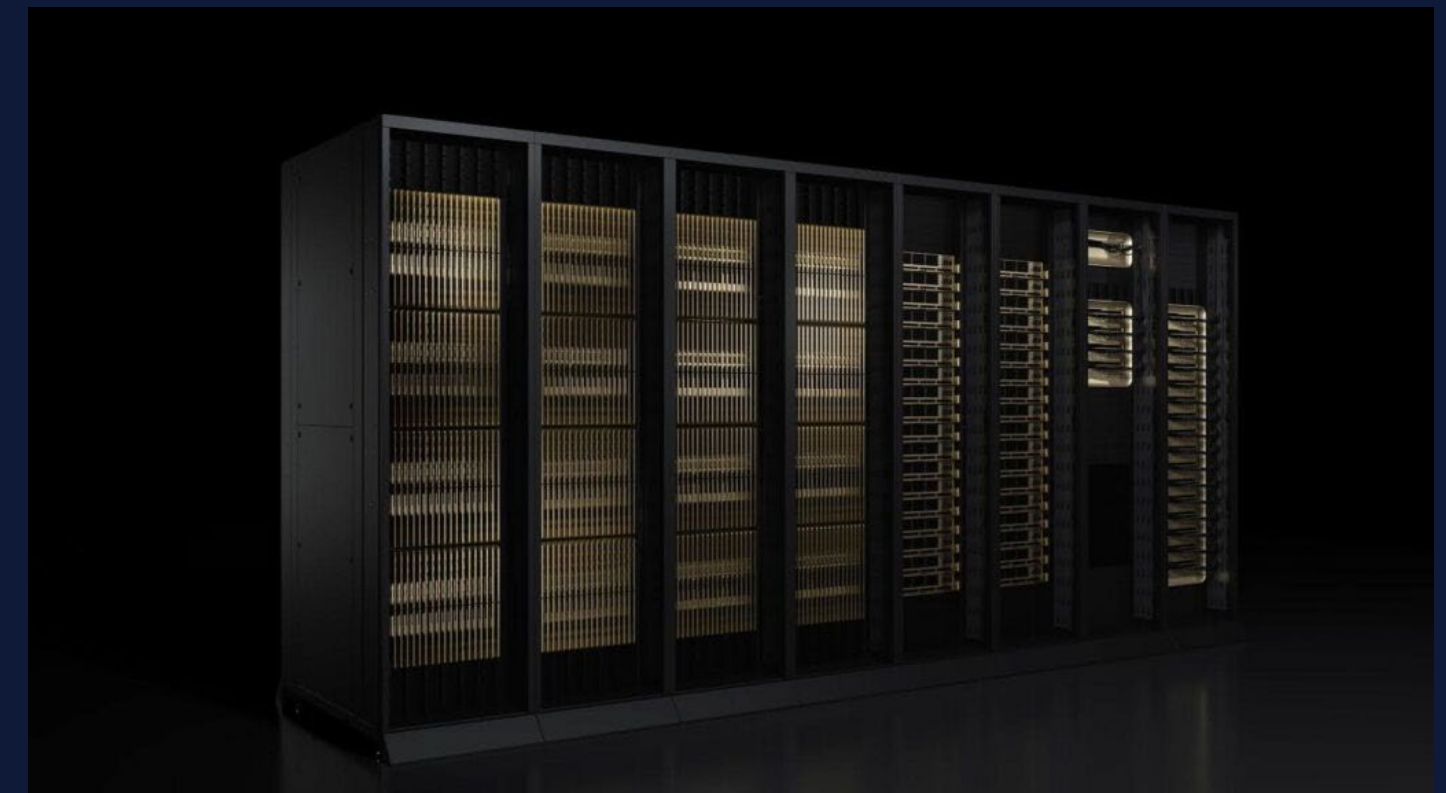
NVIDIA is leading transition to 800 VDC data center power infrastructure to support 1 MW IT racks and beyond, starting in **2027**. To accelerate adoption, NVIDIA is collaborating with key industry partners across the data center electrical ecosystem.



NVIDIA 800 VDC architecture minimizes energy conversions

Key efficiency gains

- Up to **5%** improvement in end-to-end power efficiency
- **Maintenance costs reduced by up to 70%** due to fewer PSU failures and lower labor costs for component upkeep
- **Lower cooling expenses** from eliminating AC/DC PSUs inside IT racks
- **Cost reductions** in copper usage and thermal losses across the data center backbone.

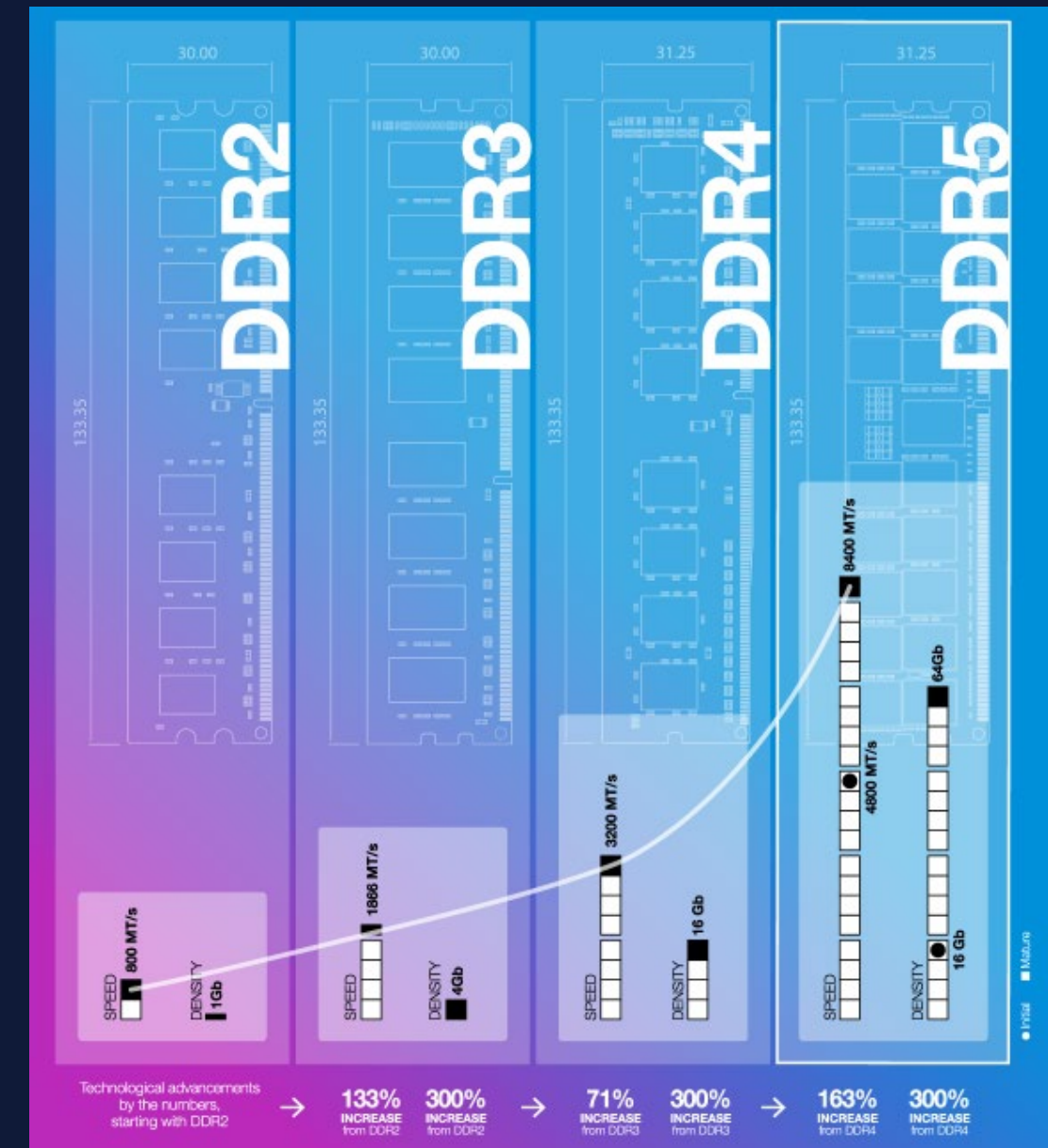


Source: NVIDIA PR: May20, 2025 <https://developer.nvidia.com/blog/nvidia-800-v-hvdc-architecture-will-power-the-next-generation-of-ai-factories/>

DDR5 Memory Architecture Benefits

- DDR5 Memory architecture offers up to **50% higher data rates**
- **10-20% better energy efficiency** – Lower I/O voltage and on-DIMM power management reduce energy per bit transferred
- **Greater capacity per DIMM** – Higher-density modules mean you can feed large models without spilling to slower storage.
- **Improved parallelism** – More independent banks allow multiple AI threads to access memory simultaneously without as many stalls.

Challenge: Complete architecture and memory power architecture re-design



Source: Micron



3

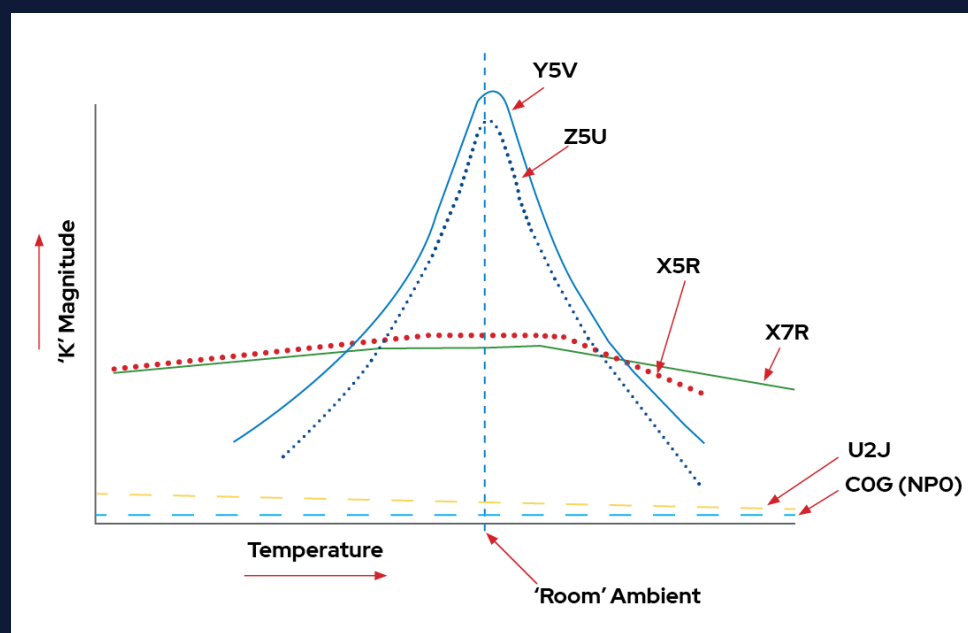
Consequences

Passive Electronic
Components Requirements

Boost in Demand for MLCC and its Development

Demand for MLCC capacitors in servers to quadruple by 2029

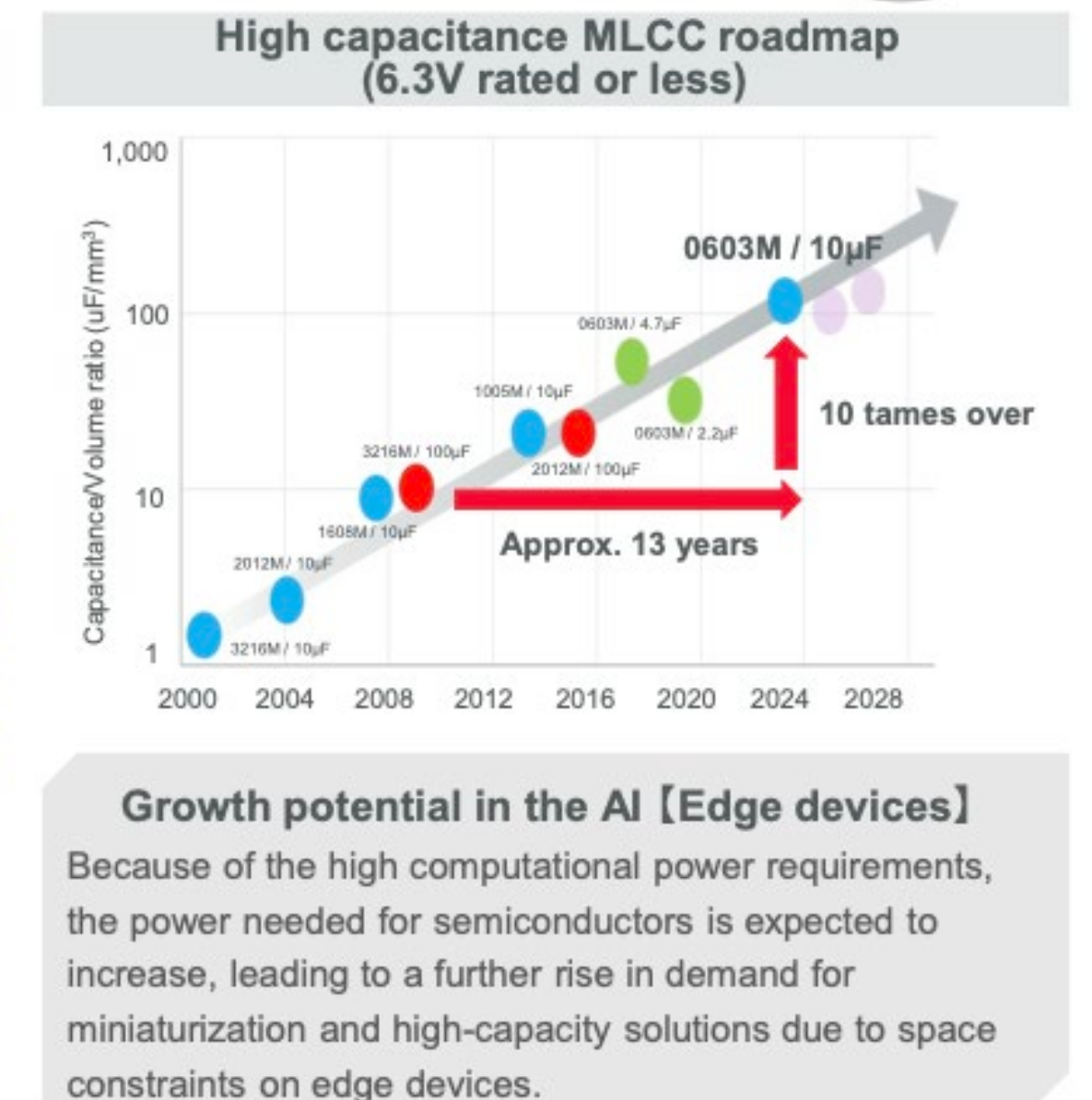
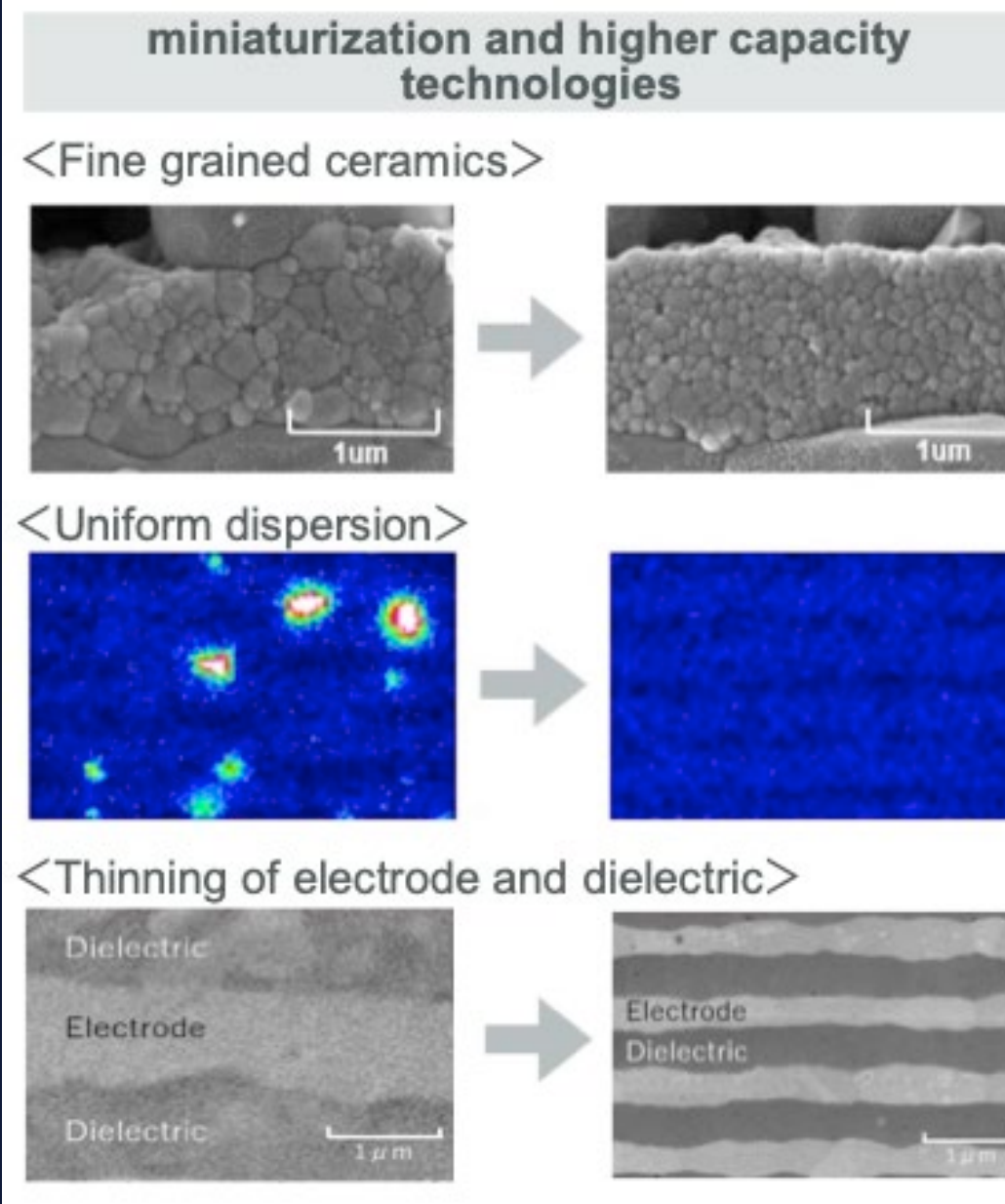
- Fine grained ceramic dielectric materials
- Uniform dispersion / purity
- Thinner layers
- New dielectric materials



Demand for AI servers and other IT infrastructure applications will grow as cloud AI becomes more prevalent

Servers (Baseboard)*	1,800–2,500
AI servers (Baseboard + AI accelerators × 8)*	10,000–20,000

- High reliability
- Increased capacity

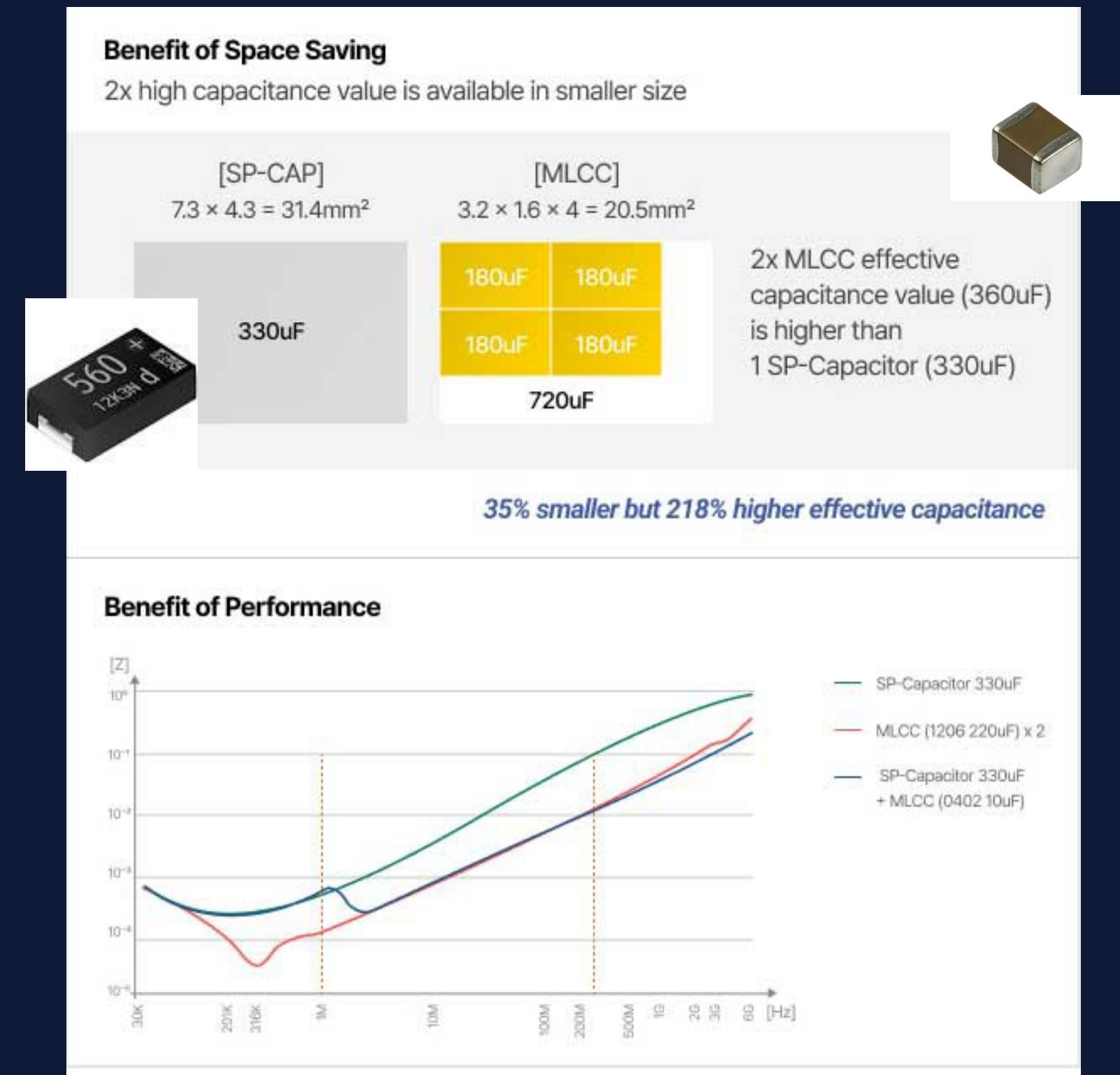


GPU/CPU Power Management Bulk Capacitors

- MLCC try to offer alternative to Al polymer chips
- Low voltage 2.5/4V capacitors
- Downsizing even at “effective capacitance”

	10uF	22uF	47uF	100uF	220uF	330uF
0402	4V X6S	4V X6S	General Server			AI Server
0603		4V X6S	4V X6S			
0805			4V X6S	4V X6S		
1206				4V X6S	4V X6S	
1210						2.5V X6S

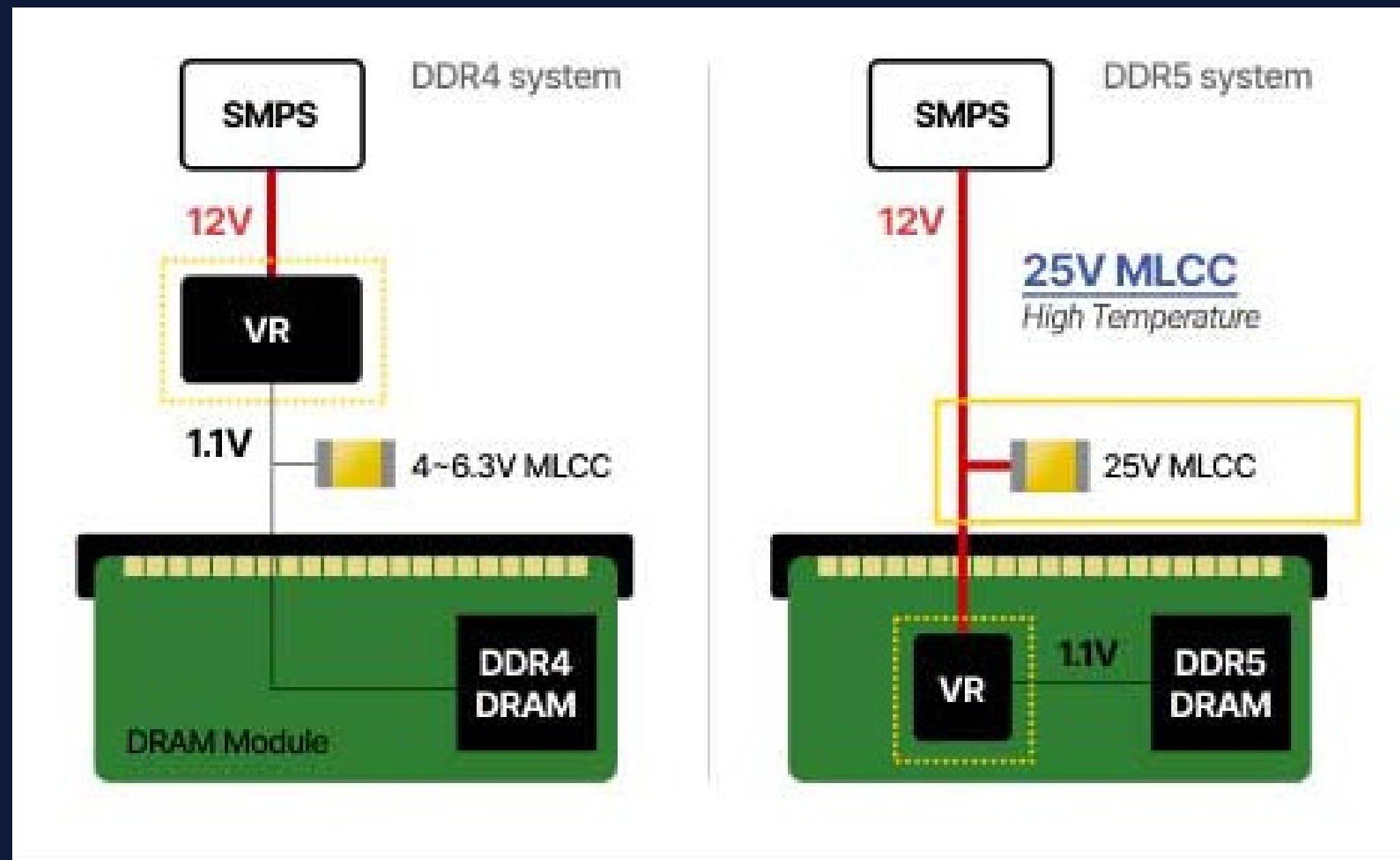
new releases 2025



source: Samsung Electro-Mechanics

DDR4 to DDR5 Memory Architecture Migration

The shift from DDR4 to DDR5 has brought significant changes to power architecture
DDR5 architecture **moves the voltage regulator (VR) to the inner part of the mainboard**



source: Samsung Electro-Mechanics

Consequences

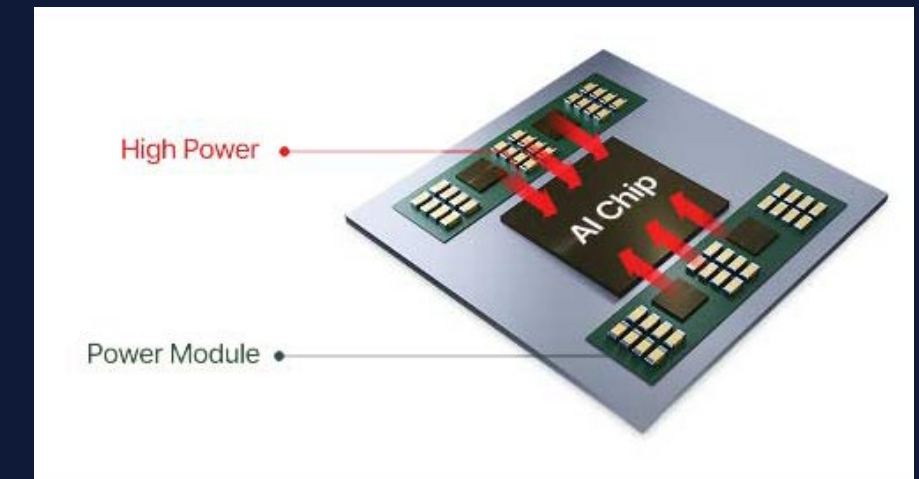
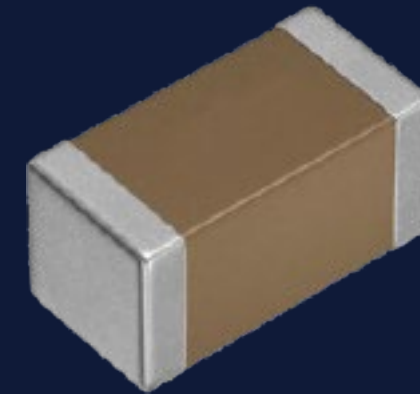
Migration of 10uF/6.3V MLCC /
Tantalum capacitors in DDR4

to 25V capacitors

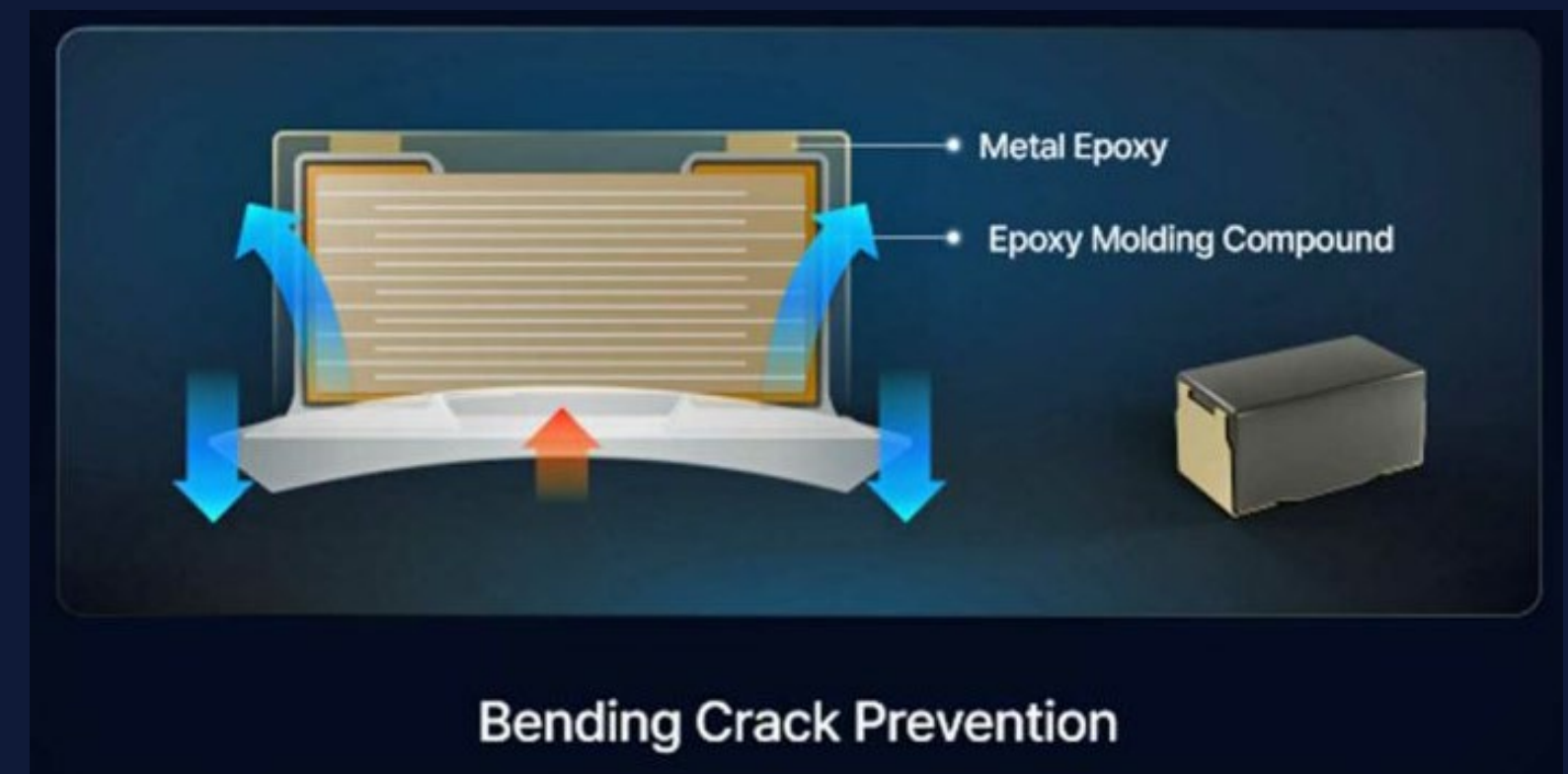
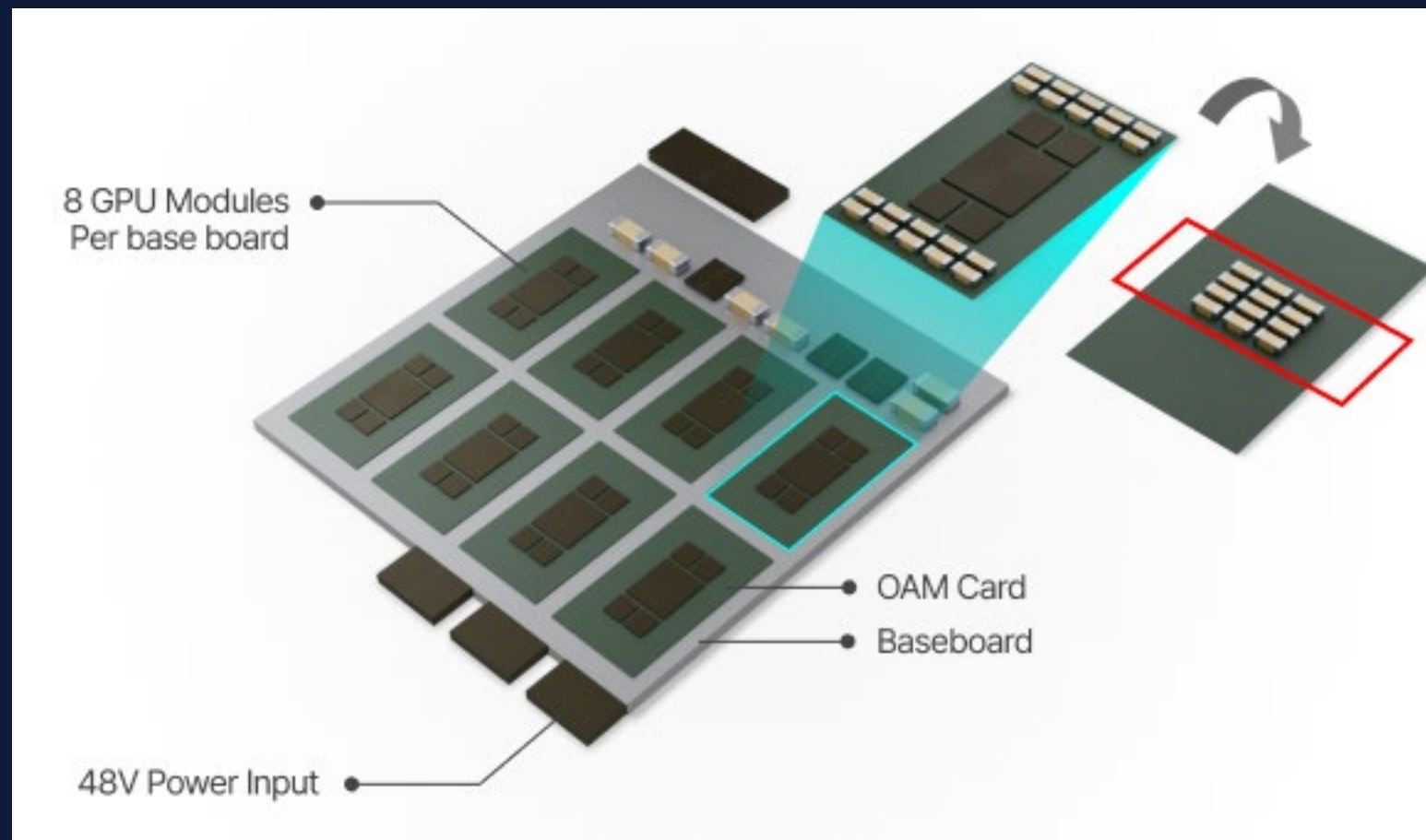
22uF (10uF effective) / 25V now
available in 0805 MLCC case size

Capacitors for 48-52V Rack Power Voltage

- Samsung Releases 4.7 μF 100V in 1206 inch 3216 metric X7S dielectric
- Murata and TDK Releases 1 μF 100V in 0603 inch 1608 metric X7R
- Murata Releases 10 μF /100V X7S MLCC in 1210 inch size 3225 metric



molded frame MLCCs to suppress flex issues in 1210 size (4.7 μF 100V)
specifically targeting AI server boards



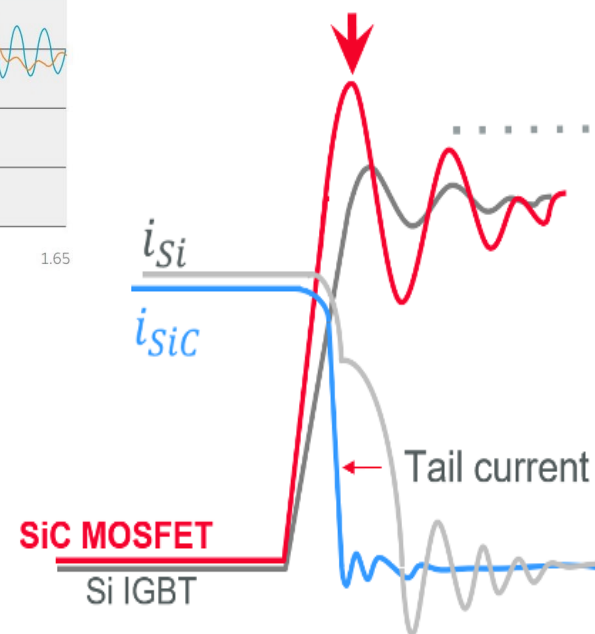
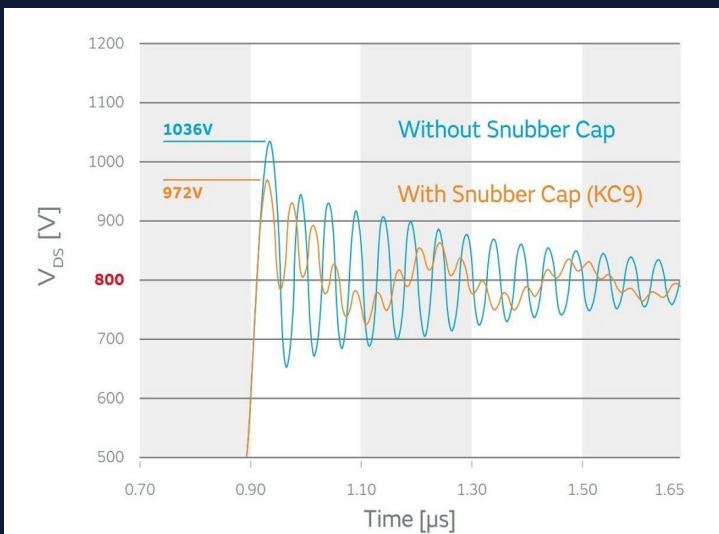
source: Samsung Electro-Mechanics, Murata, TDK

800V Power Supply Architecture



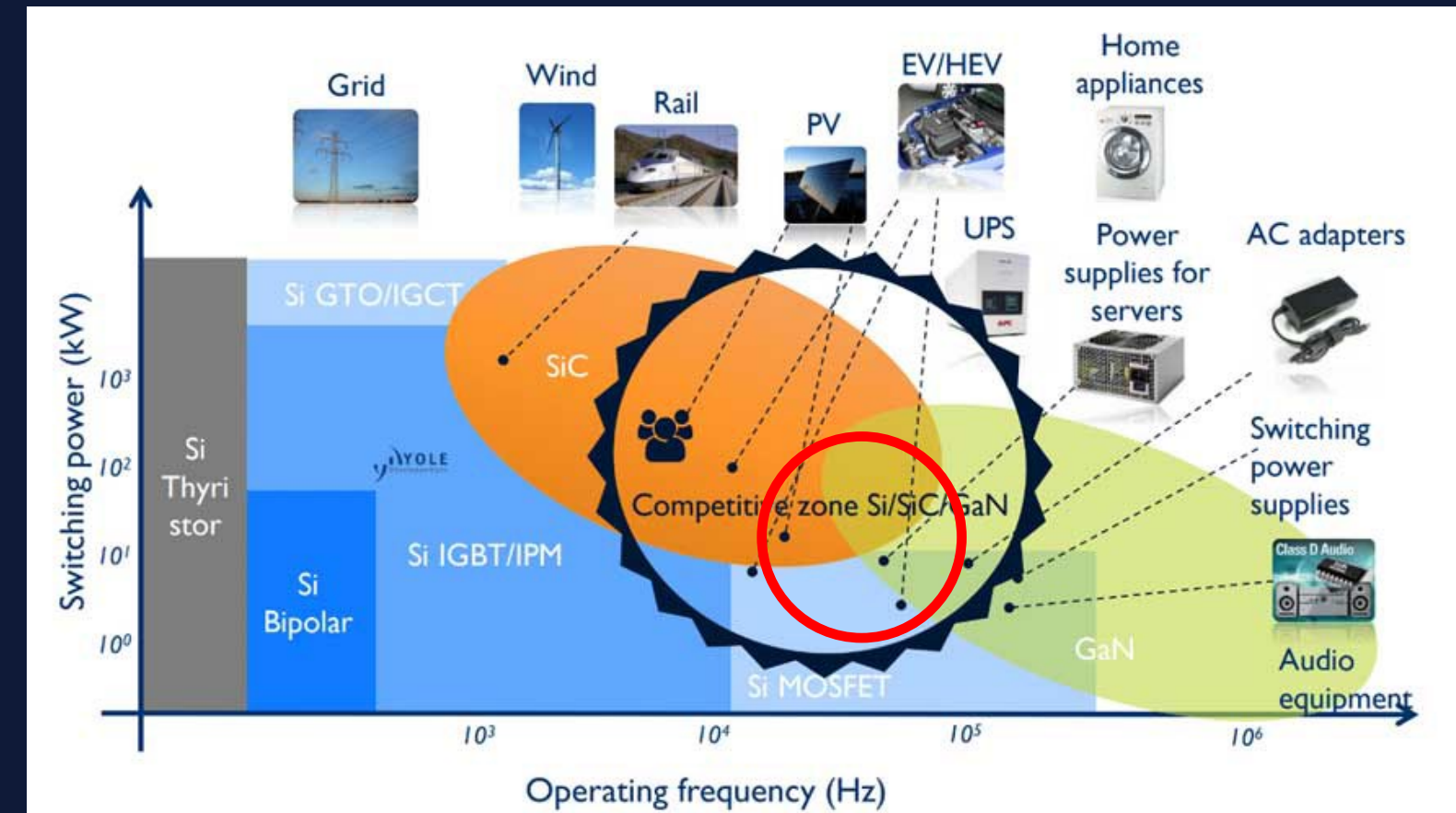
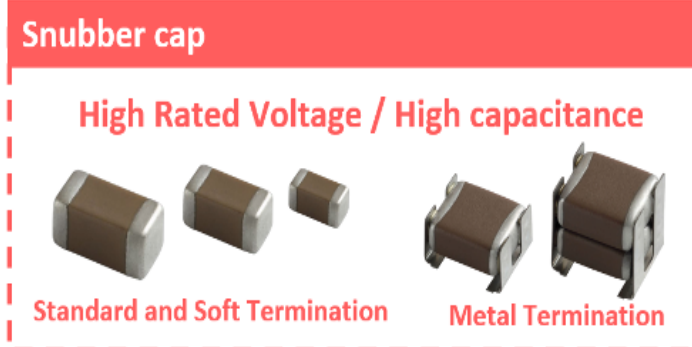
Move to Wide Gap Semiconductors – faster switching = higher surge voltage

Snubber capacitor, Safety capacitors, Circuit protection devices



Turning off (voltage/current between Drain and source)
SiC MOSFETs exhibit a higher surge voltage than traditional Si IGBTs

Snubber capacitor is more necessary



Typical operating frequency and switching power of common semiconductor materials (Source: Yole Group)

Safety Capacitors



increase creepage distances for high-voltage applications

Source: Murata

Murata's testing, utilizing double-pulse evaluation setups with typical 800V SiC configurations, demonstrates capability to mitigate VDS surge, reducing it from 1036V down to 972V.

Consequences to Passive Components

MLCC vs Silicon Capacitors

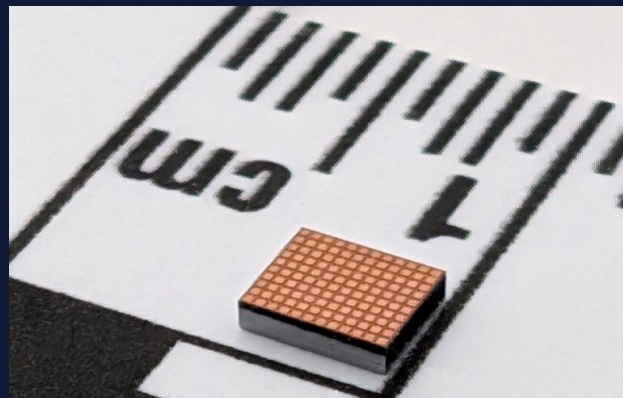
CPU/GPU Coupling Challenge

- Low parasitics, high frequency
- high ripple current

Under the GPU / Package Substrate / Embedding

close proximity placement

Empower releases 16.6 uF with sub 1 pH of ESL and 3 mΩ of ESR in a 120-pad configuration silicon capacitor for substrate core

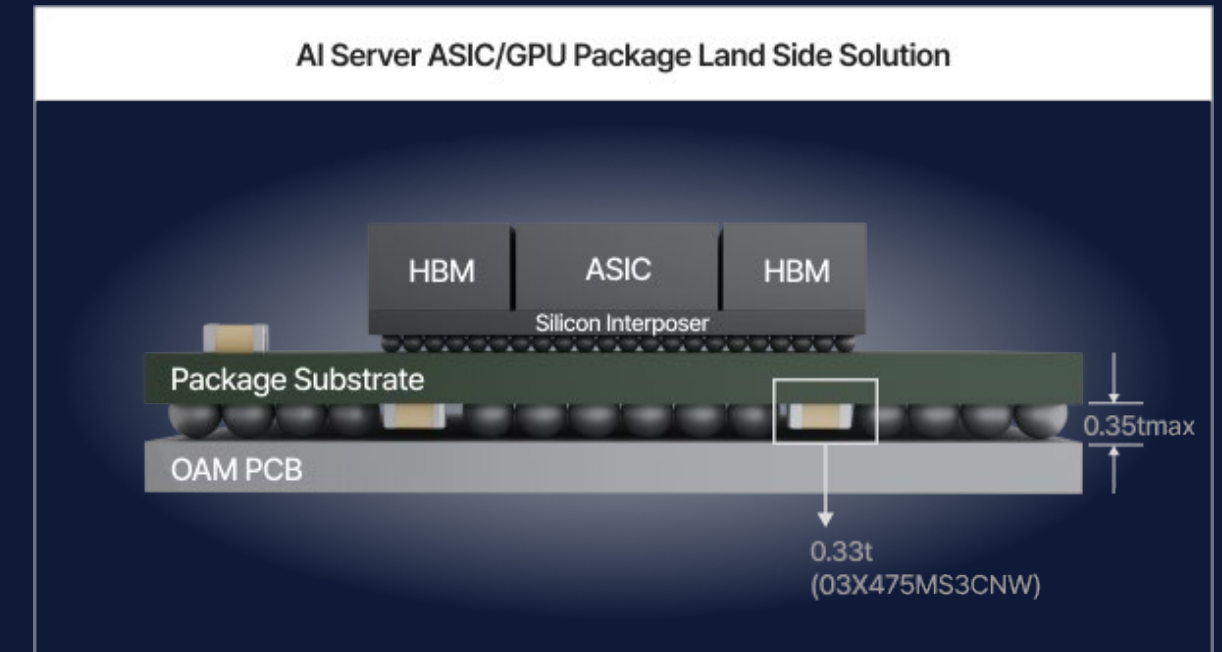


- capacitance density up to **2.5 uF/mm²**
- **stable capacitance performance**
- ultra-low ESL - single digit pH
- ultra low ESR



Silicon capacitors

MLCC capacitors



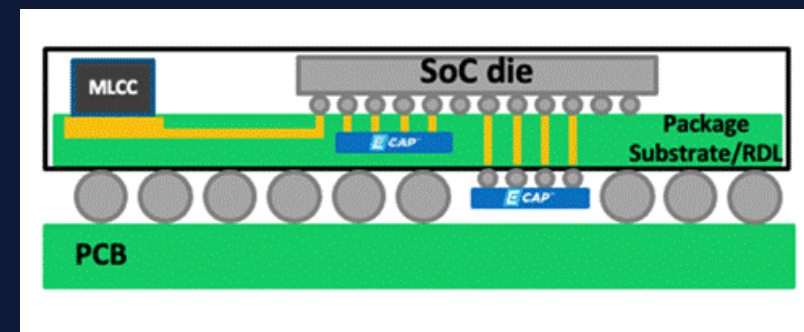
Release Aug 25 (Samsung Electro-Mechanics):
Ultra-High-Capacitance 4.7uF 2.5V MLCC in 0201, X6S for AI GPU

Capacitance density calculation

0201inch: $0.6 \times 0.3 \text{ mm} = 4.7 \text{ uF} / 0.18 \text{ mm}^2 = \mathbf{26 \text{ uF/mm}^2}$
strong dependence to DC BIAS, AC and temp

Effective Capacitance at 1.25V, small AC, 85C:
DC BIAS – 50% of rated = 13 uF/mm^2
AC small ripple – 40% = 7.8 uF/mm^2
temp 85C -7% = **7.2 uF/mm^2**

Source: Samsung CL03X475MS3CNW datasheet



CPU/GPU Power Supply Management

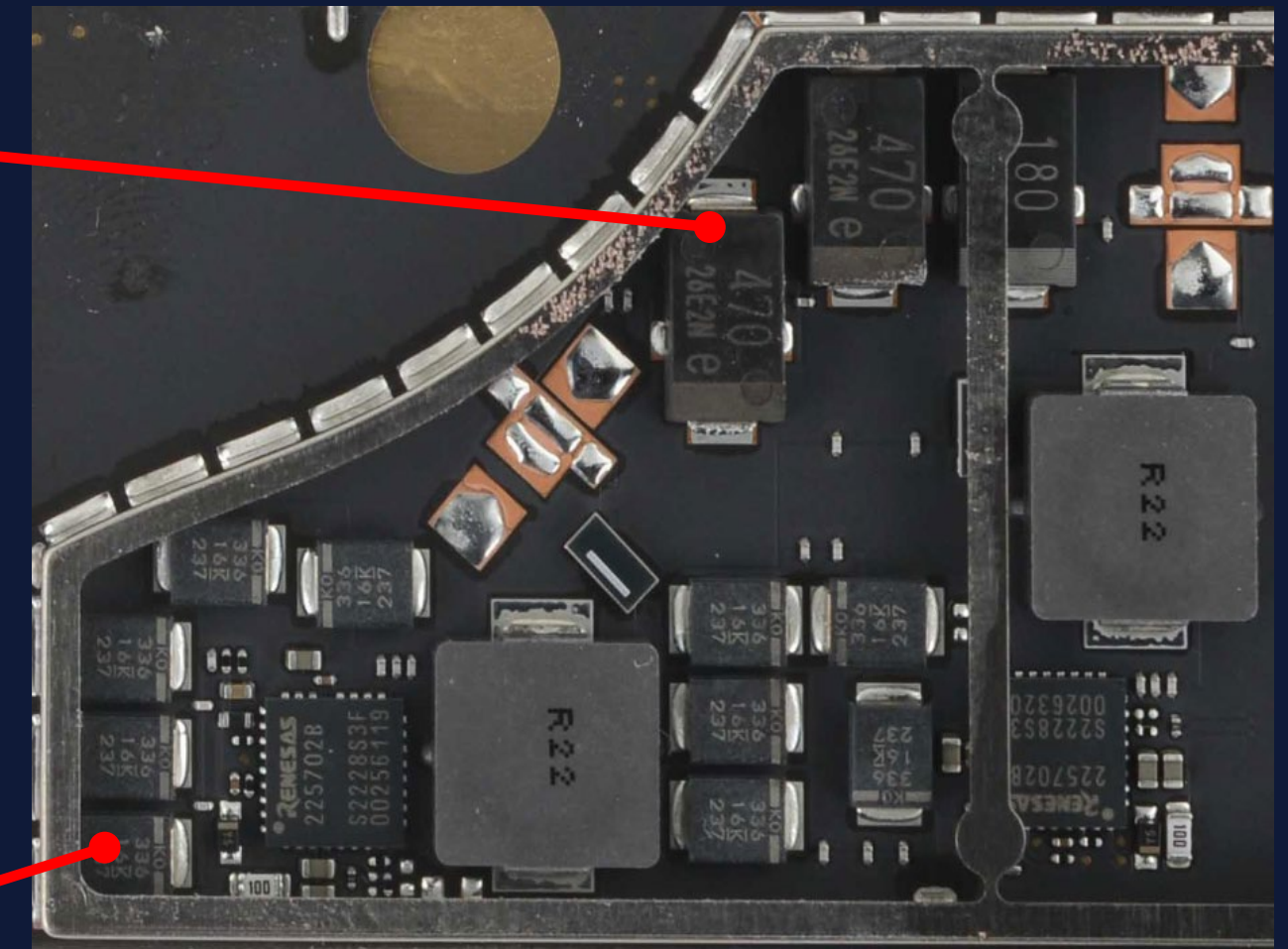
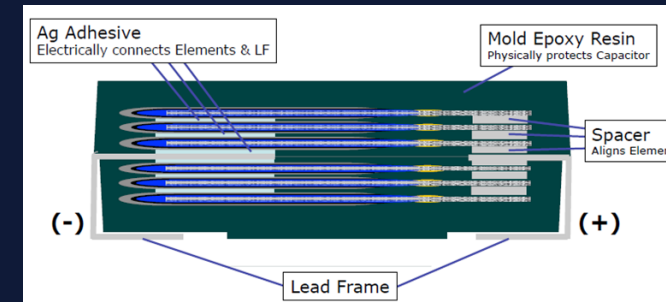
Aluminum Polymer Chip Capacitors



- 470uF 2V 9mΩ – low voltage bulk capacitors
- Extending Lifetime to 5500h@125C

Source: YAGEO KEMET

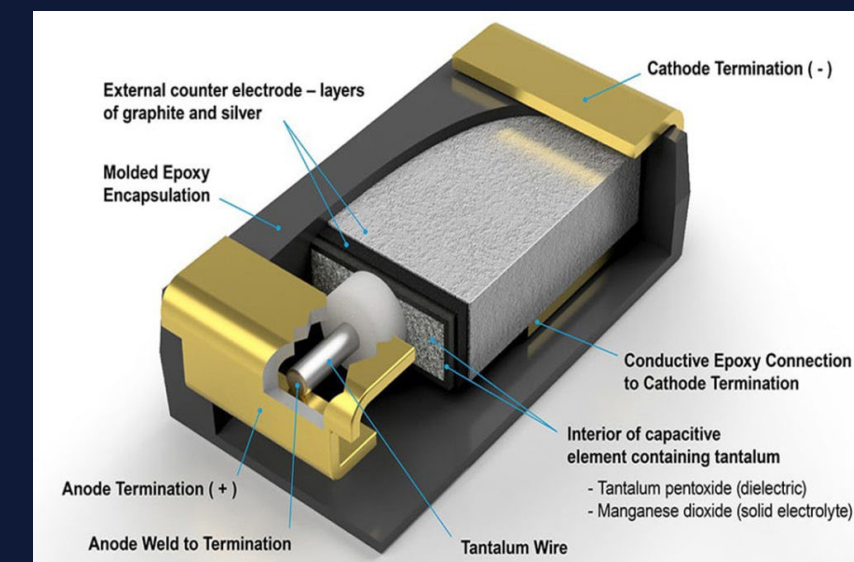
Note: for CPU decoupling circuits operating at low voltage and high current, low ESR might negatively impact system stability (such as MLCC). Carefully controlled ESR values (Polymer caps) prevent parasitic oscillations, ensuring stable operation.



Apple Macbook Pro M4 board Processor Power Management

Tantalum Polymer Chip Capacitors

- Higher CV, Lower ESR 16V capacitors for 12V rail voltage (AI NBs)
- 63/75V low ESR capacitors for 48V rail voltage (AI Servers)



Source: YAGEO KEMET, KYOCERA AVX

AI Induces Fast Transient Dynamics Challenges to Power Grids:
Sudden spikes in consumption due to AI operations can strain the grid, leading to instability or the need for additional infrastructure.

[Submitted on 9 Sep 2024]

The Unseen AI Disruptions for Power Grids: LLM-Induced Transients

Yuzhuo Li, Mariam Mughees, Yize Chen, Yunwei Ryan Li

Recent breakthroughs of large language models (LLMs) have exhibited superior capability across major industries and stimulated multi-hundred-billion-dollar investment in AI-centric data centers in the next 3-5 years. This, in turn, bring the increasing concerns on sustainability and AI-related energy usage. However, there is a largely overlooked issue as challenging and critical as AI model and infrastructure efficiency: the disruptive dynamic power consumption behaviour. With fast, transient dynamics, AI infrastructure features ultra-low inertia, sharp power surge and dip, and a significant peak-idle power ratio. The power scale covers from several hundred watts to megawatts, even to gigawatts. These never-seen-before characteristics make AI a very unique load and pose threats to the power grid reliability and resilience. To reveal this hidden problem, this paper examines the scale of AI power consumption, analyzes AI transient behaviour in various scenarios, develops high-level mathematical models to depict AI workload behaviour and discusses the multifaceted challenges and opportunities they potentially bring to existing power grids. Observing the rapidly evolving machine learning (ML) and AI technologies, this work emphasizes the critical need for interdisciplinary approaches to ensure reliable and sustainable AI infrastructure development, and provides a starting point for researchers and practitioners to tackle such challenges.

NVIDIA unveiled GB300 server, (H2 2025 launch) with integration of supercapacitors into its power system

- The exact number of supercapacitors per rack is not disclosed
- Estimation calculation based on GB300 datasheet (thanks GPT5):
the rack needs 65 - 165F at 48–52 V bus to cover 200–500 ms at 125 kW

Cells per rack (series cells × parallel strings):

EDLC scenario: 18 series cells × 3 - 6 parallel ≈ **54–108 cells**.

LIC scenario: 13 series cells × 5 -10 parallel ≈ **65–130 cells**.

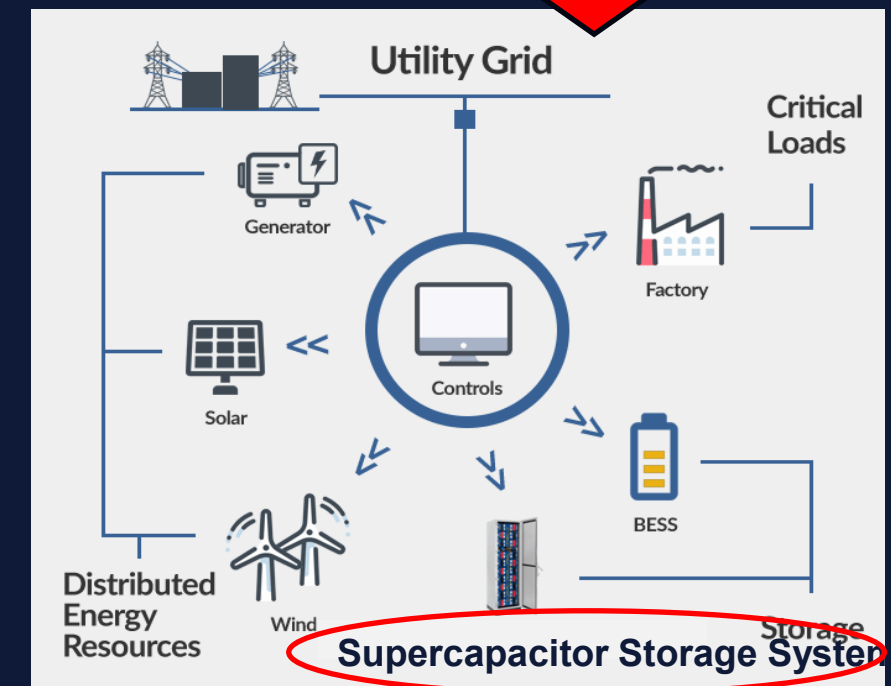
≈ over **100 SC cells per rack!**

this may boost and drive SC Demand in near future

- AI infrastructure features sharp power surge and dip, and a significant peak-idle power ratio. The power scale covers from several hundred watts to megawatts, even to gigawatts. These never-seen-before characteristics make AI a very unique load and pose threats to the power grid reliability and resilience
- Battery energy storage systems are being implemented; however, faster, long cycle lifetime stabilisation is needed



- Reduce the power connection required for AI data centres by 40% by handling fluctuating peak loads
- Reduce total electricity consumption in data centres by 1/3 by storing excess energy in supercapacitors, instead of overheating
- Increase computing power by up to 40%.



AI Server power consumption rapidly rise
Optimizing power efficiency contributes to power saving

Inductor Requirements

Higher Current Handling

- high-performance GPUs / AI accelerators running in high-current environments
- Inductors must have **higher saturation current** to avoid overheating, functional failure, or damage

Greater Power Density

- Data centers face **space limits** and **heat dissipation challenges**
- Inductors need **compact form factors** with **low DC resistance (DCR)** to cut thermal losses and boost efficiency

High-Frequency Performance

- Next-gen high-frequency DC-DC converters demand inductors with **excellent high-frequency characteristics** for stable, efficient power delivery

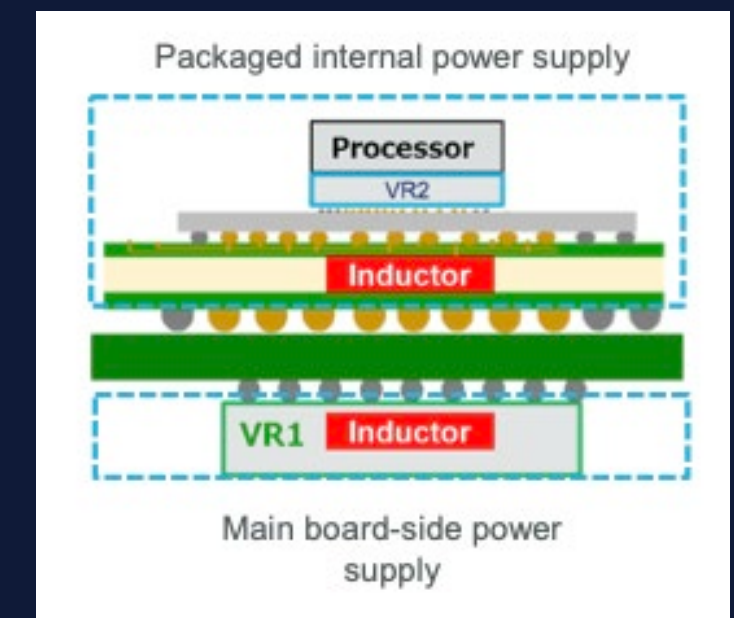
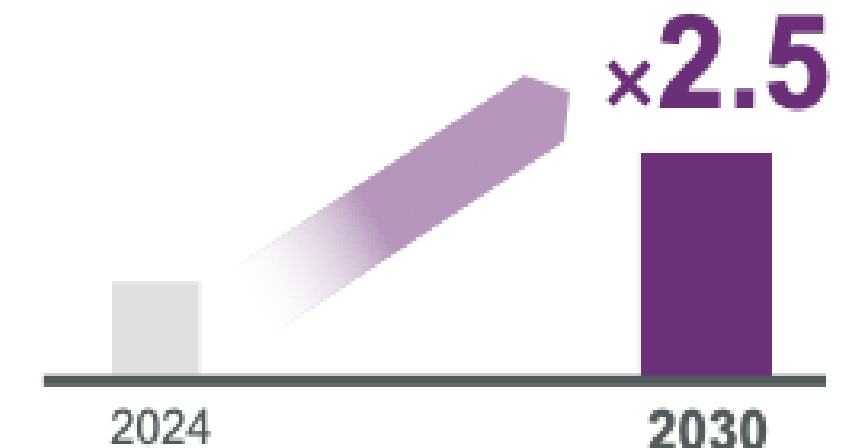
Thermal Stability & Reliability

- AI servers run at **sustained high loads** in **harsh thermal environments**
- Inductors must maintain **performance over time** without degradation

Material Innovations

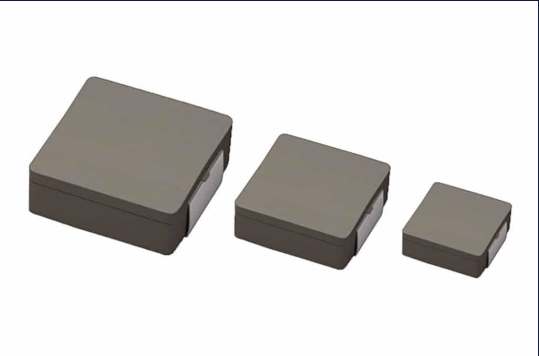
- New nanocrystalline alloys (e.g., Fe-B-P-Cu) deliver **high magnetic flux density** and **low coercivity**, enabling **smaller, more efficient inductors** for AI workloads

Demand value of power inductors for data centers



High Saturation Inductors

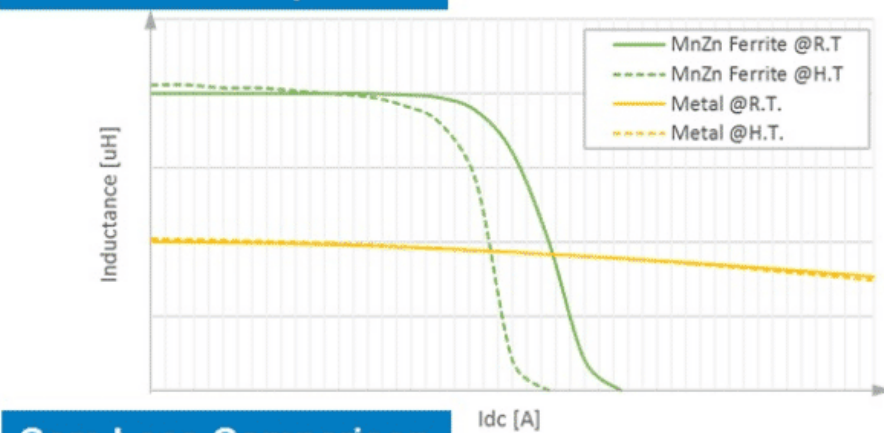
New nanocrystal metal composite magnetic materials featuring high saturation, temp stability



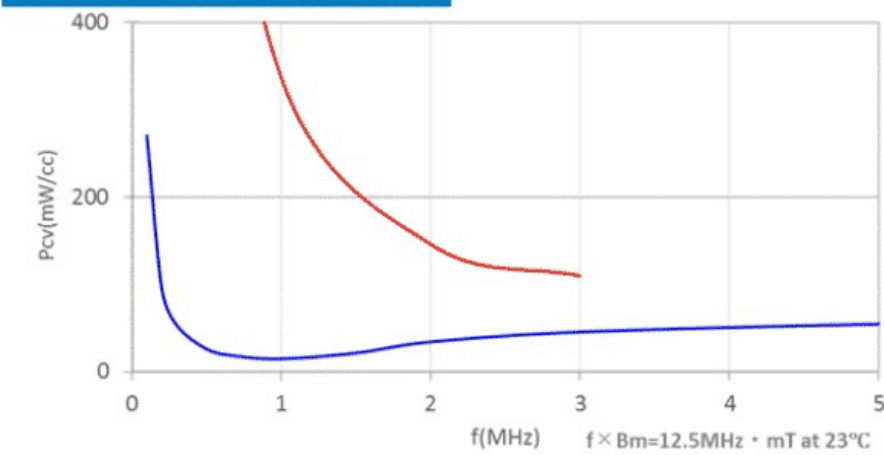
Source: Wuerth Electronic

Flat wire MnZn ferrite core construction increases saturation currents up to 25% at low core losses

Saturation Comparison



Core Loss Comparison



Advantage of Ferrite

- 1. Higher inductance with high permeability
- 2. Stable inductance in the right range

High L and low DCR capability

Advantage of Metal Composite

- 1. Very slow saturation
- 2. Very stable saturation for the thermal

Good for automotive applications

Advantage of Ferrite

Very low core loss in dynamic frequency range

Low power consumption capability

Source: Panasonic

Inductor Technology	Inductance	Magnetic Saturation	Thermal TC	Efficiency	DCR	Insulation Resistance to Core
Metal Composite	Low	Very Good (Soft)	Verry Good	Good	Low to Medium	Good
Ferrite (Mn-Zn)	High	Not as Good (Hard)	Not as Good	Very Good	Low	Not as Good
Ferrite (Ni-Zn)	Medium	Good (Hard)	Good	Good	Lot to High	Very Good

Source: KEMET YAGEO

One Turn Inductors

one-turn Mn-Zn ferrite inductor minimizes heat generation and energy loss, supporting switching frequencies up to 3 MHz

- **DC Optimization:**
hard-switching applications (multi-phase buck converters)
reduced DCR to minimize copper losses, enhancing system efficiency
- **AC Optimization**
soft-switching architectures such (STC Switched Tank Converters)
ferrite core with precisely engineered air gap **reduce AC losses** caused by eddy currents and fringing flux



Low Loss Inductors

AI adoption drive demand for high-speed and high-capacity data transfer communication within a whole communication chain

High frequency, Low loss solutions and components

Components for Optical Transceivers

Critical Role in AI Workloads – Optical transceivers enable ultra-fast, high-volume data exchange across AI compute nodes.

High-Speed Links – Power 400G/800G+ connections for GPU-to-GPU, node-to-node, and rack-to-spine traffic.

Data-Center Backbone – Support low-latency interconnects between servers, storage, and network switches.

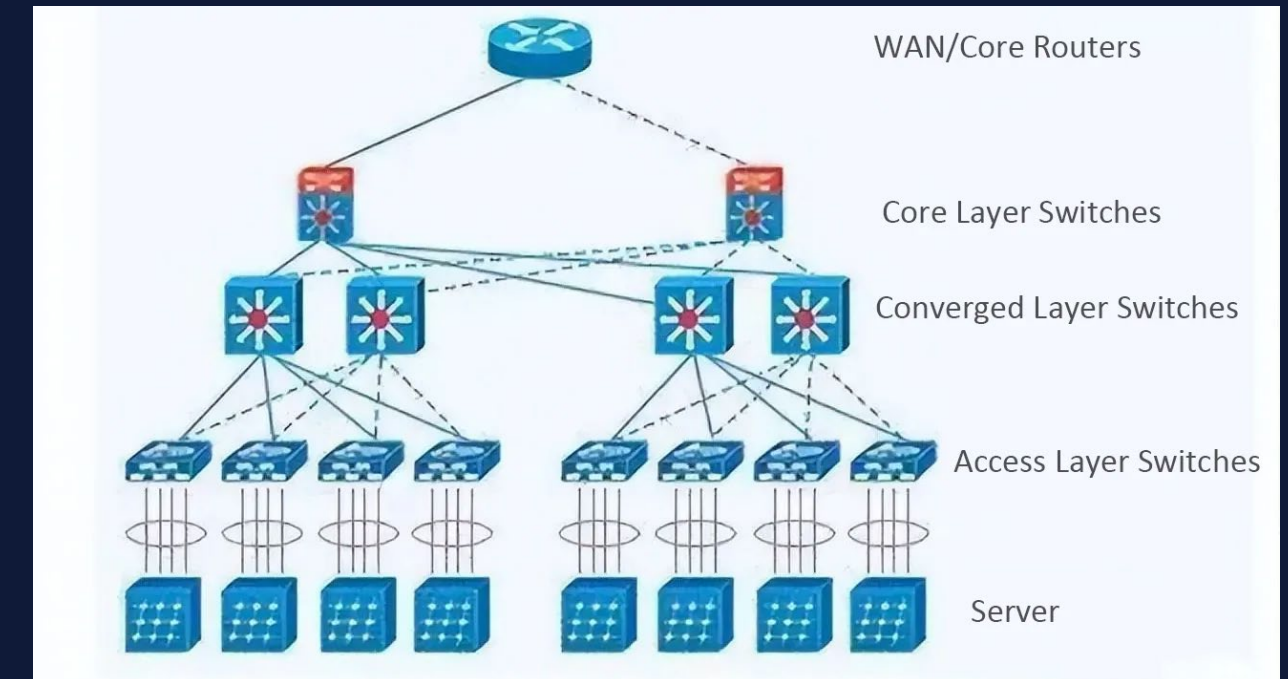
Scalable Infrastructure – Handle growing AI cluster sizes where copper cabling can't meet bandwidth or distance needs.

Long-Haul Performance – Coherent optics connect AI facilities across campuses or regions without throttling throughput.

Latest Inductor example targeting optical transceivers – release Aug 25:
thin-film chip inductor 1.2 x 0.6 x 0.95 mm

DC resistance reduced by 70%,

Rated saturation current increased by 1.7 times



Source: www.naddod.com

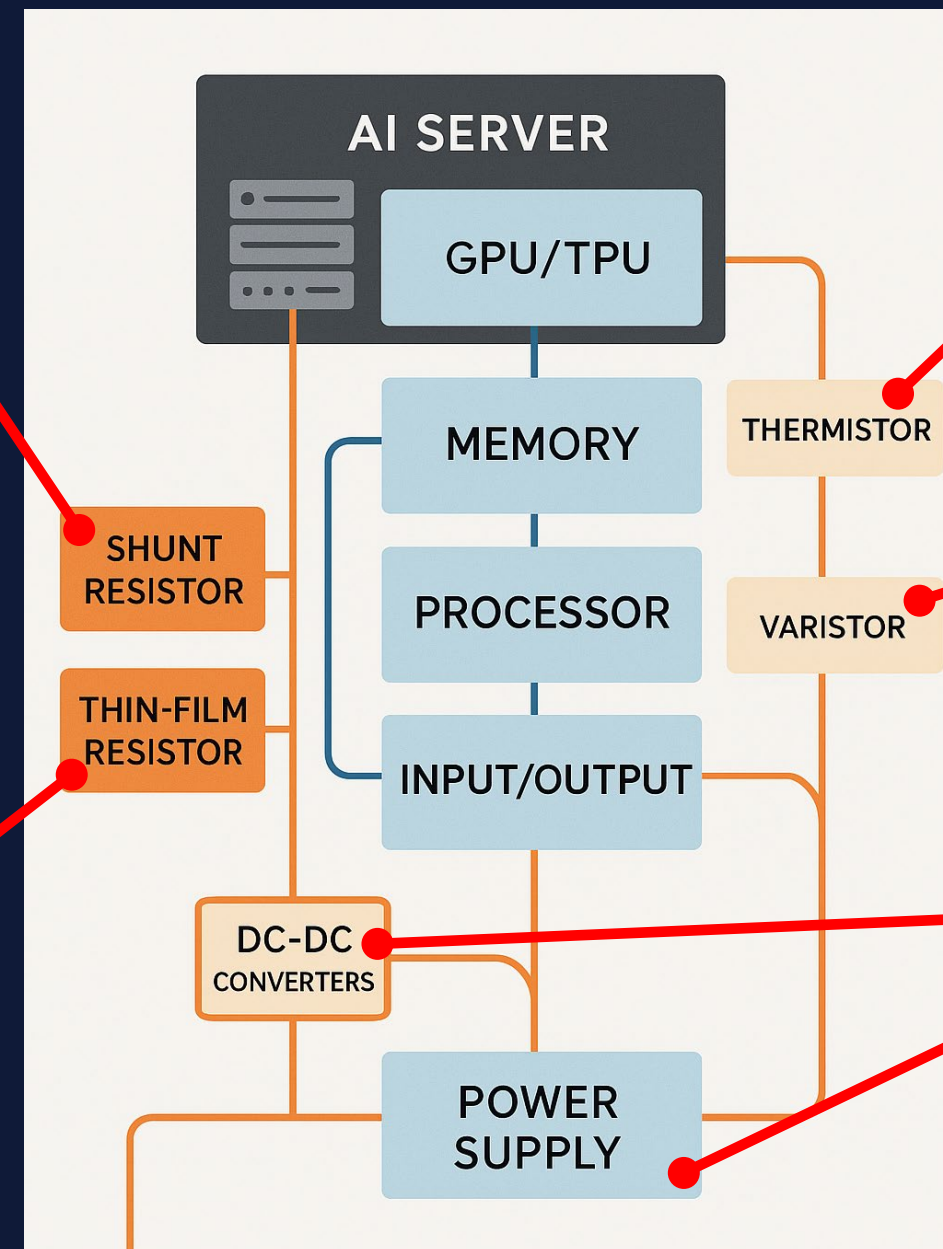


Source: TDK

Resistors in AI Servers

Current-sense resistors in hot-swap circuits and DC-DC converters must handle **higher continuous currents** and **steeper slew rates** without drifting, so low-ohmic, high-power shunt designs with excellent thermal stability are in demand

Thin-film SMD chip resistors in 0402/0201 with **higher pulse-handling** and **better heat dissipation** to fit into compact accelerator boards without sacrificing reliability



Thermistors (NTC/PTC) are critical for **inrush current limiting** and temperature monitoring in dense AI racks.

Varistors (MOVs) **protect against voltage surges** from grid fluctuations or lightning — a bigger concern as AI data centers scale power delivery

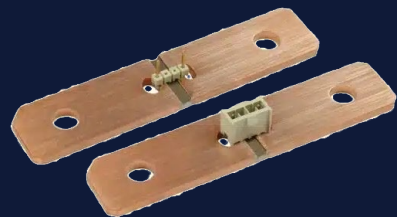
Move toward **wide-bandgap power devices** (SiC, GaN) in AI server PSUs changes switching speeds and EMI profiles, which in turn drives demand for resistors with **low parasitic inductance** and **stable high-frequency performance**

- AI hardware systems require **high precision and tight tolerance specifications** for resistor components, with tolerances down to 0.01% and TCR values of 25 ppm or better.
- Metal film resistors** are preferred in AI applications for their stability, low noise, and accuracy across environmental conditions.
- AI chips face extreme thermal challenges, necessitating **high-power thin film chip resistors** with improved heat dissipation and temperature compensation circuits.
- Noise performance is crucial, especially in high-gain amplifiers and low-level signal processing, with thin film, metal foil, and wirewound resistors offering **superior noise characteristics**.
- Miniaturization** in AI hardware challenges resistor manufacturing, with modern chip resistors ranging from 0201/0402 inch to larger sizes, requiring process and temperature compensation circuits for accuracy.

Function	Best resistor type	Key traits	Example series
High-current sense in VRMs, PDUs, hot-swap	All-metal current sense shunts (low-R)	High power density, low thermal EMF, tight TCR	Stackpole CSSH0805, 1 W in 0805
Precision high-power dividers/feedback	Thin-film on AlN high-power chips	2–6 W in small sizes, 0.1% tol, 25 ppm/°C	TT Electronics TFHP (2 W 1206, 6 W 2512)
Compact RF/power loads and precision terminations	High-power thin-film on AlN	Very high power capability in compact sizes	KYOCERA AVX CR Series (0603 up to 2.6 W; larger up to 250 W)
Ultra-precision, low noise for optics & instrumentation	Bulk metal foil precision	Ultra-low noise/TCR, superb long-term stability	VPG Foil Resistors (data center fiber-optic/networking)
General precision monitoring on server boards	Automotive-grade thin film	Low TCR, ±0.1% tolerance, server/UPS-ready	YAGEO AT thin film (servers & UPS in applications list)



1W High Power Density Current Sense Chip Resistor (source: Stackpole)



Power Metal Strip® current sense resistors TCR down to ± 10 ppm/°C designed for high-voltage systems (400 V / 800 V) (source: Vishay)

Four-terminal shunt resistors, 1% tolerance, resistance as low as 0.0002Ω. (source: YAGEO)



Thin Film High Power AlN Chip 6W in 2512 size 2W in 1206 size, 0.1%, 25ppm/°C (source: TT Electronics)



4

Future of AI Power Delivery

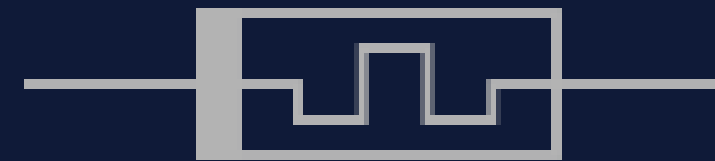
Enabling Technologies and
Developments



Memristor

“memory resistor,” is a two-terminal electrical component that regulates the flow of electrical current in a circuit and retains memory of the amount of charge that has passed through it.

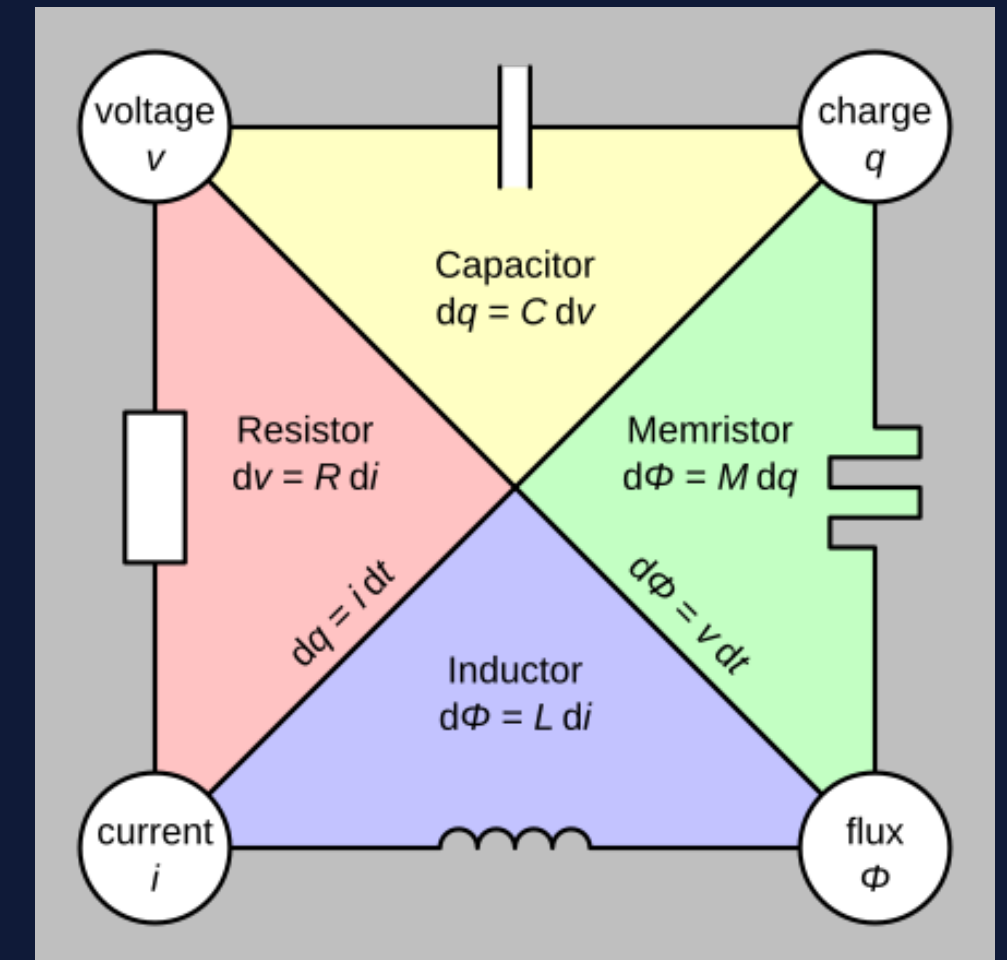
When current flows in one direction, the resistance increases; when it flows in the opposite direction, the resistance decreases. This change in resistance is retained even after the current is turned off, hence the term "memory resistor"



Typically, a memristor consists of a thin film of a material (like titanium dioxide) sandwiched between two metal electrodes. The resistance of this material changes based on the movement of ions or defects within it.

Data shuttling of conventional processors from memory to logic and back again takes a lot of energy and slows the speed of computing and requires a lot of space. If the computation and memory storage could be located in the same space, this bottleneck could be eliminated.

Memristors are proposed as memory elements in nanometre-scale logic for future brain-like networks.



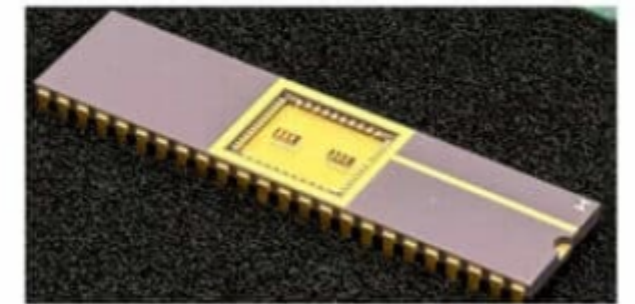
It was described and named in 1971 by Leon Chua, completing a theoretical quartet of fundamental electrical components which comprises also the resistor, capacitor, and inductor

- AI semiconductor development program, which fuses memristor and spintronics technology
- Spintronics: A technology that utilizes both the charge and spin of electrons or the spin element alone
- Novel “spin-memristor” can solve the reliability issues experienced with conventional memristors

TDK Develops Spin-Memristor for Neuromorphic Devices to Reduce AI Power Consumption to 1/100

🕒 3.10.2024 👁 31

TDK develops "spin-memristor" for neuromorphic devices, and collaborates with CEA and Tohoku University to achieve practical application of neuromorphic devices able to reduce power consumption of AI down to 1/100. TDK Corporation ...



AI circuit board equipped with four spin-memristor ceramic packages

The human brain requires around 20 W of power, which enables it to make more complex decisions than current digital AI processors, but with far lower power consumption. Therefore, TDK set itself a goal to develop a device that electrically simulates the synapses of the human brain: the memristor.

Summary



AI CHALLENGE: POWER DEMAND

AI systems, particularly GPUs and servers, are experiencing a rapid increase in power consumption, GPU already exceeding 1200W and entire racks projected to surpass 130kW by 2026.



EFFICIENCY IS PARAMOUNT

This escalating demand for a fundamental shift in power architecture. The industry is moving from 12V to 48V power rack voltage and exploring 800V power supply systems to minimize conversion and distribution losses, reducing cooling requirements and operational costs.



IMPACT ON PASSIVE COMPONENTS

The increased power density, higher frequencies, and tighter spaces in AI systems place extreme demands on passive components:

- Capacitors (MLCC, Polymer, Silicon):** Face intense pressure for lower ESR/ESL, higher current handling, improved thermal stability, and further miniaturization. The demand for MLCCs in AI servers is expected to grow significantly.
- Inductors:** Require higher saturation currents, greater power density, and superior high-frequency performance to support next-generation DC-DC converters. Technology involves nanocrystalline/metal composite powder cores, flat wire designs or one turn inductors.
- Resistors:** Need to handle higher currents and faster switching speeds, leading to a demand for low-inductance, high-power shunt and thin-film resistors with excellent thermal stability.



MATERIAL & DESIGN INNOVATION

Development of nanocrystalline alloys, utilization of wide-bandgap semiconductors (GaN/SiC), and creation of high-density inductors are enabling significant performance improvements.



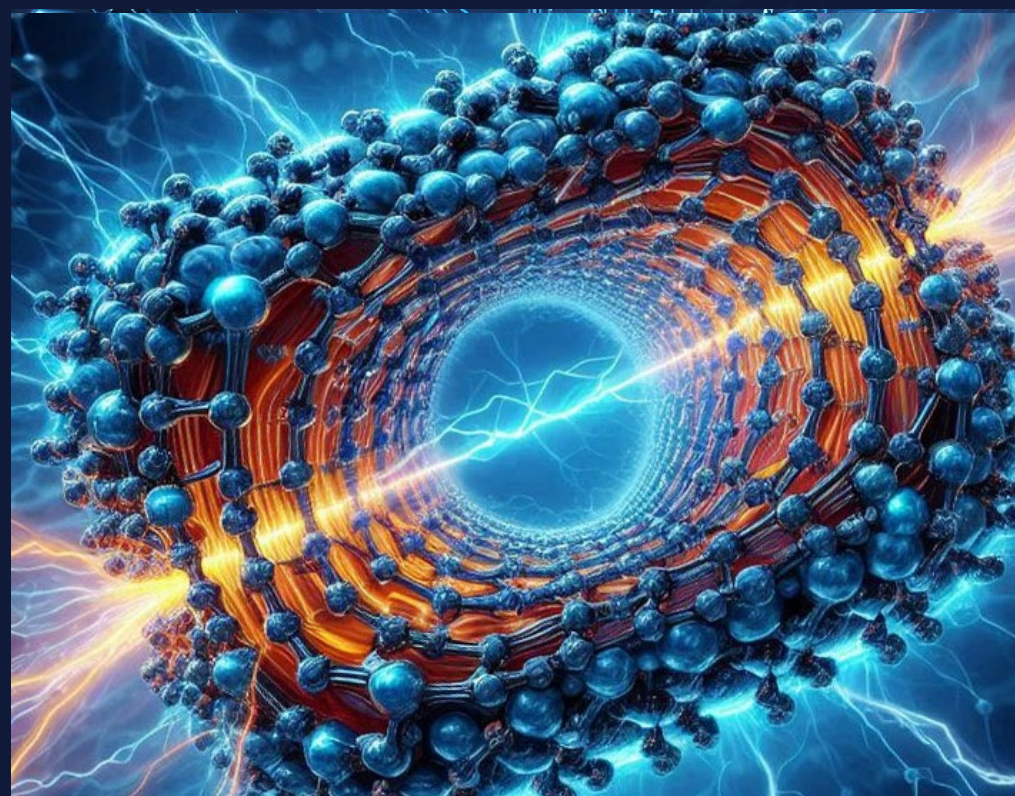
ENERGY STORAGE EVOLUTION

Supercapacitors and hybrid solutions are being integrated to manage the extreme and rapid power fluctuations (transients) characteristic of AI workloads, boosting overall efficiency.



NEXT HORIZON

Neuromorphic computing, utilizing technologies like spin-memristors, promises a revolutionary leap in efficiency, potentially reducing power consumption by up to 100 times compared to current AI compute methods.



European
Passive Components
Institute

www.passive-components.eu

Thank you!