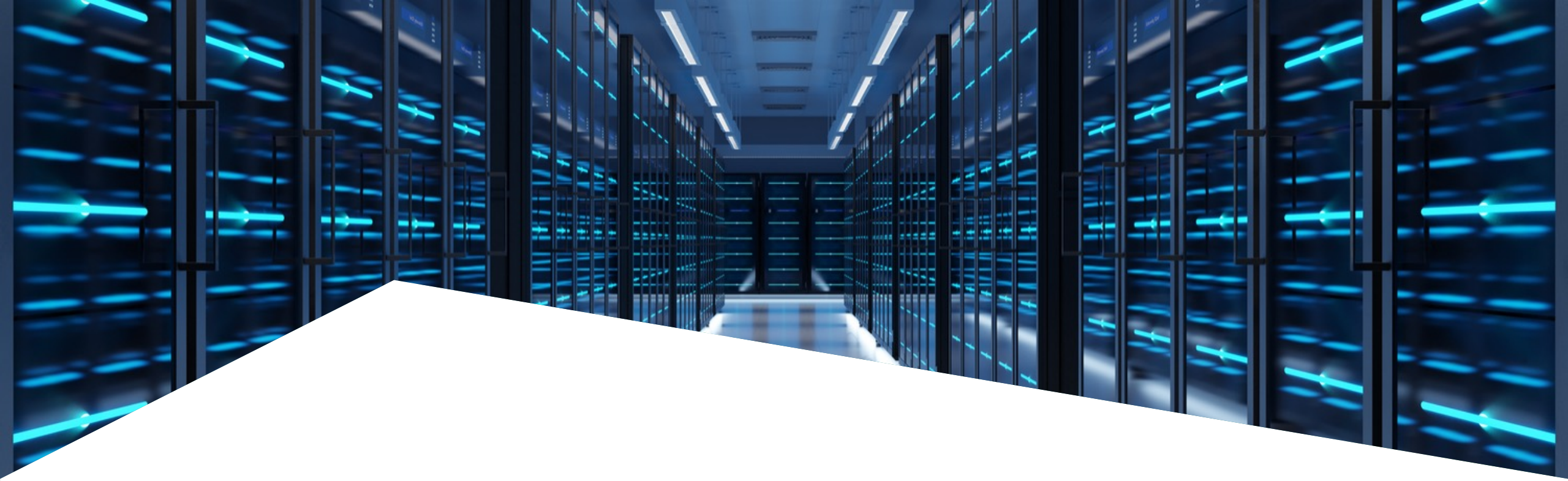


Semiconductor Industry Association
Power Management for efficient AI

Athar Zaidi
Senior Vice President Power ICs and Connectivity Systems at Infineon Technologies
May 2nd 2024

AI is a transformational technology

Every

3.4months

doubling of the amount of computing power required to train cutting-edge AI models since 2012

5days

Time it took for ChatGPT to reach 100 million users

$196bn

Value of the global AI market

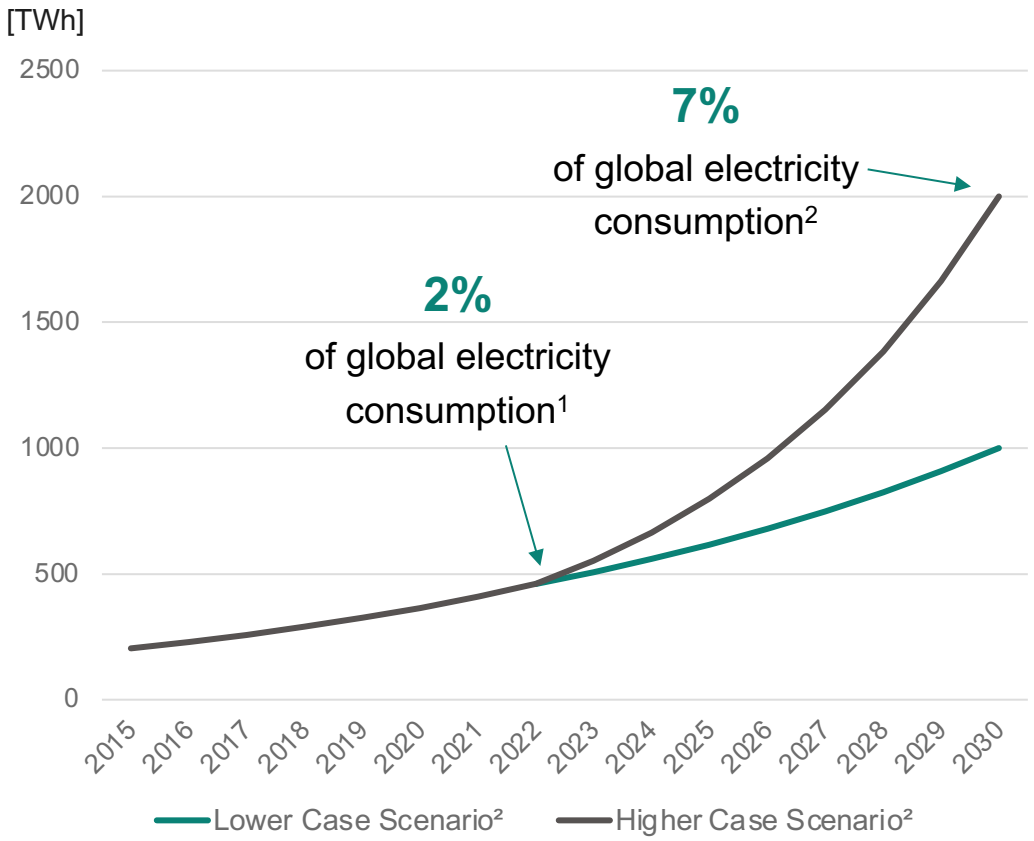
Already now 77% of the global population uses AI

Artificial Intelligence is here to stay

Sources: BMZ; Similiarweb, openAI

AI accelerates power demand in data centers, increasing the need for energy efficient solutions

Projected electricity consumption of data centers[1,2]



Sources

1 IEA; including crypto mining energy use – 2015-2022

2 Infineon assumption and calculation

3 McKinsey

Data centers' share of global final electricity demand was **2%** in 2022[1].

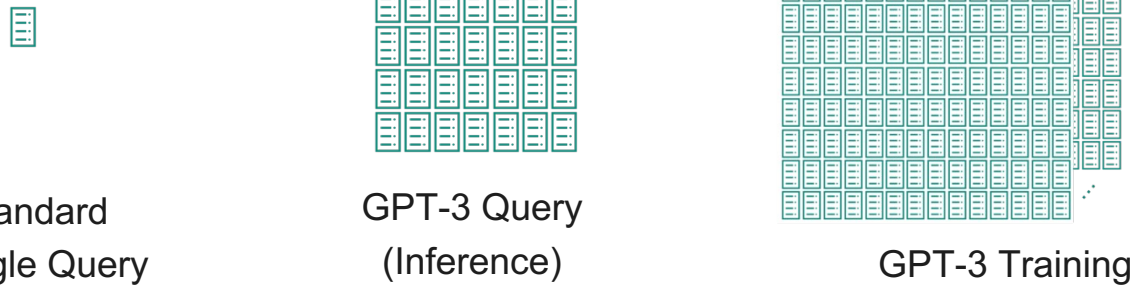
Expected to increase to **7% until 2030**[2], which corresponds to the electricity consumption of **India**.

Example US: **power consumption per Data Center** is forecasted to **grow by 10% a year until 2030**[3].

Generative AI exponentially increases electricity demand

Computing power and electricity demand in generative AI vs. a Google[1] query

	Computing power [Terra FLOPs[2]]	Energy demand [(Mega)Watt Hours]
Standard Google Query	<1 TFLOPs[2]	<1 Wh
GPT-3 Query (Inference)	350 TFLOPs[2]	1-10 Wh
GPT-3 Training	$3x10^{11}$ TFLOPs[2]	>1,000 MWh



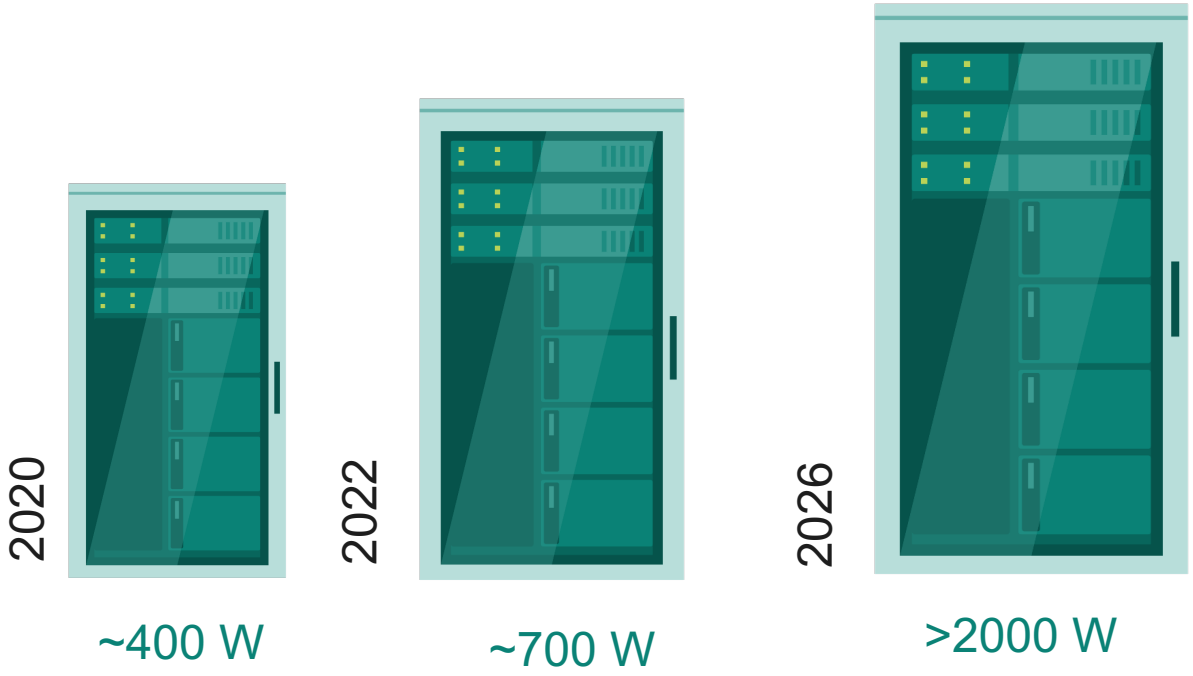
Power supply of an existing data center is limited in the medium term

Focus is required **on powering AI energy efficiently**, **w/o compromising** on **robustness** and **TCO**

Sources: Company information; Statista **1** Google BERT algorithm **2** (Tera=10^{12}) Floating Point Operations Per Second

Efficient AI is a multidimensional problem- Power management cannot be an after-thought

Exemplary development of power consumption of processors under maximum theoretical load



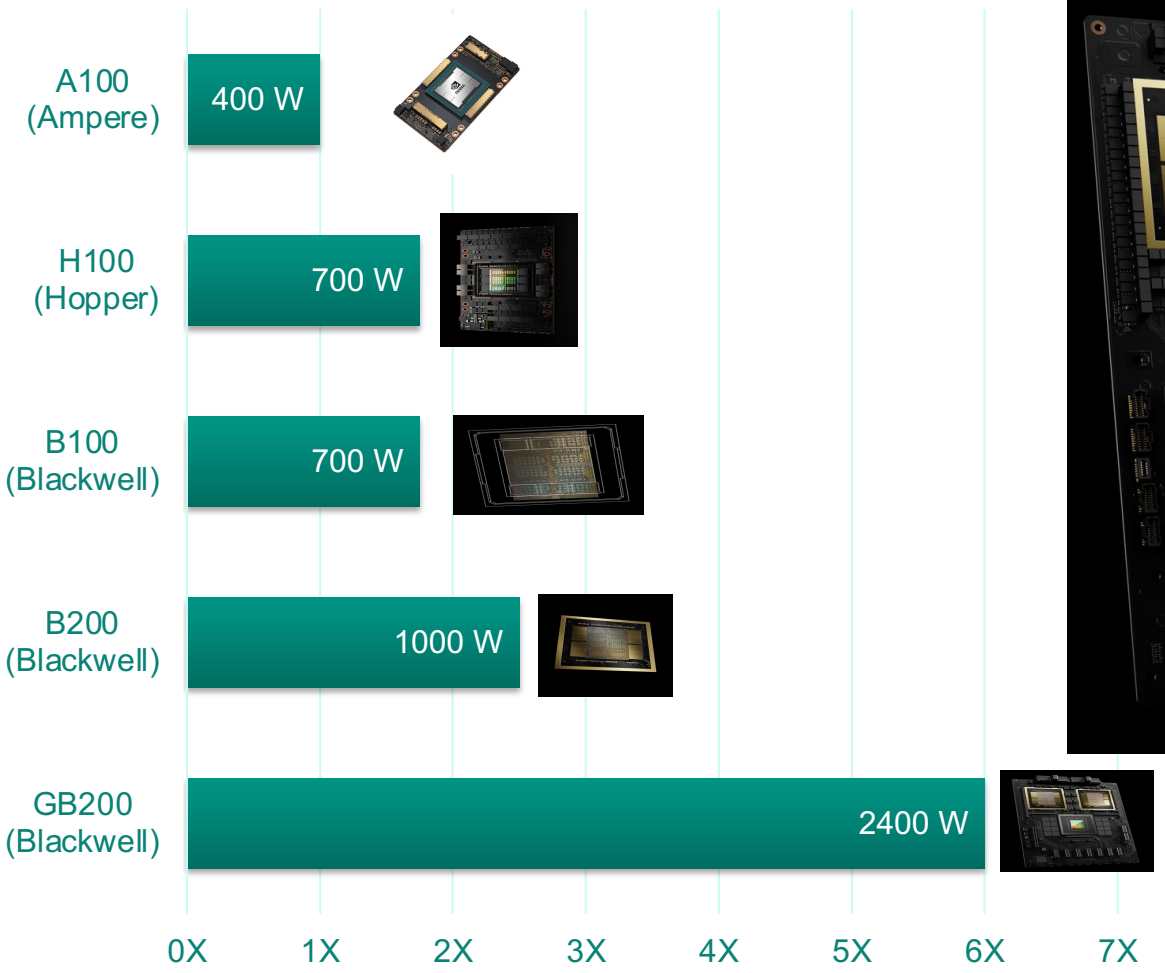
Concerns are emerging regarding the impact of escalating energy requirements linked to **newer chip technologies.**

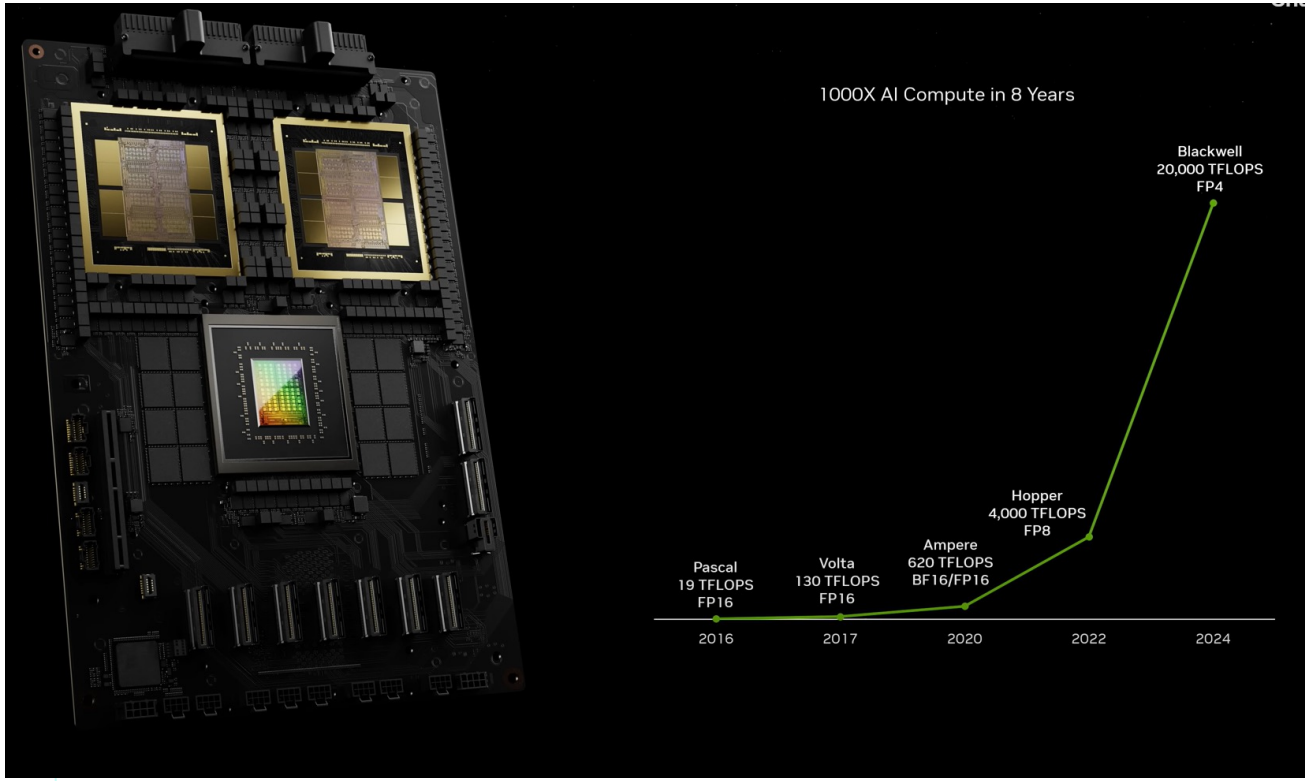
Increasing compute is only one side of the coin, efficient power consumption being the other side.

We need to **prioritize increasing power efficiency** now to reduce the drain on the grid.

Brute force power for AI could break the grid

Exponential rise in power consumption for Nvidia GPUs



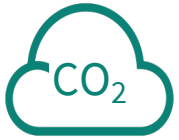


Challenges we can address by focusing on powering AI data centers more efficiently

Environmental impacts

Drain on the grid

We expect data centers' share of global final electricity demand to be 7% in 2030. Especially for data center hubs like the US this could pose a challenge.



Carbon footprint

Running AI servers is an energy-intensive process with a significant carbon footprint.

Water consumption

Around 50% of the energy consumed by data centers goes into cooling. The most common cooling systems run on chilled water or traditional air conditioning.

E-waste

E-waste from AI servers contains hazardous chemicals (i.e. lead, cadmium) that can contaminate the environment.

Sources: Infineon, Earth.org, Study Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Infineon improves current existing solutions at all fronts to increase power efficiency and robustness and minimize e-waste

Innovation fronts to improve how we power AI

- **Rearchitecting** power from the grid to the core- **48V** systems, vertical power delivery
- Designing both **Silicon and wide-bandgap** based efficient power supplies
- Make use of **advanced packaging** for **density and cooling**
- Enable **smart control & software**

Improve energy efficiency at least by 8-10%

Increased power density by 30-60%

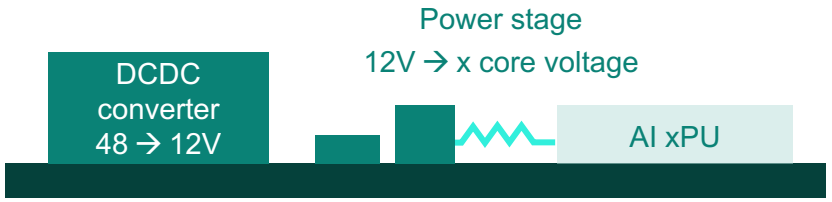
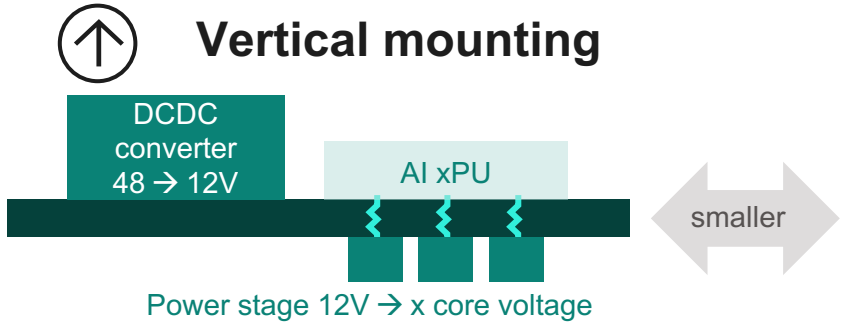
Best-in-class robustness

Best-in-class TCO

22 million metric tons CO_2 equivalent could be saved by using Infineon products in all data centers worldwide

Example: How does this look on a product level?

Infineon power modules on the AI accelerator card, powering the xPU

Power Delivery Network	Discrete solution Equivalent discrete solution	Power Design **Module solution** Infineon's dual-phase modules with inductor-on-top design
Lateral mounting 	Standard discrete down solution **10%** of input power lost[1]	• **2%** efficiency savings[3] • an up to **30%** more powerful processor[4]
Vertical mounting 	*Not applicable*	Only Infineon offers this combination with best-in-class **energy efficiency, power density & TCO** Backside mounting • **< 2%** of input power lost[2] Dual-phase module with inductor-on-top design • **2%** energy efficiency savings[3] • an up to **30%** more powerful processor[4]

[1] in motherboard interconnections through lateral mounting
[2] in motherboard interconnections through vertical mounting
[3] using Infineon's dual-phase modules with an inductor-on-top design compared to an equivalent module solution
[4] can be supplied within the same area through an up to 30% reduction of the occupation area enabling a current density increase


Infineon