We power AI
Online Media Briefing

Infineon Technologies AG

Power and Sensor Systems Division

June 2024

public

Athar Zaidi, Senior Vice President & General Manager Business Line, Power ICs and Connectivity Systems

Athar Zaidi

was born in India in 1976.

He holds dual Masters in Mechanical and Computer Engineering from North Carolina State University and an MBA from University of California at Berkeley.

2001 - 2006	2006 - 2010	2010 - 2012	2012 - 2016	2016 - 2023	Since 2023
Various position in Analog Design at Semtech Corporation.	Business Management at Maxim Integrated	Director of Business Management at Semtech Corporation	Vice President at Dialog Semiconductor	Senior Vice President & General Manager Enterprise Power at PSS DCDC	Senior Vice President & General Manager Business Line, Power ICs and Connectivity Systems PSS PCS

Table of contents

AI is a transformational technology

Every

3.4 months

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

5 days

Time it took for ChatGPT to reach 100 million users
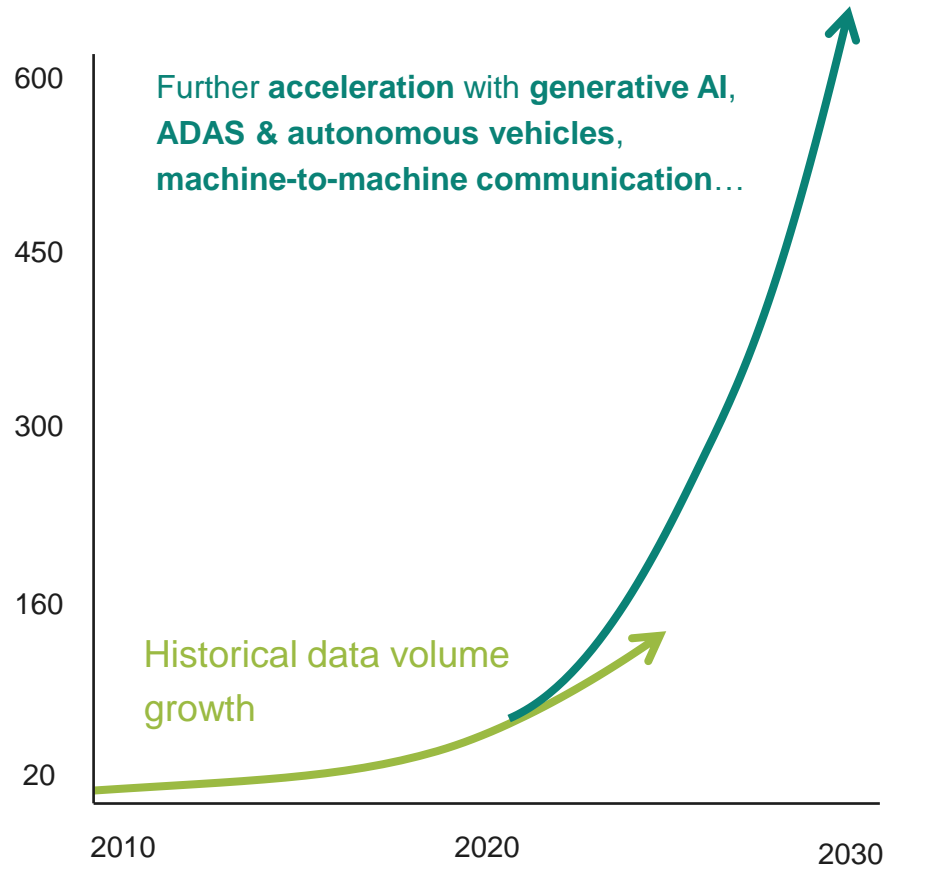
$196 bn

Value of the global AI market

Already now 77% of the global population uses AI
Artificial Intelligence is here to stay
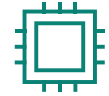
Sources: BMZ; Similiarweb, openAI

Exponential growth in global data and higher electricity demand of AI enabled systems increase the need for energy efficient solutions
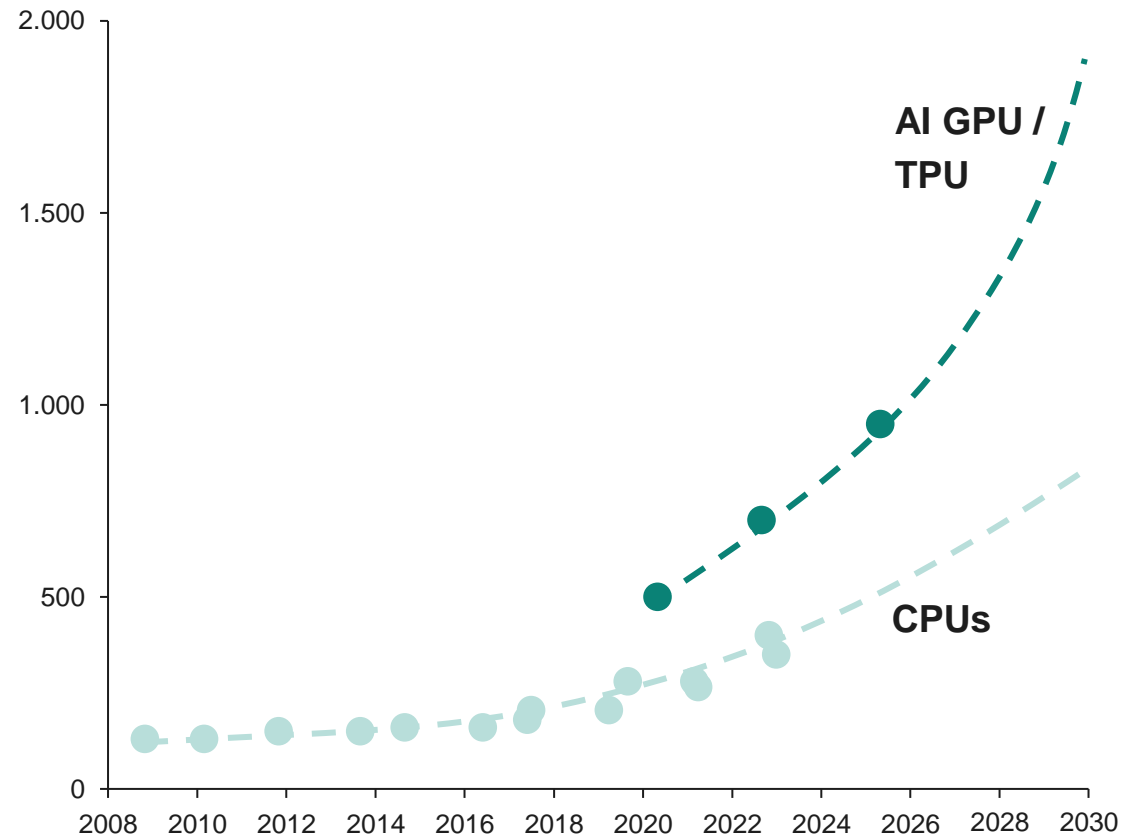

Global data volume in zettabyte (per year)

Further **acceleration** with **generative AI**, **ADAS & autonomous vehicles**, **machine-to-machine communication**…

Historical data volume growth

1 ZB = **1.000.000.000.000 GB**

Single processor electricity demand over time [Watt]

AI GPU / TPU

CPUs

Release date

Sources: Omdia; Company information; Infineon

Table of contents

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

Projected electricity consumption of data centers[1,2]

7%
of global electricity consumption[2]

2%
of global electricity consumption[1]

[TWh]

— Lower Case Scenario[2] — Higher Case Scenario[2]
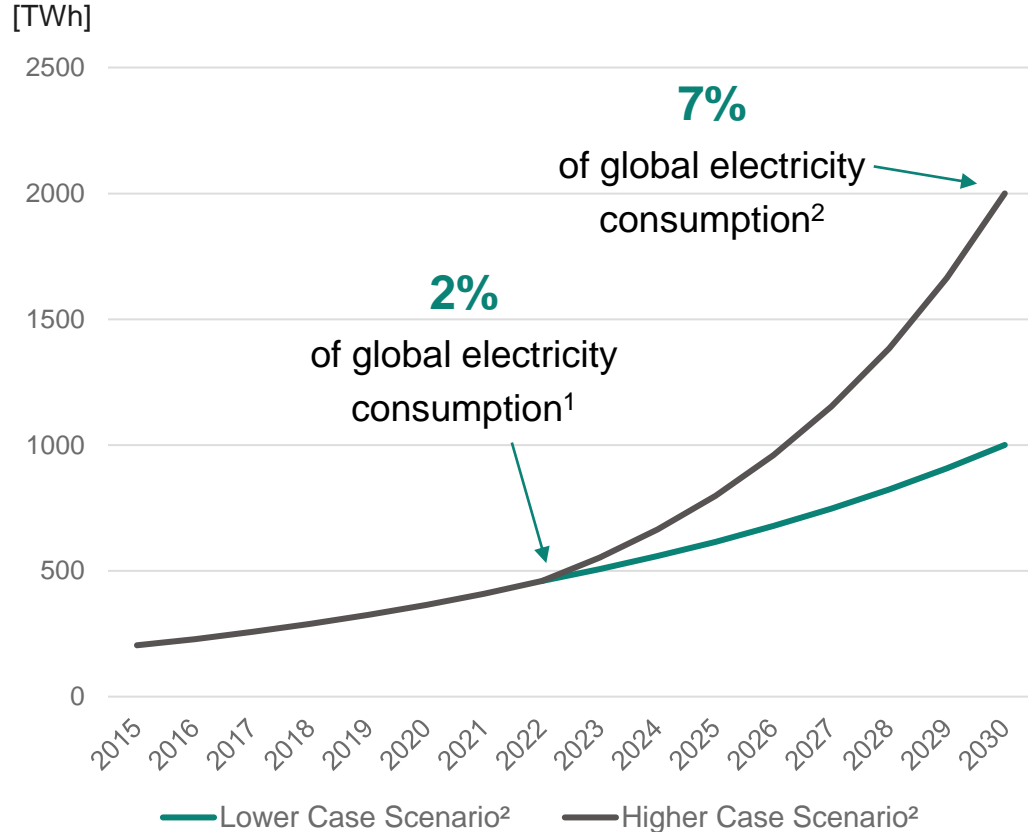
Sources
1 IEA; including crypto mining energy use – 2015-2022
2 Infineon assumption and calculation
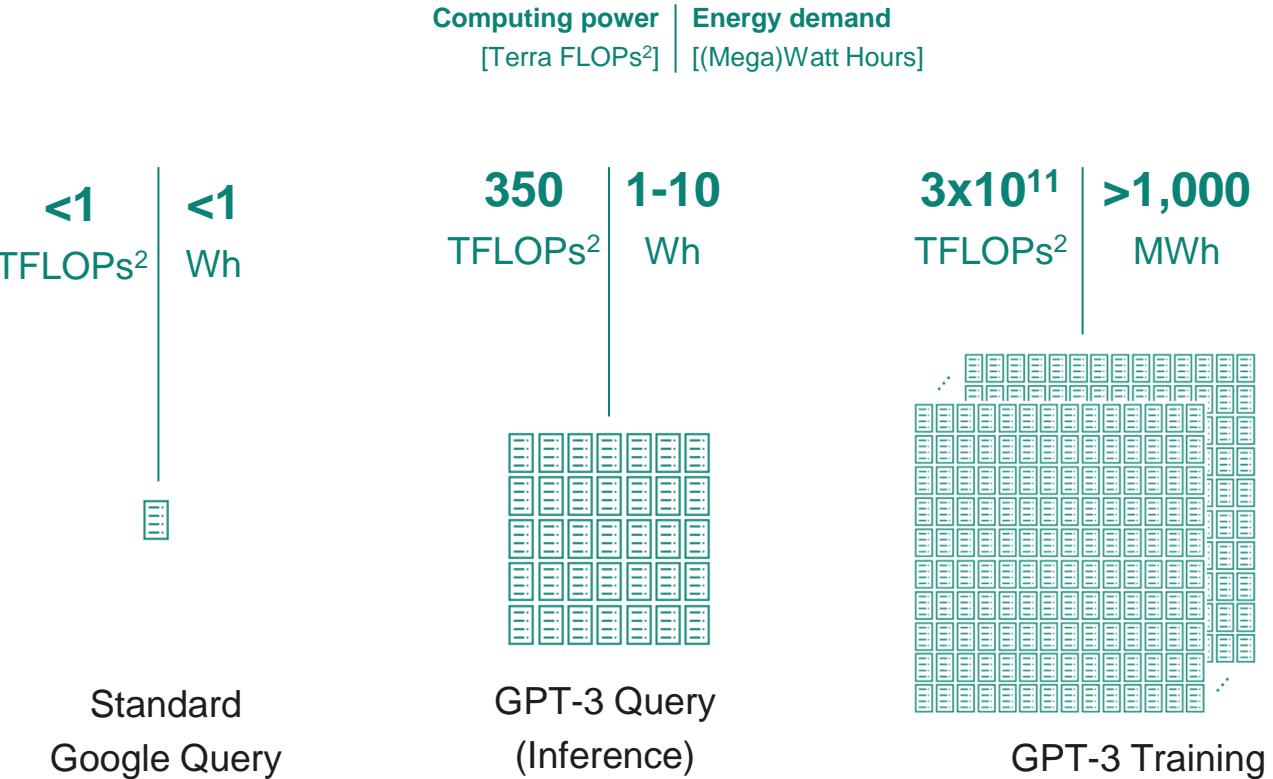**3** McKinsey

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]

<1
TFLOPs[2]

<1
Wh

Standard
Google Query

350
TFLOPs[2]

1-10
Wh

GPT-3 Query
(Inference)

$3x10^{11}$
TFLOPs[2]

>1,000
MWh

GPT-3 Training

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 multi-dimensional problem
Power management cannot be an after-thought

Exemplary development of power consumption of processors under maximum theoretical load

2020
~400 W

2022
~700 W

2026
>2000 W
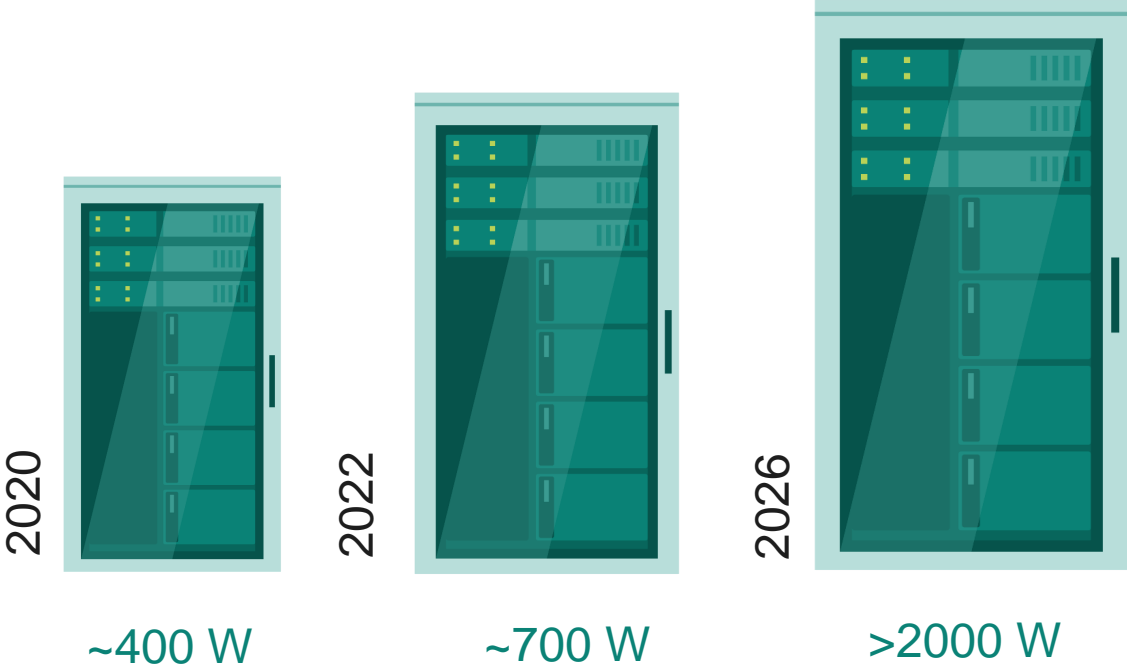
Concerns are emerging regarding the impact of escalating energy requirements linked to **newer chip technologies.**

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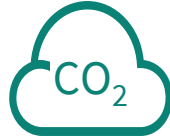
Challenges we can address by focusing on powering AI data centers more efficiently

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.

CO_2

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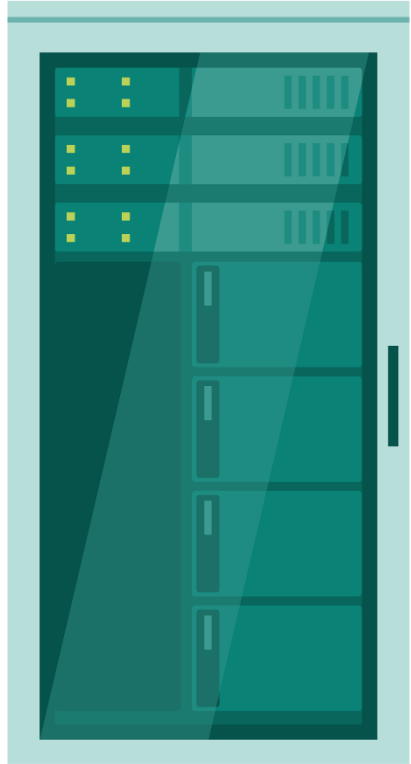
Sources: Infineon, Earth.org, Study Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models
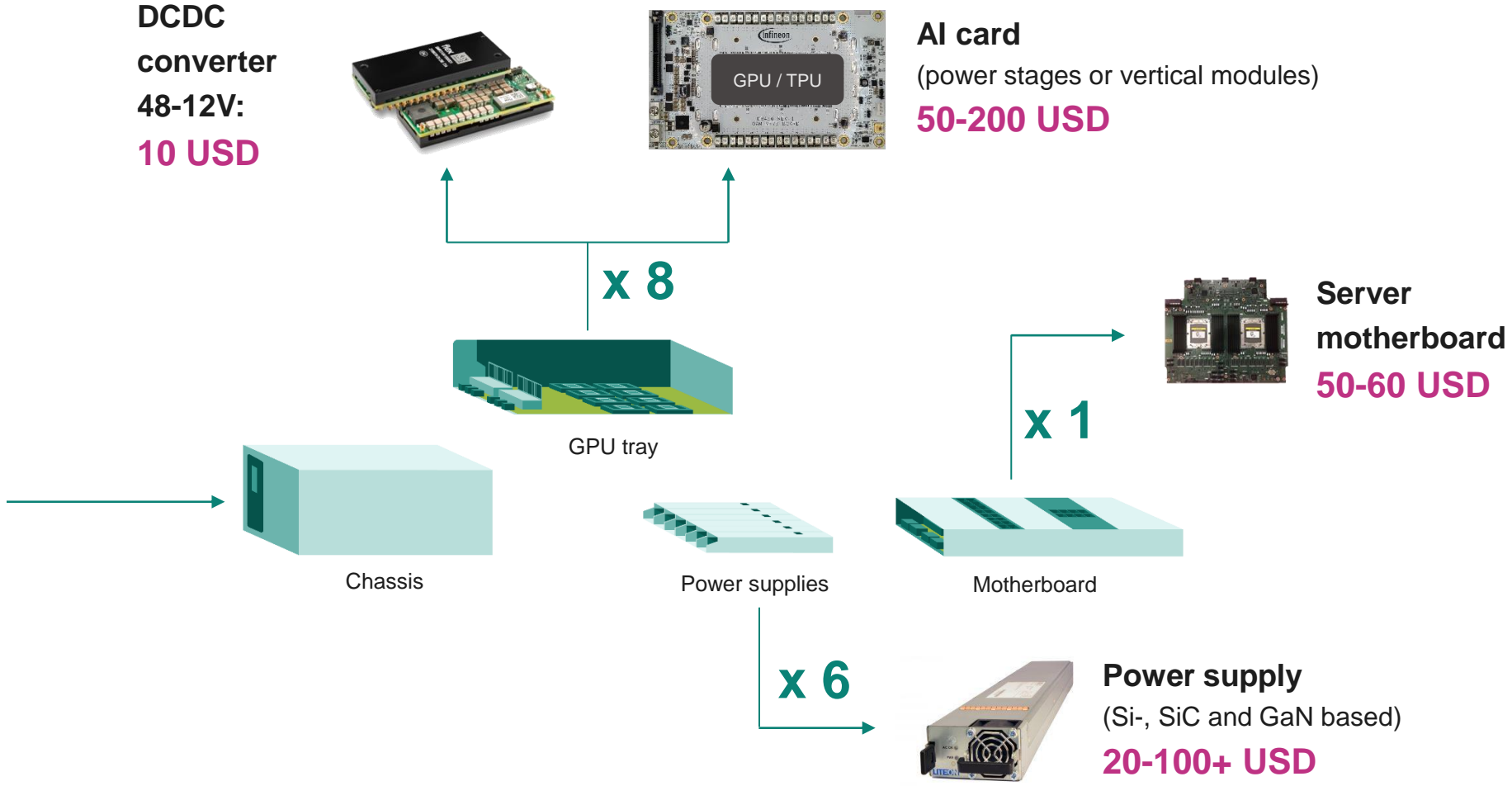
Table of contents

We power AI from the grid to the core

Average Infineon BOM per AI server about 850 to 1800 USD

DCDC converter 48-12V: **10 USD**

AI card (power stages or vertical modules) **50-200 USD**

x 8

GPU tray

Server motherboard **50-60 USD**

x 1

Chassis

Power supplies

Motherboard

x 6

Power supply (Si-, SiC and GaN based) **20-100+ USD**

1 Server rack includes
4 AI servers

USD = potential Infineon content per AI server

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%

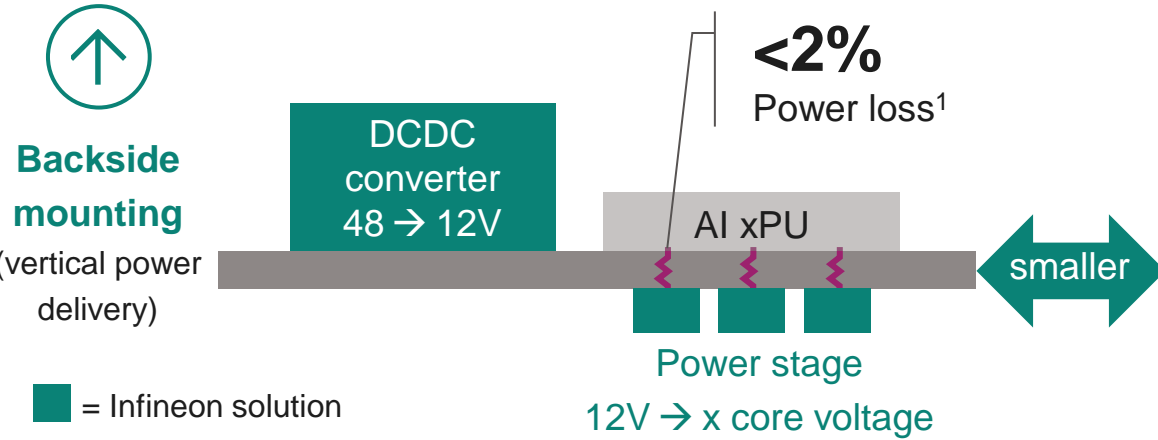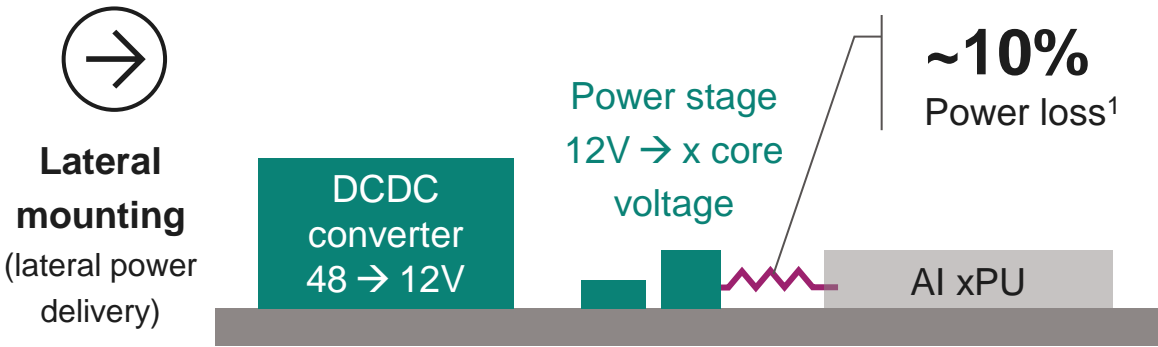Increased power density by 30-60%

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 1: How does this look on a product level?
Infineon power modules on the AI accelerator card, powering the xPU

Lateral mounting
(lateral power delivery)

DCDC converter 48 → 12V

Power stage 12V → x core voltage

AI xPU

~10% Power loss[1]

Backside mounting
(vertical power delivery)

DCDC converter 48 → 12V

AI xPU

<2% Power loss[1]

smaller

Power stage 12V → x core voltage

■ = Infineon solution

Customer benefits of vertical power delivery enabled by Infineon's vertical power modules

Increase power density via smaller size to enable further increase in compute power

Reduce power losses by >7MW for an average data center (100,000 CPU nodes)

>12% total cost of ownership saving compared to lateral power delivery networks

Source: Infineon calculation **1** Power delivery loss in % of xPU power

Example 2: How does this look on a product level?
For AC/DC, Infineon is addressing the growing power demand of AI

Power Supply Unit (PSU) solutions ranging from 3 kW – 12 kW

12 kW
- η ~97,5%
- 100 W/in³

Si   SiC   GaN

8 kW
- η ~97,5%
- 100 W/in³

Si   SiC   GaN

3.3 kW
- η 97,5%
- 95 W/in³

Si   SiC   GaN

3 kW
- η 97,5%
- 32 W/in³
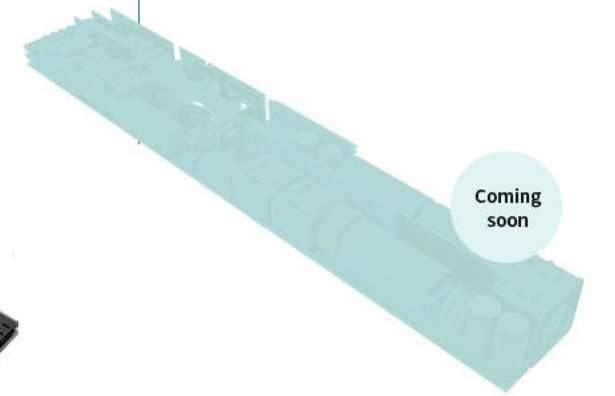
Si   SiC

Available now

Available now

Available in Q1/25

Coming soon

AI will not only be a strong driver of revenue for Infineon's server business but offers also an important lever to drive decarbonization

We have a clear differentiation potential…

⟩ **System innovation** with **leading companies**

⟩ Industry-leading **system and innovation expertise**
(i.e. digital power control, chip embedding, mastery of Si, SiC and GaN power technologies, our novel vertical power supply architecture)

⟩ Best-in-class in **efficiency** and **lowest cost of ownership**

Expected FY24 AI server revenue

Low **triple digit million** amount
(~1bn expected in the next few years)

CO_2 saving potential by using Infineon products in AI servers

22 million metric tons CO_2

Table of contents

Summary
We power greener AI, shaping the future with our solutions

AI is the next technological revolution and characterized by
an **astonishing pace of adoption**[1]

Behind the brilliance of AI lies an **energy-intensive** process with a substantial **carbon footprint**[2]

We are at the forefront of innovation offering energy-efficient and robust solutions **powering AI from the grid to the core!**

[1] Omdia
[2] IEA

Hungry for more information?
Listen to our new "We power AI podcast"

Dive into facts, challenges and innovations to power the brilliance of AI

Introduction to powering AI

Adam White, President Power & Sensor Systems

The power of efficiency: the sweet spot for Si, SiC, and GaN in data centers

with Gerald Deboy, Head of Innovation lab

Upcoming

The power struggle: tackling AI data centers' energy requirements

Athar Zaidi, Senior Vice President & General Manager Power ICs and Connectivity Systems

The importance of robustness, high efficiency & quality in data centers

with Danny Clavette, Distinguished System Architecture Engineer

Upcoming

Optimizing AI server power flow with 48 V architectures and vertical power delivery

with Carl Smith, Global Senior Director Datacenter & Computing Application

Upcoming

48 V architecture: topologies, benefits and applications

with Roberto Rizzolatti, Principal Engineer

Upcoming

... or visit our Infineon promo page