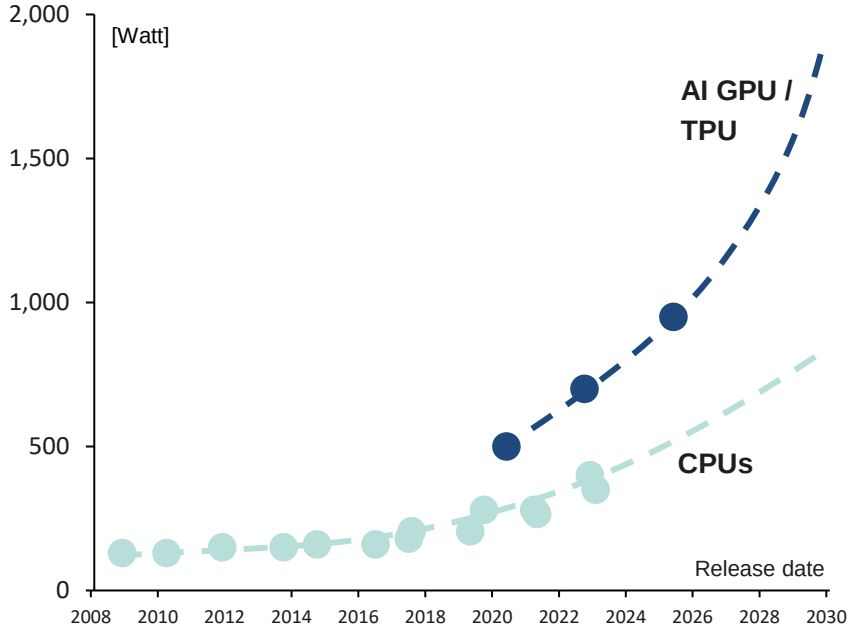
Improving cloud computing and AI power delivery efficiency with a holistic semiconductor approach

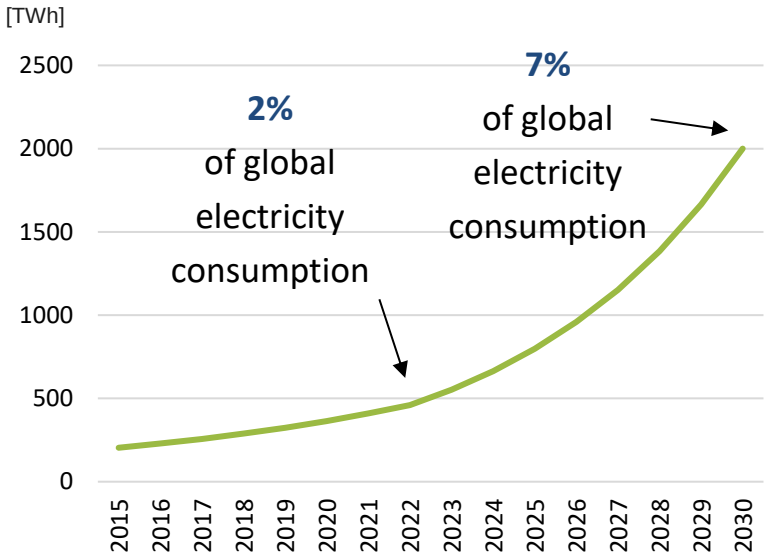
Thomas Neyer, Infineon Technologies AG

Exponential growth in global data center power consumption

Single processor electricity demand over time



Projected electricity consumption of data centers

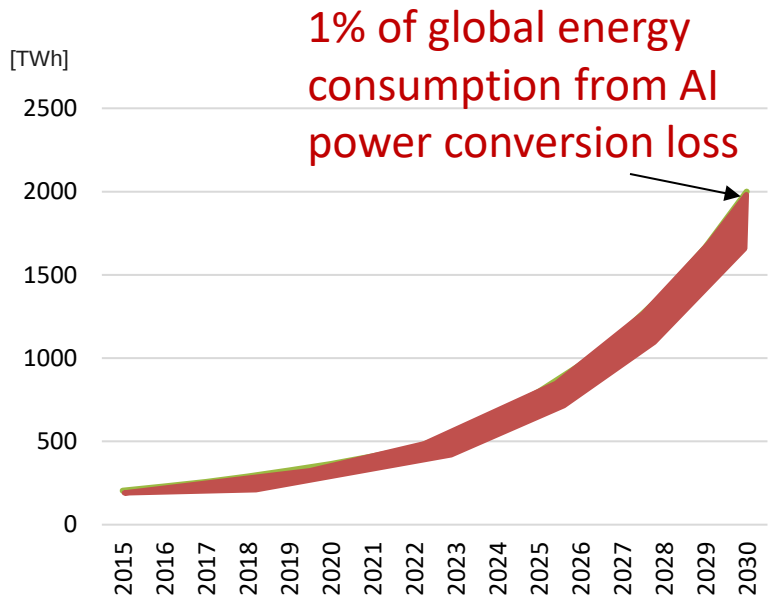


Sources: IEA including crypto mining energy use; Infineon assumption and calculation; McKinsey, BCG

What can we do about it?

- **15% total loss in power delivery chain**
- To offset only the power conversion loss with green energy would require:
  - **400 million PV panels**
  - 1000 sq. km (350 sq. mi.) solar farm in a high-irradiance area, e.g. Arizona
- If the power is generated from natural gas:
  - Losses will produce **0.5 billion metric tons of carbon dioxide** over 5 years
  - Equivalent to adding 20 million passenger cars to the road
- Saving just **1% efficiency** anywhere in the power delivery chain will reduce carbon emissions by **35 million tons**, equivalent to 1.4 million passenger cars

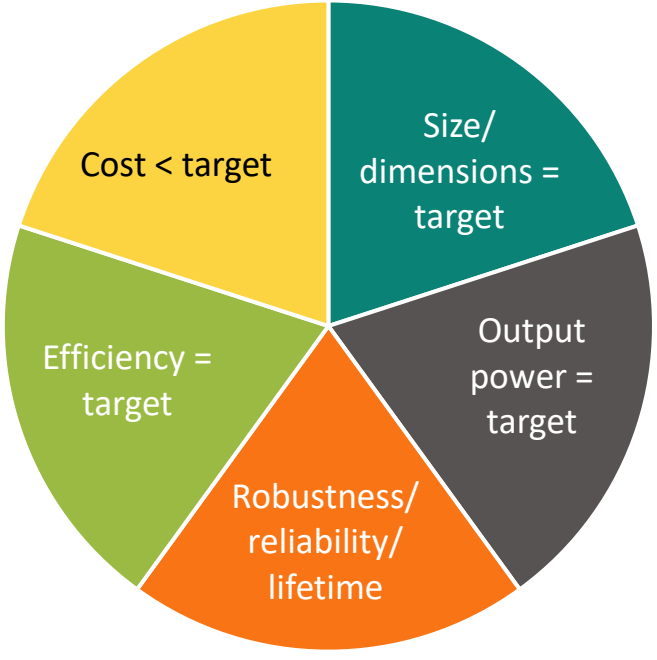
Projected electricity consumption of power loss in data center power delivery chain



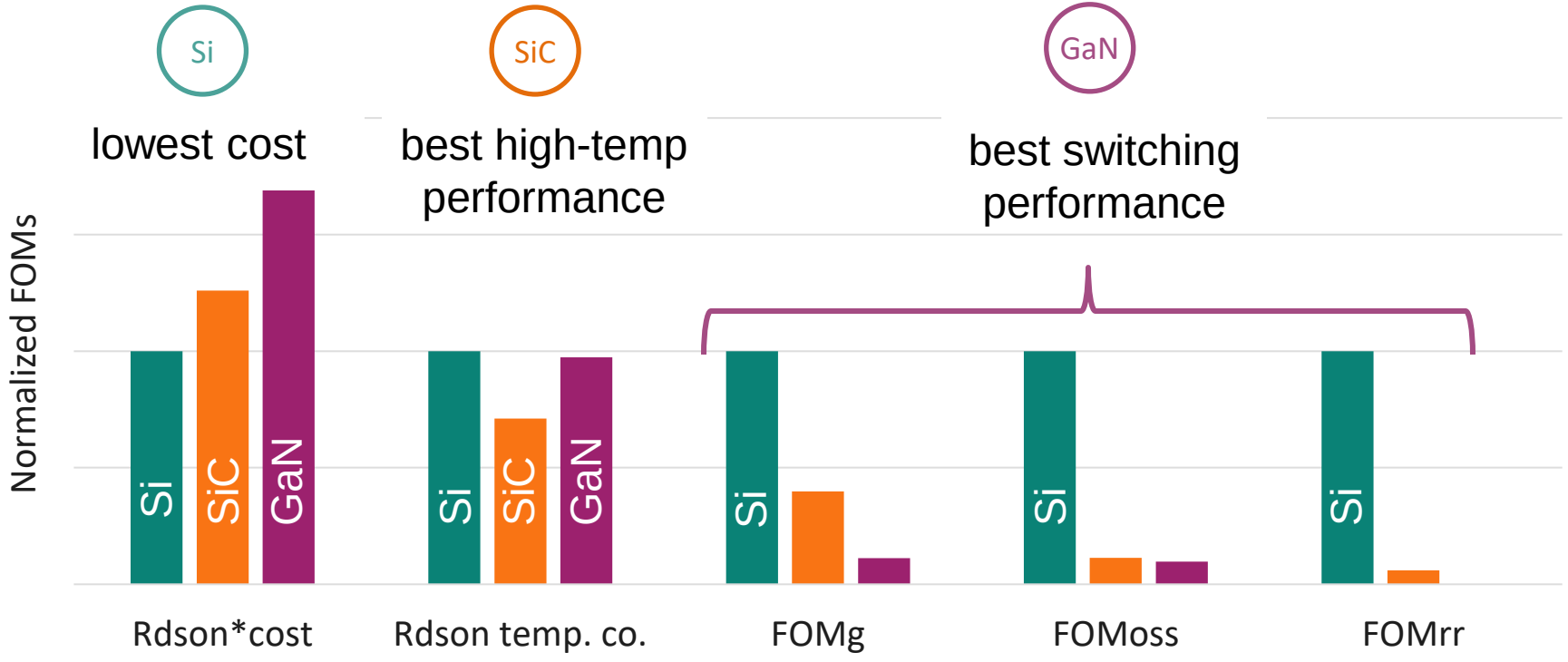
[Assuming 85% end-to-end efficiency for all years]

Design criteria in data center power systems

- Size, power, lifetime, and efficiency must meet minimum requirements/targets
- Efficiency improvement beyond the target is fine, but a lower cost alternative might be selected instead
- Highest efficiency solution doesn't always get implemented!

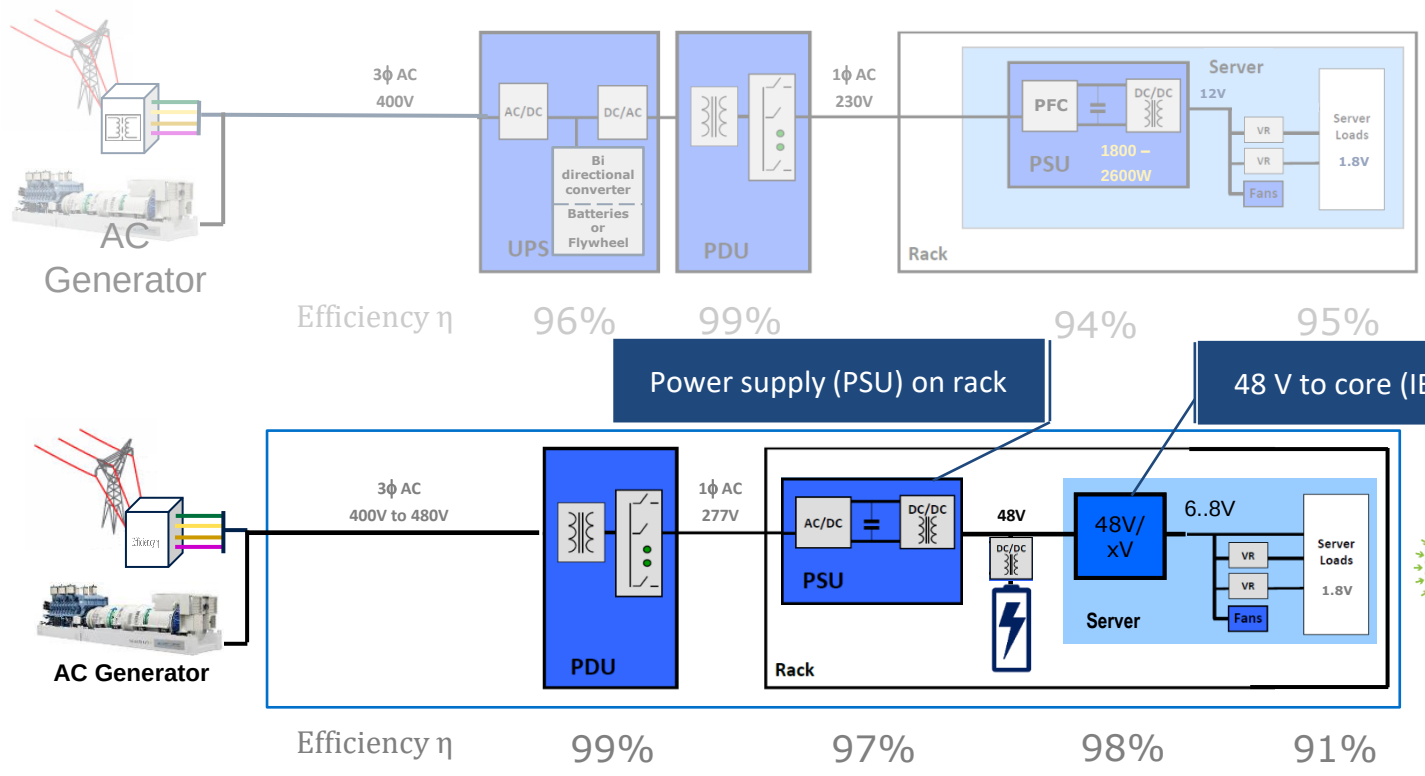


Relative merits of each semiconductor technology with today's state-of-the-art



Reality tomorrow? More on that later!

Data center power delivery system transitioning from enterprise to hyperscale architecture requires higher power density & efficiency

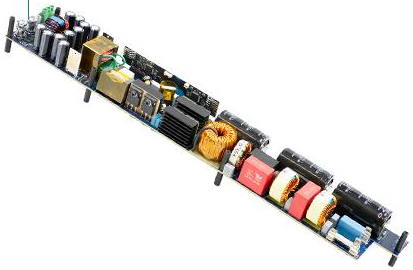


OPEN V3 Compute Project®

Addressing the growing demand of AI with higher power on-rack PSUs, using all 3 semiconductor technologies

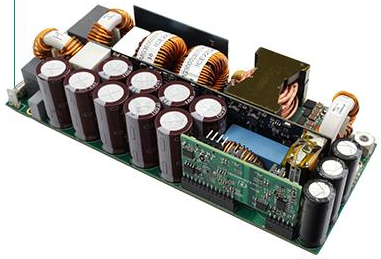
Si SiC

3 kW
η 97,5%
32 W/in³



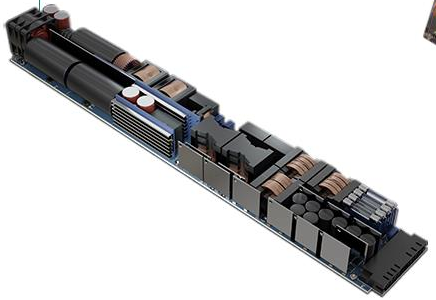
Si SiC GaN

3.3 kW
η 97,4%
98 W/in³



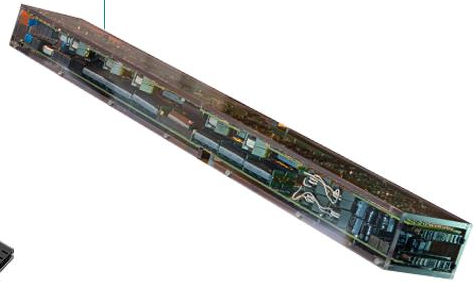
8 kW
η ~97,5%
100 W/in³

Si SiC GaN

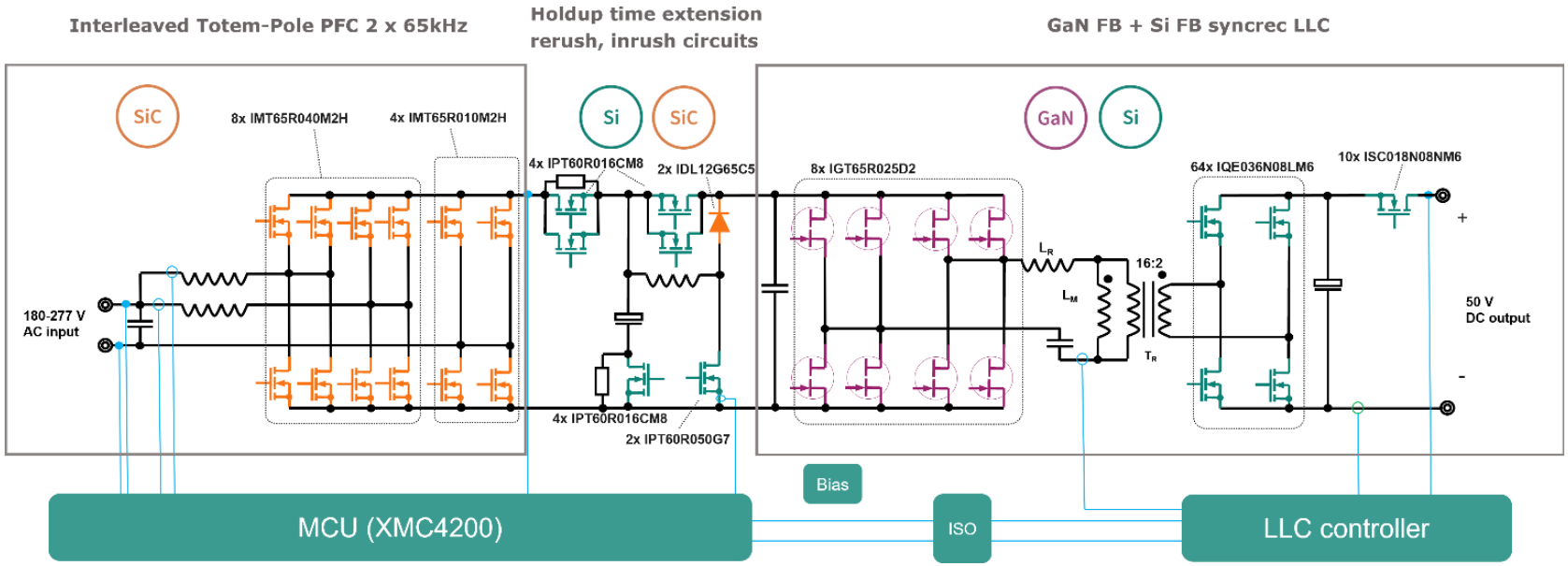


12 kW
η >97,5%
100 W/in³

Si SiC GaN



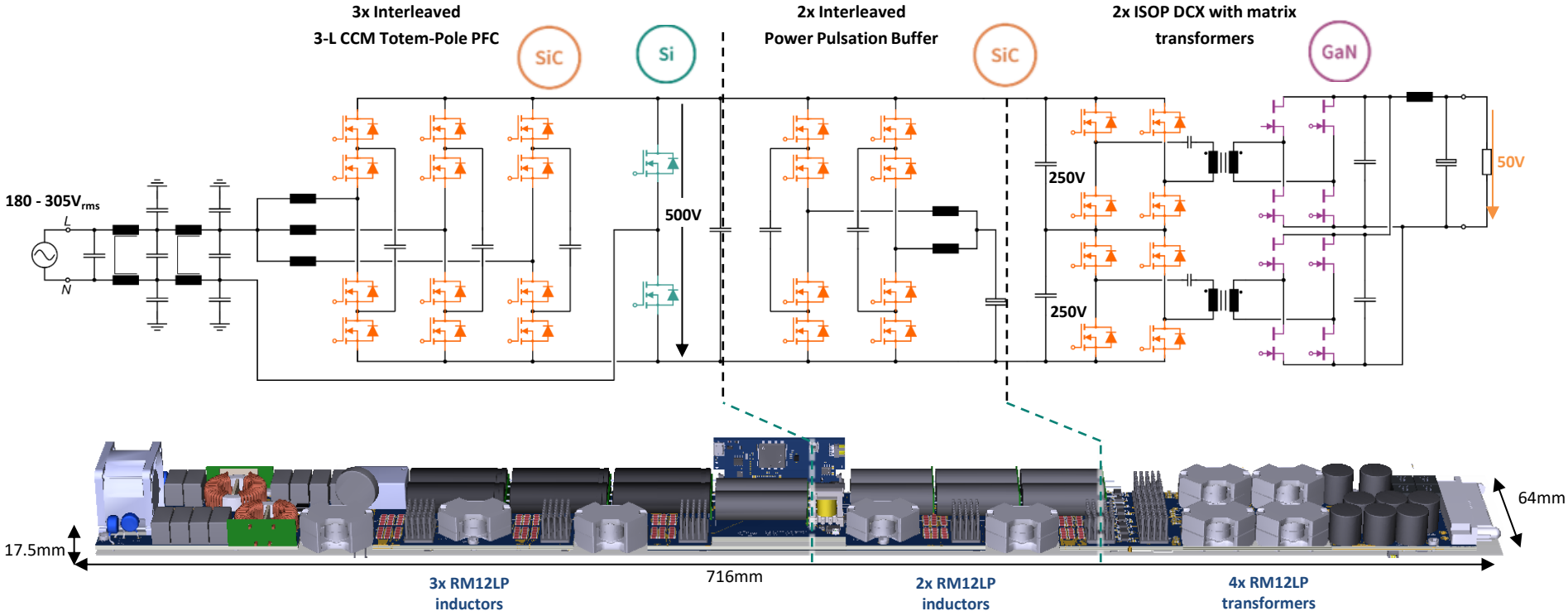
8 kW server/AI PSU with Si+SiC+GaN



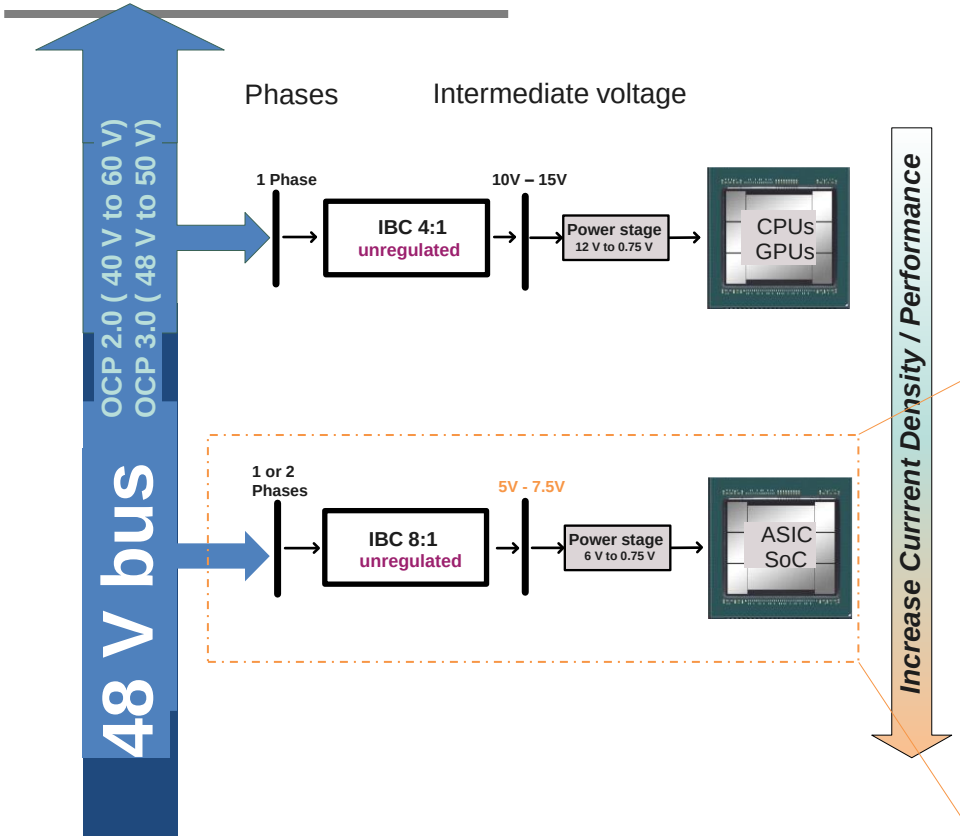
8kW in 73.5x40x446 mm^3
→ **~100W/inch^3**

12 kW server/AI PSU with Si+SiC+GaN

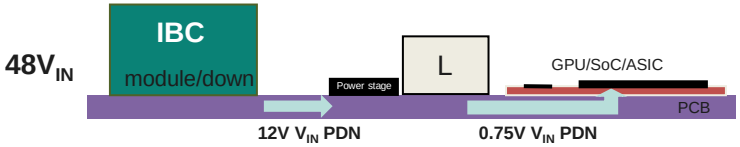
2 x **6 kW modules**, each with ½ U height, stacked vertically to fit 1U max height



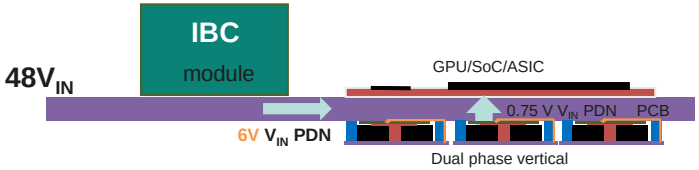
2-stage power delivery with vertical power flow

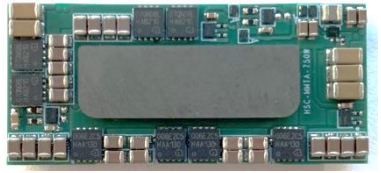


Lateral power delivery

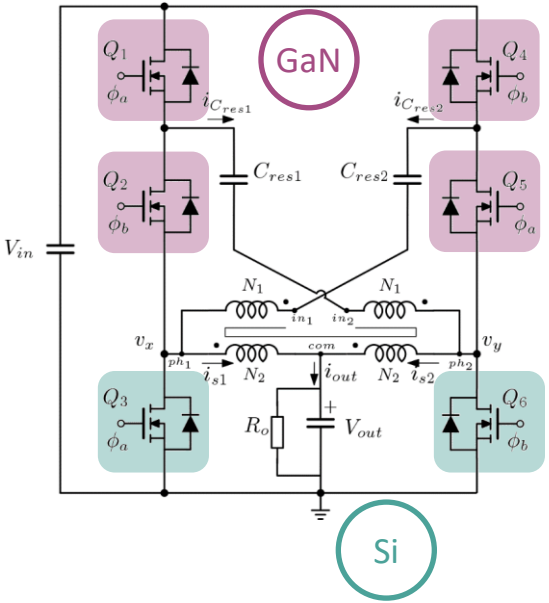


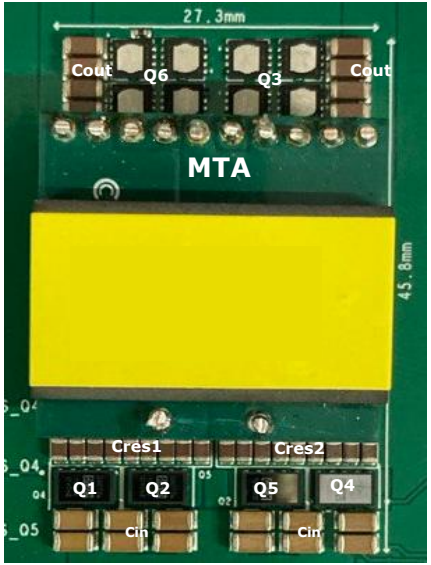
Move from *lateral* to *vertical* power delivery

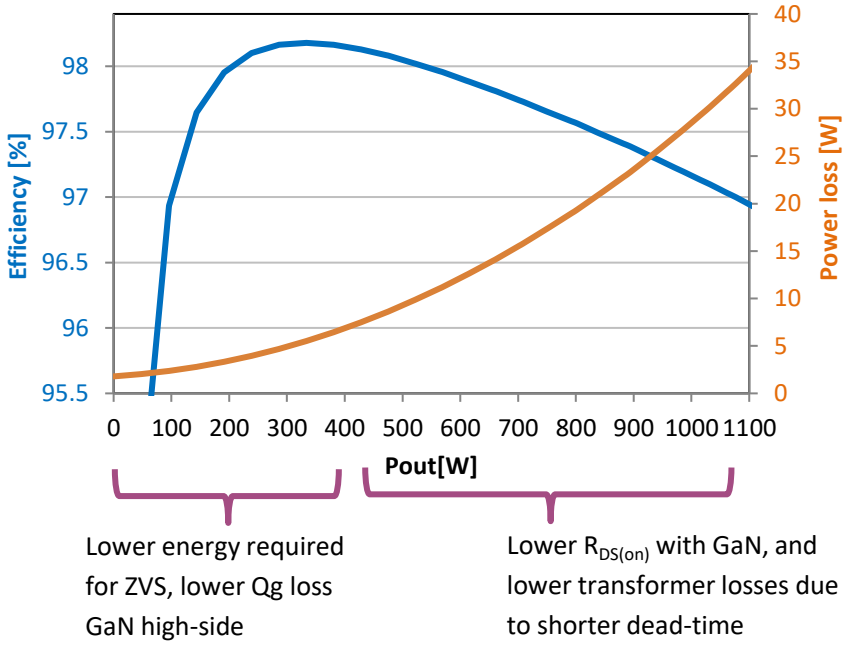




48-to-n V IBC using hybrid switched capacitor topology with GaN + Si

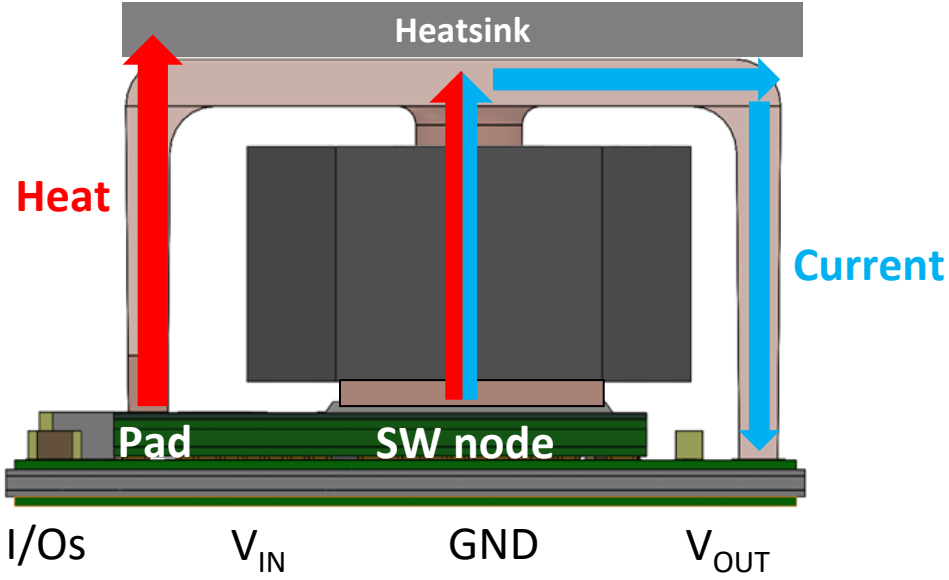


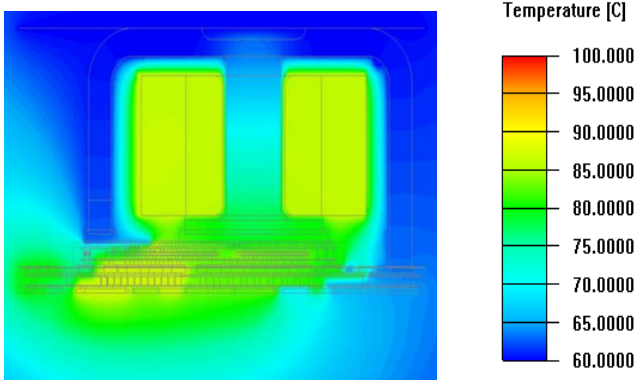




GaN takes IBC efficiency to the next level, due to lower dead-time required for ZVS, dramatically reducing transformer losses as well as transistor loss.

Vertical power delivery to GPU/CPU using chip-embedded Si and novel magnetics design, achieving >92% efficiency

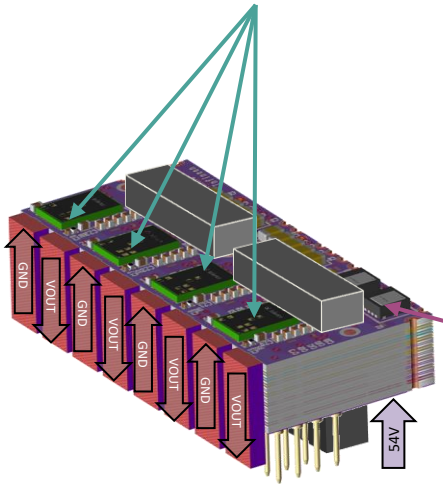




- Chip-embedding LV Si MOSFETs improves EMI and heat extraction
- Inductor has two tasks: **electrical and thermal conduction**, **tightly bound** to each other

Direct ~48 V to core conversion, 300 A_{out} with half-bridge current doubler, using GaN + Si

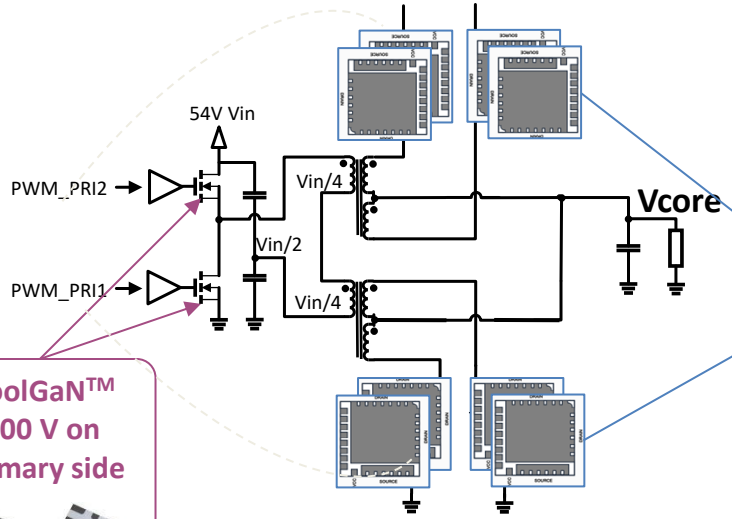
15 Volt Si MOSFETs
Secondary side SR



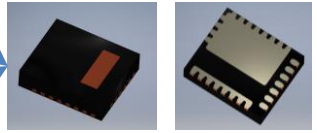
54 V to 0.7 V Module

CoolGaN™
100 V on
primary side

IGB070S10S1



Smart Driver Sync (SDS)
NFETS



- **300 A Module Capability**
- Efficiency 88% at 0.7 V / 300 A
- Combination of GaN HEMT and Si Trench MOSFET Technology

AI training will drive rack power to even higher power, triggering new architectures and topologies

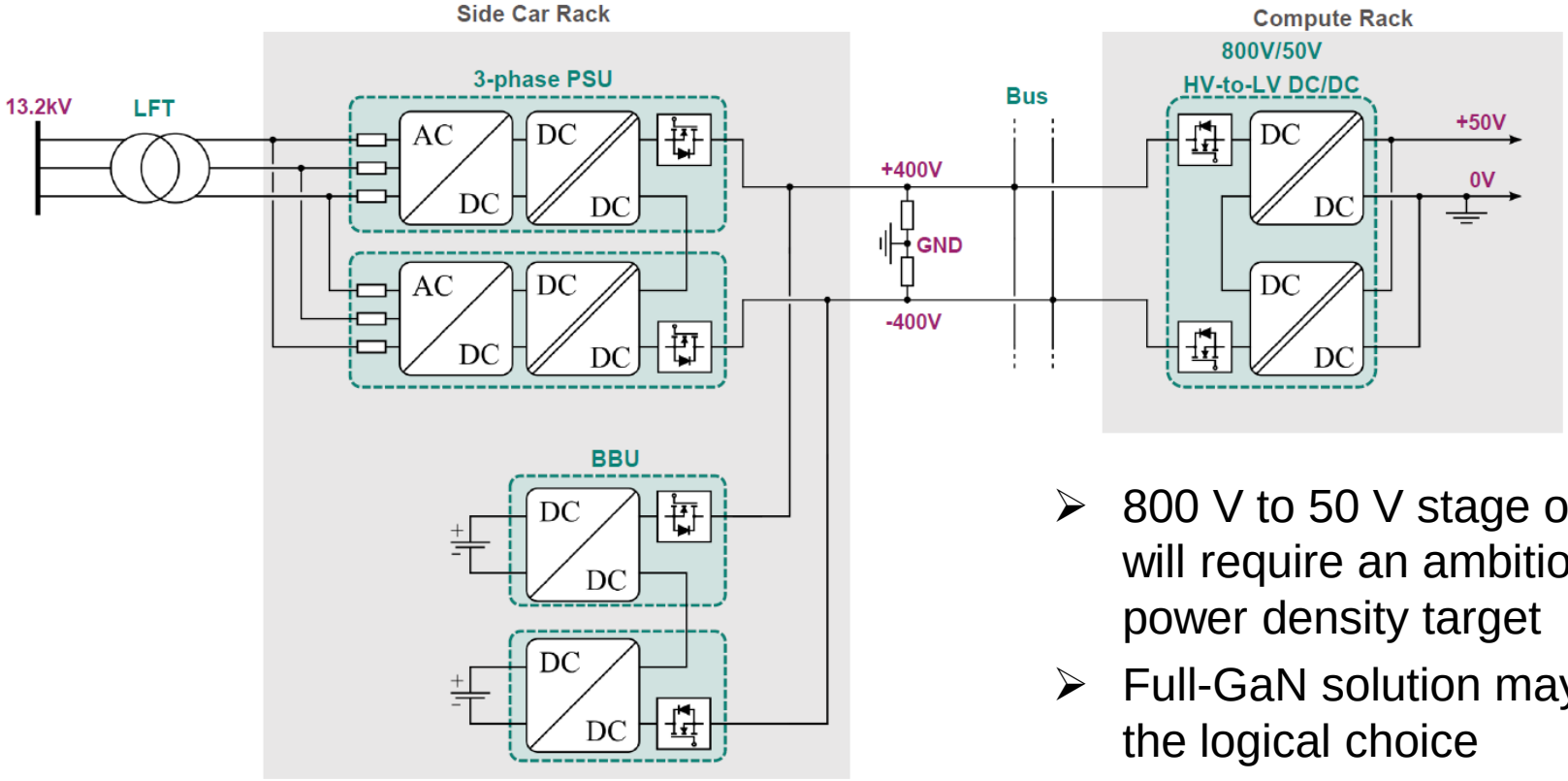
Single-phase AC
230 – 347 V AC / 48 V DC

Three-phase AC
480 – 600 V AC / 400 or 800 V DC

Facility level: SST
HV AC to **HV DC**

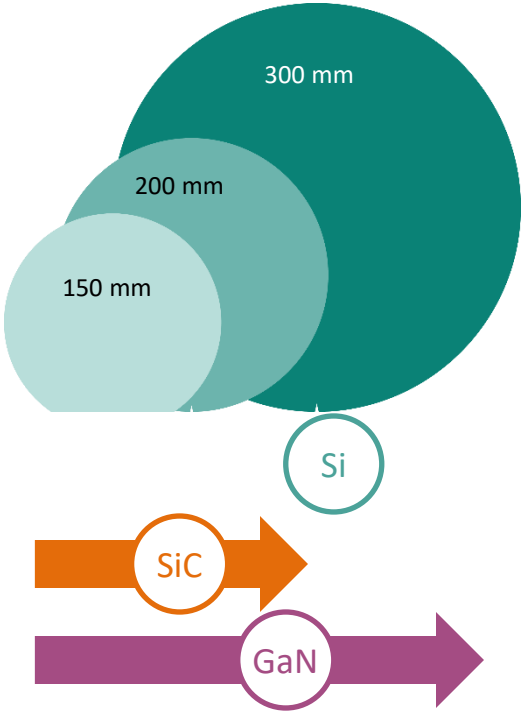
Source: OCP Global Summit, October 2024.

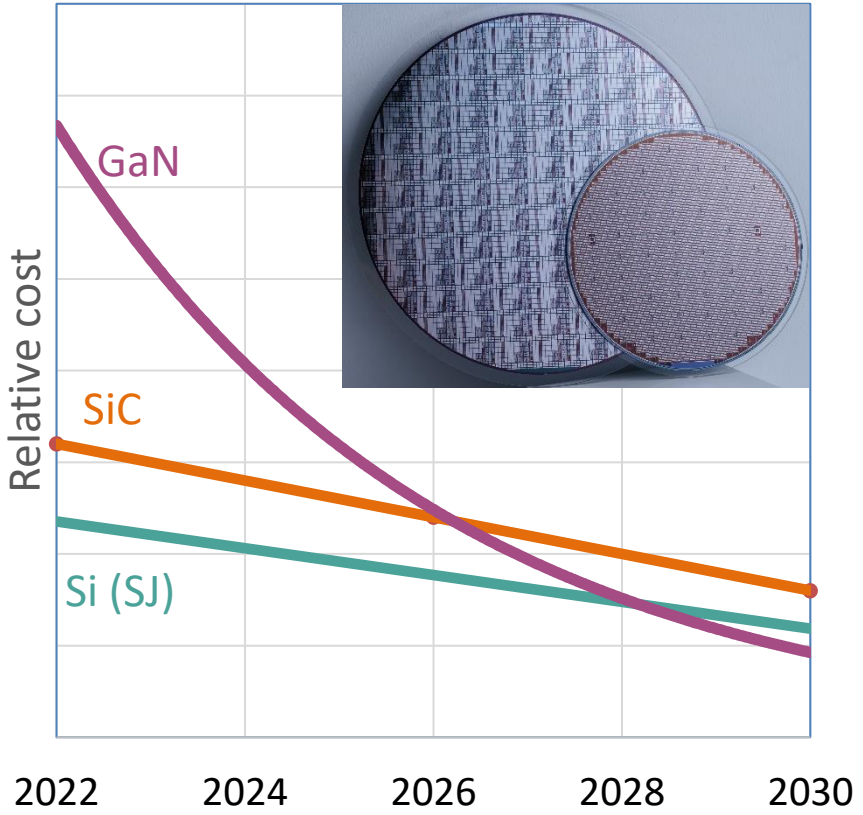
HV IBC driven by 3ph AC input architecture



- 800 V to 50 V stage on rack will require an ambitious power density target
- Full-GaN solution may be the logical choice

Si, SiC, and GaN costs dropping at different rates, approaching cost parity/crossover point by end of the decade





Conclusions

- Our mission is to improve data center power delivery efficiency with practical solutions, and the time is now!
- Holistic approach to semiconductor selection enables the best optimization to meet application requirements
- **Call to action:**
  - Develop solutions that meet the application targets, with some efficiency bonus on top of the minimum requirement
  - Keep an open mind to all three semiconductor
- **Join us to shape a sustainable AI future**

THANK YOU

Thomas Neyer, Infineon Technologies AG