
15. BECHTLE IT-FORUM
THÜRINGEN
2024
BECHTLE
15. Mai 2024 • STEIGERWALD Stadion Erfurt
Hewlett Packard Enterprise
cisco Partner
HUAWEI
intel.

Das Zeitalter der KI

Tasso Kazakidis I Supermico

AGENDA

- Über Supermicro
- Status Quo
- Was kann KI?
- KI auf technischer Ebene
- Begleiter der KI: Storage

DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

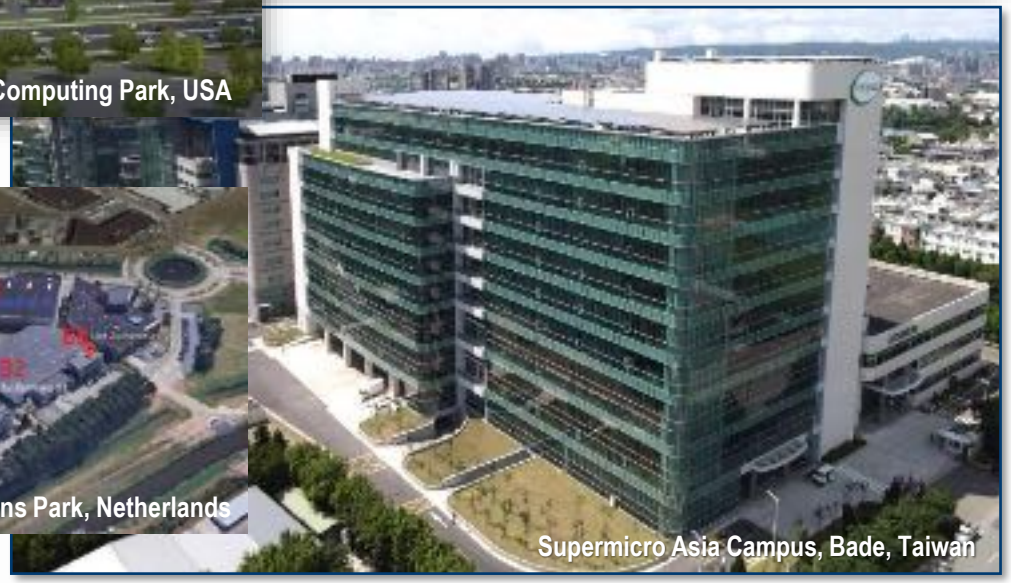
ATTRIBUTION

Über Supermicro

Supermicro Headquarters, San Jose, CA

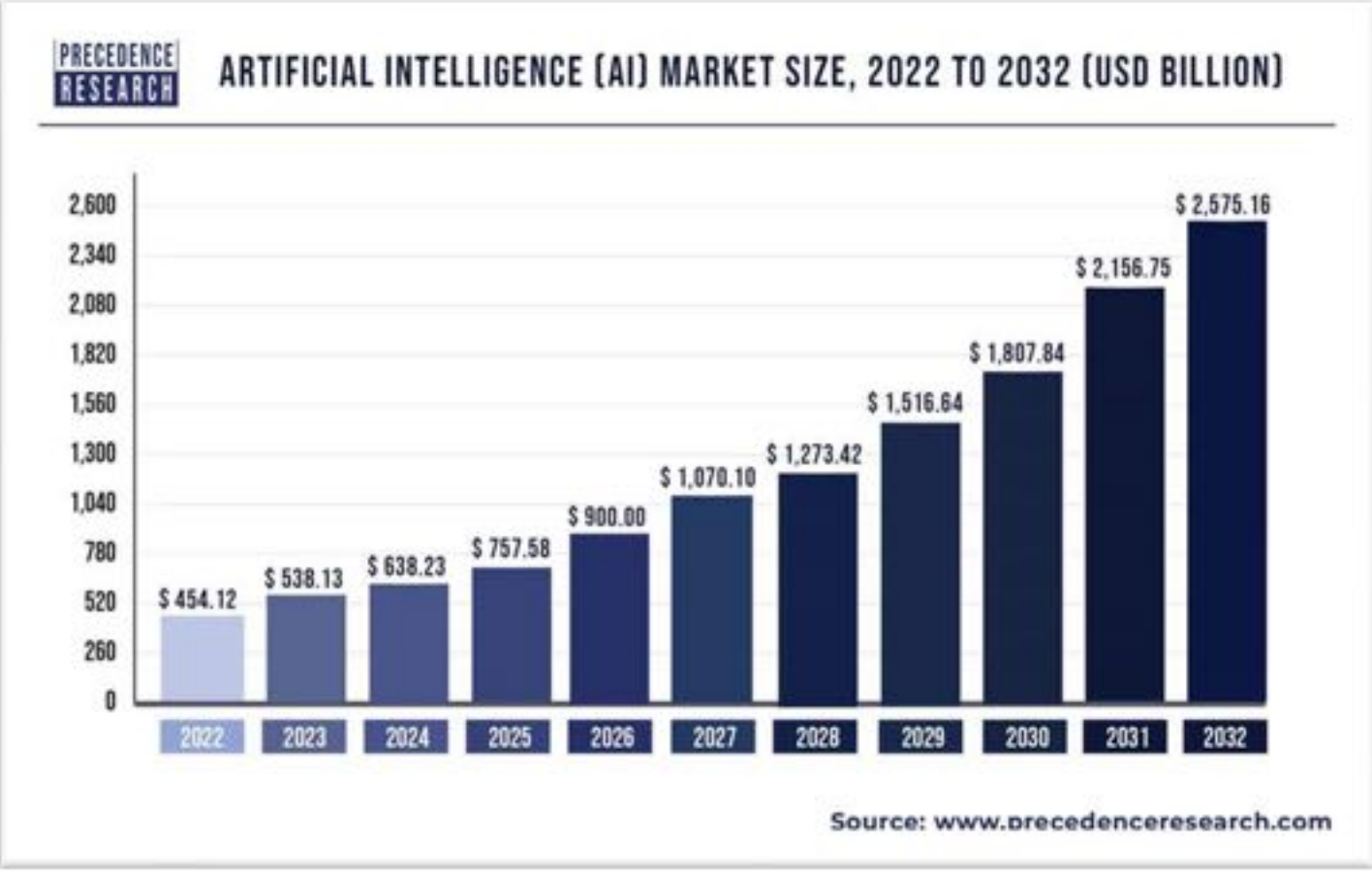

Green Computing Park, USA


EMEA Operations Park, Netherlands


Supermicro Asia Campus, Bade, Taiwan

Revenue	**$7.1B+ (FY2023)** **$5.2B (FY2022)** **$3.6B (FY2021)**
Worldwide Presence	**6M+ Sq ft. Facilities Worldwide (~ 560.000 m²)** 1. Silicon Valley (HQ) 2. Taiwan 3. The Netherlands 4. Malaysia and others
Production	**$15B/year Production Capacity (CY23)** Top 5 Largest Server System Provider Worldwide (IDC & Gartner 2022), ~1.3M units annually
Human Resource in 4 Campuses	5000+ headcount Worldwide, (>520 people in NL), ~50% Technical / R&D
Key Growth Matrix	**8 Quarters of YoY Revenue Growth** **+54% YoY Q2 FY2023 Rev.** **+79% YoY Q1 FY2023 Rev.** 100%+ YoY on *Accelerated Computing*

Der KI-Markt boomt…



The global artificial intelligence (AI) market size was valued at USD 454.12 billion in 2022 and is expected to hit around USD 2,575.16 billion by 2032, progressing with a compound annual growth rate (CAGR) of 19% from 2023 to 2032. The North America artificial intelligence market was valued at USD 167.30 billion in 2022.

https://www.precedenceresearch.com/artificial-intelligence-market

Status Quo

„China und die USA dominieren im Bereich der KI die Technologieentwicklung, während Deutschland und die EU zurückfallen“

Carolin Häussler von der Universität Passau - Expertenkommission Forschung und Innovation (EFI) Kommission

Was kann KI?

1. Medizin: Früherkennung, Entwicklung von Medikamenten
2. Archäologie: Auffinden historischer Stätten
3. Kunst: Künstlich generierte Songs
4. Kriminalität: KI im Kampf gegen Wilderei
5. Verkehr: Optimierung via KI

Was kann KI?

6. Landwirtschaft: Handlungsempfehlungen
7. Logistik: Selbstlernende Roboter
8. Wirtschaft: Entscheidungsfindung via KI
9. Bildung: individuell angepasste Inhalte

KI auf technischer Ebene - Datacenter

„Training”

Voraussetzungen:

1. Extrem performante Hardware
2. High-Speed Netzwerke
3. Optimierte Software Stacks
4. Daten, Daten, Daten
5. Zeit

Die GPU übernimmt das Kommando

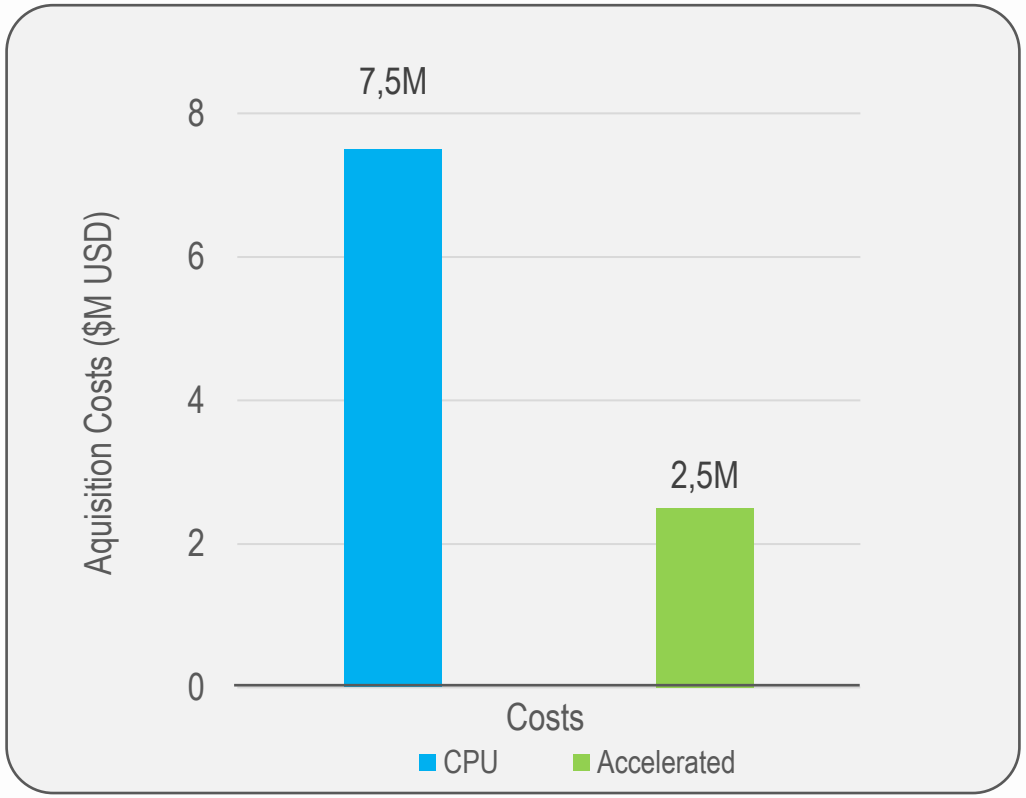
Confidential

BESCHLEUNIGTES RECHNEN IST NACHHALTIGES RECHNEN

Siemens Simcenter StarCCM+ Simulationen für Mercedes EQE verbrauchen:

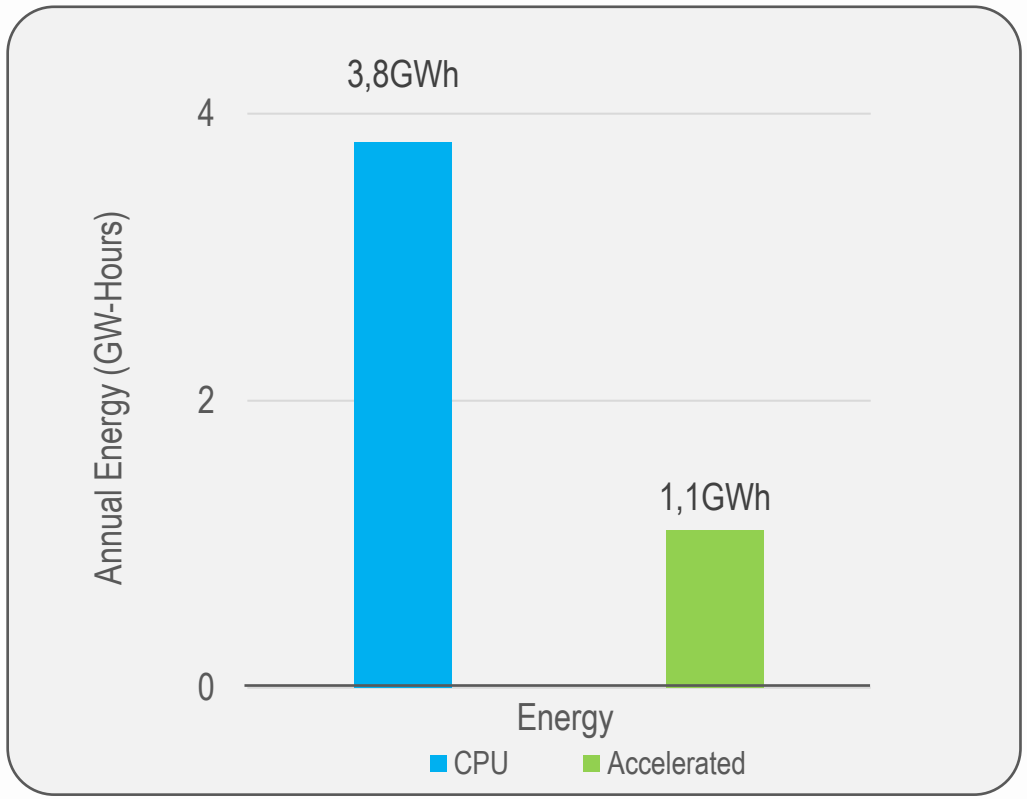
3X weniger Kosten

Gleiche Durchsatzleistung



4X weniger Energie

Gleiche Durchsatzleistung



Presented by Ian Buck, Nvidia VP Hyperscale and HPC at SC2023 in Denver, Nov. 15 - 17

Beispiel: Training System

SYS-821GE-TNHR

1. HGX H100/H200 8-GPU SXM Baseboard
2. 5th/4th Gen Intel® Xeon® Scalable processor support
3. 32 DIMM slots Up to 8TB: 32x 256 GB DRAM Memory Type: 5600MTs ECC DDR5
4. 2 PCIe Gen 5.0 X16 FHHL Slots, 2 PCIe Gen 5.0 X16 FHHL Slots (optional), 8 PCIe Gen 5.0 X16 LP
5. Flexible networking options
6. Up to 16x 2.5" Hot-swap NVMe drive bays + 3x 2.5” SATA drives
7. 10 heavy duty fans with optimal fan speed control
8. Up to 8x 3000W (4+4) Redundant Power Supplies Titanium Level

Datacenter Optimierung: Liquid Cooling

SUPERMICRO DIRECT TO CHIP LIQUID COOLING SOLUTIONS

Liquid cooled rack solution that delivers superior performance and efficiency for large scale AI and cloud scale compute infrastructure

- Full turn-key single source solution optimized from proven total solution blueprints of compute, GPU, storage, networking and power and cooling reference designs, with integrated power management tools
- Support highest densities and highest TDP CPUs and GPUs with up to 100KW power and cooling per rack
- Fully validated and tested at system (L10), rack (L11) and cluster (L12) levels
- Accelerated lead times based on in-stock inventory with deployment in weeks versus years
- Enterprise grade redundant cooling pump and power supplies, leak-proof connectors and leak detection

KI auf technischer Ebene - Edge

„Inferencing”

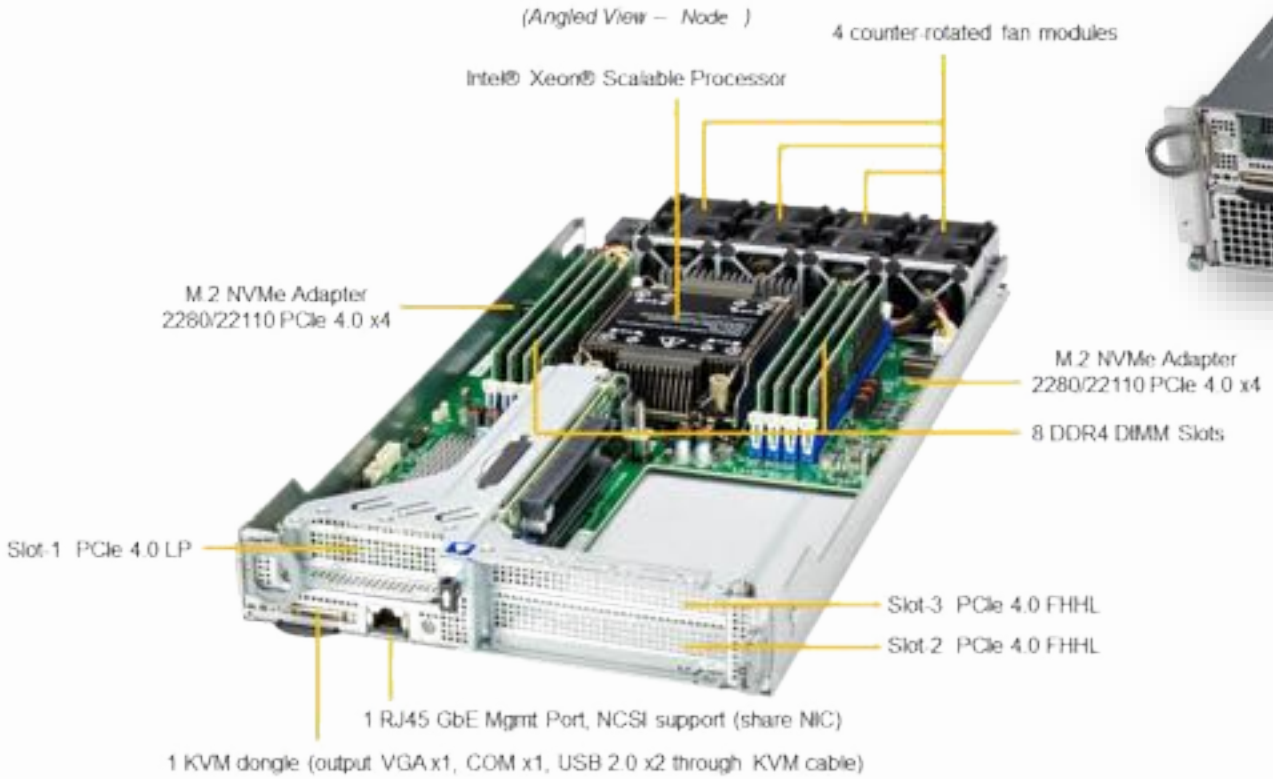
Voraussetzungen:

1. Effiziente Hardware
2. Niedrige Zugriffszeiten
3. Optimierte Software Stacks
4. Skalierbarkeit
5. Speicherkapazität

Beispiel: Inferencing System

SYS-210SE-31A

Depth *•16.9" (430 mm)*



Begleiter der KI: Storage

H13 Petascale Storage System

Extreme-Performance Storage for Data-Intensive Applications

1U A+ Server ASG -1115S-NE316R

Next-Generation Purpose-Built NVMe Storage Platform

High density for software-defined storage, in-memory computing, data-intensive HPC, private and hybrid cloud, and AI/ML applications.

- 16 hot-swap EDSFF E3.S NVMe slots for up to 480 TB of storage
- Optional 4 CXL E3.S 2T form factor memory expansion modules plus 8 E3.S NVMe storage devices
- One 4th Gen AMD EPYC™ processor—up to 128 cores
- 24 DIMMs for up to 6 TB of DDR5 memory
- 2 PCIe 5.0 Open Compute Project (OCP) 3.0 SFF-compliant AIOM slots
- 2 full-height half-length PCIe 5.0 slots with auxiliary power
- Titanium-Level efficiency power supplies


RAHMET
OBRIGADO
HVALA
GRAZIE
谢谢
KIITOS
DANKE
TAK
DANK JE
THANK YOU
SHUKRAAN
ευχαριστώ
GRACIAS
BARKA
MERCI
СПАСИБО
ARIGATO
TACK
TAKK