The RISE of up to 1MW rack densities for AI data centers.

Dr Jon Summers, Scientific Lead,
RISE Data Center Systems, SWEDEN

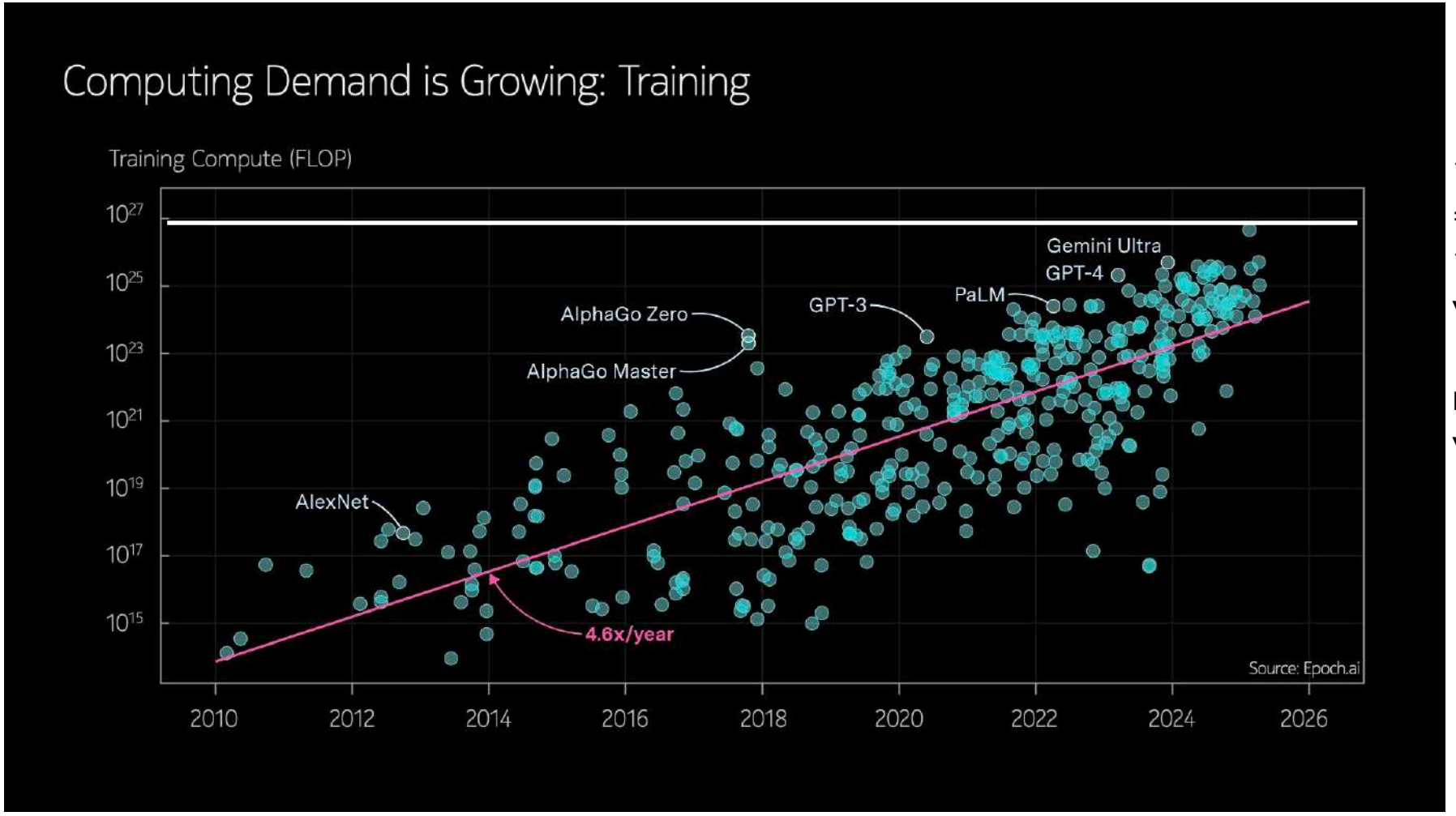
Adjunct Professor of Fluid Mechanics, LTU, Sweden
Visiting Professor of Thermofluids, Uni of Leeds, UK
Chair of Future Technologies Symposium of OCP

Data Center Forum Stockholm – 4th December 2025

Broad agenda

- Why do we need to densify?
- Growing model sizes.
- The need for more compute.
- The anticipated roadmap of the Nvidia microprocessors.
- How do we know that we can cool the next generation?

AI model compute requirements



1 ronnaFLOPs
=
1000
yottaFLOPs

ronna is 10^{27}
yotta is 10^{24}

Keynote: Anastasiia Nosova at Green Computing in the AI Era conference 2025

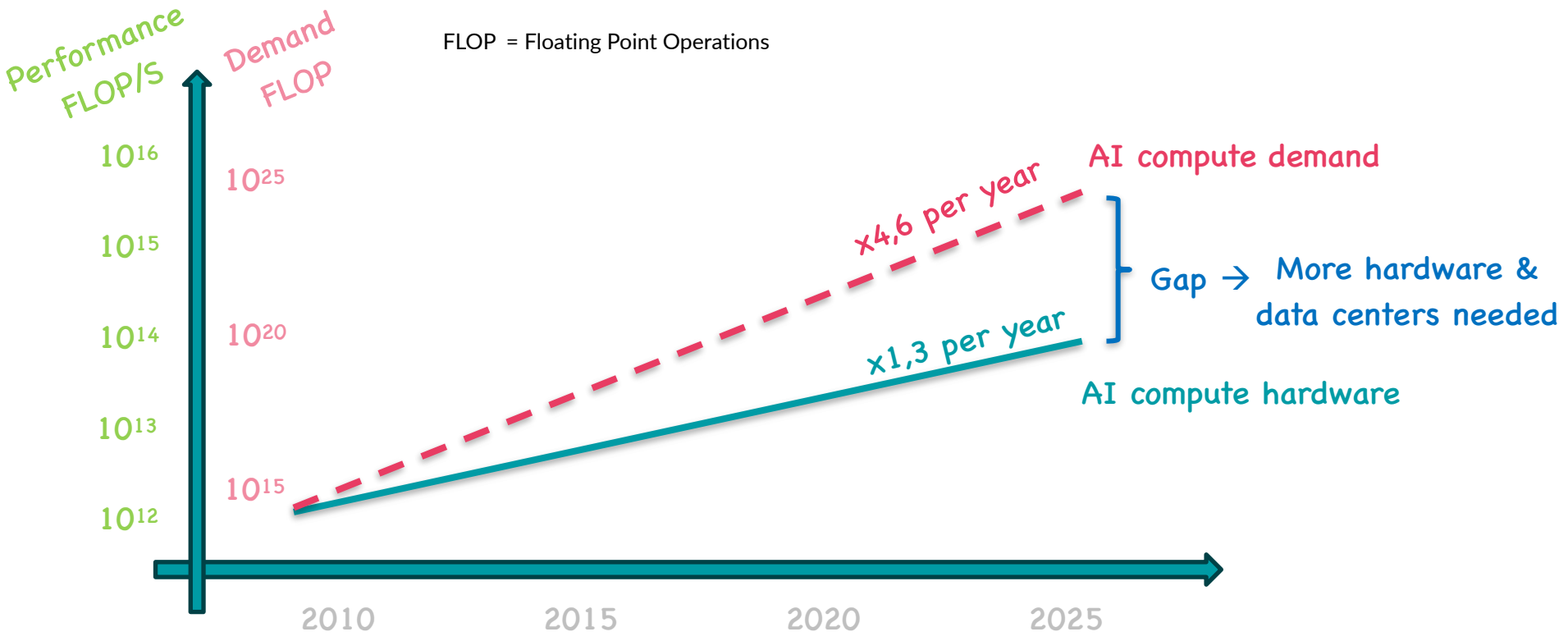
AI compute hardware capabilities

Demand is x4.6 per year

Hardware is x1.3 per year

Keynote: Anastasiia Nosova at Green Computing in the AI Era conference 2025

AI compute demand & hardware capability



Message: The widening gap between AI compute demand and the pace of hardware performance improvements is a key driver behind rising energy use and the surge in data center investments.

Sources: Keynote: Anastasiia Nosova at Green Computing in the AI Era conference 2025

Measure of compute demand

Floating point operations per joule, FLOPs/joule

$$FLOP/J = racks \times \frac{nodes}{rack} \times \frac{sockets}{node} \times \frac{cores}{socket} \times \frac{FLOP}{joule}$$

Compute energy efficiency

$$FLOP/s = racks \times \frac{nodes}{rack} \times \frac{sockets}{node} \times \frac{cores}{socket} \times \frac{cycles}{second} x \frac{FLOP}{cycle}$$

$$\frac{FLOP/s}{W} = \frac{FLOP/s\ rate\ of\ compute}{W\ power\ of\ DC}$$

Compute POWER efficiency

RI. SE

Measure of compute throughput and efficiency

Green500 Data June 2025

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	259	**JEDI** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany	19,584	4.50	67	72.733
25	1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	29,581	58.889

$$\frac{FLOPs}{joule}$$

Is not the same as

$$\frac{FLOP/s}{watt} \equiv \frac{GFlops}{watts}$$

As it involves time to solution.

Travelling vehicle analogy

RI. SE

Compute demand to train GPT-4

~1.8x10^{12} model parameters and 13x10^{12} tokens for training and 6 Floating operations (FLOP) per parameter for one token

⇒ ~140 yottaFLOP (unoptimized) [literature says ~100 yottaFLOP]

LUMI in Kajaani is ~380 petaFLOP/s at 7.1 MW (**FP64**)

⇒140 yottaFLOP/380petaFLOP/s = 368Mseconds or 11.6 years to train

⇒7.1MWx11.6years = ~2600 GWh or 260 MEuro at 0.1 Euro/kWh

An NVIDIA A100 can do 624 teraFLOP (**FP16**)

OpenAI trained on 25 000 NVIDIA A100 GPUs

⇒ Training time >100 days

⇒Cost $100 million

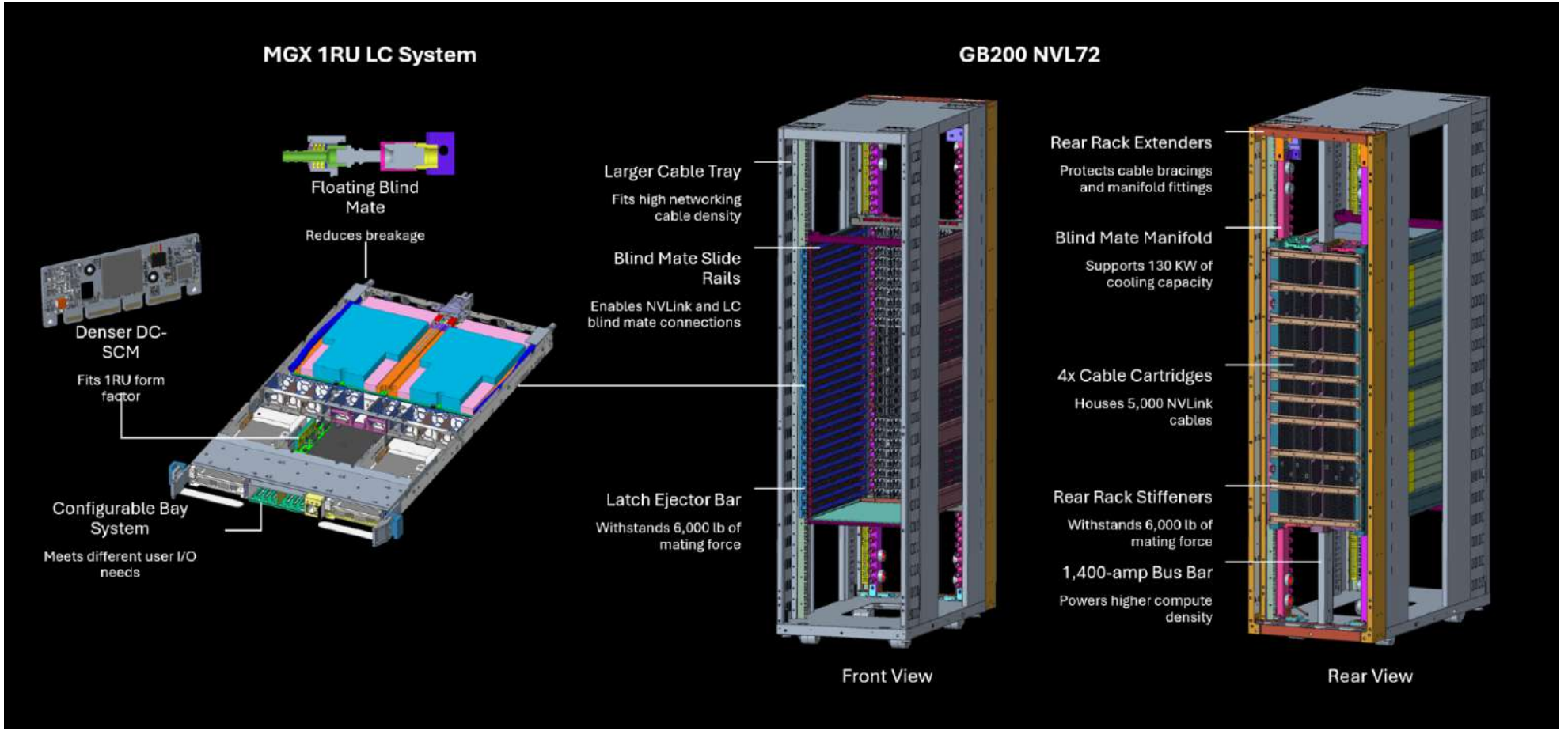
⇒ Energy ~54 GWh

yotta =10e24 = 1 000 000 000 000 000 000 000 000

RI. SE

Source: OpenAI, RISE, Wikipedia, the decoder, ChatGPT Statistics

Deep dive into the NVL72 – current rack offering.



Source: https://developer.nvidia.com/blog/nvidia-contributes-nvidia-gb200-nvl72-designs-to-open-compute-project/

Performance and DC needs for the NVL72

Into a rack

NVL72 =
18 x nodes
= 102.6 kW
+switches
and extras
=**~120kW**

Liquid
Cooled

720 petaFLOP/s
for training

The new GB200 – with 4 B200 GPUs and 2 Grace CPU
Peak power of 5.7kW
40 petaFLOP/s training OR 80 petaFLOP/s inference

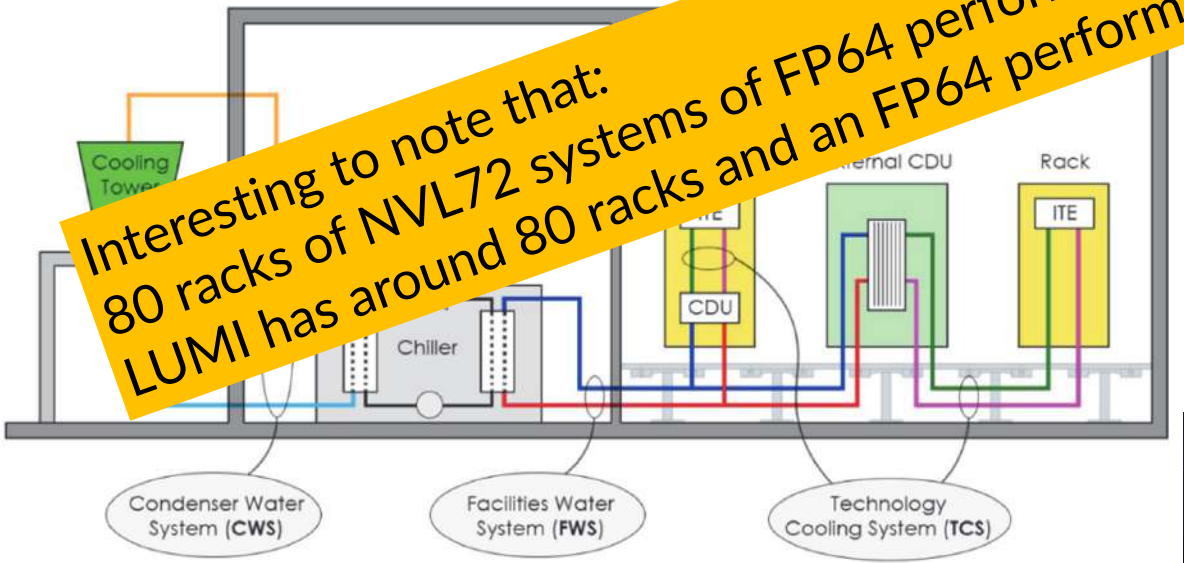
RI.
SE

Performance and DC needs for the NVL72

GPT-4 training with 80 NVL72 racks can be achieved in 28 days

8 x 32 [illegible] power shelves

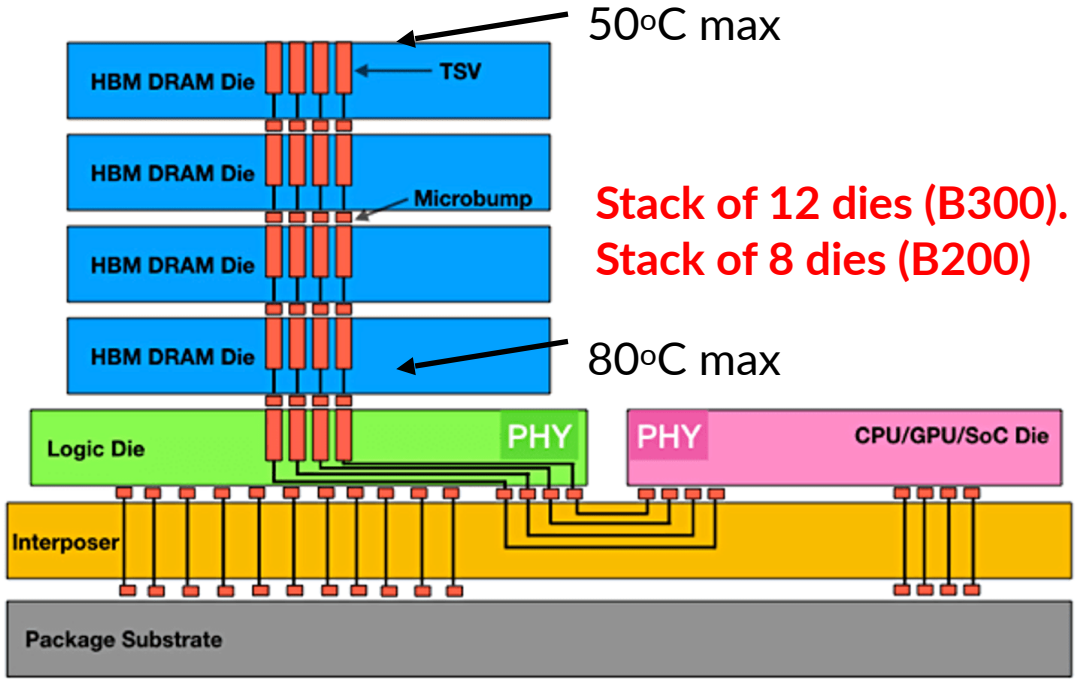
Interesting to note that:
80 racks of NVL72 systems of FP64 performance is **8 PFLOP/s** on 10.6MW
LUMI has around 80 racks and an FP64 performance of **380 PFLOP/s** on 7.1MW



Based on: Refrigerating and Air-Conditioning Engineers American Society of Heating. Liquid Cooling Guidelines for Datacom Equipment Centers. ASHRAE Publications/American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2006.

RI. SE

Current generation AI offering by NVIDIA: the B200 and B300



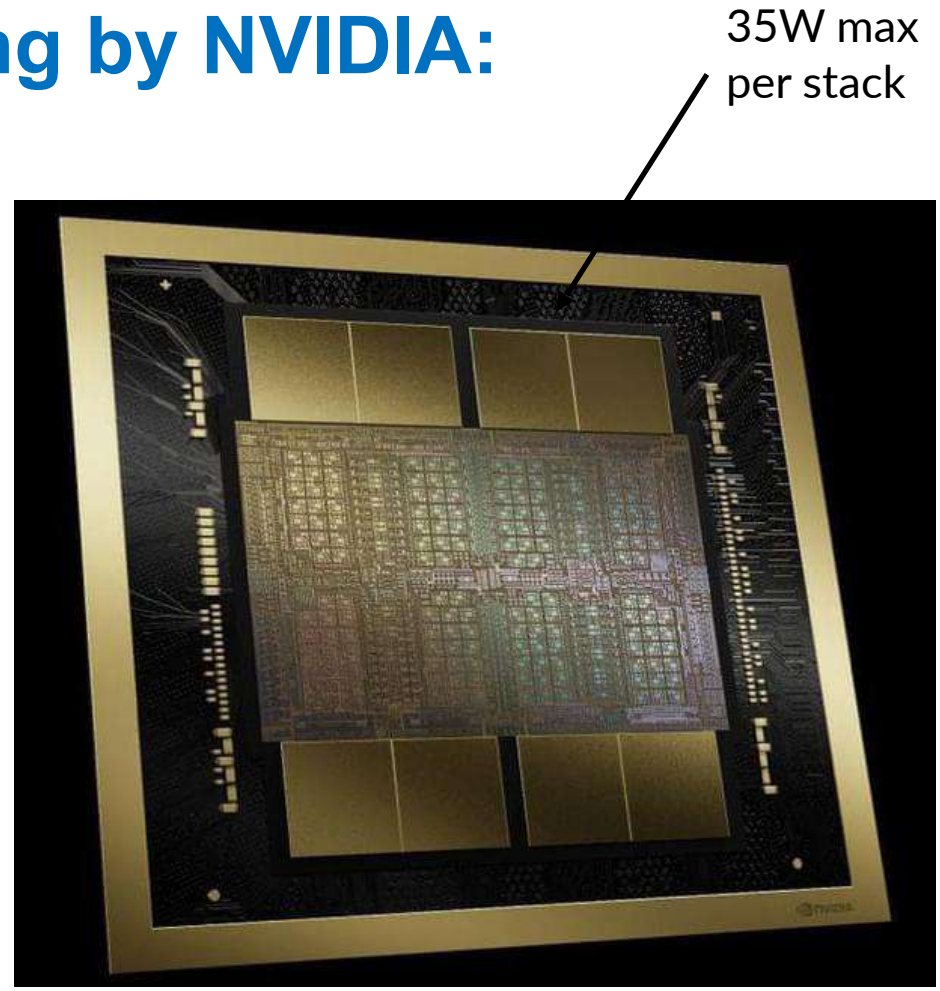


Image source: https://semiengineering.com/knowledge_centers/memory/volatile-memory/dynamic-random-access-memory/high-bandwidth-memory/

Image source: https://lifeboat.com/blog/2024/03/nvidia-reveals-blackwell-b200-gpu-the-worlds-most-powerful-chip-for-ai

RI. SE

The upcoming NVL144 – approx. 300kW per rack

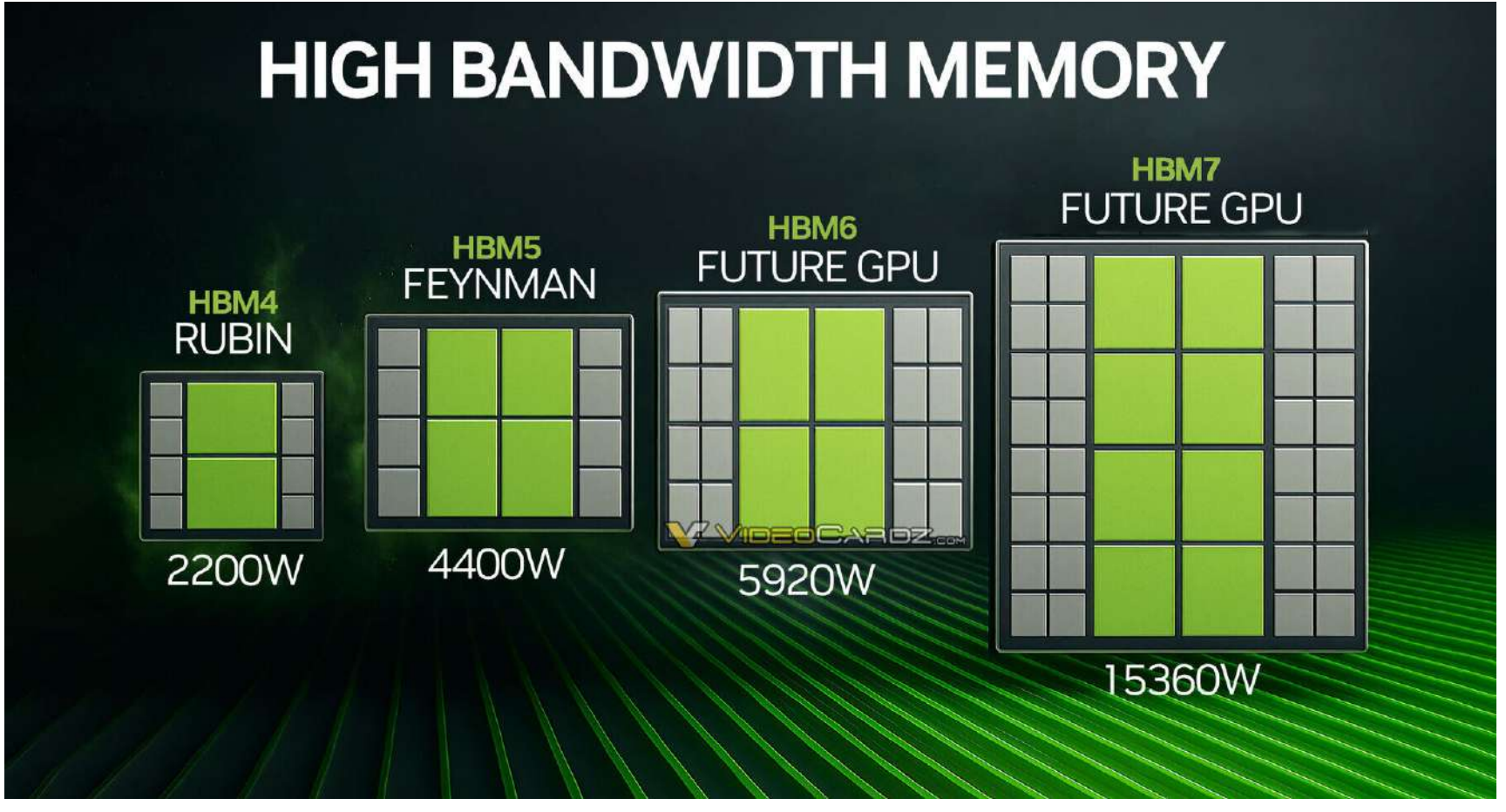
Source: https://hothardware.com/news/nvidia-announce-vera-rubin-nvl576

RI.SE

The NVL576 – expected to be 600kW per rack

Source: https://hothardware.com/news/nvidia-announce-vera-rubin-nvl576

NVIDIA GPU (AI Accelerators) specifications



Source: https://tera.kaist.ac.kr/

RI. SE

NVIDIA GPU (AI Accelerators) specifications

Next-Generation HBM Roadmap by KAIST TERALAB

HBM Spec. | Packaging/Cooling | Architecture | AI Design Agent

Ver 1.2 / updated.250521

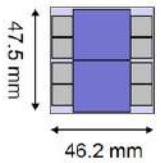
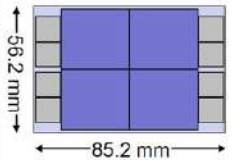
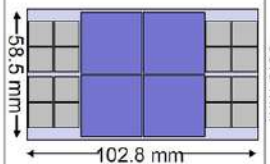
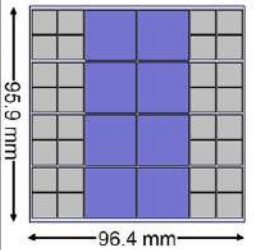
	HBM4 (2026)	HBM5 (2029)	HBM6 (2032)	HBM7 (2035)	HBM8 (2038)
Data Rate	8 Gbps	8 Gbps	16 Gbps	24 Gbps	32 Gbps
# of I/O	2,048	4,096	4,096	8,192	16,384
Bandwidth	2.0 TB/s	4 TB/s	8 TB/s	24 TB/s	64 TB/s
Capacity/die	24 Gb	40 Gb	48 Gb	64 Gb	80 Gb
# of die stack	12/16-Hi	16-Hi	16/20-Hi	20/24-Hi	20/24-Hi
Capacity /HBM	36/48 GB	80 GB	96/120 GB	160/192 GB	200/240 GB
Power/HBM	75 W	100 W	120 W	160 W	180 W
Die stacking	Microbump (MR-MUF)		Bump-less Cu-Cu Direct bonding		
Cooling Method	Direct-to-Chip (D2C) Liquid Cooling	Immersion Cooling		Embedded Cooling	
HBM Architecture	Custom HBM Base Die HBM-LPDDR	3D NMC-HBM & stacked cache / decap	Multi-tower HBM Active / Hybrid Interposer	Hybrid HBM Architecture HBM-HBF HBM-3D LPDDR	Full-3D / HBM Centric Computing Architecture
Additional Features (Patent)	NMC processor + LPDDR Ctrl	+ Cache + CXL + on-die/stacked decap + HBM shielding	+ Network switch + Bridge die + Asymmetric TSV	+ HBF/LPDDR Ctrl + Storage network	+ HBM Centric Interposer + Double side Cooling + Edge-expand Stack
AI Design Agent	ubump & TSV-array Decap placement Optimization	I/O Interface Optimization considering PSIJ	Hybrid Equalizer + Generative AI based SI/PI Metric Estimation	LLM based Human Interactive AI Design Agent	

Source: https://tera.kaist.ac.kr/

RI. SE

NVIDIA GPU (AI Accelerators) specifications

Next-Generation GPU-HBM Roadmap
: More GPU & HBM Integrated Above Interposer

GPU Architecture	Rubin (2026)	Feynman (2029)	Post Feynman (2032)	Next-Gen Architecture (2035)
GPU Die Size	728 mm²	750 mm²	700 mm²	600 mm²
GPU Power	800 W	900 W	1,000 W	1,200 W
GPU-HBM Module	R200	F400	Post Feynman GPU-HBM Module	Next-Gen GPU-HBM Module
Interposer Size GPU HBM Interposer				
# of GPU Dies	×2	×4	×4	×8
# of HBM Stack	HBM4×8	HBM5×8	HBM6×16	HBM7×32
Interposer Die Size	2,194 mm² (46.2 mm x 48.5 mm)	4,788 mm² (85.2 mm x 56.2 mm)	6,014 mm² (102.8 mm x 58.5 mm)	9,245 mm² (96.4 mm x 95.9 mm)
Total Bandwidth	16 / 32 TB/s	48 TB/s	128/256 TB/s	1,024 TB/s
Total HBM Capacity	288/384 GB	400/500 GB	1,536/1,920 GB	5,120/6,144 GB
Total Power	2,200 W	4,400 W	5,920 W	15,360 W

TERA KAIST TeraByte Interconnection and Package Laboratory

Source: https://tera.kaist.ac.kr/

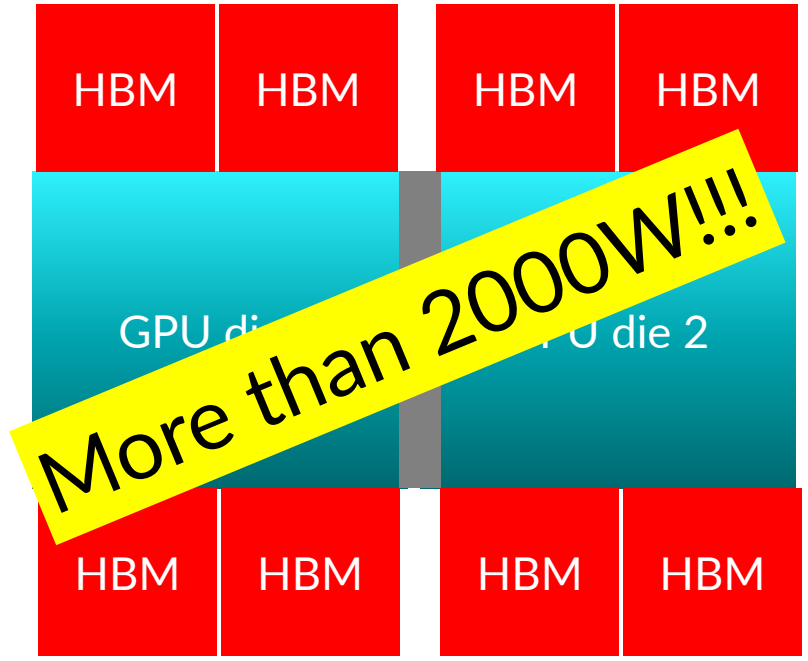
Next generation AI offering by NVIDIA
B300 (Q4, 2025) and R200 (Q3, 2026)

The energy dissipation for the HBMs is not included in the TDP.

HBM3e for the B300 and HBM4 for the R200 will be 12-Hi memory stack (12 vias).
8 x 12 x 3GB = 288GB or HBM RAM.

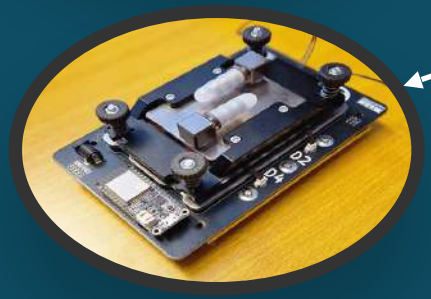
288GB is ~ 1.24 trillion transistors.
8 stacks at ~28W each = 224W

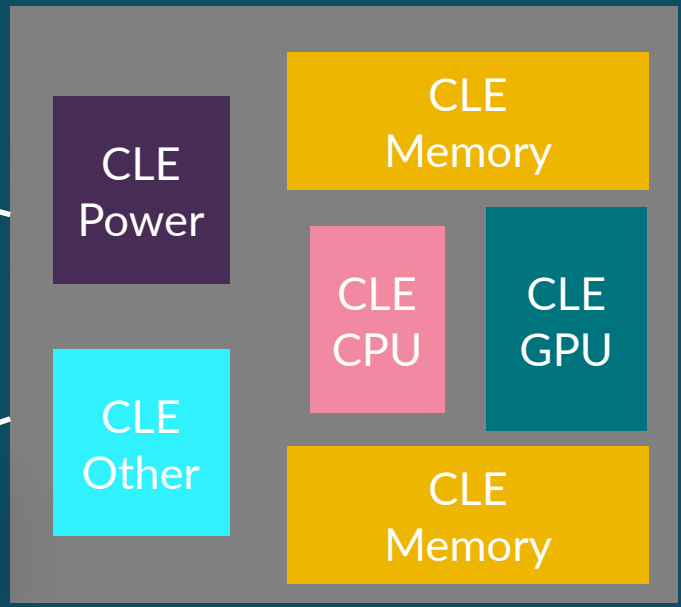
This is on top of the ~1800W for the dies.
GPU dies are ~ 800mm^2



ΔT over 12-Hi HBM is 30^{o}C
and bottom die ≤ 80^{o}C

What if we can use a Server Emulator?





- Looks like a server
- Accurately emulating a real chip
- High Power Density
- Compatible with different cooling technologies
- Modular server design
- Can meet future power demands

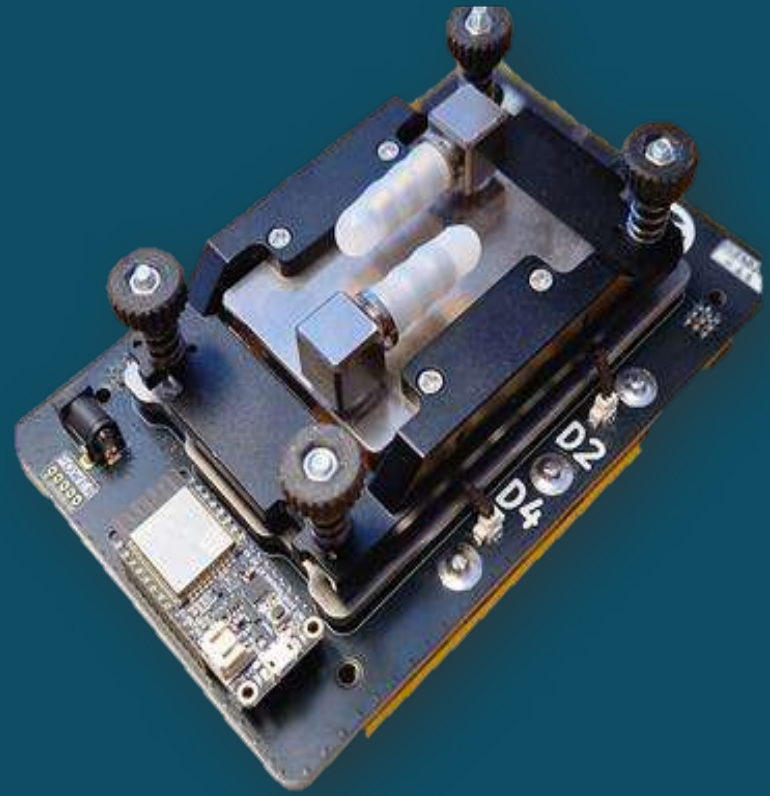
A server emulator consists of one or more CLE, Chip Load Emulators

RI. SE

Chip Load Emulator

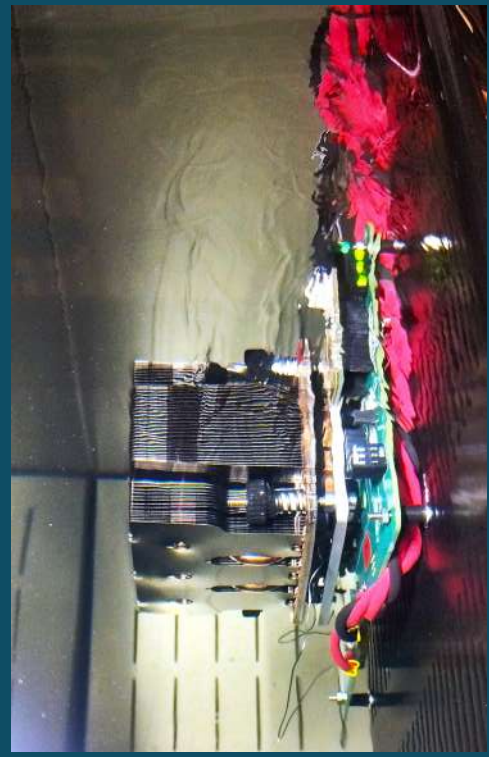
Example of use cases

- Heatsink testing
- Cold plate testing
- Thermal interface material evaluation



Compatibility with different cooling systems

Air cooling



Immersion

Direct to chip

RI.
SE

Rack Emulator

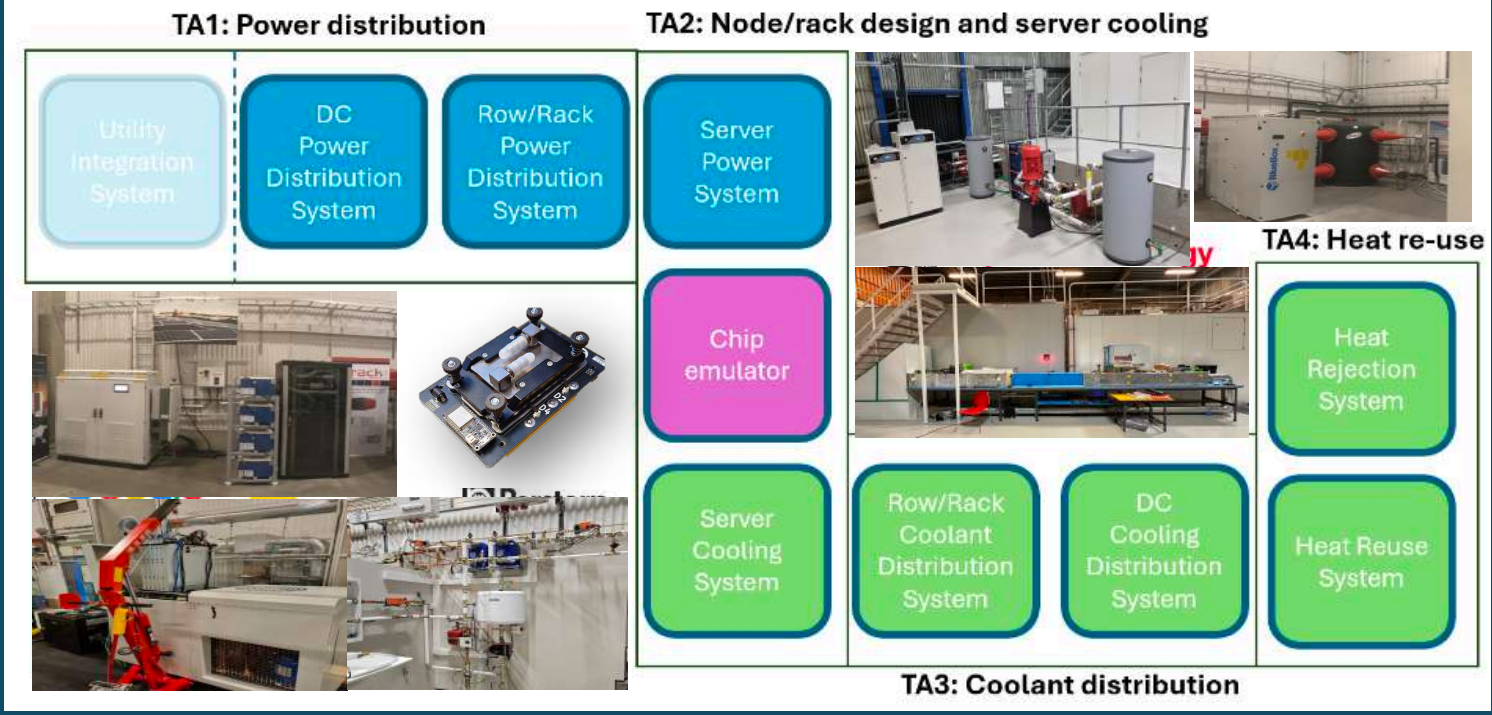
Example of use cases
19”, 48U, 0.72MW

- Coolant distribution
- Power distribution
- System stress test

A rack emulator consists of one or more server emulators

RI.
SE

The H-DINI project – going for 1MW infrastructure.



Goals

- *The project goal is to develop technologies and systems for power-dense AI compute with efficient power delivery and thermal performance.*
- *Demonstrate system solutions for selected technology areas*
- *Total budget €6M*

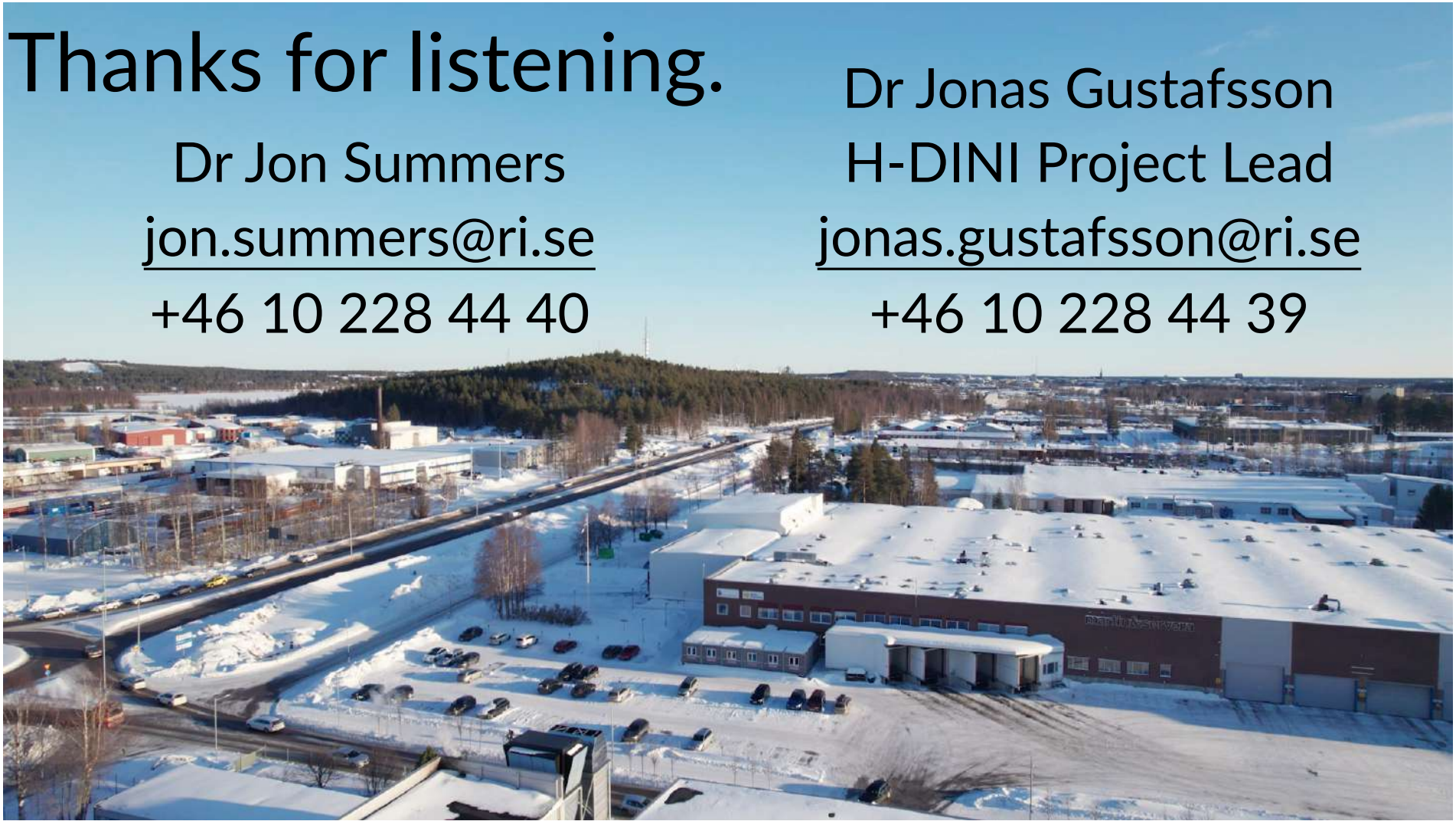
Funded by Vinnova and project partners

RI. SE

Thanks for listening.

Dr Jon Summers
jon.summers@ri.se
+46 10 228 44 40

Dr Jonas Gustafsson
H-DINI Project Lead
jonas.gustafsson@ri.se
+46 10 228 44 39



Thanks also to our colleagues at the RISE Data Center Systems.

RI.
SE