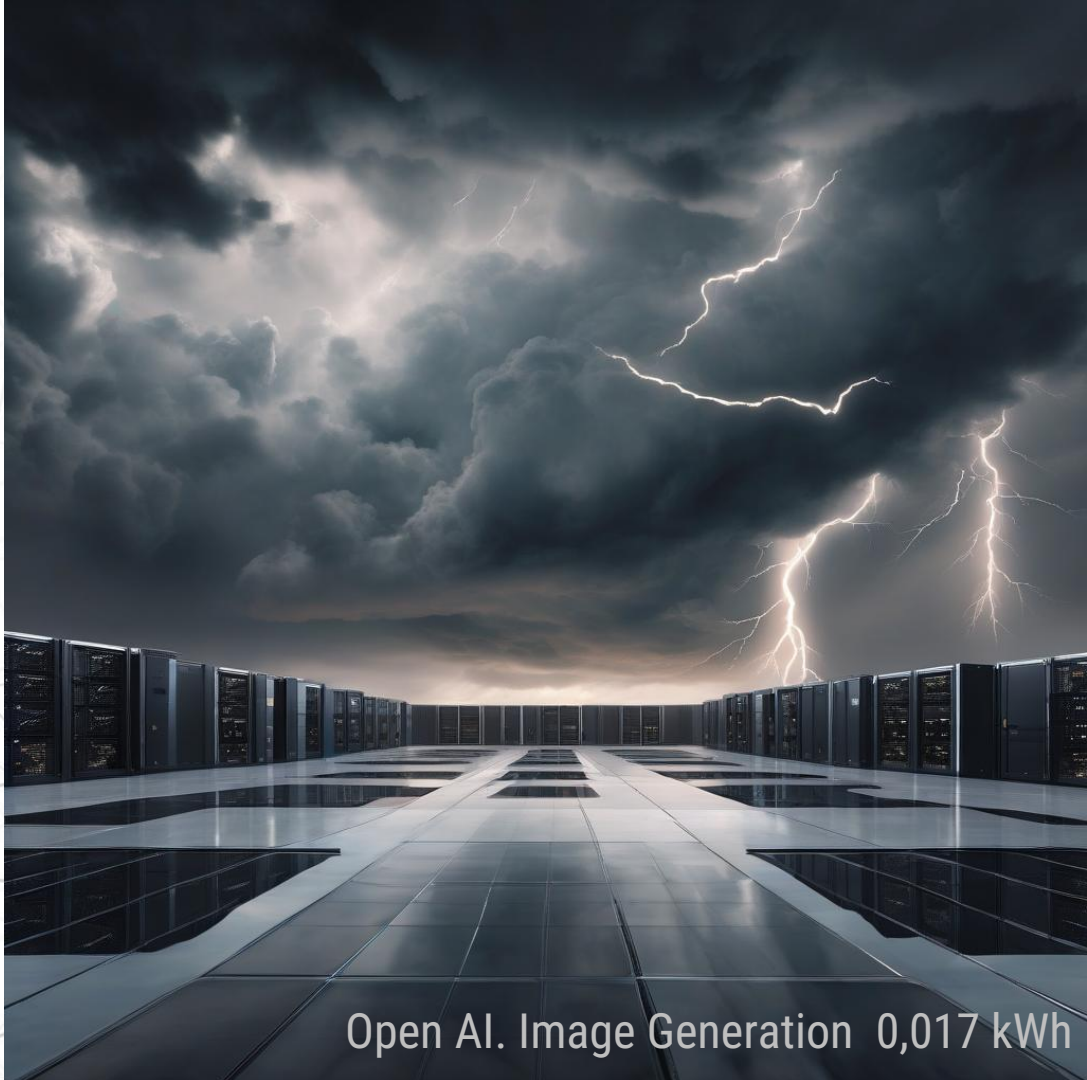
Vihar előtti csend

Katona Gerely
2024. 11.06.



Open AI. Image Generation 0,017 kWh


„Az adatközpont üzemeltetés egyáltalán nem stresszes”
John (26)
OpenArt.ai 0,026 kWh

ChatGPT- 4o

The spread of artificial intelligence (AI) will have several significant effects on data centers:

OpenAI

Increased Operational Efficiency:

Predictive Maintenance:

Energy Optimization

Enhanced Security...

Effective Resource Management

Scalability

Cost Reduction

Data Analytics and Insights

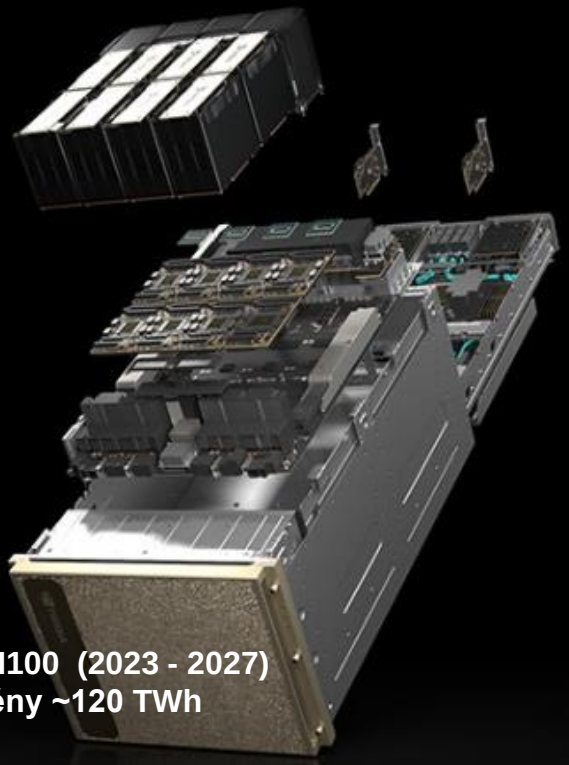
ChatGPT

Chat GPT 4o (text) 0,023 kWh

Explore DGX H100

Estimated System Power Usage **9.2kW Nominal, 10.2 kW Max**

- **8x NVIDIA H100 GPUs With 640 Gigabytes of Total GPU Memory**
  18x NVIDIA® NVLink® connections per GPU, 900 gigabytes per second of bidirectional GPU-to-GPU bandwidth
- **4x NVIDIA NVSwitches™**
  7.2 terabytes per second of bidirectional GPU-to-GPU bandwidth, 1.5X more than previous generation
- **10x NVIDIA ConnectX®-7 400 Gigabits-Per-Second Network Interface**
  1 terabyte per second of peak bidirectional network bandwidth
- **Dual Intel Xeon Platinum 8480C processors, 112 cores total, and 2 TB System Memory**
  Powerful CPUs for the most intensive AI jobs
- **30 Terabytes NVMe SSD**
  High speed storage for maximum performance



Évi: 1,5 milló DGX H100 (2023 - 2027)
Évi teljesítmény igény ~120 TWh

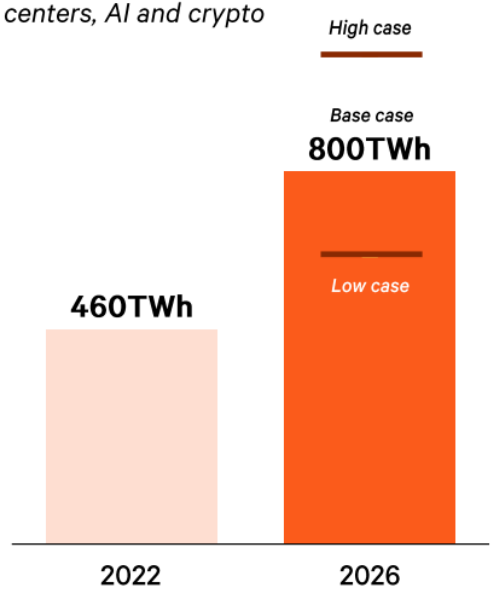
Magyarország éves **villamos energia** fogyasztása 45 TWh

Adatközpontok hazai éves **villamos energia** fogyasztása 1-2 TWh (MEKH becslés 2021)

Data center power demand is expected to grow at a rapid pace with AI.

International Energy Agency

Global electricity demand from data centers, AI and crypto



Authorities imposing limitation on data center expansion.

Connection restrictions in place since 2021.

Top colocators had permits rejected by Irish authorities 2023.

New data centers banned for three years.

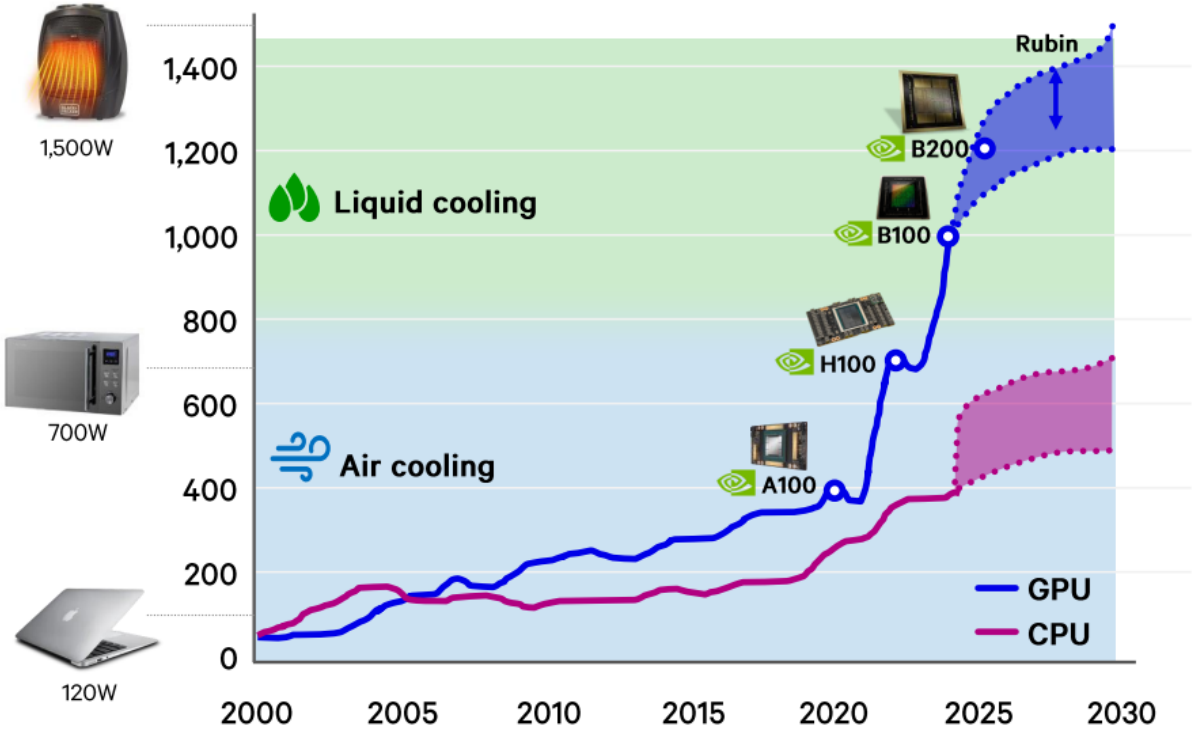
Aimed to minimize impact on Singaporean grid and carbon footprint commitments, measure was lifted with introduction of stricter rules.

Restrictions on new data centers in Virginia.

New rules impacting construction along route 7 in Loudoun county, where power infrastructure is limited.

CPU and GPU power consumption forecast

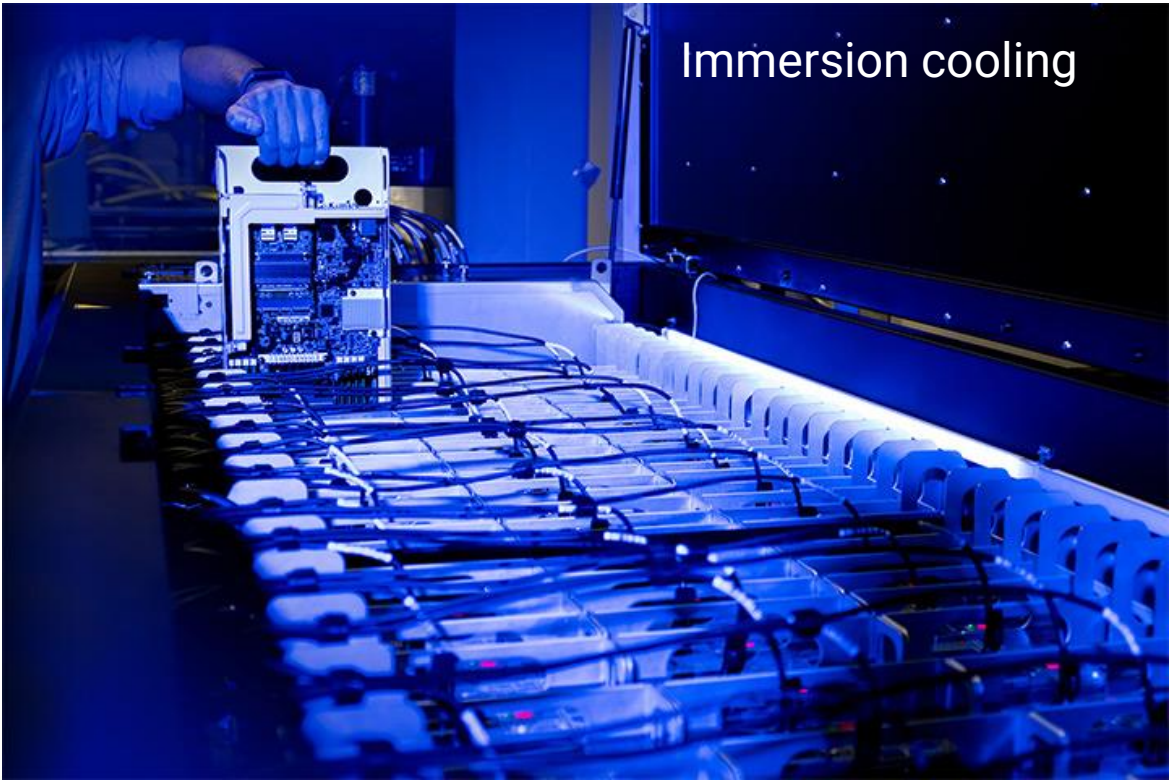
Thermal Density Power - TDP (watts)



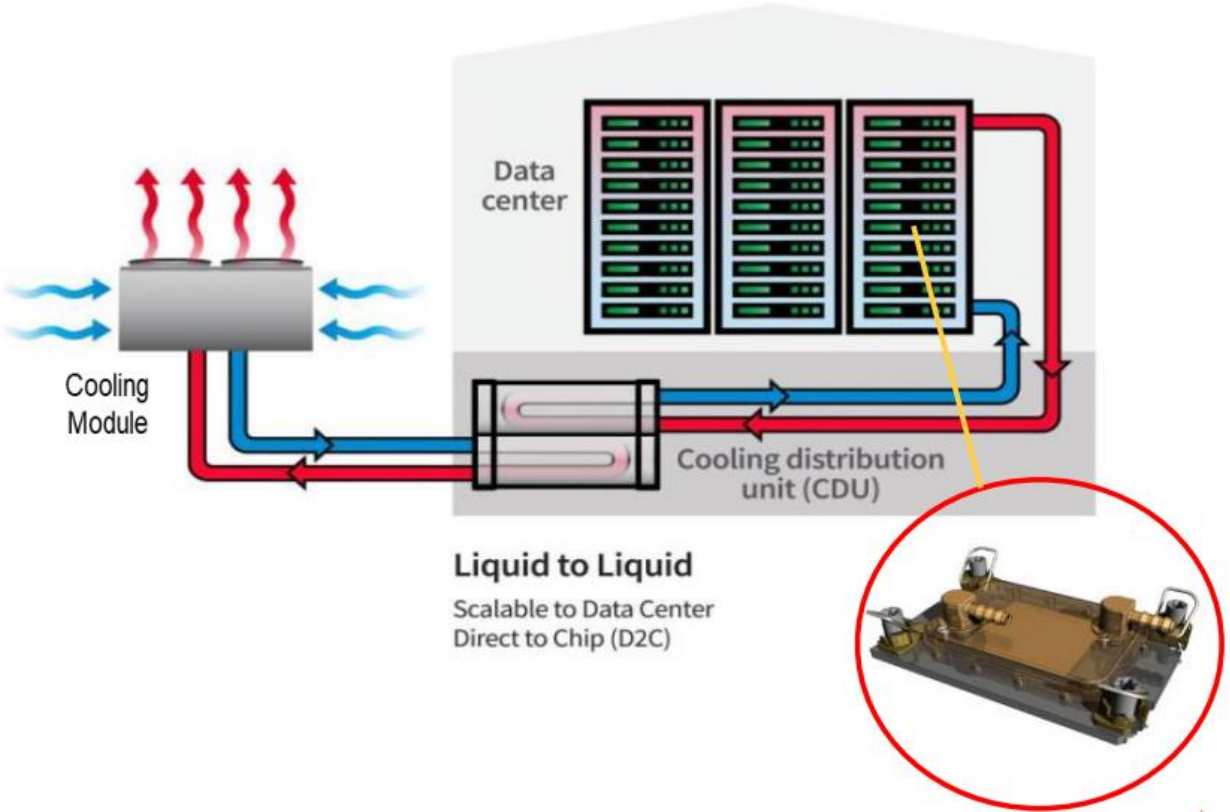
Rackenkénti rendszer csúcs fogyasztás

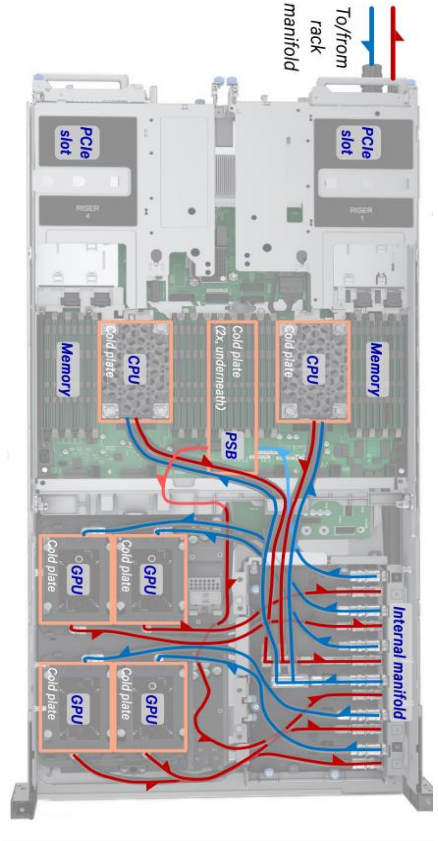
NVL288 Rubin Next	1118 kW / rack
NVL 288 Rubin	894 kW / rack
NVL 288 Blackwell	606 kW / rack
NVL 144 Rubin Next	559 kW / rack
NVL 144 Rubin	447 kW / rack
NVL72 Blackwell ultra	150 kW / rack
NVL72 Blackwell	130 kW / rack
Hopper H100	40 kW / rack

Hogyan kezelhető ez a teljesítmény sűrűség?

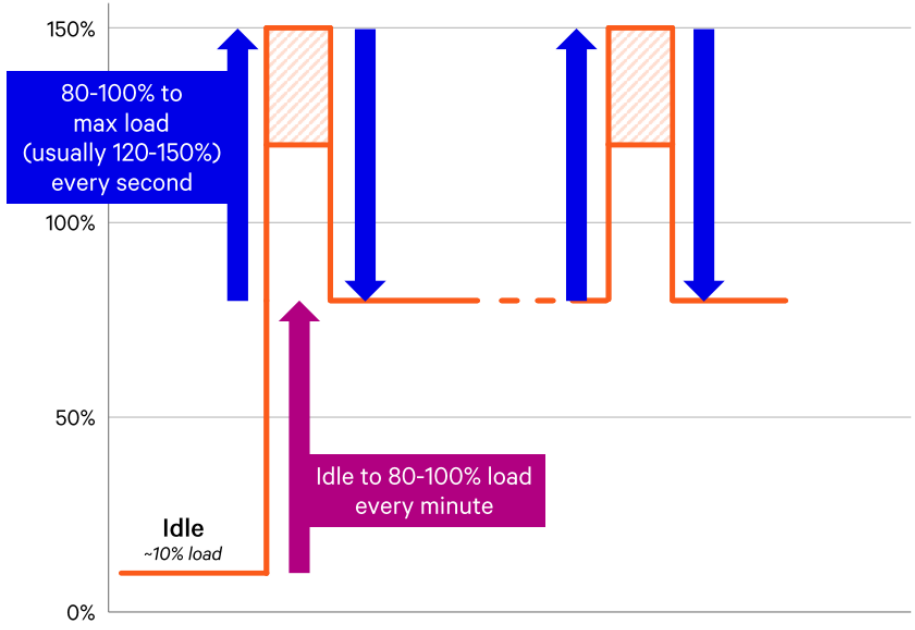


Direct Liquid Cooling

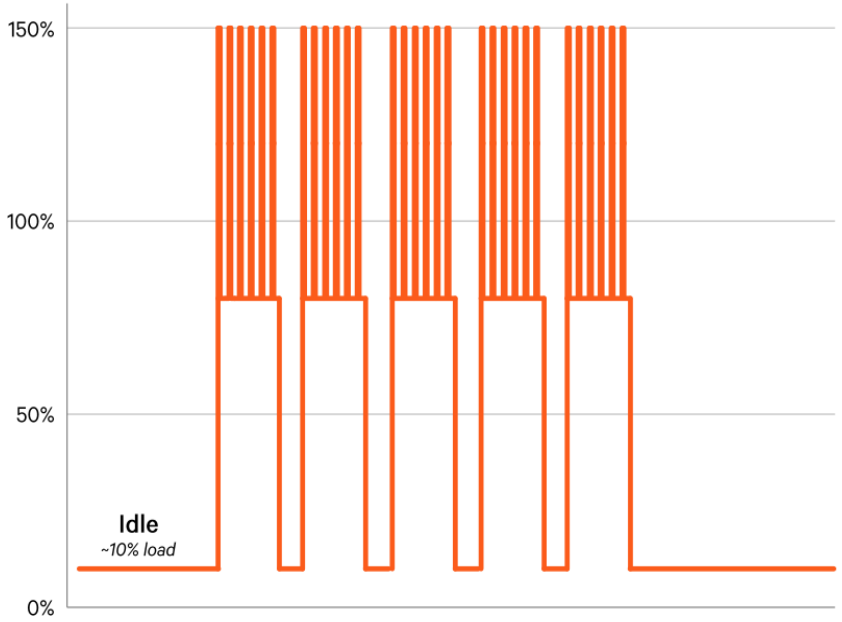




AI load profile



AI load profile



Hogy éljük együtt az AI-val?

Betanítás (GPT-4 modell: 1,75 MWh)
Elemzések riportok generálása
Nem real time video feldolgozás
Software defined power

} **Időbeli optimalizáció** (olcsó / zöld energiára)

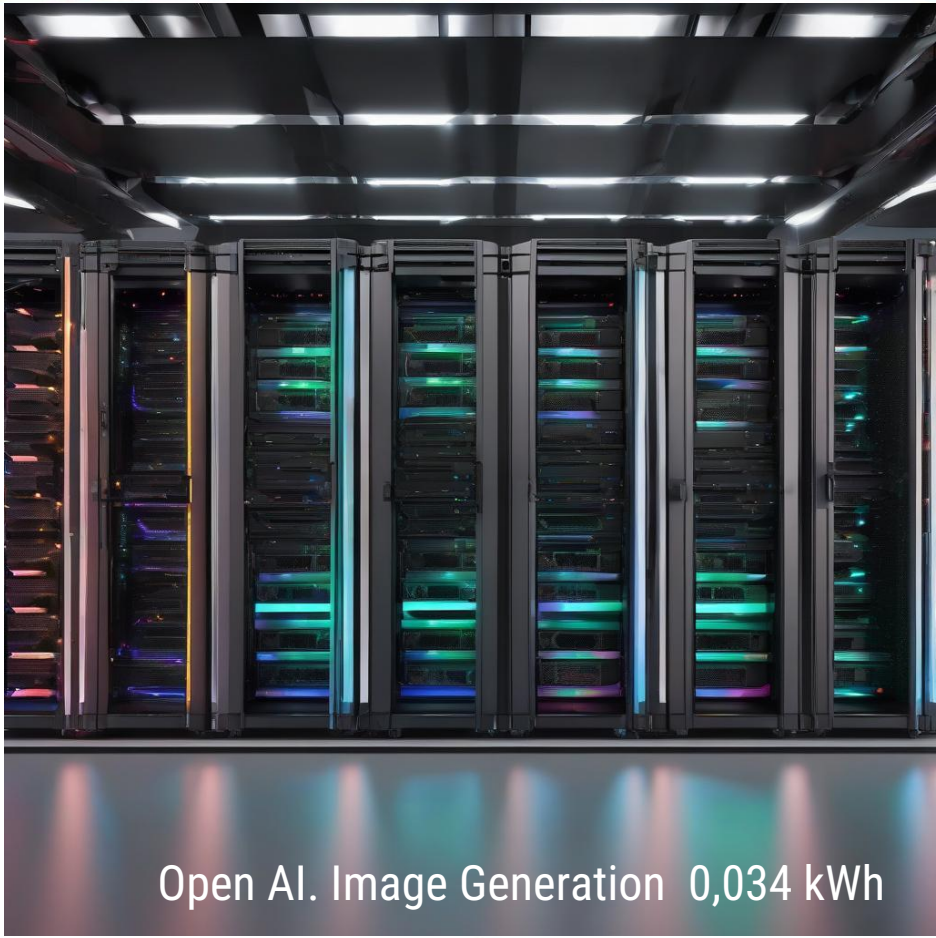
Szép új jövő?

Energiaellátó és hűtési rendszerek: **10x kapacitás** növekedés, vagy csökkenés másodpercen belül!

Nem lehet **reaktív** módszerekkel lekövetni!

Prediktív, a használatot előre jelző, akár szabályozó rendszerek kellenek

Elmosódik a határ az IT és a „Facility” között



Források:

https://arxiv.org/pdf/2311.16863
https://mekh.hu/eves-adatok
https://www.baeldung.com/cs/chatgpt-large-language-models-power-consumption
https://nvdam.widen.net/s/5kgbjq2v2t/hpc-hgx-h100-datasheet-nvidia-web
https://images.nvidia.com/aem-dam/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf
https://journal.uptimeinstitute.com/ai-will-have-a-limited-role-in-data-centers-for-now/
https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3

Köszönöm a figyelmet!

A prezentáció készítése során az alábbi AI toolokat használtam:

ChatGPT 3.5
ChatGPT 4o
ChatGPT Image generation
OpenART.ai
Midjourney

A prezentáció elkészítésének AI oldali becsült energia igénye: **234 Wh**