Dr. Stefan Fütterling – The Open Group Distinguished Architect – stefan.fuetterling@capgemini.com
Martin Bachmaier – IT Versatilist – Lenovo R&D – mbachmaier@lenovo.com

Duale Hochschule Baden-Württemberg, Stuttgart
IT Architekturen 2025-11

Teil 4.2
Scale-out Data Center Design

DHBW Duale Hochschule Baden-Württemberg Stuttgart

Gesamtsicht der Vorlesung

- **Einführung**
  - 1.1 Einführung in IT Architektur
  - 1.2 Dynamische IT Infrastrukturen
  - 1.3 Cloud Computing
- **Server Virtualisierung**
  - 2.1 Einführung in die Server Konsolidierung und Virtualisierung
  - 2.2 Virtuelle Maschinen (VMs) am Beispiel VMware vSphere (ESXi)
  - 2.3 OS Containers am Beispiel Linux LXC und Docker
  - 2.4 Deep Dive x86 Virtualisierung und LXC
- **Zentralisierter Storage**
  - 3.1 Storage Area Networks (SAN) und Network Attached Storage (NAS)
  - 3.2 RAID Levels
  - 3.3 Disksysteme und Hyperconverged Infrastructure (HCI)
- **Clusterarchitekturen**
  - 4.1 Einführung in Clusterarchitekturen (LB-Cluster, HPC Cluster, HA Cluster)
  - 4.2 Scale Out Data Center
  - 4.3 Clustersoftware am Beispiel parallele DB Systeme und Big Data Analysis Cluster
- **IT Betrieb**
  - 5.1 Überblick DevOps, Application Management und Systems Management
  - 5.2 IT Service Management (ITIL)

Topics

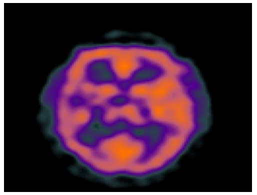
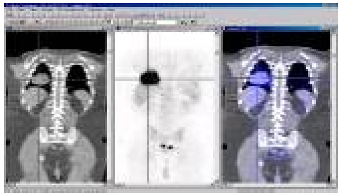
- Scale-out examples everywhere
- Challenges – Power consumption
- Challenges – Cooling
  - How to cool server and data centers
  - the path to liquid based cooling

- Scale-out for scientific workload: High Performance Computing (HPC)
  - Basics
    - Benchmarking a scale-out supercomputer
    - Architecting and operating a scale-out supercomputer
  - Examples
    - Mare Nostrum, Barcelona
    - Hawk, Stuttgart
    - COVID-19 special
- Scale-out for commercial workload
  - Public clouds
  - Social media

Scale-out examples (1/2)

- Medical area:
  - Folding @Home (Protein folding)
  - Higher quality / Faster Analysis in medical imaging: CT scan, MRI scan, PET scan ...
    - 3D image reconstruction, volume rendering, real-time rendering

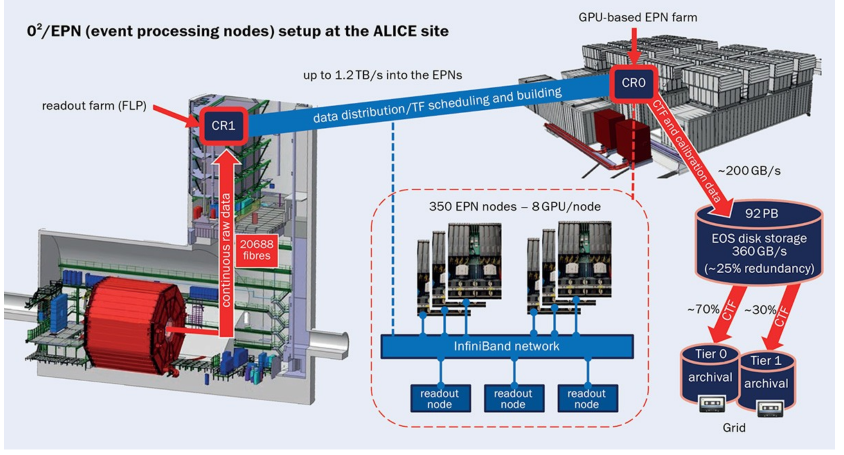


- Physics: CERN, Large Hadron Collider (LHC), analyze sub-atomic particles
    - Produces: >> 50.000 TB per year from just the collider
    - 2022: 4 TB/s real-time data processing via GPU cluster

- Weather: short-term and long-term climate
  - Daily forecast (95% for a 2-day forecast)
  - Hurricane forecasts (100m resolution)
  - Volcanic ash
  - 2020: start to model 3rd dimension of weather

- Commercial areas
  - Car crash, or engine combustion simulations
  - Material simulation
  - Banks & Insurance: financial risk assessments
  - High-Frequency Trading



Scale-out examples (2/2)

- Digital Video & Surveillance
  - 4K video processing in real time
  - Multiple cameras, every store, every corner, Edge computing
- 3D Cinema
  - Avatar (2009)
    - 4.000+ blades (HP BL2×220c)
    - 35.000+ cores, 104 TB memory
    - 10 Gigabit Ethernet
    - 3 PB storage subsystem
    - Constant processing of 7-8 GB/s data
    - Finale movie: 288 MB/s, or 17 GB/min
    - Ubuntu
    - Special water-cooled doors on the 34 racks

More than 4,000 HP blades power the processing at Weta Digital. Photo courtesy of Weta Digital.

Avatar 2 (2022): 18.5 PB storage
One NZ data center isn't enough: AWS Australia. Panel discussion with Weta on compute: https://www.youtube.com/watch?v=Oon01xMYhJ4

Challenges: Power, Cooling (and Space)

- Power
  - More servers equals higher demand
  - Existing servers draw more power (current Intel and AMD 500W, GPUs 1000W+)
  - >1kW per rack unit, 30+ kW per rack

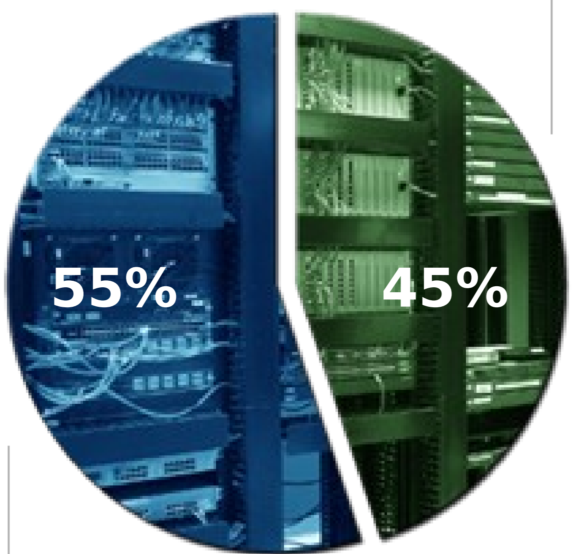
- Cooling
  - Higher Power demand = Increased Cooling demand
  - For every 1kW of power used to drive servers, on average another 0.5-1kW is necessary to drive the supporting infrastructure (such as power and air conditioning)

- Space
  - Existing data centers run out of space
  - Some data centers can't be expanded (eg on Wall Street)

Power Consumption - The food chain

Data center

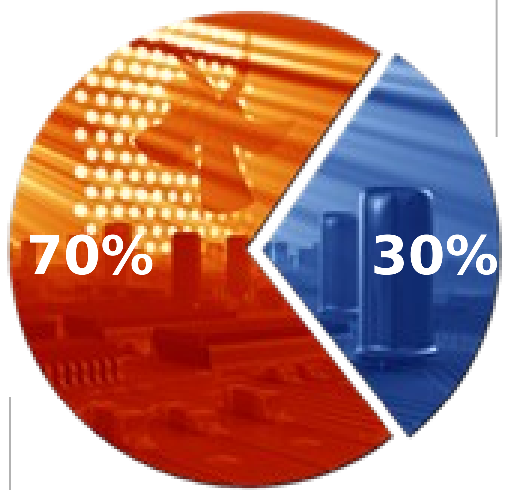
IT Load



Power and Cooling

Server hardware

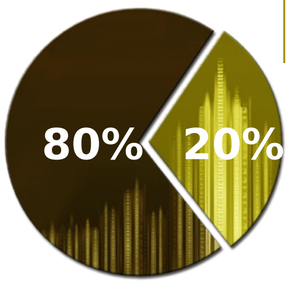
Processor



Power supply, memory, fans, mainboard, drives . . .

Server loads

Resource usage rate



Idle

Virtualized: 20-50%

Non-virtualized: 10-20%

27 W data center ← X1,7 — **16 W equiv IT power** ← x3 — **5 W equiv processor** ← x5 — **+1 W equiv. used ressource**

Power Consumption world wide

Microsoft job posting:
Nuclear power for Azure

"This senior position is tasked with leading the technical assessment for the integration of SMR and microreactors to power the data centers that the Microsoft Cloud and AI reside on."

Principal Program Manager

Multiple Locations, United States

Apply | Save | Share job

* No longer accepting applications

Date posted **Sep 25, 2023** Job number **1627555**

Overview

We're looking for a Principal Program Manager who will be responsible for maturing and implementing an energy strategy.

This senior position is tasked with leading the technical assessment for the Cloud. They will maintain a clear and adaptable roadmap for the technology's integration, diligently select and manage technology partners and solutions, and constantly evaluate the business implications of progress and implementation.

The ideal candidate will have experience in the energy industry and regulatory affairs. This role will also be responsible for research and developing other precommercial energy technologies.

You will be working with people from many different teams and backgrounds requiring an open mind-set. You must be able to identify and partner with other groups to achieve joint or complimentary goals. A proven track record of successfully managing projects, driving contractual improvements through service agreements and lower cost are skillsets needed to be demonstrated, resulting in measurable impact to be successful in this role.

In alignment with our Microsoft values, we are committed to cultivating an inclusive work environment for all employees to positively impact our culture every day.

Responsibilities

- Results oriented individual to help build and deliver the network that powers the world's largest online services
- Lead project initiatives for all aspects of energy infrastructure for global growth
- Select, onboard, and manage technology partners with a clear set of performance and outcome criteria
- Provide periodic technical assessment reports detailing progress, challenges, and implications on the broader business.
- Liaise with engineering and design teams to ensure technical feasibility and optimal integration of SMR and microreactor systems
- Own and drive commercial and contractual negotiations with vendors

Qualifications

Required Qualifications:

- Bachelor's Degree AND 6+ years' experience in nuclear industry, engineering, energy market, or related roles
  - OR equivalent experience.
- 3+ years experience managing cross-functional and/or cross-team projects.
- 3+ years' experience working on nuclear energy and associated technologies and/or nuclear regulatory affairs

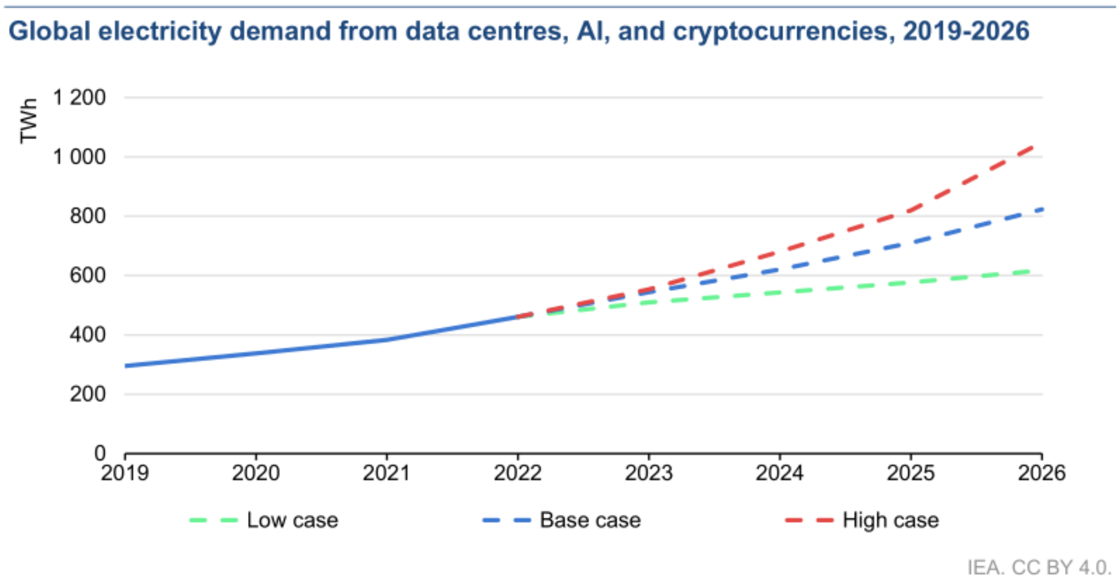
Background Check Requirements:
Ability to meet Microsoft, customer and/or government security screening requirements are required for this role. These requirements include, but are not limited to the following specialized security screenings: Microsoft Cloud Background Check: This position will be required to pass the Microsoft Cloud background check upon hire/transfer and every two years thereafter.

Preferred Qualifications:

- Master's Degree in Business, Engineering, Science, Economics, Business, or related field AND 4+ year(s) experience in engineering,

https://www.datacenterdynamics.com/en/news/microsoft-cloud-hiring-to-implement-global-small-modular-reactor-and-microreactor-strategy-to-power-data-centers/
https://jobs.careers.microsoft.com/global/en/job/1627555/Principal-Program-Manager

Power Consumption world wide



"This corresponds to an additional 160 TWh up to 590 TWh of electricity demand in 2026 compared to 2022,
roughly equivalent to adding at least one Sweden or at most one Germany."

Was ist ein Data Center?

"Unter 'IT-Betriebs-Bereich' sind Räume zu verstehen in denen die Hardware aufgebaut ist und betrieben wird, die der Bereitstellung von Diensten und Daten dient.
Das RZ umfasst neben dem IT-Betriebs-Bereich alle weiteren technischen Supportbereiche (Stromversorgung, Kälteversorgung, Löschtechnik, Sicherheitstechnik etc.), die dem bestimmungsgemäßen Betrieb und der Sicherheit des IT-Betriebsbereichs dienen."

BSI Grundschutzkompendium

> 8000 data centers worldwide:
- Carbon emission of IT industry equals aviation industry
- ICT emission is ~4% of global carbon emission (outlook 2040: 14%)

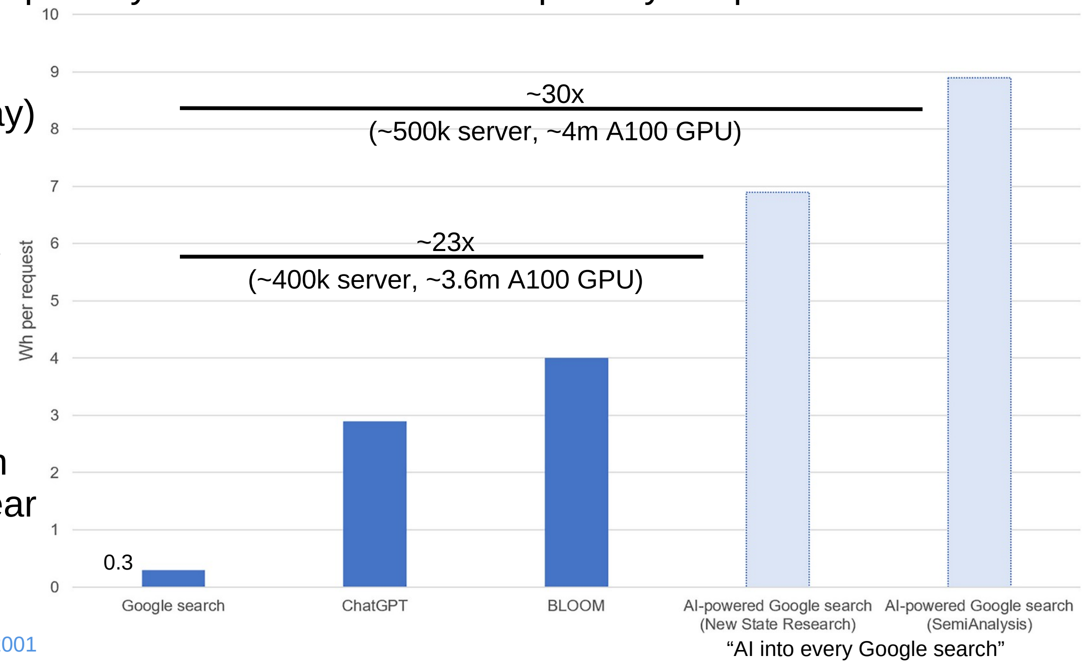
ICT consumes ~2% of world-wide electricity (EU 4%): 416 TWh in 2020, 460 TWh in 2022. Some expect 1200 TWh in 2026.
Example: Ireland expected to grow data centers from 17% (2022) to 32% (2026) of all power consumption.

https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf
https://www.researchgate.net/publication/341427004_Energy_consumption_of_data_centers_worldwide_How_will_the_Internet_become_green
https://www.nature.com/articles/d41586-018-06610-y

DHBW Duale Hochschule Baden-Württemberg Stuttgart

Power Consumption – AI and its impact

- Google
  - total data center power bill: 15 TWh in 2020, 18.3 TWh in 2021. In 2024, they demand nuclear
  - AI workload ~10-15%
- What is more expensive: training or inferencing?
  - GPT-3 training estimated 1287 MWh (using 10k V100 GPUs)
  - ChatGPT operation estimated 564 MWh per day .. translates to ~$700k per day in operational cost
    → estimated cost per request: $0.36
    (assumption: 3617 server with
    ~29k Nvidia A100 GPUs; 195m req/day)
  - BLOOM training cost: 433 MWh
    Inferencing: 914 kWh for 230k requests
    → ~4 Wh per req
  - Google 2019-2021 "60% of AI-related energy cost is from inferencing"
  - Google searches per day: up to 9 billion
    → turn it into AI search: 23...30 TWh/year
    (~ what Ireland consumes)

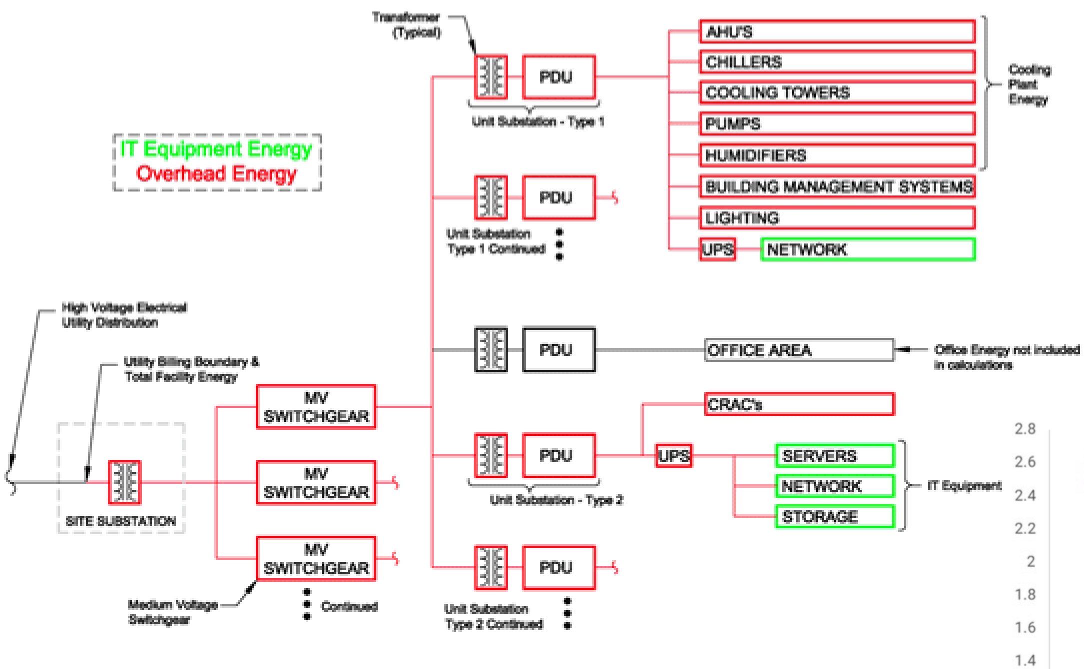


Sources:
Carbon Footprint of Bloom, 2022 - https://doi.org/10.48550/arXiv.2211.02001
Inference cost of Search disruption, 2023 - https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

Power Consumption - How to compare Data Centers

- PUE := Power Usage Effectiveness

$$PUE = \frac{\text{total data center power}}{\text{IT equipment power}}$$



Courtesy: google.com

Examples

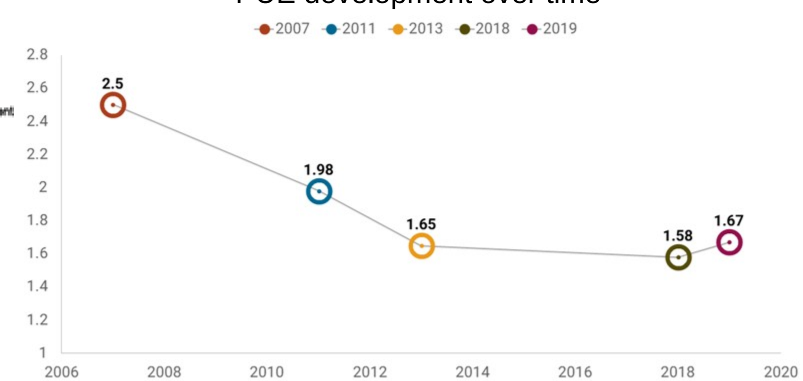
Typical legacy Data Center: 2.0

Improved Data Center: 1.9 – 1.7

Modern Data Center: 1.2 – 1.5

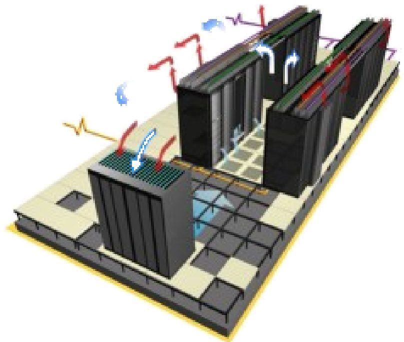
Google, Facebook: 1.0x

PUE development over time



Cooling – Technology evolves

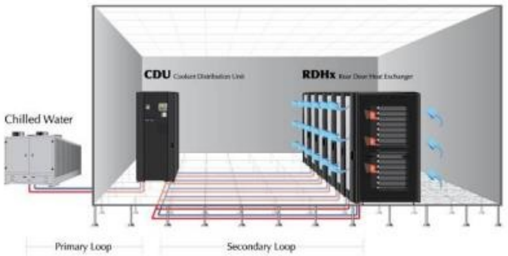
Air Cooled



- Standard air flow with internal fans cooled with the room climatization
- Broadest choice of configurable options supported
- Relatively inefficient cooling

PUE ~2.0 – 1.5

Air Cooled
with water-cooled rear door



- Air cooled but heat removed with special door through chilled water
- Retains high flexibility in choices (networking, storage, ...)
- Enables extremely tight rack placement

PUE ~1.4 – 1.2

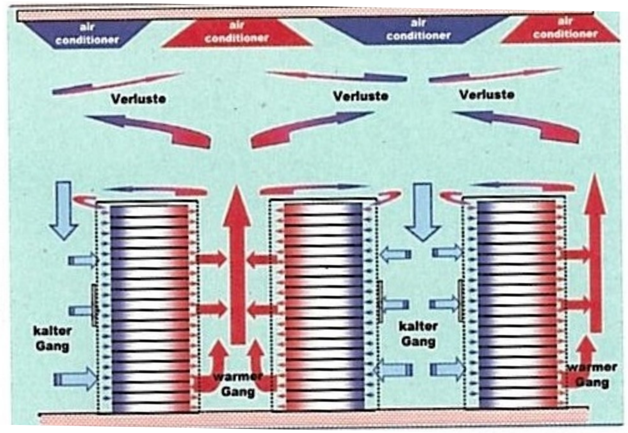
Direct Water Cooled

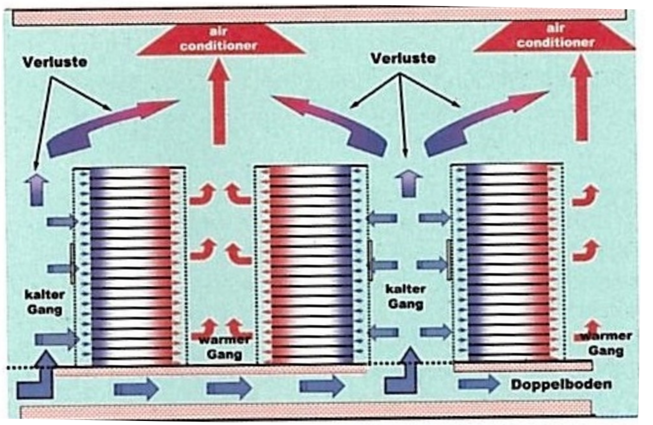
- Most heat removed by onboard-waterloop with up to 50°C temperature
- Supports highest TDP CPU at densest footprint
- Great system performance
- Free air cooling

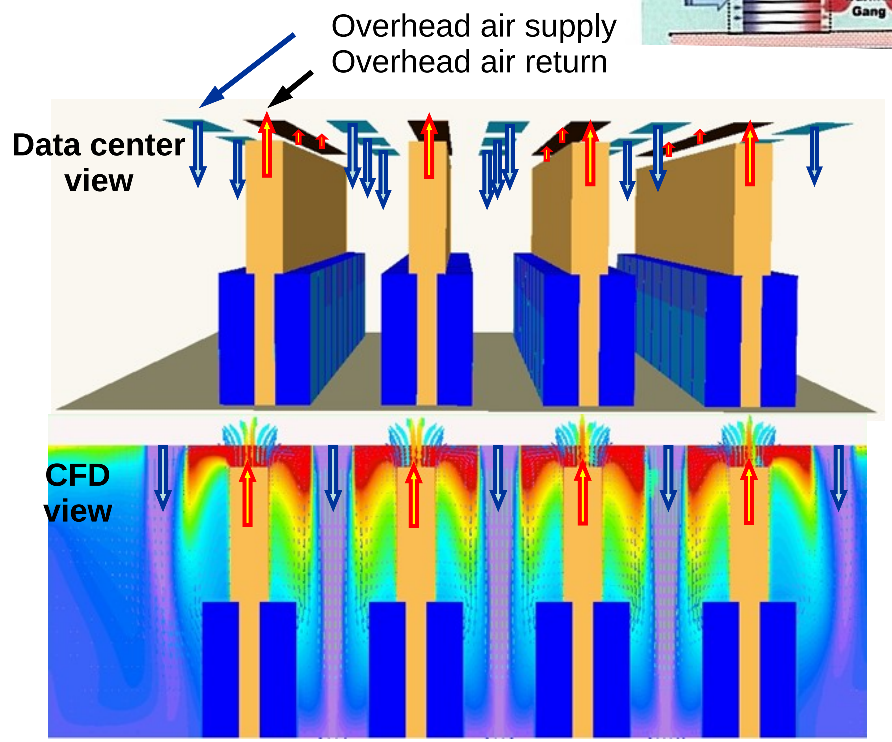
PUE <=1.1

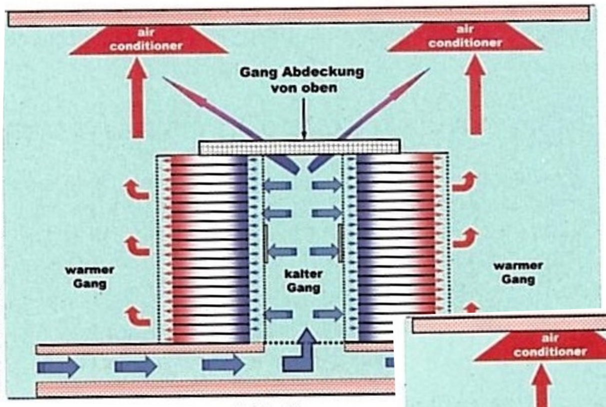
Future: Immersion liquid cooling?

Cooling – Concept of hot and cold aisles

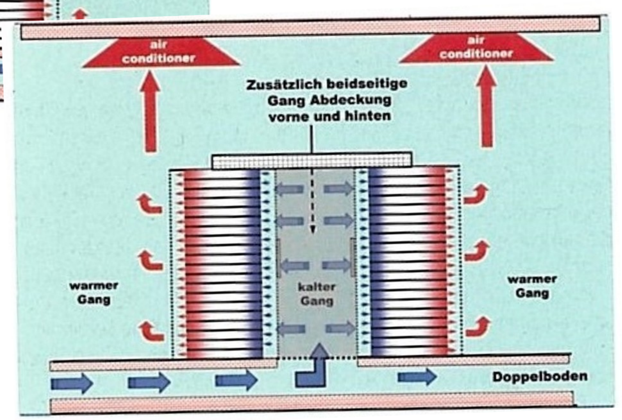








"Thermischer Kurzschluss"



Source: heise/iX

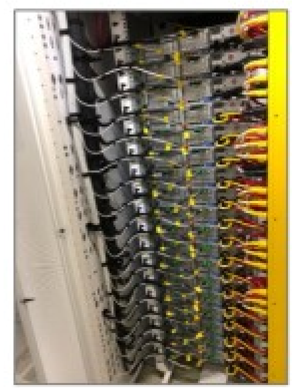
Cooling – Small but effective improvements

Blanking Panels

Cool boots
(for network cables)

Sealed outlets

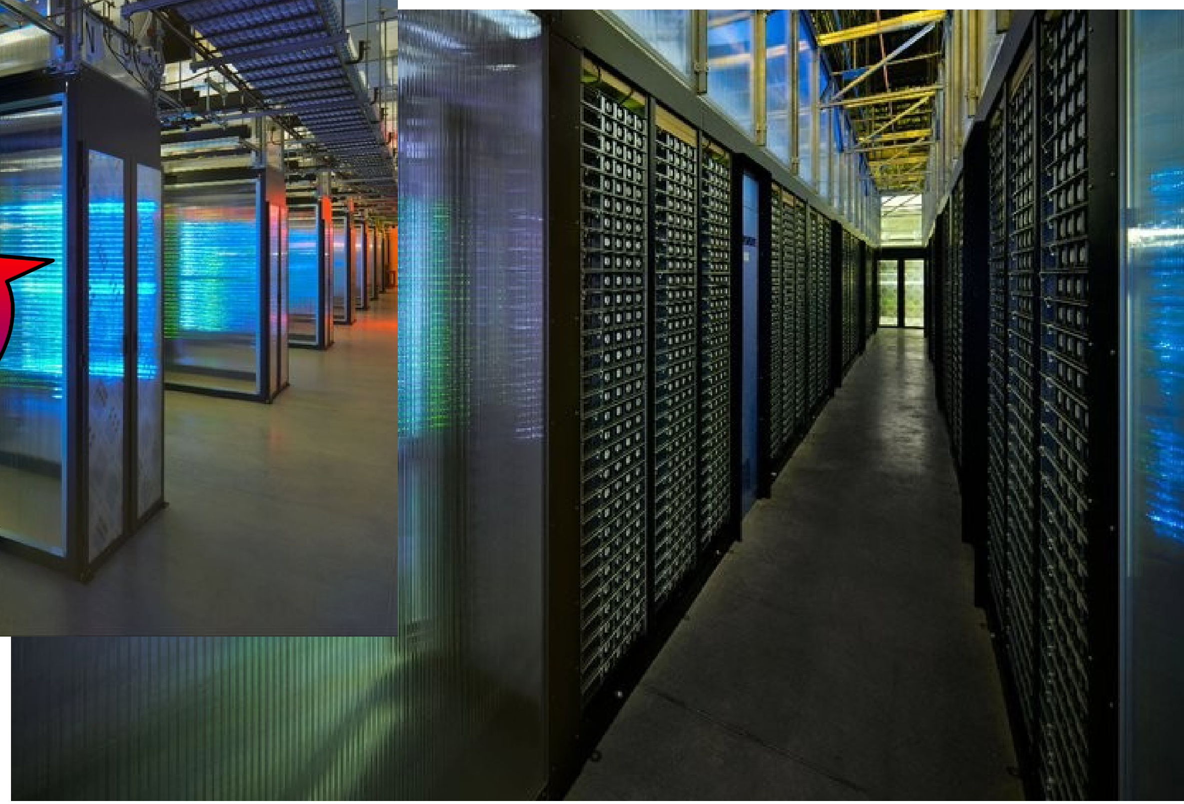
Proper rack cabling
for optimized airflow



Source: https://news.sap.com/2022/06/sap-cloud-strategy-focuses-on-green-data-centers/

Cooling @ Facebook

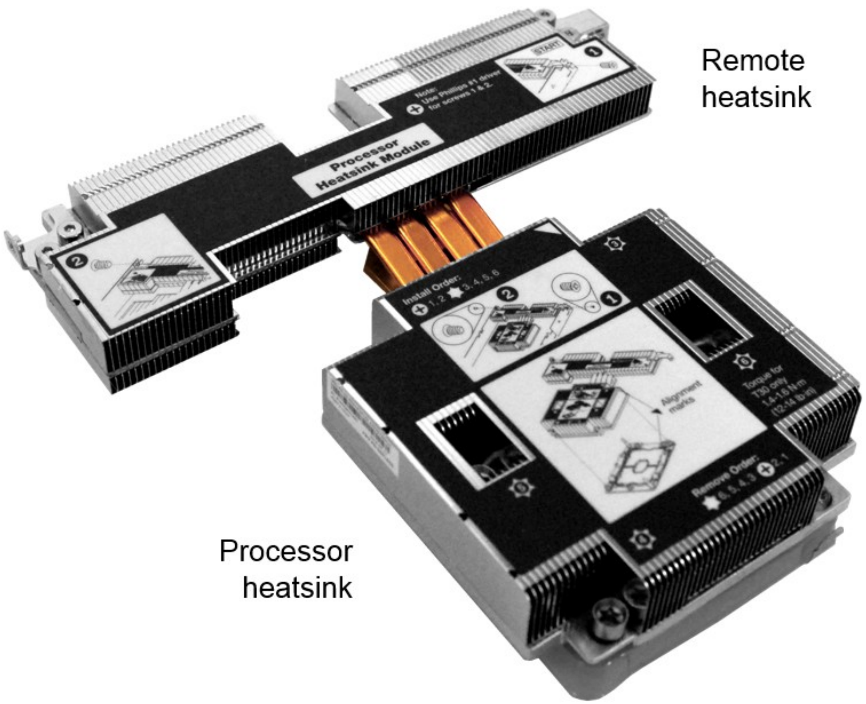
- Facebook data center in Prineville, OR (PUE < 1.1)



-> Hot / cold aisle containment
-> Controlled air flow path

Cooling – thermal transportation loops

- Idea: transport heat from hot spot to area with better/easier air cooling capabilities



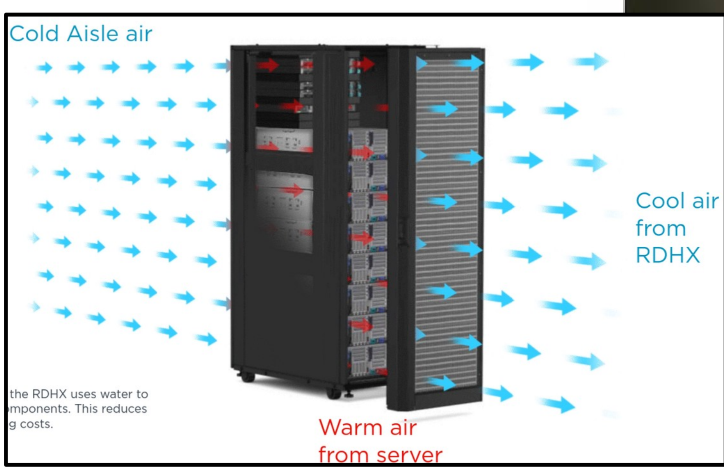
Remote
heatsink

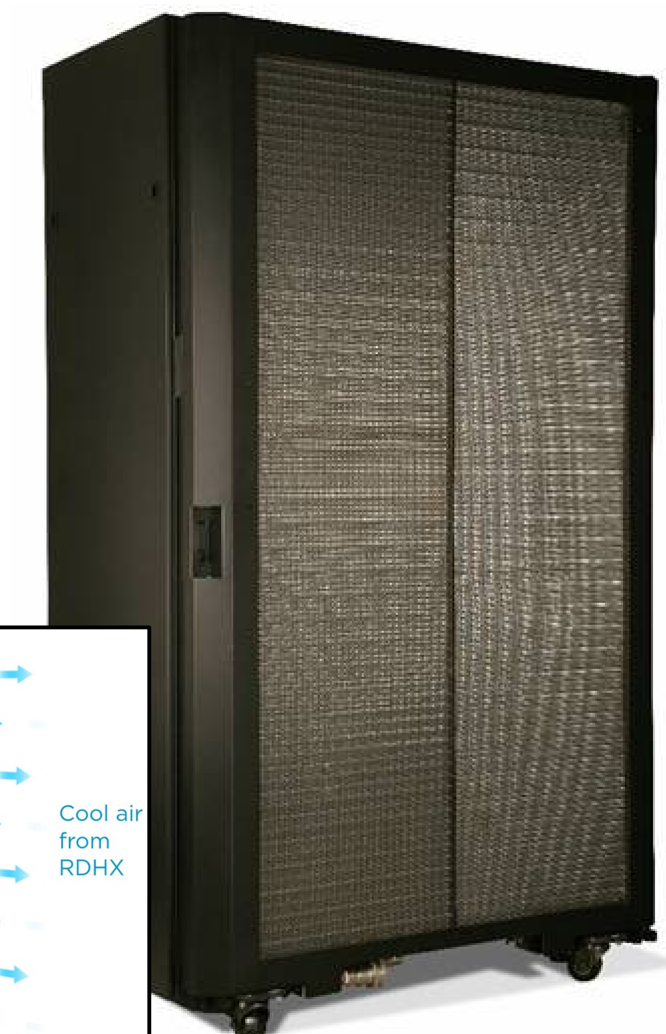
Processor
heatsink

Cooling – Liquid is better than air

Step 1:
Water to the rack

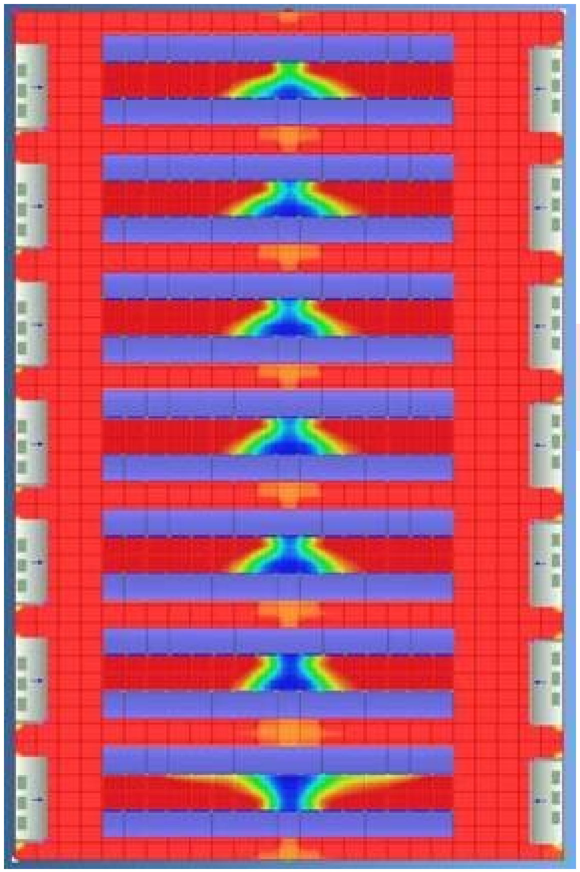
- Water is easiest
  - Water-cooled rear door
  - Rear-door heat exchanger (RDHX)

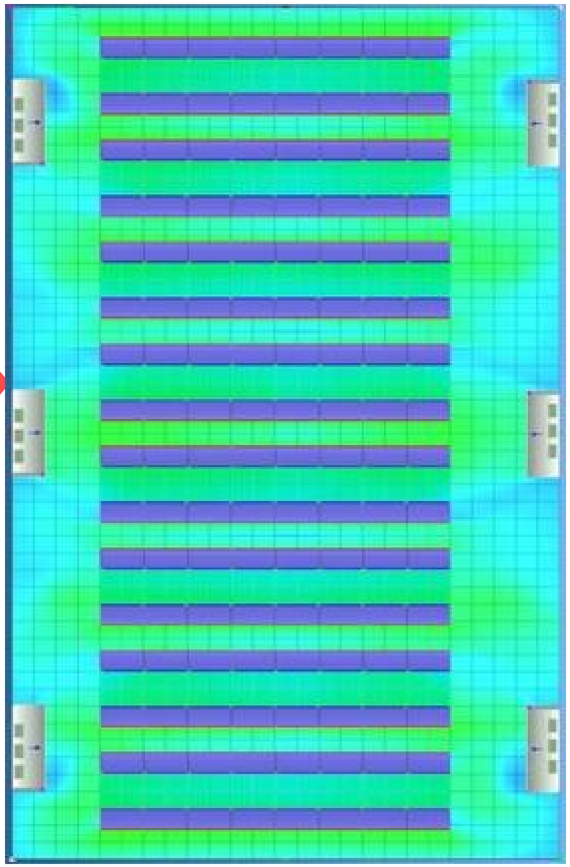




Cooling – efficiency of water-cooled doors

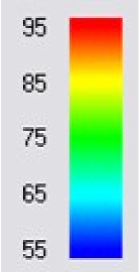
Only 12% racks are equal to or below 25C





100% racks are equal to or below 25C

less CRACs



Fahrenheit

9.400 Servers

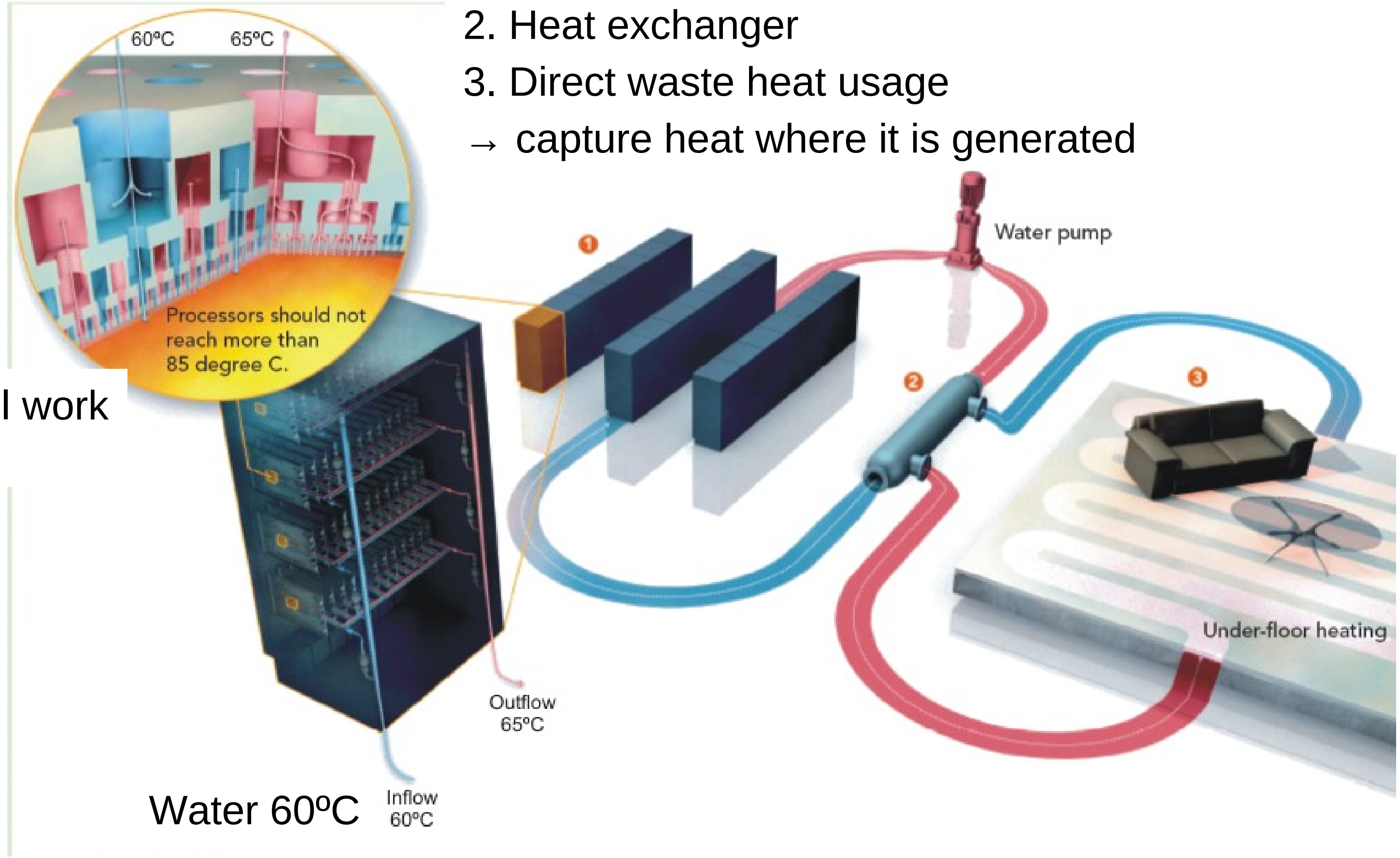
10.700 Servers

Idea of a direct-water cooled system (DWC)

Step 2:
Water to the server

1. Special heat-sinks (micro-channel liquid coolers)
2. Heat exchanger
3. Direct waste heat usage

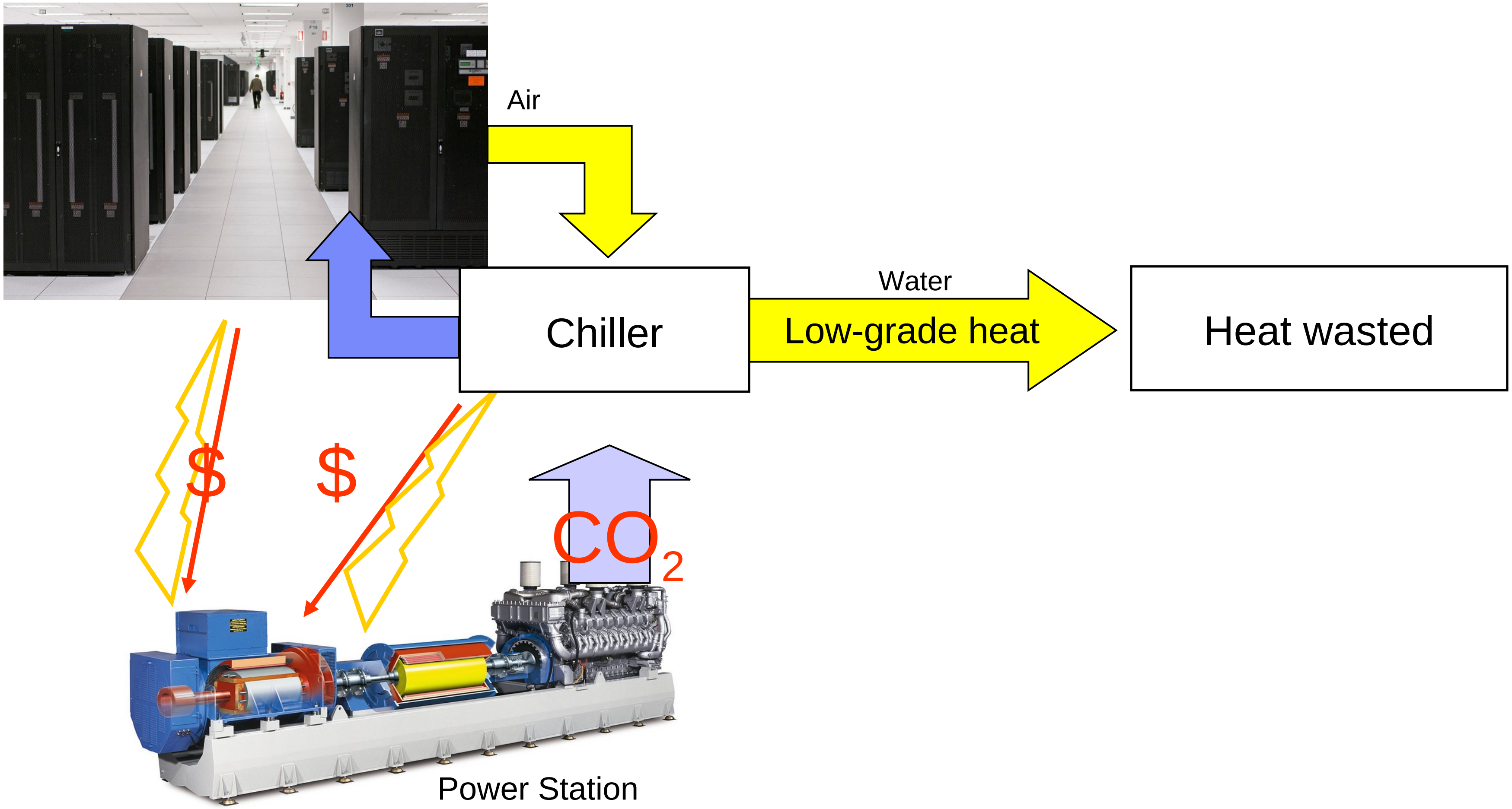
→ capture heat where it is generated



Chips still work at 80ºC

Water 60ºC

Step 1/3: Conventional Data Center

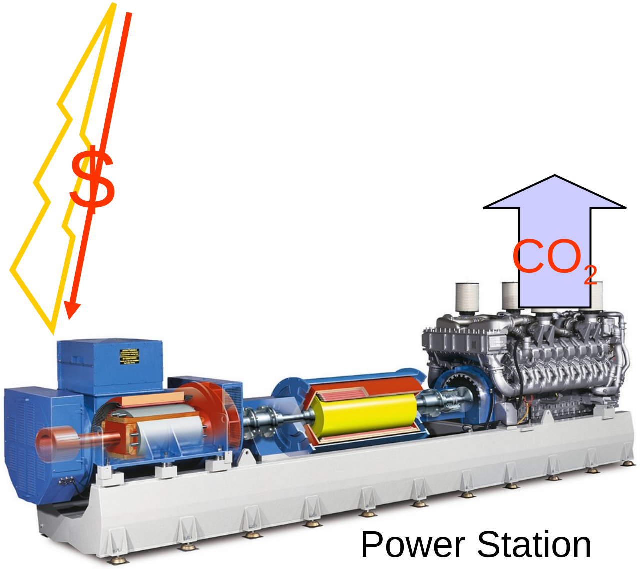


Step 2/3: Towards Zero–Emission Data Center (CO_2 ~50% reduced)

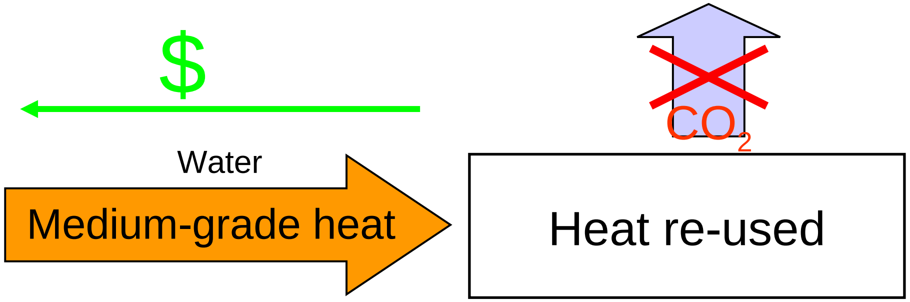
Water

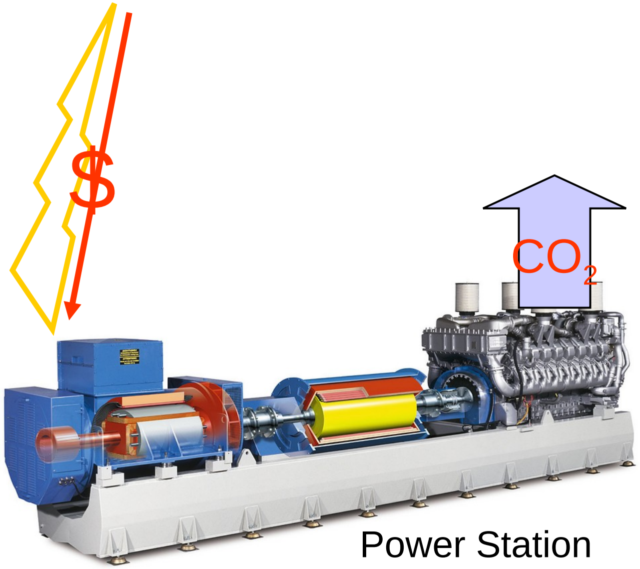
Medium-grade heat

Heat wasted



Step 3/3: Zero–Emission Data Center (CO_2 < 15 % net)

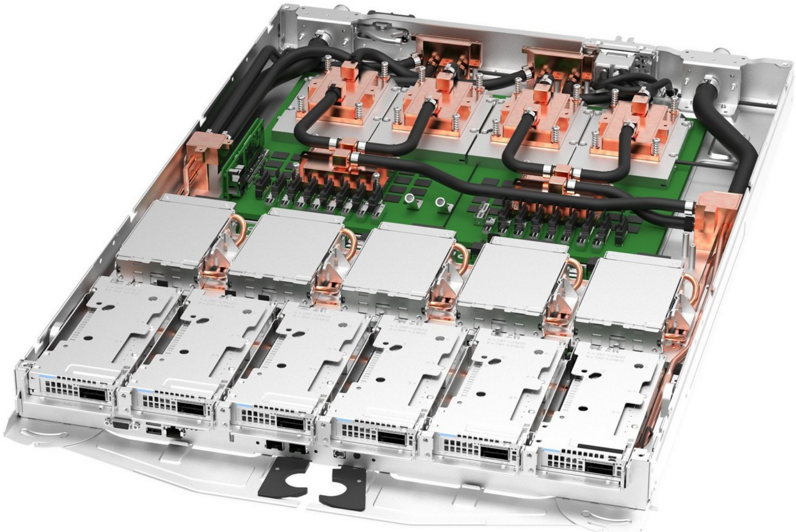


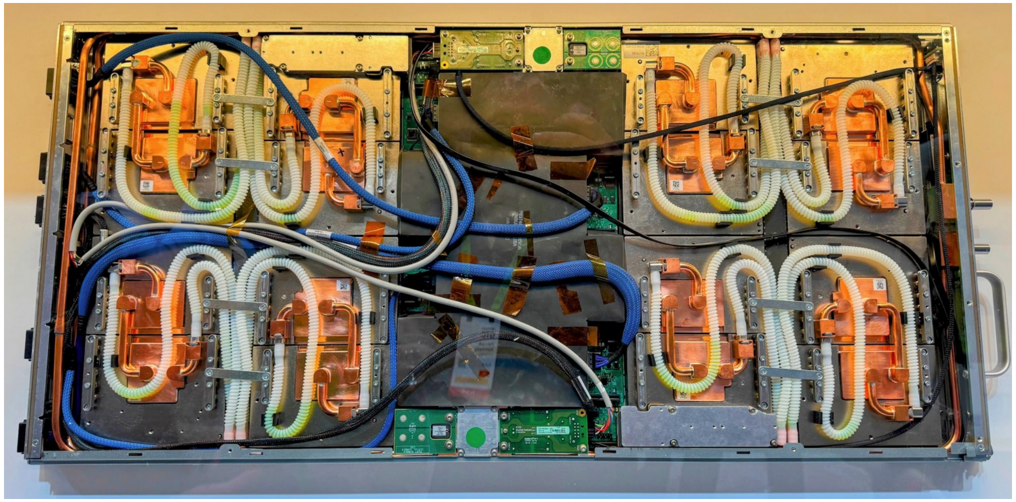


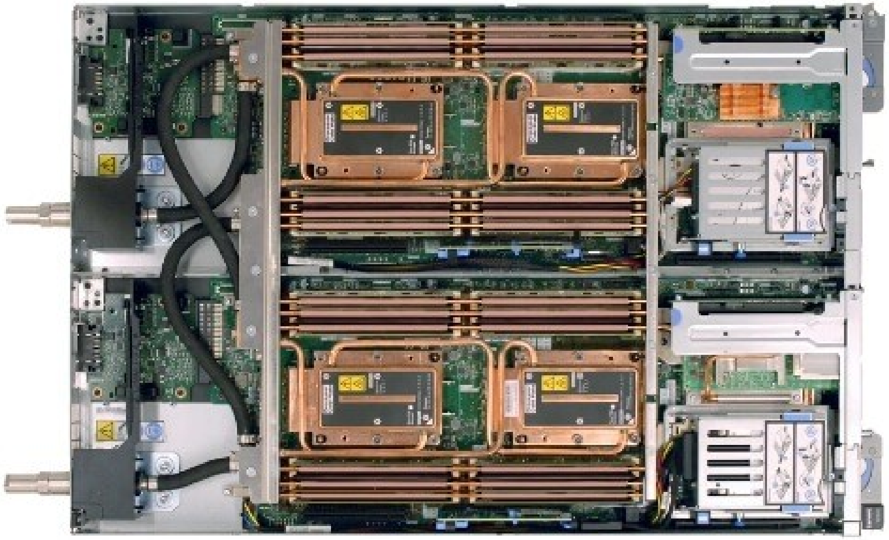
Power Station

Further reading:
https://www.climateneutraldatacentre.net/ and
EU Energy Efficiency Directive – Rules and Obligations for 2020 and 2030 efficiency targets (‘European Green Deal’)
https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en

How such a direct-water cooled system looks like







Challenges – Power and Cooling outlook

- Power
  - More cables into server :)
  - Example with GPU server:
    - 3 rack unit
    - 2x CPU + 8x GPU
    - = 600W + 6400W + x = 7kW
  - Up to 16 server in a rack
    → >**100kW per rack**

- Cooling
  - Extract 100kW++ from single rack
  - We are at maximum efficiency of air
  - More air into rack … No!

- Expectation: up to 200kW in 2025+
  - Only sustainable with direct water cooling

Big challenge these days:
- make liquid cooling mass-consumable
- make liquid cooling feasible for small installations

- Scale-out examples
- Challenges – Power consumption
- Challenges – Cooling
  - How to cool server and data centers
  - the path to liquid based cooling

- Scale-out for scientific workload: High Performance Computing (HPC)
  - Basics
    - Benchmarking a scale-out supercomputer
    - Architecting and operating a scale-out supercomputer
  - Examples
    - Mare Nostrum, Barcelona
    - Hawk, Stuttgart
    - COVID-19 special
- Scale-out for commercial workload
  - Public clouds (Google)
  - Social media

Benchmarking & Performance comparison

- HPL – High Performance Linpack Benchmark
  - Solves mathematical problem
    Random dense NxN linear system (Ax=b), using double-precision 64-bit arithmetic
  - Measures floating-point 64-bit computing power
  - Extremely scalable

  Result: floating-point operations per second (FLOPS)
    - Intel Desktop CPUs ... up to 400 GFLOPS
    - Intel Server CPUs ... up to 6500 GFLOPS
    - Nvidia Tesla H100 GPUs ... up to 60.000 GFLOPS (FP64!)

- new: HPCG – High Performance Conjugate Gradients
    Multigrid preconditioned conjugate gradient calculations, using double-precision 64-bit arithmetic
    (Verfahren der konjugierten Gradienten mit Vorkonditionierung, via Gauss-Seidel)
  - Represents more modern, real-world code
  - Result: GFLOPS, see HPL

- New AI: MLperf – suite of benchmarks for ML training and inferencing
  - Object detection, classification, recommendation, ..
  - measures 8-bit/16-bit performance

Top500 list – November 2025 (top500.org)

#	System	Manufac-turer	Site	Country	Cores	Rmax (PFlops)	Power (MW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS	HPE	LLNL, DOE	US	11,340,000	1809.00	29.685
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X	HPE	Oak Ridge, DOE	US	9,066,176	1353.00	24.607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU	HPE	Argonne, DOE	US	9,264,128	1012.00	38.698
4	**JUPITER** – Bull XH3000, GraceHopper 72C 3GHz, NVIDIA GH200, NVIDIA Infiniband	Eviden	FZ Juelich	Germany	4,801,344	1000.00	15.794
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband	MS	MS Azure	US	2,073,600	561.20	
6	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9	HPE	Eni S.p.A.	Italy	3,143,520	477.90	8,461
7	**Fugaku** - A64FX 48C 2.2GHz, Tofu interconnect	Fujitsu	Riken	Japan	7,630,848	442.01	29.899
8	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS	HPE	CSCS	Switzerland	2,121,600	434.90	7,124
9	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X	HPE	CSC	Finland	2,752,704	379.70	7.107
10	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband	Eviden	Cineca	Italy	1,824,768	241.20	7.494

10^21 Zetta // 10^18 Exa // 10^15 Peta // 10^12 Tera // 10^9 Giga

High performance benchmarking

Green500 List ranks Energy Efficiency: Performance per Watt

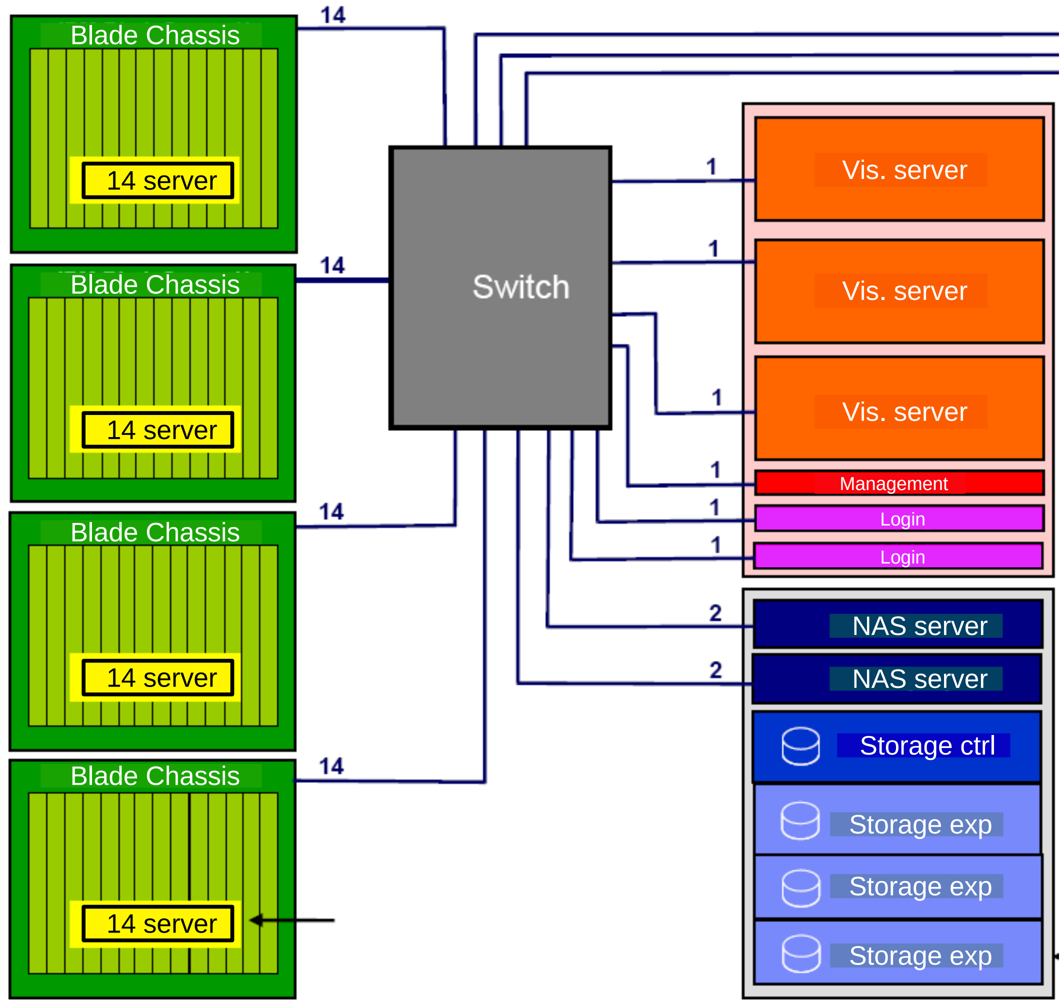
#	System	CPU	Interconnect	Accelerator	Power (kW)	GFlops/W
1	KAIRO - Bull Eviden XH3000	NVIDIA Grace 72C 3GHz	Infiniband NDR200	NVIDIA GH200 Superchip	46	**73.3**
2	ROMEO-2025 - Bull Eviden XH3000	NVIDIA Grace 72C 3GHz	Infiniband NDR200	NVIDIA GH200 Superchip	160	**70.9**
3	Levante GPU Extension – Bull Eviden XH3000	NVIDIA Grace 72C 3GHz	Infiniband NDR200	NVIDIA GH200 Superchip	110	**69.4**
4	Isambard-AI phase 1 - HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA GH200 Superchip	117	**68.8**
5	Otus – Lenovo ThinkSystem SD665-N V3	AMD EPYC 9655 96C 2.6GHz	Infiniband NDR400	Nvidia H100 80GB		**68.1**
6	Capella - Lenovo ThinkSystem SD665-N V3	AMD EPYC 9334 32C 2.7GHz	Infiniband NDR200	Nvidia H100 94GB	445	**68.1**
7	SSC24-Energy Module – HPE Cray XD670	Intel Xeon 6430 32C 2.1GHz	Infiniband NDR400	Nvidia H100 80GB	69	**67.3**
8	Helios GPU - HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA GH200 Superchip	317	**66.9**
9	AMD Ouranos – Bull XH3000	AMD 4th Gen EPYC 24C 1.8GHz	Infiniband NDR200	AMD Instinct MI300A	48	**66.4**
10	Portage – HPE Cray EX255a	AMD 4th Gen EPYC 24C 1.8GHz	Slingshot-11	AMD Instinct MI300A	370	**66.3**

Of the top 100 systems on the Green500, all leverage GPU accelerators.

→ General-purpose CPUs are not at all energy efficient

Architecting HPC cluster (general purpose, not AI/ML specific)

- This is a real life example
- What we see
  - Compute server (green)
  - Login server (pink)
  - Management server (red)
  - I/O server (blue)
  - Highspeed Interconnect (grey)
    optional:
  - Visualization server with GPUs (orange)

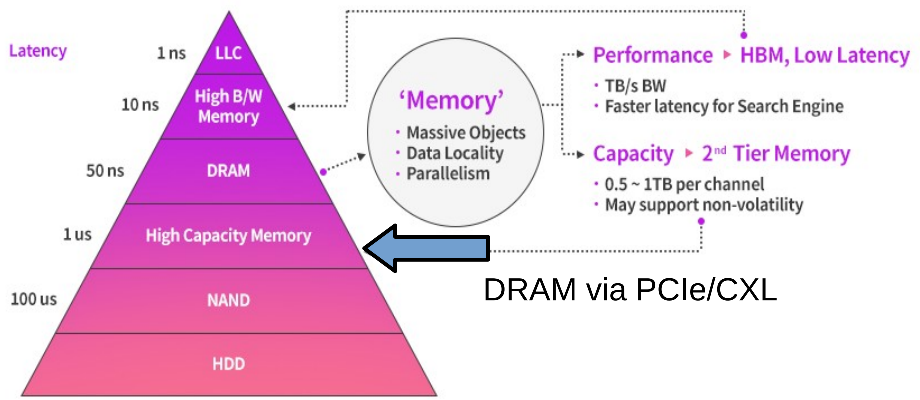


Example:
maximum Ethernet switch
- 800x 200 Gbit Eth
- 1600x 100 Gbit Eth

Different architectures:
Fat tree
Spine-leaf
Dragon Fly

How HPC impacts our 2024 IT Architectures (not just for AI)

Memory hierarchy changes: Tiering will be important



DRAM via PCIe/CXL

Source: SK Hynix

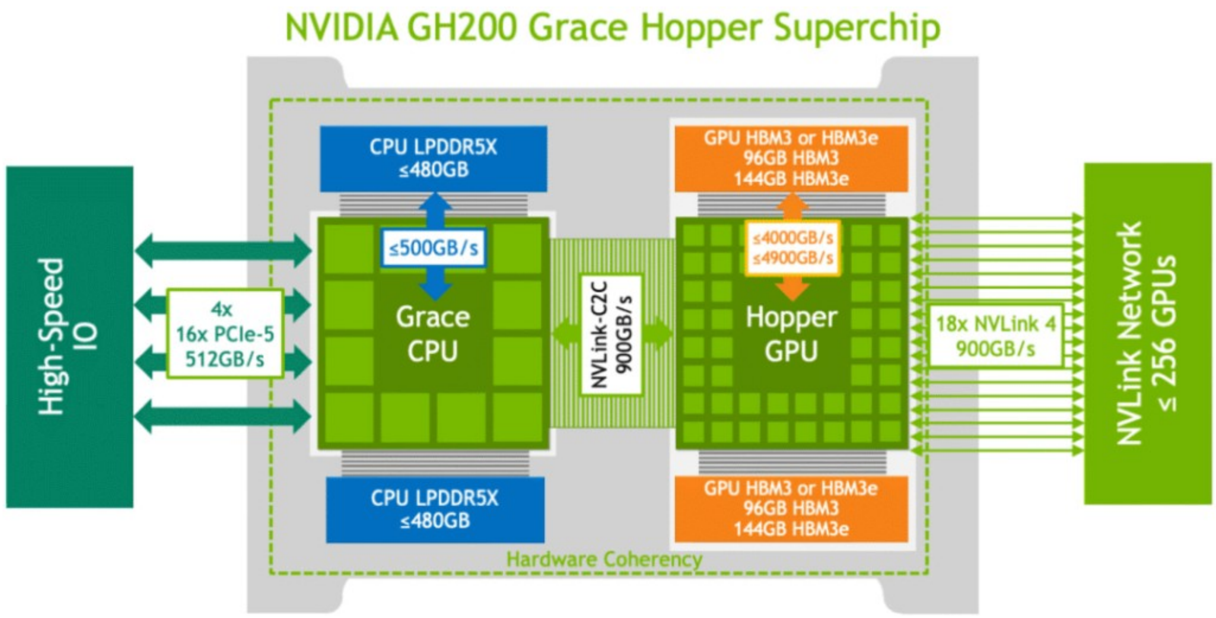
1. *HPC workload is all about speed and efficiency*
2. *to exploit modern GPU, a fast data path is required*
3. *AI wouldn't be possible without recent HPC innovations*

So much is in-memory already:
- Redis
- Aerospike Graph
- commercial, like SAP HANA or Oracle/DB2

Architecting AI cluster – NVIDIA Hopper architecture

NVIDIA sells Hopper in two architectures (products):

1. Hopper GPU (H100, H200 → standard PCIe card into Intel/AMD server)
2. Grace Hopper → heterogeneous superchip (CPU and GPU on a combined module)
   Grace CPU: ARM 72c, up to 1TB memory
   Hopper GPU: Tensor Cores



Hopper architecture details

https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

https://resources.nvidia.com/en-us-data-center-overview/hpc-datasheet-sc23-h200

NVIDIA Hopper H200 GPU

Technical Specifications	H200 SXM[1]
FP64	34 TFLOPS
FP64 Tensor Core	67 TFLOPS
FP32	67 TFLOPS
TF32 Tensor Core[2]	989 TFLOPS
BFLOAT16 Tensor Core[2]	1,979 TFLOPS
FP16 Tensor Core[2]	1,979 TFLOPS
FP8 Tensor Core[2]	3,958 TFLOPS
INT8 Tensor Core[2]	3,958 TFLOPS
GPU Memory	141GB
GPU Memory Bandwidth	4.8TB/s

Architecting AI cluster

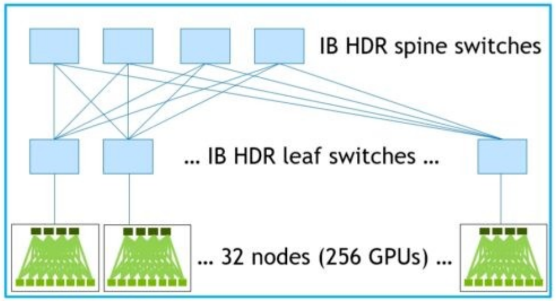
- Goal
  - connect as many GPUs (with GPU memory) as possible (→ “virtual scale-up”)
  - with the fastest interconnect and lowest latency
- Challenge
  - PCIe is slow (Gen5 ‘only’ 128 GB/s)
  - Ethernet is slow
- Different approaches
  - Open: Infiniband
  - Proprietary: NVLink

Azure 2023:
4000 GPU environment (using open architecture)

DeepL 2023:
68 H100 SuperPods

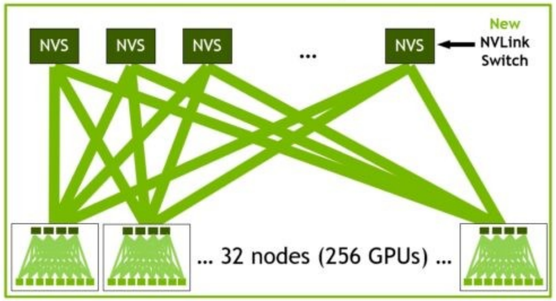
Open

DGX A100 256 SuperPOD



Proprietary

DGX H100 256 SuperPOD



Fully NVLink-connected
Massive bisection bandwidth

	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

Source: https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

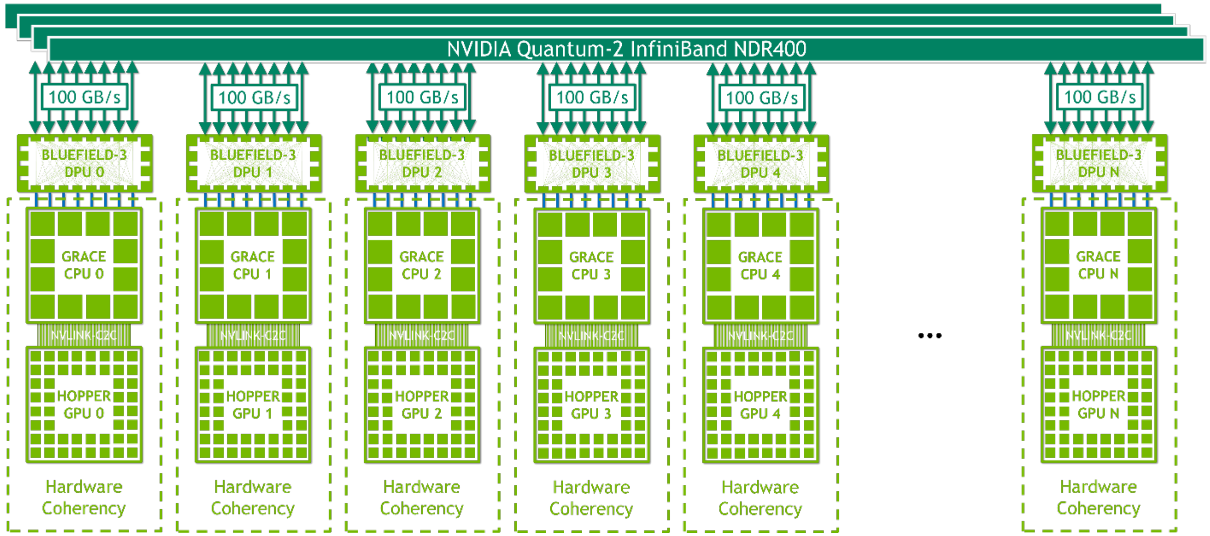
DHBW
Duale Hochschule
Baden-Württemberg
Stuttgart

Architecting AI cluster – open and proprietary (detail)

Open

Communication has to go via CPU
(GPU → CPU → NIC)

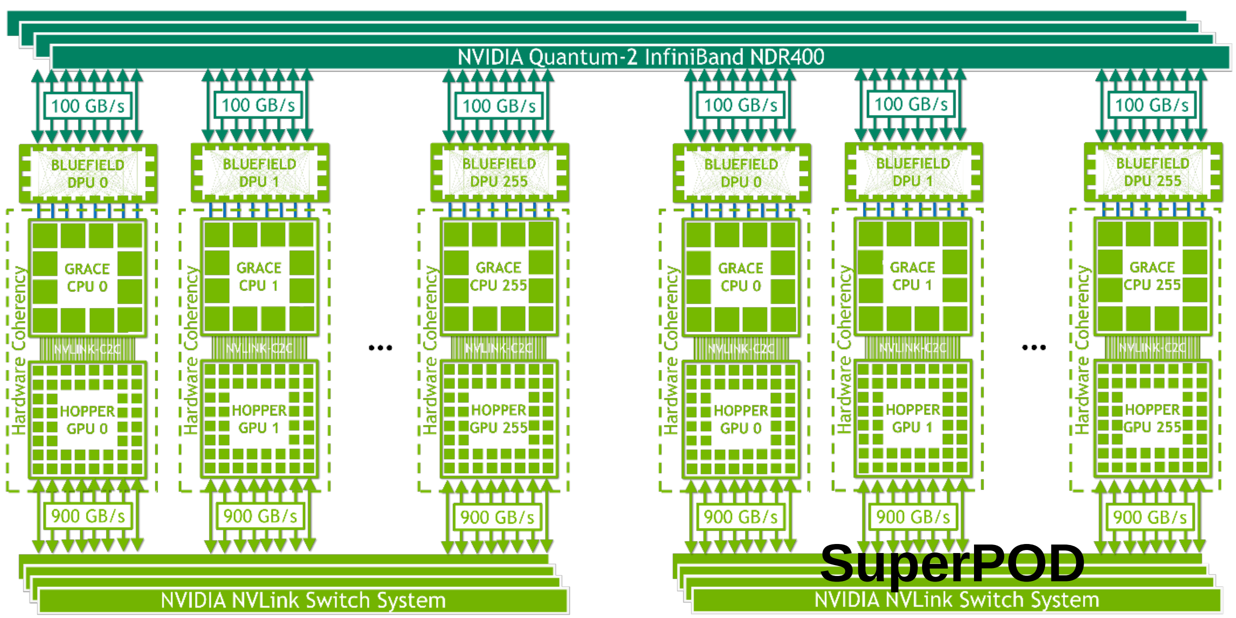
Latency: higher
Scalability: endless..



Proprietary: NVLink

Direct GPU-to-GPU communication

Latency: lower
Scalability: limited to NVLink capabilities (currently 256 GPUs with 256x 144GB = 36864 GB total memory)



Architecting AI clusters – sizing the infrastructure

1. Example: LLM inferencing aka Chatbot

GPU memory (in GB) = parameter size * quantization factor * 1.2

Parameter size (in billion)

Quantization factor (in bytes):
16-bit / FP16 → 2 bytes
8-bit / FP8 / int8 → 1 byte

Factor ~1.2: allow for some overhead

Example parameter sizes:
Llama 3.3: 70B
Llama 3.2: 90B, 11B, 3B, 1B
GPT-4o: 12B
BLOOM: 176B
Mistral Large 2: 123B

2. Example: LLM fine-tuning at 16 bit

GPU memory (in GB) = parameter size * 16 GB
Relevant information: parameter size, optimizer states, gradients, activations

Method	Precision	7B	13B	30B	70B
Full	16 bit	67 GB	125 GB	288 GB	672 GB
LoRA	16 bit	15 GB	28 GB	63 GB	146 GB
QLoRA	8 bit	9 GB	17 GB	38 GB	88 GB
QLora	4 bit	5 GB	9 GB	20 GB	46 GB

Training sizing: https://modal.com/blog/how-much-vram-need-fine-tuning

MegaScale: Scaling LLM Training to More than 10,000 GPUs (Feb 2024)
https://arxiv.org/pdf/2402.15627

Cluster Software (Operating)

- Management Software
  - mostly Open Source
  - Collection of scripts, Expect, Python, bash, ..
  - xCAT/Confluent, extremely flexible (steep learning curve)
- Resource Manager & Job Scheduler (policies, priorities, reservations, ...)
  comparable to Map-Reduce architecture
  - Slurm
  - Apache Mesos, Aurora (→ X/Twitter)
  - IBM Spectrum LSF Suite (Load Sharing Facility)
- Performance Monitoring
  - Ganglia, Grafana (time-series based databases)
  - Collects data on the compute servers (CPU load, memory util, ..)
- Programming APIs
  - Shared memory systems (single machine, NUMA): OpenMP
  - Distributed memory system (cluster nodes): MPI

→ **Many concepts flow into Kubernets and modern orchestration tools**

Barcelona – Mare Nostrum Chapel

Hunter – HLRS@Uni Stuttgart

HPE Cray EX4000
Number of compute nodes: 256
Number of GPU nodes: 188
System peak performance: 45 Petaflops

System information
CPU: AMD Epyc 9374F 24C 1.8 GHz
APU: AMD Instinct MI300A
Each server: 2x CPU + 4x APU
Interconnect bandwidth: 200 Gbit/s

Power consumption
Average power consumption: ~560 kW

Cost ~EUR 15 million
#61 on Top 500 Nov'25 list
#54 in Jun'25 list
'Herder' to be installed in 2027
EUR 100m for next-gen

Entgelt/Gebühr netto für die Aufgabengruppen Nr. 3 und 4 nach § 2 Abs. 1 in Euro / RTStd ***

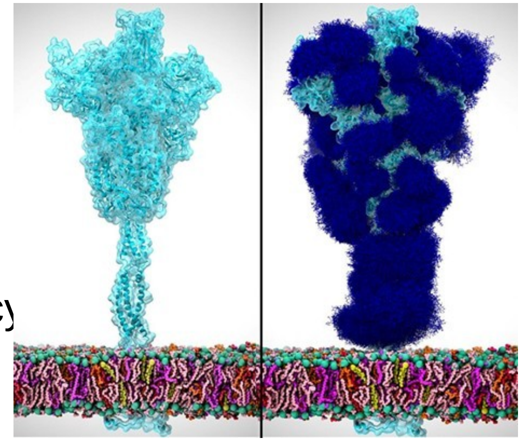
Batchbetrieb / Grundpriorität

	bwuni-cluster *	Vulcan									Hunter		Hawk			
Knotentyp	alle *	smp Knoten 1,5 TB	Skylake 192 GB	Cascade-lake 384 GB	Graphik-server AMD	Graphik-server NVIDIA	CS 500 - Spark 384 GB	Genoa Compute 768 GB	Genoa A30 GPU 768 GB	Genoa SMP 3 TB	CPU Erweiterung 768 GB	GPU 512 GB	KI Erweiterung Apollo 6500	Hawk Compute 256 GB	Pre- und Post-processing 2 TB	Pre- und Post-processing 4 TB
Rechenzeit		Resource Time 0,26 €	Knoten 0,76 €	Knoten 0,71 €	Knoten 1,09 €	Knoten 1,50 €	Knoten 0,43 €	Knoten 1,21 €	Knoten 1,32 €	Resource Time 0,06 €	Knoten 1,49 €	Knoten 8,07 €	Knoten 3,21 €	Knoten 0,39 €	Resource Time 0,01 €	Resource Time 0,03 €

HLRS Entgeltordnung:
https://www.hlrs.de/for-users/costs-fee-schedules

DHBW Duale Hochschule Baden-Württemberg Stuttgart

COVID-19 special (1/2)



- Research areas (many different)
  - Understand the virus
  - Identify a vaccine (both require lots of protein folding)
  - Model the outbreak, how does it spread, mobility of people (policy development), example: aerosole spread

Desktop performance:
~100 atoms, ~10 picosec

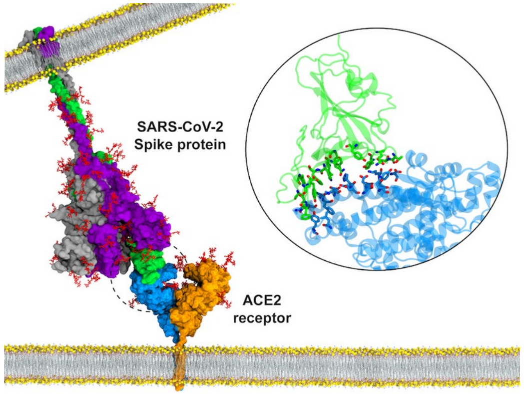
Without HPC with HPC

Corona virus is covered in a sugar shield.

Atomic-level view (~1.7 million atoms).
256 server simulate ~ 60 ns/day

- Artificial Intelligence / data mining in action
  - Brute force approach: harvest through drug data to identify possible vaccines
    Drug preselection with help from Community https://covid.postera.ai/covid
  - AI example: Illness prediction through wearables (TemPredict, UC San Diego, multiple thousand patients who donated data)

  - Folding @ home (→ more than 1 Exaflop/s total compute power)
    produced a protein data set, which Nvidia visualized
    https://youtu.be/jiMZYJ--cT8 (2min 25s)



Intel Xeon Phi accelerator, 10 Pflops
Identified an intermediate state which can be targeted to prevent spreading.

https://covid19-hpc-consortium.org/
https://covid.molssi.org/
https://www.hpcwire.com/2020/10/30/whats-new-in-computing-vs-covid-19-bars-visualizations-new-therapeutics-more/
https://www.hpcwire.com/2020/10/14/how-foldinghome-identified-and-visualized-sars-cov-2s-weak-spots/

COVID-19 special (2/2)

- Japan, Fugaku

Digital transformation of droplet/aerosol infection risk assessment realized on "Fugaku" for the fight against COVID-19

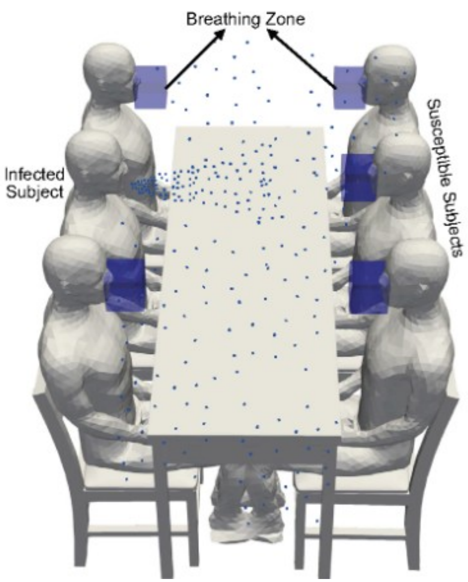


Table 8: Total simulation cases, sizes, and computational costs.

Release Date	Targets	Total mesh num.	Nodes /case	Node-hours /case	Total cases	Total Node-hour
June, 17th, 2020	Surgical masks (Gap)	29M	146	438	70	30,660
June, 17th, 2020	Humidity and droplet transport on the table	34M	206	9,888	32	316,416
June, 17th, 2020	Commuter train (windows opening)	163M	415	29,880	10	298,800
June, 17th, 2020	Four-seated table (partitioning)	34M	206	9,888	32	316,416
June, 17th, 2020	Office (windows opening etc.)	50M	511	24,528	10	245,280
June, 17th, 2020	Hospital room (droplet dispersion)	48M	291	20,952	18	377,136
August, 24th, 2020	Surgical and fabric masks	29M	146	438	65	28,470
August, 24th, 2020	Face masks to protect against infection	35M	178	534	35	18,690
August, 24th, 2020	Hospital room (A/C effect)	48M	291	20,952	32	670,464
August, 24th, 2020	Office (Partitioning and A/C)	50M	511	24,528	10	245,280
August, 24th, 2020	Classroom (windows opening)	47M	330	7,920	16	126,720
August, 24th, 2020	Concert hall (ventilation)	135M	1,369	65,712	10	657,120
August, 24th, 2020	Concert hall (audience seats)	72M	440	21,120	32	675,840
October, 13th, 2020	Masks to protect against infection	35M	178	534	10	5,340
October, 13th, 2020	Speaking/singing/coughing	21M	158	7,584	24	182,016

https://arxiv.org/abs/2110.09769

https://blogs.nvidia.com/blog/2020/11/19/covid-ai-gordon-bell-winner/

- Scale-out examples
- Challenges – Power consumption
- Challenges – Cooling
  - How to cool server and data centers
  - the path to liquid based cooling

- Scale-out for scientific workload: High Performance Computing (HPC)
  - Basics
    - Benchmarking a scale-out supercomputer
    - Architecting and operating a scale-out supercomputer
  - Examples
    - Mare Nostrum, Barcelona
    - Hawk, Stuttgart
    - COVID-19 special
- Scale-out for commercial workload
  - Public clouds (Google)
  - Social media

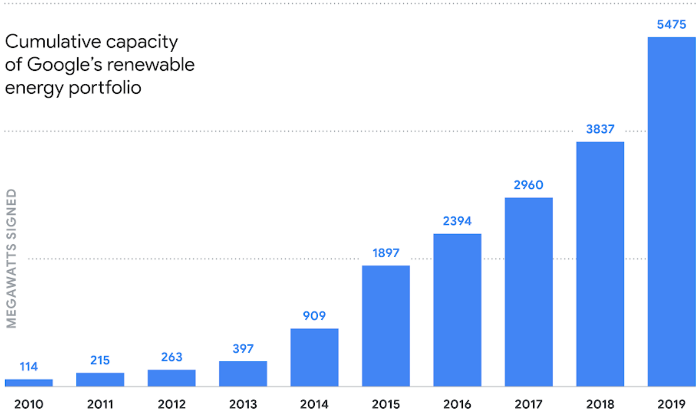
Commercial data centers – Power consumption

ByteDance: in 2020 bought data center space with >130MW

NSA: top data center Utah, 65 MW

Facebook: plans for single data center with solar power worth 440 MW in Oregon
- world-wide 586 MW (2019)

Google: bought 5475 MW renewable energy until 2019
- Finland (255 MW), Sweden (286 MW), Belgium (92 MW), Denmark (160 MW)



Microsoft: top data center Chicago, 250 MW
- 2020/21: another > $1 billion in servers
- 2020 added another 225MW of servers
- More than 1 Mio servers world-wide, 300.000 for XBox One

2025-Nov: Schwarz-Gruppe, Lübbenau 200MW

btw: how much power can you get from an average nuclear power plant ... ?!?
→ 2023, first DC colocated with nuclear power station *Cumulus Data* (cumulusinfra.com)

Dec'22 Market Research Study:
https://omdia.tech.informa.com/OM027611/Nuclear-powered-data-centers-are-on-the-horizon

Public clouds - Google

- Very few things known
- >> 1 million servers worldwide
- Shipping Containers houses up to 1160 servers, consuming 250kW
- 40 to 60 containers per data center
- Data center location chosen by power availability and pricing
- Latest trend: fuel cells at rack level

Public clouds - Google

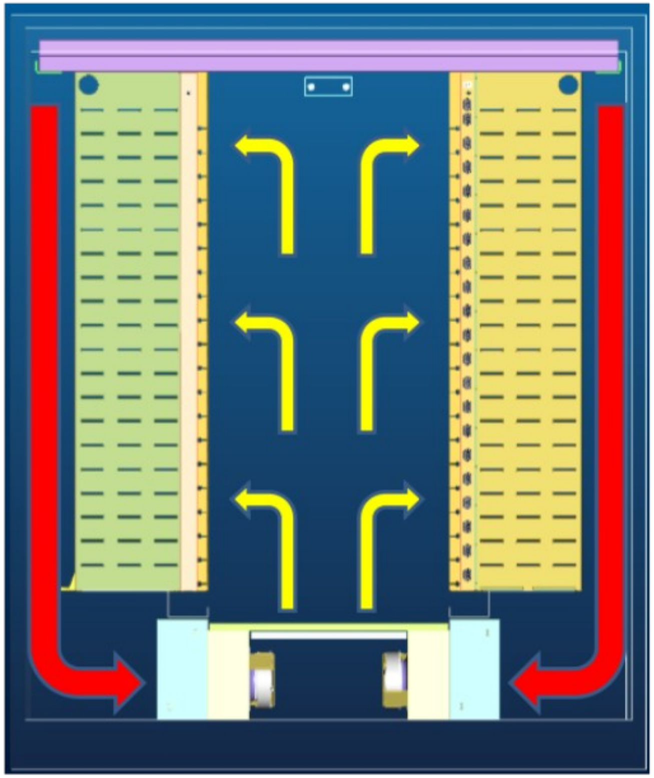
- http://www.youtube.com/watch?v=zRwPSFpLX8I

Google Data Center Best Practices Videos

(1) Intro & Measuring PUE	http://youtu.be/O96PwWkJdUo
(2) Manage Airflow	http://youtu.be/3OVqjJjNtAE
(3) Adjust Thermostat	http://youtu.be/yzJfNP5sAao
(4) Utilize Free Cooling	http://youtu.be/_uVeDwleAEo
(5) Optimize Power Distribution	http://youtu.be/dk95NV01dAY

Public clouds – Google

Container cross section

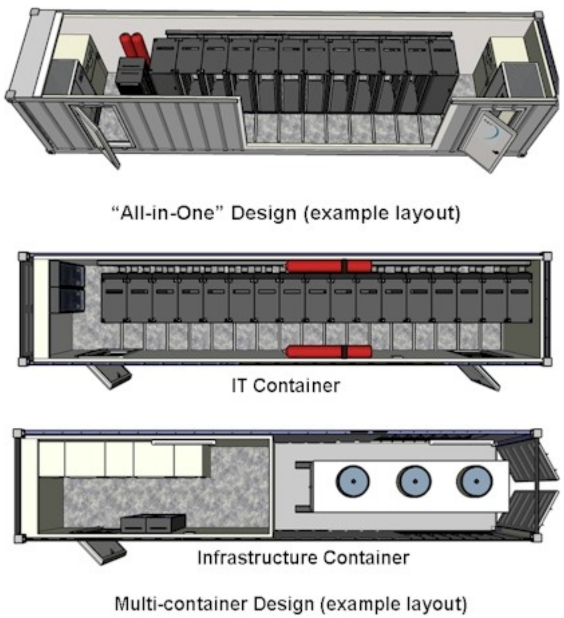


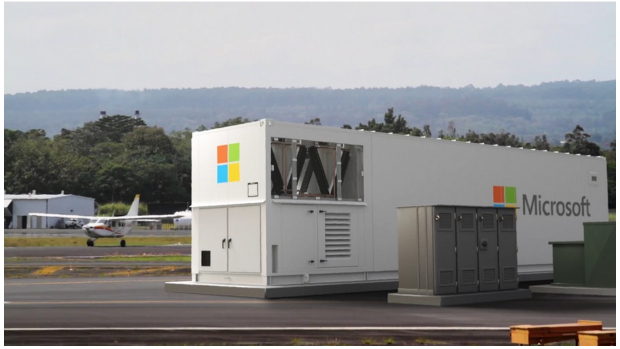
Server Design

-> Hot / cold aisle containment
-> Controlled air flow path

Containers

- Containers are hip
  - Flexible
  - Rent a data center
  - Disaster area
  - Shipping ready
  - Support for liquid-cooling PUE down to 1.14
  - MS Azure Modular Data Center





HPE Modular Data Center (MDC)

IBM Portable Modular Data Center

Social media – Facebook

Data as of 2014/2018
1 PB = 1.024 TB

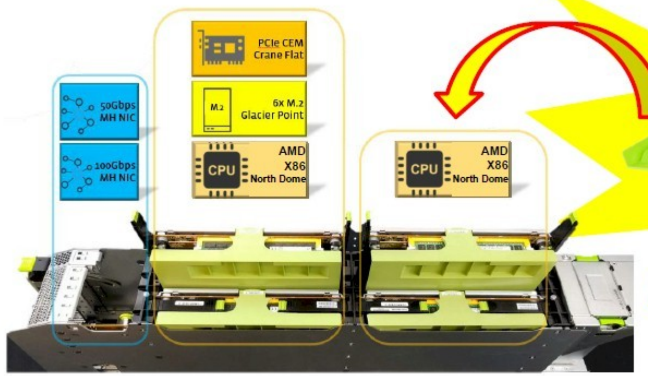
- 4 US data centers, 1 in Sweden (plus several colocations)
- Every day...
  - >>500 million photo uploads
  - 4.75 billion new content items (wall posts, status updates, ...)
  - 14.5 billion new “Like” clicks
- Photos...
  - all photos stored in 4 resolutions, every month +8 PB
  - >> 1.000.000 photos served every second
  - Special purpose file system Haystack (minimize expensive metadata operations, keep metadata in main memory, from 10 down to 1 I/O operation per photo serving)
- Estimated 180.000 servers, 200-300 PB of disk space

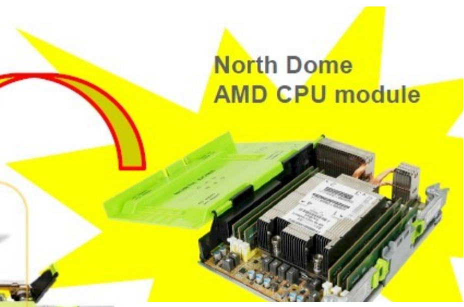
Spring 2014: 282 PB
(Fotos and videos,
Data Warehouse: extra 300 PB)

- New data every day...
  - More than 150 TB of log data
  - More than 4000 TB of user data
- Several thousand memcached servers (database caching of read requests)
  - tuned to support 200.000 UDP requests per second (50.000 with vanilla code)
  - more than 2 trillion objects in cache

Social media – Facebook

- Open Compute Project (OCP) founder and main contributor
- Latest AMD Zen3 based server:



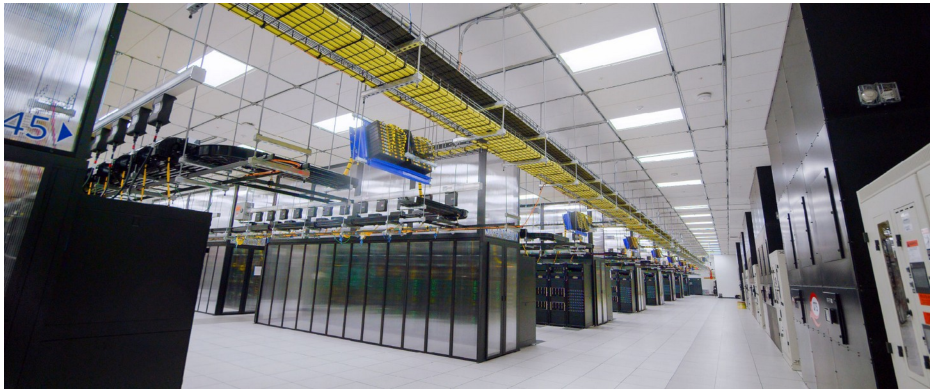


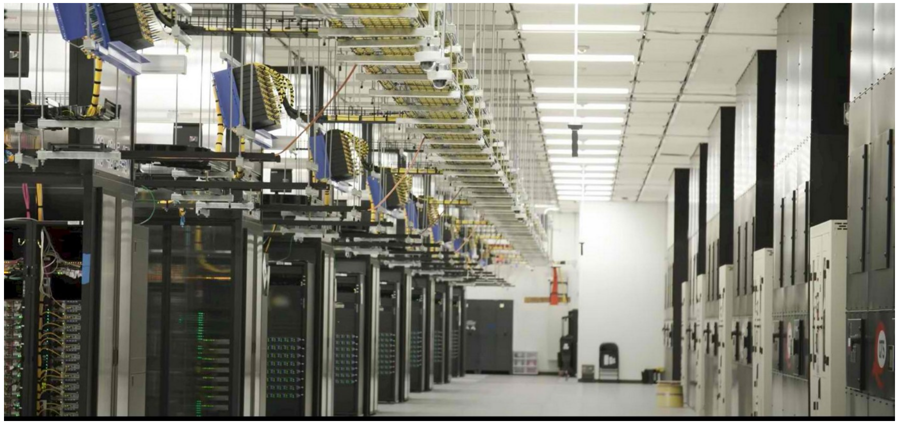
- How AI recommendation software dictated server design at Meta:
  https://arxiv.org/pdf/2104.05158.pdf

DHBW Duale Hochschule Baden-Württemberg Stuttgart

Social media – Facebook AI Research SuperCluster (2022)

- 760x Nvidia DGX A100, each
  - 2x 64-core AMD Epyc
  - 1TB main memory
  - 8x 200Gbit/s Infiniband NICs
  - 8x A100 Nvidia GPU
- ~6000 GPU (to be expanded 16.000 GPU)
- Storage
  - 46PB of cache via other x86 servers
  - 10PB object storage
  - 175PB block storage
  - To be expanded to 1EB total
  - estimated: 3TB/s bandwidth, goal 16 TB/s
- Cost: estimated $600-800 million
- Use case: Metaverse, NLP, Llama, Llama2

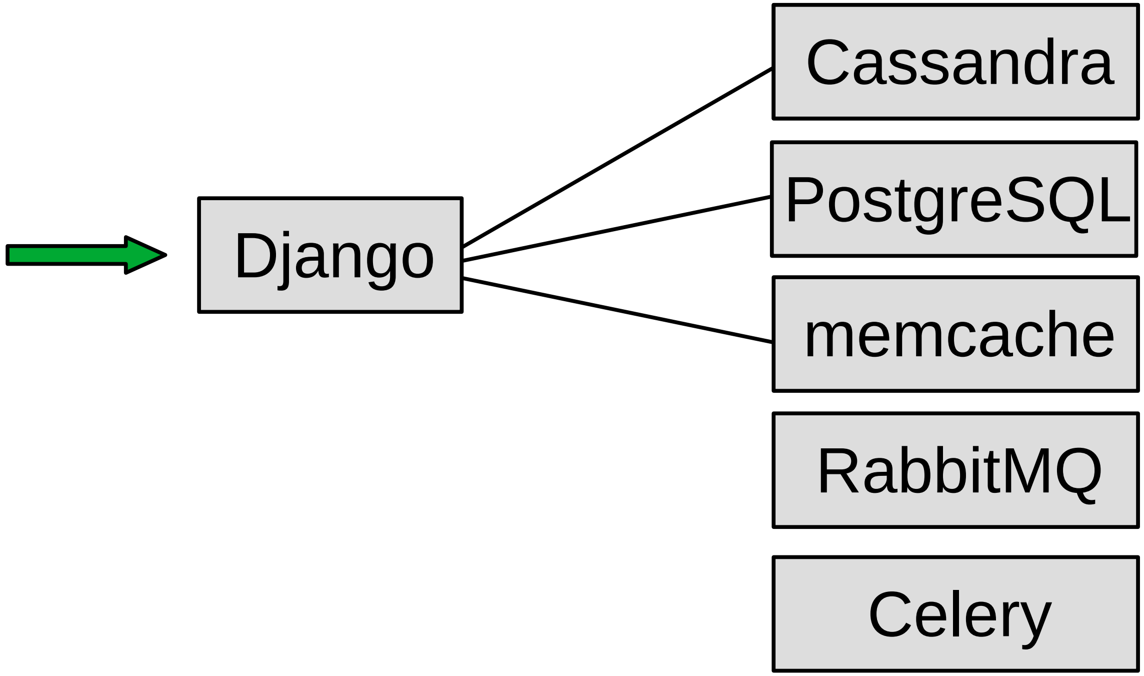




Further information: https://ai.facebook.com/blog/ai-rsc/

Latest news:
https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

Social media – Instagram stack



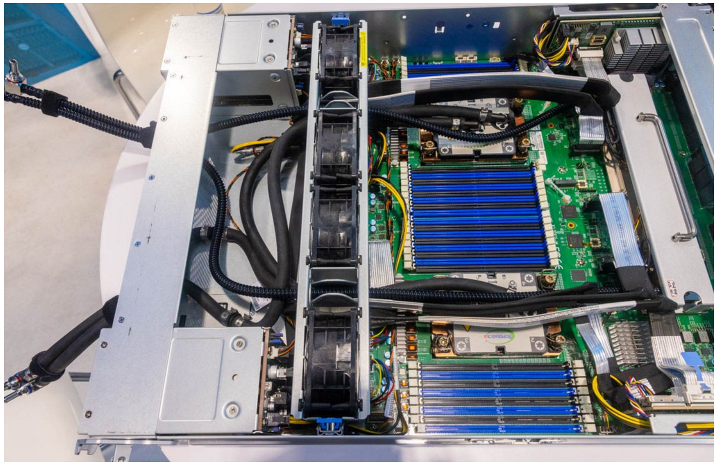
But: how to scale-out reliably and successfully this stack?

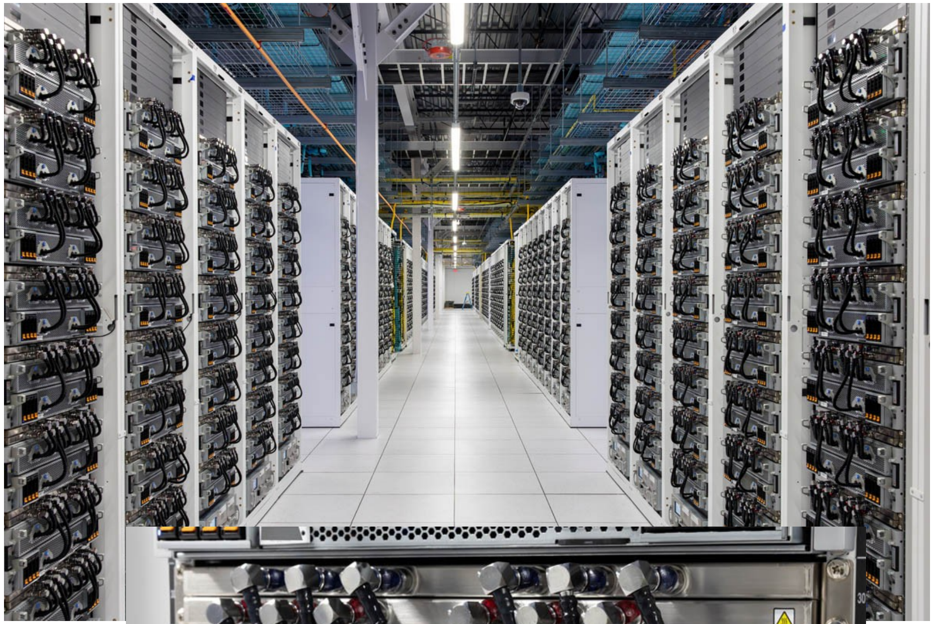
Source: Qcon 2017, presentation online https://www.youtube.com/watch?v=hnpzNAPiC0E

Social media – X / Twitter AI cluster

Colossus, built in 2024

- *100.000 H100 GPUs*
- *Built in 5 months*
- *One rack is 8x GPU server*
- *Eight racks make up a mini-cluster (with 512 GPUs)*
- *9x 400 Gb Ethernet (= 3.6 Tb/s per server)*
- *2025 expansion planned*





Source: https://www.servethehome.com/inside-100000-nvidia-gpu-xai-colossus-cluster-supermicro-helped-build-for-elon-musk/