Can generative AI run on thin air?

Generative AI impact on data centers

Mattias Vesterlund
RISE – Research Institutes of Sweden

RISE ICE Data Center

A full-scale research datacenter and test environment with the objective to increase knowledge and strengthen the global AI & Data Center ecosystems

2000 physical servers
250 kW
10 petabyte storage
240 GPUs
12,5 petaflops
HDFS clusters
OpenStack ECC
Kubernetes cluster
OCP servers
Edge modules

- 30+ projects, from the ground to the cloud
- 20+ employees
- 4+ MEUR turnover
- Established 2016
- 100+ project partners

Partners: Ericsson, ABB, Vattenfall, Meta, LTU, Region North, Vertiv, BP Castrol, Intel

2018 - 2022
BDV
GOLD

2023
BDV
i-Space PLATINUM

016, 201
BDV
i-Space SILVER

DATACLOUD
GLOBAL AWARDS
2019

BRON
C
Management
Excellence
STRIVING FOR CLUSTER EXCELLENCE

2017, 2019

Best Data Center Initiative of the Year
with RISE SICE North Research Data Center

EARTO Innovation Awards 2022

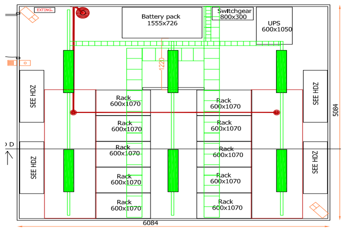
Holistic Cooling

RI. SE

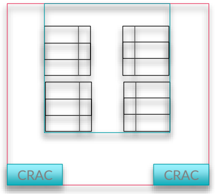
To help protect your privacy, PowerPoint has blocked automatic download of this picture.

RISE ICE Datacenter test environment

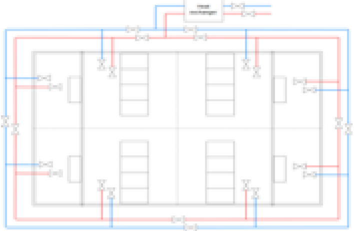
Module 1 & 2 Compute clusters

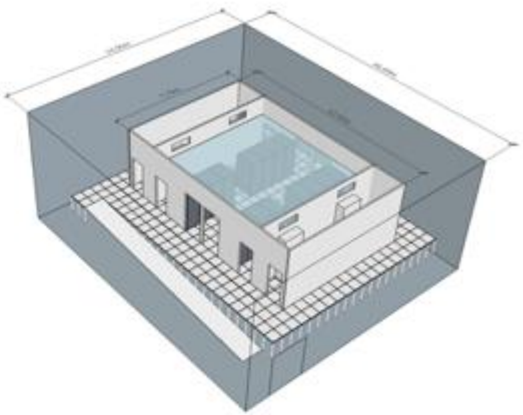


Module 3 OCP Lab + Climate box + Heat box

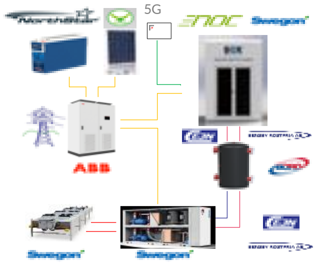


Module 4 Facility hardware test lab





Edge + wind tunnels + liquid cooling test bed



RISE ICE Datacenter public funded projects (examples)

BTDC

An EU project. The goal was the Most Energy and Cost efficient data center in the world. Holistic cooling was used to reach PUE = 1,015.

ArctiqDC

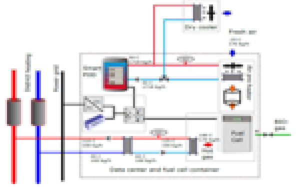
An EU Interreg Nord project . The goal was to study challenges with data centers in the arctic region. Show solutions for heat reuse & operations.

AutoDC

ITEA 3

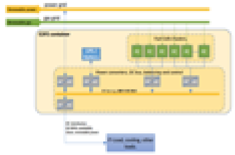
An ITEA3 project. The goal was lower OPEX by making the data center module autonomous, self-healing, self-optimizing and robust using AI.

Wedistrict



An EU project. The goal is to show a recovery process using liquid cooling and a fuel cell and use the electricity from the fuel cell for the datacenter

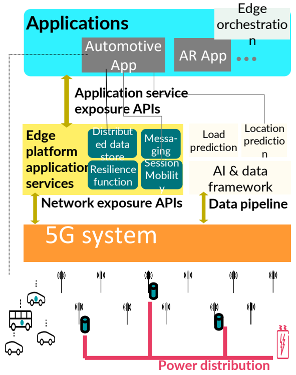
E2P2



An EU project. The goal is to define the fuel cell prime power concept for EU data centres. Demonstrate and validate a proof of concept.

Aniara

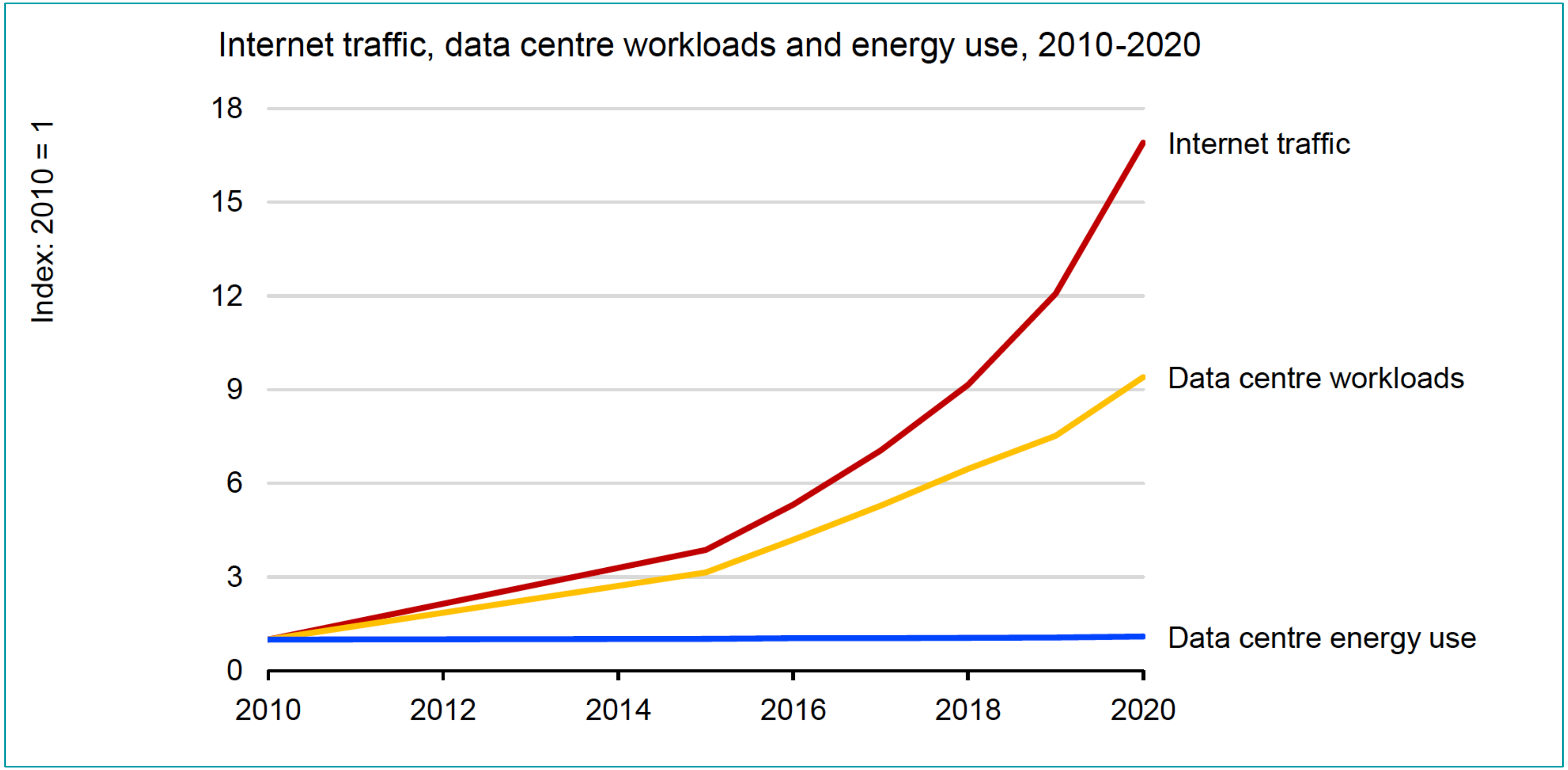
CELTIC-NEXT



A Celtic+ project. The goal is to develop automation of a network of edge nodes availability, robustness, optimization and reliability.

Use of GPT-4 running on XPUs

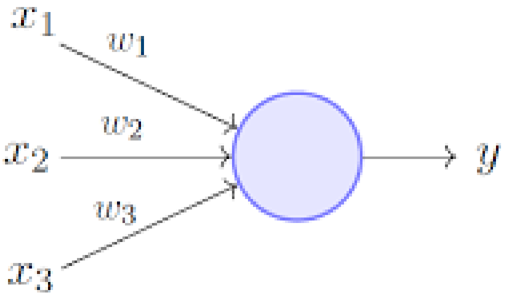
Global data center energy use trends



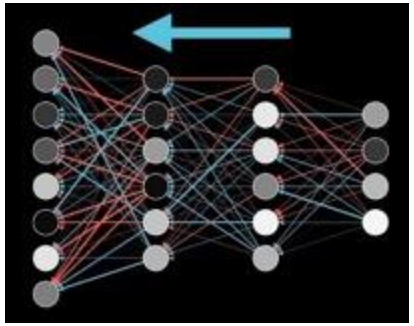
Data center energy use almost flat compared to internet traffic or data center workloads

Source: Global data center energy use trends (IEA, 2021)

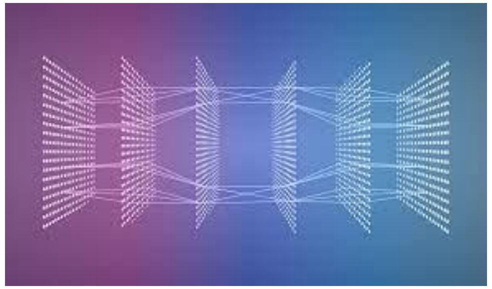
History of AI by @Torshammer



Preceptron 1950-60
(one layer, few parameters)



Backpropagation 1980-1990
(many layers, many parameters)



Deeplearning 2010-20
(many layers, billion parameters)

Simple AI - Many humans
Few control parametrar
Laptop computer

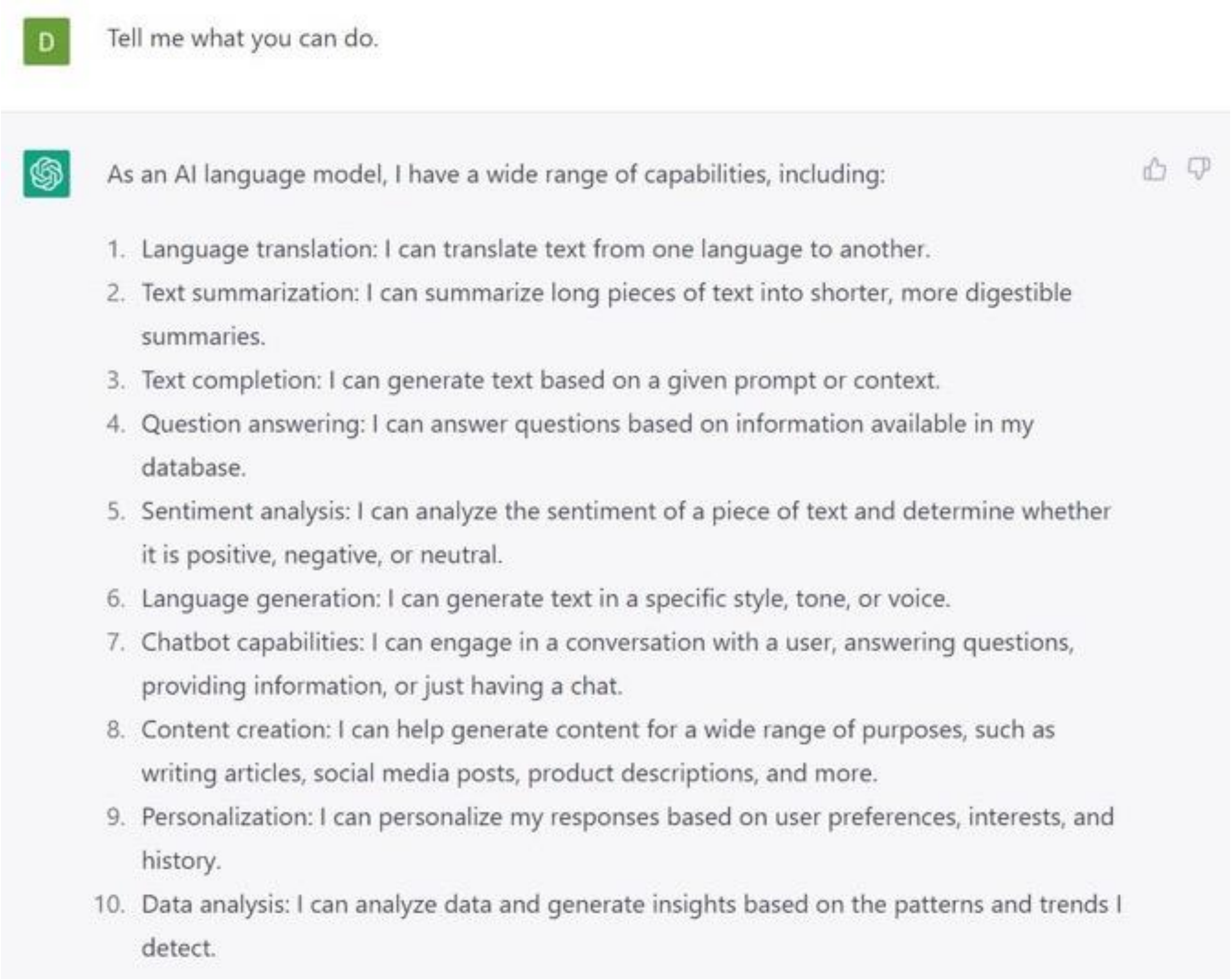
Machine learning - Few experts
Many control parameters
Advanced GPU server

Generative AI – No single human (not all tasks)
Billion control parameters
Super computer cluster

30 years

30 years

Generative AI (for ex. ChatGPT)

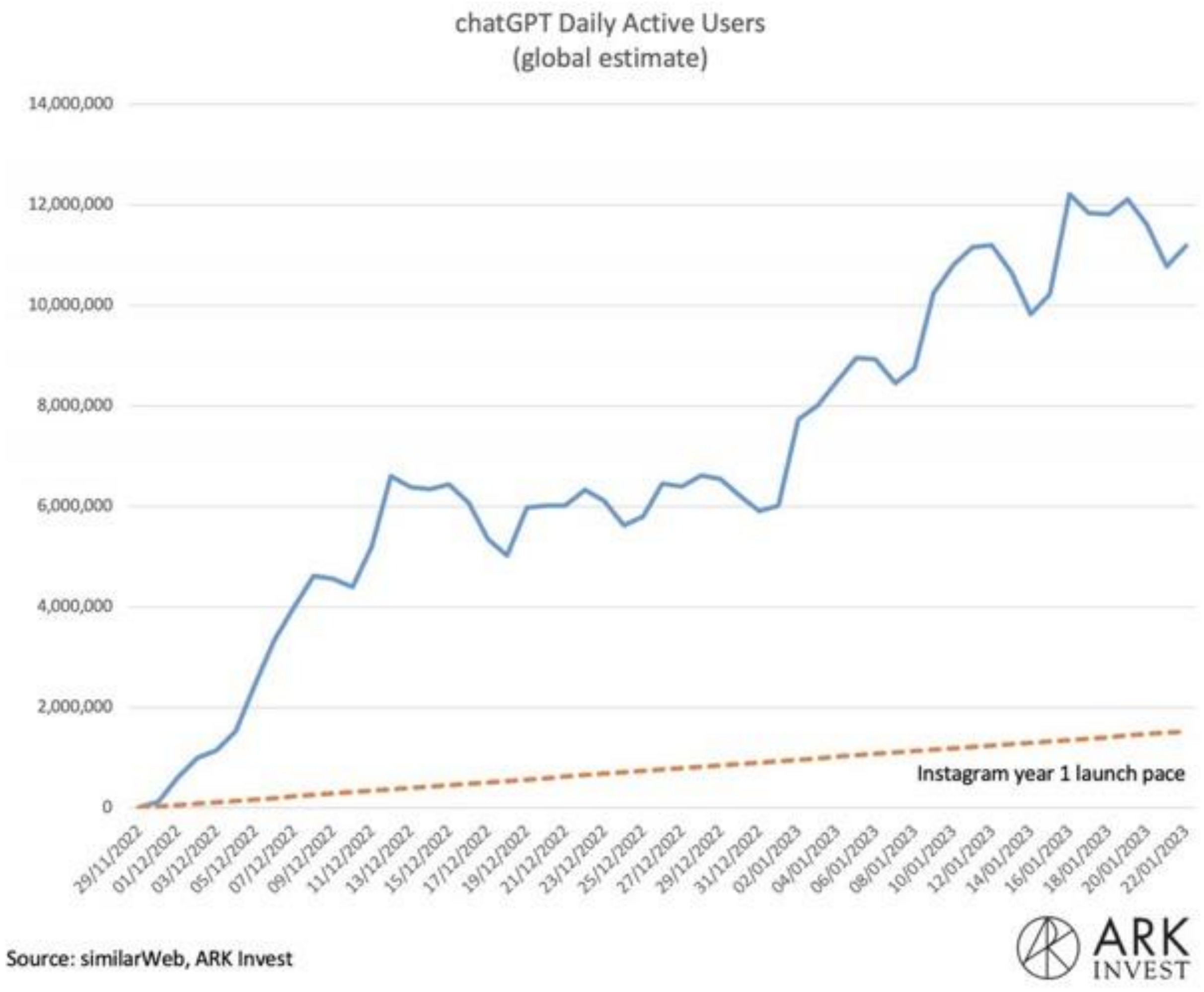


Tell me what you can do.

As an AI language model, I have a wide range of capabilities, including:

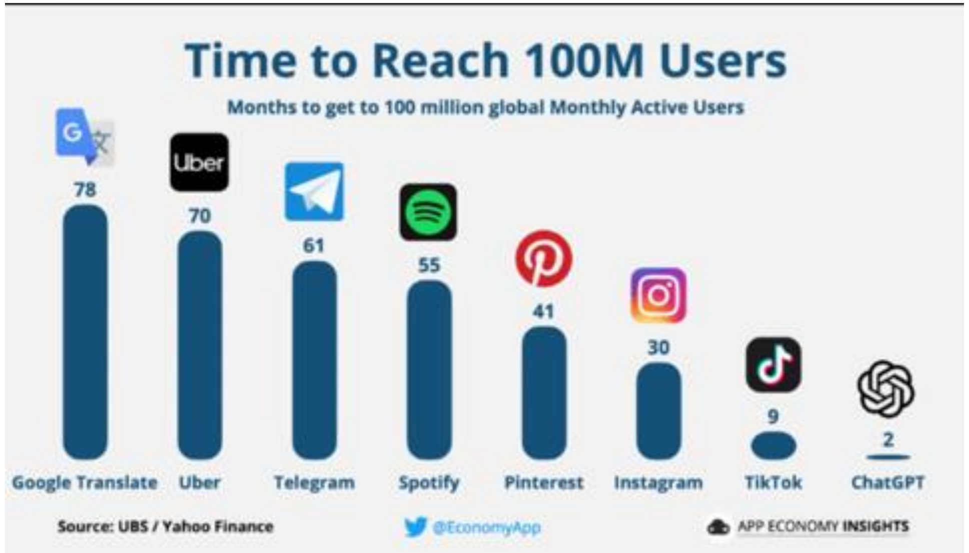
1. Language translation: I can translate text from one language to another.
2. Text summarization: I can summarize long pieces of text into shorter, more digestible summaries.
3. Text completion: I can generate text based on a given prompt or context.
4. Question answering: I can answer questions based on information available in my database.
5. Sentiment analysis: I can analyze the sentiment of a piece of text and determine whether it is positive, negative, or neutral.
6. Language generation: I can generate text in a specific style, tone, or voice.
7. Chatbot capabilities: I can engage in a conversation with a user, answering questions, providing information, or just having a chat.
8. Content creation: I can help generate content for a wide range of purposes, such as writing articles, social media posts, product descriptions, and more.
9. Personalization: I can personalize my responses based on user preferences, interests, and history.
10. Data analysis: I can analyze data and generate insights based on the patterns and trends I detect.

Growth rate ChatGPT



chatGPT Daily Active Users
(global estimate)

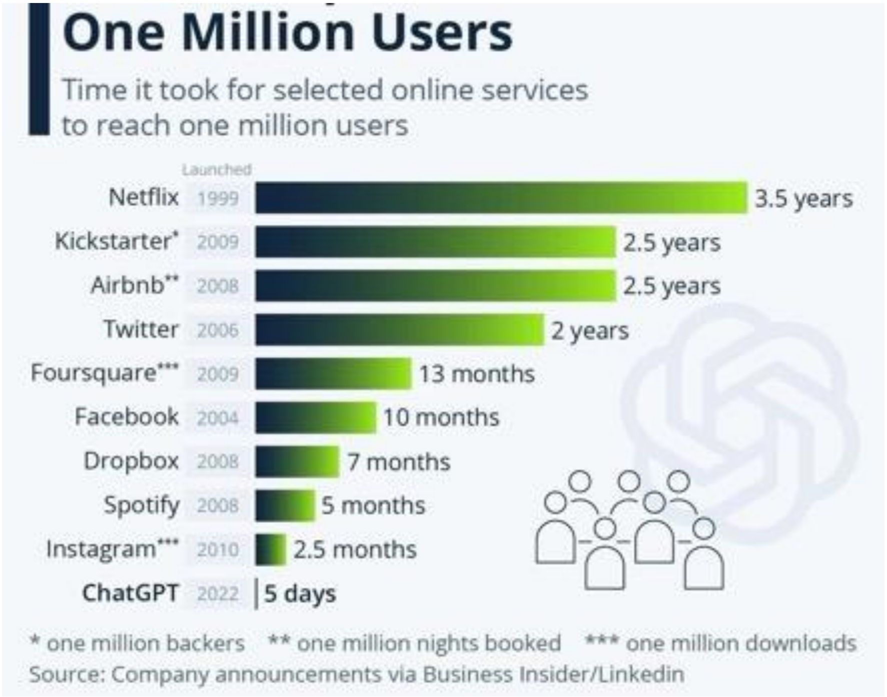
Source: similarWeb, ARK Invest



Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users

Source: UBS / Yahoo Finance



One Million Users

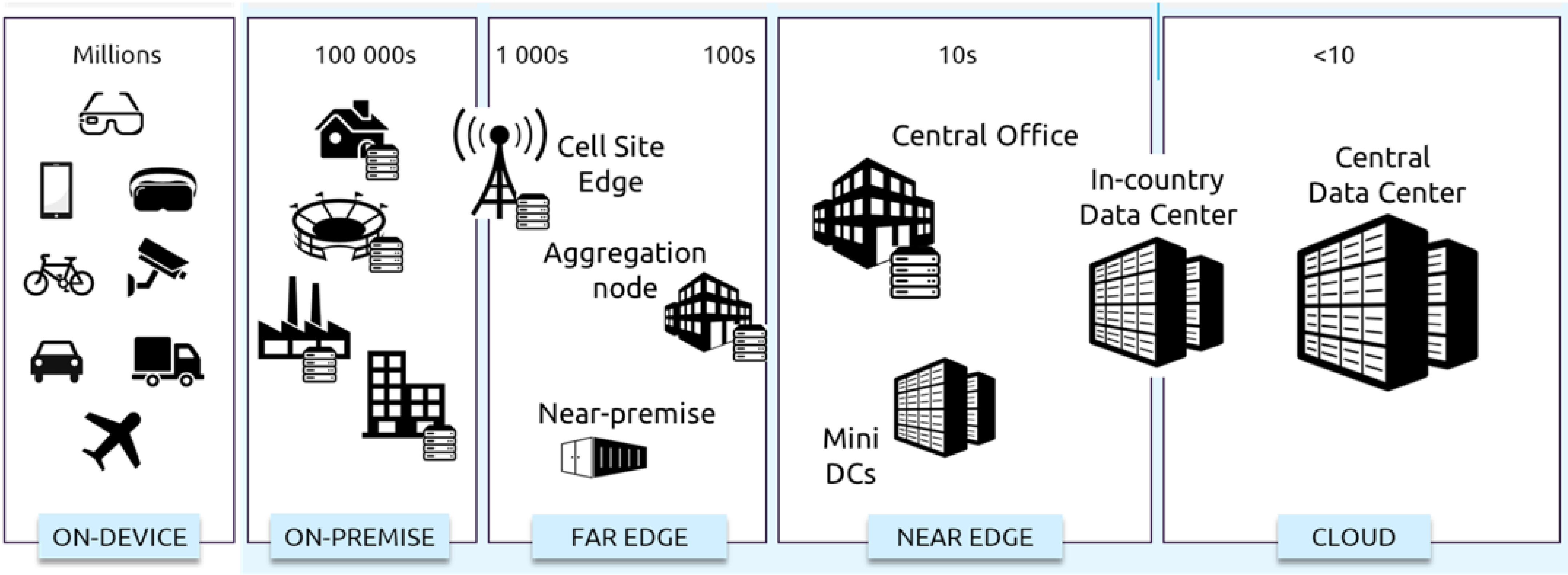
Time it took for selected online services to reach one million users

* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/Linkedin

High-end efficient and heavy AI compute

Where will it happen?



Inference (use phase)

Model learning (training phase)

Training GPT-4

1,7 trillion model parameters & 13T tokens for training and 6 Floating operations (FLOPs) per parameter for one token

⇒ 133G petaFLOPs

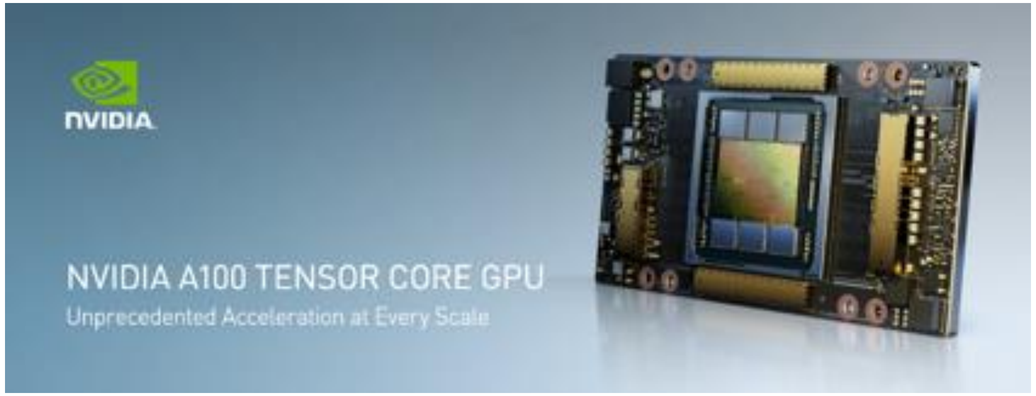
peta =10e15 = 1 000 000 000 000 000

OpenAI trained on 25 000 NVIDIA A100 GPUs

⇒ Training time >100 days

⇒ Cost $100 million

⇒ Energy 50 GWh (2% of Swedish DC energy)


NVIDIA
NVIDIA A100 TENSOR CORE GPU
Unprecedented Acceleration at Every Scale

Source: OpenAI, RISE, Wikipedia, the decoder, ChatGPT Statistics 2023

Using (inference) GPT-4

1,7 trillion model parameters & max 32k tokens and 2 Floating operations (FLOPs) per parameter for one token (assume *1k* tokens/request)

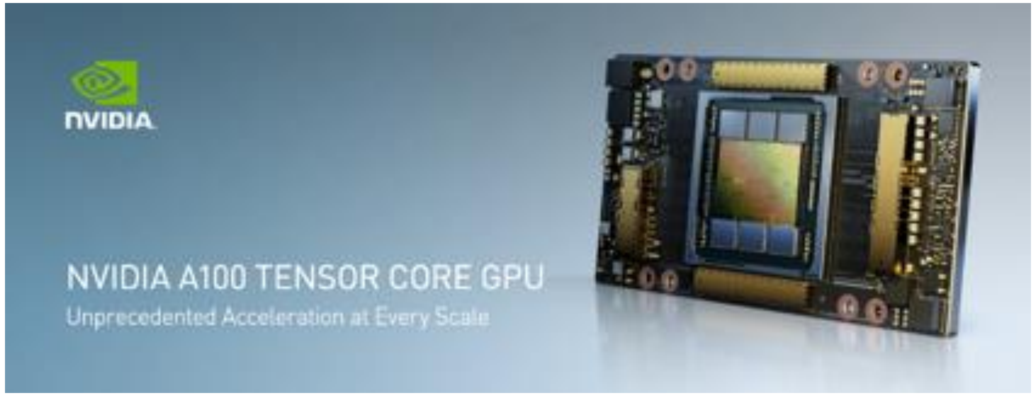
⇒ 3,4 petaFLOPs per request -> 1,2 Wh
GPT-3 – 0,13 Wh, Google search – 0,28 Wh

1 Wh = 60 W light bulb on for 1 minute.

OpenAI use 128 NVIDIA A100 GPUs for inference
100 million active users and *2 request/day*

⇒ 0,68G petaFLOPs per day -> 91 GWh/year (10MW)

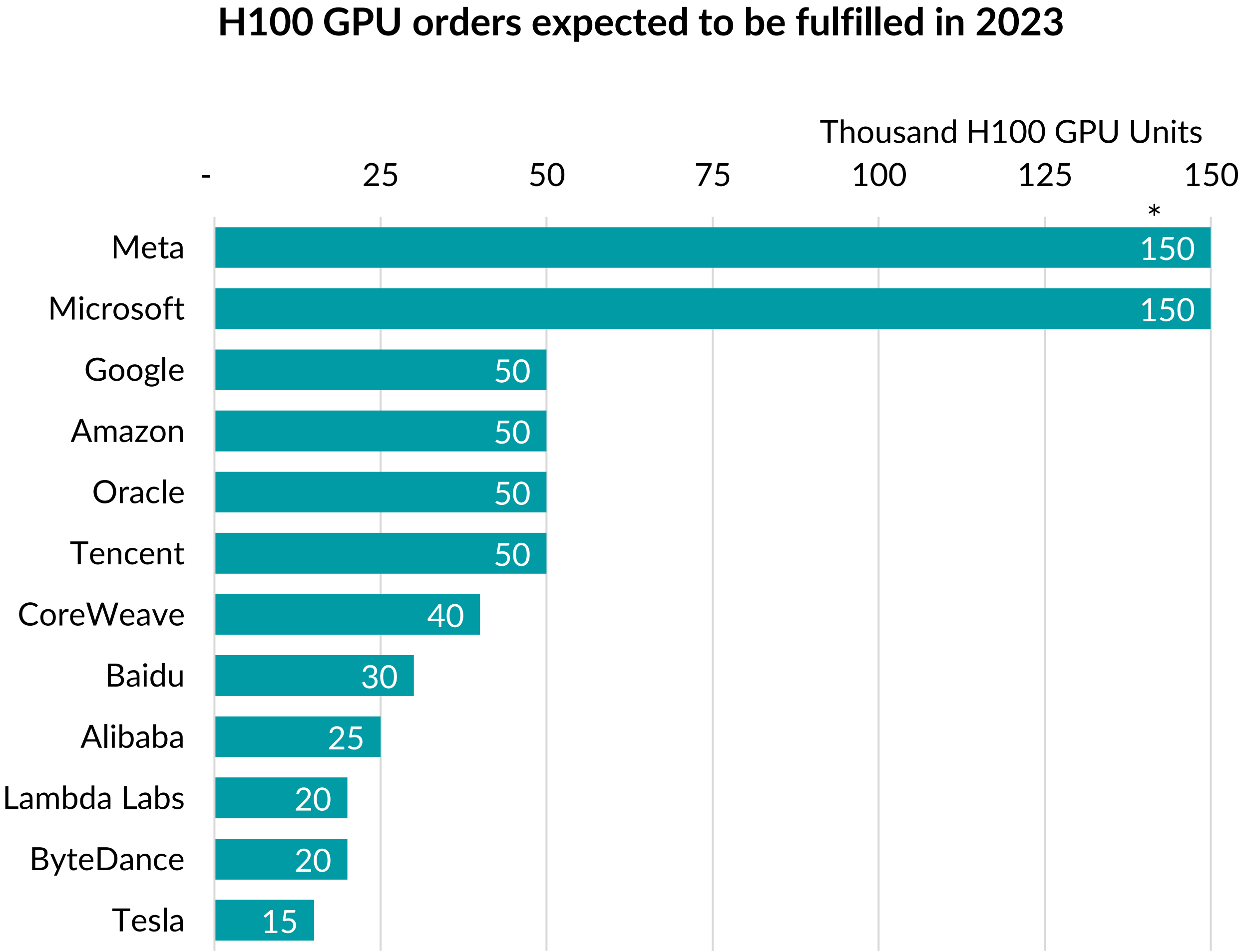
⇒ 0,1 seconds and 200 clusters for GPT-4


NVIDIA
NVIDIA A100 TENSOR CORE GPU
Unprecedented Acceleration at Every Scale

GPT-3 - 0,07G petaFLOPs/day
and 9,5 GWh/year (1 MW)

Source: OpenAI, RISE, Wikipedia, the decoder, ChatGPT Statistics 2023

AI capacity buildout continues



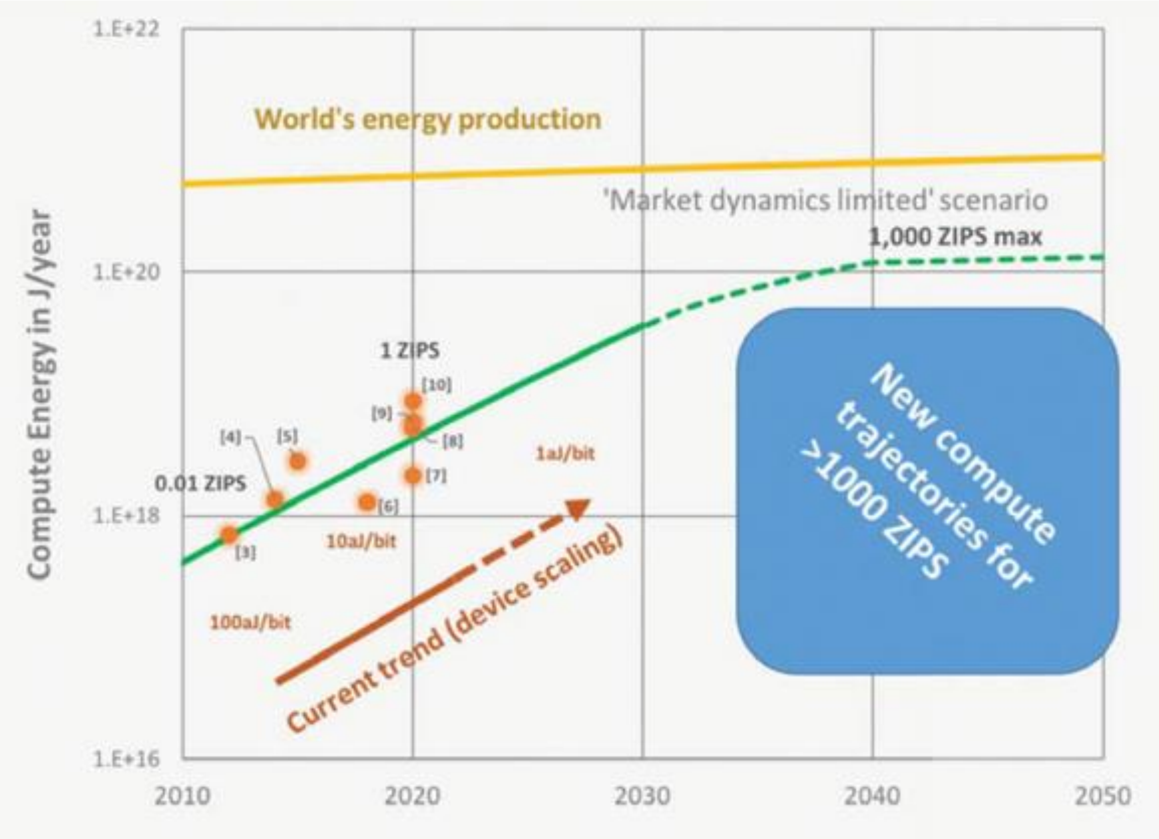
*Some Meta order fulfilment might slip into 2024 due to a delay in data center go-live dates.

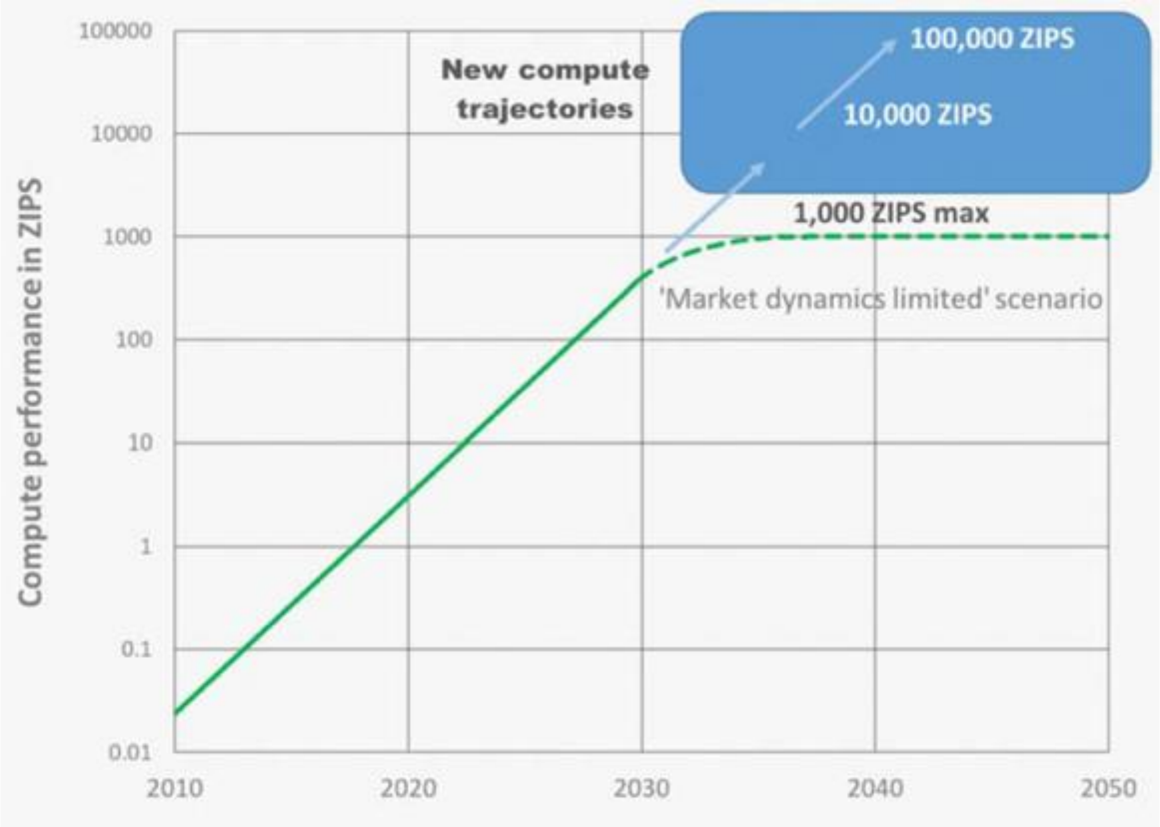
NVIDIA shipped nearly half a million H100 and A100 GPUs in 3Q23.

Source: Omdia, Cloud and Data Center Market Snapshot, Nov 2023

Thermal management in future XPUs

Information processing predictions and energy requirements

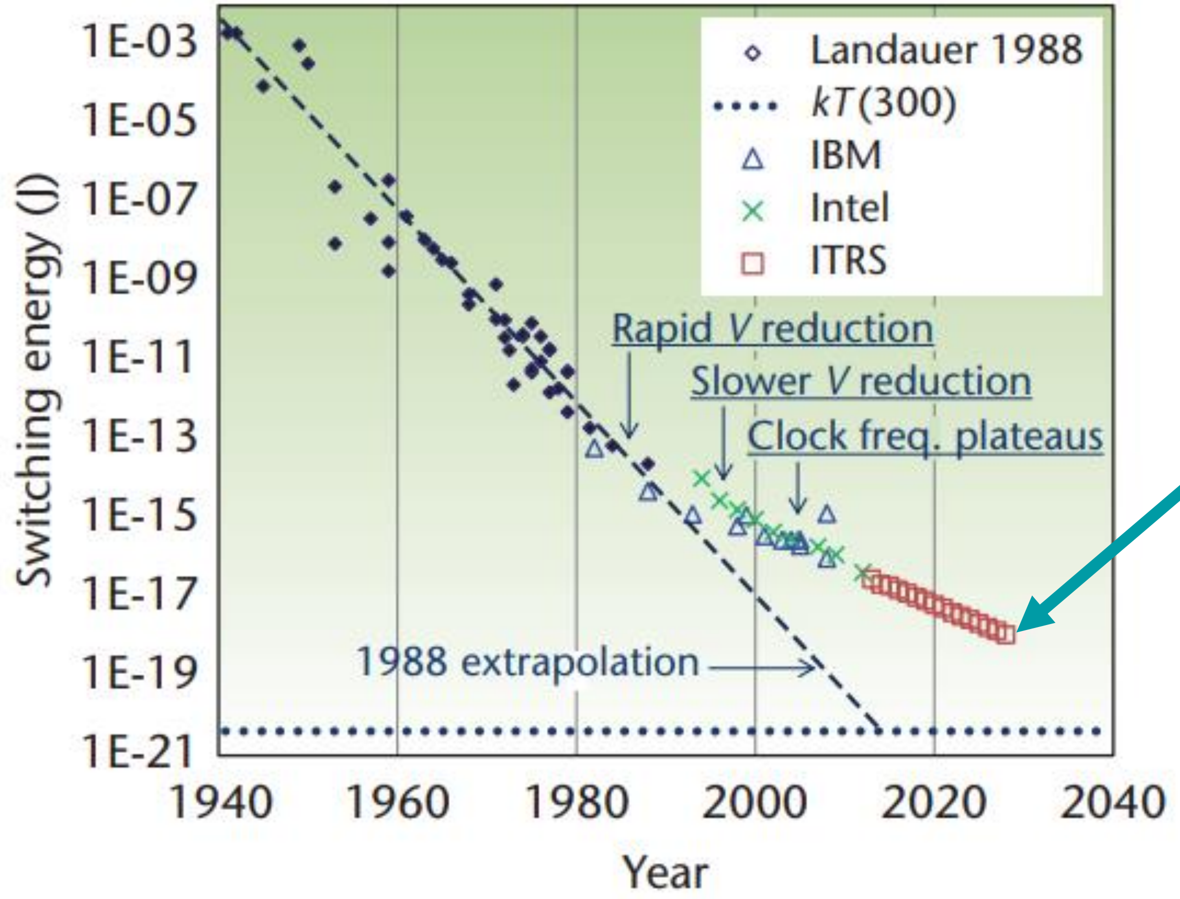




Source: Chapter 5, The Decadal Plan for Semiconductors, January 2021. https://www.src.org/about/decadal-plan/

ZIPS = Zetta Instructions Per Second
Zetta is 10^{21}

Energy efficiency in processing information.



When	Switching energy dissipation
Today	10 aJ
2030	1 aJ
Landauer limit at room temp	0.003 aJ

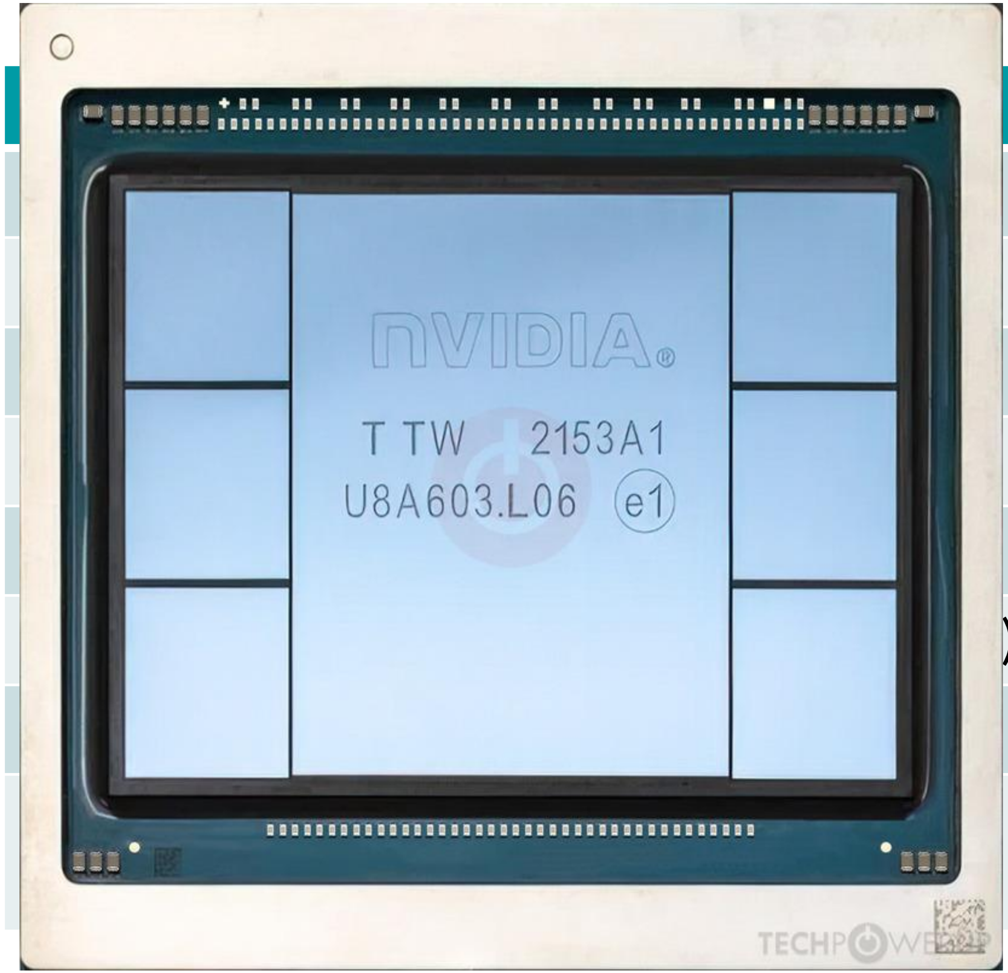
aJ is an attoJoule

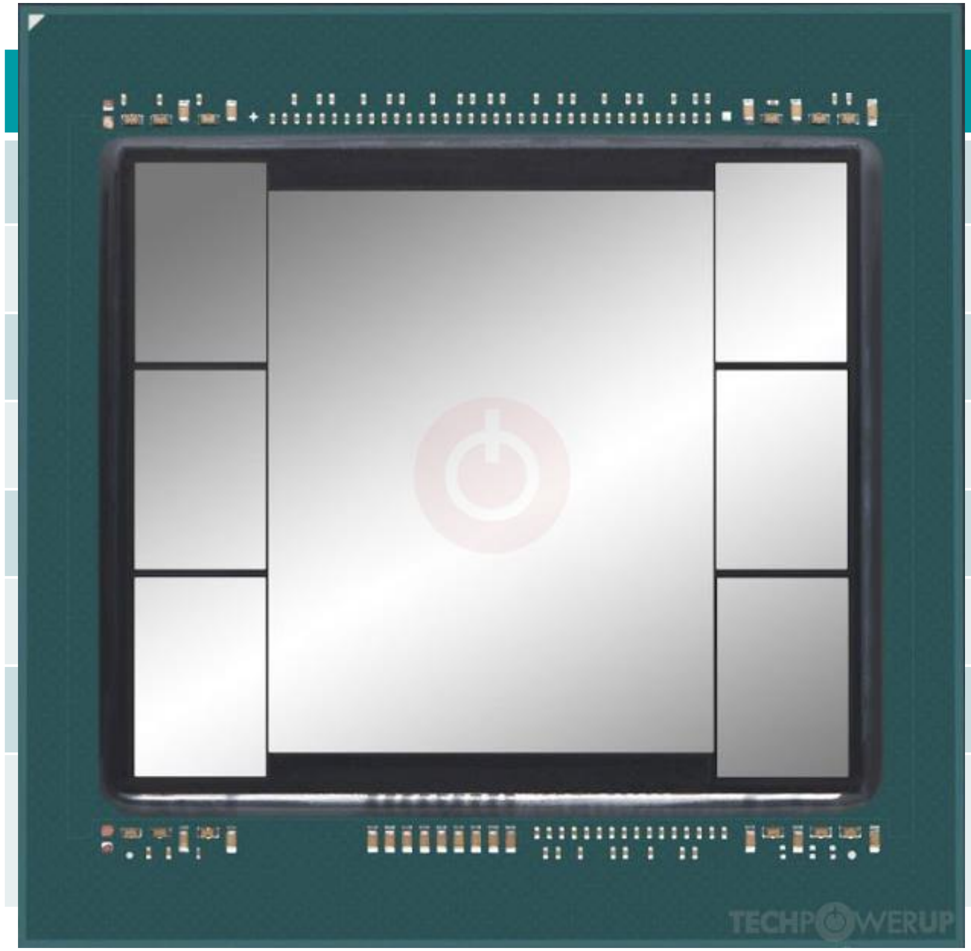
1aJ = 0.000000000000000001 J

Thomas N. Theis, H.-S. Philip Wong. "The End of Moore’s Law: A New Beginning for Information Technology," Computing in Science and Engineering, IEEE CS and AIP, March/April, 2017

Landauer, R., 1988. Dissipation and noise immunity in computation and communication. *Nature*, *335*(6193), pp.779-784.

The thermal design power (TDP)





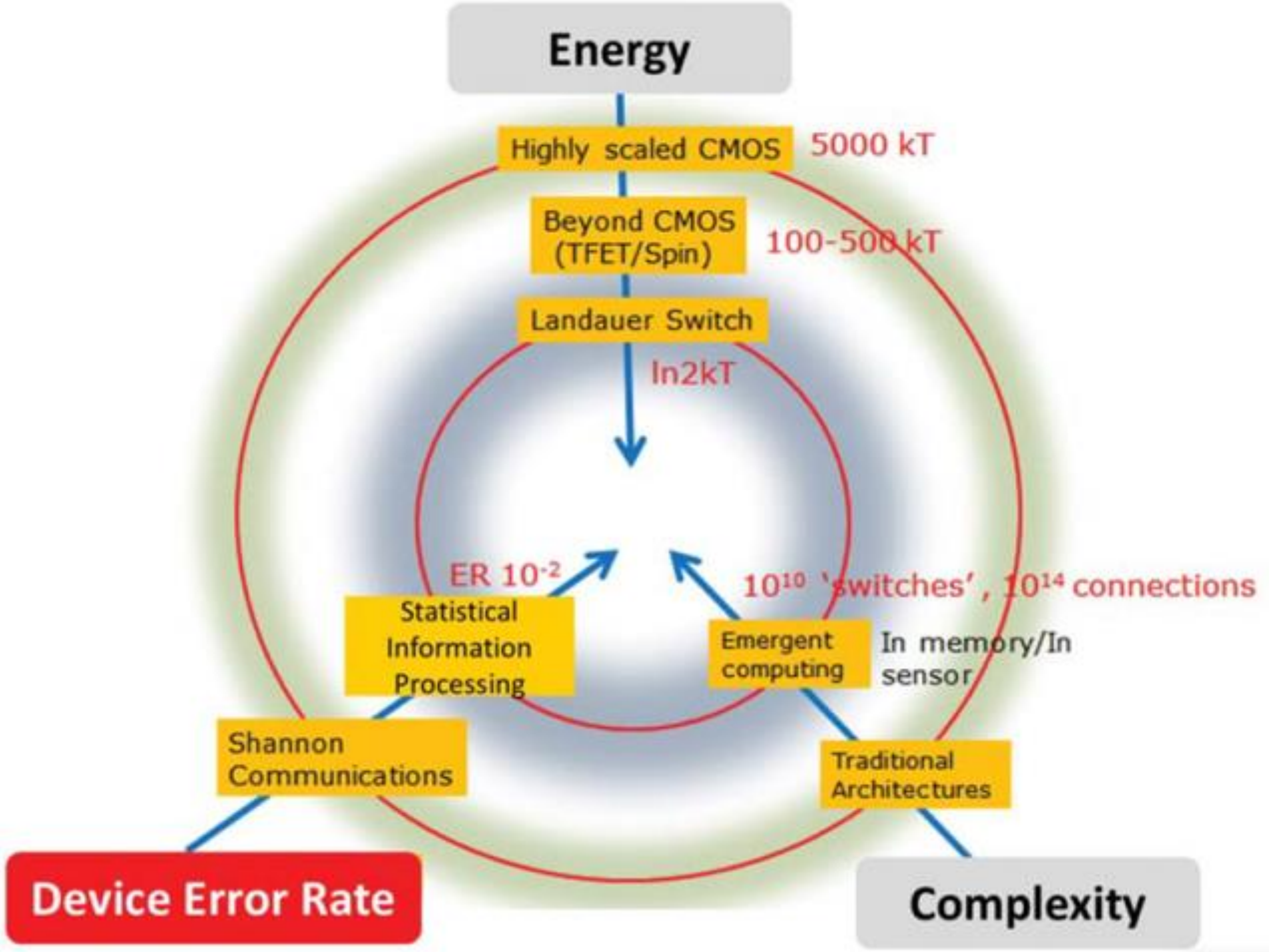
Power (W)= Switch Energy (J) x Switching Rate (s^{-1})

https://www.techpowerup.com/gpu-specs/

Where are TDPs heading?

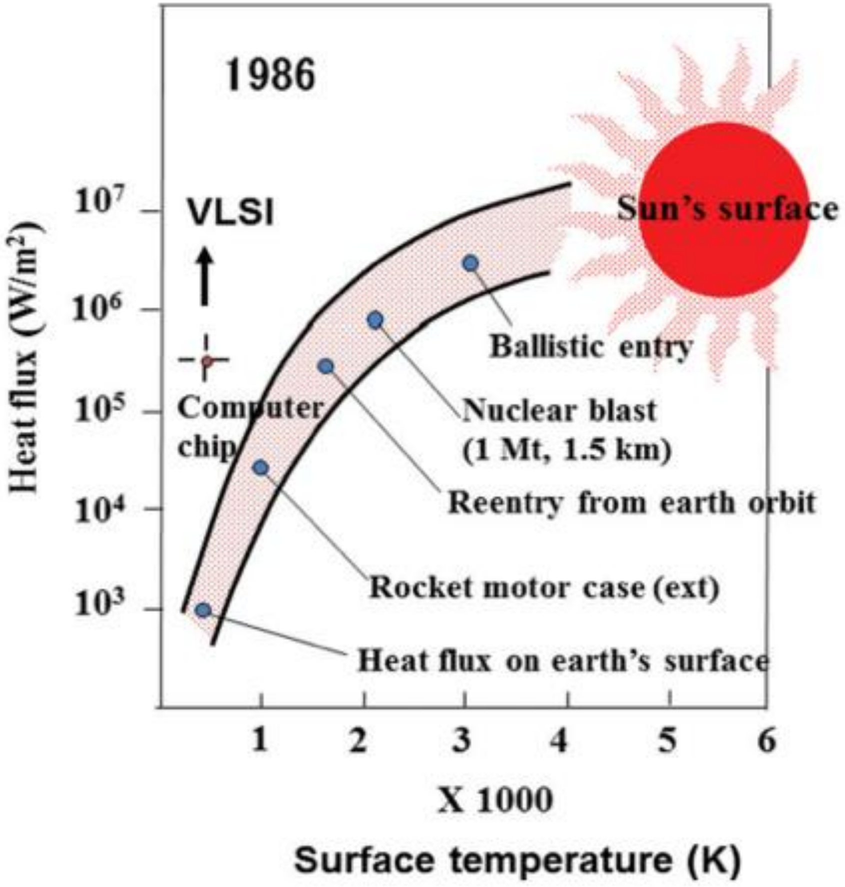
Intel microprocessor	2030
Energy dissipation (E)	1 aJ per transistor
#transistors (#T)	1 trillion
Switching frequency (f)	4 GHz (1.4 nm / A14)
Power=E x #T x f	4000 W
Dark silicon (assumption)	50%
TDP	2000 W
Die size	1250 mm2
Max heat flux HF= TDP/area	1.6 kW/m2 Tcmax < 60°C

Power (W)= Switch Energy (J) x Switching Rate (s^{-1})

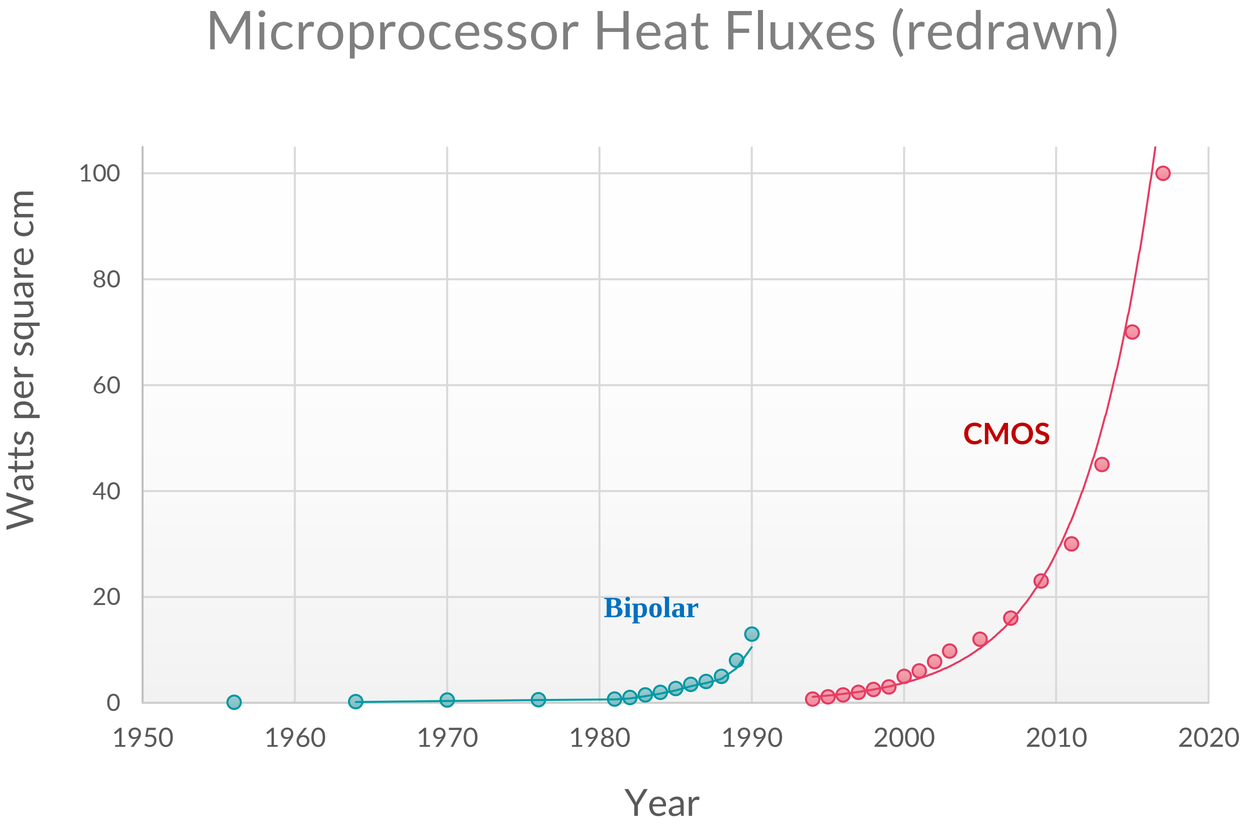


Source: Chapter 5, The Decadal Plan for Semiconductors, January 2021. https://www.src.org/about/decadal-plan/

Heat fluxes of microprocessors

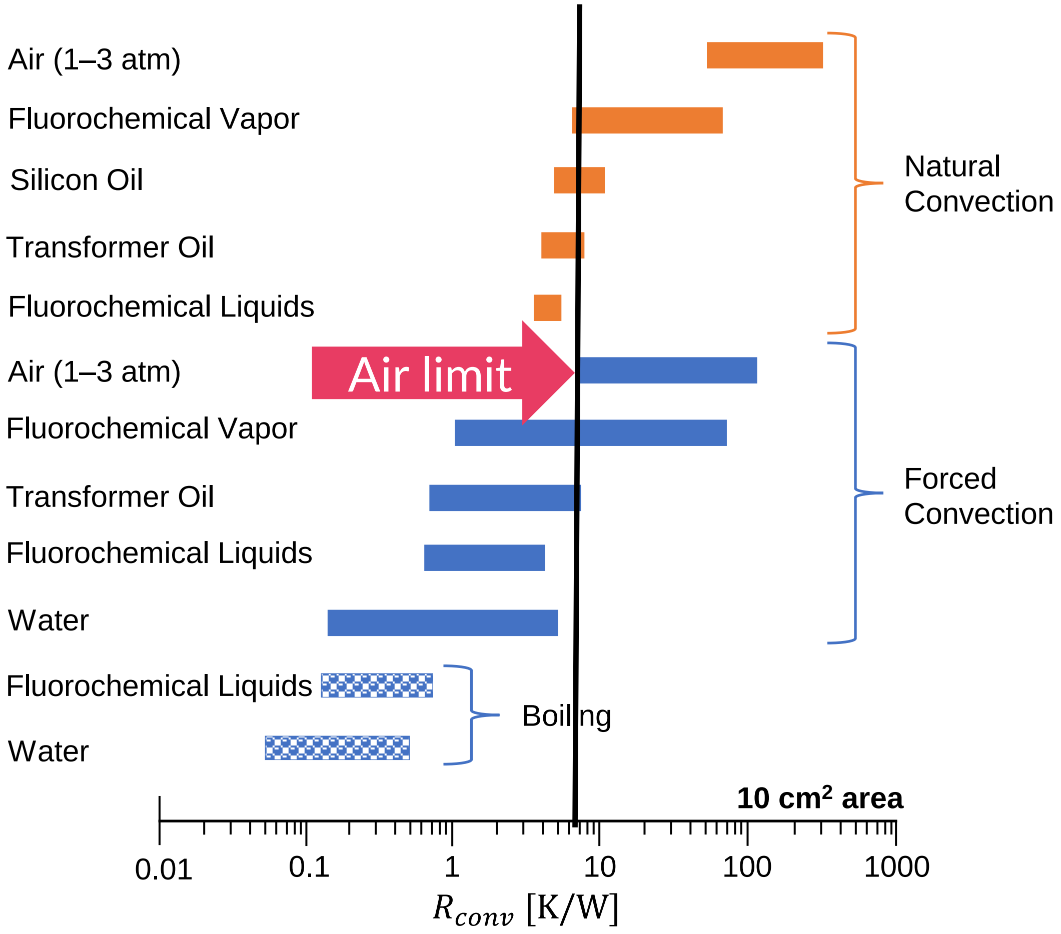


Oktay, Sevgin, Robert Hannemann, and Avram Bar-Cohen. "High heat from a small package." *Mech. Eng.;(United States)* 108, no. 3 (1986).

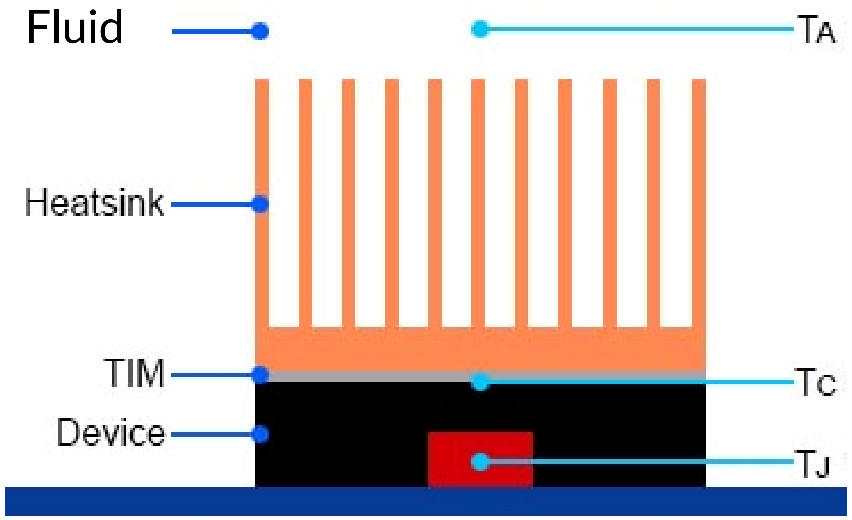


Roger Schmidt, Liquid Cooling is Back, Electronic Cooling August 2005. M.J.Ellsworth. Interpack '11 Tutorial

Fluids are needed to remove the heat, but which?



Ahmed El-Neama based on Tummala, R.R. 2001. Fundamentals of Microsystems Packaging. McGraw-Hill.



700W spread over 700cm^2 = 70 x 10cm^2

1W/cm^2 or 10W/10cm^2 from 700W/70

Temp difference between T_C and T_A is then

10W/cm2 x R_{conv}°C

$T_C - T_A = 10\ R_{conv}$ °C = 55 °C for air $R_{conv}=5.5$

$T_A = T_C - 55 = 85 - 55 = 30$ °C supply air.

Could just and does run on "thin air"!

Fluids are needed to remove the heat, but which?

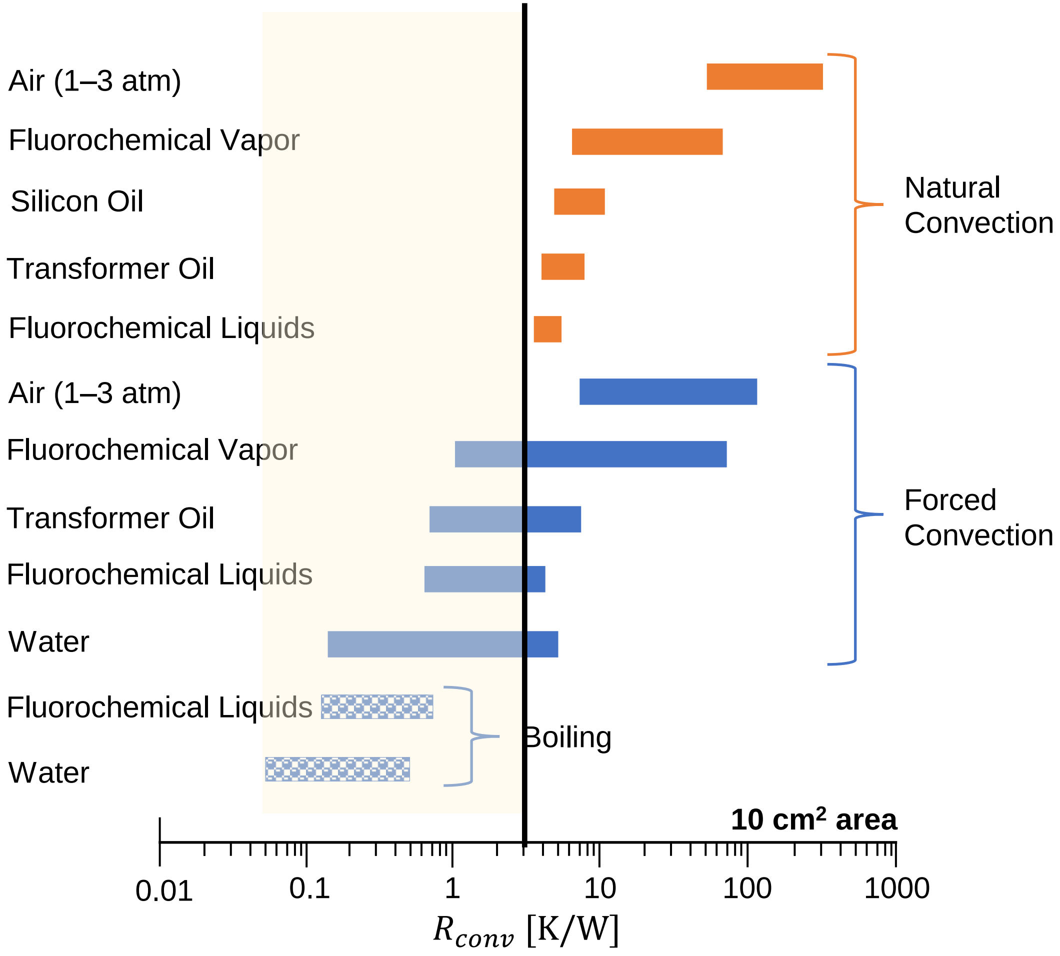
Future 2000W XPUs with T_C of 60°C and T_A at 20°C

2000W spread over $1000cm^2$ say
2000W / 100 = 20W/10cm2
Temp difference = (20 R_{conv}) °C

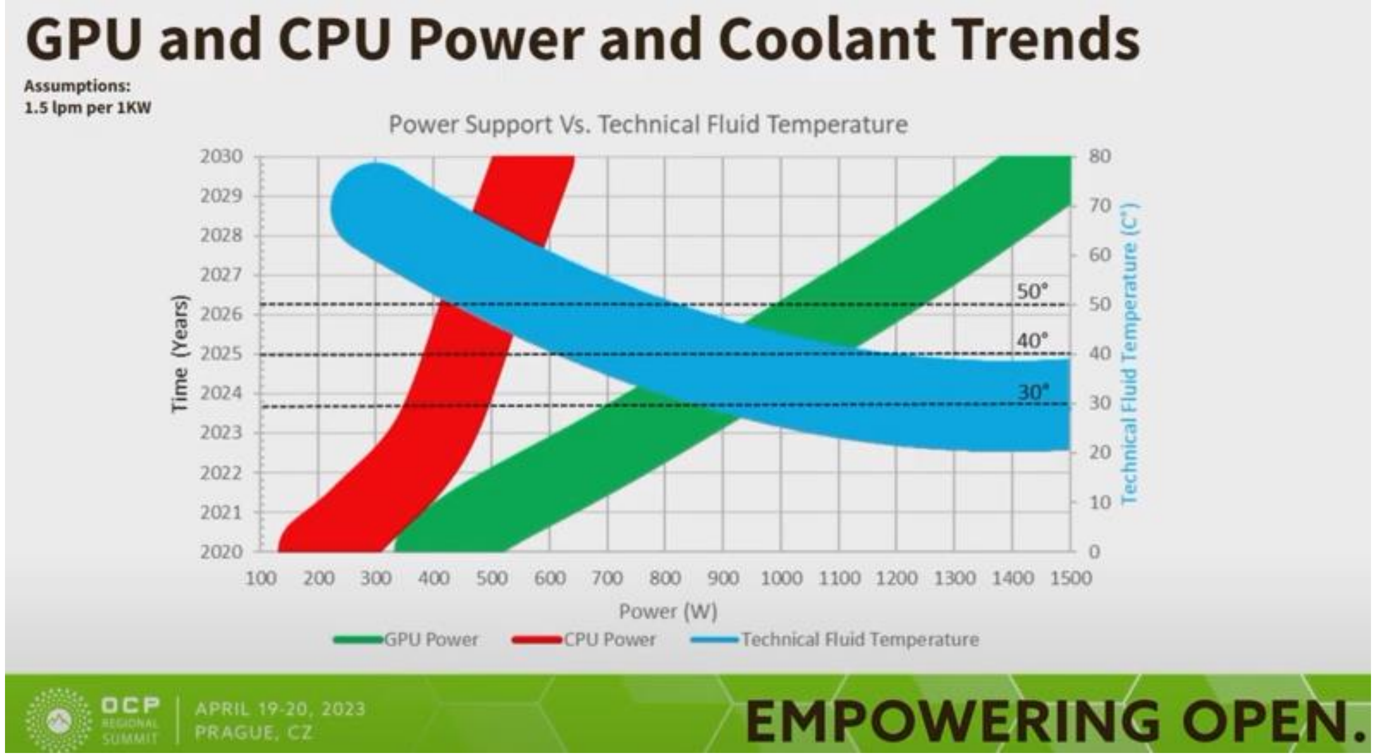
$T_A = T_C - 20\ R_{conv}$ °C

With **T_C at 60°C**
Need **R_{conv} of 3 or less.**

Cannot run on "thin air"!



OCP 2023 presentations looking at the need for liquid cooling.



GPU and CPU Power and Coolant Trends

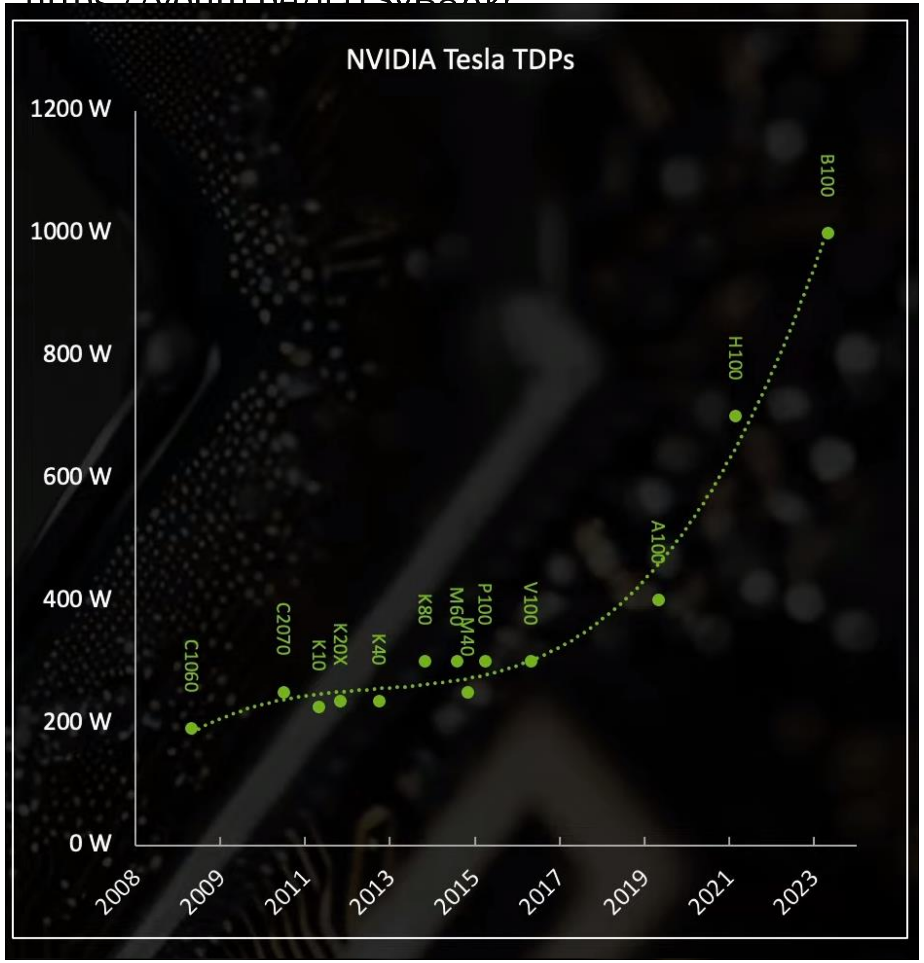
Assumptions:
1.5 lpm per 1KW

APRIL 19-20, 2023
PRAGUE, CZ

EMPOWERING OPEN.

Panel: Coolant Temperatures for Durable Data Center Designs
https://youtu.be/ToY2PBYK01Q

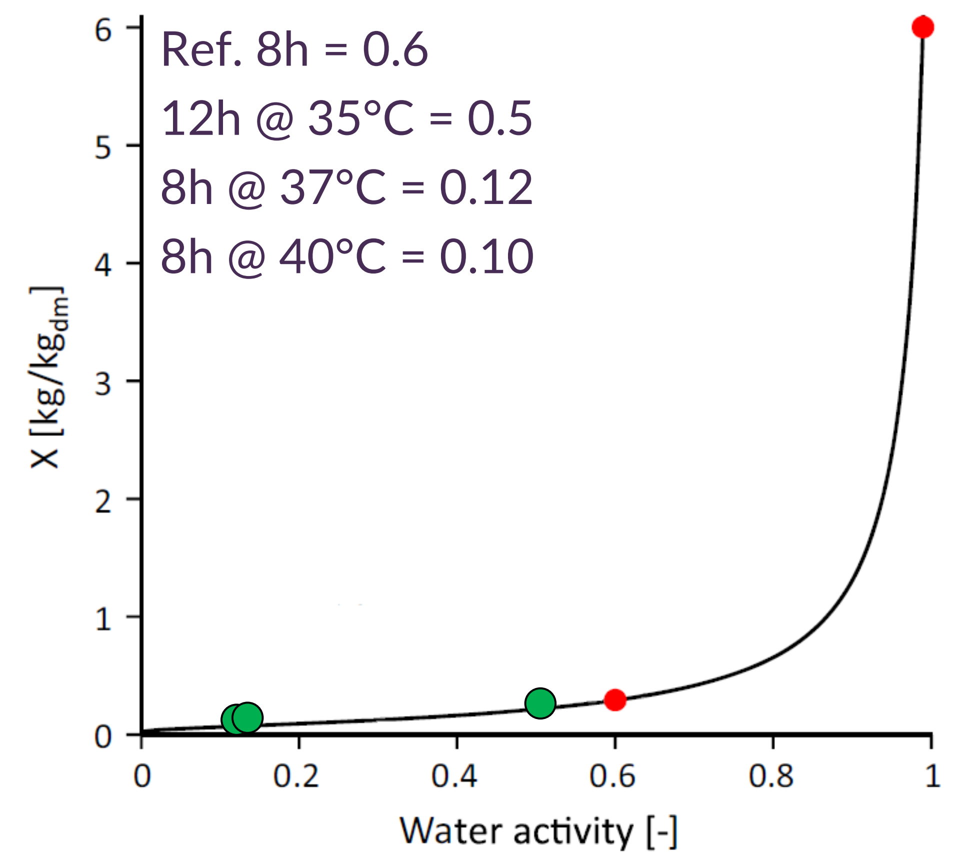
Igniting Change for the Future of Data Center Cooling
Rolf Brink, Promersion
https://youtu.be/ilaTSyB8Akc

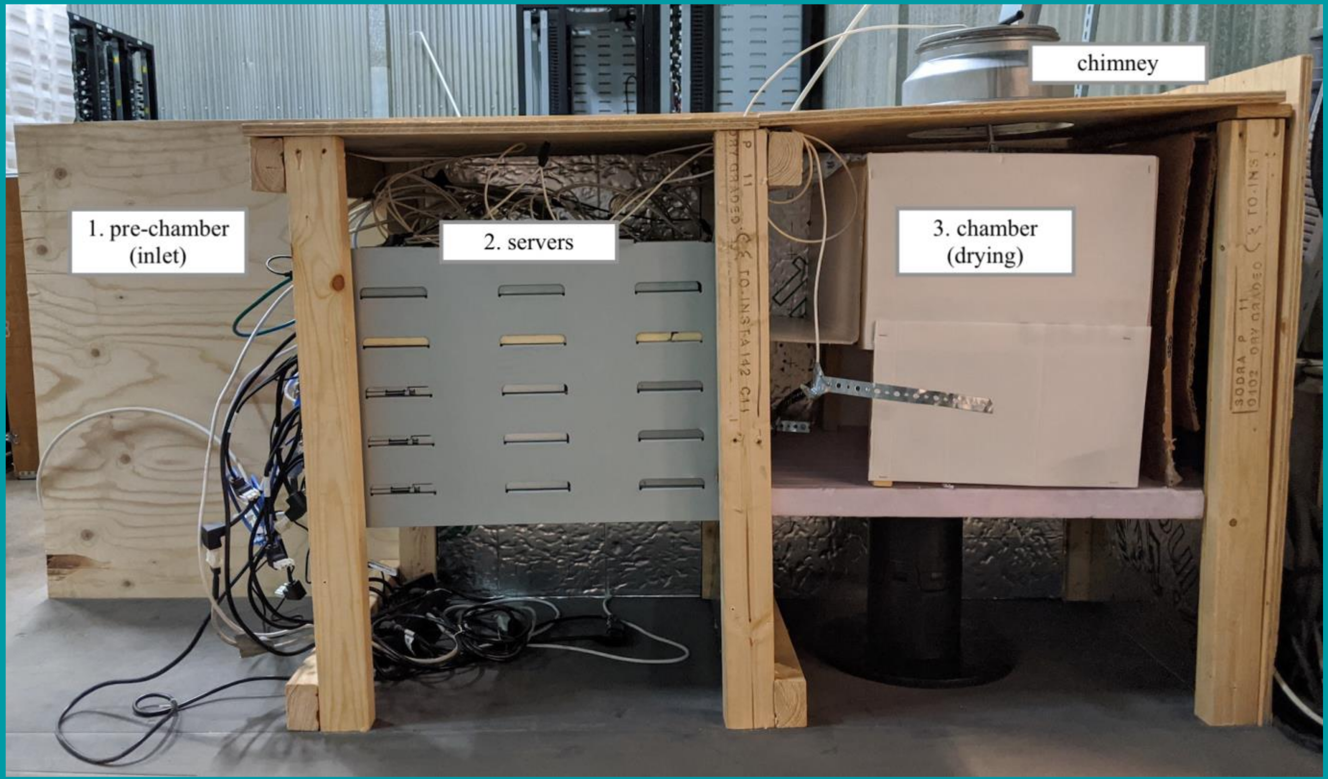


Data Center Heat Reuse

DC Excess Heat – Apple drying

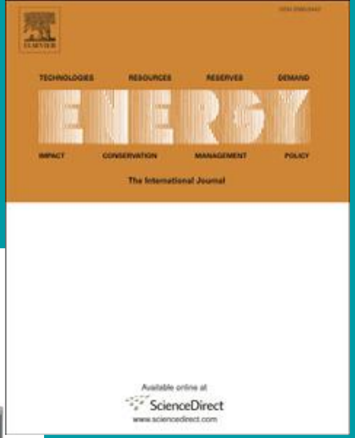
Keep the berry harvest in the region for further processing, instead of going for export.





https://doi.org/10.1016/j.energy.2020.119169

DC Excess Heat – Green House



DC power 1 MW @ 27°C

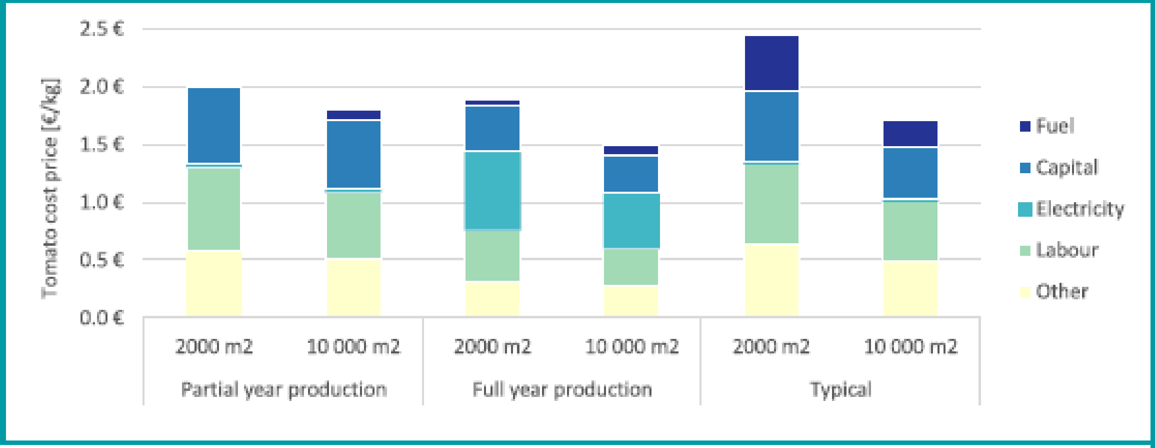
Green House area 2´000, 10´000 m^3

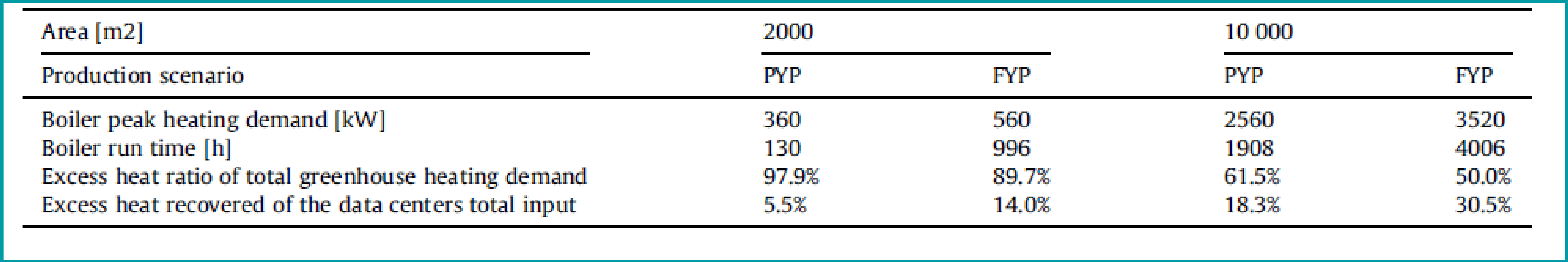
Partial and full year production

Conclusions

FYP and 10´000 m^2 most beneficial

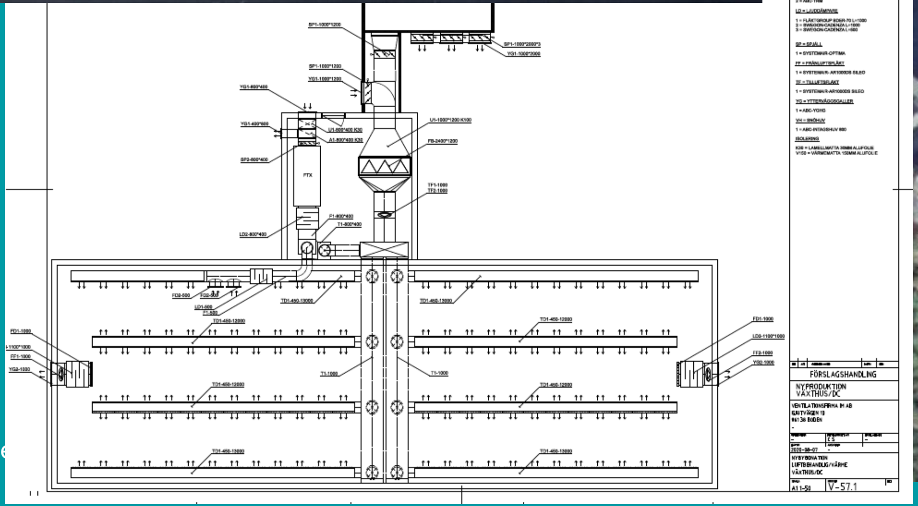
Increase self-sufficiency ≤ 8%

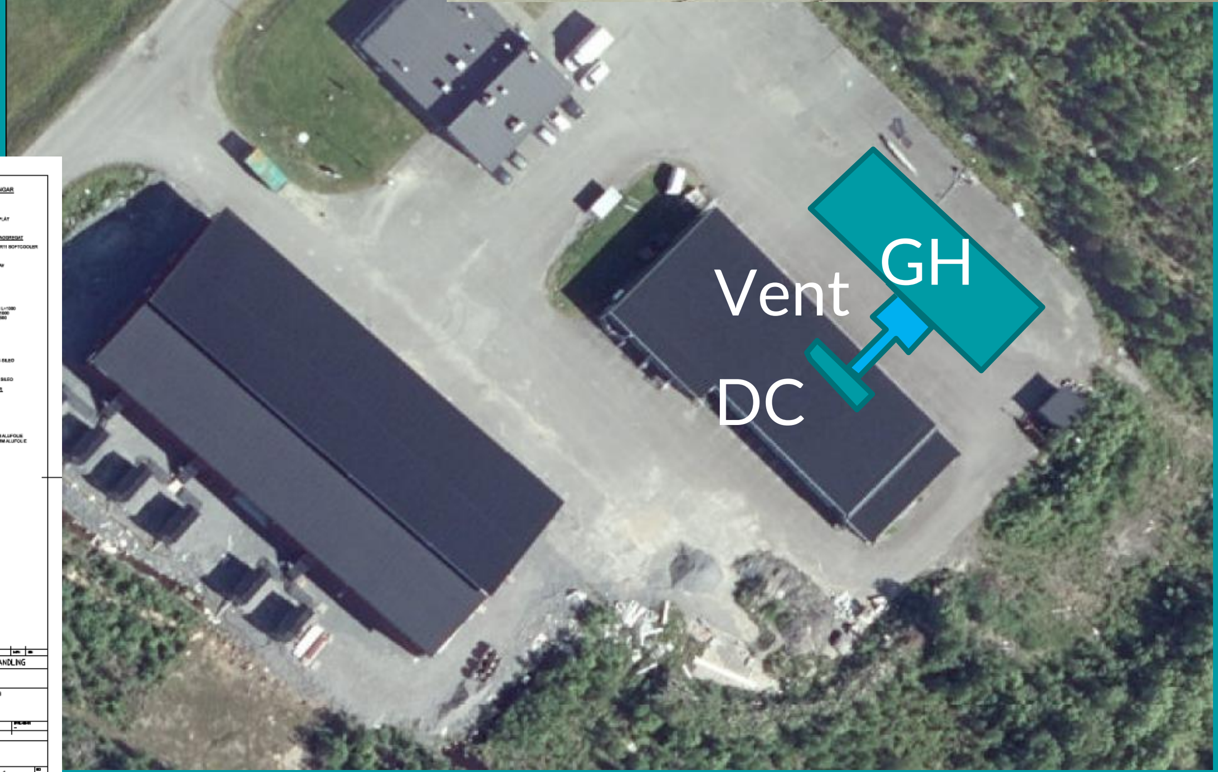




Area [m2]	2000		10 000	
Production scenario	PYP	FYP	PYP	FYP
Boiler peak heating demand [kW]	360	560	2560	3520
Boiler run time [h]	130	996	1908	4006
Excess heat ratio of total greenhouse heating demand	97.9%	89.7%	61.5%	50.0%
Excess heat recovered of the data centers total input	5.5%	14.0%	18.3%	30.5%

DC Excess Heat – Green House

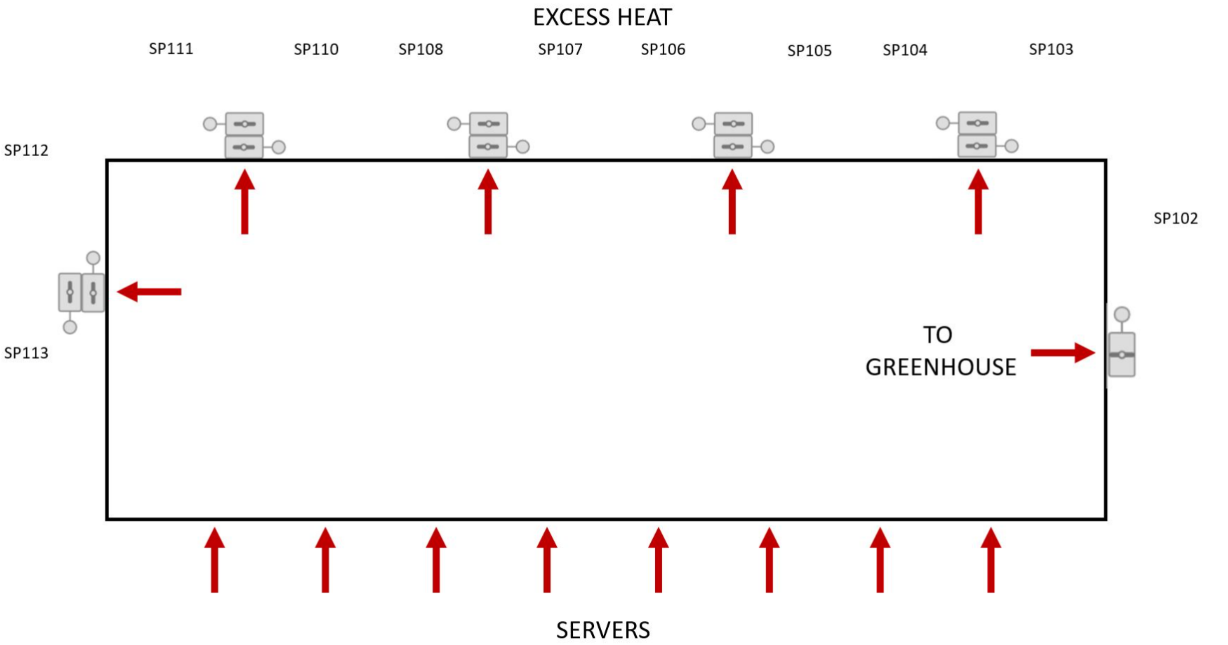


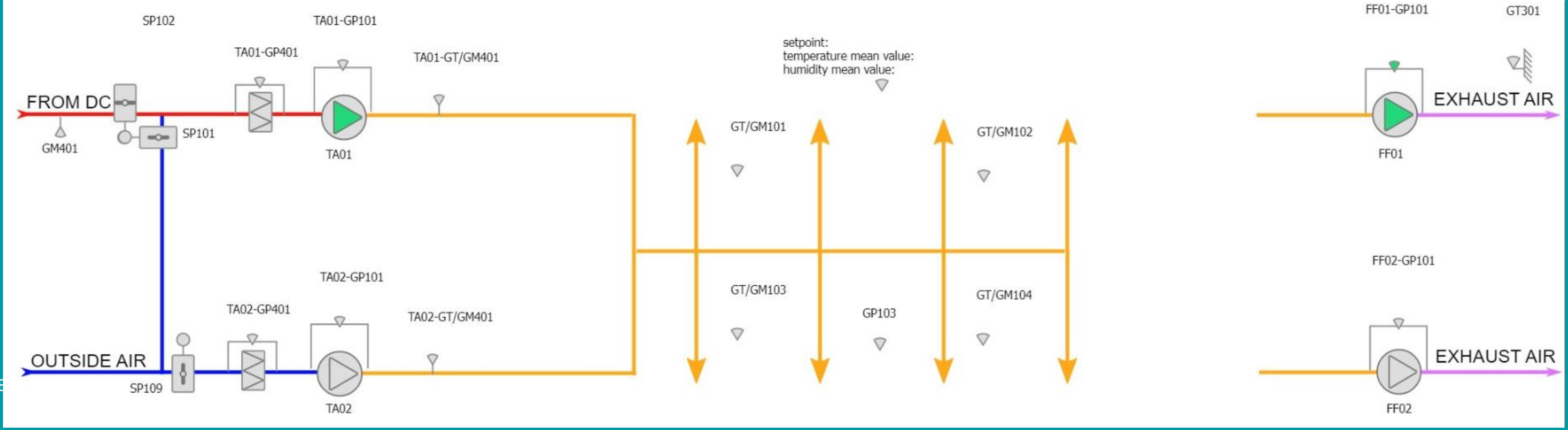


DC Excess Heat – Green House

Data collection platform measuring

- Temperatures
- Flows
- Humidity
- Pressure
- Damper position





DC Excess Heat – Mealworm Farming

10 kg of mealworms for 10 tests

DC excess heat temp (20, 30°C)

3 feeding schedules (basic, extra, tunned)

With and without access to water

Conclusions

Dominating parameters:

- DC excess heat temp
- Food
- Water

Reduction of production time 50%

https://www.nyteknik.se/premium/maskar-fods-upp-med-varme-fran-svenskt-datacenter-7005220

DC Excess Heat – Woodchip Drying

120 m^3 woodchips for 2 weeks and 10 tests

DC excess heat temp (35, 42°C)

Feeding rate (50, 75 100%)

Fan speed (50, 100%)

Conclusions

Moister reduction ≤ 10%

Energy Reuse Efficiency ≤ 42%

Dominating parameters:

- DC excess heat temp
- Fan speed

https://www.nyteknik.se/premium/de-testar-ny-roll-for-datahallar-torka-biomassa-7001305

DC Excess Heat – Hot Tap Water

DC testbed for liquid cooled servers
Connection of 2 parallel set-ups
Part of the building hot tap water system

Conclusions
Peak excess heat temp =>60 °C
Hot tap water production without heat pump

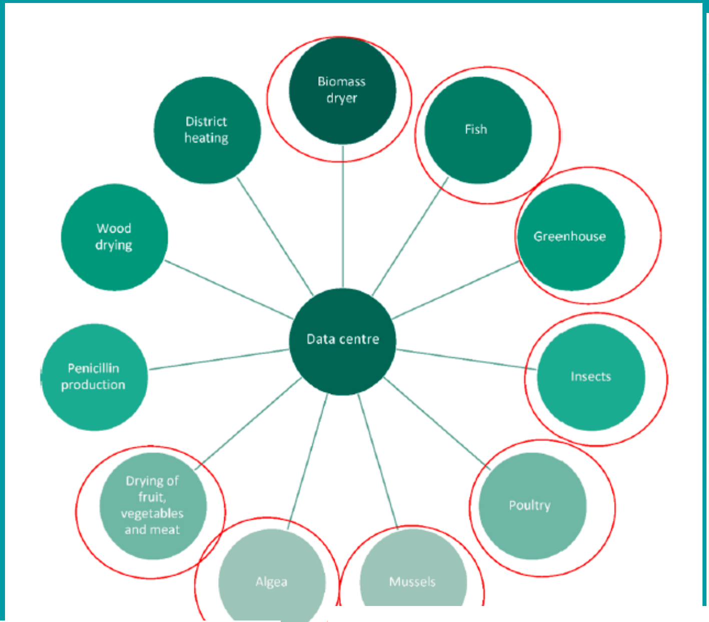
- https://doi.org/10.1016/j.ecmx.2023.100473

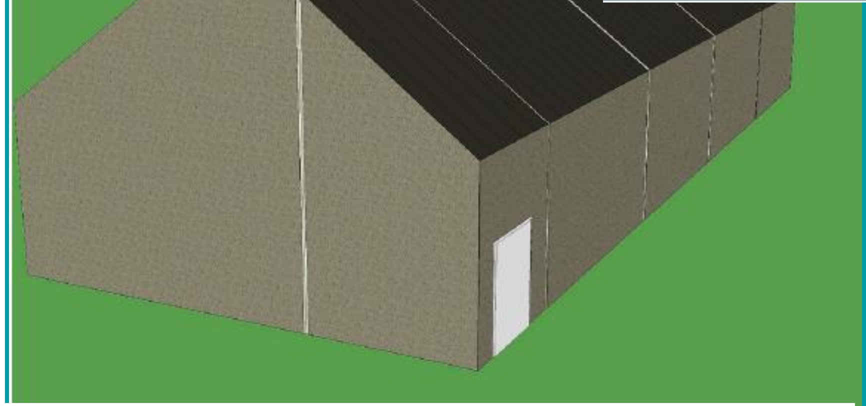
DC Excess Heat – A realistic view

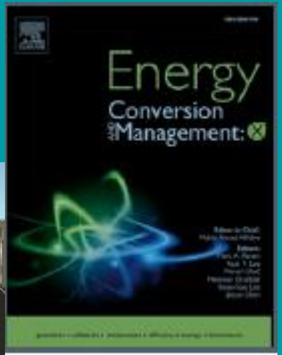
Conclusions

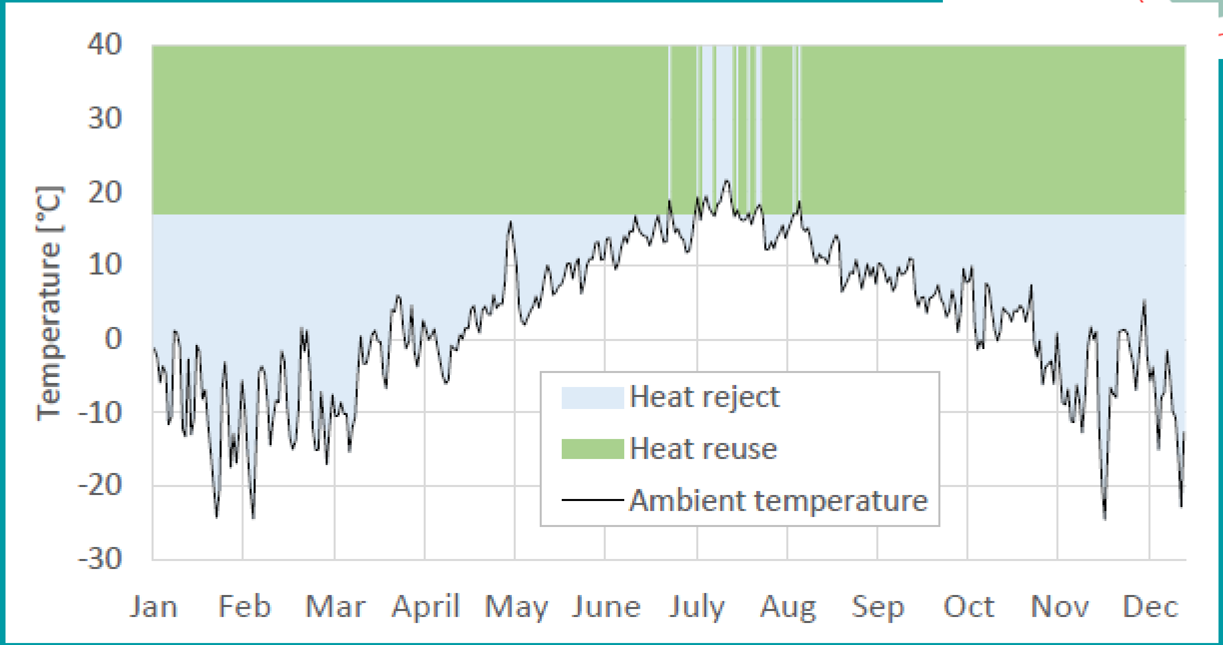
ERE; 0.5, 0,59 and 0.66 for 30, 40 ana 50°C respectively

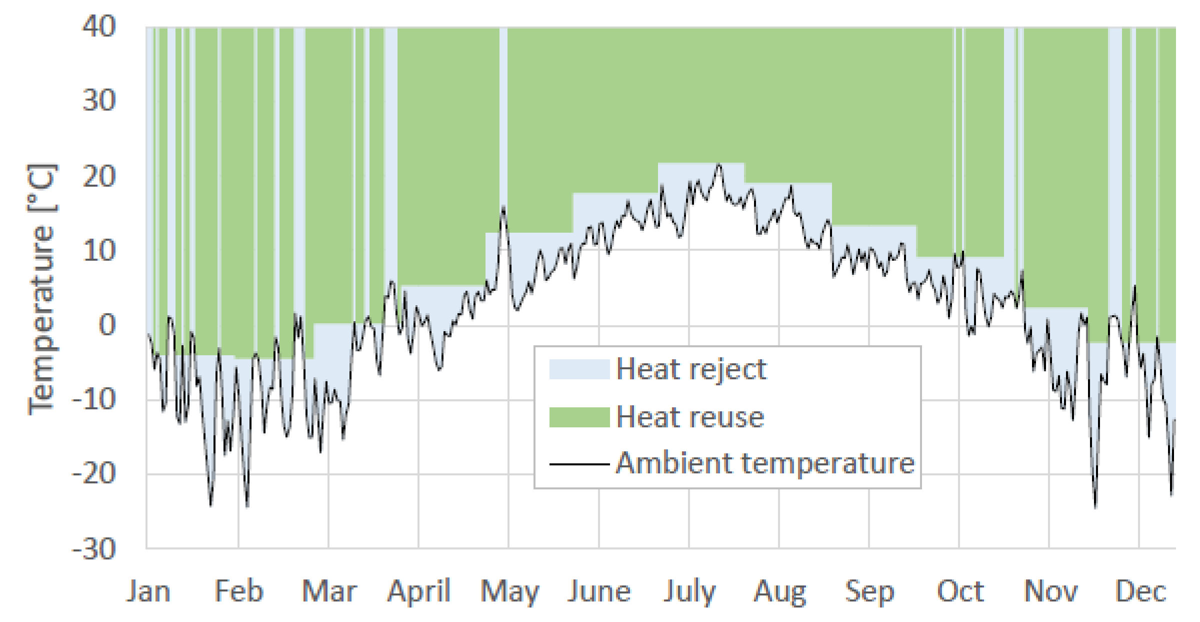
11-31% better than 1 case application.



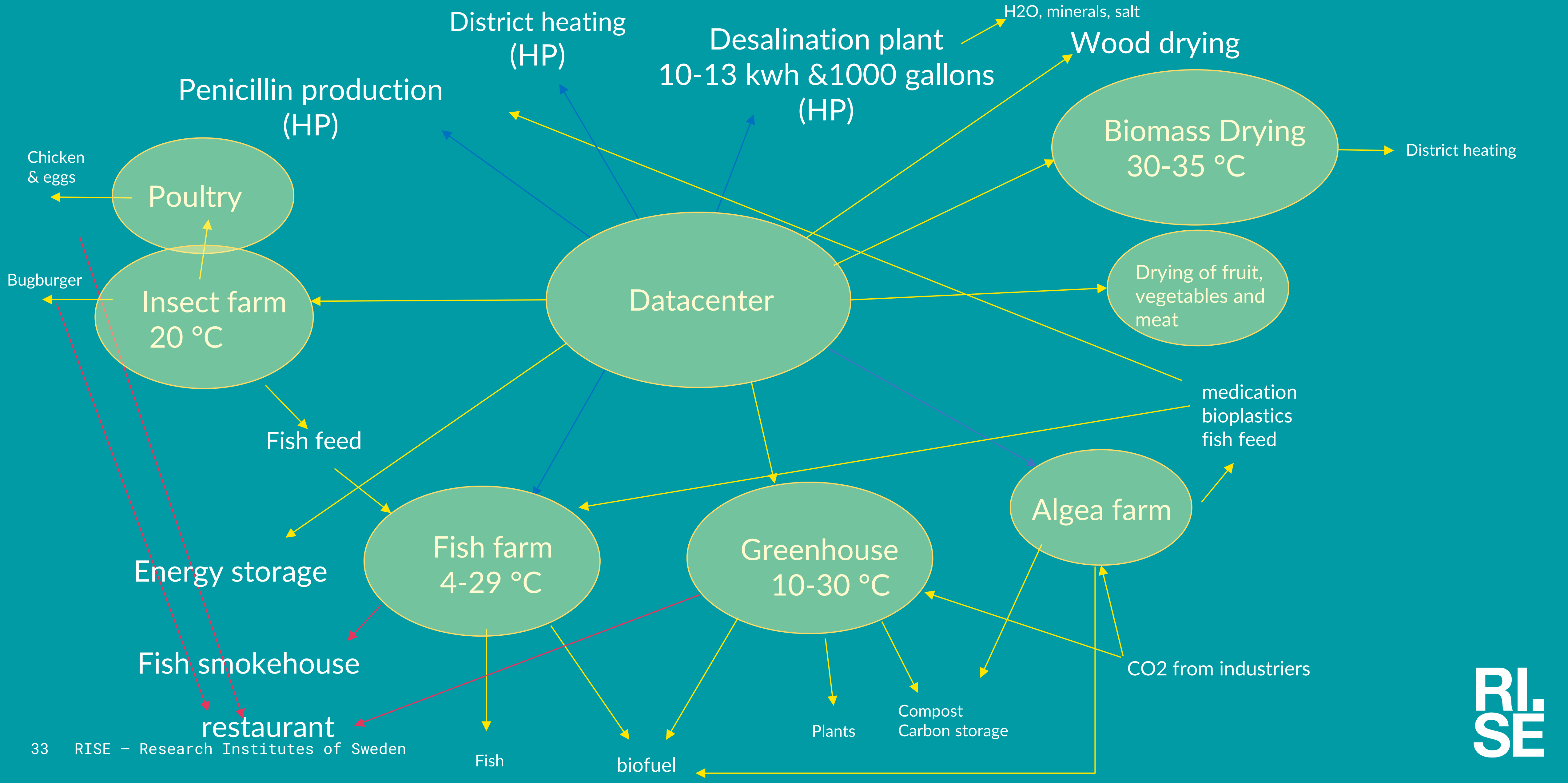








DC Excess Heat – The IUS Matrix



Thank you!

Mattias Vesterlund

mattias.vesterlund@ri.se
+46 730 418689