OVHcloud

# AI and environment: an impossible equation ?

GREGORY LEBOURG

27.11.25

Coriolis Seminars for the Environment

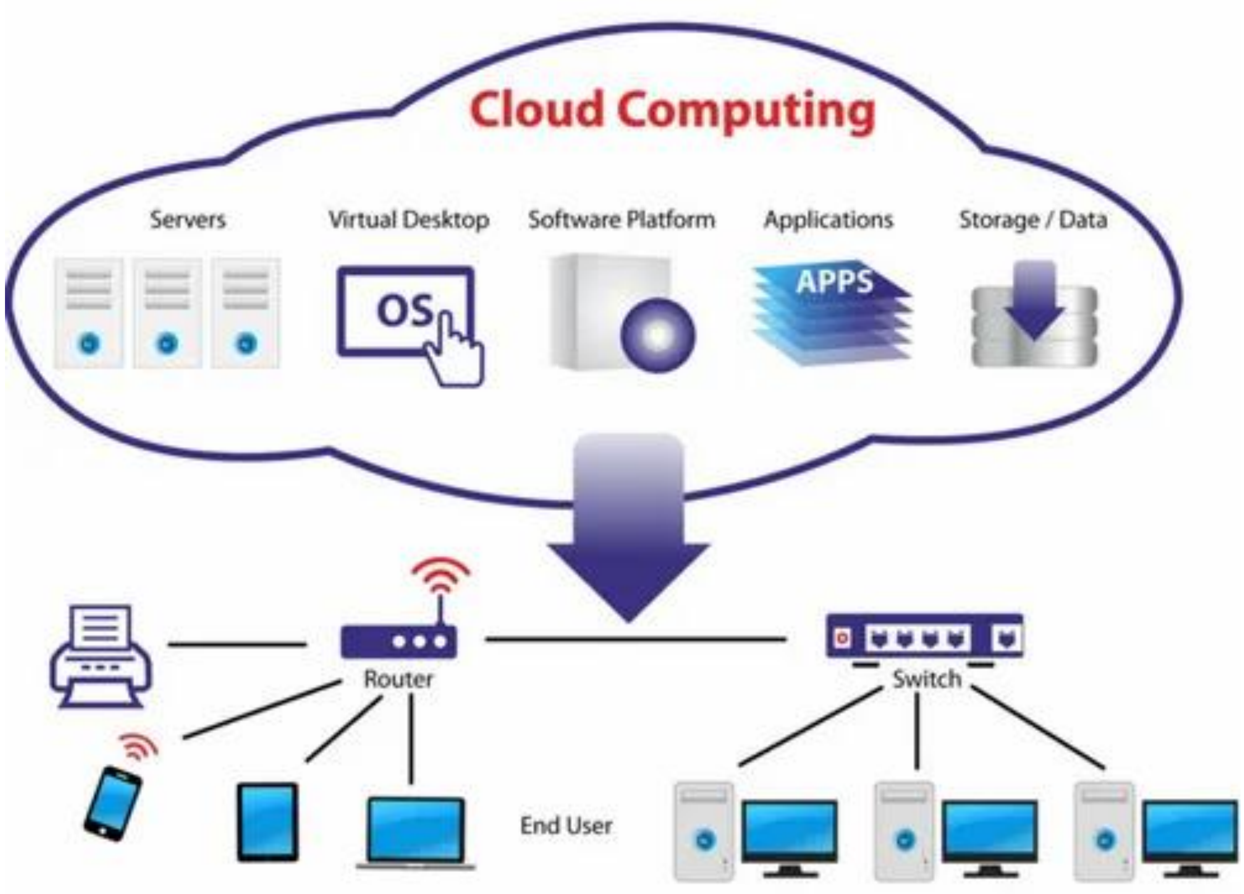
# Agenda

01

# What is Cloud AI ?

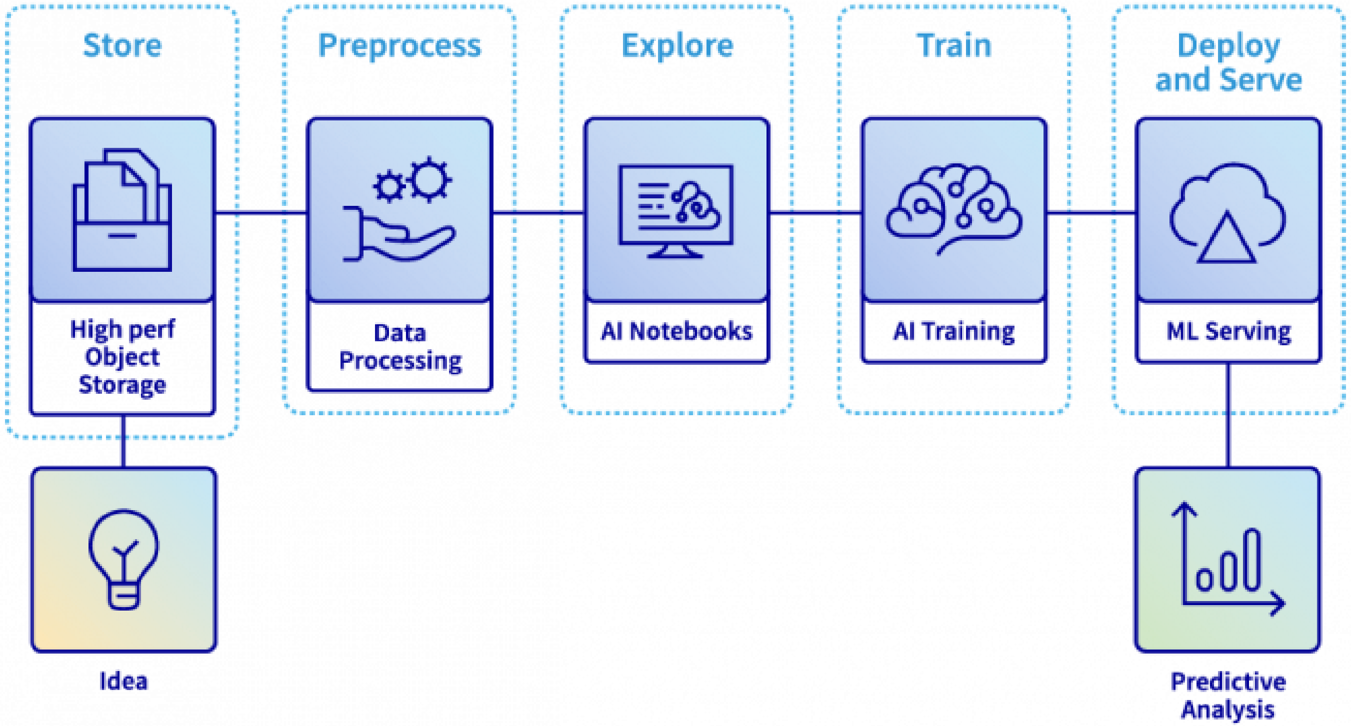
# What is Cloud ?

Cloud refers to a system that provides access to IT services (storage, computing, software) through the internet or private networks



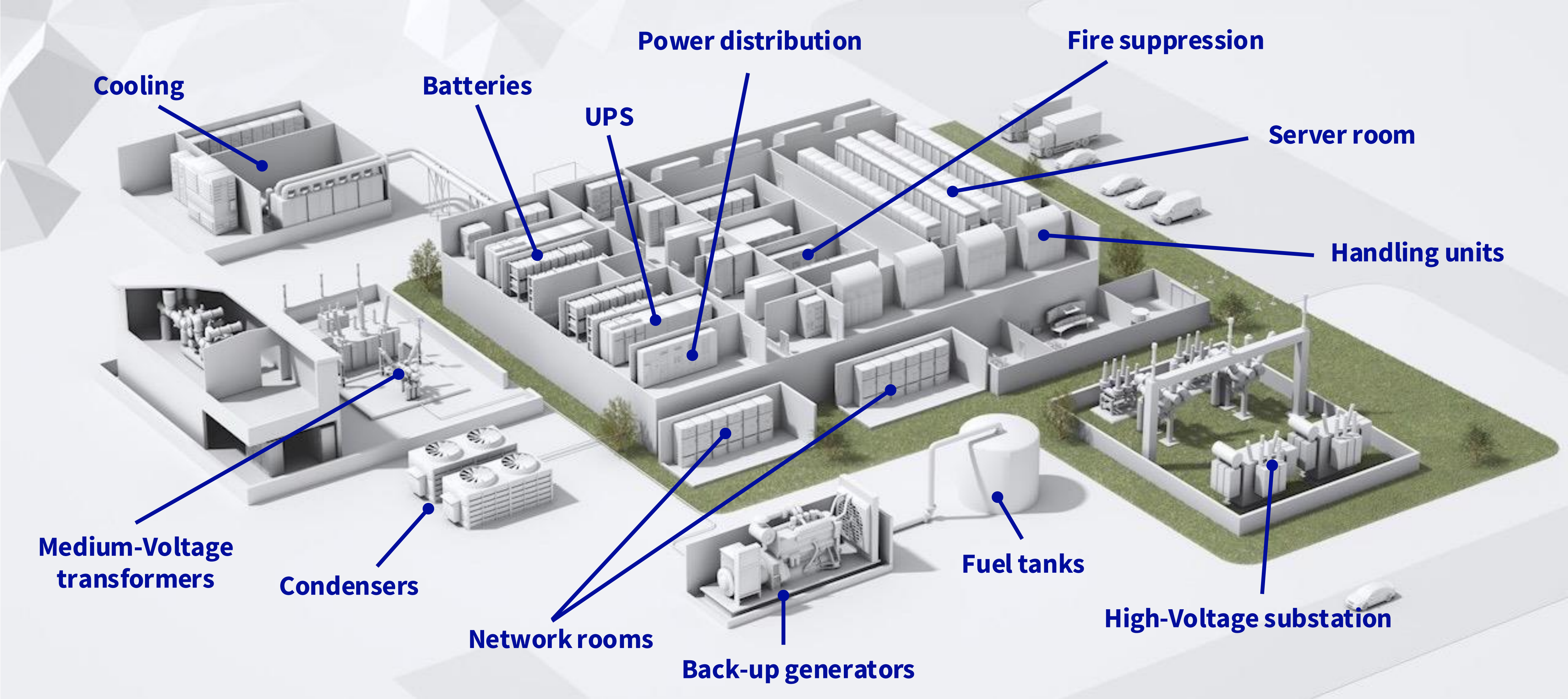
# What is Cloud AI ?

Cloud AI refers to the integration of artificial intelligence in a Public Cloud platform. It enables organisations to leverage enormous computing power and advanced AI processes without depending on costly, inefficient on-premises servers.



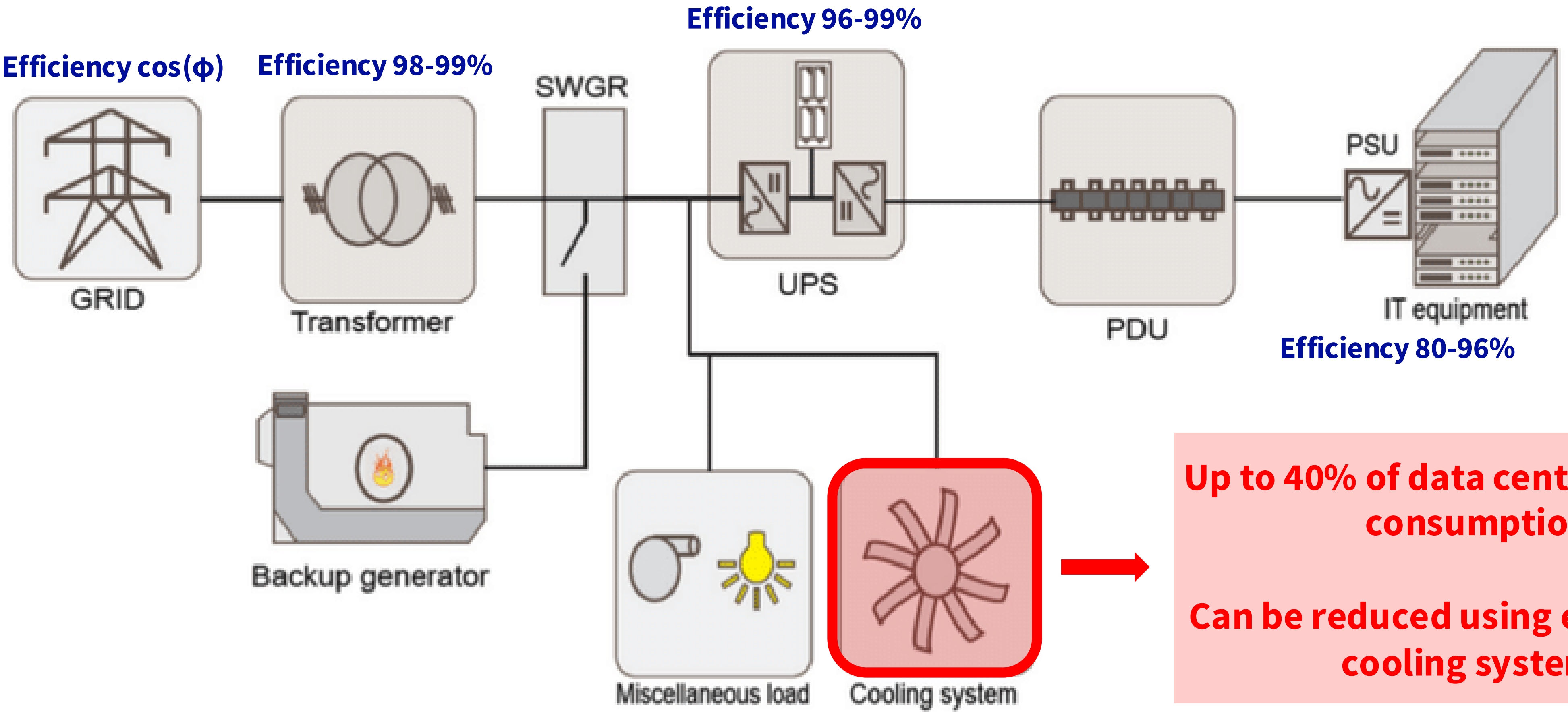
# The Cloud platforms are hosted in gigantic technical infrastructures

## Data centres = land + buildings + industrial gears + utilities (energy/water/telco)



# Whatever the design, it's all about chasing the inefficiency

2 KPIs to be optimized **PUE** (Power Usage Effectiveness) / **WUE** (Water Usage Effectiveness)

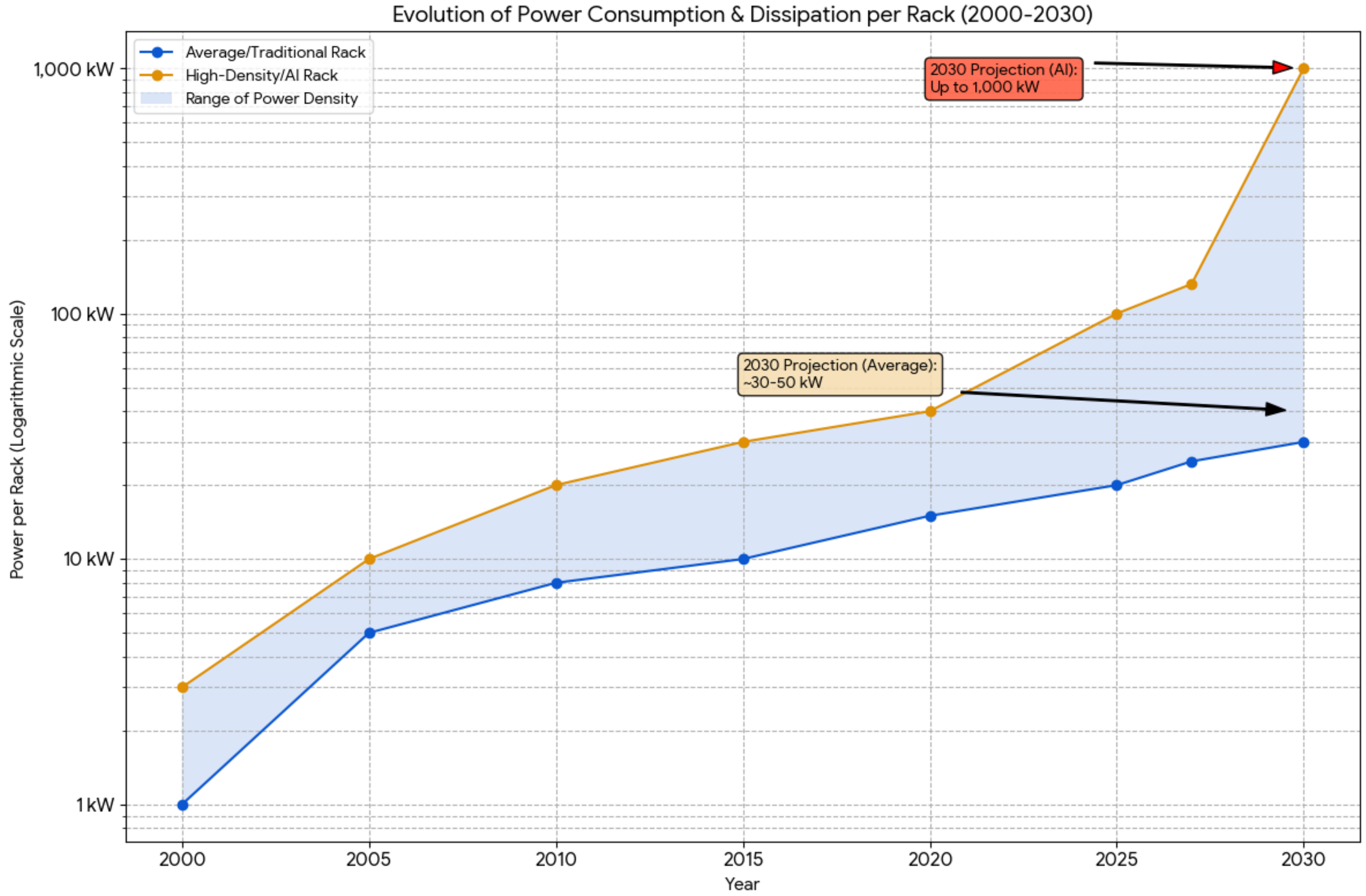



**Up to 40% of data centre electrical consumption**

**Can be reduced using evaporative cooling systems**

# The Cloud AI is built on GPU based hardware platforms

## Power density reaches a level never seen before




*Source – Lennox Data Center Solutions*

# Consequence : Cloud AI is adding to 4 main environmental issues

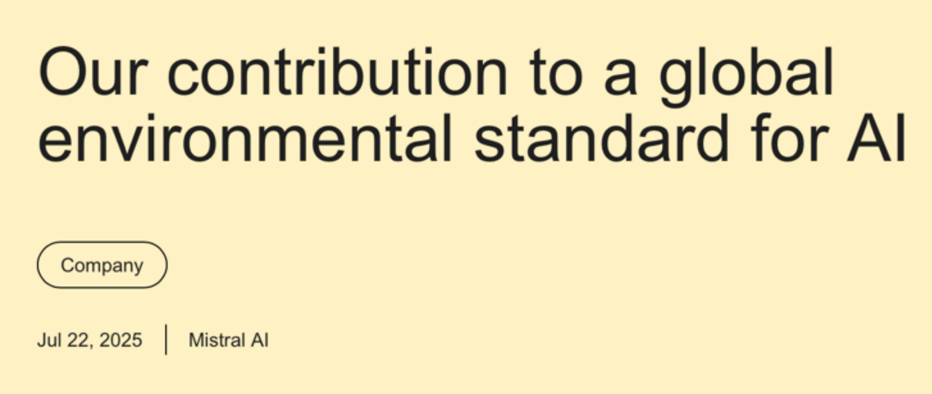


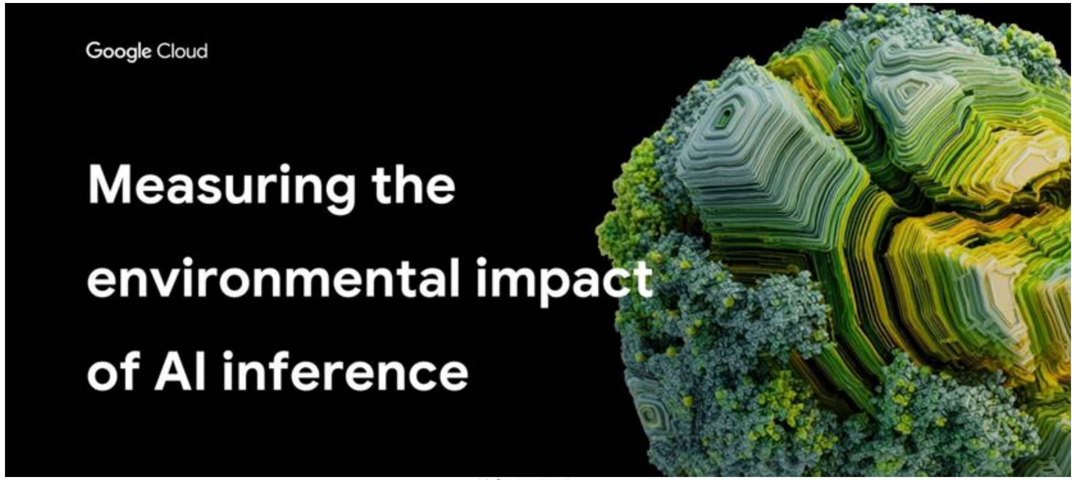
AI


# 02 Assessing the AI impact

# Only one can solve a problem by assessing it correctly…

More and more research papers being published




Infrastructure

**How much energy does Google's AI use? We did the math**

August 21, 2025

# … but we are not there yet !

Misaligned methodologies and wilful bad faith blur the picture

**MISTRAL AI_**

1.14 $gCO_2e$ /prompt

45 ml /prompt

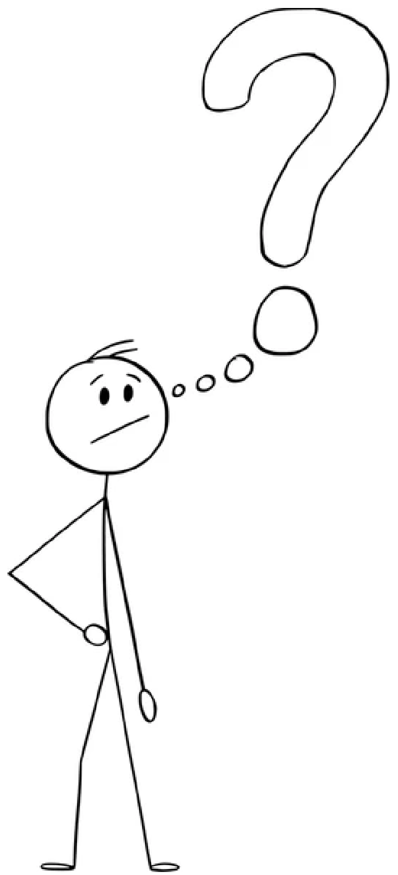
**Gemini**

0.03 $gCO_2e$ /prompt

0.26 ml /prompt

**÷ 38**

**÷ 173**



# Conclusion : Never trust the numbers before checking what's behind

Some in-depth benchmarks are available

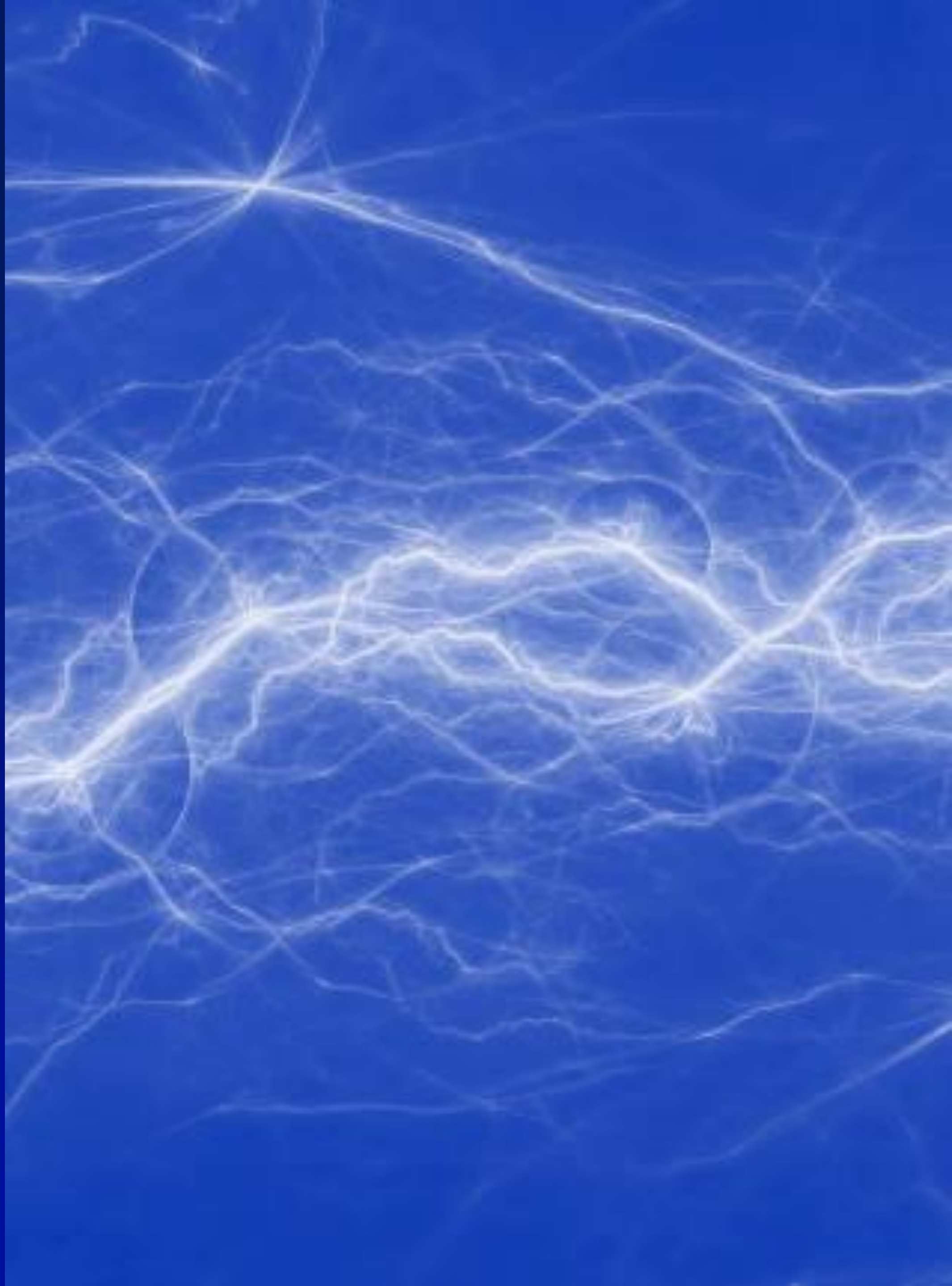
## 2. Benchmark : Carbon calculators from main Cloud providers

IJO

| | Scope 1 | Scope 2 | Scope 3 | | | |
|---|---|---|---|---|---|---|
| | Fossil energy | Electricity | Servers & equipment | Data Center | Internet | Cloud provider's Teams |
| ICT Sector Guidance (GHG Protocol) | To include | To include | Depreciation based on life span | Optional | To include | To include |
| PCR Datacenter et Cloud (ADEME) | To include | Location Based | Depreciation based on life span | Depreciation based on life span | To exclude | To include Partially |
| OVHcloud | Included | Market & Location based | 1/5 from commissioning | Included | included | Included |
| Scaleway | Included | Location based | 1/6 from commissioning | Included | included | Included |
| EXOSCALE – Outil Open Source CloudAssess | Included | Electrical mix and location to specify | Lifespan to specify | Included | not included | Partially included |
| Google Cloud Platform | Included | Market & Location based | 1/4 from commissioning | Included | not included | Partially included |
| Azure | Included | Market based, Renewable to 0 | 1/6 from commissioning | not included | included | not included |
| ORACLE / aws | Included | Market based Renewable to 0 | not included | not included | not included | not included |

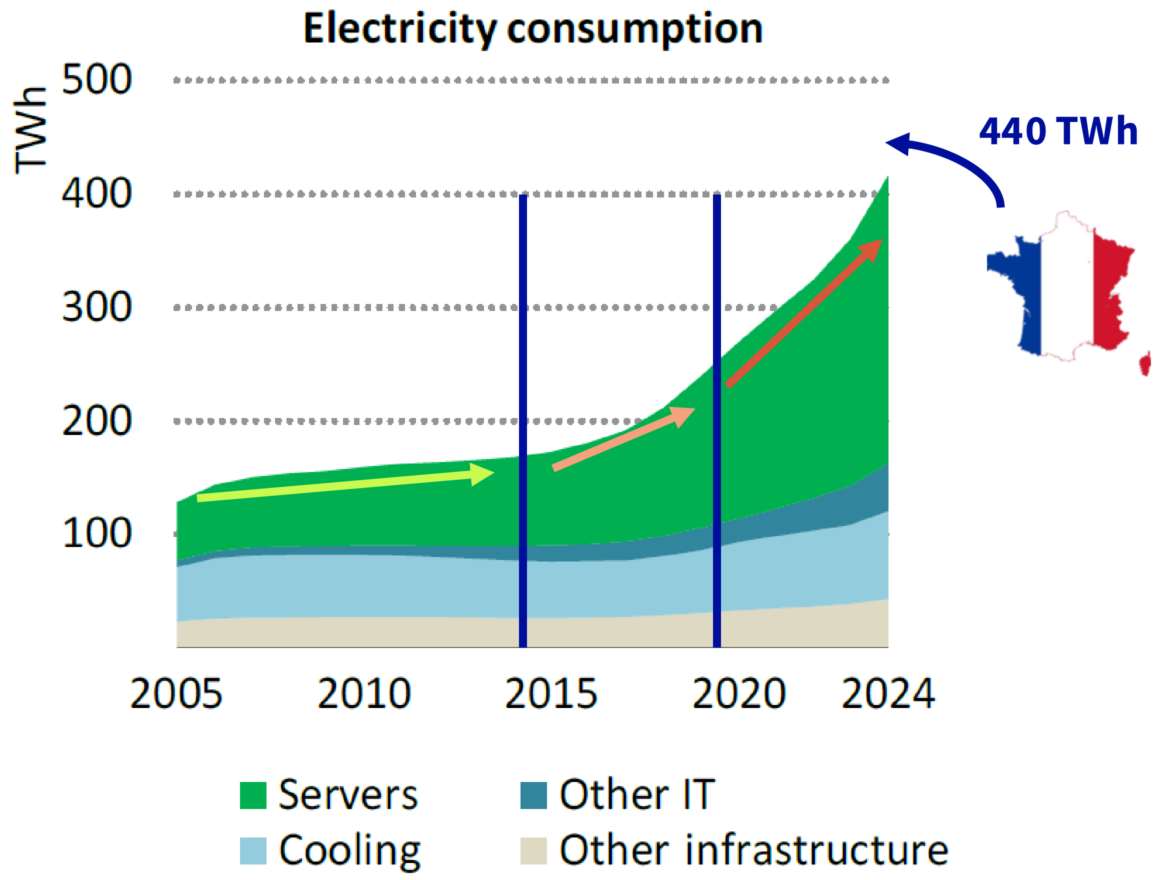
*Source – IJO study July 2025*

# 03 Data centres utilities demand



# Electricity - Let's have a look in the driving mirror

From 165 TWh in 2014, data centres energy demand has increased up to 420 TWh in 2024 (cryptocurrencies  excluded)




*Source – « Energy and AI », IEA, 2025*



YoY growth has moved from

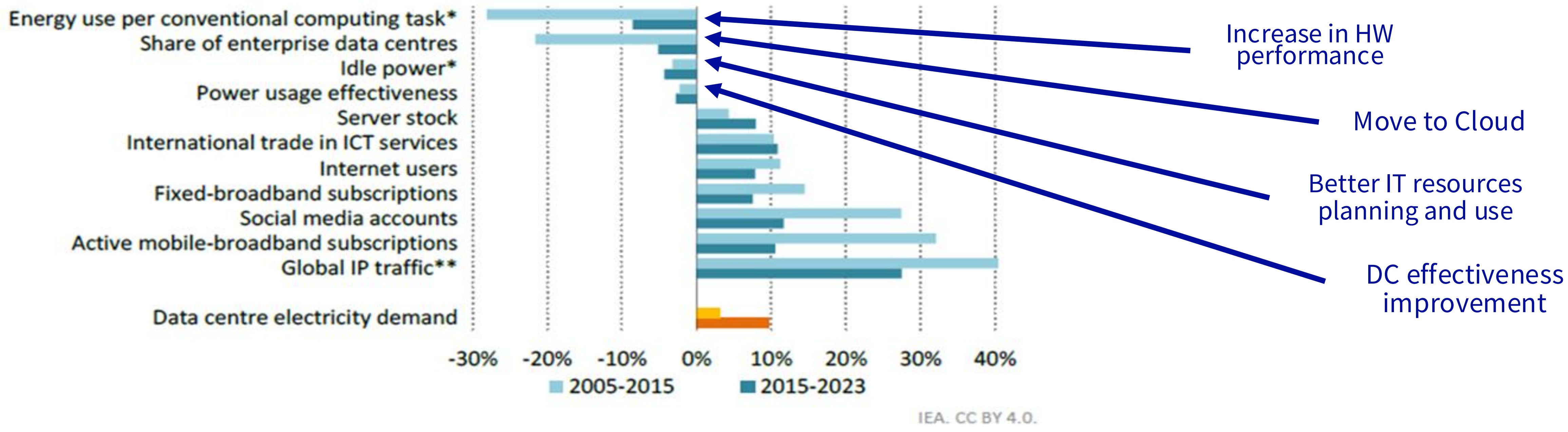
**+7 %/year** between 2014-2019

to

**+13 %/year** between 2019-2024

# Explanation : Increase in usage has overpassed the efficiency gain

Even more so that levers effects tend to decrease over time




*Robust service demand growth, an acceleration in the total number of servers and a slowdown in some efficiency indicators led to faster electricity consumption growth*

* Data starts in 2007. ** Data ends in 2022, estimated for 2022.

*Source – « Energy and AI », IEA, 2025*

# Trend : AI demand should be mitigated by chips efficiency improvement…

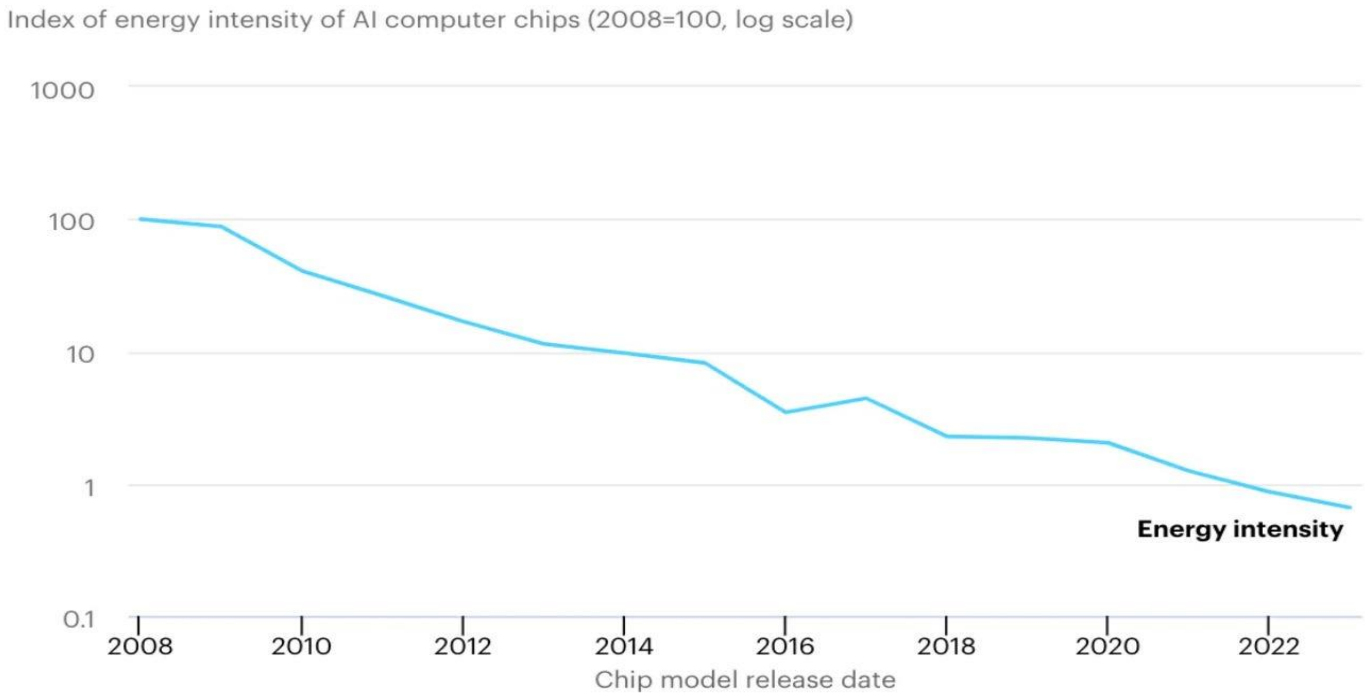
## GPUs keep on following Koomey's law

| GPU | TDP (W)[7] | TFLOPS[8] (Training) | Performance over V100 | TOPS[9] (Inference) | Performance over V100 |
|---|---|---|---|---|---|
| V100 SXM2 32GB | 300 | 15.7 | 1X | 62 | 1X |
| A100 SXM 80GB | 400 | 156 | 10X | 624 | 10X |
| H100 SXM 80GB | 700 | 500 | 32X | 2,000 | 32X |
| B200 SXM 180GB | 1,000 | 1,125 | 72X | 4,500 | 73X |
| B300 SXM 288GB | 1,400 | 1,880 | 120X | 7,500 | 121X |

*Source – Schneider Electric SPD_WP110_EN V3*

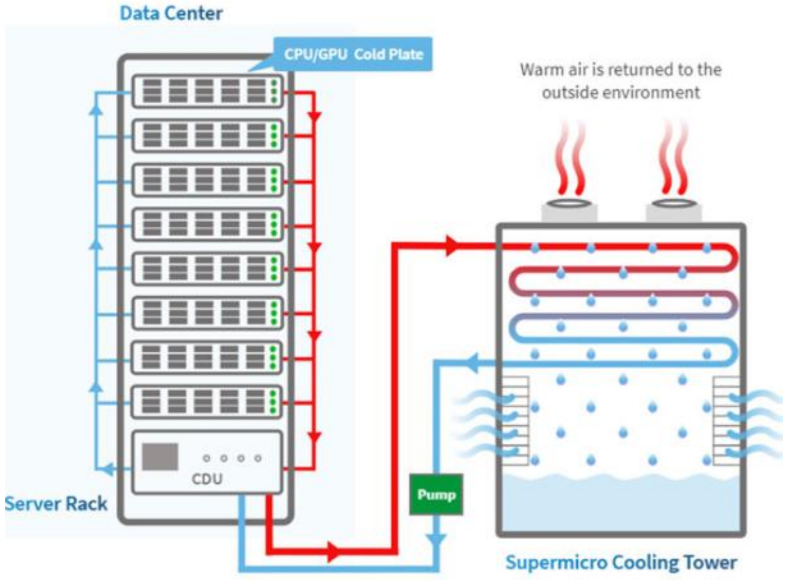
TDP (Thermal Design Power)
TFLOPS (Tera floating-point operations per second)
TOPS (Trillion of operations per second)

**Efficiency improvement of AI related computer chips, 2008-2023**

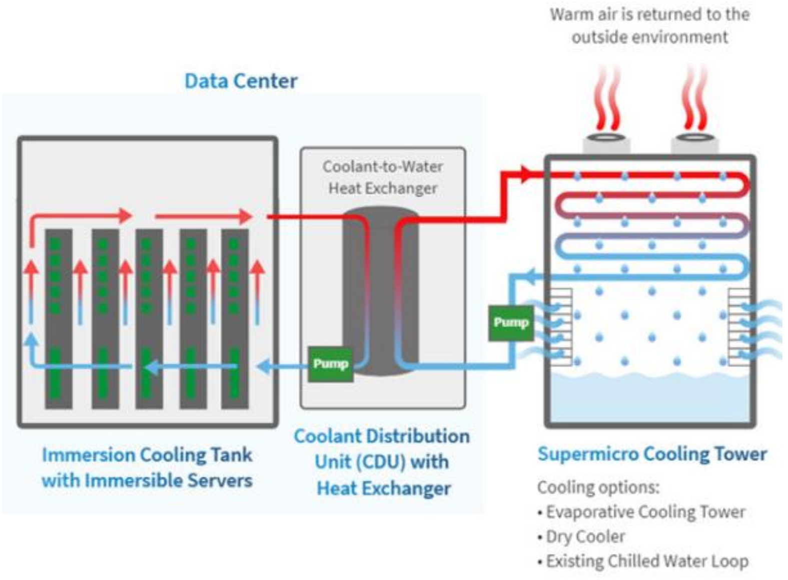


# … together with the generalization of liquid cooling for AI platforms

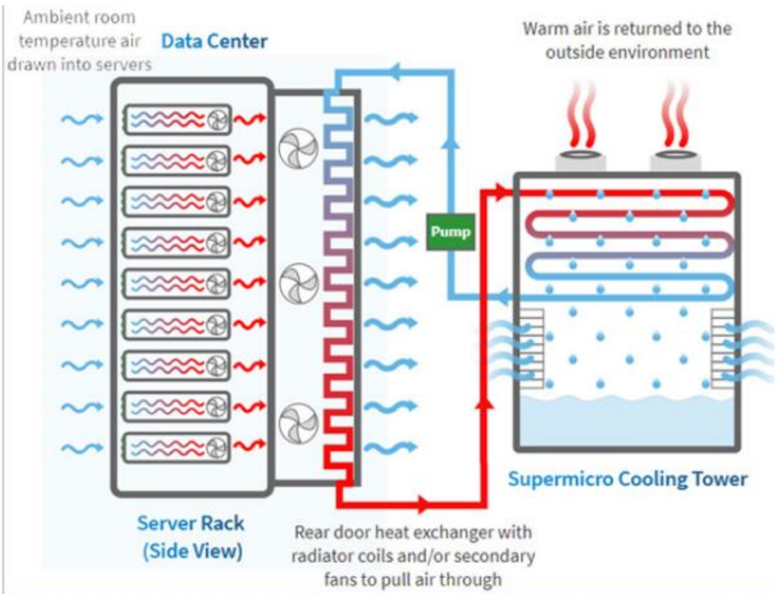
Massive adoption of liquid cooling for GPU will dramatically improve the PUE (Power Usage Effectiveness)




DIRECT TO CHIP




IMMERSIVE COOLING




REAR DOOR HEAT EXCHANGER

# Illustration : OVHcloud solution and its benefit on the PUE

Our proprietary technology allows to keep up with the rack power increase while ensuring a high-power effectiveness

**500 000 water cooled servers**
**46 Data centres**

**Introduced at scale 23 years ago**

(100+ patents)

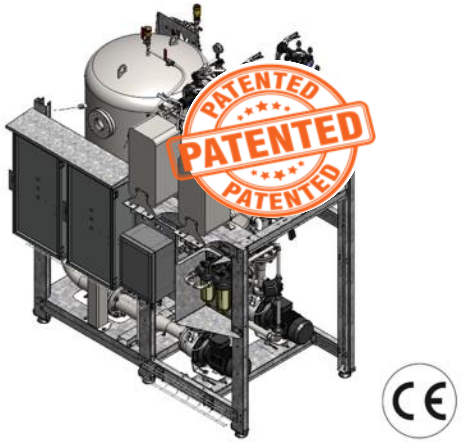
**Worldwide Power Effectiveness**

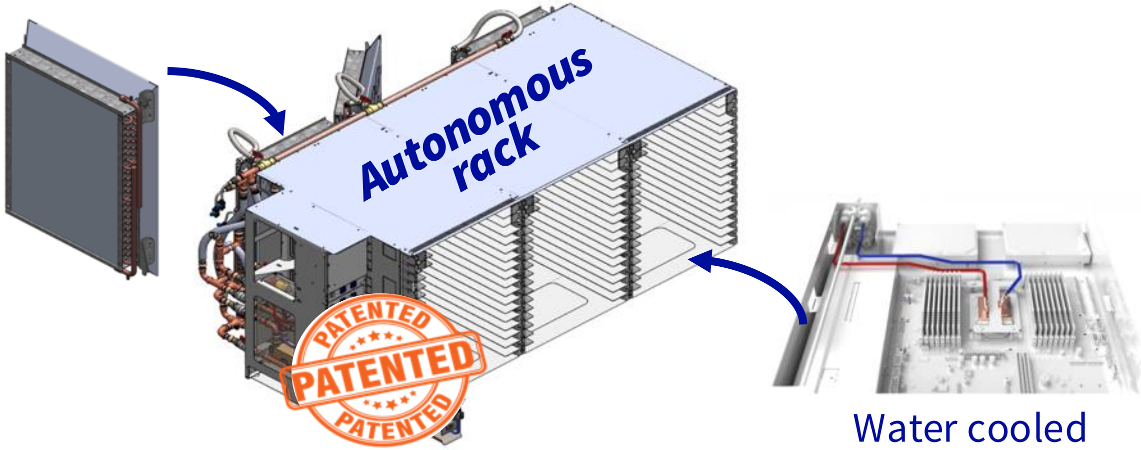
**PUE* = 1.24 (1.26 previous year)**

(vs 1.56 industry average)

**Latest Designs Performance**

**PUE as low as 1.10**



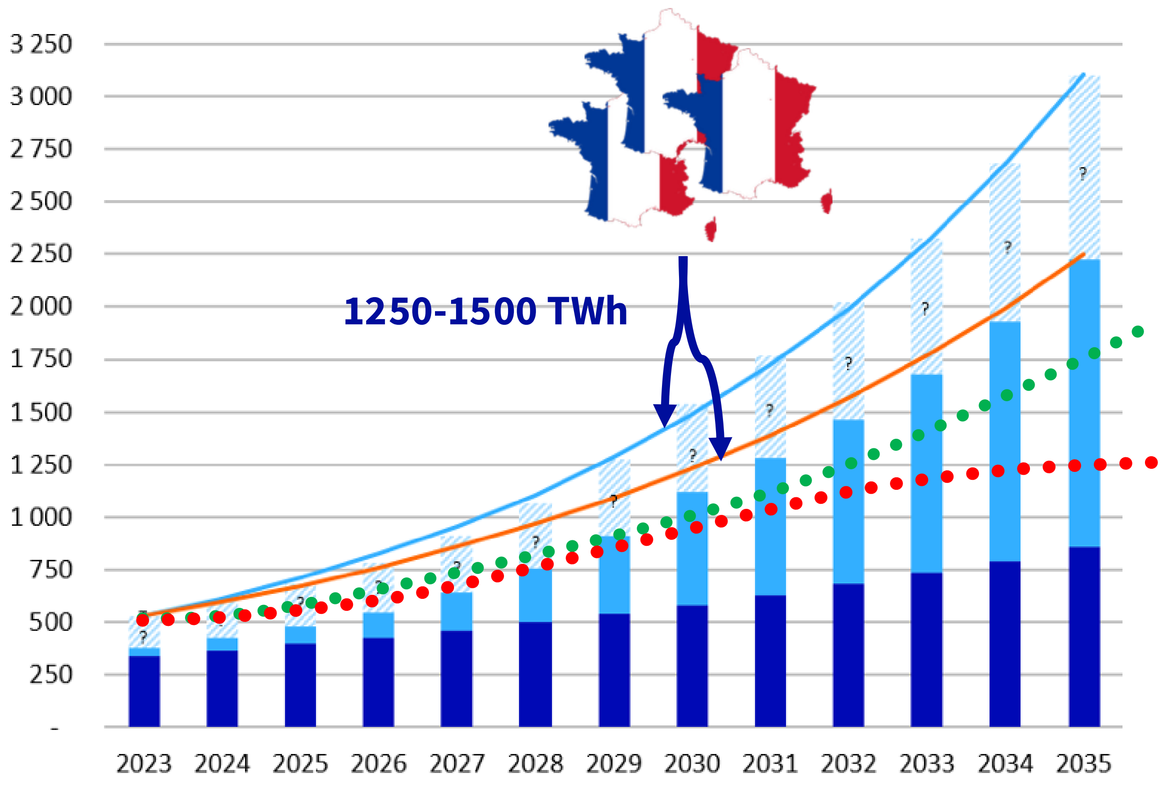





** ISO 30134 audited*

# Electricity : Let's have a look ahead but with critical thinking

Latest published forecasts are showing that energy demand is still accelerating pulling up the usage phase GHG emissions




*Source – The Shift Project 2025*

## Three dynamics

- Cryptocurrencies
- Generative AI
- Traditional usages

## Two scenarios

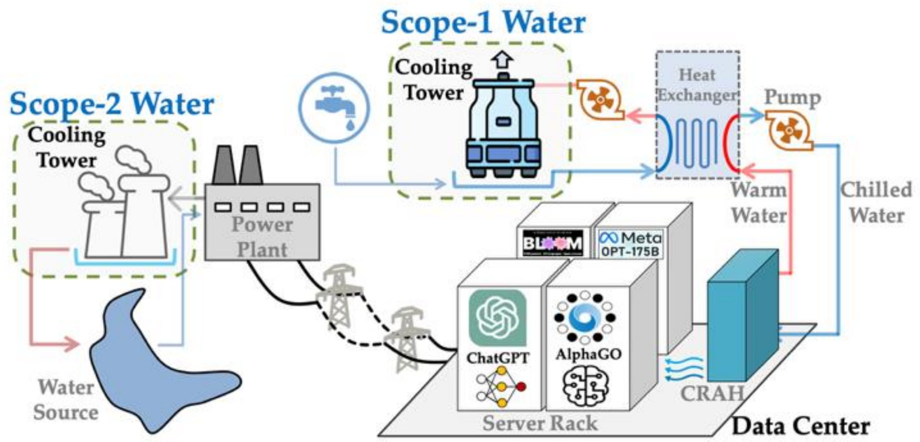
- Massive AI adoption and use of all available compute capacity
- Trend-based scenario YoY 13 % (2019-2024 value)

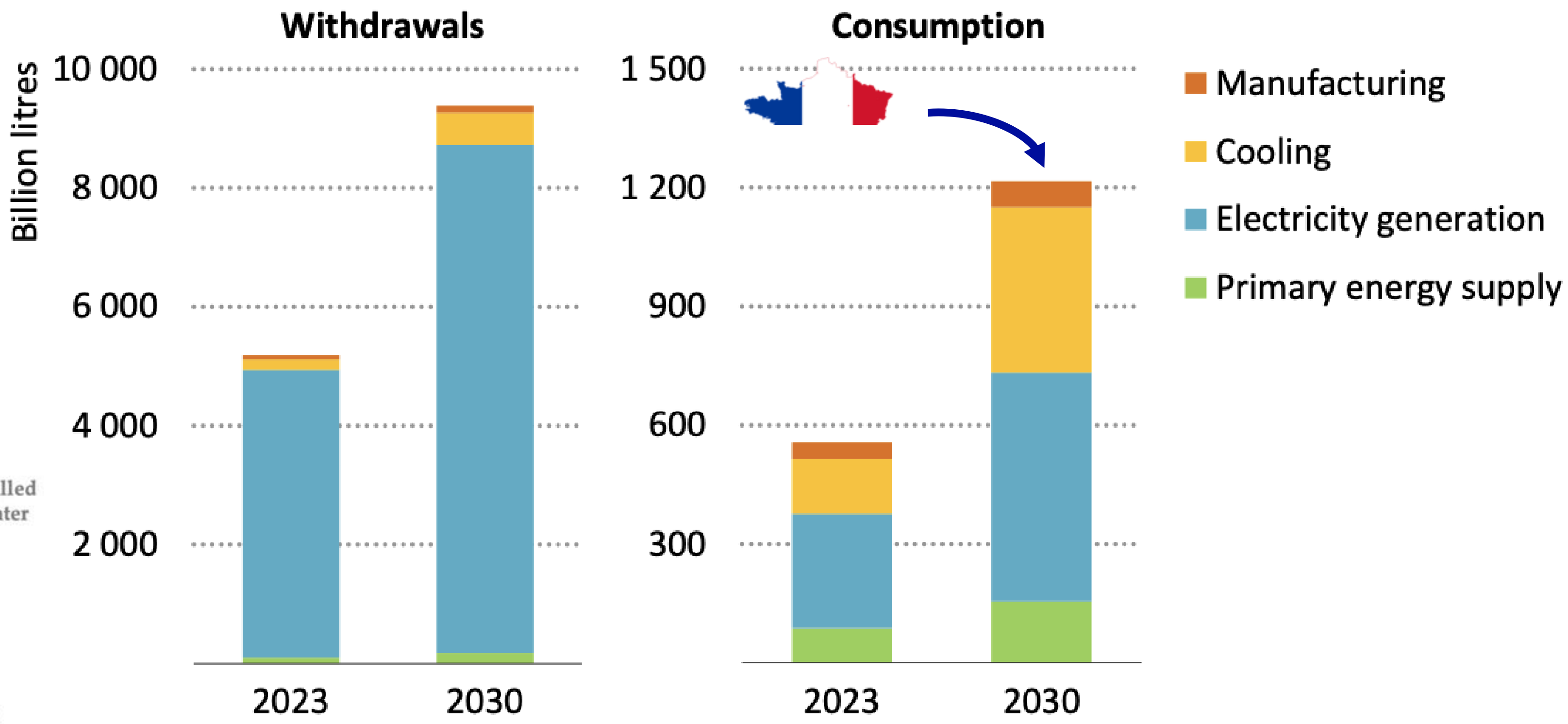
## Alternative scenarios

- Algorithms improvement / frugality
- Lack of financing / AI bubble burst / Energy crunch

# Water will logically follow the electricity trend

Massive adoption of evaporative cooling will add to the need of water for electricity generation






***Water consumption more than doubles between 2023 and 2030***

*Source – « Energy and AI », IEA, 2025*

# Illustration : OVHcloud solution and its benefit on the WUE

Our proprietary evaporative cooling technology on dry coolers allows to keep the direct water usage low

**High delta temperature and water profiles**

**Delta T at 20K (25-45°C)**

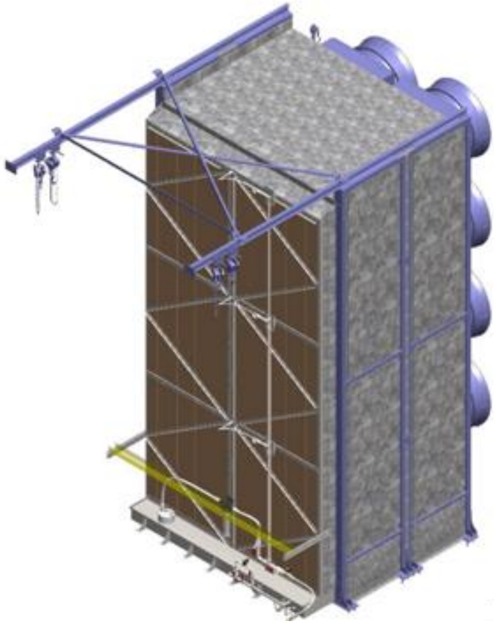
**Worldwide Water Effectiveness**

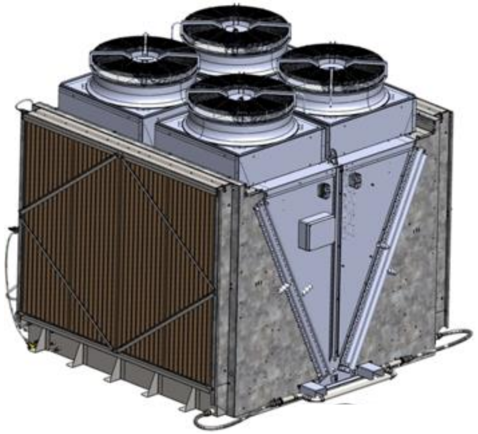
**WUE* = 0.34 l/kWh (0.37 previous year)**

(vs 1.00 to 1.50 l/kWh industry average)

**Latest Designs Performance**

**WUE as low as 0.10 l/kWh**





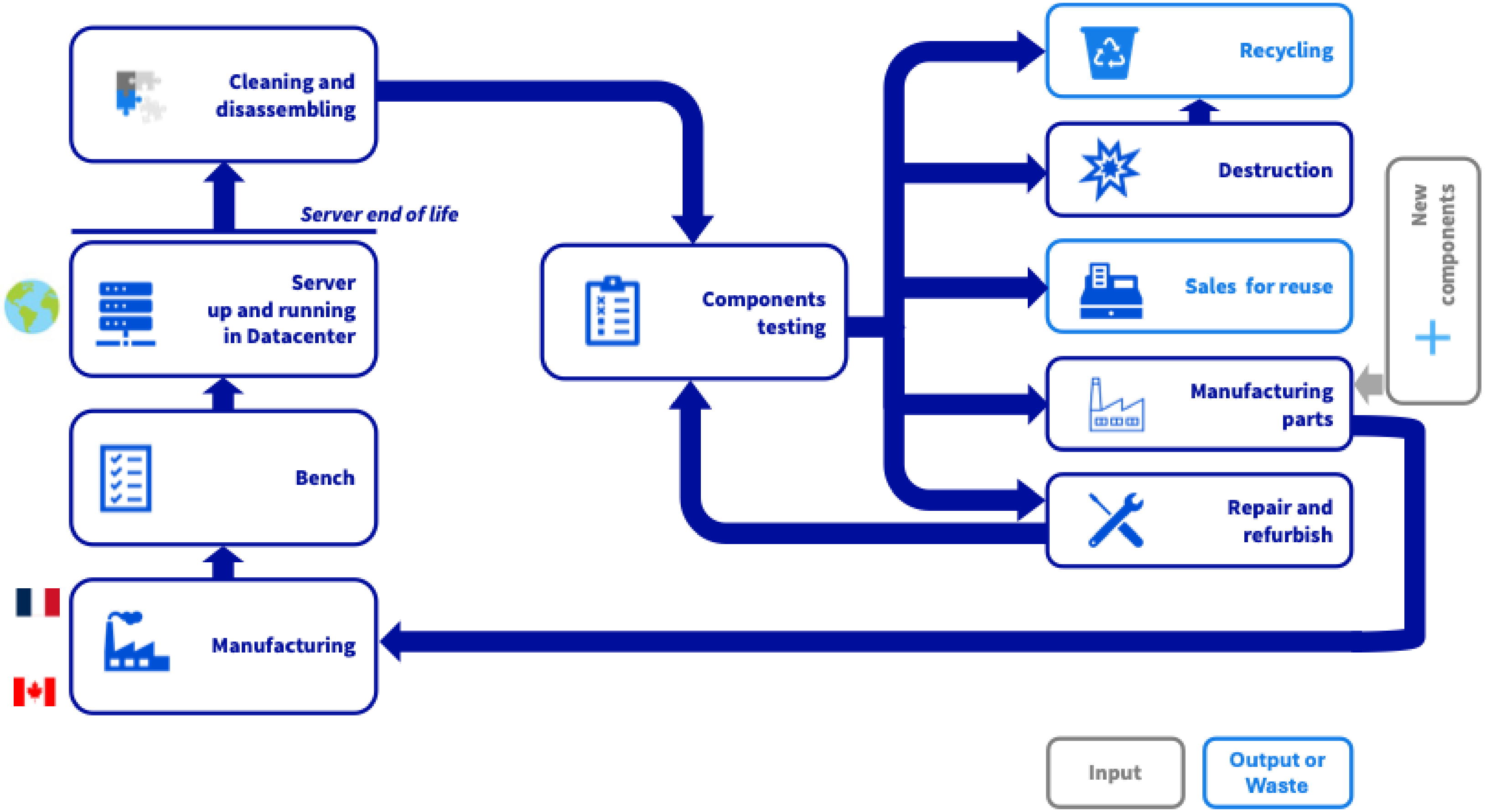
** ISO 30134 audited*

# 04 Beyond utilities

# Embodied emissions and needs for minerals are going through the roof

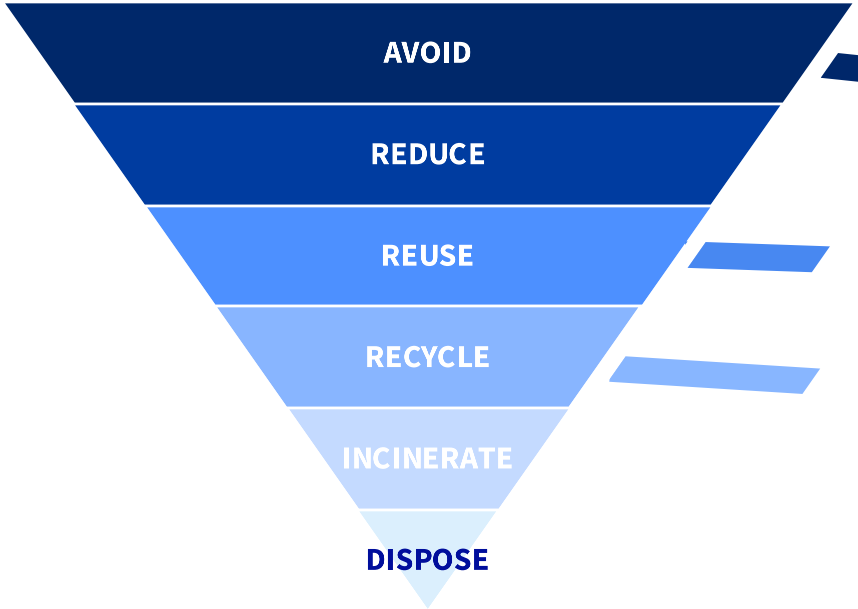
Components lifespan extension and reuse policy are more than ever necessary

## Illustration : OVHcloud reverse supply chain



# Comes last but is not the least : minerals recovery

GPUs are starting to pay off




| | CPU **GPU** (kg) | HDD (kg) | SSD (kg) | MB (kg) | RAM (kg) | TOTAL (kg) |
|---|---|---|---|---|---|---|
| Reused components (int. reverse supply chain) | 350 **277** | 2870 | 1550 | 2430 | 1320 | **8797** |
| Resold components (ext. reverse supply chain) | 1070 **80** | 15440 | 730 | 8580 | 1170 | **27 070** |
| Recovered minerals (recycling chain) | 220 **180** | 9910 | 1030 | 15620 | 70 | **27 040** |

# 05 Users and developpers role

# Choose the right GPU platform matching the target performance

Embodied emissions greatly vary from one reference to another

## Typical “cradle to gate” values of new GPUs (to be amortized over 5 years)

| | |
|---|---|
| Intel CPU range | 5 – 25 kgCO2e |
| NVIDIA GPU L4 | 50 kgCO2e |
| NVIDIA GPU L40s | 100 kgCO2e |
| NVIDIA GPU A100 | 150 kgCO2e |
| NVIDIA GPU H100 | 150 kgCO2e (163 kgCO2e**) |

## Typical “cradle to gate” values of refurbished GPUs

| | |
|---|---|
| NVIDIA Tesla V100 (2017) | 0 kgCO2e |
| NVIDIA Quadro RTX5000 (2018) | 0 kgCO2e |

**Source – Intel PCF / OVHcloud LCA*
***Source – NVIDIA LCA*

# Choose the country based on your data location constraints

Usage emissions greatly vary from one country to another

## NVIDIA H100

**kCO2e/month (Location based)**

- Manufacturing 73
- Operations 4
- Electricity 598

**- 85%**

## NVIDIA H100

**kCO2e/month (Location based)**

- Manufacturing 73
- Operations 4
- Electricity 23

# Choose the right AI instances in the portfolio of services

Pick up what you really need

## AI end-points features to be looked at

- Quantisation optimisation (FP8 to FP4 = -50% in computing needs)
- Context caching (-30% to -40% in computing needs)
- Speculative decoding (small models first, then large models if results are not accurate enough)
- Model architecture change (model split in “n” expert models, the prompt is routed to the appropriate expert model)
- Batch processing (50k prompt treated asynchronously to optimize the GPU resources planning)
- Number of parameters reduction : 7 billions parameters models now as accurate a 100s billions parameters models 2 years ago