SUSTAINABLE COMPUTING

Piotr Mańkiewicz | Format sp. z o.o.


Sustainability:
Clients
Ecosystem – Press – Analysts
World
Resources – DC
Priority?

MIT Technology Review

SUBSCRIBE

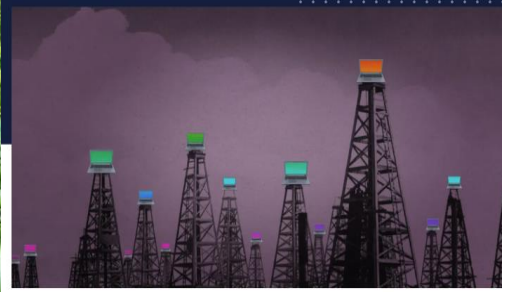
ARTIFICIAL INTELLIGENCE

AI's carbon footprint is bigger than you think

Generating one image takes as much energy as fully charging your smartphone.

By Melissa Heikkilä

December 5, 2023



STEPHANIE ARNETT/MITTR | GETTY, ENVA

This story originally appeared in The Algorithm, our weekly newsletter on AI. To get stories like this in yo

Chile partially pulls Google data center permit, seeks tougher environmental checks

Story by Reuters • 11h

FILE PHOTO: Google logos are seen during the announcement of the plans for their data centre expansion in Sa
REUTERS/Ivan Alvarado/File Photo
© Thomson Reuters

SANTIAGO (Reuters) -A Chilean environmental court partially reversed a permi
center in the country on Tuesday, asking the U.S. company to revise its applic
effects of climate change.

Google first received initial authorization for its announced $200 million Cerrillos D
2020, but the project has since drawn an outcry from residents and local officials c
capital's parched aquifer.

REUTERS® World Business Markets Sustainability Legal Breakingviews Technology Inv

Power mad: AI's massive energy demand risks causing major environmental headaches

By **Ben Payton**

December 4, 2023 2:14 PM GMT+1 · Updated 2 months ago

Industry Insight from Ethical Corporation Magazine, a part of Thomson Reuters.

The Register

Too many bytes and not enough bricks for datacenters

More than 3,000 needed to meet demand, and that won't be easy, says Aggreko

Dan Robinson

Tue 25 Jul 2023 / 09:32 UTC

The European datacenter industry is facing issues meeting the growing demand for capacity with materials and heavy equipment to build sites in short supply, among other factors.

A report produced by power generator supplier Aggreko [PDF] claims that demand is outstripping supply in the datacenter market, but there are a number of barriers to getting facilities built on schedule, such that the majority of contractors are having to extend timelines in response.

UN COP28 climate talks. As 2023 is set to become
ng is a moment of reckoning for oil and gas companies.

TECH / ARTIFICIAL INTELLIGENCE / CLIMATE

The environmental impact of the AI revolution is starting to come into focus

Security fencing around the Google Cloud data center ahead of its ceremonial opening in Hanau, Germany, on Friday, October 6th, 2023. Photo by Alex Kraus / Bloomberg via Getty Images

/ Google's use of AI alone c
use up as much electricity a
small country – but that's a
unlikely worst-case scenario
new analysis finds.

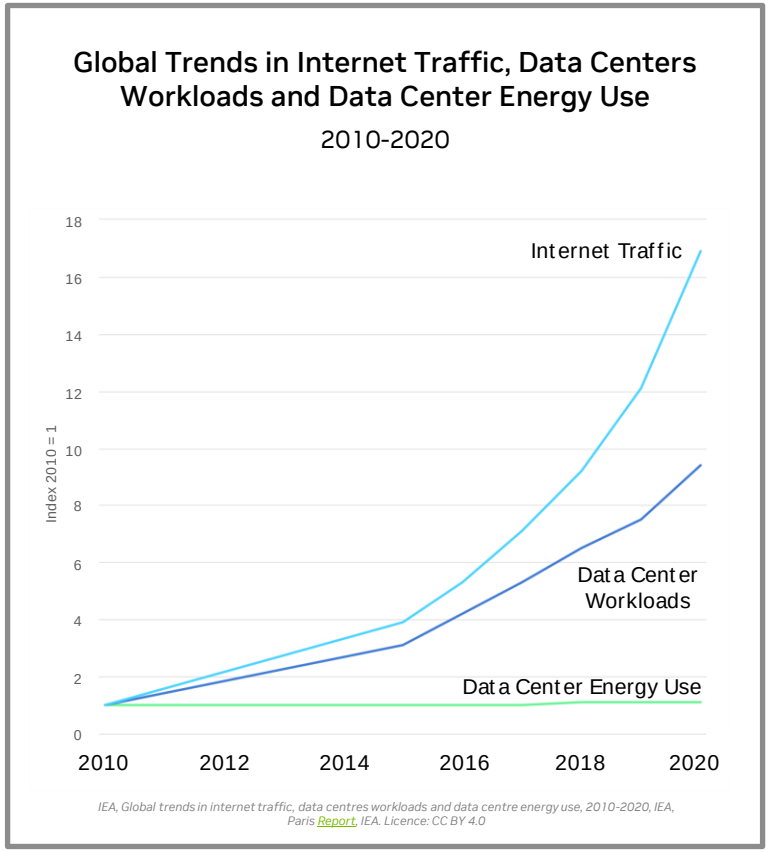
By Justine Calma, a senior science reporter covering climate change, clean energy, and environmental justice with more than a decade of experience. She is also the host of Hell or High Water: When Disaster Hits Home, a podcast from Vox Media and Audible Originals.

Oct 10, 2023, 5:00 PM GMT+2

7 Comments (7 New)

Significant Technology Trends Impacting Sustainability Efforts

Demand for Compute Grows as Data Centers are Power Limited



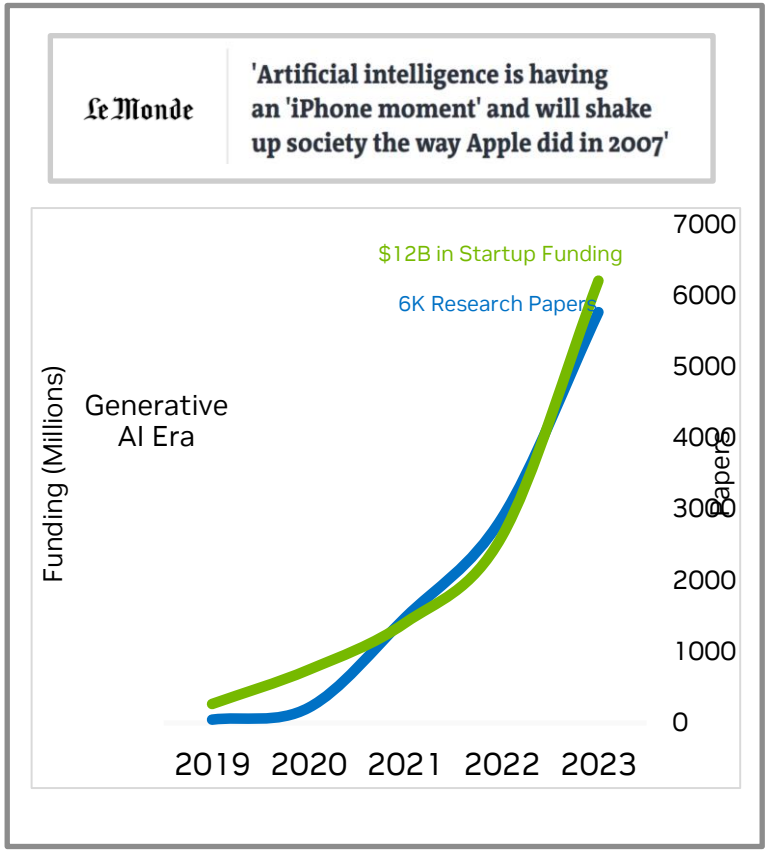
End of Moore's Law

Moore's Law Kept Data Center Energy Use Flat for Past Decade

Data Center Electricity Usage	>200 TWh/year
Share of Global Energy Usage	2%
Forecasted Share of Energy Usage	5% by 2030

Data Center Power Consumption Growing

Need to Become More Efficient



Generative AI & LLMs Fuel More Demand

Emissions Could Surge if Unchecked

Accelerated Computing is the Path Forward

"Companies are consuming ever more compute power as they embrace technologies like AI.
NVIDIA accelerated computing can meet these demands at ***lower energy consumption*** *than traditional methods." – JHH*

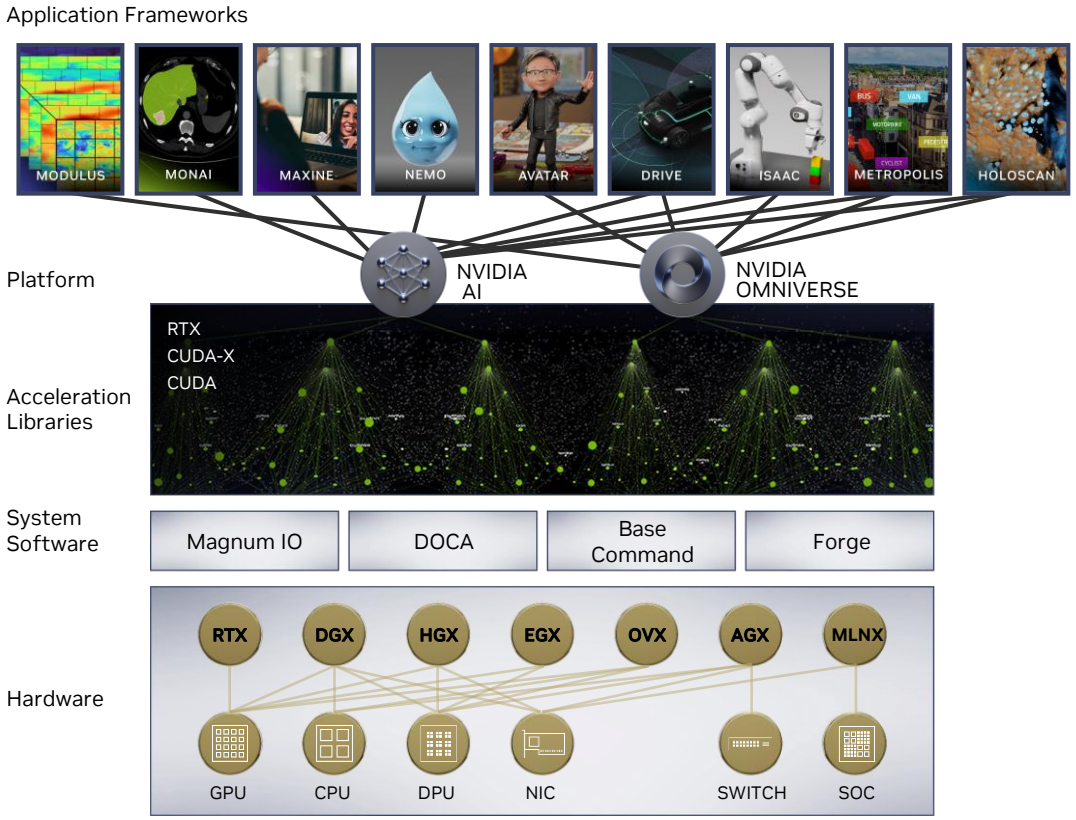
Full Stack

Data Center Scale

Cloud to Data Center to Edge

One Architecture

Million-X Speed Up Through Relentless Full Stack Invention



NVIDIA Delivers the Most Efficient Computing Platform

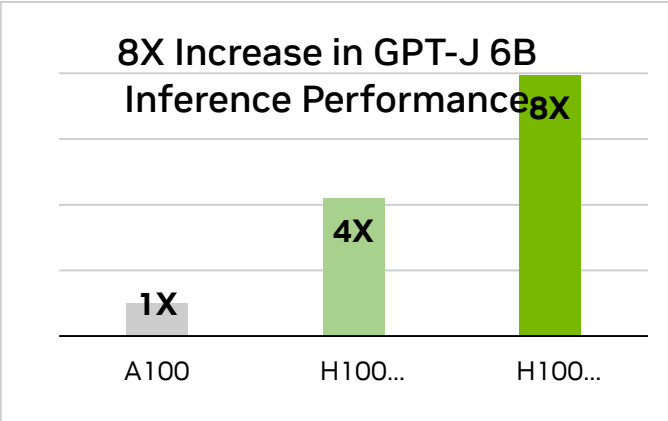
Continuous Energy Improvements Across the NVIDIA Portfolio

NVIDIA Hopper H100
Compared to Ampere A100

26X
less time to solution
for AI inference

2.7X
more energy efficient

H100 + TensorRT-LLM Software



Improving with Each Generation

Software Delivers Even Greater Performance

A100 AI Data Center
320 HGX A100

H100 AI Data Center
64 HGX A100

Equivalent AI Performance:

3X	5X	3.5X
Lower TCO	Fewer Servers	More Energy Efficient

Improving at Data Center Scale

Substantial Cost and Energy Savings Across the Data Center

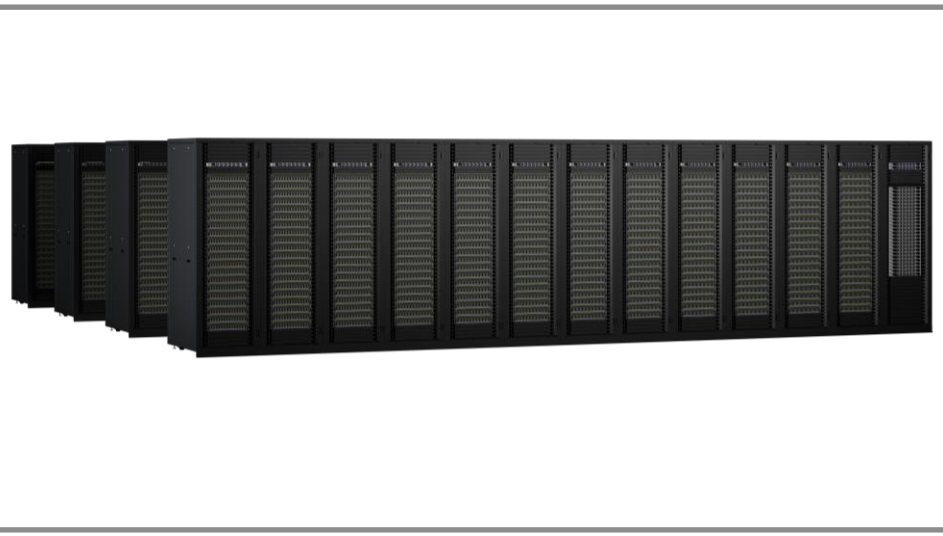
Text summarization, variable input/output length, CNN / DailyMail dataset | A100 FP16 PyTorch eager mode | H100 FP8 | H100 FP8, TensorRT-LLM, in-flight batching

Mixed AI workload of 50% training and 50% inference .
50% Large Language Models: Megatron 530B | 50% MLPerf models: computer vision (ResNet-50, Mask R-CNN, Unet-3D, RetinaNet), NLP (BERT), recommender (DLRM)
H100 data center projected workload performance, subject to change.

NVIDIA

Accelerated Computing is Efficient Computing

The Most Sustainable & Cost-Effective Way to Meet Growing Compute Needs



x86 CPU-Based Data Center for LLM Workload

Cost, Servers, & Energy Required for 1X LLM Performance

ISO-Budget

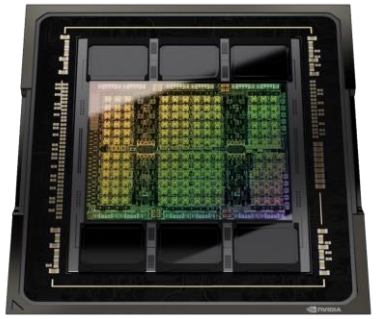
1/4 the Energy & 44X More Performance at the Same Cost

ISO-Performance

96% Less Cost & 84X More Energy Efficiency at the Same Performance

Accelerated Computing Runs on Three Foundational Elements

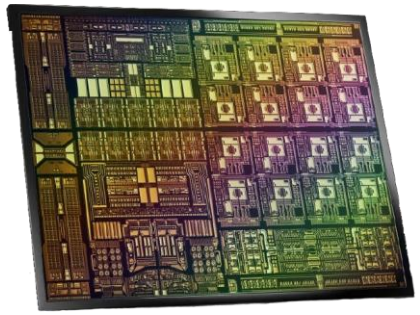
The GPU, DPU, and CPU Each Play a Critical Role in Delivering Energy Efficiency



Hopper H100
World's Most Advanced Chip

86x
more energy efficient

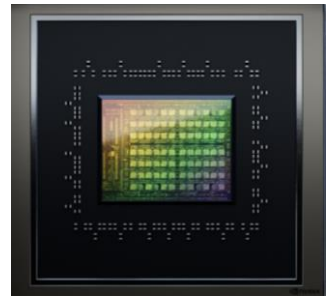
GPUs for Compute-Intensive Functions



BlueField-3
Accelerator for Networking, Storage, & Security

30%
reduce server power consumption

DPUs for Infrastructure Offloading

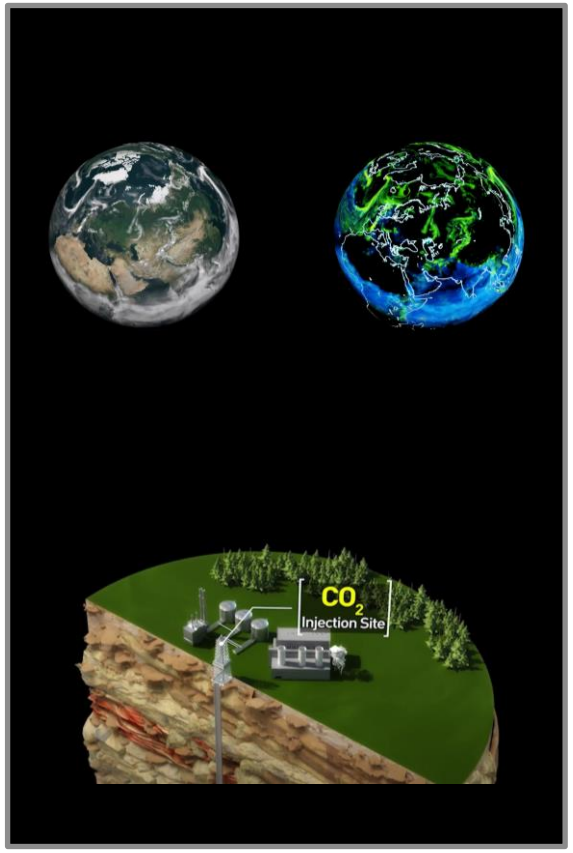


Grace CPU
High Performance CPU for HPC and AI

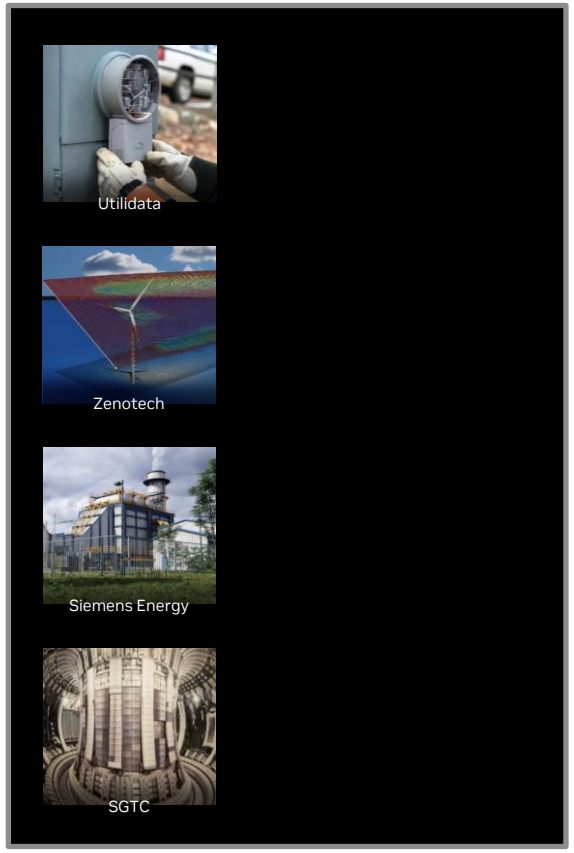
2x
up to 2X better energy efficiency

CPUs for Orchestration Management

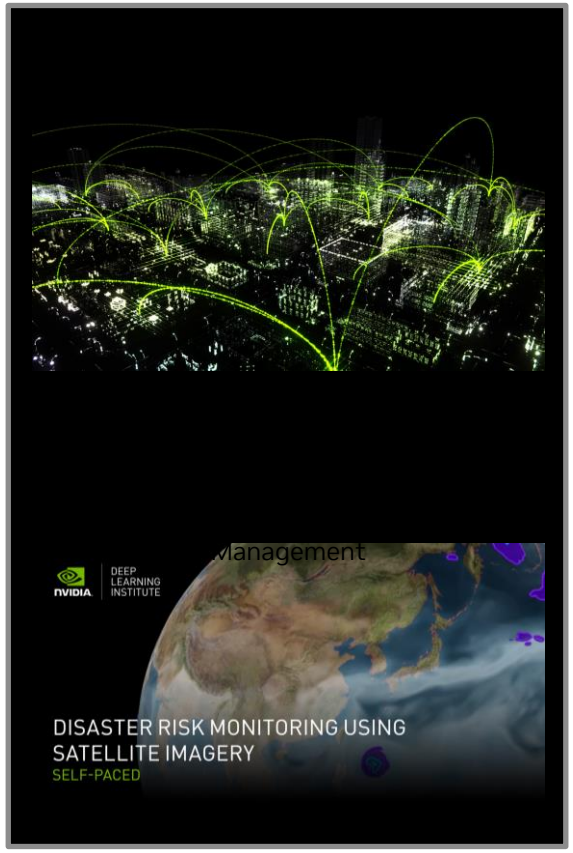
Applications of NVIDIA Technology for Climate Change Mitigation & Adaptation



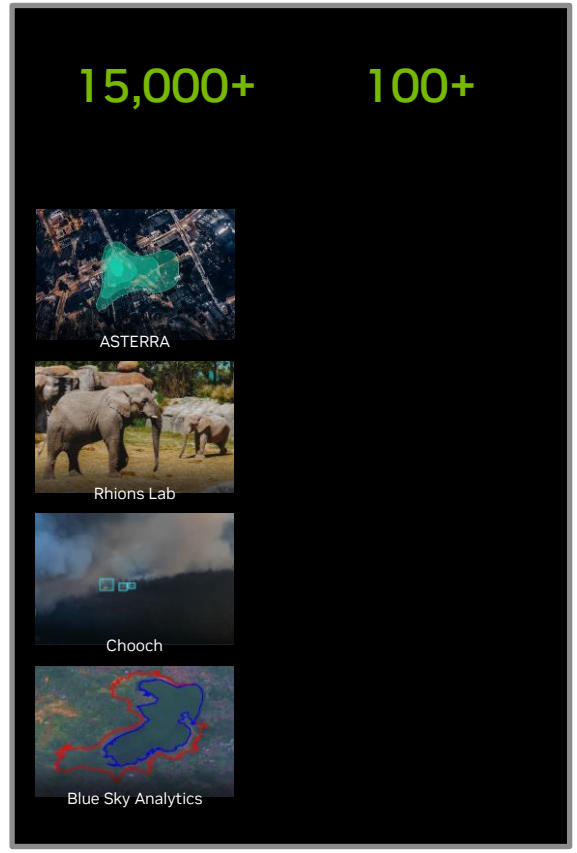
Climate Science Research



Energy Industry Innovation



Global Climate Strategies

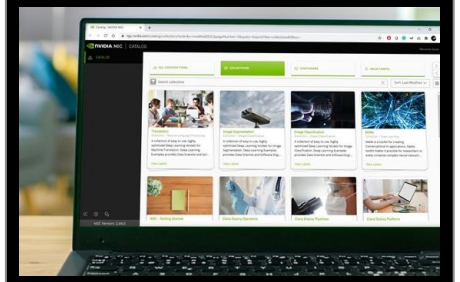


NVIDIA Inception Startups

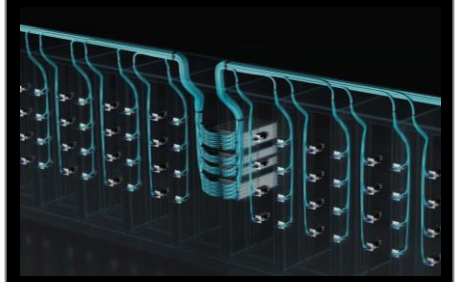
Modern, Energy-Efficient Supercomputers Run on the NVIDIA Platform

Hardware, Software, & Networking to Optimize Performance & Efficiency

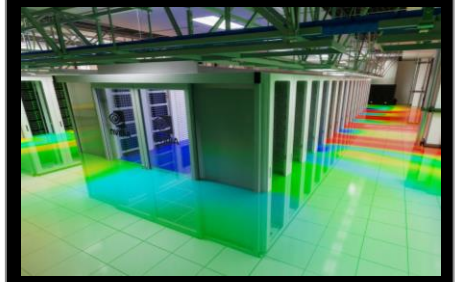
Less Rack Space & Energy



Software



Networking



Cooling

NVIDIA Sustainability Overview

NVIDIA GPUs are typically **20X more energy efficient** for certain AI and HPC workloads than traditional CPUs

Earth-2 – Building the world's most powerful AI supercomputer dedicated to predicting climate change

23 of Top 30 Supercomputers on the June 2023 Green500 are powered by NVIDIA including the #1 system, Henri

Impact of Our Technology

NVIDIA's two HQ campuses have received **LEED Gold** status

Will achieve & maintain **100% renewable electricity** for our operations & data centers by FY25 & annually thereafter

By FY26, **engage manufacturing suppliers** comprising at least 67% of our scope 3 cat. 1 GHG emissions with a goal of effecting supplier adoption of science-based targets

Initiatives Across Operations

100 Most Sustainable Companies
Barron's, 2023

S&P Global Sustainability Yearbook 2023
S&P Global, 2023

America's Most Responsible Companies
Newsweek, 2023

JUST 100 Companies
JUST Capital & CNBC, 2023

Public Recognition of Our Impact

Meta's Llama 2 models carbon footprint

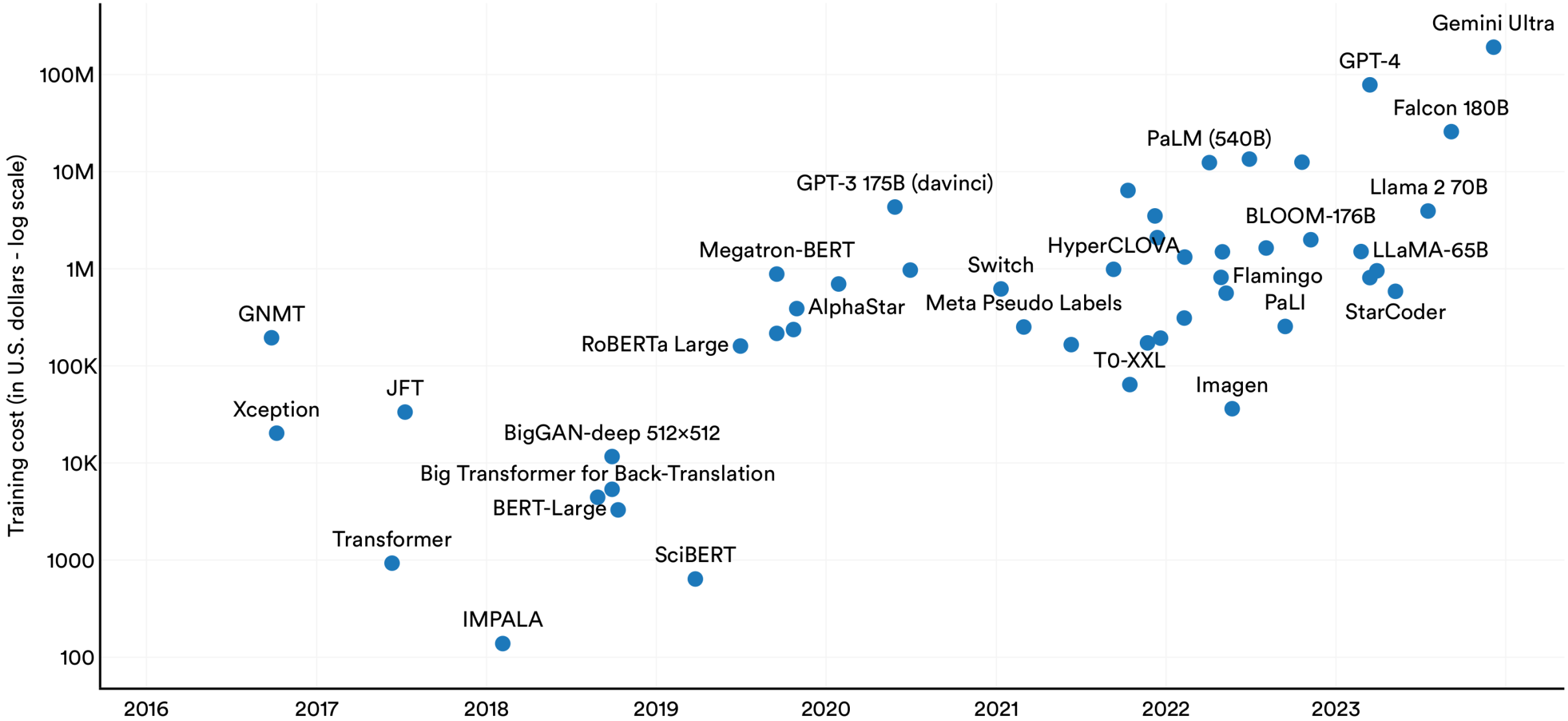
		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Llama 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

„Training Hardware. We pretrained our models on Meta's Research Super Cluster (RSC) (Lee and Sengupta, 2022) as well as internal production clusters. Both clusters use NVIDIA A100s. There are two key differences between the two clusters, with the first being the type of interconnect available: RSC uses NVIDIA Quantum InfiniBand while our production cluster is equipped with a RoCE (RDMA over converged Ethernet) solution based on commodity ethernet Switches. Both of these solutions interconnect 200 Gbps end-points. The second difference is the per-GPU power consumption cap — RSC uses 400W while our production cluster uses 350W. With this two-cluster setup, we were able to compare the suitability of these different types of interconnect for large scale training. RoCE (which is a more affordable, commercial interconnect network)"

Source: Meta, Llama 2 Whitepaper

Estimated training cost of select AI models, 2016–23

Source: Epoch, 2023 | Chart: 2024 AI Index report



Source: Stanford Univesity, AI Index Report, 2024

NVIDIA

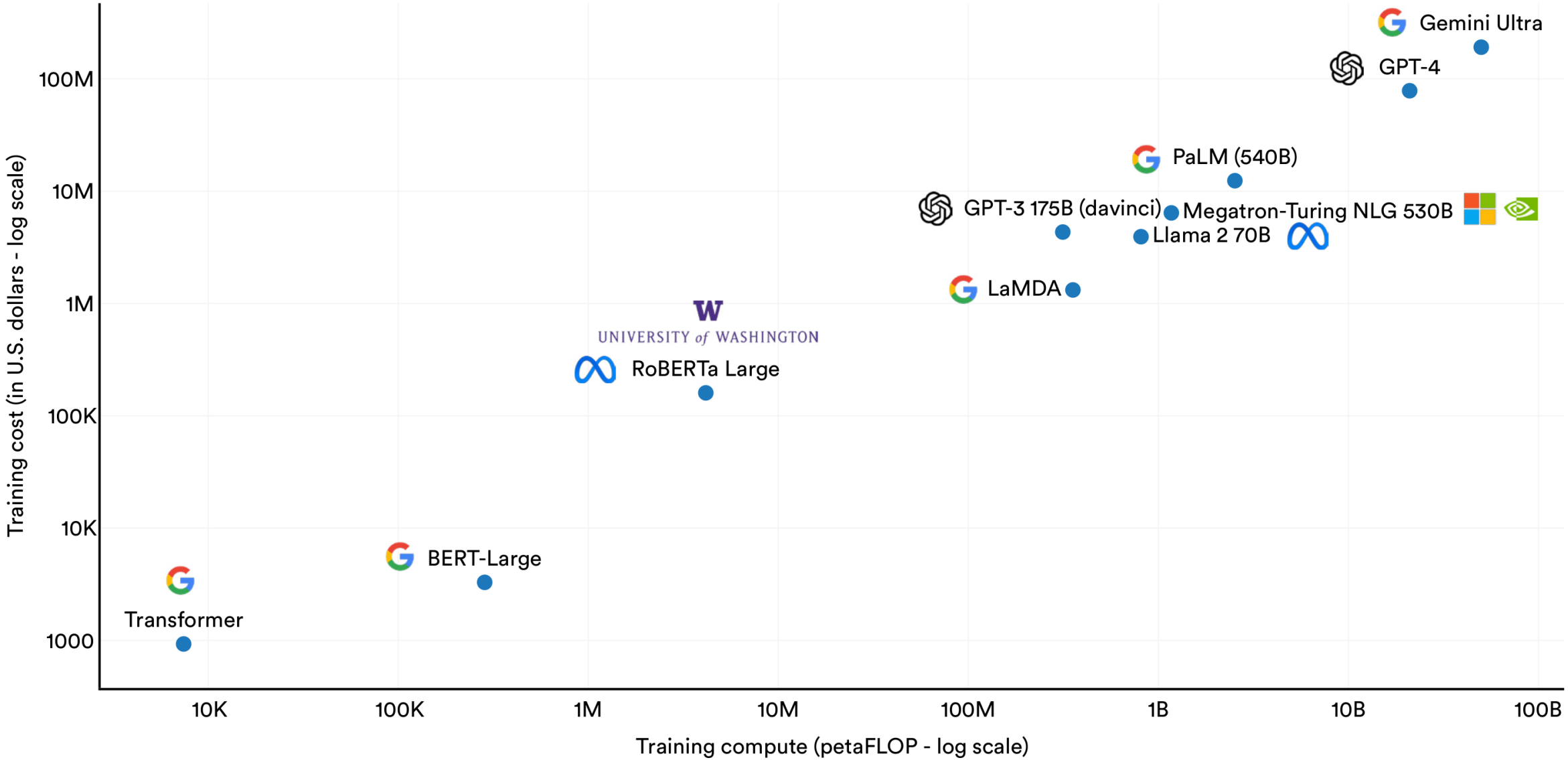
Estimated training cost – table view

Model	Year	Training cost in US-Dollar
Transformer	2017	930 US$
BERT-Large	2018	3.288 US$
RoBERTa Large	2019	160.018 US$
GPT-3 175B	2020	4.324.883 US$
Megatron-Turing NLG 530B	2021	6.405.653 US$
LaMDA	2022	1.319.586 US$
PaLM 540B	2022	12.389.056 US$
GPT-4	2023	78.352.034 US$
Llama 2 70B	2023	3.931.897 US$
Gemini Ultra	2023	191.400.000 US$

Source: Stanford Univesity, AI Index Report, 2024

Estimated training cost and compute of select AI models

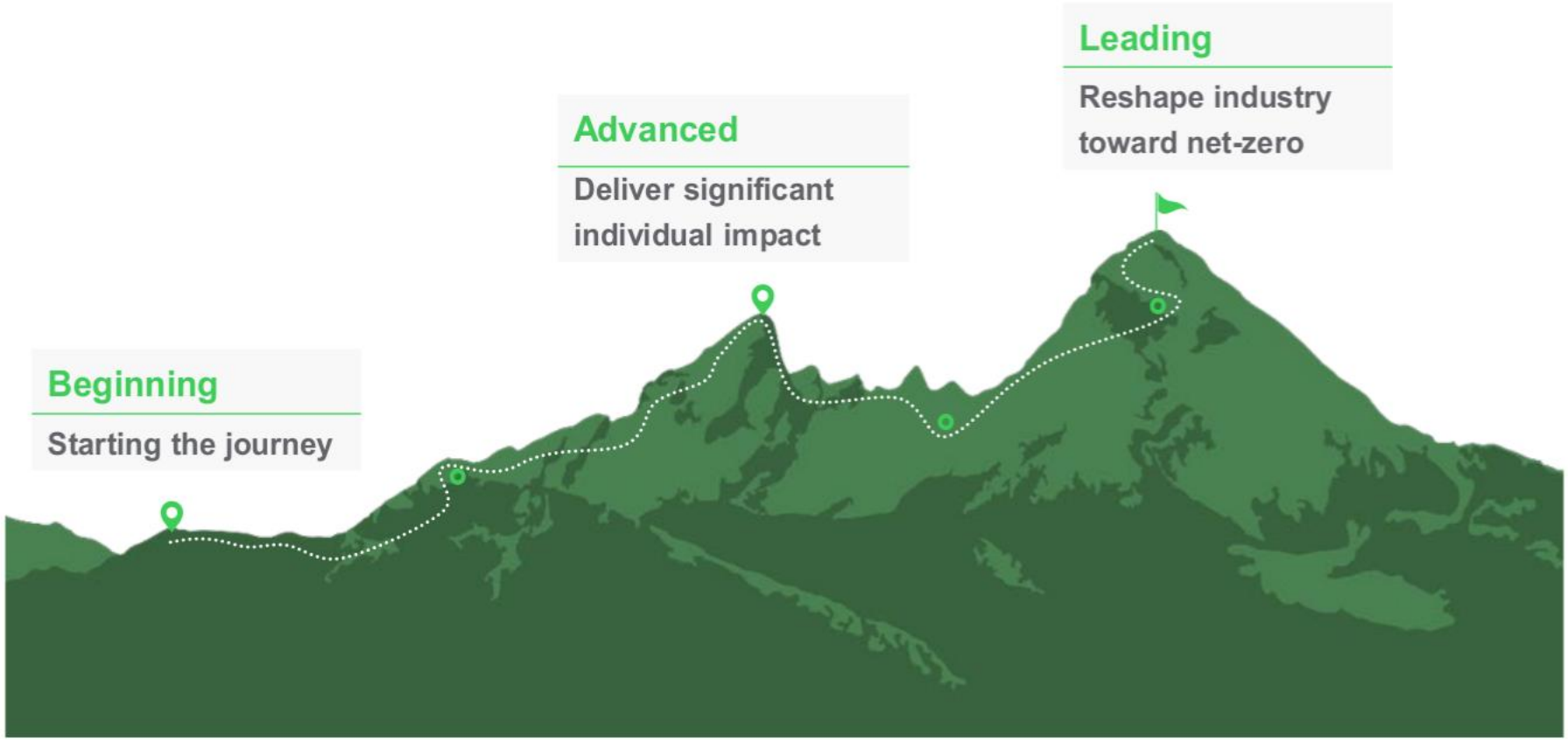
Source: Epoch, 2023 | Chart: 2024 AI Index report



Source: Stanford Univesity, AI Index Report, 2024

Guide to Environmental Sustainability Metrics for Data Centers

Schneider Electric, White Paper 67 Version 2



The World Business Council for Sustainable Development (WBCSD) identified three stages in the journey towards net-zero

Guide to Environmental Sustainability Metrics for Data Centers

Schneider Electric, White Paper 67 Version 2

Metric categories	Key metrics	Units	Recommendations		
			Beginning (6)	Advanced (18)	Leading (28)
Energy (6)	• Total energy consumption	kWh	✓	✓	✓
	• Power usage effectiveness (PUE)	Ratio	✓	✓	✓
	• Total renewable energy consumption	kWh		✓	✓
	• Renewable energy factor (REF)	Ratio			✓
	• Energy Reuse Factor (ERF)	Ratio			✓
	• Server utilization ($ITEU_{sv}$)	%		✓	✓
GHG emissions (7)	• Scope 1				
	○ GHG emissions	$mtCO_2e$	✓	✓	✓
	• Scope 2				
	○ Location-based GHG emissions	$mtCO_2e$	✓	✓	✓
	○ Market-based GHG emissions	$mtCO_2e$	✓	✓	✓
	• Scope 3				
	○ GHG emissions	$mtCO_2e$			✓
	• Carbon usage effectiveness (CUE)	kg CO_2e/kWh		✓	✓
	• Total carbon offsets	$mtCO_2e$		✓	✓
	• Hourly renewable supply & consumption matching	%			✓

Guide to Environmental Sustainability Metrics for Data Centers

Schneider Electric, White Paper 67 Version 2

Metric categories	Key metrics	Units	Recommendations		
			Beginning (6)	Advanced (18)	Leading (28)
Water (5)	• Total site water usage	m^3	✓	✓	✓
	• Total source energy water usage	m^3			✓
	• Water usage effectiveness (WUE)	m^3/MWh		✓	✓
	• Water replenishment	m^3			✓
	• Total water use in supply chain	m^3			✓
Waste (6)	• Waste generated				
	○ Total waste	Metric ton			✓
	○ E-waste	Metric ton		✓	✓
	○ Battery	Metric ton		✓	✓
	• Waste diversion rate				
	○ Total waste	Ratio			✓
	○ E-waste	Ratio		✓	✓
	○ Battery	Ratio		✓	✓
Local ecosystem (4)	• Land				
	○ Total land use	m^2		✓	✓
	○ Land-use intensity	kW/m^2		✓	✓
	• Outdoor noise	dB(A)		✓	✓
	• Mean species abundance (MSA)	MSA/km^2			✓

Guide to Environmental Sustainability Metrics for Data Centers

Schneider Electric, White Paper 67 Version 2

Key metric	Defined by	Best-in-class value	Industry target value
PUE	ISO/IEC 30134-2	1.1 (75%-85% load ratio)	1.2-1.3 (75%-85% load ratio)
REF	ISO/IEC 30134-3	1.0	0.75-1.0
CUE_2	ISO/IEC 30134-8	0.0 kg CO_2e/kWh	0.0-0.12 kg CO_2e/kWh
WUE_1	ISO/IEC 30134-9	0.0 m^3/MWh	0.3-0.45 m^3/MWh

- **Power usage effectiveness (PUE)** - The ratio of a data center's total energy consumption to IT energy consumption
- **Renewable energy factor (REF)** - The ratio of renewable energy owned and controlled by a data center organization to the data center's total energy consumption
- **Carbon usage effectiveness (CUE_2)** - The sum of data center annual Scope 1 and market-based Scope 2 carbon emissions divided by the IT energy consumption with the unit of kg CO_2e/kWh.
- **Water usage effectiveness (WUE_1)** - The onsite data center water consumption divided by the IT energy consumption with the unit of m^3/MWh.


NVIDIA
THANK YOU