AI FACTORIES_

HOW DATA CENTERS CONVERT ELECTRONS TO TOKENS_

Every AI response is a product manufactured at a token factory – a data center. But data centers aren't just buildings; they are advanced machines that use electrical energy to power IT equipment that produces tokens and generates a lot of heat in the process. Understanding how a data center works is analogous to understanding the global supply chain that underpins the AI boom. Join us as we dissect the most powerful factories humans ever built!

ENERGY: THE FOUNDATION

"You can't create an industry without energy." - Jensen Huang, NVIDIA CEO. Every token begins with an electron. OpenAI's Stargate data center in Abilene, Texas, will draw 1 gigawatt of power, enough to power the entire city of Seattle, Washington. Below: every bottleneck between the grid and the chip.

TRANSMISSION LINES

Transmission lines bring in electricity at very high voltages (115-500kV). Power grid connections can take 3-5 years to be approved.

HIGH/MEDIUM VOLTAGE (HV/MV) TRANSFORMERS

HV transformers step the grid voltage down to 20kV (lead time: 3–5 yrs); MV transformers closer to the data halls step voltage down further to ~400V (lead time: 6–17 mos).

MEDIUM VOLTAGE SWITCHGEAR

Routes power to individual data halls and includes circuit breakers. Lead time: 3–4 mos.

AUTO TRANSFER SWITCH

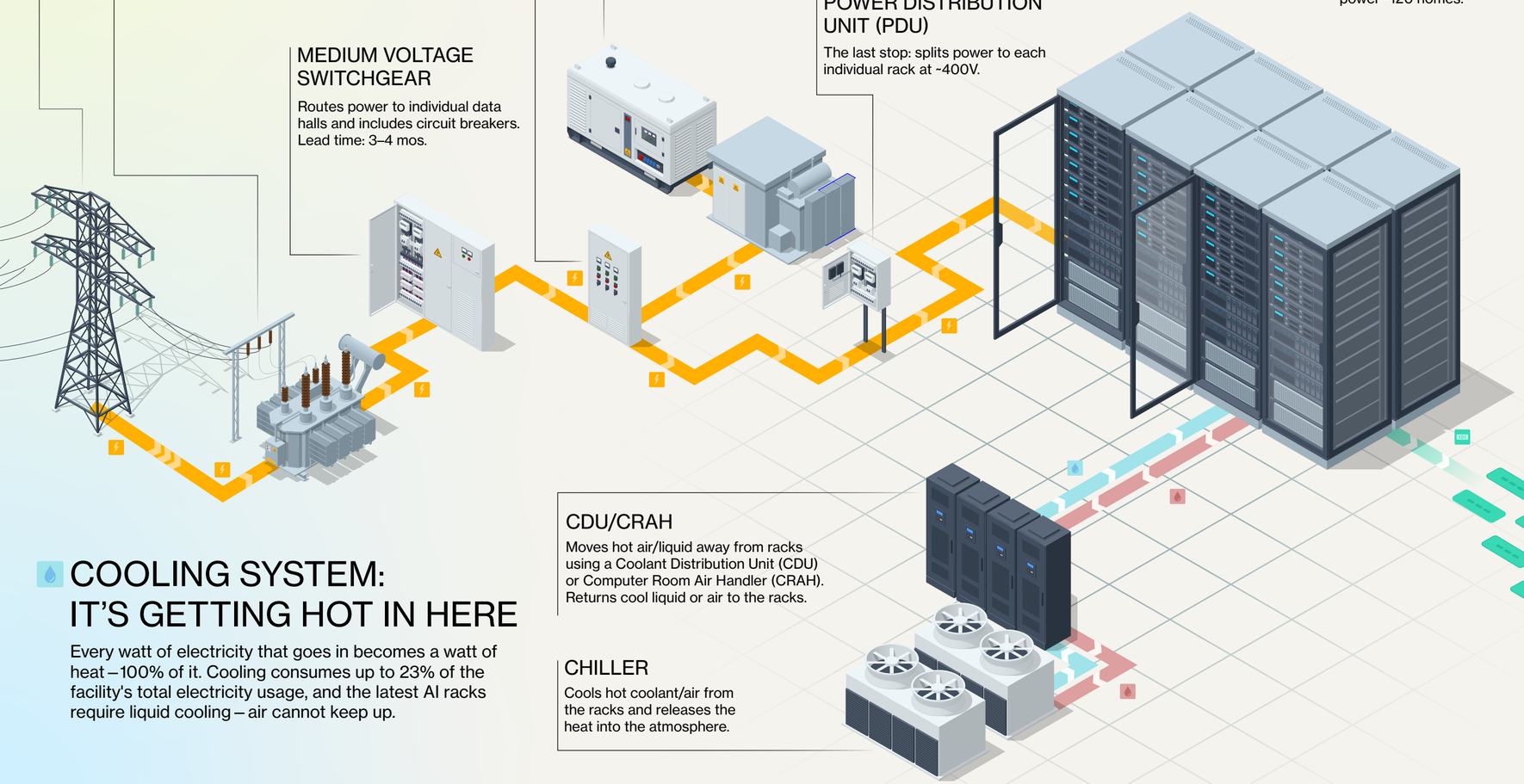
Grid failure? Switches to backup power in <20 milliseconds. Lead time: 12-18 mos.

GENERATORS + UNINTERRUPTIBLE POWER SUPPLY (UPS)

UPS bridges the power gap during a grid outage and smooths out fluctuations while generators kick in ~90s. UPS lead time: 2-3 mos. Generator lead time: 12-24 mos.

POWER DISTRIBUTION UNIT (PDU)

The last stop: splits power to each individual rack at ~400V.



COOLING SYSTEM: IT'S GETTING HOT IN HERE

Every watt of electricity that goes in becomes a watt of heat – 100% of it. Cooling consumes up to 23% of the facility's total electricity usage, and the latest AI racks require liquid cooling – air cannot keep up.

CDU/CRAH

Moves hot air/liquid away from racks using a Coolant Distribution Unit (CDU) or Computer Room Air Handler (CRAH). Returns cool liquid or air to the racks.

CHILLER

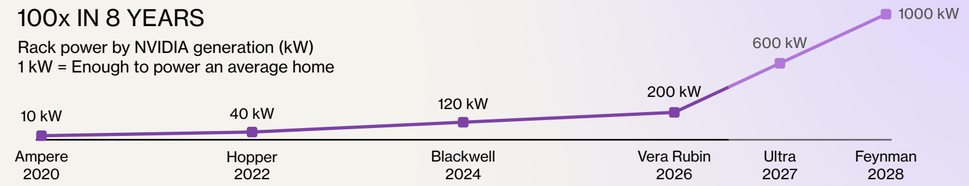
Cools hot coolant/air from the racks and releases the heat into the atmosphere.

DATA HALLS: WHERE COMPUTE HAPPENS

"Compute" means mathematical work – trillions of operations per second run by GPUs and CPUs. AI workloads require 10-100x more compute per rack than traditional servers running search and streaming.

100x IN 8 YEARS

Rack power by NVIDIA generation (kW)
1 kW = Enough to power an average home



RACKS

Cabinets of servers that can already pull up to 120 kW today – enough to power ~120 homes.

SERVERS

Individual trays in each rack, packed with GPUs, CPUs, memory and networking hardware.

CHIPS

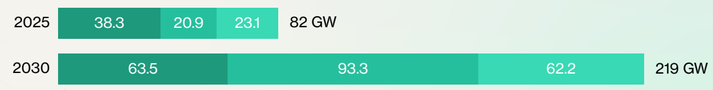
GPUs and CPUs doing the math at ~1V – where the electrons turn into tokens and heat.

TOKENS: THE PRIMARY OUTPUTS

Just as oil output is measured in barrels, AI output is measured in tokens. Tokens are the new industrial commodity sold by the millions, billed in cents, and generating real revenue. Demand could more than triple by 2030.

GLOBAL DATA CENTER DEMAND

Actual & forecast, GW by workload



NON-AI

Search, streaming, online software, gaming. Still 47% of today's capacity.

AI TRAINING

Teaching the model. The cost to train fontier LLM models has gone from ~$2M for GPT-3 to nearly $390M for the latest models. That's a lot of tokens.

AI INFERENCE

Using the model. Every AI response requires inference and creates recurring token usage for every user's query, every day. Inference is the fastest-growing workload.

WHAT IS A TOKEN?

A word, a pixel, an action. Every prompt in and response out is a stream of tokens.
A fundamental unit measuring everything an AI model reads and writes.
Tokens are priced and metered like electricity.
More gigawatts = more tokens = more revenue.

`The` `quick` `brown` `fox` = 4 tokens.

LAND USAGE

With all the electrical, computing, and cooling equipment, data centers can take up a lot of space. Stargate will take up as much space as 450 soccer fields.

WATER USAGE

Data centers can "drink" up to 334 million gallons per year, primarily to stay cool (enough to fill a little over 500 Olympic-sized pools). Operators are chasing the cold in cooler climates, underwater, and even in...space!

CRITICAL MINERALS USAGE

AI chips in data centers are built from imported minerals – gallium, germanium, indium, and tantalum, to name a few. The U.S. imports 100% of many of these.

Sources: Semianalysis, McKinsey, Epoch AI, NVIDIA, Microsoft, Google, BloombergNEF, Payden Calculations