

Cloud, Compute and Sovereignty: A Snapshot of an Industry at a Breaking Point

A retrospective and current assessment of cloud, silicon and digital sovereignty.

Nicolas Masselot
June 2026

Table of Contents

Table of Contents	2
Brief Objectives	4
Useful Vocabulary (underlined in the text)	5
Part I: The Economics of the Cloud and the Revolt Against the Hyperscalers	8
Egress Fees: the Lock-In Mechanism	8
The Rise of FinOps, the BYOC Alternative	8
The Repatriation Debate (Bare Metal)	8
From the Cloud of Abundance to the Cloud of Scarcity	9
Part II: Under the Hood of AI: Memory, Chips and Interconnects	10
The Memory Wall and Bandwidth	10
HBM: Bringing Memory and Compute Closer Together	10
Advanced Packaging: the Industry's New Lock	11
Interconnect: Scale-up versus Scale-out	12
The Software Stack and Compilers: the Real Divide	13
The End of Moore's Law: from Transistor to System	13
Part III: NVIDIA's Hegemony and Saga	15
The CUDA Moat: Why NVIDIA is a Software Company	15
From Hopper to Blackwell: from Component to System	15
The Rack as Product: GB200 NVL72 then Vera Rubin	15
DGX Cloud and Verticalisation: NVIDIA against its Own Customers	16
The Omniverse Bet and Embodied AI (GR00T)	16
Part IV: NVIDIA's Challengers	17
AMD and the Software Ball and Chain	17
The Hyperscalers and Verticalisation	17
Family 1: Replacing the GPU with another architecture	17
Family 2: The Inference Specialists	18
Family 3: The Rare Training Challengers	18
Family 4: Edge & Physical AI: a different market	18
A Platform, and Therefore a Fragmented Target	19
Part V: The Geopolitics of Silicon	20
ASML and Zeiss: the Monopoly on Etching Machines	20
American Equipment Makers: the Real Lock	20

TSMC: the Irreplaceable Foundry and the Taiwan Risk.....	20
Intel Foundry Services: the Comeback Bet	21
Samsung: the Ambivalent Giant of Memory and Foundry	21
The American Legislative Architecture and the Containment of China.....	21
The Chinese Response: War over Minerals and the Strategy of Quantity	22
Part VI: The Race for Foundation Models and the Collapse of Margins.....	23
From Training to Inference: The Great Shift	23
The DeepSeek Shock: When Algorithmic Efficiency Challenges Brute Force	23
The Collapse of Margins and the Profitability Crisis of the American Laboratories	23
Part VII: Infrastructure and the Energy Crisis	25
The Giants' Unprecedented Vertical Integration.....	25
The Neoclouds and the Risk of GPU Debt	25
xAI, Colossus and the Philosophy of Brute Force	25
Terafab: Vertical Integration Taken to the Extreme	25
Energy, the New Limit of Compute, and the Nuclear Renaissance.....	26
France and the Nuclear Card: Compute Through the Prism of Gigawatts.....	26
Part VIII: The European Awakening.....	27
The Legacy of Failures: From Andromède to the Gaia-X Illusion.....	27
The Hard Sovereignists Against the « Trusted Cloud »	27
The Mistral AI Case: Sovereignty and Pragmatism	27
InvestAI, AION and the Nuclear Card: Europe in 2026	28
Part IX: The Macro-Strategic Shock and the Global Legal Vice	29
The CLOUD Act Affair.....	29
The « Liberation Day » Shock and the Transatlantic Rupture	29
The Regulatory Tempo: Ambitions and Realities for the European Ecosystem	29
Conclusion: Infrastructure as Destiny.....	30

Brief Objectives

This document offers a retrospective and current assessment of the cloud and AI infrastructure sector, at the global, European and French levels. It was built as a personal knowledge notebook, intended to centralise my understanding of the field in a clear and evolving format, and it will be updated regularly to follow current developments.

It expanded as it was written. Starting from the economics of the cloud, it descends progressively towards the hardware layers, silicon, memory and energy, then rises back towards the questions of sovereignty. I have tried to make it accessible to a non-technical reader as well as to someone in the field, by setting out each concept before using it. This goal is not always fully achieved: the subject is so interconnected that some terms appear for the first time before being properly explained later on. When that happens, looking back or consulting the opening glossary helps to resolve the ambiguity.

The document does not claim to be exhaustive and is not a professional deliverable: it is not sourced academically and some basic concepts are not redefined. Its aim is above all to structure a personal reflection on the underlying dynamics of the sector.

Useful Vocabulary (underlined in the text)

Bare Metal : a physical server dedicated to a single customer, with no intermediate virtualisation layer. The customer has direct access to the machine's hardware resources, which generally improves performance and predictability.

Virtual machine (VM) and hypervisor: a virtual machine is a software-simulated machine running on a physical server shared among several customers. The hypervisor is the software that divides the physical server into several independent VMs.

Noisy neighbor : a phenomenon in which several customers sharing the same physical server through virtualisation see their performance degraded by one customer that consumes the shared resources excessively. It is a drawback of cloud resource pooling.

AI accelerator: a chip specialised in the training or inference of artificial intelligence models. Unlike general-purpose processors (CPUs), accelerators are optimised for the massively parallel computations required by neural networks. NVIDIA's GPUs are today the most widespread accelerators.

FLOPS : a unit measuring the number of mathematical operations that a processor or accelerator can perform each second.

Inference is the phase in which an already trained AI model is used. The model draws on its parameters to generate a response.

Lithography machine: equipment used to etch a chip's circuits onto a silicon wafer. These machines project microscopic patterns that determine the location of the billions of transistors making up modern processors. The most advanced machines, known as EUV (Extreme Ultraviolet), are produced exclusively by the Dutch group ASML.

Rack-scale AI refers to the new approach, popularised by NVIDIA with its GB200 systems, which consists of designing and selling the compute unit no longer as an individual chip or server, but as an entire preassembled cabinet (rack) integrating dozens of GPUs, their memory, their interconnects, their power supply and their liquid cooling.

CUDA (Compute Unified Device Architecture) is the proprietary software platform developed by NVIDIA since 2006 to let researchers and developers use its GPUs for computations other than graphics. Its programming language translates code into instructions executable on NVIDIA chips.

Foundation model: a large AI model trained on massive data and designed to be subsequently adapted to multiple downstream tasks (assistance, code, analysis, translation).

Project Rainier is the project by Amazon AWS and Anthropic, deployed in Indiana, bringing together more than 500,000 of Amazon's Trainium2 chips and backed by a 100 billion dollar AWS commitment over ten years.

Qualcomm is an American giant in chips for smartphones and cellular modems, which designs its chips without manufacturing them itself (Fabless). Drawing on its energy-efficiency expertise

inherited from mobile, it has recently been trying to enter the AI inference market in the data centre.

A Transformer is the type of neural-network architecture on which today's large language models, such as ChatGPT or Claude, are based. Introduced by Google in 2017, its key principle is the attention mechanism, which lets the model weigh the importance of each word relative to all the others in a text, and thus grasp context.

The CHIPS Act is a US law passed in 2022 to reshore semiconductors, providing massive subsidies to foundries that build plants on American soil. Its main beneficiaries are TSMC (for its Arizona plants), Intel, Micron and Samsung.

The 18A is the name of Intel's most advanced process node, the one meant to bring it back to TSMC's level.

The scaling law refers to the empirical law whereby a language model's performance grows predictably with the size of the model, the size of the dataset and the amount of compute invested in training. This law underpinned the entire investment strategy of the industry between 2020 and 2024: stacking up more parameters, more data, more GPUs.

Open weights: refers to a model whose « weights », that is, the billions of numbers that define the model's behaviour, are published and downloadable. Llama (Meta), Mistral and DeepSeek are the most striking examples of this approach, which stands in contrast to the purely proprietary models of OpenAI or Anthropic, accessible only via API.

SMR (Small Modular Reactor): a small nuclear reactor (50 to 300 MW) mass-produced in a factory and then assembled on site, aimed at drastically reducing costs and timelines compared with giant power plants.

SecNumCloud is a security qualification issued by the French cybersecurity agency. It imposes strict criteria such as immunity from extraterritorial laws (CLOUD Act), data storage in Europe and European ownership control. It is intended to protect French « critical » data. About ten players are qualified to date, including OVHcloud, Outscale (Dassault Systèmes), Cloud Temple and S3NS, while others such as NumSpot or the Bleu offering are still in the process of qualification.

The AI Act is a European regulation adopted in 2024 that classifies AI systems according to their level of risk and imposes increasing obligations (transparency, audit, certification) according to that level. Its application has been progressively pushed back under industry pressure, with some of the main obligations not due to take effect until December 2027.

Kyutai is a French non-profit AI research laboratory. Its ambition is « open science »: publishing its research freely, without pressure for immediate profitability. It embodies the bet that Europe can produce fundamental AI research outside American proprietary logics.

The InvestAI Facility is the European Union's financial mechanism announced in 2025, endowed with 20 billion euros, aimed at financing up to five AI gigafactories on the continent by 2030.

SaaS, PaaS, IaaS: three levels of abstraction of cloud computing.

- **With IaaS (Infrastructure as a Service), the provider rents raw infrastructure (virtual servers, storage, network) and the customer manages its own operating system, databases and applications.**
- **With PaaS (Platform as a Service), the provider delivers a platform ready to run code (managed database, managed kubernetes), and the customer only has to deploy its application.**
- **With SaaS (Software as a Service), the provider delivers a finished application that can be used directly (Salesforce, Notion, Microsoft 365).**

Container: a technology that packages an application with everything it needs to run, so that it behaves identically on any machine. Much lighter and faster than a virtual machine, the container has become the standard format for deploying modern applications.

Kubernetes: software that automatically manages hundreds or thousands of containers in production.

Part I: The Economics of the Cloud and the Revolt Against the Hyperscalers

Originally, the cloud was first presented as a flexible and economical model: companies could rent computing power without investing in servers or data centres themselves. This made it possible to substitute flexible Opex for the heavy Capex required to set up servers.

But as the cloud became widespread, the hyperscalers progressively built pricing models that reinforce customer dependence and maximise their margins.

Egress Fees: the Lock-In Mechanism

One of the main lock-in mechanisms of the cloud rests on « egress fees », the charges billed when a customer wants to retrieve its data out of the cloud.

Hyperscalers such as AWS make data entry (ingress) free, in order to draw companies onto their infrastructure. Transferring that data to the Internet or to another cloud (egress), by contrast, becomes costly.

This asymmetry creates a strong dependence: the more data a company stores with a hyperscaler, the more expensive and complex it becomes to leave.

Players such as Cloudflare have tried to call out this rent by offering zero egress fees, but the real coup de grâce came from regulation. In force since 2024, the European Data Act mandated the free movement of data, forcing Google Cloud and then AWS to abolish exit fees for customers wishing to close their accounts and migrate their workloads.

The Rise of FinOps, the BYOC Alternative

The deliberate complexity of cloud bills (billing files containing millions of lines) gave rise to the discipline of FinOps: companies pay consultants and dedicated software simply to understand what they are consuming. It is an admission of a failure of alignment between provider and customer.

Faced with this growing dependence, the BYOC (Bring Your Own Cloud) model has developed. Companies such as Snowflake or Databricks now let their customers use their tools directly within their own cloud infrastructure (AWS, Azure, Google Cloud), without moving the data out to the provider's servers. This makes it possible to keep more control over the data and to limit egress fees.

The Repatriation Debate (Bare Metal)

For years, leaving the public cloud to return to physical servers was seen as a step backwards. Several companies have nonetheless shown that this choice could be more rational at scale. Repatriation can take two forms: renting dedicated servers in **bare metal** from a player such as

OVHcloud, or buying back one's own servers and operating them in-house.

The main advantage of the public cloud is flexibility: companies can instantly increase or reduce their resources and delegate all management of the infrastructure. But this flexibility comes at a high cost. For companies with large and relatively stable compute needs, running their own servers directly often works out far cheaper than continuously renting resources from AWS, Azure or Google Cloud.

The return to bare metal also improves performance. By removing the **virtualisation layers specific to the public cloud**, **applications gain speed and no longer suffer the slowdowns linked to the sharing of resources with other customers**.

From the Cloud of Abundance to the Cloud of Scarcity

For twenty years, the cloud sold an abundant and largely interchangeable resource: conventional computing power and storage, available on demand. The balance of power was commercial: the provider held on to its customer through exit fees, pricing complexity and the lock-in described above.

The AI wave upended the business of cloud players. Training and running a model requires specialised processors, GPUs, which are rented from those same players, by the hour or by the minute, exactly as one used to rent a web server. The business is still remote capacity rental. But a new aisle has opened, and it has been stormed.

The resource that now matters is no longer generic compute but the accelerator, the GPU, and it is scarce. The investments are colossal, since the four American cloud giants, the hyperscalers, are planning 600 billion dollars of infrastructure spending in 2026, after roughly 390 billion a year earlier.

Worse, the hyperscalers are no longer constrained by demand but by their own supply capacity. Microsoft is accumulating around 80 billion dollars of Azure orders it cannot fulfil, with GPUs sitting idle in stock for want of electricity to power them. Sundar Pichai, for his part, admitted that Google's cloud revenue would have been higher if it had been able to meet demand.

This shift changes who holds power. As long as the resource was abundant, the advantage went to whoever controlled the commercial relationship, through exit fees and lock-in. Now that it is scarce, it goes to whoever controls the hardware: the chips and the energy to run them. What decides the balance of power, from now on, is access to silicon.

Part II: Under the Hood of AI: Memory, Chips and Interconnects

This part lays out the technical grammar of modern AI infrastructure. From the memory wall to compilers, it sets out the physical and software constraints that today determine the real performance of AI chips.

The Memory Wall and Bandwidth

The supreme criterion, when talking about GPUs, has often been compute power: how many operations per second. The reality is that a chip often spends its time waiting for data rather than computing. This is the « memory wall »: the compute units have become so fast that the limiting factor is no longer computation, but the speed at which data can be delivered to them from memory.

When the chip is constrained by memory bandwidth, it is said to be memory-bound; when it is constrained by computation, we speak of compute-bound. A chip can post spectacular **FLOPs** figures and remain 80% unused if it cannot be fed fast enough.

This is exactly what is at stake in the inference of large models, which breaks down into two phases with opposite profiles.

- Prefill (reading and processing the prompt), which handles all the input tokens in parallel: these are large matrix multiplications, heavily compute-bound.
- Decode (generating the response token by token): to produce a single token, the system must reread the entire model weights from memory. It reloads gigabytes of weights for a tiny amount of computation, a deeply memory-bound task.
To avoid recomputing everything at each token as the conversation lengthens, the model keeps in memory the computations of the tokens already processed (the KV cache), a store that swells with the length of the context and in turn devours bandwidth.

HBM: Bringing Memory and Compute Closer Together

To feed modern GPUs, it is no longer enough to add more memory: it has to be brought physically closer to the compute.

The industry answered this problem with HBM (High Bandwidth Memory). The idea is simple: instead of placing the memory far from the processor, it is brought as close as possible. Several layers of DRAM (random-access memory) are stacked vertically on top of one another and linked by thousands of microscopic connections passing through the silicon. This stack of memory is then installed right next to the GPU (see the diagram below).

The distances travelled by the data become much shorter and the number of connections much higher. Where a (conventional) DDR module offers a few tens of GB/s of bandwidth, an HBM

stack delivers several TB/s.

This development has an important consequence, however: memory and processor become increasingly inseparable. Future generations of HBM must be designed in close coordination with the GPUs themselves.

The development of an AI chip now depends simultaneously on three players: the processor designer (NVIDIA, AMD), the foundry responsible for assembly (TSMC) and the memory maker (SK Hynix, Samsung or Micron).

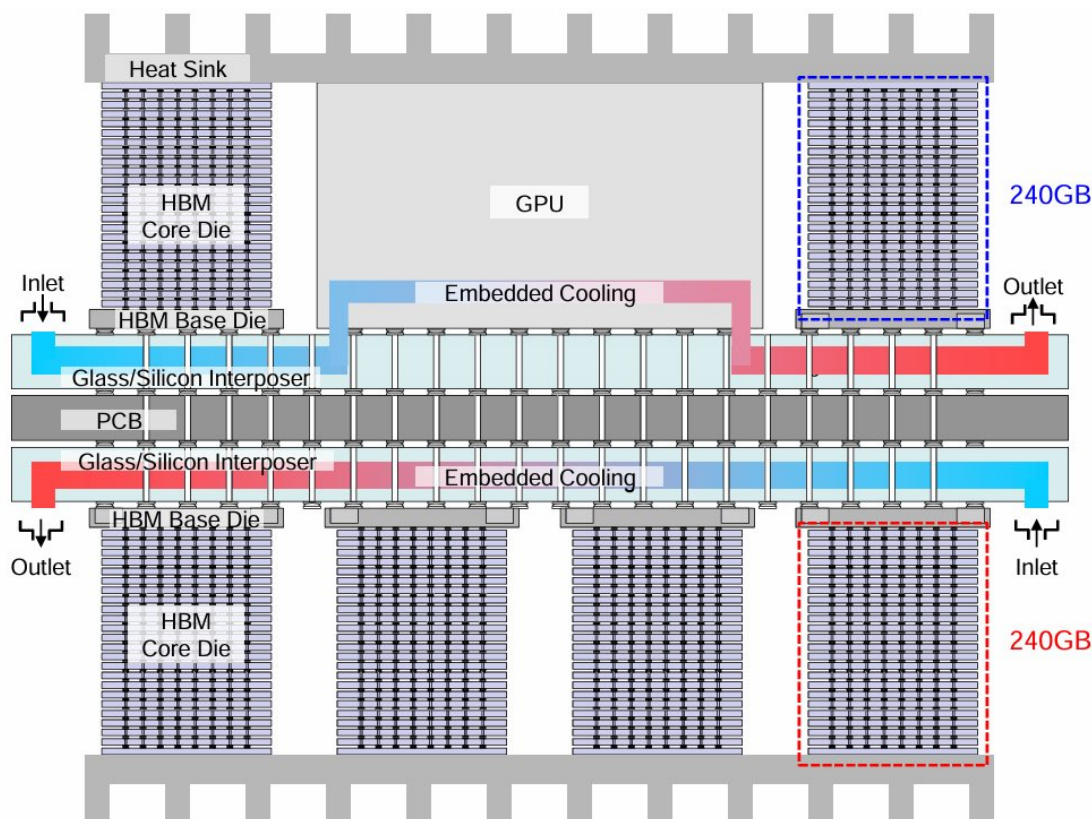
Advanced Packaging: the Industry's New Lock

With HBM, modern systems are no longer limited to an isolated GPU: they combine several specialised components, GPU, HBM memory and ultra-fast interconnects, assembled into a single unit. This integration step is called advanced packaging.

This development is reinforced by a fundamental physical constraint. The most advanced GPUs are already approaching the maximum size limits imposed by lithography machines. It therefore becomes impossible to concentrate ever more compute and memory on a single chip. The industry must now assemble several distinct components to build a complete system.

TSMC dominates this market today thanks to its CoWoS (Chip-on-Wafer-on-Substrate) technology, which makes it possible to connect GPU and HBM very closely. This assembly capability has become as important as the manufacture of the chips themselves.

The diagram below illustrates what a next-generation package might look like:



Source: KAIST TERA Lab

At the centre, the GPU. Around it, the HBM memory stacks, placed as close as possible to the compute. Each stack adds up the memory layers (the core dies) above a control chip linking it to the rest of the system (the base die).

The whole thing rests on the interposer, the plate that connects the GPU to its memory through thousands of connections. To dissipate heat, cooling is no longer a simple heatsink placed on top: a liquid circulates inside the unit itself, entering cold (in blue) and coming out hot (in red).

Interconnect: Scale-up versus Scale-out

A single GPU is not enough for large AI models. The most advanced models are too large to fit in the memory of a single chip and therefore require the computation to be spread across many accelerators.

Two complementary approaches exist.

- Scale-up consists of tightly linking a limited number of GPUs so that they behave like a single coherent system. The aim is to share memory and computation as fast as possible.
- Scale-out consists of connecting together many servers already equipped with several

GPUs. One then moves from a single system to a true cluster that can bring together thousands, or even tens of thousands, of accelerators.

The fundamental difference is that scale-up favours the speed of exchanges between a small number of GPUs, whereas scale-out favours the total size of the system. In both cases, the quality of the interconnects becomes decisive.

Indeed, when a model is spread across many chips, a growing share of the time is spent no longer on computation itself but on moving data. The real performance of a cluster therefore depends as much on its internal network as on the power of the GPUs that make it up.

This is why interconnect has become a strategic battleground. In modern AI infrastructure, moving data efficiently has become almost as important as computing it.

This scale-up logic has a physical translation: the rack. Once a simple metal cabinet in which independent servers were stacked, it has become the basic unit of AI computing, where dozens of GPUs are linked so closely that they behave like a single giant processor.

The Software Stack and Compilers: the Real Divide

A GPU is a parallel-computing machine. To use it, each computation has to be broken down into thousands of simultaneous small tasks, the kernels, distributed across the hardware, with the memory hierarchy managed and data transfers organised.

The decisive link is the compiler: the software layer that translates a model, written for example in PyTorch, into code optimised for a specific chip. It is the compiler that chooses the kernels, fuses operations and organises memory access. A manufacturer can design a chip whose compute power and bandwidth are excellent: if its compiler cannot run any given model efficiently on it, the chip remains largely unused. The limiting factor is rarely the silicon, but the software that extracts performance from it.

This is the main obstacle for NVIDIA's competitors, and it is why the industry is seeking portable software layers, independent of any single manufacturer.

The End of Moore's Law: from Transistor to System

For fifty years, progress followed Moore's law: the density of transistors was doubled every two years by miniaturising them. This dynamic is now running into physical and economic limits. Continuing to shrink requires ASML's EUV lithography, at an exorbitant cost and at the price of extreme geopolitical dependence.

The industry is therefore shifting its axis of optimisation from the transistor to the system: performance comes less from the fineness of the process node than from the way chips are assembled and data is made to flow between them (advanced packaging, chiplets, ultra-fast interconnects). The centre of gravity moves from raw computation to the movement of data.

This shift benefits first those who are denied fine lithography: unable to miniaturise, they optimise assembly and data circulation. China's Huawei, cut off from access to the most advanced

lithography machines, is the textbook example.

The same mechanism plays out on the software side, when a better-designed architecture wins out over brute force: this is what the Chinese laboratory DeepSeek achieved, obtaining with reduced compute resources results that were thought to be reserved for the largest clusters.

Part III: NVIDIA's Hegemony and Saga

NVIDIA does not owe its dominance to a chip, but to the simultaneous mastery of every link in the chain: the software that makes its chips usable, the interconnect that links them, the memory that feeds them, the rack that assembles them, and the developer ecosystem that makes the whole thing unavoidable. NVIDIA does not sell an accelerator but a complete platform.

The CUDA Moat: Why NVIDIA is a Software Company

NVIDIA's lead lies first on the software layer, where it is close to twenty years ahead. In the 2000s, Jensen Huang made a controversial decision: to embed the CUDA components in every NVIDIA chip, including those intended for gamers. This bet trained an entire generation of researchers and engineers who learned to program only on CUDA.

Today, CUDA is a complete ecosystem, continuously developed in parallel with the chips. Its strength lies above all in a set of optimised libraries, such as TensorRT-LLM, which contain years of accumulated tuning and quickly run any new model as soon as it comes out.

It is also what allows already old chips, such as the H100 series, to see their performance double or triple through a simple software update, without any hardware modification.

This is why no competitor closes the gap in a few years: moving a workload off CUDA means rewriting and revalidating millions of lines of code, giving up libraries refined over fifteen years, and retraining engineers who have never known anything else.

You can copy a chip, but not twenty years of software ecosystem nor the base of engineers trained to use it.

From Hopper to Blackwell: from Component to System

The Hopper architecture, launched in 2022, produced the H100, the absolute standard for training large models: it was on clusters of H100s that GPT-4, Claude and Llama 3 were trained. NVIDIA sold millions of them with margins above 70%.

The conceptual break came with Blackwell (announced in March 2024): instead of a monolithic chip, two giant compute dies linked by a very-high-bandwidth interconnect. This assembly complexity rested on TSMC's « CoWoS-L » advanced packaging (see Part II).

A design flaw caused a temporary crisis, opening a brief window for AMD and the hyperscalers' in-house chips, and this episode directly shaped the design of Rubin, conceived to minimise these physical fragilities. Blackwell, then Blackwell Ultra, have been in mass production since 2025.

The essential point lies elsewhere than in the teraflops: with Blackwell, the product ceases to be a chip and becomes a rack.

The Rack as Product: GB200 NVL72 then Vera Rubin

With Blackwell, NVIDIA stopped selling chips and started selling entire cabinets. The GB200 NVL72 is the first example: a complete rack bringing together 72 Blackwell GPUs and 36 Grace processors (NVIDIA's in-house CPU), linked by an ultra-fast interconnect called NVLink, delivered pre-assembled and liquid-cooled.

Vera Rubin, presented in March 2026, takes the logic further. The first rack is already running at Microsoft Azure, with partner availability in the second half of 2026. NVIDIA claims 3.6 exaflops of inference per rack, five times the performance of Blackwell, and above all a cost per token divided by ten.

Beyond performance, Rubin illustrates two underlying dynamics. The first is dependence on memory: Rubin's HBM4 is co-designed between NVIDIA, TSMC and the memory makers, which makes the HBM shortage the real bottleneck of the industry in 2026.

The second is the capacity for absorption. NVIDIA integrated into Rubin the inference technology of Groq, one of its challengers: the perfect illustration of its way of neutralising a threat by absorbing it rather than fighting it.

DGX Cloud and Verticalisation: NVIDIA against its Own Customers

NVIDIA no longer merely sells hardware. With DGX Cloud, it now rents its compute power directly to companies, entering into head-on competition with the hyperscalers (its own customers), who take a very dim view of this capture of value.

This tension is one of the underground engines of the whole sector: it is precisely to escape this dependence, and to recover the margin NVIDIA takes from them, that the hyperscalers are developing their own chips (TPU, Trainium, Maia, MTIA). We return to this in Part IV.

The Omniverse Bet and Embodied AI (GROOT)

Huang's thesis is that LLMs are only a first step. NVIDIA's real objective is « physical AI »: humanoid robots, autonomous machines and industrial systems capable of acting in the real world.

To this end, NVIDIA is developing the GROOT project, a foundation model for robots. Omniverse, in turn, is a platform that recreates physically realistic virtual environments, a kind of digital twin of the world, in which robots train. The point is to learn in simulation in order to avoid having to collect real data, which is slow and costly to produce.

Part IV: NVIDIA's Challengers

No one today can reproduce the entire NVIDIA platform. Its competitors therefore do not seek to build a « better NVIDIA », but target a segment where general-purpose GPUs are too costly or insufficiently optimised.

Training large models remains the hardest ground to attack, because it requires compute, memory, high-performance interconnects and a mature software stack all at once. Inference is more accessible, which explains why most challengers concentrate their efforts on this market.

AMD and the Software Ball and Chain

AMD is the only one taking on NVIDIA with the same weapon: a general-purpose GPU, able to train as well as run models, and programmed in a comparable way. Its chips, powerful and often cheaper than NVIDIA's, do not suffer from a silicon deficit but from a software ball and chain: the ROCm stack still struggles to rival CUDA.

Where NVIDIA's ecosystem runs any model from day one, fully exploiting an AMD chip demands an optimisation effort that only the hyperscalers can afford.

It is they, moreover, who are driving its adoption, as much to reduce their dependence on NVIDIA as for the price.

The Hyperscalers and Verticalisation

The most serious threat comes from NVIDIA's biggest customers. To stop handing over their margin, Google, Amazon, Microsoft and Meta are designing their own chips, not to sell them but to equip themselves at lower cost, especially in inference.

Google is ten years ahead with its TPUs, to the point of taking a revealing step: it has announced that it is separating a chip dedicated to training from another dedicated to inference.

Amazon is pursuing a dual-track strategy, its own Trainium and NVIDIA in parallel, deployed at scale in the supercomputer it is building for Anthropic.

Microsoft, with Maia, and Meta, with MTIA, follow the same logic. The strength of these players does not lie in the power of a single chip, but in their ability to connect tens of thousands of them at very low latency.

They do not design these chips alone, moreover: Broadcom, and to a lesser extent Marvell, co-develop most of them, starting with Google's TPUs, which makes it one of the great discreet winners of the AI wave.

Family 1: Replacing the GPU with another architecture

This is the hardest fight, because it targets the heart of the training market, and it is no longer about making a better GPU but about offering a radical alternative to it. The most striking example

is Cerebras, which etches a single chip the size of an entire wafer instead of assembling thousands of GPUs. The aim is to eliminate communication between processors, which slows down conventional clusters. It is the most credible alternative on very large models, but its software ecosystem remains far inferior to CUDA's.

Another approach, carried by Tenstorrent, bets on openness. Its chips rest on RISC-V, an open and royalty-free instruction set, that is, the basic language a chip knows how to execute, but which no one owns. Where the usual standards require a licence and royalties, RISC-V lets anyone design their chip without paying anything or asking permission.

Tenstorrent takes the logic to its end: its software is entirely public and the company licenses its intellectual property to anyone wanting to make their own chips without depending on NVIDIA. It is the anti-lock-in strategy embodied, prized by players concerned with sovereignty. Its limit remains the same as everyone's: software, since few models still run properly on its chips.

Family 2: The Inference Specialists

This is the most promising segment, because inference lends itself well to specialisation. The most instructive case is Groq, whose very-low-latency inference chip worried NVIDIA enough that the latter absorbed its technology and some of its engineers.

The most interesting technical idea comes from d-Matrix, which brings computation closer to memory to reduce the cost of moving data, the main expense of inference. Etched goes further still with a chip able to run only Transformers: specialisation is maximal and the potential gains high if the reference architecture holds.

Family 3: The Rare Training Challengers

This is the most demanding ground, and therefore the least populated. MatX is the most ambitious, with an architecture designed from the outset for the very large models of leading laboratories, but the company is young and its silicon has not yet been tested at scale. Cerebras again remains the most credible competitor here; the others focus in the short term on inference.

Family 4: Edge & Physical AI: a different market

This family does not seek to dislodge NVIDIA from the data centres, but to run AI directly on devices, as close as possible to the user. This is what is called the edge, as opposed to the centralised cloud. For phones, cars, or robots, sending the data to a distant data centre would be too slow, too costly or too sensitive. The stake there is therefore not raw power but reacting within a few milliseconds without depending on the network, and consuming very little energy.

It is a distinct market, with its own players. The European company Axelera is the most notable, and it links this battle to sovereignty, by designing low-power inference chips in Europe. There it faces above all Qualcomm, which transposes to the data centre and the car its energy-efficiency expertise inherited from the smartphone.

For NVIDIA, these players are not a direct threat, since they do not touch the large compute

centres. Each task that runs on a device, however, is a task that does not run in the cloud, and therefore one fewer GPU to rent.

A Platform, and Therefore a Fragmented Target

All these players often run into the same obstacle: software. The survivors will be those who get around it in one of the only three possible ways: owning the stack end to end, opening it up so that others write it, or shrinking the task so much that the software problem evaporates.

Part V: The Geopolitics of Silicon

The manufacturing of cutting-edge AI chips rests on an inseparable triad: the industrial mastery of the Taiwanese foundry TSMC over the 3 nm and 2 nm nodes, the optics and lithography of the European ecosystem, and the American equipment and software that govern the whole chain.

ASML and Zeiss: the Monopoly on Etching Machines

The manufacturing of the most advanced AI chips depends entirely on ASML's EUV (Extreme Ultraviolet) machines. The Dutch company itself relies on Germany's Carl Zeiss SMT, which produces the ultra-precise mirrors and optics needed for lithography. Developing this technology required decades of research, creating a monopolistic ecosystem that is almost impossible to reproduce.

This industrial dominance has become a major geopolitical issue. Until 2024, ASML's leadership opposed the American restrictions targeting China, denouncing a logic of economic warfare. Its successor had to apply a much harder line under direct pressure from Washington, which in 2026 translated into cancelled deliveries and drastic restrictions on the maintenance of previous-generation machines already installed on Chinese soil.

American Equipment Makers: the Real Lock

Whether a chip is etched in Japan, Taiwan, Korea or Europe, every leading-edge line requires American equipment (Applied Materials, Lam Research, KLA) for the critical steps of deposition, etching and metrology. By combining these monopolies with the use of American patents by foreign players, Washington locks the entire global value chain.

This structural power serves as leverage for Washington, in particular in its policy of banning the sale of the most advanced chips and machines to China, backed by sanctions against those who try to evade it.

In the last year alone, the American trade regulator imposed hundreds of millions of dollars in fines for circumvention, including a striking one against the equipment maker Applied Materials, accused of having routed its machines to China via a Korean subsidiary.

TSMC: the Irreplaceable Foundry and the Taiwan Risk

Before its challengers, a word on the central link: TSMC. The Taiwanese company designs no chips of its own, it merely manufactures them for others (NVIDIA, Apple, AMD). Etching at 3 then 2 nanometres with industrial yields requires decades of know-how that neither Intel nor Samsung manages to catch up with.

The result is a dependence of a rare kind: nearly all of the planet's cutting-edge AI chips come out of the plants of a single company, on a single island. This is the Taiwanese paradox, the « silicon shield »: Taiwan is at once the most precious and the most vulnerable point in the whole global chain, within immediate reach of China.

Under pressure from Washington, TSMC is relocating part of its production to the United States. Its Arizona site, which at 165 billion dollars has become the largest foreign investment in the country's history, already etches 4 nm for Apple and NVIDIA, and is aiming for 2 nm by the end of the decade. The chips etched in Arizona still go back to Taiwan to be assembled, however, because advanced packaging is not yet established there.

Moving the etching is not enough to move the dependence; the island remains, for a few more years, the only place where a cutting-edge chip can be built end to end.

Intel Foundry Services: the Comeback Bet

Between 2020 and 2026, Intel tried to become a central player again after years of lagging behind TSMC: it became a « foundry » open to external customers, and relocated part of its production to the United States thanks to the subsidies of the CHIPS Act.

The initial execution was difficult: Intel was investing heavily while AMD was capturing the growth of the server market. The turning point came in 2026: Apple and Intel reached a preliminary agreement, reported in May, for Intel to manufacture part of Apple's chips on its 18A node.

This signal validates the turnaround, all the more so as Intel has also attracted a 5 billion dollar investment from NVIDIA and commitments from Amazon, Microsoft and Google on the 18A.

Samsung: the Ambivalent Giant of Memory and Foundry

Samsung holds a unique position: both a foundry competing with TSMC and the world's second-largest producer of HBM.

On the foundry side, it has lost ground: its 3 nm nodes suffer from insufficient yields that pushed customers such as Qualcomm back to TSMC, relegating it to second place on the most advanced nodes.

On memory, the blow is even more symbolic: SK Hynix was selected as the priority supplier by NVIDIA for the HBM4 of Vera Rubin, relegating Samsung to the role of complementary supplier.

The American Legislative Architecture and the Containment of China

Rather than a total blockade, the American administration chose a middle path: authorising, in a trickle, the sale to China of certain advanced chips, under conditions. In January 2026, the BIS, the trade bureau that issues these authorisations, formalised this case-by-case approach.

Each chip must be tested before export, volumes are capped, and above all each sale is hit with a 25% customs tariff paid back to the US Treasury. The most advanced chips, of the Blackwell generation, remain banned.

At the same time, Congress is pushing in the opposite direction. A bill, the AI Overwatch Act, which came out of committee in January 2026, would give it the power to block, case by case, the sales of advanced chips to China that the executive has just authorised.

It is still far from being passed, but it reveals the American contradiction: while the White House charges for access, part of Congress is seeking to close it.

The Chinese Response: War over Minerals and the Strategy of Quantity

At the very moment when Washington partially reopens the door, Beijing is pushing its companies to turn away from it. In May 2026, China published for the first time a list of « safe and reliable » AI chips that its administrations must favour: nine Chinese models appear on it, those of Huawei or Cambricon for example, but none from NVIDIA. The aim is full autonomy that would render sanctions ineffective.

This goal rests on a highly vertical ecosystem. Huawei is at its centre, with its Ascend accelerator and its in-house software stack CANN, designed as the alternative to CUDA. NVIDIA's share of the Chinese market collapsed after Beijing steered its administrations and its cloud giants towards domestic silicon. Analysts see Huawei alone capturing around 60% of the Chinese market by the end of 2026.

The same quest for independence reaches down to the instruction set, the basic language a chip executes: rather than the Western standards subject to licensing, China is investing in RISC-V, open and royalty-free, through projects such as XiangShan, backed by Alibaba and the Academy of Sciences.

This strategy remains constrained by manufacturing. Deprived of the most advanced etching machines, the national foundry SMIC is stuck at the 7 nm node. Unable to miniaturise, Huawei is betting on stacking logic and shortening signal paths (see Part II). The avowed doctrine is to offset the deficit in raw power through quantity and the saturation of the domestic market. Chinese HBM remains a generation behind.

But Beijing retains a formidable lever: its near-monopoly on rare earths and the critical minerals of electronics (gallium, germanium), which it can cut off precisely where the West is most vulnerable.

Part VI: The Race for Foundation Models and the Collapse of Margins

The foundation-model layer quickly took shape around a handful of players. In 2026, three laboratories dominate the Western race: OpenAI, Anthropic and Google DeepMind. Anthropic embodies the sector's dizzying acceleration, with a revenue run-rate that went from around 1 billion dollars in early 2025 to 47 billion by mid-2026.

To understand why the sector poured such sums into raw compute, one has to go back to an idea formulated by the researcher Rich Sutton, the « bitter lesson ». His observation, drawn from several decades of AI history: in the long run, the methods that simply exploit ever more computing power almost always end up beating the clever approaches devised by humans.

This is the theoretical foundation of the scaling law and of the whole race: if adding compute almost certainly wins, then spending tens of billions on GPUs and gigawatts becomes perfectly rational. It is precisely this doctrine that DeepSeek will crack.

From Training to Inference: The Great Shift

Until 2024, AI compute was massively dominated by training: a few dozen labs and companies spent billions to train foundation models on clusters of thousands of GPUs, in long and infrequent campaigns.

Since 2025, the centre of gravity has shifted to inference. Billions of daily requests now consume more compute than training itself. The central metric is no longer the FLOP per second but the cost per token. It is now mass inference needs, rather than training campaigns, that dictate the hyperscalers' energy investments and size their data centres.

The DeepSeek Shock: When Algorithmic Efficiency Challenges Brute Force

In January 2025, the Chinese laboratory DeepSeek caused a first shockwave by proving that a better-designed architecture could compensate for a deficit in silicon. Its key technique, the « mixture-of-experts (MoE) » (MOE), activates only a small fraction of the model's parameters at each request rather than all of them: fewer weights to re-read from memory, hence a structurally lower cost per token. It is a direct answer to the memory wall described above, not merely an aggressive pricing policy.

DeepSeek proved that an open-weights model could rival the American leaders at a tiny fraction of the cost. With its new models released in May 2026, the Chinese laboratory turned the 2025 technical feat into an all-out price war. Its reasoning model shows a price per token nearly thirty times lower than that of Claude or GPT, for performance close to theirs on programming benchmarks.

The Collapse of Margins and the Profitability Crisis of the American Laboratories

By selling a million tokens at a derisory price, DeepSeek destroys the sector's margins. For OpenAI or Anthropic, matching these prices has become financially untenable: their models cost far more to run than they bring in.

The classic American models have to load the entire context into the GPUs' ultra-expensive memory. With each request processed, OpenAI and Anthropic subsidise at a loss the cost of the electricity and hardware needed to generate the answer. Without a complete software overhaul aimed at imitating Chinese algorithmic efficiency, the growth in user volume mechanically widens their financial losses and makes their dependence on raw compute hard to sustain.

Faced with this simultaneous pressure on margins and on compute capacity, the American laboratories have only one way out: securing infrastructure at any cost. This is what explains the giants' vertical integration and the race for gigawatts that now structures the sector.

Part VII: Infrastructure and the Energy Crisis

The Giants' Unprecedented Vertical Integration

The race for frontier compute has shattered the classic scales of technology investment. Launched in early 2025, the Stargate project is a joint venture bringing together OpenAI, Japan's SoftBank and Oracle to build in the United States a network of giant data centres dedicated to training models. With commitments exceeding 100 billion dollars, it aims at an unprecedented concentration of computing power.

Faced with this gigantism, Amazon responds with full vertical integration through its Project Rainier, the supercomputer it is deploying in Indiana for Anthropic. The point for Amazon is to run its own in-house chips there at scale, the Trainium, rather than NVIDIA.

The Neoclouds and the Risk of GPU Debt

To meet the initial GPU shortage, new specialised players such as CoreWeave, Lambda Labs or Crusoe emerged. Their model rested on a simple logic: borrow heavily to buy NVIDIA servers, then rent out that computing power for AI (as opposed to the generalist hyperscalers). CoreWeave pushed this financial arrangement to the extreme by using its H100 chips directly as collateral to raise billions of dollars of debt.

In 2026, this model is showing its systemic limits. The arrival on the market of next-generation chips (Blackwell) abruptly accelerated the obsolescence of H100 fleets, while DeepSeek-style algorithmic efficiency caused rental prices for computing power to collapse. The Neoclouds that did not secure very long-term contracts with hyperscalers find themselves trapped.

xAI, Colossus and the Philosophy of Brute Force

xAI, Elon Musk's AI laboratory, embodies a philosophy diametrically opposed to that of DeepSeek: where the Chinese engineers optimise, Musk builds by brute force. Colossus, his supercomputer set up in Memphis, went from 100,000 to more than 200,000 NVIDIA GPUs in a few months.

This bet produced an excess of capacity that Grok, xAI's model, was never enough to absorb. It is in this context that one of the most improbable deals in AI history was struck. In May 2026, Anthropic signed a contract with xAI worth 1.25 billion dollars per month until 2029, about 40 billion in total, to rent the entirety of Colossus 1. The reason: Anthropic saw revenue and usage growth eight times higher than its forecasts at the start of the year, creating a major compute crisis.

In February 2026, Musk publicly called Anthropic « misanthropic and evil ». When access to compute is at stake, ideological rivalries give way to contracts.

Terafab: Vertical Integration Taken to the Extreme

The most radical project of 2026 was unveiled by Elon Musk in March: Terafab, a joint venture between Tesla and SpaceX/xAI. The idea is a mega-foundry that would bring under one roof every stage of production, from design to lithography, memory, packaging and testing, for an announced cost of around 25 billion dollars.

The stated ambition is to produce a terawatt of computing power per year, a volume on a scale unlike anything in the current industry, a large part of which is designed to operate in space, on solar-powered satellites.

Behind the scenes, it is also a bargaining lever with TSMC: by threatening to break free of it, Musk's group is seeking to improve its position in the foundry's queue. For now the project is more an outsized announcement than an established fact, since building a leading-edge foundry is one of the hardest industrial feats there is.

Energy, the New Limit of Compute, and the Nuclear Renaissance

The explosion in AI's energy needs triggered a major strategic reversal: absolute dependence on the electricity grid and the revival of civil nuclear power. Data centres now consume such colossal volumes of electricity that intermittent renewables (solar, wind) are unable to guarantee the required stability (around the clock).

To secure their growth, Big Tech is directly financing the nuclear sector through long-term electricity purchase agreements. Microsoft has secured more than 800 MW through its contracts. Google and AWS have both signed landmark deals to deploy fleets of SMRs by 2030-2035.

France and the Nuclear Card: Compute Through the Prism of Gigawatts

International competition has shifted from the terrain of silicon to that of electricity grids. The acceleration of mass inference and reasoning models consumes such gigantic volumes of electricity that immediate access to stable megawatts has become the main asymmetric advantage.

On this point, the world map is fracturing. While American hyperscalers buy up nuclear plants across the Atlantic and the Gulf deploys solar parks coupled with gas turbines, continental Europe suffers from high energy costs and saturated grids.

In this gigawatt war, France has a key advantage: its civil nuclear infrastructure, stable and decarbonised, is now establishing itself as its main asset for attracting European compute centres, turning its energy policy into a pillar of its digital sovereignty.

Part VIII: The European Awakening

The Legacy of Failures: From Andromède to the Gaia-X Illusion

The history of European sovereign cloud is one of strategic misunderstanding. As early as 2009, the French Andromède project (split into Cloudwatt and Numergy) failed because the State conceived it as a classic, passive hosting infrastructure, where AWS was building a platform betting on API-driven automation and designed for developers. Ten years later, the collective Gaia-X attempt bogged down in bureaucracy: by bringing American hyperscalers into its governance, the project lost its substance, prompting Scaleway's departure in 2021.

This historic split still divides Europe into two camps in 2026: on one side the hard sovereigntists, on the other the pragmatists of hybrid alliances.

The Hard Sovereigntists Against the « Trusted Cloud »

The camp of full technological sovereignty is led by OVHcloud and Scaleway. OVHcloud remains the industrial champion of bare metal, but the transition to AI requires massive investment in NVIDIA chips. Scaleway, by contrast, bet on hyper-specialisation by buying large GPU clusters very early to become the refuge of European AI startups. In the public sector, it is NumSpot that is pushing to equip administrations and healthcare, aiming for SecNumCloud qualification, the sovereignty label issued by the State.

This head-on model stands against the pragmatic doctrine of the « Trusted Cloud », embodied by Bleu (Orange and Capgemini with Microsoft) and S3NS (Thales with Google). The starting idea: to control American technologies by operating them under licence, behind European legal protection. These players are also aiming for SecNumCloud, and S3NS has already obtained it, proof that the boundary between the two camps is more porous than it appears.

In 2026, the verdict on these alliances remains open. For its defenders, it is a pragmatic first step towards gradual sovereignty; for its critics, a legally softened dependence with no real technological break.

The Mistral AI Case: Sovereignty and Pragmatism

The trajectory of Mistral AI illustrates the continent's complex position. The Paris startup waged intense lobbying against the excesses of the AI Act in the name of sovereignty, while leaning on a partnership with Microsoft and its Azure infrastructure: a dependence openly accepted, for as long as it takes to grow.

Since then, it has diversified its support. Its leading shareholder is now the equipment maker ASML, with about 11 % of the capital, and it is building its own compute capacity in France, with a cluster hosted by Scaleway near Paris.

Rather than competing on raw compute, the French ecosystem (around Mistral, Station F and the Kyutai laboratory) is betting on efficiency: open, ultra-optimised models whose lightweight

architecture aims to match the closed American models without requiring Silicon Valley's outsized infrastructure.

InvestAI, AION and the Nuclear Card: Europe in 2026

The real operational turning point of 2026 is energy. Faced with the logistical difficulties of the United Kingdom and Germany, where AI mega-data-centre projects are on hold because of the cost of electricity and grid saturation, France is using its civil nuclear fleet as a competitive advantage.

It is in this context that the European InvestAI Facility programme (20 billion euros) and the emergence of the French AION consortium take shape. Driven by Scaleway and major industrial groups, AION aims to build in France a world-class compute infrastructure connected directly to EDF's nuclear plants. France's ambition is clear: to become the energy and physical hub of European compute.

This ambition has just found its most spectacular expression. In May 2026, at the Choose France summit, SoftBank announced an investment of up to 75 billion euros to build as much as five gigawatts of data centres dedicated to AI, presented as the largest commitment ever made to a digital infrastructure in Europe.

Part IX: The Macro-Strategic Shock and the Global Legal Vice

The CLOUD Act Affair

This debate over sovereignty stems from a historic standoff. In 2013, the American courts demanded access to emails stored by Microsoft on servers in Ireland. The firm refused, protecting data hosted abroad. To settle the matter once and for all, the US Congress passed the CLOUD Act in 2018. The text compels any technology company under American jurisdiction (such as the hyperscalers) to hand over its users' data to the authorities, wherever in the world it is physically stored.

In the age of artificial intelligence, this mechanism turns cloud infrastructure into a tool of direct interference, rendering the very idea of localised hosting void if the service provider remains American.

The « Liberation Day » Shock and the Transatlantic Rupture

The major political shift came in spring 2025. During « Liberation Day », the Trump administration imposed a wave of massive, undifferentiated tariffs on imports from China (34 %) and the European Union (20 %).

Beyond the immediate economic shock, these measures caused a psychological rupture in Europe. They demonstrated the danger of depending on cloud infrastructure controlled by giants subject to the decrees of an unpredictable American government. In the eyes of European decision-makers, this episode legitimised the introduction of technological-preference policies.

The Regulatory Tempo: Ambitions and Realities for the European Ecosystem

Faced with this vice, the European Union's response took shape through the rollout of the AI Act. The outcome of this regulatory tempo, however, is deeply asymmetric. The American giants have the resources to absorb the cost of European compliance, which they see simply as a tax for access to a market of 450 million consumers.

This regulatory burden, by contrast, weighs disproportionately on European startups and SMEs. Deprived of massive capital and lacking access to cheap compute, they must devote a significant share of their resources to compliance rather than R&D. Having to adapt constantly to changing rules slows their growth and reduces their competitiveness against the global leaders.

Conclusion: Infrastructure as Destiny

The history of cloud, compute and sovereignty between 2006 and 2026 confirms a simple thesis: lasting value lies not in software but in the physical layers that run it. An application can be copied over a weekend. No one duplicates over a weekend a rack of 72 GPUs, a stack of HBM memory, a 2 nm fabrication plant or a nuclear reactor. This hyper-materialisation of the digital has reshuffled the cards at every scale.

At the company scale, the hyperscalers had made pricing complexity and exit fees an instrument of capture. But AI shifted the battle: power no longer rests on commercial tricks, it rests on access to chips and electricity, now so scarce that the providers themselves run short.

At the industry scale, power has lodged itself in the links hardest to replicate. NVIDIA proved it: its dominance rests not on a chip but on twenty years of software ecosystem and on mastery of the entire system. This near-monopoly will not necessarily last. The centre of gravity is moving from training towards inference, where the software lead matters less and where specialised architectures attack cost per token rather than raw power. Silicon remains the decisive terrain, but the question is no longer only who can build the best chip; it is who will be able to run the models at the lowest price.

At the scale of States, access to compute is now a matter of sovereignty on a par with energy or territory, and it merges with them: training and running the models consumes so much electricity that the map of power is being redrawn in gigawatts.

In June 2026, three blocs contend for this sovereignty. The United States dominates through the depth of its capital, the density of its ecosystem and control of the chain's critical nodes. China resists through quantity, its algorithmic efficiency and its leverage over critical minerals, building an entire stack out of Washington's reach. Europe is looking for its place. It does, however, hold one irreplaceable link in the whole chain, ASML and the lithography monopoly, but remains dependent for the rest, torn between hybrid alliances and the ambition of a few players such as Mistral, Scaleway or the AION project.

This reopening holds at every level. NVIDIA's software lock pushes Tenstorrent to bet on open architectures; its raw power, cut out for training, leaves Cerebras to attack it on inference; the Western lithography monopoly pushes China towards RISC-V and its own plants. Every point of dominance gives rise to those who want to break it, and it is often the harshest constraint that gives birth to the breakthrough, as DeepSeek showed.

The future will be decided neither by a Brussels directive nor by yet another model. It will be decided by the capacity to mobilise gigawatts, to master 2 nm lithography and to face an uncomfortable idea: in the economy of AI, infrastructure is not a means, it is the end.