A FIELD GUIDE | EDITION 1 | OCTOBER 2026

THE COMPUTE DECADE

THE AI ADVANTAGE

A field guide to the money, chips and power reshaping the next ten years

Written for curious beginners. No engineering degree required.

ABOUT THIS BOOK

Before you begin

This book exists because one question keeps coming up in 2026: **why is the whole world suddenly spending trillions of dollars on computers, chips and power plants?** If you have never studied technology or finance, the headlines can feel like a foreign language. Hyperscalers. Capex. HBM. Gigawatts. Inference. This guide translates all of it.

It is organised as a journey. We start with the basic idea of "compute", then follow the money, then the chips, then the electricity, then the politics, and finally the question that matters most: what could happen over the next ten years, and what should an ordinary person or small business do about it?

How to use this book

- **Complete beginner?** Read in order. Every new term is explained the first time it appears, and there is a glossary at the back.
- **Short on time?** Read the "In one minute" box that opens each chapter, then the "Key takeaway" at the end.
- **Want the numbers?** Turn to the Cheat Sheet on the last pages, and the Sources list for where each figure came from.

A note on the numbers

All figures were gathered from public company filings, agency reports and financial news up to early October 2026. Forecasts differ between banks and analysts, sometimes by hundreds of billions of dollars, so this book shows ranges and names the source type where it matters. Treat every forecast as an informed estimate, not a fact. This book is educational and is **not financial or investment advice**.

How this book was made

This edition was researched on the open internet and written with Claude, an AI assistant made by Anthropic. Company results (such as Nvidia's quarterly numbers) come from the companies' own filings. Industry forecasts come from published bank, agency and analyst reports. Where sources disagreed, the book says so rather than hiding it. Please verify any figure before using it in an investment decision, a client report or a published article.

CONTENTS

What is inside

Before you begin 2

What is inside 4

Why compute is the story of the decade 5

What is "compute", really? 8

The AI stack: five layers 10

The trillion-dollar buildout 13

Who pays, and what could go wrong 16

Nvidia and the GPU kingdom 19

The choke points: TSMC, packaging and memory 21

The challengers: custom chips 23

Power: the new binding constraint 26

Chips as statecraft 29

India's compute moment 31

2026 to 2036: three scenarios 34

What it means for you 36

Glossary 38

Cheat sheet: ten numbers to remember 40

Sources and further reading 41

INTRODUCTION

Why compute is the story of the decade

In one minute

"Compute" means computing power: the chips, servers and electricity that let AI think. In 2026, the biggest technology companies are spending roughly $700 billion in a single year to build more of it. Nvidia, the leading AI-chip company, reported $96.2 billion of revenue in one quarter. Electricity, not software, is becoming the limit. Whoever controls compute shapes the economy.

In the 2000s the scarce resource of the internet was attention. In the 2010s it was data. In the 2020s, the scarce resource is **compute**. You can have the cleverest AI idea in the world, but if you cannot rent enough chips and power to run it, it stays an idea.

Look at what happened in only three years. After ChatGPT arrived in late 2022, demand for AI chips exploded. Nvidia's data-centre revenue for its fiscal 2026 reached about $194 billion, roughly thirteen times the level before ChatGPT, according to its results. Then in the quarter ending July 2026, Nvidia alone brought in $96.2 billion, more than double the same quarter a year earlier. Its chief executive summed up the new mood in one sentence: *"Now, compute is revenue."*

That sentence is the heart of this book. For AI companies, compute is no longer a cost hidden in the back office. It is the factory floor. More compute means more AI "tokens" produced, and more tokens produced means more products sold. That is why the money, the chips and the power are all being pulled into the same giant project at once.

The three forces: money, chips, power

- **Money.** Tech giants, governments and investors are committing sums that rival the budgets of entire countries. Part II explains where it comes from and what could go wrong.
- **Chips.** Only a handful of companies on earth can make the processors and memory that AI needs. Part III shows the bottlenecks.
- **Power.** Every AI data centre is an electricity customer on the scale of a small city. Part IV shows why the grid has become the new frontier.

A fourth force, **politics**, sits on top of the other three, because governments now treat chips like strategic weapons. Part V covers it, including what all this means for India.

PART I

Foundations

What compute is, and how the whole machine fits together.

CHAPTER 1

What is "compute", really?

In one minute

Compute is the raw processing capacity used to train and run AI. It is measured physically by chips and by electricity (megawatts and gigawatts), and commercially by how many AI answers ("tokens") you can produce per dollar.

Think of a bakery. The recipe is the AI model. The flour is data. The ovens are the chips. And the gas or electricity heating those ovens is the power. Compute is the combination of ovens and fuel: how much bread you can bake per hour.

Training and inference

AI has two big jobs. **Training** is teaching the model, like a student studying for years. It is enormous but happens in bursts. **Inference** is using the model to answer questions, write code or generate images, like the student working a job every day. Early AI spending focused on training. By 2026 the weight is shifting to inference, because millions of people and thousands of companies now use AI constantly, and "AI agents" that complete multi-step tasks can use many times more computation than a single chat reply.

What is a token?

A token is a small chunk of text, roughly three-quarters of an English word. AI models read and write in tokens. When executives say "the cost per token is falling", they mean the same AI answer costs less to produce. Nvidia says its new Rubin platform aims to cut inference token cost by up to ten times compared with Blackwell, its previous generation.

Cheaper tokens do not shrink spending, they usually encourage far more usage.

The units you will keep seeing

Term	Plain meaning
GPU	Graphics processing unit. A chip that does thousands of calculations in parallel. Originally for video games, now the workhorse of AI.
Data centre	A warehouse-sized building full of servers, cooling and power equipment.
Megawatt (MW)	One million watts. A mid-sized data centre may use 20 to 100 MW.
Gigawatt (GW)	1,000 MW. Roughly the output of one large power plant, or the electricity use of a big city.
Hyperscaler	A giant cloud company: Amazon, Microsoft, Google, Meta, Oracle.
Capex	Capital expenditure. Money spent on building things (data centres, chips) rather than daily running costs.

Key takeaway: Compute is not magic. It is chips plus buildings plus electricity, and each of those can run short.

CHAPTER 2

The AI stack: five layers

In one minute

Every AI product sits on a stack: power, buildings, chips, models, applications. The lower layers are physical and slow to expand. The upper layers are software and fast. Bottlenecks at the bottom set the speed limit for everything above.

Picture a five-storey building. The ground floor is power. On top of it sits the data-centre shell. Above that, the chips. Above that, the AI models. On the roof, the apps people actually use.

Layer	What it is	Example players	Typical constraint
5. Applications	Chatbots, coding tools, search, agents	Thousands of firms	Customer demand
4. Models	The trained AI "brains"	OpenAI, Anthropic, Google, Meta, DeepSeek	Talent, compute
3. Chips	GPUs, custom chips, memory	Nvidia, AMD, Broadcom, TSMC, SK hynix	Manufacturing, memory
2. Data centres	Buildings, cooling, networking	Equinix, hyperscalers, neoclouds	Land, equipment, permits
1. Power	Grid, gas, nuclear, solar	Utilities, energy firms	Grid connection, turbines

Notice something important. Layers 4 and 5 can change in months. Layers 1 to 3 take years. A new AI app can launch in a weekend, but a new power plant or chip factory takes five to ten years. That mismatch of speeds explains almost every headline in this book.

Key takeaway: When you see an AI story, ask which layer it is about. The scarce layer usually decides who earns the profit.

PART II

The Money

Where a trillion dollars comes from, where it goes, and what could break.

CHAPTER 3

The trillion-dollar buildout

In one minute

The largest US cloud companies are on track to spend roughly $700 billion or more on capital expenditure in 2026, and forecasts for 2027 and 2028 are higher still. UBS projects about $1 trillion in 2026 rising to about $1.6 trillion in 2028. Estimates keep being revised upward.

To understand the scale, compare with history. The US interstate highway system, the Apollo programme and the dot-com fibre boom were each enormous. Many commentators now compare the AI buildout with them, and it is being paid for mostly by private companies rather than governments.

What the forecasts say

Different banks count different companies, which is why numbers vary. Reported ranges for 2026 include about $600 billion for the five largest hyperscalers (a January 2026 estimate), about $725 billion (mid-year analysis), and $775 to $800 billion after first-quarter earnings, with Goldman Sachs tracking $732.5 billion for the big four after Q2. Forecasters have repeatedly been too low: Goldman's own charts show estimates for 2026 climbing again and again.

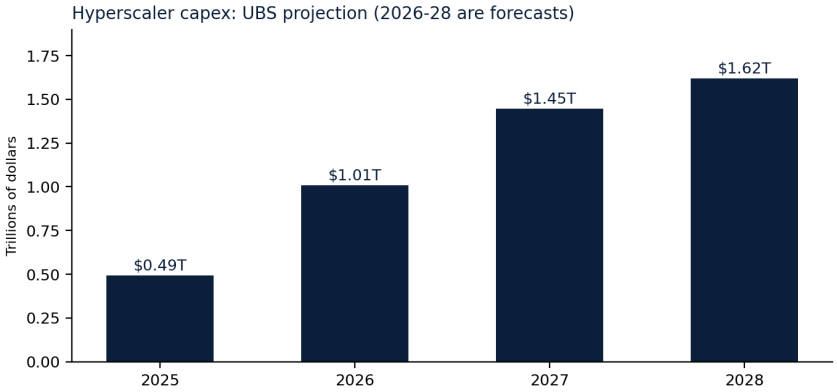


Figure 1. UBS projection of total hyperscaler capital spending. Source: UBS, as reported by Yahoo Finance (2026).

UBS also estimates that hyperscalers will spend about $4.1 trillion from 2026 to 2028, three times the $1.3 trillion spent over the previous six years. Goldman Sachs projects AI-related capex of $765 billion in 2026 climbing to $1.64 trillion in 2031, a cumulative $7.6 trillion.

Where does a data-centre dollar go?

Chips are the headline, yet they are a shrinking share. One 2026 analysis found that more than 60 percent of spending now goes to power, cooling, buildings and networking rather than accelerators. Another breakdown suggests about 75 percent of 2026 hyperscaler capex, roughly $545 billion, is tied to AI-specific infrastructure. Memory makers, networking firms, electrical-equipment suppliers and construction firms all share in the flood.

Why not just spend less?

Because each company fears losing the race more than it fears overspending. Executives argue that if demand is real and they have no capacity, rivals take their customers. That logic, shared by all of them at once, is what keeps the numbers rising.

Key takeaway: The buildout is bigger, faster and more repeatedly under-forecast than almost anyone expected. Treat any single forecast as a floor in a rising market, but remember forecasts have been wrong in both directions in past booms.

CHAPTER 4

Who pays, and what could go wrong

In one minute

For years the giants paid for AI out of profits. Now spending is overtaking cash flow, so debt, private credit and "circular" deals are growing. That makes the boom more fragile if AI revenue disappoints.

Here is the uncomfortable arithmetic. According to analysis from Epoch AI, combined cash capex of the five major hyperscalers was on track to exceed their combined operating cash flow for the first time around the third quarter of 2026. UBS calculated that Amazon, Alphabet and Microsoft together would spend 102 percent of their cloud revenue on capex in 2026. In plain words, they are plowing back every rupee, dollar or euro their cloud businesses earn, and then some.

Three ways the boom is financed

- **Cash flow.** Still the biggest source. These companies earn extraordinary profits.
- **Debt.** AI-related debt issuance jumped 112 percent in 2025, according to Capital Group, and 2026 is on pace to be larger. One analysis projects around $1.5 trillion of new borrowing over three years.
- **Creative deals.** Short-term leases with "neoclouds" (specialist AI-hosting companies), private credit funds, and vendor financing, where a chip supplier effectively helps fund its own customer.

The bubble debate

Is this a bubble? Honest experts disagree. Here is each side's best case.

The bull case	The bear case
Demand is real: chip supply is sold out, and Nvidia says its next year is supply constrained.	Spending is running ahead of proven AI revenue for many customers.
The biggest spenders are highly profitable and have low leverage so far.	Debt, vendor financing and circular deals add hidden risk.
Agents and daily use push inference demand far higher than chat did.	One or two AI labs are huge customers for many suppliers (concentration risk).
Falling token costs open new uses, a pattern seen in earlier technologies.	Power, memory and equipment shortages could delay returns, raising costs.

A balanced reading: the technology is genuinely useful, and demand today is real, but the financing is increasingly aggressive, and timing matters. Even a good technology can be a bad investment if you pay too much or borrow too heavily to buy it. History (railways, fibre optics) shows that real revolutions can still produce painful busts for investors.

Key takeaway: Watch three signals: whether cloud revenue keeps growing, whether debt costs rise, and whether chip orders get cancelled or delayed.

PART III

The Chips

Who designs them, who builds them, and why they are so hard to get.

CHAPTER 5

Nvidia and the GPU kingdom

In one minute

Nvidia designs the dominant AI chips and the software and networking around them. In its quarter ended July 26, 2026 it reported $96.2 billion revenue, $89.0 billion from data centres, with 75 percent gross margins. Its market value passed $5 trillion. Its dominance is real but contested.

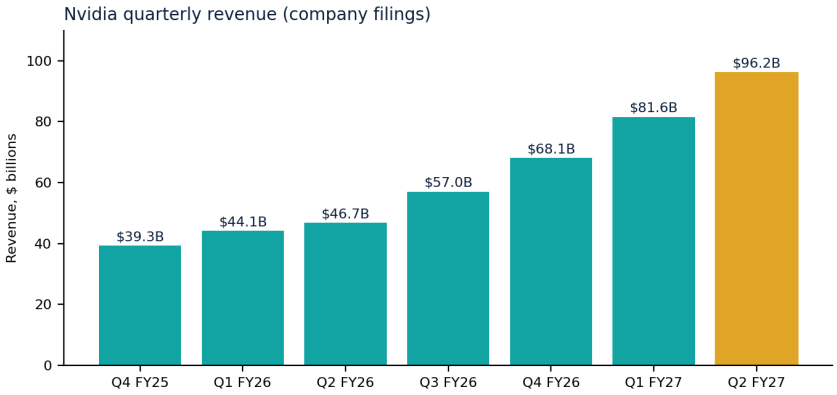


Figure 2. Nvidia quarterly revenue, fiscal Q4 2025 to Q2 2027. Source: Nvidia filings and results releases.

Look at Figure 2. In the quarter ending January 2025, Nvidia earned $39.3 billion. Eighteen months later it earned $96.2 billion. Its guidance for the next quarter is about $108 billion, and its finance chief described the following fiscal year as a "supply-constrained" outlook with revenue growth of about 70 percent, well above the 44 percent analysts had expected, according to LSEG data reported by CNBC.

Why Nvidia wins

- **Performance.** Each generation (Hopper, then Blackwell, now Vera Rubin) delivers big leaps. Nvidia says Rubin is in full production and expected to reach about 20 percent of data-centre revenue.
- **Software.** Its CUDA platform has millions of developers. Switching away is painful and slow.
- **Networking.** Nvidia sells the cables and switches that connect thousands of chips into one machine. Networking revenue was about $11 billion in a single quarter by early 2026.
- **Customer concentration.** A few cloud giants account for a little over half of data-centre sales, which is both a strength and a risk.

Rivals

AMD sells competing accelerators. Intel is trying to recover. Chinese firms such as Huawei are building a separate ecosystem (Chapter 9). But the most serious pressure comes from Nvidia's own customers, who are designing their own chips (Chapter 7).

Key takeaway: Nvidia is the toll booth of the AI age, and the toll is high. But no one enjoys paying a toll, so the customers are building bypass roads.

CHAPTER 6

The choke points: TSMC, packaging and memory

In one minute

Designing a chip is not the hard part. Manufacturing the most advanced ones, packaging them, and supplying high-bandwidth memory (HBM) is. Each step is dominated by a few firms, and all are sold out well into 2027.

Nvidia does not own factories. Its chips are made by **TSMC** in Taiwan, the world's leading contract chipmaker. TSMC's most advanced manufacturing lines were reported to be booked through 2027.

Step 1: Make the chip

An AI chip is etched on a silicon wafer using machines that cost hundreds of millions of dollars each. Only TSMC, and to a smaller degree Samsung and Intel, can do this at the leading edge.

Step 2: Package it

Modern AI processors are not one chip but several pieces bonded together on a tiny silicon "bridge", using a TSMC process called **CoWoS**. For several years, CoWoS capacity limited how many AI chips could ship at all. TSMC has been expanding it quickly, and some analysts expect packaging to ease sooner than memory.

Step 3: Feed it memory

An AI chip is only as fast as the memory next to it. **HBM** (high-bandwidth memory) is stacks of memory chips layered vertically and placed right beside the processor. Three companies make it: SK hynix, Samsung and Micron. In late 2025, SK hynix's finance chief said

the company had already sold out its entire 2026 HBM supply, and Micron said its HBM for 2025 and 2026 was fully booked.

The squeeze spread. HBM uses so much manufacturing capacity that ordinary memory for phones, laptops and cars has gotten scarcer and pricier. In July 2026, SK hynix's chief warned that the memory market was headed toward its worst supply shortage in 2027, with tightness that could last until 2030, according to industry reports. Nvidia and SK Group also announced a large multi-year partnership in July 2026 covering HBM4 memory for Vera Rubin. A separate Bank of America analysis, cited in industry coverage, suggested that bottlenecks could limit new memory capacity through 2028 to a fraction of plans. These items come from trade reports, so treat the exact figures cautiously.

Why memory matters for your wallet

If you have noticed smartphone and laptop prices creeping up, or RAM getting expensive, AI is one reason. The same factories that make consumer memory are being redirected to HBM, where profits are higher.

Key takeaway: The AI industry runs through a few narrow doors: TSMC's factories, advanced packaging, and HBM. Whoever holds the door holds the profit.

CHAPTER 7

The challengers: custom chips

In one minute

Google, Amazon, Meta, Microsoft and OpenAI are designing their own AI chips with help from Broadcom and Marvell. Custom chips are growing faster than Nvidia GPUs, mostly for inference. Nvidia still holds roughly 70 percent or more of the market by most estimates.

If you spend $50 billion a year on someone else's chips, designing your own starts to look sensible. That is the logic behind **custom silicon** (also called ASICs: application-specific integrated circuits). A custom chip does fewer things than a GPU but does its chosen job more cheaply and with less power.

Who is doing what

Company	Custom chip	Note
Google	TPU (Tensor Processing Unit)	The oldest and most mature custom AI chip programme.
Amazon	Trainium and Inferentia	Amazon said its chip business reached a $20B revenue run rate.
Meta	MTIA	Hundreds of thousands deployed for inference on its apps.
Microsoft	Maia	Second generation arrived in early 2026.
OpenAI	Built with Broadcom	Part of a push to diversify away from one supplier.

Trade analysts report that custom AI server shipments could reach about 28 percent of the market in 2026, and that Broadcom has a very large AI order backlog, with its CEO describing line of sight to over $100 billion

of AI chip revenue in 2027. Some forecasts even suggest Nvidia's share of inference could fall substantially by 2028. Those are projections, not facts, and Nvidia keeps moving the goalposts with each generation.

Key takeaway: Expect a two-track future: Nvidia GPUs for flexibility and cutting-edge training, custom chips for huge, repetitive workloads.

PART IV

The Power

Why electricity, not silicon, is becoming the real limit.

CHAPTER 8

Power: the new binding constraint

In one minute

Data-centre electricity use rose 17 percent in 2025 while global demand rose about 3 percent. The IEA expects data-centre demand to roughly double by 2030, with AI-specific demand about tripling. Gas turbines, transformers and grid connections are the new bottlenecks.

On its February 2026 earnings call, Nvidia's management said that every data centre is power constrained. It is hard to overstate how much that changes the picture. For a decade, tech firms scaled by buying more servers. Now they have to become energy developers.

The numbers, from the International Energy Agency

- Data centres used roughly 460 terawatt-hours (TWh) of electricity in 2024, heading for over 1,000 TWh by 2030, about 3 percent of global electricity, comparable to Japan's entire usage today, according to the IEA.
- In advanced economies, data centres could drive over 20 percent of electricity-demand growth to 2030.
- The IEA warns that as many as one in five planned data centres could face delays from power and grid limits.
- Gas turbine orders now involve lead times of several years, which can push some projects beyond 2030.

Where will the power come from?

Source	Role in the buildout	Catch
Solar, wind and storage	The IEA expects renewables to supply roughly half of added data-centre demand; tech firms signed around 40% of corporate renewable power deals in 2025.	Intermittent. Needs batteries and grid upgrades.
Natural gas	Fast to deploy, widely used for on-site power in the US.	Emissions. Turbine shortages.
Nuclear	Tech firms are signing deals, including small modular reactors (SMRs).	Most SMRs only arrive around 2030 or later.
Geothermal	Emerging, steady, clean.	Small scale so far.

Because the grid is slow, many US developers now build **on-site generation** next to the data centre. The IEA expects roughly 15 to 27 GW of on-site gas to serve data centres by 2030. Another idea is making data centres "flexible" so they can reduce load when the grid is stressed.

> **Why this matters to ordinary people**
>
> Electricity prices, local water use, land and noise are all becoming political issues wherever big data centres are built. Expect more debates, regulations and community pushback over the decade, in India too.

Key takeaway: The cheapest, fastest route to AI leadership may not be the best chip but the nearest abundant electricity. Countries and states with spare power have a new bargaining chip.

PART V

The World

Politics, rivalry and India's place in the compute race.

CHAPTER 9

Chips as statecraft

In one minute

The US restricts which AI chips can go to China. Policy has swung back and forth, and Nvidia now says its China data-centre share is near zero while Huawei fills the gap. Controls slow China but also push it toward self-reliance.

Governments realised that AI chips are both economic engines and military assets. The US has tightened and loosened export rules several times. In January 2026, the US moved to case-by-case licensing for some Nvidia H200 and AMD chips headed to China, paired with a 25 percent federal surcharge, according to industry reports. Yet by 2026 Nvidia was still guiding for no China data-centre compute revenue, and its CEO said the company's direct share there had fallen to zero, arguing that the policy had backfired.

Meanwhile, Huawei is ramping its Ascend chips. Reports suggest its AI chip revenue could rise from about $7.5 billion in 2025 to $12 billion in 2026. A bottleneck for China is memory: its domestic HBM production reportedly supports only a few hundred thousand Ascend chips in 2026.

Two competing viewpoints

Supporters of tight controls say	Critics say
Compute is the key military and economic input. Do not hand rivals an edge.	Controls push China to build alternatives faster and shrink US firms' revenue and influence.
Loosening rewards the very regime the controls target.	The US should sell its technology widely so the world builds on its stack.

Reasonable people disagree, and the policy keeps changing, so check current news before relying on any specific rule. Whatever happens, chips will remain a geopolitical lever for the entire decade.

Key takeaway: Compute has become like oil in the 20th century: a resource countries compete for, restrict and weaponise.

CHAPTER 10

India's compute moment

In one minute

India has about 1.5 to 1.9 GW of data-centre capacity, tripling or more since 2020, with forecasts of roughly 5 to 7.5 GW by 2030. The IndiaAI Mission provides subsidised GPUs. India is data-rich but still compute-poor compared with the US and China.

India produces about a fifth of the world's data but holds only a small share of global AI compute. That gap is both a weakness and an opportunity.

Where India stands in 2026

- **Capacity.** The IT ministry reported around 1,575 MW of installed data-centre capacity in August 2026, more than four times 2020. KPMG puts FY26 capacity near 1.9 GW, with a pipeline taking it to about 7.0 to 7.5 GW by 2030. Other analysts forecast closer to 5 GW. Definitions differ, so use these as a range.
- **Public compute.** The IndiaAI Mission (outlay about 10,372 crore rupees) had expanded shared GPU capacity to over 45,000 GPUs by June 2026, offered to startups and researchers at subsidised prices.
- **Private gigawatt projects.** Reports describe planned AI "factories" by Reliance in Jamnagar, Google in Visakhapatnam and TCS, and Larsen and Toubro with Nvidia. Hyderabad, Mumbai and Chennai are key hubs.
- **Investment pledges.** After the AI Impact Summit in New Delhi in February 2026, large commitments were announced. Treat pledges as intentions, not installed capacity.

India's challenges

- Reliable, affordable, clean electricity at gigawatt scale.
- Water and cooling in hot regions such as Rajasthan and Gujarat.
- Access to the most advanced chips and memory, which are scarce worldwide.
- Moving from hosting other countries' AI to building Indian models and products.

For a country with India's talent base, the opportunity is real: build local-language AI, serve the home market, and offer reliable, lower-cost capacity. Rajasthan's abundant sun and open land also make it a natural place to watch for solar-backed data-centre projects.

Key takeaway: India will not out-spend the US in this decade, but it can win on talent, language, cost and applications, if power and chips keep up.

PART VI

The Road Ahead

Three scenarios for the next ten years, and what you can do about them.

CHAPTER 11

2026 to 2036: three scenarios

In one minute

Nobody knows the future. A useful way to think is in scenarios: the boom continues (bull), growth continues with a bumpy correction (base), or the cycle breaks (bear). Each leaves different winners and losers.

	Bull: compute supercycle	Base: bumpy growth	Bear: bust and digest
What happens	Agents and AI products create huge demand. Capex stays above $1.5T a year.	Spending keeps rising but a market correction hits weaker firms and leveraged builders.	AI revenue lags, credit tightens, projects are cancelled, capex falls sharply.
Chips	Nvidia and custom chips both grow strongly.	Nvidia share erodes slowly; memory stays tight.	Chip makers see falling orders; inventories pile up.
Power	Grid investment boom, nuclear and gas build-out.	Grid delays slow some sites; power-rich regions win.	Stranded assets; cheaper power for others.
Winners	Chips, memory, power, cooling, AI-native apps	Diversified, cash-rich firms	Cash-rich survivors buy distressed assets

A fair summary: most mainstream forecasts lean toward the base-to-bull range through 2028, because chips and memory are sold out and capex guidance keeps rising. But the further out you look, the less reliable any forecast becomes, and the financing risks in Chapter 4 are real.

Signposts to watch

- Quarterly capex guidance from Amazon, Microsoft, Alphabet, Meta and Oracle.
- Nvidia's guidance and whether orders are described as supply constrained.
- Memory prices and HBM contract news from SK hynix, Samsung and Micron.
- Corporate bond spreads and private-credit stress tied to data centres.
- Power: turbine lead times, grid-connection queues, nuclear milestones.
- Policy: US-China chip rules and any new global export framework.

Key takeaway: Prepare for a range of outcomes rather than betting your plans on one forecast.

CHAPTER 12

What it means for you

In one minute

You do not need to own a data centre to benefit from this decade. Individuals and small businesses can use cheap, powerful AI, build products on top of it, and learn the language of the shift. Be cautious about investing based on hype.

This chapter offers general ideas, not personalised advice. Whether you run a blog, a service business or a software project, the compute boom touches you in three ways.

1. Use the cheap layer

While chips and power are scarce for giants, the price of AI for users has been falling because each generation produces tokens more cheaply. A small team can now automate writing, research, customer support, translation and coding in ways that needed a large staff a few years ago. The advantage goes to people who learn to apply AI well, not to those who merely know it exists.

2. Build where compute is not the bottleneck

Because the physical layers are crowded, the best opportunities for small players are at the top: niche apps, local-language content, industry-specific tools, and services that help other businesses adopt AI. India's language diversity is a good example of a gap that global firms do not cover well.

3. If you invest, be humble

- Understand what you own. Stocks tied to AI can swing wildly.

- Avoid borrowing to buy a theme. Leverage turns a normal dip into a disaster.
- Diversify. Even true revolutions have produced losing investors.
- Beware anyone promising certainty. This book's scenarios exist because nobody can offer it.
- Consider speaking to a qualified, registered financial adviser before big decisions.

A 30-day starter plan

Week	Action
1	Read this book. Follow quarterly results from two companies (say Nvidia and one cloud giant) to see the numbers move.
2	Pick one daily task in your work and try an AI assistant for it. Record time saved.
3	Learn the free dashboards: IEA reports on energy and AI, and company investor pages.
4	Decide one small project that uses AI for a real audience, and ship it.

Final takeaway: The next ten years will reward people who understand the physical foundations of AI. You now know the vocabulary, the players and the pressure points. Keep learning, stay sceptical, and keep building.

REFERENCE

Glossary

Term	Meaning
Accelerator	A chip designed to speed up AI calculations, such as a GPU or TPU.
AI agent	An AI system that carries out multi-step tasks with some autonomy.
ASIC	Application-specific integrated circuit. A custom chip for one purpose.
Blackwell / Rubin	Nvidia's current and next-generation AI chip platforms.
Capex	Capital expenditure on long-lived assets such as data centres and chips.
CoWoS	TSMC's advanced packaging technology that joins processors and memory.
CUDA	Nvidia's software platform for programming its GPUs.
Foundry	A factory that makes chips designed by other companies (TSMC is the leader).
GPU	Graphics processing unit; the main AI chip type.
Gigawatt (GW)	One billion watts of power.
HBM	High-bandwidth memory; stacked memory placed beside AI chips.
Hyperscaler	A very large cloud provider, such as Amazon, Microsoft, Google, Meta or Oracle.
Inference	Running a trained AI model to produce answers.
Neocloud	A specialist cloud company renting out AI computing capacity.
SMR	Small modular reactor; a smaller, factory-built nuclear reactor design.

Term	Meaning
Token	A small unit of text that AI models read and write.
TWh	Terawatt-hour; one billion kilowatt-hours of electricity.
Training	Teaching an AI model by exposing it to large amounts of data.

REFERENCE

Cheat sheet: ten numbers to remember

Number	What it tells you	Source type
$96.2B	Nvidia revenue in the quarter ended July 26, 2026, up 106% year on year	Nvidia filing
$89.0B	Nvidia data-centre revenue that quarter	Nvidia filing
~$108B	Nvidia's guidance for the next quarter (plus or minus 2%)	Nvidia filing
$600B to $800B	Range of estimates for 2026 hyperscaler capex	Banks, analysts
~$1.0T to ~$1.6T	UBS capex projection, 2026 to 2028	UBS via press
102%	Amazon, Alphabet and Microsoft capex as share of cloud revenue in 2026	UBS via press
460 to 1,000+ TWh	Global data-centre electricity use, 2024 to 2030	IEA
17%	Growth in data-centre electricity demand in 2025 (vs 3% overall)	IEA
~1.5 to 1.9 GW	India's installed data-centre capacity (varies by source)	Ministry, KPMG
45,000+	GPUs in India's shared public compute pool, June 2026	IT Ministry

REFERENCE

Sources and further reading

The figures in this book were drawn from the following kinds of public sources, accessed in early October 2026. Search for these titles to read the originals.

- NVIDIA, press releases and Form 8-K filings: fourth quarter and fiscal 2026 results (Feb 25, 2026) and second quarter fiscal 2027 results (Aug 26, 2026), on SEC EDGAR and investor.nvidia.com.
- International Energy Agency (IEA), *Key Questions on Energy and AI* (2026) and related coverage of data-centre electricity demand.
- UBS hyperscaler capex projections, as reported by Yahoo Finance (2026): "AI's Absurd Spending Boom? Hyperscalers Are Spending 102% of Cloud Revenue on Capex."
- Goldman Sachs AI capex estimates, as summarised by the IO Fund (2026): "AI Capex to Hit $1 Trillion."
- Epoch AI analysis of hyperscaler capex versus operating cash flow, as summarised in Q3 2026 data-centre coverage.
- Capital Group, "How to spot an AI bubble" (2026), on AI debt issuance and vendor financing.
- Tom's Hardware, "The custom AI ASIC state of play" (May 2026), on Broadcom, Google TPU and Meta MTIA.
- Trade and industry reporting on HBM supply from SK hynix, Micron and Samsung, including Q3 2025 earnings-call commentary and July 2026 industry coverage.
- Reporting on US export policy for AI chips to China (January 2026 licensing changes) and Nvidia's comments on its China share.

- Ministry of Electronics and Information Technology (India) statement on data-centre capacity and IndiaAI Mission GPUs (Aug 2026); KPMG India data-centre report (Jul 2026); Avendus Capital (May 2026).

Verify before you rely

Some figures above come from analyst blogs or trade newsletters that summarise other sources. Where a number matters for money, legal, or publishing decisions, find the original filing or report and confirm it.

The decade is still being written.

Money builds the machines, chips give them power, and power makes them run. Understand those three, and you will understand most of what comes next.