FT Alphaville Artificial intelligence

The next wave of AI hype will be geopolitical. You're paying

The wealth of hallucinations

© FTAV montage / Frinkiac

Bryce Elder MAY 29 2024

A popular view of generative AI is that it's unjustifiably expensive, chronically wasteful, rarely useful, and is being foisted on the general public for ideological reasons even though it makes the services they rely on worse. Governments are sure to be all over it.

So says Barclays:

> The initial wave of AI is well under way, largely driven by billions of hyperscale dollars. The concern with Nvidia, and ultimately with the second wave of AI ecosystem, is where the next pocket of dollars comes from once hyperscale capex can't shift further toward AI or grow meaningfully year-on-year.
>
> In recent months, we have seen a growing initiative from nations around the world to quickly educate themselves and remain at the forefront of AI's powerful potential. In a practical sense, this has amounted to public announcements for multi-hundred million and even billion-dollar spending plans from several nations (Saudi Arabia, Singapore, Germany, UK, India), which will go toward supporting the AI hardware ecosystem.

Private-sector AI cash burn is already sovereign-sized. The combined capex of Amazon, Meta, Google and Microsoft will be around $200bn this year, according to Bernstein Research.

These sunk costs need to deliver some kind of return by the time the depreciation charges reach their income statements. A continued acceleration of capex growth depends on companies finding something that the public wants to buy. Their need for revenue may soon become urgent, and concepts so far have not been encouraging.

But because political leaders care more about one-upmanship than ROIC, taxpayer-subsidised AI can continue to boom even if the corporate bubble bursts.

Barclays estimates in a recent note that if nations ex-China matched US's $4bn of AI spending, scaled to GDP, it'd add another $3.5bn:

Sovereigns	GDP ($Bs)	AI Spend ($Ms)
USA	28,781	4,000
Germany	4,591	638
Japan	4,110	571
India	3,937	547
UK	3,495	486
France	3,130	435
Canada	2,242	312
Australia	1,790	249
South Korea	1,761	245
TOTAL		$7,482

For Nvidia, that adds up to barely a month's revenue. The more important thing is the replacement cycle.

Barclays estimates that hardware purchases will become obsolete within two years. As AI hardware becomes bigger and more expensive, aggregate annual sovereign spending can easily exceed $25bn very quickly:

> Overall, we see AI as the most powerful enabler of technological progress, as well as a major security risk as adversarial nations increase capabilities, ultimately justifying our estimated spend and giving us confidence that the numbers should move materially higher.

The US has taken an early lead because its government has been relatively enthusiastic about AI. There was a Federal Use Case Inventory published in September that identifies more than 700 possible applications, and a Senate Roadmap for Artificial Intelligence Policy earlier this month proposed an R&D budget of $32bn.

Though such figures may prove fanciful, the bigger costs are much less scrutinised. The Senate road map does not include defence, which appears to account for nearly all current US federal AI spending.

A study of government tenders published by the Brookings Institution in March found that the US Department of Defense was ramping up AI investment aggressively in 2022. By maximum potential contract value, just over $4bn of the $4.56bn in AI procurement costs last year were for the defence agency, Brookings calculates.

Senate majority leader Chuck Schumer has said the US AI defence budget needs to increase by approximately eightfold. The exact purpose of all this investment will remain classified information.

Countries following the America's lead will want something built in-house that's at least equivalent to OpenAI's GPT-4, says Barclays. Last year's best technology represents "the minimum starting point for nations attempting to remain at the forefront of AI for both economic and security purposes".

Processor blades for such a rig costs will cost $600mn at current prices, plus the same again to cover interconnects, storage, power costs, etc. What such a set-up won't do is scare the enemy. That requires staying at AI's bleeding edge, which will be a *lot* more expensive.

Training GPT-4 is said to have used 25,000 accelerator cards, whereas GPT-3 — released less than three years earlier — needed just 1,000. The below grid gives an approximate idea of current all-in build costs in increments of ten thousand accelerators, or XPUs.

		XPU ASP ($k)				
	($Ms)	25	24	23	22	21
Cluster size (k XPU)	10	250	240	230	220	210
	20	500	480	460	440	420
	30	750	720	690	660	630
	40	1000	960	920	880	840
	50	1250	1200	1150	1100	1050

Projected AI Accelerator Spend by Cluster Size © Barclays

If hardware cost inflation continues at the current pace, the cost of one best-in-class AI computing cluster could easily exceed $5bn, says Barclays:

		XPU ASP ($k)				
	($Ms)	38	36.5	35	33.5	32
Cluster size (k XPU)	25	950	913	875	838	800
	50	1900	1825	1750	1675	1600
	75	2850	2738	2625	2513	2400
	100	3800	3650	3500	3350	3200
	500	19000	18250	17500	16750	16000

Projected Next-Gen AI Accelerator Spend by Cluster Size © Barclays

The infowars arms race will escalate so quickly, only about 15 nations can afford to take part, says Barclays. And for those able to pay there's no option to back down, it says, because "AI capabilities have become one of the most important, if not the most important, national initiatives globally":

> In our view, global development of AI applications will undoubtedly become a national security issue no different than how the government views domestic leading-edge chip production, and under the lens of the ~$39bn CHIPS Act passed several years ago, we see adequate room in the government budget for increased spending on new clusters and more advanced hardware once readily available.
>
> Furthermore, we believe new AI/compute investment plans put forth by Saudi Arabia ($40bn AI investment fund according to the New York Times), Singapore, Germany, and even India could push policymakers to act sooner rather than later on writing more robust AI investment plans into policy

So buy Nvidia, Barclays tells clients. The stock might look expensive, with an attached heap of risks related to sanctions and antitrust, but civil servants won't know any better than to buy servers off the shelf:

> We see NVDA as the largest beneficiary of Sovereign AI given its already dominant share of the merchant AI accelerator installed base, performance leadership, and preference from the developer community. We also see the Sovereign AI market as a strong potential adopter of the company's full rack solution [. . .] given government agencies' lack of engineering know-how and resources required to assemble custom solutions around merchant hardware. Overall, we view the projected spend from Sovereign AI as additive to the entire AI ecosystem, and thus believe it will trickle down to the broader AI ecosystem as well.

And sure. Why not. Once a trailing PE goes above 300x, anything goes.

Being an ESG darling was a big part of last year's Nvidia buy case. A year earlier it was adjacent to the shitcoin bubble, and before then it was mostly about *Cyberpunk 2077* frame refresh rates. Now it's a buy because it's the de facto weapons supplier for World War GPT.

One common thread that links Nvidia's customers and shareholders is that they don't know what they are buying, or why they need it, but are sure they have to have it. An international arms race for billion-dollar boondoggles and tiger-repelling rocks would fit this description perfectly.

Further reading:
— 'Sell Nvidia' (FTAV)