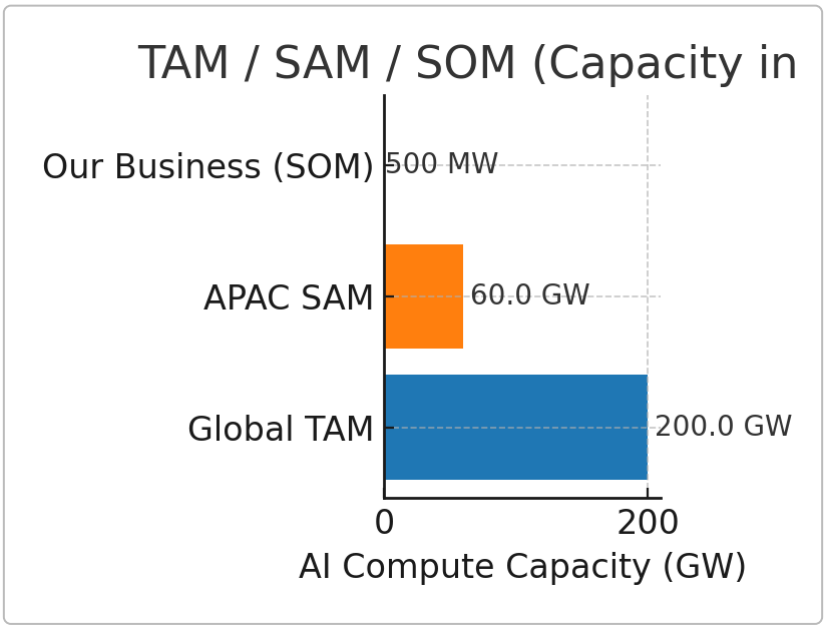
ChatGPT

AI GPU Compute Infrastructure Market Analysis

TAM, SAM, and SOM Overview



Global TAM (Total Addressable Market), regional SAM (Serviceable Available Market in India/APAC), and potential SOM for the proposed business (initial 3 MW scaling to 500 MW). The global demand for AI compute is enormous – projected to reach ~200 GW of data center capacity by 2030 [1] (up from ~60 GW in 2023 [1]) and a market value well over $200 billion [2]. Even a 500 MW facility would constitute only a fraction of a percent of worldwide AI capacity. This highlights the sizable headroom for growth: in a TAM of hundreds of gigawatts globally, the SAM in India/APAC is on the order of tens of gigawatts, and a 500 MW SOM captures only a small slice [1] [3].

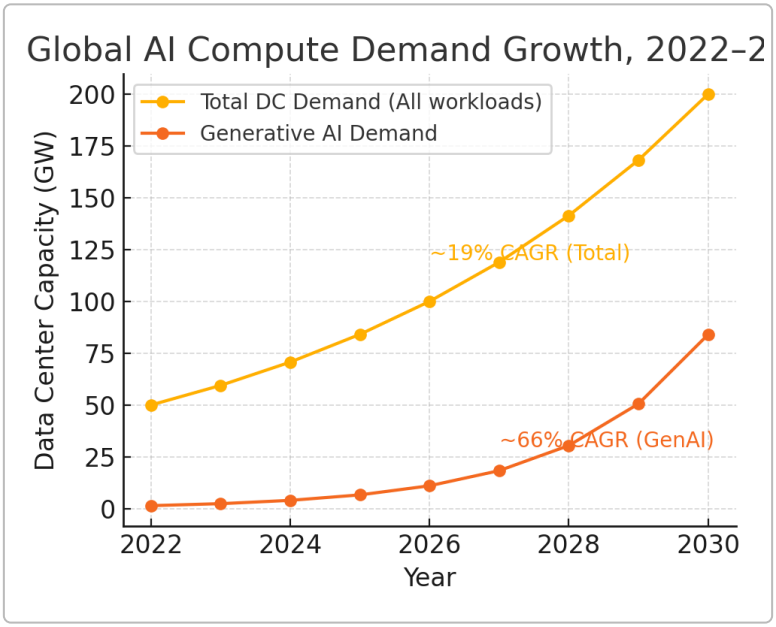
Global TAM (Total AI Compute Market): The worldwide market for AI GPU computing is expanding explosively. In terms of **power capacity**, total data center demand (all workloads) is expected to **more than triple** from ~60 GW in 2023 to roughly **171–219 GW by 2030** in a midrange scenario [1]. Crucially, **advanced AI workloads** (training and inference on GPUs/accelerators) will drive ~70% of that capacity by 2030 [4]. In value terms, the **global AI infrastructure market** (hardware, software, services) was ~$35.4 billion in 2023 and is forecast to reach **$223 billion by 2030** (30.4% CAGR) [2]. Some analysts predict even larger TAM figures as the AI boom accelerates – for example, AMD's CEO revised her estimate for **data center AI accelerators** from ~$30 billion in 2023 to over **$400 billion by 2027** [5]. Likewise, the data center GPU hardware market is projected at **$228 billion by 2030** (up from ~$120 billion in 2025) [6]. In short, the total global opportunity for AI compute is measured in the **hundreds of billions of dollars** and continues to expand rapidly alongside AI adoption.

India/APAC SAM (Serviceable Available Market): The addressable market in India and the broader Asia-Pacific is one of the fastest-growing segments of this global TAM. Asia-Pacific's AI infrastructure demand is surging, supported by a large digital user base and government initiatives [7] [8]. China alone accounted for ~35.6% of APAC's AI infrastructure spending in 2023 [9], and India is anticipated to have the **highest growth rate** in the region going forward [8]. India's data center capacity (all uses)

expanded from **~540 MW in 2019 to ~1,011 MW in 2023** [10], and a recent analysis projects an **additional 500 MW by 2028** driven largely by AI workloads [3]. This implies India's installed capacity could exceed ~1.5 GW around 2028, double the 2023 level. Major investments underpin this growth – e.g. Yotta and NVIDIA partnering on an AI-centric data center in Gujarat, and CtrlS committing $2 billion for 350 MW of AI-ready capacity [11]. In fact, Reliance Industries (Mukesh Ambani) has announced plans for the **world's largest AI data center** in Gujarat, targeting **3 GW** (3,000 MW) of capacity – far above today's largest sites (~600 MW) – at an estimated $20–30 billion cost [12] [13]. This underscores the vast **SAM** in India/APAC: on the order of **tens of gigawatts** of AI data center capacity and tens of billions of dollars in revenue potential in the coming decade.

Serviceable Market (Proposed Business SOM): For the proposed data center project scaling from **3 MW to 500 MW**, the attainable market share is modest relative to the massive TAM/SAM. A 3 MW initial deployment is a drop in the bucket globally. Even at the full **500 MW** scale, it would represent only about **0.25%** of projected global AI data center capacity (~0.5 GW out of ~200 GW) and perhaps on the order of **0.5–1%** of the Asia-Pacific market by 2030 [1]. However, within India such a capacity is significant – 500 MW is roughly half of the new AI-driven capacity India is expected to add by 2028 [3]. Capturing this SOM would mean the business becomes a leading regional player, while still being a small fraction of the **global** demand (which is good news in terms of room to grow). In sum, the funnel is extremely wide at the top: a vast global TAM, a rapidly expanding APAC/India SAM, and a feasible SOM that, at 500 MW, could generate substantial revenue yet still only meet a tiny portion of overall AI compute needs.

AI Compute Demand Growth (2022–2030)



Global AI compute demand is on a steep growth curve. Total data center capacity (all workloads) is projected to grow ~3–4× from 2022 to 2030 (~19% CAGR), while the subset for AI workloads (especially generative AI) is growing much faster – on the order of 60%+ CAGR in current scenarios. The chart illustrates a mid-range scenario where global data center demand rises from ~50 GW in 2022 to ~200 GW in 2030 [1], and the portion attributable to generative AI (orange line) surges from near-zero to ~84 GW by 2030 [14].

The demand for AI GPU compute is experiencing **exponential growth**, fueled by the rise of deep learning, generative AI models, and cloud AI services. In 2022, AI workloads were a relatively small part of data center usage, but this changed drastically by 2023–2024 with the advent of large-scale

generative AI (e.g. ChatGPT-like services). **Global data center capacity** (all uses) grew from ~50 GW in 2022 to about **60 GW in 2023** [1]. Looking ahead, forecasts suggest total capacity will reach **~200 GW by 2030** in a mid-case scenario [1] – an ~19% annual growth rate. Critically, **AI-focused compute** is the engine of this growth. McKinsey estimates that ~**70% of data center capacity in 2030 will be dedicated to AI** workloads [4], up from a much smaller share today. Generative AI (e.g. large language model training and inference) is the fastest-growing segment, expected to account for ~**40% of total data center capacity by 2030** [4]. This means roughly *80 GW out of 200 GW* serving generative AI by 2030, a stunning rise from virtually negligible levels just a few years ago.

Multiple analyses reinforce how steep the **AI demand curve** is. One scenario (from CSIS) projects **global generative AI computing demand jumping from ~4 GW in 2024 to ~84 GW in 2030** [14] – an annual growth rate of ~66%, far outpacing general IT growth. Another study suggests that by 2029, running generative AI could consume about **1.5% of all electricity in the world** [15] (up from a tiny fraction in 2022), highlighting the rapid increase in GPU-powered workloads. In power terms, 1.5% of global electricity equates to on the order of 40–50 GW of continuous demand – an astonishing figure for a single category of compute [15]. Even more conservative forecasts show huge growth: Deloitte cites Lawrence Berkeley Lab research predicting **US data center power demand** (driven by AI) will roughly **double from ~40 GW in 2024 to 74–132 GW by 2028** [16]. Overall, the **demand for AI GPU compute (in MW)** is growing ~**30–35% yearly in baseline scenarios, and potentially much faster (50–70%+) in aggressive adoption scenarios** [1] [14].

This surging demand is predominantly for **generative AI and "neo-AI" cloud services** – i.e. the new wave of AI models for language, vision, and decision support. Training these models and serving inference at scale requires massive compute-hours. For example, Meta's latest Llama 3 (70B parameter model) required **6.4 million H100 GPU-hours** just for training [17]. Bloomberg's BloombergGPT model reportedly used **1.3 million GPU-hours** in training [18]. As enterprises and cloud providers race to deploy such models, the aggregate GPU time being consumed is exploding. By our estimates, global generative AI workloads in 2030 could be utilizing **trillions of GPU-hours per year**, up from only billions of GPU-hours in the early 2020s. This is consistent with the jump in power demand: ~84 GW of AI load in 2030 (if realized) would equate to on the order of *10^12 GPU-hours* annually. In financial terms, **AI compute spend** is rising commensurately – one analysis found the **data center GPU market ballooned from ~$17 billion in 2022 to ~$125 billion in 2024** [19] as hyperscalers snapped up accelerators for AI, and it's projected to continue growing strongly (e.g. ~$228 billion by 2030) [6].

Market Value and Capacity Trends

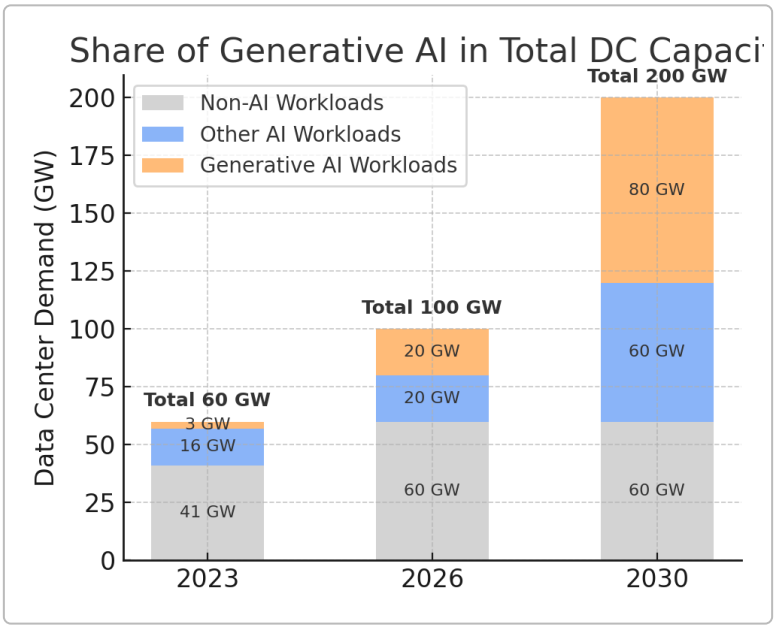
To support these projections, we see clear evidence from market data and industry investments:

- **GPU Shipments & Spending:** The volume of AI accelerator hardware being deployed is skyrocketing. In 2023, an estimated **5.5 million data center GPUs** were shipped (nearly double the 2022 volume), driving about **$37 billion** in GPU sales [20]. Forecasts show continued growth – by 2027, annual shipments are expected to reach **~13.5 million units**, with data center GPU revenues around **$95 billion** [20]. NVIDIA currently dominates this space with ~90%+ market share [21]. This surge reflects the scramble by cloud providers and enterprises to secure AI computing capacity. Nvidia's CEO noted that demand is effectively "unlimited" in the near term for AI GPUs, and AMD has similarly pointed to unprecedented growth in the accelerator TAM [5].

- **Hyperscaler and Cloud CAPEX:** The hyperscale cloud firms (AWS, Microsoft, Google, etc.) and large AI labs are pouring capital into AI infrastructure. For example, a consortium of Oracle, SoftBank, and OpenAI announced project "**Stargate**," aiming to invest **$100–$500 billion** in new

AI supercluster data centers in the U.S. [22]. Microsoft, Google, and Meta each have multi-billion-dollar programs to expand AI computing clusters (often in partnership with NVIDIA for H100 GPUs). In India, Reliance's planned **3 GW AI data center** (mentioned above) is a $20+ billion endeavor [23] – a single project that would dwarf the country's current total capacity. These investments indicate that major players foresee a **continuing boom in AI compute demand** and are racing to build out capacity to meet it.

- **Generative AI Services Growth:** On the demand side, the proliferation of generative AI services (GPT-based cloud APIs, AI-enhanced software, etc.) is driving cloud usage to new highs. AWS and Azure have reported significant uptake of their AI instances; Nvidia's data center revenue (largely AI GPUs) jumped ~4x year-on-year in 2023 due to this trend [19]. Enterprise adoption of AI is also accelerating – IoT Analytics reports the combined **generative AI software/platform market** grew from ~$0.2 billion in 2022 to $25.6 billion in 2024 [24], which in turn fuels more demand for GPU compute on the back-end. All of this suggests the **monetary value of AI compute** (TAM in USD) will track the capacity growth, with estimates of **hundreds of billions of dollars by 2030** in cloud AI services, hardware, and supporting infrastructure.



The portion of data center capacity devoted to AI is rapidly increasing. In 2023, generative AI was only a small fraction of workloads (orange segment), but by 2030 it could be ~40% of total capacity [4]. This illustrative chart shows total data center demand (stack height) growing from 60 GW in 2023 to 200 GW in 2030, and the share of AI (blue + orange) expanding dramatically. Non-AI workloads (grey) become a minority by 2030 as AI — especially generative AI (orange) — dominates new capacity additions.

In summary, **the AI GPU compute infrastructure market is experiencing unprecedented growth** across all metrics: total power (MW) deployed, number of GPUs, and market value in USD. The **TAM** is global and immense, set to reach into the hundreds of GW and hundreds of billions of dollars this decade. The **SAM** in India/APAC is likewise growing fast, propelled by digital transformation and government support, making the region a key battleground for AI cloud expansion. And for the proposed business, even capturing a **SOM** of a few hundred MW will translate into a healthy enterprise with ample runway, given the overall demand far exceeds current supply. All data and forecasts point to a sustained **demand growth curve** for AI compute (especially generative AI) through 2030, reinforcing the opportunity for new AI-focused data center capacity in India. With credible sources ranging from market research firms to hyperscaler disclosures and analyst forecasts backing these trends [1] [3] [14]

[5], the business plan can be grounded in a robust understanding that the **AI compute boom** is not a short-term spike but a long-term, secular increase in the infrastructure needed to power the AI revolution.

Sources: Key market forecasts and data have been drawn from McKinsey's analysis of data center capacity trends [1] [4], industry research reports (Grand View Research, MarketsandMarkets) on AI infrastructure value [2] [6], hyperscaler and chipmaker insights (AMD, Nvidia) on AI hardware TAM [5] [20], and region-specific studies such as the Avendus report on India's data centers [3]. Analyst commentary on GPU shipments [20] and power consumption projections [14] [15] further validate the explosive CAGR of AI compute demand. All these sources consistently indicate a **massive and rapidly growing market** for AI GPU computing, with particularly strong momentum in generative AI and cloud AI services. The data has been compiled into charts and projections above to support the business plan with a clear, visual narrative of the TAM/SAM/SOM and growth trajectory. [1] [3] [14] [19]

---

[1] [4] AI data center growth: Meeting the demand | McKinsey
https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

[2] [7] [8] [9] AI Infrastructure Market Size, Share & Growth Report, 2030
https://www.grandviewresearch.com/industry-analysis/ai-infrastructure-market-report

[3] [10] [11] AI-led demand for Data Centres to add 500 MW of capacity by 2028
https://www.techcircle.in/2024/08/21/ai-led-demand-for-data-centres-to-add-500-mw-of-capacity-by-2028/

[5] [20] [21] Ongoing Saga: How Much Money Will Be Spent On AI Chips?
https://www.nextplatform.com/2024/07/15/ongoing-saga-how-much-money-will-be-spent-on-ai-chips/

[6] Data Center GPU Market worth $228.04 billion by ... - Barchart.com
https://finbets.websol.barchart.com/?module=topNews&storyID=32420245&symbol=&selected=news

[12] [13] [22] [23] India to get world's largest data center in AI push
https://coingeek.com/india-to-get-world-largest-data-center-in-ai-push/

[14] [16] The AI Power Surge: Growth Scenarios for GenAI Datacenters Through 2030
https://www.csis.org/analysis/ai-power-surge-growth-scenarios-genai-datacenters-through-2030

[15] Data center sustainability | Deloitte insights
https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/genai-power-consumption-creates-need-for-more-sustainable-data-centers.html

[17] NVIDIA Sets New Generative AI Performance and Scale Records in MLPerf Training v4.0 | NVIDIA Technical Blog
https://developer.nvidia.com/blog/nvidia-sets-new-generative-ai-performance-and-scale-records-in-mlperf-training-v4-0/

[18] Bloomberg used 1.3 Million GPU hours and 600 Billion documents ...
https://www.fanaticalfuturist.com/2023/05/bloomberg-used-1-3-million-gpu-hours-and-600-billion-documents-to-train-bloomberggpt/

[19] [24] The leading generative AI companies
https://iot-analytics.com/leading-generative-ai-companies/