TIRIAS RESEARCH

Forecasting the Growth of Generative AI Usage, Compute Requirements, and Infrastructure Costs

Forecast Growth of AI Usage 2024 to 2028

How fast will demand for core generative AI (GenAI) services grow, and what is the corresponding requirement for computing infrastructure?

151x The forecast growth in **Token Generation** and its use in emerging applications from **Large Language Models by 2028**

167x The forecast growth in **Image Generation**, with about half of image generation from the creation of video by 2028

$84 Billion Forecast Total Operating Costs in 2028 Growing from $1.75 Billion in 2024

The raw cost of inference processing supporting generative AI services. Includes amortized servers, power & cooling, electricity, operations; does not include software or facility construction costs, and does not include network training.

It starts with investments in training neural networks, setting network parameter values, and running the models with the learned parameters to provide services. Consumers and businesses then consume the outputs of these models – words, images, video, sound and ultimately fusions of models to create ever increasing levels of capability.

Global GenAi Output (Billions)	2023	2024	2028	2023 vs. 2024	2023 vs. 2028
Images + Video Frames	15	59	2,500	4x	167x
Tokens	6,900	19,900	1,034,000	3x	151x

Forecast TCO of GenAI Inference – At a Glance

The amortized cost of servers (4 years) plus power and cooling hardware (8 years) contribute to capital cost. Power and data center operations contribute to the operating costs.

Additional costs not incorporated include building costs, software, and other costs associated with operating GenAI services.

	2024	2028
Total Cloud GPU & Lean TPU Servers	15,144	737,644
Total Amortized Server Capital Cost (Sum)	$ 1,556,447,251	$ 75,818,490,358
Total Power & Cooling Hardware Cost	$ 56,507,131	$ 2,752,555,651
Total Server Power Cost	$ 109,220,114	$ 5,320,665,459
Total Server Operations	$ 2,825,357	$ 137,627,783
FORECAST TCO, TODAY'S DOLLARS	$ 1,174,999,852	$ 84,029,339,250

Usage and Neural Network Complexity Drive Demand For Compute Infrastructure

The amount of computing work per token and per image is expected to increase every year as neural network models grow in complexity. Even as service providers work to optimize neural network size and improve computing efficiency, the Forecast Total Operating Cost (FTCO) increases dramatically over time as we move toward billions of users and everyday usage of GenAI-driven services.

Global GenAI TCO Forecast	2024	2025	2026	2027	2028
Accelerators: Required Cloud GenAI Accelerators for Service Operations e.g Inference	362,000	1,322,000	2,916,000	8,376,000	17,638,000
Power: Total Data Center GenAI KWh	1,375,000,000	5,016,000,000	11,058,000,000	31,794,000,000	66,978,000,000
Forecast Total Operating Costs - Includes Amortized Servers Power, Operations	$ 1,725,000,000	$ 6,285,000,000	$ 13,837,000,000	$ 39,859,000,000	$ 84,029,000,000

Moving 20% of the GenAI workload to the edge would save $ 16 billion dollars in 2028

GenAI inference will scale creating massive incentives to distribute workloads to edge devices

By 2028, Cloud GenAI power consumption is forecast to rise over 66 billion KWh

For perspective, cloud-based GenAI by 2028 is anticipated to consume the same power annually as 19 billion flagship smartphones

Powerful servers, operating in public or private clouds, will be necessary for larger neural networks requiring large amounts of memory and computing performance. However, smartphones and PCs can also make a dent in the workload, taking on the processing load for smaller and more specialized models.

The Trias Research GenAI FTCO Model Forecasts Demand, Compute Requirements, Server Compute Capacity, and TCO

Demand Forecast

2023 MAU's & usage estimated utillizing validated with multiple public source & interviews

NN Compute Requirements Forecast

Projected cost, technology, and demand for more capable GenAI services

GPU/TPU Server Capacity Forecast

Internal bechmarks are validated gainst public benchmark data & interviews. Performance gains are countered by increasing complexity

Capital & Operating Cost Forecast

Forecasts capital cost for configured servers, power and cooling infrastructure. Forecast operating costs including data center operating costs, power costs

GenAI FTCO Forecast Overview

Trends that intersect or are direct implications of the pace of growth of cloud GenAI

What is the GenAI FTCO model?

- The forecast of the total operating costs of hardware running GenAI services in the cloud
- Includes GenAI inference or running of models, not model training e.g. the forecast includes operations, not R&D
- Today, looking at accelerator sales vs. the needs of GenAI services, training and forward-looking buildout dominates the use of accelerators being sold today, but this is expected to flip as operating those services at scale outpaces the requirements of training and growth, still expected to remain high, normalizes

What are the major factors driving cloud GenAI operating costs?

- Projected proliferation of useful GenAI application services driven by academic and corporate R&D
- Demand by businesses and consumers for these services as they come to market
- The total operating costs of the hardware running these services in the cloud

For more information, visit **www.tiriasresearch.com** or email **simon@tiriasresearch.com** TIRIAS RESEARCH