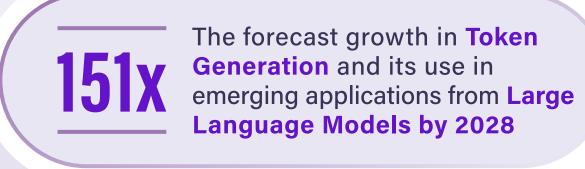


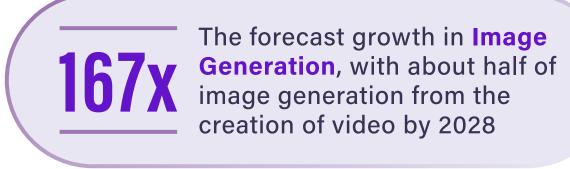
Forecasting the Growth of Generative Al Usage, Compute Requirements, and Infrastructure Costs



Forecast Growth of Al Usage 2024 to 2028

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







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 GenAl 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.

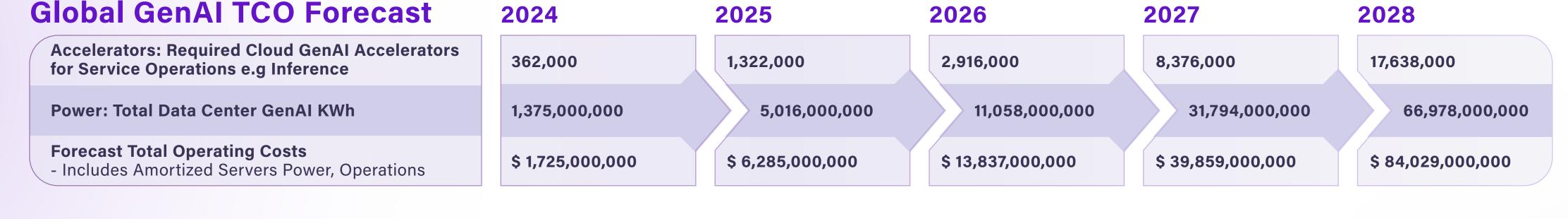
Additional costs not incorporated include building costs, software, and other costs associated with operating GenAl services.

	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 GenAl-driven services.





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



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



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



consume the same power annually as 19 billion flagship smartphones

For perspective, cloud-based GenAI by 2028 is anticipated to



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.

Powerful servers, operating in public or private clouds, will be





The Trias Research GenAl 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

GenAl services

GPU/TPU Server Capacity Forecast Internal bechmarks are validated gainst public benchmark

data & interviews. Performance gains are countered by increasing complexity

Forecasts capital cost for configured servers, power and

Capital & Operating Cost Forecast

cooling infrastructure. Forecast operating costs including data center operating costs, power costs



GenAl FTCO Forecast Overview

of the pace of growth of cloud GenAl

Trends that intersect or are direct implications

What is the GenAl FTCO model?



• 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 forecast of the total operating costs of hardware running GenAI services in the cloud

• 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 GenAl operating costs?

• Projected proliferation of useful GenAl 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