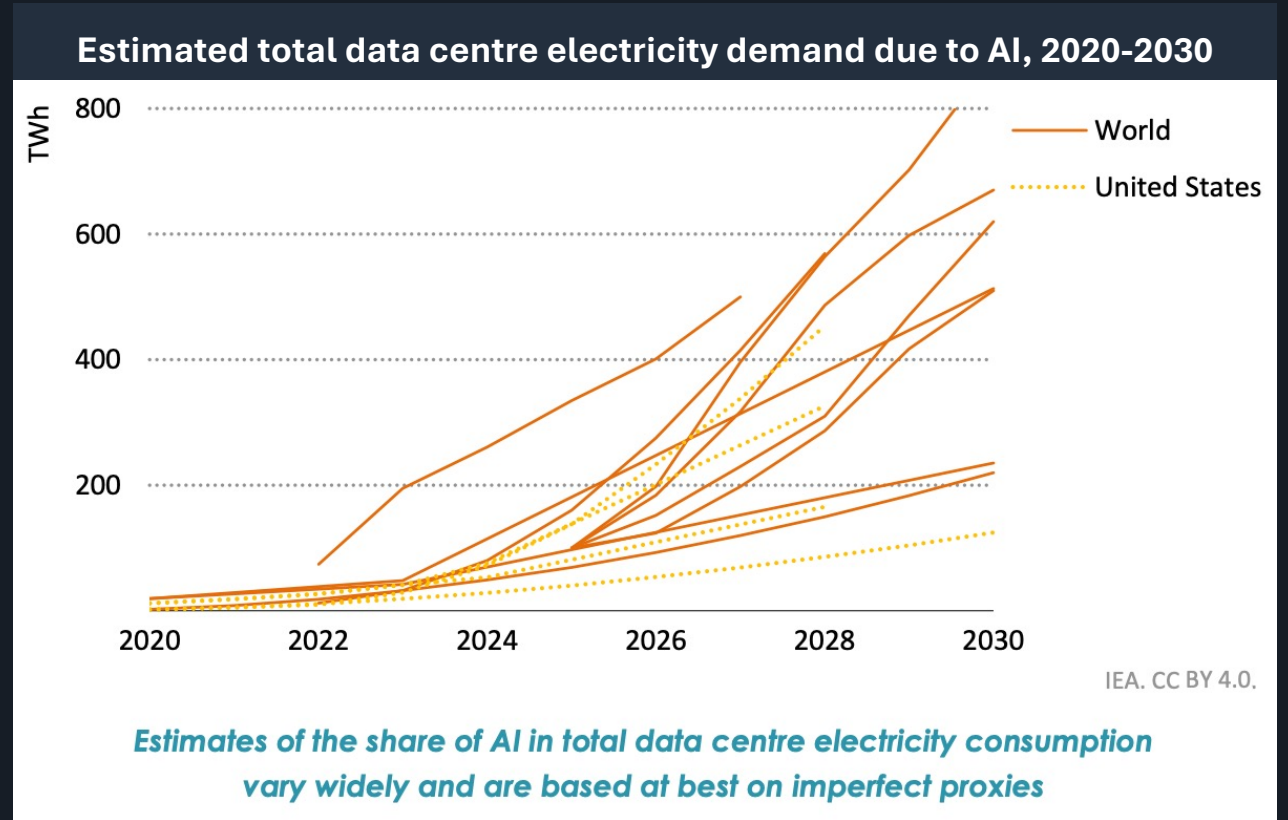
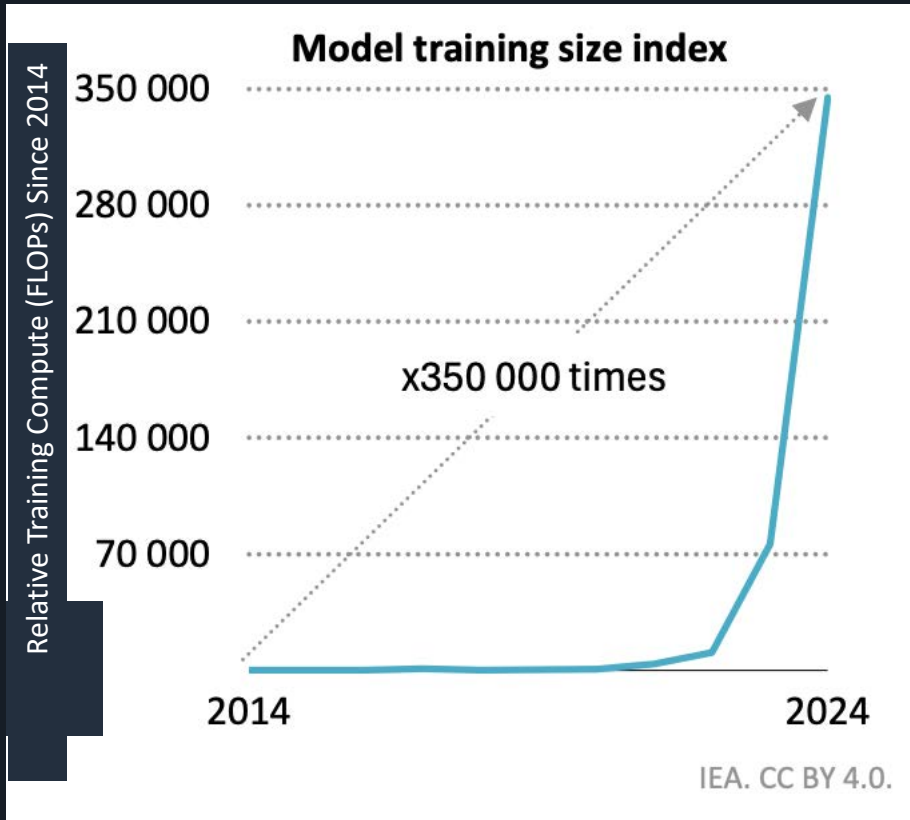


Building GenAI that doesn't cost the earth

Mohamed Najaaf
Solutions Architect
Amazon Web Services

AI models are getting bigger... and resources consumption is increasing



Goldman Sachs Research forecasts that about 60% of the increasing electricity demands from data centers will be met by burning fossil fuels, which could increase global carbon emissions by approximately **220 million tons**.

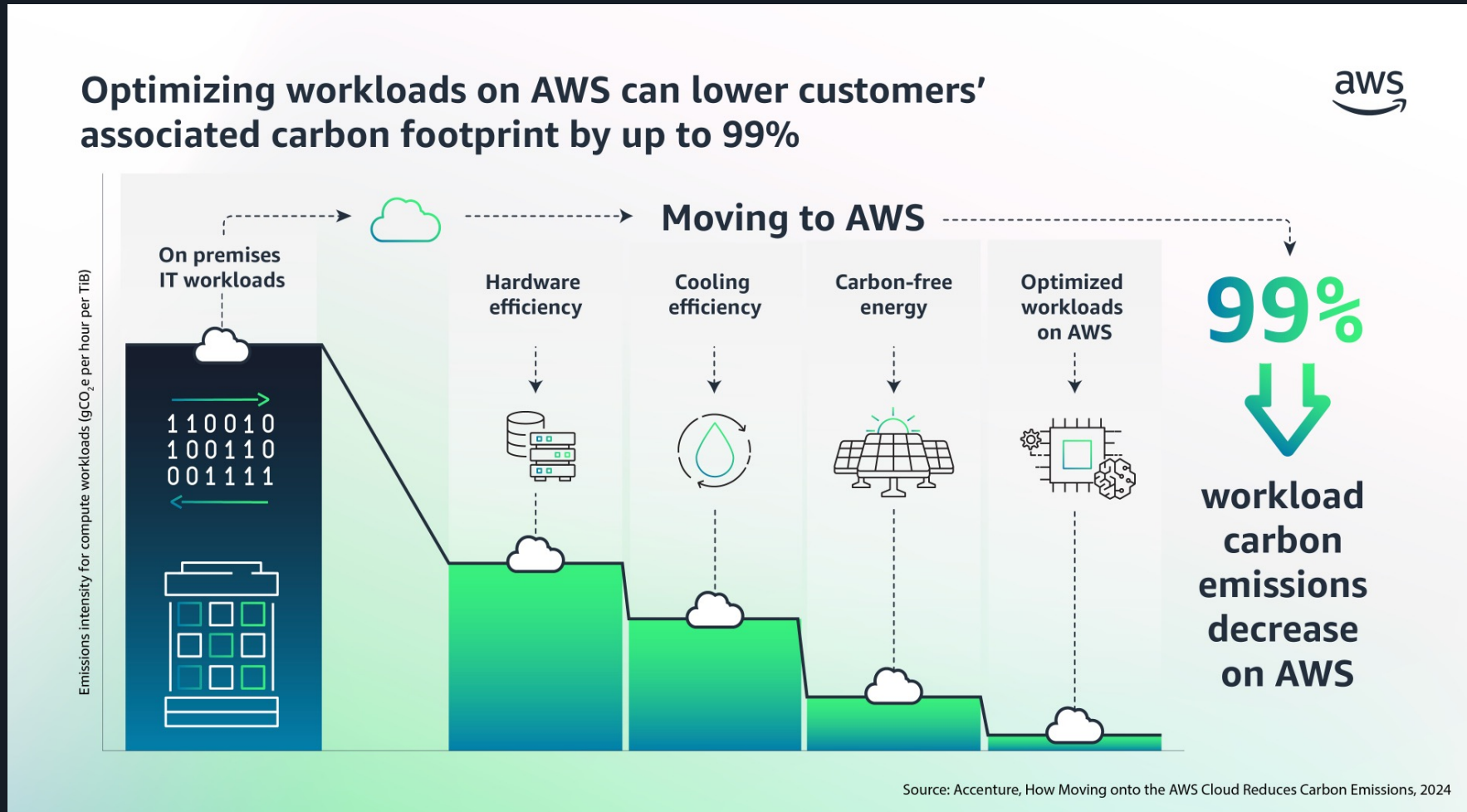


A 2024 article from the **World Economic Forum** notes that the computational power required for AI is doubling approximately every **100** days.



Carbon Reduction Opportunity

Optimizing workloads on AWS can lower customers' associated carbon footprint **by up to 99%**



Find the report at https://sustainability.aboutamazon.com/carbon_reduction_aws.pdf



Managed Services



Amazon SageMaker Model Training



Amazon SageMaker HyperPod



Amazon EKS

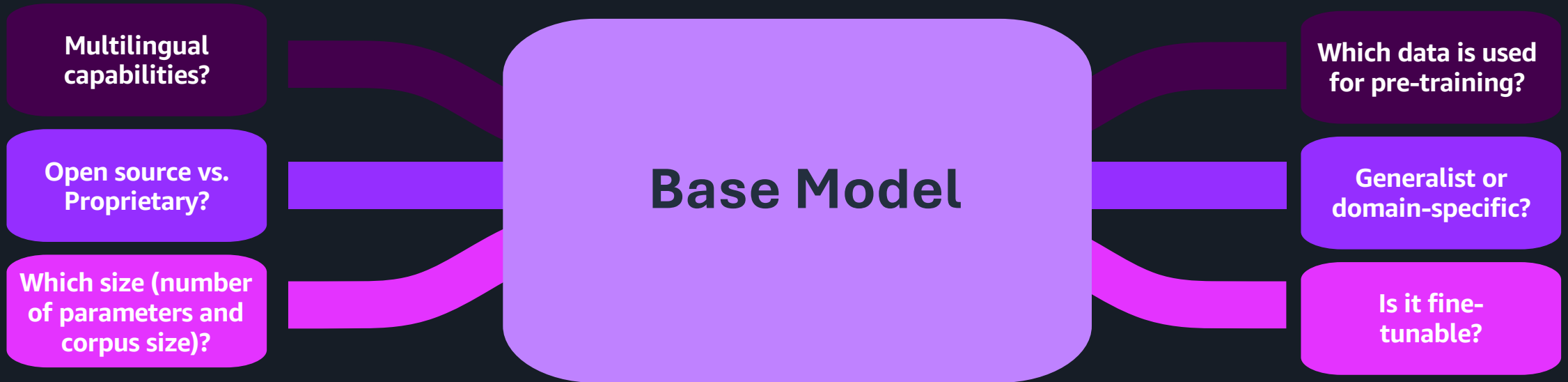


Amazon Bedrock



Amazon Bedrock AgentCore

Model Selection



Amazon Bedrock Evaluations

Amazon Bedrock > Evaluation

Evaluation Info

Models | **Knowledge Bases**

Knowledge Base evaluation Info

Assess the performance or effectiveness of your Knowledge Base using LLMs. [Learn more](#)

► How it works

Knowledge Base evaluations (6)

<input type="checkbox"/>	Evaluation ...	Status	Evaluated res...
<input checked="" type="checkbox"/>	Travelapp_Eval	Complete	Travelapp_Eval
<input checked="" type="checkbox"/>	Travel&leisure	Complete	knowledge-ba...
<input type="checkbox"/>	Financeapp	Complete	knowledge-ba...
<input type="checkbox"/>	Financeapp_1	Complete	knowledge-ba...
<input type="checkbox"/>	eval_1	Complete	knowledge-ba...
<input type="checkbox"/>	eval_1	Complete	knowledge-ba...

Compare Knowledge Bases evaluation metrics

Select 2 Knowledge Bases to compare their performance.

► Evaluation overview (2)

At a glance | Evaluation details

Comparison: [Travelapp_Eval](#) and [Travel&leisure](#)

The radar chart compares two knowledge bases across eight metrics. The scale ranges from 0 to 1.0. The blue series represents 'Travelapp_Eval' and the red series represents 'knowledge-base-quick-start-fwvzy'. 'Travelapp_Eval' shows higher scores in Helpfulness, Faithfulness, and Completeness, while 'knowledge-base-quick-start-fwvzy' shows higher scores in Correctness and Logical coherence. Both perform similarly on Refusal, Stereotyping, and Harmfulness.

Metric	Travelapp_Eval	knowledge-base-quick-start-fwvzy
Helpfulness	0.8	0.6
Faithfulness	0.7	0.5
Correctness	0.5	0.8
Completeness	0.7	0.5
Logical coherence	0.5	0.8
Harmfulness	0.5	0.5
Stereotyping	0.5	0.5
Refusal	0.5	0.5

Travelapp_Eval
 knowledge-base-quick-start-fwvzy

Model Customization and Adaptation

Training from Scratch

Full Fine Tuning

Parameter Efficient
Fine Tuning (PEFT)

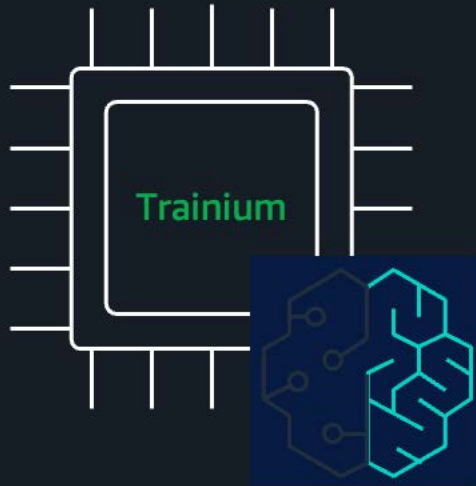
Retrieval Augmented
Generation (RAG)

PE

Increasing Energy Consumption
and Carbon Emissions

Choosing the right silicon

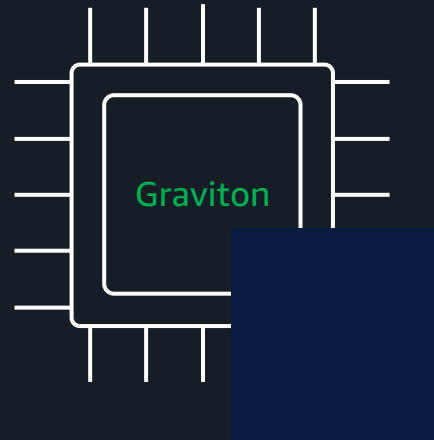
Training & Inference



Trn1 instances are up to **25% more energy efficient** than comparable accelerated EC2 instances.

Trn2 instances are **3x more energy efficient** than Trn1 instances.

Inference for non-GPU workloads



Best performance per watt in Amazon EC2

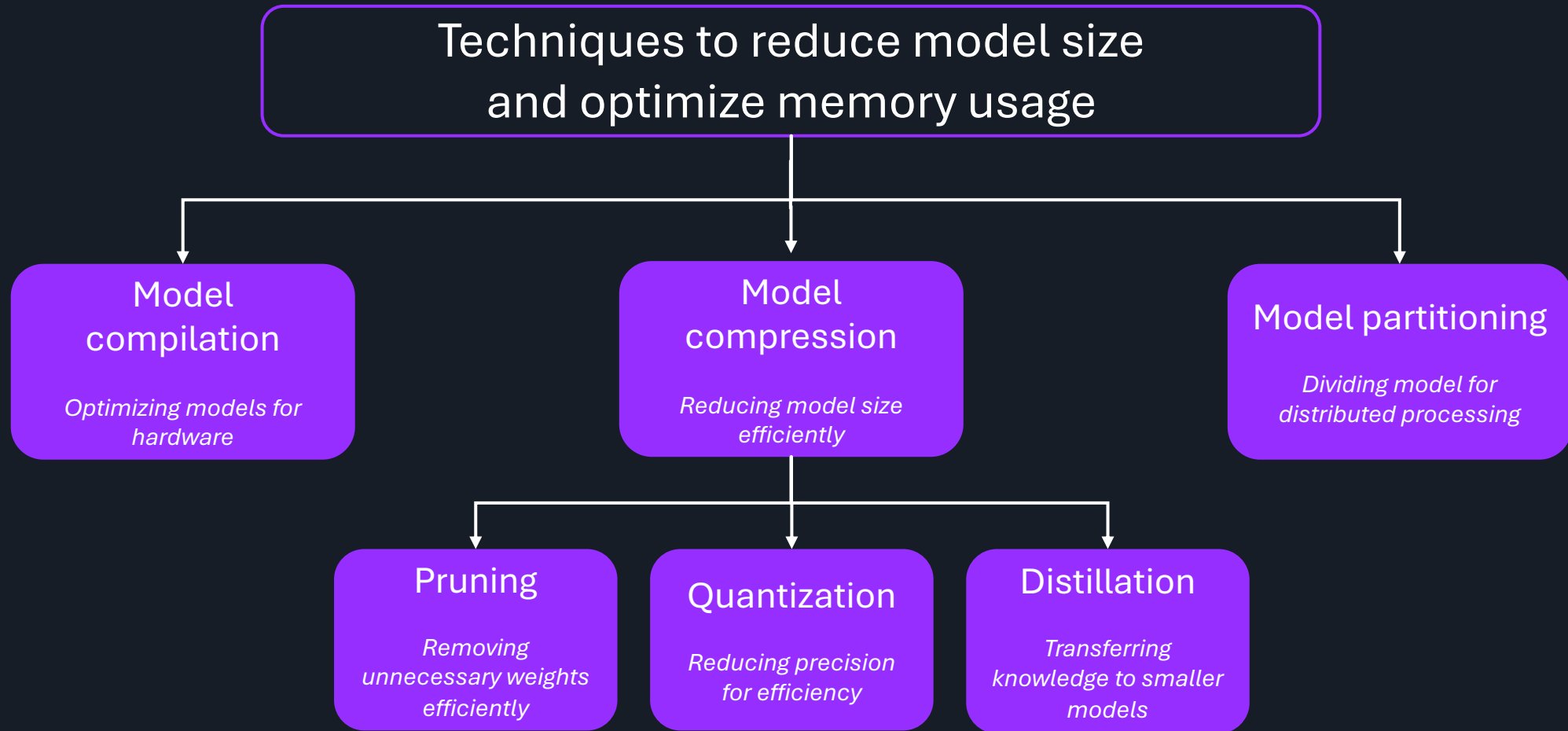
Up to **60% less energy** than comparable EC2 instances

4x

Performance increase since the first generation Graviton1 in 2018



Model Deployment/Inference Optimization



Tools available in Large Model Inference (LMI) Containers

Continuous Monitoring and Optimization

**Amazon
CloudWatch**



**SageMaker
Profiler**



**AWS Neuron
Monitor**



**NVIDIA System
Management
Interface**



Monitor Amazon Bedrock Agents
using CloudWatch Metrics



Memory Info
GPU utilization
Neuroncore utilization
Memory Utilization
...

Data and model quality monitoring
with Amazon SageMaker Model
Monitor

Key Techniques to consider

Managed Services

Base Model
Selection

Model
Customization
Techniques

Deployment and
Inference
Optimization

Right Silicon Choice

Continuous
Improvement

- Ask questions about sustainability in planning discussions and set goals
- Identify a workload and start a sustainability architecture review
- Learn how to optimize for sustainability

Blogs

[Optimize AI/ML workloads for sustainability](#)

[Optimizing Generative AI workloads for sustainability](#)

[Guidance for Optimizing MLOps for Sustainability on AWS](#)

Whitepapers

[AWS Well-Architected Framework – Machine Learning Lens](#)

[AWS Well-Architected Framework – Sustainability Improvement Process](#)

