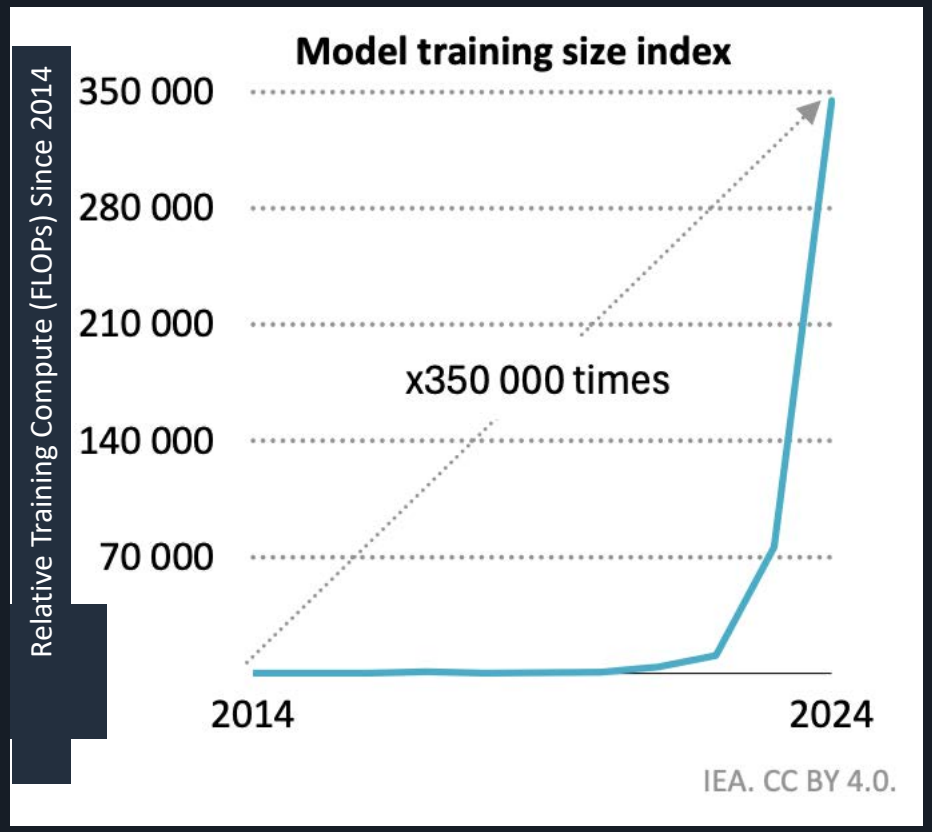
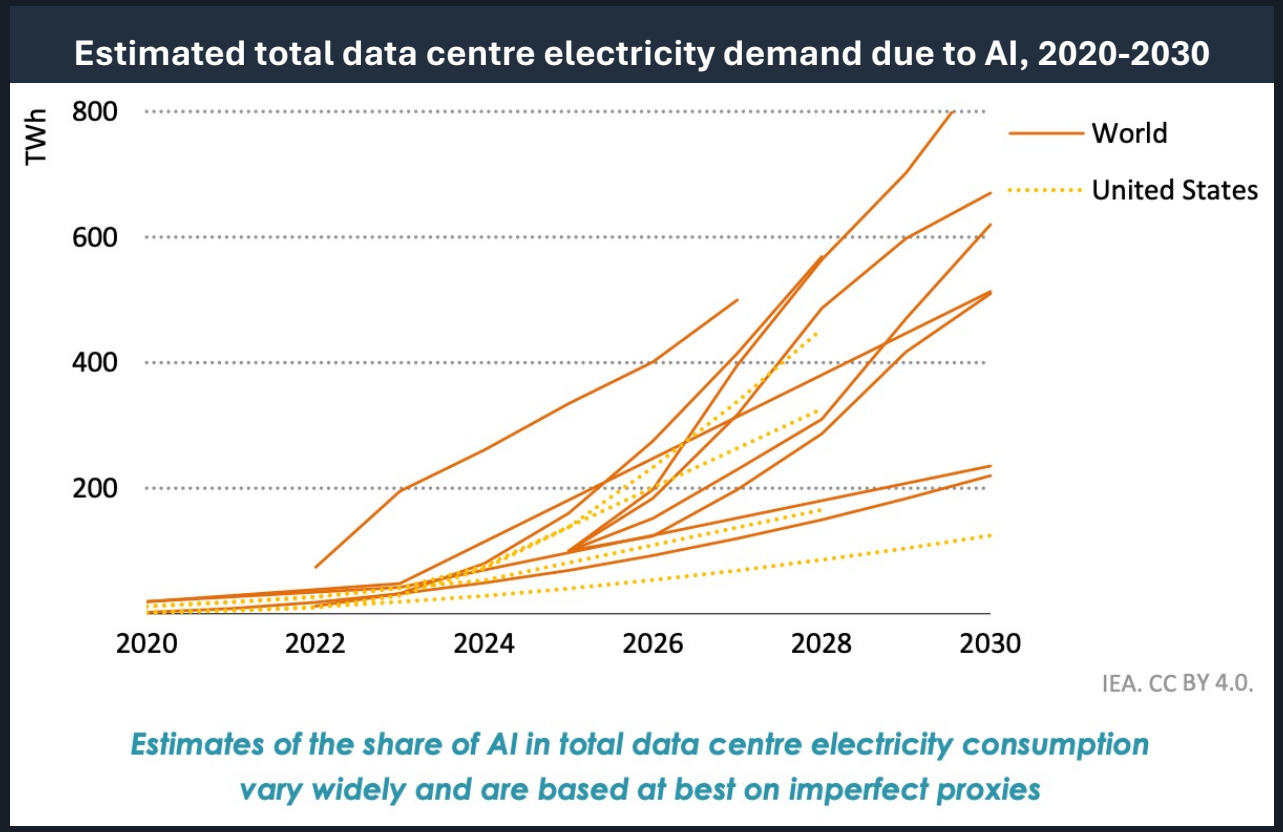
Building GenAI that doesn’t cost the earth

Mohamed Najaaf
Solutions Architect
Amazon Web Services

AI models are getting bigger…
and resources consumption is increasing





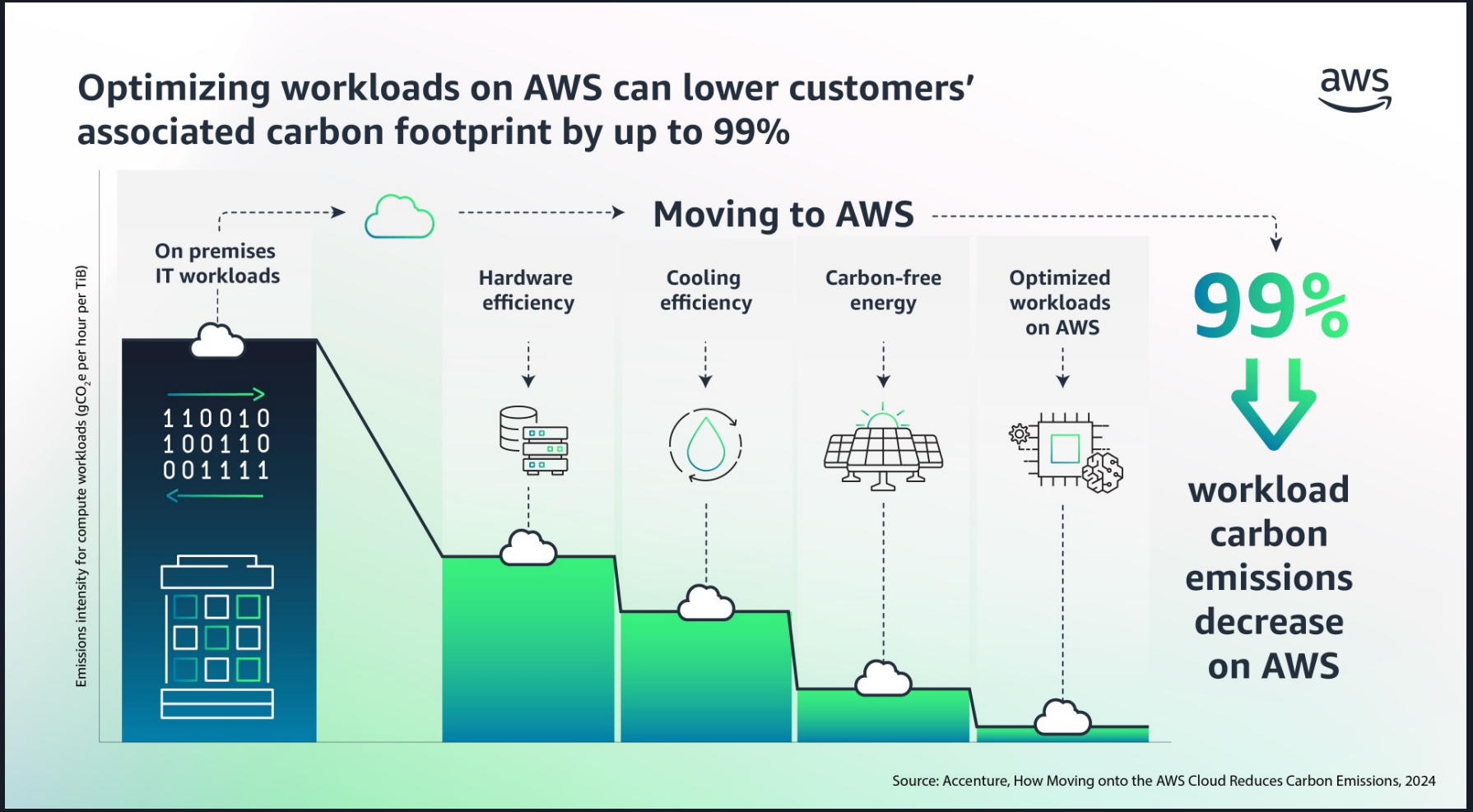
Sources: IEA analysis based on data from Deloitte (2024), Gartner (2024), Goldman Sachs (2024), Schneider Electric (2024), SemiAnalysis (2024), and Shehabi, et al., (2024).

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

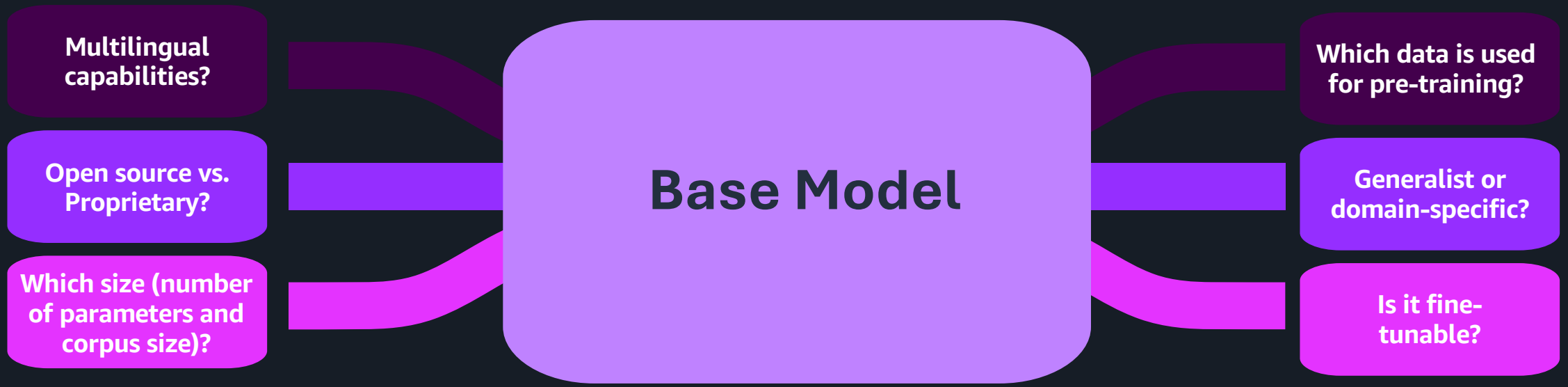
Optimizing workloads on AWS can lower customers' associated carbon footprint **by up to 99%**



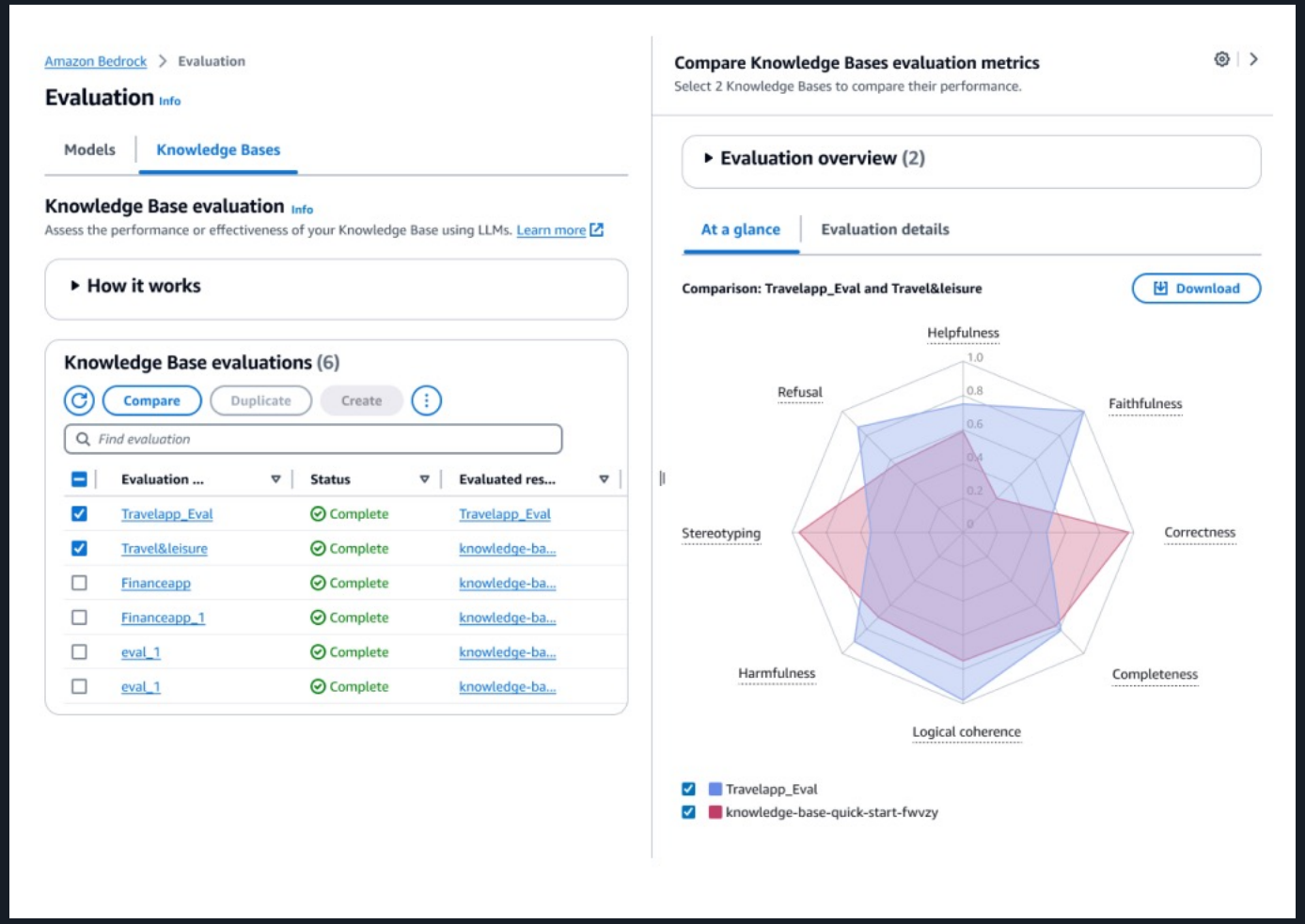
Find the report at https://sustainability.aboutamazon.com/carbon_reduction_aws.pdf

Managed Services

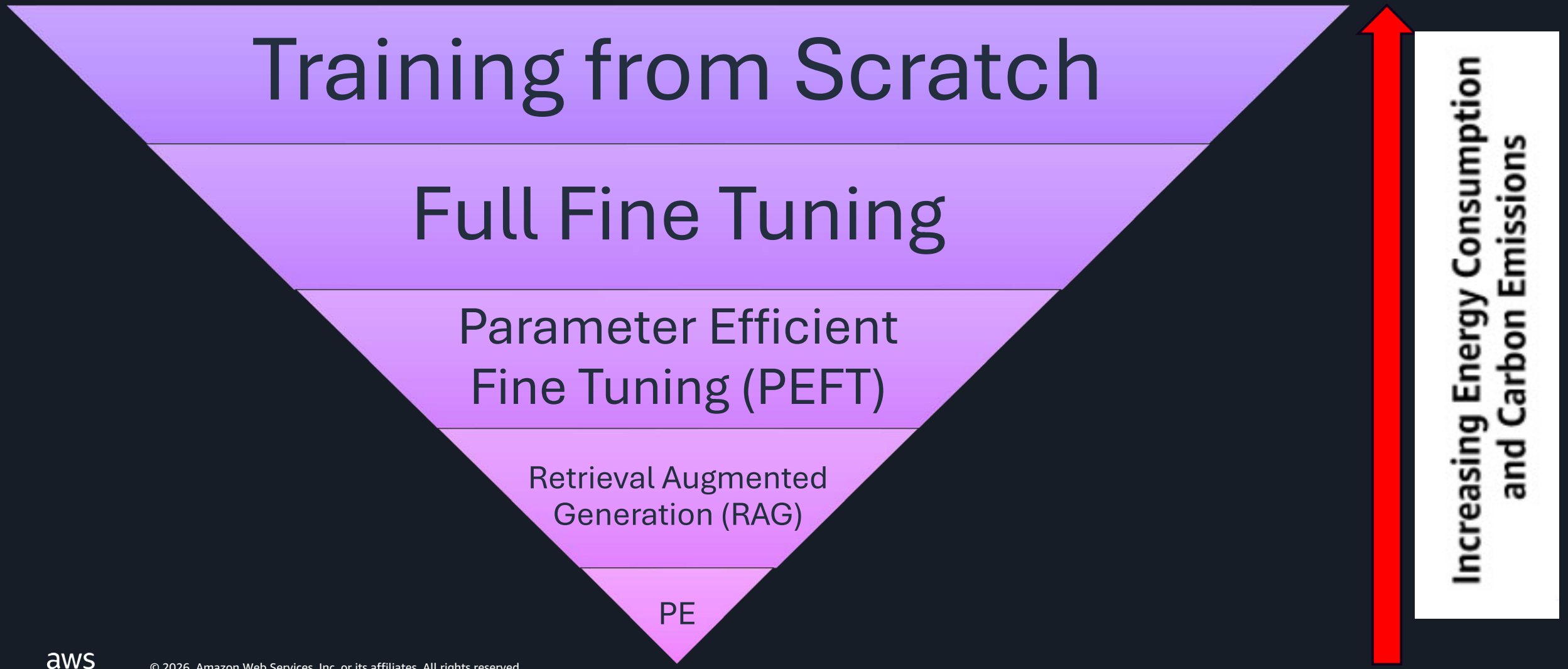
Model Selection



Amazon Bedrock Evaluations

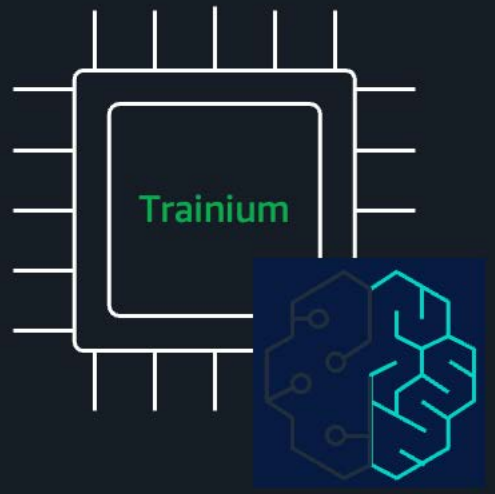


Model Customization and Adaptation



Choosing the right silicon

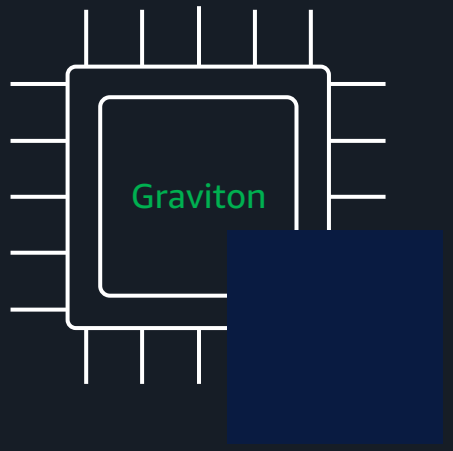
Training &Inference



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



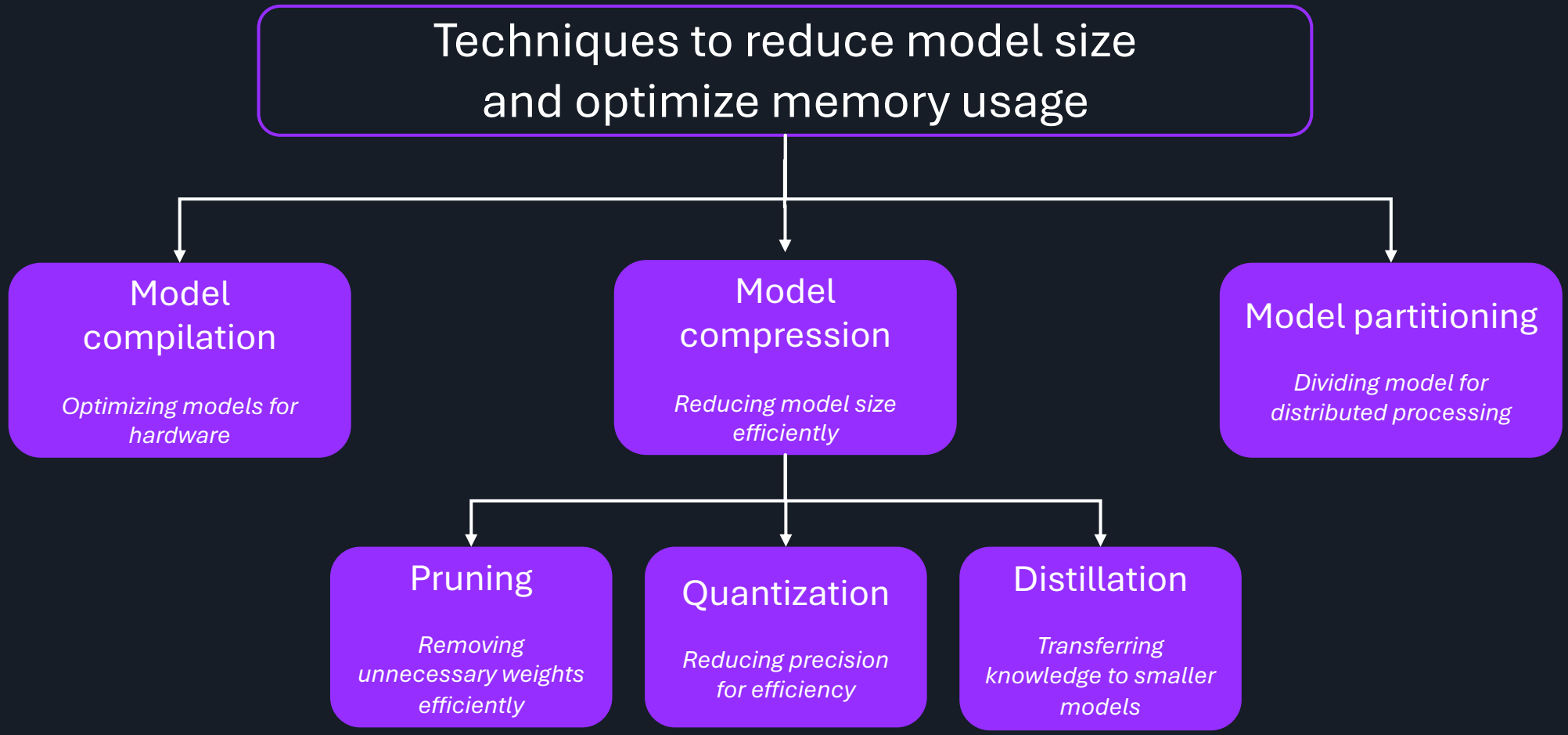
Best performance per watt in Amazon EC2

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

Monitor Amazon Bedrock Agents using CloudWatch Metrics

SageMaker Profiler

Memory Info
GPU utilization
Neuroncore utilization
Memory Utilization
…

AWS Neuron Monitor

Data and model quality monitoring with Amazon SageMaker Model Monitor

NVIDIA System Management Interface

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