

Hewlett Packard
Enterprise

Optimizing deep neural network inference workloads

Lindsey Hillesheim, Head of HPE Tech Advance Program

Hana Malha, AI Technologist at HPE

November 15, 2023

Agenda



1. Motivation
2. Overview of optimization methods
3. Generative AI & LLMs

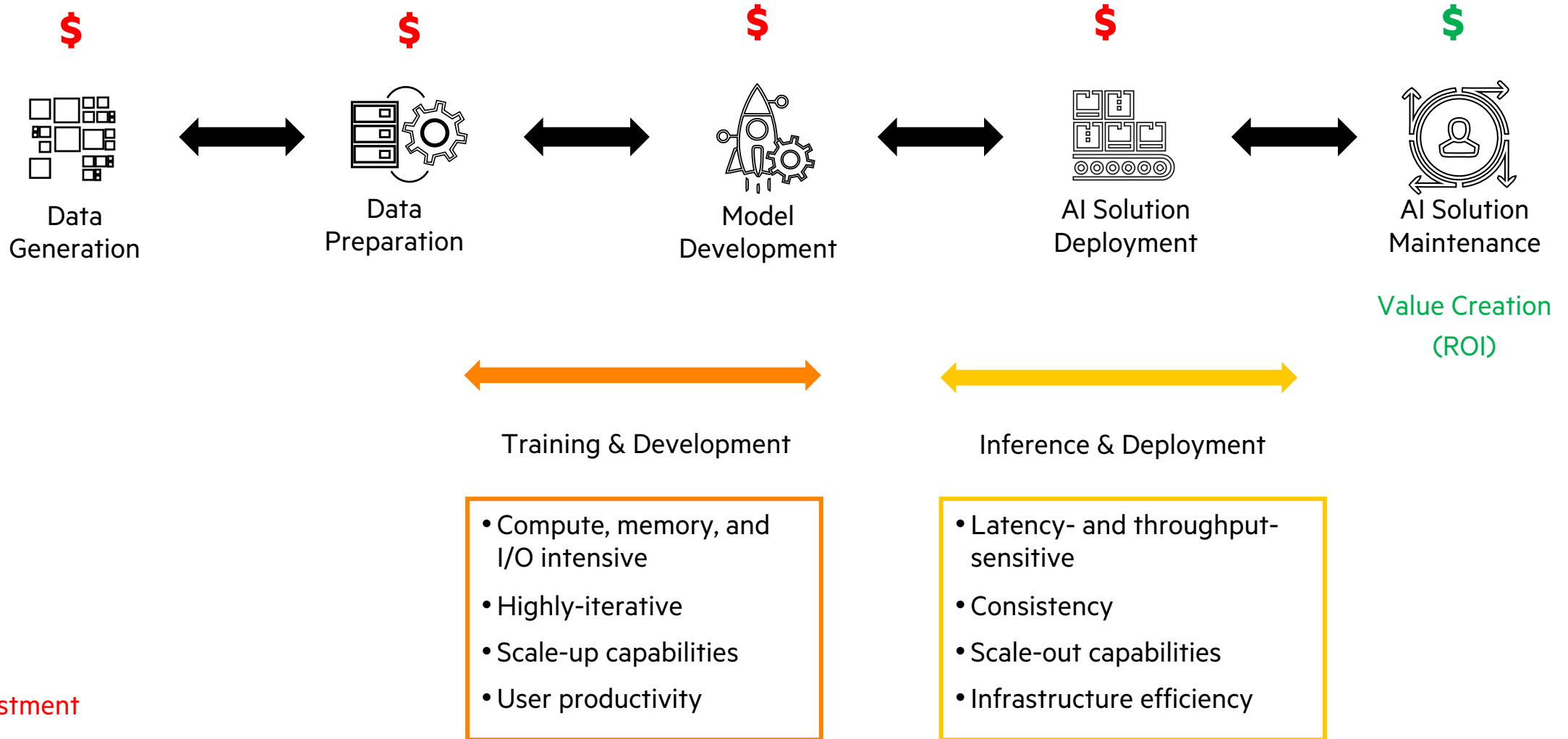


Training vs. Inference

Apples & Oranges



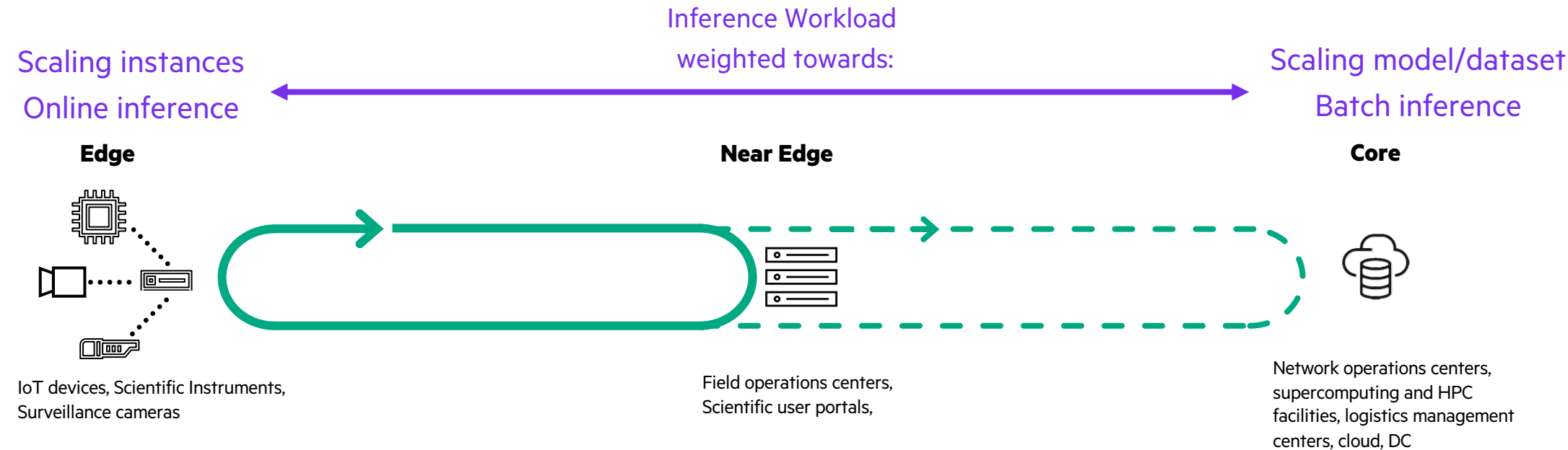
Inference & training drive different system requirements



\$ Investment







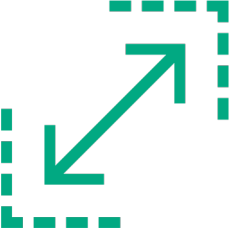

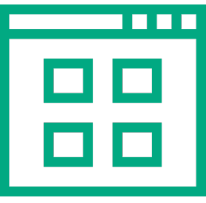
Inference occurs at edge and core



Example Use Cases	Intelligent Cameras Wearables Predictive Maintenance TTS/STT	Network Intrusion Detection Quality Assurance	NLP Large # compute cycles dedicated to inference vs. training jobs once model is in production (batch inference)
	Autonomous Driving Telco Network Management Smart Grid		



Key performance metrics for inference

Important Metrics for Inference				Other Factors (Production & Long Term)		
 Cost	 Power	 Throughput	 Latency	 Scalability	 OS Support	 Software Stack
Total \$	Total Watts	Total Inf/Sec	Total Time (single inference)			
Performance/\$ & TCO	Performance/W & TCO	Performance/\$ Performance/W & TCO	Time to Decision (Single threaded)	Large models Many Params	Market Adoption	TCO & Market Adoption

Which metrics matter vary by application.
Meeting metrics = ROI



Why should you care about inference optimization?

Optimize a model to target HW **that meets KPIs at development time** with fewer iterations

Choose alternative hardware for inference workloads due to HW cost

Optimization approach that works **across use cases from edge to core**

ROI

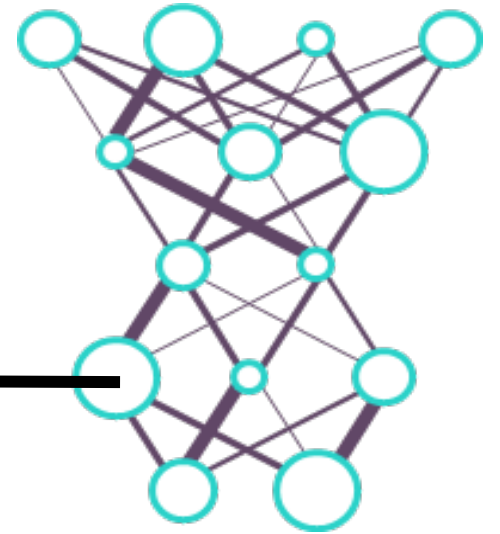
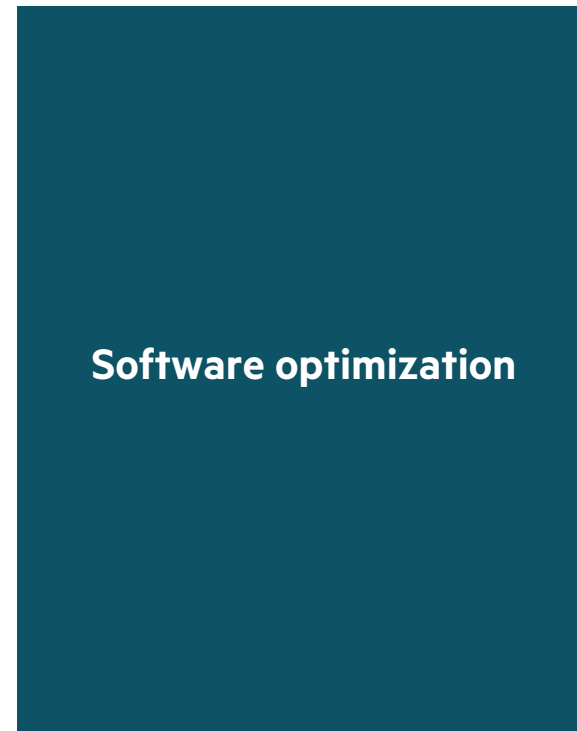
more AI Models
in production



Inference Optimization Methods



Inference optimization:

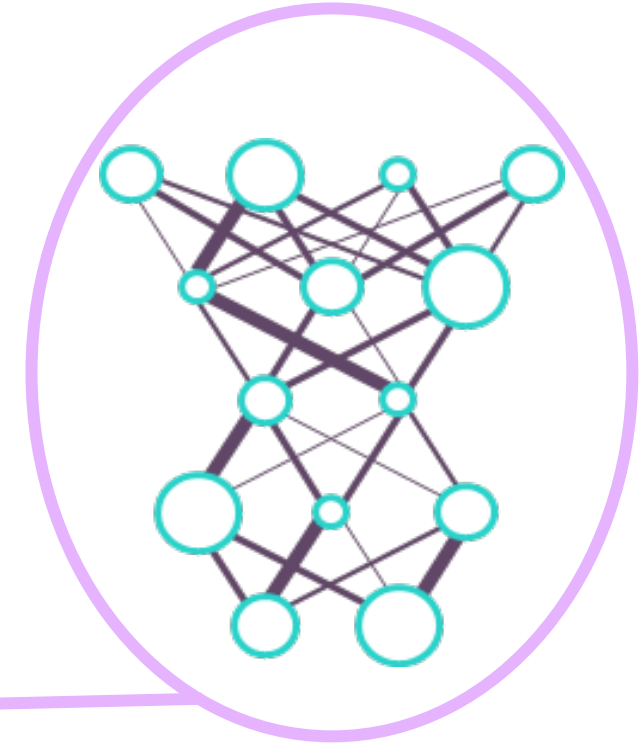


Vendor specific drivers

Target selection
Heterogenous/multi-device execution



Inference optimization:



**Converting Model (graph)
to Machine Code**
Optimizing, Lowering

Compilation

Optimally feeding the Accelerator
Batching, model caching

Runtime

Vendor specific drivers
Target selection
Heterogenous/multi-device execution

Execution



Inference optimization:

Pruning: Eliminate nodes
Quantization*: Reduce Precision
Knowledge Distillation
(require retraining)

**Converting Model (graph)
to Machine Code**
Optimizing, Lowering

Optimally feeding the Accelerator
Batching, model caching

Vendor specific drivers
Target selection
Heterogenous/multi-device execution

**Training-aware
optimizations**

Compilation

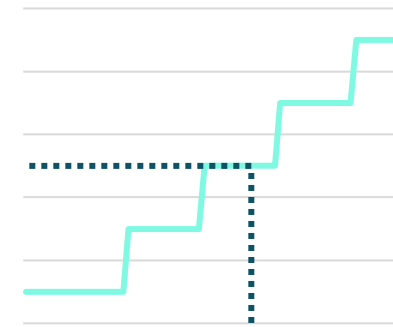
Runtime

Execution



Quantization

Quantized values

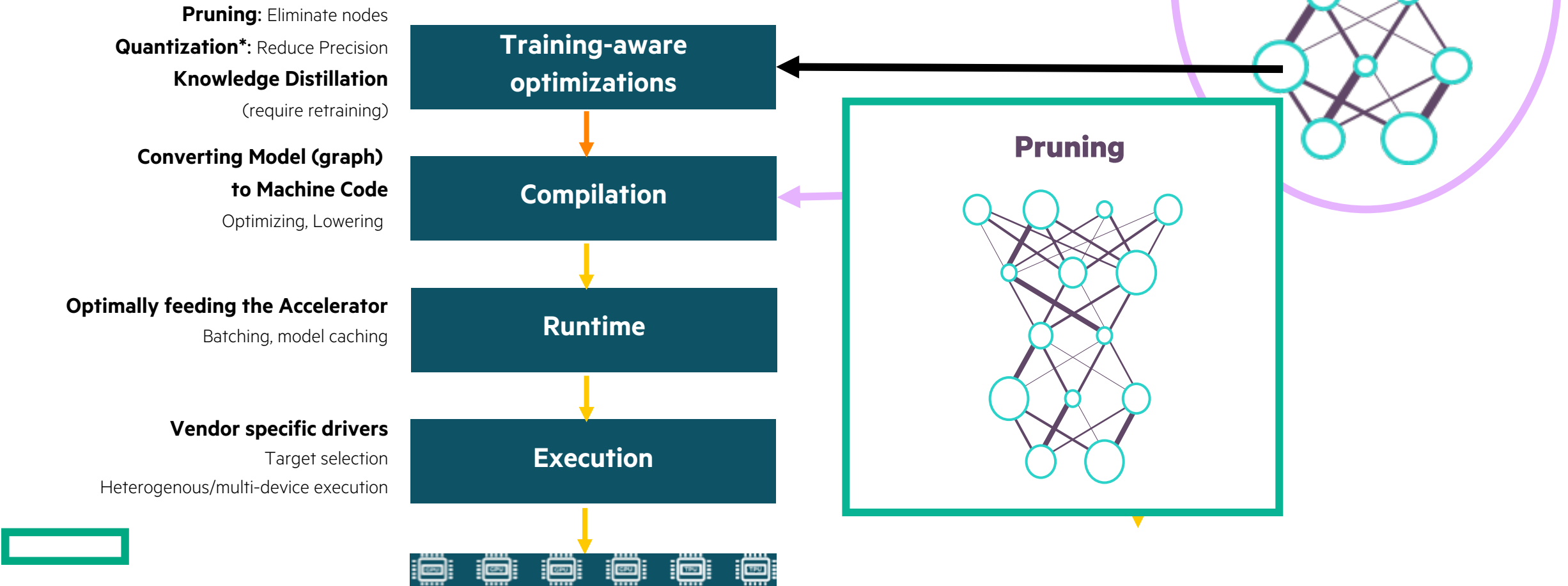


FP32 range

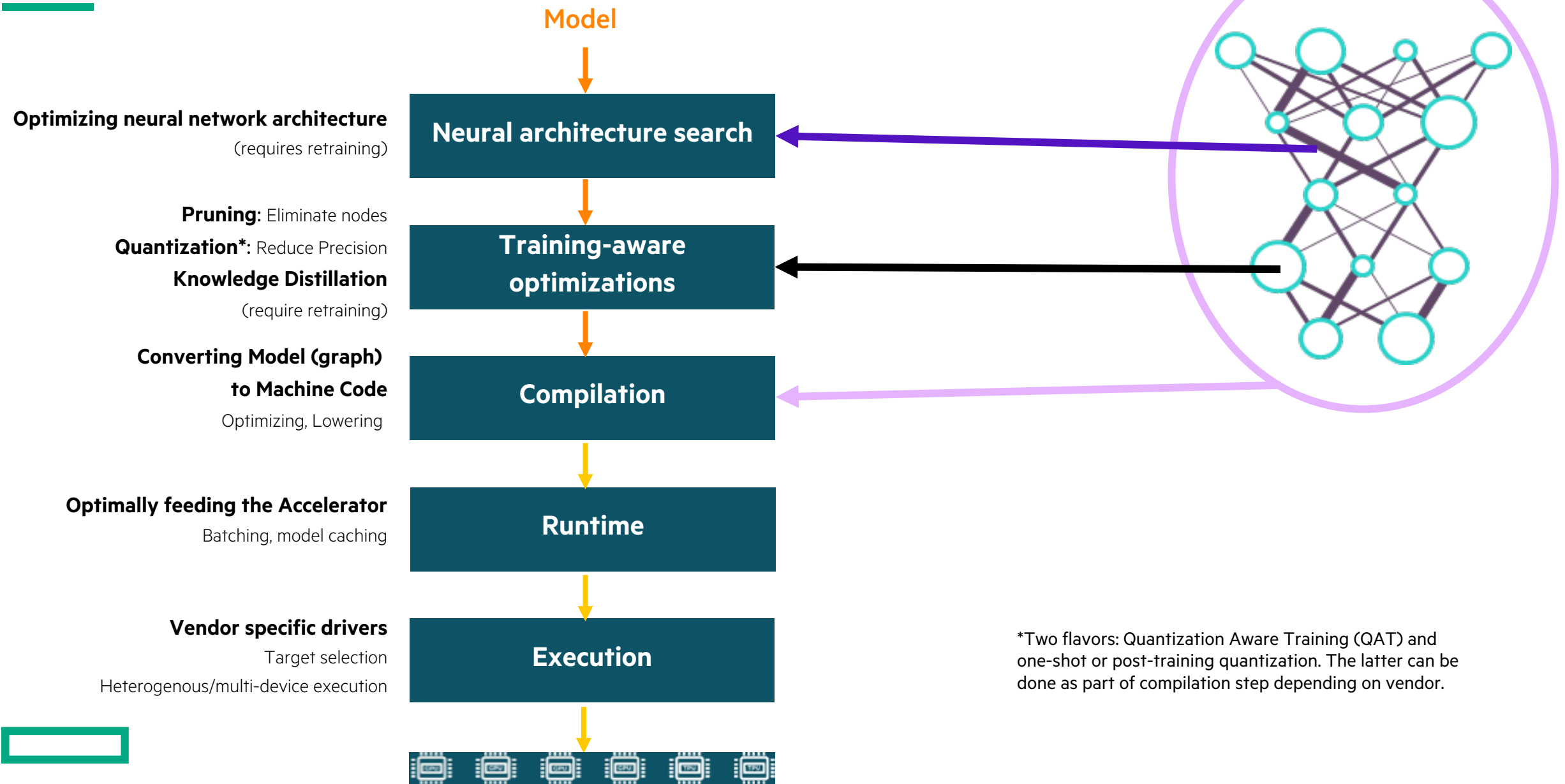
- FP32 to 8bits or lower
- PTQ, QAT, ...



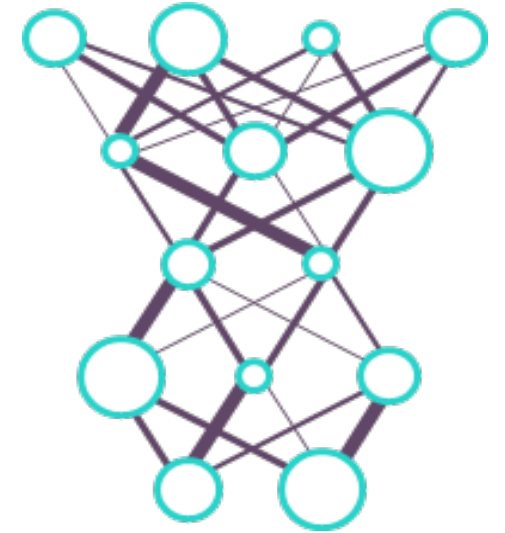
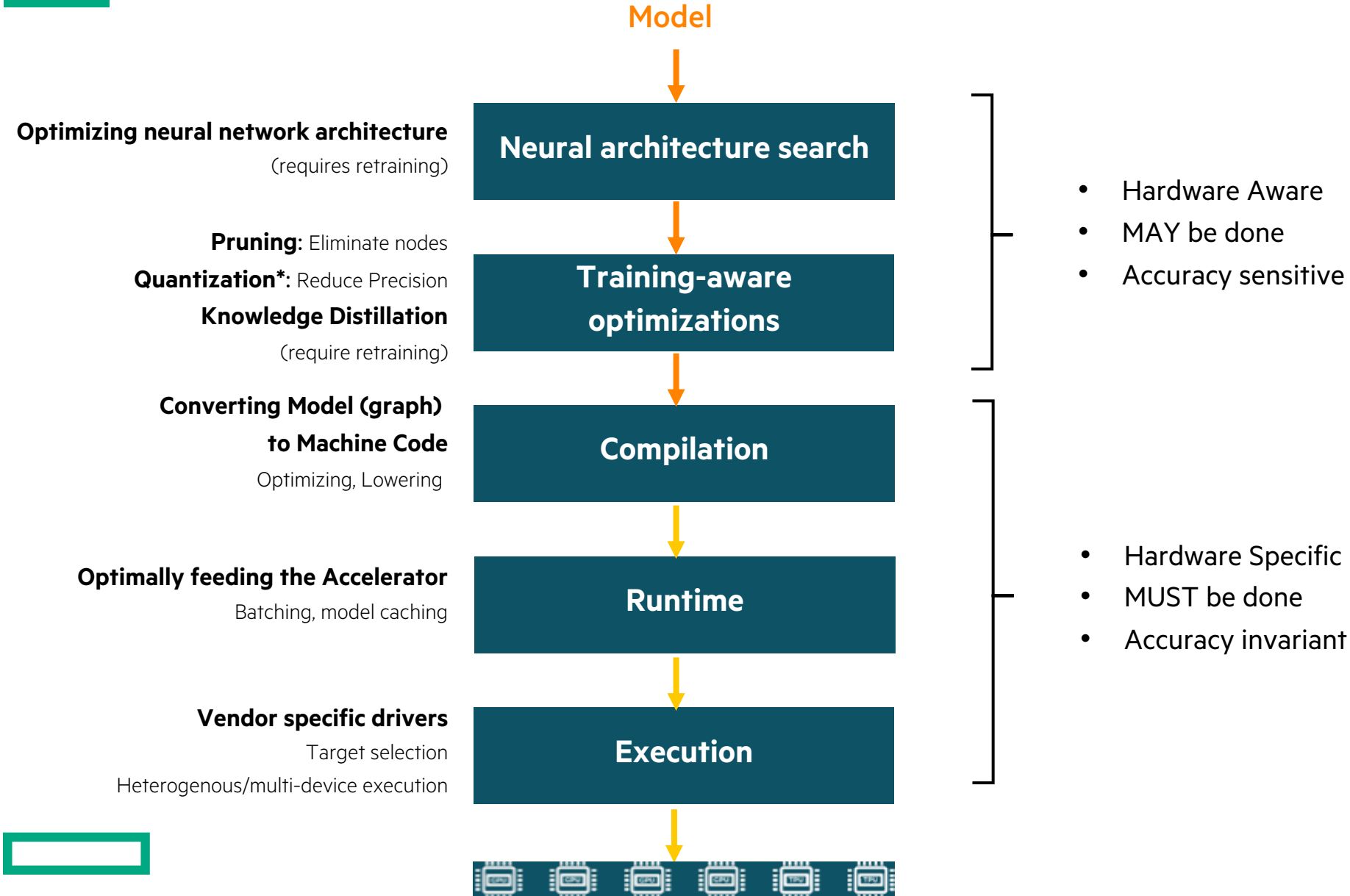
Inference optimization:



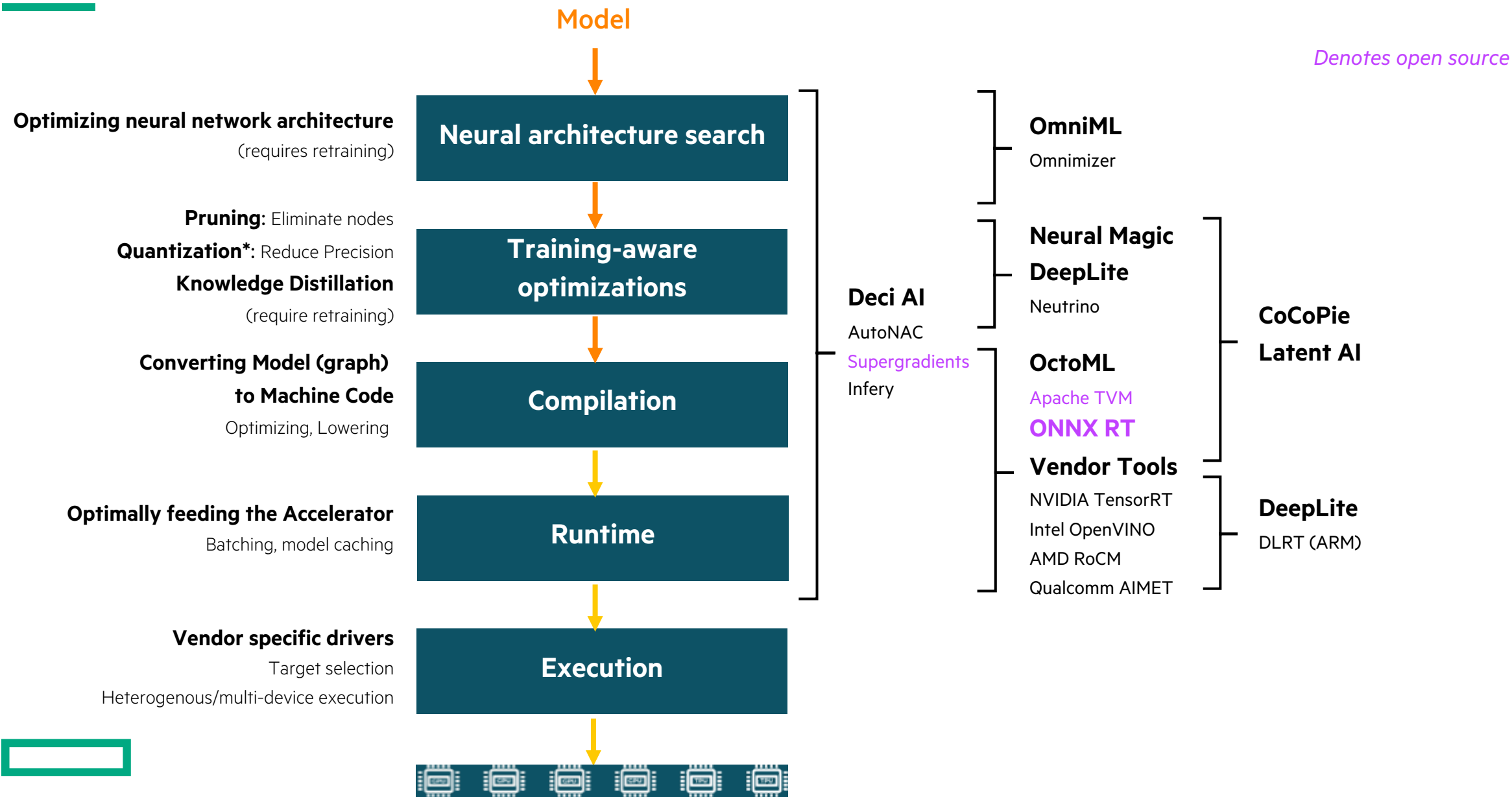
Inference optimization: technology taxonomy



Inference optimization: technology taxonomy



Inference optimization: start-up, OSS & vendor landscape



Generative AI / Large Language Models

Elephant in the room: What happens when the DNN gets really big?



Key inference differences for Gen AI models



Iterative nature

- Inference is more complex
- Model performs several inference iterations to generate a new sample
- Ex: Text is generated token by token.

Overprovisioned infrastructure

- Each inference requires more compute due to model size & iterative nature.
- A lot of infrastructure required to maintain high availability and low latency when demand spikes.

Variable cost per prompt

- Content generated depends on users' prompt so cost per prompt is not constant
- Difficult to estimate and predict cost of running and scaling models in production.



Generative AI: Biggering and biggering

Time

2 mo

For Chat GPT to reach 100 million active users¹

Cost

\$40m

Estimated costs for Open AI to process prompts in January²

Infrastructure

\$4b

Required infrastructure to support serve Microsoft's Bing AI chatbot²

CO2 emissions

2%

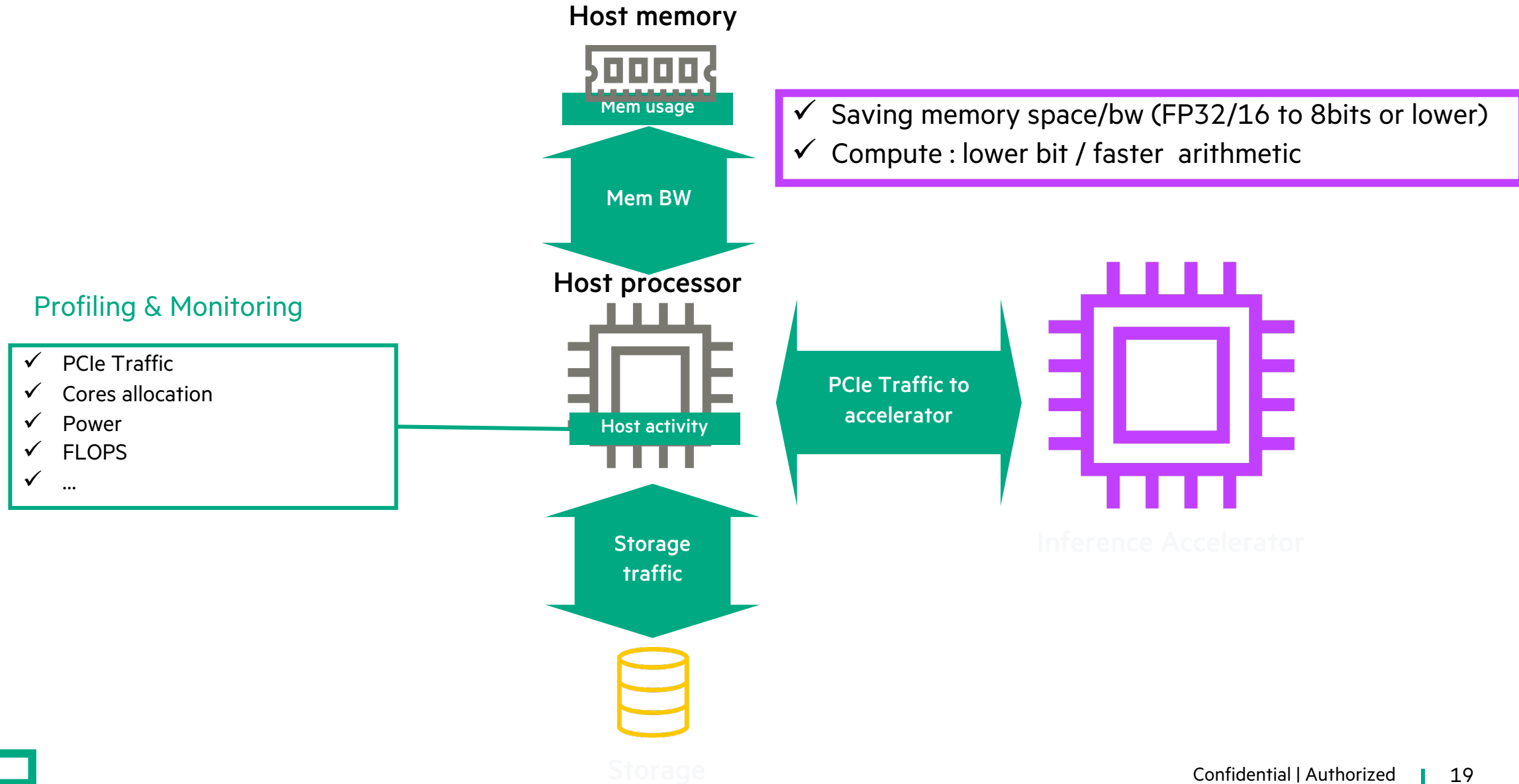
Share of global CO2 emissions attributed to the Information and Communications Technology (ICT) sector in 2020³

¹ <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>

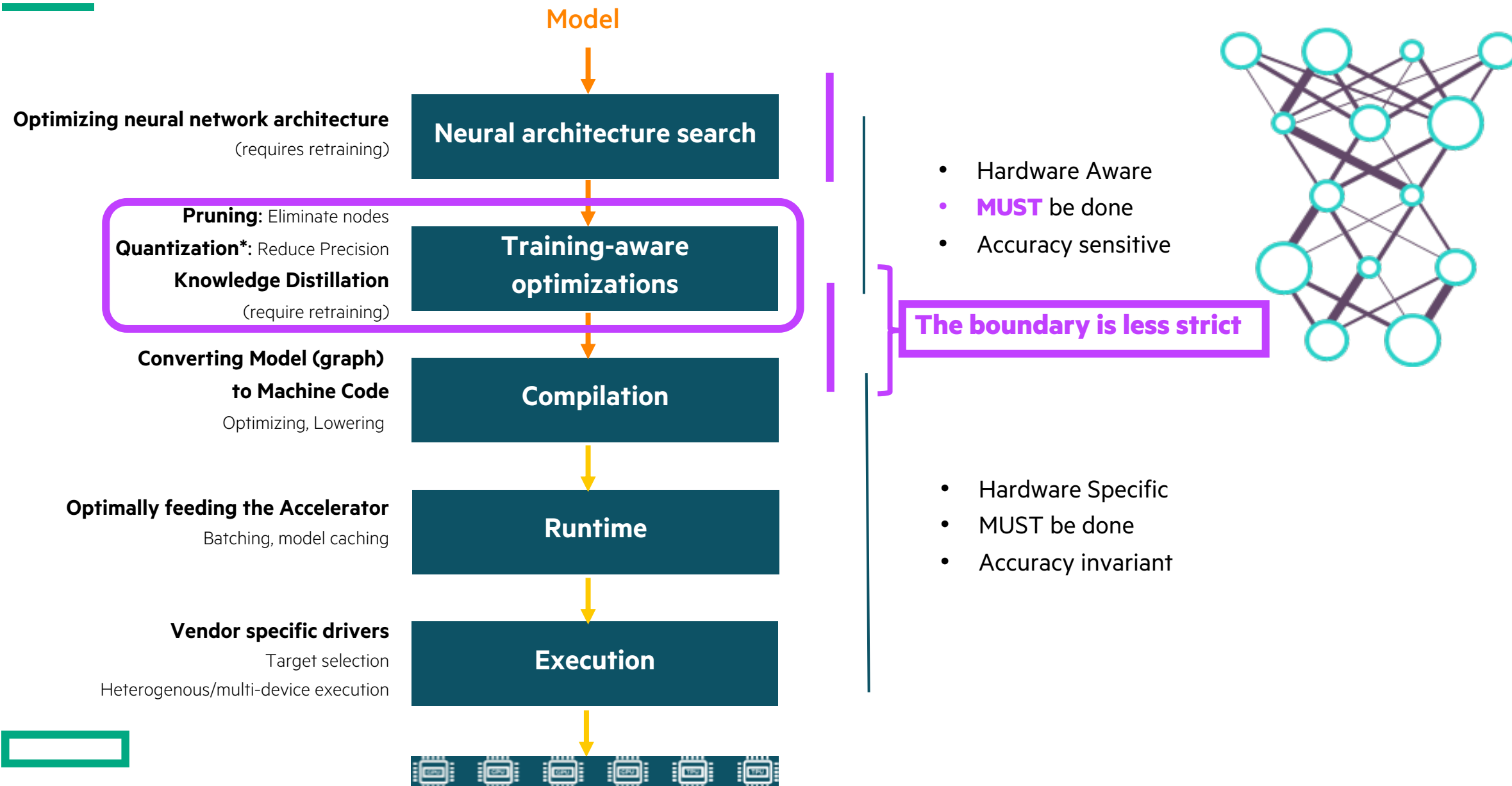
² <https://www.cnn.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html>

³ <https://www.spiceworks.com/tech/artificial-intelligence/quest-article/hidden-costs-of-generative-ai/>

Hardware resources bottleneck



LLM Inference optimization



Key Takeaways

Inference workloads are complex

- Inference workload profiles depend on the model, underlying software, and hardware.
-

Multiple Methods

- Multiple software optimization methods are needed to get large performance gains.
-

Performance, Power, TCO

- Optimization can improve performance and allow inference to be run on lower TCO hardware.
-

Parka not Lipstick

- Inference optimization is not lipstick on the model; it is the snow parka when it is -5 F.
-



Thank you

lindsey.hillesheim@hpe.com

hana.malha@hpe.com

Creating a curated and trusted innovation ecosystem

HPE Tech Advance Mission

We build deep partnerships with the most innovative & promising technology and solution providers in data, edge, sustainability, and AI to address current and emerging HPE customer needs.

To build a trusted and mutually beneficial relationship, we take a phased approach.

