



Measure, Then Compress: Scale- and Architecture-Dependent Routing of Compression for Edge AI

Brian Jalaian¹ · Venkat R. Dasari²

¹University of West Florida, Pensacola, FL ²DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD

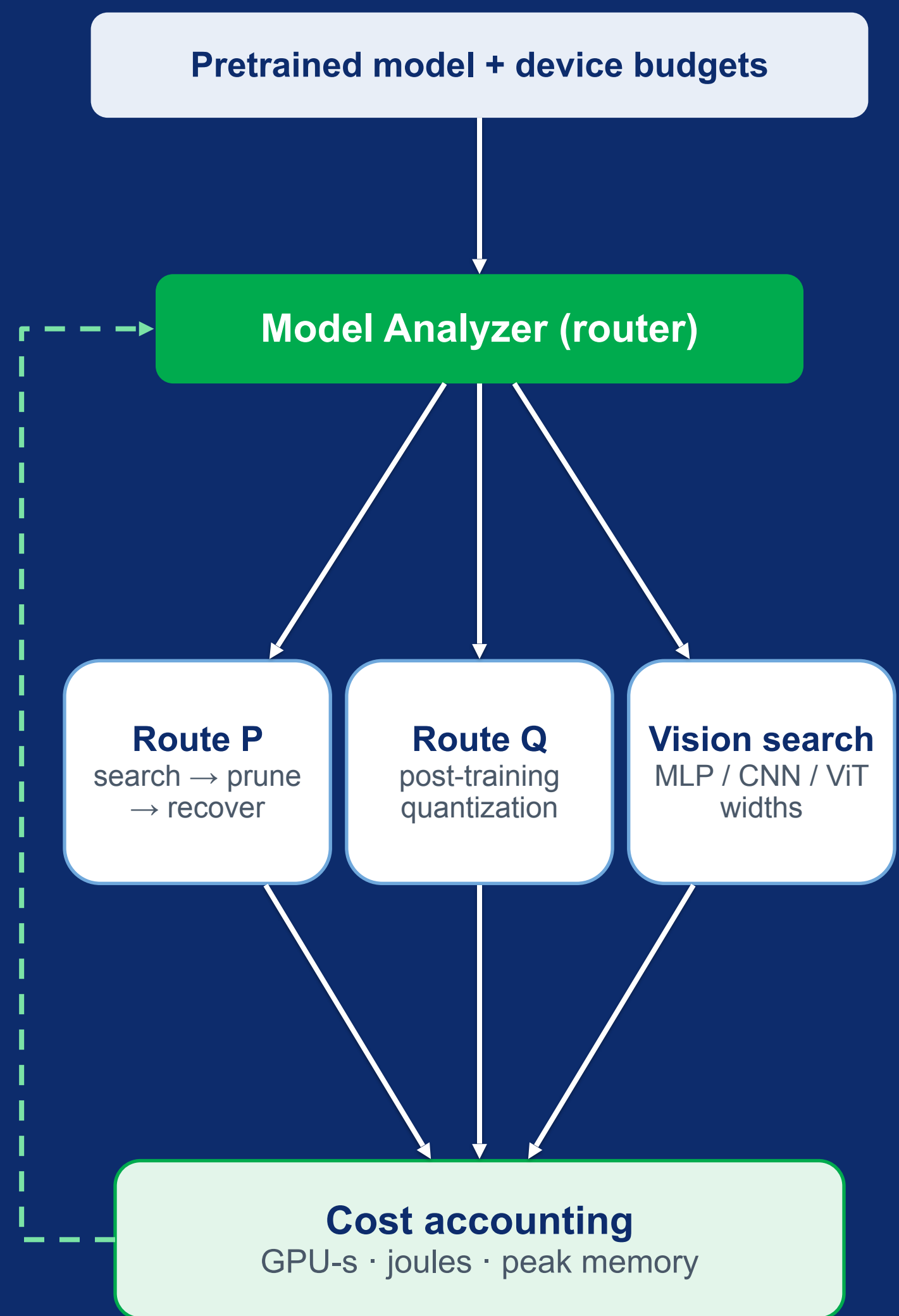


MOTIVATION

- Edge accelerators differ in compute, memory, and power. Each compression method wins only in a narrow context.
- The choice of pruning, quantization, or search is made by hand and **defended by assertion**, e.g. that supernet search is intractable for LLMs.
- We add no new operator. We **measure which route to apply** and what it costs.



METHOD: AN INSTRUMENTED ROUTER



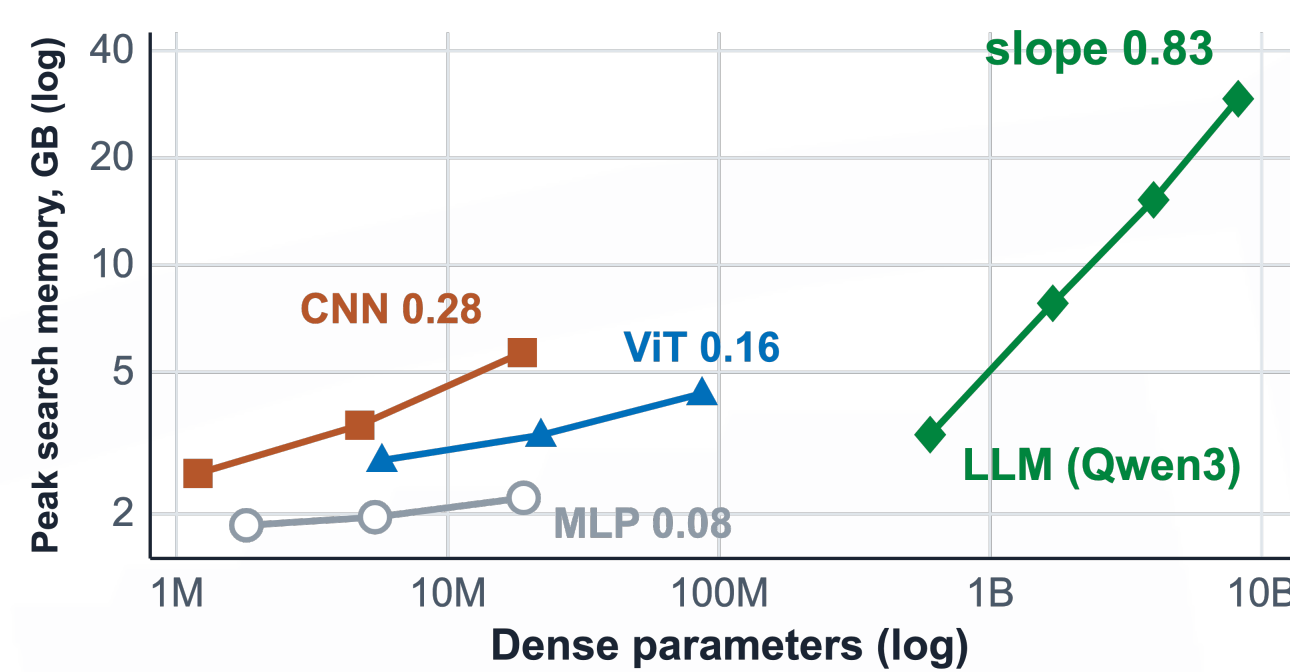
- One **unchanged** supernet search on MLP, CNN, ViT, and Qwen3 0.6B–8B (**13.7×** params).
- 4%** accuracy-loss budget on CommonsenseQA + ARC-Challenge; all runs on NVIDIA A40.



No static compression recipe survives contact with model scale, family, and architecture. Measured, the right route changes: so **measure, then compress**.

RESULT 1

Search is bound by memory, not time

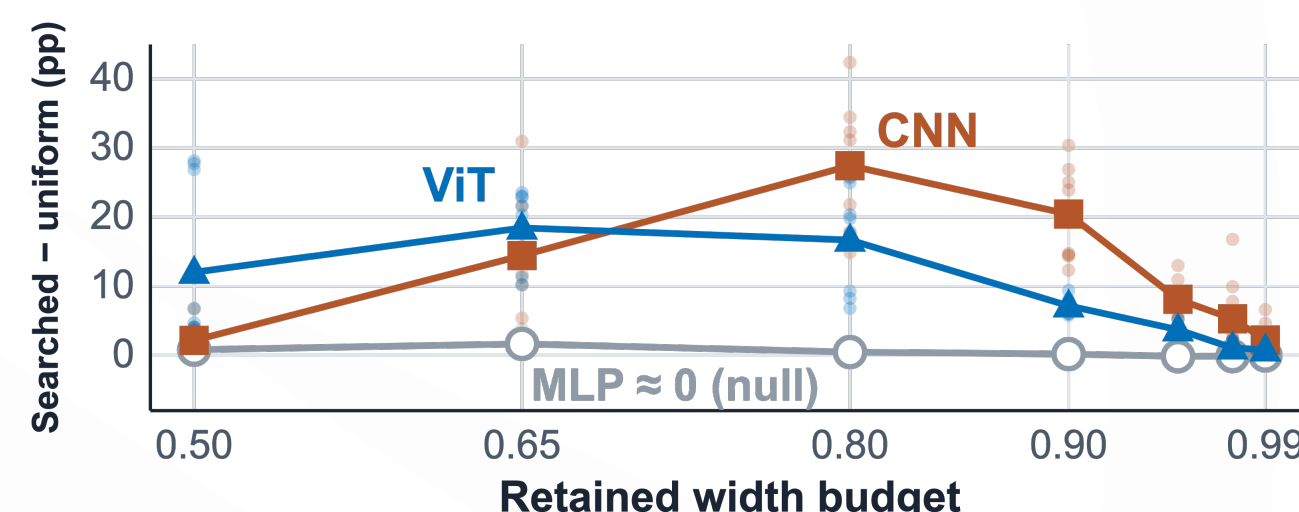


Peak search memory vs. parameters; labels give log–log slopes fitted within each family.

- Qwen3 0.6B→8B: GPU-s grow 5.9×, but peak memory grows **8.9×**.
- A full search at Qwen3-8B takes **under one GPU-hour** (2888 GPU-s), so search is affordable.
- Memory exponent is **0.83** for the LLM vs. 0.08–0.28 for vision. Compute is sublinear everywhere.

RESULT 2

Search pays only on some architectures



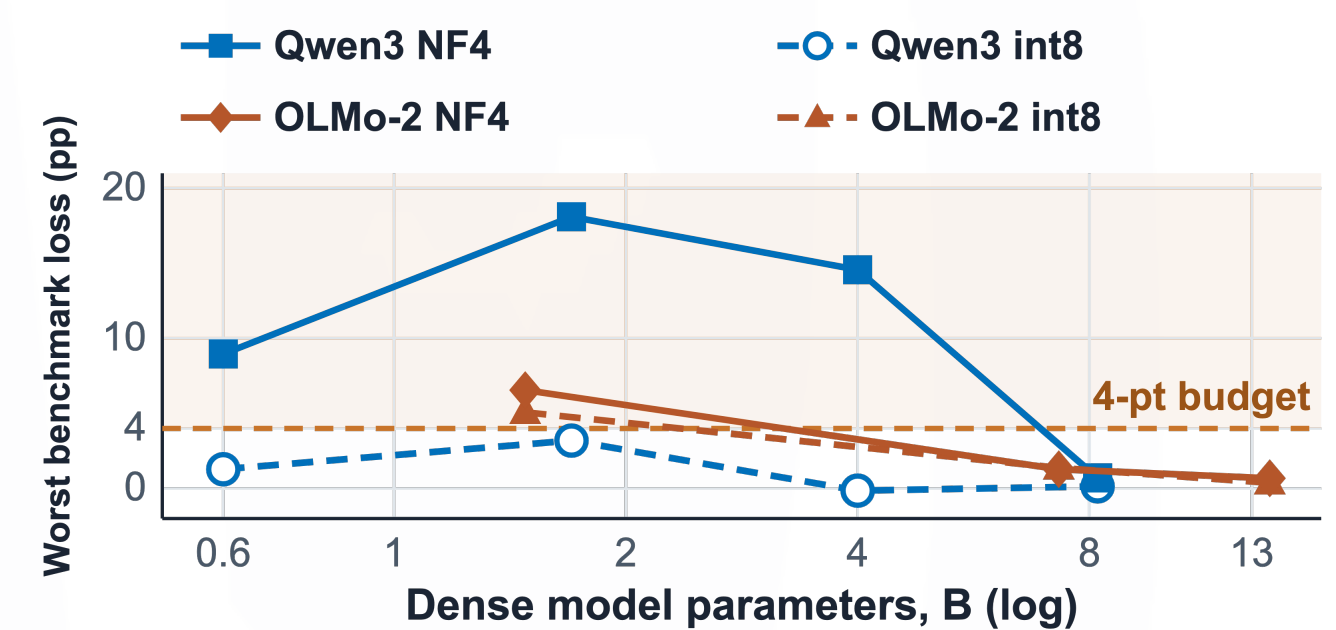
Paired top-1 gain of searched over uniform width allocation (same budget, held-out split). Lines are means; dots are rung–seed pairs.

- CNN / ViT:** search wins by **+6.6–12.8 pts**. CNNs won 56/63 pairs while *smaller*.
- MLP:** +0.3–0.5 pts, a coin flip. A null result.

RESULT 3

Quantize, but the crossover is no scale law

Qwen3	Search GPU-s	Route P CQA / ARC	Route Q
0.6B	490	1.05× / 1.02×	int8 1.59×
1.7B	849	1.06× / 1.01×	int8 1.69×
4B	1868	1.19× / 1.08×	int8 1.82×
8B	2888	1.18× / 1.09×	NF4 2.75×



Worst benchmark loss per rung. NF4 becomes admissible between 4B and 8B on Qwen3 but between 1B and 7B on OLMo-2.

- Quantization (Q) beats prune + recover (P) at every scale: 2.75×** vs. 1.18× at 8B, with zero search.
- A Qwen3-calibrated precision rule is **wrong at 2 of 3** OLMo-2 rungs.



ROUTING GUIDANCE



Key the router on peak memory, not GPU-seconds.



LLMs: quantize first; measure int8 vs. NF4 per model.



Vision: search CNN/ViT widths; skip MLPs.

Limits: A40 GPUs, not edge devices; multiple choice and image classification only.

