

Not Every Question Needs the Whole GPU

When to use BF16, FP8, INT4, or speculative decoding, and when a router earns its keep.

Markell Rawls

AI Developer Advocate, Red Hat

- **Track B: Qwen3.8-27B on H200 MIG slices**

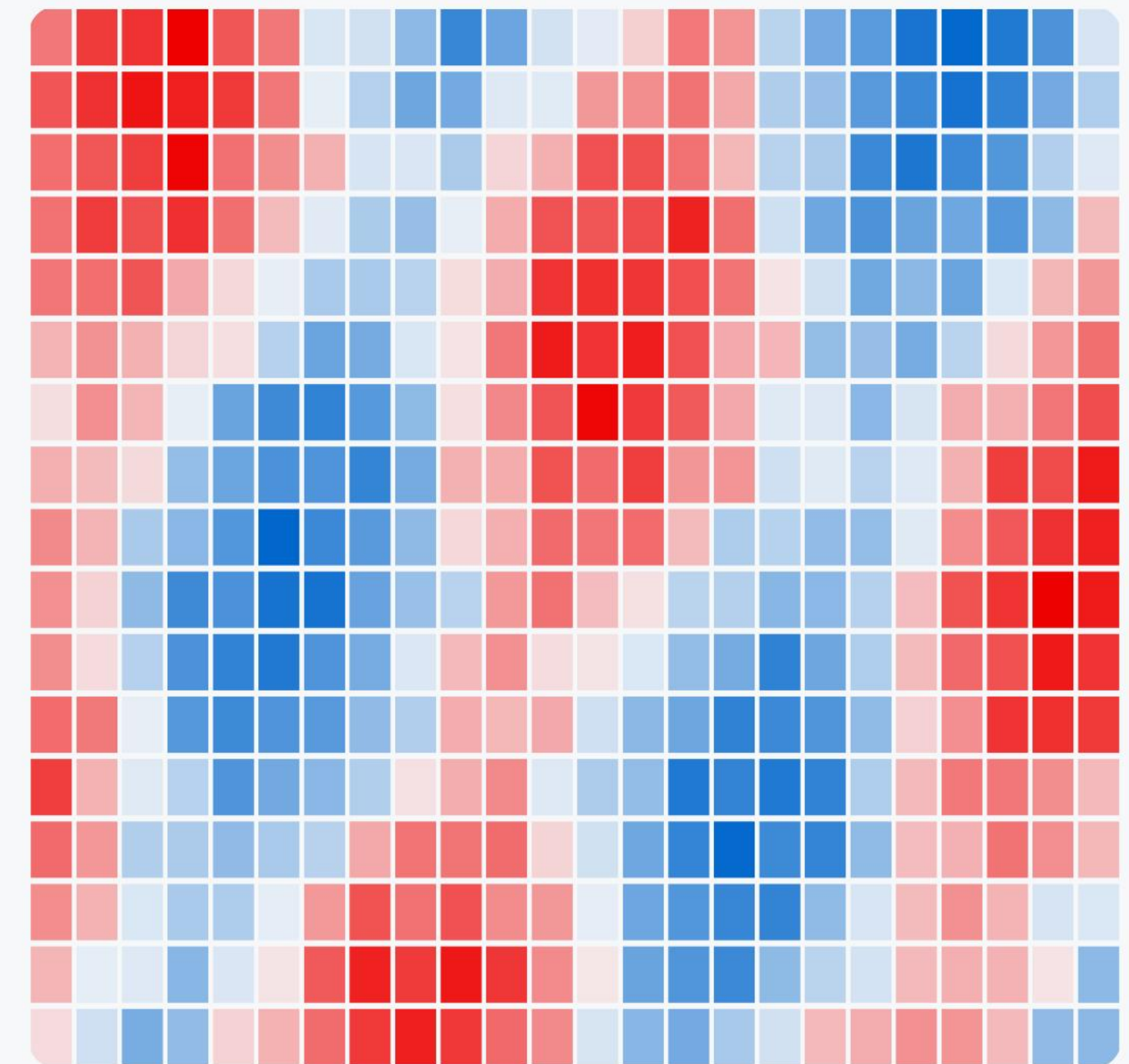
👋 Hands up: which one is your shop?

1 70B-class, spread across GPUs

Llama 3.1 70B on two H200s

2 27B-class, one GPU or a slice

Qwen3.8-27B on H200 MIG slices



A weight matrix getting rounded to 4 bits, one of the tricks in this story.

QUANTIZATION SHOWDOWN?

This is a **toolbox tour**, not a shootout

What you'll walk away with

- 1 **When each option earns its spot** – BF16, FP8, INT4, and speculative decoding each solve a different problem
- 2 **How to measure the trade-off** – speed, memory, accuracy, and cost per token
- 3 **A live demo** – all four running the same question side by side

How I'll show you

- ? A story – three people, five chapters, one live demo
- ? Same model, same prompts, same hardware, different trade-offs
- ? One thing to keep in mind: a small accuracy drop compounds over thousands of tokens



Auto

Notes

THE STORY

One model, one platform team, and a GPU bill that keeps growing



Priya platform team

Runs the model the whole company uses, and is a little naive about it



Marcus data science

Needs answers he can trust for deep analysis



Dana finance

Watches the GPU bill, and has questions



Priya ships a 27B at full precision



It gets popular, so she adds replicas



The bill climbs, so she grabs the first fix



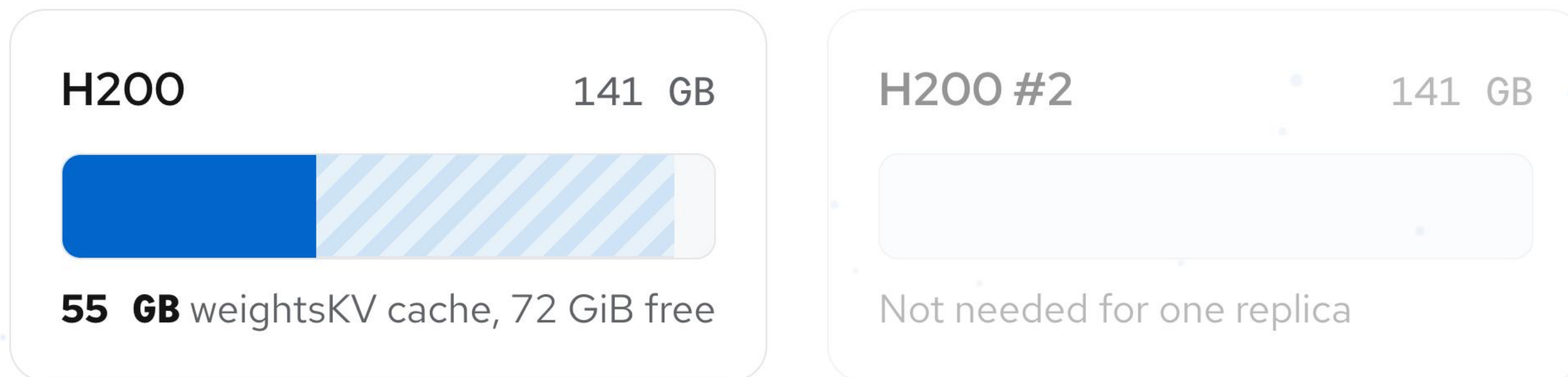
Marcus thinks the answers got worse



She stops looking for one winner

CHAPTER 1 MEET PRIYA

Priya shipped Qwen3.8-27B at BF16 on a full H200, and most of the card sits empty.



Then it got popular: 1 replica → 4, so  4 H200s

Direct message



Dana Finance · 9:02 AM

hey Priya, quick question 😊 why did the chatbot go to **2,920** GPU-hours a month?

Dana is typing a follow-up...

It is never a quick question.



Auto

Notes

CHAPTER 2 THE OBVIOUS FIX

Priya is a little naive, so she grabs **the first INT4 file she finds** and swaps it in.

BF16

2 bytes / weight



~55 GB a full H200

one 3g slice = 71 GB

FP8

1 byte / weight



~31 GB fits a 71 GB slice

INT4

½ byte + scales



~19 GB fits a 71 GB slice, or a 35 GB one

Four INT4 replicas fit on 71 GB slices, two per H200, so two H200s instead of four, Dana replies with 🎉, and Priya is a hero for about two weeks.

Weights as loaded by vLLM 0.24. BF16 51.1 GiB, FP8 28.5 GiB, INT4 17.7 GiB

CHAPTER 3 TWO WEEKS LATER

SUP-4471 · from Marcus on the data science team 🙄

PRIORITY: HIGH

“Did the model get worse? My analysis answers have felt off since the switch.”

Assigned to Priya · 3 more like it this week

👋 Who has quantized a model and then just **hoped the quality was fine?**



Auto

Notes

Three precisions Priya can try, each one solves a different problem

● BF16

THE BASELINE



1 full H200

BUYS YOU

The reference answer

BEST FOR

Your hardest questions, until the others are tested on them

● FP8

QWEN'S OWN FP8 BUILD



1 slice, 2 per H200

BUYS YOU

96% of a full card's speed, on a 71 GB slice

The slice has half the card's memory bandwidth

BEST FOR

Everyday traffic, two replicas per H200

● INT4

GPTQ VIA LLM COMPRESSOR



1 slice, 2 per H200

BUYS YOU

86% of a full card's speed, on a 71 GB slice

Fits a 35 GB slice too, but slower in eager mode

BEST FOR

Long contexts and big batches: 1.2M tokens of FP8 KV cache

No measurable quality loss: FP8 and INT4 matched BF16 on GSM8K (1,319) and MMLU-Pro (280)

● On top of BF16, she can enable **MTP (speculative decoding)**: about 2.3× faster answers, with BF16 quality. Best for Anyone waiting on the answer.

Let's find out.

 Pick a question, and you'll see how every setup answered it.

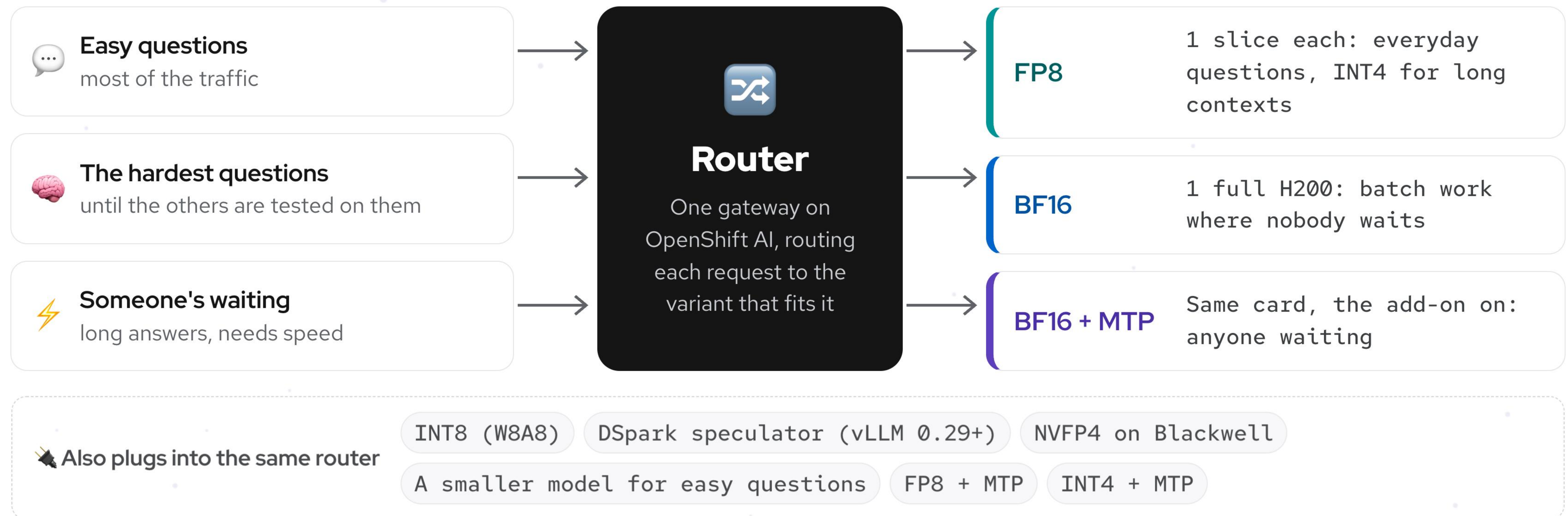
● **RECORDED RUNS**



Auto

Notes

There wasn't one winner, so Priya added a router



Only the hard questions get a whole H200, so she needs fewer cards than running everything on BF16.

Marcus closes the ticket  and Dana sends another 

WRAP-UP

What to take home

The points we made

- 1 **On Hopper, try FP8 first.** A 71 GB slice, 96% of a full card's speed for one request, and no measurable loss.
- 2 **INT4 is for long contexts.** Small weights leave room for 1.2M tokens of KV cache on one slice, at 86% of the speed.
- 3 **Spec decode is an add-on, built in here.** On BF16, 2.3× faster for one request; 80 to 90% of BF16's throughput under load, which is draft overhead.
- 4 **No measurable quality loss, but verify.** FP8 and INT4 matched BF16 on GSM8K and MMLU-Pro. A small gap compounds over thousands of tokens, so test on your own prompts.
- 5 **A router** lets each one run where it's best, once your traffic is big and mixed enough to run more than one.

Things to think about

- ? How much of your traffic really needs full precision?
- ? Is your traffic big enough to justify more than one pool?
- ? Is your target cost per token, or time to answer?
- ? Do you have an eval set built from real prompts?

 **Technical questions**, answered in the repo: github.com/MarkellR-RedHat/pytorch-quantization-demo#technical-questions



Auto

Notes

Here's all the source code I used

 github.com/MarkellR-RedHat/pytorch-quantization-demo

 If you want to see how I did this, it's all in there

 Two tracks in this deck, Llama 3.1 70B on H200s and Qwen3.8-27B on MIG slices. **T** on the title slide switches.

 Model checkpoints on huggingface.co/RedHatAI – every build linked in the repo README

Thank you 🙏

Markell Rawls · AI Developer Advocate, Red Hat · mrawls@redhat.com



 **SCAN FOR THE REPO**