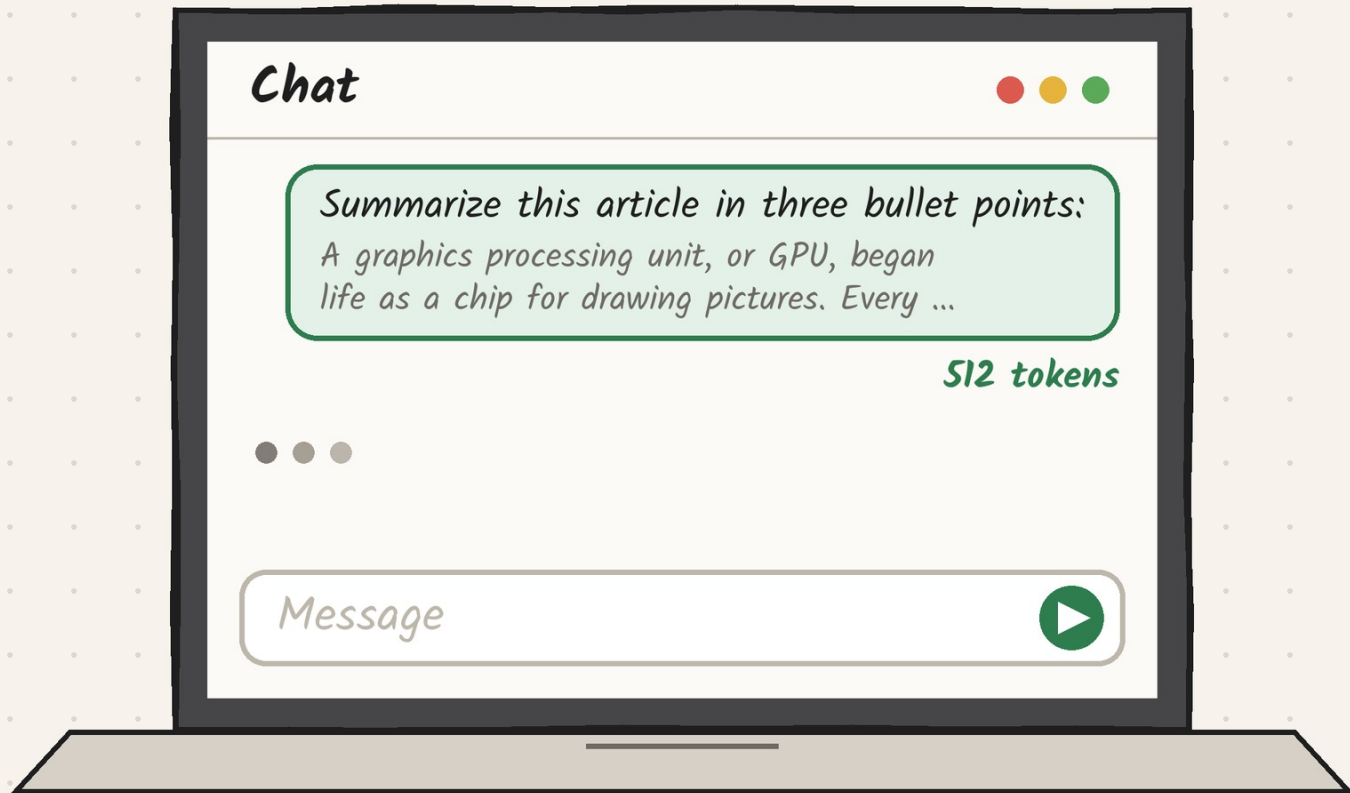


To a GPU, Qwen3.8-27B isn't a model.
It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU



the prompt travels from your laptop to the 1x NVIDIA H100 GPU server
(network time is not counted: every time in this video is measured on the server)

1/3

It starts with your prompt
an article to summarize: 512 tokens

To a GPU, Qwen3.8-27B isn't a model.
It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

tokenizer (CPU)

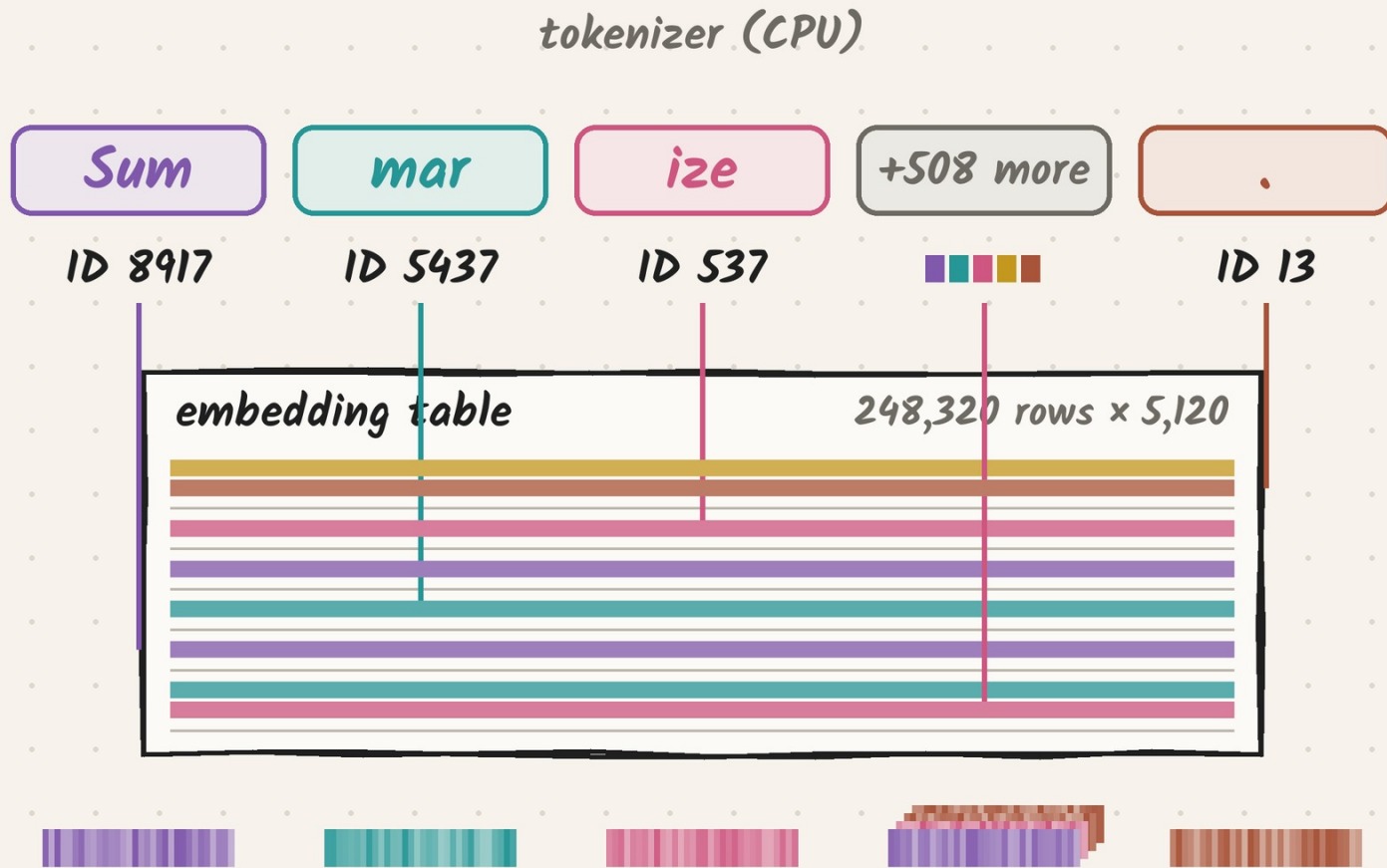


2/3

The tokenizer (CPU) splits it
into 512 token IDs: Sum · mar · ize ...

To a GPU, Qwen3.8-27B isn't a model.
It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU



3/13

Each ID picks a row of 5,120 numbers
from the embedding table

To a GPU, Qwen3.8-27B isn't a model.
It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

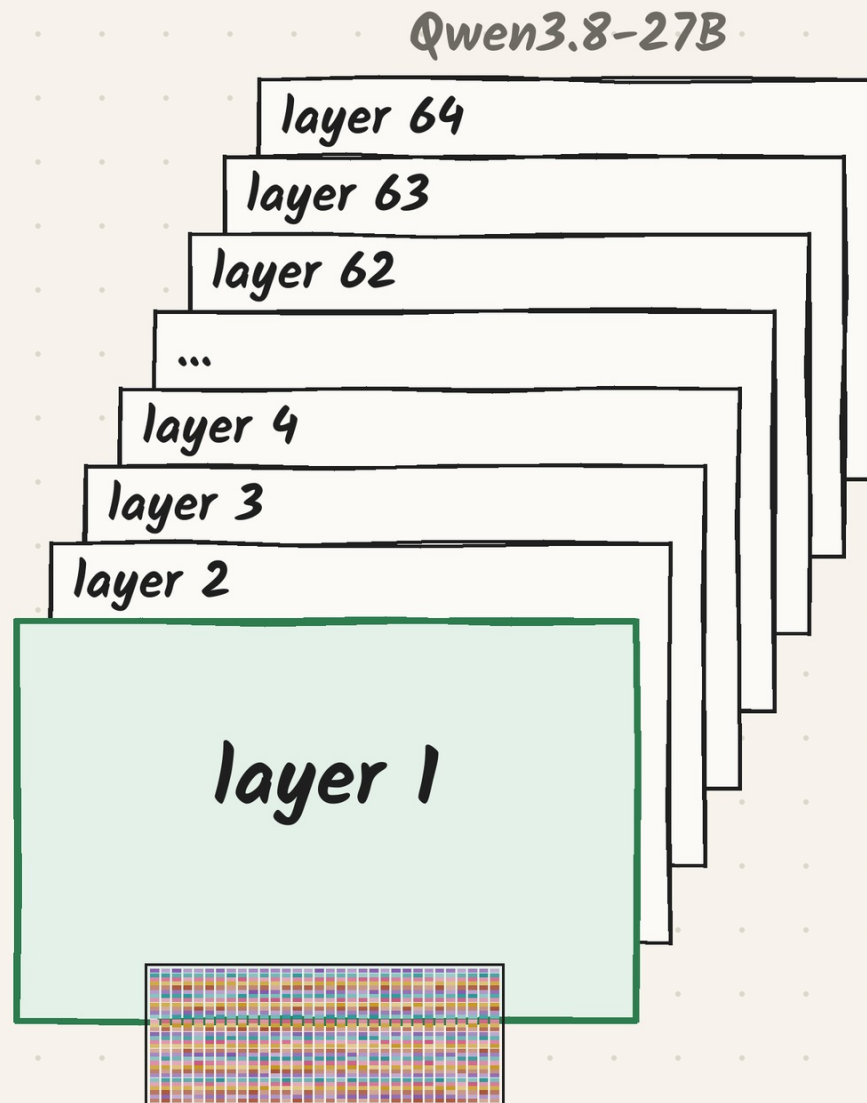


4/13

Stack them: 512 x 5,120
one row per token

To a GPU, Qwen3.8-27B isn't a model.
It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU



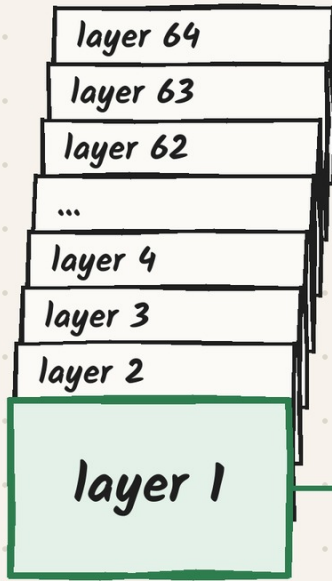
5/13

Prompt phase: all 512 go in together
the whole matrix enters layer 1

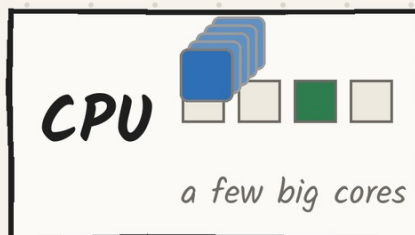
To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



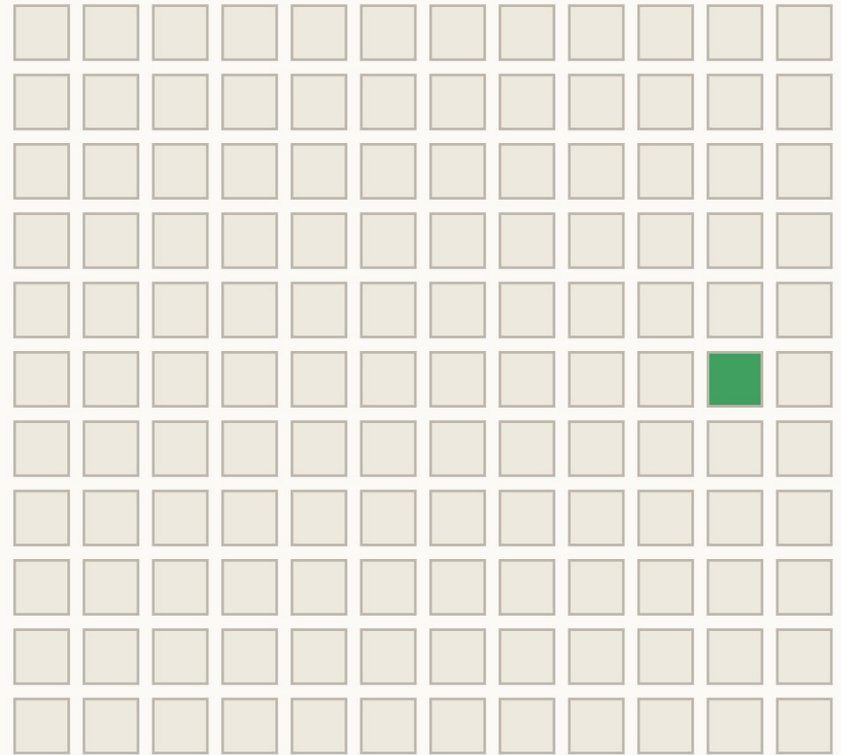
-  RMSNorm
-  GEMM Q, K, V
-  Attention
-  GEMM output
-  RMSNorm
-  GEMM MLP up
-  SiLU
-  GEMM MLP down



1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



6/13

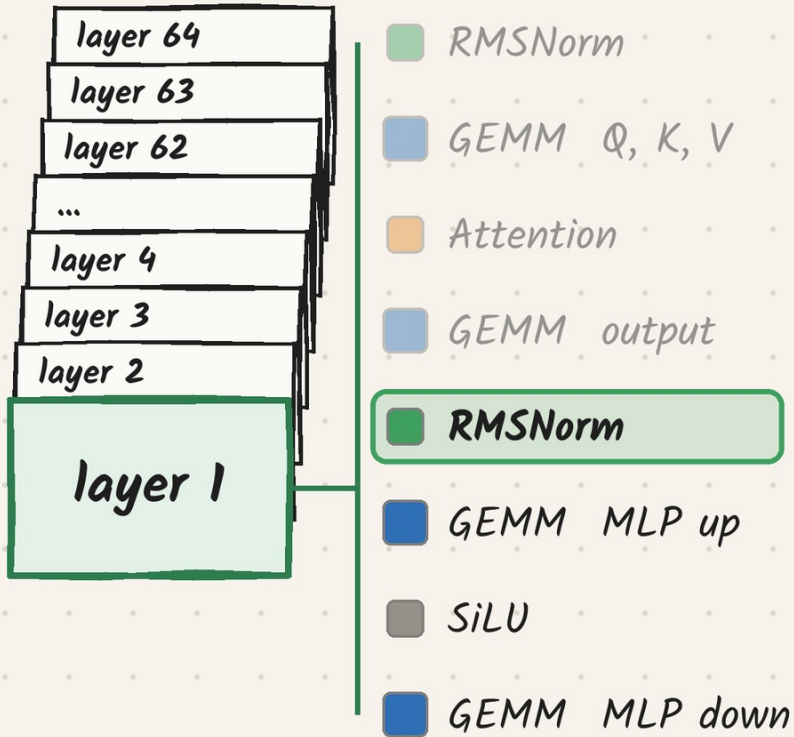
Layer 1 = a chain of kernels

the CPU sends them, the GPU runs them in order

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



CPU
a few big cores

1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



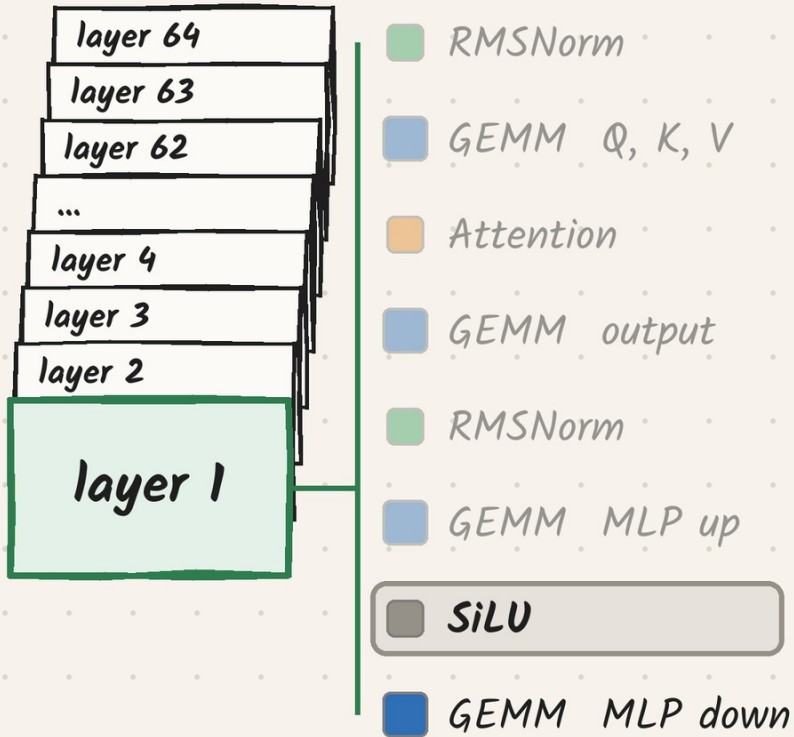
7/13

Q, K, V GEMM saves K and V of all 512 tokens, in the KV cache

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:

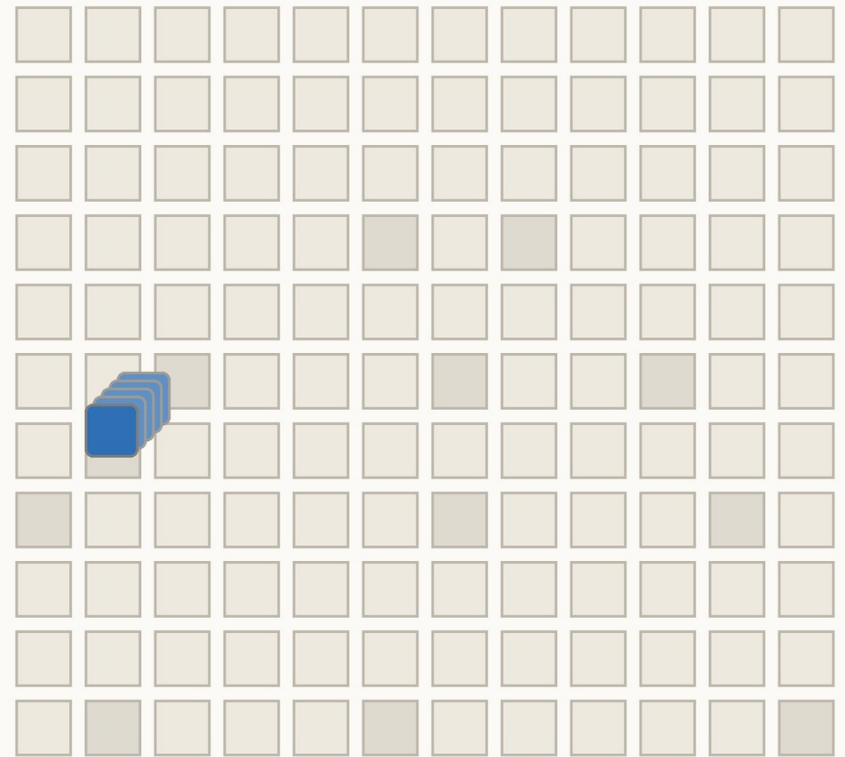


CPU
a few big cores

1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



8/13

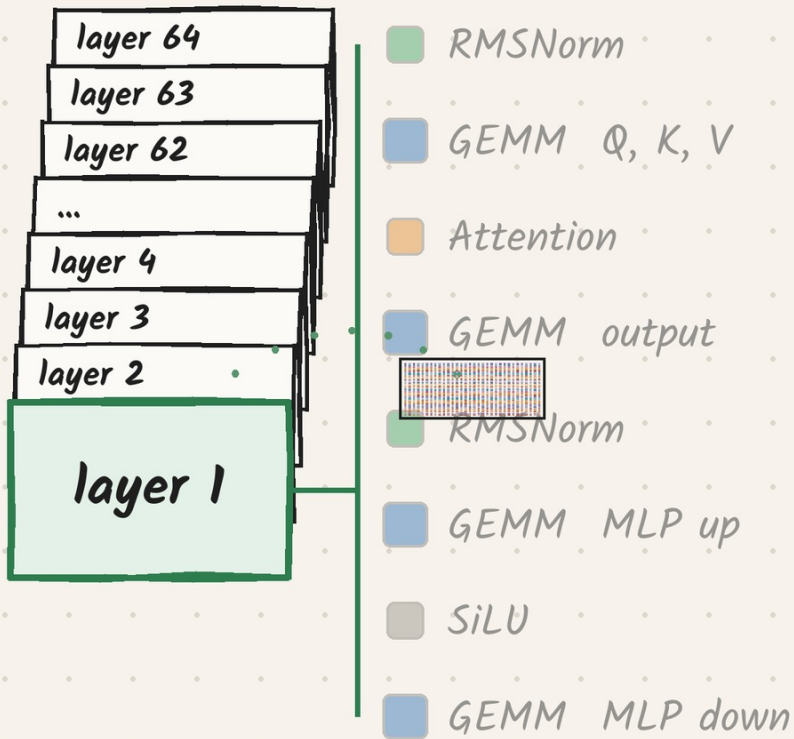
Each kernel spreads over 132 SMs

GEMMs (blue): 92.6% of the GPU time

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



CPU
a few big cores

1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



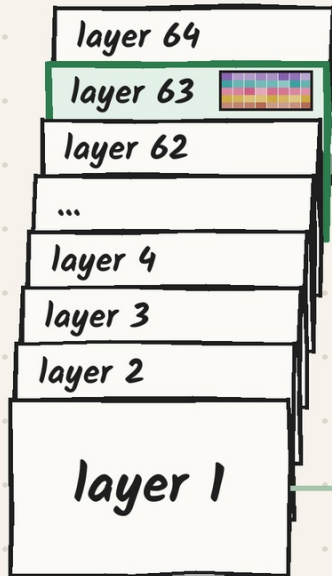
9/13

Layer 1's output = layer 2's input still 512 × 5,120

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



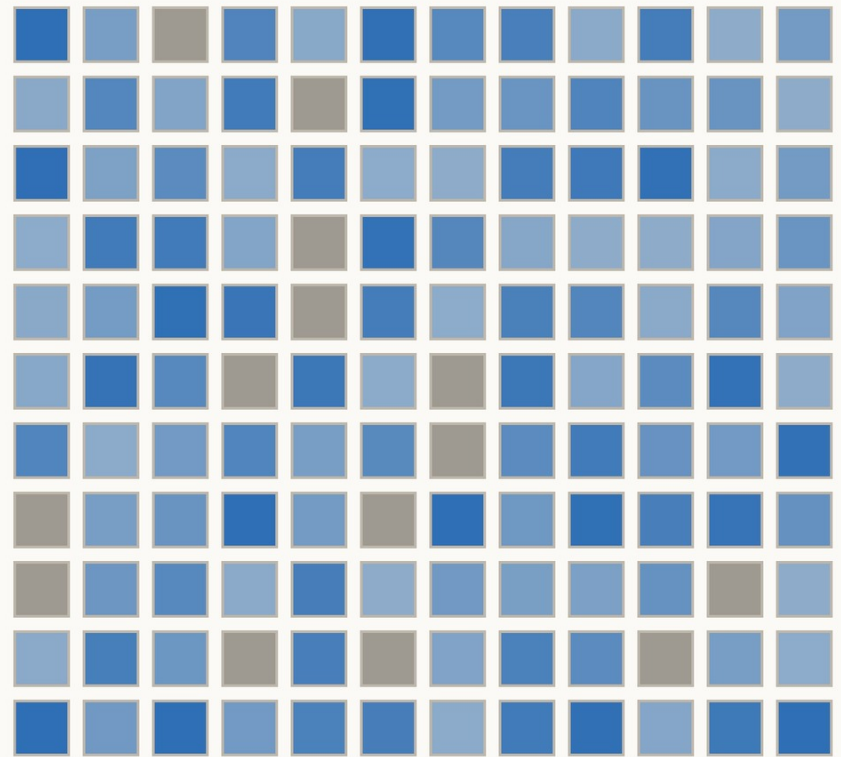
- RMSNorm
- GEMM Q, K, V
- Attention
- GEMM output
- RMSNorm
- GEMM MLP up
- SiLU
- GEMM MLP down

CPU 
a few big cores

1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache
in GPU memory



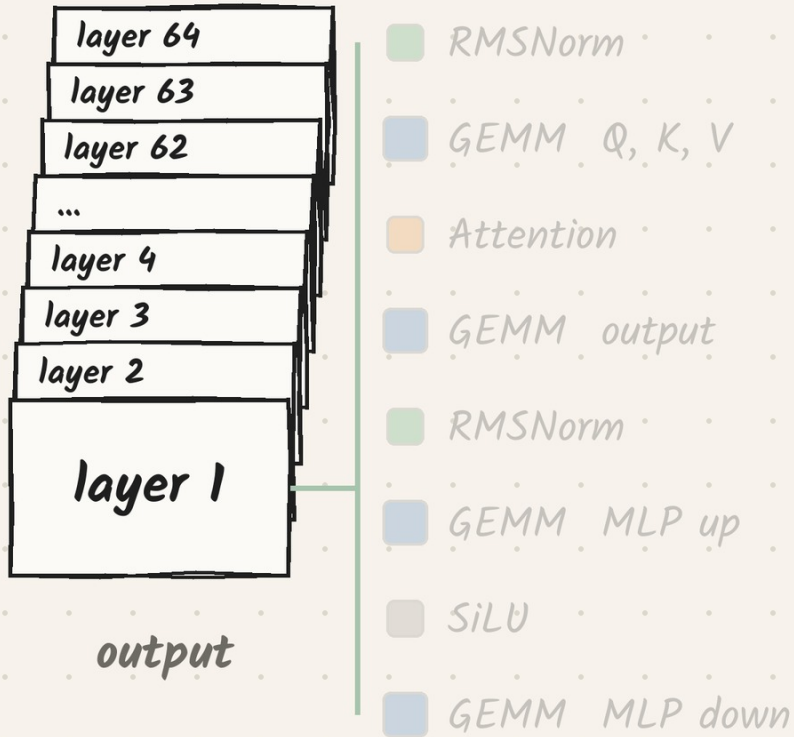
10/13

Then layers 2, 3 ... 64 now: layer 63

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



CPU 
a few big cores

1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache
in GPU memory



11/13

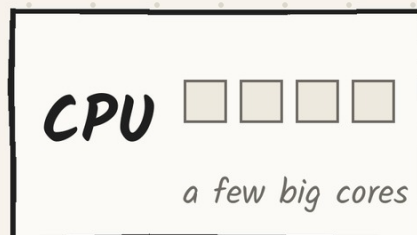
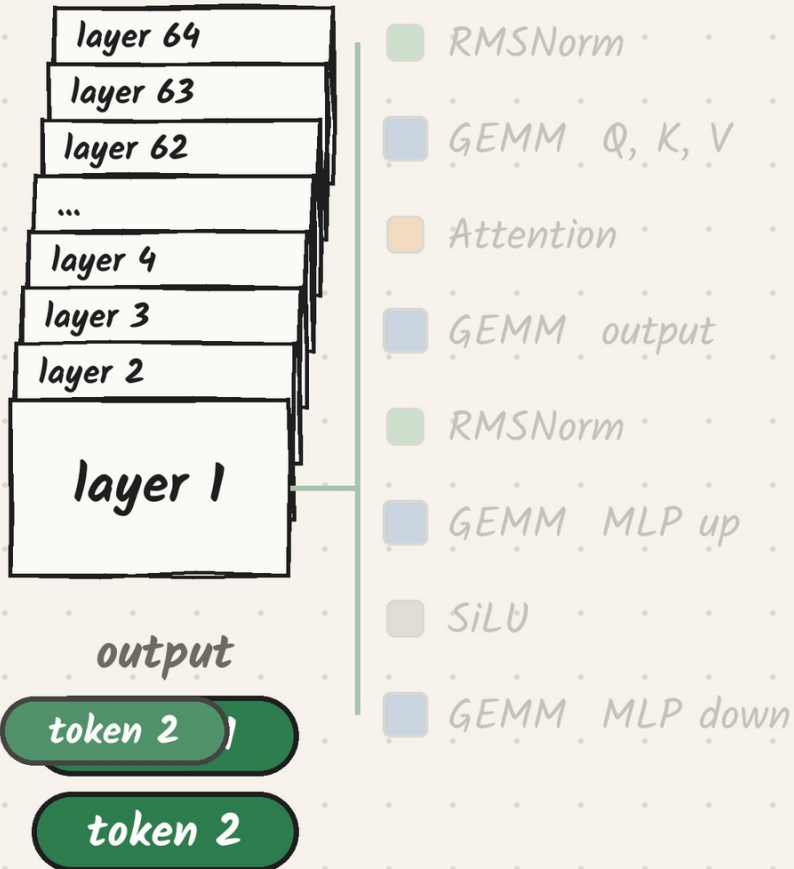
Output head picks token 1

prompt phase: 512 tokens in 54.3 ms

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



512 prompt tokens

new tokens

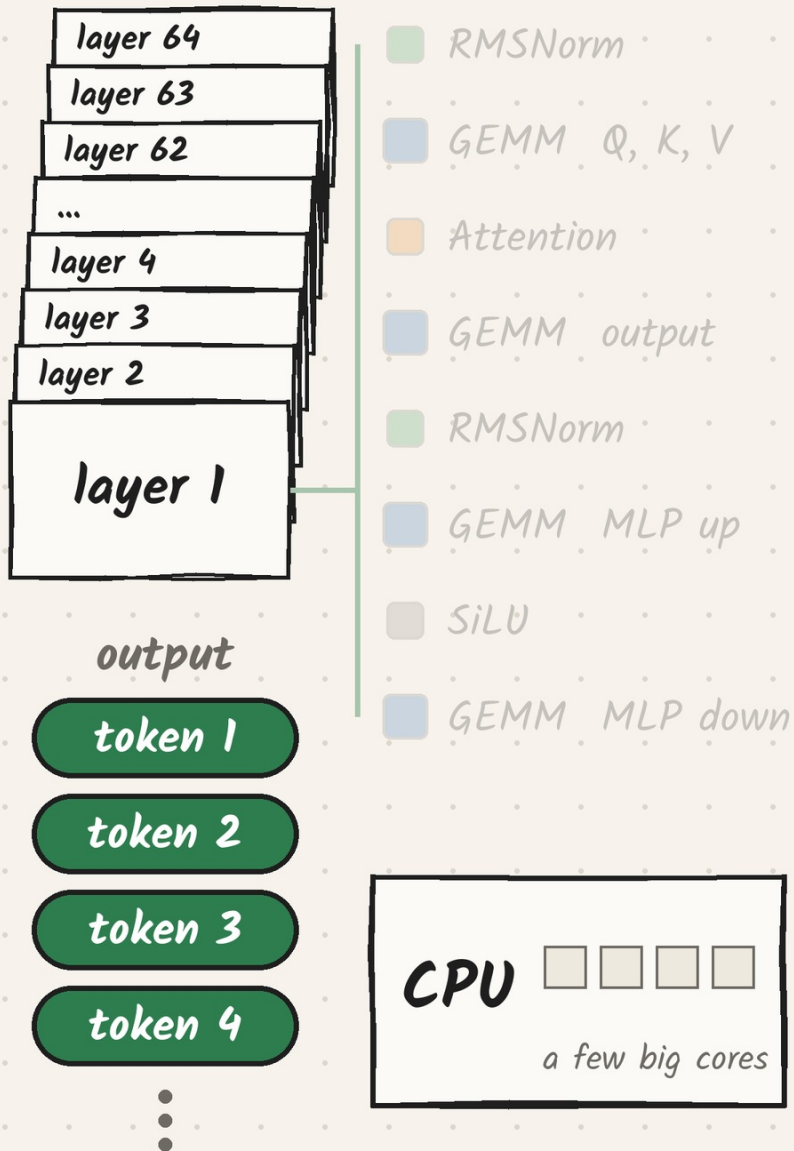
12/13

Decode: only the new token goes in, with K, V of 512 prompt tokens from KV cache, stored in prefill

To a GPU, Qwen3.8-27B isn't a model. It's a loop of 902 kernels.

how Qwen3.8-27B makes tokens on 1x NVIDIA H100 GPU

kernels in a layer:



1x NVIDIA H100 GPU

132 SMs

989 TFLOP/s BF16 · 3.35 TB/s memory · 528 Tensor Cores



KV cache

in GPU memory



One decode step: 902 kernels

20.55 ms per token, 48.7 tokens/s

13/13