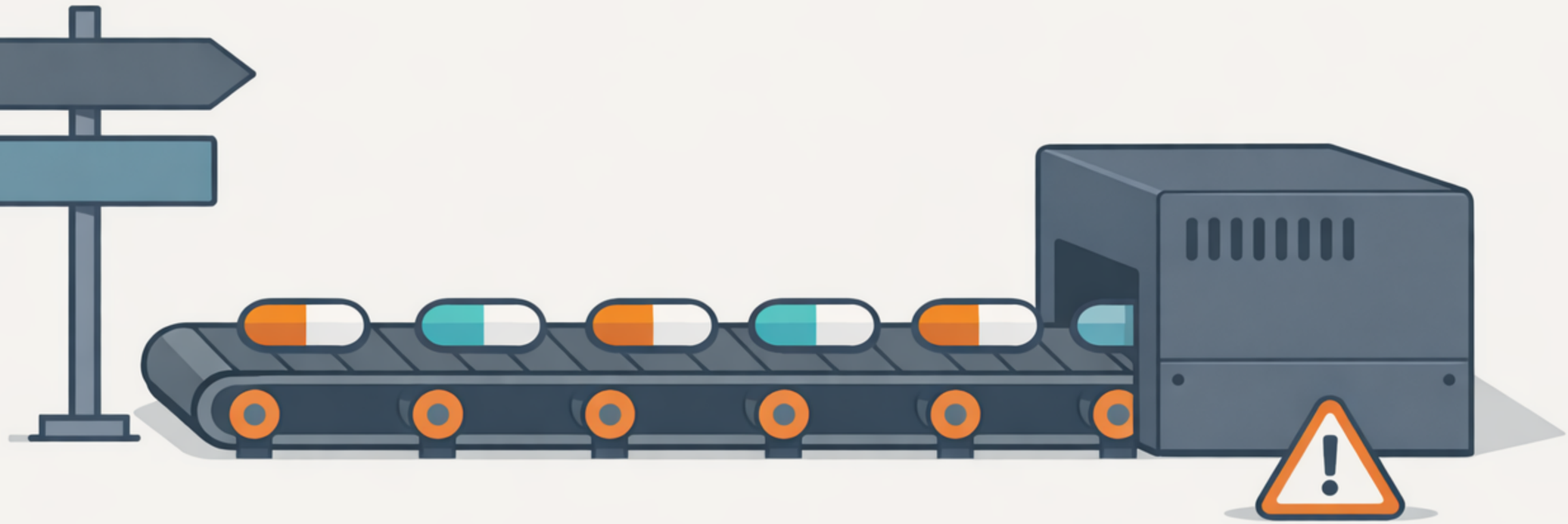
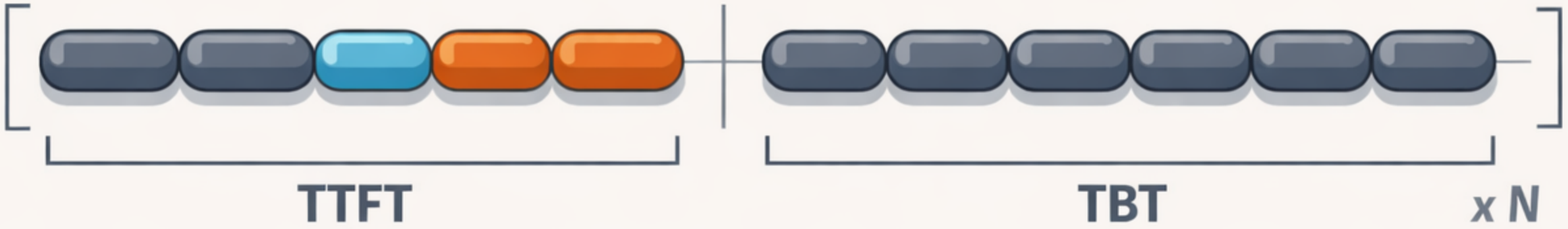
What matters

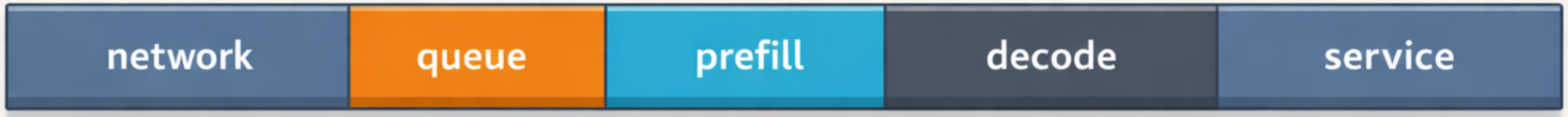
What makes it slow?



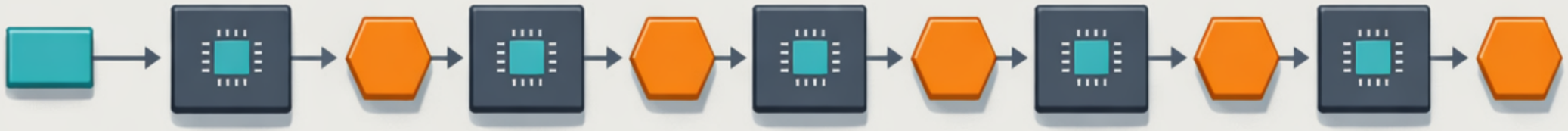
TTFT vs TBT



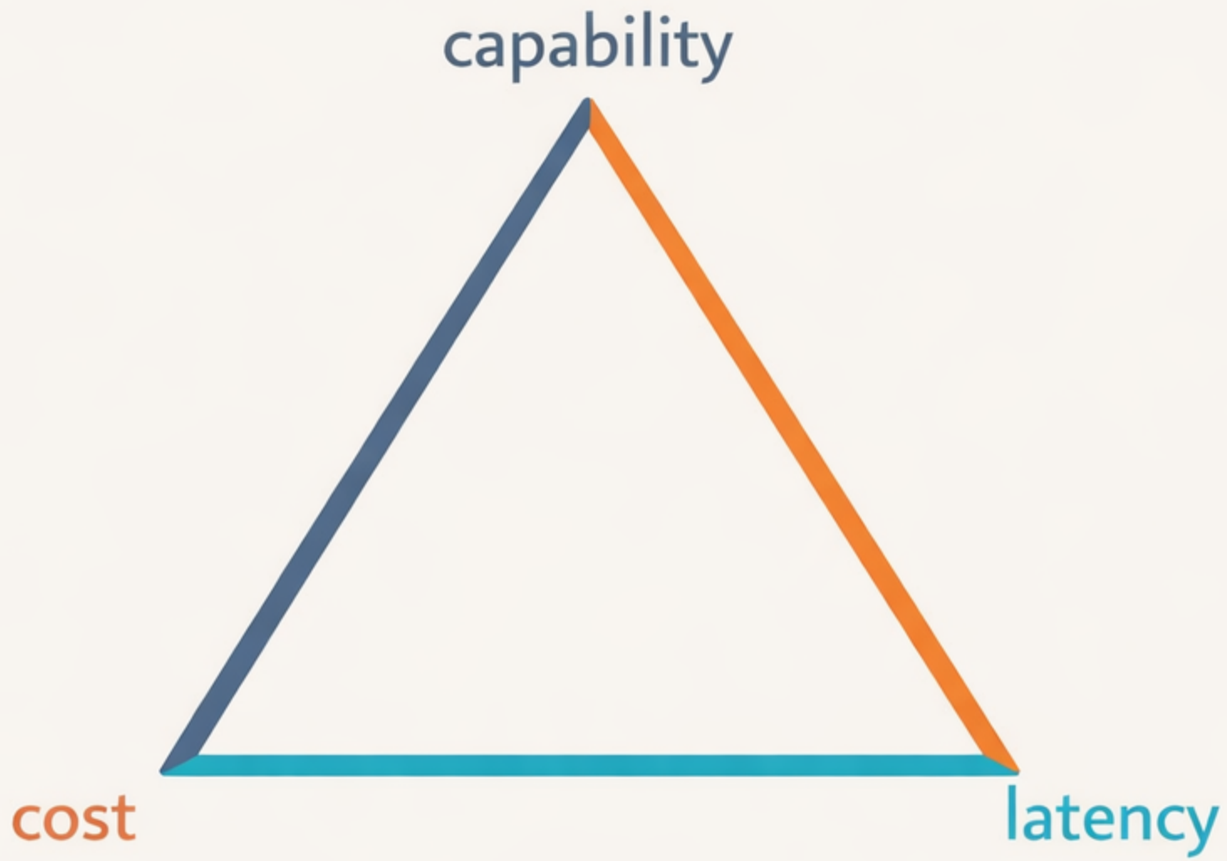
End-to-end latency



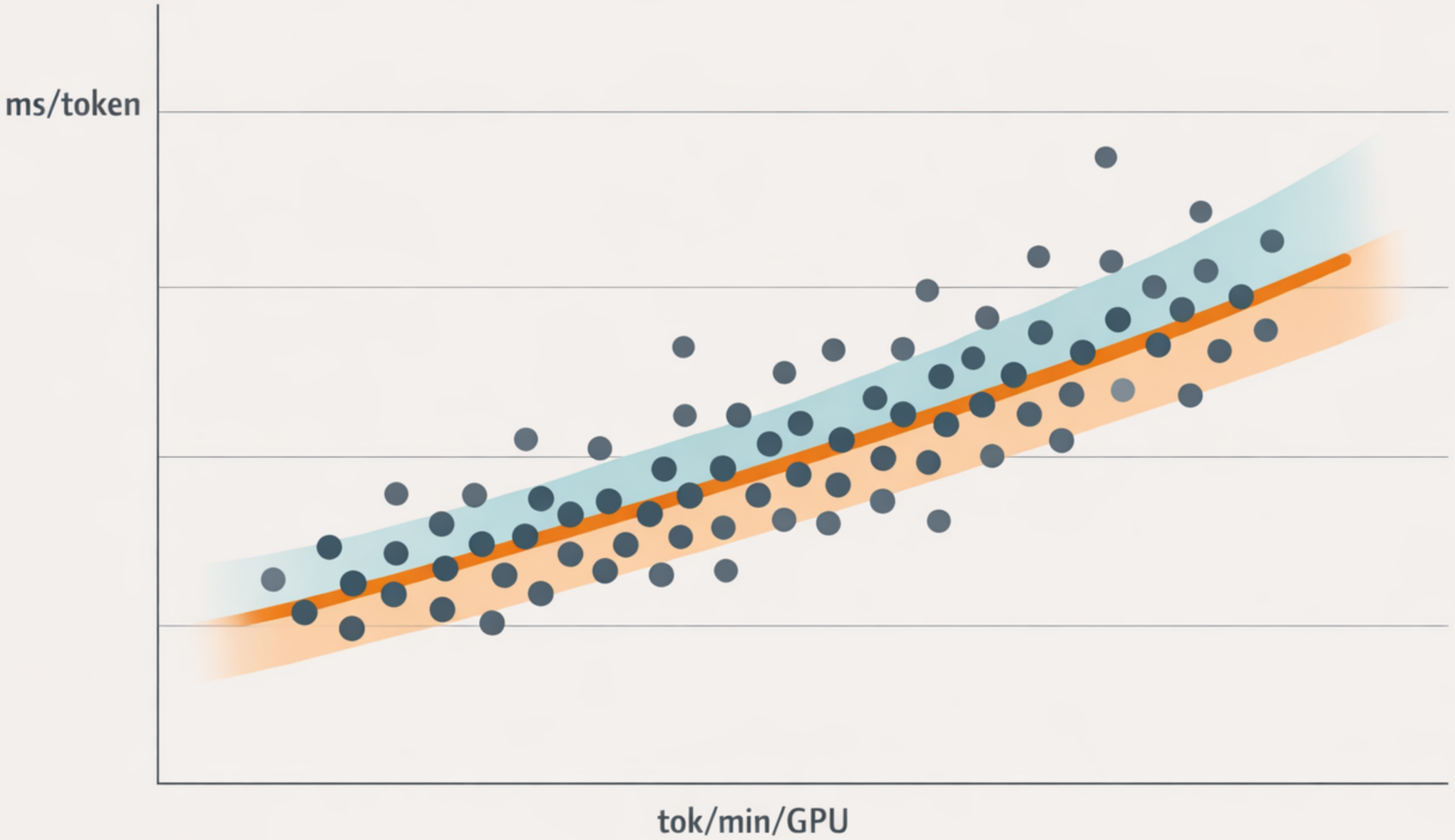
Why output dominates



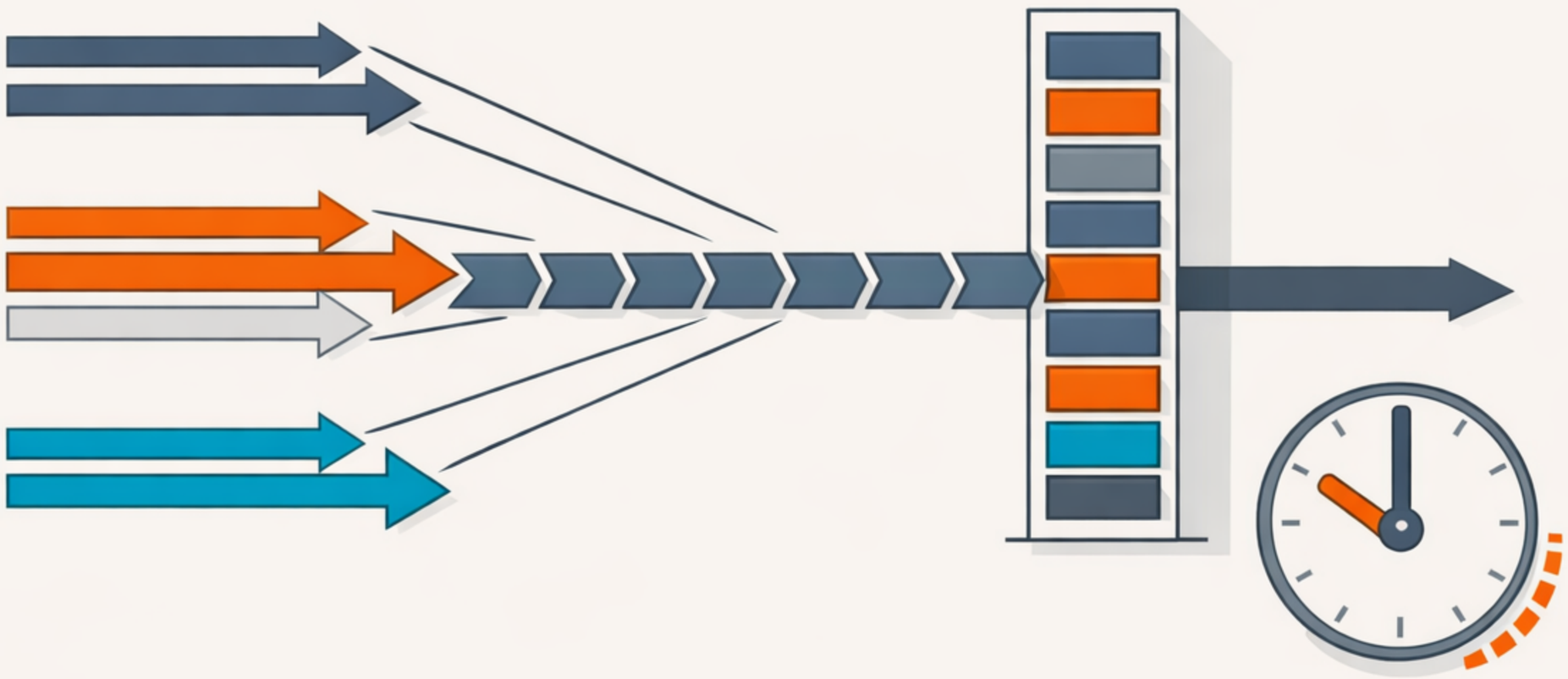
Tradeoffs



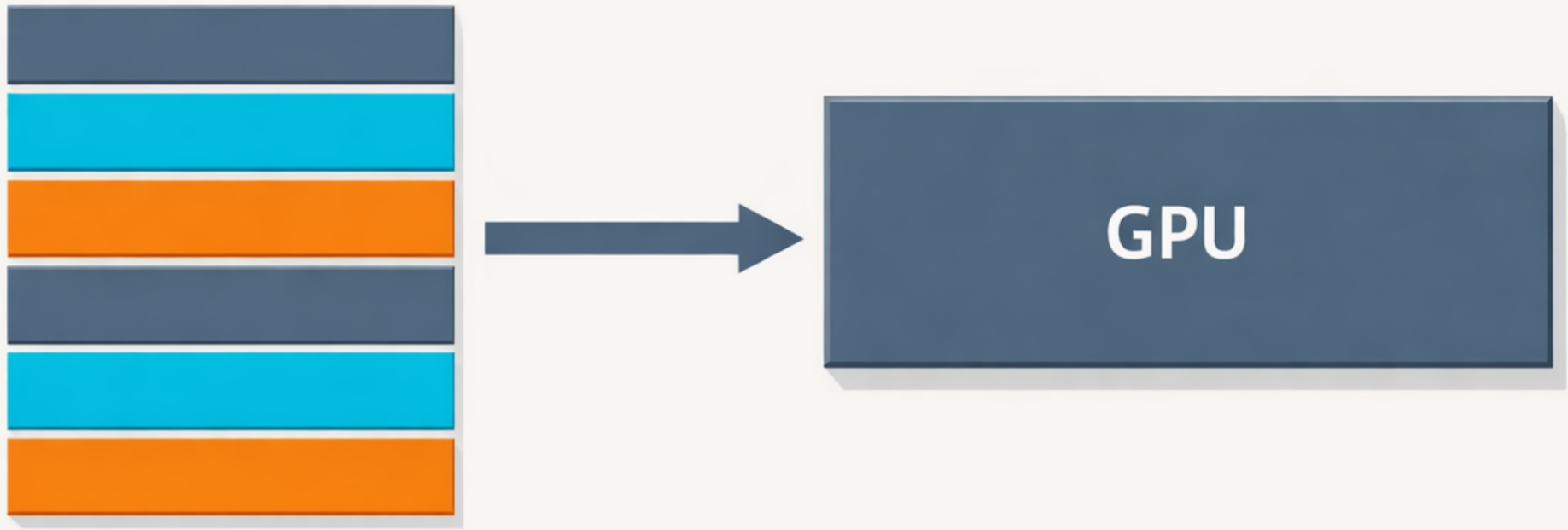
Trade-off Between Latency and Throughput



Congestion

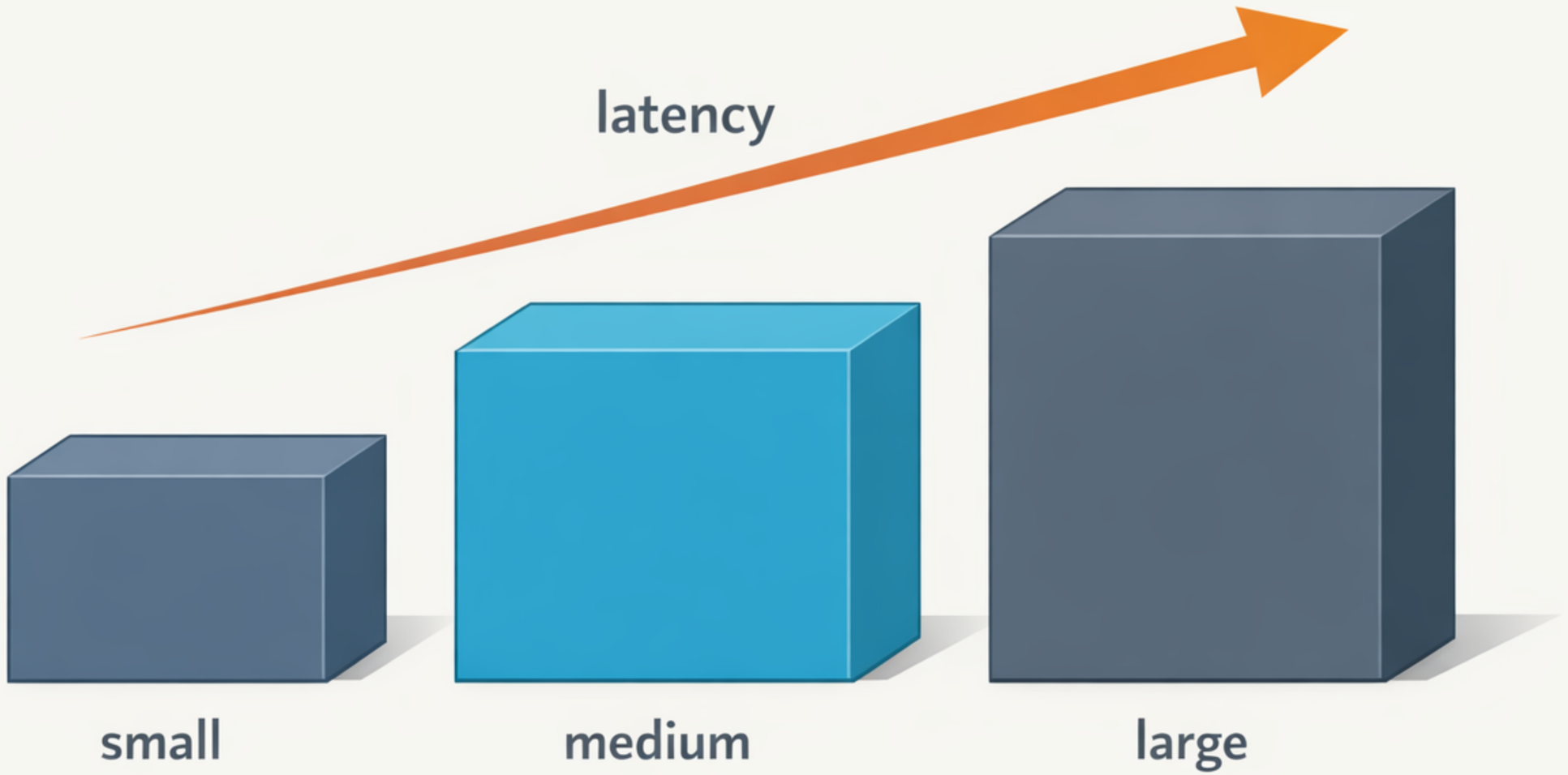


Batching



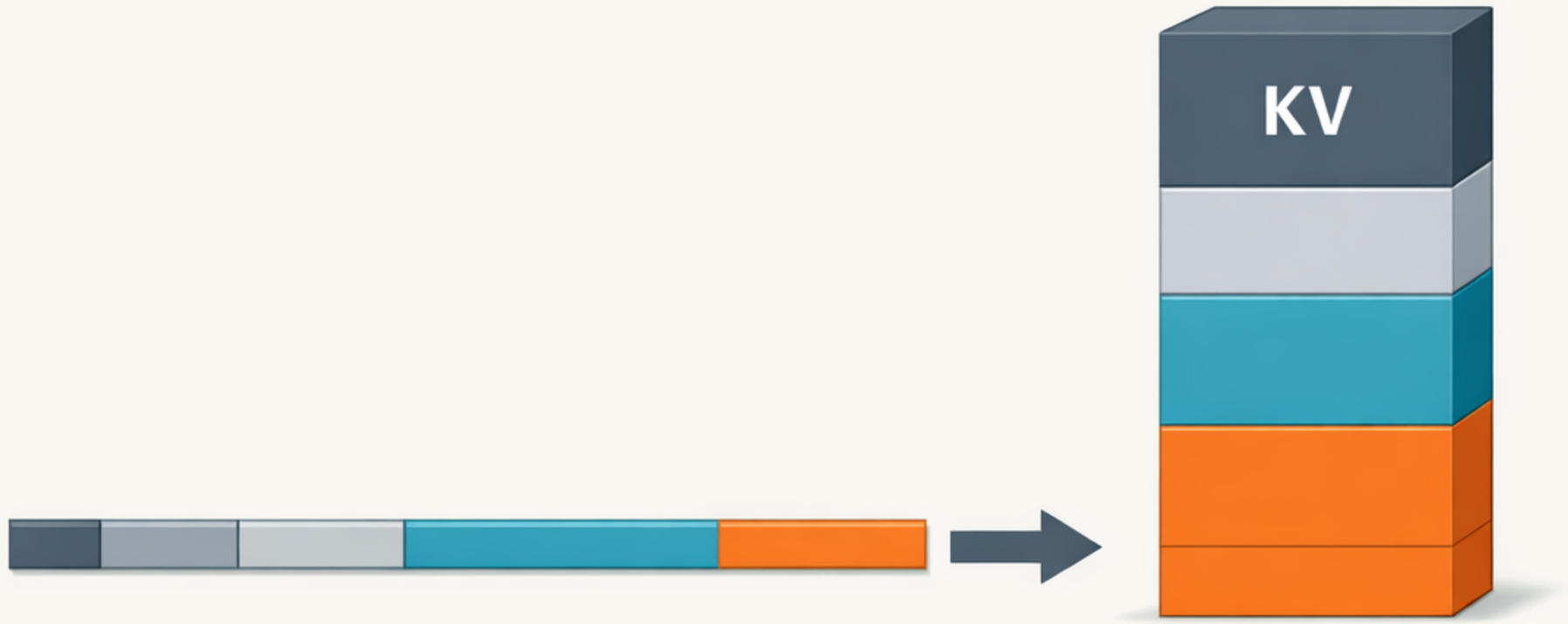
Routing

Model size



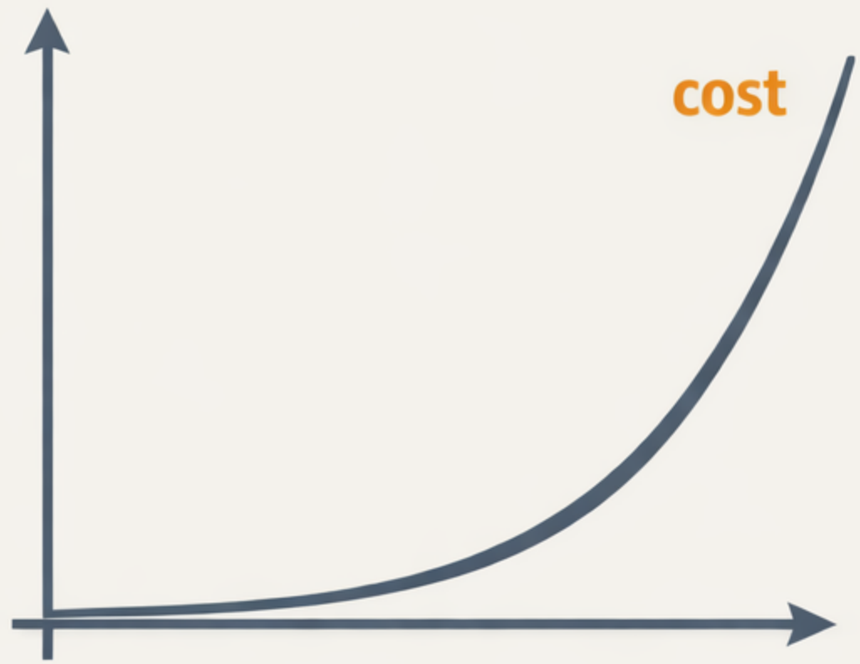
Input vs output

Context + KV



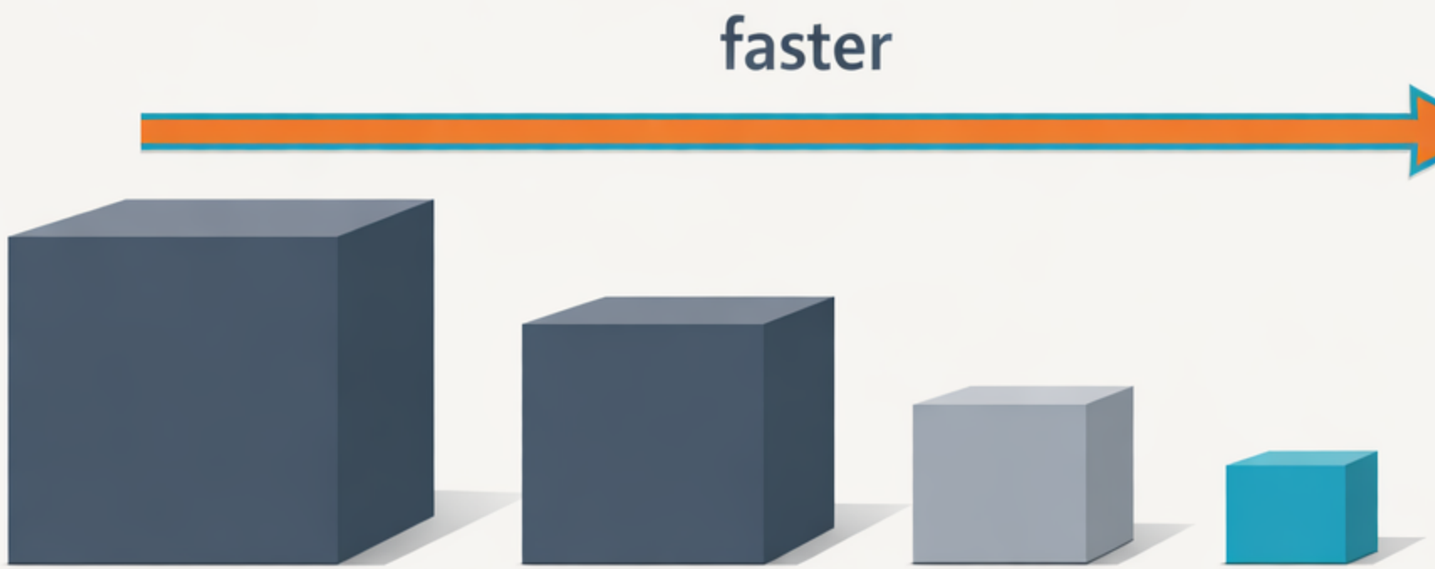
Cached prefix

Long requests

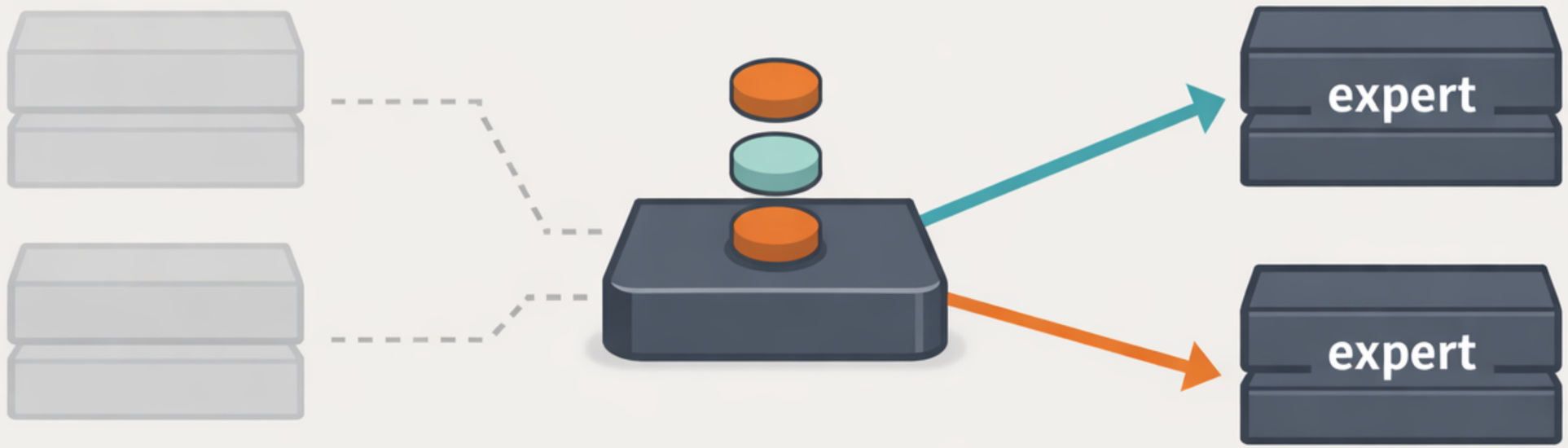


Speedups

Smaller models



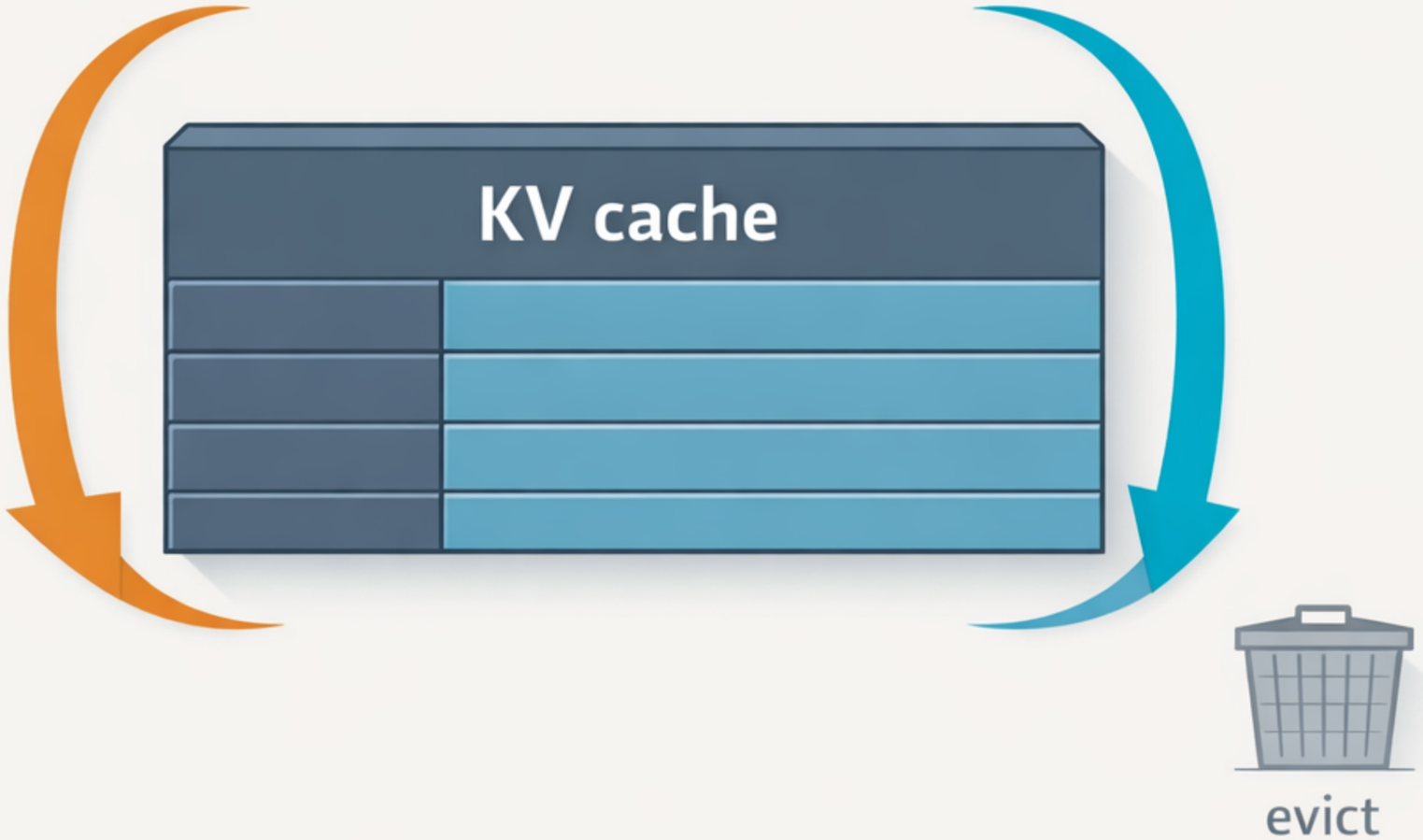
MoE



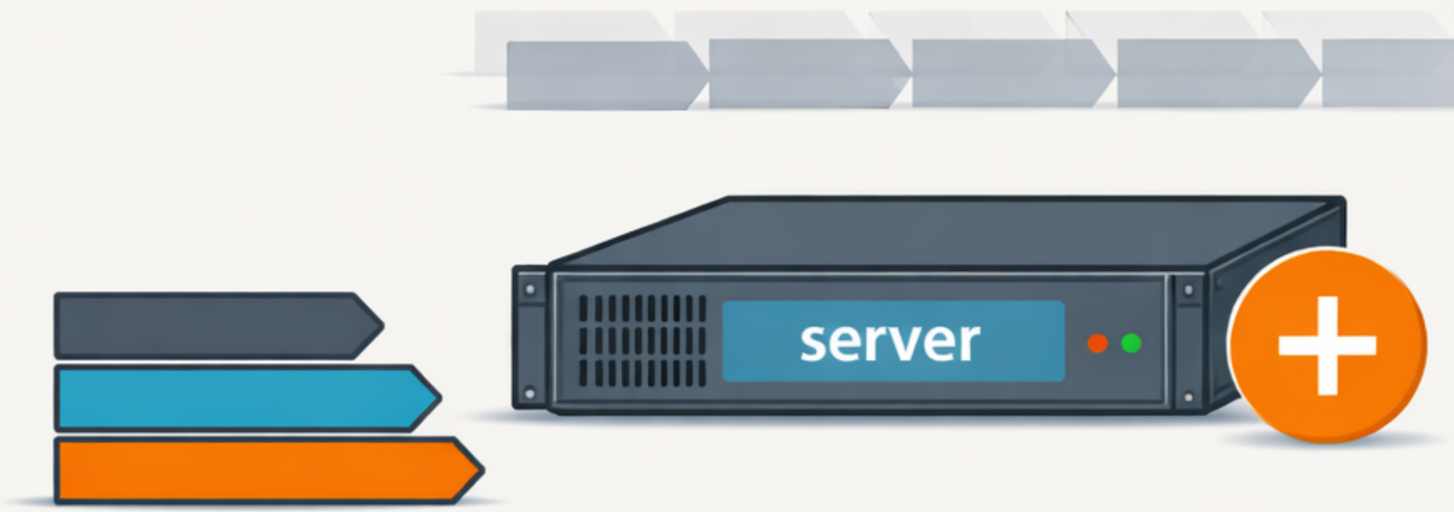
Proxy model

Quantization

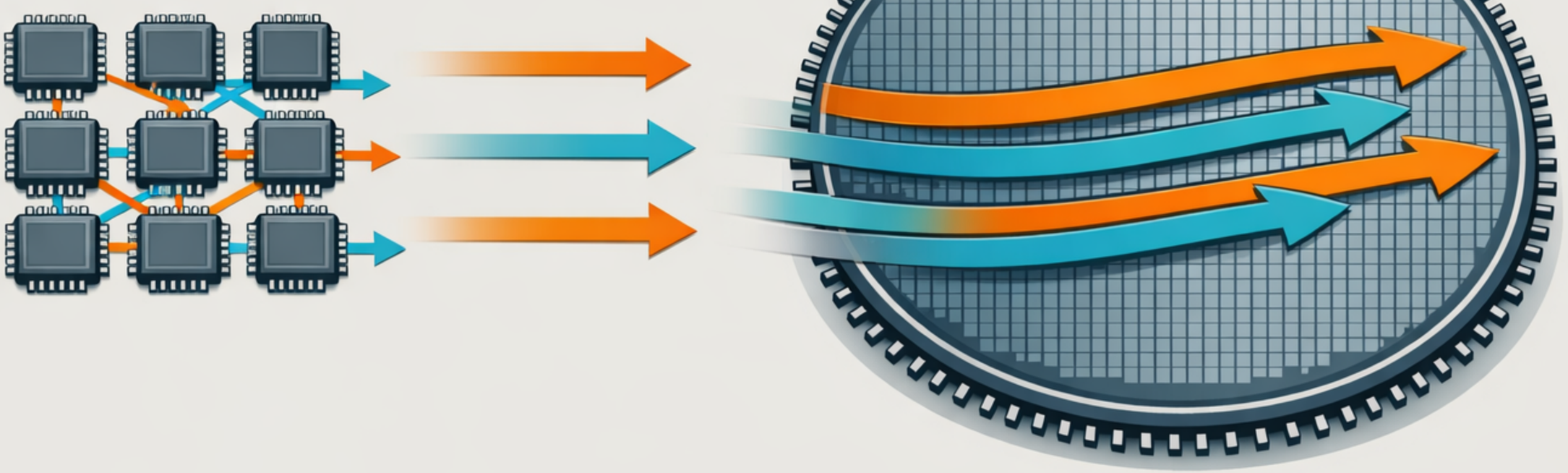
KV cache



Capacity



Wafer-scale



What matters

Do the work