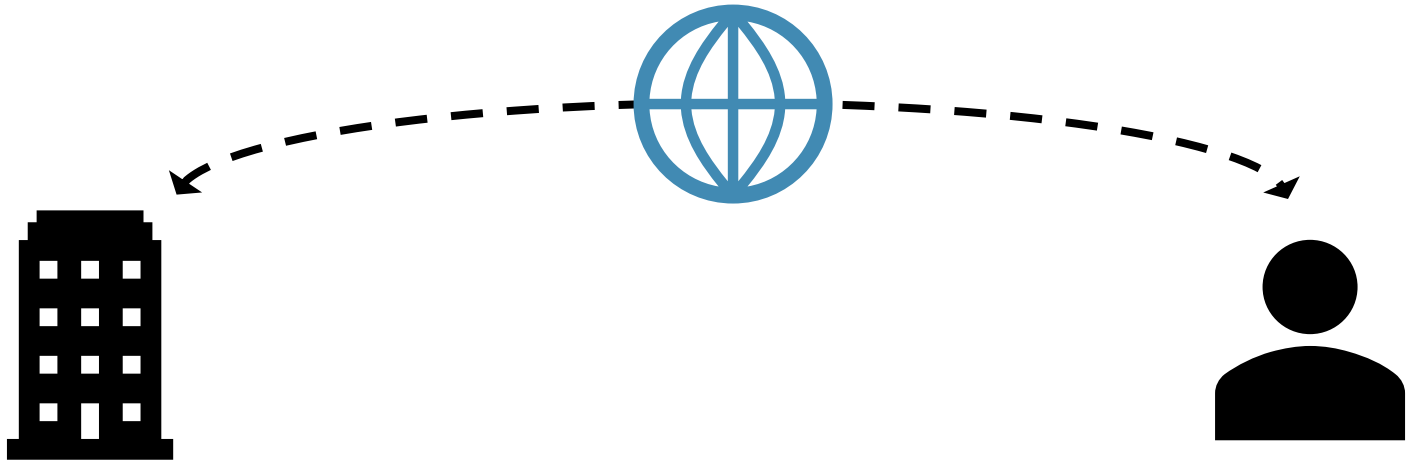
SpecEdge: Scalable Edge-Assisted Serving Framework for Interactive LLMs

Jinwoo Park, Seunggeun Cho, Dongsu Han

KAIST

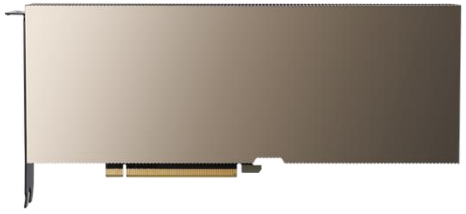
Opportunity lies in edge devices

- Can we utilize available edge (user-side) devices for efficient LLM serving?



Server-side

User-side



NVIDIA A100 80GB

$5.05/h

312 TFLOPS for FP 16 operations

GeForce RTX 4090

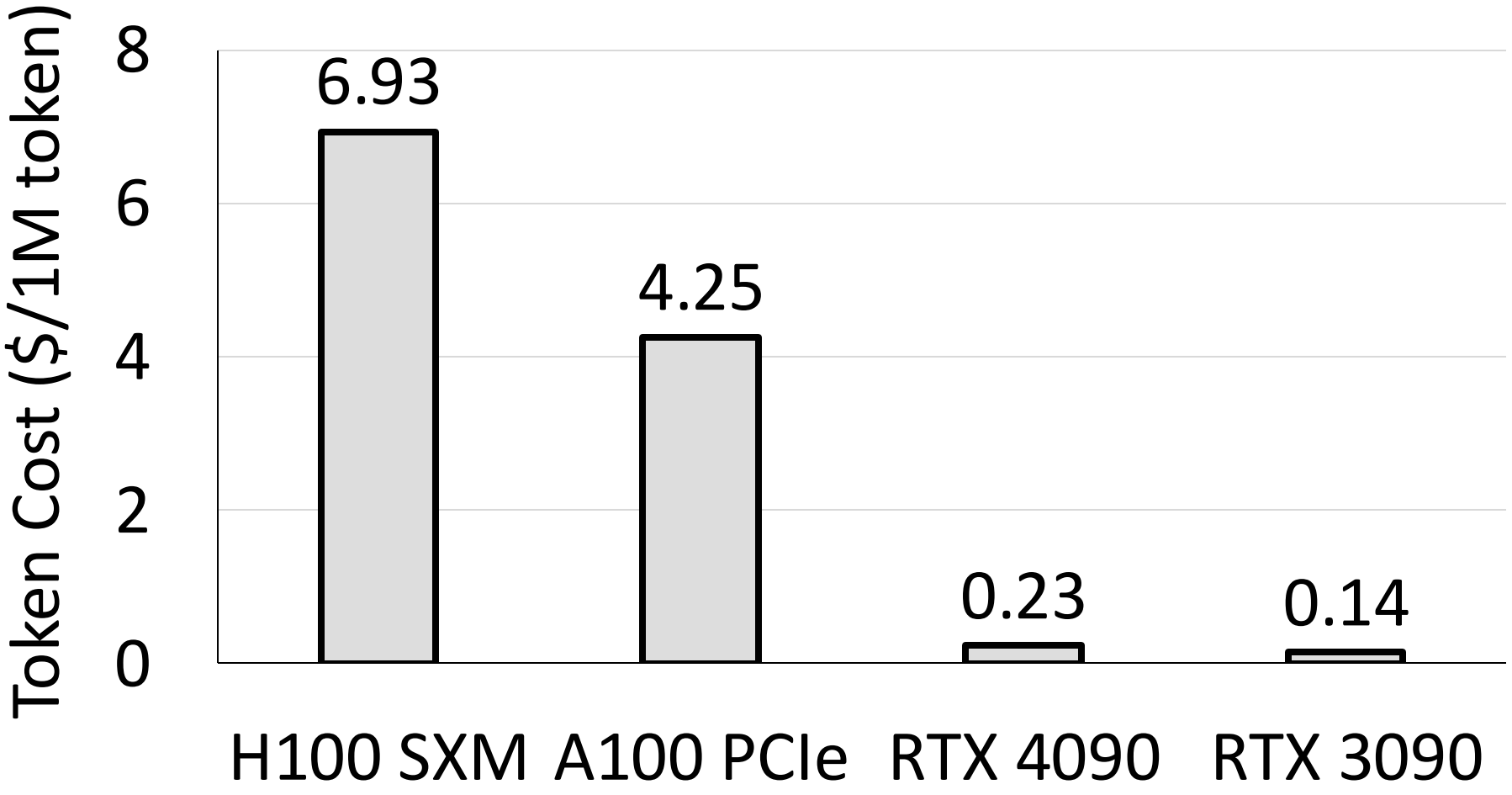
$0.35/h*

330.3 TFLOPS for FP 16 operations

* https://www.vast.ai/

Cost efficiency of Consume-Grade GPUs

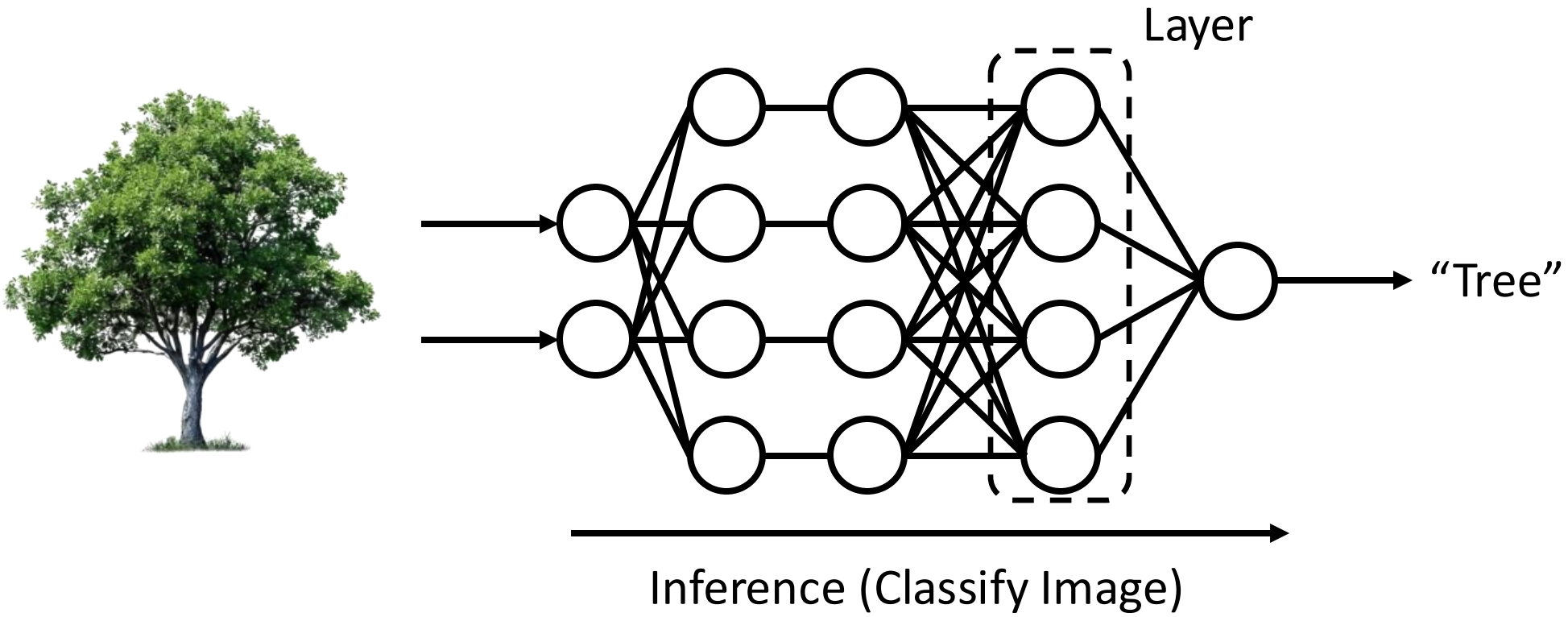
- Benchmarks run on GPUs with the Qwen2-0.5B model (RunPod, 2025).



* https://www.runpod.io/dir/gpu-benchmarks

Conventional split computing

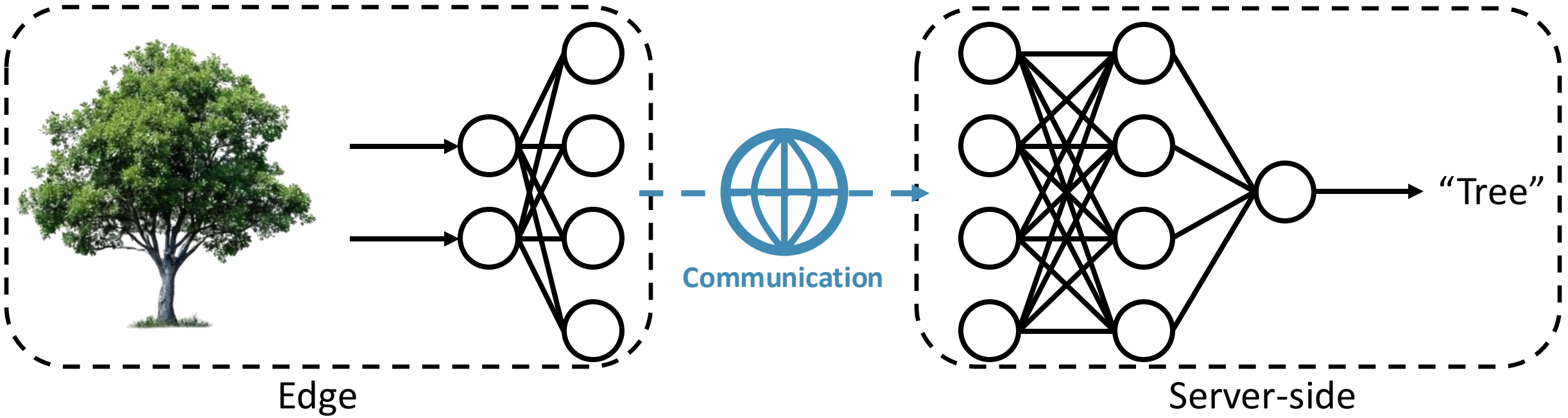
- Divides layers or neural networks between cloud and the edge
  - Reduces computational load on the server
  - Minimize network communication costs by transmitting intermediate tensors instead of raw data



* Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge, Kang et al. ASPLOS '17

Conventional split computing

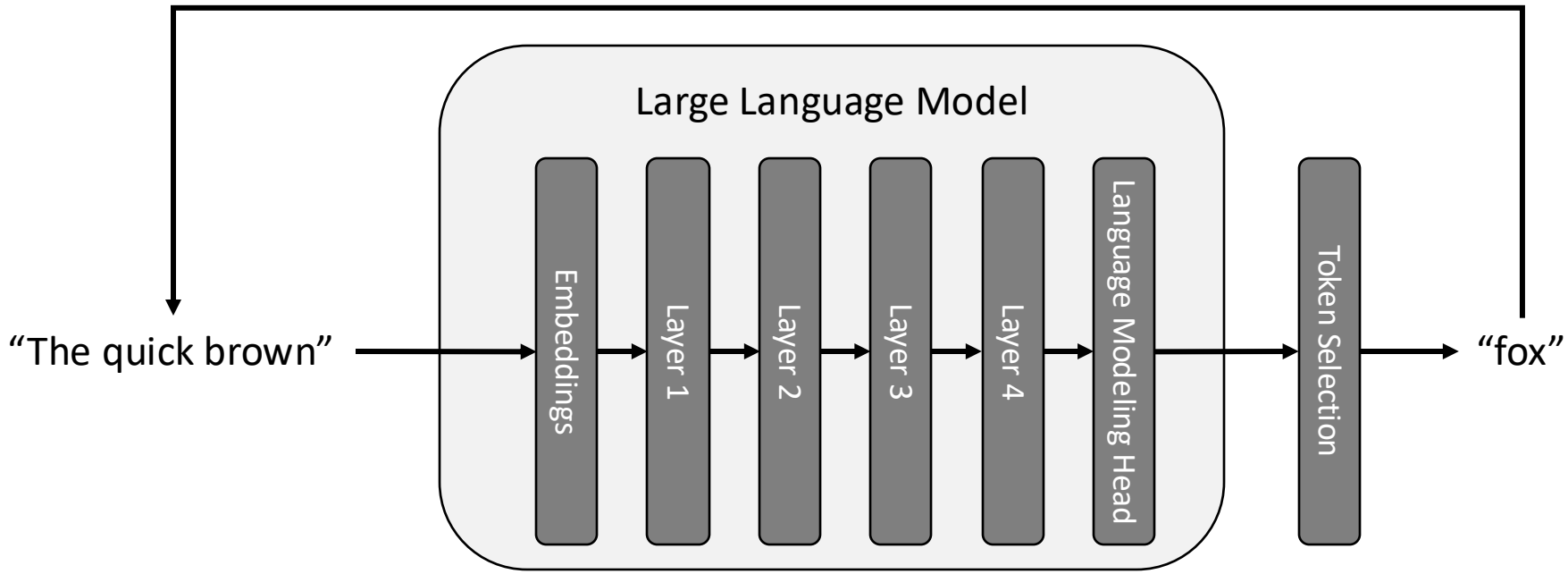
- Divides layers or neural networks between cloud and the edge
  - Reduces computational load on the server
  - Minimize network communication costs by transmitting intermediate tensors instead of raw data



* Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge, Kang et al. ASPLOS '17

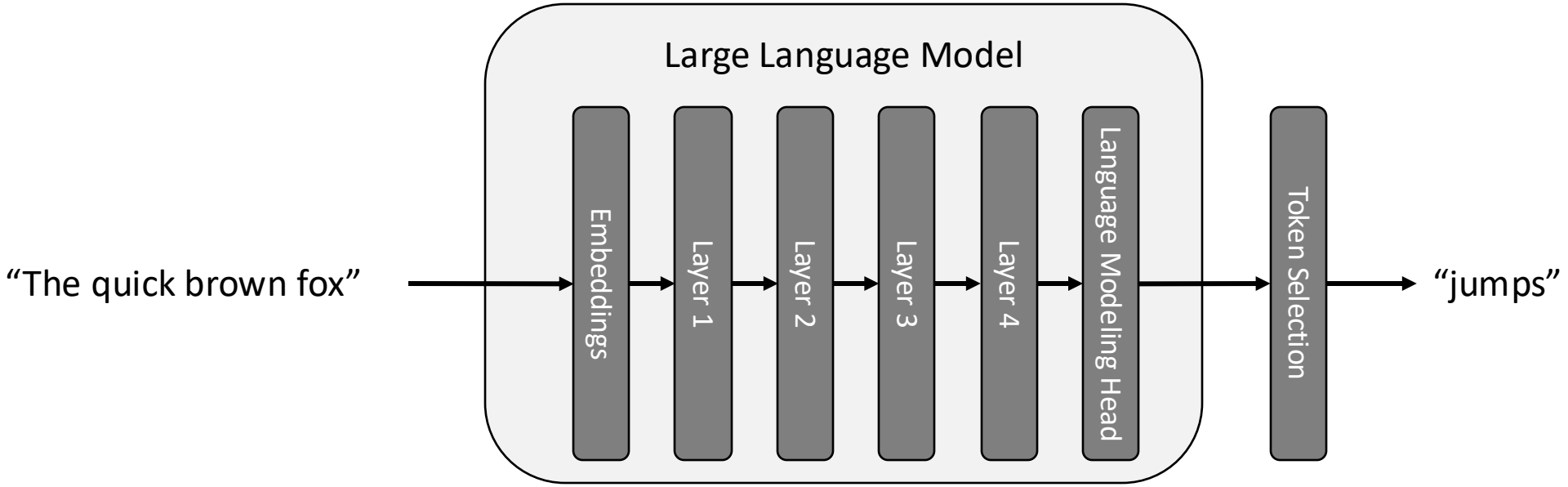
Split computing in LLM serving

- LLM generates token one by one in autoregressive manner



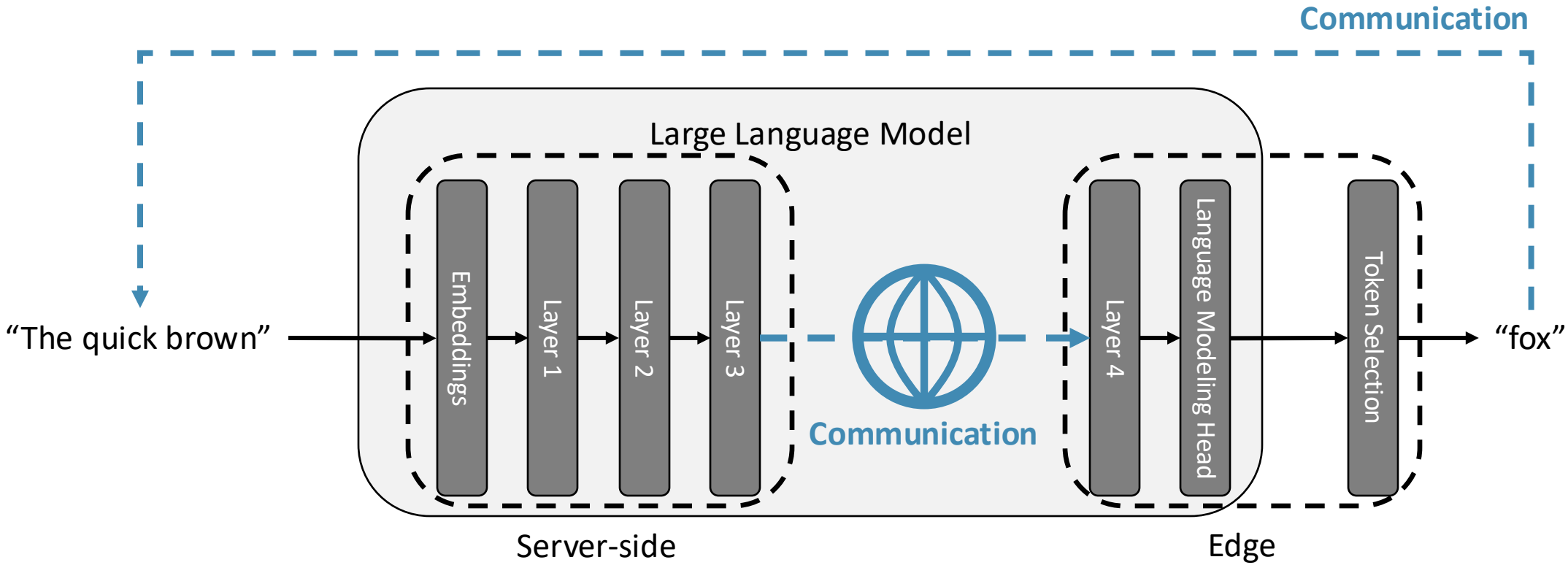
Split computing in LLM serving

- LLM generates token one by one in autoregressive manner



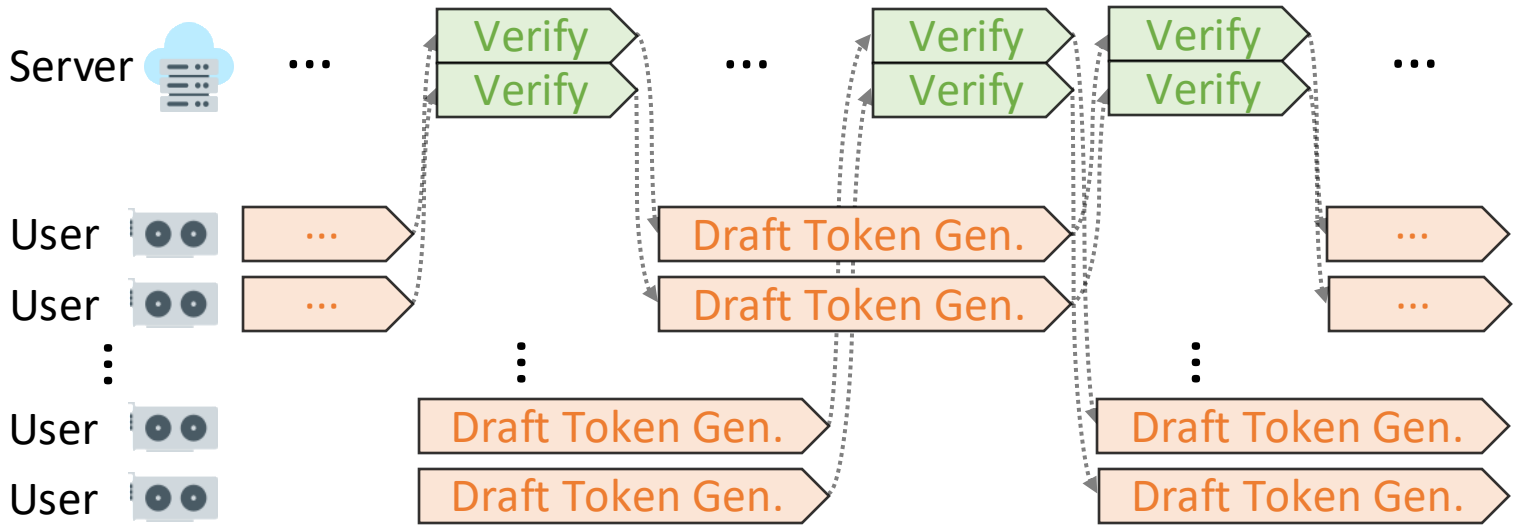
Split computing in LLM serving

- Key Limitation: Excessive Latency



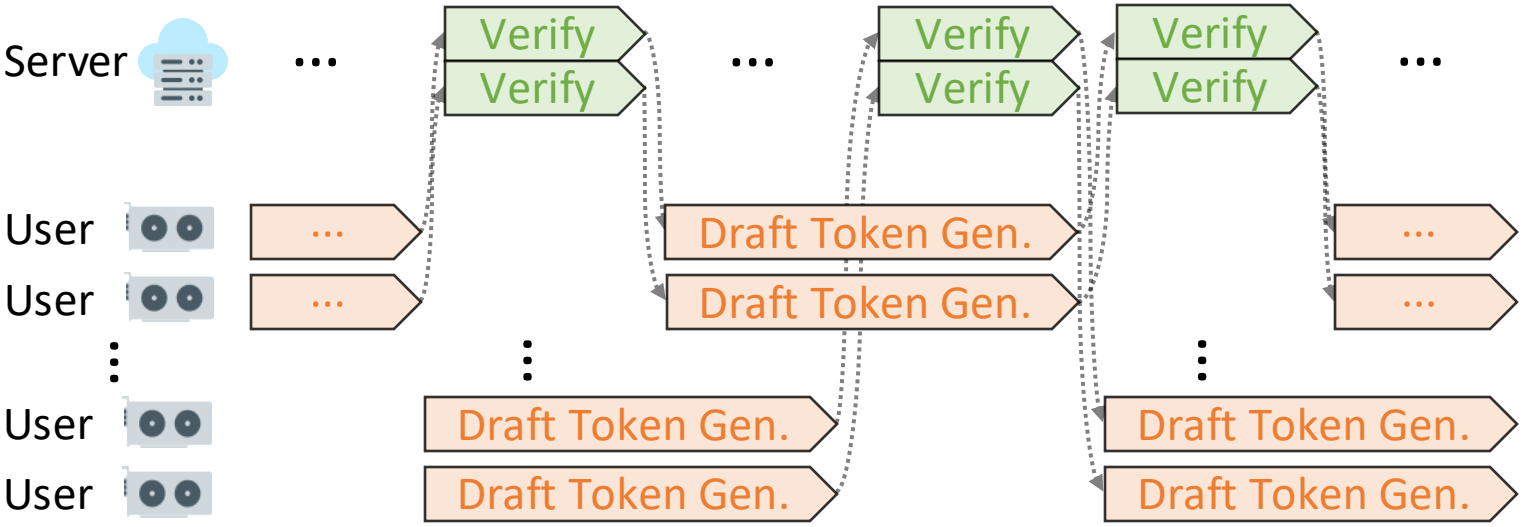
Disaggregated LLM decoding

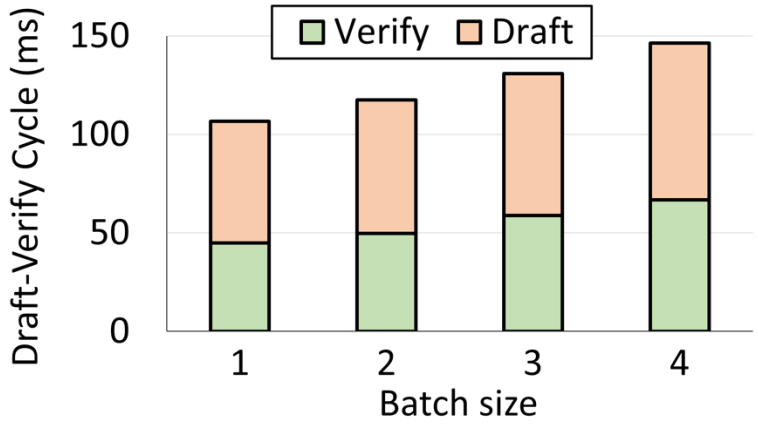
- Disaggregates speculative decoding pipeline into two distinct stages
  - User-side edge drafting & Server verification



Disaggregated LLM decoding

- Disaggregates speculative decoding pipeline into two distinct stages
  - Less frequent communication
  - Reduced network bandwidth requirement
  - Server-side Memory I/O reduction
  - Efficient use of edge GPUs





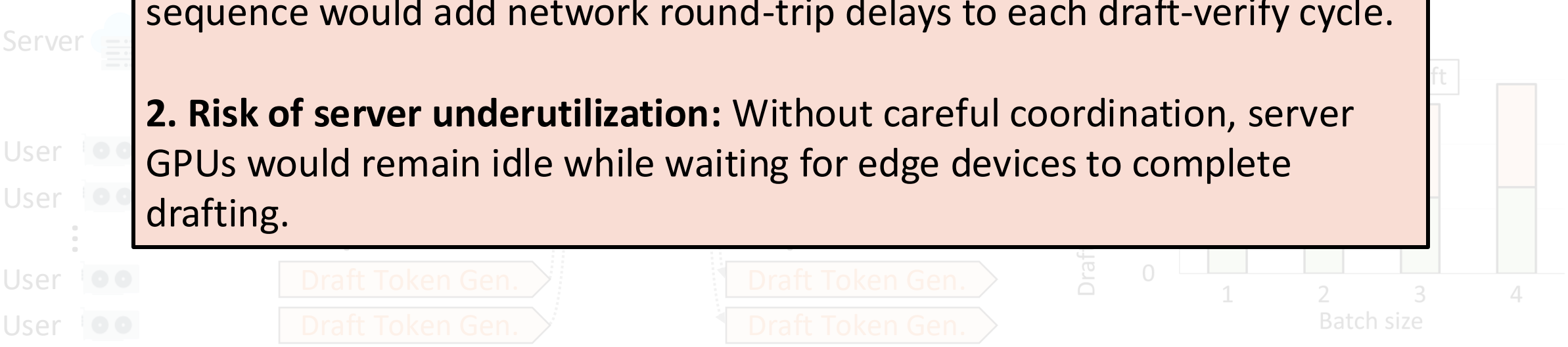
Disaggregated LLM decoding

- Disaggregates speculative decoding pipeline into two distinct stages
  - Less frequent communication
  - Reduced network bandwidth requirement

Key challenges:

1. Potential latency increase: Naïvely disaggregating speculative decoding sequence would add network round-trip delays to each draft-verify cycle.

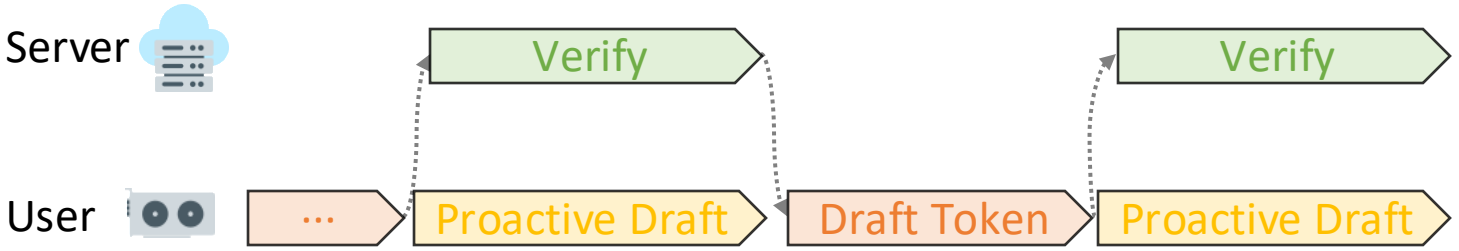
2. Risk of server underutilization: Without careful coordination, server GPUs would remain idle while waiting for edge devices to complete drafting.



Proactive draft generation at the edge

- Instead of waiting for the server's verification results, the edge GPU continues generating draft tokens proactively.

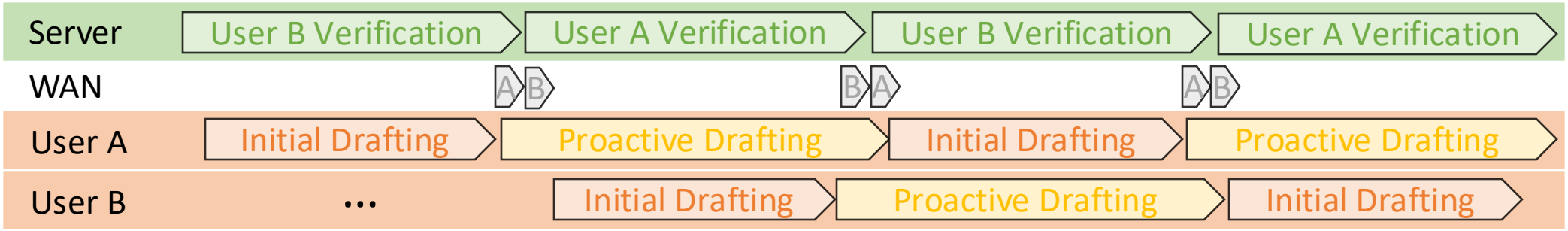
<Conventional speculative decoding>



<SpecEdge with proactive draft generation>

Pipeline-aware verification scheduling

- Interleaving verification requests from multiple users
  - While the edge is busy drafting the next set of tokens, the server would idle if it only awaits verification tasks from the same request, creating “bubbles”.
- Pipeline-aware scheduling
  - SpecEdge calibrates the server verification time with edge drafting time and network latency.
  - We adjust draft depth – the number of forward passes of draft models.



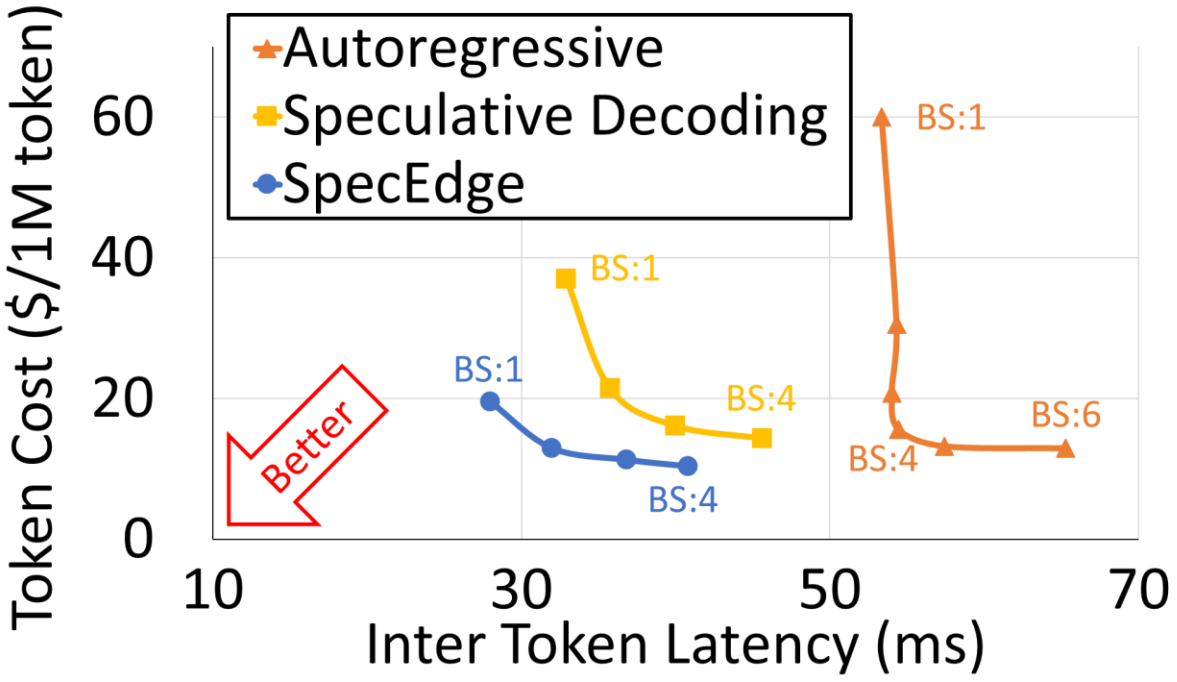
End-to-end performance evaluation

- Gain on average
  - **1.91x** better cost efficiency
  - **2.22x** throughput
- Baseline:
  Tree-based server-only speculative decoding
- SpecEdge Network RTT:
  14.07ms

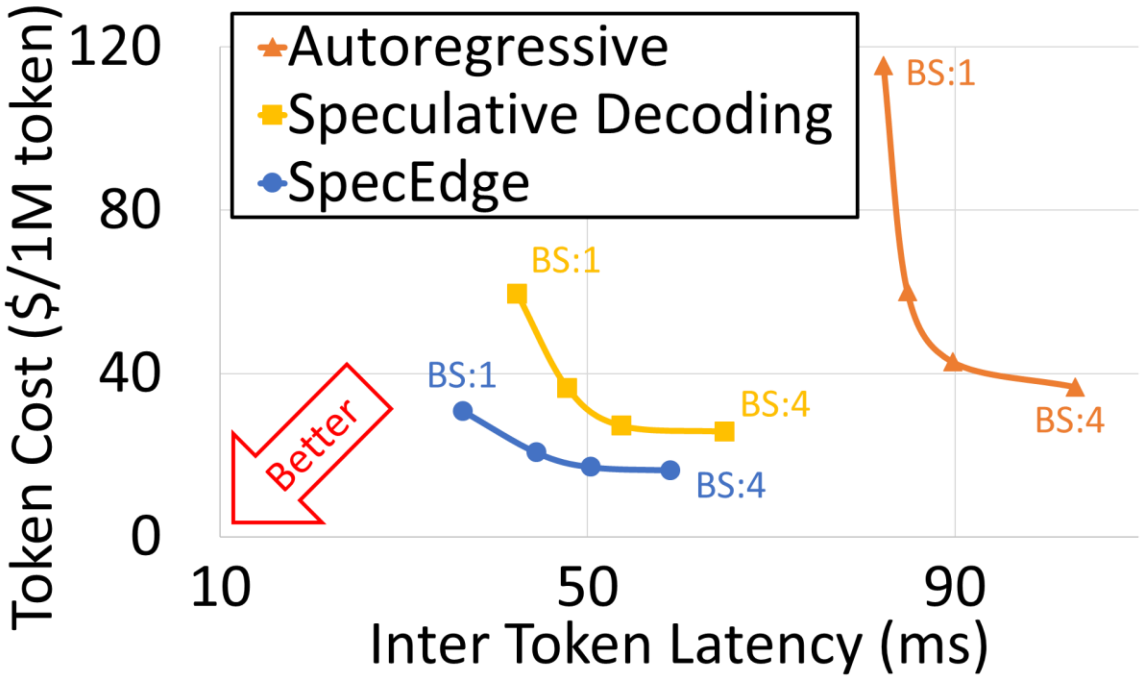
		Gen. tokens per verify		Server Throughput (tok/s)		Cost Efficiency (1k toks/$)	
Target/Draft	Task	Server-only	**SpecEdge**	Server-only	**SpecEdge**	Server-only	**SpecEdge**
Qwen3 14B/1.7B	Multi-turn bench	3.92±1.51	3.98±1.57	31.78	66.54 (**2.09x**)	28.25	50.60 (**1.79x**)
	Translation	3.95±1.47	4.25±1.45	32.24	65.25 (**2.02x**)	28.66	49.47 (**1.73x**)
	Summarization	3.73±1.60	3.95±1.61	29.70	67.53 (**2.27x**)	26.40	51.22 (**1.94x**)
	QA	3.42±1.57	3.59±1.56	27.30	62.04 (**2.27x**)	24.26	47.09 (**1.94x**)
	Math.	4.10±1.48	4.25±1.49	32.84	72.93 (**2.22x**)	29.19	55.28 (**1.89x**)
	RAG	3.73±1.53	3.83±1.56	29.89	64.04 (**2.14x**)	26.57	48.78 (**1.84x**)
Qwen3 14B/0.6B	Multi-turn bench	3.87±1.41	4.41±2.25	33.45	69.58 (**2.08x**)	29.73	52.97 (**1.78x**)
	Translation	3.79±1.48	4.67±2.34	32.88	69.00 (**2.10x**)	29.22	52.23 (**1.79x**)
	Summarization	3.68±1.49	4.21±2.16	31.17	68.60 (**2.20x**)	27.71	52.16 (**1.88x**)
	QA	3.33±1.41	3.79±1.94	28.89	61.90 (**2.14x**)	25.68	46.98 (**1.83x**)
	Math	3.90±1.57	5.27±2.27	33.53	83.88 (**2.50x**)	29.80	63.56 (**2.13x**)
	RAG	3.53±1.52	4.29±2.16	30.07	69.51 (**2.31x**)	26.73	52.76 (**1.97x**)
Qwen3 32B/1.7B	Multi-turn bench	4.22±1.93	4.71±2.66	24.96	56.47 (**2.26x**)	17.80	35.38 (**1.99x**)
	Translation	4.08±1.97	5.24±2.84	24.33	58.79 (**2.42x**)	17.34	36.83 (**2.12x**)
	Summarization	4.19±2.01	4.52±2.68	24.33	54.07 (**2.42x**)	17.65	33.90 (**2.12x**)
	QA	3.62±1.95	3.93±2.49	21.59	46.14 (**2.14x**)	15.39	28.99 (**1.88x**)
	Math.	4.60±1.93	5.40±2.78	27.52	64.01 (**2.33x**)	19.62	40.18 (**2.05x**)
	RAG	3.89±2.05	4.19±2.69	22.67	49.46 (**2.18x**)	16.16	31.05 (**1.92x**)

End-to-end performance evaluation

- Per token cost and inter token latency comparison between server-only baselines and SpecEdge with varying batch size (BS).



14B-1.7B



32B-1.7B

Demo Video

- Scan QR below to checkout SpecEdge in action!!