EasySpec: Layer-Parallel Speculative Decoding for Efficient Multi-GPU Utilization

Yize Wu[1,2], Ke Gao[1], Ling Li[1,2], Yanjun Wu[1*]

[1]Institute of Software, CAS [2]University of Chinese Academy of Sciences

The 39th Annual Conference on Neural Information Processing Systems

Background

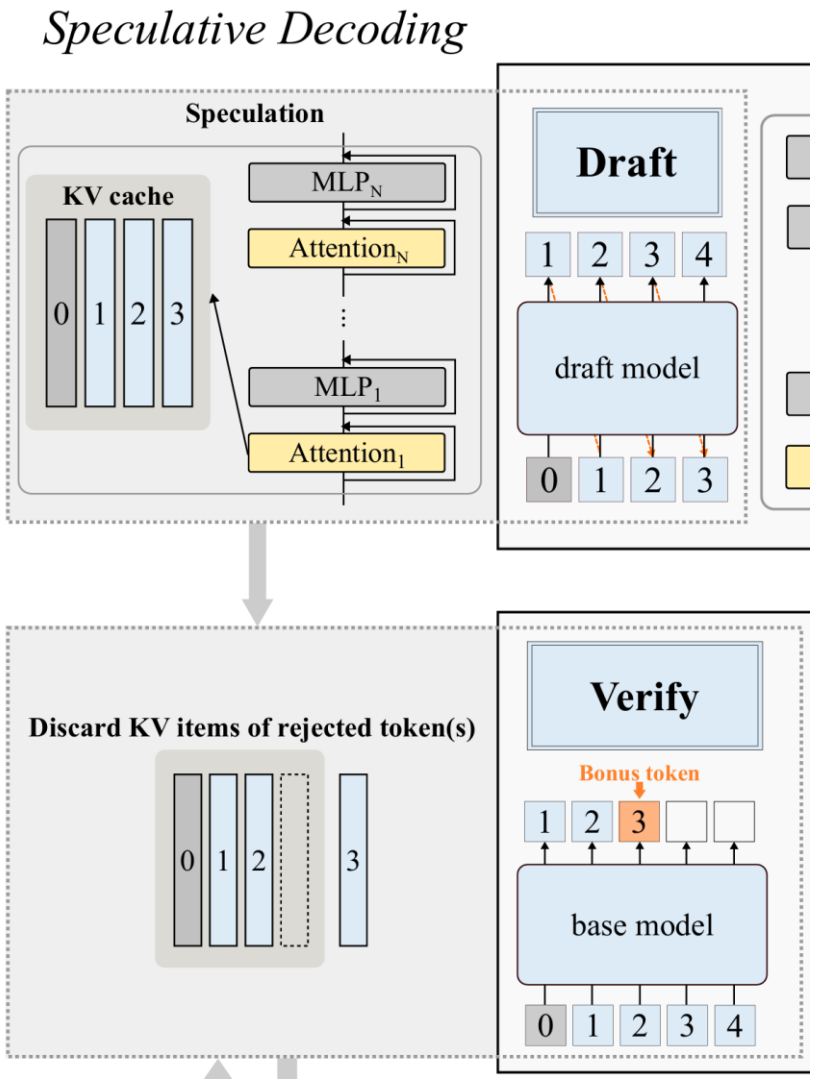
Speculative Decoding

- **Draft**: a smaller model generates a draft token sequence
- **Verify**: the base model conducts token-level parallel and non-autoregressive verification

Tensor Parallelism

- Partitioning workloads across multiple devices

Both are lossless



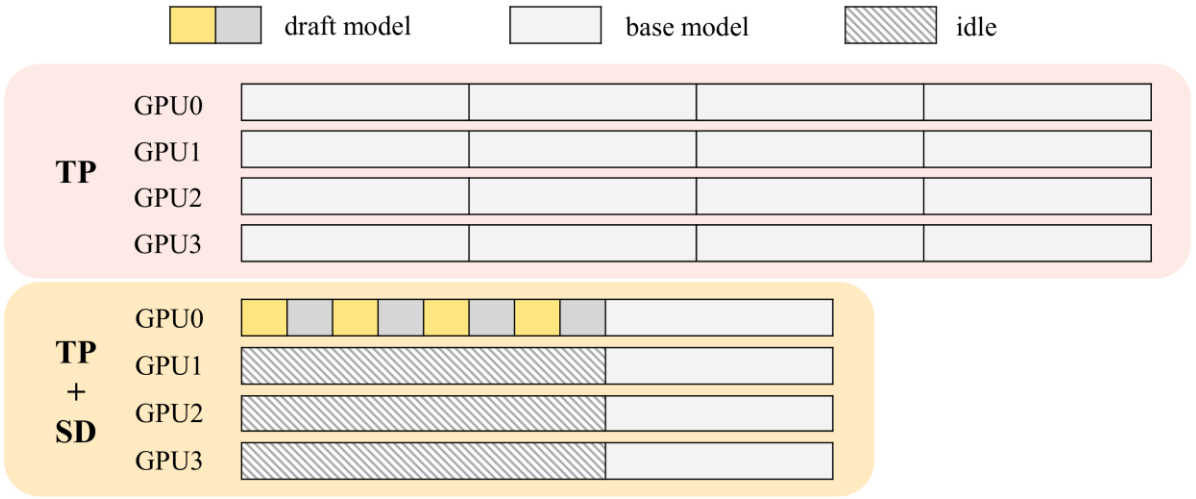
Problem

SD + TP = further lossless acceleration,
BUT…

Inefficient Multi-GPU Utilization

- Optimal TP size:
  - Verification > Drafting
- **GPU idling** during drafting

Model	TP=1	TP=2	TP=4	TP=8
L3-70B	8.39	13.00	13.23	13.23
L3-70B*	OOM	19.5	28.47	28.25
Q2-72B	8.49	13.03	13.01	12.89
Q2-72B*	OOM	18.91	28.63	29.39
L3-8B	36.76	33.90	32.31	32.39
L3-8B*	79.39	68.86	68.35	67.76
Q2-7B	37.16	36.46	37.12	-
Q2-7B*	83.38	73.9	73.81	-



Problem

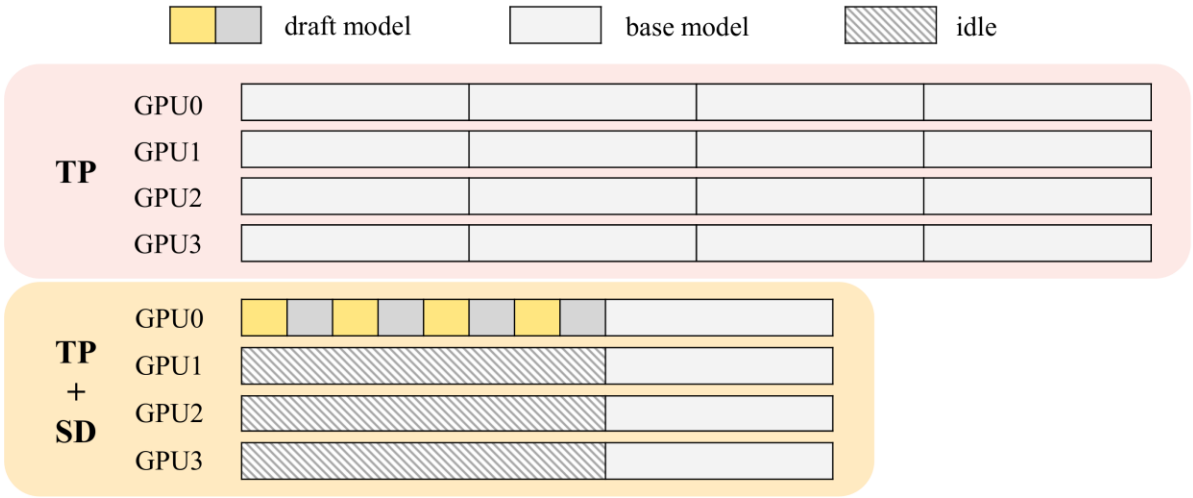
SD + TP = further lossless acceleration,
BUT…

Inefficient Multi-GPU Utilization

- Optimal TP size:
  - Verification > Drafting
- **GPU idling** during drafting

Cause: layer-level data dependency

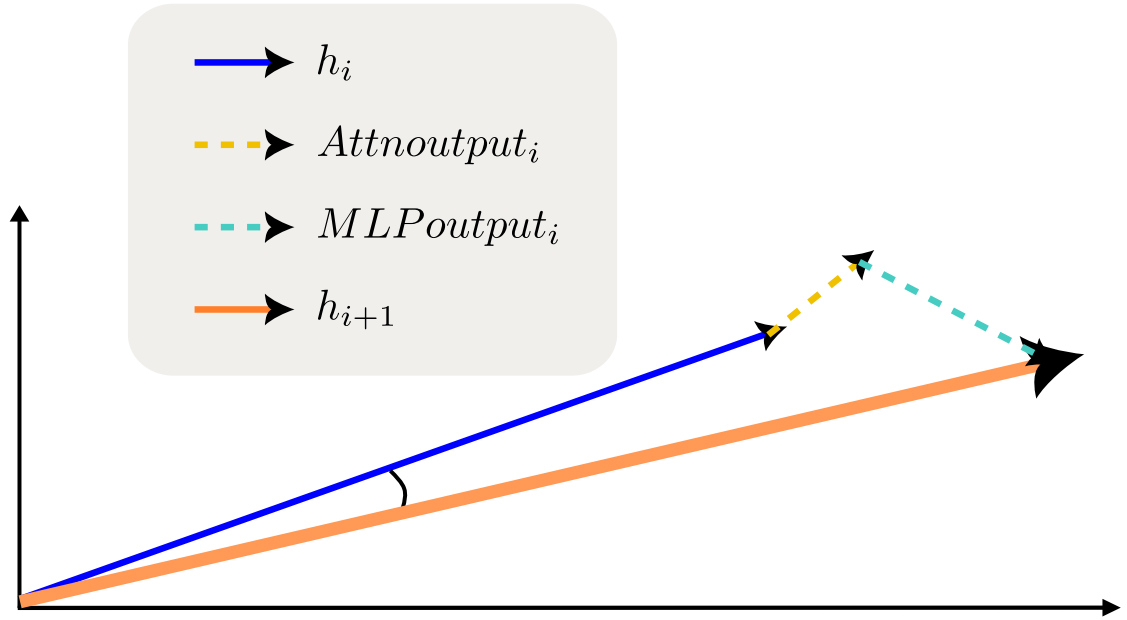
- $h_{i+1} = h_i + Attnoutput_i + MLPoutput_i$
- Layers have to be run **sequentially**, limiting parallelism to **intra-layer** level



Observation

Cross-Layer Hidden-State Approximation

- h_{i+1} can be **well approximated** by h_i , and therefore $h_{i+\mathrm{N}}$ can also be well approximated by h_i



Method

1.Layer-Parallel Fuzzy Speculation

- Directly input h_i to attention layer i+1, …, i+N
  - Eliminate N-layer data dependencies, utilizing idling GPUs by **inter-layer** parallelism
  - Well approximate $Attnoutput_{i+j}$, maintaining **high precision** of the fuzzy results
- Fuzzy speculation **does not** impact output quality
- MLP layers are still run sequentially

Algorithm 2 Layer-Parallel Fuzzy Speculation

Input: hidden state h, N consecutive attention layers $Attn_1,\cdots,Attn_N$ and MLP layers $MLP_1,\cdots,MLP_N$

$h_1 = h$

for $i = 1$ **to** N **do**

  $Attnoutput_i = Attn_i(h_1)$ (parallel)

end for

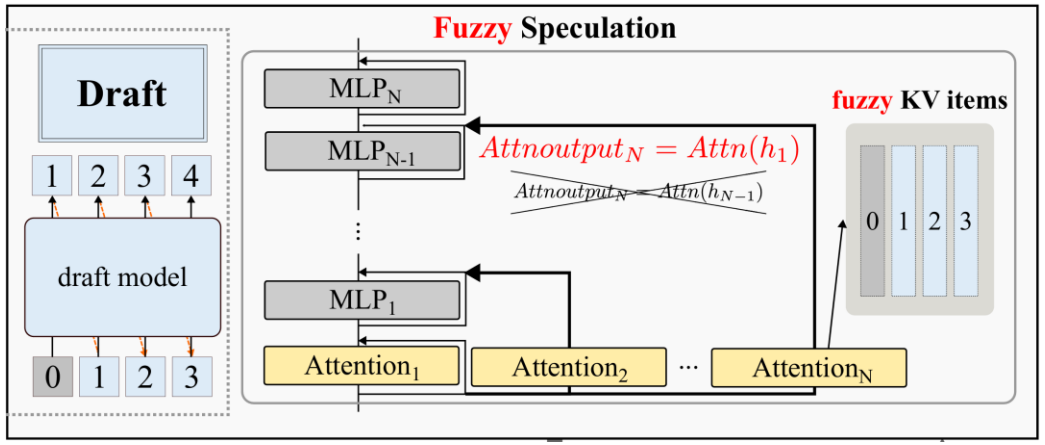
for $i = 1$ **to** N **do**

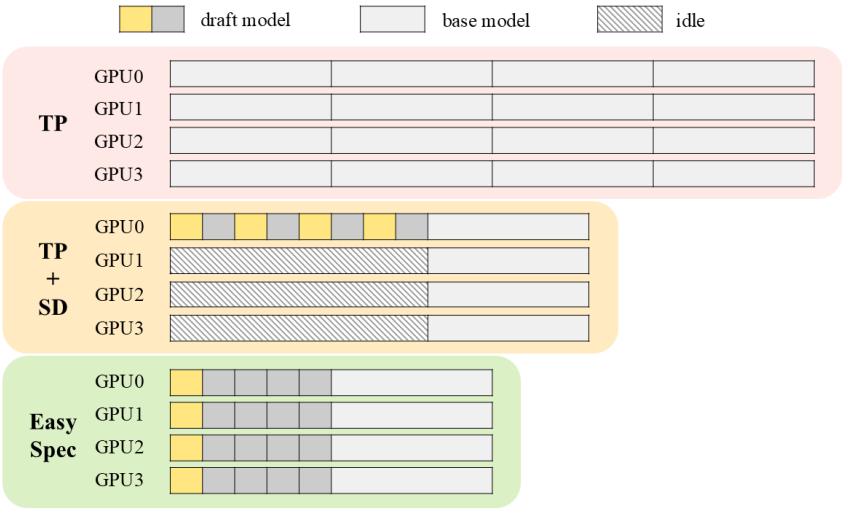
  $h'_i = h_i + Attnoutput_i$

  $MLPoutput_i = MLP_i(h'_i)$

  $h_{i+1} = h'_i + MLPoutput_i$

end for

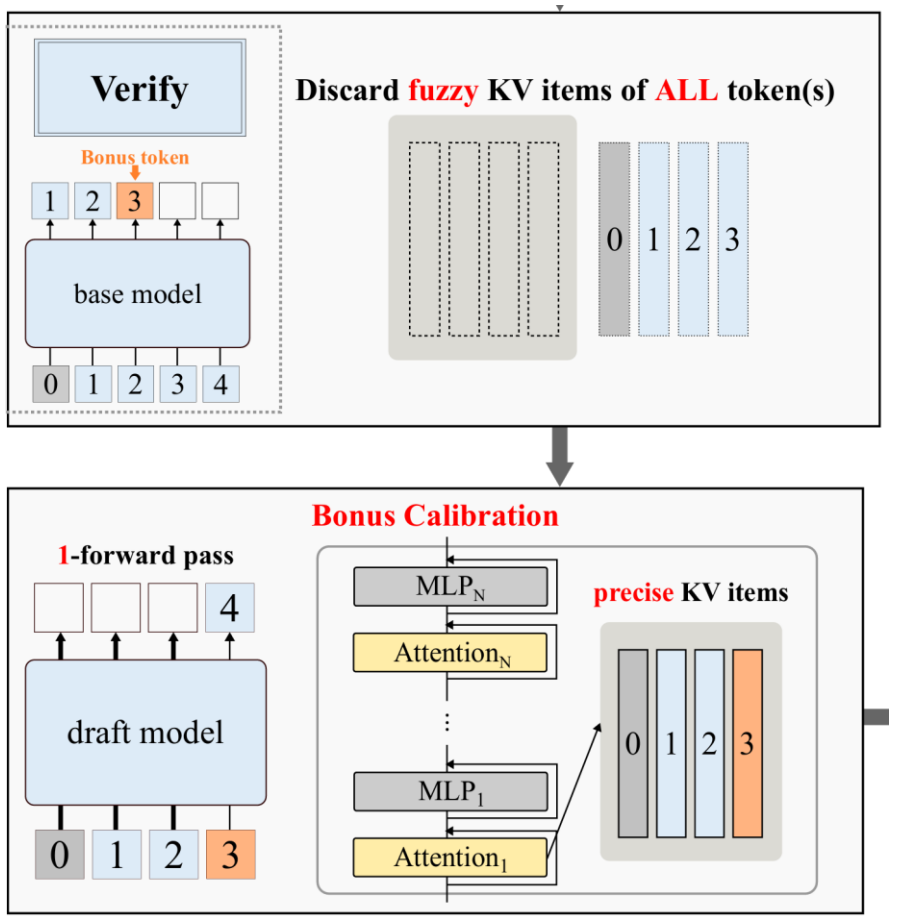




Method

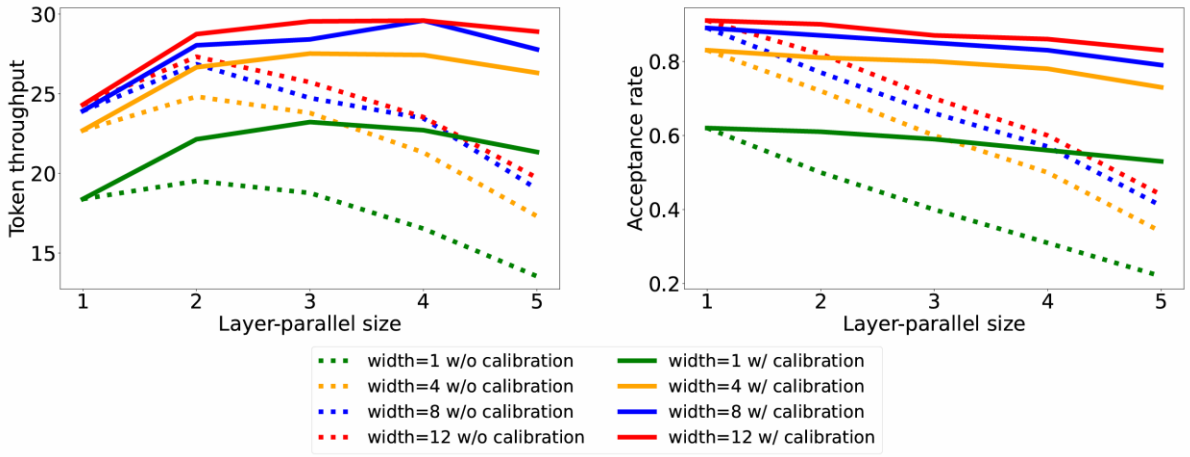
1.Layer-Parallel Fuzzy Speculation
2.Bonus Calibration

- Fuzzy KV items can accumulate **long-term** imprecisions.
- Bonus calibration discards **all** KV items after verification, and **refill** the KV cache with **1-forward layer-sequential** pass of the draft model, with the bonus token



Experiment

Dataset	Method	Llama-3-70B(8B)-Instruct d	v	total	α	Qwen2-72B(7B)-Instruct d	v	total	α
				temperature=0					
MMLU	TP	-	-	1.53x	-	-	-	1.56x	-
	+sd	3.52	2.15	2.05x	0.57	3.32	2.25	2.13x	0.52
	+tree	2.52	1.59	2.82x	0.88	2.38	1.62	2.96x	0.85
	EasySpec	1.70(↑1.48x)	1.73	**3.38x**	0.82	1.65(↑1.44x)	1.70	**3.55x**	0.80
HE	TP	-	-	1.55x	-	-	-	1.57x	-
	+sd	2.93	1.79	2.50x	0.74	2.82	1.83	2.58x	0.69
	+tree	2.53	1.58	2.87x	0.92	2.26	1.51	3.18x	0.95
	EasySpec	1.61(↑1.57x)	1.63	**3.64x**	0.87	1.48(↑1.52x)	1.54	**3.97x**	0.91
MATH	TP	-	-	1.52x	-	-	-	1.54x	-
	+sd	2.96	1.74	2.45x	0.73	2.48	1.65	2.86x	0.78
	+tree	2.50	1.47	2.90x	0.95	2.20	1.45	3.24x	0.96
	EasySpec	1.58(↑1.58x)	1.55	**3.68x**	0.91	1.44(↑1.52x)	1.47	**4.06x**	0.95
IFEval	TP	-	-	1.50x	-	-	-	1.52x	-
	+sd	3.68	2.16	1.93x	0.55	4.24	2.80	1.64x	0.39
	+tree	2.53	1.55	2.76x	0.89	2.80	1.84	2.49x	0.72
	EasySpec	1.68(↑1.51x)	1.65	**3.39x**	0.82	1.87(↑1.50x)	1.94	**3.04x**	0.67
MGSM	TP	-	-	1.54x	-	-	-	1.56x	-
	+sd	2.66	1.61	2.73x	0.80	2.62	1.75	2.73x	0.72
	+tree	2.45	1.48	2.96x	0.96	2.12	1.50	3.29x	0.94
	EasySpec	1.55(↑1.58x)	1.51	**3.80x**	0.93	1.54(↑1.37x)	1.57	**3.83x**	0.88



Models	MMLU		HumenEval		MATH		IFEval		MGSM	
Q2-72B-1.5B	24.02	29.94	26.28	33.78	27.54	36.11	20.65	25.17	25.92	33.26
Q2-72B-0.5B	24.57	29.29	28.80	33.15	29.61	35.82	20.57	24.08	27.48	32.63
L3-70B-3B	25.41	31.32	26.81	32.52	27.49	35.78	25.20	30.29	27.23	35.18
L3-70B-1B	32.10	34.37	34.25	37.25	35.70	40.00	28.60	34.04	35.78	40.25
L3-8B-1B	47.97	56.49	50.99	61.88	54.04	64.97	49.54	56.01	52.68	63.68

The cosine similarities between h_i and h_{i+j} are approaching 1, indicating high approximation precision.

LP size	h	q	k	v	$Attnoutput$
2	0.93	0.98	0.99	0.92	0.93
3	0.89	0.97	0.98	0.86	0.88
4	0.86	0.96	0.97	0.82	0.83

Check our paper and code for more information!

Paper

https://openreview.net/forum?id=RGUcF6pIZN

Code

https://github.com/Yize-Wu/EasySpec

Contact: wuyize2021@iscas.ac.cn