

----- SLO      —●— FasterTransformer      —▲— vLLM      —■— FastServe-FCFS      —◆— FastServe

