Inference System Performance Benchmark (P100)

CARC P100 perf

		7.04 tokens/s	9.01 tokens/s	9.77 tokens/s
input len=256, output len=32		batch size=1	batch size=8	batch size=16
with KV cache	Peak Mem	3230.12 MB	5755.00 MB	8641.54 MB
	Runtime	4.54 s	28.41 s	52.42 s
without KV cache	Peak Mem	3230.12 MB	5755.00 MB	8641.54 MB
	Runtime	11.61 s	29.16 s	52.46 s
		2.76 tokens/s	8.78 tokens/s	9.76 tokens/s

Table 1: Inference System Performance Benchmark

Finetuning System Performance (P100)

		Grad. Accumulation	Grad. Checkpoint	Mixed Precision	LoRA
Memory	parameter	No Change	No Change	Increase	Increase
	activation	No Change	Decrease	Decrease	No Change
	gradient	No Change	No Change	No Change	Decrease
	optimizer state	Decrease	No Change	No Change	Decrease
Computation		0.188 s	0.486	0.413 s	0.223 s

vanilla: 0.335 s

Table 2: Fine-Tuning System Performance Analysis

GC	OFF				ON			
MP	OFF		ON		OFF		ON	
LoRA	OFF	ON	OFF	ON	OFF	ON	OFF	ON
Peak Mem	11954	8396	11954	10395	11953	OOM	11954	OOM
Runtime	0.336	0.223	0.413	0.257	0.486	OOM	0.598	OOM

Table 3: Fine-Tuning System Performance Benchmark