Llama 3.1 8B Latency P95 (s) Configurations 60 Hardware Type Short (8) AMD MI250X P95 AMD MI300X Intel XPU -atency **NVIDIA A100 NVIDIA GH200** NVIDIA H100 Precision bfloat16 fp8 25 50 100 125 75

Batch Size