Qwen MoE: System Scalability Analysis Latency Scaling Characteristics System Throughput vs Batch Size On-Demand On-Demand Oracle -- Oracle **─** Multi-Look Multi-Look Top-K Top-K Intelligent Intelligent 10^{1} --- Ideal Linear Throughput (requests/sec) **Normalized Latency** 10^{0} 10⁰ 12 10 14 16 2 6 10 12 14 16 **Batch Size Batch Size Memory Efficiency Expert Loading Efficiency** 0.85 400 0.80 Hit Rate per 1K Expert Loads **Hit Rate per GB** 0.70 0.70 On-Demand Oracle Oracle Multi-Look Multi-Look Тор-К Тор-К Intelligent Intelligent 0.60 0.55 12 10 2 8 14 16 8 10 12 14 16 6

Batch Size

Batch Size