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 10¹ ---- Ideal Linear Throughput (requests/sec) Normalized Latency 10⁰ 10^{0} 12 10 14 16 2 6 10 12 14 16 **Batch Size Batch Size Memory Efficiency Expert Loading Efficiency** Oracle On-Demand 1.0 Multi-Look Oracle Top-K Multi-Look 400 Intelligent Top-K Intelligent Hit Rate per 1K Expert Loads 0.9 Hit Rate per GB 8.0 8.7 0.7 0.6 12 12 16 2 8 10 14 16 6 8 10 14 **Batch Size Batch Size**