

Cache-aware routing is a larger lever than PD separation on agentic workload

