

Decode-side KV concentration explains the PD-sep memory wall (8x H20, KV pool  $\approx$  28 GB/GPU, 8 concurrent decodes)

