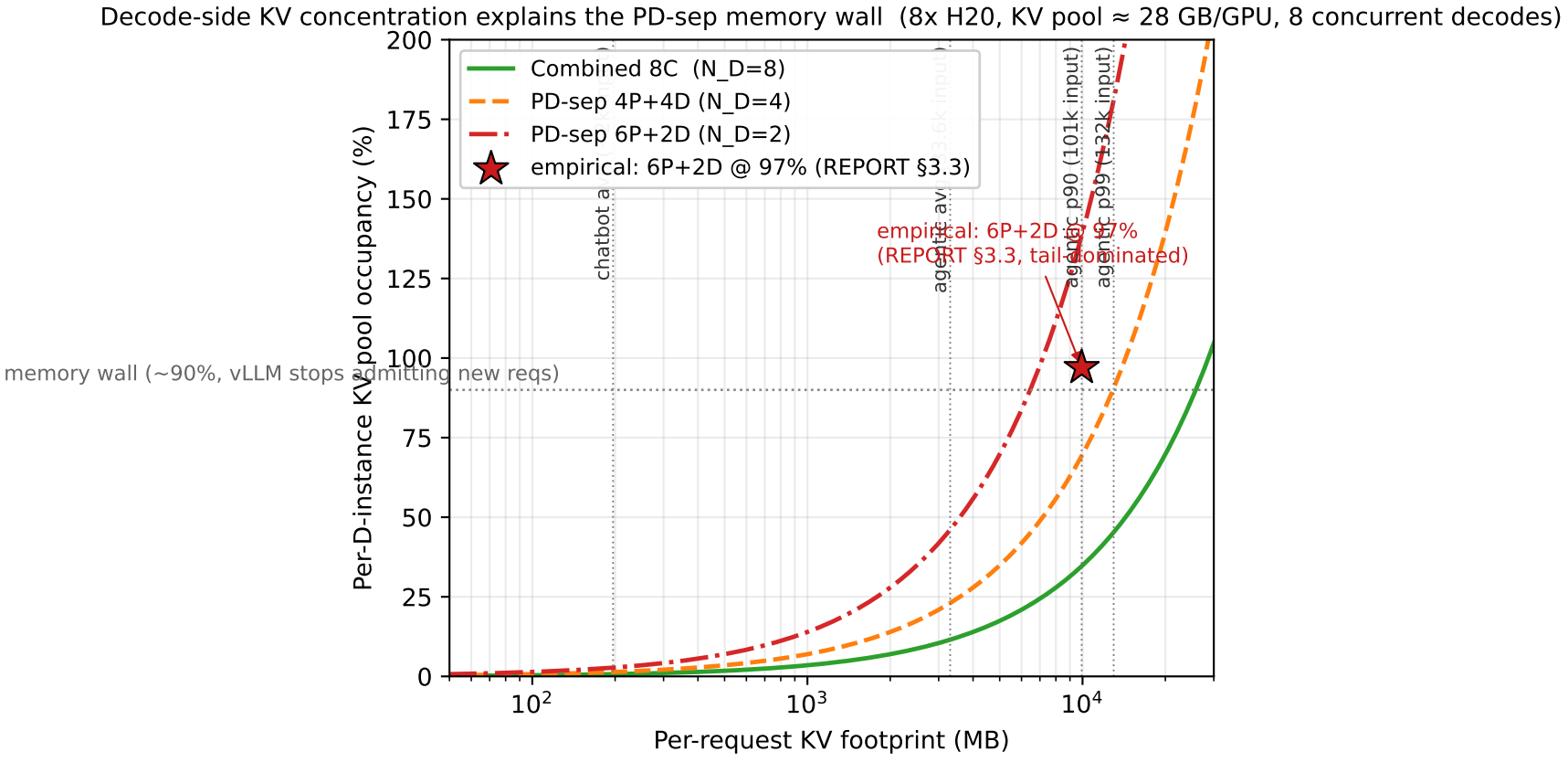
Decode-side KV concentration explains the PD-sep memory wall (8x H20, KV pool ≈ 28 GB/GPU, 8 concurrent decodes)
Combined 8C (N_D=8)
PD-sep 4P+4D (N_D=4)
PD-sep 6P+2D (N_D=2)
empirical: 6P+2D @ 97% (REPORT §3.3)
chatbot a
agentic av
agentic p90 (101k input)
agentic p99 (132k input)
empirical: 6P+2D @ 97%
(REPORT §3.3, tail-dominated)
memory wall (~90%, vLLM stops admitting new reqs)
Per-D-instance KV pool occupancy (%)
Per-request KV footprint (MB)
0
25
50
75
100
125
150
175
200
$10^2$
$10^3$
$10^4$