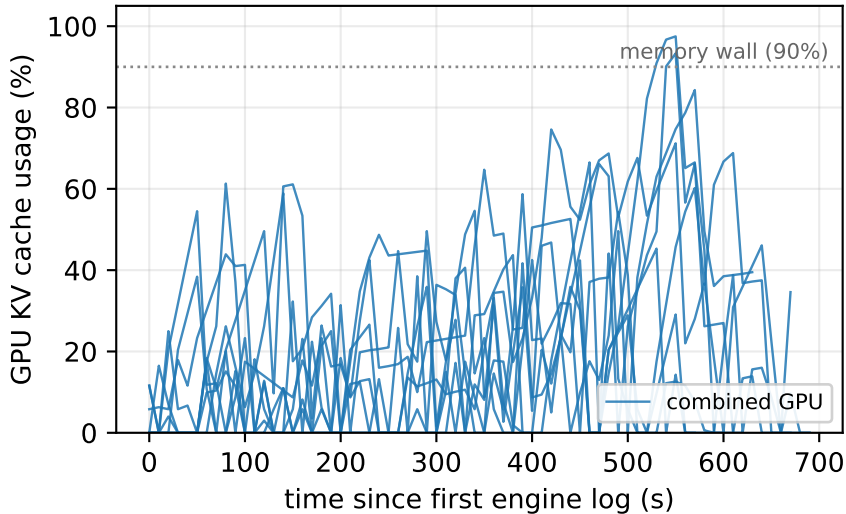
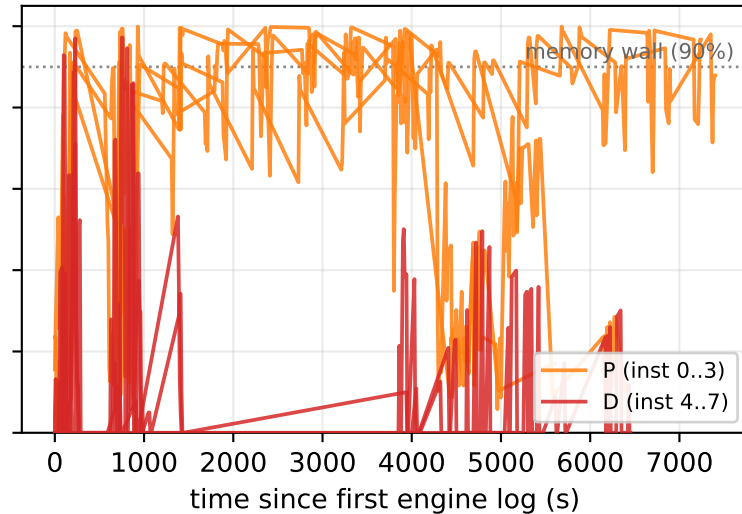


# KV cache utilization: PD-sep concentrates pressure on whichever side has fewer GPUs

combined-ca  
peaks 61..98%



pdsep-4p4d  
peaks 72..100%



pdsep-6p2d  
peaks 99..100%

