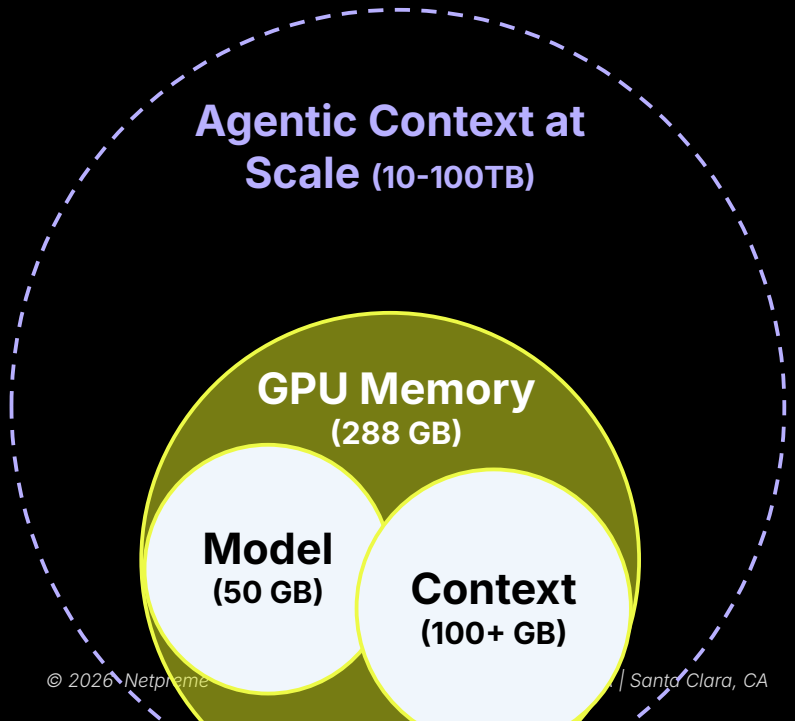
`torch.OutOfMemoryError`

-- finally, an exception you'll never catch again!

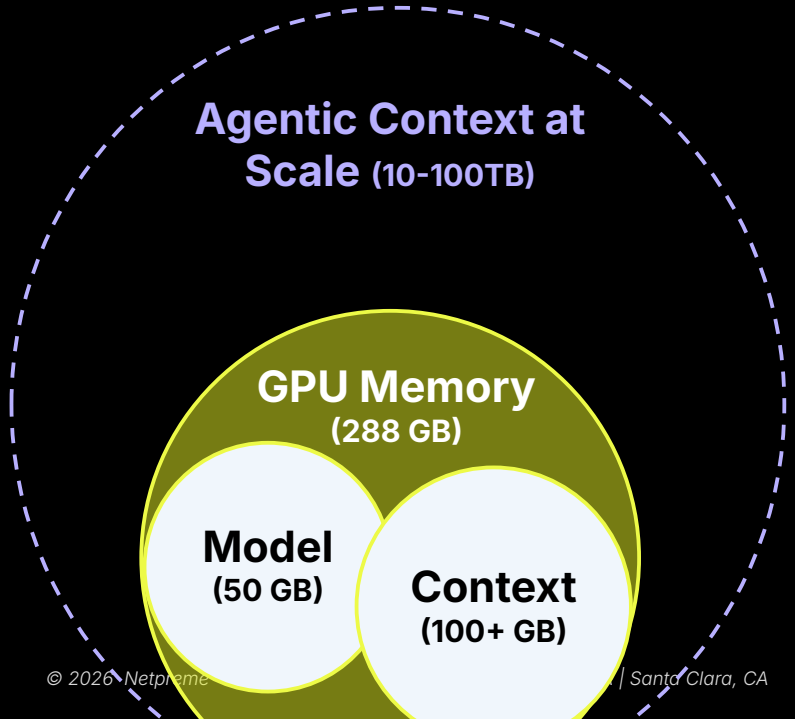
The AI Memory Problem

GPU Memory Capacity is NOT ENOUGH

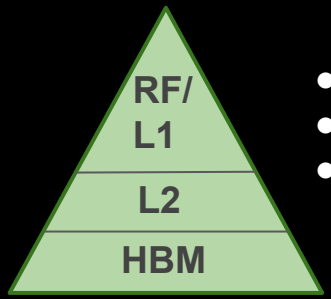


The AI Memory Problem

GPU Memory Capacity is NOT ENOUGH



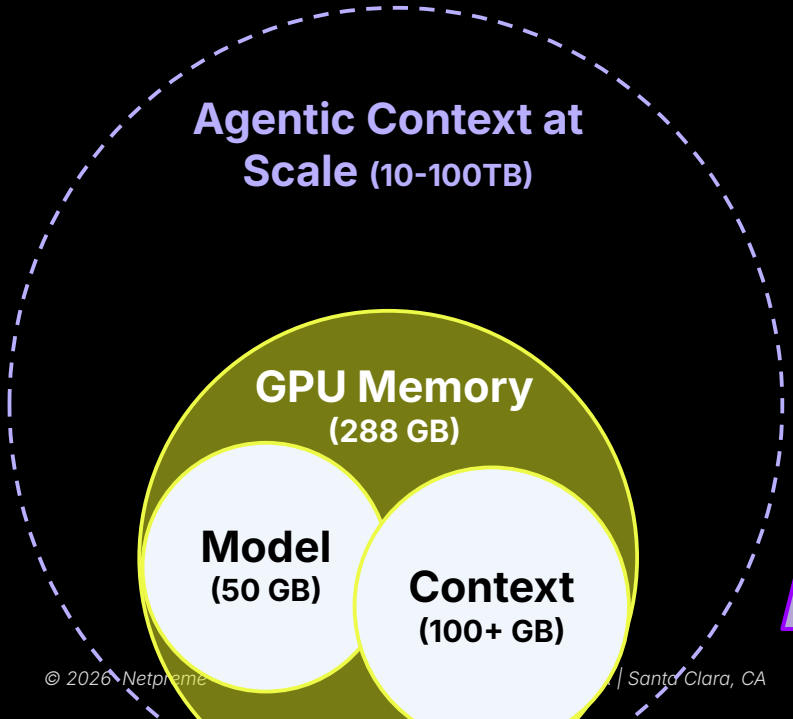
AI Memory Hierarchy Goes Beyond GPUs



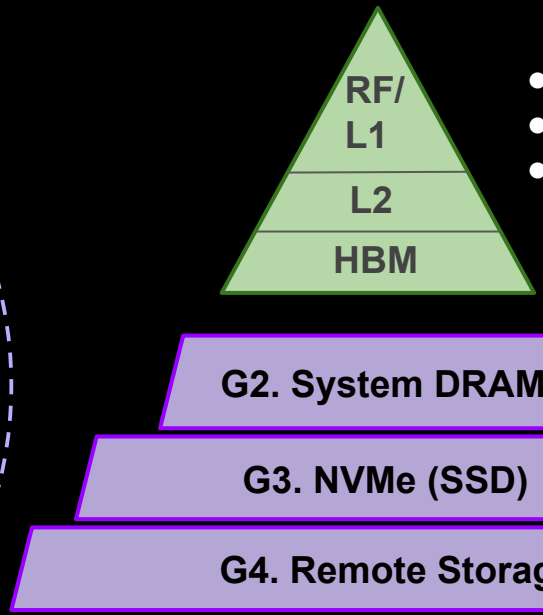
- model
- current context
- weights (training)

The AI Memory Problem

GPU Memory Capacity is NOT ENOUGH



AI Memory Hierarchy Goes Beyond GPUs

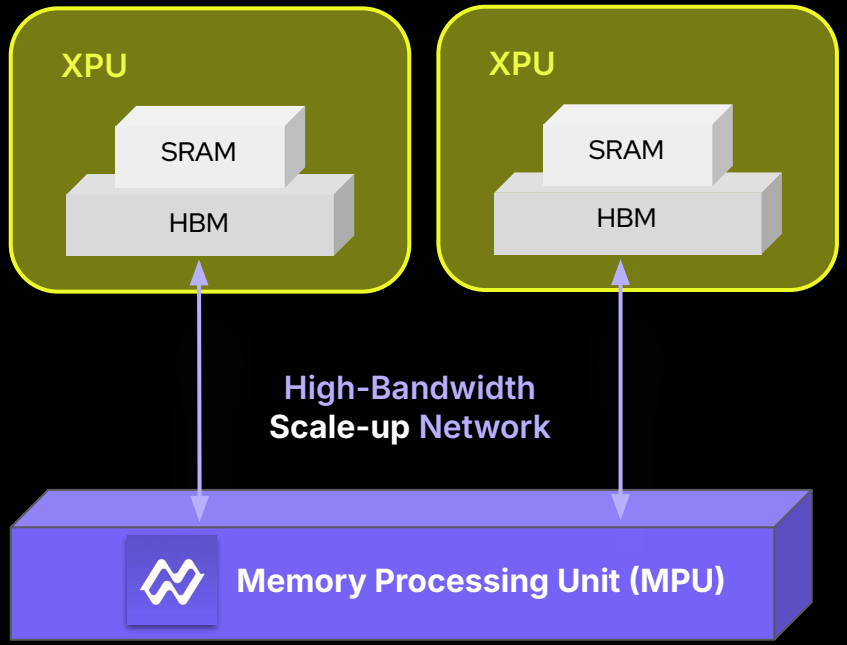


- model
- current context
- weights (training)

- inactive ctx
- sparse attention
- serverless
- activations/ optimizer state

Netpreme™ X-Mem™ Technology

The world's fastest memory-to-compute fabric



US Patent App. 63/793,877 US Patent App. 63/838,102 US Patent App. 63/951,047

Now Integrated with

vLLM SGL

What you get

10x faster than RDMA/System Memory

High-Capacity near-memory computing

50% efficient at scale

Find us at

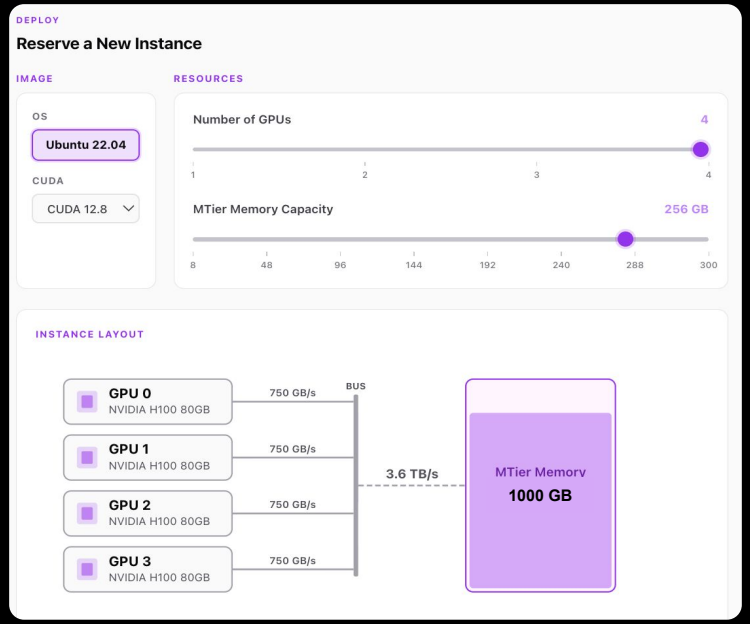
BOOTH #10

Try X-Mem™ Pilot

If you are....

- ML Serving Engine Builder
- AI Infrastructure Engineer

Struggling with long contex?
with complex agentic flows?
with large batches?
Or you tired to see
`torch.OutOfMemoryError`



We're hiring

- ML System
- ML Kernels
- ASIC and more.....