Accelerating AI & ML with CXL-Attached Memory

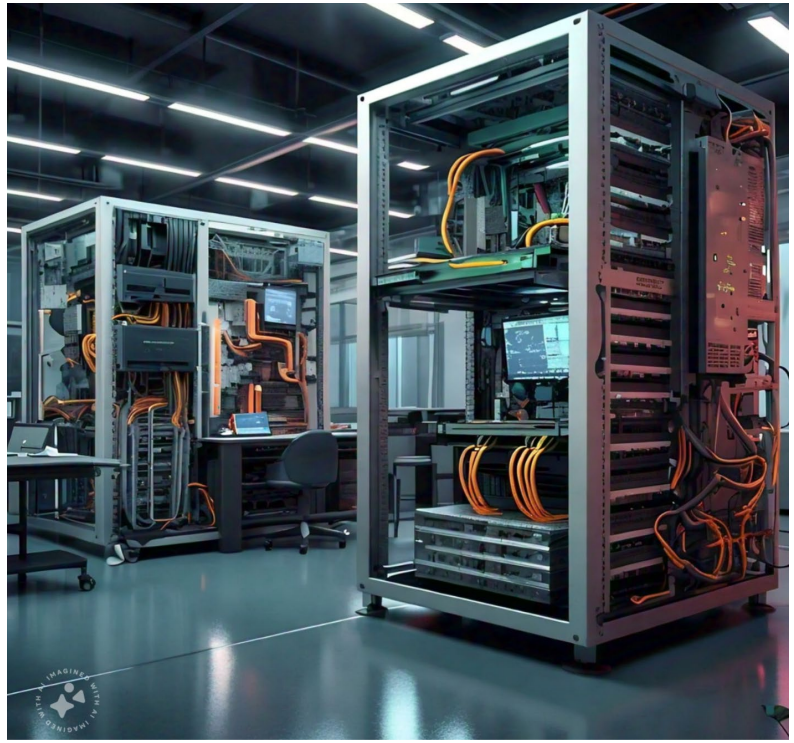
Michael Ocampo, Ecosystem Alliance Manager, Astera Labs

the Future of Memory and Storage

8/6/2024

Agenda

Improving GPU Utilization with CXL



Created by LLaMA 3

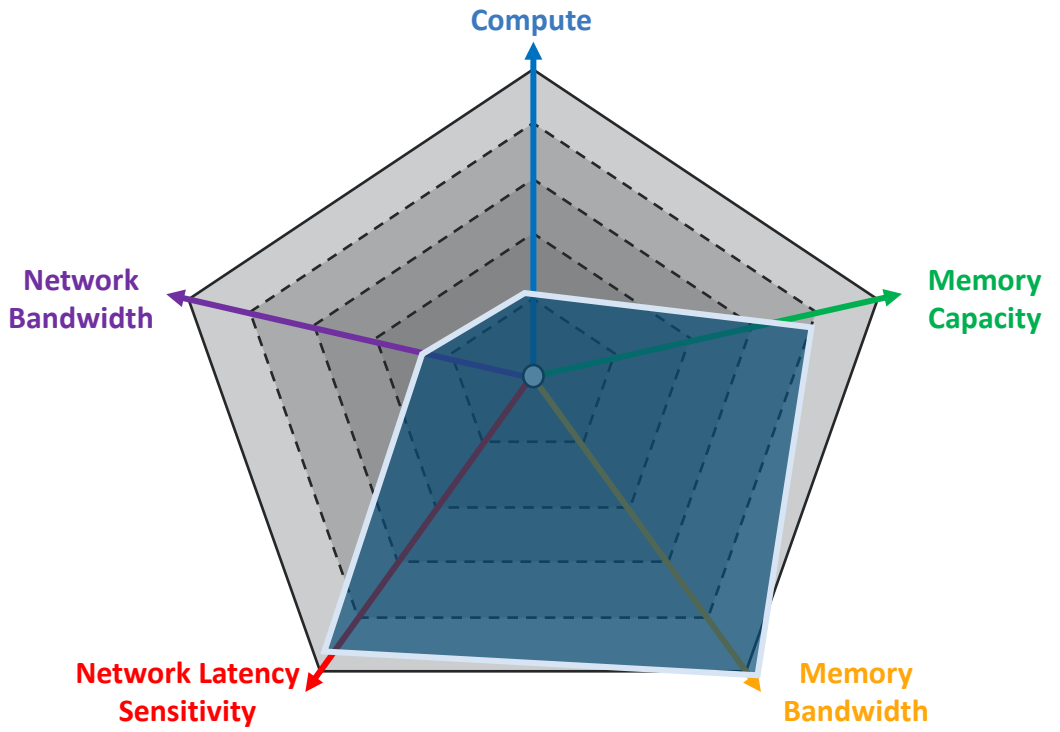
Topics

- AI Inferencing Memory Requirements
- Evolution of Inference Server Architecture
- IO-Efficiency with CXL Increases GPU Utilization
- CXL Optimized AI Inference Performance
- Scaling LLM Instances with CXL
- Key Takeaways

AI Inferencing Memory Requirements

LLM Inferencing

- Focuses on decoding data efficiently
- More memory intensive and network latency sensitive
- GPT-like apps require more RAM for larger context windows
- Examples: GPT, OPT, LLaMA, Mistral, etc



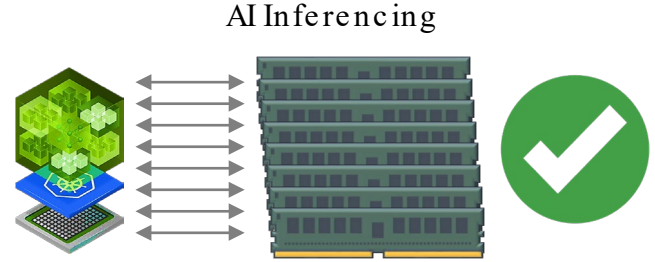
Key components for AI inferencing

- *Context Window* – sets the boundary for the KV Cache
- *KV Cache* – stores keys and values of all previous tokens

KV cache consumes a significant amount of memory

- Attention models may consume on the order of **~1MB/token[2]**
- KV Cache size depends on precision, ie: FP32, BF16, FP16, INT8, FP8, etc

KV Cache stores Keys & Values of all Previous Tokens[3]

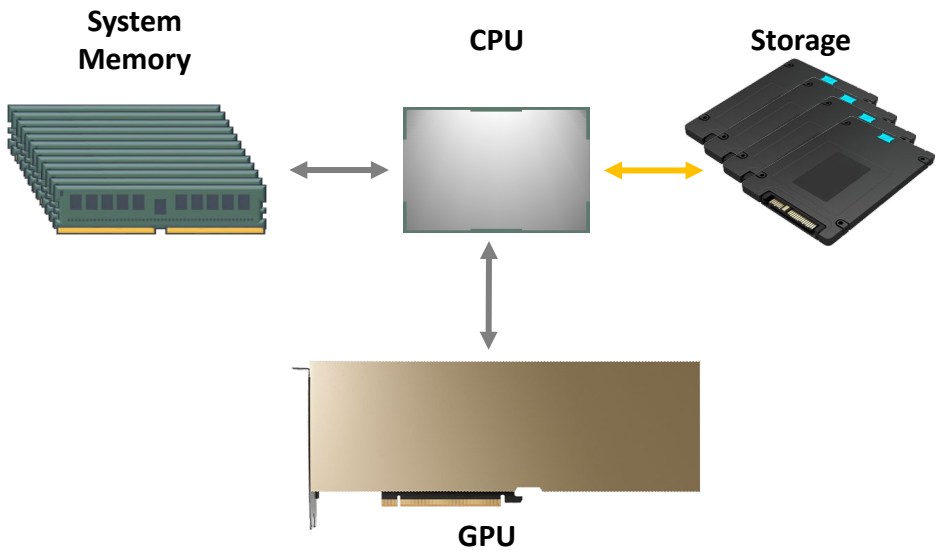


An input the size of 10 novels would require a ~1000K context window, requiring about 1TB of memory

[1] Presented at OCP Global Summit 2023 by Dan Rabinovitsj (Meta)
[2] Pierre Lienhart, "LLM Inference Series: 4. KV caching, a deeper look", www.medium.com, 7/16/2024

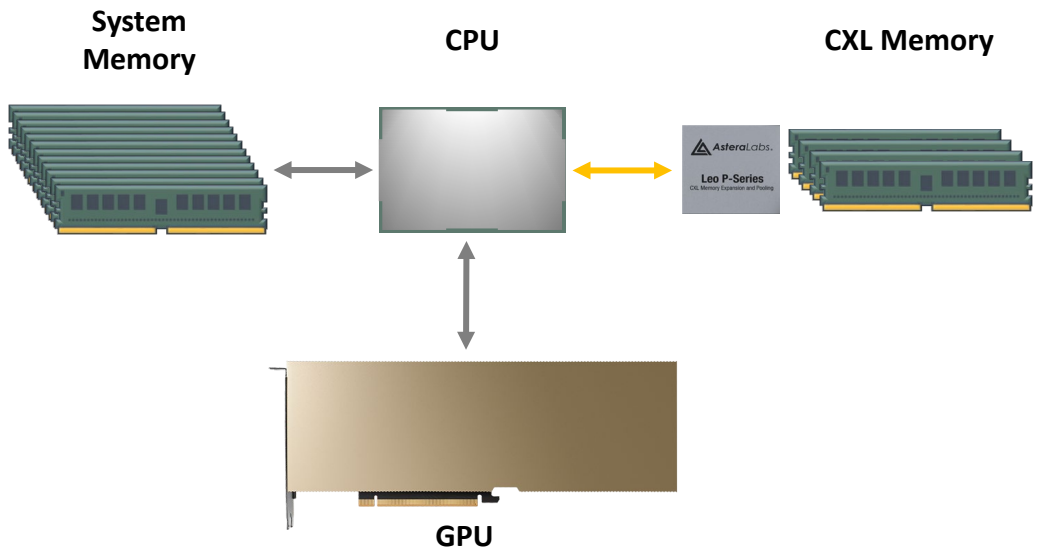
Evolution of Inference Server Architecture

Server with NVMe SSDs



LLM and KV Cache stored in:
System Memory (High performance)
Storage (Low performance cache)

Server with CXL Memory



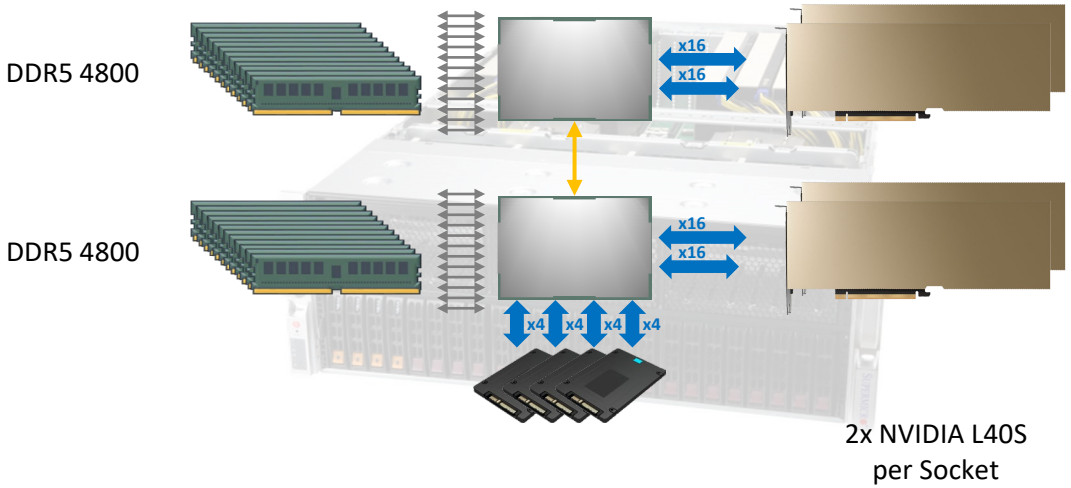
LLM and KV Cache stored in:
System Memory (High performance)
CXL Memory (High performance)

CXL Can Help Alleviate Memory Bottlenecks

CXL Optimized AI Inference Server Performance Results

Four GPUs without CXL (24 DIMMs)

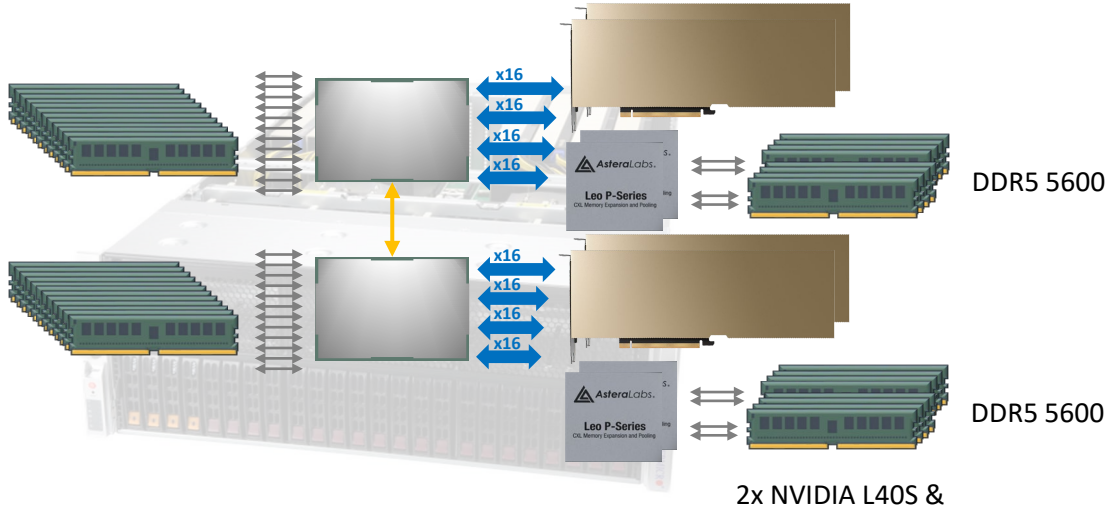
Supermicro AS -4125GS-TNRT



- **Slower** time to insights with NVMe cache
- **Higher** CPU utilization
- **Limited** concurrent LLM instances per server

Four GPUs with Four Leo CXL Controllers (40 DIMMs)

Supermicro AS -4125GS-TNRT with Astera Labs Leo CXL Memory



- **40%** faster time to insights
- **40%** lower CPU utilization
- **2x** concurrent LLM instances per server

Test Results Show Significant Benefits with CXL

Note: Actual results based on single-socket configuration with 1x Genoa AMD EPYC 9534, 12x DDR5-4800, 2x NVIDIA L40S, 2x Aurora A1000 AICs, and 4x DDR5-5600 CXL-attached memory

IO-Efficiency with CXL Increases GPU Utilization

GPU Utilization of OPT-66B

Hardware Configuration

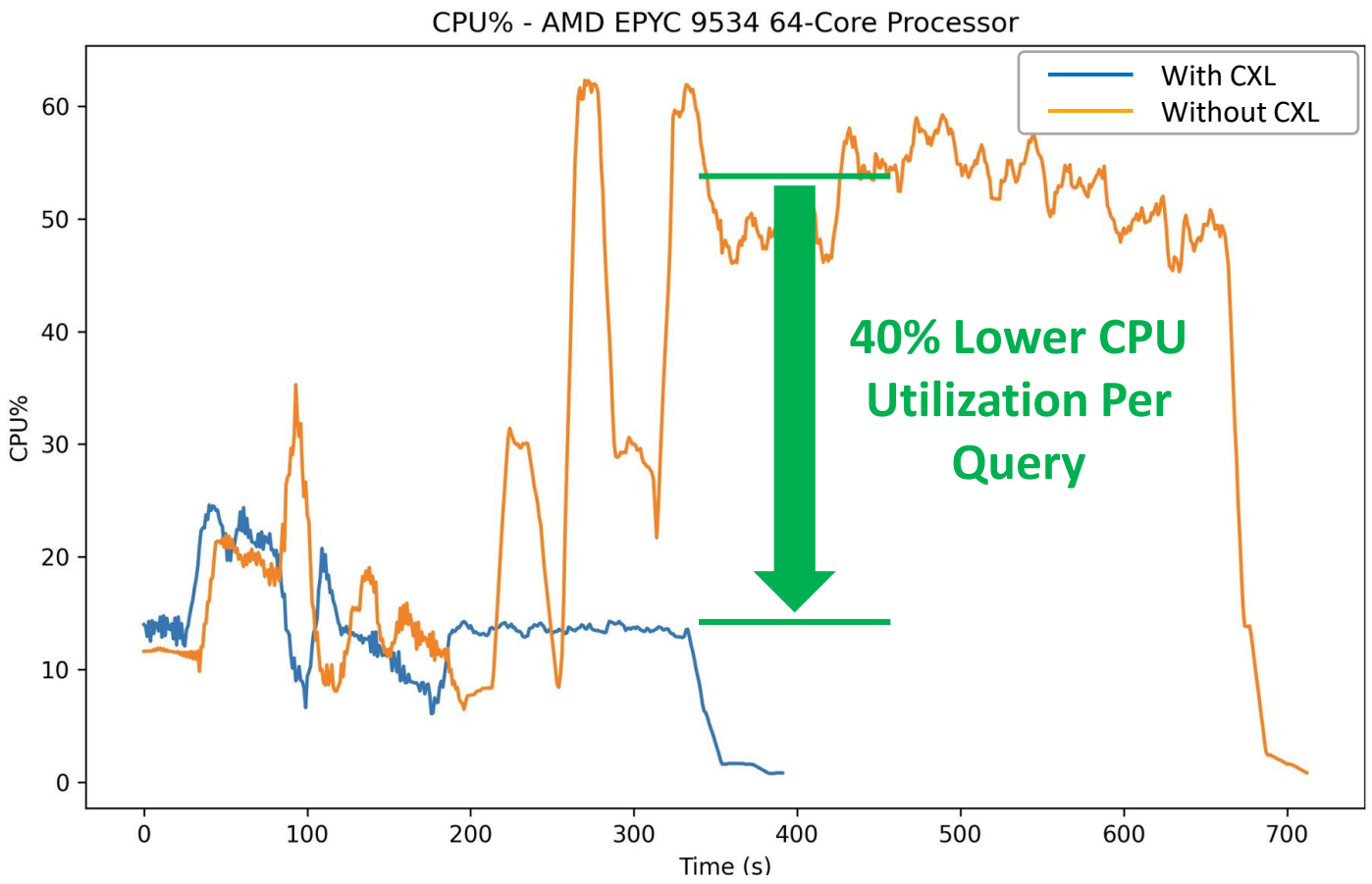
- System Configuration without CXL
  - System: Supermicro 4U GPU System
  - CPU: 5th Gen AMD EPYC Scalable Processor (9534)
  - GPU: 2x NVIDIA L40S (96GB GDDR6)
  - Native Memory: 12x 64GB **DDR5-4800 (768GB)**
  - Storage: 2x 2TB PCIe 5.0 SSDs (RAID0)
- System Configuration with CXL
  - System: Supermicro 4U GPU System
  - CPU: 5th Gen AMD EPYC Scalable Processor (9534)
  - GPU: 2x NVIDIA L40S (96GB GDDR6)
  - CXL: 2x Aurora A1000 add-in cards
  - Native Memory: 12x 64GB **DDR5-4800 (768GB)**
  - CXL-Memory: 4x 64GB **DDR5-5600 (256TB)**

Software Configuration

- OS: Ubuntu 22.04, Kernel 5.15.0
- LLM Engine: FlexGen
- Model Size: 122.375GB (OPT-66B)
- Run Parameters:
  - PROMPT LENGTH = 512
  - GEN_LENGTH = 8
  - GPU_BATCH_SIZE = 24
  - NUM_BATCHES = 12

CXL Improves Overall System Performance

CPU Utilization of OPT-66B



Hardware Configuration

- System Configuration without CXL
  - System: Supermicro 4U GPU System
  - CPU: 5th Gen AMD EPYC Scalable Processor (9534)
  - GPU: 2x NVIDIA L40S (96GB GDDR6)
  - Native Memory: 12x 64GB **DDR5-4800 (768GB)**
  - Storage: 2x 2TB PCIe 5.0 SSDs (RAID0)
- System Configuration with CXL
  - System: Supermicro 4U GPU System
  - CPU: 5th Gen AMD EPYC Scalable Processor (9534)
  - GPU: 2x NVIDIA L40S (96GB GDDR6)
  - CXL: 2x Aurora A1000 add-in cards
  - Native Memory: 12x 64GB **DDR5-4800 (768GB)**
  - CXL-Memory: 4x 64GB **DDR5-5600 (256TB)**

Software Configuration

- OS: Ubuntu 22.04, Kernel 5.15.0
- LLM Engine: FlexGen
- Model Size: 122.375GB (OPT-66B)
- Run Parameters:
  - PROMPT LENGTH = 512
  - GEN_LENGTH = 8
  - GPU_BATCH_SIZE = 24
  - NUM_BATCHES = 12

Scaling LLM Instances with CXL

System Workload for 1 Instance (~1TB)

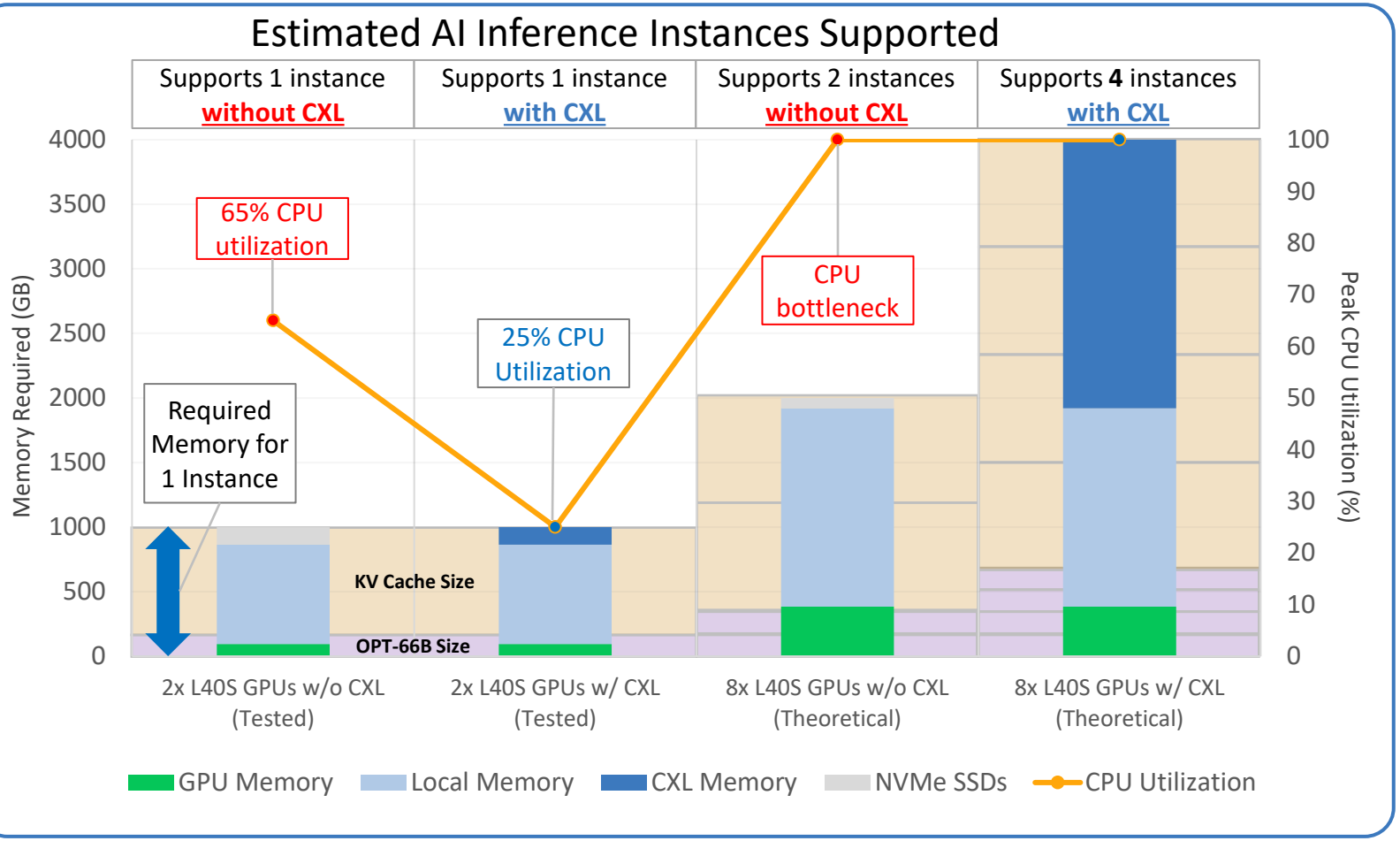
- Without CXL, OPT-66B & KV Cache cannot fit into RAM (864GB, Local RAM + L40S GPUs)
- CPU utilization **without CXL:** 65%
- CPU utilization **with CXL:** 25%

System Workload for 2-4 Instances

- CPU utilization **without CXL** **(~2TB)**
  - 2 instances: 100% **(CPU bottleneck)**
- CPU utilization **with CXL** **(~4TB)**
  - 4 instances: 100% **(Theoretical)**

CXL Optimized AI System

- 8x L40S GPUs (384GB)
- 24x 128GB DDR5-5600[1] (3TB)
- 4x Leo CXL Memory Expansion Cards **(2TB)**



Fixed FlexGen Parameters | Batch Size = 12, Prompt Length = 512, Generation Length = 8, GPU_Batch_Size = 24

Deploying More LLM Instances per Server with CXL Improves TCO

[1]128GB DDR5-5600 memory modules is the highest capacity supported on Supermicro AS -4125GS-TNRT

Accelerating AI & ML with CXL-Attached Memory

40% faster time to insights

40% lower CPU utilization

2x concurrent LLM instances per server

Where To Learn More

Featured FMS Demos & Resources

- AI Inference Demo: Supermicro Booth, #846
  - 4U Supermicro AMD System, AS-4125GS-TNRT
  - Astera Labs' Aurora A1000 PCIe Add-in Cards
  - NVIDIA L40S GPU PCIe Add-in Cards
- CXL Benchmark: MemVerge, #1251
- CXL Interop Reports: All major DDR vendors
  - QR code for Leo Interop Reports

Supermicro Booth, #846

Questions & Answers

AsteraLabs®

FMS
the Future of Memory and Storage

Thank You

Check us out on

www.asteralabs.com