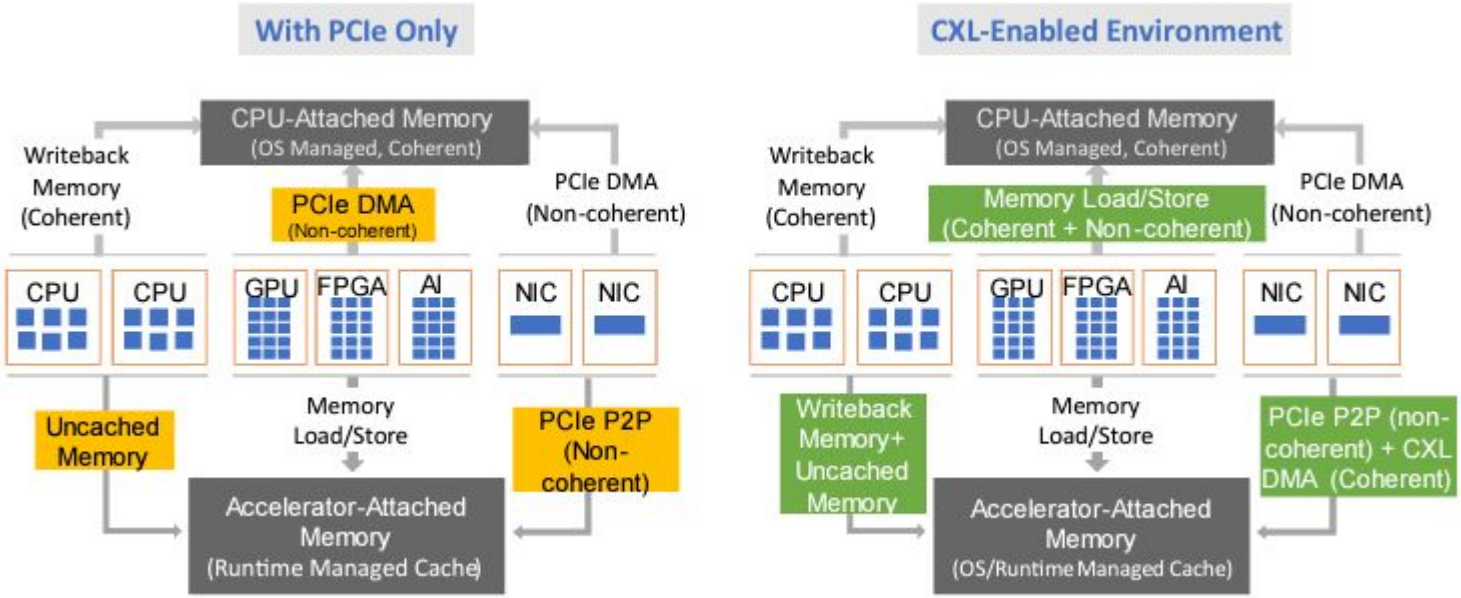
CXL and why networking people should care

Shrijeet Mukherjee, CTO Enfabrica.
shrijeet@enfabrica.net

Netdev 0x17, 2023
Vancouver

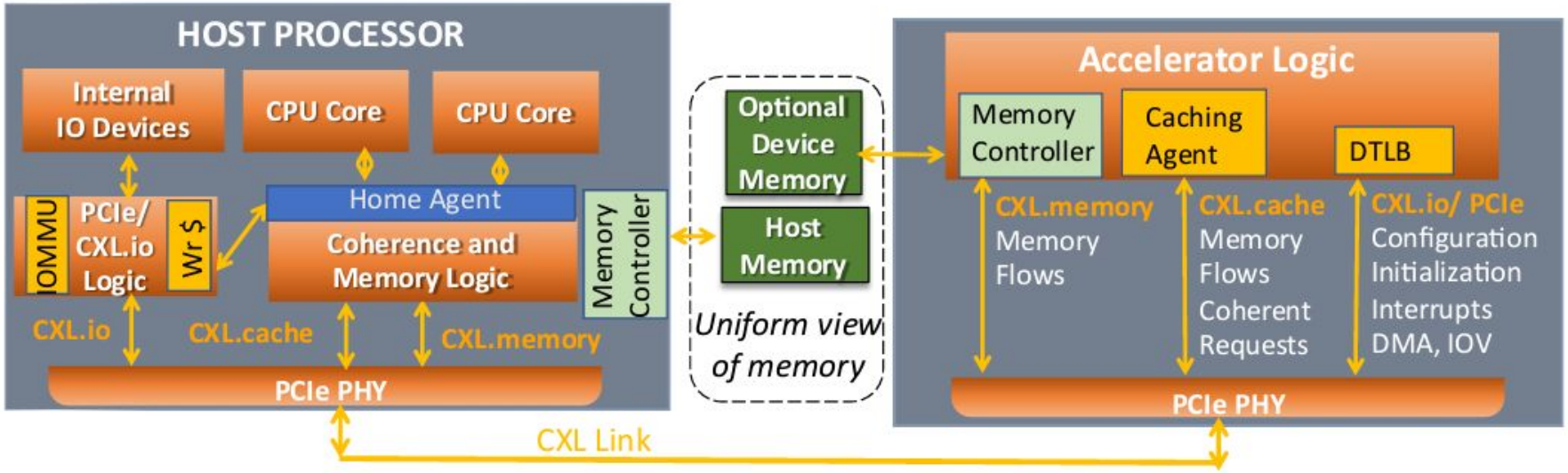
enfabrica

:: CXL : A new beginning

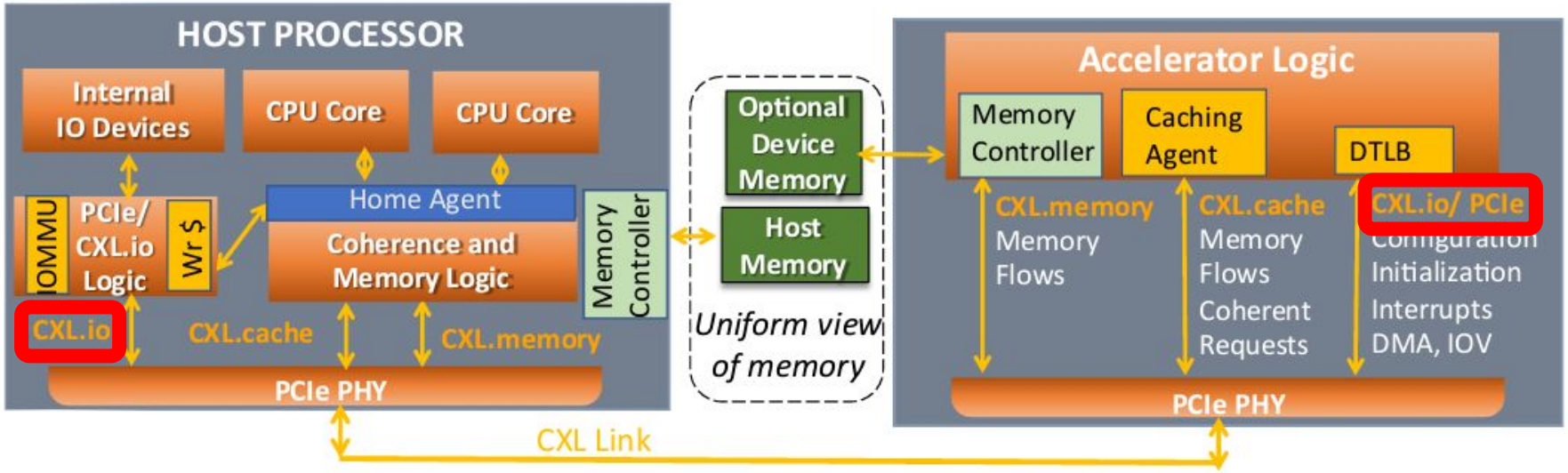


Ref : https://arxiv.org/pdf/2306.11227.pdf

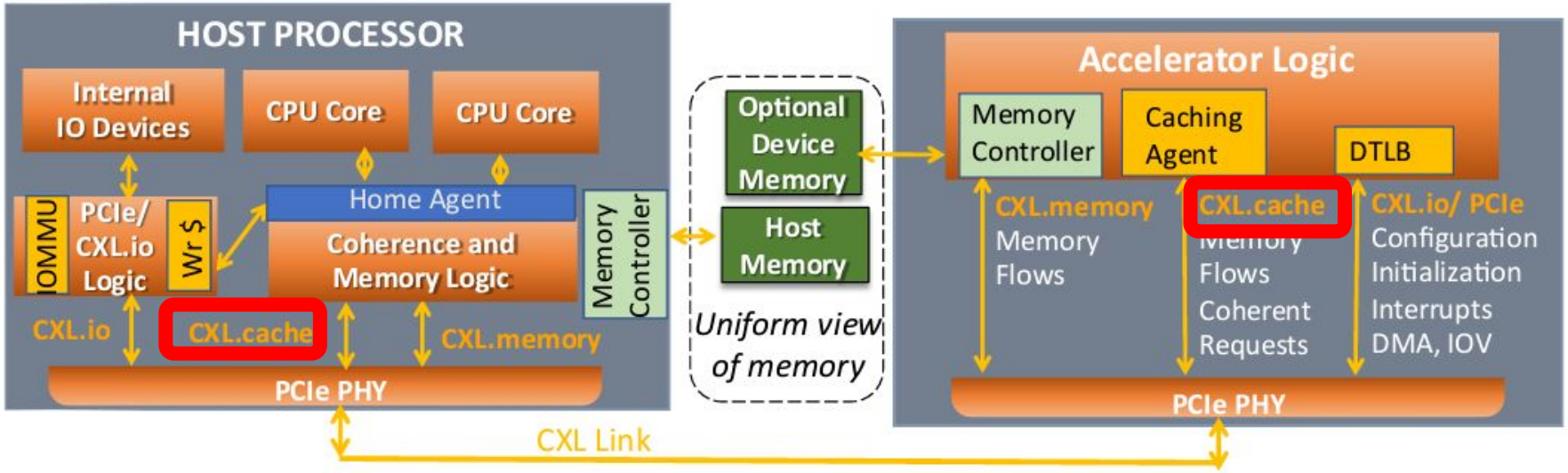
:: the 3 horsemen of CXL



:: the 3 horsemen of CXL : io - PCIe with relaxed ordering @64B

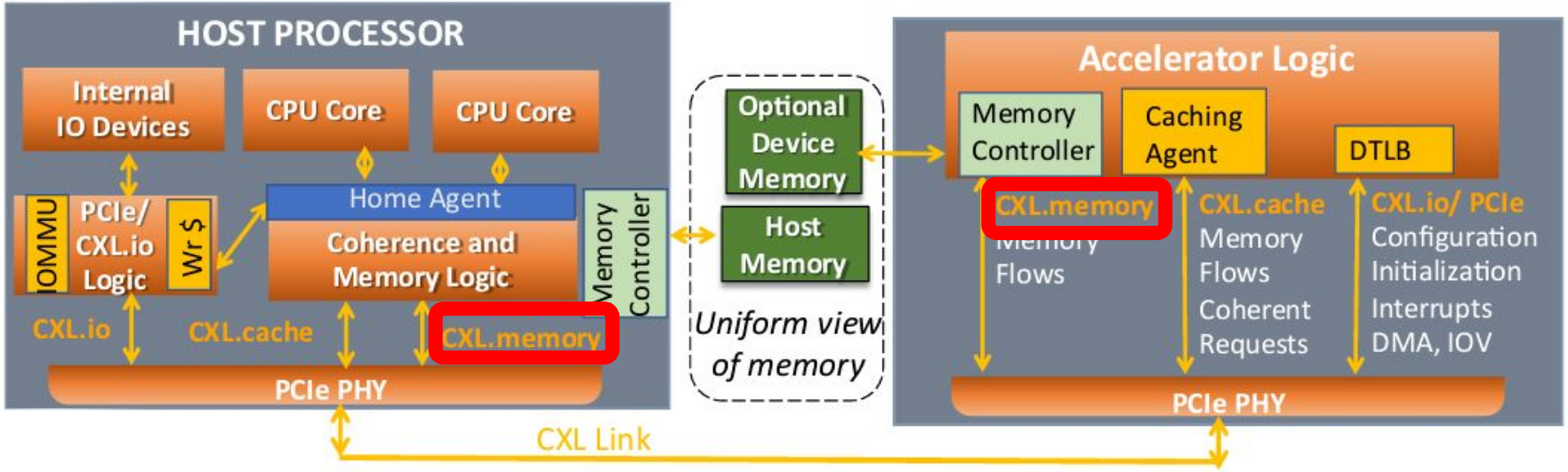


:: the 3 horsemen of CXL : cache - distributed control plane/state



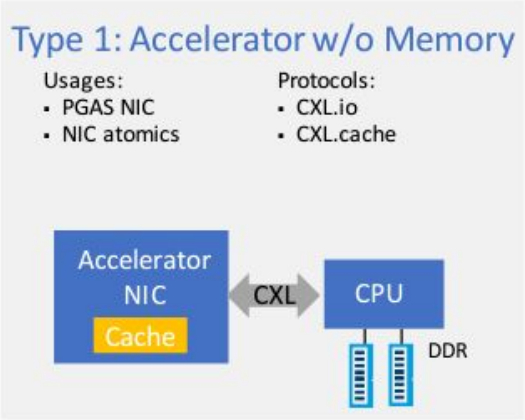
enfabrica

:: the 3 horsemen of CXL : memory - expansion of memory/buffers



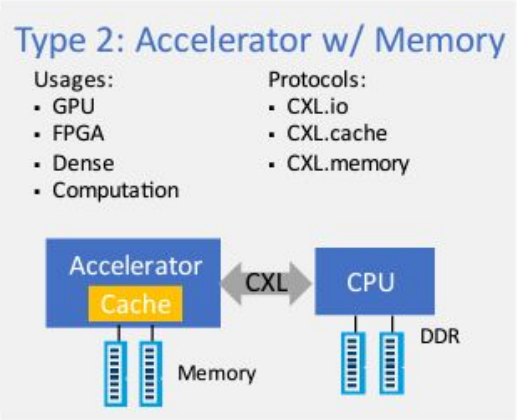
enfabrica

:: cxl target usage models : type 1



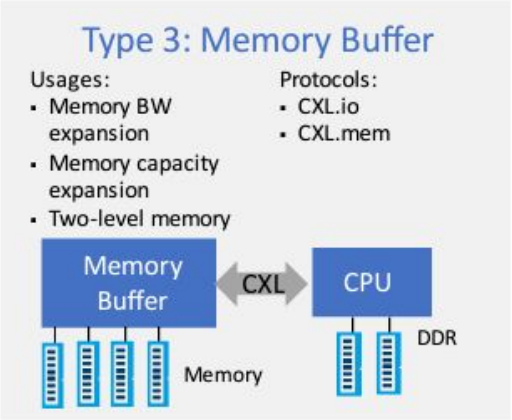
enfabrica

:: cxl target usage models : type 2

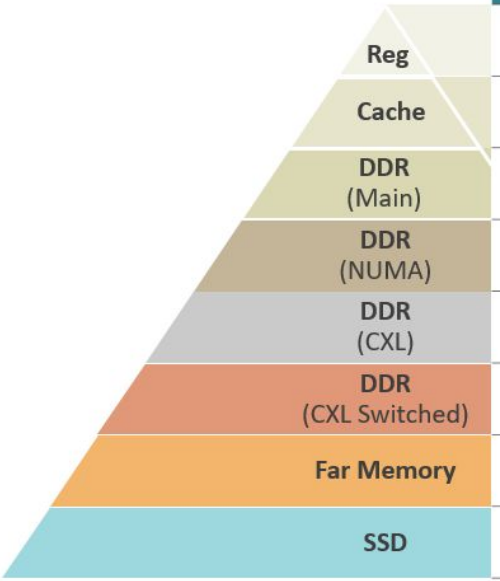


enfabrica

:: cxl target usage models : type 3

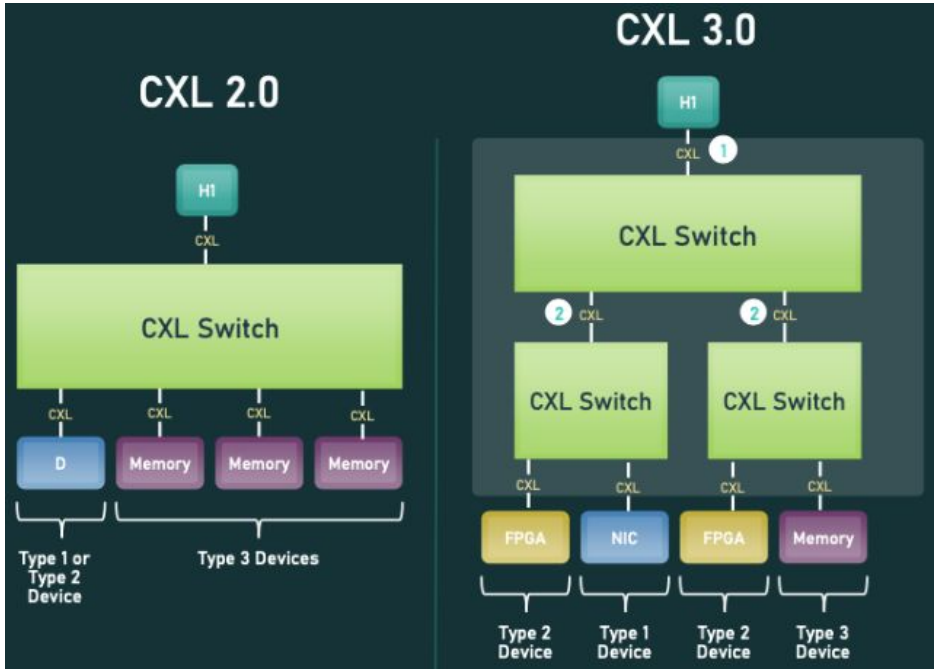


:: so where does it fit in the memory solar system



	Latency	Bandwidth / Channel	Max Capacity*	Significance	Programmers View
Reg	0.2ns		KB	In CPU	L1 - dereference pointer
Cache	40ns		KB		
DDR (Main)	80-140ns	32-51.2 GB/s (DDR5)	Up to 4TB		L2 - dereference pointer high perf memcpy
DDR (NUMA)	170-250ns	32-51.2 GB/s (DDR5)	Up to 8TB		
DDR (CXL)	170-250ns	32-51.2 GB/s (DDR5)	2-4 TB	CPU independent but local	L3 - dereference pointer high perf memcpy, swap
DDR (CXL Switched)	300-400ns	32-51.2 GB/s (DDR5)	64TB		
Far Memory	2-4us	100 GB/s (800g ethernet)	infinite	Network attached	L4 - memcpy, swap
SSD	50-100us				L5 - memcpy, swap

:: cxl - upwards, onwards and every direction possible



2.0 :

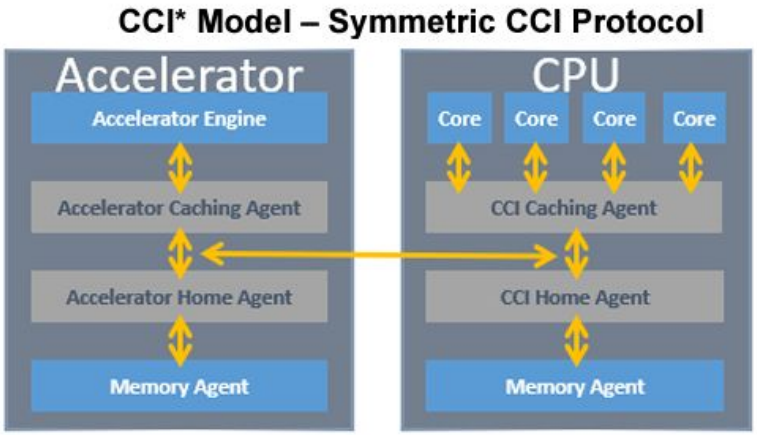
- Resource migration, memory scaling, and device fanout.
- Support for memory pooling. Adds SLD and MLD support
- Link-level Integrity and Data Encryption (CXL IDE)

3.0 :

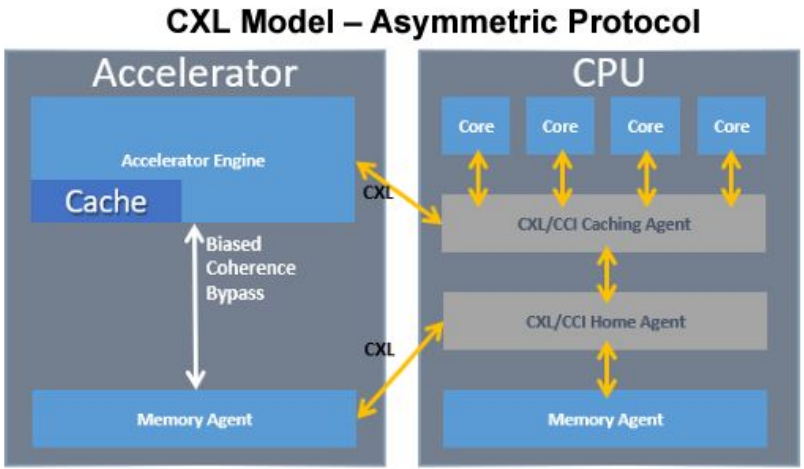
- Fabric capabilities
  - Multi-head and Fabric Attached Devices
  - Enhanced fabric management
  - Composable disaggregated infrastructure
  - Enhanced memory pooling
  - Multi-level switching
  - New enhanced coherency capabilities

Ref : https://community.cadence.com/cadence_blogs_8/b/breakfast-bytes/posts/hot-chips-cxl-tutorial

:: cxl is asymmetric and that is a good thing



*Cache Coherent Interface

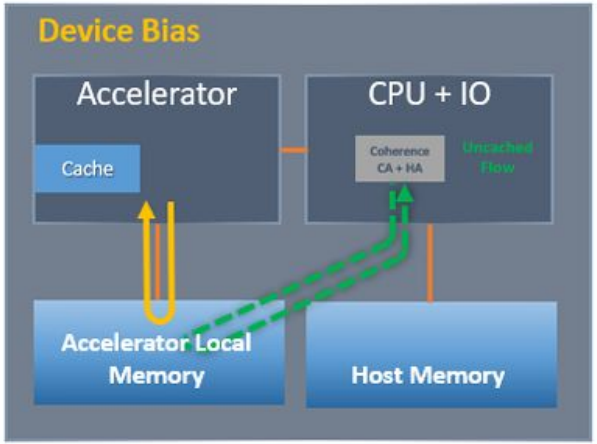


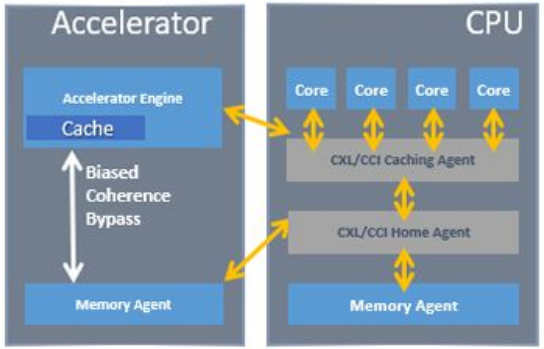
CXL Key Advantages:

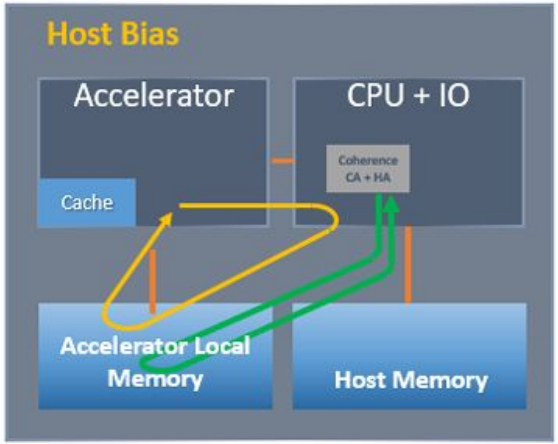
- Avoid protocol interoperability hurdles/roadblocks
- Enable devices across multiple segments (e.g. client / server)
- Enable Memory buffer with no coherency burden
- Simpler, processor independent device development

Ref : https://www.dmtf.org/sites/default/files/CXL_Overview_Virtual_DMTF_APTS_July_2020.pdf

:: asymmetry needs bias







Critical access class for accelerators is "device engine to device memory"

"Coherence Bias" allows a device engine to access its memory coherently without visiting the processor

Two driver managed modes or "Biases"

HOST BIAS: pages being used by the host or shared between host and device

DEVICE BIAS: pages being used exclusively by the device

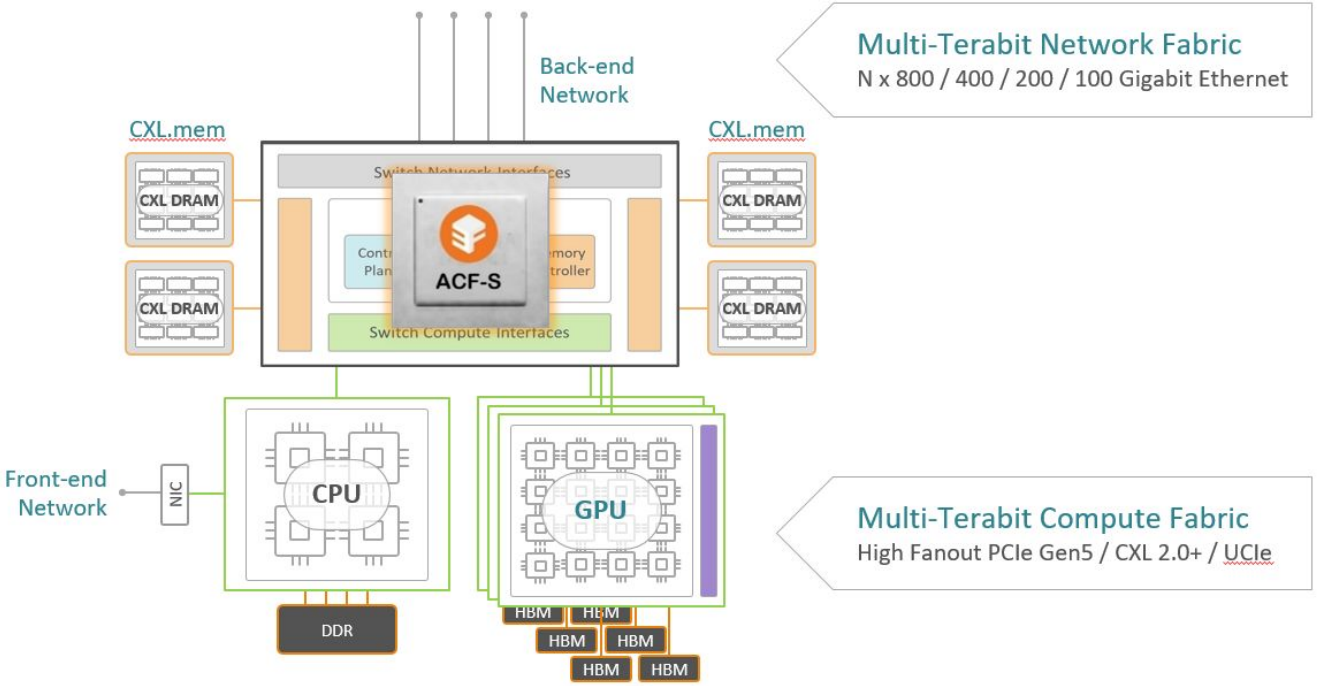
Both biases guaranteed correct/coherent

Guarantee applies even when software bugs or speculative accesses unexpectedly access device memory in the "Device Bias" state.

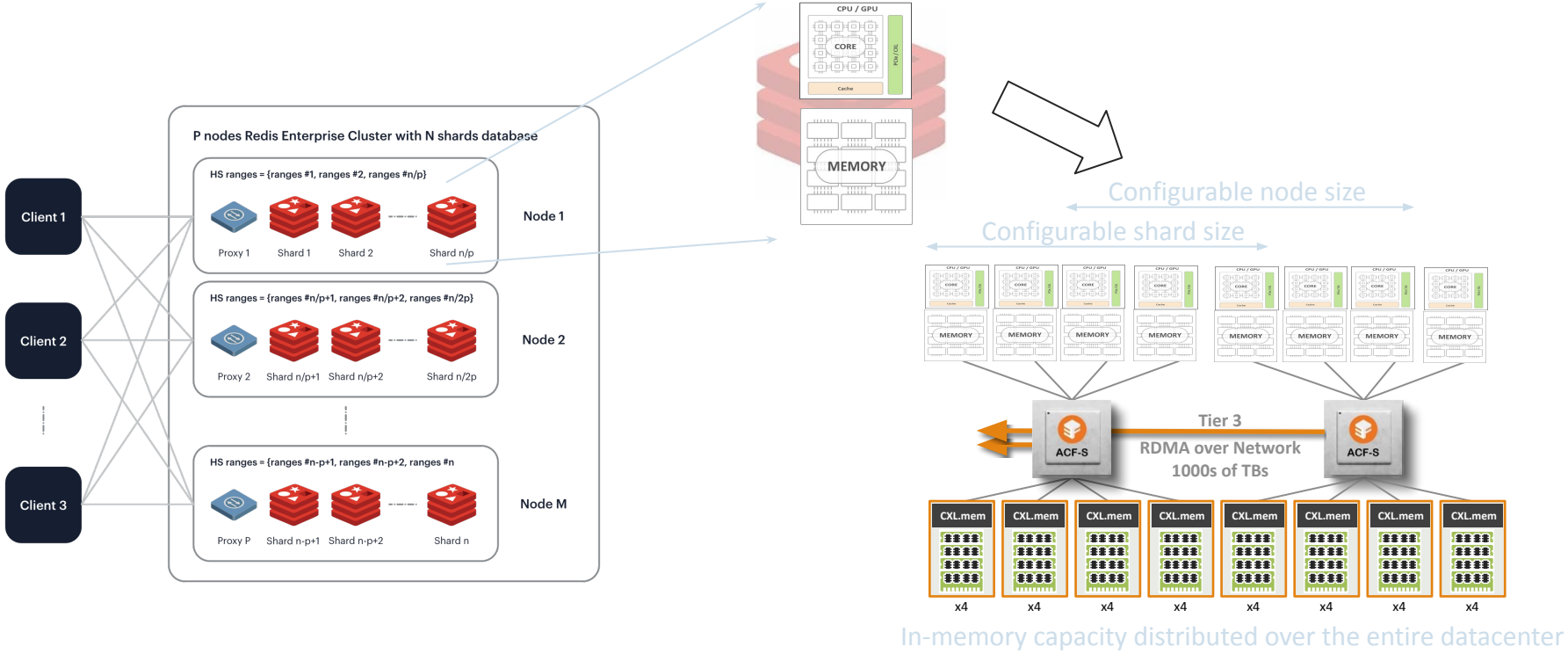
Ref : https://www.dmtf.org/sites/default/files/CXL_Overview_Virtual_DMTF_APTS_July_2020.pdf

enfabrica

:: cxl as imagined by enfabrica

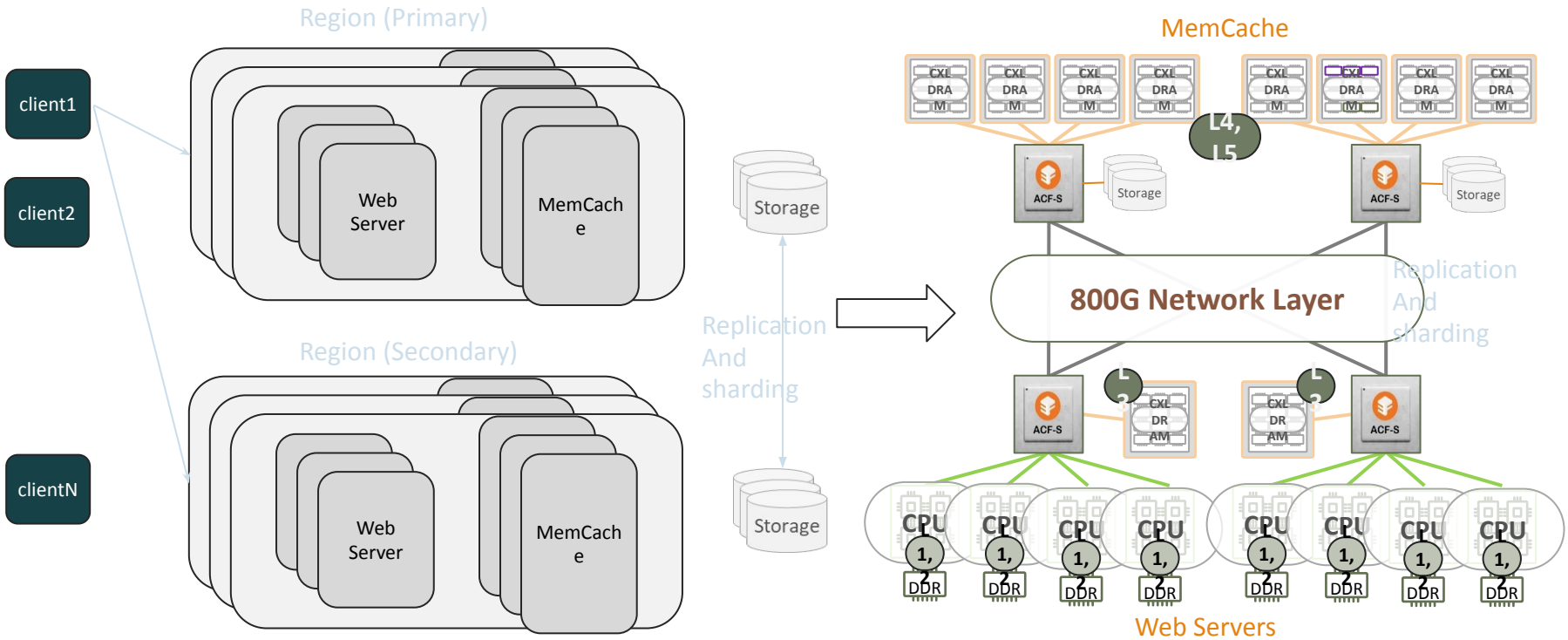


:: in-memory databases

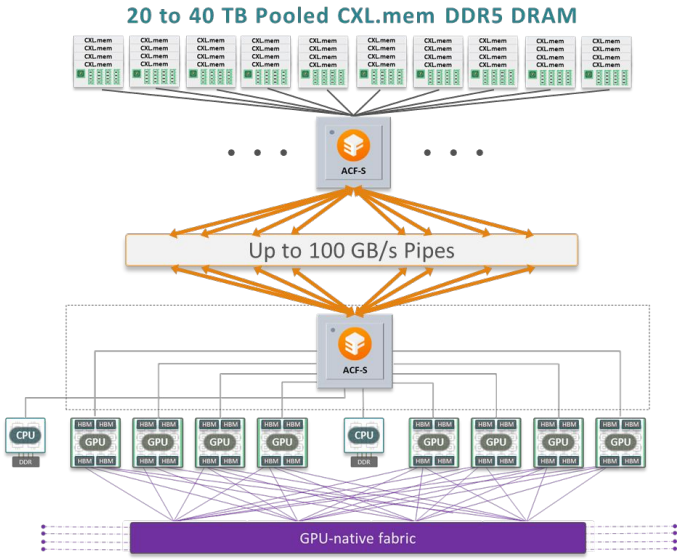


https://redis.com/redis-enterprise/technology/linear-scaling-redis-enterprise/

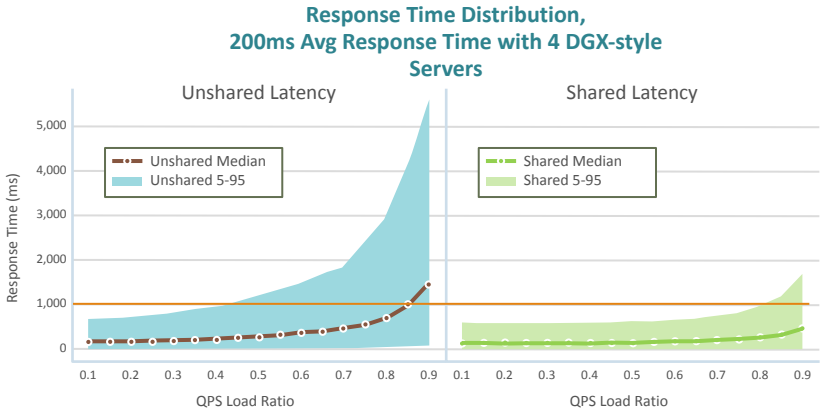
:: memcached and horizontal scaling



:: reducing inference fleet tco : 50% fewer gpus



Example LLM	Required QPS	# Required GPUs	# Required CPUs	User Context Capacity
User Context in CPU DRAM	**1K**	**128**	**16**	**80K**
User Context in ACF DRAM	**1K**	**64**	**8**	**80K**



Scaling memory headlessly with ACF-S reduces total xPU spend for LLM inference