東京 2025

LINUX PLUMBERS CONFERENCE

TOKYO, JAPAN / DECEMBER 11-13, 2025

To Online or Not To Online CXL Memory

東京 2025
LINUX
PLUMBERS CONFERENCE
TOKYO, JAPAN / DEC. 11-13, 2025

Performance: Is it good enough?

1-Socket System w/ 3:1 DRAM:CXL capacity

CXL Perf is ~1.6x latency, 1/10th BW compared to Socket attached DRAM

1) Latency differential between +35ns and +203ns depending on bandwidth load.
2) Latency only +35ns when DRAM loaded and CXL unloaded.
3) Idle memory from common workloads suggest this scenario is very common.
4) In practice, bandwidth limits were not an issue with even basic tiering (reclaim).
5) In practice, where workloads were capacity-bound, +CXL implied +performance.

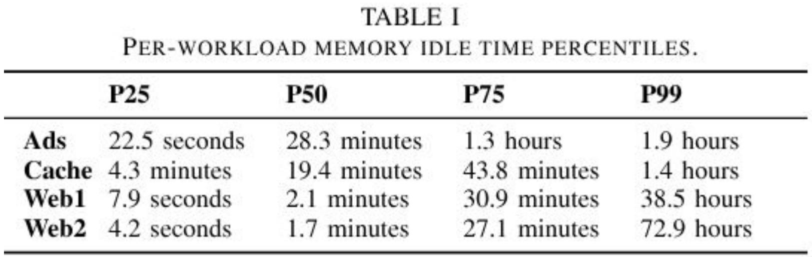


TABLE I

PER-WORKLOAD MEMORY IDLE TIME PERCENTILES.

	P25	P50	P75	P99
Ads	22.5 seconds	28.3 minutes	1.3 hours	1.9 hours
Cache	4.3 minutes	19.4 minutes	43.8 minutes	1.4 hours
Web1	7.9 seconds	2.1 minutes	30.9 minutes	38.5 hours
Web2	4.2 seconds	1.7 minutes	27.1 minutes	72.9 hours

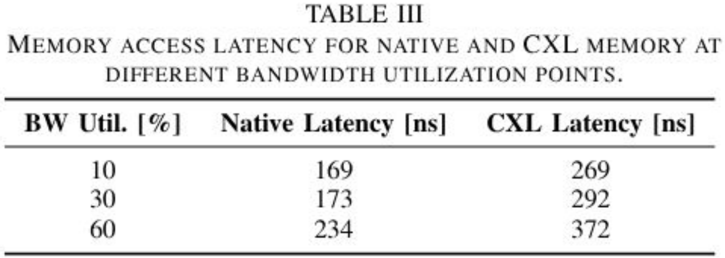


TABLE III

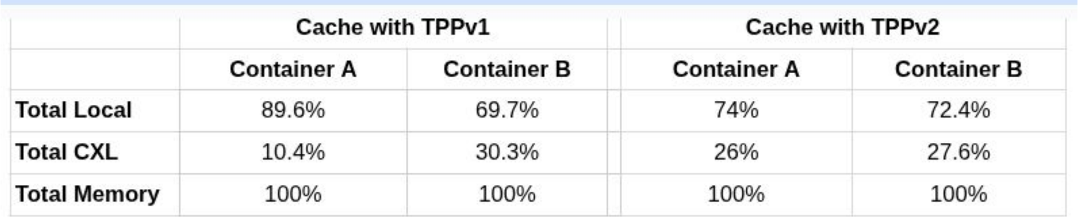
MEMORY ACCESS LATENCY FOR NATIVE AND CXL MEMORY AT DIFFERENT BANDWIDTH UTILIZATION POINTS.

BW Util. [%]	Native Latency [ns]	CXL Latency [ns]
10	169	269
30	173	292
60	234	372

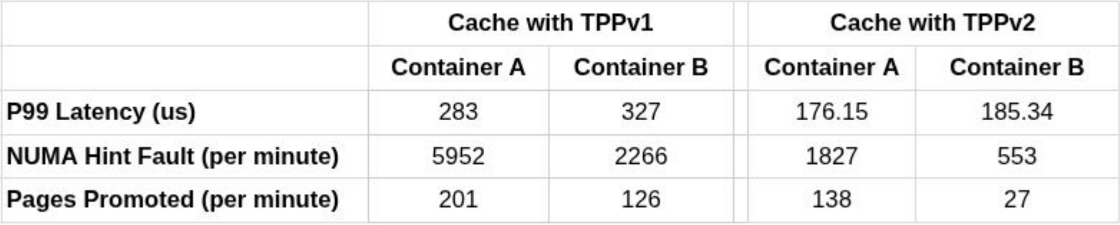
* preliminary data from paper expected early 2026

Performance: What about multi-tenant?

With some reclaim fixes*, it can be reliable.
Normal noisy neighbor issues apply. Careful not to over-sub CXL bandwidth.



	Cache with TPPv1		Cache with TPPv2	
	Container A	Container B	Container A	Container B
Total Local	89.6%	69.7%	74%	72.4%
Total CXL	10.4%	30.3%	26%	27.6%
Total Memory	100%	100%	100%	100%



	Cache with TPPv1		Cache with TPPv2	
	Container A	Container B	Container A	Container B
P99 Latency (us)	283	327	176.15	185.34
NUMA Hint Fault (per minute)	5952	2266	1827	553
Pages Promoted (per minute)	201	126	138	27

* TPPv2 fairness improvements, new RFC soon

ZONE_MOVABLE: More than just hotplug

ZONE_MOVABLE can be used to prevent kernel and other system-critical allocations from landing on farther away nodes. Maybe not it’s “intended use”.

- Enhances TPP’s ability to promote anything
- Makes multi-tenant more reliable as jobs come and go.
- Make failure not system-fatal (SIGBUS instead of MCE)
- Hotplug path makes driver-signal useful for avoiding online

This has all been useful for systems in production environments.

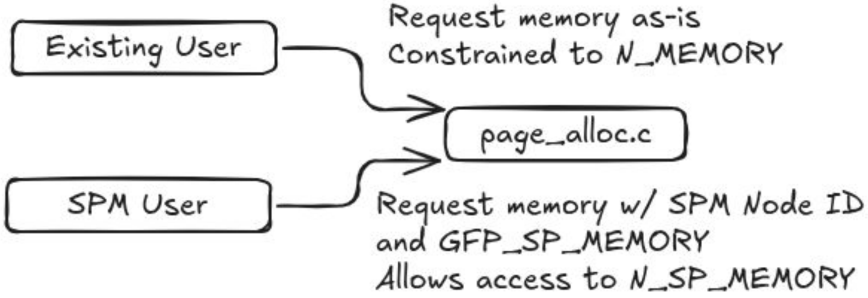
Specific Purpose Memory

For significantly slower memory, or memory with “Special Features”...

Change the definition of “Online”.

- Still have `struct page`
- Do not place in N_MEMORY node
- Make new “Private” memory nodes (SPM)
- Maybe use ZONE_DEVICE instead

Come to MM session for full details



東京 2025

LINUX PLUMBERS CONFERENCE

TOKYO, JAPAN / DECEMBER 11-13, 2025