Linux
Plumbers
Conference
Dublin, Ireland September 12-14, 2022

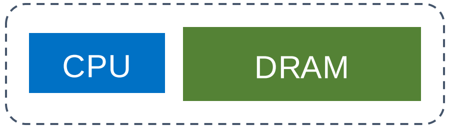
Live In a World With Multiple Memory Types

Huang, Ying

Updates In Last Year

Items	Contributor	Status
NUMA balancing based promotion – basic support	Ying	v5.18
NUMA balancing based promotion – hot page selection	Ying	mm-unstable
Explicit memory tiers	Aneesh	mm-unstable
Memory tiers user space interface	Aneesh, Wei	mm-unstable
Interleave among memory tiers	Johannes	WIP
Respect NUMA policy/cpuset in demotion	Feng	WIP
Partition a type of memory (DRAM) among cgroups	Tim	WIP

Initially…



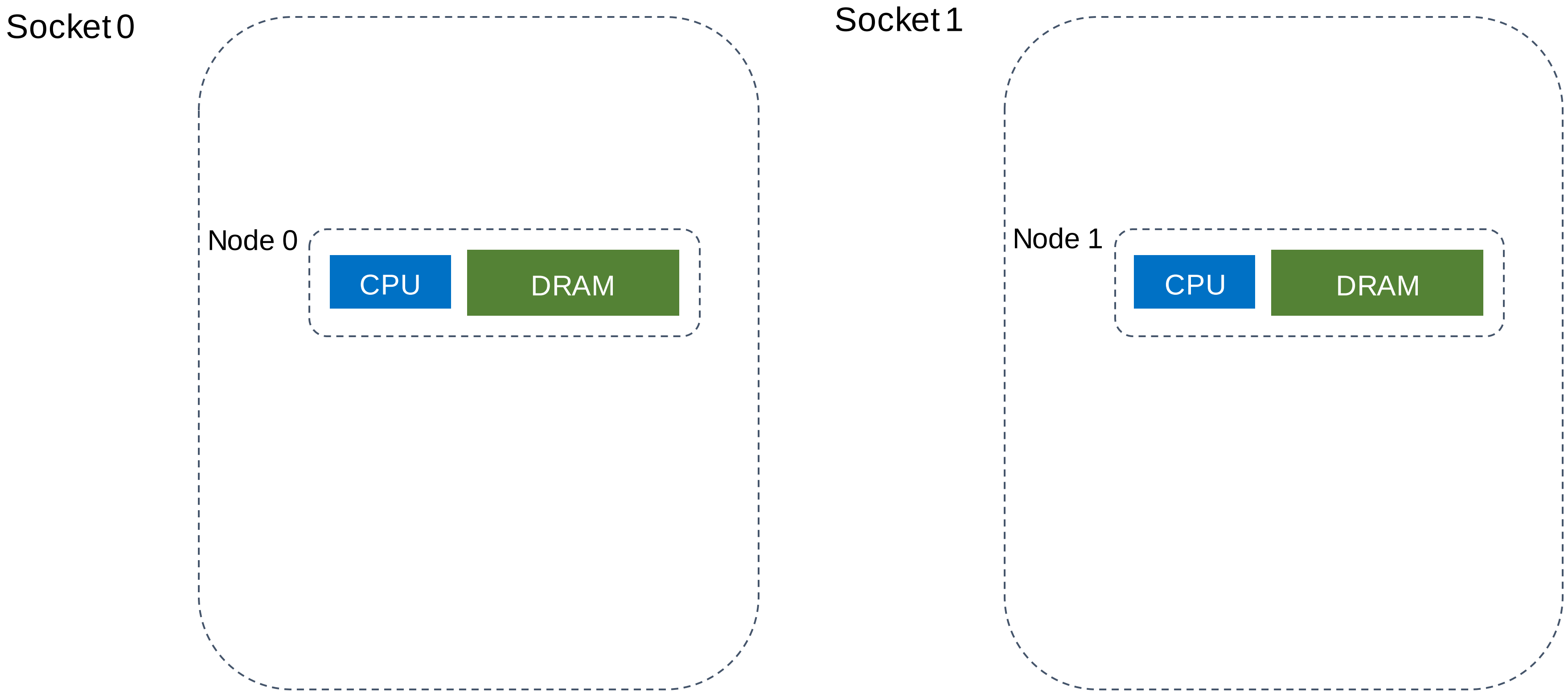
CPU

DRAM

- Initially, all memory are just simple DRAM

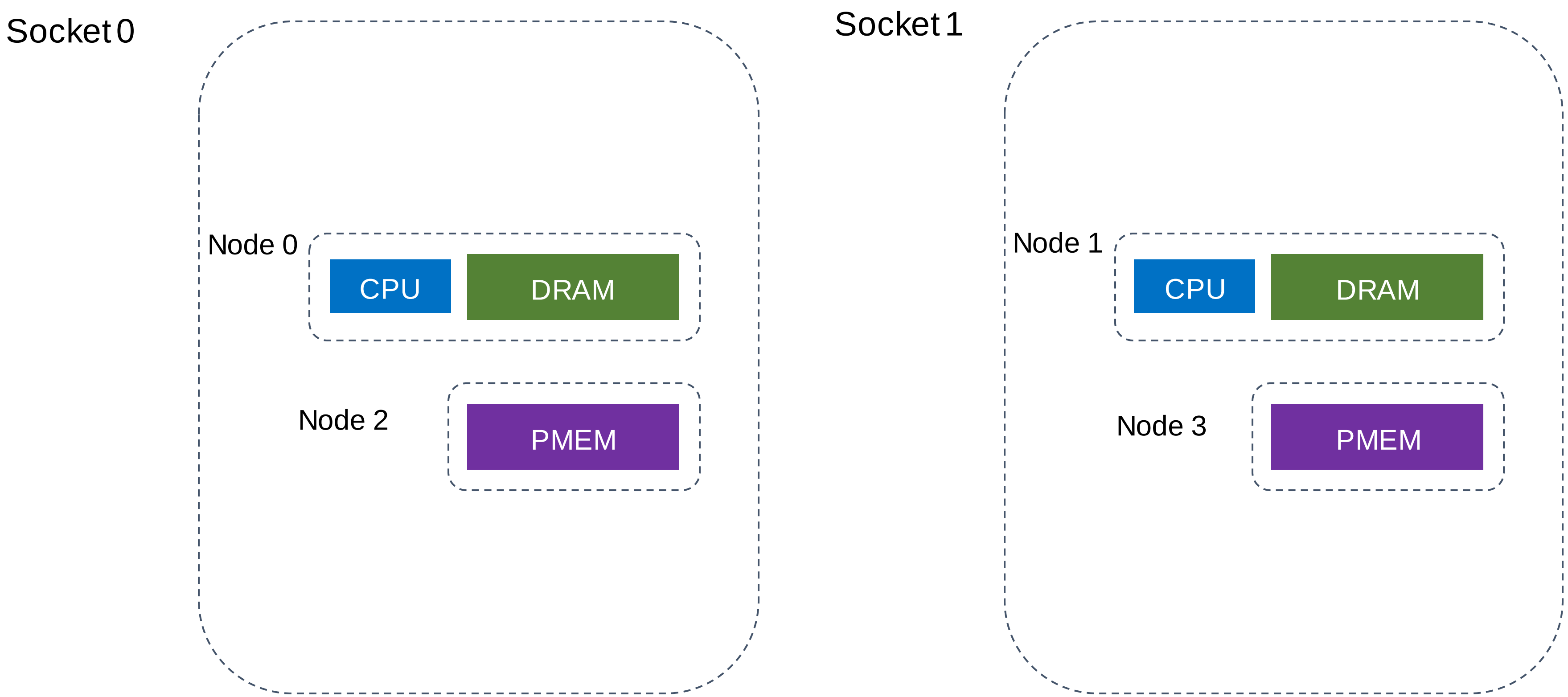
Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

NUMA



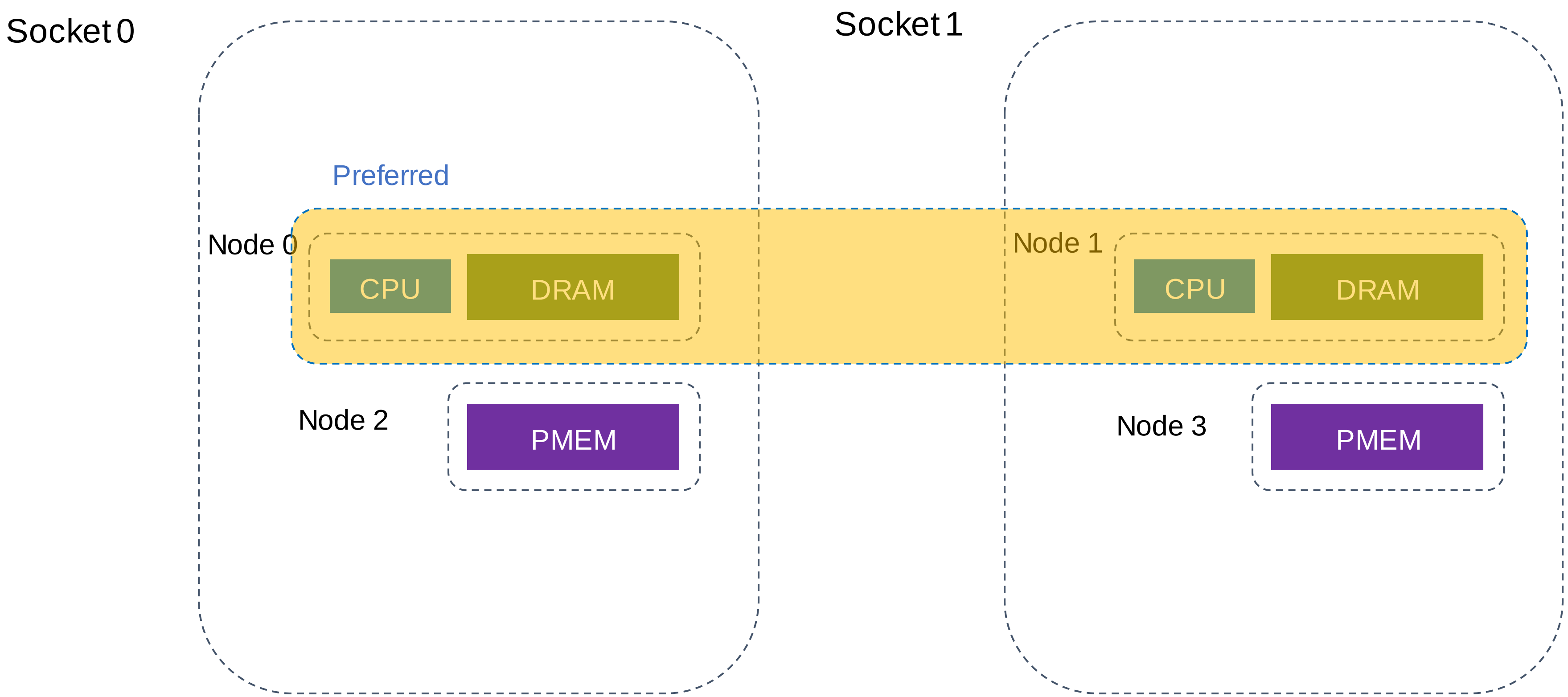
- Then, we get remote DRAM
- We manage it with NUMA policy, cpuset, NUMA balancing, etc.

PMEM



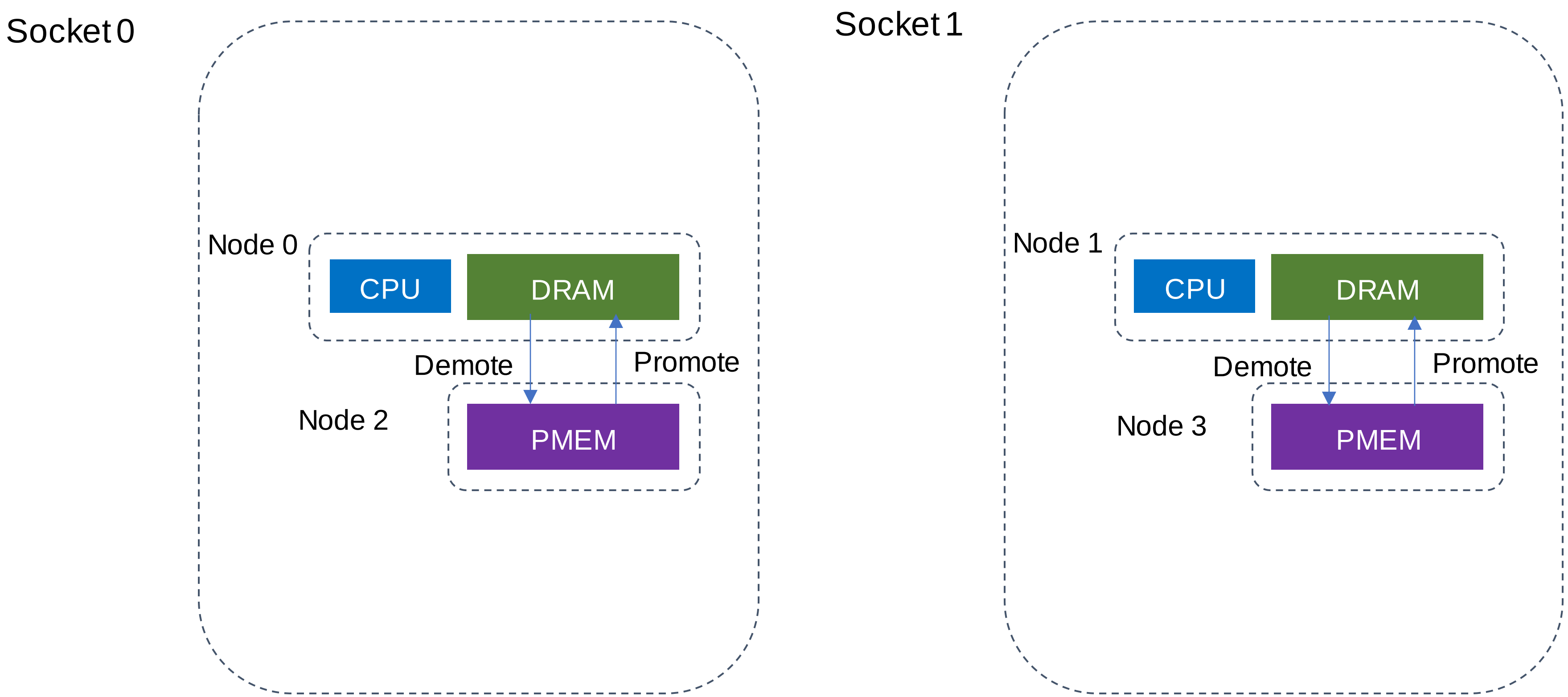
- Then, we get PMEM
- We put them in separate NUMA nodes to use NUMA mechanism/API

MPOL_PREFERRED_MANY



- NUMA mechanism/API are extended
- E.g., prefer remote DRAM over local PMEM

Page Placement Optimization

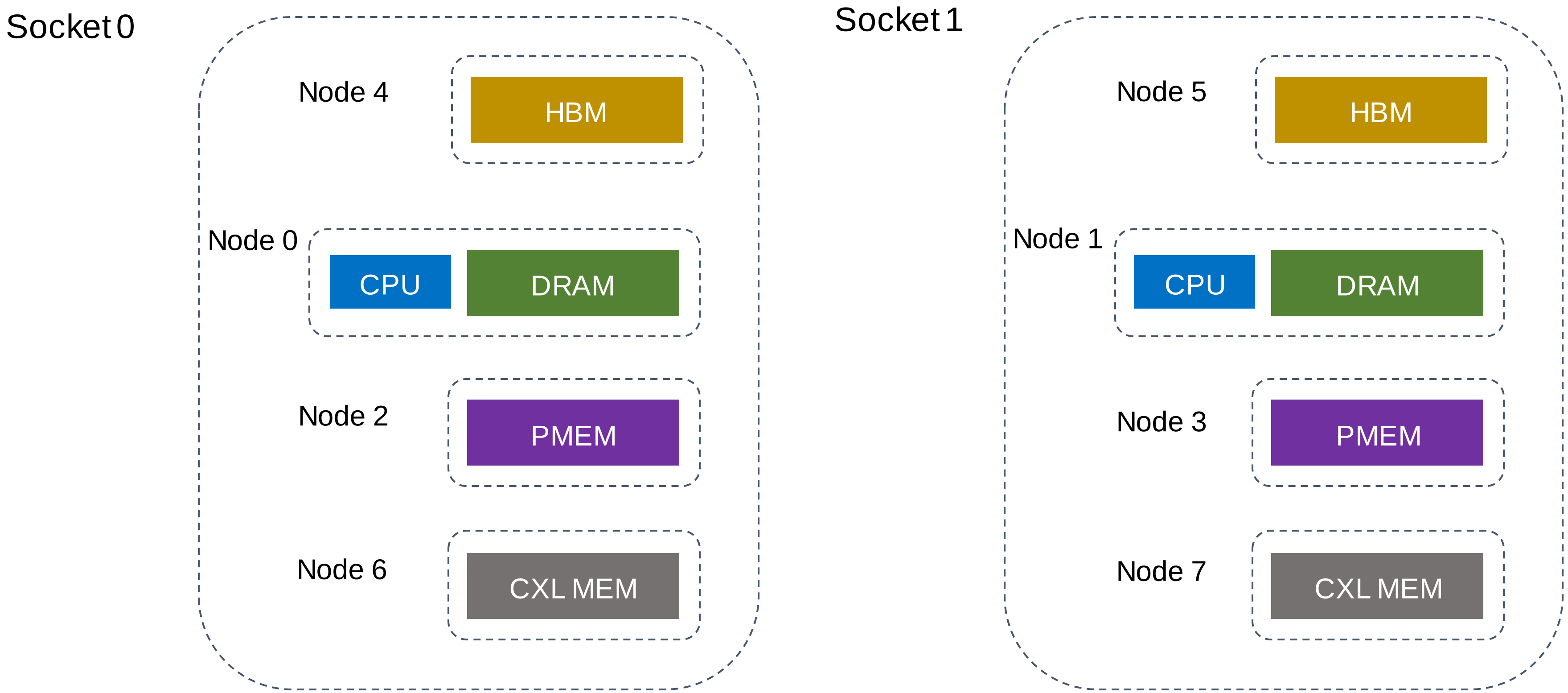


- Demote: per-node page reclaiming
- Promote: NUMA balancing

Linux Plumbers Conference

Dublin, Ireland September 12-14, 2022

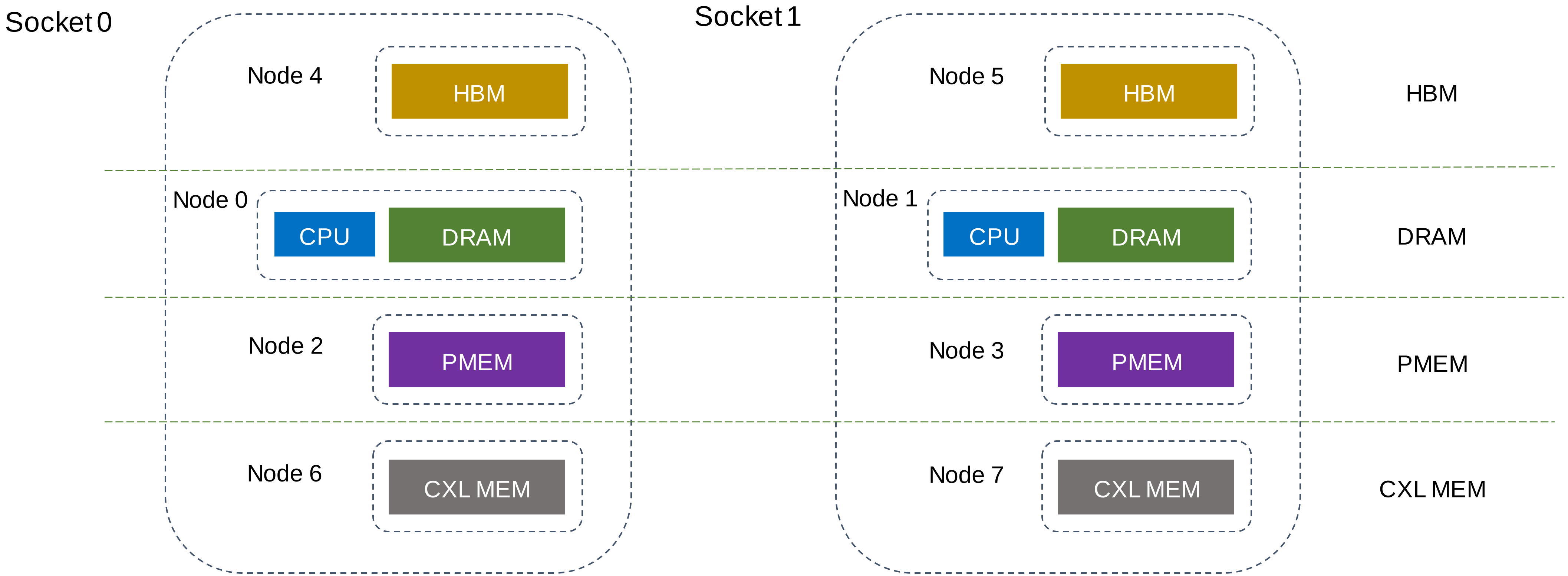
Even More Memory Types



- Even More memory types are coming, HBM, and CXL memory devices, etc.
- How to manage them?

Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

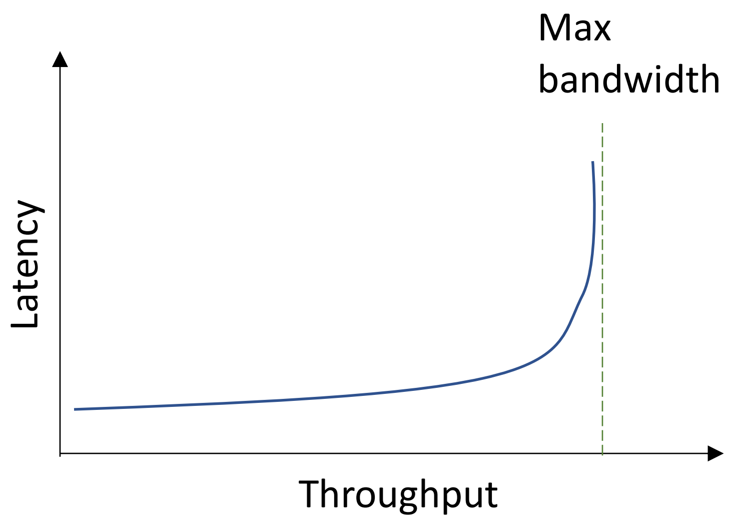
Memory Types



- Memory devices with same driver, link, media, etc.
  - Same performance.

Memory Types: Abstract Distance

- Distance from CPUs to a type of memory in the same socket
  - Inspired by NUMA distance
  - Orthogonal with NUMA topology
  - Smaller is better
- Performance metric of a memory type
  - Latency + bandwidth: how to **combine**?
  - One possibility: latency under expected access throughput
    - Workload dependent



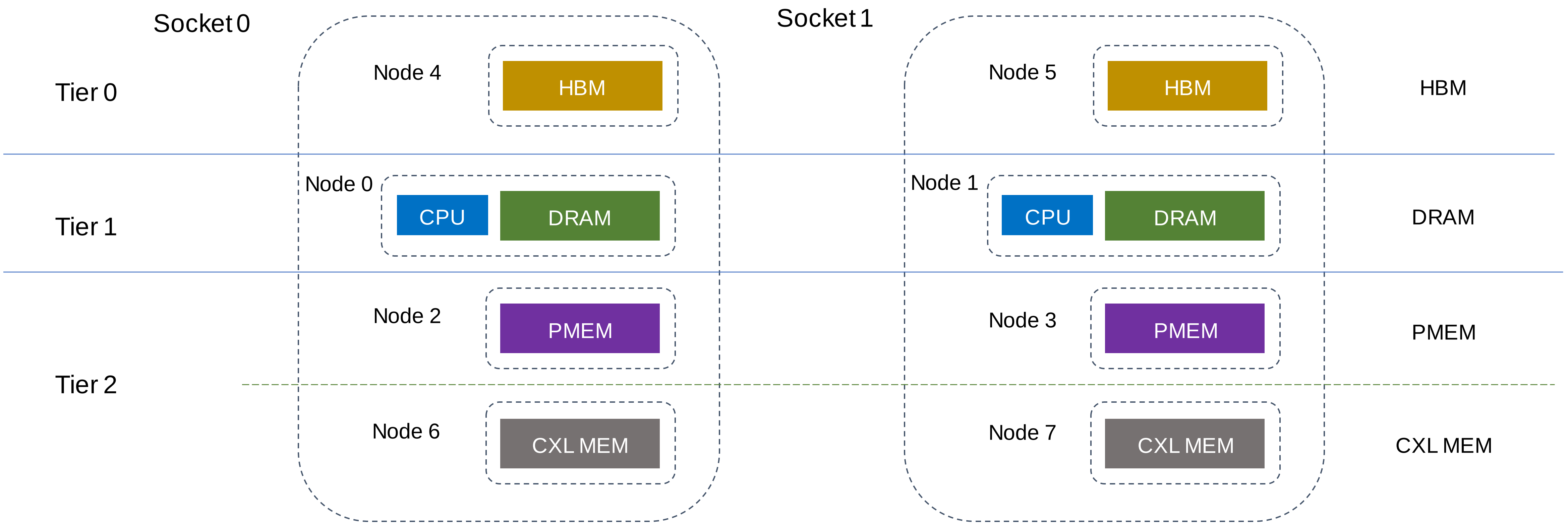
Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

Memory Types: Sysfs Interface

- /sys/devices/virtual/memory_type/memory_typeN
    - name: HBM, DRAM, PMEM, CXL MEM, etc.
    - nodeX: symbol links to the NUMA nodes of the memory type
    - default_abstract_distance: default provided by driver
    - **abstract_distance_offset**: override by users
        - Deal with firmware issue
        - Reflect actual latency under expected access throughput
    - latency, bandwidth?: performance metrics (ACPI HMAT, CXL CDAT)
- Subsystem?
    - “system” is considered legacy now. What is the appropriate subsystem? virtual?

Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

Memory Tiers



- Memory tier: memory types in a range of abstract distance
- Performance and policy

Memory Tiers: Sysfs Interface

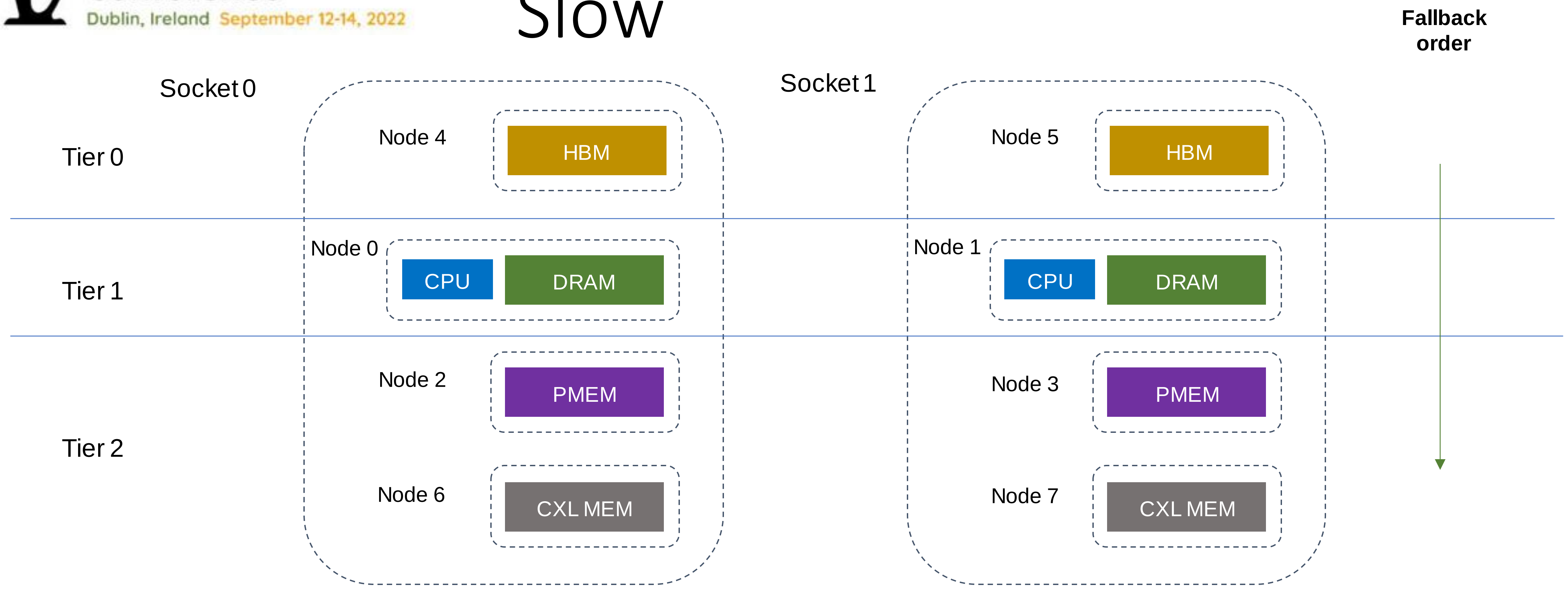
- /sys/devices/virtual/memory_tier
  - memory_tierN
    - nodes: list of nodes of the memory tier
    - memory_typeM: symbol links to the memory types in the memory tier
    - abstract_distance_start: start of abstract distance range
    - abstract_distance_end: end of abstract distance range
  - default_memory_tier: symbol link to memory tier of normal DRAM
  - **abstract_distance_chunk_size**: customize abstract distance range
    - Abstract distance chunks: [0, chunk_size); [chunk_size, 2*chunk_size); …
    - Apply users' policy to group memory types
      - Alternative method: customize the abstract distance of memory type

Memory Tiers: Sysfs Interface - 2

- Memory tier device ID
  - 0, 1, 2, …
    - Intuitive to understand
    - ID may change with node online/offline
  - abstract_distance_start / abstract_distance_chunk_size or abstract_distance_start
    - ID may change with abstract distance ranges customization
    - Memory tiers relationship via sorting IDs

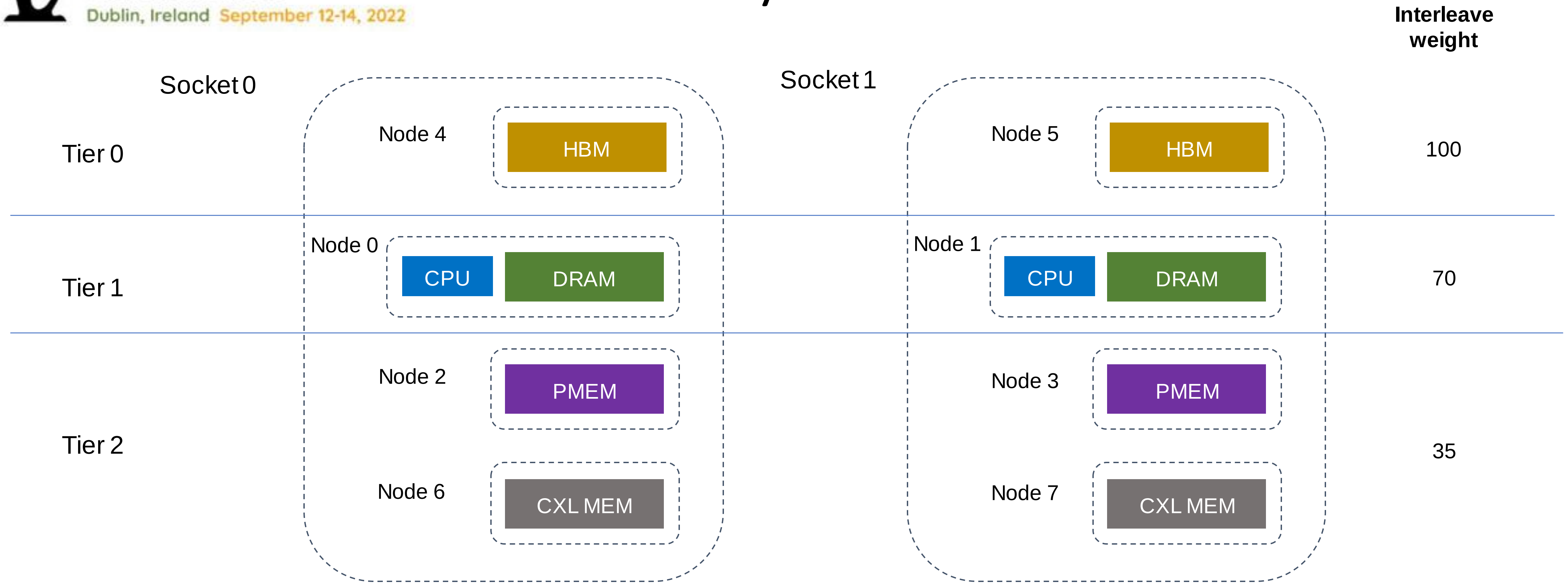
Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

Memory Tiers: From Fast to Slow



- Default memory allocation fallback order: from fast to slow
- Take full advantage of faster memory, hot pages are allocated first

Memory Tiers: Interleave

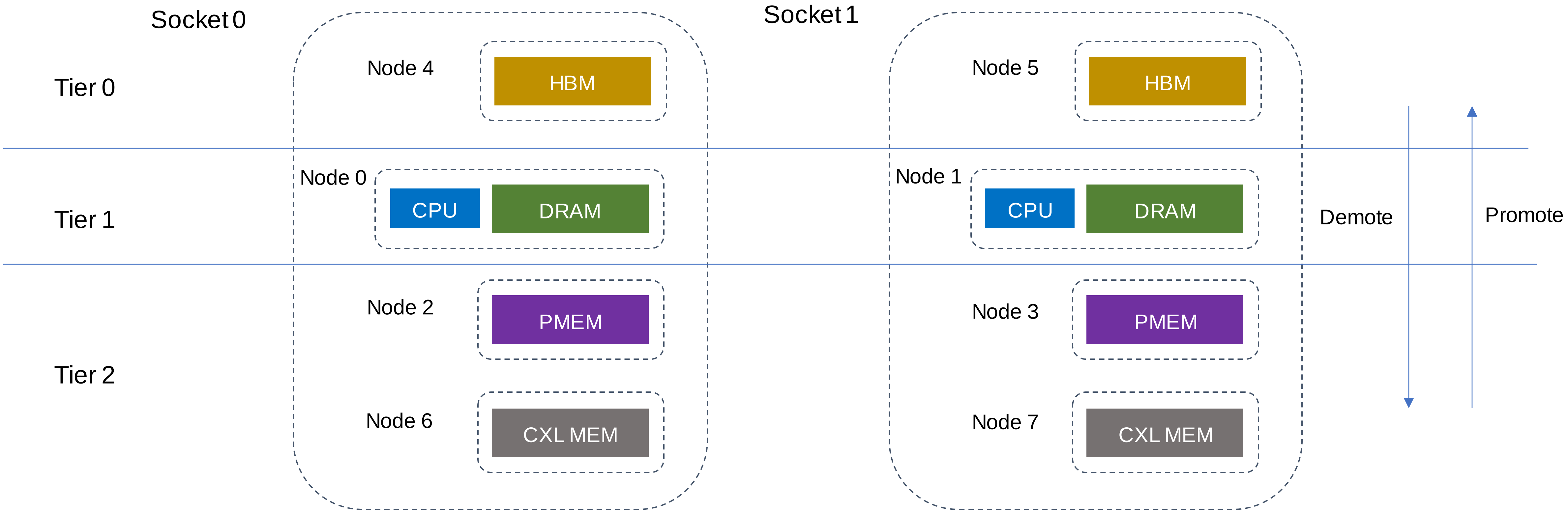


	Socket 0	Socket 1	Interleave weight
Tier 0	Node 4 HBM	Node 5 HBM	100
Tier 1	Node 0 CPU DRAM	Node 1 CPU DRAM	70
Tier 2	Node 2 PMEM Node 6 CXL MEM	Node 3 PMEM Node 7 CXL MEM	35

- Interleave among memory tiers: maximize memory throughput
- /sys/devices/virtual/memory_tier/memory_tierN/interleave_weight

Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

Memory Tiers: Page Placement Optimization



- Demotion was rebased on explicit memory tiers
- Promotion wasn't changed much

Memory Tiers: Demotion and Explicit NUMA Policy

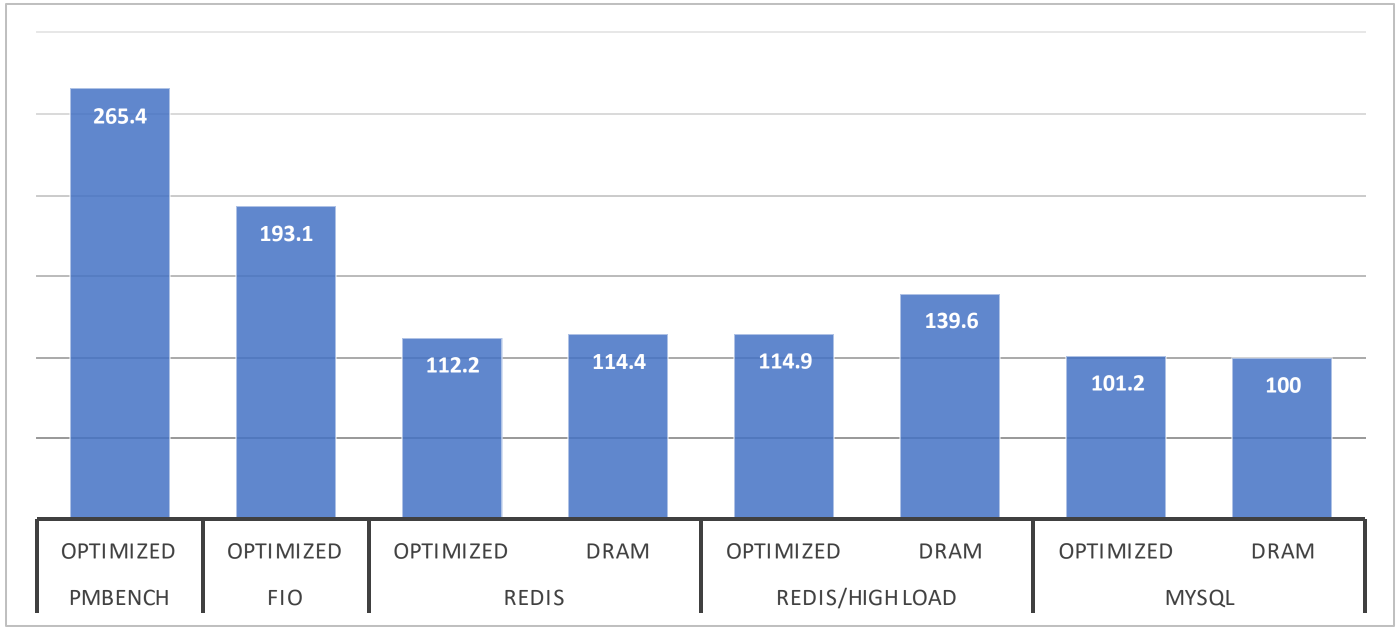
- NUMA policy/cpuset needs to be respected during demotion
  - To avoid cross-socket memory accessing
  - To implement placement control: e.g., run in normal DRAM
- Cpuset
  - Cgroupv2: via unified hierarchy
    - page -> memcg -> cgroup -> cpuset
  - Cgroupv1**?**
- VMA NUMA policy: mbind()
  - page -> rmap -> VMA -> policy
- Task NUMA policy: set_mempolicy()
  - **Not** all information is available during demotion
  - Best effort: page -> rmap -> VMA -> mm -> owner (task) -> policy

Memory Tiers: Performance Evaluation

- Hardware
  - 2-socket server with DRAM + Optane DCPMM
  - DRAM to PMEM ratio: 1:4
- Configurations
  - Base: DRAM + PMEM with demotion/promotion disabled
  - Optimized: DRAM + PMEM with demotion/promotion enabled
  - DRAM: DRAM only, same total memory size as base/optimized

Linux Plumbers Conference
Dublin, Ireland September 12-14, 2022

Memory Tiers: Performance Evaluation – Test Results



- Score of base configuration: 100
- Micro-benchmarks show effectiveness of the optimization
- Redis results are good if load isn’t too high
- The bottleneck of MySQL is disk random sync write latency

TODOs

- Finish the memory tiers user space interface. **More Review**!
- Build memory types from various information (ACPI HMAT, SLIT, etc.)
- Unmapped file cache pages promotion
- Page demotion/promotion thrashing control solution
- Avoid to reclaim too many reclaimable/unmovable pages (inode/dentry cache, etc.) during demotion
- Promoting ahead of accessing
- Further improve the demotion/promotion algorithm

Thanks!