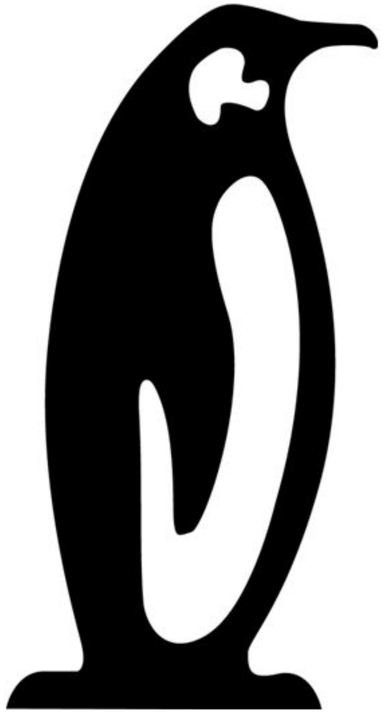


Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024

CXL Dynamic Capacity Devices

Ira Weiny, Jonathan Cameron, Navneet Singh

v2

LINUX PLUMBERS CONFERENCE | Vienna, Austria
Sept. 18-20, 2024

Current Status

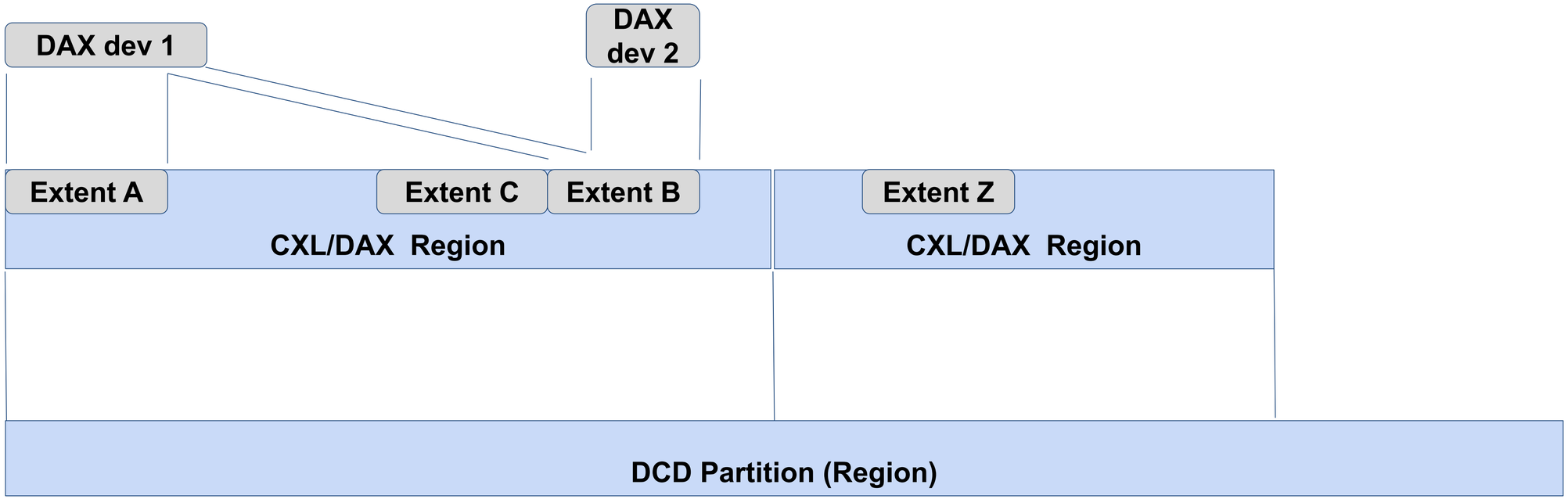
- Patch set v3 (really v2) is on the list
  - Comments have been mild
- ndctl support
  - DCD region creation
  - extents shown in regions
  - cxl-test
- QEMU support upstream
- Discussion at last community forum
  - Tag handling
- qos class has been added since v3
  - sysfs entries changed

Desired use case

- Two modes exist for Extents which are added
  - Online as system memory - “system-ram”
  - Leave as DAX memory - “devdax”
- “system-ram” has issues
- “devdax” is where it’s at

Extent Tag handling (Current)

- Ignore tags
  - Current implementation
  - Hard to change policy later

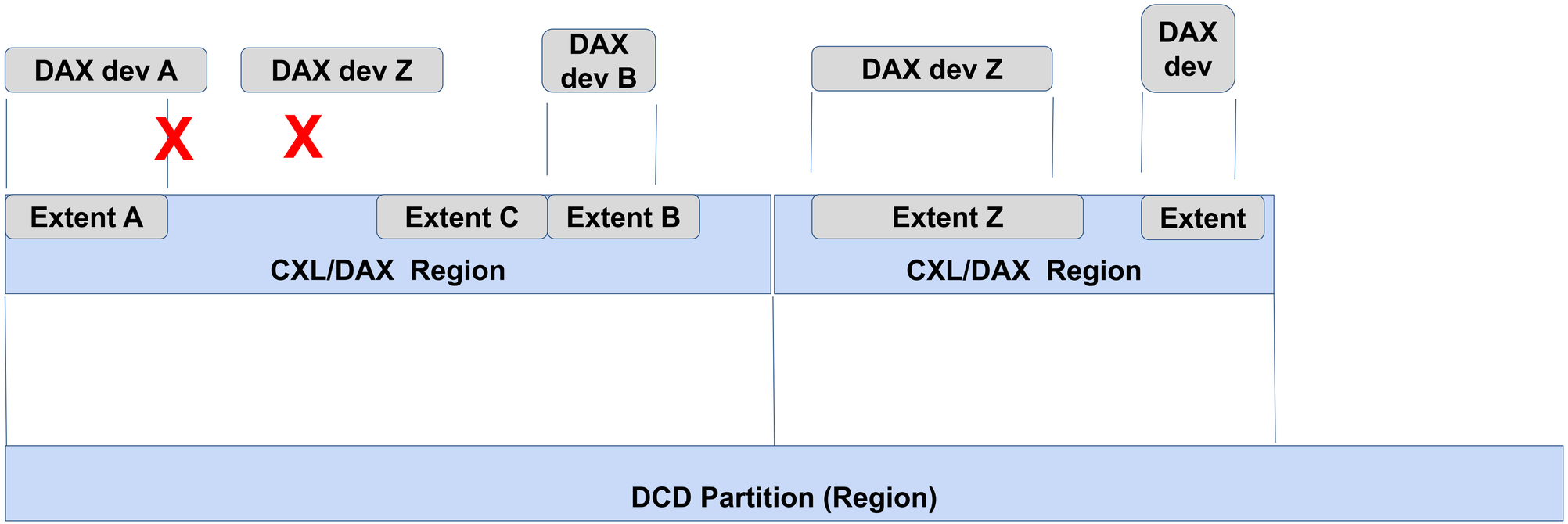


Extent Tag handling (Possible)

- Reject tags
  - correlates well with the lack of tag for region
  - Allows current region interfaces to work ‘as expected’ as ‘sans tag’ regions
- Assign tags to DAX devices which only allocate from matching extents
  - Issue with region available size calculations
- ~~Assign region tag based on first tag surfaced~~
  - ~~Weird interface~~
- ~~Assign region tag based on user input~~
  - ~~How to correlate DPA to tag?~~
  - ~~reject tags which don’t match?~~

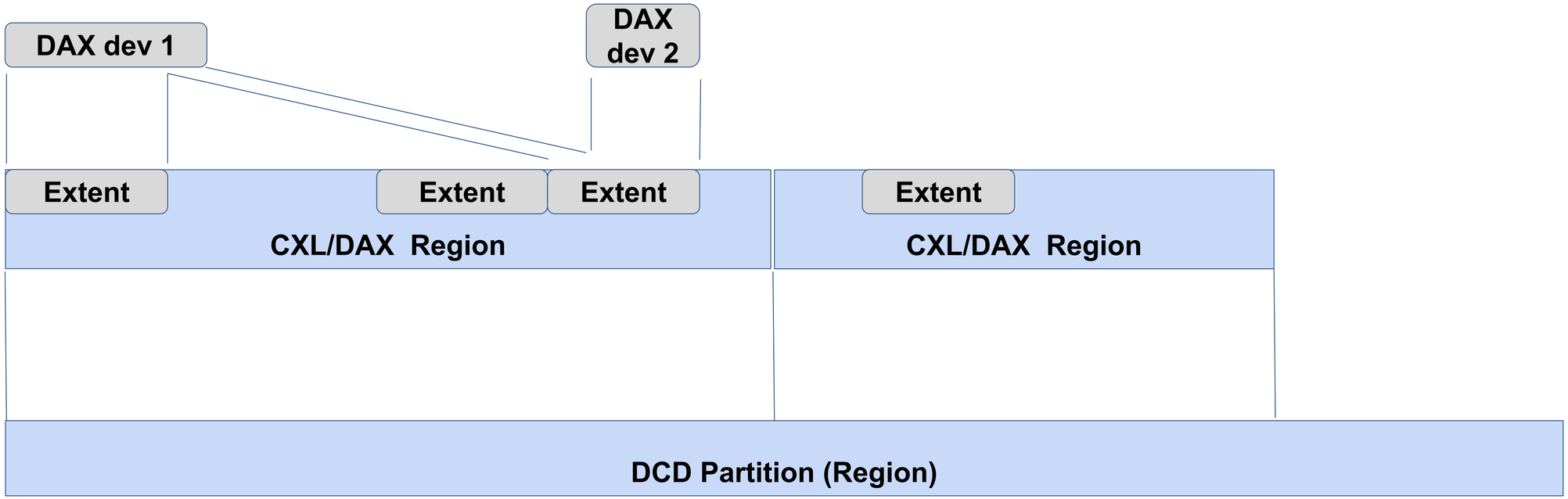
DAX devices attach to matching tagged extents

- Extents are tagged A, B, C, Z, and no tag
- Requires new interface for DAX devices to be tagged
- Follow sequence requirements



Extent Tag V4 proposal

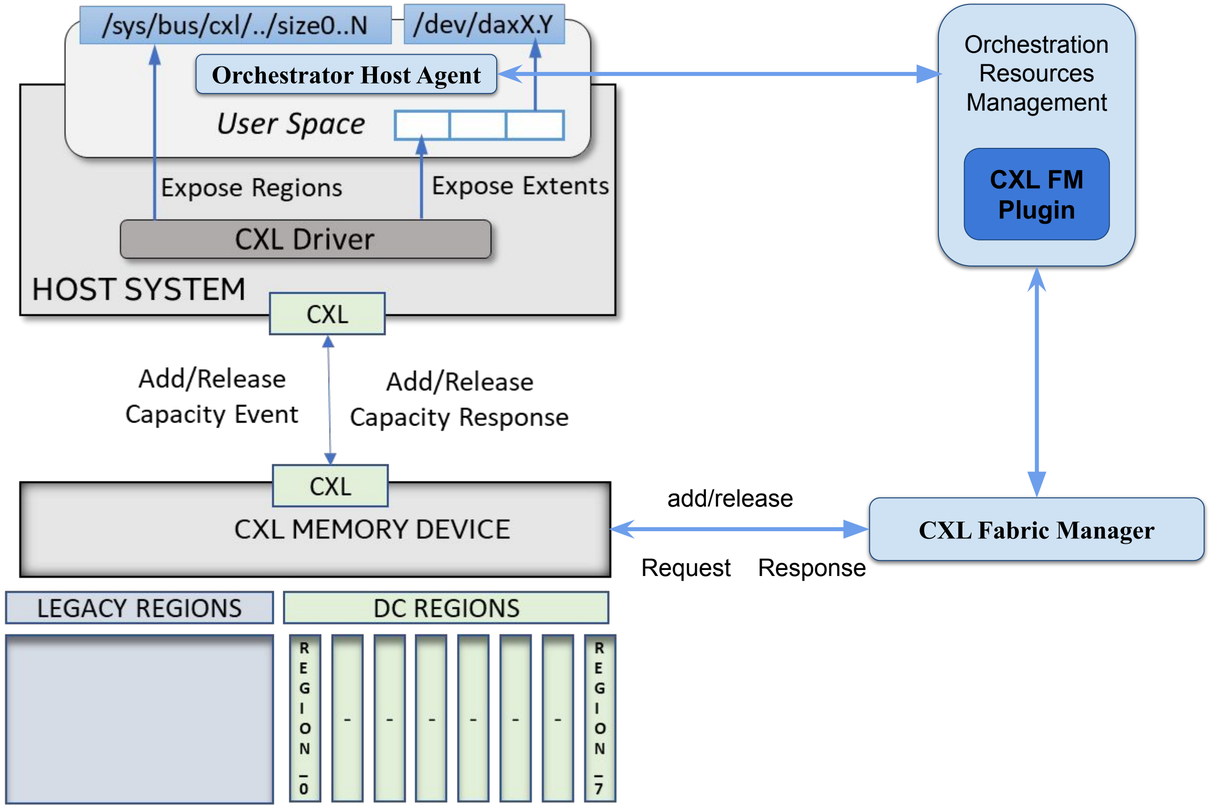
- Reject extent tags
  - correlates well with the lack of tag for region
  - Allows current region interfaces to work ‘as expected’ as ‘sans tag’ regions - or 0 tag regions



Fabric Manager/Orchestrator interactions

DCD Data Center - Open Questions

- Host/Orchestrator User Agent?
- Orchestrator functionality to support dynamic allocation
- Orchestrator ⇔Fabric Manager interfaces/communication
- Who are the right folks to work on this?



Future Support

- QoS work
  - Added in V4
- Interleaving (no use cases yet, development in back ground)
- Mapping of tagged extents
  - previous slide
- manage split/merge extents?
  - splitting extents will be a more difficult problem to resolve which parts of the extent are mapped.
  - merging new extents will be easier (but not necessary)
- DAX device interfaces
  - memfd? (Still researching)

Comparisons to RDMA

- DCD shared memory work loads are similar to RDMA
- Similarities
  - Direct access to ‘remote node memory’
- Not the same
  - message set up
  - message sychronization
- Kernel support is minimal but important
  - Setup/tear down of shared memory
  - Should be manageable from current interfaces

Linux Plumbers Conference

Vienna, Austria | September 18-20, 2024