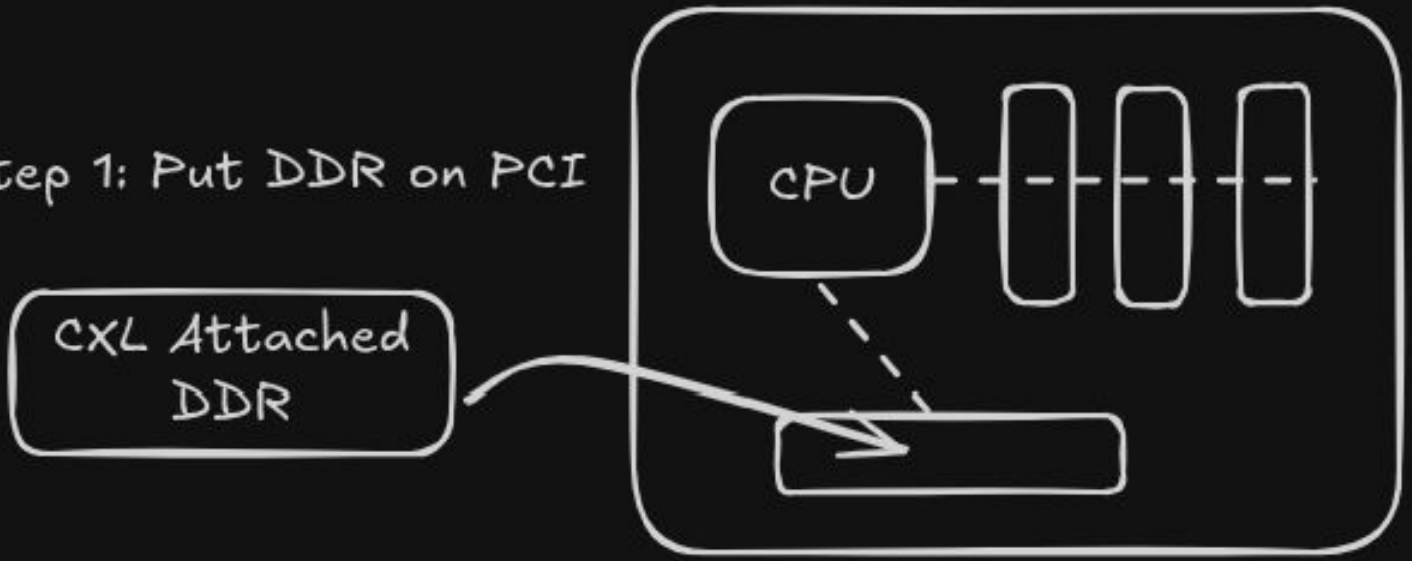
Managing CXL Devices

Use-case analysis

Gregory Price

This discussion is not about tiering.

Step 1: Put DDR on PCI

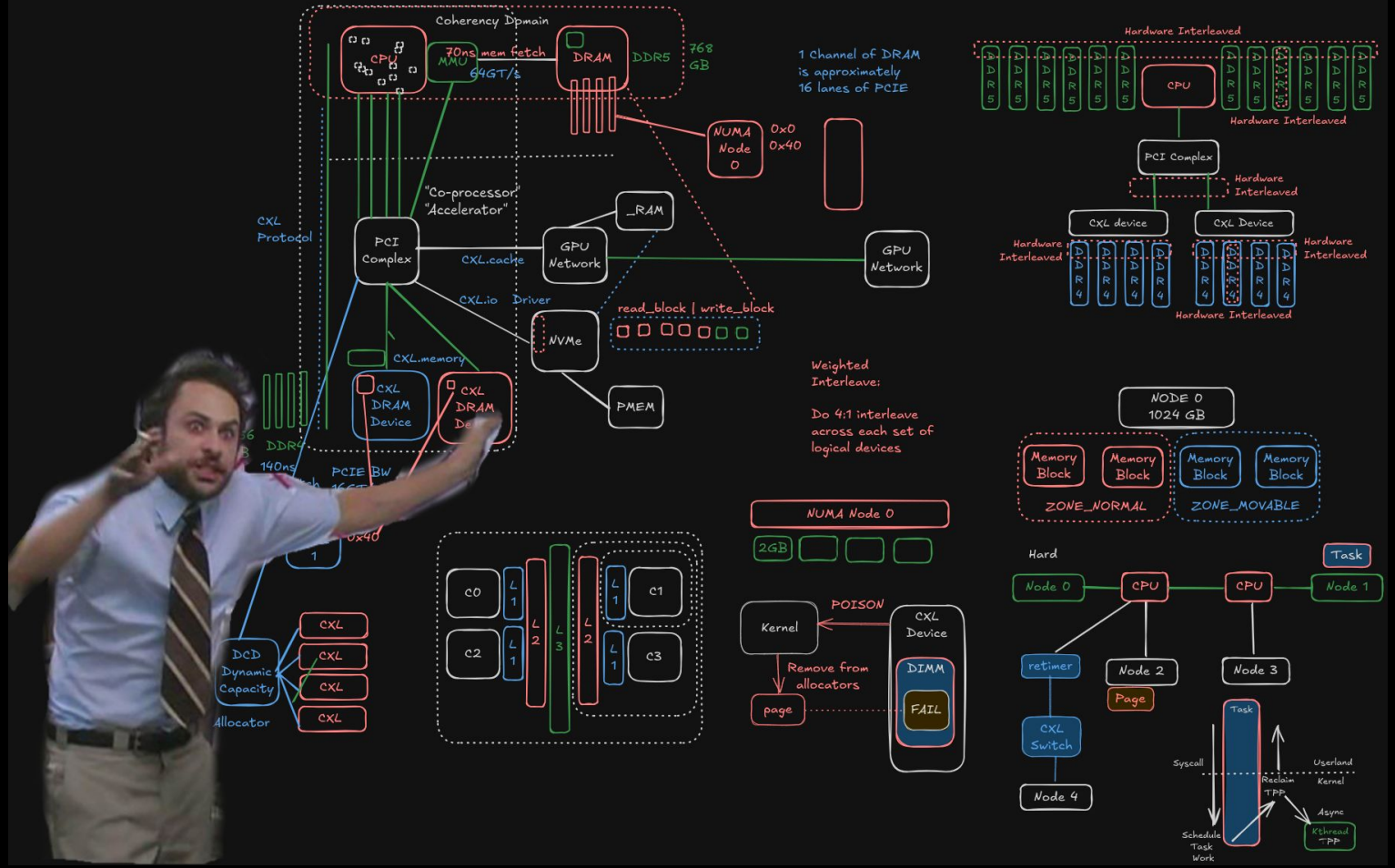


Step 2: Expose to Page Allocator

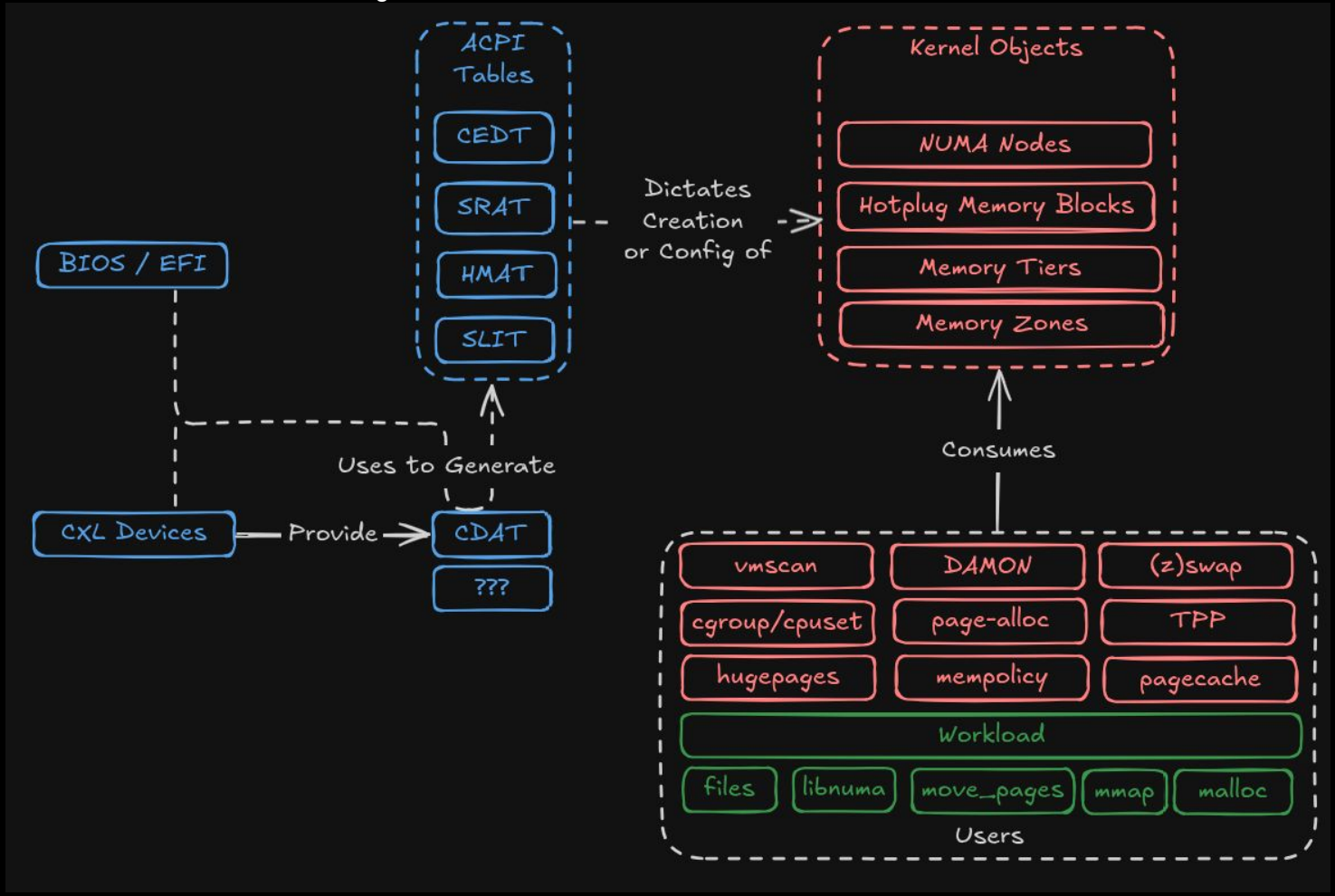
Step 3: ????

Step 4: Profit

A depiction of Gregory teaching people how CXL works



BIOS dictates Linux object creation



Flexible NUMA Topologies

Why: Reasons ™

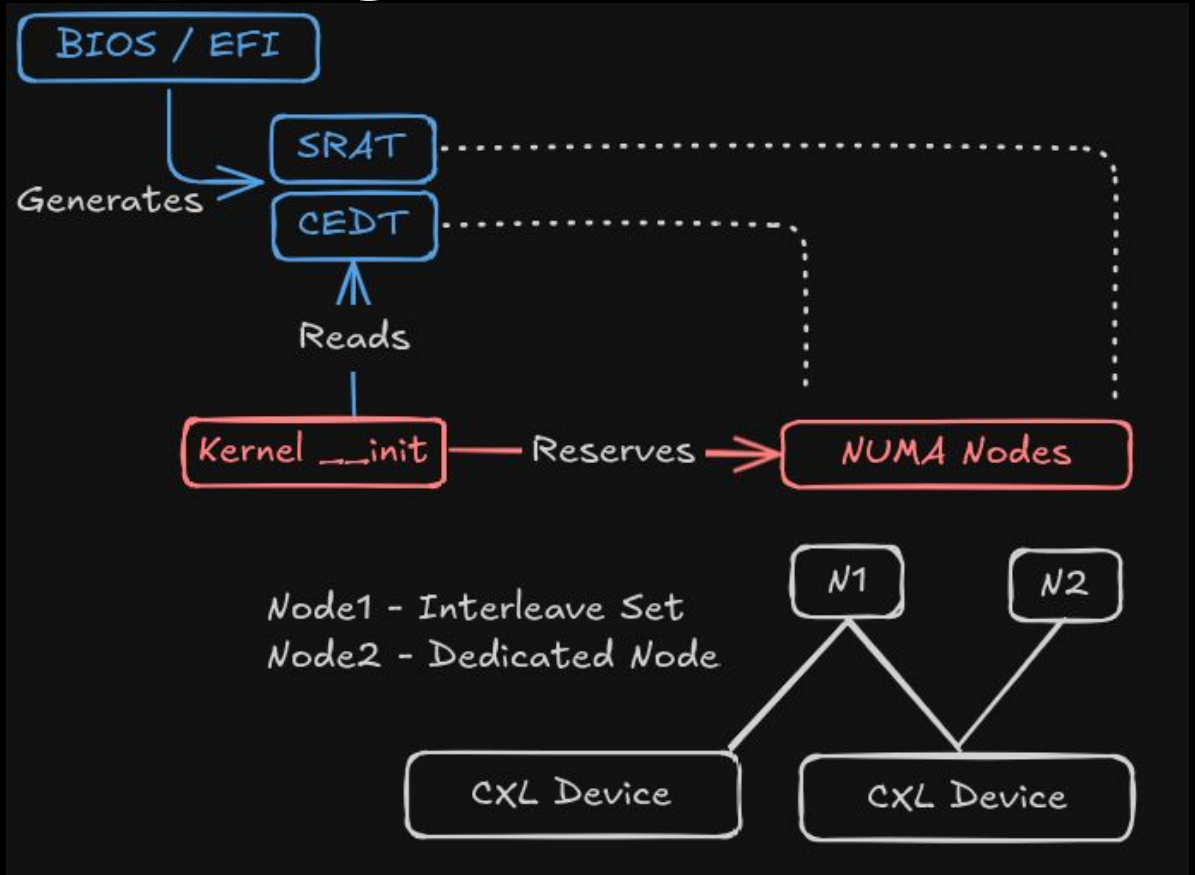
How: ACPI Tables
CXL Decoders

Fun: Multi-Node/Device

Thoughts:

- Known issues, best bet is some modicum of theory of operation docs.

Give platforms something to build toward.



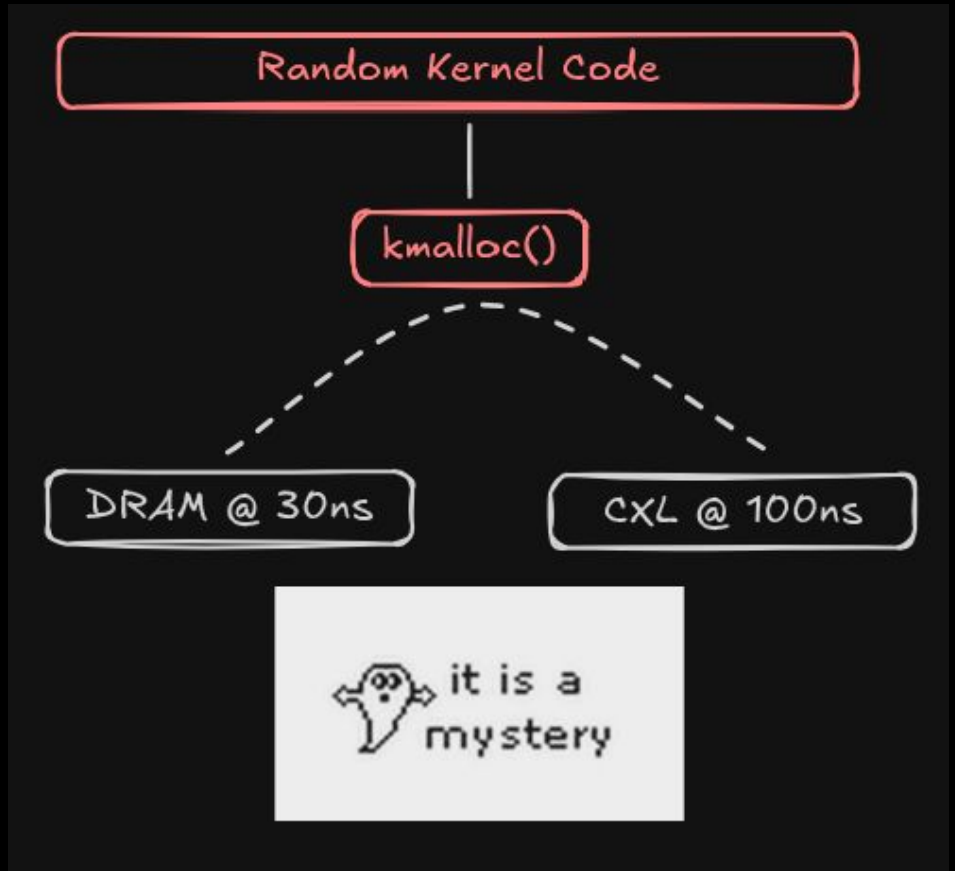
Isolate Kernel from CXL (mostly)

Why: Performance, Reliability

How: EFI_MEMORY_SP + ZONE_MOVABLE

Implications: Memmap Cost
1GB compatibility
Reclaim (Z-N Pressure)

Questions: Different lever?
ZONE_PONIES?

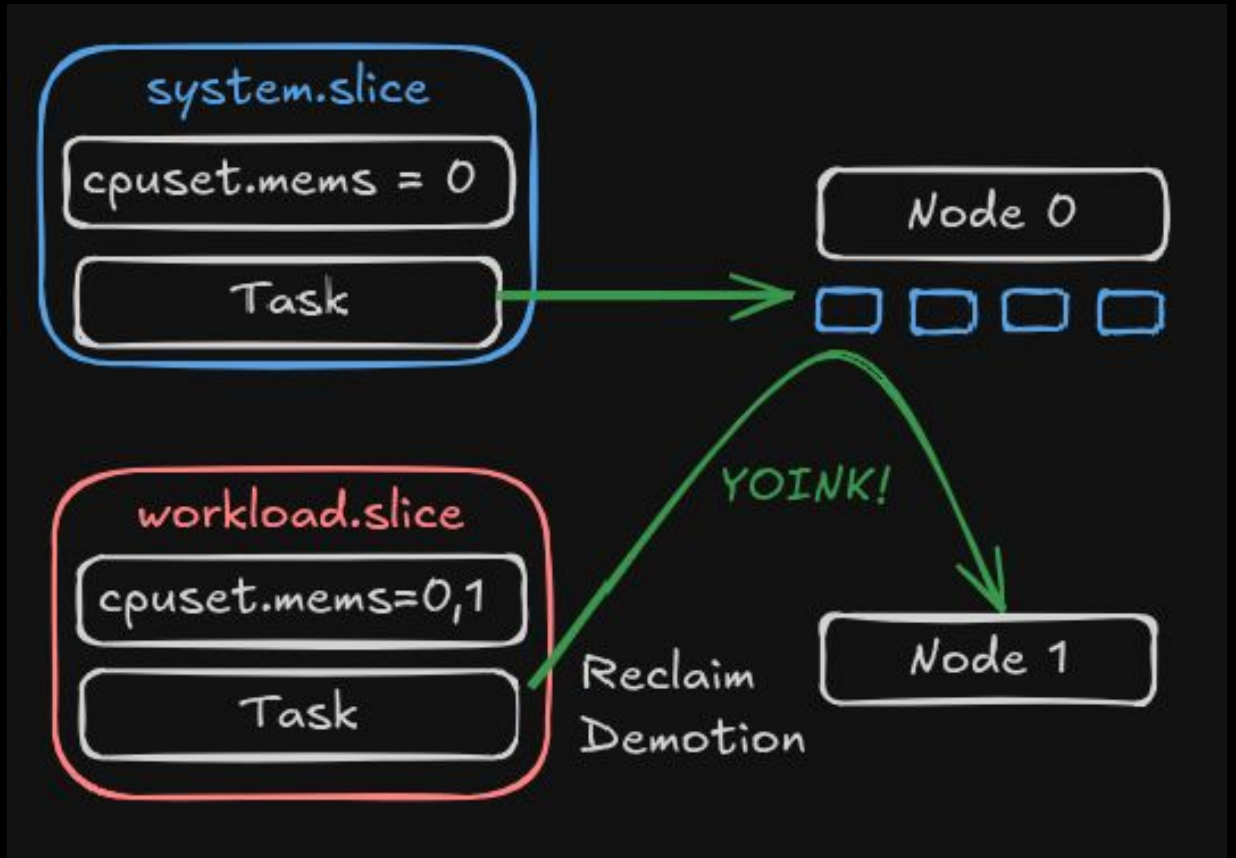


Isolate Workloads from/to CXL

Why: Workload Priority

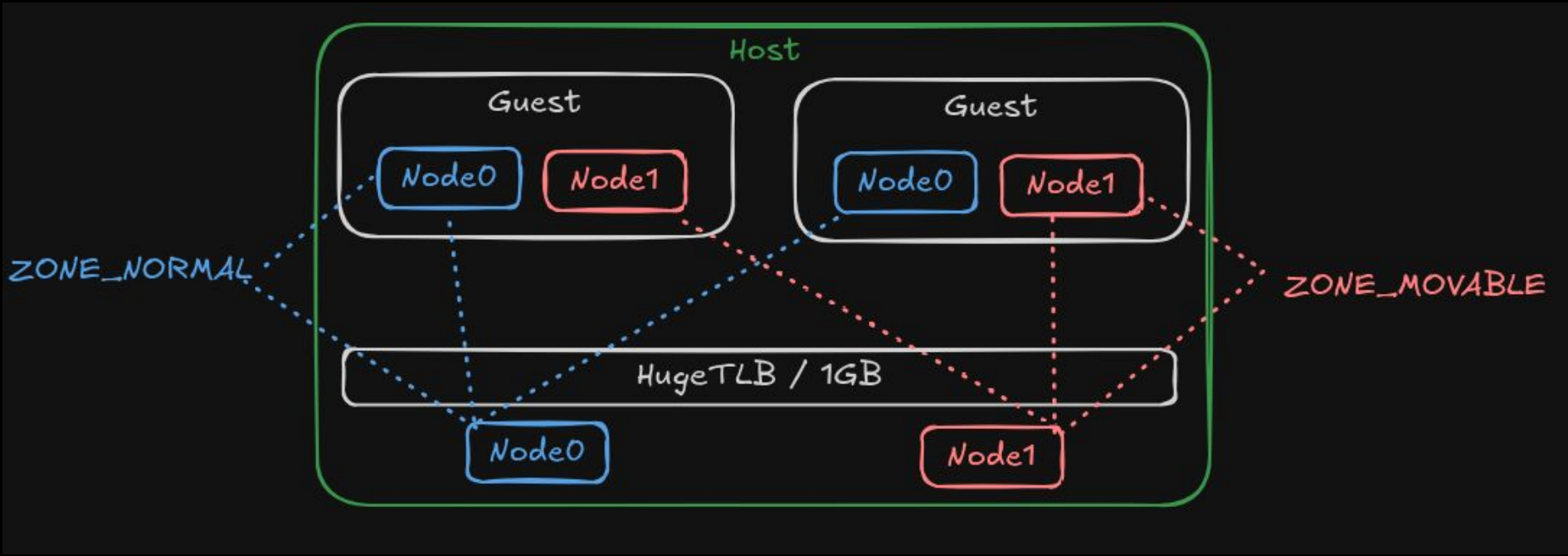
How: cgroup/cpusets

Issues: Demotion (RFC)
Shared pages
Racey



Host/Guest Kernel Parity

Why: Single-Kernel management How: HugeTLB? QEMU Hack? Dax w/ 1GB pg? guest_memfd?

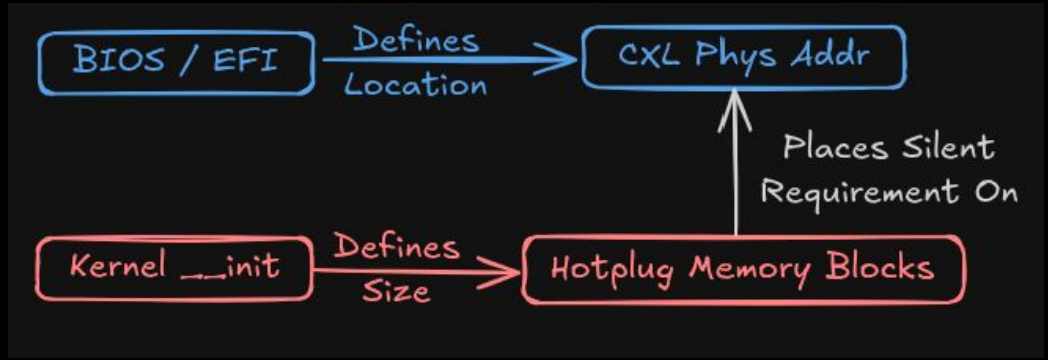


Memory Hotplug

Why: Reasons ™

Subtlety: Block size alignment

Problem: Stranded Capacity
1GB Page compat
Arch-Defined Block Size



Fix: ¯_(ツ)_/¯

- ACPI informed block size (PATCH since November ‘24)
- Variable size blocks?
- Warn? (PATCH)

```
node 1 size: 258048 MB
node 1 free: 254450 MB
```

-1.5% Capacity
4GB / 256 GB

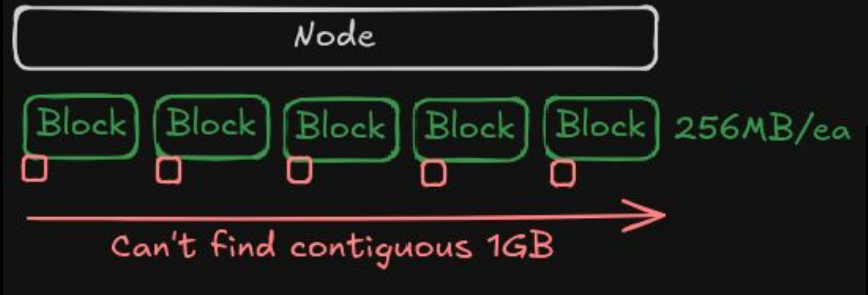
1GB Huge Pages

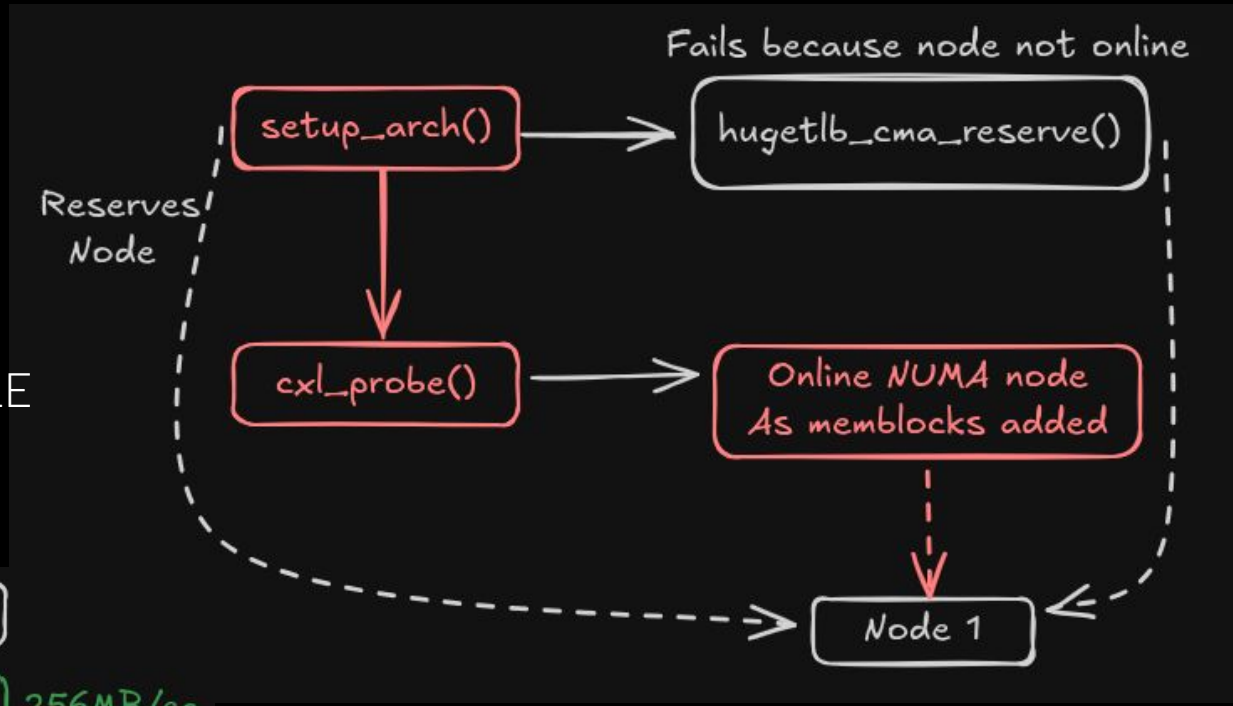
Why: Performance (VMs)

How: CMA / HugeTLB

Issues:
- EFI _MEMORY_SP vs CMA
- HugeTLB vs ZONE_MOVABLE

Memory Block Size Issue

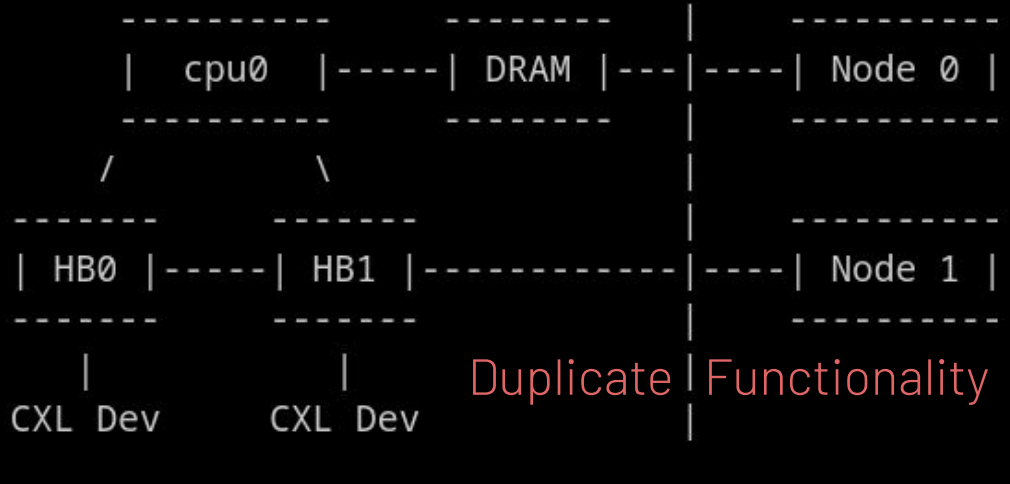




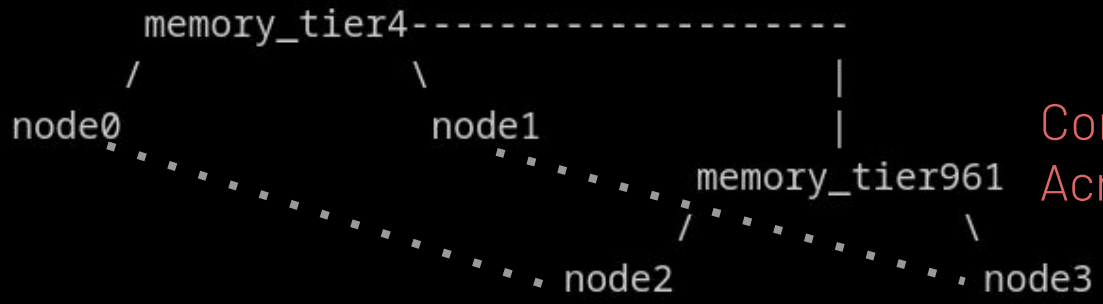
Memory Tiers (memory-tier.c)

Why: Reasons™
What: Logically group nodes
Issues: Cross-socket
H/W Interleave

Question: Rethink?
next_demotion_target(nid)
node_is_toptier(nid)



Duplicate Functionality



Confusing / Actively Harmful
Across Sockets