Linux Plumbers Conference | Richmond, VA | Nov. 13-15, 2023

Intel Low Power Mode Daemon on Hybrid CPUs

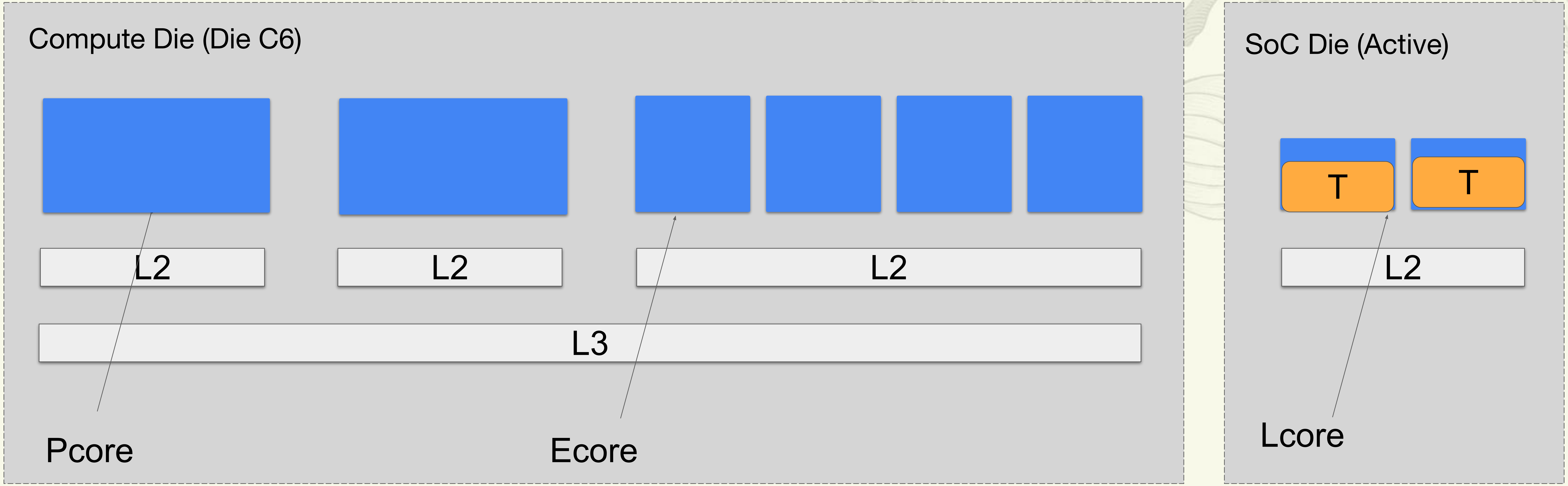
Zhang Rui <rui.zhang@intel.com>

Srinivas Pandruvada <srinivas.pandruvada@linux.intel.com>

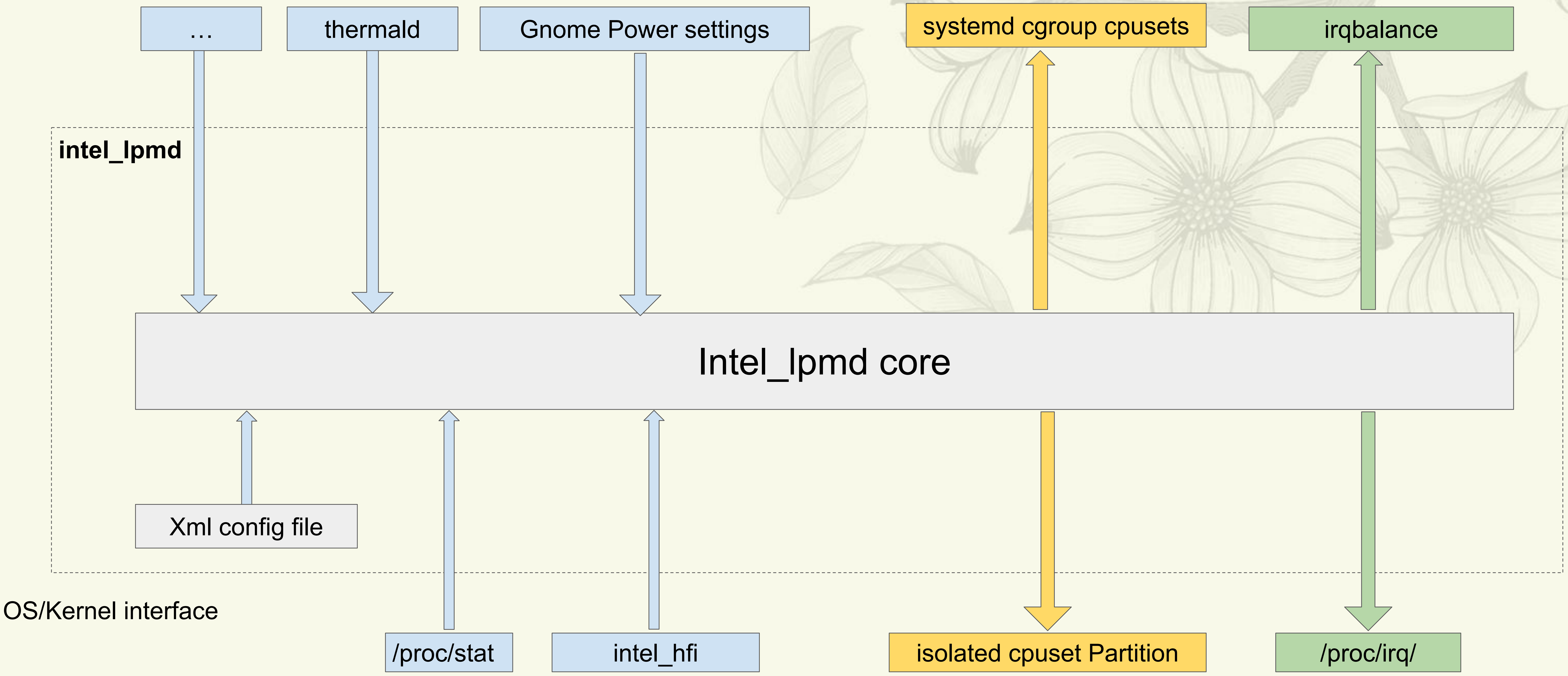
Background

- multiple CPU types within the same processor
- Different CPUs have different power efficiencies
- Power save can be achieved by running tasks on a set of most power efficient CPUs only (**Low Power Mode**)

Low Power Mode on Intel Hybrid CPUs (MTL)



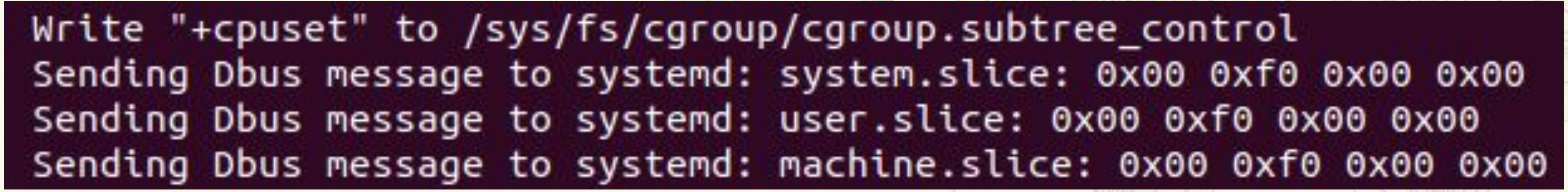
Intel Low Power Mode Daemon Diagram



Github repo: https://github.com/intel/intel-lpmd

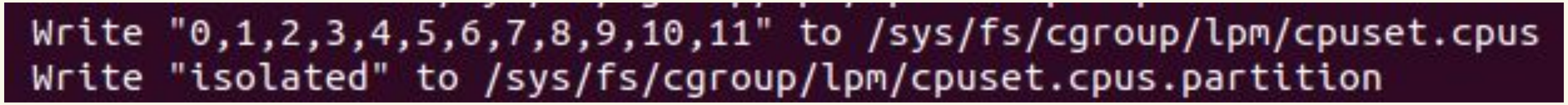
Task placement with cgroup

- cgroup cpuset controller allows restricting tasks to certain cpus
  - system cgroup cpusets



```
Write "+cpuset" to /sys/fs/cgroup/cgroup.subtree_control
Sending Dbus message to systemd: system.slice: 0x00 0xf0 0x00 0x00
Sending Dbus message to systemd: user.slice: 0x00 0xf0 0x00 0x00
Sending Dbus message to systemd: machine.slice: 0x00 0xf0 0x00 0x00
```

  - cgroup isolated partition



```
Write "0,1,2,3,4,5,6,7,8,9,10,11" to /sys/fs/cgroup/lpm/cpuset.cpus
Write "isolated" to /sys/fs/cgroup/lpm/cpuset.cpus.partition
```

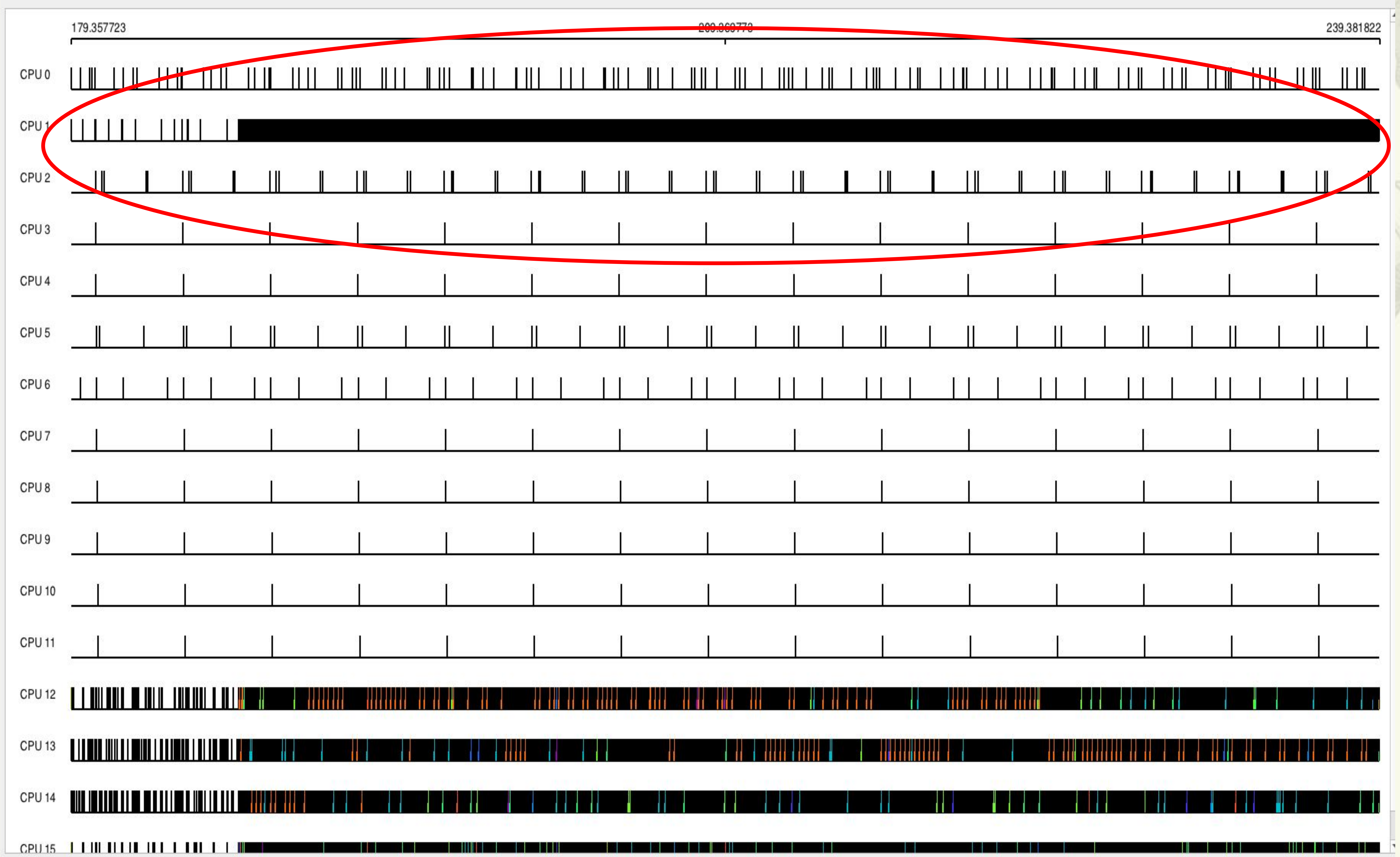
Challenges: Task placement

- Scheduler chooses cpus in isolated partition to do idle load balance
  - [PATCH] sched/fair: Skip cpus with no sched domain attached during NOHZ idle balance
- Unbound workqueues run on CPUs in isolated partition
  - [PATCH v2 0/4] cgroup/cpuset: Improve CPU isolation in isolated partitions
- Timers fire on CPUs in isolated partition
  - [PATCH v8 00/25] timer: Move from a push remote at enqueue to a pull at expiry model
- Hard to detect Lcores on MTL due to missing cache sysfs
  - [PATCH v3 0/3] x86/cacheinfo: Set the number of leaves per CPU

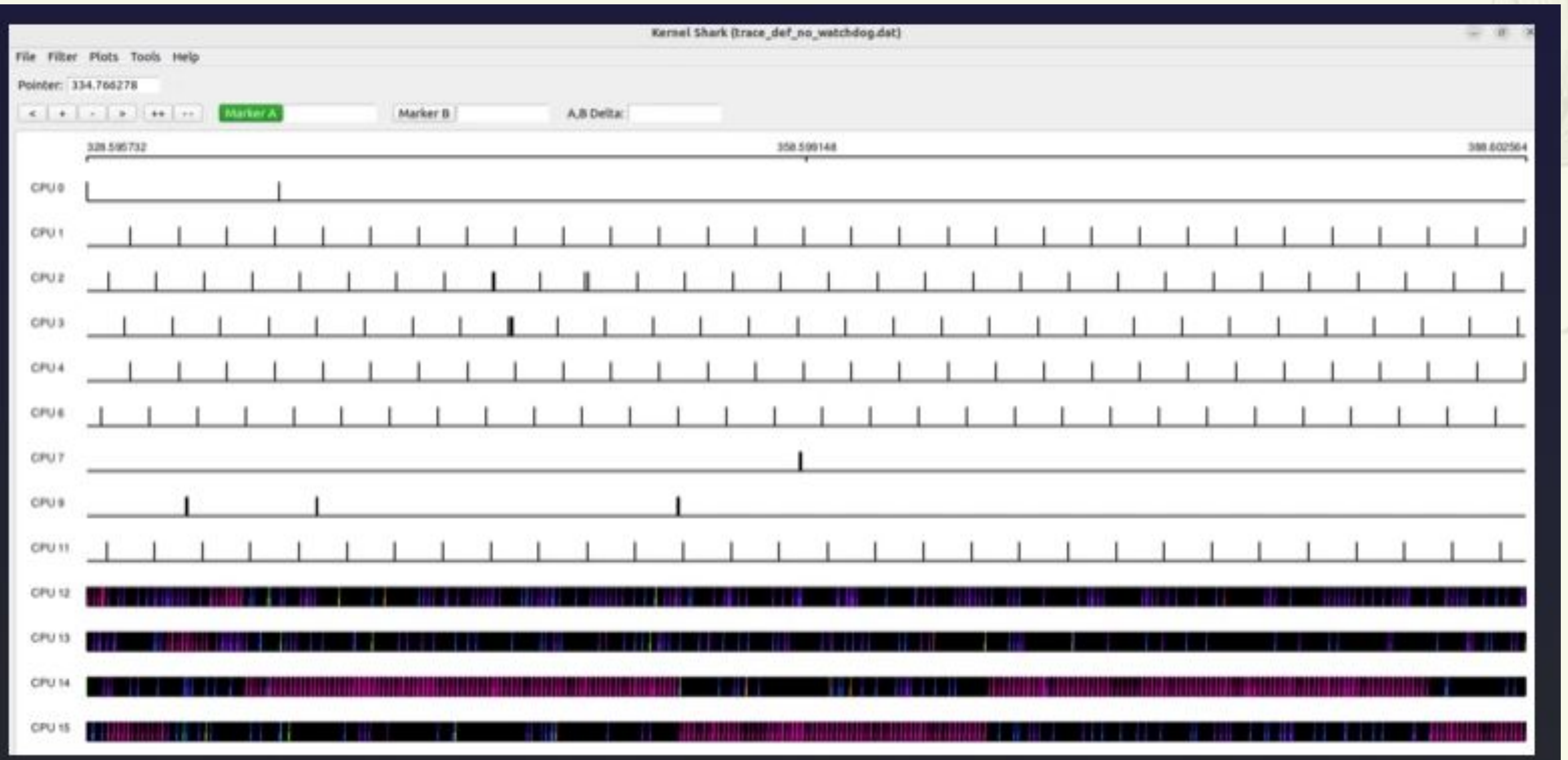
Challenges: Response time

- Need polling to enter/exit Low Power Mode
  - firmware work load type interrupts helps to LPM exit only
  - **Can kernel scheduler provide an event when utilization is low?**
- irq placement (enter/exit Low Power Mode)
  - irqbalance (socket message) takes time to respond
  - or need to handle a large number of irq procfs entries
  - **Can cgroup isolated partition handle irqs?**

Backup: Idle Load balance on isolated partition CPUs



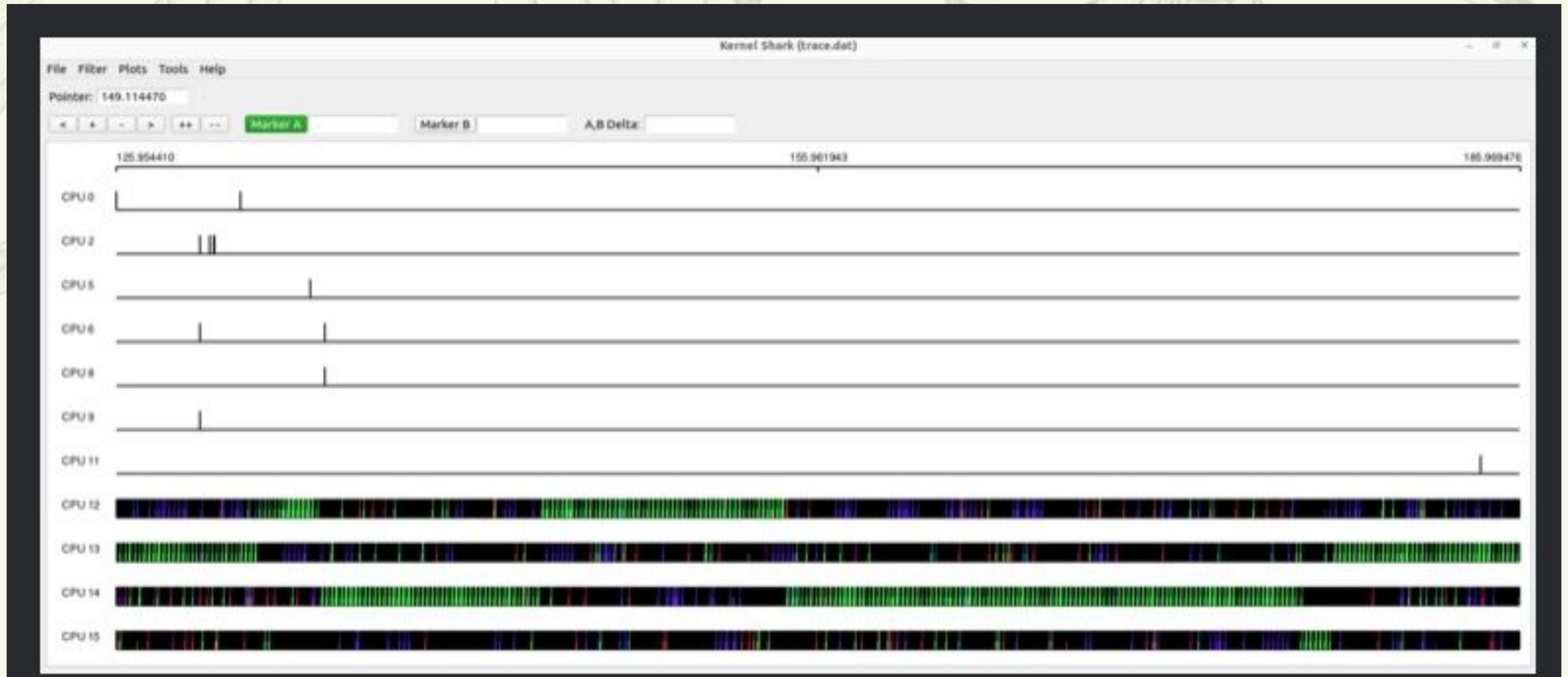
Backup: Timers fired on isolated partition CPUs



Kernel Shark (trace_def_no_watchdog.dat)

Pointer: 334.766278

Without timer patches (sched+timer+irq trace flags)



Kernel Shark (trace.dat)

Pointer: 149.114470

With timer patches (sched+timer+irq trace flags)

Backup: userspace tools could break

- Problem:
  - Userspace tool does not handle CPUs in cgroup isolated partition
    - E.g. turbostat still reads counters on isolated CPUs.
- Solution:
  - /sys/fs/cgroup/cpuset.cpus.effective
- Question:
  - Any other tools that could break?

Linux Plumbers Conference

Richmond, Virginia | November 13-15, 2023