Linux Plumbers Conference

Richmond, Virginia | November 13-15, 2023

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

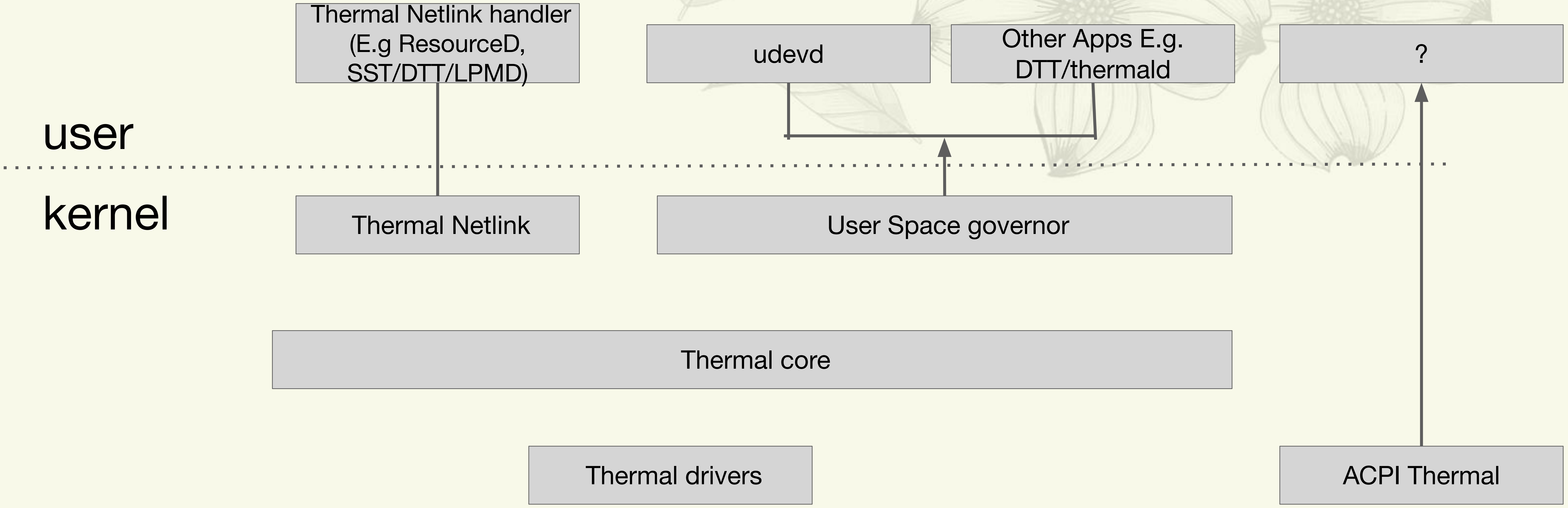
Pitfalls of using Netlink in Thermal Subsystem

Srinivas Pandruvada

Overview

- Netlink
  - User<->kernel communication
  - Socket based
  - Datagram oriented service (SOCK_DGRAM, SOCK_RAW)
  - Unicast/multicast capability
- Netlink families
  - NETLINK_KOBJECT_UEVENT : User space governor
  - NETLINK_GENERIC : Thermal events and samples
  - NETLINK_GENERIC : ACPI Thermal events

Block Diagram on Intel platforms



Requirements for user-kernel Interface

- Low overhead
- Low usage of resources
- Fast enough to mitigate thermals

Issues (userspace gov)

- Freeze user space
  - Each message results in two messages (UDEV KERNEL, UDEV USER)
  - On systems with
    - high number of CPUs with many zones and low swap
    - High traffic with Constant trip change
      - Consumes lots of system memory and CPU time

        User reported 100 MB usage on a system
        https://github.com/intel/thermal_daemon/issues/399
        Udev workers can exhaust system memory
        https://www.suse.com/support/kb/doc/?id=000019156
        Multiple udevd processes causing high load average
        https://access.redhat.com/solutions/457313

- Double reporting with user space governor

Solution

- Rate control of events
  - At firmware level
  - Kernel level: Coalesce events of same type
- Deprecate user space governor
  - Legacy issues
  - Replace with thermal-netlink
    - Not a complete set of events
      - Add additional events to thermal netlink

Issues (Thermal Netlink)

- There is no subscription: Too much noise
    - All handlers gets all events
        - All zones
        - All policies (with user space governor, there is some filter)
    - One multicast group “event”
    - Without consumer, wastes several cycle to multicast
- Not fast enough

Max round trip response Times

KOBJECT UEVENT:300+ us
GENERIC NETLINK: 100+us
Character device: 15+ us

Measured on Tiger Lake system with 4 CPUs ,
2.3GHz base and 4.2GHz Max turbo

Solution (Thermal Netlink)

- Create a filter command
  - Filter of zones
  - Policy filter
  - Multiple clients : Use the last setting
- Check user space presence
- Fine grain event multicast groups
  - Separate for trips and non trips

Improve response time

- Special need to response firmware events (hot trip, keep alive)
  - Introduce one cdev for thermal subsystem
  - Option to callers unicast/multicast

ACPI thermal notifications

- To avoid need for one more subscription, can we deprecate this?

Linux Plumbers Conference

Richmond, Virginia | November 13-15, 2023