Impact of recent CPU topology changes on Android's phantom domains

Android MC - LPC 2022

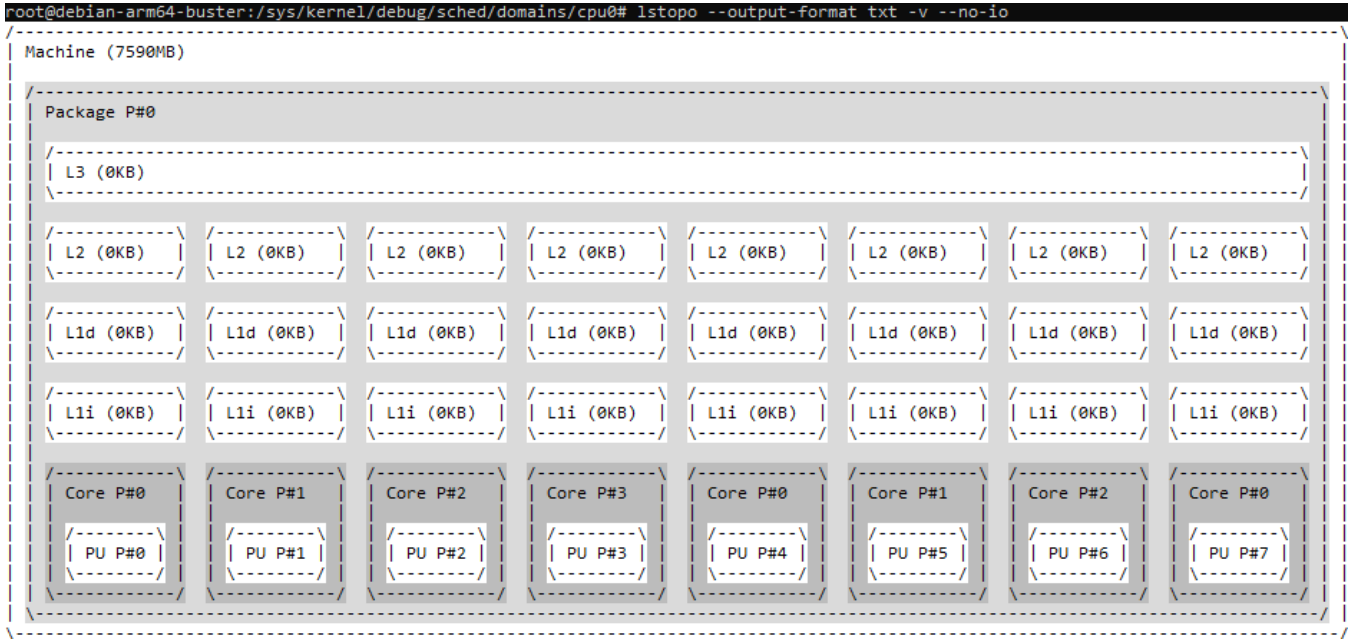
Ionela Voinescu & Dietmar Eggemann

September, 2022

Agenda

- Introduction
- Recent CPU topology changes
- Phantom domains
- Impact on phantom domains users

RB5 - Qualcomm QRB5165 Robotics SoC



Introduction

- Topology information - exposed via sysfs
  - CPU topology (1)
  - Cache topology (2)
- Task scheduler topology (3) – exposed via debugfs
  - Sched domain cpumask functions (arch_topology driver)

    struct cpumask *cpu_smt_mask(cpu): &cpu_topology[cpu].thread_sibling
    struct cpumask *cpu_clustergroup_mask(cpu): &cpu_topology[cpu].cluster_sibling
    struct cpumask *cpu_coregroup_mask(cpu): the smaller of NUMA, core(package), LLC
    struct cpumask *cpu_cpu_mask(cpu): cpumask_of_node(cpu_to_node(cpu))
  - Sched topology table

    { cpu_smt_mask, cpu_smt_flags, SD_INIT_NAME(SMT) },
    { cpu_clustergroup_mask, cpu_cluster_flags, SD_INIT_NAME(CLS) },
    { **cpu_coregroup_mask**, cpu_core_flags, SD_INIT_NAME(**MC**) },
    { cpu_cpu_mask, SD_INIT_NAME(DIE) },

(1) CPU topology

```
/sys/devices/system/cpu/cpu0/topology
cluster_cpus_list:0
cluster_id:-1
core_cpus_list:0
core_id:0
core_siblings_list:0-7
package_cpus_list:0-7
physical_package_id:0
thread_siblings_list:0
```

(2) Cache topology

```
/sys/devices/system/cpu/cpu0/cache
index1/shared_cpu_list:0
index2/shared_cpu_list:0
index3/shared_cpu_list:0-7 - LLC (Last Level $)
```

(3) Task scheduler topology

```
/sys/kernel/debug/sched/domains/cpu0
domain0/busy_factor:16
domain0/cache_nice_tries:1
domain0/flags:SD_BALANCE_NEWIDLE, ...
domain0/imbalance_pct:117
domain0/max_interval:16
domain0/max_newidle_lb_cost:38463
domain0/min_interval:8
domain0/name:MC
```


arm

Recent CPU topology changes

- The `**arch_topology: Updates to add socket support and fix cluster ids**` patch-set (v6.0-rc1) has led to changes in topology information:
  - CPU topology that better describes hardware: the use of socket number versus cluster index as physical package ID
  - Device-tree (DT) CPU topology better aligned with ACPI topology description – improvements towards a consistent view
  - Improved detection of shared caches (e.g., Last Level Cache) for DT systems

ACPI: PPTT: Use table offset as fw_token instead of virtual address
cacheinfo: Use of_cpu_device_node_get instead cpu_dev->of_node
cacheinfo: Add helper to access any cache index for a given CPU
cacheinfo: Move cache_leaves_are_shared out of CONFIG_OF
cacheinfo: Add support to check if last level cache(LLC) is valid or shared
cacheinfo: Allow early detection and population of cache attributes
cacheinfo: Use cache identifiers to check if the caches are shared if available
cacheinfo: Align checks in cache_shared_cpu_map_{setup,remove} for readability
arch_topology: Add support to parse and detect cache attributes
arch_topology: Use the last level cache information from the cacheinfo
arm64: topology: Remove redundant setting of llc_id in CPU topology
arch_topology: Drop LLC identifier stash from the CPU topology
arch_topology: Set thread sibling cpumask only within the cluster
arch_topology: Check for non-negative value rather than -1 for IDs validity
arch_topology: Avoid parsing through all the CPUs once a outlier CPU is found
arch_topology: Don't set cluster identifier as physical package identifier
arch_topology: Set cluster identifier in each core/thread from /cpu-map
arch_topology: Add support for parsing sockets in /cpu-map
arch_topology: Warn that topology for nested clusters is not supported
ACPI: Remove the unused find_acpi_cpu_cache_topology()
cacheinfo: Use atomic allocation for percpu cache attributes
ACPI: PPTT: Leave the table mapped for the runtime usage
arch_topology: Fix cache attributes detection in the CPU hotplug path

https://lore.kernel.org/lkml/20220704101605.1318280-1-sudeep.holla@arm.com/
https://lore.kernel.org/lkml/20220720-arch_topo_fixes-v3-0-43d696288e84@arm.com/

arm

Phantom domains

```
cpu-map {
        cluster0 {
                core0 {
                        cpu = <&CPU0>;
                };
                core1 {
                        cpu = <&CPU1>;
                };
                core2 {
                        cpu = <&CPU2>;
                };
                core3 {
                        cpu = <&CPU3>;
                };
        };
        cluster1 {
                core0 {
                        cpu = <&CPU4>;
                };
                core1 {
                        cpu = <&CPU5>;
                };
                core2 {
                        cpu = <&CPU6>;
                };
        };
        cluster2 {
                core0 {
                        cpu = <&CPU7>;
                };
        };
};
```

*DynamIQ system **w/** phantom domains*

- **Phantom domains**: device tree clusters used to group CPUs of the same micro-architecture.
- When old style, out-of-tree Energy Model (EM) was used for DynamIQ systems, Android topologies were represented with phantom domains to mimic classical big.LITTLE.
- After moving to a simplified EM (mainline) that is no longer attached to the sched domain hierarchy, and with DynamIQ support added [1], **phantom domains are no longer needed**.

```
cpu-map {
        cluster0 {
                core0 {
                        cpu = <&CPU0>;
                };
                core1 {
                        cpu = <&CPU1>;
                };
                core2 {
                        cpu = <&CPU2>;
                };
                core3 {
                        cpu = <&CPU3>;
                };
                core4 {
                        cpu = <&CPU4>;
                };
                core5 {
                        cpu = <&CPU5>;
                };
                core6 {
                        cpu = <&CPU6>;
                };
                core7 {
                        cpu = <&CPU7>;
                };
        };
};
```

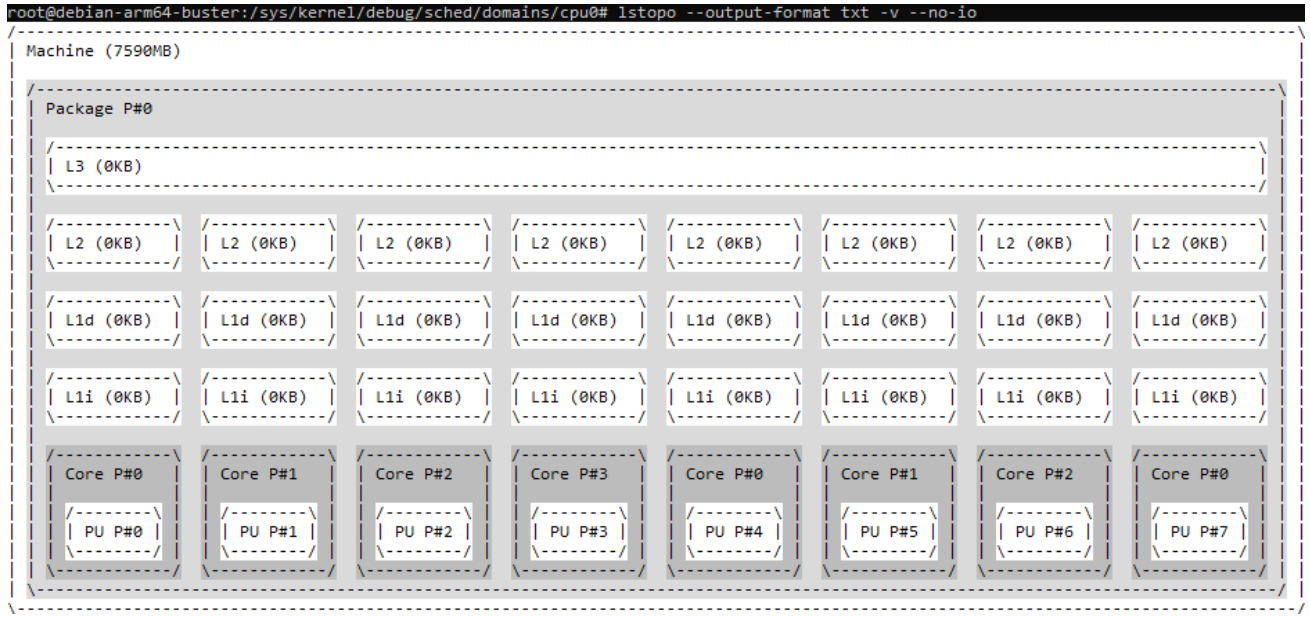
*DynamIQ system **w/o** phantom domains*

[1] https://lore.kernel.org/lkml/20200206191957.12325-2-valentin.schneider@arm.com (v5.5)

Impact on userspace

package_cpus_list:0-3

Topology with phantom domains presented CPUs of the same micro-arch as belonging to different packages.



package_cpus_list:0-7

Even if DT presents phantom domains, the topology parsing code shows those as clusters, not packages, as if there were no phantom domains present.

Summary: sysfs files **package_cpus** and **package_cpus_list** (and deprecated equivalents core_siblings and core_siblings_list) can no longer be used to obtain same micro-arch CPU groups !!!

Impact on task scheduler

- No impact on Energy-Aware Scheduling (EAS)
- Completely Fair Scheduler (CFS) load-balance between all CPUs happens more often
- Android customization: Any functionality behind vendor hooks relying on phantom domains will no longer work as expected !!!

Summary: After the recent DT topology changes, the use of phantom domains will result in scheduler topologies akin to a non-phantom domains setup. These changes lead to correct hardware description and are consistent with the recommendation for phantom domains to be removed.

arm

Thank You
Danke
Gracias
Grazie
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה