Contribution ID: **301** Type: **not specified**

Scheduler domains and cache bandwidth

Monday, 9 September 2019 17:45 (15 minutes)

The Linux Kernel scheduler represents a system's topology by the means of scheduler domains. In the common case, these domains map to the cache topology of the system.

The Cavium ThunderX is an ARMv8-A 2-node NUMA system, each node containing 48 CPUs (no hyperthreading). Each CPU has its own L1 cache, and CPUs within the same node will share a same L2 cache.

Running some memory-intensive tasks on this system shows that, within a given NUMA node, there are "socklets" of CPUs. Executing those tasks (which involve the L2 cache) on CPUs of the same "socklet" leads to a reduction of per-task memory bandwidth.
On the other hand, running those same tasks on CPUs of different "socklets" (but still within the same node) does not lead to such a memory bandwidth reduction.

While not truly equivalent to sub-NUMA clustering, such a system could benefit from a more fragmented scheduler domain representation, i.e. grouping these "socklets" in different domains.

This talk will be an opportunity to discuss ways for the scheduler to leverage this topology characteristic and potentially change the way scheduler domains are built.

I agree to abide by the anti-harassment policy

Yes

I confirm that I am already registered for LPC 2019

Primary author: SCHNEIDER, Valentin (Arm Ltd)

Presenter: SCHNEIDER, Valentin (Arm Ltd)

Session Classification: Scheduler MC