Contribution ID: **113** Type: **not specified**

Taming the Chiplet: High Performance CCX scheduling via BPF

Friday 12 December 2025 17:36 (18 minutes)

Thread placement on machines with complex cache hierarchies (such as AMD CPU Core Complexes (CCX'es)) requires careful management for optimal performance. Unlike NUMA domains, which are large enough that hard partitioning is a viable strategy, these chiplet domains are too small to schedule efficiently without a means of enforcing some degree of **soft affinity**. Spillover of threads to remote CCXs should be a managed exception, permitted only under specific load conditions, and the resulting spread must be minimized and carefully managed.
This presentation will detail an application of extensible scheduling that leverages **BPF** and a **userspace agent** to achieve this fine-grained control. Our policy implements a **per-CCX runqueue** and utilizes an asynchronous **bin-packing algorithm** to dynamically assign and manage the soft affinity of thread groups. Additionally, the scheduler also employs heuristics to intelligently decide when and where threads should spill out from their preferred CCX during load spikes. While the concepts here are largely generic to any workload, we will primarily consider how it intersects with VM scheduling, which is our current application. We will demonstrate how the userspace component enables sophisticated policy management through tunable parameters and complex accounting. The architecture ensures **fairness** among competing threads while maintaining high **throughput** and **cache locality**. We will discuss the design, the BPF-userspace interaction, and the performance benefits of this approach.

Finally, we wish to open a discussion with the community on best practices for tuning such policies and to explore potential improvements to our design.

Primary author: GATTANI, Aniket (Google)

Co-author: DON, Josh (Google)

Presenters: GATTANI, Aniket (Google); DON, Josh (Google)

Session Classification: sched_ext: The BPF extensible scheduler class MC

Track Classification: sched_ext: The BPF extensible scheduler class MC