Contribution ID: **119** Type: **not specified**

Paravirt-Scheduling: Limit CPU resources using dynamic capacity based on the steal time

Wednesday, 18 September 2024 15:22 (22 minutes)

CPU capacity is a software construct to reflect underlying physical CPU capacity. Load balancer uses the CPU capacity to choose an optimal CPU for performance and energy efficiency. CPU Capacity can be affected by frequency, higher level sched classes, guest preemption etc. Steal time is an indicator of guest preemption by the host hypervisor. Current Linux scheduler, updates the CPU capacity based on the steal time.

In IBM powerpc, PowerVM hypervisor schedules at the SMT8 core level, but not at individual thread level and steal time is uniform across the cores. In an overcommitted and overutilized shared environment such as multiple Shared Processor Logical PARrtitions (SPLPAR) it would be optimal to schedule the tasks on limited set of cores instead of spreading across all the cores. Number of cores to schedule can be derived from the percentage of steal time. If the steal time is more, reduce the number of cores which have high capacity and vice versa.

In this talk, we would like to talk about, why current mechanism of updating CPU capacity doesn't work in the above use case and why we need a different way of updating the CPU capacities by steal time in the paravirtualized environment for effective usage of CPU resources among the guests. We would discuss advantages and disadvantages of different approaches considered such as cgroup cpuset, cpu offline etc. We would discuss the issues present currently, when capacity values are very far reaching such as 1024 vs 1.

Primary authors: HEGDE, Shrikanth; Mr DRONAMRAJU, Srikar

Presenter: HEGDE, Shrikanth

Session Classification: Sched MC