


Slide 1



intel[®]
experience
what's inside™

Core scheduling

Agata Gruza

Disclaimer

Intel provides these materials as-is, with no express or implied warranties.

All products, dates, and figures specified are preliminary, based on current expectations, and are subject to change without notice.

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at <http://intel.com>.

Some results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results are based on testing as of October 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

*Other names and brands may be claimed as the property of others.

© Intel Corporation

Takeaways – core scheduler

- Core scheduling performs better than turning off HT in all overcommitting scenarios. In certain cases up to 20% performance drop.
- Impact of core scheduling depends on the workload and thread scheduling intensity.
- Core scheduling requires cgroups. Single cgroup per VM. Each VM should run on its own independent cgroup.

Overview

Motivation: core scheduling is required to protect against leakage of sensitive data allocated on a sibling thread. We want to measure performance impact of core scheduling across different workloads.

Patch description: Patch changes core scheduling in a way that doesn't allow two processes from different cgroups to be executed on a sibling thread.

Experiments: A fixed configuration running a benchmark toggling the following: HT ON/HT OFF, default kernel/core-sched (v3), .5 overcommit/1 overcommit/2 overcommit. Each VM has its own cgroup.

Overcommitting: The ratio of total number of virtual CPUs in VM to CPU threads.

- .5 overcommit: number of vCPUs = half of the number of CPU threads
- 1 overcommit: number of vCPUs = number of CPU threads
- 2 overcommit: number of vCPUs = twice the number of CPU threads

Below you will find data for core scheduling. Data presented here are based on previous version of core-sched (v3) plus additional kernel patches added by tim.c.chen@intel.com and load balancer made by aubrey.li@intel.com that are now in (v4) core scheduler kernel <https://github.com/digitalocean/linux-coresched/commits/coresched/v4-v5.4.y>

Experiment: Multiple VMs (2 in most cases) bound to a socket, taskset for binding processes to CPU for [over/under] committing purposes and to reduce run-to-run variance, cgroups for isolating CPUs.

Slide 5

System setup

SPECint_rate_base 2017
(1 copy for HT ON and HT OFF)
core-sched-v4_rc6

Platform	NUC6i7KYB
Number of Sockets	1
Processors	Intel(R) Core(TM) i7-6770HQ CPU
Number of Cores per Socket	4
Last Level Cache	6 MB
Processor Base Frequency	2.60 GHz
Memory	8 GB (2 x 4 GB DDR4 @2400)
Memory speed achieved	2400 MT/s
OS Distribution	Ubuntu 18.04
Host kernel	5.1.5-1-core-sched-v4_rc6
Guest kernel	kernel-5.1.5
BIOS version	KYSLi70.86A.0061.2019.0222.1748
Microcode version	0xCC
Storage	1 x 240 GB SSD
Networking	1Gbe NIC
Benchmark version	SPECint_rate_base 2017

Measurements provided for research purposes (on Intel internal reference platforms)

5

<http://mark.intel.com/products/93341/Intel-Core-i7-6770HQ-Processor-6M-Cache-up-to-3-50-GHz-SKX>

runcpu --config=core-sched.cfg --size=refrate --copies=1 --noreportable --iterations=3 intrate
1 copy (default value) - 3 iterations

Slide 6

SPECint_rate_base 2017
core_sched_v4_rc6



Core scheduling performs the same (within up to 0.5% run-to-run variance) as turning off HT in all overcommitting scenarios.

System setup

Cassandra
core-sched-v4-rc8

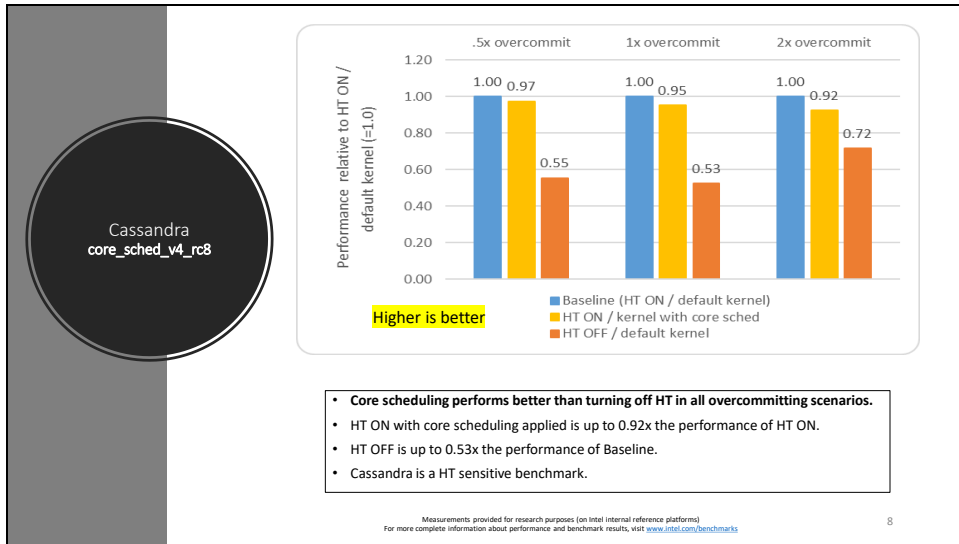
Platform	Purley
Number of Sockets	2
Processors	Intel(R) Xeon(R) Platinum 8173M CPU
Number of Cores per Socket	28
Last Level Cache	38.5 MB
Processor Base Frequency	2.00 GHz
Memory	384 GB (12 x 32 GB DDR4 @2666)
Memory speed achieved	2666 MHz
OS Distribution	RHEL 7.5 GA
Host kernel	core_sched_v4-rc8
Guest kernel	core_sched_v4-rc8
BIOS version	PLYXCRB1.HON.0580.D04.1904301428
Microcode version	0x2000064
Storage	4x1.8TB P4500 for VMs 1x380G SSD for OS
Networking	1Gbe NIC
Benchmark version	Cassandra 4.0

Note: Further slides will show different (2nd) Cassandra experiment (different kernel) but same system set up.

Measurements provided for research purposes (on Intel internal reference platforms)

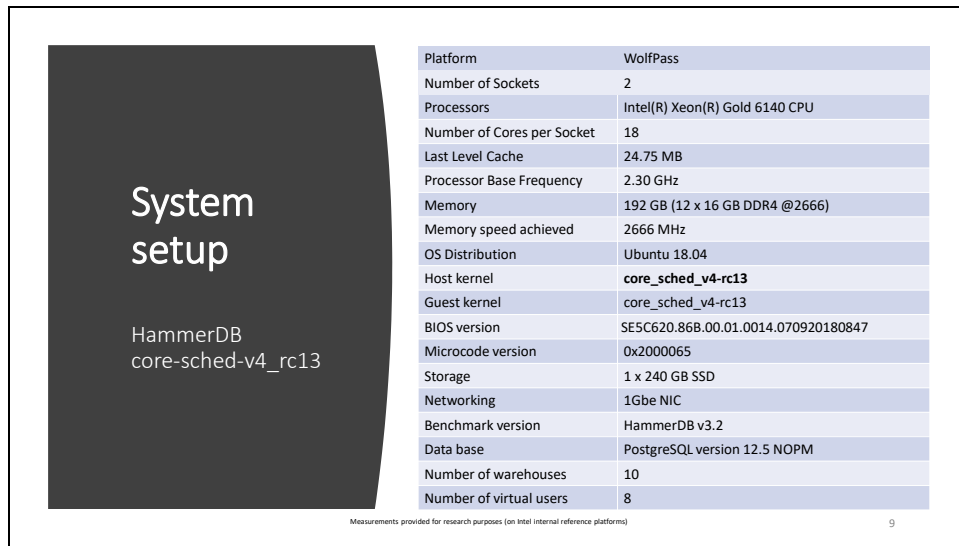
7

Slide 8



When applying core scheduling we are slower by up to 7.63%.
With HT OFF we are slower than with HT ON, by up to 47.47%.

Slide 9



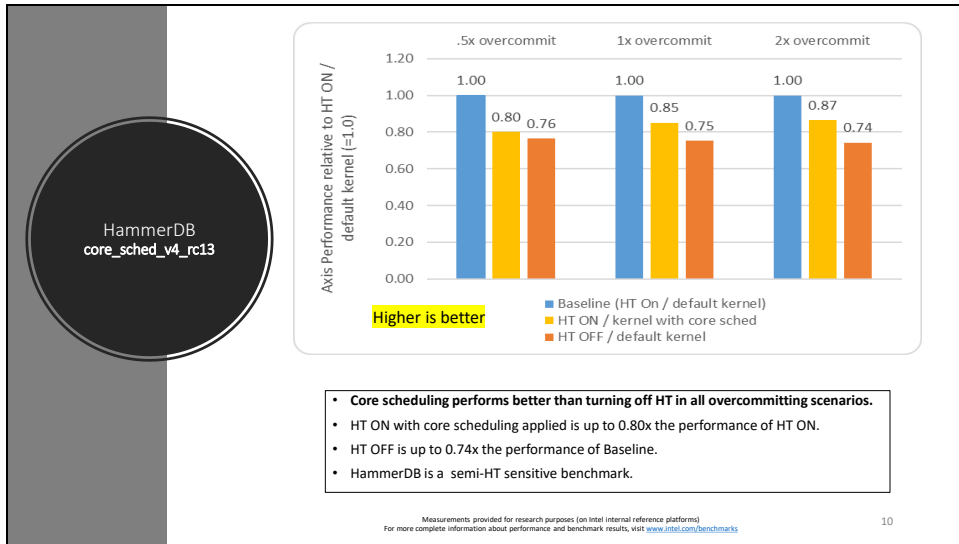
System setup
HammerDB
core-sched-v4_rc13

Platform	WolfPass
Number of Sockets	2
Processors	Intel(R) Xeon(R) Gold 6140 CPU
Number of Cores per Socket	18
Last Level Cache	24.75 MB
Processor Base Frequency	2.30 GHz
Memory	192 GB (12 x 16 GB DDR4 @2666)
Memory speed achieved	2666 MHz
OS Distribution	Ubuntu 18.04
Host kernel	core_sched_v4-rc13
Guest kernel	core_sched_v4-rc13
BIOS version	SE5C620.86B.00.01.0014.070920180847
Microcode version	0x2000065
Storage	1 x 240 GB SSD
Networking	1Gbe NIC
Benchmark version	HammerDB v3.2
Data base	PostgreSQL version 12.5 NOPM
Number of warehouses	10
Number of virtual users	8

Measurements provided for research purposes (on Intel internal reference platforms)

9

<http://mark.intel.com/products/120485/Intel-Xeon-Gold-6140-Processor-24-75M-Cache-2-30-GHz-10-warehouses>
10 warehouses
8 virtual users
DB: **PostgreSQL version 12.5**
System info based on system info for 5.0.0-rc7-4.peterz-sched-core-scheduling-4
The performance metc is NOPM



When applying core scheduling we are slower by up to 19.99%.
With HT OFF we are slower by up to 25.71%.

SPECvirt Core Scheduler_v4_rc15

- SPECvirt workloads:

VMS	vCPU
App Server	4
DB Server	14
Infra Server	1
Web Server	4
Mail Server	1
Batch Server	1

- Tile topology:

- 1 tile contains:
 - 1 appserver
 - 1 webservice
 - 1 infraserver
 - 1 mailservice
 - 1 batchserver
- 4 tiles share 1 DB

core_sched_v4-rc15
with SPECvirt

- Overall score is the main metric used for SpecVirt performance (the higher score, the better performance). It's commonly used in conjunction with a total # of VMs.
- Single cgroup per VM. Each VM is running on it's own independent cgroup.
- Each tile has a unique number of VMs.
- Tiles # vary (from 10 to 14 tiles) based on an experiment.
- Overcommitting:
74VMs = 210 vCPU overcommit $210/112=1.875$
63VMs = 174 vCPU overcommit $174/112=1.554$
53VMs = 152 vCPU overcommit $152/112=1.357$

Measurements provided for research purposes (on intel internal reference platforms)

12

VMWare has 14 tiles on SKX

53 VMs → 10 tiles

63 VMs → 12 tiles

74 VMs → 14 tiles

System setup

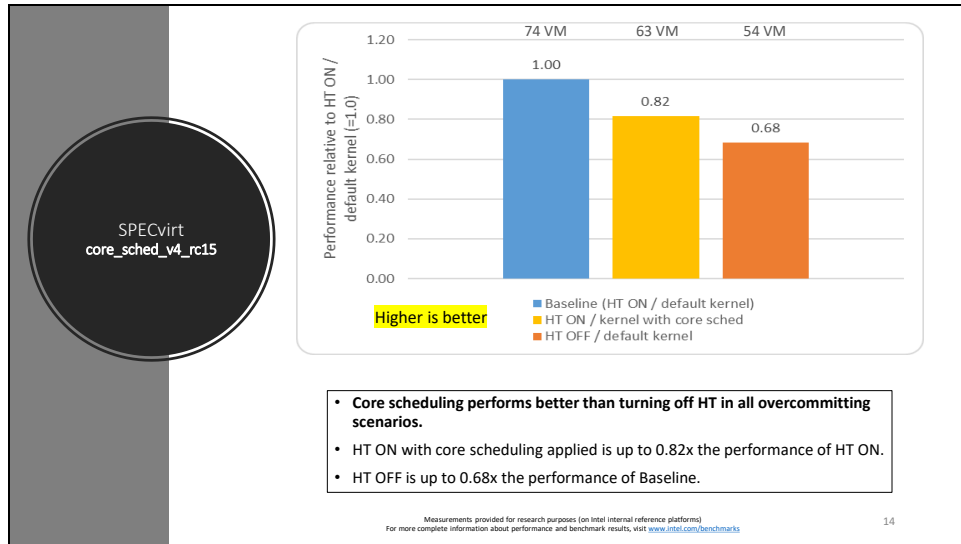
SPECvirt
core_sched_v4-rc15

Platform	Wolfpass Cascade 8280
Number of Sockets	2
Processors	CLX B1 8280
Number of Cores per Socket	28
Last Level Cache	38.5 MB
Processor Base Frequency	2.70 GHz
Memory	768 GB (24 x 32 GB DDR4 @2666)
Memory speed achieved	2666 MTs
OS Distribution	RHEL 8.0 GA
Host kernel	core_sched_v4-rc15
Guest kernel	kernel-5.1.5
BIOS version	SE5C620.86B.0D.01.438.041620190736
Microcode version	0x5e
Storage	3x4TB P4500 for VMs 1x400G SSD for OS
Networking	2x 82599 dual port Ethernet
Benchmark version	SPECvirt_sc2013 version V1.1

Measurements provided for research purposes (on Intel internal reference platforms)

13

<http://mark.intel.com/products/192478/Intel-Xeon-Platinum-8280-Processor-38-5M-Cache-2-70-GHz->



When applying core scheduling we are slower by up to 18.00%.
With HT OFF we are slower by up to 32%.

System setup

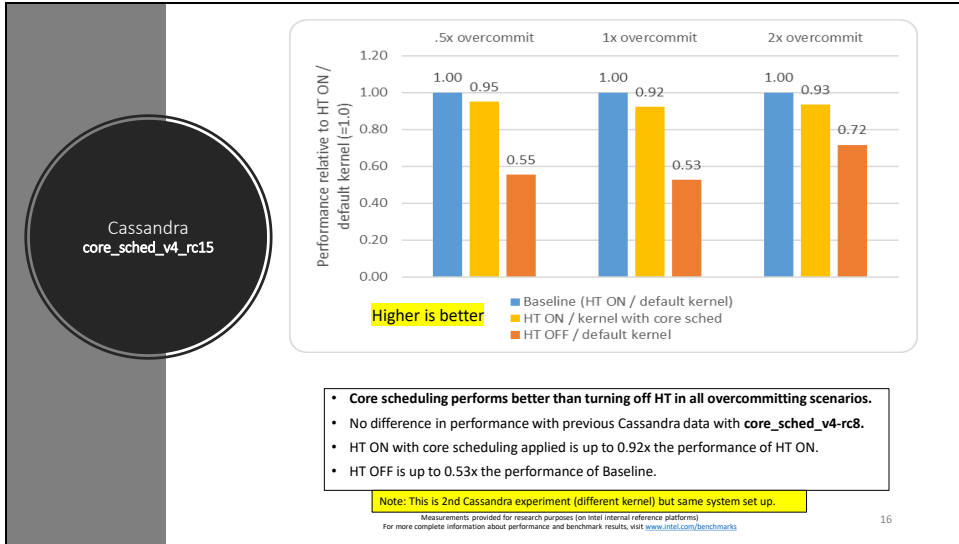
Cassandra
core-sched-v4-rc15

Platform	Purley
Number of Sockets	2
Processors	Intel(R) Xeon(R) Platinum 8173M CPU
Number of Cores per Socket	28
Last Level Cache	38.5 MB
Processor Base Frequency	2.00 GHz
Memory	384 GB (12 x 32 GB DDR4 @2666)
Memory speed achieved	2666 MHz
OS Distribution	RHEL 7.5 GA
Host kernel	core_sched_v4-rc15
Guest kernel	core_sched_v4-rc15
BIOS version	PLYXCRB1.HON.0580.D04.1904301428
Microcode version	0x2000064
Storage	4x1.8TB P4500 for VMs 1x380G SSD for OS
Networking	1Gbe NIC
Benchmark version	Cassandra 4.0

Note: This is 2nd Cassandra experiment (different kernel) but same system set up.

Measurements provided for research purposes (on Intel internal reference platforms)

15



When applying core scheduling we are slower by up to 7.9%.
With HT OFF we are slower by up to 47.47%.