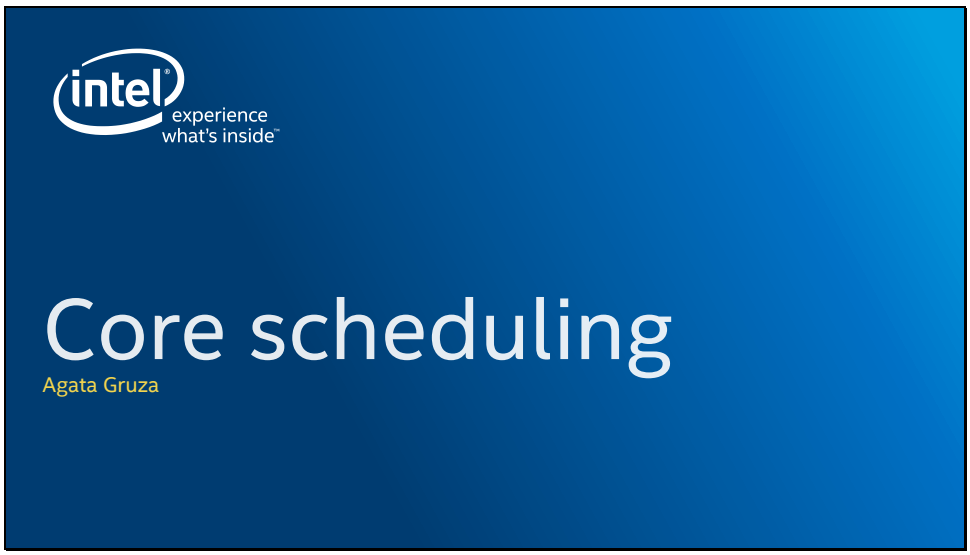
Slide 1


intel
experience
what's inside
Core scheduling
Agata Gruza

Slide 2

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.

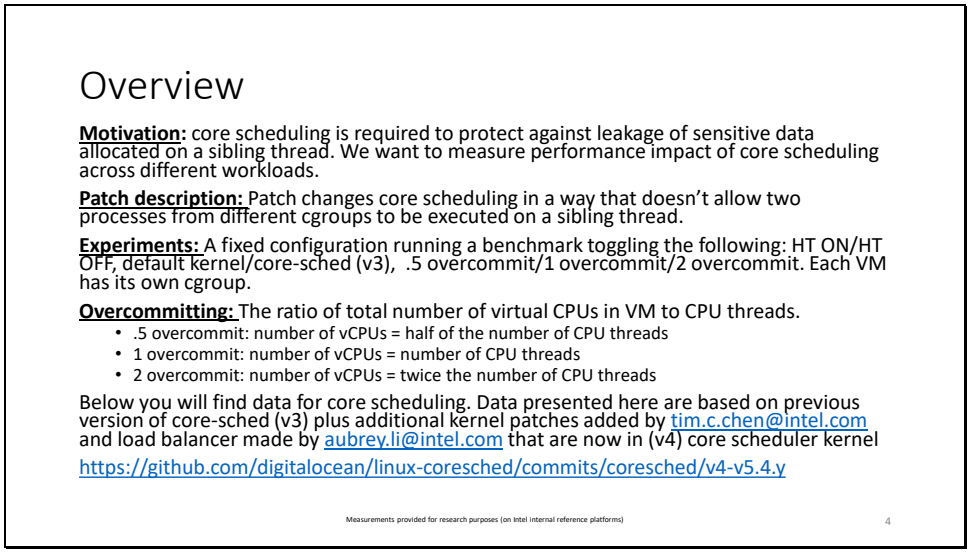
Slide 3

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.

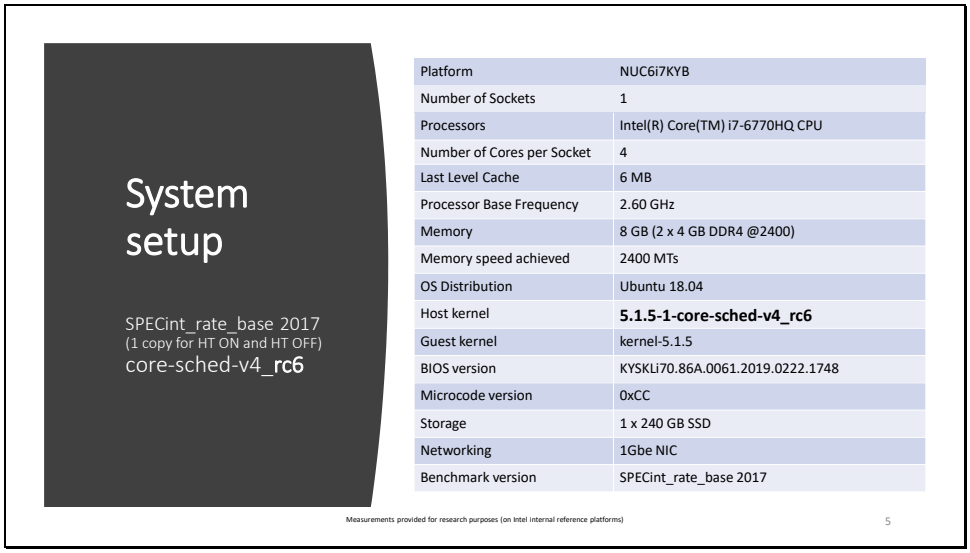
Measurements provided for research purposes (on Intel internal reference platforms)

Slide 4



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

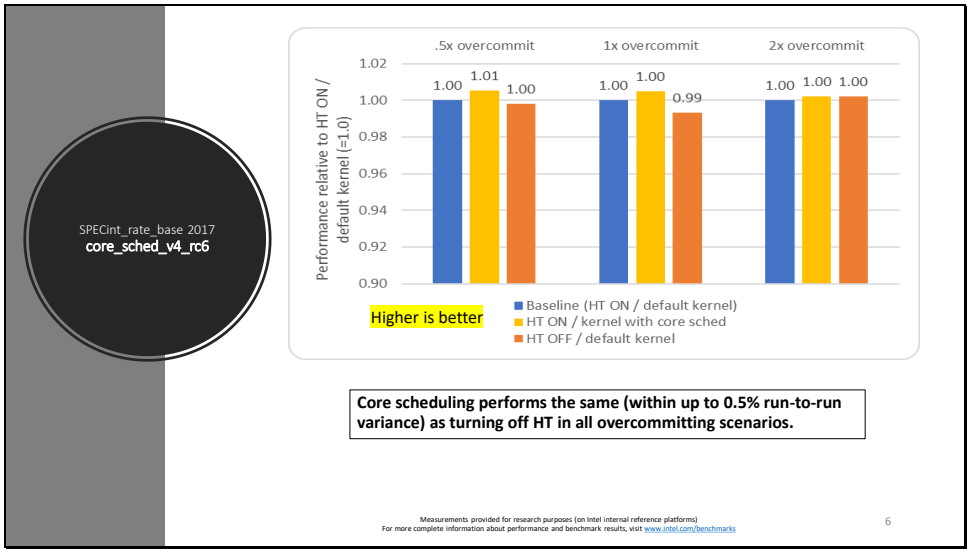


http://mark.intel.com/products/93341/Intel-Core-i7-6770HQ-Processor-6M-Cache-up-to-3-50-GHz-SKX

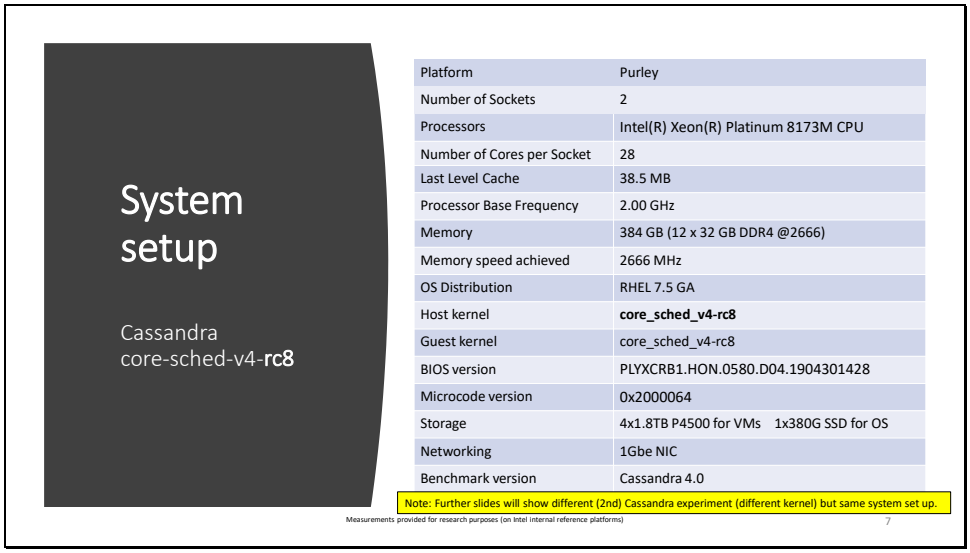
runcpu --config=core-sched.cfg --size=refrate --copies=1 --noreportable --iterations=3 intrate

1 copy (default value) - 3 iterations

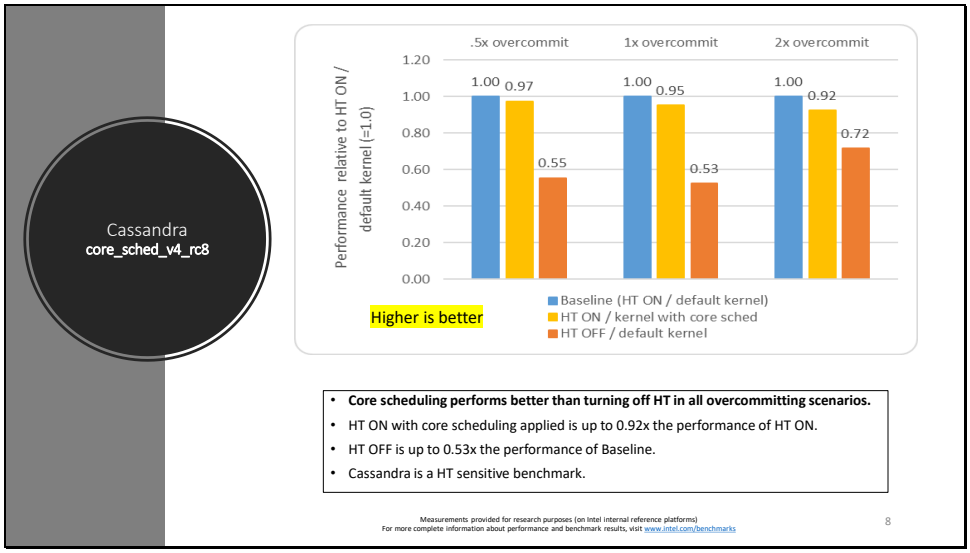
Slide 6


SPECint_rate_base 2017
core_sched_v4_rc6
.5x overcommit
1x overcommit
2x overcommit
Performance relative to HT ON / default kernel (=1.0)
1.02
1.00
0.98
0.96
0.94
0.92
0.90
1.00
1.01
1.00
1.00
1.00
0.99
1.00
1.00
1.00
Higher is better
Baseline (HT ON / default kernel)
HT ON / kernel with core sched
HT OFF / default kernel
Core scheduling performs the same (within up to 0.5% run-to-run variance) as turning off HT in all overcommitting scenarios.
Measurements provided for research purposes (on Intel internal reference platforms)
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
6

Slide 7

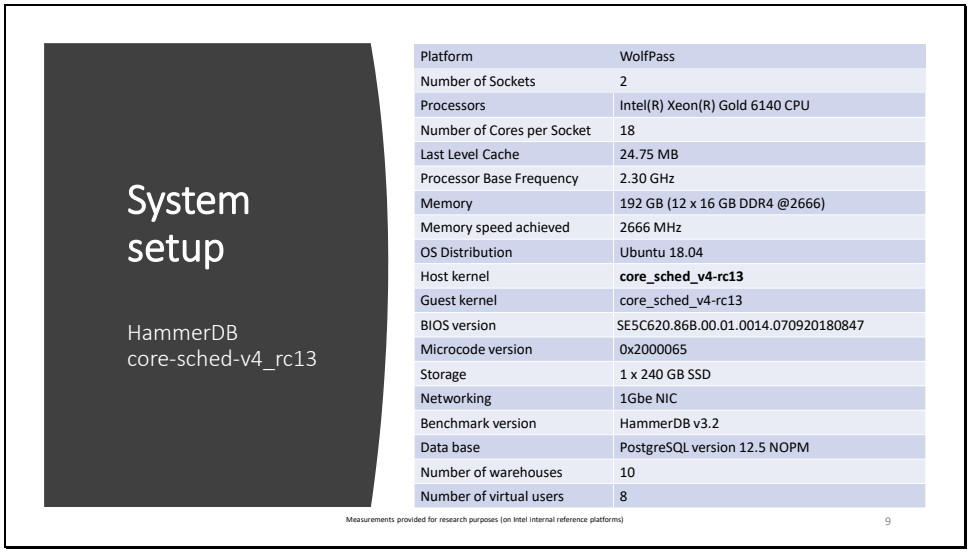

System setup
Cassandra core-sched-v4-rc8
Platform Purley
Number of Sockets 2
Processors Intel(R) Xeon(R) Platinum 8173M CPU
Number of Cores per Socket 28
Last Level Cache 38.5 MB
Processor Base Frequency 2.00 GHz
Memory 384 GB (12 x 32 GB DDR4 @2666)
Memory speed achieved 2666 MHz
OS Distribution RHEL 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



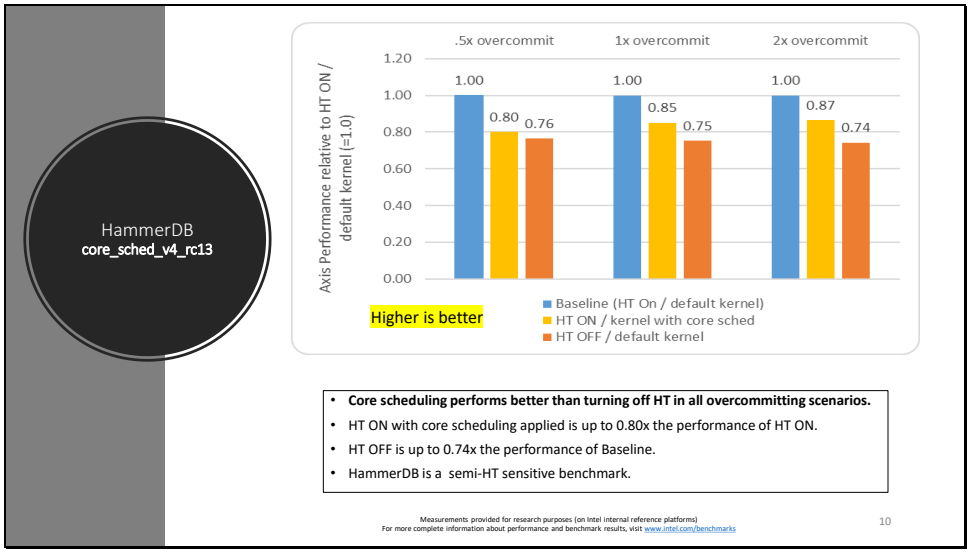
When applying core scheduling we are slower by up to 7.63%.
With HT OFF we are slower than with HT ON, by up to 47.47%.

Slide 9



http://mark.intel.com/products/120485/Intel-Xeon-Gold-6140-Processor-24-75M-Cache-2-30-GHz-
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

Slide 10



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

Slide 11

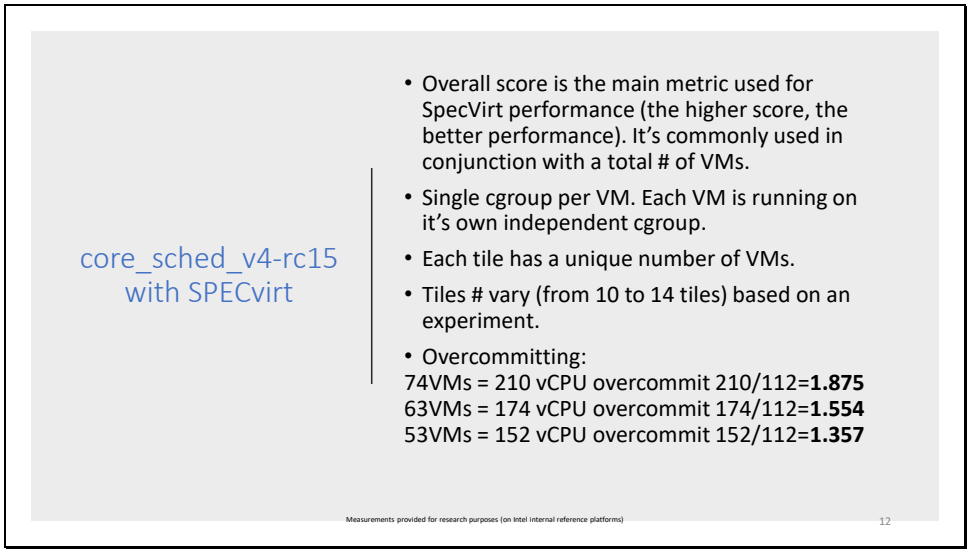
SPECvirt Core Scheduler_v4_rc15

- SPECvirt workloads:

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

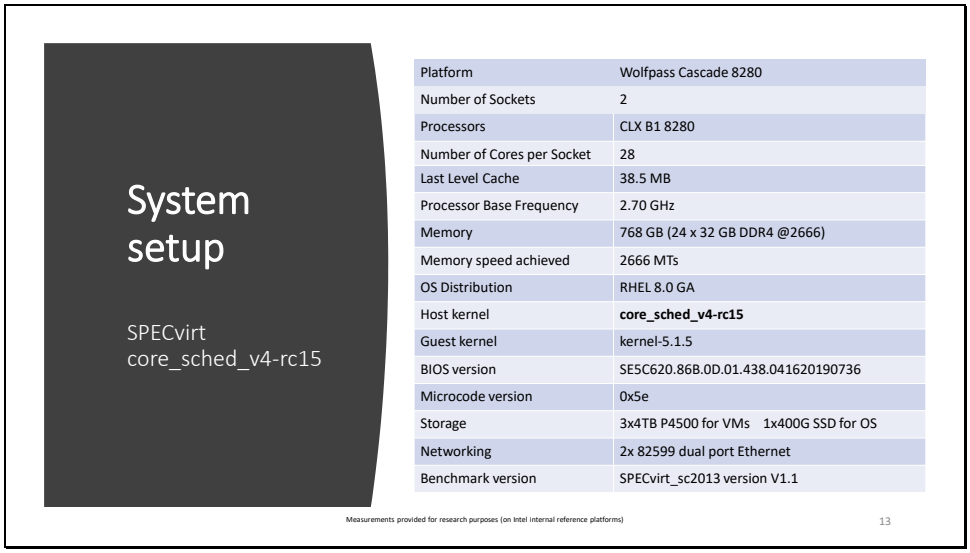
- Tile topology:
  - 1 tile contains:
    - 1 appserver
    - 1 webserver
    - 1 infraserver
    - 1 mailserver
    - 1 batchserver
  - 4 tiles share 1 DB

Slide 12



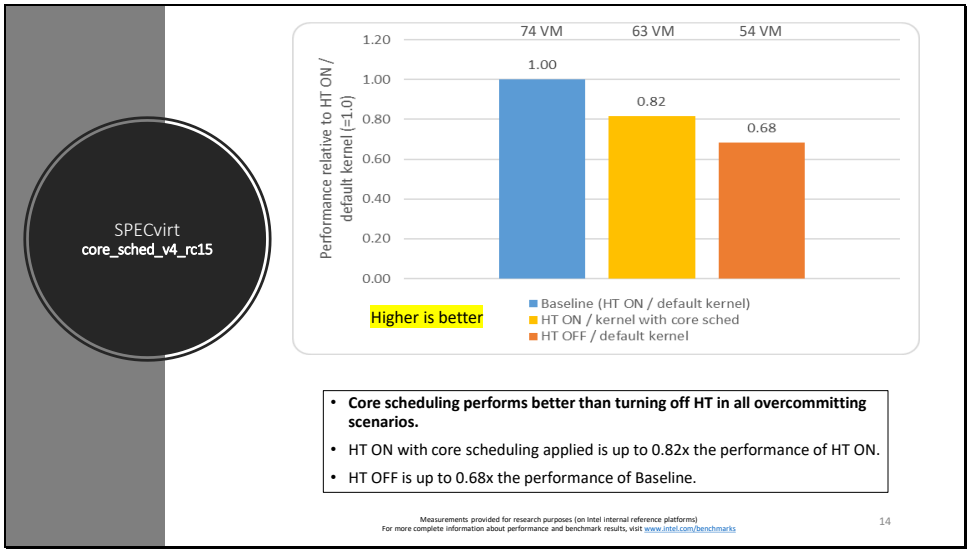
VMWare has 14 tiles on SKX
53 VMs → 10 tiles
63 VMs → 12 tiles
74 VMs → 14 tiles

Slide 13



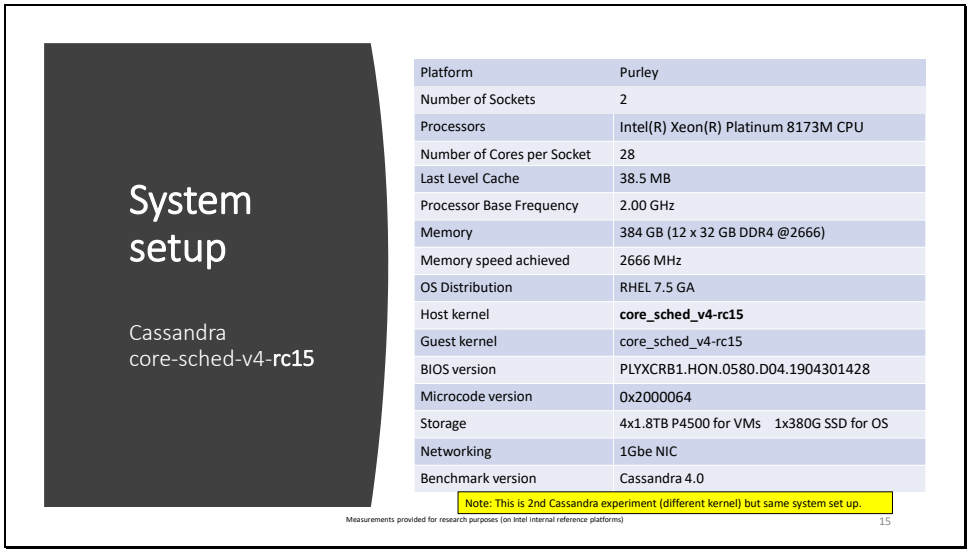
http://mark.intel.com/products/192478/Intel-Xeon-Platinum-8280-Processor-38-5M-Cache-2-70-GHz-

Slide 14

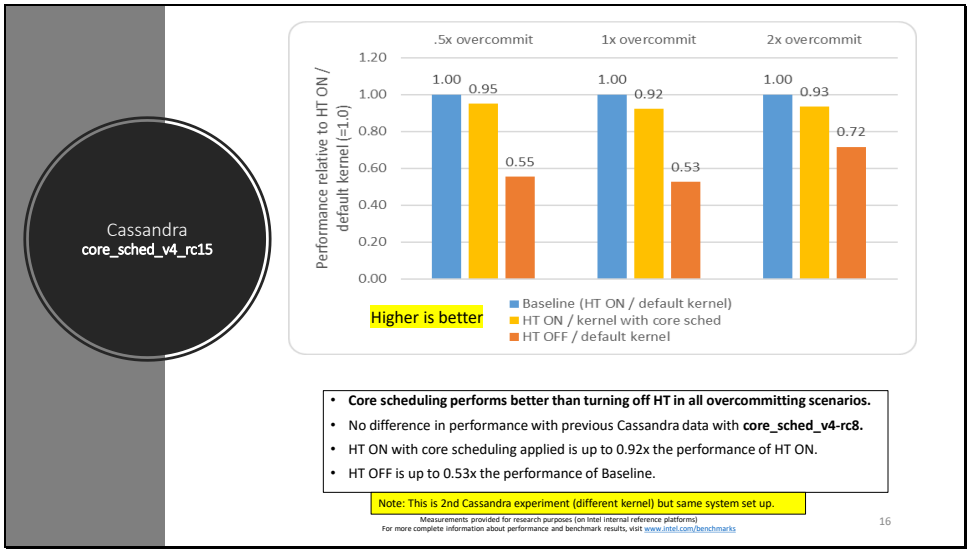


When applying core scheduling we are slower by up to 18.00%.
With HT OFF we are slower by up to 32%.

Slide 15


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

Slide 16



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