

# Vector API based Spark SQL Query Optimization in Parquet on SPR

Internal-Use Only  
Do Not Share outside of Intel

Template Revision #: DCPCOMP-002  
June 2023

Author: Xin Dong, Fang Xie, Haitao Li



# Executive Summary

- What is the objective of benchmarking?
  - Benefit from Java Vector API (based on AVX-512) optimizes Spark SQL Query in Parquet reader on SPR
  - Showcase Spark SQL Query improvement based on optimized Parquet reader on SPR
- What are the key takeaways?
  - SSB (Star Schema Benchmark) gains **13%** throughput improvement with Parquet optimization
  - SQL String “LIKE” Operator gains up to **96%** throughput improvement
- Where is this claim going to be used?
  - The data will be presented outside to demonstrate SPR advantages and scale SPR solutions
  - It may be quoted by partners or public events.

# Workload Description

## ■ SSB (Star Schema Benchmark)

- “The SSB (Star Schema Benchmark) is designed to measure performance of database products in support of classical data warehousing applications and is based on the TPC-H benchmark.” ([Reference-1](#))
- The dataset imports SSB generated tables (customer, part, supplier, lineorder) and converts “star schema” to denormalized “flat schema”. 13 query SQLs derived from SSB examine the performance of database. ([Reference-2](#))

## ■ Spark query SQL

- Spark is used to query the SSB’s “flat schema” table. 11 query SQLs are selected to examine the performance. (Q1.3 and Q3.4 are removed due to incompatible Spark’s syntax), which are derived from SSB SQL queries <sup>\*</sup>.
- A SQL String “LIKE” Operator is added in this workload to demonstrate the advantage of string type optimization.

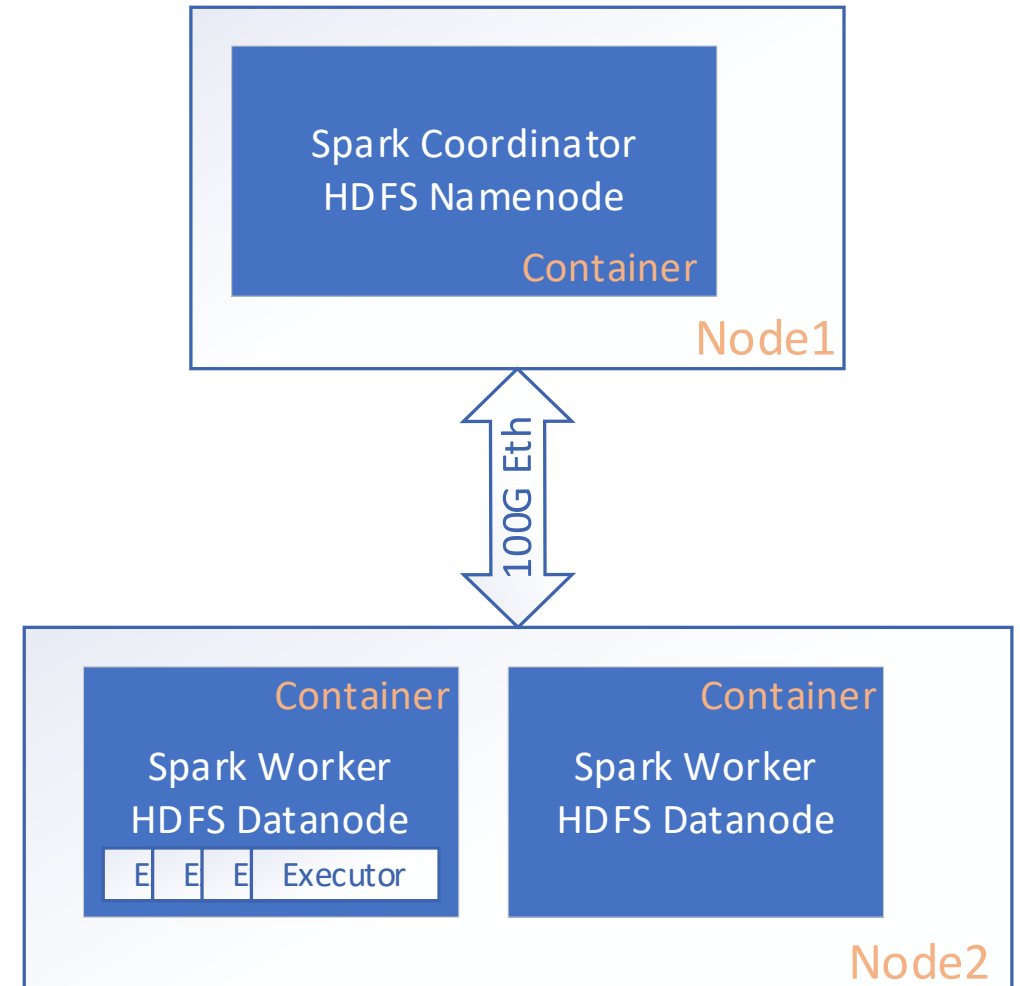
## ■ Hadoop HDFS

- The generated data is stored into Hadoop HDFS.

<sup>\*</sup> The reason for slightly modifying the SSB query SQL is to adapt the syntax of Spark

# Workload Description

- Working scenario of the workload
  - Two servers are required in this scenario.
  - Spark workers and coordinator are in different servers.
  - Spark workers are in a same node to shield the network influence.
  - Hadoop follows the similar deployment.
- Optimized source code
  - The source code can be found on [Intel-innersource](#).
  - Part of the code has upstreamed to Apache: [PARQUET-2159](#)



# System Configuration

	COORDINATOR	WORKER
Name	r06u30-dcai-spr-cd	r06u36-dcai-spr-cd
Time	Tue Jun 27 12:04:36 AM UTC 2023	Tue Jun 27 12:04:29 AM UTC 2023
System	M50FCP2SBSTD	M50FCP2SBSTD
Baseboard	M50FCP2SBSTD	M50FCP2SBSTD
Chassis	Rack Mo4unt Chassis	Rack Mount Chassis
CPU Model	Intel(R) Xeon(R) Platinum 8480+	Intel(R) Xeon(R) Platinum 8480+
Microarchitecture	SPR_XCC	SPR_XCC
Sockets	2	2
Cores per Socket	56	56
Hyperthreading	Disable	Disable
CPUs	112	112
Intel Turbo Boost	Enabled	Enabled
Base Frequency	2.0GHz	2.0GHz
All-core Maximum Frequency	3.0GHz	3.0GHz
Maximum Frequency	3.8GHz	3.8GHz
NUMA Nodes	2	2
Prefetchers	L2 HW, L2 Adj., DCU HW, DCU IP	L2 HW, L2 Adj., DCU HW, DCU IP
Installed Memory	512GB (16x32GB DDR5 4800 MT/s [4800 MT/s])	512GB (16x32GB DDR5 4800 MT/s [4800 MT/s])
Hugepagesize	2048 kB	2048 kB
Transparent Huge Pages	madvise	madvise
Automatic NUMA Balancing	Enabled	Enabled
NIC	Intel Corporation Ethernet Controller E810-C	Intel Corporation Ethernet Controller E810-C
Disk	INTEL SSDPE2KX010T8	INTEL SSDPE2KX010T8
BIOS	SE5C7411.86B.9525.D03.2301041329	SE5C7411.86B.9525.D03.2301041329
Microcode	0x2b000181	0x2b000181
OS	Ubuntu 22.04.2 LTS	Ubuntu 22.04.2 LTS
Kernel	5.15.0-75-generic	5.15.0-75-generic
TDP	350 watts	350 watts
Power & Perf Policy	Performance	Performance
Frequency Governor	performance	performance
Frequency Driver	intel_pstate	intel_pstate
Max C-State	9	9

# Security Mitigations

Intel® Xeon® Platinum 8480+ CPU (Sapphire Rapids) / coordinator-node:

CVE-2017-5753	OK (Mitigation: usercopy/swapgs barriers and __user pointer sanitization)
CVE-2017-5715	OK (Enhanced IBRS + IBPB are mitigating the vulnerability)
CVE-2017-5754	OK (Not affected)
CVE-2018-3640	OK (your CPU microcode mitigates the vulnerability)
CVE-2018-3639	OK (Mitigation: Speculative Store Bypass disabled via prctl and seccomp)
CVE-2018-3615	OK (your CPU microcode mitigates the vulnerability)
CVE-2018-3620	OK (Not affected)
CVE-2018-3646	OK (your kernel reported your CPU model as not affected)
CVE-2018-12126	OK (Not affected)
CVE-2018-12130	OK (Not affected)
CVE-2018-12127	OK (Not affected)
CVE-2019-11091	OK (Not affected)
CVE-2019-11135	OK (your CPU vendor reported your CPU model as not affected)
CVE-2018-12207	OK (this system is not running a hypervisor)
CVE-2020-0543	OK (your CPU vendor reported your CPU model as not affected)
CVE-2022-0001	OK (unprivileged eBPF disabled)
CVE-2022-0002	OK (unprivileged eBPF disabled)

Intel® Xeon® Platinum 8480+ CPU (Sapphire Rapids) / worker-node:

CVE-2017-5753	OK (Mitigation: usercopy/swapgs barriers and __user pointer sanitization)
CVE-2017-5715	OK (Enhanced IBRS + IBPB are mitigating the vulnerability)
CVE-2017-5754	OK (Not affected)
CVE-2018-3640	OK (your CPU microcode mitigates the vulnerability)
CVE-2018-3639	OK (Mitigation: Speculative Store Bypass disabled via prctl and seccomp)
CVE-2018-3615	OK (your CPU microcode mitigates the vulnerability)
CVE-2018-3620	OK (Not affected)
CVE-2018-3646	OK (your kernel reported your CPU model as not affected)
CVE-2018-12126	OK (Not affected)
CVE-2018-12130	OK (Not affected)
CVE-2018-12127	OK (Not affected)
CVE-2019-11091	OK (Not affected)
CVE-2019-11135	OK (your CPU vendor reported your CPU model as not affected)
CVE-2018-12207	OK (this system is not running a hypervisor)
CVE-2020-0543	OK (your CPU vendor reported your CPU model as not affected)
CVE-2022-0001	OK (unprivileged eBPF disabled)
CVE-2022-0002	OK (unprivileged eBPF disabled)

# Software/Wokrload Configuration

SOFTWARE CONFIGURATION	SSB-Spark-Standalone
Orchestration	Kubernetes + Containerd
Kubernetes	v1.24.4
Containerd	Version: 1.6.15
Spark	Version: 3.3.2
Spark deploy mode	Standalone
JDK	Openjdk-17.0.5+8
SSB	Version: 2.1.8.5
SCALE_FACTOR	1000 *
Bind_Core number	32
Worker number	2
Stream number	4
Core number per executor	4
Memory size per executor	40g
Hadoop	Version: 3.3.2
Numctl parameter for Spark	Coordinator@server1:-- numactl -C 0-7 --localalloc Worker1@server2:-- numactl -C 0-15 --localalloc Worker2@server2: -- numactl -C 56-71 --localalloc

\* SCALE\_FACTOR = 1000, means the data size is around 600G and the amount of Parquet files is 4400.

# Raw Performance Results

SSB throughput:

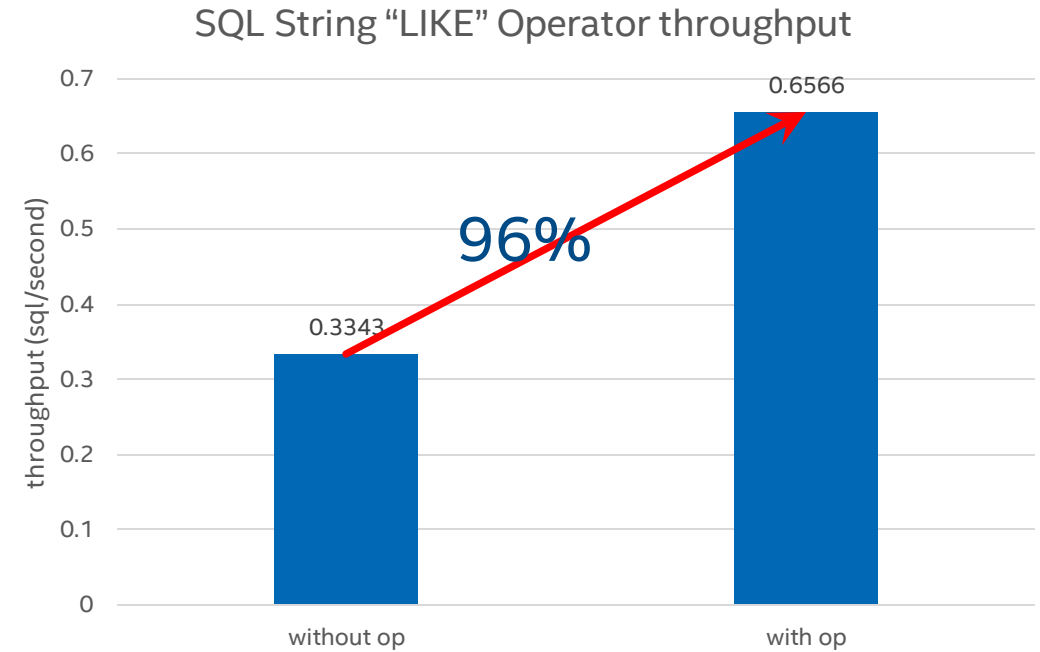
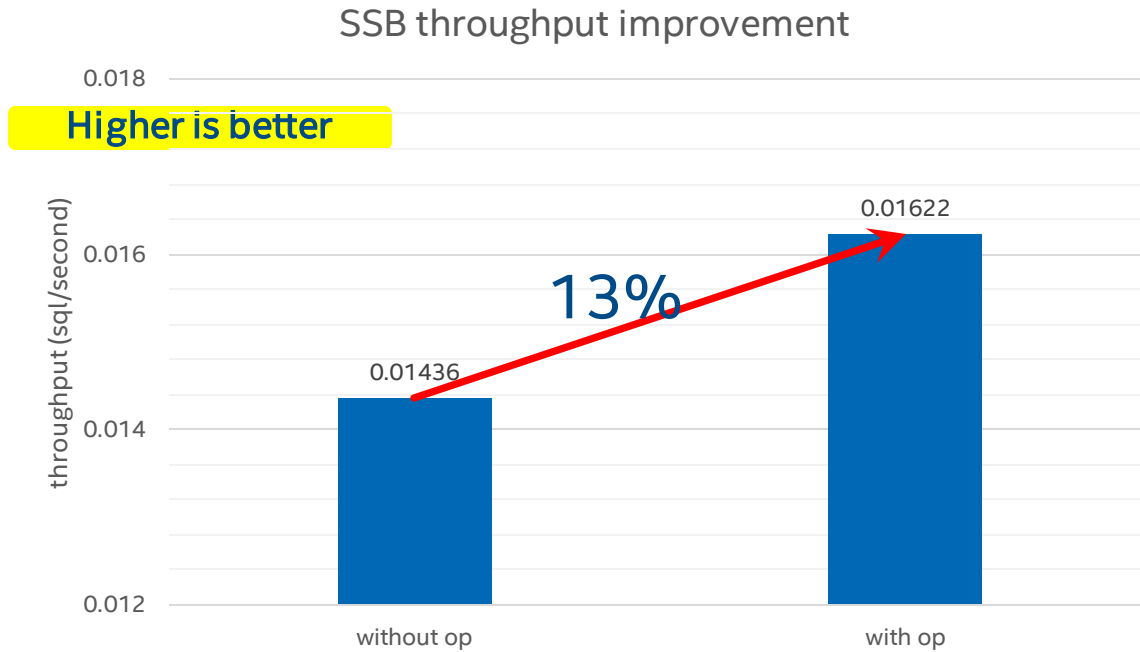
	With Parquet reader optimization	Without Parquet reader optimization
Execution time (second)	678.14	765.95
Throughput (sql/sec)	0.01622	0.01436
CPU frequency (GHz)	2.98	2.99
CPU utilization	91.47%	92.45%

SQL String “LIKE” Operator throughput:

	With Parquet reader optimization	Without Parquet reader optimization
Execution time (second)	15.23	29.91
Throughput (sql/sec)	0.6566	0.3343



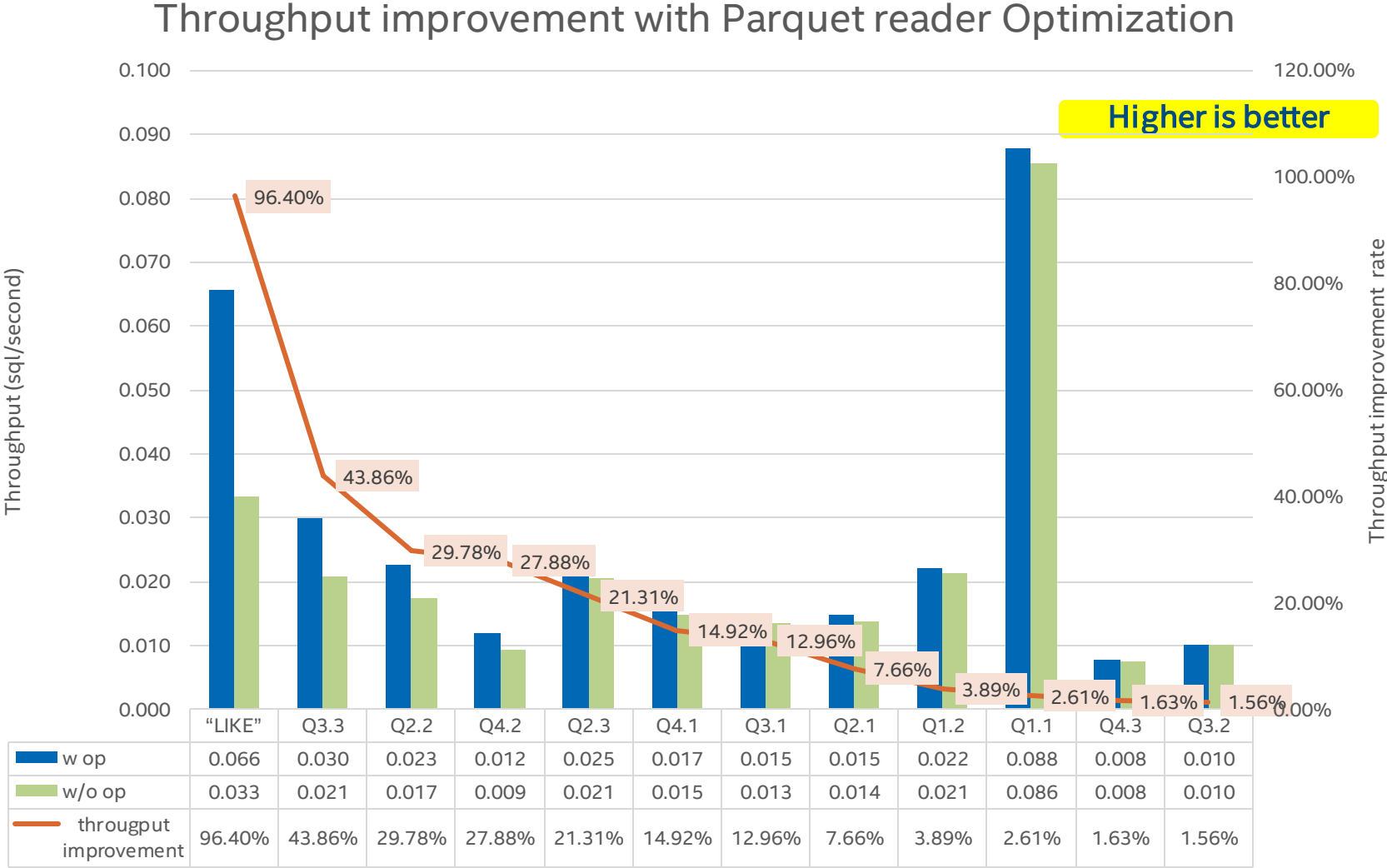
# Throughput improvement with Optimization



## Summary:

- With optimized Parquet reader, SSB has an **13%** throughput improvement.
- The SQL string "LIKE" Operator can have an **96%** throughput improvement.

# Each query improvement with Optimization



# Emon data Conclusion

- With Vector optimized Parquet reader, value of *metric\_core % cycles in license level 3* is not zero.
- Compared two optimized SQL queries (“LIKE” Operator and SSB Q3.3), the higher the throughput improvement, the higher the value of *metric\_core % cycles in license level 3*.
- With Vector optimized Parquet reader, value of *metric\_core initiated local dram read bandwidth (MB/sec)* is much smaller, new algorithm reads less data from memory.
- With Vector optimized Parquet reader, value of *INST\_RETIRED.ANY* is much smaller, path length is shorter.

# Emon data for SQL “LIKE” Operator

	With Parquet reader Optimization	Without Parquet reader Optimization
(EDP 5.4) name (sample #1 - #241)	cpu 0 (S0C0T0)	cpu 0 (S0C0T0)
metric_CPU operating frequency (in GHz)	2.9788	2.9901
metric_CPU utilization %	83.3707	93.3551
metric_CPU utilization% in kernel mode	12.7002	7.2981
metric_CPI	0.3960	0.2826
metric_kernel_CPI	1.3343	1.3980
metric_EMON event mux reliability% (>95% good)	99.9423	99.8467
metric_core % cycles in license level 1	29.8382	25.3762
metric_core % cycles in license level 2	24.7400	74.6238
metric_core % cycles in license level 3	45.4218	0.0000
metric_core % cycles in license level 4	0.0000	0.0000
metric_core % cycles in license level 5	0.0000	0.0000
metric_core % cycles in license level 6	0.0000	0.0000
metric_core initiated local dram read bandwidth (MB/sec)	1,678.4735	2,662.9337
metric_core initiated remote dram read bandwidth (MB/sec)	41.2989	31.1625
INST_RETIRED.ANY	6271817321.16038	9878763851.12823
throughput (sql/second)	0.065665347	0.033434149
Throughput improvement	96.40%	

# Emon data for SSB Q3.3

	With Parquet reader Optimization	Without Parquet reader Optimization
(EDP 5.4) name (sample #1 - #241)	cpu 0 (S0C0T0)	cpu 0 (S0C0T0)
metric_CPU operating frequency (in GHz)	2.98	2.9932
metric_CPU utilization %	91.7994	92.6950
metric_CPU utilization% in kernel mode	18.2359	13.8043
metric_CPI	0.4244	0.3089
metric_kernel_CPI	1.2552	1.3351
metric_EMON event mux reliability% (>95% good)	99.7439	99.7985
metric_core % cycles in license level 1	51.3870	37.8577
metric_core % cycles in license level 2	24.4715	62.1423
metric_core % cycles in license level 3	24.1416	0.0000
metric_core % cycles in license level 4	0.0000	0.0000
metric_core % cycles in license level 5	0.0000	0.0000
metric_core % cycles in license level 6	0.0000	0.0000
metric_core initiated local dram read bandwidth (MB/sec)	1,739.2339	2,230.5799
metric_core initiated remote dram read bandwidth (MB/sec)	104.7535	67.4988
INST_RETIRED.ANY	6450999450.88388	8980937356.43984
throughput (sql/second)	0.029987447	0.020845164
Throughput improvement	43.86%	



# Backup







Intel Confidential – Internal Use Only

# Configuration of the workload

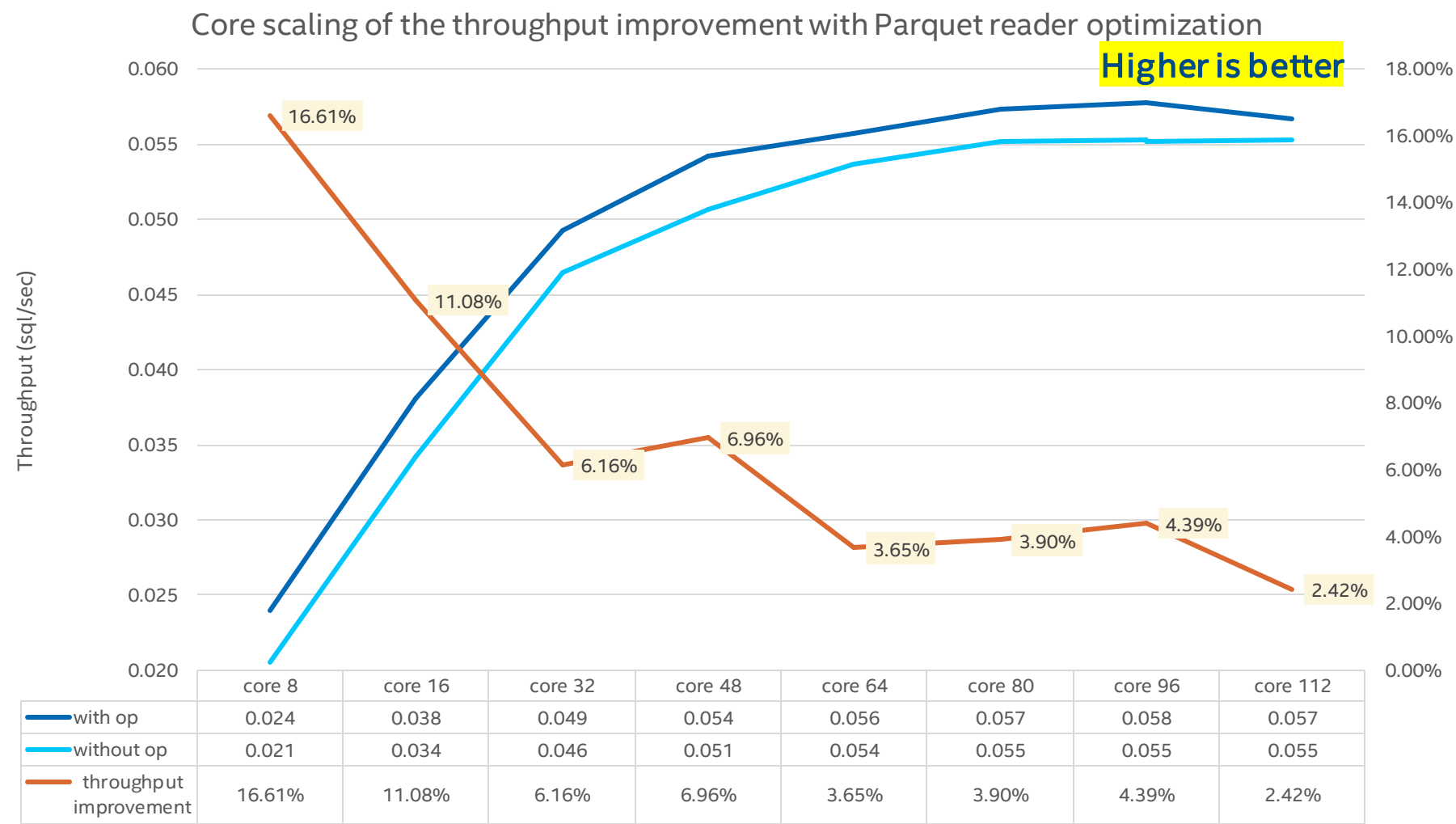
Keys	Values
Workload Name	ssb-spark-standalone_jdk17
Test Case	test_static_ssb-spark-standalone_jdk17_multi_parquet
Command Line	./ctest.sh -R '^test_static_ssb-spark-standalone_jdk17_multi_parquet\$' -V --set TERRAFORM_CONFIG TF=\$(readlink -f 5416341b-8848-48ed-9bc6-f3f05c798c3e-terraform-config.tf)
OS_VER	22.04
OS_IMAGE	ubuntu
OPENJDK_VER	jdk-17.0.5+8
OPENJDK_PKG	<a href="https://github.com/adoptium/temurin17-binaries/releases/download/jdk-17.0.5%2B8/.tar.gz">https://github.com/adoptium/temurin17-binaries/releases/download/jdk-17.0.5%2B8/.tar.gz</a>
SSB_VERSION	0741e06d4c3e811bcec233378a39db2fc0be5d79
SSB_REPO	<a href="https://github.com/vadimtk/ssb-dbgen.git">https://github.com/vadimtk/ssb-dbgen.git</a>
SPARK_VER	spark-3.3.2
SPARK_PKG	<a href="https://archive.apache.org/dist/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz">https://archive.apache.org/dist/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.tgz</a>
PARQUET_OPTIMIZE_PKG	<a href="https://af01p-igk.devtools.intel.com/artifactory/platform_hero-igk-local/hero_features_assets/data_services/SSB/">https://af01p-igk.devtools.intel.com/artifactory/platform_hero-igk-local/hero_features_assets/data_services/SSB/</a>
HADOOP_VER	hadoop-3.3.2
HADOOP_PKG	<a href="https://archive.apache.org/dist/hadoop/core/hadoop-3.3.2/hadoop-3.3.2.tar.gz">https://archive.apache.org/dist/hadoop/core/hadoop-3.3.2/hadoop-3.3.2.tar.gz</a>
spark_master_port	29000
spark_workers_num	2
jdk_version	jdk17
parquet_opt	TRUE
spark_mode	multi
scale_factor	1000
sql_num	999
regen_data	FALSE
stream	4
hadoop_replication	TRUE
spark_executor_core	4
spark_executor_mem	40
bind_core_flg	TRUE
bind_core_num	16



# Emon Files

	"LIKE" SQL	SSB Q3.3
with Vector Opt	 summary_w_LIKE. xlsx	 summary_w_Q33. xlsx
without Vector Opt	 summary_wo_LIK E.xlsx	 summary_wo_Q3 3.xlsx

# Core scaling of throughput improvement\*



\* The stream number of these performance data is 1.