Improvements to OpenSearch using Intel Sapphire Rapids and hardware accelerators

OpenSearchCon
NORTH AMERICA | San Francisco 2024

Presenter

Akash Shankaran

Lead Software Architect, Intel

Partner Speaker : Michelle Tabirao, Canonical

AGENDA

- Hardware acceleration within OpenSearch
  - Vector Search acceleration – AVX512
  - Compression acceleration – QAT
- OPEA
- Partner collaboration

These accelerators are available on Intel instances in Public Clouds !!

CHALLENGES

- Growth in data
- Explosion in vector capabilities

MOTIVATION

- Hardware innovation will become the driving force for data management systems
- Creates additional hardware-software codesign opportunities.
- Further facilitate disaggregation of compute, storage and memory.

HISTORY

Intel’s contributions to OpenSearch

- Zstandard compression (2023)
- Indexing tuning guides (2023)
- QAT compression (2024)
- Vector Search using AVX512 (2024)

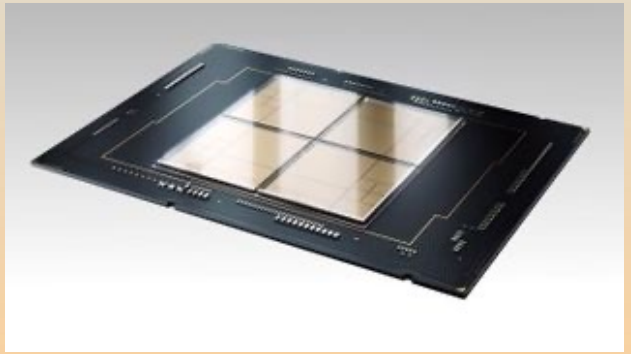
BENEFITS

Why hardware acceleration?

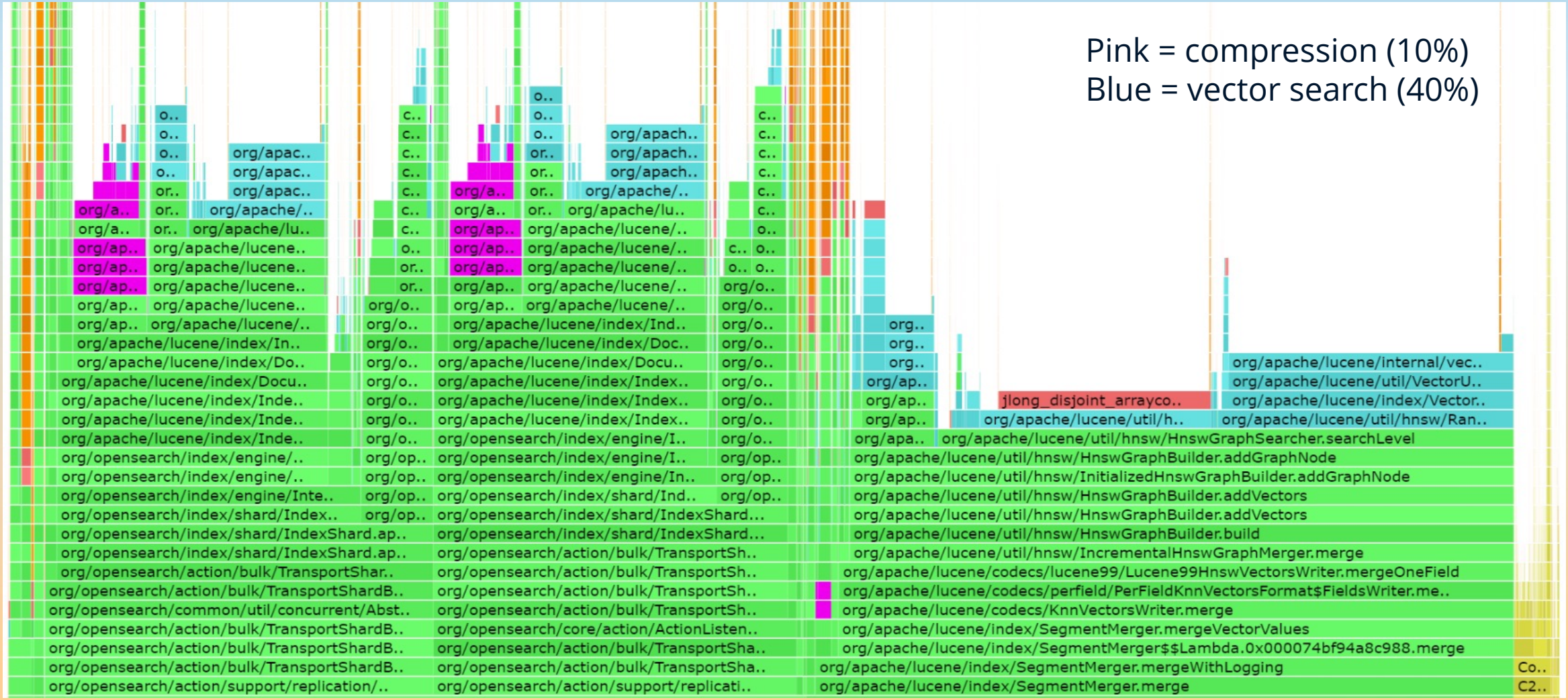
PERFORMANCE

ENERGY EFFICIENCY

BETTER COMPUTE UTILIZATION

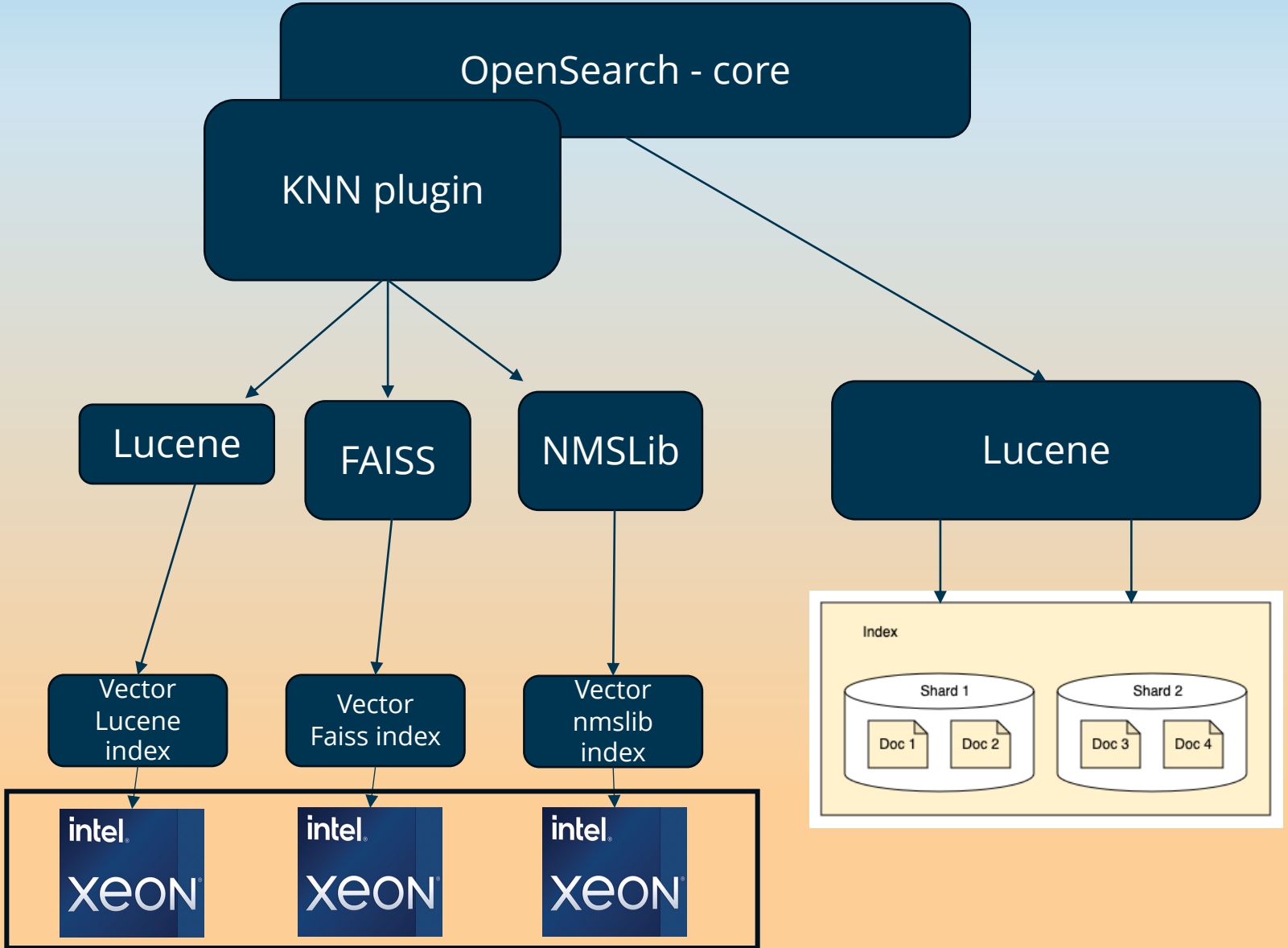


Why hardware acceleration in OpenSearch?



Results of using hardware accelerators in open search workloads

Where is acceleration happening?



VECTOR SEARCH

Benchmark setup:

Workload	Vectorsearch: cohere dataset, documents: 10M , Dimensions: 768
Application Server	OpenSearch docker image 2.16.0
Application Client	OpenSearch-Benchmark 1.6
Tools/Compilers	OpenJDK 21
Drivers	Default OS Drivers (Ubuntu 22.04)
K-nn engine	Lucene

Number of data nodes	3
Number of coordinating nodes	1
Number of cluster management nodes	1
Java heap size (GB)	64
HNSW ef_construction	256
HNSW ef_search	256
HNSW m	16
Space type	innerproduct
Max number of segment per shard	10
Bulk indexing clients	20
Number of queries	50000
Number of searched neighbors (k)	100

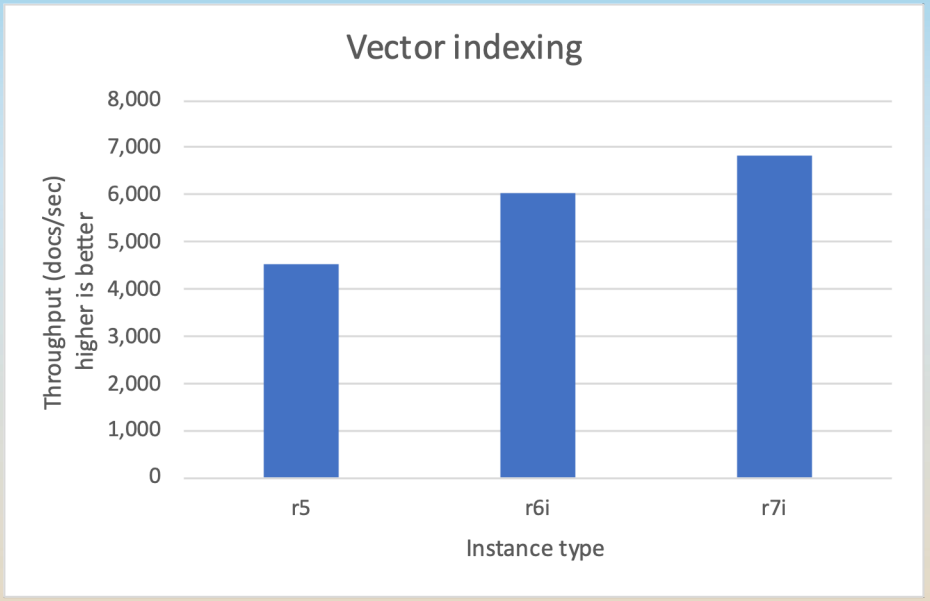
VECTOR SEARCH – SAPPHIRE RAPIDS (R7I - AWS) PERFORMANCE

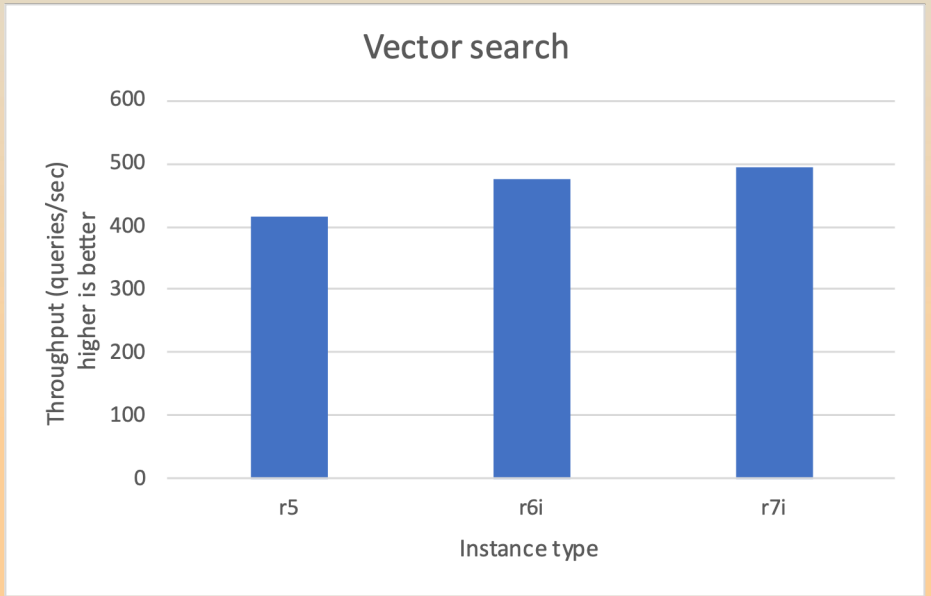
Intel instance gen over gen performance on AWS cloud:

Generation over generation:

Instance change	% improvement
r5 -> r6i	14 - 34%
r6i -> r7i	5 – 13%

r5 => r7i == ~43% better TCO



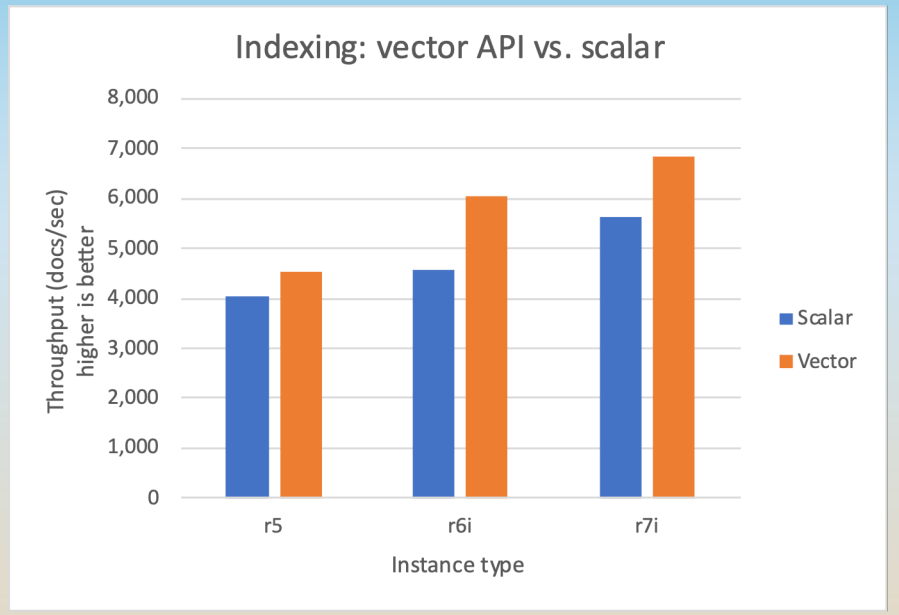


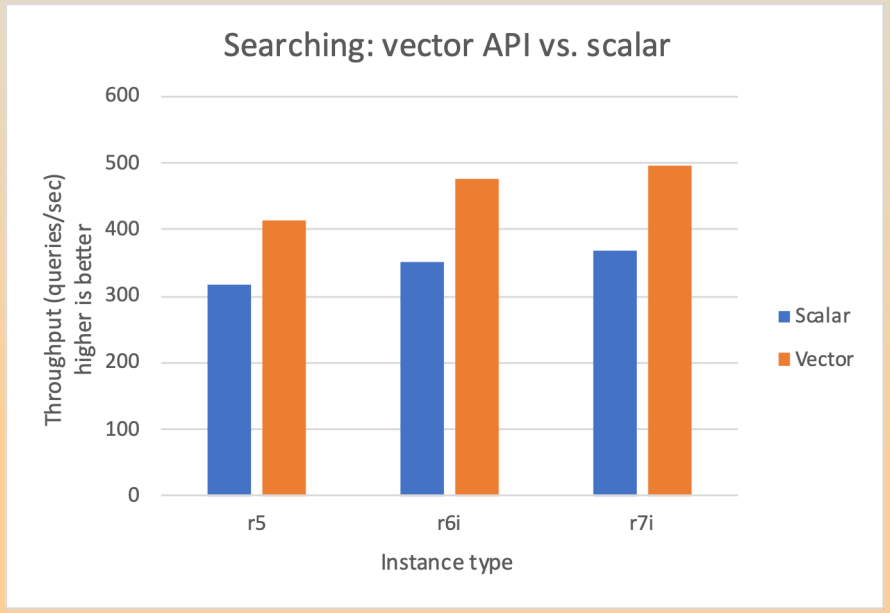
VECTOR SEARCH – AVX512 ACCELERATION

Intel instance performance using AVX512:

Same product, benefits due to vectorization:

Instance family	Benefits due to AVX512
r5	31%
r6i	35%
r7i	34%





What is Intel QuickAssist Technology (QAT)

Integrated onboard encryption and compression accelerator

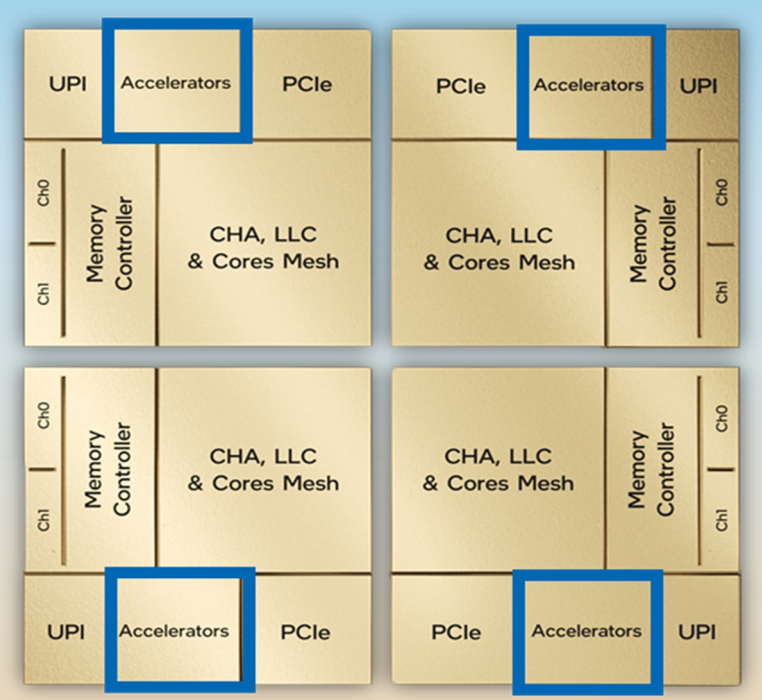
Where can I get this?

Available on 4th generation Intel® Xeon® Scalable processors (Sapphire Rapids) and newer processors.

What does it do?

Offloads computationally intensive operations from the CPU cores, allowing the CPU to perform other tasks more efficiently for greater overall system performance, efficiency, and power.

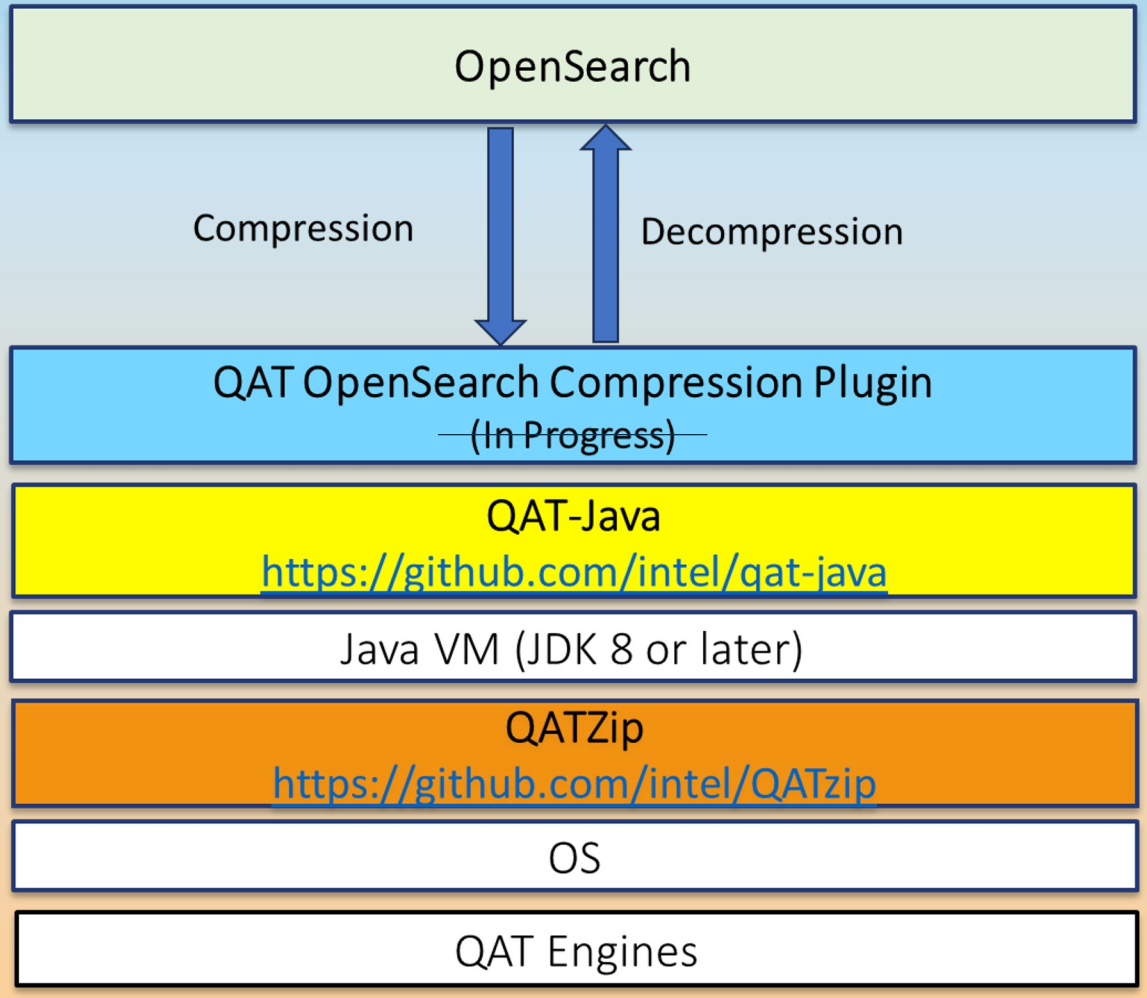
Supported algorithms: Deflate, LZ4 and Zstd*



As of today, QAT is available in OpenSearch as a custom-codec.

Can be used on any AWS Bare-Metal c7i, m7i, r7i instance

QAT SOFTWARE STACK



QAT INDEXING

Benchmark setup:

OpenSearch indexing workload with StackOverflow corpus.
OpenSearch-Benchmark used for benchmarking
Single node setup with 4 data nodes on 1 socket and benchmark driver on the other.

	BEST_COMPRESSION	QAT_DEFLATE	%
Total time (s)	739	655	11.4
Mean throughput	220,027	272,089	23.7
Store size	162.6	166.7	2.5
CPU%	52.3	49.4	5.7

With QAT:
~24% throughput improvement
~12% store size reduction
~5-7% CPU offloaded

Main reason for improvement:
Path Length reduction

	LZ4	QAT_LZ4	%
Total time (s)	702	685	2.5
Mean throughput	252806	261407	3.4
Store size	216.3	189.3	12.5
CPU%	49.5	45.9	7.1

Intel instances relevant to OpenSearch

Cloud	Instance families
AWS	m7i, c7i, r7i, i4i
Azure	Edv5, Edsv6, Dsv6, Ddsv6
GCP	c4, n4, x4, c3, n3, h3

Cloud:

Emerald Rapids – available on Azure and GCP

Sapphire Rapids – available on AWS, Azure and GCP

OPEA

Open Platform for Enterprise AI

What is OPEA?

Framework of composable building blocks for Generative AI solutions

For example: LLMs, vector databases, and prompt engines

Blueprints of end-to-end workflows

For example: RAG applications, Agentic systems, AI avatar

Construction

Assessment for **grading Gen AI systems**

For example: performance, features, trustworthiness and enterprise-grade readiness

Evaluation

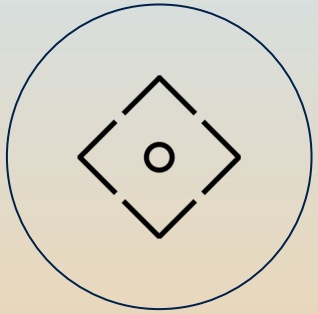
https://opea.dev/

OPEA VALUE

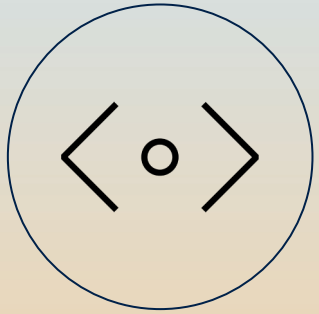
Helps Enterprises unlock value from their data using Generative AI (LLM, RAG) faster and easier

Reduces complexities of fragmented ecosystem and helps solutions to scale in production

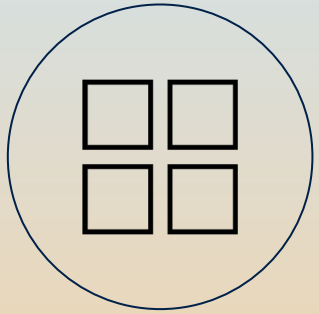
Ignites collaboration and contribution across industry leaders partnering with the Linux Foundation



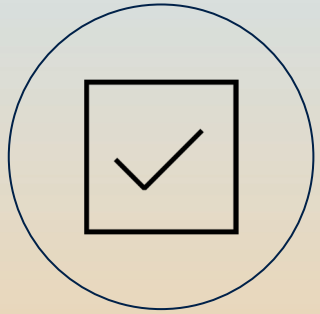
Efficient
Harnesses existing infrastructure, the AI accelerator or other hardware of your choosing.



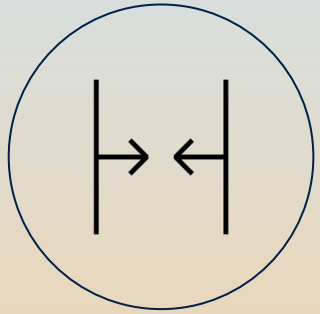
Seamless
Integrates with enterprise software, with heterogeneous support and stability across system & network.



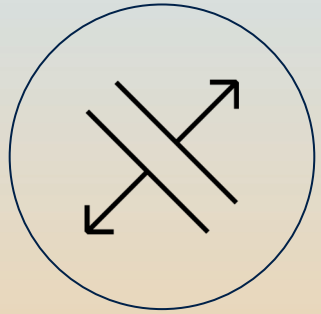
Open
Brings together best of breed innovations and is free from proprietary vendor lock-in.



Ubiquitous
Runs everywhere through a flexible architecture built for cloud, data center, edge and PC.



Trusted
Features a secure enterprise-ready pipeline and tools for responsibility, transparency, and traceability.



Scalable
Provides access to a vibrant ecosystem of partners to help build and scale your solution.

https://opea.dev/

OPEA Partners

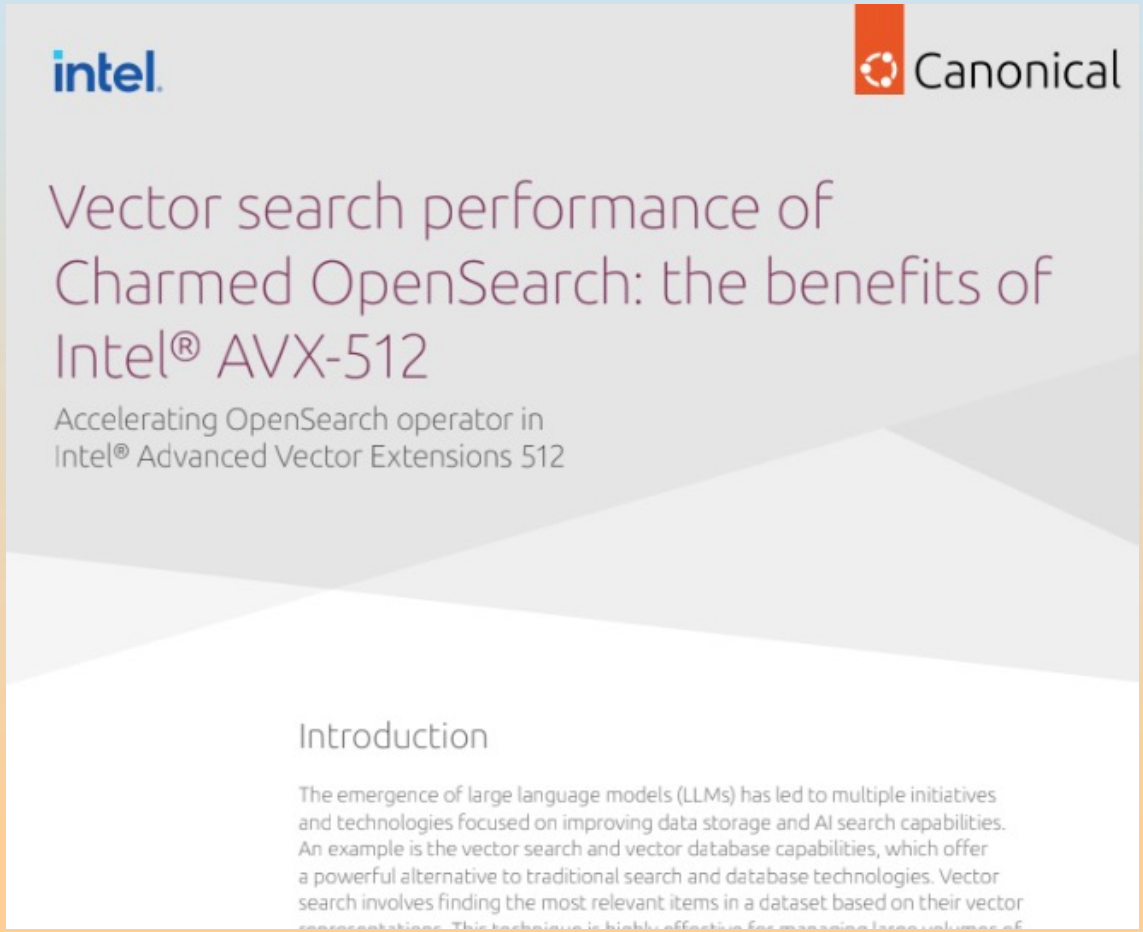
Partners as of Sep 9, 2024

Intel – canonical collaborations

- Enabling foundational technologies in the Ubuntu OS.
- Enabling hardware acceleration capabilities on Charmed OpenSearch
- Joint presentation on deploying LLMs with Retrieval Augmented Generation, featuring Intel Accelerators and Charmed OpenSearch

Data & AI Masters: Oct 1-2
https://events.ringcentral.com/events/canonical-data-and-ai-masters/registration

Charmed opensearch – avx512 acceleration


intel.

Canonical

Vector search performance of Charmed OpenSearch: the benefits of Intel® AVX-512

Accelerating OpenSearch operator in Intel® Advanced Vector Extensions 512

Introduction

The emergence of large language models (LLMs) has led to multiple initiatives and technologies focused on improving data storage and AI search capabilities. An example is the vector search and vector database capabilities, which offer a powerful alternative to traditional search and database technologies. Vector search involves finding the most relevant items in a dataset based on their vector

Taking the non-AVX benchmark results as a baseline, the tests show that both indexing and searching throughput increased by a double-digit percentage. The 90th percentile of latency was also reduced when compared with a non-AVX setup. Searching improved considerably in P99 and P99.9 by 20% and 15%, respectively. The relative change in performance can be found here:

	Bulk Indexing	Searching
Median throughput increase	34%	18%
90th Percentile Latency Reduction	35%	20%

The results above mean that Charmed OpenSearch 2.14.0 is enhanced by Vector API support. Indexing and searching get a third and fifth faster than the non-AVX baseline.

https://ubuntu.com/engage/opensearch-intel-performance

Next steps

- Vectorization: opportunities in k-nn FAISS library.
- Quantization: FP16, Byte quantized vectors
- Compression acceleration to vector search.
- AVX10: https://cdrdv2-public.intel.com/784343/356368-intel-avx10-tech-paper.pdf

Thank you!

Akash Shankaran
akash.Shankaran@intel.com
Github: akashsha1

Acknowledgements

Mulugeta Mammo
Vesa Pehkonen
Olasoji Denloye
Arun Gupta

<first.lastname@intel.com>

Sarthak Aggarwal
Naveen Tatikonda
Vamshi Nakkirtha
Dylan Tong
William Beckler
Andriy Redko
Michelle Tabirao

Questions?