Efficiency of First-Stage Retrieval

Katelyn Harlan
Supervisors: Andrew Trotman, Veronica Liesaputra
University of Otago, Dunedin, New Zealand

Background

First-Stage Retrieval

- Sparse retrieval, particularly using learned sparse representations (LSR), is seeing a rise in popularity.
- Hybrid approaches proving quite effective.
- More efficient and accurate results at the low end of the pipeline gives more time to refine results at the higher levels.

(1) Indexing → (2) Retrieval → (3) Reranking

Impact-Ordered Indexes // Score-at-a-Time

- Typical postings lists: $\langle d_1, f_{t,d_1} \rangle, \langle d_2, f_{t,d_2} \rangle, \ldots, \langle d_n, f_{t,d_n} \rangle$
- Instead of term frequencies, what if we stored a pre-computed score?
- Impact-ordered: $\langle i_t : d_1, d_2, \ldots, d_n \rangle$
- SaaT: process query in decreasing impact order, so effective results even with early termination.

$$S_d = \sum_{t \in Q \cap D} i_t$$

Score-at-a-Time

fruit	*i*=5	1	2	5	9	*i*=1	3	4	11	14
bat	*i*=3	5	9	*i*=2	1	2	12	16		

i=5	1	2	5	9
i=3	5	9		
i=2	1	2	12	16
i=1	3	4	11	14

JASSv2 & IOQP

To my knowledge, currently there are only two open-source SAAT search engines.

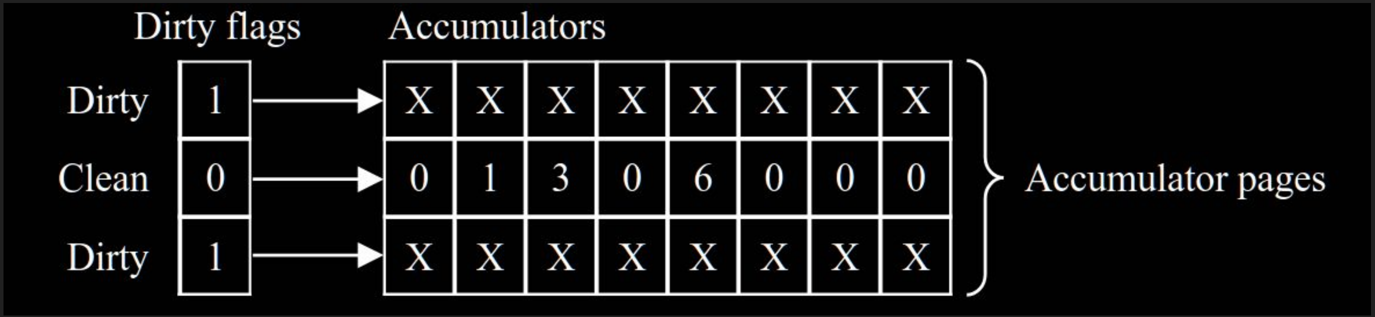
- Compression
- Accumulator Management
- Query Processing
- Early Termination

JASSv2

- Elias Gamma SIMD VB & QMX.
- 2D accumulator array.
- Maintains a heap of the top-k during search, interruptible.
- Process up to ϱ postings.

IOQP

- SIMD BP-128 & StreamVByte.
- Zeroes table at start of each query.
- Uses a heap to find top-k documents at the end of search.
- Process at least ϱ postings.



Preliminary Results

JASSv2 vs IOQP

- Reproducibility study.
- Anserini -> CIFF -> ciffTools -> FGB -> JASSv2/IOQP.
- Collections: MSMARCO, Gov2, Robust04.
- 16-bit accumulators, 8-bit quantization.

Table 3: Mean, median, and 99th percentile latency (ms) and RR@10 scores for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over JASSv2 are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2				IOQP							
	Mean	P_{50}	P_{99}	RR	Mean		P_{50}		P_{99}		RR	
Exhaustive												
BM25	8.2	6.7	28.2	0.188	6.8†	(17.3)	5.6†	(15.4)	21.1†	(25.0)	**0.188**	(0.1)
BM25-T5	33.4	18.7	481.4	**0.274**	16.2†	(51.5)	15.6†	(16.3)	42.9†	(91.1)	**0.274**	(0.0)
DeepCT	3.2	2.9	9.1	0.243	3.1†	(4.0)	2.8†	(4.1)	7.9†	(13.5)	**0.244**	(0.0)
DeepImpact	23.6	25.0	63.7	**0.327**	19.1†	(19.0)	19.5†	(21.9)	51.6†	(18.9)	**0.327**	(0.0)
uniCOIL-TILDE	58.2	48.7	197.3	**0.350**	46.8†	(19.6)	39.4†	(19.1)	156.0†	(20.9)	**0.350**	(0.0)
SPLADEv2	231.2	227.9	443.5	**0.369**	187.2†	(19.0)	184.5†	(19.0)	358.3†	(19.2)	**0.368**	(0.2)
Approximate												
BM25	5.7	6.4	8.2	**0.187**	5.2†	(9.9)	5.7†	(11.6)	7.3†	(9.9)	**0.187**	(0.0)
BM25-T5	4.8†	4.9†	18.1†	0.272	7.1	(−48.8)	5.1	(−4.5)	18.7	(−3.3)	**0.274**	(0.4)
DeepCT	3.2	2.9	7.8	**0.243**	3.0†	(4.4)	2.8†	(4.8)	6.9†	(11.5)	**0.243**	(0.0)
DeepImpact	6.1	6.5	8.3	**0.319**	5.4†	(12.0)	5.6†	(13.0)	7.4†	(11.7)	**0.319**	(0.0)
uniCOIL-TILDE	7.2	7.3	8.8	0.335	6.5†	(9.0)	6.7†	(9.3)	8.1†	(7.9)	**0.336**	(0.4)
SPLADEv2	7.7	7.7	9.2	**0.319**	6.5†	(16.2)	6.5†	(16.4)	7.8†	(14.8)	0.318	(−0.4)

JASSv2 vs IOQP

JASSv2 vs IOQP

- Reproducibility study.
- Anserini -> CIFF -> ciffTools -> FGB -> JASSv2/IOQP.
- Collections: MSMARCO, Gov2, Robust04.
- 16-bit accumulators, 8-bit quantization.
- Our results matched the previous study.
- IOQP outperforms JASSv2 out of the box.

Compression

- JASSv2 – Elias Gamma SIMD VB, QMX
- IOQP – SIMD BP-128 (and StreamVByte)
- Partially replicate previous work:
    - QMX vs SIMD BP-128
    - Elias Gamma SIMD VB vs QMX
- For consistency, we used the original implementations of the algorithms as found in each search engine.

Table 7: Median and 99th percentile latency (ms) for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over EGVB are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2										IOQP							
	EG VB		QMX				BP-128				QMX				BP-128			
	P_{50}	P_{99}	P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}	
Exhaustive																		
BM25	6.7	28.2	5.6	(16.5)	22.5	(20.3)	5.6	(15.8)	22.4	(20.4)	5.5	(17.7)	22.1	(21.4)	5.6	(15.4)	21.1†	(25.0)
BM25-T5	18.7	481.4	16.6	(11.3)	481.0	(0.1)	14.6†	(22.0)	505.0	(−4.9)	27.7	(−48.1)	67.9	(85.9)	15.6	(16.3)	42.9†	(91.1)
DeepCT	2.9	9.1	2.5†	(13.4)	7.6	(16.5)	2.6	(9.2)	7.8	(14.3)	2.6	(10.0)	7.4†	(19.0)	2.8	(4.1)	7.9	(13.5)
DeepImpact	25.0	63.7	19.8	(21.0)	51.5	(19.1)	19.2†	(23.2)	50.1	(21.3)	20.3	(18.8)	50.9	(20.0)	19.5	(21.9)	51.6	(18.9)
uniCOIL-TILDE	48.7	197.3	41.2	(15.3)	168.1	(14.8)	41.7	(14.3)	169.8	(13.9)	37.1	(23.8)	149.7†	(24.1)	39.4	(19.1)	156.0†	(20.9)
SPLADEv2	227.9	443.5	188.1	(17.5)	376.3	(15.2)	181.7	(20.3)	354.6	(20.1)	173.8	(23.7)	336.0†	(24.3)	184.5	(19.0)	358.3	(19.2)
Approximate																		
BM25	6.4	8.2	5.4†	(16.0)	7.0†	(14.3)	5.3†	(16.7)	6.9†	(15.6)	5.6	(13.5)	7.1	(13.0)	5.7	(11.6)	7.3	(9.9)
BM25-T5	4.9	18.1	3.9	(19.1)	17.2†	(4.9)	4.0	(18.2)	17.7	(2.1)	5.4	(−9.8)	42.2	(−133.2)	5.1	(−4.5)	18.7	(−3.3)
DeepCT	2.9	7.8	2.6	(11.9)	6.7	(13.8)	2.6	(10.1)	6.7	(14.2)	2.6	(10.5)	6.4†	(16.9)	2.8	(4.8)	6.9	(11.5)
DeepImpact	6.5	8.3	5.5	(14.3)	7.7	(7.3)	5.3	(18.0)	7.2	(14.2)	5.5	(15.3)	6.9	(16.9)	5.6	(13.0)	7.4	(11.7)
uniCOIL-TILDE	7.3	8.8	6.2	(14.9)	7.6	(13.5)	6.3	(14.4)	7.7	(12.9)	5.9†	(19.4)	7.0†	(20.9)	6.7	(9.3)	8.1	(7.9)
SPLADEv2	7.7	9.2	6.3	(18.5)	7.7	(16.0)	6.6	(14.3)	8.3	(10.0)	6.0†	(21.9)	7.4†	(20.0)	6.5	(16.4)	7.8	(14.8)

Compression // Efficiency

Table 7: Median and 99th percentile latency (ms) for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over EGVB are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2										IOQP							
	EG VB		QMX				BP-128				QMX				BP-128			
	P_{50}	P_{99}	P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}	
Exhaustive																		
BM25	6.7	28.2	5.6	(16.5)	22.5	(20.3)	5.6	(15.8)	22.4	(20.4)	5.5	(17.7)	22.1	(21.4)	5.6	(15.4)	21.1†	(25.0)
BM25-T5	18.7	481.4	16.6	(11.3)	481.0	(0.1)	14.6†	(22.0)	505.0	(−4.9)	27.7	(−48.1)	67.9	(85.9)	15.6	(16.3)	42.9†	(91.1)
DeepCT	2.9	9.1	2.5†	(13.4)	7.6	(16.5)	2.6	(9.2)	7.8	(14.3)	2.6	(10.0)	7.4†	(19.0)	2.8	(4.1)	7.9	(13.5)
DeepImpact	25.0	63.7	19.8	(21.0)	51.5	(19.1)	19.2†	(23.2)	50.1	(21.3)	20.3	(18.8)	50.9	(20.0)	19.5	(21.9)	51.6	(18.9)
uniCOIL-TILDE	48.7	197.3	41.2	(15.3)	168.1	(14.8)	41.7	(14.3)	169.8	(13.9)	37.1	(23.8)	149.7†	(24.1)	39.4	(19.1)	156.0†	(20.9)
SPLADEv2	227.9	443.5	188.1	(17.5)	376.3	(15.2)	181.7	(20.3)	354.6	(20.1)	173.8	(23.7)	336.0†	(24.3)	184.5	(19.0)	358.3	(19.2)
Approximate																		
BM25	6.4	8.2	5.4†	(16.0)	7.0†	(14.3)	5.3†	(16.7)	6.9†	(15.6)	5.6	(13.5)	7.1	(13.0)	5.7	(11.6)	7.3	(9.9)
BM25-T5	4.9	18.1	3.9	(19.1)	17.2†	(4.9)	4.0	(18.2)	17.7	(2.1)	5.4	(−9.8)	42.2	(−133.2)	5.1	(−4.5)	18.7	(−3.3)
DeepCT	2.9	7.8	2.6	(11.9)	6.7	(13.8)	2.6	(10.1)	6.7	(14.2)	2.6	(10.5)	6.4†	(16.9)	2.8	(4.8)	6.9	(11.5)
DeepImpact	6.5	8.3	5.5	(14.3)	7.7	(7.3)	5.3	(18.0)	7.2	(14.2)	5.5	(15.3)	6.9	(16.9)	5.6	(13.0)	7.4	(11.7)
uniCOIL-TILDE	7.3	8.8	6.2	(14.9)	7.6	(13.5)	6.3	(14.4)	7.7	(12.9)	5.9†	(19.4)	7.0†	(20.9)	6.7	(9.3)	8.1	(7.9)
SPLADEv2	7.7	9.2	6.3	(18.5)	7.7	(16.0)	6.6	(14.3)	8.3	(10.0)	6.0†	(21.9)	7.4†	(20.0)	6.5	(16.4)	7.8	(14.8)

Compression // Efficiency

Table 7: Median and 99th percentile latency (ms) for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over EGVB are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2										IOQP							
	EG VB		QMX				BP-128				QMX				BP-128			
	P_{50}	P_{99}	P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}	
Exhaustive																		
BM25	6.7	28.2	5.6	(16.5)	22.5	(20.3)	5.6	(15.8)	22.4	(20.4)	5.5	(17.7)	22.1	(21.4)	5.6	(15.4)	21.1†	(25.0)
BM25-T5	18.7	481.4	16.6	(11.3)	481.0	(0.1)	14.6†	(22.0)	505.0	(−4.9)	27.7	(−48.1)	67.9	(85.9)	15.6	(16.3)	42.9†	(91.1)
DeepCT	2.9	9.1	2.5†	(13.4)	7.6	(16.5)	2.6	(9.2)	7.8	(14.3)	2.6	(10.0)	7.4†	(19.0)	2.8	(4.1)	7.9	(13.5)
DeepImpact	25.0	63.7	19.8	(21.0)	51.5	(19.1)	19.2†	(23.2)	50.1	(21.3)	20.3	(18.8)	50.9	(20.0)	19.5	(21.9)	51.6	(18.9)
uniCOIL-TILDE	48.7	197.3	41.2	(15.3)	168.1	(14.8)	41.7	(14.3)	169.8	(13.9)	37.1	(23.8)	149.7†	(24.1)	39.4	(19.1)	156.0†	(20.9)
SPLADEv2	227.9	443.5	188.1	(17.5)	376.3	(15.2)	181.7	(20.3)	354.6	(20.1)	173.8	(23.7)	336.0†	(24.3)	184.5	(19.0)	358.3	(19.2)
Approximate																		
BM25	6.4	8.2	5.4†	(16.0)	7.0†	(14.3)	5.3†	(16.7)	6.9†	(15.6)	5.6	(13.5)	7.1	(13.0)	5.7	(11.6)	7.3	(9.9)
BM25-T5	4.9	18.1	3.9	(19.1)	17.2†	(4.9)	4.0	(18.2)	17.7	(2.1)	5.4	(−9.8)	42.2	(−133.2)	5.1	(−4.5)	18.7	(−3.3)
DeepCT	2.9	7.8	2.6	(11.9)	6.7	(13.8)	2.6	(10.1)	6.7	(14.2)	2.6	(10.5)	6.4†	(16.9)	2.8	(4.8)	6.9	(11.5)
DeepImpact	6.5	8.3	5.5	(14.3)	7.7	(7.3)	5.3	(18.0)	7.2	(14.2)	5.5	(15.3)	6.9	(16.9)	5.6	(13.0)	7.4	(11.7)
uniCOIL-TILDE	7.3	8.8	6.2	(14.9)	7.6	(13.5)	6.3	(14.4)	7.7	(12.9)	5.9†	(19.4)	7.0†	(20.9)	6.7	(9.3)	8.1	(7.9)
SPLADEv2	7.7	9.2	6.3	(18.5)	7.7	(16.0)	6.6	(14.3)	8.3	(10.0)	6.0†	(21.9)	7.4†	(20.0)	6.5	(16.4)	7.8	(14.8)

Compression // Efficiency

Table 7: Median and 99th percentile latency (ms) for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over EGVB are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2										IOQP							
	EG VB		QMX				BP-128				QMX				BP-128			
	P_{50}	P_{99}	P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}	
Exhaustive																		
BM25	6.7	28.2	5.6	(16.5)	22.5	(20.3)	5.6	(15.8)	22.4	(20.4)	5.5	(17.7)	22.1	(21.4)	5.6	(15.4)	21.1†	(25.0)
BM25-T5	18.7	481.4	16.6	(11.3)	481.0	(0.1)	14.6†	(22.0)	505.0	(−4.9)	27.7	(−48.1)	67.9	(85.9)	15.6	(16.3)	42.9†	(91.1)
DeepCT	2.9	9.1	2.5†	(13.4)	7.6	(16.5)	2.6	(9.2)	7.8	(14.3)	2.6	(10.0)	7.4†	(19.0)	2.8	(4.1)	7.9	(13.5)
DeepImpact	25.0	63.7	19.8	(21.0)	51.5	(19.1)	19.2†	(23.2)	50.1	(21.3)	20.3	(18.8)	50.9	(20.0)	19.5	(21.9)	51.6	(18.9)
uniCOIL-TILDE	48.7	197.3	41.2	(15.3)	168.1	(14.8)	41.7	(14.3)	169.8	(13.9)	37.1	(23.8)	149.7†	(24.1)	39.4	(19.1)	156.0†	(20.9)
SPLADEv2	227.9	443.5	188.1	(17.5)	376.3	(15.2)	181.7	(20.3)	354.6	(20.1)	173.8	(23.7)	336.0†	(24.3)	184.5	(19.0)	358.3	(19.2)
Approximate																		
BM25	6.4	8.2	5.4†	(16.0)	7.0†	(14.3)	5.3†	(16.7)	6.9†	(15.6)	5.6	(13.5)	7.1	(13.0)	5.7	(11.6)	7.3	(9.9)
BM25-T5	4.9	18.1	3.9	(19.1)	17.2†	(4.9)	4.0	(18.2)	17.7	(2.1)	5.4	(−9.8)	42.2	(−133.2)	5.1	(−4.5)	18.7	(−3.3)
DeepCT	2.9	7.8	2.6	(11.9)	6.7	(13.8)	2.6	(10.1)	6.7	(14.2)	2.6	(10.5)	6.4†	(16.9)	2.8	(4.8)	6.9	(11.5)
DeepImpact	6.5	8.3	5.5	(14.3)	7.7	(7.3)	5.3	(18.0)	7.2	(14.2)	5.5	(15.3)	6.9	(16.9)	5.6	(13.0)	7.4	(11.7)
uniCOIL-TILDE	7.3	8.8	6.2	(14.9)	7.6	(13.5)	6.3	(14.4)	7.7	(12.9)	5.9†	(19.4)	7.0†	(20.9)	6.7	(9.3)	8.1	(7.9)
SPLADEv2	7.7	9.2	6.3	(18.5)	7.7	(16.0)	6.6	(14.3)	8.3	(10.0)	6.0†	(21.9)	7.4†	(20.0)	6.5	(16.4)	7.8	(14.8)

Compression // Efficiency

Table 7: Median and 99th percentile latency (ms) for JASSv2 and IOQP on the MSMARCO passage collection. Percentage improvements over EGVB are given in parentheses. † indicates statistically significant improvement over other scores.

Model	JASSv2										IOQP							
	EG VB		QMX				BP-128				QMX				BP-128			
	P_{50}	P_{99}	P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}		P_{50}		P_{99}	
Exhaustive																		
BM25	6.7	28.2	5.6	(16.5)	22.5	(20.3)	5.6	(15.8)	22.4	(20.4)	5.5	(17.7)	22.1	(21.4)	5.6	(15.4)	21.1†	(25.0)
BM25-T5	18.7	481.4	16.6	(11.3)	481.0	(0.1)	14.6†	(22.0)	505.0	(−4.9)	27.7	(−48.1)	67.9	(85.9)	15.6	(16.3)	42.9†	(91.1)
DeepCT	2.9	9.1	2.5†	(13.4)	7.6	(16.5)	2.6	(9.2)	7.8	(14.3)	2.6	(10.0)	7.4†	(19.0)	2.8	(4.1)	7.9	(13.5)
DeepImpact	25.0	63.7	19.8	(21.0)	51.5	(19.1)	19.2†	(23.2)	50.1	(21.3)	20.3	(18.8)	50.9	(20.0)	19.5	(21.9)	51.6	(18.9)
uniCOIL-TILDE	48.7	197.3	41.2	(15.3)	168.1	(14.8)	41.7	(14.3)	169.8	(13.9)	37.1	(23.8)	149.7†	(24.1)	39.4	(19.1)	156.0†	(20.9)
SPLADEv2	227.9	443.5	188.1	(17.5)	376.3	(15.2)	181.7	(20.3)	354.6	(20.1)	173.8	(23.7)	336.0†	(24.3)	184.5	(19.0)	358.3	(19.2)
Approximate																		
BM25	6.4	8.2	5.4†	(16.0)	7.0†	(14.3)	5.3†	(16.7)	6.9†	(15.6)	5.6	(13.5)	7.1	(13.0)	5.7	(11.6)	7.3	(9.9)
BM25-T5	4.9	18.1	3.9	(19.1)	17.2†	(4.9)	4.0	(18.2)	17.7	(2.1)	5.4	(−9.8)	42.2	(−133.2)	5.1	(−4.5)	18.7	(−3.3)
DeepCT	2.9	7.8	2.6	(11.9)	6.7	(13.8)	2.6	(10.1)	6.7	(14.2)	2.6	(10.5)	6.4†	(16.9)	2.8	(4.8)	6.9	(11.5)
DeepImpact	6.5	8.3	5.5	(14.3)	7.7	(7.3)	5.3	(18.0)	7.2	(14.2)	5.5	(15.3)	6.9	(16.9)	5.6	(13.0)	7.4	(11.7)
uniCOIL-TILDE	7.3	8.8	6.2	(14.9)	7.6	(13.5)	6.3	(14.4)	7.7	(12.9)	5.9†	(19.4)	7.0†	(20.9)	6.7	(9.3)	8.1	(7.9)
SPLADEv2	7.7	9.2	6.3	(18.5)	7.7	(16.0)	6.6	(14.3)	8.3	(10.0)	6.0†	(21.9)	7.4†	(20.0)	6.5	(16.4)	7.8	(14.8)

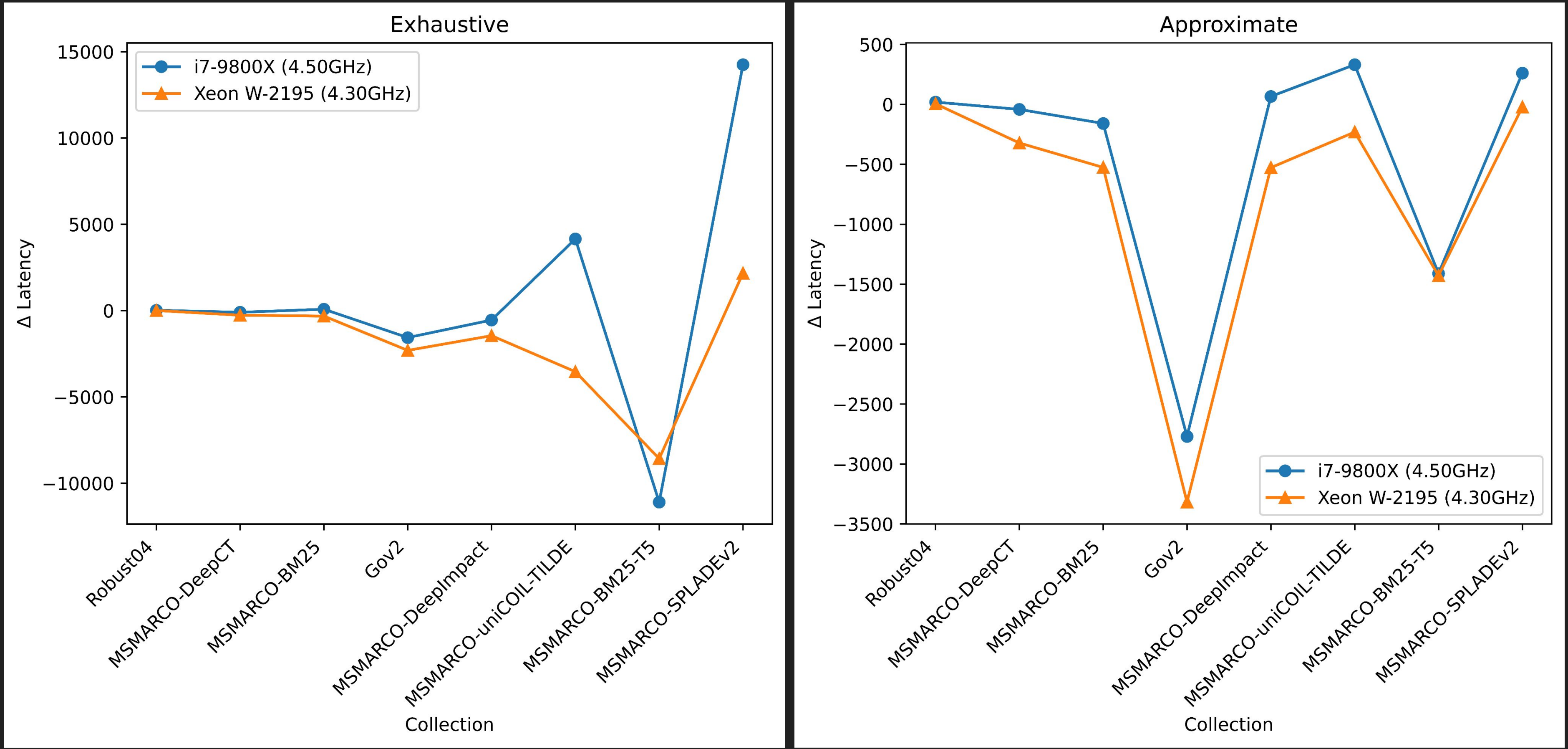
Compression // Efficiency

Compression

- In terms of space, there is no substantial difference between the codecs. But, JASSv2 indexes are always smaller than their IOQP counterparts.
- Elias Gamma SIMD VB is outperformed by QMX and SIMD BP-128.
- We could not determine if QMX or SIMD BP-128 was more efficient.
- IOQP has faster tail latency.
- JASSv2 has faster median latency.

CPU

- Up to this point we used a desktop-grade CPU (Intel i7-9800X/4.50GHz), but prior work used dual server-grade CPUs (Intel Xeon Gold 6144/4.20GHz).
- We introduce a second machine with a server-grade Intel Xeon W-2195 (4.30GHz) to investigate the impact of the CPU.
- We found that the Xeon generally decreased latency. But what of the performance gap?

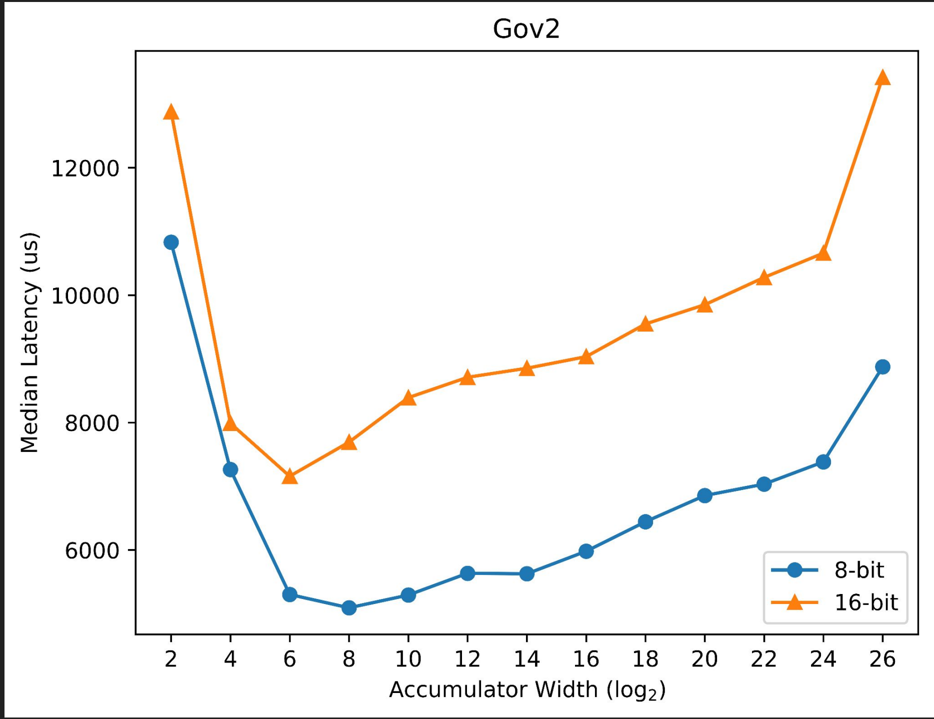


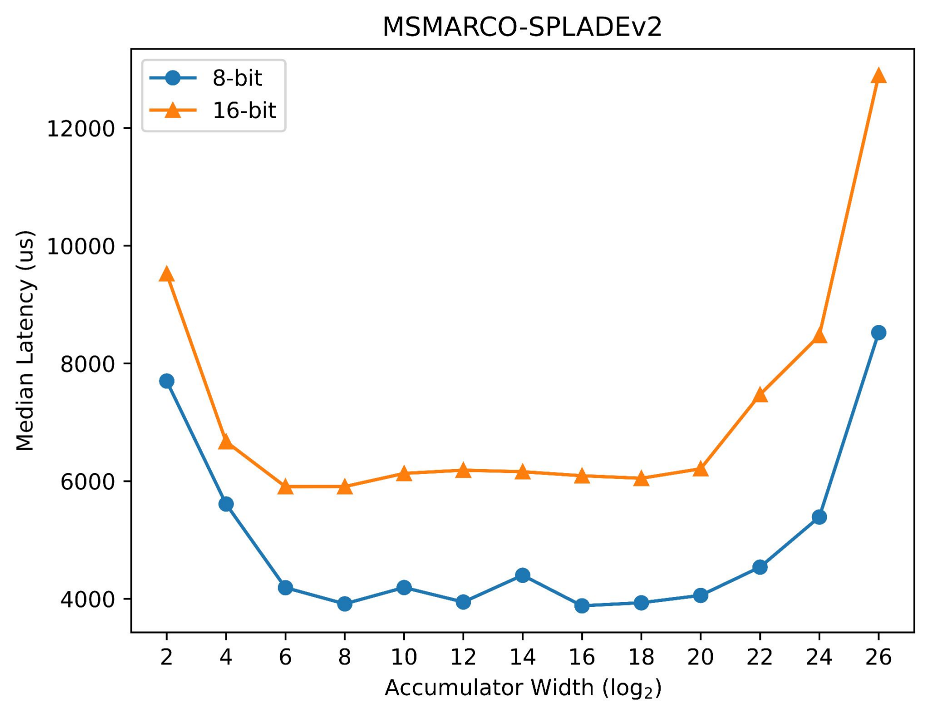
Exhaustive
i7-9800X (4.50GHz)
Xeon W-2195 (4.30GHz)
15000
10000
5000
0
−5000
−10000
Δ Latency
Robust04
MSMARCO-DeepCT
MSMARCO-BM25
Gov2
MSMARCO-DeepImpact
MSMARCO-uniCOIL-TILDE
MSMARCO-BM25-T5
MSMARCO-SPLADEv2
Collection
Approximate
500
0
−500
−1000
−1500
−2000
−2500
−3000
−3500
Δ Latency
i7-9800X (4.50GHz)
Xeon W-2195 (4.30GHz)
Robust04
MSMARCO-DeepCT
MSMARCO-BM25
Gov2
MSMARCO-DeepImpact
MSMARCO-uniCOIL-TILDE
MSMARCO-BM25-T5
MSMARCO-SPLADEv2
Collection

CPU

CPU

- Overall, we found JASSv2 was faster on the Xeon.
- IOQP was typically faster on the i7.
- The query latency is affected by hardware — but the effects are not equal across search engines.





Accumulators // Approximate

Extending the Research

Future Work

- Accumulator management:
  - Is this sensitive to the CPU?
  - When to use 2D array?
  - Finding the ideal width.
  - Exploring other strategies.
- Early termination:
  - A quick examination of the different termination logic.
- Seismic and Block-Max Pruning?
  - Comparing SaaT to other developments in LSR.

Acknowledgements

University of Otago

SIGIR