

# Variable Byte Encoding vs Elias Gamma Encoding

## Ukuran Indexing

Pada slide, dikatakan bahwa (perhatikan, ini merupakan quote dari slide)

“Bit-level coding biasanya menghasilkan hasil kompresi yang lebih baik”. Akan tetapi pada kasus saya, hal ini tidak terjadi, bahkan ukuran indexing cukup berbeda signifikan (walaupun masih termasuk compressed):

- Ukuran normal encoded postings: 20 bytes
- Ukuran VBE encoded postings: 9 bytes
- Ukuran Elias Gamma encoded postings: 16 bytes.

Dapat dilihat, walaupun Elias Gamma merupakan bit level coding, namun compression yang dihasilkan elias gamma ini tidaklah lebih kecil dibandingkan byte level coding seperti VBE Encoding. Hal ini menurut saya terjadi karena implementasi bitstream menjadi byte. Hal ini membuat bitstream yang sebelumnya (56 Bit -> 7 Byte) perlu diubah menjadi byte agar memenuhi kriteria TP ini (karena kita perlu menulis posting list dalam bentuk byte pada InvertedIndexWriter).

## Waktu Indexing

Waktu indexing untuk Variable Byte Encoding yang saya jalankan adalah sekitar **17 menit**. Sedangkan, waktu indexing untuk Elias Gamma Encoding yang saya jalankan adalah sekitar **25 menit**. Hal ini disebabkan karena Elias Gamma membutuhkan lebih banyak langkah untuk encoding dan decoding angka dibandingkan Variable Byte. Misalnya, Elias Gamma melibatkan representasi dalam bentuk unary dan binary, sementara Variable Byte cukup membagi data menjadi byte-byte dan menambahkan flag untuk penanda bit terakhir. Hal ini juga konsisten dengan penjelasan yang diberikan di slide:

“Namun, bit-level coding memerlukan operasi bit manipulation yang banyak (karena terbentur machine word, yang biasanya 8 bit, 16 bit, ...). Jadi, biasanya proses coding jadi lebih lambat.”

## Ukuran File Indices

Berdasarkan gambar yang diberikan dibawah, dapat dilihat bahwa secara keseluruhan VB Encoding memberikan kompresi yang lebih kecil di kasus saya. Walaupun berdasarkan slide Bit level coding seperti Elias Gamma memberikan kompresi yang lebih baik, namun berdasarkan implementasi saya secara keseluruhan VB Encoding masih lebih baik dalam memberikan performa kompresi dalam bentuk ukuran (lebih kecil).

Seperti yang saya jelaskan sebelumnya, hal ini terjadi karena proses konversi bitstream menjadi bytes membuat lebih banyak bytes yang didapatkan dari bit-bit yang diperoleh sebelumnya.

Atas: Elias Gamma  
Bawah: VB Encoding

This PC > DevDrive (D:) > IR > TP1 > index_eg				
Sort View ...				
Name	Date modified	Type	Size	
intermediate_index_12.dict	18/09/2024 22:21	DICT File	374 KB	
intermediate_index_12.index	18/09/2024 22:09	INDEX File	750 KB	
intermediate_index_13.dict	18/09/2024 22:21	DICT File	372 KB	
intermediate_index_13.index	18/09/2024 22:10	INDEX File	756 KB	
intermediate_index_14.dict	18/09/2024 22:21	DICT File	378 KB	
intermediate_index_14.index	18/09/2024 22:11	INDEX File	769 KB	
intermediate_index_15.dict	18/09/2024 22:21	DICT File	385 KB	
intermediate_index_15.index	18/09/2024 22:11	INDEX File	769 KB	
intermediate_index_16.dict	18/09/2024 22:21	DICT File	394 KB	
intermediate_index_16.index	18/09/2024 22:12	INDEX File	771 KB	

Name	Date modified	Type	Size	
intermediate_index_11.dict	16/09/2024 15:34	DICT File	367 KB	
intermediate_index_11.index	16/09/2024 15:23	INDEX File	685 KB	
intermediate_index_12.dict	16/09/2024 15:34	DICT File	374 KB	
intermediate_index_12.index	16/09/2024 15:23	INDEX File	690 KB	
intermediate_index_13.dict	16/09/2024 15:34	DICT File	372 KB	
intermediate_index_13.index	16/09/2024 15:24	INDEX File	699 KB	
intermediate_index_14.dict	16/09/2024 15:34	DICT File	378 KB	
intermediate_index_14.index	16/09/2024 15:24	INDEX File	705 KB	
intermediate_index_15.dict	16/09/2024 15:34	DICT File	385 KB	
intermediate_index_15.index	16/09/2024 15:25	INDEX File	698 KB	
intermediate_index_16.dict	16/09/2024 15:34	DICT File	394 KB	
intermediate_index_16.index	16/09/2024 15:25	INDEX File	699 KB	

## Perbedaan Hasil

Difference Link: <https://www.diffchecker.com/LcFHW1dB/>

Variable Byte Encoding vs Elias Gamma Encoding		Created now	Diff never expires	Clear	Share
2 removals	20 lines	Copy	0 additions	18 lines	Copy
1 arxiv_collections\0\0704.0732.txt					
2 arxiv_collections\0\0704.1383.txt				1 arxiv_collections\0\0704.1383.txt	
3 arxiv_collections\0\0706.0283.txt					
4 arxiv_collections\10\0901.2750.txt				2 arxiv_collections\10\0901.2750.txt	
5 arxiv_collections\11\0904.0465.txt				3 arxiv_collections\11\0904.0465.txt	
6 arxiv_collections\12\0905.3357.txt				4 arxiv_collections\12\0905.3357.txt	
7 arxiv_collections\13\0906.3547.txt				5 arxiv_collections\13\0906.3547.txt	
8 arxiv_collections\14\0909.1074.txt				6 arxiv_collections\14\0909.1074.txt	
9 arxiv_collections\19\1005.4089.txt				7 arxiv_collections\19\1005.4089.txt	
10 arxiv_collections\2\0710.0610.txt				8 arxiv_collections\2\0710.0610.txt	
11 arxiv_collections\22\1010.5513.txt				9 arxiv_collections\22\1010.5513.txt	
12 arxiv_collections\24\1101.5536.txt				10 arxiv_collections\24\1101.5536.txt	
13 arxiv_collections\24\1103.2569.txt				11 arxiv_collections\24\1103.2569.txt	
14 arxiv_collections\3\0712.1770.txt				12 arxiv_collections\3\0712.1770.txt	
15 arxiv_collections\4\0712.3324.txt				13 arxiv_collections\4\0712.3324.txt	
16 arxiv_collections\5\0802.3266.txt				14 arxiv_collections\5\0802.3266.txt	
17 arxiv_collections\5\0802.3642.txt				15 arxiv_collections\5\0802.3642.txt	
18 arxiv_collections\7\0808.3074.txt				16 arxiv_collections\7\0808.3074.txt	
19 arxiv_collections\8\0809.2022.txt				17 arxiv_collections\8\0809.2022.txt	
20 arxiv_collections\8\0809.4335.txt				18 arxiv_collections\8\0809.4335.txt	

### **Variable Byte Encoding (VBE)**

Query : (cosmological AND (quantum OR continuum)) AND geodesics

Results:

arxiv\_collections\0\0704.0732.txt  
arxiv\_collections\0\0704.1383.txt  
arxiv\_collections\0\0706.0283.txt  
arxiv\_collections\10\0901.2750.txt  
arxiv\_collections\11\0904.0465.txt  
arxiv\_collections\12\0905.3357.txt  
arxiv\_collections\13\0906.3547.txt  
arxiv\_collections\14\0909.1074.txt  
arxiv\_collections\19\1005.4089.txt  
arxiv\_collections\2\0710.0610.txt  
arxiv\_collections\22\1010.5513.txt  
arxiv\_collections\24\1101.5536.txt  
arxiv\_collections\24\1103.2569.txt  
arxiv\_collections\3\0712.1770.txt  
arxiv\_collections\4\0712.3324.txt  
arxiv\_collections\5\0802.3266.txt  
arxiv\_collections\5\0802.3642.txt  
arxiv\_collections\7\0808.3074.txt  
arxiv\_collections\8\0809.2022.txt  
arxiv\_collections\8\0809.4335.txt

### **Elias Gamma Encoding**

Query : (cosmological AND (quantum OR continuum)) AND geodesics

Results:

arxiv\_collections\0\0704.1383.txt  
arxiv\_collections\10\0901.2750.txt  
arxiv\_collections\11\0904.0465.txt  
arxiv\_collections\12\0905.3357.txt  
arxiv\_collections\13\0906.3547.txt  
arxiv\_collections\14\0909.1074.txt  
arxiv\_collections\19\1005.4089.txt  
arxiv\_collections\2\0710.0610.txt  
arxiv\_collections\22\1010.5513.txt  
arxiv\_collections\24\1101.5536.txt  
arxiv\_collections\24\1103.2569.txt  
arxiv\_collections\3\0712.1770.txt  
arxiv\_collections\4\0712.3324.txt  
arxiv\_collections\5\0802.3266.txt  
arxiv\_collections\5\0802.3642.txt  
arxiv\_collections\7\0808.3074.txt  
arxiv\_collections\8\0809.2022.txt  
arxiv\_collections\8\0809.4335.txt