Compression Benchmark Results

February 4, 2018

1 Compression

This section shows compression rates achieved, measured in bits per symbol.[1]

File	bcm	bwz	bzip2	gzip	tbcm	tbwz	twt	wt	xz-extreme	xz	zpaq
dickens	1.761	2.113	2.197	3.036	**1.752**	2.109	2.677	2.700	2.222	2.222	1.775
mozilla	2.495	2.939	2.798	2.975	2.450	2.915	3.732	3.903	**2.089**	2.109	2.120
mr	**1.699**	2.083	1.958	2.961	1.699	2.083	2.349	2.345	2.208	2.206	1.795
nci	0.292	0.339	0.432	0.762	**0.275**	0.328	0.542	0.608	0.345	0.424	0.362
ooffice	3.306	3.821	3.722	4.027	3.265	3.797	5.138	5.332	3.156	3.155	**2.931**
osdb	1.784	2.250	2.223	2.966	**1.770**	2.238	2.510	2.952	2.256	2.260	1.886
reymont	1.186	1.470	1.504	2.243	**1.184**	1.470	1.966	1.996	1.588	1.589	1.271
samba	1.488	1.805	1.684	2.021	1.419	1.750	2.445	2.696	1.384	1.402	**1.196**
sao	5.155	5.949	5.450	5.882	5.155	5.949	6.234	6.227	4.882	**4.870**	4.980
webster	1.239	1.481	1.668	2.354	1.236	1.480	2.074	2.075	1.614	1.665	**1.209**
xml	0.590	0.664	0.660	1.035	0.559	0.643	1.163	1.334	0.650	0.678	**0.530**
x-ray	**3.452**	4.244	3.824	5.699	3.452	4.244	4.473	4.470	4.239	4.238	3.560
sources	1.222	1.383	1.493	1.804	1.165	1.349	2.082	2.286	1.184	1.264	**0.989**
pitches	2.664	2.838	2.858	2.422	2.420	2.658	3.624	4.441	1.980	2.080	**1.963**
proteins	2.331	2.289	3.451	3.575	**1.949**	2.003	2.715	3.971	2.222	2.575	2.609
dna	1.720	1.829	2.060	2.249	**1.696**	1.808	2.027	2.047	1.778	1.818	1.859
english.1024MB	1.561	1.837	2.266	3.037	**1.337**	1.658	1.987	2.450	1.925	2.098	1.641
dblp.xml	0.628	0.751	0.911	1.393	0.617	0.744	1.163	1.266	0.819	0.942	**0.605**
Escherichia-Coli	0.796	0.776	2.158	2.312	0.583	0.586	0.918	1.488	**0.368**	0.719	1.941
cere	0.238	0.237	2.017	2.171	0.170	0.168	0.313	0.377	**0.087**	1.863	1.771
coreutils	0.229	0.232	1.281	1.961	0.163	0.172	0.379	0.690	**0.144**	0.934	0.618
einstein.de.txt	0.022	0.015	0.345	2.494	0.015	0.013	0.178	0.399	0.008	0.008	**0.007**
einstein.en.txt	0.015	0.007	0.413	2.806	0.010	0.006	0.233	0.394	0.005	0.005	**0.004**
influenza	0.119	0.120	0.526	0.897	0.112	0.116	0.297	0.314	**0.082**	0.117	0.351
kernel	0.136	0.130	1.738	2.160	0.094	0.095	0.223	0.517	0.064	0.201	**0.058**
para	0.315	0.308	2.091	2.244	0.211	0.209	0.424	0.490	**0.113**	1.925	1.854
world-leaders	0.126	0.121	0.555	1.428	0.105	0.106	0.287	0.396	**0.088**	0.103	0.093

[1] Example: let `f.txt` be a file with size 2 MB, `f.comp` be the compressed file of `f.txt` with size 1 MB, then the respective bits per symbol of the used compression algorithm applied to `f.txt` is $\frac{8 \cdot 1 \text{ MB}}{2 \text{ MB}} = 4$ bps.

2 Coding Speeds

This section shows speeds for encoding and decoding, measured in $\frac{\text{MB}}{\text{s}}$.[2]

2.1 Encoding

File	bcm	bwz	bzip2	gzip	tbcm	tbwz	twt	wt	xz-extreme	xz	zpaq
dickens	6.51	9.16	13.11	**18.30**	4.35	5.45	6.39	12.44	1.61	1.64	0.83
mozilla	7.30	8.59	13.83	**26.10**	4.42	4.86	6.14	14.36	1.94	2.41	0.74
mr	7.36	9.89	18.97	**20.18**	6.25	8.05	10.67	14.38	1.59	2.43	0.83
nci	7.82	17.19	7.74	**72.56**	5.94	9.03	9.27	18.27	1.17	3.33	0.97
ooffice	7.51	7.60	13.30	**20.87**	4.37	4.54	6.23	14.27	2.60	2.66	0.70
osdb	7.28	10.11	13.52	**31.95**	7.17	8.21	10.79	14.33	2.13	2.28	0.77
reymont	7.25	9.56	14.01	**20.32**	5.17	6.37	8.19	14.33	1.67	1.77	0.87
samba	7.65	10.56	13.91	**41.96**	3.90	4.38	5.09	15.96	1.77	3.12	0.83
sao	6.51	6.06	10.78	**18.15**	5.23	4.90	8.22	11.70	2.75	2.61	0.64
webster	6.63	9.90	12.12	**28.83**	4.52	5.99	6.70	12.16	1.40	1.73	0.85
xml	8.62	18.81	9.25	**56.01**	6.13	8.48	9.25	20.30	2.33	4.46	0.90
x-ray	6.90	7.47	14.40	**25.98**	5.76	6.30	9.38	12.41	2.56	2.53	0.66
sources	6.82	9.23	12.50	**35.96**	2.89	3.18	3.48	12.96	1.67	2.49	0.85
pitches	6.73	8.93	10.54	**42.22**	2.47	2.70	2.88	12.79	1.91	2.85	0.75
proteins	5.00	6.86	11.31	**29.51**	1.49	1.58	1.60	7.67	0.97	1.53	0.72
dna	5.83	8.10	**12.91**	10.10	3.11	3.83	4.18	9.78	0.75	1.02	0.96
english.1024MB	5.63	6.41	12.96	**18.56**	1.50	1.54	1.64	9.16	1.13	1.55	0.80
dblp.xml	6.83	11.91	10.39	**53.27**	5.23	6.98	7.40	13.06	1.52	2.91	0.87
Escherichia-Coli	6.33	10.28	**12.83**	9.69	2.04	2.19	2.22	11.01	0.92	1.21	0.94
cere	6.36	11.06	**13.79**	10.02	3.36	3.72	3.67	11.25	1.01	1.07	0.96
coreutils	6.76	12.13	12.24	**35.20**	3.29	3.47	3.52	12.92	2.19	3.41	0.87
einstein.de.txt	7.04	14.59	9.75	**24.77**	8.26	11.02	10.97	14.12	4.86	9.43	0.91
einstein.en.txt	6.65	13.14	9.47	**24.39**	7.18	10.31	10.26	12.78	4.77	9.01	0.90
influenza	6.76	12.64	8.80	**24.80**	4.41	5.99	6.01	12.62	1.76	4.34	0.96
kernel	6.75	12.71	13.75	**32.11**	4.43	4.70	4.74	12.85	2.28	2.98	0.89
para	6.26	10.66	**13.38**	9.68	3.75	4.07	4.01	10.91	1.00	1.03	0.96
world-leaders	8.57	22.05	20.63	**50.27**	6.17	8.27	8.30	22.27	2.19	5.86	0.94

2.2 Decoding

[2]Example: let `f.txt` be a file with size 2 MB, which is encoded/decoded in 1 second. The respective encoding/decoding speed thus is $\frac{2\ \text{MB}}{1\ \text{s}} = 2\ \frac{\text{MB}}{\text{s}}$.

File	bcm	bwz	bzip2	gzip	tbcm	tbwz	twt	wt	xz-extreme	xz	zpaq
dickens	5.71	4.45	34.59	**190.59**	5.19	4.17	3.84	3.99	80.33	80.33	0.81
mozilla	5.66	3.02	34.61	**157.06**	5.78	3.07	3.39	3.19	54.21	54.82	0.72
mr	5.33	3.35	45.06	**186.44**	5.49	3.33	4.46	4.42	59.06	55.60	0.82
nci	6.24	10.45	58.07	**288.27**	6.55	9.60	7.84	7.65	244.26	226.94	0.93
ooffice	5.98	2.36	26.54	**143.10**	5.69	2.39	2.93	2.84	41.61	38.85	0.70
osdb	5.31	3.58	35.49	**157.67**	7.17	3.90	4.38	3.09	59.74	56.24	0.76
reymont	5.90	3.39	39.25	**203.87**	5.53	3.22	4.44	4.44	103.60	103.60	0.86
samba	6.95	4.10	46.72	**204.01**	7.65	4.45	4.54	3.87	89.20	89.20	0.80
sao	4.73	2.06	22.97	**168.68**	4.48	2.00	2.43	2.41	25.52	24.61	0.63
webster	5.79	5.52	39.10	**178.90**	5.46	5.46	4.37	4.42	87.66	87.66	0.81
xml	8.34	12.71	56.01	242.74	11.30	15.40	7.07	5.36	**242.74**	164.44	0.88
x-ray	5.21	2.43	28.76	**113.82**	4.89	2.40	2.87	2.88	26.84	26.84	0.63
sources	6.35	3.87	45.90	**195.05**	6.89	4.14	4.68	4.15	103.60	97.57	0.81
pitches	6.16	4.03	33.67	**156.14**	6.92	4.57	4.23	3.42	55.98	52.66	0.76
proteins	5.46	7.54	25.57	**124.62**	5.85	7.51	4.77	3.96	52.34	43.81	0.74
dna	5.63	7.17	32.58	**169.62**	4.80	6.23	5.16	5.82	75.22	80.40	0.94
english.1024MB	5.20	2.59	32.42	**155.59**	5.51	2.92	4.34	3.84	77.68	72.36	0.79
dblp.xml	6.71	9.23	48.35	**217.07**	7.54	9.40	6.51	5.84	145.50	130.68	0.84
Escherichia-Coli	5.91	8.98	32.26	173.05	7.29	8.74	7.38	5.78	**261.48**	167.65	0.94
cere	6.09	10.75	33.96	174.50	7.73	9.84	7.89	6.71	**477.65**	89.76	0.94
coreutils	5.73	7.08	47.62	188.06	10.72	10.49	8.93	4.61	**415.65**	130.42	0.84
einstein.de.txt	7.89	20.05	61.38	173.11	10.41	15.48	9.54	7.36	675.27	**675.27**	0.85
einstein.en.txt	6.78	14.36	60.17	158.64	7.98	12.65	8.19	7.31	**654.86**	636.18	0.87
influenza	7.33	15.50	55.48	254.10	7.05	12.62	7.84	8.32	**459.92**	408.96	0.94
kernel	5.49	7.95	41.68	178.14	10.84	11.84	9.50	4.49	**511.45**	350.94	0.86
para	6.09	10.52	33.17	172.66	8.15	9.89	8.12	6.47	**425.99**	85.26	0.94
world-leaders	5.72	9.40	75.79	222.84	7.18	9.86	7.36	5.90	492.22	**492.22**	0.92

3 Memory Peaks

This section shows peaks of memory usage during construction, measured in bits per symbol.

3.1 Encoding

File	bcm	bwz	bzip2	gzip	tbcm	tbwz	twt	wt	xz-extreme	xz	zpaq
dickens	42.20	42.27	6.17	**2.83**	99.83	98.33	106.53	43.27	128.95	78.58	86.48
mozilla	40.47	40.48	1.36	**0.55**	88.50	88.59	98.57	40.69	83.73	15.65	17.84
mr	42.38	42.39	6.37	**2.83**	84.51	84.45	93.65	43.43	130.33	80.18	88.30
nci	40.69	40.70	1.93	**0.87**	40.74	40.70	41.05	41.07	89.40	23.88	27.17
ooffice	43.60	43.62	10.90	**4.54**	115.95	115.87	131.11	45.58	167.17	99.56	137.67
osdb	42.20	42.31	6.26	**2.90**	80.51	80.25	88.53	43.40	131.14	79.29	85.76
reymont	43.46	43.37	9.59	**4.23**	84.46	83.56	85.19	45.11	161.25	99.88	127.86
samba	41.01	41.01	2.99	**1.26**	66.64	67.19	74.51	41.60	99.53	37.09	42.33
sao	43.14	43.21	8.76	**3.95**	157.95	157.16	172.56	44.65	154.14	97.46	119.14
webster	40.58	40.59	1.56	**0.66**	74.38	74.54	81.74	40.83	86.48	19.33	22.03
xml	44.35	44.51	12.97	**5.23**	49.95	52.20	56.12	46.40	182.16	103.94	155.35
x-ray	42.74	42.72	7.55	**3.33**	120.75	120.65	133.20	43.98	141.17	92.62	101.89
sources	40.11	40.11	0.33	**0.13**	63.28	63.25	70.00	40.16	26.87	3.80	4.33
pitches	40.41	40.41	1.24	**0.50**	95.18	94.87	102.92	40.60	82.33	14.36	16.38
proteins	40.02	40.01	0.05	**0.02**	86.75	86.76	92.29	40.02	4.78	0.67	0.76
dna	40.06	40.06	0.15	**0.07**	124.34	124.30	128.56	40.08	14.02	1.98	2.21
english.1024MB	40.02	40.02	0.05	**0.02**	80.83	80.84	86.42	40.03	5.27	0.74	0.85
dblp.xml	40.08	40.07	0.21	**0.09**	48.70	48.72	54.08	40.11	19.13	2.70	3.08
Escherichia-Coli	40.22	40.20	0.57	**0.25**	48.19	48.18	50.05	40.31	50.28	7.11	8.06
cere	40.05	40.05	0.13	**0.06**	40.05	40.05	40.07	40.07	12.28	1.73	1.94
coreutils	40.12	40.11	0.34	**0.13**	40.12	40.11	40.16	40.17	27.60	3.90	4.45
einstein.de.txt	40.26	40.26	0.72	**0.29**	40.26	40.25	40.37	40.37	60.53	8.63	9.85
einstein.en.txt	40.05	40.04	0.14	**0.06**	40.05	40.05	40.07	40.07	12.11	1.71	1.95
influenza	40.15	40.15	0.44	**0.18**	40.15	40.16	40.23	40.22	36.60	5.17	5.81
kernel	40.09	40.09	0.24	**0.10**	40.09	40.09	40.13	40.13	21.96	3.10	3.54
para	40.05	40.05	0.14	**0.06**	40.05	40.05	40.08	40.08	13.20	1.86	2.09
world-leaders	40.49	40.50	1.33	**0.58**	40.52	40.53	40.73	40.74	84.48	17.05	19.46

3.2 Decoding

File	bcm	bwz	bzip2	gzip	tbcm	tbwz	twt	wt	xz-extreme	xz	zpaq
dickens	42.18	42.08	3.95	**2.85**	43.71	43.53	44.60	43.03	10.84	8.62	84.34
mozilla	40.45	40.41	0.77	**0.56**	39.09	39.07	39.29	40.65	8.69	1.71	17.81
mr	42.13	42.20	4.01	**2.87**	44.30	44.24	45.35	43.37	10.07	8.86	88.22
nci	40.62	40.67	1.21	**0.84**	35.92	35.94	36.23	40.91	8.65	2.61	27.02
ooffice	43.47	43.36	6.56	**4.55**	43.58	43.05	45.31	45.17	13.26	12.29	137.52
osdb	42.25	42.03	3.98	**2.80**	30.67	30.68	31.90	43.30	11.44	8.77	87.27
reymont	43.26	43.09	5.95	**4.29**	44.40	44.16	45.85	44.92	12.84	11.11	127.66
samba	41.04	40.94	1.85	**1.35**	34.37	34.27	34.84	41.54	8.97	4.08	42.25
sao	43.17	42.98	5.56	**4.05**	45.37	45.04	46.59	44.48	11.87	11.58	118.92
webster	40.49	40.51	0.96	**0.67**	41.76	41.75	42.04	40.79	8.74	2.11	22.00
xml	44.16	44.02	7.33	**5.44**	34.06	34.20	36.06	46.17	13.55	14.28	158.32
x-ray	42.50	42.38	4.74	**3.32**	44.73	44.67	46.09	43.78	10.87	10.41	101.80
sources	40.10	40.10	0.19	**0.13**	34.89	34.88	34.98	40.15	2.64	0.41	4.33
pitches	40.39	40.38	0.72	**0.50**	31.59	31.57	31.78	40.58	8.39	1.58	16.35
proteins	40.01	40.01	0.03	**0.02**	28.38	28.38	28.43	40.02	0.47	0.07	0.76
dna	40.05	40.05	0.09	**0.07**	41.27	41.26	41.31	40.07	1.38	0.21	2.23
english.1024MB	40.02	40.02	0.03	**0.02**	33.20	33.20	33.21	40.03	0.51	0.08	0.85
dblp.xml	40.07	40.07	0.13	**0.09**	32.98	32.99	33.03	40.11	1.88	0.30	3.08
Escherichia-Coli	40.20	40.18	0.35	**0.24**	17.09	17.08	17.22	40.28	4.95	0.77	8.01
cere	40.04	40.04	0.08	**0.06**	16.64	16.64	16.67	40.06	1.21	0.19	1.92
coreutils	40.10	40.11	0.19	**0.14**	11.59	11.58	11.64	40.15	2.71	0.42	4.44
einstein.de.txt	40.24	40.23	0.42	**0.30**	17.90	17.87	18.00	40.35	6.01	0.95	9.84
einstein.en.txt	40.04	40.04	0.08	**0.06**	24.53	24.53	24.56	40.06	1.19	0.18	1.95
influenza	40.13	40.13	0.25	**0.17**	36.72	36.71	36.79	40.20	3.60	0.56	5.88
kernel	40.09	40.08	0.15	**0.10**	9.29	9.29	9.33	40.13	2.16	0.34	3.53
para	40.04	40.04	0.09	**0.06**	12.56	12.56	12.60	40.07	1.30	0.20	2.09
world-leaders	40.46	40.43	0.81	**0.61**	24.42	24.38	24.66	40.69	8.75	1.89	19.45

4 Tunneled BWT estimator quality

This section shows quality measurements for estimations done in the tunneled BWT. To this end, the gross-net-benefit ratio is defined as the fraction of the gross benefit (benefit not including the aux-encoding) and net benefit (benefit minus size of the encoding for aux). Use is motivated by two facts:

- Small variations in the aux-encoding can be tolerated if the benefit is orders of magnitude bigger.
- The ratio is not affected by the efficiency of a compressor, making it nicely comparable.

To get a measure how well the theoretical model fits onto the given compressor, the model-fit is defined as the minimax-distance of the gross-net-benefit ratio of both model and compressor, i.e. $\frac{\min\{\widehat{\mathsf{gnb}},\mathsf{gnb}\}}{\max\{\widehat{\mathsf{gnb}},\mathsf{gnb}\}}$.

Furthermore the tunneling potential is defined as the theoretical gross-net-benefit divided by the size of the same `bwz`-encoded text, indicating how much benefit theoretically should be possible.

File	tunnelingpotential	bcmzip-model-fit	bwzip-model-fit	wtzip-model-fit
dickens	0.41 %	98.99 %	76.34 %	83.48 %
mozilla	1.93 %	89.37 %	80.59 %	75.17 %
mr	0.01 %	0.00 %	0.00 %	0.00 %
nci	6.15 %	94.75 %	88.89 %	83.45 %
ooffice	1.58 %	93.49 %	75.59 %	73.98 %
osdb	0.51 %	98.93 %	99.32 %	95.44 %
reymont	0.22 %	91.31 %	26.58 %	68.43 %
samba	5.26 %	96.23 %	91.16 %	84.51 %
sao	0.00 %	0.00 %	0.00 %	0.00 %
webster	0.29 %	97.99 %	35.83 %	17.51 %
xml	5.87 %	91.22 %	90.89 %	77.11 %
x-ray	0.00 %	0.00 %	0.00 %	0.00 %
sources	5.55 %	89.72 %	82.75 %	75.19 %
pitches	10.30 %	94.50 %	93.01 %	86.02 %
proteins	15.52 %	95.35 %	99.29 %	91.84 %
dna	1.63 %	96.71 %	89.40 %	62.66 %
english.1024MB	11.54 %	99.92 %	98.70 %	99.88 %
dblp.xml	2.45 %	92.72 %	79.63 %	65.08 %
Escherichia-Coli	31.18 %	96.38 %	97.99 %	91.91 %
cere	30.04 %	97.90 %	99.72 %	97.42 %
coreutils	34.35 %	99.68 %	98.58 %	96.83 %
einstein.de.txt	20.53 %	92.88 %	98.88 %	89.64 %
einstein.en.txt	17.04 %	91.60 %	97.92 %	89.51 %
influenza	7.82 %	94.13 %	76.72 %	83.04 %
kernel	34.71 %	99.84 %	98.45 %	97.28 %
para	38.81 %	98.13 %	98.42 %	89.48 %
world-leaders	18.37 %	92.76 %	95.56 %	83.11 %

5 System Information

Following list shows information about processor, installed memory and operating system.

CPU	`Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz`
Memory	`16313516 kB`
OS Info	`Linux #121~14.04.1-Ubuntu SMP Wed Oct 11 11:54:55 UTC 2017 x86_64`