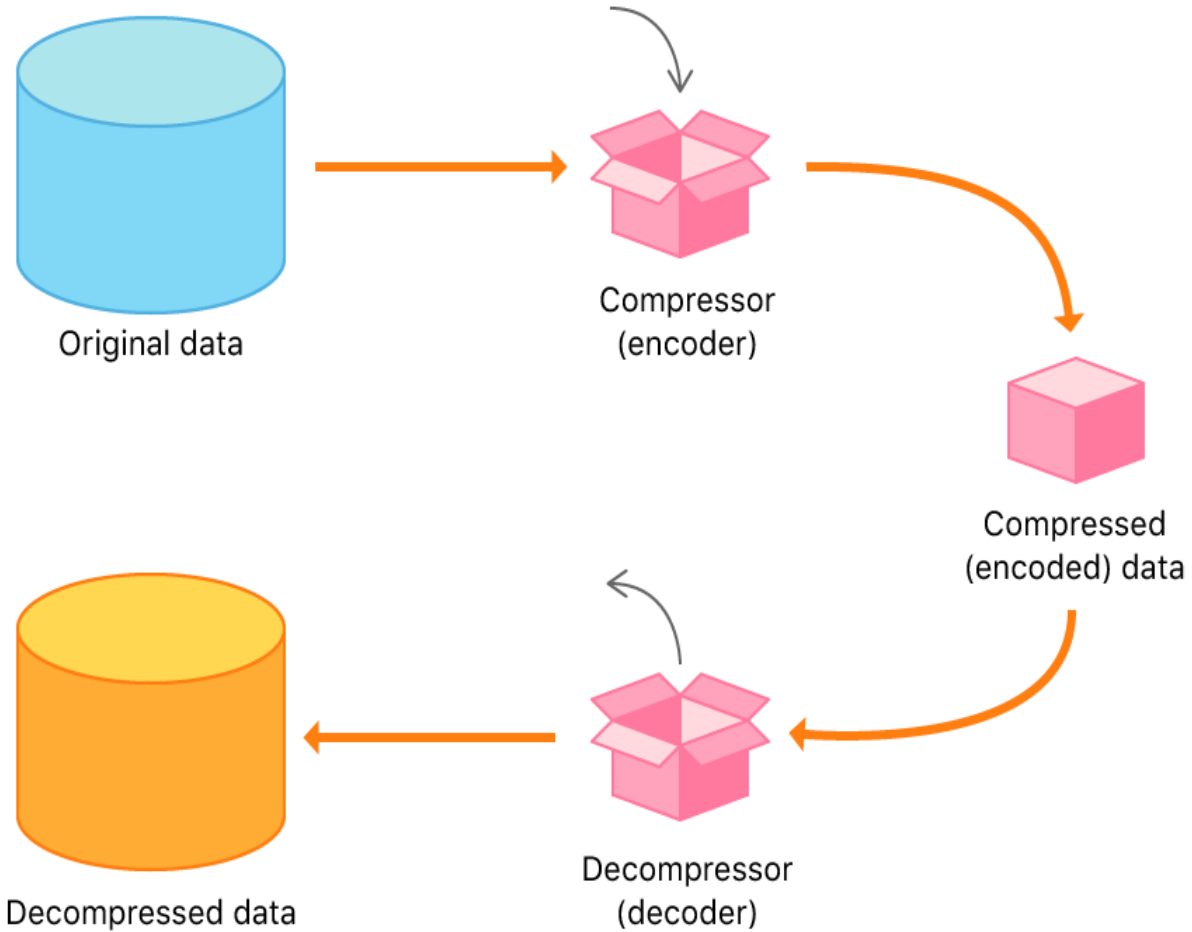
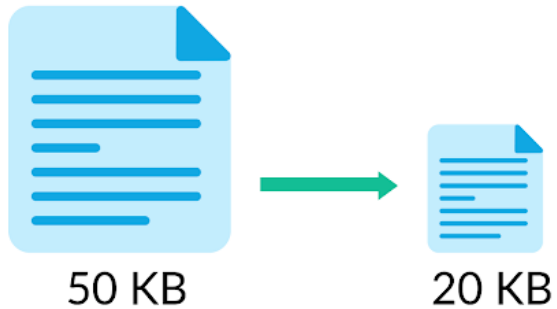
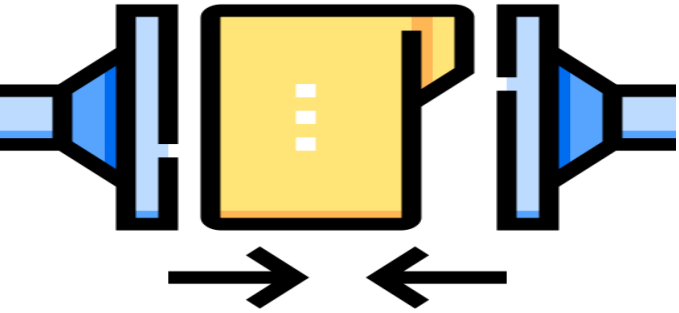


COMPRESSION ALGORITHMS WITH NEAR SORTEDNESS

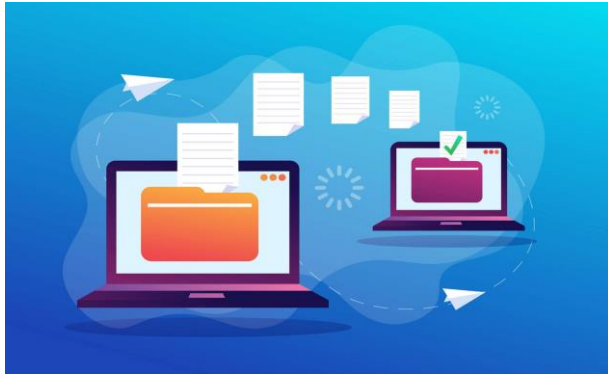
HARSHITHA TUMKUR KAILASA MURTHY - U00683580
VISHWAS BHAKTAVATSALA - U74206902

WHAT IS COMPRESSION?



WHY IS COMPRESSION IMPORTANT ?

REDUCES THE FILE SPACE ON HARD DRIVE

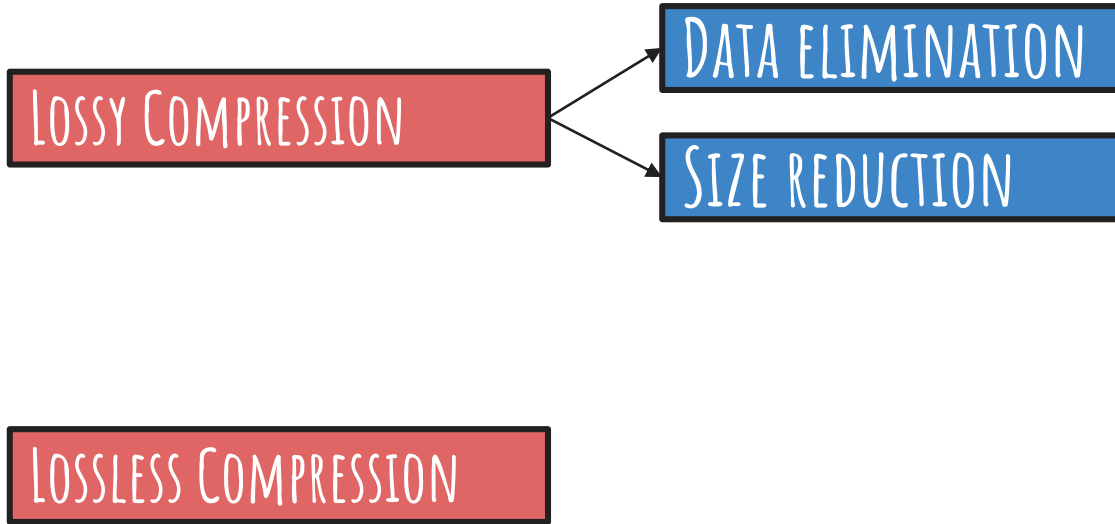


REDUCES FILE TRANSFER TIME

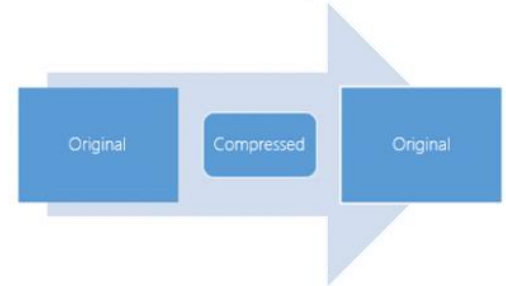
REDUCES REDUNDANCY



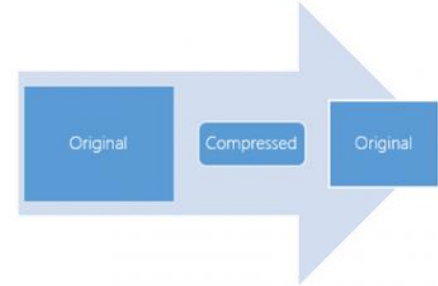
TYPES OF COMPRESSION



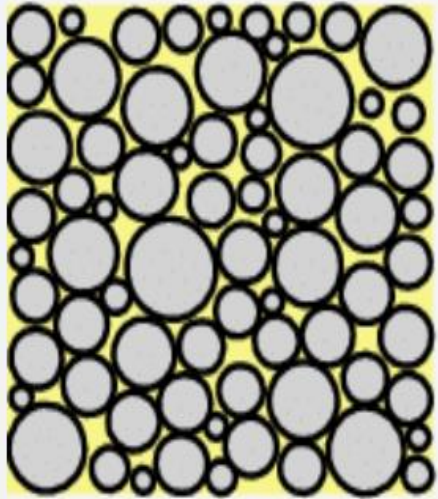
Lossless Compression



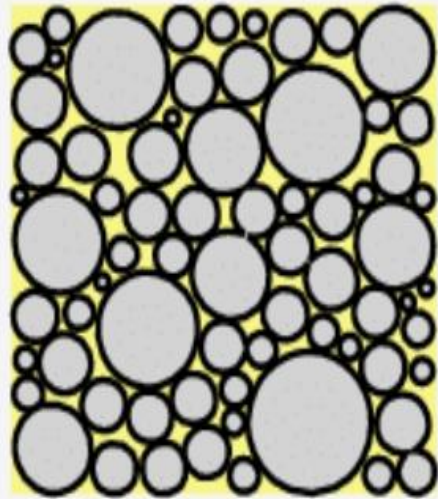
Lossy Compression



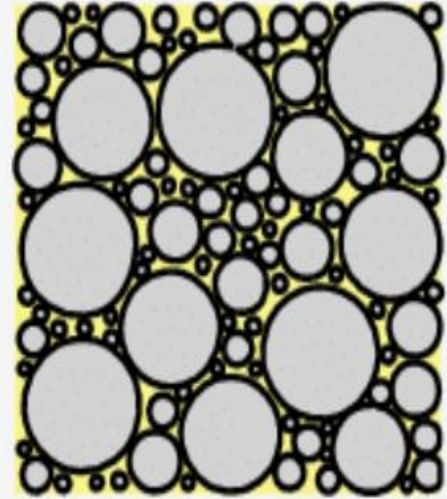
WHAT IS SORTEDNESS?



well sorted



moderately sorted



poorly sorted

ALGORITHMS WE CHOSE FOR OUR RESEARCH

1. RUN LENGTH ENCODING
2. DELTA COMPRESSION
3. ZSTANDARD
4. SNAPPY COMPRESSION
5. LZ77

RUN LENGTH ENCODING

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 4 | 0 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|



RUN LENGTH ENCODER



| | | | | | |
|----|----|----|----|----|----|
| 13 | 24 | 32 | 41 | 01 | 41 |
|----|----|----|----|----|----|

DELTA COMPRESSION ALGORITHM

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 56030 | 56031 | 56044 | 56040 | 56000 | 56003 | 56004 | 56005 | 56012 | 56007 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|

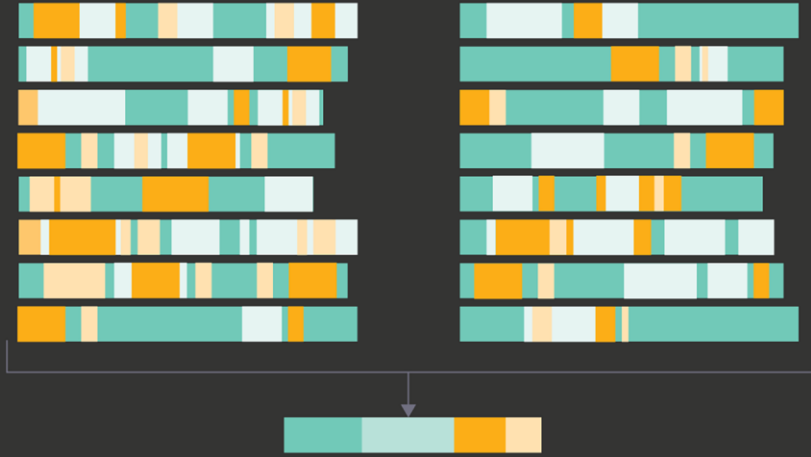
(MEMORY - 40 BYTES)

Delta Compression

| | | | | | | | | | |
|-------|---|----|----|-----|---|---|---|---|----|
| 56030 | 1 | 13 | -4 | -40 | 3 | 1 | 1 | 7 | -5 |
|-------|---|----|----|-----|---|---|---|---|----|

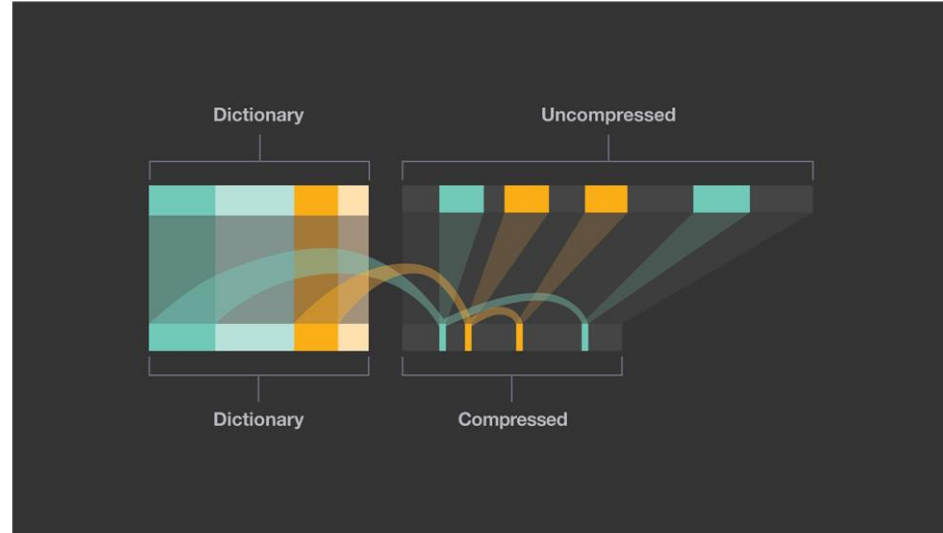
(MEMORY - 13 BYTES)

META'S ZSTANDARD ALGORITHM



DICTIONARY

COMPRESSION



LZ77 & SNAPPY COMPRESSION

| | | | | | | | | | | | |
|---|---|----|----|----|----|---|---|----|----|----|----|
| 4 | 8 | 15 | 16 | 23 | 42 | 4 | 8 | 15 | 16 | 23 | 42 |
|---|---|----|----|----|----|---|---|----|----|----|----|

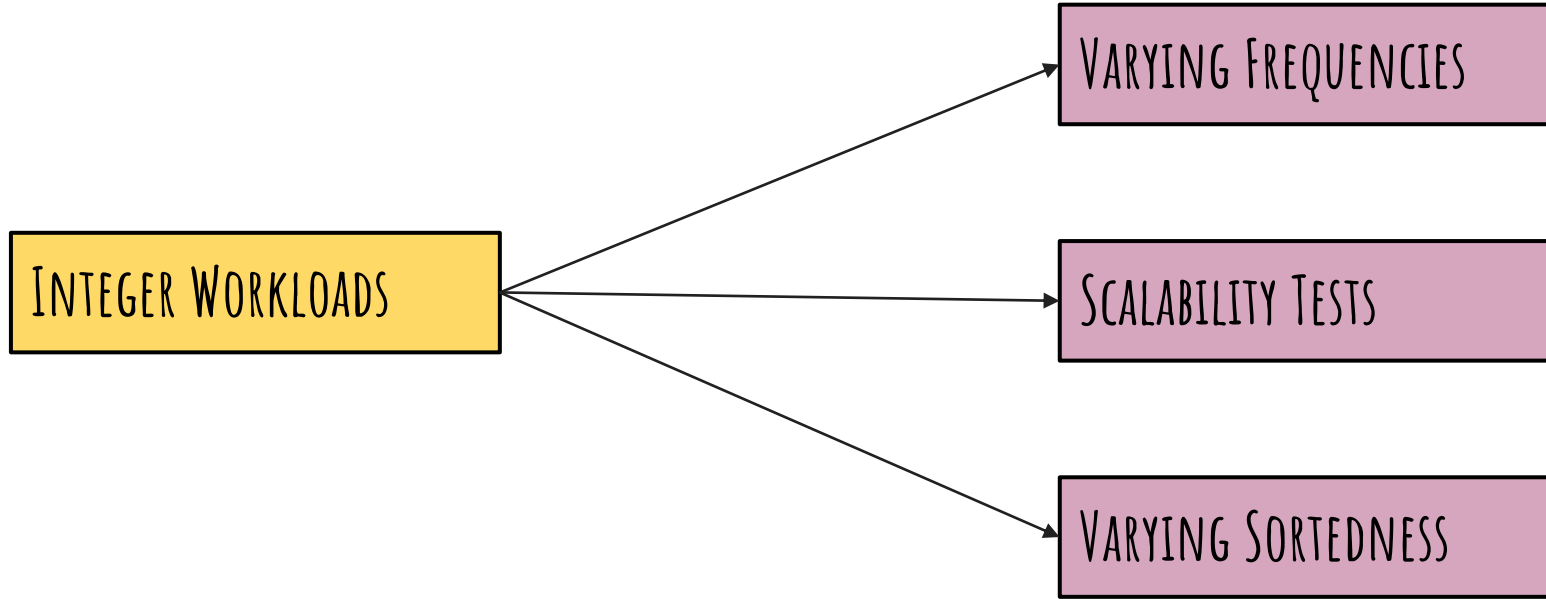


LZ77 Compression



| | | | | | | | | | | | |
|---|---|----|----|----|----|---|---|----|----|----|----|
| 4 | 8 | 15 | 16 | 23 | 42 | 4 | 8 | 15 | 16 | 23 | 42 |
| 0 | 0 | 0 | 0 | 0 | 0 | 6 | 6 | 6 | 6 | 6 | 6 |

EVALUATIONS



TYPES OF WORKLOADS CONSIDERED FOR OUR EXPERIMENTS

SCALABILITY TESTS

400KB - 400 MB with 10X increases in workloads.

NEAR SORTEDNESS TESTS

Ranges are referred from BoDS research paper.

Workload Length: 5M

Workload Size ~40Mb

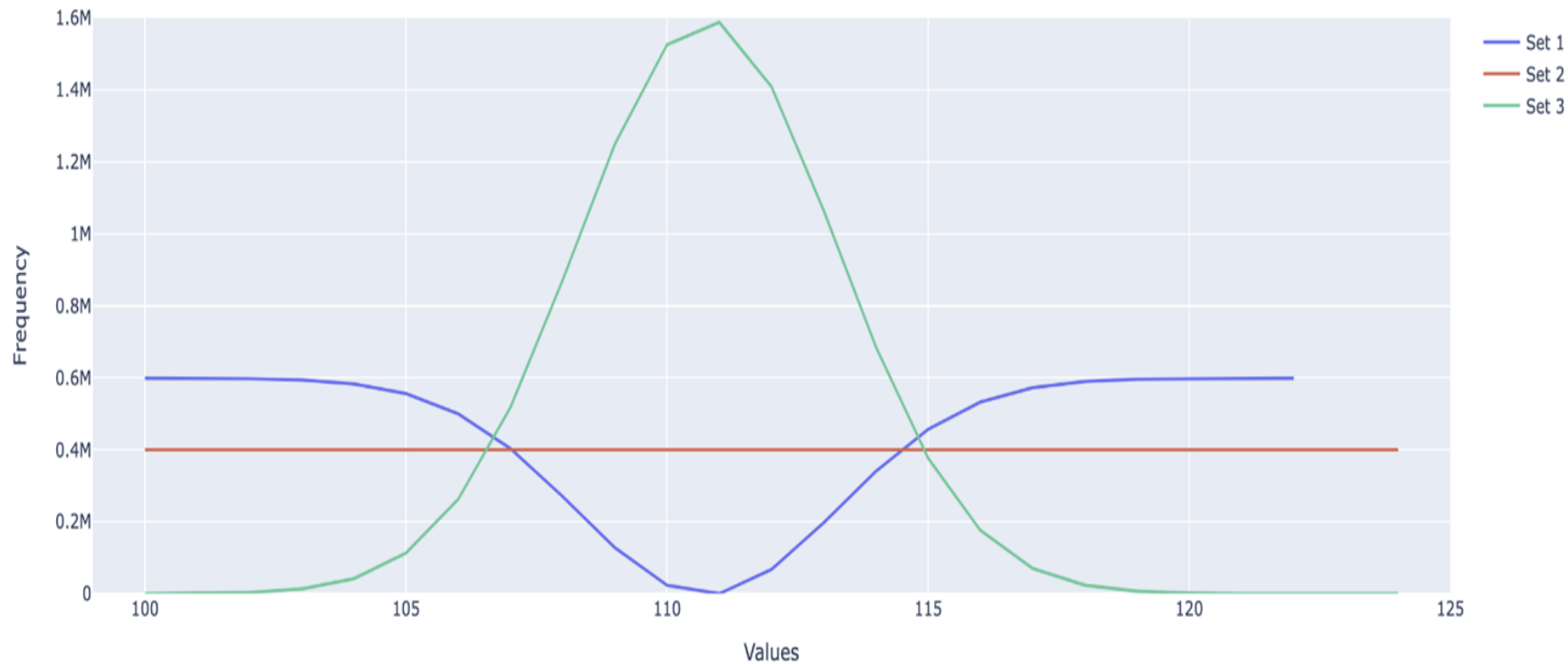
FREQUENCY TESTS

0.2M - 1.6M data with increase of 0.2M on each workload.

Workload Length: 10M

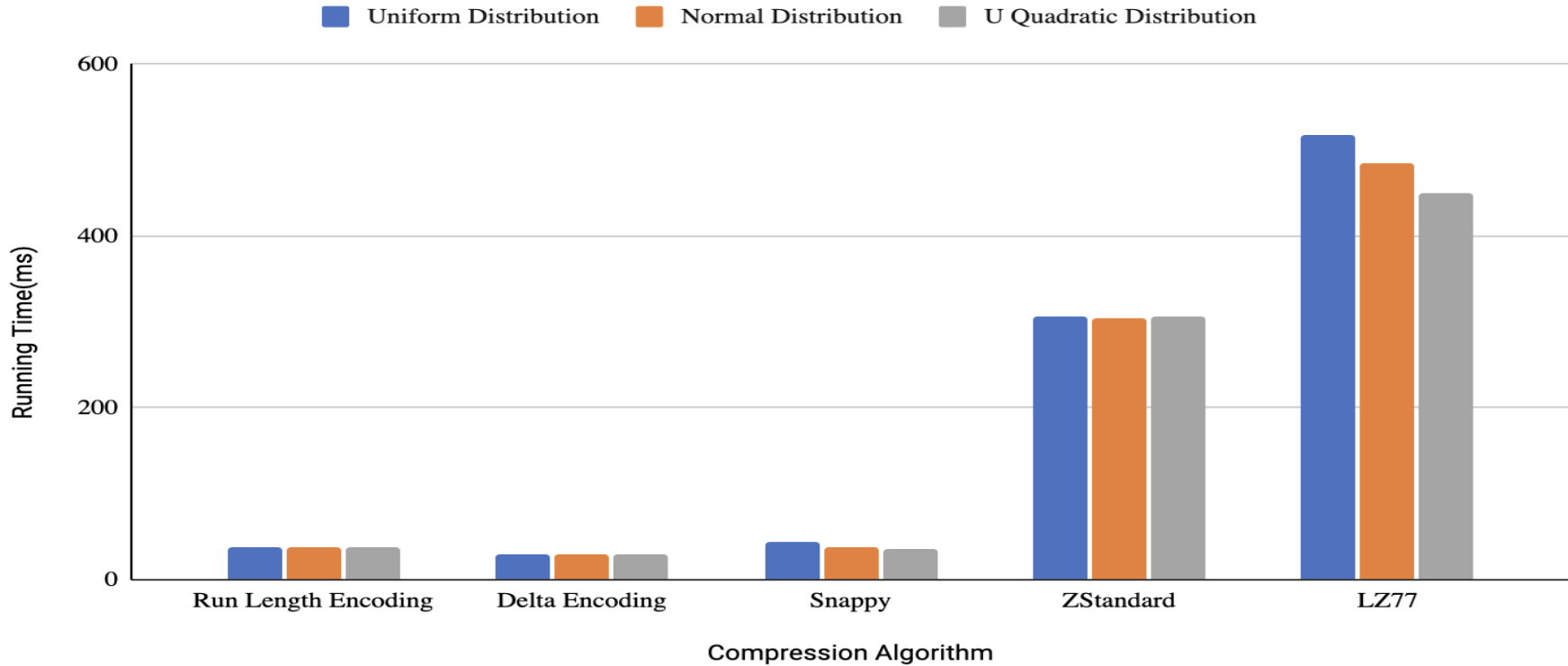
Workload Size: 40Mb

VARYING FREQUENCY



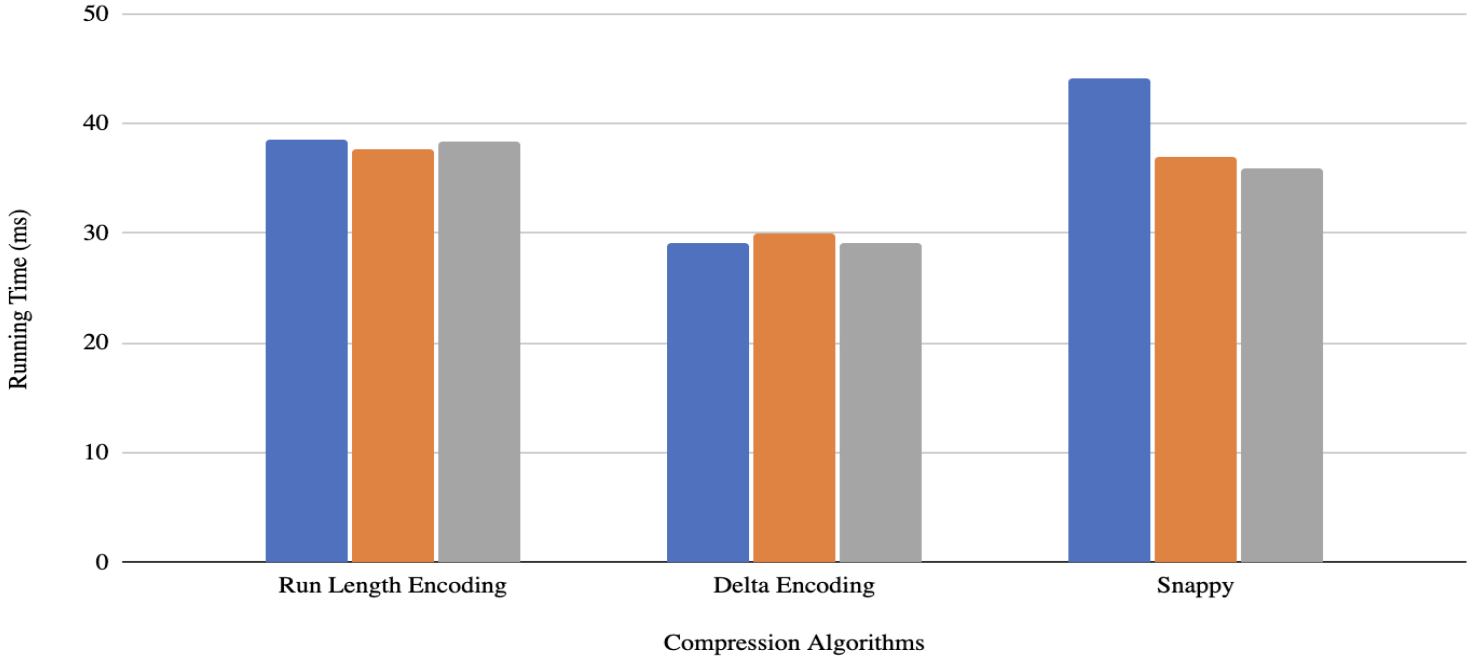
VARYING FREQUENCY

Running times for varying workload distributions



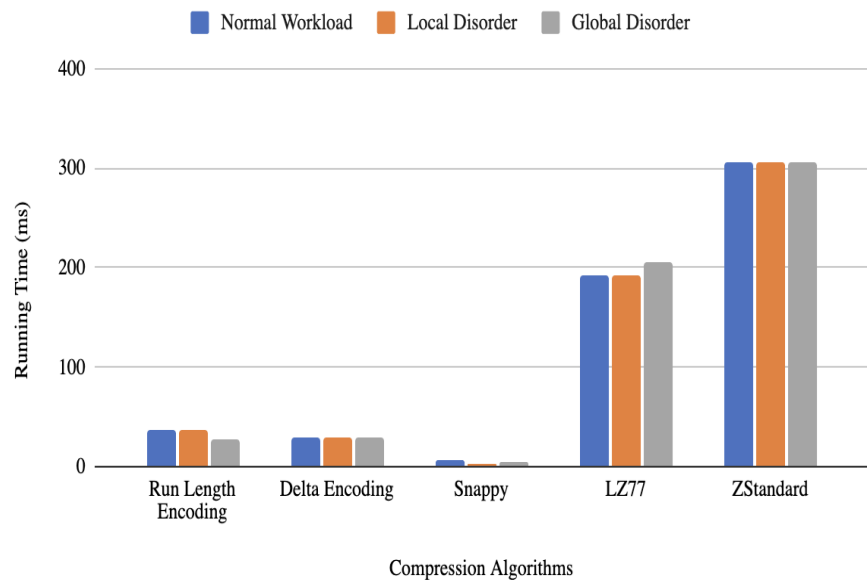
Running times for varying workload distributions

Uniform Distribution Normal Distribution U Quadratic Distribution

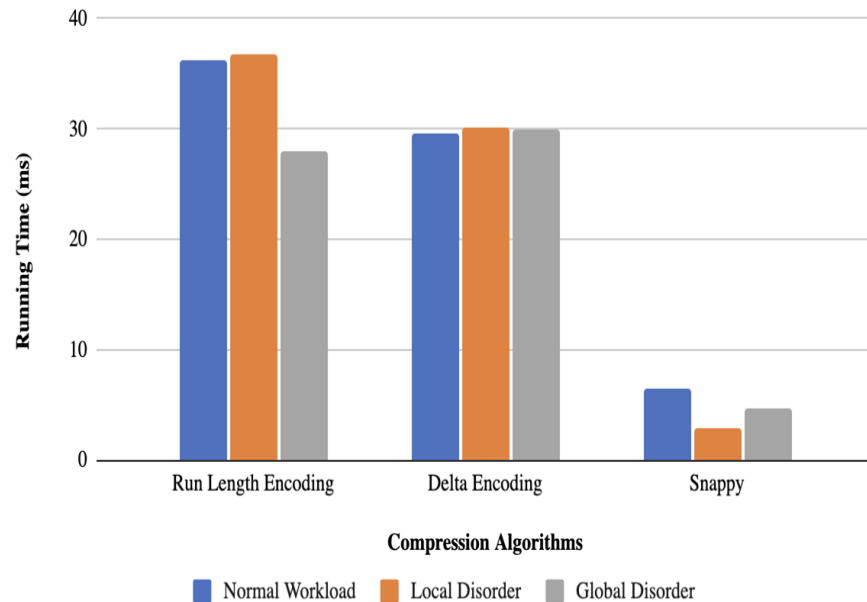


VARYING FREQUENCY

Normal Workload, Local Disorder and Global Disorder

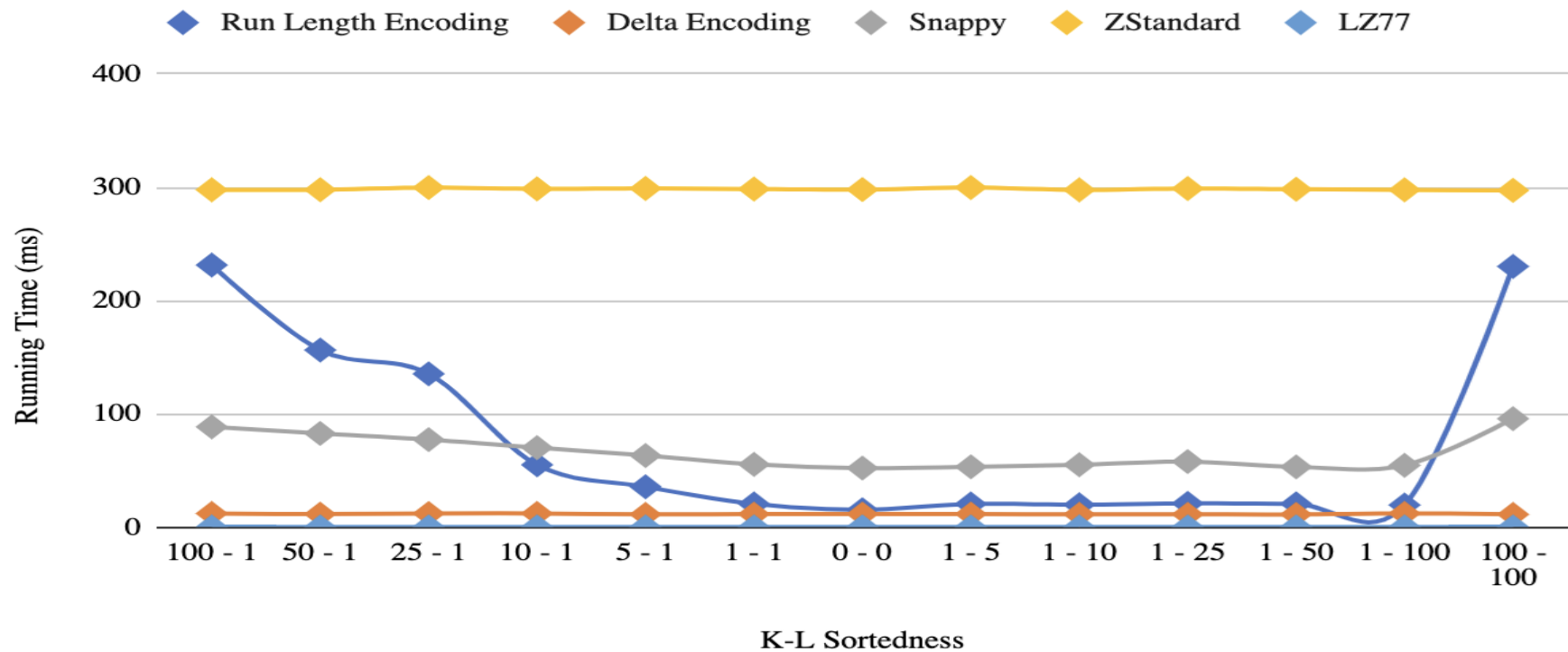


Normal Workload, Local Disorder and Global Disorder

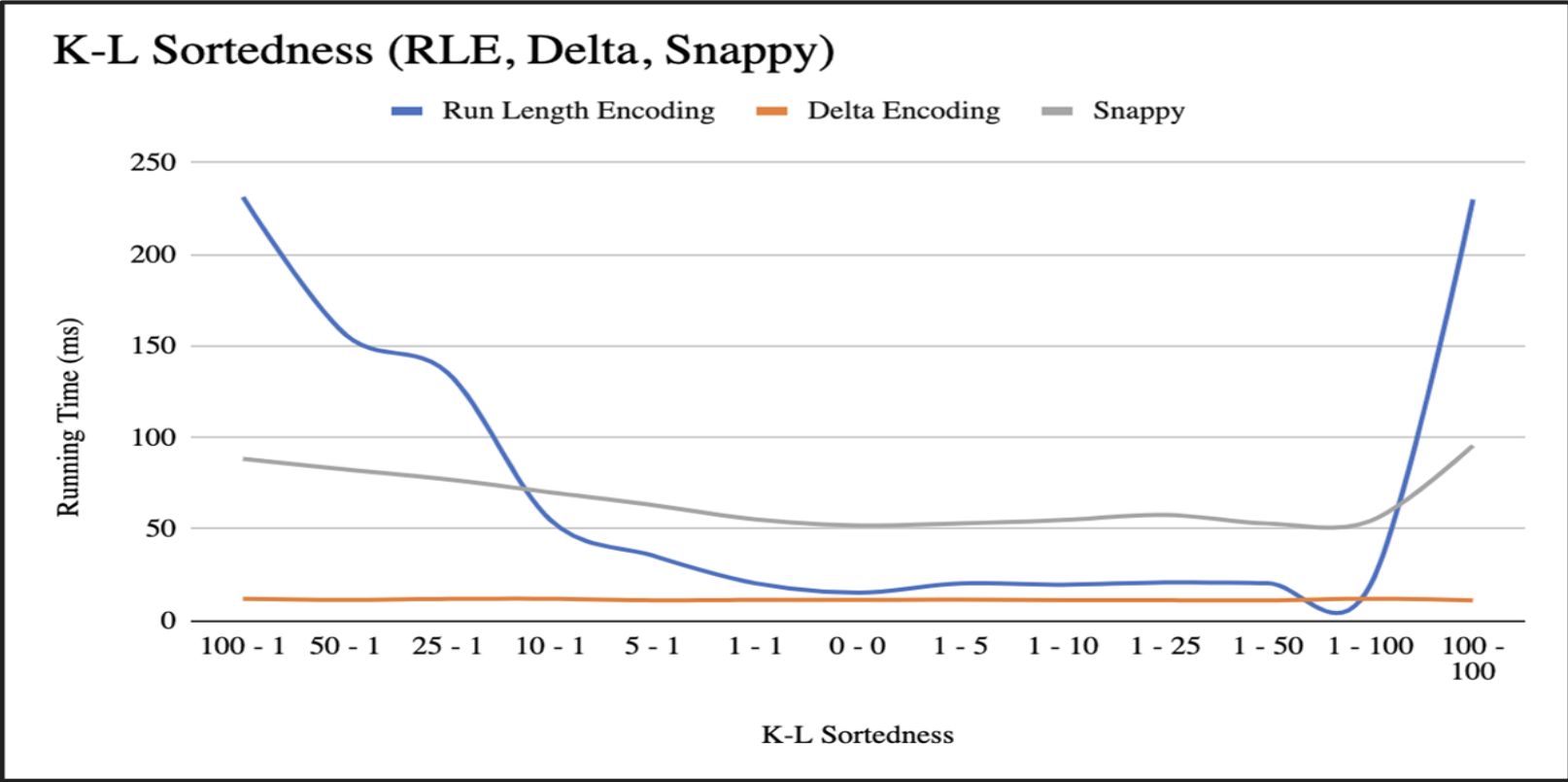


SORTEDNESS EXPERIMENTS

K-L Sortedness Benchmark



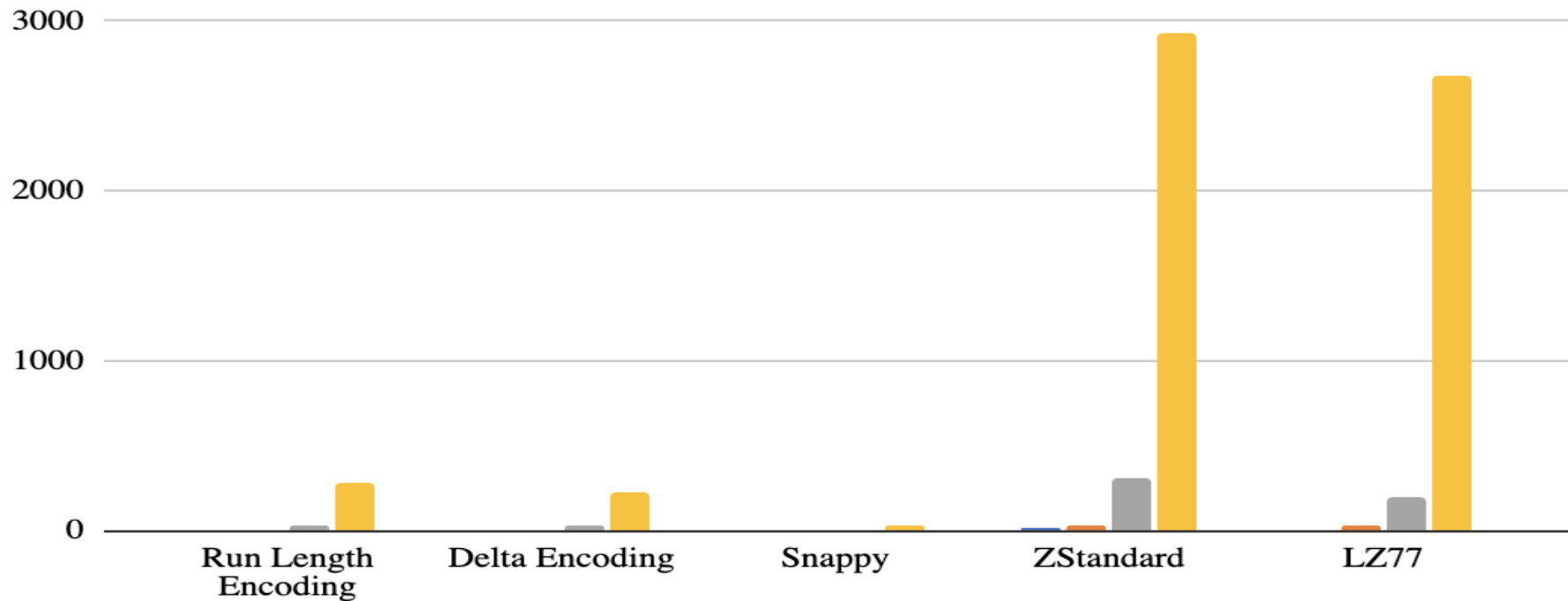
SORTEDNESS EXPERIMENTS



SCALABILITY EXPERIMENTS

Scalability Tests - Running Time

400 Kb 4 Mb 40 Mb 400 Mb



CONCLUSION AND FUTURE WORK

- ❑ DELTA COMPRESSION IS ONE AMONG THE BEST COMPRESSION ALGORITHMS FOR REDUNDANT AS WELL AS NEAR SORTED DATA BUT THE COMPRESSION RATIO TAKES A HIT.
- ❑ GOOGLE'S SNAPPY ALGORITHM IS FAST AND EFFICIENT AND HAS A BETTER COMPRESSION THAN DELTA.
- ❑ OVERALL, SNAPPY COMPRESSION HAS BETTER PERFORMANCE AND METRICS THAN OTHER COMPRESSION ALGORITHMS IN FOCUS
- ❑ AS PART OF THE FUTURE WORK, WE CAN PERFORM MORE INTRICATE EXPERIMENTS AND RESEARCH NEW ALGORITHMS LIKE LZ4, HUFFMAN ETC.

THANK YOU