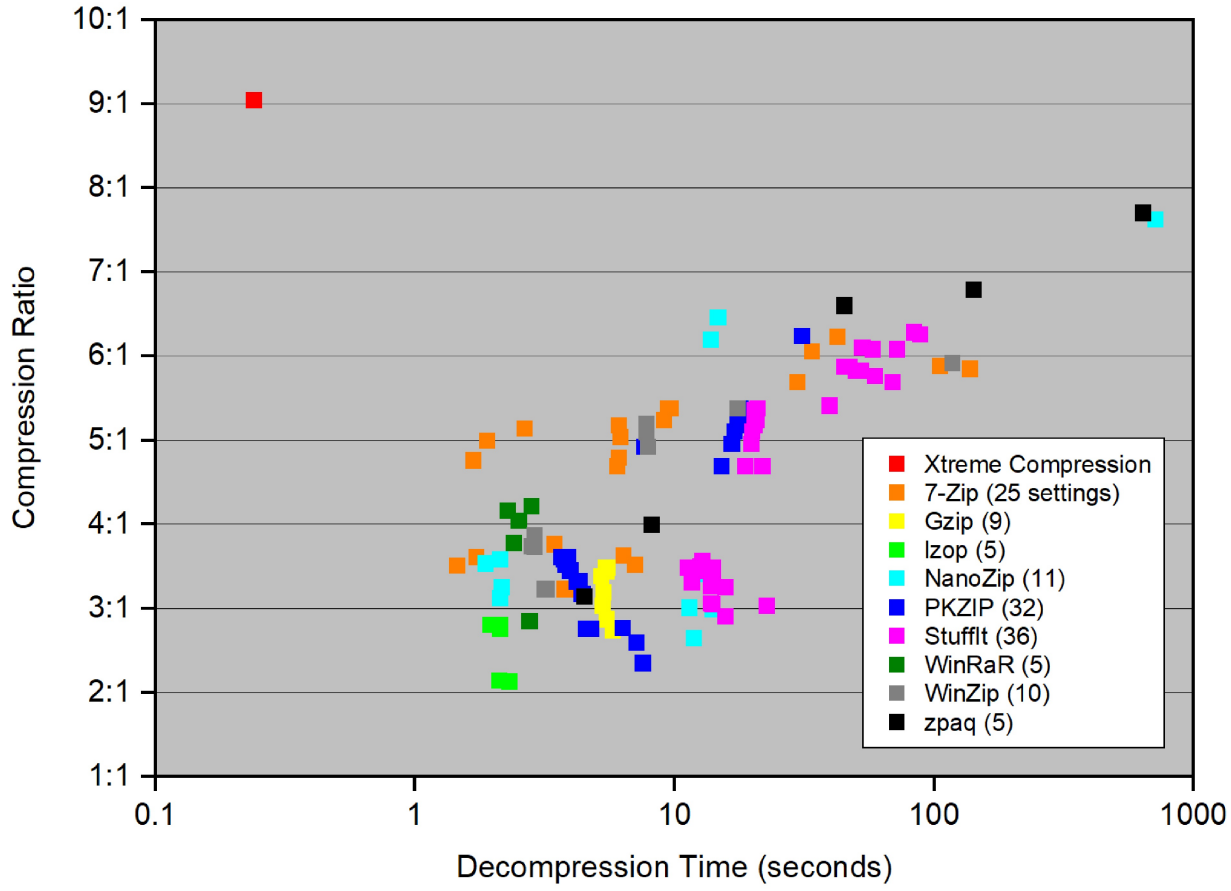
Xtreme Compression Technical Brief

October 1, 2020



Xtreme Compression vs 9 conventional compression utilities on
6 million TPC-H lineitem table records (765,864,502 bytes)

WHAT IS XTREME COMPRESSION?

Xtreme Compression is a specialized software design consultancy. Under contract, we create custom software around our suite of proprietary algorithms.

Our niche is the lossless compression of discrete multidimensional structured data (i.e., tables) for fast transmission, real-time query, and archiving.

Unlike conventional data compression methods, ours do not operate on serial text, images, audio, or video data. Instead, each targets a particular class of structured data -- primarily row/column-organized tables and access data structures -- used in databases.

Common examples of target data are financial transaction information, census files, GIS data, postal address files, clickstream data, metafiles, and similar forms of *Big Data*.

WHY XTREME COMPRESSION?

Performance *beyond the reach of conventional methods* is the reason. That means typically twice the compression ratio with an order of magnitude faster decompression.

Such extraordinary performance greatly extends the realm of products, services, systems, devices, operations, and applications in which data compression's well-known system and economic advantages can be fully realized and exploited. There are two categories:

- Where existing methods' inadequate compression ratios had made capturing sufficient economic benefit impossible
- Where existing methods' inadequate decompression rates had made the use of compression infeasible

THE UNMET NEED

The growth of the world's data in volume and structural complexity has historically been accompanied and facilitated by progress in data management strategies, such as compression. There is now abundant theory, literature, and software for compressing text, images, and video data, but not for discrete multivariate structured data, i.e., table data with row/column structure.

That unmet need persists largely because the state of the art of multivariate data compression has not kept pace. That is due to a lack of principles, insight, and design guidance from multivariate information theory, a historically underdeveloped missing piece. Not until the 2010 advent of the 'partial information decomposition' axiomatic framework did a useful perspective appear.

Cover illustration: On data from a standard data set used by IBM, Oracle, and others for real-time DBMS evaluation, the scatter plot shows the performance advantages of Xtreme Compression when compared to 9 general-purpose compression utilities, each invoked with all valid combinations of command line parameters.

Xtreme Compression achieves 2.5 times the compression ratio with 6 times the decompression rate of the compressor with the next-fastest decompression (7-Zip with "a -m0=LZMA2:x=1"), and 15 percent more compression with over 2500 times the decompression rate of the one with the next-best compression ratio (zpaq with "add -m5").

Timing measurements used Windows 10 PowerShell Measure-Command cmdlet on a Hewlett-Packard 570-p047c Pavilion desktop with a 3.6-GHz Intel 'Kaby Lake' Core i7-7700 processor, H270 chipset, and 16 GB of RAM.

In the absence of suitable methods, low-dimensional, 'industry standard', conventional techniques are still being brought to bear *ad-hoc* on structured data, most often in conjunction with dimensionality reduction. In such a role, they generally perform poorly for several reasons:

- Being structurally mismatched to the data, they cannot recognize interdependencies or remove redundancies whose origin is the data's dimensional structure
- Lacking semantic comprehension, they model data at a single, low level of abstraction
- They have no means to distinguish among, or make efficient use of, the qualitatively distinct kinds of information: unique information, synergistic information, source redundancy, and mechanistic redundancy

Consequently, there is an unmet need, and opportunity, for new compression techniques that deliver performance *beyond the reach of conventional methods* on structured data.

Creating algorithms and software to fill that need is the mission of Xtreme Compression.

PROBLEM, CHALLENGE, AND OBJECTIVE

In general, the redundant information bound up in a set of structured data records is distributed as innumerable discrete but potentially overlapping high-dimensional redundancies that are small individually but substantial when taken together. With high dimensionality, the number of discrete redundancy terms can be prohibitive -- in the millions or billions. That makes obtaining a full quantitative understanding of the data in the conventional way -- by identifying and measuring individual source-target dependencies through statistical analysis -- computationally intractable.

So the challenge is to develop a methodology for systematically making high-dimensional predictions along and across dimensions, and across data types and levels of abstraction, while avoiding having to perform the level of statistical analysis that would otherwise be required. Doing that would allow, in principle, any set of fields from any data record to predict any set of fields in any subsequent record. That is the objective.

THE PROPRIETARY METHODS

Attribute Vector Coding. Attribute vector coding, the centerpiece of Xtreme Compression's proprietary technology, is a fixed-to-variable-length row-oriented vector transform methodology for compressing tables with row/column structure.

Wordencoding. Wordencoding is a 0-order variable-to-variable-length algorithm for compressing text strings.

Repopulation. Repopulation is a structural method for compressing integer sequences in hash tables and similar data structures.

Superpopulation. Superpopulation is a variable-to-variable-length algorithm targeting index tables, lists, arrays, and the like. It may be used alone or in conjunction with repopulation.

Further information can be found at www.xtremecompression.com.