



Compression Algorithm Testing

Making Things Smaller:

An analysis of data compression algorithms

by
Larry Lummis
Product Line Manager,
Quantum DLTtape Group

SDLT
Super DLTtape™ Technology

Table of Contents

Testing Objective.....	2
Testing Objective.....	3
Tape Drive Specifications.....	3

Disclaimer Statement

The testing data referenced in this document was derived from testing performed by an independent laboratory, Percept Technology, Inc., in a controlled environment using specific systems and data sets. Actual results in other environments may vary. These results do not constitute a guarantee of performance. Testing was conducted using Linear Tape-Open (LTO) drives and media from several different LTO consortium manufacturers.

While the information contained in this document may have been reviewed by Quantum for accuracy, there is no guarantee that the same or similar results will be obtained elsewhere. Use of this information is for illustrative purposes and may not represent actual results in a specific application. Product data is accurate as of initial publication and is subject to change without notice.

The information in this document is provided "As Is" without any warranty of any kind.

For use by Quantum and authorized Quantum partners only. Quantum, DLTape, and Super DLTape are trademarks or registered trademarks of Quantum Corporation in the USA and other countries. Other company, product, and service names mentioned may be trademarks or registered trademarks of their respective companies.

© 2002 Quantum Corporation. All rights reserved.

Testing Objective

This Whitepaper describes the results of testing the data compression algorithms of Super Tape Drives. The testing requirements were to compare the data compression algorithm performance of the SDLT 320 manufactured by Quantum against the Linear Tape Open (L.T.O.) drives from HP, IBM, and Seagate and the Advanced Intelligent Tape drive from Sony.

Tape Drive Specifications

The specific tape drives used for the test were the SDLT 320, IBM 3580 Ultrium LTO, HP Ultrium 230 LTO, Seagate Viper 200 LTO and the Sony AIT-3. A summary of each tape drive's specifications is shown in the table below.

		IBM 3580 Ultrium	HP Ultrium 230	Seagate Viper 200	Sony AIT-3
Capacity (native)	160 GB	100 GB	100 GB	100 GB	100 GB
Transfer Rate (native)	16 MB/sec	15 MB/sec	15 MB/sec	16 MB/sec	12 MB/sec
8 TB Library Storage Density	50 Cartridges	80 Cartridges	80 Cartridges	80 Cartridges	80 Cartridges
Interfaces	LVD or HVD	LVD or HVD, FC	LVD or HVD	LVD or HVD	LVD or HVD
Tape Format	Linear serpentine	Linear serpentine	Linear serpentine	Linear serpentine	Helical Scan
Servo Method	Optical Servo	Magnetic Servo	Magnetic Servo	Magnetic Servo	Magnetic Servo
Encoding Method	PRML	RLL 1,7	RLL 1,7	RLL 1,7	PRML
Data Compression Algorithm	DLZ	LTO DC	LTO DC	LTO DC	ALDC
Backward Compatibility	Yes	No	No	No	Yes
MTBF	250,000 @ 100%	250,000 @ 100%	250,000 @ 100%	250,000 @ 100%	400,000 @ 100% <small>POH</small>
Uncorrected Bit Error Rate	< 1 in 10 ¹⁷ bits read	< 1 in 10 ¹⁷ bits read	< 1 in 10 ¹⁷ bits read	< 1 in 10 ¹⁷ bits read	< 1 in 10 ¹⁷ bits read

Tape Drive Data Compression Specifications

In an effort to increase capacity performance all of the tape drives in the test incorporate hardware data compression. Depending on the type of data being backed up, data compression can typically increase the overall capacity of a tape drive 2 to 5 times its native capacity.

The tape drives in the benchmark use different data compression algorithms. All of these algorithms, however, are based on the LZS lossless compression algorithm. LZS stands for Lempel Ziv Stac. LZS is named for its inventors Abraham Lempel, Jacob Ziv and Stac Electronics. Lempel and Ziv developed the mathematical foundations of LZS compression in 1977. Stac improved upon their work and developed engineering solutions for file compression on disk, tape and other computer media.

A brief description of the data compression algorithm used by each of the tape drives is given below.

DLZ

This algorithm is used by the SDLT 320 tape drive. DLZ stands for Digital Lempel Zif. It is the digital hardware equivalent of the LZS standard.

ALDC

This algorithm is used by the Sony AIT-3 tape drive. ALDC stands for Adaptive Lossless Data Compression. ALDC is a slight variation of LZS. It differs from LZS in that it is tuned to provide a higher compression ratios for highly compressible data. The trade-off is a lower compression ratio for less compressible data.

LTO-DC

This algorithm is used by LTO tape drives. LTO-DC is the same as the ALDC algorithm with a incompressible data pass-thru. That is LTO-DC does not apply data compression to input data that has already been compressed such as encrypted data.

Test Setup and Process

The test setup consisted of one Dell Windows 2000 server, one Dell CPi latitude laptop, one Quantum SDLT 320, IBM 3580 Ultrium, HP Ultrium 230, Seagate Viper 200 LTO and a Sony AIT-3 tape drives and Calgary Corpus, Canterbury Corpus, Biomedical Corpus and Binary Corpus files sets. To ensure the best possible performance, all the tape drives were direct attached to the server via a separate Ultra SCSI host adapter to isolate the tape drive bus from the source data bus.

Test setup configuration information is given below:

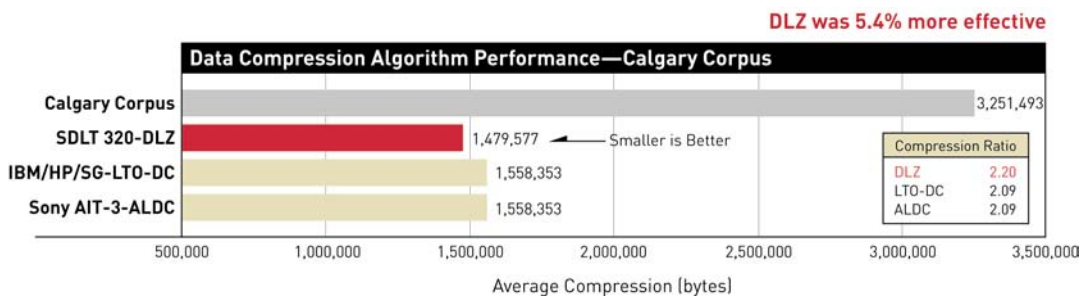
SERVER:	LAPTOP:
Dell PowerEdge 1300/400 400 MHz Pentium 2 128 MB RAM 18GB SCSI Hard disk Windows 2000 server Adaptec 29160 Ultra SCSI host bus adapter -Internally attached source data drive On-board Ultra SCSI host bus adapter - Externally attached tape drive	Dell CPi Latitude 366 MHz Pentium 2 256 MB RAM 6GB ATA Hard disk Windows 2000 Professional
Tape Drives:	Corpi:
SDLT 320: Firmware revision 46 Hewlett Packard Surestore Ultrium 230: Firmware revision E15D Seagate Ultrium Viper 200: Firmware revision 1360 IBM Ultrium StorageSmart 3585: Firmware revision 0BN1	Calgary Corpus Canterbury Corpus Biomedical Corpus Binary Corpus

The test process was designed to get the best objective representative performance for each data compression algorithm used by the tape drives. Because data compression performance is highly dependent on the type of data being compressed, great care had to be taken in selecting the file sets used for the benchmark.

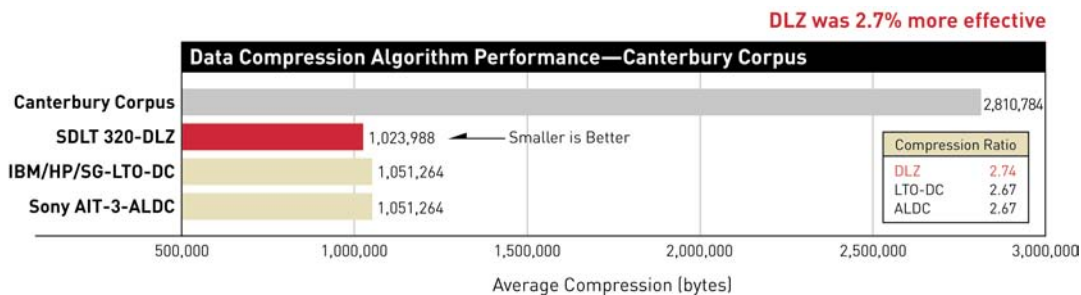
The Calgary Corpus, Canterbury Corpus, Biomedical Corpus and Binary Corpus files sets were chosen for the benchmark tests. The Calgary and Canterbury Corpora were selected because they represent the de-facto standard file sets used by the data compression research community and industry to evaluate the effectiveness of lossless data compression algorithms. The Biomedical and Binary Corpora were selected because they expertly evaluate a data compression algorithm's ability to handle highly compressible data. For each input file set, the overall data compression ratio of the algorithm was measured and recorded.

Testing Results

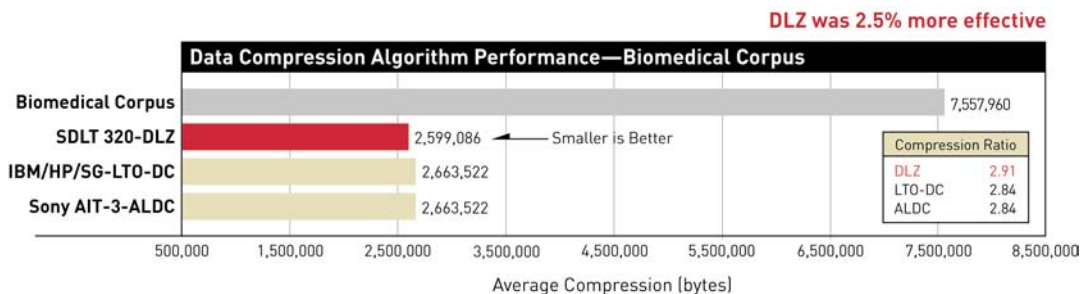
For the Calgary Corpus, the DLZ algorithm used by the SDLT 320 had the best compression performance. DLZ compressed the Calgary file set 5.4% more than either LTO-DC, the algorithm used by LTO tape drives, or ALDC, the algorithm used by AIT-3.



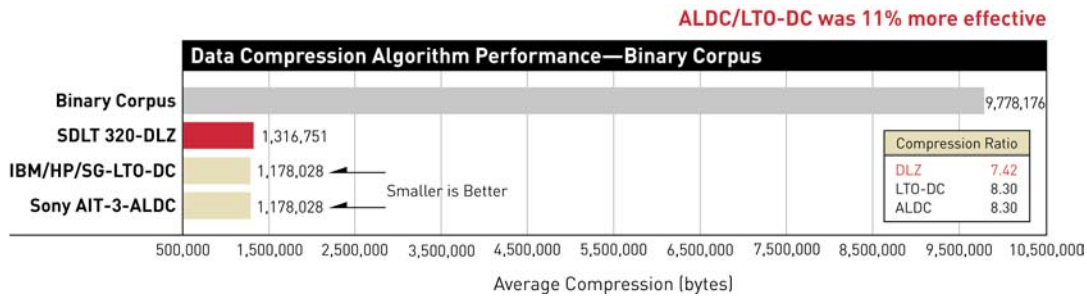
For the Canterbury Corpus, the DLZ algorithm used by the SDLT 320 had the best compression performance. DLZ compressed the Canterbury file set 2.7% more than either LTO-DC, the algorithm used by LTO tape drives, or ALDC, the algorithm used by AIT-3.



For the Biomedical Corpus, the DLZ algorithm again had the best compression performance. DLZ compressed the Biomedical file set 2.5% more than either LTO-DC, or ALDC algorithms.



The LTO-DC and ALDC algorithms performed best on the highly compressible Binary Corpus.



Conclusions

DLZ, the algorithm used by the SDLT 320, had the best compression results for the Calgary, Canterbury and Biomedical Corpi. The DLZ algorithm showed a 5.4%, 2.7% and 2.5% increase, respectively, in compression ratio over the LTO-DC and ALDC, algorithms used by LTO and AIT-3 tape drives.

Corpus Data Compression Test Results

Corpus	DLZ	LTO-DC	ADLC
Calgary	2.198	2.086	2.086
Canterbury	2.745	2.674	2.674
Biomedical	2.908	2.838	2.838
Binary	7.426	8.300	8.300

These results are consistent with what one would expect. The DLZ algorithm was designed to provide excellent data compression for a “typical” average mix of files. The Calgary and Canterbury Corpora have been the de-facto standard for data compression algorithm evaluation because they were designed to represent a “typical” mix of files.

The LTO-DC and ALDC algorithms were designed to provide better results for highly compressible data files, which are represented by the Binary Corpi.

For an end-user, the SDLT 320’s increase in data compression efficiency translates into a real lower total cost of ownership advantage. Consider the case of a SDLT 320 based-8TB library. Because of the SDLT 320’s superior DLZ data compression algorithm a customer would typically be able to backup 400 GB more data than either a LTO or AIT-3 tape drive.

References

T.C. Bell, J. G. Cleary, and I. H. Witten. Text Compression. Prentice Hall, Englewood Cliffs, NJ, 1990.

Editors. Data Compression Conference (DCC 91), Snowbird, Utah, 1995. IEEE Computer Society Press.

T.A. Welch. A technique for high performance data compression. IEEE Computer, 17:8-20, June 1984

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. IEEE Transactions on Information Theory, IT-23:337-343, 1977.

J. Ziv and A. Lempel. Compression of Individual sequences via variable rate coding. IEEE Transactions on Information Theory, IT-24:530-536, 1978.

###