Compression Module

Justin

Algorithm

1. Divide input data into segments equivalent to 20 to 40% of the available memory. (Larger buffer size)
2. Divide the segments into variable chunks for data de duplication.
3. De duplicate chunks based on fingerprinting and hashing techniques.
4. Supply the de duplicated chunks to a Diff tree generator: this generates a maximum spanning tree based on the diffability among chunks (All pair shortest path approach). Uses the root as the base for diff (Needs a little clarification) .
5. Do a delta compression using the root from the above step, and arrange the data into a single continuous block for final compression.
6. Do a final compression on the block generated from above step (lz or bzip2).