A Survey of Content-Defined Chunking Algorithms

Content-defined chunking (CDC) is a fundamental technique in data deduplication systems. Unlike fixed-size chunking, CDC uses a rolling hash function to determine chunk boundaries based on the content itself, making it resilient to byte insertions and deletions.

The Rabin fingerprint algorithm is commonly used for CDC boundary detection. When the fingerprint matches a predetermined pattern, a chunk boundary is declared. This approach enables efficient deduplication even when data is modified.

Recent advances include FastCDC, which improves throughput by using normalized chunking and gear-based rolling hash. Experiments show FastCDC achieves 10x higher throughput than traditional Rabin-based approaches while maintaining similar deduplication ratios.