

A Thorough Investigation of Content-Defined Chunking Algorithms for Data Deduplication

Marcel Gregoriadis, Leonhard Balduf, Björn Scheuermann and Johan Pouwelse

Abstract—Data deduplication emerged as a powerful solution for reducing storage and bandwidth costs in cloud settings by eliminating redundancies at the level of chunks. This has spurred the development of numerous Content-Defined Chunking (CDC) algorithms over the past two decades. Despite advancements, the current state-of-the-art remains obscure, as a thorough and impartial analysis and comparison is lacking. We conduct a rigorous theoretical analysis and impartial experimental comparison of several leading CDC algorithms. Using four realistic datasets, we evaluate these algorithms against four key metrics: throughput, deduplication ratio, average chunk size, and chunk-size variance. Our analyses, in many instances, extend the findings of their original publications by reporting new results and putting existing ones into context. Moreover, we highlight limitations that have previously gone unnoticed. Our findings provide valuable insights that inform the selection and optimization of CDC algorithms for practical applications in data deduplication.

Index Terms—Data deduplication, content-defined chunking, storage systems, performance evaluation.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.



1 Introduction

IN the era of Big Data, cloud storage systems have become indispensable for managing the explosive growth of digital information [1]. As storage is costly, these systems require efficient data reduction techniques. Concurrently, the advent of the Internet of Things has underscored the importance of minimizing data transfers between edge devices and central servers, often located in the cloud. In large-scale systems, as data accumulates, it is typical for content to appear redundantly. This results in an inefficient utilization of both bandwidth and storage.

Data deduplication emerges as a solution to this issue. The strategy is to eliminate redundant content at a chunk-level, for instance, in blocks of 8 kB. To this end, files are split into chunks and each chunk is indexed and identified by its cryptographic fingerprint. Thereafter, a file is described as a sequence of such fingerprints. Hence, duplicated blocks of data need to be stored or transferred only once, and instances of the blocks can be referred to by their fingerprint. As the size of a fingerprint is much smaller than the content it represents, this results in effective deduplication on systems where redundant content is prevalent. Large-scale studies by Microsoft [2], [3] and EMC [4] report space savings of up to 83% using this technique.

The algorithm by which the files are chunked has an important effect on deduplication. The most straightforward solution is *Fixed-Size Chunking (FSC)*, where files are split

into equal-sized chunks. This strategy, however, suffers from the *boundary-shift problem*. It describes the situation that two (or more) files share similar content, but the misalignment of their chunk boundaries hinders the detection of the existing redundancies (cf. Figure 1). This problem is addressed by *Content-Defined Chunking (CDC)* algorithms, which yield variable-sized chunks based on *content* rather than *position*. To do this, CDC algorithms often rely on rolling hash functions, whose first application was the Rabin fingerprinting scheme [5]–[7].

Over the years, numerous algorithms have been proposed, claiming better efficiency (*i.e.*, higher throughput), lower chunk-size variance, or better deduplication efficacy [8]–[12]. However, a comprehensive and unbiased evaluation of these methods remains elusive. Each study typically presents its algorithm as the superior solution, often using curated datasets and assumptions that favor their approach. This fragmented landscape obscures a clear understanding of the true state-of-the-art in CDC. In our study, we select a set of CDC algorithms for rigorous evaluation, including Rabin [7], Buzhash [13], Gear [14], AE [15], RAM [10], MII [16], PCI [11], and BFBC [12], and the *Normalized Chunking (NC)* technique proposed for FastCDC [8]. We reimplement these algorithms efficiently and compare their performance on four realistic datasets. Our evaluation encompasses throughput, average chunk size and variance, and deduplication ratio. We report new results and contrast them with existing literature. In addition, we derive new theoretical insights, including novel formulas relating algorithm parameters to the expected average chunk size for AE, RAM, MII, and BFBC. Moreover, we improve upon the existing formula for AE. In summary, our research provides a comprehensive and unbiased evaluation, shedding new light on the capabilities and limitations of these algorithms.

The remainder of this article is structured as follows: In

- *M. Gregoriadis and J. Pouwelse are with the Data-Intensive Systems Lab at Delft University of Technology, The Netherlands. E-mail: m.gregoriadis@tudelft.nl; j.a.pouwelse@tudelft.nl*
- *L. Balduf and B. Scheuermann are with the Communication Networks Lab at Darmstadt University of Technology, Germany. leonhard.balduf@tu-darmstadt.de; scheuermann@tu-darmstadt.de*

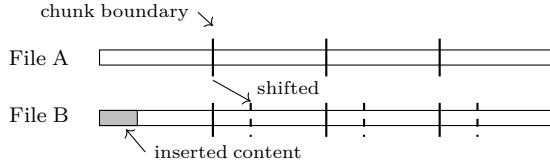


Fig. 1: Boundary-shift problem.

Section 2, we give an overview of the relevant algorithms and key techniques that shaped the field of CDC and define today’s state-of-the-art. Following this, in Section 3 we outline related works in the field of empirical measurements of CDC algorithms. In Section 4, we provide a detailed exposition of all the chunking algorithms that are subject to our in-depth analysis and comparison. This includes remarks on their expected behavior and performance, such as the effect of their parameters on the expected chunk size. Following this, we commence the experimental evaluation of the selected algorithms, its detailed procedure described in Section 5. The subsequent three sections present and analyze the results of our experiments with respect to our key metrics: throughput (Section 6), chunk size distribution (Section 7), and deduplication (Section 8). In Section 9, we interpret interrelations between the results and contemplate their implications in a broader context. Finally, in Section 10, we arrive at our conclusions.

2 Background

In the face of exponential data growth, data deduplication has emerged as a pivotal strategy for efficient data management [17]. The algorithm by which chunking is performed poses the crucial feature by which efficacy and efficiency of the data deduplication process is determined. This section provides an overview of the evolution of CDC and the seminal innovations that shaped it.

2.1 Inception of CDC

CDC algorithms avoid boundary shifting by setting chunk cut-points based not on position but *content*. Traditionally, this has been the result of hash-based comparisons on a sliding window that is iterated over a file byte-by-byte (cf. Figure 2). The fingerprint on each iteration of the sliding window is compared against a bitmask to determine new chunk boundaries. Since hash functions are deterministic, this results in chunk boundaries that are set in a content-dependent manner. This idea, which we refer to as *Basic Sliding Window (BSW)*, marks the advent of data deduplication. It can be attributed to two pioneering studies from the early 2000s [18], [19].

2.2 Chunk-Size Variance

Early on, BSW-based CDC was criticized for two shortcomings: low throughput and high chunk-size variance. While the problem of low throughput was ameliorated by more efficient hash algorithms [13], [14], high chunk-size variance remained a problem inherent to the BSW approach. High chunk-size variance gives rise to the issue of pathological chunk sizes. Very large chunks can be the product of a recurring pattern in

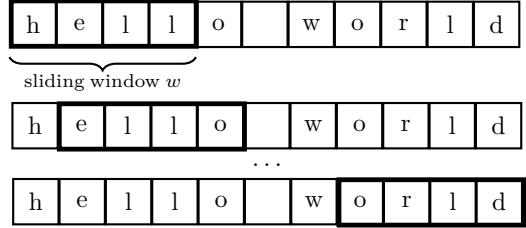


Fig. 2: Window sliding over a file byte-by-byte.

the byte sequence which happens to not meet the criteria for setting a chunk boundary [20]. These chunks are undesirable because they impair deduplication: First, large chunks are generally more difficult to deduplicate as the chances for smaller chunks of data to be redundant are higher. Second, when an existing file is modified, this modification is more likely to affect a larger chunk than it is to affect a small chunk, which in turn leads to a higher number of bytes affected by the modification, hence a negative effect on deduplication. Pathologically small chunks are undesirable as well because they produce more metadata, more computation overhead, and, in distributed settings, greater overhead due to round trip times.

2.3 Modern CDC Algorithms

High chunk-size variance and low throughput led to the emergence of an alternative approach. Starting in 2009, researchers started proposing CDC algorithms based on the identification of local extrema in the input data [9], [10], [21], [22]. By using byte comparisons rather than hash functions, these algorithms claim to achieve higher throughput than BSW algorithms. Furthermore, they are attributed with a significantly lower chunk-size variance [21], [22].

Later on, researchers focused on the specific application of CDC for incremental synchronization, as for example in rsync [23]. This use case does not consider data reduction in storage systems, but the incremental synchronization of data between machines. Files are chunked on both ends to determine the segments of data that need to be transmitted. However, the produced chunks are never stored. This condition puts a relaxation on the constraint for low chunk-size variance. We find this focus particularly in the works of Zhang *et al.* in 2019 [16] and 2020 [11]. Those algorithms do not fundamentally differ, however, from the BSW or extremum-based approaches.

Lastly, the third and most recent approach is chunking based on a dynamically predetermined set of divisors, and emerged in 2020 [12]. The technique relies on a statistical analysis of the expected dataset. Specifically, the algorithms will conduct a statistical frequency estimation of byte pairs [12], [24] (or triplets [25]), and, based on that, determine a set of byte pair/triplet divisors. The matching condition is thus reduced to a simple table lookup, promising superior throughput compared to classical CDC algorithms.

3 Related Work

Previous works have studied the state-of-the-art of data deduplication and chunking algorithms [17], [26]. However, these studies primarily focus on theoretical discussions based

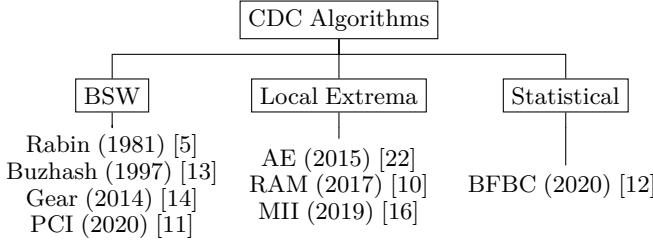


Fig. 3: Taxonomy of evaluated chunking algorithms.

on the original works that introduced the algorithms. Few researchers have attempted to reproduce the results or investigate the state-of-the-art through experimental means. Most experimental evaluations and comparisons of chunking algorithms are part of works introducing new algorithms [10], [11], [22], [27]. We find that there is no consistent set of datasets and comparable evaluation methodologies to judge these algorithms based on their original publications.

In their work [27], Ellapan *et al.* present the superior throughput of their own algorithm. In their measurements, however, they did not isolate the effects of the chunking algorithm from the computationally expensive SHA-1 fingerprinting applied to the produced chunks. Consequently, the throughput is heavily skewed in favor of algorithms that generate a smaller number of chunks. Notably, their own algorithm produces the fewest chunks across all evaluated algorithms and datasets.

The authors of PCI [11] compare their algorithm against Rabin, LMC, AE, RAM, and MII. The datasets used in the experiments are artificial, based on sequences of zeros with random byte insertions or deletions in defined intervals. Ultimately, chunking algorithms are applied to real-world data, which is not reflected in the datasets used by the authors. For instance, the issue of low-entropy strings is disregarded within this experiment.

We find one study that compares specifically the throughput of Rabin, LMC, AE, RAM, and PCI on random datasets [28]. In their experiment, RAM significantly outperformed the other algorithms, followed by AE and PCI; results that our own experimental study confirms as well. While useful, this work lacks *realistic* datasets, making it difficult to derive actionable recommendations from it.

4 Chunking Algorithms

We conducted a thorough investigation of related literature to identify a set of state-of-the-art algorithms to carry on for our theoretical analysis and experimental comparison. This includes recent as well as traditional CDC algorithms (cf. Figure 3). In this section, we reintroduce these algorithms with detailed technical descriptions. We examine the algorithms through the lens of theoretical behavior, extending the descriptions and derivations made by the original authors where possible. The pseudocode to our implementations of the chunking algorithms can be found in Appendix A.

4.1 Basic Sliding Window (BSW)

BSW algorithms operate by sliding a fixed-size window of size w over the stream of input data, deriving a fingerprint for

the current window using a function H , and emitting a chunk boundary if the calculated fingerprint fulfills a given condition. The BSW variants differ in the hash function producing the fingerprint, the function judging the fingerprint, and the choice of window size. Typically, *rolling hash* functions are used for their efficiency. These functions can update their output in constant time when used in a sliding window. Further, the judging function usually checks for a number b of least-significant bits to be zero. If H is distributed uniformly at random, this can then be expected to occur for any window with a probability of 2^{-b} . Therefore, an average chunk size of μ can be aimed by setting $b = \log_2(\mu)$.

BSW variants differ in the hash function they utilize. Rabin-based chunking is the first prominent application of the BSW algorithm, and moreover of CDC in general [18], [29], [30]. It is rooted in the fingerprinting schema presented in [5], [7]. Rabin was often criticized for being slow [31], [32], which spurred the development of more efficient rolling hash functions. Hashing by cyclic polynomials [13], or *Buzhash*, presents a more efficient rolling hash function. Another efficient implementation was presented as *Gear* [14]. Due to its shifting behavior, the matching condition for *Gear* uses the *most* significant bits of H . In Table 1, we show how to compute both the first and consecutive hashes for a sliding window over a stream of bytes. In the presented formulas, x is a prime number, $\rho^b(x)$ denotes a binary rotation of x by b bits, and $h : [0, 255] \mapsto [0, 2^{32}]$ denotes a predefined table. We note that we experimentally verified uniform hash value distributions for these functions, as this poses a vital criterion for effective CDC [33].

Another optimization applicable to BSW algorithms is data-parallelism. This is a common optimization technique with dedicated instructions on all modern instruction sets [34], [35]. It is generally not obvious how to parallelize CDC algorithms to operate on different parts of the input data, due to the boundaries of *previous* blocks affecting the *current* state of the algorithm. BSW algorithms have a property that can be exploited in this regard: They operate on a fixed window, *i.e.*, the number of input bytes affecting their current state is limited. Using this property, it is possible to write data-parallel versions of these algorithms using *single instruction multiple data (SIMD)* instructions. For *Gear*, in particular, this is made easy due to the simplicity of the algorithm itself, and SIMD implementations exist.¹

4.2 Asymmetric Extremum (AE)

AE [15], [22] emerged from the alternative line of extremum-based approaches, *i.e.*, it does not rely on hash functions. It determines chunk boundaries based on an extreme value within an asymmetric window. The window is comprised of a fixed-size horizon of length h (to the right) and another dynamically sized horizon (to the left). A chunk boundary is declared at index $i + h$ if the byte at i is the local extremum from the previous chunk cut-point to $i + h$; more precisely, if $B_i > \max\{B_j\}_{j=1}^{i-1} \wedge B_i \geq \max\{B_j\}_{j=i+1}^{i+h}$. We illustrate the chunking mechanism in Figure 4.

The average chunk size in AE largely depends on the parameter setting h . Larger values yield larger chunks. In order to produce chunks of specific average size μ , understanding the

1. *e.g.*, <https://crates.io/crates/gearhash>

TABLE 1: Rolling Hash Functions With Initial Computation (H_{prev}) and Update Method (H_{next})

Rolling Hash	$H_{\text{prev}} = H(B_1, \dots, B_w)$	$H_{\text{next}} = H(B_2, \dots, B_{w+1})$
Rabin [5]	$B_1x^{w-1} + B_2x^{w-2} + \dots + B_w$	$(H_{\text{prev}} - B_1x^{w-1})x + B_{w+1}$
Buzhash [13]	$\rho^{w-1}(h(B_1)) \oplus \rho^{w-2}(h(B_2)) \oplus \dots \oplus h(B_w)$	$\rho(H_{\text{prev}}) \oplus \rho^w(h(B_1)) \oplus h(B_{w+1})$
Gear [14]	$h(B_1) \cdot 2^{w-1} + h(B_2) \cdot 2^{w-2} + \dots + h(B_w)$	$(H_{\text{prev}} \ll 1) + h(B_{w+1}) \bmod 2^w$

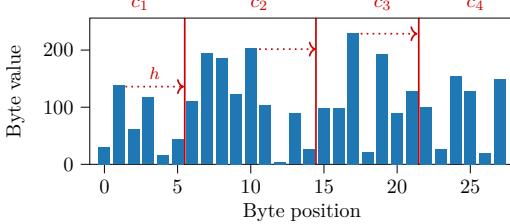


Fig. 4: Illustration of the AE chunking algorithm on a sequence of 27 bytes and a horizon $h = 4$. The vertical red lines mark the cut points which then determine the resulting chunks c_i .

relationship between μ and h is crucial. The authors suggest that the average chunk size on random data input is expected to be $(e-1) \cdot h$, therefore $h = \mu/(e-1)$. This formula has been implemented in their open-source testing framework Destor², which was used in various works as the basis for the experimental evaluation of AE in comparison to other algorithms [8], [27], [36], [37], as well as the more recent benchmarking tool DedupBench [38]. Our experimental as well as theoretical analysis, however, suggests that this formula cannot be used to accurately predict the average size of produced chunks. When employing it to determine the parameter h , AE yields means significantly lower than the target. We observed these results on uniformly distributed random data, motivating us to revisit the mechanism to determine h . Hereby, we notice that the authors disregard not only the discrete value range of the random variables in $[0, 255]$, but also the conditioning of the probabilities based on dependent events. Thus, we conduct our own stochastic analysis and come to the following conclusions: If h is large, the extreme value is likely to be 255, and therefore the probability for an unknown random byte to match this value is $\frac{1}{256}$. Based on this reasoning, we propose using the approximation $h \approx \mu - 256$ where $\mu \geq 2 \text{ KiB}$. For smaller target chunk sizes, we rely on empirical evaluations, the results which we list in Table 2.

TABLE 2: Empirical Results for Parameter h in AE for Target Chunk Sizes $\mu < 2 \text{ KiB}$

μ	h
512	348
770	563
1024	793

4.3 Rapid Asymmetric Extremum (RAM)

The throughput of AE has been improved even further in RAM [10]. This algorithm essentially swaps the order of the fixed and dynamically sized windows. It first employs a fixed-sized window, in which it finds the maximum value x ; it then determines the first byte that is larger or equal x as the next cut-point. This requires fewer comparison operations than AE and is stated to be $\approx 25\%$ faster.

The authors note that the performance of RAM can suffer when the given data has low entropy. This is apparent for large x , as it becomes ever more unlikely to find any byte $\geq x$. To counteract this behavior, the authors recommend setting a maximum chunk size, trading off deduplication efficacy. As our objective is to measure the inherent characteristics of the algorithms, and give them a fair comparison, we deliberately do not impose any such limit on our implementation.

The authors do not explicitly state how to tune the algorithm's parameter h for desired average chunk sizes. We analytically derive the relationship between μ and h as expressed in Equation 1. By solving for h numerically, this formula can be used to tune RAM for different target chunk sizes. Our experimental evaluation supports this derivation with near-perfect empirical means on uniformly distributed random data.

$$\mu = h + \left(1 - \frac{\sum_{m=0}^{255} \left(m \left(\left(\frac{m+1}{256} \right)^h - \left(\frac{m}{256} \right)^h \right) \right)}{256} \right)^{-1} \quad (1)$$

4.4 Minimal Incremental Interval (MII)

MII [16] is another instance of extremum-based algorithms, with less emphasis on the achievement of low chunk-size variance. It applies a fixed window w and determines a chunk cut-point after each byte position i for which the predicate $B_i > B_{i-1} > \dots > B_{i-w}$ is true. In simpler words, MII sets a chunk boundary after an incremental interval of length w .

Larger intervals, *i.e.*, higher w , will accordingly lead to larger chunks. An exact formula is not provided by the authors. For the purpose of our own experiments, we propose Equation 2. The rationale for this formula is that in every window of w bytes, there exist 256^w possible combinations. For the above-stated predicate to be true, every byte in the window must occur uniquely. Furthermore, for every possible combination of w distinct values, there exists exactly one order in which they are ascending. The total number of this possibility is captured in $\binom{256}{w}$.

$$\mu(w) = \left(\frac{\binom{256}{w}}{256^w} \right)^{-1} + w \quad (2)$$

This equation shows, as also the authors have mentioned, the factorial growth of the average chunk size with growing

2. <https://github.com/fomy/destor>

TABLE 3: Empirical Results for PCI Parameters w, θ to Approximate Specific Targets μ as Obtained in Simulation

μ	w bytes	$8w$ bits	θ	$\frac{\theta}{8w}$
512	58	464	253	0.545
770	40	320	181	0.566
1024	34	272	157	0.577
2048	61	488	273	0.559
4096	39	312	183	0.587
5482	56	448	256	0.571
8192	57	456	262	0.575

w and the weak control over it that comes with that. As our experiments reveal (cf. Table 11), this formula predicts the average chunk size with an offset of $\approx -15\%$. We speculate this is due to dependent probabilities of not matching the previous window, which we have not regarded in our formula.

4.5 Parity Check of Interval (PCI)

The authors of MII later proposed PCI [11] as an improvement to MII, which they critiqued for having weak control over the average chunk size. PCI works schematically similar to the BSW algorithm, but uses the popcount, *i.e.*, the number of 1-bits in the sliding window, instead of a hash function. A chunk boundary is set if the popcount exceeds a specific threshold θ . It shares properties of rolling hash algorithms, as subsequent iterations require only the removal of the popcount of the leftmost byte and the addition of the popcount of the rightmost byte. We note that the pseudocode in the original publication omits this optimization.

Contrary to the statement in the original paper, the popcount is not subject to a discrete *uniform* but rather a discrete *binomial* distribution for uniformly distributed random byte sequences. Specifically, this leads to a probability of $\binom{8w}{\theta} \cdot 2^{-8w}$ for every popcount $\theta \in [0, 8w]$ in a window of w bytes.

The average chunk size is determined by the ratio between θ and w , rather than their absolute value. Because of the binomial probability distribution for θ bits in $8w$ bits to be 1-bits, the average chunk size grows superlinearly with increasing $\frac{\theta}{8w}$, if $\frac{\theta}{8w} > 0.5$. However, if w can be chosen freely, any granularity of $\frac{\theta}{8w}$, and thereby virtually any average chunk size, can be targeted. Note, w simultaneously sets an implicit lower bound on the chunk size.

Since the expectation of the popcount in a sliding window is influenced by the validation of the matching condition on preceding bytes, the aforementioned formula does not conclusively inform about the frequency of chunk cut-points given w and θ . As an exact solution is out of scope of our work, we determine the parameters empirically for our experiments. We run a simulation of the algorithm on a sequence of 10 MB of data for all possible parameters in the range $w = [32, 64]$. The results are shown in Table 3. In the last column, we show the ratio between popcount-threshold and window size. This leads to an interesting observation: The naive assumption is that the average chunk size is subordinate to the ratio $\frac{\theta}{w}$, rather than their absolute values. However, one must also acknowledge the role of w as an implicit lower bound on the chunk size. Ultimately, both w itself as well as the ratio $\frac{\theta}{w}$ influence the produced chunk size.

4.6 Bytes-Frequency-Based Chunking (BFBC)

BFBC [12] operates differently than the other algorithms in that it is tailored to the dataset. The initialization is composed of a statistical frequency analysis. This analysis identifies the top- k frequent byte pairs, which are then used as divisors in the chunking process, alongside minimum and maximum chunk size thresholds. This approach aims to be faster but also achieve superior deduplication compared to traditional CDC algorithms.

As the original publication does not indicate implementation details, the data structure to hold the set of divisors presents an interesting design choice. In our implementation we utilize a 8 KiB bitset, which results in constant-time lookups, regardless of the number of divisors.

Meeting a desired chunk size is challenging and depends on the distribution of byte pairs and therefore the content of the dataset itself. In our experiments, we noticed that the most frequent byte pairs in realistic datasets tend to occur excessively (*e.g.*, NULL-NUL, or `/>` in HTML files). Using such a byte pair as a divisor leads to chunk boundaries often created a few bytes after the minimum threshold. As we require comparable average chunk sizes for our analysis, and using the minimum chunk size as a means to control it negatively impacts the deduplication ratio, we design an algorithm to select a set of divisors that would result in the desired average chunk size. In our experiments, we run this modification of BFBC as an additional algorithm, denoted BFBC*. It only differs in the procedure which determines its divisors, explained in the following.

4.6.1 Determining BFBC* Divisors

We propose an algorithm that finds a set of divisors in the list of frequent byte pairs that match a given target chunk size μ w.r.t. a minimum chunk length λ_{\min} . For our explanations, we will denote F as an ordered list of only the frequencies of the most-frequently occurring byte pairs in descending order, *e.g.*, $F = (10112, 8435, 8003, \dots)$. Furthermore, D represents the set of indices of F used as divisors.

We can further say, without consideration of λ_{\min} , that the subjected file will be split in as many chunks as the accumulated frequencies over all divisors, plus 1. Knowing the file's length l , we can determine the definitive average chunk size. When considering $\lambda_{\min} > 0$, we can set up the assumption that the divisors are uniformly distributed across the file, and then simply subtract the number of hypothetical chunks multiplied with λ_{\min} to get an adjusted average chunk size. Finally, this leaves us with Equation 3 for the calculation of the expected average chunk size.

$$\mu(D) = \frac{l}{1 + \sum_{i \in D} F_i} + \lambda_{\min} \quad (3)$$

Our algorithm uses Equation 3 while iterating over potential divisors to make predictions about the outcome of the average chunk size. The algorithm is greedy, minimizing the difference between target mean and achieved mean.

5 Experiment Setup

We implement all algorithms from Section 4 in a standardized environment, tuned for performance and efficiency. In addition, we created a test framework that analyzes the

throughput, average chunk size and chunk size distribution, and deduplication ratio for a range of target chunk sizes and datasets. We have been careful to isolate the performance of the algorithms from factors that are not determined by the characteristics of the chunking algorithms themselves and would therefore lead to false conclusions or unfair comparisons. Our analysis excludes additional overhead from the processes of fingerprinting and disk I/O as much as possible. We detail the measures taken to ensure reproducibility.

In all our experiments, we use target chunk sizes from 512B to 8KiB in exponential steps as values in that range are common in literature and practice [16], [22], [30], [31]. MII represents a special case among the set of evaluated algorithms, as adjustments of its parameter w result in target chunk sizes of $\dots, 130, 770, 5482, 45037, \dots$ bytes. This makes aligning w to our range of target chunk sizes impossible. Therefore, we use w where $\mu(w)$ is closest to μ in our analysis of throughput, where the chunk sizes do not have a notable effect on the performance. Further, in order to be able to make a fair assessment of MII in the other experiments, we add target chunk sizes 770B and 5482B. Note, however, that those targets cannot be met by BSW algorithms.

Our testing framework³, as well as our implementations of the algorithms⁴, are published on GitHub under permissive licenses.

5.1 Parameters Settings

The parameters in each algorithm ultimately adjust the target chunk size. All parameter settings in our experiments are informed by either the literature, our own stochastic analysis, or, in some cases, empirical findings. They align with the algorithm descriptions provided in Section 4. An overview is shown in Table 4.

In order to determine optimal window sizes for Rabin and Buzhash, we run an experiment to measure their deduplication performance on randomly distributed data. The best results were achieved with $w = 32$ and $w = 64$, respectively (cf. Figure 10). Recall that Gear has an implicit window size equal to the width of its hash, *e.g.*, 32B for 32-bit Gear.

Similar to MII and PCI, finding parameter values for BFBC is not trivial. We follow the recommendation of the authors: We employ a minimum chunk size λ_{\min} shortly before the target and set $k = 3$, aiming for chunks being created within a short interval after λ_{\min} . In addition, we extend the set of algorithms by BFBC*, by which we refer to BFBC using our improved algorithm for determining divisors (cf. Section 4.6.1). BFBC* does not enforce a minimum chunk length.

5.2 Datasets

As the deduplication ratio and the chunk size distribution highly depend on the given dataset, we have collected multiple real-world datasets in addition to one artificially generated dataset with maximal entropy. Our choice of datasets is inspired by the kind that is typically used when evaluating CDC algorithms [10], [12], [31], [39], [40]. In our repository, we include the exact scripts used to craft the datasets. Those encompass:

3. <https://github.com/mrd0ll4r/cdc-algorithm-tester>
4. <https://github.com/mrd0ll4r/cdchunking-rs>

TABLE 4: Algorithmic Parameter Settings

Alg.	Parameter Settings
BSW	$w = 32, b = \log_2(\mu - w)$
AE	$h = \begin{cases} \mu - 256 & \text{if } \mu < 2 \text{ KiB}, \\ \text{cf. Table 2} & \text{otherwise} \end{cases}$ h such that
RAM	$h + \left(1 - \frac{\sum_{m=0}^{255} \left(m \left(\left(\frac{m+1}{256}\right)^h - \left(\frac{m}{256}\right)^h\right)\right)}{256}\right)^{-1} = \mu$
MII	w such that $\left(\binom{256}{w} \cdot 256^{-w}\right)^{-1} \approx \mu$
PCI	cf. Table 3
BFBC	$k = 3$ and $\lambda_{\min} = \mu - 128$
BFBC*	D such that $\frac{l}{1 + \sum_{i \in D} F_i} + \lambda_{\min} \approx \mu$

- **LNX:** A selection of 14 Linux ISO images representing various distributions, as well as different versions of the same distribution. Note that ISO is an uncompressed file format.
- **PDF:** Over 2000 PDF files (scientific articles) retrieved from arXiv⁵. PDF is a complex format and includes both textual content and binary content (*e.g.*, to represent images).
- **WEB:** Daily snapshots of the website nytimes.com for the entire year of 2022, downloaded from the Internet Archive and recursively crawling three levels of links. As such, this dataset is a mix of textual (HTML, CSS, and JavaScript) as well as binary files (images, videos, font files, etc.). The latter makes up 89 % of the content.
- **CODE:** Source code distributions of various releases of the open-source projects GCC, GDB, and Emacs (94 versions in total). The content in this dataset is 97 % textual and thereby the most likely to benefit from CDC.

The datasets were intentionally chosen to be suitable beneficiaries for CDC because of textual file format or the high intra-correlation given by series of consecutive versions and their incremental changes. We still chose to include an artificial dataset of random data, **RAND**, which corresponds to the theoretical considerations about the expected behavior of the algorithms as laid out in Section 4. All datasets are approximately 10 GiB in size, which is important for comparable results. An overview of the datasets and their characteristics is given in Table 5. The entropy of a dataset is represented as its size after GZIP compression relative to its original size.

TABLE 5: Experimental Datasets

Name	Entropy	Description
RAND	100.0 %	Randomly generated binary.
LNX	98.6 %	Linux ISO images.
PDF	87.5 %	Collection of research papers.
WEB	72.8 %	Daily website snapshots.
CODE	22.1 %	Consecutive versions from code repositories.

Apart from RAND, each dataset represents a collection of multiple files. However, CDC algorithms operate on only

5. https://info.arxiv.org/help/bulk_data

one data stream. In order to make valid and comparable claims and observations, it is necessary to collect each dataset into one single file. We do this through simple concatenation. This comes with a caveat when interpreting the results: They might not represent the performance that would be found for the same datasets in real storage systems, as those systems usually deduplicate on the level of individual files. This can be advantageous because similar (or related) files will have their first chunks' starting position aligned at byte index 0. With our method, chunks can be formed starting in one file and ending in another. This means that even files that are identical might not be detected as duplicates, especially when the target chunk size is large relative to the file size. In order to get comparable results on various target chunk sizes, datasets, and algorithms, we have to contemplate the datasets as streams of data of a specific type (*e.g.*, source code). For the same reasons we omit the last chunk produced for every dataset in the evaluation of our experiments.

5.3 Benchmark Program

In order to measure the algorithms, both in terms of throughput as well as the chunks produced, we present a framework and implementation of a benchmark program, written in the Rust programming language, with a focus on performance and efficiency. The framework makes it easy to implement new chunking algorithms and test their performance. It consists of a *driver*, which reads the input file in large blocks and uses them to drive a selected algorithm. The algorithms are presented with consecutive blocks of file data, on which they are to find a boundary, advancing their internal state as they ingest the blocks. This allows for a performant, real-world oriented implementation. Ultimately, the algorithms operate on sequences of bytes, closely following the pseudocode laid out in their descriptions. The algorithms are collected into a single benchmark program, which is compiled with optimizations enabled, targeted at the executing machine. The benchmark program operates as follows: The selected input file is read, fed into the selected algorithm, produced chunks are fingerprinted, and the fingerprint and size of the chunks is output.

When evaluating the throughput of an algorithm, however, the resulting chunks are not fingerprinted. In this case, the benchmark program tracks and outputs a single value, the sum of all chunks' sizes, to prevent compiler optimizations from removing the chunking code altogether. The entire file is served from a RAMdisk, ensuring that the speed of reading the file is not a limiting factor, which we verify by implementing and evaluating FSC using the same framework. We ensure ample memory remains for program execution. Apart from our benchmark program, the system is in an idle state. All benchmarks are evaluated sequentially, in order not to influence each other. We execute the the benchmark program for each dataset, algorithm, and target chunk size a total of $n = 10$ times. In addition, we “warm up” the system once per dataset/algorithm combination.

We execute our benchmarks on a machine running Ubuntu 22.04 with an Intel Xeon Gold 6154 CPU at 3.00 GHz, capture performance counters using `perf`, and report on statistics derived from these results. While we execute our benchmarks on one specific system, we believe that general trends are

transferable to other systems. Properties of the algorithms, such as cache utilization or ease of branch prediction, are influential on all modern systems. Note also that the results for deduplication performance and chunk size distribution are independent of the executing machine, as all algorithms are deterministic.

6 Computational Efficiency

Much of the discussion and development around chunking algorithms has been fueled by the need to achieve higher throughput while maintaining good deduplication. Rabin, one of the oldest CDC algorithms, is widely known to be slow. This has spurred the development of newer, faster CDC algorithms. We dedicate this section to investigate the performance of the algorithms in terms of throughput, or computational efficiency. We report on overall achievable throughput of the algorithms, as well as microarchitectural details to explain certain behaviors.

6.1 Setup and Methodology

We execute our benchmark program as described in Section 5.3 and collect both the execution time as well as `perf` counters. We normalize results by the size of the dataset where applicable. We expect different distributions for the recorded metrics: 1) For all microarchitectural performance counters, such as *instructions per cycle (IPC)*, number of instructions, number of branches, etc., we assume a normal distribution. As such, we derive the mean value and corresponding standard error for these metrics. We expect very little spread in most of these metrics through multiple runs, as we compile our program just once, and all algorithms are deterministic. The data confirms these expectations, with a standard error for *all* reported metrics of ≤ 0.01 , which we thus omit. 2) For the runtime of the benchmarks, and conversely the throughput as a measure of input processed per execution time, we assume a skewed distribution in accordance with [41]. As such, we report the median and *interquartile range (IQR)*, calculated as $Q_3 - Q_1$.

We also evaluate the SIMD optimization for BSW algorithms discussed in Section 4.1. For this, we use an existing implementation⁶ of the 64-bit Gear algorithm with manual vectorization, which we refer to as Gear64+. The instruction set utilized by the implementation depends on the support of the current machine. On our system, the algorithm uses AVX2 instructions. Note that the original Gear algorithm uses 32-bit. For transparency about what contributed to the difference in results, we therefore additionally implement 64-bit Gear without the use of manual vectorization, denoted Gear64.

6.2 Overview on Synthetic Dataset

We first provide an overview of the achievable throughput using the various algorithms. To that end, we evaluate each of them on the RAND dataset with a target chunk size of 2 KiB (Figure 5). Due to the specific content-dependence of BFBC with regard to its efficiency, we additionally present measurements for BFBC on the CODE dataset, referred to as BFBC-L. Evaluating on the random dataset corresponds to the same theoretical considerations on expected chunk size

6. <https://crates.io/crates/gearhash>

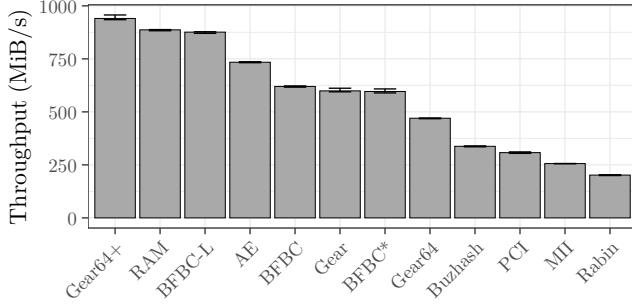


Fig. 5: CDC throughput, median values and quartiles, $\mu = 2 \text{ KiB}$, RAND dataset. BFBC-L indicates BFBC on the CODE dataset.

TABLE 6: Computational Performance of Chunking Algorithms ($\mu = 2 \text{ KiB}$, RAND Dataset; BFBC-L Indicates BFBC on the CODE Dataset; Br./B = Branches/Input Byte. BM-% = Branch Misprediction Percentage)

Alg.	Throughput (MiB/s)					
	Median	IQR	Inst./B	IPC	Br./B	BM-%
FSC	2529	51	0.40	0.31	0.06	0.99
Gear64+	941	22	8.03	2.33	0.73	0.38
RAM	887	6	8.45	2.40	1.75	0.46
BFBC-L	876	8	10.93	3.05	1.79	0.11
AE	734	5	11.42	2.56	2.56	0.19
BFBC	620	6	16.78	3.24	2.49	0.05
Gear	599	17	13.39	2.50	1.73	0.09
BFBC*	596	19	17.38	3.24	2.56	0.06
Gear64	470	3	15.41	2.24	2.57	0.06
Buzhash	338	4	29.34	3.06	5.02	0.10
PCI	308	7	31.25	3.03	2.54	0.07
MII	256	1	12.43	0.99	2.16	15.06
Rabin	202	4	34.27	2.19	5.04	0.05

as given in Section 4. To better understand these results, we furthermore provide relevant microarchitectural counters in Table 6. We investigate the results in detail in the following.

6.2.1 BSW Algorithms

The BSW algorithms (Rabin, Buzhash, Gear, and PCI) generally make up the lower end of the performance scale, with the exception of Gear. Although these algorithms all operate in $\mathcal{O}(1)$ per input byte, we can see that they differ substantially in constant complexity: Rabin and PCI place past 30 instructions per input byte, Buzhash places just shy of that, which results the highest throughput of the three at $\approx 340 \text{ MiB/s}$. Gear uses only ≈ 13 instructions per byte, leading to a much higher throughput of $\approx 600 \text{ MiB/s}$.

The SIMD implementation (Gear64+) uses multiple “heads”, spaced out by a number of bytes dependent on the instructions supported on the target machine. Each head then performs the Gear algorithm as usual, although all heads execute in parallel using SIMD instructions. Once any of the heads finds a boundary, the code falls back to a scalar variant in order to ensure none of the *previous* heads detects a chunk point in any of the yet-unprocessed data. The SIMD

TABLE 7: Performance of Scalar and SIMD Implementations of Gear64 on the RAND Dataset (Br. = Branches)

Alg.	μ (B)	Throughput (MiB/s)				
		Median	IQR	Inst./B	IPC	Br. ($\times 10^9$)
Scalar	512	452.2	6.1	15.57	2.27	27.82
SIMD	512	732.4	23.9	11.28	2.57	11.25
Scalar	8192	473.6	16.6	15.38	2.23	27.50
SIMD	8192	1024.1	13.2	7.02	2.21	6.30

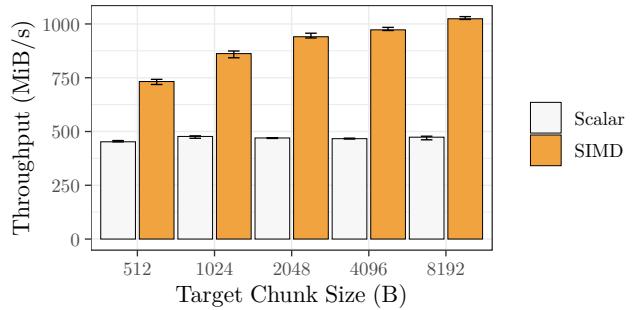


Fig. 6: Throughput of scalar and SIMD implementations of Gear64, median and quartiles, RAND dataset.

implementation is around twice as fast as the scalar variant (Gear64). This difference becomes more pronounced the larger the target chunk sizes (cf. Figure 6). This is expected from the implementation, *i.e.*, the code needs to fall back to a scalar version less frequently for larger target chunk sizes since fewer chunk boundaries are found. In terms of microarchitectural performance (Table 7), we can see that the larger the target chunk size, the fewer instructions per byte are utilized by the SIMD version, which again follows from the implementation falling back to scalar code less frequently. The same applies to branches. Finally, by comparing Gear with Gear64, we can also see that there is an expected efficiency drawback that comes with generating larger hashes. Note that it should also be possible to apply manual vectorization to the 32-bit variant, as well as to other BSW algorithms, but this is not the focus of our work. In conclusion, while not always possible, data-parallelism can offer a large increase in performance.

6.2.2 Extremum-Based Algorithms

The algorithms utilizing local extrema (AE, RAM, and MII) generally perform very well, with the exception of MII. Of the three, RAM performs the best at $\approx 870 \text{ MiB/s}$, markedly better than AE at $\approx 730 \text{ MiB/s}$. We thus conclude that RAM achieves its goal of being a faster AE [10]. MII, although algorithmically simple, performs poorly at only $\approx 260 \text{ MiB/s}$. It requires only slightly more instructions per input byte than AE and RAM. However, we can see that it utilizes the CPU poorly at only $\approx 1 \text{ IPC}$. A closer look shows that this is due to the poor predictability of the branch comparing the current input byte to the previous one. We find that 15% of branches are mispredicted, the worst behavior of all algorithms examined.

6.2.3 BFBC

As the results in Section 7 will demonstrate, the original BFBC algorithm is not well-suited for high-entropy datasets. We compare BFBC on RAND to BFBC on CODE (denoted BFBC-L). BFBC struggles on datasets like RAND because the top-3 most frequent byte pairs occur as frequently as any other. This causes the algorithm to skip the minimum chunk size, yet fail to find a boundary quickly. This is evident in the higher number of instructions and branches per input byte. On low-entropy datasets like CODE the top-3 most frequent byte pairs are usually encountered shortly after skipping the minimum chunk size. We can also compare BFBC to BFBC* to judge the effects of skipping. Recall that BFBC* uses a different set of divisors and does not skip, but otherwise functions the same as standard BFBC. In particular, comparison against the chosen byte pairs happens in constant time in our implementation, *i.e.*, the number of selected byte pairs should have no effect on throughput. The absence of skipping results in a slight elevation in number of instructions and branches per input byte, which leads to a correspondent decrease in throughput.

6.2.4 Fixed-size Chunking

Finally, if content-defined chunking is not a concern, fixed-size chunking unsurprisingly outperforms all CDC contenders at more than 2.5 GiB/s of processed input. These results also show that our benchmarking system is not limited by I/O.

6.3 Key Takeaways

In this section, we focused on performance in terms of throughput as a measure of computational efficiency for a range of algorithms. In summary, we can derive the following conclusions:

6.3.1 Complexity in Pseudocode versus Performance on Real-world Systems

Modern CPUs are complex. It is not always obvious how an algorithm, given in pseudocode, performs on a real system. We were able to show that various microarchitectural factors, in particular caching and branch predictability, play an important role in achieving higher throughput. Some algorithms (*e.g.*, MII) are algorithmically simple yet difficult for the machine to execute. Others (*e.g.*, Gearhash) are both simple and fast to execute.

6.3.2 BFBC Variants

We found that the BFBC variants are relatively fast, with differences stemming from skipping over parts of a new chunk and the dataset. We did not evaluate the time it takes or the feasibility of deriving byte pair frequencies before chunking. In general, while the algorithms perform well, their application is potentially limited by this requirement.

6.3.3 BSW Algorithms

Of all the BSW algorithms, Gearhash presents itself with low algorithmic complexity and good real-world performance. If supported on the machine, and an implementation is feasible, vectorization of these algorithms, in particular of Gear, achieves very high throughput. In our evaluation, SIMD-accelerated Gear outperforms *all* other CDC algorithms. We

postulate whether it would be more fruitful in the future to develop algorithms with this in mind, instead of developing new CDC algorithms.

6.3.4 Algorithms Using Local Extrema

Of the algorithms utilizing extrema (MII, RAM, AE), we find that MII performs poorly due to the difficulty of predicting its branch. Both AE and RAM perform very well overall. Between the two, we find that RAM indeed outperforms AE.

7 Chunk Size Distribution

In this section, we explore the chunk size distribution of the selected algorithms. Our analysis is based on empirical data collected from running these algorithms on diverse datasets, reflecting both high-entropy and low-entropy scenarios. The chunk size distributions produced by any CDC algorithm can be characterized by two relevant statistics: 1) The empirical mean chunk size \bar{c}_s produced, which should be close to the target μ . Intuitively, this reflects how easy it is to configure an algorithm for a target and how predictable its behavior is. 2) The spread of the distribution around the mean, calculated as the empirical standard deviation, s . As laid out in Section 2.2, pathological chunk sizes are undesirable. On the other hand, an algorithm must show some flexibility in the size of chunks produced in order to effectively combat the boundary shift problem. Additionally, it is helpful to not just evaluate an algorithm based on \bar{c}_s and s , but also examine the shape of the distribution function, and understand the mechanics behind it.

7.1 Distributions

In Figure 7, we show the distribution of the produced chunk sizes for a selection of algorithms. For ease of interpretation, and because their distributions resemble Rabin's, we do not show Buzhash and Gear, although we discuss minor differences in Section 7.2. For a comprehensive overview of the distributions across all settings, and to better distinguish between individual algorithms, please visit <https://mrd0ll4r.github.io/cdc-algorithm-tester>, where we provide the data through interactive charts.

Strikingly, almost all algorithms exhibit a similar shape. The reason is that they underlie the same stochastic property: Each position of the data stream, looked upon independently, is equally likely to become a chunk cut-point (cf. Section 2). However, each position's probability also depends on all previous positions not having fulfilled the same matching condition. Therefore, we observe that almost all distributions peak at a minimal size, determined by the window size or other feature of the algorithm, and then decay, forming a heavy-tailed distribution. The BSW algorithms, utilizing small windows of typically 32 B to 256 B, form a large number of small chunks, but then compensate by producing a long tail that shifts the mean closer to the target. This inevitably results in a large chunk-size variance. The local maxima-based algorithms, in contrast, present themselves with a minimum chunk size, bound by the horizon size in AE and RAM, close to the target μ . Their distributions drop much more rapidly, with most of the chunk sizes forming within a smaller region around the target. The only exception to this pattern is the distribution of BFBC on RAND. Recall how BFBC operates on a fixed

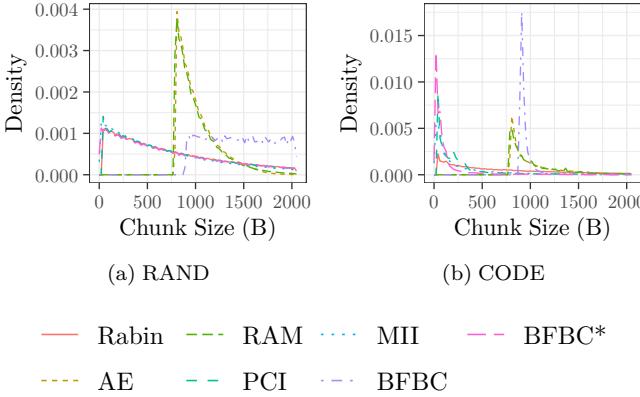


Fig. 7: Chunk size distributions for target chunk size $\mu = 1024$ (in case of MII, $\mu = 770$).

set of $k = 3$ most popular byte pairs of the dataset (cf. Section 4.6), as well as skipping $\mu - 128$ B. This is a valid strategy for low-entropy datasets, as the most frequent byte pairs are expected to occur very frequently, leading to chunk boundaries closely aligning with μ . Because the top frequent byte pairs in those datasets occur with such high frequency, chunks never grow much beyond the minimum chunk size. For example, we find that in CODE the most frequent byte pair were two spaces, presumably for indentation; in LNX, PDF, and WEB, we find that it is two null bytes, presumably for padding. On pseudorandom data, however, the top most frequent byte pairs occur as frequently as any other. This leads to an almost uniform distribution of chunk sizes after the skipped minimum size, with single chunks sized up to 300 KB.

7.2 Quantitative Results

We now move to present the mean chunk size $\bar{c}s$ and standard deviation s for all datasets and algorithms evaluated. To that end, we include an extensive table in the appendix (Table 11), and an aggregated version here (Table 8). The cell colors follow a continuous scale, where a light color indicates values near the optimum. For the mean $\bar{c}s$, this is μ . The color range reaches its maximum where the mean deviates $\pm 100\%$ from the target μ . For the standard deviation s , the color range is relative to the empirical mean $\bar{c}s$, in a range $[0, 2\bar{c}s]$. Recall (cf. Section 4.3) that we examine all algorithms without a limit on the chunk size.

We make the following observations: Firstly, judging by results rendered for RAND, we see our formulas for determining parameters for AE and RAM confirmed, as means are observed very closely to the desired target. The empirical means for MII, on the other hand, exceed the target by 13–15 %. Additionally, we observe that, for almost all algorithms, chunk-size variance seemingly correlates inversely with dataset entropy, *i.e.*, datasets with low entropy tend to lead to higher variance in chunk sizes produced. Often, with BSW algorithms, we furthermore observe an increase in variance when target chunk sizes are higher. With respect to the mean, the performance varies with no obvious pattern.

TABLE 8: Aggregated Overview of the Relative Performance of Chunk-size Variance and Mean

Algorithm	Mean				SD					
	RAND	LNX	PDF	WEB	CODE	RAND	LNX	PDF	WEB	CODE
Rabin	Optimal	Pathological								
Buzhash	Pathological									
Gear	Pathological									
Gear NC-1	Pathological									
Gear NC-2	Pathological									
Gear NC-3	Pathological									
AE	Pathological									
RAM	Pathological									
PCI	Pathological									
MII	Pathological									
BFBC	Pathological									
BFBC*	Pathological									

Optimal Pathological

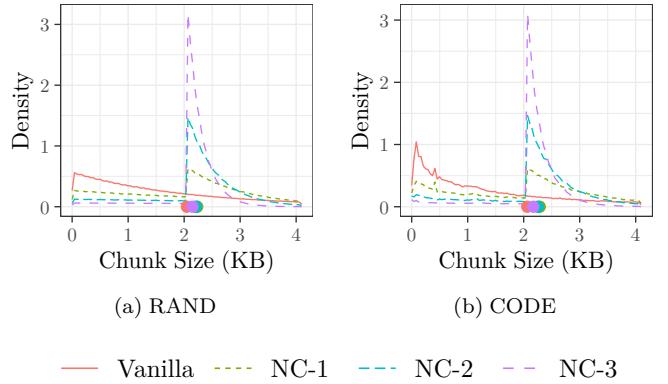


Fig. 8: Effects of NC on the chunk size distribution of Gear, with target chunk size 2 KiB. Dots beneath the x-axis mark the mean chunk size produced.

7.2.1 BSW Algorithms and Normalized Chunking

Different hash functions within BSW algorithms yield different distributions. While Gear seems to be better at maintaining means close to the target, all BSW algorithms notoriously struggle with chunk-size variance. This effect is gradually reduced with increasing levels of NC, see Figure 8.

7.2.2 Local Extrema-based Algorithms

The lowest levels of variance are produced by AE. Historically, this was also the motivation behind local extrema approaches in general. In contrast, RAM fulfills this promise only on the RAND dataset. On the realistic datasets, the results are pathological as RAM fails to find chunk boundaries. The authors warned against poor performance on files containing low-entropy. In their paper [10], the experiments yielded very similar results in comparison with AE. Surprisingly, we find also pathological performance on LNX and PDF, which still represent realistic datasets with fairly high entropy.

7.2.3 Algorithms for Data Synchronization

PCI exhibits a significant deviation from the target chunk size and high chunk-size variance across most settings. Its predecessor, MII, although restrictive in the targets it can tune to, performs slightly better in this regard. However, since chunk-size variance is not a critical concern in the application for data synchronization, this may not be a substantial drawback.

7.2.4 BFBC

Finally, BFBC attains pathological means if datasets have high entropy, for reasons explained previously. The BFBC* variant fixes this issue. However, high chunk-size variance remains a problem as the uniform distribution of the divisors within the datasets is never given, neither with BFBC nor with BFBC*.

7.3 Key Takeaways

In this section, we focused on the distribution of chunk sizes produced by CDC algorithms on both synthetic as well as real-world datasets. In summary, we arrive at the following conclusions.

7.3.1 Heavy-Tailed Distributions

Almost all algorithms produce chunk sizes with a heavy-tailed distribution. We find a large number of small chunks shortly after a minimum defined by the algorithm, *e.g.*, the size of the window, which is always smaller than the target chunk size. This is followed by a heavy tail, which moves the produced mean closer to the target.

7.3.2 Chunk-Size Variance

We find that aforementioned skew in the distributions is more pronounced for BSW algorithms, which operate on a relatively small window. AE and RAM produce a distribution of similar shape, although forming much more closely around the target mean, leading to lower chunk-size variance.

7.3.3 Normalized Chunking

Gear with normalized chunking presents a variant with lower chunk-size variance than plain Gear through its use of two matching conditions. This presents itself in the distribution as two pronounced peaks.

8 Deduplication Ratio

The degree of achievable deduplication is a key metric in the evaluation of CDC algorithms. It determines how effectively the algorithm can identify and eliminate redundant data, thereby minimizing storage requirements. In this section, we investigate the comparative deduplication performance of the selected algorithms. Additionally, we explore how the characteristics of the dataset, such as entropy, influence the deduplication achieved by each algorithm. We proceed similar to earlier analyses: We will apply each algorithm to each dataset, report on the results, and derive insights into the characteristics of each algorithm and dataset. As key metric we use the deduplication ratio. It indicates the ratio of storage space that can be saved due to the elimination of redundant chunks, as a value in $[0, 1]$. For instance, a deduplication ratio of 0.4 on a dataset of 1 GB indicates that the same dataset

can be represented in 600 MB of unique chunks. Note that we do not consider the overhead of metadata here. We evaluate the algorithms on the four realistic datasets CODE, WEB, LNX, and PDF. Evaluations on the RAND dataset are moot, as no deduplication is possible with realistic chunk sizes. For the other datasets, we determined optimal window sizes for the BSW algorithms (Figure 10).

8.1 General Overview

In Figure 9, we give an overview of the deduplication performance of each chunking algorithm on each dataset and over a range of target chunk sizes. We summarize Rabin, Buzhash, and Gear into one category of BSW algorithms. These algorithms mainly differ in their choice of hash function. As investigated in Section 7, this *does* lead to differences in chunk size distributions. Interestingly, however, we find only minuscule (< 0.01) differences in the deduplication ratios achieved. Generally, for all algorithms, the deduplication ratio drops with growing target chunk sizes, although to varying degrees. This is expected as smaller chunks have a higher chance of being duplicates. It is particularly noticeable in the CODE dataset. We suspect that this is due to the nature of the content: Minor modifications to source code (which is purely text-based) affect only a few bytes, while modifications to binary files (a large portion of the files in WEB) potentially affect a longer sequence of bytes, up to the entire file.

The top performers across all settings are the BSW algorithms and AE, with no considerable differences among each other. However, there is a slight divergence as the target chunk size increases. Presumably, AE's deduplication performance lags behind as the window size increases. We speculate that this difference may become more pronounced for target chunk sizes larger than those tested. Notably, although PCI occasionally emerges as the top performing (LNX and PDF on high targets), it is essential to consider this in the context of the mean chunk sizes it generates (*e.g.*, $< 1 \text{ KiB}$ on $\mu = 8 \text{ KiB}$). Furthermore, its degradation for larger chunk sizes is less pronounced than in the other algorithms. RAM also establishes itself among the top performers, with the exception of the CODE dataset. The problem with RAM on CODE likely stems from the fact that it forms very large chunks and thus fails to find duplicates (cf. Section 7). Finally, we also observe MII with competitive deduplication results in most scenarios. BFBC and BFBC* both fall behind the competition. Despite being among the top performers on CODE on a target of 512 B, their performance declines rapidly when run with higher target chunk size configurations. This is in contrast to results presented in the original publication [12], which we discuss in Section 9.1.3.

8.2 Normalized Chunking

The goal of NC, as proposed for FastCDC [31], is the reduction of chunk-size variance. We investigated this claim in Section 7. The question remains as to how this affects deduplication performance. Because the same byte sequences at different positions will be subject to inconsistent matching rules, it is plausible to expect a potentially degrading effect. We measure the attained deduplication ratio with three levels of normalized chunking applied to Gear, in comparison to vanilla

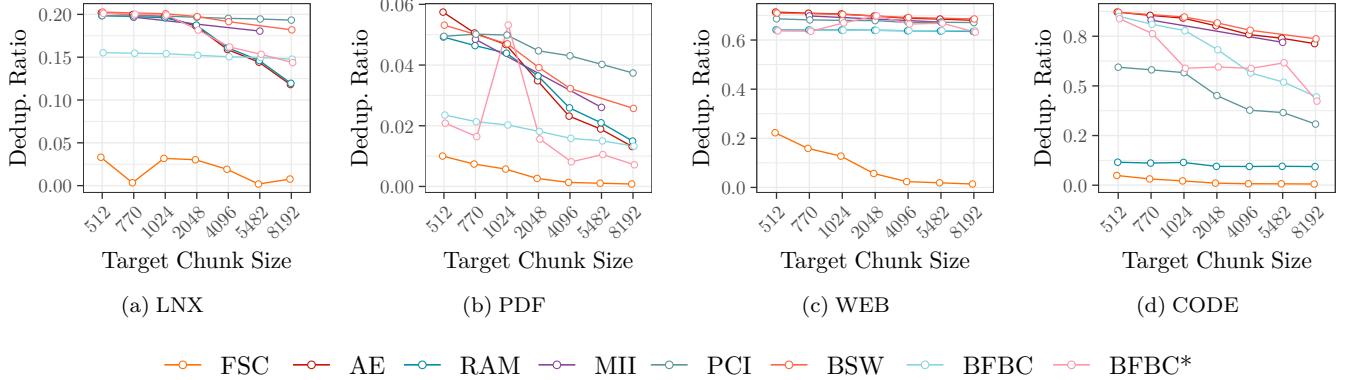


Fig. 9: Overview of deduplication ratios. Note the varying scales on the y-axis.

Gear without NC. The results are presented in Figure 11. Surprisingly, even with NC-3, we see only marginal detrimental effects. The largest difference can be observed at a target of 8 KiB on CODE (Figure 11d), where NC-2 surpasses Vanilla by two percentage points. While generally, the performance differences are minor, the implications of this are large since NC can drastically reduce chunk-size variance.

9 Discussion

Our theoretical analyses and experimental evaluations conducted throughout this study reproduce existing results but also reveal new insights. The theoretical analysis uncovered novel aspects and corrects existing formulas, extending the original contributions of the algorithms’ developers. Through rigorous and impartial experimentation, we identified previously unreported performance characteristics while also validating many of the original claims. With our discussion, we aim to synthesize those findings, providing a comprehensive overview of the deduplication landscape and also offering practical recommendations based on our findings.

9.1 Contrasting Results

While our study confirms many of the claims made by the algorithm developers, we also encountered several discrepancies. These differences arose from factors such as variations in dataset characteristics, experimental setups, and implementation details. Thus, this section is dedicated to contrasting our results with the ones from the literature, and reflecting on their implications.

9.1.1 AE

As we illuminate in Section 4.2, the formula by which the parameterization in AE is derived does not reflect the actual behavior of the algorithm fully, but has been disseminated in this form throughout several studies [8], [15], [22], [27], [36], [37], and moreover, the open-source chunking algorithm evaluation platform Destor and DedupBench, whose authors we have notified about the issue. According to our own experiments on RAND, the previously established formula renders chunks 13 % smaller than the target when it is set to 512 B, and 39 % smaller with a target of 8 KiB. In fact, the higher the target, the more pronounced the deviation. Because

smaller chunks are easier to deduplicate, earlier results on the deduplication ratio of AE must be taken with a grain of salt. In the original paper [22], the authors suggest that AE is superior to BSW in terms of deduplication. While AE has shown superior performance in some instances of our experiment, and is competitive in most, its performance sometimes degrades with higher target chunk sizes. Our results suggest that the AE algorithm cannot robustly handle high target chunk sizes, compared to, e.g., BSW. It is likely that this has been overlooked previously due to aforementioned incorrect target chunk size calculations and evaluations on lower targets.

9.1.2 RAM

We demonstrate the incapability of RAM to deal with low-entropy datasets. This is especially pronounced on CODE (cf. Table 11). The authors of RAM noticed this flaw [10]; however, they did not present the extent of this shortcoming. In their study, they evaluate RAM on high-entropy datasets that, as we argue, do not represent realistic candidates for data deduplication. Indeed, deduplication is only evaluated in the form of “bytes saved per second”, which counterintuitively conflates throughput with deduplication. Moreover, the paper suggests chunk-size variances similar to AE, while the experiments on none of our real-world datasets support this statement. Even with our LNX dataset, which is composed of similar files than their “Dataset 1”, we cannot confirm comparable results.

9.1.3 BFBC

The authors of BFBC [12] compare their performance to Rabin’s in an experiment that uses two source code-based datasets, similar to our CODE dataset. On both datasets, the deduplication achieved by BFBC outperforms Rabin’s, whereas in our experiment, Rabin emerges as the superior algorithm. We note that their experiment differs from ours in enforcing minimum and maximum chunk size lengths. That is, they evaluate the algorithms on chunk size ranges instead of targets. In these experiments, BFBC performed best in the range 128–256 B. However, we note that the result they report is similar for the range 128–8192 B, which is the setting closest to our experiment on target 512 B. In Table 9, we contrast their results with ours. Even when considering the best results we achieved using BFBC, it is still outperformed

TABLE 9: Comparison of Deduplication Ratio Results Presented in [12] vs. in Our Own Experiments

	Their Results [12]		Ours
	Dataset 1	Dataset 2	CODE
BFBC	79.8 %	96.7 %	85.0 %
Rabin	68.2 %	92.6 %	87.1 %

by Rabin. We postulate that the chunk size limits imposed on Rabin in their work impacted the results negatively. Since the exact settings chosen for Rabin in their analysis are not disclosed, we are unable to confirm this hypothesis. Our comprehensive experiments and results ultimately do not support the claims of improved deduplication ratios for BFBC. In terms of throughput, as well, we find that more efficient BSW algorithms, *e.g.* Gear, outperform BFBC.

9.2 Summary

After an extensive performance evaluation of various algorithms across multiple settings, we now attempt to give a conclusive summary of the results and findings.

Our research indicates that traditional BSW algorithms are still unbeaten in terms of deduplication. However, we find that similar performance can also be attained with AE and MII. RAM proved to be very sensitive w.r.t. entropy in a dataset: On datasets with particularly low entropy, such as text-based data (CODE), the algorithm fails to find chunk boundaries, resulting in the generation of pathologically large chunks. Ultimately, these effects render the slight efficiency gains of RAM over AE negligible.

Only slightly behind AE in terms of throughput, Gear proves to be the fastest BSW algorithm and simultaneously maintains the healthiest levels of chunk size mean and variance among that group. Furthermore, with the help of NC, we find that the otherwise high variance in chunk size can be mitigated without impairing the deduplication ratio. This technique becomes especially powerful when combined with minimum chunk size skipping, which can improve throughput significantly [40].

Although the issue of chunk-size variance is of lesser importance in the context of data synchronization, which is the intended use case for MII and PCI, we were unable to identify any advantages over alternative solutions such as AE or Gear, which offer comparable or superior deduplication and significantly higher throughput.

Finally, we also find that BFBC and BFBC* offer no real advantage over algorithms such as Gear or AE. Moreover, whereas the algorithm for BFBC does not consider higher entropy datasets, both BFBC and BFBC* have no mechanism to ensure consistent chunk sizes. Furthermore, they add complexity through the initial process of collecting statistics over a dataset. We note, however, that the optimized divisor selection algorithm in BFBC* successfully rectifies chunk size averages, and thus offers a real improvement over the original implementation of BFBC.

These findings are summarized in simplified form in Table 10. A checkmark indicates that the algorithm was among the top performers with regard to the respective metric. The table reveals AE as the only CDC algorithm competitive on

TABLE 10: Performance Summary of CDC Algorithms Based on Our Experiments

Algorithm	Dedup.	Throughput	Chunk Sizes	
			Mean	SD
Rabin, Buzhash	✓			
Gear	✓	✓	✓	
Gear with NC	✓	✓	✓	✓
AE	✓	✓	✓	✓
RAM	✓ [†]	✓		
PCI				
MII		✓		
BFBC			✓	
BFBC*			✓	✓

[†] except on the CODE dataset

all metrics. However, this result requires a nuanced interpretation. While AE performed admirably within our tested range of target chunk sizes, its deduplication efficacy may degrade with larger chunk sizes more strongly than the more robust BSW algorithms. Additionally, our analysis reveals that NC significantly reduces Gear’s chunk-size variance, though not to the level achieved by AE, without corrupting deduplication performance. This reduction also enables the safe skipping of a minimum chunk size, which in turn boosts throughput. Given these considerations, Gear with NC emerges as a robust and efficient alternative, making it an equally attractive choice for various applications.

10 Conclusion

In this work, we present a comprehensive and impartial evaluation of state-of-the-art CDC algorithms. Furthermore, we provide an analytical framework, as well as a set of benchmarks for the evaluation of future advancements in CDC. Our rigorous theoretical analysis and extensive experimental validation yield both reproducible results and novel insights. Our comparison highlights several limitations and shortcomings that are not apparent from previous studies. We find that many researchers promote their algorithm under conditions or assumptions that fail to hold up in realistic scenarios, often relying on biased datasets, misleading metrics, or narrowly defined test cases. Finally, we recognize Gear with NC and AE as the most attractive choice for CDC, despite more recent advancements in algorithm development. We believe that our findings and methodologies will significantly contribute to the optimization of storage and bandwidth efficiency in cloud computing infrastructures.

Acknowledgments

This work was supported by the Dutch national NWO/TKI science grant BLOCK.2019.004, as well as the German Research Foundation (DFG) within the Collaborative Research Center (CRC) SFB 1053: MAKI.

References

- [1] D. R.-J. G.-J. Rydning, J. Reinsel, and J. Gantz, “The digitization of the world from edge to core,” *Framingham: International Data Corporation*, vol. 16, 2018.

- [2] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," *ACM Transactions on Storage (ToS)*, vol. 7, no. 4, pp. 1–20, 2012.
- [3] A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data deduplication—large scale study and system design," in *USENIX Annual Technical Conference*, vol. 2012, no. 2012, 2012, pp. 285–296.
- [4] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, "Characteristics of backup workloads in production systems," in *FAST*, vol. 12, 2012, pp. 4–4.
- [5] M. O. Rabin, "Fingerprinting by random polynomials," *Technical report*, 1981.
- [6] R. M. Karp and M. O. Rabin, "Efficient randomized pattern-matching algorithms," *IBM journal of research and development*, vol. 31, no. 2, pp. 249–260, 1987.
- [7] A. Z. Broder, "Some applications of rabin's fingerprinting method," in *Sequences II*. Springer, 1993, pp. 143–152.
- [8] W. Xia, X. Zou, H. Jiang, Y. Zhou, C. Liu, D. Feng, Y. Hua, Y. Hu, and Y. Zhang, "The design of fast content-defined chunking for data deduplication based storage systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 9, pp. 2017–2031, 2020.
- [9] A. Anand, C. Muthukrishnan, A. Akella, and R. Ramjee, "Redundancy in network traffic: findings and implications," in *Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, 2009, pp. 37–48.
- [10] R. N. Widodo, H. Lim, and M. Atiquzzaman, "A new content-defined chunking algorithm for data deduplication in cloud storage," *Future Generation Computer Systems*, vol. 71, pp. 145–156, 2017.
- [11] C. Zhang, D. Qi, W. Li, and J. Guo, "Function of content defined chunking algorithms in incremental synchronization," *IEEE Access*, vol. 8, pp. 5316–5330, 2020.
- [12] A. S. M. Saeed and L. E. George, "Data deduplication system based on content-defined chunking using bytes pair frequency occurrence," *Symmetry*, vol. 12, no. 11, p. 1841, 2020.
- [13] J. D. Cohen, "Recursive hashing functions for n-grams," *ACM Transactions on Information Systems (TOIS)*, vol. 15, no. 3, pp. 291–320, 1997.
- [14] W. Xia, H. Jiang, D. Feng, L. Tian, M. Fu, and Y. Zhou, "Ddelta: A deduplication-inspired fast delta compression approach," *Performance Evaluation*, vol. 79, pp. 258–272, 2014.
- [15] Y. Zhang, D. Feng, H. Jiang, W. Xia, M. Fu, F. Huang, and Y. Zhou, "A fast asymmetric extremum content defined chunking algorithm for data deduplication in backup storage systems," *IEEE Transactions on Computers*, vol. 66, no. 2, pp. 199–211, 2016.
- [16] C. Zhang, D. Qi, Z. Cai, W. Huang, X. Wang, W. Li, and J. Guo, "Mii: A novel content defined chunking algorithm for finding incremental data in data synchronization," *IEEE Access*, vol. 7, pp. 86 932–86 945, 2019.
- [17] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, M. Fu, Y. Zhang, and Y. Zhou, "A comprehensive study of the past, present, and future of data deduplication," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1681–1710, 2016.
- [18] A. Muthitacharoen, B. Chen, and D. Mazieres, "A low-bandwidth network file system," in *Proceedings of the eighteenth ACM symposium on Operating systems principles*, 2001, pp. 174–187.
- [19] S. Quinlan and S. Dorward, "Venti: A new approach to archival data storage," in *Conference on file and storage technologies (FAST 02)*, 2002.
- [20] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: local algorithms for document fingerprinting," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, pp. 76–85.
- [21] N. Bjørner, A. Blass, and Y. Gurevich, "Content-dependent chunking for differential compression, the local maximum approach," *Journal of Computer and System Sciences*, vol. 76, no. 3-4, pp. 154–203, 2010.
- [22] Y. Zhang, H. Jiang, D. Feng, W. Xia, M. Fu, F. Huang, and Y. Zhou, "Ae: An asymmetric extremum content defined chunking algorithm for fast and bandwidth-efficient data deduplication," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 1337–1345.
- [23] A. Tridgell, P. Mackerras *et al.*, "The rsync algorithm," 1996.
- [24] H. B. Jehlol and L. E. George, "Big data de-duplication using classification scheme based on histogram of file stream," in *2022 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE)*. IEEE, 2022, pp. 1–7.
- [25] ———, "Enhancing deduplication efficiency using triple bytes cutters and multi hash function," *International Journal of Intelligent Engineering & Systems*, vol. 16, no. 5, 2023.
- [26] A. Goel and C. Prabha, "A detailed review of data deduplication approaches in the cloud and key challenges," in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2023, pp. 1771–1779.
- [27] M. Ellappan *et al.*, "Dynamic prime chunking algorithm for data deduplication in cloud storage," *KSII Transactions on Internet & Information Systems*, vol. 15, no. 4, 2021.
- [28] D. Viji and S. Revathy, "Comparative analysis for content defined chunking algorithms in data deduplication," *Webology*, vol. 18, no. SpecialIssue2, pp. 255–268, 2021.
- [29] N. T. Spring and D. Wetherall, "A protocol-independent technique for eliminating redundant network traffic," in *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2000, pp. 87–95.
- [30] A. El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, "Primary data {Deduplication—Large} scale study and system design," in *2012 USENIX Annual Technical Conference (USENIX ATC 12)*, 2012, pp. 285–296.
- [31] W. Xia, Y. Zhou, H. Jiang, D. Feng, Y. Hua, Y. Hu, Q. Liu, and Y. Zhang, "FastCDC: A fast and efficient Content-Defined chunking approach for data deduplication," in *2016 USENIX Annual Technical Conference (USENIX ATC 16)*. Denver, CO: USENIX Association, Jun. 2016, pp. 101–114. [Online]. Available: <https://www.usenix.org/conference/atc16/technical-sessions/presentation/xia>

- [32] J. Sun, H. Chen, L. He, and H. Tan, "Redundant network traffic elimination with gpu accelerated rabin fingerprinting," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7, pp. 2130–2142, 2015.
- [33] B. Chapuis, B. Garbinato, and P. Andritsos, "Throughput: A key performance measure of content-defined chunking algorithms," in *2016 IEEE 36th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2016, pp. 7–12.
- [34] F. Ni, X. Lin, and S. Jiang, "Ss-cdc: A two-stage parallel content-defined chunking for deduplicating backup storage," in *Proceedings of the 12th ACM International Conference on Systems and Storage*, 2019, pp. 86–96.
- [35] W. Xia, L. Pu, X. Zou, P. Shilane, S. Li, H. Zhang, and X. Wang, "The design of fast and lightweight resemblance detection for efficient post-deduplication delta compression," *ACM Transactions on Storage*, vol. 19, no. 3, pp. 1–30, 2023.
- [36] X. Jin, H. Liu, C. Ye, X. Liao, H. Jin, and Y. Zhang, "Accelerating content-defined chunking for data deduplication based on speculative jump," *IEEE Transactions on Parallel and Distributed Systems*, 2023.
- [37] M. Ellappan and A. Murugappan, "A smart hybrid content-defined chunking algorithm for data deduplication in cloud storage," *Soft Computing*, pp. 1–16, 2023.
- [38] A. Liu, A. Baba, S. Udayashankar, and S. Al-Kiswany, "Dedupbench: A benchmarking tool for data chunking techniques," in *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2023, pp. 469–474.
- [39] E. Kruus, C. Ungureanu, and C. Dubnicki, "Bimodal content defined chunking for backup streams." in *Fast*, 2010, pp. 239–252.
- [40] Z. Xu and W. Zhang, "Quickcdc: A quick content defined chunking algorithm based on jumping and dynamically adjusting mask bits," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BD-Cloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 288–299.
- [41] (2023) Are your memory-bound benchmarking timings normally distributed? Visited 2024-08-25. [Online]. Available: <https://lemire.me/blog/2023/04/06/are-your-memory-bound-benchmarking-timings-normally-distributed/>



Marcel Gregoriadis received the B.Sc. degree from Stuttgart Media University in 2020 and the M.Sc. degree from Humboldt University of Berlin in 2023, both in computer science. He is currently working toward a Ph.D. in the Data-Intensive Systems group at Delft University of Technology, where he is conducting research on decentralized information retrieval with intersections to artificial intelligence.



Leonhard Balduf is a Ph.D. candidate at the Communication Networks Lab at TU Darmstadt in his third year. He received his B.Sc. from Munich University of Applied Sciences and the M.Sc. from Humboldt University of Berlin in 2021. His research interests are in measurement studies of distributed and peer-to-peer systems.



Björn Scheuermann is a professor of Communication Networks at TU Darmstadt, Germany. He obtained his Ph.D. from the University of Düsseldorf, Germany, in 2007. After professorships in Düsseldorf, Würzburg, Bonn and at Humboldt University of Berlin he joined TU Darmstadt in 2021. His research interests include network protocol design and analysis, networked systems security and network hardware engineering.



Johan Pouwelse is an associate professor at Delft University of Technology, specialized in large-scale cooperative systems. During his Ph.D., he created the first system for cooperative resource management. The resulting driver got accepted into the Linux kernel and is still used by every Android and iOS device. Also, he conducted the first resource usage measurements for IEEE 802.11b, known now as wifi. After receiving his Ph.D., he conducted one of the largest measurements of the BitTorrent P2P network. He founded the Tribler video-on-demand client in 2005, which has been installed by 1.8 million people over the past decade.

Appendix A: Chunking Algorithms

Here, we include the pseudocode to the algorithms outlined in Section 4.

Algorithm 1 BSW(d, l)

Input: Data stream d , data length l
Predefined: Window size w , bitmask length b

```

1: for  $i \leftarrow w$  to  $l$  do
2:    $f \leftarrow H(d[i-w], \dots, d[i])$ 
3:   if  $f \& (2^b - 1) = 0$  then
4:     return  $i$ 
5:   end if
6: end for

```

Algorithm 2 AE(d, l)

Input: Data stream d , data length l
Predefined: Horizon length h

```

1:  $x_{\text{val}} \leftarrow 0$ 
2:  $x_{\text{pos}} \leftarrow 0$ 
3: for  $i \leftarrow 1$  to  $l$  do
4:   if  $d[i] \leq x_{\text{val}}$  then
5:     if  $i = x_{\text{pos}} + h$  then
6:       return  $i$ 
7:     end if
8:   else
9:      $x_{\text{val}} \leftarrow d[i]$ 
10:     $x_{\text{pos}} \leftarrow i$ 
11:   end if
12: end for

```

Algorithm 3 RAM(d, l)

Input: Data stream d , data length l
Predefined: Horizon length h

```

1:  $x \leftarrow 0$ 
2: for  $i \leftarrow 1$  to  $l$  do
3:   if  $i \leq h$  then
4:     if  $d[i] > x$  then
5:        $x \leftarrow d[i]$ 
6:     end if
7:   else
8:     if  $d[i] \geq x$  then
9:       return  $i$ 
10:    end if
11:   end if
12: end for

```

Algorithm 4 MII(d, l)

Input: Data stream d , data length l

Predefined: Window size w

```

1:  $c \leftarrow 0$ 
2: for  $i \leftarrow 2$  to  $l$  do
3:   if  $d[i] > d[i-1]$  then
4:      $c \leftarrow c + 1$ 
5:   if  $c = w$  then
6:     return  $i$ 
7:   end if
8:   else
9:      $c \leftarrow 0$ 
10:  end if
11: end for

```

Algorithm 5 PCI(d, l)

Input: Data stream d , data length l

Predefined: Window size w , threshold θ

```

1:  $v \leftarrow (0, 0, \dots, 0)$ ,  $|v| = w$                                  $\triangleright$  Current window.
2:  $p \leftarrow 0$                                                $\triangleright$  Popcount in  $v$ .
3: for  $i \leftarrow 1$  to  $l$  do
4:    $p \leftarrow p - \text{POPCOUNT}(v[i \bmod w]) + \text{POPCOUNT}(d[i])$ 
5:    $v[i \bmod w] \leftarrow d[i]$ 
6:   if  $i \geq w$  and  $p \geq \theta$  then
7:     return  $i$ 
8:   end if
9: end for

```

Algorithm 6 BFBC(d, l)

Input: Data stream d , data length l

Predefined: Divisors D , minimum chunk length λ_{\min}

```

1: for  $i \leftarrow 1$  to  $l$  do
2:   if  $i > \lambda_{\min}$  then
3:     for each  $(b_0, b_1) \in D$  do
4:       if  $(d[i-1], d[i]) = (b_0, b_1)$  then
5:         return  $i$ 
6:       end if
7:     end for
8:   end if
9: end for

```

Algorithm 7 DETERMINEBFBCDIVISORS(F, μ, l)

Input: Frequencies of top-frequent byte pairs F , target chunk size μ , file length l

Output: Set of divisors as indices of F

```

1:  $D \leftarrow \{\}$ 
2: for  $i \leftarrow 1$  to  $|F|$  do
3:   if  $D = \emptyset$  then
4:     if  $\mu(\{i\}) \geq \mu$  then
5:        $D \leftarrow D \cup \{i\}$ 
6:     end if
7:   else
8:     if  $|\mu - \mu(D \cup \{i\})| < |\mu - \mu(D)|$  then
9:        $D \leftarrow D \cup \{i\}$ 
10:    end if
11:   end if
12: end for
13: return  $D$ 

```

Appendix B: Extended Results

TABLE 11: Means and Standard Deviation of Chunks Produced by Chunking Algorithms on Different Target Chunk Sizes and Datasets

Algorithm	512 B		737 B		1024 B		2048 B		4096 B		5152 B		8192 B	
	$\bar{c}s$	s												
RAND														
Rabin	542	511			1055	1024	2078	2047	4127	4091			8220	8179
Buzhash	576	512			1089	1026	2115	2053	4166	4101			8267	8213
Gear	512	510			1024	1023	2048	2047	4098	4096			8184	8175
Gear NC-1	558	339	993	613	1116	680	2233	1362	4464	2725	5045	3134	8928	5451
Gear NC-2	553	210	913	349	1105	420	2209	841	4420	1685	5391	2140	8842	3372
Gear NC-3	538	130	852	203	1076	261	2150	522	4300	1047	5479	1432	8603	2093
AE	512	136	769	179	1024	209	2048	252	4095	255	5480	255	8191	255
RAM	544	234	785	247	1030	252	2048	255	4096	255	5482	255	8192	255
PCI	506	465	761	731	1017	991	2085	2036	4292	4255	5388	5341	8690	8642
MII			887	882							6238	6228		
BFBC	22019	21604	22280	21604	22535	21603	23555	21602	25595	21598	26983	21596	29703	21594
BFBC*	509	507	765	764	1016	1015	2031	2028	4062	4062	5414	5413	8121	8128
LNX														
Rabin	507	512			923	1018	1617	1996	2628	3810			3842	6897
Buzhash	556	512			1016	1022	1848	2027	3235	3980			5247	7581
Gear	514	1235			1029	1902	2058	3075	4115	5288			8223	9512
Gear NC-1	560	1214	997	1680	1121	1792	2242	2727	4482	4339	5063	4761	8972	7338
Gear NC-2	555	1178	917	1531	1109	1691	2218	2477	4438	3719	5412	4238	8873	5808
Gear NC-3	540	1154	855	1452	1080	1636	2157	2342	4315	3410	5496	3928	8631	5063
AE	505	138	761	184	1019	219	2053	287	4112	354	5501	397	8217	455
RAM	544	1154	788	1383	1037	1589	2067	2212	4133	3788	5527	4365	8252	5319
PCI	238	2236	272	2413	310	2593	452	3211	583	3783	631	4009	740	4528
MII			920	1909							6239	9107		
BFBC	15625	33037	18427	35154	20115	36270	23905	38418	28185	40406	30447	41325	34288	42727
BFBC*	579	1464	895	2077	1195	2720	2448	6127	4961	14980	6474	21447	7794	28928
PDF														
Rabin	516	658			957	1213	1728	2281	2880	4304			4341	7778
Buzhash	539	577			946	1080	1612	2038	2555	3768			3645	6652
Gear	508	667			1018	1271	1971	2360	3838	4620			8352	9399
Gear NC-1	559	535	985	866	1103	939	2172	1766	4506	3240	5088	3644	8964	6351
Gear NC-2	553	433	911	634	1096	718	2232	1225	4450	2269	5416	2721	8823	4181
Gear NC-3	538	336	859	508	1084	584	2165	946	4314	1616	5484	2002	8575	3017
AE	507	146	772	202	1039	247	2109	372	4209	568	5628	743	8399	1047
RAM	618	4188	890	5052	1166	5779	2283	8091	4486	11300	5972	13043	8855	15834
PCI	250	2486	270	2628	297	2780	445	3636	530	4012	602	4394	700	4862
MII			685	3343							2397	8825		
BFBC	2445	8026	3447	9364	4253	10252	6652	12286	10116	14309	12068	15171	15803	16520
BFBC*	481	6907	997	10484	1025	9552	2051	14521	4106	31292	5482	25065	8242	40550
WEB														
Rabin	451	516			744	998	1113	1834	1511	3184			1832	5120
Buzhash	547	520			969	1018	1648	1992	2660	3839			3793	6652
Gear	516	550			1046	1111	2075	2184	4123	4267			8166	8497
Gear NC-1	564	356	1010	665	1129	728	2264	1461	4514	2889	5068	3290	9055	5710
Gear NC-2	557	218	922	368	1112	436	2229	889	4529	1788	5474	2266	8952	3595
Gear NC-3	539	135	856	212	1078	269	2166	531	4340	1109	5525	1482	8672	2211
AE	490	137	744	209	1002	253	2063	380	4164	554	5576	662	8339	990
RAM	601	12033	882	14679	1166	17054	2332	24339	4650	34631	6213	39897	9245	48800
PCI	495	3321	598	3796	674	4255	1090	6270	1426	11024	1704	11118	2043	13215

MII		856	1981					5018	11503		
BFBC	1256	18359	1872	22389	2422	25448	4357	34050	7558	44714	9516
BFBC*	453	11319	763	13940	1875	13780	2026	12155	4107	24116	5492
CODE											
Rabin	542	3006			1051	4281	2018	6210	3741	9208	6447
Buzhash	574	3084			1041	4229	1943	6014	3406	8638	5454
Gear	519	2951			1037	4261	2062	6278	4249	9906	8385
Gear NC-1	568	3045	1021	4118	1138	4352	2286	6278	4568	9201	9247
Gear NC-2	558	3004	934	3895	1123	4273	2252	6097	4495	8738	9146
Gear NC-3	540	159	859	3723	1083	4180	2171	5934	4387	8475	5565
AE	484	145	758	217	1043	289	2242	570	4647	1136	6268
RAM	15989	759876	26865	975111	33879	1070323	66846	1733430	129537	2407519	169158
PCI	4358	47128	5124	52722	5978	58003	12021	115074	17925	173403	21797
MII		563	3917								2334
BFBC	458	3046	734	3853	1002	4501	2078	6466	4207	9167	5642
BFBC*	506	7706	764	16540	1019	32168	2824	24637	4084	48854	5473

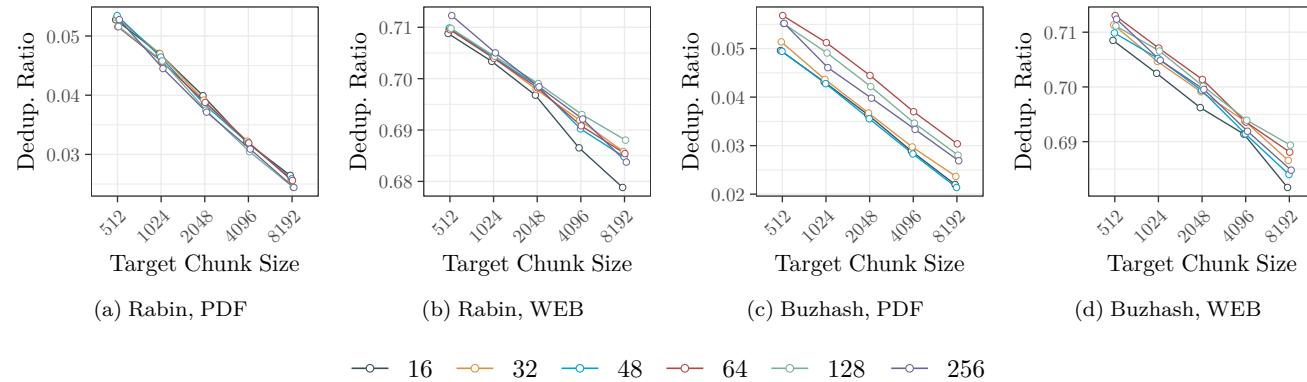


Fig. 10: Deduplication ratio of Rabin and Buzhash on different window sizes.

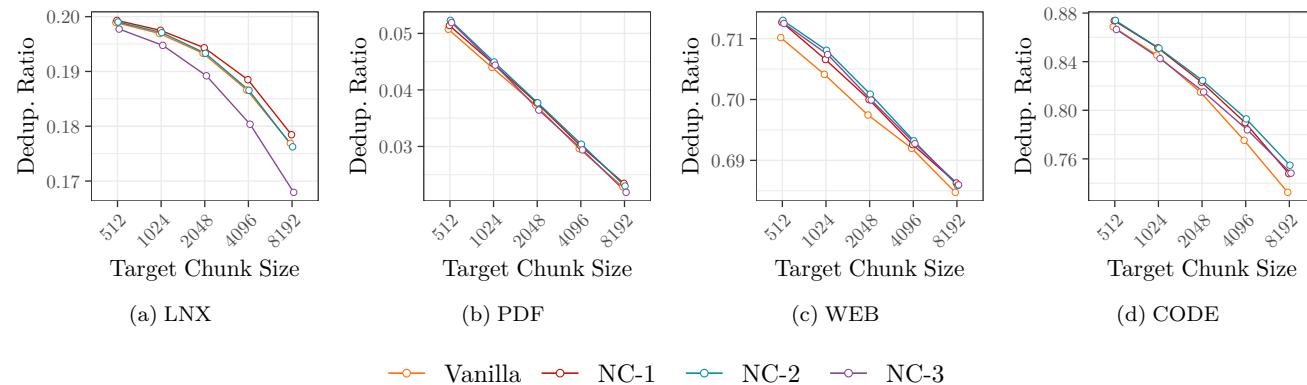


Fig. 11: Deduplication performance of Gear with different levels of NC, as well as without (Vanilla).