

# CIRCLLHIST

A LOG-LINEAR HISTOGRAM DATA STRUCTURE FOR IT INFRASTRUCTURE MONITORING

**Heinrich Hartmann**

heinrich.hartmann@circonus.com  
Circonus

**Theo Schlossnagle**

theo.schlossnagle@circonus.com  
Circonus

January 22, 2020

## ABSTRACT

The circllhist histogram is a fast and memory efficient data structure for summarizing large numbers of latency measurements. It is particularly suited for applications in IT infrastructure monitoring, and provides nano-second data insertion, full mergeability, accurate approximation of quantiles with a-priori bounds on the relative error.

Open-source implementations are available for C/lua/python/Go/Java/JavaScript.

## 1 Introduction

Latency measurements have become an important part of IT infrastructure and application monitoring. The latencies of a wide variety of events like requests, function calls, garbage collection, disk IO, system-call, CPU scheduling, etc. are of great interest of engineers operating and developing IT systems.

There are a number of technical challenges associated with managing and analyzing latency data. The volume emitted by a single data source can easily become very large. Furthermore, data has to be collected and aggregated from a large number of different sources. The data has to be stored over long time periods (months, years), in order to allow historic comparisons and long-term service quality estimations (SLOs).

In order to address these challenges a compression scheme has to be applied, that drastically reduces the size of the data to be stored and transmitted. Such a compression scheme needs to allow at minimum (1) arbitrary aggregation of already compressed data, (2) accurate quantile approximations, with a-priori bounds on the relative error (3) accurate counting of requests larger or lower than a given threshold. Furthermore it's beneficial if (4) information about the full distribution is retained, so that general probabilistic modeling techniques can be applied.

Traditionally monitoring tools, either store raw data on which calculations are performed (e.g. ELK<sup>1</sup>) or compute latency quantiles on each host separately and store them as numeric time series (e.g. statsd<sup>2</sup>). Both approaches have obvious drawbacks. The high volume of data makes raw data storage uneconomical for sources like request latencies, and impractical for high volume sources like function-call or system-call latencies. Direct calculation of quantiles does not allow further aggregation, so that accurate quantiles for the total population can not be calculated.

Circonus has addressed this problem with the circllhist data-structure, that we describe in this document. It has been in and production use since 2011. Figure 1 shows an example of a visualized circllhist in the Circonus product. We have talked about this at various conferences and blogs (e.g. [10],[11],[12], [7]). An open source implementation is available at [13]. However, no academic paper was published until now.

This document describes the circllhist data-structure and compares it to other methods that have been adopted by other monitoring vendors since then.

## 2 Related Work

There is also a fair bit of work in the academic literature on the problem of efficiently calculating aggregated quantiles since 1980. This runs under the name “mergeable summaries” and “quantile sketches”. A good summary of these methods can be found in [1], Section 1.2. Here we focus on methods that have been adopted in practice.

<sup>1</sup><https://www.elastic.co/what-is/elk-stack>

<sup>2</sup><https://github.com/statsd/statsd>

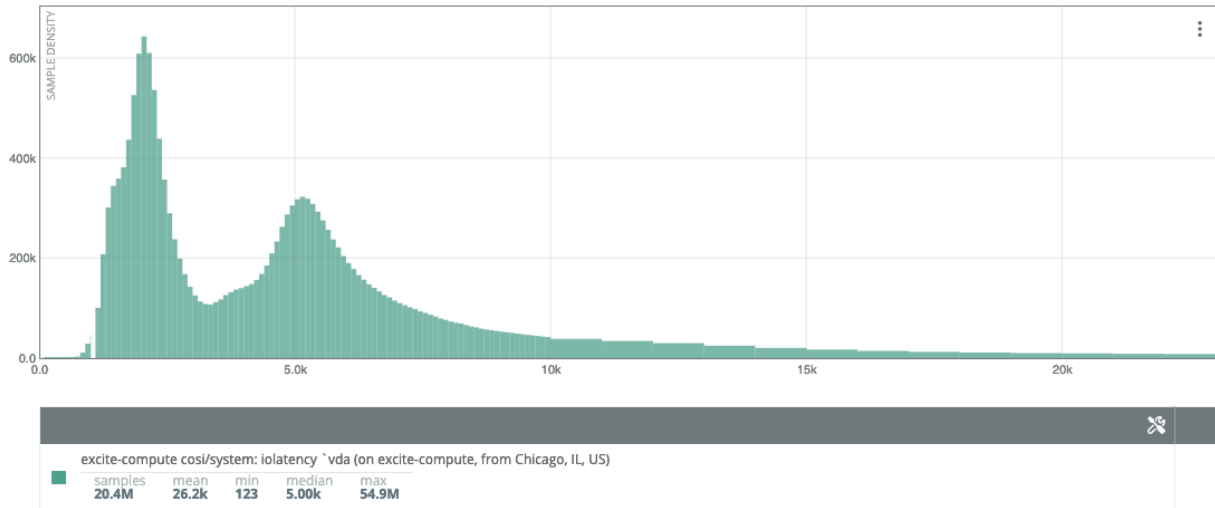


Figure 1: Circllhist representation of 20M block-level disk I/O latencies measured over the course of one month.

A very similar approach to ours has been suggested by G. Tene from Azul Systems who developed a High Definition Range (HDR) Histogram data-structure [5] to capture latency data for benchmarking applications.

T. Dunning and O. Ertl developed the t-digest data structure in [3], which is used in the Wavefront monitoring product. The t-digest aggregates nearby points into clusters of adaptive size, in such a way, that high resolution data is available at the tails of the distribution, where it's most critical for applications.

The Prometheus monitoring system [4] has added a simple histogram data-type that allows a rough summarization of the distribution with a set of numeric time series.

Most recently Data Dog has published a logarithmic histogram data structure DDSketch in [1].

In the next section we will develop some theory around general histogram summaries, that allows us to precisely define HDR Histograms, DDSketches and the circllhist in section 5.

### 3 Theory

In this section we develop an abstract theory of histograms to a degree that allows us to formally define circllhist as a linear refinement of a logarithmic histogram structure.

The basic idea behind log-linear histograms like the circllhist is illustrated in Figure 2. We start with a logarithmic binning of the real axes, that has bins at the powers of ten.

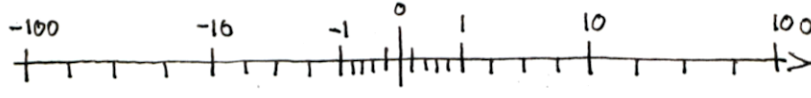
$$\dots 0.01, 0.1, 1, 10, 100, \dots$$

We divide each logarithmic bin into  $n = 90$  equally spaced segments. In this way the bin boundaries are precisely the base-10, precision-2 floating point numbers:

$$\begin{array}{ccccccc} \dots & 1.0, & 1.1, & 1.2, & \dots & 9.9, & \\ & 10, & 11, & 12, & \dots & 99, & \\ & 100, & 110, & 120, & \dots & 990, & \dots \end{array}$$

Those are the bin boundaries for the circllhist data structure. When samples are inserted into the circllhist, we retain counts of the number of samples in each bin. This information allows us to approximate the original location of the inserted samples with a maximal relative error less than 5%.

$b=10/n=4$  Log-Linear Binning



$b=10$  Logarithmic Binning



$n=4$  Linear Binning of  $[0, 1]$

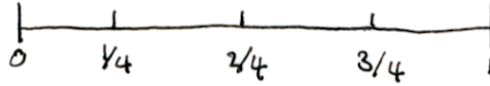


Figure 2: Construction of the Log-Linear Binning

### 3.1 Binnings

**Definition 3.1.** Let  $D \subset \mathbb{R}$  be a connected subset of the real axes (e.g.  $D = \mathbb{R}, D = [0, 1)$ ). A binning of  $D$  is a collection of intervals  $\text{Bin}[i], i \in I$ , that are disjoint and collectively cover the binning domain  $D$ :

$$D = \bigcup_{i \in I} \text{Bin}[i] \quad \text{and} \quad \text{Bin}[i] \cap \text{Bin}[j] = \emptyset \quad \text{for} \quad i \neq j.$$

The map that associates to each  $x \in D$  the unique index  $i$  so that  $x \in \text{Bin}[i]$  is called binning map and is denoted as  $\text{bin}(x) = i$ .

*Remark 3.2.* The binning map  $\text{bin} : D \rightarrow I$  determines the binning via  $\text{Bin}[i] = \{x \in D \mid \text{bin}(x) = i\}$ .

**Example 3.3.** The linear binning of  $\mathbb{R}$  is given by  $I = \mathbb{Z}$ , with

$$\text{Bin}[i] = [i, i + 1) \quad \text{and} \quad \text{bin}(x) = \lfloor x \rfloor$$

**Example 3.4.** The length  $n$  linear binning of  $[0, 1)$  is given by  $I = \{0, \dots, n - 1\}$ , with

$$\text{Bin}_n^{\text{Lin}}[i] = \left[ \frac{i}{n}, \frac{i+1}{n} \right) \quad \text{and} \quad \text{bin}_n^{\text{Lin}}(x) = \lfloor x \cdot n \rfloor$$

**Example 3.5.** The logarithmic binning with basis  $b > 0$  of  $\mathbb{R}_{>0}$  is given by  $I = \mathbb{Z}$ , with

$$\text{Bin}_b^{\text{Log}}[i] = [b^i, b^{i+1}) \quad \text{and} \quad \text{bin}_b^{\text{Log}}(x) = \lfloor \log_b(x) \rfloor$$

**Definition 3.6.** Given a map  $\alpha : I \rightarrow J$ , and a binning  $(I, \text{Bin})$ , we can define a new binning  $(J, \text{Bin}^*)$  by setting:

$$\text{Bin}^*[j] = \bigcup_{i, \alpha(i)=j} \text{Bin}[i], \quad \text{and} \quad \text{bin}^*(x) = \alpha(\text{bin}(x))$$

In this situation we call  $(J, \text{Bin}^*)$  a coarsening of  $(I, \text{Bin})$ , and  $(I, \text{Bin})$  a refinement of  $(J, \text{Bin}^*)$ .

**Definition 3.7.** Given a binning  $\text{Bin}[i], i \in I$  with half-open bins  $\text{Bin}[i] = [a_i, b_i)$ . The length- $n$  linear refinement of  $(I, \text{Bin})$ , is given by the index set  $I \times \{0, \dots, n - 1\}$ , bins

$$L_n \text{Bin}[i, j] = \left[ a_i + \frac{j}{n}(b_i - a_i), a_i + \frac{j+1}{n}(b_i - a_i) \right) \quad (1)$$

**Lemma 3.8.** The binning map of the length- $n$  linear refinement is given by:

$$L_n \text{bin}(x) = (\text{bin}(x), \lfloor \frac{x - a_{\text{bin}(x)}}{b_{\text{bin}(x)} - a_{\text{bin}(x)}} \cdot n \rfloor). \quad (2)$$

*Proof.* We have to show that  $x \in L_n \text{Bin}[L_n \text{bin}(x)]$  for all  $x \in D$ . Let  $(i, j) = L_n \text{bin}(x)$ . Since  $i = \text{bin}(x)$ , we know that  $x \in [a_i, b_i)$ . Now we consider the linear map  $\phi(x) = (x - a_i)/(b_i - a_i)$  which maps  $\text{Bin}[i]$  bijectively to  $[0, 1)$ . We have  $j = \lfloor n\phi(x) \rfloor$  by definition of  $L_n \text{bin}(x)$ . To show that  $x \in L_n \text{Bin}[L_n \text{bin}(x)]$  it suffices to verify that  $\phi(x) \in \phi(L_n \text{Bin}[i, j]) = [\frac{j}{n}, \frac{j+1}{n})$ . And indeed,

$$\frac{j}{n} = \frac{\lfloor n\phi(x) \rfloor}{n} \leq \phi(x) < \frac{\lfloor n\phi(x) \rfloor + 1}{n} = \frac{j+1}{n}.$$

□

**Lemma 3.9.** *The length- $n$  linear refinement of a binning  $(I, \text{Bin})$  is a refinement in the sense of Definition 3.6. The index map is given by  $\alpha(i, j) = i$  for  $i \in I, j \in \{0, \dots, n-1\}$ .*

*Proof.* We have to show that  $\bigcup_j L_n \text{Bin}[i, j] = \text{Bin}[i]$ , for all  $i \in I$ . Again we consider the linear bijection  $\phi(x) = (x - a_i)/(b_i - a_i)$ , which maps  $\text{Bin}[i]$  to  $[0, 1)$  and  $L_n \text{Bin}[i, j]$  to  $[j/n, (j+1)/n)$ . Hence it suffices to show that  $[j/n, (j+1)/n)$  cover  $[0, 1)$  for  $j = 0, \dots, n$  which is evident. □

Now we are in a position to define log-linear binnings.

**Definition 3.10.** *The base  $b$ , length  $n$  log-linear binning of  $\mathbb{R}_{>0}$  is the length- $n$  linear refinement of the base- $b$  Logarithmic binning of  $\mathbb{R}_{>0}$ .*

**Proposition 3.11.** *Let  $b, p$  be positive integers. The boundaries of the base  $b$ , length  $n = b^p - b^{p-1}$  log-linear binning of  $\mathbb{R}_{>0}$  are precisely the base- $p$  precision- $p$  floating point numbers:*

$$\text{float}_{b,p}(e, d) = \frac{d}{b^{p-1}} \cdot b^e = d \cdot b^{e-p+1} \quad \text{with } e \in \mathbb{Z}, d \in \{b^{p-1}, \dots, b^p - 1\}$$

The binning map is given by

$$\text{bin}(x) = (e(x), d(x) - b^{p-1}), \quad \text{with } e(x) = \lfloor \log_b(x) \rfloor, d(x) = \lfloor x \cdot b^{-e(x)+p-1} \rfloor$$

with values  $e(x) \in \mathbb{Z}$  and  $d(x) \in \{b^{p-1}, \dots, b^p - 1\}$ . This binning is also called base- $b$  precision- $p$  log-linear binning.

**Example 3.12.** *According to Proposition 3.11, the base-10 precision-1 binning has bin boundaries at*

$$\{d \cdot 10^e \mid e \in \mathbb{Z}, d \in \{1, \dots, 9\}\} = \{\dots 0.8, 0.9, 1, 2, \dots, 8, 9, 10, 20, \dots\}$$

binning map

$$\text{bin}(x) = (e(x), d(x) - 1), \quad \text{with } e(x) = \lfloor \log_{10}(x) \rfloor, d(x) = \lfloor x/10^{e(x)} \rfloor.$$

*Proof.* To proof Proposition 3.11, we compute the log-linear bin boundaries using equation 1 with  $a_i = b^i, b_i = b^{i+1}$  and  $n = b^p - b^{p-1}$ :

$$\begin{aligned} \text{Bin}[e, j] &= [b^e + \frac{j}{b^p - b^{p-1}}(b^{e+1} - b^e), b^e + \frac{j+1}{b^p - b^{p-1}}(b^{e+1} - b^e)] \\ &= [b^{e-p+1}(b^{p-1} + j), b^{e-p+1}(b^{p-1} + j + 1)] \\ &= [db^{e-p+1}, (d+1)b^{e-p+1}] \end{aligned}$$

where we set  $d = b^{p-1} + j$ . If  $j$  runs through  $1 \dots n$ , then  $d$  runs through  $b^{p-1}, \dots, b^p - 1$ . This shows that the lower boundaries are exactly the base- $b$  precision- $p$  floating point numbers. For the upper boundary note, that the if  $d = b^p - 1$  then  $(d+1)b^{e-p+1} = b^{p-1}b^{(e+1)-p+1}$  is again a base- $b$  precision- $p$  floating point number (with a larger exponent).

The binning map can be explicitly calculated using Equation 2 as:

$$\begin{aligned} \text{bin}(x) &= (e(x), k(x)), \quad \text{with } e(x) = \lfloor \log_b(x) \rfloor \\ k(x) &= \lfloor \frac{x - b^e}{b^{e+1} - b^e} (b^p - b^{p-1}) \rfloor = \lfloor x \cdot b^{-e+p-1} \rfloor - b^{p-1} = d(x) - b^{p-1} \end{aligned}$$

as claimed. □

**Definition 3.13.** The circllhist binning is the base-10 precision-2 log-linear binning extended to the real axes, with bins:

$$\begin{aligned} \text{Bin}[+1, e, d] &= [d \cdot 10^{e-1}, (d+1)10^{e-1}), \quad e \in \mathbb{Z}, d \in \{10, \dots, 99\} \\ \text{Bin}[0, 0, 0] &= \{0\} \\ \text{Bin}[-1, e, d] &= [-d \cdot 10^{e-1}, -(d+1) \cdot 10^{e-1}). \end{aligned}$$

The binning map is given by:

$$\begin{aligned} \text{bin}(x) &= (+1, e, d), e = \lfloor \log_{10}(x) \rfloor, d = \lfloor x \cdot 10^{-e-1} \rfloor \\ \text{bin}(0) &= (0, 0, 0) \\ \text{bin}(-x) &= (-1, e, d) \end{aligned}$$

for  $x > 0$ .

**Proposition 3.14.** The binning map of the circllhist can be recursively computed with the following algorithm:

```
function bin(x)
  if x == 0:
    return (0, 0, 0)
  if x < 0:
    (s, e, d) := bin(-x)
    return (-1, e, d)
  if x < 10:
    return bin(x * 10) - (0, 1, 0)
  if x > 100:
    return bin(x / 10) + (0, 1, 0)
  else: # 10 <= x < 100:
    return (+1, 1, floor(x))
end
```

In particular, the circllhist binning map can be computed without use of the logarithm function.

*Proof.* The recursion terminates since every positive number  $x$  can be brought into range  $10 \leq x < 100$  with a finite number of divisions or multiplications by 10. It's straight forward to verify that each case computes valid results assuming that the results of the recursive call are correct.  $\square$

*Remark 3.15.* If  $x, e$  are integers, then the binning map of the number  $x \cdot 10^e$  can be computed without the use of floating point arithmetic as  $\text{bin}(x) + (0, e, 0)$  using Algorithm 3.14.

This is of practical relevance when used in an environment which do not have floating point arithmetic available. One example being nano-second latencies measured in the Linux kernel, or embedded devices.

## 3.2 Pareto Midpoints

**Proposition 3.16.** Given an interval  $[a, b]$  in  $\mathbb{R}_{>0}$  the unique point  $m$  in  $[a, b]$  so that the maximal relative distance  $rd(m, y) = |m - y|/y$  to all other points in  $[a, b]$  is minimized by the pareto midpoint  $m = 2ab/(a + b)$ .

The maximal relative distance to the pareto midpoint is assumed at the interval boundaries  $rd(m, a) = rd(m, b) = (b - a)/(a + b)$ .

*Proof.* We have to minimize the function  $\text{maxrd}(x) = \max_{y \in [a, b]} rd(x, y)$  over  $[a, b]$ . The maximum  $\max_{y \in [a, b]} rd(x, y)$  is attained either for  $y = a$  or  $y = b$ , hence  $\text{maxrd}(x) = \max\{rd(x, a), rd(x, b)\}$ .

Note that the function  $f(x) = rd(x, a)$  is continuous and strictly monotonically increasing on  $[a, b]$ , with  $f(a) = 0, f(b) > 0$ , and the function  $g(x) = rd(x, b)$  is continuous and strictly monotonically decreasing on  $[a, b]$ , with  $g(a) > 0$  and  $g(b) = 0$ . The point  $m$  is the unique point in  $[a, b]$  where both functions are equal, with

$$rd(m, a) = \frac{b - a}{a + b} = rd(m, b)$$

Now if  $a \leq x < m$  then  $rd(x, b) = g(x) > g(m)$  and so  $\text{maxrd}(x) > \text{maxrd}(m) = g(m)$ . Similarly if  $m < x \leq b$  then  $rd(x, a) = f(x) > f(m)$  and so  $\text{maxrd}(x) > \text{maxrd}(m) = f(m)$ .

This shows that  $x = m$  is the unique minimum of  $\text{maxrd}(x)$  on  $[a, b]$ .  $\square$

The following proposition gives a probabilistic interpretation of the location of relative distance minimizing midpoint. Recall, that the expected value of a uniformly distributed random variable  $X \sim U[a, b]$  is the midpoint  $\mathbb{E}[X] = (a+b)/2$ .

**Proposition 3.17.** *Given an interval  $[a, b]$ , and an  $a=2$  pareto distributed random variable  $X$ , then the pareto midpoint is the conditional expectation:*

$$\mathbb{E}[X | X \in [a, b]] = 2ab/(a + b).$$

*Proof.* The pareto distribution has density  $p(x) = C \cdot 1/x^{a+1}$ , for some positive constant  $C$ , so for  $a = 2$  we get

$$\mathbb{E}[X | X \in [a, b]] = \frac{\int_a^b xp(x)dx}{\int_a^b p(x)dx} = \frac{\int_a^b x^{-2}dx}{\int_a^b x^{-3}dx} = 2 \frac{b^{-1} - a^{-1}}{b^{-2} - a^{-2}} = \frac{2}{b^{-1} + a^{-1}} = 2 \frac{ab}{a + b}$$

which proves the claim.  $\square$

**Proposition 3.18.** *The maximal relative distance to a pareto midpoint in the circllhist binning is  $1/21 \approx 4.76\%$ .*

*Proof.* Substituting the bin boundaries into the formula given in Proposition 3.16 we find  $(b - a)(a + b) = 1/(2d + 1)$  for the bin  $\text{Bin}[e, d], e \in \mathbb{Z}, d \in \{10, \dots, 99\}$ . This is minimized for  $d = 10$  with a value of  $1/21$  as claimed.  $\square$

### 3.3 Histograms

Once we have established the binnings, histograms are easy to define.

**Definition 3.19.** *A histogram with domain  $D \subset \mathbb{R}$  is a binning  $\text{Bin}[i], I$ , together with a count function  $H : I \rightarrow \mathbb{N}_0$ .*

Given a dataset and a binning, we can associate a histogram.

**Definition 3.20.** *Given binning  $\text{Bin}[i], I$  of  $D \subset \mathbb{R}$ , and a dataset  $X = (x_1, \dots, x_n)$  with values in  $D$ , we define the histogram summary of  $X$  as the histogram with binning  $\text{Bin}[i], I$  and count function*

$$H_X(i) = \#\{j | x_j \in \text{Bin}[i]\}.$$

*This means, that  $H_X(i)$  counts the number of points of  $X$  lying in  $\text{Bin}[i]$ .*

Histograms can be freely merged without losing information.

**Definition 3.21.** *Let  $H_1, H_2$  be histograms for the same binning. The merged histogram  $H_1 + H_2$  has count function:*

$$(H_1 + H_2)(i) = H_1(i) + H_2(i).$$

The merge operation is clearly associative and commutative.

We can easily see that the histogram merge computes is compatible with merge (concatenation) of datasets:

**Proposition 3.22.** *Given binning  $\text{Bin}[i], I$  of  $D \subset \mathbb{R}$ , and two datasets  $X = (x_1, \dots, x_n), Y = (y_1, \dots, y_m)$  with values in  $D$ . Let  $Z = (x_1, \dots, x_n, y_1, \dots, y_m)$  be the merged dataset, then  $H_Z = H_X + H_Y$ .*

*Proof.* We have

$$\begin{aligned} H_Z(i) &= \#\{j | z_j \in \text{Bin}[i]\} = \#\{j \leq n | x_j = z_j \in \text{Bin}[i]\} + \#\{j > n | z_j = y_{j-n} \in \text{Bin}[i]\} \\ &= H_X(i) + H_Y(i). \end{aligned} \quad \square$$

*Remark 3.23.* Let  $H$  be a histogram summary of a dataset  $X$ , and a threshold value  $y$ . The count functions  $\text{count-below}_X(y) = \#\{i | x_i < y\}$  and  $\text{count-above}_X(y) = \#\{i | x_i \geq y\}$  are of great practical interest. If the bin boundaries line-up with the threshold, we can get exact approximations of those functions from the histogram.

In the case of the circllhist, we get exact counts for every 2-digit precision decimal floating point number.

For logarithmic histograms we get exact counts at the powers of the base  $b^e, e \in \mathbb{Z}$ .

### 3.4 Histogram Operations

Given a histogram we can try to approximate statistics (means, percentiles, etc.) of the original dataset. There are three strategies for doing so, that we have found valuable in practice:

1. Derive a probability distribution from a histogram (with uniform distribution inside the bins), and apply the statistics to this distribution.
2. Resample the dataset by placing all points inside a bin at equally spaced distances (*fair resampling*).
3. Resample the dataset by placing all points inside a bin into the (pareto) midpoints (*midpoint resampling*).

From a theoretical perspective the probabilistic strategy is the most natural, and gives generally good results for data sampled from continuous distribution (with density). Midpoint resampling is the simplest and gives a low a-priori error bounds independent of the distribution of  $X$ . Fair resampling mitigates between both approaches, and gives lower expected errors on a wide variety of datasets and while avoiding large relative errors in case of single sample bins.

The current implementation of circllhist uses midpoint resampling to define sum, mean, stdev and moment estimation. For percentile calculations fair resampling is used.

**Proposition 3.24.** *Let  $X = (x_1, \dots, x_n)$ ,  $n > 0$  be a dataset with values in  $\mathbb{R}_{>0}$ , let  $H$  be the circllhist summary of  $X$ . For each bin  $\text{Bin}[i]$ , let  $c_i \in \text{Bin}[i]$  be the pareto midpoint. Let*

$$\text{count}[H] = \sum_i H(i), \quad \text{sum}[H] = \sum_{i \in I} H(i) \cdot c_i, \quad \text{and} \quad \text{mean}[H] = \text{sum}[H] / \text{count}[H].$$

then

1.  $\text{count}[H] = \text{count}(X) = n$ .
2. The relative error  $|\text{sum}[H] - \text{sum}(X)| / \text{sum}(X)$  is smaller than  $1/21 \approx 4.76\%$ .
3. The relative error  $|\text{mean}[H] - \text{mean}(X)| / \text{mean}(X)$  is smaller than  $1/21 \approx 4.76\%$ .

*Proof.* The first claim follows immediately from the definition of  $\text{count}[H]$ , and  $H(X)$ . For the second claim, we have

$$\text{sum}[H] - \text{sum}(X) = \sum_{i \in I} H(i) \cdot c_i - \sum_j x_j = \sum_{i \in I} (H(i) \cdot c_i - \sum_{j, x_j \in \text{Bin}[i]} x_j) = \sum_{i \in I} \sum_{j, x_j \in \text{Bin}[i]} (c_i - x_j)$$

And hence,

$$|\text{sum}[H] - \text{sum}(X)| \leq \sum_{i, j, x_j \in \text{Bin}[i]} |c_i - x_j| \leq \sum_j \frac{1}{21} |x_j| = \frac{1}{21} \cdot \text{sum}(X),$$

where we used 3.18 in second step. This shows the second claim. For the third claim follows from the second by extending the fraction with  $1/n$ .  $\square$

### 3.5 Quantiles

Quantiles can be approximated from histograms using any of the strategies outlined at the beginning of the last section. The probabilistic strategy has the downside, that for  $q$  close to 1, the  $q$ -quantile will always be near the high end of the largest bin with at least one sample in it. Often times this bin contains only a single sample. In this case the worst case error of the full bin size can be easily assumed.

Midpoint resampling has the downside, that expected errors for percentiles in the center of the distribution are much larger than needed. For densely populated bins the uniform distribution is often a good approximation, and can be used to estimate percentiles with high precision in typical cases.

Fair resampling offers a welcome middle route. Instead of placing all samples at the (pareto) midpoint, or smoothing them out with a uniform distribution, we place the samples at equally spaced position within each bin. In this way, densely populated bins will be approximated by a near uniform distribution and bins with only a single sample will be replaced by a single sample at the midpoint, reducing the worst-case error to half the bin size.

**Definition 3.25.** Given a histogram  $H$  on a binning  $\text{Bin}[i], i \in I$  we define the fair resampling  $X$  of  $H$  as follows. For each bin  $\text{Bin}[i]$ , with boundaries  $a, b$  and count  $H(i) = n$ , we consider the points

$$x_{i,k} = a + \frac{k}{n+1}(b-a) \quad \text{for } k = 1, \dots, n$$

and let  $\hat{X} = (x_{i,k} \mid i \in I, k = 1 \dots, H(i))$ . This is a dataset with  $\text{count}[H]$  points.

Similarly, we define the midpoint resampling of  $H$ , with  $H(i)$  samples at the midpoint of  $\text{Bin}[i]$ , and the pareto midpoint resampling of  $H$ , with  $H(i)$  samples at the pareto midpoint of  $\text{Bin}[i]$ .

We now want to define the quantiles of a histogram as quantiles of the fair resampling of  $H$ . Before we can do so we first need to clarify the quantile definition for datasets. While there is only a single established quantile definition for probability distributions, there are multiple different quantile definitions for datasets found in the wild. A comprehensive list was compiled by Hyndman-Fan in 1996 [2]. For our discussion we will need type-1/7 quantiles from the Hyndman-Fan list, as well as two other quantile functions, used by data-structures covered in our evaluation.

**Definition 3.26.** Given a dataset  $X = (x_1, \dots, x_n), n > 0$  of real numbers, and a number  $q \in [0, 1]$ . Let  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  be the ordered version of  $X$ . We define the minimal type-1  $q$ -quantile as

$$Q_0^1(X) = x_{(1)} \quad \text{and} \quad Q_q^1(X) = x_{(\lceil q \cdot n \rceil)} \quad \text{for } q > 0.$$

We define the minimal type-7 quantile as

$$Q_q^7(X) = x_{(\lfloor q \cdot (n-1) \rfloor + 1)}.$$

The interpolated type-7 quantile is given by:

$$Q_q^{7i}(X) = (1 - \gamma) \cdot x_{(\lfloor q \cdot (n-1) \rfloor + 1)} + \gamma \cdot x_{(\lceil q \cdot (n-1) \rceil + 1)},$$

where the interpolation factor  $\gamma \in [0, 1]$  is given by  $\gamma = q(n-1) - \lfloor q(n-1) \rfloor$ .

The type-hdr quantile  $Q_q^{\text{hdr}}(X)$  is given by  $x_{(1)}$  if  $qn \leq \frac{1}{2}$ , by  $x_{(n)}$  if  $qn \geq n - \frac{1}{2}$ , and otherwise, by

$$Q_q^{\text{hdr}}(X) = x_{(\lfloor qn - \frac{1}{2} \rfloor)}.$$

Similarly, the type-tdigest quantile is given by  $x_{(1)}$  if  $qn \leq \frac{1}{2}$ , by  $x_{(n)}$  if  $qn \geq n - \frac{1}{2}$ , and otherwise, by

$$Q_q^{\text{tdigest}}(X) = (1 - \gamma) \cdot x_{(\lfloor qn - \frac{1}{2} \rfloor)} + \gamma \cdot x_{(\lceil qn - \frac{1}{2} \rceil)}$$

where  $\gamma = qn - \frac{1}{2} - \lfloor qn - \frac{1}{2} \rfloor$ .

Figure 3a shows a plot of these quantile definitions as functions on  $q$ . From a theoretical perspective type-1 quantiles are the most natural, since they correspond the probabilistic quantiles for the empirical distribution function of the dataset  $X$ . Also they allow to formulate precise statements about lower counts (cf. Proposition 3.27 below). For this reason the circllhist implementation uses minimal type-1 quantiles on the fair resampling of the histogram as for quantile calculations.

Somewhat surprisingly, type-7 quantiles are most commonly found in software packages. E.g. numpy [6] computes interpolated type-7 quantiles by default. Also DDSketch [1] and t-digest [3] implement variants of type-7 quantiles. A detailed comparison between type-1 and type-7 quantiles can be found at [9]. The t-digest and the HDR Histograms use custom quantile functions that were not covered in Hyndman-Fan. We call them ‘‘type-hdr’’ and ‘‘type-tdigest’’ here. Figure 3b shows a plot of the quantile functions, produced by all relevant implementations.

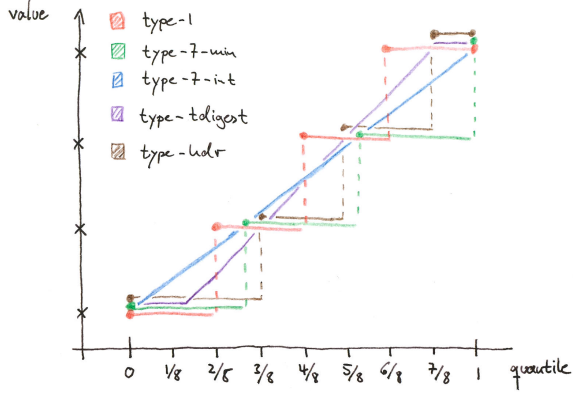
**Proposition 3.27.** Let  $X$  be a dataset of length  $n > 1$ ,  $y \in \mathbb{R}$  a threshold value and  $0 < q \leq 1$ . Then the following statements are equivalent:

- (1) There are at least  $qn$  datapoints  $x$  in  $X$  with  $x \leq y$ .
- (2) We have  $Q_q^1(X) \leq y$ .

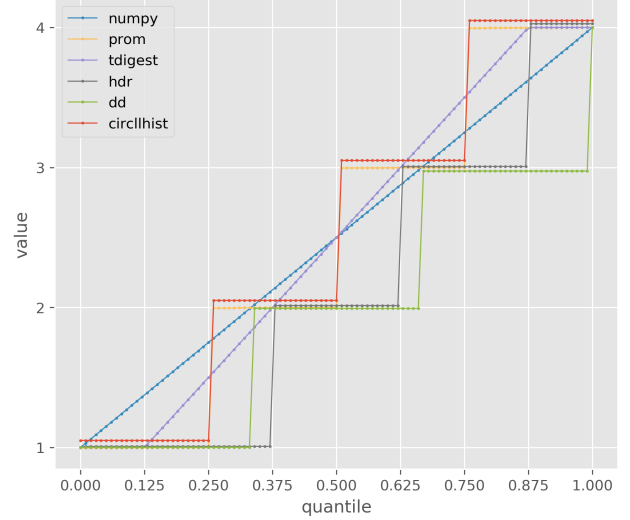
*Proof.* To show (1) implies (2), assume that there are at least  $qn$  datapoints below  $y$ . Since the number of datapoints is always an integer, the same holds for  $r = \lceil qn \rceil$ . This implies that the ordered values  $x_{(1)}, \dots, x_{(r)}$  all lie below  $y$ , and hence  $x_{(r)} = Q_q^1(X) \leq y$ .

To show (2) implies (1), assume that  $Q_q^1(X) \leq y$  then the  $r = \lceil qn \rceil$  values  $x_{(1)}, \dots, x_{(r)} = Q_q^1(X)$  all lie below  $y$ . This shows there are at least  $nq$  datapoints lie below  $y$ .  $\square$





(a) Theoretical Quantile Functions



(b) Computed Quantile Functions

Figure 3: Quantile Function Comparison for the Dataset [1, 2, 3, 4]

**Proposition 3.28.** Let  $X$  a dataset with  $n > 0$  positive values and  $0 < q \leq 1$ . Let  $H$  be the circllhist summary of  $X$ . We denote by  $\hat{X}^f(H)$  the fair resampling, and by  $\hat{X}^p(H)$  the pareto-midpoint resampling of  $H$ .

1. The relative error of the estimated type-1 quantile is less than  $1/21 \approx 4.76\%$  for the pareto-resampling:

$$|Q_q^1(X) - Q_q^1(\hat{X}^p(H))| \leq \frac{1}{21} Q_q^1(X)$$

2. The relative error of the estimated type-1 quantile is less than  $1/10 = 10\%$  for the fair-resampling:

$$|Q_q^1(X) - Q_q^1(\hat{X}^f(H))| \leq \frac{1}{10} Q_q^1(X)$$

Furthermore, in case the quantile value falls into a bin with a single sample (typical for outliers), the maximal relative error is 5%.

*Proof.* Let  $r = \lceil qn \rceil$ , so that  $Q_q^1(X) = X_{(r)}$ . By construction of the pareto resampling, the  $r$ -th ordered value in the  $Q_q^1(\hat{X}^p) = \hat{X}_{(r)}^p$ , is the pareto midpoint of the bin containing  $X_{(r)}$ . This shows that  $|Q_q^1(X) - Q_q^1(\hat{X}^p)| < 1/21 Q_q^1(X)$  by Proposition 3.18.

Similarly for the fair resampling, the  $r$ -th ordered value in the  $Q_q^1(\hat{X}^f) = \hat{X}_{(r)}^f$ , lies in the bin containing  $X_{(r)}$ . Now we claim, that the maximal relative distance to any point in the same circllhist bin is 10%. Indeed, the worst case relative difference is assumed for a bin  $[10 \cdot 10^e, 11 \cdot 10^e)$  with  $x = 10 \cdot 10^e$  and  $y \rightarrow 11 \cdot 10^e$  so that  $|x - y|/x \rightarrow 1/10 = 10\%$ .

In case that there is only a single sample in the bin  $[a, b)$  containing  $X_{(r)}$ , the value  $\hat{X}_{(r)}^f$  will be at the (arithmetic) midpoint of the bin  $m = a + \frac{1}{2}(b - a)$ . Hence, in this case  $|X_{(r)} - \hat{X}_{(r)}^f|/X_{(r)} \leq \frac{1}{2}(b - a)/X_{(r)}$ . This number is maximized for  $X_{(r)} = a$ , and  $a = 10 \cdot 10^e$  in the circllhist binning. In which case  $\frac{1}{2}(b - a)/a = \frac{1}{2} \frac{1}{10} = 5\%$ .  $\square$

**Example 3.29.** With the notation from the last proposition. Let  $X = (10, 10, 10, \dots, 10)$  of length  $n$ , then the fair resampling is given by  $\hat{X}^f = (10 + \frac{1}{n+1}, \dots, 10 + \frac{n}{n+1})$ . So  $Q_1^1(\hat{X}^f) = 10 + \frac{n}{n+1}$  which converges to 11 for  $n \rightarrow \infty$ . On the other hand  $Q_1^1(X) = \max(X) = 10$ . So the worst case error of 10% is assumed in the asymptotic case.

This proposition shows, that the worst-case relative error for the circllhist is 10%. The theoretical worst-case is only realized in cases were a large number of samples falls at the lower end of a bin as shown in the last example. This is very rare to happen. In practice we usually see a 5% worst-case relative error at the tails of the distribution, and high accuracy quantiles ( $< 1\%$  relative error) in the body of the distribution.

## 4 The Circllhist Implementation

The data-structure implemented in [13], is a circllhist in the sense of Definition 3.13 with exponent range limited to  $-128 \leq e < 127$ . In this way, both exponent  $e$  and mantissa  $d$  can be represented by 8-bit integers. The counts  $H(i)$  are represented as 64-bit unsigned integers.

The largest representable number of the circllhist is  $99 \cdot 10^{127}$ . This number is larger than the age of the universe measured in nano-seconds (13.8 billion years) The smallest representable positive number is  $10 \cdot 10^{-128}$ . This number is smaller than the Plank time measured in years ( $5.39 \cdot 10^{-44}$  s). We have found this value range to be sufficient for all practical purposes.

Within that range the circllhist bins have a relative size of 1%-10% of the values contained in them. Hence, the original values can be reconstructed with a relative error of no more than 10%. If the reconstructed values are placed at the paretro midpoint of the bin, the relative reconstruction error can be reduced to 4.76% (see Proposition 3.18).

The accuracy of approximations of statistical quantities like sums, means and quantiles is commonly much better than 10%, since the individual reconstruction errors cancel across the dataset. In the case of quantiles, we will see this in the evaluation below.

The counts of samples above or below a threshold, can be reconstructed accurately for threshold that fall onto bin boundaries. In the case of the circllhist, those lie at decimal floating point numbers with two digits of precision (e.g. 0.23, 1.5, 110). In practice, those are the numbers that humans choose, if they have to come up with thresholds.

A circllhist is internally represented as a sparse list of (bin, count) pairs. If a bin has a count of 0, then it can be skipped in the representation. The serialized form of the histogram will only contain bins with non-zero counts. The theoretical maximum of used bins is  $2 \times 256 \times 90 + 1 = 46081$  (for sign, exponent, mantissa and zero bucket) which makes for a maximal size of 461kb.

In practice we have never seen histogram data structures getting close to this size. Even in extreme cases, when capturing billions of samples from a nano-second to minute scale, the number of allocated bins never exceeded 1000, and the total size stayed below 10kb. Typical histograms in our system, have anywhere from 0-200 allocated bins and occupy <2kb before compression.

A notable design goal of the circllhist is it's use for measurements inside the kernel or low-powered embedded devices. In those environments floating point arithmetic is not available, and insertion performance is particularly critical. For these purposes the circllhist provides a highly optimized insertion function that avoids floating point arithmetic entirely (cf. Proposition 3.14).

The C implementation of libcircllhist includes a number of performance optimizations. It comes with an optional index structure, that avoids iteration over bins when retrieving and inserting data and uses static branch annotations to aid CPU branch predictions. With these optimizations we can get raw insertion latencies down to  $\sim 10ns$  for integer, and  $\sim 80ns$  for double values.<sup>3</sup>

Implementations of the circllhist are available for a variety of languages including:

- C/C++ – <https://github.com/circonus-labs/libcircllhist>
- Python – <https://github.com/circonus-labs/libcircllhist/tree/master/src/python>
- Go – <https://github.com/circonus-labs/circonusllhist>
- Lua – <https://github.com/circonus-labs/libcircllhist/tree/master/src/lua>
- JavaScript – <https://github.com/circonus-labs/circllhist.js>
- C# .NET – <https://github.com/circonus-labs/netcircllhist>.

## 5 Evaluation

In this section we present a numerical comparison between the quantile aggregation methods described in Section 2. We will evaluate the precision and performance of quantile calculations, performance of insertion and merge operations, as well as the size of the resulting data-structures.

The evaluation proceeds in three phases: insertion, merge and quantile calculation. In the insertion phase, raw data is inserted the data-structures. Each data-set is split into a number of batches. We create individual data-structures for

<sup>3</sup> These latencies were measured in a tight C loop with the provided `test/histogram_perf.c` script, on a 2Ghz Intel Xeon CPU. The evaluation in section 5 is Python based and uses different iteration counts and data.

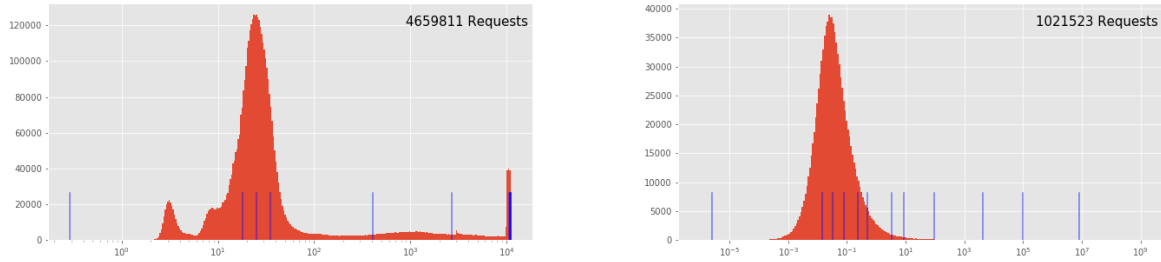


Figure 4: Total Distribution of the 'API Latency' and 'Simulated Latencies' Datasets, with quantile markers.

each batch. In the merge phase, data-structures created for each batch are merged into a single one. In the quantile phase, we perform quantile calculations on the merged data-structure.

The evaluation was performed using a set of Jupyter notebooks, which are available at

<https://github.com/circonus-labs/circllhist-paper>.

This repository contains datasets and source code used to generate the exact tables and graphics we are using this document. We also provide docker images and instructions, that should aid the reproduction of the evaluation results on other machines. We are open to extension and improvements to the evaluation setup. Please contact us via email or open a pull request, if you identified flaws or found improvements.

## 5.1 Methods

The following data-aggregation methods are considered for evaluation.

- exact Exact quantile computation based on numpy arrays [6].
- prom Quantile estimation based on Prometheus Histograms [4].
- hdr The HDR Histogram data-structure introduced in [5].
- dd The DDSketch data-structure introduced in [1].
- t-digest The t-digest data-structure introduced in [3].
- circllhist The circllhist data-structure described in this document.

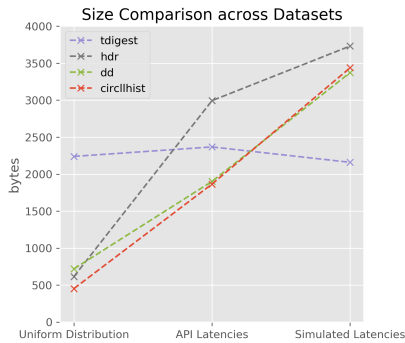
The Prometheus monitoring system provides a histogram data-type that consists of a list of “less than” metrics, which count how many samples were inserted that are below manually configured threshold values. This datum of a Prometheus histogram is equivalent to a histogram in the sense of Definition 3.19, with bin-boundaries at the configured thresholds. Prometheus Histograms are included here for their wide use in practice. The method itself is not really comparable, since its results are highly dependent on the number and location of the chosen thresholds. For the evaluation, we follow the recommended practice of using a total ten bins at locations which cover the whole data range with emphasis on the likely quantile locations. We use a hand-written Python translation of the original quantile functions written in go.

The HDR Histogram data structure is a log-linear Histogram in the sense of Definition 3.10. It uses a base of 2, and a configurable range and precision. It’s notable that the exposed API let’s the user specify decimal precision. The internal precision is then chosen in such a way that the resulting base-2 bins are smaller than the specified base-10 accuracy. Using base-2 arithmetic has the advantage of allowing bit-wise manipulation of floating points numbers to determine the bin location. For our evaluation used the Python implementation<sup>4</sup>, and configured the HDR Histograms to cover the same range of the circllhist ( $10^{-128} \dots 10^{+127}$ ) with two digits of decimal precision.

The DDSketch is a histogram for the logarithmic binning introduced in Example 3.5. It allows arbitrary positive real numbers as basis, to configure the desired precision. We use the Python implementation<sup>5</sup>, with the default precision of 1% (corresponding to  $b = 101/99$ ).

<sup>4</sup>[https://github.com/HdrHistogram/HdrHistogram\\_py](https://github.com/HdrHistogram/HdrHistogram_py)

<sup>5</sup>[github.com/DataDog/sketches-py](https://github.com/DataDog/sketches-py)



(a) Size Comparison

	Uniform Distribution	API Latencies	Simulated Latencies
exact	800000	37278488	8172184
prom	88	88	88
tdigest	2240	2368	2160
hdr	615	2994	3732
dd	720	1902	3373
circllhist	453	1866	3438

(b) Tabulated sizes in bytes

Figure 5: Size Comparison

The t-digest is the only evaluated method, that is not internally using a histogram data structure. Instead the inserted data is aggregated into clusters of adaptive size. The sizes are chosen in such a way that, high resolution data is available at the tails of the distribution. We use the original Java implementation<sup>6</sup> called from python using pyjnius<sup>7</sup>. The digests are configured with a compression parameter of 100, which is described to be a “common value for normal uses” in the source code.

For the circllhist we used the Python binding<sup>8</sup>. It does not allow any configuration of bin sizes or accuracy.

## 5.2 Datasets

For the evaluation we choose three different datasets: “Uniform Distribution”, “API Latencies” and “Simulated API Latencies” each containing over 1 million samples, split into more than 1000 batches.

For the “Uniform Distribution” dataset, random samples are generated for a uniform distribution on  $[10, 100]$ . These are split in to 1000 batches each containing 100 values, for a total of 100.000 samples.

The “API Latencies” dataset contains data collected at one of our internal APIs collected in 10 minute batches over the course of 6 weeks. The data is spread between 0.29 and 11.000 (ms). There is a cutoff value around 11sec, that was caused by timeout logic getting triggered. This results in the higher quantile values being close together. It consists of a total of 4.65 million samples in 6048 batches.

For the “Simulated Latencies” dataset, we generate a total of 1000 batches of randomized sizes with randomized data. The size of the batches follows a geometric distribution. The samples themselves are generated as a randomly displaced and scaled pareto distribution. The data is spread between on between  $10^{-5}$  and  $10^{10}$  with an extremely long tail. The dataset contains a total of close to 1.000.000 samples.

Figure 4 contains a histogram visualization of the total distribution of these datasets.

## 5.3 Size

Figure 5a and Table 5b show the sizes of the data structures after all samples from the respective datasets have been inserted and the data structures have been merged.

With our choice of parameters, the size of the HDR Histogram, DDSketch, circllhist follow each other quite closely. Note that the size increases with the spread of the data, not with the size of the inserted samples.

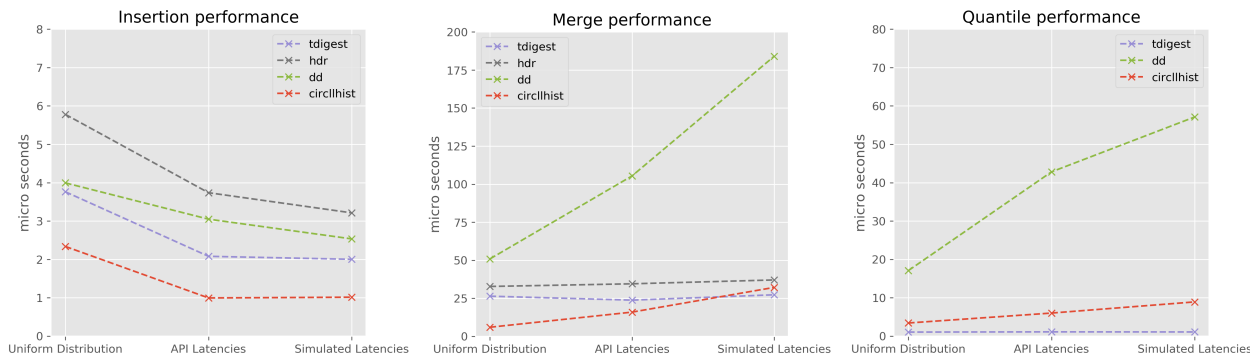
It’s important to note, that we compare the sizes of the serialized versions of the data-structures, as they would be consumed on disk or on the network, not the in-memory size (as this is harder to estimate). The in-memory size might be larger than the serialized size. This is particularly true for the HDR Histogram with pre-allocates all bins in memory, and skips empty bins for the serialization.

The Prometheus histogram stores a count for each of the ten configured threshold value, as well as the total count (infinite bin), and hence consumes exactly 88 bytes.

<sup>6</sup>[github.com/tdunning/t-digest](https://github.com/tdunning/t-digest)

<sup>7</sup><https://github.com/kivy/pyjnius>

<sup>8</sup>[github.com/circonus-labs/libcircllhist](https://github.com/circonus-labs/libcircllhist)



(a) Performance Comparison

Phase Dataset	Insertion			Merge			Quantile		
	Unif. D.	API L.	Sim. L.	Unif. D.	API L.	Sim. L.	Unif. D.	API L.	Sim. L.
exact	1.3	0.0	0.1	33.0	3351.9	472.6	825.1	25199.5	8789.9
prom	8.9	7.2	6.6	0.9	0.9	0.9	10.0	8.6	10.2
tdigest	3.8	2.1	2.0	26.4	23.8	27.3	1.1	1.1	1.1
hdr	5.8	3.7	3.2	32.8	34.5	37.1	1384.3	1981.2	1604.7
dd	4.0	3.0	2.5	50.9	105.4	184.0	17.1	42.8	57.1
circllhist	2.3	1.0	1.0	6.0	15.9	32.2	3.4	6.0	8.9

(b) Tabulated performance data in usec<sup>9</sup>

Figure 6: Performance Comparison

The size of the t-digest is relatively constant across all three data-sets, which reflects the fact that the number of clusters is kept constant when additional data is inserted.

## 5.4 Performance

Figure 6a and Table 6b show the measured performance for insertion, merge and analysis phase for the three considered datasets. As always, performance measurements should be taken with a grain of salt, since they are heavily influenced by the implementation, configuration and hardware choices. In our case, we choose to perform all performance measurements in python, with the most popular python implementations available. In the case of the t-digest we ran into accuracy (and performance) problems with the the python version so we used the official Java version via pyjnius.

The measurements itself were performed with the `timeit`<sup>10</sup> package, with 5 consecutive batches of runs each taking longer than 0.2 seconds. As recommended, we report the minimal duration that was attained in any of the runs.

We ran these experiments on a server with Intel(R) Xeon(R) D-1540 CPU and 128GB RAM running Linux 5.3.13 and python 3.7.3.

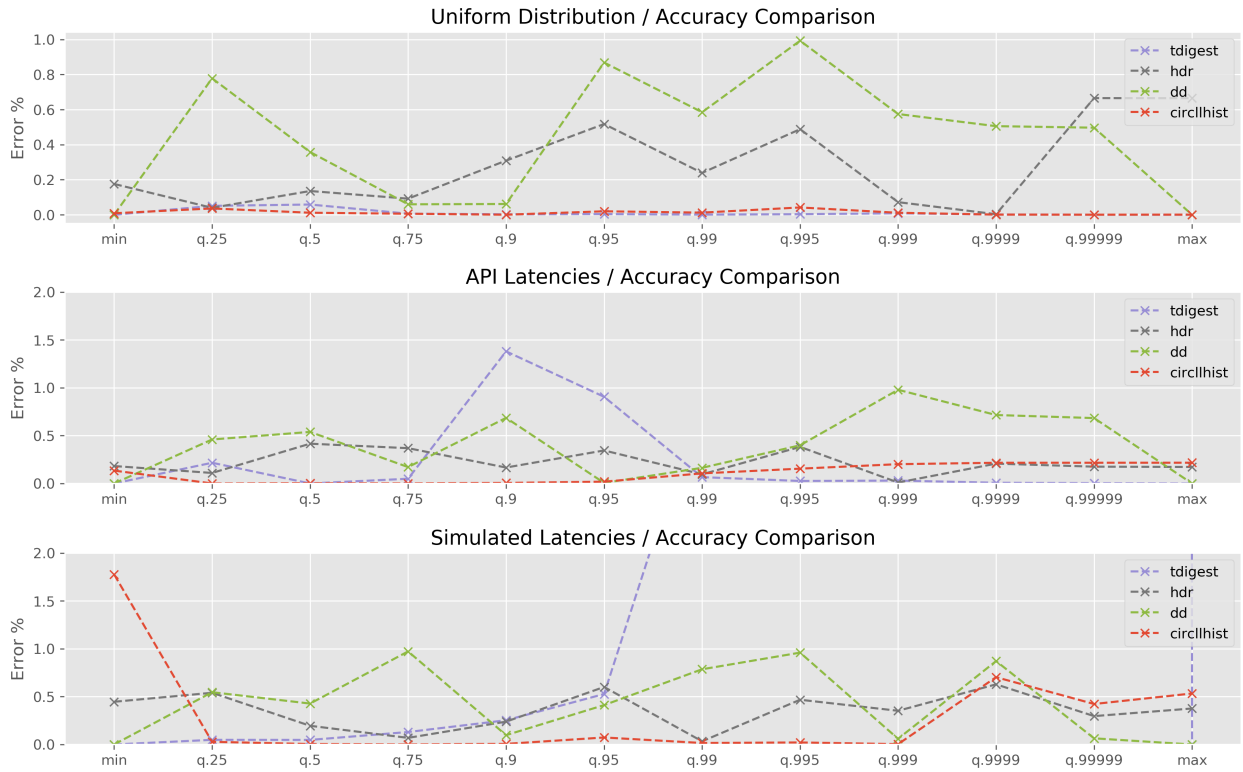
With this configuration, the circllhist was consistently the fastest method when it comes to insertion and merge. For the quantile calculations the t-digest is very efficient, followed by circllhist and DDSketch. Quantile calculations for the HDR Histogram take a lot longer, which probably indicates optimization potential.

## 5.5 Accuracy

As we have seen in section 3.5, there are a number of different quantile definitions circulating in the wild. When evaluating quantile accuracy, we have to make sure we are comparing the computed numbers to the theoretical quantiles the methods are approximating. With the notation of Definition 3.26, the circllhist and Prometheus approximate type-1 quantiles, DDSketch approximates minimal type-7 quantiles. t-digest and HDR Histograms approximate custom quantile functions, that we called “type-tdigest” and “type-hdr” in Definition 3.26.

<sup>9</sup> The insertion time is reported as per inserted sample. The merge time is the time per merged batch. The quantile time is reported per calculated quantile.

<sup>10</sup><https://docs.python.org/3/library/timeit.html>



(a) Accuracy Comparison

	Quantile	0	0.25	0.5	0.75	0.9	0.95	0.99	0.995	0.999	0.9999	0.99999	1
Uniform Distribution	prom	100.00	0.01	0.00	0.01	0.06	0.01	0.01	0.04	0.01	0.00	0.00	0.00
	tdigest	0.00	0.05	0.06	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
	hdr	0.18	0.04	0.14	0.09	0.31	0.52	0.24	0.49	0.07	0.00	0.67	0.67
	dd	0.00	0.78	0.36	0.06	0.06	0.87	0.59	0.99	0.57	0.51	0.50	0.00
	circllhist	0.01	0.04	0.01	0.01	0.00	0.02	0.01	0.04	0.01	0.00	0.00	0.00
API Latencies	prom	100.00	52.67	126.47	145.16	38.95	10.63	5.66	7.31	8.58	8.86	8.89	8.89
	tdigest	0.00	0.22	0.00	0.05	1.38	0.91	0.07	0.03	0.03	0.01	0.00	0.00
	hdr	0.18	0.11	0.42	0.37	0.17	0.34	0.10	0.38	0.01	0.21	0.18	0.17
	dd	0.00	0.46	0.54	0.17	0.69	0.01	0.16	0.40	0.98	0.71	0.68	0.00
	circllhist	0.13	0.00	0.00	0.00	0.01	0.02	0.11	0.15	0.20	0.22	0.22	0.22
Simulated Latencies	prom	100.00	1711.93	1526.34	858.07	294.85	93.91	129.75	14.93	4.81	13.55	437.21	87.35
	tdigest	0.00	0.05	0.05	0.13	0.25	0.53	3.69	11.88	41.43	122.94	1062.53	0.00
	hdr	0.45	0.54	0.20	0.07	0.24	0.60	0.03	0.47	0.35	0.63	0.30	0.38
	dd	0.00	0.55	0.43	0.97	0.10	0.41	0.79	0.96	0.06	0.87	0.06	0.00
	circllhist	1.78	0.03	0.00	0.00	0.01	0.07	0.02	0.02	0.00	0.70	0.42	0.53

(b) Relative errors for quantile calculation in percent.

In all cases, the following quantile values were considered:

$$0, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99, 0.995, 0.999, 0.9999, 0.99999, 1.$$

For each value, the quantile is computed once with the evaluated data-structure and once with the respective theoretical function on the raw data. The relative difference between those value is reported in Figure 7a and Table 7b.

The first thing to note, is that the three histogram-based methods (HDR, DDSketch and circllhist) all compute quantiles with a relative error of below 2%.

In the body of the distribution accuracy of the circllhist is generally a little better than that of the DDSketch and the HDR Histogram. This is due to fact that circllhist uses fair resampling for quantile calculations, whereas DDSketch uses pareto midpoint resampling. It’s also visible, that DDSketch tracks min and max values separately, and reports those values exactly.

The accuracy of the t-digest is very high for the “Uniform Distribution” dataset and on the tails of the “API Latency Dataset”. However, the high quantiles of the “Simulated Latencies” dataset have relative errors of more than 100%. This example illustrates, that the t-digest does not give any accuracy guarantees for quantile approximation after multiple

merging steps have been performed. There are a-priori error bounds for the initial data ingestion, but those are not guaranteed to hold after merging steps. This particular dataset involves a challenging merge of 1000 batches with highly variable distribution and size. As the authors of [3] write:

We can force a digest formed by merging other digests to be fully merged by combining centroids wherever consecutive clusters taken together meet the size bound. The resulting t-digest will not necessarily be the same as if we had computed a t-digest from all of the original data at once even though it will meet the same size constraint. [...] This loss of strictly ordering makes it difficult to compute rigorous error bounds.

Our example suggests, that general rigorous error bounds are unlikely to exist.

Prometheus quantiles are only accurate for the “Uniform Distribution” dataset. For the others errors of >100% are not uncommon.

## 6 Conclusion

In this article we have introduced the circllhist as a data-structure for summarizing data which allows accurate quantile calculations on aggregates. To do so, we developed a general theory of log-linear histograms, and established a-priori error bounds for reconstructed samples (4.76%) and computed quantiles (10%).

We compared this data-structure to alternative data-structures which are employed in practice for aggregated quantile calculations: Prometheus Histograms [4], t-digest [3], HDR Histograms [5], and DDSketches [1].

We have seen that, like the circllhist, also HDR Histograms and DDSketches arise as special cases of abstract log-linear histograms as described in this document. As a result the differences between these methods come down to configuration choices and implementation details. In particular, all three methods operate on essentially unbounded data ranges, offer fast insertion and merge performance, approximate quantiles with comparable high accuracy (< 2%).

The Prometheus histogram, is a basic histogram data-structure that requires explicit configuration of bin boundaries. It’s reliance on numeric time-series as backing data-structures makes using large numbers of bins impractical. With the recommended number of 10 bins, the quantile accuracy was not competitive with the other considered methods.

The t-digest, is the only considered method that is not based on histograms. It brings competitive insertion and merge performance, as well as very fast quantile calculations. It also delivered precise quantile estimates for most of the considered cases. However, it does not guarantee a-priori bounds on the relative error, as do the log-linear histogram methods. This could be seen in one of the synthetic data-sets where large deviations of quantile values could be experienced after multiple merges had taken place.

In comparison to the other methods, the circllhist is the oldest available method, dating back to 2011 when it was first introduced in the Circonus product. It’s a one-size fit’s all method that does not require any configuration and delivers unbounded data range, full mergeability, and good accuracy (typically < 1%, worst-case < 10%) for various summary statistics including quantile calculations. It comes with a mature and polished implementation that delivers best-in class performance for insertion and merge operations. The circllhist has been used in the last decade at countless internal and third-party sites, for a highly diverse set of applications including high volume load balancing<sup>11</sup>, in-kernel latency measurements[8] and general application performance monitoring.

## References

- [1] M. Charles, J.E. Rim, H.K. Lee. DDSketch: A fast and fully-mergeable quantile sketch with relative-error guarantees. Proceedings of the VLDB Endowment 12.12 (2019): 2195-2205.
- [2] Hyndman, R. J. and Fan, Y. Sample quantiles in statistical packages American Statistician 50, 361–365. 10.2307/2684934, 1996
- [3] T. Dunning and O. Ertl. Computing extremely accurate quantiles using t-digests. <https://github.com/tdunning/t-digest>, 2017.
- [4] Prometheus. An open-source systems monitoring and alerting toolkit. <https://prometheus.io/>
- [5] G. Tene. HdrHistogram: A high dynamic range (hdr) histogram <http://hdrhistogram.org/>, 2012
- [6] Travis E. Oliphant. A guide to NumPy Trelgol Publishing, (2006). <https://numpy.org/>

<sup>11</sup>as part of Envoy Proxy <https://www.envoyproxy.io/>

- [7] H. Hartmann. Latency SLOs done right. SRECon 2018. <https://www.usenix.org/conference/srecon19emea/presentation/hartmann-latency>
- [8] H. Hartmann. Linux System Monitoring with eBPF, Blog, 2018. <https://www.circonus.com/2018/05/linux-system-monitoring-with-ebpf/>
- [9] H. Hartmann. Quantiles Blog, 2019. <https://www.heinrichhartmann.com/math/quantiles.html>
- [10] T. Schlossnagle. What's in a number? NYCDevOps 2011. Slides: <https://www.slideshare.net/postwait/whats-in-a-number>
- [11] T. Schlossnagle. It's all about telemetry. Velocity 2012. Slides: <https://www.slideshare.net/postwait/its-all-about-telemetry>
- [12] T. Schlossnagle. Understanding data with Histograms. Blog 2012. <https://www.circonus.com/2012/09/understanding-data-with-histograms/>
- [13] Circonus. libcircllhist: An implementation of Circonus Log-Linear Histograms <https://github.com/circonus-labs/libcircllhist>, 2016