


# Reproducible Floating-Point Aggregation in RDBMSs

**Conference Paper****Author(s):**

Müller, Ingo ; Arteaga, Andrea; Hoefler, Torsten; Alonso, Gustavo

**Publication date:**

2018

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000304330>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

<https://doi.org/10.1109/ICDE.2018.00098>

# Reproducible Floating-Point Aggregation in RDBMSs

Ingo Müller<sup>1\*</sup>Andrea Arteaga<sup>2</sup>Torsten Hoefler<sup>1</sup>Gustavo Alonso<sup>1</sup>

<sup>1</sup>*Systems Group, Dept. of Computer Science, ETH Zurich*  
 {ingo.mueller, torsten.hoefler, alonso}@inf.ethz.ch

<sup>2</sup>*Federal Institute of Meteorology and Climatology MeteoSwiss*  
 andrea.arteaga@meteoswiss.ch

**Abstract**—Industry-grade database systems are expected to produce the same result if the same query is repeatedly run on the same input. However, the numerous sources of non-determinism in modern systems make reproducible results difficult to achieve. This is particularly true if floating-point numbers are involved, where the order of the operations affects the final result.

As part of a larger effort to extend database engines with data representations more suitable for machine learning and scientific applications, in this paper we explore the problem of making relational GROUPBY over floating-point formats *bit-reproducible*, i.e., ensuring any execution of the operator produces the same result up to every single bit. To that aim, we first propose a numeric data type that can be used as drop-in replacement for other number formats and is—unlike standard floating-point formats—associative. We use this data type to make state-of-the-art GROUPBY operators reproducible, but this approach incurs a slowdown between  $4\times$  and  $12\times$  compared to the same operator using conventional database number formats. We thus explore how to modify existing GROUPBY algorithms to make them bit-reproducible and efficient. By using vectorized summation on batches and carefully balancing batch size, cache footprint, and preprocessing costs, we are able to reduce the slowdown due to reproducibility to a factor between  $1.9\times$  and  $2.4\times$  of aggregation in isolation and to a mere 2.7% of end-to-end query performance even on aggregation-intensive queries in MonetDB. We thereby provide a solid basis for supporting more reproducible operations directly in relational engines.

## I. INTRODUCTION

The continued progress of all areas of computer science has led to digitization and automation of many everyday processes. In particular, algorithms are responsible for making decisions in an increasing number of commonplace situations [16]. Consequently—and understandably—, society has started to demand accountability from the algorithms by which it is affected. For example, the General Data Protection Regulation of the European Union [15] has recently given the “right to explanation” to individuals affected by automated decision-making. Similarly, the ACM published a *Statement on Algorithmic Transparency and Accountability* [1] including explainability as a principle that algorithmic decision-making should follow.

In this paper we take steps towards improving explainability of today’s data processing systems, namely the reproducibility of algorithms based on floating-point arithmetic. This problem

\*Part of this work was carried out while this author was working part-time at Oracle Labs, Zurich.

An extended version of this paper is available at <http://arxiv.org/abs/1802.09883>.

```
CREATE TABLE R      (i int, f float);
INSERT INTO R VALUES (1, 2.5e-16);
INSERT INTO R VALUES (2, 0.9999999999999999);
INSERT INTO R VALUES (3, 2.5e-16);
SELECT SUM(f) FROM R; -- Returns 0.9999999999999999
UPDATE R SET i = i + 1 WHERE i = 2;
-- 'f' is unchanged, but rows are physically reordered
SELECT SUM(f) FROM R; -- Returns 1.0!
```

Algorithm 1: Example of non-reproducible SQL query.

was brought to our attention by several of our industry partners. In addition to more classical use-cases like debugging, testing, certification, and redundant computations, where reproducibility can be helpful or necessary [7, 9], they observe that many users, in particular non-experts, are confused by non-reproducible or, in general, non-predictable behavior.

Today’s data management systems often become non-reproducible if floating-point arithmetic is used. The problem with floating-point arithmetic is that, unlike arithmetic on real numbers, it is not associative, so the order in which operations are carried out may change their outcome [17]. The order of operations, in turn, may be affected by a large list of mechanisms: For example, concurrent execution of multiple threads may be non-deterministic, the number of processing elements may influence how the work is split into sub-tasks, and the data storage layer may physically reorder data for a number of reasons. As soon as floating-point numbers are involved, most of today’s systems do not follow the principle of *data independence*, which demands that changes on the physical level shall not have an impact on the result of queries.

Algorithm 1 illustrates how the order of records in the storage layer may affect query results in a subtle and potentially surprising way. The situation shown was produced on a fresh installation of PostgreSQL 9.5.1. The same query summing up three floating-point numbers returns two different results before and after the update of an unrelated attribute. Since, internally, the update is implemented as the creation of a new record and the masking of the old one, the physical order is different in the two queries, which, consequently, yields two different results (differing on all digits of the decimal representation).

One may be tempted to brush away the problem of non-reproducibility with the argument that the underlying rounding errors are rather small and can, hence, be ignored. However, these small errors may still lead to very different outcomes for

individual records, which are hard to explain to the affected individual. For instance, the GROUPBY query of Algorithm 1 extended with a HAVING SUM( $f$ )  $\geq 1$  clause could end up returning specific records in some runs but not in others.

Such misclassifications can affect applications in obvious ways: We ran PageRank on different permutations of a small web graph with 900k pages.<sup>1</sup> We observed that, from one run to the next, the ranks of about 10-20 pages would be *different enough to swap ranks with another page*.

In this paper we show how to make GROUPBY aggregation using SUM reproducible. This essentially solves the problem for SQL: With a reproducible aggregate function for floating-point SUM, *all* aggregate functions in SQL can be made reproducible as well, including non-standard ones such as VARIANCE, STDDEV, and some statistical functions, all of which can be computed using SUM. Furthermore, many projections are intrinsically reproducible and window functions can either be solved with our approach, define an execution order, or they are not reproducible even with integers. Finally, GROUPBY aggregation is not only used in relational database systems, but in virtually every data processing system (possibly under a different name including REDUCE or REDUCEBYKEY), to which our results apply as well.

We start with proposing a format for floating-point numbers that is—unlike formats typically supported by hardware—associative. The format can be implemented in software and builds on techniques from high-performance computing (HPC) [3, 12, 13]. The key to this approach is to anticipate rounding errors by subtracting lower-order terms from each value before it is added to the aggregate of its group.

While this makes it possible to make any algorithm on floating-point numbers bit-reproducible with little to no modification, it comes at a high price: We show that state-of-the-art GROUPBY operators become about  $4\times$  and  $12\times$  slower using this approach, depending on the desired precision. The challenge is, hence, to keep the overhead of the additional calculations at an acceptable level. However, the tuning techniques used in HPC work for the sum of a single vector, while in a SQL GROUPBY, there is potentially a very large number of sums involved. This is not compatible with known data processing techniques, which usually aggregate input tuples as early as possible instead of physically “grouping” them. State-of-the-art aggregation algorithms used with our data type, hence, switch between the summation of different groups *for every input tuple*, which explains the high overhead.

To remedy this problem, we design a novel GROUPBY algorithm based on a concept we call *summation buffer*. The main idea is to buffer input values for each group and delay their aggregation until it can be done efficiently for the whole buffer. As we need a summation buffer for every group, the number of groups that we can process efficiently at the same time is limited by how many buffers we can keep in cache. We thus tune the summation routine to small buffer sizes and use highly-tuned partitioning routines as preprocessing. This

reduces the slowdown of aggregations due to reproducibility to a factor between  $1.9\times$  and  $2.4\times$  over non-reproducible aggregation on built-in floating-point numbers, depending on the number of groups in the input. Integration into a real system, MonetDB [8], shows that we can bring the overhead of end-to-end query performance down to as little as 2.7%. Since our implementation hides almost all computations behind memory accesses, we can even increase accuracy with minimal additional cost and, hence, as a side effect provide higher accuracy than IEEE numbers at essentially the same price, which is crucial in many scientific applications.

To summarize, the paper makes the following contributions:

- We propose a highly tuned algorithm for reproducible summation of floating-point numbers using SIMD instructions (Section III).
- We show how state-of-the-art algorithms for aggregation with GROUPBY can be made bit-reproducible and more accurate with relatively little effort if compromises in performance can be made (Section IV).
- We design a novel grouping algorithm that improves upon this approach, which reduces the slowdown of reproducibility to a  $1.9\times$  and  $2.4\times$  (Section V).
- We show the trade-offs offered by the different algorithms in extensive experiments and quantify their impact on end-to-end query performance in a real system (Section VI).

## II. PROBLEM DEFINITION

We start by illustrating the cause of non-reproducibility of floating-point summation and by discussing potential solutions for bit-reproducibility, which, unfortunately, either do not actually solve the problem or have prohibitive costs.

### A. Reproducible Floating-Point Aggregation with GROUPBY

For the purpose of this paper, we define aggregation with GROUPBY as the operation that turns a sequence of  $\langle key, value \rangle$  pairs into the  $\langle key, aggregate \rangle$  pairs where each key of the input occurs exactly once in the output and the aggregate stored with a key is equal to the sum of all values with that key in the input. We say that it runs on floating-point numbers if the *value* fields of the input pairs are floating-point numbers. An aggregation algorithm is bit-reproducible, or reproducible for short, if the *aggregate* of each group has exactly the same bit pattern for any execution.

### B. Floating-Point Numbers and Associativity

Floating-point values are numbers of the form  $x = M \cdot 2^E$ , where  $M \in [1, 2)$  is called *mantissa* and  $E \in \{E_{min}, E_{max}\}$  is the *exponent*. As the number of relevant bits  $m$  in the mantissa as well as the exponent are finite, only a finite subset of real numbers can be represented exactly. Hence, a rounding function  $rd$  is required in order to map real numbers to representable floating-point values. This includes the results of arithmetic expressions, which may not be representable even if the operands are. For example, the floating-point sum of two floating-point numbers  $a$  and  $b$  is defined as  $a \oplus b = rd(a + b)$ .

<sup>1</sup><https://snap.stanford.edu/data/web-Google.html>

To understand why this can be a problem, consider the numbers  $a = b = 1.01_2 \cdot 2^0$  and  $c = 1.11_2 \cdot 2^1$  in a toy format for floating-point numbers with a mantissa of  $m = 2$  (given as binary number) and truncation for  $\text{rd}$ . To compute the sum of the three numbers, we can compute  $(a \oplus b) \oplus c = \text{rd}(\text{rd}(a+b)+c)$ . Since  $\text{rd}(a+b) = \text{rd}(1.010_2 \cdot 2^1) = 1.01_2 \cdot 2^1$  and  $\text{rd}(1.01_2 \cdot 2^1 + c) = \text{rd}(1.100_2 \cdot 2^2) = 1.10_2 \cdot 2^2$ , no rounding errors occur and the sum is accurate. However, we can compute the sum as well as  $a \oplus (b \oplus c) = \text{rd}(a + \text{rd}(b+c))$ . In this case  $\text{rd}(b+c) = \text{rd}(1.0011_2 \cdot 2^2) = 1.00_2 \cdot 2^2$  and  $\text{rd}(a + 1.00_2 \cdot 2^2) = \text{rd}(1.0101_2 \cdot 2^2) = 1.01_2 \cdot 2^2$ , so rounding errors occur in both operations (typeset in bold). Note that the *sum* of the rounding errors is  $1.00_2 \cdot 2^0$ , which could be added to  $a \oplus (b \oplus c)$  without rounding error.

Rounding errors are larger if the exponents of the two summands are different. Therefore, if we compute the sum of many numbers, the rounding error incurred during the addition of a particular input value depends on the current value of the accumulator, which depends on the order of execution. Furthermore, even though each error is small, their sum may be big enough to change the final result.

The problem also occurs in the most common floating-point formats, which are the ones defined by the IEEE-754 standard [32] (even if the absolute error is obviously smaller due to the higher precision than our toy format) and which we use throughout this paper.

### C. Non-Solutions for Reproducibility

We now discuss a number of naive approaches for making aggregation with GROUPBY reproducible, but which do not provide satisfactory solutions to the problem.

**Higher precision.** Using a truncated or rounded result produced by operations with a higher floating-point precision (i.e., using doubles instead of floats) is not sufficient, as it does not make it more reproducible: Even tiny rounding errors can make significant bits flip (such as from 0.999999... to 1).

**Deterministic order of operations.** It is possible to make the order of the operations deterministic. For linear algebra, the cuBLAS library [21] and the Intel Math Kernel Library [25] follow this approach. However, this does not solve the entire problem for database systems, which aim for *data independence* as discussed above. In addition to the example given in the introduction, the physical order of the input may also change due to compression, data placement on distributed machines, backup and restore operations, and other mechanisms, which in turn changes the order of operations. The only way to make the order of the operations deterministic is thus to use a static and deterministic schedule *and* to sort the input, which may be more than an order of magnitude slower [4, Figure 5.8] than state-of-the-art AGGREGATION algorithms based on hashing.

**Fixed-point arithmetic.** In traditional workloads, it is often possible to use fixed-point arithmetic for fractional numbers (also called *binary scaling* or *decimal-scaled binaries* if the scale has base ten). This is the case if all input numbers are integer multiple of some common denominator and come from

Name	Meaning
$n$	Number of input values
$L$	Number of levels
$V$	Size of vector register
$W$	Bit width of error-free transformation
$NB$	Block size between carry-bit operations
$m$	Size of the mantissa
$f$	Exponent of the first transformation level
$b_i$	Input value
$q_i^{(l)}$	Contribution of $b_i$ at level $l$
$r_i^{(l)}$	Remainder of $b_i$ at level $l$
$S^{(l)}$	Running sum at level $l$
$C^{(l)}$	Carry-bit count at level $l$
$Q^{(l)}$	Sum of the contributions at level $l$
$\text{ufp}(x)$	Unit in the first place
$\text{ulp}(x)$	Unit in the last place
$\text{rd}(x)$	Rounding function
$x \oplus y$	$\equiv \text{rd}(x + y)$ , floating-point sum of $x$ and $y$
$x \ominus y$	$\equiv \text{rd}(x - y)$ , floating-point subtraction of $x$ and $y$

Table I: Summary of the parameters, variables, and functions used in Section III.

similar orders of magnitude, such as all values in a *salary* field are multiples of 1¢ and range between some thousand and some million dollars. Operations can then be executed using integer operations internally, which are cheap to execute and reproducible most of the time. However, in many modern data processing applications, these assumptions do not apply: values from some domains, such as measurements or scientific data, cannot be expressed as multiple of some smallest unit and the values of different orders of magnitude such as those handled in machine learning and other scientific applications require a floating-point representation.

**Arbitrary-precision operations.** It is possible to push the limits of fixed-point arithmetic by using high-precision or even arbitrary-precision operations in software. Examples implementing this approach include the GNU MPFR Library [28], the BigDecimal class of Java [23], and numeric data types offered by some database systems (such as PostgreSQL). However, this not only requires many hardware instructions for each arithmetic operation, but also variable-width storage, which is much more difficult to handle than the fixed-width built-in types.

## III. REPRODUCIBLE ACCURATE SUM

In this section, we explain an algorithm that solves the problem of reproducible floating-point summation where all inputs are summed up to produce a *single* number (i.e., aggregation *without* grouping). For the sake of exposition, we develop the algorithm in three iterations of increasing completeness. In a fourth iteration, we show how to speed up this algorithm using vector instructions. Table I summarizes the variables and function names we use.

### A. Definitions

Two numbers related to any floating-point number  $x$  are of importance in this section:  $\text{ufp}(x)$  and  $\text{ulp}(x)$ . They were first defined by Goldberg [17].  $\text{ufp}(x)$  designates the *unit in the*

first place, i.e., the numeric value of the first bit in the mantissa. If  $x = M \cdot 2^E$  as above, then  $\text{ufp}(x) = 2^E$ . All floating-point numbers with the same exponent as  $x$  have an absolute value in  $[\text{ufp}(x), 2 \cdot \text{ufp}(x))$ . Similarly,  $\text{ulp}(x)$  designates the *unit in the last place*, i.e., the value of the last bit in the mantissa. For  $x = M \cdot 2^E$ ,  $\text{ulp}(x) = 2^{E-m}$ . This value represents the difference between  $x$  and its closest representable values.

### B. Error-Free Transformation

Our reproducible summation algorithm is based on the observation that the floating-point sum of two numbers  $a$  and  $b$  can be performed *exactly* if the one with the smaller absolute value has sufficiently many zeroes at the end of its mantissa. To understand when this is the case, let us define  $a$  and  $b$  as integer multiples of the same power of two:  $a := \alpha_a \cdot 2^p$ ,  $b := \alpha_b \cdot 2^p$ , where  $\alpha_a, \alpha_b$ , and  $p$  are integer numbers, and  $|a| > |b|$ , WLOG. If the values  $|\alpha_a|, |\alpha_b|$ , and  $|\alpha_a + \alpha_b|$  do not exceed  $2^m$ , then  $a, b$ , and  $a + b$  can be represented in the floating-point format and, consequently, the floating-point sum is exact:  $a \oplus b = a + b$ . In other words, the sum is exact if these three numbers can have a (possibly denormalized) floating-point representation with the same exponent and without losing information.

As an example, the values  $a := 26.046875$  and  $b := 2.8125$  can be represented exactly with an 11-bit mantissa (which corresponds to IEEE half-precision, where  $m = 10$ ). Also, they can be represented as  $\alpha_a \cdot 2^p$  and  $\alpha_b \cdot 2^p$ , with  $\alpha_a = 1667$ ,  $\alpha_b = 180$ , and  $p = -6$  and their sum is  $a \oplus b = a + b = (\alpha_a + \alpha_b) \cdot 2^p = 28.859375$ . It can be computed exactly with this format because said conditions are met.

Let us now consider any representable  $a$  and  $b$  with  $|a| > |b|$ . One can split  $b$  as  $b = q + r$  so that  $q$  is an integer multiple of  $\text{ulp}(a)$ , namely  $q := (a \oplus b) \ominus a$  and  $r := b \ominus q$ . The sum  $a \oplus q$  can be computed exactly because the three aforesaid conditions are met between  $a$  and  $q$ , with  $2^p = \text{ulp}(a)$ . Also,  $b$  can be recovered exactly through  $q + r = q \oplus b$ . This procedure, named *error-free transformation*, was defined by Ogita et al. [22].

For instance,  $a = 1.010_2 \cdot 2^0$ ,  $b = 1.101_2 \cdot 2^{-2}$ .  $q$  is computed as  $(a \oplus b) \ominus a = 1.101_2 \cdot 2^0$ , while  $r = b \ominus q = 1_2 \cdot 2^{-5}$ .

We can now perform an order-independent summation of a sequence of values  $b_1, b_2, b_3$  by applying the same error-free transformation to all values  $b_i$ . Figure 1 illustrates the idea. The transformation produces  $q_i := (a \oplus b_i) \ominus a$ ,  $r_i := b_i \ominus q_i$ . We call the values  $q_i$  and  $r_i$  *contributions* and *remainders*, respectively, of the input values  $b_i$  and the value  $a$  the *extractor* of the error-free transformations. Since all contributions  $q_i$  are integer multiples of the same power of two, and that their sum is representable with the same format, their floating-point summation is free of rounding error and thus order-independent:  $q_1 \oplus q_2 \oplus q_3 = q_1 + q_2 + q_3 = 265$ .

While this procedure solves the problem of order-independent bit-reproducibility, it has two problems: First, it only works under the assumption that the absolute value of every intermediate result of this summation, including the final result, is strictly bounded by  $2 \cdot \text{ufp}(a)$ . An obvious, yet suboptimal solution would be to do two passes over the data, the first one computing the maximum absolute value

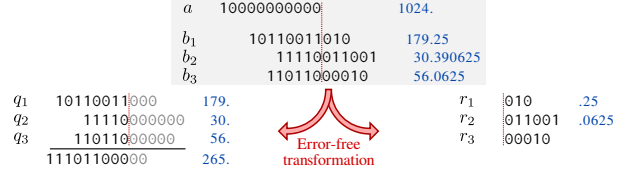


Figure 1: Error-free transformation.

for which all assumptions are fulfilled, the second one for the actual summation. Second, the sum is inaccurate because relevant parts of the input values, namely the remainders, are discarded. We present an algorithm that solves both problems.

### C. Accurate Reproducible Scalar Summation

In HPC, the problem of summing up long vectors of numbers has been studied in detail. In this section, we explain RSUM [13]. In the subsequent section, we discuss why it does not work well with SQL GROUPBY.

In order to address the problem of accuracy, we recall that the error-free transformation produces two outputs: the *contribution*  $q$  and the *remainder*  $r$ . In the previous section, we made use of the contributions to obtain an order-independent summation. Now we make use of the remainders to improve the accuracy of this summation: we perform an error-free transformation on the remainder, this time using the smaller extractor  $a^{(2)} := 2^{-W} \cdot a$  (with  $W \in \mathbb{N} \setminus \{0\}$ ). We thus obtain the second-level contribution and remainder of each input value:  $q_i^{(2)} := (a^{(2)} \oplus r_i) \ominus a^{(2)}$ ,  $r_i^{(2)} := r_i \ominus q_i^{(2)}$ . These second-level contributions can be summed up to obtain a second-level result:  $Q^{(1)} := \sum_{i=1}^n q_i$ ,  $Q^{(2)} := \sum_{i=1}^n q_i^{(2)}$ . As we show in Section VI-B, the final result  $Q := Q^{(1)} \oplus Q^{(2)}$  is of comparable accuracy as a standard, non-reproducible floating-point summation. If higher accuracy is needed, an arbitrary number of levels  $L$  can be used. The value  $W$  expressing the logarithm of the ratio of two consecutive extractors  $a^{(l)}, a^{(l+1)}$  is bounded by  $m - 2$  and it affects the result (the higher, the more accurate) and the cost (the higher, the slower) of the algorithm. Good choices are 18 and 40 for single and double precision respectively and we use these values in this work.

So far, the extractors  $a$  are never assumed to be powers of two. The example in Figure 1 shows a power of two as the extractor, but this does not necessarily have to be the case. The only important factor is that the exponent of the extractor never changes, nor do the intermediate results of the error-free transformation. For this reason, the role of the error-free extractor is taken by the running sums  $S^{(l)}$  in the algorithm. The running sum will never change its exponent, as is explained later in this section.

We start with the values  $S^{(l)} = 1.5 \cdot 2^{f-(l-1) \cdot W}$ . The value  $f$  can be chosen arbitrarily, as long as the resulting extractor is large enough for the transformation of the first value  $b_1$ , i.e.,  $f > \log_2 |b_1| + m - W + 1$ . Each input value  $b_i$  is transformed using  $S^{(1)}, S^{(2)}, \dots, S^{(L)}$  as extractors. The resulting contributions  $q_i^{(1)}, q_i^{(2)}, \dots, q_i^{(L)}$  are added to  $S^{(1)}, S^{(2)}, \dots, S^{(L)}$  respectively. For  $|b_i| \geq 2^{W-1} \cdot \text{ulp}(S^{(1)})$ , the first level is not large enough to contain its contribution. In

```

1: Load state  $S^{(l)}, C^{(l)} \forall 1 \leq l \leq L$ 
2: for  $i = 1$  to  $n$  do
3:    $\triangleright$  Check extractor validity, update levels if needed:
   necessary
4:   while  $|b_i| \geq 2^{W-1} \cdot \text{ulp}(S^{(1)})$  do
5:     for  $l = L$  to  $2$  do
6:        $S^{(l)} \leftarrow S^{(l-1)}; C^{(l)} \leftarrow C^{(l-1)}$ 
7:        $S^{(1)} \leftarrow 1.5 \cdot 2^W \cdot \text{ufp}(S^{(2)}); C^{(1)} \leftarrow 0$ 
8:      $\triangleright$  Load and transform value  $b_i$ , update  $S^{(l)}$ :
9:      $r_i^{(0)} \leftarrow b_i$ 
10:    for  $l = 1$  to  $L$  do
11:       $q_i^{(l)} \leftarrow (r_i^{(l-1)} \oplus S^{(l)}) \ominus S^{(l)}$ 
12:       $S^{(l)} \leftarrow S^{(l)} \oplus q_i^{(l)}$ 
13:       $r_i^{(l)} \leftarrow r_i^{(l-1)} \ominus q_i^{(l)}$ 
14:     $\triangleright$  Carry-bit propagation:
15:    for  $l = 1$  to  $L$  do
16:      Find  $d \in \mathbb{Z}$  s.t.  $S^{(l)} \ominus d \cdot 0.25 \cdot \text{ufp}(S^{(l)}) \in$ 
       $[1.5 \cdot \text{ufp}(S^{(l)}), 1.75 \cdot \text{ufp}(S^{(l)})]$ 
17:       $S^{(l)} \leftarrow S^{(l)} \ominus d \cdot 0.25 \cdot \text{ufp}(S^{(l)})$ 
18:       $C^{(l)} \leftarrow C^{(l)} \oplus d$ 
19: Store state  $S^{(l)}, C^{(l)} \forall 1 \leq l \leq L$ 

```

Algorithm 2: RSUM SCALAR.

this case, the last level  $S^{(L)}$  is discarded, all other levels are demoted (e.g., the first-level sum becomes the second-level sum,  $S^{(1)} := S^{(2)}$ , and the new first-level sum is initialized to  $S^{(1)} := 1.5 \cdot 2^W \cdot \text{ufp}(S^{(2)})$ ). This ensures that all input values can be included in the summation without breaking the assumptions for reproducible results, also avoiding the need for a first pass to find the maximum absolute value in the input.

In order to avoid that the running sums  $S^{(l)}$  change their exponent (which would affect the error-free transformation), a check is performed before its usage.  $S^{(l)}$  is *usable*, if it lies in the range  $[1.5 \cdot \text{ufp}(S^{(l)}), 1.75 \cdot \text{ufp}(S^{(l)})]$ . If it does not, a multiple of  $0.25 \cdot \text{ufp}(S^{(l)})$  is added to or removed from it, and the corresponding value is subtracted from or added to the carry-bit counter  $C^{(l)}$ , which is initialized to 0. For instance, if  $S^{(l)} = 1.84 \cdot \text{ufp}(S^{(l)})$ , then  $1 \cdot (0.25 \cdot \text{ufp}(S^{(l)}))$  is subtracted from it, so that the running sum after this operation is  $S^{(l)} = 1.59 \cdot \text{ufp}(S^{(l)})$ ; the value 1 is added to  $C^{(l)}$ . The complete state of the summation is given by the running sums  $S^{(l)}$  and the corresponding carry-bit counts  $C^{(l)}$ .

Algorithm 2 summarizes this procedure. In this version of the algorithm we assume that the summation state has been initialized. A number of input values are added to this summation state, and its state is stored to main memory again, so that the summation can be resumed later. In order to finalize the summation, the following sum has to be performed:

$$Q := \sum_{l=1}^L \left( (S^{(l)} \ominus 1.5 \cdot \text{ufp}(S^{(l)})) \oplus 0.25 \cdot \text{ufp}(S^{(l)}) \cdot C^{(l)} \right) \quad (1)$$

This sum is not order-independent, thus a predefined order has to be imposed. In order to avoid cancellation, we perform it in reverse order, i.e., we start from the last level.

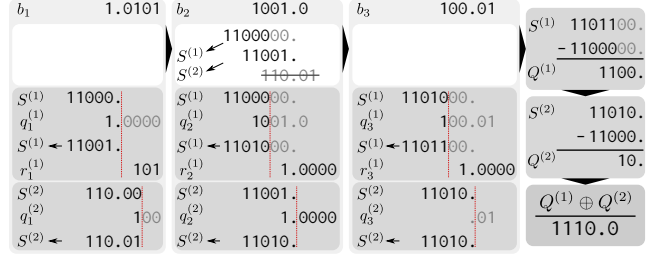


Figure 2: Application of the RSUM algorithm on three values.



Figure 3: Example of a carry-bit propagation.

Figure 2 shows the application of Algorithm 2 on the values  $b_1 = 1.3125$ ,  $b_2 = 9$ , and  $b_3 = 4.25$ , with the format defined by  $m = 4$ , the parameter  $W = 2$ , the first extractor chosen with  $f = 4$ , and two extraction levels. The figure uses only binary digits. In the first iteration, the value  $b_1$  is added to the first-level running sum  $S^{(1)}$  (incrementing it by the contribution  $q_1^{(1)}$ ). The remainder  $r_1^{(1)}$  is added to the second-level running sum  $S^{(2)}$ . In the second iteration,  $|b_2| \geq 2^{W-1} \cdot \text{ulp}(S^{(1)})$ , thus triggering an adjustment of the levels, shown in the white box: The second-level sum is discarded, the first-level sum is moved to the second level, and a new extractor is set as first level. Then, the extraction is performed normally. The third value does not trigger the level adjustment. Finally, the sum for each level is computed ( $Q^{(1)}$  and  $Q^{(2)}$ ) and these values are summed to give the final result  $1110_2 = 14$ .  $C^{(l)}$  variables are never shown in this example because their value is always zero, as  $S^{(l)} \in [1.5 \cdot \text{ufp}(S^{(l)}), 1.75 \cdot \text{ufp}(S^{(l)})]$  at all times. Figure 3 illustrates a carry-bit operation on one level when the running sum  $S^{(1)}$  has the value  $11011_2$ , the carry-bit counter  $C^{(1)} = 0$ , and the value  $b_4 = 3.125$  is processed: after the summation the running sum exceeds  $1.75 \cdot \text{ufp}(S^{(l)})$ . The adjusting number is found to be  $d = 1$ , and the sum and the carry-bit counter are modified accordingly.

#### D. Vectorization of the Summation Algorithm

RSUM [13] was originally introduced in a MIMD context, where each process performs the full summation of the local data and the results are finally summed up globally using `MPI_Reduce`. As a first step to make it suitable to `GROUPBY`, we propose a SIMD variant of RSUM. In this variant, the running sum of each level is represented in the registers as a tuple of values  $\langle S_1^{(l)}, \dots, S_V^{(l)} \rangle$ , where  $V$  is the width of the register (e.g., for double-precision values on AVX architectures,  $V = 4$ ). Similarly, the carry-bit counts are represented by a tuple of  $V$  elements and  $V$  input values are transformed and added to the running sums concurrently. Moreover, a tiling optimization is performed: the extractor validity and the carry-bit propagation are performed just once every  $NB$  iterations. This is bounded by  $NB \leq 2^{-m-W-1}$ [13].

- 1: Load state  $S_1^{(l)} \leftarrow S^{(l)}$ ;  $C_1^{(l)} \leftarrow C^{(l)} \quad \forall 1 \leq l \leq L$
- 2:  $S_v^{(l)} \leftarrow 1.5 \cdot \text{ufp}(S^{(l)})$ ;  $C_v^{(l)} \leftarrow 0 \quad \forall 2 \leq v \leq V, 1 \leq l \leq L$
- 3: **for**  $i = 1$  **to**  $n$ , **increment by**  $V \cdot NB$  **do**
- 4:    $\triangleright$  Check  $\max_{i \leq j < i+V \cdot NB} |b_j|$ , update levels if necessary. See Algorithm 2, lines 3-7
- 5:   **for**  $j = i$  **to**  $i + V \cdot NB$ , **increment by**  $V$  **do**
- 6:      $\triangleright$  Load and transform values  $\langle b_j, \dots, b_{j+V-1} \rangle$ , update  $\langle S_1^{(l)}, \dots, S_V^{(l)} \rangle$ . See Algorithm 2, lines 8-13
- 7:      $\triangleright$  Carry-bit propagation. See Algorithm 2, lines 14-18
- 8:      $\triangleright$  Horizontal summation
- 9:   **for**  $l = 1$  **to**  $L$  **do**
- 10:      $S^{(l)} \leftarrow 1.5 \cdot \text{ufp}(S_1^{(l)}) \oplus \sum_{v=1}^V (S_v^{(l)} \ominus 1.5 \cdot \text{ufp}(S_v^{(l)}))$
- 11:      $C^{(l)} \leftarrow \sum_{v=1}^V \langle C_1^{(l)}, \dots, C_V^{(l)} \rangle$
- 12: Store state  $S^{(l)}, C^{(l)} \quad \forall 1 \leq l \leq L$

Algorithm 3: RSUM SIMD.

The summation state does not change format: one running sum and one carry-bit count per level are stored in main memory. When loading a summation state from memory into the registers, the first element of the registers is set to the value read from memory (thus, e.g.,  $S_1^{(1)} = S^{(1)}$ ), while the other elements of each register are initialized to  $1.5 \cdot \text{ufp}(S^{(l)})$  for running sums and to 0 for carry-bit counts. A horizontal summation has to be performed at the end of the algorithm when storing the state to the main memory: The resulting running sum and carry-bit count of each level are:

$$S^{(l)} := 1.5 \cdot \text{ufp}(S_1^{(l)}) \oplus \sum_{v=1}^V (S_v^{(l)} \ominus 1.5 \cdot \text{ufp}(S_v^{(l)})), \quad (2)$$

$$C^{(l)} := \sum_{v=1}^V C_v^{(l)} \quad (3)$$

with order-independent sums. Algorithm 3 summarizes this procedure. For brevity and clarity, parts of the algorithm constitute references to the equivalent lines of Algorithm 2.

#### IV. A REPRODUCIBLE FLOATING-POINT TYPE

As a first solution for reproducible floating-point aggregation with GROUPBY, we propose a data type that can be used as drop-in replacement for intermediate aggregates of floating-point numbers in any state-of-the-art aggregation algorithm with little to no modification.<sup>2</sup> We base this type on the reproducible summation algorithm from the previous section: It simply consists of an  $\langle \vec{S}, \vec{C} \rangle$  pair, where the symbols  $\vec{S} = \langle S^{(1)}, \dots, S^{(L)} \rangle$  and  $\vec{C} = \langle C^{(1)}, \dots, C^{(L)} \rangle$  are, respectively, the  $L$  levels of the running sum and carry-bit counter as introduced in Section III. In languages such as C++, we can implement this data type as a class with member variables  $S[L]$  and  $C[L]$  and overload its `operator+=` for summation with scalars and

<sup>2</sup>The only arithmetic operation that this type supports is addition, so in a real system it would most likely be an internal type of the execution layer not exposed to the user.

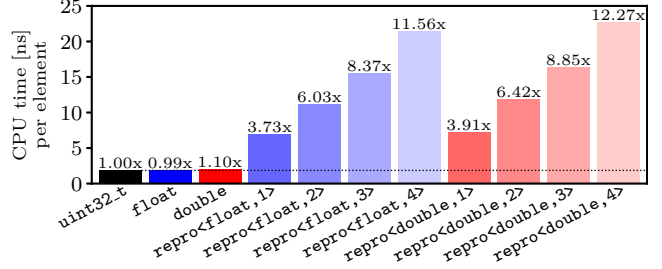


Figure 4: HASHAGGREGATION with different reproducible data types and 16 groups.

instances of that type. We refer to this data type as `repro<ScalarT,L>`, where `ScalarT` is either `float` or `double`.

Figure 4 shows the performance of a start-of-the-art HASHAGGREGATION algorithm<sup>3</sup> instantiated with different variants of the `repro` data type. This algorithm looks up the aggregate of the corresponding group in a hash table using the `key` field of the input pair and adds the `value` field to that aggregate. We choose the small number of 16 groups to eliminate effects not related to the data types themselves (such as cache effects or pre-processing costs). As the plot shows, the algorithm is between  $4 \times$  and  $12 \times$  slower with the reproducible data types than with integers or IEEE floats and the more so the more levels of summation we use (i.e., the higher the precision). This is not surprising considering the computational overhead: in the `operator+=` of reproducible types, each level of summation requires about 12 floating-point operations and 4 load and store instructions, while the `operator+=` of standard data types only requires a single one. Finally, there is virtually no difference between single and double precision. This is due to the fact that the algorithm is heavily compute bound and the latency of most instructions does not depend on the operand width.

We have learned two things from this section. First, it requires relatively little development effort to make a large class of algorithms for aggregation with GROUPBY bit-reproducible on floating-point numbers. Second, this approach comes at a high price: If we want to match the precision of IEEE floats, we need the types with  $L = 2$ , which, in the situation shown above, makes the algorithm more than 6 times slower. In the next section, we show a more involved algorithm that improves the slowdown to between  $1.9 \times$  and  $2.4 \times$  for any `repro<ScalarT,L>`.

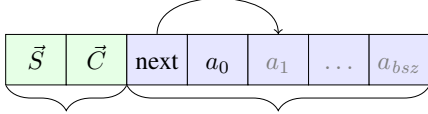
#### V. AGGREGATION WITH SUMMATION BUFFERS

In this section, we present a novel algorithm for aggregation with GROUPBY that achieves bit-reproducibility on floating-point numbers efficiently. We first describe the main idea, summation buffers, which allows to use our efficient, vectorized summation routine, and then incorporate that into a state-of-the-art AGGREGATION algorithm.

##### A. Summation Buffers

The main idea for batching the aggregation of input values can be used in any AGGREGATION algorithm: we store

<sup>3</sup>The experimental setup as described in Section VI-A.



repro<ScalarT,L> accumulation buffer  
 Figure 5: Memory layout of an intermediate aggregate.

a reproducible float along with a buffer of input values as intermediate aggregates, which we call *summation buffer*. A summation buffer consists of an array of input values and the offset of the next free slot in the array (“next”). The layout of intermediate aggregates is thus as shown in Figure 5.

For example, the textbook HASHAGGREGATION [19] with summation buffers works as follows: Whenever we process a  $\langle key, value \rangle$  pair, we first use the key to lookup the entry of the group in the hash table and then use the offset to append the value to the buffer of the group (incrementing the offset accordingly). Only when a buffer is full, we aggregate the buffered values and reset the offset to 0, i.e., to the beginning of the buffer. This allows us to use our vectorized summation algorithm RSUM SIMD (Algorithm 3), which is much more efficient than the per-element summation from the previous section. In languages such as C++, we can implement this as new data type again, where the summation operators contain the logic just described, and use this new data type in any existing AGGREGATION algorithm transparently.

One important tuning parameter is now the buffer size ( $bsz$ ). On the one hand, the larger the buffer is, the better the constant costs associated with a call to the summation algorithm can be amortized. On the other hand, the larger each buffer is, the larger is also the cache footprint of the algorithm, which may decrease performance significantly. The rest of this section shows how to design an algorithm that makes the best trade-off in all situations.

### B. High-Level Algorithm Structure

The overall structure of our AGGREGATION algorithm is illustrated in Algorithm 4: We partition the input on the hash value of the keys (Line 1), which can be done very efficiently on modern hardware [8, 20, 26]. Since all input records for a particular group are copied into the same partition, each partition can be processed independently, which we do using HASHAGGREGATION (Lines 2 to 3). Finally, we combine the intermediate results into a single hash table shared among all threads (Lines 4 to 6). We call this algorithm PARTITIONANDAGGREGATE.

The partitioning effectively divides the number of groups per partition and, hence, the cache footprint of HASHAGGREGATION by the partitioning fan-out  $F$ , which may out-weigh the additional costs for partitioning. Depending on the number of groups in the input, a larger or smaller fan-out is needed in order to fit the working set of HASHAGGREGATION into the cache. For a small number of groups, no partitioning may be required. In this case, i.e., if  $F = 1$ , PARALLELPARTITION is a no-op that forwards its input. Since modern hardware can run PARTITIONING efficiently only up to a certain fan-out [8, 20,

```

1: partitions  $\leftarrow$  PARALLELPARTITION(input, key,  $F = f^d$ )
2: for each p in partitions with index i parallel do
3:   privateTables[i]  $\leftarrow$  HASHAGGREGATION(p)
4: for each t in privateTables parallel do
5:   for each  $\langle key, value \rangle$  in t do
6:     sharedTable[key] += value

```

Algorithm 4: PARTITIONANDAGGREGATE.

26], we implementing it recursively using zero or more levels of partitioning i.e., we partition with  $F = f^d$  for  $f = 256$  and  $d = 0, 1, \dots$

All phases of the algorithm can be fully parallelized: The partitioning routine called in Line 1 can be parallelized by splitting the input in an arbitrary way (e.g., into equally-sized chunks, using a work queue, or using work stealing) and logically concatenating the corresponding output partitions produced by different threads. After the partitioning, each thread gets a subset of the partitions, which it aggregates into a private hash table independently of the other threads (Lines 2 to 3). With some care, the subsequent transfer to the shared hash table (Lines 4 to 6) can be implemented without synchronization since the threads work on non-overlapping subsets of its content. If no partitioning needs to be done, i.e., if  $F = 1$ , the shared hash table needs some form of synchronization such as locks. However, since in this case there are only few groups and each of them only appears once in the hash table of each thread, this last phase takes a negligible amount of time, so the overhead of locking is acceptable.

In order to make this algorithm reproducible, we use summation buffers as the data type for the intermediate aggregates produced by HASHAGGREGATION in Line 3. In the process of aggregating its share of the input or its partitions, each thread calls `operator+=(ScalarT)` on the appropriate intermediate aggregates, which makes it effectively alternate between probing the hash table in order to append input values at the end of their corresponding buffers and summing up the content of buffers as they become full. The shared hash table used in Line 6 has aggregates of type `repro<ScalarT,L>`, i.e., it does not use summation buffers and `operator+=(repro<ScalarT,L>)` is used for merging the intermediate aggregates of the different threads.

The careful reader may wonder why the data of a particular partition is first aggregated into a private hash table and then transferred into a part of the shared hash table that the thread has exclusive access to. It seems possible to save the transfer by aggregating into summation buffers in the shared hash table directly. However, this would have several disadvantages: (1) for all but almost distinct inputs, the transfer is negligible as argued earlier, so there is no need to speed up this phase, (2) the result would consist of summation buffers, which take up more space than needed, and (3) to finalize the computation, we need to iterate over the results anyway in order to flush the buffers.



### C. Tuning Buffer Size and Partitioning Depth

PARTITIONANDAGGREGATE with summation buffers has thus two parameters that influence its cache footprint: the size of the summation buffers  $bsz$  and the partitioning depth  $d$ . We now show how to choose these two parameters.

We first determine the size of the summation buffers given a fixed partitioning depth. Since the access to the summation buffers follows the random pattern given by the hash values of the keys of the input records, the cache footprint of the algorithm consists of the size of the hash table, which we can quantify as  $n_{groups} \cdot \text{sizeof}(\text{ScalarT}) \cdot bsz$ , where  $n_{groups}$  is the number of groups in the input. The buffers should be as large as possible in order to amortize constant costs of calling the summation routine. We thus set the buffer size such that they use the entire cache, which is given by the following equation:

$$bsz = \min \left\{ \left\lceil \frac{|cache|}{(n_{groups}/F \cdot \text{sizeof}(\text{ScalarT}))} \right\rceil, bsz_{max} \right\}, \quad (4)$$

where  $|cache|$  is the size of the last-level cache corresponding to one thread,  $bsz_{max}$  the largest buffer size available in the system, and  $F$  the partitioning fan-out.

The optimal number of levels of partitioning  $d$  depends on the number of groups: It must be large enough to reduce the number of groups per partition to a point where the subsequent, final level of aggregation can be done in cache. It should not be larger, otherwise, the partitioning has no benefit and its execution only constitutes overhead. In earlier work [20], we propose to select this depth adaptively: start with a private hash table of fixed size; while the number of groups is lower than the threshold, process all input this way; if and when the threshold is crossed, add a level of partitioning and recurse. This has virtually no overhead, so the resulting runtime essentially corresponds to the optimal partitioning depth for any given input. Since the adaptation mechanism is orthogonal to the topic of this paper and incorporation into our algorithm is only a matter of implementation time, we simply determine the optimal number of levels offline and use that in the remainder of this paper.

### D. System Integration

We envision the integration of our algorithm into real systems either as a “fix” for SUM on floating-point numbers or as an alternate aggregate function  $\text{Rsum}(\langle expression \rangle, L)$ , which would give the user control on the desired precision.

## VI. EXPERIMENTAL EVALUATION

In this section, we show micro-benchmarks that justify design decisions taken in the previous sections, evaluate the performance of our algorithms experimentally, and quantify their impact on end-to-end query performance.

### A. Experimental Setup

We run the experiments on a system with 256 GiB RAM and two Intel Xeon E5-2630 v3 CPUs, which belong to the Haswell-EP product line. The CPUs have 8 physical cores each clocked at 2.4 GHz, each with private first and second-

level data caches of size 32 KiB and 256 KiB, respectively, as well as a 20 MiB last-level cache shared among all cores. The system runs Debian 8 (jessie) with a Linux kernel v3.18.14. HyperThreading and frequency scaling are switched off.

Unless otherwise mentioned, we use  $n = 2^{30} \langle key, value \rangle$  pairs as input, where the key is of type `uint32_t` and the type of the value is as follows: It is of type `ScalarT` if we say that an algorithm runs on `repro<ScalarT,L>` (i.e., it is of type `float` or `double`) and of type `DECIMAL(p)` if we say that the algorithm runs on one of these types. We implement the `DECIMAL` types as built-in integers of size 32, 64, and 128 bit<sup>4</sup> for  $p = 9, 19, 38$ , respectively, which is a typical way to implement them. The keys are drawn uniformly at random from the range  $[0, n_{groups})$ . Due to the nature of random distributions, this means that there are actually less than  $n_{groups}$  groups in the input if  $n_{groups} \approx n$ . We omit experiments on other data distributions as known techniques to handle data skew [10, 20] are orthogonal to the topic of this paper and can be included into our algorithms. All presented numbers are averages of ten identical runs, among which we observed low variance.<sup>5</sup> We express the runtime as “CPU time per element” =  $T \cdot P/n$ , where  $T$  is the total running time,  $P = 8$  the number of processing elements, and  $n$  the number of input elements, which simplifies comparison across different machines.

In experiments not shown in this paper, we compared our baseline implementation to that of Cieslewicz and Ross [10]. Our implementation is at least as fast as theirs and up to 4 times faster for large numbers of groups because we use the highly-tuned partitioning routine used in other work [8, 26]. Back-of-the-envelope calculations suggest that we achieve the same performance as the implementations used in [20], as well, thereby ensuring our baseline for `GROUPBY` matches the state of the art.

### B. Vectorized Summation Algorithm

We first evaluate accuracy and performance of our summation routine presented in Section III.

1) *Accuracy*: The accuracy of our summation routine `Rsum SIMD` (Algorithm 3) depends on the number of levels of running sums and carry-bit counters,  $L$ . To quantify the accuracy of our routine, we compare its absolute error with that of the non-reproducible summation of conventional floats. According to Demmel and Nguyen [12], the latter error can be bounded by:

$$e_{conv} := (n - 1) \cdot \varepsilon \cdot \sum_{i=1}^n |b_i|, \quad (5)$$

where  $\varepsilon$  represents a machine constant [17] and  $b_i, i = 1..N$ , represent the summands. The error of our routine can be bounded the following expression, which is due to Demmel and Nguyen [13] and the same for their and our algorithm:

$$e_{\text{Rsum SIMD}} := n \cdot 2^{(1-L) \cdot W - 1} \cdot \max_{1 \leq i \leq n} |b_i|, \quad (6)$$

<sup>4</sup>For 128-bit integers, we use the `__int128` type available in recent versions of GCC on our CPU.

<sup>5</sup>The relative standard deviation is mostly below 1% and never above 5%.

	$n = 10^3$		$n = 10^6$	
	U[1, 2)	Exp(1)	U[1, 2)	Exp(1)
Conventional	$1.7 \cdot 10^{-10}$	$1.1 \cdot 10^{-10}$	$1.7 \cdot 10^{-4}$	$1.1 \cdot 10^{-4}$
RSUM ( $L = 1$ )	$1.0 \cdot 10^3$	$1.1 \cdot 10^4$	$1.0 \cdot 10^6$	$1.1 \cdot 10^7$
RSUM ( $L = 2$ )	$9.1 \cdot 10^{-10}$	$1.0 \cdot 10^{-8}$	$9.1 \cdot 10^{-7}$	$1.0 \cdot 10^{-5}$
RSUM ( $L = 3$ )	$8.3 \cdot 10^{-22}$	$9.1 \cdot 10^{-21}$	$8.3 \cdot 10^{-19}$	$9.1 \cdot 10^{-18}$

Table II: Maximum absolute error of conventional and reproducible summation algorithms in double precision.

where  $f$  and  $W$  are the exponent of the first extractor used and the ratio of two consecutive extractors, respectively.

To quantify these expressions in a simple way, we consider the summation of random arrays, with uniformly-distributed values in the range  $[1, 2)$  and exponentially-distributed values with  $\lambda = 1$ , with varying number of values  $n$ . With the latter distribution, the probability of an input set with  $10^6$  values to contain a value larger than 22 is lower than 0.03%, thus we choose 22 as the maximum expected input value and we use it to give a reasonable error bound. Table II shows the expected values of the error bounds for the different algorithms and different distributions in double precision. The error bound for the single-level reproducible summation can be surprisingly large. The reason for such large bounds is the low control on the magnitude of the levels the algorithm has. The largest extractor used for the summation could have a much larger magnitude than the result (up to  $2^{W-1}$  times larger). In this unfortunate case, only one significant bit of the result is kept by the first summation level. If this is the only level used, the uncertainty on the result is as large as the result itself. In many cases, the one-level summation can deliver reasonably accurate results and, for large input sizes, its accuracy can be comparable to the conventional summation. Nevertheless, it gives no guarantee of accurate results beyond the error bounds listed in the table. All error bounds for the reproducible algorithm are up to  $2^{W-1}$  times more pessimistic than the actual error of the summations.

**Conclusion:** Our summation routine RSUM SIMD with  $L = 2$  has comparable accuracy as conventional floating-point summation and achieves much higher accuracy with  $L > 2$ .

2) *Performance:* We also measure the performance of several variants of the reproducible summation algorithm (RSUM) for summing up a large array of random numbers. Figure 6 shows the result. For RSUM SCALAR and RSUM SIMD (Algorithms 2 and 3, respectively), we sum up the input in chunks of  $c$  values for various values of  $c$ , i.e., we call the respective algorithm for each chunk of  $c$  values. This mimics the pattern of how the summation algorithms are used in the aggregation algorithms of Section V, where they switch between the summation of inputs of different groups. For brevity, the algorithms are simply called SCALAR and SIMD in the figure. We plot the performance in terms of slowdown compared to a conventional summation algorithm on the same input, which is called CONV in the figure and implemented as a single call to `std::accumulate`. Finally, as “lower bound”,

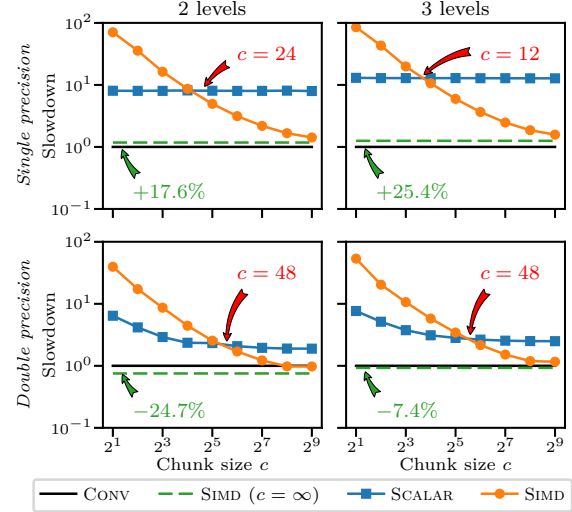


Figure 6: Relative performance of RSUM algorithms compared to a conventional sum using `std::accumulate` (CONV).

we also plot the performance of a single call to RSUM SIMD, which we call SIMD ( $c = \infty$ ).

As the figure shows, RSUM SIMD is slower than RSUM SCALAR for small chunk sizes, but faster for large chunk sizes. For small chunk sizes, it suffers from a higher start-up overheads because the state it loads and stores from memory into registers and back is a factor  $V$  times larger than that of the scalar version. For large chunk sizes, vectorization pays off. The cross-over point (annotated in red) is somewhere between  $c = 12$  and  $c = 48$  depending on the number of levels  $L$  and the precision. As the chunk size reaches  $c = 512$ , the start-up overhead of the multiple calls to RSUM SIMD is amortized and the performance reaches that of SIMD ( $c = \infty$ ), which consists of a single call. At this point, RSUM SIMD is at most 25% slower than CONV and even somewhat faster in case of double precision. We attribute this to the fact that the compiler is not able to fully vectorize `std::accumulate`, whereas RSUM SIMD ( $c = \infty$ ) is optimized to the point that it is memory-bound.

**Conclusion:** Our summation routine is faster, the larger the buffer size  $bsz$ , as constant start-up costs can be amortized better. For  $bsz \geq 2^6$ , SIMD is always better than SCALAR and for  $bsz \geq 2^9$  or earlier, the difference to the maximum throughput is negligible, where the routine is memory-bound.

### C. Aggregation without Summation Buffers

We first present experiments of unmodified state-of-the-art aggregation algorithms, i.e., we do not use aggregation buffers. We compare two categories of data types: (1) The reproducible `repro<ScalarT,L>` types presented in Section IV, which corresponds to making AGGREGATION reproducible with minimal development effort, and (2) `DECIMAL(p)` with various precisions  $p$  (where  $p$  is the number of decimal digits), which may be a good enough alternative for some applications. We emphasize that, as discussed in Section II-C, the `DECIMAL` types are not flexible enough for many modern applications, which cannot

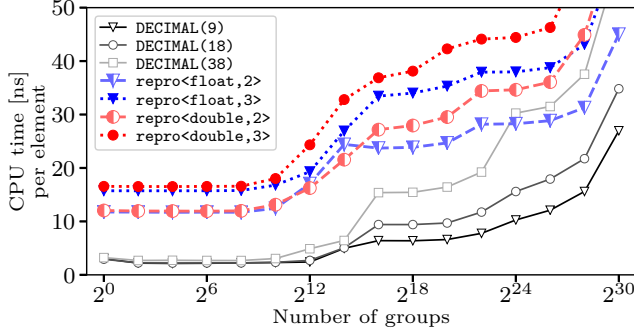


Figure 7: PARTITIONANDAGGREGATE on various  $\text{repro}\langle\text{ScalarT},L\rangle$  without summation buffers compared to the same algorithm on  $\text{float}/\text{DECIMAL}(8)$ .

determine the scale of the involved numbers statically and thus need a *floating*-point representation. We still include them as a reference point. We also ran the algorithms with built-in floating-point types, but observed exactly the same performance as  $\text{DECIMAL}(9)$  and  $\text{DECIMAL}(18)$  for  $\text{float}$  and  $\text{double}$ , respectively, so we omit them in the plots shown below.

Figure 7 summarizes our findings. For better readability, we only show the results for  $L = 2$  and  $L = 3$ , which are the most interesting configurations in practice. As in the experiment shown in Figure 4, the different levels are about equidistant from each other, which gives an idea of the omitted data points.

The plot shows that the runtime for all data types follows the expected pattern: fast, in-cache processing for small numbers of groups and more and more overhead for the partitioning as the number of groups increases, each “step” corresponding to an additional level of partitioning. For  $\text{DECIMAL}(p)$ , the steps are higher the larger  $p$ , which is due to the higher memory traffic for wider data types in that phase. Furthermore, the larger  $L$ , i.e., the higher the accuracy, the longer the runtime of  $\text{repro}\langle\text{ScalarT},L\rangle$  by quite a large difference, slightly more so for doubles than for floats. This represents a slowdown of up to factor<sup>6</sup> 4 to 10 for small numbers of groups compared to built-in  $\text{float}$ s (which has the same performance as  $\text{DECIMAL}(9)$ ).

**Conclusion:** Using reproducible floats as drop-in replacement in unmodified state-of-the-art aggregation algorithms has an overhead of up to factor 6 compared to built-in floats with comparable accuracy.

#### D. Aggregation with Summation Buffers

We now turn our attention to PARTITIONANDAGGREGATE with summation buffers.

We first confirm Equation 4, which we define to choose the buffer size  $bsz$ . To that aim, we compare the value predicted by the equation with the buffer size performing best found through exhaustive search. We find that 75 % of all combinations of number of groups and data types deviate by less than 1 % from the optimal performance, 90 % of them deviate by less than 5 %, and the largest deviation is 20 %. We can thus predict a

<sup>6</sup>For the omitted configuration of  $L = 4$ , the slowdown is even up to 12 ×.

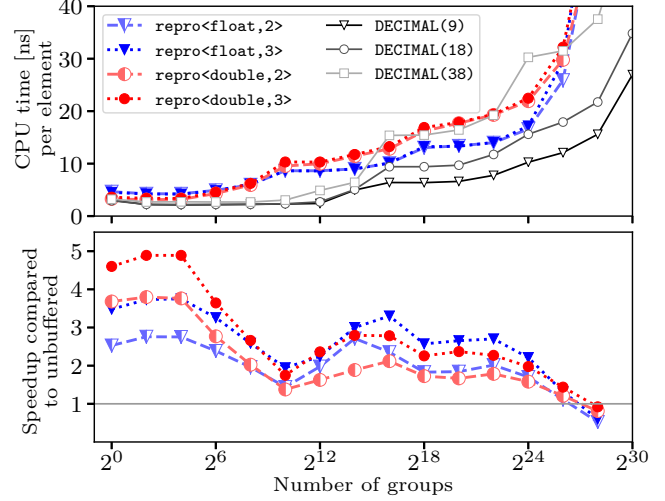


Figure 8: PARTITIONANDAGGREGATE with summation buffers on various  $\text{repro}\langle\text{ScalarT},L\rangle$  compared to the same algorithm on unbuffered  $\text{DECIMAL}$ .

close-to-optimal buffer size using Equation 4, which we use for the remainder of this paper.

Figure 8 shows the performance of PARTITIONANDAGGREGATE using  $\text{repro}\langle\text{ScalarT},L\rangle$  and summation buffers in comparison with unbuffered  $\text{DECIMAL}$  types. The upper diagram shows the absolute running time, which exhibits the same pattern of increasing cost due to more levels of partitioning for increasing numbers of groups. Compared to the algorithm without summation buffers, however, the running time is generally lower and, in particular, there is now little difference between different configurations of  $\text{repro}\langle\text{ScalarT},L\rangle$ . The largest difference is caused by the fact that the reproducible data types based on  $\text{double}$  are slower than those based on  $\text{float}$ . This is mainly due to the fact that PARTITIONING, which is memory bound, needs to move twice as much data in the former case. This is the same effect that slows down the  $\text{DECIMAL}$  types, which makes them about as slow or slower as our reproducible types for  $2^{16}$  groups and more (in addition to being less flexible).

Most importantly, we can see that the summation buffers have significantly narrowed the gap between our new reproducible data types and the built-in floating-point types. In many cases, this slowdown is in the range of 1.3 to 2.5. Only for almost distinct data (more than  $2^{26}$  groups) and for the range of  $2^8$  to  $2^{12}$  groups, the slowdown may be larger than that. In the latter range, the number of groups are such that the algorithm on built-in floats can still fit its working set into the last-level cache, while the algorithm using summation buffers cannot, and thus needs to pay the additional price of partitioning the input first (on top of the overhead of buffering and more expensive summation). Overall, the slowdown is still reasonable: its geometric mean of all numbers of groups ranges from 1.87 to 2.35 for types based on  $\text{float}$  and from 2.12 to 2.41 for types based on  $\text{double}$ . We believe that this is an affordable price for full reproducibility.

	double	repro<d,4> without buffer	repro<d,4> with buffer	double (sorted)
Aggregations	34.2	51.3	38.7	45.1
Other	65.8	63.1	64.0	682.1
Total	100.0	114.4	102.7	727.2

Table III: CPU time of different approaches for TPC-H Query 1 relative to the total CPU time on built-in doubles in %.

Finally, the lower diagram shows the speedup of our algorithm with summation buffers compared to the naïve approach without them. In particular for small data sets, the speedup is considerable (between factor 2 and more than 5 for the shown configurations and up to factor 6 for the omitted  $L = 4$ ). As expected, it is the higher, the larger  $L$ . The speedup drops slightly below 1 for the largest number of groups, i.e., using summation buffers is actually slower than not using them. Since the difference is not large, we leave this as a small open problem.

**Conclusion:** Thanks to efficient partitioning routines, careful cache-management, and vectorized summation on summation buffers, the overhead of reproducibility on floating-point numbers can be reduced to a slowdown of about a factor two.

#### E. End-to-End Query Performance

We integrated our reproducible data types into MonetDB [8] v11.25.23 in order to quantify their impact on end-to-end query performance. To that aim, we modified MonetDB’s aggregation operator for sum on built-in doubles such that it first aggregates its input into a locally allocated array using our reproducible data types (with or without summation buffers) and then copies the result converted to doubles into the result array allocated by the system.<sup>7</sup> We run a modified TPC-H benchmark as workload where we replaced all DECIMAL columns by DOUBLE.

Table III shows the CPU time of different approaches on Query 1 relative to the CPU time of an unmodified MonetDB. As additional baseline, we include the CPU time of modified queries that sort the input to the grouping and aggregation operators, which is the only way to make them reproducible across input permutations without modifying the system.

As the table shows, using repro<double,4> without summation buffers takes about 14% longer due to a 50% CPU time increase of the aggregation operators. This increase is lower than the  $10\times$  increase observed with our own aggregation operator in Section VI-C due to the slower baseline of MonetDB’s operator, which performs several overflow checks for each input element. With summation buffers, however, the overhead of reproducibility is a negligible 2.7%. Sorting, in contrast, is more than  $7\times$  slower, which shows the importance of a numeric solution such as the one we propose.

<sup>7</sup>This does not technically make MonetDB reproducible because it parallelizes query plans as independent subplans on parts of the input whose intermediate results are merged. Our changes make the aggregation operators of each subplan as well as the merging reproducible, but the splitting of the input remains non-deterministic. We argue that this still gives a good approximation of the performance impact. A full integration would require the introduction of a new type, which is a development effort out of the scope of this paper.

## VII. RELATED WORK

Aggregation with GROUPBY on conventional data types is a well understood problem. In recent years, it has been studied extensively for in-memory, multi-core database systems [10, 20, 29, 30]. The focus of that work was contention-free parallelization and cache efficiency. Which strategy is best to achieve the latter goal mainly depends on the size of the result, which directly depends on the number of groups. The consensus [10, 20] is that algorithms similar to PARTITIONANDAGGREGATE with various levels of partitioning are best for different numbers of groups, which is why we build on them in this paper. For the case where the result is larger than a private cache, but smaller than the combined shared cache of all threads, Cieslewicz and Ross [10] show that SHAREDAGGREGATION may be a better solution than the other two, which uses a shared (lock-free) hash table, at least in the absence of skew. Similar techniques have been proposed for JOIN and SORT operators [4, 5, 6, 8]. As we show in this paper, these techniques alone are not sufficient for reproducible floating-point numbers. Variants of SORTAGGREGATION, have not been found competitive by recent studies [4] (except for presorted inputs [10]), mainly due to much higher computational costs and the difficulty to combine early aggregation with vectorization.

There is a line of research in High-Performance Computing that studies the problem of reproducibility of numerical computations using floating-point numbers [3, 12, 13]. However, as argued throughout the paper, the proposed solutions are not applicable to aggregation with GROUPBY. No work on numeric reproducibility in the field of data processing is known to us.

Our work may seem to contradict attempts to speed up query execution either by approximating the computation or reducing the precision. Prominent examples of the first class include DBO [18], BlinkDB [2], and Sample+Seek [14], as well as recent summarization techniques based on the principle of Maximum Entropy [24]. We consider this work somewhat orthogonal to ours since many sampling techniques are based on deterministic pseudo-random number generators and could require a technique similar to ours to be completely reproducible. Maybe more importantly, from a user’s perspective, it is clearly less surprising to get different results from a system that gives approximate answers by design rather than from one that is assumed to be deterministic. Examples of the second class include work on neural networks with 8-bit and even 1-bit precision by Suda et al. [27] and Courbariaux et al. [11], respectively, as well as the low-precision machine learning framework ZipML [31]. However, as discussed in Section II-C, precision and reproducibility are completely orthogonal (our algorithms could be implemented based on lower-precision floating-point types as well).

## VIII. SUMMARY AND CONCLUSION

In this paper we have addressed the problem of bit-reproducible aggregation in database systems. The main challenge is that achieving reproducibility is expensive and cannot be efficiently done with existing algorithms for reproducible

summation and for GROUPBY. Any naïve combination of existing results in the two areas leads to prohibitive overheads.

The main insights from the work include identifying the bottlenecks that result from the bookkeeping needed to keep track of rounding errors and the effects that it has in cache locality for high cardinality aggregation. Based on these insights, we have proposed ways to extend existing aggregation operators with bit-reproducibility in a way that the resulting overhead is acceptable and comparable to that of conventional aggregation over built-in types.

With these results, we establish the basis for exploring more complex data types and operations inside the database engine providing the same guarantees and meeting the same requirements as those imposed on regular code processing floating-point data. As part of future work we intend to look into operators for machine learning, vector manipulation, and series analysis based on the algorithms presented in this paper.

#### ACKNOWLEDGMENTS

We thank Eric Sedlar (Oracle Labs) for bringing the problem of reproducibility in the context of database systems to our attention, as well as for valuable feedback on this work. We also thank Lefteris Sidirourgos for his feedback on the integration of our algorithm into MonetDB.

#### REFERENCES

- [1] ACM US Public Policy Council. *Statement on Algorithmic Transparency and Accountability*. 2017.
- [2] S. Agarwal et al. “BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data.” In: *EuroSys*. 2013.
- [3] A. Arteaga, O. Fuhrer, and T. Hoefler. “Designing Bit-Reproducible Portable High-Performance Applications.” In: *IPDPS* (2014).
- [4] Ç. Balkesen. “In-Memory Parallel Join Processing on Multi-Core Processors.” PhD thesis. ETH Zurich, 2014.
- [5] C. Balkesen, J. Teubner, and G. Alonso. “Main-Memory Hash Joins on Multi-Core CPUs : Tuning to the Underlying Hardware.” In: *ICDE*. 2013.
- [6] C. Balkesen et al. “Multi-Core, Main-Memory Joins: Sort vs. Hash Revisited.” In: *PVLDB*. 2013.
- [7] G. E. Blelloch et al. “Internally Deterministic Parallel Algorithms Can Be Fast.” In: *PPoPP*. Vol. 47. 8. 2012.
- [8] P. A. Boncz, M. L. Kersten, and S. Manegold. “Breaking the Memory Wall in MonetDB.” In: *CACM* 51.12 (2008).
- [9] W.-f. Chiang and G. L. Lee. “Determinism and Reproducibility in Large-Scale HPC Systems.” In: *WoDet* (2013).
- [10] J. Cieslewicz and K. Ross. “Adaptive Aggregation on Chip Multiprocessors.” In: *PVLDB*. 2007.
- [11] M. Courbariaux et al. “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1.” In: *CoRR* abs/1602.02830 (2016).
- [12] J. Demmel and H. D. Nguyen. “Fast Reproducible Floating-Point Summation.” In: *ARITH*. 2013.
- [13] J. Demmel and H. D. Nguyen. “Parallel Reproducible Summation.” In: *IEEE Trans. Comput.* 64.7 (2015).
- [14] B. Ding et al. “Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee.” In: *SIGMOD*. 2016.
- [15] European Parliament and European Council. *General Data Protection Regulation*. 2016. URL: <http://data.europa.eu/eli/reg/2016/679/oj>.
- [16] Federal Trade Commission. *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (FTC Report)*. Tech. rep. 2016.
- [17] D. Goldberg. “What Every Computer Scientist Should Know About Floating-Point Arithmetic.” In: *ACM CSUR* 23.1 (1991).
- [18] C. Jermaine et al. “Scalable Approximate Query Processing With The DBO Engine.” In: *TODS* 33.4 (2008).
- [19] I. Müller. “Engineering Aggregation Operators for Relational In-Memory Database Systems.” PhD thesis. Karlsruhe Institute of Technology, 2016.
- [20] I. Müller et al. “Cache-Efficient Aggregation: Hashing Is Sorting.” In: *SIGMOD*. 2015.
- [21] NVIDIA Corporation. *CUDA Toolkit v8.0: cuBLAS*. 2016. URL: [http://docs.nvidia.com/cuda/cublas/#cublasApi\\_reproducibility](http://docs.nvidia.com/cuda/cublas/#cublasApi_reproducibility).
- [22] T. Ogita, S. M. Rump, and S. Oishi. “Accurate Sum and Dot Product with Applications.” In: *ICRA* 26.6 (2004).
- [23] Oracle. *Java Platform, Standard Edition 8: API Specification*. 2016. URL: <https://docs.oracle.com/javase/8/docs/api/java/math/BigDecimal.html>.
- [24] L. Orr, M. Balazinska, and D. Suciu. “Probabilistic Database Summarization for Interactive Data Exploration.” In: *PVLDB*. Vol. 10. 10. 2017.
- [25] T. Rosenquist and S. Story. “Using the Intel Math Kernel Library (Intel MKL) and Intel Compilers to Obtain Run-to-Run Numerical Reproducible Results.” In: *The Parallel Universe* 11 (2012).
- [26] F. M. Schuhknecht, P. Khanchandani, and J. Dittrich. “On the Surprising Difficulty of Simple Things: the Case of Radix Partitioning.” In: *PVLDB*. Vol. 8. 9. 2015.
- [27] N. Suda et al. “Throughput-Optimized OpenCL-based FPGA Accelerator for Large-Scale Convolutional Neural Networks.” In: *FPGA*. 2016.
- [28] *The GNU MPFR Library*. URL: <http://www.mpfr.org/>.
- [29] J. Wen. “Revisiting Aggregation Techniques for Data Intensive Applications.” PhD thesis. University of California, Riverside, 2013. ISBN: 978-1-303-71220-3.
- [30] Y. Ye, K. A. Ross, and N. Vesdapunt. “Scalable Aggregation on Multicore Processors.” In: *DaMoN*. 2011.
- [31] H. Zhang et al. “The ZipML Framework for Training Models with End-to-End Low Precision: The Cans, the CANNOTs, and a Little Bit of Deep Learning.” In: *ICML*. 2016.
- [32] D. Zuras et al. “IEEE standard for floating-point arithmetic.” In: *IEEE Std 754-2008* (2008).