



Partial Sums on the Ultra-Wide Word RAM

Philip Bille^(✉), Inge Li Gørtz, and Frederik Rye Skjoldjensen

DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
{phbi, inge}@dtu.dk

Abstract. We consider the classic partial sums problem on the ultra-wide word RAM model of computation. This model extends the classic w -bit word RAM model with special ultrawords of length w^2 bits that support standard arithmetic and boolean operation and scattered memory access operations that can access w (non-contiguous) locations in memory. The ultra-wide word RAM model captures (and idealizes) modern vector processor architectures.

Our main result is a new in-place data structure for the partial sum problem that only stores a constant number of ultrawords in addition to the input and supports operations in doubly logarithmic time. This matches the best known time bounds for the problem (among polynomial space data structures) while improving the space from superlinear to a constant number of ultrawords. Our results are based on a simple and elegant in-place word RAM data structure, known as the Fenwick tree. Our main technical contribution is a new efficient parallel ultra-wide word RAM implementation of the Fenwick tree, which is likely of independent interest.

Keywords: Ultra-wide word RAM model · Partial sums · Fenwick tree

1 Introduction

Let $A[1, \dots, n]$ be an array of integers of length n . The *partial sums problem* is to maintain a data structure for A under the following operations:

- **sum**(i): return $\sum_{k=1}^i A[k]$.
- **update**(i, Δ): set $A[i] \leftarrow A[i] + \Delta$.

The partial sums problem is a classic and well-studied data structure problem [1–4, 9, 12, 14, 16–19, 21–24, 31, 32, 38]. Partial sums is a natural range query problem with applications in areas such as list indexing and dynamic ranking [12], dynamic arrays [3, 32], and arithmetic coding [14, 34]. From a lower bound perspective, the problem has been central in the development of new techniques for proving lower bounds [29]. In classic models of computation the complexity of the partial sums problem is well-understood with tight logarithmic upper and lower bounds on the operations [31]. Hence, a natural question is if

practical models of computation capturing modern hardware advances will allow us to overcome the logarithmic barrier.

One such model is the *RAM with byte overlap* (RAMBO) model of computation [6, 7, 17]. The RAMBO model extends the standard w -bit word RAM model [20] with special words where individual bits are shared among other words, i.e., changing a bit in a word will also change the bit in the words that share that bit. The precise model depends on the layout of shared bits. This memory architecture is feasible to design in hardware and prototypes have been built [27]. In the RAMBO model Brodnik et al. [8] gave a time-space trade-off for partial sums that uses $O(n^{w/2^\tau} + n)$ space and supports operations in $O(\tau)$ time and for a parameter τ , $1 \leq \tau \leq \log \log n$. Here, the n term in the space bound is for the special words with shared bits (organized in a tree layout) and the $O(n^{w/2^\tau})$ term is for standard words. Plugging in constant τ , this gives an $O(n^{\epsilon w} + n)$ space and constant time solution, for any $\epsilon > 0$. At the other extreme, with $\tau = \log \log n$, this gives an $O(n)$ space and $O(\log \log n)$ time solution.

More recently, Farzan et al. [13] introduced the *ultra-wide word RAM* (UWRAM) model of computation. The UWRAM model also extends the word RAM model, but with special *ultrawords* of length w^2 bits. The model supports standard arithmetic and boolean operations on ultrawords and *scattered* memory access operations that access w locations in memory specified by an ultraword in parallel. The UWRAM captures modern vector processor architectures [11, 28, 33, 36]. We present the details of the UWRAM model in Sect. 2. Farzan et al. [13] showed how to simulate algorithms on RAMBO model on the UWRAM model at the cost of slightly increasing space. Simulating the above solution for partial sums they gave a time-space trade-off for partial sums that uses $O(n^{w/2^\tau} + nw \log n)$ space and supports operations in $O(\tau)$ time and for a parameter τ , $1 \leq \tau \leq \log \log n$. For constant τ , this is $O(n^{\epsilon w} + nw \log n)$ space and constant time, for any $\epsilon > 0$, and for $\tau = \log \log n$ this is $O(nw \log \log n)$ space and $O(\log \log n)$ time.

1.1 Setup and Results

We revisit the partial sums problem on the UWRAM and present a simple new algorithm that significantly improves the space overhead of the previous solutions. Let A be an array of n w -bit integers. An *in-place data structure* for the partial sums problem is a data structure that modifies the input array A , e.g., by replacing some of the entries in A , to efficiently support operations. In addition to the modified array the data structure is only allowed to store $O(1)$ ultrawords. This definition extends the standard in-place/implicit data structure concept [10, 15, 30, 35, 37] to the UWRAM, by allowing a constant number of ultrawords to be stored instead of (standard) words. Clearly, without this modification computation on ultrawords is impossible. As in Farzan et al. [13] we distinguish between the *restricted UWRAM* that supports a minimal set of instructions on ultrawords consisting of addition, subtraction, shifts, and bitwise boolean operations and the *multiplication UWRAM* that extends the instruction

set of the restricted UWRAM with a multiplication operation on ultrawords. We show the following main result:

Theorem 1. *Given an array A of n w -bit integers, we can construct in-place partial sums data structures for A that support sum and update operations in $O(\log \log n)$ time on a restricted UWRAM.*

Compared to the previous result, Theorem 1 matches the $O(\log \log n)$ time bound of Farzan et al. [13] (with parameter $\tau = \Theta(\log \log n)$) while improving the space overhead from $O(nw \log n)$ to a constant number of ultrawords. This is important in practical applications since modern vector processors have a very limited number of ultrawords available.

Technically, our solution is based on a simple and elegant in-place word RAM data structure, called the *Fenwick tree* (see Sect. 3 for a detailed description). The Fenwick tree support operations in $O(\log n)$ by sequentially traversing an implicit tree structure. We show how to efficiently compute the access pattern on the tree structure in parallel using prefix sum computations on ultrawords. Then, given the locations to access we use scattered memory operations to access them all in parallel. In total, this leads to the exponential improvement of Fenwick trees. The main bottleneck in our algorithm is the prefix sum computation. Interestingly, if we allow multiplication we can compute prefix sums in constant time leading to the following Corollary for the multiplication UWRAM:

Corollary 1. *Given an array A of n w -bit integers, we can construct in-place partial sums data structures for A that support sum and update operations in constant time on a multiplication UWRAM.*

Multiplication (or prefix sum computation) is not an AC^0 operation (it cannot be implemented by a constant depth, polynomial size circuit) and therefore likely not practical to implement on ultraword. However, Corollary 1 shows that we can achieve significant improvements on the UWRAM with special operations. Since UWRAM capture modern processors, we believe it is worth investigating further, and that our work is a first step in this direction.

1.2 Outline

The paper is organized as follows. In Sect. 2 and 3 we review the UWRAM model of computation and the Fenwick tree. In Sect. 4 we present our UWRAM implementation of the Fenwick tree. Finally, in Sect. 4.4 we discuss extensions of the result and open problems.

2 The Ultra-Wide Word RAM Model

The *word RAM* model of computation [20] consists of an infinite memory of w -bit words and an instruction set of arithmetic, boolean, and memory access instructions such as the ones available in standard programming languages such

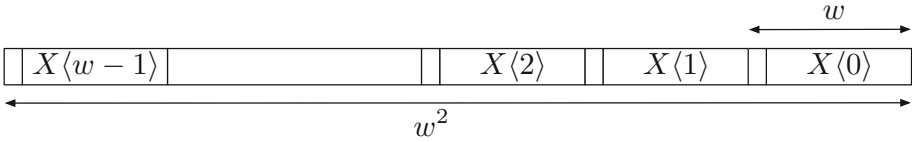


Fig. 1. The layout of an ultraword of w^2 divided into w words each of w bits. The leftmost bit of each word is reserved to be a test bit.

as C . We assume that we can store a pointer into the input in a single word and hence $w \geq \log n$, where n is the size of the input. The time complexity of a word RAM algorithm is the number of instructions and the space complexity is the number of words used by the algorithm.

The *ultra-wide word RAM* (UWRAM) model of computation [13] extends the word RAM model with special ultrawords of w^2 bits. We distinguish between the *restricted UWRAM* that supports a minimal set of instructions on ultrawords consisting of addition, subtraction, shifts, and bitwise boolean operations and the *multiplication UWRAM* that additionally supports multiplication. The time complexity is the number of instruction (on standard words or ultrawords) and the space complexity is the number of (standard) words used by the algorithm. The restricted UWRAM captures modern vector processor architectures [11, 28, 33, 36]. For instance, the Intel AVX-512 vector extension [33] support similar operations on 512-bit wide words (i.e., a factor of 8 compared to $64^2 = 4096$).

2.1 Word-Level Parallelism

Due to their similarities, we can adopt many word-level parallelism techniques from the word RAM to the UWRAM. We briefly review the key primitives and techniques that we will use.

Let X be an ultraword of w^2 bits. We often view X as divided into w words of w consecutive bits each. See Fig. 1. We number the words in X from right-to-left starting from 0 and use the notation $X\langle j \rangle$ to denote the j th word in X . Similarly, the bits of each word $X\langle j \rangle$ are numbered from right-to-left starting from 0. If only the rightmost $\ell \leq w$ words in X are non-zero, we say that X has *length* ℓ . For simplicity in the presentation, we reserve the leftmost bit of each word to be a *test bit* for word-level parallelism operations. One may always remove this assumption at no asymptotic cost, e.g., by using two words in an ultraword to simulate each single word.

We now show how to implement common operations on ultrawords that we will use later. Most of these are already available in hardware on modern vector processor architectures. Componentwise arithmetic and bitwise operation are straightforward to implement using standard word-level parallelism techniques from the word RAM. For instance, given ultrawords X and Y , we can compute the componentwise addition, i.e., the ultraword Z such that $Z\langle j \rangle = X\langle j \rangle + Y\langle j \rangle$ for $j = 0, \dots, w - 1$ by adding X and Y and $\&$ 'ing with the mask $(01^{w-1})^w$ to

clear any test bits (we use exponentiation to denote bit repetition, i.e., $0^3 1 = 0001$). We can also compare X and Y componentwise by \mid 'ing in the test bits of X , subtracting Y , and masking out the test bits by $\&$ 'ing with $(10^{w-1})^w$. The j th test bit of the result contains a 1 iff $X\langle j \rangle \geq Y\langle j \rangle$. Given X and another ultraword T containing only test bits, we can extract the words in X according to the test bits, i.e., the ultraword E such that $E\langle j \rangle = X\langle j \rangle$ if the j th test bit of T is 1 and $E\langle j \rangle = 0$ otherwise. To do so we copy the test bits by a subtracting $(0^{w-1} 1)^w$ from T and $\&$ 'ing the result with X . All of the above mentioned operation take constant time on a restricted UWRAM. Given an ultraword X of length ℓ , the *prefix sum* of X is the ultraword P of length ℓ , such that $P\langle j \rangle = \sum_{k \leq j} X\langle k \rangle$. We assume here that the integers computed in the prefix sum never exceed the maximum size available in a word such that $P\langle j \rangle$ is always well-defined. We need the following result.

Lemma 1. *Given an ultraword X of length ℓ we can compute the prefix sum of X in $O(\log \ell)$ time on a restricted UWRAM and in $O(1)$ time on a multiplication UWRAM.*

Proof. First consider the restricted UWRAM. We implement a standard parallel prefix-sum algorithm [25] (see also the survey by Blelloch [5]). For simplicity, we assume that ℓ is a power of two. The algorithm consists of two phases that conceptually construct and traverse a perfectly balanced binary tree T of height $\log \ell$ whose leaves are the ℓ words of X .

Given an internal node v in T , let v_{left} and v_{right} denote the left and right child of v , respectively. The first phase performs a bottom-up traversal of T and computes for each node v an integer $b(v)$. If v is a leaf, $b(v)$ is the corresponding integer in X and if v is an internal node $b(v) = b(v_{\text{left}}) + b(v_{\text{right}})$. The second phase performs a top-down traversal of T and computes an integer $t(v)$. If v is the root then $t(v) = 0$ and if v is an internal node then $t(v_{\text{left}}) = t(v)$ and $t(v_{\text{right}}) = t(v_{\text{left}}) + b(v_{\text{right}})$. After the second phase the integers at the leaves is the prefix sum shifted by a single element and missing the last element. We shift and add the last element to produce the final prefix sum. Since T is perfectly balanced we can implement each level of a phase in constant time using shifting and addition. The final shift and addition of the last element takes constant time. It follow that the total time is $O(\log \ell)$. During the computation we only need to maintain all of the values in a constant number of ultrawords.

Next consider the multiplication instruction set. We can then simply multiply X with the constant $(0^{w-1} 1)^w$ and mask out the ℓ rightmost words of the result to produce the prefix sum. See Hagerup [20] for a detailed description of why this is correct. In total this uses $O(1)$ time.

2.2 Memory Access

The UWRAM supports standard memory access operation to read or write a single word or a sequence of w contiguous words. More interestingly, the UWRAM also supports *scattered* access operations that access w memory locations (not

necessarily contiguous) in parallel. Given an ultraword A containing w memory addresses, a *scattered read* loads the contents of the addresses into an ultraword X , such that $X\langle j \rangle$ contains the contents of memory location $A\langle j \rangle$. Given two ultrawords A and X *scattered write* sets the contents memory location $A\langle j \rangle$ to be $X\langle j \rangle$. Scattered memory accesses captures the memory model used by IBM’s *Cell* architecture [11]. Scattered memory access operations were also proposed by Larsen and Pagh [26] in the context of the I/O model of computation.

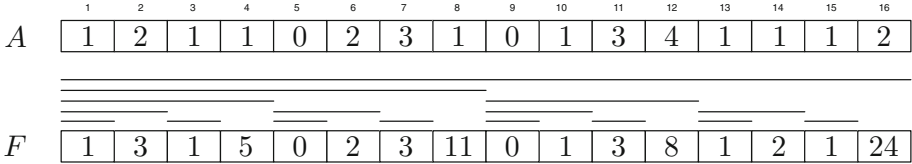


Fig. 2. A array A and the Fenwick tree F . The lines above F indicate the partial sum of A stored at the rightmost endpoint of the line. For instance, the $F[12] = A[9] + A[10] + A[11] + A[12] = 0 + 1 + 3 + 4 = 8$.

3 Fenwick Trees

Let A be an array of n w -bit integers and assume for simplicity that n is a power of two. The Fenwick tree [14, 34] is an in-place data structure that replaces the array A as follows. If $n = 1$, then leave A unchanged. Otherwise, replace all values at even entries $A[2i]$ by the sum $A[2i - 1] + A[2i]$. Then, recurse on the subarray $A[2, 4, \dots, n]$. The resulting array F stores a subset of the partial sums of A organized in a tree layout (see Fig. 1).

To answer $\text{sum}(i)$ query, we compute a sequence of indices in F and add the values in F at these indices together. Let $\text{rmb}(x)$ denote the position of the rightmost bit in an integer x . Define the *sum sequence* i_1^s, \dots, i_r^s given by $i_1^s = i$ and $i_j^s = i_{j-1}^s - 2^{\text{rmb}(i_{j-1}^s)}$, for $j = 2, \dots, r$. The final element i_r^s is 0. We compute and return $F[i_1^s] + F[i_2^s] + \dots + F[i_{r-1}^s]$. For instance, for $i = 13 = (1101)_2$ the sum sequence is 13, 12, 8, 0 = $(1101)_2, (1100)_2, (1000)_2, (0000)_2$. Hence, $\text{sum}(13) = F[13] + F[12] + F[8] = 1 + 8 + 11 = 20 = A[1] + \dots + A[13]$. We access at most $O(\log n)$ entries in F and hence the total time for sum is $O(\log n)$. Note that we can always recover the original array A using the sum operation, since $A[i] = \text{sum}(i) - \text{sum}(i - 1)$.

To compute $\text{update}(i, \Delta)$, we compute a sequence of indices in F and add Δ to the values in F at each of these indices. Define the *update sequence* i_1^u, \dots, i_t^u given by $i_1^u = i$ and $i_j^u = i_{j-1}^u + 2^{\text{rmb}(i_{j-1}^u)}$, for $j = 2, \dots, t$. The final element i_t^u is $2n$. We set $F[i_1^u] = F[i_1^u] + \Delta, \dots, F[i_t^u] = F[i_{t-1}^u] + \Delta$. For instance, for $i = 13$ the update sequence is 13, 14, 16, 32. Hence, $\text{update}(13, 5)$ adds 5 to $F[13]$, $F[14]$, and $F[16]$. Similar to the sum operation, the total running time for update is $O(\log n)$.

4 Partial Sums on the Ultra-Wide Word RAM

We now present an efficient implementation of Fenwick trees on the UWRAM model of computation. We only store the Fenwick tree, as the array F described in Sect. 3 and a constant number of ultraword constants that we use for computation. We first show some basic properties of the sum and update sequences in Sect. 4.1, before presenting our UWRAM implementation of the operations in Sects. 4.2 and 4.3.

4.1 Computing Sum and Update Sequences

To compute the sum and update sequences we cannot directly apply the recursive definitions, since this would need $\Omega(\log n)$ steps. Instead, we show how to express the sequences as a prefix sum that we can efficiently derive from the input integer i . Then, using Lemma 1 we will show how to compute it in on the UWRAM in the following sections.

Let i_1^s, \dots, i_r^s and i_1^u, \dots, i_t^u be the sum sequence and update sequences, respectively, for i as defined in Sect. 3. Define the *offset sum sequence* o_1^s, \dots, o_{r-1}^s and *offset update sequence* o_1^u, \dots, o_{t-1}^u for i to be the sequences of differences of the sum and update sequences, respectively, that is, $o_j^s = i_{j+1}^s - i_j^s$, for $j = 1, \dots, r - 1$ and $o_j^u = i_{j+1}^u - i_j^u$, for $j = 1, \dots, t - 1$. By definition, we have that

$$i_j^s = i + \left(\sum_{k < j} o_k^s \right) \quad i_j^u = i + \left(\sum_{k < j} o_k^u \right) \quad (1)$$

We also have that $o_j^s = -2^{\text{rmb}(i_j^s)}$ and hence each sum offset is a power of 2 corresponding to the rightmost 1 bit in i_j^s . Thus, o_1^s corresponds to the rightmost 1 in $i_1^s = i$. Adding $o_1^s = -2^{\text{rmb}(i)}$ (i.e., subtracting $2^{\text{rmb}(i)}$) “clears” the rightmost 1 bit in i . Thus, o_2^s corresponds to the 1 bit in i immediately to left of the rightmost 1 bit. In general, we have that $o_j^s = -2^b$, where b is the position of the j th rightmost bit in i , for $j = 1, \dots, r - 1$. For instance, for $i = 13 = (1101)_2$ the offset sum sequence is $-1, -4, -8$ corresponding to the three 1 bits in the binary representation of i .

Similarly, for the update offsets, we have that $o_j^u = 2^{\text{rmb}(i_j^u)}$. Hence, o_1^u corresponds to rightmost 1 in i . Adding $o_1^u = 2^{\text{rmb}(i_1^u)}$ clears the rightmost consecutive group of 1 bits in i and flips the following 0 bit to 1. In general, we have that $o_j^u = 2^b$, where b is the position of the j th rightmost 0 to the left of $\text{rmb}(i)$, for $j = 2, \dots, t - 1$. For instance, for $i = 13 = (01101)_2$ the offset update sequence is $1, 2, 16$.

4.2 Sum

To compute the $\text{sum}(i)$, the main idea is to first construct the sum sequence in an ultraword, then use a scattered read to retrieve the entries from F in parallel into another ultraword, and finally sum the entries of this ultraword to compute

the final result. We do this in 3 steps as follows. See Fig. 3 for an example of the computed ultrawords during the algorithm.

I	13	13	13	13
M	8	4	2	1
O	8	4		1
P	13	5		1
P'	5		1	
S	8		12	13

Fig. 3. Computing the sum sequence for $i = 13 = (1011)_2$. Words with 0 are left blank. I contains duplicates of i . M is a precomputed mask. O is the bitwise & of I and M . P is the prefix sum of the non-zero words in O . P' is P shifted left by one word. S is the sum sequence obtained by componentwise subtraction of P' from I .

Step 1: Compute Offsets Compute the ultraword O such that $O\langle j \rangle = 2^j$ if -2^j is an offset for i and 0 otherwise, i.e., the non-zero entries of O is the offset sequence for i . To do so we first construct the ultraword I consisting of $\log n$ duplicates of i , i.e., $I\langle j \rangle = i$ for $j = 1, \dots, \log n$. We then compute the bitwise & of I and a mask M , such that $M\langle j \rangle = 2^j$ for $j = 1, \dots, \log n$, i.e., bit j of $M\langle j \rangle = 1$ and the other bits of $M\langle j \rangle$ are 0. By the discussion in Sect. 4.1 the resulting ultraword is O .

On the multiplication UWRAM we can construct I in constant time by multiplying i with $(0^{w-1}1)^w$. On the restricted UWRAM we can construct I in $O(\log \log n)$ time by repeatedly doubling using shifts and bitwise $|$. The rest of the computation takes constant time in both models.

Step 2: Compute Sum Sequence Compute an ultraword S of length $\log n$ whose non-zero entries is the sum sequence i_1^s, \dots, i_{r-1}^s . To do so we first compute the prefix sum P of the non-zero words of O , i.e., we compute the prefix sum of O and then extract the words corresponding to non-zero words in O . Then we shift P by 1 word to the left to produce an ultraword P' and finally subtract P' from I to produce an ultraword S . By (1) the non-zero words in S is the sum sequence for i .

By Lemma 1 the prefix sum computation takes constant time on a multiplication UWRAM and $O(\log \log n)$ time on a restricted UWRAM. The remaining steps take constant time.

Step 3: Compute Sum Finally, we compute $F[i_1^s] + F[i_2^s] + \dots + F[i_{r-1}^s]$. To do so we do a scattered read on S to retrieve $F[i_1^u], \dots, F[i_{s-1}^u]$ into a single

ultraword F' and compute a prefix sum on F' . The sum is then the last word in the result. The scattered read takes constant time. The prefix sum computation takes constant time on a multiplication UWRAM and $O(\log \log n)$ time on a restricted UWRAM. We assume here that $F[0] = 0$. If not we may simply temporarily set $F[0] = 0$ during the computation. Also note that it suffices to perform the first phase of the prefix sum computation as discussed in the proof of Lemma 1 since we only need the sum of all of the retrieved entries.

In total, the `sum` operation takes constant time on a multiplication UWRAM and $O(\log \log n)$ time on a restricted UWRAM.

4.3 Update

We compute `update(i, Δ)` similar to our algorithm for `sum`. We describe how to modify each step of `sum`.

In step 1, we modify the computation of the ultraword O such that it now contains the update offsets, that is, $O\langle j \rangle = 2^j$ if 2^j is an update offset for i and 0 otherwise. To do so we now construct a mask M such that $M\langle j \rangle$ contains a 0 in bit j if j is to the left of `rmb(i)` and 1 elsewhere. We then compute a bitwise \mid of M and I and negate the result. Finally, we set word `rmb(i)` of the result to be $2^{\text{rmb}(i)}$. By the discussion in Sect. 4.1 the resulting ultraword is O .

In step 2, since O now contains the offsets and not the negative offsets, we change the final subtraction to an addition to produce the update sequence stored in a single ultraword U .

In step 3, we do a scattered read on U to retrieve $F[i_1^u], \dots, F[i_{s-1}^u]$ into a single ultraword F' . We then duplicate Δ to all words in an ultraword D and add D to F' to produce an ultraword F'' . Finally, we do a scattered write on U and F'' to update F .

The changes are straightforward to implement in the same time as above. Hence, the `update` operation takes constant time on a multiplication UWRAM and $O(\log \log n)$ time on a restricted UWRAM.

In summary, we use $O(\log \log n)$ time on a restricted UWRAM and $O(1)$ time on a multiplication UWRAM for both operation. We only store the Fenwick tree in the array F and a constant number of ultrawords. This completes the proof of Theorem 1 and Corollary 1.

4.4 Extensions and Open Problems

We sometimes also consider the following operations in the context of partial sums:

- `access(i)`: return $A[i]$.
- `select(j)`: return the smallest i such that $\text{sum}(i) \geq j$

As mentioned `access` is trivial to support since $A(i) = \text{sum}(i) - \text{sum}(i - 1)$. In contrast, the `select` operation do not seem to easily lend itself to an efficient parallel implementation on the UWRAM. While it is straightforward to implement

in $O(\log n)$ time by “top-down” traversal of the Fenwick tree our techniques do not appear be useful to speed up this solution on the UWRAM. We leave it as an open problem to investigate the complexity of the `select` operation on the UWRAM.

Our results leave the precise relation between UWRAM and RAMBO model of computation open. While Farzan et al. [13] show how to simulate RAMBO algorithms with a small overhead in space our results show that a direct approach to designing UWRAM algorithms can produce significantly better results. We wonder what the precise relation between the models are and if stronger simulation results are possible.

References

1. Ben-Amram, A.M., Galil, Z.: A generalization of a lower bound technique due to Fredman and Saks. *Algorithmica* **30**(1), 34–66 (2001)
2. Ben-Amram, A.M., Galil, Z.: Lower bounds for dynamic data structures on algebraic RAMs. *Algorithmica* **32**(3), 364–395 (2002)
3. Bille, P., et al.: Dynamic relative compression, dynamic partial sums, and substring concatenation. *Algorithmica* **80**(11), 3207–3224 (2018). Announced at ISAAC 2016
4. Bille, P., Christiansen, A.R., Prezza, N., Skjoldjensen, F.R.: Succinct partial sums and Fenwick trees. In: Fici, G., Sciortino, M., Venturini, R. (eds.) SPIRE 2017. LNCS, vol. 10508, pp. 91–96. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67428-5_8
5. Belloch, G.E.: Prefix sums and their applications. In: *Synthesis of Parallel Algorithms* (1990)
6. Brodnik, A.: Searching in constant time and minimum space (*Minimae res magni momenti sunt*). Ph.D. thesis, University of Waterloo (1995)
7. Brodnik, A., Carlsson, S., Fredman, M.L., Karlsson, J., Munro, J.I.: Worst case constant time priority queue. *J. Syst. Softw.* **78**(3), 249–256 (2005)
8. Brodnik, A., Karlsson, J., Munro, J.I., Nilsson, A.: An $O(1)$ solution to the prefix sum problem on a specialized memory architecture. In: Navarro, G., Bertossi, L., Kohayakawa, Y. (eds.) TCS 2006. IIFIP, vol. 209, pp. 103–114. Springer, Boston, MA (2006). https://doi.org/10.1007/978-0-387-34735-6_12
9. Burkhard, W.A., Fredman, M.L., Kleitman, D.J.: Inherent complexity trade-offs for range query problems. *Theor. Comput. Sci.* **16**(3), 279–290 (1981)
10. Chan, T.M., Chen, E.Y.: Optimal in-place algorithms for 3-D convex hulls and 2-D segment intersection. In: *Proceedings of the 25th SOCG*, pp. 80–87 (2009)
11. Chen, T., Raghavan, R., Dale, J.N., Iwata, E.: Cell broadband engine architecture and its first implementation—a performance view. *IBM J. Res. Dev.* **51**(5), 559–572 (2007)
12. Dietz, P.F.: Optimal algorithms for list indexing and subset rank. In: Dehne, F., Sack, J.-R., Santoro, N. (eds.) WADS 1989. LNCS, vol. 382, pp. 39–46. Springer, Heidelberg (1989). https://doi.org/10.1007/3-540-51542-9_5
13. Farzan, A., López-Ortiz, A., Nicholson, P.K., Salinger, A.: Algorithms in the ultrawide word model. In: Jain, R., Jain, S., Stephan, F. (eds.) TAMC 2015. LNCS, vol. 9076, pp. 335–346. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-17142-5_29
14. Fenwick, P.M.: A new data structure for cumulative frequency tables. *Softw. Pract. Exp.* **24**(3), 327–336 (1994)

15. Franceschini, G., Muthukrishnan, S., Pătraşcu, M.: Radix sorting with no extra space. In: Arge, L., Hoffmann, M., Welzl, E. (eds.) ESA 2007. LNCS, vol. 4698, pp. 194–205. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75520-3_19
16. Frandsen, G.S., Miltersen, P.B., Skyum, S.: Dynamic word problems. *J. ACM* **44**(2), 257–271 (1997)
17. Fredman, M., Saks, M.: The cell probe complexity of dynamic data structures. In: Proceedings of the 21st STOC, pp. 345–354 (1989)
18. Fredman, M.L.: A lower bound on the complexity of orthogonal range queries. *J. ACM* **28**(4), 696–705 (1981)
19. Fredman, M.L.: The complexity of maintaining an array and computing its partial sums. *J. ACM* **29**(1), 250–260 (1982)
20. Hagerup, T.: Sorting and searching on the word RAM. In: Morvan, M., Meinel, C., Krob, D. (eds.) STACS 1998. LNCS, vol. 1373, pp. 366–398. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0028575>
21. Hampapuram, H., Fredman, M.L.: Optimal biweighted binary trees and the complexity of maintaining partial sums. *SIAM J. Comput.* **28**(1), 1–9 (1998)
22. Hon, W.K., Sadakane, K., Sung, W.K.: Succinct data structures for searchable partial sums with optimal worst-case performance. *Theor. Comput. Sci.* **412**(39), 5176–5186 (2011)
23. Husfeldt, T., Rauhe, T.: New lower bound techniques for dynamic partial sums and related problems. *SIAM J. Comput.* **32**(3), 736–753 (2003)
24. Husfeldt, T., Rauhe, T., Skyum, S.: Lower bounds for dynamic transitive closure, planar point location, and parentheses matching. In: Karlsson, R., Lingas, A. (eds.) SWAT 1996. LNCS, vol. 1097, pp. 198–211. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61422-2_132
25. Ladner, R.E., Fischer, M.J.: Parallel prefix computation. *J. ACM* **27**(4), 831–838 (1980)
26. Larsen, K.G., Pagh, R.: I/O-efficient data structures for colored range and prefix reporting. In: Proceedings of the 23rd SODA, pp. 583–592 (2012)
27. Leben, R., Miletic, M., Špegel, M., Trost, A., Brodnik, A., Karlsson, J.: Design of high performance memory module on PC100. In: Proceedings of the ECSC, pp. 75–78 (1999)
28. Lindholm, E., Nickolls, J., Oberman, S., Montrym, J.: NVIDIA Tesla: a unified graphics and computing architecture. *IEEE Micro* **28**(2), 39–55 (2008)
29. Miltersen, P.B.: Cell probe complexity—a survey. In: Proceedings of the 19th FSTTCS, p. 2 (1999)
30. Munro, J.I., Suwanda, H.: Implicit data structures for fast search and update. *J. Comput. Syst. Sci.* **21**(2), 236–250 (1980)
31. Pătraşcu, M., Demaine, E.D.: Logarithmic lower bounds in the cell-probe model. *SIAM J. Comput.* **35**(4), 932–963 (2006). Announced at SODA 2004
32. Raman, R., Raman, V., Rao, S.S.: Succinct dynamic data structures. In: Dehne, F., Sack, J.-R., Tamassia, R. (eds.) WADS 2001. LNCS, vol. 2125, pp. 426–437. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44634-6_39
33. Reinders, J.: AVX-512 Instructions. Intel Corporation, Santa Clara (2013)
34. Ryabko, B.Y.: A fast on-line adaptive code. *IEEE Trans. Inf. Theory* **38**(4), 1400–1404 (1992)
35. Salowe, J., Steiger, W.: Simplified stable merging tasks. *J. Algorithms* **8**(4), 557–571 (1987)

36. Stephens, N., et al.: The ARM scalable vector extension. *IEEE Micro* **37**(2), 26–39 (2017)
37. Williams, J.W.J.: Algorithm 232: heapsort. *Commun. ACM* **7**, 347–348 (1964)
38. Yao, A.C.: On the complexity of maintaining partial sums. *SIAM J. Comput.* **14**(2), 277–288 (1985)