A GMP-based implementation of Schönhage-Strassen's large integer multiplication algorithm

Pierrick Gaudry, Alexander Kruppa, Paul Zimmermann

A GMP-based Implementation of Schönhage-Strassen's Large Integer Multiplication Algorithm

Pierrick Gaudry Alexander Kruppa Paul Zimmermann
LORIA, CACAO project-team, Campus scientifique, 54506 Vandœuvre-lès-Nancy

ABSTRACT

Schönhage-Strassen's algorithm is one of the best known algorithms for multiplying large integers. Implementing it efficiently is of utmost importance, since many other algorithms rely on it as a subroutine. We present here an improved implementation, based on the one distributed within the GMP library. The following ideas and techniques were used or tried: faster arithmetic modulo $2^n + 1$, improved cache locality, Mersenne transforms, Chinese Remainder Reconstruction, the $\sqrt{2}$ trick, Harley's and Granlund's tricks, improved tuning.

Categories and Subject Descriptors

I.1.2 [**Computing methodologies**]: Algorithms—*Symbolic and algebraic manipulation*

General Terms

Algorithms, Performance

Keywords

Integer multiplication, multiprecision arithmetic

INTRODUCTION

Since Schönhage and Strassen presented in 1971 a method to multiply two N-bit integers in $O(N \log N \log\log N)$ time [19], several authors have shown how to reduce other operations — inverse, division, square root, gcd, base conversion, elementary functions — to multiplication, possibly with $\log N$ multiplicative factors [5, 7, 15, 16, 18, 21]. It has now become common practice to express complexities in terms of the cost $M(N)$ to multiply two N-bit numbers, and many researchers tried hard to get the best possible constants in front of $M(N)$ for the above-mentioned operations (see for example [6, 14]).

Strangely, much less effort was made for decreasing the implicit constant in $M(N)$ itself, although any gain on that constant will give a similar gain on all multiplication-based operations. Some authors reported on implementations of large integer arithmetic for specific hardware or as part of a number-theoretic project [2, 10]. In this article we concentrate on the question of an optimized implementation of Schönhage-Strassen's algorithm on a classical workstation.

In the last few years, the multiplication of large integers has found several new applications in "real life", and not only in computing billions of digits of π. One such application is the segmentation method (called Kronecker substitution in [23]) to reduce the multiplication of polynomials with integer coefficients to one huge integer multiplication; this is used for example in the GMP-ECM software [25]. Another example is the multiplication or factorization of multivariate polynomials [21, 22].

In this article we detail several ideas or techniques that may be used to implement Schönhage-Strassen's algorithm (SSA) efficiently. As a consequence, we obtain what we believe is the best existing implementation of SSA on current processors; this implementation might be used as a reference to compare with other algorithms based on the Fast Fourier Transform, in particular those using complex floating-point numbers.

The paper is organized as follows: §1 revisits the original SSA and defines the notation used in the rest of the paper; §2 describes the different ideas and techniques we tried; finally §3 provides timing figures and graphs obtained with our new GMP implementation, and compares it to other implementations.

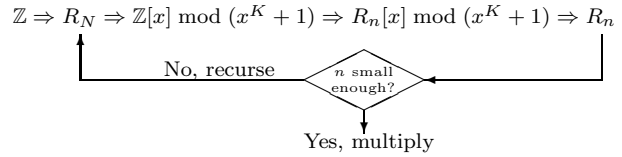
1. THE ALGORITHM OF SCHÖNHAGE AND STRASSEN

Throughout the paper we use w for the computer word size in bits — usually 32 or 64 — and denote by N the number of bits of the numbers we want to multiply.

Several descriptions of SSA can be found in the literature, see [11, 19] for example. We recall it here to establish the notations.

Let R_N^+ — or simply R_N — be the ring of integers modulo $2^N + 1$. SSA reduces integer multiplication to multiplication in R_N, which reduces to polynomial multiplication in $\mathbb{Z}[x] \bmod (x^K + 1)$, which in turn reduces to polynomial multiplication in $R_n[x] \bmod (x^K + 1)$, which finally reduces to multiplication in R_n. The reason for choosing R_N as the ring to map the input integers to is that the multiplications of elements of R_n can use SSA recursively, skipping the first step of mapping from integers to R_N again.

ISSAC'07, July 29–August 1, 2007, Waterloo, Ontario, Canada.



The first reduction — from $\mathbb{Z}$ to R_N — is simple: to multiply two non-negative integers of u and v bits, it suffices to compute their product mod 2^N+1 for $N \geq u+v$.

The second step — a map from R_N to $\mathbb{Z}[x] \bmod (x^K+1)$ — works as follows. Assume $N = 2^k M$ for integers k and M, and define $K := 2^k$. An integer $a \in [0, 2^N]$ can be uniquely written $a = \sum_{i=0}^{K-1} a_i 2^{iM}$, with $0 \leq a_i < 2^M$ for $i < K-1$, and $0 \leq a_{K-1} \leq 2^M$, that is, we cut a into K pieces of M bits each, except the last piece can be equal to 2^M. Now the integer a is the value at $x = 2^M$ of the polynomial $A(x) = \sum_{i=0}^{K-1} a_i x^i$. Assume we decompose an integer $b \in R_N$ in the same manner, and let $C(x)$ be the product $A(x)B(x)$ over $\mathbb{Z}[x]$: $C(x) = \sum_{i=0}^{2K-2} c_i x^i$. One now has $ab = A(2^M)B(2^M) = C(2^M)$, thus $ab = \sum_{i=0}^{2K-2} c_i 2^{iM}$.

Now what we really want is $ab \bmod (2^N+1)$, i.e.,

$$(c_0 - c_K) + \cdots + (c_{K-2} - c_{2K-2}) 2^{(K-2)M} + c_{K-1} 2^{(K-1)M} \quad (1)$$

which comes from $C^+(x) := \sum_{i=0}^{K-1} \bar{c}_i x^i = A(x)B(x) \bmod (x^K+1)$, since $x = 2^M$ and $N = KM$. To determine $C^+(x)$, one uses a *negacyclic convolution* over the ring R_n, i.e., modulo 2^n+1, where n is taken large enough so that the $\bar{c}_i$ can be recovered exactly. For $0 \leq i \leq K-1$, one has $0 \leq c_i = \sum_{j=0}^{i} a_j b_{i-j} < (i+1)2^{2M}$. Similarly for $K \leq i \leq 2K-3$, one has $0 \leq c_i < (2K-1-i)2^{2M}$ and finally $0 \leq c_{2K-2} \leq 2^{2M}$. With the convention that $c_{2K-1} = 0$, according to (1), we have

$$((i+1)-K)2^{2M} \leq \bar{c}_i = c_i - c_{i+K} < (i+1)2^{2M} \quad (2)$$

for $0 \leq i < K$. Hence each coefficient of $C(x) \bmod (x^K+1)$ is confined to an interval of length $K2^{2M}$, and so it suffices to have $2^n + 1 \geq K2^{2M}$, i.e., $n \geq 2M + k$[1].

The negacyclic convolution $A(x)B(x) \bmod (x^K+1)$ can be performed efficiently using the Fast Fourier Transform (FFT). More precisely, SSA uses here a simple case of the Discrete Weighted Transform (DWT) [10]. Assume $\omega = \theta^2$ is a primitive Kth root of unity in R_n. (All operations in this paragraph are in R_n.) Given $(a_i)_{0 \leq i < K}$, the weight signal is $(a'_i := \theta^i a_i)_{0 \leq i < K}$. The forward transform computes $(\hat{a}_i := \sum_{j=0}^{K-1} \omega^{ij} a'_j)_{0 \leq i < K}$, and similarly for $(\hat{b}_i)$. One then multiplies $\hat{a}_i$ and $\hat{b}_i$ together in R_n (pointwise products): let $\hat{c}_i = \hat{a}_i \hat{b}_i$. The backward transform computes $(c'_i := \sum_{j=0}^{K-1} \omega^{-ij} \hat{c}_j)_{0 \leq i < K}$:

$$\begin{aligned} c'_i &= \sum_{j=0}^{K-1} \omega^{-ij} \hat{a}_j \hat{b}_j = \sum_{j=0}^{K-1} \omega^{-ij} \left(\sum_{\ell=0}^{K-1} \omega^{j\ell} a'_\ell \right) \left(\sum_{m=0}^{K-1} \omega^{jm} b'_m \right) \\ &= \sum_{\ell,m=0}^{K-1} a_\ell b_m \theta^{\ell+m} \sum_{j=0}^{K-1} \omega^{j(\ell+m-i)}. \end{aligned}$$

Since ω is a primitive Kth root of unity, $\sum_{j=0}^{K-1} \omega^{j(\ell+m-i)}$ is zero unless $\ell + m - i \equiv 0 \bmod K$, which holds for $\ell + m = i$ or $\ell + m = i + K$. Since $\theta^K \equiv -1 \bmod (2^n+1)$, it follows:

$$c'_i = K\theta^i \sum_{\ell=0}^{K-1} (a_\ell b_{i-\ell} - a_\ell b_{i+K-\ell}) = K\theta^i (c_i - c_{i+K}),$$

where b_m is assumed zero for m outside the range $[0, K-1]$.

SSA thus consists of five consecutive steps, where all computations in steps (2) to (4) are done modulo 2^n+1:

(1) the "decompose" step extracts from a the M-bit parts a_i, and multiplies them by the weight signal θ^i, obtaining a'_i (similarly for b_i);

(2) the "forward transform" computes $(\hat{a}_0, \ldots, \hat{a}_{K-1})$ from $(a'_0, \ldots, a'_{K-1})$ (similarly for $\hat{b}_i$);

(3) the "pointwise product" step computes $\hat{c}_i = \hat{a}_i \hat{b}_i$, for $0 \leq i < K$;

(4) the "backward transform" computes $(c'_0, \ldots, c'_{K-1})$ from $(\hat{c}_0, \ldots, \hat{c}_{K-1})$;

(5) the "recompose" step divides c'_i by $2^k \theta^i$, and constructs the final result as $\bar{c}_0 + \bar{c}_1 2^M + \cdots + \bar{c}_{K-1} 2^{(K-1)M}$. Some $\bar{c}_i$, defined in Eq. (2), may be negative, but the sum is necessarily non-negative.

For a given input bit-size N, several choices of the FFT length K may be possible. SSA is thus a whole family of algorithms: we call FFT-K — or FFT-2^k — the algorithm splitting the inputs into $K = 2^k$ parts. For a given input size N, one of the main practical problems is how to choose the best value of the FFT length K, and thus of the bit-size n of the smaller multiplies (see §2.6).

1.1 Choice of n and Efficiency

SSA takes for n a multiple of K, so that $\omega = 2^{2n/K}$ is a primitive Kth root of unity, and $\theta = 2^{n/K}$ is used for the weight signal. This ensures that all FFT butterflies only involve additions/subtractions and shifts on a radix 2 computer (see §2.1).

In practice one may additionally require n to be a multiple of the word size w, to make the arithmetic in 2^n+1 simpler. Indeed, a number from R_n is then represented by n/w machine words, plus one additional bit of weight 2^n. We call this a *semi-normalized* representation, since values up to $2^{n+1}-1$ can be represented.

For a given bit size N divisible by $K = 2^k$, we define the *efficiency* of the FFT-K scheme:

$$\frac{2N/K + k}{n},$$

where n is the smallest multiple of K larger than or equal to $2N/K + k$. For example for $N = 1,000,448$ and $K = 2^{10}$, we have $2N/K + k = 1964$, and the next multiple of K is $n = 2048$, therefore the efficiency is $\frac{1964}{2048} \approx 96\%$. For $N = 1,044,480$ with the same value of K, we have $2N/K + k = 2050$, and the next multiple of K is $n = 3072$, with an efficiency of about 67%. The FFT scheme is close to optimal when its efficiency is near 100%.

Note that a scheme with efficiency below 50% does not need to be considered. Indeed, this means that $2N/K + k \leq \frac{1}{2}n$, which necessarily implies that $n = K$ (remember n has to be divisible by K). Then the FFT scheme of length $K/2$ can be performed with the same value of n, since

[1] One might use $n \geq 2M + k + 1$ to get a lifting algorithm from R_n to $\mathbb{Z}$ which is independent of i.

$2(N/(K/2))+(k-1) < 4N/K+2k \leq n$, and n is a multiple of $K/2$.

From this last remark, we can assume $2N/K \geq \frac{1}{2}n$ — neglecting the small k term —, which together with $n \geq K$ gives:

$$K \leq 2\sqrt{N}. \quad (3)$$

2. OUR IMPROVEMENTS

We describe in this section the ideas and techniques we have tried to improve the GMP implementation of SSA. We started from the GMP 4.2.1 implementation, and used the graph of the multiplication time up to $1,000,000$ words on an Opteron as benchmark. After encoding each idea, if the new graph was better than the old one, the new idea was validated, otherwise it was discarded. Each technique saved only 5% up to 20%, but all techniques together saved a factor of about 2 with respect to GMP 4.2.1.

2.1 Arithmetic Modulo $2^n + 1$

Arithmetic operations modulo 2^n+1 have to be performed during the forward and backward transforms, when applying the weight signal, and when unapplying it. Thanks to the fact that the primitive roots of unity are powers of two, the only needed operations are additions, subtractions, and multiplications by a power of two. Divisions by 2^k can be reduced to multiplications by 2^{2n-k}.

We recall that we desire n to be a multiple of the number w of bits per word. Since n must also be a multiple of $K = 2^k$, this is not a real constraint, unless $k < 5$ on a 32-bit computer, or $k < 6$ on a 64-bit computer. Let $m = n/w$ be the number of computer words corresponding to an n-bit number. A residue mod $2^n + 1$ has a semi-normalized representation with m full words and one carry of weight 2^n:

$$a = (a_m, a_{m-1}, \ldots, a_0),$$

with $0 \leq a_i < 2^w$ for $0 \leq i < m$, and $0 \leq a_m \leq 1$.

The addition of two such representations is done as follows (we give here the GMP code):

```
c = a[m] + b[m] + mpn_add_n (r, a, b, m);
r[m] = (r[0] < c);
MPN_DECR_U (r, m + 1, c - r[m]);
```

The first line adds $(a_{m-1}, \ldots, a_0)$ with $(b_{m-1}, \ldots, b_0)$, puts the low m words of the result in $(r_{m-1}, \ldots, r_0)$, and adds the out carry to $a_m + b_m$; we thus have $0 \leq c \leq 3$. The second line yields $r_m = 0$ if $r_0 \geq c$, in which case we simply subtract c from r_0 at the third line. Otherwise $r_m = 1$, and we subtract $c - 1$ from r_0: a borrow may propagate, but at most to r_m. In all cases $r = a + b \bmod (2^n + 1)$, and r is semi-normalized. The subtraction is done in a similar manner.

The multiplication by 2^e is more tricky to implement. However this operation mainly appears in the butterflies $[a, t] \leftarrow [a+b, (a-b)2^e]$ of the forward and backward transforms, which may be performed as follows:

```
Bfy(a, b, t, e)
   1. Write e = d*w + s with 0 <= s < w,
         where w is the number of bits per word
   2. Decompose a = (ah, al),
         where ah contains the upper d words
      Idem for b
   3. t <- (al - bl, bh - ah)
   4. a <- a + b
   5. t <- t * 2^s
```

Step 3 means that the most significant words from t are formed with `al - bl`, and the least significant words with `bh - ah`, where we assume that borrows are propagated, so that t is semi-normalized. Thus the only real multiplication by a power of two is that of step 5, which may be efficiently performed with GMP's `mpn_lshift` routine.

If one has a combined `addsub` routine which computes simultaneously $x + y$ and $x - y$ faster than two separate calls, then step 4 can be written `a <- (bh + ah, al + bl)`, which shows that t and a may be computed with two `addsub` calls.

2.2 Cache Locality During the Transforms

When multiplying large integers with SSA, the time spent in accessing data for performing the Fourier transforms is non-negligible. The literature is rich with papers dealing with the organization of the computations in order to improve the locality. However most of these papers are concerned with contexts which are different from ours: usually the coefficients are small and most often they are complex numbers represented as a pair of `double`'s. Also there is a variety of target platforms, from embedded hardware implementations to super-scalar computers.

We have tried to apply several of these approaches in our context where the coefficients are modular integers that fit in at least a few cache lines and the target platform is a standard PC workstation.

In this work, we concentrate on multiplying large, but not huge integers. By this we mean that we consider only 3 levels of memory for our data: L1 cache, L2 cache, and standard RAM. In the future we might consider also the case where we have to use the hard disk as a 4th level.

Here are the orders of magnitude for these memories, to fix ideas: on a typical Opteron, a cache line is 64 bytes; the L1 data cache is 64 kB; the L2 cache is 1 MB; the RAM is 8 GB. The smallest coefficient size (i.e., n-bit residues) we consider is about 50 machine words, that is 400 bytes. For very large integers, a single coefficient hardly fits in the L1 cache.

The very first FFT algorithm is the iterative one. In our context this is a really bad idea. The main advantage of it is that the data is accessed in a sequential way. In the case where the coefficients are small enough so that several of them fit in a cache line, this saves many cache misses. But in our case, contiguity is irrelevant due to the size of the coefficients compared to cache lines.

The next very classical FFT algorithm is the recursive one. In this algorithm, at a certain level of recursion, we work on a small set of coefficients, so that they must fit in the cache. This version (or a variant of it) was implemented in GMP up to version 4.2.1. This behaves well for moderate sizes, but when multiplying large numbers, everything fits in the cache only at the tail of the recursion, so that most of the transform is already done when we are at last in the cache. The problem is that before getting to the appropriate recursion level, the accesses are very cache unfriendly.

In order to improve the locality for large transforms, we have tried three strategies found in the literature: the Belgian approach, the radix-2^k transform, and Bailey's 4-step algorithm.

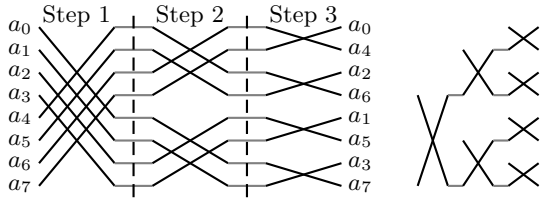


Figure 1: FFT circuit of length 8 and butterfly tree of depth 3.

2.2.1 The Belgian Transform

In [9], Brockmeyer et *al.* propose a way of organizing the transform that reduces cache misses. In order to explain it, let us first define a tree of butterflies as follows (we don't mention the root of unity for simplicity):

```
TreeBfy(A, index, depth, stride)
    Bfy(A[index], A[index+stride])
    if depth > 1
        TreeBfy(A, index-stride/2, depth-1, stride/2)
        TreeBfy(A, index+stride/2, depth-1, stride/2)
```

An example of a tree of depth 3 is given on the right of Figure 1. Now, the depth of a butterfly tree is bounded by a value that is not the same for every tree. For instance, on Figure 1, the butterfly tree that starts with the butterfly between a_0 and a_4 has depth 1: one can not continue the tree on step 2. Similarly, the tree starting with the butterfly between a_1 and a_5 has depth 1, the tree starting between a_2 and a_6 has depth 2 and the tree starting between a_3 and a_7 has depth 3. More generally, the depth can be computed by a simple formula.

One can check that by considering all the trees of butterflies starting with an operation at step 1, we cover the complete FFT circuit. It remains to find the right ordering for computing those trees of butterflies. For instance, in the example of Figure 1, it is important to do the tree that starts between a_3 and a_7 in the end, since it requires data from all the other trees.

One solution is to perform the trees of butterflies following the `BitReverse` order. For an integer i whose binary representation fits in at most k bits, the value `BitReverse(i,k)` is the integer one obtains by reading the k bits (maybe padded with zeros) in the opposite order. One obtains the following algorithm, where `ord_2` stands for the number of trailing zeros in the binary representation of an integer (together with the 4-line `TreeBfy` routine, this is a recursive description of the 36-line routine from [9, Code 6.1]):

```
BelgianFFT(A, k)
    K = 2^{k-1}
    for i := 0 to K-1
        TreeBfy(A, BitReverse(i, k-1), 1+ord_2(i+1), K)
```

Inside a tree of butterflies, we see that most of the time, the butterfly operation will involve a coefficient that has been used just before, so that it should still be in the cache. Therefore an approximate 50% cache-hit is provided by construction, and we can hope for more if the data is not too large compared to the cache size.

We have implemented this in GMP, and this saved a few percent for large sizes, thus confirming the fact that this approach is better than the classical recursive transform.

2.2.2 Higher Radix Transforms

The principle of higher radix transforms is to use an atomic operation which groups several butterflies. In the book [1] the reader will find a description of several variants in this spirit. The classical FFT can be viewed as a radix-2 transform. The next step is a radix-4 transform, where the atomic operation has 4 inputs and 4 outputs (without counting roots of unity) and groups 4 butterflies of 2 consecutive steps of the FFT.

We can then build a recursive algorithm upon this atomic operation. Of course, since we perform 2 steps at a time, the number of steps in the recursion is reduced by a factor of 2, and we have to handle separately the last step when the FFT level k is odd.

In the literature, the main interest for higher radix transforms comes from the fact that the number of operations is reduced for a transform of complex numbers (this is done by exhibiting a free multiplication by i). In our case, the number of operations remains the same. However, in the atomic block each input is used in two butterflies, so that the number of cache misses is less than 50%, just as for the Belgian approach. Furthermore, with the recursive structure, just as for the classical recursive FFT, at some point we deal with a number of inputs which is small enough so that everything fits in the cache.

We have tested this approach, and this was faster than the Belgian transform by a few percent.

The next step after radix 4 is radix 8 which works in the same spirit, but grouping 3 levels at a time. We have also implemented it, but this saved nothing, and was even sometimes slower than the radix 4 approach. Our explanation is that for small numbers, radix 4 is close to optimal with respect to cache locality, and for large numbers, the number of coefficients that fit in the cache is rather small and we have misses inside the atomic block of 12 butterflies. Further investigation is needed to validate this explanation.

More generally, radix 2^t groups t levels together, with a total of $t2^{t-1}$ butterflies, over 2^t residues. If all those residues fit in the cache, the cache miss rate is less than $1/t$. Thus the optimal strategy seems to choose for t the largest integer such that $2^t n$ bits fit in the cache (either L1 or L2, in fact the smallest cache where a single radix 2 butterfly fits).

2.2.3 Bailey's 4-step Algorithm

The algorithm we describe in this section can be found in a paper by Bailey [3]. In there, the reader will find earlier references tracing back the original idea. For simplicity we stick to the "Bailey's algorithm" denomination.

A way of seeing Bailey's 4-step transform algorithm is as a radix-$\sqrt{K}$ transform, where $K = 2^k$ is the length of the input sequence. In other words, instead of grouping 2 steps as in radix-4, we group $k/2$ steps. To be more general, let us write $k = k_1 + k_2$, where k_1 and k_2 are to be thought as close to $k/2$, but this is not really necessary. Then Bailey's 4-step algorithm consists in the following phases:

1. Perform 2^{k_2} transforms of length 2^{k_1};
2. Multiply the data by weights;
3. Perform 2^{k_1} transforms of length 2^{k_2}.

There are only three phases in this description. The fourth phase is usually some matrix transposition, but this is irrelevant in our case: the coefficients are large so that we keep

a table of pointers to them, and this transposition is just pointer exchanges which are basically for free, and fit very well in the cache.

The second step involving weights is due to the fact that in the usual description of Bailey's 4-step algorithm, the transforms of length 2^{k_1} are exactly Fourier transforms, whereas the needed operation is a twisted Fourier transform where the roots of unity involved in the butterflies are different (since they involve a 2^k-th root of unity, whereas the classical transform of length 2^{k_1} involves a 2^{k_1}-th root of unity). In the classical FFT setting this is very interesting, since we can then reuse some small-dimension implementation that has been very well optimized. In our case, we have found it better to write separate code for this twisted FFT, so that we merge the first and second phases.

The interest of this way of organizing the computation is again not due to a reduction of the number of operations, since they are exactly the same as with the other FFT approaches mentioned above. The goal is to help locality. Indeed, assume that $\sqrt{K}$ coefficients fit in the cache, then the number of cache misses is at most $2K$, since each call to the internal FFT or twisted FFT operates on $\sqrt{K}$ coefficients.

Of course we are interested in numbers for which $\sqrt{K}$ coefficients do not fit in the L1 cache, but for all numbers we might want to multiply, they do fit in the L2 cache. Therefore the structure of the code follows the memory hierarchy: at the top level of Bailey's algorithm, we deal with the RAM vs L2 cache locality question, then in each internal FFT or twisted FFT, we can take care of the L2 vs L1 cache locality question. This is done by using the radix-4 variant inside our Bailey-algorithm implementation.

We have implemented this approach (with a threshold for activating Bailey's algorithm only for large sizes), and combined with radix-4, this gave us our best timings. We have also tried a higher dimensional transform, in particular 3 steps of size $\sqrt[3]{K}$. This did not help for the sizes we considered.

2.2.4 Mixing Several Phases

Another way to improve locality is to mix different phases of the algorithm in order to do as much work as possible on the data while they are in the cache. An easy improvement in this spirit is to mix the pointwise multiplication and the backward transform, in particular when Bailey's algorithm is used. Indeed, after the two forward transforms have been computed, one can load the data corresponding to the first column, do the pointwise multiplication of its elements, and readily perform the small transform of this column. Then the data corresponding to the second column is loaded, multiplied and transformed, and so on. In this way, one saves one full pass on the data. Taking the idea one step further, assuming that the forward transform for the first input number has been done already (or that we are squaring one number), after performing the column-wise forward transform on the second number we can immediately do the point-wise multiply and the backward transform on the column, so saving another pass over memory.

Following this idea, we can also merge the "decompose" and "recompose" steps with the transforms, again to save a pass on the data. In the case of the "decompose" step, there is more to it since one can also save unnecessary copies by merging it with the first step of the forward transform.

The "decompose" step consists of cutting parts of M bits from the input numbers, then multiplying each part a_i by θ^i modulo 2^n+1, giving a'_i. If one closely looks at the first FFT level, it will perform a butterfly between a'_i and $a'_{i+K/2}$ with θ^{2i} as multiplier. This will compute $a'_i + a'_{i+K/2}$ and $a'_i - a'_{i+K/2}$, and multiply the latter by θ^{2i}. It can be seen that the M non-zero bits from a'_i and $a'_{i+K/2}$ do not overlap, thus no real addition or subtraction is required: the results $a'_i + a'_{i+K/2}$ and $a'_i - a'_{i+K/2}$ can be obtained with just copies and ones' complements. As a consequence, it should be possible to completely avoid the "decompose" step and the first FFT level, by directly starting from the second FFT level, which for instance will add $a'_i + a'_{i+K/2}$ to $(a'_j - a'_{j+K/2})\theta^{2j}$; here the four operands $a'_i, a'_{i+K/2}, a'_j, a'_{j+K/2}$ will be directly taken from the input integer a, and the implicit multiplier θ^{2j} will be used to know where to add or subtract a'_j and $a'_{j+K/2}$. This example illustrates the kind of savings obtained by avoiding trivial operations like copies and ones' complements, and furthermore improving the locality. This idea was not used in the results from §3.

2.3 Fermat and Mersenne Transforms

The reason why SSA uses negacyclic convolutions is because the algorithm can be used recursively: the "pointwise products" modulo 2^n+1 can in turn be performed using the same algorithm, each one giving rise to K' smaller pointwise products modulo $2^{n'}+1$. (In that case, n must satisfy an additional divisibility condition related to K'.) A drawback of this approach is that it requires a weighted transform, i.e., additional operations before the forward transforms and after the backward transform. However, if one looks carefully, power-of-two roots of unity are needed only at the "lower level", i.e., in R_n^+. Therefore one can replace R_N by R_N^- — i.e., the ring of integers modulo 2^N-1 — in the original algorithm, and replace the weighted transform by a classical cyclic convolution, to compute a product mod 2^N-1. This works only at the top level of the algorithm, and not recursively. We call this a "Mersenne transform", whereas the original SSA performs a "Fermat transform"[2]. This idea of using a Mersenne transform is already present in [4] where it is called "cyclic Schönhage-Strassen trick".

Despite the fact that it can be used at the top level only, the Mersenne transform is nevertheless very interesting for the following reasons:

• a Mersenne transform modulo 2^N-1, combined with a Fermat transform modulo 2^N+1 and CRT reconstruction, can be used to compute a product of two N-bit integers;

• a Mersenne transform can use a larger FFT length $K=2^k$ than the corresponding Fermat transform. Indeed, while K must divide N for the Fermat transform, so that $\theta = 2^{N/K}$ is a power of two, it only needs to divide $2N$ for the Mersenne transform, so that $\omega = 2^{2N/K}$ is a power of two. This improves the efficiency for K near $\sqrt{N}$, and enables one to use a value of K close to optimal. (The constraint on the FFT length can still be decreased by using the "$\sqrt{2}$ trick", see §2.4.)

The above idea can be generalized to a Fermat transform mod $2^{aN}+1$ and a Mersenne transform mod $2^{bN}-1$ for small integers a, b.

[2] In the whole paper, a Fermat transform, product, or scheme is meant modulo 2^N+1, without N being necessarily a power of two as in Fermat numbers.

LEMMA 1. *Let a, b be two positive integers. Then at least one of* $\gcd(2^a+1, 2^b-1)$ *and* $\gcd(2^a-1, 2^b+1)$ *is* 1.

PROOF. Let $g = \gcd(a,b)$, $r = 2^g$, $a' = a/g$, $b' = b/g$. Denote by $\mathrm{ord}_p(r)$ the multiplicative order of $r \pmod p$. In the case of b' odd, $p \mid r^{b'}-1 \Rightarrow \mathrm{ord}_p(r) \mid b' \Rightarrow 2 \nmid \mathrm{ord}_p(r)$, and $p \mid r^{a'}+1 \Rightarrow \mathrm{ord}_p(r) \mid 2a'$ and $\mathrm{ord}_p(r) \nmid a' \Rightarrow 2 \mid \mathrm{ord}_p(r)$, hence no prime can divide both $r^{b'}-1$ and $r^{a'}+1$. In the other case of b' even, a' must be odd, and the same argument holds with the roles of a' an b' exchanged, so no prime can divide both $r^{a'}-1$ and $r^{b'}+1$. □

It follows from Lemma 1 that we can use one Fermat transform of size aN (respectively bN) and one Mersenne transform of size bN (respectively aN). However this does not imply that the reconstruction is easy: in practice we used $b = 1$ and made only a vary (see §2.6.2).

2.4 The $\sqrt{2}$ Trick

Since all prime factors of 2^n+1 are $p \equiv 1 \pmod 8$ if $4 \mid n$, 2 is a quadratic residue (mod n), and it turns out that $\sqrt{2}$ is of a simple enough form to make it useful as a root of unity with power-of-two order. Specifically, $\left(2^{3n/4}-2^{n/4}\right)^2 \equiv 2 \pmod{2^n+1}$, which is easily checked by expanding the square. Hence we can use $\sqrt{2} = 2^{3n/4}-2^{n/4}$ as a root of unity of order 2^{k+2} in the transform to double the possible transform length for a given n. In the case of the negacyclic transform, this allows a length 2^{k+1} transform, and $\sqrt{2}$ is used only in the weight signal. For a cyclic transform, $\sqrt{2}$ is used normally as a root of unity during the transform, allowing a transform length of 2^{k+2}. This idea is mentioned in [4, §9] where it is credited without reference to Schönhage, but we have been unable to track down the original source. In our implementation, this $\sqrt{2}$ trick saved roughly 10% on the total time of integer multiplication.

Unfortunately using higher roots of unity for the transform is not feasible as prime divisors of 2^n+1 are not necessarily congruent to 1 $\pmod{2^{k+3}}$, deciding whether they are or not requires factoring 2^n+1, and even if they are as in the case of the eighth Fermat number $F_8 = 2^{256}+1$ [8], there does not seem to be a simple form for $\sqrt[4]{2}$ which would make it useful as a root on unity in the transform.

2.5 Harley's and Granlund's Tricks

Rob Harley [13] suggested the following trick[3] to improve the efficiency of a given FFT scheme. Assume $2M+k$ is just above an integer multiple of K, say λK. Then we have to use $n = (\lambda+1)K$, which gives an efficiency of only about $\frac{\lambda}{\lambda+1}$. Harley's idea is to use $n = \lambda K$ instead, and recover the missing information from a CRT-reconstruction with an additional computation modulo the machine word 2^w.

A drawback of Harley's trick is that when only a few bits are missing, the K^2 word products may become relatively expensive. When only a few bits are missing, we can multiply $A(x)$ and $B(x)$ over $\mathbb{Z}[x]$ modulo a small power of 2 using the segmentation method. That way, if $h \le w$ bits are missing, one trades K^2 word products for the product of two large integers of $(2h+k)K/w$ words, which can in turn use fast multiplication[4].

[3]Bernstein attributes a similar idea to Karp in [4, §9].

[4]Harley's trick extends to $h > w$: together with the segmentation method, the exact same reasoning holds.

Torbjörn Granlund [12] found that this idea — combining computations mod 2^n+1 with computations mod 2^h — can also be used at the top-level for the plain integer multiplication, and not only at the lower-level as in Harley's trick. Assume one wants to multiply two integers u and v whose product has m bits, where m is just above an "optimal" Fermat scheme $(2^N+1, K)$, say $m = N+h$. Then first compute $uv \bmod (2^N+1)$, and second compute $uv \bmod 2^h$, by simply computing the plain integer product $(u \bmod 2^h)(v \bmod 2^h)$, again possibly in turn with fast multiplication. The exact value of uv can be efficiently reconstructed by CRT from both values. We call this idea "Granlund's trick".

Let us denote $M(N)$, $M^+(N)$ and $M^-(N)$ the cost of the multiplication of two N-bit integers, multiplication modulo 2^N+1 and multiplication modulo 2^N-1 respectively. Granlund's trick can be written $M(N+h) = M^+(N) + M(h)$, or $M(N+h) = M^+(N) + M^+(2h)$ if one reduces the plain product modulo 2^h to a modular product modulo $2^{2h}+1$. Marco Bodrato (personal communication) discovered that Granlund's trick can be applied simultaneously to the low and high ends of the product, giving $M(N+h) = M^+(N) + 2M(h/2)$.

We use neither Harley's nor Granlund's trick in our current implementation. We believe Granlund's trick is less efficient than the generalized Fermat-Mersenne scheme we propose (§2.3), which yields $M(a+b) = M^+(a) + M^-(b)$, with $M^-(N)$ the cost of a multiplication modulo 2^N-1, if a good efficiency is possible for $M^+(a)$ and $M^-(b)$. As for Harley's trick, we tried it only at the word level, i.e., for $\lambda K < 2M+k \le \lambda K+w$, which happens in rare cases only, and it made little difference. However, when multiplying two numbers modulo a Fermat number $2^{2^n}+1$, Harley's trick becomes very attractive, since $2M$ is a power of two in that case.

2.6 Improved Tuning

We found that significant speedups could be obtained with better tuning schemes, which we describe here. All examples given in this section are related to an Opteron.

2.6.1 Tuning the Fermat and Mersenne Transforms

Until version 4.2.1, GMP used a naive tuning scheme for the FFT multiplication. For the Fermat transforms modulo 2^N+1, an FFT of length 2^k was used for $t_k \le N < t_{k+1}$, where t_k is the smallest bit-size for which FFT-2^k is faster than FFT-2^{k-1}. For example on an Opteron, the default `gmp-mparam.h` file uses $k = 4$ for a size less than 528 machine words, then $k = 5$ for less than 1184 words, and so on:

```
#define MUL_FFT_TABLE  { 528, 1184, 2880, 5376, 11264,
  36864, 114688, 327680, 1310720, 3145728, 12582912, 0 }
```

A special rule is used for the last entry: here $k = 14$ is used for less than $m = 12582912$ words, $k = 15$ is used for less than $4m = 50331648$ words, and then $k = 16$ is used. An additional single threshold determines from which size upward — still in words — a Fermat transform mod 2^n+1 is faster than a full product of two n-bit integers:

```
#define MUL_FFT_MODF_THRESHOLD          544
```

For a product mod 2^n+1 of at least 544 words, GMP 4.2.1 therefore uses a Fermat transform, with $k = 5$ until 1183 words according to the above `MUL_FFT_TABLE`. Below

the 544 words threshold, the algorithm used is the 3-way Toom-Cook algorithm, followed by a reduction mod $2^n + 1$.

This scheme is clearly not optimal since the FFT-2^k curves intersect several times, as shown by Figure 2.

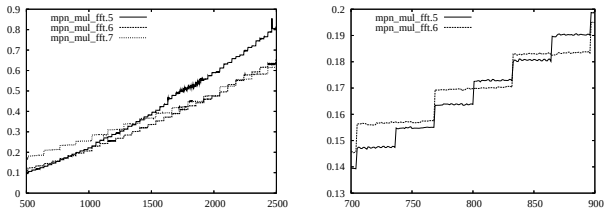


Figure 2: Time in milliseconds needed to multiply numbers modulo 2^n+1 with an FFT of length 2^k for $k = 5, 6, 7$. On the right, the zoom (with only $k = 5, 6$) illustrates that two curves can intersect several times.

To take into account those multiple crossings, the new tuning scheme determines word-intervals $[m_1, m_2]$ where the FFT of length 2^k is preferred for Fermat transforms:

```
#define MUL_FFT_TABLE2 {{1, 4 /*66*/}, {401, 5 /*96*/},
  {417, 4 /*98*/}, {433, 5 /*96*/}, {865, 6 /*96*/}, ...
```

The entry `{433, 5 /*96*/}` means that from 433 words — and up to the next size of 865 words — FFT-2^5 is preferred, with an efficiency of 96%. A similar table is used for Mersenne transforms.

2.6.2 *Tuning the Plain Integer Multiplication*

Up to GMP 4.2.1, a single threshold controls the plain integer multiplication:

```
#define MUL_FFT_THRESHOLD                7680
```

This means that SSA is used for a product of two integers of at least 7680 words, which corresponds to about 148,000 decimal digits, and the Toom-Cook 3-way algorithm is used below that threshold.

We now use the generalized Fermat-Mersenne scheme described in §2.3 with $b = 1$ (in our implementation we found $1 \le a \le 7$ was enough). Again, for each size, the best value of a is determined by our tuning program:

```
#define MUL_FFT_FULL_TABLE2 {{16, 1}, {4224, 2},
  {4416, 6}, {4480, 2}, {4608, 4}, {4640, 2}, ...
```

For example, the entry `{4608, 4}` means that to multiply two numbers of 4608 words — or whose product has 2×4608 words — the new algorithm uses one Mersenne transform modulo $2^N - 1$ and one Fermat transform modulo $2^{4N} + 1$. Reconstruction is easy since $2^{aN} + 1 \equiv 2 \bmod (2^N - 1)$.

3. RESULTS AND CONCLUSION

On July 1st, 2005, Allan Steel wrote a web page [20] entitled "*Magma V2.12-1 is up to 2.3 times faster than GMP 4.1.4 for large integer multiplication*". This was actually our first motivation for improving GMP's implementation.

Magma V2.13-6 takes 2.22s to multiply two numbers of 784141 words, whereas our GMP development code takes only 0.96s. Thus our GMP-based code is clearly faster than Magma by a factor of 2.3. Note that this does not mean that we have gained a factor $2.3^2 = 5.29$ over GMP 4.1.4. In both

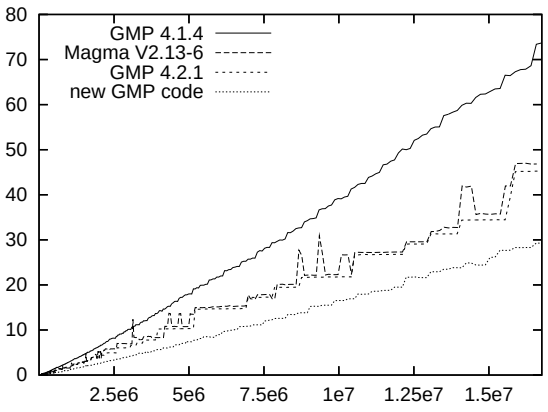


Figure 3: Comparison of GMP 4.1.4, GMP 4.2.1, Magma V2.13-6 and our new code for the plain integer multiplication on a 2.4Ghz Opteron (horizontal axis in 64-bit words, vertical axis in seconds).

cases, 2.3 is the maximal ratio between Magma V2.12-1 and GMP 4.1.4, and between our code and Magma V2.13-6 respectively, following the well known "benchmarketing" strategy[5] (both versions of Magma give very similar timings).

We have tested other freely available packages providing an implementation for large integer arithmetic. Among them, some (`OpenSSL/BN`, `LiDiA/libI`) do not go beyond Karatsuba algorithm, some do have some kind of FFT, but are not really made for really large integers: `arprec`, `Miracl`. Two useful implementations we have tested are `apfloat` and `CLN`. They take about 4 to 5 seconds on our test machine to multiply one million-word integers, whereas we need about 1 second. Bernstein mentions some partial implementation `Zmult` of Schönhage-Strassen's algorithm, with good timings, but right now, only very few sizes are handled, so that the comparison with our software is not really possible.

A program that implements a complex floating-point FFT for integer multiplication is George Woltman's `Prime95`. It is written mainly for testing large Mersenne numbers $2^p - 1$ for primality in the in the Great Internet Mersenne Prime Search [24]. It uses a DWT for multiplication mod $a2^n \pm c$, with a and c not too large, see [17]. We compared multiplication modulo $2^{2wn} - 1$ in `Prime95` version 24.14.2 with multiplication of n-word integers using our SSA implementation on a Pentium 4 at 3.2 GHz, and on an Opteron 250 at 2.4 GHz, see Figure 4. It is plain that `Prime95` beats our implementation by a wide margin, in fact usually by more than a factor of 10 on a Pentium 4, and by a factor between 2.5 and 3 on the Opteron. Some differences between `Prime95` and our implementation need to be pointed out: due to the floating point nature of `Prime95`'s FFT, rounding errors can build up for particular input data to the point where the result are incorrectly rounded to integers. The floating point FFT can be made provably correct, see again [17], but at the cost of using larger FFT lengths. For example, for a length 2^{25} FFT, [17] allows 9 bits per `double`, whereas `Prime95` uses up to 17.76. To eliminate any chance of fatal round-off-error, the transform length and hence run-time would need to be about doubled. Also, the implementation of the FFT

[5]The word "benchmarketing" has been suggested to us by Torbjörn Granlund.

in `Prime95` is done in hand-optimized assembly for the x86 family of processors, and will not run on other architectures.

Another implementation of complex floating point FFT is Guillermo Ballester Valor's `Glucas`. The algorithm it uses is similar to that in `Prime95`, but it is written portably in C. This makes it slower than `Prime95`, but still faster than our code on both the Pentium 4 and the Opteron, as shown in Figure 4.

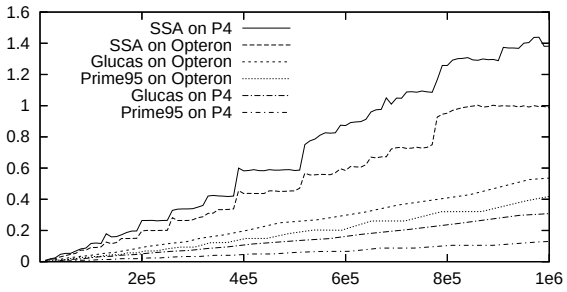


Figure 4: Time in seconds for multiplication of different word lengths with our implementation, Prime95 and Glucas on a 3.2 GHz Pentium 4 and a 2.4 GHz Opteron.

Acknowledgments.

This work was done in collaboration with Torbjörn Granlund, during his visits as invited professor at INRIA Lorraine in March-April and November-December 2006; we also thank him for proof-reading this paper. This work would probably not have been achieved without the initial stimulation from Allan Steel; the authors are very grateful to him. Thanks to Markus Hegland for pointing to the Belgian paper, and to William Hart for finding a typo in a preliminary version. We also thank the anonymous referees for many valuable remarks on the manuscript.

4. REFERENCES

[1] Arndt, J. *Algorithms for programmers (working title)*. Draft version of 2007-January-05, `http://www.jjj.de/fxt/`.

[2] Bailey, D. The computation of π to 29,360,000 decimal digits using Borwein's quartically convergent algorithm. *Math. Comp. 50* (1988), 283–296.

[3] Bailey, D. FFTs in external or hierarchical memory. *J. Supercomputing 4* (1990), 23–35.

[4] Bernstein, D. J. Multidigit multiplication for mathematicians. `http://cr.yp.to/papers.html#m3`, 2001.

[5] Bernstein, D. J. Fast multiplication and its applications. `http://cr.yp.to/papers.html#multapps`, 2004.

[6] Bernstein, D. J. Removing redundancy in high-precision Newton iteration. `http://cr.yp.to/fastnewton.html`, 2004.

[7] Brent, R. P. Multiple-precision zero-finding methods and the complexity of elementary function evaluation. In *Analytic Computational Complexity* (1975), J. F. Traub, Ed., Academic Press, pp. 151–176.

[8] Brent, R. P., and Pollard, J. M. Factorization of the eighth Fermat number. *Math. Comp. 36* (1981), 627–630.

[9] Brockmeyer, E., Ghez, C., D'Eer, J., Catthoor, F., and Man, H. D. Parametrizable behavioral IP module for a data-localized low-power FFT. In *Proc. IEEE Workshop on Signal Processing Systems (SIPS)* (1999), IEEE Press, pp. 635–644.

[10] Crandall, R., and Fagin, B. Discrete weighted transforms and large-integer arithmetic. *Math. Comp. 62*, 205 (1994), 305–324.

[11] Crandall, R., and Pomerance, C. *Prime Numbers: A Computational Perspective*. Springer-Verlag, 2000.

[12] Granlund, T. Personal communication, Dec. 2006.

[13] Harley, R. Personal communication, Jan. 2000.

[14] Karp, A. H., and Markstein, P. High-precision division and square root. *ACM Trans. Math. Softw. 23*, 4 (1997), 561–589.

[15] Knuth, D. The analysis of algorithms. In *Actes du Congrès International des Mathématiciens de 1970* (1971), vol. 3, Gauthiers-Villars, pp. 269–274.

[16] Moenck, R., and Borodin, A. Fast modular transforms via division. In *Proceedings of the 13th Annual IEEE Symposium on Switching and Automata Theory* (Oct. 1972), pp. 90–96.

[17] Percival, C. Rapid multiplication modulo the sum and difference of highly composite numbers. *Math. Comp. 72*, 241 (2003), 387–395.

[18] Schönhage, A. Schnelle Berechnung von Kettenbruchentwicklungen. *Acta Inform. 1* (1971), 139–144.

[19] Schönhage, A., and Strassen, V. Schnelle Multiplikation großer Zahlen. *Computing 7* (1971), 281–292.

[20] Steel, A. Magma V2.12-1 is up to 2.3 times faster than GMP 4.1.4 for large integer multiplication. `http://magma.maths.usyd.edu.au/users/allan/intmult.html`, July 2005.

[21] Steel, A. Reduce everything to multiplication. Computing by the Numbers: Algorithms, Precision, and Complexity, Workshop for Richard Brent's 60th birthday, Berlin, July 2006. `http://www.mathematik.hu-berlin.de/~gaggle/EVENTS/2006/BRENT60/`.

[22] van der Hoeven, J. The truncated Fourier transform and applications. In *Proceedings of the 2004 international symposium on symbolic and algebraic computation (ISSAC)* (2004), J. Gutierrez, Ed., pp. 290–296.

[23] von zur Gathen, J., and Gerhard, J. *Modern Computer Algebra*. Cambridge University Press, 1999.

[24] Woltman, G., and Kurowski, S. *The Great Internet Mersenne Prime Search*. `http://www.gimps.org/`.

[25] Zimmermann, P., and Dodson, B. 20 years of ECM. In *Proceedings of the 7th Algorithmic Number Theory Symposium (ANTS VII)* (2006), F. Hess, S. Pauli, and M. Pohst, Eds., vol. 4076 of *Lecture Notes in Comput. Sci.*, Springer-Verlag, pp. 525–542.