LibTomMath v0.04
A Free Multiple Precision Integer Library

Tom St Denis
tomstdenis@iahu.ca

February 28, 2003

1 Introduction

"LibTomMath" is a free and open source library that provides multiple-precision integer functions required to form a basis of a public key cryptosystem. LibTomMath is written entire in portable ISO C source code and designed to have an application interface much like that of MPI from Michael Fromberger.

LibTomMath was written from scratch by Tom St Denis but designed to be drop in replacement for the MPI package. The algorithms within the library are derived from descriptions as provided in the Handbook of Applied Cryptography and Knuth's "The Art of Computer Programming". The library has been extensively optimized and should provide quite comparable timings as compared to many free and commercial libraries.

LibTomMath was designed with the following goals in mind:

1. Be a drop in replacement for MPI.
2. Be much faster than MPI.
3. Be written entirely in portable C.

All three goals have been achieved. Particularly the speed increase goal. For example, a 512-bit modular exponentiation is four times faster[1] with LibTomMath compared to MPI.

Being compatible with MPI means that applications that already use it can be ported fairly quickly. Currently there are a few differences but there are many similarities. In fact the average MPI based application can be ported in under 15 minutes.

Thanks goes to Michael Fromberger for answering a couple questions and Colin Percival for having the patience and courtesy to help debug and suggest optimizations. They were both of great help!

2 Building Against LibTomMath

Building against LibTomMath is very simple because there is only one source file. Simply add "bn.c" to your project and copy both "bn.c" and "bn.h" into your project directory. There is no configuration nor building required before hand.

If you are porting an MPI application to LibTomMath the first step will be to remove all references to MPI and replace them with references to LibTomMath. For example, substitute

```
#include "mpi.h"
```

with

```
#include "bn.h"
```

[1] On an Athlon XP with GCC 3.2

Remove "mpi.c" from your project and replace it with "bn.c". Note that currently MPI has a few more functions than LibTomMath has (e.g. no square-root code and a few others). Those are planned for future releases. In the interim work arounds can be sought. Note that LibTomMath doesn't lack any functions required to build a cryptosystem.

3 Programming with LibTomMath

3.1 The mp_int Structure

All multiple precision integers are stored in a structure called **mp_int**. A multiple precision integer is essentially an array of **mp_digit**. mp_digit is defined at the top of bn.h. Its type can be changed to suit a particular platform.

For example, when **MP_8BIT** is defined[2] a mp_digit is a unsigned char and holds seven bits. Similarly when **MP_16BIT** is defined a mp_digit is a unsigned short and holds 15 bits. By default a mp_digit is a unsigned long and holds 28 bits.

The choice of digit is particular to the platform at hand and what available multipliers are provided. For MP_8BIT either a $8 \times 8 \Rightarrow 16$ or $16 \times 16 \Rightarrow 16$ multiplier is optimal. When MP_16BIT is defined either a $16 \times 16 \Rightarrow 32$ or $32 \times 32 \Rightarrow 32$ multiplier is optimal. By default a $32 \times 32 \Rightarrow 64$ or $64 \times 64 \Rightarrow 64$ multiplier is optimal.

This gives the library some flexibility. For example, a i8051 has a $8 \times 8 \Rightarrow 16$ multiplier. The 16-bit x86 instruction set has a $16 \times 16 \Rightarrow 32$ multiplier. In practice this library is not particularly designed for small devices like an i8051 due to the size. It is possible to strip out functions which are not required to drop the code size. More realistically the library is well suited to 32 and 64-bit processors that have decent integer multipliers. The AMD Athlon XP and Intel Pentium 4 processors are examples of well suited processors.

Throughout the discussions there will be references to a **used** and **alloc** members of an integer. The used member refers to how many digits are actually used in the representation of the integer. The alloc member refers to how many digits have been allocated off the heap. There is also the β quantity which is equal to 2^W where W is the number of bits in a digit (default is 28).

3.2 Calling Functions

Most functions expect pointers to mp_int's as parameters. To save on memory usage it is possible to have source variables as destinations. For example:

```
mp_add(&x, &y, &x);          /* x = x + y */
mp_mul(&x, &z, &x);          /* x = x * z */
mp_div_2(&x, &x);            /* x = x / 2 */
```

[2] When building bn.c.

3.3 Basic Functionality

Essentially all LibTomMath functions return one of three values to indicate if the function worked as desired. A function will return **MP_OKAY** if the function was successful. A function will return **MP_MEM** if it ran out of memory and **MP_VAL** if the input was invalid.

Before an mp_int can be used it must be initialized with

```
int mp_init(mp_int *a);
```

For example, consider the following.

```
#include "bn.h"
int main(void)
{
   mp_int num;
   if (mp_init(&num) != MP_OKAY) {
      printf("Error initializing a mp_int.\n");
   }
   return 0;
}
```

A mp_int can be freed from memory with

```
void mp_clear(mp_int *a);
```

This will zero the memory and free the allocated data. There are a set of trivial functions to manipulate the value of an mp_int.

```
/* set to zero */
void mp_zero(mp_int *a);

/* set to a digit */
void mp_set(mp_int *a, mp_digit b);

/* set a 32-bit const */
int mp_set_int(mp_int *a, unsigned long b);

/* init to a given number of digits */
int mp_init_size(mp_int *a, int size);

/* copy, b = a */
int mp_copy(mp_int *a, mp_int *b);

/* inits and copies, a = b */
int mp_init_copy(mp_int *a, mp_int *b);
```

The **mp_zero** function will clear the contents of a mp_int and set it to positive. The **mp_set** function will zero the integer and set the first digit to a value specified. The **mp_set_int** function will zero the integer and set the first 32-bits to a given value. It is important to note that using mp_set can have unintended side effects when either the MP_8BIT or MP_16BIT defines are enabled. By default the library will accept the ranges of values MPI will (and more).

The **mp_init_size** function will initialize the integer and set the allocated size to a given value. The allocated digits are zero'ed by default but not marked as used. The **mp_copy** function will copy the digits (and sign) of the first parameter into the integer specified by the second parameter. The **mp_init_copy** will initialize the first integer specified and copy the second one into it. Note that the order is reversed from that of mp_copy. This odd "bug" was kept to maintain compatibility with MPI.

3.4 Digit Manipulations

There are a class of functions that provide simple digit manipulations such as shifting and modulo reduction of powers of two.

```
/* right shift by "b" digits */
void mp_rshd(mp_int *a, int b);

/* left shift by "b" digits */
int mp_lshd(mp_int *a, int b);

/* c = a / 2^b */
int mp_div_2d(mp_int *a, int b, mp_int *c);

/* b = a/2 */
int mp_div_2(mp_int *a, mp_int *b);

/* c = a * 2^b */
int mp_mul_2d(mp_int *a, int b, mp_int *c);

/* b = a*2 */
int mp_mul_2(mp_int *a, mp_int *b);

/* c = a mod 2^d */
int mp_mod_2d(mp_int *a, int b, mp_int *c);
```

Both the **mp_rshd** and **mp_lshd** functions provide shifting by whole digits. For example, mp_rshd(x, n) is the same as $x \leftarrow \lfloor x/\beta^n \rfloor$ while mp_lshd(x, n) is equivalent to $x \leftarrow x \cdot \beta^n$. Both functions are extremely fast as they merely copy digits within the array.

Similarly the **mp_div_2d** and **mp_mul_2d** functions provide shifting but allow any bit count to be specified. For example, mp_div_2d(x, n, y) is the

same as $y = \lfloor x/2^n \rfloor$ while mp_mul_2d(x, n, y) is the same as $y = x \cdot 2^n$. The **mp_div_2** and **mp_mul_2** functions are legacy functions that merely shift right or left one bit respectively. The **mp_mod_2d** function reduces an integer mod a power of two. For example, mp_mod_2d(x, n, y) is the same as $y \equiv x \pmod{2^n}$.

3.5 Basic Arithmetic

Next are the class of functions which provide basic arithmetic.

```
/* b = -a */
int mp_neg(mp_int *a, mp_int *b);

/* b = |a| */
int mp_abs(mp_int *a, mp_int *b);

/* compare a to b */
int mp_cmp(mp_int *a, mp_int *b);

/* compare |a| to |b| */
int mp_cmp_mag(mp_int *a, mp_int *b);

/* c = a + b */
int mp_add(mp_int *a, mp_int *b, mp_int *c);

/* c = a - b */
int mp_sub(mp_int *a, mp_int *b, mp_int *c);

/* c = a * b */
int mp_mul(mp_int *a, mp_int *b, mp_int *c);

/* b = a^2 */
int mp_sqr(mp_int *a, mp_int *b);

/* a/b => cb + d == a */
int mp_div(mp_int *a, mp_int *b, mp_int *c, mp_int *d);

/* c = a mod b, 0 <= c < b  */
int mp_mod(mp_int *a, mp_int *b, mp_int *c);
```

The **mp_cmp** will compare two integers. It will return **MP_LT** if the first parameter is less than the second, **MP_GT** if it is greater or **MP_EQ** if they are equal. These constants are the same as from MPI.

The **mp_add**, **mp_sub**, **mp_mul**, **mp_div**, **mp_sqr** and **mp_mod** are all fairly straight forward to understand. Note that in mp_div either c (the quotient) or d (the remainder) can be passed as NULL to ignore it. For example, if you only want the quotient $z = \lfloor x/y \rfloor$ then a call such as mp_div(&x, &y, &z, NULL) is acceptable.

There is a related class of "single digit" functions that are like the above except they use a digit as the second operand.

```
/* compare against a single digit */
int mp_cmp_d(mp_int *a, mp_digit b);

/* c = a + b */
int mp_add_d(mp_int *a, mp_digit b, mp_int *c);

/* c = a - b */
int mp_sub_d(mp_int *a, mp_digit b, mp_int *c);

/* c = a * b */
int mp_mul_d(mp_int *a, mp_digit b, mp_int *c);

/* a/b => cb + d == a */
int mp_div_d(mp_int *a, mp_digit b, mp_int *c, mp_digit *d);

/* c = a mod b, 0 <= c < b  */
int mp_mod_d(mp_int *a, mp_digit b, mp_digit *c);
```

Note that care should be taken for the value of the digit passed. By default, any 28-bit integer is a valid digit that can be passed into the function. However, if MP_8BIT or MP_16BIT is defined only 7 or 15-bit (respectively) integers can be passed into it.

3.6 Modular Arithmetic

There are some trivial modular arithmetic functions.

```
/* d = a + b (mod c) */
int mp_addmod(mp_int *a, mp_int *b, mp_int *c, mp_int *d);

/* d = a - b (mod c) */
int mp_submod(mp_int *a, mp_int *b, mp_int *c, mp_int *d);

/* d = a * b (mod c) */
int mp_mulmod(mp_int *a, mp_int *b, mp_int *c, mp_int *d);

/* c = a * a (mod b) */
int mp_sqrmod(mp_int *a, mp_int *b, mp_int *c);

/* c = 1/a (mod b) */
int mp_invmod(mp_int *a, mp_int *b, mp_int *c);

/* c = (a, b) */
int mp_gcd(mp_int *a, mp_int *b, mp_int *c);
```

```
/* c = [a, b] or (a*b)/(a, b) */
int mp_lcm(mp_int *a, mp_int *b, mp_int *c);

/* d = a^b (mod c) */
int mp_exptmod(mp_int *a, mp_int *b, mp_int *c, mp_int *d);
```

These are all fairly simple to understand. The **mp_invmod** is a modular multiplicative inverse. That is it stores in the third parameter an integer such that $ac \equiv 1 \pmod{b}$ provided such integer exists. If there is no such integer the function returns **MP_VAL**.

3.7 Radix Conversions

To read or store integers in other formats there are the following functions.

```
int mp_unsigned_bin_size(mp_int *a);
int mp_read_unsigned_bin(mp_int *a, unsigned char *b, int c);
int mp_to_unsigned_bin(mp_int *a, unsigned char *b);

int mp_signed_bin_size(mp_int *a);
int mp_read_signed_bin(mp_int *a, unsigned char *b, int c);
int mp_to_signed_bin(mp_int *a, unsigned char *b);

int mp_read_radix(mp_int *a, unsigned char *str, int radix);
int mp_toradix(mp_int *a, unsigned char *str, int radix);
int mp_radix_size(mp_int *a, int radix);
```

The integers are stored in big endian format as most libraries (and MPI) expect. The **mp_read_radix** and **mp_toradix** functions read and write (respectively) null terminated ASCII strings in a given radix. Valid values for the radix are between 2 and 64 (inclusively).

4 Function Analysis

Throughout the function analysis the variable N will denote the average size of an input to a function as measured by the number of digits it has. The variable W will denote the number of bits per word and c will denote a small constant amount of work. The big-oh notation will be abused slightly to consider numbers that do not grow to infinity. That is we shall consider $O(N/2) \neq O(N)$ which is an abuse of the notation.

4.1 Digit Manipulation Functions

The class of digit manipulation functions such as **mp_rshd**, **mp_lshd** and **mp_mul_2** are all very simple functions to analyze.

4.1.1 mp_rshd(mp_int *a, int b)

If the shift count "b" is less than or equal to zero the function returns without doing any work. If the the shift count is larger than the number of digits in "a" then "a" is simply zeroed without shifting digits.

This function requires no additional memory and $O(N)$ time.

4.1.2 mp_lshd(mp_int *a, int b)

If the shift count "b" is less than or equal to zero the function returns success without doing any work.

This function requires $O(b)$ additional digits of memory and $O(N)$ time.

4.1.3 mp_div_2d(mp_int *a, int b, mp_int *c, mp_int *d)

If the shift count "b" is less than or equal to zero the function places "a" in "c" and returns success.

This function requires $O(2 \cdot N)$ additional digits of memory and $O(2 \cdot N)$ time.

4.1.4 mp_mul_2d(mp_int *a, int b, mp_int *c)

If the shift count "b" is less than or equal to zero the function places "a" in "c" and returns success.

This function requires $O(N)$ additional digits of memory and $O(2 \cdot N)$ time.

4.1.5 mp_mod_2d(mp_int *a, int b, mp_int *c)

If the shift count "b" is less than or equal to zero the function places "a" in "c" and returns success.

This function requires $O(N)$ additional digits of memory and $O(2 \cdot N)$ time.

4.2 Basic Arithmetic

4.2.1 mp_cmp(mp_int *a, mp_int *b)

Performs a **signed** comparison between "a" and "b" returning **MP_GT** is "a" is larger than "b".

This function requires no additional memory and $O(N)$ time.

4.2.2 mp_cmp_mag(mp_int *a, mp_int *b)

Performs a **unsigned** comparison between "a" and "b" returning **MP_GT** is "a" is larger than "b". Note that this comparison is unsigned which means it will report, for example, $-5 > 3$. By comparison mp_cmp will report $-5 < 3$.

This function requires no additional memory and $O(N)$ time.

4.2.3 mp_add(mp_int *a, mp_int *b, mp_int *c)

Handles the sign of the numbers correctly which means it will subtract as required, e.g. $a + -b$ turns into $a - b$.

This function requires no additional memory and $O(N)$ time.

4.2.4 mp_sub(mp_int *a, mp_int *b, mp_int *c)

Handles the sign of the numbers correctly which means it will add as required, e.g. $a - -b$ turns into $a + b$.

This function requires no additional memory and $O(N)$ time.

4.2.5 mp_mul(mp_int *a, mp_int *b, mp_int *c)

Handles the sign of the numbers correctly which means it will correct the sign of the product as required, e.g. $a \cdot -b$ turns into $-ab$.

For relatively small inputs, that is less than 80 digits a standard baseline or comba-baseline multiplier is used. It requires no additional memory and $O(N^2)$ time. The comba-baseline multiplier is only used if it can safely be used without losing carry digits. The comba method is faster than the baseline method but cannot always be used which is why both are provided. The code will automatically determine when it can be used. If the digit count is higher than 80 for the inputs than a Karatsuba multiplier is used which requires approximately $O(6 \cdot N)$ memory and $O(N^{lg(3)})$ time.

4.2.6 mp_sqr(mp_int *a, mp_int *b)

For relatively small inputs, that is less than 80 digits a modified squaring or comba-squaring algorithm is used. It requires no additional memory and $O((N^2 + N)/2)$ time. The comba-squaring method is used only if it can be safely used without losing carry digits. After 80 digits a Karatsuba squaring algorithm is used whcih requires approximately $O(4 \cdot N)$ memory and $O(N^{lg(3)})$ time.

4.2.7 mp_div(mp_int *a, mp_int *b, mp_int *c, mp_int *d)

The quotient is placed in "c" and the remainder in "d". Either (or both) of "c" and "d" can be set to NULL if the value is not desired.

This function requires $O(4 \cdot N)$ memory and $O(N^2 + N)$ time.

4.3 Modular Arithmetic

4.3.1 mp_addmod, mp_submod, mp_mulmod, mp_sqrmod

These functions take the time of their host function plus the time it takes to perform a division. For example, mp_addmod takes $O(N + (N^2 + N))$ time. Note that if you are performing many modular operations in a row with the same modulus you should consider Barrett reductions.

NOTE: This section will be expanded upon in future releases of the library.

4.3.2 mp_invmod(mp_int *a, mp_int *b, mp_int *c)

This function is technically only defined for moduli who are positive and inputs that are positive. That is it will find $c = 1/a \pmod b$ for any $a > 0$ and $b > 0$. The function will work for negative values of a since it merely computes $c = -1 \cdot (1/|a|) \pmod b$. In general the input is only **guaranteed** to lead to a correct output if $-b < a < b$ and $(a, b) = 1$.

NOTE: This function will be revised to accept a wider range of inputs in future releases.

5 Timing Analysis

5.1 Observed Timings

A simple test program "demo.c" was developed which builds with either MPI or LibTomMath (without modification). The test was conducted on an AMD Athlon XP processor with 266Mhz DDR memory and the GCC 3.2 compiler[3]. The multiplications and squarings were repeated 100,000 times each while the modular exponentiation (exptmod) were performed 50 times each. The "inversions" refers to multiplicative inversions modulo an odd number of a given size. The RDTSC (Read Time Stamp Counter) instruction was used to measure the time the entire iterations took and was divided by the number of iterations to get an average. The following results were observed.

[3]With build options "-O3 -fomit-frame-pointer -funroll-loops"

Operation	Size (bits)	Time with MPI (cycles)	Time with LibTomMath (cycles)
Inversion	128	264,083	172,381
Inversion	256	549,370	381,237
Inversion	512	1,675,975	1,212,341
Inversion	1024	5,237,957	3,114,144
Inversion	2048	17,871,944	8,137,896
Inversion	4096	66,610,468	22,469,360
Multiply	128	1,426	847
Multiply	256	2,551	1,848
Multiply	512	7,913	3,505
Multiply	1024	28,496	9,097
Multiply	2048	109,897	29,497
Multiply	4096	469,970	112,651
Square	128	1,319	883
Square	256	1,776	1,895
Square	512	5,399	3,543
Square	1024	18,991	8,692
Square	2048	72,126	26,792
Square	4096	306,269	103,263
Exptmod	512	32,021,586	7,096,687
Exptmod	768	97,595,492	14,849,813
Exptmod	1024	223,302,532	27,826,489
Exptmod	2048	1,682,223,369	142,026,274
Exptmod	2560	3,268,615,571	292,597,205
Exptmod	3072	5,597,240,141	452,731,243
Exptmod	4096	13,347,270,891	941,433,401

Note that the figures do fluctuate but their magnitudes are relatively intact. The purpose of the chart is not to get an exact timing but to compare the two libraries. For example, in all of the tests the exact time for a 512-bit squaring operation was not the same. The observed times were all approximately 3,500 cycles, more importantly they were always faster than the timings observed with MPI by about the same magnitude.

5.2 Digit Size

The first major constribution to the time savings is the fact that 28 bits are stored per digit instead of the MPI defualt of 16. This means in many of the algorithms the savings can be considerable. Consider a baseline multiplier with a 1024-bit input. With MPI the input would be 64 16-bit digits whereas in LibTomMath it would be 37 28-bit digits. A savings of $64^2 - 37^2 = 2727$ single precision multiplications.

5.3 Multiplication Algorithms

For most inputs a typical baseline $O(n^2)$ multiplier is used which is similar to that of MPI. There are two variants of the baseline multiplier. The normal

and the fast variants. The normal baseline multiplier is the exact same as the algorithm from MPI. The fast baseline multiplier is optimized for cases where the number of input digits N is less than or equal to $2^w/\beta^2$. Where w is the number of bits in a **mp_word**. By default a mp_word is 64-bits which means $N \leq 256$ is allowed which represents numbers upto 7168 bits.

The fast baseline multiplier is optimized by removing the carry operations from the inner loop. This is often referred to as the "comba" method since it computes the products a columns first then figures out the carries. This has the effect of making a very simple and paralizable inner loop.

For large inputs, typically 80 digits[4] or more the Karatsuba method is used. This method has significant overhead but an asymptotic running time of $O(n^{1.584})$ which means for fairly large inputs this method is faster. The Karatsuba implementation is recursive which means for extremely large inputs they will benefit from the algorithm.

MPI only implements the slower baseline multiplier where carries are dealt with in the inner loop. As a result even at smaller numbers (below the Karatsuba cutoff) the LibTomMath multipliers are faster.

5.4 Squaring Algorithms

Similar to the multiplication algorithms there are two baseline squaring algorithms. Both have an asymptotic running time of $O((t^2 + t)/2)$. The normal baseline squaring is the same from MPI and the fast is a "comba" squaring algorithm. The comba method is used if the number of digits N is less than $2^{w-1}/\beta^2$ which by default covers numbers upto 3584 bits.

There is also a Karatsuba squaring method which achieves a running time of $O(n^{1.584})$ after considerably large inputs.

MPI only implements the slower baseline squaring algorithm. As a result LibTomMath is considerably faster at squaring than MPI is.

5.5 Exponentiation Algorithms

LibTomMath implements a sliding window k-ary left to right exponentiation algorithm. For a given exponent size L an appropriate window size k is chosen. There are always at most L modular squarings and $\lfloor L/k \rfloor$ modular multiplications. The k-ary method works by precomputing values $g(x) = b^x$ for $0 \leq x < 2^k$ and a given base b. Then the multiplications are grouped in windows of k bits. The sliding window technique has the benefit that it can skip multiplications if there are zero bits following or preceding a window. Consider the exponent $e = 11110001_2$ if $k = 2$ then there will be a two squarings, a multiplication of $g(3)$, two squarings, a multiplication of $g(3)$, four squarings and and a multiplication by $g(1)$. In total there are 8 squarings and 3 multiplications.

MPI uses a binary square-multiply method. For the same exponent e it would have had 8 squarings and 5 multiplications. There is a precomputation

[4]By default that is 2240-bits or more.

phase for the method LibTomMath uses but it generally cuts down considerably on the number of multiplications. Consider a 512-bit exponent. The worst case for the LibTomMath method results in 512 squarings and 124 multiplications. The MPI method would have 512 squarings and 512 multiplications. Randomly every $2k$ bits another multiplication is saved via the sliding-window technique on top of the savings the k-ary method provides.

Both LibTomMath and MPI use Barrett reduction instead of division to reduce the numbers modulo the modulus given. However, LibTomMath can take advantage of the fact that the multiplications required within the Barrett reduction do not have to give full precision. As a result the reduction step is much faster and just as accurate. The LibTomMath code will automatically determine at run-time (e.g. when its called) whether the faster multiplier can be used. The faster multipliers have also been optimized into the two variants (baseline and comba baseline).

As a result of all these changes exponentiation in LibTomMath is much faster than compared to MPI.