

### Introduction

This purpose of this document is to suggest ways of porting standard quantitative financial codes to the ClearSpeed Advance™ accelerator card. The discussion is motivated with examples using the Black-Scholes pricing model and a variety of numerical methods to price vanilla options. Each example has an accompanying source code implementation.

### Assumptions

- Familiarity with ClearSpeed products, in particular the Advance™ card.
- Familiarity with Cn code, CSAPI and the vector math library.
- Familiarity with quantitative finance concepts.

## Table of contents

<b>1</b>	<b>Overview of examples</b> .....	<b>4</b>
1.1	Host-card communication .....	5
1.2	Utility files .....	5
<b>2</b>	<b>Black-Scholes analytic pricing formula</b> .....	<b>6</b>
2.1	Background .....	6
2.2	Initial code description .....	7
2.3	Parallelizing sequential code .....	8
2.4	Incorporating the vector math library .....	9
2.5	Performance and possible improvements .....	10
<b>3</b>	<b>European option pricing via binomial method</b> .....	<b>12</b>
3.1	Background .....	12
3.2	Initial code description .....	13
3.3	Parallelizing sequential code .....	14
3.4	Incorporating the vector math library .....	15
3.5	Performance and possible improvements .....	16
<b>4</b>	<b>European Option Pricing via Monte Carlo</b> .....	<b>17</b>
4.1	Background .....	17
4.2	Initial code description .....	18
4.3	Parallelizing sequential code .....	19
4.4	Incorporating the vector math library .....	19
4.5	Performance and possible improvements .....	20
<b>5</b>	<b>Asian option pricing via Monte Carlo</b> .....	<b>21</b>
5.1	Background .....	21
5.2	Initial code description .....	22
5.3	Parallelizing sequential code .....	22
5.4	Incorporating the vector math library .....	23
5.5	Performance and possible improvements .....	23
5.6	Pricing a portfolio of Asian options using C++ and CSPX .....	24

---

<b>6</b>	<b>Pricing American options via explicit finite differences</b>	<b>25</b>
6.1	Background	25
6.2	Initial code description	26
6.3	Parallelizing sequential code	27
6.4	Incorporating the vector math library	27
6.5	Performance and possible improvements	28
<b>7</b>	<b>Broadie-Glasserman random tree method</b>	<b>29</b>
7.1	Background	29
7.2	Initial code description	31
7.3	Parallelizing sequential code	32
7.4	Incorporating the vector math library	32
7.5	Performance and possible improvements	32
<b>8</b>	<b>Summary</b>	<b>34</b>
<b>9</b>	<b>Bibliography</b>	<b>35</b>
<b>Appendix A Reduction methods</b>		<b>36</b>
<b>Revision history</b>		<b>37</b>

# 1 Overview of examples

Each of the examples described in this document is driven from a host executable run from the command-line. Each host executable is built using makefile included in the example directory. A top-level makefile exists to build each of the examples and an associated library. Navigate to the finance examples install directory and then into **option-pricing**. You can build the examples on Windows by typing:

```
csmake -f Makefile
```

from a DOS prompt.

For Linux, type:

```
make -f Makefile
```

Each of the examples consists of a top-level driving routine that parses any command-line options supplied when the executable is run. Depending on the command-line options supplied, either a reference implementation, a mono implementation, a poly implementation or an optimized vector implementation will be run. Calling the executable with a `-h` option prints a help message with a list of command-line options applicable for the example executable.

Each of the examples provides a reference ANSI-C implementation, along with detailed explanation of how the reference code reflects the quoted pricing formula. The reference implementation is not optimized for any specific computer architecture, its purpose is to describe the algorithm for the pricing method.

The mono implementation runs solely on the MTAP processor and does not make use of the 96 processing elements (PEs). Consequently, mono implementations take approximately an order of magnitude longer to execute than the equivalent program executing on the host processor. The examples allow you to pass in command-line options to control the execution. See the examples' help messages for details.

The poly implementation makes use of the 96 PEs available on each CSX600 chip. Unless otherwise noted, all examples utilize both CSX600 chips on the ClearSpeed Advance card.

The vector implementation not only utilizes the 96 PEs, but also uses the vector math library that has been optimized for the ClearSpeed CSX600 architecture.

The code was not profiled during the process of porting the code to the ClearSpeed Advance card. It is always recommended that you profile your application before starting the porting process. The ClearSpeed Visual Profiler documentation [\[1\]](#) contains details of how to profile your application code both on your host and on the ClearSpeed Advance card.

**Note:** *Each example should be run from within the example's directory to ensure that the host executable can correctly locate the ClearSpeed executables. For example, on Windows one should run:*

```
cd c:\progra~1\clearspeed\cs_ev5m512_le\examples\finance\option-pricing
cd analytic\host
blackscholes_analytic.exe
```

## 1.1 Host-card communication

Each of the examples contains code to drive the Advance card and run programs on the CSX600 processor. The host source code is written in C and uses the CSAPI to set up and run the CSX600 processor and control the data transfer to and from the board.

Each run is timed using a timer accurate to one millisecond. Runtimes may vary depending on the platform used and what other tasks the OS is running. For details on the CSAPI and Cn please see [2], [3] and [4]. The performance figures listed within this document are for the following system:

Operating system	RedHat Enterprise Linux 4 64-bit
Processor	Intel Xeon 3 GHz, 2 Gb RAM
C compiler	GCC 3.4.6
SDK/runtime	3.0
Advance™ card	X620F

Unless otherwise noted, all of the examples presented here use a single Advance board and both CSX processors on the Advance™ card.

### Programming Host-card communication

The accepted way to pass data values between programs running on the host processor and the CSX600 processor is to use global mono variables in the Cn program. Once the Cn program has been compiled to csx executable file, we can use CSAPI calls to programmatically interrogate the csx file to find the address of the variable symbol. We can then load appropriate data into the address. This approach works for aggregates, allowing the host to fill structures and arrays on the host and bitwise copy the data into the mono memory.

## 1.2 Utility files

Files common to each of the examples are placed in the **Utilities** directory of the installation. These files include cross-platform timing functions, math functions and functions layered on top of CSAPI to drive the Advance card. Calling the makefile generates a library file that is linked in when building each of the examples.

## 2 Black-Scholes analytic pricing formula

### 2.1 Background

In 1973 Fischer Black and Myron Scholes published the development of the Black-Scholes (Black-Scholes-Merton) model. The assumptions underlying this model have proved hugely influential within the finance community and they still form the basis of many modern models. Based on these assumptions, Black, Scholes and Merton also derived a differential equation that describes the price of any derivative dependent on a stock.

The key assumption involved the risk-free rate of return of a portfolio consisting only of a long position on a stock and a number of short call options on that stock. To simplify, you buy a stock and simultaneously sell the option to buy that stock. In the Black-Scholes world, the value of such a portfolio (assuming that the portfolio can be instantaneously rebalanced) will change at the risk-free interest rate [6].

The full assumptions are outlined as follows.

1. The stock price follows the Ito process described by:

$$\Delta x = a(x, t)\Delta t + b(x, t)\varepsilon\sqrt{\Delta t}$$

Where  $\varepsilon$  is a random drawing from a standardized normal distribution.

2. The underlying securities are liquid and can be perfectly hedged.
3. No transaction costs are involved with buying or selling securities.
4. No dividends are paid during the lifetime of the option.
5. No arbitrage opportunities are available.
6. Trading is continuous.
7. Risk-free interest rate is constant over the lifetime of the derivative.

These assumptions led to the Black-Scholes-Merton differential equation, where the option price  $V(r, S, \sigma, t)$  is written as:

$$\frac{\partial V}{\partial t} + rS\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV$$

The authors also supplied analytic solutions to this differential for a European style derivative that may only be exercised at the termination date defined in the option contract.

$$call = S_0 N(d_1) - Ke^{-rT} N(d_2)$$

$$put = Ke^{-rT} N(-d_2) - S_0 N(-d_1)$$

Where

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(\frac{r + \sigma^2}{2}\right)T}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(\frac{r - \sigma^2}{2}\right)T}{\sigma\sqrt{T}}$$

$N(x)$  is the cumulative probability distribution function for a standardized normal distribution. In other words, it is the probability that a variable drawn from a normal distribution will be less than  $x$ .

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx$$

For the simple circumstances described (non-dividend-paying stock, and so on), the given analytic formulae can be used to provide a fair value for a **call** or **put** option.

## 2.2 Initial code description

The analytic pricing formulae are simple to translate into C code. Most standard implementations of the C math library will not provide an implementations of  $N(x)$ , the cumulative normal distribution. The approximation used for  $N(x)$  can produce dramatically varying results towards the limits of the domain, but a widely used approximation comes from Abramowitz and Stegun [7] via Hull [6].

In this example, the number of options to evaluate can be set on the command-line. The default is to evaluate 100,000 options. You can type:

```
$ ./blackscholes_analytic -s 100000 -v reference
```

to execute the host reference code. To run the mono, poly or vector versions, you substitute "mono", "poly" or "vector" for "reference".

An appropriately sized buffer is allocated for each parameter and filled with randomly generated values. Depending on the command-line options supplied, the parameters will be passed to a reference implementation running on the host or to the ClearSpeed Advance board for processing.

Since the evaluation of each option is entirely independent of every other, both CSX processors on the Advance card can be used to run 50,000 evaluations on each processor. Half of each input buffer is transferred to each CSX600 processor. The application code calls CSAPI routines that allocate memory in CSX processor's memory space and copy data into that memory space. The exact implementation can be found in `\utilities\host_run_example.c`.

The code listed in `blackscholes_analytic_mono.cn` differs from the host reference implementation only in the addition of mono global variables that constitute the interface between the Advance card and the host. The symbols for these variables can be programmatically interrogated using CSAPI calls. Once the symbol is found, data can be written to and read from these symbols. Note that in the Cn code we use global pointers that

are seemingly never initialized. We rely on the host allocating memory and initializing the pointers to point at the memory block. This is equivalent to calling malloc in the Cn code, but more efficient. The Cn implementation of malloc actually calls back to the host to request the allocation of memory (as it is the host processor that controls memory allocation) before returning control to the CSX processor.

## 2.3 Parallelizing sequential code

Although the code listed in `blackscholes_analytic_mono.cn` now runs on the Advance card, it does not use the SIMD capability of the ClearSpeed architecture. We now need to parallelize the code across the array of PEs. The easiest way to do this is to notice the data parallel aspect of this problem. We can execute the same instructions (call the same functions) on every PE, with different data sets on each PE. This way we perform 96 evaluations in parallel (per CSX600).

### Exploiting parallelism

In order to execute function or arithmetic operations on each of the PEs, rather than the mono execution unit, simply alter the qualifier for each of the variables involved in the expression. Looking at the parallelized version (listed in `blackscholes_analytic_poly.cn`), you can see that the functions are identical, except for `poly` keyword qualifying each of the appropriate variables and functions. You must also remember to include the appropriate poly header files. For each of the standard system header files such as `<math.h>` there is an equivalent `<mathp.h>`.

### Unrolling the loop

We are now executing the loop kernel 96 as many times per loop iteration. We must unroll the loop to reduce the number of times the kernel is executed. The most straightforward way to do this is to increment the loop index by `__NUM_PES__`, rather than by 1.

```
for( i = 0; i < MAX_ITERATIONS; i += __NUM_PES__ )
{
    // kernel...
}
```

### Copying data onto the PEs

Data transferred from the host into the card's memory banks is not immediately accessible from the PE array. You must copy the data into poly memory space. Likewise, you must copy data from poly memory back to the mono memory before allowing the host to attempt to read it. The data copying can be achieved by using the `memcpym2p` function for mono to poly transfers, or `memcpyp2m` for poly to mono transfers. The transfer of data from mono memory to poly can cause delays so always profile the code and attempt to minimize transfers.

```
...
src_addr = &sigma[0] + (penum+j*__NUM_PES__);
memcpym2p( &psigma, src_addr, sizeof(double) );
for( i =0; i<NUM_PASSES; i++)
{
    pvalue = BlackScholes(pCallPutFlag, pS, pX, pT, pr, psigma);
}
```

```
dst_addr = &value[0] + (penum+j*__NUM_PES__);
memcpy2m(dst_addr, &pvalue, sizeof(pvalue));
...
```

## 2.4 Incorporating the vector math library

The code is now parallelized across the PE array. However, better performance can be achieved by aggregating four operations into a single instruction issue, increasing the amount of parallelism. Performance gains in the range of 2-4x can be gained with this further vectorization. In practice, this is achieved by unrolling the inner loop. When evaluating the Black-Scholes pricing kernel this is easy, since each evaluation is independent of the others.

### Using the vector math library

The vector math library (VML) provides optimized versions of standard libm maths functions. This means that certain range-checking and exception handling has been removed to allow maximum performance. The functions supplied also operate on chunks of four words at a time. In order to use these, you must alter the variable datatype from `poly double` to `__DVECTOR`. You must also use the VML functions for each of the maths functions. The following code snippet shows how the functions must be changed.

The poly version,

```
w = 1.0 / sqrt(2.0 * PI) *
      exp(- (X*X) / 2) *
      (a1*K +
       a2*K*K +
       a3*powp(K,three) +
       a4*powp(K,four) +
       a5*powp(K,five));
```

The version using the vector math library and optimized math functions,

```
t = cs_exp((*X) * (*X) * -0.5) * 0.3989422804;
powK2 = K * K;
powK3 = powK2 * K;
powK4 = powK3 * K;
powK5 = powK4 * K;
w = t * ((K * a1) + (powK2 * a2) + (powK3 * a3) +
         (powK4 * a4) + (powK5 * a5));
```

### Unrolling the loop for vectorization

Each PE now performs four evaluations per loop cycle. We have effectively unrolled the loop a further four times.

```
for( i = 0; i < MAX_ITERATIONS; i += __NUM_PES__ * 4)
{
    // kernel...
}
```

## 2.5 Performance and possible improvements

### Runtime performance

Here we are looking at the most important indicator of performance; run time. In general, how many seconds it takes to calculate “N” samples is important, but other factors, such as power, capacity and scalability may also be significant.

You can experiment with running the `blackscholes_analytic` host executable. The version to run can be specified using the “-v” parameter and the number of samples to evaluate can be set with “-s”. Systems with more than one board can set the number of boards to use with “-b”. For example, to run the “vector” version with 1,000,000 samples on 2 boards, call:

```
$ ./blackscholes_analytic -b 2 -v vector -s 1000000
```

The version parameter accepts four values, “reference”, “mono”, “poly”, and “vector”, each corresponds to the appropriate version to run.

It is important to note the mono version simply runs the C code on the mono execution unit and therefore we see a fifty fold increase in the run time. This is not surprising; the mono execution unit is not designed for high performance arithmetic and runs around one-tenth of the clock speed of a modern processor.

```
$ ./blackscholes_analytic -v reference -s 1000000
Black-Scholes Analytic value 0.002334
1000000 samples in 3.422664 secs
$ ./blackscholes_analytic -v mono -s 1000000
Black-Scholes Analytic value 0.002334
1000000 samples in 150.904971 secs
```

The parallel nature of this problem means we can easily distribute the compute over two CSX600 processors and we can see that the poly version is already nearly as fast as the processor (and running at far lower power). Finally, exploiting the optimized maths libraries improves performance around ten times. It could be expected that moving to vector types improve the performance by a factor of four, yet it actually causes a factor of 10 improvement. This improvement is in part due to using optimized versions of math functions. Functions such as `exp`, `log` and `sin` are functions from `libm` that work on poly data-types. The vector math library provides optimized versions that work on poly or `__DVECTOR` datatypes. The optimized versions have the “cs\_” prefix.

### Improvements

There are currently 6 individual calls to `memcpym2p`, copying a total of 6\*8 bytes to each PE. Calling `memcpym2m` or `memcpym2p` has a setup overhead, so we should aim to reduce this. More importantly, we can exploit the multi-threaded nature of the array processor and overlap the compute and IO. This leads to a double-buffer approach. The outline of such an approach is detailed in the following pseudo-code.

```
// prime the pipeline by fetching 8 words from input array
async_memcpym2p( SEM_M2P,
                 &pe_buf[active_buffer][0],
                 ((mono_buf)+pe_offset),
                 sizeof(double));
__sem_wait(SEM_M2P);
```

```
while(1)
{
    // prefetch the next chunk
    async_memcpy2p(SEM_M2P,
                  &pe_buf[!active_buffer][0],
                  ((mono_buf)+pe_offset+stride) );

    __sem_wait(SEM_M2P);

    //
    // bulk of processing here...
    //

    async_memcpy2m(SEM_P2M,
                  ((mono_results)+pe_offset),
                  &pe_buf[active_buffer][0] );
    __sem_wait(SEM_P2M);

    // swap the write buffers
    active_buffer = !active_buffer;
}
}
```

After introducing this approach, the I/O time should be completely overlapped with compute and so the run time will only reflect the compute time of the algorithm.

## 3 European option pricing via binomial method

### 3.1 Background

Whilst the Black-Scholes-Merton pricing formula was an important milestone in options pricing, its applicability was limited. Particularly because it is not generally possible to accurately price early-exercise options such as American and Bermudan (in specific circumstances such as continuously-dividend-paying option it is possible to find closed-form solutions). In 1979, Cox, Ross and Rubenstein published a paper [8] applying the numerical method of Binomial Trees to pricing derivatives. They considered the pricing of an option on a non-dividend-paying stock. Assuming the option lifetime is divided into small time intervals, and at each time interval the stock value  $S$  can increase to  $uS$  with probability  $p$ , or decrease to  $dS$  with probability  $q = 1-p$ . Coefficients  $u$  and  $d$  are chosen to be recombining such that  $Sud == Sdu == S$ .

The expected value after one time step is:

$$(1) \quad pSu + (1-p)Sd$$

The expected variance of the return over  $\Delta t$ :

$$(2) \quad pu^2 + (1-p)d^2 - [pu + (1-p)d]^2$$

Setting (1) equal to the risk-free rate of return ( $Se^{r\Delta t}$ ), and (2) equal to  $\sigma^2\Delta t$ , and with rearranging the algebra, we find:

$$p = \frac{e^{r\Delta t} - d}{u - d}$$

$$u = e^{\sigma\sqrt{t}}$$

$$d = e^{-\sigma\sqrt{t}}$$

A European option can be priced using the following process.

1. Generate a tree of depth  $n$ , with the underlying stock prices at each node.
2. At expiry time, in other words at the leaves of the tree, optimally exercise the option at time  $T (=t_n)$ .
3. Step backwards in time, using the option value at  $t_{i+1}$  to find the value at  $t_i$ .
4. Repeat until the option price at  $t_{i=0}$  is found.

It is apparent that a similar algorithm can be used to price early exercise options. At each time-step we check whether it is optimal to exercise at that time.

*Note: The definitions of  $u$  and  $d$  create a recombining tree. Recombining trees have helpful properties in that it is simple to calculate the price at each node and there are  $n+1$  nodes at depth  $n$  rather than  $2^n$  nodes. The depth of the tree can be chosen according to the required accuracy of the pricing. As the number of steps approaches infinity, the value converges to the analytic price.*

### 3.2 Initial code description

When generating the pricing tree, it is unnecessary to simultaneously hold all the values in memory. By working on a single stripe at any time step, the amount of memory required can be reduced to just the applicable “number of leaves”.

To simplify the code, allocate a maximum possible array size at compile time. We can check at run time that the requested number of time steps does not exceed this.

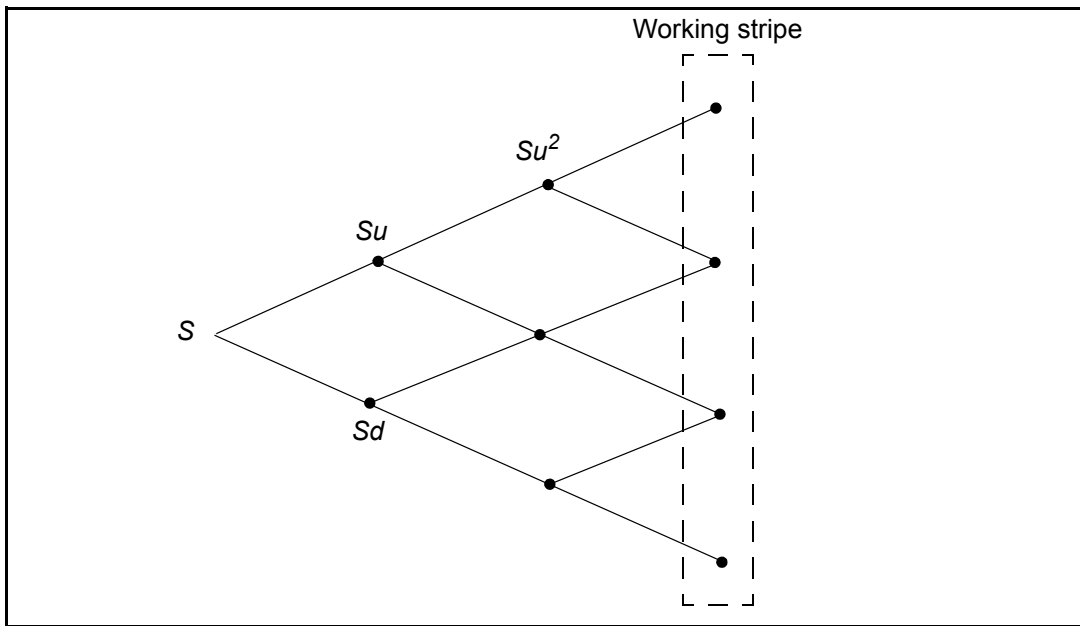
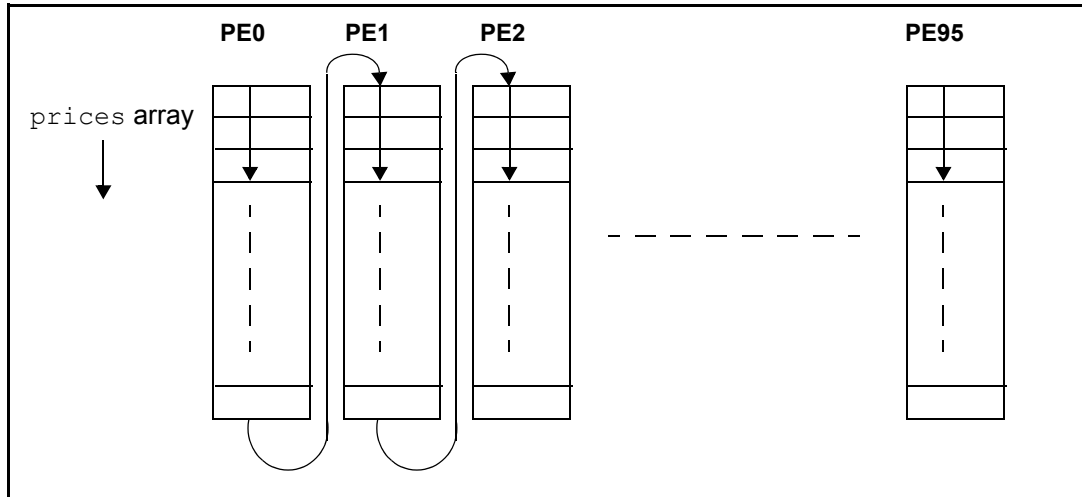


Figure 1. Tree configuration

The first two loops set up the underlying price at the leaves of the tree, and then fill in the call payoff for each of the prices. Once the boundary conditions contain the expiry date, the algorithm steps backwards in time updating the price array, holding the possible payoff values. As the time regresses to  $t=0$ , the width of the tree narrows. At time  $t=0$  the result is held in `call_values[0]`.

### 3.3 Parallelizing sequential code

The array of `prices` and `call_values` is blocked onto the PE array.

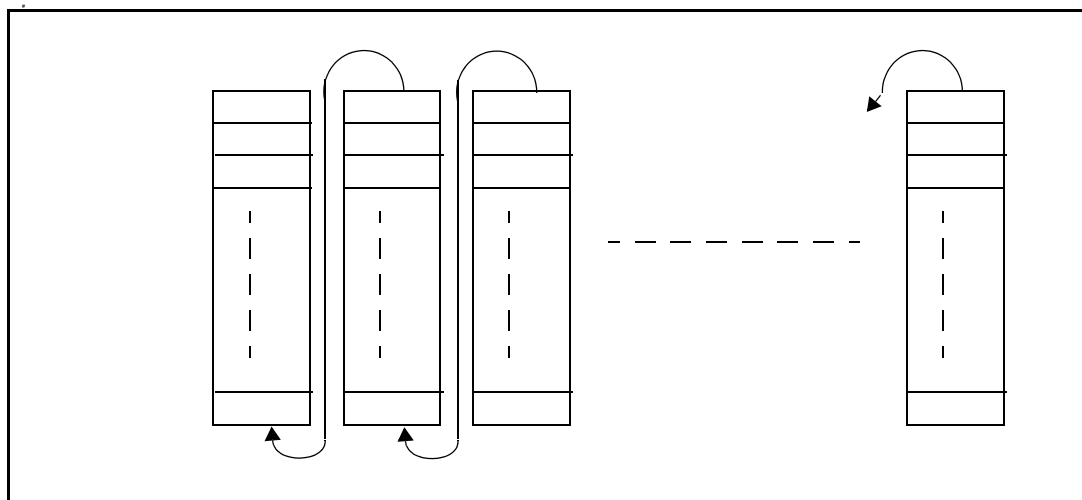


**Figure 2. Blocking the array of prices onto the PE array**

The data dependency introduced by the code:

```
for (i=0; i<=step; ++i)
{ // value of call_values[i] dependent on call_values[i+1]
  call_values[i] = (p_up*call_values[i+1] +
    p_down*call_values[i])*Rinv;
  prices[i] = d*prices[i+1];
  // ... exercise payoff
}
```

means that the data must move across the PE array:



**Figure 3. Moving data across the array**

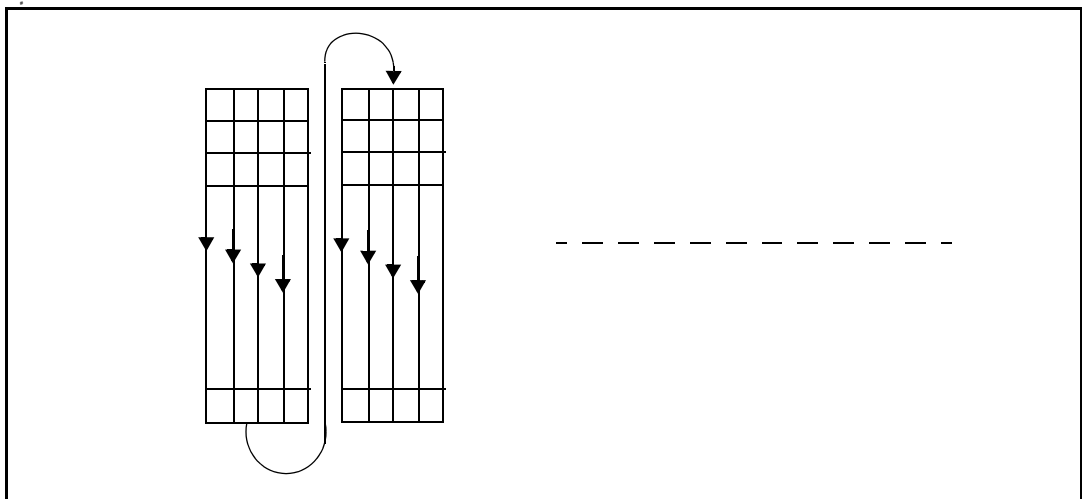
This is achieved using “swizzle” functions. Swizzling transfers data between adjacent PEs, so a `swizzle_down_double` call transfers a double word from PE `n` to PE `n-1`.

The basic parallelism strategy is intuitive and the most efficient. It is possible to consider other strategies that incorporate pipelining between PE elements but these require moving  $O(n)$  data between each PE.

There are other possibilities to consider when laying the data across the PEs. For instance, keeping data local to PEs rather than striping it across the PE array. For large trees this is effective because when processing the leaves of the tree, local data can be worked on without the need for swizzling. However, there are several arguments against this data-layout tactic. Firstly, data will always need to be communicated through the PEs at some point in tree processing. So whilst initially forgoing swazzles, the eventual swazzling of data means marginally more complicated code. Secondly, typical tree sizes are not large enough to warrant this tactic.

### 3.4 Incorporating the vector math library

In order to make use of the vector instructions, the data layout must be altered again. Making best use of the vector instructions means that there can not be any data dependencies within a variable of vector datatype. The data can, therefore, be laid out as in [Figure 4: Moving data across the array](#).



**Figure 4. Moving data across the array**

Where the data in the first element of the price array consists of the tuple  $\{0, chunk\_size, chunk\_size*2, chunk\_size*3, chunk\_size*4\}$  and so on. The result ends up in the first vector element of the first element of the `call_values` array.

## 3.5 Performance and possible improvements

### Performance

Performance of the binomial tree method is bounded by the data transfer of input parameters to the Advance card. However, as the number of steps is increased (increasing the amount of computation relative to the initial data transfer) the performance improves.

It is worth noting that only one of the CSX600 processors on the Advance card is used in this example. In theory the problem is parallelizable across multiple chips (or even multiple cards). However, in practice, data-coherency issues associated with spreading the tree over multiple chips may mean that it is not efficient. The binomial tree can be sliced across multiple processors, but data will always need to be exchanged across the boundary. For processors on the same Advance card, memory can be shared, but this is complex. For CSX600 processors on different cards, exchanging data (across a PCI-X or PCIe bus) can be a costly operation. An alternative way to use both CSX600 processors and double the performance, would be to run two separate problems at the same time, one on each CSX600.

After running the executable, you should see the following output.

```
$ ./euro_binomial -v reference
European Put (Binomial Method) value: 1.030803
European Put (Binomial Method) runtime: 3.979188 secs
$ ./euro_binomial -v vector -b 1
Note:      Using 1 of 2 available processors
European Put (Binomial Method) value: 1.030803
European Put (Binomial Method) runtime: 0.101121 secs
```

This corresponds to a 40x acceleration. However, this acceleration depends on the number of on the number of steps in the binomial tree. Increasing the number of steps in the tree increases the accuracy of the option calculation, converging to the Black-Scholes analytic price. As the number of steps in the tree is increased, the data transfer time becomes a smaller proportion of the run time. For vanilla options in the Black-Scholes world there is little need for very fine-grain trees, but for complicated models, options or calculating the "Greeks", it can be necessary to use more steps.

## 4 European Option Pricing via Monte Carlo

### 4.1 Background

Thus far we have examined two techniques for pricing a European option under a Black-Scholes framework. We will now turn to a third technique, that of Monte Carlo integration (or simulation). Monte Carlo integration techniques are often used when traditional numerical integration techniques fail because of the high dimensionality of the underlying function or because the underlying behaviour is stochastic in nature.

Performing a Monte-Carlo simulation is conceptually very simple; it involves drawing a series of sample points randomly from a particular distribution. By the law of large numbers, in the limit, the arithmetic average of the function at these points converges to the true integral.

Imagine a function  $f(x)$ , to be integrated over the range  $[0,1)$ .

$$(1) \quad I = \int_0^1 f(x) dx$$

If we approximate the integral as  $E[f(\mathbf{U})]$ , we can write an equivalent Monte Carlo estimate.

$$(2) \quad E[f(\mathbf{U})] = \frac{\sum_{i=1}^N f(u_i)}{N}$$

where  $\mathbf{U}$  is a set of uniformly distributed random numbers.

The error of the estimate is normally distributed and diminishes as  $O(N^{-1/2})$ . To simplify, increasing the quantity of random numbers used improves the estimate. Adding an extra decimal place of accuracy requires 100 times as many points.

#### The Monte Carlo Algorithm

In general, running a Monte-Carlo simulation to price an option is simple.

1. Generate a uniformly distributed random value,  $u$ .
2. The underlying stock price is a stochastic variable and its variation is lognormal. Therefore transform the uniform variate to a normally distributed variate (with mean = 0.0 and standard deviation = 1.0).
3. Use the normal variate in the discrete Stochastic Differential Equation, and apply the payoff to give a value.
4. Accumulate the value. Repeat steps 1. to 4. for a suitably large number,  $N$ .
5. Scale the accumulated values and apply the discount factor.

Steps 1. to 4. constitute the simulation kernel.

We use the Euler discretization of the stochastic differential equation:

$$S(t + dt) = S(t) \exp[(r - \sigma^2)dt + (\sigma \cdot \varepsilon) \cdot \sqrt{dt}]$$

A European option can only be exercised at the expiry date, so the dt is equal to the option period. eps is a normally distributed random variable, and sigma is the appropriately scaled volatility.

## 4.2 Initial code description

### Generating random numbers

At the core of any Monte Carlo simulation is a random number generator. We require good quality random numbers, that have certain properties and must pass certain statistical tests. We use the Mersenne Twister, a Pseudo Random Number Generator (PRNG) developed by Makoto Matsumoto and Takuji Nishimura - widely accepted to be one of the most robust PRNGs available. The reference implementation uses code taken directly from the originators' website [14].

PRNGs tend to generate uniform random numbers, that is, the random values are uniformly distributed over a range. In order to simulate an asset that follows a log normal process, we must transform the uniform distribution into a normal distribution. We achieve this through the Polar Box-Muller transform [8]. This is implemented in the function gaussrand().

The remainder of the code in the kernel performs the math for the discretisation equation above.

### Calculating the error

The estimate returned from a Monte Carlo simulation should always be accompanied by an error value. This gives an indication of how closely the estimate has converged to its true value. To this end, we sum the squared estimate of the option price. Using sum-of-squared - estimates and the sum-of-estimates allows us to calculate a measure of confidence. The confidence measure is a range within which we are 95% certain the true value will fall.

$$\text{Error} = \left( \frac{(1.96 \cdot \text{stddev})}{\sqrt{N}} \right)$$

## 4.3 Parallelizing sequential code

The mono version of this example uses identical C code to that list in the host reference implementation. Parallelizing most of this application code is trivial; the hard part is modifying Mersenne Twister algorithm (MT) for the CSX architecture. To this end ClearSpeed provide the Mersenne Twister PRNG as part of the ClearSpeed Random Number Library. The code is listed in `euro_mc_poly.cn`.

By the originators' own admission, MT is not suitable for parallel architectures. They suggest using "Dynamic Creation" of multiple MTs. More details can be found in [14]. It is this approach that the ClearSpeed Random Number Library takes. As a consequence of this approach MT streams generated on the CSX600 processor will not be bit identical to a single MT stream generated on the host. What is guaranteed, is that each stream running on the CSX600 is independent and identical distributed (i.i.d) so no bias will be introduced.

The approach for parallelizing the application is very simple and efficient. In this Monte-Carlo simulation each individual simulation is independent of the others. Therefore, we can perform the simulation kernel on each PE simultaneously. This is can be thought of as "unrolling" the outer-loop and performing 96 simulations at a time.

In order to run a kernel function on each PE, the functions must be converted to poly functions to operate exclusively on poly variables. For the most part we will be using math functions supplied by the ClearSpeed Standard Math Library or the ClearSpeed Vector Math Library; all the functions have a poly equivalent.

We have also inlined the call to the `gaussrand` function. Inlining will often improve performance, especially in loop kernels. The `gaussrand`, particularly so, because it is generates two variates every other call, and stores one of them. If we assume there will always be a sufficient number of simulations to perform, we can unroll the simulation loop and generate and consume two random variates per pass of the loop.

### Result reduction

The simulation loop accumulates results on each PE. At the end of the loop, 96 sub-results remain. In order to obtain a final result, the sub-results must be reduced to a single result. There are many ways to do this, but an efficient method is to use the swizzle path and accumulate data in mono memory as it is moved sideways. Once the fully accumulated result is in mono memory it can be transferred to the host. See Appendix A for more on reduction methods.

## 4.4 Incorporating the vector math library

The code is now parallelized across the PE array. However, better performance can be achieved by aggregating four operations into a single instruction issue, increasing the amount of parallelism. In practice, this can be achieved by further unrolling the inner loop. For a Monte-Carlo simulation this is simple as each evaluation is independent of the others.

### Using the vector math library

The VML provides functions that operate on chunks of four words at a time. In order to use these, the variable data-type must be altered from `poly double` to `__DVECTOR`. You must also use the VML functions for each of the math functions.

### Unrolling the loop

So long as the number of simulation paths is further divisible by two, we can further unroll the outer loop. The random number generator library supplies the function `cs_vdgaussian(...)` which generates two normally distributed random variates via the Box-Muller method.

The last statement in the simulation kernel reduces the two results for this individual simulation from a `__DVECTOR` to a simple poly variable. It is this poly variable that keeps the running total over the simulation loop.

## 4.5 Performance and possible improvements

### Performance

The mono version of the example uses a single chip on the Advance card. This is to show that the results are identical to the reference implementation, calculated on the host. It does mean that the mono version will take a very long time.

```
$ ./euro_mc -v reference
Confidence interval = (12.848152, 12.848157)
European Call (Monte-Carlo Method) value: 12.848155
European Call (Monte-Carlo Method) runtime: 4.708328 secs

$ ./euro_mc -v mono
Using 1 of 2 available processors
Confidence interval = (12.848152, 12.848157)
European Call (Monte-Carlo Method) value: 12.848155
European Call (Monte-Carlo Method) runtime: 344.840954 secs
```

For the poly and vector implementations, we swap out the serial Mersenne Twister implementation and use ClearSpeed library implementation.

*Note: As described above, we use the dynamic creation of Mersenne Twisters so at this point the results may be different from the reference implementation.*

```
$ ./euro_mc -v vector
Confidence interval = (12.845707, 12.845712)
European Call (Monte-Carlo Method) value: 12.845709
European Call (Monte-Carlo Method) runtime: 0.125705 secs
```

However, the performance of Monte-Carlo examples again shows good improvement once fully converted to use the high performance vector math libraries.

The vector version shows an increase of speed of approximately 40x speedup over the host version. Due to the nature of this algorithm, this scales perfectly onto multiple cards.

### Improvements

ClearSpeed supplies a random number library with various common PRNGs, including `rand48`, `mcg59` and Mersenne Twister 19937; see [2] and [3] for details.

## 5 Asian option pricing via Monte Carlo

### 5.1 Background

An Asian option is a path-dependent option; that is, its value depends on the values that the underlying asset takes during the course of the option lifetime. This is in contrast to a European-style option that only depends on the value of the asset at expiry time. The Asian option has a payoff that is a function of the average of the underlying stock over some specified period of time. If the average is geometric and the stock follows a lognormal process, analytic solutions can be found, since the geometric average of  $N$  lognormal values is itself lognormal. However, if the average is arithmetic such solutions can not be found and numerical approximations must be used. A commonly used numerical technique is Monte-Carlo integration (or simulation). Details of performing a Monte Carlo simulation can found in the [Section 4: European Option Pricing via Monte Carlo](#).

Many techniques exist for improving the convergence of the error bound of the estimate, including antithetic variables, control variates, and stratified sampling [9].

Pricing of an arithmetically averaged Asian option can be helped significantly by the addition of a control variate. Assume derivative A is similar to derivative B, but an analytic solution is available for B. If both derivatives are simulated in parallel using the same random number stream (ie they have the same variance and perfect correlation); the resultant estimate for A can be improved upon by adding the error between the different prices of B. In this example, control variate is the analytically determined geometric average Asian option price.

$$V_A = V^*A - (V^*B - V_B)$$

where the estimate of the option prices are denoted with  $*$ .

We follow a similar set of steps to those used in pricing a European Option via Monte Carlo; this time, however we must also price the control variate so we end up doing approximately twice the computation.

1. Price the control variate analytically
2. Generated a normally distributed random variate.
3. Use the normal variate to evolve the asset and apply the payoff
4. Use the same normal variate to price the control variate
5. Calculate the difference between the analytic control variate value and its simulated value
6. Add this difference to the simulated value of option we are pricing

Repeat steps 2-6

With the addition of path dependency, average options can be priced using the same framework. In order to generate the average, sample the stock's path over  $M$  points in time.

The Monte Carlo algorithm for pricing Asian option is implemented in two examples, `monte-carlo` and `monte-carlo-cspx`. You can browse the reference implementation in `monte-carlo/asian_mc_reference.c` or `monte-carlo-cspx/portfolio/asianoption.cpp`. The C version models a single Asian option and uses the CSAPI interface to communicate with the ClearSpeed accelerator. The C++ version models a portfolio of Asian options and uses the CSPX interface to communicate with the accelerator.

## 5.2 Initial code description

### Unique RNG Seeds

It is desirable to have each simulation use a unique number, in order to ensure that no bias is introduced into the Monte-Carlo estimate. To this end, each chip uses a different seed for `srand48(. . .)`. It is possible to find seeds that will ensure unique values are generated over the entire simulation.

### Generating time-varying paths

The inner loop generates the M-point path for each of the N paths simulated. The inner loop simulates both the geometric and the arithmetic price path evolution. The geometric average is defined as:

$$\sqrt[M]{\prod_i S(i)}$$

To remove the need for an Mth root calculation, we can accumulate the logarithm of the stock price path instead.

As well as accumulating the call payoff, we also accumulate the square of the payoff to enable an estimate of the error bound of the simulation. The actual computation of the confidence level<sup>(1)</sup> is best executed on the host, rather than on the slow mono unit of the CSX600.

## 5.3 Parallelizing sequential code

The simplest strategy for parallelizing the code is also the most efficient. In this Monte-Carlo simulation each individual simulation is independent of the others. Therefore, we can perform the simulation kernel on each PE simultaneously, so that each PE performs a number of simulations. In order to do this, the functions called on each PE must be converted to poly functions. In this example we use a very simple PRNG, `drand48`. In comparison with the Mersenne Twister used in the European Option pricing via Monte Carlo example, `drand48` can be parallelized relatively easily. Again we use the version from the ClearSpeed Random Number Library. The code can be found in `asian_mc_poly.cn`.

### Result reduction

We can use the same reduction method as was used in the European Option Pricing via Monte Carlo example. Forth-coming versions of the SDK will include standard, optimised reduce functions.

---

1. The confidence level is chosen as a representation of the error on the Monte Carlo simulation,

## 5.4 Incorporating the vector math library

The code is now parallelized across the PE array. However, better performance can be achieved by aggregating four operations into a single instruction issue, increasing the amount of parallelism. The process is explained in the previous example, European Option Pricing via Monte Carlo.

## 5.5 Performance and possible improvements

### Performance

The performance of Monte-Carlo examples again shows good improvement once fully converted to use the high performance vector math libraries. The host program also prints a confidence level associated with the final answer.

The test shows a greater speed-up for an Asian option compared with the speed-up for a European option. Why might this be? The answer lies in the amount of computation. The ClearSpeed accelerator is primarily a math accelerator and thus the more compute available the greater the speed-up. When simulating an Asian option we actually simulate two options; the arithmetic average option and its control variate, the geometric option.

The values printed for each version may not be exactly the same. These discrepancies arise from the differences in random number generators. This can be tested by overriding `drand48()` to return a constant in the range  $[0.0,1.0)$ . The discrepancies should not be large, but if further accuracy is needed, simply increase the number of simulations. As previously stated, to improve the accuracy by one decimal place requires 100 times as many simulations. The quality of random numbers produced by `rand48` is not high, so it may be preferable to use a Mersenne Twister function. ClearSpeed provide such a function, see [3] for details. Alternatively, generate random variates on the host and stream them across to the Advance card.

```
$ ./asian_mc -b 1 -v vector
Confidence interval with control variate = (5.378237, 5.378252)
Asian Call (Monte-Carlo Method) value: 5.378244
Asian Call (Monte-Carlo Method) runtime: 0.143239 secs
```

The vector version shows an increase of speed of approximately 50 times over the host version. Due to the nature of this algorithm, this scales perfectly onto multiple cards.

### Improvements

ClearSpeed supplies a random number library with various common PRNGs, including `rand48`, `mcg59` and Mersenne Twister 19937. For further details of PRNGs provided by ClearSpeed, see [2] and [3]. The `monte-carlo-cspx` example uses the `mcg31m1` PRNG and the optimized Gaussian transformation included in the ClearSpeed random number library.

## 5.6 Pricing a portfolio of Asian options using C++ and CSPX

The `monte-carlo-cspx` example uses the same underlying algorithm as the standard `monte-carlo` example, however it does this in a more generic fashion intended to illustrate how one might go about pricing a portfolio of options. It also uses the CSPX interface library to provide a layer of abstraction from the communications with the ClearSpeed accelerator, see the CSPX documentation [5] for more details.

The example begins by creating a Portfolio object and adding a number of AsianOption objects to this portfolio. The portfolio consists of both call and put options with a variety of strike prices.

To value the portfolio using the reference code the example will call the `Portfolio::value()` method. This in turn calls the abstract `Instrument::value()` method on each of the options, in this case the options are all AsianOption objects and hence `AsianOption::value()` method, which implements the reference algorithm is called.

To value the portfolio using the ClearSpeed accelerator, the example creates an `AccelerationEngine` object and calls `AccelerationEngine::value()` passing the portfolio as a parameter.

The `AccelerationEngine::value()` method first extracts the options from the portfolio as structures of parameters ready in an array (`std::vector`). The example then creates a `CSPX::State` object to connect to the board(s) and creates `CSPX::Object` objects to allow data to be transferred to/from the board.

The `CSPX::Object` objects are migrated to the board and the 'valuePortfolio' function on the board is called. This call is blocking (CSPX also allows for non-blocking calls to the ClearSpeed accelerator) and so once the coprocessor has completed the valuation function the host code will continue. The objects are migrated back to the host and the results are extracted into the correct locations.

Note that the valuation of the portfolio takes each option in turn and runs independent Monte Carlo simulations. An alternative, faster approach would be to share the random numbers and value all the options along the same path.

## 6 Pricing American options via explicit finite differences

### 6.1 Background

Previous examples in this document have used the Black-Scholes model to find a numerical solution for pricing an option. Another obvious approach is to solve the partial differential equation itself. Solving PDEs has been the topic of research for many years in many disciplines, including fluid dynamics and structural mechanics, so many techniques are available. Please refer to the large quantity of literature available, particularly in the field of quantitative finance. An excellent introduction is found in [11], extensively examined in [12].

We want to solve the Black-Scholes differential equation

$$\frac{\partial V}{\partial t} + rS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV$$

At time  $t=0$ , discretize asset price and time onto a Cartesian grid.

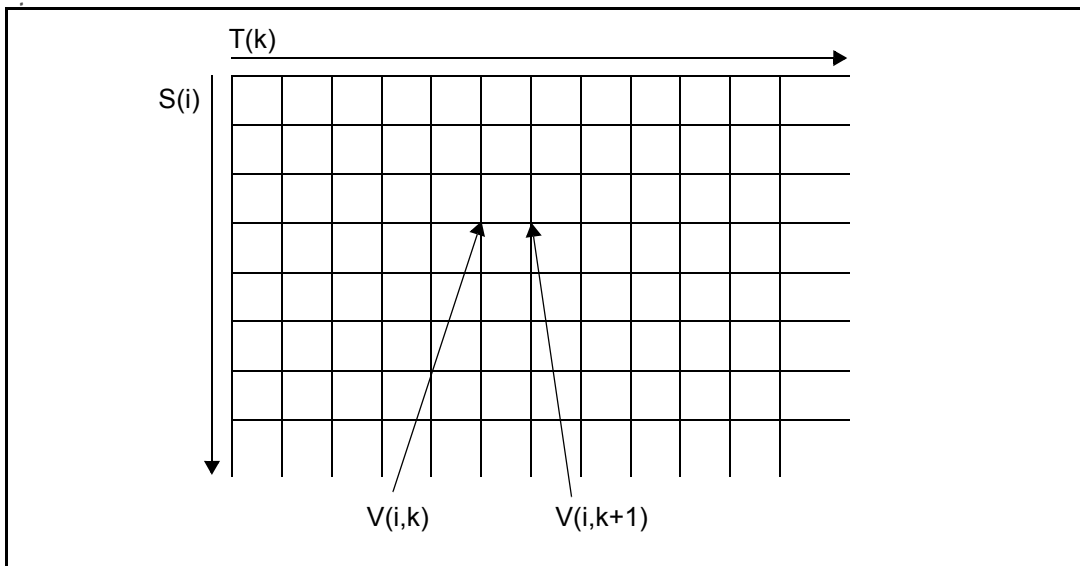


Figure 5.Asset and time discretization grid

The first order approximation to the time-derivative on the grid is

$$\frac{\partial V}{\partial t}(S, t) \approx \frac{V_i^k - V_i^{k+1}}{\delta t}$$

Similarly, the approximations for Delta (the first derivative with respect to S) and Gamma (the second derivative with respect to S) using central differences are as follows

$$\frac{\partial V}{\partial S}(S, t) \approx \frac{V_{i+1}^k - V_{i-1}^k}{2\delta t}$$

$$\frac{\partial^2 V}{\partial S^2}(S, t) \approx \frac{V_{i+1}^k - 2V_i^k + V_{i-1}^k}{\delta t^2}$$

The overall approximation for the Black-Scholes equations is:

$$(1) \quad \frac{V_i^k - V_i^{k+1}}{\delta t} + \alpha_i^k \left( \frac{V_{i+1}^k - V_{i-1}^k}{2\delta S} \right) + \beta_i^k \left( \frac{V_{i+1}^k - 2V_i^k + V_{i-1}^k}{\delta S^2} \right) + \gamma_i^k V_i^k = O(\delta t, \delta S^2) \equiv 0$$

The coefficients  $\alpha$  and  $\beta$  are functions of the discretized asset value ( $i\delta S$ ). This approximation can be rearranged to find the value of the option at the next time-step  $V_i^{k+1}$  from values of the option at the current time-steps  $V_i^k$ ,  $V_{i-1}^k$  and  $V_{i+1}^k$ . This is the explicit finite differences method. Analysis of the convergence of parabolic PDEs (of which this is an example) under the explicit method shows that the time-step size cannot be chosen arbitrarily. Convergence depends on the values of coefficients  $\alpha$ ,  $\beta$ ,  $\gamma$  and the size of asset grid step. The following inequality must be observed:

$$\delta t \leq \frac{1}{\sigma^2} \left( \frac{\delta S}{S} \right)^2$$

Typically the accuracy of the option price result is chosen by setting the size of the asset-price step. This places a bound on the size of time-step and the number of iterations necessary to price the option.

## 6.2 Initial code description

The size of the 1-d grid is set by `NumSteps`, and first two loops set up the grid of asset prices and the values in coefficients  $a$ ,  $b$ , and  $c$ . Equation (1) has been re-arranged to group the coefficients of  $V_i^k$ ,  $V_{i-1}^k$  and  $V_{i+1}^k$ . These coefficients are represented by the arrays  $a$ ,  $b$  and  $c$  in the reference code (`american_pde_reference.c`) and the mono code (`american_pde_mono.cn`).

The third loop sets up the boundary condition that corresponds to the application of a put at the expiry date.

$$\frac{\partial V}{\partial t}(S, T) = \max(0.0, X - S)$$

Finally, the main loop iterates backwards, from the solution at  $t=T$  to  $t=0$ . In this specific example, the loop could iterate forwards, as there are no time dependencies in coefficients. However, if we were to extend the model to include time variations in the coefficients  $a$ ,  $b$ ,  $c$ , such as using data from a yield curve, we must realize we are solving backwards in time and therefore iterate backwards. When pricing an American style option that allows exercise at any point in time, the put option must be exercised at each time-step.

The price of the option at  $t=0$ , is chosen at the centre of the grid, because that point corresponds to the original starting value of  $S$ .

### 6.3 Parallelizing sequential code

The simplest way to make use of the PE array is to "block" the 1-d grid over the PE array in as explained in the binomial tree example. `NumSteps` are chosen such that each PE holds a unique value in the grid. The time stepping kernel calculation requires option values from two adjacent points in the grid. These are supplied via swazzles up and down the array. The boundary conditions at each time-step are applied to the PEs at both ends of the PE array.

The code to apply the boundary conditions requires a `poly-if` statement. Instructions within a branch conditional on a poly value are predicated rather than conditionally executed. This means that the instructions are always executed on every PE, but on PEs where the poly condition is false, variable values are not updated. In the example below, the number of instructions executed on every PE is a sum of the instructions executed in all the branches, that is,  $2 + 1 + 20 = 23$ .

```
if( penum == 0 )
{
    // 2 instructions
}
else if( penum == __NUM_PES__-1)
{
    // 1 instruction
}
else
//if( (penum != 0) && (penum !=95) )
{
    // 20 instructions
}
```

### 6.4 Incorporating the vector math library

There is very little point in using vector instructions in this example in its current form. Using vector instructions normally requires unrolling loops to expose further parallelism. However, the process of converting code to run on the PE array has exposed all available parallelism.

## 6.5 Performance and possible improvements

The explicit method PDE solver is similar in some ways to trinomial tree methods (which are a direct extension of the binomial tree methods covered in the binomial tree example). The explicit method also suffers from similar problems. Once the inner loop (that updates the discretized grid) has been unrolled, there is no more parallelism that can be exploited as it is impossible to unroll the time-stepping loop. Parallelizing the problem across two CSX600s on an Advance card can also cause problems because of the data sharing needed at the boundary between data on each processor.

```
$ ./american_pde -v reference
American Put (Explicit Finite Differences) value: 4.068666
American Put (Explicit Finite Differences) runtime: 0.063052 secs
$ ./american_pde -v poly
Note:      Using 1 of 2 available processors
American Put (Explicit Finite Differences) value: 4.068666
American Put (Explicit Finite Differences) runtime: 0.077492 secs
```

The problem for this particular grid size is that it runs at about 60% of the host speed. However, because only one of the CSX600 processors is utilized, it is possible to run a separate query on the other processor, therefore doubling the effective processing rate.

It is possible to increase the accuracy of the result by increasing the resolution of grid. We can block the data arrays across the PE array just as we did in the binomial example and this would show a performance improvement on a par with the binomial example.

For certain options (digitals, or barriers) increasing the step size can prevent oscillations, giving a stable result. However, the explicit method still places a bound on the time step size with respect to the asset step size. Doubling the asset step size means the time step size must quarter in size.

## 7 Broadie-Glasserman random tree method

### 7.1 Background

Pricing an American option in a Monte-Carlo framework is a substantially more complex problem compared with pricing path-dependent options ([Section 5: Asian option pricing via Monte Carlo on page 21](#)). The extra difficulty comes because when attempting to optimally solve the option, we must solve a problem known as an optimal stopping problem. This introduces not only conceptual difficulties, but requires significantly more computation. Most methods work with Bermudan options and assume that in the limit these will tend to an approximation of an American option.

The value of an American option at time  $t=0$  over the lifetime  $0 < t < T$ , with the strike  $K$ , and  $S$  the underlying asset, is:

$$V(0) = \sup E[\max(e^{-rt}(K - S(t)), 0)]$$

Where  $\sup$  is the maximum value over the range of the function.

Given the expected value of the option at time  $t_{i+1}$ , and the value if the payoff is exercised at  $t_i$ , a decision can be made at each time-step to exercise if  $t_i$  is optimal. The value of the option when not exercising is called the continuation value.

There are various approaches to finding the expected future (time-discounted) value of the option given its current state. Longstaff and Schwartz use a series of fitted polynomial basis functions to give the predicted future value [\[13\]](#). We use Broadie and Glasserman's random tree approach. This scheme is quite complex so is only covered briefly in this document. Readers are referred to [\[14\]](#) for further information.

In any approach to estimating the continuation value, there are two sources of systematic error leading to bias. The first results from the use of future information in making the decision to exercise (which is inevitable in schemes that work backwards along simulated paths); this leads to high bias. The second approach however, provides a low bias. This results from a sub-optimal exercise decision.

The valuation of the option at the expiry date is insignificant, the item of interest ultimately is its value at  $t = 0$ . A dynamic programming formulation can be used to recursively determine the value at  $t = 0$ . The option  $V_i(\chi)$  is the value of the option at  $t_i$  given the state  $X_i = \chi$ .

$$V_m = h_m(\chi)$$

$$V_{i-1}(\chi) = \max\{h_{i-1}(\chi), E[D_{i-1,i}(\chi_i)V_i(\chi_i) | X_{i-1} = \chi]\}$$

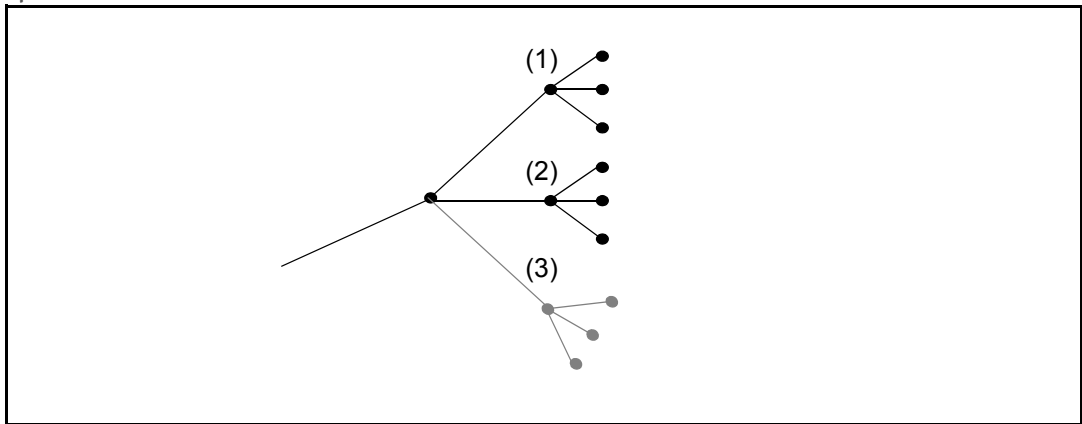
$$i = 1, \dots, m$$

$D_{i-1,i}$  is the discount factor from  $t_{i-1}$  to  $t_i$  and  $h_m(\chi)$  is the payoff.

The random tree method involves simulating a tree of paths; at each node  $b$  independent successor states are generated, where  $b$  is the branching factor ( $\geq 2$ ). The high and low estimators are defined by backward induction, starting at  $t = t_m$  - the expiry date. At the terminal nodes

$$\tilde{V}_m = h_m(\tilde{X}_m)$$

Where the  $\sim$  notation implies payoff is applied over each of the terminals. The tree can be visualized as in [Figure 6](#).



**Figure 6. Multinomial tree with branching = 3**

For the high estimator, apply backward induction and the nodes at level  $m-1$  in the tree are valued as

$$\tilde{V}_{m-1} = \max \left\{ h_{m-1}(\tilde{X}_{m-1}), \frac{1}{b} \sum_{j=1}^b \tilde{V}_m \right\}$$

This can be intuitively understood as the arithmetic average of the future value over each successor branch. The low estimator is more complicated and calculating it is split into two parts. The continuation value is defined as

$$v = \alpha, \text{ if } Y_1 \leq \alpha \text{ else } Y_2$$

Where  $Y_1, Y_2$  are the average of disjoint subsets of the successor nodes. In practice, Broadie and Glasserman use  $(b-1)$  values to calculate  $Y_1$  and use the remaining value as  $Y_2$ . They then average the result over all  $b$  ways of leaving out one of the successor nodes.

Each estimator is calculated backwards recursively until the root node is reached. The steps described form the Monte-Carlo simulation kernel, so this kernel must be repeated a number of times and estimator values averaged.

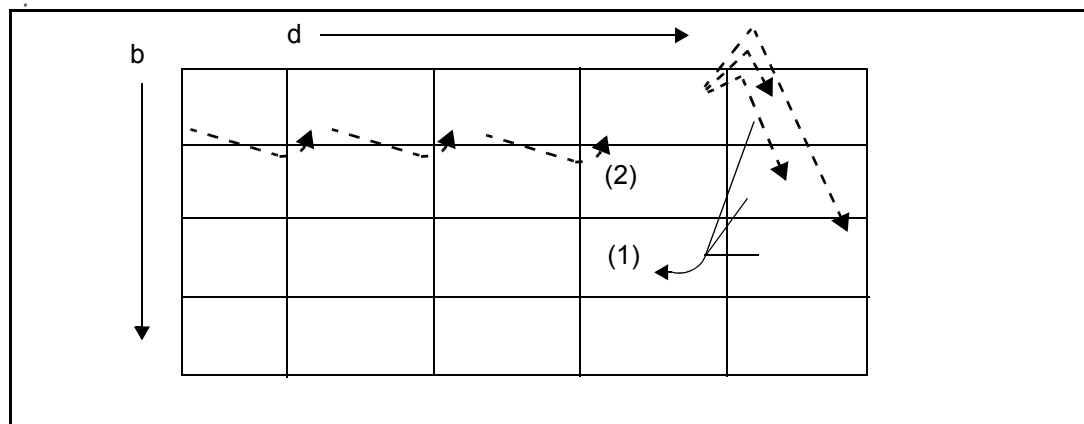
## 7.2 Initial code description

The underlying asset dynamic is modeled as geometric Brownian motion (modeled by the `GBM(...)` function in the code following), so implicitly this is a lognormal process with mean,  $\mu = 0$ . and standard deviation,  $\sigma = 1$ . The generation of random variates is controlled by the `NormInv()` function. Uniform variates are generated using calls to `drand48()` and transformed to normal variates using the inverse cumulative normal distribution. The approximation used for inverse cumulative normal distribution is important, the implementation from Peter Acklam has full machine precision. Other approximations' precision may vary.

The storage requirements for a multinomial or "bushy" tree (with branching factor, 'B') grow exponentially with the number of exercise days, 'D'. This can quickly lead to excessive memory use. However, as Broadie and Glasserman themselves point out, it is not necessary to store the entire tree. The processing of the tree implicitly occurs in depth-first order (processing the leaves first), so the maximum memory requirements can be reduced to  $B \cdot D$  elements, rather than  $B^D$ . The current path in the tree is held in the array `v`. The branch being processed at tree depth, `d`, is indexed using the array `w`. The value at a leaf is generated with the code;

```
// exercise the option
y = z * (v[w[d]][d] - X);
v[w[d]][d] = (y > 0.0) ? y : 0.0;
// whilst not generated all leaves
if (w[d] < B)
{
    v[w[d] + 1][d] = v[w[d - 1]][d - 1] *
        exp(GBM(drift, SigSqrt));
    w[d] = w[d] + 1;
}
```

The tree held in `v` can be visualized as follows;



**Figure 7. Multinomial tree with branching = 3**

The dashed lines show the evolution of the price path before the high estimator is calculated for node (1). Calculating the estimators for node (2) requires re-evolving the price path onto the leaves and placing the results in the element adjacent to (1).

## 7.3 Parallelizing sequential code

Now use a technique similar to the previous Monte-Carlo example. This includes converting all functions to act on poly variables (or to be a poly function) and performing a reduction in one form or another to get the data back into mono memory.

Each PE now performs a fraction of the number of simulations. The number of simulations has been chosen such that it is not a multiple of 96. Unfortunately, this now means that more than `nSimulations` will be performed. In order to guard against this, restrict the execution of the simulation kernel on some of the PEs. This is a common technique.

```
current_sim = get_penum();
for (Simulation = 0; Simulation < nSimulations;
     Simulation += __NUM_PES__)
{
  // rest of code..

  // ensure that only nSimulations are performed
  if( current_sim < nSimulations )
  {
    EstimatorSump = EstimatorSump + v;
  }
  current_sim += __NUM_PES__;
}
```

## 7.4 Incorporating the vector math library

As with most Monte-Carlo schemes, the code can take advantage of massive data-parallelism available by using the vector functions supplied in the Vector Math Library. Follow the same conversion process outlined in the Asian Monte-Carlo example:

1. Unroll the simulation loop a further four times.
2. Convert poly variables to `__DVECTORs`.

## 7.5 Performance and possible improvements

### Performance

As in the previous Monte-Carlo method example, the performance increase depends upon the number of simulations. A small number of simulations numbering just thousands is the cut-off, below which no acceleration is seen. As the number of simulations is increased, to increase the accuracy of the result, the observed speedup increases.

```
$ ./broadieglasserman -v reference
American Call (Broadie-Glasserman Tree Method) value: 7.728031
American Call (Broadie-Glasserman Tree Method) runtime: 4.98629 secs
$ ./broadieglasserman -v vector
American Call (Broadie-Glasserman Tree Method) value: 7.7538555
American Call (Broadie-Glasserman Tree Method) runtime: 0.50800 secs
```

As you can see, for 38400 trials (the default, if not specified on the command-line) the speed-up is 9.8x. Increasing the number of trials to 3,840,000 shows a 12x speed-up. This increase in speedup is due to the decreasing proportion of time spent in transferring data to and from the Advance board.

```
$ ./brodieglasserman -v reference -s 3840000
American Call (Broadie-Glasserman Tree Method) value: 7.729163
American Call (Broadie-Glasserman Tree Method) runtime: 49.8567 secs
$ ./brodieglasserman -v vector -s 3840000
American Call (Broadie-Glasserman Tree Method) value: 7.535152
American Call (Broadie-Glasserman Tree Method) runtime: 4.10476 secs
```

## 8 Summary

These examples should provide confidence in approaching other financial algorithms with a view to exploiting the CSX600 processor. It is important to find parallelism strategies that minimize communication between processing elements and minimize the movement of data between the mono memory and poly memory.

Monte Carlo pricing approaches almost always parallelize because each simulation is independent of the others. Lattice methods, for example binomial trees and PDEs, can benefit from SIMD parallelism, but the grid sizes must be chosen judiciously. If spreading the algorithm across two CSX600 processors requires excessive communication, two independent problems can be run, one on each processor.

Lastly, a high capacity approach can be used for computationally intensive analytic solutions to the Black-Scholes differential equation. Although running a single Black-Scholes analytic solution on the Advance card may be no faster than running it on an Intel Xeon (or AMD Opteron), when attempting to run 1,000,000 solutions back-to-back the Advance card can be significantly faster.

Once the correct coarse-grain parallelism strategy has been found, ensure the optimized math libraries are used. Try to find even more parallelism by unrolling loops and using `__DVECTOR` data-types that perform four calculations at a time. Finally, investigate the use of the ClearSpeed Visual Profiler to profile your code.

## 9 Bibliography

1. ClearSpeed Visual Profiler User Guide
2. ClearSpeed SDK Reference Manual
3. The Cn Standard Library
4. CSX600 Runtime Software User Guide
5. CSPX User Guide
6. Options, Futures and Other Derivatives (Fourth Ed.) - John C. Hull (Pub Prentice Hall 2000)
7. Abramowitz and Stegun: A Handbook of Mathematical Functions (<http://www.math.sfu.ca/~cbm/aands/>)
8. Cox, J., Ross, S., & Rubinstein M., - "Option Pricing: A Simplified Approach." Journal of Financial Economics, 7. (Sept '79)
9. Monte Carlo Methods in Finance - Peter Jaeckel (Pub. Wiley Finance Reprint 2003)
10. Options: Approach for Parallel Implementation of Boyle's Monte Carlo Method - R. Mirani (<http://www.datasimfinancial.com/articles.php>)
11. Paul Wilmott on Quantitative Finance 2nd Ed - Paul Wilmott (Pub. Wiley Finance Reprint 2003)
12. Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach - Daniel J. Duffy (Pub Wiley Finance 2006)
13. Valuing American Options by Simulation: A Simple Least-Squares Approach - Francis A. Longstaff & Eduardo S. Schwartz (<http://galton.uchicago.edu/~mykland/346W05/Longstaff.pdf>)
14. Monte Carlo Methods in Financial Engineering v.53 - Paul Glasserman (Pub. Springer 2000)
15. <http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/DC/dgene.pdf>

## Appendix A Reduction methods

Reduction operations are common in multi-processing environments. It is often necessary to find the sum or product (or some other simple 2-to-1 function) of a list of values distributed over the processors. For instance, at the end of a calculation, we may have 96 results, one on each PE. In order to sum these partial results to a single value, a sum-reduce must be performed.

There are two main reduction methods for the CSX600 architecture, and they vary in their efficiency. Release 3.0 of the SDK provides reduction functions as demonstrated in the examples, these optimized functions are based on the second method described below.

The first relies on copying the values from poly memory into mono memory and performing the sum in mono memory. The pseudo-code follows:

```
poly double partial_result;
double p_result[96];
double result = 0.;
// ... perform calculation and write to partial_result
memcpy2m( p_result, &partial_result, sizeof(double);
for(i=0;i<96;i++) result += p_result[i];
// result contains the sum-reduce value
```

Even allowing for slight inefficiencies in the code this method takes thousands of clock cycles to complete. This is an inefficient method because only the mono arithmetic unit is used and this disregards the processing power of the PE array.

The second method presented is more efficient and with sensible optimization can be made extremely fast. This method relies on the swizzle path connecting adjacent PEs. For more details on swizzling, see [\[1\]](#) and [\[2\]](#).

```
poly double partial_result;
double result = 0.;
// ... perform calculation and write to partial_result
for (i = 0; i < __NUM_PES__; i++)
{
    partial_result = swizzle_down_double(partial_result);
    result += get_swizzle_low_double();
}
```

The `swizzle_down_*` function takes value in the register file on PE(n) and writes it into the register file on PE(n-1). Zeros are shifted into the register on PE95. As values are shifted out PE0, they can be picked up using `get_swizzle_low_*`. The swizzle path is extremely fast, taking only 2 clock cycles to transfer a double word between PEs.

## Revision history

Date	Revision	Changes
September 2010	1.F	Amendment to copyright statement.
September 2008	1.E	Amendment to copyright statement.
July 2008	1.D	Minor updates to 1.C amendments.
June 2008	1.C	Text and code fragment amendments.
March 2008	1.B	Update to equations and minor text amendments.
January 2008	1.A	Updated with minor bug fixes and template changes.
July 2007	1.0	Details of revision status.

**Table 1. Document revision history**

**ClearSpeed Technology Ltd**

130 Aztec West  
Park Avenue  
Bristol BS32 4UB  
United Kingdom

Tel: +44 (0)1454 629 623

Fax: +44 (0)1454 629 624

**Email:** [info@clearspeed.com](mailto:info@clearspeed.com)

**Web:** <http://www.clearspeed.com>

**Support:** <http://support.clearspeed.com>

1. Information and data contained in this document, together with the information contained in any and all associated ClearSpeed documents including without limitation, data sheets, application notes and the like ('Information') is provided in connection with ClearSpeed products and is provided for information only. Quoted figures in the Information, which may be performance, size, cost, power and the like are estimates based upon analysis and simulations of current designs and are liable to change.
2. Such Information does not constitute an offer of, or an invitation by or on behalf of ClearSpeed, or any ClearSpeed affiliate to supply any product or provide any service to any party having access to this Information. Except as provided in ClearSpeed Terms and Conditions of Sale for ClearSpeed products, ClearSpeed assumes no liability whatsoever.
3. ClearSpeed products are not intended for use, whether directly or indirectly, in any medical, life saving and/ or life sustaining systems or applications.
4. The worldwide intellectual property rights in the Information and data contained therein is owned by ClearSpeed. No license whether express or implied either by estoppel or otherwise to any intellectual property rights is granted by this document or otherwise. You may not download, copy, adapt or distribute this Information except with the consent in writing of ClearSpeed.
5. The system vendor remains solely responsible for any and all design, functionality and terms of sale of any product which incorporates a ClearSpeed product including without limitation, product liability, intellectual property infringement, warranty including conformance to specification and or performance.
6. Any condition, warranty or other term which might but for this paragraph have effect between ClearSpeed and you or which would otherwise be implied into or incorporated into the Information (including without limitation, the implied terms of satisfactory quality, merchantability or fitness for purpose), whether by statute, common law or otherwise are hereby excluded.
7. ClearSpeed reserves the right to make changes to the Information or the data contained therein at any time without notice.

© Copyright ClearSpeed Technology Ltd 2010. All rights reserved.

Advance is a registered trademark of ClearSpeed Technology Ltd

ClearSpeed, ClearConnect, Advance and the ClearSpeed logo are trade marks or registered trade marks of ClearSpeed Technology Ltd. All other brands and names are the property of their respective owners.