A Portable and Fast Stochastic Volatility Model Calibration using Multi and Many-Core Processors

Matthew Dixon  
Department of Analytics  
University of San Francisco  
San Francisco, CA  
mfdixon@usfca.edu

Jörg Lotze  
Xcelerit  
28–32 Upper Pembroke Street  
Dublin 2, Ireland  
jorg.lotze@xcelerit.com

Mohammad Zubair  
Department of Computer Science  
Old Dominion University  
Norfolk, VA  
zubair@cs.odu.edu

ABSTRACT
Financial markets change precipitously and on-demand pricing and risk models must be constantly recalibrated to reduce risk. However, certain classes of models are computationally intensive to robustly calibrate to intraday price-stochastic volatility models being an archetypal example due to the non-convexity of the objective function. In order to accelerate this procedure through parallel implementation, financial application developers are faced with an ever growing plethora of low-level high-performance computing frameworks such as OpenMP, OpenCL, CUDA, or SIMD intrinsics, and forced to make a trade-off between performance versus the portability, flexibility and modularity of the code required to facilitate rapid in-house model development and productionization.

This paper describes the acceleration of stochastic volatility model calibration on multi-core CPUs and GPUs using the Xcelerit platform. By adopting a simple dataflow programming model, the Xcelerit platform enables the application developer to write sequential, high-level C++ code, without concern for low-level high-performance computing frameworks. This platform provides the portability, flexibility and modularity required by application developers. Speedups of up to 30x and 293x are respectively achieved on an Intel Xeon CPU and NVIDIA Tesla K40 GPU, compared to a sequential CPU implementation. The Xcelerit platform implementation is further shown to be equivalent in performance to a low-level CUDA version. Overall, we are able to reduce the entire calibration process time of the sequential implementation from 6,189 seconds to 183.8 and 17.8 seconds on the CPU and GPU respectively without requiring the developer to reimplementation in low-level high performance computing frameworks.

1. INTRODUCTION
Financial markets change precipitously and on-demand pricing and risk models must be constantly recalibrated to reduce model risk. The problem of how to robustly calibrate stochastic volatility models to quoted option prices is a topic of great importance in financial modeling. Stochastic volatility models are used to build implied volatility surfaces which are subsequently used to price more exotic options which trade in thin markets. Hence the mis-calibration of stochastic volatility models to liquid vanilla options can lead to significant pricing anomalies in relatively illiquid exotic options.

A serial implementation of the Heston model calibration on a CPU does not meet the performance requirements for on demand calibration to sub-minute best bid-offer snapshots of liquid option chains. Financial application developers seek to deploy computationally intensive calculations on accelerator platforms without concern for low-level high performance implementations. This preference arises from an “implementation skillset gap” - oftentimes high level programmers, with application domain expertise, lack the low-level high performance programming skills to be able to deploy their code on multi and many-core accelerator platforms [4][8]. Moreover, platform optimized variants of the code, reduce maintainability and portability of the production code, and lengthen software testing and thus time to deployment.

In this paper we describe our latest work on speeding up the calibration of a Heston stochastic volatility model on multi-core CPUs and GPUs using the Xcelerit platform. The calibration of a Heston model is performed over n option data points (referred to as a “chain”) which remains fixed during the calibration computation. The calibration

Categories and Subject Descriptors
G.4 [MATHEMATICAL SOFTWARE]: Parallel and vector implementations

General Terms
Algorithms, Performance

Keywords
Calibration, Stochastic Volatility, GPGPU, C++
algorithm starts with an initial guess of the model parameters and iteratively improves the guess using an optimization algorithm until it meets the convergence criteria. A typical organization of this computation involves calling an optimization routine with a pointer to ErrorFunction(), which estimates the error between market observed option prices and model prices for the current guess of the parameter set. This paper describes the efficient implementation of the ErrorFunction() on multi- and many-core architecture using the Xcelerit platform.

Speedups of up to 29.5x and 293x are respectively achieved on an Intel Xeon CPU and NVIDIA Tesla K40 GPU, compared to a sequential CPU implementation. The Xcelerit platform implementation is further shown to be equivalent in performance to a low-level CUDA version. Overall, we are able to reduce the entire calibration process time of the sequential implementation from 6,189 seconds to 183.8 and 17.8 seconds on the CPU and GPU respectively.

1.1 Literature review

Heston’s stochastic volatility [9] has drawn widespread usage for its ability to capture the volatility smile and skew due to, for example, leverage effects. However the robust and fast calibration of this model requires judicious choice of numerical approximation and optimization schemes.

Mikhailov and Nögel [15] address the problem of calibrating Heston’s stochastic volatility model by providing guidance on how to calibrate the model to vanilla call options. The authors draw attention to the fact that the calibration procedure is non-trivial – it is a non-linear programming problem with a non-linear constraint and non-convex objective function. Since multiple local-minima may exist, Mikhailov et al. propose using a combination of global search and local optimizers. The authors further note that the use of common stochastic algorithms for global search, such as simulated annealing, generally renders the calibration problem more computationally burdensome.

To improve the performance of this calibration approach, Aichinger, Binder, Fürst and Kletzmayr [14] implement a shared memory parallelization of the Heston model calibration routine on a multicore CPU SGI Altix 4700 and a GPU server with two C1060 cards and a GTX260 card. Aichinger et al. compare the stability and performance of various off-the-shelf global optimizers before concluding that the best performance can be obtained by using a hybrid composed of one of a variety of global optimizers with a Levenberg-Marquardt unconstrained local optimizer. The global optimizers that the authors consider include the differential evolution (DE) algorithm and simulated annealing (SA), both of which have been employed elsewhere in the quantitative finance literature [1]. Finally, the stability of the hybrid calibration method is evaluated on a snapshot of the option chain on the FTSE-100 index.

The use of an unconstrained local optimizer is unsatisfactory for two reasons. First, bounds on the model parameters are necessary to ensure that the solution lies in a feasible region which is consistent with the modeling parameter definitions. Second, the non-linear constraint enforces the Feller condition which prevents the stochastic volatility process from reaching zero. Neither of these conditions are satisfied a priori by the use of the Levenberg-Marquardt optimizer and the latter condition is not satisfied by using an off-the-shelf global solver either. The authors address a violation of the Feller condition by a posteriori adjustment of the parameter set to satisfy the Feller condition. However, as the authors further note, this adjustment reduces the performance of the local solver.

Dixon and Zubair [6] and Dixon, Khan and Zubair [5] approach the problem from a parallel architecture and software engineering perspective motivated by [10]. Starting from the premise that implementation of financial models in high level programming environments and languages such as R and Python is conducive to design exploration, rapid prototyping and model validation, the authors evaluate parallel implementations of the calibration designed for clusters of multi-core CPUs and GPUs.

For the R implementation, they show how to efficiently off-load the error function into the CUDA programming environment using a map-reduce design pattern [10]. While the implementation is highly efficient, the performance increase comes at the price of compromising model flexibility, modularity and portability because the stochastic volatility model must be reimplemented in CUDA instead of enabling modeling specific code to remain implemented in R. Furthermore, the parallel implementation is specific to GPU architectures with more cores than there are options in the chain, rather than hardware independent.

While R and Python are very convenient high level modeling languages, C++ and C# still remain the primary languages used by quants in investment banks because many performance critical models are often already available. The contribution of this paper is to demonstrate and evaluate how the Xcelerit platform can be used to efficiently deploy the Heston model calibration on either GPUs or multi-core CPUs with a single code base written in C++. We begin by revisiting the Heston model specification in the next Section, followed by introducing the calibration procedure in Section 3. Section 4 briefly describes the Xcelerit platform and demonstrates how the tool can be used to implement portable accelerated stochastic volatility calibration algorithms. Performance benchmarks of the Xcelerit implementation are presented in Section 5. Section 7 concludes.

2. HESTON MODEL

The calibration problem and implementation described later in this paper generalize to a wide range of stochastic volatility models. However, to fix notation and detail the model which shall be used for benchmarking the Xcelerit implementation, a brief introduction to the Heston stochastic volatility model is provided here.

The Heston model describes the evolution of a stock price \( S_t \), whose variance \( V_t \) is given by a mean reverting square root process:

\[
\begin{align*}
\frac{dS_t}{S_t} &= \mu dt + \sqrt{V_t} dW_t^1, \\
\frac{dV_t}{V_t} &= \kappa(\theta - V_t)dt + \sigma \sqrt{V_t} dW_t^2,
\end{align*}
\]

A key characteristic of the model is that the Wiener processes are correlated \( dW_t^1 \cdot dW_t^2 = \rho dt \). This feature enables the model to exhibit the ‘leverage effect’. In the notation of Equation (6), the parameter set \( \mathbf{z} := [\theta, \sigma, \kappa, \rho, \nu_0] \) and the additional non-linear constraint (the Feller condition) \( 2\kappa \theta - \sigma^2 > 0 \) is imposed during the calibration to ensure that \( V_t \) is positive. 

2.1 Pricing

With marginal loss of generality, we will restrict the scope of this section to European option prices. The Heston stochastic volatility model permits closed-form solutions for computing risk neutral European option prices. The price can be represented as a weighted sum of the delta of the European call option $P_1$ and $P_2$ - the probability that the asset price will exceed the strike price at maturity. Adopting standard option pricing notation, the call price of a vanilla European option is

$$C(S_0, K, \tau; z_0) = S_0 P_1 - K e^{-(r - q)\tau} P_2,$$

where $\phi_\tau$ are Heston analytic characteristic functions and are given in a convenient form in [11], and $z_0$ is the vector of Heston model parameters. Following Fang and Oosterlee [7], the entire inverse Fourier integral in Equation (4) is reconstructed from Fourier-cosine series expansion of the integrand to give the following approximation of the call price

$$C(S_0, K, \tau; z_0) \approx K e^{-(r - q)\tau} \sum_{k=0}^{N} \phi\left(\frac{k\pi}{b - \alpha}; z_0\right) e^{ik\ln b} U_k,$$

where $x := \ln(S_0/K)$ and $\phi(w; z_0)$ denotes the Heston characteristic function of the log-asset price, $U_k$ the payoff series coefficients and $N$ denotes the number of terms in the cosine series expansion (typically 128 will suffice). The Fourier-Cosine approach is shown to be superior in convergence properties to other FFT and quadrature based methods in [6].

3. CALIBRATION

Calibration of a stochastic volatility model to option prices involves finding the parameters which minimize the error between the model prices and the observed prices across a set of options on the same underlying instrument, but whose contract maturities $\tau$ and strikes $K$ differ. This is formulated as a constrained non-linear least squares optimization problem of the form

$$\min_\mathbf{x} f(\mathbf{x}) = \left(\sum_{i=1}^{|\mathbf{K}|} \sum_{j=1}^{|\mathcal{T}|} w_{ij} [V(S_0, K_i, \tau_j; \mathbf{x}) - \hat{V}_{ij}]^2 \right)^{1/2},$$

subject to the bound constraints $a_i \leq x_i \leq b_i$ (an additional non-linear constraint is imposed by specific models). $V(S_0, K_i, \tau_j; \mathbf{x})$ denotes the model option price and $\hat{V}_{ij}$ denotes the quoted mid-price of the option with an underlying price $S_0$, maturity $\tau_j$ and strike $K_j$. The overall quality of fit is sensitive to the choice of weights. An intuitive choice is to emphasize the most liquid contracts in the chain by choosing the weights to be the reciprocal of the bid-ask spread $w_{ij} = 1/\left(\hat{V}_{i,ask} - \hat{V}_{i,bid}\right)$.

In order to apply off-the-shelf global and local solvers to Equation (6) with a priori enforcement of the Feller constraint, we apply the following transformation $\kappa' = 2\kappa\theta - \sigma^2$, with bounds $\epsilon \leq \kappa' \leq b'$ where $\epsilon << 1$ and $b'$ is given by twice the product of the upper bounds on $\kappa$ and $\theta$ minus the square of the lower bound on $\sigma$. The parameter set in Equation (6) becomes $\mathbf{z}' := [\theta, \sigma, \kappa', \rho, \nu_0]$ subject to

$$a' \leq \mathbf{z}' \leq b'.$$

The original parameter $\kappa$ is recovered from $\kappa = \frac{\kappa' + \sigma^2}{2\theta}$ and since $\kappa', \theta > 0$ it follows trivially that $\kappa > 0$. We note however that we have surrendered direct control of the bounds of $\kappa$, although in practice this can be accommodated provided there is flexibility on the choice of bounds of $\sigma$ and $\theta$. The other constraints are unaffected, $\rho \in [-1, 1], \nu_0 > 0, a > -1, \sigma_j > 0$ and $\lambda > 0$.

4. IMPLEMENTATION

For calibrating the option price model we consider a sample chain of $n$ option data $\mathbf{ch}[n]$, where the $i$th chain data has the following key information:

- $\mathbf{ch}[i].u$: Underlying asset price
- $\mathbf{ch}[i].s$: Strike price
- $\mathbf{ch}[i].m$: Time to maturity
- $\mathbf{ch}[i].p$: Option price

The calibration algorithm starts with an initial guess of five parameters and iteratively improves the guess using an optimization algorithm until it meets the convergence criteria. A typical organization of this computation involves calling an optimization routine with a pointer to the ErrorFunction() given by Equation (6), which estimates the error between market observed option prices and prices calculated using the model, $V()$, for the current guess of the parameter set $p$. More specifically, the ErrorFunction($p$) computes option prices using the option model for a list of tuples $< \mathbf{ch}[i].s, \mathbf{ch}[i].m >$, $0 \leq i < n$ using the current estimates of the five parameters $p$ and compares it with the corresponding data in $\mathbf{ch}[i].p$. In our discussion, we focus on the parallel implementation of the ErrorFunction($p$) as it dominates the overall computation.

A high level description of the sequential version of the ErrorFunction($p$) is given in Algorithm 1. Note that for reasons of keeping the description simple, we have avoided some subtleties of the implementation.

We implemented Algorithm 1 together with the Fourier-Cosine method and model pricing function in C++. The ErrorFunction() is called by non-linear optimization packages NLopt and DEOptim. More specifically, we combine the DEOptim global optimizer with one of three constrained local optimization solvers provided in the NLopt package. These optimizers are (i) the Sequential Least Squares Programming (SLSQP) method; (ii) the L-BFGS-B algorithm; and (iii) the Truncated Newton (TNC) method. Each method exploits the smoothness of the error function over the feasible region by approximating the Jacobian with first order

1We use the average price of the bid and ask (mid-price) as the option price.
forward differences under perturbations of each parameter. A small number of Hessian vectors are also computed at each main iteration in the L-BFGS-B algorithm. The SLSQP method has the advantage that the change of parameterization detailed in Section 3 is not necessary as the method already incorporates non-linear inequality constraints.

The number of function evaluations per iteration is thus dependent on the number of model parameters. The global optimizer is terminated if either the objective function is below a threshold or the number of iterations exceeds a limit. The specifiable stopping criterion varies for each of the local optimizers. However, for ease of comparison of convergence properties between each, it is possible to terminate if either the absolute difference in function values between successive iterations is within a tolerance or the number of function evaluations exceeds a limit. In practice, a tolerance on the absolute difference of the function value is neither intuitive or ideally suited to calibration. In further experiments, not reported here, we find that specifying the tolerance on the norm of the difference in solution iterates leads to more stable parameters over successive calibrations. Of the three aforementioned local solvers, only the TNC method permits a tolerance of this form.

4.1 Xcelerit Implementation

The Xcelerit platform is a tool which makes it easy for financial application developers to develop high performance applications in C++ [12, 16]. Xcelerit-enabled applications efficiently make use of multi-core CPUs, GPUs, and combinations of these in a grid from a single high-level code-base. The source code is free from parallel constructs, and the efficient execution is managed automatically.

To use the Xcelerit API, compute-intensive algorithms are expressed as a dataflow graph – a succession of processing stages (termed actors) which are connected together. A series of actors are continuously applied to streams of data, flowing through the graph. Source actors generate the data, generic actors take data from their input ports and compute the output data to be placed into their output ports, and sink actors consume data (for example, save it into a file).

Figure 1 shows a dataflow graph for the SV model calibration. The SequenceSource source actor generates all combinations of i, j indices, the HestonFC actor evaluates each Heston Fourier-Cosine series term over all options in the chain, the PartialSumReduce actor sums the Fourier-Cosine series to obtain a Heston model price for each option, and the Error sink actor computes the root mean square of the error between the model and quoted option prices.

5. EXPERIMENTAL RESULTS

All performance results reported in this section are obtained using an Intel Xeon CPU and an NVIDIA Tesla K40 GPU (Kepler architecture). The details of the configuration are given in Table 1.

In our earlier experiments [6, 5] we observed that the overall time to calibrate the Heston model is dominated by the ErrorFunction() routine. For this reason, we report the execution time of a single iteration of ErrorFunction() in Tables 2 and 3. Table 2 compares the timing results in milliseconds of the Xcelerit and CUDA implementations of the Heston error function applied to the AAPL chain along with the speedups against the serial version (shown in the parentheses). The Table shows the comparative performance of the Xcelerit and CUDA versions with varying numbers of Fourier-Cosine terms. Table 3 shows the performance results of the serial, CUDA and Xclerit implementations applied to six different single-name equity option chains. The number of Fourier-Cosine terms is kept fixed at 4096. Option chain sizes vary in size, both across instrument and also over time because the number of contracts is dependent on the size of the mid-market. AAPL is the largest option chain at n = 1546.

We observe that the Xcelerit version outperforms the CUDA
Table 2: This table compares the elapsed wall-clock time of the error function in milliseconds between the serial, CUDA and Xcelerit CPU and Xcelerit GPU using the APPL chain data. The error function is evaluated using varying number of Fourier-Cosine terms. Speedups relative to the sequential version are shown in parentheses.

<table>
<thead>
<tr>
<th>#terms</th>
<th>Serial</th>
<th>Xcelerit CPU</th>
<th>CUDA</th>
<th>Xcelerit GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>819.0</td>
<td>31.15 (26.3x)</td>
<td>4.50 (182x)</td>
<td>5.16 (159x)</td>
</tr>
<tr>
<td>1,024</td>
<td>1,642.8</td>
<td>61.71 (26.0x)</td>
<td>8.13 (202x)</td>
<td>7.94 (207x)</td>
</tr>
<tr>
<td>2,048</td>
<td>3,323.2</td>
<td>123.06 (27.0x)</td>
<td>15.24 (218x)</td>
<td>15.50 (246x)</td>
</tr>
<tr>
<td>4,096</td>
<td>7,112.7</td>
<td>248.3 (28.6x)</td>
<td>29.40 (242x)</td>
<td>24.48 (291x)</td>
</tr>
</tbody>
</table>

Table 3: This table compares the elapsed wall-clock time of the error function in milliseconds between the serial, CUDA and Xcelerit CPU/GPU implementations for six single-name equity option chains for 4,096 Fourier terms. Speedups relative to the sequential version are shown in parentheses.

Table 5: This table compares the overall time (in seconds) to calibrate the Heston model across the local solvers and sequential and parallel implementations. The number of iterations $N$ is shown in parentheses.

<table>
<thead>
<tr>
<th>#terms</th>
<th>Solver</th>
<th>Serial time (N)</th>
<th>Xcl CPU time (N)</th>
<th>Xcl GPU time (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>SLSQP</td>
<td>733 (448)</td>
<td>24.18 (448)</td>
<td>4.0 (448)</td>
</tr>
<tr>
<td></td>
<td>LBFGSB</td>
<td>903 (454)</td>
<td>32.8 (459)</td>
<td>6.8 (487)</td>
</tr>
<tr>
<td></td>
<td>TNC</td>
<td>4,109 (810)</td>
<td>107.0 (678)</td>
<td>49.0 (1,209)</td>
</tr>
<tr>
<td>1,024</td>
<td>SLSQP</td>
<td>1,515 (454)</td>
<td>49.2 (448)</td>
<td>6.2 (448)</td>
</tr>
<tr>
<td></td>
<td>LBFGSB</td>
<td>2,598 (497)</td>
<td>76.8 (478)</td>
<td>7.6 (453)</td>
</tr>
<tr>
<td></td>
<td>TNC</td>
<td>7,750 (783)</td>
<td>331.3 (860)</td>
<td>28.1 (689)</td>
</tr>
<tr>
<td>2,048</td>
<td>SLSQP</td>
<td>3,093 (444)</td>
<td>94.4 (444)</td>
<td>10.2 (444)</td>
</tr>
<tr>
<td></td>
<td>LBFGSB</td>
<td>4,047 (459)</td>
<td>126.9 (458)</td>
<td>14.1 (461)</td>
</tr>
<tr>
<td></td>
<td>TNC</td>
<td>18,123 (838)</td>
<td>357.9 (459)</td>
<td>38.5 (627)</td>
</tr>
<tr>
<td>4,096</td>
<td>SLSQP</td>
<td>6,189 (441)</td>
<td>183.8 (441)</td>
<td>17.8 (441)</td>
</tr>
<tr>
<td></td>
<td>LBFGSB</td>
<td>7,451 (447)</td>
<td>332.6 (486)</td>
<td>33.7 (490)</td>
</tr>
<tr>
<td></td>
<td>TNC</td>
<td>21,349 (627)</td>
<td>805.3 (663)</td>
<td>261.6 (1,339)</td>
</tr>
</tbody>
</table>

We find that the SLSQP solver not only requires the least number of iterations to converge, but this number stays relatively stable with varying data and choice of parallel implementation. Using the Xcelerit platform (with 4096 Fourier-Cosine terms), the overall time for calibrating the sequential implementation of the Heston model with the SLSQP solver is reduced from 6,189 seconds to 183.8 seconds and 17.8 seconds on the CPU and GPU respectively. On the other hand, the TNC solver requires the most iterations to achieve convergence and the number of iterations is sensitive to the ordering of the RMSE evaluation, and hence the parallel implementation. This sensitivity explains the high variations in timings between varying numbers of Fourier-Cosine terms and between the serial and parallel implementations for the TNC solver.

6. EXTENSION TO OTHER MODELS

In this paper we employed the Heston stochastic volatility model for all experiments. However, there are many other stochastic volatility models in use in practice, such as the Bates Jump-Diffusion model [2], the Variance Gamma model [3], or CGMY model [13]. As outlined in [5], these models can be calibrated using the same approach as presented here for Heston, but they require changes in the implementation of the ErrorFunction itself.

With the Xcelerit SDK, as dataflow actors are implemented as C++ classes, other models can be implemented by use of polymorphism. That is, by defining a polymorphic base class for the Fourier-Cosine evaluation actor, replacing actor HestonFC in Figure 1, and implementing the model calculation in a derived class, one for each model. With this approach, new models can be added easily, benefiting from GPU acceleration and fast multi-core CPU execution similarly to the Heston model described here.

7. CONCLUSIONS

Financial markets change precipitously and on-demand pricing and risk models must be constantly recalibrated to reduce model risk. This paper describes the acceleration of stochastic volatility model calibration on multi-core CPUs and GPUs using the Xcelerit platform. By adopting a sim-
ple dataflow programming model, the Xcelerit platform enables the application developer to write sequential, high-level C++ code, without concern for low-level high performance computing frameworks such as OpenMP, OpenCL, CUDA, or SIMD intrinsics. This platform provides the portability, flexibility and modularity required to facilitate rapid in-house model development and productionization. Speedups of up to 29.5x and 292.8x are respectively achieved on an Intel Xeon CPU and NVIDIA Tesla K40 GPU, compared to a sequential CPU implementation. The Xcelerit platform implementation is further shown to be equivalent in performance to a low-level CUDA version. Overall, we are able to reduce the entire calibration process time of the sequential implementation from 6,189 seconds to 183.8 and 17.8 seconds on the CPU and GPU respectively without requiring the developer to reimplement in low-level high performance computing frameworks.

8. ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the support of Hanweck Associates and the International Securities Exchange in providing access to a data sample from the ISE/Hanweck Premium Hosted Database.

9. REFERENCES


