Financial Software on GPUs: Between Haskell and Fortran

Cosmin E. Oancea[1], Christian Andreetta[1], Jost Berthold[1], Alain Frisch[2], Fritz Henglein[1]

HIPERFIT, [1]Department of Computer Science, University of Copenhagen (DIKU) and [2]LexiFi

cosmin.oancea@diku.dk, christian.andreetta@diku.dk, berthold@diku.dk, alain.frisch@lexifi.com, henglein@diku.dk

Abstract

This paper presents a real-world pricing kernel for financial derivatives and evaluates the language and compiler tool chain that would allow expressive, hardware-neutral algorithm implementation and efficient execution on graphics-processing units (GPU). The language issues refer to preserving algorithmic invariants, e.g., inherent parallelism made explicit by map-reduce-scan functional combinators. Efficient execution is achieved by *manually* applying a series of generally-applicable compiler transformations that allows the generated-OpenCL code to yield speedups as high as 70× and 540× on a commodity mobile and desktop GPU, respectively.

Apart from the concrete speed-ups attained, our contributions are twofold: First, from a *language perspective*, we illustrate that even state-of-the-art auto-parallelization techniques are incapable of discovering all the requisite data parallelism when rendering the functional code in Fortran-style imperative array processing form. Second, from a *performance perspective*, we study which compiler transformations are necessary to map the high-level functional code to hand-optimized OpenCL code for GPU execution. We discover a rich optimization space with nontrivial trade-offs and cost models. Memory reuse in map-reduce patterns, strength reduction, branch divergence optimization, and memory access coalescing, exhibit significant impact individually. When combined, they enable essentially full utilization of all GPU cores.

Functional programming has played a crucial double role in our case study: Capturing the naturally data-parallel structure of the pricing algorithm in a transparent, reusable and entirely hardware-independent fashion; and supporting the correctness of the subsequent compiler transformations to a hardware-oriented target language by a rich class of universally valid equational properties. Given the observed difficulty of automatically parallelizing imperative sequential code and the inherent labor of porting hardware-oriented and -optimized programs, our case study suggests that functional programming technology can facilitate *high-level* expression of leading-edge performant *portable* high-performance systems for massively parallel hardware architectures.

Categories and Subject Descriptors D.1.3 [*Concurrent Programming*]: Parallel Programming; D.3.4 [*Processors*]: Compiler

General Terms Performance, Design, Algorithms

Keywords autoparallelization, tiling, memory coalescing, strength reduction, functional language

FHPC'12, September 15, 2012, Copenhagen, Denmark.

1. Introduction

The financial system is facing fundamental challenges because of their complexity, interconnectedness and speed of interaction. International banking and insurance regulations increasingly focus on analyzing and reducing the systemic effects of financial institutions on the financial system as a whole. For this reason, such institutions are asked to evaluate their reliability and stability in a large number of economic scenarios, with some of the scenarios presenting critical conditions that require large scale modeling efforts. In this context, Monte Carlo simulations, originally developed by physicists to efficiently investigate the stochastic behavior of complex, multidimensional spaces, have emerged as tools of choice in critical applications like risk modeling and pricing of financial contracts. These simulations are paradigmatic *Big Compute* problems that transcend the domain of *embarrassingly parallel* problems. From a hardware architecture perspective, they require employing and effectively exploiting massive *parallelism*. Interesting results have been achieved by efficient management of processes on grid farms and expert use of specialized hardware such as graphic processing units (GPUs) [26]. In particular, the latter unite the advantages of parallelization, low power consumption, and low latency in data transfer to efficiently execute a large number of single instructions on multiple data (SIMD). This kind of massively parallel hardware requires programming practices that differ from conventional imperative von-Neumann-machine-style programming, however.

The desirability of a programming model that supports high-level description of large-scale data transformations for modeling purposes, coupled with the need to target rapidly evolving massively parallel hardware architectures without letting these infiltrate the programs themselves has led us to concentrate on the well-established practices of *functional programming*. Functional languages are renowned for their good modularity, testability and code reuse [24], which drastically improves maintainability and transparency – crucial properties in areas where the success of a company depends on the correctness and reliability of its software. Furthermore, the purity of functional languages largely facilitates reasoning about the inherent parallelism of an algorithm, and effective parallelizations exist for common higher-order functions [22].

Functional languages are increasingly employed in financial institutions for modeling and high-productivity programming purposes, for instance DSLs for finance [4, 40]. Additionally, functional solutions have demonstrated their ability to exploit novel hardware, such as GPUs and FPGAs, without letting hardware specifics encroach on the programming model [12, 32]. It is the double match of functional programming with modeling in quantitative finance and with naturally expressing data parallelism that motivates our research into architecture-independent parallelization of financial code using a functional approach.

In the remainder of this section we provide a rationale for our case study and an overview of the optimization techniques evaluated. In the following sections we present the functional formulation of the pricing algorithm (Section 2), the optimizations for compiling it to OpenCL (Section 3), the empirical evaluation of the optimizations' impact (Section 4), a review of related work on imperative and functional parallelization (Section 5), and finally our conclusions as to what has been accomplished so far and which future work this suggests (Section 6).

1.1 Notations

Throughout the paper, we denote by $\odot$ a binary-associative operator with neutral element $e_\odot$, `fold` $\odot\ e_\odot\ [a_1,..,a_n] \equiv a_1\odot..\odot a_n$, `scan` $\odot\ e_\odot\ [a_1,..,a_n] \equiv [e_\odot, a_1, a_1\odot a_2, ..., a_1\odot a_2\odot..\odot a_n]$, and `map` $f\ [a_1,..,a_n] \equiv [f\ a_1,..,f\ a_n]$. We also write `(red` $\odot$`)` as a shortcut for `(fold` $\odot\ e_\odot$`)`. We use common-helper functions (i) $\texttt{dist}_p$`::[a]`$\rightarrow$`[[a]]` to split the input list into a list of `p` lists of nearly equal lengths, and (ii) $\texttt{tile}_t$`::[a]`$\rightarrow$`[[a]]` to chunk the list into a list of lists containing each roughly `t` elements.

1.2 Bird's Eye View

While speeding up the runtime of financial software by hand-parallelizing the code for GPU execution is in itself of pragmatic importance, this paper takes a broader view, in which we use the gained insights to evaluate the language and compiler infrastructure needed to automate the process. The main objectives are twofold:

Language. We take the perspective that the language should provide what is necessary for the user (i) to express algorithmic invariants explicitly in the language, and, in general, (ii) to write an implementation that comes as close as possible to the "pure" algorithmic form. If the algorithm is inherently parallel, then we expect the implementation to preserve this property. In this sense, without having parallelism in mind, we have written a sequential, functional (Haskell) version of the generic-pricing algorithm to provide a baseline for comparison against the original imperative (C) code.

Not surprisingly, we find that the *functional* style, with better support for mathematical abstraction, makes parallelism (almost) explicit by means of higher-order functions such as `map`, `fold` and `scan` (i.e., `do-all`, reduction and prefix sum). On the other hand, *imperative*, production code is often optimized for sequential execution but obfuscates the inherent algorithmic parallelism to an extent that makes it difficult to recognize for both programmer and compiler. The latter was observed not only on our case study, but also on benchmarks in PERFECT-CLUB and SPEC suites [21, 39].

We demonstrate this perspective throughout the paper by presenting side-by-side examples of imperative *vs.* functional code and surveying the vast literature of autoparallelizing techniques. Section 1.3 highlights the programming-style differences via a contrived, but still illustrative, example.

Performance. While we have argued that algorithmic clarity should come first, we also take the view that this should not be achieved by compromising performance. The second objective of this paper, outlined in Section 1.4, is to explore the compiler optimizations that have proved most effective for our case study, although they have been implemented by hand: *First*, we present evidence of how user-specified invariants can drive powerful high-level optimizations (e.g. strength reduction). *Second*, we reveal a rich optimization space that exhibits non-trivial cost models, which are best left in the care of the compiler. *Third*, we discuss several lower-level, GPU-related optimizations that have to be the compiler's responsibility if we require hardware transparency (i.e. write once - run anywhere).

1.3 Language Perspective

Figure 1 presents two semantically-equivalent functions, written in Fortran77 and Haskell, which are our instances of imperative and functional languages, respectively. The example is telling in that it combines several interesting coding patterns that appear in implementations of Sobol quasi-random sequences [10], and contrived in that it does not produce random numbers.

```
CC FORTRAN CODE
1  SUBROUTINE example ( D, N, M, dirVs, ret )
2    INTEGER i, j, k, D, N, M, len
3    INTEGER ia(M), ret(D,N), dirVs(M,D)
4    DO i = 1, N
5      len = 0
6      DO k = 1, M
7        IF( test(i,k) ) THEN
8          len     = len + 1
9          ia(len) = k
10     ENDIF  ENDDO
11     DO j = 1, D
12       ret(j, i) = 0
13       DO k = 1, len
14         ret(j,i) = ret(j,i) XOR dirVs(ia(k), j)
15       ENDDO
16       IF(i .GT. 1)
17         ret(j,i) = ret(j,i) XOR ret(j,i-1)
18     ENDDO
19 ENDDO END

-- HASKELL CODE
20 example :: Int -> Int -> Int -> [[Int]] -> [[Int]]
21 example n m dirVs =          --  d × m     n × d
22   let lbody:: Int -> [Int]
23       lbody i =
24         let ia     = filter (test i) [0..m-1]
25             xorV v = fold xor 0 [v!j | j<-ia]
26         in map xorV dirVs
27       ret = map lbody [1..n]
28       e   = replicate (length dirVs) 0
29   in tail (scan (zipWith xor) e ret)
```

Figure 1. Contrived, but illustrative example: Fortran77 *vs* Haskell

Haskell Code. Let us examine first the `lbody` function at lines $22-26$: Indexes in $0..m-1$ are filtered based on the `test` predicate, e.g., testing whether index $k \in [0..m-1]$ in the bit-representation of i is set. Next, (i) the `xorV` function reduces the elements corresponding to the filtered indexes of a `dirVs`'s row with the `xor` operator (i.e., `fold` at line 25), and (ii) this is applied to each row of `dirVs`, i.e., the `map` at line 26. The result of `lbody` is thus a list of the same length (denoted `d`) as `dirVs`.

The rest of `example`'s implementation is straightforward: (i) at line 27 `lbody` is mapped to each integer in $[1..n]$, resulting in a list representation of a $n \times d$ matrix, named `ret`, and finally (ii) prefix-sum with operator `xor` is applied to aggregate the elements in the same position in each row of `ret`, i.e., the `scan` at line 29.

One can observe that parallelism is made (almost) explicit in the implementation by the sequence of `map` and `scan` at lines 27 and 29. The latter has depth $log(n)$, while the former is embarrassingly parallel and exhibits nested[1] parallelism that could be further optimized via flattening [8, 11].

Fortran Code. Examining the Fortran code, an experienced imperative programmer might recognize that (i) the `do k` loop at lines $6-10$ implements the filtering of indexes based on the `test` predicate, and (ii) the `do k` loop at lines $13-15$ corresponds to the `fold` at line 25. (Note that Fortran uses column-major arrays). The outermost loop and the `do j` loop at lines $11-18$ (minus line 17) correspond to the Haskell `map`s at lines 27 and 26, which compute the result array `ret`. The code is arguably less obvious than the one in Haskell, due to the lack of higher-order functions such as `filter`,`fold`, and due to the explicit array indexing.

However, even the experienced programmer might have difficulties understanding that in fact, line 17 implements a prefix-sum computation, i.e., the `scan` at line 29. While the destructive update

[1] Since `lbody` is in itself a `map`, line 27 exhibits the composition of two `map`, which, if merged, would improve the parallelism degree from n to $n \times d$.

to ret(j,i) optimizes[2] the sequential execution time, we note that, at least to some degree, it affects readability.

There are two main impediments to proving parallelism for the outermost loop do i. *The first issue* refers to array ia: the algorithm's logic is that each iteration i works with its own (independent) set of filtered indexes, i.e., ia should be logically declared/allocated inside the loop. The implementation optimizes the sequential case by promoting ia's declaration outside the loop.

However, this results in bogus cross-iteration read-after-write (RAW), write-after-read (WAR) and write-after-write (WAW) dependencies. To enable parallelism, one has to prove the validity of the reverse transformation, known as privatization, which reduces to proving that every read from ia is covered by a write to ia from the same iteration. A programmer might observe that loop do k at line 13 iterates precisely on the set of values computed by loop do k at line 6. However, most compiler solutions [21, 42] cannot establish this property, as their dependency analysis is restricted to cases where the array subscript can be expressed as a closed-form, typically affine, formula in the loop indexes. In our case, the conditional increment of len at line 8 does not satisfy this requirement.

The second issue is even more discouraging: the prefix-sum pattern of line 17 appears as a cross-iteration dependency of constant distance 1, which forms a dependency cycle that cannot be easily broken. Furthermore, prefix-sum can be written imperatively in a number of ways, and we are not aware of compiler technique that would effectively parallelize this pattern. In contrast, parallel reduction is effectively supported by pattern-matching techniques [31].

1.4 Performance Perspective

The previous section hinted that it is significantly more difficult to uncover parallelism from an imperative program than it is to optimize a nearly-parallel functional version via imperative-like optimizations. This section outlines several such optimizations.

Space-Reuse of Map-Reduce Functions. It is well known that (red ⊙) . (map f) can be formally transformed, via list homomorphism (LH) promotion lemma [6], to its equivalent form:

$$(\texttt{red}\odot).(\texttt{map f}) \equiv (\texttt{red}\odot).(\texttt{map}\ ((\texttt{red}\odot).(\texttt{map f}))).\texttt{dist}_p \quad (1)$$

i.e., the input list is split into number-of-processor lists, on which each processor performs the original computation (sequentially), and finally, the local results are reduced in parallel across processors. Note that the map-reduce in the middle does not need to instantiate the list result of map, i.e., destructive updates can be used to accumulate each output of f to the local result. This requires a total memory space proportional to p, rather than N (the list's length).

The latter form is typically preferred on massively parallel systems to optimize the communication cost, while the former is preferred on SIMD (vector) systems, which typically exhibit a rather uniform memory and very limited per-processor resources. GPUs are, in a sense, a mix of both: a GPU is pseudo-SIMD, but features a non-homogeneous memory, in which the local memory close to the core is several orders of magnitude faster than the global one. We identify an interesting trade-off: if the result of applying f fits in the fast (local) memory, then the application becomes compute-bound rather than memory-bound. The downside is that increasing the per-core resources decreases the parallelism degree of the system, and, as such, its effectiveness at hiding various kinds of latencies. Section 3.2 explores this trade-off in detail.

Strength Reduction is a transformation that replaces an expensive operation (*) with a recurrence that uses a cheaper operation (+). In the code snippet below, k0+2*(i-1) has been replaced with the cheaper recurrence k=k+2. The inverse transformation, induction variable substitution, replaces the recurrence with a closed-form formula in the loop index, and thereby enables parallelism extraction: (i) it eliminates the cross-iteration RAW dependency on k and allows the compiler to disprove cross-iteration WAW dependences on array A, i.e., $\texttt{k0+2*(i}_1\texttt{-1) = k0+2*(i}_2\texttt{-1)} \Rightarrow \texttt{i}_1 = \texttt{i}_2$.

```
k = k0                                   k = k0
do i = 1,N        Strength Red.          doall i = 1,N
  A[k] = ..       <------------            A[k0+2*(i-1)]=..
  k = k + 2       ------------>          enddo
enddo             Ind.Var.Subst.         k = k0 +MAX(2*N,0)
```

Compilers typically support simple algebras that, for example, allow replacing multiplication/exponentiation with recurrent addition/multiplication formulas in the sequential case, and the reverse for the parallel case. Section 2.2 shows a more complex example of strength reduction and advocates that such invariants should be captured at language level, since they reveal a nontrivial and impactful optimization space, which is explored in Section 3.3.

Branch Divergence. Consider the target code map fun, where fun i = if(test i) then ($\texttt{f}_1$ i) else ($\texttt{f}_2$ i). When evaluating the *parallel* application of fun to all elements of an array on a symmetric multiprocessor (SMP), an asymptotic worst-case time-cost estimate is $\mathcal{C}(\texttt{map fun}) = \mathcal{C}(\texttt{test}) + MAX(\mathcal{C}(\texttt{f}_1),\ \mathcal{C}(\texttt{f}_2))$.

In contrast, when the code runs on a SIMD machine, in case the if branch diverges for at least one core (one element of the array), the runtime effectively corresponds to all cores executing both branches, i.e., $\mathcal{C}(\texttt{map fun}) = \mathcal{C}(\texttt{test}) + \mathcal{C}(\texttt{f}_1) + \mathcal{C}(\texttt{f}_2)$.

Our solution is to tile the computation via LH promotion lemma:

$$\texttt{map fun} \equiv (\texttt{red}\,{+}{+})\ .\ (\texttt{map}\ (\texttt{map fun}))\ .\ \texttt{tile}_t \qquad (2)$$

where map fun in the middle is intended to be executed sequentially, and to replace it (map fun) with a semantics-preserving, efficient, imperative code that permutes map's iteration space 1..t such that the elements for which test succeeds are computed via $\texttt{f}_1$ before the ones for which test fails (via $\texttt{f}_2$).

To see how this approach optimizes divergence, consider the case in which two GPU cores process tiled lists $[1, 2, 2]$ and $[4, 5, 5]$, where test = odd. Without the transformation, the two cores execute different branches for each pair of elements. With the transformation, the lists are processed in the orders $[2, 2, 1]$ and $[4, 5, 5]$, and only the middle elements cause branch divergence.

This technique, described in detail in Section 3.4, is complemented by copying-in and out the tiled lists to and from fast memory in order to not introduce un-coalesced accesses.

Memory Coalescing is achieved on GPU when a group of neighboring cores (e.g., 16) access, in the same instruction, a contiguous chunk of memory (e.g. 64 bytes). Since the virtual memory is implemented interleaved on different memory banks, the whole chunk is brought to registers in one memory transfer (and in parallel with the accesses of all such groups of neighboring cores).

As explained in Section 3.5, one can transparently restructure arrays and indexes to enable coalesced accesses. For example, consider the code map (red ⊙)::[[Int]]->[Int], where the input list represents a $N \times 32$ matrix and map is parallelized. In each instruction, each group of 16 cores accesses addresses 128 bytes apart from each other, which requires 16 memory transfers, resulting in inefficient bandwidth utilization. For example, if 16 divides N, the layout can be changed to a three-dimensional array $N/16 \times 32 \times 16$, and an access to row x and column y is mapped to index (x `div` 16, y, x `mod` 16), achieving coalesced accesses.

1.5 Main Contributions

We consider the following main contributions of this paper:

- A side-by-side comparison of functional *vs* imperative code patterns that provides evidence that parallelism is easier to recog-

[2] ret(j,i) is locally computed at lines 12-15; line 17 xor-aggregates, across same-row-position elements, the local contribution of iteration i to the "sum" of the previous i-1 iterations, available in ret(j,i-1).

nize in the former style, while the latter style often requires the compiler to reverse-engineer sequentially-optimized code,

- Four optimizations that (i) take advantage of the map-reduce functional style to derive simple yet powerful imperative-style program transformations, and (ii) seem well-suited for integration into the repertoire of a GPU-optimizing compiler,
- An empirical evaluation on a real-world financial kernel that demonstrates (hints) that (i) the proposed optimizations have significant impact, and that (ii) the rich trade-off space is effectively exploited by the simple (proposed) cost models,
- From a pragmatic perspective, we show speedups as high as $70\times$ and on average $43\times$ against the sequential CPU execution on a mobile GPGPU, and $\sim 8\times$ that on a mid-range GPGPU.

2. Generic Pricing Algorithm and Invariants

Section 2.1 provides the algorithmic background of our generic-pricing software, outlining the Monte Carlo method used and its salient configuration data. We then illustrate how the computational steps in the algorithm translate to a composition of functional basic blocks that expose the inherent parallelism of the algorithm as instances of well-known higher-order functions. We refer the interested reader to Hull (2009) [25] and Glasserman (2004) [18] for a more detailed description of the financial model and the employment of Monte Carlo methods in finance, respectively.

Section 2.2 advocates the need to express high-level invariants at language level: In the context of the Sobol quasi-random-number generator, we identify a strength-reduction pattern and demonstrate that (i) its specification can trigger important performance gains, but (ii) the latter should be compiler's responsibility.

2.1 A Generic Pricing Kernel for Liquid Markets

Financial Semantics. Financial institutions play a major role in providing stability to economic activities by reallocating capital across economic sectors. Such crucial function is performed by insuring and re-balancing risks deriving from foreseeable future scenarios, quantified by means of (i) a probabilistic description of these (yet unknown) scenarios, and (ii) a method to evaluate at present time their economic impact. Risk management is then performed by allocating capital according to the foreseen value of the available opportunities for investment, while at the same time insuring against outcomes that would invalidate the strategy itself.

Option contracts are among the most common instruments exchanged between two financial actors in this respect. They are typically formulated in terms of trigger conditions on market events, mathematical dependences over a set of assets (*underlyings* of the contract), and a set of exercise dates, at which the insuring actor will reward the option holder with a payoff depending on the temporal evolution of the underlyings. A vanilla European call option is an example of contract with one exercise date, where the payoff will be the difference, if positive, between the value of the single underlying at exercise date and a threshold (strike) set at contract issuing. Options with multiple exercise dates may also force the holder to exercise the contract before maturity, in case the underlyings crossed specific barrier levels before one of the exercise dates.

Three option contracts have been used to test our pricing engine: a European vanilla option over a market index, a discrete barrier option over multiple underlyings where a fixed payoff is function of the trigger date, and a barrier option monitored daily with payoff conditioned on the barrier event and the market values of the underlyings at exercise time. The underlyings of the latter contracts are the area indexes Dj Euro Stoxx 50, Nikkei 225, and S&P 500, while the European option is based on the sole Dj Euro Stoxx 50. The number of monitored dates is one for the European case, and 5 and 367 for the two barrier contracts, respectively.

```
mc_pricing n = sum gains / fromIntegral n
  where c     = init_pricing n
        gains = map ( payoff c           -- $\mathbb{R}^{u\times d} \to \mathbb{R}$
                    . black_scholes c    -- $\mathbb{R}^{u\times d} \to \mathbb{R}^{u\times d}$
                    . brownian_bridge c  -- $\mathbb{R}^{u\cdot d} \to \mathbb{R}^{u\times d}$
                    . gaussian           -- $[0,1)^{u\cdot d} \to \mathbb{R}^{u\cdot d}$
                    . sobolInd c         --  Int  $\to [0,1)^{u\cdot d}$
                    ) [0..n-1]
```

Figure 2. Functional Basic Blocks and Types

Two key components are necessary for the appreciation at current time of the future value of these contracts: (i) a stochastic description of the underlyings, allowing to explore the space of possible trigger events and payoff values, and (ii) a technique to efficiently estimate the expected payoff by aggregating over the stochastic exploration. The kernel studied here uses the quasi-random population Monte Carlo method [18] for the latter. Samples are initially drawn from an equi-probable, homogeneous distribution, and are later mapped to the probability distributions chosen to model the underlyings. Since these exhibit very good liquidity and present no discontinuities, with good approximation they can be independently modeled as continuous stochastic processes following Normal distributions (*Brownian* motions) [7]. Further, these stochastic processes express an intrinsic regular behavior that can allow sampling of the value of the underlyings at the sole exercise dates. As a final step, correlation between the underlyings is imposed via Cholesky composition, by means of a positive-definite correlation matrix L provided as input parameter.

The stochastic exploration proceeds as following: first, the Sobol multidimensional quasi-random generator [10] draws samples from a homogeneous coverage of the sampling space. Then, these samples are mapped to Normally distributed values (by quantile probability inversion [48]), which model the value of each underlying at the exercise dates. A Browian Bridge scales these samples to ensure conservation of the properties of the stochastic processes also in non-observed dates, preserving modeling consistency [25]. These samples, once again scaled to express the expected correlations among the underlyings, now mimic a market scenario. They can therefore be provided as input to the payoff function, which returns the future gain from the contract given this particular scenario. This procedure is repeated for a large number of initial random samples, and an average gain calculated, which estimates the future payoff in its first order statistics. Finally, the impact of this aggregated future payoff at present time is estimated using a suitable discount model [25].

Functional Composition. Figure 2 shows how these algorithmic steps directly translate into composition of essential functions. The first step (given last in the function composition) is to generate independent pseudo-random numbers for all underlyings, u, and dates, d, by means of the Sobol's quasi-random number algorithm (function `sobolInd`). This method is known to provide homogeneous coverage of the sampling space, and thus to a stochastically efficient exploration of such space with a relatively low number of samples [18]. Additionally, it exhibits a strength-reduction invariant that enables an efficient parallel implementation providing identical semantics to its sequential algorithm. The uniform samples are then mapped to Normally distributed values by quantile probability inversion (function `gaussian`).

The next step, `brownian_bridge`, maps the list of random numbers to Brownian bridge samples of dimension $u \cdot d$. This step induces a dependency between the columns of the samples matrix, i.e. in the date dimension d. In the following function, `black_scholes`, the underlyings, stored in the rows of the matrix, are mapped to their individual stochastic process and correlated by Cholesky composition, inducing a dependency in the row dimension. Finally, the `payoff` function computes the payoff from the monitored values of the underlyings.

The outer Monte Carlo level, expressed by the `map` function, repeats this procedure for each of the input samples, and first-order statistics are collected by averaging over the payoff values.

From a developer perspective, this functional outline allows to fully appreciate the composition possibilities and the reusability of this solution. In fact, the only function strictly dependent on the contract type is the `payoff` function, while all the other modules can be freely employed to price options having underlyings modeled with similar stochastic processes. Furthermore, Figure 2 evidences the inherent possibilities for parallelism: distribution (`map`) and reduction (`sum`) are immediately evident, and the functional purity allows to easily reason about partitioning work and dependencies. The Haskell code shown here has in fact been written as a prototype for reasoning about potential parallelization strategies for a C+GPU version; while at the same time providing the basis for an optimized Haskell version for multicore platforms.

2.2 Algorithmic Invariants: Sobol sequences

Algorithm. A *Sobol sequence* [10] is an example of a *quasi-random*[3] or *low-discrepancy* sequence of values $[x_0, x_1, \ldots, x_n, \ldots]$ from the unit hypercube $[0,1)^s$. Intuitively this means that any prefix of the sequence is guaranteed to contain a representative number of values from any hyperbox $\prod_{j=1}^{s}[a_j, b_j)$, so the prefixes of the sequence can be used as successively better representative uniform samples of the unit hypercube. Sobol sequences achieve a discrepancy of $O(\frac{\log^s n}{n})$, which means that there is a constant c (which may depend on s, but not n) such that, for all $0 \leq a_j < b_j \leq 1$:

$$|\#\{x_i \mid x_i \in \prod_{j=1}^{s}[a_j, b_j) \wedge i < n\} - n\prod_{j=1}^{s}(b_j - a_j)| \leq c \log^s n$$

Let us denote the canonical bit representation of non-negative integer n by $B(n)$, with B^{-1} mapping bit sequences back to numbers. The algorithm for computing a Sobol sequence for $s = 1$ starts by choosing a *primitive* polynomial $P = \sum_{i=0}^{d} a_i X^i$ of some degree d over the Galois Field $GF(2)$, with $a_0 \neq 0, a_d \neq 0$. The second step is to compute a number of *direction vectors* m_k via a recurrent formula that uses P's coefficients:

$$m_k = (\bigoplus_{i=1}^{d} a_{d-i} m_{k-i}) \oplus 2^d m_{k-d}$$

for $k \geq d$, where $m \oplus n = B^{-1}(B(m)\,\texttt{xor}\,\,B(n))$ and `xor` denotes the exclusive-or on bit sequences. The values of m_i for $0 \leq i < d$ can be chosen freely such that $2^i \leq m_i < 2^{i+1}$. In the third step, we compute *Sobol proxies* via the *independent* (as opposed to recurrent) formula

$$x'_i = \bigoplus_{j \geq 0} B(i)_j m_j$$

where $B(i)_j$ denotes the j-th bit of $B(i)$. (The 0-th bit is the least significant bit.) Finally, reading the binary representation of Sobol proxies as a fixed point number yields the Sobol number x_i:

$$x_i = \sum_{j \geq 0} B(x'_i)_j 2^{-j-1}.$$

```
-- Independent Formula
sobolInd :: Config -> Int -> [ Int ]
sobolInd c i = map xorVs (sobol_dirs c)
         where
         inds     = filter (bitSet (grayCode i)) [0 .. numbits-1]
         xorVs vs = fold xor 0 [ vs!i | i <- inds ]
-- Generating the first n numbers using the independent formula:
-- map (sobolInd c) [1..n]

-- Recurrent Formula INVAR: i ≥ 0 ⇒
-- sobolInd (i + 1) ≡ sobolRec (sobolInd i) i
sobolRec :: Config -> [Int] -> Int -> [Int]
sobolRec c prev i = zipWith xor prev dirVs
    where dirVs = [ vs!bit | vs <- sobol_dirs c]
          bit   = least_sig_0bit i
-- Generating the first n numbers using the recurrent formula:
-- scan (sobolRec c) (sobolInd c 0) [1..n-1]
```

Figure 3. Sobol Generator: Independent *vs* Recurrent Formulas.

```
REAL zd(u,d), wf(u,d)
DO i = 1, N ...
  DO m = 1, u
    wf( m, bb_bi(0)-1 ) = bb_sd(1) * zd(m, 1);
    DO j = 2, d
      wk = wf( m, bb_ri(j) - 1 );
      zi = zd( m, j );
      wf( m, bb_bi(j) - 1 ) = bb_rw(j) * wk + bb_sd(j) * zi
      IF (bb_li(j) - 1 .NE. -1)
        wf( m, bb_bi(j) - 1 ) += bb_lw(j) * wf( m, bb_li(j) - 1)
    ENDDO
  ENDDO
  ...    res = res + wf(..,..) ...
ENDDO
```

Figure 4. Brownian-Bridge Code Snippet

Instead of using $B(n)$ in the definition of Sobol proxies we can use the *reflected binary Gray code* $G(n)$, which can be computed by taking the exclusive or of n with itself shifted one bit to the right: $G(n)_j = B(n)_j \oplus B(n)_{j+1}$. This changes the sequence of numbers produced, but does not affect their asymptotic discrepancy. It enables the following *recurrence formula* for Sobol proxies:

$$x'_{n+1} = x'_n \oplus m_c$$

where c is the position of the least significant zero bit in $B(n)$.

A Sobol sequence for s-dimensional values can be constructed by s-ary zipping of Sobol sequences for 1-dimensional values.

Invariants. Figure 3 shows the essential parts of our Haskell implementation for s-dimensional quasi-random Sobol proxies.[4] The function `sobolInd` implements the independent formula with the optimization that n's bits set to one are filtered and the result is reduced via `xor`. The recurrent formula is implemented by `sobolRec`: the least significant zero bit is used to select the set of direction vectors (`dirVs`) that are `xor`ed with the corresponding entries of the previous vector (`zipWith xor prev`).

Section 1.4 has outlined an example of *strength reduction*, in which a repeated multiplication was replaced via a computationally cheaper, plus-recurrence formula. We observe that `sobolInd` and `sobolRec` match the strength reduction pattern: Computing the first n vectors via `sobolInd` is embarrassingly parallel, i.e., the `map` in Figure 3, while the strength-reduced `sobolRec` is significantly cheaper but requires a $\log n$-depth algorithm (`scan`).

The imperative Sobol code, not presented here, exhibits the patterns discussed in Section 1.3 that would preclude parallelism discovery. Another illustrative example corresponds to the Brownian-bridge implementation, shown in Figure 4: each iteration `i` reuses the space of array `wf` and accumulates the result in `res`. This space-saving technique, together with the indirect indexing makes it very difficult to prove that each read from `wf` in iteration `i` is covered by a corresponding read to `wf` in the same iteration `i`, i.e., the loop `do i` can be parallelized by privatizing array `wf`. The functional

[3] The nomenclature is misleading since a quasi-random sequence is neither random nor pseudo-random: It makes no claim of being hard to predict.

[4] Our code actually computes the integers corresponding to the *reverse* bit representation of Sobol proxies. functions involved in pricing.

style would likely expand array wf with an outermost dimension of size N, and express the loop as an easily-parallelizable map-reduce pattern, in which map's function is given by the do m loop.

Discussion. This paper takes the perspective that the compiler should be the depositary of the knowledge of how best to optimize a program, while the user should primarily focus on the algorithmic invariants that (i) are typically beyond the compiler's analytical abilities and (ii) would enable the application of such optimizations. There are several reasons that support this view:

First, specifying such invariants requires minimal effort, e.g., sobolInd (i+1) c $\equiv$ sobolRec c (sobolInd c i) i documents the strength reduction invariant: the independent formula can be described via a recurrence.

Second, the optimization strategy is often hardware-dependent, hence it is impossible for the user to write an optimal hardware-agnostic program. For instance, scan sobolRec is well suited to the sequential case, while map sobolInd can be better on a massively parallel machine that exhibits high communication costs.

Finally, program-level transformations are often nontrivial, and at least tedious even for the experienced user to do by hand: e.g., Section 3.3 presents how to optimize both the parallelism depth and time overhead: the computation is tiled via a factor t, where the tile amortizes the cost of one sobolInd over $t-1$ (fast) executions of sobolRec. Another good example is *flattening* [8].

3. Optimizations

This section describes in detail several compiler optimizations that had a strong impact on the pricing algorithm, and that we believe are likely to prove effective in a general context. Sections 3.2 and 3.3 describe optimizations and trade-offs related to exploiting coarse-grained parallelism and strength-reduction invariants. These are high-level transformations demonstrated using functional code snippets. Sections 3.4 and 3.5 present lower-level optimizations, related to branch divergence and memory coalescing, that are demonstrated on a Fortran intermediate representation.

3.1 Language Assumptions

Throughout the paper, we use Haskell to illustrate the functional programming style, but disregard laziness issues and use lists instead of performance-oriented special types like vectors or arrays for the sake of clarity.

When discussing the imperative programming model, we use Fortran77 uniformly, because: (i) it accurately illustrates the original C code of the pricing algorithm, and, if anything, (ii) it eliminates the maybe-aliasing issue, which is a major hindrance to automatic parallelization. Furthermore, (iii) a vast amount of work in autoparallelization targets Fortran77. As a fourth point, Fortran77 code resembles the GPU API OpenCL which we use, in that it supports neither recurrence nor dynamic allocation (static arrays only).

Another aspect to be taken into account when discussing optimizations is data locality and thread grouping on GPUs. A GPU operates in thread *blocks*, and threads are grouped to SIMD groups (so-called *warp*s) executed on one SIMD unit comprising multiple cores. To simplify our argument, we consider that each SIMD unit comprises 32 hardware cores. Technically this is not correct, as a warp resides on only 8 cores, which execute four-cycle instructions and need four threads to amortize the cost, but the analogy is valid for the points we are making. A block of size B yields $B/32$ hardware threads per core, which we call "virtual cores".

3.2 Vectorized *vs* Coarse-Grained parallelism

Section 1.4 has outlined the tradeoff related to selecting one of (at least) two possible implementations of a *map-reduce* computation. Figure 5 illustrates these two choices in the context of the generic-pricing algorithm shown in Figure 2.

```
-- vt1, vt2 $\in \mathbb{R}^{n\times(u\cdot d)}$              CC t1, t2 $\in \mathbb{R}^{u\cdot d}$
-- vt3, vt4 $\in \mathbb{R}^{n\times u\times d}$, vt5 $\in \mathbb{R}^n$     CC t3, t4 $\in \mathbb{R}^{u\times d}$, t5 $\in \mathbb{R}$
let                                      do i = 0, n-1
  vt1 = map sobolInd  c [0..n-1]           t1 = sobolInd        c i
  vt2 = map gaussian           vt1         t2 = gaussian          t1
  vt3 = map brownian_bridge c vt2          t3 = brownian_bridge c t2
  vt4 = map black_scholes   c vt3          t4 = black_scholes   c t3
  vt5 = map payoff          c vt4          t5 = payoff          c t4
in                                         res = res + t5
  sum vt5                                enddo
-- Memory Complexity: $O(n\cdot u\cdot d)$         CC $O(P\cdot u\cdot d)$,$P$ = core num
```

Figure 5. Vectorized (Haskell) *vs* Coarse parallelism (Fortran)

The vectorized version distributes the outer map across each of the basic-block kernels, and reduces the result vector in parallel via the plus operator. (This transformation is the inverse of fusion and is known as loop distribution in the imperative context.)

On GPU, vectorization exhibits the advantage that each kernel requires fewer resources per virtual core than the fused version. This potentially increases the parallelism degree, which can be used for hiding latencies. In addition, vectorization enables each kernel to be further optimized, e.g. the gaussian kernel applies function map gaussian_elem, hence map gaussian exhibits nested parallelism that can be flattened to increase the degree of parallelism.

The downside is that the memory complexity is nonoptimal, i.e., proportional to n, because all intermediate vectors need to be instantiated. It follows that vt1..5 have to be allocated in global storage, which is several order of magnitude slower than the local memory. (The superior parallelism degree hides to a certain level, but typically does not eliminate memory latency, i.e., spawning more computation may stress too much the memory system.)

The coarse-grained version is obtained via the transformation: $(\texttt{red}\ \odot).(\texttt{map f}) \equiv (\texttt{red}\ \odot).(\texttt{map}\ ((\texttt{red}\ \odot).(\texttt{map f}))).\texttt{dist}_p$ that distributes the input list among processors, performs the original computation $(\texttt{red}\ \odot).(\texttt{map f})$ sequentially on all processors and post-reduces the local results in parallel. Space consumption is optimized via privatization: t1..t5 are allocated per virtual-core, and memory is reused via destructive updates for both the privatized variables and the (accumulated) result res. Note that (i) res needs to be replicated for each sub-list before a final (parallel) reduction, and (ii) the iteration scheduling policy, i.e., the list distribution, is omitted in Figure 5, since it is handled automatically by GPU's programming interface (OpenCL compiler).

The main advantage of the coarse-grained version is that the memory consumption is (asymptotically) optimal: its size is proportional to the number of virtual cores rather than to the data size. When all local variables fit in the fast memory this leads to a computational, rather than memory-bound behavior (our example eliminates global-memory latency by encoding the input list via an affine formula on the loop index; this is not always possible). The downside is that (i) it requires more per-virtual-core resources than vectorization, hence exhibits a lower parallelism degree, and (ii) it is not applicable when the local resources do not fit in fast memory.

The Cost Model must be able to compute a maximum size of per-virtual-core resources, as an upper limit from which on the benefits of using local memory are eliminated by the reduced parallelism degree failing to optimize other kinds of latency (e.g., cache and instruction latencies, register dependencies). An accurate model is difficult to implement because latencies are in general both program and data sensitive, e.g., global-memory latency depends on whether memory accesses are coalesced. In principle, this could be addressed via machine-learning and/or profile-guided techniques, but that study is beyond the scope of this paper.

```
-- USER SPECIFICATION
sobolInd :: Config -> Int -> [ Int ]

-- Recurrent Formula INVAR: i ≥ 0 ⇒
-- sobolInd c (i + 1) ≡ sobolRec c (sobolInd c i) i
sobolRec :: Config -> [Int] -> Int -> [Int]
sobolRec Config{..} prev i = ...

-- COMPILER GENERATED CODE
sobolRecMap conf (l,u) = scanl (sobolRec conf) fst [l..u-1]
    where fst = sobolInd conf l

tile_segm :: ((Int,Int)->[a]) -> Int -> Int -> Int -> [a]
tile_segm fun l u t = red (++) [] (map fun iv)
    where divides = (u-l+1) ‘mod‘ t == 0
          last    = if (divides) then [] else [u]
          iv      = zip [l,l+t..] ([l+t-1, l+2*t-1 .. u] ++ last)
-- COMPILER TRANSFORMS map (SobolInd conf) [n..m] TO:
sobolGen conf n m = case (cost_model conf) of
    1 -> tile_segm (sobolRecMap conf) n m tile
    2 -> map (sobolInd conf) [n..m]
    3 -> sobolRecMap conf (n,m)
```

Figure 6. Sobol Generator: Independent *vs* Recurrent Formulas.

A simple heuristic is to define the cutoff point by computing the per-virtual-core resources associated to a reasonably-minimal concurrency ratio $\mathcal{CR}^{min}$. Since the technique eliminates the global-memory latency, $\mathcal{CR}^{min}$ is related to arithmetic latency, which, on our GPU hardware requires a ratio of virtual to hardware cores between 9 and 18, depending on the existence of register dependencies. We choose $\mathcal{CR}^{min} = (9+18)/2 = 14$, and compute the associated per-virtual-core resources as $\mathcal{R}_{th} = \mathcal{M}^{sm}_{fast}/(\mathcal{CR}^{min} \cdot 32)$, where $\mathcal{M}^{sm}_{fast}$ and the denominator denote the fast-memory size and the number of virtual cores per multiprocessor, respectively.

Our hardware exhibits $\mathcal{M}^{sm}_{fast} =$ 112kB, thus $\mathcal{R}_{th}=$ 256 bytes. In our example, each virtual core (iteration) requires storage for three vectors, each of (flattened) size $u \cdot d$: the first two are necessary because some kernels cannot do the computation in-place and the third is necessary to record the previous quasi-random vector required by the strength-reduction optimization. In addition we need about 16 integers to store various scalars, such as loop vectors. It follows that the cutoff point is $u \cdot d = 16$, which is close to the optimal in our case, but warrants a systematic validation. The cost model is implemented via a runtime test, and we observe speedups as high as 2× when the coarse-grained version is selected.

3.3 Strength Reduction

This section demonstrates how strength reduction can trigger a code transformation that combines the advantages of both independent and recurrent formula. In essence, the user-specified invariant:

`sobolInd c (i+1)` ≡ `sobolRec c (sobolInd c i) i`,

allows one to derive that the `(i+k)`th random number, `sobolInd c (i+k)`, can be written as a reduction of the previous `k-1` numbers: `fold (sobolRec c) (sobolInd c i) [i..i+k-1]`, and similarly, the `i`th`..(i+k)`th random numbers can be computed as a prefix sum: `scan (sobolRec c) (sobolInd c i) [i..i+k-1]`. This is synthesized in Figure 6 by the `sobolRecMap` function that computes the (consecutive) samples indexed from `l` to `u`.

The idea is that tiling a `map` computation would allow to use `sobolRecMap` to efficiently (sequentially) compute tile-size consecutive random numbers, where tiles are computed in parallel. More formally, on the domain of lists holding consecutive numbers, one can derive that `map (sobolInd c)` is equivalent to `(red++).(map (map (sobolInd c))).tile`$_t$. The last step is to replace `map (sobolInd c)` with the more efficient `sobolRecMap`, i.e., `(red++).(map (sobolRecMap c)).tile`$_t$.

In Figure 6 we use `tile_segm` to implement tiling, with the difference that we encode a list of consecutive numbers via a pair (l, u) denoting the lower and upper bound of the set, hence `tile_segm` returns a list of such lower-upper bound pairs. Finally, `sobolGen` selects, based on a cost model, one of the (at least) three ways to compute the `n`th to `m`th random numbers.

The Cost Model needs to select between the independent I^f, recurrent R^f and tiled T^f formulas. For the sequential execution, R^f is the most efficient. For the parallel case, we first compare I^f and R^f. Computing N elements with I^f and R^f exhibits depths (i.e., asymptotic runtime) C_{If} and $log(N) \cdot C_{Rf}$, where C_{If} and C_{Rf} are the (average) costs of one execution of I^f and R^f, respectively. It follows that I^f prevails when $log(N) > C_{If}/C_{Rf}$. On GPU, this means that I^f is superior in most cases of practical interest, because N is typically large.

Finally, to compare T^f and I^f in the parallel case, one has to model the tradeoff between the cheaper computational cost of T^f and the negative impact various tile sizes may have on the parallelism degree, and thus on the effectiveness with which latency is hidden. (A detailed exploration is beyond the scope of this paper.)

A simple model that works well on our case study and may prove effective in practice is to compute a maximal tiling size t_{max} such that it still allows for a (fixed) parallelism degree CR^{fix}, high enough to hide all latencies. For example, we pick the virtual-to-hardware-core ratio between extreme values 18 and 64 for compute and memory bound kernels, respectively.

For a input size N, we compute $t_{max} \geq 1$ as the closest power of two less or equal to N/CR^{fix}, and bound it from above via a convenient value, e.g., 128. In essence, we have circumvented the difficult problem of modeling the relation between tile sizes and latency hiding, by computing the maximal tile size that would not negatively impact on T^f. One can observe now that T^f is always superior to I^f (i.e., in terms of the work to compute N elements): $N * C_{If} \geq (C_{If} + (T-1) \cdot C_{Rf}) \cdot N/T \Leftrightarrow C_{If} \geq C_{Rf}$.

Strength reduction exhibits speed-ups as high as 4×, and allows an efficient Sobol implementation that computes the same result as the sequential version, modulo float associativity issues.

3.4 Branch-Divergence Optimization

Intuition. On SIMD hardware, branches that are not taken in the same direction by all cores exhibit a runtime equivalent to each core executing both targets of the branch. This section proposes an inspector-executor approach to alleviate this overhead: (i) the (parallel) loop is tiled, then (ii) the inspector computes a permutation of the iteration space of a tile that groups iterations corresponding to the `true` (`false`) branches together, and, finally, (iii) the executor processes the tile in the new (permuted) order. As outlined in introductory Section 1.4, organizing the (sequential) execution of a tile in this way minimizes the branch divergence across different tiles, which are processed in parallel (SIMD).

Consider the Haskell code `map f ginp`, where `f` is defined as:

```
f a =if (cond a) then (fun1 a)
     else let m = a * a
          in if(cond m) then (fun2 m) else (fun3 m)
```

The top-left part of Figure 7 shows the Fortran version of this code, where the outer loop has been tiled, and, for simplicity we assume that `TILE` divides `N`. The bottom-right part of Figure 7 shows the inspector, `itPerm`, associated to one branch target. The inspector executes the slice of the original code, i.e., `cloned_code`, that is necessary to find the direction taken by the original branch, and replaces the bodies of the branch with code that aligns the indexes of `true`/`false` iterations contiguously in the first/last part of σ, respectively. Finally, the split index is returned. Note that the input σ is not required to be ordered, any input permutation of the iteration space will be transformed in a permutation that groups the

```
CC Tiled Code; TILE | N
DO i = 1, N, TILE
  DO j = i, i+TILE-1
    IF (cond(ginp(j))) THEN
      gout(j) = fun1(ginp(j))
    ELSE
      m = ginp(j) * ginp(j)
      IF ( cond(m) ) THEN
        gout(j) = fun2(m)
      ELSE
        gout(j) = fun3(m)
ENDIF ENDIF ENDDO ENDDO

CC Inspector-Executor Code
CC initially σ=[1..TILE]
PRIVATE i,s1,s2,inp,out,σ
DO i = 1, N, TILE
  inp[1:TILE]=ginp[i:i+TILE-1]
  s1 = itPerm(id,σ,inp,TILE)
  DO j = 1, s1
    out(σ(j))=fun1(inp(σ(j)))
  ENDDO
  s2 = itPerm( sq, σ(s1+1),
               inp, TILE-s1 )
  DO j = s1+1, s1+s2
    m = inp(σ(j)) * inp(σ(j))
    out(σ(j)) = fun2( m )
  ENDDO
  DO j = s1+s2+1, TILE
    m = inp(σ(j)) * inp(σ(j))
    out(σ(j)) = fun3( m )
  ENDDO
  gout[i:i+TILE-1]=out[1:TILE]
ENDDO
```

```
CC Clone: identity (no cloning)
FUNCTION id(x)
  id = x
END
CC Clone: square
FUNCTION sq(x)
  sq = x * x
END

CC Inspector Code: computes
CC an iter-space permutation
CC that groups the true and
CC false iterations together
FUNCTION itPerm(σ, inp, size,
                cloned_code)
  INTEGER beg,end,size,σ(size)
  beg = 1
  end = size
  DO j = 1, size-1
    m = cloned_code( inp(σ(j)) )
    IF( cond( m ) ) THEN
      beg = beg + 1
    ELSE
      tmp    = σ(beg)
      σ(beg) = σ(end)
      σ(end) = tmp
      end = end - 1
    ENDIF
  ENDDO
  if( cond( inp(size) ) )
    beg = beg + 1
  itPerm = beg - 1
END
```

Figure 7. Branch Divergence Example

true and false iterations contiguously, hence σ, once initialized, does not need to be reset in the program.

The bottom-left part of Figure 7 shows the transformed code: The global-memory input associated to the tile is first copied to private space inp. Then, inspector itPerm is called to compute the permutation of the iteration space. Loop DO j = 1, s1 executes the true iterations of the outer if, and a similar loop was intermediary generated for the false iterations. The latter loop was recursively transformed to disambiguate its (inner) if branch.

This corresponds to the second call to itPerm on the remaining indexes σ(s1+1..TILE), in which the cloned code refers to the square-root computation of m in the original code that is used in branch condition cond(m). Finally, the loop is distributed across the true and false iterations of the inner if, and the result is copied out to global memory. (Without the copy-in/out to and from private storage, the permutation of the iteration-space may introduce non-coalesced, global-memory accesses).

Implementation. We observe that the transformation is valid only on independent loops (i.e., parallel, no cross-iteration dependencies), otherwise the iteration-space permutation is not guaranteed to preserve the original program semantics.

Consider the case when an independent loop contains only *one* outermost if branch. To apply the transformation: First, inline the code after the if-then-else construct inside each branch, or separate that code via loop-distribution to form another loop.

Second, extract the inspector by computing the transitive closure of loop statements necessary to compute the branch condition, and by inserting the code that computes the permutation.

Third, generate the (distributed) loops corresponding to the true/false iterations by cloning the loop, replacing the if construct with the body of the true/false branch, substituting the loop index j with σ(j), and simplifying, e.g., dead-code elimination. The procedure can be repeated to optimize inner branches in the two formed loops, where each loop further refines its iteration space recorded in its corresponding (contiguous) part of σ.

```
CC Uncoalesced Layout
DOALL i = 1, N
  DO j = 1, M
    ... ARR(σ(j), i) ...
  ENDDO
ENDDOALL
```

```
CC Coalesced Layout
DOALL i = 1, N/B
  DOALL k = 1, B
    DO j = 1, M
      ... ARR(i%B, σ(j), i/B)
ENDDO ENDDOALL ENDDOALL
```

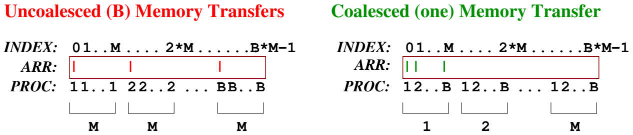
Uncoalesced (B) Memory Transfers
INDEX: 01..M....2*M......B*M-1
ARR:
PROC: 11..1 22..2 ... BB..B
M
M
M
Coalesced (one) Memory Transfer
INDEX: 01..M....2*M......B*M-1
ARR:
PROC: 12..B 12..B ... 12..B
1
2
M

Figure 8. Memory-Coalescing Code Example

If the independent loop contains two branches at the same level, then one can distribute the loop around the two branches and apply the procedure for each branch, i.e., map $(f_1.f_2)$ can be rewritten as $(\text{map } f_1).(\text{map } f_2)$, and the if branches of f_1 and f_2 can be treated individually. Furthermore, the transformation can be applied uniformly via a top-down traversal of the control-flow (and call) graph of the original loop, where each if-branch target is transformed in the context of its enclosing loop.

Cost Model. One can observe that optimizing branch divergence exhibits both fast-memory and instructional overhead: The memory overhead is related to the size of the tile, which typically dictates the size of the private input and output buffers, and the size of σ. Splitting the computation into an inspector-executor fashion may introduce instructional overhead because both the if condition and the if body may be data-dependent on the same statements, e.g., the statement that computes m in Figure 7. Enabling transformations, such as loop distribution, may also require either statement cloning or array expansion to fix potential data-dependencies between the two distributed loops.

To determine the profitability of this transformation, static analysis should first identify good branch candidates, i.e., if statements that exhibit high computational granularity for at least one of their true and false branches (fun1 and fun2). Then, similar to Section 6, the maximal tile size can be computed so that the associated fast-memory overhead does not significantly affect latency hiding.

To improve precision, runtime profiling can be used to measure the divergence ratio and to what degree the transformation would reduce divergence. Finally, the instructional overhead should be taken into account to determine whether this optimization is profitable for the target branch. With our case study, this optimization exhibits speedups (slowdowns) as high (low) as $1.3\times$ $(0.95\times)$.

3.5 Memory-Coalescing Optimization

This section presents a transformation that fixes potential uncoalesced accesses of a map construct, such as map fun inp, where the elements of inp are arrays of similar dimensionality.

One can observe that since map hides the iteration space, any array indexing inside fun would likely be invariant to the loop that implements map. For example, in the left side of Figure 8, DOALL i corresponds to the original map, and the DO j loop implements fun, which processes an inner array of dimension M, indexed by σ(j). Executing the i loop on GPU leads to the access pattern depicted in the left-bottom part of Figure 8, in which B cores in a SIMD-group access in one instruction elements that are 4*M bytes apart from each other, where we assumed for simplicity $\sigma \equiv$ id.

We fix this behavior by reshaping uniformly such arrays via transformation $\mathbb{T}$([x,y]) = [x/B,y,x%B], in which x and y correspond to the row and column index in the original matrix (since Fortran uses column order, we would write ARR(y,x)). Since B is a power of two the new index is computed using *fast arithmetic*.

In essence, we have trimmed the outermost dimension and added an innermost (row) dimension of size B, the size of the SIMD group, such that one SIMD instruction exhibits coalesced access. The top-right part of Figure 8 shows the transformed code, where we made explicit the SIMD grouping via the `DOALL k` loop, while the outer `DOALL i` loop expresses the parallelism among SIMD groups. The bottom-right part of Figure 8 demonstrates that after transformation `B` consecutive cores access contiguous locations.

We observe that this transformation is effective for arrays of any dimensions, as long as the internal indexing is `map`-loop invariant. For example, assuming that the Brownian-bridge code of Figure 4 is written in map-reduce style, i.e., array expansion is applied to `wf` and `zd`, this transformation results in coalesced accesses for arrays `wf` and `zd`, despite the indirect indexing exhibited on the dates (`d`) dimension. Finally, assuming that all computational-intensive kernels are executed on GPU, it is beneficial to reshape all relevant arrays in this fashion, since the potential overheads of the CPU-executed code are in this case negligible.

We conclude by observing that this technique (i) transparently solves any uncoalesced accesses introduced by other compiler optimizations such as tiling, and (ii) yields speed-ups as high as 28×.

4. Experimental Results

Experimental Setup. We study the impact of our optimizations on two heterogenous commodity systems: a desktop[5] and an integrated mobile[6] solution. We compile (i) the sequential-CPU kernel with the `gcc` compiler versions `4.6.1` and `4.4.3`, respectively, with compiler option `-O3`, and (ii) a very similar version of the CPU code with NVIDIA's `nvcc` compiler for `OpenCL` version `4.2` and `4.1`, respectively, with default compiler options. Reported speed-ups were averaged among 20 independent runs.

We estimate the three contracts described in Section 2.1: (i) an European option, named `Simple`, (ii) a discrete barrier option, named `Medium`, and (iii) a daily-monitored barrier option, named `Complex`. These contracts are written in terms of a number of underlyings, u, and dates, d: 1×1, 3×5 and 3×367, respectively. This amounts to very different runtime behavior, since u and d dictate (i) the amount of data processed per iteration and (ii) the weight each basic-block kernel has in the overall computation.

In addition, we estimate the contracts with both single precision (`SimpleF`) and double precision (`SimpleD`) floating points. From a compute perspective this accentuates the different runtime behavior, as `double` are more expensive than `float` operations (and require twice the space). From a financial perspective we note that the results of our parallel versions are equal to the sequential one, with precision higher than 0.001%. This is a consequence of the Sobol quasi-random generator being modeled as described in Section 2.2, where the parallel implementation preserves the modulo associativity semantics exhibited by the sequential version.

Figures 9, 10 and 11 show the speed-up measurements for the described contracts under different optimization conditions. Readings for the *gaming system* are reported as *vertical labels* over plain area bars, while readings for the *mobile solution* are reported as *horizontal, white labels* over crossed regions. All histograms present error bars indicating the standard deviation of the measurements, which seem mostly affected by bus transfer delays between host system and GPU. Missing histograms in the `ComplexF` and `ComplexD` cases are due to the `Complex` contract model exceeding the available fast memory. The remaining of this section discusses the impact of the proposed optimizations.

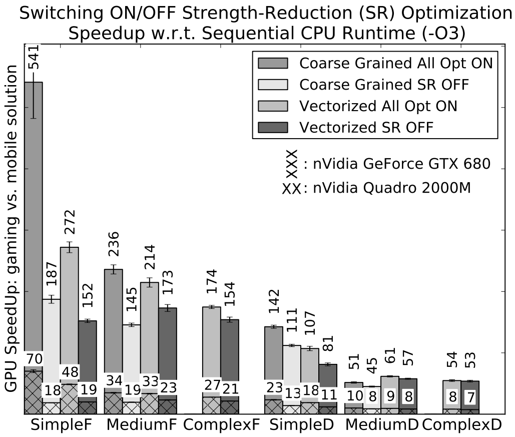

Figure 9. Impact of Strength-Reduction Optimization.

Vectorization *vs* Coarse-Grained. One of the main optimization *choices* the compiler has to make is whether to employ coarse-grained parallelism over vectorization, as described in Section 3.2. Figure 9, in which the reader should ignore for the moment the `SR OFF` bars, demonstrates the tradeoff: The coarse-grained version on `Simple` contract exhibits a small $u \cdot d$ value (1 `float/double`), which results in (all) data fitting well in fast memory, while still allowing a good parallelism degree. It follows that the coarse-grained `SimpleF/D` is significantly faster than its vectorized analog.

As the per-core fast-memory consumption, i.e., $u \cdot d$, increases, latency is less efficiently hidden: (i) `MediumF/D` ($u \cdot d = 15$) is very close to the cutoff point between the two versions, and (ii) `ComplexF/D` cannot run the coarse-grained version simply because $u \cdot d = 3 \cdot 365$ does not fit in fast memory.

We remark that the cutoff point is (surprisingly) well estimated by the simple cost model of Section 3.2, and that, albeit tested (only) on the same application, it is consistent among the two hardware. At large, the top-end hardware exhibits similar behavior but superior speedups for coarse-grained and vectorized versions. The rest of this section evaluates the impact of the other three optimizations for both the coarse-grained and vectorized code.

Strength Reduction. The `SR OFF` bars in Figure 9 correspond to the obtained speed-up when all but the strength-reduction optimization were used. Comparing the `SR OFF` bars with their left neighbor, which correspond to the fully-optimized code, one can observe speed-ups as high as $3-4\times$ for `SimpleF`'s coarse-grained and vectorized code, respectively. As $u \cdot d$ increases, i.e., in `Medium` and `Complex` contracts, the optimization's impact decreases because: (i) the weight of the Sobol kernel in the overall computation decreases and (ii) the tile sizes computed by the cost model also decrease. The latter corresponds to how many times we apply the recurrence formula to amortize the more expensive independent formula, and also explains the smaller impact on the code version that uses `doubles`. For `ComplexF/D` the ratio is four and two, respectively, and the gain is smaller. We remark that the empirical data seem to validate the cost model in that sequentializing some computations via the recurrent formula never generates slowdowns.

Branch Divergence. The results shown in Figure 10 correspond to optimizing the divergence of the only `if` branch, located in the `gaussian` kernel, that exhibits enough computational-granularity to trigger the branch-divergence (BD) optimization. `Simple` ex-

[5] A four-core `Intel Core2 Quad@2.40GHz` with 4 GB global memory, and a NVIDIA `GeForce GTX 680` GPU that exhibits `1536` CUDA cores running at `706MHz`, `2 GB` global memory with a clock-rate of `3GHz`, 48 `kB` fast (local) memory, 64 `kB` constant memory, and `65536` registers per block. GPU and CPU are connected with a PCI EXPRESS V. 1.0 interface.

[6] A four-core `Intel i7-2820QM@2.30GHz` with `8 GB` global memory, and a NVIDIA `Quadro 2000M` GPU that exhibits 192 CUDA cores running at `1100MHz`, `2 GB` global memory, `48 kB` fast (local) memory, `64 kB` constant memory, and `32768` registers per block. GPU and CPU are connected with a SANDY BRIDGE interface supporting PCI EXPRESS V. 2.0.

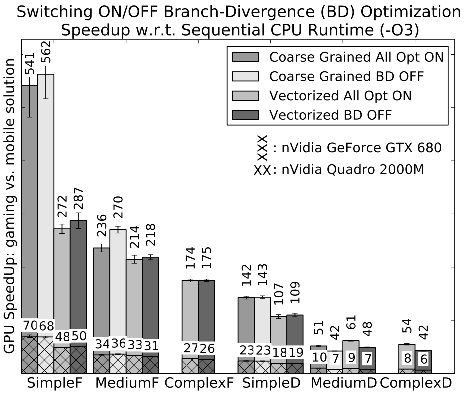

Figure 10. Impact of Branch-Divergence Optimization.

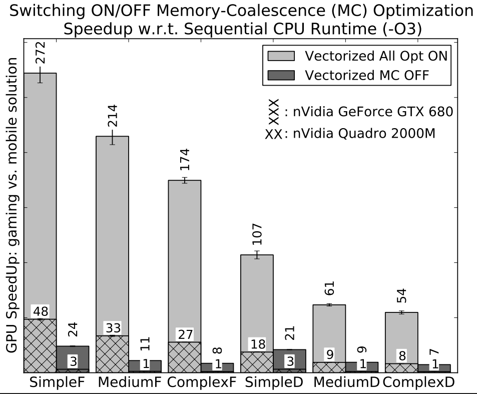

Figure 11. Impact of Memory-Coalescing Optimization.

hibits little divergence overhead, i.e., less than 2% of the `gaussian` kernel runtime, and thus the optimization results in about 5% slowdown due to the computational and memory overheads introduced by BD. `Medium` and `Complex` exhibit roughly 61% divergence overhead at the level of the `gaussian` kernel. Examining `MediumF` we observe an interesting behavior: while the coarse-grained version exhibits slowdown, the vectorized version exhibits either no change, or a speedup. The reason is that the extra memory needed for the appliction of BD exacerbates the reduced degree of parallelism of the coarse-grained version, which is already using a significant amount of fast memory. Finally, we note that the use of `double` increases the computational granularity of the `if` branch, and, as such, the `double` version exhibits significantly better speedups (e.g., $1.3\times$), than the `float`-based one.

Memory Coalescing. Figure 11 demonstrates that achieving coalesced accesses is fundamental for extracting reasonable performance from the GPU hardware, and that the proposed transformation is effective in enabling well-coalesced accesses. We observe that in the case of `SimpleF/D` the uncoalesced accesses refer to the ones introduced by the strength-reduction optimization.

Discussion. Figure 11 shows that Memory Coalescing has by far the largest impact, enabling a factor $5\times$-$28\times$ speedup. This is followed by Strenght Reduction, with a factor $1.3\times$-$4\times$. Choosing between Vectorized and Corse-Grained approaches delivers a speed increase up to $2\times$. The simple cost models implemented here are effective for this application; we would however like to proceed to more extensive empirical evaluations to assess opportunities of generalization. In particular, Branch Divergence seems to express its potential in applications with larger computation granularity, like the `double` case in the `Medium` and `Complex` contracts.

For a fixed set of optimizations, the ratio between the speedups obtained on the two hardware platforms is at large in the interval $5.7\times$-$9\times$, which correlates well with the ratio of the number of available cores in the `GeForce GTX 680` and `Quadro 2000M` GPUs (1536 and 192 respectively). Full utilization of these computing units is achieved by instanciating the entire algorithm on GPU, with little data transfer between host system and GPU. As result of this, the discussed hardwares allow to obtain speedups as high as $70\times$ and $540\times$ compared to the sequential CPU case.

5. Related Work

A considerable amount of work has been published on parallelizing financial computations on GPUs, reporting impressive speedups (see Joshi [26] or Giles [27], for example), or focusing on production integration in large banks' IT infrastructure [36]. Our work differs in that we aim at systematizing and eventually *automating* low-level implementation and optimization by taking a architecture-independent functional language and compilation approach.

Imperative Auto-Parallelization. Classical static dependency analysis [1, 16] examines an entire loop nest at a time and accurately models both the memory dependencies and the flow of values between every pair of read-write accesses, but the analysis is restricted to the simpler affine domain. Dependencies are represented via systems of linear (in)equations, disambiguated via Gaussian-like elimination. These solutions drive powerful code transformations to extract and optimize parallelism [41], e.g., loop distribution, interchange, skewing, tiling, etc., but they are most effective when applied to relatively small intra-procedural loop nests exhibiting simple control flow and affine accesses.

Issues become more complicated with larger loops, where symbolic constants, complex control flow, array-reshaping at call sites, quadratic array indexing, induction variables with no closed-form solutions hinder parallelism extraction [21, 39].

Various techniques have been proposed to partially address these issues: Idiom-recognition techniques [29] disambiguate a class of subscripted subscripts and push-back arrays, such as array `ia` in Figure 1, which is indexed by the conditionally-incremented variable `len`. The weakness of such techniques is that small code perturbations may render the access pattern unrecognizable and yield very conservative results. Another direction has been to refine the dependency test to qualify some non-affine patterns: for example Range Test [9] exploits the monotonicity of polynomial indexing, and similarly, extensions of Presburger arithmetic [42] may solve a class of irregular accesses and control flow.

The next step has been to extend analysis to program level by using various set algebras to summarize array indexes interprocedurally, where loop independence is modeled via an equation on (set) summaries of shape $S = \emptyset$. The set representation has taken the form of either (i) an array abstraction [21, 39], e.g., systems of affine constraints, or (ii) a richer language [43] in which irreducible set operations are represented via explicit $\cap, -, \cup_{i=1}^{N}$ constructors. Array abstractions have been refined further to exploit (simple) control-flow predicates [34, 42] (i) to increase summary precision or (ii) to predicate optimistic results for undecidable summaries. The language-representation approach allows an accurate classification of loop independence at runtime, e.g., it can prove that array `wf` in Figure 4 is write first, hence privatizable, but the runtime cost may be prohibitive in the general case. This issue has been further addressed by a translation $\mathbb{T}$ to an equally-rich language of predicates [37], i.e., $\mathbb{T}(S) \Rightarrow S = \emptyset$, where the extracted predicates $\mathbb{T}(S)$ has been found to solve uniformly a number of difficult cases under negligible runtime overhead. While these important techniques are successful in disambiguating a large number of imperative (Fortran) loops, there still remain enough parallel benchmarks that are too difficult to solve statically [3]. In these cases, techniques that track dependencies at runtime [14, 38] may extract parallelism on multi-core systems, albeit at significant runtime overhead, but they has not been validated (yet) on GPU.

Imperative GPGPU work follows two main directions. *The first* one aims at ease of programming: CudaLite [47] abstracts over the complex GPGPU memory hierarchy, Lime [15] extends the type system of a subset of Java to express desirable invariants such as isolation and immutability, and finally, the popular OpenMP-annotated loops are translated [28] to CUDA, to mention only a few.

The second direction refers to GPGPU performance. Main principles are [44]: (i) achieving memory-coalescing via block tiling, where threads cooperatively copy-in/out the data block to/from on-chip (fast) memory, (ii) optimizing register usage via loop unrolling and (iii) prefetching data to hide memory latency at the expense of register pressure. Implementation of these principles as compiler optimizations ranges from (i) heuristics based on pattern-matching of code or array indexes [15, 47, 49], to (ii) the more formal modeling of affine transformations via the polyhedral model [2, 5], e.g., multi-level tiling code generation, or host-to-accelerator memory-transfer optimizations, to (iii) aggressive techniques that may be applicable even for irregular control-flow/accesses [28], e.g., loop collapsing/interchange exploits a statically-assumed and runtime-verified monotonicity of array values.

In comparison, we take the view that a (hardware-neutral) functional language presents opportunities for both automatic GPU translation and optimization, due to the better preservation of algorithmic invariants. For example, all our optimizations rely on properties of the map-reduce constructs, which appear naturally in the functional code and drive our higher-level (and perhaps simpler) code transformations: *Memory-coalescing* exploits the fact that the array-indexing used inside the mapped function is likely invariant across the mapped elements. Our (novel) technique is complementary to the thread-cooperating block tiling, in that it requires neither block-level synchronization nor the use of shared memory, but it exhibits restructuring overhead when the same (nested) array is traversed on different axes in different kernels.

Similarly, optimizing *branch-divergence* relies on map's parallelism to ensure the validity of the employed iteration-space permutation. The closest related work is Strout's inspector-executor technique [45] that improves cache locality by permuting the array layout to match the order in which elements are accesses at runtime. We have not found stated elsewhere the trade-offs related to the *strength-reduction* invariant and to the *coarse-grained vs vectorized* code, albeit they show significant impact in our case study.

Functional Parallelization. Functional languages and their mathematical abstraction allow for more expressive algorithm implementations, in which data parallelism appears naturally by means of higher-order functions like map, fold, and scan, for which efficient parallel implementations are known. Consequently, research work has focused less on completely automating the process, but rather on studying in a formal manner what classes of algorithms allow asymptotically efficient (parallel) implementations.

Previous research we draw upon here is the Bird-Meertens Formalism (BMF) [6]. Functions f (x++y)=(f x) $\odot$ (f y) are homomorphisms between (i) the monoid of lists with concatenation operator and empty list as neutral element, and (ii) the monoid of the result type with operator $\odot$ and neutral element f [], and can be rewritten in the map-reduce form which provides an efficient parallel implementation (at least when the reduction does not involve concatenation). List invariants like the promotion lemmas (map f).(red ++) $\equiv$ (red ++).(map (map f)) and (red $\odot$).(red ++) $\equiv$ (red $\odot$).(map (red $\odot$)), can be used to transform programs to a higher degree of parallelism and load balancing ($\leftarrow$ direction), or to distribute computations to available processors for reduced communication overhead ($\rightarrow$ direction).

Other work follows this research strand and (i) studies how to extend a class of functions [13] to become list-homomorphisms (LH), or (ii) show how to use the third LH theorem [17] to formally derive the LH definition from its associated (and simpler) leftwards and rightwards forms [19, 35], or (iii) formulate a class of functions [20], such as scan, for which an asymptotically-optimal hypercube implementation can be formally derived, despite the fact that concatenation appears inside $\odot$, or (iv) extend the applicability of BMF theory to cover programs of more general form [23].

All these techniques rely on a functional computation language, where referential transparency and the absence of side-effects allow for vast transformations and rewriting. Such program transformations (with known operators) play a major role in compilation of functional programs, for example in the implementation of data-parallel Haskell [11]. Other work targets GPU platforms [12] using two-stage execution techniques and JIT compilation. All Haskell's data-parallel approaches rely heavily on fusion to adjust task granularities and to justify parallel overhead for the particular platform.

Our work is informed by the same reasoning for the high-level optimization, e.g., coarse-grain vs. vectorised code, strength reduction, but also addresses other important hardware-specific optimizations, e.g., memory coalescing and branch divergence.

It is a general problem that functional approaches can lead to excess parallelism and too fine-grained tasks. More task-oriented parallelization techniques today follow a semi-explicit programming model of annotations (GpH [46]), or make parallelism completely explicit (Eden [30] and the Par monad [33]). Automatic parallelism in these approaches relates mainly to runtime system management, and to functional libraries that capture algorithmic patterns at a high abstraction level. Our work is more narrow in the algorithmic selection, and thereby allows for very specific optimizations.

6. Conclusions and Future Work

This paper (i) has shown evidence that real-world financial software exhibits computationally-intensive kernels that can be expressed in a list-homomorphic, map-reduce fashion, and (ii) has presented and demonstrated four relatively-simple optimizations that allowed substantial speedups to be extracted on commodity GPUs.

While functional languages have often been considered elegant but slow, GPU's enticing parallelism and this paper's results motivates a systematic investigation of what is necessary to transparently and efficiently extract parallelism from functional programs.

As future work we plan to implement and explore such imperative-style optimizations and their cost models, in the context of an array-calculus functional language, such as Single Assignment C.

We believe that code transformations can be guided by interprocedural analysis that summarizes array read/write accesses, in which the trade-off (cost model) can be modeled as equations on these summaries. When the trade-off cannot be answered statically, (higher-order) predicates (i) can be derived as sufficient-satisfiability conditions of the corresponding equation, and (ii) can be evaluated in parallel on GPU to select the most efficient offline-generated kernel. Such an approach has been validated for the (more difficult) problem of classifying loop parallelism in the Fortran context [37], and we believe it also suits our context well.

Acknowledgments

This research has been partially supported by the Danish Strategic Research Council, Program Committee for Strategic Growth Technologies, for the research center 'HIPERFIT: Functional High Performance Computing for Financial Information Technology' (http://hiperfit.dk) under contract number 10-092299.

References

[1] R. Allen and K. Kennedy. *Optimizing Compilers for Modern Architectures*. Morgan Kaufmann, 2002. ISBN 1-55860-286-0.

[2] M. Amini, F. Coelho, F. Irigoin, and R. Keryell. Static Compilation Analysis for Host-Accelerator Communication Optimization. In *Int. Work. Lang. and Compilers for Par. Computing (LCPC)*, 2011.

[3] B. Armstrong and R. Eigenmann. Application of Automatic Parallelization to Modern Challenges of Scientific Computing Industries. In *Int. Conf. Parallel Proc. (ICPP)*, pages 279–286, 2008.

[4] L. Augustsson, H. Mansell, and G. Sittampalam. Paradise: A Two-Stage DSL Embedded in Haskell. In *Int. Conf. on Funct. Prog. (ICFP)*, pages 225–228, 2008.

[5] M. M. Baskaran, J. Ramanujam, and P. Sadayappan. Automatic C-to-CUDA Code Generation for Affine Programs. In *Int. Conf. on Compiler Construction (CC)*, pages 244–263, 2010.

[6] R. S. Bird. An Introduction to the Theory of Lists. In *NATO Inst. on Logic of Progr. and Calculi of Discrete Design*, pages 5–42, 1987.

[7] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy*, pages 637–654, 1973.

[8] G. Blelloch. Programming Parallel Algorithms. *Communications of the ACM (CACM)*, 39(3):85–97, 1996.

[9] W. Blume and R. Eigenmann. The Range Test: A Dependence Test for Symbolic, Non-Linear Expressions. In *Procs. Int. Conf. on Supercomp*, pages 528–537, 1994.

[10] P. Bratley and B. L. Fox. Algorithm 659 Implementing Sobol's Quasirandom Sequence Generator. *ACM Trans. on Math. Software (TOMS)*, 14(1):88–100, 1988.

[11] M. Chakravarty, R. Leshchinskiy, S. P. Jones, G. Keller, and S. Marlow. Data parallel Haskell: A status report. In *Int. Work. on Declarative Aspects of Multicore Prog. (DAMP)*, pages 10–18, 2007.

[12] M. M. Chakravarty, G. Keller, S. Lee, T. L. McDonell, and V. Grover. Accelerating Haskell Array Codes with Multicore GPUs. In *Int. Work. on Declarative Aspects of Multicore Prog. (DAMP)*, pages 3–14, 2011.

[13] M. Cole. Parallel Programming, List Homomorphisms and the Maximum Segment Sum Problem. In *Procs. of Parco 93*, 1993.

[14] F. Dang, H. Yu, and L. Rauchwerger. The R-LRPD Test: Speculative Parallelization of Partially Parallel Loops. In *Int. Par. and Distr. Processing Symp. (PDPS)*, pages 20–29, 2002.

[15] C. Dubach, P. Cheng, R. Rabbah, D. F. Bacon, and S. J. Fink. Compiling a High-Level Language for GPUs. In *Int. Conf. Prg. Lang. Design and Implem. (PLDI)*, pages 1–12, 2012.

[16] P. Feautrier. Dataflow Analysis of Array and Scalar References. *Int. Journal of Par. Prog*, 20(1):23–54, 1991.

[17] J. Gibbons. The Third Homomorphism Theorem. *Journal of Functional Programming (JFP)*, 6(4):657–665, 1996.

[18] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004. ISBN 0387004513.

[19] S. Gorlatch. Systematic Extraction and Implementation of Divide-and-Conquer Parallelism. In *PLILP'96*, pages 274–288, 1996.

[20] S. Gorlatch. Systematic Efficient Parallelization of Scan and Other List Homomorphisms. In *Ann. European Conf. on Par. Proc. LNCS 1124*, pages 401–408. Springer-Verlag, 1996.

[21] M. W. Hall, S. P. Amarasinghe, B. R. Murphy, S.-W. Liao, and M. S. Lam. Interprocedural Parallelization Analysis in SUIF. *Trans. on Prog. Lang. and Sys. (TOPLAS)*, 27(4):662–731, 2005.

[22] K. Hammond and G. Michaelson, editors. *Research Directions in Parallel Functional Programming*. Springer, London, 2000.

[23] Z. Hu, M. Takeichi, and H. Iwasaki. Diffusion: Calculating Efficient Parallel Programs. In *Int. Work. Partial Eval. and Semantics-Based Prg. Manip. (PEPM)*, pages 85–94, 1999.

[24] J. Hughes. Why Functional Programming Matters. *The Computer Journal*, 32(2):98–107, 1989.

[25] J. Hull. *Options, Futures And Other Derivatives*. Prentice Hall, 2009.

[26] M. Joshi. Graphical Asian Options. *Wilmott J.*, 2(2):97–107, 2010.

[27] A. Lee, C. Yau, M. Giles, A. Doucet, and C. Holmes. On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods. *J. Comp. Graph. Stat*, 19(4):769–789, 2010.

[28] S. Lee, S.-J. Min, and R. Eigenmann. OpenMP to GPGPU: a Compiler Framework for Automatic Translation and Optimization. In *Int. Symp. Princ. and Practice of Par. Prog. (PPoPP)*, pages 101–110, 2009.

[29] Y. Lin and D. Padua. Analysis of Irregular Single-Indexed Arrays and its Applications in Compiler Optimizations. In *Procs. Int. Conf. on Compiler Construction*, pages 202–218, 2000.

[30] R. Loogen, Y. Ortega-Mallén, and R. Peña-Marí. Parallel Functional Programming in Eden. *J. of Funct. Prog. (JFP)*, 15(3):431–475, 2005.

[31] B. Lu and J. Mellor-Crummey. Compiler Optimization of Implicit Reductions for Distributed Memory Multiprocessors. In *Int. Par. Proc. Symp. (IPPS)*, 1998.

[32] G. Mainland and G. Morrisett. Nikola: Embedding Compiled GPU Functions in Haskell. In *Int. Symp. on Haskell*, pages 67–78, 2010.

[33] S. Marlow, R. Newton, and S. Peyton Jones. A Monad for Deterministic Parallelism. In *Int. Symp. on Haskell*, pages 71–82, 2011.

[34] S. Moon and M. W. Hall. Evaluation of Predicated Array Data-Flow Analysis for Automatic Parallelization. In *Int. Symp. Princ. and Practice of Par. Prog. (PPoPP)*, pages 84–95, 1999.

[35] K. Morita, A. Morihata, K. Matsuzaki, Z. Hu, and M. Takeichi. Automatic Inversion Generates Divide-and-Conquer Parallel Programs. In *Int. Conf. Prog. Lang. Design and Impl. (PLDI)*, pages 146–155, 2007.

[36] F. Nord and E. Laure. Monte Carlo Option Pricing with Graphics Processing Units. In *Int. Conf. ParCo*, 2011.

[37] C. E. Oancea and L. Rauchwerger. Logical Inference Techniques for Loop Parallelization. In *Int. Conf. Prog. Lang. Design and Impl. (PLDI)*, 2012.

[38] C. E. Oancea, A. Mycroft, and T. Harris. A Lightweight, In-Place Model for Software Thread-Level Speculation. In *Int. Symp. on Par. Alg. Arch. (SPAA)*, pages 223–232, 2009.

[39] Y. Paek, J. Hoeflinger, and D. Padua. Efficient and Precise Array Access Analysis. *Trans. on Prog. Lang. and Sys. (TOPLAS)*, 24(1): 65–109, 2002.

[40] S. Peyton Jones, J.-M. Eber, and J. Seward. Composing Contracts: an Adventure in Financial Engineering (functional pearl). In *Int. Conf. on Funct. Prog. (ICFP)*, pages 280–292, 2000.

[41] L. Pouchet and et al. Loop Transformations: Convexity, Pruning and Optimization. In *Int. Conf. Princ. of Prog. Lang. (POPL)*, 2012.

[42] W. Pugh and D. Wonnacott. Constraint-Based Array Dependence Analysis. *Trans. on Prog. Lang. and Sys.*, 20(3):635–678, 1998.

[43] S. Rus, J. Hoeflinger, and L. Rauchwerger. Hybrid Analysis: Static & Dynamic Memory Reference Analysis. *Int. Journal of Par. Prog*, 31(3):251–283, 2003.

[44] S. Ryoo, C. I. Rodrigues, S. S. Baghsorkhi, S. S. Stone, D. B. Kirk, and W.-m. W. Hwu. Optimization Principles and Application Performance Evaluation of a Multithreaded GPU Using CUDA. In *Int. Symp. Princ. and Practice of Par. Prog. (PPoPP)*, pages 73–82, 2008.

[45] M. M. Strout. *Performance transformations for irregular applications*. PhD thesis, 2003. AAI3094622.

[46] P. Trinder, K. Hammond, J. Mattson Jr., A. Partridge, and S. Peyton Jones. GUM: a Portable Parallel Implementation of Haskell. In *Int. Conf. Prg. Lang. Design and Implem. (PLDI)*, pages 78–88, 1996.

[47] S.-Z. Ueng, M. Lathara, S. S. Baghsorkhi, and W.-M. W. Hwu. CUDA-Lite: Reducing GPU Programming Complexity. In *Int. Work. Lang. and Compilers for Par. Computing (LCPC)*, pages 1–15, 2008.

[48] M. Wichura. Algorithm AS 241: The percentage points of the Normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(3):477–484, 1988.

[49] Y. Yang, P. Xiang, J. Kong, and H. Zhou. A GPGPU Compiler for Memory Optimization and Parallelism Management. In *Int. Conf. Prog. Lang. Design and Implem. (PLDI)*, pages 86–97, 2010.