Multitier Reactive Programming in High Performance Computing

Short Paper

Daniel Sokolowski
Technische Universität Darmstadt
Germany
sokolowski@cs.tu-darmstadt.de

Philipp Martens
Technische Universität Darmstadt
Germany
philipp.martens@stud.tu-darmstadt.de

Guido Salvaneschi
Technische Universität Darmstadt
Germany
salvaneschi@cs.tu-darmstadt.de

Abstract
High Performance Computing (HPC) is crucial in a number of sectors, including weather forecasts, particle simulations and fluid dynamics. Existing programming frameworks for HPC expose developers to low-level details such as message passing and explicit memory management, which are hard to program and error-prone.

In this paper, we present ongoing work on increasing the level of abstraction for HPC. We tackle this problem with an approach based on a combination of multitier programming and reactive programming which enables the development of complex processor configurations in a uniform way using event streams as communication pattern. We report our experience with LULESH, a well known HPC benchmark, and we outline our research roadmap.

CCS Concepts · Software and its engineering → Grid computing; · Computer systems organization → n-tier architectures;

Keywords High Performance Computing, Reactive Programming, Tierless Programming

1 Introduction
High Performance Computing (HPC) [14] enables running scientific and technical computations at high speed and it is crucial in a number of sectors including weather forecasts, fluidodynamics and particle simulation. The programming models adopted in HPC aim to take advantage of modern hardware architectures, which can scale to up to millions of processors in the case of supercomputers. Such programming models are based on languages like C, C++ or Fortran. These feature explicit memory management and get augmented with libraries to execute instructions in parallel on multiple processors as well as to support inter-process communication via message passing (MPI [13]/OpenMP).

While the current approach to program HPC applications maximizes efficiency, it also forces developers to deal with a kind of programming experience deeply tangled with low-level details that in many other areas of software development have been abstracted away by high-level languages.

In this paper, we present our ongoing research on increasing the level of abstraction for programming frameworks in HPC. Our approach stems from the observation that PGAS languages [10] share a number of similarities with multitier languages [6, 7, 27, 32, 33]. The latter, have been recently explored in combination with managed runtimes (e.g., Java and Scala) and reactive abstractions, which promise to conveniently relieve developers from both manual memory management and message passing [35, 36]. We present our preliminary results based on the LULESH HPC benchmark, and discuss a research roadmap to improve the abstraction level of current HPC programming frameworks. In summary, we make the following contributions:

• We analyze multitier programming and reactive programming from the perspective of HPC applications and we show how their combination has the potential to improve the design of HPC software.
• We perform a case study and showcase this approach based on the LULESH benchmark, a real world HPC application consisting of more than 6k lines of C++.
• We identify open problems and we present an outlook for future research on combining multitier programming and reactive programming for HPC.

The paper is structured as follows. In section 2, we motivate and introduce our approach to apply ScalaLoci to HPC, present the framework and point out challenges. Section 3 presents a first case study, which highlights necessary future work. Section 4 shows our preliminary performance results. Section 5 presents our research plans, section 6 discusses related work. Section 7 concludes.

2 Multitier Reactives in HPC
In this section, we explain the potential benefits of adopting multitier reactive programming in HPC, present the frame-
works we adopt in this paper, and we analyze the challenges of using these paradigms in the context of HPC.

### 2.1 A Case for Multiter Reactive Programming

Today’s HPC applications require the development of distributed software in which components communicate to exchange partial results of a computation. Explicit handling of code assigned to each processor and of inter-process communication is error-prone and can severely increase the complexity of applications.

Multiter programming languages enable the joint development of distributed system components within one compilation unit. This approach makes the distributed communication flow more explicit and does not break the distributed logic of the application at the network boundaries. In addition, multiter programming reduces the code overhead for communication, because communication code is injected by the compiler. This solution allows developers to focus on the application logic rather than low-level details like data transmission among hosts.

Ideally, the execution of components in distributed systems is parallelizable. In this context, flexibility in expected execution times is crucial to avoid waiting time. Asynchronous communication enables this kind of flexibility. Reactive programming provides support for asynchronous communication and it improves code readability regarding execution and data flows [22].

These observations suggest to combine multiter languages and reactive programming for HPC applications. As suitable multiter language we identified ScalaLoci, because it supports multiter programming in a high-level language (Scala) in user-defined, complex system architectures. Moreover, through the integration of REScala, ScalaLoci features multitier reactive programming across nodes. These two libraries are briefly introduced in the following.

### 2.2 ScalaLoci

ScalaLoci [35, 36] is a general purpose multitier language, which supports various architectures of distributed systems. Those are specified by peer types, which represent endpoints, e.g., processes, of the application. Their connections are specified by so called ties, which setup bindings and their arities as optional, single or multiple. These are examples for a homogeneous peer-to-peer and a client-server architecture:

```scala
1 // peer-to-peer
2 @peer type Node <: { type Tie <: Multiple[Node] }
3 // client-server
4 @peer type Server <: { type Tie <: Multiple[Client] }
5 @peer type Client <: { type Tie <: Single[Server] }
```

Based on these peers, ScalaLoci allows the located placement of values and functions using placement types. These augment a data type with the peer type, on which the value or function is placed on. At compile time, the code is split up for the various executables of the program along these types. Moreover, ScalaLoci checks statically, whether remote function calls and value transports are correct according to the architecture specification. This makes the dataflow visible and enforces correctness across processes. Examples for the placement, remote value accesses and remote procedure calls are shown in Listing 1.

While ScalaLoci enables the integration of distributed control flows, recent versions also introduced a module concept. The modularization can be used to separate sub-systems in the program from each other. The crucial advantage over other module systems is that these separation boundaries do not have to be at the inter-process communication points, but can be chosen in a problem oriented and domain specific way to encapsulate a distributed functionality into a single multitier module.

We see great potential in using ScalaLoci for HPC applications, because system architecture and modularization can be performed in a domain specific way, which is much more intuitive than today’s HPC programming techniques. Moreover, the control and data flow are visible across peers. This is further improved through reactive abstractions of which we describe the applied implementation next.

### 2.3 REScala

REScala [29] is a Scala library, which integrates reactive values seamlessly with an event system. REScala provides data types for events (Listing 2) and so called signals (Listing 3). Events can be used to propagate changes imperatively. Signals update themselves and propagate changes automatically whenever one of their functional dependencies changes.

Both abstractions have the advantage of helping the developer to write encapsulated and composable code, with a clear visualization of the data flow [30, 31]. Moreover, the implied update propagation of values can save a lot of manual implementation work for the exchange of changing data.

REScala events and signals are available for ScalaLoci and can be used seamlessly across peers. This enables simple
2.4 Challenges of Multitier Reactives

These are the challenges we see for multitier reactive programming in HPC.

Architecture Representation. In traditional programming languages the code that belongs to different peers is developed separately, forcing the separation of logic at inter-process communication and introducing error-prone boilerplate code for communication. MPI aims to simplify inter-process communication with a message passing model that is especially helpful in Single Program Multiple Data (SPMD) applications. However, this solution addresses processes via (multidimensional) indexes, which is very abstract and not related to the underlying architecture of the program. This complexity can cause undetected and hard-to-debug errors in the code. In contrast, multitier abstractions allow domain specific architecture definitions, which improve the understandability of the communication code and introduce static correctness guarantees for inter-process communication.

Asynchronous Control Flow. MPI assumes a coordinated startup of all processes and is not designed to scale the system dynamically or support fault-tolerance mechanisms. Traditionally, most HPC applications do not require those features, but recent and future hardware developments lead to HPC application deployments onto hardware clusters, in which the probability of a partial hardware failure is higher. Also, the execution speed variability of today’s hardware processors is significant. In parallel applications written in a synchronous style – like many MPI and OpenMP applications are – this leads to wasted computational time at faster CPUs at each synchronization barrier. Reactive programming can greatly help addressing these issues, providing abstractions for work decomposition into asynchronous tasks with asynchronous communication patterns that can account for varying execution times on distributed processors and avoid unnecessary synchronization barriers.

Performance. A disadvantage of a high-level language with virtualization and garbage collection is that it is usually less performant than hand-optimized low-level C code. For a first step, we consider to implement often executed core functions in optimized C and orchestrate them using ScalaLoci by leveraging the Java Native Interface (JNI). We believe this might be a valid and practicable compromise to combine the best of both approaches. In further steps we can also consider heap management optimizations or manual memory management in the JVM as done for HPC, e.g., in [3].

3 ScalaLoci in HPC

In this section we introduce the LULESH Benchmark and discuss our implementation of LULESH using the ScalaLoci multitier programming language.

3.1 The LULESH Benchmark

Livermore Unstructured Lagrangian Explicit Shock Hydrodynamics (LULESH) [18] is a hydrodynamics proxy application [17] developed at Lawrence Livermore National Laboratory (LLNL). LULESH is a smaller, but representative example of scientific applications in HPC, including the way it implements numerical algorithms and data transfer among processors. Initially developed as a problem for the DARPA UHPC program, which aimed for sustaining the design of extreme scale computers, it is also commonly used as case study in HPC software design, e.g., by [15, 16, 21].

LULESH implementations exist for many technologies. In our work, we refer to the baseline implementation for CPU based computation in the version 2.0.3. This code is implemented in C++ and supports single-threaded sequential execution, multi-threaded execution using OpenMP, distributed execution with MPI, and the combination of these modes. Thus, it is a suitable reference for various setup sizes.

After initialization, the simulation in LULESH is achieved by a central main loop. Each iteration simulates a discrete time step and consists of decentralized calculations for the state update of the simulated nodes and elements as well as the size of the simulated time step. The execution terminates after a predefined timespan has been simulated.

Topology and Communication. LULESH simulates expansion in 3-dimensional space. Thereby, the entire simulation space is a cube. For shared-nothing parallelization this cube

---

1 `val` fruits = Evt[String]() // Event declaration
2 fruits.observe(printIn(_))
3 fruits.fire("Apple") // Prints "Apple"
4 fruits.fire("Banana") // Prints "Banana"

Listing 2. REScala event example.

1 `val` name = Var("Anna") // Signal value declaration
2 `val` s = Signal("Hey " + name()) // Dependent signal declaration
3 s.changed.observe(printIn(_)) // Prints "Hey Anna" directly
4 name.set("Berta") // Prints "Hey Berta"
5 name.set("Chris") // Prints "Hey Chris"

Listing 3. REScala signal example.

implementation of complex data flows in HPC programs. Thereby, the communication is asynchronous and potentially allows the implementation of algorithms that do not waste computational time due to barrier synchronization.
is sub-divided into a mesh of hexahedrons, which are initially equally sized cubes, but deform over time. All processes simulate one hexahedron, called a domain. The number of processes \( n \) has to be \( 3^l, l \in \mathbb{N}_{\geq 0} \).

In each iteration, all processes need to decide on a common time step, which requires global communication among all processes. Furthermore, processes exchange node or element data with the processes simulating adjacent domains. Through the 3-dimensional arrangement into planes, rows, and columns, a domain exchanges data with up to 26 other domains over different communication directions: 6 faces, 12 edges and 8 corners (Figure 1).

### 3.2 ScalaLoci Implementation

![Figure 2. High-level call graph of LULESH.](image)

We ported the main program logic, which contains all inter-process communication, to ScalaLoci while keeping the optimized computation kernels in C++. The detailed separation is visualized in Figure 2, which resembles a high-level call graph of the baseline LULESH implementation. The white methods are re-implemented in ScalaLoci, while the black methods are kept in C++ and called via JNI. Thereby, OpenMP based parallelization is kept in the C++ methods.

**Communication Setup.** A core advantage of the ScalaLoci adaption is the explicit definition of the peer architecture. For the global communication, we define a single Master peer, while all processes simulating a domain are abstract Worker peers connected to Master:

```scala
1 @peer type FaceRowPos <: Worker (  
2     type Tie <: Single[Master] with Single[FaceRowNeg]  
3   )
4 @peer type FaceRowNeg extends Worker (  
5     type Tie <: Single[Master] with Single[FaceRowPos]  
6   )
```

Listing 4. Example neighborhood peers for LULESH.

While such explicit definitions require a significant amount of code, they also grant a significant advantage to developers. Since all communication channels and their directions are explicit, communication between adjacent domains can now be implemented without any syntactic overhead. In addition, the communication flow is represented explicitly in the code, helping developers to understand the application logic.

**Communication Patterns.** For global synchronization after each main loop iteration, the C++ implementation of LULESH uses the synchronous MPI::Allreduce function, which collects values from all domain processes before continuation. Our ScalaLoci implementation instead uses reactive events, on which each domain publishes its values, and the master aggregates the worker events to a central event. This is done for example for the global time step evaluation:

```scala
1 val elapsedTime: Evt[Double] on Worker = // [...]  
2 def mainIteration(): Unit = localOn Worker = placed ( // [...]  
3   elapsedTime fire duration  
4 )
5 val elapsedTimeGlobal: Event[Double] on Master = placed (  
6   elapsedTime.asLocalFromAllSeq.fold(// [...] +) (// [...] +)  
7 )
```

Listing 5. Example of a Domain peer for LULESH.

While this code looks in first place more complicated, it decouples the aggregation communication, which might be necessary for future optimizations of the program.

For communication between neighbors, the C++ implementation uses asynchronous MPI message passing functions. However, this requires a lot of dynamic checks, whether a domain has a neighbor in the direction it wants to communicate, because there is no explicit architecture definition. Moreover, the addressing is based on indexes, which is not intuitive, provides no static correctness guarantees and is as such potentially error-prone. The explicit peer architecture in ScalaLoci with peers for each communication direction improves this by addressing directions by intuitive names, which makes dynamic checks for partners obsolete, and by ensuring correctness of the communication pattern statically.
For instance, the nodal mass value of the adjacent neighbor located at the face in the positive row direction is transferred:

```scala
1 val nmFaceRowPos: Event[Double] on FaceRowPos = // [...]  
2 placed[FaceRowNeg].local {  
3 nmFaceRowPos.asLocal observe {  
4  /* Processing with nodal mass from adjacent domain */  
5 }
```

This example shows the decoupled communication and visualizes the intuitive way in which communication partners are addressed in the code. Moreover, through the explicit architecture specification this implied value transfer is executed automatically on all workers, which have such a neighbor, and not deployed to any worker without such a neighbor.

## 4 Preliminary Performance Evaluation

We evaluate the run time of the C++ LULESH implementation and our ScalaLoci version. All measurements are taken on a c5.metal [34] AWS EC2 cloud computing instance with 96 cores of Intel’s Cascade Lake processors at 3.6/3.9 GHz turbo frequencies, and 192 GB of memory.

We considered the execution in several configurations AXY and BXY (Table 1) which are defined as follows. Each configuration defines the number of processes $n$, the number of OpenMP threads per process $o$, the total thread count $t = n \cdot o$, the LULESH size parameter $s$, and the resulting computational problem size $p = n \cdot s^2$.

Figure 3 shows the results for the configurations defined in Table 1; for single run time is better. Generally, the ScalaLoci version is slower. For the single process executions, we can see that the ScalaLoci implementation benefits from OpenMP parallelization and is single threaded roughly 4 times slower than the C++ reference.

For the measurements with 8 processes, both implementations have increasing run time with higher number of threads due to OpenMP overhead at smaller per-domain complexity. Moreover, the inter-process communication is less efficient for ScalaLoci.

As major bottlenecks in our ScalaLoci adaption we identified that **TCP based communication** causes significant overhead compared to many MPI implementations, for example shared-memory communication in this case study. In contrast to TCP, MPI also enables the usage of faster networking standards like Infiniband in computing clusters. To make use of these techniques, a MPI based communication backend for ScalaLoci is planned. **Immutable objects** for the domain data structure are changed frequently and cause a lot of short living objects. In this LULESH case study these objects cause expensive overhead, because the size of the domain causes the short lived objects to move out of Eden heap space before they are released again. Therefore, garbage collection is very expensive. This can be solved with mutable data structures. **JNI calls** introduce a small time overhead.

### 5 Outlook

The observations we collected through the LULESH case study indicate our directions for future work.

**Architectures.** The LULESH case study adopts a well defined architecture specification, as it is the case for HPC applications which need to refer to the configuration of the processors available for data processing. Future work shall explore how code overhead to express complex architectures can be reduced while maintaining the same static guarantees, and how semi-dynamic architecture specifications could help with this approach.

**Communication Patterns.** The experience with gained in the development of the LULESH benchmark suggests that a set of reusable distributed communication patterns should be made available. We plan to work on this direction by extending ScalaLoci. These shall be solutions for load-balancing, scalability and fault-tolerance for simplified development of high quality HPC algorithms. Plugin implementations of these patterns could be provided as ScalaLoci modules.

**Automated Placement.** Explicit placement like in ScalaLoci offers the highest level of safety, but might hinder performance optimizations based on automatic runtime placement of computations. We plan to explore controlled and automated placement of values and functions – statically at compile time as well as dynamically at run time. Both can also be leveraged for improvements on load-balancing, scalability and fault-tolerance.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>A01</th>
<th>A02</th>
<th>A08</th>
<th>B01</th>
<th>B04</th>
<th>B12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processes $n$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>OMP Threads $o$</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Total Threads $t$</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>32</td>
<td>96</td>
</tr>
<tr>
<td>Size Parameter $s$</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Problem Size $p$</td>
<td>8 $k$</td>
<td>8 $k$</td>
<td>8 $k$</td>
<td>8 $k$</td>
<td>8 $k$</td>
<td>8 $k$</td>
</tr>
</tbody>
</table>

**Table 1.** Run time evaluation configurations.

**Figure 3.** Run time results for LULESH.
Performance Improvements. Language techniques like virtual methods dispatching or GC are common in managed languages but introduce significant overhead. Yet, we strongly believe their benefits should be accessible to HPC developers. We want to combine our approach with native code that is then integrated into the ScalaLoci programming model. Such native functions could be generated automatically, e.g., with Scala Native, or they could be provided by the programmer.

6 Related Work
We discuss related work regarding PGAS languages, multitier programming and reactive programming.

PGAS/APGAS Languages. The Partitioned Global Address Space (PGAS) model is a parallel programming model similar to both the shared memory and the message passing communication models. While cooperating processes may not physically use the same memory space, they can access data that resides on other processes – the memory space appears shared. Remote accesses usually require a special syntax, but, in contrast to message passing, the actual communication to access data from a remote memory space is transparent for programmers. The survey by De Wael et al. [10] classifies PGAS programming languages.

X10 [5] is an object-oriented programming language from IBM Research aiming to increase programmer productivity in HPC. X10 features distribution of arrays over so-called places, abstractions that define a partitioning of the global address space. For parallel programming, X10 introduces asynchronous activities that can be spawned individually or by using parallel loops. Activities are associated with a place. Hence remote data can be accessed by spawning on a place an activity which reads and returns the value or modifies it.

Chapel [4] has a similar data distribution model and supports both data and task parallelism. The language provides a more general way of expressing distribution: Besides arrays, other data structures, like sets or graphs, can be distributed.

In addition to distributed data structures, Fortress [1] features constructs for explicit and implicit parallelism. To bring programs closer to their mathematical counterpart and thus increase productivity in scientific computing, the Fortress syntax is designed to adopt unicode characters (e.g., the sum operator $\Sigma$).

Multitier Languages. Multitier languages aim to remove the separation of code between communicating processes. In the Web context, this is achieved by compiling the client code to JavaScript or by using JavaScript on the server, too. In Links [7] and Opa [28], function annotations specify whether a function is executed on the client or on the server. Similarly, in StiP.Js [25], annotations assign code fragments to the client or the server. In addition, a program slicing mechanism detects dependencies between fragments and the rest of the program and decides where functionalities are placed.

Ur/Web [6] is a multitier language for the Web similar to Haskell. Eliom is a multitier extension of ML [26, 27]. Both assign code placement based on user annotations and check the correctness of placements in the compiler.

The approaches discussed so far focus on the Web. Instead, ML5 [23] is a multitier language for generic software architectures. Similarly, ScalaLoci is a Scala extension that supports distributed systems whose architecture is specified by the user via an architectural specification. The compiler then checks that the application complies with the specified architecture.

Functional Reactive Programming. FRP [12] allows the definition of data flows declaratively and concisely via discrete time-changing values (events) and continuous time-changing values (signals). The runtime of reactive languages automatically propagates the changes to dependent values, which usually form a dependency graph. Over time a number of variants have been proposed, with different strategies for the change propagation over the graph. Bainomugisha et al. [2] provide an overview of the existing solutions.

Flapjax [20] introduced the use of FRP in Web applications, which has recently inspired a number of reactive libraries such as RxJS [19] and Bacon.js [24]. REScala [29] integrates event streams and time-changing values with object-oriented abstractions. Its runtime provides an event propagation system supporting dynamic dependencies for adding event queries during the execution.

Recent research on RP focused on different propagation strategies to achieve properties such as glitch avoidance [8], efficient propagation over remote network connections [11] or concurrent propagation in a Web environment [9].

7 Conclusion
In this paper, we present our ongoing approach to increase the level of abstraction in HPC programming. We propose to apply multitier programming and reactive programming techniques to access complex processor configurations uniformly and to specify their communication patterns. Our work bases on ScalaLoci, which we use for a first case study based on the LULESH benchmark, and shows that this research has the potential to improve the software design of HPC applications.

Acknowledgments
This work has been supported by the German Research Foundation (DFG) within the Collaborative Research Center (CRC) 1053 MAKI and 1119 CROSSING, by the DFG projects SA 2918/2-1 and SA 2918/3-1, by the Hessian LOEWE initiative within the Software-Factory 4.0 project, by the German Federal Ministry of Education and Research and by the Hessian Ministry of Science and the Arts within CRISP, and by the AWS Cloud Credits for Research program.
Multitier Reactive Programming in High Performance Computing

REBLS’19, October 21, 2019, Athens, Greece

References


Languages, and Applications (Companion to OOPSLA ’06). ACM, New York, NY, USA.


