PAEAN: Portable and Scalable Runtime Support for Parallel Haskell Dialects

JOST BERTHOLD*
Commonwealth Bank of Australia, Sydney

and

HANS-WOLFGANG LOIDL
School of Mathematical and Computer Sciences
Heriot-Watt University

and

KEVIN HAMMOND
School of Computer Science
University of St. Andrews

(e-mail: jberthold@acm.org, hwloidl@macs.hw.ac.uk, kevin@kevinhammond.net)

Abstract

Over time, several competing approaches to parallel Haskell programming have emerged. Different approaches support parallelism at various different scales, ranging from small multicores to massively parallel high-performance computing systems. They also provide varying degrees of control, ranging from completely implicit approaches to ones providing full programmer control. Most current designs assume a shared memory model at the programmer, implementation and hardware levels. This is, however, becoming increasingly divorced from the reality at the hardware level. It also imposes significant unwanted runtime overheads in the form of garbage collection synchronisation etc. What is needed is an easy way to abstract over the implementation and hardware levels, while presenting a simple parallelism model to the programmer.

The PAEAN (PArallEl shAred Nothing) runtime system design aims to provide a portable and high-level shared-nothing implementation platform for parallel Haskell dialects. It abstracts over major issues such as work distribution and data serialisation, consolidating existing, successful designs into a single framework. It also provides an optional virtual shared-memory programming abstraction for (possibly) shared-nothing parallel machines, such as modern multicore/manycore architectures or cluster/cloud computing systems. It builds on, unifies, and extends, existing well-developed support for shared-memory parallelism that is provided by the widely-used GHC Haskell compiler. This paper summarises the state-of-the-art in shared-nothing parallel Haskell implementations, introduces the PAEAN abstractions, shows how they can be used to implement three distinct parallel Haskell dialects, and demonstrates that good scalability can be obtained on recent parallel machines.

* Corresponding author. Reported work performed while at the University of Copenhagen (DIKU).
1 Introduction

This paper studies how to implement parallelism portably and abstractly for different dialects of parallel Haskell targeting modern multicore/manycore/cluster/cloud systems. In order to improve energy efficiency and performance, there is a strong trend towards reducing sharing in the underlying parallel hardware (Lameter, 2013). However, exposing sharing concerns to the programmer moves away from the traditional functional programming strengths of increasing programmability and maintaining very high levels of abstraction. In this paper, we describe a runtime system approach that enables us to separate sharing at the language, implementation and hardware levels. In this way, we can provide a possibly shared-nothing implementation for a possibly logically-shared language design, for example. We can thus obtain the benefits of reducing or eliminating sharing at the hardware level, without losing the benefits of high-level abstraction at the language level. By providing a coherent framework of sophisticated runtime system operations, it becomes possible to develop implementations of new language extensions quickly and robustly. The same approach also enables extensibility of the runtime system. This helps overcome two of the key obstacles to developing new parallel runtime systems, while maintaining functional programming properties of abstraction and ease of use at the language level. Our approach is evaluated against three existing and contrasting dialects of Parallel Haskell: GpH (Hammond & Peyton Jones, 1990; Trinder et al., 1995), Eden (Loogen et al., 2005) and EdI (Berthold & Loogen, 2007; Berthold, 2008).

1.1 Contributions

This paper makes the following contributions:

• We expose and evaluate the key design decisions underlying the GpH, Eden and EdI Parallel Haskell dialects, focusing particularly on sharing, parallelism control and load distribution;
• We describe a high-level, portable runtime system framework, PAEAN (PArallEl shAred-Nothing), that builds on the widely-used GHC\(^1\) implementation and that can be used to implement a variety of parallel dialects of Haskell, for possibly shared-nothing parallel systems;
• We demonstrate how PAEAN can be used to implement different dialects of Parallel Haskell, in particular GpH and Eden, capturing all of the requirements of the previously-constructed specialised GUM and DREAM runtime systems;
• We evaluate the performance and scalability of the parallel Haskell implementations that follow PAEAN ideas and principles; and
• We discuss the state-of-the-art for distributed-memory parallel Haskell implementations, relating different programming models to the respective runtime support and assessing the merits of runtime-system versus library support for coordinating parallel programs.

\(^1\) The Glorious Haskell Compiler, https://www.haskell.org/ghc/
Table 1. Library based and runtime-system based parallel Haskell

<table>
<thead>
<tr>
<th></th>
<th>Library</th>
<th>RTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code Bloat</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Abstraction Level</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Language Integration</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td>Communication Abstraction</td>
<td>Pervasive</td>
<td>Localised</td>
</tr>
<tr>
<td>Performance</td>
<td>Hard to tune</td>
<td>Easily tunable</td>
</tr>
<tr>
<td>Implementation Cost</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Extensibility</td>
<td>Easy</td>
<td>Hard</td>
</tr>
</tbody>
</table>

1.2 Library versus Runtime System

A number of recent designs have chosen to implement parallel Haskell constructs exclusively at the library level, e.g. (Maier & Trinder, 2012; Foltzer et al., 2012). In contrast, we have deliberately taken a runtime system (RTS) approach. As summarised in Table 1, there are many advantages to our approach. Firstly, less coordination code is required to manage parallelism and load distribution because an underlying RTS will take care of all the details without programmer intervention. This significantly improves programmability and productivity. RTS-based approaches can abstract most of the coordination into language constructs, which simplifies programming, reduces code clutter and gives better integration with the programming language. This is especially true for communication abstractions, which are pervasive in any library-based approach for shared-nothing approaches. Secondly, performance-critical code is much easier to fine-tune and avoids the functionality limitations that are inherent to a library approach (for instance, custom memory management is simple and direct). The key disadvantages to an RTS approach are: firstly, that the cost of implementation can be significantly higher, since features must be considered throughout the language implementation; and secondly, that it can be difficult to incorporate new features. By taking a modular approach, PAEAN addresses both of these problems, making it easier to design and build new parallel runtime systems.

1.3 Why Sharing Matters

In Haskell, the sharing of the program and its data through graph reduction is fundamental to achieving lazy evaluation. In a parallel setting, this creates a tension since inter-thread sharing can impose unwanted synchronisation. At a hardware level, shared memory is increasingly a performance bottleneck. In order to improve time and energy efficiency, there is therefore a strong trend towards non-uniform memory architectures (NUMA) in multicore/manycore processors (Lameter, 2013), where memory is not treated as a single homogeneous, easily addressable array. At a software level there is a correspondingly strong trend to avoid locks, wherever possible, and to minimise expensive cache coherency by exploiting non-shared memory. We can obtain a more efficient and more scalable implementation by not assuming the existence of shared memory at a hardware or implementation level. Scalability is enhanced by reducing shared hotspots and memory bottlenecks. Such a shared-nothing approach (Gray, 1985) is also easier to deploy on large-scale systems, including clusters and distributed systems. We would also argue that it is easier to implement such an approach, even on a relatively small scale, since issues of
cache coherency, synchronisation etc. are exposed in the implementation, and not hidden through shared memory accesses.

If memory cannot be assumed to be shared, then a number of implementation issues need to be addressed. These include:

1. how to serialise data;
2. how to logically share values between physically disjoint heaps;
3. how to manage global garbage collection while minimising synchronisation costs;
4. how to migrate data, potential parallelism and threads.

Some recent approaches such as Cloud Haskell (Epstein et al., 2011) and HdPH (Maier & Trinder, 2012) delegate much of this work to the applications programmer, making parallelism control fully explicit in the program. While this minimises the work of the systems implementor, such an approach is likely to suffer from replication, errors, and unexpected feature interaction, as well as to cause incompatibilities between applications. In our opinion, fully explicit parallelism control entangled with the application flies in the face of the usual functional programming philosophy of good abstraction, effectively “throwing out the baby with the bathwater”. This paper takes the opposite approach, by focusing on a mostly implicit model of parallelism at the source level, where parallelism is specified in an abstract, declarative way and by supporting this with a sophisticated, well-engineered and flexible runtime system.

1.4 Paper Structure

The remainder of the paper is structured as follows. Section 2 describes the high-level programming models for Parallel Haskell that provide the requirements for the PAEAN implementation. Section 3 describes the PAEAN design and implementation. Section 4 evaluates the utility and scalability of the PAEAN approach. Section 5 covers related work, including extensions to PAEAN. Finally, Section 6 concludes.

2 Parallel Programming models for Haskell

A wide variety of parallel extensions to Haskell have been proposed. A key design question is how explicit to make the coordination of the parallel execution. One extreme is a purely implicit approach, as taken by pH (Aditya et al., 1995), where there are no extensions to the language and all parallelisation is performed by the compiler. Data parallel extensions to Haskell, such as DpH (Chakravarty et al., 2007), take a similar approach, but are limited to data-parallelism on specific data structures. While implicit parallelism is appealing to the programmer, it also limits the scope for performance tuning. The other extreme is to make most aspects of the coordination explicit in the program and thus give the programmer more control on how the parallelism is used. In the Haskell world, examples of this approach include the Par-Monad (Marlow et al., 2011) and Cloud Haskell (Epstein et al., 2011). While they are easier to implement and tune than implicit approaches, such approaches have the disadvantage of introducing all the complexities and lack of abstraction from mainstream parallel programming models. In such an approach, applications programmers must become systems programmers. The approach that we will mainly focus on is therefore
PAEAN: Portable, Scalable Runtime Support for Parallel Haskells

sumEuler :: Int → Int
sumEuler n = sum [euler i | i ← [n..n-1]]

euler :: Int → Int
-- Euler phi function:
euler n = length (filter (relprime n) [1..n-1])

relprime :: Int → Int → Bool
-- are x and y coprime
relprime x y = hcf x y == 1

hcf :: Int → Int → Int
-- highest common factor of x and y
hcf x 0 = x
hcf x y = hcf y (rem x y)

Fig. 1. The basic Euler totient function, sumEuler

a middle one of semi-explicit parallelism, where the programmer only needs to identify (potential) parallelism, but where synchronisation and communication is managed by the runtime system. This greatly simplifies the parallel programming task, but still gives the programmer a concrete handle on how to improve parallel performance. Supporting such an approach requires an elaborate runtime system, however. The focus of this paper is on how to design such a runtime system in a generic, flexible and reusable way. A more detailed discussion of related systems, including ones taking both more explicit and more implicit approaches, is given in Section 5.

In this paper, we will consider three Haskell dialects: Glasgow parallel Haskell (GpH), Eden, and EdI (the implementation language for Eden). The approach taken by GpH uses programmer-specified annotations (par) to create “sparks” that may, or may not, be subsequently transformed into parallel threads. In contrast, in Eden and EdI, parallelism is introduced by process constructs that always create new threads. While Eden provides a purely functional interface, EdI uses monadic constructs in a similar fashion to the Par Monad. We will use the sumEuler function to characterise these three dialects. For a given n, sumEuler computes the Euler \( \phi \)-function\(^2\) for values up to n, and sums the results. The sequential Haskell code for this is shown in Figure 1.

2.1 GpH

GpH (Hammond & Peyton Jones, 1990; Trinder et al., 1995) uses a semi-explicit parallelism model, where the programmer simply indicates potentially parallel closures and all synchronisation, coordination etc. issues are delegated to the runtime system. It uses two basic primitives: par and pseq.

par, pseq :: a -> b -> b -- parallel/sequential composition

The par primitive identifies potential parallelism using lazy futures (Mohr et al., 1991). Its first argument is a closure that is marked for possible parallel execution (sparked). It

\( \phi(i) \) counts how many numbers smaller than \( i \) are coprime to \( i \). It can of course be computed more efficiently using a prime factorisation of \( i \). The naïve version shown here is just for benchmarking.
returns the value of its second argument. So, \( x \; \text{par} \; e \) marks \( x \) for parallel execution, and then returns \( e \). The \text{pseq} primitive sequences the evaluation of its arguments, returning the value of its second argument. So, \( x \; \text{pseq} \; e \) first evaluates \( x \), and then returns \( e \).

Experience has shown that unstructured use of \text{par} and \text{pseq} can quickly obscure programs. A higher level of abstraction is provided by evaluation strategies (Trinder et al., 1998; Marlow et al., 2010), which cleanly separate coordination from computation. Evaluation strategies are lazy, polymorphic, higher-order functions that control the evaluation degree and parallelism of a Haskell expression. Some primitive strategies are shown in Figure 2. The \text{rseq} and \text{rpar} strategies are analogous to \text{pseq} and \text{par}. The \text{rdeepseq} strategy is similar to \text{rseq}, except that it fully evaluates its argument\(^3\). The \text{parList} strategy is an example of a higher-order strategy. This takes another, possibly parallel, strategy as its argument, and applies it in parallel to all elements of a list. So, for example, \text{parList rdeepseq} is the strategy that evaluates a list in parallel, forcing all the results to be completely evaluated. Finally, the using function applies a strategy to a Haskell expression. So, \( e \; \text{using} \; \text{parList rdeepseq} \) will evaluate \( e \) using the \text{parList rdeepseq} strategy.

\begin{verbatim}
sumEulerGpH0 :: Int \rightarrow Int
sumEulerGpH0 n = sum (map euler [n, n-1..1])
    \'using\' parList rdeepseq

sumEulerGpH :: Int \rightarrow Int \rightarrow Int
sumEulerGpH z n = sum (map workF (unshuffle z [n, n-1..1]))
    \'using\' parList rdeepseq
where workF = sum \circ map euler
\end{verbatim}

Fig. 3. GpH versions of \text{sumEuler}

Our first GpH version of \text{sumEuler}, \text{sumEulerGpH0} (Figure 3), simply uses \text{parList rdeepseq} to compute every application of \text{euler} in parallel. However, the granularity of

\(^3\) i.e. it evaluates to \textit{full normal form} rather than the Haskell default of \textit{weak head normal form}. The \texttt{NFData} class defines the recursive evaluation strategy to implement this full evaluation.
the parallelism is too fine, since each spark only describes a single function application.
sumEulerGpH therefore uses chunking to overcome this. We therefore introduce an additional parameter $z$, the desired number of chunks (i.e. sparks). The $n$ list elements are then distributed into $z$ sub-lists. Each sub-list becomes a spark that may then potentially be evaluated in parallel. Alternative chunking functions can be easily defined as standard Haskell definitions⁴. For example, as shown in Figure 4, splitAtN splits a list into sub-lists of size $n$, and unshuffle distributes elements of a list into exactly $n$ sub-lists, in a round-robin fashion. For sumEulerGpH, we have chosen to use unshuffle rather than splitAtN because the computation time for the $\phi$-function increases with its argument. If splitAtN was used instead, then the largest argument values would be gathered into the first sparks that were created, creating load-imbalance between the sparks.

⁴ Chunking functions can also be provided abstractly for lists and other data types via a Cluster type class (Totoo & Loidl, 2014).

### 2.2 Eden

In contrast to GpH, Eden provides explicit process abstraction and process instantiation operations (Figure 5). The expression \( \text{process } ( \lambda x \rightarrow e ) \) of type \( \text{Process } a \rightarrow b \) denotes a process abstraction over a function \( \lambda x \rightarrow e \) of type \( a \rightarrow b \). The (#) operator allows a new child process to be instantiated, i.e. evaluated in parallel with the calling parent process. process \( ( \lambda x \rightarrow \text{map factorial x} ) \) # [1..100] creates a process that maps the factorial function over the list [1..100]. This list is evaluated on the parent, and passed in a fully evaluated form to the child. Once computed, the result list is then returned to the parent. Eden’s spawn function lifts process to a list of processes and inputs. All processes are created eagerly in parallel. This avoids the sequential demand-driven
evaluation that the lazy Haskell list would imply. Once instantiated, the new child process will be executed in parallel with the parent process. The parent and child processes have independent heaps, and will only communicate by exchanging (implicit) messages. Data that is exchanged between Eden processes is always in normal form (i.e. fully evaluated). Therefore, process instantiation is roughly equivalent to hyper-strict function application. Lists are communicated as streams, element by element. Tuple components are evaluated by concurrent threads. This allows for circular programs and infinite data streams.

Figure 6 shows how \texttt{sumEuler} can be implemented in Eden by spawning a number of independent processes that each apply \texttt{euler} to part of the original input list. Having created these processes, the main process will block until the results become available, and then compute the final sum. Since, unlike GpH, Eden will create threads for every process construct that is specified by the programmer, we introduce a use of the \texttt{unshuffle} function to balance the load between child processes. This will evenly divide the inputs among the number of available processors, \texttt{noPe} (provided as a constant by the Eden RTS).

\begin{verbatim}
spawnProcessAt :: Int \to IO () \to IO ()
data ChanName' a
createC :: IO (ChanName' a, a)
createCs :: Int \to IO ([ChanName' a], [a])
sendWith :: Strategy a \to ChanName' a \to a \to IO ()
sendStreamWith :: Strategy a \to ChanName' [a] \to [a] \to IO ()

noPe, selfPe :: IO Int
\end{verbatim}

Fig. 7. The complete API for the Eden implementation language, EdI

### 2.3 EdI

The Eden implementation language EdI (Berthold & Loogen, 2007; Berthold, 2008) (Figure 7) takes the approach of explicit process control even further, requiring the programmer to take full control of, and responsibility for, all data transfers. As suggested by its name, EdI was designed to implement Eden, but can also be used in its own right as a monadic language for parallelism. As in Eden, a new parallel process can be spawned on a specific instance of a running system but, unlike Eden, an EdI process is simply an IO action. EdI processes communicate explicitly using typed one-to-one channels. Channels are created using \texttt{createC}, which returns a \textit{channel name}, of type \texttt{ChanName' a} and a placeholder result of type \texttt{a}. A channel name can be explicitly communicated to another
sumEulerEdi :: Int \rightarrow IO Int
sumEulerEdi n = do pes ← noPe
  (cs, rs) ← createCs pes
  zipWithM_ (spawnEulerW pes) [1..pes] cs
  return (sum rs)

where spawnEulerW :: Int \rightarrow Int \rightarrow ChanName' Int \rightarrow IO ()
  spawnEulerW stride k c = spawnProcessAt k
    (sendWith rseq c
      (sum (map euler [n-k+1, n-k+1-stride..1])))

Fig. 8. EdI implementation for sumEuler

process, using \texttt{sendWith}, or implicitly passed to a new process when it is created using \texttt{spawnProcessAt}. The other process can use the channel to return data (of a suitable type) to the original process. Note that the \texttt{sendWith} operation takes an evaluation strategy argument. This controls where the evaluation happens and is applied to the data argument before it is transmitted. Finally, the \texttt{sendStreamWith} operation extends \texttt{sendWith} to streams. This enables the definition of circular process networks and pipelining.

Figure 8 shows how \texttt{sumEuler} can be implemented in EdI. A set of channels, \texttt{cs}, and placeholders, \texttt{rs}, are created using \texttt{createCs}. One channel and one placeholder are created for each available processor (\texttt{noPe}). Each channel is used in one call to \texttt{spawnEulerW}, which creates a child process on processor \texttt{k} using \texttt{spawnProcessAt}. The new process will generate its own part of the input list, apply the \texttt{euler} function to this partial input, and sum the resulting list. This partial sum is returned to the parent by calling \texttt{sendWith} on the given channel. EdI represents the lowest level of abstraction that we will consider in this paper, exposing the full control, and burden, of parallel evaluation and explicit communication to the programmer. It constitutes a minimal set of runtime-system support functionality for sending and receiving data and computations, exploiting the underlying GHC runtime system to achieve synchronisation.

3 Haskell on shared-nothing systems: The PAEAN runtime framework

The language characteristics of Eden, GpH, and EdI can be classified into the four key concepts of parallelism, binding, synchronisation and workload management (Table 2). Each of these will lead to a number of important implementation requirements.

Parallelism. As we have seen, in Eden and EdI, all threads that are specified by the programmer must be created (parallelism is mandatory). In contrast, in GpH, the runtime system decides both when and whether to instantiate the potential parallelism that has been indicated by the programmer (parallelism is advisory). The corresponding requirements for thread management are:

- to provide mechanisms to support the creation of parallelism;
- to decide on whether or not to instantiate potential parallelism;
- to decide on when to instantiate potential parallelism.
Table 2. Language characteristics of GpH, Eden, and EdI, and implementation concepts

<table>
<thead>
<tr>
<th>Concept</th>
<th>Eden</th>
<th>EdI</th>
<th>GpH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallelism</td>
<td><strong>Explicit process constructs</strong></td>
<td><strong>Explicit process constructs</strong></td>
<td><strong>Implicit par annotations</strong></td>
</tr>
</tbody>
</table>
| Threads            | implicit threads
                     mandatory parallelism
                     explicit task distribution | explicit threads
                     mandatory parallelism
                     explicit task distribution | implicit threads
                     advisory parallelism
                     implicit task distribution |
| Binding            | data exchange through process application                            | variables act as futures                                            | virtual-shared memory                                                |
| Memory mgt.        | distributed heaps (per process)                                       |                                                                      |                                                                      |
| Synchronisation    | **Implicit through process arguments and results**                     | explicit, including evaluation control                               | implicit variable sharing
                     implicit and hyper-strict
                     (on process arguments)    |                                                                      | implicit and lazy (on shared data)                                  |
| Communication      |                                                                      |                                                                      |                                                                      |
| Workload Management| **Explicit logical processor ids**                                    |                                                                      | **Implicit**                                                        |
| Workload distribution | default (round-robin) or explicit process placement                  |                                                                      | automatic spark distribution (work-stealing)                        |

**Binding.** The concept of variable binding is realised in Eden and EdI via data transfer, initiated by a process creation, and in GpH via lazy futures and implicit synchronisation. This concept is implemented by the memory management component. For Eden, this links remote data access with the language concept of process abstraction, requiring explicit data transfer. For GpH, the main requirements are completely transparent access to bindings, which leads to automatic and distributed memory management over several physical heaps, and hence to a virtual-shared memory abstraction. More specifically, the requirement on the runtime system is to ensure that all data is available at the location where the computation is to be executed. This implicitly invokes communication, so establishing a tighter link between these components in the GpH implementation. In contrast, the runtime system for Eden and EdI must implement a model of distributed heaps that are connected through communication channels. The requirements for parallel memory management are:

- to identify points of data exchange;
- to implement a logically shared address space on top of distributed memory;
- to interact with the communication component based on the need for remote data.

**Synchronisation.** Synchronisation is required to i) avoid evaluating a value multiple times; ii) communicate values between threads; and iii) ensure that results are shared between the evaluating thread and the threads that use those results. In EdI, the programmer is responsible for all communication and for ensuring that evaluation is performed appropriately. In Eden, the argument to a process is fully evaluated to full normal form by the parent process and communicated incrementally. The child process will block if it requires some part of the data that has not yet been evaluated and communicated to it. In contrast, in
GpH, some data may be passed to a child thread when it is created, but the majority will be fetched on a demand-driven basis. This increases sharing and reduces startup costs, but may add some delay if a significant volume of data is fetched during execution. Having created one or more child processes, in Eden the parent process will typically block until the results become available. In GpH, however, execution continues normally. If any GpH thread requires a value that is being evaluated by another thread, it will block until the result is produced. When the value of a virtually shared node becomes available, it will be returned to all blocked threads. This reduces synchronisation delays, but once again, shared data may be fetched incrementally. This uses an extension of the standard “black-holing” mechanism that is normally used to detect cycles in sequential code. Rather than raising an exception, as would normally happen, the parallel implementation treats an attempt to evaluate a node that is currently under evaluation as a synchronisation request. The virtual shared graph model used in GpH will ensure that the results of computations will be shared as they become available, so avoiding repeated evaluation. In contrast, Eden and EdI make no attempt to avoid repeated evaluation other than through transmitting only normal forms.

The requirements for communication are therefore:

- to check whether necessary data is available;
- to realise data exchange when necessary;
- to update values with results as they are produced;
- to notify blocked threads of the availability of a result;
- to interact with the thread management component, based on data availability.

Workload Management. A key runtime decision is when and where to execute parallelism. At the language level, Eden and EdI both provide a notion of virtual processor, which can be used by the programmer to offload processes to specific processors. GpH abstracts this even further: threads are instantiated from sparks and then mapped to the available processors by the runtime system. In order to get a good load balance, it may be necessary to migrate sparks and/or threads between processors. Finally, where multiple threads exist on a single processor, it is necessary to schedule them appropriately both for efficiency and, in the case of Eden and EdI, to ensure fairness. The requirements for workload distribution are:

- to decide on the order of execution of available threads (scheduling);
- to determine how to distribute the available parallelism (load balancing);
- to decide whether and when to offload parallelism.

As shown in Figure 9, PAEAN realises each of these issues as its own modular component. This makes PAEAN much more generally applicable. Although they have formed our starting point, the PAEAN system design is not restricted to the GpH and Eden parallelism models. Neither is it necessarily limited to Haskell: the concepts and realisations that we have shown in Table 2 are characteristic of a much wider class of languages, including some parallel dialects of Lisp, Prolog and C++. Any language that is more prescriptive in terms of thread management can easily be mapped to the underlying PAEAN realisations, though it may not make full use of the PAEAN mechanisms.

A full description of this mechanism can be found in (Hammond, 2011)
3.1 PAEAN Thread management

PAEAN coordinates multiple instances of the standard GHC runtime system. Each PAEAN instance is typically mapped onto a separate processor in its own operating system process. When run on a multicore, an instance may comprise multiple Haskell Execution Contexts (HECs) which share the same heap, i.e. the standard shared-memory parallel Haskell implementation of GpH, GHC-SMP (Marlow et al., 2009). These instances are connected to form a collaborating parallel system with a distributed heap, as shown in Figure 10. Each PAEAN instance manages its own thread pool and, if required, a spark pool, providing its own independent thread scheduler. In addition, PAEAN introduces the concept of a process, a thread group that shares a region of the Haskell heap and communication channels (explained in Section 3.4.3). As shown in Table 3, PAEAN routines are provided to: spark closures and add them to the local spark pool (createSpark); select a spark from the local spark pool (findSpark); turn a spark into a thread and move it from the spark pool to the thread pool (activateSpark); add an IO thread directly to the local thread pool (createIOThread); and start a new process (startNewProcess).

3.2 Virtual shared heap management

Following our shared-nothing design, each PAEAN instance maintains its own independent heap, with its own address space. Each instance undertakes its own local garbage collection (GC) on this heap, using, for example, the usual GHC garbage collection mechanism. By default, this is a stop-and-copy generational garbage collector (Appel, 1989). On a shared-memory system, the GHC-SMP shared-memory garbage collection implementation may be used, if desired, though this may degrade performance. It is possible to combine the heaps from multiple PAEAN instances to form a single virtually-shared heap that can be accessed by any PAEAN instance. Shared heap nodes are accessed through a shared global address space.
Table 3. PAEAN runtime system routines for sparks and threads

<table>
<thead>
<tr>
<th>PAEAN Routine and Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>void createSpark(StgClosure *la)</td>
<td>Generate a spark and put it into the local spark pool</td>
</tr>
<tr>
<td>StgClosure* findSpark()</td>
<td>Select and return a spark from the the local spark pool</td>
</tr>
<tr>
<td>void activateSpark(StgClosure *spark)</td>
<td>Turn a spark into a thread, allocating a thread, and put it into the thread pool</td>
</tr>
</tbody>
</table>

Table 4. PAEAN runtime system routines for global heap management

<table>
<thead>
<tr>
<th>PAEAN Routine and Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RtsGA* makeGlobal()</td>
<td>Create a new unique global address (GA) on the current node</td>
</tr>
<tr>
<td>StgClosure* lookupGA(RtsGA *ga)</td>
<td>Look up global address ga in the GIT, return a local address (LA) if found</td>
</tr>
<tr>
<td>RtsGA* lookupLocal(StgClosure *la)</td>
<td>Look up local address la, return GA if found</td>
</tr>
<tr>
<td>void commonUp(RtsGA *ga, StgClosure *la)</td>
<td>Check whether a ga already exists on this node; if so, keep only the further evaluated of the two (a data duplicate is eliminated)</td>
</tr>
<tr>
<td>void splitWeight(RtsGA *ga1, RtsGA *ga2)</td>
<td>Split weight attached to ga1, into 2 components adding up to the original, one attached to ga1 and one attached to ga2 (a reference is created)</td>
</tr>
<tr>
<td>RtsGA* addWeight(RtsGA *ga)</td>
<td>Add the weight in ga to the weight of the GA in the local GIT table (a reference is deleted)</td>
</tr>
<tr>
<td>void markGIT()</td>
<td>Traverse in-pointers section of GIT during GC (local closures with GAs are roots)</td>
</tr>
<tr>
<td>void rebuildGIT</td>
<td>Traverse GIT to fix references to local addresses (out-pointers), return weight for dead entries</td>
</tr>
</tbody>
</table>

3.2.1 Global addresses

The virtual shared heap forms a subset of the individual PAEAN instance heaps. Each (potentially) shared heap object is identified by a unique global address (GA), created by makeGlobal. Unshared objects exist only in the local heap for the PAEAN instance and so do not have a global address. Only selected unevaluated objects need to be shared, other objects referenced from them will be handled transparently, and evaluated data can simply be copied. This minimises the number of globally addressed objects and brings benefits in terms of performance (quicker lookup), space leaks (fewer external pointers), and memory usage (smaller pointers). This design is based on the assumption that only a small fraction of any given instance heap is globally visible, as verified on the GRIP architecture (Peyton Jones et al., 1987; Hammond & Peyton Jones, 1992). Each PAEAN instance maintains tables of in-pointers from other PAEAN instances (a Global In-Pointer Table, or GIT),
3.2.2 Transferring program graph between PAEAN instances

Every object which is moved from one PAEAN instance to another is given a new, unique GA in the target instance’s GIT. The original local heap node is replaced with a special FetchMe closure that refers to this GA. When the value of a FetchMe node is needed, the communication subsystem is invoked to obtain the value of the node from the remote PAEAN instance. In this way, local heap nodes are linked to remote heap nodes. The commonUp routine is used to avoid replicating data where the same GA is communicated to an instance more than once. An example is shown in Figure 11. Here, two thunks (nodes 1 and 3) are moved between heaps, replacing the original nodes with FetchMe nodes. The packing routine checks whether the thunks are already global. If they are not, GAs are allocated locally for each thunk (GA1.1–GA1.2). These GAs are used to refer to the thunks while they are in transit. When the thunks are unpacked, new GAs (GA2.1–GA2.2) are generated to indicate the transfer of ownership to Instance 2. An acknowledgement message informs Instance 1 that GA1.1 and GA1.2 have been superseded by GA2.1 and GA2.2. The latter become the permanent remote references, whereas the former become garbage. Figure 11 shows the final state, with FetchMes on Instance 1 referring to GAs on Instance 2 (GA2.1–GA2.2).
3.2.3 Global Garbage Collection

A weighted reference counting scheme (Bevan, 1987) is used to garbage collect the independent PAEAN heaps. As we have seen, the GIT indicates the possibly live in-pointers for each instance. These are marked as roots during GC using the markGIT routine. Weights are associated with each GA. Whenever a GA is shared between instances, e.g. because a FetchMe has been replicated, then its weight is split between the two relevant GIT entries using splitWeight. Conversely, when an instance frees a remote closure during its local garbage collection, the weight is returned to the original GIT using addWeight. The same happens when old GAs are replaced by new ones. When all the weight is concentrated in the original GIT, then the closure is no longer referenced globally and the GIT entry can be removed. The local closure can then be freed if it is not referenced locally. One general weakness of this scheme is that it will not collect cycles across independent heaps. Various schemes can be used to deal with this issue, if needed, such as collating cross-heap cycles onto a single instance, where they can be collected (Bevan, 1987). Such a mechanism, however, is not currently implemented in PAEAN.

3.3 Haskell program graph serialisation

Haskell program graphs are communicated between runtime system instances as serialised sub-graphs (see Figure 12). These are packed into one or more packets for efficient communication. PAEAN provides methods to traverse a graph in a local heap, packing it for efficient transmission (packGraph), and to unpack a serialised graph into the remote heap (unpackGraph), restoring its original structure and building any inter-instance links using new FetchMe closures. The serialisation concept used in PAEAN is a breadth-first traversal that is designed to limit the amount of communicated graph to a single efficiently-transmitted message. This was first used in (Trinder et al., 1995) as a modification of the original Graph for PVM mechanism (Loidl & Hammond, 1994; Hammond, 1993) that packed entire sub-graphs. (Loidl & Hammond, 1996) studies the best strategies for efficiently packing shared program graph. We have followed this work here.

<table>
<thead>
<tr>
<th>PAEAN Routine and Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rtsPackBuffer* packGraph(StgClosure *la)</td>
<td>serialise a sub-graph rooted at la</td>
</tr>
<tr>
<td>StgClosure* unpackGraph(rtsPackBuffer *buf)</td>
<td>de-serialise a buffer, returning a local address</td>
</tr>
</tbody>
</table>

3.3.1 Closure Layout and Serialisation

Most GHC closures are laid out in the heap as a header section containing the key metadata, followed by any pointers to other closures, and then by all the non-pointers. The first header field is an info pointer, which points to a record containing information such as the number of pointers and non-pointers for a particular heap closure, plus a method that can be used to evaluate the closure. During packing, the Haskell program graph is traversed and serialised so that it can be written into the communication packet. Every
Fig. 12. Example serialisation of the five-closure graph from Figure 11 with two shared closures.

The packet on the left encodes the graph on the right, using REF tags to describe sharing.

unique closure in the packet is identified using a CLO tag, followed by the global address of the packed closure. The header and all non-pointer data is packed directly into the packet immediately after the GA. Pointers into the local instance heap will, however, need to be re-established once the program graph has been successfully transmitted to the remote heap. They are therefore omitted from the packet. The order in which the graph is reconstructed allows them to be correctly re-established. Where a packed graph contains shared closures, explicit REF tags are used to preserve sharing. They also allow cyclic structures to be transmitted. Finally, top-level program constants (constant applicative forms) do not need to be transmitted between PAEAN instances, since each instance maintains its own identical collection of such constants, embedded in the binary program that is executed by all instances. These are identified using a CAF tag. Clearly, the graph referenced from the root might be larger than could fit into a single fixed-size packet. In this case, PAEAN will either use FetchMe closures to reference the additional heap, or will raise an error.

A few special types of GHC closures cannot be packed using the mechanism described here. These include primitive GHC data structures for synchronisation: Haskell mutable variables, MVars, and transactional variables that used to implement software transactional memory, STM, as well as some system-level data structures. Packing these nodes would not be sensible since they encode state that is local to a specific PAEAN instance, such as operating system file handles.

3.3.2 Example: Packing/Unpacking Graph

Figure 12 shows how the graph from Figure 11 can be serialised into a single packet and transferred between PAEAN instances. Here, closures 2 and 5 are shared using REF tags. The packGraph function traverses the graph in a breadth-first manner. This means that if the complete sub-graph cannot be packed, then a closure will generally be packed with all its children. In the example, closures 1 and 3 are thunks, and are therefore replaced by FetchMe closures in the sender’s heap. The packGraph method assigns temporary global addresses, GA 1.1 and GA 1.2, to the two thunks and includes them in the serialised structure. During de-serialisation, these temporary addresses will be replaced by the final addresses that are assigned by the receiver, GA 2.1 and GA 2.2. The same serialisation process can also be used without taking advantage of the PAEAN virtual shared heap mechanism. In this case, the GA fields are omitted and no GIT table is maintained by
**Table 6. PAEAN runtime system communication routines.**

<table>
<thead>
<tr>
<th>PAEAN Routine</th>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rtsBool MP_start</td>
<td>Program arguments</td>
<td>Start/synchronise all nodes</td>
</tr>
<tr>
<td>rtsBool MP_sync</td>
<td>Program arguments</td>
<td>Start/synchronise all nodes</td>
</tr>
<tr>
<td>rtsBool MP_quit</td>
<td>error code</td>
<td>disconnect from system</td>
</tr>
<tr>
<td>rtsBool MP_send</td>
<td>receiver, tag, data</td>
<td>(non-blocking) data sending</td>
</tr>
<tr>
<td>int MP_recv</td>
<td>destination buffer</td>
<td>(blocking) data reception</td>
</tr>
<tr>
<td>rtsBool MP_probe</td>
<td>void</td>
<td>(non-blocking) probe for available messages</td>
</tr>
</tbody>
</table>

**Table 7. PAEAN message tags and protocols**

<table>
<thead>
<tr>
<th>Message tag</th>
<th>Meaning / Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Sync:</td>
<td>READY</td>
</tr>
<tr>
<td>PETIDS</td>
<td>Sync. message from main to non-main instances</td>
</tr>
<tr>
<td>FINISH</td>
<td>Shutdown signal (from one non-main instance to main instance, or broadcast from main to all non-main instances)</td>
</tr>
<tr>
<td>FAIL</td>
<td>Failure indicator (middleware to main instance)</td>
</tr>
<tr>
<td>Data transfer:</td>
<td>CONNECT</td>
</tr>
<tr>
<td>DATA</td>
<td>receiving instance replaces heap placeholder by arriving data</td>
</tr>
<tr>
<td>STREAM</td>
<td>receiving instance modifies heap placeholder, adding arriving data into a list</td>
</tr>
<tr>
<td>FETCH</td>
<td>receiving instance sends back requested data (identified by a global address)</td>
</tr>
<tr>
<td>Global Addresses:</td>
<td>ACK</td>
</tr>
<tr>
<td>NACK</td>
<td>sender indicates failure to receive thunks, receiver will revert thunk</td>
</tr>
<tr>
<td>Process control:</td>
<td>RFORK</td>
</tr>
<tr>
<td>TERMINATE</td>
<td>receiving instance terminates a Haskell thread (identified by thread ID)</td>
</tr>
<tr>
<td>Work distribution:</td>
<td>FISH</td>
</tr>
<tr>
<td>SCHEDULE</td>
<td>receiver will create new thread to evaluate data (answered by ACK or NACK)</td>
</tr>
</tbody>
</table>

the receiving instance. In place of global addresses, global references are established by communication channels connected to heap cells. These can be created from within Haskell source programs using the createC primitive operation (Section 3.4.3).

### 3.4 The PAEAN communication subsystem

PAEAN provides a number of communication primitives that allow instances to communicate using asynchronous messages. The RTS implementation provides event-handling routines that allow messages to be received at certain safe execution points, and that link to one of several low-level transport libraries.

#### 3.4.1 Low-level Communication Primitives

The PAEAN design is based on simple asynchronous point-to-point message passing communication, plus a mechanism for structured startup and shutdown of PAEAN instances,
Table 8. PAEAN primitive operations (callable from Haskell)

| data Mode = Stream | Data Connect | Instantiate Int | data modes: Stream or Single data special modes: Connection, Instantiation |
| createC :: IO ( ChanName' a, a ) | create a channel and placeholder |
| connectToPort :: ChanName' a → IO () | receiver registration |
| sendData :: Mode → a → IO () | send data to registered receiver |
| fork :: IO () → IO () | new thread in same process (Conc.Haskell) |
| noPe :: IO Int | number of instances |
| selfPe :: IO Int | ID of own instance (1..N) |

where one PAEAN instance acts as the main instance. As shown in Table 6, message passing is implemented by MP_send (non-blocking send) and MP_recv (blocking receive). These are complemented by the (non-blocking) MP_probe, which tests whether messages are available to receive. MP_start and MP_sync start and synchronise PAEAN instances. MP_quit performs a controlled shutdown of the system. Different middleware can be used to implement this communication sub-system. Existing implementations can use both PVM (Geist, 2011) and MPI (MPI Forum, 2012). The shared-nothing message-passing mechanism can also be implemented on top of physically-shared memory. The Eden implementation supports POSIX shared memory, Windows shared memory, and Windows Mailslots.

3.4.2 High-level messages

A high-level set of messages is built on this low-level implementation. Table 7 summarises all the PAEAN message types and their semantics. Many kinds of messages are sent internally by the runtime system itself: system messages are exchanged during startup and shutdown to synchronise and coordinate instances; data transfer messages (Section 3.2.2) request globalised data (FETCH); global address messages ACKnowledge a new GA or indicate failures (NACK); and work distribution messages (Section 3.5) search for work (FISH) and pass on new work (SCHEDULE).

3.4.3 Explicit communication via channels

PAEAN also provides facilities to enable explicit communication and process control directly from within Haskell, using a number of new primitive operations (Table 8). These primitives may be preferred where a full shared heap is not required, and allow direct communication between two instances, forming a dual to the FetchMe closures described above. EdI uses the primitives directly as a simple explicit language for parallel coordination. Figure 13 shows a simple example of using the explicit communication primitives, which provide functionality for creating and using typed channels, as well as instantiating computations on nodes.
PAEAN: Portable, Scalable Runtime Support for Parallel Haskells

---

Fig. 13. Communicating computations and values explicitly. A thread on Instance 1 creates a channel (1) and passes the created `ChanName` to a new process on Instance 2 (2). A newly created thread connects to the channel (3) and sends evaluated `Data` (4).

---

```
-- parent process
do -- create reply channel (1)
    (ch,bkh) ← createC
-- spawn remote process (2)
sendData (Instantiate 2)
    (thread2 ch)
-- next function call will
-- block until response arrives
print bkh
```

---

```
-- child process
-- thread2 ch =
do connectToPort ch -- (3)
    let dat = fib 42
    rnf dat `pseq` -- (4)
    sendData Data dat
-- terminated
```

---

```
-- insert a closure into a local HEC's spark pool:
par# :: a → b → b
-- as above, but with locality control, specifying min and max distances:
parDist# :: Int → Int → a → b → b
```

---

Fig. 14. Spark insertion primitives in GHC-SMP and PAEAN

---

### 3.5 Work distribution through Offloading and Work Stealing

In a shared-nothing system, work will need to be moved from instances that have excessive amounts of work to those that are idle. When and how this is done is an important policy choice that can significantly affect performance. There are two basic approaches: work can be offloaded *eagerly* at the time of creation (active load distribution); or it can be offloaded *lazily*, e.g. in response to changes in system load (passive load distribution).

Eager work offloading aims to quickly saturate the parallel system. This is beneficial in cases of regularly structured parallelism where a single main task generates all the available parallelism at the start of the execution. Such a policy is easy to implement using PAEAN’s remote execution mechanisms, as used by Eden. Typically, under such an approach, all the data that a thread requires will also be sent with the thread when it is offloaded. This may, of course, add significant delay before a thread can start. When used with virtual shared memory, it also runs the risk of significant heap fragmentation and increased communication costs. In contrast, lazy work offloading is typically implemented by some kind of *work stealing* mechanism, where idle instances attempt to obtain work from busy instances. PAEAN provides flexible support for distributing work lazily across
3.5.1 Example of Work Stealing

Figure 15 illustrates this work-stealing load-balancing mechanism. When Instance 2 becomes idle, it sends a work request message (or FISH) to another processor. In this case Instance 1 is targeted. Since Instance 1 has some sparks in its spark pool, it responds by sending one of these sparks to Instance 2 in a SCHEDULE message, serialising the graph that it refers to, as described in Section 3.3, and replacing the original sparked closure by a FetchMe. The program graph will be de-serialised when it is received on Instance 2. The spark is added to its local spark pool, whence it can be turned into a thread using the usual thread scheduling mechanisms, or even passed on to another PAEAN instance if Instance 2 has gained sufficient work. If a targeted instance has no sparks, it forwards the FISH message a fixed number of times. If no instance has any sparks, the FISH message is sent back to the originator. A back-off delay is introduced, and a new FISH message will be sent (Trinder et al., 1995).

Typically, execution of the thread that created the spark on Instance 1 depends on the result of this spark that has been moved to Instance 2, and other threads might also depend on it. When execution requires the exported spark, the thread will be added to a local queue of blocked threads. The FetchMe will then be followed, and a FETCH message will be sent

---

6 This helps avoid starvation. It would be possible, but highly unlikely, for a spark to be transferred between the same instances multiple times, but only where the instances gain sufficient work elsewhere. Although some effort would then be wasted, progress would have been made, so termination is ensured.
Fig. 16. Low- and high-watermark mechanisms for spark distribution. (showing a single spark pool). By default work-stealing (passive) load distribution is used. If the size exceeds the higher-watermark, work is actively offloaded. The system aims to always fill the pool up to the low-watermark.

3.5.2 Work Distribution Mechanisms for Hierarchical Systems

We have made several enhancements to our work distribution mechanism in order to improve performance for large-scale systems, which typically involve clusters of multi-core processor nodes connected using high-latency networks. Such systems are usually hierarchical with widely varying inter-node latencies, i.e. they are genuinely non-uniform memory architectures. The largest system we have studied is Edinburgh Parallel Computer Centre’s HECToR, which has over 90,000 cores and a complex hardware interconnect. Section 4 shows scalability results from a smaller, but still realistic, hierarchical cluster.

Watermarks: One simple, but flexible, mechanism that gives better control of spark distribution is to use low- and high-watermarks for each spark pool. Using this approach, work offloading decisions are based on the sizes of each spark pool, as shown in Figure 16. The low-watermark specifies a minimum number of sparks that should be held in the local spark pool. If the number of sparks falls below this watermark, no sparks will be exported, and the instance will try to obtain additional sparks from other instances. The high-watermark indicates the maximum number of sparks that should be held in a spark pool. If the number of sparks exceeds this limit, the instance will use SCHEDULE messages to actively offload sparks to other instances without being asked for work. In other words, the instance will temporarily and locally switch from lazy load distribution to eager load distribution, until the spark pool size drops below the high-watermark. If all instances have large numbers of sparks, a back-off mechanism is used to delay SCHEDULE messages, as described above for FISH messages. Our experiments identify this mechanism as the most important one for tuning cluster performance (Aljabri et al., 2013). Measurements on a 32-node cluster with up to 100 cores show a reduction in runtime on a set of micro-benchmarks of up to 57%, i.e. in the best case, we can gain a speedup of more than 2 over the standard FISHing policy that we described above. For larger benchmarks, improvements range from 16%
to 28%. Further improvements are possible, by dynamically adapting the low-watermark over the runtime of the program. In more recent work (Aljabri, 2015) (Section 5.4), we have extended the scalability measurements to a combination of two remote clusters with 300 cores in total. Using our sumEuler example, we have obtained speedups of up to 140 on this more heterogeneous, distributed architecture.

**Locality control:** Controlling locality is important for ensuring good performance on hierarchical shared-nothing systems, especially those with high latency connections (Loidl, 1998; Loidl, 2001). PAEAN provides several enhanced primitives to support locality control. The most powerful of these primitives is `parDist` (Aswad, 2012) (Figure 14), which takes two additional integer arguments, specifying a lower and an upper limit on the distance of work distribution from the current instance. If these are both 0, then this forces local evaluation. Setting a minimum limit forces work to be distributed through the system. This is useful in low-load situations, such as system startup.

### 4 Evaluation of scalability and performance

We have carried out a series of experiments to assess the scalability and performance of our GpH, Eden and EdI implementations using the versions of `sumEuler` from Section 2. Our experiments used a commodity cluster of 32 nodes, each with an 8-core x86_64 CPU. The cluster therefore provides a total of 256 processor cores. Since the full machine was not available for our exclusive use, the measurements reported here are limited to a maximum of 64 cores on 16 nodes (i.e. using at most four cores per node). We delegate the placement of PAEAN instances to the communication middleware. Both MPI and PVM prefer using different nodes to using different cores on the same node. This balances the load at the cost of increased network communication. For our experiments, we have mainly used PVM, but, where indicated, some of the EdI results have also used MPI. The implementations used in our measurements are the Eden/EdI system version 7.8.3 (mid 2014), and a research version of GpH on clusters, GUM-SMP, based on GHC 6.12.3.

#### 4.1 Measured Speedup

Figure 17 shows measured speedups for each `sumEuler` variant, with \( n = 90,000 \). We can clearly see that the sumEuler program gives near-linear speedup for all three implementations, even when large numbers of cores are used. Overall, the EdI implementation shows the best speedups. The MPI implementation is slightly better than PVM. The Eden version comes close to EdI. However, since the GpH version requires more management (i.e. overhead), it delivers worse speedup. It does, however, scale robustly (we have confirmed this up to 180 cores). The difference between the EdI and the Eden versions can be explained by differences in communication behaviour: while the Eden version communicates the full list of arguments to each child process as a stream, the EdI version requires the child processes

---

7 The beowulf cluster at Heriot-Watt University Edinburgh. Each node is built from 2 quad-core Xeon E5506 2.13GHz, with 256kB L2 and 4MB shared L3 cache, connected by gigabit Ethernet, and running CentOS 6.4.
to generate their arguments from parameters (size, offset, stride), and communicates only the results. This reduces communication costs. The GpH version suffers in this example due to the higher overheads associated with managing the virtual shared heap. Since the work is already well distributed through the structure of the problem, the automatic load balancing mechanism does not improve performance in this case. It would, however, deal better with situations where some nodes were executing external workloads, and were not dedicated to the Haskell program. Overall, the sumEuler example clearly demonstrates the ease of parallelisation and simplicity that is claimed for parallel functional programming. It was straightforward to achieve the near-perfect speedup reported here, since the problem structure itself is intuitive and easy to parallelise. Note that the speedup we obtain is dependent on the problem size. For smaller problem sizes, where \( n < 90,000 \), it would scale less well with increasing number of cores, since the constant overhead of instantiating the program and using the middleware would have an increasing effect.

### 4.2 Scalability

While the sumEuler example is easy to decompose into independent parallel tasks, many parallel algorithms have more complex structures with inter-task dependencies. One example is the N-Body simulation, which simulates the movement and gravitational acceleration of a number of “particles” (physical bodies) in a 3-dimensional space. The simulation is typically approximated iteratively in a number of discrete time-steps. A simple straightforward parallelisation of this problem will exploit parallelism within each iteration, assigning subsets of particles to different threads, where each thread computes the new velocity and position for its own particles. In order to update the velocities for its own particles, each thread needs the position and mass of all the other particles. This is approximated by exchanging new information after each iteration, in a potentially costly global exchange of data. We have measured the performance of an Eden implementation of N-Body based on iteration skeletons (Dieterle et al., 2013). These use the allGather skeleton (Dieterle
et al., 2010) to implement the global data exchange. Figure 18 shows the speedups that we obtain with this implementation for a problem size of 20000 particles. For this problem size, the program scales reasonably well, achieving a maximum speedup of around 30 on 64 cores. Performance degrades when more than two cores per node are used. Once again, the MPI back-end performs similarly to, but slightly better than, its PVM counterpart.

The iterative structure of this algorithm means that each iteration is performed sequentially. Figure 19 shows the overhead when the algorithm performs more iterations while the amount of computation is kept constant ($N^2 \cdot k$ computations for $N$ particles and $k$ iterations). We can see that speedup degrades earlier and is generally lower when more iterations are performed, since each iteration requires a global synchronisation.
As a baseline for our shared-memory results, we again use the \texttt{sumEuler} program on a 48-core machine\footnote{Four AMD Opteron-based processors, each with two NUMA regions, where each region has six 2.8GHz cores; total RAM is 512GB, with 64GB per region; there is a 2MB L2 cache, shared between every 2 cores in each region, and a globally shared 6MB L3 cache.} to compare the performance of the GpH and Eden implementations. Figure 20 shows the results that we obtain. While the speedups are fairly modest for this simple parallel program, we observe a steady increase in speedup for all versions, up to 28 on 46 cores in the best case for Eden with an ShMem shared-memory back-end. All implementation show a drop in speedup when using all 48 cores. This is consistent with previous experimental results on similar systems, where we have found that allocating all available processors to the parallel execution can interfere with system process execution. As with the cluster results, the GpH speedups are slightly worse than those for Eden, but notably better than those for the GHC-SMP shared-memory implementation. This matches with our previous study of the impact of NUMA architectures on parallel Haskell performance (Aljabri et al., 2014). We summarise our findings below. These findings are also consistent with parallel benchmarks, across languages, on this NUMA architecture. Due to the different memory latencies to different memory banks in NUMA architectures, even simple parallel programs struggle to efficiently exploit this kind of architectures.

For a more detailed evaluation of the performance of the PAEAN shared-nothing approach on a physically shared-memory system, we have chosen to use the \texttt{MatMult} benchmark program from the \textsc{nofib} suite\footnote{http://git.haskell.org/nofib.git}. Matrix multiplication can be easily decomposed, with independent calculations, so should give good scalability, in principle, provided that the thread granularity is managed carefully. We compared the scalability of the original GHC-SMP implementation against that of an equivalent Eden version. The Eden version uses a variant of Cannon’s algorithm, communicating matrix blocks in a toroidal topology,

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{sumEuler.png}
\caption{Speedup of \texttt{sumEuler} Variants on an 48-core machine: $n = 90000$}
\end{figure}
using a recursive toroid skeleton (Berthold & Loogen, 2005). The GHC-SMP version uses a similar blockwise task distribution, but tasks communicate via the shared heap without using a specific communication topology. Hence, the GHC-SMP version can freely choose the block size to improve load balancing, but might encounter bus contention when doing so. Since we are evaluating scalability on shared-memory architectures, we have used a 64-core physically shared-memory system\textsuperscript{10}. Our speedup results are shown in Figure 21. They clearly show that the Eden version can exploit the available cores, delivering a peak speedup of approximately 54 on 64 cores. The GHC-SMP shared-memory version delivers the best speedups on small numbers of cores, but flattens out at about 16 – 17×. The Eden versions scale better when using more than 25 cores; all PAEAN shared-nothing implementations have speedup that is equivalent to or better than GHC-SMP from that size onwards. This is probably an effect of shared-memory garbage collection. Note that the speedup for the Eden versions does not scale linearly with the number of cores. The toroidal algorithm delivers a “stepped” speedup profile, that is ideally matched to a square number of cores. The Eden program attempts to utilise all the cores by placing more than one process on some of them when approaching square numbers. This however, hardly yields any better speedup.

We note, in passing, two interesting phenomena: Firstly, the GHC-SMP version is faster than the equivalent sequential version without any spark creation when executed on only one core. However, it uses a considerably larger heap. Creating sparks reserves heap space early, and therefore accelerates execution. This effect is intensified when using column-wise instead of block-shaped sparks: column division exposes the worst access locality. On more cores, this effect quickly disappears, and the spark strategies only have a minor influence on the runtimes. Secondly, the Eden version has been measured with all three possible back-ends for data transfer: the shared-memory back-end, MPI, and PVM. As

\textsuperscript{10} Compute server \textit{lovelace}, University of St. Andrews: 4 \times 16 core Opteron 6376 1.4GHz with 2MB L2-cache (per core), 16MB shared L3 cache, 640GB shared RAM.
expected, PVM performs worse than MPI, and the shared-memory back-end yields the best performance. This exposes the overhead that is imposed by PAEAN communication sub-system middleware: the bespoke shared-memory back-end was tailor-made for its specific message-passing needs, without any additional features which would worsen performance.

Finally, we have recently performed a systematic evaluation of shared-nothing versus shared-memory parallel Haskell systems on a number of modern NUMA architectures (Aljabri et al., 2014). We have explored a range of configurations, including purely distributed memory, classic GUM; a mixture of shared-memory and distributed-memory, GUM-SMP (Aljabri et al., 2013; Aljabri, 2015); and purely shared-memory, GHC-SMP. Our measurements show that the shared-memory implementation consistently performs worst on these architectures, repeatedly hitting memory access and locking bottlenecks. A pure distributed-memory implementation improves runtimes by a factor of about 3 compared with the shared memory implementation. Further improvements can be achieved by configuring the system to use one shared heap for each NUMA region on the machine. These results support the findings reported above, underlining that even on the current generation of moderately parallel systems, a shared-nothing design delivers consistent and tangible performance benefits.

5 Related work

5.1 Built-in Parallelism Support in GHC: GHC-SMP

The standard GHC runtime system provides built-in support for parallelism (Marlow et al., 2009), but only on physically shared-memory systems. Drawing on experience with e.g. the GUM design for GpH (Trinder et al., 1995), GHC-SMP supports both the GpH and Par-Monad models of parallelism. It uses a physically-shared heap with a common address space. Synchronisation is achieved using a “black-holing” mechanism. Memory management uses a shared-memory implementation of GHC’s standard generational garbage collector, with independent nursery areas, but shared old generations. Garbage collection is currently stop-and-copy: all threads must be suspended during garbage collection. While there has been some work on independent garbage collection (Marlow & Peyton Jones, 2011), the complexity of supporting this efficiently and effectively on a physically shared-memory architecture, without adopting a shared-nothing implementation, means that this has not yet been integrated into the production version of GHC.

5.2 Implementations of GpH and Eden

The first implementation of GpH was designed to work on the novel GRIP multiprocessor (Peyton Jones et al., 1987; Hammond & Peyton Jones, 1990; Hammond & Peyton Jones, 1992). This was first adapted to yield a message-passing implementation of virtual-shared memory using PVM, GRAPH for PVM (Loidl & Hammond, 1994) and subsequently to both physically shared-memory and shared-nothing parallel machines in the form of GUM (GRAPH on a Unified Memory Model) (Trinder et al., 1995). GUM uses a two-level message-passing communication system, where high-level messages for synchronisation, workload distribution etc. are built on a basic communication system. Low-level communication can be realised via either via shared-memory or via an explicit
message-passing communication library such as PVM or MPI. This gives significant flex-
ibility. GUM has been extended using a variety of workload management systems (Loidl
& Hammond, 1996; Loidl, 2001; Al Zain et al., 2008; Du Bois et al., 2002; Aljabri et al.,
2012). Using these mechanisms, it has been deployed on a variety of systems ranging from
small-scale multicores to hierarchical clusters of up to 90,000 cores (Maier et al., 2014a).

The DREAM runtime system (Breitinger et al., 1998; Klusik et al., 1999) was the first
working implementation of Eden. It was originally implemented by adapting the GUM
implementation to comply with the the DREAM abstract machine model (Breitinger et al.,
1997). It has subsequently seen several complete rewrites and major revisions. Some parts
of the implementation which were originally implemented directly in the runtime system
have now been lifted to the Haskell level, relying on simpler and more modular runtime
system support (Berthold et al., 2003). This refinement ultimately led to factoring out
the EdI language: all Eden language constructs are now implemented in Haskell, using
the simpler primitive operations that implement EdI directly (Berthold & Loogen, 2007;
Berthold, 2008).

5.3 Distributed Systems: Cloud Haskell and HdpH

A contrasting approach to that described here is taken by Cloud Haskell (Epstein et al.,
2011) and HdpH (Maier et al., 2014b). Rather than providing flexible and general runtime-
system support, which can be used to build appropriate high-level parallel programming
constructs for different language designs, Cloud Haskell and HdpH both use a fully explicit
approach to parallelism on distributed memory platforms, relying on the programmer to
build any necessary higher-level abstractions. Cloud Haskell provides a minimal abstrac-
tion level, while HdpH also includes ideas of work stealing and virtual sharing that were
previously developed for GpH. As with Eden or GpH, higher-level parallel abstractions
can be built on these primitives, in the form of structured parallel approaches such as
algorithmic skeletons (Cole, 1989) or evaluation strategies (Trinder et al., 1998). In our
opinion, this is likely to be the best way for functional applications programmers to use
libraries such as Cloud Haskell or HdpH, since many problematic details can then be
encapsulated into higher-level abstractions. This is also consistent with the functional
programming philosophy of good abstraction and encapsulation.

5.4 Work distribution for Large-Scale Hierarchical PAEAN Systems

The basic PAEAN work distribution model (Section 3.5) has been extended to provide
better support for large-scale, hierarchical architectures. Grid-GUM2 (Al Zain et al.,
2008) systematically attaches load information to FISH messages in order to refine the search for
available work. Unlike the standard PAEAN mechanism, it delivers a variable number of
work items in order to amortise the cost of transferring work and data between sub-clusters
in a Grid infrastructure. Grid-GUM2 also supports migration of live threads (Du Bois et al.,
2002): threads are serialised in their current execution state, and can be migrated across
processors in order to improve load balance. Migration should, however, be used carefully
since it is expensive and increases heap fragmentation. (Aljabri et al., 2013; Aljabri, 2015)
describe GUM-SMP, which combines our shared-nothing work-stealing mechanisms with
those in the standard GHC-SMP. This gives additional freedom in configuring PAEAN systems for large-scale clusters of multicore machines.

### 5.5 Scheduling Support in Haskell

Both the Par-monad (Marlow et al., 2011) and Meta-par (Foltzer et al., 2012) have adopted the idea of defining parallel scheduling policies in Haskell. However, these approaches do not consider more general runtime-system support and system architecture, as we have done in this paper. Moreover, unlike our approach, the Par-monad is restricted to shared memory parallelism. Two separate lightweight implementations of concurrent Haskell have also been produced that lift scheduling and other concurrency features to the Haskell level (Li et al., 2007; Sivaramakrishnan et al., 2013). This complements our work, that instead focuses on the introduction and control of parallelism. It would, however, be both straightforward and useful to incorporate similar concurrency features into PAEAN, and we intend to investigate this in future work. Similar support for scheduling operating system level threads and processes has been provided in the prototype House operating system (Hallgren et al., 2005), which was written entirely in Haskell.

### 5.6 The ARToP micro-kernel

ARToP\(^\text{11}\) is a highly modular micro-kernel design for parallel Haskell (Berthold et al., 2008) that is closely related to PAEAN. The kernel level implements simple and generic routines for asynchronous communication, basic execution management, and system information. System Modules written in Haskell restrict and combine these kernel routines to provide higher-level coordination constructs, narrowing the generic runtime support to a particular programming model. Coordination constructs can be used at both a Library and an Application level either to implement algorithms directly or to implement parallel algorithmic patterns and skeletons (Cole, 1989). System modularity is achieved similarly to PAEAN components: driven by a workload distribution logic, tasks are distributed to processors and controlled by a scheduling component; scheduling relies on a communication subsystem; and a monitoring component informs dynamic and adaptive scheduling decisions. In (Berthold, 2008; Berthold et al., 2008), we presented a Haskell scheduling framework in the spirit of the ARToP design. Complex scheduling policies and heuristics can be defined in Haskell, based on a system of type classes for parallel jobs and scheduler states. The scheduling framework can express complex mechanisms such as the FISH mechanism discussed in Section 3 and Grid-GUM2.

### 5.7 Haskell-level serialisation for PAEAN

The PAEAN serialisation and de-serialisation routines allow data to be exchanged between different heaps in a running system. In doing so, however, they also provide type- and evaluation-agnostic data persistence for Haskell. This has several important uses (Berthold, 2008),
data Serialized a = Serialized { packetData :: ByteArray# }

serialize :: a → IO (Serialized a)
deserialize :: Serialized a → IO a

−− Packing and unpacking exceptions:
(data PackException
  = P_BLACKHOLE | P_NOBUFFER | P_CANNOTPACK | ... | P_GARBLED  -- from RTS
  | P_ParseError  | P_BinaryMismatch | P_TypeMismatch  -- from Haskell

−− Serialized instances:
instance Typeable a ⇒ Binary (Serialized a) ...
instance Typeable a ⇒ Show (Serialized a) ...
instance Typeable a ⇒ Read (Serialized a) ...  -- read ◦ show == id

Fig. 22. Proposed Haskell types for serialised data

2011), including supporting persistence for memoised functions and checkpointing. Figure 22 shows how Haskell-level serialisation can be realised, supported by PAEAN primitive operations. Serialisation is orthogonal to evaluation: any Haskell program graph can be serialised. It is also orthogonal to Haskell types: as shown, type-safety can be re-established by phantom types. When Serialized data is externalised (written to files or sent over the network), phantom type information is lost and must be recovered dynamically, using Typeable. This is achieved using the representations provided by the Binary, Show and Read instances. To make our work usable in other settings, e.g. networked systems, we have factored out this serialisation functionality, so that it can be installed as a Haskell library12 using cabal. This library is a good example of the targeted bespoke runtime-system support for parallelism that we advocate for PAEAN: well-tailored operations enable a simple and straightforward API and avoid the pervasive complications of library-only solutions to Haskell data communication, as typified by e.g. Cloud Haskell or HdpH.

5.8 Kali Scheme

As with A RTCOP, Kali Scheme (Cejtin et al., 1995) aimed to provide user-level scheduling operations, including load balancing and thread migration. Like PAEAN, Kali Scheme is a shared-nothing implementation, but for Scheme rather than Haskell. Kali Scheme operates at a similar level to EdI, HdpH or Cloud Haskell, using explicit message-passing communication. Unlike the PAEAN approach, but like Cloud Haskell and HdpH, it provides explicit primitives to create and manage independent heap spaces, including serialisation and deserialisation for transmitting closures. Kali Scheme assigns unique identifiers (UIds) to all heap objects, and uses a mechanism similar to our FetchMe to fetch all data referenced by UId in a received message before taking any action on it. The PAEAN system, in contrast, assigns global addresses only to thunks that were actually communicated before being evaluated, and lazily fetches other references only when the data is required. Kali Scheme

12 Available at http://github.com/jberthold/packman/ and on hackage.
also provides a proxy mechanism, which allows remote values to be accessed using Kali’s strong code mobility capabilities. As with GUM and PAEAN, Kali Scheme supports a two-level garbage collection scheme, where local collections can proceed independently. Global collection uses a combination of lightweight distributed reference counting with a separate mechanism for global cycle detection. A similar mechanism to our ACK/NACK messages is used to ensure that messages are not lost in transit, and that in-flight data is not garbage collected prematurely. At the time of writing, the Kali Scheme system seems to be no longer maintained. However, there are calls to revive the project.

5.9 Manticore

Finally, while it does not aim to provide a reusable library as with PAEAN, the Manticore (Fluet et al., 2007; Fluet et al., 2010) implementation for parallel ML provides similar implicitly-threaded parallelism and follows the same spirit of semi-explicit adaptivity as our earlier implementations of GpH. Futures and data parallel constructs are used to manage local parallelism. Explicit synchronisation and coordination is used at a larger scale (Reppy et al., 2009). At the runtime level, Manticore provides a notion of virtual processor similar to a HEC, which can be similarly mapped to physical cores by the operating system. A process, comprising multiple lightweight threads, is mapped to the required number of virtual processors using a provision operation. There is a simple built-in mechanism for basic load-balancing, extended using a low-latency work-stealing mechanism (Acar et al., 2013). Manticore also provides a low-level internal scheduling interface (Acar et al., 2012), which provides preemption and interrupt via a signal-based approach. Processes are scheduled in a round-robin fashion, and schedulers may be nested, so that, for example, a data parallel scheduler may be used within a process-level scheduler.

6 Conclusions

This paper has introduced PAEAN, a runtime-system framework that supports efficient and scalable execution of parallel Haskell programs on a variety of (possibly NUMA) parallel architectures, and shown how it can be used to implement three distinct parallel Haskell dialects: GpH, Eden and EdI. PAEAN is based on the widely-used and heavily-tuned GHC runtime system. It offers a flexible and general API for workload distribution, virtual shared memory, communication and serialisation that can be used to implement a variety of further Haskell dialects, on a “pick-and-mix” basis.

The central design decision in PAEAN is its shared-nothing heap, which greatly enhances the scalability of the system. PAEAN abstracts over low-level communication libraries, linking to standards such as PVM and MPI. We believe that with hardware scalability limitations imposed by upcoming many-core architectures, such a design is the only way forward for modern, massively-parallel systems. These concerns are already visible in modern NUMA hardware architectures, and threaten to disrupt current programmer models of memory behaviour. It is no longer tenable to assume that access to an array is uniform cost, for example.

This provides an opportunity for more structured and flexible approaches to data as promoted by Haskell and other functional languages, that will better fit limited sharing
or shared-nothing settings. In order to achieve this potential, it is important to provide
good abstractions over this rapidly increasing complexity at the hardware memory level.
We have built a programmer-level virtual shared heap abstraction on top of a set of low-
level message passing primitives. PAEAN also supports physically-shared memory, but as
a special case of our shared-nothing design. Our performance results show that we can
obtain good, scalable and easy-to-use parallelism for multiple Haskell dialects beyond the
usual shared-memory limitations of standard parallel Haskell implementations.

6.1 Limitations and further work

While we have considered reasonably sized clusters in this paper (of up to 64 cores), in
order to fully study scalability, we will need to consider the use of PAEAN implementations
on larger machines such as EPCC’s 90,000 core HECToR. Our results with a challenging
symbolic computation application on up to 1024 cores of HECToR show good scalabil-
ity for a program with complex data dependencies (Maier et al., 2014a). However, we
anticipate that larger, hierarchical systems will highlight the need for improved control
of locality, hierarchical workload management, and other mechanisms of work and data
distribution that we have discussed above.

One other obvious limitation is that PAEAN does not currently provide support for
fault tolerance. This is not a major issue in a typical parallel computing setting (whether
shared-something or shared-nothing), since such systems can generally be assumed either
to be reliable or to provide systems-level recovery mechanisms. Likewise, cloud and other
managed distributed servers typically provide good system-level mechanisms for check-
pointing and recovery, which should deal with system-level failures. A language-level
mechanism is more useful for dealing with programmer-level failures, where the cause
of any error can be predicted and corrected by the programmer. In such a situation, the
direction taken by Cloud Haskell and HdpH, using Erlang-style supervisor processes to
monitor the health of large distributed systems (Stewart et al., 2012), would seem to offer
a relative simple and effective mechanism that can easily be added to PAEAN. We intend
to investigate this in due course.

Parallel programming is increasingly mainstream, and functional languages like Haskell
are, in principle, an excellent fit to modern massively parallel architectures. However,
the chief advantages of functional programming, namely composability and referential
transparency, can easily be lost, for example, by using explicit communication primiti-
tives to link shared-nothing machines. Advanced runtime support such as we provide in
PAEAN aims to bridge and structure necessary performance tweaks in a carefully designed
and modular runtime system. In our opinion, such an approach is crucially important to
achieving large-scale parallel programming without losing key semantic advantages of
the functional approach. Ultimately, we intend to develop a revised PAEAN system that
captures the spirit of our ArTCoP design: with the necessary runtime system internals
largely programmed in Haskell, while still achieving good scalable performance through
flexible runtime-system support. We anticipate that this will allow us to address both near-
term and more distant parallel hardware developments, while providing support for high-
level parallelism abstractions in Haskell.
**Acknowledgements**

This work has been partially supported by the European Union grants IST-248828 ADVANCE: Asynchronous and Dynamic Virtualisation through performance ANalysis to support Concurrency Engineering, IST-288570 ParaPhrase: Parallel Patterns for Adaptive Heterogeneous Multicore Systems, and IST-644235 RePhrase: Refactoring Parallel Heterogeneous Systems, a Software Engineering Approach; by EU COST Action IC1202: Timing Analysis On Code-Level (TACLE); and by the UK’s Engineering and Physical Sciences Research Council grant EP/G055181/1 HPC-GAP: High Performance Computational Algebra. Jost Berthold was partially supported by the Danish Council for Strategic Research DSF under contract number 10-092299 (HIPERFIT). We would like to thank the Scottish SICSA initiative for funding a visit of Jost Berthold to Heriot-Watt and St Andrews Universities, and for supporting numerous events promoting research on (parallel) programming languages and implementations. We would also like to thank Malak Aljabri for her contributions in terms of performance measurements for the GUM-SMP system, as well as all the PhD students and others who have contributed to extending the GUM and Eden systems in many interesting directions.

**Systems availability**

Source code for the latest version of the GpH runtime system, including GUM and GUM-SMP, is available at [http://www.macs.hw.ac.uk/~hwloidl/hackspace/GUMSMP](http://www.macs.hw.ac.uk/~hwloidl/hackspace/GUMSMP). This version covers both GpH and Eden. The full Eden source code, which implements a subset of the PAEAN design, plus pre-compiled binary releases for selected GHC versions, are available for download at [http://www.mathematik.uni-marburg.de/~eden/](http://www.mathematik.uni-marburg.de/~eden/). Binaries are provided for Windows, Linux, and Mac OS versions. All Haskell modules are available on hackage. RTS source code is available at [http://github.com/jberthold/ghc](http://github.com/jberthold/ghc).

**References**


J. Berthold, H. W. Loidl, K. Hammond


PAEAN: Portable, Scalable Runtime Support for Parallel Haskells


