

# Asynchronous Runtime with Distributed Manager for Task-based Programming Models

Jaume Bosch, Carlos Álvarez, Daniel Jiménez-González, Xavier Martorell, Eduard Ayguadé

---

## Abstract

Parallel task-based programming models, like OpenMP, allow application developers to easily create a parallel version of their sequential codes. The standard OpenMP 4.0 introduced the possibility of describing a set of data dependences per task that the runtime uses to order the tasks execution. This order is calculated using shared graphs, which are updated by all threads in exclusive access using synchronization mechanisms (locks) to ensure the dependence management correctness. The contention in the access to these structures becomes critical in many-core systems because several threads may be wasting computation resources waiting their turn.

This paper proposes an asynchronous management of the runtime structures, like task dependence graphs, suitable for task-based programming model runtimes. In such organization, the threads request actions to the runtime instead of doing them directly. The requests are then handled by a distributed runtime manager (DDAST) which does not require dedicated resources. Instead, the manager uses the idle threads to modify the runtime structures. The paper also presents an implementation, analysis and performance evaluation of such runtime organization. The performance results show that the proposed asynchronous organization outperforms the speedup obtained by the original runtime for different benchmarks and different many-core architectures.

*Keywords:* OmpSs, OpenMP, Task-Based, Task-Graph, Dependence Manager, Runtime

---

### LICENCE

© 2020 Elsevier. This manuscript version is made available under the CC-BY-NC-ND 4.0 license  
<http://creativecommons.org/licenses/by-nc-nd/4.0/>

DOI: 10.1016/j.parco.2020.102664.

## 1. Introduction

The multicore processors popularization started due to the end of Dennard scaling law which states that the power density of an integrated circuit can stay constant meanwhile the transistors get smaller. Until 2006, Dennard's law and Moore's law have guided processor manufacturers to periodically reduce the transistors length and increase the clock frequency which also increases the processors performance. However, the leakage current grows much faster at small transistor sizes; therefore the clock frequency cannot increase without impacting the overall power consumption. Since the transistor still reduces its size periodically as Moore's law states, processor manufacturers started to introduce multiple cores in their processors to keep the processors performance increase.

As multicore processors have become popular, parallel programming has become a need to take advantage of these processors. Instead of dealing with complex applications programmed for one specific processor architecture, parallel programming

models decouple applications from hardware. Their goal is to allow programmers to indicate the potential parallelism in the applications source code without directly managing it. There are several examples like MapReduce [1], OpenMP [2], OpenCL [3], StarSs [4], etc. The exposed parallelism is then managed by a runtime library that coordinates the application execution transparently to the application programmer. Similarly, there are parallelized libraries that can be used from sequential applications that implement several parallel skeletons for commonly used operations. Some examples of these libraries are: Spark [5], OpenBLAS [6], Intel MKL, etc.

The task oriented paradigm is one powerful way to define potential parallelism in one application. Programmers only have to annotate code regions called tasks that can run in parallel. Additionally, developers can provide additional task information like data requirements. This information defines the task execution order enforced by the runtime libraries at execution time. The OpenMP standard introduced task dependences in the 4.0 version greatly influenced by the OmpSs programming model which extends the standard syntax with additional features.

The runtimes of these models are responsible for guaranteeing the task execution order correctness defined by the task data requirements. Therefore, the runtime updates a task graph when a task is created and when a task finalizes its execution. Usually, these modifications require to read and write the information in the task graph atomically to ensure the order correctness.

In a processor with a lot of cores, the probability of collisions between threads trying to access the task dependence graph increases. Each collision implies that a thread is wasting its computation time waiting for another one modifications.

---

*Email address:* [jbosch@bsc.es](mailto:jbosch@bsc.es) (Jaume Bosch)

This problem that has currently started arising is expected to be an important bottleneck as the number of cores in the future processors is expected to keep growing [7]. Thereby, the access contention on some runtime structures will kill the application performance if runtimes do not redesign their internals to tackle the problem.

To improve the current task-based parallel programming runtimes and avoid the contention expected in the many-core processors, we propose an asynchronous runtime organization where the runtime threads do not update the runtime structures directly. Instead, the threads request the needed actions to the runtime and this request will be handled in the future. This asynchronous approach avoids the problem of actively waiting for the exclusive access and allows the threads to return immediately to the application code. Moreover, such structure tries to maximize the utilization of the processor cores to run application code and avoid active waiting on the locks.

The threads requests to the runtime are handled by a runtime manager who updates the runtime structures. Initially, we proposed a centralized implementation based on an extra thread (DAS Thread, DAST) together with a mechanism to avoid the manager saturation [7]. In this work, we present a new distributed implementation of the runtime manager (Distributed DAST, DDAST) based on a mechanism where any thread may become a runtime manager thread. Therefore, the runtime tries to use all the available threads in a smart way to restrict the accesses to the runtime structures.

The proposed asynchronous runtime model provides similar performance to the original runtime when the application has a small number of tasks or when the execution uses a reduced amount of threads. However, when the number of tasks and/or the number of threads is large, the new runtime achieves better performance due to the better thread utilization, data locality and contention reduction. All the changes proposed here are transparent for application developers and are general enough to be used in a wide range of task-based parallel programming models. Even more, the design could be adapted for particular heterogeneous architectures, like big.LITTLE [8], allowing a subset of the worker threads to become manager threads.

The remainder of the paper is organized as follows. Section 2 describes the OmpSs task-based programming model, as an example of task-based programming model whose runtime takes care of the dependences management. Section 3 describes the design and implementation of the new distributed runtime manager (DDAST). Section 4 presents the experimental setup. Section 5 presents the tuning results for the manager internal parameters. Section 6 shows the performance of the new runtime and analyzes its behavior during the executions. Finally, section 7 presents the related work and section 8 concludes.

## 2. Background

The implementation of the asynchronous runtime model has been developed using the OmpSs programming model, which is a forerunner of the standard OpenMP parallel programming model. Therefore, the following sections introduce the OmpSs programming model (section 2.1) and Nanos++ (section 2.2),

which is the runtime library used to run OmpSs applications. The OmpSs programming model is also supported by the source-to-source Mercurium compiler for C, C++ and Fortran. However, the change proposed in this work is only to modify the runtime internals which is transparent for the compiler and the application developers. Besides, the design fits any runtime for task-based parallel programming models because they share the same execution model.

### 2.1. OmpSs programming model basics

The OmpSs programming model is a task-based parallel programming model developed at BSC and composed by a set of directives and library routines. The name OmpSs comes from two other programming model names, OpenMP [2] and StarSs [4]. The goal of the programming model was to extend the OpenMP syntax with some of the StarSs features to provide a productive environment for HPC applications development. Productive means that the applications developed in OmpSs achieve reasonable performance compared to similar solutions for the same architectures and that the development cost is small and does not require huge changes in the applications [9] [10].

On one hand, OmpSs takes from OpenMP the philosophy of providing a way to produce a parallel version of the application adding annotations that do not require modifications in the source code. These annotations allow the compiler to generate a parallel version of the application by replacing the annotations with runtime API calls. This philosophy is intended to simplify the development process leading to a better productivity. On the other hand, OmpSs takes from StarSs the thread-pool model. In contrast to the fork-join model used by OpenMP, StarSs model has an implicit parallelism during all the execution so, programmers do not have to annotate the parallel regions [9] [10].

#### 2.1.1. Task annotation

The main annotation in OmpSs is the task clause which defines a code region that will be asynchronously executed. Tasks can be concurrently executed by any thread when they are ready. The task execution order can be defined by the programmers using the `in(...)`, `out(...)` and `inout(...)` clauses that extend the task annotation. These clauses define the input, output and input+output data dependences for each task (similar to OpenMP `depend` clause), which implicitly define the dependences between tasks. The runtime is responsible for synchronizing task executions to guarantee the data dependences.

In addition to the implicit synchronization created by the data dependences, the programmer can introduce explicit synchronization points using the `taskwait` annotation. This ensures that after the annotation all tasks created before it are executed and all data generated by tasks is available with the latest values.

```

1 void foo (int *a, int *b) {
2   for (int i = 1; i < N; i++) {
3     #pragma omp task in(a[i-1]) inout(a[i]) out(b[i])
4     propagate(&a[i-1], &a[i], &b[i]);
5     #pragma omp task in(b[i-1]) inout(b[i])
6     correct(&b[i-1], &b[i]);
7   }
8   #pragma omp taskwait
9 }

```

Listing 1: OmpSs code annotation example

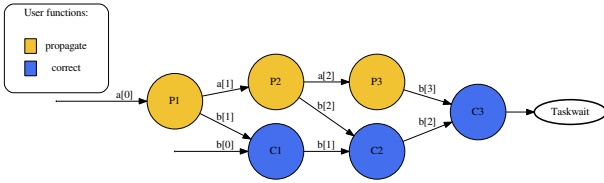


Figure 1: OmpSs task graph for listing 1 (N=3)

An example code [11] of a C function parallelized with OmpSs is shown in listing 1. The function contains two function calls that are annotated with the task directive: *propagate* and *correct*. Thus, the calls to those functions are asynchronously executed when the data dependences defined in the task annotation are satisfied. The resulting task dependence graph (for  $N = 3$ ) is shown in figure 1 where the nodes are tasks and the edges true dependences among them. There is a true dependence between the  $i^{\text{th}}$  *propagate* and the  $i^{\text{th}}$  *correct* tasks due to the  $i^{\text{th}}$  element of  $b$ . There is a true dependence between the  $i^{\text{th}}$  *propagate* and the  $i^{\text{th}} + 1$  *propagate* tasks due to the  $i^{\text{th}}$  element of  $a$ . Finally, there is a true dependence between the  $i^{\text{th}}$  *correct* and the  $i^{\text{th}} + 1$  *correct* tasks due to the  $i^{\text{th}}$  element of  $b$ .

## 2.2. Nanos++ runtime

Nanos++ is a runtime library designed to serve as runtime support in parallel environments. The runtime is developed at the Barcelona Supercomputing Center within the Programming Models group, and its main target is to support the OmpSs programming model. Apart from OmpSs, Nanos++ also supports most of the OpenMP 3.1 features and includes some additional extensions (some of them also introduced in following OpenMP releases) [12].

The runtime provides the required services to support task parallelism based on data dependences. Data parallelism is also supported by means of services mapped on top of its task support. Tasks are implemented as user-level threads when possible. It also provides support for maintaining coherence across different address spaces (such as nodes with FPGAs or cluster nodes) by means of a directory/cache mechanism [12].

The main purpose of Nanos++ is to be used in research of parallel programming environments. The runtime tries to enable easy development of different parts, so researchers have a platform that allows them to try different mechanisms. As such it is designed to be extensible by means of plugins. The

scheduling policy, the throttling policy, the dependence approach, the barrier implementations, slicers and workshar-ing mechanisms, the instrumentation layer and the architectural dependant level are examples of plugins that developers may easily implement using Nanos++ [12].

### 2.2.1. Task life cycle

Task representation inside Nanos++ is made by one Work Descriptor (WD) for each task. Each WD contains all needed information to manage the task during its life cycle. For instance, the WDs store the data dependences of each task. The parent task, which is the task being run when the child task is created, contains the task graph with the relations of its children. This limits the tasks to depend on only sibling tasks, but the global order is guaranteed because parent task dependences must be a super-set of its child tasks dependences. Despite this distributed model, actions in each graph are protected by spinlocks because different sibling tasks can finalize at the same time and/or collision with another sibling task creation.

The different steps in the task life cycle are summarized following:

1. Task creation. At this step, the WD structure is allocated and initialized with the information provided in the annotations related to the task. Moreover, the values of function arguments or local variables are stored in order to execute the code asynchronously.
2. Task submission. At this step, the data dependences of the task are stored in the WD and introduced in the task graph to compute the predecessor WDs. If no predecessors are found, the task can immediately become ready. How the predecessors are computed depends on the dependences plugin which can be changed in each execution.
3. Task becomes ready. At this step, task data dependences have been satisfied or task blocking condition has been fulfilled. Consequently, the task execution can start. How the task will be executed depends on the used scheduling policy that can be changed in each execution.
4. Task becomes blocked. At this step, the task cannot proceed its execution until some condition becomes true. For example, when a task contains a `taskwait` annotation it becomes blocked until its children tasks finish.
5. Task finalization. At this step, the task has finished its execution and the successor WDs may become ready if they only depend on the finalized task. Therefore, the WD can be deleted if it does not have children tasks. Otherwise, the children tasks might further reference the parent WD in their finalization to access the task graph.
6. Task deletion. At this step, the WD can be safely deleted because no more references to it will be done.

Each task state is mainly related with one runtime component, therefore the WDs flow over the different runtime structures during its life cycle. Figure 2 shows a scheme of the tasks flow where each circle represents a task. Each circle color is associated with a task state: yellow for a task being created or

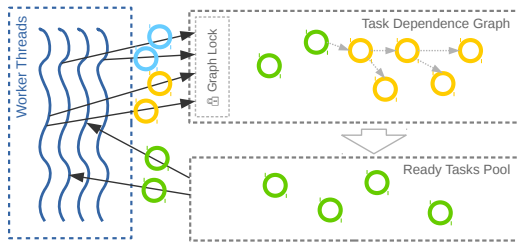


Figure 2: Nanos++ task flow over runtime structures

a submitted task, green for a ready task and blue for a finished task. First, a thread pushes the created tasks into the task dependence graph to determine the task order. Then, other threads “push” the finalized tasks into the task dependence graph to notify the successor tasks. In addition, this action removes the finished task from the graph and adds the tasks that become ready into the ready tasks pool. Both dependence graph operations (push a created or finished task) require acquisition of the graph lock to safely perform the modifications. Finally, the worker threads try to acquire ready tasks from the ready tasks pool to execute them. The management (insertion, deletion, etc.) of tasks into the ready tasks pool may require acquiring some lock, but it is not shown on the figure 2 as the pool implementation depends on the scheduling policy.

### 3. Distributed Runtime Manager

The design of the asynchronous runtime with the distributed manager is based on the idea that any worker thread can become a manager thread and start executing only runtime code. With this approach, all threads can cooperate to execute the pending runtime operations when there are several of them. Correspondingly, all the threads can execute application tasks when the number of pending runtime operations is small. The implementation of the newer design is based on the knowledge acquired in our previous design where almost all the runtime operations that modify the runtime structures were centralized in an additional thread called DAST (DAS Thread)[7]. However, the development of the new distributed runtime manager (DDAST, Distributed DAST) is done over a new and fresh runtime version and it is based on general modules that can be extended to support other runtime functionalities. The three main parts of the new asynchronous runtime design with the DDAST manager can be seen in figure 3 and are explained as follows:

- Section 3.1 explains the messages (requests of runtime operations) that the worker threads send to the DDAST manager, instead of directly doing those operations, and the queues used to transmit/store them.
- Section 3.2 explains the Functionality Dispatcher module introduced in the runtime to mediate between its components and used by the DDAST manager.

- Section 3.3 explains the module that implements the DDAST callback registered into the Functionality Dispatcher.

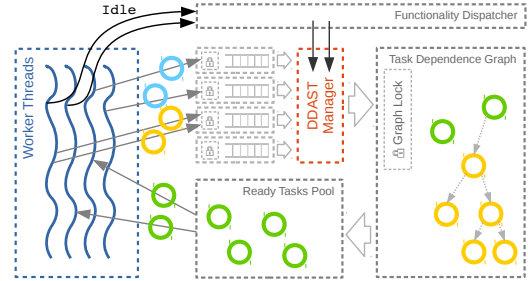


Figure 3: Task flow over runtime structures for the asynchronous runtime with DDAST manager

#### 3.1. Messages and Queues

The messages (request of runtime operations) sent by the worker threads to the runtime manager can be of two types: the *Submit Task Message* and the *Done Task Message*. The first one, the *Submit Task Message* is sent when a worker thread wants to submit a new task into the runtime structures to find out its predecessor tasks. The second one, the *Done Task Message* is sent when a worker thread finishes the execution of a task and wants to notify the successor tasks, scheduling them if they become ready.

There is another synchronization point between the worker threads and the runtime manager that could be managed with a third message. That is during the task deletion step when the task information is deleted. At that point, the worker threads do not know if the task *Done Task Message* has been handled or not, so they cannot safely delete the task information. However, based on our previous experience [7], this synchronization can be handled by means of an additional task state. This allows to safely remove the task information with less overhead than using a third message to the runtime manager.

The two message types are stored in a queuing system until the runtime manager does not handle them. There are two independent queues (one for each message type) replicated for each worker thread where only itself can insert messages, and only the DDAST manager can pop messages, as shown in figure 3. It is important to stress that the queue for the *Submit Task Messages* respects the insertion order to create the right task dependence graph. Otherwise, the sequential execution order of the tasks cannot be guaranteed. Also, only one manager thread can pop and process *Submit Task Messages* from a worker thread queue at the same time. Otherwise, a newer message could enter in the task dependence graph, before an older one, creating a wrongly computed task dependence graph. In contrast, the *Done Task Messages* can be processed by any manager thread concurrently without any restriction, as there is no implicit guaranteed finalization order for the tasks under execution.

### 3.2. Functionality Dispatcher

The Functionality Dispatcher is a new module introduced in the runtime core that mediates between different runtime parts. This module easily allows both using the idle resources to execute any runtime operation and implementing some runtime functionalities without having computational resources exclusively dedicated to them.

Any runtime module can register a callback function in the Functionality Dispatcher during the runtime initialization or the application execution. Those callbacks are listed into the new module, which is also notified by the worker threads when they become idle. In current Nanos++ implementation, the worker threads make a busy waiting loop until they obtain tasks to execute. Therefore, the Functionality Dispatcher tries to take advantage of those idle resources and uses them to execute the different registered callbacks.

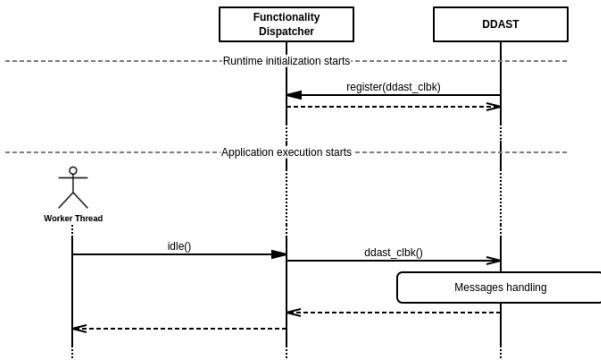


Figure 4: Functionality Dispatcher Sequence Diagram Example

Figure 4 shows the sequence diagram for the implemented Functionality Dispatcher with the DDAST manager. During the runtime initialization, the DDAST module registers a callback function into the Functionality Dispatcher. Therefore, during the application execution, a worker thread that becomes idle notifies the Functionality Dispatcher, which instructs the worker to execute the DDAST callback function that starts handling the messages in the queuing system. This final change of a worker thread to a manager thread through the Functionality Dispatcher is also shown in figure 3.

### 3.3. DDAST Callback

The distributed runtime manager is implemented in a callback function registered in the Functionality Dispatcher. Therefore, the callback is executed when a worker thread becomes idle and the Functionality Dispatcher calls the registered function. That a worker thread becomes idle usually means that the pending messages in the queues must be processed to submit more tasks into the task graph or trigger the scheduling of some new ready tasks.

The behavior of the DDAST callback is parametrized by different constants defined at the beginning of the application execution. The performance impact and the default values for these variables are analyzed in section 5. Here follows a brief list and explanation of these variables::

- `MAX_DDAST_THREADS`. Maximum number of threads allowed to execute the DDAST callback concurrently.
- `MAX_SPINS`. Number of times that the thread will try to get messages without success before leaving the callback.
- `MAX_OPS_THREAD`. Maximum number of messages satisfied from the same worker thread queue before changing to another worker thread queue.
- `MIN_READY_TASKS`. Minimum number of ready tasks available before exiting the callback.

```

1 if (numThreads >= MAX_DDAST_THREADS) return
2 ++numThreads
3 spins = MAX_SPINS
4 do {
5   totalCnt = 0
6   foreach(worker: workers) {
7     if (readyTasks >= MIN_READY_TASKS) break
8     cnt = 0
9     if (worker.queueSubmit.acquire()) {
10      while (cnt < MAX_OPS_THREAD &&
11             (msg = worker.queueSubmit.pop())) {
12        msg.satisfy()
13        ++cnt
14      }
15      worker.queueSubmit.release()
16    }
17    while (cnt < MAX_OPS_THREAD &&
18           (msg = worker.queueOthers.pop())) {
19      msg.satisfy()
20      ++cnt
21    }
22    totalCnt += cnt
23  }
24  spins = totalCnt == 0 ? (spins - 1) : MAX_SPINS
25 } while (spins != 0 && readyTasks < MIN_READY_TASKS)
26 --numThreads
  
```

Listing 2: DDAST callback pseudo-code

Listing 2 shows the pseudo-code of the callback function. First, the number of threads running the DDAST callback is checked and the function returns if the maximum number is reached (listing 2, line 1). After that, the idle thread tries to retrieve messages and satisfy them. This is done until the minimum number of ready tasks is reached or the thread iterates `MAX_SPINS` times without finding any message (listing 2, line 25). The way to retrieve messages is to iterate through all worker threads and try to retrieve and satisfy up to `MAX_OPS_THREAD` messages for the same worker thread (prioritizing the *Submit Task Messages*). Note that not all worker threads are iterated if the number of ready tasks becomes higher than `MIN_READY_TASKS` (listing 2, line 7).

## 4. Experimental Setup

The evaluation of the asynchronous runtime implementation has been done in different architectures and under different benchmarks to show the adaptability and capabilities in different contexts. They are introduced and explained in the following sections: section 4.1 for the different machines/architectures and section 4.2 for the different benchmarks. However, a common criteria for all the architectures and benchmarks has been followed to enhance the evaluation quality, avoid external inferences and facilitate the reproducibility of the results.

1. The machine nodes used during all executions are exclusively reserved for the tests.
2. The applications are compiled with optimization flags and tuning for each architecture.
3. The shown time measurements are the best execution time of 5 repetitions. This way the most stressful conditions, where runtime overheads become crucial, are analyzed. However, the variability between executions is small and the results will be qualitatively the same using the average execution time.
4. The scheduling policy used in all the OmpSs executions is the Distributed Breadth First (DBF). The DBF policy uses a queue of ready tasks for each thread with a stealing mechanism [11].

#### 4.1. Machines/Architectures

The tests have been run in different processor architectures summarized in table 1, the characteristics and software stack of each machine is explained in the next subsections.

Table 1: Machine resources summary

Machine	Num. Cores	Threads x core	CPU Ghz	Mem. GB	Other
KNL	64	4	1.3	96	16GB HBM
ThunderX	48	1	1.8	64	
Power8+	10+10	8	4	256	
Power9	20+20	4	3	512	

##### 4.1.1. Intel Xeon Phi (KNL)

The Knights Landing (KNL) is the second generation of a series of processors manufactured by Intel and characterized by its high parallelism and vectorization capacity [13]. The model used in our evaluation is the *Intel® Xeon® Phi(TM) CPU 7230* configured in Quadrant mode [13]. The executions use up to 64 worker threads, 1 thread per core, as the used machine has the hyper-threading disabled. In the executions with less than 64 cores, the use of the processor resources is maximized. For instance, the cores 0 and 2 are used in a 2-cores execution because cores 0 and 1 share part of the cache.

The version of the Intel® Math Kernel Library (Intel(R) MKL) used by the applications is 2017.0.2. The compiler used to compile the applications and the runtimes is the GNU C Compiler Collection (GCC) version 6.3.0.

##### 4.1.2. ThunderX (ARM)

The ThunderX is a family of processors developed by Cavium based on the 64-bit ARMv8 architecture [14]. The compiler used to natively compile the applications and the runtimes is the GNU C Compiler Collection (GCC) version 5.3.0. The version of the ARM Performance Libraries (ARM PL) used by the applications is 2.0.0.

##### 4.1.3. Power8+

The Power8+ is a revision of the Power8 processors developed by IBM. The nodes used in our evaluation have 2 IBM PowerNV 8335-GTB processors with 10 cores each. The executions only use 1 and 2 threads per core because more than that does not benefit the performance of the evaluated benchmarks. The compiler used to natively compile the applications and the runtimes is the GNU C Compiler Collection (GCC) version 6.3.0.

##### 4.1.4. Power9

The Power9 is the latest version of the architecture developed by IBM [15]. The nodes used in our evaluation have 2 IBM Power9 8335-GTG processors with 20 cores each. The executions only use 1 thread per core because more than that does not benefit the performance of the evaluated benchmarks. The compiler used to natively compile the applications and the runtimes is the GNU C Compiler Collection (GCC) version 8.1.0.

#### 4.2. Benchmarks

The used benchmarks are explained in the following subsections. For each one, its execution arguments are explained and provided with the number of created tasks in each configuration and any other remarks that may be valuable for reproducibility. In all of them, some timing instructions are added after the sequential initialization and after the final global taskwait. The elapsed time between these two points is defined as the execution time in the rest of the paper.

For each benchmark, two different sets of execution parameters are used to create two tasks granularities: coarse grain (CG) tasks and fine grain (FG) tasks. Besides, the benchmark execution parameters are selected considering the following:

- Problem size. Have a big enough problem size to gather significant results.
- Task size (CG). Smallest task size that has enough parallelism to feed all processor cores, delivering almost the best performance hiding the Nanos++ runtime overheads.
- Task size (FG). Solve the same problem with tasks that use half the coarse grain value.

##### 4.2.1. Matrix Multiply

The Matrix Multiply (Matmul) benchmark [16] computes the product of two blocked matrices in parallel. The application takes two main arguments: the matrix dimension (MS) and the block dimension (BS). The task dependences follow a regular pattern with several independent chains that group all tasks working with the same output block. The used values for MS and BS arguments are summarized in table 2.

##### 4.2.2. N-Body

The N-Body benchmark [17] simulates movements of particles under some physic forces, such as gravity. The application takes three arguments: the number of particles (NUM\_PARTICLES), the number of time steps (NUM\_TIMESTEPS)

Table 2: Matmul execution arguments

Machine	MS	Coarse Grain		Fine Grain	
		BS	#Tasks	BS	#Tasks
KNL	8.192	512	4.096	256	32.768
ThunderX	4.096	128	32.768	64	262.144
Power8+/9	8.192	512	4.096	256	32.768

to be simulated and the number of particles per block (BS). Therefore, the particles are spread into blocks with BS particles, which are used as task input/output. The tasks follow a regular chained pattern similar to the Matmul one but this benchmark has nested tasks. This nesting make more critical some of the requests to the DDAST manager because they may block the application parallelism until they are processed. The values for the arguments used in each machine are summarized in table 3 with the number of tasks created in each configuration.

Table 3: N-Body execution arguments

Machine	Num. Particles	Num. Timesteps	Coarse Grain		Fine Grain	
			BS	#Tasks	BS	#Tasks
KNL	16.384	16	128	262.176	64	1.048.608
ThunderX	16.384	16	128	262.176	64	1.048.608
Power8+/9	16.384	16	256	65.568	128	262.176

#### 4.2.3. Sparse LU

The Sparse LU benchmark [16] computes the Lower Upper (LU) decomposition of a sparse matrix in parallel. The application takes two arguments: the matrix dimension (MS) and the block dimension (BS). Therefore, the matrix with MS\*MS elements is divided into sub-matrices with BS\*BS elements. The task dependences follow a much more complex and irregular pattern than the Matmul and N-Body benchmarks. The used values for MS and BS arguments are summarized in table 4.

Table 4: Sparse LU execution arguments

Machine	MS	Coarse Grain		Fine Grain	
		BS	#Tasks	BS	#Tasks
KNL	8.192	128	11.472	64	89.504
ThunderX	8.192	128	11.472	64	89.504
Power8+/9	8.192	128	11.472	64	89.504

## 5. DDAST Tuning

The initial executions with the new runtime structure were intended to find good default values for the callback parameters explained in section 3.3. To this end, some initial values are defined, based on a reasonable approximation to the expected optimal ones, and the same execution is repeated changing only one parameter value. The executions for each parameter are done with two benchmarks that have different task dependence

patterns: Matmul and Sparse LU. Each execution set is doubling the parameter value from 1 up to 128. Also, different amounts of threads and threads per core are considered depending on the architectures: KNL, ThunderX and Power8+. However, the results only consider the two configurations with the largest amount of threads in each architecture as they are the most interesting. Moreover, the larger the number of threads, the bigger the influence of parameter modifications in the execution time.

Table 5: DDAST parameters values

Parameter	Initial Value	Tuned Value
MAX_DDAST_THREADS	$\infty$	$[num\_threads/8]$
MAX_SPINS	20	1
MAX_OPS_THREAD	6	8
MIN_READY_TASKS	4	4

The predefined values for each parameter before (*Initial value*) and after (*Tuned Value*) the tuning are shown in table 5. Using the initial values as default, sections 5.1 to 5.4 present the results obtained when one of the parameters is modified. Finally, section 5.5 briefly explains how we verified the correctness of *Tuned Values*.

### 5.1. Maximum number of DDAST threads

The plots of figure 5, one for each benchmark and architecture, show the speedup over the default parameter value (y-axis) when changing the MAX\_DDAST\_THREADS parameter value (x-axis). The number of threads used in each configuration appears in the legend of the plots.

The results show different behavior depending on the architecture/application combination. The executions with fine grain tasks are more influenced by the parameter value than coarse grain. One one hand, some fine grain results (figure 5b, figure 5c and figure 5d) show that only one manager thread may not be able to handle the incoming messages increasing the execution time and limiting the application performance. In these cases, the results show that the parameter value does not influence the execution time when it goes above 2-4 depending on the number of worker threads. On the other hand, the SparseLU results in Power8+ show that more than one thread in the DDAST callback is needed to handle the messages but restricting the number of threads to 2-4 may provide a better performance (figure 5f). The improvement comes from the better exploitation of the runtime structures' locality as they are mainly accessed by only 2-4 threads.

Considering all results in figure 5, keeping the number of manager threads in the DDAST callback restricted but avoiding creating a bottleneck in the messages processing is the best option. The smaller the number of manager threads, the better data locality in the execution of runtime operations. In addition, the number of manager threads must be large enough to process the peaks of messages without harming the application performance. These peaks are related to the number of worker threads (*num\_threads*), which is known during the runtime initialization, because it shapes the throughput of messages creation.

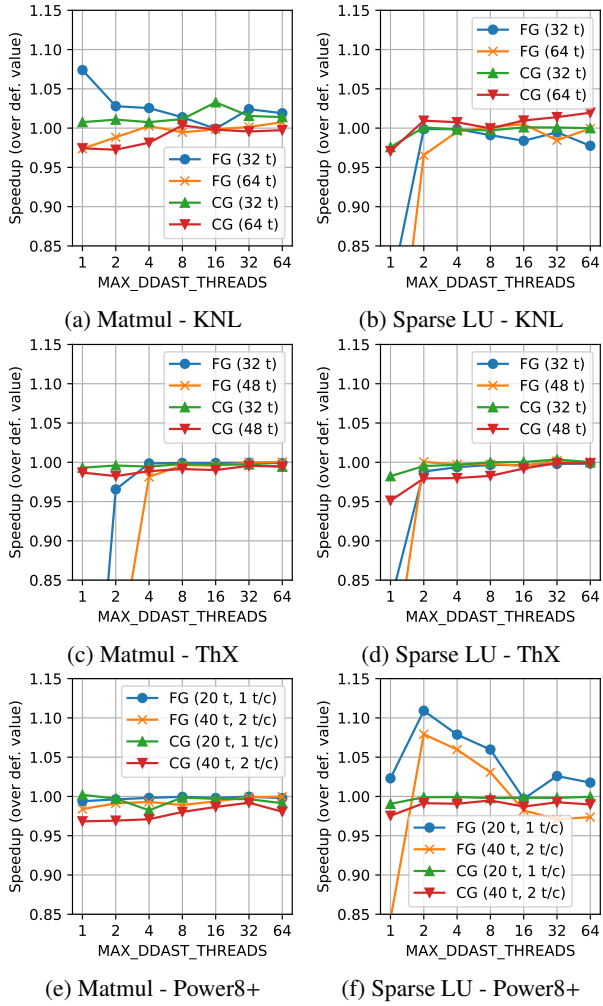


Figure 5: Speedup changing the MAX\_DDAST\_THREADS

Therefore, the new default value for MAX\_DDAST\_THREADS based on the number of available worker threads is defined as:  $\lceil \text{num\_threads} / 8 \rceil$ .

### 5.2. Maximum number of spins

Figure 6 shows the speedup over the default parameter value (y-axes) when changing the MAX\_SPINS parameter value (x-axes). The figure contains the equivalent results of figure 5 but summarized in only one plot as all benchmarks and architectures yield very similar results. Figure 6a and figure 6b contain the same results but using different scales in the y-axes to better observe the behavior.

The results show that the execution time is almost not affected by the value of MAX\_SPINS. Regardless of the values, the speedup is near to 1 in all the cases and the variability of the results is small ( $\pm 0.5\%$ ) and can be associated with the usual variability between executions rather than the performance difference between the different values. On one hand, if no other callback functions are registered, the best approach may be use a large value and retain the threads in the DDAST callback until there are some ready tasks. On the other hand, if there are other

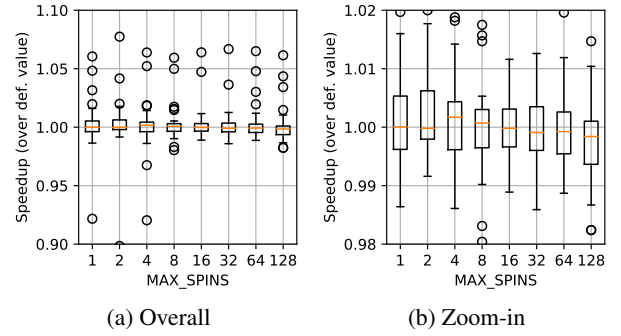


Figure 6: Speedup changing the MAX\_SPINS

registered callback functions, a small parameter value avoids retaining the threads in the DDAST callback until any other break condition is satisfied. Considering the gathered results and a future scenario where the Functionality Dispatcher is used to manage more runtime functionalities, the new default value for the MAX\_SPINS parameter is set to 1.

### 5.3. Maximum operations per thread

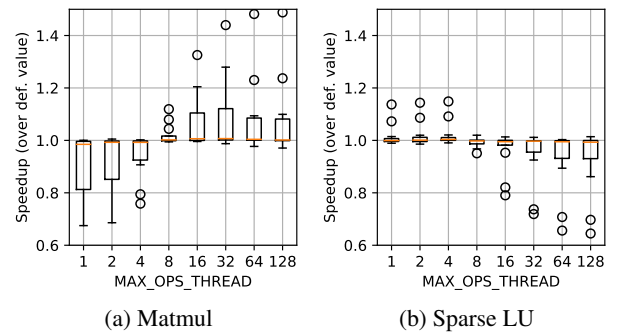


Figure 7: Speedup changing the MAX\_OPS\_THREAD

Figure 7 shows the speedup over the default parameter value (y-axes) when changing the MAX\_OPS\_THREAD parameter value (x-axes). The results are joined by benchmark as each one yields different results, but they behave similarly regardless of the architecture.

The results show that the average speedup for all values in all benchmarks is near to 1 but the variability changes with the parameter value. On one hand, the Matmul (figure 7a) have slowdowns when the MAX\_OPS\_THREAD is below 8 and improvements when the value is larger than so. On the other hand, the Sparse LU benchmark (figure 7b) shows opposite behavior with slowdowns when the MAX\_OPS\_THREAD is above 8.

The opposed behaviors shown are due to the different dependence patterns that each application has. The execution time may increase with a large value of MAX\_OPS\_THREAD if the ready tasks depend on a *Done Task Message*. In this case, a critical runtime operation request may be delayed because the DDAST manager will before process a large number of messages from different worker threads. However, the execution time may decrease with large values of MAX\_OPS\_THREAD if most of the messages trigger the scheduling of a new ready



tasks. In this case, the manager threads may benefit from the data locality when executing the runtime operation request from the same queue.

Considering the results, a reasonable default value for the `MAX_OPS_THREAD` parameter is 8. This value is the one with less observations below 1 and without any observation which significantly decrease the previous value performance. In addition, it still has some positive speedup observations.

#### 5.4. Minimum number of ready tasks

Figure 8 shows the speedup over the default parameter value (y-axis) when changing the `MIN_READY_TASKS` parameter value (x-axis). The results are joined in different plots by benchmark.

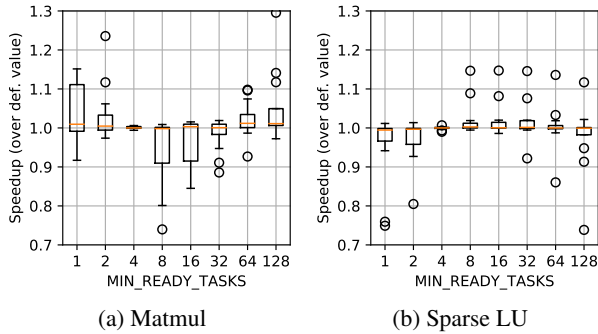


Figure 8: Speedup changing the `MIN_READY_TASKS`

The different task dependence patterns result in a different behavior depending on the benchmark, as figure 8 shows, like for the `MAX_OPS_THREAD` parameter. However, the average speedup in all applications and values is near to 1 but with different variabilities depending on the parameter value. In the Matmul benchmark, the values with positive variation and without slowdowns are the more extreme. The values in the middle pollute the cache to discover some ready tasks but don't benefit from the data locality of loaded information to discover the largest possible number of ready tasks. In contrast, the SparseLU benchmark has negative variations in the extreme values. In this benchmark, discovering one ready task is not as easy as in the Matmul so the cache usually becomes polluted. Moreover, discovering a large number of ready tasks is also complicated, so the manager threads are spinning a lot of time.

Considering all the results, the predefined default value for the `MIN_READY_TASKS` parameter, which is 4, is the best configuration. The alternatives that outperform this value in some cases are really bad in other ones.

#### 5.5. Verification

To confirm the results obtained in the previous evaluation, the analysis has been repeated but using the *Tuned Values* as defaults and also evaluating the N-Body benchmark. These new results are not shown in the paper as they behave similar to presented ones, that confirms the correctness of *Tuned Values* defined after the initial analysis. In addition, the results obtained in this second exploration are used to measure the *DDAST tuned* performance in the next section.

## 6. Performance Comparison

The following sections present the performance comparison of the new runtime model against other task-based parallel programming models. Section 6.1 shows the scalability results for the different runtimes and section 6.2 presents the most relevant executions where the main differences between both runtime approaches can be seen.

### 6.1. Scalability results

The obtained scalability results are shown in figure 9, figure 10 and figure 11 for Matmul, Sparse LU and N-Body respectively. As we have used Matmul and Sparse LU benchmarks for the tuning, the N-Body results can be used as a control of the performance obtained with a new application. The results are shown for the KNL, ThunderX and Power9 architectures and for different runtime versions/configurations:

- *Nanos++*. Baseline OmpSs runtime (version 0.11a).
- *DDAST*. Runtime with the distributed runtime manager implementation (DDAST manager) and using the tuned values for the DDAST parameters, which are summarized in table 5 and are the same for all the runs. This version is implemented on top of Nanos++ runtime (version 0.11a).
- *DDAST tuned*. Same runtime as *DDAST* but with the best values of the DDAST parameters found during the tuning verification for each combination: benchmark, task granularity and architecture.
- *GOMP*. OpenMP implementation for the GNU Compiler Collection. This is a production runtime which performance can be used as a reference of the potential of our approach.

The speedup over the sequential version of each benchmark is shown in all plots of figure 9, figure 10 and figure 11 (y-axis). All of them show strong scalability of Nanos++, DDAST and GOMP runtimes for Matmul, Sparse LU and N-Body benchmarks. Therefore, we can observe the performance evolution when the runtimes must manage more computational resources. The label of each plot describes the architecture and the task granularity (fine grain, FG, or coarse grain, CG) of those results. *DDAST tuned* results are included because they show the potential of our proposal. Also, although it is out of the scope of this work, DDAST manager parameters may be dynamically tuned at runtime to fit each application as shown in [18].

Figure 9a and figure 9b show a significant performance improvement (~40% for fine grain and ~30% for coarse grain) when using *DDAST* in comparison to *Nanos++* in KNL architecture for the Matmul benchmark. The same figures show even a larger improvement for *DDAST tuned* that goes up to ~90%.

The difference is due to better data locality which reduces the cache misses and consequently the execution time of each task, which is ~33% shorter in *DDAST* than in *Nanos++*. The improvement can be explained by the way each approximation accesses the data: *Nanos++* is accessing the runtime data

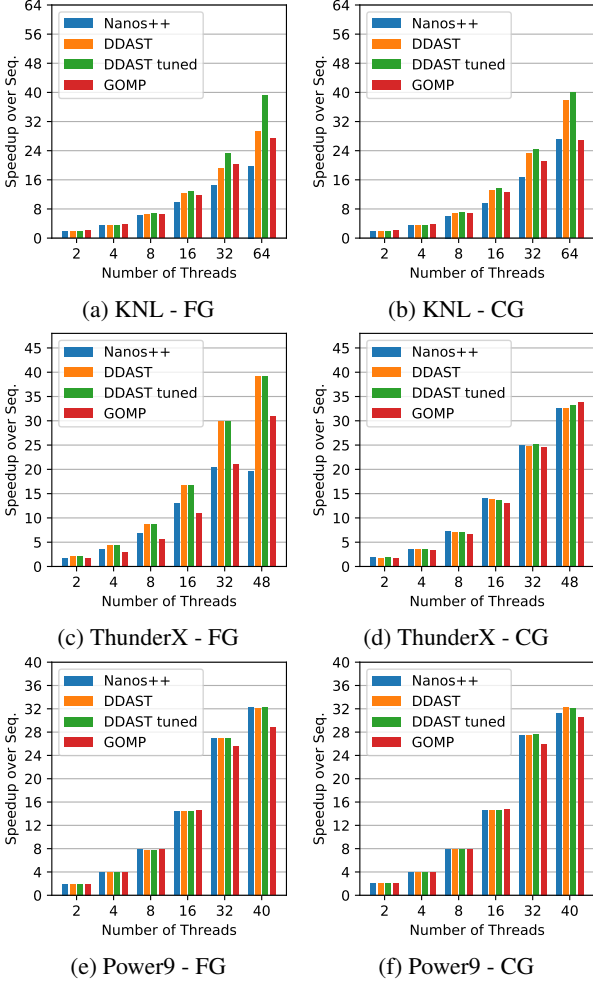


Figure 9: Matmul scalability

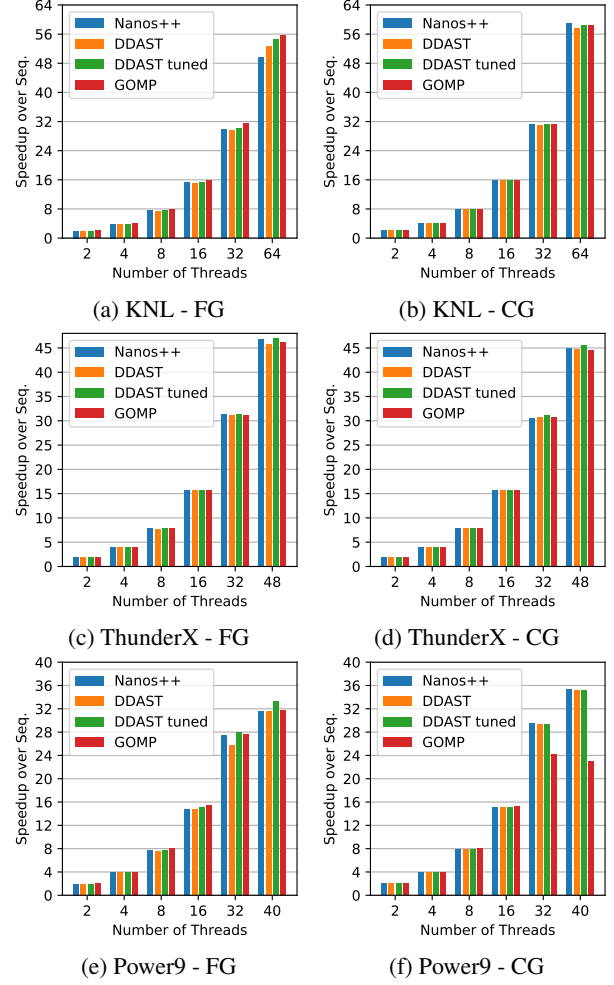


Figure 10: Sparse LU scalability

structures between two task executions, and therefore the runtime is "polluting" the thread caches with its data. In contrast, *DDAST* avoids those accesses by its asynchronous approach. The *GOMP* results behave similar to *Nanos++* with better performance in the fine grain results as the GNU runtime has a smaller footprint than *Nanos++*. The asynchronous approach that drives *DDAST* would also benefit *GOMP* in these situations because the accesses to the shared runtime structures between the execution of tasks will be avoided.

Figure 9c (Matmul, ThunderX, FG) shows improvement when using *DDAST*, similar to the KNL results. However, *DDAST* gets the same performance as *DDAST tuned* in this architecture. This means that default values of *DDAST* manager parameters are the best configuration for such combinations. Figure 9d, figure 9e and figure 9f show similar results between all runtimes because those configurations do not suffer the contention problem that our proposal tackles.

Regardless of the task granularity, all runtimes provide very good scalability in the Sparse LU benchmark (figure 10). The data dependences in this benchmark create an irregular task graph that usually requires processing multiple requests from different worker threads to discover a single ready task. This

creates a challenging situation for the *DDAST* manager where all possible ready tasks depend on a message which is hidden by several other requests in a queue. However, the results show that even with this type of applications *DDAST* is able to achieve a performance similar to *Nanos++* and *GOMP*.

Figure 11a, figure 11c and figure 11e (fine grain results of N-Body in the three architectures) show a performance standstill when moving to the larger amount of worker threads for *Nanos++*. In contrast, *DDAST* maintains or increases the performance compared to the baseline runtime. The difference between both runtimes is the cost of task submission, which is smaller in *DDAST* due to its asynchronous approach. This allows the application to create a huge amount of tasks faster in the new runtime model than in the baseline implementation. In those cases, *GOMP* creates tasks faster than *Nanos++* based runtimes for small amounts of worker threads (up to 16 threads in figure 11a and up to 32 threads in figure 11b). However, *GOMP* suffers great contention from the idle worker threads when tasks are executed faster than created, which happens in figure 11a with 32/64 threads and figure 11b with 64 threads. In ThunderX architecture, *GOMP* does not reach the point where there are several idle worker threads, and consequently, it per-

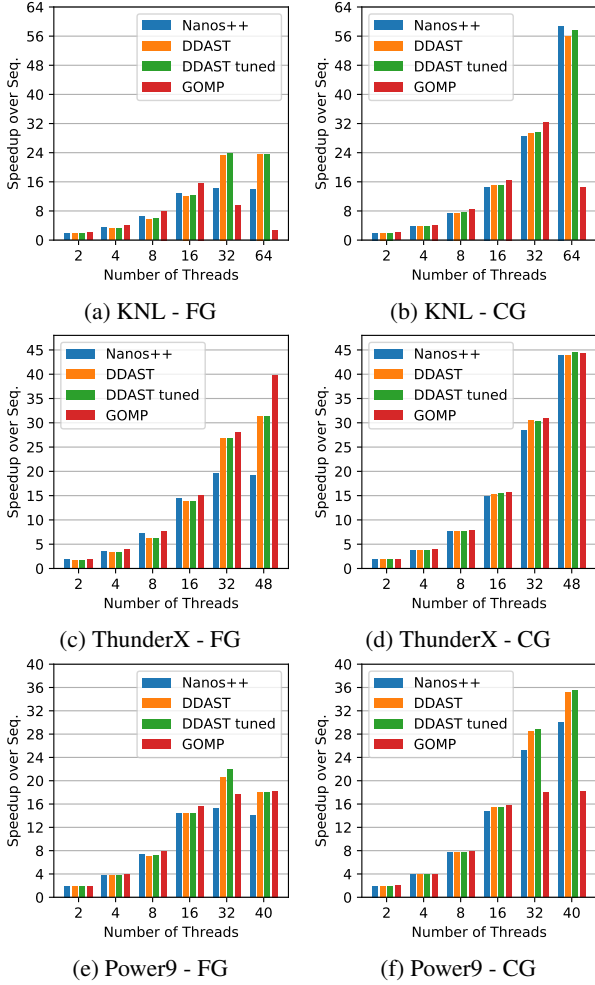


Figure 11: N-Body scalability

forms better than both Nanos++ based runtimes.

The coarse grain results in figure 11b, figure 11d and figure 11f do not have such a large performance gap between runtimes as the number of created tasks is a quarter of the amount in fine grain. Nevertheless, the new runtime organization keeps the performance of the original one.

## 6.2. Execution analysis

The Paraver [19] execution traces in this section show the behavior of *DDAST* and *Nanos++* among different benchmarks executions to see the main differences between both runtimes. All execution traces contain the initial and end timestamps in the x-axes (time). Despite the initial and end timestamps may not match between different runtime executions, any pair of traces that are intended to be compared have the same time duration. These differences are due to the variable startup overheads that may change between executions, thereby the initial time is adjusted to match the point where the execution of the first task starts.

Figure 12 shows the evolution of the number of tasks in the dependence task graph (figure 12a) and the number of ready tasks (figure 12b) along the execution time (x-axis). Both ex-

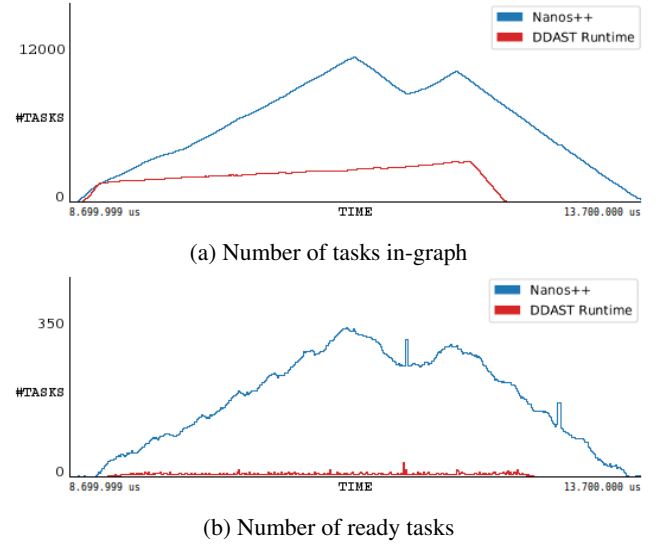


Figure 12: Execution traces of fine grain Matmul on KNL with 64 threads

ecution traces are from the same execution of Matmul benchmark in KNL architecture with fine grain tasks. They have the same duration for the x-axes (5 seconds) but different scales in the y-axes (parameter evolution). On one hand, *Nanos++* has almost a pyramid shaped evolution where a huge amount of tasks are concurrently managed in the task graph (upper line in figure 12a) and ready queues (upper line in figure 12b). In fact, the evolution is not a perfect pyramid due to a trace flush to disk at those points, temporally interrupting the main thread and the task creation. On the other hand, *DDAST* has a roof shaped evolution (bottom lines in figure 12a and figure 12b) where only the minimum amount of tasks needed to discover some parallelism are used, the rest are kept in the manager queues. This difference greatly influences the runtime overheads which are related to the number of elements that should be managed in the runtime structures.

Figure 13 shows different execution traces of N-Body with coarse grain tasks in ThunderX for *Nanos++* and *DDAST*. However, for clarity purposes, only two timesteps are simulated (instead of the value provided in table 3). In figures 13a and 13c, each line in the y-axes represents one of the 48 worker threads. In each line, the different colors represent the thread state: sky-blue for IDLE state and other colors for the different task types execution. Finally, figure 13b shows the number of tasks in graph (y-axes) evolution for both runtimes.

The tasks execution traces show that the threads execute the tasks at the same throughput that are created. During each timestep, the execution of the top level task (two long dark-blue regions in figure 13a) ends almost at the same time as children tasks (small pink regions), which are created by the first one. In this case, the number of ready tasks is near to zero all the time. Therefore, the faster the task creation, the better performance and resource usage. So, the reduction in the execution time is from the faster execution of the top level tasks and not from a shorter execution time of leaf tasks like in the Matmul case.

In *DDAST*, the requests to the runtime manager are quickly

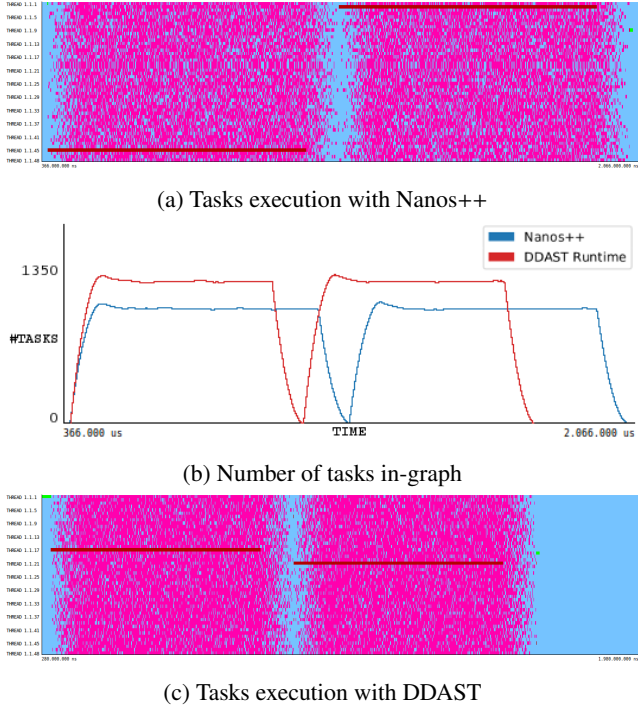


Figure 13: Coarse grain N-Body execution traces on ThunderX with 48 threads

processed by the manager because the worker threads become idle very frequently. This increases the throughput of task submission in comparison the the baseline runtime, as the same amount of tasks is created with less time. Therefore, the number of tasks in the dependence graph increases as can be seen in figure 13b. In addition, the same behavior has been observed in the fine grain executions of the same benchmark explaining the better performance results of *DDAST* against *Nanos++* results.

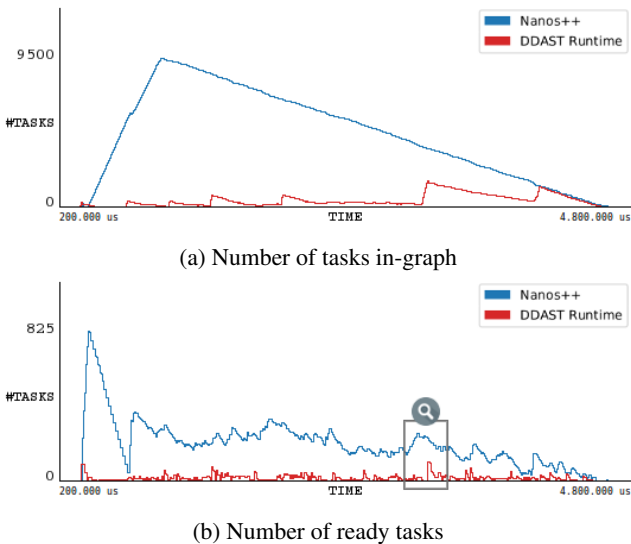


Figure 14: Execution traces of coarse grain Sparse LU on ThunderX with 48 threads

The evolution among all the execution time of Sparse LU in ThunderX for the number of tasks in-graph and ready tasks is

shown in figure 14. Each figure contains a trace for *Nanos++* and *DDAST* runtimes for the entire execution. The evolution of those parameters is equivalent to the observed in the Matmul benchmark. The number of in-graph tasks has a pyramid shaped evolution in *Nanos++* and a plain shape with small peaks in *DDAST*.

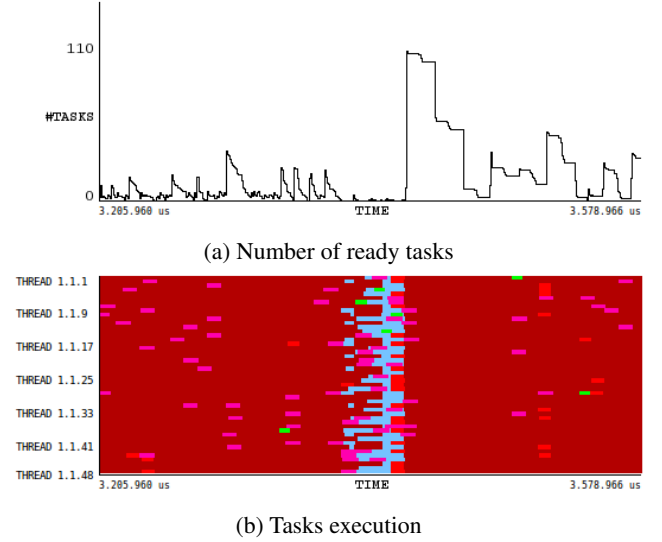


Figure 15: Coarse grain Sparse LU partial execution traces on ThunderX with 48 threads and DDAST

Figure 15 shows two execution traces when almost all threads become idle for the small portion marked in figure 14b. Figure 15b shows the tasks that are executed in each thread and figure 15a shows the number of ready tasks (y-axis), both among the execution time (x-axes). Figure 15a shows that the number of ready tasks becomes nearly zero for a relatively long portion of time.

Therefore, almost all threads are idle as can be seen in figure 15b, so they start to process the pending requests to the runtime manager. In addition, several tasks are added to the task dependence graph but their data dependences are not satisfied so the tasks do not become ready. They do so when the *Task Finalization* requests of the critical tasks are processed, at this point the number of ready tasks suddenly increases from zero to more than 100 as can be seen in figure 15a.

## 7. Related Work

Several works exist about characterization and improvement of parallel programming models. They are over different models working at different levels and with different approaches. OmpSs tools (Mercurium and Nanos++), which are open source and are the ones used to test our model, can execute inter-node and intra-node applications [20] and are under constant development introducing new features. Moreover, several people use this programming model as a base to develop different prototypes or extend its functionality.

Previous works discussed the task scheduling and dependences resolution overheads in data-driven task-based models

like OpenMP and OmpSs. TurboBLYSK [21] is a framework which implements the OpenMP 4.0 with a custom compiler and a highly efficient runtime schedule of tasks with explicit data-dependence annotations. Its objective is also to reduce the dependence management overheads of the runtime. However, TurboBLYSK approach requires extra information in the task dependences definition to allow the runtime to re-use previously resolved dependence patterns and to reduce the overall overhead. In contrast, our proposal only uses the information provided by default to reduce the overall task management overhead, so our optimization is transparent to the programmers.

Other task-based programming models like Intel Threading Building Blocks [22] and Charm++ [23] use an execution model that is more pure data-flow than the OmpSs/OpenMP model which has a hybrid (control/data flow) model [24]. This execution model usually allows to exploit better the parallelism of the applications but requires a specific structure and an application redesign. In the context of the Intel Threading Building Blocks, there is a previous work discussing the cost of the synchronization inside its runtime, but they focus the problem in the work distribution [25] instead of the task graph management that is implicitly done in their execution model. The Charm++ programming model is intended to provide some valuable features for executions in large computation systems like migratability, checkpoint application restarting, process failure tolerance, malleability, etc. They have previous work about optimizing the communications inside their runtime [25], but the dataflow model that they have moves the complexity of task graph management into the application development process like in TBB.

HPX (High Performance ParalleX) [26] and STAPL (Standard Template Adaptive Parallel Library) [27] are general purpose frameworks for parallel and distributed applications of any scale. Both use the same asynchronous philosophy that we have used in the DDAST design to avoid starvation and reduce the runtime overheads. However, they require changes in the application implementation to make use of the library. Also, there are OpenMP implementations over HPX: hpxMP [? ], which has assembly parts only developed for x86 architectures; and OMPX [? ]. The OMPX implementation and our work achieve a similar asynchronous task management. They use the `async` and `future` constructs to achieve the asynchronous task management. In contrast, we implemented the asynchronous task management in the runtime core. That allows us to parametrize and control where the runtime management happens.

The DAGuE framework [28] offers an architecture aware scheduling and management of micro-tasks on distributed many-core heterogeneous architectures. It uses a mixed control-data flow similar to OmpSs programming model and it also optimizes the tasks management minimizing the amount of tasks created in the system during the execution. The main difference with our approach is that DAGuE auto-parallelizes the application based on a static analysis of the application at compile time, instead of dynamically building a task dependence graph at runtime. This difference reduces the amount of applications that can use DAGuE as not all of them have a static dependence pattern that can be formally expressed at compile time.

Other works, which also try to accelerate current runtimes,

propose moving part of the runtime into a specific hardware of FPGAs. Some examples are Nexus# [29] and Picos [30]. They present different hardware designs that can manage tasks dependences of task-based programming models. Besides these, there is active research in new computer architectures able to manage efficiently tasks in StarSs family. For example, some research aims to look for a new Runtime-Aware Architecture to overcome current multi-core restrictions like power, programmability and resilience [31]. The main difference between those works and the one proposed in this project is the way to improve the existing system. They proposed new hardware to work in harmony with the software in order to improve the performance. In contrast, this project improves the existing parallel programming model runtimes with software ideas that do not require additional hardware.

## 8. Conclusions

The multicore processors have become popular and are present in almost any electronic device nowadays. Task-based parallel programming models, like OmpSs, facilitate programmers to use such processor architectures by simply annotating the sequential applications source code. However, the runtime libraries that support such models present a contention problem when the number of threads grows to some tens. As current many-core processors, the future processors are expected to have several cores; thereby the runtimes may become a performance bottleneck.

This paper presents a design of an asynchronous runtime structure based on a distributed runtime manager. It is based on requests from the worker threads to the runtime manager, which modifies the runtime structures handling the requests. The distributed manager design is based on the idea that any thread should be allowed to become a manager thread if needed. Despite the fact that we implemented the design over one runtime, it can be applied to any task based runtime system like any of the OpenMP compliant runtimes. This is because the task life cycle should be really close in all of them and we just change the transition and management of those states. Even more, the design could be adapted for particular heterogeneous architectures, like big.LITTLE, allowing a subset of the worker threads to become manager threads.

As a proof-of-concept, the paper also presents an implementation of the asynchronous runtime design based on the Nanos++ runtime library. In such implementation, the runtime core is extended with new generic modules that provide the possibility of asynchronously executing other runtime services by idle worker threads. These new modules could be used for other runtime actions like sending tasks to accelerators or processing the finished ones. Finally, the performance results show that the current implementation outperforms the baseline runtime for large amounts of threads and that it has the same performance for small amounts of threads or when the tackled problem does not arise. As a future work, the runtime manager will dynamically tune its parameters to fit the application requirements and increase even more the performance of our proposal.

## Acknowledgements

This work is partially supported by the European Union H2020 Research and Innovation Action (projects 801051, 754337 and 780681), by the Spanish Government (projects SEV-2015-0493 and TIN2015-65316-P, grant BES-2016-078046), and by the Generalitat de Catalunya (contracts 2017-SGR-1414 and 2017-SGR-1328).

## References

- [1] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, *Commun. ACM* 51 (1) (2008) 107–113. doi:10.1145/1327452.1327492.
- [2] L. Dagum, R. Menon, OpenMP: an Industry Standard API for Shared-Memory Programming, *Computational Science Engineering*, IEEE 5 (1) (1998) 46–55. doi:10.1109/99.660313.
- [3] J. E. Stone, D. Gohara, G. Shi, OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems, *Computing in Science & Engineering* 12 (3) (2010) 66–73.
- [4] J. Planas, R. M. Badia, E. Ayguadé, J. Labarta, Hierarchical Task-Based Programming With StarSs, *Int. J. High Perform. Comput. Appl.* 23 (3) (2009) 284–299. doi:10.1177/1094342009106195.
- [5] A. Spark, Apache Spark<sup>®</sup> - Unified Analytics Engine for Big Data, <https://spark.apache.org/> (Sep 2018).
- [6] Z. Xianyi, W. Qian, W. Saar, OpenBLAS: An optimized BLAS library, <https://www.openblas.net/> (Sep 2018).
- [7] J. Bosch, X. Tan, C. Álvarez, D. Jiménez-González, X. Martorell, E. Ayguadé, Characterizing and Improving the Performance of Many-Core Task-Based Parallel Programming Runtimes, in: *Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2017 IEEE International, IEEE, 2017, pp. 1285–1292.
- [8] P. Greenhalgh, big.LITTLE Processing with ARM Cortex-A15 & Cortex-A7, ARM White paper 17.
- [9] A. Duran, E. Ayguadé, R. M. Badia, J. Labarta, L. Martinell, X. Martorell, J. Planas, OmpSs: A PROPOSAL FOR PROGRAMMING HETEROGENEOUS MULTI-CORE ARCHITECTURES, *Parallel Processing Letters* 21 (02) (2011) 173–193. doi:10.1142/S0129626411000151.
- [10] J. Bueno Hedro, Run-time support for multi-level disjoint memory address spaces.
- [11] Programming Models Group BSC, OmpSs User Guide, <https://pm.bsc.es/ompss-docs/user-guide/> (May 2017).
- [12] Programming Models Group BSC, Nanos++ Runtime Library, <https://github.com/bsc-pm/nanox> (May 2017).
- [13] A. Sodani, Knights landing (KNL): 2nd Generation Intel® Xeon Phi processor, in: *2015 IEEE Hot Chips 27 Symposium (HCS)*, 2015, pp. 1–24. doi:10.1109/HOTCHIPS.2015.7477467.
- [14] L. Gwennap, ThunderX Rattles Server Market, *Microprocessor Report* 29 (6) (2014) 1–4.
- [15] S. K. Sadasivam, B. W. Thompto, R. Kalla, W. J. Starke, IBM Power9 Processor Architecture, *IEEE Micro* 37 (2) (2017) 40–51. doi:10.1109/MM.2017.40.
- [16] Computer Science Department BSC, BSC Application Repository, <https://pm.bsc.es/projects/bar/wiki/Applications> (April 2017).
- [17] Programming Models Group BSC, BAR-Benchmarks [at] INTER-TWinE, <https://pm.bsc.es/gitlab/ompss/bar-benchmarks/> (April 2017).
- [18] X. Li, M. J. Garzarán, D. Padua, A dynamically tuned sorting library, in: *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization*, CGO '04, IEEE Computer Society, Washington, DC, USA, 2004, pp. 111–. URL <http://dl.acm.org/citation.cfm?id=977395.977663>
- [19] V. Pillet, J. Labarta, T. Cortes, S. Girona, Paraver: A tool to visualize and analyze parallel code, in: *Proceedings of WoTUG-18: transputer and occam developments*, Vol. 44, IOS Press, 1995, pp. 17–31.
- [20] J. Bueno, L. Martinell, A. Duran, M. Farreras, X. Martorell, R. M. Badia, E. Ayguadé, J. Labarta, Productive cluster programming with ompss, in: *Euro-Par 2011 Parallel Processing*, Springer, 2011, pp. 555–566.
- [21] A. Podobas, M. Brorsson, V. Vlassov, TurboBLYSK: scheduling for improved data-driven task performance with fast dependency resolution, in: *Using and Improving OpenMP for Devices, Tasks, and More*, Springer, 2014, pp. 45–57.
- [22] C. Pheatt, Intel® Threading Building Blocks, *J. Comput. Sci. Coll.* 23 (4) (2008) 298–298. URL <http://dl.acm.org/citation.cfm?id=1352079.1352134>
- [23] L. V. Kale, S. Krishnan, CHARM++: A Portable Concurrent Object Oriented System Based on C++, in: *Proceedings of the Eighth Annual Conference on Object-oriented Programming Systems, Languages, and Applications*, OOPSLA '93, ACM, New York, NY, USA, 1993, pp. 91–108. doi:10.1145/165854.165874.
- [24] F. Yazdanpanah, C. Álvarez, D. Jiménez-González, Y. Etsion, Hybrid dataflow/von-Neumann architectures, *Parallel and Distributed Systems*, *IEEE Transactions on* 25 (6) (2014) 1489–1509.
- [25] G. Contreras, M. Martonosi, Characterizing and improving the performance of intel threading building blocks, in: *Workload Characterization*, 2008. IISWC 2008. IEEE International Symposium on, IEEE, 2008, pp. 57–66.
- [26] H. Kaiser, T. Heller, B. Adelstein-Lelbach, A. Serio, D. Fey, HPX: A Task Based Programming Model in a Global Address Space, in: *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*, PGAS '14, ACM, New York, NY, USA, 2014, pp. 6:1–6:11. doi:10.1145/2676870.2676883.
- [27] I. Papadopoulos, N. Thomas, A. Fidel, D. Hoxha, N. M. Amato, L. Rauchwerger, Asynchronous nested parallelism for dynamic applications in distributed memory, in: X. Shen, F. Mueller, J. Tuck (Eds.), *Languages and Compilers for Parallel Computing*, Springer International Publishing, Cham, 2016, pp. 106–121.
- [28] G. Bosilca, A. Bouteiller, A. Danalis, T. Herault, P. Lemarinier, J. Dongarra, DAGuE: A generic distributed DAG engine for High Performance Computing, *Parallel Computing* 38 (1) (2012) 37 – 51, extensions for Next-Generation Parallel Programming Models. doi:<https://doi.org/10.1016/j.parco.2011.10.003>.
- [29] T. Dallou, A. Elhossini, B. Juurlink, N. Engelhardt, Nexus#: A Distributed Hardware Task Manager for Task-Based Programming Models, in: *International Parallel and Distributed Processing Symposium (IPDPS)*, 2015 IEEE International, 2015, pp. 1129–1138. doi:10.1109/IPDPS.2015.79.
- [30] X. Tan, J. Bosch, M. Vidal, C. Álvarez, D. Jiménez-González, E. Ayguadé, M. Valero, General Purpose Task-Dependence Management Hardware for Task-based Dataflow Programming Models, in: *Parallel and Distributed Processing Symposium (IPDPS)*, 2017 IEEE International, IEEE, 2017, pp. 244–253.
- [31] M. Valero, M. Moreto, M. Casas, E. Ayguadé, J. Labarta, Runtime-Aware Architectures: A First Approach, *Supercomputing frontiers and innovations* 1 (1).