Composing Parallel Science Packages and the Dominance of Performance Characteristics

Matt Martineau and Simon McIntosh-Smith

Merchant Venturers Building, Woodland Road, Bristol, BS82AA, United Kingdom
{m.martineau,cssnmis}@bristol.ac.uk

Abstract. In this paper we discuss the composition of parallel packages, and how this affects performance on highly parallel supercomputing hardware. We introduce four new mini-apps, Fast, Hot, Wet, and Bright, which perform a parallel FFT, solve heat diffusion and hydrodynamics, and track neutral particles. The mini-apps represent four of the seven Dwarfs of parallel programming, and have been developed with the purpose of evaluating the impact of coupling those algorithms together. Although much research has been performed that evaluates the performance profile of the seven Dwarfs, production applications rarely utilise only a single package for solving a scientific problem. We start by presenting a motivational comparison of the Dwarfs, considering their invidual performance characteristics and which of those characteristics are dominant. Following this we present performance results of coupling Hot, Wet, and Fast together, using them to solve test problems on modern hardware including Intel Xeon Broadwell, Xeon Phi, IBM Power8, NVIDIA Kepler, and NVIDIA Pascal.
The paper ends with a discussion of some of the most important performance issues that we can forsee when coupling applications together, motivating a whole stream of future work in the area.

Keywords: performance portability,mini-apps,high performance computing,openmp 4

1 Introduction

A great amount of work is being devoted to understanding how scientific production applications will be ported to run on modern supercomputers. Some context of the challenges currently faced by the community is important to demonstrate the motivation of this paper.

1.1 The Use of Mini-Apps to Optimise for Modern Architectures

It is simply not possible

It is considered an onerous and complicated task to program GPUs and until recently there has been little traction within the community to port existing large-scale applications, even though some of the largest supercomputers in the world, for instance Titan, include both CPUs and GPUs. We have shown in our previous research that GPUs can be programmed effectively, in most cases, using existing programming models. We have shown that productivity can be an important deciding factor in the choice of programming model, and further showed that directive-based models can offer a highly productive interface, that can be portable given good implementations, whilst not significantly sacrificing performance.

1.2 Extending the Single Mini-App Focus

This paper is a window into a set of work that is attempting to extend the success of mini-apps for evaluating modern architectures and programming environments. Our hope is that we can begin to capture the features of production scientific codes that have been yet unexplored.

Much of the mini-app optimisation focus has been limited to individual applications which have, in general, performed well on modern hardware, using a range modern parallel programming models. Up until this point the individual mini-apps have failed to capture some important aspects of large-scale scientific applications, in that they are isolated instances of algorithms that do not encapsulate the multi-package hierarchies that scientists are truly interested in seeing results for.

We recognise that many of the issues we discuss or uncover will have been discovered by application developers attempting to combine multiple packages together in the past. However, it is essential to the relevance of results obtained by mini-apps that those features are considered.

In order to limit our focus to the most important algorithms that might be encountered within scientific production codes, we will follow the seven Dwarfs of parallel programming. In the 2006 View from Berkely paper [**?**], a number of important classes of parallel algorithm were discussed and categoried as computation Dwarfs. Each of the Dwarfs exposes a diverse computational and communicational pattern and can encompass many of the algorithms you are likely to see in production scientific applications.

Their paper eluded to the necessity for those Dwarfs to be composed together and investigated as coupled application. Our intention is to use a new suite of mini-apps, tailored for composition, to expose the dominant performance characteristics when combining multiple packages, as would be seen in a common scientific production application. Each mini-app represents on of a subset of the Dwarfs, and each was developed to ensure they are general, with great care taken to ensure that they capture the desired performance profiles of their proxied applications.

2 Contributions

This research is primarily focussed upon discussing the implication of composing a subset of the parallel application Dwarfs, and we present a number of important contributions in order to support the discussion:

- We have developed four mini-apps that serve as proxy applications for important scientific algorithms. An FFT solver for Poisson equations (Fast), a heat conduction application (Hot), a fluid dynamics application (Wet), and a Monte Carlo Particle Transport application (Bright).
- We present results for those algorithms in isolation on modern supercomputing devices: an NVIDIA Tesla P100, an Intel Xeon Phi Knights Landing processor, an IBM POWER8 CPU, and a Intel Xeon Broadwell.
- We naively couple the Hot, Wet, and Fast mini-apps together to demonstrate their composed performance profiles.
- We present the results of running those applications independently and composed at scale.
- We offer some insights into the issues and opportunities that might be experienced when composing algorithms that goes beyond the scope of our current experiments.

3 Background

In this section we will present some basic details about each application, in order to make clear what each application is intended to simulate. Our intention was to choose the simplest methods possible, whilst achieving a reasonable level of accuracy, and in particuler focused upon capturing the performance profile of their respective classes of algorithms.

In order to give a clear discussion, we will explain the initial development of each application prior to discussing our efforts to couple them together. This means that we will explain how each application was optimally developed, but with the caveat that this would likely have to change once coupling was required. We did not particularly consider the coupling up-front, instead opting to choose the best strategy for each application in isolation, maximising the potential outcomes of the coupling exercise.

3.1 Composition of Dwarfs

When considering the coupling of parallel algorithms, there is the potential to consider any number of existing packages. In this research we have chosen to continue on from the work discussed by Asanovic et al. [?], who outlined a number of Dwarfs, each of which describes a different computational class that features in modern supercomputing applications. They continue to describe an additional 6 Dwarfs, but this paper will only investigate the original seven.

Each of the Dwarfs has a unique set of computational and communicational characteristics that mean that they have different requirements of modern computing hardware. While many of those characteristics have been analysed and optimised heavily on existing architectures, we hope to uncover the extent to which those classes can co-exist within an application. At this stage we will reiterate that that the majority of production applications that solve significant scientific problems will require packages encompassing a number of the classes described by the Dwarfs.

The classes described by the Dwarfs are broad, and they encapsulate a subset of possible performance characteristics, which may be present in other Dwarfs. We hope to discover which is of those characteristics is most important, in the context of parallel performance on modern hardware, by considering the Structured Grid, Monte Carlo, Sparse Linear Algebra, and Spectral Methods Dwarfs. In particular, we consider four applications of those classes, hydrodynamics (Wet), neutral particle tracking (Bright), implicit heat diffusion (Hot), and the Fast Fourier Transform (Fast).

Some of the most important performance characteristics of those particular applications are listed in Table 1.

Property	**Hot**	**Wet**	**Fast**	**Bright**
Nearest Neighbour	Yes	Yes	No	No
Dynamic Locality	No	No	No	Yes
Inner Iterations	Yes	No	No	No
Iterative	No	No	No	Yes
Mesh-based	Yes	Yes	Yes	Yes

Table 1. Performance characteristics of the four mini-apps.

Our intention is to investigate how the computational and communicational profiles of those applications interact with eachother. Prior to discussing the coupling stage and performance evaluations, we will describe the individual applications and how they are developed.

We can pre-emptively assess some of the likely effects that running multiple mini-apps at once might observe. For instance, we would expect that there could be an impact on cache utilisation, and a generally increased memory footprint, while some state could be shared potentially improving the overall performance. It is also possible that data structure between the individual applications will require some form of transformation, which we expect to have a major influence on performance. Aside from harming the performance overall, we are hopeful that there will be opportunities to exploit overlapping in some cases, and hide the cost of expensive operations.

3.2 Fast - Fast Fourier Transform for Poisson's Equation

Later in the paper we will see that there are certain dominant characteristics of packages or algorithms that can have a major impact on the overall perforamce of an application.

The fast fourier transform (FFT) is an important algorithm to a number of scientific and engineering domains, and the literature is extensive on optimising for parallel execution. As part of this research we have developed a highly simplified FFT mini-app, which solves Poisson's equation using an FFT.

The mini-app is flexible and can use a hand-rolled algorithm for the actual FFT and inverse FFT steps, but also provides an option to use the optimised Intel MKL library version of the FFT. We do not believe that the actual implementation of the fundamental FFT is particularly important to the overall impact that including an FFT solver in a multi-package application. This point will be discussed extensively throughut the paper, but the most dominant feature of the FFT is not the computational bound that is created by it's arithmetic intensity, as the communicational patterns and data layouts are far more dominant factors in the performance of the algorithm.

Following on from the discussions in Gholami et al., we utilise the FFT to solve Poisson's equation for some dependent data passed to the solver. The general solve phase is as follows:

$$f = -\Delta u \tag{1}$$

$$\mathrm{fft}(f) = f' \tag{2}$$

$$(f' \circ \Delta^{-1}) = f'' \tag{3}$$

$$\mathrm{fft}^{-1}(f'') = f''' \tag{4}$$

Clearly the mathematical requirement of this package is quite low, especially when compared to Wet, which requires a cohesion between many numerical methods. The resulting algorithm purposefully simple however, whilst capturing the dominant performance characteristics exhibited by algorithms that take advantage of the FFT.

Parallelisation The parallel decomposition of the FFT operation is architecture dependent, and in particular varies depending on how many computational elements the problem will be executed on. This is particularly important when composing an FFT package with other packages, as different decompositions might negatively impact on the performance of the whole application leading to a larger search space of decisions when optimising an application.

3.3 Hot - Heat Conduction via a CG Solver

We have developed a simple conjugate gradient (CG) solver that implicitly solves the heat conduction equation in order to solve the problem within a reasonable time-frame with acceptable fidelity. Our implementation uses a standard preconditioner, but otherwise adopts the simplest CG approach.

The conjugate gradient method is iterative, and reduces the error of a guessed solution by walking along the eigenvectors of the solution vector. We will not develop or explain the mathematics underpinning the CG solver, and direct interested readers to explore the wealth of existing literature [] (PUT A LOAD OF CG REFERENCES IN). The only physical feature of the application is manifested in the density calculations, where the densities are stored at the cell centers and interpolated to the edges.

The best decomposition for this algorithm is essentially regular, minimising the surface area to volume of each rank to reduce communication overheads. In terms of boundary conditions we selected reflective, as this was simple, and improves the verifiability of the results. It is also useful when coupling to the other systems.

Parallelisation The algorithm does not expose any load imbalance, and the MPI communication was simply nearest neighbour halo exchanges, performed after each timestep. As each of the core kernels is essentially a simple linear algebra method, such as a dot product or sparse matrix-vector multiply, it was straightforward to parallelise with OpenMP and CUDA, given the extent of the data-parallelism exposed by the algorithms.

At each iteration an alpha and beta value is calculated, which needs to be distributed amongst all of the ranks. The implication is that each iteration requires two calls to MPI_Allreduce, an obvious performance bottleneck at scale.

3.4 Wet - Fluid Dynamic via Direct Lagrangian-Eulerian Flux Calculations

Our fluid dynamics simulation maintains the simplest possible methods, whilst keeping second order in space for all dependent variables. We chose to stay first order in time because algorithms that are second-order in time are somewhat more complicated, and the benefit is primarily reduce wallclock time. Our initial instinct is that the relative runtimes of each of the applications was not particularly important at this stage, but may revisit this issue in a later publication.

The direct Lagrangian-Eulerian algorithm takes Euler's equations, and explicitly solves them across a staggered structured grid. The application uses simple smoothing to simulate artificial viscosity, with shock heating accounted for as part of the mechanical work update. For the mass flux calculations, a Van-Leer flux limiter is used to maintain a monotonic profile at shock boundaries, and second-order interpolations are performed for energy and momentum to ensure monotonic behaviour for all dependent variables.

Directional splitting is used in order to make the application two-dimensional, and we chose to maintain symmetry by alternating the directional calculations with each timestep. Our explicit timestep controls limit the timestep based on the artificial viscosity coefficients and stop any flux extending beyond a single cell.

As with the heat condution application, the boundary conditions we chose for this application are reflective. This is particularly useful for fluid dynamics as tracking of conservation is important in verifying the results.

Parallelisation As with the heat conduction application, the algorithm does not have any inherent load imbalance, and only requires nearest neighbour communication. The explicit nature of the algorithm means that multiple kernels are invoked in order, and the independent kernels are inherently data parallel, allowing them to be easily threaded for CPU and GPU.

3.5 Bright - Monte Carlo Particle Transport

Our transport application uses a Monte Carlo particle tracking method, making it quite different to the other two algorithms. Our initial effort split the problem into batches of particles, each of which had an individual history that was followed in a time-dependent manner.

A major aspect of the application is the data structure used to describe particles. Given that we did not have prior experience with this particular algorithmic class, we were not exactly sure which of the potential approaches would end up being the most performant, so we explored the space and performed some testing which we describe in detail later. Ultimately, we described a single particle using a data structure that contains the particle's position in space, direction, energy, and initially tracked the particles location by cell on the mesh.

Our simple tracking procedure would maintained several running timesteps to different event types: (1) a Collision event, which is where the particle would either change direction or have some of its weight absorbed, (2) a Facet event, where the particle reached the boundary of a cell or zone, and (3) a Census event, where the particle reached the end of the current timestep. Of course, the census events for the application in isolation were simply set to some small constant size, with the intention for this timestep to be later controlled by the requirement of the CFL condition exposed by the fluid dynamics simulation.

In order to determine if a Collision event has occured, we perform a lookup to cross sectional data taken from the ENDF database. In our particular approximation, we only consider absorption and scattering, ignoring any contributions due to fission. In order to reduce the overall variance of the simulation, we extend the lifespan of particles by giving them a weight, and reducing this weight each time an absorption event occurs, rather than declaring them as dead particles. Once the weight has reduced past a certain point, or the particle has reached a low enough energy, we will only then consider the particle to be dead.

When a Collision event occurs and the result is a scattering of the particle a random number is generated to determine the angle of scattering, and this is also used to determine the level of energy dampening that occurs using well known relations [].

To perform facet intersection checking, we are able to leverage the simple geometry of the structured grid to calculate the new direction using a simple method of our own devising. We first check which of the edges would be hit first if the particle were travelling at the same speed but only along a single axis, and use this to solve the system of equations that arises to determine the intersection point. We have considered that this might not be optimal, and doesn't generalise beyond the basic geometry presented by our structured grid case.

In terms of boundary conditions we did select reflective at this stage, as this made it easier to track some conservation in the system, and matches the approach taken with the hydrodynamics solver, although we did recognise that it would likely be more representative if the boundary conditions were transmissive.

Of course, all of this requires a sufficient random number generator with a large period. In order to fulfil this requirement we leveraged an existing lightweight Mersenne-Twister algorithm, to generate random numbers for the whole system.

Parallelisation Unlike the other two applications, the Monte Carlo method exposes some load imbalance. There are a number of approaches to improving this problem, and we explored some of them to select the best for our particular set of requirements.

In terms of parallelising the algorithm there were several factors to consider, for instance it is important to maintain good data locality in particular for vectorisation on CPUs and coalescense on GPUs. Our initial strategy was to perform a coarse-grained parallelism, splitting the particles into batches in order for them to be processed in parallel. Of course this has implications in terms of the memory layout as the particles are tracked through space.

4 Coupling

As this application is developed by computer scientists and primarily focusses upon the evaluation of performance profiles on modern architecture, the exact physical solution to our test problems is not necessarily important. Of course, we are conscious of the fact that couplings of certain algorithms and for modelling particular scientific concepts could lead to different performance profiles.

We could see two different approaches to coupling the applications, the first of which involves directly coupling their constituent equations and solving the whole system at once. This approach has two important drawbacks: (1) the complexity of this approach is significant, especially considering that we plan to extend this project into a suite of seven algorithms, and (2) boiling the algorithms under a single solve wouldn't capture the potential difficulties of executing diverse routines together.

Given this, our choice of applications allows for a simple coupling stage that requires very little scientific knowledge to achieve a reasonable, albeit inaccurate, outcome. Although some variations of all four mini-apps (Hot, Wet, Fast, and Bright) exist, this research presents the results of coupling Hot and Wet together. We chose to start with those applications as they have similar communication patterns but significantly different computational profiles (SOMETHING

STRONGER?). The inputs of the hydro package are density and energy, and momentum in each dimension, while the inputs for the diffusion package are density and temperature. As such we performed a simple conversion between internal energy and temperature between the packages, and ran them sequentially.

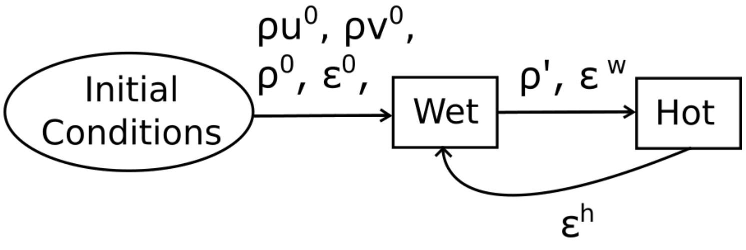

Fig. 1. The flow of our coupled Hot and Wet applications.

Figure 1 demonstrates the current coupling, as well as the prospective coupling we plan to extend this work to accommodate.

5 Asynchronicity Among Packages

Modern hardware presents a number of challenges and some important opportunities. The balance of different packages within a full application might allow new avenues for overlapping long communications or IO times, and even the potential for overlapping the compute of packages within the full application. We are interested to explore a number of these issues and expect that this could be a future direction for work but have not investigated that point in this research.

6 Results

The results presented in this

6.1 Experimental Setup

To fully assess the impact of composing the Dwarfs, we will test our coupled applications on a range of modern HPC hardware. We recognise that the Hot and Wet mini-apps are both memory bandwidth bound, and the Fast application is compute bound, however the

6.2 Hot'n'Wet

As we are presenting two applications that have been developed from scratch, we provide some simple scaling studies to demonstrate that each application

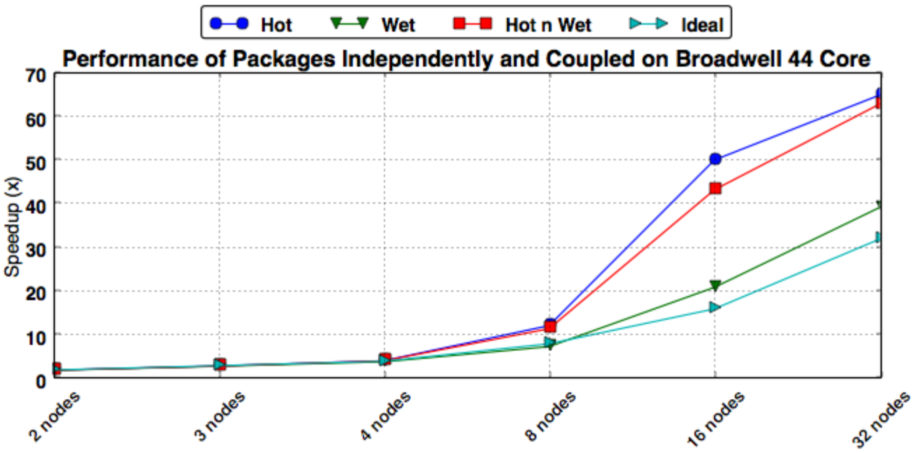

Fig. 2. Scaling the paired applications on multiple nodes of 44 core Broadwell CPU.

performs to an acceptable level. This data gives us some baseline for comparison as we consider the applications within a multi-algorithmic execution setting.

As can be seen from Figure 2, the results of scaling the application across 32 nodes of 44 core Broadwell CPUs has resulted in a relatively consistent performance profile. Each of the application beats the ideal scaling performance over a single node due to the improved utilisation of cache that occurs as the problem is decomposed into smaller chunks. There is a noticeable gap in the scaling of the two solver at 16 nodes, that greatly reduces by the time that 32 nodes have been reached. This appears to be an effect of the increased data requirement that comes from executing two applications sequentially, where the overall capacity required to fit within cache is increased and so the improvements due to cache have been slightly delayed.

Overall this scaling profile suggests that the composition of the two application on a CPU has achieved an expected level of performance cohesion. We wanted to continue this investigation on some other types of hardware to ensure that there weren't any other effects that we had not anticipated.

The scaling was generally quite poor on this node and we only achieved a maximum of 50% efficiency even for the invidual applications. We were surprised to see that the composition of the two applications had lead to some performance bug, which had totally limited scaling on the multi-GPU node.

Our results here are fairly unremarkable, and support our hypothesis that the two applications investigated are performance compatible, meaning they do not interfere with the performance profile of eachother. In order to increase the positive or negative influence of

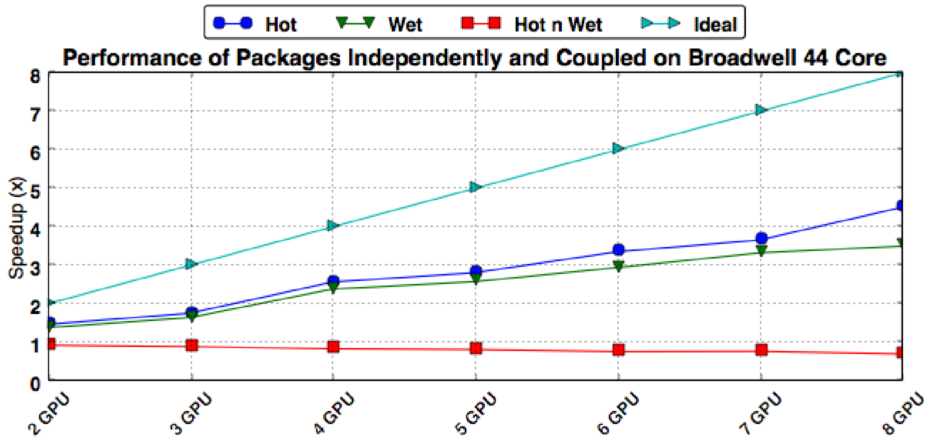

Fig. 3. Scaling the paired applications on a CS STORM node containing 8 NVIDIA K40 GPUs.

7 Fast'n'Hot or Wet'n'Fast

8 Conflict Resolution

In the event that the data structures within two packages conflict, we can theorise some potential resolutions:

- **Pick the best decomposition for the worst package** - This may or may not be easy depending on the restrictions imposed by the other packages. This is likely only suitable when you have a significant difference between the relative costs of the packages.
- **Choose some mutually beneficial layout** - This could be a complicated decision to make, and might lead to specific novel layouts for particular couplings. If the relative costs of the two packages are very similar, then we expect this to be a good choice.
- **Perform a transposition of the data between packages** - Considering a simple two package application, with a victim that requires it's data transposed, this would lead to two transposition at entry to and exit from the victim package. Ideally this transposition could be overlapped with independent work.

9 Patterns Affecting Performance for Multi-Package Applications

Although both Hot and Wet combined together successfully, our research has uncovered a number of areas that we recognise present risks or opportunities to the performance of multi-package applications.

- **Dynamic Meshes -** In the event that mesh data has to move between applications, there is significant potential.
- **Competing Decompositions -** It is possible that different packages within an application might require different decompositions for optimal performance. We extend this discussion in Section 9.1 to predict a key bottleneck to multi-package scaling.
- **Capacity Requirements -** As you increase the number of packages that are included in an application, the capacity requirements can be inflated significantly. It will be important to overlap as much of the capacity as possible in order to minimise this effect.

We hope that our future research can extend the results of this paper to begin to understand the issues we hypthesise here.

9.1 Diminishing Scalability for Increasing Core Counts

Modern supercomputing has reached a scale where the core counts at the node level are increasingly significantly each year, and this has major implications for writing portable and performant code. In particular, many of the largest supercomputers in the world include heterogeneous devices, such as accelerators, which greatly increase the core count, and require that significant data-parallelism is exposed within scientific application's algorithms in order to perform well.

Accelerator devices like the Intel Xeon Phi processors, have greater numbers of cores than existing Intel CPUs, for instance the Knights Landing has between 64 and 72, each with 4 hardware threads, whereas the Intel Xeon Broadwell can have up to 22 cores per socket, each core supporting up to 2 hardware threads. The IBM POWER8 CPU also exposes a large core count, with up to 12 cores per socket and 8 hardware threads for a total of 192 threads.

This continual increase in the number of cores available on a node has major implications for the scalability of codes, and this is amplified by the increasing number of nodes that are hosting those devices. It is well known that many algorithms reach a turnover point for strong scaling, past which communication costs begin to overwhelm the performance, and decomposing a problem will experience diminishing returns. Regardless of node-level application performance, as communication becomes the limiting factor, we are reaching a stage where the scalability of applications demands that the highly parallel applications exploit shared memory as much as possible.

We can see here a significant risk to the overall performance of

10 Future Couplings

We believe that it will be useful to extend this work to investigate the performance of diverse applications types, in particular we plan to develop new applications that cover the seven dwarves of parallel programming. There are many possible combinations that can occur within a multi-package applications depending on the requirements of the particular scientific question being solved. It is possible to consider many potential

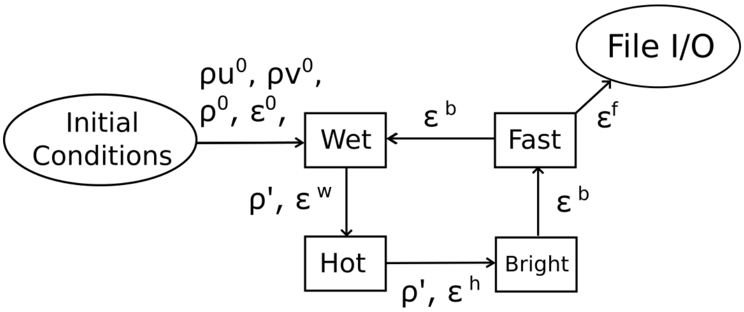

Fig. 4. The flow of a prospective multi-package application.

11 Future Work

We further recognise that there are many requirements not handled by our existing applications, such as three-dimensional solves, multiple materials, and complex unstructured geometries. Each of these has a significant influence on the performance of individual algorithms, and likely introduces new issues for coupling packages together. As such it will be essential to consider new mini-apps that can proxy those particular features and consider their performance on new hardware.

12 Related Work

(SPEAK TO W ABOUT THIS SECTION, WHAT IS ACCEPTABLE TO BE TIED TO THIS PAPER)

13 Conclusion

Importantly we have been able to observe an important fundamental rule with multi-package applications, that the complexity and computational cost of the individual package matters less to the overall performance of the application that the dominant conflicting performance characteristics present amongst the applications.

References