Design and Analysis of Two Highly Scalable Sparse Grid Combination Algorithms

Peter E. Strazdins  
Research School of Computer Science  
Australian National University  
Canberra, Australia  
Peter.Strazdins@cs.anu.edu.au

Md Mohsin Ali  
Research School of Computer Science  
Australian National University  
Canberra, Australia  
md.ali@anu.edu.au

Brendan Harding  
Mathematical Sciences Institute  
Australian National University  
Canberra, Australia  
brendan.harding@anu.edu.au

Abstract—Many petascale and exascale scientific simulations involve the time evolution of systems modelled as Partial Differential Equations (PDEs). The sparse grid combination technique (SGCT) is a cost-effective method for solving time-evolving PDEs, especially for higher-dimensional problems. It consists of evolving PDE over a set of grids of differing resolution in each dimension, and then combining the results to approximate the solution of the PDE on a grid of high resolution in all dimensions. It can also be extended to support algorithmic-based fault-tolerance, which is also important for computations at this scale.

In this paper, we present two new parallel algorithms for the SGCT that supports the full distributed memory parallelization over the dimensions of the component grids, as well as across the grids as well. The direct algorithm is so called because it directly implements a SGCT combination formula. We give details of the design and implementation of a ‘partial’ sparse grid data structure, which is needed for its efficient implementation. The second algorithm converts each component grid into their hierarchical surpluses, and then uses the direct algorithm on each of the hierarchical surpluses. The conversion to/from the hierarchical surpluses is also an important algorithm in its own right. It requires a technique called sub-gridding in order to correctly deal with the combination of very small surpluses. An analysis of both indicates the direct algorithm minimizes the number of messages, whereas the hierarchical surplus minimizes memory consumption and offers a reduction in bandwidth by a factor of $d$, where $d$ is the dimensionality of the SGCT. However, this is offset by its incomplete parallelism (70–80%) and a factor of $2^d$ load imbalance in practical scenarios. Our analysis also indicates both are suitable in a bandwidth-limited regime and that the direct algorithm is scalable with respect to $d$. Experimental results including the strong and weak scalability of the algorithms indicates that, for scenarios of practical interest, both are sufficiently scalable to support large-scale SGCT but the direct algorithm has generally better performance, at least by a factor of 2 in most cases. Hierarchical surplus formation is much less communication intensive, but shows less scalability with increasing core counts. Altering the layout of processes in the process grids and the mapping of processes affects the performance of the 2D SGCT by less than 10%, and affects even less the application part of an SGCT advection application.

I. INTRODUCTION

Large scale scientific simulations at the petascale and exascale are capable of making great advances in computational science. Usually, these involve the numerical solution of Partial Differential Equations (PDEs). There are two main obstacles for scientific applications in order to reach this goal: achieving sufficient scalability, and ensuring their reliable completion. Both of these problems arise due the immense size, and hence number of components, of supercomputing systems at these scales. The latter problem arises due to the fact that these simulations are typically long-running, and that the failure rate of the system is proportional to the number of components.

PDEs are normally solved on a regular grid. With uniform discretization across all its dimensions, the number of grid points increases exponentially with the increase of dimensionality. This behavior makes the high-dimensional PDE solver computationally expensive. In order to address this issue, the PDE can be solved on a grid with substantially fewer grid points than the regular full grid. These grids are called sparse grids [1]. An example of a sparse grid is shown in Figure 1 where the sparse grid is represented as the union of smaller regular grids.

A numerical method called the Sparse Grid Combination Technique (SGCT) [2], [3] can be employed to approximate the solutions to PDEs on the sparse grid. Instead of solving the PDEs on a full grid, it solves them on several anisotropic grids with substantially fewer grid points, called the sub-grids or component grids, as shown in Figure 1. Solutions on these sub-grids are then linearly combined to approximate the solution on the sparse grid. The technique can be applied in principle to any PDE, but sufficient ‘smoothness’ of the solution is required for good accuracy.

An important technique associated with sparse grids is the formation of hierarchical surpluses [4], [5] on regular grids. These facilitate error analysis over the various dimensions at different resolutions, and can be used to determine when refinement is required on adaptive grids [5].

As well as making high-dimensional PDEs tractable, employing the SGCT can result in computational efficiencies even for lower-dimensional problems [2], [3], [6], [7]. Fur-
thermore, it inherently supports parallelism in the sense that component grids can be solved completely independently. The SGCT can also be extended to support algorithm-based fault-tolerance [8, 9, 10]. Upon detection of a fault, the SGCT can be applied in such a way that the component grids associated with the failed nodes are avoided. The lost data can be recovered by down-sampling from the combined (sparse) grid, and the computation can continue sustainably. The ability to repeatedly apply the SGCT over a computation is also useful in improving the accuracy of the sparse grid approximation in certain cases [6, 7].

Thus, there is an urgent need to have highly parallel SGCT and hierarchical surplus formation algorithms for where the component grids themselves are large and must be solved in parallel. The contributions of this paper are to (i) present two new algorithms for the SGCT which support fault-tolerance and are suitable for computations over a large number of cores and (ii) perform an analysis and detailed experimental evaluation of their performance.

To the best of our knowledge, this is the first work that does so for a SGCT algorithm which supports distributed memory parallelism both across and within the component grids. The direct algorithm is furthermore highly general in terms of the process grid sizes used for the component grids (the hierarchical surplus algorithm requires these to be powers of 2). We also believe that we also give the first general distributed memory algorithm for the conversion of a grid to/from its hierarchical surpluses. We also are the first to discuss the concept of a partial sparse grid, and give the first implementation of a distributed sparse grid data structure. In order to do so, we introduce the concept of the sparse block distribution, which we believe is novel. Furthermore, we show how the hierarchical surpluses can be coalesced and scheduled, and show that the former is crucial for its efficiency. We give a deep analysis of the scaling behaviour of these algorithms, followed by a comprehensive experimental evaluation. Finally, we expect that our approach to algorithm design and implementation using vector arithmetic and associated distribution mappings will be applicable to other communication-intensive algorithms over high dimensional spaces.

The paper is organized as follows. Section II describes the SGCT technique. Details of our implementation of a highly scalable SGCT algorithm are given in Section III followed by an analysis in Section IV. Experimental results are given in Section V related work is discussed in Section VI and conclusions are given in Section VII.

II. THE SPARSE GRID COMBINATION TECHNIQUE

Consider the SGCT for the 2D case. Each component sub-grid $G_i$, where $i = (i_x, i_y)$, is assumed to have $(2^{i_x} + 1) \times (2^{i_y} + 1)$ grid points with a grid spacing of $h_1 = 2^{-i_x}$ and $h_2 = 2^{-i_y}$ in the x- and y-directions, respectively, where $i_x, i_y \geq 0$. If we consider a square domain, then the grid points of $G_i$ are $\{(x', y') | x' = 0, 1, \ldots, 2^{i_x}, y' = 0, 1, \ldots, 2^{i_y}\}$.

In the more general case, the index space for the grids will be some finite set $I \subset \mathbb{N}^d$. If $u_i$ denotes the approximate solution of a PDE on $G_i$, the combination solution $u_I^c$ generally takes the form

$$u_I^c = \sum_{i \in I} c_i u_i,$$

where $c_i \in \mathbb{R}$ are the combination coefficients. Clearly, the accuracy of the combination technique approximation depends on the choice of the index space $I$ of the subgrids and their respective coefficients. For the 2D case, good choices of the coefficients are $\pm 1$. For instance, in the classical case, we have for level $l$ the set $I = \{(i_x, i_y) | i_x, i_y \geq 0, l - 1 \leq i_x + i_y \leq l\}$ and the combination coefficients are $c_i = 1$ if $i_x + i_y = l$ and $c_i = -1$ if $i_x + i_y = l - 1$ where $i = (i_x, i_y)$. This provides the following combination formula

$$u_I^c = \sum_{i_x + i_y = l} u_i - \sum_{i_x + i_y = l - 1} u_i,$$

which is depicted in Figure 2.

In contrast to the full grid approach which needs $O(h_n^{-d})$ grid points, the SGCT works with only $O(h_n^{-1} \log_2(h_n^{-1})^{d-1})$ grid points, where $h_n$ denotes the employed grid spacing and $d$ is the dimension. The accuracy of the solution obtained from the SGCT deteriorates only slightly from $O(h_n^2)$ to $O(h_n^{2} \log_2(h_n^{-1})^{d-1})$ for a sufficiently smooth solution.
A. Hierarchical Surpluses

Approximations on component grids are typically computed and stored using a nodal basis representation, that is each element \( v_k \) in the vector \( v \) gives the function’s value at the grid point \( x_k = k \cdot 2^{-i} \), that is \( v_k = f(x_k) \). An alternative representation is via the hierarchical basis \( \phi \) whereby each element of the vector is the difference between the function’s value’s at the corresponding grid point and the function’s values at the hierarchical neighbours. Writing \( v \) using the level notation \( v_{l,k} \) where \( l \in \{0, ..., i\} \) and

\[
k \in B_l = \begin{cases} \{i \in \mathbb{N}: 0 < i < 2^l, i \text{ odd}\} & \text{if } l > 0 \\ \{0, 1\} & \text{if } l = 0 \end{cases}
\]

one has the hierarchical representation

\[
v_{l,k} = \begin{cases} f(x_{l,k}) - \frac{1}{2} \left( f(x_{l-1,(k-1)/2}) + f(x_{l-1,(k+1)/2}) \right) & \text{for } l > 0 \\ f(x_{l,k}) & \text{for } l = 0 \end{cases}
\]

where \( x_{l,k} := x_{k,2^l-1} \). These \( v_{l,k} \) correspond to the heights of the hierarchical basis functions depicted in Figure 3. The collections \( \{v_{l,k}\}_{k \in B_l} \) for each \( l \) are referred to as hierarchical surpluses. The hierarchical basis function in 2 or more dimensions is obtained by taking the tensor product of the one dimensional hierarchical basis functions. For full details we refer the reader to \[4\], \[5\]. Rather than perform the combination technique directly over the component grid approximations (in nodal basis) one may instead convert each component grid to hierarchical basis, combine each of the hierarchical surpluses (distributing the result) and then reconstruct the nodal basis for each component grid. This is advantageous over direct combination for two reasons: 1) it reduces the communication volume \[10\]; 2) it allows one to avoid interpolation onto a full grid or sparse grid structure thus reducing compute cycles and memory requirements.

The formation of the hierarchical surpluses involves applying Equation 3 across each of the dimensions of a component grid \( G_i \). Each element in a grid will correspond to a hierarchical surplus of index \( j \), where \( j \leq i \). This is illustrated for \( i = (3, 3) \) in Figure 4. The hierarchization process occurs in-place, with the surpluses computed from the initial grid values.

Consider the hierarchization across the \( x \) dimension in row 0. The elements with hierarchy 30 are formed by subtracting the average of the elements at distance 1 to the left and right (‘30’ is a shorthand for ‘(3,0)’). Then elements with 20 are formed by subtracting the average of the elements at distance 2 from right or left. Finally, the single element of hierarchy 10 is computed by subtracting the average of the elements in hierarchy 00 at a distance 4 to the left and right.

B. Truncated Combinations

Our SGCT algorithms support so-called ‘truncated’ combinations \[11\], which avoids using the highly anisotropic grids (e.g. \( G_{(1,i)} \)) close to the axes of the grid index space. This avoids the problem of minimum dimension size imposed by some applications \[7\]. Furthermore, the highly anisotropic grids have been known to contribute least towards the accuracy of the sparse grid solution or cause convergence problems \[11\], enabling us to concentrate process resources on the more accurate sub-grids. Finally, it removes the limitation that the approximated full grid need not be isotropic, i.e. have an index of \((l,l)\).

In this context we use a different notion of level to that described above in describing how much smaller the sub-grids are relative to some full grid \( G_i \). In particular, a level \( l \leq \min\{i_x, i_y\} \) combination in this context consists of sub-grids from the index set:

\[
I = \left\{(i_x, i_y) : (i_x' - l, i_y' - l) < i \leq i_x' + i_y' + 1 - l \right\}.
\]

It can be noted from the above that the truncated combination formula is a generalization of the standard SGCT formulation, which has \( i' = (l, l) \).

III. Parallel SGCT Algorithms

In this section, we present first the direct and then the hierarchical surplus based algorithms. Finally, we discuss their current limitations and how they may be extended.
These algorithms have a succinct description when expressed via d-dimensional vector arithmetic, and we use the following notations (explained for the case \( d = 2 \)). Given \( M, N \in \mathbb{N}^d \), and \( a \in \mathbb{N} \), then \( M \leq N \) means \((M_x \leq N_x) \land (M_y \leq N_y)\) and \( a \leq N \) means \((a \leq N_x) \land (a \leq N_y)\). Also \( M \ast N = (M_x \ast N_x, M_y \ast N_y)\) and \( a \ast N = (a \ast N_x, a \ast N_y)\). This applies to all other (integer) arithmetic operators, which include \% and \(==\) (which have the same meaning as in the C language). We also use the vector to scalar notations \( \Sigma(a) \) and \( \Pi(a) \) for the sum and product of elements of \( a \), i.e., \( \Sigma((a_x, a_y)) = a_x + a_y \) and \( \Pi((a_x, a_y)) = a_x \ast a_y \) in the 2D case.

Crucial to their description (and to the reliable construction of their implementations [12]) are the mappings of global grid lengths and offsets to local lengths and offsets (which break down in turn to process offsets and index offsets within a process). If \( p \in \mathbb{N}^d \) is the id of a process on a process grid of size \( P \), with \( 0 \leq p < P \), and given \( \hat{N} \in \mathbb{N}^d \) with \( 0 \leq \hat{N} < N \), and letting \( n = N/P \), we define the following vector functions for a block-distribution of \( N \) over \( P \):

\[
l(N, p, P) = n \ast (p == P - 1) \ast (N\%P) \\
g_b(N, p, P) = p \ast n \\
p(\hat{N}, N, P) = \min(\hat{N}/n, P - 1) \\
o(\hat{N}, N, P) = (\hat{N}/n == P - 1) ? \hat{N} - n \ast (P - 1) : \hat{N} \% n
\]

where \( l(N, p, P) \) is the local size of \( N \) on process \( p \), \( g_b(N, p, P) \) is the global index of the 0th local point on \( p \), \( p(\hat{N}, N, P) \) is the id of the process holding global point \( \hat{N} \), and \( o(\hat{N}, N, P) \) is the local offset of global point \( \hat{N} \) on that process.

The above assumes that \( N \geq P \). We now introduce what we will call the sparse block distribution which covers also the case \( N < P \):

\[
l^*(N, p, P) = (N >= P) ? l(N, n, P) : l^*(N, p/2, P/2) \\
g^*_b(N, p, P) = (N >= P) ? g_b(N, p, P) : g^*_b(N, p/2, P/2) \\
\]

where \( r = \frac{P}{N-1} \). The above formulations assumes \( P \) and \( N - 1 \) are powers of 2. The sparse block distribution arises when we take power of two sub-vectors from a block distribution, as for example when forming hierarchical surpluses (see Figure 4). When the sub-vectors become smaller than the process grid, we effectively form a new process grid where the processes holding no elements of the sub-vector are removed.

A. Direct SGCT

The presentation of our SGCT algorithm begins first with the scaled addition of a single sub-grid component into the combination solution. Each sub-grid is assumed to be distributed over an independent set of processes, arranged in a d-dimensional logical grid. The combination solution is similarly distributed over a logical process grid. The parallelization of this over each component grid forms a gather stage. This is illustrated in Figure 5. For computations performing the SGCT repeatedly, it is necessary to reverse this process by sending the respective points of the combined grid back to the component sub-grids. This is called the scatter stage, and can be visualized by reversing the arrows in Figure 5.

1) Single Sub-grid Components: Consider the scaled addition of a single grid into a full grid representation of the combination \( u_i^C \leftarrow u_i^C + cu_i \) in Equation 1. For brevity, we write this as \( u' \leftarrow u' + cu \). Let \( N, N' \in \mathbb{N}^d \) represent the global sizes of \( u \) and \( u' \), respectively. For the SGCT, the elements of \( N - 1, N' - 1 \) will be powers of 2 and \((N' - 1) = r \ast (N - 1)\), where \( r \in \mathbb{N}^d \).

Assume \( u \) and \( u' \) are distributed over process grids of size \( P \) and \( P' \), respectively \((P, P' \in \mathbb{N}^d \), \( 0 < P \leq N \) and \( 0 < P' \leq N' \)).

On process \( p \), the gather of \( u \) to the appropriate process(es) on \( P'' \) is described in Algorithm 1. The loop structure is shown for \( d = 2 \); this generalizes in the obvious way for \( d > 2 \). The number of messages sent is bounded by \( \Pi((P + P')/P + 1) \). On line 9, note that a volume of grid points of size \( dn + 1 \) is sent: an extra uppermost
row and column of points is required for interpolation. It is assumed that $u$ has storage for its $n$ points plus a ‘halo’ of neighbouring points to the positive direction, and that these have been filled by a halo exchange operation from the corresponding processes in $P$ prior to the start of the algorithm.

The receipt on process $p'$ of the grid $u$ (with combination coefficient $c$) and its interpolation into $u'$ is essentially the mirror image of this and is described in Algorithm 2. The number of messages received is bounded by $\Pi((P + P' - 1)/P' + 1)$.

These algorithms are general for arbitrary $P$; as expressed here, it assumes that on $P'$, $g_0(N',p',P') \% r = 0$, that is, the global indices of the 0th points on $p'$ are a multiple of $r$. This is satisfied for the SGCT if the elements of $P'$ are powers of 2.

2) Overall SGCT Algorithm and Repeated SGCT Support: Algorithm 1 is run in parallel on each $P_i$, and Algorithm 2 is run iteratively over $i$ (i.e. for each sub-grid) on $P'$. To support the latter, arrays over $i$ containing the process grid ($P_i$), grid index and combination coefficient of each grid are maintained over all processes in $P'$.

Our implementation can take advantage of MPI_Irecv() semantics by allocating different buffers ($u$) for each grid $i$, and overlapping the interpolation with the receipt of messages.

Where repeated combinations are performed, it is necessary to scatter back the points of the combined grids (with appropriate down-sampling) on $P'$ iteratively to each sub-grid on $P_i$. It is assumed that the processes in $P_i$ have been restored by this stage if there was a failure in $u_i$ detected for the gather stage. The scatter send stage on $P'$ is identical to Algorithm 2 except that lines 10–11 are replaced with:

```c
sample points $i : i + dn' - 1$ of $u'$ into $u$;
```

send the $n$ points of $u$ to $p$;

The receipt of the scatter occurs in parallel over each process grid $P_i$. The algorithm is identical to Algorithm 1 except line 9 becomes:

```c
receive $dn$ points from $p'$ into $u$;
```

store $v$ into points $i : i + dn - 1$ of $u$;

3) Partial Sparse Grid Data Structure for Efficient Interpolation: On the receive stage of the gather operation, if we use a full grid representation of the combined grid $u'$, the interpolation stage (line 11 of Algorithm 2) will involve interpolating on $O(2^l - 1)$ points in the full grid which are not represented by any point in any of the component grids. This is wasteful as these points are never used (in the sample step, above), in terms of both time and space, and furthermore this overhead is exponential in $l$.

This can be avoided by interpolating onto a sparse representation of $u'$. However, when using the truncated combination formula (Equation 4), the resulting data structure will be a generalization of both a regular full grid ($l = 1$) and a (classical) sparse grid ($l = i_l' = i_0'$). We call this representation a partial sparse grid, which represents exactly the union of points in all component grids. This section will describe how to create a distributed data structure for a partial sparse grid and the resulting interpolation and sample algorithms.

Figure 6 gives a visualization of this data structure for level $l = 3$ onto grid $G^i$ where $i' = (4,4,4)$. This introduces a ‘fill-in’ factor $f = i' - l = (1,1,1)$ where any gaps of size $2^l$ in from the corresponding sparse grid...
are filled in (as far as this is possible). Note that the smaller planes must be compressed first so that there are no gaps in their boundary points.

We use concepts similar to the CSR format for sparse matrices, except we take advantage of the regularity of the partial sparse grid to avoid column indexes. Recall that this corresponds to a full grid of size \( N^t = 2^t + 1 \). In the 2D case, global row \( i, 0 \leq i < N^t_y \), will have global length \( N^t_y(i) = 2^{t+1}i + 1 \), where \( l(i) = Z(2(t+1) + 2^t) \) and \( Z(n) \) gives the number of rightmost zeroes before the first 1 bit in \( n > 0 \). The stride between the consecutive elements in this row that must be stored is \( 2^{t-1}(i) \).

The local length in the \( x \)-dimension on process \( P' \) of process grid \( P' \) of the partial sparse grid is then given by

\[
n_x(i) = l^t(N^t_y(i), P'_x, P'_x)
\]

where we are now using the scalar \( (d = 1) \) version of Equation 2. We need to use this equation as now \( N^t_y(i) < P'_x \) is possible and hence we now need to assume \( P'_x \) is a power of 2.

Process \( P' \) computes these lengths for the \( n_y = l(N^t_y, P'_x, P'_y) \) rows beginning from local point 0 (global point \( i = g_0(N^t_y, P'_x, P'_y) \)) and creates the required storage vector \( u' \) and a row index vector \( \text{rx}[0..n_y-1] \). It caches the corresponding row strides \( s[0..n_y-1] \).

Algorithm 3 gives the resulting code for the local interpolation step. It requires both alignment and scaling to take into account the grid row stride \( (s[i']) \). These calculations in turn assume that the first local element in each row corresponds to a global (full grid) index which is a multiple of \( s[i'] \); this will be the case if \( P'_x \) is a power of 2, as assumed previously.

The sample operation (Algorithm 4) iterates over the elements of the component grid instead of the partial sparse grid.

To generalize this to 3D, the \( z \) and \( y \) dimensions take the place of \( y \) and \( x \), respectively, in the above. For the \( j \)-th row in \( y \) of \( i \)-th plane in \( z \), the global length of this row is given by \( N^t_x(i, j) = 2^{t+1} + l(i, j) + 1 \), where \( l(i, j) = Z(2(t+1) + 2^t) \).

![Figure 6. Partial Sparse Grid for level \( l = 3 \) SGCT onto grid (4,4,4), shown in slices across the z plane. 'Filled in' points (and planes) are shown in green, with points from the sparse grid (4,4,4) in black. The compression factor \( c \) indicates the actual distance between neighbouring xy-points on the edges.](image-url)
effectively oblivious to the extra requirements for a fault-tolerant SGCT.

B. Hierarchical Surpluses-Based Algorithm

An alternate approach to the SGCT is to form the hierarchical surpluses on each component grid. This can be done in-place, as observed in Section I-A and this can occur in parallel over each component grid’s process grid (P). Then, for each surplus \( h \in \mathbb{N}^d \), where \( h \leq i’ \), the (direct) SGCT is performed on each. Only the processes holding part of surplus \( h \) need participate, that is those holding the component grid \( G_i \), where \( i \geq h \). Finally the original component grids can be recovered from the combined surpluses, thus performing an overall SGCT.

Only surpluses common to more than one grid need be combined. For 2D, this is for hierarchy \( h \) satisfying:

\[
H2(x, y) : h_x + h_y <= i'_x + i'_y - l
\]

This is illustrated in Figure 7. For 3D, the condition becomes:

\[
\Sigma(h) <= \Sigma(i') - 2l + 1 \land H2(x, y) \land H2(x, z) \land H2(y, z)
\]

As the number of hierarchies is \( O(\Pi(i')) \), applying the SGCT on them individually is likely to suffer from high startup overheads.

In this section, we describe a parallel algorithm for forming the surpluses (and recovering the original grid from its surpluses). This is followed by a description of how to coalesce these hierarchies, in order to try to reduce startup overheads, and how to schedule them where possible in parallel. We then describe how to overcome an issue with this approach, that is how to handle the case where a hierarchy becomes too small for a process grid \( P > 1 \), noting hierarchy \( h \) is of size \( 2^h - 1 \) for \( h > 0 \). Our solution is called process sub-gridding. Finally, we present how the direct SGCT algorithm was extended to operate on coalesced surpluses.

1) Formation of Hierarchical Surpluses: Let \( P \in \mathbb{N}^d \) denote the process grid over which component grid \( G_i \) is distributed and \( N \in \mathbb{N}^d \) be the global size of this grid (\( N = 2^d + 1 \)). Algorithm 5 gives the corresponding parallel algorithm for the 3D case (which can easily be generalized for higher dimensions). The notation \( N \setminus d, x \in \mathbb{N}^d \) denotes a vector with the same elements as \( N \) in all dimensions except \( d \), where it has a value of \( x \in \mathbb{N} \).

First, we calculate the global left and rightmost global indices of points on process \( p \) (line 1). We iterate over each dimension (line 2) and then over each stride length of the hierarchies of that dimension (line 3). Note that the index for the hierarchy is given by \( h = l_d - s \). Then, we calculate the global indices of the leftmost and rightmost source points for this surplus (lines 4–5). With this, we calculate the surpluses for local points which do not need points from neighbouring processes (lines 6–7). “Plane \( j \) of \( u \)” should be read as “plane of global index \( j \) in dimension \( d \) of \( u \).” Note that the local index on \( p \) is given simply by \( j - N_d \). If \( p \) owns the leftmost source point \( j^l \) and another process on the left holds the next destination point \( k \), we send plane \( j^l \) to that process, and similarly for \( j^r \) (lines 8–10). If \( p \) owns \( k \) then we receive the plane \( k^l - 2^r \) from the owning process on the left and plane \( k^l \) can now be hierarchized, and similarly for \( k^r \) (lines 13–18).

To restore the original grid from the surpluses, we simply reverse the order of the stride loop (line 4), and add instead of subtract the averages of the source points (line 7, 15 and 18).

2) Scheduling and Coalescing of Surpluses: A set of hierarchies may be coalesced without introducing redundant communications if they are present on a common set of component grids. This is illustrated for a 2D example in Figure 7 which shows the index spaces for each surplus can be coalesced under this principle. The component grids are indicated in yellow; indices on the upper boundaries (dotted) are common to only one grid and need not be combined.

The surpluses may be combined in the following order. Surpluses \((i'_x - l + 1, i'_y - l + 1), \ldots, (i'_x - l + 1, i'_y - l + 1)\) are combined (separately) first, then those of the next diagonal in the triangular area, until finally surplus \((i'_x - l + 1, i'_y - l + 1)\) is added. It should be noted that each surplus in this triangular area operates on different sets of component grids, and so may not be coalesced with another (without causing redundant communications). Then we combine the surpluses coalesced

\[
\begin{align*}
\text{Algorithm 5: Hierarchization of grid } u \text{ of global size } N = 2^d + 1 \text{ by process } p \text{ in process grid } P. \int \\
\end{align*}
\]
in one dimension, \((0 : i'_{x}−l, i'_{y}−l), ...,(0 : i'_{x}−l, i'_{y}−l+1)\) and \((i'_{x}−1, 0 : i'_{y}−l), ...,(i'_{x}−l+1, 0 : i'_{y}−l)\), and finally those coalesced in both \((0 : i'_{x}−l, 0 : i'_{y}−l)\). This coalescing is possible as they reside on common sets of the component grids.

While the avoidance of communicating the surpluses common to only one grid \(((i'_{x}−1, 0 : i'_{y}), ...,(i'_{x}−l+1, 0 : i'_{y}−l))\) is an advantage of this approach, it has a drawback in that there is limited parallelism over the whole system. This can be minimized if surpluses \(i\) along each diagonal in the triangular area at a distance \(\delta(i) = i'_{x} + i'_{y} + i_{x} + i_{y}\) apart are scheduled in parallel, as they share no common component grids. For example, in Figure[7] along the diagonal beginning with \((8, 5)\), \(\delta(i) = 2\), and surpluses marked ‘1’ may run in parallel, as may those marked ‘2’. Similarly, for the surpluses coalesced in one dimension, if these are scheduled from biggest to smallest, i.e. \((0 : i'_{x}−l, i'_{y}−l), (i'_{x}−1, 0 : i'_{y}−l), (0 : i'_{x}−l, i'_{y}−2), (i'_{x}−2, 0 : i'_{y}−l),...\), then the first \([l/2]−1\) of these pairs operate on disjoint component grids and hence will run in parallel. This is illustrated in in Figure[7] with the coalesced surpluses marked ‘3’. It should be noted that (coalesced) surpluses need to have the same size if they are to be effectively scheduled together.

For 3D, the generalization involves the combination of \((l − 1)(l − 2)(l − 3)/3, 3(l − 1)(l − 2)/2, 3!\) and 1 groups coalesced in 0, 1, 2, and 3 (respectively) dimensions. The scheduling distance \(\delta(i)\) increases by 1. This, combined with the fact that the average diagonal is shorter, means that the average scope of parallelization of coalesced surpluses markedly diminishes from the 2D case.

3) **Extension of Direct Algorithm to Support Surpluses:**

We now consider how the SGCT can be applied to the hierarchical surplus \(h \in \mathbb{N}^{d}\), for which the boolean vector \(\vec{c} \in \{0, 1\}^{d}\) indicates that it is not coalesced with all lower indices in each dimension.

Let the current process hold part of grid \(G_{i}\) of size \(N = 2^{i} + 1\), with an associated process grid \(P\), and this process has id \(p\) on this grid. If \(h \leq i\) does not hold, surplus \(h\) is not in \(G_{i}\) and the process participates no further. Otherwise, the process proceeds with the corresponding grid index for the SGCT given by \(h' = h − \vec{c}\).

The (local portion of the) surplus must first be extracted. The relevant parameters of are the stride \(s\), global index offset \(O\), local index offset \(o\) and the local length \(n\):

\[
\begin{align*}
  s &= 2^{i−h−\vec{c}} \\
  O &= \vec{c} * s / 2 \\
  o &= \min\{v \geq 0\}[(g_{0}(N, p, P) − O + v) % s = 0] \\
  n &= (l(N, p, P) − o + s − 1) / s + (p == P − 1) * \vec{c}
\end{align*}
\]

The last term for \(n\) represents a padding required for an uncoalesced surplus, since these have a global size which is exactly a power of 2. We can also compute \(n\) by

\[
n = l(N', p, P), \quad \text{where} \quad N' = 2^{h'} + 1
\]

The elements of the extracted surplus may be now supplied to the direct SGCT algorithm, which has \(h'(N')\) set to both the component and combined grid index (size) for the current combination.

The SGCT algorithm is also set into hierarchical mode. In this mode, computations over all the original component grids must skip any grid not containing the hierarchy \(h'\). Here, the grid ratio \(r = 1\) and we send \(dn\) instead of \(dn+1\) points (line 9 of Algorithm[1]), as interpolation and hence boundary points are not required. Finally, the combination process grid \(P'\) in this case may only involve processes participating in the current combination, which will be a subset of the processes operating on all component grids.

4) **Process Sub-griding:**

Even with the coalescing of hierarchical surpluses, a component grid \(G_{h}\) that is distributed of a process grid \(P\) can have a hierarchy \(h\) that is, in some dimensions, smaller than the process grid. In this case, the regular block distribution breaks down and the direct SGCT algorithm cannot be applied. Figure[8] illustrates this case where \(P = (8, 8)\). It can be observed however that here we still have a sparse block distribution (see Equations[9][10]), and that, if we take a sub-grid of \(P\) beginning at process offset \((0, 2)\) and stride \((2, 4)\), the hierarchy will retain a block distribution over that sub-grid. An exception occurs for the rightmost elements of the hierarchy, where the process stride is 1.
Our current implementation can accommodate 2D or 3D combinations. The main engineering issues in extending it to a higher dimension $d$ are (i) for a given level $l$, the number of sub-grids increases rapidly with $d$ (so load balance in the underlying SGCT application will become an increasing concern), (ii) enumerating the grid indices for these sub-grids across the hyperplanes in grid index space becomes increasingly complex and (iii) the sparse grid data structure becomes increasingly complex. For the hierarchical surplus algorithm, the enumeration of the coalesced surpluses subspaces (Section III-B2) also becomes increasingly complex.

As mentioned in [7], the implementation of our algorithms supports a 2D or 3D SGCT on 4D or higher-dimensional grids. A lower dimension (say of size $b$) than those used for the SGCT is supported by extending the implementation to handle blocks of $b$ elements. Higher dimensions can be dealt with by applying the SGCT iteratively. The dimensions chosen for the SGCT must be contiguous, and a local transposition can be applied in order to ensure this.

Our direct SGCT algorithm supports parallelization over arbitrary process grids ($P$) in the SGCT dimensions of a data grid. As also explained in [7], parallelization in a non-SGCT dimension can be simply performed by applying the whole SGCT in parallel over this dimension.

IV. ANALYSIS

We now will give an analysis of our SGCT algorithms under typical operating conditions imposed it with conjunction with an application [9], [7].

In order to perform a SGCT over the desired set of sub-grids $\{u_i\}$, where $u_i$ is distributed over the process grid $P$, we form the combined process grid $P'$ over a subset of these processes. That is, $\Pi(P') = 2^k$ for the largest $k$ such that $2^k \leq \Sigma_i \Pi(P_i)$. The aspect ratio of $P'$ will be set to match the grid index of the combined (full) grid. In this case $P' \geq P_i$. For simplicity (and because it is a case of strong practical interest [9], [7]), we assume each $P_i$ is a power of 2 also.

In terms of load balancing, we allocate the same number ($p \in \mathbb{N}$) of processes on each of the $P_i$ on each diagonal in the grid index space (see Figure 2); in the 3D case, the diagonal becomes a plane. The next lower diagonal receives $\lfloor p/2 \rfloor$ processes. This strategy balances the amount of data points and hence work across each processes, which approximates to a first order the load for that process.

We denote $g = g(d, l)$ to be the number of sub-grids given dimensionality $d$ and level $l$. We have $g(d, l) = O(l^{d-1})$ and more specifically:

\[ g(2, l) = 2l - 1 \quad (11) \]
\[ g(3, l) = 3l^2 - 2l + 1 \quad (12) \]

Let $m$ denote the number of data points in a sub-grid in a given process. By the above assumptions $m$ should be constant across all processes. Let $\alpha$ and $\beta$ denote the cost

| 00 30 20 30 10 30 20 30 |
| 03 33 23 33 13 33 23 33 |
| 02 32 22 32 12 32 22 32 |
| 05 35 25 35 15 35 25 35 |
| 01 31 21 31 11 31 21 31 |
| 03 33 23 33 13 33 23 33 |
| 05 35 25 35 15 35 25 35 |
| 00 30 20 30 10 30 20 30 |

Figure 8. Hierarchical surplus indices corresponding to the elements of a $l = 3$ grid, highlighting coalesced hierarchy $(0:2,2)$ on an $8 \times 8$ process grid.

Algorithm 6: Determination of sub-grid process ids

1 $q = r; \quad // \ r \ is \ the \ 1D \ rank \ of \ this \ process \ in \ P$
2 for $d = x, y, z$ do
3 \hspace{1cm} $p'_d = (q - O_d)\%P_d; \quad \delta = (p'_d == P_d - 1) * Z_d;$
4 \hspace{1cm} if $(p'_d + \delta)\%sd == 0 \quad / * \ is \ part \ of \ sub-grid */ \ then$
5 \hspace{2cm} $p'_d = (p'_d + \delta)/sd; \quad q = q/P_d;$

As noted at the end of Section III-A1, the elements on sparse grid process grid ($P'$) currently need to be powers of 2 for the direct algorithm. This can be relaxed if the halo points of $u$ have also been filled in the negative direction, and points $i - 1: i + \delta n$ are sent at line 7 of Algorithm 1 and $n + 2$ points are received on line 10 of Algorithm 2 and included in the interpolation (line 11).

The hierarchical surplus algorithm however requires all component process grids sizes ($P$) to be powers of 2. This is because the surpluses have a power of two stride (see Figure 4); thus the elements of the surpluses will only retain a block distribution (Equations (5-8)) in this case. This could be overcome at the expense of redistribution.
of starting up a message and the transmission cost per unit data.

A. Direct SGCT Algorithm

For this algorithm, in the worst case, each process in \( P \) will receive \( 2m \) words of data. Each process in each \( P_t \) sends and receives \( 2P_t/P \) messages, and each process in \( P \) receives and then sends \( g \) messages. Since the whole operation should be completely parallelized, the total cost of the direct gather-scatter is given by:

\[
t^d \leq 2gn + 3mβ \quad (13)
\]

This indicates that the algorithm should be efficient in bandwidth-constrained (large \( m \)) circumstances, which will be the case for most large-scale simulations. Otherwise, when \( g \) is large, it may be less efficient compared to an algorithm based on multistage reductions with a constant fan-in [10], with a cost of \( 2\log_2(g)(α + mβ) \).

B. Interpolation Overheads

Consider a level \( l \) SGCT on a \( d \)-dimensional isotropic grid of grid index \( i' \in \mathbb{N}, i' ≥ l \). The full grid will have \( O(2^d l) \) points. The partial sparse grid will have a fill in factor of \( f(d, l) = 2^{d-l} \) and hence will have

\[
f(d, l)2^dO(l^{d-1}) = 2^{d-l}O(l^{d-1})
\]
data points, noting a classical sparse grid (\( i' = l \)) has \( 2^l O(l^{d-1}) \) points. Each component grid will have \( 2^l (d-l)(l-1)O(l^{d-1}) \) points.

Assume the upper level grids are distributed across \( p \) processors, and partial sparse grid is distributed over the total number of processors, \( g(d, l)/p = O(l^{d-1})p \) using Equations [11][12].

Then the ratio of partial sparse grid to component grid points per process is given by:

\[
\frac{2^i2^{d-i}O(l^{d-1})(l-1)/(l-1)/p}{2^i2^{d-i}(l-1)/p} = O(1)
\]

This indicates that the interpolation stage should scale with \( p, l, \) and \( d \). However, the number of interpolation calculations per received point is \( O(l^{d-1}) \) and hence, for sufficiently large \( l \) and \( d \), interpolation time will dominate communication time.

By contrast, if a full grid is used for interpolation, the ratio of full grid to component grid points per process is

\[
\frac{2^i2^{d}O(l^{d-1})}{2^i2^{d-i}(l-1)/p} = 2^{d-l}O(l^{d-1})
\]

which is essentially exponential in \( l \).

C. Hierarchical Surplus-based Algorithm

This approach avoids communication of surpluses shared by only one grid, which reduces the total communication volume by a fraction \( 1/4 \). This is however offset by both the limited parallelization and load imbalance imposed. Furthermore, even with coalescing, the overhead of communication startups will be considerably increased over the direct method. This will however be partially offset by the average effective value of \( g \) being \( ≈30\% \) on average less. For example, the direct algorithm should should perform fairly optimally, given Equation [13] on the largest surpluses, which have a minimal \( g = 3 \).

Roughly the load imbalance factor will be \( ≈ 2^{d-1} \), due to the fact that process grids on the inner diagonal (see the highlighted areas of Figure [7]) have half the number of processes, but the sizes of the surpluses is the same across all grids. The average degree of parallelization will be \( \approx 2/3 \) for the scenario of Figure [7]. Consider the largest surpluses, labelled ‘1’ and ‘2’. Those labelled ‘2’ cannot run in parallel with those marked ‘1’. Those marked ‘3’ can run in parallel, but then only \( ≈ 2/3 \) of the processes can be utilized. As we move closer to the origin, the overall utilization increases, till it is 1 when the indices are less than \( i' - 1 + 1 = (5,5) \) in the figure. However, these areas represent only a small fraction of the overall data.

A more precise analysis needs to be done empirically. Included in the code release for this work is a Python script calculating the overall degree of parallelization for the combination of all surpluses. This generated the data for the following table, for the 2D case \( i' = (9,9) \) with coalesced surpluses:

<table>
<thead>
<tr>
<th>scheduling</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.80</td>
<td>0.82</td>
<td>0.74</td>
<td>0.75</td>
<td>0.72</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td>no</td>
<td>0.80</td>
<td>0.69</td>
<td>0.62</td>
<td>0.58</td>
<td>0.54</td>
<td>0.52</td>
<td>0.50</td>
</tr>
</tbody>
</table>

This indicates that scheduling should make a worthwhile difference, achieving a 70–80% degree of parallelization. An experiment where the grid size grows with \( l \), i.e. \( i' = (9 + l, 9 + l) \), gave essentially identical results.

Taking all these factors together, it is unclear if this approach, even with scheduling, has any overall inherent advantage in speed even for \( d = 2 \).

V. Experimental Results

In this section, we give performance results for our algorithms, under the operating conditions mentioned in Section IV.

All experiments were conducted on the Raijin cluster located at the Australian National University. It has 3,592 compute nodes, consisting of dual 8-core Intel Xeon (Sandy Bridge 2.6 GHz) processors (i.e. 16 cores) with on average 44 GB memory per node, and with an Infiniband FDR interconnect (56Gbs). Version 1.4.3 of Open MPI was used
and our codes were compiled with gcc 4.6.4 with the -O3 option.

A. Direct SGCT Performance

In order to demonstrate the effect of our parallel SGCT algorithm on an application, we use an advection solver capable of solving 2D and 3D problems. As the amount of work per timestep for the advection problem is relatively small compared to most real applications, any inefficiencies in the SGCT will be easily seen.

Figure [9] shows strong scaling results. The plots BSend indicate an MPI_BSend based implementation was used: this imposes a fixed order of message receipt. The plots ISend indicate that an MPI_ISend/IRecv implementation was used, which does not have this limitation and allows the overlap of communication with the interpolation of the gather stage. For the 3D problem, results for small core counts are missing for the former, as the MPI implementation was not able to handle the very large BSend buffers that were required.

The results indicate that the application scales even with the SGCT applied relatively frequently, with the ISend implementation lending \( \approx 20\% \) extra performance to the overall execution time.

Figure [10] indicates the weak scaling performance of the direct 2D SGCT in isolation using the partial sparse grid data structure (the default). We chose \( 2^{14} \) points per process as this represents a reasonable amount of data, and is sufficiently large so that the bandwidth components of the SGCT (see Equation [13]) becomes significant. It indicates that once communication is predominantly inter-node, the execution time remains relatively flat, indicating good scaling performance to the limit of our tests at around 3000 cores.

It can be noted that execution time increases slowly with increasing \( l \) when the sparse grid data structure is used (Figure [10](b)). At \( l = 3 \) and \( l = 11 \), the total core counts are an exact power of two; this is the best case for the algorithm, as all cores can hold part of the partial sparse grid. We have some performance loss at \( l = 5, l = 7 \) and \( l = 9 \) where only 80%, 76% and 88% of the cores hold part of the sparse grid. There is approximately a factor of 2 between the performance at \( l = 3 \) and \( l = 11 \) cases. Much of this difference is due to the interpolation time (‘in’), whose proportion of overall execution time increases with \( l \). This is largely due to load imbalance effects: the sparser the (partial) sparse grid becomes, the less evenly the points divide over a regular process grid. As can be seen from Figure [11] the processes at the corners may have significantly more points than their neighbours towards the interior.

Figure [10](a) shows the results if the much simpler full grid data structure is used for interpolation. This introduces a factor of \( \approx 2^{l-1} \) redundant data points. As expected, we get an exponential increase in execution time with respect to \( l \), with interpolation time quickly accounting for almost all execution time. It indicates that the sparse grid data structure is essential for good scaling with \( l \).

While interpolation appears to account for much of the total time, it should be noted that message receipt is overlapped with it, hence message receipt time is attributed to interpolation time. Results corresponding to Figure [10](a) for the BSend implementation indicate that interpolation never consumes \( > 50\% \) of execution time. They also show poorer scaling with core count than the ISend results. Other experiments indicate the partial sparse grid interpolation function achieves near-ideal thread-level parallelism.

Figure [11] gives a comparison of the 2D and 3D algorithms using the sparse grid data structure. In terms of the number of component grids, \( l = 3, 5, 7 \) in 3D is equivalent to the cases \( l = 5, 17, 31 \) in 2D. We can see that 2D and 3D performance is very similar for \( l = 3 \) (3D) and \( l = 5 \) (2D), and 3D scales similarly well with increasing core count. We also see a gradual decrease in performance as the number of component grids increases becomes large, due to the increased effects of startup latency.

All data shown so far are for ‘warmed’ timings when an SGCT was performed before the timing was taken. Figure [12] shows the time for the direct SGCT without this warmup; we see a degradation in performance by a factor of approximately 20. This is caused by OpenMPI setting up new (inter-grid) connections between processes in order to perform the SGCT. While not part of the SGCT itself, this overhead needs to be taken into account by any application using the SGCT. The results of our previous work [13] do not take this effect into account.

B. Hierarchical SGCT Performance

Figure [12] indicates the performance of the hierarchical surplus based algorithms. A comparison of sub-figures (a) to (b) and (c) to (d) indicates the clear benefit of coalescing the surpluses. Comparing sub-figures(a) and (c) to Figure [11] we see that this version is slightly slower for smaller \( l \) and core counts. We also see that it scales less well with \( l \), due to the fact there are \( O(l^2) \) sub-operations and the number of cores (especially as \( g(l) \) increases). Somewhat surprisingly, scheduling of the surpluses gave no clear overall benefit.

The forming and restoring from the hierarchical surpluses formed a small factor (\( \approx 1/10 \)) of this. As expected, the results vary little with \( l \) for 2D, but show some increase with \( l \) for 3D.

It should also be noted that there are no other SGCT codes available that can be used to perform the above experiments for comparison.

C. Load Balance and Communication Profiles

Figure [13] gives IPM profiles [14] for the 2D advection application with a \( l = 11 \) SGCT. We can clearly see the structure of the advection sub-problems. MPI_Send and
<table>
<thead>
<tr>
<th>number of cores</th>
<th>overall execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>88</td>
</tr>
<tr>
<td>176</td>
<td>352</td>
</tr>
<tr>
<td>704</td>
<td>1408</td>
</tr>
</tbody>
</table>

(a) 2D problem with level $l = 4$ and $2^{13} \times 2^{13}$ grid points.

(b) 3D problem with level $l = 3$ and $2^9 \times 2^9 \times 2^8$ grid points.

Figure 9. Overall execution time of the advection application with the direct SGCT running over 1024 timesteps (MPI warm-up time excluded). Results shown are an average of two experiments.

<table>
<thead>
<tr>
<th>number of cores</th>
<th>overall execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>62</td>
</tr>
<tr>
<td>124</td>
<td>248</td>
</tr>
<tr>
<td>496</td>
<td>992</td>
</tr>
<tr>
<td>1984</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of cores</th>
<th>overall execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>1024</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of cores</th>
<th>overall execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td>1024</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of cores</th>
<th>overall execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>400</td>
<td>800</td>
</tr>
<tr>
<td>1024</td>
<td></td>
</tr>
</tbody>
</table>

D. Effect of Process Grid Aspect Ratio and Process Layout on Performance

The aspect ratios of each component grid and its process grid determines the shape of the decomposed domain. Changing the decomposed domain shape and changing how processes are aid out across nodes changes the communication pattern for both the computation of application on the grid and the SGCT. Each application has its own communication requirements, and so process grid aspect ratio and layout may affect them differently. We analyze the effect for the 2D advection application with different process layouts.

The default SGCT process grid layout is made to match the shape of the component grid $G_i$. Allocating 64 processes to the largest component grids, the default process layout for 2D $G_i$ with $i = (3, 0), (2, 1), (1, 2), (0, 3), (2, 0), (1, 1), \text{and } (0, 2)$ are $16 \times 4$, $8 \times 8$, $8 \times 8$, $4 \times 16$, $8 \times 4$, $8 \times 4$, and $4 \times 8$, respectively. That of the 2D full grid is $16 \times 16$.

How process are mapped to nodes also affects communication. The process-to-node mapping of all the previous benchmarks were done by mapping the processes to cores in nodes linearly (default), as shown in Figure 14a for a 2D process grid. This will minimize inter-node communication horizontally at the expense of inter-node communication in the vertical dimension. A block-wise mapping, such is as
shown in Figure 14b will balance inter-node communication in both dimensions.

The effect of process aspect ratio and layout for a 2D advection application is presented in Tables I and II in isolation. It is observed from Table I that 8 \times 8, 8 \times 4 process grid with block map shows a small advantage in advection computation time. Applying the same configuration also improves the combination time slightly as shown in Table II.

VI. RELATED WORK

Early work in the parallelization of the SGCT for Laplace’s equation and the 3D Navier-Stokes system was reported in [2] and [15], respectively. However, the algorithm assumed each node had multiple sub-grids and concentrated on load-balancing aspects, whereas our algorithm is designed for the many processes per sub-grid case.

More recently, work on a parallel SGCT algorithm has also involved allocating multiple grids per process group of size limited to the number of cores in a socket, with OpenMP parallelism applied within each process group [8]. The algorithm was targeted towards higher dimensional problems and the concern was in load balancing within subgroups of \( O(1000) \) grids.

Work on load-balancing of component grid instances for the SGCT and a parallel SGCT algorithm using both the direct and hierarchical surplus implementations has been reported in [16] and [10], respectively. The latter paper also gives an analysis of their algorithms’ communication overheads. However, it is assumed that there is only one process per component sub-grid (i.e. \( P_l = 1 \)), severely limiting its applicability to large-scale simulations. Their algorithm also does not support the truncated combination formula. While in both cases, the (direct) algorithm is based on a gather-scatter approach at the top level, our algorithm has only two stages, whereas theirs has \( O(\log_2 l) \) stages. Furthermore, their version of the direct algorithm involves a significant amount of redundant communications, resulting in their conclusion that the hierarchical surplus approach is \( 10 \times \) (3.5 \times) faster for a high SGCT level (\( l \)) for 2D (3D) combinations. For our algorithms, we reach the opposite conclusion. The paper mentions the idea of “merging” (coalescing?) hierarchical surpluses only in the context of future work.

Efficient data structures for sparse grids usually rely on tree or hash methods [17], [18]. More recently, a more compact representation involving multi-level indexing of a flat array, ranged in the order of the hierarchical surpluses [19]. It also offers faster indexing. Our sparse grid data structure based on sparse matrix ideas shares these advantages and moreover has potentially better data locality, as the flat array is based on the spatial layout of the original component grid. Furthermore our implementation is more general than any, as it is fully distributed and covers partial sparse grids as well.

[9] used our 2D direct SGCT implementation to build a fault-tolerant 2D advection solver. As its emphasis was on fault-tolerance it gave no details of the SGCT algorithm nor did it examine its performance in isolation. Similarly for [12], which studied the computational complexity of the parallel SGCT and described software engineering techniques to reduce the complexity and ensure the reliability of the 2D direct SGCT implementation. Indeed, these techniques proved even more crucial to the more complex 3D and hierarchical surplus based implementation of the algorithms presented here. However, these papers give no details about the parallel SGCT algorithm itself.

This paper is an extended version of our preliminary work [13]. The algorithms have been perfected, with details for 3D
Figure 12. Execution time of the average of 10 repetitions of the hierarchical SGCT in isolation (MPI warm-up time excluded). The workload of each core is $2^{14}$ grid points. 'l', 'gs', and 'hc' denote the level of the SGCT, combination time, and the time to form and restore from the hierarchical surpluses, respectively. As the scheduling of hierarchical surpluses does not offer significant improvement, those are not included in the graph. ISend-based implementations are used. Results shown are an average of two experiments.

(a) 2D problem with coalescing and no load balancing on coalesced hierarchies.
(b) 2D problem with no coalescing and no load balancing on coalesced hierarchies.
(c) 3D problem with coalescing and no load balancing on coalesced hierarchies.
(d) 3D problem with no coalescing and no load balancing on coalesced hierarchies.

Figure 13. Analysis of IPM generated load balance of 2D direct SGCT solving advection on level 11 with a single combination (MPI warm-up time excluded). Core count on each upper layer grid is 16. Workload per core is $2^{14}$ points and number of timesteps is $2^{14}$. The total execution time is 4.06 sec, combination time is 0.0058s.

(a) computing whole application: MPI task
(b) performing combination: MPI task
added, and we have introduced the sparse block distribution, the partial sparse grid data structure and the process subgridding technique. The analysis is more comprehensive and more precise. The experimental results are all significantly improved and extended results have been included.

VII. CONCLUSIONS

In this paper, we have presented two general SGCT combination algorithms both supporting parallelism within and across process grids and the more general SGCT ‘truncated combination’ formula. They can also support fault-tolerant SGCT combination formulae. The algorithms are inherently complex but can be formulated (and implemented) tractably when expressed in terms of d-dimensional vector arithmetic and the associated mappings for a block distribution. The direct algorithm is highly efficient in bandwidth-limited scenarios and is general in terms of the process grid configurations of the component grids. It is particularly suited to SGCT where the number of component grids is small to moderate, i.e. $d \leq 5$. It does however need a partial sparse grid data structure, a generalization of the sparse grid, to be scalable with the SGCT level $l$ and the dimensionality $d$. Our distributed implementation of this data structure, based on sparse matrix concepts, is comparably efficient to existing non-distributed implementations, in terms of both space and access time, and scales in the same sense as the SGCT.

The hierarchical surplus based algorithm uses the direct algorithm for combining the individual and coalesced hierarchical surpluses. It thus inherits some of the performance characteristics of the direct algorithm. It offers a bandwidth reduction by a factor of $1 - 2^{-d}$; however, this seems to be offset by an incomplete parallelization factor $\geq 2/3$ for 2D (and even this requires careful scheduling of the computation), a factor of $\approx 2^{d-1}$ load imbalance (arising from the uneven number of processes across the component grids) and overall increased startup times. The latter is reduced by our technique of coalescing the surpluses for combination. It appears to have an advantage in memory consumption, with the extra memory for the SGCT being reduced to the size of the largest common surpluses, which is a factor of $2^{-d}$ of the largest component grids. However, the direct algorithm can match this by partitioning the SGCT

![Figure 14](image-url) 2D linear and block mapping of $32 \times 4$ process grid onto cores of Raijin nodes. Processes with the same color are mapped onto the cores of the same node. Different color represents mapping onto different nodes.

Table I

<table>
<thead>
<tr>
<th>combination</th>
<th>SGCT default</th>
<th>$8 \times 8, 8 \times 4$</th>
<th>block map</th>
<th>$8 \times 8, 8 \times 4$ and block map</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>4.34</td>
<td>4.35</td>
<td>4.32</td>
<td>4.28</td>
</tr>
<tr>
<td>maximum</td>
<td>4.49</td>
<td>4.55</td>
<td>4.44</td>
<td>4.34</td>
</tr>
</tbody>
</table>

Table II

<table>
<thead>
<tr>
<th>combination</th>
<th>SGCT default</th>
<th>$8 \times 8, 8 \times 4$</th>
<th>block map</th>
<th>$8 \times 8, 8 \times 4$ and block map</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>2.04E-02</td>
<td>2.03E-02</td>
<td>2.02E-02</td>
<td>2.02E-02</td>
</tr>
<tr>
<td>maximum</td>
<td>2.20E-02</td>
<td>2.23E-02</td>
<td>2.18E-02</td>
<td>2.17E-02</td>
</tr>
</tbody>
</table>
into a similar number of stages.

Our algorithm for hierarchical surplus formation is much less communication-intensive. It hoes however show weaker scalability than the4 SGCT over increasing core counts.

Experimental results with 2D and 3D advection solvers indicate that, even with a relatively frequent application of the direct SGCT algorithm, the application will scale to at least 1500 cores. The MPI ISend based implementation, imposing no order on message receipt, gave worthwhile benefits of up to 15% for the whole application. A weak scalability analysis shows that the execution time remains relatively flat, indicating good scaling performance to limit of our tests at around 3000 cores. The hierarchical algorithm requires coalescing of the surpluses for satisfactory performance, and, surprisingly, scheduling the computation creates no appreciable benefit. We find that increasing the SGCT level increased execution time only by a factor of 2 from $l = 3$ to $l = 11$ for the direct algorithm. The hierarchical algorithm scaled slightly less well in this respect, due to increase startup costs.

Future work will can still improve the direct algorithm by overcoming the limitation of a power of two process grid for the combined grid. This can provide a potential increase in performance by a factor of up to 2.


ACKNOWLEDGMENTS

The authors thank Markus Hegland and Jay Larson for valuable advice and support. We also thank the NCI National Facility for the use of the Raijin cluster. This research was supported under the Australian Research Council’s Linkage Projects funding scheme (project number LP110200410).

REFERENCES


