

Asynchronous One-Level and Two-Level Domain Decomposition Solvers

Christian Glusa*, Paritosh Raman^{†‡}, Erik G. Boman*, Edmond Chow[†] and Sivasankaran Rajamanickam*

*Center for Computing Research,

Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

E mail: {caglusa, egboman, srjama}@sandia.gov

[†]School of Computational Science and Engineering, College of Computing,
Georgia Institute of Technology, Atlanta, Georgia, USA

Email: paritoshpr@gatech.edu, echow@cc.gatech.edu

[‡]School of Industrial and Systems Engineering, College of Engineering,
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract—Parallel implementations of linear iterative solvers generally alternate between phases of data exchange and phases of local computation. Increasingly large problem sizes on more heterogeneous systems make load balancing and network layout very challenging tasks. In particular, global communication patterns such as inner products become increasingly limiting at scale.

We explore the use of asynchronous communication based on one-sided MPI primitives in a multitude of domain decomposition solvers. In particular, a scalable asynchronous two-level method is presented. We discuss practical issues encountered in the development of a scalable solver and show experimental results obtained on state-of-the-art supercomputer systems that illustrate the benefits of asynchronous solvers in load balanced as well as load imbalanced scenarios.

I. INTRODUCTION

Multilevel methods such as multigrid and domain decomposition are among the most efficient and scalable solvers developed to date. Adapting them to the next generation of supercomputers and improving their performance and scalability is crucial in the push towards exascale. Domain decomposition methods subdivide the global problem into subdomains, and then alternate between local solves and boundary data exchange. This puts a significant stress on the network interconnect, since all processes try to communicate at once. On the other hand, during the solve phase, the network is underutilized. The use of non-blocking communication can only alleviate this issue, but not solve it. In asynchronous methods, on the other hand, computation and communication occur at the same time, with some processes performing computation while others communicate, so that the network is consistently in use.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract de-na0003525.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND NO. SAND2018-9402 O

Unfortunately, the term “asynchronous” can have several different meanings in the literature. In computer science, it is sometimes used to describe communication patterns that are non-blocking, such that computation and communication can be overlapped. Iterative algorithms that use such “asynchronous” communication typically still yield the same iterates (results), just more efficiently. In applied mathematics, on the other hand, “asynchronous” denotes parallel algorithms where each process (processor) proceeds at its own speed without synchronization. Thus, asynchronous algorithms go beyond the widely used bulk-synchronous parallel (BSP) model. More importantly, they are mathematically different than synchronous methods and generate different iterates. The earliest work in this area was called “chaotic relaxation” [1]. Both approaches are expected to play an important role on future supercomputers. In this paper, we focus on the mathematically asynchronous methods, and we will use the terms “asynchronous” and “synchronous” to distinguish between mathematically asynchronous and mathematically synchronous methods.

Domain decomposition solvers [2]–[4] are often used as preconditioners in Krylov subspace iterations. Unfortunately, the computation of inner products and norms widely used in Krylov methods requires global communication. Global communication primitives, such as `MPI_Reduce`, asymptotically scale as the logarithm of the number of processes involved. This can become a limiting factor when very large process counts are used. The underlying domain decomposition method, however, can do away with globally synchronous communication, assuming the coarse problem in multilevel methods can be solved in a parallel way. Therefore, we will focus on using domain decomposition methods purely as iterative methods in the present work. We will note, however, that the discussed algorithms could be coupled with existing pipelined methods [5] which alleviate the global synchronization requirement of Krylov solvers.

Another issue that is crucial to good scaling behavior is load imbalance. Load imbalance might occur due to heterogeneous hardware in the system, or due to local, problem specific causes, such as iteration counts for local sub-solves that vary

from region to region. Especially the latter are difficult to predict, so that load balancing cannot occur before the actual solve. Therefore, a synchronous parallel application has to be idle until its slowest process has finished. In an asynchronous method, local computation can continue, and improve the quality of the global solution.

An added benefit of asynchronous methods is that, since the interdependence of one subdomain on the others has been weakened, fault tolerance [6], [7] can be more easily achieved. When one process has to be stopped, be it for a hard or a soft fault, it can be replaced without having to halt every other process.

The main drawback of asynchronous iterations is the fact that deterministic behavior is sacrificed. Consecutive runs do not produce the same result. (But they obviously are at most a distance proportional to the convergence tolerance apart from each other.) This also makes the mathematical analysis of asynchronous methods significantly more difficult than the one of its synchronous counterparts. Analytical frameworks for asynchronous linear iterations have long been available [1], [8]–[10], but generally cannot produce sharp convergence bounds except for in the simplest of cases.

The main contributions of our work are:

- A novel asynchronous two-level domain decomposition method, scalable to thousands of processors.
- Empirical comparisons of synchronous and asynchronous methods on three domain decomposition solvers on a state-of-the-art parallel computer.
- An empirical study of one-sided MPI performance in a scientific computing setting.

Our work demonstrates that asynchronous methods have the potential of outperforming conventional synchronous solvers and offer a viable alternative in the push towards exascale.

The present work is structured as follows: In Section II, we present three overlapping domain decomposition methods, and explain their use in synchronous and asynchronous fashion. For a general introduction to domain decomposition methods we refer the reader to [2]–[4]. Section III is dedicated to a description of the presently available mechanisms in MPI and hardware to achieve truly asynchronous communication. Numerical experiments using the presented methods are given in Section IV, where we compare the strong and weak scaling behavior of synchronous and asynchronous solvers with and without load imbalance.

A. Related work

A one-level domain decomposition solver with optimized artificial boundary conditions was proposed in [11]. An optimization package that leverages asynchronous coordinate updated is presented in [12]. Synchronization reducing Krylov methods have a long history [13]. However, preconditioning such methods is unresolved apart from some simple preconditioners [14]. Recent work extends the preconditioners to one level domain decomposition preconditioning [15]. Pipelined Krylov methods [5] reduce synchronization costs in addition

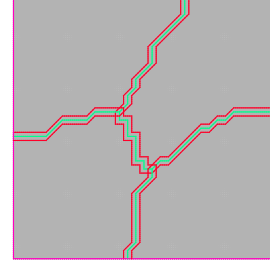


Fig. 1. Partitioning of a uniform triangular mesh of the unit square into 4 overlapping subdomains. The non-overlapping partitioning produced using METIS [16] is shown in green, the extended overlapping subdomains in red.

to overlap computation and communication, and can be used with any preconditioner.

II. DOMAIN DECOMPOSITION METHODS

A. One-level Restricted Additive Schwarz (RAS)

We want to solve the global system

$$\mathbf{A}\vec{u} = \vec{f},$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$. Informally speaking, one-level domain decomposition solvers break up the global system of equations into overlapping sub-problems that cover the whole global system. The iteration then alternates between computation of the global residual, which involves communication, and local solves for solution corrections. Special attention needs to be paid to the unknowns in the overlap, in order to avoid over-correction. Below, we describe the different methods considered in this work in detail in order to understand the what data is required to be exchanged and how the methods can be executed in asynchronous fashion.

Based on the graph of \mathbf{A} or geometric information for the underlying problem the unknowns are partitioned into P overlapping sets \mathcal{N}_p of size $N_p, p = 1, \dots, P$. An example of such a partitioning is given in Figure 1.

The notation throughout this section is based on [2]. We call the restriction to the p -th set $\mathbf{R}_p \in \mathbb{R}^{N_p \times N}$. The entries of the matrices \mathbf{R}_p are all either one or zero, with exactly one entry per row and at most one entry per column being non-zero. The local parts of \mathbf{A} are given by

$$\mathbf{A}_p = \mathbf{R}_p \mathbf{A} \mathbf{R}_p^T \in \mathbb{R}^{N_p \times N_p}.$$

Furthermore, we require a partition of unity, represented by diagonal weight matrices \mathbf{D}_p , such that the discrete partition of unity property holds

$$\mathbf{I} = \sum_{p=1}^P \mathbf{R}_p^T \mathbf{D}_p \mathbf{R}_p. \quad (1)$$

In what follows, we will assume that \mathbf{D}_p are Boolean, i.e. their entries are either zero or one. This means that every (potentially shared) unknown has a special attachment with exactly one subdomain. Consequently,

$$\mathbf{D}_p \mathbf{R}_p^T \mathbf{R}_q \mathbf{D}_q = \mathbf{0} \quad \text{for } p \neq q \quad (2)$$

and

$$D_p R_p R_p^T D_p = D_p. \quad (3)$$

We will furthermore require that $(D_p)_{jj} = 0$ for all $j \in \mathcal{N}_p$ such that there is a $k \in \mathcal{N}_p^c$ with $A_{jk} \neq 0$. (I.e. j is on the boundary of partition p .) An importance consequence is the following identity:

$$R_p A R_q^T D_q = R_p R_q^T R_q A R_q^T D_q. \quad (4)$$

This holds, since for any $\vec{u}_q \in \mathbb{R}^{N_q}$, $D_q \vec{u}_q$ is supported on the interior unknowns, and hence $A R_q^T D_q \vec{u}_q$ is supported in \mathcal{N}_q . But on \mathcal{N}_q , $R_q R_q^T$ acts as identity.

A stationary preconditioned iterative method based on the splitting $A = M - N$ is given globally as

$$\vec{u}^{n+1} = \vec{u}^n + M^{-1} (\vec{f} - A \vec{u}^n),$$

where M^{-1} is a preconditioner for A .

This means that we need to calculate the residual $\vec{r}^n = \vec{f} - A \vec{u}^n$. Its local part on node p is given by

$$\begin{aligned} R_p \vec{r}^n &= R_p \vec{f} - R_p A \vec{u}^n \\ &= R_p \left(\sum_{q=1}^P R_q^T D_q R_q \right) \vec{f} - R_p A \left(\sum_{q=1}^P R_q^T D_q R_q \right) \vec{u}^n \\ &= \sum_{q=1}^P R_p R_q^T D_q R_q \vec{f} - \sum_{q=1}^P R_p R_q^T A_q D_q R_q \vec{u}^n \\ &= \sum_{q=1}^P R_p R_q^T (D_q R_q \vec{f} - A_q D_q R_q \vec{u}^n), \end{aligned}$$

where we used (1) and (4). This means that in order to obtain the local part of the global residual, we first compute locally $D_p R_p \vec{f} - A_p D_p R_p \vec{u}^n$ on every node p , and then communicate and accumulate all the values in the overlap. The latter operation is represented by the operator $\sum_{q=1}^P R_p R_q^T$.

The *restricted additive Schwarz (RAS) preconditioner* [17], [18] is given by

$$M_{RAS}^{-1} = \sum_{p=1}^P R_p^T D_p A_p^{-1} R_p.$$

It is widely used, and is the default option for overlapping domain decomposition preconditioners in PETSc [19]. It can be thought of as a variant of the additive Schwarz preconditioner

$$M_{AS}^{-1} = \sum_{p=1}^P R_p^T A_p^{-1} R_p$$

that is convergent as an iterative method, since the damping by D_p in the overlapping parts avoids over-correction. Note that for a natural choice of D , the number of communication steps is cut in half as there is no communication associated with $R_p^T D_p$.

Now, the local part of the RAS iteration is given by

$$\begin{aligned} R_p \vec{u}^{n+1} &= R_p \vec{u}^n + R_p M_{RAS}^{-1} \vec{r}^n \\ &= R_p \vec{u}^n + \sum_{q=1}^P R_p R_q^T D_q A_q^{-1} R_q \vec{r}^n. \end{aligned}$$

If we set $\vec{u}_p^n = R_p \vec{u}^n$ and $\vec{r}_p^n = R_p \vec{r}^n$ the local parts of solution and residual respectively, the RAS iteration reads as

$$\begin{aligned} \vec{r}_p^n &= \sum_{q=1}^P R_p R_q^T (D_q R_q \vec{f} - A_q D_q \vec{u}_q^n), \\ \vec{u}_p^{n+1} &= \vec{u}_p^n + \sum_{q=1}^P R_p R_q^T D_q A_q^{-1} \vec{r}_q^n. \end{aligned}$$

This seems to suggest that the update step requires neighborhood communication as well. But in fact, in the next iteration, computation of the residual only requires $D_p \vec{u}_p^{n+1}$. From (2), (3), we see that the iterative scheme without the communication step in the update

$$\vec{r}_p^n = \sum_{q=1}^P R_p R_q^T (D_q R_q \vec{f} - A_q D_q \vec{w}_q^n), \quad (5)$$

$$\vec{w}_p^{n+1} = \vec{w}_p^n + A_p^{-1} \vec{r}_p^n \quad (6)$$

is equivalent because $D_p \vec{u}_p^n = D_p \vec{w}_p^n$ for all n . The solution \vec{u}_p^n can be recovered from \vec{w}_p^n in the post-processing step

$$\vec{u}_p^n = R_p \vec{u}^n = \sum_{q=1}^P R_p R_q^T D_q R_q \vec{u}^n = \sum_{q=1}^P R_p R_q^T D_q \vec{w}_q^n.$$

Finally, we use the norm of the residual in the stopping criterion. The norm can be computed from local quantities as

$$\begin{aligned} \|\vec{r}^n\|^2 &= \vec{r}^n \cdot \vec{r}^n = \vec{r}^n \cdot \left(\sum_{p=1}^P R_p^T D_p R_p \vec{r}^n \right) \\ &= \sum_{p=1}^P (R_p \vec{r}^n) \cdot (D_p R_p \vec{r}^n) \\ &= \sum_{p=1}^P \vec{r}_p^n \cdot (D_p \vec{r}_p^n). \end{aligned}$$

In conclusion, we can give the local form of RAS as in Figure 2, where we have dropped the superscript n for the iteration number. In fact, Figure 2 describes both the synchronous and the asynchronous version of RAS. In the synchronous version Line 4 is executed in lock step by all subdomains using non-blocking two-sided communication primitives. This communication step could be overlapped by computation. However, in established frameworks such as Trilinos, such overlapping requires major changes to the framework¹. PetSc allows some overlap of computation and communication with two phase assembly [19]. It is possible to modify such established libraries for the asynchronous iterations that are focus of

¹<https://github.com/trilinos/Trilinos/issues/767>

```

1:  $\vec{w}_p \leftarrow \vec{0}$ 
2: while not converged do
3:   Local residual:  $\vec{t}_p \leftarrow \mathbf{D}_p \mathbf{R}_p \vec{f} - \mathbf{A}_p \mathbf{D}_p \vec{w}_p$ 
4:   Accumulate:  $\vec{r}_p \leftarrow \sum_{q=1}^P \mathbf{R}_p \mathbf{R}_q^T \vec{t}_q$ 
5:   Solve:  $\mathbf{A}_p \vec{v}_p = \vec{r}_p$ 
6:   Update:  $\vec{w}_p \leftarrow \vec{w}_p + \vec{v}_p$ 
7: end while
8: Post-process:  $\vec{u}_p \leftarrow \sum_{q=1}^P \mathbf{R}_p \mathbf{R}_q^T \mathbf{D}_q \vec{w}_q$ 

```

Fig. 2. Restricted additive Schwarz (RAS) in local form

this paper. However, in order to keep the focus on algorithmic development, we developed a library that supports the one-sided communication primitives, simple mesh generation and discretization options and build the new solvers using the communication primitives. In the asynchronous variant, each subdomain exposes a memory region to remote access. On execution of Line 4, the relevant components of current local residual vector $\vec{t}_p = \mathbf{D}_p \mathbf{R}_p \vec{f} - \mathbf{A}_p \mathbf{D}_p \vec{w}_p$ are written to the neighboring subdomains, and the latest locally available data \vec{t}_q from every neighbor q is used.

B. One-level Jacobi-Schwarz (JS)

A simple modification of RAS results in the *Jacobi-Schwarz iteration*. We replace the single discrete partition of unity (1) with P different ones which, for $p, q \in \{1, \dots, P\}$, $q \neq p$, are given by the Boolean matrices

$$\begin{aligned} \left(\mathbf{D}_p^{(p)} \right)_{jj} &= 1, \quad \forall j \in \mathcal{N}_p, \\ \left(\mathbf{D}_q^{(p)} \right)_{jj} &= \begin{cases} 1 & \text{if } j \in \mathcal{N}_q^{(0)} \setminus \mathcal{N}_p, \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Here, we let $\mathcal{N}_p^{(0)}$, $p = 1, \dots, P$, be a non-overlapping partition of the unknowns such that $\mathcal{N}_p^{(0)} \subset \mathcal{N}_p$, $p = 1, \dots, P$. In practice, this can be chosen to be the partitioning of unknowns before the overlaps are constructed. Then, for each $p \in \{1, \dots, P\}$, it holds that

$$\sum_{q=1}^P \mathbf{R}_q^T \mathbf{D}_q^{(p)} \mathbf{R}_q = \mathbf{I}.$$

By replacing the partition of unity in (5), (6), we obtain the local form of the Jacobi-Schwarz iteration:

$$\begin{aligned} \vec{r}_p^n &= \sum_{q=1}^P \mathbf{R}_p \mathbf{R}_q^T \left(\mathbf{D}_q^{(p)} \mathbf{R}_q \vec{f} - \mathbf{A}_q \mathbf{D}_q^{(p)} \vec{w}_q^n \right), \\ \vec{w}_p^{n+1} &= \vec{w}_p^n + \mathbf{A}_p^{-1} \vec{r}_p^n. \end{aligned}$$

C. Two-level synchronous RAS

In order to improve the scalability of the solver, a mechanism of global information exchange is required. Let $\mathbf{R}_0 \in \mathbb{R}^{n \times n_0}$ be the restriction from the fine grid problem to a coarser mesh, and let the coarse grid matrix \mathbf{A}_0 be given by the Galerkin relation $\mathbf{A}_0 = \mathbf{R}_0 \mathbf{A} \mathbf{R}_0^T$. The coarse grid solve

```

1:  $\vec{w}_p \leftarrow \vec{0}$ 
2: while not converged do
3:   On subdomains
4:     Local residual:  $\vec{t}_p \leftarrow \mathbf{D}_p \mathbf{R}_p \vec{f} - \mathbf{A}_p \mathbf{D}_p \vec{w}_p$ 
5:     Send  $\mathbf{R}_0 \mathbf{R}_p^T \vec{t}_p$  to coarse grid
6:     Accumulate:  $\vec{r}_p \leftarrow \sum_{q=1}^P \mathbf{R}_p \mathbf{R}_q^T \vec{t}_q$ 
7:     Solve:  $\mathbf{A}_p \vec{v}_p = \vec{r}_p$ 
8:     Update:  $\vec{w}_p \leftarrow \vec{w}_p + \frac{1}{2} \vec{v}_p$ 
9:     Receive  $\vec{c}_p = \mathbf{R}_p \mathbf{R}_0^T \vec{v}_0$  from coarse grid
10:    Update:  $\vec{w}_p \leftarrow \vec{w}_p + \frac{1}{2} \vec{c}_p$ 
11:   On coarse grid
12:     Receive  $\mathbf{R}_0 \mathbf{R}_p^T \vec{t}_p$  from subdomains
13:     Accumulate  $\vec{r}_0 = \sum_{p=1}^P \mathbf{R}_0 \mathbf{R}_p^T \vec{t}_p$ 
14:     Solve  $\mathbf{A}_0 \vec{v}_0 = \vec{r}_0$ 
15:     Send  $\vec{c}_p = \mathbf{R}_p \mathbf{R}_0^T \vec{v}_0$ ,  $p = 1, \dots, P$  to subdomains
16:   end while
17:   On subdomains
18:     Post-process  $\vec{u}_p \leftarrow \sum_{q=1}^P \mathbf{R}_p \mathbf{R}_q^T \mathbf{D}_q \vec{w}_q$ 

```

Fig. 3. Synchronous RAS with additive coarse grid in local form

can be incorporated in the RAS iteration either in additive fashion:

$$\vec{u}^{n+1} = \vec{u}^n + \left(\frac{1}{2} \mathbf{M}_{RAS}^{-1} + \frac{1}{2} \mathbf{R}_0^T \mathbf{A}_0^{-1} \mathbf{R}_0 \right) \left(\vec{f} - \mathbf{A} \vec{u}^n \right), \quad (7)$$

or in multiplicative fashion:

$$\begin{aligned} \vec{u}^{n+1/2} &= \vec{u}^n + \mathbf{M}_{RAS}^{-1} \left(\vec{f} - \mathbf{A} \vec{u}^n \right), \\ \vec{u}^{n+1} &= \vec{u}^{n+1/2} + \mathbf{R}_0^T \mathbf{A}_0^{-1} \mathbf{R}_0 \left(\vec{f} - \mathbf{A} \vec{u}^{n+1/2} \right). \end{aligned}$$

In what follows, we focus on the additive version, since it lends itself to asynchronous iterations: subdomain solves and coarse-grid solves are independent of each other.

Again, we determine the local form of the global algorithm. For simplicity of exposition we do not describe the solution of the coarse grid problem itself in local form, i.e. we will simply write \mathbf{A}_0^{-1} . The local part of the coarse grid update is

$$\begin{aligned} &\frac{1}{2} \mathbf{R}_p \mathbf{R}_0^T \mathbf{A}_0^{-1} \mathbf{R}_0 \left(\vec{f} - \mathbf{A} \vec{u}^n \right) \\ &= \frac{1}{2} \left(\mathbf{R}_p \mathbf{R}_0^T \right) \mathbf{A}_0^{-1} \sum_{p=1}^P \left(\mathbf{R}_0 \mathbf{R}_p^T \right) \left(\mathbf{D}_p \mathbf{R}_p \vec{f} - \mathbf{A}_p \mathbf{R}_p \vec{u}^n \right). \end{aligned}$$

Here, the operators $\left(\mathbf{R}_0 \mathbf{R}_p^T \right)$ and $\left(\mathbf{R}_p \mathbf{R}_0^T \right)$ encode the communication from subdomain p to the coarse grid and vice versa. In conclusion, the local form of RAS with additive coarse grid is given in Figure 3. Again, we have dropped the superscript for the iteration number.

D. Two-level asynchronous RAS

From the mathematical description (7) of two-level additive RAS, one might be tempted to see the coarse-grid problem simply as an additional subdomain. From Figure 3

the fundamental differences between the subdomains and the coarse-grid problem become apparent. Subdomains determine the right-hand side for their local solve and correct it by transmitting boundary data to their neighbors. The coarse-grid, on the other hand, receives its entire right-hand side from the subdomains, and hence has to communicate with every single one of them.

In order to perform asynchronous coarse-grid solves, we therefore need to make sure that all the right-hand side data necessary for the solve has been received on the coarse grid. Moreover, corrections sent by the coarse grid should be used exactly once by the subdomains. This is achieved by not only allocating memory regions to hold the coarse grid right-hand side on the coarse grid rank and the coarse grid correction on the subdomains, but also Boolean variables that are polled to determine whether writing or reading right-hand side or solution is permitted. More precisely, writing of the local subdomain residuals to the coarse grid memory region of \vec{r}_0 is contingent upon the state of the Boolean variable `canWriteRHSp`. (See Figure 4.) When `canWriteRHSp` is True, right-hand side data is written to the coarse grid, otherwise this operation is omitted. Here, the subscripts are used to signify the MPI rank owning the accessed memory region. As before, index 0 corresponds to the coarse grid and indices $1, \dots, P$ correspond to the subdomains. To improve readability, we show access to a memory region on the calling process in blue, while remote access is printed in red.

In a similar fashion, the coarse grid checks whether every subdomain has written a right-hand side to \vec{r}_0 by polling the state of the local Boolean array `RHSisReady0`. We notice that the algorithm is asynchronous despite the data dependencies. Coarse grid and subdomain solves do not wait for each other.

Since we determined by experiments that overall performance is adversely affected if the coarse grid constantly polls the status variable `RHSisReady0`, we added a sleep statement into its work loop. The sleep interval should not be chosen too large, since this effectively results in under-usage of the coarse grid. Keeping the ratio of attempted coarse grid solves to actual performed coarse grid solves at around 1/20 has been proven effective to us. This can easily be achieved by an adaptive procedure that counts solves and solve attempts and either increases or decreases the sleep interval accordingly.

III. ONE-SIDED MESSAGE PASSING INTERFACE

In order to drive the asynchronous method on a distributed memory setup, we use a one-sided approach wherein the remote process incurs minimal overhead for servicing received messages from the sender process. The one-sided approach is achieved in MPI using the Remote Memory Access (RMA) semantics, wherein every process exposes a part of its local memory window to remote processes for read as well as write operations. However, in reality, a synchronization between the source and the target process is required for progress of the underlying application. This active synchronization step, while still preserving the asynchronous nature of the algorithm, is expensive and might erode the natural gains

```

1: while not converged do
2:   On subdomains
3:     Local residual:  $\vec{t}_p \leftarrow D_p R_p \vec{f} - A_p D_p \vec{w}_p$ 
4:     if canWriteRHSp then
5:        $\vec{r}_0 \leftarrow \vec{r}_0 + R_0 R_p^T \vec{t}_p$ 
6:       canWriteRHSp  $\leftarrow$  False
7:       RHSisReady0[p]  $\leftarrow$  True
8:     end if
9:     Accumulate asynchronously:  $\vec{r}_p \leftarrow \sum_{q=1}^P R_p R_q^T \vec{t}_q$ 
10:    Solve:  $A_p \vec{v}_p = \vec{r}_p$ 
11:    Update:  $\vec{w}_p \leftarrow \vec{w}_p + \frac{1}{2} \vec{v}_p$ 
12:    if solutionIsReadyp then
13:      Update:  $\vec{w}_p \leftarrow \vec{w}_p + \frac{1}{2} \vec{c}_p$ 
14:      solutionIsReadyp  $\leftarrow$  False
15:    end if
16:  On coarse grid
17:    if RHSisReady0[p]  $\forall p = 1, \dots, P$  then
18:      Solve  $A_0 \vec{v}_0 = \vec{r}_0$ 
19:      for  $p = 1, \dots, P$  do
20:        RHSisReady0[p]  $\leftarrow$  False
21:        canWriteRHSp  $\leftarrow$  True
22:         $\vec{c}_p \leftarrow R_p R_0^T \vec{v}_0$ 
23:        solutionIsReadyp  $\leftarrow$  True
24:      end for
25:    else
26:      Sleep
27:    end if
28:  end while
29:  On subdomains
30:    Post-process synchronously  $\vec{u}_p \leftarrow \sum_{q=1}^P R_p R_q^T D_q \vec{w}_q$ 

```

Fig. 4. Asynchronous RAS with additive coarse grid in local form. Variables printed in blue are exposed memory regions that are local to the calling process. Red variables are remote memory regions.

obtained from the asynchronous method. Therefore in order to extract the maximum gains from an asynchronous method, a passive approach is required. A passive approach entails transmission of messages which causes little to no interference to the target process. As a result, the target process does not need to yield its operating system time for servicing incoming message interrupts and therefore does not participate in the communication process. The RMA framework on MPI implements passive target synchronization with the help of two sets of primitives `MPI_Lock/MPI_Unlock` and `MPI_Lockall/MPI_Unlockall`. While, the former involves opening and closing the exposure epoch on remote nodes for each access operation, the latter only requires opening and closing of access epoch once during the application lifetime incurring less target synchronization overhead.

A. Remote Direct memory Access

RMA's passive one-sided communication can leverage a hardware mechanism known as Remote Direct Memory Access (RDMA) [20] when available. It allows RMA to directly map memory windows to the RDMA engine, allowing mes-

sages written by remote processes can be directly read by each process at its perusal. This leads to minimum disturbance to the remote process and achieving a truly passive, one-sided communication scheme.

B. Asynchronous Progress Control

RDMA is usually a hardware characteristic that may not be supported by all machines. Though we expect one-sided communication of RMA to be able to handle progress of communication in an entirely asynchronous manner, it generally fails to do so since MPI does not guarantee asynchronous progress. In such a case, asynchronous progress may be enforced by allocating certain auxiliary cores to ghost processes that solely perform the task of asynchronous progress control. As a consequence we obtain an RDMA agnostic system while simultaneously obtaining the benefits of RDMA. Even in the presence of RDMA, asynchronous progress control mechanism can be complementary since the low level RDMA engine may not be capable to handle high volumes of communication. Casper [21] and Intel Asynchronous Progress Control (APC) are two such implementations that provide ghost processes for asynchronous progress control.

IV. NUMERICAL EXPERIMENTS

A. Performance metrics

The performance of linear iterative methods is typically measured [22, Chapter 3.2.5] by the average contraction factor per iteration $\tilde{\rho} = \left(\frac{r_{\text{final}}}{r_0}\right)^{\frac{1}{K}}$, where r_0 is the norm of the initial residual vector, r_{final} the norm of the final residual vector, and K the number of iterations that were taken to decrease the residual from r_0 to r_{final} . For an asynchronous method, the number of iterations varies from subdomain to subdomain, and hence $\tilde{\rho}$ is not well-defined. The following generalization permits us to compare synchronous methods with their asynchronous counterpart $\hat{\rho} = \left(\frac{r_{\text{final}}}{r_0}\right)^{\frac{\tau_{\text{sync}}}{T}}$. Here, T is the total iteration time, and τ_{sync} is the average time for a single iteration in the synchronous case. In the synchronous case, since $T = \tau_{\text{sync}}K$, $\hat{\rho}$ recovers $\tilde{\rho}$. The approximate contraction factor $\hat{\rho}$ can be interpreted as the average contraction of the residual norm in the time of a single synchronous iteration.

B. Test problem

As a test problem, we solve

$$-\Delta u = f \quad \text{in } \Omega = [0, 1]^2, \quad u = 0 \quad \text{on } \partial\Omega,$$

where the right-hand side is $f = 2\pi^2 \sin(\pi x) \sin(\pi y)$ and the corresponding solution is $u = \sin(\pi x) \sin(\pi y)$. We discretize Ω using a uniform mesh composed of triangles and approximate the solution using piece-wise linear finite elements.

C. Convergence detection

In classical synchronous iterative methods, a stopping criterion of the form $\|\vec{r}\| < \varepsilon$ is evaluated at every iteration. Here, \vec{r} is the residual vector, ε is a prescribed tolerance (that might be chosen as a function of the discretization error), and $\|\cdot\|$ is an appropriate norm. The global quantity $\|r\|$ needs to be computed as the sum of local contributions from all the subdomains. This implies that convergence detection in asynchronous methods is not straightforward, since collective communication primitives require synchronization. In the numerical examples below, we use of the following options to terminate the iteration:

- Prescribe the total solve time.
(Examples in Section IV-F)
- Use a simplistic convergence criterion.
(Examples in Sections IV-G and IV-H)

The simplistic convergence criterion consists in writing the local contributions to a master rank, say rank 0. This master rank sums the contributions, and exposes the result through another window. Each subdomain can retrieve this approximation of the global residual norm from the master rank, and terminates if it is smaller than the prescribed tolerance. This simplistic convergence detection mechanism has several drawbacks. For one, the global residual is updated by the master rank, which might not happen frequently enough. Hence it is possible that the iteration continues despite the true global residual norm already being smaller than the tolerance. Moreover, the mechanism puts an increased load on the network connection to the master rank, since every subdomain writes to its memory region. Finally, since the local contributions to the residual norm are not necessarily monotonically decreasing, the criterion might actually detect convergence when the true global residual is not yet smaller than the tolerance. The delicate topic of asynchronous convergence detection has been treated in much detail in the literature, and we refer to [23], [24] for an overview of more elaborate approaches.

D. Convergence histories

In order to plot convergence histories for an asynchronous method, we employed the strategy illustrated in Figure 5. We start by discarding a fixed number of iterations, compute the average solve time per iteration and proceed in timed increments of this value and collect the global error for each step. The idea is to execute the algorithm from the beginning each time increasing the maximum run time linearly. In doing so, even though we avoid the use of termination detection scheme, this approach can become computationally intensive for larger values of r . We note however that the sole purpose of employing such a scheme is to collect snapshots of the global error or residual without corrupting the convergence behavior with the utilization of a termination detection scheme. In practice however, the objective is to directly reach the final tolerance goal without the requirement of recording the global state.

```

1: Initialize system
2: Execute asynchronous method for 5 iterations
3:  $\alpha \leftarrow$  Time taken per iteration during first 5 iterations
4: for  $r = 1, 2, \dots$  do
5:   Reinitialize system
6:    $T \leftarrow \text{GetWallClockTime}()$ 
7:   while  $\text{GetWallClockTime}() - T < r\alpha$  do
8:     Execute asynchronous method
9:   end while
10:  record global error  $e$ 
11:  if  $e \leq \epsilon$  then
12:    exit
13:  end if
14: end for

```

Fig. 5. Stop and repeat strategy for obtaining convergence history

E. Platform and implementation details

All runs are performed on Cori at NERSC. While all of the code was written from scratch, the differences between the synchronous and the asynchronous code path are limited, since only communication layer and stopping criterion need to be changed. (E.g. compare Figures 3 and 4.) Both the Intel Knights Landing (KNL) and the Haswell partition were used. Unless otherwise stated, one MPI rank is used per core, i.e. 64 ranks per KNL node and 32 ranks per Haswell node. In the case of the two-level method, the coarse grid solve is performed on a single MPI rank. The underlying mesh is partitioned either into uniformly sized rectangular subdomains or using the METIS library [16]. In the latter case, the option to minimize the overall communication volume is used. Our solvers handle general unstructured matrices, and the structure of the mesh is not exploited. Both subdomain and coarse grid problems are factored and solved using the SuperLU [25], [26] direct solver. This choice is guided by the desire to eliminate the impact that inexact solves such as preconditioned iterative sub-solves might have on the overall convergence.

F. One-level Jacobi Schwarz

In this section we present results from experiments pertaining to the Asynchronous Jacobi Schwarz solver. Our objective is to shed light on the computational characteristics of the Asynchronous Jacobi Schwarz method and explore the different options to achieve asynchronous communication. We particularly examine the following cases:

- *Case 1:* Effects of computational behavior with respect to increase in processes per compute node.
- *Case 2:* Effects of strong scaling conducted on a regular 2D mesh.

All experiments in this section were conducted using the KNL partition on Cori. For the experiments in this section APC was active and RDMA support was also leveraged.

1) *Effects of varying number of processes per node:* In this experiment, we fix the number of KNL nodes and perform a strong scaling study to observe convergence behavior among

the two different ways of achieving asynchronous communication with RMA mentioned in Section III. Within the algorithm described in Figure 5, the iteration was terminated using the prescribed solve times. In Figure 6, we show convergence histories obtained using MPI_Lock/Unlock represented by the curve labeled LOCK_ASYNC and MPI_Lockall/Unlockall represented by the curve labeled ASYNC. Moreover, we also show results obtained using non-blocking two-sided communication. We immediately observe that LOCK_ASYNC is very inefficient compared to ASYNC owing to the opening and closing of exposure epochs at each update step to communicate with remote neighbors. We further observe that the performance of the asynchronous method improves with more processes per node. This can be attributed to more communication among subdomains assuming a shared memory characteristic as more processes become local to each node. Since it can also be observed that the LOCK_ASYNC method is not as efficient as ASYNC in achieving asynchronous communication, we omit experiments pertaining to LOCK_ASYNC for the rest of this paper and focus on comparisons of ASYNC with the synchronous method as the benchmark.

2) *Strong scaling effects for JS:* In a second experiment, we perform a strong scaling study of the Jacobi-Schwarz iteration. The results are summarized in Figures 7 and 8. Again, we show convergence histories in Figure 7. Figure 8 compares total solution times, communication times and asynchronous degree of the runs shown in Figure 7. Communication time refers to the maximum total time taken by any single process to return from its communication routines during the algorithm run. In case of synchronous communication, this would be the time incurred by two-sided MPI primitives and in case of the asynchronous this would refer to the time taken by RMA driven MPI_Puts amidst MPI_Lockall/MPI_Unlockall operations. From Figure 8b we can see that the synchronous and asynchronous method scale reasonably well, and that the asynchronous method consistently improves relative to the synchronous one as the number of processes are increased. It should be noted that perfect scaling cannot be achieved for a one-level method, since there is no global mechanism of information exchange. In order to quantify how asynchronous the iteration is, we measure the asynchronous degree which is the ratio between the minimum number of updates performed by any process versus the maximum number of updates by any process upon convergence. The more imbalance there is, the greater the asynchrony in the system leading to a lower ratio. On the other hand, a perfectly synchronous system would have an asynchronous degree of 1 since the number of updates remain same across all processes. We note that the asynchronous degree is potentially oblivious to transient imbalance which could arise from adaptive throttling of processor speeds or from noise in the system, as these imbalances could cancel out over the time of one run. From Figure 8a we observe that the asynchronous degree decreases with increase in the number of processes. This is expected as there is greater inter node communication among subdomains with decreasing sizes. From Figure 8b we observe that the maximum total

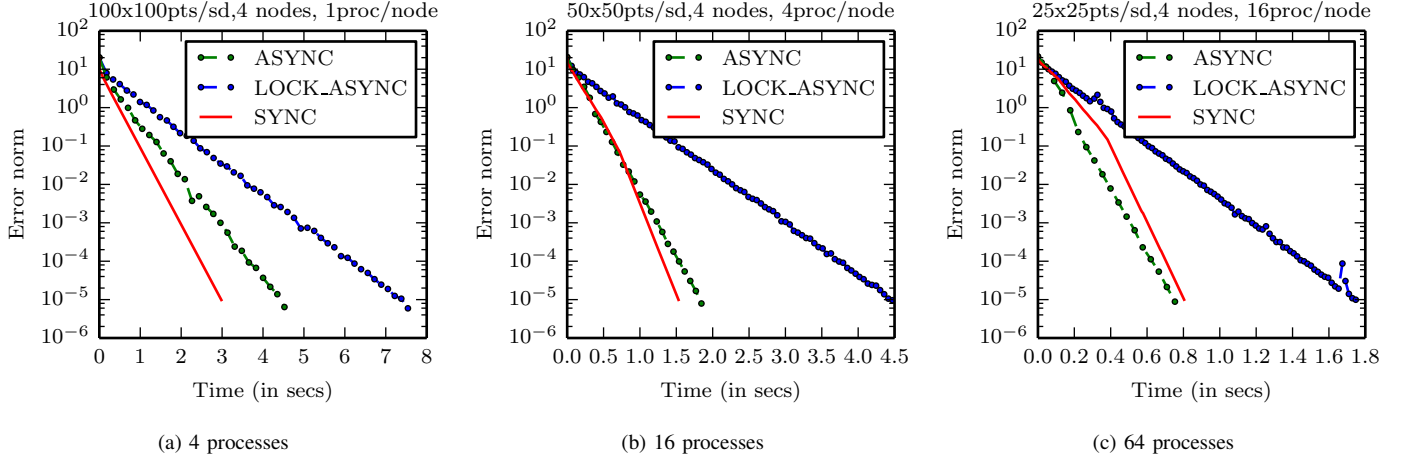


Fig. 6. Strong Scaling study of synchronous and asynchronous JS with varying number of processes per node: From left to right, the number of processes per node keeps varying while the number of nodes remains constant at 4. The global problem size is also fixed at $200 \times 200 = 40,000$ unknowns. The tolerance is set to 10^{-5} .

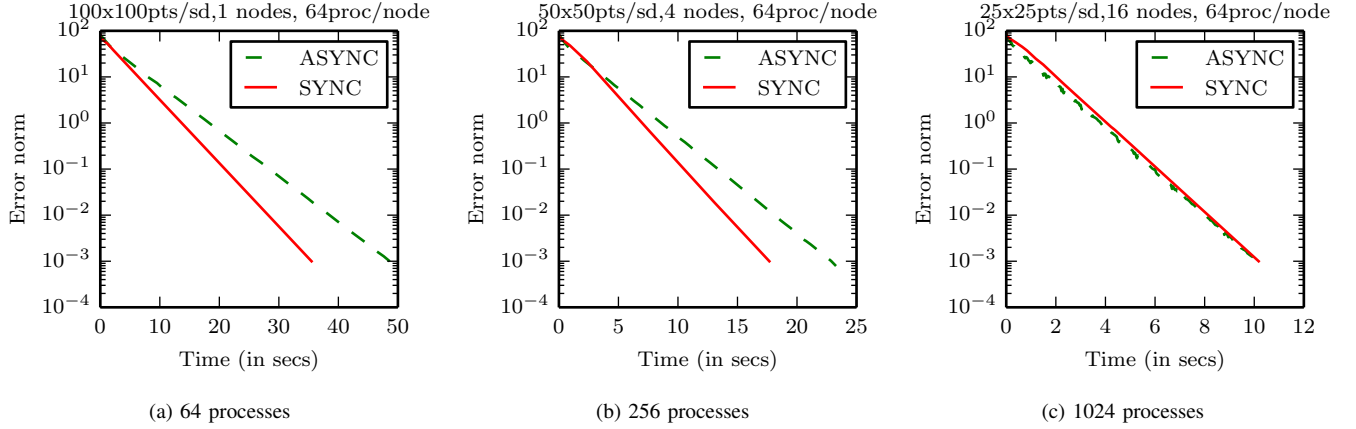


Fig. 7. Study of convergence behavior with strong scaling for synchronous and asynchronous JS: The global problem size is fixed at $800 \times 800 = 640,000$ unknowns, the number of nodes varies from 1, 4 and 16 with 64 processes per node. The tolerance is set to 10^{-3} .

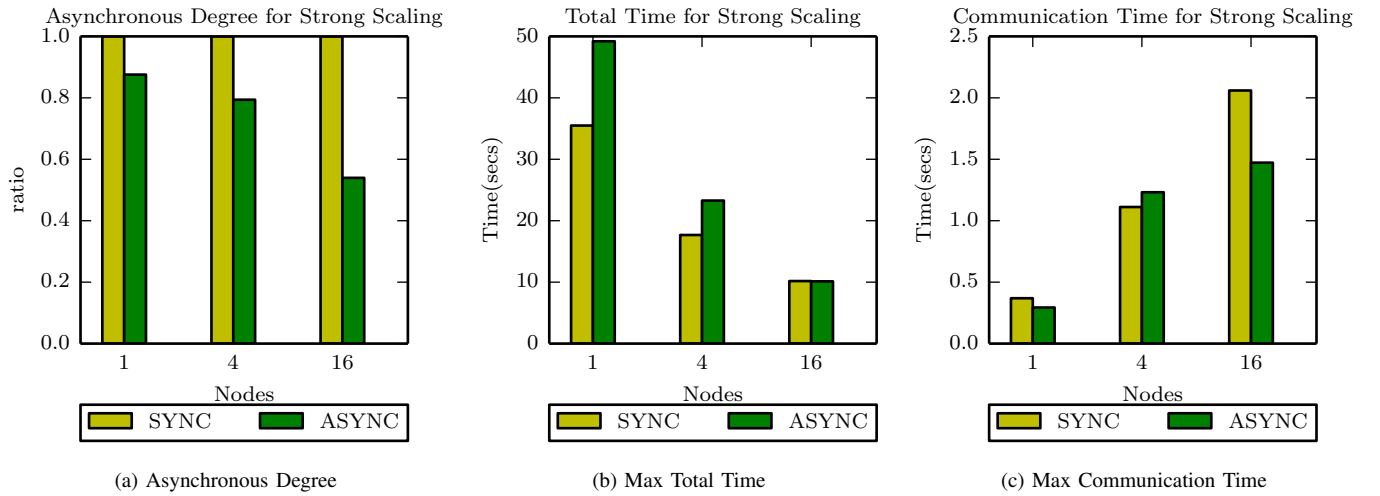


Fig. 8. Study of strong scaling effects for JS with respect to the asynchronous degree, the maximum total time and the maximum communication time.

time for the asynchronous method decreases at a faster pace than the synchronous one. This can be attributed to more efficient communication arising out of hardware support for asynchronous communication referred to in Section III. This fact is reinforced by results depicted in Figure 8c where we see that with an increase in number of processes, the growth rate in asynchronous communication costs is less than that of its synchronous counterpart. The disparity between synchronous and asynchronous methods can be attributed to stronger effects of RDMA and asynchronous progress control. From the results depicted in Figures 8b and 8c, we can say that in asynchronous JS subdomains are able to fully utilize new information from a subset of neighbors along with the latest available information from the rest to converge faster. This ability to take advantage of newest updates from neighboring subdomains without waiting for all subdomains to finish coupled with a reduced communication cost drives the system towards faster convergence.

G. One-level RAS

We compare synchronous and asynchronous one-level RAS in a strong scaling experiment, where we fix the global problem size to about 261,000 unknowns, and vary the number of subdomains between 4 and 256. We obviously cannot expect good scaling behavior for this one level method, since the number of subdomains adversely affects the rate of convergence. The iteration is terminated based on the simplistic convergence criterion described in Section IV-C. All experiments were performed on the Haswell partition of Cori. In Figure 9 we display solve time, final residual norm and approximate rate of convergence. It can be observed that the synchronous method is faster for smaller numbers of subdomains, yet comparatively slower for larger core counts. The crossover point between the two regimes appears to be at 64 subdomains.

An important question is whether the asynchronous method happens to converge because every subdomain performs the same number of local iterations, and hence the asynchronous method just mirrors the synchronous one, with merely the communication method exchanged. The histogram in Figure 10 shows that this is not the case. The number of local iterations varies significantly. The slowest subdomain performs barely more than 11,000 iterations, whereas the fastest one almost reaches 16,000. The problem was load balanced by the number of degrees of freedoms, thus the local solves are also approximately balanced but the communication is likely slightly imbalanced.

The advantage of asynchronous RAS becomes even clearer when the experiment is repeated under load imbalance. We create an artificial load imbalance by choosing one of the subdomains to be 50% larger than the rest. In Figure 11 it is observed that the asynchronous method outperforms the synchronous one in all but the 4 subdomain case.

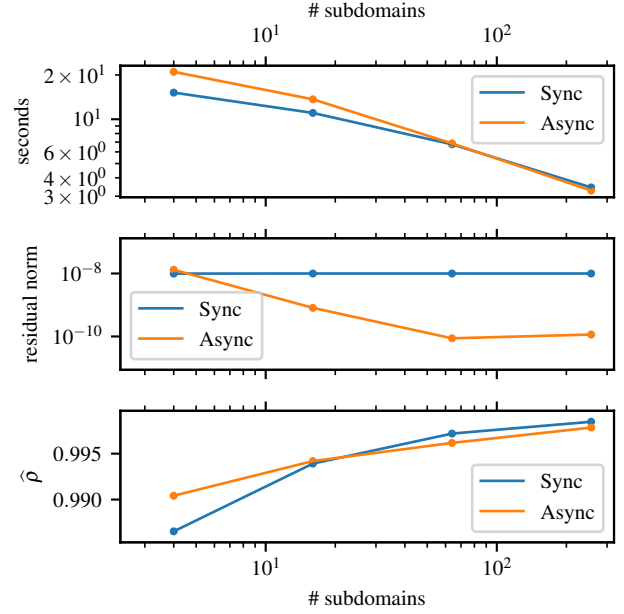


Fig. 9. Performance of synchronous and asynchronous one-level RAS for a system size of approximately 261,000 unknowns. The subdomains are load balanced. From top to bottom: Solution time, final residual norm, and the resulting approximate contraction factor. It can be observed that the synchronous method is significantly faster for smaller numbers of subdomains, yet comparatively slower for larger core counts, as shown by the contraction factor.

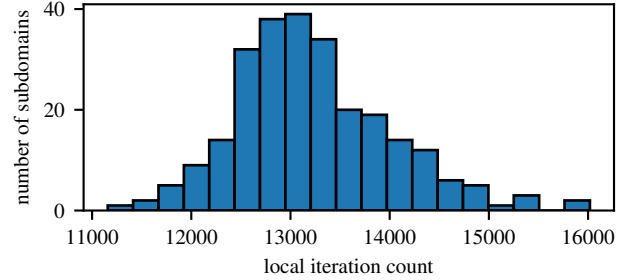


Fig. 10. Histogram of local iteration counts asynchronous one-level RAS with 256 subdomains.

H. Two-level RAS

In order to gauge the performance and scalability of the synchronous and asynchronous two-level RAS solvers, we perform weak and strong scaling experiments.

1) *Weak scaling*: In the weak scaling experiment the number of subdomains P and the global number of degrees of freedom (DoFs) are increased proportionally. We use 16, 64, 256 and 1024 subdomains. The local number of unknowns on each subdomain is kept constant at almost 20,000. The coarse grid problem increases in size proportionally to the number of subdomains, with approximately 16 unknowns per subdomain. Again, the iteration is terminated based on the simplistic convergence criterion described in Section IV-C.

In Figure 12 we plot the solution time, the achieved residual norm and the average contraction factor $\hat{\rho}$ depending on the global problem size. Both the synchronous and the

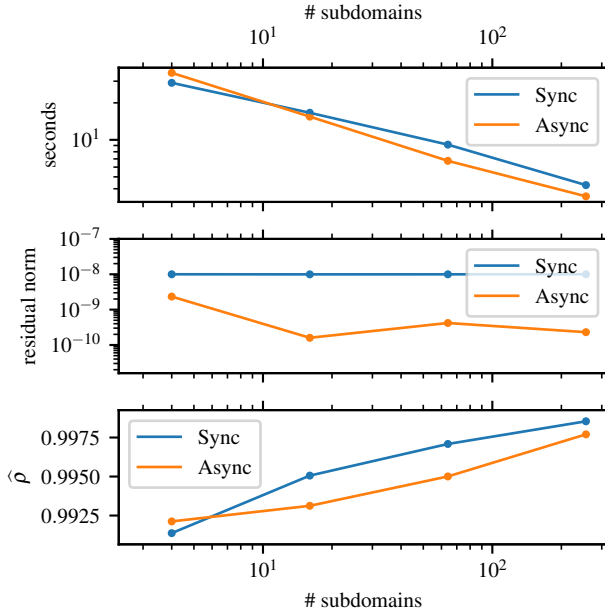


Fig. 11. Performance of synchronous and asynchronous one-level RAS for a system size of approximately 261,000 unknowns under load imbalance: one subdomain is 50% larger than the rest. From top to bottom: Solution time, final residual norm, and approximate contraction factor. It can be observed that the asynchronous method outperforms the synchronous one in all but the 4 subdomain case, as shown by the contraction factor. The advantage of the asynchronous method over the synchronous one is increased, as compared to Figure 9.

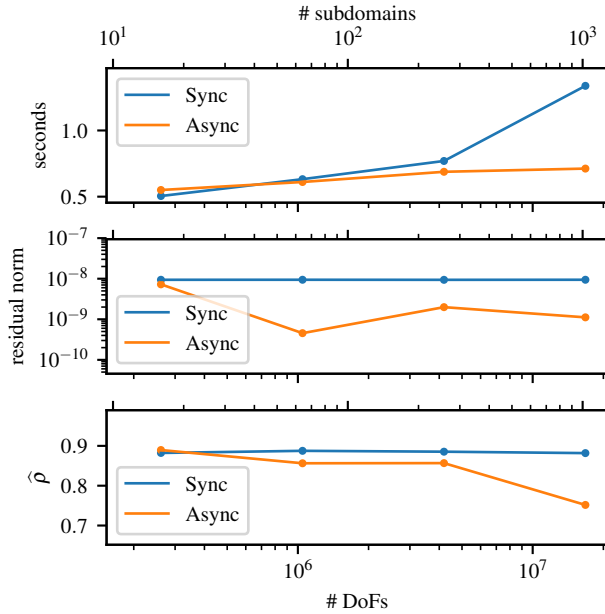


Fig. 12. Weak scaling of synchronous and asynchronous two-level additive RAS, load balanced case. From top to bottom: Total solution time, final residual norm, and approximate contraction factor. One can observe that for 16, 64 and 256 subdomains, the asynchronous and the synchronous method take almost the same time for the solve, with a slight advantage of the asynchronous method. For 1024 subdomains, however, the synchronous method is seen to take drastically more time, since the coarse grid, due to its size, starts to be the limiting factor. The asynchronous method is not affected by this.

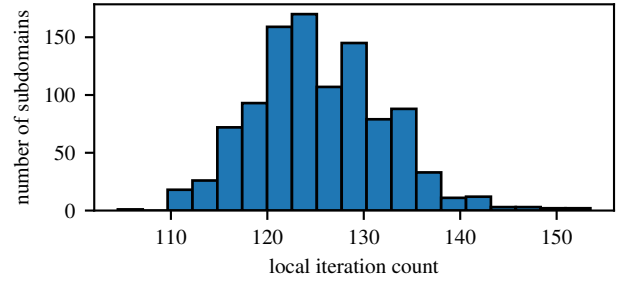


Fig. 13. Histogram of local iteration counts asynchronous two-level additive RAS with 1024 subdomains.

asynchronous method reach the prescribed tolerance of 10^{-8} . Due to the lack of an efficient mechanism of convergence detection, the asynchronous method ends up iterating longer than necessary, so that the final residual norm often is smaller than 10^{-9} . The number of iterations in the synchronous case is about 110, whereas the number of local iterations in the asynchronous case varies between 110 and 150. (See Figure 13.) Both iteration counts are significantly lower than the ones encountered in the one-level methods. One can observe that for 16, 64 and 256 subdomains, the asynchronous and the synchronous method take almost the same time for the solve. For 1024 subdomains, however, the synchronous method is seen to take drastically more time. This can be explained by the fact that for 1024 subdomains, the size of the coarse grid is comparable to the size of the subdomains, and hence the coarse grid solve which exchanges information with all the subdomains slows down the overall progress. For the asynchronous case this is not observed, since the subdomains do not have to wait for information from the coarse grid. This explains why we see better weak scalability for the asynchronous method than for the synchronous variant. The third subplot of Figure 12 shows that the asynchronous method outperforms its synchronous equivalent in all but the smallest problem.

To further illustrate the effect of load imbalance, we repeat the previous experiment with one subdomain being 50% larger than the rest. The results are shown in Figure 14. While the results are mostly consistent with the previous case, it can be seen that, as expected, the performance advantage of the asynchronous method over the synchronous one has increased. Even before the size of the coarse grid system is comparable to the size of the typical subdomain problem, the asynchronous method outperforms its synchronous counterpart.

2) *Strong scaling*: For the strong scaling experiment the global number of degrees of freedom is fixed at about 4 million. The coarse grid problem consists of approximately 4,000 unknowns. The number of subdomains used on the fine level takes values in $\{4, 16, 64, 256\}$. This means that the coarse grid problem is always smaller than the typical subdomain problem, and no load imbalance arises between the two levels.

The timing results are shown in the top of Figure 15. Both synchronous and asynchronous method display good strong

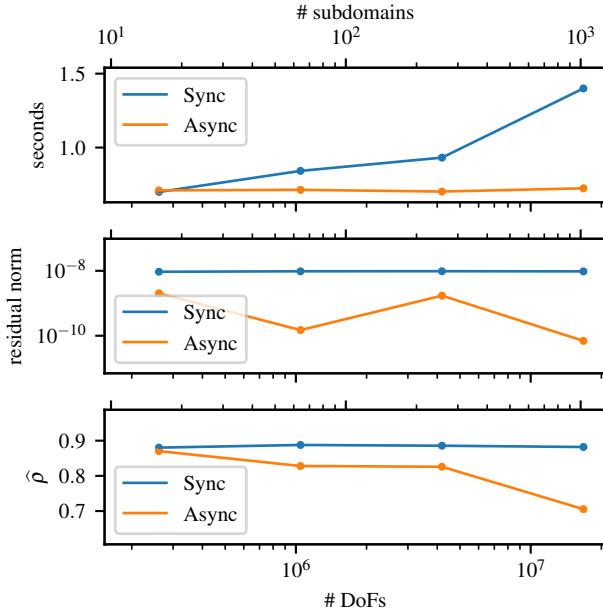


Fig. 14. Weak scaling of synchronous and asynchronous two-level additive RAS under load imbalance: one subdomain is 50% larger than all the other ones. From top to bottom: Total solution time, final residual norm, and approximate contraction factor. The advantage of the asynchronous method over the synchronous one is increased, as compared to Figure 12.

scaling behavior. It is observed that the synchronous method is faster than the asynchronous method for smaller subdomain count. But already for 64 subdomains this behavior is reversed, and the asynchronous method outperforms the synchronous one. This suggests that synchronization is an important factor already at modest core count.

At the bottom of Figure 15, we show the timing results in the case of load imbalance. It can be seen that the asynchronous method is faster than the synchronous one independent of the number of subdomains, and that its performance advantage increases as more processes are used.

V. CONCLUSION

In the present work, we have explored the use of asynchronous alternatives to conventional (synchronous) one-level and two-level domain decomposition solvers. To the best of our knowledge, we proposed the first truly asynchronous two-level method, where each processor can do different number of updates (iterations). Previous “asynchronous” methods could only solve the two levels in parallel but still require synchronization after each iteration. Several options to achieve asynchronous communication were tested, and we found that our use case benefited most from using `MPI_Lockall/MPI_Unlockall`. The numerical results presented demonstrate that asynchronous iterations can be considered a viable alternative to synchronous methods, despite partial availability of information from neighbors. Asynchronous methods seem to be beneficial already at modest core count, even for load balanced scenarios. In the presence of load imbalance, their performance advantage becomes even

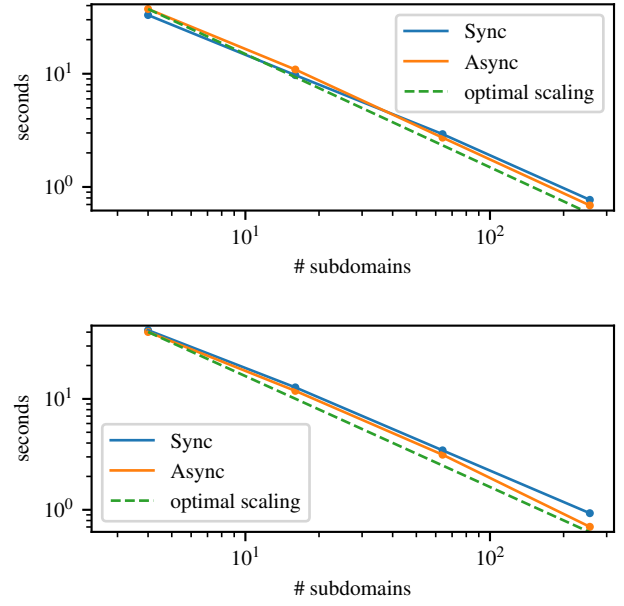


Fig. 15. Strong scaling of synchronous and asynchronous two-level additive RAS. On top: load balanced subdomains. At the bottom: load imbalance, one subdomain is 50% larger than all the other ones.

clearer. The inclusion of a novel asynchronous coarse grid correction paves the way for asynchronous methods to be used in extremely scalable parallel solvers.

ACKNOWLEDGMENT

We thank Daniel Szyld for helpful discussions on asynchronous methods. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Numbers DE-SC-0016564. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

- [1] D. Chazan and W. Miranker, “Chaotic relaxation,” *Linear algebra and its applications*, vol. 2, no. 2, pp. 199–222, 1969.
- [2] V. Dolean, P. Jolivet, and F. Nataf, *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation*. SIAM, 2015, vol. 144.
- [3] A. Toselli and O. Widlund, *Domain decomposition methods: algorithms and theory*. Springer, 2005, vol. 3.
- [4] B. Smith, P. Bjorstad, and W. Gropp, *Domain decomposition: parallel multilevel methods for elliptic partial differential equations*. Cambridge university press, 2004.
- [5] P. Ghysels, T. J. Ashby, K. Meerbergen, and W. Vanroose, “Hiding global communication latency in the GMRES algorithm on massively parallel machines,” *SIAM Journal on Scientific Computing*, vol. 35, no. 1, pp. C48–C71, 2013.
- [6] F. Cappello, A. Geist, B. Gropp, L. Kale, B. Kramer, and M. Snir, “Toward exascale resilience,” *International Journal of High Performance Computing Applications*, vol. 23, no. 4, pp. 374–388, 2009.
- [7] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, and M. Snir, “Toward exascale resilience: 2014 update,” *Supercomputing frontiers and innovations*, vol. 1, no. 1, pp. 5–28, 2014.
- [8] G. M. Baudet, “Asynchronous iterative methods for multiprocessors,” *Journal of the ACM (JACM)*, vol. 25, no. 2, pp. 226–244, 1978.

- [9] D. P. Bertsekas, "Distributed asynchronous computation of fixed points," *Mathematical Programming*, vol. 27, no. 1, pp. 107–120, 1983.
- [10] A. Frommer and D. B. Szyld, "On asynchronous iterations," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1, pp. 201–216, 2000.
- [11] F. Magoulès, D. B. Szyld, and C. Venet, "Asynchronous optimized Schwarz methods with and without overlap," *Numerische Mathematik*, pp. 1–29, 2017.
- [12] Z. Peng, Y. Xu, M. Yan, and W. Yin, "ARock: An Algorithmic Framework for Asynchronous Parallel Coordinate Updates," *SIAM Journal on Scientific Computing*, vol. 38, no. 5, pp. A2851–A2879, 2016.
- [13] A. T. Chronopoulos and C. W. Gear, "s-step iterative methods for symmetric linear systems," *J. Comput. Appl. Math.*, vol. 25, no. 2, pp. 153–168, 1989.
- [14] —, "On the efficient implementation of preconditioned s-step conjugate gradient methods on multiprocessors with memory hierarchy," *Parallel computing*, vol. 11, no. 1, pp. 37–53, 1989.
- [15] I. Yamazaki, S. Rajamanickam, E. G. Boman, M. Hoemmen, M. A. Heroux, and S. Tomov, "Domain decomposition preconditioners for communication-avoiding Krylov methods on a hybrid CPU/GPU cluster," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2014, pp. 933–944.
- [16] G. Karypis and V. Kumar, "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 359–392, 1998.
- [17] X.-C. Cai and M. Sarkis, "A restricted additive Schwarz preconditioner for general sparse linear systems," *SIAM Journal on Scientific Computing*, vol. 21, no. 2, pp. 792–797, 1999.
- [18] X.-C. Cai, M. Dryja, and M. Sarkis, "Restricted additive Schwarz preconditioners with harmonic overlap for symmetric positive definite linear systems," *SIAM Journal on Numerical Analysis*, vol. 41, no. 4, pp. 1209–1231, 2003.
- [19] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, D. A. May, L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B. F. Smith, S. Zampini, H. Zhang, and H. Zhang, "PETSc users manual," Argonne National Laboratory, Tech. Rep. ANL-95/11 - Revision 3.9, 2018. [Online]. Available: <http://www.mcs.anl.gov/petsc>
- [20] J. Liu, J. Wu, and D. K. Panda, "High performance RDMA-based MPI implementation over InfiniBand," *International Journal of Parallel Programming*, vol. 32, no. 3, pp. 167–198, 2004.
- [21] M. Si, A. J. Pena, J. Hammond, P. Balaji, M. Takagi, and Y. Ishikawa, "Casper: An asynchronous progress model for MPI RMA on many-core architectures," in *Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International*. IEEE, 2015, pp. 665–676.
- [22] W. Hackbusch, *Iterative solution of large sparse systems of equations*, ser. Applied Mathematical Sciences. New York: Springer-Verlag, 1994, vol. 95.
- [23] J. M. Bahi, S. Contassot-Vivier, R. Couturier, and F. Vernier, "A decentralized convergence detection algorithm for asynchronous parallel iterative algorithms," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 1, pp. 4–13, 2005.
- [24] F. Magoulès and G. Gbikpi-Benissan, "Distributed convergence detection based on global residual error under asynchronous iterations," *IEEE Transactions on Parallel and Distributed Systems*, 2017.
- [25] X. Li, J. Demmel, J. Gilbert, iL. Grigori, M. Shao, and I. Yamazaki, "SuperLU Users' Guide," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-44289, September 1999, <http://crd.lbl.gov/~xiaoye/SuperLU/>. Last update: August 2011.
- [26] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu, "A supernodal approach to sparse partial pivoting," *SIAM J. Matrix Analysis and Applications*, vol. 20, no. 3, pp. 720–755, 1999.