Asynchronous scalable version of the Global-Local non-invasive coupling

Ahmed EL KERIM*,**,†, Pierre GOSSELET** and Frédéric MAGOULES***

*Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS, LMPS - Laboratoire de Mécanique Paris-Saclay 91190, Gif-sur-Yvette, France
ahmed.elkerim@ens-paris-saclay.fr

**Université de Lille , CNRS, Centrale Lille / LaMcube
pierre.gosselet@univ-lille.fr

***Université Paris-Saclay, CentraleSupélec / MICS, Gif-sur-Yvette, France.
{ahmed.el-kerim,frederic.magoules}@centralesupelec.fr

†Corresponding author

Abstract

The Global-Local non-invasive coupling is an improvement of the submodeling technique, which permits to locally enhance structure computations by introducing patches with refined models and to take into accounts all the interactions. In order to circumvent its inherently limited computational performance, we propose and implement an asynchronous version of the method. The asynchronous coupling reduces the dependency on communications, failures, and load imbalance. We present the theory and the implementation of the method in the linear case and illustrate its performance on academic cases inspired by actual industrial problems.

1. Introduction

The non-invasive global-local coupling is an iterative technique that aims at making accurate the well-known submodeling technique. It is strongly related to many reanalysis techniques and domain decomposition methods. Starting from a global simplified model, this technique allows inserting local alterations (geometry, material, load and mesh) and evaluate their effect without heavy intervention on the initial model. The method is non-invasive in the sense that it is adapted to coupling commercial (closed) and research software, its first implementation in Abaqus was proposed in. This philosophy was successfully applied in many contexts like: the introduction of local plasticity and geometrical refinements, the computation of the propagation of cracks in a sound model, the taking into account of the exact geometry of connectors in an assembly of plates. In the method was used in order to implement a nonlinear domain decomposition method in a non-invasive manner in code_aster. Extension of the approach to explicit dynamics was proposed in, improved in and applied to the prediction of delamination under impact loading in. All the above applications were developed in a synchronous framework which has been taken advantage of by the use of accelerators (Aitken, quasi-Newton, Krylov), sec were the method is proved to be an implementation of an alternating Dirichlet-Robin Schwarz domain decomposition method where the Robin parameter corresponds to the condensation of the coarse domain covered by the patch.

Unfortunately, even if fast convergence is often observed, the method possesses inherent limitations in terms of computational performance. The objective of this work is to propose an asynchronous version of the algorithm. Asynchronous parallel computation, first introduced by, was then the subject of several theoretical works to prove its convergence in linear and nonlinear situations. Recent works show that domain decomposition methods are well adapted to asynchronous parallel computation and may lead to fascinating results. The idea is to allow each processor to move at its speed without waiting for the others, only considering the latest version of the data available.

The paper is organized as follows: Section 2 gives the basics of the Global-Local algorithm, Section 3 presents its asynchronous version and Section 4 provides illustrations in the linear case.
2. Global-Local coupling

We consider linear elliptic equations, typically arising from static thermal or elastic (small deformations and small displacements) problems. As in the submodeling approach, the starting point of the Global-Local coupling is a Global (index \(G\)) model which is a simplification of the problem to be solved adapted to a fast calculation and able to give a correct representation of the long-distance fluxes (see Fig. 1a). The Global model is insufficient in certain zones of interest (denoted by \(\Omega^{(i)}\) with \(s > 0\)) which need refined geometry, material law and adapted meshes (see Fig. 1b). A local reanalysis is then conducted in the Fine models (index \((i)\)) with Dirichlet conditions inherited from the Global computation. The submodeling approach would stop at this point, but large errors could be obtained, in particular if the patches evolve nonlinearly, compared to the Reference computation (Fig 1c, index \(R\)) where the Fine zones of interest replace their Global counterparts (index \((0)\), \(G\), \(s > 0\)) and a complete computation is run. Indeed, in the submodeling Local-to-Global and Local-to-other-Local interactions are totally omitted.

The error committed during the coupling can be measured by the lack of balance between the Fine zone of interest and their Global neighborhood (index \((0)\), \(G\)). The Global-Local method consists in evaluating this residual and re-imposing it in the Global model as an extra load, in a Richardson iteration manner.

![Figure 1: Models and subdomains for the Global/Local coupling](image)

For simplicity reason, we derive the method in after finite element discretization. We note \(K\) the stiffness matrices (or their thermal equivalent), which are symmetrical positive semidefinite (definite as soon as enough Dirichlet conditions are given), \(f_{\text{ext}}\) the generalized load vector, and \(u\) the unknown vector (temperature or displacement).

The interface \(\Gamma\) is defined as the boundary of patches: for each subdomain \(\Gamma^{(i),G} = \partial \Omega^{(i),G} \setminus \partial \Omega^{(i),F}\) and globally \(\Gamma^G = \bigcup \Gamma^{(i),G}\). The Fine version of the local interfaces \(\Gamma^{(i),F} = \partial \Omega^{(i),F} \setminus \partial \Omega^R\) is geometrically conforming with \(\Gamma^{(i),G}\) but can be meshed more finely. We introduce the global trace operator \(T^G : \Omega^G \rightarrow \Gamma^G\), and the local fine traces \(T^{(i),F} : \Omega^{(i),F} \rightarrow \Gamma^{(i),F}\), note that their transpose is the extension-by-zero operator. We note \(J^{(i)} : \Gamma^{(i),G} \rightarrow \Gamma^{(i),F}\) the interpolation operator between the meshes. Finally, \(A^{(i),G} : \Gamma^{(i),G} \rightarrow \Gamma^G\) is the assembly operators which inject the subdomain’s interface into the global interface.

Let \(p\) be a nodal effort defined on the interface \(\Gamma\), initialized by 0, whose role will soon be made clear: the Global problem can be written as:

\[
K^G u^G = f^G_{\text{ext}} + T^G p,
\]

and the interface nodal reaction can be post-processed in the Global zone not covered by patches, on referred to as complement zone, and numbered \((0)\):

\[
X^{(0),G} = T^{(0),G} (K^{(0),G} u^{(0),G} - f^{(0),G}_{\text{ext}}).
\]

The Fine problems can be written as:

\[
\begin{pmatrix}
K^{(i),F} & T^{(i),F} \\
T^{(i),F} & 0
\end{pmatrix}
\begin{pmatrix}
u^{(i),F} \\
-J^{(i),F}
\end{pmatrix}
= \begin{pmatrix}
f^{(i),F}_{\text{ext}} \\
0
\end{pmatrix},
\]

\(X^{(i),F}\) is the Lagrange multiplier associated with the Dirichlet condition. The minus sign is used to make it interpretable as the nodal reaction exerted by the surrounding of the patch.
It is then possible to evaluate the compatibility between the models as the lack of balance at the interface:

\[
\mathbf{r} = - \left( \mathbf{A}^{(0)} \mathbf{G}^{(0)} + \sum_{s>0} \mathbf{A}^{(s)} J^{(s)} \mathbf{F}^{(s)} \right)
\]  

(4)

If \( \mathbf{r} \) is not small enough, it is injected in the computation in a modified Richardson iteration way:

\[
\mathbf{p} \leftarrow \mathbf{p} + \omega \mathbf{r}
\]  

(5)

where \( \omega \) is a relaxation parameter to be determined.

Under the chosen hypothesis, it can be proved that the method converges to the solution of the reference problem for a sufficiently small relaxation. In fact convergence is more general and can be improved by acceleration, in particular Aitken, see among others,18,33

The Global-Local coupling is recalled in Algorithm 1.

**Algorithm 1**: Non-invasive synchronous stationary iterations

| Arbitrary initialization \( \mathbf{p}_0^G \), Relaxation parameter \( \omega \) |
| for \( j \in [0, \ldots, m] \) do |
| Global solve for \( \mathbf{u}_j^G \) for given \( \mathbf{p}_j \), eq. (1) |
| Postprocess \( \mathbf{A}^{(0)} \mathbf{G}^{(0)} \), eq. (2) |
| Parallel fine solve, for \( s > 0 \) obtain \( \mathbf{u}_j^{(s)} \mathbf{F}^{(s)} \) and \( \mathbf{A}^{(s)} \mathbf{J}^{(s)} \mathbf{F}^{(s)} \) from imposed \( \mathbf{T} \mathbf{u}^G \), eq. (3) |
| Assemble residual: \( \mathbf{r}_j \), eq. (4) |
| Update: \( \mathbf{p}_{j+1} = \mathbf{p}_j + \omega \mathbf{r}_j \) |
| end |

Figure 2a presents the time sequence of the classical approach applied to a two-patch case as the one of Figure 1. The global analysis is carried out in alternation with the local ones and generates waiting and inactivity times on both sides, which seriously affects the method’s performance. Problems due to load balancing, communication delays, or machine failures are always related to this synchronization.

![Figure 2: Time sequence of the synchronous and asynchronous Global-Local coupling.](figure2.png)
3. Asynchronous Coupling

3.1 Principle and algorithm

We investigate the potential of the asynchronous paradigm in order to obtain higher parallel performance than the classical approach. The idea is to launch a computation as soon as some a processor is idling and new input is available. The asynchronous time sequence is presented in Figure 2b, it is clear that getting rid of synchronization increase the intensity of the computation.

The iteration is detailed in Algorithm 2. It can be observed that the Global model is always assembling a residual, which makes it easy to tell convergence, while this point can be troublesome in other methods.\textsuperscript{16, 31}

\begin{algorithm}
\caption{Non-invasive asynchronous iterations}
\begin{algorithmic}
\STATE \textbf{Initialization} \textbf{p} = 0 and \langle q^{(s)} = 0 \rangle on the \textit{global} model (rank 0)
\IF{\textbf{Rank} 0 is available and detects at least one new \langle q^{(s)} \rangle}
\STATE Assemble residual: \textbf{r} = − Σ_{s=0}^{N} A^{(s)}q^{(s)}
\IF{\|\textbf{r}\| is small enough (initialization excluded)}
\STATE break
\ENDIF
\STATE Update \textbf{p} = \textbf{p} + \omega\textbf{r}
\STATE \textit{Global} solves for \textbf{u}^{G}_{j} for given \textbf{p}^{j}, eq. (1)
\STATE \textit{Global} puts \langle A^{(s)}u^{G}_{j} \rangle on the patches
\IF{there is a complement, compute \lambda^{(0),G} and set \langle q^{(0)}(s) = \lambda^{(0),G} \rangle}
\ENDIF
\STATE \textbf{if Subdomain} \textbf{s} > 0 \textbf{is available and detects new} \langle A^{(s)^{T}}u^{G} \rangle \textbf{then}
\STATE Parallel \textit{fine} solve, for \textbf{s} > 0 obtain \textbf{u}^{(s)^{F}}_{j} and \langle A^{(s)^{F}} \rangle from imposed \textbf{T}^{G}\textbf{u}^{G}, eq. (3)
\STATE Subdomain \textbf{s} > 0 puts \langle q^{(s)} = J^{(s)^{T}}A^{(s)^{F}} \rangle on the \textit{Global} model (rank 0)
\ENDIF
\end{algorithmic}
\end{algorithm}

The theoretical proof of the convergence of the asynchronous Global-Local iteration can be derived by the framework of paracontractions.\textsuperscript{14} The main result is that for a given relaxation parameter, the synchronous iteration convergences then so does the asynchronous iteration.

One drawback of the asynchronous iteration is that for now no acceleration strategy is available.

3.2 Implementation

Implementing an asynchronous communication protocol has been the subject of several research works, generally based on MPI as in,\textsuperscript{28} where the idea is the use of classical two-sided communication. However, new works based on one-sided communication,\textsuperscript{7, 23} also known as MPI-RDMA, have proven the efficiency of these techniques and their adaptation to asynchronous communication.

The performance of these techniques depends on the MPI version used and the network. We tested on several configurations: OPENMPI, INTELMPI, MPICH, and several network architectures like the classic Ethernet and more developed Infiniband or Intel OPA. The influence has been observed in asynchronous with sometimes implicit synchronization imposed by the network or minor performing communication operations depending on the MPI version.

The general idea of RDMA is to allow access to data on other machines without the need to involve the target machine. We create a part of the memory called a window in which we place the searched data. The other machines will be able to perform operations of type PUT or GET to update this information in the window or to recover it and use it afterward. The idea is thus well adapted to the asynchronous calculation because we are in a procedure where we do not need to stop computing to a Send or a Receive operations.

Figure 3 corresponds to a communication between a processor 0 and processor 1, and this last one applies two operations of communication PUT and GET, on the processor 0, to receive the value X and to send the value Y, these two operations as mentioned before have been made without involving the processor 0.

These communications are usually followed by a synchronization step. We distinguish two types of synchronization in RDMA the active synchronization, where we make a collective operation to update everyone before going from one iteration to another with the command: MPI\_WIN\_Fence(). The second technique, called passive synchronization, is used in asynchronous. This technique consists of synchronizing each processor without necessarily doing
4. Numerical results

Our code is realized in python. It uses several other tools and software like GMSH\textsuperscript{8} to generate the geometries and meshes of the studied cases. For the finite element approximation, we use the Getfem library.\textsuperscript{37} For the parallel side, we use the mpi4py library.\textsuperscript{11}

The study was carried out with the cluster of the LMPS simulation center using several workstations with an ethernet network. These machines are quite heterogeneous with 4 different generation of CPUs:(Intel(R) Xeon(R) CPU E5-1660 v3 (Haswell) @ 3.00GHz, Intel(R) Xeon(R) CPU E5-2630 v4 (Broadwell) @ 2.20GHz, Intel(R) Xeon(R) Silver 4116 CPU (Skylake) @ 2.10GHz, Intel(R) Xeon(R) W-2255 CPU (Cascade Lake) @ 3.70GHz.

We propose 3 illustrations to analyze the performance of the asynchronous coupling.

4.1 2D academic cases

To quickly show the effect and the improvements brought in asynchronous, we study the performances in synchronous and asynchronous of the simple problem presented previously in Figure 1. The global problem contains 701 nodes, the first patch 381 and the second 379. We are interested in a thermal problem (Poisson equation with constant source term) and a linear elasticity problem (under gravitational load). Zero Dirichlet boundary conditions are applied on the bottom side of the Global domain.

Table 1 gives the performance for the thermal problem while Table 2 deals with the (plane stress isotropic) elasticity problem. We show both computation time and number of iterations, in the asynchronous case, we distinguish between the number of Global iterations and Local iterations for each patch.

<table>
<thead>
<tr>
<th>Models</th>
<th>Synchronous</th>
<th>Aitken</th>
<th>Asynchronous</th>
<th>Relaxed asynchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.22 s</td>
<td>0.12 s</td>
<td>0.3 s</td>
<td>0.19 s</td>
</tr>
<tr>
<td>Iterations</td>
<td>23</td>
<td>12</td>
<td>45[96 - 97]</td>
<td>29[64 - 65]</td>
</tr>
</tbody>
</table>

Table 1: Performance for the 2D thermal problem.

<table>
<thead>
<tr>
<th>Models</th>
<th>Synchronous</th>
<th>Aitken</th>
<th>Asynchronous</th>
<th>Relaxed asynchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0.67 s</td>
<td>0.3 s</td>
<td>0.6 s</td>
<td>0.52 s</td>
</tr>
<tr>
<td>Iterations</td>
<td>43</td>
<td>16</td>
<td>53[112 - 119]</td>
<td>48[100 - 107]</td>
</tr>
</tbody>
</table>

Table 2: Performance for the 2D elasticity problem.

We compare synchronous and asynchronous versions with and without relaxation. In the synchronous case, relaxation is dynamically adapted using Aitken acceleration. In the asynchronous case, the relaxed iteration corresponds to the best result that we obtained in a trial-and-error campaign to fit the relaxation coefficient.
ASYNCHRONOUS GLOBAL-LOCAL COUPLING

The simplicity and well-balanced zones of interest in the problems are against the asynchronous approach. We can see that without relaxation, the asynchronous is faster in the case of linear elasticity. However, (synchronous) Aitken’s method provides a good acceleration which cannot be beaten even with carefully chosen relaxation. Anyhow, it is interesting to observe how getting rid of synchronization unleashes computational power: in comparison, much more computations are done with the asynchronous iteration in not as much more time.

4.2 3D Weak scalability

For this study, we generate a simple geometry problem which can easily be extended. The basic pattern is a cube. For the Global model, it is homogeneous and coarsely meshed while the Fine models have a heterogeneous spherical inclusion (the matrix has the same material properties as the Global model) and adapted meshes, see Figure 5.

![Figure 4: 64 subdomains test case for the weak scalability study.](image)

Figure 4: 64 subdomains test case for the weak scalability study.

![Figure 5: Detail of the geometry for the weak scaling test case](image)

Figure 5: Detail of the geometry for the weak scaling test case

The idea of the study is to increase the number of subdomains while keeping the same cubic geometry, so we consider the cases with $n^3$ subdomains with $n = 2, ..., 7$. Figure 4 presents the 64-subdomain global model ($n = 4$). Note that the Fine subdomains are perfectly balanced, which is favorable for the synchronous iteration. One face of the Global cube is submitted to zero Dirichlet boundary condition, while a constant source term is applied in the domain.

In Table 3, we summarize the number of nodes for each case. The Fine discretization is kept constant, while naturally the size of the Global mesh increases when subdomains are added.

<table>
<thead>
<tr>
<th># of subdomains</th>
<th>8</th>
<th>27</th>
<th>64</th>
<th>125</th>
<th>216</th>
<th>343</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>233</td>
<td>667</td>
<td>1449</td>
<td>2681</td>
<td>4465</td>
<td>6903</td>
</tr>
<tr>
<td>Local (1 subdomain)</td>
<td>1858</td>
<td>1858</td>
<td>1858</td>
<td>1858</td>
<td>1858</td>
<td>1858</td>
</tr>
</tbody>
</table>

Table 3: Number of nodes in the meshes.

We first consider a thermal problem where the inclusions are ten time more insulating than the matrix. A comparison of the time to solution for the synchronous (accelerated with Aitken) and asynchronous (with well-chosen relaxation) is presented in Figure 6a while Table 4 gives the number of iterations for the Global domain and for the max and min number of iterations for the Fine subdomains.

We then consider the same geometry for linear elasticity problem. The inclusions are 100 times more flexible than the matrix. Table 4 summarizes the number of iterations for each case, while Figure 6b compares the CPU time.
Asynchronous Global-Local Coupling

(a) Thermal problem

(b) Linear elasticity

Figure 6: Weak scalability study.

<table>
<thead>
<tr>
<th># subdomains</th>
<th>8</th>
<th>27</th>
<th>64</th>
<th>125</th>
<th>216</th>
<th>343</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aitken</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 4: Linear thermal (heterogeneity ratio 10) weak scalability study: # iterations

The behavior of the solver is roughly the same as for the thermal problem, except for small numbers of subdomains for which the synchronous iteration is more efficient.

In broad terms, we observe that despite the unfavorable configuration, the asynchronous algorithm is roughly two-time faster than the synchronous one. We can see that this performance is achieved despite the tremendously larger number of iterations.

<table>
<thead>
<tr>
<th># subdomains</th>
<th>8</th>
<th>27</th>
<th>64</th>
<th>125</th>
<th>216</th>
<th>343</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aitken</td>
<td>22</td>
<td>21</td>
<td>25</td>
<td>25</td>
<td>26</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 5: Linear elasticity (heterogeneity ratio 100): # iterations

4.3 3D test case

In this part, we are interested in studying a test case inspired by an industrial problem. The geometry corresponds to the turbine blade of an aircraft engine. The Global model makes use of a simplified geometry which omits cooling micro-perforations. The two zones of interest are two critical regions of the domain where the precise geometry (with the perforations) is taken into account, see Figures 7 and 8. The number of nodes of the meshes are given in Table 6, we can see than one zone of interest is about two times larger than the other one which is roughly of the same size as the Global model.

Details about the actual industrial problem can be found in. Here, we adopt a simplified version: a thermal problem is considered, with constant source term. In order to make the problem more complex (else very few iterations are needed), we artificially introduce a heterogeneity ratio of 10 (in the conductance coefficient) between the Global and the Fine models.

The parallel analysis is conducted using 3 CPUs: one for the global problem and the other two for each zone of interest. The performance is summed up in Table 7. The asynchronous iteration is about 30% faster than the accelerated (Aitken) synchronous iteration. Amazingly, we see that the largest subdomains needs fewer iterations in the asynchronous case (20) than in the synchronous case (22); in fact what mattered was having the Global and other subdomain sufficiently converged.
5. Conclusion

In this paper, an asynchronous version of the non-invasive Global-Local coupling has been presented. The MPI-RDMA parallelization techniques with passive synchronization have been used for the programming aspect. The presented results showed that the asynchronous version may allow better performance than the synchronous Aitken accelerator on a heterogeneous cluster.

Acknowledgments

This work was partly funded by the French National Research Agency as part of project ADOM, under grant number ANR-18-CE46-0008.
References


ASYNCHRONOUS GLOBAL-LOCAL COUPLING


