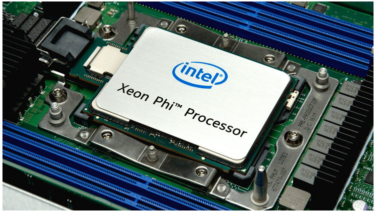
On the Ensemble Propagation for Efficient Uncertainty Quantification of Mechanical Contact Problems



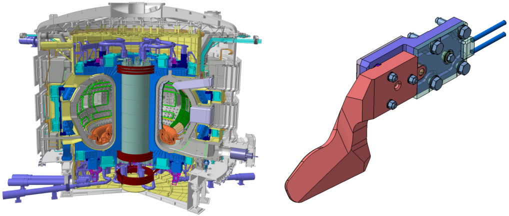
K. Liegeois[1], R. Boman[1], E. T. Phipps[2], T. A. Wiesner[2] and M. Arnst[1]

[1]Aerospace and Mechanical Engineering, Université de Liège, Belgium
[2]Sandia National Laboratories, USA

SIAM Conference on Uncertainty Quantification
Garden Grove, USA
April 17, 2018

Ongoing PhD: New methods for parametric computations with multiphysics models on HPC architectures with applications to design of opto-mechanical systems

ITER



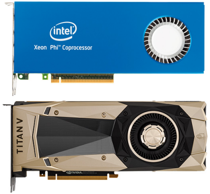
Ph. Mertens, A. Panin, FZ. Jülich

High performance computing library

Trilinos

Clusters

Emerging architectures

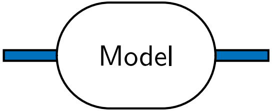


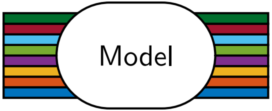
Motivation and context

Previous work on the Ensemble propagation (EP) [Phipps, 2017], [D'Elia, 2017]:

- Symmetric positive definite system ⇒ **Conjugate Gradient**,
- Reduced and not reduced norms and inner products, BLAS and preconditioners.

In sampling-based uncertainty quantification (UQ), instead of individually evaluating each instance of the model, EP consists of **simultaneously evaluating** a **subset of samples** of the model.



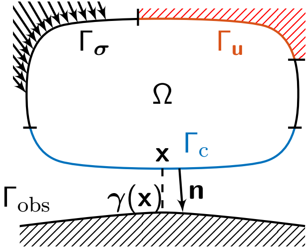


This work: going towards EP for **mechanical contact problems**

- Samples of a same ensemble can have different activities,
- Non-symmetric saddle-point system ⇒ **GMRES**,
- Reduced and not reduced norms and inner products, BLAS and preconditioners,
- Towards industrial problems.

Outline

Mechanical contact problem



Algorithm 1: Active set strategy

$k \leftarrow 0$

Choose an initial guess for the active set $\mathcal{A}_k$

do

  Given $\mathcal{A}_k$, compute the solution of

$$\left[\begin{array}{cc|cc} \mathbf{K}_{\mathrm{ii}} & \mathbf{K}_{\mathrm{ic}} & \mathbf{0} & \mathbf{0} \\ \mathbf{K}_{\mathrm{ci}} & \mathbf{K}_{\mathrm{cc}} & \mathbf{D}_{\mathcal{I}_k}^{\mathrm{T}} & \mathbf{D}_{\mathcal{A}_k}^{\mathrm{T}} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}_{\mathcal{I}_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{\mathcal{A}_k} & \mathbf{0} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{u}_{\mathrm{i}}^{k+1} \\ \mathbf{u}_{\mathrm{c}}^{k+1} \\ \hline \boldsymbol{\lambda}_{\mathcal{I}_k}^{k+1} \\ \boldsymbol{\lambda}_{\mathcal{A}_k}^{k+1} \end{array}\right] = \left[\begin{array}{c} \mathbf{f}_{\mathrm{i}} \\ \mathbf{f}_{\mathrm{c}} \\ \hline \mathbf{0} \\ \mathbf{g}_{0,\mathcal{A}_k} \end{array}\right]$$

  $\mathcal{A}_{k+1} \leftarrow \left\{ q \in P_{\mathrm{c}}^{h,\mathrm{s}} : \lambda_q^{k+1} + c\, \mathbf{e}_q^{\mathrm{T}} \left(\mathbf{D}\mathbf{u}_{\mathrm{c}}^{k+1} - \mathbf{g}_0 \right) > 0 \right\}$

  $k \leftarrow k+1$

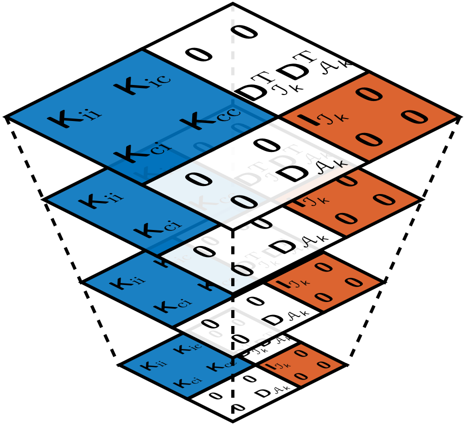
while $\mathcal{A}_k \neq \mathcal{A}_{k-1}$

Inner nodes: i, potential contact nodes: c, at iteration k, inactive set: $\mathcal{I}_k$, and active set: $\mathcal{A}_k$.

Preconditioners: Full multigrid approach

Introduced in [Wiesner, 2015] for **contact problem**.

- Main idea: use **coarser representations** of fine level problems in order to **speed up** the solution process,
- Uses the multigrid approach on **the full matrix**, preserving the **saddle-point structure** on all levels,
- Algebraic multigrid: **no special information** is necessary to build the mutligrid **hierarchies**,
- Mutligrid hierarchies are **independent of the activity** of the Lagrange multipliers.
- Allows the use of a **direct solver** on **the coarsest level**.



Outline

Ensemble propagation for mechanical contact problem

Instead of individually solving the mechanical contact problem for each instance of the model, we have to **solve simultaneously** the mechanical contact problem for **a subset of samples** of the model.

Advantages of the EP:

- Reuse of common variables,
- Improved probability of auto-vectorization,
- Improved memory usage,
- Reduction of Message Passing Interface (MPI) latency per sample.

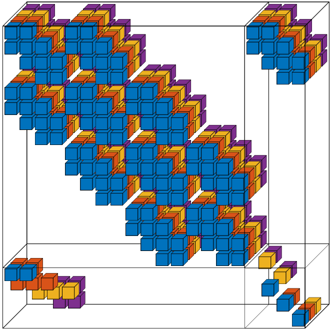
} Improve **throughput**.

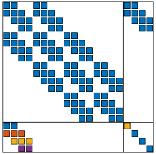
Difficulties of the EP for mechanical contact problem:

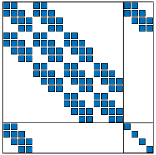
- Different samples can have **different active Lagrange multipliers**,
- Samples may require a **different number** of **active set iterations**,
- For a given active set iterations, they may require **different number** of **Krylov iterations**.

The algebraic full form as a way to handle activities

The matrix of the system:





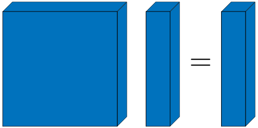


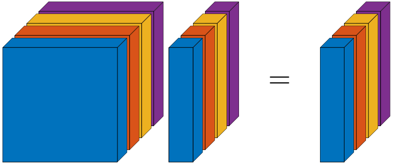
- has a **constant size** but its **graph varies** with the active set,
- can be stored using an **extended graph** which is the union of all the possible graphs,
- has a **saddle-point** structure,
- is **not positive definite** (if at least one Lagrange multiplier is active).

Outline

GMRES with Ensemble propagation

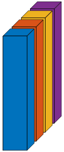
Instead of individually solving the GMRES for each instance of the model, we have to **solve simultaneously** the GMRES for **a subset of samples** of the model.





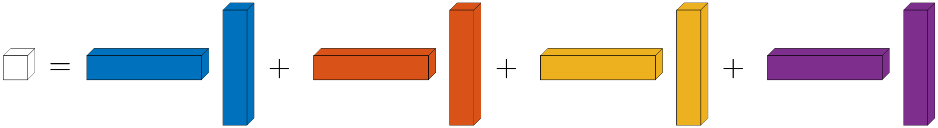
GMRES is based on the notion of **inner products** and **norms**.

What is an inner product (and its associated norm) of vectors of ensemble type?

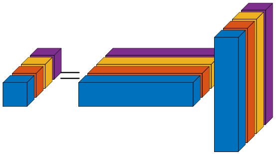


Reduced and not reduced inner products

- **Reduced inner product** and its associated norm were the first one introduced, implemented, and tested in the EP [Phipps, 2017]:



- **Not reduced inner product** and its associated norm were first introduce for grouping purpose [D'Elia, 2017]:



GMRES using reduced and not reduced inner products

Algorithm 2: Not reduced norm GMRES

$\boldsymbol{r}_\ell = \boldsymbol{b}_\ell - \boldsymbol{A}_\ell\, \boldsymbol{x}_\ell^{(0)}, \quad \ell = 1, \dots, s$

$\beta_\ell = \|\boldsymbol{r}_\ell\|, \quad \ell = 1, \dots, s$

$\boldsymbol{v}_{1,\ell} = \boldsymbol{r}_\ell / \beta_\ell, \quad \ell = 1, \dots, s$

for $j = 1, \dots, m$ **do**

- $\boldsymbol{w}_\ell = \boldsymbol{A}_\ell\, \boldsymbol{v}_{j,\ell}, \quad \ell = 1, \dots, s$
- $h_{ij,\ell} = \langle \boldsymbol{v}_{i,\ell}, \boldsymbol{w}_\ell \rangle, \quad \ell = 1, \dots, s, \quad i = 1, \dots, j$
- $\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \sum_{i=1}^{j} h_{ij,\ell} \boldsymbol{v}_{i,\ell}, \quad \ell = 1, \dots, s$
- $h_{(j+1)j,\ell} = \|\widehat{\boldsymbol{v}}_\ell\|, \quad \ell = 1, \dots, s$
- $\boldsymbol{v}_{(j+1),\ell} = \widehat{\boldsymbol{v}}_\ell / h_{(j+1)j,\ell}, \quad \ell = 1, \dots, s$
- **if** $h_{(j+1)j,\ell}/\beta_\ell \le \epsilon, \quad \forall \ell \in \{1, \dots, s\}$ **then**
  - $m = j$
  - **break**

$\min_{\boldsymbol{y}_\ell} \|\beta_\ell\, \boldsymbol{e}_1 - \boldsymbol{H}_\ell \boldsymbol{y}_\ell\|, \quad \ell = 1, \dots, s$

$\boldsymbol{x}_\ell^{(m)} = \boldsymbol{x}_\ell^{(0)} + \boldsymbol{V}_\ell\, \boldsymbol{y}_\ell, \quad \ell = 1, \dots, s$

Algorithm 3: Reduced norm GMRES

$\boldsymbol{r}_\ell = \boldsymbol{b}_\ell - \boldsymbol{A}_\ell\, \boldsymbol{x}_\ell^{(0)}, \quad \ell = 1, \dots, s$

$\beta = \sqrt{\sum_{\ell=1}^{s} \|\boldsymbol{r}_\ell\|^2}$

$\boldsymbol{v}_{1,\ell} = \boldsymbol{r}_\ell / \beta, \quad \ell = 1, \dots, s$

for $j = 1, \dots, m$ **do**

- $\boldsymbol{w}_\ell = \boldsymbol{A}_\ell\, \boldsymbol{v}_{j,\ell}, \quad \ell = 1, \dots, s$
- $h_{ij} = \sum_{\ell=1}^{s} \langle \boldsymbol{v}_{i,\ell}, \boldsymbol{w}_\ell \rangle, \quad i = 1, \dots, j$
- $\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \sum_{i=1}^{j} h_{ij}\, \boldsymbol{v}_{i,\ell}, \quad \ell = 1, \dots, s$
- $h_{(j+1)j} = \sqrt{\sum_{\ell=1}^{s} \|\widehat{\boldsymbol{v}}_\ell\|^2}$
- $\boldsymbol{v}_{(j+1),\ell} = \widehat{\boldsymbol{v}}_\ell / h_{(j+1)j}, \quad \ell = 1, \dots, s$
- **if** $h_{(j+1)j}/\beta \le \epsilon$ **then**
  - $m = j$
  - **break**

$\min_{\boldsymbol{y}} \|\beta\, \boldsymbol{e}_1 - \boldsymbol{H} \boldsymbol{y}\|$

$\boldsymbol{x}_\ell^{(m)} = \boldsymbol{x}_\ell^{(0)} + \boldsymbol{V}_\ell\, \boldsymbol{y}, \quad \ell = 1, \dots, s$

GEMM operations in orthogonalization of the GMRES

Componentwise orthogonalization:

Algorithm 4: Not reduced orthogonalization

$h_{ij,\ell} = \langle \boldsymbol{v}_{i,\ell}, \boldsymbol{w}_\ell \rangle, \quad \ell = 1, \dots, s, \quad i = 1, \dots, j$

$\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \sum_{i=1}^{j} h_{ij,\ell} \boldsymbol{v}_{i,\ell}, \quad \ell = 1, \dots, s$

Algorithm 5: Reduced orthogonalization

$h_{ij} = \sum_{\ell=1}^{s} \langle \boldsymbol{v}_{i,\ell}, \boldsymbol{w}_\ell \rangle, \quad i = 1, \dots, j$

$\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \sum_{i=1}^{j} h_{ij} \boldsymbol{v}_{i,\ell}, \quad \ell = 1, \dots, s$

Writting with matrix vector multiplications:

Algorithm 6: Not reduced orthogonalization

$\boldsymbol{h}_{j,\ell} = \boldsymbol{V}_\ell^{\mathrm{T}} \boldsymbol{w}_\ell, \quad \ell = 1, \dots, s$

$\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \boldsymbol{V}_\ell \boldsymbol{h}_{j,\ell}, \quad \ell = 1, \dots, s$

Algorithm 7: Reduced orthogonalization

$\boldsymbol{h}_j = \sum_{\ell=1}^{s} \boldsymbol{V}_\ell^{\mathrm{T}} \boldsymbol{w}_\ell$

$\widehat{\boldsymbol{v}}_\ell = \boldsymbol{w}_\ell - \boldsymbol{V}_\ell \boldsymbol{h}_j, \quad \ell = 1, \dots, s$

these operations can be implemented with **GEMV** routines or, more generally, with **GEMM** to support multiple right-hand sides, however, these implementations are not trivial and require to take into account the **memory layout** of the **ensemble type**.

Pros and cons of both approaches

Not reduced norm:

Pros:

- At the end of the GMRES, the stop criterion **is fulfilled** by every sample individually.
- The spectrums **are not** gathered.
- Convergence rates **controlled** by the slowest sample.

Cons:

- Divisions by norms need **to be done with caution** to avoid underflow and division by zeros due to happy breakdown.
- **No current** implementation of the needed BLAS routines in the MKL.

Reduced norm:

Pros:

- **No division** by zero when we divide by the norm of a non-zero residual.
- Use of **standard libraries** such as MKL.

Cons:

- At the end of the GMRES, the stop criterion **may not** be fulfilled by every sample individually.
- The spectrum of the ensemble matrix **is the union** of the spectrum of each sample matrix: to have a good preconditioner is more complex.
- **Increase** the number of iterations.

Outline

Code

- The full mechanical contact simulation is implemented and **fully templated** in a homemade code heavily based on **Trilinos** [Heroux, 2005] which provides a full-templated solver stack.
- The **C++** code is embedded in a **Python** interface [Boman]. This eases the looping around samples, group samples together, etc.
- The software has **hybrid parallelism** based on **Tpetra** with **MPI** for distributed memory and **Kokkos** [Edwards, 2012] with **OpenMP** for shared memory.
- It uses **Gmsh** [Geuzaine, 2009] to import 3D meshes and **VTK** to write the output files.
- The code has already generated preliminary results for **industrial thermomechanical contact problems**.

Outline

(1) Motivation

(2) Mechanical contact problem

(3) Ensemble propagation for mechanical contact problem

(4) GMRES with Ensemble propagation

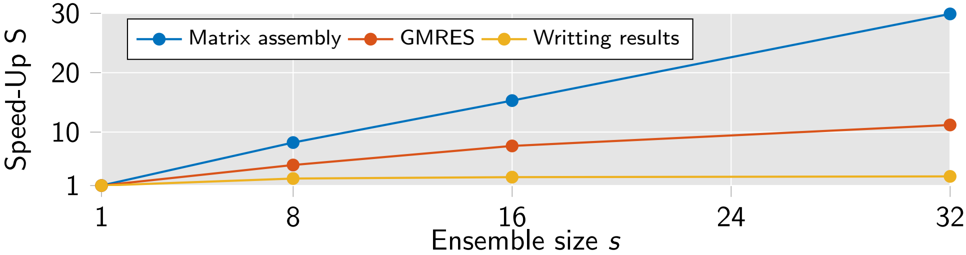
(5) Code

(6) First numerical results

Speed-Up and R

- **Speed-Up:** relative gain in CPU cost (architecture dependent):

$$\mathrm{S}(e) = \frac{\sum_{\ell \in e} \mathrm{Time}_\ell}{\mathrm{Time}_e}, \qquad \mathrm{S} = \frac{\sum_e \sum_{\ell \in e} \mathrm{Time}_\ell}{\sum_e \mathrm{Time}_e}.$$

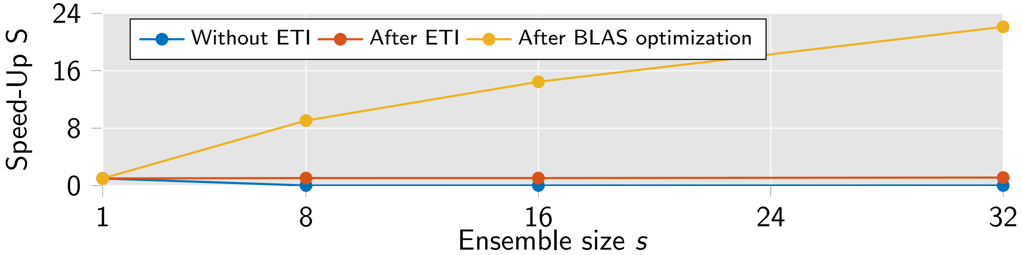


- **R:** relative increase in computational work (architecture independent):

$$\mathrm{R}(e) = \frac{s\,\#\mathrm{iterations}_e}{\sum_{\ell \in e} \#\mathrm{iterations}_\ell}, \qquad \mathrm{R} = \frac{s \sum_e \#\mathrm{iterations}_e}{\sum_e \sum_{\ell \in e} \#\mathrm{iterations}_\ell}.$$

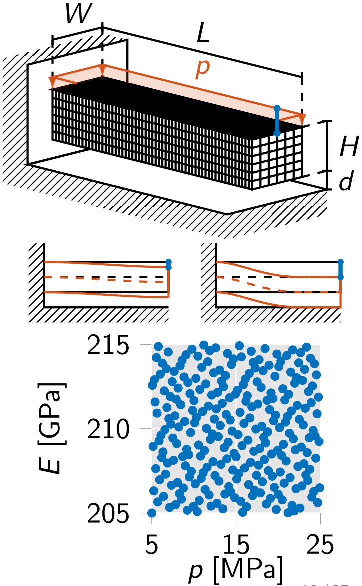
Speed-Up of the GMRES with not reduced norm

- **Default GEMM**: naive implementation with three nested loops.
- **Explicit Template Instantiation** (ETI): improves optimization of the code by the compiler.
- **BLAS optimization** made using threaded loops around the **vector kernel** [Kim, 2017]. The matrices are split into submatrices sufficiently small to be loaded in **higher level caches**, each thread treats one submatrix at a time with the kernel.
- Tested on a SPD problem of size 14 739 (local balance of momentum on a cube).
- One MPI process on a Xeon Phi KNL with 256 OpenMP threads.
- **Replicated samples** without preconditioner.

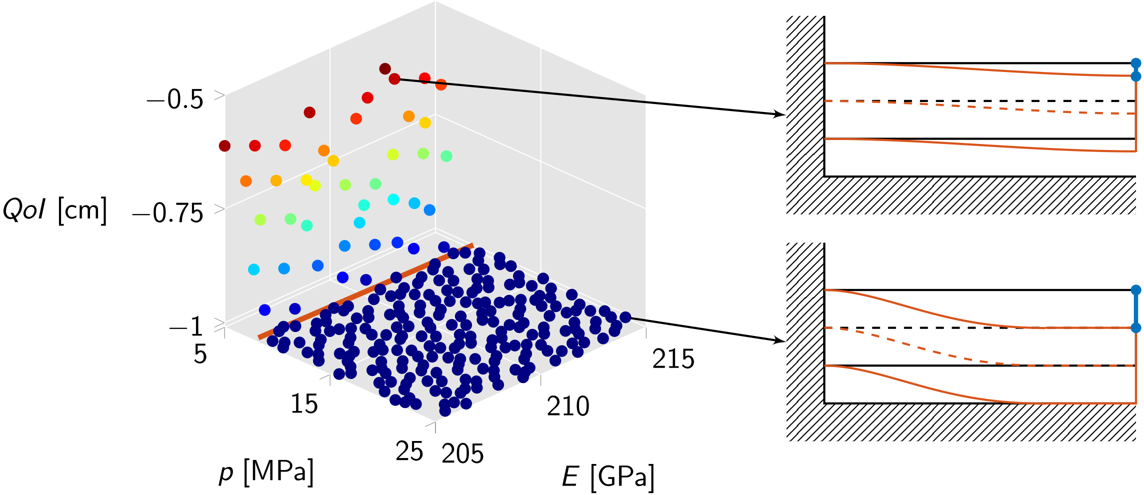


Beam contact problem

- Size: $L = 50\,\text{cm}, W = 5\,\text{cm}, H = 5\,\text{cm}, d = 1\,\text{cm}$,
- Elements: $60 \times 6 \times 6$ hexahedra,
- Number of Dofs: $9\,394 = 3 \times 61 \times 7^2 + 61 \times 7$,
- Depending on the pressure $p \sim \mathcal{U}(5, 25)$ [MPa], the contact is fully open or partially closed.
- Material:
  - Young's modulus: $E \sim \mathcal{U}(205, 215)$ [GPa].
  - Poisson coefficient: 0.29.
- Quantity of Interest: displacement along z on the center point of the face $x = L$,
- 256 Halton Quasi Monte Carlo samples,
- One MPI process on a Xeon Phi KNL with 256 OpenMP threads.

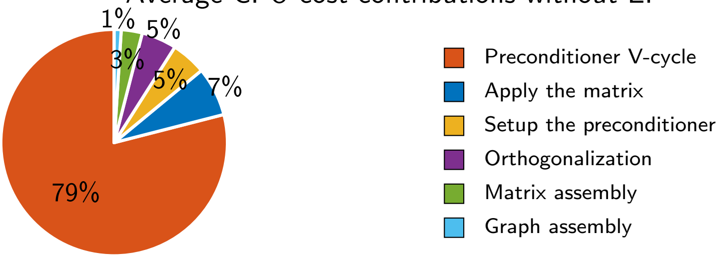


Quantity of Interest

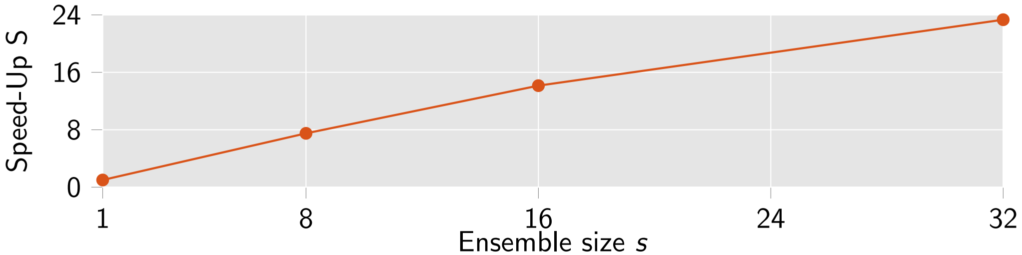


Speed-Up of the preconditioner: main average CPU cost contribution

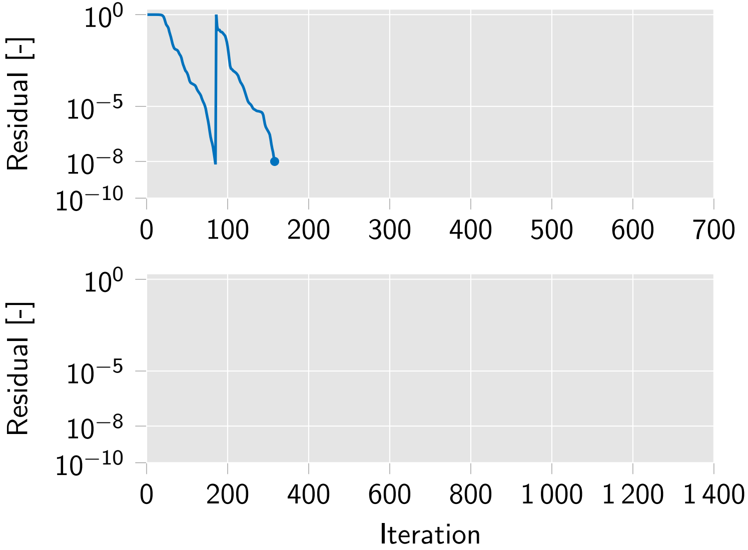
Average CPU cost contributions without EP

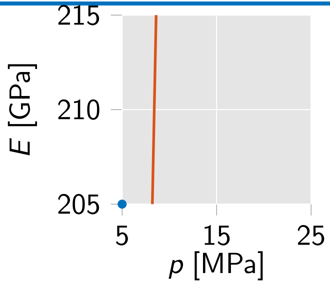


Speed-Up of one preconditioner V-cycle



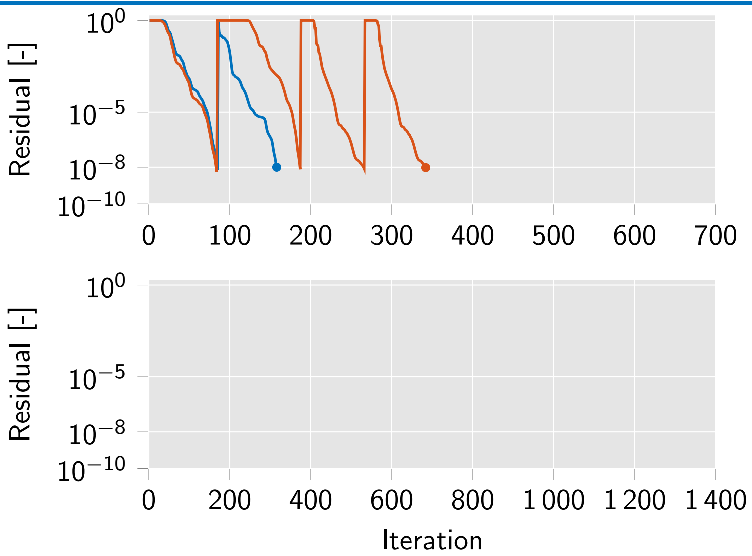
Convergence: reduced norm

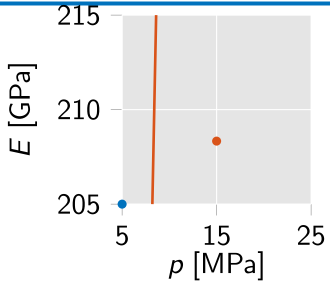




	Active	Iterations	CPU
1	2	158	21.6 s

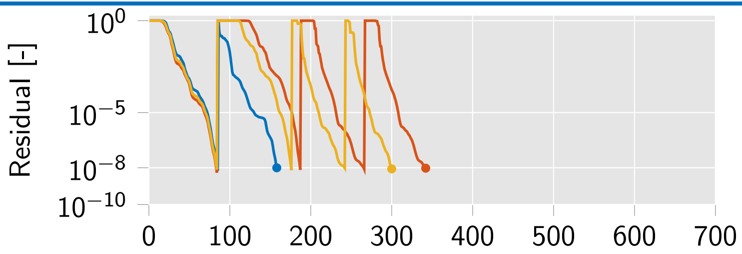
Convergence: reduced norm

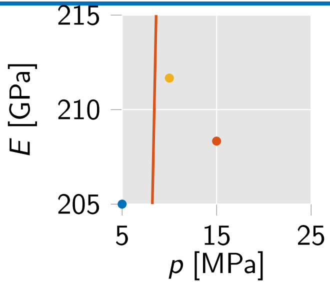


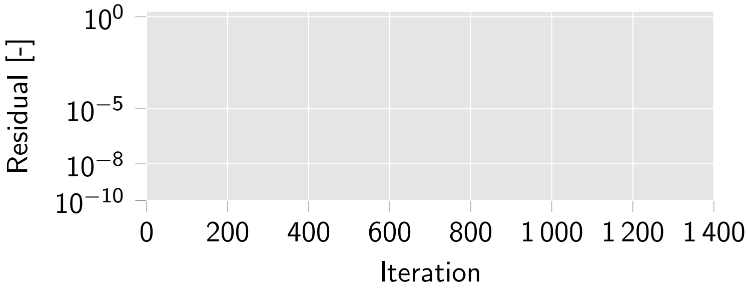


	Active	Iterations	CPU
1	2	158	21.6 s
2	4	342	45.9 s

Convergence: reduced norm

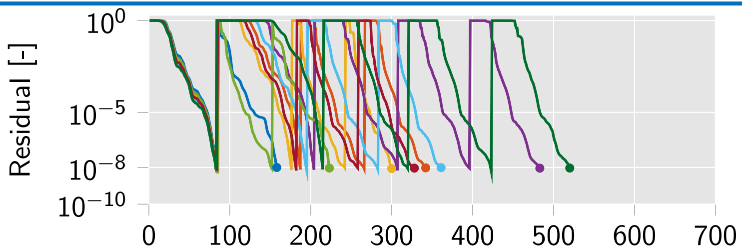


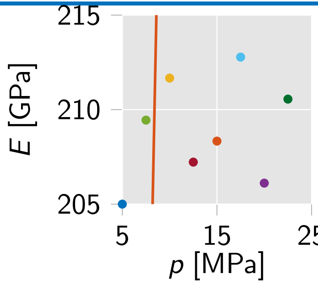


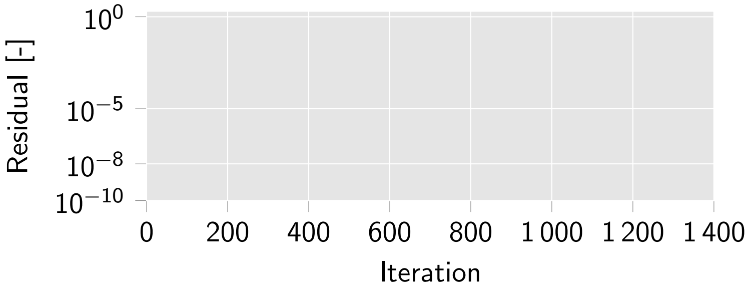


	Active	Iterations	CPU
1	2	158	21.6 s
2	4	342	45.9 s
3	4	300	41.5 s

Convergence: reduced norm

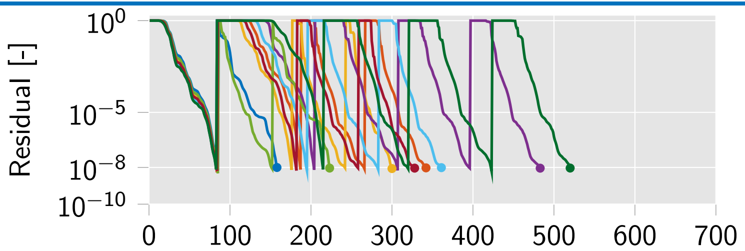


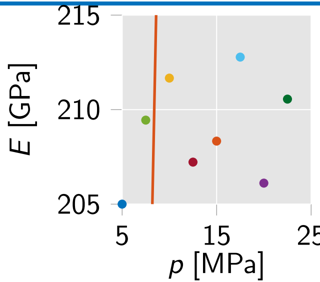


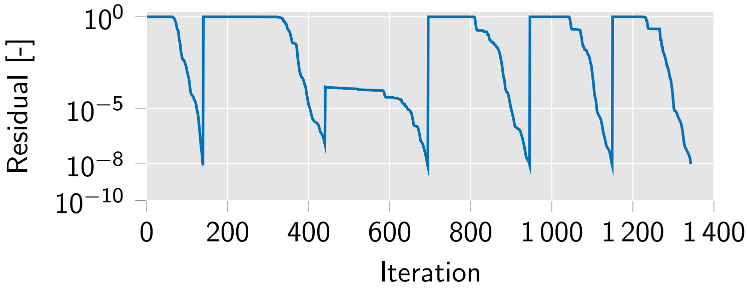


	Active	Iterations	CPU
1	2	158	21.6 s
2	4	342	45.9 s
3	4	300	41.5 s
⋮	⋮	⋮	⋮
8	5	520	72.3 s
Sum	-	2715	375.1 s

Convergence: reduced norm

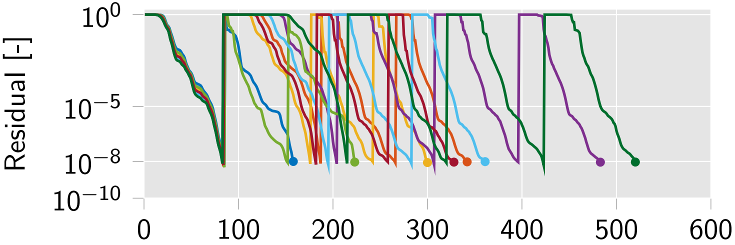


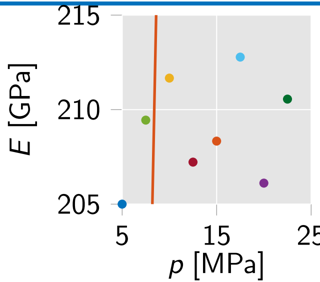


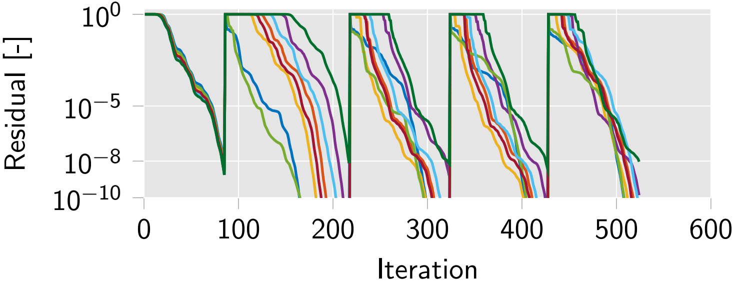


	Active	Iterations	CPU
1	2	158	21.6 s
2	4	342	45.9 s
3	4	300	41.5 s
⋮	⋮	⋮	⋮
8	5	520	72.3 s
Sum	-	2715	375.1 s
EP	5	1343	180.3 s
R	-	3.957	-

Convergence: not reduced norm

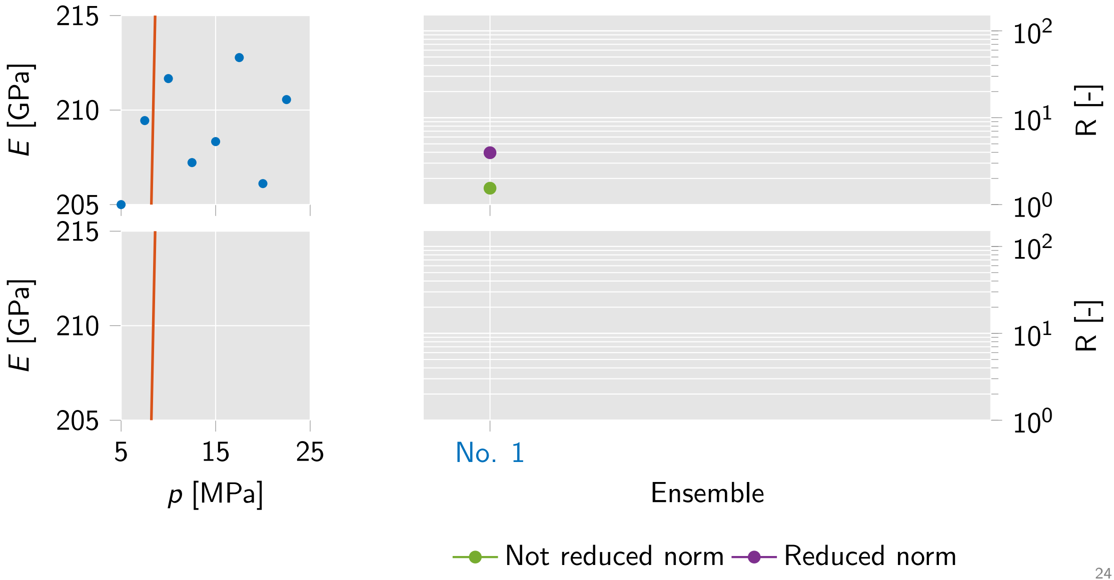




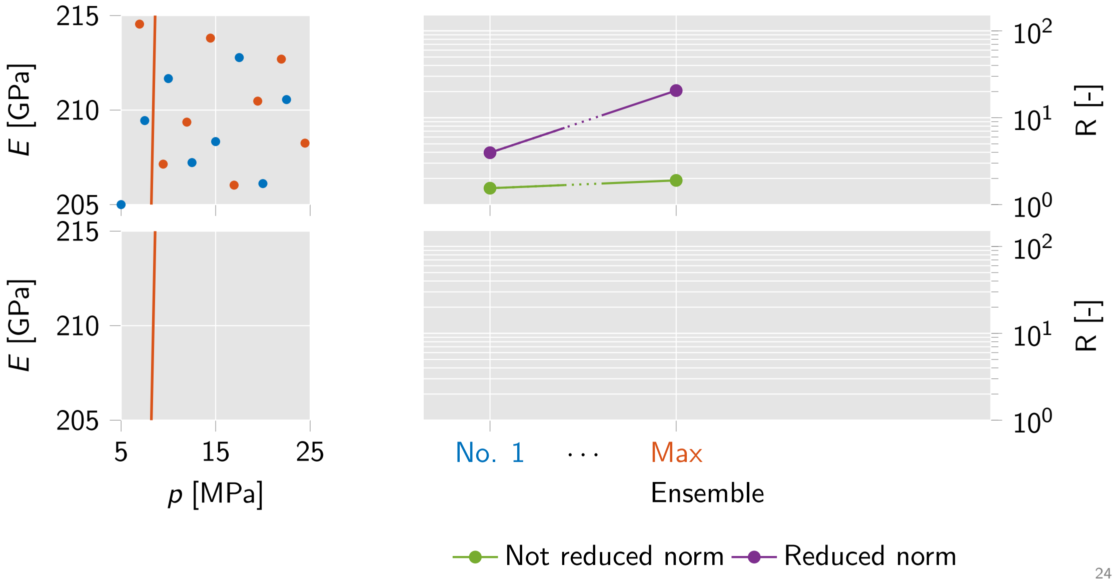


	Active	Iterations	CPU
1	2	158	21.6 s
2	4	342	45.9 s
3	4	300	41.5 s
⋮	⋮	⋮	⋮
8	5	520	72.3 s
Sum	-	2715	375.1 s
EP	5	524	77.7 s
R	-	1.544	-

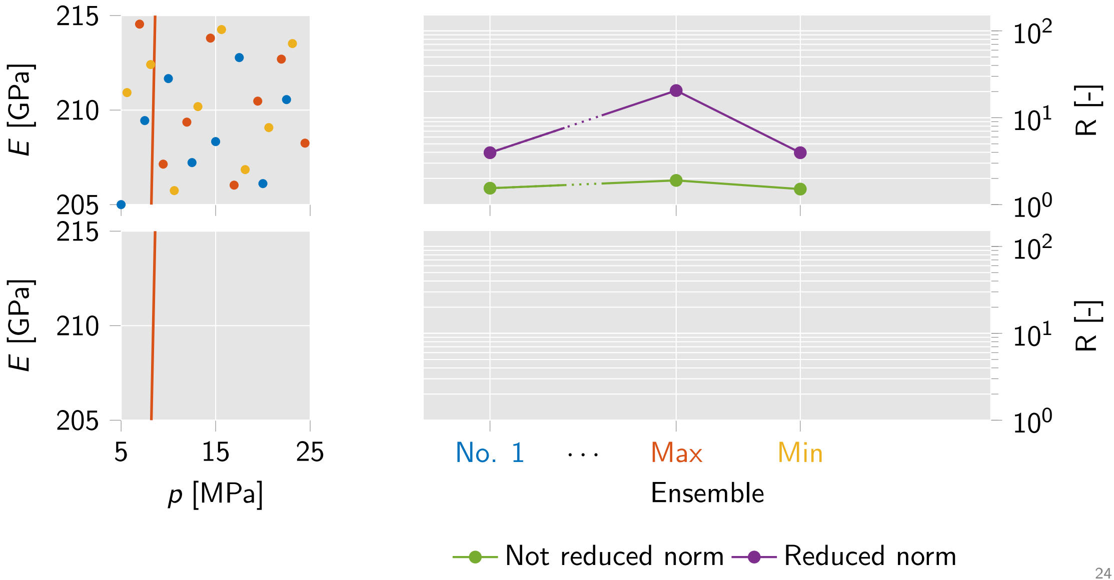
Ensemble size: 8 and 16



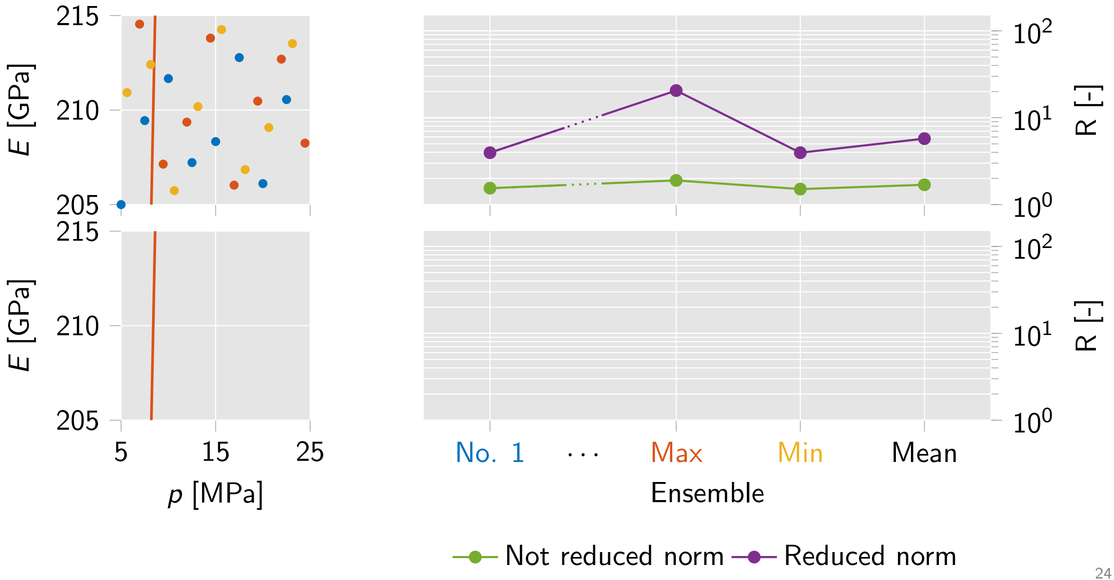
Ensemble size: 8 and 16



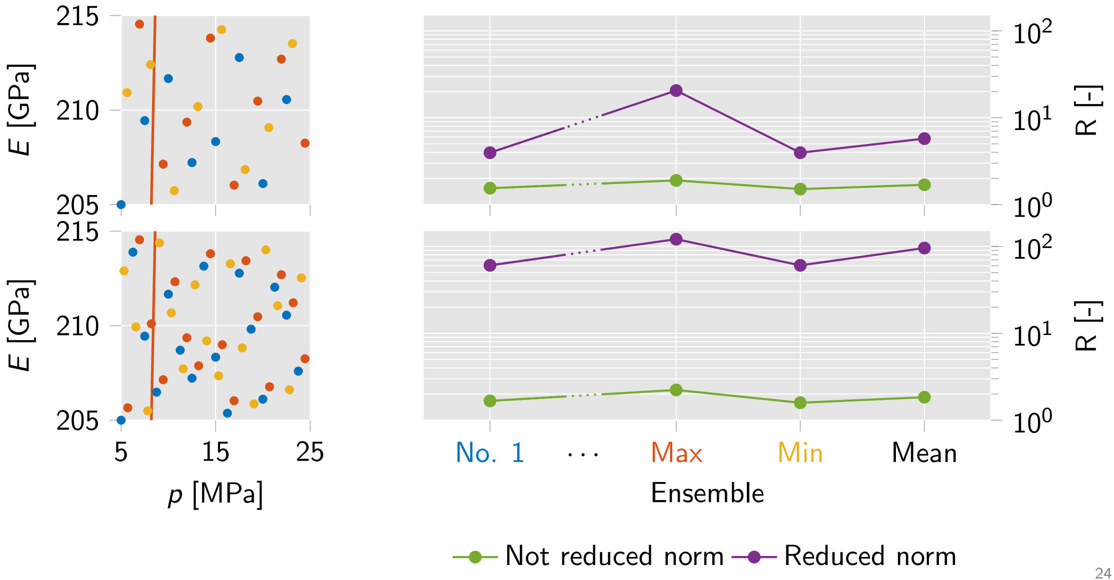
Ensemble size: 8 and 16



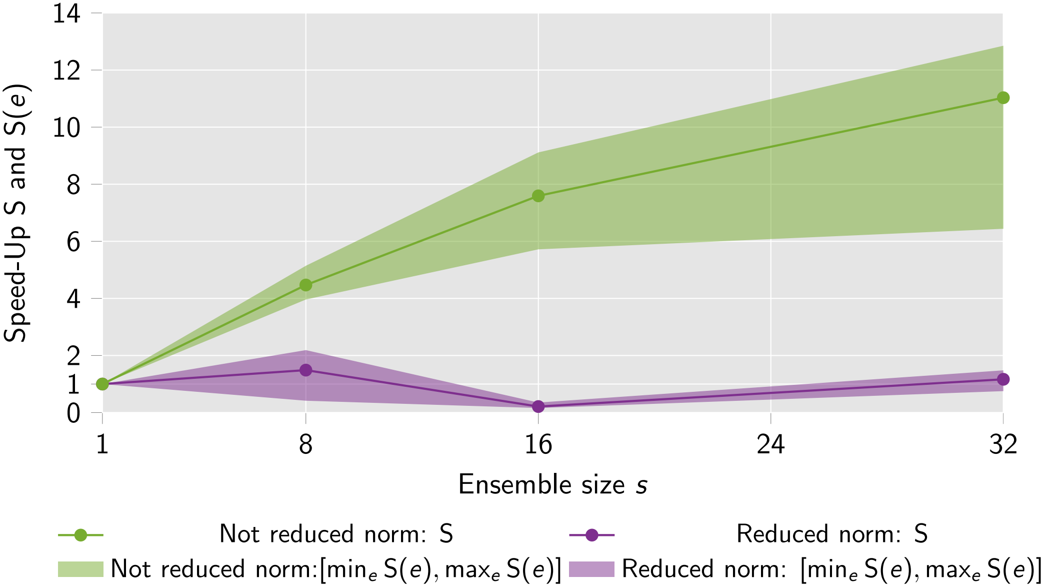
Ensemble size: 8 and 16



Ensemble size: 8 and 16



Speed-Up of the full simulation



Conclusion and future work

Conclusion:

- Contributions towards EP for **mechanical contact problems** including strategy to handle **activities** and the influence of the **norms** on the **GMRES**.
- **Two norms** can currently be used in the **GMRES**: the reduced and the not reduced,
- Promising first results: the choice of the **norm** influences the **performance** and the **precision** of the solutions,
- The convergence of the reduced norm is not already fully understood.

Future work:

- Finish the **optimization** of the **BLAS** implementation for ensemble type,
- Continue to study **theoretically** how the **norm** influences the **convergence** of GMRES,
- Study how to use this method in **uncertainty quantification** of contact problems with **local surrogate model** and **grouping**,
- Apply the method on **engineering problems** relevant for **ITER** in collaboration with FZ. Jülich.

References

- M. D'Elia, E. T. Phipps, A. Rushdiz, and M. S. Ebeida, Surrogate-based Ensemble Grouping Strategies for Embedded Sampling-based Uncertainty Quantification. arXiv preprint arXiv:1705.02003, 2017.
- K. Kim, T. B. Costa, M. Deveci, A. M. Bradley, S. D. Hammond, M. E. Guney, S. Knepper, S. Story, and S. Rajamanickam, Designing vector-friendly compact BLAS and LAPACK kernels. In : Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2017. p. 55.
- E.T. Phipps, M. D'Elia, H C. Edwards, M. Hoemmen, J. Hu, and S. Rajamanickam, Embedded ensemble propagation for improving performance, portability, and scalability of uncertainty quantification on emerging computational architectures. SIAM Journal on Scientific Computing, 2017, vol. 39, no 2, p. C162-C193.
- T. A. Wiesner, M. W. Gee, and W. A. Wall, Algebraic multigrid for contact problems in saddle point formulation, International Journal for Numerical Methods in Engineering, 2015.

Acknowledgement

The first author, Kim Liegeois, would like to acknowledge the Belgian National Fund for Scientific Research (FNRS-FRIA) and the Federation Wallonia-Brussels (FW-B) for their financial support.

fnrs
FREEDOM TO RESEARCH