Wave days in South-West - Pau
March 9-11, 2016

On some parallel linear algebra tools for wave propagation simulations

Luc Giraud

joint work with
Inria HiePACS and Nachos project members

HiePACS Inria Project
Inria Bordeaux Sud-Ouest

Sparse linear solver

Goal: solving $\mathcal{A}x = b$, where $\mathcal{A}$ is sparse

Usual trades off

Direct

- Robust/accurate for general problems
- BLAS-3 based implementations
- Memory/CPU prohibitive for large $3D$ problems
- Limited weak scalability

Iterative

- Problem dependent efficiency / accuracy
- Sparse computational kernels
- Less memory requirements and possibly faster
- Possible high weak scalability

Outline

Hybrid linear solver for Maxwell applications

Block GMRES method with inexact breakdowns and deflated restarting

Hybrid direct-iterative solver with application to Maxwell in the frequency domain

E. Agullo and M. Kuhn (PD)

S. Lanteri and L. Moya (PD)

A. Falco (PhD) and Y. Harness (PD)

S. Nakov and G. Pichon (PhD)

L. Poirel (PhD)

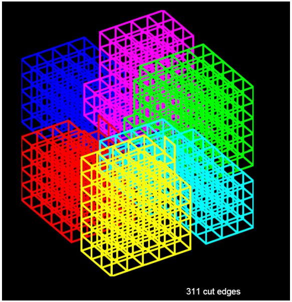
Hybrid Linear Solvers

Develop robust scalable parallel hybrid direct/iterative linear solvers

- Exploit the efficiency and robustness of the sparse direct solvers
- Develop robust parallel preconditioners for iterative solvers
- Take advantage of the natural scalable parallel implementation of iterative solvers

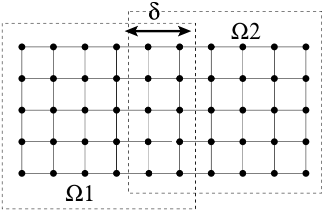
Domain Decomposition (DD)

- Natural approach for PDE's
- Extend to general sparse matrices
- Partition the problem into subdomains, subgraphs
- Use a direct solver on the subdomains
- Robust preconditioned iterative solver



Overlapping Domain Decomposition [H. Schwarz - 1870]

Classical Additive Schwarz preconditioners



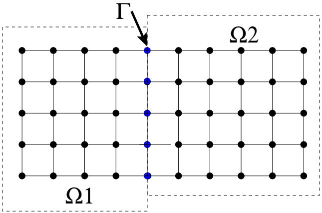
- *Goal:* solve linear system $\mathcal{A}x = b$
- Use iterative method
- Apply the preconditioner at each step
- The convergence rate deteriorates as the number of subdomains increases

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_{1,1} & \mathcal{A}_{1,\delta} & \\ \mathcal{A}_{\delta,1} & \mathcal{A}_{\delta,\delta} & \mathcal{A}_{\delta,2} \\ & \mathcal{A}_{\delta,2} & \mathcal{A}_{2,2} \end{pmatrix} \Longrightarrow \mathcal{M}^{\delta}_{\mathcal{AS}} = \left(\boxed{\begin{matrix} \mathcal{A}_{1,1} & \mathcal{A}_{1,\delta} \\ \mathcal{A}_{\delta,1} & \mathcal{A}_{\delta,\delta} \end{matrix}}^{-1} + \boxed{\begin{matrix} \mathcal{A}_{\delta,\delta} & \mathcal{A}_{\delta,2} \\ \mathcal{A}_{\delta,2} & \mathcal{A}_{2,2} \end{matrix}}^{-1} \right)$$

Classical Additive Schwarz preconditioners N subdomains case

$$\mathcal{M}^{\delta}_{AS} = \sum_{i=1}^{N} \left(\mathcal{R}_i^{\delta}\right)^T \left(\mathcal{A}_i^{\delta}\right)^{-1} \mathcal{R}_i^{\delta}$$

Non-overlapping Domain Decomposition



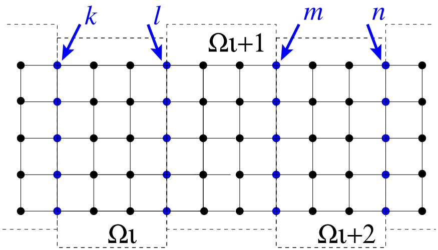
- *Goal:* solve linear system $\mathcal{A}x = b$
- Apply partially Gaussian elimination
- Solve the reduced system $\mathcal{S}x_\Gamma = f$
- Then solve $\mathcal{A}_i x_i = b_i - \mathcal{A}_{i,\Gamma} x_\Gamma$

$$\begin{pmatrix} \mathcal{A}_{1,1} & 0 & \mathcal{A}_{1,\Gamma} \\ 0 & \mathcal{A}_{2,2} & \mathcal{A}_{2,\Gamma} \\ 0 & 0 & \mathcal{S} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_\Gamma \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_\Gamma - \sum_{i=1}^{2} \mathcal{A}_{\Gamma,i} \mathcal{A}_{i,i}^{-1} \; b_i \end{pmatrix}$$

Solve $\mathcal{A}x = b \Longrightarrow$ solve the reduced system $\mathcal{S}x_\Gamma = f \Longrightarrow$ then solve $\mathcal{A}_i x_i = b_i - \mathcal{A}_{i,\Gamma} x_\Gamma$

where $\mathcal{S} = \mathcal{A}_{\Gamma,\Gamma} - \sum_{i=1}^{2} \mathcal{A}_{\Gamma,i} \mathcal{A}_{i,i}^{-1} \mathcal{A}_{i,\Gamma}$, and $f = b_\Gamma - \sum_{i=1}^{2} \mathcal{A}_{\Gamma,i} \mathcal{A}_{i,i}^{-1} \; b_i$.

Distributed Schur complement



$$\Gamma = k \cup \ell \cup m \cup n$$

$$\overbrace{\begin{pmatrix} \mathcal{S}_{kk}^{(\iota)} & \mathcal{S}_{k\ell} \\ \mathcal{S}_{\ell k} & \mathcal{S}_{\ell\ell}^{(\iota)} \end{pmatrix}}^{\Omega_\iota} \quad \overbrace{\begin{pmatrix} \mathcal{S}_{\ell\ell}^{(\iota+1)} & \mathcal{S}_{\ell m} \\ \mathcal{S}_{m\ell} & \mathcal{S}_{mm}^{(\iota+1)} \end{pmatrix}}^{\Omega_{\iota+1}} \quad \overbrace{\begin{pmatrix} \mathcal{S}_{mm}^{(\iota+2)} & \mathcal{S}_{mn} \\ \mathcal{S}_{nm} & \mathcal{S}_{nn}^{(\iota+2)} \end{pmatrix}}^{\Omega_{\iota+2}}$$

In an assembled form: $\mathcal{S}_{\ell\ell} = \mathcal{S}_{\ell\ell}^{(\iota)} + \mathcal{S}_{\ell\ell}^{(\iota+1)} \Longrightarrow \mathcal{S}_{\ell\ell} = \sum_{\iota \in adj} \mathcal{S}_{\ell\ell}^{(\iota)}$

Algebraic Additive Schwarz preconditioner

[L.Carvalho, L.G., G.Meurant - 01]

$$\mathcal{S} = \sum_{i=1}^{N} \mathcal{R}_{\Gamma_i}^T \mathcal{S}^{(i)} \mathcal{R}_{\Gamma_i}$$

$$\mathcal{S} = \begin{pmatrix} \ddots & & & & \\ & \mathcal{S}_{kk} & \mathcal{S}_{k\ell} & & \\ & \mathcal{S}_{\ell k} & \mathcal{S}_{\ell\ell} & \mathcal{S}_{\ell m} & \\ & & \mathcal{S}_{m\ell} & \mathcal{S}_{mm} & \mathcal{S}_{mn} \\ & & & \mathcal{S}_{nm} & \mathcal{S}_{nn} \end{pmatrix} \Rightarrow \mathcal{M} = \begin{pmatrix} \ddots & & & & \\ & \mathcal{S}_{kk} & \mathcal{S}_{k\ell} & ^{-1} & \\ & \mathcal{S}_{\ell k} & \mathcal{S}_{\ell\ell} & \mathcal{S}_{\ell m} & ^{-1} \\ & & \mathcal{S}_{m\ell} & \mathcal{S}_{mm} & \mathcal{S}_{mn} \\ & & & \mathcal{S}_{nm} & \mathcal{S}_{nn} \end{pmatrix}$$

$$\mathcal{M} = \sum_{i=1}^{N} \mathcal{R}_{\Gamma_i}^T (\bar{\mathcal{S}}^{(i)})^{-1} \mathcal{R}_{\Gamma_i}$$

where $\bar{\mathcal{S}}^{(i)}$ is obtained from $\mathcal{S}^{(i)}$

$$\underbrace{\mathcal{S}^{(i)} = \begin{pmatrix} \mathcal{S}_{kk}^{(\iota)} & \mathcal{S}_{k\ell} \\ \mathcal{S}_{\ell k} & \mathcal{S}_{\ell\ell}^{(\iota)} \end{pmatrix}}_{\text{local Schur}} \Rightarrow \underbrace{\bar{\mathcal{S}}^{(i)} = \begin{pmatrix} \mathcal{S}_{kk} & \mathcal{S}_{k\ell} \\ \mathcal{S}_{\ell k} & \mathcal{S}_{\ell\ell} \end{pmatrix}}_{\text{local assembled Schur}}$$

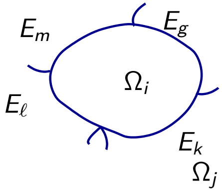
$$\searrow \quad \sum_{\iota \in adj} \mathcal{S}_{\ell\ell}^{(\iota)} \quad \nearrow$$

Similarity with Neumann-Neumann preconditioner [J.F Bourgat, R. Glowinski, P. Le Tallec and M. Vidrascu - 89] [Y.H. de Roek, P. Le Tallec and M. Vidrascu - 91]

Parallel preconditioning features

$$S^{(i)} = A^{(i)}_{\Gamma_i\Gamma_i} - A_{\Gamma_i I_i} A^{-1}_{I_i I_i} A_{I_i \Gamma_i}$$

$$M_{AS} = \sum_{i=1}^{\#domains} R_i^T (\bar{S}^{(i)})^{-1} R_i$$



$$\bar{S}^{(i)} = \begin{pmatrix} S_{mm} & S_{mg} & S_{mk} & S_{m\ell} \\ S_{gm} & S_{gg} & S_{gk} & S_{g\ell} \\ S_{km} & S_{kg} & S_{kk} & S_{k\ell} \\ S_{\ell m} & S_{\ell g} & S_{\ell k} & S_{\ell\ell} \end{pmatrix} \quad S^{(i)} = \begin{pmatrix} S^{(i)}_{mm} & S_{mg} & S_{mk} & S_{m\ell} \\ S_{gm} & S^{(i)}_{gg} & S_{gk} & S_{g\ell} \\ S_{km} & S_{kg} & S^{(i)}_{kk} & S_{k\ell} \\ S_{\ell m} & S_{\ell g} & S_{\ell k} & S^{(i)}_{\ell\ell} \end{pmatrix}$$

Assembled local Schur complement

local Schur complement

$$S_{mm} = \sum_{j \in adj(m)} S^{(j)}_{mm}$$

Parallel implementation

- Each *subdomain* $\mathcal{A}^{(i)}$ is handled by one *processor*

$$\mathcal{A}^{(i)} \equiv \begin{pmatrix} \mathcal{A}_{\mathcal{I}_i\mathcal{I}_i} & \mathcal{A}_{\mathcal{I}_i\Gamma_i} \\ \mathcal{A}_{\mathcal{I}_i\Gamma_i} & \mathcal{A}^{(i)}_{\Gamma\Gamma} \end{pmatrix}$$

- Concurrent partial factorizations are performed on each processor to form the so called "local Schur complement"

$$\mathcal{S}^{(i)} = \mathcal{A}^{(i)}_{\Gamma\Gamma} - \mathcal{A}_{\Gamma_i\mathcal{I}_i}\mathcal{A}^{-1}_{\mathcal{I}_i\mathcal{I}_i}\mathcal{A}_{\mathcal{I}_i\Gamma_i}$$

- The reduced system $\mathcal{S}x_\Gamma = f$ is solved using a distributed Krylov solver
  - One matrix vector product per iteration each processor computes $\mathcal{S}^{(i)}(x_\Gamma^{(i)})^k = (y^{(i)})^k$
  - One local preconditioner apply $(\mathcal{M}^{(i)})(z^{(i)})^k = (r^{(i)})^k$
  - Local neighbor-neighbor communication per iteration
  - Global reduction (dot products)
- Compute simultaneously the solution for the interior unknowns

$$\mathcal{A}_{\mathcal{I}_i\mathcal{I}_i}x_{\mathcal{I}_i} = b_{\mathcal{I}_i} - \mathcal{A}_{\mathcal{I}_i\Gamma_i}x_{\Gamma_i}$$

Current Software software implementation of MaPHyS

Partitioner

- Scotch

Dense direct solver

- Multi-threaded MKL library

Sparse direct solvers

- MUMPS
- Multi-threaded PASTIX

Iterative Solvers

- CG/GMRES/FGMRES using multi-threaded MKL library

Current Software software implementation of MaPHyS

Partitioner

- Scotch

Dense direct solver

- Multi-threaded Mkl library

Sparse direct solvers

- Mumps
- Multi-threaded PaStiX

Iterative Solvers

- Cg/Gmres/FGmres using multi-threaded Mkl library
- Challenge
  - Composability
  - Performance

TECSER project

Goal:

- Novel high performance numerical solution techniques for Radar cross-section computations

Challenges:

- Very large problems, irregular geometric structures, heterogeneous and anisotropic propagation mediums

Solutions:

- Hybridizable Discontinuous Galerkin method (HDGM: Nachos),
- Massively Parallel Hybrid Solver (MaPHyS: HiePACS)

Partners:

Agence Nationale de la Recherche ANR — DGA — AIRBUS GROUP — Inria INVENTEURS DU MONDE NUMÉRIQUE

The HDG method

Attractive features of DG methods

Thanks to the discontinuity DG methods have many advantages

- Easily obtained high order accuracy
- p-adaptivity (approximation is purely local)
- h-adaptivity (conforming or non-conforming grid refinement)
- Natural parallelism

One main drawback of DG methods particularly sensitive for stationary problems

- The excessive number of globally coupled DOFs
  $\Rightarrow$ DG methods are expensive both in terms of CPU time and memory consumption

Hybridization of DG methods is devoted to address this issue while keeping all the advantages of DG methods

The HDG method

The HDG method can be decomposed in two steps

1. A conservativity condition is imposed on the numerical trace, whose definition involved the hybrid variable at the interface between neighboring elements. As result we obtain a global linear system in terms of the DOFs of the hybrid variable.
2. Once the DOFs of the hybrid variable are known, the local values of the electromagnetic fields can be obtained by solving local linear systems element-by-element.

DG vs HDG

Assuming a uniform interpolation degree p, the number of globally coupled DOFs is then

$$\begin{aligned} \text{DG} &: \quad (p+1)(p+2)(p+3)|\mathcal{T}_h|, \\ \text{HDG} &: \quad (p+1)(p+2)|\mathcal{F}_h|. \end{aligned}$$

For a simplicial mesh $|\mathcal{F}_h| \approx 2|\mathcal{T}_h|$, the ratio of the globally coupled DOFs is roughly $2/(p+3)$ for HDG method over DG method.

Propagation of a plane wave in vacuum

- Computational domain: the unit cube $[0,\ 1]^3$
- First order Silver-Müller boundary condition
- Plane wave:
  - Wave vector: $(k_x, k_y, k_z) \simeq (12.6,\ 0.0,\ 0.0)$
  - Polarization: $(0,\ 0,\ 1)$
  - Frequency: $f = 600$ MHz
  - Angular frequency: $\omega = 2\pi f \simeq 12.6$ rad/m
  - Wavelength: $\lambda \simeq 0.4997$ m
- Electromagnetic parameters: $\varepsilon = \mu = 1$ (vacuum)
- Characteristics of the meshes used for numerical convergence:

	# elements	# faces	h
M1	2 692	5 544	0.2500
M2	6 144	12 928	0.1875
M3	12 000	25 000	0.1500
M4	20 736	42 912	0.1250

Propagation of a plane wave in vacuum

Numerical convergence of the HDG method (Error $= \|\mathbf{E} - \mathbf{E}_h\|_2$)

	Error	Order
M1	$7.10\,e^{-02}$	–
M2	$4.27\,e^{-02}$	1.8
M3	$2.85\,e^{-02}$	1.8
M4	$2.03\,e^{-02}$	1.9

HDG-$\mathbb{P}_1$

	Error	Order
M1	$6.78\,e^{-03}$	–
M2	$2.90\,e^{-03}$	2.9
M3	$1.49\,e^{-03}$	3.0
M4	$8.68\,e^{-04}$	3.0

HDG-$\mathbb{P}_2$

	Error	Order
M1	$3.89\,e^{-04}$	–
M2	$1.24\,e^{-04}$	4.0
M3	$5.09\,e^{-05}$	4.0
M4	$2.46\,e^{-05}$	4.0

HDG-$\mathbb{P}_3$

	Error	Order
M1	$2.05\,e^{-05}$	–
M2	$4.89\,e^{-06}$	5.0
M3	$1.61\,e^{-06}$	5.0
M4	$6.48\,e^{-07}$	5.0

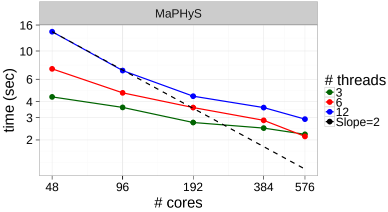
HDG-$\mathbb{P}_4$

$\Rightarrow$ Optimal convergence order (similar results for $\|\mathbf{H} - \mathbf{H}_h\|_2$)

Propagation of a plane wave in vacuum: performances

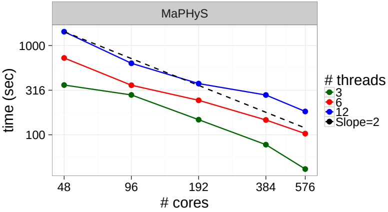
n≃257.5k, nnz≃10.5M

Execution times, interpolation P1

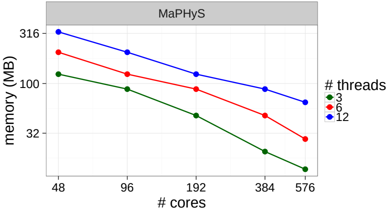


n≃1.3M, nnz≃263.8M

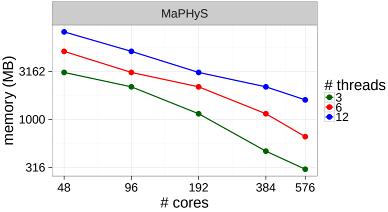
Execution times, interpolation P4



Maximum local memory, interpolation P1

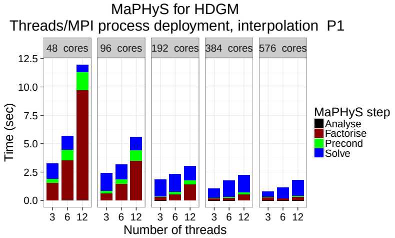


Maximum local memory, interpolation P4

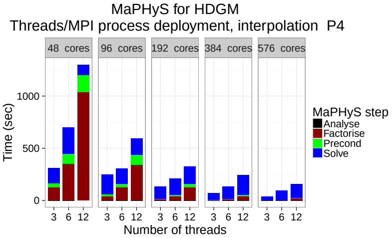


Propagation of a plane wave in vacuum: performances

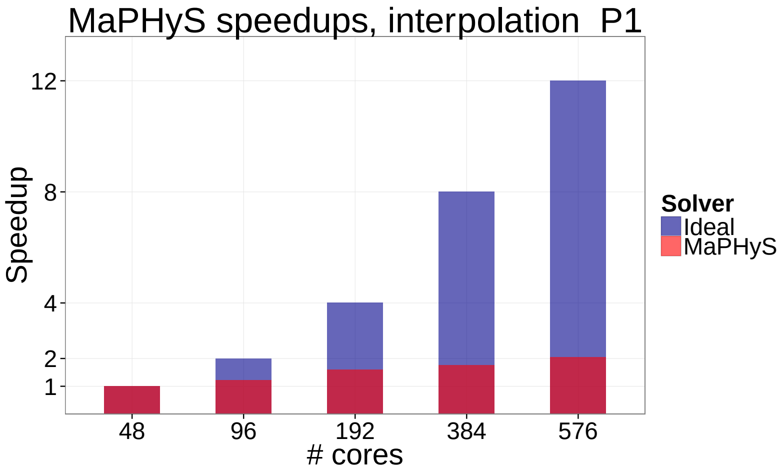
n≃257.5k, nnz≃10.5M



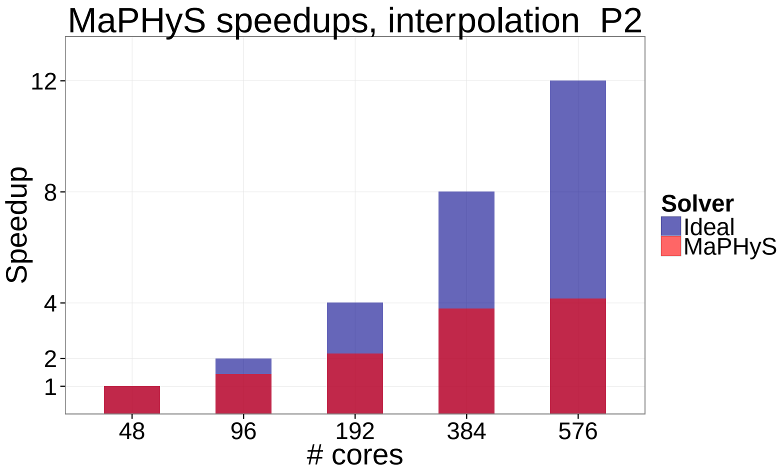
n≃1.3M, nnz≃263.8M



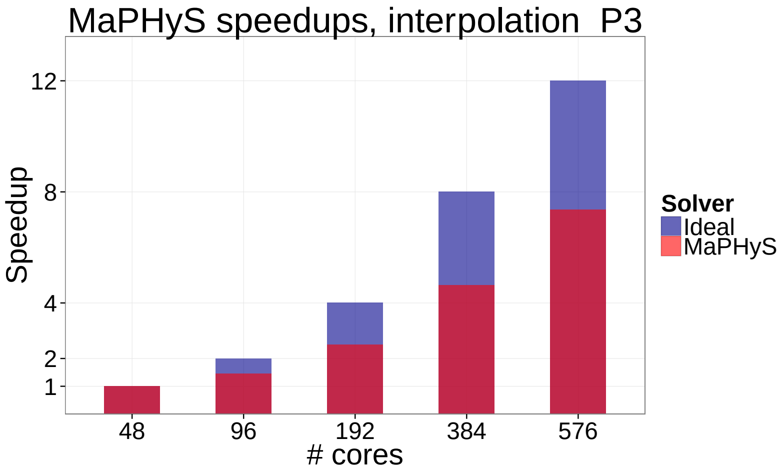
Strong speed-ups



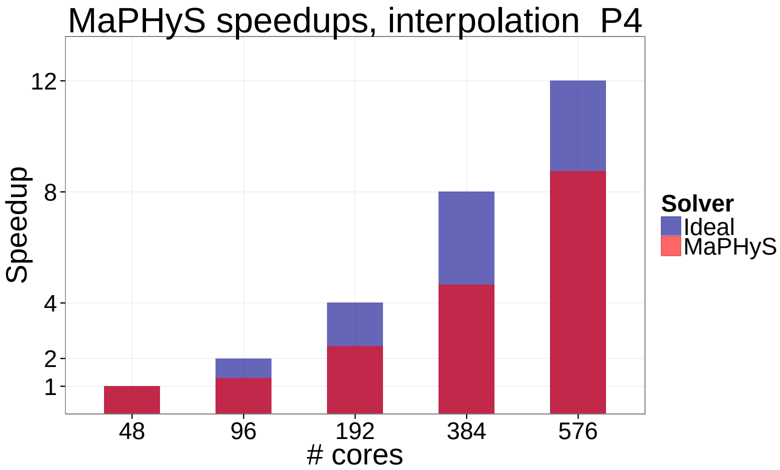
Strong speed-ups



Strong speed-ups



Strong speed-ups



Exposure of head tissues to a plane wave

- Computational domain:
  - sphere of radius $r = 0.3$ m, centered at $(0, 0, 0)$
  - heterogeneous geometrical model of the head tissues (namely, the skin, the skull, the CSF - Cerebro Spinal Fluid and the brain)
- Characteristics of the mesh:
  - 725 136 faces and 361 848 tetrahedra
  - $h_{min} = 0.002$ m and $h_{max} = 0.045$ m
- First order Silver-Müller boundary condition
- Plane wave:
  - Wave vector: $(k_x, k_y, k_z) \simeq (37.7,\ 0.0,\ 0.0)$
  - Polarization: $(0,\ 0,\ 1)$
  - Frequency: $f = 1800$ MHz
  - Angular frequency: $\omega = 2\pi f \simeq 37.7$ rad/m
- Electromagnetic parameters:

	Vacuum	Skin	Skull	CSF	Brain
ε	1.00	38.66	11.60	68.25	43.88
σ (S·m^{-1})	0.00	1.18	0.27	2.28	0.97
λ (mm)	166.67	26.79	48.90	20.16	25.14
ρ	1.00	1 100.00	1 200.00	1 000.00	1 050.00

Exposure of head tissues to a plane wave

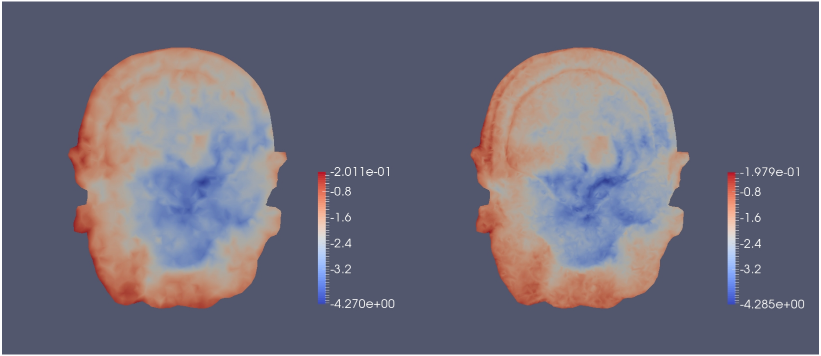
- Statistics of the global matrix

	Matrix order	nnz
HDG-$\mathbb{P}_1$	$4.3M$	$184M$
HDG-$\mathbb{P}_2$	$8.7M$	$736M$
Nonzero per row: 42 ($\mathbb{P}_1$), 84 ($\mathbb{P}_2$)		

- Value of interest the SAR (Specific Absorption Rate)

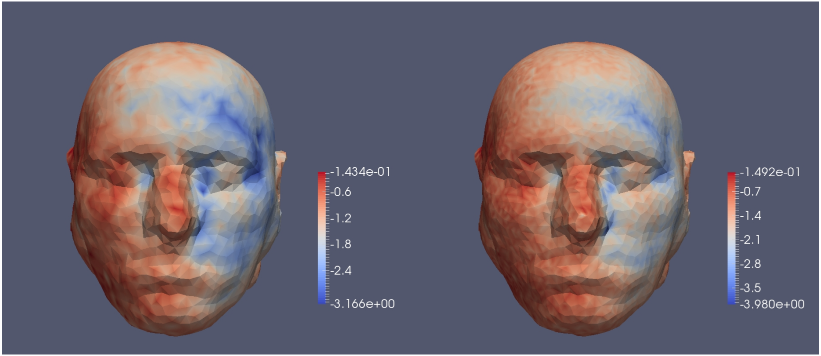
  The SAR is a measure of the rate at which electric energy is absorbed by the tissues when exposed to a radio-frequency electromagnetic field. For instance, it involved in the definition of international norms for mobiles phones. This quantity represents the power absorbed per mass of tissues and has units of watts per kilogram (W$\cdot$kg^{-1}), it is defined by $\sigma|\mathbf{E}|^2/\rho$

Exposure of head tissues to a plane wave



Contour lines of the local SAR over the maximal local SAR (logarithmic scale), HDG-$\mathbb{P}_1$ - HDG-$\mathbb{P}_2$ methods (left - right)

Exposure of head tissues to a plane wave

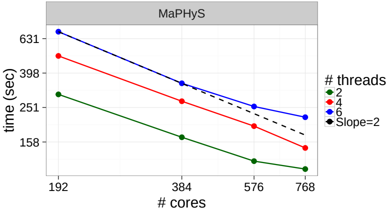


Contour lines of the local SAR over the maximal local SAR (logarithmic scale), HDG-$\mathbb{P}_1$ - HDG-$\mathbb{P}_2$ methods (left - right)

Exposure of head tissues to a plane wave: performances

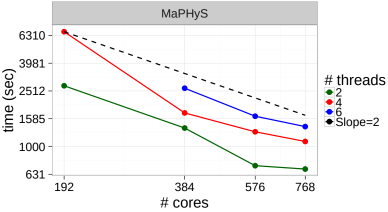
n≃4.4M, nnz≃184.1M

Execution times, interpolation P1

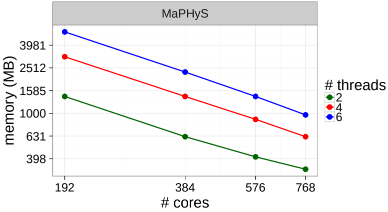


n≃8.7M, nnz≃736.3M

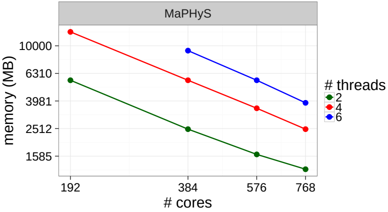
Execution times, interpolation P2



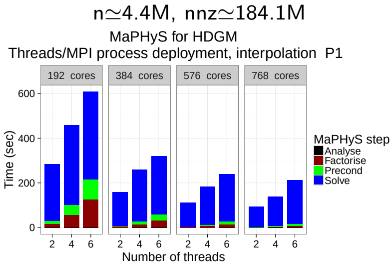
Maximum local memory, interpolation P1

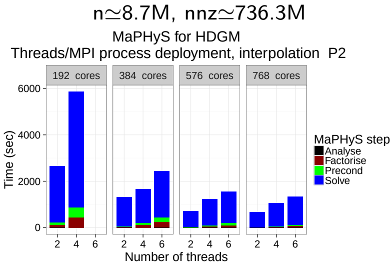


Maximum local memory, interpolation P2

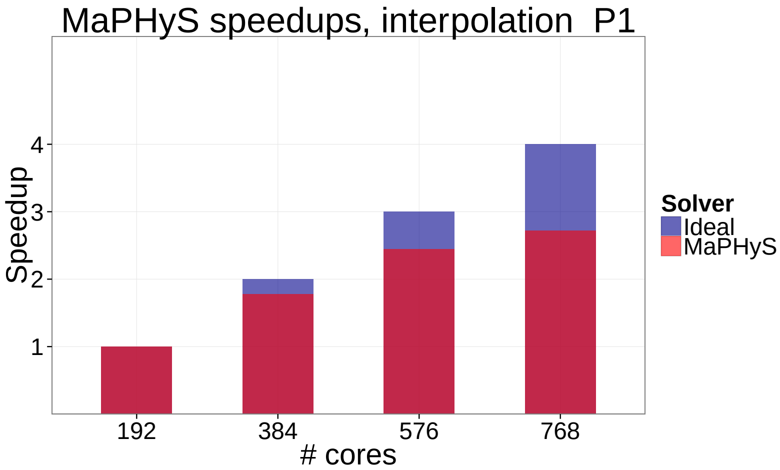


Exposure of head tissues to a plane wave: performances

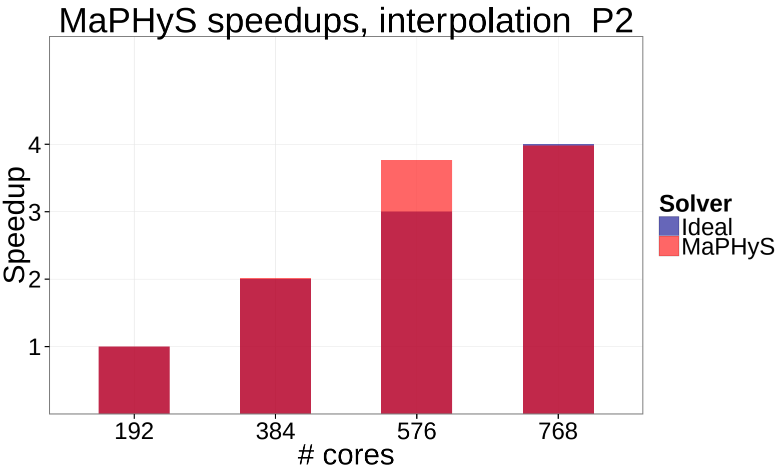




Strong speed-ups



Strong speed-ups



Related activities

- Recent/Ongoing efforts
  1. Partioning/balancing both interface and interior vertices (A. Cassadei)
  2. Parallel analysis and FEM API (M. Kuhn)
  3. Deflation/augmentation via local spectral calculation (L. Poirel)
  4. $\mathcal{H}$-arithmetic for local solve ($\mathcal{H}$-PasTiX) and preconditioner (A. Falco, G. Pichon, Y. Harness)
  5. Numerical resilience policies (M. Zounon)
  6. Experiments on large 3D elastodynamic problems (S. Nakov - Magique 3D)
- Future step: Full task based implemenation on top of runtime systems

Block Krylov linear solver

E. Agullo and C. Piacibello

Y.F. Jing, Chengdu University, China

Some basic ingredients in classical GMRES - $Ax = b$

$$x_\ell = \underset{z \in \mathcal{K}_\ell(b,A)}{\operatorname{argmin}} \|b - Az\|_2$$

with $\mathcal{K}_\ell(b, A) = \text{span}(b, Ab, ..., A^{\ell-1}b)$:

1. Construction of an orthonormal basis of the Krylov space
2. Minimum norm solution

Computational facts

1. Happy breakdown
2. Simple restarting mechanism

Construction of the orthonormal basis

Arnoldi with modified Gram-Schmidt orthogonalization

$\beta = \|b\|$ set $v_1 = b/\beta$
for $j = 1, 2, \ldots, m$ **do**
    Compute $w_j = Av_j$
    for $i = 1, 2, \ldots, j$ **do**
        $h_{i,j} = v_i^H w_j$
        $w_j = w_j - v_i h_{i,j}$
    end for
    $w_j = v_{j+1} h_{j+1,j}$
end for

Key equalities :

$$AV_j = V_j H_j + [0_{n\times(j-1)}, w_j] = V_{j+1} \underline{H}_j$$

$$\text{with } V_j^H V_j = I_j \text{ and } V_{j+1}^H V_{j+1} = I_{j+1} \text{ where } V_j = [v_1, ..., v_j]$$

Minimum norm solution

- What we want

$$x_\ell = \underset{z \in \mathcal{K}_\ell(b,A)}{\operatorname{argmin}} \|b - Az\|_2 \quad x_\ell = V_\ell y_\ell$$

- Key equality

$$\begin{aligned} \|b - Ax_\ell\| &= \quad \|b - AV_\ell y_\ell\| = \|b - V_{\ell+1}\underline{H}_\ell y_\ell\| \\ &= \|V_{\ell+1}(\beta e_1 - \underline{H}_\ell y_\ell)\| = \|\beta e_1 - \underline{H}_\ell y_\ell\| \end{aligned}$$

- Key features that make it works
  1. Arnoldi equality $AV_\ell = V_{\ell+1}\underline{H}_\ell$
  2. Orthonormal basis $V_{\ell+1}^H V_{\ell+1} = I_{\ell+1}$
  3. Right-hand side in search space $b \in \text{span}(V_{\ell+1})$

Happy breakdown

This situation occurs when $w_j = 0$ in Arnoldi, meaning the algorithm cannot extend the space

$$AV_j = V_j H_j + [0_{n\times(j-1)}, w_j] = V_j H_j$$

Consequences

- Happy breakdown: the solution $x \in \text{span}(V_j)$
- b can be expressed as a linear combination of j eigenvectors
  Remark: all eigenvectors are not revealed at the same speed in the Krylov space (argument will come back later)

Basic restart mechanism

- Computation per iteration and storage increase linearly with iteration
- Restart mechanism when maximum search space dimension m is attained
- Set $x_0 = x_m$, solve

$$Ae = r_0$$

using GMRES where $r_0 = b - Ax_0$ so that $x_j \in x_0 + \mathcal{K}_j(r_0, A)$
Remark: all spectral information captured in the Krylov space is lost at restart

Some key ingredients for block GMRES - $AX = B$

$$X_\ell = \underset{Z \in \mathcal{K}_\ell(V_1, A)}{\operatorname{argmin}} \|B - AZ\|_F$$

with $\mathcal{K}_\ell(V_1, A) = \text{span}(V_1, AV_1, ..., A^{\ell-1}V_1)$:

1. Construction of an orthonormal basis of the Krylov space, where $B = V_1\Lambda_1$ is the reduced QR factorisation of B
2. Minimum residual norm solution

Computational challenges

1. Numerical deficiency in W_j - inexact breakdown [Robbé, Sadkane]
2. More sophisticated restarting mechanism [R. Morgan]

Construction of the orthonormal basis

Arnoldi with modified Gram-Schmidt orthogonalization

Choose a unitary matrix V_1 of size $n \times p$
for $j = 1, 2, \ldots, m$ **do**
  Compute $W_j = AV_j$
  for $i = 1, 2, \ldots, j$ **do**
    $H_{i,j} = V_i^H W_j$
    $W_j = W_j - V_i H_{i,j}$
  end for
  $W_j = V_{j+1} H_{j+1,j}$ (reduced $QR-$factorization)
end for

$$A\mathscr{V}_j = \mathscr{V}_j \mathscr{H}_j + [0_{n \times n_{j-1}}, \quad W_j] = \mathscr{V}_{j+1} \underline{\mathscr{H}}_j$$

with $\mathscr{V}_{j+1}^H \mathscr{V}_{j+1} = I_{n_{j+1}}$ where $\mathscr{V}_{j+1} = [V_1, \ldots, V_{j+1}]$

Minimun norm solution

$$\|B - AX_j\|_F = \min_{Y \in \mathbb{C}^{n_j \times p}} \left\| \mathscr{V}_{j+1} \left(\Lambda_j - \underline{\mathscr{H}}_j Y \right) \right\|_F = \min_{Y \in \mathbb{C}^{n_j \times p}} \left\| \Lambda_j - \underline{\mathscr{H}}_j Y \right\|_F$$

because $\mathscr{V}_{j+1}$ forms an orthonormal basis and

$$\Lambda_j = \begin{bmatrix} \Lambda_1 \\ 0 \end{bmatrix} \in \mathbb{C}^{n_{j+1} \times p}$$

Remark: we minimize the Frobenius norm of the block that translates in 2-norm for the individual column residual

Numerical rank deficiency in W_j

- For reasons to be made clear later but related to stopping criterion we decompose

$$W_j = V_{j+1} H_{j+1,j} + Q_j$$

  with $(Q_j \perp V_{j+1}) \perp \mathscr{V}_j$.
  We still have

$$A\mathscr{V}_j = \mathscr{V}_j \mathscr{H}_j + [\mathscr{Q}_{j-1}, \quad W_j],$$

  where $\mathscr{Q}_{j-1} = [Q_1, \ldots, Q_{j-1}] \in \mathbb{C}^{n \times n_{j-1}}$ accounts for all the abandoned directions.
- To characterize a minimum norm solution in $\mathscr{V}_j$ we need to have an orthonormal basis of $[\mathscr{V}_j, \mathscr{Q}_{j-1}, \quad W_j]$ so that

$$A\mathscr{V}_j = \left[\mathscr{V}_j, [P_{j-1}, \tilde{W}_j]\right] \mathscr{F}_j$$

Shortcut for deriving the extended Arnoldi equality I

[M. Robbé and M. Sadkane, LAA, 2006]

$$A\mathscr{V}_j = \mathscr{V}_j\mathscr{H}_j + [\mathcal{Q}_{j-1}, \quad W_j]$$

- $\tilde{\mathcal{Q}}_{j-1} = (I - \mathscr{V}_j{\mathscr{V}_j}^H)\mathcal{Q}_{j-1}$, $\mathscr{L}_j = \mathscr{H}_j + {\mathscr{V}_j}^H \left[\mathcal{Q}_{j-1}, 0_{p_j}\right]$ (~~Hessenberg~~)
- $\mathcal{Q}_{j-1}$ is low rank so is $\tilde{\mathcal{Q}}_{j-1} = P_{j-1}G_{j-1}$

  $$\begin{cases} P_{j-1} \in \mathbb{C}^{n\times \tilde{q}_{j-1}} \text{ has orthonormal columns with } {\mathscr{V}_j}^H P_{j-1} = 0, \\ G_{j-1} \in \mathbb{C}^{\tilde{q}_{j-1}\times n_{j-1}} \text{ is of full rank.} \end{cases}$$

- W_j orthogonalized against P_{j-1} with $W_j - P_{j-1}C_j$ where $C_j = P_{j-1}^H W_j$
- $\tilde{W}_j D_j = \text{QR}\ (W_j - P_{j-1}C_j)$.
- $[\mathscr{V}_j, P_{j-1}, \tilde{W}_j]$ form an orthonormal basis of $[\mathscr{V}_j, \mathcal{Q}_{j-1}, W_j]$.

Shortcut for deriving the generalized Arnoldi equality II

[M. Robbé and M. Sadkane, LAA, 2006]

- Extended Arnoldi equality

$$
\begin{aligned}
A\mathscr{V}_j = & \ \mathscr{V}_j\mathscr{L}_j + \left[P_{j-1}G_{j-1}, \left[P_{j-1}, \tilde{W}_j\right] \begin{bmatrix} C_j \\ D_j \end{bmatrix}\right] \\
= & \ \left[\mathscr{V}_j, P_{j-1}, \tilde{W}_j\right] \begin{bmatrix} & \mathscr{L}_j & \\ G_{j-1} & & C_j \\ 0 & & D_j \end{bmatrix} \\
= & \ \left[\mathscr{V}_j, [P_{j-1}, \tilde{W}_j]\right] \mathscr{F}_j
\end{aligned}
$$

- Least-squares problem reads

$$
Y_j = \operatorname*{argmin}_{Y \in \mathbb{C}^{n_j \times p}} \left\| \Lambda_j - \mathscr{F}_j Y \right\|_F, \text{with } \Lambda_j = \begin{bmatrix} \Lambda_1 \\ 0 \\ 0 \end{bmatrix}
$$

Numerical rank deficiency in $\tilde{W}_j$ vs convergence

- Based on SVD of least-squared residual

$$\Lambda_j - \mathscr{F}_j Y_j = \mathbb{U}_1 \Sigma_1 \mathbb{V}_1^H + \mathbb{U}_2 \Sigma_2 \mathbb{V}_2^H \text{ with } \epsilon^{(R)} \leq ||\Sigma_1||$$

- Decompose

$$\mathbb{U}_1 = \begin{pmatrix} \mathbb{U}_1^{(1)} \\ \mathbb{U}_1^{(2)} \end{pmatrix} \text{ in accordance with } \left[\mathscr{V}_j, [P_{j-1}, \tilde{W}_j]\right]$$

- Consider $[\mathbb{W}_1, \mathbb{W}_2]$ unitary so that $\text{Range}(\mathbb{W}_1) = \text{Range}(\mathbb{U}_1^{(2)})$
- Define and update

$$V_{j+1} = \left[P_{j-1}, \tilde{W}_j\right] \mathbb{W}_1$$

$$P_j = \left[P_{j-1}, \tilde{W}_j\right] \mathbb{W}_2$$

$$G_j = \mathbb{W}_2^H \begin{bmatrix} G_{j-1} & C_j \\ 0 & D_j \end{bmatrix}$$

Rank deficiency threshold vs stopping criterion

Assuming p inexact breakdowns

$$||\Lambda_\ell - \mathscr{F}_\ell Y_\ell|| = ||B - AX_\ell||_2 \le \epsilon^{(R)}$$

$$\frac{||b^{(i)} - Ax_\ell^{(i)}||_2}{||b^{(i)}||_2} \le \frac{||B - AX_\ell||_2}{||b^{(i)}||_2} \le \frac{||B - AX_\ell||_2}{\min\limits_{i=1,\ldots,p} \left\|b^{(i)}\right\|_2} \le \frac{\epsilon^{(R)}}{\min\limits_{i=1,\ldots,p} \left\|b^{(i)}\right\|_2}$$

It follows that the choice

$$\epsilon^{(R)} = \varepsilon \times \min_{i=1,\ldots,p} \left\|b^{(i)}\right\|_2$$

ensures convergence below the threshold ϵ for individual $b^{(i)}$ if same accuracy required for all the righ-hand sides

A few definitions

Definition

Harmonic Ritz pair. Consider a subspace $\mathcal{U}$ of $\mathbb{C}^n$. Given a matrix $B \in \mathbb{C}^{n\times n}$, $\lambda \in \mathbb{C}$ and $y \in \mathcal{U}$, (λ, y) is a harmonic Ritz pair of A with respect to $\mathcal{U}$ if and only if

$$Ay - \lambda\, y \perp A\mathcal{U}$$

The vector y is a harmonic Ritz vector associated with the harmonic Ritz value λ.

Lemma

The harmonic Ritz pairs $(\tilde{\theta}_i, \tilde{g}_i)$ associated with $\mathcal{U} = \mathrm{span}(\mathscr{V}_m)$ satisfy the following property

$$\mathscr{F}_m^H \left(\mathscr{F}_m \tilde{g}_i - \tilde{\theta}_i \begin{bmatrix} \tilde{g}_i \\ 0_p \end{bmatrix} \right) = 0, \ \ (i = 1, \ldots, n_m),$$

$\tilde{g}_i \in \mathbb{C}^{n_m}$, and $\mathscr{V}_m \tilde{g}_i$ are the harmonic Ritz vectors associated with the corresponding harmonic Ritz values $\tilde{\theta}_i$.

An interesting fact for augmentation at restart

Lemma

Assume that $\mathscr{L}_m$ is of full rank after performing a first cycle of IB-BGMRES, then the column vectors $\left(\mathscr{F}_m\tilde{g}_i - \tilde{\theta}_i \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix}\right) \in \mathbb{C}^{n_m+p}$ *($i = 1, \ldots, n_m$) are all contained in the subspace spanned by the least-squares residuals* $R_{LS_m} = (\Lambda_m - \mathscr{F}_m Y_m) \in \mathbb{C}^{(n_m+p)\times p}$*, i.e.,* $\exists \alpha_i \in \mathbb{C}^p$ *so that*

$$\mathscr{F}_m\tilde{g}_i - \tilde{\theta}_i \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix} = R_{LS_m}\alpha_i.$$

Proposition

The harmonic residual vectors are all linear combinations of the residual vectors from the minimum residual solutions of the linear equation problem after performing a first cycle of the IB-BGMRES.

An interesting fact for augmentation at restart

Lemma

Assume that $\mathscr{L}_m$ is of full rank after performing a first cycle of IB-BGMRES, then the column vectors $\left(\mathscr{F}_m\tilde{g}_i - \tilde{\theta}_i \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix}\right) \in \mathbb{C}^{n_m+p}$ $(i = 1, \ldots, n_m)$ are all contained in the subspace spanned by the least-squares residuals $R_{LS_m} = (\Lambda_m - \mathscr{F}_m Y_m) \in \mathbb{C}^{(n_m+p)\times p}$, i.e., $\exists \alpha_i \in \mathbb{C}^p$ so that

$$\mathscr{F}_m\tilde{g}_i - \tilde{\theta}_i \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix} = R_{LS_m}\alpha_i.$$

Proposition

The harmonic residual vectors are all linear combinations of the residual vectors from the minimum residual solutions of the linear equation problem after performing a first cycle of the IB-BGMRES.

Some harmonic vectors can be kept in the search space at restart with the residual vector that must be in the space

Restarting mechanism I

Let $\tilde{G} = [\tilde{g}_1, \ldots, \tilde{g}_k] \in \mathbb{C}^{n_m \times k}$ and form $\underline{G} = \begin{bmatrix} \tilde{G} & R_{LS_m} \\ 0_{p\times k} & \end{bmatrix}$ We denote $\underline{G} = Q_{\underline{G}} R_{\underline{G}}$ the reduced $QR-$factorization of $\underline{G}$,

$$Q_{\underline{G}} = \begin{bmatrix} \Gamma_1 & \Gamma_2 \\ 0_{p\times k} & \end{bmatrix} \in \mathbb{C}^{(n_m+p)\times(k+p)},$$
$$R_{\underline{G}} = \begin{bmatrix} \Theta_1 & \Theta_2 \\ 0_{p\times k} & \end{bmatrix} \in \mathbb{C}^{(k+p)\times(k+p)},$$

so that

$$\tilde{G} = \Gamma_1 \Theta_1,$$
$$R_{LS_m} = Q_{\underline{G}} \Theta_2$$

We can define an orthonormal basis for the restarting search space that contains spectral information

$$\mathscr{V}_1^{\text{new}} = \mathscr{V}_m \Gamma_1$$

and an orthonormal encompassing basis that contains the residuals

$$\left[\mathscr{V}_1^{\text{new}}, [P_0, \tilde{W}_1]^{\text{new}}\right] = \left[\mathscr{V}_m, [P_{m-1}, \tilde{W}_m]\right] Q_{\underline{G}}$$

Restarting mechanism II

Extended Arnoldi relation

$$A\mathscr{V}_1^{\text{new}} = \left[\mathscr{V}_1^{\text{new}}, [P_0, \tilde{W}_1]^{\text{new}}\right] \mathscr{F}_1^{\text{new}} \quad A\mathscr{V}_1^{\text{new}} = \mathscr{V}_2^{\text{new}} \underline{\mathscr{L}}_1^{\text{new}} + \tilde{\mathcal{Q}}_1^{\text{new}},$$

with

$$\left[\mathscr{V}_1^{\text{new}}, [P_0, \tilde{W}_1]^{\text{new}}\right] = \left[\mathscr{V}_m, [P_{m-1}, \tilde{W}_m]\right] Q_{\underline{G}}, \quad R_0 = \left[\mathscr{V}_1^{\text{new}}, [P_0, \tilde{W}_1]^{\text{new}}\right] \Lambda_1^{\text{new}} \text{ with } \Lambda_1^{\text{new}} = \Theta_2,$$

$$\mathscr{L}_1^{\text{new}} = \Gamma_1^H \mathscr{L}_m \Gamma_1, \quad \mathbb{H}_1^{\text{new}} = \Gamma_2^H \mathscr{F}_m \Gamma_1, \quad \mathscr{F}_1^{\text{new}} = \begin{bmatrix} \mathscr{L}_1^{\text{new}} \\ \mathbb{H}_1^{\text{new}} \end{bmatrix},$$

$$V_2^{\text{new}} = [P_0, \tilde{W}_1]^{\text{new}} \mathbb{W}_1^{\text{new}}, \quad P_1^{\text{new}} = [P_0, \tilde{W}_1]^{\text{new}} \mathbb{W}_2^{\text{new}}, \quad \mathscr{L}_{2,:}^{\text{new}} = \mathbb{W}_1^{\text{new}\,H} \mathbb{H}_1^{\text{new}}, \quad G_1^{\text{new}} = \mathbb{W}_2^{\text{new}\,H} \mathbb{H}_1^{\text{new}},$$

$$\mathscr{V}_2^{\text{new}} = \left[\mathscr{V}_1^{\text{new}}, V_2^{\text{new}}\right], \quad \underline{\mathscr{L}}_1^{\text{new}} = \begin{bmatrix} \mathscr{L}_1^{\text{new}} \\ \mathscr{L}_{2,:}^{\text{new}} \end{bmatrix}, \quad \tilde{\mathcal{Q}}_1^{\text{new}} = P_1^{\text{new}} G_1^{\text{new}},$$

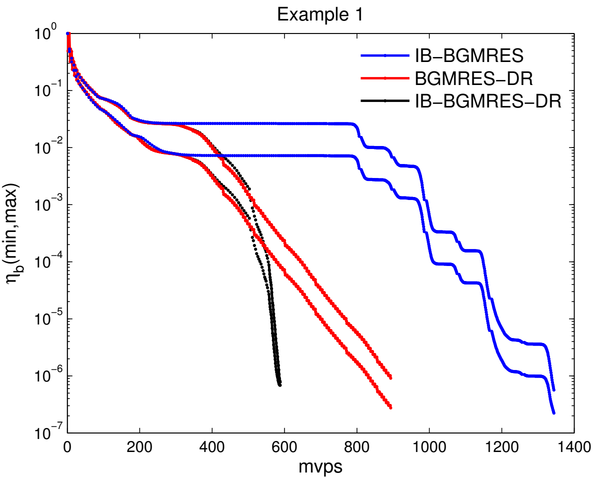
where $\text{Range}(\mathbb{W}_1^{\text{new}}) = \text{Range}(\mathbb{U}_1^{\text{new}(2)})$ with $\mathbb{U}_1^{\text{new}} = \begin{bmatrix} \mathbb{U}_1^{\text{new}(1)} \\ \mathbb{U}_1^{\text{new}(2)} \end{bmatrix}$ and $\left[\mathbb{W}_1^{\text{new}}, \quad \mathbb{W}_2^{\text{new}}\right]$ is unitary with

$$\Lambda_1^{\text{new}} - \mathscr{F}_1^{\text{new}} Y_1^{\text{new}} = \mathbb{U}_1^{\text{new}} \Sigma_1^{\text{new}} \mathbb{V}_1^{\text{new}\,H} + \mathbb{U}_2^{\text{new}} \Sigma_2^{\text{new}} \mathbb{V}_2^{\text{new}\,H}, \text{ with SVD trheshold } \epsilon^{(R)}$$

the SVD to detect inexact breakdown in the restarting block residual where

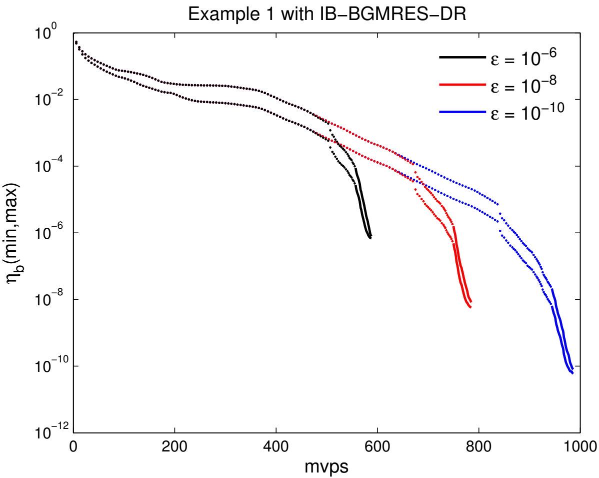
$$Y_1^{\text{new}} = \underset{Y \in \mathbb{C}^{n_1 \times p}}{\text{argmin}} \left\| \Lambda_1^{\text{new}} - \mathscr{F}_1^{\text{new}} Y \right\|_F .$$

Comparative covergence rate



IB-BGMRES [M. Robbé and M. Sadkane, LAA, 2006], BGMRES-DR [R. Morgan, APNUM, 2005]

Inexact breakdown vs targeted accuracy



Concluding remarks

- The new algorithm IB-BGMRES-DR inherits the positive genes of its parents IB-BGMRES [M. Robbé and M. Sadkane, LAA, 2006] and BGMRES-DR [R. Morgan, APNUM, 2005]
- Flexible variants can be designed to accomodate resiliency or mixed precision calculation
- Possible extension to handle massive number of right-hand sides (deflation between sequences)
- Flexible implementation in the framework of the Hi-Box project in collaboration with Airbus Group Innovations and IMACS

"Personal" advert

Parallel Matrix Algorithms and Applications
http://pmaa16.inria.fr

Merci for your attention
Questions ?

https://team.inria.fr/hiepacs/

Comparisons with cousins and parents

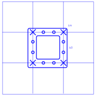
Iso-memory comparison for basis storage

Example	GMRES	GMRES-DR	IB-BGMRES	BGMRES-DR	IB-BGMRES-DR
1	2536	1077	1344	892	**588**
2	1069	856	788	667	**538**
3	378	378	372	341	**335**
4	**412**	**412**	446	447	440
5	845	694	617	474	**386**
6	464	464	357	294	**248**
7	3154	**2003**	3291	3090	2104
8	10643	3110	-	4426	**2202**

Table: Number of *mvps* for regular GMRES, GMRES-DR, IB-BGMRES, BGMRES-DR and IB-BGMRES-DR with $\varepsilon = 10^{-6}$.

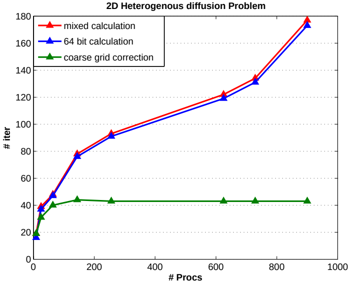
Numerical alternative: numerical scalability in $3D$

Domain based coarse space : $M = M_{AS} + R_O^T A_O^{-1} R_0$ where $A_0 = R_0 S R_O^T$

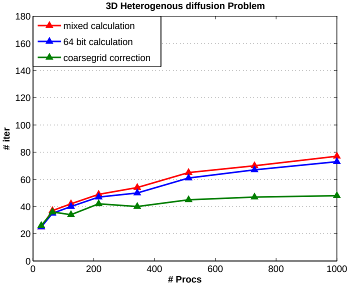


- "As many" dof in the coarse space as sub-domains [Carvalho, Giraud, Le Tallec, 01]
- Partition of unity : R_0^T simplest constant interpolation

$2D$ Heterogenous diffusion



$3D$ Heterogenous diffusion



Experimental set up

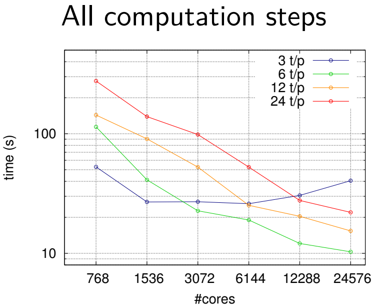
Hopper - LBNL platform

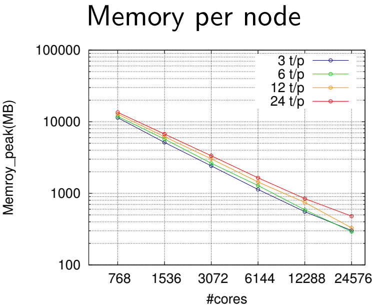
- Two twelve-core AMD 'MagnyCours' 2.1-GHz
- Memory: 32 GB GDDR3
- Double precision

Matrices

Matrix	Matrix211	Nachos4M
n	801K	4,147K
nnz	129,4M	256,4M
Preconditioner	dense	sparse02

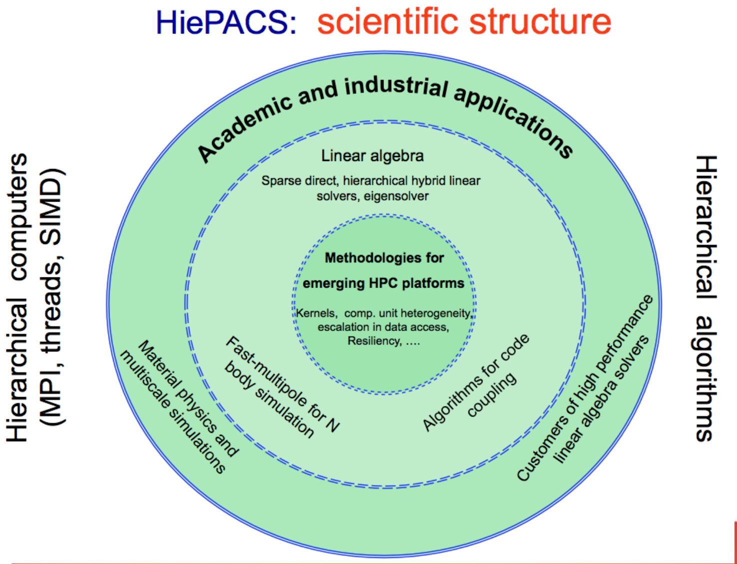
Nachos4M matrix on the Hopper platform





Forewords

HiePACS objectives: Contribute to the design of effective tools for frontier simulations arising from challenging research and industrial multi-scale applications towards extreme computing



Team members

- Permanent Researchers
  - E. Agullo - Inria
  - O. Coulaud - Inria
  - A. Esnard - Bordeaux University
  - M. Faverge - Bordeaux INP
  - L. Giraud - Inria - team leader
  - A. Guermouche - Bordeaux University
  - P. Ramet - Bordeaux University
  - J. Roman - Inria (Bdx INP)
- Post-Doctoral Fellows
  - Y. Harness - Inria-Région Aquitaine
  - M. Kuhn - RAPID DGA
  - E. F. Yetkin - G8-ECS/FP7 Exa2CT
- Technical Staff
  - M. Hastaran - Inria Prace 4IP
  - Q. Khan - Inria
  - C. Piacibello - DGA HiBox
  - F. Pruvost - Inria ADT HPC-Collective
- PhD Students
  - P. Blanchard - ENS Cachan
  - B. Bramas - Airbus-Inria-R. Aquitaine
  - J.M. Couteyen - ASTRIUM/ANRT
  - A. Durocher - Mds/CEA
  - A. Falco - Airbus-Inria-R. Aquitaine
  - C. Fournier - CERFACS
  - L. Poirel - ANR DEDALES
  - G. Pichon - DGA
  - M. Predari - Inria-Région Aquitaine
- Research scientist (partners)
  - P. Brenner - Airbus Defence and Space
  - G. Latu - CEA Cadarache
  - G. Sylvand - Airbus Group Innovation

Current collaborations witin Associate Teams: MORSE (UTK, ICL, UCL, Kaust), FASTLA (LNBL, Stanford), IPL C2S@Exa and industrial partners