Extreme-Scale Solver for Earth’s Mantle –
Spectral-Geometric-Algebraic Multigrid Methods for
Nonlinear, Heterogeneous Stokes Flow

Johann Rudi$^1$ Cristiano Malossi$^2$ Tobin Isaac$^3$ Georg Stadler$^4$
Michael Gurnis$^5$ Peter Staar$^2$ Yves Ineichen$^2$ Costas Bekas$^2$
Alessandro Curioni$^2$ Omar Ghattas$^{1,6}$

$^1$Institute for Computational Engineering and Sciences,
The University of Texas at Austin, USA
$^2$Foundations of Cognitive Solutions, IBM Research – Zurich, Switzerland
$^3$Computing Institute, The University of Chicago, USA
$^4$Courant Institute of Mathematical Sciences, New York University, USA
$^5$Seismological Laboratory, California Institute of Technology, USA
$^6$Jackson School of Geosciences and Department of Mechanical Engineering,
The University of Texas at Austin, USA
Outline

Driving scientific problem & computational challenges

w-BFBT and improved robustness of over established state of the art

HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+w-BFBT

Parallel scalability and performance for HMG+w-BFBT
Incompressible Stokes flow with heterogeneous viscosity

Commonly occurring problem in CS&E:

Creeping non-Newtonian fluid modeled by incompressible Stokes equations with power-law rheology yields spatially-varying and highly heterogeneous viscosity $\mu$ after linearization.

Nonlinear incompressible Stokes PDE:

$$-\nabla \cdot [\mu(u, x) (\nabla u + \nabla u^\top)] + \nabla p = f$$

viscosity $\mu$, RHS forcing $f$

$$-\nabla \cdot u = 0$$

seek: velocity $u$, pressure $p$

Linearization, then discretization with inf-sub stable finite elements yields:

$$\begin{bmatrix} A_\mu & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

→ poor conditioning due to heterogeneous $\mu$

Iterative scheme with upper triangular block preconditioning:

$$\begin{bmatrix} A_\mu & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} \tilde{A}_\mu & B^\top \\ 0 & \tilde{S} \end{bmatrix}^{-1} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

$\tilde{A}_\mu^{-1} \approx A_\mu^{-1}$

$\tilde{S}^{-1} \approx S^{-1} := (BA_\mu^{-1}B^\top)^{-1}$
Severe challenges for parallel scalable PDE solvers

... arising, e.g., in Earth’s mantle convection:

- Severe nonlinearity, heterogeneity, and anisotropy of the Earth’s rheology
- Sharp viscosity gradients in narrow regions (6 orders of magnitude drop in ~5 km)
- Wide range of spatial scales and highly localized features, e.g., plate boundaries of size $O(1 \text{ km})$ influence plate motion at continental scales of $O(1000 \text{ km})$
- Adaptive mesh refinement is essential
- High-order finite elements $Q_k \times P_{k-1}^{\text{disc}}$, order $k \geq 2$, with local mass conservation; yields a difficult to deal with discontinuous, modal pressure approximation
Outline

Driving scientific problem & computational challenges

w-BFBT and improved robustness of over established state of the art

HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+w-BFBT

Parallel scalability and performance for HMG+w-BFBT
Propose: \( w \)-BFBT inverse Schur complement approx.

\[
\begin{bmatrix}
    A_\mu & B^\top \\
    B & 0 \\
\end{bmatrix}
\begin{bmatrix}
    \tilde{A}_\mu & B^\top \\
    0 & \tilde{S} \\
\end{bmatrix}^{-1}
\begin{bmatrix}
    u \\
    p \\
\end{bmatrix} =
\begin{bmatrix}
    f \\
    0 \\
\end{bmatrix}
\]

\( \tilde{A}_\mu^{-1} \approx A_\mu^{-1} \)

\( \tilde{S}^{-1} \approx S^{-1} := (BA_\mu^{-1}B^\top)^{-1} \)
Propose: \( w \)-BFBT inverse Schur complement approx.

\[
\begin{bmatrix}
A_\mu & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{A}_\mu & B^T \\
0 & \tilde{S}
\end{bmatrix}^{-1}
\begin{bmatrix}
u \\
p
\end{bmatrix}
= 
\begin{bmatrix}
f \\
0
\end{bmatrix}
\]

\( \tilde{A}_\mu^{-1} \approx A_\mu^{-1} \)

\( \tilde{S}^{-1} \approx S^{-1} := (BA_\mu^{-1}B^T)^{-1} \)

Underlying principle of BFBT / Least Squares Commutators (LSC): find a commutator matrix \( X \) s.t. (denote unit vectors by \( e_j \))

\[
A_\mu D^{-1}B^T - B^T X \approx 0 \quad \text{or} \quad \min_X \left\| A_\mu D^{-1}B^T e_j - B^T X e_j \right\|_C^2 \quad \forall j
\]

\[
\Rightarrow \quad \tilde{S}_{\text{BFBT}}^{-1} := \left( BC^{-1}B^T \right)^{-1} \left( BC^{-1}A_\mu D^{-1}B^T \right) \left( BD^{-1}B^T \right)^{-1} .
\]

Choice of matrices \( C, D \) is critical for convergence and robustness.
Propose: w-BFBT inverse Schur complement approx.

\[
\begin{bmatrix}
A_\mu & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{A}_\mu & B^T \\
0 & \tilde{S}
\end{bmatrix}^{-1}
\begin{bmatrix}
u \\
p
\end{bmatrix}
= 
\begin{bmatrix}
f \\
0
\end{bmatrix}
\tilde{A}_\mu^{-1} \approx A_\mu^{-1}
\tilde{S}^{-1} \approx S^{-1} := (BA_\mu^{-1}B^T)^{-1}
\]

Underlying principle of BFBT / Least Squares Commutators (LSC):
find a commutator matrix \( X \) s.t. (denote unit vectors by \( e_j \))

\[
A_\mu D^{-1} B^T - B^T X \approx 0 \quad \text{or} \quad \min_X \| A_\mu D^{-1} B^T e_j - B^T X e_j \|_C^{-2} \quad \forall j
\]

\[
\Rightarrow \quad \tilde{S}^{-1}_{\text{BFBT}} := \left( B C^{-1} B^T \right)^{-1} \left( B C^{-1} A_\mu D^{-1} B^T \right) \left( B D^{-1} B^T \right)^{-1}.
\]

Choice of matrices \( C, D \) is critical for convergence and robustness.

\[
\tilde{S}^{-1}_{w-\text{BFBT}} := \left( B C_\mu^{-1} B^T \right)^{-1} \left( B C_\mu^{-1} A_\mu D^{-1}_\mu B^T \right) \left( B D^{-1}_\mu B^T \right)^{-1}
\]

where \( C_\mu = D_\mu := \tilde{M}_u(\sqrt{\mu}) \) are responsible for efficacy and robustness.
Robustness of w-BFBT over established state of the art

<table>
<thead>
<tr>
<th>Problem difficulty (number of sinkers)</th>
<th>Number of GMRES iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#iterations with $M_p(1/\mu)$</th>
<th>$10^4$</th>
<th>$10^6$</th>
<th>$10^8$</th>
<th>$10^{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-rand</td>
<td>29</td>
<td>31</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>S8-rand</td>
<td>64</td>
<td>79</td>
<td>93</td>
<td>165</td>
</tr>
<tr>
<td>S16-rand</td>
<td>85</td>
<td>167</td>
<td>231</td>
<td>891</td>
</tr>
<tr>
<td>S24-rand</td>
<td>117</td>
<td>286</td>
<td>3279</td>
<td>5983</td>
</tr>
<tr>
<td>S28-rand</td>
<td>108</td>
<td>499</td>
<td>2472</td>
<td>$&gt;10000$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#iterations with $w$-BFBT</th>
<th>$10^4$</th>
<th>$10^6$</th>
<th>$10^8$</th>
<th>$10^{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1-rand</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>S8-rand</td>
<td>38</td>
<td>40</td>
<td>41</td>
<td>44</td>
</tr>
<tr>
<td>S16-rand</td>
<td>40</td>
<td>45</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td>S24-rand</td>
<td>31</td>
<td>32</td>
<td>39</td>
<td>55</td>
</tr>
<tr>
<td>S28-rand</td>
<td>29</td>
<td>31</td>
<td>42</td>
<td>60</td>
</tr>
</tbody>
</table>
Outline

Driving scientific problem & computational challenges

w-BFBT and improved robustness of over established state of the art

HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+w-BFBT

Parallel scalability and performance for HMG+w-BFBT
HMG: Hybrid spectral-geometric-algebraic multigrid

- Multigrid hierarchy of nested meshes is generated from an **adaptively refined octree-based mesh** via spectral-geometric coarsening.
- **Re-discretization** of PDEs at coarser levels.
- **Parallel repartitioning** of coarser meshes for load-balancing (crucial for AMR); sufficiently coarse meshes occupy only **subsets of cores**.
- **Coarse grid solver**: AMG (PETSc’s GAMG) invoked on small core counts.
HMG: Hybrid spectral-geometric-algebraic multigrid

- **High-order** $L^2$-projection onto coarser levels; restriction & interpolation are adjoints of each other in $L^2$-sense
- **Chebyshev accelerated Jacobi smoother** (Cheb. from PETSc) with tensorized matrix-free high-order stiffness apply; assembly of high-order diagonal only
- Efficacy, i.e. error reduction, of HMG V-cycles is independent of core count
- No collective communication needed in spectral-geometric MG cycles
**p4est:** Parallel forest-of-octrees AMR library [p4est.org]

Scalable geometric multigrid coarsening due to:

- **Forest-of-octree** based meshes enable fast refinement/coarsening
- Octrees and **space filling curves** used for fast neighbor search, mesh repartitioning, and 2:1 mesh balancing in parallel

*Colors* depict different processor cores.
Geometric coarsening: Repartitioning & core-thinning

- Parallel repartitioning of locally refined meshes for **load balancing**
- **Core-thinning** to avoid excessive communication in multigrid cycle
- **Reduced MPI communicators** containing only non-empty cores
- **Ensure coarsening across core boundaries**: Partition families of octants/elements on same core for next coarsening sweep

Colors depict different processor cores, *numbers* indicate element count on each core.
Outline

Driving scientific problem & computational challenges

$w$-BFBT and improved robustness of over established state of the art

HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+$w$-BFBT

Parallel scalability and performance for HMG+$w$-BFBT
Algorithmic scalability for HMG+w-BFBT (decreasing $h$)

Discretization parameters to test algorithmic scalability:
- Finite element order $k = 2$ is fixed ($Q_k \times P_{k-1}^{\text{disc}}$)
- Vary mesh refinement level $\ell$

Multigrid parameters for $A_\mu$ and $K_d := BC_\mu^{-1}B^\top$:
- 1 HMG V-cycle with 3+3 smoothing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.11</td>
<td>18</td>
<td>0.02</td>
<td>8</td>
<td>0.12</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>0.82</td>
<td>18</td>
<td>0.13</td>
<td>7</td>
<td>0.95</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>6.44</td>
<td>18</td>
<td>1.05</td>
<td>6</td>
<td>7.49</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>50.92</td>
<td>18</td>
<td>8.39</td>
<td>6</td>
<td>59.31</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>405.02</td>
<td>18</td>
<td>67.11</td>
<td>6</td>
<td>472.12</td>
<td>34</td>
</tr>
<tr>
<td>9</td>
<td>3230.67</td>
<td>18</td>
<td>536.87</td>
<td>6</td>
<td>3767.54</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>25807.57</td>
<td>18</td>
<td>4294.97</td>
<td>6</td>
<td>30102.53</td>
<td>34</td>
</tr>
</tbody>
</table>
Algorithmic scalability for HMG+w-BFBT (increasing $k$)

Discretization parameters to test algorithmic scalability:

- Vary finite element order $k$ ($Q_k \times P_{k-1}^{\text{disc}}$)
- Mesh refinement level $\ell = 5$ is fixed

Multigrid parameters for $A_\mu$ and $K_d := BC_\mu^{-1}B^\top$:

- 1 HMG V-cycle with 3+3 smoothing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.82</td>
<td>18</td>
<td>0.13</td>
<td>7</td>
<td>0.95</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>2.74</td>
<td>20</td>
<td>0.32</td>
<td>8</td>
<td>3.07</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>6.44</td>
<td>20</td>
<td>0.66</td>
<td>7</td>
<td>7.10</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>12.52</td>
<td>23</td>
<td>1.15</td>
<td>12</td>
<td>13.67</td>
<td>43</td>
</tr>
<tr>
<td>6</td>
<td>21.56</td>
<td>23</td>
<td>1.84</td>
<td>12</td>
<td>23.40</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>34.17</td>
<td>22</td>
<td>2.75</td>
<td>10</td>
<td>36.92</td>
<td>54</td>
</tr>
<tr>
<td>8</td>
<td>50.92</td>
<td>22</td>
<td>3.93</td>
<td>10</td>
<td>54.86</td>
<td>67</td>
</tr>
</tbody>
</table>
Algorithmic scalability of nonlinear solver (decreasing $h$)

<table>
<thead>
<tr>
<th>Max level of refinement $\ell_{\text{max}}$</th>
<th>Finest resolution [m]</th>
<th>DOF $\times 10^6$</th>
<th>Newton iterations</th>
<th>Total GMRES iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2443</td>
<td>0.96</td>
<td>14</td>
<td>1408</td>
</tr>
<tr>
<td>11</td>
<td>1222</td>
<td>2.67</td>
<td>18</td>
<td>1160</td>
</tr>
<tr>
<td>12</td>
<td>611</td>
<td>5.58</td>
<td>21</td>
<td>1185</td>
</tr>
<tr>
<td>13</td>
<td>305</td>
<td>11.82</td>
<td>21</td>
<td>1368</td>
</tr>
<tr>
<td>14</td>
<td>153</td>
<td>36.35</td>
<td>27</td>
<td>1527</td>
</tr>
</tbody>
</table>

- Finite element order fixed at $Q_2 \times P_1^{\text{disc}}$
- Locally refined mesh with aggressive refinement at plate boundaries
- Multigrid parameters: 1 HMG V-cycle with 3+3 smoothing
Outline

Driving scientific problem & computational challenges

w-BFBT and improved robustness of over established state of the art

HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+w-BFBT

Parallel scalability and performance for HMG+w-BFBT
Implementation optimizations for Blue Gene/Q

(A) Before optimizations

(B) Reduction of blocking MPI communication

(C) Minimization of integer operations & cache misses

(D) Optimization of element-local derivatives; SIMD vectorization

(E) OpenMP threading of matrix-free apply loops (e.g. multigrid smoothing, intergrid projection)

(F) MPI communication reduction, overlapping with computations, OpenMP threading in intergrid operators

(G) Finite element kernel optimizations (e.g. increase of flop-byte ratio, consecutive memory access, pipelining)

(H) Low-level optimizations (e.g. boundary condition enforcement, interpolation of hanging finite element nodes)
Global mantle convection problem for scalability tests

Discretization parameters to test parallel scalability:

- Finite element order $k = 2$ is fixed ($Q_k \times P_{k-1}^{\text{disc}}$)
- Vary max mesh refinement $\ell_{\text{max}}$ for weak scalability
- Refinement down to $\sim 75 \text{ m}$ local resolution
- Resulting mesh has 9 levels of refinement

Multigrid parameters for $A_\mu$ and $K_d$:

- 1 HMG V-cycle with 3+3 smoothing
Blue Gene/Q node performance in weak scaling

- **1 rack (7.5 TFlops/s):**
  - 25.9% A_μ
  - 14.1% K_d
  - 37% B/B^T
  - 3.6% Stokes
  - 8.6% Intergrid
  - 9.9% Total solve

- **32 racks (239 TFlops/s):**
  - 25.9% A_μ
  - 14% K_d
  - 37.4% B/B^T
  - 3.7% Stokes
  - 8.8% Intergrid
  - 9.1% Total solve

- **64 racks (445 TFlops/s):**
  - 25.6% A_μ
  - 13.6% K_d
  - 35.6% B/B^T
  - 3.5% Stokes
  - 9.1% Intergrid
  - 9.8% Total solve

- **96 racks (687 TFlops/s):**
  - 25.1% A_μ
  - 13.7% K_d
  - 35.7% B/B^T
  - 3.5% Stokes
  - 9.8% Intergrid
  - 9.6% Total solve

Time & GFlops/s for MatVec and intergrid operators within Stokes solves

- Highly optimized matrix-free MatVecs dominate with \(\sim 80\%\) of time
- MatVecs and intergrid times consistent across 1 to 96 racks
Extreme weak scalability for HMG+w-BFBT on Sequoia

Performed on LLNL’s Sequoia (Vulcan used for up to 65,536 cores):
IBM Blue Gene/Q architecture with 96 racks resulting in 98,304 nodes,
each node contains 16 compute cores and 16 GBytes of memory.
Extreme strong scalability for HMG + w-BFBT on Sequoia

Performed on LLNL’s Sequoia (Vulcan used for up to 65,536 cores): IBM Blue Gene/Q architecture with 96 racks resulting in 98,304 nodes, each node contains 16 compute cores and 16 GBytes of memory.
References

Preconditioning Stokes problems with variable viscosity:
- Rudi, Stadler, and Ghattas, in preparation.

Octree-based AMR and geometric multigrid on adaptive meshes:

Extreme-scale Earth mantle convection:
- Burstedde, Ghattas, Gurnis, Tan, Tu, Stadler, Wilcox, and Zhong, Proceedings of SC08 (2008), Gordon Bell finalist.
Outline

Appendix: Parallel scalability for HMG+w-BFBT on TACC’s Lonestar 5
Multi-sinker problem for scalability tests on Lonestar 5

Discretization parameters to test parallel scalability:

- Finite element order $k = 2$ is fixed ($\mathbb{Q}_k \times \mathbb{P}^\text{disc}_{k-1}$)
- Vary mesh refinement level $\ell$ for weak scalability

Multigrid parameters for $A_\mu$ and $K_d := BC_\mu^{-1}B^T$:

- 1 HMG V-cycle with 3+3 smoothing
Weak scalability for HMG+w-BFBT on Lonestar 5

Performed on TACC’s Lonestar 5: Cray XC40 with 1252 compute nodes, each contains 2 Intel Haswell 12-core processors and 64 GBytes of memory.
Strong scalability for HMG+$w$-BFBT on Lonestar 5

Performed on TACC’s Lonestar 5: Cray XC40 with 1252 compute nodes, each contains 2 Intel Haswell 12-core processors and 64 GBytes of memory.