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 μ after linearization.

Nonlinear incompressible Stokes PDE:

$$-\nabla \cdot \left[\mu(\boldsymbol{u}, \boldsymbol{x}) \left(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top\right)\right] + \nabla p = \boldsymbol{f} \qquad \text{viscosity } \mu, \text{ RHS forcing } \boldsymbol{f}$$

$$-\nabla \cdot \boldsymbol{u} = 0 \qquad \text{seek: velocity } \boldsymbol{u}, \text{ pressure } p$$

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

$$\begin{bmatrix} \mathbf{A}_\mu & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \qquad \rightarrow \text{poor conditioning due to heterogeneous } \mu$$

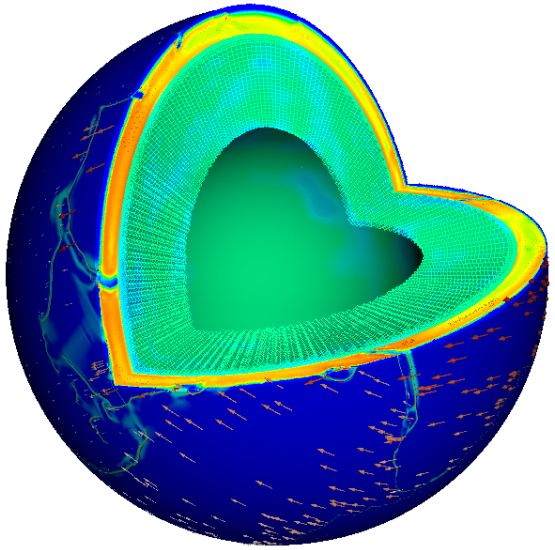
Iterative scheme with upper triangular block preconditioning:

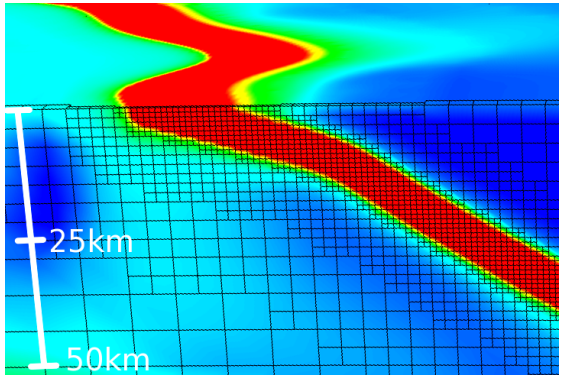
$$\begin{bmatrix} \mathbf{A}_\mu & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_\mu & \mathbf{B}^\top \\ \mathbf{0} & \tilde{\mathbf{S}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \qquad \begin{aligned} \tilde{\mathbf{A}}_\mu^{-1} &\approx \mathbf{A}_\mu^{-1} \\ \tilde{\mathbf{S}}^{-1} &\approx \mathbf{S}^{-1} := (\mathbf{B}\mathbf{A}_\mu^{-1}\mathbf{B}^\top)^{-1} \end{aligned}$$

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 $\sim$5 km)
- Wide range of spatial scales and highly localized features, e.g., plate boundaries of size $\mathcal{O}(1\,\text{km})$ influence plate motion at continental scales of $\mathcal{O}(1000\,\text{km})$
- Adaptive mesh refinement is essential
- High-order finite elements $\mathbb{Q}_k \times \mathbb{P}_{k-1}^{\text{disc}}$, order $k \geq 2$, with local mass conservation; yields a difficult to deal with discontinuous, modal pressure approximation





Viscosity *(colors)*, surface velocity at sol. *(arrows)*, and locally refined mesh.

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.

$$\begin{bmatrix} \mathbf{A}_\mu & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_\mu & \mathbf{B}^\top \\ \mathbf{0} & \tilde{\mathbf{S}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \qquad \begin{aligned} \tilde{\mathbf{A}}_\mu^{-1} &\approx \mathbf{A}_\mu^{-1} \\ \tilde{\mathbf{S}}^{-1} &\approx \mathbf{S}^{-1} := (\mathbf{B}\mathbf{A}_\mu^{-1}\mathbf{B}^\top)^{-1} \end{aligned}$$

Propose: w-BFBT inverse Schur complement approx.

$$\begin{bmatrix} \mathbf{A}_\mu & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_\mu & \mathbf{B}^\top \\ \mathbf{0} & \tilde{\mathbf{S}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \qquad \begin{aligned} &\tilde{\mathbf{A}}_\mu^{-1} \approx \mathbf{A}_\mu^{-1} \\ &\tilde{\mathbf{S}}^{-1} \approx \mathbf{S}^{-1} := (\mathbf{B}\mathbf{A}_\mu^{-1}\mathbf{B}^\top)^{-1} \end{aligned}$$

Underlying principle of BFBT / Least Squares Commutators (LSC): find a commutator matrix $\mathbf{X}$ s.t. (denote unit vectors by $\mathbf{e}_j$)

$$\mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top - \mathbf{B}^\top \mathbf{X} \approx \mathbf{0} \quad \text{or} \quad \min_{\mathbf{X}} \left\| \mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top \mathbf{e}_j - \mathbf{B}^\top \mathbf{X} \mathbf{e}_j \right\|_{\mathbf{C}^{-1}}^2 \quad \forall j$$

$$\Rightarrow \quad \tilde{\mathbf{S}}_{\text{BFBT}}^{-1} := \left(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top\right)^{-1} \left(\mathbf{B}\mathbf{C}^{-1}\mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top\right) \left(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top\right)^{-1}.$$

Choice of matrices $\mathbf{C}, \mathbf{D}$ is critical for convergence and robustness.

Propose: w-BFBT inverse Schur complement approx.

$$\begin{bmatrix} \mathbf{A}_\mu & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_\mu & \mathbf{B}^\top \\ \mathbf{0} & \tilde{\mathbf{S}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix} \qquad \begin{aligned} &\tilde{\mathbf{A}}_\mu^{-1} \approx \mathbf{A}_\mu^{-1} \\ &\tilde{\mathbf{S}}^{-1} \approx \mathbf{S}^{-1} := (\mathbf{B}\mathbf{A}_\mu^{-1}\mathbf{B}^\top)^{-1} \end{aligned}$$

Underlying principle of BFBT / Least Squares Commutators (LSC): find a commutator matrix $\mathbf{X}$ s.t. (denote unit vectors by $\mathbf{e}_j$)

$$\mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top - \mathbf{B}^\top \mathbf{X} \approx \mathbf{0} \quad \text{or} \quad \min_{\mathbf{X}} \left\| \mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top \mathbf{e}_j - \mathbf{B}^\top \mathbf{X} \mathbf{e}_j \right\|_{\mathbf{C}^{-1}}^2 \quad \forall j$$

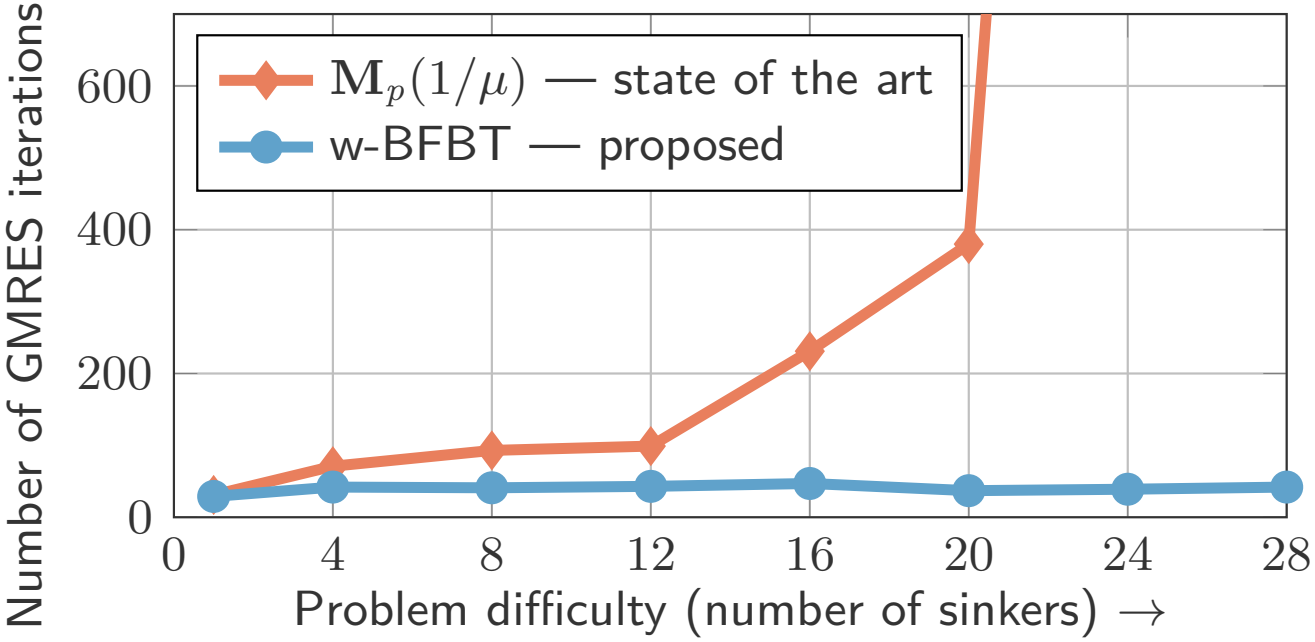
$$\Rightarrow \quad \tilde{\mathbf{S}}_{\text{BFBT}}^{-1} := \left(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top\right)^{-1} \left(\mathbf{B}\mathbf{C}^{-1}\mathbf{A}_\mu \mathbf{D}^{-1}\mathbf{B}^\top\right) \left(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top\right)^{-1}.$$

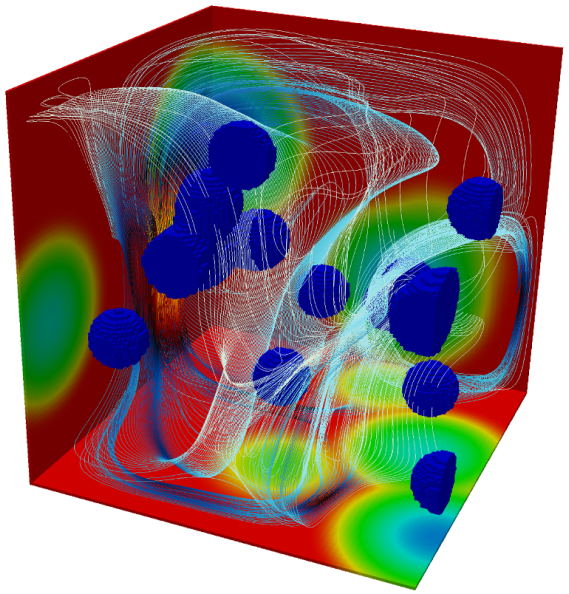
Choice of matrices $\mathbf{C}, \mathbf{D}$ is critical for convergence and robustness.

$$\tilde{\mathbf{S}}_{\text{w-BFBT}}^{-1} := \left(\mathbf{B}\mathbf{C}_\mu^{-1}\mathbf{B}^\top\right)^{-1} \left(\mathbf{B}\mathbf{C}_\mu^{-1}\mathbf{A}_\mu \mathbf{D}_\mu^{-1}\mathbf{B}^\top\right) \left(\mathbf{B}\mathbf{D}_\mu^{-1}\mathbf{B}^\top\right)^{-1}$$

where $\mathbf{C}_\mu = \mathbf{D}_\mu := \tilde{\mathbf{M}}_{\boldsymbol{u}}(\sqrt{\mu})$ are responsible for efficacy and robustness.

Robustness of w-BFBT over established state of the art





#iterations with $\mathbf{M}_p(1/\mu)$

$\mathrm{DR}(\mu) = \ldots$	10^4	10^6	10^8	10^{10}
S1-rand	29	31	31	29
S8-rand	64	79	93	165
S16-rand	85	167	231	891
S24-rand	117	286	3279	5983
S28-rand	108	499	2472	>10000

#iterations with w-BFBT

$\mathrm{DR}(\mu) = \ldots$	10^4	10^6	10^8	10^{10}
S1-rand	29	29	29	30
S8-rand	38	40	41	44
S16-rand	40	45	47	48
S24-rand	31	32	39	55
S28-rand	29	31	42	60

Outline

Driving scientific problem & computational challenges

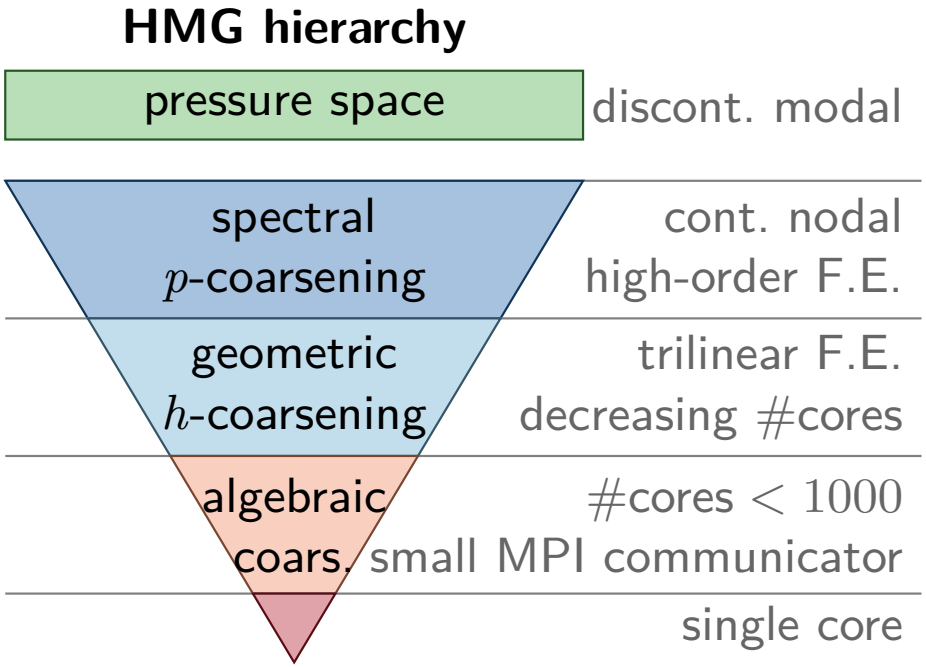
w-BFBT and improved robustness of over established state of the art

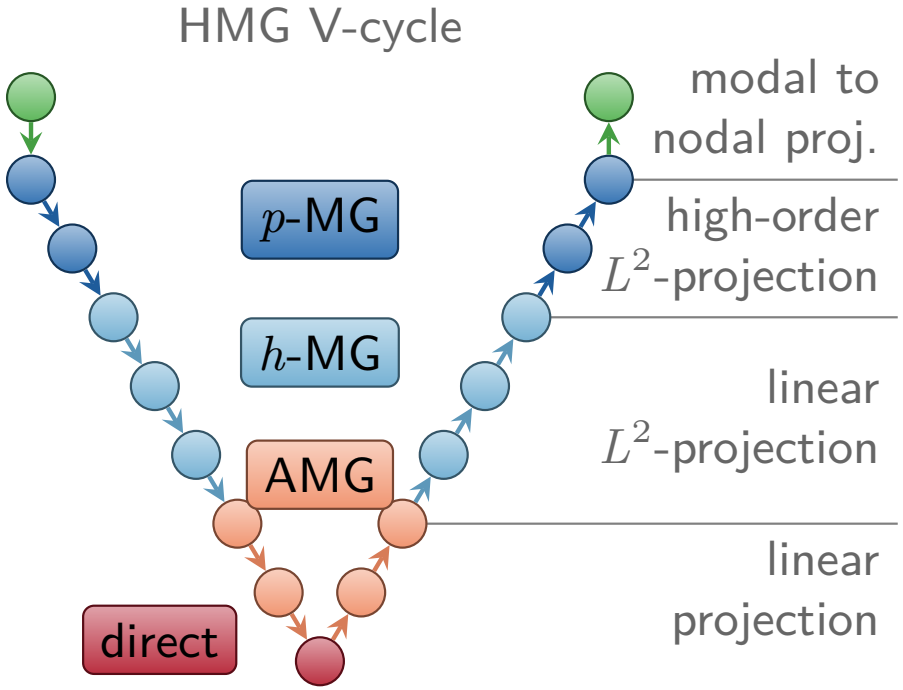
HMG: Hybrid spectral-geometric-algebraic multigrid

Algorithmic scalability for HMG+w-BFBT

Parallel scalability and performance for HMG+w-BFBT

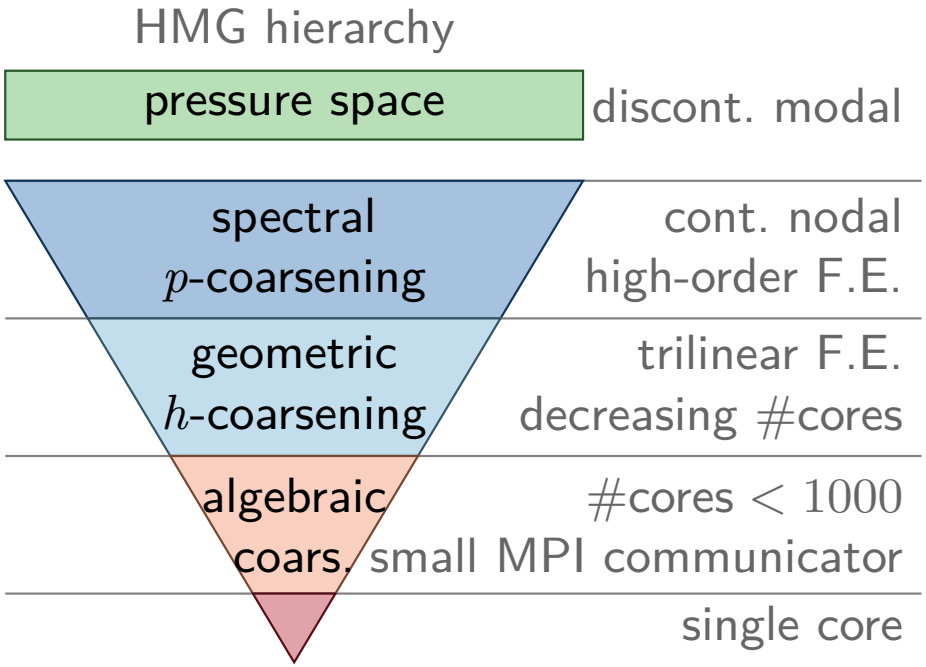
HMG: Hybrid spectral-geometric-algebraic multigrid

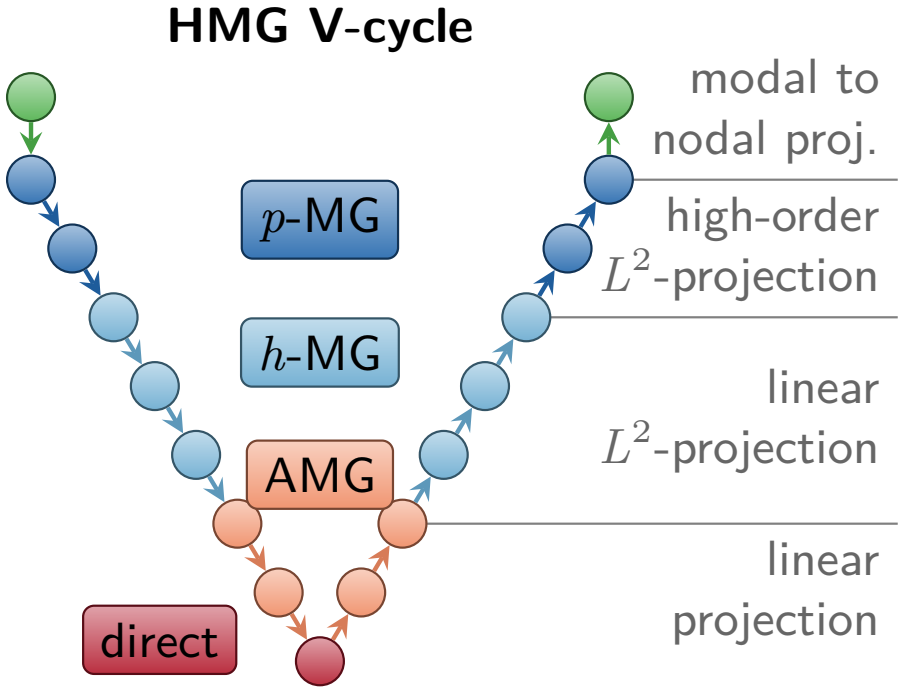




- Multigrid hierarchy of nested meshes is generated from an adaptively refined octree-based mesh via spectral-geometric coarsening
- Re-discretization of PDEs at coarser levels
- Parallel repartitioning of coarser meshes for load-balancing (crucial for AMR); sufficiently coarse meshes occupy only subsets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

HMG: Hybrid spectral-geometric-algebraic multigrid



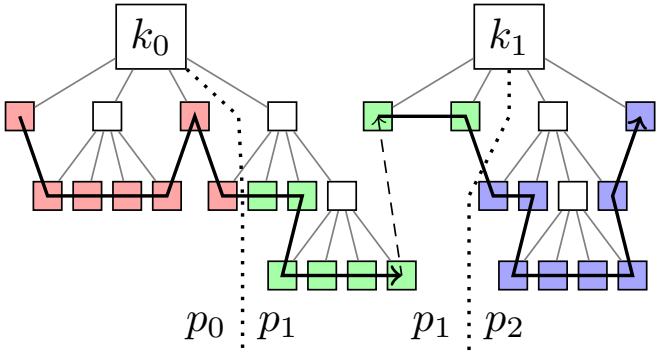


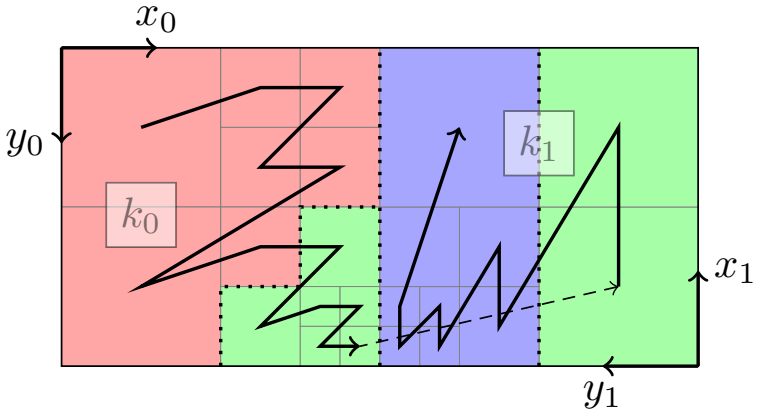
- High-order L^2-projection onto coarser levels;
  restriction & interpolation are adjoints of each other in L^2-sense
- Chebyshev accelerated Jacobi smoother (Cheb. from PETSc) with tensorized matrix-free high-order stiffness apply; assembly of high-order diagonal only
- Efficacy, i.e. error reduction,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

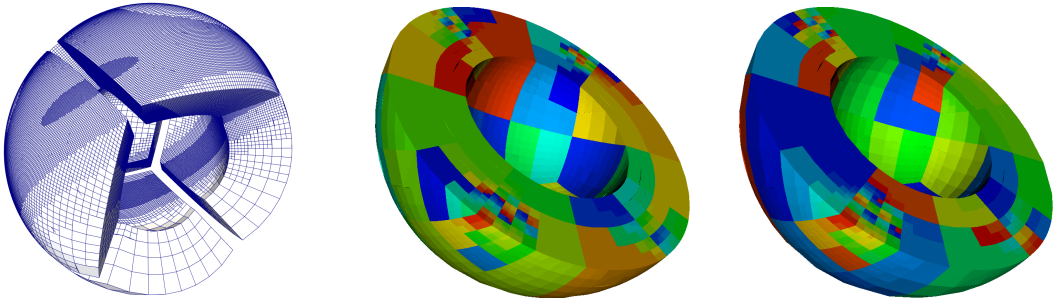
p4est: Parallel forest-of-octrees AMR library [p4est.org]

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



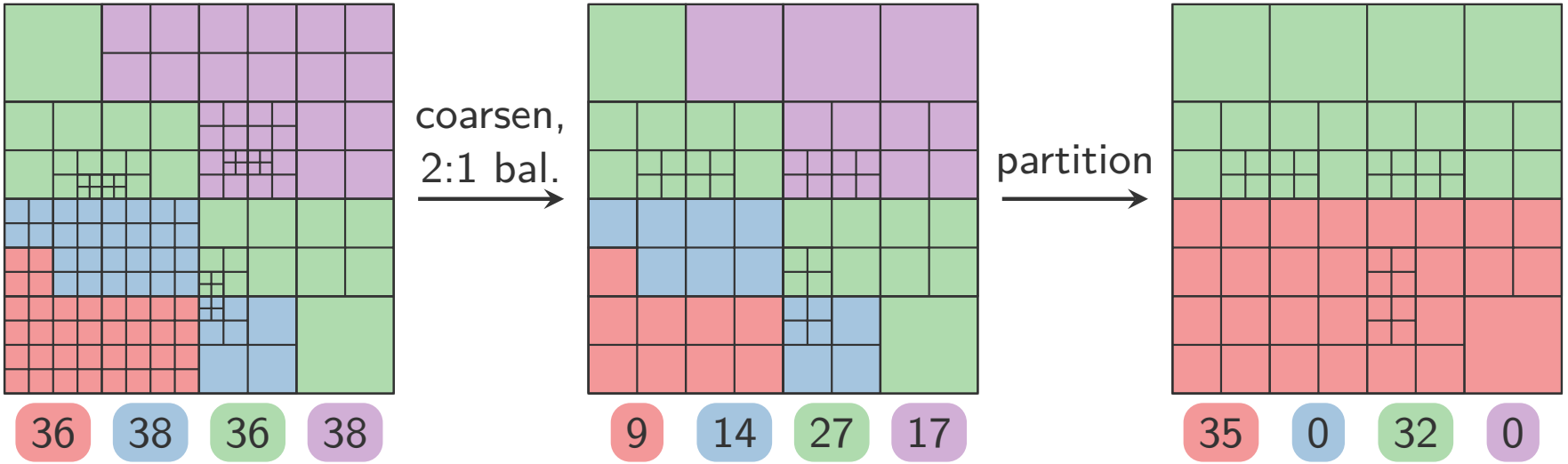




Colors depict different processor cores.

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

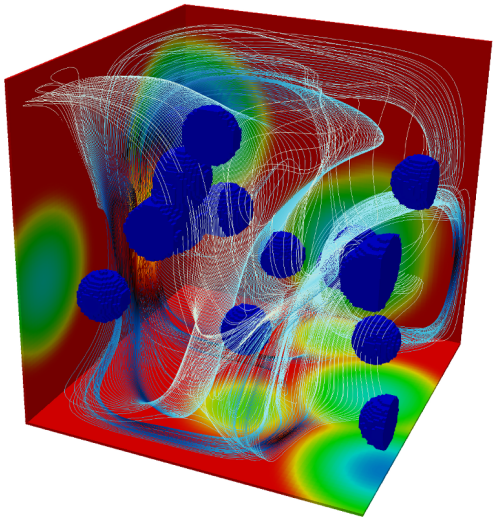
w-BFBT and improved robustness of over established state of the art

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 ($\mathbb{Q}_k \times \mathbb{P}_{k-1}^{\text{disc}}$)
- Vary mesh refinement level ℓ

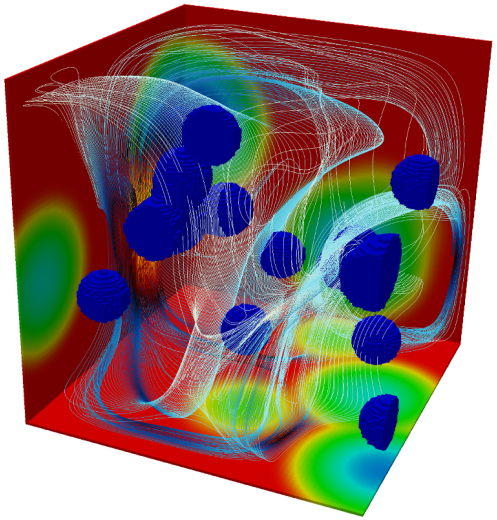
Multigrid parameters for $\mathbf{A}_\mu$ and $\mathbf{K}_\text{d} := \mathbf{B}\mathbf{C}_\mu^{-1}\mathbf{B}^\top$:

- 1 HMG V-cycle with 3+3 smoothing

#iterations for solving sub-systems $\mathbf{A}_\mu \mathbf{u} = \mathbf{f}$, $\mathbf{K}_\text{d}\mathbf{p} = \mathbf{g}$, and full Stokes system

ℓ	$\boldsymbol{u}$-DOF $[\times 10^6]$	It. $\mathbf{A}_\mu$	p-DOF $[\times 10^6]$	It. $\mathbf{K}_\text{d}$	DOF $[\times 10^6]$	It. **Stokes**
4	0.11	18	0.02	8	0.12	40
5	0.82	18	0.13	7	0.95	33
6	6.44	18	1.05	6	7.49	33
7	50.92	18	8.39	6	59.31	34
8	405.02	18	67.11	6	472.12	34
9	3230.67	18	536.87	6	3767.54	34
10	25807.57	18	4294.97	6	30102.53	34

Algorithmic scalability for HMG+w-BFBT (increasing k)



Discretization parameters to test algorithmic scalability:

- Vary finite element order k ($\mathbb{Q}_k \times \mathbb{P}_{k-1}^{\text{disc}}$)
- Mesh refinement level $\ell = 5$ is fixed

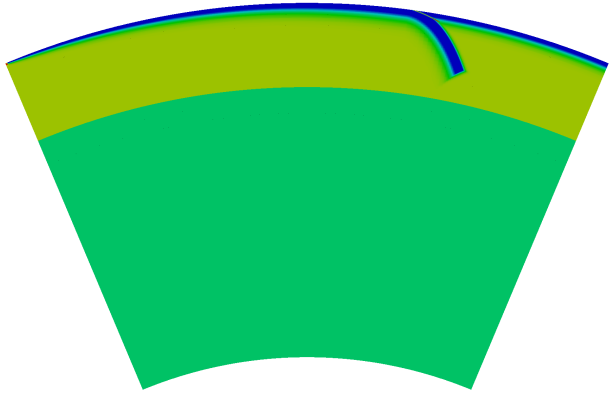
Multigrid parameters for $\mathbf{A}_\mu$ and $\mathbf{K}_\text{d} := \mathbf{B}\mathbf{C}_\mu^{-1}\mathbf{B}^\top$:

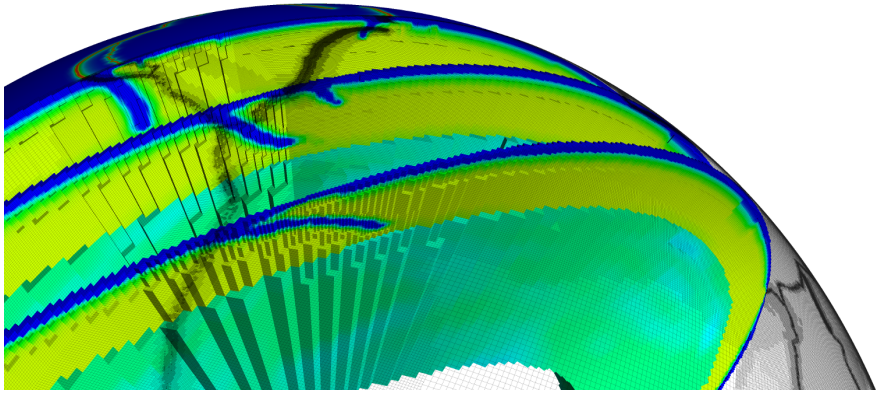
- 1 HMG V-cycle with 3+3 smoothing

#iterations for solving sub-systems $\mathbf{A}_\mu \mathbf{u} = \mathbf{f}$, $\mathbf{K}_\text{d}\mathbf{p} = \mathbf{g}$, and full Stokes system

k	$\boldsymbol{u}$-DOF $[\times 10^6]$	It. $\mathbf{A}_\mu$	p-DOF $[\times 10^6]$	It. $\mathbf{K}_\text{d}$	DOF $[\times 10^6]$	It. **Stokes**
2	0.82	18	0.13	7	0.95	33
3	2.74	20	0.32	8	3.07	37
4	6.44	20	0.66	7	7.10	36
5	12.52	23	1.15	12	13.67	43
6	21.56	23	1.84	12	23.40	50
7	34.17	22	2.75	10	36.92	54
8	50.92	22	3.93	10	54.86	67

Algorithmic scalability of nonlinear solver (decreasing h)





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

- Finite element order fixed at $\mathbb{Q}_2 \times \mathbb{P}_1^{\text{disc}}$
- 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

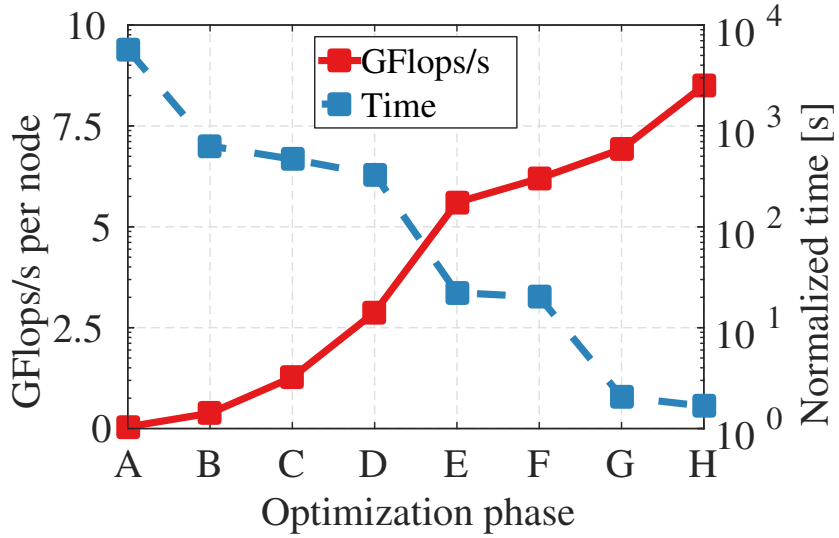
Implementation optimizations for Blue Gene/Q

(A) Before optimizations

(B) Reduction of blocking MPI communication

(C) Minimization of integer operations & cache misses

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



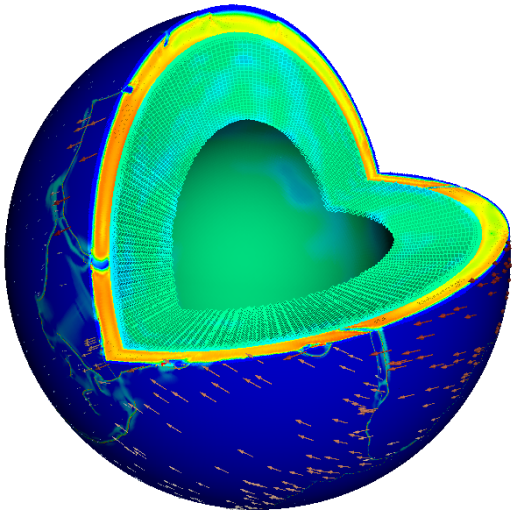
(E) OpenMP threading of matrix-free apply loops (e.g. multigrid smoothing, intergrid projection)

(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



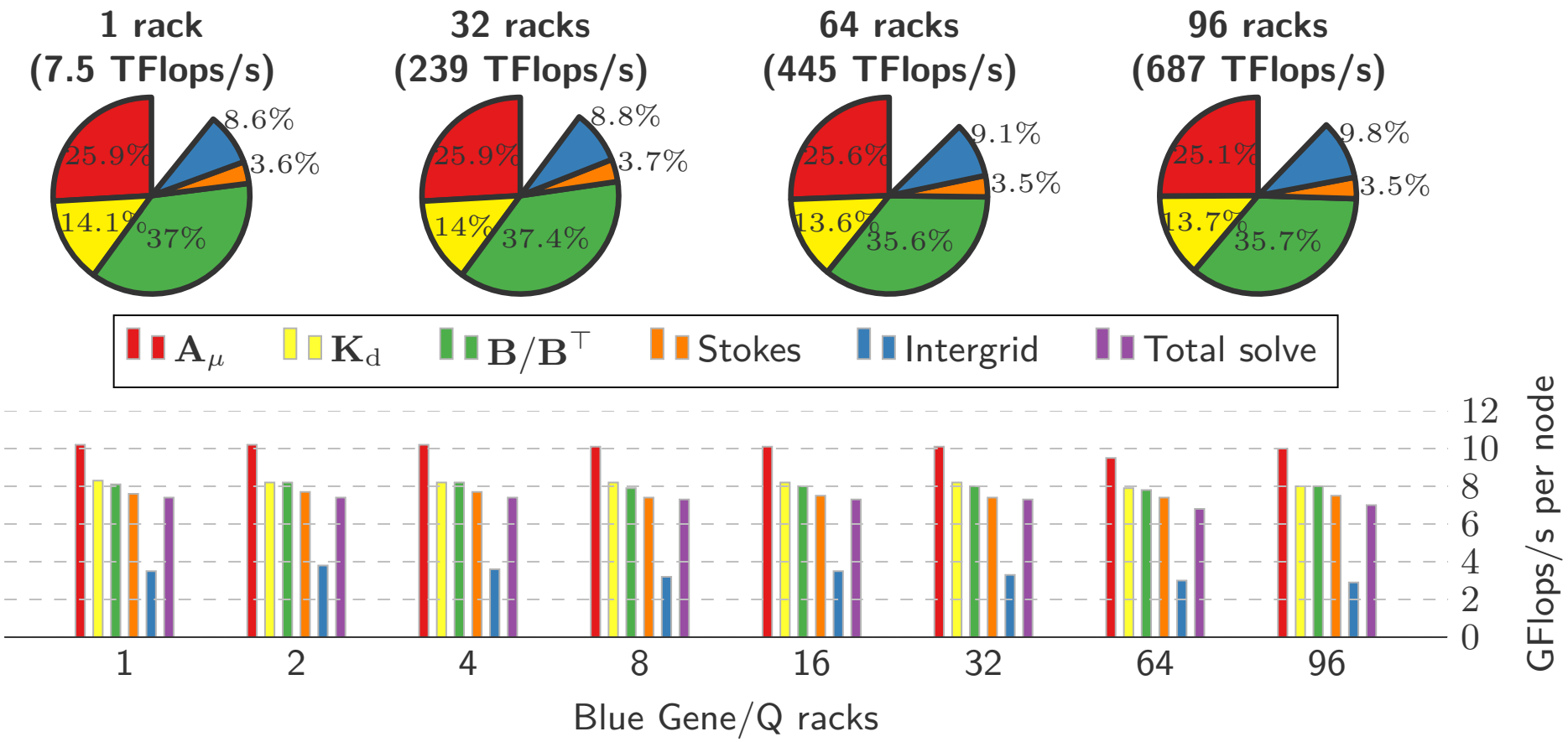
Discretization parameters to test parallel scalability:

- Finite element order $k = 2$ is fixed $(\mathbb{Q}_k \times \mathbb{P}_{k-1}^{\mathrm{disc}})$
- Vary max mesh refinement $\ell_{\max}$ for weak scalability
- Refinement down to ∼75 m local resolution
- Resulting mesh has 9 levels of refinement

Multigrid parameters for $\mathbf{A}_\mu$ and $\mathbf{K}_{\mathrm{d}}$:

- 1 HMG V-cycle with 3+3 smoothing

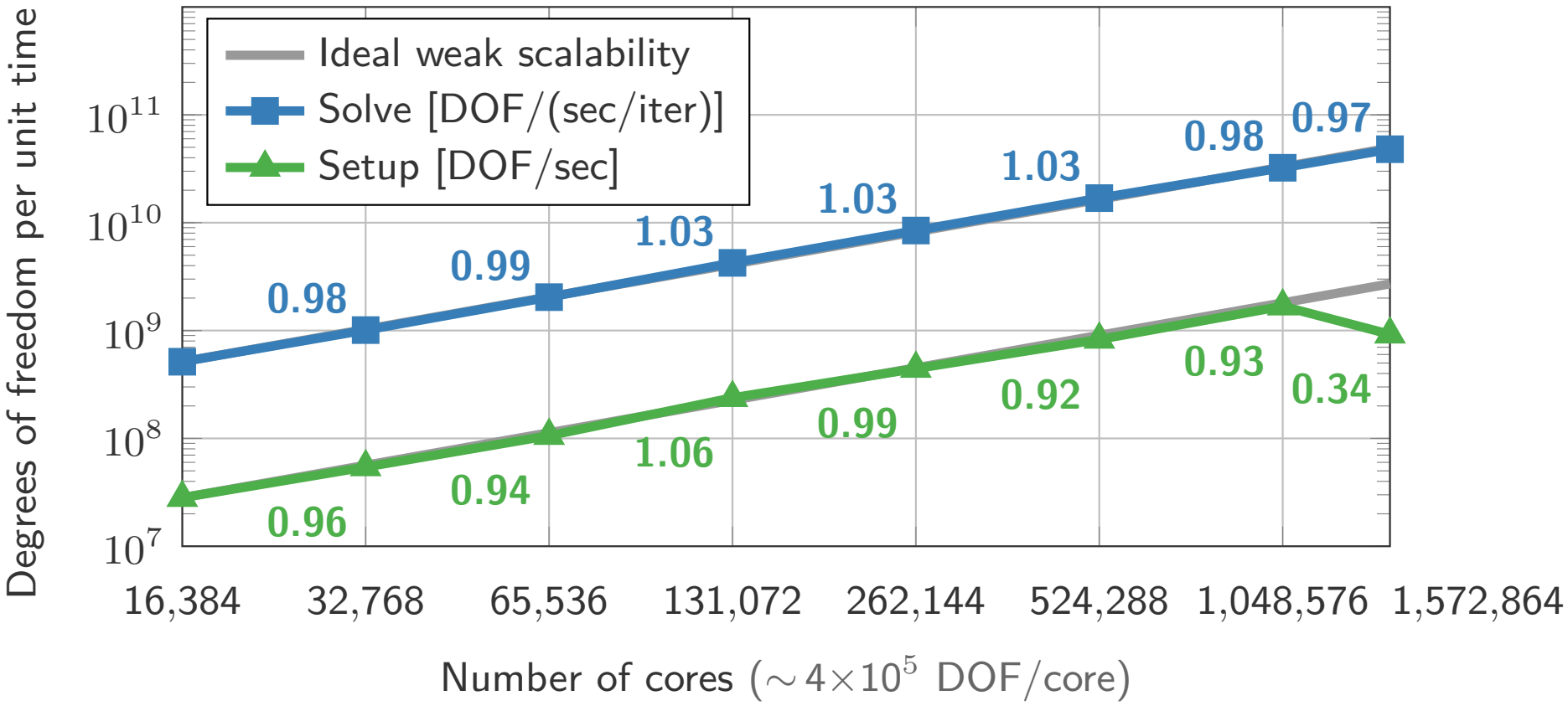
Blue Gene/Q node performance in weak scaling



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

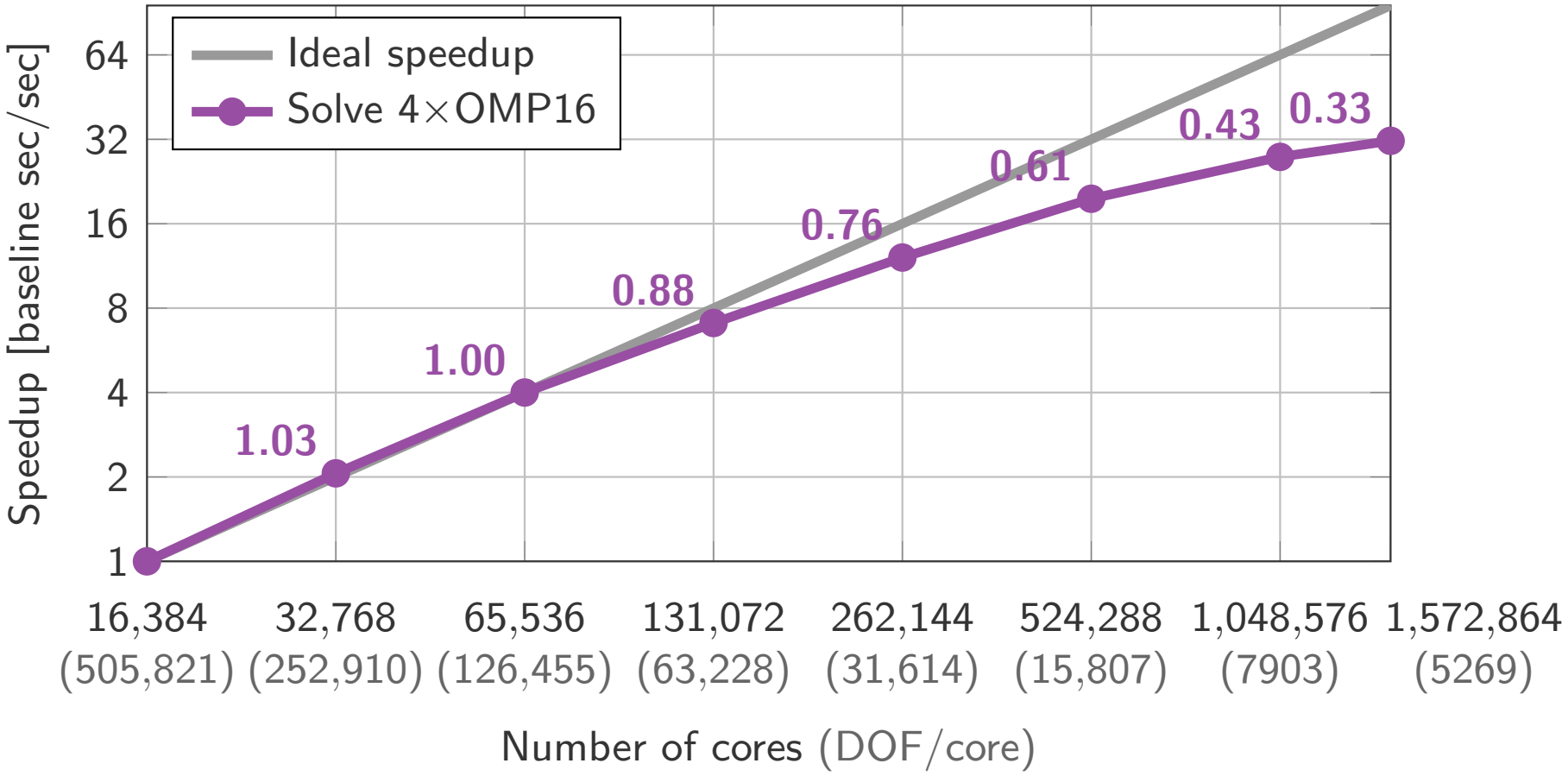
- Highly optimized matrix-free MatVecs dominate with ∼80% of time
- MatVecs and intergrid times consistent across 1 to 96 racks

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:

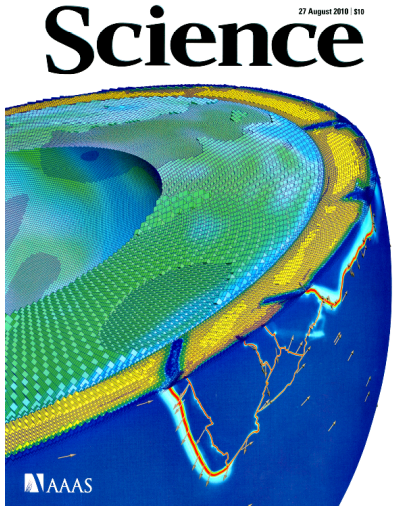
- May and Moresi, Phys. Earth Planet. In., 171 (2008).
- Burstedde, Ghattas, Stadler, Tu, and Wilcox, Comput. Method. Appl. M., 198 (2009).
- Grinevich and Olshanskii, SIAM J. Sci. Comput., 31 (2009).
- Rudi, Stadler, and Ghattas, in preparation.

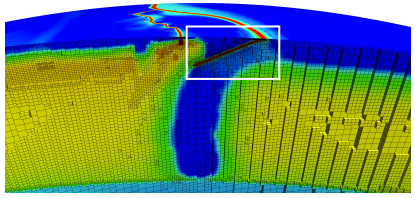
Octree-based AMR and geometric multigrid on adaptive meshes:

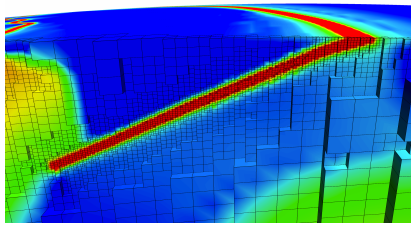
- Burstedde, Ghattas, Gurnis, Isaac, Stadler, Warburton, and Wilcox, Proceedings of SC10 (2010), Gordon Bell finalist.
- Burstedde, Wilcox, and Ghattas, SIAM J. Sci. Comput., 33 (2011).
- Sundar, Biros, Burstedde, Rudi, Ghattas, and Stadler, Proceedings of SC12 (2012).

Extreme-scale Earth mantle convection:

- Burstedde, Ghattas, Gurnis, Tan, Tu, Stadler, Wilcox, and Zhong, Proceedings of SC08 (2008), Gordon Bell finalist.
- Stadler, Gurnis, Burstedde, Wilcox, Alisic, and Ghattas, Science, 329 (2010).
- Rudi, Malossi, Isaac, Stadler, Gurnis, Ineichen, Bekas, Curioni, and Ghattas, Proceedings of SC15 (2015), Winner of Gordon Bell Prize.



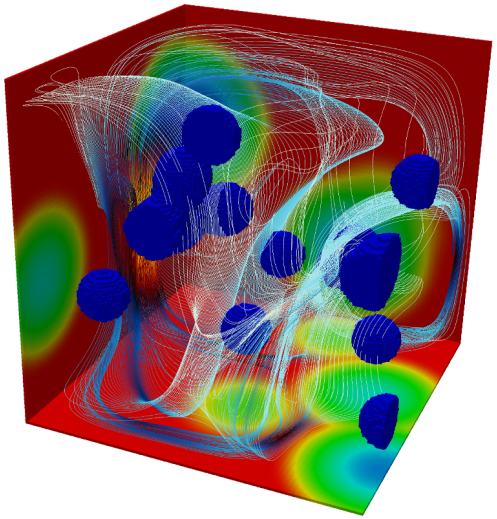




Outline

Appendix: Parallel scalability for HMG+w-BFBT on TACC's Lonestar 5

Multi-sinker problem for scalability tests on Lonestar 5



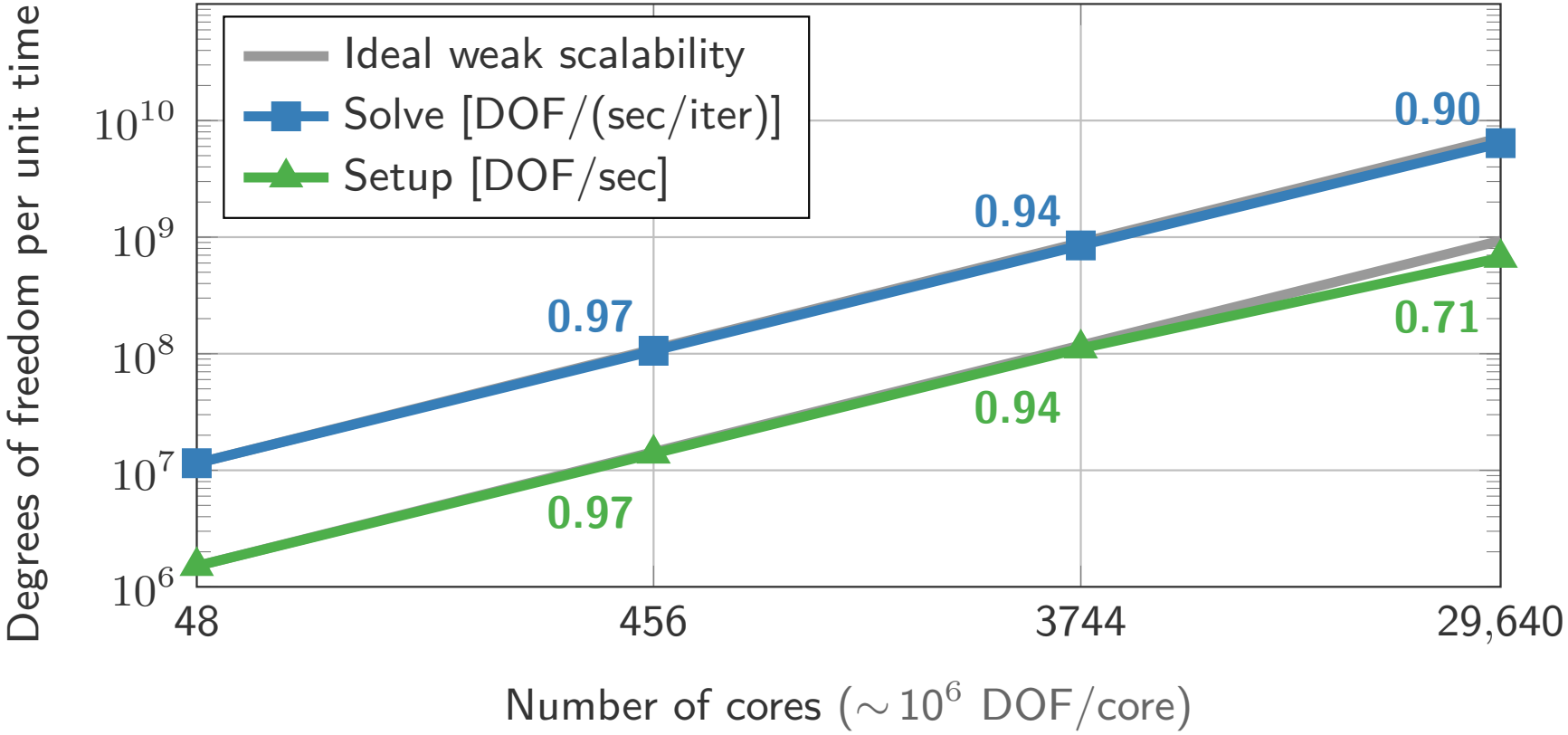
Discretization parameters to test parallel scalability:

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

Multigrid parameters for $\mathbf{A}_\mu$ and $\mathbf{K}_\mathrm{d} := \mathbf{B}\mathbf{C}_\mu^{-1}\mathbf{B}^\top$:

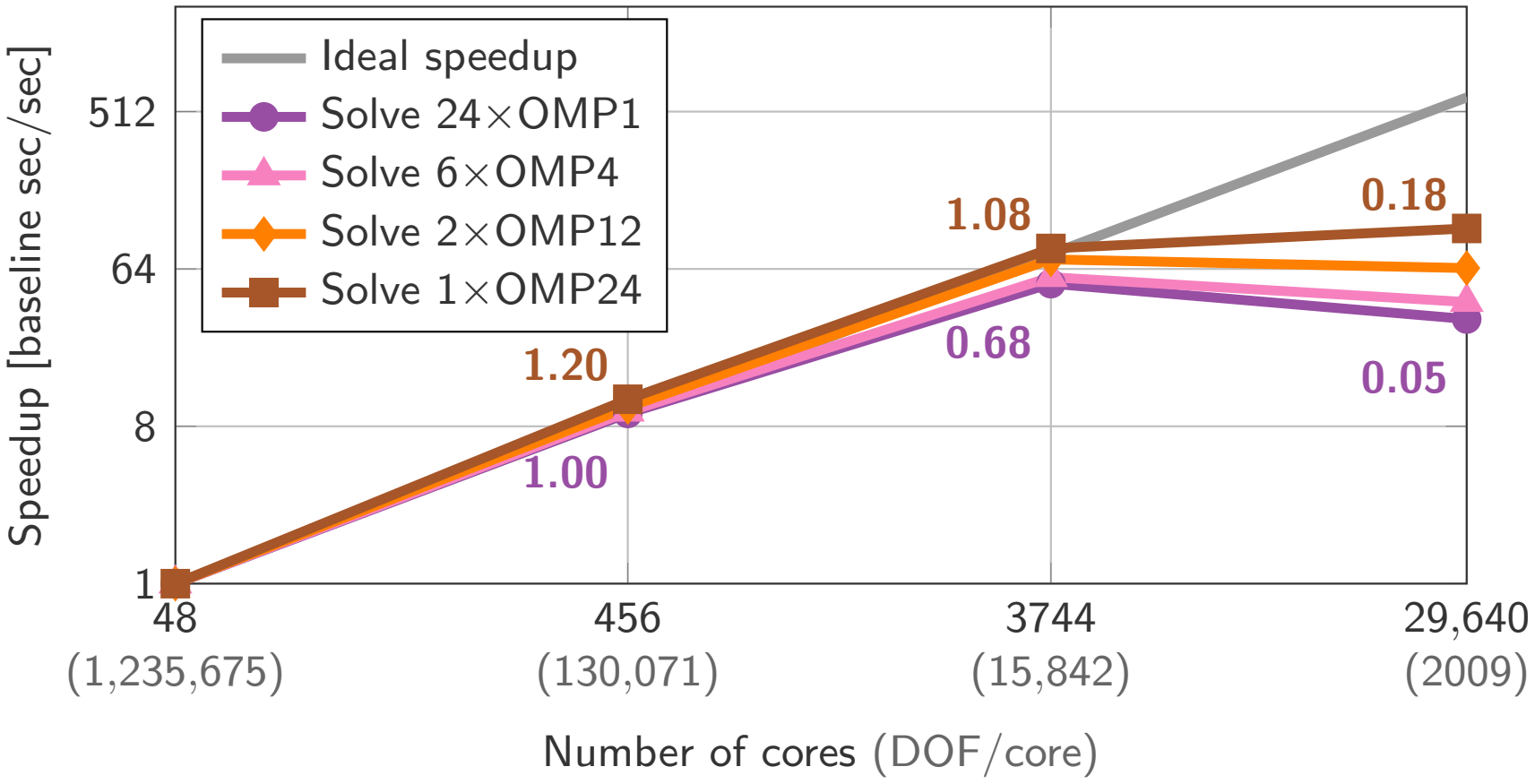
- 1 HMG V-cycle with 3+3 smoothing

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.