Flexible GMRES with Analog Accelerators

A. Gupta†, V. Kalantzis†, M.S. Squillante†, C.W. Wu†, H. Avron§, S. Ubaru†, T. Gokmen†, M. Rasch†, T. Nowicki†, L. Horesh†

†IBM Research §Tel-Aviv University

Sparse linear system solvers

- The iterative solution of linear algebraic equations of the form

$$Ax = b,\ A \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n,$$

  is among the most important computational kernels in optimization, science, and engineering.

- In this work, we focus on solving general sparse linear systems by preconditioned Krylov subspace methods, which consider approximate solutions x_m from the subspace

$$\mathcal{K}_j(A, b) = \{r_0, AM^{-1}r_0, (AM^{-1})^2 r_0, \ldots, (AM^{-1})^{m-1} r_0\},$$

  where $M \in \mathbb{R}^{n\times n}$ approximates A^{-1} and $r_0 = b - Ax_0$.

- While research on sparse iterative solver libraries for exascale computing is active, alternatives that increase concurrency of compute nodes are highly desirable.

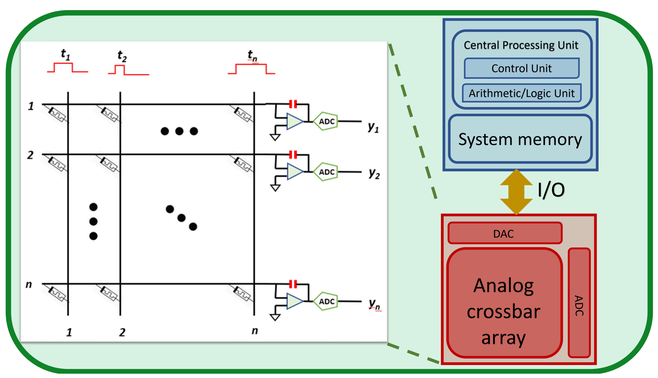
Analog crossbar arrays

Analog crossbar arrays form an alternative computing paradigm in which matrix values are stored in an array of non-volatile memory.

- They offer high degrees of parallelism with low energy consumption by employing memristive elements to store information and execute operations such as a multiply-and-add.
- Performing Matrix-Vector Multiplications (MVM) is then possible in a time that is independent of the number of nonzero entries in the operand matrices.
- Such devices have been recently advocated in machine learning, due to the abundance of MVMs and outer-product updates in the training of neural networks.
- Analog crossbar arrays introduce stochastic noise, and reaping their benefits requires algorithms that can absorb these errors in a flexible manner.

Hybrid architecture for Flexible GMRES (FGMRES) We illustrate the proposed hybrid architecture to solve sparse linear systems with approximate inverse preconditioners:

- Perform all steps of FGMRES but step 4 using a digital substrate.
- Perform step 4 in an analog crossbar array through an MVM product.

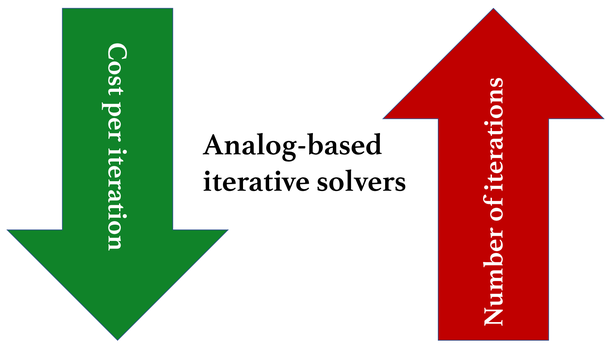


Preconditioning FGMRES by approximate inverses on a hybrid architecture

- Since analog crossbar arrays can execute MVM very fast, we focus on approximate inverse preconditioners (e.g., SPAI) constructed in digital space and mapped to analog hardware.
- Due to stochastic noise, the application of the preconditioner is non-deterministic, and standard preconditioned GMRES can not be employed.
- We propose the combined use of analog computing technology and approximate inverses, paired with Richardson iterations, as preconditioners in FGMRES.
- The flexible nature of FGMRES provides an ideal setting for analog hardware since the time-varying inaccuracies of these devices are incorporated directly in the iterative solver.

Algorithm 1 Flexible GMRES

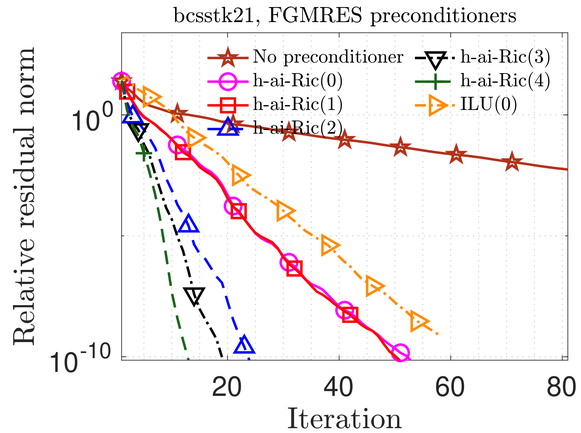
1: **input**: $A \in \mathbb{R}^{n\times n}$; $b, x_0 \in \mathbb{R}^n$; $t_{\text{ol}} \in \mathbb{R}$; $m \in \mathbb{N}$.
2: Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|$, and $v_1 = r_0/\beta$
3: **for** $j = 1$ **to** m **do**
4: Compute $z_j = M_j v_j$
5: Compute $w = Az_j$
6: For $i = 1, \ldots, j$: $\begin{cases} h_{i,j} = w^T v_i \\ w = w - h_{i,j} v_i \end{cases}$
7: **end for**
8: Define $Z_m = [z_1, \ldots, z_m]$
9: Compute $x_m = x_0 + Z_m y_m$ where $y_m = \arg\min_{y\in\mathbb{R}^m} \|\beta e_1 - \overline{H}_m y\|$ and $e_1 = [1, 0, \ldots, 0]^T$
10: If $\|r_m\|/\|r_0\| \le t_{\text{ol}}$, exit; else, restart from Step 2 with $x_0 = x_m$;

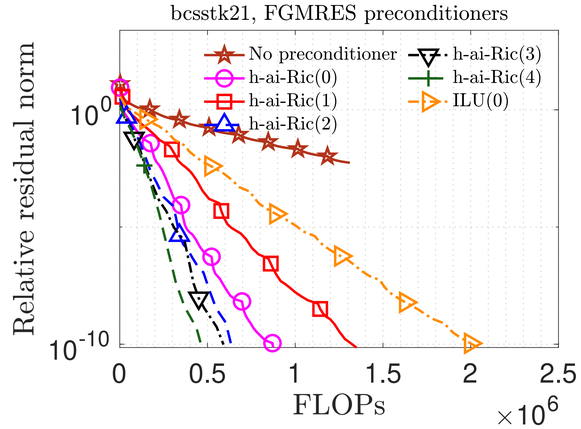


Experiments

We consider the matrix `bcsstk21` and plot the relative residual norm achieved by FGMRES preconditioned by up to 5 steps of Richardson iteration paired with an approximate inverse preconditioner. Comparisons with standard GMRES and GMRES preconditioned by ILU(0) are also provided. Top: Results as function of number of iterations; Bottom: Results as function of number of digital FLOPs.

Experiments (Cont.)





Our simulation experiments on a small realistic problem show orders of magnitude speedup over standard GMRES and nearly an order of magnitude speedup over GMRES preconditioned with ILU(0) running on a microprocessor for attaining the same level of accuracy. The speedup would likely widen even further with larger problems, or if we considered the memory inefficiencies stemming from substitutions with sparse triangular ILU(0) factors. It is important to note that "h-ai-Ric(0)" runs at essentially the same digital cost as unpreconditioned GMRES but its convergence rate is considerably improved. This shows that even without Richardson smoothing, simply using analog approximate inverses as preconditioners in FGMRES can lead to major improvements.

References

Vasileios Kalantzis, Anshul Gupta, Lior Horesh, Tomasz Nowicki, Mark S Squillante, Chai Wah Wu, *Solving sparse linear systems with approximate inverse preconditioners on analog devices.* 2021 IEEE High Performance Extreme Computing Conference (HPEC 2021).