ASP-DAC 2023



Parallel Incomplete LU Factorization Based Iterative Solver for Fixed-Structure Linear Equations in Circuit Simulation

Lingjie Li, Zhiqiang Liu, Kan Liu, Shan Shen, Wenjian Yu Dept. Computer Science & Tech., BNRist, Tsinghua University, Beijing, China Jan. 18, 2023

Speaker's bio

- Lingjie Li is a Ph.D candidate from the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His supervisor is Prof. Wenjian Yu.
- His research interests include
 - > High-performance numerical algorithms
 - Circuit simulation
 - > Tensor computation
 - > Data compression





- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions







- Core problem of circuit simulation
 - > Solving the systems of linear equation generated from circuits
- Features of the systems
 - > Large
 - ~100K nodes, even more
 - Very sparse, nonzero diagonal
 - Fixed-structure
 - Linear circuit: values don't change
 - Nonlinear circuit: same pattern, changed values

Ax = b



Background

- Direct solver
 - LU factorization
 - > Good for linear circuit
 - Only need to factorize once
 - Potentially huge number of fill-ins
 - Even after reorder





Background

Iterative solver

> Preconditioned GMRES

$$Ax = b \longrightarrow P^{-1}Ax = P^{-1}b$$

- > Preconditioner
 - diag(*A*)
 - Incomplete LU factorization (ILU)
 - graph spectral sparsification
 - •

Background

- ◆ ILU(*k*) factorization
 - > Level of fill-ins



> Ignore fill-ins whose level > k
> ILU(0): ignore all fill-ins

		1	2	3	4	5	6	7	8	9	10
	1	X		X	X					X	
	2	X	X	1	X					1	x
	3	X	X	X	X					1	
	4	X	X	1	X					1	
	5					X	x		x		
l)	6			X	1	X	x		1	2	
	7							x			x
	8					x	1	x	x	4	1
	9		X	2	1		x		2	x	1
	10		X	2	X			X	1	2	X



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

• Up-looking algorithm for ILU(0)

Algo	orithm 1 Conventional ILU(0) factorization	l
Inpu	at: $A \in \mathbb{R}^{N \times N}$.	
1: 1	for $i = 2, \ldots, N$ do	⊳ up-looking(i)
2:	for $k = 1,, i - 1$ do	▷ row-update(i, k)
3:	if $a_{ik} \in S$ then	
4:	$a_{ik} \neq a_{kk}$.	
5:	for $j = k + 1,, N$ do	
6:	if $a_{ij} \in S$ and $a_{kj} \in S$ then	
7:	$a_{ij} = a_{ik}a_{kj}.$	

◆ ILU(*k*)? Add a symbolic factorization

	1	2	3	4	5	6	7	8	9	10
1	x		X	X						
2		x	X	X					X	X
3	x	x	x	X						
4	X	x	X	x						X
5					X	X		X		
6					X	X		X	X	
7							X	X		X
8					X	X	X	X	X	
9		X				X		X	x	
10		X		X			X			X

Row-level parallelism and data dependencies





Figure 1: The dependency DAG of a matrix with given lower triangular structure. $k \rightarrow i$ means row *i* depends on row *k*.

How to schedule?

- Dynamic task pool
 - Too much overhead
- Static scheduling
 - Fixed-structure
 - Pre-analyze the dependency at setup stage
 - > Overhead can be amortized



- Level-scheduling [Dutto, IJNMF 99]
 - > No dependencies in same level
 - > Barriers between levels
- ♦ NISCLU [Chen, TCAD 13]
 - > Dual-mode level scheduling
 - Same for upper levels
 - > Pipeline mode for bottom levels with very few nodes



Pipeline mode

- > Remove barriers between levels
- > Add dependency handling before row-update

```
Algorithm 1 Conventional ILU(0) factorization
Input: A \in \mathbb{R}^{N \times N}.
 1: for i = 2, ..., N do
                                                          ▷ up-looking(i)
        for k = 1, ..., i - 1 do
                                                       ▷ row-update(i, k)
  2:
                                                                              Wait until row k finished
             if a_{ik} \in S then
  3:
                 a_{ik} \neq a_{kk}.
  4:
                 for j = k + 1, ..., N do
  5:
                     if a_{ij} \in S and a_{kj} \in S then
  6:
                         a_{ij} = a_{ik}a_{kj}
  7:
```



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

Proposed Parallel ILU based Iterative Solver

Algorithm 3 ILU-GMRES solver (sketch)

Input: Equation Ax = b, initial guess $x^{(0)}$, $I_{restart}$. 1: Incomplete LU factorization $A \approx LU$. \rightarrow set up preconditioner 2: repeat Parallel ILU factorization $r = b - Ax^{(0)}.$ ▶ *compute initial residue* 3: $v^{(1)} = r/||r||_2, s_1 = ||r||_2.$ 4: for $i = 1, 2, ..., I_{restart}$ do 5: Parallel forward/backward $z^{(i)} = U^{-1}L^{-1}v^{(i)}$. right preconditioning6: substitutions $w = Az^{(i)}$. \triangleright sparse matrix-vector multiplication 7: ▶ modified Gram-Schmidt for k = 1, ..., i do 8: $h_{ki} = (\mathbf{w}, \mathbf{v}^{(k)}), \mathbf{w} = \mathbf{w} - h_{ki} \mathbf{v}^{(k)}.$ 9: $h_{i+1,i} = ||w||, v^{(i+1)} = w/h_{i+1,i}.$ **Others: Parallel BLAS** 10: Apply J_1, \ldots, J_{i-1} on $h_{:,i}$, and construct Givens rotation 11: J_i acting on h_{ii} and $h_{i+1,i}$ so that $(J_i h_{:,i})_{i+1} = 0$. $s = J_i s$. • E.g. Intel MKL Solve triangular system Hy = s. $x^{(0)} = x^{(0)} + \sum_i y_i z^{(i)}$. 12:13: **until** residual tolerence satisfied



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

Problems of level-based scheduling

Load-balancing

- > Barriers between levels
- > More levels to pipeline mode?
 - Overhead for dependency handling

Data locality

- Memory access order: 1 2 5 7 | 3 6 | 4 8 | 9 10
- Larger matrix -> poorer cache efficiency



Assume we have partitioned the rows into *P* balanced private queue and 1 small shared queue

- P: number of threads
- All tasks in one private queue have no dependencies on tasks in any other queue



private queue 1	1234
private queue 2	5678
shared queue	9 10

- Each thread works on its private queue
 - > In natural order: Better data locality
 - > No inter-thread communication
- After one finishing its private queue
 - > Work on shared queue in pipeline mode
 - > No barrier before shared queue
 - > Better load-balancing



private queue 1	1234
private queue 2	5678
shared queue	9 10

How to obtain the partition?

- ◆ Hard on arbitrary DAG, but easy on tree
- Elimination tree (ETree) [Gilbert, SIMAX 93]
 - Generated from DAG
 - > The dependencies of LU factorization for $L + L^{T}$
 - > Overestimates the dependencies
- Subtree partitioning



Algorithm 5 Build private queues based on ETree	
Input: Dependency DAG G, P threads, relaxing parameter α . 1: Build ETree T based on G.	- Setun
2: Subtree set $S = \{T\}$.	Secup
 3: Let w_S denote the total weight of subtrees in S. 4: Let l_S denote the root of the subtree with largest weight in S. 	$O(N \log N)$
5: \triangleright subtree partitioning 6: while $S \neq \emptyset$ and weight(subtree(l_S)) $\times P\alpha > w_S$ do	– Divide subtree
7: $u = l_S$. 8: $S = S \setminus {\text{subtree}(u)}.$	In most cases:
9: for all $v \in$ children of u do 10: $S = S \cup {$ subtree $(v) }.$	shared queue < 5%
11: Evenly assign each subtree in <i>S</i> to one of the <i>P</i> threads based on their weights.	– Assign task
12: Build <i>P</i> private queues according to the subtree assignment, and sort each queue in natural order.	– Build private queues
13: Build the shared queue with the remaining nodes, and sort the \triangleleft queue in level order of <i>G</i> .	 Build shared queues



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

- Very similar row-level parallelism to the ILU factorization
 - > Forward substitution: structure of *L*
 - \succ Backward substitution: structure of U
- Similar static scheduling
- Fewer FLOPs
 - Synchronization and memory access affects more
 - Subtree-based > level-based

• Parallel efficiency in shared queue remains to be a problem

• Task granularity: improve the data locality of the shared queue

- > At most g adjacent tasks are aggregated into a *task pack*
- > each thread takes a pack of tasks instead of one task

◆ To minimize the loss of parallelism: 2-round execution

Algo	orithm 1 Conventional ILU(0) factorization	n	
Inpu	it: $A \in \mathbb{R}^{N \times N}$.		
1: 1	for $i = 2, \ldots, N$ do	▶ up-looking(i)	
2:	for $k = 1,, i - 1$ do	▷ row-update(i, k)	If row k not finished
3:	if $a_{ik} \in S$ then		
4:	$a_{ik} \neq a_{kk}$.		- 1 st round: skip to next task
5:	for $j = k + 1,, N$ do		2 nd round wait
6:	if $a_{ij} \in S$ and $a_{kj} \in S$ then		
7:	$a_{ij} = a_{ik}a_{kj}$.		26

- Node stripping: minimize the size of shared queue
 - When reordering matrix to get a better ETree of the factorization, ETree of U tends to be badly degenerated



Figure 4: The dependency DAG and ETree for *U* in backward substitution. Note that ETree(*U*) is badly degenerated.

Node stripping: minimize the size of the shared queue

- > At most $\beta \times N$ nodes are stripped from the top levels of DAG
- > Top levels: level-based scheduling
- Rest of DAG: subtree-based scheduling



10 9
4321
8765
empty

 $\beta_1 = 20\%, \beta_2 = 50\%$ shared queue < 2%



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

Overview



- 1 setup + 100 factorization
 - > S: sequential version
 - L: dual-mode level scheduling (NISCLU)
 - P: proposed subtree-based scheduling
 - 8 threads
- Better speedup for larger/denser matrices

Table 1: Runtime of performing ILU(*k*) factorization for 100 times (including the static schedule for "L" and "P")

matrix	N	nnz		$T_{\rm ILU}$ (s)		Sp.		
	1	N	S	L	Р	P/S	P/L	
case9	3993	878	75.8	37.5	14.5	5.23	2.58	
case10	12155	691	156	45.5	30.3	5.15	1.50	
mult_dcop_01	25187	7.7	0.149	0.0824	0.0751	1.98	1.10	
rajat15	37261	20.2	1.07	0.288	0.232	4.62	1.24	
ckt11752_tr_0	49702	18.8	1.62	0.468	0.392	4.14	1.20	
ASIC_100ks	99190	5.8	0.372	0.105	0.0958	3.88	1.10	
ASIC_100k	99340	9.6	0.686	0.296	0.266	2.58	1.11	
trans4	116835	6.6	0.40	0.238	0.208	1.90	1.14	
transient	178866	6.0	0.636	0.276	0.242	2.63	1.16	
Raj1	263743	11.4	2.95	1.41	0.800	3.69	1.77	
nxp1	414604	10.7	4.42	2.46	1.13	3.90	2.17	

Level of fill (i.e. k) in ILU(k) varies among matrices for GMRES convergence. transient, rajat15, nxp1: ILU(1); Raj1, ckt11752_tr_0: ILU(5); the others: ILU(0).

- GMRES until convergence
 - ➤ relative tolerance: 10⁻⁴
- Higher speedup for substitution than factorization
- Around 4X (up to 4.5X) overall speedup with 8 cores

Table 2: Time for executing GMRES iterations with the ILU(k)preconditioners (same k as the experiment for Table 1)

matrix	matrix #iter		T _{subs} (ms)		Sp.		T _{GMRES} (ms)			Sp.	
	, 1001	S	L	Р	P/S	P/L	S	L	Р	P/S	P/L
case9	109	725	505	163	4.4	3.1	1107	649	309	3.6	2.1
case10	25	420	288	106	4.0	2.7	641	375	192	3.3	2.0
mult_dcop_01	17	10.1	4.78	3.24	3.1	1.5	20.8	8.10	6.51	3.2	1.2
rajat15	168	268	135	78.4	3.4	1.7	632	217	160	3.9	1.4
ckt11752_tr_0	51	111	52.4	24.2	4.6	2.2	237	81.5	53.2	4.5	1.5
ASIC_100ks	5	9.51	3.16	2.54	3.7	1.2	20.5	5.51	5.03	4.1	1.1
ASIC_100k	5	12.9	5.09	4.17	3.1	1.2	27.1	9.61	8.78	3.1	1.1
trans4	45	97.4	48.4	35.9	2.7	1.4	343	105	91.1	3.8	1.2
transient	180	561	286	187	3.0	1.5	2290	674	566	4.0	1.2
Raj1	427	3300	1955	1024	3.2	1.9	9096	3210	2273	4.0	1.4
nxp1	339	4087	2503	1372	3.0	1.8	11694	4078	2958	4.0	1.4

Results on power grid transient simulation

Table 3: Computational time for power grid transient simulation (unit in second)

circuit	KLU	S	L		Sp.				
	T_{tr}	T_{tr}	T_{tr}	#iter	Err(mV)	T_{tr}	P/K	P/S	P/L
ibmpg3t	115	172	79.7	14.7	2.0/1.0	38.9	3.0	4.4	2.1
ibmpg4t	138	124	56.7	15.1	0.20/0.07	35.7	3.9	3.5	1.6
ibmpg5t	96.9	177	89.8	11.2	0.84/0.56	46.0	2.1	3.8	2.0
ibmpg6t	117	247	123	11.2	1.4/0.82	61.4	1.9	4.0	2.0

 T_{tr} : time for linear equation solutions in simulation; #iter: average GMRES iterations; Err(mV): maximum/average errors of the solutions.



Figure 6: The transient simulation results at a VDD (up) and a GND (down) node in case "ibmpg3t", obtained with direct equation solver and the proposed ILU-GMRES solver.

- Results on nonlinear circuit simulation
 - > ngspice*(P): modified ngspice that uses proposed parallel iterative solver

Table 4: Time for DC analysis and each time-point advancing for transient simulation of nonlinear circuits (unit in second)

circuit	Ν	#Dev.	HS	PICE	ngs	spice	ngspice*(P)	
			$T_{\rm DC}$	<i>T</i> trans	$T_{\rm DC}$	<i>T</i> _{trans}	$T_{\rm DC}$	<i>T</i> trans
ckt1	164899	47614	32.8	0.741	945	45.8	2.9	0.190
ckt2	1043446	201056	38.4	2.25	fail	fail	25.9	1.90
ckt3	1214290	265946	128	15.3	fail	fail	29.8	2.03

#Dev: the number of nonlinear devices; T_{DC} : time for DC analysis including setup time; T_{trans} : average time for a time step.

Results on nonlinear circuit simulation

Table 5: The total time for equation solutions with "S", "L" and "P" in the transient simulation based on ngspice^{*}

circuit	ILU-GMRES time (s)			Sp.	
	S	L	Р	P/S	P/L
ckt1	10.0	3.08	2.55	3.92	1.21
ckt2	170	79.5	44.7	3.80	1.78
ckt3	209	96.8	52.7	3.97	1.84



- Background
- Related work
- Proposed Parallel ILU based Iterative Solver
 - Subtree-based parallel ILU factorization
 - Parallel forward/backward substitutions
- Numerical Experiments
- Conclusions

Conclusions

- In this work, we have presented a subtree-based parallel ILU factorization and forward/backward substitution algorithm for fixed-structure matrices in circuit simulation, and an ILU-GMRES solver incorporating them.
- With matrices from circuit simulation, the proposed ILU-GMRES solver using 8 threads is up to 5.23X faster than the sequential baseline for solving fixedstructure equations.
- Experiments involving transient simulation of linear and nonlinear circuits show that the proposed parallel algorithms are faster than the parallel baselines derived from the dual-mode level scheduling in NICSLU with up to 2.1X speedup.
- The large benefit over HSPICE for solving large nonlinear circuits is also demonstrated.

References

- R. Akhunov, S. Kuksenko, V. Salov, and T. Gazizov. Optimization of the ILU(0) factorization algorithm with the use of compressed sparse row format. *Journal* of Mathematical Sciences, 191(1):19–27, 2013.
- [2] E. Anderson and Y. Saad. Solving sparse triangular linear systems on parallel computers. *International Journal of High Speed Computing*, 1(01):73–95, 1989.
- [3] M. Benzi, W. Joubert, et al. Numerical experiments with parallel orderings for ILU preconditioners. *Electronic Trans. Numeric Analysis*, 8:88–114, 1999.
- [4] L. Blackford, A. Petitet, et al. An updated set of basic linear algebra subprograms (BLAS). ACM Trans.s on Mathematical Software, 28(2):135–151, 2002.
- [5] X. Chen, Y. Wang, and H. Yang. NICSLU: An adaptive sparse matrix solver for parallel circuit simulation. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 32(2):261–274, 2013.
- [6] X. Chen, Y. Wang, and H. Yang. *Parallel Sparse Direct Solver for Integrated Circuit Simulation*. Springer, 2017.
- [7] E. Chow and A. Patel. Fine-grained parallel incomplete LU factorization. *SIAM J. Scientific Comput.*, 37(2):C169–C193, 2015.
- [8] T. Davis and Y. Hu. The University of Florida sparse matrix collection. ACM Trans. Math. Softw., 38(1):1–25, 2011.
- [9] T. A. Davis and E. Palamadai Natarajan. Algorithm 907: KLU, A Direct Sparse Solver for Circuit Simulation Problems. *ACM Trans. Math. Softw.*, sep 2010.
- [10] L. Dutto and W. Habashi. Parallelization of the ILU(0) preconditioner for CFD problems on shared-memory computers. *International J. numerical methods in fluids*, 30(8):995–1008, 1999.
- [11] J. Gilbert and J. Liu. Elimination structures for unsymmetric sparse LU factors. SIAM J. Matrix Analysis and Applications, 14(2):334–352, 1993.
- [12] M. Heroux, P. Vu, and C. Yang. A parallel preconditioned conjugate gradient package for solving sparse linear systems on a Cray Y-MP. *Applied Numerical Mathematics*, 8(2):93–115, 1991.
- [13] D. Hysom and A. Pothen. Efficient parallel computation of ilu (k) preconditioners. In Proc. ACM/IEEE conference on Supercomputing, pages 29–es, 1999.
- [14] D. Hysom and A. Pothen. A scalable parallel algorithm for incomplete factor preconditioning. SIAM J. Scientific Comput., 22(6):2194–2215, 2001.

- [15] Intel Corporation. Developer Reference for Intel® oneAPI Math Kernel Library -C. https://www.intel.com/content/www/us/en/develop/documentation/onemkldeveloper-reference-c/top.html, 2022.
- [16] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Scientific Comput., 20(1):359–392, 1998.
- [17] S. Nassif. IBM power grid benchmarks. https://web.ece.ucsb.edu/~lip/ PGBenchmarks/ibmpgbench.html, 2008.
- [18] Ngspice. Open source spice simulator. http://ngspice.sourceforge.net/, 2022.
- [19] Y. Saad. ILUT: A dual threshold incomplete LU factorization. *Numerical linear algebra with applications*, 1(4):387–402, 1994.
- [20] Y. Saad and M. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Scientific and Statistical Computing*, 7(3):856–869, 1986.
- [21] K. Shen, X. Jiao, and T. Yang. Elimination forest guided 2D sparse LU factorization. In Proc. ACM SPAA, pages 5–15, 1998.
- [22] Synopsys, Inc. PrimeSim HSPICE. https://www.synopsys.com/implementationand-signoff/ams-simulation/primesim-hspice.html, 2022.
- [23] H. Thornquist, E. Keiter, R. Hoekstra, D. Day, and E. Boman. A parallel preconditioning strategy for efficient transistor-level circuit simulation. In *Proc. ICCAD*, pages 410–417, 2009.
- [24] A. Vladimirescu. The SPICE Book. Wiley New York, 1994.
- [25] X. Xu. OpenMP parallel implementation of stiffly stable time-stepping projection/GMRES(ILU(0)) implicit simulation of incompressible fluid flows on sharedmemory, multicore architecture. *Applied Mathematics and Computation*, 355:238– 252, 2019.
- [26] J. Zhao, Y. Wen, et al. SFLU: Synchronization-free sparse LU factorization for fast circuit simulation on GPUs. In *Proc. ACM/IEEE DAC*, pages 37–42, 2021.
- [27] X. Zhao and Z. Feng. GPSCP: A general-purpose support-circuit preconditioning approach to large-scale SPICE-accurate nonlinear circuit simulations. In *Proc. IEEE/ACM ICCAD*, pages 429–435, 2012.

Thank You !