量子输运模拟中的并行二次选择性求逆

1st Vincent Maillou*D-ITET*，苏黎世联邦理工学院苏黎世，瑞士 0000-0003-4861-3298

2nd Matthias Bollhofer数值分析研究所，*TUBS*不伦瑞克，德国 0000-0002-8093-5812

3rd Olaf Schenk计算研究所，*USI*卢加诺，瑞士 0000-0001-8636-1023

4th Alexandros Nikolaos Ziogas*D-ITET*，苏黎世联邦理工学院苏黎世，瑞士 0000-0002-4328-9751

5th Mathieu Luisier*D-ITET*，苏黎世联邦理工学院苏黎世，瑞士 0000-0002-2212-7972

arXiv:2601.04904v1 [cs.DC] 8 Jan 2026

摘要——受摩尔定律推动，晶体管的尺寸已缩小至纳米尺度。要准确仿真此类纳米器件，需要先进的量子输运（QT）求解器。非平衡格林函数（NEGF）形式非常适用于这些任务，但计算开销极大，涉及矩阵的选定求逆（SI）以及二次矩阵方程的选定解（SQ）。现有用于解决这些数值问题的算法非常适合采用 GPU 加速，例如所谓的递归格林函数（RGF）技术；然而，这些算法通常是顺序执行的，需要以块三对角（BT）矩阵作为输入，而且迄今为止，其实现仅限于共享内存并行，因此限制了可仿真的器件尺寸。为解决这些不足，我们提出了基于 RGF 的分布式方法，可实现选定求逆和二次矩阵方程选定解的并行计算。我们还将这些方法扩展至处理带箭头块的 BT 矩阵，从而能够研究多端晶体管结构。我们使用一个来自纳米带晶体管量子输运仿真的真实数据集评估了所提方法的性能，并将其与稀疏直接求解软件包 PARDISO 进行比较。在扩展至 16 个 GPU 时，我们融合后的 SI 和 SQ 求解器比 PARDISO 的 SI 模块在短 16× 的器件上快 5.2× 倍。这些结果凸显了我们的方法在加速基于 NEGF 的纳米器件仿真方面的潜力。

索引词——选定求逆、半导体器件建模、分布式算法、线性代数

I. 引言

为了跟上摩尔定律，并逐代提升纳米尺度晶体管的性能，器件工程师提出了许多创新方案，例如应变工程、高-κ 介电层或三维鳍式场效应晶体管（FinFET）。至关重要的是，晶体管的栅电极应能有效控制带电载流子（电子或空穴）在源极和漏极接触之间的流动，同时防止所谓的“栅极漏电流”。满足这些条件有助于最大限度地降低待机功耗，并实现逻辑 0 和 1 之间的高速数字开关。硅基 FinFET自2011年以来一直是半导体行业的主力产品，十多年来都实现了这些目标，但其三栅极结构

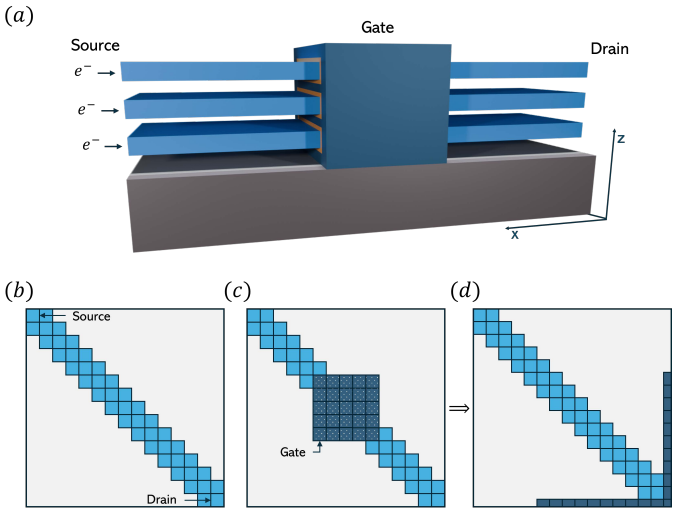


图 1：(a) 环栅结构纳米带场效应晶体管的示意图 [1]–[3]。三条纳米带上下堆叠，构成器件沟道。电子（或空穴）可从源极（左侧）、漏极（右侧）和栅极（中间）接触处注入器件结构，随后从这些位置之一流出。(b)–(d) 此类器件在量子输运模拟中产生的典型哈密顿矩阵稀疏模式。(b) 描述经典双端晶体管沟道（仅有源极和漏极）的块三对角矩阵。(c) 含有较大但稀疏的中心块的块三对角矩阵，进一步考虑了中央栅极接触处的电子/空穴注入与收集。(d) 将 (c) 中的矩阵重新排序为具有箭头形稀疏模式的块三对角矩阵。

这种架构在栅长为 15 nm 及以下时无法提供所需的控制能力。因此，它们正逐渐被采用环栅（GAA）结构的纳米带（或纳米片）场效应晶体管（NR-FET）所取代 [3]–[5]，如图 1 所示。

	求解器库	运算			稀疏性		后端		
		分解	选定求逆	选定解	BT	BTA	CPU	GPU	分布式内存
	本工作	✓	✓	✓	✓	✓	✓	✓	✓
Libraries	SERINV [6]	PO	PO		✓	✓	✓	✓	✓
Libraries	PARDISO* [7], [8]	✓	✓		通用		✓		
Libraries	MUMPS [9]–[11]	✓	~		通用		✓		✓
Libraries	SCALAPACK [12]	✓			带状 / 稠密		✓		✓
Libraries	CUSOLVERMP* [13]	✓			稠密			✓	✓
Alg. w. impl	RGF [14], [15]	✓	✓	✓	✓		✓	✓	
Alg. w. impl	$INLA_{BTA}$ [16]	PO	PO		✓	✓		✓	
Alg. w. impl	SPLITSOLVE [17]	✓	~		✓			✓	✓
Alg. w. impl	PSELLINV [18], [19]	✓	✓		通用		✓		✓
Alg. w.o. impl	FIND [20]–[23]	✓	✓	DIAG	通用		✓		✓
Alg. w.o. impl	PSR、P-DIV/SPIKES [24]–[27]	✓	✓		✓		✓		✓

注释：PO：正定矩阵，DIAG：仅限对角线条目，$\sim$：不适用于此任务。

表I：BT 和 BTA 结构化稀疏矩阵的选定求逆（$AX = I$）和二次方程（$AXA^\dagger = B$）选定解问题中可用算法和/或库的总结。文中标明了中央处理器或图形处理器实现的可用性，以及并行化模型。我们的工作以蓝色突出显示，相较于成熟的 RGF 算法所作的创新则以虚线箭头标示。我们用 $*$ 表示闭源软件包。

多项晶体管实验开发工作得到了基于物理的技术计算机辅助设计（TCAD）工具的支持，在某些情况下甚至由其促成 [28]–[30]。这类器件模拟器能够揭示影响已制造器件功能的内在机制，同时为设计新器件提供可靠的指导。要具备实际应用价值，现代 TCAD 软件包应基于第一性原理（不使用经验输入参数）[31], ，考虑所有接触处的电子/空穴注入 [32],，提供仿真域的量子力学描述，纳入限制纳米尺度晶体管性能的主要散射机制，尤其是表面粗糙度散射 [33], 、电子-声子 [34] 和电子-电子 [35] 相互作用，并逐一处理构成目标结构的每个原子。

非平衡格林函数（NEGF）方法能够在单一框架内自然地结合所有这些效应 [36]。然而，由于其计算负担较重，通常仅限于由数百个原子组成的小型系统，但也有少数值得注意的例外，例如 [37]。NEGF 的核心运算是计算不同类型的格林函数：$AG^R = I$ 推迟分量 G^R 和 $AG^<, >A^\dagger = \Sigma^<, >$ 其小于 $G^<$和大于 $G^>$ 对应进程的格林函数，由此可以推导出大多数可观测量（载流子和电流密度）[36]。其中，大小为 A 的矩阵 $N_{AO} \times N_{AO}$（N_{AO} 是器件结构中的原子轨道总数）与哈密顿矩阵 H 直接成正比，后者涵盖了所有材料属性。大小同为 $N_{AO} \times N_{AO}$ 的自能矩阵 $\Sigma^<, >$考虑了与半无限接触的耦合（开放边界条件）以及散射机制。

假设电子（或空穴）仅在图 1(a) 所示的源极和漏极接触处进入仿真域，且任意两个相互作用粒子（电子-声子或电子-电子）之间的最大距离不超过截断半径 r_{cut}，A 和 $\Sigma^<, >$ 是具有块三对角（BT）结构的结构化稀疏矩阵，如图 1(b) 所示。此外，计算所需的观测量时，只需用到格林函数矩阵 $G^R, <, >$ 的选定元素，即与 A和 $\Sigma^<, >$ [38]的初始稀疏模式相对应的元素。然而，若采用标准方法求值，格林函数在计算过程中会变成稠密矩阵。例如，对 A求逆以得到 G^R会产生一个完整矩阵，其中大多数元素都是不必要的。由于所考虑晶体管的尺寸可能很大（$N_{AO} \propto O(10^5 - 10^6)$），计算 $G^R, <, >$ 时出现的稠密化，会使此类运算在计算和内存方面都变得不可行。因此，从数值角度看，采用 NEGF 对纳米器件建模，能从选定线性代数方法中显著获益，例如选定求逆和选定解，它们能够直接且仅生成所需的格林函数矩阵元素。这也是器件建模领域开发多种方法来计算$G^R, <, >$选定元素的原因，其中的黄金标准是所谓的递归格林函数（RGF）算法 [14], [39]。值得注意的是，RGF 可联合求解小于和大于格林函数的二次方程 $AG^<, >A^\dagger = \Sigma^<, >$ 的选定解，同时对推迟格林函数对应的 A 进行选定求逆。这些特性将在第二节进一步讨论。

RGF 可直接应用于矩阵 A 中各块大小不同的晶体管结构。例如，当器件的横截面在源极与漏极之间发生变化，或者电子穿过将晶体管沟道与栅极接触隔开的绝缘层发生隧穿的概率不可忽略时，就会出现这种情况。为避免待机功耗过高，应尽量减小由此产生的栅极漏电流。为考虑这些漏电流，必须在栅电极处注入并收集电子，这会在矩阵 A 的中央引入一个大型、高度稀疏的块，

如图1(c)所示。该块源于开放边界条件，这些条件将所有连接到栅极区域的原子与中央器件区域 [32]相连。由于 RGF 需要对 A的所有对角块求逆，因此矩阵中央存在一个大型单元会导致计算性能欠佳，并造成较高的内存消耗。不过，经过适当的置换，该矩阵仍可转换为箭头形块三对角（BTA）结构（图1(d)），从而形成有利的计算模式。

尽管 BT 和 BTA 矩阵在量子输运模拟中都发挥着重要作用，但 RGF 算法仅适用于 BT 稀疏模式，从而无法研究具有两个以上接触的真实原子系统。如果在现代图形处理器上实现，对于由最多 $N_{A0} \simeq$40,000 个轨道组成的均匀器件，它可以返回$_{G^{R,<,>}}$ 的选定元素，具体取决于结构横截面（图 1(a) 中的 y-z 平面）与其长度（x 轴）之比。不过，由于其块顺序特性，RGF 在块级别上仅限于共享内存（SM）并行，这妨碍了对 A 矩阵无法装入单个节点内存的原子系统进行模拟。

A. 相关工作

人们已尝试并行计算推迟格林函数和小于/大于格林函数。用于并行求解推迟格林函数的分布式内存（DM）算法包括 PSELLINV [18],PSR [24], 和 SPLITSOLVE [17]。尽管这些方法效率较高，但由于无法返回 $G^<$ 和 $G^>$，因此仅适用于弹道输运模拟。这些量是描述散射机制所必需的 [36]。人们也尝试并行求解二次矩阵方程，该方程可返回$G^{<,>}$；但所提出的方法假设散射自能矩阵 $\Sigma^{<,>}$ 具有（块）对角稀疏模式。典型示例包括 FIND [23] 和 P-DIV [26]。此外，这些方法均未移植到 GPU 上，因此在当今的混合超级计算机上的实用性有限。另一种选择是采用 PARDISO [7] 或 MUMPS [9]等先进的稀疏直接求解器，处理量子输运中遇到的大型线性系统。然而，在求解二次矩阵方程的选定解时，这两种求解器都存在明显局限：它们不支持 GPU 加速，且 PARDISO 缺乏 DM 能力（见表I）。另一方面，SCALAPACK 的块带状对角占优例程（PZGBSV）等分布式结构化求解器或许不具备选定求逆能力，但通常具有较好的可扩展性 [12]。尽管如此，由于其计算复杂度极高且内存需求很大，它们并不适合执行量子输运模拟。第II-C节将进一步详述这些问题，表I则全面概述了本文介绍的算法。

B. 我们的贡献

本工作首先介绍将 RGF 算法扩展至 BTA稀疏矩阵。随后，我们提出由 RGF 推导而来的新型并行算法，用于在分布式内存环境中执行选定求逆，并计算二次矩阵方程的选定解。所提出的方法利用分块高斯消元和区域分解技术，可对大型原子结构进行基于非平衡格林函数的量子输运模拟，同时考虑散射效应，并支持多端注入。本求解器适用于中央处理器和图形处理器架构，其主要特性在表I中以蓝色标出。我们的方法对于突破单节点/GPU的 RGF 内存限制、扩展纳米器件模拟规模至关重要。本工作的主要贡献如下：

- 将 RGF 算法扩展到 BTA稀疏矩阵；
- 针对 BT 和 BTA稀疏矩阵情况下二次矩阵方程的选定解，推导分布式内存算法；
- 将所推导的方法移植到图形处理器上；
- 在 NR-FET 晶体管数据集上对我们实现的性能进行评估（弱扩展），并与 PARDISO 进行比较。
- 评估我们基于图形处理器加速的实现的并行效率（弱扩展和强扩展性），最多使用 32 个 GPU。

本文共分为五个部分。第二节介绍所需的数学和算法背景，并给出 RGF 算法。第三节推导我们提出的新方法，随后在第四节对已实现的顺序代码和分布式代码进行数值评估。最后，第五节总结全文。

II. 背景

A. 问题表述

选定线性代数指一类数值方法，这类方法不仅考虑输入数据的稀疏性（例如， A 中的系统矩阵 $AX = B$），也考虑解空间的稀疏性（例如，输出矩阵 X）。这类方法的主要优势在于，它们只返回 X 中所需的元素，从而避免对解进行稠密表示，否则其计算和/或存储可能难以实现。在 NEGF 框架下进行量子输运模拟时，有两个方程尤为重要： $AX = I$和 $AXA^\dagger = B$。在这两种情况下， A 和 B 都是稀疏矩阵，通常呈现块三对角或 BTA 稀疏模式。通常，只需提取 A 和 B （这两个矩阵具有相同结构）的非零稀疏模式所对应的 X 中的元素，即可获得所需信息。A、 B和 X 的块三对角（带箭头）模式可以理想地划分为块三对角（带箭头）模式。此划分假设 n 为方形对角块。为简单起见，这里仅考虑所有对角块大小相同且尺寸为 b 的情况。该参数化如图2(a)所示。对于 BTA，箭头块的尺寸

Name	描述
	矩阵参数
n	构成矩阵主对角线的方块矩阵块数量
b	BT(A) 矩阵（不包括箭头尖端块）。 BT(A) 矩阵的对角块和上／下对角块的大小（不包括箭头块）。
a	BTA 矩阵中箭头尖端块的大小。
N	BT（BTA）矩阵的总大小： nb （$nb + a$）。
P	并行矩阵分区的数量。当置于 名称前时， P 表示该算法的并行版本。
	算法概念
块顺序	指以矩阵块 运算表示的算法和例程，彼此之间存在顺序依赖关系。
前向/后向 Pass	从矩阵左上角块开始，或从右上角块开始执行的块顺序循环（分块后的）矩阵的右下角块，向 右下角块或左上角块执行。
融合运算	指一种算法，它能够同时从同一组操作数中生成多个结果，从而有效融合这些运算 并降低总体复杂度。
/ (\)	前向（后向）三角求解
	矩阵类型与算法
DD	一般对角占优（或块对角占优） 矩阵。
块三对角（带箭头）	块三对角（带箭头块）矩阵。
SI	系统 $AX = I$ 的选定求逆
SQ	二次矩阵方程的选定解 $AXA^H = B$.
SC	舒尔补算法。
SCI（SCQ）	基于选定求逆（二次矩阵方程解） 的舒尔补计算。
RGF	递归格林函数算法。

表 II：本工作中使用的符号和术语列表。

为 a （见图 3）。本工作中使用的这些符号及其他所有符号均在表 II 中定义。

以下各节所述算法通常由两个阶段组成：矩阵分解，称为前向遍历；随后是三角求解例程的组合，称为后向遍历。一般情况下，如果系统矩阵具有（块）对角占优（DD）性质，则可确保数值稳定性 [41]，；量子输运问题中的矩阵以及更广泛地描述最近邻类型相互作用的矩阵都满足这些性质。

B. RGF 算法

递归格林函数（RGF）算法对基于 BT 矩阵的系统执行选定求逆，并求解二次矩阵方程。它是非平衡格林函数形式下量子输运计算的事实标准方法，因为它只需计算 $X = A^{-1}$ 和 $X = A^{-1}BA^{-\dagger}$的选定元素。

从数值角度来看，RGF 依赖于系统矩阵的稀疏舒尔补 S_A （前向遍历）和选定求逆（反向遍历），这相当于将 Takahashi [42] 推导的算法限制于 A的 BT 稀疏模式。为此，在前向遍历中进行块分解$A = LDU$，其中 L，U 具有单位块对角部分。出于效率考虑，只存储块对角矩阵 D，而 L和 U 的元素则在反向遍历期间临时计算。对 D 求逆可得到 $S_A = D^{-1}$。RGF 可扩展为采用第二个舒尔补 S_B，它指的是对角

算法 1 RGF 舒尔补（前向过程）

输入： 块三对角矩阵： A和 B。

1: **对于** $n_i = 0$到 $n - 2$ **执行**
2: $S_{A_{i,i}} \leftarrow A_{i,i}^{-1}$
3: $S_{B_{i,i}} \leftarrow S_{A_{i,i}} B_{i,\ i} S_{A_{i,i}}^{\dagger}$
4: $\mathrm{T}_1^A \leftarrow A_{i+1,i} S_{A_{i,\ i}}$
5: $A_{i+1,i+1} \leftarrow A_{i+1,i+1} - \mathrm{T}_1^A A_{i,i+1}$
6: $B_{i+1,i+1} \leftarrow B_{i+1,i+1} + A_{i+1,i} S_{B_{i,i}} A_{i+1,i}^{\dagger} - B_{i+1,i}\mathrm{T}_1^{A\dagger} - \mathrm{T}_1^A B_{i,\ i+1}$
7: **结束循环**
8: $S_{A_{n-1,\ n-1}} \leftarrow A_{n-1,\ n-1}^{-1}$
9: $S_{B_{n-1,\ n-1}} \leftarrow S_{A_{n-1,n-1}} B_{n-1,\ n-1} S_{A_{n-1,n-1}}^{\dagger}$

选定求逆　选定解二次方程

算法2 递归格林函数选定求逆与选定求解（反向遍历）

输入： 舒尔补矩阵： S_A 和 S_B。

1: $X_{A_{n-1,n-1}} \leftarrow S_{A_{n-1,n-1}}$
2: $X_{B_{n-1,\ n-1}} \leftarrow S_{B_{n-1,\ n-1}}$
3: **对于** $i = n - 2$到 0 **执行**
4: $\mathrm{T}_1^A \leftarrow S_{A_{i,\ i}} A_{i,\ i+1}$, $\mathrm{T}_2^A \leftarrow X_{A_{i+1,\ i+1}} A_{i+1,\ i}$,
5: $X_{A_{i+1,\ i}} \leftarrow -\mathrm{T}_2^A S_{A_{i+1,\ i}}$
6: $X_{A_{i,i+1}} \leftarrow -\mathrm{T}_1^A X_{A_{i+1,i+1}}$
7: $X_{A_{i,i}} \leftarrow S_{A^{i,i}} - \mathrm{T}_1^A X_{A_{i+1,i}}$
8: $\mathrm{T}_3^B \leftarrow X_{B_{i+1,\ i+1}} \mathrm{T}_1^{A\dagger}$, $\mathrm{T}_2^B \leftarrow S_{B_{i,\ i}} \mathrm{T}_2^{A\dagger}$, $\mathrm{T}_3^B \leftarrow \mathrm{T}_2^A S_{B_{i,\ i}}$
9: $\mathrm{T}_4^B \leftarrow S_{A_{i,i}} B_{i,i+1} X_{A_{i+1}^{\dagger} i+1}^{\dagger}$, $\mathrm{T}_5^B \leftarrow X_{A_{i+1,i+1}} B_{i+1,i} S_{A_{i}^{\dagger} i}^{\dagger}$
10: $X_{B_{i,i+1}} \leftarrow -\mathrm{T}_1^A X_{B_{i+1,i+1}} - \mathrm{T}_2^B + \mathrm{T}_4^B$
11: $X_{B_{i+1,i}} \leftarrow -\mathrm{T}_1^B - \mathrm{T}_3^B + \mathrm{T}_5^B$
12: $X_{B_{i,\ i}} \leftarrow S_{B_{i,\ i}} + \mathrm{T}_1^A \mathrm{T}_1^B + \mathrm{T}_1^A \mathrm{T}_3^B + \mathrm{T}_2^B \mathrm{T}_1^{A\dagger} - \mathrm{T}_1^A \mathrm{T}_5^B - \mathrm{T}_4^B \mathrm{T}_1^{A\dagger}$
13: **结束循环**

选定求逆　选定解二次方程

块 $S_A L^{-1} B L^{-\dagger} S_A^{\dagger}$，将二次矩阵方程的求解与选定求逆融合，从而仅返回解空间中的选定元素，降低整体计算复杂度。RGF 算法被称为递归算法，因为它在舒尔补（选定求逆+二次矩阵方程）的前向（反向）过程中，沿 BT 矩阵的对角块迭代。在步骤 i中，计算一个舒尔补，并将其作为更新传递到下一个对角块。图2(a) 展示了 BT 矩阵的这一中间步骤。块运算以及 RGF 前向遍历的完整描述见算法1。与 A的选定求逆相关的运算以绿色标出，而与二次矩阵方程选定解相关的运算则以紫色标出。

在前向遍历过程中，RGF 在每一步都会对对角块 A_i 进行 LU 分解$_{i,}$，对下（上）对角块A_{i+1}进行三角矩阵求解$_{i}$ （$A_{i,\ i+1}$），并执行若干次通用矩阵乘法（*gemm*）。为处理二次矩阵方程的补项 S_B，还需要额外的三角求解和矩阵乘法。在 GPU 上实现 RGF 时，最好计算对角块的逆矩阵，并通过矩阵乘法而非三角求解例程来执行前向更新。这一“技巧”使得可以在块级别进行主元选取

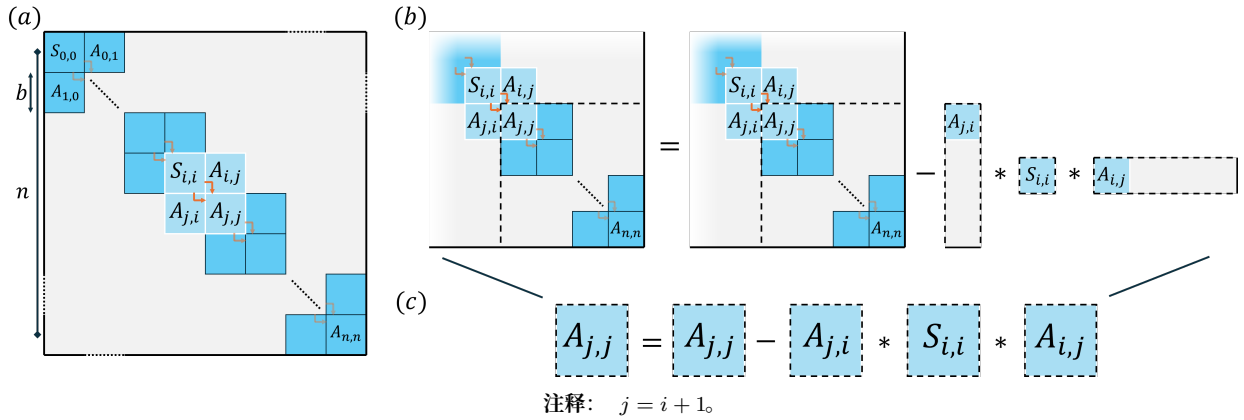


图2：针对块三对角（BT）结构稀疏矩阵，RGF 算法前向过程推导的高层概览。(a) 输入 A 矩阵的 BT 稀疏模式，以及前向过程第 i 步所涉及的块。(b) 在第 i 步对矩阵进行切片，并放大显示前向更新中涉及的块运算。(c) 通过避免先验零值计算，简化 (b) 中的块运算。

而无需进行面板置换 [43],，并能增加矩阵乘法的数量，这对高性能 GPU 实现至关重要。算法1采用了这种形式。

反向选定求逆以及二次矩阵方程的求解，是从最后一个对角块（右下方）朝向第一个对角块（左上方）进行的逐步过程，反转了前向过程中的依赖关系。逆矩阵（二次方程）的非对角块$X_{A/B_{i,\,j}}$ 以及 $X_{A/B_{j,\,i}}$ 可以显式计算（算法2第 6、7、9 和 10 行）；或者，当只需要对角块时，也可以舍弃这些块，转而采用经过约简的暂态结果。

RGF 算法面临两项主要限制，阻碍了量子输运模拟的可扩展性以及可处理的原子数量。首先，它仅适用于 BT 稀疏模式。在大多数情况下，可以将稠密 BT 分块适配于描述器件的系统矩阵，但代价是产生额外的填充和针对零元素的冗余计算。在采用合适的置换后，这些结构通常可以置换为 BTA 稀疏模式，从而减少前向过程引入的填充。其次，RGF 算法在其前向过程和反向过程中表现出块级顺序依赖，这使其实现受限于共享内存（细粒度）并行，因此可研究的纳米器件规模也受限于单个图形处理器的内存容量。

C. 求解 $AXA^{\dagger} = B$ 的其他方法

如第 I 节所述，除了 RGF 之外，还有几种方法可以对矩阵 A 进行选定求逆，PARDISO 和 MUMPS 中也提供了相应的专用功能。对于二次方程 $AXA^{\dagger} = B$，情况则有所不同。一种贪心方法是先对系统矩阵 A 进行 LU 分解，再对右端矩阵 B 进行一系列三角矩阵求解。我们考虑这种方法的两种变体：第一种见第II-C1节，采用稀疏方法分解 A，并对前向/后向代入进行批处理。第二种见第II-C2节，先对 A 进行稠密分解，再进行简洁的前向/后向代入，从而得到完整解 $X = A^{-1}BA^{-\dagger}$。

1) 稀疏求解器： 通用稀疏求解器（包括用于带状矩阵的求解器）在分解过程中保持稀疏性，但可能会产生额外的填充。对于带状矩阵或块三对角（BT）矩阵，这种填充为零。对于 $AX = I$ 和 $AXA^{\dagger} = B$ 的选定解，假设需要包含 X 的所有列的条目，则基于 LU 对 A 的 LU 分解获取 X 会得到一个稠密结果。为缓解这一问题，可以采用批处理方法，如算法 3 所述。对于给定切片 $b_i = B[:, i]$ 的每次三角求解之后，可以将不需要的条目屏蔽掉，并重复该运算，直到考虑完 B 的所有列。该方法的计算复杂度等价于先求 A 的稠密逆矩阵，再进行矩阵乘法 $A^{-1}BA^{-\dagger}$。不过，批处理可确保所需总内存受限于所需的解空间。

此外，最先进的通用稀疏求解器缺乏 GPU 加速能力。这类求解器通常基于填充重排序方法产生的消元图 [44], [45],，并且往往依赖稠密但尺寸较小的块 [46]（例如，PARDISO 中的超节点，MUMPS 中的前沿），在这些块上可以使用三级 BLAS 内核。这些块的维度使此类求解器适合在多线程中央处理器上执行，但限制了可利用的并行度，因而影响其在 GPU 上的性能。

算法 3 稀疏/结构化批量求解

1: $L,\ U \leftarrow$ LU 分解(A)2: **对** $i = 0$ 到 N **执行**3: $b_i = B[:$ $i]$4: $y_i = U \backslash (L / b_i)$ 5: $x_i = ($ $y_i \backslash L^{-\dagger}) / U^{-\dagger}$6: **结束循环**

2) 稠密求解器： 稠密方法涉及对系统矩阵进行稠密（或带状）分解，并显式表示 A 的逆矩阵。所得的 A^{-1} 随后可用于通过与右端矩阵 B 相乘来求解二次矩阵方程，如算法 3 所述。通过移除所得稠密解中不需要的条目，可以将结果稀疏化。在这种方法中，计算复杂度为 $O(N^3)$，内存占用为 $O(N^2)$。

算法 4 稠密求解

1: $L, U \leftarrow$ LU 分解(A)2: $Y = U \backslash (L/B)$ 3: $X = (Y \backslash L^{-\dagger})/U^{-\dagger}$

3) 总结： 上述用于求解 $AXA^{\dagger} = B$ 的 RGF 替代方法存在局限性，因此不适用于实际应用场景。表 III 总结了它们的复杂度，并将其与 RGF 算法进行了比较。为完整起见，我们还介绍了这些方法的分布式内存变体的特性。如果 $b \simeq N$，稀疏和稠密方法在计算上都具有吸引力。然而，在典型的纳米器件模拟中，我们处理的是 $2b \ll N$ 矩阵，即其带宽远小于矩阵尺寸。为处理这类情况，本文提出了新方法，可在 CPU 和 GPU 上实现类 RGF 算法前向和后向遍历的高分布式内存可扩展性，包括二次矩阵方程的选定解。

	方法	分解 (前向过程)		选定解二次方程 (反向遍历)	
		计算	内存	计算	内存
Seq.	RGF	$O(nb^3)$	$O(nb^2)$	$O(nb^3)$	$O(nb^2)$
	稀疏	$O(nb^3)$	$O(nb^2)$	$O(N^3)$	$O(nb^2)$
	稠密	$O(N^3)$	$O(N^2)$	$O(N^3)$	$O(N^2)$
Dist.	本工作	$O(\frac{nb^3}{P} + Pb^3)$	$O(\frac{nb^2}{P} + Pb^2)$	$O(\frac{nb^3}{P} + Pb^3)$	$O(\frac{nb^2}{P} + Pb^2)$
	稀疏	$O(\frac{nb^3}{P} + Pb^3)$	$O(\frac{nb^2}{P} + Pb^2)$	$O(N^3/P)$	$O(\frac{nb^2}{P})$
	稠密	$O(N^3/P)$	$O(N^2/P)$	$O(N^3/P)$	$O(N^2/P)$

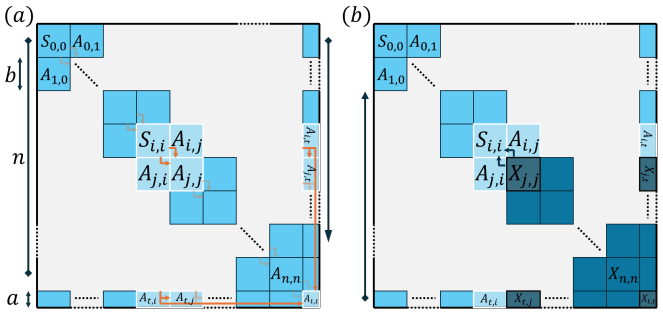
表 III：稠密、稀疏和 RGF 方法在计算 $AXA^{\dagger} = B$ 时的复杂度和内存占用概述，其中对象是具有参数 $[n, b]$ 的块三对角矩阵。以红色 突出显示的是会导致不可行性的复杂度。

III. 方法

我们介绍两种计算二次矩阵方程 $AXA^{\dagger} = B$ 的选定解的方法。首先，我们将 RGF 算法扩展到 BTA 稀疏模式，从而能够对更广泛的一类纳米器件进行建模。其次，我们提出基于区域分解和约化系统组装的分布式方法，适用于 BT 和 BTA 稀疏模式。

A. 将 RGF 扩展到 *BTA* 矩阵

RGF 算法适用于具有 BT 稀疏性的矩阵。为处理 BTA 稀疏模式，我们给出一种与此前介绍的方法类似的高层次推导

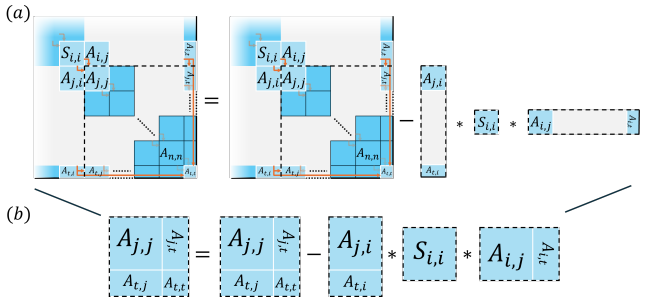


注： $j = i + 1$。

图 3：将 RGF 算法扩展到具有块三对角箭头形稀疏模式的矩阵。(a) 前向舒尔补运算的第 $i - th$ 步中系统矩阵的切片。(b) 反向选定求逆运算的第$i - th$ 步中系统矩阵的切片。橙色和海军蓝箭头展示了步进方向，并指出所执行的块运算。

见第 II-B 节，重点关注前向舒尔补阶段。该推导基于给定步骤 i 中块计算的可视化表示。我们考虑 BTA 矩阵 A 的切片 $A_{i:,\ i:}$，如图 3(a) 所示。橙色箭头表示迭代方向，从左上角的对角块向下指向箭头尖端。浅蓝色叠加部分突出显示了在步骤 i 中进行 $A_{j:,\ j:}$ 的舒尔补形成和前向更新时涉及的块。二次矩阵方程的选定求逆/求解采用镜像切片，从矩阵的最后一个块（箭头块的尖端 $A_{t,\ t,}$）开始，向后运行至 $A_{0,\ 0}$（图 3(b)）。深蓝色块表示逆矩阵/解中的精确元素，而深色叠加部分突出显示了步骤 j 中反向选定求逆$X_{i:,\ i:}$ 涉及的块。

图 4 展示了将 RGF 算法扩展到 BTA 矩阵的块推导过程。我们在子图 (a) 中列出了步骤 i 中对每个对角块执行的三级 BLAS 运算。对于 BTA 稀疏模式，这些运算只会产生固定数量的零块运算，即非



注释： $j = i + 1$。

图4：扩展 RGF 算法前向更新的块运算（第 i 步）。（a）包含所有零运算。（b）压缩后，避免先验 零运算。

零块运算，具体包括箭头块 $A_{t,\ j}$ 和$A_{j,\ t}$ 以及尖端块 $A_{t,\ t}$。图4(b) 中，我们给出了扩展 RGF 算法步骤 i 中待执行运算的压缩块表示，避免了 BTA 稀疏模式中特有的显式零计算。二次矩阵方程的前向舒尔补每步涉及 34 次 gemm、1 次 getrf（LU 分解）和 2 次 trsm 块运算。相应的反向选定二次求解每步需要 76 次 gemm 块运算。

B. 二次方程选定解的并行算法

我们在此推导用于求解 $AXA^{\dagger} = B$ 的并行方法，适用于 BT 矩阵和 BTA 矩阵。假设 A 和 B 具有相同的稀疏模式。注意，BT 矩阵可视为满足 a=0 的 BTA 矩阵的特例。

1) 概述： 尽管 RGF 的块方法非常适合通过第 3 级 BLAS 运算利用块内并行度，但该算法本身在相邻块运算之间存在固有的顺序依赖。为了打破块级别的顺序依赖并降低算法深度，我们引入一种置换方案，以增加块运算和工作负载为代价，产生易并行的部分。这些易并行部分分配给 P 个进程（中央处理器或图形处理器）。

我们的方案通过置换方案引入的填充以及需要求解、连接各易并行分区的约化方程组来增加工作量。我们将所有向上（向下）与分区 P_{i-1} （P_{i+1}）相连的 P_i 称为“中间分区”。首尾分区 P_0 和 P_{P-1} 分别仅与 P_1 和 P_{P-2} 相连。由于中间分区与首尾分区的连接方式不同，因此额外工作量也不同。因此，分布式算法需要进行负载均衡，可通过调整中间分区中的对角块数量来实现。

2) 置换： 对 A 和 B 矩阵进行置换，是为了在算法层面重新排序块访问模式，从而隐式地打破 RGF 的块级顺序依赖；这一过程无需内存复制。置换矩阵 P用于实现这一目的。在图5(a)所示的具体情形中，它将所考虑的 BTA 矩阵进行区域分解或剖分，分为由虚线界定并分别以蓝色、橙色和绿色标示的三个分区。连接 A 两个分区的块充当分隔块。这些块以紫色表示，并将例如a 和b 这两个块 A连接起来。

图5(b)直观展示了重新排序后的矩阵 PAP^T。中间分区中的置换所引起的分解填充以阴影线标出。最后，在图5(c)中，对置换后的矩阵重新排序，以分离出已显现的易并行部分（左上象限）、顺序约化系统（右下象限）以及分解填充（右上和左下象限）。在二次

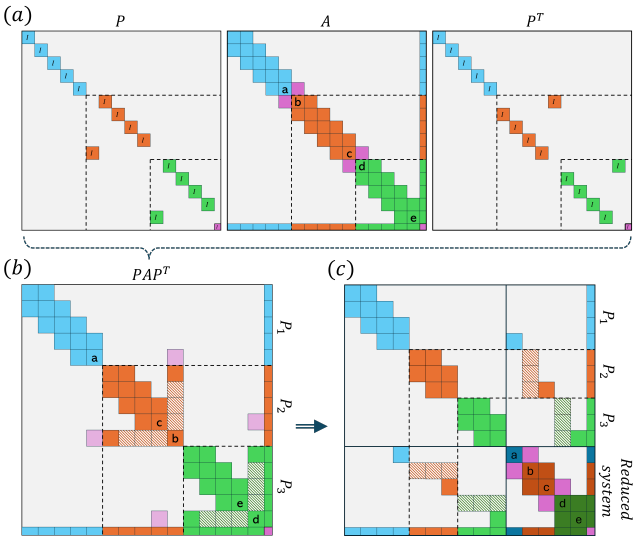


图5：（a）置换 PAP^T的图形表示，其中矩阵 A分布在三个进程上。所有分区均以不同颜色表示。（b）根据（a）中的运算，对分区矩阵 A 进行置换和重新排序后的结果。阴影线表示舒尔补运算期间置换引起的填充。

矩阵方程中，分区和隐式置换同时应用于系统矩阵 A 和右端矩阵 B。

3）直观推导：与 RGF 算法的 BTA 扩展类似，完整的算法推导包含许多相互依赖的块运算。这里我们通过直观的方式展示本工作所介绍的分布式选定求逆/选定解算法的推导过程。我们的方法能够计算舒尔补和 $AXA^{\dagger} = B$ 的选定解，同时还能对 $AX = I$进行选定求逆。在图 6 中，将上一节介绍的置换应用于二次矩阵方程的舒尔补。在置换的基础上，高亮显示了算法的某一步 i 中间分块所需执行的块运算。这些运算可以计算矩阵 B的舒尔补：$S_{B_{j:}\ j:} = S_{A_{i:i:}} S_{A_{i:i:}} S^{\dagger}_{A_{i:}}$

按照第 III-B2 节介绍的切片与重排序方式，我们在图 6 中识别出三种运算：(i) 更新下一个对角块 $B_{j,\ j}$；(ii) 更新约化系统内部对应的对角块 $B_{p,\ p}$；(iii) 更新箭头块的尖端 $B_{t,\ t}$，该部分同样属于约化系统。除中间进程外，所有进程的访问模式都相同，因为中间进程分区边界处的块（即连接其上/下邻居的块）也需要更新。沿用图 4 中的思路，我们通过避免计算先验零元素来压缩块运算的表示。这种压缩块表示如图 7 所示。以较深色调突出显示的更新块，表示当前进程对约化系统的局部贡献。

图 6：置换和重新排序后的二次矩阵方程 $AXA^{\dagger} = B$ 。在步骤 i 中，舒尔补涉及浅橙色、浅棕色和浅紫色区域中的切片块乘法。灰色箭头表示前向更新，为下一步的补运算做准备。

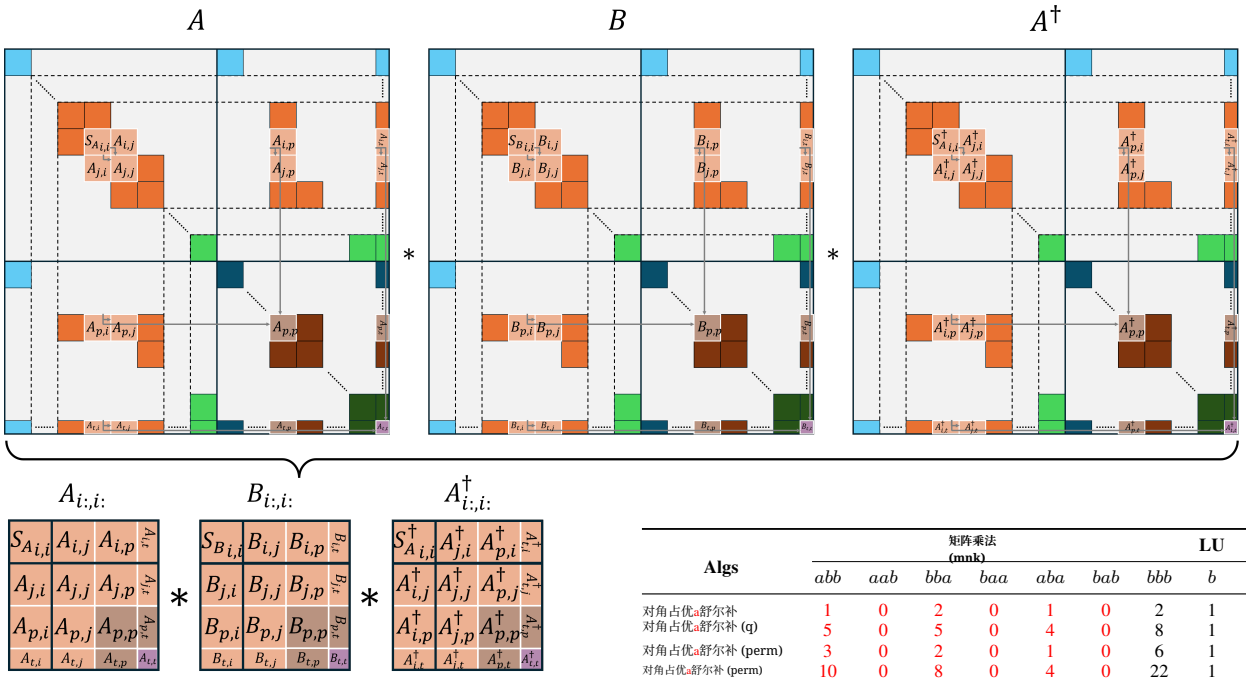


图 7：从图 6 中提取的压缩结构，突出显示了每次迭代 i 中执行的舒尔补运算和前向更新。

在所有进程上完成易于并行的舒尔补计算后，可以在本地收集约化系统。随后，可以使用第 III-A 节介绍的 BTA RGF 算法求解，也可以通过递归调用该算法的分布式版本来求解。这一运算最终构成了该算法的串行瓶颈。最后，恢复切片方式和块依赖关系，并执行易于并行的选定求逆。每个进程都会生成与其系统矩阵分区的非零模式相对应的逆矩阵/解的块。总体而言，这就完成了对原始线性系统的完整分布式二次选定解。

C. 分析

我们在表 V 中给出了分布式内存算法的渐近计算复杂度和内存占用。我们分别考察易于并行处理的部分的贡献，以及由置换引起的额外工作。分布式算法的计算效率在 $\frac{n}{P} \gg P$ 时达到最优，即每个分区的大小大于进程数。

—

我们在表 IV 中列出了本文介绍的所有算法每一步（对角块数量）所需的块运算次数。将 RGF 算法扩展到 BTA 矩阵需要付出额外代价：每一步 i 都要计算多个形状为 abb 的矩阵乘积，

Algs	矩阵乘法 (mnk)							LU	TRSM
	abb	*aab*	*bba*	*baa*	*aba*	*bab*	*bbb*	*b*	*b*
对角占优a舒尔补	1	0	2	0	1	0	2	1	2
对角占优a舒尔补 (q)	5	0	5	0	4	0	8	1	2
对角占优a舒尔补 (perm)	3	0	2	0	1	0	6	1	2
对角占优a舒尔补 (perm) (q)	10	0	8	0	4	0	22	1	2
ddbtasci	2	1	2	1	0	3	7	0	0
对角占优块三对角a科学（q）	6	3	9	4	0	13	34	0	0
ddbta科学（置换）	3	1	5	1	0	5	14	0	0
ddbtasci（置换）（q）	10	3	14	5	0	22	72	0	0

表 IV：本工作介绍的不同算法所需的块运算次数。关键词 *perm*表示在分布式版本中使用了置换算法，而关键词 *q*表示二次矩阵方程的选定解。以红色突出显示的运算仅在算法的 BTA 版本中需要。

aab, bba, baa, aba和 bab。实际上，对于许多科学应用，箭头块 a 相较于对角块大小 b较小，因此计算开销主要来自形状为 bbb的矩阵乘积。

稀疏性模式	分布式	
	计算	内存
BT	$O(\frac{nb^3}{P} + Pb^3)$	$O(\quad + \quad)$
BTA	$O(\quad + \quad)$	$O(\quad + \quad)$

易并行。　约化系统工作。

表 V：所提出的并行算法的渐近复杂度和内存占用分析。

分布式算法依赖于将每个易并行分区中的更新后的块聚合到一个必须按顺序处理的约化系统中。每个中间进程都会向所有对应进程发送 2 个大小为 b^2 的对角块，以及 2 个大小为 b^2 的上/下对角块

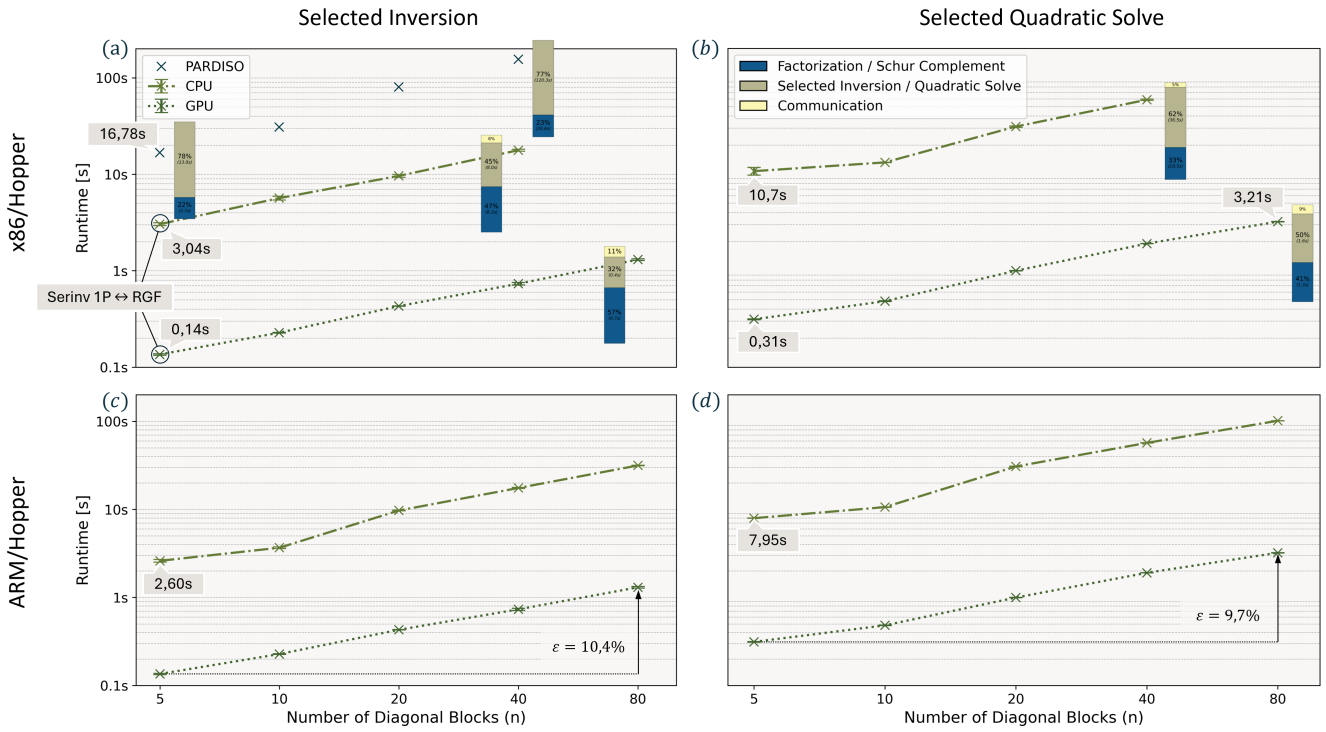


图 8：我们的方法在量子输运仿真的 NR-FET 晶体管真实数据集上的弱扩展性能，随对角块数量 n的变化而变化。对角块大小 b 设为 3408。（a）和（c）选定求逆。（b）和（d）融合选定求逆以及二次问题的选定解。在可能的情况下，提供了与其最佳多线程配置（$OMP = 64$）下 PARDISO 的比较。评估分别在（a-b）x86/Hopper 和（c-d）ARM/Hopper 架构上进行。堆叠条形图显示了因子分解、选定求逆 / 二次求解以及（如适用）通信所占的运行时间百分比。

，还会通过*AllGather* 集合通信发送 4 个大小为 ab 的下（上）箭头块。约化系统的尖端包含更新后的箭头块，它由来自所有进程的数据构建，并通过大小为 a^2 的 *AllReduce* 集合通信进行聚合。

IV. 评估

A. 方法

我们通过两种方式对我们的方法的性能进行了基准测试。首先，我们使用一个标记为 NR_{3408}的真实量子输运数据集作为基准。该数据集对应于一个 NR-FET 晶体管，类似于图 1 中所示的晶体管。器件沟道的长度随进程数量成比例增加，以测试我们的分布式算法。其次，我们基于一个标记为 SD、涵盖各种可能沟道长度的合成数据集，确定了我们方法的并行效率。数据集总结见表 VI。

参考	n	b	N
NR_{3408}	$\{5, 10, 20, 40, 80\}$	3408	$17k - 272k$
SD	$\{32, 64, 128, 256, 512, 1024\}$	1024	$32k - 1M$

表 VI：用于评估我们分布式算法性能的数据集。其中，n 是主对角线上的对角块数量，b 是对角块大小，N 是矩阵总维数。

Name	架构	描述	内存
FAU	x86 (Sapphire Rapids)	2 x 英特尔至强白金 8470（2 × 52 核 @ 2.0 GHz）	2TB DDR5
ALPS	ARM（Grace）Hopper	NVIDIA GH200（72 核中央处理器 + 1 图形处理器）	128GB LPDDR5X 96GB HBM3

表 VII：用于评估我们算法实现性能的硬件总结。

在这两项研究中，我们测量了选定求逆和选定二次方程求解方法的求解时间，并在可能的情况下将结果与 PAR- DISO 进行了比较。所有情况下均报告至少 10 次运行的平均值以及 95% 置信区间。我们采用了两种不同的硬件架构：x86 CPU 搭配 Hopper GPU，以及 ARM CPU 搭配 Hopper GPU（GH200 超级芯片）；前者由埃尔朗根国家高性能计算中心（NHR@FAU）提供，后者由瑞士国家超级计算中心（CSCS）提供。其特性总结于表 VII。我们使用 MPI

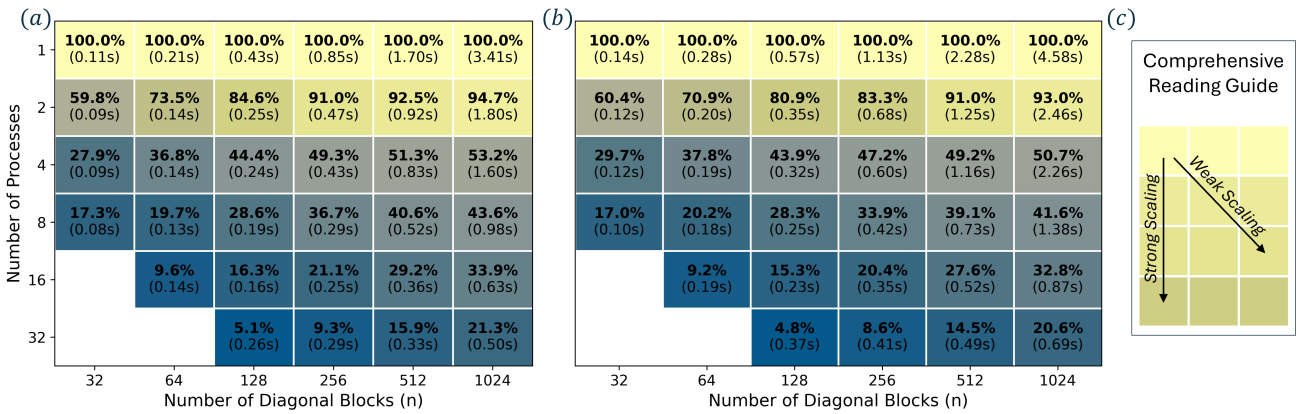


图 9：在合成数据集SD上，评估我们的 GPU 实现在 GH200 超级芯片上的并行效率。(a) 选定求逆。(b) 融合选定求逆与选定二次求解。问题规模包括 $n = [32, 64, 128, 256, 512, 1024]$ 个对角块，每个块的大小均为 $b = 1024$。(c) 用于解读 (a) 和 (b) 的阅读指南。

（中央处理器）和 NCCL（图形处理器）作为分布式代码的通信库。

CPU 评估采用了最有利的多线程设置：我们的方法在 x86 系统上使用 32 个 OpenMP (OMP) 线程，在 ARM 上使用 64 个 OMP 线程；PARDISO 在 x86 上使用 64 个 OMP 线程。对于 PARDISO，使用 METIS 软件包中的嵌套剖分对输入矩阵进行重排序 [44]，，该方法被认为最适用于所考察的稀疏模式。此外，还分配了最大 2 TB 的内存池。

B. NR-FET 器件的弱扩展

图 8 展示了我们针对 NR_{3408} 数据集所开发的分布式算法的弱扩展评估：子图 (a) 和 (c) 中为选定求逆，子图 (b) 和 (d) 中为融合选定逆和选定二次方程求解。报告了中央处理器（x86 和 ARM）以及 Hopper GPU 上的求解时间。我们还给出了 PARDISO 选定逆求解器在 x86 CPU 上的测量结果，这是该求解器唯一可用的配置，并采用性能最佳的线程数（64）。由于线程数并未根据矩阵大小进行调整，因此图 8 中的 PARDISO 耗时并不代表弱扩展实验。

1) 选定求逆：对于具有 $n = 5$ 个块的NR_{3408} 器件，其最短沟道长度下，PARDISO 完成选定求逆需要 16.78 秒，而我们的方法在一台 x86（ARM）中央处理器上分别耗时 3.04（2.6）秒，在单个节点上分别实现了 $5.5\times$（$6.45\times$）的加速。图形处理器实现相同计算仅需 0.14 秒，即比 PARDISO 快 $120\times$，比其 x86 CPU 版本快 $22.5\times$。

当将对角块数量增加到 $n = 40$、节点数量增加到 8 时，我们的方法在 x86（ARM）CPU 上用 17.76 秒（17.72 秒）完成选定求逆，耗时与 PARDISO 处理最小器件所需的时间大致相同。我们的实现将 47% 的时间用于因子分解，45% 用于求逆，8% 用于通信。

使用 PARDISO 对同一器件进行选定求逆，在 1 个节点上需要 156 秒。我们的 GPU 代码在 16 个节点上用 1.3 秒完成所研究最大器件（$n = 80$）的求逆，其中 57% 的时间用于因子分解，32% 用于选定求逆，11% 用于通信。从 $n = 5$（单个 CPU 或 GPU）扩展到 $n = 80$（16 个 CPU 或 GPU）时，我们的方法在 ARM CPU 和 Hopper GPU 上分别实现了 $\eta = 8.26\%$ 和 $\eta = 10.4\%$ 的并行效率。

2) 选定二次方程求解：对于最短器件（$n = 5$），与选定求逆融合的选择性二次方程求解，在单个 x86（ARM）中央处理器上耗时 10.7 秒（7.95 秒），在单个图形处理器上耗时 0.31 秒。因此，这类能够考虑散射的模拟，运行速度比 PARDISO 及其选定逆求解器快 1.57 至 $2.11\times$ 倍（中央处理器）和 $54.1\times$ 倍（图形处理器）。使用我们的方法，在 $n = 80$（16 个图形处理器）上，选定二次方程求解耗时 3.21 秒；对于长度短 $16\times$ 的器件，其速度比 PARDISO 的选定求逆快 $5.22\times$ 倍。

C. 并行效率矩阵

图 9 展示了我们的方法在 ALPS 超级计算机的 NVIDIA GH200 超级芯片上运行的图形处理器实现的并行效率矩阵：(a) 对数据集 SD 进行选定求逆；(b) 对二次矩阵方程进行选定求逆与选定解融合计算。保持对角块大小不变（$b = 1024$），逐步增加对角块数量（横轴）和图形处理器数量（纵轴）。子图 (c) 提供了性能矩阵的解读指南。

选定求逆的弱扩展从单个图形处理器上的 $n = 32$（$0.11s$）扩展到 32 个 GPU 上的 $n = 1024$（$0.5s$），并行效率达到 $\eta = 21.3\%$。针对二次问题的融合选定求逆和选定解进行的相同实验，并行效率达到 $\eta = 20.6\%$。这一性能展示了将现有的 BT 矩阵选定求逆分布式方法成功扩展到二次矩阵方程选定解的成果。

V. 讨论

我们介绍了分布式内存算法，用于求解基于非平衡格林函数（NEGF）形式的纳米器件量子输运模拟中产生的二次矩阵方程的选定求逆和选定解。在参考文献 [24] 的基础上，并扩展成熟的递归格林函数算法，我们的方法能够处理块三对角矩阵以及带箭头块的块三对角矩阵。

我们的实现能够高效利用中央处理器和图形处理器架构上的分布式内存并行性。在真实的 NR-FET 晶体管数据集上，与用于选定求逆的最先进稀疏直接求解器 PARDISO 相比，性能最高可提升 120× 倍。此外，这些实现还支持纳入散射效应；散射效应需要对二次问题进行选定求解，而即便如此，与将 PARDISO 的选定求逆求解器用于尺寸16× 较小的器件相比，仍然 5.2× 更快。针对最多 32 个 GPU 的弱扩展性和强扩展性分析表明，并行效率超过 20%，证实我们的算法可用于实际器件模拟，其中求解时间和内存消耗是关键因素。因此，所提出的方法显著拓展了基于非平衡格林函数的量子输运研究的适用范围，使其能够处理超出传统求解器能力范围的器件尺寸。

致谢

本工作得到瑞士国家科学基金会（SNSF）的资助，资助项目包括编号为 n° 209358 （QuaTrEx）和编号为 n° 200021 （NumESC）的项目；同时还得到瑞士先进科学计算平台（BoostQT）的资助。我们感谢 CSCS 项目 sm96 和 lp16 提供的科学支持和高性能计算资源，也感谢埃尔朗根国家高性能计算中心（NHR@FAU）在 NHR 项目 80227 下提供的支持；该中心隶属于埃尔朗根-纽伦堡弗里德里希-亚历山大大学（FAU）。

参考文献

[1] A. Agrawal, W. Chakraborty, W. Li, H. Ryu, B. Markman, S. H. Hoon, R. K. Paul, C. Y. Huang, S. M. Choi, K. Rho, A. Shu, R. Iglesias, P. Wallace, S. Ghosh, K. L. Cheong, J. L. Hockel, R. Thorman, L. Baumgartel, L. Shoer, V. Mishra, S. Berrada, A. Ashita, C. Weber, B. Obradovic, A. A. Oni, Z. Brooks, N. Franco, J. Kavalieros, and G. Dewey, “硅纳米带FET CMOS，栅长为 6 nm”，载于*2024 IEEE*国际电子器件会议*(IEDM)*，2024 年 12 月，第 1–4 页，ISSN：2156-017X。 [在线]。可访问：https://ieeexplore.ieee.org/document/10873367

[2] Z. Zhang, Z. Lin, P.-Y. Liao, V. Askarpour, H. Dou, Z. Shang, A. Charnas, M. Si, S. Alajlouni, A. Shakouri, H. Wang, M. Lundstrom, J. Maassen, and P. D. Ye, “一种漏极电流接近 20 mA/m 的全环绕栅氧化铟纳米带FET”，*IEEE*电子器件快报，第43卷，第11期，第1905–1908页，2022年11月。 [在线]。可访问：https://ieeexplore.ieee.org/abstract/document/9903594

[3] C. Y. Huang, G. Dewey, E. Mannebach, A. Phan, P. Morrow, W. Rachmady, I. C. Tung, N. Thomas, U. Alaan, R. Paul, N. Kabir, B. Krist, A. Oni, M. Mehta, M. Harper, P. Nguyen, R. Keech, S. Vishwanath, K. L. Cheong, J. S. Kang, A. Lilak, M. Metz, S. Clendenning, B. Turkot, R. Schenker, H. J. Yoo, M. Radosavljevic, and J. Kavalieros, “用于延续摩尔定律缩放的3-D自对准堆叠式NMOS-on-PMOS纳米带晶体管”，载于*2020 IEEE*国际电子器件会议*(IEDM)*，2020年12月，第20.6.1–20.6.4页，ISSN：2156-017X。 [在线]。可访问：https://ieeexplore.ieee.org/document/9372066

[4] N. Loubet, T. Hook, P. Montanini, C.-W. Yeung, S. Kanakasabapathy, M. Guillom, T. Yamashita, J. Zhang, X. Miao, J. Wang, A. Young, R. Chao, M. Kang, Z. Liu, S. Fan, B. Hamieh, S. Sieg, Y. Mignot, W. Xu, S.-C. Seo, J. Yoo, S. Mochizuki, M. Sankarapandian, O. Kwon, A. Carr, A. Greene, Y. Park, J. Frougier, R. Galatage, R. Bao, J. Shearer, R. Conti, H. Song, D. Lee, D. Kong, Y. Xu, A. Arceo, Z. Bi, P. Xu, R. Muthinti, J. Li, R. Wong, D. Brown, P. Oldiges, R. Robison, J. Arnold, N. Felix, S. Skordas, J. Gaudiello, T. Standaert, H. Jagannathan, D. Corliss, M.-H. Na, A. Knorr, T. Wu, D. Gupta, S. Lian, R. Divakaruni, T. Gow, C. Labelle, S. Lee, V. Paruchuri, H. Bu, 和 M. Khare，《堆叠纳米片环绕栅晶体管：实现超越鳍式场效应晶体管的缩放》，载于2017年超大规模集成电路技术研讨会，2017年6月，第T230–T231页，iSSN：2158-9682。 [在线]。获取地址：https://ieeexplore.ieee.org/document/7998183

[5] S. Mukesh 和 J. Zhang，《环绕栅纳米片场效应晶体管工艺机遇综述》，《电子学》，第11卷，第21期，第3589页，2022年1月，期号：21，出版方：多学科数字出版研究所。 [在线]。获取地址：https://www.mdpi.com/2079- 9292/11/21/3589

[6] V. Maillou、L. Gaedke-Merzhaeuser、A. N. Ziogas、O. Schenk 和 M. Luisier，《Serinv：用于箭头块矩阵块三对角选定求逆的可扩展库》，2025年3月，arXiv:2503.17528 [cs]。 [在线]。获取地址：http://arxiv.org/abs/2503.17528 。

[7] O. Schenk、K. Gartner、W. Fichtner 和 A. Stricker，《PARDISO：半导体器件仿真中的高性能串行和并行稀疏线性求解器》，《未来一代计算机系统》，第18卷，第1期，第69–78页，2001年9月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S0167739X00000765

[8] O. Schenk 和 K. Gartner，《使用PARDISO求解非对称稀疏线性方程组》，《未来一代计算机系统》，第20卷，第3期，第475–487页，2004年4月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S0167739X03001882

[9] P. R. Amestoy、I. S. Duff、J.-Y. L’Excellent 和 J. Koster，《MUMPS：通用分布式内存稀疏求解器》，载于《应用并行计算：工业界与学术界高性能计算的新范式》，T. Sørevik、F. Manne、A. H. Gebremedhin 和 R. Moe 编。柏林，海德堡：施普林格，2001年，第121–130页。[10] P. R. Amestoy、I. S. Duff 和 J. Y. L’Excellent，《多波前并行分布式对称与非对称求解器》，《计算机在应用力学与工程中的方法》，第184卷，第2期，第501–520页，2000年4月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S004578259900242X

[11] P. R. Amestoy、I. S. Duff、J.-Y. L’Excellent 和 J. Koster，《采用分布式动态调度的全异步多波前求解器》，《SIAM矩阵分析与应用杂志》，第23卷，第1期，第15–41页，2001年1月，出版方：工业与应用数学学会。 [在线]。获取地址：https://epubs.siam.org/doi/abs/10.1137/S0895479899358194 。

[12] L. S. Blackford、J. Choi、A. Cleary、E. D’Azevedo、J. Demmel、I. Dhillon、J. Dongarra、S. Hammarling、G. Henry、A. Petitet 等，《ScaLAPACK用户指南》。SIAM，1997年。[13] “cuSOLVERMp：用于分布式稠密线性代数的高性能CUDA库 — cuSOLVERMp。” [在线]。获取地址：https://docs.nvidia.com/cuda/cusolvermp/index.html

[14] A. Svizhenko、M. P. Anantram、T. R. Govindan、B. Biegel 和 R. Venugopal，《纳米晶体管的二维量子力学建模》，《应用物理学杂志》，第91卷，第4期，第2343–2354，2002 [ ] 10 1063 1 1432117 2008年2月。在线。获取地址：https://doi.org/。/。

[15] D. E. Petersen、H. H. B. Sørensen、P. C. Hansen、S. Skelboe 和 K. Stokbro，《块三对角矩阵求逆与快速传输计算》，《计算物理学杂志》，第227卷，第6期，第3174–3190页，2008年3月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S0021999107005177

[16] L. Gaedke-Merzhaeuser、E. Krainski、R. Janalik、H. Rue 和 O. Schenk，《用于大规模时空贝叶斯建模的集成嵌套拉普拉斯近似》，2023年3月，arXiv:2303.15254 [cs, math, stat] [ ] 2303 15254 。在线。获取地址：http://arxiv.org/abs/。

[17] M. M. Calderara，《SplitSolve：一种用于第一性原理量子输运模拟的算法》，博士论文，苏黎世联邦理工学院，2016年，接收日期：2017-09-22T13:18:14Z。 [在线]。获取地址：https://www.research-collection.ethz.ch/handle/20.500.11850/125842

[18] L. Lin、C. Yang、J. C. Meza、J. Lu、L. Ying 和 W. E，《Selinv——一种用于稀疏对称矩阵选定求逆的算法》，ACM

数学软件汇刊，第37卷，第4期，2011年2月。 [在线]。获取地址：https://doi.org/ 10 1145 1916461 1916464 。/。M. Jacquelin、L. Lin 和 C. Yang, [19]《PSelInv——一种用于选定求逆的分布式内存并行算法：非对称情形》，《并行计算》，第74卷，第84–98页，2018年。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S 0167819117301941 [20] S. Li、S. Ahmed、G. Klimeck 和 E. Darve, 《使用FIND算法计算稀疏矩阵逆矩阵的元素》，《计算物理学杂志》，第227卷，第22期，第9408–9427页，2008年11月。[在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S 0021999108003458 [21] S. Li 和 E. Darve, 《优化FIND算法以计算稀疏矩阵的逆矩阵》，载于2009年第13届国际计算电子学研讨会，2009年5月，第1–4页。 [在线]。获取地址：https://ieeexplore.ieee.org/abstract/document/5091136 [22] ——, 《FIND算法的扩展与优化：计算格林函数和小于格林函数》，《计算物理学杂志》，第231卷，第4期，第1121–1139页，2012年2月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S002199911100338X [23] S. Li、W. Wu 和 E. Darve, 《一种用于与求逆相关的稀疏矩阵计算的快速算法》，《计算物理学杂志》，第242卷，第915–945页，2013年6月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S 0021999113000892 [24] D. E. Petersen、S. Li、K. Stokbro、H. H. B. Sørensen、P. C. Hansen、S. Skelboe 和 E. Darve, 《一种用于并行计算格林函数的混合方法》，《计算物理学杂志》，第228卷，第14期，第5020–5039页，2009年8月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S 0021999109001648 [25] E. Polizzi 和 A. H. Sameh, 《并行混合带状系统求解器：SPIKE算法》，《并行计算》，第32卷，第2期，第177–194页，2006年2月。 [在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S 0167819105001353 [26] S. Cauley、M. Luisier、V. Balakrishnan、G. Klimeck 和 C.-K. Koh, 《用于模拟含散射纳米电子器件的分布式非平衡格林函数算法》，《应用物理学杂志》，第110卷，第4期，第043713页，2011年8月。 [在线]。获取地址：https://doi.org/ 10 1063 1 3624612 。/。L. Spellacy 和 D. Golden, 《复数块三对角矩阵的部分逆》，载于《并行处理与应用数学》，R. Wyrzykowski、J. Dongarra、E. Deelman 和 K. Karczewski 编。尚：施普林格国际出版，2018年，第634–645页。[28] J. Wu 和 C. H. Diaz, 《预测性技术计算机辅助设计在先进技术开发中的作用拓展》，载于2013年国际半导体工艺与器件仿真会议（SISPAD），2013年，第167–171页。[29] U. Kwon、Y. Park、Y.-S. Kim、J. Yoo 和 D. S. Kim, 《技术计算机辅助设计面临的挑战与机遇：寻找适用于亚3纳米缩放的可行器件架构》，载于2021年国际半导体工艺与器件仿真会议（SISPAD），2021年，第5–10页。[30] M. A. Stettler、S. M. Cea、S. Hasan、L. Jiang、P. H. Keys、C. D. Landon、P. Marepalli、D. Pantuso 和 C. E. Weber, 《工业技术计算机辅助设计：从原子到芯片建模》，《IEEE电子器件汇刊》，第68卷，第11期，第5350–5357页，2021年。[31] M. Brandbyge、J.-L. Mozos、P. Ordej´on、J. Taylor 和 K. Stokbro, 《非平衡电子输运的密度泛函方法》，《物理评论B辑》，第65卷，第165401页，2002年3月。 [在线]。获取地址：https://link.aps.org/doi/ 10 1103 65 165401 。/PhysRevB.。M. Luisier 和 A. Schenk, [32]《单一介质和高 κ 栅极堆栈金属氧化物半导体场效应晶体管的二维隧穿效应及其漏电流》，《IEEE电子器件汇刊》，第55卷，1501，2008年6月，第1494页。[33]J. Wang、E. Polizzi、A. Ghosh、S. Datta 和 M. Lundstrom, 《硅纳米线晶体管中表面粗糙度散射的理论研究》，《应用物理快报》，第 87， 4， 043101, 卷，第。页。在线。07 2005 [ ] 10 1063 1 获取地址：https://doi.org/。/。 2001158 [34] T. Frederiksen、M. Paulsson、M. Brandbyge 和 A.- P. Jauho, 《基于第一性原理的非弹性输运理论：方法及其在纳米尺度器件中的应用》，《物理评论B辑》，第75卷，第205413页，2007年5月。[在线]。获取地址：https://link.aps.org/doi/ 10 1103 75 205413 。/PhysRevB.。K. S. Thygesen 和 A. Rubio, [35]《用于纳米尺度接触量子输运的非平衡GW方法》，《化学物理学杂志》，第126卷，第9期，第091101页，2007年3月。 [在线]。获取地址：https://doi.org/10.1063/1.2565690 [36] S. Datta, “纳米尺度器件建模：格林函数方法”，《超晶格与微结构》，第28 卷，第 4 期，页 253–278，2000 [ ] 10 1006 2000 0920 年 10 月。[37]在线。可访问：http://dx.doi.org/ 。 /spmi. 。A. N. Ziogas、T. Ben-Nun、G. I. Fernandez、T. Schneider、M. Luisier 和 T. Hoefler, “面向极端规模从头算耗散量子输运模拟的数据中心方法”，载于国际高性能计算、网络、存储与分析会议论文集，丛书 SC ’19。美国纽约州纽约：美国计算机协会，2019年。 [在线]。可访问：https://doi.org/ 10 1145 3295500 3357156 。 / 。L. Deuschle、J. Cao、A. N. Ziogas、A. Winka、A. Maeder、N. Vetsch 和 M. Luisier, [38]“通过非平衡格林函数和GW近似处理器件仿真中的电子-电子相互作用”，物理评论*B*辑，第 111 卷，页 195421，2025 年 5 月。 [在线]。可访问：https://link.aps.org/doi/ 10 1103 111 195421 。 /PhysRevB. 。R. Lake、G. Klimeck、R. C. Bowen 和 D. Jovanovic, [39]“层状半导体器件中量子电子输运的单能带与多能带建模”，《应用物理学杂志》，第 81 卷，第 12 期，页 7845–7869，1997年 06 月。 [在线]。可访问：https://doi.org/ 10 1063 1 365394 。 / 。L. Deuschle、A. Maeder、V. Maillou、N. Vetsch、A. Winka、J. Cao、A. N. Ziogas 和 M. Luisier, [40]“迈向GW近似下纳米电子器件的百亿亿次级模拟”，载于*SC24*：国际高性能计算、网络、存储与分析会议，2024年，页 1–16。[41] D. G. Feingold 和 R. S. Varga, “块对角占优矩阵及盖尔圆定理的推广。”1962年。[42] F. J. K Takahashi、Chin M, “稀疏母线阻抗矩阵的构建及其在短路分析中的应用”，第八届*PICA*会议论文集，1973年。[43] A. Haidar、H. Ltaief、A. YarKhan 和 J. Dongarra, “多核架构上稠密线性代数动态调度分块算法的分析”，《并发与计算：实践与经验》，第 24 卷，第 3 期，页 305–321，2012年。 [在线]。可访问：https://onlinelibrary.wiley.com/doi/abs/ 10 1002 1829 。 /cpe. G. Karypis 和 V. Kumar, [44]*METIS*：用于非结构化图划分、网格划分以及稀疏矩阵填充减少排序计算的软件包，1998年 9 月。[45] G. KARYPIS, “Metis：非结构化图划分与稀疏矩阵排序系统”，技术报告，1997年。 [在线]。可访问：https://cir.nii.ac.jp/crid/ 1574231874810284416 [46] C. C. Ashcraft、R. G. Grimes、J. G. Lewis、B. W. Peyton、H. D. Simon 和 P. E. Bjørstad, “向量超级计算机上大型线性系统稀疏矩阵方法的进展”，《国际超级计算应用杂志》，第 1 卷，第 4 期，页 10–30，1987年。 [在线]。可访问：https://doi.org/10.1177/109434208700100403