中山大学计算机学院本科生实验报告

一、 课程名称：超级计算机组成原理　　任课教师：吴迪

年级	2019	专业（方向）	计科（超级计算）
学号	18324034	姓名	林天皓
开始日期	2021.6.10	完成日期	2021.6.25

二、实验题目

阅读国际会议 IEEE SC 、 IEEE Cluster 或 ACM PpoPP 等 近三年发表的一篇正式科研论文（正式论文一般是英文双栏，大于 10 页），并撰写阅读报告（约 8 页左右，若图片较多，可增加页数）:

选择论文为 SC 19

Full-State　　antum Circuit Simulationby Using Data Compression

Full-state quantum circuit simulation by using data compression (acm.org)

三、相关背景

量子计算是一种全新的计算模式，它以量子位为信息单元，遵循量子力学的规律，通过调控量子来完成密码学中的相关计算，其中还可以应用于多体系统量子力学模拟和量子机器学习(等其他领域的诸多复杂算法) 。但是现在量子计算机的实现非常困难，目前仍然没有实用化的量子计算机可以用来求解大规模的科学计算问题，因此使用电子计算机来模拟量子计算系统成为最常用的研究手段，量子线路模拟器应运而生，用来测试量子领域复杂计算问题的模拟，如 projectQ,QuEST 等。这些模拟器作为重要的研究工具受到了量子计算研究者的高度重视，但是目前开发出的量子线路模拟器往往会消耗更多的存储资源带来更多的计算成本。

下面介绍量子模拟的相关基本概念

在量子系统中量子状态不仅可以是二进制 1 或 0，而且可以是同时持有"和#的多种组合，从而形成一个量子位,如下列公式所示，一个量子位的状态可以用一个二维的状态向量来表示，其中 a 和 b 都是复数，多个量子位的表示同理，只要增加状态向量的维数即可。

$$|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle$$

$$|\varphi\rangle = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$$

图 1-2 位量子状态表示

如果一个量子系统有 n 位，那么他的状态由 2^N 个数表示

$$|\psi\rangle = a_{0\ldots00}|0\ldots00\rangle + a_{0\ldots01}|0\ldots01\rangle + \ldots + a_{1\ldots11}|1\ldots11\rangle$$

图 2-n 位量子状态表示

$$\sum_i |a_i|^2 = 1.$$

图 3-所有 a 的平方和为 1

量子门是在一组量子位之间进行操作的基本量子线路，是量子线路的基础。在进行量子线路模拟时，多个量子门的运算是串行完成的，每个量子门都会对全部的状态向量进行一次遍历，完成状态向量的更新计算。

$$A = I^{\otimes n-k-1} \otimes U \otimes I^{\otimes k},$$

图 4-量子电路描述

例如上述公式可以描述一个量子电路，其中 I 和 U 是一种量子门电路。

在量子模拟器中，模拟器记录了被模拟量子系统的状态，以及各个进程的相关信息。它用一个复数数组来存放全部状态向量，每个状态向量由 2 个 double 类型的浮点数构成，分别表示状态向量的实部和虚部，模拟器进行计算的基本数据单元。状态向量的总数与被模拟的量子位数有关，假设对拥有 N 个量子位的量子系统进行模拟,就需要 2^N 个状态向量来进行表示.例如当 N=30 个状态向量共需要 16G 内存空间。被模拟的量子系统每增加 1 个量子位状态向量需要的 存储空间就会随之翻倍可见量子电路模拟器的内存消耗很大。

四、问题是什么

量子电路的经典模拟对于量子计算机的研发至关重要，保证量子计算的算法的正确性和行为费非常重要，通过这些量子仿真的方式，仿真允许研究人员和开发人员评估新量子算法的复杂性并验证目前所设计的量子算法。在建立量子计算机的过程中，如 IBM 的 50 Qubit 量子计算机和谷歌的 72 个 Qubit 量子计算机将需要通过量子电路模拟器验证硬件的正确性。

在量子模拟器中，模拟器记录了被模拟量子系统的状态，以及各个进程的相关信息。它用一个复数数组来存放全部状态向量，每个状态向量由 2 个 double 类型的浮点数构成，分别表示状态向量的实部和虚部，模拟器进行计算的基本数据单元。状态向量的总数与被模拟的量子位数有关，假设对拥有 N 个量子位的量子系统进行模拟,就需要 2^N 个状态向量来进行表示.例如当 N=30 个状态向量共需要 16G 内存空间。被模拟的量子系统每增加 1 个量子位状态向量需要的 存储空间就会随之翻倍可见量子电路模拟器的内存消耗很大。

不幸的是，今天的我们能进行的模拟限制是 47 量子比特，这是因为量子计算模拟所需要的内存是随着量子比特数量的上升是指数级别上升的，使得计算机内存的大小是

制约我们量子计算模拟的重要因素，

Table 1: Examples of supercomputers, their total memory capacity, and the maximum number of qubits they can simulate for arbitrary circuits.

System	Memory (PB)	Max Qubits
Summit	2.8	47
Sierra	1.38	46
Sunway TaihuLight	1.31	46
Theta	0.8	45

图 5-不同量子比特电路模拟理论所需内存大小

使用当今运行内存最大的超级计算机也智能最大模拟具有 47 个量子比特的量子电路，因此我们需要相应的数据压缩方法，推进更多量子比特的量子电路模拟。

五、现有解决方案

现有的量子电路模拟方法有以下几种。

1.薛定谔方法：此策略在内存中维护全部状态的状态向量，并在每次步骤中更新状态向量。由于空间呈指数增长，因此物理内存限制了模拟大小。该模拟方法时间复杂度与量子门电路的数量是多项式相关的，因此该方法能够模拟任意深度电路。这种仿真方法可以使用 500TB 的内存模拟 45 个量子比特电路。同时经过一定的优化，Li【43】等人优化，使得 49 个量子比特的电路也成功模拟。

2.费曼路径方法：该方法根据量子的最终状态计算所有的量子路径是否符合初始状态，因此该方法的时间复杂度较高，随着门电路数量与电路层数的增加为指数级别，因此该方法只能用于层数较低的量子电路模拟

3.张量网络伸缩：这种方法使用张量网络表示量子电路，这种方法的表示占据的空间复杂度为与量子电路树底层的宽度指数相关。常常用来量子比特较多的量子模拟。

六、作者的核心思想、创新点是在哪里

为了模拟具有较高量子位数和深度的通用电路，作者提出的量子电路仿真技术可以通过在运行时通过压缩质量状态幅度来减少存储量子全部状态的内存要求。由于该技

术目的是模拟中间尺度的普通量子电路，因此必须尽可能实现数据压缩比的提高，因为压缩比是对模拟中的量子比特数量的关键。作者方法使用了无损与有损压缩和自适应误差限度，降低了模拟的内存要求。通常有损的压缩算法导致显着高于损耗压缩机的压缩算法，同时在某种程度上引入误差。为了最大限度地减少误差传播和保证高保真仿真结果，作者利用 Zstandard 无损压缩器和根据人为设置的误差来构建模拟框架

自然的，人们可能担心有损压缩带来的误差，作者在仿真输出中的相关误差会与物理机器发生体验的误差非常不同。然而，使用有损压缩以不相关的方式应用，这缩短了压缩时间。作者的方法能够对内存的不同进行计算期望的保真度。作者实现了由 intel 中开发的全态量子电路模拟器的技术 intel-QS 。 Intel-QS 是一种基于 MPI 的分布式高性能 Quantum 电路模拟器，可以在超级计算系统上运行。该方法集成了对量子计算的了解减少量子电路模拟的存储器需要的数据压缩技术，使该技术允许模拟具有相同存储库的较大量子系统。在不同大小的量子电路仿真中提供了一个选项控制误差。

通过使用数据压缩来提出一种新技术，以减少一般量子电路的全状态模拟的内存要求。减少内存要求可以增加全状态模拟中的量子比特的数量。

1.设计一种有损压缩方法，用于优化压缩比和量子模拟的压缩速度。这种有损的压缩技术可以结合几种现有的仿真技术，减少了内存的使用。

2.在 Argonne 国家实验室（ANL）的 Theta 超级计算机上实施了一般量子电路仿真框架。

3.实验结果表明该方法的内存要求，模拟的量子电路从 32EB 降低到 768 TB 内存。基于最先进的仿真技术，结果表明，该技术可以将模拟大小升高 2 至 16 量子比特，平均模拟保真度为 0.976

1. 数据压缩方法

凭借极端缺乏的研究和应用产生的巨大数据，多年来已经开发了各种数据压缩技术。

通常科学研究人员采用两种类型的压缩机：无损压缩器或者误差有界有损压缩器。无损压缩器通常分析可变长度编码算法（例如霍夫曼码和算术编码）和作为 LZ77 / 78 的算术编码。然而，在大多数情况下，无损压缩机作为 GZIP，ZSTD 和 BLOSC 不能有效地进行科学数据，因为浮点值的结束尾数位是随机的，使得很难找到恰好数据流中的图案。一些研究表明，现在低压缩比（大多数情况下大约 2：1），远远不满足现在高性能计算应用程序的需求。根据，有界有损的压缩方法被广泛作为这种大科学数据问题的最佳解决方案，因为它不仅可以显着减少数据大小，而且还可以根据用户的控制压缩的误差。错误的损耗压缩机可能具有不同的设计和实现，因此选择最合适的压缩算法对该研究至关重要。所有现有的错误界限损耗量压缩机都可以分为两种型号：基于数据预测和域变换，如下所述。

1.基于数据预测的压缩模型。该模型尝试尽可能准确地在空间或时间维度中准确地预测每个数据点，并且通过一些编码算法压缩数据大小。典型的实例压缩机是 SZ ，其涉及四个压缩步骤：（1）数据预测，（2）线性缩放量化，（3）熵编码，和（4）无损压缩。误差界限在步骤（2）介绍并控制。其他例子包括 isabela 和 fpzip。

2.基于域变换的压缩模型。该模型需要将所有原始数据值转换为另一种非正交系数域进行去序，然后通过应用一些优化的编码算法作为嵌入式编码来缩小数据大小。一个典型的示例压缩器为 zfp ，它执行在每个 4dblock 中利用三种技术的经典纹理压缩步骤为：（1）指数对齐，（2）（非）正交块变换，和（3）嵌入式编码。仅在步骤（3）时具有压缩损失。

所有现有的最先进的有界有损压缩主要设计或评估为了视觉上的误差，因此在量子计算模拟的背景下，它们被要求要求数据保真度和压缩质量的要求。作者首先测试现有最先进的损坏压缩机对量子电路仿真结果的有效性，然后利用超出了超出了超出了损耗压缩机的相当有效的压缩方法。

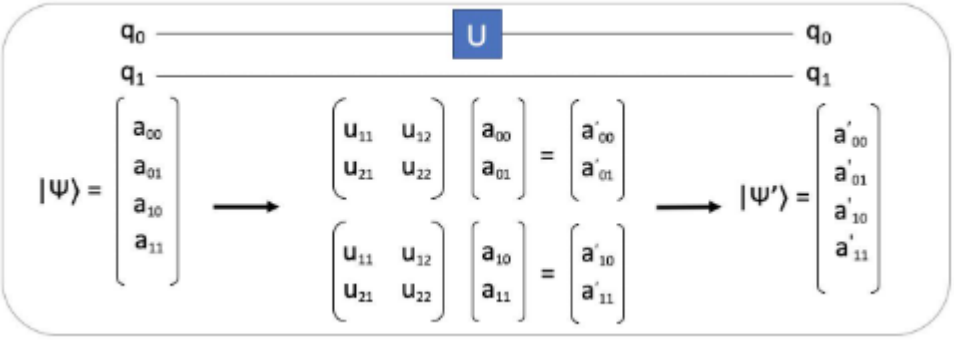

图 6－ 一个量子门电路的压缩过程

2.时间与空间的动态调整

作者目的是使用高保真进行仿真一般量子电路，将压缩技术集成到量子电路调度中，使得可以随着存储器容量增加模拟的精度。该技术允仿真通过应用有损的压缩技术来权衡计算时间和模拟精度。较低的压缩误差界限提供了更高的模拟保真度，但会出现更高的误差，同时具有更高的压缩比，以便可以模拟具有较大数量量子比特的量子电路。

2. 模拟流程

首先将量子的状态矢量平均划分。为了减少内存的使用，进一步划分了每个等级的部分状态矢量为内部区块。每个块都以压缩格式存储。为了完成门操作，需要将矩阵乘法应用于其区块的幅度为 0 和 1 目标时量子位位置，因此在大多数两个块上被解压缩，计算，然后更新状态向量。在所有的子区块被更新后，压缩结果向量并转移到下一个块。一旦更新了所有的区块，就完成了量子们的计算操作。

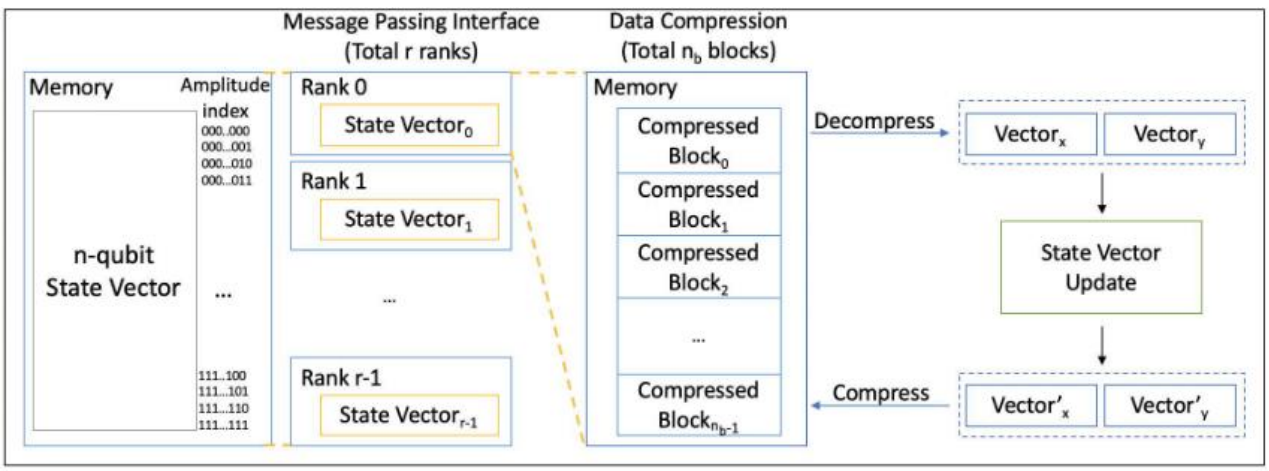

图 7-量子计算流程

计算缓存优化：

对于大部分的量子电路，在门电路中获得的增益额可能是一样的，通过将计算得到的冗余量子计算状态存放在内存中暂时缓存，如果在后续的计算中遇到了相同的量子状态和操作符，可以直接返回系统中已经存在的缓存，从而提高量子电路模拟的速度。

可变压缩精度：

为了保持模拟精度，当压缩率仍然足够高时，在模拟开始时使用无损压缩算法应用于存储器中的所有压缩块。在仿真过程中，量子状态变得更复杂，因此无损压缩比越耐力。当压缩比需要降低以适合内存的大小时，仿真将使用有损压缩来计算状态向量。如果要控制误差，使用误差有界压缩以压缩状态向量。该压缩阶段保证了原始数据和解压缩数据界限的误差范围内。在这项实验中，作者使用了有五个不同的误差界限：1E-5，1E-4，1E-3,1E-2 和 1E-1。每当一个压缩比率不足时，误差界限都会放宽到此级别（更大的误差）。

保真度大小评估：

由于在模拟中使用了有损压缩，因此信息损失了每一个有损压缩，导致量子模拟的整体准确性降低。仿真精度可以通过状态保真度量化，衡量两个量子状态的相似性。保真度是在 0 和 1 之间介于 1 之间的值，具有较高的保险密度在两个状态之间更大的相似性。保真度为 1 将表示两个量子状态是相同的。从仿真中的理想输出状态，在模拟中控制了误差的界限，通过计算每个门电路的误差，累乘计算它们对整体最大损耗的综合影响，将所有门的贡献相结合，将模拟应用程序计算为计算仿真保真度的下限。在作者选择的误差算法的界限为 1E-5,1E-4,1E-3,1E-2 和 1E-1 时候。谢图显示了当应用不同的误差级别时，不同数量的门电路对保真度的影响。

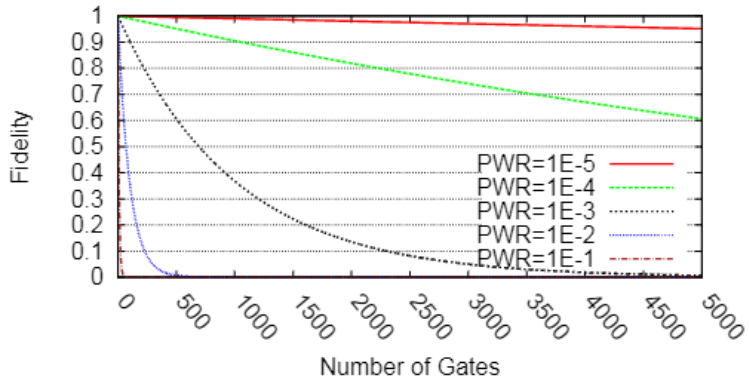

图 8 - 不同数量的门电路对保真度的影响

七、通过什么样的实验进行验证

对于实验的验证，使用 Thetheta 超级计算机进行了模拟。该超级计算机具有 4,392 个节点的，每个节点包含 64 核 Intel Xeon phi processor 7230，具有 16G 字节的高带宽内存存储器（MCDRAM）和 192 GB 的 DDR4 RAM 。 MCDRAM 的带宽为 400gb / s，平均延迟为 154 ns 。在 Theta 上，每个工作都有限制不同数量的节点上限。使用节点的时间限制分别为 128 个节点 3 小时，256 个节点 6 小时和 1,024 个节点 24 小时。在运行超过运行时限的作业中，通过暂时保存结果，然后继续计算。

为了展示针对真实应用的结果，作者选择一些不同数量量子应用作为基准。选择基板标记以具有不同的程序特征，以表明该仿真可以适用于与任意量子电路。

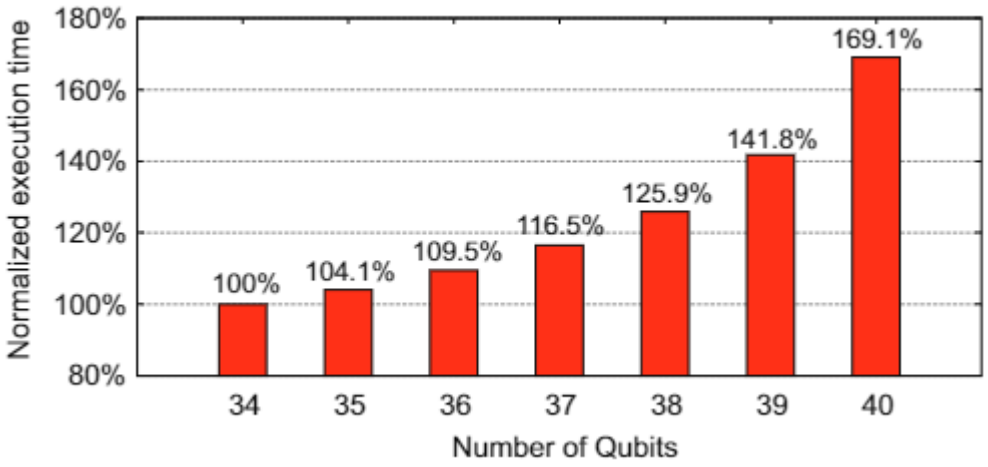

图 9 -该算法时间与传统算法比较

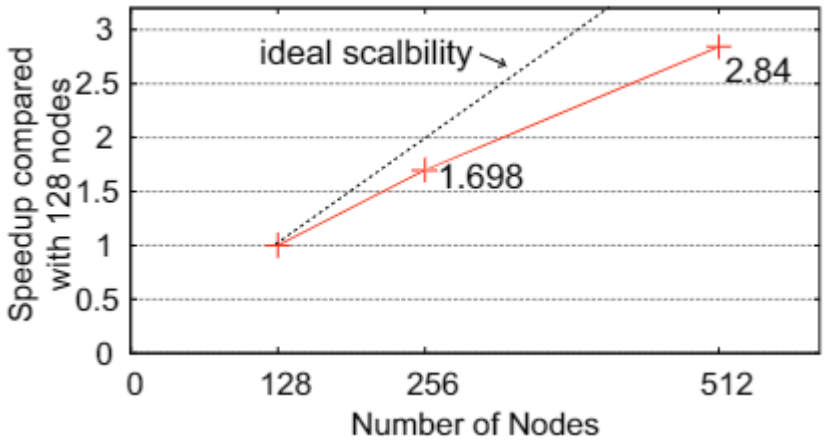

图 10 -该算法并行运算效率

该方法对浅层电路和深层电路应用进行了良好。初始状态为全为 0。用于的基准测试的程序使用包括以下内容。

1.GROVER：GROVER 的搜索算法用于数据库搜索。基准测试使用 GROVER 的搜索算法查找方根。

2.随机电路采样：Google 提出了电路，测试量子规则上的随机电路。由于该算法未对随机量子电路做优化，因此不打算使用深层量子电路运行测试。使用电路深度为 11 个进行实验。

3.qaoa：量子近似优化算法是混合量子分析算法。使用 qaoa 在随机 4 个图上解决最大流问题。

4.QFT：这是用于量子傅里叶逆变形式的量子电路，这是许多量子算法的基本功能（Shor 的算法，相位估计算法等），这是一个深层电路。因此随机将 X 门设置初始状态作为 QFT 在实验中的输入

实验结果

运行基准测试总系统内存大小远低于理论要求，以表明该方法对于内存需求的减少。首先是基准应用程序是 Grover 的搜索算法 47,59 和 61 量子比特。以前，由于内存不足被认为是不可能运行的。使用该方法，用高压缩比压缩，因为这种应用的状态向量的幅

度较为常规，因此该方法可以仅使用 768 TB 的内存成功执行的模拟 61 量子比特的 Grover 的搜索算法，这意味着原本需要 32 EB 的内存的量子电路被完整执行。还使用该技术完成 47 个量子比特的的 Grover 搜索算法仿真，如果不使用压缩方法，需要 2 PB 的内存。接下来，测试谷歌的量子随机电路的模拟，由于该方法在计算中利用均匀的结构，在随机电路上效果不好。因此，测试了 11 个随机电路深度的仿真结果。虽然该技术不是专门为随机化电路设计的，但该方法可以使用 48 TB 内存运行 42 量子比特模拟电路，并使用 192TB 模拟 45 量子比特的电路。下一个基准是 Qaoa。由于 Qaoa 是一个重要的量子应用，因此对 Qaoa 量子电路的模拟压缩内存容量非常关键。因为 Qaoa 对低保真度的要求不高。最后，使用 QFT 电路的结果显示该技术对于模拟深层次电路的技术是有效的，用于使用超过 21 倍的压缩状态，仍然具有 0.962 的保真度。对于这些量子应用，该技术可以模拟高保真度的电路。在的基准中，随机电路使用更多的量子纠缠。因此，当量子纠缠较少时，该技术效果更好。

Benchmark	Grover			Random Circuit Sampling				QAOA			QFT
Number of Qubits (Memory Requirement)	61 (32 EB)	59 (8 EB)	47 (2 PB)	5×9 (512 TB)	6×7 (64 TB)	6×6 (1 TB)	7×5 (512 GB)	45 (512TB)	43 (128 TB)	42 (64 TB)	36 (1 TB)
Number of Gates	314	310	305	227	261	165	208	394	344	336	3258
Number of Nodes	4096	4096	128	1024	128	1	1	1024	256	128	1
Total System Memory (Sys Mem / Req.)	768 TB (0.002%)	768 TB (0.009%)	24 TB (1.17%)	192 TB (37.5%)	24 TB (37.5%)	192 GB (18.75%)	192 GB (37.5%)	192TB (37.5%)	48 TB (37.5%)	24 TB (37.5%)	192 GB (18.75%)
Total Time (Hour)	8.14	3.48	0.49	4.87	8.64	7.96	6.23	13.34	5.83	8.65	78.98
Compression Time	1.87%	4.59%	2.04%	55.79%	40.26%	59.10%	58.57%	50.66%	44.97%	41.02%	57.86%
Decompression Time	1.87%	3.73%	4.08%	31.47%	22.19%	33.78%	30.59%	26.46%	27.64%	25.52%	37.68%
Communication Time	32.7%	20.98%	36.73%	0.12%	0.57%	0.02%	0.03%	3.03%	0.22%	0.23%	2.56%
Computation Time	63.47%	70.70%	57.15%	12.60%	36.97%	7.08%	10.8%	19.84%	27.16%	33.22%	1.9%
Time per Gate (Sec)	93.34	40.49	5.78	64.69	119.22	173.65	107.86	121.91	61.02	92.64	87.27
Simulation Fidelity	0.996	0.996	1	0.987	0.993	0.933	0.985	0.895	0.999	0.999	0.962
Compression Ratio	7.39×10^4	8.26×10^4	1.06×10^4	6.03	9.40	8.16	10.05	5.38	4.85	9.25	21.34

图-运行量子模拟电路实验结果

八、对你的启发

在该论文中，作者通过研究量子计算中状态的可压缩性，以及学习一些经典的科学计算中压缩的算法，通过两者的结合，通过时间换取空间的方式，大大优化了量子计算中的空间复杂度，使得现在的超级计算机器可以运行以前需要千倍内存才能运行的量子电路模拟，为将来的真实量子计算机研发与验证打下良好的基础。在该论文中，作者对

于量子计算具有深入的理解，同时也能和跨学科的“数据压缩算法”很好的结合在一起。除此之外，作者在实验应用中基础扎实，在超级计算的服务器上通过设置检查点，编写具有良好并行度的代码，在 intel 的量子模拟器基础上做改造等操作，使得该算法能够应用在实践中，最后通过大规模的实验验证了该算法的实用性。这启发我们在学习理论知识的同时，要融会贯通，积极实践，才能才能使自己的科学追求学有所用，为国家的现代化建设添砖加瓦。