·“双清论坛”专题：理论化学家视角中的仪器创制·

异构计算量子化学软件的研发恰逢其时

田英齐[1,2,3]　马英晋[1,2]　索兵兵[2,4]　金　钟[1,2*]

（1. 中国科学院计算机网络信息中心，北京 100190；
2. 中国科学院计算科学应用研究中心，北京 100190；
3. 中国科学院大学，北京 100049；4. 西北大学现代物理研究所，西安 710069）

[摘　要] 量子化学计算软件在化学研究中扮演着日益重要的角色。利用前沿的计算机架构开发高性能量子化学软件，也受到软件开发者的广泛关注。近年来，随着 GPU 等加速部件展现出越来越强大的计算能力，异构计算也逐渐成为了量子化学软件的研究热点。本文介绍了异构计算技术的软硬件基础；综述了异构计算技术在计算化学，尤其是量子化学领域最新的研究进展和应用，涉及基于图形处理器（GPU）、协处理器（co-processor）的计算加速；并简要介绍了应用异构计算技术的相关量子化学软件。从这些研究的发展中我们可以看出，异构计算已经成为加速量子化学计算的巨大推动力量。对于诸多著名量子化学软件，其辉煌成就已成为了某种意义上的“历史负担”，而异构计算为量子化学软件跨越式发展提供的机遇使得该应用领域国产软件的研发恰逢其时。

[关键词] 量子化学；异构计算；双电子积分；图形处理器；协处理器

1　引　言

量子化学是以量子力学为手段研究化学问题的理论化学分支学科。该学科以公式推导、编程实现、程序计算为研究手段，通过对分子体系中粒子（包含原子核、电子）行为的描述来研究分子的物理化学性质、探讨化学反应机理。由于量子力学仅能精确计算含极少数目粒子的体系，理论化学家提出了原理模型化的方案如休克尔分子轨道方法等，并在早期的冯·诺伊曼结构（程序存储体系结构）的电子计算机上进行了实现，获得了丰硕的成果。由于量子力学的核心是粒子波动方程的求解（薛定谔方程实为二阶偏微分方程，需要使用计算机来求解），以此为基础的量子化学学科的发展跟同时代计算硬件的发展有着很强的相关性。七八十年代以后，随着计算机能力的快速提高（如至今为止计算机性能的提升仍然符合“摩尔定律”）和计算方法的不断发展（如自洽场方法、密度泛函方法等），化学理论和计算研究得到了迅速发展，量子化学方法也逐步可以处理比较复杂的实际化学问题。1998 年和 2013 年诺贝尔化学奖两次授予理论化学家，代表着整个化学已经开始经历着一场革命性的变化，“量子化学将化学带入了一个实验和理论相互配合来研究分子体系性质的新时代”（1998 年诺贝尔化学奖颁奖通告）；“如今对化学家来说，计算机同试管一样重要”（2013 年诺贝尔化学奖颁奖通告）。

量子化学的成功引起了整个化学学科及其它相关学科的关注和重视，反过来也对量子化学提出了更高的要求。比如化学科学家往往期望量子化学可以描述更加实际的大分子体系，如蛋白质分子、金属螯合体系。这些需求也直接促进了量子化学方法的发展。比如从头算（*ab initio*）量子化学能够精确计算的分子体系，已经由最先的几个原子到数十个原子，发展到当前数百个原子的化学团簇。对于上千原子的体系，基于分子片的方法也能实现精准的从头计算甚至达到线性标度[1]。化学科学家的另一个需求就是算的更准，尤其是针对激发态等含复杂电子相互作用的体系。这类体系往往需要精准的电子

收稿日期：2017-10-10；修回日期：2017-12-20

* 通信作者，Email：zjin@sccas.cn

相关计算，计算标度很高甚至会随体系的电子数目呈指数增长，如完全活性空间自洽场方法（CASSCF）以及基于此方法的多参考（MR）电子相关方法。针对此类体系，量子化学家也发展出了多种可靠的计算方案，如基于图形酉群（GUGA）的组态相互方案[2]，基于密度矩阵重整化群的方案等[3-5]。在方法研究的同时，科学家们也十分重视量子化学软件的开发。目前有诸多流行的量子化学软件，如拥有众多用户的 Gaussian、GAMESS-US、Q-Chem，侧重于多组态/多参考量子化学的 Molcas、Molpro，专注于大规模并行的 NWChem，以及新近研发出的开源量子化学软件 Psi4、完全基于 GPU 的量子化学软件 Tera Chem 等[6]。这些量子化学软件极大的方便了化学研究人员，为他们提供了强有力的研究工具。

量子化学家在不断改进算法的同时，也越来越意识到应该最大限度利用计算机软硬件的进展成果，来取得事半功倍的效率提升[7-8]。当前通用的中央处理器（Central Processing Unit，CPU）的构架不断优化、频率不断提高，早已从单核到发展到多核。值得注意的是，CPU 并行效率往往受制于数据通信速度，会随着 CPU 处理器数量的增加而快速衰减。相对于通用的 CPU 处理器，最初用于计算机图形显示的图形处理器（Graphic Processing Unit，GPU）约 10 年前被引进到高性能计算领域。由于 GPU 使用跟 CPU 不同的构架设计思路，合理使用 GPU 可以进一步提高计算速度。当前与 GPU 通用计算功能相匹配的软件开发环境如 Compute Unified Device Architecture（CUDA）、Open Computing Language（OpenCL）、Open Accelerators（OpenACC）等的出现更是极大地提升了开发效率。此外，INTEL 近年来也发布了新一代的协处理器部件如 Xeon Phi 至强融核处理器，D. E. SHAW 公司发布了 ANTON 系列处理器，Google 发布了新的张量处理器（Tensor Processing Unit，TPU），以及中科寒武纪发布了智能处理器。这些新型的计算处理器也已经或者将有可能用在量子化学方面，也值得量子化学程序开发人员的关注。

本文介绍了异构加速计算在量子化学领域的最新应用和进展。综述的第一部分概要介绍 CPU、GPU、Xeon Phi、TPU 等异构计算平台硬件构架的特点，以及不同硬件平台上适合量子化学方法开发的软件环境。第二部分介绍了异构加速主要涉及的量子化学求解算法及其最新进展。第三部分将会简要介绍常见的几种支持异构计算的量子化学软件。最后总结并展望了异构计算在量子化学软件开发的发展前景并指出支持异构计算将是国产量子化学软件的重要特征。若无特殊说明，文中的加速比较均为 GPU 卡与多核 CPU 而非单个 CPU 核的比较。

2 异构计算软硬件平台的发展

在介绍异构计算前，我们首先简介同构计算（Homogeneous Computing）。顾名思义，同构计算是指在相同硬件构架上进行的计算。传统的同构并行计算设计架构分为共享内存的设计架构和消息传递的设计架构两类，分别以 Open Multi-Processing（OpenMP）和 Message Passing Interface（消息传递接口，MPI）为代表。用于共享内存并行系统的 OpenMP 当前已经成为最为普及的标准，并已经被各种量子化学软件广泛使用。OpenMP 提供了基于 CPU 的多线程并行程序设计的一套指导性注释。这种对于并行描述的高层抽象降低了并行编程的难度和复杂度。这样量子化学软件开发人员可以把更多的精力投入到并行算法本身，而非其具体实现细节。信息传递接口（MPI）是基于消息传递机制的编程接口标准。它定义了一系列的编程接口，并且有着 Intel MPI、Open MPI、MVAPICH 等多种实现。它通过 send、receive 等函数进行线程间的消息传递，在大规模并行应用中很有优势，更适合于集群架构。MPI 在量子化学程序中也有着广泛的应用，如 GAMESS、Molcas、VASP 等都应用了 MPI 技术。

异构计算（Heterogeneous Computing）是指使用不同类型指令集和体系构架的计算单元组成系统的计算方式。常见的计算单元类别包括 CPU、GPU、可编程逻辑装置（FPGA）和专用集成电路（ASIC）等（表 1）。异构计算近几年来得到更多关注，主要是因为通过提升 CPU 时钟频率和内核数量而提高计算能力的传统方式遇到了散热和能耗极限；与此同时，GPU 等专用计算单元虽然工作频率较低，具有更多的内核和并行计算能力，总体性能—芯片面积比和性能—功耗比都很高，却远远没有得到充分利用。当前量子化学软件对于异构平台的利用主要集中在 CPU 和加速部件（GPU、集成众核 MIC 构架）平台。后两者（FPGA 和 ASIC）芯片的功能是固定的，实现的算法直接用门电路实现，软硬件一体化的特点决定了 FPGA 和 ASIC 设计中极端重要的资源利用率特征，但也大幅提高了软硬件开发人员的门槛。目前为止在计算化学领域，只有 D. E. Shaw 在 2007 年 ISCA（计算机体系结构顶级

会议）上展示的自研分子动力学专用机 Anton（最新为 Anton2 且 Anton3 也正在开发中）采用了 ASIC 平台，其分子模拟的计算速度为当时普通计算机群的 1 000—10 000 倍[9]。对于高性能计算机的研发领域，异构与众核架构正成为现代超级计算机的发展潮流。表 2 中我们列出了 TOP500 排名前十的超级计算机的情况[10]。其中星号标出的六台超级计算机均为众核、异构架构。

对于计算化学和量子化学软件开发者来说，主处理器（CPU）＋加速部件（如 GPU、MIC）的平台是最为实用的异构计算开发平台。当前 MIC 和 GPU 的双精度浮点计算性能可以 2—8 倍于 CPU 的计算性能。当前的加速部件大多基于单指令多数据（Single Instruction Multiple Data，SIMD）结构设计，即一个控制器来控制多个处理器，同时对一组数据（又称“数据矢量”）执行相同的操作从而实现空间上的并行性的技术，因而在程序设计上与传统算法有所不同。无论是 NVIDIA 的 Tesla 系列 GPU 还

表 1　流行的异构计算平台（截至 2017 年底）

平台	架构特点	芯片工艺	器件代表	单精度浮点（TFLOPS）	双精度浮点（TFLOPS）	功耗（W）	参考价格（CNY）
CPU	约 70％晶体管用来构建缓存，还有一部分控制单元，计算单元少。运算复杂度高，逻辑复杂。	14 nm	E7-8890v4（24 核 48 线程）	3.2	1.6	165	4 万
			E5-2620 v4（8 核心 16 线程）	0.66	0.33	85	0.3 万
CPU（Sunway）	主从核设计，众核构架；简化的缓存结构，集成内存	28 nm	申威 26010	6.12	3.06	300	—
CPU（MIC）	X86 众核构架，集成内存	14 nm	INTEL Xeon phi 7250	6.092	3.046	215	1.5 万
GPU	提供大量的计算单元（多达几千个计算单元）和大量的高速内存，适合并行处理。	16 nm	Tesla P100	9.2	4.6	250	5.0 万
		12/16 nm	Tesla V100	14.9	7.45	300	＞10 万
		14 nm	Radeon RX Vega 64	12.7	0.8	295	0.5 万
		16 nm	GeForce GTX 1080 Ti	11.3	0.35	250	0.8 万
FPGA	高性能、低功耗，接近底层 IO。通过底层晶体管和连线实现逻辑可编程。	20 nm	Virtex UltraScale 系列	1.5	—	20	—
			Arrial 10 系列	1.5	—	35	—
ASIC	晶体管根据算法定制，不会有冗余，功耗低，计算性能高。	65 nm	寒武纪 DianNao 系列	0.5	—	0.5	—
		16 nm	寒武纪 MLU 系列	—	—	—	—
		—	Google TPU	—	—	75	—
		—	ANTON2	—	12.7	—	—

注：“—”代表没有明确报道。

表 2　TOP 500 前 10 名（2017 年 6 月）

排名	计算系统	性能（PF）	安装地	处理器体系架构
1	太湖之光	93.01	无锡超级计算中心，中国	申威 *
2	天河二号	33.86	广州超级计算中心，中国	Intel Xeon ＋ Xeon Phi*
3	Piz Daint	19.59	瑞士国家超级计算中心	Intel Xeon ＋ NVIDIA Tesla P100*
4	Titan	17.59	橡树岭国家实验室	Opteron ＋ NVIDIA K20x*
5	Sequoia	17.17	劳伦斯利福摩尔国家实验室	Power BQC
6	Cori	14.01	劳伦斯伯克利国家实验室	Intel Xeon Phi*
7	Oakforest-PACS	13.55	JCAHPC，日本	Intel Xeon Phi*
8	京	10.51	RIKEN，日本	SPARC64 VIIIfx
9	Mira	8.59	阿贡国家实验室	Power BQC
10	Trinity	8.10	劳伦斯利福摩尔国家实验室	Intel Xeon

* 异构加速硬件。

是 AMD 的 Vega 系列 GPU，都有配套的专用 SDK（软件开发工具包），如 NVIDIA 的 CUDA（计算统一设备构架），AMD 的 APP（加速并行处理技术）。此外，近几年的 CPU 设计中也开始加入了 SIMD 的结构，尤其是英特尔（INTEL）2013 年发布的集成众核（MIC）架构的至强融核（Xeon Phi）协处理器（co-processor），使得相关研究更为关键。此类协处理器于 2013 年配备在了“天河二号”超级计算机上，使得“天河二号”达到了 33.86 PetaFLOPS 的浮点运算能力，成为了当年性能最强的超级计算机并连续 6 次蝉联 TOP500 榜单的排名第一（2012—2015 年）。与 GPU 类似，如何将计算任务向量化，并使得有效计算比例提高，同时提高内存访问效率，是 MIC 优化的关键。

CUDA 架构是英伟达（Nvidia）公司专门为开发 GPU 加速的高性能异构程序而提供的开发环境，在相关的 GPU 加速的量子化学软件开发中起到了巨大的作用，如第一个完全基于 GPU 的量子化学软件 TeraChem 即是基于 CUDA 进行的开发。CUDA 架构下 GPU 的硬件分为了流处理器（Stream Processors，SP）和流多处理器（Streaming Multiprocessor，SM）两层。一个 SM 中会包含多个 SP，可以比较快速地访问 SM 中的共享内存（Shared Memory）。同时所有的 SM 共享卡上的全局内存（Global Memory），访问全局内存的速度相对较慢。每 32 个 SP 会组合在一起成为一个 warp（GPU 执行程序时的调度单位），一个 warp 在一个时钟周期内进行完全相同的操作，这是一种类似 SIMD 的架构。GPU 加速优化的关键就是提高一个 warp 内有效计算的比例，同时提高共享内存的使用效率，减少访问全局内存。

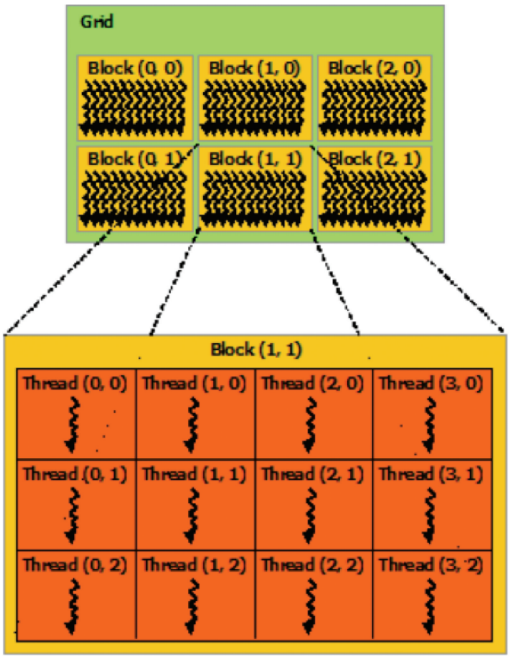

图 1 CUDA 线程架构[11]

图 2 NVIDIA GF100 硬件架构[12]

由苹果公司发起并最终被业界认可的开放式计算语言 OpenCL 旨在通过定义一套机制来实现硬件独立的软件开发环境。利用 OpenCL 可以充分利用设备的并行特性和支持不同级别的并行，并能有效映射到由 CPU，GPU，FPGA，ASIC 和将来出现的设备所组成的同构或异构，单设备或多设备的系统。新近发展的如 FERMIONS＋＋软件、ORCA 的积分加速模块都已经提供对于 OpenCL 的支持。

事实上，无论 CUDA 或者 OpenCL 都需要平台和设备初始化代码以及繁琐冗余的并行计算和数据移动代码。为了解决此类问题，类似 OpenMP 的针对异构平台的指导性注释语言 OpenACC（开放加速器）被开发出来，当前已经提供了对于多种 CPU、GPU、MIC 甚至国产申威系列处理器的支持。这也极大地方便了含量子化学在内的相关计算软件在混合异构平台上的开发。

3　异构加速技术在量子化学上的应用

从头算（*ab initio*）量子化学的基础是自洽场理论，其中单参考量子化学起始于 Hartree-Fock（HF）自洽场计算（由于代码上的相似性，密度泛函理论（Density Functional Theory，DFT）计算普遍调用 HF 自洽场来执行），多参考量子化学起始于多组态自洽场。自洽场计算的速度决定步骤普遍被认为是双电子排斥积分（Electron Repulsion Integrals，ERI）、Fock 矩阵生成以及矩阵的对角化操作。最早提出可使用 GPU 来提高双电子排斥积分计算速度的是日本名古屋大学的 Yasuda 教授。他使用当时最先进的 GeForce 8800 GTX GPU 卡和 Gaussian03 程序，证明 GPU 仅需 CPU 常规计算时间的 1/3 即可计算出同样的双电子积分（受制于当时的硬件架构，该结果是在单精度浮点水平上计算出的）[13]。

同年伊利诺伊大学的 Ufimtsev 和 Martinez（现为斯坦福大学）教授也发表了使用 GPU 加速双电子积分计算的相关研究结果[14]。在文中，他们利用英伟达公司生产的 8800 GTX 显卡以及 CUDA 架构，提出了 3 种计算任务分配方式，即“One Block One Contracted Integral”（1B1CI）、“One Thread One Contracted Integral”（1T1CI）和“One Thread One Primitive Integral”（1T1PI），并比较了这 3 种方式。经过试验测试，1T1CI 和 1T1PI 都取得了不错的效果，而考虑到 direct SCF 方法后续的计算，1T1PI 的任务分配方式更为合适。

在后续工作中，Ufimtsev 和 Martinez 教授研究了如何在 GPU 上实现 direct SCF 方法[15]。Direct SCF 方法的核心是 prescreening、构造 J 矩阵和构造 K 矩阵。他们设计了 presort 的方法，预先对左矢（bra）和右矢（ket）进行排序，使得积分的有效计算更密集，提高了计算的效率，也使 prescreening 起到了最大的效果。在计算积分、构造矩阵的过程中，忽略了部分积分对称性，牺牲了部分计算，使得积分值在共享内存上即可完成矩阵元构造的过程。基于此，他们重新设计了构造 J 矩阵的算法，虽然增加了约 1 倍的计算量，但使得计算在 GPU 的 block 内即可完成。对于构造 K 矩阵，由于此步骤的对称性与 J 矩阵不同，因此设计了不同的构造算法。虽然基于相同的思想，然而 K 矩阵的构造计算量更大。尽管这一算法计算量比传统算法大很多，在实际的运行中却比 CPU 上的传统算法快的多。他们的试验表明，GPU 上的程序比 CPU 模快了几十甚至上百倍。

在以上工作基础之上，Ufimtsev 和 Martinez 完成了能量梯度的计算、结构优化，并实现了第一原理的分子动力学模拟[16]。对天冬氨酸分子溶解于 147 个水分子中的体系（457 个原子，2 014 个基函数），可以达到 0.7 ps/天的模拟速度。由于 GPU 硬件的特点，单精度的计算速度远远超过了双精度的计算速度。因此他们后来又发展了可变精度的计算方法[17]。该工作显示，经过合适的设置，对于积分值比较大的使用双精度而对其他的使用单精度，可以较好地将计算结果误差控制在一定范围内。他们设计的算法在 20 到接近 2 000 个原子的体系上，针对需要的精度，可以使用最少的双精度计算。后续他们又研究了激发态的计算方法等，如在 Configuration Interaction Single（CIS）、Time-Dependent Density Functional Theory（TDDFT）、Multi-Configurational Self-Consistent Field（MCSCF）等方法在 GPU 上的实现[18-21]。上述工作均已集成于 TeraChem 软件并取得了很多成功的应用。如采用 CIS 结合 TDDFT 的方法，光敏黄蛋白在水溶液（487 个水分子）中的激发态达到了上百倍的加速（推测为对比单个 CPU 核）[18]。他们结合 *ab initio* multiple spawning 和态平均的完全活性空间自洽场（State-Averaged Complete Active Space SCF，SA-CASSCF）方法，研究了光化学反应中非常重要的维他命 D_3（Provitamin D_3）的激发态动力学，并证实了实验上观测到光转换过程中的双指数衰减是由于非平衡

动力学过程中环己二烯开环和闭环过程所致[18]。

目前计算高斯基组的双电子积分主要有 3 种递推方法来实现，分别是 McMurchie-Davidson 方法、Obara-Saika 方法和 Rys 多项式方法[13-15, 22, 23]，这几种递推方法得到的积分计算在数学上是等价的，但不同的递推方法实现在硬件上却有不同的效率。Ufimtsev 和 Martinez 之前的工作采用了 McMurchie-Davidson 方法，并开发了程序自动根据递推公式得到积分代码[14, 15]，取得了不错的结果。但是受限于采用的递推方法，该程序仅能实现到 p 函数的积分。而 Asadchev 则采用了 Rys 多项式的方法，使得在 GPU 上实现了最高到 g 函数的积分[22]。在此基础上，他们发展了 CPU/GPU 混用的多线程 Hartree-Fock 方法。他们采取的针对 GPU 的优化策略有：针对不同角动量的积分采取了不同的方法；积分计算完成，立即与密度矩阵进行收缩构造 Fock 矩阵元，使得积分值不会存入内存中；在 host 端构造积分 batch。相比于 CPU、GPU 优化后取得了 17.5 倍的加速[23]。

与 Ufimtsev 和 Asadchev 不同，Miao 和 Merz 采用了 Obara-Saika 方法实现了 GPU 上的双电子积分[24]。Obara-Saika 方法由平行递推和垂直递推两部分构成，平行递推关系更适合与在 GPU 上并行进行，而垂直递推关系在收缩构造矩阵中使用。他们也加入了 presort（预分类）过程，并根据角动量、基组函数的数量和 Schwarz Cutoff 的大小进行了排序，提高 warp 中的有效计算任务比例，取得了不错的成果。根据 O-S 递推方法的特性，他们充分使用了积分的对称性，尽管使用全局内存来构造 Fock 矩阵，仍然取得了不错的效果，相对于单个 CPU 模，最高达到了 143 倍的加速。

2011 年，Karl A. Wilkinson 等人发表了 GAMESS-UK 在双电子积分 GPU 加速方面的研究[25]。他们针对双电子积分的（ss|ss）积分进行了 GPU 加速优化，利用 Rys 多项式的方法，设计了 GPU 上的积分算法。传统的 CPU 程序在更新 Fock 矩阵后会重新生成 bra 和 ket 用于计算，而新的 GPU 算法取消了这一步骤。在预先生成并排序之后，每次更新 Fock 矩阵之后直接生成要进行的积分，并进行下一次的积分计算。测试结果显示基于 GPU 的算法对比 CPU 上成熟的算法取得了 8 倍的加速。

受限于 GPU 内存容量的限制，大基组尤其是涉及高角动量的部分，单纯使用 GPU 存在难以构造的问题。针对此问题，德国马克斯—普朗克研究所的 J. Kalinowski、F. Wennmohs、F. Neese 等建议采用内存占用最少的 McMurchie-Davidson 的方法来计算，基于 OpenCL 实现了任意角动量积分求算中 CPU 和 GPU 的混合使用[26]。慕尼黑大学的 J. Kussmann 和 C. Ochsenfeld 开发了一套深度混合 CPU/GPU 的积分引擎[27]。该引擎的核心技术就是将积分计算的工作负载实时分配到单独的线程（线程类别可以为 CPU 线程、CUDA 线程和 OpenCL 线程），而非静态分配。任务分配的原则是 CPU 线程优先被分配高角动量积分的计算，并由高角动量到低角动量的顺序计算积分；GPU 线程优先被分配低角动量积分的计算，并由低角动量到高角动量的顺序计算积分。

截至目前，量子化学领域绝大多数的异构计算均是基于 NVIDIA 的 GPU，新近才有基于 AMD GPU 的文献报道。在 J. Kussmann 和 C. Ochsenfeld 基于 OpenCL 扩展了他们开发的 FERMIONS++从头算密度泛函计算软件[26]。由于 OpenCL 编程具有跨平台的特征，扩展后的 FERMIONS++可以支持 CPUs、Intel Xeon Phi 和 GPUs。考虑到 OpenCL 内核执行的效率问题，他们在此工作中针对于 GPU 进行了积分内核的预编译和优化，以提高执行效率。测试结果表明，理论双精度浮点性能相同的 NVIDIA 和 AMD 的 GPU 在密度泛函计算的时候仍然具有相当的性能。而值得注意的是，相比于 McMurchie-Davidson 算法，AMD 的 GPU 可能更加适合 Rys-quadrature 的积分求解方案。此外，作者也指出了 OpenCL 和 CUDA 均是类似 C 的编程语言，编程风格极为接近，故基于 CUDA 的量子化学软件通过少量修改即可以提供基于 OpenCL 的异构计算支持。

除了针对 GPU 的异构计算研究，近年来基于 Intel MIC 构架的异构计算系统也受到越来越多的重视。2015 年，Shan 等人的针对 MIC 架构进行了 NWChem 程序的线程级并行优化。他们利用了 TEXAS 积分库取得了 65 倍加速，并在构造 Fock 矩阵的过程中取得了 1.6 倍加速[27]。同年 Edmond Chow 等人在天河二号上进行了 Hartree-Fock 计算的并行优化[30, 31]。基于天河二号超级计算机的架构，可扩展的大规模并行算法和基于 MIC 架构的向量化算法是优化重点。他们发展了可扩展的构造 Fock 矩阵的方法和设计了负载平衡优化的算法，并将 Global Array 应用到“天河二号”上。同时他们也

利用了 ERD 积分库，并对该积分库进行了向量化优化，使其可以运行在 Xeon Phi 融核处理器上。最终可计算的体系达到了 2 938 个原子和 27 394 个基函数，并可运行在 8 100 个节点上。

如上所述，向量化是 MIC 优化的关键环节。Benjamin P. Pritchard 和 Edmond Chow 针对双电子积分的向量化，开发了新的积分库 SIMINT[32]。在这一工作中，他们利用了 Obara-Saika 递推方法，将平行递推关系进行了向量化，并设计了地址连续、负载平衡的计算算法。在 Intel 支持 AVX 指令集的 CPU 上取得了 2—4 倍的加速。

4　异构计算相关软件介绍

如前所述，目前已经有许多量子化学软件进行了异构计算的相关研究。从 NVIDIA 官网上可以了解到，进行了 NVIDIA GPU 加速相关研究的量子化学软件有 26 款之多，涉及到矩阵运算、Hartree-Fock、DFT 自洽场计算、Møller-Plesset perturbation theory（MP2）、Coupled Cluster Singles and Doubles（CCSD）等多种计算方法，以及单 GPU、多 GPU 等多种硬件架构。可以说相关领域的研究正如火如荼的进行。以下我们对较为流行的几种量子化学软件的异构计算研究进展进行简要介绍。

（1）Gaussian

GAUSSIAN 16 版程序完成了 100% 的 PGI OpenACC 移植，可以在 GPU 上进行 Hartree-Fock 和 DFT 的能量、一阶梯度、二阶梯度的计算以及 CCSD(T)的计算。GPU 与 CPU 的计算结果没有精度上的差异。程序支持多 GPU 并行，相对于 CPU 取得了一定的加速。

（2）GAMESS-US

GAMESS-US 程序可以利用 GPU 加速进行 Hartree-Fock 计算、MP2 方法和 CCSD(T)的计算，使用了基于 Rys 多项式方法的 libqc 积分库，并支持多节点与多 GPU 计算。

（3）NWChem

NWChem 在 6.3 版本中加入了 GPU 加速的支持，可以进行高可扩展的 multi-reference coupled cluster 计算。程序同样支持多 GPU 以及部分 Xeon Phi 协处理器加速。在 2013 年和 2014 年分别有文献介绍了 NWChem 针对 CCSD(T)方法在 GPU 和 Xeon Phi 上的加速优化情况[33,34]。根据 NWChem 官方测试数据，在计算 CCSD(T)方法的非迭代部分时，Intel Xeon Phi 协处理器与 NVIDIA Kepler GPU 有相当的加速效果[35]。

（4）TeraChem

TeraChem 程序是完全基于 GPU 的量子化学程序，可以进行 Hartree-Fock 和 Kohn-Sham 能量和梯度的计算，完全支持 s、p、d 型基函数，可以进行几何构型优化和过渡态搜素，可以进行从头算分子动力学模拟。

5　总结和展望

目前，高性能计算机已进入了异构计算时代，而大数据、机器学习和人工智能在量子化学研究中也初露峥嵘。作为理论、实验外的“第三科研范式”，计算模拟已获得了广泛应用。量子化学计算软件已成为了重要的、软的“实验仪器”，在科学研究中发挥着举足轻重的作用。在欧美高性能计算应用领域，继承已有很广泛应用的量子化学软件代码并提高其性能是非常热门、却似乎有些“无解”的课题。彻底重构不失为一个非常好的解决方案，但代价昂贵。如对于诸多著名的量子化学程序来说，过去和现在的辉煌反而成为了某种意义上的“历史负担”，而新的完全基于异构计算平台的 TeraChem 等愈发崭露头角。这也为国产量子化学软件的跨越式发展提供了良好的战略机遇期。

值得注意的一点是，我国的高性能计算机硬件取得了长足的进步，已经连续取得了七次 TOP500 排名第一。然而作为计算量排名前三位的国产量子化学软件的缺失已成为了大问题。它不但制约着我国量子化学及应用的发展，而且也大大缩小了高性能计算的应用空间。因此，在量子化学软件研发过程中，充分发挥现代高性能计算机的性能应当引起足够重视，而适应现代高性能计算机体系架构也应成为国产量子化学计算软件的重要特色。

致谢　本研究得到了中国科学院科研信息化应用工程项目（XXH13506-403）和国家自然科学基金青年科学基金项目（21703260）的支持。在本文撰写过程中得到了复旦大学徐昕教授的大力支持，在此表示感谢。

参　考　文　献

[1] Gordon MS, Fedorov DG, Pruitt SR, et al. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. Chemical Reviews, 2012, 112(1): 632－672.

[2] Sherrill CD, Schaeffer HF. The configuration interaction method: Advances in highly correlated approaches. Advances in Quantum Chemistry, 1999, 34: 143－269.

[3] Knecht S, HedegÅerd ED, Keller S, et al. New Approaches for ab initio calculations of molecules with strong electron correlation. Chimia, 2016, 70, 244.

[4] Chan, GKL, Sharma S. The density matrix renormalization group in quantum chemistry. Annual Review of Physical Chemistry, 2011, 62: 465 - 481.

[5] Chan, GKL, Keselman A, Nakatani N, et al. Matrix product operators, matrix product states, and ab initio density matrix renormalization group algorithms. The Journal of Chemical Physics, 2016, 145(1): 014102.

[6] https://en.wikipedia.org/wiki/List_of_quantum_chemistry_and_solid-state_physics_software. 2017-11-28.

[7] 鲍建樟,丰鑫田,于建国. GPU 引发的计算化学革命. 物理化学学报, 2011, 27: 2019 - 2026.

[8] 刘松,鲍建樟,李长瑜,等. GPU 计算:突破制约计算化学发展的瓶颈. 科研信息化技术与应用, 2014, 5: 73 - 81.

[9] Shaw DE, Grossman JP, Bank JA, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC14), Piscataway, NJ: IEEE, 2014: 41 - 53. Gordon Bell Prize.

[10] http://www.top500.org.

[11] NVIDIA. CUDA C Programming Guide. http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf.

[12] NVIDIA. GF100 Whitepaper, V1.5.

[13] Yasuda K. Two-electron integral evaluation on the graphics processor unit. Journal of Computational Chemistry, 2008, 29(3): 334 - 342.

[14] Ufimtsev IS, Martinez TJ. Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. Journal of Chemical Theory and Computation, 2008, 4(2): 222 - 231.

[15] Ufimtsev IS, Martinez TJ. Quantum chemistry on graphical processing units. 2. direct self-consistent-field implementation. Journal of Chemical Theory and Computation, 2009, 5(4): 1004 - 1015.

[16] Ufimtsev IS, Martinez TJ. Quantum chemistry on graphical processing units. 3. analytical energy gradients, geometry optimization, and first principles molecular dynamics. Journal of Chemical Theory and Computation, 2009, 5(10), 2619 - 2628.

[17] Luehr N, Ufimtsev IS, Martinez TJ. Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs). Journal of Chemical Theory and Computation, 2011, 7(4): 949 - 954.

[18] Isborn CM., Ufimtsev IS, Martinez TJ. Excited-state electronic structure with configuration interaction singles and Tamm-Dancoff time-dependent density functional theory on graphical processing units. Journal of Chemical Theory and Computation, 2011, 7(6): 1814 - 1823.

[19] Snyder JW, Curchod BFE, Martínez TJ. GPU-accelerated state-averaged casscf interfaced with ab initio multiple spawning unravels the photodynamics of provitamin D3. The Journal of Physical Chemistry Letters, 2016, 7(13): 2444 - 2449.

[20] Snyder JW, Fales BS, Hohenstein EG, et al. A direct-compatible formulation of the coupled perturbed complete active space self-consistent field equations on graphical processing units. The Journal of Chemical Physics, 2017, 146(17): 174113.

[21] Snyder JW, Parrish RM, Martínez TJ. α-CASSCF: An efficient, empirical correction for sa-casscf to closely approximate MS-CASPT2 potential energy surfaces. The Journal of Physical Chemistry Letters, 2017, 8: 2432 - 2437.

[22] Asadchev A, Allada V, Felder J, et al. Uncontracted Rys quadrature implementation of up to g functions on graphical processing units. Journal of Chemical Theory and Computation, 2010, 6(3): 696 - 704.

[23] Asadchev A, Gordon MS. New multithreaded hybrid cpu/gpu approach to hartree-fock. journal of chemical theory and computation, 2012, 8(11): 4166 - 4176.

[24] Miao Y, Merz JrKM. Acceleration of electron repulsion integral evaluation on graphics processing units via use of recurrence relations. Journal of Chemical Theory and Computation, 2013, 9(2): 965 - 976.

[25] Wilkinson KA, Sherwood P, Guest MF, et al. Acceleration of the GAMESS-UK electronic structure package on graphical processing units. Journal of Computational Chemistry, 2011, 32(10): 2313 - 2318.

[26] Kalinowski J, Wennmohs F, Neese F. Arbitrary angular momentum electron repulsion integrals with graphical processing units: application to the resolution of identity hartree fock method. Journal of Chemical Theory and Computation, 2017, 13: 3160 - 3170.

[27] Kussmann J, Ochsenfeld C. Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. Journal of Chemical Theory and Computation, 2017, 13(7): 3153 - 3159.

[28] Kussmann J, Ochsenfeld C. Employing OpenCL to accelerate ab initio calculations on graphics processing units. Journal of Chemical Theory and Computation, 2017, 13: 2712 - 2716.

[29] Shan H, Williams S, Jong W, et al. Thread-level parallelization and optimization of nwchem for the intel mic architecture, in PMAM'15 Conference, February 07-11 2015, San Francisco, CA, USA.

[30] Chow E, Liu X, Smelyanskiy M, et al. Parallel scalability of Hartree-Fock calculations. The Journal of Chemical Physics, 2015, 142(10): 104103.

[31] Chow E, Liu X, Misra S, et al. Scaling up Hartree-Fock calculations on Tianhe-2. The International Journal of High Performance Computing Applications, 2016, 30(1): 85 - 102.

[32] Pritchard BP, Chow E. Horizontal vectorization of electron repulsion integrals. Journal of Computational Chemistry, 2016, 37(28): 2537 - 2546.

[33] Bhaskaran NK., Ma W, Krishnamoorthy S, Villa O, van Dam HJJ, Aprà E, Kowalski K. Noniterative Multireference Coupled Cluster Methods on Heterogeneous CPU-GPU Systems. Journal of Chemical Theory and Computation, 2013, 9: 1949 - 1957.

[34] Apra E, Klemm M, Kowalski K. Efficient Implementation of Many-Body Quantum Chemical Methods on the Intel® Xeon Phi Coprocessor. In SC '14 Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New Orleans, Louisana: IEEE Press, 2014: 674 - 684.

[35] http://www.nwchem-sw.org/index.php/Benchmarks

The development of heterogeneous computational quantum chemistry software meets the time

Tian Yingqi[1,2,3]　Ma Yingjin[1,2]　Suo Bingbing[2,4]　Jin Zhong[1,2]

(1. *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190;
2. *Center of Scientific Computing Applications & Research, Chinese Academy of Sciences, Beijing* 100190;
3. *University of Chinese Academy of Sciences, Beijing* 100049 *Institute of Modern Physics, Northwestern University, Xi'an* 710069))

Abstract　Quantum chemistry software plays an important role in chemistry researches. Developing high performance computing software with cutting-edge computer architecture is a widely concerned issue for chemistry researchers and software developers. In recent years, accelerators like Graphic Processing Unit (GPU) are showing significant power in computing, and aroused general interest. In this article, we introduced hardware and software basics of heterogeneous computing; summarized researches and applications dealing with heterogeneous computing (both GPU and co-processor are involved) in quantum chemistry; and briefly introduced related quantum chemistry software. We can conclude form these researches that heterogeneous computing is becoming a great power in accelerating quantum chemistry calculation. However, achievements of many popular quantum chemistry software packages restrained their further development in adapting to modern supercomputer. Heterogeneous computing provides development of domestic quantum chemistry software a good opportunity. This leads good time to develop such kind of software package.

Key words　quantum chemistry; heterogeneous optimization; two-electron repulsion integrals; GPU; co-processor