

杰出青年人才发展专项计划年度进展报告书

(2010 年度)

姓名	张云泉	课题名称	并行算法与并行软件
资助金额	160.00 万元	课题起止时间	2009.9-2013.9
资助类别	应用基础	所在部门	并行软件与计算科学实验室

研究工作主要进展和阶段性成果（含论文、研究生培养等）

2010 年，在所杰出青年人才发展专项计划的支持下，开展了如下研究工作：

1. 并行计算模型和并行编程模型的研究进展

- 1) 调研国际上基于层次存储或层次并行的计算模型和编程模型研究进展，对主流并行程序设计语言的局部性设计机制进行了深入的分析和分类，总结了其优缺点，通过调研综述，把握了国内国际基于层次存储或层次并行的并行计算模型和并行编程模型的研究进展，指出了新一代并行编程语言所应具有若干语言特点，重点提出了应综合考虑横向局部性和纵向局部性支持的设计机制的研究的观点。图 1 为主流并行编程语言和面向高生产率并行编程语言在地址空间以及横向局部性的位置图，图 2 为 Cell 编程模型的横向和纵向局部性位置图。相关研究工作整理出综述技术报告一篇[17]。

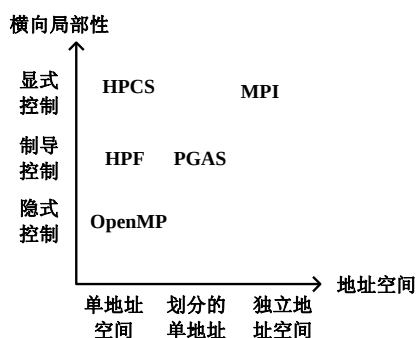


图 1 编程语言的横向局部性及地址空间

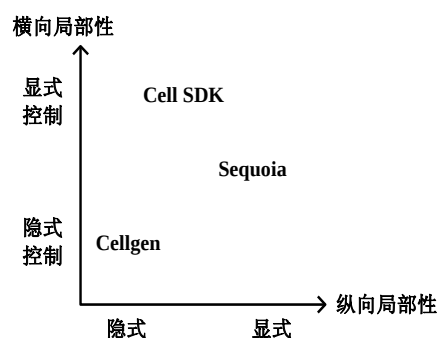


图 2 Cell 编程模型的横向和纵向局部性

- 2) 提出一个新的基于延迟隐藏因子的 GPU 并行计算模型：针对 GPU 上编程和调优过程的繁琐和 GPU 低延迟高带宽的特性，提出了基于延迟隐藏因子的 GPU 计算模型，图 3 是 GPU 模型计算流程图。图 4 是 GPU 上的 Volkov 矩阵乘算法在不同规模下的预测结果和实测结果图，实验结果表明了模型的正确性。相关工作在《软件学报》发表文章一篇[11]。

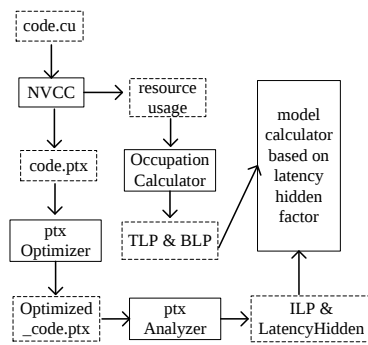


图 3 GPU 计算模型计算流程图

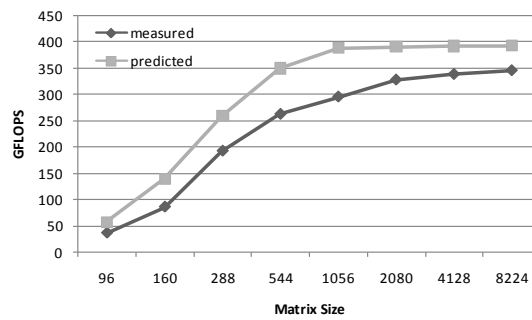


图 4 模型预测值与实测数值

- 3) 层次, Allgather 的平均预测误差由原始 LogGP 模型的 36%下降到 12%。相关工作发表在 IEEE CSE 2010 国际会议上[6]。

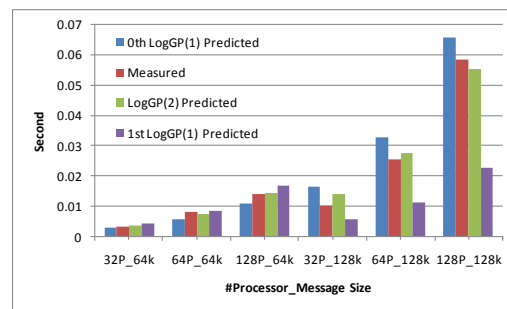


图 5 魔方上对邻居交换算法采用 LogGP 和 LogGPH 模型的预测和实测值对比

2. 以 FFT 算法和 PDE 求解算法等基础算法为研究目标，在国产百万亿次机器进行几千处理器的可扩展性测试，发现并行算法设计的瓶颈，分析其原因。

- 1) SpMV 是求解 PDE 的一个核心算法，根据应用矩阵非零元的对角线式的分布特点创新性地提出了 CVRSD 矩阵存储格式。该格式能够基于运行平台的特点自适应的选择最优的 SpMV 实现。测试结果表明，所提出的新存储格式针对不同问题规模的矩阵，在不同线程下均优于 MKL 中基于 CSR 的 SpMV 实现，测试的性能与加速比结果如下图 6 所示。可以看出 CVRSD 在单线程下，相对经典存储格式 CSR 运行效率，加速比平均为 3.05 倍，最高可以到达 4.38 倍；相对于 MKL 的 DIA 存储格式，对于不适合 DIA 存储格式的矩阵，CVRSD 同样有很好的加速效果，如图中编号为 4 和 5 的矩阵，加速比可以达到 50 倍以上，除此之外，其余最高加速比为 2.02，平均加速比为 1.74。在多核情况下，CVRSD 的加速效果在相同处理器下，是 CSR 存储格式的 2.16 倍，最高可以达到 4.61 倍。相关工作申请发明专利一项（申请号：201010594672.1），正在撰写高水平国际会议文章，拟投 EuroPar 2011。

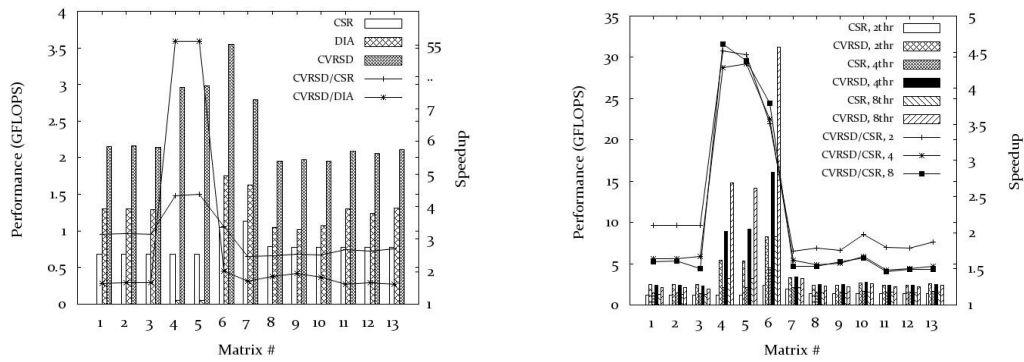


图 6 不同矩阵存储格式的性能比较

- 2) 基于对角线数据结构的 SpMV 性能优化: 针对对角线模式的稀疏矩阵存储方法最主要为 **DIAG**, 这种方法会多存储矩阵范围外的非零元, 对于远离主对角线的对角线, 增加了存储矩阵外虚位的负担, 提出了一种新的 **DDD-SPLIT** 存储方法。该方法只存储矩阵范围内的非零元及相应对角线上的零元, 根据过对角线与矩阵两侧交点的水平线对矩阵进行切分, 使得每个子稀疏矩阵行块内所需要的对角线相同, 因此计算模式一致, 可以重用 **y**。图 7 为 **DDD-SPLIT** 对 **CSR** 的加速比。相关工作申请发明专利一项 (申请号: 201010594672.1), **IEEE HPCC 2010** 国际会议发表文章一篇[5]。

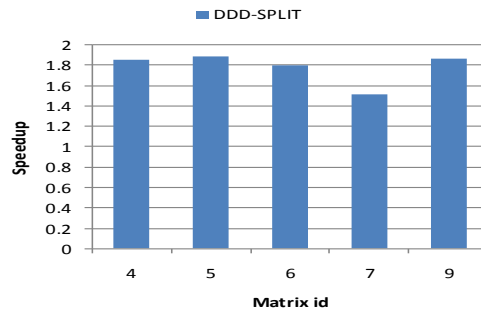


图 7 DDD-SPLIT 加速比

- 3) 通过在计算模型、大型稀疏线性方程组高效并行解法器、稀疏矩阵向量乘稀疏矩阵存储格式、MPI 与 OpenMP 的混合并行和可扩展消息传递集合通信算法等多个方面开展关键技术的研究工作, 在国产百万亿次平台曙光 5000A (魔方) 和深腾 7000 上, 使地球外核热流动的数值模拟程序的强可扩展性达到了 2048 处理器核以上。相关工作在 **IEEE HPCC 2010** 和 **HPCA2010** 国际会议上发表文章各一篇[2, 4], 国内一级学报《数值计算与计算机应用》发表文章一篇[10], 国内会议发表文章两篇[18, 19]。

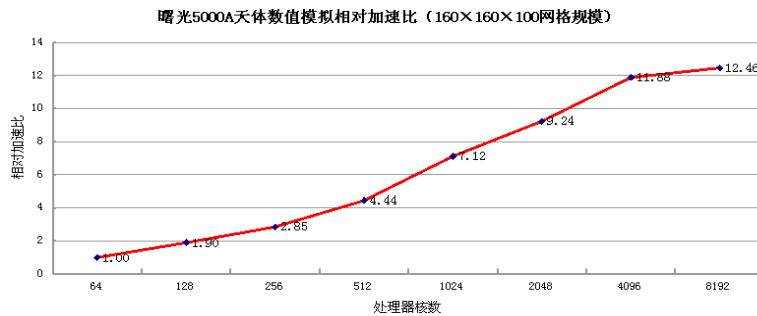


图 8 曙光 5000A 天体数值模拟加速比随处理器核数的变化 (网格规模 $160 \times 160 \times 100$)

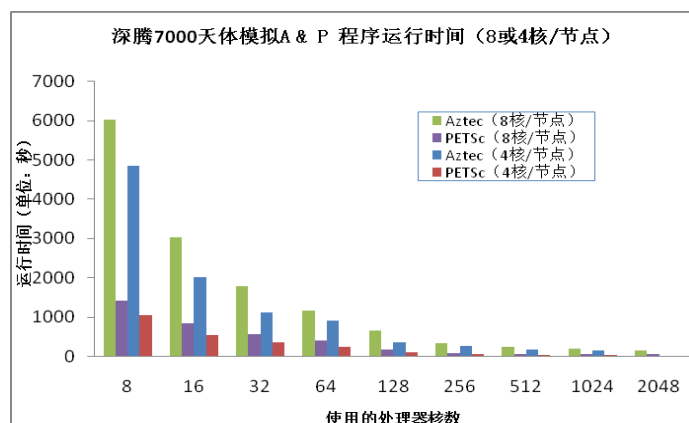


图 9 深腾 7000 天体模拟 Aztec 与 PETSc 程序运行时间 (8 或 4 核/节点)

3.集合通信算法在国产百万亿次计算平台上的大规模测试，分析及优化

- 1) 基于新的 LogGPH 模型设计了新的 MPI_Allgather 映射算法，平衡了 MPI_Allgather 邻居交换映射算法中偶数步通信层次，在魔方单节点上测试效果平均加速比为 6%。图为魔方上新旧映射方法对比的加速比。

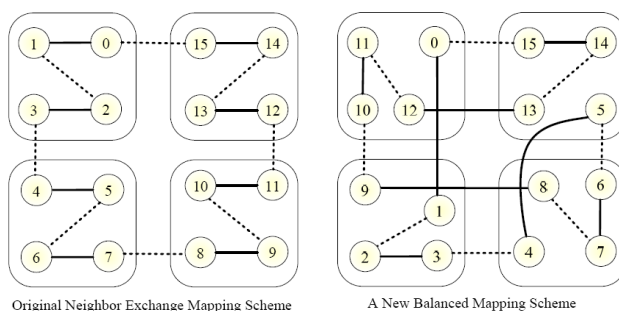


图 10 邻居交换算法新进程映射模式

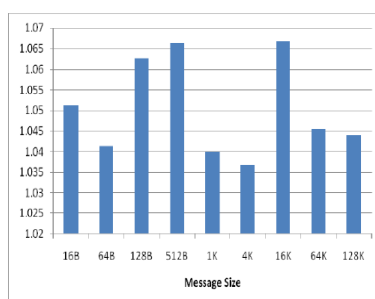


图 11 魔方上新映射方法加速比

- 2) 在魔方上，采用 MPICH 进程绑定工具 hydra 测试分析了 6 种不同进程绑定策略对点集合通信 (IMB3.2) 性能的影响，并分析了 Broadcast, Gather/Scatter, Allgather, Alltoall, Allreduce 算法在不同消息长度的实际运行程序[性能可以提高的空间。图 12 显示，长消息 ring 算法的 Allgather 性能至少可以提升 27% (性能最优的 CASE5 相对于性能最坏的 CASE3)。图 13, 中短消息 Bcast 性能至少可以提升 31%。

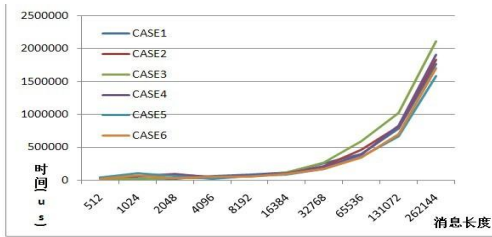


图 12 1024 进程不同绑定策略的 Allgather

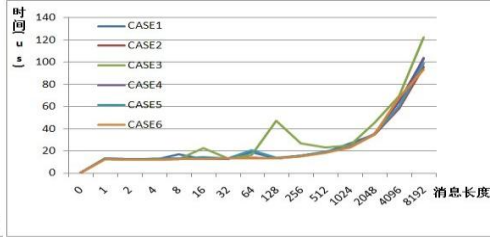


图 13 1024 进程不同绑定策略的 Bcast

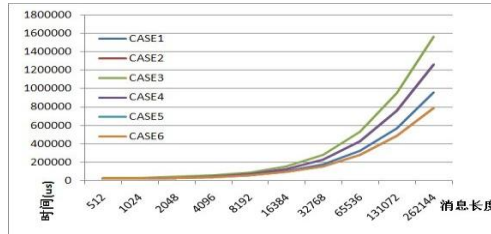


图 14 1024 进程不同绑定策略的 Allgather 模拟

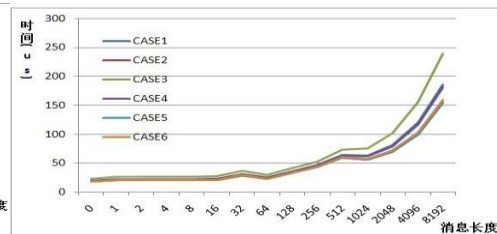


图 15 1024 进程不同绑定策略的 Bcast 模拟

- 3) 通过分析不同进程绑定策略下 CPU 内、CPU 间、结点间通信的次数和消息大小，计算模拟了集合通信的实际运行时间，图 14 为 1024 进程 6 种不同进程映射策略下长消息的 ring 算法的 Allgather 模拟时间，图 15 为短消息的 Bcast 的模拟时间。相关工作的初步成果在国内核心期刊发表文章一篇[16]，进一步的工作正在撰写国际会议文章。

4. 龙芯 3 国产多核处理器的国产高性能扩展数学库研发进行关键技术预研；

目前，基于龙芯 3A 国产四核 CPU 进行关键技术研发，已完成龙芯 3A 体系结构调研报告和优化技术报告。具体来说，针对高性能扩展数学库的特点，研究了自适应性能优化技术，递归分块技术，地址交错技术和多核并行化技术等，以提高数学库性能。

设计和初步实现了针对龙芯 3A CPU 的国产高性能扩展数学库 CLeXML，包含 BLAS，LAPACK，FFT，直接解法器和迭代解法器等 5 个模块，已申请计算机软件著作权登记：龙芯 CPU 多核高性能数学库软件（受理号：2010R11L099597）。进行了数学库的正确性、性能和健壮性测试。部分应用了所预研的性能优化关键技术，比如：自适应性能优化，地址交错技术和多核并行化等。在龙芯 3A 平台，CLeXML 库 5 个模块，2 线程和 4 线程相对于单线程的平均加速比，分别为 1.43 和 1.91。相关工作发表 IEEE NAS 2010 和 CIT 2010 国际会议文章各一篇[3, 7]，国内一级学报文章三篇[8.9.11]，国内核心期刊一篇[14]。国内会议文章二篇[21, 22]。

综上所述，本课题 2010 年支持毕业硕士生 3 名，其中一名硕士生获得优秀毕业生，本人获得优秀导师。相关工作发表在 FCSC 国际期刊文章 1 篇，IEEE HPCC，IEEE CSE，IEEE NAS 和 IEEE CIT 等国际会议发表文章 6 篇，《计算机研究与发展》、《数值计算与计算机应用》、《软件学报》、《计算机科学》和《计算机工程》等国内期刊发表文章 10 篇，国内会议发表文章 5 篇，申请发明专利 2 项，申请软件著作权 1 项。

- 专利发明人：孙相征，张云泉，王婷，刘芳芳，袁良；专利申请名称：针对稀疏矩阵的数据存储方法及基于该方法的 SpMV 实现方法；申请号：201010594672.1。
- 专利发明人：袁良，张云泉，孙相征，王婷，刘芳芳；专利申请名称：稀疏矩阵的对角线数据存储方法及基于该方法的 SpMV 实现方法；申请号：201110004075.3。
- 计算机软件著作权登记，龙芯 CPU 多核高性能数学库软件，受理号：2010R11L099597。

1. Yunquan Zhang, Jiachang Sun, Guoxing Yuan and Linbo Zhang. Perspectives of China's HPC system development: a view from the 2009 China HPC TOP100 list. *Frontiers of Computer Science in China*, Volume 4, Number 4, pp.437-444, Springer, DOI: 10.1007/s11704-010-0372-0.
2. Chao Yang, Ligang Li, Yunquan Zhang, Development of a scalable solver for the Earth's core convection, in proceedings of HPCA 2009, *Lecture Notes in Computer Science*, vol. 5938, pp 497-502, Springer, 2010.
3. Lei Wang, Yunquan Zhang, Xianyi Zhang, Fangfang [Liu](#). Accelerating Linpack Performance with Mixed Precision Algorithm on CPU+GP GPU Heterogeneous Cluster. 10th IEEE International Conference on Computer and Information Technology, CIT 2010, Bradford, West Yorkshire, UK, June 29-July 1, 2010. IEEE Computer Society 2010, ISBN 978-0-7695-4108-2: 1169-1174.
4. [Chao Yang](#), Yunquan Zhang, [Ligang Li](#). Numerical Simulation of the Thermal Convection in the Earth's Outer Core. 12th IEEE International Conference on High Performance Computing and Communications, HPCC 2010, 1-3 September 2010, Melbourne, Australia: 552-555.
5. [Liang Yuan](#), Yunquan Zhang, [Xiangzheng Sun](#), [Ting Wang](#). Optimizing Sparse Matrix Vector Multiplication Using Diagonal Storage Matrix Format. 12th IEEE International Conference on High Performance Computing and Communications, [HPCC 2010](#), 1-3 September 2010, Melbourne, Australia: 585-590.
6. Liang Yuan, Yunquan Zhang, Yuxin Tang, Li Rao, Xiangzheng Sun. LogGPH: A Parallel Computational Model with Hierarchical Communication Awareness. The 13th IEEE International Conference on Computational Science and Engineering (CSE-2010). Dec. 11-13, Hang Kong, China: 269-274.
7. Yan Li, Yunquan Zhang, Ke Wang, Wenhua Guan. Heterogeneous Multi-core Parallel SGEMM Performance Testing and Analysis on Cell/B.E Processor. IEEE NAS 2010, JULY 15-17, Macau, China: 202-207.
8. 孙相征, 张云泉, 王宣强, 王磊。数值软件自适应性能优化搜索过程评价技术研究。《计算机研究与发展》, 2010, Vol. 47, (4):679-686。
9. 刘胜飞, 张云泉, 孙相征。一种改进的 OpenMP Guided 调度策略研究。《计算机研究与发展》, 2010, Vol. 47, (4):687-694。
10. 余元, 张云泉, 李会元。一类非张量区域快速傅立叶变换算法在国产并行机上的可扩展性测试, 《数值计算与计算机应用》, 2010, 31(2): 123-130。
11. 陈少虎, 张云泉, 张先轶, 程豪。BLAS 库在多核处理器上的性能测试与分析, 软件学报, 2010 年增刊
12. 袁良, 张云泉, 龙国平, 王可, 张先轶, 基于延迟隐藏因子的 GPU 计算模型, 软件学报, 2010 年增刊
13. 孙相征, 张云泉, 王婷, 杨超, 李力刚。天体大规模数值模拟软件性能优化。《华中科技大学学报(自然科学版)》。2010 年 6 月, Vol. 38, 增刊 I: 69-76。
14. 程豪, 张云泉, 张先轶, 李玉成。CPU-GPU 混合并行 DGEMM 算法实现与性能分析。《计算机工程》, Vol. 36, No.13, pp.24-26, 2010 年 7 月。
15. 王磊, 张云泉, 刘芳芳, 张先轶, 基于混合精度算法的改进 HPL 软件包, 《计算机工程》, Vol. 36, No.19, pp.47-49, 2010 年 10 月。
16. 饶立, 张云泉, 李玉成。国产百万亿次机群系统 Alltoall 性能测试与分析。《计算机科学》, Vol. 8, 2010 年。
17. 袁良, 张云泉, 并行程序设计语言中局部性机制的研究, 《高性能计算发展与应用》, 2010 年第一期:34-44。

18. 王婷, 张云泉, 孙相征, 杨超, 李力刚。天体大规模数值模拟软件性能优化初探。第二届中国国家网格学术年会 (CNGRID2010) 论文集, 2010 年 1 月 15-17, 北京: 87-92。
19. 王婷, 张云泉, 孙相征, 杨超。行星流体动力学大规模计算的性能测试与分析。HPC China 2010, 2010 年 10 月 28-30 日, 北京。
20. 费辉, 张云泉, 王可。分子动力学模拟在 OpenCL 框架的 GPU 上的实现。HPC China 2010, 2010 年 10 月 28-30 日, 北京。
21. 李焱, 张云泉, 王可, 赵美超。异构平台上基于 OpenCL 的 FFT 测试与分析。HPC China 2010, 2010 年 10 月 28-30 日, 北京。
22. 解庆春, 张云泉, 王可, 李焱。SIMD 技术与向量数学库研究。HPC China 2010, 2010 年 10 月 28-30 日, 北京。

下一年度工作计划, 包括国内外合作与交流计划

1. 继续改进和优化新的基于层次存储和层次并行的并行计算模型，并进行模型分析和验证
 - 1) 2010 年 12 月，经张云泉研究员邀请，美国阿贡国家实验室助理计算机科学家 Pavan 博士来访中科院软件所并行软件与计算科学实验室 2 个月。基于访问成果，将和 Pavan 合作，在 GPU 上扩展 MPICH 通信函数，提高 MPICH 在 GPU 上的编程效率，并进行优化。
 - 2) 继续优化改进 GPU 计算模型，用更多 kernel 对模型进行测试，基于测试结果，对模型进行优化。
 - 3) 基于提出的 LogGPH 模型，对更多的集合通信算法，更多的实际应用进行算法改进，基于算法改进经验，设计映射优化算法，实现算法映射自适应。
2. 在已有国产百万亿次机器找到并行算法设计的瓶颈问题，依据进行测试的实验结果，进行详细分析，基于测试经验对 FFT 算法和 PDE 求解算法等基础算法设计实现有潜力扩展到几千处理器的并行算法，并进行初步测试。
3. 在国产百万亿次计算平台上发现 MPI_Alltoall 和 MPI_Allgather 等算法的性能瓶颈，尝试提出新的适合国产计算平台的大规模集合消息传递算法；
 - 1) 根据不同进程绑定策略性能不同的分析设计出 Hardware-aware 的最优算法，采用 PMI 获取用户的绑定信息（默认为随机绑定），并将其映射为最优算法以获取较优（存在不可能映射的可能）性能。
 - 2) 进一步利用提出的 LogGPH 模型，对更多具有明显时间步特征的集合通信算法以及实际应用进行算法。
4. 继续完善和优化国产高性能扩展数学库 CLeXML，支持将于 2011 年推出的龙芯 3B CPU，针对其上的 SIMD 指令进行优化；
5. 相关成果撰写高水平学术文章，并投稿著名国际会议；