格点 QCD 计算的软硬件基础

宫明

中科院高能所
CLQCD 合作组

中科院理论所
2018.10.29

提纲

提纲

格点 QCD 计算的基本算法

蒙特卡罗算法

1. 生成组态（HMC）
2. 计算物理量（夸克传播子与强子关联函数）
3. 统计（误差分析）

计算规模的来源

1. 统计量要足够多
2. 物理体积要足够大
3. 最小格距要足够小
4. ~~夸克质量足够轻~~

- 瓶颈问题：自由度太多导致费米子矩阵 M 太大（维度约 $O(10^5 - 10^8)$）

问题的降解

费米子矩阵 M 本身

- 局域费米子作用量的 M 是稀疏且规则的
- 非局域费米子作用量 (比如 overlap) 的 M 不稀疏
  - 可以展开为稀疏矩阵的多项式或分式

HMC 算法

- MD 算法
  - 需要计算费米力（RHMC/PHMC）
    - 涉及求解线性方程组
- Metropolis 算法
  - 需要用伪费米子场计算费米子矩阵行列式
    - 涉及求解线性方程组

问题的降解

夸克传播子

- 费米子矩阵的逆（或逆矩阵列的线性组合）
  - 求解线性方程组

求解线性方程组

- Krylov 子空间迭代算法
  - 迭代算法
  - 预处理技术
  - Multi-shift 技术

具体操作

- 矩阵乘以向量：4 维（或 5 维）Stencil 运算
- 向量内积：全局归约

提纲

格点 QCD 的计算任务

计划中的任务

1. 产生组态
2. 计算夸克传播子
3. 计算关联函数
4. 拟合、分析误差

实践中的任务

1. 讨论、推公式
2. 写代码：硬件适配、并行优化、算法实现、具体的计算过程……
3. debug、debug、debug……
4. 提交和管理任务、整理数据、转格式
5. 拟合、画图、拟合、画图……
6. 写文章、与审稿人吵架

整理一下：

计算的逻辑层级

5 **调度管理**：工作流：自动化脚本、数据可视化、针对格点 QCD 设计的 DSL 等

4 **数学算法**：数据处理：拟合算法、重取样算法等

3 **物理图景**：物理层优化：作用量选取、多尺度高阶积分算法、关联函数缩并、蒸馏算法、集团分解算法等

2 **数学算法**：Krylov 子空间求解：预处理类算法、域分解算法、Deflation 算法、混合精度技术等

1 **硬件环境**：4 维（或 5 维）stencil $\emptyset$：不同硬件环境的适配与极限优化

国外的软件基础

USQCD 框架

Chroma CPS FUEL MILC QLua

Inverters MDWF QOPQDP **QUDA** **QPhiX**

QDP++ QDP/C **QDP-JIT** QIO

QLA **QMP** QMT

一些“散在”代码包

- **Grid**、DDHMC、**cl2QCD**、χQCD/GWU codes、XMBF ⋯⋯
- 各合作组的大量“祖传代码”

国内的代码现状

CLQCD 合作组

- 大量零散“祖传代码”，包括在 Quenched 近似下的组态产生代码、计算介子散射、胶球等各类物理问题的专用代码等。
- 修改版的 QUDA 代码，来自 ETMC 合作组、χQCD 合作组的共享代码等。
- 用于大规模复杂数据处理和工作流控制的 QScheme 语言。
  - 基于 scheme 语言的领域特定编程语言（DSL）。
  - 支持自动 MPI 并行化，适合大规模复杂数据处理。
  - 扩展性强，可以方便地作为胶水语言使用。
  - 目前仅有解释器，编译器在研发中。
- 针对国产申威架构设计的 SimpleQCD 代码包（正在开发中）。
  - 代码能够高效地运行在申威独特的片上异构硬件上。
  - 目前版本实现了基本的 $\not{D}$ 与 MinRes 求解算法。
  - 未来版本将与 Chroma 对接，从而移植更丰富的功能。
- ……

提纲

计算机计算能力的发展

摩尔定律

这里一般会放一张图

但其实没必要

- 我们应该真正关注的问题是什么？

是什么在推动计算能力的发展？

- 制程、主频
- 多核
- 异构

是什么在推动 (制约) 格点 QCD 软件的发展？

- 人力、资源
- 合作
- 有远见的顶层设计

E 级超级计算机前瞻

中国

- 神威
  - 江南计算所
  - 片上异构
- 曙光
  - AMD
  - 异构，x86 CPU + DCU
- 天河
  - 国防科大
  - 众核异构？

E 级超级计算机前瞻

美国 (Coral-2)

- Aurora A21
  - Cray + Intel
  - x86 协处理器 ? GPU ?
- Frontier/El Capitan ??
  - IBM + Nvidia
  - 异构，Power CPU + GPU
- Frontier/El Capitan ??
  - HPE + AMD
  - 共享内存模式 ??

E 级超级计算机前瞻

日本

- Post-K
  - Fujitsu
  - Arm 异构

欧洲

- ??
  - Atos ??
  - Risc-V ? Arm ?

是什么在推动计算能力的发展？

- 制程、主频
- 多核
- 异构

是什么在推动 (制约) 格点 QCD 软件的发展？

- 人力、资源
- 合作
- 有远见的顶层设计

格点 QCD 软件的顶层设计

USQCD SciDAC 项目的设计问题

- 一个中心：以 QDP 接口为核心构建四层软件
  - 不按格子切分并行怎么办？(Grid)
  - 四层乱掉怎么办？(QUDA)
- 两个基本点：把进程模型和线程模型封装起来
  - 不得不拆封怎么办？(Chroma/openmp、QUDA/GPU direct)
  - 主从模型怎么办？(GPU)
  - 多级并行结构怎么办？(众核异构)
  - 其他奇葩模型怎么办？(申威从核)

CLQCD 的顶层设计原则

- 一张白纸，博采众长，弯道超车
- 基于国内的硬件环境、物理实验方向，发挥比较优势

提纲

什么是基础设施？

- 硬件资源
- 软件开发
- 数据积累
- 人才培养

硬件资源

- 国内大型超算
  - 资源丰富、价格不确定
  - 国产超算需要代码移植或重新开发
  - 大多数情况下内存带宽是瓶颈
- 单位中小型超算
  - 对于本单位的人近水楼台
  - 一般都是 GPU 集群，代码成熟
- 非主流的可能性？
  - 志愿计算平台
  - 新硬件架构测试（FPGA ?)
  - 量子计算机？

软件开发

在申威处理器上的代码开发

- 个人猜测：下一代 E 级计算机肯定有一台申威架构的
- 长期资源丰富，价格有商谈空间，也有合作可能
- 目前还在开发中，完成了一些阶段性的版本
- 申威架构目前的特殊性可能是未来超算的发展方向之一，相关的理论研究也值得跟进
- 在自动代码生成、自动优化等方面可以做一些尝试，为下一代软件的顶层设计方案打下基础

软件开发：关于 SimpleQCD 代码包

版本 0

- 在普通 cpu 上运行的版本，作为对照的原型程序。

版本 1.*

- 实现了从核上的 $\not{D}$ 和 MinRes 求解代码，也实现了主核上的 MPI 并行。
- 每个核组负责 8^4 子格子，从核切分方案固定为 xy 平面切片。

版本 2.*

- 作为一系列实验性的版本，验证自动代码生成与遗传算法优化的可行性。
- 每个核组负责 8^4 子格子，从核切分方案可以任意指定。
- 利用 DAG 的拓扑排序生成 64 组代码，分别用于 64 个从核。用任意长度的两列整数作为输入，不同数列导致不同的排序结果。
- 用遗传算法优化那两列整数，采用代码实际运行时间的函数作为生存权重。
- 这个方法的确可以优化代码的效率，但跑了几十个小时的优化后仍不如手工编写的代码效率高。

软件开发：关于 SimpleQCD 代码包

版本 3.*

- 手工编写的代码，每个核组负责 16^4 子格子，切分方案是固定的。
- 根据测算，每个核组负责超过 8^4 子格子时，必须多次在主存与从核局存间交换数据，而且有一部分边界数据需要交换两次。
- 根据测算，每个核组负责达到 16^4 子格子时，MPI 开销的大部分可以被计算掩盖。
- 这个版本正在紧张开发中，预计年内完成。

未来版本

- 每个核组应该负责超过 16^4 的子格子且这个大小应可调整。
- 根据测算，每个核组负责超过 16^4 子格子时，有至少 $1/6$ 的数据需要交换至少两次。代码会比较复杂，可能需要采用自动化生成和优化技术。
- 未来版本需要增加一个兼容 QDP 协议的接口，从而嵌入到 Chroma 软件内，从而无缝移植大量代码到申威架构上。

软件开发

向曙光和天河架构移植 USQCD 软件包和 χQCD 软件包

- 曙光采用 x86 CPU + DCU 架构，与 GPU 类似。
- 我们正在曙光 E 级原型机上尝试把 QUDA 进行 HIP 转码。
- 我们已经在天河 E 级原型机上成功运行并测试了 Chroma 软件。
- 未来天河 3 的实际结构可能与原型机不同，到时候还需要一些移植工作。

软件开发

数据文件格式标准

- 我们设计了 xdata(也称 io_general) 文件格式
- 其实就是带元信息的多维数组
- 元信息包括数组的维度和每个维度的指标
- 方便数据交换和数据处理

通用脚本环境

- QScheme 语言
- 可以直接操作 xdata 格式的数据
- 支持 MPI 并行、各种软件的接口调用等

软件开发：QScheme 语言

QScheme 是什么

- QScheme 是一个 scheme 的方言，scheme 是 lisp 的一个方言
- 引入了 xdata 数据类型和 xfile 文件格式，支持外部扩展，支持 MPI 并行。

QScheme 可以用来做什么

- 数据处理：批量运算、数据拟合、误差分析……
- 画图、提交与管理任务、与其他软件交互……
- 文本处理、自动代码生成……
- 作为胶水语言把其他所有工具联合在一起

QScheme 好在哪里

- 隐藏了实现细节，专注于物理，减少 bug
- 利用高阶函数自动“升级”代码，真正的模块化
- 极高的灵活性，方便组织调配外部工具、元编程等

举个栗子：xdata

数据 a[mass][operator][complex]：

- mass：1 2 3 5 6
- operator：0 1 2
- complex：0 1

数据 b[t][mass]：

- t：0 1 2 3 4 5 6 7 8 9
- mass：2 3 4 5

(x.+ a b) 的结果：c[mass][operator][complex][t]

- mass：2 3 5
- operator：0 1 2
- complex：0 1
- t：0 1 2 3 4 5 6 7 8 9

再举个栗子：MPI 并行化

I have a function

```
(define (my_func par1 par2 par3)
        ...
)
```

I have an MPI wrapper

```
(define (x.grind/mpi f . dims)
        ...
)
```

MPI function!

```
(define my_mpi_func
        (x.grind/mpi my_func 'mass 't))
```

数据积累

规范场组态生成

- 目的：
  - 根据研究目标选择格子和作用量参数，有的放矢
  - 培养产生组态的经验和能力，不再依赖于国外合作组的二手组态
- 目前计划：
  - 2 味 Wilson-Clover 费米子作用量
  - 各向异性格子，不对称率为 5
  - 中等规模格子，大统计量
  - 质量目标暂定约 $m_\pi = 300\text{MeV}$ 左右
- 有关的物理问题：
  - XYZ 粒子的研究
  - 粲偶素的衰变
  - 胶球的性质以及与粲偶素的耦合
  - 把 t 轴用于空间方向可以用来研究 PDF 等需要大动量的物理量
  - ……

数据积累

用蒸馏算法计算 perambulator

- 蒸馏是一种前期投入较高，后续测量免费的方案
- 保存下来的 perambulator 可以用于计算各种算符的两点函数，可以在很多物理问题中直接使用
- 可以考虑同时计算一些 “半边的”perambulator，用于计算三点函数

谢谢！

格点 QCD 计算的软硬件基础

宫明

中科院高能所
CLQCD 合作组

中科院理论所
2018.10.29