Simulation of Lattice QCD with a GPU Cluster

Ting-Wai Chiu[1,2], Tung-Han Hsieh[3], and Kenji Ogawa[1] (for the TWQCD Collaboration)

[1]Physics Department, National Taiwan University, Taipei, Taiwan

[2]Center for Quantum Science and Engineering, National Taiwan University, Taipei, Taiwan.

[3]Research Center for Applied Sciences, Academia Sinica, Taiwan.

Quantum Chromodynamics (QCD)

QCD is the quantum field theory for the strong interaction, describing the interactions of the quarks and gluons making up hadrons (e.g. proton, neutron, and pion). It accounts for the nuclear energy inside the nucleus of an atom, and plays an important role in the evolution of the early universe.

Salient features:

- Gauge group $SU(3) \Rightarrow$ gluons have self-interactions.
- Asymptotic freedom: $g(r) \to 0$ as $r \to 0$.
- IR slavory: $g(r) \simeq 1$ at $r \simeq 10^{-15}$ m $\Rightarrow$ quark/color confinement
- No exact analytic solutions

Quarks

Quarks are spin $\frac{1}{2}$ fermions carrying color, and there are 6 species (flavors) of quarks.

u c t	u c t	u c t
d s b	d s b	d s b

Hadrons are color singlets composed of quarks

$$P = uud + \text{antisym. in color}$$
$$N = udd + \text{antisym. in color}$$
$$\pi^+ = \bar{d}u + \bar{d}u + \bar{d}u$$

The nuclear force between nucleons emerges as residual interactions of QCD

The Challenge of QCD

At the hadronic scale, $g(r) \simeq 1$, perturbation theory is incapable to extract any quantities from **QCD**, nor to tackle the most interesting physics, namely, the **spontaneously chiral symmetry breaking** and the **color confinement**

To extract any physical quantities from the first principles of **QCD**, one has to solve **QCD** nonperturbatively.

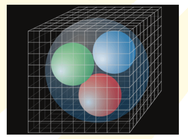
A viable nonperturbative formulation of **QCD** was first proposed by **K. G. Wilson** in 1974.

But, the problem of lattice fermion, and how to **formulate exact chiral symmetry on the lattice had not been resolved until 1992-98. [Kaplan, Neuberger, Narayanan,…], i.e., Lattice QCD with exact Chiral Symmetry.**

Basic notions of Lattice QCD

1. Perform Wick rotation: $t \to -ix_4$, then $\exp(iS) \to \exp(-S_E)$, and the expectation value of any observable O

$$\langle O \rangle = \frac{1}{Z}\int [dA][d\psi][d\bar{\psi}] O(A,\psi,\bar{\psi}) e^{-S_E}$$
$$Z = \int [dA][d\psi][d\bar{\psi}] e^{-S_E}$$



2. **Discretize the space-time as a 4-d lattice** $L^4 = (Na)^4$ with lattice spacing a. Then the path integral in QFT becomes a well-defined multiple integral which can be evaluated via Monte Carlo

$$\langle O \rangle = \frac{1}{Z}\int \prod_i dA_i \prod_j d\psi_j \prod_k d\bar{\psi}_k \, O(A,\psi,\bar{\psi}) e^{-S_E}$$

Gluon fields on the Lattice

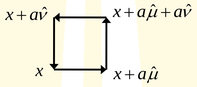
The $SU(3)$ color gluon field $A_\mu(x)$ are defined on each link connecting x and $x+a\hat{\mu}$, through the link variable

$$U_\mu(x) = \exp\left[iagA_\mu\left(x + \frac{a}{2}\hat{\mu}\right)\right]$$

Then the gluon action on the lattice can be written as

$$S_g[U] = \frac{6}{g^2}\sum_{\text{plaquette}}\left[1 - \frac{1}{3}\text{Re}\, tr(U_p)\right] \xrightarrow{a \to 0} \int d^4x \frac{1}{2} tr\left[F_{\mu\nu}(x) F_{\mu\nu}(x)\right]$$

where $U_p = U_\mu(x) U_\nu(x+a\hat{\mu}) U_\mu^\dagger(x+a\hat{\nu}) U_\nu^\dagger(x)$



Lattice QCD

The QCD action $S = S_G(U) + \bar{\psi} D(U) \psi$

where $S_G(U)$ is the action of the gluon fields

$$\bar{\psi} D(U) \psi \equiv \bar{\psi}^f_{\alpha\alpha x} D_f(U)_{\alpha\alpha x, b\beta y} \psi^f_{b\beta y}$$

$f = u,d,s,c,b,t$	flavor index
$a,b = 1,2,3$	color index
$\alpha,\beta = 1,2,3,4$	Dirac index
$x,y = 1,\cdots,N_{\text{sites}} = N_x N_y N_z N_t$	site index

For example, on the $16^3 \times 32$ lattice, D is a complex matrix of size $1{,}572{,}864 \times 1{,}572{,}864$

$$\langle \mathcal{O}(\bar{\psi},\psi,U) \rangle = \frac{\int dU d\bar{\psi} d\psi \mathcal{O}(\bar{\psi},\psi,U) e^{-S}}{\int dU d\bar{\psi} d\psi e^{-S}} = \frac{\int dU \Theta(D^{-1},U) \det(D) e^{-S_G}}{\int dU \det(D) e^{-S_G}}$$

The Challenge of Lattice QCD

- To have lattice volume large enough such that $m_\pi L \gg 1$.
  So far, the lightest u/d quark cannot be put on the lattice.
  Rely on ChPT to extrapolate lattice results to physical ones.
- To have lattice spacing small enough such that $m_q a \ll 1$.
- To meet the above two conditions:
  The lattice size should be at least $100^3 \times 200$.
  The computing power should be at least Petaflops × year

The state-of-the-art in the simulations of unquenched LQCD with exact chiral symmetry

- **RBC and UKQCD Collaborations** (~20-25 members)
  QCDOC, ~(10+10) Tflops(peak), **~(2+2) Tflops(sustained)**
  lattice fermion: **Domain-Wall Fermion**
  lattices: $16^3 \times 32 \times 16$, $24^3 \times 64 \times 16$, $32^3 \times 64 \times 16$
- **JLQCD Collaboration** (~10-15 members)
  IBM BlueGene/L, ~ 57 Tflops (peak), **~ 8 Tflops (sustained)**
  lattice fermion: **Overlap Fermion (with fixed topology)**
  lattices: $16^3 \times 32$, $16^3 \times 48$, $24^3 \times 48$
- **TWQCD Collaboration** (~7-10 members)
  GPU cluster, ~ 120 Tflops (peak), **~ 14 Tflops(sustained)**
  lattice fermion: **Optimal Domain-Wall Fermion**
  lattices: $16^3 \times (32,10,8,6,4) \times (16,32)$

The Hardware of TWQCD

- **16 units of Nvidia Tesla S1070** (total **64 GPUs, 64 x 4 GB**) connected to **16 servers** (total **32 Intel QC Xeon, 16 x 32 GB**)
- **56 Nvidia GTX285** (total **56 GPUs, 56 x 1 GB**), connected to **40 servers** (total **40 Intel i7, 40 x 12 GB**)
- Hard disk storage > **300 TB** (**Lustre** cluster file system)
- Peak performance is **120 Tflops**
- Developed efficient **CUDA** codes for unquenched lattice QCD.
- Attained **14 Tflops (sustained) with a price $200,000.**

Optimal Domain-Wall Fermion

[T.W. Chiu, Phys. Rev. Lett. 90 (2003) 071601]

$$A_{\text{odwf}} = \sum_{s,s'=1}^{N_s}\sum_{x,x'} \bar{\psi}_{x,s}\left[(I + \omega_s D_w)_{x,x'}\delta_{s,s'} - (I - \omega_s D_w)_{x,x'}\left(P_-\delta_{s',s+1} + P_+\delta_{s',s-1}\right)\right]\psi_{x',s'}$$
$$\equiv \bar{\Psi} D_{\text{odwf}} \Psi$$

$$D_w = \sum_{\mu=1}^{4}\gamma_\mu t_\mu + W - m_0, \quad m_0 \in (0,2)$$

$$t_\mu(x,x') = \frac{1}{2}\left[U_\mu(x)\delta_{x',x+\mu} - U_\mu^\dagger(x')\delta_{x',x-\mu}\right]$$

$$W(x,x') = \sum_{\mu=1}^{4}\frac{1}{2}\left[2\delta_{x,x'} - U_\mu(x)\delta_{x',x+\mu} - U_\mu^\dagger(x')\delta_{x',x-\mu}\right]$$

boundary conditions $\quad P_+\psi(x,0) = -\frac{m_q}{2m_0}P_+\psi(x,N_s)$

$$P_-\psi(x,N_s+1) = -\frac{m_q}{2m_0}P_-\psi(x,1)$$

The weights $\{\omega_s\}$ are fixed such that the effective 4D Dirac operator possesses the optimal chiral symmetry,

$$\omega_s = \frac{1}{\lambda_{\min}}\sqrt{1 - \kappa'^2 sn^2(v_s;\kappa')}, \quad s = 1,\cdots,N_s$$

$sn(v_s;\kappa')$ is the Jacobian elliptic function with argument v_s and modulus $\kappa' = \sqrt{1 - \lambda_{\min}^2/\lambda_{\max}^2}$

$\lambda_{\min}^2$ and $\lambda_{\max}^2$ are lower and upper bounds of the eigenvalues of H_w^2

Hybrid Monte Carlo (HMC) for 2 flavor QCD

1. Initial gauge configuration $\{U_l\}$
2. Generate $\{P_l^a\}$ with probability distribution $\propto \exp[-(P_l^a)^2/2]$
3. Generate ξ with probability distribution $\propto \exp(-\xi^\dagger\xi)$
   Recall: $\exp[-\phi^\dagger C_{PV}^\dagger (CC^\dagger)^{-1} C_{PV}\phi] = \exp[-\xi^\dagger\xi]$
4. Fixing the pseudofermion field $\phi = C_{PV}^{-1} C\xi \equiv D\xi$
5. Molecular dynamics (Omelyan integrator with multiple-time scale)
   $\eta(\tau) = \left(DD^\dagger(U(\tau))\right)^{-1}\phi$ ← the most expensive part of HMC
   $\dot{U}_l(\tau) = iP_l(\tau)U_l(\tau), \quad P_l(\tau) \equiv P_l^a(\tau)T^a$
   $\dot{P}_l^a(\tau) = -D_l^a\left[A_{\text{gauge}}(U(\tau))\right] + \eta^\dagger(\tau)D_l^a\left[DD^\dagger(U(\tau))\right]\eta(\tau)$
6. Accept $\{U_l'\}$ with the probability $P_A = \min[1, \exp(-H' + H)]$
7. Go to 2.

CG Algorithm with Mixed Precision

$Ax = b \Leftrightarrow (CC^\dagger)\eta' = \phi'$

1. $r_k = b - Ax_k$
2. If $|r_k| < \varepsilon |b|$, then stop
3. Solve $At_k = r_k$ in single precision to an accuracy $\varepsilon_1 < 1$
4. $x_{k+1} = x_k + t_k$
5. Go to 1.

Features of CG Kernel for ODWF

$$(CC^\dagger)\eta' = \phi'$$

$$C \equiv I - D_w^{oe} Y X^{-1} D_w^{eo} Y X^{-1}$$

One thread takes care of all computations at each (s,x,y,z) with a loop over all t (even/odd).

- 1-dim Grid $\quad Nthread_X \times Nthread_Y \times Nblock = N_s N_x^3$
- 2-dim Block $\quad Nthread_X = N_s = 16 \times (1,2,...)$
  $Nthread_X \times Nthread_Y = 64$

Tuning the CG Kernel for ODWF

- $M \equiv YX^{-1}$ in constant memory space
- Use texture memory for link variables and vectors
- Reorder data in the device memory (coalescing; Ns threads share the same link variables)
- Reuse forward/backward data (in t) for neighboring sites
- Unroll short loops
- …

Attaining 140/120/100 Gflops for GTX285/GTX280/T10

Conclusions and Outlook

- **GPU is a revolutionary device for lattice QCD.** Using a GPU cluster consisting of 120 GPUs, the TWQCD Collaboration in Taiwan has emerged as one of the most computationally powerful lattice QCD groups around the world, achieving 14 Tflops (sustained) with a price $200,000.
- We have shown that the Optimal Domain-Wall Fermion provides a viable framework to simulate unquenched QCD with exact chiral symmetry.
- Current production runs for 2-flavor QCD (T=0, and T > 0) on $16^3 \times (32,\ 10,\ 8,\ 6,\ 4) \times (16,\ 32)$ lattices will be completed before the end of 2009, and these gauge configurations are expected to yield interesting physical results.
- **GPU has created a sensation in the lattice QCD community.** Currently, there are many lattice QCD groups around the world building large GPU clusters dedicated to QCD. This will have significant impacts to lattice QCD, leading to new discoveries in the strong interaction physics.