Probing the QCD Vacuum with Lattice Simulations on a GPU Cluster

Ting-Wai Chiu (趙挺偉)
Physics Department
National Taiwan University

2012 Cross Strait Meeting on Particle Physics and Cosmology
May 7-12, Chong-Qing, China.
Quantum Chromodynamics (QCD)

• The quantum field theory for the strong interaction between quarks and gluons which build up the hadrons (e.g., neutron, proton, pion, etc.).

• QCD provides the framework to understand the nuclear force/energy from the first principles.

• QCD plays an important role in the evolution of the early universe, from the quark gluon "plasma" phase to the hadron phase.
Salient features:

- Gauge group $SU(3) \Rightarrow$ gluons have self-interactions.
- Asymptotic freedom: $g(r) \to 0$ as $r \to 0$.
- IR slavery: at $r \approx 1 \text{ fm}$, $g(r) \approx 1 \Rightarrow$ quark (color) confinement (Nonperturbative !)
- Spontaneously chiral symmetry breaking (Nonperturbative !)
- At high temperature (in the early universe), the chiral sym. is restored, and quarks and gluons become deconfined, and form the so-called quark-gluon "plasma".
To solve QCD is a grand challenge among all sciences. The most promising approach to solve QCD nonperturbatively is to discretize the continuum space-time into a 4 dimensional lattice (lattice QCD), and to compute physical observables by Monte Carlo simulation.

\[
\Delta x = a \cdot 1.2 \times 10^{-16} \text{ m}
\]

\[
L^3 \equiv (2 \times 10^{-15} \text{ m})^3
\]

\[
\Delta t \equiv 3.3 \times 10^{-24} \text{ sec}
\]

\[
Q_t = -1
\]

\[
16^3 \times 32
\]
➢ It took 23 years (1974 ~1997) to realize that **Lattice QCD with Exact Chiral Symmetry** is the ideal theoretical framework to study the nonperturbative physics from the first principles of QCD.

➢ **It is challenging to perform the HMC simulation** such that the chiral sym. is preserved to very high precision and all topological sectors are sampled ergodically.

➢ Since 2009, the TWQCD collaboration has been using a **GPU cluster** to simulate lattice QCD with **optimal domain-wall quarks**. The chiral sym. is preserved to a good precision with $m_{res}a \approx 0.0004$, and all topological sectors are sampled ergodically.
A graphic card (e.g., Nvidia GTX580) is capable to deliver > 300 Gflops (sustained) with the price less than US$400. It gives a speed up 10x –100x comparing with a single CPU.

- This opens up a great opportunity for many scientific and engineering problems which require enormous amount of number-crunching power.

- Recall that in the past 50 years, each 10x jump in computing power motivated new ways of computing, which in turn led to many scientific breakthroughs.

T.W. Chiu, May 7, 2012
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer/Year Vendor</th>
<th>Cores</th>
<th>$R_{\text{max}}$</th>
<th>$R_{\text{peak}}$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RIKEN Advanced Institute for Computational Science (AICS) Japan</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu</td>
<td>705024</td>
<td>10510.00</td>
<td>11280.38</td>
<td>12659.9</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Center in Tianjin China</td>
<td>NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT</td>
<td>186368</td>
<td>2566.00</td>
<td>4701.00</td>
<td>4040.0</td>
</tr>
<tr>
<td>3</td>
<td>DOE/SC/Oak Ridge National Laboratory United States</td>
<td>Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.</td>
<td>224162</td>
<td>1759.00</td>
<td>2331.00</td>
<td>6950.0</td>
</tr>
<tr>
<td>4</td>
<td>National Supercomputing Centre in Shenzhen (NSCS) China</td>
<td>Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning</td>
<td>120640</td>
<td>1271.00</td>
<td>2984.30</td>
<td>2580.0</td>
</tr>
<tr>
<td>5</td>
<td>GSIC Center, Tokyo Institute of Technology Japan</td>
<td>HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP</td>
<td>73278</td>
<td>1192.00</td>
<td>2287.63</td>
<td>1398.6</td>
</tr>
</tbody>
</table>

GPU supercomputing is a viable way to realize EXAFLOPS by 2020?
QCD GPU Cluster at NTU

- 86 [M2090/C2070/C2050/GTX580 (Fermi)]
- 96 [S1070/C1060 (Tesla)] + 120 [GTX285]
- QDR InfiniBand switch (M2090/C2070)
- Lustre cluster file system > 300 TB

Total 300 NVIDIA GPUs

- Peak Performance: 325 TFLOP/s

- Efficient CUDA codes for lattice QCD
  - 320 / 250 / 156 / 180 / 132 GFLOP/s
  on GTX580 / GTX480 / C2070 / GTX285 / C1060

- Sustained Performance: 85 TFLOP/s

Nvidia CEO Jensen Huang visited NTU on Nov 19, 2009
Outline

- Introduction
- Lattice QCD with Exact Chiral Symmetry
- Lattice QCD in GPU
- Physical Results
  (i) Topological Susceptibility
  (ii) Pseudoscalar Mass & Decay Constant
- Conclusions and Outlooks
The QCD action

\[ S = S_G(U) + \bar{\psi} D(U) \psi \]

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

\[ \bar{\psi} D(U) \psi \equiv \bar{\psi}^f_{\alpha x} D^f_{\alpha x} (U)_{\alpha x, \beta y} \psi^f_{\beta y} \]

\( f = u, d, s, c, \ldots \) flavor index

\( a, b = 1, 2, 3 \) color index

\( \alpha, \beta = 1, 2, 3, 4 \) Dirac index

\( x, y = 1, \ldots, N_{\text{sites}} = N_x N_y N_z N_t \) site index

For example, on the \( 16^3 \times 32 \) lattice, for each flavor, \( D \) is a complex matrix of size \( 1,572,864 \times 1,572,864 \)

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

$$U_\mu (x) = \exp \left[ i a g A_\mu \left( x + \frac{a}{2} \hat{\mu} \right) \right] , \quad 3 \times 3 \text{ unitary matrix}$$

Then the action of gluon fields on the lattice can be written as

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

where

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

T.W. Chiu, May 7, 2012
Any gauge covariant Dirac operator \( D \) on the lattice must violate at least one of following properties:

- **Chiral symmetry** \( D\gamma_5 + \gamma_5 D = 0 \)
- **Locality** \( \|D(x, y)\| \leq \exp(-|x - y|/l) \) with \( l \ll a \);
  or \( D(x, y) = 0 \) for \( |x - y| > z \), where \( z \ll L \)
- **Free of species doublings.**
  The free fermion propagator \( D^{-1}(p) \) has only one simple pole at \( p = 0 \) in the Brillouin zone
- **Correct continuum behavior at \( p = 0 \).**
  In the free fermion limit and \( a \to 0 \), \( D(p) \square i\gamma_\mu p_\mu \) around \( p = 0 \).
Wilson Quark Matrix (1975)

\[
[D_W(x, y)]_{ab}^{\alpha\beta} = \sum_{\mu=1}^{4} \gamma_{\mu}^{\alpha\beta} t_{ab}^{\mu}(x, y) + \delta_{\alpha\beta} W_{ab}(x, y) - m_0 \delta_{xy} \delta_{ab} \delta_{\alpha\beta},
\]

\[
t_{ab}^{\mu}(x, y) = \frac{1}{2} \left( [U_{\mu}(x)]_{ab} \delta_{y, x+\hat{\mu}} - [U_{\mu}^\dagger(y)]_{ab} \delta_{y, x-\hat{\mu}} \right),
\]

\[
W_{ab}(x, y) = \sum_{\mu} \frac{1}{2} \left( 2 \delta_{ab} \delta_{y, x} - [U_{\mu}(x)]_{ab} \delta_{y, x+\hat{\mu}} - [U_{\mu}^\dagger(y)]_{ab} \delta_{y, x-\hat{\mu}} \right),
\]

\(U_{\mu}(x)\) is the link variable \((x, x + \hat{\mu})\), \(3 \times 3\) special unitary matrix, \(a, b = 1, 2, 3\) color indices, \(\alpha, \beta = 1, 2, 3, 4\) Dirac indices, The Wilson term \(W_{ab}(x, y)\) breaks the chiral symmetry explicitly!
The proper way to break the chiral symmetry at finite lattice spacing is to impose the Ginsparg-Wilson relation (1982)

\[ D\gamma_5 + \gamma_5 D = D\gamma_5 D \]

equivalently,

\[ D^{-1}(x, y)\gamma_5 + \gamma_5 D^{-1}(x, y) = \gamma_5 \delta_{x,y} \]

which is realized by the Domain-Wall Fermion (Kaplan, 1992), and the overlap Dirac operator (Neuberger, 1998)

\[ D = \left( I + \gamma_5 \frac{H}{\sqrt{H^2}} \right), \quad H^\dagger = H, \]

\( D \) is exponentially local for sufficiently smooth gauge field. In the continuum limit \( a \to 0, \ D \to \gamma^\mu (\partial_\mu + igA_\mu) \).
Current status in the simulations of unquenched QCD with exact chiral symmetry

- **RBC and UKQCD Collaborations**
  Machine: QCDOC, IBM BlueGene/P
  Lattice fermion: Domain-Wall Fermion
  Lattice sizes: $16^3 \times 32 \times 16, 24^3 \times 48 \times 16, 32^3 \times 64 \times 16$

- **JLQCD Collaboration**
  Machine: IBM BlueGene/L
  Lattice fermion: Overlap Fermion (with fixed topology $Q_t=0$)
  Lattice sizes: $16^3 \times 32, 16^3 \times 48,$

- **TWQCD Collaboration**
  Machine: GPU cluster (300 GPUs)
  Lattice fermion: Optimal Domain-Wall Fermion (ODWF)
  Lattice sizes: $16^3 \times 32 \times 16, 20^3 \times 40 \times 16, 24^3 \times 48 \times 16$
Current status in the simulations of unquenched QCD with exact chiral symmetry

- RBC and UKQCD Collaborations
  - Machine: QCDOC, IBM BlueGene/P → IBM BlueGene/Q
  - Lattice fermion: Domain-Wall Fermion → ?
  - Lattice sizes: $16^3 \times 32 \times 16, 24^3 \times 48 \times 16, 32^3 \times 64 \times 16 \rightarrow$ larger lattices

- JLQCD Collaboration
  - Machine: IBM BlueGene/L → IBM BlueGene/Q
  - Lattice fermion: Overlap Fermion (with fixed topology $Q_t=0$) → ?
  - Lattice sizes: $16^3 \times 32, 16^3 \times 48 \rightarrow$ larger lattices

- TWQCD Collaboration
  - Machine: GPU cluster (300 GPUs)
  - Lattice fermion: Optimal Domain-Wall Fermion (ODWF)
  - Lattice sizes: $16^3 \times 32 \times 16, 20^3 \times 40 \times 16, 24^3 \times 48 \times 16 \rightarrow$ larger lattices
Central Problems in Lattice QCD

- To simulate full QCD with dynamical quarks, $\det(D)$
- To compute the (all-to-all) quark propagator, $D^{-1}$
- To compute the (low-lying) eigenmodes of $D$

The matrix $D$ is prohibitively large for exact solvers. Iterative algorithms involve the matrix-vector multiplication

$$\frac{H}{\sqrt{H^2}} \cdot Y$$

The inverse square-root cannot be computed exactly.
Nested Conjugate Gradient

To compute quark propagator requires nested CG

\[ D \cdot Y \equiv \left( I + \gamma_5 \frac{H}{\sqrt{H^2}} \right) \cdot Y = |b\rangle \]

\[ \frac{H}{\sqrt{H^2}} \cdot Y = HR_{Z}^{(n-1,n)}(H^2) \cdot Y \]

\[ R^{(n-1,n)}(H^2) \cdot Y = \sum_{l=1}^{n} \frac{b_l}{H^2 + d_l} Y = \sum_{l=1}^{n} b_l Z^{(l)} \]

\[ \left( H^2 + d_l \right) Z^{(l)} = Y \quad \text{solved by CG with multi-shifts} \]

and the 2-pass algorithm
Difficulties of HMC with Overlap Fermion

- The action is discontinuous at the boundary between different topological sectors.

- It is very costly to tunnel through the topological boundary using the refraction-reflection algorithm (Fodor, Katz, Szabo, JHEP 2004)
\(D_{\text{dwf}}\) is a local op. with the nearest neighbor coupling along \(\hat{s}\)

\[
\int [d\bar{\psi}][d\psi] \exp(-\bar{\psi}D_{\text{dwf}}\psi) = \det D_c \quad D_c = \frac{1 + \gamma_5 S}{1 - \gamma_5 S}
\]

\(N_s \to \infty, \quad S \to \frac{H}{\sqrt{H^2}}, \quad D_c\gamma_5 + \gamma_5 D_c = 0, \quad \text{Exact Chiral Sym.}\)

But at finite \(N_s\), \(S\) is NOT equal to the optimal rational approx.
Optimal Domain-Wall Fermion

\[ A_{\text{odwf}} = \sum_{s,s'=1}^{N_s} \sum_{x,x'} \overline{\psi}_{x,s} \left[ (I + \rho_s D_w)_{x,x'} \delta_{s,s'} - (I - \sigma_s D_w)_{x,x'} \left( P_{-\delta_{s',s+1}} + P_{+\delta_{s',s-1}} \right) \right] \psi_{x',s'} \]

\[ \equiv \overline{\Psi} D_{\text{odwf}} \Psi \]

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

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

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

with boundary conditions

\[ P_+ \psi(x,0) = -mP_+ \psi(x,N_s), \quad m \propto m_q \text{ (bare quark mass)} \]

\[ P_- \psi(x,N_s+1) = -mP_- \psi(x,1), \quad P_{\pm} = \frac{1}{2} (1 \pm \gamma_5) \]
Optimal Domain-Wall Fermion (cont.)

The action for Pauli-Villars fields is similar to $A_{odwf}$

$$A_{PV} = \sum_{s,s'=1}^{N_s} \sum_{x,x'} \phi_{x,s} \left[ (I + \rho_s D_w)_{x,x'} \delta_{s,s'} - (I - \sigma_s D_w)_{x,x'} \left( P_{-\delta_{s',s+1}} + P_{+\delta_{s',s-1}} \right) \right] \phi_{x',s'}$$

but with boundary conditions:

$$P_+ \phi(x,0) = -P_+ \phi(x,N_s),$$

$$P_- \phi(x,N_s + 1) = -P_- \phi(x,1)$$

➢ For optimal chiral symmetry, $\rho_s = \sigma_s = \omega_s$

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

where $\text{sn}(v_s;\kappa')$ is the Jacobian elliptic function with argument $v_s$ and modulus $\kappa' = \sqrt{1 - \frac{\lambda_{\min}^2}{\lambda_{\max}^2}}$, $\lambda_{\min}$ and $\lambda_{\max}$ are lower and upper bounds of the eigenvalues of $H_w^2$.
Optimal Domain-Wall Fermion (cont.)

\[ \int [d\bar{\psi}] [d\psi] [d\phi] [d\phi] \exp(-A_{odwf} - A_{PV}) = \det D(m_q) \]

The effective 4D Dirac operator

\[ D(m_q) = m_q + \left( m_0 - m_q / 2 \right) \left[ 1 + \gamma_5 S_{opt}(H_w) \right] \]

\[ S_{opt}(H_w) = \frac{1 - \prod_{s=1}^{N_s} T_s}{1 + \prod_{s=1}^{N_s} T_s}, \quad T_s = \frac{1 - \omega_s H_w}{1 + \omega_s H_w} \]

\[ = \begin{cases} 
H_w R_Z^{(n-1,n)} (H_w^2), & N_s = 2n \\
H_w R_Z^{(n,n)} (H_w^2), & N_s = 2n + 1 
\end{cases} \]

Zolotarev optimal rational approximation for \( \frac{1}{\sqrt{H_w^2}} \)
Lattice setup for 2-flavor QCD with ODWF

- Lattice Size: $16^3 \times 32 \times 16$
- Quark Action: Optimal Domain-Wall Fermion (ODWF)
- Gluon Action: Plaquette (beta = 5.95)
- Lattice Spacing: $a = 0.1032(2) [\text{fm}]$, $1/a = 1.911(4) [\text{GeV}]$
- Lattice Volume: $\sim (1.7 \ \text{fm})^3$
- 8 sea quark masses, with pion masses 230 – 580 MeV.

Each mass has $30 \times 400$ traj. After discarding $30 \times 300$ traj. for thermalization, measurements are performed every 10 traj., with a total of $30 \times 10 = 300$ confs.

For each conf, zero modes plus 80+80 conjugate pairs of low-lying eigenmodes of the overlap operator are projected.
First Physical Results

- The topology of the QCD vacuum:
  To determine its topological structure and fluctuations, and their relationship with the spontaneous chiral symmetry breaking, and the color (de)confinement.

- To compute the pseudoscalar meson mass and decay constant, and to check whether their sea-quark mass dependence agree with the ChPT.

- To determine the low-energy constants of ChPT, $\Sigma, F, \bar{l}_3, \bar{l}_4$

- To determine the u/d quark mass.
Theoretically, topological susceptibility is defined as

$$\chi_t = \int d^4 x \langle \rho(x) \rho(0) \rangle, \quad \rho(x) = \frac{1}{32\pi^2} \varepsilon_{\mu\nu\lambda\sigma} \text{tr} \left[ F_{\mu\nu}(x) F_{\lambda\sigma}(x) \right]$$

Topological susceptibility is the most crucial quantity to measure the topological charge fluctuation of the QCD vacuum.

$$\chi_t = \frac{1}{\Omega} \langle Q_t^2 \rangle, \quad Q_t = \int_\Omega d^4 x \rho(x) = \text{integer}$$

However, on a lattice, it is difficult to extract $\rho(x)$ unambiguously from the link variables (gauge fields)!
We turn to the Atiyah-Singer index theorem

\[ Q_t \equiv \int d^4x \rho(x) = \text{index}(D) = n_+ - n_- \]

where \( n_\pm \) is the number of zero modes of \( D \) with \( \pm \) chirality. Thus, to preserve exact chiral symmetry on the lattice is vital for studying the topology of the QCD vacuum, which is relevant to many important nonperturbative physics.
Chiral Perturbation Theory (ChPT) for Topological Susceptibility

ChPT - an effective field theory of QCD for the low-energy physics of the (pseudo-)Nambu–Goldstone bosons.

- LO (tree-level) ChPT

\[ \chi_t = \frac{\langle Q_t^2 \rangle}{V} = \bar{m} \Sigma + O(m_i^2), \]
\[ \frac{1}{\bar{m}} = \sum_{i=1}^{N_f} \frac{1}{m_i} \]

Leutwyler-Smilga relation (1992)

\[ \chi_t = \sum \left( \frac{1}{m_u} + \frac{1}{m_d} \right)^{-1} \]
\[ N_f = 2 \]

\[ \chi_t = \frac{\Sigma m_q}{2}, \quad N_f = 2 \quad \text{(isospin limit)} \]
ChPT for Topological Susceptibility (cont)

- NLO ChPT [Y. Mao, TWC, PRD 80, 034502 (2009)]

A general formula of $\chi_t$ for any $N_f$ has been derived.

$N_f = 2$ in the isospin limit $m_u = m_d$

$$\frac{\chi_t}{m_q} = \sum \left\{ 1 - 3 \left( \frac{\sum m_q}{16\pi^2 F_\pi^4} \right) \ln \left( \frac{2\sum m_q}{F_\pi^2 \mu^2_{\text{sub}}} \right) + 32 \left( \frac{\sum}{F_\pi^4} \right) (2L_6 + 2L_7 + L_8) m_q \right\}$$

$L_i$ are renormalized low-energy coupling constants defined at $\mu_{\text{sub}}$

$\mu_{\text{sub}} = 770$ MeV
Topo. Susceptibility of 2-flavor QCD with ODWF


\[ \chi_t = \frac{\langle Q_t^2 \rangle}{V} \]

fitting to LO ChPT

\[ \chi_t = \frac{\sum m_q}{2} \Rightarrow \Sigma^{\overline{\text{MS}}} (2 \text{ GeV}) = [259(6)(7) \text{ MeV}]^3 \]

[ Leutwyler-Smilga (1992) ]
Topological Susceptibility of 2-flavor $QCD$ (cont)

TWC, Hsieh, Mao [TWQCD Collaboration], PLB 702 (2011) 131

fitting to NLO ChPT of $\chi_t$ \(\Rightarrow F_\pi = 92(12)(2) \text{ MeV} \)

[ Mao, TWC, PRD (2009) ]

\[ \Sigma_{\overline{MS}}^{\overline{MS}} (2 \text{ GeV}) = [259(6)(7) \text{ MeV}]^3 \]

\[ 2L_6 + 2L_7 + L_8 = -0.0001(3) \]
Pseudoscalar Meson

\[
\langle 0 | \pi^- (\bar{x}, t) \pi^+ (\bar{0}, 0) | 0 \rangle = -\langle 0 | (\bar{u} \gamma_5 d)(\bar{x}, t)(\bar{d} \gamma_5 u)(\bar{0}, 0) | 0 \rangle = \text{tr} \left[ (D_c + m_u)^{-1}_{0,x} \gamma_5 (D_c + m_u)^{-1}_{x,0} \gamma_5 \right]
\]

Fitting \( C_\pi (t) = \sum_\bar{x} \langle 0 | \pi^- (\bar{x}, t) \pi^+ (\bar{0}, 0) | 0 \rangle \) to

\[
\frac{\left| \langle \pi^+ (\bar{p} = 0) | \pi^- (\bar{0}, 0) | 0 \rangle \right|^2}{2M_\pi} \left( e^{-M_\pi t} + e^{-M_\pi (T-t)} \right) + \text{excited states}
\]

to extract \( M_\pi \) and

\[
F_\pi = \frac{(m_u + m_d)}{\sqrt{2M_\pi^2}} \left| \langle \pi^+ (\bar{p} = 0) | \pi^- (\bar{0}, 0) | 0 \rangle \right|
\]
Pion in 2-flavors QCD with ODWF

TWC, Hsieh, Mao [TWQCD Collaboration], arXiv: 1109.3675

NLO ChPT
(Gasser & Leutwyler, 1985)

\[
\frac{M_{\pi}^2}{m_q} = 2B \left( 1 + \frac{B m_q}{(4\pi F)^2} \ln \frac{2B m_q}{\Lambda_3^2} \right),
\]

\[
B \equiv \frac{\Sigma}{F^2}
\]

\[
F_{\pi} = F \left( 1 - \frac{2B m_q}{(4\pi F)^2} \ln \frac{2B m_q}{\Lambda_4^2} \right)
\]

\[
\bar{l}_3 \equiv \ln \frac{\Lambda_3^2}{m_{\pi^\pm}^2}, \quad \bar{l}_4 \equiv \ln \frac{\Lambda_4^2}{m_{\pi^\pm}^2},
\]

\[
m_{\pi^\pm} = 0.14 \text{ GeV}
\]

230 MeV < \( M_{\pi} \) < 580 MeV
Simultaneous fit of 8 pairs of \((M_\pi, F_\pi)\) to NLO ChPT, with correlation between \(M_\pi\) and \(F_\pi\) at the same \(m_q\), we obtain

\[
F = 0.08339(35)(38) \text{ GeV}
\]

\[
\sum^{\overline{\text{MS}}} (2 \text{ GeV}) = \left[ 235(8)(4) \text{ MeV} \right]^3
\]

\[
\overline{l}_3 = 4.149(35)(14)
\]

\[
\overline{l}_4 = 4.582(17)(20)
\]
Pion in 2-flavors QCD with ODWF (cont)

With the fitted parameters, we use the NLO ChPT formulas to solve for the physical (bare) quark mass.

\[
\frac{M_{\pi}(m_q)}{F_\pi(m_q)} = \frac{0.135 \text{ GeV}}{0.093 \text{ GeV}} = 1.45 \implies m_q^{\text{phys}}(\text{bare}) = 0.00505(13) \text{ GeV}
\]

At the physical point, the NLO ChPT formulas give

\[
F_\pi = 0.090(4)(2) \text{ GeV}
\]

\[
m_{ud}^{\text{MS}}(2 \text{ GeV}) = 4.06(10)(12) \text{ MeV}
\]

\[
M_\pi = 0.130(5)(3) \text{ GeV}
\]
Conclusions and Outlooks

- Pion mass and decay constant in 2-flavors QCD with ODWF are in good agreement with the sea-quark mass dependence predicted by NLO ChPT, and provide the first principles determination of the following physical quantities:

\[ F_\pi = 90(4)(2) \text{ MeV} \]
\[ M_\pi = 0.130(5)(3) \text{ GeV} \]
\[ m_{ud}^{\overline{\text{MS}}} (2 \text{ GeV}) = 4.06(10)(12) \text{ MeV} \]
\[ \Sigma^{\overline{\text{MS}}} (2 \text{ GeV}) = \left[ 235(8)(4) \text{ MeV} \right]^3 \]
The topological susceptibility in 2-flavors QCD with ODWF is in good agreement with sea-quark mass dependence predicted by NLO ChPT, and provide the first principles determination of

\[ F_\pi = 92(12)(2) \text{ MeV} \]

\[ \Sigma^{\text{MS}} (2 \text{ GeV}) = [259(6)(7) \text{ MeV}]^3 \]

These results imply that the nonperturbative chiral dynamics of the sea quarks are well under control in the TWQCD’s HMC simulations with ODWF.
Conclusions and Outlook (cont.)

- ODWF provides a viable framework for the simulation of QCD, which not only preserves the chiral symmetry to a good precision, but also samples all topological sectors ergodically.

- For 2-flavors QCD on the $16^3 \times 32$ lattices, we have completed two ensembles, $\beta = 5.90$, and $\beta = 5.95$. For each ensemble, 8 sea quark masses, each of 5000 trajectories after thermalization.

- Currently, we are simulating 2-flavors QCD on the $20^3 \times 40$, and $24^3 \times 48$ lattices, and also finite temperature QCD on $24^3 \times 8$ lattices.
Backup slides
Quarks

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

\[
\begin{align*}
\text{u} & \quad \text{c} & \quad \text{t} & \quad \text{u} & \quad \text{c} & \quad \text{t} & \quad \text{u} & \quad \text{c} & \quad \text{t} \\
\text{d} & \quad \text{s} & \quad \text{b} & \quad \text{d} & \quad \text{s} & \quad \text{b} & \quad \text{d} & \quad \text{s} & \quad \text{b}
\end{align*}
\]

Hadrons are color singlets composed of quarks

\[
P = uud + \text{antisym. in color, Proton}
\]
\[
N = udd + \text{antisym. in color, Neutron}
\]
\[
\pi^+ = \bar{d}u + \bar{u}d + \bar{d}u, \quad \text{Pion}
\]

The nuclear force between nucleons emerges as residual interactions of QCD
Some important relations

Veneziano-Witten relation  \[ \chi_t (\text{quenched}) = \frac{f_\pi^2 m_{\eta'}^2}{4N_f} \]

Leutwyler-Smilga relation

\[ \chi_t = \frac{\langle Q_t^2 \rangle}{V} = \sum \left( \frac{1}{m_u} + \frac{1}{m_d} + \frac{1}{m_s} \right) + O(m_u^2), \quad N_f = 2 + 1 \]

Banks-Casher relation

\[ \sum = \pi \rho(0) \]

\( \rho(0) \) is the density of near-zero modes of the massless Dirac operator \( D = \gamma^\mu (\partial_\mu + igA_\mu) \)
ChPT for Topological Susceptibility (cont)

- **NLO ChPT** [Y. Mao, TWC, PRD 80, 034502 (2009)]

\[ N_f = 2 \]

\[ \chi_t = \sum \left( \frac{1}{m_u} + \frac{1}{m_d} \right)^{-1} \left[ 1 - \frac{3}{2F^2_\pi} \frac{M^2_\pi}{16\pi^2} \ln \frac{M^2_\pi}{\mu^2_{sub}} + K_6(m_u + m_d) \right. \]

\[ + 2(2K_7 + K_8) \frac{m_u m_d}{m_u + m_d} \]

\[ N_f = 2 + 1 \]

\[ \chi_t = \sum \bar{m} \left\{ 1 - \frac{1}{2F^2_\pi} \left[ \sum_{i\neq j} \left( \frac{\bar{m}}{m_i} + \frac{\bar{m}}{m_j} \right) \frac{B_0(m_i + m_j)}{16\pi^2} \ln \frac{B_0(m_i + m_j)}{\mu^2_{sub}} \right. \right. \]

\[ + \left( \frac{\bar{m}}{m_u} + \frac{\bar{m}}{m_d} \right) M^2_{\pi_0} \ln \frac{M^2_{\pi_0}}{\mu^2_{sub}} + \frac{1}{3} \left( \frac{\bar{m}}{m_u} + \frac{\bar{m}}{m_d} + 4 \frac{\bar{m}}{m_s} \right) M^2_\eta \ln \frac{M^2_\eta}{\mu^2_{sub}} \right] \]

\[ + K_6(m_u + m_d + m_s) + 3(3K_7 + K_8)\bar{m} \} \]
To convert $\Sigma$ and $m_{ud}$ to the MS scheme, we compute
the renormalization factor $Z_s^{\text{MS}}(2 \text{ GeV})$ using the
nonperturbative renormalization technique through
the RI/MOM scheme.

$$Z_s^{\text{MS}}(2 \text{ GeV}) = 1.244(18)(39)$$

This gives

$$m_{ud}^{\text{MS}}(2 \text{ GeV}) = 4.06(10)(12) \text{ MeV}$$

$$\Sigma^{\text{MS}}(2 \text{ GeV}) = \left[235(8)(4) \text{ MeV}\right]^3$$
Question:
To what extent the physical observables extracted from lattice QCD with exact chiral symmetry agree with the chiral perturbation theory (ChPT) ?

I try to answer this question with the following physical observables:

- Topological susceptibility
- Pion mass and decay constant

in the framework of 2-flavors QCD in the isospin limit $m_u = m_d$
Chiral Perturbation Theory (ChPT)

- An effective field theory for the low-energy physics of the (pseudo-)Nambu–Goldstone bosons of QCD.

- ChPT provides a useful guideline to extrapolate lattice QCD results to the physical regime.

- On the other hand, lattice QCD results can be used for the determination of low-energy constants in ChPT.

- Here we focus on the following physical quantities:
  - Topological susceptibility
  - Pion mass and decay constant
Hybrid Monte Carlo (HMC) for 2 flavor QCD

1. Initial gauge configuration \( \{U_i\} \)
2. Generate \( \{P_i^a\} \) with probability distribution \( \propto \exp[-(P_i^a)^2 / 2] \)
3. Generate \( \xi \) with probability distribution \( \propto \exp(-\xi^\dagger \xi) \)

   Recall: \( \exp[-\phi^\dagger C_{PV} (CC^\dagger)^{-1} C_{PV} \phi] = \exp[ -\xi^\dagger \xi ] \)

4. Fixing the pseudofermion field \( \phi = C_{PV}^{-1} C \xi \equiv D \xi \)
5. Molecular dynamics (Omelyan integrator with multiple-time scale)

\[
\eta(\tau) = \left( DD^\dagger (U(\tau)) \right)^{-1} \phi \quad \leftarrow \text{the most expensive part of HMC}
\]

\[
\dot{U}_i(\tau) = iP_i(\tau)U_i(\tau), \quad P_i(\tau) \equiv P_i^a(\tau)T^a
\]

\[
\dot{P}_i^a(\tau) = -D_i^a \left[ S_G (U(\tau)) \right] + \eta^\dagger(\tau)D_i^a \left[ DD^\dagger (U(\tau)) \right] \eta(\tau)
\]

6. Accept \( \{U'_i\} \) with the probability \( P_A = \min[1, \exp(-H' + H)] \)
7. Go to 2.

\[
D_i^a \left[ f (U) \right] = i \sum_{ij} (T^a U^i_1)_{ij} \frac{\partial f (U)}{\partial (U^i_1)}_{ij}
\]

T.W. Chiu, May 7, 2012
Mixed-Precision CG (1)

High precision

\[ \hat{A} \hat{x} = \hat{b}, \quad \hat{A} = \hat{C} \hat{C}^\dagger \]

\[ \hat{x} := \text{initial guess} \]
\[ \hat{r} := \hat{b} - \hat{A} \hat{x} \]

while \( (|\hat{r}|^2 > \varepsilon) \)

\[ r := \hat{r} \]
\[ p := \hat{r} \]
\[ x := 0 \]

Low-precision CG

\[ \hat{x} := \hat{x} + x \]
\[ \hat{r} := \hat{b} - \hat{A} \hat{x} \]

Low precision

\[ A \hat{x} = r(= \hat{r}), \quad A = CC^\dagger \]

\[ \rho := (r, r) \]

while \( (\beta_0 > \epsilon) \)

\[ v_0 := C^\dagger p \]
\[ \alpha := \rho/(v_0, v_0) \]
\[ r := r - \alpha C v_0 \]
\[ \rho' := \rho \]
\[ \rho := (r, r) \]
\[ x := x + \alpha p \]
\[ p := r + (\rho/\rho') p \]

Single-precision operations are much faster than double-precision ones on GPU
Hybrid Monte Carlo Simulation of Lattice QCD

\[
\langle O(\bar{\psi}, \psi, U) \rangle = \frac{\int dU \Theta(D^{-1}, U) \det(D) e^{-S_G}}{\int dU \det(D) e^{-S_G}} \approx \frac{1}{N_{\text{conf}}} \sum_{i=1}^{N_{\text{conf}}} \Theta(D_i^{-1}, U_i),
\]

provided the probability of sampling \( U \) is proportional to \( \det(D) e^{-S_G} \). However, to compute \( \det(D) \) is very demanding.

To circumvent, introduce pseudofermions (scalar fields) \( \varphi \) and \( \varphi^\dagger \) carrying color and Dirac indices, and rewrite

\[
\det(D) \square \int d\varphi^\dagger d\varphi \exp\left( -\varphi^\dagger \frac{1}{D} \varphi \right),
\]
GPU/CUDA at National Taiwan University

- NTU is the most prominent university in Taiwan.

- The (under)graduate students attain the highest academic performance in the world-wide standard.

- Many renowned scientists/engineers did their undergraduate study at NTU.

- Students: 35,000 = 20,000(57%)(M) + 15,000(43%)(F)

- Faculty: 2,500 = 2,000(80%)(M) + 500(20%)(F)

- Since 2009, NTU is one of the Nvidia CUDA Center of Excellence (CCOE), the first one in Asia-Pacific.
## Nvidia GPUs

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Memory</th>
<th>Memory bandwidth</th>
<th>CUDA cores</th>
<th>Peak (SP)</th>
<th>Peak (DP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTX285</td>
<td>2009</td>
<td>1.0 GB</td>
<td>159 GB/s</td>
<td>240</td>
<td>1063 GF/s</td>
<td>89 GF/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTX480</td>
<td>2010</td>
<td>1.5 GB</td>
<td>177 GB/s</td>
<td>480</td>
<td>1345 GF/s</td>
<td>168 GF/s</td>
</tr>
<tr>
<td>GTX580</td>
<td>2010</td>
<td>1.5 GB</td>
<td>192 GB/s</td>
<td>512</td>
<td>1581 GF/s</td>
<td>198 GF/s</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>3.0 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1060</td>
<td>2008</td>
<td>4 GB</td>
<td>102 GB/s</td>
<td>240</td>
<td>933 GF/s</td>
<td>78 GF/s</td>
</tr>
<tr>
<td>C2050</td>
<td>2010</td>
<td>3 GB/ECC</td>
<td>144 GB/s</td>
<td>448</td>
<td>1030 GF/s</td>
<td>515 GF/s</td>
</tr>
<tr>
<td>C2070</td>
<td>2010</td>
<td>6 GB/ECC</td>
<td>144 GB/s</td>
<td>448</td>
<td>1030 GF/s</td>
<td>515 GF/s</td>
</tr>
<tr>
<td>M2090</td>
<td>2011</td>
<td>6 GB/ECC</td>
<td>177 GB/s</td>
<td>512</td>
<td>1331 GF/s</td>
<td>665 GF/s</td>
</tr>
</tbody>
</table>
A link variable is represented by a 3 × 3 unitary matrix

\[ U^\mu (x, y, z, t) = \begin{pmatrix} U_{11}^\mu & U_{12}^\mu & U_{13}^\mu \\ U_{21}^\mu & U_{22}^\mu & U_{23}^\mu \\ U_{31}^\mu & U_{32}^\mu & U_{33}^\mu \end{pmatrix}, \quad \mu = \hat{x}, \hat{y}, \hat{z}, \hat{t} \]

\[ U_{\mu} U^{\mu\dagger} = U^{\mu\dagger} U^{\mu} = I \]
The Quark Matrix (Lattice Dirac Operator)

\[ D = \left( D_{ij} \right) \]

\[ i = (a, \alpha, x) \quad j = (b, \beta, y) \]

\( a, b = 1, 2, 3 \) (color indices)

\( \alpha, \beta = 1, 2, 3, 4 \) (Dirac spinor indices)

\( x, y = 1, \ldots, N_{\text{sites}} \) (Lattice site indices)
Optimal Rational Approximation for Square Root

In general, for any function, rational approximations are better than polynomial approximations.

For the inverse square root function, the optimal rational approx. was obtained by Zolotarev in 1877.

\[
\frac{1}{\sqrt{x}}, \ x \in [1, b]
\]

\[
R_Z^{(n-1,n)} (x) = \frac{2\lambda}{1 + \lambda} \frac{1}{M} \prod_{l=1}^{n-1} \left(1 + \frac{x}{c_{2l}}\right)
\]

\[
\prod_{l=1}^{n} \left(1 + \frac{x}{c_{2l-1}}\right)
\]

where \( \lambda, M, c_{2l-1}, \) and \( c_{2l} \) are expressed in terms of the Jacobian Elliptic functions.

Yegor Ivanovich Zolotarev (1847 – 1878)
Salient Feature of Optimal Rational Approximation

$$1 - \sqrt{x} R_Z^{(n,m)} (x)$$

Has \((n + m + 2)\) alternate change of sign in \([x_{\text{min}}, x_{\text{max}}]\), and attains its max. and min. (all with equal magnitude)

In the figure, \(n = m = 6\) it has 14 alternate change of sign in \([1,1000]\)
DWF with even-odd preconditioning

\[
\mathcal{D}(m_q) = S_1^{-1} \left( \begin{pmatrix} 1 & M_5 D_w^{EO} \\ M_5 D_w^{OE} & 1 \end{pmatrix} \right) S_2^{-1}
\]

Schur decomposition

\[
\mathcal{D}(m_q) = S_1^{-1} \left( \begin{pmatrix} 1 & 0 \\ M_5 D_w^{OE} & 1 \end{pmatrix} \right) \left( \begin{pmatrix} 1 & 0 \\ 0 & C \end{pmatrix} \right) \left( \begin{pmatrix} 1 & M_5 D_w^{EO} \\ 0 & 1 \end{pmatrix} \right) S_2^{-1}
\]

\[
C \equiv 1 - M_5 D_w^{OE} M_5 D_w^{EO}
\]

For 2-flavors QCD, the pseudofermion action is

\[
A_{PF} = \phi^\dagger C_{PV}^\dagger (CC^\dagger)^{-1} C_{PV} \phi \quad \quad C_{PV} \equiv C(m_q = 2m_0)
\]
Conjugate Gradient Method

Conjugate Gradient is an iterative method for solving the inverse of a sparse positive-definite Hermitian matrix.

\[ Ax = b, \quad A = CC^\dagger \]

\[
\begin{align*}
x_0 & := \text{initial guess} \\
r_0 & := b - Ax \\
p_0 & := r_0
\end{align*}
\]

Iteration to convergence

\[
\alpha_k = \frac{(r_k, r_k)}{(p_k, Ap_k)} = \frac{(r_k, r_k)}{(C^\dagger p_k, C^\dagger p_k)}
\]

\[
r_{k+1} = r_k - \alpha_k Ap_k
\]

\[
\beta_{k+1} = \frac{(r_{k+1}, r_{k+1})}{(r_k, r_k)}
\]

\[
x_{k+1} = x_k + \alpha_k p_k
\]

\[
p_{k+1} = r_{k+1} + \beta_{k+1} p_k
\]

CG is used for calculating fermion force, which is the most time-consuming part in the HMC simulation.
Mixed-Precision CG

1. \( r_k = b - Ax_k, \ A \equiv CC^\dagger \)
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.

Proof:

Let \( u_k = r_k - At_k, \ |u_k| < \varepsilon_1 |r_k| \)

then \( |r_{k+1}| = |b - Ax_{k+1}| = |b - Ax_k - At_k| = |u_k| < \varepsilon_1 |r_k| < |r_k| \)
CUDA Programming Model
Thread/Block Management (1)

Basic ideas of the execution mode:

- Parallelize a loop by designating the value of the loop counter to each thread.
- Try to have as many threads as possible, which means more threads working at the same time.
- All threads inside the same block can access (r/w) the shared memory.
Thread/Block Management (2)

- Number of threads per block
  - Should be tested to find the best value (may be limited by resource in one block)
  - Must be a multiple of half-warp.
- Memory bandwidth bound $\Rightarrow$ try to reuse data.
  - Larger number of blocks does NOT necessarily mean better performance!
  - Using loop inside kernel to reduce the number of blocks sometimes runs faster.
  - For Dw multiplication, loop can further help to reuse one of the hopping data.
Memory Management (1)

Basics of the memory hierarchy:

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Access</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>Large</td>
<td>r/w by all threads and host</td>
<td>Slow</td>
</tr>
<tr>
<td>Constant</td>
<td>Small</td>
<td>Read only by all threads</td>
<td>Fast</td>
</tr>
<tr>
<td>Texture</td>
<td>Small cache</td>
<td>Read only by all threads</td>
<td>Fast</td>
</tr>
<tr>
<td>Shared</td>
<td>Very small</td>
<td>r/w by all threads within one block</td>
<td>Very fast</td>
</tr>
<tr>
<td>Register</td>
<td>Very small</td>
<td>r/w by only one thread</td>
<td>Very fast</td>
</tr>
</tbody>
</table>

- Shared memory may have bank conflicts
- Texture can take care of the locality.
- GPU computing is memory bandwidth bound!
CG Kernels Overview (single-prec.)

\[ C \equiv 1 - M_5 D_w^{OE} M_5 D_w^{EO} \]

The multiplication of M5 and Dw are implemented in different kernels.

\[ v_0 := C^\dagger p \]
\[ \alpha := \rho / (v_0, v_0) \]
\[ r := r - \alpha C v_0 \]
\[ \rho' := \rho \]
\[ \rho := (r, r) \]
\[ x := x + \alpha p \]
\[ p := r + (\rho / \rho') \rho \]

Each line below is implemented in one kernel.

\[ v_1 := M_5^\dagger p \]
\[ v_0 := (D_w^{OE})^\dagger v_1 \]
\[ v_1 := M_5^\dagger v_0 \]
\[ v_0 := p - (D_w^{EO})^\dagger v_1 \]
\[ \alpha := \rho / (v_0, v_0) \]

\[ v_1 := D_w^{EO} v_0, \quad r := r - \alpha v_0 \]
\[ v_0 := M_5 v_1 \]
\[ v_1 := D_w^{OE} v_0 \]
\[ r := r + \alpha M_5 v_1 \]
\[ \rho' := \rho, \quad \rho := (r, r) \]
\[ x := x + \alpha p, \quad p := r + (\rho / \rho') \rho \]

T.W. Chiu, May 7, 2012
CG Kernels Overview (double-prec.)

Each line below is implemented in one kernel.

\[ A \equiv CC^\dagger \]
\[ C \equiv 1 - M_5 D_w^{OE} M_5 D_w^{EO} \]

The multiplication of M5 and Dw are implemented in different kernels.

\[ \hat{x} := \hat{x} + x \]
\[ \hat{r} := \hat{b} - \hat{A}\hat{x} \]

\[ v_1 := M_5^\dagger p \]
\[ v_0 := (D_w^{OE})^\dagger v_1 \]
\[ v_1 := M_5^\dagger v_0 \]
\[ v_0 := (D_w^{EO})^\dagger v_1 \]
\[ v_1 := p - v_0 \]
\[ v_2 := D_w^{EO} v_1 \]
\[ v_0 := M_5 v_2 \]
\[ v_2 := D_w^{OE} v_0 \]
\[ v_1 := v_1 - v_2 \]
\[ r := b - v_1 \]
CG Kernels ($D_w$ multiplication)

$$(D_w^{OE})_{xx'} = -\frac{1}{2} \sum_{\mu} \left[ (1 - \gamma_\mu) U_\mu(x) \delta_{x+a\hat{\mu},x'} + (1 + \gamma_\mu) U_\mu^\dagger(x') \delta_{x-a\hat{\mu},x'} \right]$$

- **Hopping terms**
  - Texture is used for caching data
  - Internal loop is used to reuse read-in data

- **Link variables multiplication**
  - For a given $\mu$, $U$ is the same for all $s$
    - Use shared memory

- **Gamma matrices multiplication**
  - Only left-handed Dirac indices are calculated
Dw multiplication Implementation

From thread/block indices, calculate x, y, z, s

Iteration for t

Iteration for 8 directions

Load vector (texture → register)

Load link variable (texture → shared memory)

Gamma matrices multiplication (only left-handed components)

Link variable multiplication

Restore right-handed components

Write output to global memory

GPU main feature

To reuse data; significant performance enhancement

memory bandwidth bound

All loops here expanded

T.W. Chiu, May 7, 2012
CG Kernels (M5 multiplication)

\[ M_5 = \left[ (d - m_0) + \omega^{-1/2} (1 - L)(1 + cL)^{-1} \omega^{-1/2} \right]^{-1} \]

- Block diagonal in the chiral basis.
- Does not depend on x, y, z, t, or color index.
- It is a constant matrix-vector multiplication in the 5\textsuperscript{th}-dim space.
- Use shared memory for storing source vector

\[ v_{s'} = \sum_s (M_5)_{s's} v_s \]
M5 multiplication Implementation

From thread/block indices, calculate x, y, z, s → GPU main feature

Iteration for t → To reuse data

Load M5 matrix (global → register)

Iteration for 24 Dirac/color/complex indices → loop expanded

Load vector (global → shared memory)

M5 matrix multiplication → loop expanded

Write output to global memory
CG Kernels (More Tunings)

- Try to **reuse data** as much as possible!
- When doing parallel reduction (calculating norm), do partly in the previous kernel:
  
  \[ v_0 := p - (D_w^{EO})^\dagger v_1 \quad \text{Do a “pre-parallel reduction” within each block} \]
  
  \[ \alpha := \rho / (v_0, v_0) \quad \text{Parallel reduction of } v_0 \]

- Addition/subtraction: try to combine these simple operations with existing multiplication kernels, for examples:
  \[ v_1 := D_w^{EO} v_0, \quad r := r - \alpha v_0 \]
Memory Management (2)

- Reorder array indices such that adjacent threads will access adjacent memory spaces.

→ better coalesce!

When $N_s = 4$, for a given point $(x, y, z, t)$:

- $s = 0$
- $s = 1$
- $s = 2$
- $s = 3$

$4 \times 3 \times 2 = 24$ real numbers

Reorder the indices

- 4 real numbers = one float4

Neighboring threads access neighboring memory spaces

T.W. Chiu, May 7, 2012
Memory Management (3)

- Use **texture** and **shared memory**

Used by Dw(single-prec.) and x p update:
- Global Memory -> Texture Cache -> Shared Memory
- Shared Memory -> program utilization
- Register Memory -> program utilization

Used by M5(single-prec.), and all double prec. kernels:
- Global Memory -> Shared Memory
- Shared Memory -> program utilization
- Register Memory -> program utilization

T.W. Chiu, May 7, 2012
Benchmarks

- **CG (mixed prec.)** attains 317 Gflops on GTX580

<table>
<thead>
<tr>
<th></th>
<th>Dw(Single)</th>
<th>M5(Single)</th>
<th>Dw(Double)</th>
<th>M5(Double)</th>
<th>CG(Mixed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTX285</td>
<td>177</td>
<td>346</td>
<td>33</td>
<td>69</td>
<td>181</td>
</tr>
<tr>
<td>C1060</td>
<td>128</td>
<td>290</td>
<td>29</td>
<td>61</td>
<td>132</td>
</tr>
<tr>
<td>C2070</td>
<td>171</td>
<td>244</td>
<td>22</td>
<td>96</td>
<td>156</td>
</tr>
<tr>
<td>GTX480</td>
<td>293</td>
<td>309</td>
<td>37</td>
<td>116</td>
<td>252</td>
</tr>
<tr>
<td>GTX580</td>
<td>338</td>
<td>445</td>
<td>41</td>
<td>150</td>
<td>317</td>
</tr>
</tbody>
</table>

All numbers are in unit of Gflops, tested with ODWF on $16^3 \times 32 \times 16$ lattice

- **The bottleneck is Dw single-precision multiplication**
In general, \( \rho_s = c\omega_s + d \), \( \sigma_s = c\omega_s - d \), \( c,d \) (constants)

The effective 4D Dirac operator becomes

\[
D(m_q) = m_q + \left( m_0(1-dm_0) - \frac{m_q}{2} \right) \left[ 1 + \gamma_5 S_{opt}(H) \right], \quad H = \frac{cH_w}{1+d\gamma_5H_w}
\]

\[
S_{opt}(H) = \frac{1 - \prod_{s=1}^{N_s} T_s}{1 + \prod_{s=1}^{N_s} T_s}, \quad T_s = \frac{1 - \omega_s H}{1 + \omega_s H}
\]

\[
HR_Z^{(n-1,n)}(H^2), \quad N_s = 2n
\]

\[
HR_Z^{(n,n)}(H^2), \quad N_s = 2n + 1
\]
CUDA

Compute Unified Device Architecture

- Multicore CPUs and manycore GPUs means that the processor chips are parallel systems.

- The challenge is to develop application software that transparently scales its parallelism to leverage the increasing number of processor cores.

- CUDA is a scalable parallel programming model and software environment designed to meet this challenge, for programmers familiar with C.
Salient Features of the Quark Matrix

◆ $D$ is prohibitively large for exact solvers.

◆ In general, $D$ is a sparse matrix, since it only involves (next-to-)nearest neighbor interactions in 4-dim or 5-dim lattice.

◆ Iterative algorithms (conjugate gradient, Lanczos, etc.) are used, which involve the matrix-vector multiplication.

◆ CUDA kernels can be optimized for the matrix-vector multiplication in QCD.
FOR THE SPECIAL CASE $\rho_s = 1, \sigma_s = 0$

It reduces to the conventional DWF which has been using by RBC-UKQCD, which does NOT have the optimal chiral sym.

$$D(m_q) = m_q + \left( \frac{m_0}{2} (2 - m_0) - \frac{m_q}{2} \right) \left[ 1 + \gamma_5 S_{\text{polar}}(H) \right], \quad H = \frac{H_w}{2 + \gamma_5 H_w}$$

$$S_{\text{polar}}(H) = \frac{1 - T^{N_s}}{1 + T^{N_s}}, \quad T = \frac{1 - H}{1 + H}$$

$$b_l = \sec^2 \left[ \frac{\pi}{N_s} \left( l - \frac{1}{2} \right) \right]$$

$$d_l = \tan^2 \left[ \frac{\pi}{N_s} \left( l - \frac{1}{2} \right) \right]$$

Polar approximation

$$b_l = \frac{2}{N_s} \sum_{l=1}^{n} \frac{b_l}{H^2 + d_l}, \quad N_s = 2n$$

$$b_l = \frac{1}{N_s} + \frac{2}{N_s} \sum_{l=1}^{n} \frac{b_l}{H^2 + d_l}, \quad N_s = 2n + 1$$