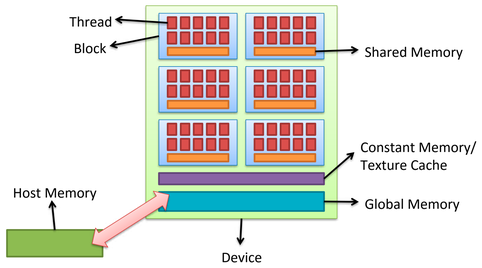
GPU and CUDA Architecture

Since 2009, researchers at National Taiwan University have successfully set up a GPU cluster which currently constitutes of 350 GPUs. This is the first GPU supercomputer in Taiwan. We have developed highly efficient CUDA codes for the most computationally challenging problems in high energy physics, condensed matter physics, and astrophysics. In 2015, our GPU cluster attains 200 Teraflops (sustained) for lattice QCD. During 2009-2015, we have developed efficient algorithms and CUDA codes for the ground-breaking simulation of lattice QCD with exact chiral symmetry. Now we are one of the three lattice QCD groups (RBC-UKQCD, JLQCD, TWQCD) around the world who can perform such a demanding large-scale lattice QCD simulation incorporating dynamical quarks with exact chiral symmetry. Remarkably, we have succeeded in performing our simulations using a GPU cluster, rather than expensive supercomputers (e.g., IBM BlueGene/Q).

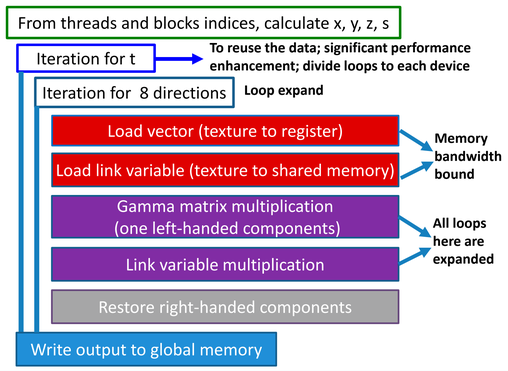
One of the most crucial part in the simulation program is the multi-GPU Conjugate Gradient (CG) solver with OpenMP. In the followings, the implementation and optimization of the two main kernels in matrix-vector multiplication in our CG calculation are discussed.



CG Kernels (D_w Multiplication)

$$(D_w)_{x,x'} = \frac{-1}{2}\sum_{\hat{\mu}}[(1-\gamma_\mu)U_\mu(x)\delta_{x+\hat{\mu},x'} + (1+\gamma_\mu)U_{-\mu}(x)\delta_{x-\hat{\mu},x'}]$$

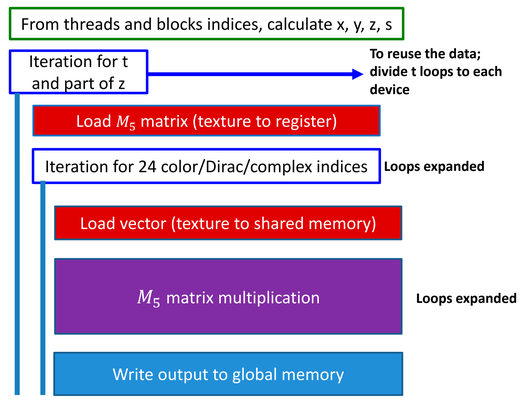
- Hopping terms
  - Texture is used for caching data
  - Internal loop is used to reuse the read-in data
  - Peer-to-Peer access is enabled to load the hopping term from other devices
- Link variables multiplication
  - For a given $\hat{\mu}$, U is the same in fifth dimension, hence the shared memory is used
- Gamma matrix multiplication
  - Only the left-handed Dirac indices are calculated.



CG Kernels (M_5 Multiplication)

$$M_5 = \left[(4-m_0) + \omega^{\frac{-1}{2}}\left[c(1+L)(1-L)^{-1} + d\omega^{-1}\right]^{-1}\omega^{\frac{-1}{2}}\right]^{-1}$$

- **Block diagonal in chiral basis**
- **Does not depend on x, y, z, t or color-Dirac indices**
- **It is the constant matrix multiplication in the 5th space**
- **Use share memory to store source vectors**
- **Internal loop to reuse the read-in M_5 matrix**



Benchmarks

CG (mixed prec.) attains 550 GFLOPS on GTX TITAN-X

	Dw(Single)	M5(Single)	Dw(Double)	M5(Double)	CG(Mixed)
C2070	171	244	22	96	156
GTX480	293	309	37	116	252
GTX580	338	445	41	150	317
GTX970	417	495	74	96	375
GTX TITAN	440	578	53	195	410
GTX TITAN-X	784	899	123	132	550

All numbers are in unit of GFLOPS, tested with DWF on 16^3 x 32 x 16 lattice

	1 GPU	2 GPU	Speedup
GTX680	248	453	1.83
K20c	286	535	1.87
GTX TITAN	410	781	1.90
GTX TITAN -X	550	1050	1.90

All numbers are in unit of GFLOPS, tested with DWF on 24^3 x 48 x 16 lattice

	2 GPU	4 GPU	Speedup
GTX690	475	942	1.98
GTX TITAN-Z	780	1350	1.73

All numbers are in unit of GFLOPS, tested with DWF on 32^3 x 64 x 16 lattice

TWQCD COLLABORATION

GPU Accelerated Computation of the β-Function of the SU(3) Gauge Theory with 10 Massless Fermions in the Fundamental Representation

Ting-Wai Chiu[1,2]
[1] Physics Department, National Taiwan University, Taipei 10617, Taiwan
[2] Center for Quantum Science and Engineering, National Taiwan University, Taipei 10617, Taiwan

Abstract

Recent experiments in the Large Hadron Collider at CERN have discovered the Higgs scalar at the mass ~125 GeV. Even though the Higgs scalar is an elementary particle in the Standard Model, there is still a possibility that such a light scalar might arise as a composite particle in non-abelian gauge theories with many fermions, provided that it is not too far below the conformal window. This study focuses on the SU(3) gauge theory with 10 massless domain-wall fermions in the fundamental representation, using a GPU cluster at National Taiwan University. Our result of the beta-function suggests that this theory is infrared conformal.

Lattice QCD with Domain-Wall Fermion

Quantum Chromodynamics (QCD) is the SU(3) gauge theory for the interaction between quarks and gluons. It manifests as the short-range strong interaction in the nucleus, and plays an important role in the evolution of the early universe, from the quark-gluon "plasma" phase to the hadron phase. To solve QCD is a grand challenge, since it requires the largest scale numerical simulation of the discretized action of QCD on the 4-dimensional space-time lattice [1].

For the QCD action $S = S_G(U) + \bar{\psi}D(U)\psi$, any physical observables $\mathcal{O}(\bar{\psi},\psi,U)$ can be obtained from

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

Then we can put this integral on the lattice and use Hybrid Monte Carlo (HMC) method to compute this integral. The most time-consuming part in HMC is to solve a linear system by the conjugate gradient algorithm(CG). By using GPU, we can boost our simulation dramatically.

Moreover, since quarks are relativistic fermions, the 5-th dimension is introduced such that massless quarks with exact chiral symmetry can be realized at finite lattice spacing, on the boundaries of the fifth dimension, the so-called domain-wall fermion (DWF) [2]. In this study, we use the optimal DWF [3]. The effective action of DWF can be written as

$$D_m(U) = D_w(U) + M_5(m)$$

1. K. G. Wilson, Phys. Rev. D 10, 2445 (1974).
2. D. B. Kaplan, Phys. Lett. B 288, 342 (1992).
3. T.W. Chiu, Phys. Rev. Lett. 90, 071601 (2003).

Salient Features of the Quark Matrix $D_m(U)$

- D_m is a sparse matrix, only involving the nearest neighbor interactions.
- Iterative algorithms (conjugate gradient, Lanczos, etc.) are used, which involve the matrix-vector multiplication.
- CUDA kernels can be optimized for the matrix-vector ops. in QCD.

Theoretical Background

Introduction

- To understand the beta-function of the SU(3) gauge theory with 10 massless fermions is a fundamental problem in quantum field theory, which requires nonperturbative (lattice) study.
- The strong interaction physics around IRFP may have significant impacts to the phenomenological models beyond the Standard Model, e.g., models of composite Higgs boson.
- It is vital to use lattice fermions with exact chiral symmetry (domain-wall /overlap) fermion, having exactly the same flavor symmetry as their counterparts in the continuum.

Finite Volume Gradient Flow Scheme

The gradient flow (Wilson flow) amounts to solving

$$\frac{dB_\mu}{dt} = D_\nu G_{\nu\mu}, \quad G_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu + [B_\mu, B_\nu], \quad D_\nu = \partial_\mu + [B_\mu, \ ],$$

in the flow time t with the initial condition $B_\mu\big|_{t=0} = A_\mu$.

The renormalized coupling on the lattice L^4 is obtained from

$$g_c^2(L,a) \propto \langle t^2 E(t)\rangle, \quad c = \frac{\sqrt{8t}}{L} = \text{constant}, \quad E(t) = \frac{1}{4}G^a_{\mu\nu}G^a_{\mu\nu}(t), \text{ energy density}$$

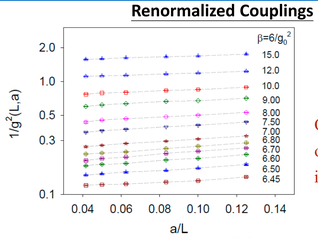
The proportional constant is fixed by requiring $g_c^2(L,a) = g^2_{\overline{MS}}$ to the 1-loop order.

Simulations

- Gauge action: Wilson plaquette action
  $\beta = 6/g_0^2$ = 6.45, 6.5, 6.6, 6.7, 6.8, 7.0, 7.5, 8.0, 9.0, 10.0, 12.0, 15.0
- Lattice sizes: **8^4, 10^4, 12^4, 16^4, 20^4, 24^4**
- For each (β, L), after thermalization, **2000-4000** trajectories have been accumulated. Sampling one configuration every 10 trajectories gives **200-400** configurations for measurement.

Discrete β-Function

Renormalized Couplings



$c = \frac{\sqrt{8t}}{L} = 0.3$

Other values of $g^2(L,a)$ are obtained by cubic spline interpolation.

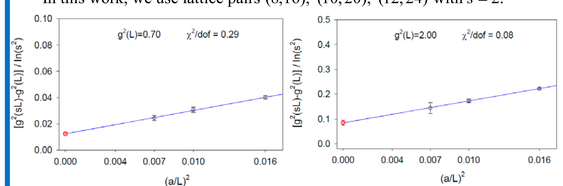
Discrete β-function

$$\beta(s,a/L) = \frac{g^2(sL,a) - g^2(L,a)}{\ln(s^2)}, \text{ for each pair of lattices } (L, sL).$$

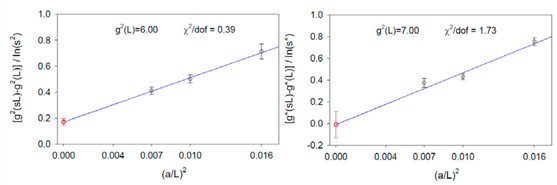
Given a set of lattice pairs with a fixed ratio s, $\lim_{a\to 0}\beta(s,a/L)$ can be obtained by extrapolation. Moreover, if $\lim_{a\to 0}\beta(s,a/L)$ is available for several s, then the limit $s \to 1$ can be taken, $\lim_{s\to 1}\lim_{a\to 0}\beta(s,a/L) = -\beta(g^2) = -\frac{dg^2}{d\ln\mu^2}$

If $\lim_{a\to 0}\beta(s,a/L)$ has IRFP, then $\beta(g^2)$ also has IRFP, and vice versa.

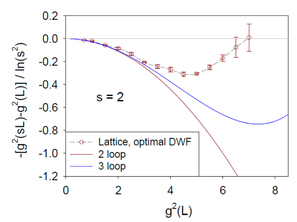
In this work, we use lattice pairs $(8,16)$, $(10,20)$, $(12,24)$ with $s = 2$.



Results and Discussions



Discrete β-function of the SU(3) Gauge Theory with Nf=10



Concluding Remarks

- Our data of the discrete β-function (with s=2) of the SU(3) gauge theory with 10 massless optimal DWFs suggests that the theory may possess IRFP at g_c^2 ~7.0, and it is infrared conformal.
- The statistical+systematic errors of the discrete β-function around the IRFP can be further reduced by increasing the statistics as well as more data points around the regime of strong couplings. Moreover, it is instructive to simulate L=32 , and add the fourth lattice pair (16,32) to the existing data of discrete β-function.