

GPU-Based Conjugate Gradient Solver for Lattice QCD with Domain-Wall Fermions

Ting-Wai Chiu^{1,2}, Tung-Han Hsieh³, Yao-Yuan Mao^{*,1}, Kenji Ogawa¹ (for the TWQCD Collaboration)

¹ Department of Physics, and Center for Theoretical Sciences, National Taiwan University, Taipei 10617, Taiwan

² Center for Quantum Science and Engineering, National Taiwan University, Taipei 10617, Taiwan

³ Research Center for Applied Sciences, Academia Sinica, Taipei 115, Taiwan

We present a GPU-based conjugate gradient (CG) solver for lattice QCD with domain-wall fermions (DWF). It is well-known that CG is the most time-consuming part in the Hybrid Monte Carlo simulation of unquenched lattice QCD, which becomes even more computational demanding for lattice QCD with exact chiral symmetry. We have implemented the CG for the general 5-dimensional DWF operator on NVIDIA[®] CUDA[™] architecture with mixed-precision, using the defect correction as well as the reliable updates algorithms. We optimize our computation by even-odd preconditioning and several innovative techniques for our CUDA kernels. For NVIDIA GeForce[®] GTX 285/480, our CG solver attains 180/233 Gflops (sustained).

*The XXVIII International Symposium on Lattice Field Theory, Lattice2010
June 14-19, 2010
Villasimius, Italy*

*Speaker.

1. Introduction

Simulation of unquenched lattice QCD with exact chiral symmetry is a grand challenge among all sciences. Even for a modest $16^3 \times 32$ lattice with lattice spacing $a \sim 0.1$ fm, it often requires a supercomputer with peak computing power more than 50 Teraflops (e.g., 10 racks of IBM BlueGene/L). Therefore, only 2-3 lattice QCD groups around the world could afford to perform the simulation of unquenched lattice QCD with the domain-wall fermion [1], or the overlap-Dirac fermion [2]. However, this scenario has been undergoing a dramatic change during the last 12 months. With the emergence of low-cost and computationally powerful Graphic Processing Unit (GPU), now plugging a graphic card with NVIDIA GTX 285 (240 cores, one Teraflops peak) into a PC immediately turns the system into a powerful device, attaining a sustained 180 Gflops for the conjugate gradient (CG) with mixed precision.

Since 2009, the Taiwan Lattice QCD Collaboration (TWQCD) has been using a GPU cluster (currently constituting of 250 NVIDIA GPUs with 40 Teraflops (sustained)) to simulate unquenched lattice QCD with optimal domain-wall quarks [3, 4]. We have met the challenge of preserving the chiral symmetry to a high precision and sampling all topological sectors ergodically. For our recent physical results on the 2-flavors QCD, we refer the readers to [5] and [6].

In this paper, we present our design of the CG solver for the general 5-dimensional DWF operator on NVIDIA CUDA architecture with mixed-precision, using the defect correction as well as the reliable updates algorithms. Our CG solver is optimized with even-odd preconditioning on the 4-dimensional space-time lattice, plus several innovative tuning techniques for the CUDA kernels. For NVIDIA GeForce GTX 285/480, our CG solver achieves 180/233 Gflops (sustained).

2. Conjugate Gradient Solver for Domain-Wall Fermions

2.1 Optimal Domain-Wall Fermions

For a given N_s (the number of sites in the fifth dimension), the mathematically maximal chiral symmetry can be attained by the optimal domain-wall fermion (ODWF) [3] with the operator

$$[\mathcal{D}(m_q)]_{xx';ss'} = (\omega_s D_w + 1)_{xx'} \delta_{ss'} + (\sigma_s D_w - 1)_{xx'} L_{ss'}, \quad (\sigma_s = \omega_s) \quad (2.1)$$

Here D_w is the standard Wilson Dirac Operator plus a negative parameter $-m_0$ ($0 < m_0 < 2$),

$$(D_w)_{xx'} = -\frac{1}{2} \sum_{\mu} [(1 - \gamma_{\mu}) U_{\mu}(x) \delta_{x+\hat{\mu},x'} + (1 + \gamma_{\mu}) U_{\mu}^{\dagger}(x') \delta_{x-\hat{\mu},x'}] + (d - m_0), \quad (2.2)$$

where $U_{\mu}(x)$ denotes the link variable, d is the dimension of the space-time ($d = 4$ for QCD),

$$L = P_+ L_+ + P_- L_-, \quad P_{\pm} = (1 \pm \gamma_5)/2, \quad (2.3)$$

and

$$(L_{\pm})_{ss'} = \begin{cases} \delta_{s-1,s'}, & 1 < s \leq N_s \\ -(m_q/2m_0) \delta_{N_s,s'}, & s = 1 \end{cases}, \quad L_- = (L_+)^{\dagger}. \quad (2.4)$$

The weights $\{\omega_s\}$ along the fifth dimension are fixed according to the formula derived in [3] such that the mathematically maximal chiral symmetry is attained. In general, for other DWF with non-maximal chiral symmetry, the weights $\{\sigma_s\}$ and $\{\omega_s\}$ have different values, e.g., for the conventional (Shamir) DWF, $\sigma_s = 0, \omega_s = 1, \forall s$, and for the Borici DWF [7], $\sigma_s = \omega_s = 1, \forall s$.

2.2 Even-Odd Preconditioning

Since D_w commutes with $(\omega)_{ss'} \equiv \omega_s \delta_{ss'}$ and $(\sigma)_{ss'} = \sigma_s \delta_{ss'}$, one can re-order the terms in Eq. (2.1) as

$$\mathcal{D}(m_q) = D_w(\omega + \sigma L) + (1 - L). \quad (2.5)$$

Separating the even and the odd sites on the 4D space-time lattice, Eq. (2.5) can be written as

$$\mathcal{D}(m_q) = \begin{pmatrix} d - m_0 & D_w^{\text{EO}} \\ D_w^{\text{OE}} & d - m_0 \end{pmatrix} (\omega + \sigma L) + (1 - L) = \begin{pmatrix} X & D_w^{\text{EO}} Y \\ D_w^{\text{OE}} Y & X \end{pmatrix}, \quad (2.6)$$

where

$$X \equiv (d - m_0)\omega(1 + cL) + (1 - L), \quad Y \equiv \omega(1 + cL), \quad (c)_{ss'} \equiv (\sigma_s / \omega_s) \delta_{ss'}. \quad (2.7)$$

Now we further rewrite it in a more symmetric form by defining

$$M_5 \equiv \sqrt{\omega}^{-1} Y X^{-1} \sqrt{\omega} = \left[(d - m_0) + \sqrt{\omega}^{-1} (1 - L) (1 + cL)^{-1} \sqrt{\omega}^{-1} \right]^{-1}, \quad (2.8)$$

and

$$S_1 \equiv \sqrt{\omega}^{-1} Y X^{-1} = M_5 \sqrt{\omega}^{-1}, \quad S_2 \equiv Y^{-1} \sqrt{\omega}. \quad (2.9)$$

then Eq. (2.6) becomes

$$\mathcal{D}(m_q) = S_1^{-1} \begin{pmatrix} 1 & M_5 D_w^{\text{EO}} \\ M_5 D_w^{\text{OE}} & 1 \end{pmatrix} S_2^{-1} = S_1^{-1} \begin{pmatrix} 1 & 0 \\ M_5 D_w^{\text{OE}} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & C \end{pmatrix} \begin{pmatrix} 1 & M_5 D_w^{\text{EO}} \\ 0 & 1 \end{pmatrix} S_2^{-1}, \quad (2.10)$$

where the Schur decomposition has been used in the last equality, with the complement

$$C \equiv 1 - M_5 D_w^{\text{OE}} M_5 D_w^{\text{EO}}. \quad (2.11)$$

Obviously, the most time-consuming task in the HMC is to solve the linear system $CC^\dagger|x\rangle = |b\rangle$ by the conjugate gradient (CG), namely, in the computation of the fermion force in the molecular dynamics. In this work, we implement the entire conjugate gradient inside the NVIDIA GPU, which can be used for the HMC as well as for computing the valence quark propagators.

2.3 Algorithm

Conjugate Gradient (CG) method [8] is a widely-used numerical algorithm for iteratively solving a linear system $Ax = b$ to a certain precision ε , where A is a positive-definite Hermitian matrix. With the CG algorithm (see Algorithm 1), the problem is turned into a task dominated by the matrix-vector multiplication. In this work, we utilize CUDA to implement the 5D domain-wall fermion operator (2.10) matrix-vector multiplications of the CG on NVIDIA GPUs. For the GPU, the single-precision operations are several times faster than the double-precision ones, thus it is advantageous to use the mixed-precision CG. In the so-called *defect correction* algorithm, one solves x in the low-precision, and updates the solution \hat{x} and the residue \hat{r} in the high-precision. (see Algorithm 2, where the hatted symbols represent variables in the high-precision). In this fashion, most of the floating-point operations are in the low-precision, thus it is advantageous for the GPU. Theoretically, the *defect correction* algorithm is mathematically sound, and it always works in prac-

Algorithm 1 Conjugate Gradient

```

 $x_0 :=$  initial guess
 $r_0 := b - Ax_0$ 
 $p_0 := r_0$ 
 $k := 0$ 
while  $|r_k| > \varepsilon|b|$  do
   $\alpha_k := (r_k, r_k) / (p_k, Ap_k)$ 
   $r_{k+1} := r_k - \alpha_k Ap_k$ 
   $\beta_{k+1} := (r_{k+1}, r_{k+1}) / (r_k, r_k)$ 
   $x_{k+1} := x_k + \alpha_k p_k$ 
   $p_{k+1} := r_{k+1} + \beta_{k+1} p_k$ 
   $k := k + 1$ 
end while

```

Algorithm 2 Mixed-Precision Conjugate Gradient (Defect Correction)

```

 $\hat{x} :=$  initial guess
 $\hat{r} := \hat{b} - \hat{A}\hat{x}$ 
while  $|\hat{r}_k| > \hat{\varepsilon}|\hat{b}|$  do
   $r := \hat{r}$ 
   $p := \hat{r}$ 
   $x := 0$ 
  Use Algorithm 1 to solve  $x = A^{-1}r$  in the low-precision to a precision  $\varepsilon$ 
   $\hat{x} := \hat{x} + x$ 
   $\hat{r} := \hat{b} - \hat{A}\hat{x}$ 
end while

```

tice. However, the seemingly drawback is that one has to build up the Krylov space every time it restarts the CG in the low precision. On the other hand, if one does not reset the low-precision p vector inside the while loop of Algorithm 2 (i.e., skipping the step ($p := \hat{r}$) except at the first time), the “warm-up” time in re-building the Krylov space could be reduced. This so-called *reliable updates* algorithm [9, 10] would run faster than the *defect correction*. Although the reliable updates in this fashion may not converge for all cases theoretically due to the non-orthogonality of p and r , in practice it still works for most cases. We have implemented both algorithms in our CG solver, and it automatically switches to the *defect correction* if the *reliable updates* does not converge in the first place.

3. CUDA Kernels and Optimization

The CUDA architecture developed by NVIDIA enables us to do parallel computations on NVIDIA’s GPUs. (For more detailed programming models and strategies, see “CUDA Programming Guide for CUDA Toolkit” [11].)

In CUDA, a *thread* is the smallest unit to execute a *kernel*, and a certain number of threads form a *block*. Each thread will be assigned a 3-dimensional index, and each block a 2-dimensional

index. Inside a block, a *warp* of threads will execute the *kernel* concurrently. Each thread has its own *register* memory, and threads in the same block share a *shared* memory. The space of register and shared memory is very limited, but they have the fastest access speed. The *global* memory on the device can be accessed by any thread, but it has the slowest bandwidth.

To implement the mixed-precision CG for ODWF with CUDA, we perform all matrix-vector multiplication, vector reduction (inner product), and vector addition/subtraction on the GPU (device), while the CPU (host) is used to do the flow control, memory assignment, and device control.

The CUDA kernels in our CG solver can be divided into five different catalogs. We will discuss each catalog and their optimization in the following subsections.

3.1 Vector Index Conversion

These kernels are used to change the indexing schemes between the device (GPU) and the host (CPU). To store the Dirac spinors in an one-dimensional array, we need to map the multi-dimensional indices to the 1D array index. One needs $4 \times 3 \times 2 = 24$ real numbers to store one Dirac spinor on one site of the 5D lattice. On the CPU, this color-spinor index c which runs from 0 to 23 is the inner-most (fastest-running) index, which is followed by the fifth-dimension index s , and then x, y, z, t indices, where t is the outer-most (slowest-running) index. If i_{host} denotes the one-dimensional array index of the CPU scheme, then we have

$$i_{host} = c + s \times 24 + x \times 24N_s + \cdots + t \times 24N_x N_y N_z N_s. \quad (3.1)$$

However, for computation on the GPU, we assign each thread a five-dimensional site index. This implies that adjacent threads have consecutive s indices. Thus we want to arrange the data such that optimal coalescence is attained when loading the vector from device global memory to the register and/or the shared memory of the threads. Since the GPU provides vector data types such as `float4` and `double2` which allow us to move 4 float numbers (or 2 double numbers) at one time, a simple way to map to the one-dimensional array index on GPU (for single precision) is

$$i_{dev} = c \bmod 4 + s \times 4 + [c/4] \times 4N_s + x \times 24N_s + \cdots + t \times 24N_x N_y N_z N_s, \quad (3.2)$$

and similarly for the double precision. So every time when we transfer data between the host and the device, we convert the index accordingly.

3.2 Matrix-Vector Multiplication for $D_w^{OE}(D_w^{EO})$

$D_w^{OE}(D_w^{EO})$ is the usual D_w without mass term,

$$[D_w^{OE}(D_w^{EO})]_{xx'} = -\frac{1}{2} \sum_{\mu} [(1 - \gamma_{\mu}) U_{\mu}(x) \delta_{x+a\hat{\mu}, x'} + (1 + \gamma_{\mu}) U_{\mu}^{\dagger}(x') \delta_{x-a\hat{\mu}, x'}]. \quad (3.3)$$

From this expression we see that the multiplication of $D_w^{OE}(D_w^{EO})$ with a vector involves the link variables $U_{\mu}(x)$ and γ -matrices. We have used the following tricks for optimization.

Firstly, since the γ -matrices in Eq. (3.3) are in the combination $(1 \pm \gamma_{\mu})$, the left-handed and the right-handed Dirac components are related to each other. Also, since the link variables do not have Dirac indices, we can just multiply $U_{\mu}(x)$ to the left-handed components, and then restore the right-handed components.

Secondly, $U_\mu(x)$ has no fifth-dimension dependence, so threads having the same x but different s can share the same $U_\mu(x)$. So we put the link variables in the shared memory.

Thirdly, because GPU computation is memory bandwidth bound, one must try to reuse the data. For example, the hopping term ($\delta_{x-a\hat{\mu},x'}$) in D_w , all neighboring sites of x are involved in the calculation. If we assign each x to one thread, then there must be overlapping data loading for neighboring sites. To reduce this overlapping data transfer, we distribute each (x, y, z) to one thread, with a loop in the t -direction. Then the neighboring data in the t -direction can be reused, and the efficiency is enhanced.

Besides above tuning techniques, we also expand small loops, and to use the texture memory for caching data, Here texture is used for loading the vectors and link variables. We use Python [12] to expand small loops, and to generate the set of (D_w multiplication) kernels.

3.3 Matrix-Vector Multiplication for M_5

The matrix M_5 is given by Eq. (2.8). One can see that M_5 is block diagonal in the chiral basis and it does not depend on the space-time nor the color indices. In fact, it can be divided into two constant matrices in the fifth dimension, i.e., the left-handed and the right-handed ones. So the multiplication of M_5 with a vector can be regarded as $u_s = \sum_{s'} (M_5)_{ss'} v_{s'}$. Here we use the shared memory to store the source vector (v). Since M_5 only takes $2N_s^2$ real numbers, we can put M_5 into the register of each thread (with different s and x, y, z). Again, a loop in t is inserted to reuse the M_5 data in the register. Also, we use Python to generate these kernels.

3.4 Vector Reduction

To calculate the norm of a vector, we use the well-known *parallel reduction* with the shared memory. However, due to the limitation on the number of threads per block, it is inevitable to use global memory when the size of a vector becomes very large. Our solution is to perform the *block reduction* in prior kernels, i.e., to sum up vector elements (already stored in registers/shared memory) within each block. Then these partial sums can be added with a parallel reduction.

3.5 Vector Addition and Subtraction

We can combine the simple addition/subtraction with other kernels in which one vector has been loaded. For example, to multiply $C \equiv 1 - M_5 D_w^{\text{OE}} M_5 D_w^{\text{EO}}$ to a vector, we can combine the last subtraction with the last M_5 multiplication.

4. Performance

We present some benchmarks of our CG solver, using NVIDIA GeForce GTX 285, GeForce GTX 480, Tesla™ C1060, and Tesla C2050. Note that our code has not yet been well-tuned for the Fermi architecture (GTX 480 and C2050). From Table 1, we see that the bottleneck of our program is in the single-precision D_w matrix-vector multiplication. Due to the mixed-precision CG, the time used in the double-precision operations are almost negligible. For the Fermi architecture, due to the larger L1 cache, there is a significant improvement in the single-precision D_w matrix-vector multiplication, and also in the double-precision M_5 matrix-vector multiplication.

Table 1: Benchmark of our CG solver for DWF on a $16^3 \times 32 \times 16$ lattice, numbers in units of Gflops)

	D_w (single)	M_5 (single)	D_w (double)	M_5 (double)	CG (mixed)
GTX 285	177	346	33	69	181
GTX 480	248	331	32	116	233
C1060	128	290	29	61	132
C2050	160	239	22	100	156

5. Summary

We have implemented an efficient GPU-based CG solver for generalized domain-wall fermions. Our CUDA kernels are tuned with several innovative schemes. On NVIDIA GeForce GTX 285/480, our CG solver achieves 180/233 Gflops (sustained). This efficient CG solver constitutes the most crucial part in TWQCD's HMC code for simulation of unquenched lattice QCD with the optimal domain-wall fermion.

Acknowledgments

This work is supported in part by the National Science Council (Nos. NSC96-2112-M-002-020-MY3, NSC99-2112-M-002-012-MY3, NSC96-2112-M-001-017-MY3, NSC99-2112-M-001-014-MY3, NSC99-2119-M-002-001) and NTU-CQSE (Nos. 99R80869, 99R80873).

References

- [1] D. B. Kaplan, Phys. Lett. B **288**, 342 (1992)
- [2] H. Neuberger, Phys. Lett. B **417**, 141 (1998); R. Narayanan and H. Neuberger, Nucl. Phys. B **443**, 305 (1995)
- [3] T. W. Chiu, Phys. Rev. Lett. **90**, 071601 (2003); Nucl. Phys. Proc. Suppl. **129**, 135 (2004)
- [4] T. W. Chiu *et al.* [TWQCD Collaboration], PoS **LAT2009**, 034 (2009) [arXiv:0911.5029 [hep-lat]].
- [5] T. W. Chiu *et al.* [TWQCD Collaboration], PoS **LAT2010**, 099 (2010)
- [6] T. H. Hsieh *et al.* [TWQCD Collaboration], PoS **LAT2010**, 085 (2010)
- [7] A. Borici, Nucl. Phys. Proc. Suppl. **83**, 771 (2000) [arXiv:hep-lat/9909057].
- [8] M. R. Hestenes and E. Stiefel, Journal of Research of the National Bureau of Standards **49**, 6 (1952).
- [9] G. L. G. Sleijpen, and H. A. van der Vorst, Computing **56**, 141-164 (1996).
- [10] M. A. Clark, R. Babich, K. Barros, R. C. Brower and C. Rebbi, Comput. Phys. Commun. **181**, 1517 (2010) [arXiv:0911.3191 [hep-lat]].
- [11] <http://developer.nvidia.com/object/gpucomputing.html>
- [12] <http://www.python.org>