Optimization of the Oktay-Kronfeld Action Conjugate Gradient Inverter

Yong-Chull Jang, Jon A. Bailey, Carleton DeTar, Andreas Kronfeld, Weonjong Lee, M. B. Oktay

Lattice Gauge Theory Research Center,
Seoul National University
(SWME, MILC, and Fermilab Lattice)

LATTICE 2013
Mainz, Germany
JUL. 29 – AUG. 3

Heavy Flavor Physics with Lattice QCD

Motivation

- 3σ tension in the neutral Kaon indirect CP violation parameter

$$
\begin{aligned}
|\epsilon_K|^{\text{exp}} &= 2.228(11) \times 10^{-3} && \text{(PDG)} \\
|\epsilon_K|^{\text{SM}} &= 1.6(2) \times 10^{-3} && (\text{SWME } \hat{B}_K, \text{FNAL/MILC } V_{cb})
\end{aligned}
$$

- Error budgets

$$
\sigma(X)^2/\sigma(|\epsilon_K^{SM}|)^2 = \begin{cases} 14\% & , \hat{B}_K \\ 51\% & , V_{cb} \end{cases}
$$

- A way of reducing the V_{cb} error is increasing the precision of lattice form factor calculation.

$$
B \to D^{(*)} l \nu_l
$$

- Heavy quark discretization is a dominant error source.

$$
m_b > a^{-1}, m_c \sim a^{-1}
$$

- Theoretical improvement is needed in contrast to take a brute force approach of reducing a lattice spacing continuously.

OK Action [M. B. Oktay and A. S. Kronfeld, PRD 78, 014504 (2008)]

Construction of LE$\mathcal{L}$

- OK action is the improved Fermilab action,
  [A. El-Khadra, A. S. Kronfeld and P. B. Mackenzie, PRD **55**, 3933 (1997)].
  - Building blocks

$$B_i, E_i, D_\mu, \psi, \bar{\psi}, \gamma_\mu$$

  - full set of $d = 6$ bilnear operators
  - part of $d = 7$ bilinear operators which commensurate to $d = 6$ operators by the power counting
  - No four-fermion operators (tree-level)

- The improvement terms are suppressed by up to
  - λ^3 of HQET power counting for heavy-light meson

$$\lambda \sim a\Lambda_{\mathrm{QCD}} \;\text{ or }\; \Lambda_{\mathrm{QCD}}/m_Q$$

  - v^6 of NRQCD power counting for heavy quarkonium

$$v = p/m_Q$$

OK Action

Power Counting

- Number of bilinears

d	4	5	6	7
λ^0	1			
λ^1	1	3		
λ^2		1	2(0)	
λ^3			5(4)	12(2)

HQET

$\lambda \sim a\Lambda_{QCD}$ or Λ_{QCD}/m_Q

d	4	5	6	7
v^2	2	2		
v^4		2	4(2)	2(1)
v^6			3(2)	8(1)
v^8				2(0)

NRQCD

$v = p/m_Q$

- 4 of 7 dimension 6 operators and

  2 of 12 dimension 7 operators have non-zero coupling after the tree-level matching.

OK Action

Matching: Tree Level

[A. El-Khadra, A. S. Kronfeld and P. B. Mackenzie, PRD **55**, 3933 (1997)]

- On-shell improvement amounts to an expansion in $\mathbf{p}a$
- Each matched coupling has full bare mass dependence, $c_i(m_0 a)$.

[M. B. Oktay and A. S. Kronfeld, PRD **78**, 014504 (2008)]

- Match the on-shell quantities
  - Energy (quark dispersion relation)
  - Current (quark-gluon vertex)
  - Quark-quark scattering
  - Compton scattering

- The tree-level matched action has 12 **operators**.

- **Use this action with the tree-level tadpole improvement.**

Form of the Dirac Operator

$$M_{x,y}\psi_y = \xi_x$$

- In temporal direction, only the *nearest neighbors* are involved.

$$M_{x,x\pm 4} \neq 0$$

- All of the *next-nearest neighbors* in spacial directions participate in. $(i, j = 1, 2, 3)$

$$M_{x,x\pm i},\ M_{x,x\pm i\pm j},\ M_{x,x\pm i\mp j} \neq 0$$

- Dirac operator receives *on-site* contribution from the clover term and the mass term.

$$M_{x,x} \neq 0$$

- Each terms $M_{x,y}$ consist of products of $1 \sim 5$ gauge links and γ matrices.

Optimization of the CG Inverter

Strategies

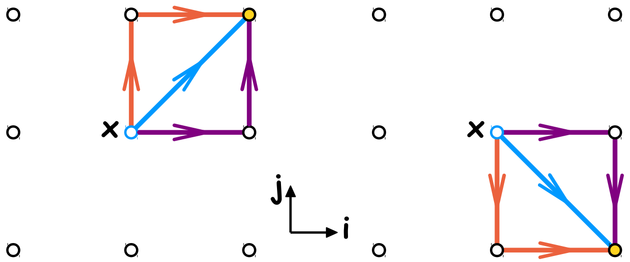
- Combinations of gauge link product(color matrix) in the Dirac operator are precalculated.
- Reflecting the γ matrix structure, the Dirac operator is represented by 4×4 block matrix, the precalcuation matrix.
- Each block is color matrix.

- $M^\dagger M$ preconditioning
- Even-Odd preconditioning
- Spin projection

Precalculation

Reduce the Floating Point Operations

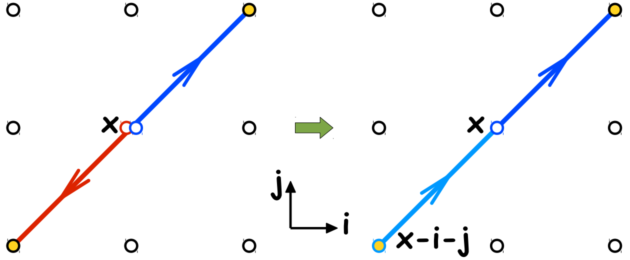
- The precalculation matrices connect the off-diagonal sites are symmetrized.



Precalculation

Toward a Memory Budget Solver

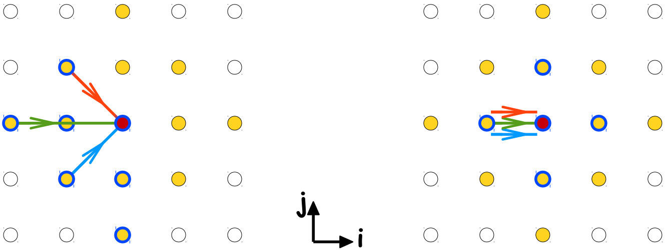
- The precalculation matrix which is pointing to the backward direction from the site x, is the *conjugate* of the forward precalculation matrix defined on the site $x - i - j$.
  - Hermitian conjugate in the spin-color space
  - followed by the sign correction



Precalculation

Simplify the Off-node Communication Pattern

- To update the fermion field on the red spot, one needs
  the fermion fields defined on the yellow sites and
  the precalcuation matrices defined on the blue-circled sites.

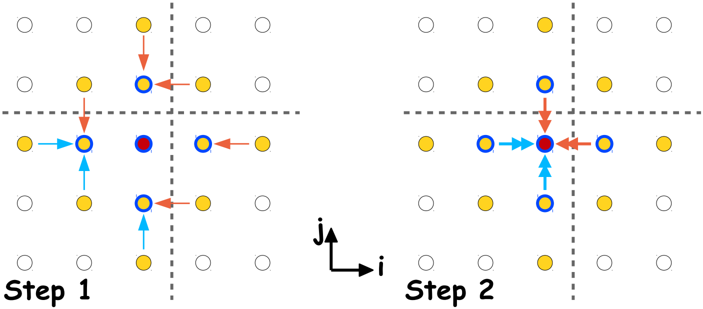


- To simplify this field access pattern, the shifted precalculation matrices are constructed.

Precalculation

Simplify the Off-node Communication Pattern

- Only the nearest neighbor off-node communications are required.
- Matrix multiplications are isolated from the communications.



- Only the nearest neighbor fermion fields are gathered and multiplied to the precalculation matrices.(Step 1)
- Then, the multipication results are gathered from the nearest neighbor sites and added up.(Step 2)

Libraries

Introduction

- **QOPQDP**:

  the implementation of QCD OPerations using QDP
  e.g., fermion inverters

- **QDP**:

  a C implementation of the LQCD suitable Data Parallel interface

- **CUDA**:

  the parallel computing architecture(Compute Unified Device Architecture) and/or programming language implemented by the NVIDIA GPU

- **MILC**:

  a LQCD application package which is hosted by the MILC collaboration

- **QUDA**:

  the QOPQDP analogy which is implemented by using CUDA

Libraries

Development Environments

- Bi-Stabilized CG solver in the QOPQDP is used.
- Mixed precision CG: Single precision iteration followed by a few double precision update

- $\not{D}$ function:
  - matrix multiplication
$$M_{x,y}\psi_y = \xi_x$$
  - precalculation
  - communication

- For CPU cluster,
  - The $\not{D}$ function in the QOPQDP is optimized.
  - To test the solver, heavy meson correlators are calculated by using the MILC code.

Libraries

Development Environments

- For GPU cluster,
  - Only the matrix mutiplication module in the optimized QOPQDP $\not{D}$ function is replaced by CUDA.
  - This matrix multiplication module for GPU is not fully optimized.
  - Precalculation and communication modules belong to the optimized QOPQDP code.

Performance

Time Table

- MILC coarse lattice $20^3 \times 64$ / 4 Nodes

	Naive	Optimized	CUDA
CG total (s)	11814.8	3036.8	891.0
Gain	1	3.9	13.3

- CPU: Intel i7 920@2.67GHz
- GPU: NVIDIA GTX 480
- Network: QLogic InfiniBand, 1 Rail

Timing (s)	Optimized	CUDA
Matrix Multiplication [SP]	2469.3	79.7
Matrix Multiplication [DP]	109.7	10.1
FLOPS (GF/Nodes)	2.1	58.8
CUDA Memory Copy, W		44.1
CUDA Memory Copy, ψ		191.6
FLOPS (GF/Nodes)		18.4
QOPQDP Preparation		138.9
Precalculation	67.1	
Communication	7.3	
Gamma Basis Change	45.2	
Spin Decomposition	62.5	

- Overhead(374.6s) exceeds the floating point calculation(89.8s).

Performance

Memory Requirement

- GTX 480 has 1.5GB global memory.
- Precalculation matrix is quite memory demanding.
- Only the single precision precalculation matrix is saved on the GPU global memory.

nodes	nx	ny	nz	nt	CPU(GBytes/node)	GPU(GBytes/node)
4	20	20	20	64	1.837	0.704
12	28	28	28	96	2.520	0.966
32	40	40	40	96	2.755	1.056
64	48	48	48	144	3.571	1.368
192	64	64	64	192	3.762	1.441

- Option: GTX Titan has 6GB global memory.

Future Work

- Optimize the GPU version of the OK action CG inverter further
- It is expected that the QOPQDP side overhead can be removed by using QUDA.
- The total memory transfer between CPU and GPU should be reduced.
- Use this inverter for the V_{cb} calculation

Thank you for your attention !