Large scale Domain Wall Fermion simulations on GPUs: Techniques and Properties

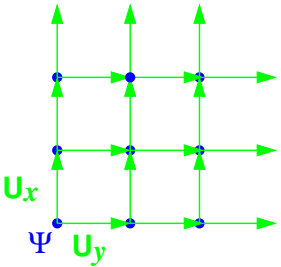
Chulwoo Jung, for RBC/UKQCD collaborations

David Murphy, Jiqun Tu, Robert Mawhinney, Christoph Lehner, Peter Boyle, ...

Feb. 3, 2021

Introduction to lattice QCD

Quantum ChromoDynamics (QCD): Theory of strong interaction which governs interaction between quarks and gluons.
In contrast to Quantum Electrodynamics (QED), The effective coupling of QCD decreases in high energy, hence is calculable by hand, but not in low energy. $\rightarrow$ Nonperturbative techniques such as lattice QCD is needed for *ab initio* calculations.



$$(\psi(x), A_\mu(x)) \rightarrow (\psi(n), U_\mu(n) = \exp(-iA_\mu))$$

$$\begin{aligned} Z &= \int [dU] \det(\not{D} + m) e^{-(S_g)} \\ &= \int [dU][d\bar{\psi}][d\psi] \ \exp[-(S_g + S_f)] \\ S_f &= \bar{\psi}(D^\dagger D)^{-1}\psi, \quad S_{eff} = S_g + S_f \\ S_g &= \beta \sum \left[(U_\mu(x) U_\nu(x+\hat{\mu}) U_\mu^\dagger(x+\hat{\nu}) U_\nu^\dagger(x) \right] \end{aligned}$$

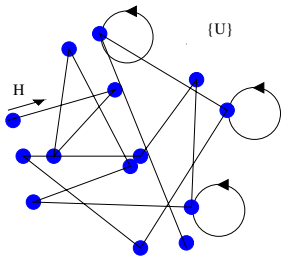
Current "typical" calculation: $V = 64^3 \times 128$, rank(D) $\sim 10^{10}$, nonzero element per row $=\sim 10^2$

Configuration generation, measurements

(Hybrid → Hamiltonian) Monte Carlo first devsied for LQCD (Duane,Kennedy,Pendleton/Gottlieb,Toussaint, ...)

Importance sampling using $\exp(-(S = (S_g + S_f)))$ as the probability distribution, requires S_f to be real & nonnegative. Introduce fictitious momenta H to evaluate the path integral. $S_f = \det(\not{D} + m)$ is evaluated by Pseudofermions.

Numerical integrators does not preserve $h = (\frac{1}{2}H^2 + S)$ exactly. Calculate h at each end of trajectory and accept or reject according to $max[1, \exp[-(h' - h)]$. This process achieves $\pi(U) \propto \exp(-S)$ if reversible and $\pi(U_1)P(U_1 \rightarrow U_2) = \pi(U_2)P(U_2 \rightarrow U_1)$ (detailed balance).



$$Z = \int [dU][d\bar{\psi}][d\psi][dH] \ \exp(-(\frac{1}{2}H^2 + S_2)),$$

$$\frac{dU_\mu(x)}{dt} = iH_\mu(x)U_\mu(x), \frac{d(\frac{1}{2}H^2 + S_1)}{dt} = 0$$

$$< O >= \frac{\int O \exp(-S_3)}{\int \exp(-S_3)} = \frac{\sum_i O_i w_i}{\sum_i w_i} \qquad w_i = e^{-(S_3 - S_2)}$$

Different Hamiltonians used in Monte Carlo simulations:
S_1: Guiding Hamiltonian during MD (Shadow Hamiltonian)
S_2: Hamiltonian for Accept/reject step
S_3: Hamiltonian for ensemble averaging
Reweighting: $(S_3 - S_2) \neq 0$. In principle, $S_1.S_2, S_3$ don't have to be the same, but it is hard to find $S_{2,3}$ which the acceptance/reweighting factor $e^{-\Delta S}$ is close enough to 1 while gives significant benefit (Overlap problem). Many ML based approaches aim to find S_2 which is numerically cheap with tolerable ΔS.

LQCD workflow and characteristics

1. Configuration generation: generate U_i according to $\exp(-S)$ evaluation of $\det(\not{D} + m)$ requires inversion of spase matrix M, a discretized version of $(\not{D} + m)$ for evolving gauge configurations.
2. Measurements: often requires multiple inversions, but on the same gauge configurations $\rightarrow$ allows for effective use of algorithms and techniques for multiple RHS

M from typical LGT simulations exhibits a separation of scales:

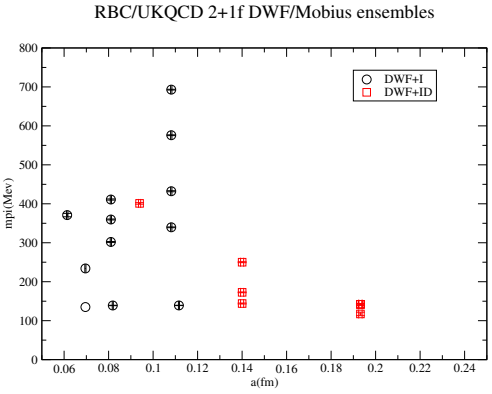
- Larger eigenvalues ($> m_s$): Dense. Usually straight inversion.
- Small eigenvalue ($\sim< m_s$): relatively small, slowly evolving in MD , affects the condition number , can be separated by deflation, Low-mode/All-mode averaging (LMA/AMA), multi-timescale integration algorithm, etc...

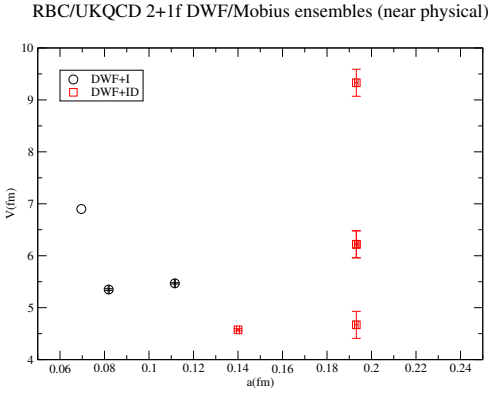
RBC/UKQCD Domain Wall Fermion program

RBC/UKQCD has chosen DWF discretization of Dirac operators, which realizes 4d fermion as boundary modes of 5d fermion.

- Good chrial symmetry: Remnant symmetry breaking (residual mass) can be controlled separately from lattice spacing by increasing the extent of the 5th dimension (L_s) and the coupling between 4d slices ((z)Möbius, Möbius Accelerated DWF (MADWF)..)
- Protected zero mode: In contrast to Wilson fermions where the discretized Dirac operator can have poles near the valence mass (exceptional configurations), DWF formalism guarantees safety as long as valence mass is positive. Allowing simulation at physical point for moderate lattice spacing without the need for chrial extrapolation or link smearing for fermions, etc. $\rightarrow$ Focus on physical point. Avoid relying on ChPT.
- Focus on generating one longest possible Markov chain (vs. 'farming') : While evolving multiple chain can give practical advantages, this needs additional thermalization and potentially obscures autocorrelation, ergodicity issues.

RBC/UKQCD 2+1f ensembles





DWF+I: Iwasaki gauge action

DWF+ID: Iwasaki + Dislocation Suppressing Determinant Ratio : Supresses the chiral symmetry breaking on larger lattice spacing

- Performance optimization and algorithmic advances has brought computaitional cost for finer(smaller a) and lighter (near physical m_l) per trajectory somewhat manageable, despite challenges from computing environement (severe inter-node bandwidth decrease compared to compute capabilities)
  - Mass preconditioning (Hasenbusch), N-th root trick: decrease the condition number and force for each pseudofermions, which allows to use larger step size and reduce the number of light quark inversions per each trajectory.

$$\left|\frac{D(m_f)}{D(1)}\right| = \left|\frac{D(m_1)}{D(1)}\right| \left|\frac{D(m_2)}{D(m_1)}\right| \cdots \left|\frac{D(m_f)}{D(m_{n-1})}\right| = \left[\left|\frac{D(m_f)}{D(1)}\right|^{1/N}\right]^N$$

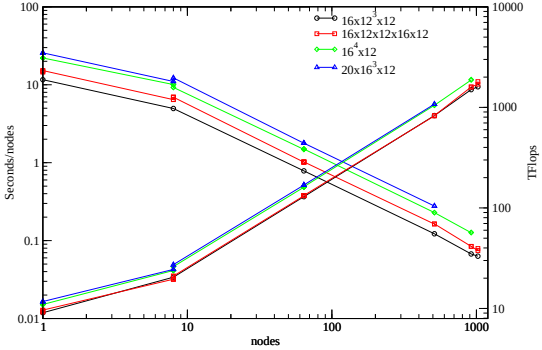
  - Exact One flavor algorithm (T.W. Chiu, D. Murphy,..): Especially useful on GPU, as it significantly reduces the memory traffic. $\sim 40\%$ reduction in time per MD.
  - Multisplitting preconditioned CG (J. Tu): Utilizes Tensor core. $\sim 10 - 20\%$ gain over CG
- Various deflation techniques, mostly built on efficient generation of exact eigenvectors (All-to-All (A2A), All mode averaging (AMA)), achieves significant reduction in numerical cost for generation of necessary propagators (Disconnected HVP, etc....)

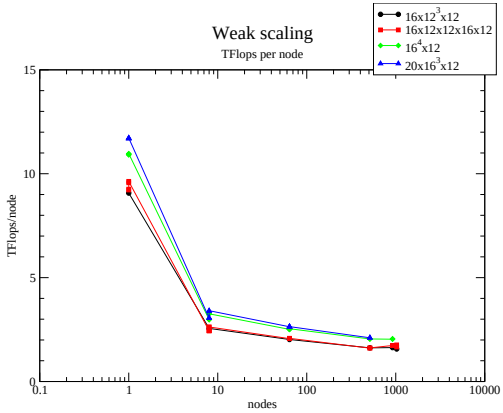
OLCF SUmmit (https://www.olcf.ornl.gov/summit/)

(6 Nvidia V100 2 Power9 CPU, 512GB DDR4 + 96GB HBM2) $\times$ 4608 nodes
Infiniband EDR
2+1 flavor physical , $96^3 \times 192 \times 12$, a^{-1} $\sim$2.77Gev, $L \sim 6.8$fm
$16 \times 12^3 \times 12$ on $(1 \times 8 \times 8 \times 16 = 1024)$ ndoes $\times$ 6 GPUs
Software: CPS(https://github.com/RBC-UKQCD/CPS),
QUDA(https://github.com/lattice/quda), Grid (https://github.com/paboyle/Grid),
GPT(https://github.com/lehner/gpt)

- Tuning with (currently 7) Hasenbusch masses, Force Gradient...
- Started from a thermalized $32^3 \times 64$ lattice duplicated in all directions
- QUDA inverter (CG+Multimass) interface (re)checked against CPS/BFM. Exact One Flavor Algorithm(EOFA) added.
- Previously only had interface to asymmetric preconditioner. Symmetric added for Multisplitting-preconditioned CG (MSPCG: arXiv:1804.08593).

QUDA Möbius inverter performance on Summit at the INCITE submission





Performance limited by network bandwidth.

Not-so-great weak scaling: weak scaling QUDA inverter performance drops by factor of ~ 4

Exact one-flavor(EOFA) TWQCD, D. Murphy, J. Tu,*et al.* arXiv:1403.1683, 1611.00298, 1706.05843

$$D_{Mob}(m) = (D_w)_{xx'}\left[(c+d)\delta_{ss'} + (c-d)L_{ss'}\right] + \delta_{xx'}(1-L)_{ss'} = D_{EOFA}\cdot\tilde{D} \tag{1}$$

$$\tilde{D} = d(1-L) + c(1+L) = \begin{pmatrix} c+d & (c-d)P_- & & & -m(c-d)P_+ \\ (c-d)P_+ & c+d & (c-d)P_- & & \\ & (c-d)P_+ & c+d & \ddots & \\ & & \ddots & \ddots & \ddots \\ -m(c-d)P_- & & & (c-d)P_+ & c+d \end{pmatrix}$$

$$D_{EOFA}(m) = (D_w)_{xx'}\delta_{ss'} + \delta_{xx'}(P_+M_+)_{ss'} + \delta_{xx'}(P_-M_-)_{ss'}, H_{EOFA}(m) = \gamma_5 R_5 D_{EOFA}(m),$$

1 flavor Mobius fermion action with PV (det $\left|\frac{D_{EOFA}(m_1)}{D_{EOFA}(m_2)}\right|$) can be simulated by

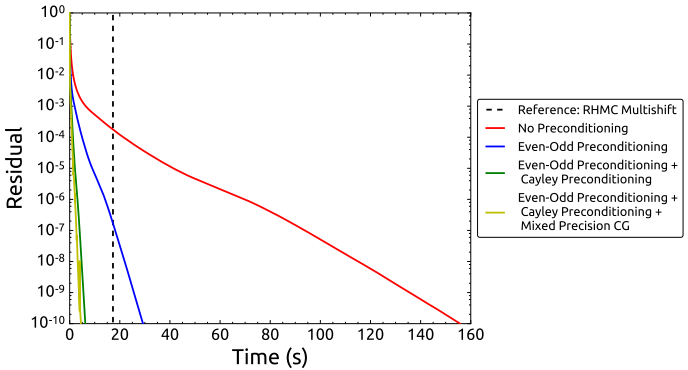
$$\begin{aligned} S_{pf} \quad = \quad & \begin{pmatrix} 0 & \phi_1^\dagger \end{pmatrix}\left[I - k\Omega_-^T \frac{1}{H_T(m_1)}\Omega_-\right]\begin{pmatrix} 0 \\ \phi_1 \end{pmatrix} + \\ & \begin{pmatrix} \phi_2^\dagger & 0 \end{pmatrix}\left[I + k\Omega_+^T \frac{1}{H_T(m_2) - \Delta_+(m_1,m_2)P_+}\Omega_+\right]\begin{pmatrix} \phi_2 \\ 0 \end{pmatrix}, \end{aligned} \tag{2}$$

Advantages:

Allow mixed precision, etc to reduce time on inversion

Less memory footprint: improve overall arithmetic intensity, especially significant on GPUs

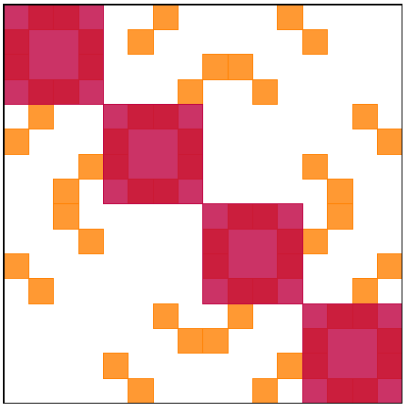
Eq. (1) suggests $\tilde{D}$ can be used as a preconditioner, to allow using D^{-1}_{MOB} instead of D^{-1}_{EOFA} with dense 5D matrix.



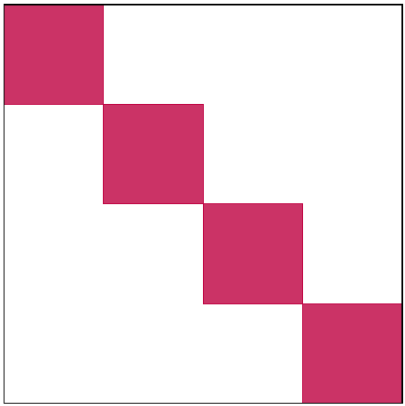
Factor of ~ 4 savings overall for $32^3 \times 64$ G-parity ensemble (32ID-G).

Multisplitting-preconditiond Conjugate Gradient(MSPCG) arXiv:1811.08488

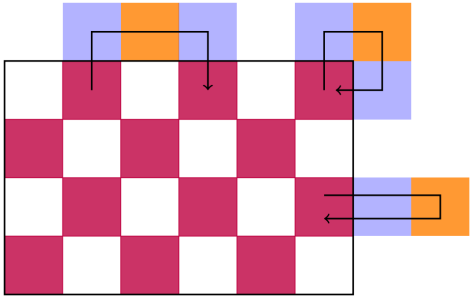
Additive Schwarz 'done right' for Möbius CGNE ($D^\dagger D\phi = D^\dagger \chi$).
Fixed iteration preconditioner per outer iteration.



A

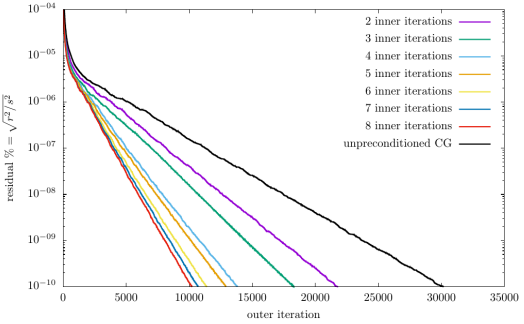


$P = \bigoplus_s A_s$



from J. Tu, arXiv:1811.08488

The preconditioner inversion P^{-1} does not need to be solved to arbitrarily high precision for the algorithm to work.



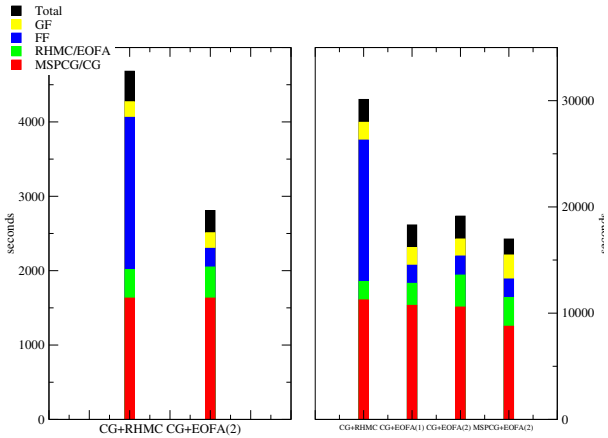
Up to factor of 3 reduction in outer iteration.

- QUDA Dslash rewrite (K. Clark,..),
- 5D part fused to achieve better overlap with communication (J. Tu) + Tensor core for 5D part

Time in CG	$< 10\%$	25-35%	$> 60\%$	
$64^3 \times 128 \times 12$	Double	Half	Precon	min/traj.
$(4\times)4^2 \times 8 = 128$	160	570	3230	86
$(4\times)4^2 \times 16 = 256$	260	861	6230	53
$(4\times)4 \times 8 \times 16 = 512$	360	1165	11630	36
$96^3 \times 192 \times 12$				
$(6\times)4^2 \times 16 = 256$	420	1340	9400	
$(6\times)4 \times 8 = 512$	770	2300	18810	79
$(6\times)8^2 \times 16 = 1024$	1140	3700	36300	47

Table 1: Aggregate QUDA Möbius performance on summit, in TFLOPS/s

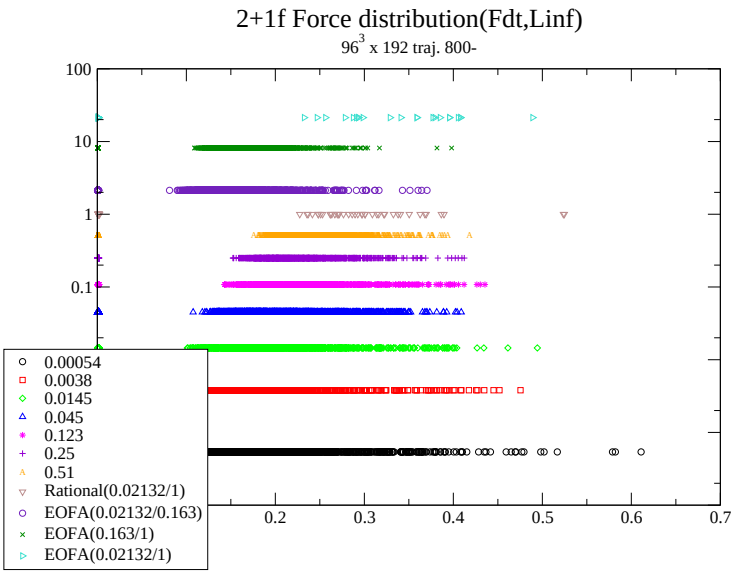
96I evolution on Summit



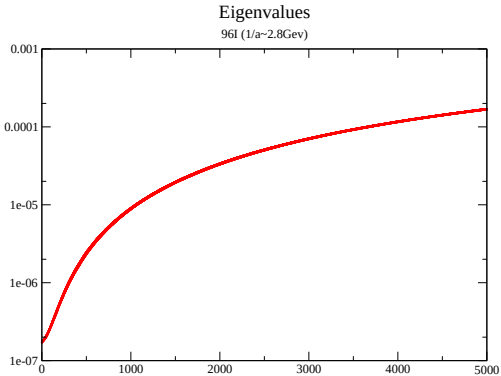
1024 nodes
CG+RHMC/CG+EOFA(2)

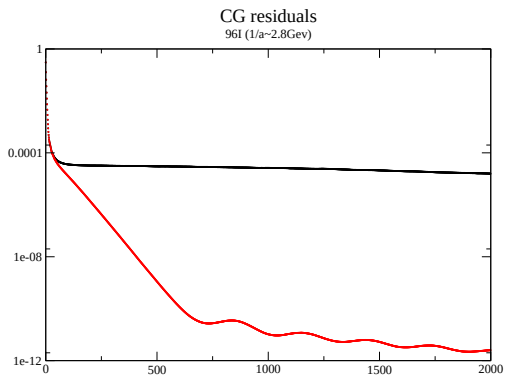
128 nodes
CG+RHMC/CG+EOFA(1)/
CG+EOFA(2)/MSPCG+EOFA(2)

While Multimass solver does not take much time, fermion force term calculation becomes time consuming. Arithmetic intensity low (no smearing).
EOFA with the Cayley preconditioner allows effective use of mixed precision solvers, efficient mass preconditioning, and reduce the number of pseudofermions significantly.


2+1f Force distribution(Fdt,Linf)
96^3 x 192 traj. 800-
100
10
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.00054
0.0038
0.0145
0.045
0.123
0.25
0.51
Rational(0.02132/1)
EOFA(0.02132/0.163)
EOFA(0.163/1)
EOFA(0.02132/1)

Eigenvector generation & deflation





From coarse grid Lanczos(arXiv:1710.06884) implemented in Grid. Factor of $\sim$1000 reduction in condition number turning to $\sim$ 30 reduction in iteration count for each inversion. Coarse grid also reduces memory footprint by a factor of $\times \sim 30$ over normal eigenvectors.

- All mode averaging (AMA):

$$\left\langle \mathcal{O}^{(\mathrm{imp})} \right\rangle = \left\langle \frac{1}{N_E} \left(\mathcal{O} - \mathcal{O}^{(appx)} \right) \right\rangle + \left\langle \frac{1}{N_G} \sum_{g \in G} \mathcal{O}^{(appx),g} \right\rangle . N_E \ll N_G$$

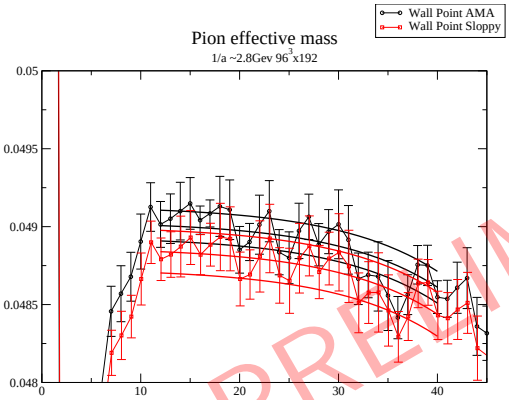
  $\mathcal{O}^{(appx)}$: $\mathcal{O}$ with 'sloppy' propagators, often much relaxed stopping condition with deflation

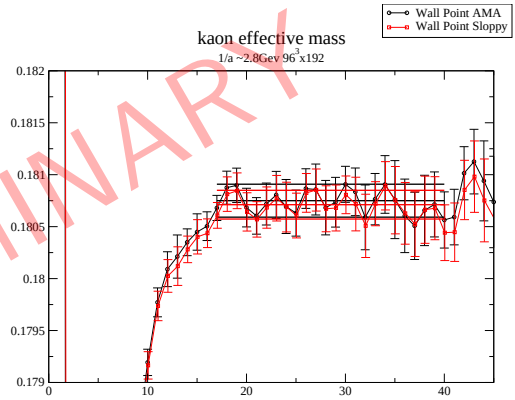
- All-to-all(A2A): A2A approximates an arbitrary component of the M^{-1} as a sum of the outer product of pre-calculated vectors

$$M^{-1} \simeq \sum_{i}^{N_l} |\lambda_i\rangle \frac{1}{\lambda_i} \langle \lambda_i| + \sum_{i'}^{N_h} \left(M^{-1} - \sum_{i}^{N_l} |\lambda_i\rangle \frac{1}{\lambda_i} \langle \lambda_i| \right) |\eta_{i'}\rangle \langle \eta_{i'}|$$

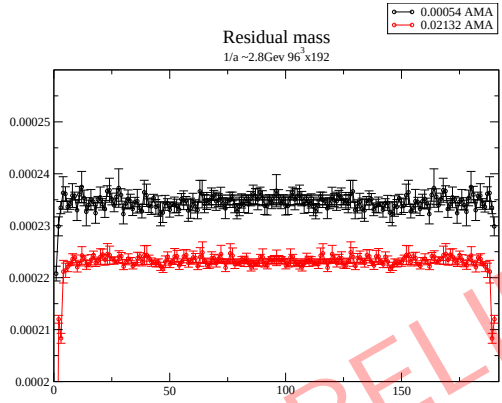
$$= \sum_{i=N_l+1}^{N_l+N_h} |v_i\rangle \langle w_i|$$

Pion, Kaon effective masses on 96I ensemble (preliminary)

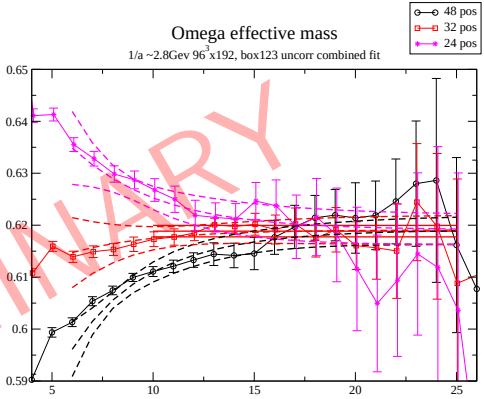




$(((48\mathit{sloppy} + 8\mathit{exact}) \times 12)$ light+strange) inversion per configuration



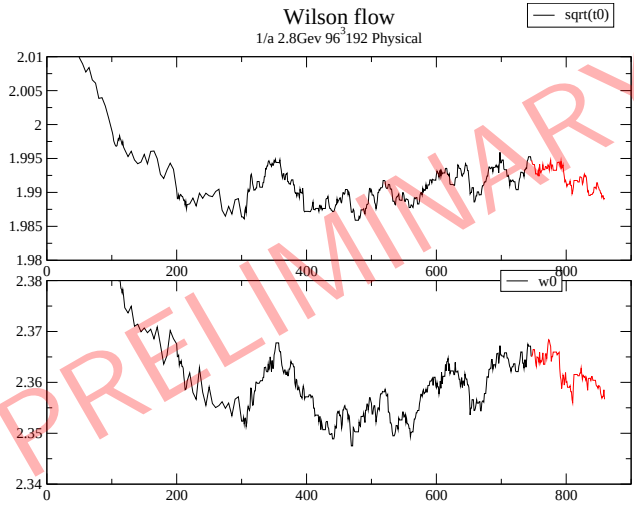
$m_{res}a \sim 2 \times 10^{-4}$



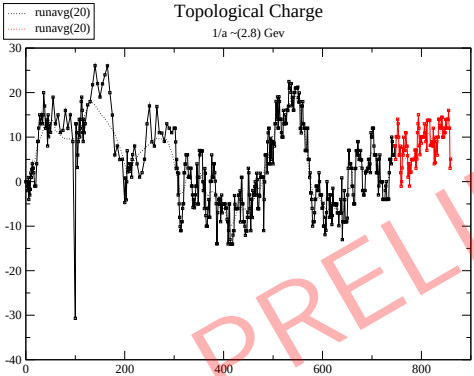
$m_\Omega a \sim 0.62$

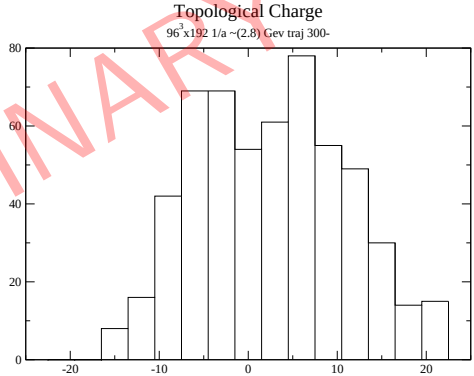
Z(3) guage-fixed box source

Wilson flow scale on 96I ensemble



Global topology on 96I ensemble





Me

autocorrelation $\sim$ 50MD

Discussion & future plans

- $96^3 \times 192$ DWF+Iwasaki ensemble (96I) will provide RBC/UKQCD with the physical mass at 3rd lattice spacing.
- Current lattice spacing was chosen to achieve reasonable topology tunneling. Going further requires a significant increase in the computing resource required to generate similar number of independent configurations ($cost$) $\sim a^{-10}$ (Critical slowing down).
- DWF ensemble generation on new and upcoming machines are likely to continue to be limited by the memory and internode bandwidth. While EOFA and MSPCG has helped mitigating these issues, evolution is more vulnerable compared to measurements, where various techniques (exact deflation, AMA, A2A, Split Grid...) are already devloped to mitigate, if not overcome, bandwidth issues.
- Algorithmic improvements for reducing autocorrelation time and numerical cost for evolution are critical.

Thank you!

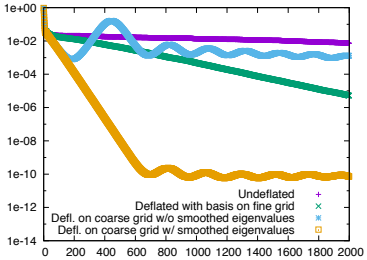
Local coherence: Eigenvector compression and Multi-Grid Lanczos

LQCD operator has relatively low number of near-nullspace, which allows for efficient deflation via multigrid or exact deflation with Chebyshef-accelerated Lanczos. 1-2000 low modes can result in factor of 10-100 reduction in subsequent measurements. Storing and retrieving these eigenvectors pose a signifiant challenge for exact deflation approach.

Local coherence ('smoothness') of the low modes suggests the bases formed from spatially blocked eigevectors effectively spans eigenspace just above already generated ones $\rightarrow$ necessary to only save coefficients. (https://github.com/lehner/eigen-comp)

$48^3 \times 96 \rightarrow 4^3 \times 3$, 2000 evecs : 9.3 $\rightarrow$ 1.5 TB (84% compression)
$64^3 \times 128 \rightarrow 4^4$, 2000 evecs : 36 $\rightarrow$ 3.5 TB (90% compression)
$96^3 \times 192 \rightarrow 4^4$, 5000 evecs : 500 $\rightarrow$ 15 TB (97% compression)
Multi-Grid Lanczos(arXiv:1710.06884):
Using the same blocking procedure to generate eigenvectors reduces memory usage significantly. Allows more efficient eigenvector generation

Deflated CG residual from Multi-Grid Lanczos



Split Grid/Domain

Most time consuming LQCD operations has flops/byte ~ 1, L = 10 - 20 $\rightarrow$ Maximum performance per node is often limited by internode bandwidth.

Most available "communication avoiding" algorithms focus more on minimizing number of global operation or overhead from them, rather than minimizing the amount of communication.

For some applications, there are multiple, duplicate routines needed (Dirac operator inversion on multiple sources, etc).

Within single binary, switch between 1 domain (MPI/QMP Communicator) and multiple small domains $\rightarrow$ improve the surface/volume.

Split CG: Inversion on multiple sources done on split grid.

$\sim$4X gain on 256-node Cori at NERSC.

$\sim$2X gain on 16-split Summit at ORNL.

Eigenvector generation (Y. Jang, C. Jung): Lanczos with block size 8 converge with a similar numerical cost as IRL (Lehoucq and Sorensen). Split domain implementation in Grid done.