Performance of Staggered Fermion Kernels using Grid

Patrick Steinbrecher[1], Swagato Mukherjee[1]

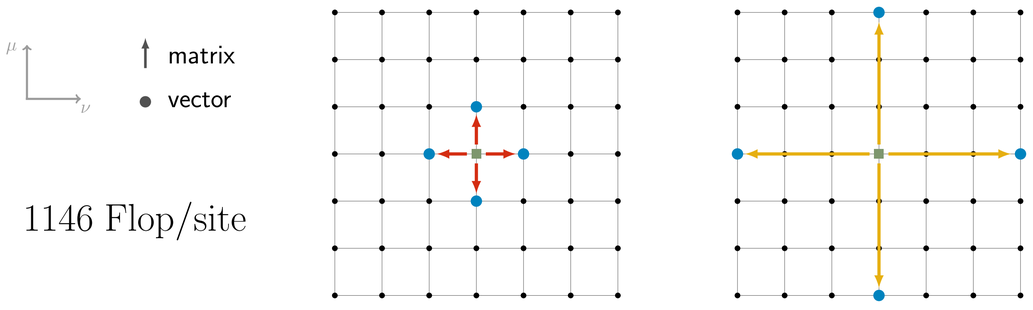
[1] Brookhaven National Laboratory - USA

HISQ Dslash Operator

The key operation in many Lattice QCD simulations is the inversion of the fermion matrix. It requires a 4-dimensional stencil which calculates the product of a vector ν by a sparse matrix known as the Dslash operator and stems from a discretized 4-dimensional derivative.

$$w_n = \sum_{\mu=0}^{4} \left[\left(U_{n,\mu} \nu_{n+\mu} - U^{\dagger}_{n-\mu,\mu} \nu_{n-\mu} \right) + \left(N_{n,\mu} \nu_{n+3\mu} - N^{\dagger}_{n-3\mu,\mu} \nu_{n-3\mu} \right) \right]$$

ν: complex 3-dim vector; U: complex 3×3 matrix; N: $U(3)$ matrix

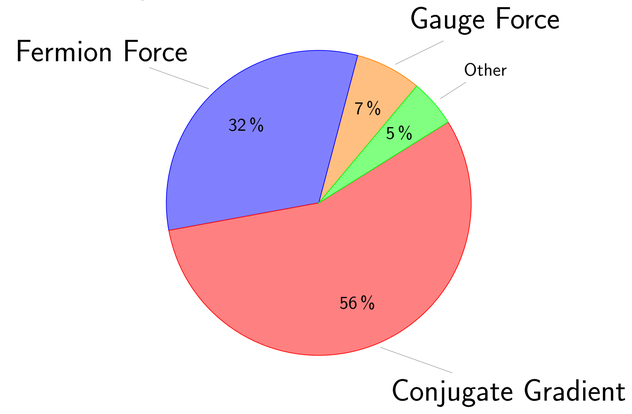


1146 Flop/site

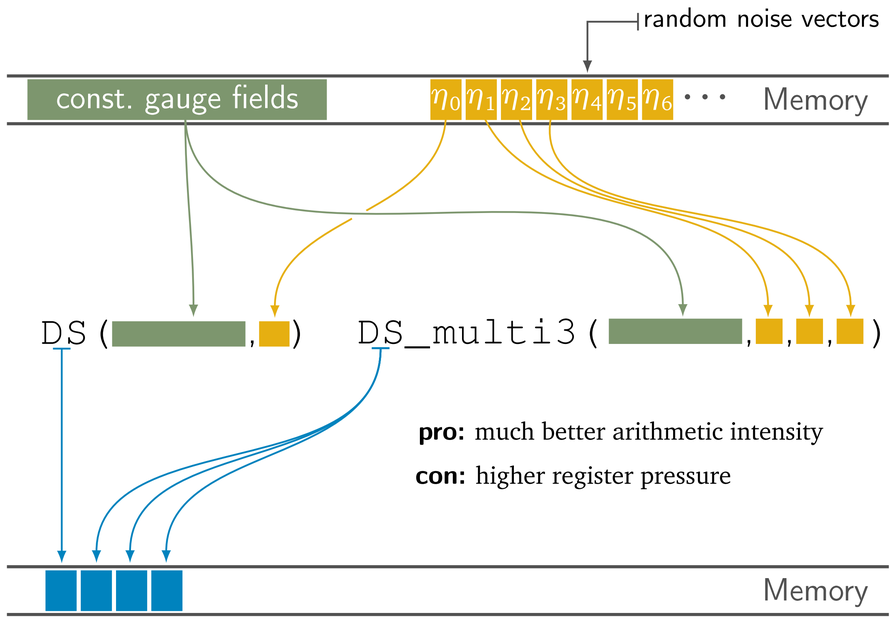
HISQ Link Smearing

$$c_{1S} \; \overset{\mu}{\underset{n}{\bullet\!\!\rightarrow\!\!\bullet}} \; + c_{3S} \sum_{\nu \neq \mu} \; + c_{5S} \sum_{\rho \neq \nu \neq \mu} \; + c_{7S} \sum_{\sigma \neq \rho \neq \nu \neq \mu} \; + c_{L} \sum_{\nu \neq \mu}$$

The HISQ action suppresses unphysical interactions of quarks by smoothing each link with a weighted sum of neighboring links referred to as smearing. It is mainly used in the force calculation for the Hybrid Monte Carlo algorithm and can take up to 40% of the total runtime.

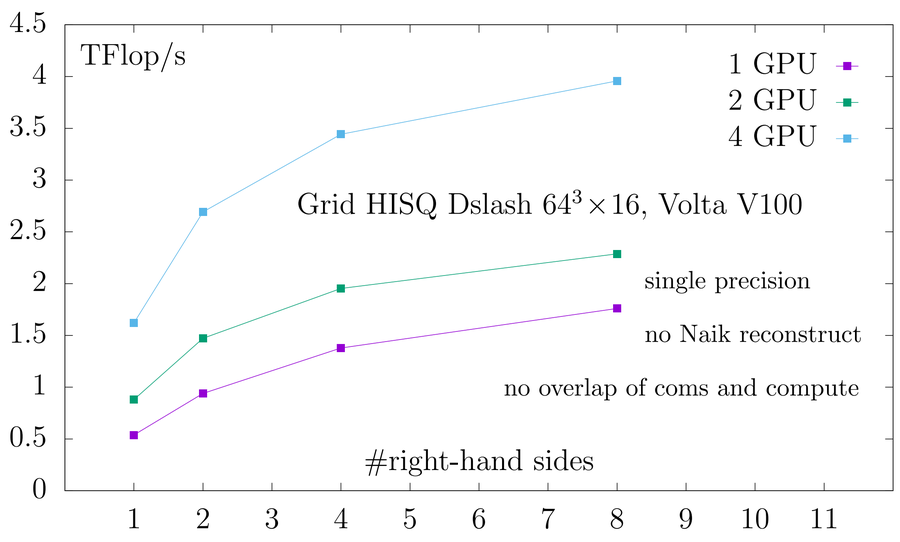


Multiple Right-hand Side Dslash



For many Lattice QCD applications, a large number of fermion matrix inversions are performed on a single gauge field. In order to exploit reuse of these gauge fields, we can apply the Dslash operation for multiple right-hand sides (rhs) at once. Increasing the number of rhs from one to four more than doubles the arithmetic intensity (Flop/byte) of the Dslash operation.

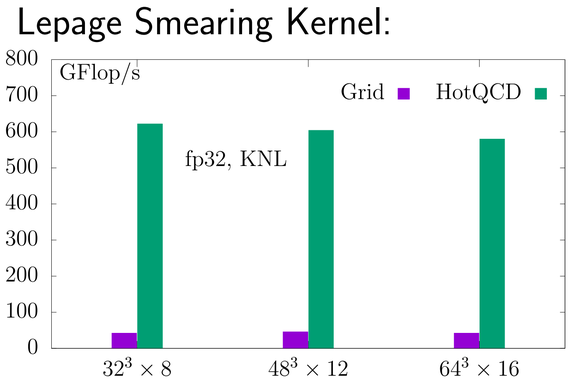
Multi-GPU Dslash Performance



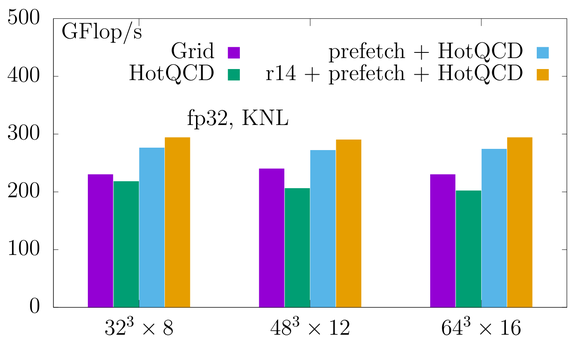
- Grid [1] available at: **github.com/paboyle/Grid**
  - architecture independent code
  - high and low level interface written in C++
  - support for multiple Lattice actions

Solvable Performance Problems

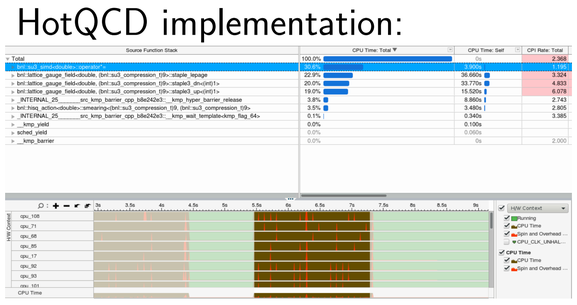
Lepage Smearing Kernel:



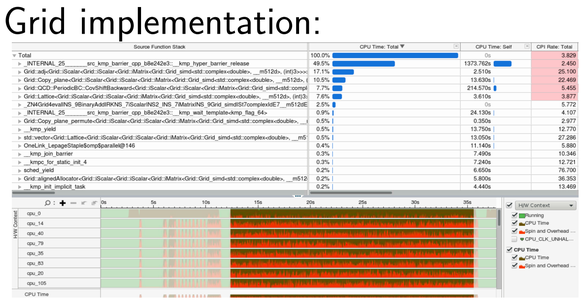
Dslash Kernel:



HotQCD implementation:

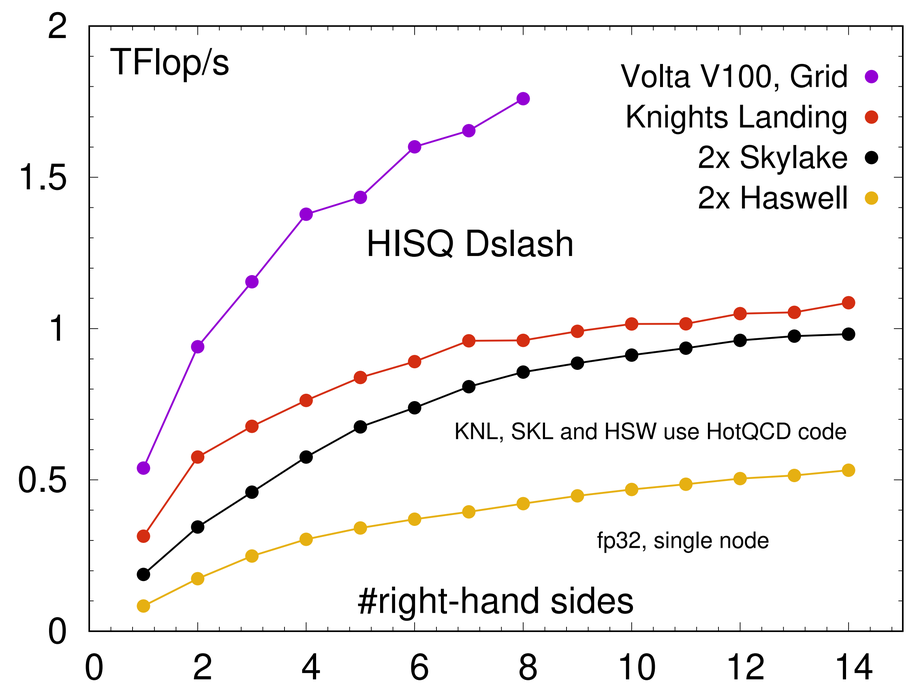


Grid implementation:



- Grid Dslash achieves optimal performance
- Grid Lepage smearing performance is limited by shortcomings of parallel transport approach
  - one parallel region for each link multiply in a staple
  - too many synchronization points
  - unnecessary loads and stores of intermediate results

Single-GPU Dslash Performance Comparison



#rhs	1	2	3	4	5	6
Flop/byte	0.73	1.16	1.45	1.65	1.80	1.91

[1] P. Boyle et al., Grid: A next generation data parallel C++ QCD library. arXiv:1512.03487

SciDAC Scientific Discovery through Advanced Computing

NESAP

ECP EXASCALE COMPUTING PROJECT