Usage Report for Fiscal Year 2017

Project Title:

Lattice QCD calculation of direct CP violation in kaon decays

Name: ○Christopher Kelly, Peter Boyle, Norman Christ, Taku Izubuchi, Chulwoo Jung

Laboratory at RIKEN: RIKEN Nishina Center for Accelerator-Based Science, RIKEN BNL Research Center, Computing Group
Other affiliations: Columbia University (CK, NC), University of Edinburgh (PB), Brookhaven National Laboratory (CJ,TI)

Introduction

The observable Universe contains much more matter than antimatter, but the reason for this remains a mystery. Such an asymmetry requires the breaking of the CP (Charge-Parity) symmetry. While mechanisms for this breaking are present in the Standard Model, it appears that the predicted amount is far too small to account for the observed asymmetry. This strongly suggests that new, Beyond the Standard Model physics awaits discovery.

Direct CP violation - the breaking of CP in particle decays - is heavily suppressed in the Standard Model and therefore offers a sensitive window to search for new physics. A precise measurement of the amount of CP violation in the decays of neutral kaons was performed in the late 1990s at CERN and FermiLab, but until recently no reliable theoretical determination existed with which it could be compared. The reason is that these decays have large contributions from low-energy QCD effects which cannot be reliably computed using traditional perturbative techniques.

In 2015 we produced the first direct Standard Model calculation of ε', the measure of direct CP-violation in kaon decays. We used lattice QCD, a technique for studying non-perturbative physics directly using simulations performed on massively parallel computers. Our result was compatible with experiment at the 2.1σ level, but the possibility of a tension has created significant interest in the physics community.

The error on our 2015 calculation is 30% with respect to the experimental measurement, and is divided roughly equally between statistical and systematic error. A significant effort is presently underway to improve the statistical errors by 2.5x by increasing the number of measurements from 216 to O(1500). On BW-MPC we aim to contribute significantly to this running. Similar efforts are underway to improve the systematic errors.

Code and Running

We perform our measurements using the CPS library built on top of Grid, a library for QCD calculations that takes maximal advantage of the SIMD, thread and node parallelism of modern supercomputers. The most computationally expensive aspect of our calculation is the inversion of the Dirac operator, an eight-point stencil operation acting between nearby lattice sites, using Krylov space methods. Our Dirac operator kernel is highly tuned for AVX512 machines, and on a single 64-core Intel KNL node achieves 330 Gflops sustained single-precision performance. This performance figure is significantly faster than the 240 Gflops we achieved on KNL at the time of our BW-MPC proposal, and results from carefully hand-coding the kernel using AVX512 intrinsics.

The Intel Skylake nodes employed by BW-MPC are both simpler and more complex than the KNL: while the processors themselves have more powerful cores with an architecture that is typically more friendly to optimize for, the use of two sockets per node introduces difficulties as the node performance of our stencil operation is heavily reliant upon strong intranode communication. With a traditional model of two MPI ranks per node we have observed significant performance reduction associated with communication via MPI through the memory system. This forced us to adopt a hybrid model whereby intranode communication is performed by directly reading and writing to a shared memory buffer, bypassing MPI, whereas off-node communications are performed using the traditional MPI route.

Even with the above optimization our performance on a single node of BW-MPC is presently only 250 Gflops, significantly lower than we had hoped. For a slightly different action we have also a hand-coded AVX512 assembly implementation of the kernel which performs at a more respectable 360 Gflops, hence there is hope that our kernel performance can be improved with further work.

We have also encountered significant performance issues with the Infiniband EDR network. On 128 nodes of BW-MPC our performance per node is only 80 Gflops, a 3x reduction over the single-node performance, resulting from poor network performance. This result is comparable to that on Cori II which has a supposedly weaker Cray Aries network.

Our measured MPI bandwidth is only 7 GB/s for single-precision communications, and 4 GB/s for half-precision (2x smaller packet size). The linear scaling of bandwidth with packet size suggests it is latency driven, a phenomenon we have also observed on machines with the Omnipath network. Unfortunately the result is that we do not gain any significant advantage through the use of mixed-precision algorithms with half-precision communication, which have benefited us considerably on Cori II.

As a result of these difficulties we have poured our effort into the development of a so-called 'split Grid' algorithm, whereby the many inversions necessary for our measurements are 'farmed out' to independent, smaller sub-partitions of the machine and run in parallel. This effort has now been completed for traditional single-rank-per-node machines, and has resulted in a 2.7x and 4x speed-up on 128 and 256 nodes of Cori II, respectively. Unfortunately additional effort was required to merge this work with our hybrid MPI/shared memory model for Skylake and as such we are only now able to begin benchmarking on BW-MPC.

Another unexpected issue is that, while the machine itself has 840 nodes, the present queue limits allow a maximum of only 128 nodes for a job. The contraction component of our calculation is very heavily memory dependent, and we have only recently succeeded in running within memory on 256 nodes of Cori II, the nodes of which have the same 96 GB as BW-MPC. To do so we needed to introduce an explicitly-managed block allocator on top of the existing, already sophisticated, distributed memory model, in order to suppress issues with memory fragmentation. Even with this optimization however it may prove to be challenging to perform the job on just 128 nodes. A related issue is the 24 hour job time limit, which coupled with the small number of nodes necessitates splitting our job by saving intermediate results and restarting; while this is something we have experience with, the correct partitioning will require some tuning.

Conclusion

In summary we have encountered a large number of unexpected issues that have thus far prevented us from running our measurements. Given the rapidly approaching end of our allocation and the fact that our goal of O(1500) configurations is now only weeks from completion due to our continued running on other machines, we feel it prudent to retarget our BW-MPC programme to aid our understanding of the systematic errors on the calculation.

As part of our calculation we perforce compute the I=0 $\pi\pi$ scattering phase shift. Our value of $23.8(5.0)^\circ$ resulting from the 2015 study is 2.8σ below the value of $38.3(1.3)^\circ$ obtained using dispersion theory combined with experimental input, and shows little sign of convergence despite significant statistical improvement. The origin of this discrepancy is not yet understood, and while there remains the potential for error in the dispersive calculation, it may instead arise due to an unexpected systematic error in our calculation; for example the existence of a nearby excited state. In order to understand this better we propose to measure the $\pi\pi$ correlation function with additional operators, which would enable the use of sophisticated methods such as the Generalized Eigenvalue approach to isolate the desired ground state as well as to study the spectrum of excited states.

On BW-MPC we intend perform the Dirac operator inversions for this study, saving the results to disk for later analysis. In this way we perform the bulk of the computation that actually requires a strongly parallel machine while both shortening the total job time and avoiding the complex memory usage pattern of the contractions stage. We can thus easily predict our total memory usage at peak will be 64 GB per node for a 128-node job, which will fit comfortably within the available 96 GB. This component of the job on 256 nodes of Cori II requires around 10 hours. With linear scaling this would then suggest around 20 hours per configuration, although in practise we expect it will be less due to better strong scaling at smaller node counts. However assuming 20 hours, the per-job cost is 102,400 core hours. Thus with our 13.6M core hours we can potentially perform 132 measurements, an amount sufficient for exploring and optimizing the choice of $\pi\pi$ operators.

Unfortunately the disk requirement for this modified proposal is quite significant; each configuration requires 1.125 TB of storage space, thus 150 TB total storage will be required. As such we request an allocation of 150 TB on the hierarchical storage solution as well as 54 TB of online storage for staging and temporary storage.

We intend to submit a Quick Use Project for FY18 to analyze the data we collect.