FINITE TEMPERATURE LATTICE QCD WITH GPUS*

Nuno Cardoso, Marco Cardoso, Pedro Bicudo

CFTP, Instituto Superior Técnico
Avenida Rovisco Pais, 1, 1049-001 Lisboa, Portugal

(Received July 27, 2011)

Graphics Processing Units (GPUs) are being used in many areas of physics, since the performance *versus* cost is very attractive. The GPUs can be addressed by CUDA which is a NVIDIA's parallel computing architecture. It enables dramatic increases in computing performance by harnessing the power of the GPU. We present a performance comparison between the GPU and CPU with single precision and double precision in generating lattice SU(2) configurations. Analyses with single and multiple GPUs, using CUDA and OPENMP, are also presented. We also present SU(2) results for the renormalized Polyakov loop, colour averaged free energy and the string tension as a function of the temperature.

DOI:10.5506/APhysPolBSupp.4.697
PACS numbers: 11.15.Ha, 12.38.Gc

1. Introduction

Since the first release of CUDA (Compute Unified Device Architecture) by NVIDIA, the GPUs (Graphics Processing Units) are being addressed for physics computing in different areas where the performance is relevant. CUDA gives developers access to the GPU by virtual instruction set and memory of computational elements. Whereas the CPU was projected for executing a single thread very quickly, the GPU architecture was projected to execute many concurrent threads slowly.

The most successful theories that describe elementary particle physics are the so-called gauge theories. SU(2) is an interesting gauge group, either to simulate the electroweak theory, or to use as a simplified case of the SU(3) gauge group of the strong interaction.

However, generating SU(N) lattice configurations is a highly computationally demanding task and requires advanced computer architectures such as CPU clusters or GPUs.

* Presented at the Workshop "Excited QCD 2011", Les Houches, France, February 20–25, 2011.

Nevertheless, GPUs are easier to access and maintain, as they can run on a local desktop computer, compared with CPU clusters.

This paper is divided in three sections. In Section 2, we present the performance results and the results of the Polyakov loop, the colour averaged free energy and the string tension as well as a brief description how to calculate them. For a more detailed description on how to generate lattice SU(2) configurations in GPUs see [1]. In Section 3, we conclude.

2. Results

We implemented our code in CUDA language to run in one GPU or in several GPUs with OPEMMP. The code was tested in two different architectures, NVIDIA 295 GTX and NVIDIA 480 GTX cards, see Table I.

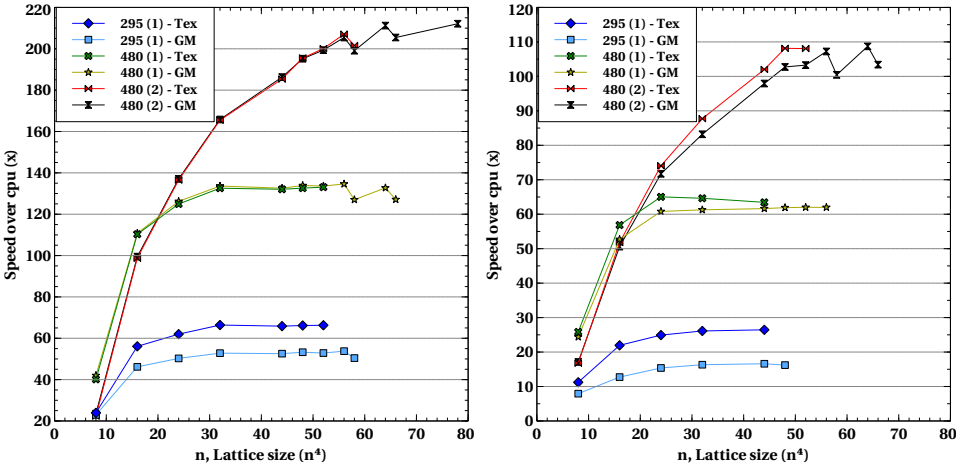
TABLE I

NVIDIA's architecture specifications (SM means Streaming Multiprocessor).

NVIDIA Geforce GTX	295 (GT200)	480 (Fermi)
Number of GPUs	2	1
CUDA capability	1.3	2.0
Number of cores	2×240	480
Global memory	896 MB per GPU	1536 MB
Number of threads per block	512	1024
Registers per block	16384	32768
Shared memory (per SM)	16KB	48KB or 16KB
L1 cache (per SM)	None	16KB or 48KB
L2 cache (per SM)	None	768KB
Clock rate	1.37 GHz	1.40 GHz

2.1. Performance

In order to test the GPU performance, we measure the execution time for the CUDA code implementation in one, two GPUs and the serial code in one CPU core (CPU Intel$^{(R)}$ Core$^{(TM)}$ i7 CPU 920, 2.67GHz, 8 MB of L2 Cache and 12GB of RAM) for different lattice sizes at $\beta = 6.0$ with random SU(2) matrix initialization followed by 100 iterations of the heat bath method and the calculation of the mean average plaquette at each iteration, see Fig. 1. For a more detailed overview see [1].



(a) Single precision. (b) Double precision.

Fig. 1: Performance results. 295 — NVIDIA Geforce 295 GTX; 480 — NVIDIA Geforce 480 GTX; (1) — with 1 GPU; (2) — with 2 GPUs; Tex — using textures; GM — using global memory.

2.2. Finite temperature

The Polyakov loop, $\langle L \rangle$, is an order parameter for the deconfinement transition, [2], it measures the free energy, F_q, of a single static quark at temperature T,

$$\langle L \rangle \propto \exp\left(-\frac{F_q}{T}\right) , \tag{1}$$

where T is connected to the lattice spacing a by $T = 1/(aN_t)$.

The results for the Polyakov loop, Fig. 2 (a), show a dependence on the extension of the lattice in time direction. This is due to the self-energy contribution of the static quark source used as order parameter.

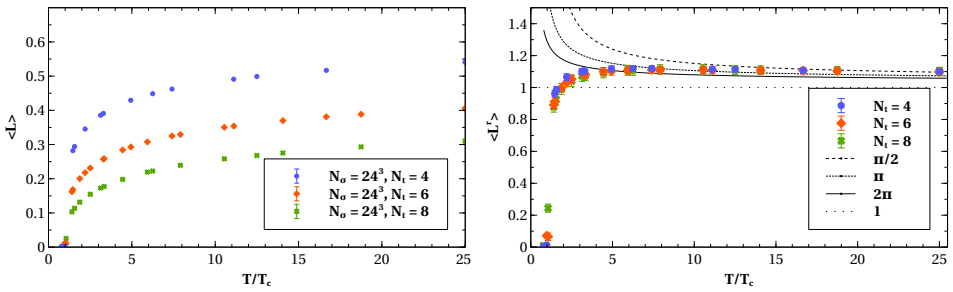
Elimination of this self energy term is necessary to obtain an order parameter which is a function of the temperature alone.

This can be done using the renormalization procedure described in [3] and using the values of [4] obtained for the effective potential as the seed values. The renormalized Polyakov loop can be written as

$$\langle L^{\mathrm{r}} \rangle = \left(Z\left(g^2\right)\right)^{N_t} \langle L \rangle , \tag{2}$$

where the renormalization constants $Z(g^2)$ should only depend on the bare coupling and fitting the values of $Z(g^2)$ obtained with this procedure with $Z(g^2) = \exp\left(A\, g^2 + B\, g^4\right)$, we obtain $A = 0.0637(18)$ and $B = 0.0731(16)$

with $\chi^2/\text{d.o.f.} = 1.16613$ for $g^2 < 1.3$. Applying this last results to all of our results in Fig. 2 (a), we obtain a renormalized Polyakov loop, Fig. 2 (b), which is independent of the extension of the lattice in the time direction. At high temperatures, the renormalized Polyakov loop approaches their corresponding HTL result.



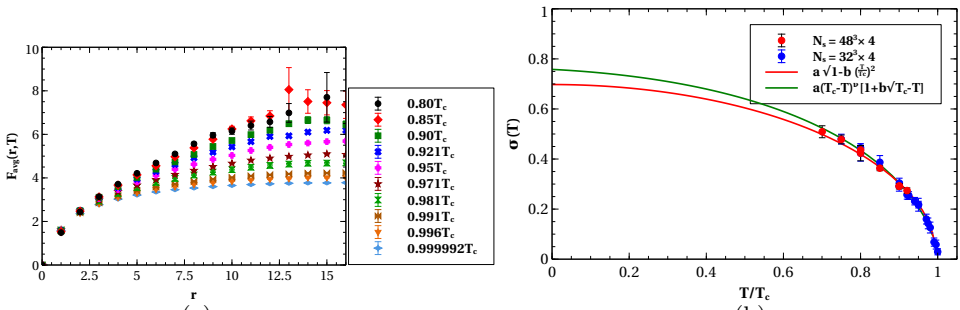
(a) (b)

Fig. 2: Polyakov loop. (a) — unrenormalized SU(2) Polyakov loop, $\langle L \rangle$, at finite temperature. (b) — SU(2) Polyakov loop renormalized, the dotted lines correspond to the pure gauge Polyakov loop in HTL perturbation theory for $\pi/2,\ \pi,\ 2\pi$.

The colour averaged free energy is defined as the correlation between two Polyakov loops,

$$e^{-F_{\text{avg}}(r,T)/T+C} = \tfrac{1}{4}\left\langle \text{Tr}\, L(y)\, \text{Tr}\, L^{\dagger}(x) \right\rangle \tag{3}$$

which is gauge invariant. To eliminate the trivial temperature dependence due to the colour trace normalization, we apply $F_{\text{avg}}(r,T) \to F_{\text{avg}}(r,T) - T\ln 4$. Fitting the $F_{\text{avg}}(r,T)$ data in Fig. 3 (a) with $F_{\text{avg}}(r,T)$ with $a_0(T) - \frac{a_1(T)}{r} + \sigma(T)r$, we show in Fig. 3 (b) the results for $\sigma(T)$ as a function of the temperature. Although the string tension in SU(2) was already addressed by [5], the number of data points is too low to have a clear overview. We fit our results with two different ansatz, $a\sqrt{1-b\,(T/T_{\text{c}})^2}$ and $a\,(T_{\text{c}}-T)^{\nu}[1 + b\sqrt{T_{\text{c}}-T}]$ and obtain a reasonable $\chi^2/\text{d.o.f.}$ for the both fits. For the first ansatz, we obtain $a = 0.6976 \pm 0.0176$, $b = 0.9990 \pm 0.0059$ and $\chi^2/\text{d.o.f.} = 0.732$. In the second, we fix $\nu = 0.63$ according the 3D Ising exponent for the correlation length and obtain $a = 1.5541 \pm 0.0435$, $b = -0.5122 \pm 0.0576$ and $\chi^2/\text{d.o.f.} = 0.598$. Nevertheless, we need more data for $T < 0.7\,T_{\text{c}}$.



(a) (b)

Fig. 3: (a) — SU(2) color averaged free energy, F_{avg}. (b) — SU(2) string tension, $\sigma(T)$.

3. Conclusions

With 2 NVIDIA GTX 480, we were able to obtain more than 200× the performance over one CPU core in single precision. It is not possible to generate SU(2) configurations using only the GPU shared memory due to the limited amount of shared memory available. The limited number of registers also affects the GPU performance. Using texture memory in this problem, we were able to achieve high performance, both in the GPU without cache memory and in the GPUs with cache memory. However, in the GPUs with cache memory the difference is bigger in double precision than in single precision. The occupancy and performance of the GPUs is strongly connected to the number of threads per block, registers per thread, shared memory per block, memory access, read and writing, patterns. To maximize performance it is necessary to ensure that the memory access is coalesced and to minimize copies between GPU and CPU memories.

The renormalized Polyakov loop for $N_t \geq 4$ shows very small dependence on the lattice time direction for $N_t = 4$ and low T. The string tension as a function of the temperature, $\sigma(T)$, extracted from the colour averaged free energy, for two different spatial lattice sizes does not reveal any volume dependence. The string tension for $T > T_c$ is zero, however for $T < T_c$ is temperature dependent. We fit the string tension with two different ansatz, however, we need more data for $T < 0.7\,T_c$. Future work will be dedicated to the study of this case.

This work was financed by the FCT contracts POCI/FP/81933/2007, CERN/FP/83582/2008, PTDC/FIS/100968/2008 and CERN/FP/109327/2009. Nuno Cardoso is also supported by the FCT under the contract SFRH/BD/44416/2008.

REFERENCES

[1] N. Cardoso, P. Bicudo, *J. Comput. Phys.* **230**, 3998 (2011) [arXiv:1010.4834 [hep-lat]].

[2] L.D. McLerran, B. Svetitsky, *Phys. Rev.* **D24**, 450 (1981).

[3] S. Gupta, K. Huebner, O. Kaczmarek, *Phys. Rev.* **D77**, 034503 (2008) [arXiv:0711.2251 [hep-lat]].

[4] G.S. Bali *et al.*, *Phys. Rev. Lett.* **71**, 3059 (1993) [arXiv:hep-lat/9306024].

[5] S. Digal, S. Fortunato, P. Petreczky, *Phys. Rev.* **D68**, 034008 (2003) [arXiv:hep-lat/0304017].