PUBLISHED VERSION

http://hdl.handle.net/2440/77575

GPUs: An Oasis in the Supercomputing Desert

Waseem Kamleh[*]

University of Adelaide, Australia

E-mail: waseem.kamleh@adelaide.edu.au

A novel metric is introduced to compare the supercomputing resources available to academic researchers on a national basis. Data from the supercomputing Top 500 and the top 500 universities in the Academic Ranking of World Universities (ARWU) are combined to form the proposed "500/500" score for a given country. Australia scores poorly in the 500/500 metric when compared with other countries with a similar ARWU ranking, an indication that HPC-based researchers in Australia are at a relative disadvantage with respect to their overseas competitors. For HPC problems where single precision is sufficient, commodity GPUs provide a cost-effective means of quenching the computational thirst of otherwise parched Lattice practitioners traversing the Australian supercomputing desert. We explore some of the more difficult terrain in single precision territory, finding that BiCGStab is unreliable in single precision at large lattice sizes. We test the CGNE and CGNR forms of the conjugate gradient method on the normal equations. Both CGNE and a modified form of CGNR (with restarts) provide reliable convergence for quark propagator calculations in single precision.

[*]Speaker.

1. The 500/500 Metric for Academic HPC Resources

Lattice QCD has traditionally and continues to be one of the most computationally demanding research fields within quantitative science. Progress in Lattice QCD has closely tracked advances in high performance computing (HPC). It is unsurprising then that the semi-annual supercomputing Top 500 list[1] is closely watched by many researchers within lattice QCD. The Top 500 provides a straightforward answer to those wanting to know which country has the biggest and the best machines. It is arguable that such a simple comparison is not always the most relevant. In certain circumstances, it may be more pertinent to ask a different question: *How much supercomputing access do I have relative to my competitors overseas?*

In an attempt to provide an answer, our starting point is the Academic Ranking of World Universities (ARWU) list compiled by Shanghai Jiao Tong University, China, also known as the Shanghai Ranking[2]. This survey lists the top 500 ranked universities in the world, which we shall simply refer to as the ARWU 500. Table 1 lists the top 6 countries, as ranked by the Academic Ranking of World Universities (ARWU) in 2012. The national rankings are determined in a similar manner to those based on the Olympic medal tallies. Countries are first ranked in descending order by the number of university entries they have in the ARWU Top 20, then by the number of Top 100 universities, followed by the number of Top 200, 300, 400 and 500 entries respectively.

COUNTRY	Top 20	Top 100	Top 200	Top 300	Top 400	Top 500
USA	17	53	85	109	137	150
UK	2	9	19	30	33	38
Japan	1	4	9	9	16	21
Australia	–	5	7	9	16	19
Germany	–	4	14	24	30	37
Canada	–	4	7	17	18	22

Table 1: Top 6 countries, as ranked by the Academic Ranking of World Universities (ARWU) in 2012.

These 6 countries will form the basis of our study of the HPC resources to available to academics in Australia, in comparison to our overseas competitors. The list includes Japan, Germany and the USA, the traditional leaders of the supercomputing field. Canada has broadly similar socioeconomic characteristics to Australia and hence provides a useful point of comparison.

We now turn our attention to the June 2012 Top 500 Supercomputer list. We filter the Top 500 supercomputing data by restricting ourselves to the aforementioned top 6 countries in the ARWU ranking. The top 3 entries for each country in the Academic and Research segments of the Top 500 supercomputer list are displayed in Table 2. Also shown are the total number of entries, number of compute cores, and combined computing power for all Academic/Research entries in the list for that country. The quantity that we will be interested in is the combined R_{max} value for each country, which is an indicator of the total number of Teraflops available to the Academic/Research segments in that country. R_{max} is the LINPACK benchmark and provides a measure of the supercomputer's speed in Teraflops.

---

[1]http://top500.org/
[2]http://www.shanghairanking.com/

| RANK	COUNTRY/SITE	N_{cores}	R_{max}	R_{peak}
	Australia			
31	VLSCI/Avoca	65536	690.2	838.9
139	NCI-NF/Vayu	11936	126.4	139.9
248	iVEC	9600	87.2	107.5
	Total: 3 Academic/Research entries	**87072**	**903.8**	**1086.3**
	Canada			
66	SciNet/U. Toronto/Compute Canada/GPC	30912	261.6	312.82
71	Calcul Canada/Calcul Québec/Sherbrooke	37728	240.3	316.9
90	Environment Canada	8192	185.1	251.4
	Total: 9 Academic/Research entries	**137872**	**1342.5**	**1751.3**
	Germany			
4	Leibniz Rechenzentrum/SuperMUC	147456	2897.0	3185.1
8	Forschungszentrum Juelich/JuQUEEN	131072	1380.4	1677.7
25	Forschungszentrum Juelich/JUGENE	294912	825.5	1002.7
	Total: 16 Academic/Research entries	**753944**	**7062.6**	**8471.0**
	Japan			
2	RIKEN/K computer	705024	10510.0	11280.4
12	IFERC/Helios	70560	1237.0	1524.1
14	GSIC/Tokyo Inst. of Tech./TSUBAME 2.0	73278	1192.0	2287.6
	Total: 23 Academic/Research entries	**1184258**	**17089.0**	**20430.9**
	United Kingdom			
13	STFC/Daresbury Laboratory/Blue Joule	114688	1207.8	1468.0
20	U. Edinburgh/DiRAC	98304	1035.3	1258.3
32	U. Edinburgh/HECToR	90112	660.2	829.0
	Total: 16 Academic/Research entries	**455584**	**5875.3**	**7553.0**
	United States			
1	DOE/NNSA/LLNL/Sequoia	1572864	16324.8	20132.7
3	DOE/SC/Argonne/Mira	786432	8162.4	10066.3
6	DOE/SC/Oak Ridge/Jaguar	298592	1941.0	2627.6
	Total: 87 Academic/Research entries	**5063813**	**44953.9**	**56928.4**

Table 2: Selected entries in the June 2012 Top 500 Supercomputer list in the Academic and Research segments. The top 3 entries are listed for each of the chosen countries, as well as the total number of entries and the aggregate computing capacity of the entries. N_{cores} is the number of compute cores. R_{max} (the LINPACK benchmark score) and R_{peak} (the theoretical peak) are in Teraflops.

The most straightforward measure of the supercomputing power available to researchers in a given country would be to compare the integrated R_{max} values in the academic segment. However, this simple measure doesn't reflect the level of competition for those resources. In order to provide

a better estimate of the HPC resources available to a given research group, we propose a novel measure called the 500/500, which is calculated for each country by taking the combined Teraflops of the Academic and Research entries in the Top 500 supercomputer list and dividing by the number of institutions in the ARWU 500. A summary of the data is presented in Table 3. The measure assumes that the number universities in the ARWU 500 is a good representation of the number of academic supercomputing groups in the country.

COUNTRY	TOP 500	TOTAL R_{max}	ARWU 500	500/500
Australia	3	903.8	19	47.6
Canada	9	1342.5	22	61.0
Germany	16	7062.6	37	190.9
Japan	23	17089.0	21	813.8
UK	16	5875.3	38	154.6
USA	87	44953.9	150	299.7

Table 3: Data of interest for the selected countries in 2012. Listed are the number of Academic/Research Top 500 entries, the combined R_{max} (in Teraflops) under the Academic and Research segments, the number of ARWU 500 entries and our proposed 500/500 measure of academic HPC resources (in Tflops/institution).
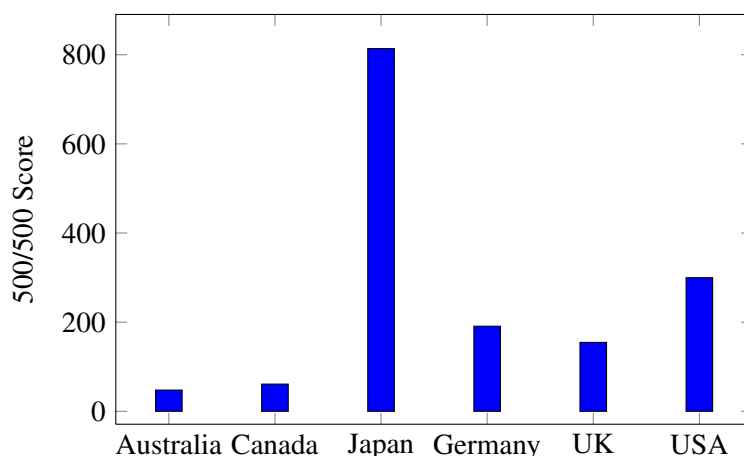
Figure 1: The 500/500 scores (in units of Tflops/institution) for the selected countries in 2012.

As demonstrated in Figure 1, Japanese researchers are the clear winners, with 500/500 score more than double that of second-placed USA, and nearly twenty times that of Australia! While the USA easily has the highest integrated R_{max} score, they are ranked second on the basis of their 500/500 score, a reflection of the intense competition for those resources as indicated by their place at the top of the ARWU 500. Of the six selected countries, Australia ranks last according to the 500/500 metric.

2. GPU Computing

As demonstrated in the previous section, Australian researchers are disadvantaged with regard to HPC resources when compared to our overseas competitors. The lack of HPC resources is par-

Architecture	GPU	Cores	Peak (SP)	Peak (DP)	ECC
Fermi	GTX 580	512	1581 Gflops	166 Gflops	No
	Tesla M2090	512	1331 Gflops	665 Gflops	Yes
Kepler	GTX 680	1536	3090 Gflops	95 Gflops	No
	Tesla K20	2496	3520 Gflops	1170 Gflops	Yes

Table 4: Previous (Fermi) and current (Kepler) generation NVIDIA GPUs. Shown are the number of CUDA cores, the peak floating point performance in single and double precision, and the ECC memory capability.

ticularly acute in our field of Lattice QCD, where some of our competitors have access to dedicated lattice machines capable of hundreds of Teraflops.

The ILDG program allows for the sharing of gauge field configurations within a group or with the lattice QCD community at large[1]. The PACS-CS collaboration in Japan generously released several gauge field ensembles of large volume and light quark mass suitable for cutting edge calculations to the general lattice community[2]. Through the use of these configurations we have been able to bypass the unaffordable gauge field generation process and devote our limited cycles towards the production of quark propagators.

It should come as no surprise that with the relatively scarce level of HPC resources available to us when compared to our competitors, we have turned to GPUs as a cost-effective way of competing with overseas groups. Lattice QCD has a geometric parallelism that makes it ideally suited to be put on GPUs[3, 4]. NVIDIA has two distinct GPU product lines that are relevant to HPC. The Tesla line of cards specifically targets HPC users, whereas the commodity GeForce graphics cards target the much bigger computer gaming market. The specific cards that we are interested in are listed in Table 4. As we can see, in comparison to the GTX cards, the Tesla GPUs feature improved double precision performance and ECC memory. These features come at a cost however, with a Tesla card costing roughly 4 times as much as a top-end GTX card.

Fortunately, the numerical requirements for quark propagator generation are much less strict than those for gauge field generation. The need to preserve unitarity during gauge field generation typically requires double precision, and as the generated gauge fields are not easily "checked" one also requires ECC memory. In contrast, for quark propagators the tolerance when calculating the application of the fermion matrix inverse is typically $\sim 10^{-5}$, which means single precision is sufficient. Furthermore, the solution to the linear system is easily verified, avoiding the need for ECC memory. Hence, GTX cards are perfectly viable for quark propagator calculation.

3. Adventures in Single Precision

To obtain the action of the the inverse fermion matrix D^{-1} on a vector we calculate the solution to the linear system

$$D\mathbf{x} = \mathbf{b}. \tag{3.1}$$

As the fermion matrix D is non-Hermitian the most common algorithm for obtaining the solution is BiCGStab[5] or some variant thereof. In double precision BiCGStab usually converges to a solution, even though the typical convergence is not smooth but rather 'spiky'. However, in single

precision we find that BiCGStab is numerically unstable. When attempting to invert the fermion matrix on large lattices and light quark masses BiCGStab frequently fails to converge. To avoid this, we propose to use an algorithm that minimises the residual and hence will converge smoothly.

The conjugate gradient (CG) algorithm[6] minimises the residual, but is only applicable to cases where the matrix being inverted is Hermitian positive-definite (Hpd). There are two simple ways to convert our original problem into a form suitable for the CG algorithm. The first is to simply multiply (3.1) by $D^\dagger$ to obtain the CGNR form of the normal equations,

$$D^\dagger D\mathbf{x} = D^\dagger \mathbf{b}. \tag{3.2}$$

The second is to solve the CGNE form of the normal equations

$$DD^\dagger \mathbf{x}' = \mathbf{b}, \tag{3.3}$$

where the solution to the original equations is given by $\mathbf{x} = D^\dagger \mathbf{x}'$.

When solving the CGNE form of the normal equations, the residual for the normal form $|DD^\dagger \mathbf{x}' - \mathbf{b}|$ and the residual for the original form $|D\mathbf{x} - \mathbf{b}|$ coincide by construction, so when CGNE converges we have obtained the solution to the original equation to the desired tolerance δ_{tol}. Furthermore, we find that even in single precision the estimated residual $|\mathbf{r}|$ and the true residual coincide for the CGNE process.

In double precision, when the CGNR process converges this usually implies that we have obtained the desired solution. However, in single precision, the solution to (3.2) converges well before we have obtained the solution to (3.1). To work around this, we propose a simple modification of the CGNR process. When the CGNR normal equation converges with tolerance δ_{ne}, check if we have a solution to the original equation within δ_{tol}. If not, adjust δ_{ne} and restart CGNR with the current solution. Our modified CGNR algorithm with restarts is presented in Figure 2.

A comparison of the typical behaviour of CGNE and our CGNR with restarts is shown in Figure 3. We can see that the estimated residual $|\mathbf{r}_{\text{ne}}|$ and the true residual for the CGNR normal equations $\varepsilon' = |D^\dagger(D\mathbf{x} - \mathbf{b})|$ coincide until the CGNR system (3.2) has converged, after which they diverge due to hitting the limits of single precision. What is interesting is that even though the CGNR process undergoes restarts, the true residual for the original system $\varepsilon = |D\mathbf{x} - \mathbf{b}|$ decreases smoothly until it has converged to within the desired tolerance. Tests comparing CGNR and CGNE were performed at several quark masses and we found that the modified CGNR process (with restarts) consistently converges significantly faster than CGNE, requiring $\sim 10\% - 30\%$ less iterations to reach the desired tolerance.

References

[1] M. G. Beckett, B. Joo, C. M. Maynard, D. Pleiter, O. Tatebe, *et al.*, *Building the International Lattice Data Grid*, *Comput.Phys.Commun.* **182** (2011) 1208–1214, [0910.1692].

[2] **PACS-CS** Collaboration, S. Aoki, K.-I. Ishikawa, N. Ishizuka, T. Izubuchi, D. Kadoh, K. Kanaya, Y. Kuramashi, Y. Namekawa, M. Okawa, Y. Taniguchi, A. Ukawa, N. Ukita, and T. Yoshié, $2+1$ *flavor lattice QCD toward the physical point*, *Phys. Rev. D* **79** (Feb, 2009) 034503. arXiv:0807.1661 [hep-lat].

[3] G. I. Egri, Z. Fodor, C. Hoelbling, S. D. Katz, D. Nogradi, *et al.*, *Lattice QCD as a video game*, *Comput.Phys.Commun.* **177** (2007) 631–639, [hep-lat/0611022].

Initialise $\delta_{\text{ne}} := \delta_{\text{tol}}$ to the desired solution tolerance.
loop
    Set $\mathbf{y} := \mathbf{r}_{\text{ne}} := D^\dagger D \mathbf{x} - D^\dagger \mathbf{b}, \quad \rho := |\mathbf{r}_{\text{ne}}|^2$.
    while $\sqrt{\rho} > \delta_{\text{ne}}$ **do**
        Set $\beta := \langle \mathbf{y}, D^\dagger D \mathbf{y} \rangle, \quad \omega := \rho/\beta$.
        Set $\mathbf{x} := \mathbf{x} + \omega \mathbf{y}, \quad \mathbf{r}_{\text{ne}} := \mathbf{r}_{\text{ne}} - \omega D^\dagger D \mathbf{y}$.
        Set $\rho' := \rho, \quad \rho := |\mathbf{r}_{\text{ne}}|^2, \quad \theta := -\rho/\rho'$.
        Set $\mathbf{y} := \mathbf{r}_{\text{ne}} - \theta \mathbf{y}$.
    end while
    Set $\varepsilon := |D\mathbf{x} - \mathbf{b}|$ to the true residual for the original equation.
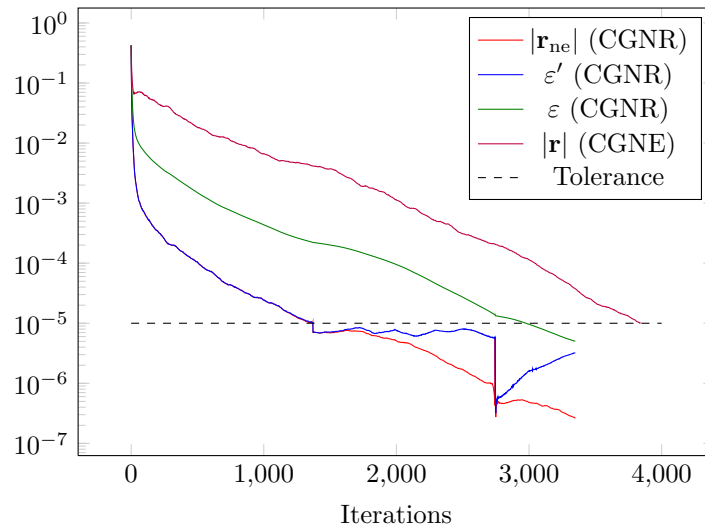   **if** $\varepsilon < \delta_{\text{tol}}$ **then exit** {*We are finished.*}
    Set $\varepsilon' := |D^\dagger D \mathbf{x} - D^\dagger \mathbf{b}|$ to the true residual for the normal equation.
    Update $\delta_{\text{ne}} := \tau \cdot \delta_{\text{tol}} \cdot (\varepsilon'/\varepsilon)$. {*Restart CGNR.*}
end loop

Figure 2: The modified CGNR algorithm with restarts. The constant $\tau \sim 0.9$ controls the restart frequency.

Figure 3: Typical behaviour of the CGNE process and the CGNR process with restarts. Shown for CGNR are the estimated residual $|\mathbf{r}_{\text{ne}}|$ and true residual ε' for the normal equation, as well as the true residual ε for the original equation. For CGNE we show the estimated residual $|\mathbf{r}|$ (which coincides with the true residual).

[4] M. Clark, R. Babich, K. Barros, R. Brower, and C. Rebbi, *Solving Lattice QCD systems of equations using mixed precision solvers on GPUs*, Comput.Phys.Commun. **181** (2010) 1517–1528, [0911.3191].

[5] H. A. van der Vorst, *Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems*, SIAM J. Sci. and Stat. Comput. **13(2)** (1991) 631–644.

[6] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of research of the National Bureau of Standards **49** (1952) 409–436.