UM-P-92/94
OZ-92/34
November, 1992

UNRAVELLING THE STRUCTURE OF MATTER ON HIGH-PERFORMANCE COMPUTERS

T D Kieu* and B H J McKellar
School of Physics,
University of Melbourne,
Parkville VIC 3052

Abstract

The various phenomena and the different forms of matter in Nature are believed to be the manifestation of only a handful set of fundamental building blocks —the elementary particles— which interact through the four fundamental forces. In the study of the structure of matter at this level one has to consider forces which are not sufficiently weak to be treated as small perturbations to the system, an example of which is the strong force that binds the nucleons together. High-performance computers, both vector and parallel machines, have facilitated the necessary non-perturbative treatments. We introduce and explain the principles and the techniques of computer simulations applied to Quantum Chromodynamics, the widely accepted theory of strong interactions, to calculate, among many other things, the mass of nucleons and their decay rates. Some commercial and special-purpose high-performance machines for such calculations are also mentioned.

Introduction

The various simple and complicated phenomena in Nature —from that of falling rocks to those giving rise to conscious life— over a big range of length scales —from that of the entire universe to the tiny scale of the elementary particles, the fundamental building blocks of matter as can be resolved at present-day high-energy microscopes— are the consequences of four fundamental interactions. In the increasing order of strength, they are the gravitational, the weak (nuclear), the electromagnetic, and the strong (nuclear) forces. Those forces are not only different in the strength, the range but also in the kinds of particles that experience them. All matter is subjected to the long-range gravitational forces and the 'combined' electromagnetic and weak interactions. But only *quarks* can feel

*kieu@tauon.ph.unimelb.edu.au; tel: (03) 344 5435; fax: (03) 347 4783.

the strong interactions which bind them together, forming nuclear entities like protons and neutrons.

The framework of gauge theories has enabled a description, and unification of those forces. *Quantum Electrodynamics* (QED) is the fine example of gauge theories. It describes the interactions between charged particles such as electrons through the fields of yet another kind of particles known as the gauge particles. In this example, the gauge particles are the photons.

The gauge theory for the strong interactions, on the other hand, is *Quantum Chromodynamics* (QCD). The gauge particles of this theory are the *gluons* gluing together the quarks which are characterised by a different kind of charges, namely the colour charges. Each quark participates through its colour charge, either blue, red or green. Unlike QED where the photons are neutral and do not interact with each other directly, the gluons do carry some colour charges and hence can feel the direct influence of other gluons. The interactions are thus very non-linear. But the distinction from other forces that makes it difficult to treat the strong force is in its intense strength. The force is so strong that we cannot observe quarks as individual but only their composite products called *hadrons* like the nucleons and mesons, until we reach for a much higher energy analyser or a higher temperature. The mathematics of perturbative approximations, where the interactions are treated as small disturbances to the otherwise free particles, is not applicable here. Non-perturbative techniques are thus called for; and among those available, computer simulations offer the most systematic approach from the first principles and have been exploited extensively.

We next introduce the mathematical framework leading to the computer simulations. Some recent results are presented for illustrations. We also review some high-performance machines, both commercial and home-built, which have been available to or are going to be built for QCD tasks.

The mathematical framework

Quantum physics takes into account the extra fluctuations which are negligible at everyday scale but at atomic and smaller scale become dominantly important. Those quantum fluctuations are most clearly displayed in the *path-integral* formulation of *quantum field theory* where the integrating variables are functions of space and time describing the field configurations. Thus not only the classically dominant configurations but others are also included with weightings appropriate to their importance at the scale.

All the experimentally measurable quantities can be expressed in terms of certain averages in the path integral with the weights proportional to a phase factor whose argument is the *action* of that particular field configuration. (The action completely determines the theory.)

$$< Q > \; = \; \frac{1}{Z} \int Q[\phi(x)] \exp\{iS'[\phi]\} D\phi(x); \qquad (1)$$

Q is the quantity to be averaged, $\mathcal{Z}$ the normalising factor, $\mathcal{S}'$ the action, and $D\phi(x)$ denotes the functional measure. The formal measure needs to be defined more precisely in practice.

One approach is to approximate the continuous space and time by some regular or random lattice grid on which the fields reside. (To be precise, matter fields describing the elementary particles reside on the lattice sites; and the gauge fields on the links in order to preserve certain nice properties of gauge interactions.) In so doing the path integral now turns into ordinary multi-dimensional integral, with dimensions equal to the total number of sites and links of the lattice. By making the lattice spacing small enough, but still non-zero, continuum physics can be extracted in certain manner. This *Lattice Field Theory* approach is suitable for computer implementation and is the one we will take here.

The multi-dimensional integral is still prohibitively large to be evaluated directly, even on a modest 16^4-site hypercubic lattice. Simple random Monte Carlo sampling to choose a finite set of configurations approximating the integral does not work because the factor $\exp\{i\mathcal{S}\}$ fluctuates too much. This led to the idea of *importance sampling* [1] where the field configurations $\{\phi(x_{\text{lattice}})\}$ are selected according to certain probability. And in order to interpret the weights in (1) as probability, we use the trick of imaginary time in transforming $t \rightarrow it$, upon which $\exp\{i\mathcal{S}'\} \rightarrow \exp\{-\mathcal{S}\}$, to obtain real and positive weight factors.

The numerical techniques

If the configurations $\{\phi_i\}$, $i = 1, C$, are selected with some probability $P(\phi_i)$ then the functional average (1) can be approximated as

$$< Q > \approx \frac{\sum_{i=1}^{C} Q(\phi_i) P^{-1}(\phi_i) \mathrm{e}^{-\mathcal{S}(\phi_i)}}{\sum_{i=1}^{C} P^{-1}(\phi_i) \mathrm{e}^{-\mathcal{S}(\phi_i)}}. \tag{2}$$

It is possible to construct a random walk in the configurational space via a Markov process such that after many steps $P(\phi_i) \rightarrow P_{\text{eq}}(\phi_i)$. If the thermally equilibrated distribution is chosen as $P_{\text{eq}}(\phi_i) \sim \exp\{-\mathcal{S}(\phi_i)\}$, the average (2) is then reduced to a simple form

$$< Q > \approx \frac{1}{C} \sum_{i=1}^{C} Q(\phi_i). \tag{3}$$

The Markov process is determined by a *transition probability*, $W[\{\phi\} \rightarrow \{\phi'\}]$, from one configuration to another, and is characterised by a 'lack of memory' i.e. the probability of reaching the new configuration depends only on the preceding one. Imposing the condition of *detailed balance*,

$$P_{\text{eq}}(\phi) W[\{\phi\} \rightarrow \{\phi'\}] = P_{\text{eq}}(\phi') W[\{\phi'\} \rightarrow \{\phi\}], \tag{4}$$

is then sufficient (but not necessary) to ensure $P \to P_{\rm eq}$. This does not completely specify the transition probability; a common choice is the *Metropolis algorithm*

$$W[\{\phi\} \to \{\phi'\}] \;=\; \begin{cases} \exp\{-\partial\mathcal{S}\} & \text{if } \partial\mathcal{S} = \mathcal{S}(\phi') - \mathcal{S}(\phi) > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

As the lack of memory of the Markov process allows, an initial configuration is chosen arbitrarily first. Another configuration is then generated from this by some random modifications. The above algorithm is used to either accept or reject the configuration. If it is accepted then from it the next configuration is generated, and so on. A sequence of thermally equilibrated configurations, after many steps, is thus obtained for measurements. It is more efficient to average only over widely separated configurations since successive configurations in the sequence are highly correlated. It should also be mentioned that closer to the regime where continuum physics can be extracted, it is difficult to reach thermal equilibrium and time-consuming to generate uncorrelated configurations, an effect called *critical slowing down*.

The motivation for computer simulation is that it provides some controlled approximation to the non-perturbative solution, as some errors always afflict all calculations. It is therefore desirable to identify and analyse the sources of systematic errors. The typical size of a finite sample of configurations to represent the path integral rarely exceeds few hundred 'independent' configurations. Since the statistical error only falls as $1/\sqrt{\#\text{configurations}}$, substantial algorithmic improvements are needed to efficiently produce decorrelated configurations and to make a significant reduction in this error.

Another source of error is the failure to achieve fine enough lattice as this requirement of small lattice spacing results in small overall volume of the lattice when the total number of lattice sites is bounded above by limitations of the computer hardware. In many current simulations, particles are squeezed into a box smaller than their size: linear dimension of the box is typically in the region of about 2 fm (the smallest lattice spacing so far is 0.06 fm). A desirable goal would be something like 5 fm before some reliable finite-volume extrapolation method can be applied to derive the physics in infinite extent.

One further obstacle is that of the quark fields. Because of their peculiar statistical properties, the generation of lattice field configurations is a highly non-local problem, and consequently very time-consuming. The *quenched* approximation, where the non-valence quarks are discarded as if they had infinitely large masses, is usually adopted to reduce the computing time to a more manageable size. This is simply an expedient, with little or no physical justification, and is yet another source of systematic errors. And only recently *dynamical quarks* have been fully incorporated into QCD simulations thanks to some progress in exact algorithms. Nevertheless, other kinds of errors are presently too big for the 'defects' of the quenched approximation to be detected when compared to full simulations.

Those quark matter fields are also beset by a long-standing problem of the extraneous generation of species on the lattice. As the continuum is approached, certain desirable lattice theories give a result that is an integer multiple of the expected continuum result,

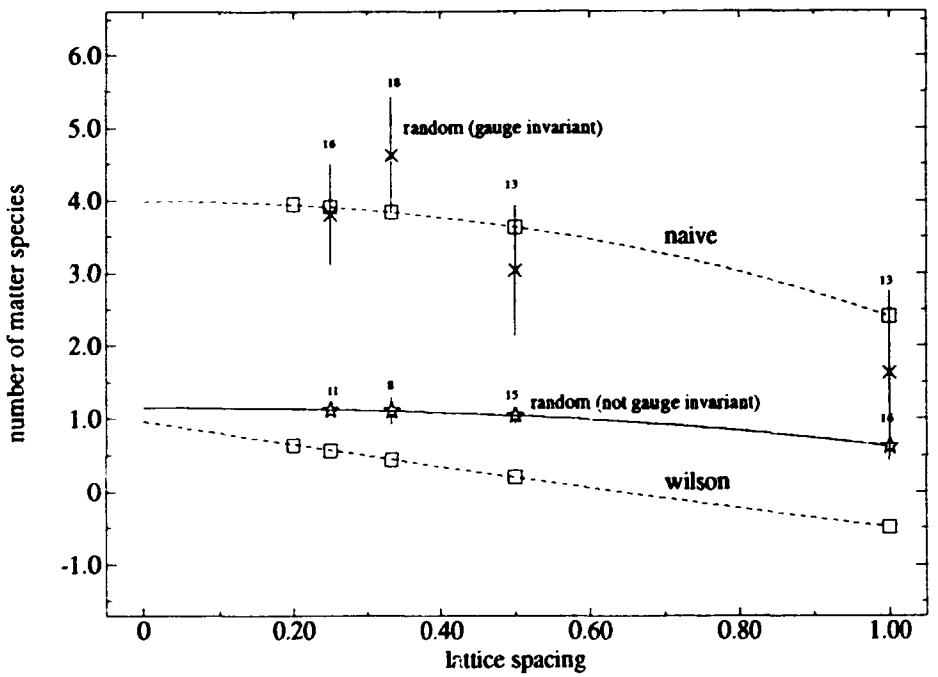

Figure 1: Some lattice formulations generate more than one species of matter fields. Those having the right number of species, on the other hand, lack certain desirable properties.

as if there are multiple quark species on the lattice. Figure 1, from Griffin and Kieu in [3], gives an illustration of this *doubling* problem. Far from being a technicality only inherent to the lattice approximation, this long-standing problem lies deep at the heart of quantum field theory. However, the classes of results given in the next section are not seriously afflicted by this as for them there are ways to suppress the doubling.

Some recent results

Among many lattice QCD results deriving various experimentally confirmed nuclear processes and predicting new phenomena yet to be observed, we can only mention in the restricted space here the two classes concerning the mass spectrum of hadrons and their decays. These are the simplest calculations but yet interesting and crucial to demonstrate directly the value and power of QCD without any model-dependent parameters to be fiddled with. A truly remarkable feature of QCD is that the theory has no dimension*ful* parameter, but yet from a single dimension*less* parameter it can give rise to physical scales of mass and length. Appropriately this property is named *dimensional transmutation.*

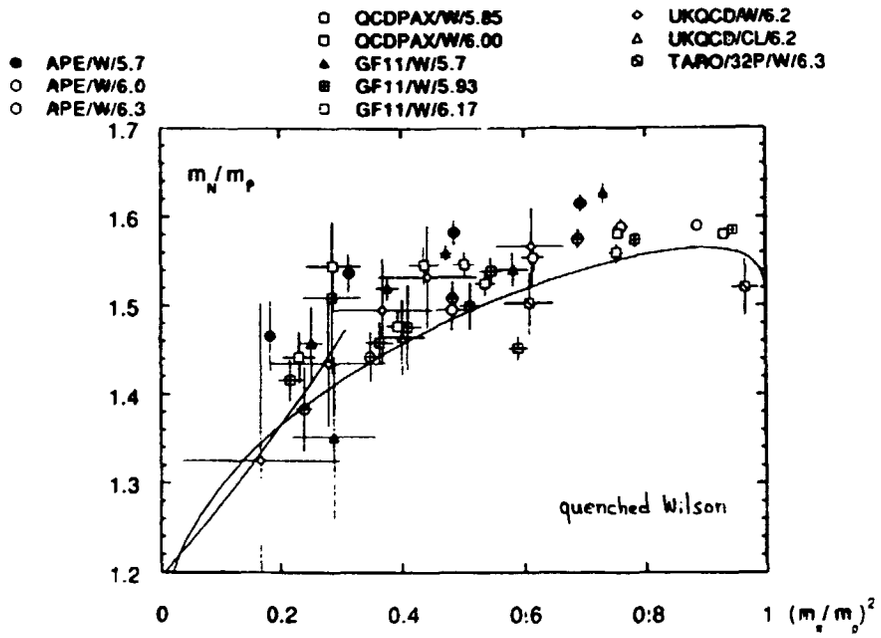

Figure 2: Results from various research groups show the trend towards the experimental point at the intersection at the lower left-hand corner [coordinate value $(0.03, 1.22)$] as the quark masses are tuned to their vanishingly small values.

Hadron mass

To measure the mass of the proton, say, we put the proton represented by its three valence quarks at some point on the lattice then measure at some distance away in the time direction the 'likelihood' that the proton propagates to there. The heavier a particle the shorter a distance it can travel. The mass of the particle can then be extracted by fitting to a decay exponential curve. In practice, one may have to fit to a suitable combination of exponentials taking into account the finite extent of and the boundary conditions imposed on the lattice. The effects of other particles, which are also excited from the three quarks and are heavier than the proton, only die away at sufficiently long times. To take care of these, one would need a large lattice in the time direction or suitable fitting functions.

All calculations are performed at unphysically large quark masses to avoid severe finite-size effects and reduce the computing effort at small quark mass. Apart from the pion, it is not known how to extrapolate hadron masses in terms of quark masses. Thus the results are usually displayed as the mass ratios so that they become independent of lattice scales when close enough to the continuum limit. The key test for QCD is the ratio of the nucleon mass to the ρ meson mass, which is shown in Figure 2 (from Ukawa in [3]) plotted against the square of the ratio of pion mass to ρ meson mass. The data show a general trend towards the experimental point but the quark mass and some error bars are

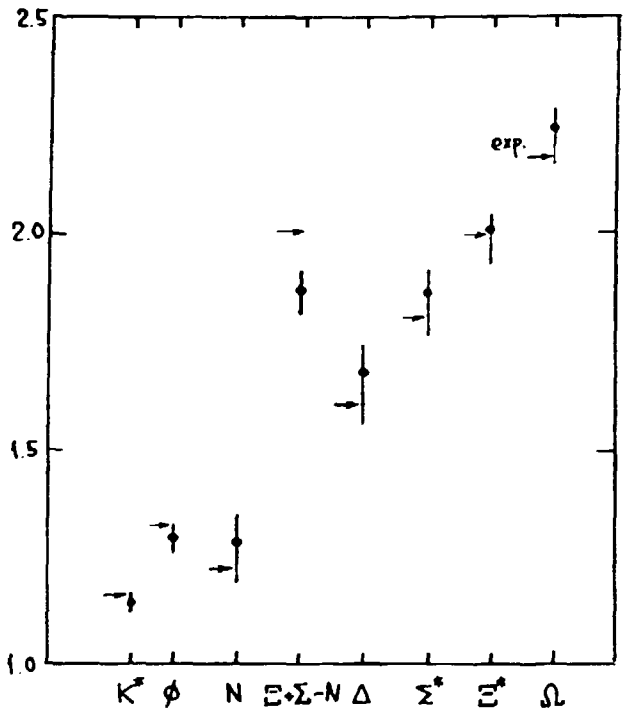

Figure 3: Reasonable agreement between lattice results, with lattice spacing extrapolated to zero, and experimental data pointed to by the arrows. The symbols denote various hadronic quantum states. The mass on the vertical axis is measured in Gev.

still unacceptably large.

Another way to reach the real world is to extrapolate to vanishing lattice spacing using an ensemble of lattices of different values for lattice spacings and total lattice sizes. Figure 3 (adapted from Weingarten as reported in [3]) shows encouraging lattice results as compared to experimental data pointed to by the arrows.

Some decaying processes

From the fitting described in the last section we can also extract other useful numbers beside the mass such as the experimentally measurable *decay constants* in decaying processes. The errors for these are from non-zero lattice spacing contributing sizable amount, $\sim 20-30\%$, and from the matching with continuum physics. Results for some decay constants are shown in Table 1 (for pseudoscalar mesons), from Martinelli in [2]. They are favourably compared with empirical data.

	f_M (MeV) Latt.	f_M (MeV) Exp.
π	$(140 \pm 20 \pm 20)$	132
K	(160 ± 10)	160
D	210 ± 15	< 290

Table 1: Pseudoscalar meson decay constants.

Some commercial and special-purpose machines

From the first large scale numerical QCD at the beginning of 1980's on commercially available machines capable of about, on Cray 1 say, 1.2×10^{15} operations per year, there has been much progress on better machines with faster chips or with parallel architecture allowing a large number of arithmetic chips to be applied to a single problem. QCD calculations are suitable to be implemented on parallel computer as QCD laws are local and homogeneous in space and time. Seen as a computer, the universe has independent arithmetic unit at each point of space time executing the same program and communicating only with its nearest neighbours. So both Single Instruction Multiple Data (SIMD) and Multiple Instruction Multiple Data (MIMD) architecture can be exploited efficiently.

The big potential gain, up to 1000-fold in cost effectiveness over commercially available machines, together with the fact that it requires only moderate memory speed and inter-processor communication speed has prompted several groups to build their own QCD-dedicated machines.

- **Columbia** University machines. The latest has a peak speed of 16 Gflops and can perform both 32-bit (7 Gflops sustained) and 64-bit IEEE standard arithmetic. It has 256 nodes of Intel 80286 microprocessors and arithmetic units distributed on a two-dimensional square with nearest neighbour communication and periodic boundary conditions. Although node execution is not synchronised, communication is. Thus it can be characterised as having MIMD execution combined with SIMD communication.

- **GF11** at IBM Research in Yorktown Heights. The machine has 512 nodes of purely arithmetic units capable of a peak rate of 20 Gflops and of a sustained rate of 7 Gflops. Standard IEEE 32-bit floating point format is used. Each processor has data channels to and from a central switching networking, which is capable of 1024 permutations on up to 576 words sent to and from the processors. The architecture is SIMD.

- **APE** at the University of Rome. The latest is a 16-processor machine with the peak rate of 1 Gflops and sustained rate for QCD is about 200 Mflops. It is SIMD with only arithmetic units of 32-bit operations. The processors are arranged in a closed ring with data channels between nearest neighbours.

machine	architecture	peak Gflops	sustained Gflops	memory Gbytes
CM2	SIMD	32	6.5	2
Meiko	MIMD	5	1.5	0.5
AP1000	MIMD	4	1.5	8
iPSC/860	MIMD	5	1.0	0.5

Table 2: Some commercially available machines.

- **ACPMAPS** at Fermi Lab. The latest has 256 nodes based on Weitek xl chip set. Using 32-bit data, it has a peak of 5 Gflops (to be upgraded to Intel i860 microprocessors to achieve 50 Gflops) and a sustained rate of about 1Gflops. This is a MIMD machine where the processors communicate through a general message passing network.

- **QCDPAX** at the University of Tsukuba. It is rather similar to the Columbia machine but with Motorola 68020 microprocessors. Its peak rate is 14 Gflops, sustained 2.8 Gflops.

- **New Projects.** APE100 is a combination of SIMD and MIMD with expected peak speed of 102.4 Gflops (over 80 Gflops sustained) and cost of US$ 10M. US Teraflops, expected sustained rate of 1 Tflops of 32-bit arithmetic by 1995 with cost over US$40 M; it is largely a 4096 node version of Thinking Machines' new CM5 family of MIMD computers. New Japanese machine, developed in conjunction with Hitachi and to be finished in 1996 with peak rate of over 300 Gflops. European Community also has its own Teraflop project.

Some of the commercial machines which have been used for QCD calculations are summarised in table 2 (adapted from Weingarten in [2]).

A common feature of all the machines above is the use of distributed memory, since shared memory seems to offer no advantages for QCD. It is also apparent that the global instruction and clock distribution system for very large SIMD design becomes very hard to be implemented. Therefore MIMD design is increasingly in favour despite the fact that it is harder to program on such machines.

Concluding remarks

The exploitation of computing power in the quest for the structure of matter is still a young and fast-moving field. Much progress has been achieved and yet there are more works to be done to overcome many limitations. The challenge for better control of systematic errors of present calculations is to be met by better algorithms, bigger memory, and faster processing

speed, or even some new architecture. As more powerful, and thus more complex, machines are in the pipeline there also arises the need for standardisation in programming language, particularly for parallel computers, to enable cross collaboration and full exploitation of resources.

We thank Chris Griffin for preparing the figures. TDK acknowledges support from the Australian Research Council and the Pamela Todd Award.

References

[1] Kenway R D, *Rep Prog Phys* **52** (1989) 1475.

[2] Proceedings of the 1991 International Symposium on Lattice Field Theory, Tsukuba, Japan, *Nucl Phys B (Proc Suppl)* **26** (1992).

[3] Proceedings of the 1992 International Symposium on Lattice Field Theory, Amsterdam, the Netherlands, to be published.