Quarks, Lattices and Transputers

Ken Bowler, Edinburgh

(Physics Department, University of Edinburgh)

Quarks

What is an 'elementary particle'? Ever since the beautiful series of highly inelastic electron scattering experiments performed at Stanford in the late 1960's, there has been a steady accumulation of evidence to suggest that the proton and the neutron are not elementary, as once thought, but are composite objects. Along with the many other strongly-interacting particles produced in high energy accelerators, the proton and neutron behave as bound states of more elementary objects, the quarks. A simple picture, in which the fermions are built from three quarks whereas the bosons consist of a quark and an anti-quark, has been remarkably successful in providing us with an understanding of the observed multiplet structure in the spectrum of particles. However, understanding the mechanism responsible for the binding of quarks into the observed strongly-interacting particles, known collectively as hadrons, is a much tougher proposition. Indeed, it seems that quarks are permanently confined within hadrons and that it may be impossible to liberate them!

The past twenty years has seen a sustained theoretical attack on the problem of quark binding and confinement. The outcome of this effort has been the formulation of a quantum field theory, known as quantum chromodynamics or QCD, which is widely believed to be the correct description of the interactions between quarks. The field quanta which mediate the inter-quark forces, in a manner analogous to photons in quantum electrodynamics or QED, are known as gluons. Unlike photons, however, gluons interact directly with each other even in the absence of other particles, a fact which has dramatic consequences. It has been established that when quarks are very close together, the inter-quark forces become vanishingly small, a property of QCD known as asymptotic freedom. This weakness of the coupling between quarks means that it is possible to do calculations on the short-distance properties of quarks using perturbation theory, a method which has been incredibly successful in QED. Asymptotic freedom has thus provided theoretical understanding of the inelastic electron scattering results by validating a picture in which, at very short distance scales, the proton just looks like a collection of three non-interacting quarks.

Unfortunately, the methods of perturbation theory, expressed through the Feyman diagram calculus, fail to address the crucial questions of confinement and the spectrum of bound states, which concern the long-distance behaviour of inter-quark forces. The corollary of asymptotic freedom is that as two quarks are separated the force between them grows until the coupling is so strong that the perturbative treatment of QCD, the main analytical tool available, is no longer valid. It has been conjectured that what then happens is that as the inter-quark force continues to grow, it becomes energetically more favourable for a quark-antiquark pair to pop out of the vacuum. If we imagine trying to pull a quark from a proton, we see that this process of pair creation allows the quark we are extracting to combine with the antiquark of that pair; instead of a free quark appearing, an additional hadron is created. This picture may be intuitively appealing as a description of confinement, but what is really needed is a non-perturbative treatment of QCD, able to make quantitative statements about the long-distance regime.

Lattices

The major break-through came in 1974 with the suggestion by Ken Wilson that, as a calculational trick, quantum field theories such as QCD should be defined on a discrete lattice of space-time points instead of on a continuum. The quark fields are associated with the points of the lattice, whereas the gluon fields are associated with the links joining neighbouring lattice sites. To recover continuum physics, the lattice spacing must be small compared with any important physical length scale. Fortunately, it is not necessary to go all the way to zero lattice spacing to get back continuum results. It is sufficient to reduce the lattice spacing until physical quantities become independent of the lattice spacing, the so-called scaling regime of the lattice theory.

What makes the Wilson approach so useful is that by formulating the theory in this way, it is possible to use calculational methods developed in statistical mechanics and condensed matter physics for dealing with real lattices, methods which do not depend upon perturbation theory. To be more specific for a moment, by using a formulation known as the path integral method, devised by Richard Feynman, the expectation value of an operator, O, in a simple field theory with just a single scalar field, ϕ, is given by

$$<O> = \frac{\int [d\phi] O \exp[-S(\phi)]}{\int [d\phi] \exp[-S(\phi)]}$$

where S is the action of the field theory, and the integral is a functional integral over the space of configurations of the field ϕ. A configuration simply means specifying the value of the field, ϕ, at every space-time point, and the integral represents the sum over all possible configurations. This expression can also be thought of as the expectation value in the canonical ensemble of classical statistical mechanics in four dimensions, with the factor $\exp(-S)$ replacing the usual Boltzmann factor $\exp(-\beta H)$. Replacing space-time by a finite lattice of points makes the configuration space susceptible to enumeration and, most importantly, opens up the possibility of calculating expectation values numerically on a computer.

A number of difficulties immediately present themselves. First, it is an impossible task to calculate all possible configurations of fields, even on the tiniest lattices. However, the idea of importance sampling comes to the rescue. What is needed is an algorithm that generates field configurations with probability proportional to the Boltzmann factor, $\exp(-S)$, so that only the more important contributions to the integral are sampled. We can then simply compute values of O over a large ensemble of

configurations and average them to obtain an estimate for the expectation value, $<O>$. Monte Carlo methods provide just such a technique and have been used with great success in condensed matter physics for many years. Second, for QCD, matters are complicated by the fact that the fields which appear in the action which defines the theory are no longer simple scalar functions of the space-time co-ordinates. Each gluon field is represented by a complex 3 × 3 matrix, U, whose elements are functions of space-time! Even worse, the quarks are described by anti-commuting variables, ψ and $\bar{\psi}$, for which there is no obvious representation in terms of ordinary numbers in a computer. The expectation value of an operator, O, is given in this more complicated situation by:

$$<O> = \frac{\int [dU][d\bar{\psi}][d\psi] O(U,\bar{\psi},\psi) \exp[-S(U,\bar{\psi},\psi)]}{\int [dU][d\bar{\psi}][d\psi] \exp[-S(U,\bar{\psi},\psi)]}.$$

There is no immediately obvious way to evaluate this integral using Monte Carlo methods. However, what rescues the situation is that is is actually possible to perform the integral over the quark variables ψ and $\bar{\psi}$ analytically, leaving only an integral over the gluon fields, U, to be evaluated by Monte Carlo methods, employing a modified Boltzmann weight. The denominator, for example, becomes

$$\int [dU] \det M(U) \exp[-S(U)]$$

where the matrix, $M(U)$, arises from discretising the derivatives that occur in the action of QCD. This is not the end of all difficulties; the matrix has dimension of order $(L/a)^4$, where L is the edge length of the lattice and a is the lattice spacing, so that it is completely out of the question to evaluate the determinant directly for any but the smallest lattices. Fortunately it seems to be a good approximation for many problems in lattice QCD to neglect the determinant factor by setting it equal to unity, the so-called 'quenched' or 'valence' approximation.

We now turn to the problem of recovering continuum physics from a lattice calculation. Requiring that the lattice spacing be small compared to any physical length scale means that, in the language of condensed matter physics, correlation lengths associated with physical quantities must be large in lattice units. In other words, the lattice field theory must be close to a continuous phase transition, where the correlation length, ξ, diverges. For a computer simulation, we are forced to approximate space-time by a finite box, because there is a limit to the number of variables that we can store in the computer memory. This in turn means that we are restricted in the range of length scales that are accessible. Thus when we to try to ensure that $\xi \gg a$, there is an upper limit provided by the size of the box, L. Indeed, we really need L to be large compared to ξ to be sure that the results are not sensitive to the proximity of the edges of the box. Ideally the calculation should be performed for a range of box sizes to enable finite-size scaling analysis and extrapolation to the infinite-volume limit. Thus we require:

$$L \gg \xi \gg a.$$

State-of-the-art simulations of lattice QCD typically have $L \cong 16a$. Monte Carlo simulations thus have to deal with a number of variables of order 16^4. Furthermore, successive configurations generated by a Monte Carlo algorithm are in general highly correlated. Since $<O>$ is obtained by averaging over a number of samples, the error in the estimate is proportional to $1/\sqrt{N}$, where N is the number of statistically independent samples. which will be far less than the number of configurations generated. This difficulty is compounded by the phenomenon known as critical slowing down, which occurs as ξ/a grows. For local updating algorithms, the number of updatings of each variable required to generate an independent configuration typically grows as ξ^z with $z \cong 2.0$.

What kinds of results have been obtained to date for QCD? In Fig. 1 we show results obtained by the Edinburgh group for the masses of various hadrons in the quenched approximation on a lattice of 16^3 space points and 24 time points. The main qualitative features of the results are encouraging. The relative ordering of the masses is correct, with the nucleon heavier than the ρ meson, which is in turn heavier than the pion. Furthermore, the pion mass appears to go to zero as the square root of the quark mass, in line with theoretical expectations. However, we see from the plot that the nucleon and pion masses, expressed in terms of the ρ mass, are still rather far from the experimental values. Part of the problem is that the physical value of the quark mass is very small — the pion is light on the scale of other hadron masses. To run the numerical computation at such small quark masses is not feasible because the numerical algorithms take an inordinately long time to converge. A second problem is that finite-size effects start to become significant as we decrease the quark mass. So it seems that we need larger lattices, runs at smaller quark masses and higher statistical precision to determine whether or not there are significant

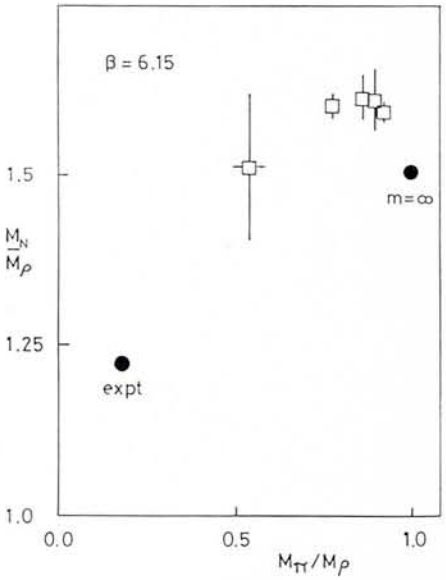


Fig. 1 — The computed masses of the nucleon and pion expressed in units of the meson, for different values of the quark mass. The quark masses used range from 1 GeV/c² to approximately 20 MeV/c². The solid point labelled expt corresponds to the experimental value.

deviations between the predictions of the quenched approximation to QCD and experiment.

By now it should be clear to the reader that Monte Carlo simulation of lattice gauge theories such as QCD is computationally an enormous task. However, the rewards are potentially great — we are freed from the straitjacket of perturbation theory, and the computations have the great virtue of being systematically improvable; as more powerful computers become available, larger systems can be simulated and the statistical errors, which are intrinsic in a Monte Carlo calculation, can be reduced.

Transputers

Where is the necessary computer power to come from? The gain in speed of computers — by roughly a factor of one million in the past 30 years — is due only in part to increases in the intrinsic speed of their components, which accounts for a factor of roughly 1000. The other factor of 1000 is due to the implementation of parallelism. For example, on the large scale, input to, and output from the computer are dealt with separately from the actual computation, and, on a finer scale, the multiplication of each of the digits of one number into the other can be done simultaneously. The logical extension of this idea is to harness the power of many computers to work concurrently on a single problem. Many problems in computational science, in particular the lattice QCD

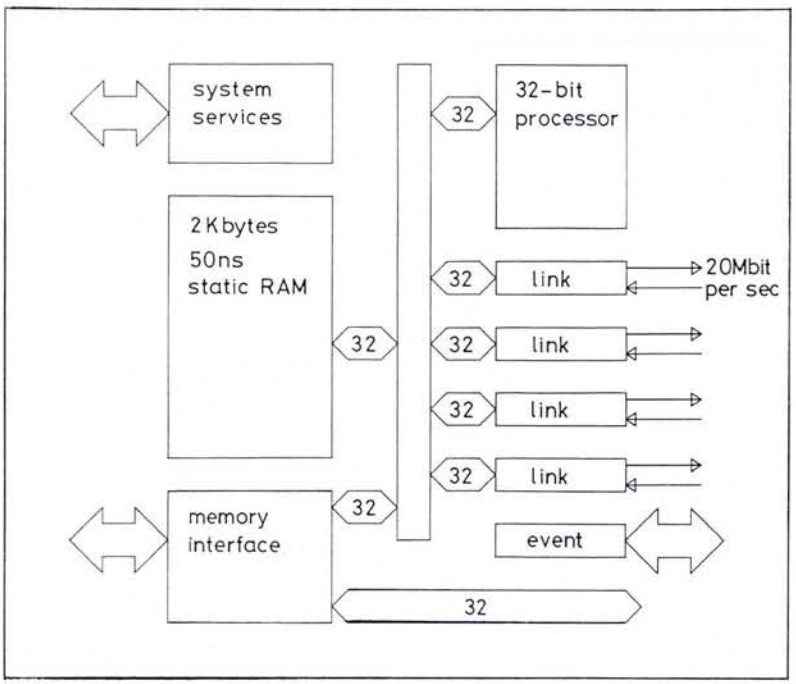


Fig. 2 — Schematic diagram of a transputer.

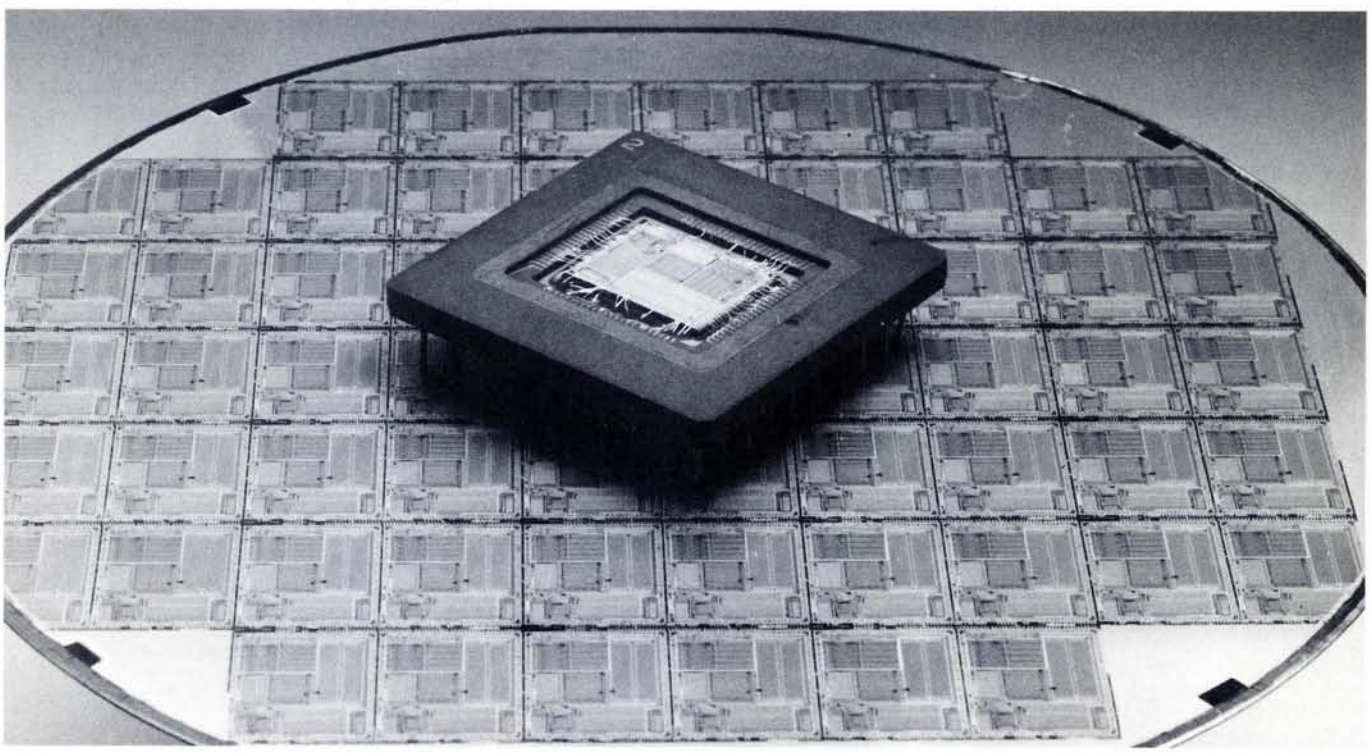
calculations which feature in this article, involve calculations which are inherently parallel; that is, a set of identical computations is performed for a large number of separate data. Thus there is a very natural way in which the intrinsic parallelism of the problem can exploit the multi-processing capabilities of parallel computers. This strategy, in which each processor executes the same program on data corresponding to a sub-region of the physical system which is being simulated and communicates boundary data to neighbouring processors handling neighbouring sub-regions, is usually referred to as geometric parallelism.

The transputer was designed by the British company INMOS as a programmable component for building systems with a far higher degree of parallelism than has hitherto been possible. The term 'transputer' reflects this new device's ability to be used as a system building block. The word is derived from 'transistor' and 'computer', since the transputer is both a computer on a chip and a silicon component like a transistor. What differentiates the transputer from conventional microprocessors is that, as well as processor and memory, the single chip also contains four bi-directional communications links which provide point-to-point connection between transputers. This allows transputer networks of arbitrary size and topology to be constructed. No additional components are required to interconnect transputers on the same board.

Fig. 2 is a schematic diagram of a transputer. From the point of view of large scale simulations, the most interesting member of the transputer family is the T800 which has, as well as the features just described, an on-chip arithmetic unit capable of well over 1 million floating-point operations per second. Fig. 3 shows a T800 transputer in its 84-pin package, sitting atop a silicon wafer.

The principal tool for exploiting the features of the transputer is the occam programming language. Occam enables a program to be described as a collection of concurrent processes, which communicate with each other through channels. Each channel provides a one-way connection between two concurrent processes. Communication is synchronised; if a channel is used for input in one process, and output in another, communication only takes place when both processes are ready. An occam program may be executed by a network of transputers, with the inter-processor links corresponding to occam channels, but equally, the same program may be

Fig. 3 — The photograph shows a wafer of T800 transputers. Sitting on top is the finished product — a single T800 in its 84-pin package.



executed virtually unchanged by a single transputer, in which case the occam channels connect to memory locations. Of course processing is only truly parallel when more than one processor is involved in executing the program, but the transputer has special instructions that share the processor time between concurrent processes and inter-process communication, enabling a single processor to operate in quasi-parallel.

One of the first companies to put together complete systems built from transputers was the Bristol-based firm, Meiko Ltd. Their 'Computing Surface' is a modular system based on a hardware library of different board types, such as compute boards and graphics boards, which can be configured by means of electronic switching chips into a machine whose topology is appropriate to any given problem. At the University of Edinburgh, a large Computing Surface containing, at the time of writing, 200 T800 processors has been established in collaboration with Meiko. The Edinburgh Concurrent Supercomputer, as the machine is known, is already being exploited for a wide range of problems in science and engineering, including lattice QCD. We hope that this radically new machine will enable us to continue the attack on the problem of solving quantum field theories, to give us new insights into the sub-nuclear world.

We conclude by pointing out the existence of a number of other projects elsewhere in the world which are also exploiting the idea of parallel processing to attack the problems of lattice QCD. Special purpose computers have been constructed by groups at Columbia University and at Fermilab in the USA, at the University of Rome, and at Tsukuba in Japan. In each of these projects the computer architecture, although not exploiting transputers, is nevertheless based on a set of interconnected nodes, each containing fast floating-point hardware and large amounts of memory, allowing the geometric decomposition strategy outlined earlier to be adopted.

FURTHER READING

Physics on parallel computers

Bowler K.C. and Kenway R.D., *Contemporary Physics* **28** (1987) 573 and **29** (1988) 33.

Lattice QCD

Creutz M., *Quarks, Gluons and Lattices* (Cambridge University Press) 1983.

Kogut J.B., *Rev. Mod. Phys.* **55** (1983) 775.

Montvay I., *Rev. Mod. Phys.* **59** (1987) 263.

Transputers and occam

Bowler K.C. *et al.*, *An Introduction to Occam 2 Programming* (Chartwell-Bratt, Bromley) 1987.

Green Light for ESRF Red Book

After sheltering under the skirts of ILL for some 12 months the European Synchrotron Radiation Facility has finally emerged into the light. The formal Convention setting it up was signed in Paris on 16 December 1988 by the research ministers of eleven countries (see panel). France was represented by Hubert Curien the Minister for Research and Technology a most appropriate delegate as he was President of the European Science Foundation (ESF) for much of the "political" phase of the project.

The story of the ESRF goes back to 1975 when it became one of the first major studies of the ESF following its foundation in 1974. Through the initiative of H. Maier-Leibnitz a group was formed to consider the feasibility and usefulness of building a special purpose facility to span the X-ray region, based on a high current electron storage ring of 5-7 GeV. Their report (colour dark grey) which was presented to the ESF Assembly in November 1977 identified two light regions as being important — the UV and the X-ray and placed great hopes on the use of wigglers as a means of getting very high intensity in a narrow frequency range.

In addition to encouraging the parasitic use of high energy physics facilities, the report also recommended the establishment of a group to study the possibility of building a new (western) European X-ray synchrotron radiation laboratory starting in the 1980s. The report was endorsed and an *ad-hoc* Committee set up under Y. Farge, then Director of LURE at Orsay. This Committee, with its two Sub-committees — one for the machine chaired by D.J. Thompson from Daresbury, UK and the other for instrumentation chaired by B. Buras, from Copenhagen, Denmark, in turn produced its report, which came to be known as the Blue Book. In four volumes it established the general feasibility of the project, set out the scientific case, the basic machine design and the instrumentation required.

National projects seemed more important however, and while the Committee continued to improve the design and in particular develop the possibilities of undulators and wigglers, the project as a whole languished. Clearly a major political effort was needed and this was confided to an ESRF Progress Committee under Dr. Paul Levaux of Belgium. The Committee continued to work towards a solution to the funding question and a choice of site amongst the three put forward, namely Risø in Denmark, Trieste in Italy and Strasbourg in France. Later, Dortmund and (unofficially) ILL in Grenoble were added. A report was published (the Yellow Book) which updated the Blue Book and to avoid dispersing the technical group it was agreed to fund an ESR Project Group. This was formed in mid 1983 under B. Buras and

Contracting Parties

and contribution to construction costs exclusive of TVA

France	34%
Fed. Republic of Germany	24%
Italy	14.5%
United Kingdom	12.5%
Belgium	3%
Spain	4%
Nordic countries (DK, SF, N, S)	4%
Switzerland	4%

Machine Characteristics

Storage ring for electrons or positrons:
845 m round, 32 straight sections with ≥ 6 m between quadrupoles.

Experimental hall over ring accommodating beam lines up to 75 m long.

Beams at 6 GeV of *ca* 100 mA in multibunch mode, 5 mA single bunch;
No. of bunches 1-992; horizontal emittance 7×10^{-9} mrad, vertical 6×10^{-10} mrad.

Storage capability: > 8 hours for smooth fall to 1/e initial value.

Brilliance from undulator: > 10^{17} ph/s mrad^{2}mm^2 per 0.1% bandwidth and metre of undulator at 14 keV.

Flux: 8×10^{12} ph/s mrad per 0.1% bandwidth.