Defect Migration in Solids: Microscopic Calculation of Diffusion Rates

A. M. Stoneham

Theoretical Physics Division B424.4, Harwell Laboratory, Didcot, Oxfordshire OX11 0RA, UK

Abstract

Diffusion includes some of the commonest processes on an atomic scale, in which uncorrelated atomic jumps bring about changes in solids. The many predictions of activation energies characterising the temperature dependence of diffusion have high accuracy when good interatomic potentials are known, and a continually improving accuracy from direct quantal calculations. Much more is needed to find the absolute rate at some temperatures. Recent work for both classical and quantum diffusion shows that rates too can be estimated with respectable accuracy. Such calculations highlight basic problems of solid-state defect physics. They also offer predictions of technological value for diffusion rates in cases where the timescale or physical conditions are too hard for direct experiment. This talk will discuss (a) various ways of obtaining absolute rates, (b) rates of cation diffusion in oxides and their implications for the so-called Compensation rule (c) issues of charge state in oxides and semiconductors, (d) transients in diffusion, as for recombination-enhancement in semiconductors or the motion of short-lived species, and (e) quantum diffusion of hydrogen in metals.

1. Introduction: The phenomena of solid-state processes

My topic is one of the most basic of processes on an atomic scale by which change can occur in solids, and my main concern is the theoretical prediction of the rates of such processes.

Diffusion provides a link to the commercial exploitation of science, a feature which was remarked on in the very first studies in 1684: Boyle [1], noting the way sulpher could "soak into the pores of copper" to "give it a durable colour" such that a copper coin changed its colour to gold with "the former stamp, that was impressed on the coin, continuing visible", decided it was best to leave out the precise details of his experiments "because of the bad use that may be made of it". This also provides an early example of the threat to good science of the wrong sort of application. The practical role of diffusion is now widespread (see Ref. [2]) for an interesting range of early studies), occurring in the oxidation of metals or of silicon, the doping of semiconductors (either directly by diffusion or by the anneal of implants), the undesirable electromigration in aluminium interconnects, or oxygen exchange in gas sensors or in ceramic superconductors. The importance of predictive theory lies partly in the way that irreversible phenomena and solid state processes are understood, and partly in complementing experiment to cover those regimes (e.g., extremes of temperature or pressure, or timescales which may vary from the geologically long to the exceptionally short) which are necessary technologically but inconvenient experimentally.

Boyle's observations were possible because the diffusion rate was just right to show striking effects on a timescale practical for experiment. It is worth emphasising just how widely diffusion rates can vary: the same class of process can encompass the so-called "superionic" conduction in solid electrolytes and geological dating. Observed diffusion rates vary over nearly twenty orders of magnitude (and indeed over 11 orders [3] for at least one single system, Si : 0) so that it is a formidable task to predict the rate correctly for a real system to within experimental error.

This difficulty can be seen from the first of the major phenomenological rules of diffusion, the Arrhenius law:

$$X = X_0 \exp\left[-E/kT\right] \qquad (1)$$

where X is the diffusion coefficient D, or the mobility μ, or the conductivity, or the jump rate Γ of some atom; E is an activation energy and the prefactor X_0 may have some slow temperature dependence (indeed must do so for certain of these quantities if the Nernst–Einstein relation between mobility and diffusion coefficient holds [4]; see also [5]). The form of eq. (1) is known to theorists as a well-known observation and to experimenters as a fully-established prediction. It does have exceptions, of course: behaviour will be more complex if there are several parallel processes (or a hierarchy of processes); certain related processes may show rates which are power laws with temperature. Yet the Arrhenius expression (1) is central to any discussion of diffusion rates, and the link between theory and experiment is usually achieved through the two characteristic parameters E and X_0.

There is a second general rule, less well-known and less clear-cut. It concerns the relationship between the activation energy E and the prefactor X_0 in a range of related processes (e.g. perhaps for equivalent processes in a range of oxides of the same structure, though the rule is surprisingly widespread). The "Compensation" or Meyer–Neldel [6] Rule is:

$$X_0 = X_{00} \exp\left(E/kT_0\right) \qquad (2)$$

$$\ln(X_0) = A + BE \qquad (3)$$

i.e., large activation energies (which would give low rates) tend to be compensated by large prefactors (which raise the rates) [17]. Again, qualification is needed; as for (1), a single mechanism should operate in each case; moreover, what "related" processes are allowed is not too well defined. Nor should the rule be forced by data selection, even if unintentional, as when only those data are included which can be collected in a convenient time (like a working day; see [8] for a more subtle variant). That the rule is not universal is conveniently illustrated by work on a-Si : H prepared at various deposition temperatures [9]. The dark conductivity showed a compensation rule if the samples were fully annealed after illumination but — even in the same samples — showed no sign of compensation if they had been fully optically-irradiated before measurement.

The Arrhenius rule (1), supplemented to some degree by

equations (2, 3) is the basis for the systematic description of diffusion, and gives the source of data for modelling on an atomic scale. Theoretically, there are several key ideas, which include the relationship between diffusion and molecular reaction rates and some classes of non-radiative transition. I shall concentrate on the way that the rate of diffusion can be calculated; indeed, more specifically, I shall concentrate on the prediction of the basic jump rates, so as to avoid the extra complications like combining jump probabilities to give more-directly observed properties, or the numbers of diffusing particles, or geometric factors [4, 10]. Some of the broader issues will emerge, like the status of the Arrhenius and Compensation rules, the situations for which transient behaviour is important, and the circumstances in which classical descriptions are not enough and it is necessary to return to the Schrödinger equation.

2. Approaches to absolute rates

2.1. *General issues*

Diffusion theory raises problems of several distinct sorts, including scientific questions such as:

(I) Issues of basic statistical physics, like the origin of macroscopic irreversibility from reversible microscopic equations [11, 12]. The idea of ergodicity (so that the system, if left to itself for a sufficient length of time, would pass through every dynamical state consistent with its energy; time averages are then equivalent to ensemble averages) is another basic issue, which is not true for some important systems, e.g., an assembly of harmonic oscillators.

(II) Problems of idealised systems, like the ideas of "friction" within Kramers' theory [13, 14], or the use of a single reaction coordinate for a particle coupled to a heat bath (cf. [15]), the other coordinates treated as friction. Here there are implied assumptions about energy transfer between modes and about dephasing (see later), and there are further issues if there are bath frequencies slower than the hopping frequency or its equivalent;

(III) Aspects of special systems, like the existence or not of diffusion in some one-dimensional systems [16, 17];

(IV) Questions of quantum physics, notably the status of the common view that quantal effects occur only through $h\Omega/kT$ factors (i.e., in statistics and, by implication, not in dynamics) so that the classical limit $h \to 0$ and the high temperature limit $T \to \infty$ are the same [18].

(V) Questions of whether the electrons need be considered in detail, beyond their involvement in determining the total energy (which, incidentally, need not be single-valued) as a function of nuclear coordinates [19].

Points IV and V will emerge again later. I shall assume any problems under I–III have been solved, and I shall concentrate on the best available atomistic models for which serious comparison with experiment is possible. Morover, I shall note some issues of applied science, involving situations where the system of interest is not chosen for the convenience of the theorist.

2.2. *Simple reference systems*

Two simple examples of diffusion suffice for many of the points to be made. One is the diffusion of an impurity interstitial, with negligible host lattice diffusion (though, of course, the interstitial and host interact stongly). The second

is vacancy diffusion, in which an atom or ion makes a jump into a vacant site from one of the neighbour sites. For either system, we shall need to define an energy surface, in practice the adiabatic surface (i.e., for each position of the diffusing particle, all other atoms are allowed to relax to their static equilibrium positions) for each of many positions of the diffusing particle as it moves from one site to another. We shall also wish to identify (and quantify) the saddle point on this path, and define a "reaction coordinate" Q which follows a favourable route from one site to the next (choosing this favourable route is not easy, cf. [20–22], nor is it necessarily simply related to the most important atomic motions of diffusion). Further, we shall need to know about the thermal motions of the atoms, at least the harmonic vibrations about selected points, and sometimes even more information about vibrational eigenstates.

2.3. *Molecular dynamics*

This is the most direct way to study atomic motion, and of great importance. The classical equations of motion are integrated numerically for some large cell of atoms (perhaps several hundred per cell) with periodic boundary conditions. Diffusion jumps are rare events in thermal equilibrium for most systems, so ingeneous ways are used to deduce jump rates. Molecular dynamics has real advantages in investigating particular aspects, like "return jumps", in which a particle passes through the saddle point (or saddle surface) only to return immediately [23, 24], or transient processes (see Section 5 here), or situations in which the barriers are small, so the separation of atomic motion into vibration and diffusion is hard. It is also especially helpful if the jump mechanism is not obvious. However, there are some aspects which are best done by the other methods I shall be discussing, notably (a) when one wants an activation energy — i.e. a temperature dependence — accurately over a wide temperature range, since the statistics become poor for lower temperatures and (b) if one wants to use the energy calculation beyond the simpler ones, e.g., the shell model to describe ionic systems, since there is a significant cost penalty. Fortunately, molecular dynamics and other approaches are complementary and also in agreement in those cases which have been tested [24, 25].

2.4. *Rice–Slater dynamical theory [26, 27]*

Suppose the reaction coordinate Q can be regarded as dynamically independent*. How often does one find the system with $Q > Q_s$ and $dQ/dT > 0$, with Q_s the saddle point and the sign of dQ/dt corresponding to motion toward a new final site?

With certain working assumptions, e.g., that the energy surface is harmonic on each side of the saddle point, this can be related to the "first passage" problem; hence, with some generalisation [4, 28–30] to an expression for the jump rate of the form:

$$\Gamma = \Gamma_0 \exp\left(-E/\langle \varepsilon \rangle\right) \tag{4}$$

---

* In a harmonic solid, the normal modes are dynamically independent, i.e. the (non-diffusive) vibrational motion can be described in terms of uncoupled harmonic oscillators. However, any chosen motion (e.g., of one atom or in the reaction coordinate Q) is usually not dynamically independent, even though the coordinate may be convenient to describe some process. The distinction between normal modes and reaction coordinates is of great importance, often hidden in one-mode models.

in which $\langle \varepsilon \rangle$ is simply proportional to the energy associated with the reaction coordinate, scaled here to be kT for thermal equilibrium, and Γ_0 is an average of vibrational frequencies for the (defect) lattice. There are two corollaries. First, any "first passage" theory will yield a jump rate bounded by the highest lattice frequency, and this is usually confirmed by experiment (H in liquid Fe may be an exception, but one assumes the jump frequency bound is still good though jumps are longer than usual [31]). In simple cases the frequency prefactor Γ_0 might be expected to have the same bound, though experiment often disagrees (so for interstitial O in Si [3], the prefactor is about 500 times greater than the highest vibrational frequency for Arrhenius behaviour over eleven orders or magnitude). The second corollary is that this picture presumes negligible transfer of energy between degrees of freedom, a reasonable assumption for some of the molecular reactions for which the description was devised. If a fixed amount of energy $\delta\varepsilon$ can be added in excess of the kT value in thermal equilibrium (see Section 5) then $\langle \varepsilon \rangle$ becomes $kT + \delta\varepsilon$, leading to non-Arrhenius behaviour.

2.5. Reaction rate theory and Vineyard theory

The original ideas of Marcelin, generalised in several ways by Hirschenfelder, Wigner, and others [32–34], concentrate on a system in thermal equilibrium. For the system of classical particles, the probability of some particular combination of momenta and positions is defined in thermal equilibrium (with assumed boundary conditions, usually constant volume) for each such configuration. How the system reached that state is not asked: when one talks of the relative populations of "ground" and "saddle-point" states, there is no assumption about the path used, merely more general assumptions like ergodicity [12] (so that, for instance, there is no selection rule preventing the saddle point being reached from a chosen initial state) and that the rate of transfer of energy between degrees of freedom is never a limiting step (the opposite assumption to dynamical theory), an assumption usually adequate for solids (not always, cf the phonon bottleneck [35]) but not always so for molecules. In this approach, one concentrates on a plane (which need not be the saddle plane) and looks at the flux of those particles moving across the plane in a particular direction (there will be fluxes in both directions of course, balancing on average in thermal equilibrium).

These ideas yield a rate which, in general terms, includes three factors: a population factor, i.e., the probability that the system is close to the saddlepoint (more strictly to the saddle hypersurface in the phase space of the system); a flux factor describing the motion near the saddle point; and an efficiency factor, allowing for transient motions which cross the barrier without being countable as jumps [25]; in our cases we should expect the efficiency factor to differ from unity by only a few percent if the guidelines from ref [25] apply. The practical version of reaction rate theory is that of Vineyard [36], who assumed locally-harmonic energy surfaces. In the initial state (and the equivalent final state) the normal modes have frequencies v_i; at the saddlepoint the real frequencies are v_j', there being at least one imaginary frequency. The key frequency which enters the jump rates is:

$$\Gamma' = \prod_{i=1}^{N} v_i \bigg/ \prod_{j=1}^{N-1} v_j' \tag{5}$$

Note that, at this level, these are products over modes.

If one mode needs special treatment (perhaps because of some anharmonicity which does not mix modes) this can be handled separately. Clearly an practical application of eq. (5) needs a technique for the accurate prediction of at least a representative sample of frequencies v, v' [37, 38]. Note too that if even one v' is small, as when the energy surface normal to the jump path at the saddle point is especially flat, then Γ' may have a very high value. This is not the explanation of the high prefactors mentioned earlier but (as for vacancy diffusion in MgO) a sign that one should go beyond Vineyard theory [21].

2.6. Polaronic and similar mechanisms

So far it has been implied that the saddle point in the adiabatic energy surface is the key to the jump process. Yet this is not always so, and especially when quantum, tunnelling can occur. Here, if the quantum particle is light (and so faster-moving than other atoms) the coincidence state has analogous significance [39, 40]. In the simplest cases the activation energy is the strain energy needed to move the host atoms from their positions relaxed about the initial site (from the self-trapping distortion in some cases) to a geometry in which the light particle would have equal energy at either site. This description is important in the quantum diffusion of hydrogen (Section 6), as well as in the motion of small polarons (Section 4).

3. Cation mobility in MgO

3.1. Experiment

MgO is a typical ceramic, similar in many respects to the even commoner alumina, a high melting-point oxide with applications ranging from relatively coarse technology to being a high-quality thin-film dielectric for microelectronics and the substrate on which have been prepared to specimens of the new superconductors with the highest critical currents yet achieved. It is also simple oxide, having the NaCl structure, and it is ionic in a well-defined sense [41] in which one may concentrate on closed-shell ions Mg^{2+} and O^{2-}. Whilst single crystals are easily grown, most diffusion data are complicated by the role of impurities, notably C, Fe and possibly H. Yet, as Sempolinski and Kingery showed [42], the mobility $\mu(T)$ of the cation vacancy can be measured near 1400° C. This, in turn, can be expressed in terms of the probability $\Gamma(T)$ per unit time that one of the ($N = 12$) neighbouring Mg^{2+} ions jumps into the vacancy:

$$\mu(T) = [Na^2/3][e/kT]\Gamma(T) \tag{6}$$

with a the nearest neighbour distance. Direct analysis of the experimental data yields an activation energy 2.3 $\pm$ 0.2 eV at 1400° C and a value of the prefactor Γ_0 of 210 $\pm$ 80 THz, a value which needs comment when compared with the maximum lattice frequency of around 20 THz.

3.2. Some basic relationships

The explanation of the value of Γ needs two types of general relation, one algebraic and the others thermodynamic. The algebraic component is simple. Suppose that in an activated process with rate $A \exp(-\varepsilon/kT)$ the activation energy decreases with temperature, perhaps because thermal expansion causes a lowering of energy barriers, or (for electronic transport in amorphous crystals) because the Fermi level

shifts with temperature:

$$A \exp(-\varepsilon/kT) = A \exp(-[\varepsilon_0 - \alpha kT]/kT)$$

$$= [A \exp(\alpha)] \exp(-\varepsilon_0/kT). \qquad (7)$$

Experiment will measure $A \exp(\alpha)$, not A: there is no way to separate the two from observation along. It may make sense to choose A from the maximum lattice frequency; if so, the prefactor observed will still be $A \exp(\alpha)$ and may be much larger: high prefactors suggest (as will be discussed) activation energies decreasing with temperature. Many authors describe αk as an entropy. This can be useful, though it is only $A \exp(\alpha)$ which is defined, not A and α separately, so (contrary to recent assertions) there is no significance whatsoever in the temperature $T = \varepsilon_0/\alpha k$. However, there may be correlations between ε_0 and α, and these offer a way of describing the Compensation rule. The temperature dependence of the activation energy can, of course, be more complicated than linear, and indeed some complex behaviour near phase transitions has been described.

The thermodynamic relationships [43, 44] can make these qualitative arguments more precise. Experimentally, the concern is with constant pressure Gibbs free energies g_p. Theoretically, it is usually easier to estimate constant volume (or strictly constant lattice parameter) quantities; moreover, Vineyard theory works with a constant volume ensemble. The key expressions for the enthalpy h_p and entropy s_p are these, where in each case the second form assumes the thermal expansion is sensibly linear in temperature for the regime of interest:

$$h_p = u_V - TV(\mathrm{d}V/\mathrm{d}T)_p(\mathrm{d}f/\mathrm{d}V)_T \qquad (8a)$$

$$\approx u_V(0) + T^2(\mathrm{d}s_V/\mathrm{d}V)_T(\mathrm{d}V/\mathrm{d}T)_p \qquad (8b)$$

$$s_p = s_V - (\mathrm{d}V/\mathrm{d}T)_p(\mathrm{d}f/\mathrm{d}V)_T \qquad (9a)$$

$$\approx s_V(0) - (\mathrm{d}V/\mathrm{d}T)_p\{(\mathrm{d}u_V/\mathrm{d}V)_T - 2T(\mathrm{d}s_V/\mathrm{d}V)_T\} \qquad (9b)$$

The fact that h_p has as its leading term $u_V(0)$ is the reason for the relative success of those many calculations which merely estimate the zero-temperature internal energy. For a vibrational entropy, however, the constant pressure and constant volume forms may differ greatly, even in sign [44]. The correction from the volume dependence of u_V through thermal expansion can be large, analogous to the term in α mentioned above.

3.3. Explicit calculations for MgO

In all the calculations to be described, the energy surfaces are obtained using the shell model [45] with interatomic potentials obtained from perfect crystal data only [46]. For ionic crystals this approach is both accurate and convenient; further, because of major, widely used codes (see [48]; the main examples are the Harwell HADES [47] and SHEOL [38, 44] codes used for the calculations described here, or the Harwell/SERC CASCADE code) the approach is both efficient and economical.

The strategy of the calculation is shown in the Fig. 1. We can usefully discuss the results in parallel with an earlier treatment by Sangster and Stoneham [37, 22] who concentrated on u_V, s_V and $(\mathrm{d}u_V/\mathrm{d}V)_T$, i.e., the leading terms in the earlier expressions, and evaluated (with some generalisation) the Vineyard expression for the prefactor in the jump rate. They predicted a migration energy of 2.26 eV, agreeing well
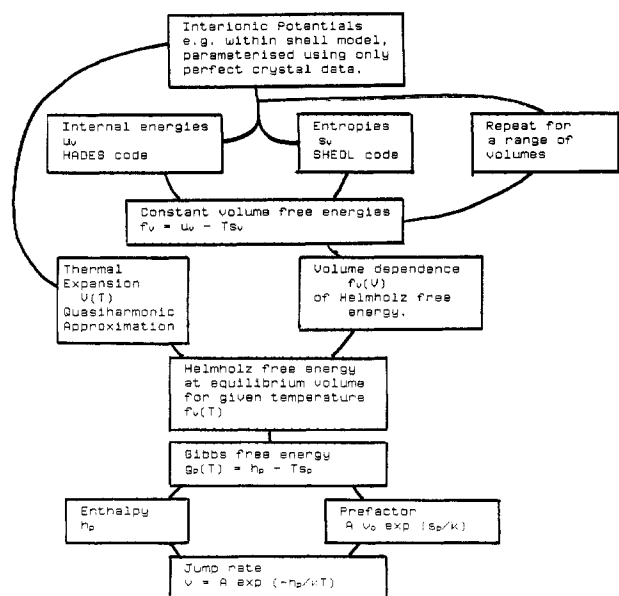
Fig. 1. Flowchart for calculations of the diffusion jump rate: classical case.

with experiment, the associated activation volume (-37% of the atomic volume) and the implied value of α (2.46 using the experimental thermal expansion, 1.53 with their theoretical value) being reasonable. The Vineyard prefactor was, however, surprisingly large: 32.9 THz, more than the 20 THz maximum lattice frequency. This is precisely the problem mentioned earlier: the energy surface at the saddle point (shown in [22]) is rather flat near the saddle point. However, as noted earlier, the way the Vineyard expression splits into a product over modes can be exploited. Sangster and Stoneham did this in two ways: by direct use of the energy surface (an approach improved in [45]) and by a simple analytical expression. Since the flat region is local, the potential getting much harder as one moves from the jump path near the saddle point, it suffices to replace these harder regions by infinite walls. This yields a correction $\mathrm{erf}\,(E^s/kT)^{1/2}$, where E^s is the strain energy in the soft mode to just reach the wall. The correction given approximately is 0.76 ± 0.1; the fuller treatment gives 0.72.

The following table shows just how well these calculations do work. Despite their sensitivity (for the rate depends exponentially on calculated parameters) the jump rates at quite high temperatures are predicted to within the limits of experimental error.

3.4. Implications for other systems

We may now use these calculations for cation motion in MgO to understand several other systems. First, we remarked that an activation energy which fell rapidly with temperature (e.g., from thermal expansion of the host, giving large α)

Table I. *Jump rate, jump rate prefactor and activation energy for MgO*

	Theory [37]	Theory [49]	Experiment [42]
Γ_0, THz (without exp (α))	23	12	19 ± 7
ε_0, eV	2.26	2.14	2.3 ± 0.2
Γ (1400°C) jumps/sec	4.34×10^7	1.48×10^7	2.84×10^7

Table II. *Compensation: cation motion in the alkaline earth oxides [49]*

	MgO	CaO	SrO	BaO
Prefactor $\Gamma_0 \exp(\alpha)$ THz	31.29	20.55	11.60	6.51
ε_0 eV	2.14	1.87	1.68	1.48

suggested a high prefactor ($A \exp(\alpha)$). Even without explicit calculation, one might guess that expansion should have a strong effect on the oxygen interstitial in silicon, where the prefactor is so much larger than the highest frequency: for $Si:O_I$ the interstitial is associated with (but displaced from) the bond-centre position, and the geometry (and hence bond angles and jump distance) and energy are likely to be sensitive to sensitive to expansion. Perhaps more important, we can see how it is that prefactors might bear little relation to simple "ballistic" or "approach frequency" pictures, and can be much larger [50] or much smaller [51] than those imply.

Secondly, we may remark on the Meyer–Neldel rule. The simplest explanation merely supposes that $u_V(0)$ and $(du_V/dV)_T$ correlate, i.e., that large barriers imply large deformations along the jump path, and that these are relieved effectively by expansion. This can be made quantitative [49] for a series of alkaline earth oxides (see Table II).

Thirdly, we have noted the flatness in the saddle plane normal to the jump path. This can be regarded as a move towards a bifurcated path as routes somewhat away from the direct one becomes favoured. Indeed, for the Fe^{3+} in MgO, bifurcation does occur ([22, 52]; ref 22 shows the energy surface), and such behaviour seems typical for $3+$ ions in MgO. Bifurcation makes it hard to apply the dynamical theory, since the reaction coordinate is not unique; moreover, we note that dynamical theory gives rather poor values for the Mg^{26} jump rate in MgO too. Molecular dynamics and Vineyard theory, however are found to agree well [24].

Fourthly, one severe test of diffusion dynamics is the isotope effect. Without going into details the rather complete knowledge of frequencies and their dependence on ionic masses suffice [53] to give good estimates of the so-called kinetic energy factor, δK. Implicit here is the assumption that the potential energy surface does not depend on isotope, an assumption which will need to be questioned in discussing quantum effects.

Finally, we may remark on the differences in going to the transition metal oxides like CoO or NiO, with the same crystal structure. These differences take several forms [52]:

(a) The cation can exist in several charge states, notably the $3+$ and perhaps the $+$ states, i.e., as small polarons which themselves hop from site to site in a diffusive motion;

(b) There is usually a cation deficiency, so that more complex defects involving cation vacancy clusters can occur [54, 55]

(c) For a $3+$ ion near a cation vacancy, the energetically favoured site is the (200) second neighbour, not the (110) first neighbour, so that diffusion of $3+$ ions involves correlated polaron and ion motion (see Section 4);

(d) There are additional terms in the expression for the energy, beyond those in the shell model. The crystal field splitting and Jahn Teller terms are recognised, but usually

small in effect [56]; quadrupole terms are another matter [57], and seem to be significant in understanding differences from one host to another (the effects are from the host ion quadrupoles, rather than those of the diffusing particle).

4. Charge states and related matters

Comments in the last section drew attention to the differences in behaviour between $3+$ species, like Fe^{3+}, and $2+$ species like Fe^{2+}, even when diffusing by the same vacancy mechanism. We may relate these differences to other cases in oxides and semiconductors. We begin with the transition metal oxides MnO, FeO, CoO, NiO, which readily go non-stoichiometric (FeO especially to a degree that defect clusters must be considered). Diffusion jumps are possible by a variety of combinations of purely ionic and purely electronic jumps [see [50] for examples]. Moreover, the energy surface is no longer single-valued: for given nuclear positions, there are various ways of disposing the electronic carriers over the ions. In particular, an ion stable as $2+$ next to a vacancy might be more stable as $3+$ at the saddle point. Such multi-valued energy surfaces raise questions of ergodicity unless transitions between them are effective. Consideration shows two main types of transition between surfaces: (i) Equilibria between the alternative electronic energy surfaces established at the stable site; (ii) "Ballistic" transitions with charge state changes (transitions occuring between energy surfaces) as the diffusing ion moves towards the saddle point (this is an odd but common type of description; in reaction rate theory the question does not arise, since we assert an equilibrium without needing to worry about rates of certain special processes, and especially electronic processes). For transition metal oxides the hopping integrals are large enough for both processes to appear possible; for the Al interstitial in silicon [57] and for Fermi level effects on dislocation motion in silicon (see [58]), (i) alone seems a reasonable description from known carrier capture rates and populations.

It is known from doping studies that charged defects dominate silicon self-diffusion at high temperatues [59]; this does not depend on the question of which species is mobile:

	1000 K (lower temperatures)	1670 K (higher temperatures)
$D(V^-)/D(V^0)$	0.56	11.5
$D(V^+)/D(V^0)$	0.07	18.4

It is also well known that silicon self-diffusion appears to have a very high entropy, $\approx 10\,k$, and this has led to much speculation. The $10\,k$ cannot come easily from vibration alone (e.g., soft saddle-point modes) partly because, as discussed for MgO, the apparent enhancement is misleading; also, whilst it is easy to imagine a very soft mode for some particular lattice parameter, silicon self-diffusion is observed at constant pressure, so the unit cell size changes with temperature. Nor is it adequate to say the several types of self-interstitial offer enough (N say) alternative paths for a configurational entropy term to give $10\,k$: even if the degeneracy were exact (it is not) the degeneracy would contribute the inadequate $\ln(N)$ factor. Complex defects ("extended interstitials") have been proposed too [60], but specific models which work [61] are yet to be identified. The most obvious

explanation makes a direct link between the entropy and the charge state change. The rise in importance of the charged species is presumably partly a result of the changes in electron and hole concentrations; these in turn depend on the band gap. The band gap itself is strongly temperature-dependent (0.6 eV fall from 0 K to the melting point, i.e., dE/dT of about 6 k), and it is natural to assume the major part of the 10 k in self-diffusion comes simply from the band gap change. The conclusion is that any theory which fails to predict this band gap change will also fail to predict the entropy for self-diffusion.

5. Transients in diffusion

Both the classical and quantal cases discussed so far refer to jump rates in thermal equilibrium. The equilibrium is not static, of course, as every particle continually changes its position and velocity. However, two notable cases of non-equilibrium behaviour occur, for which transient behaviour matters:

(1) When extra energy is suddenly made available locally, as in recombination-enhanced diffusion, in which an electron and hole recombine at a site and, in one possible process the electronic energy released gives a "phonon kick" to a particle. This is a gentler process than radiation damage, with energies of 1–2 eV typically involved.

(2) When the particle number is not conserved, as for muon or exciton diffusion, where there are at least three timescales, namely (a) that for self-trapping, i.e., establishing the substantial local distortion, (b) diffusive hopping, and (c) the decay of the particle.

I shall discuss (1) only here; for (2) see [62].

Recombination enhanced diffusion is of real practical importance for semiconductor systems, since it speeds up degradation of devices like semiconductor lasers when they are in use, rather than merely on the shelf (for references to 1981 see [63]).

Discussions of the phenomenon (e.g., [64]) have normally been along these lines. Suppose normal thermal hopping occurs at a rate $A \exp(-E/kT)$. If the recombination energy available per event is R, then the effective activation energy is reduced to $E' = E - \phi R$ because the thermal fluctuations have to overcome an effective lower barrier. Here ϕ is an "efficiency factor" depending on how well the phonon kick (determined by the gradients of the electronic energy surfaces) and the reaction path (determined by the ground state energy surfaces) correspond, etc. The prefactor A will be altered too by some factor defining the fraction of time for which the excitation is available as enhanced local motion, and will depend on the intensity I of excitation. Overall, there will be a new term $AI\beta \exp(-[E - \phi R]/kT)$, still thermally activated. With some qualifications (e.g., ϕ has to be quite large, even > 1 at times, and β may have to be big too) this description works quite well.

The problem lies in the efficiency factors ϕ and β. High efficiencies mean little loss of coherent motion in the reaction coordinate. Now reaction coordinates are rarely dynamically-independent normal modes, as assumed in simple models. If they were, the lack of energy loss would lead us to Rice–Slater dynamical theory. This would indeed predict a change in rate, but not in the Arrhenius form: the vibrational energy in the key mode changes to give a rate $A\beta'I \exp(-E/[kT + \theta R])$.
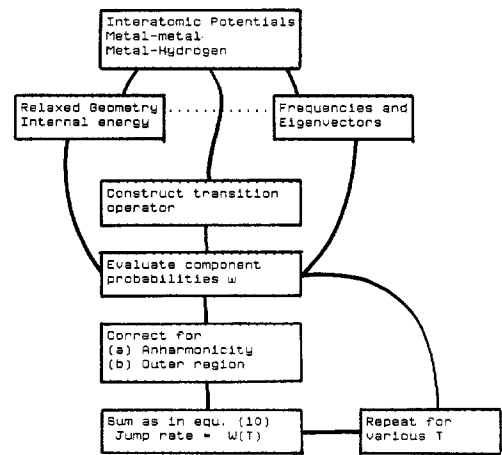

Fig. 2. Flowchart for the calculation of the diffusion jump rate: quantum case. (Note this is given for low temperatures, for which thermal expansion can be ignored.)

Such behaviour is indeed seen for molecules. To resolve this problem for solids, molecular dynamics seems essential. The only study appears to be [65], examining the consequences of a phonon kick on interstitial motion. The statistics for their results did not, however, allow the two expressions mentioned to be distinguished, but sufficed to identify several other points. One was that the direction of the phonon kick and of the reaction path had to be matched well for reasonable efficiency. Another point was the striking effect of phonon dispersion (for most simple models assume a single frequency): coherent motion in the reaction coordinate is damped with a characteristic time which is short. If the mean frequency is $\langle \omega \rangle$ and if we use $f = [\langle \omega^2 \rangle - \langle \omega \rangle^2]^{1/2} \langle \omega \rangle$ as a measure of the dispersion, then the efficiency α did not exceed $\exp(-2f)$ and the motion will be damped after a few (typically $\pi/2f$) vibrations in the reaction coordinate.

6. Quantum diffusion of light interstitials:hydrogen in metals

Quantum diffusion and clasical diffusion differ in concepts as well as in the statistics used [18, 66]. Some changes are obvious: zero-point motion will modify the local distortion near a light interstitial in a way which depends on isotope mass; light interstitials will have discrete energy levels, not a continuum up to some saddle point, etc. But the use of the Schrödinger equation instead of Newton's laws of motion means that the dynamics of the diffusing particle, not just statistics, are changed: the correspondence principle ($h \rightarrow 0$) is not equivalent to high temperatures ($T \rightarrow \infty$) as it would be if statistics alone (involving just $h\Omega/kT$) were significant. Moreover, the saddle-point configuration loses its special significance. Just as in polaron theory the analogue (like the coincidence configuration, cf Section 2) is distinct. This was recognised initially by Flynn and Stoneham [40], whose approach is discussed in relation to more recent methods shortly.

The idea that the classical saddle point is no longer sufficient has also been exploited in an interesting molecular dynamics simulation by Gillan [67]. Current quantum simulations do not permit direct calculation of the jump rate but, in the spirit of classical reaction rate theory, another similar type of estimate is possible. One seeks the probability that the interstitial has equal weight on initial and final sites, in the spirit of the coincidence site picture, but using quantum

molecular dynamics and classical interatomic potentials. This probability can be compared with that for the classical saddle point, so that a crossover between mechanisms may be identified. There are, of course, implicit assumptions (I shall note later some contributions to quantum diffusion which this method misses) but it is a very promising general approach. Like other molecular dynamic methods, it has the advantage of finding the diffusion mechanism without necessarily being told it in advance.

Some of the alternative ideas of the pioneering study in [40] will be needed in describing later treatments. First, in a crystal with translational symmetry, the eigenstates will have the same translational symmetry and, one might assume, would be delocalised. The hopping transitions in diffusion, however, presume localised states. This problem is resolved when the local lattice distortion due to the interstitial is recognised: the deformation effectively immobilises the particle to give "sensibly-localised" states between which transitions can occur. These are, of course, vibronic states in which both the site of the interstitial (p) and any phonon variables (v) must be specified. The relevant jump rate observed can be expressed in terms of component jumps $\omega_{pp'}(v, v')$ by averaging over initial vibrational states and summing over all final states:

$$W_{pp'} = \langle \Sigma_{v'} \omega_{pp'}(\{v\}, \{v'\}) \rangle_T \qquad (10)$$

Diffusion — i.e., the incoherent motion of interest here — occurs whenever W is dominated by those processes ($\{v\}, \{v'\}$ not identical) which redistribute energy over the vibrational degrees of freedom. These ideas are fairly general. Flynn and Stoneham [40] made additional working approximations in a theory which generalised small polaron methods. The key assumptions for hydrogen in bcc metals were:

A: The usual Born–Oppenheimer approximation that electrons follow nuclei adiabatically;

B: An adiabatic approximation that separates the rapid hydrogen motion from the sluggish motion of the host metal nuclei (like Nb or Ta);

C: A linear coupling approximation, so that hydrogen displaces the neighbouring atoms without changing modes or frequencies;

D: (for bcc hosts only) the Condon approximation that a (hydgrogen) tunnelling matrix element is independent of the positions of the host atoms.

The limitations of these assumptions were discussed in some detail in the early papers [18, 40, 66]. Yet they allowed analytical results to be derived, and these results were in striking agreement with many experiments: they pointed to links between activation energies and volumes of solution, to the differences between bcc and fcc hosts, to the Zener relations, at least a partial explanation of the isotope effect, and a successful description of the curvature of the Arrhenius plot for bcc hosts without introducing extra parameters.

Neverthless, the story was not complete; some of the assumptions (especially B, C, D, though even A proves inadequate in some situations [19]) are questionable or worse, and the method does not lend itself too well either to the use of the best interatomic potentials or to extensions to systems other than light interstitials. Many papers made improvements in specific aspects, usually at the expense of some complexity. Schober and Stoneham [68, 69] realised that there were advantages in an embedded cluster model, exploit-

ing the fact that the degrees of freedom for which the assumptions are worst are highly-localised local modes (e.g., of H) or resonances (e.g., of host atoms neighbouring H) involving only a few atoms. Again the jump rate can be calculated by explicitly summing the contributions to eq. (10), the contributions being obtained by (a) the calculation of (harmonic) initial and final vibronic wavefunctions for an assumed interatomic potential (including the self-trapping distortion at each site) by explicit calculation of the modes for a cluster of up to 21 atoms; (b) the calculation of the component probabilities ω as transitions induced by that part of the Hamiltonian not diagonalised; (c) thermal average and summation, as in eq. (10); (d) correction for sundry points, like the changes in the jump of the displacements of atoms outside the cluster, and the several ways in which anharmonic corrections must be incorporated.

This approach avoids the previous restrictive assumptions and, moreover, can be extended to heavy interstitials too. Further, the individual contributions of specific processes ($\{v\}, \{v'\}$) can be identified. Thus at lower temperatures (say 50–250 K) the transitions are mainly those without the excitation of local mode quanta. At higher temperatures transitions with local mode excitation (e.g., from $v_L = 0$ to $v_L' = 1$) become important; in due course transitions between excited vibronic states (e.g., $v_L = 1$ to $v_L' = 1$) become dominant. This is in qualitative accord with the simpler occurence probability approach, and is the source of the "knee" in $D(T)$, the change in Arrhenius energy around 250 K.

The explanation of the "knee" from local mode excitation is different from the quantal-classical transition suggested by Gillan [67], whose approach includes only symmetric initial and final states (and different again from generalisations of the curvature of the Arrhenius plot from Flynn–Stoneham theory (e.g., [70, 71])). The several explanations are not contradictory (i.e., all could contribute) and direct comparison is hard because different interatomic potentials were used. However, the quantum molecular dynamics approach does not give the rate directly, and does omit transitions for which the initial and final states are not equivalent (e.g., those with $v_L = 0$ to $v_L' = 1$) which do seem important in the embedded cluster approach.

The Schober–Stoneham results make several predictions, though any detailed comparisons with experiment must recognise the limited accuracy of the interatomic potentials, to which the predictions are very sensitive. The rates and qualitative trends of activation energy with temperature and with isotope are in respectable agreement with experiment. The tunnel frequency for Nb:H can also be compared with that observed for Nb:O, H. Here one can measure the "undressed" value J from internal friction data [71] and the "dressed" value J exp $(-S)$ with polaron corrections from neutron data [72]. The prediction of 0.01 meV (which does not correct for the nearby oxygen) for the dressed value is rather less than the observed 0.2 meV, but the dressing factor of 0.15 is reasonable. The errors are arguably mainly from the limitations of the potentials, in line with the great sensitivity noted in earlier work [69].

7. Conclusions

The major conclusion is that it is possible to predict the rates

of key solid state processes, and not merely their characteristic energies. Let us look back at what the important features are. The most convenient route involves (a) some algorithm for predicting an energy surface, i.e., the energy for any desired positions of nuclei, and (b) the solution of a dynamical problem (at least vibrational frequences, sometimes vibrational wavefunctions, and occasionally the integration of the equations of motion). The algorithm for the energy must be good enough to give constant pressure results — including entropies — and to give the relative energies of different charge states well. There will usually be other thermodynamic and kinetic relationships needed to relate results to experiment.

For the present paper the energy surface has been obtained from interatomic potentials which, of course, can be based on various levels of use of the Schrödinger equation. For ionic crystals, the shell model is especially convenient (though not so conveniently for molecular dynamics, which restricts the use of that method) and of high demonstrated accuracy. For metals other than simple metals and for semiconductors there is no model of comparable accuracy yet, and one should be cautious about some simplistic analyses.

What emerges from this work is the realisation that, in the best cases, diffusion rates can be predicted at least as well as they can be measured. This seems to be true even when there are quantum effects in the dynamics, and provides a basis for understanding some of the successes and failures of the main empirical rules of diffusion. This understanding and ability to make quantitative predictions is important when one notes the range of phenomena of broader interest to which these ideas apply: to the extreme conditions of temperature and pressure involved in geophysics and in studies related to reactor safety, to engineering on the nanometer scale and the direct observations of diffusion jumps by scanning tunnelling microscopy, and indeed to open systems (where there are further complexities [73]) including biological processes where there can be circumstances in which theory has advantages over experiment alone.

Acknowledgements

References

1. Boyle, R., Experiments and Considerations about the Porosity of Bodies. London: St Paul's Churchyard (1684).
2. Boer, J. M., Disc Farad Soc **23**, 171 (1957).
3. Watkins, G. D., Corbett, J. W. and MacDonald, R. S., J. Appl. Phys. **53**, 7097 (1982).
4. Flynn, C. P., Defects and Diffusion in Solids. Oxford University Press (1972).
5. Larsen, U., Phys. Rev. **B32**, 1772 (1985).
6. Meyer, W. and Neldel, H., Z. Tech. Phys. **18**, 588 (1937).
7. Harding, J. H., Sangster, M. J. L. and Stoneham, A. M., J. Phys. **C20**, 5281 (1987).
8. Banks, B. E. C., Damjanovic, V. and Vernon, C. A., Nature **240**, 147 (1972).
9. Irsigler, P., Wagner, D. and Dunstan, D. J., J. Phys. **C16**, 6605 (1983).
10. Allnatt, A. and Lidiard, A. B., Rep. Prog. Phys. **50**, 373 (1987); Howard, R. E. and Lidiard, A. B., Rep. Prog. Phys. **27**, 161 (1964).
11. Penrose, O., Foundations of Statistical Mechanics. Pergamon, Oxford (1979).
12. Leibowitz, J. L. and Penrose, O., Physics Today (February) 23 (1973).
13. Kramers, H. A., Physica **7**, 284 (1940).
14. Carmell, B. and Nitzan, A., Phys. Rev. Lett. **25**, 233 (1983).
15. Caldiera, A. O. and Leggett, A. J., Phys. Rev. Lett. **46**, 211 (1981).
16. Gillan, M. J. and Holloway, R. W., J. Phys. **C18**, 4903 (1985).
17. Movaghar, B., Murray, D. W., Pohlmann, B. and Wurtz, D., J. Phys. **C17**, 1677 (1984).
18. Stoneham, A. M., Berichte der Bunsenges. **76**, 816 (1972).
19. Kondo, J., Physica **124B**, 25 (1984).
20. Muller, K., Angew. Chem. Int. Ed. **19**, 1 (1980).
21. Jug, K., Theor. Chim. Acta **54**, 263 (1980).
22. Stoneham, A. M., Sangster, M. J. L. and Rowell, D. K., Adv. Ceram. **10**, 79 (1984).
23. Jacucci, G., Diffusion in Crystalline Solids (Edited by G. E. Murch and A. S. Nowick), Academic Press, New York (1984).
24. Gillian, M., Harding, J. H. and Tarento, R., J. Phys. **C20**, 2331 24 (1987).
25. de Lorenzi, G. and Jacucci, G., Phys. Rev. **B33**, 1993 (1986); See also: Jacucci, G., Toller, M., De Lorenzi, G. and Flynn, C. P., PRL **52**, 295 (1984); Marchese, M., De Lorenzi, G., Jacucci, G. and Flynn, C. P., Phys. Rev. Lett. **57**, 3280 (1986); Flynn, C. P. and Jacucci, G., Phys. Rev. **B25**, 6225 (1982); De Lorenzi, G., Jacucci, G. and Flynn, C. P., Phys. Rev. **B36**, 9461 (1987); Marchese, M., Jacucci, G. and Flynn, C. P., Phys. Rev. **B36**, (1987).
26. Slater, N. B., Theory of Unimolecular Reactions. Cornell University Press (1959).
27. Rice, S. A., Phys. Rev. **112**, 804 (1958).
28. Feit, M., Phys. Rev. **B3**, 1223 (1971).
29. Kac, M., Am. Math. J. **65**, 609 (1943).
30. Doll, J. D., J. Chem. Phys. 2760 (1980).
31. Oates, W. A., Mainwood, A. M. and Stoneham, A. M., Phil. Mag. **A38**, 607 (1977).
32. Hirschenfelder, J., Ann. Rev. Phys. Chem. **34**, 1 (1983).
33. Glasstone, S., Laidler, K. J. and Eyring, H., Theory of Rate Processes. McGraw Hill, New York (1941).
34. Wert, C. A. and Zener, C. Phys. Rev. **76**, 1169 (1949).
35. Stoneham, A. M., Theory of Defects in Solids. Oxford University Press, Oxford (1985).
36. Vineyard, G. H., J. Phys. Chem. Sol. **3**, 121 (1957).
37. Sangster, M. J. L. and Stoneham, A. M., J. Phys. **C17**, 6093 (1984).
38. Harding, J. H., Physica **131B**, 15 (1985).
39. Holstein, T., Ann. Phys. (NY) **8**, 343 (1959).
40. Flynn, C. P. and Stoneham, A. M., Phys. Rev. **B31**, 3966 (1970).
41. Catlow, C. R. A. and Stoneham, A. M., J. Phys. **C16**, 4321 (1983).
42. Sempolinski, D. R. and Kingery, W. D., J. Am. Ceram. Soc. **63**, 11 (1980).
43. Catlow, C. R. A., Corish, J., Jacobs, P. W. M. and Lidiard, A. B., J. Phys. **C14**, L121 (1981).
44. Harding, J. H., Phys. Rev. **B32**, 6861 (1985); Adv. Ceram. (1988) in press.
45. Dick, B. G. and Overhauser, A. W., Phys. Rev. **112**, 90 (1958).
46. Stoneham, A. M. and Harding, J. H., Ann. Rev. Phys. Chem. **37**, 53 (1987).
47. Norgett, M. J. and Lidiard, A. B., Computational Solid State Physics (Edited by F. Herman, N. W. Dalton and T. R. Koehler), p. 385, Plenum Press (1972); Norgett, M. J., AERE Report R7650 (1974); AERE Report M2880 (1977).
48. Catlow, C. R. A. and Mackrodt, W. C. (Editors) Computer Simulation of Solids, Lecture Notes in Physics 166, Springer (1982).
49. Harding, J. H., Sangster, M. J. L., Stoneham, A. M. and Tarento, R. -J., to be published.
50. Almond, D. P. and West, A. R., Nature **306**, 456 (1983).
51. Huberman, B. A. and Boyce, J. B., Solid State Commun. **25**, 843 (1978).
52. Catlow, C. R. A., Non-Stoichiometric Oxides (Edited by O. T. Sørensen), Academic Press, New York (1981).
53. Stoneham, A. M., Tomlinson, S. M., Catlow, C. R. A. and Harding, J. H., Physics of Disordered Materials (Edited by D. Adler, H. Fritsche and S. R. Ovshinksy), p. 243, Plenum Press, NY (1985).
54. Stoneham, A. M. and Sangster, M. J. L., Phil. Mag. **B52**, 717 32 (1985).
55. Stoneham, A. M. and Sangster, M. J. L., Phil. Mag. **B43**, 609 (1980).
56. Harding, J. H., J. Phys. **C19**, L731 (1986).
57. Baraff, G., Schluter, M. and Allan, G., Phys. Rev. Lett. **47**, 601 (1981).

58. Stoneham, A. M., Phil. Mag. **B51**, 161 (1985).
59. Fair, R. B., Impurity Doping Processes in Silicon, Chap. 7, North Holland, New York (1981).
60. Seeger, A. and Chik, K. P., Phys. Stat. Sol. **29**, 455 (1968).
61. Stoneham, A. M., Mat. Sci. Forum. **10-12**, 9 (1986).
62. Stoneham, A. M., Hyp. Int. **17-19**, 53 (1984); Browne, A. M. and Stoneham, A. M., J. Phys. **C15**, 2709 (1982).
63. Stoneham, A. M., Rep. Prog. Phys. **44**, 1251 (1981).
64. Kimerling, L. C., Solid State Electronics **21**, 1391 (1978).
65. Masri, P. M. and Stoneham, A. M., J. Electonic Mat. **14**, 205 (1984).

66. Stoneham, A. M., J. Nucl. Mat. **69/70**, 109 (1977).
67. Gillan, M. J., J. Phys. **C20**, 3621 (1987).
68. Schober, H. R. and Stoneham, A. M., Phys. Rev. Lett. **60**, 2307 (1988); Schober, H. R. and Stoneham, A. M., Hyp. Int. **31**, 141 (1986).
69. Schober, H. R. and Stoneham, A. M., Phys. Rev. **B26**, 1819 (1982).
70. Stoneham, A. M., J. Phys. **F2**, 417 (1972).
71. Zapp, P. E. and Birnbaum, H., Acta Met. **28**, 1523 (1980).
72. Wipf, H. and Neumaier, K., Phys. Rev. Lett. **52**, 1308 (1984).
73. Landauer, R., Helv. Phys. Acta **56**, 847 (1983); (also Phys. Today, Nov 1978 p. 23).