Existence of Two Types of Insulating States in the Hubbard Model*

Tadashi Arai
Argonne National Laboratory, Argonne, Illinois 60439
(Received 1 April 1974)

If a band is narrow and is not nearly empty nor nearly full, the Hubbard lattice is shown to be an insulator. The lattice containing about one electron per atom will be a Hubbard-type insulator with a wide energy gap but if the number of electrons decreases or increases beyond a certain range, the gap is suddenly reduced, yielding a small-gap insulator of a completely new type.

Hubbard[1] has shown that, in the presence of a strong intra-atomic repulsive interaction I, a narrow conduction band splits into two and that, if the lower band is filled and the upper band empty, the lattice is an insulator. Although the Hubbard treatment is regarded as one of the most promising approaches for d electrons in transition metals, many puzzling questions on the nature of the Hubbard solution are unanswered.

If the lower (or upper) band is partly filled, the lattice is metallic. According to Herring,[2] however, the ratio between the Fermi-surface volume and the number of electrons deviates from the value predicted for weakly interacting electrons as much as by a factor of 2 when the number of electrons N is nearly equal to the number of atoms N_a in the lattice. This is contradictory to Luttinger's theorem[3] that the Fermi-surface volume is unchanged by electron interaction to all orders of perturbation. The Hubbard insulating state appears only when the ratio N/N_a is exactly equal to a certain number, say n. Since n can be less or greater than 1, the lattice with one electron per atom may not be an insulator. As the density of electrons per volume increases, the energy gap disappears, yielding a metallic state with two overlapping bands. However this state again maintains the abnormal behavior of the Fermi-surface volume.

In this Letter, we shall show that the Hubbard lattice, having excess charges properly compensated, remains insulating unless the original conduction band is nearly empty or nearly full. The lattice with nearly one electron per atom will be in the Hubbard insulating state with a large energy gap, Δ_I, even if a fractional number of electrons per atom are added to (or removed from) the lattice. If the Hubbard lower (or upper) band is partly filled, the lattice is in a coherent (localized) state and a small, but finite, energy Δ_ϵ ($\Delta_\epsilon < \Delta_I$) is required to remove an electron, making the lattice a small-gap semiconductor. This is in line with Mott's insulating state,[4] and the Hubbard metallic states with the improper behavior of the Fermi-surface volume may never exist.

To find the Hubbard insulating state, it is essential to include explicitly the strong interaction I which appears whenever two electrons with opposite spins σ and $\bar{\sigma}$ meet at the same atomic site. Hubbard has calculated it by using the Green's functions of the type $\langle\langle C_{R\sigma}(t)N_{R\bar{\sigma}}(t); C_{R'\sigma}{}^{\dagger}(t')\rangle\rangle$, where $C_{R\sigma}$ and $C_{R\sigma}{}^{\dagger}$ are destruction and creation operators of an electron σ at site R and $N_{R\sigma}=C_{R\sigma}{}^{\dagger}C_{R\sigma}$. As simultaneous motion of other electrons is included by calculating higher order terms such as Hubbard's electron-scattering and resonance-broadening correlations, there appear posibilities that the same electron σ returns to the same site after various collisions. If it is essential to calculate the interaction explicitly at the first time an electron σ appears at a site R, it is equally important to include it every time the same electron σ comes back to R.[5] Even in the most advanced Hubbard treatment, the successive interactions are neglected because of decoupling approximations used. To include the successive interaction, the higher-order Green's functions have to be treated on the same basis as the single-particle Green's function G.

As has been discussed previously,[6] the higher-order Green's function can be reduced to functional derivatives of the one-particle Green's function G in the presence of a small external field. If the simple Green's function G^0 obtained by Hubbard is used in calculating the derivatives and hence the self-energy correction to G^0, the seemingly difficult task of handling the higher-order Green's functions on the same basis as G may be carried out satisfactorily, and the result obtained will retain the Hubbard interaction explicitly up to infinite order in hoppings.

The resulting Green's function is

$$2\pi G^{-1}(k\sigma,\omega)=\frac{(\omega\mp\omega_1)(\omega\mp\omega_2)}{[\omega\mp(1-n_{\bar\sigma})I]}\mp\frac{I^2 g(k\sigma)}{[\omega\mp(1-n_{\bar\sigma})I]^2}, \quad (1)$$

with the upper and lower signs for the electron and hole Green's functions, respectively, and

$$g(k\sigma)=\lim_{\tau=0^{\pm}} iG_{RR\sigma}(\tau)\frac{1}{N_a}\sum_q \epsilon_q n_{q\bar\sigma}+n_{\bar\sigma}(1-n_{\bar\sigma})\epsilon_k +\frac{1}{N_a^2}\sum_{q_1}\sum_{q_2}\epsilon_{k-q_1+q_2}n_{q_1\bar\sigma}n_{q_2\bar\sigma}, \quad (2)$$

where ω_1 and ω_2 are the Hubbard simplest solutions for the lower and upper bands, respectively, ϵ_k or ϵ_q is the Fourier transform of the hopping matrix elements $\epsilon_{RR'}$, $n_{q\sigma}=\langle C_{q\sigma}{}^{\dagger}C_{q\sigma}\rangle$, and $n_\sigma=\langle N_{R\sigma}\rangle$. The origin of energy is chosen at the mean of ϵ_k so that $\sum_k\epsilon_k=0$. The most important aspect of the above result is that the self-energy correction represented by $g(k\sigma)$ contains the Green's function $G_{RR\sigma}(\tau)\equiv\langle\langle C_{R\sigma}(\tau)C_{R\sigma}{}^{\dagger}(0)\rangle\rangle$. If the electron Green's function $G_{\rm el}$ is being calculated, τ should remain positive and $iG_{RR\sigma}(\tau=0^+)=1-n_\sigma$; while for the hole Green's function $G_{\rm hole}$, $iG_{RR\sigma}(\tau=0^-)=-n_\sigma$. Hence poles $\omega_i{}^{N+1}$ of $G_{\rm el}$ will be different from the corresponding poles $-\omega_i{}^{N-1}$ of $G_{\rm hole}$, yielding the distinct energy spectra, $\xi_i{}^{N+1}$ for an electron and $-\xi_i{}^{N-1}$ for a hole, where

$$\pm\xi_i{}^{N\pm1}=\pm\omega_i{}^{N\pm1}-\mu(N) \quad (3)$$

and $\mu(N)$ is the chemical potential.

The solutions $\omega_i{}'\equiv\pm\omega_i{}^{N\pm1}$ of the cubic equation $G^{-1}=0$ may be estimated by the intersections of two curves $h(k\sigma,\omega)=(\omega-\omega_1)(\omega-\omega_2)[\omega-(1-n_{\bar\sigma})I]$ and $I^2g(k\sigma)=$ const as is shown in Fig. 1. In the narrow-band limit, $g(k\sigma)$ is small and the solutions $\omega_i{}'$ and $\omega_2{}'$ are not different from the Hubbard results ω_1 and ω_2. Even if the bandwidth $2D$ increases, the inequality $\omega_1<(1-n_{\bar\sigma})I<\omega_2$ remains valid and the shape of the h curve will not change drastically. If $2D$ exceeds a certain limit, say λI, however, two of the three intersections disappear, being replaced by two complex solutions, where λ is of the order 0.27 for an electron added to the lattice with a nearly half-filled lower band ($N\approx N_a/2$).

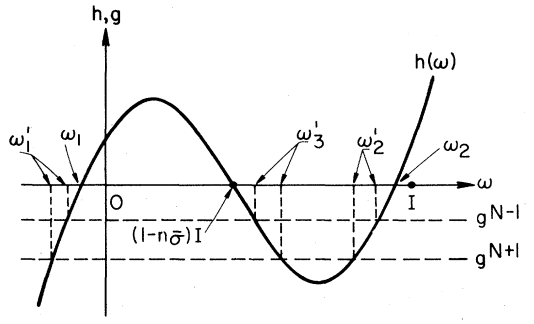

FIG. 1. Plot of two curves $h(k\sigma,\omega)$ and $g(k\omega)=$const. The intersections give the solutions of $G^{-1}=0$.

Let us consider the case where the lower band is partly filled ($N<N_a$). Since, in the narrow-band limit, $n_{q\bar\sigma}$ is $\sim\frac{1}{2}$ for occupied states and zero otherwise, the gap parameter $d\equiv N_a{}^{-1}\sum_q\epsilon_q n_{q\bar\sigma}$ involved in $g(k\sigma)$ will be of the order $(N/2N_a)\epsilon_{\rm av}$, where $\epsilon_{\rm av}$ is the density average of ϵ_q for occupied states. Since the origin of energy is chosen at the center of ϵ_k, d will be negative and of the order $1/N_a$ for a nearly empty lower band, of the order $D/2\pi$ for a half-filled lower band, and a small but finite negative value for a nearly filled lower band.

Let $g^{N+1}(k\sigma)$ and $g^{N-1}(k\sigma)$ be the values of g involved in $G_{\rm el}$ and $G_{\rm hole}$, respectively. Because of $iG_{RR\sigma}(\tau)$ involved, the first term on the right-hand side of Eq. (2) is negative for $g^{N+1}(k\sigma)$ and positive for $g^{N-1}(k\sigma)$, yielding that $g^{N-1}(k\sigma)-g^{N+1}(k\sigma)=|d|=(N/2N_a)|\epsilon_{\rm av}|>0$. Hence, $-\omega_1{}^{N-1}>\omega_1{}^{N+1}$, $-\xi_1{}^{N-1}>\xi_1{}^{N+1}$, and the difference, $\delta\xi_1{}^N=(-\xi_1{}^{N-1})-\xi_1{}^{N+1}\approx I^2|d|/(\delta h/\delta\omega)_{\omega=\omega_1}$, becomes of the order $1/N_a$ for a nearly empty lattice, and of the order $(4/3)|d|\approx(2/3\pi)D$ for a nearly half-filled lower band. The spectral weight function of the lower band for an electron,

$$A_1{}^{N+1}(k\sigma)=[\omega_1{}^{N+1}\mp(1-n_{\bar\sigma})I]^2(\omega_1{}^{N+1}-\omega_2{}^{N+1})^{-1}\times(\omega_1{}^{N+1}-\omega_3{}^{N+1})^{-1},$$

is also smaller than the corresponding spectral weight function for a hole, $A_1{}^{N-1}(k\sigma)$, since $(\delta h/\delta\omega)_{\omega=\omega_3}$ is negative and its value is less than $(\delta h/\delta\omega)_{\omega=\omega_1}$ by the factor $1-n_{\bar\sigma}$. The difference $\delta A_1{}^N\equiv A_1{}^{N-1}-A_1{}^{N+1}$ is of the order $(40/9)\times(d/I)$ for a nearly half-filled lower band.

The foregoing calculation has profound consequences. The N-electron lattice in the ground state may be constructed by adding electrons one by one and the chemical potential $\mu(N)=\omega_1{}^{N+1}(k_F\sigma)$ and the Fermi momentum k_F are determined by $G_{\rm el}$ by the relation

$$2N_a-N=-\frac{1}{\pi}\lim_{\tau\to0^+}\sum_{k,\sigma}\int_{-\infty}^{\infty}d\omega\,{\rm Im}G_{\rm el}(k\sigma,\omega)=\sum_{k\sigma}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}A_1{}^{N+1}(k\sigma)[1-f(\omega)], \quad (4)$$

or

$$N=\sum_{k\sigma}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}A_1^{N+1}(k\sigma)f(\omega), \qquad (4')$$

where $f(\omega)=1/\{\exp[(\omega-\mu)/k_BT]-1\}$. The excitation spectrum of an electron *added to* the lattice is calculated by G_{el} and the ground-state energy is given by $\int_0^N\mu(N)dN$. However, the energy spectrum of an electron *removed from* the lattice has to be computed by G_{hole}. Since $A_1^{N-1}(k\sigma) > A_1^{N+1}(k\sigma)$, the right-hand side of Eq. (4') becomes greater than N if A_1^{N+1} is replaced by A_1^{N-1}. Therefore the hole state $k_{max}{}^{N-1}$ with the maximum energy $\xi_{max}{}^{N-1}$ cannot be k_F but $k_{max}{}^{N-1} < k_F$ and it should be calculated by

$$N=\sum_{k,\sigma}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}A_1^{N-1}(k\sigma)f(\omega). \qquad (5)$$

This is possible because, after losing an electron, the lattice of $N-1$ electrons is no longer in the ground state but in the excited state with a hole in $k_{max}{}^{N-1}$. The energy $-\xi_1^{N-1}(k\sigma)$ of a removed electron calculated by G_{hole} is then higher by the amount $\delta\xi_1^N$ than the corresponding energy $\xi_1^{N+1}(k\sigma)$ obtained by G_{el}, suggesting that an activation energy $\Delta_\epsilon(N)=\delta\xi_1^N$ supplied externally is needed to remove the electron.[7]

In conclusion, the ground state of the Hubbard lattice is a coherent (localized) state with a forbidden band $k_{max}{}^{N-1}\sim k_F{}^N$ between the occupied and unoccupied spectra; the activation energy $\Delta_\epsilon(N)$ and the momentum $k_F-k_{max}{}^{N-1}$ are needed to excite an electron into an unoccupied state, making the lattice an insulator with a small indirect gap $\Delta_\epsilon(N)$. The temperature dependence of the conductivity will be proportional to $\exp(-\Delta_\epsilon/2k_BT)$ of the carrier concentration.

Even if the number of electrons increases, the lattice will remain insulating. When the lower band is nearly half-filled the gap $\Delta_\epsilon(N)$ will attain its maximum value of $(2/3\pi)D$ (which may be ~0.3 eV if the original bandwidth $2D$ is of the order 3 eV) and then decrease gradually. Even if the lower band for electrons is filled with N' electrons, the "hole" band will be terminated at the state $k_{max}{}^{N'-1}$ and the small gap parameter d for G_{hole} will remain negative and finite. Consequently, the energy required to remove an electron from the lattice is equal to the activation energy $\Delta_\epsilon(N)$ but the lowest energy level into which the electron can be excited is in the upper band with energy $\xi_2^{N'+1}(k_0\sigma)$, thus suddenly increasing the activation energy to a large value $\xi_2^{N'+1}-\xi_1^{N'+1}\approx\Delta_I$. This energy gap is the same as the gap involved in the Hubbard result in magnitude as well as in physical origin. Even if more electrons are added to the lattice, however, the lattice remains a Hubbard insulator. Only after $\Delta N'$ electrons are added and the lower "hole" band is extended to the zone boundary does the lattice return to the small-gap semiconductor. By this time, $\Delta N'$ states in the lower part of the upper band for electrons are filled, yielding a finite negative value for the gap parameter d involved in G_{el}, and the arguments for the small-gap semiconductor developed for the lower band will be applied to this case. Only if the lattice is nearly empty or full does the gap parameter d vanish as $1/N_a$ and the lattice become metallic with the conventional Fermi surface.

As the bandwidth $2D$ increases, the gap parameter d increases, stabilizing the small-gap insulating state. If $2D$ becomes much greater than λI, however, energies in the upper half of the lower band and in the lower half of the upper band become complex, introducing finite widths in the spectra and, at the same time, reducing their spectral weights. Then the above tendency is reversed and eventually the lattice will become a normal metal with a single band, but a more precise calculation is needed to discuss the metal-nonmetal transition.

So far we have neglected explicit consequences of the third solution ω_3' since, in the narrow-band region, its spectral weight is of the order $(d/I)^2$ and smaller than the quantities δA and $\delta\xi$ of the order d/I needed in obtaining the foregoing conclusion. If the bandwidth increases, this is no longer the case. We have calculated the inverse Green's function correctly up to terms linear in ϵ_k by extending the perturbation up to an infinite order and found that the effect of ω_3' can be included properly and that the present conclusions remain correct.

*Based on work performed under the auspices of the U. S. Atomic Energy Commission.

[1]J. Hubbard, Proc. Roy. Soc., Ser. A 276, 238 (1963), and 281, 401 (1964).

[2]C. Herring, in *Magnetism*, edited by G. T. Rado and H. Suhl (Academic, New York, 1966).

[3]J. M. Luttinger, Phys. Rev. 119, 1153 (1960).

[4]N. F. Mott, Phil. Mag. 6, 287 (1961).

[5]This mechanism is parallel to the mechanism leading to Anderson's localizability of electrons in random lattices. See P. W. Anderson, Phys. Rev. 109, 1492 (1958); E. N. Economou and M. H. Cohen, Phys. Rev. B 5, 2931 (1972); F. Yonezawa and M. Watabe, Phys. Rev. B 8, 4551 (1973).

[6]T. Arai and M. Tosi, Solid State Commun. 14, 947 (1974).

[7]If $N=1$, the energy of an electron added to the lattice is $\epsilon_k-\Delta$ with $\Delta>0$, because of the attractive interaction between two electrons involved, but the energy of an electron removed from the lattice with $N=1$ is ϵ_k since only a single electron is present. This illustrates that the difference Δ is the activation energy needed to break up the bound pair. This two-electron N-site problem was originally solved exactly by J. C. Slater, H. Statz, and G. F. Koster, Phys. Rev. 91, 1323 (1953). The present result suggests that the same attractive interaction appears between the lattice with $N\rightarrow\infty$ and an electron added.

Observation of Recombination-Enhanced Defect Reactions in Semiconductors

D. V. Lang and L. C. Kimerling
Bell Laboratories, Murray Hill, New Jersey 07974
(Received 26 June 1974)

Recombination-enhanced annealing of defects in semiconductors has been observed directly for the first time. The defects were produced in GaAs by 1-MeV electron irradiation and observed by transient-junction-capacitance techniques. The data clearly relate the enhanced defect annealing rate to electron-hole recombination processes at the defect.

We have observed a new mechanism for the enhancement of defect reactions in semiconductors, namely enhancement by electron-hole recombination at the defect. A direct correlation of defect reactions with electron-hole recombination processes at the defect has not previously been observed. The details of such recombination-enhanced processes are crucial to the basic understanding of both defect motion and nonradiative recombination phenomena in semiconductors.

The recombination-enhanced mechanism was observed to produce a significant increase in the annealing rate of 1-MeV-electron irradiation defects in n-GaAs under conditions of minority-carrier injection. The specific identity of these defects is at this time unknown, but it is reasonable to expect that they are isolated vacancies, interstitials, and/or simple complexes. These radiation-induced defects are observed by a new junction-capacitance technique, deep level transient spectroscopy (DLTS).[1] With this technique it is possible to measure for each defect the activation energy for thermal emission of a carrier to the nearest band edge, the concentration, and the capture rates for electrons and holes. These properties are highly specific to a particular defect and allow us to resolve five levels whose activation energies for electron emission to the conduction band are 0.08, 0.19, 0.45, 0.76, and 0.96 eV and three levels with activation energies for hole emission to the valence band of 0.32, 0.44, and 0.76 eV. A typical DLTS spectrum is shown in Fig. 1. The full details of the properties of these defects will be published elsewhere.[2] For the purpose of demonstrating recombination enhancement we will focus on the 0.45-eV electron trap shown in Fig. 1.

Junction-capacitance techniques such as DLTS are ideal for these studies since it is possible to independently vary both the average charge state and the electron-hole (e-h) recombination rate at a particular defect. The charge state can be var-

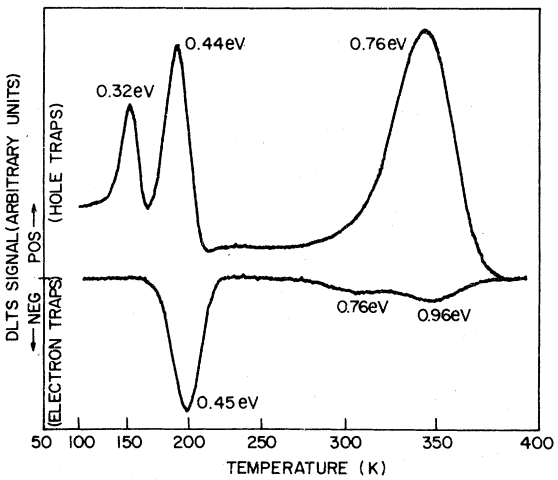

FIG. 1. A typical DLTS spectrum of 1-MeV electron irradiated n-GaAs. The energies shown are the measured activation energies for emission of a carrier to the nearest band edge. The positive signals are due to hole traps (injection pulse scan) while the negative signals are due to electron traps (majority-carrier pulse scan).