MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

REACTION AT A REFRACTORY METAL SEMICONDUCTOR INTERFACE: V/GaAs(110)

Microelectronic and Information Sciences Center

Technical Report #21

M. Grioni
Department of Chemical Engineering
and Materials Science
University of Minnesota

J. Joyce
Materials Science Program
University of Wisconsin, Madison

J. H. Weaver
Department of Chemical Engineering
and Materials Science
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

REACTION AT A REFRACTORY METAL
SEMICONDUCTOR INTERFACE: V/GaAs(110)

MEIS Technical Report #21

M. Grioni
J. Joyce
J. H. Weaver

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

[1984?]

ABSTRACT

Synchrotron radiation photoemission spectroscopy has been used to study the formation of the reactive V/GaAs(110) interface. Valence band and core level results indicate that metal deposition produces an extended intermixed phase involving the formation of both V-Ga and V-As bonds. In this reacted region the Ga 3d core line exhibits a continuous shift to lower binding energy (total shift 1.55 eV over band bending) indicative of a variable chemical environment, while analysis of the As 3d lineshape suggests that As is present in two well-defined chemical states. Core level intensity profiles show preferential outdiffusion of arsenic, with As present at ~6% of the original level with coverages of 110 Å. Comparison to previous results for Cr/GaAs(110) shows similar Ga and As attenuation profiles.

PACS: 73.40.+y
73.40.-c
73.40.Ns
79.60.Gs

In the last decade, metal-semiconductor interfaces have been intensely studied for both fundamental and technological reasons[1,2]. Among the goals of this research are the description of electronic interactions that take place in the critical transition region between bulk metal and bulk semiconductor and the understanding of how experimental conditions affect the products, the kinetics and even the existence of such reactions. Synchrotron radiation photoemission is a particularly suitable technique for this purpose because it is an intrinsically local probe, having the possibility of modulating the depth of the region under investigation and achieving high resolution for core level studies.

In this paper we examine the formation of the interface between a compound semiconductor and a refractory metal. A motivation for this experiment comes from the fact that, while metal/semiconductor junctions with noble and near-noble metals have been extensively studied, much less is known about systems involving refractory metal overlayers.[1] Moreover, technological interest in refractory metal overlayers is growing because of applications in integrated circuit technology.

Recent studies employing surface analytical techniques have shown that refractory metals can induce reactions on Si and GaAs substrates even at room temperature[3-7]. On the other hand, Clabes _et al_.[8] observed layer by layer growth of V on Si(111) and Si(100) with no evidence for room temperature chemical reaction. Hence it is interesting to determine whether V exhibits the same low reactivity on the GaAs(110) surface.

Photoemission studies were conducted at the Wisconsin Synchrotron Radiation Center using radiation dispersed by a 3m toroidal grating monochromator or a Grasshopper monochromator for $12<h\nu<85$ eV. The photoelectrons were energy analyzed by a double pass CMA to give a total energy resolution (monochromator + analyzer) of 280meV and 400 meV at hν=60eV and hν=85eV, respectively. Clean GaAs(110) surfaces 4mmx4mm were obtained by cleaving prenotched posts of n-type GaAs (Si doped at $4x10^{18}$ cm^{-3}) in a UHV system operating at $3x10^{-11}$ Torr. The quality of each cleave was first judged visually and then spectroscopically by checking the sharpness of the valence band features and the absence of pinning of the Fermi level at the surface[9].

Vanadium was evaporated from a resistively-heated 5 mil Mo boat requiring 100-110 amps from a regulated power supply. Stable rates of approximately 1Å/min were typically obtained. For the lowest coverages, slower rates were chosen and the sample surface was exposed to the evaporation only after the rate was proven to be stable over several minutes. After suitable outgassing, the pressure never rose above $1.5x10^{-10}$ Torr during evaporation. The thickness of the deposited overlayer was monitored by an oscillating quartz crystal. Vanadium coverages ranging from 0.25Å to 140Å were studied with repetitive cleavage and deposition cycles. Coverages will be defined as 1Å = $6.85x10^{14}$ at/cm^2 = 0.77 ml, were 1 ml is referred to the surface density of the substrate $8.9x10^{14}$at/cm^2.

In Fig.1 we show valence band photoelectron energy distribution curves (EDCs) at hν=21eV after subtraction of the inelastic background. The bottom-most curve represents the clean GaAs(110) surface and the EDCs are shifted upwards with increasing V coverage. Each is aligned at the valence band maximum using the characteristic bulk feature "A" as a reference. The vertical scale is adjusted for incident photon flux so that comparisons of relative intensities are meaningful.

The deposition of the vanadium overlayer is accompanied by smearing and rapid attenuation of the emission from the top of the valence band and by the growth of the d-band derived emission within 2 eV of E_F. However, the cutoff of these metal derived states is not coincident with E_F at the lowest coverages and a metallic Fermi edge is clearly seen only above 3Å. Structure "B", which is not present in the clean V valence band[8,10], is seen to grow near -3.4eV for Θ>3Å and is suggestive of hybrid V-GaAs states. All features of the clean surfaces have disappeared by 16Å and the valence band is dominated by V d-states near E_F and a prominent peak at -3.4eV. Subsequent evolution of the valence band is slow and convergence to that of bulk vanadium occurs only above 100Å, when the V feature "C" at -2.1 eV dominates the reaction-derived peak "B".

Fig.2 shows the evolution of the Ga 3d core lines. Again the spectra are referred to the top of the valence band so that the measured shifts are of chemical nature and above variations in band bending. The bottom-most curve represents the clean surface; the Ga 3d spin-orbit doublet is not resolved because of the contribution of the surface-shifted cores (0.28 eV to greater binding energy[11]). The EDCs have been scaled to approximately the same height to demonstrate changes in lineshape. A slight broadening of the peak is seen on the low-binding energy side at Θ=0.5Å. By Θ=1Å, a shoulder develops which is indicative of reaction, band bending is fully established (0.65 eV), and the Fermi level is pinned at midgap. The reaction-induced shoulder is clearly visible by 2Å and, in the same spectrum, quenching of the surface contribution is evident from the decrease of emission in the region between the two spin-orbit-split components. The reacted peak disperses with coverage to lower binding energy and, at the same time, the contribution from the substrate is rapidly attenuated. The position of the reacted peak saturates at $\Theta \simeq 30Å$, corresponding to a total shift of 1.55 eV.

The coverage dependence of the As 3d core line is reported in Fig. 3. The EDC for the clean surface reveals the characteristic shoulder on the low binding energy side which originates from As atoms at the surface (surface core shift=0.37 eV[11]). At $\Theta=0.5Å$ this shoulder is already considerably attenuated. A new structure becomes visible by 2Å, shifted ~600meV from the main $3d_{5/2}$ peak, and is clearly resolved by $\Theta=10Å$. At coverages $6<\Theta<30$ Å, the lineshape is complex. The presence of two relatively sharp peaks and a well-resolved shoulder on the high energy side suggest, however, that a fit of the EDCs can be attempted using only two doublets, separated by 600 meV. A semi-quantitative lineshape analysis, based on doublets with strengths indicated by the length of the arrows of Fig. 3, reproduced the experimental lineshapes quite well.

The attenuation of the total Ga and As core emission is shown in Fig. 4 where we plot the normalized integrated intensities, $\ell n[I(\Theta)/I(0)]$, obtained after background subtraction and correction for photon flux. The Ga 3d attenuation indicates that the Ga content of the surface region for $\Theta \simeq 50Å$ is approximately 2% of that for the clean surface. For both Ga and As, the observed deviation from exponential attenuation indicates substantial atom intermixing in the V overlayer. As shown, the As 3d core intensity decreases much more slowly than Ga, indicating greater outdiffusion of As than Ga.

The best evidence that reaction is occurring at low coverages is given by the low-binding-energy component in the Ga 3d core for coverages larger than ~1Å. This can be compared to detailed results for the Cr/GaAs(110) interface[3], where we demonstrated the existence of a critical coverage

(~2ml≈2Å) for the interface reaction. This critical coverage separated the low-coverage regime, characterized by weak metal-substrate interaction, from an intermediate coverage region where extensive interdiffusion takes place. The results of Fig. 1 show that such a precursor stage for V/GaAs is confined to coverages smaller than 0.5Å.

Reaction proceeds above 1Å with the disruption of the GaAs surface. The appearance in the Ga 3d spectrum of a low-binding energy component can be associated with reaction. As was the case for Cr/GaAs, this second component shifts continuously throughout the coverage range shown in Fig. 2 and the final position is well beyond that of metallic Ga (1.55 eV with respect to bulk GaAs, vs.~0.9 eV for Ga droplets[12]). We propose therefore, that a V/Ga intermetallic forms, analogous to Cr/GaAs. The rapid attenuation of the Ga 3d core intensity shows that Ga diffusion through the V overlayer is small at room temperature and that the Ga concentration in the intermixed phase changes quickly with coverage. As a consequence the local environment of the Ga atoms changes from Ga-rich to V-rich. On the basis of simple arguments based on Pauling's electronegativity differences between Ga(1.81) and V(1.63), we expect charge transfer from V to Ga, and indeed the shift of the Ga 3d peak to lower binding energy indicates that such charge transfer exists. Moreover, the continuous shift in binding energy away from the metallic Ga position allows one to follow the progressive dilution of Ga in the V film and the continuous change in the local chemical environment.

Analysis of the As 3d EDCs of Fig. 3 clearly shows that the interaction of V with As is substantially different from Va with Ga. The main effect of metal deposition above 1-2Å is the growth of a low-binding energy peak at fixed position. The energy shift of this peak relative to the clean surface is opposite to that expected for the formation of As-As covalent bonds and can instead be associated with the replacement of Ga-As bonds by V-As in the interface region. Its invariance in energy suggests that the character of the V-As bond remains unchanged throughout the reacted region and a well defined interface product is formed. From Fig. 3 we see that the shifted doublet becomes a major component between 4Å and 6Å. At these coverages it amounts to ~25% of the original signal. Further, its relative importance grows with coverage and it is the only component present above 30Å.

A satisfactory description of the As 3d core EDCs of Fig. 3 can be based on two doublets at the fixed positions marked by arrows. For each component, the branching ratio was 1.55, the spin orbit splitting was 680 meV, and an experimental lineshape was used with a FWHM of 650 meV. This FWHM is somewhat larger than expected from the experimental resolution but can be rationalized as allowing for contributions from slightly inequivalent sites.

The fact that a good match to the lineshape of the As cores is obtained with just two components should not be taken to imply that one is reacted and one is the unshifted substrate component. The attenuation of the unshifted component is in fact much slower than one would expect for the covering of the unreacted portion of the substrate. Instead, the attenuation of the substrate As contribution can be easily obtained from the analysis of the unshifted Ga 3d peak, taking advantage of the larger separation of the reacted and unreacted components and noting that no substrate emission is visible for $\Theta \gtrsim 18$Å. For the Ga core in the GaAs bonding configuration, we obtain the exponential attenuation shown by the dashed line in Fig. 4. The dash-dot line represents the reacted phase. The same attenuation plot can then be used for the As 3d since the kinetic energy of the photoelectrons, and therefore the escape depth, was fixed to be the same by our choice of photon energy. As shown in Fig. 4, the contribution from the unreacted substrate is ~2% of the initial intensity at Θ=18Å, whereas the unshifted As doublet accounts for ~10% of the initial intensity. This, and analogous comparisons for all coverages, shows that a third component is present at the same position as As in GaAs. It does not shift with coverage, within the experimental resolution and the accuracy of the lineshape analysis, and is related to the disruption of the GaAs surface. The origin of this peak is not clear but, because of its binding energy, we associate it with an As-rich configuration where the V-As local coordination is lower than the "fully reacted" bulk V-As phase, very possibly a surface phase with weak V-As bonding. Since it is not shifted ~0.3 eV to greater binding energy relative to GaAs[13], we do not associate it with segregated covalently-bonded As. The relative contribution from this doublet is maximum at 10-14 Å (8-10% of the initial As 3d intensity) and decreases thereafter with respect to the shifted component (~3-4% of the initial signal vs. ~20% at Θ=30Å).

The progress of the reaction at the interface can be followed from the dash-dot curves of Fig. 4 which represent the total contribution from the reacted Ga and As components. From these curves we can conclude that breaking of the GaAs bonds and the growth of the reacted configuration proceeds to coverages of ~8-10Å. At higher coverages, the V-Ga compound is quickly covered and only slight outdiffusion of Ga occurs through the overlayer, as discussed above. Arsenic, on the other hand, is always present in substantial amounts within the probed region, mainly in the form of the fully reacted V-As configuration.

In this paper we have presented experimental evidence for the existence of chemical reaction and the formation of a broad intermixed layer at the V/GaAs(110) interface. Analysis of the core level shifts and intensity attenuation made possible the identification of reacted species and the microscopic modeling of the interface, but additional information relative to the concentration profiles is reqired. Work in this direction is in progress.

This work was supported by the Office of Naval Research under ONR-N00014-83-K-0579 and the Minnesota Microelectronic~~s~~ and Information Sciences Center. The Wisconsin Synchrotron Radiation Center is supported by the National Science Foundation and we gratefully acknowledge the expert assistance of its staff. Stimulating discussions with M. Ruckman, M. del Giudice and S.A. Chambers contributed to this research.

References

* Materials Science Program, University of Wisconsin, Madison, WI 53706

1. Exhaustive reviews of metal-semiconductor junctions studied with surface-sensitive techniques can be found in L.J. Brillson, Surf. Sci. Rep. 2, 123 (1982).
2. K.N. Tu and J.W. Mayer in Thin Films - Interdiffusion and Reactions, ed. byJ.M. Poate, K.N. Tu, and J.W. Mayer (Wiley - Interscience, NY 1978).
3. J.H. Weaver, M. Grioni, and J. Joyce, Phys. Rev. B (submitted 8/84).
4. A. Franciosi, D.J. Peterman, J.H. Weaver, and V.L. Moruzzi, Phys. Rev. B25, 4981 (1982).
5. G. Rossi, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, J. Vac. Sci. Technol. 21, 617 (1982).
6. J.R. Waldrop, S.P. Kowalczyk, and R.W. Grant, J. Vac. Sci. Technol. 21, 607 (1982).
7. M. Iwami, S. Hashimoto, and A. Hiraki, Solid State Commun. 49, 459 (1984).
8. J.G. Clabes, G.W. Rubloff, and T.W. Tan, PHys. Rev. B29, 1540 (1984).
9. P. Pianetta, I. Lindau, P.E. Gregory, C.M. Garner, and W.E. Spicer, Surface Sci. 72, 298 (1978).
10. L. Ley, O.B. Dabbuosi, S.P. Kowalczyk, F.R. Mc Feely, and D.A. Shirley, Phys. Rev. B16, 5372 (1977).
11. D.E. Eastman, T.-C. Chiang, P.Heimann, and F.J.Himpsel, Phys. Rev. Lett. 45, 656 (1980).
12. P. Skeath, I. Lindau, C.Y. Su, and W.E. Spicer, Phys. Rev. B 28, 7051 (1983).
13. J.F. van der Veen, L. Smit, P.K. Larsen, and J.H. Neave, Physica 117B 118B, 822 (1983).

Figure Captions

Fig. 1 Photoemission energy distribution curves (EDCs) for V overlayers on GaAs(110) at hν=21 eV after subtraction of the inelastic background and photon flux normalization. The cutoff of the V 'd' states is below E_F until ~ 3Å. Formation of hybrid states upon reaction is apparent from the growth of peak "B". Convergence to bulk V occurs only above 100 Å.

Fig. 2 Core level results for the Ga 3d core at hν=60 eV. The low-binding energy component corresponds to Ga liberated upon the disruption of the GaAs surface. This peak disperses continuously to lower binding energy and its final position is beyond that of metallic Ga, suggesting the formation of a V-Ga intermetallic. Binding energies are relative to E_V.

Fig. 3 Core level results for the As 3d core at hν=85 eV. The change in lineshape demonstrates strong chemical interaction. For each coverage the length of the arrows is proportional to the relative intensity of the two components used in the semi-quantitative fit (see text). The high-binding energy component is the superposition of contributions from the attenuating substrate and an As rich phase. Above ~30 Å only the low binding energy doublet, corresponding to As strongly bonded to V, is present. Binding energies are relative to E_V.

Fig. 4 Attenuation curves for Ga 3d and As 3d integrated intensities. Solid line = total intensity; dashed line=substrate contribution; dash-dot line = difference curve (total - substrate). The reaction proceeds with the disruption of Ga-As bonds to ~ 8 Å, corresponding to the maximum of the dash-dot lines for both As and Ga. Above ~ 8 Å the interface is progressively covered. Substantial outdiffusion of As in the V-bonded configuration occurs, while Ga atoms are more effectively trapped in the reacted region.

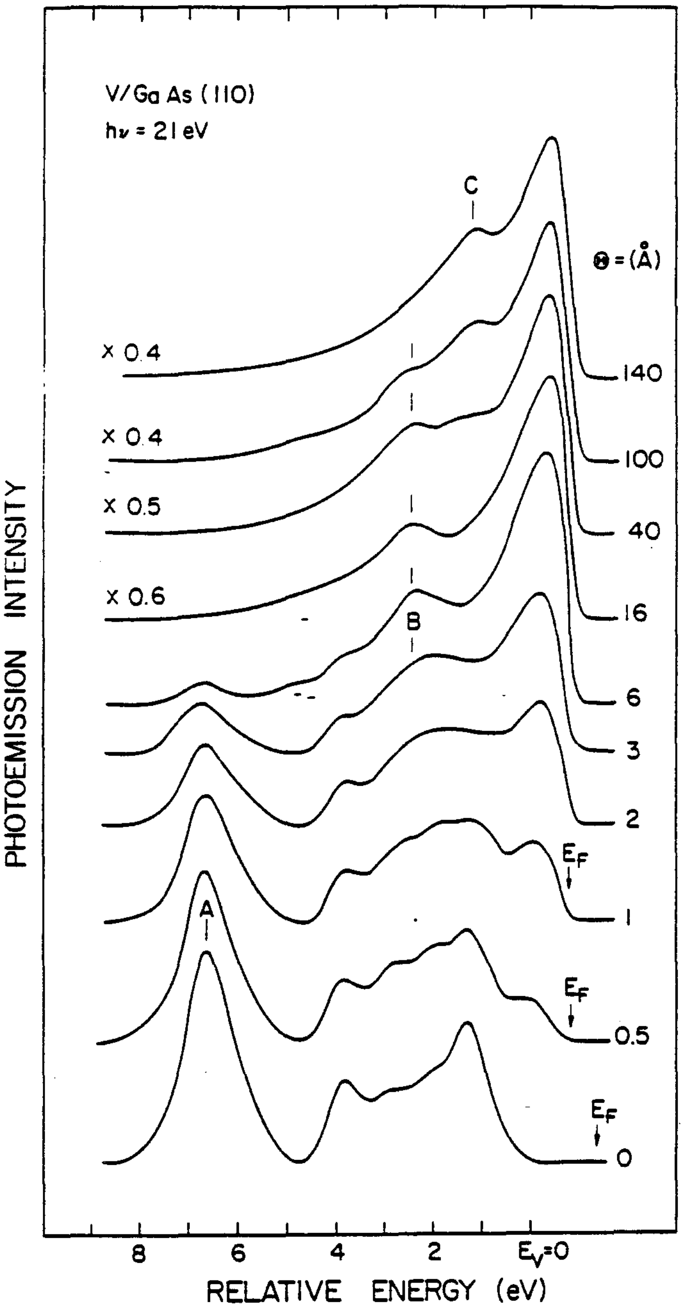

V/GaAs (110)
hν = 21 eV
C
A
B
Θ = (Å)
× 0.4
× 0.4
× 0.5
× 0.6
140
100
40
16
6
3
2
1
0.5
0
E_F
E_F
E_F
PHOTOEMISSION INTENSITY
8
6
4
2
E_V=0
RELATIVE ENERGY (eV)

Fig. 1

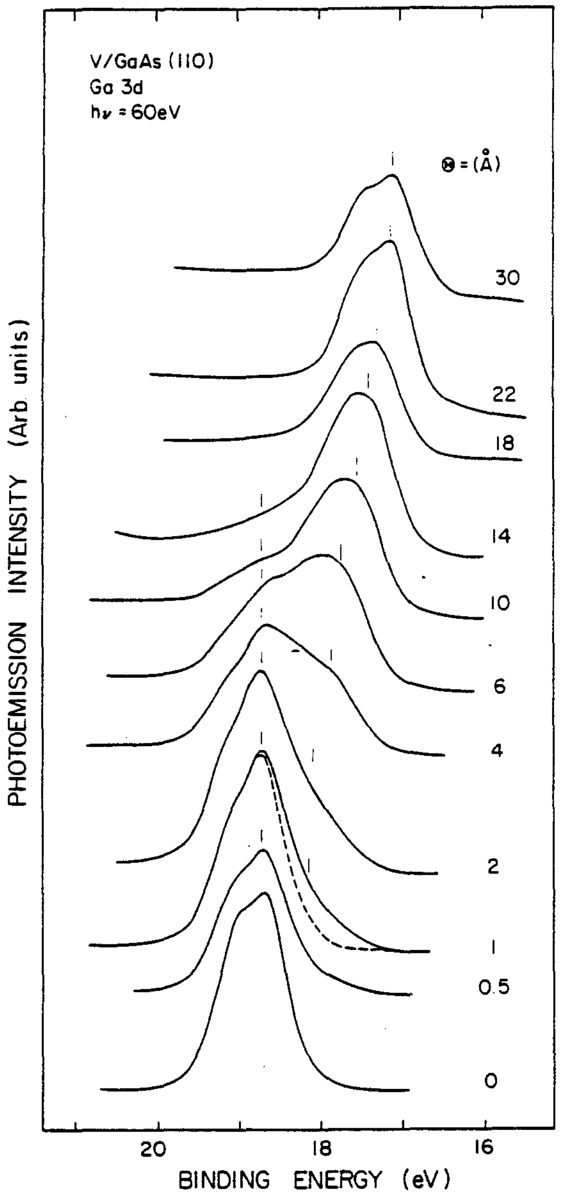

V/GaAs (110)
Ga 3d
hν = 60eV
Θ = (Å)
30
22
18
14
10
6
4
2
1
0.5
0
PHOTOEMISSION INTENSITY (Arb. units)
20
18
16
BINDING ENERGY (eV)

Fig. 2

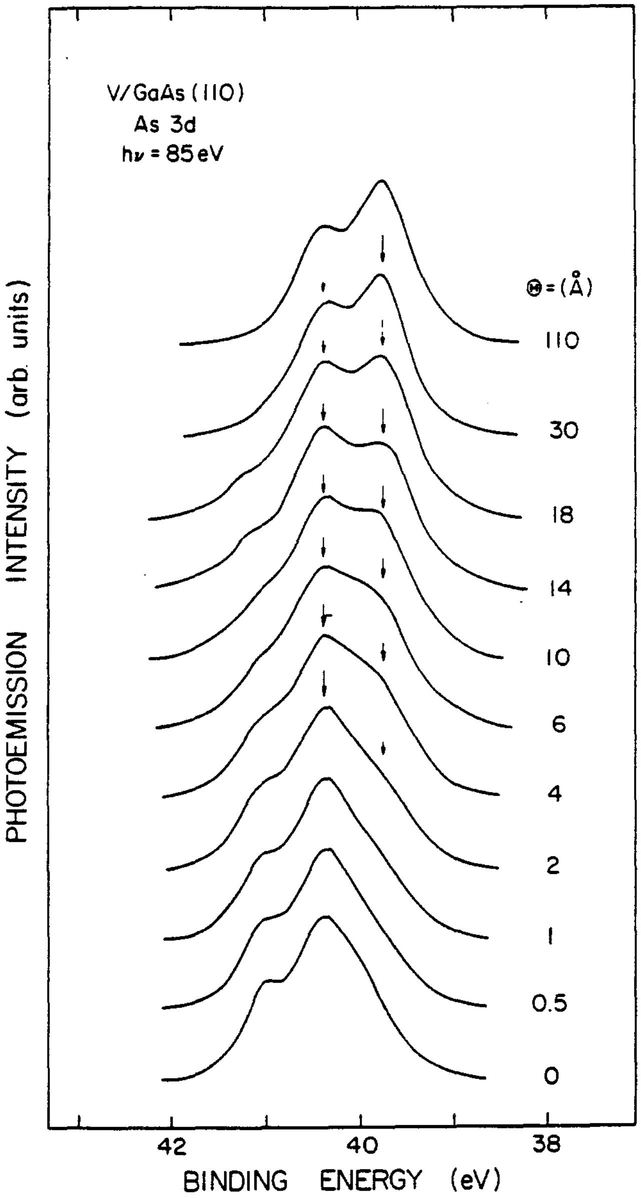

V/GaAs (110)
As 3d
hν = 85 eV
Θ = (Å)
110
30
18
14
10
6
4
2
1
0.5
0
PHOTOEMISSION INTENSITY (arb. units)
42
40
38
BINDING ENERGY (eV)

Fig. 3

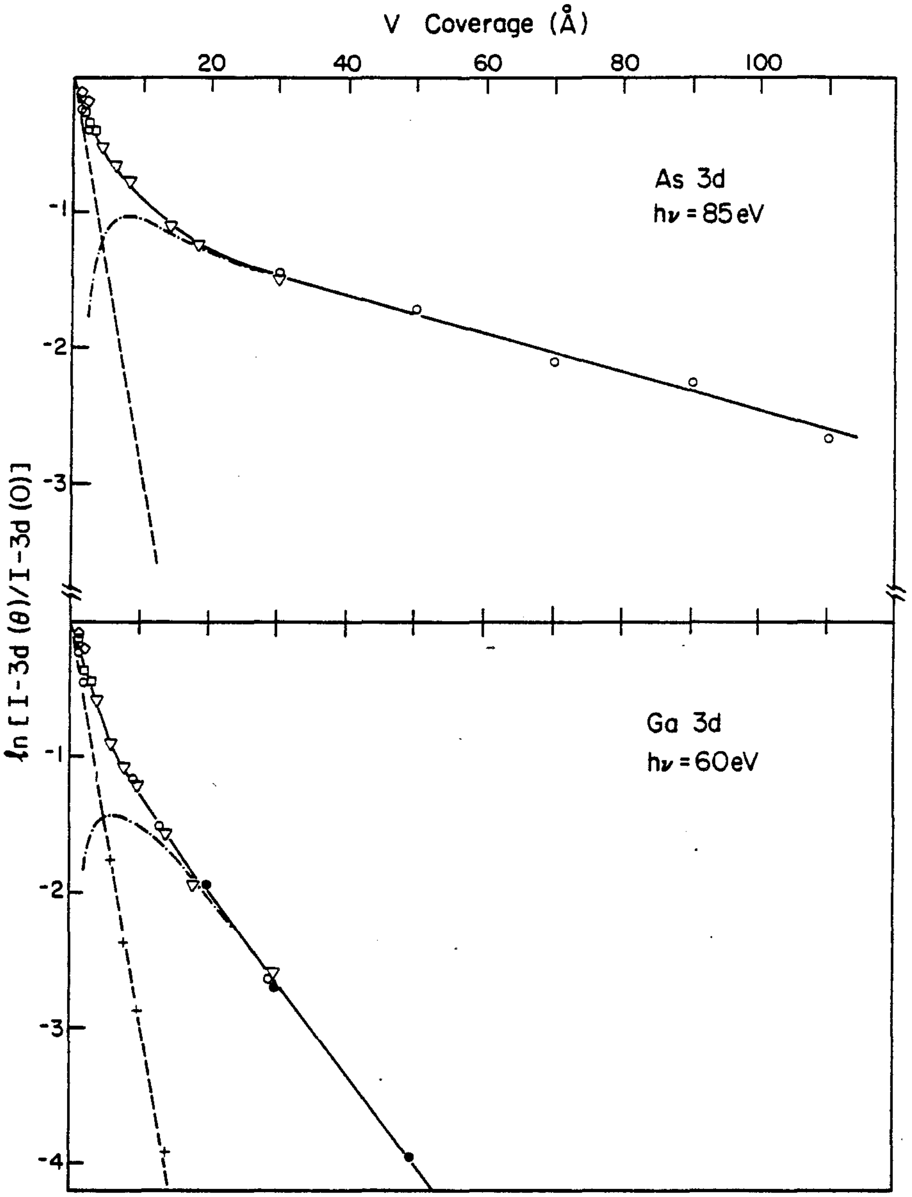

V Coverage (Å)
20
40
60
80
100
As 3d
hν = 85 eV
-1
-2
-3
Ga 3d
hν = 60 eV
-1
-2
-3
-4
ℓn [I-3d (θ)/I-3d (0)]

Fig. 4

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

HETEROJUNCTION DISCONTINUITIES: THE CURRENT POSITION

Microelectronic and Information Sciences Center

Technical Report #22

A. Nussbaum
Department of Electrical Engineering
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

HETEROJUNCTION DISCONTINUITIES:
THE CURRENT POSITION

MEIS Technical Report #22

A. Nussbaum

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

[1984?]

In accordance with the doctrine of equal time, and as a consequence of comments received during an evening presentation at the recent Device Research Conference (Santa Barbara, 18-20 June 1984) as well as those in correspondence from H. Kroemer(1), it would seem desirable to complete a series(2-4) of letters concerning the nature of the band structure of heterojunctions by presenting a simple, analytic proof of the position taken in previous publications(3,5,6). The basis of this proof is the form of the Poisson equation

$$d^2V/dx^2 = -(e/\varepsilon)\ (p - n + N_D - N_A) \qquad (1)$$

applicable to semiconductors, where the symbols have their well-known meanings. Using Boltzmann statistics for non-degenerate silicon or germanium, the electron and hole concentrations are

$$n = n_i\ e^{(E_F - E_I)/kT},\ p = n_i\ e^{-(E_F - E_I)/kT} \qquad (2)$$

where E_F is the Fermi level (as positioned by the impurity concentration) and E_I is the intrinsic level. Substituting (2) into (1) gives the familiar Poisson-Boltzmann equation.

We now consider a heterojunction, one side of which could be intrinsic germanium and the other intrinsic gallium arsenide. The band structure of Fig. 1 is based on the assumption that the vacuum level E_{vac} is everywhere continuous, as proposed by Anderson(7). This implies that the bands will bend in the space-charge region, where E_F lies above or below E_I. (The energies shown in this diagram may not be precise, but the structure is qualitatively correct.) Following Shockley[8], we identify E_I as the reference for electrostatic potential V, so that

$$E_I = -eV \tag{3}$$

It is convenient to introduce the dimensionless variable

$$u = (E_F - E_I)/kT \tag{4}$$

Then (3) and (4) combine to give

$$d^2V/dx^2 = -(1/e)d^2E_I/dx^2 = (kT/e)d^2u/dx^2 \tag{5}$$

and the Poisson-Boltzmann equation in the space-charge region of Fig. 1 takes on the very simple form (for $N_D = N_A = 0$)

$$d^2u/dr^2 = \sinh u \tag{6}$$

where $r = x/L_D$

and $L_D = (\varepsilon kT/2n_i e^2)^{\frac{1}{2}}$ is the intrinsic Debye length. The first integral of (6) is

$$du/dr = \pm [2(\cosh u + C)]^{\frac{1}{2}} \tag{7}$$

and the integration constant C has the value C = -1 since du/dr = 0 for u = 0 in the two neutral regions. Thus

$$du/dr = \pm 2 \sinh (u/2) \tag{8}$$

To deal with the sign ambiguity in (8), we recognize that the slopes in Fig. 1 of the energy levels in the space-charge region, as measured by du/dr (or dE_I/dx) are positive on either side of the physical junction. Since the vacuum level E_{vac} determines the shape of the bands, and in particular, the behavior of the intrinsic level E_I, we also see that the quantity u, as defined by (4), must be positive in the GaAs and negative in the Ge. It would then be necessary, as suggested by Parrott and others[9], to resolve this ambiguity by choosing a positive sign to the right of the junction and a negative one to the left; the fact that (8) defines an odd function then makes the analysis consistent with Fig. 1. But this is incorrect, as may be simply shown on the basis of a hypothetical experiment: add a small amount of donor atoms to the GaAs and an equally small amount of acceptor atoms to the Ge. The intrinsic levels will shift away from the constant Fermi level and the vacuum level will have the barrier of 0.27eV lowered somewhat; otherwise, there will be no drastic change. What is significant is that the respective donor and acceptor concentrations N_D and N_A enter (7) to give

$$du/dr = \pm [2(\cosh u - au + C)]^{\frac{1}{2}} \qquad (9)$$

where $a = (N_D - N_A)/2n_i$ is the relative net concentration in each region. The crucial point about (9) is that it can not be converted into the odd function (8); the quantity in the radical must always be positive to avoid an imaginary solution. Therefore, the positive sign must be chosen on both sides of the junction to match the energy band diagram predicted by the Anderson model. It is physically inconceivable that the addition of a single impurity atom to intrinsic Ge would reverse the sign associated with the radicals in going from (7) to (9). Therefore, the only way to resolve this difficulty is to require that u = 0 in (8), so the E_I is everywhere congruent with E_F; i.e. the intrinsic level in a heterojunction is continuous.

An objection(10) that has been raised to the above argument is that the identification of the intrinsic level as a measure of the electrostatic energy, as accomplished by eq. (3), represents the reason why we have been able to demonstrate the continuity of E_I and any level so chosen would have this property. We now show why this is not the case. The conduction band edge E_C, as an example(11) , is often used for such a purpose. Equations (2) are then replaced by

$$n = N_C e^{(E_F - E_C)/kT}, \; p = N_V e^{(E_C - E_G - E_F)/kT} \qquad (2a)$$

and (3) by

$$E_C = -eV \qquad (3a)$$

Because of the asymmetric form of the Boltzmann statistics in (2a), we do not obtain an odd function as in (8), and no corresponding conclusion about the continuity of E_C is possible. It is well-known that electrostatic potential can be defined only to within an arbitrary constant which drops out of the left-hand side of (1); it should be possible to derive an equation analogous to (6) for this new reference arrangement, and then use it to prove our original conclusion concerning E_I. However, the algebra will be rather complicated and the analysis given above is by far the simplest.

Another point that needs addressing is the statement(4) that the identification of the intrinsic level as a measure of the electrostatic potential on both sides of the heterojunction is incorrect; once it has been made for a given component, the potential on the other side can be obtained only by integrating the Poisson-Boltzmann equation twice across the interface. The reply to this is based on the well-known equivalence of the Fermi level and the equilibrium electrochemical potential(6). Since the latter is simply the algebraic sum of the chemical potential (which is $E_F - E_I$ at the any point in the structure) and the electrostatic potential, it follows that a constant value of E_F combined with a knowledge of the doping as a function of position defines -eV everywhere. Kroemer is correct in saying that two integrations are necessary to find the chemical potential as a function of position, but for a doubly-intrinsic structure (and only such a device), this can be done analytically(6) .

The final point of Kroemer's that needs to be discussed is the assertion that the direct experimental measurements of valence band discontinuities which appear to confirm the continuity of the intrinsic level are in fact of not much value, because almost all of these have been made in junctions combining column IV non-polar elements with polar III-V or ternary compounds. These combinations are believed to be of sufficient complexity to make measurements of small differences in large quantities have doubtful validity. One example that overcomes this difficulty is the lattice-matched system GaAs/$Al_xGa_{1-x}As$. The well-known Dingle rule(14) for x = 0.2 predicts that 85% of the total discontinuity at the junction lies in the conduction band and 15% in the valence band. Not enough is known about electron affinity values to say that this result confirms the Anderson rule, but it at least is consistent with the general pattern of highly unequal discontinuties. The continuous intrinsic level model,(5) on the other hand, predicts virtual equality of the discontinuities in the two bands, so that the Dingle results definitely do not support it. However, recent measurements by Miller et al(15) replace the 85/15 rule with one that divides the discontinuities equally between the two bands. For this system at least, which has been extensively studied and is of great practical value, it would appear that the arguments given above have validity.

This work was done under the auspices of the Microelectronics and Information Sciences Center of the University of Minnesota, whose support is gratefully acknowledged.

References

(1) H. Kroemer (personal communication, 3 August 1983).

(2) H. Kroemer, IEEE Electron Device Lett., vol. EDL-4, p. 25, 1983.

(3) A. Nussbaum, IEEE Electron Device Lett., vol. EDL-4, p. 267, 1983.

(4) H. Kroemer, IEEE Electron Device Lett., vol. EDL-4, p. 365, 1983.

(5) M. J. Adams and A. Nussbaum, Solid-State Electron. vol. 22, p. 783, 1979.

(6) A. Nussbaum, "Theory of Semiconducting Junctions," in Semiconductors and Semimetals 15, R. A. Willardson & A. C. Beer, Eds. New York: Academic Press, 1981.

(7) R. L. Anderson, Solid-State Electron. vol. 5, p. 341, 1962.

(8) W. Shockley, Electrons and Holes in Semiconductors, New York: D. van Nostrand, 1950.

(9) J. E. Parrott (personal communication to M. J. Adams, 24 September 1979); H. Beneking (personal communication, 12 July 1983), K. M. vanVliet (personal communication, 29 December 1978).

(10) M. S. Lundstrom (personal communication, 19 June 1984).

(11) A. Nussbaum, Semiconductor Device Physics, New York: PRentice-Hall, 1962.

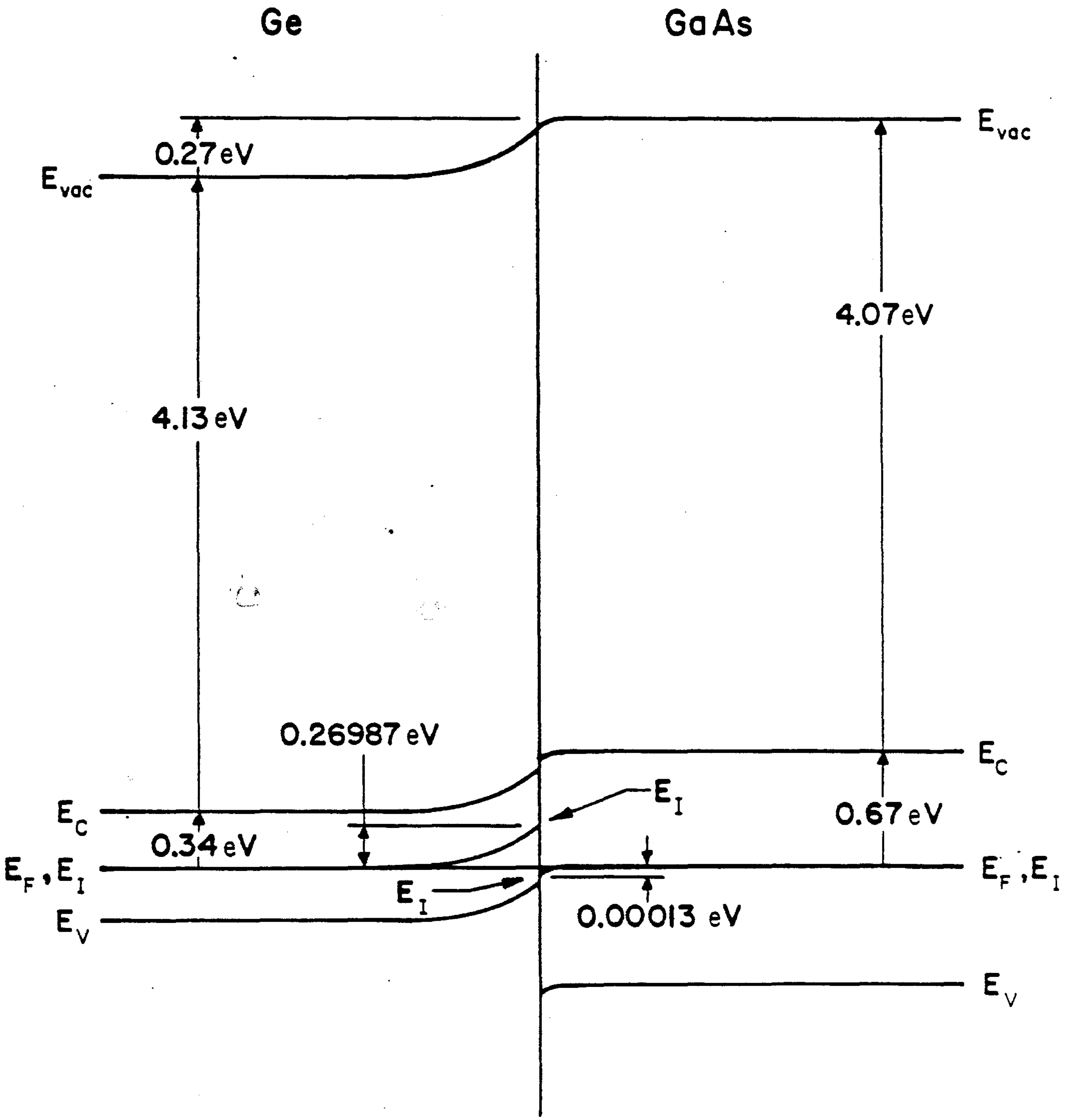
(12) A. Nussbaum, Solid-State Electron. 25, p. 1201, 1982.

(13) H. Kroemer (personal communication, 6 January 1982).

(14) R. Dingle, "Confined Carrier Quantum Wells in Ultrathin Semiconductor Heterostructures, "Festkörperprobleme/Advances in Solid State Physics, H. J. Queisser ed. (Vieweg, Braunschweig, vol. 15, pp. 21-48, 1975.

(15) R. C. Miller, A. C. Gossard, D. A. Kleinman, and O. Munteanu, Phys. Rev. B29, p. 3740, 1984.

Figure 1



Ge
GaAs
E_{vac}
0.27 eV
4.13 eV
E_C
0.34 eV
E_F, E_I
E_V
0.26987 eV
E_I
E_I
0.00013 eV
E_{vac}
4.07 eV
E_C
0.67 eV
E_F, E_I
E_V

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

THE ROLE OF KNOWLEDGE IN THE ARCHITECTURE OF A ROBUST ROBOT CONTROL

Microelectronic and Information Sciences Center

Technical Report #23

M. Gini
Department of Computer Science
University of Minnesota

R. Doshi
Department of Computer Science
University of Minnesota

M. Gluch
Department of Computer Science
University of Minnesota

R. Smith
Department of Computer Science
University of Minnesota

I. Zualkernan
Department of Computer Science
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

THE ROLE OF KNOWLEDGE IN THE
ARCHITECTURE OF A ROBUST ROBOT CONTROL

MEIS Technical Report #23

M. Gini
R. Doshi
M. Gluch
R. Smith
I. Zualkernan

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center and the Graduate School.

[1984?]

ABSTRACT

We would like robots to recognize and handle situations that do not conform with normal operating conditions. We want to be able to do this without having to consider explicitly errors caused by missing or defective parts, or by malfunctioning. To this end we present the detailed design of a system in which the controller of the robot takes advantage of a large knowledge bases to ensure proper execution of the robot task. Real time considerations played a large role in our design.

1. INTRODUCTION

Robots are today operating on a wide variety of tasks such as object handling, painting, and welding. Even though assembly is still considered a difficult area, more and more robots are used in manufacturing to perform assembly tasks. New areas outside manufacturing, like exploration of unknown environments and medical applications, are being considered. Future growth areas are predicted to involve highly complex tasks.

Robot programming, which was very easy when tasks were simple, becomes one of the central issues. Programming a computer controlled robot is really different than programming a computer. Robot programs run in a world which is incompletely known and imperfectly modeled. This requires strategies to detect and prevent potential catastrophes like collisions and to recover from errors. Many actions are irreversible. After the arm has crashed there is no way to undo the action that produced the crash. Actions are not exactly reproducible, making it difficult to detect causes of errors and to fix them.

Currently, programmers of robot tasks must depend on their experience, intuition, and common sense to decide what errors to watch for. Errors are difficult to identify because of their unpredictability. The same program can work well hundred of times and then stop because of a minimal variation in size of one part. The causes of failures are often associated with parts that are not in the right position and orientation. Sometimes the problem is a malfunction in the hardware of the robot.

Sensors can be used to detect information about real world. Proximity sensors or other control strategies can be applied to avoid catastrophes. Since it is not that easy to restart the program after an emergency stop most of the times human intervention is required. Unless the errors are explicitly considered by the programmer there is no way with current programming systems to do error recovery.

Since computer controlled manipulators have been introduced the methodology of controlling and of programming them for new tasks has seen a great deal of development (Albus, 1981), (Binford, 1979), (Nitzan, 1976), (Paul, 1981).

Two completely different approaches to robot programming have been considered in the past. On one side within the Artificial Intelligence community a lot of research has been done on plan formation systems to provide robots with autonomous reasoning capabilities (Sacerdoti, 1977). None of these systems have been used to control a real robot, with the exception of STRIPS at SRI (Fikes, 1971).

On the other side the need to control industrial robots has pushed the development of simple but effective methods for robot programming (Luh, 1983). Complex systems have been designed over the years to cope with increasing demands. None of them requires yet the reasoning capabilities provided by Artificial Intelligence. Many robot languages have been designed; among those are AML (Taylor, 1983), AL (Binford, 1979), PAL (Takase, 1981), VAL (Shimano, 1979), and WAVE (Paul, 1977). A good classification and comparison of many of them can be found in (Bonner, 1982) and (Lozano-Perez, 1983).

The gap between these two approaches is becoming smaller as more powerful languages are designed and Artificial Intelligence techniques are applied to the solution of specific problems in robotics (Brooks, 1982), (Brooks, 1983), (Lozano-Perez, 1981), (Lozano-Perez, 1984), (Poplestone, 1980).

Most of the research in Artificial Intelligence has so far concentrated on the problem of developing the task algorithm starting from very-high level specifications of the task. When programs will be generated from a task description language some strategies for error recovery will be incorporated into them. Even though very interesting results are expected in the next years it will take a long time before seeing these languages in day to day operation.

Presently industrial robots do not have any world model (Gini, 1983a); their knowledge is encrypted into variables and data structures which have a meaning only for their human programmer. Everything the robot has to do is precisely defined in the program.

From an industrial point of view the idea of programming robots off-line is becoming more and more appealing. Robots could do useful work while new programs are being developed. This would reduce at the minimum the period of time in which the production has to be interrupted. When robots are components of complex industrial automation systems this aspect is particularly important (Ambler, 1982).

Unfortunately it is well known that it is impossible to completely test the program off-line. Problems come from the lack of sensor data since there is no simple way to simulate sensors (Smith, 1983). Even though the use of sensors is still primitive in industrial robots, the simulation of even these simple sensory environments has not been included in any simulation system. The world model in the computer is not the same as the real world. It becomes important to be able to identify unexpected situations or discrepancies with the model and to take appropriate actions without the need for human intervention.

The problem of dealing with errors has been approached in various ways and with different objectives in Artificial Intelligence research. Most of the work has been done in solving errors during the planning (Sacerdoti, 1977), (Sussman, 1975). Srinivas (Srinivas, 1976), (Friedman, 1977) has designed a system for analyzing the causes and kinds of failures in robot programs and for replanning. A major limitation derives from the extensive use of plan formation as the basis for constructing robot programs.

The problem of automatic error recovery has not yet been fully addressed. By automatic error recovery we mean the fact that errors are identified every time they arise and a recovery procedure is carried on without the need for the user to consider them.

The main reason why automatic error recovery is difficult is because it requires precise knowledge about the environment, operations, resource usage, and the intent of the user program. Since the environment is changing in time a dynamic model of the environment should be used. The ability of interpreting information gathered by sensors, and of deciding appropriate recovery actions are also needed.

Automatic error recovery may play an important role in industrial robotics. According to a recent Japanese Delphi forecast programming languages based on world models will have an industrial impact before 1990. The results of the Japanese forecast are in agreement with a study of manufacturing requirements for a high-level off-line programming language performed as part of the ICAM project. Long term developments (three to six years) should include automatic collision avoidance, recovery after unexpected events, force sensing, and task-oriented language.

The aim of this paper is to present the design of a system in which the controller of the robot is able to identify errors and to recover from them whenever they happen (Gini, 1983b). Our methodology relies on an extensive use of knowledge bases. It requires monitoring, interpreting, diagnosing, and planning, as shown in the next section.

Before going into details we will present some of the assumptions that we made.

We assume that the task is described by a working program in a high-level robot programming language. We have selected the AL language (Binford, 1979), (Finkel, 1975), (Mujtaba, 1979), but other languages could work as well. Everything needed to control the robot is defined in the program. We also assume that the program does not have logic errors. For example, the program might incorrectly command the robot to try to move through objects; we can try to recover after such an error but we can't expect to be able to perform the robot's intended task. This is not too strong of a requirement if the program can be developed and tested offline where we may easily verify the logical correctness of the robot's sequence of actions.

2. OVERVIEW

Our research has been examining the problem of error recovery in robots. We are trying to develop a system which enables the robot to detect and recover from errors caused by unexpected conditions in its environment. The system's primary goal is to make the robot more robust so that there is less need for operator intervention. Another important objective is to reduce the robot's programming time by shifting the burden of error detection from the programmer onto the robot system itself. We intend our system to be compatible with existing systems of automated manufacturing.

Figure 1 presents a diagram of the robust robot control. The system consists of two parts: the *offline phase* and the *online* phase. The task performed by the robot is programmed in the robot programming language AL. The offline phase consists of a *Preprocessor* that uses the AL program along with the Global Knowledge Base (labeled GKB in the diagram) to obtain the semantic structure of the program. This information is placed in the Local Knowledge Base (labeled LKB in the diagram). The Preprocessor also modifies the robot task written in AL to produce the Augmented Program (labeled AP in the diagram). The Augmented Program is then executed by the Monitor which in turn interacts with the robot. The Monitor uses sensor information from the robot to detect errors and other unexpected conditions. During execution of the robot task, the Monitor maintains information about the robot's operation in the Dynamic Knowledge Base (labeled DKB in the diagram). A catalog of objects in the robot's work cell is a key part of the Dynamic Knowledge Base. If an error is detected, the Monitor passes the control to the *Recoverer* which interprets the error by using the Dynamic Knowledge Base, sensor information, and

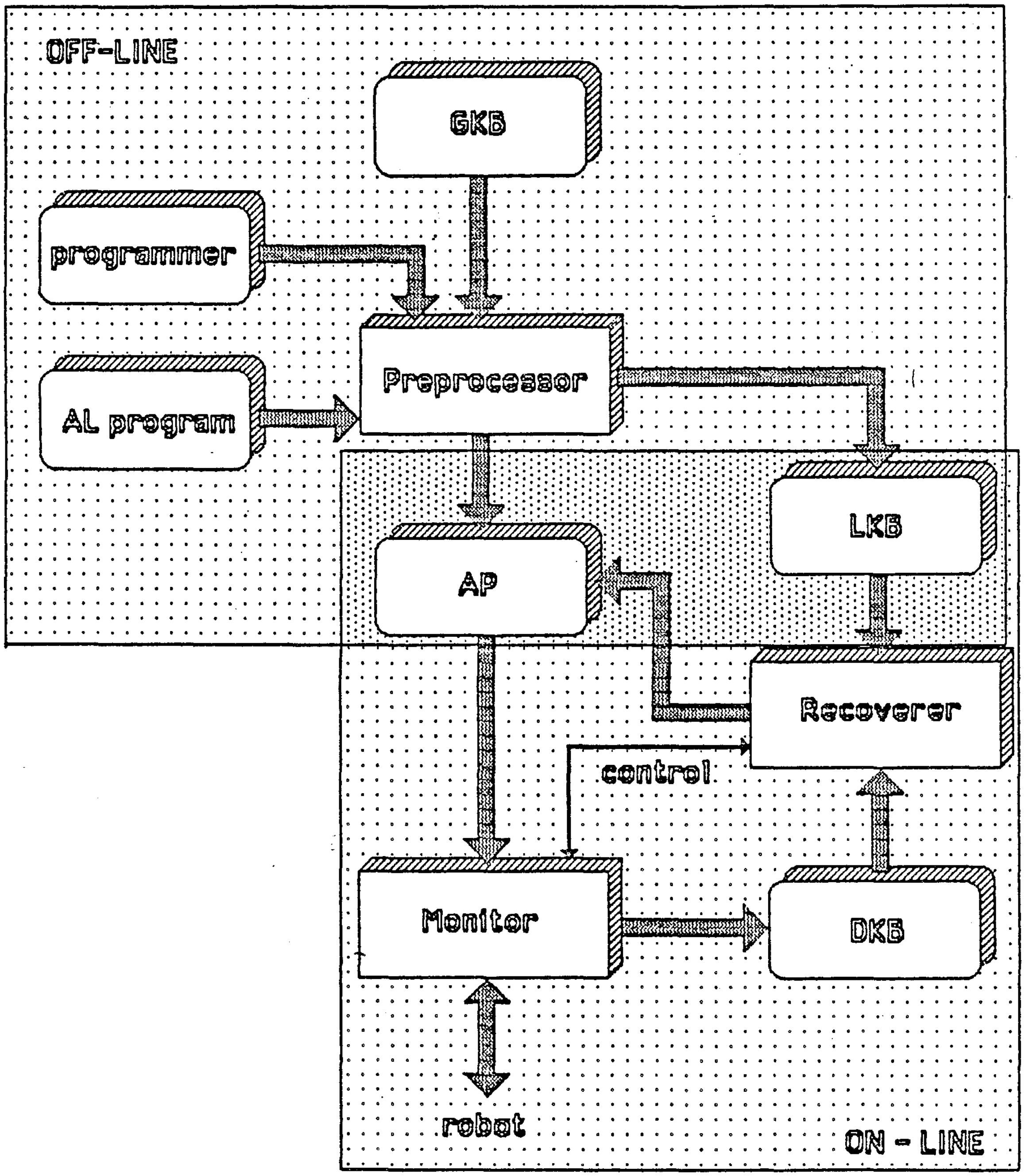


Figure 1: Diagram of the Robust Robot

the Local Knowledge Base. The Recoverer also uses this information to devise a strategy for recovering from the error. In the process of error interpretation the Recoverer can also run some tests through the Monitor. If the Recoverer finds a plausible recovery strategy then modifications to the Augmented Program are passed to the Monitor. The Monitor resumes normal execution of the Augmented Program at an appropriate point after executing the recovery steps. If a recovery strategy is not found then the system informs the robot's operator of the error along with information about the cause of the error.

2.1. Major Issues

The interpretation of the AL program's semantic structure plays a very important role. During the course of our research we found this interpretation to be one of the harder problems. It seems that aside from the obvious semantics of the program, a lot more additional information is needed for reasonable performance of the system. Checking the robot's expected sensor conditions entails a special strategy since it is not feasible to check all the possible conditions after each instruction. To obviate this problem we have developed a *filtration* technique to limit the amount of error checking performed on each instruction. This strategy, combined with demands of the recovery process, requires us to keep a trace of the robot's runtime actions and environment. This is also bound by space constraints.

The offline part entails the development of a world model for the robot, its task, and its environment. This model is developed by using the knowledge from the Global Knowledge Base and the given AL program. An important piece of information derived from the program is the identification of the objects, their positions, and how they are manipulated by the robot. It is not possible to establish the intent of the whole program just from its syntax, but syntactic information does help us develop a partial description of the task's general semantics.

We derive another higher level of abstraction in our partial description by using pattern matching on the AL program. For example, the three AL instructions

```
MOVE xarm TO y
CENTER xarm
MOVE xarm TO z
```

can be transformed into a macro instruction of the form

move (OB, y, z, x)

which means "*move* object *OB* from *y* to *z* using arm *x*."

The second level of description is also useful in achieving a relatively macro view of the intent of the task as compared to the instruction level view.

The Global Knowledge Base includes the most general forms of post- and preconditions associated with each instruction. We call these general forms *generic instructions.* Depending on each instruction's context and environment in the AL program the Preprocessor attaches a more specific form of conditions to instructions. This process eliminates unnecessary checking of error conditions. This also introduces some uncertainty in the system; we might miss detecting an error by not checking all possible post- and preconditions. The same technique is repeated for the next level of hierarchy. Instead of checking all conditions on all instructions this process can at times enable the system to do the checking on a higher level, thus saving time.

The idea of *filtration* is also necessitated by the real time environment. The system may miss some errors since it only checks for the most probable ones at runtime. To avoid missing an unprobable condition which caused an error, the Local Knowledge Base keeps a list of these relatively improbable errors. This list is used by the Recoverer to determine the cause when an error does occur.

The interpretation of the sensory information is another important issue. The raw data is interpreted as qualitative information by relating information in the Local Knowledge Base to the sensor information provided in the Dynamic Knowledge Base. The qualitative interpretation is derived from the environment and the intent of each part of the program. This qualitative information is used by the Recoverer's error interpreter to fire appropriate rules in order to interpret the error.

Another issue is the computation of the expected values of sensors for post- and preconditions in the Augmented Program. Some of these values are specified in the program, some can be derived from program semantics, but there are some cases where expected values can not be easily determined offline. In some cases the values of specific sensors (values of force sensors in a critical assembly, for example) can be derived by the programmer by manually manipulating the robot. In many cases the system must compute the expected sensor values dynamically. These computations must be kept as simple as possible to meet the robot's time constraints.

2.2. The Knowledge Bases

The Global Knowledge Base contains information about the world in general, the specific robot, and the robot's environment in general. There is also knowledge which enables the Preprocessor to extract specific qualitative physical laws relevant to the AL program being processed; the relevant laws are then included in the corresponding Local Knowledge Base. The Global Knowledge Base also contains *generic forms* for each instruction. There is also corresponding knowledge about how to generate specific pre- and postconditions from the generic form of each instruction. Further, the Global Knowledge Base contains information about possible errors and a measure of probability of their occurrence.

The Augmented Program contains the actions generated by the original AL instructions along with the post- and preconditions for each instruction or set of instructions. An example of such a program is given later in Figure 5. The Augmented Program is structured as finite state automaton. Events in the automaton are the result of evaluating the post- and preconditions; each event triggers some action and a transition to a new state.

The Local Knowledge Base contains information about the robot environment restricted to the particular robot configuration and the specific AL program. The Local Knowledge Base also contains the unlikely error conditions which are not checked in the Augmented Program, but might be necessary for the recovery part. There is also qualitative physical knowledge appropriate to the robot's task.

The Dynamic Knowledge Base contains dynamic information about the objects being manipulated by the program. This consists of a catalog of information about objects in the robot's workspace. There is also a trace of the robot's actions and recent sensor readings. This information is essential to the recovery process.

3. THE OFFLINE PHASE

In this section we will describe the functions of the offline phase, how to derive the intent from the syntax of an AL program and an example of an Augmented Program. The Preprocessor performs four major functions:

Function 1: Extract the semantics from the AL program.

The Monitor has to know what to check after executing an instruction. This is the only way through which the Monitor can detect an error or unexpected situation. During error interpretation the semantics and the operation's intent help in constraining the search for the possible causes of an error. During recovery the Recoverer has to know the intent of an action (or groups of actions) at various levels of abstraction. For example, an AL instruction could be:

```
MOVE xarm TO y
```

At a local level of abstraction, the meaning might be "the arm will move from point-A to point-B." At some intermediate level of abstraction, the meaning might be "move to point-B, so as to pick up a fragile object." At the topmost level of abstraction, the instruction may or may not have a lot of significance.

Other examples of intent could include the following:

- we are opening the hand to grasp an object.
- we are opening the hand to drop an object.
- we are pushing an object.
- we are pulling an object.

As another example, suppose that the finger sensors were activated during arm movement. It is important to know whether or not the hand is supposed to be empty in order to interpret the sensor reading. If so, then a collision probably occurred; otherwise an object may have slipped out of the hand, either partially or completely. This process is only performed during the offline phase. It would be inefficient to try to derive this information in real time.

Function 2: Give specifications to the Monitor for pre- and postcondition checking

The Monitor has to know the acceptable range of sensor values. It is also important to know which conditions were *not* checked; this helps the Recoverer reason about the possible causes of the error. A list of these unverified conditions is put into the Local Knowledge Base.

We may choose to ignore certain sensor readings. In some cases it may be very expensive or time consuming since the system could be checking for many things. Also, certain sensors may not be necessary for the immediate task. For example, there is no need to check finger sensors when an empty hand is not moving. The Augmented Program tells the Monitor what sensors to check for before executing each instruction.

Function 3: Compute the expected sensor values and tolerances

To detect errors, the Monitor must know what constitutes an abnormal or unexpected situation. The expected sensor values can be derived from information in the Global Knowledge Base or in the AL program. Additional information such as tolerances may also have to be obtained The acceptable tolerances may have to be obtained from from the programmer, extracted from a CAD database describing the assembly parts, by numerical computation on assembly models, or during the manual teaching of the robot. The expectation of sensor values and tolerances are inserted into the Augmented Program.

Function 4: Suggest what errors are likely to occur and when they may occur

The Preprocessor uses its knowledge of the intent and importance of particular subtasks to determine which errors are likely to occur and where. Appropriate directives are inserted into the Augmented Program at points of likely errors to insure more thorough checking. Heuristic knowledge is also helpful during error interpretation since it helps focus attention on more likely errors first. The list of errors and the locations where they may occur are placed in the Local Knowledge Base.

Since the programmer has the best understanding of the constraints and characteristics of the task, the Preprocessor can interact with the programmer to focus attention on those subtasks that are most important or need more care. This information can be encoded into special directives in the AL program or acquired interactively by the Preprocessor.

3.1. Extracting the Task's Intent

The robust robot relies on general knowledge contained in the Global Knowledge Base about robots, robot programming, and the errors that robots encounter. By applying this knowledge the Preprocessor analyzes the AL program to determine what actions the robot must perform and what objects it must manipulate. Some of the knowledge about AL programs is in the form of program pattern that imply the program's intent. This knowledge allows the Preprocessor to apply *pattern matching* in many cases to extract the program's semantics. For example, if the instructions sequence consists of:

```
OPEN xhand
.
.
MOVE xarm TO y
.
.
CENTER xarm
```

then the intention of this sequence is to *grasp* an object. The initial location of this object being grasped is the destination of the last MOVE instruction in the sequence, right before the CENTER instruction that actually grasps the object. If the sequence is

```
.
.
grasp
.
.
MOVE xarm TO z
```

then the intention of the MOVE instruction is to *carry* the object. There are several different subcategories of *carry*, like *insert* or *screw in*, depending on the context of the MOVE instruction. If the sequence is

```
.
.
carry
.
.
OPEN xhand
```

then the purpose of the OPEN instruction is to *dispose* of the carried object at the location specified in the previous MOVE instruction in the *carry* sequence. If the instruction sequence is

```
.
.
dispose
.
.
MOVE xarm TO y
.
.
CENTER xarm
```

then the purpose of the CENTER instruction is to grasp another object.

For a larger example, let us consider the flowchart in Figure 2 and the corresponding AL program in Figure 3†. The program repeatedly performs a sequence of *grasp - carry - carry - dispose* operations. Since each execution of the loop ends with *dispose*, we can conclude that each execution of the loop deals with a different object. Furthermore, all these objects are grasped at the same location, which might indicate the existence of a part feeder. This fact may be important for error analysis since a feeder failure could prevent the robot from *grasp*ing a part before the loop count is exhausted.

There are two destinations of final *carry* operations, so the program evidently sorts the objects into two categories. One of these destinations is fixed while the other is parametrized by variables controlled in the loop. This suggests that two different kinds of assembly operations are

† This program is taken from the "AL User's Manual" (Mujtaba, 1979).

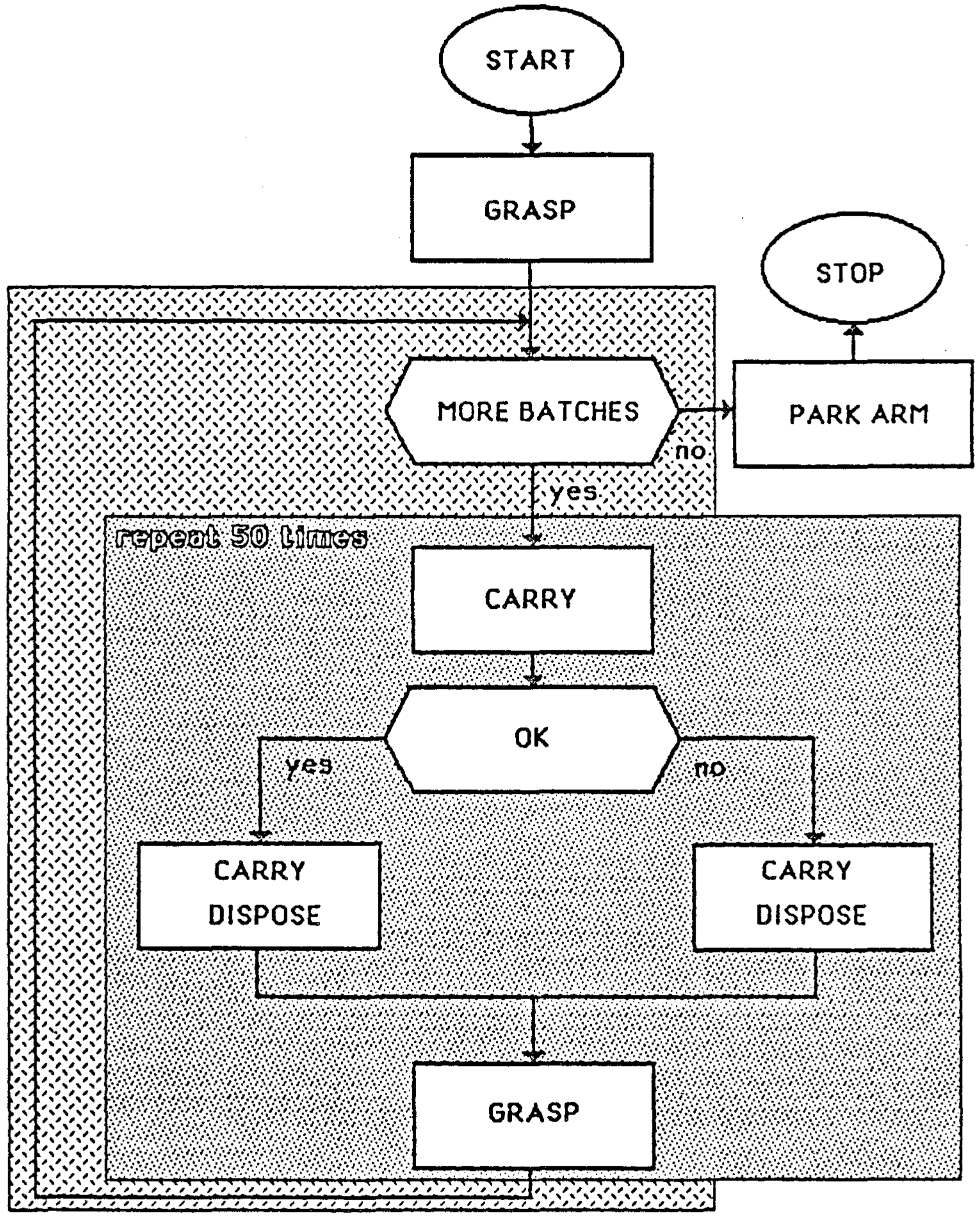


Figure 2: Flow Chart of Example Robot Task

```
BEGIN
{ declarations and initializations }
OPEN bhand TO 3*inches;
MOVE barm TO pickup DIRECTLY;                                grasp
CENTER barm;
IF(bhand<1.5*inches) THEN more_batches ← FALSE
  ELSE more_batches ← TRUE;
WHILE more_batches DO
  BEGIN
  FOR casting_number ← 1 STEP 1 UNTIL 50 DO
    BEGIN
    ok ← FALSE;
    MOVE casting TO pickup+3*zhat*inches                   carry (lift up)
      ON FORCE(zhat)>=20*ounces DO ok ← TRUE;                (with test)
    IF ok THEN
      BEGIN
        good ← good + 1;
        IF pallet_column=4
          THEN BEGIN
            pallet_column ← 0;
            pallet_row    ← pallet_row+1;
          END
        ELSE pallet_column ← pallet_column+1;                carry
        MOVE casting TO pallet+VECTOR(pallet_column,packing_distance,
            pallet_row*packing_distance,0 * inches)
          WITH APPROACH = #*zhat*inches;
        OPEN bhand TO 3*inches:                              dispose
        IF(pallet_column=4)AND(pallet_row=6) THEN BEGIN
            pallet_column ← 0;
          pallet_row    ← 1;
            move_conveyor();
        END;
        MOVE barm TO pickup;                               beginning grasp
      END
      ELSE BEGIN
        bad ← bad+1;
        MOVE casting TO garbage_bin DIRECTLY;                carry
        OPEN bhand TO 3*inches;
        MOVE barm TO pickup;                               beginning grasp
      END;
      casting ← pickup;
      CENTER barm;                                         grasp
    END {of FOR loop}
  IF(bhand<1.5*inches)THEN more_batches ← FALSE;
END; {of WHILE loop}
MOVE barm TO bpark;          return robot to PARK position
END.
```

Figure 3: AL Program to sort castings

being carried out on the parts. We can safely conclude that the parts going to the parameterized locations are being used in an assembly where the parts are being put together in a fixed pattern. The parts being taken to the same location are either being rejected or possibly being transported elsewhere by conveyor belt.

3.2. The Augmented Program

The Augmented Program is a rewriting of the robot's task program that more effectively exploits the robot's sensors and tracks the robot's operation. The control flow of the original program as restructured as a finite state automaton:

- Statements in the original program are split and regrouped into *actions* caused by state transitions in the automaton. The actions are constructed so that each transition causes the robot to make only one motion or perform only one operation.
- A given transition is fired when the conditions are satisfied for the robot's corresponding action. These conditions are a combination of the specific preconditions of that action plus the postconditions of the immediately preceding action.

For example, consider this representation of a simple program that lifts an object:

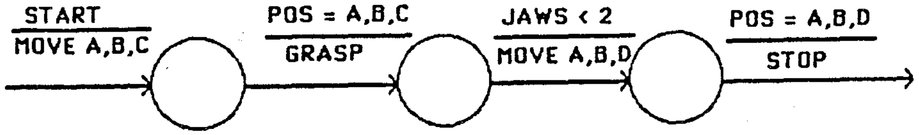


Figure 4: Finite Automaton Representation

We represent sensor readings here as the values of variables such as POSITION and JAWS. A transition, then, is caused when a continuously monitored sensor variable passes some test. We generalize this so that we may also cause transitions by testing user defined variables in the original program. This makes it a simple matter to represent the control structures that occur in typical procedural programming languages.

The values being tested are important. The Augmented Program contains sensor tests for the qualitatively interesting sensor values. These values are used to fire transitions in the automaton. The sensor values are also used to filter the sensor data: we only look for the relevant value.

4. THE ONLINE SECTION

The online section takes care of controlling the robot's actual operation. It uses the Augmented Program generated by the offline section to interpret sensory information and to determine the appropriate action. If the online section detects an error, it uses the Local Knowledge Base generated by the offline section and the Dynamic Knowledge Base to diagnose the problem and plan the recovery strategy.

The online section has two jobs. The Monitor interprets the Augmented Program, using the information from it to direct the robot according to sensory input. The Recoverer is invoked when an error is detected and is responsible for analyzing the error and developing the recovery plan.

The Monitor tracks the robot's activities by monitoring data from the robot's sensors. The sequence of sensor data yields an event trace from which the Recoverer can determine the robot's recent history when an error occurs. Error isolation and recovery planning depend heavily on the location of objects in the workspace. Iterative loops and other control structures make it impossible to derive information about the objects in the workspace from a static analysis of the program. For this reason the online phase keeps track of the objects in the workspace in real time. The details of which robot activities manipulate which objects is included in the Augmented Program by the Preprocessor. The real time monitor uses this information to keep a catalog of

```
begin: event:  !startq
       action: compute:
               packing_distance ← 4"
               pallet_row ← 1
               pallet_column ← 0
               good ← 0
               bad ← 0
               casting ← pickup
             open bhand to 3"
             next_state s1

s1:  event:  !bhand >= 3"
     action: move barm to pickup directly
              next_state s2
     event:  !bhand_stopped
     action: signal_error

s2:  event:  !barm = pickup
     action: center barm
              next_state s3
     event:  !barm_stopped
     action: signal_error

s3:  event:  !barm_centered Λ !bhand < 1.5"
     action: compute:  more_batches ← false
              move barm to park
              next_state s10
     event:  !barm_centered Λ !bhand_stopped
     action: compute:  more_batches ← true
              acquire_object casting
              move_object casting
              move barm to pickup + zhat*3"
                   next_state s4

s4:  event:  !force(zhat) >= 20oz
     action: compute:
                   ok ← true
                   good ← good + 1
                   if pallet_column = 4
                     then begin
                       pallet_column ← 0
                       pallet_row ← pallet_row+1
                     end
                   else pallet_column ← pallet_column+1
              move_object casting
              move barm to approach pallet + vector(
                       pallet_column*packing_distance,
                       pallet_row*packing_distance,0")
                       at zhat*3"
              next_state s7
     event:  !force_different
     action: compute:  ok ← false; bad ← bad + 1
              move_object casting
              move barm to garbage_bin directly
              next_state s5
     event:  !untouch
     action: signal_error

s5:  event:  !barm = garbage_bin
     action: open bhand to 3"
              next_state s6
     event:  !barm_stopped
     action: signal_error
     event:  !untouch
     action: signal_error

s6:  event:  !bhand >= 3"
     action: discard_object casting
              move barm to pickup
              next_state s2
     event:  !bhand_stopped
     action: signal_error

s7:  event:  !barm = approach pallet + vector(
                   pallet_column*packing_distance,
                   pallet_row*packing_distance,0)
                   at zhat*3"
     action: move_object casting
              move barm to pallet + vector(
                   pallet_column*packing_distance,
                   pallet_row*packing_distance,0)
              next_state s8
     event:  !barm_stopped
     action: signal_error
     event:  !untouch
     action: signal_error

s8:  event:  !barm = pallet + vector(
                   pallet_column*packing_distance,
                   pallet_row*packing_distance,0)
     action: open bhand to 3"
              next_state s9
     event:  !barm_stopped
     action: signal_error
     event:  !untouch
     action: signal_error

s9:  event:  !bhand >= 3" Λ pallet_column=4 Λ
                        pallet_row=6
     action: dropped_object casting
              move barm to park
              next_state s10
     event:  !bhand >= 3" Λ ¬ (pallet_column=4 Λ
                        pallet_row=6)
     action: dropped_object casting
              move barm to pickup
              next_state s2
     event:  !bhand_stopped
     action: signal_erro

s10: event:  !barm = park
     action: stop
```

Figure 5: Augmented Program for Example Robot Task

objects in the workspace.

The Recoverer is invoked by the Monitor when it detects an error. The Recoverer consists of two major parts: the *Interpreter* and the *Strategian.* The Interpreter uses information about the robot's activities and sensor readings in the Dynamic Knowledge Base to determine the nature of the error. Information about the robot's task and about the nature of errors from the Local Knowledge Base is also used to interpret the error. The Strategian then tries to come up with a recovery strategy.

4.1. The Monitor

All activities that run concurrently with the robot's actual task are part of the Monitor. These activities are dedicated to tracking the robot's activities and keeping an accurate model of what the robot has been doing. It is essential that the Monitor keep in step with the robot; it must not fall behind in its monitoring task.

The Monitor consists of three separate processes: the *sensor handler,* the *dynamic knowledge base update,* and the real time monitor itself. As seen in Figure 1, the real time monitor works exclusively from information provided in the Augmented Program generated by the offline section. For speed, the monitor itself is given the simplest of testing and sequencing tasks; other tasks are offloaded onto the other two processes.

The sensor handler filters the robot sensor data so that the Monitor only sees significant events. The sensor handler has to keep up with the sensors in real time. Much of this work can probably be offloaded onto a separate processor or performed by special interface hardware. The sensor monitoring tasks are easily performed concurrently and in hardware.

The dynamic knowledge base update process uses the event trace from the monitor to maintain an accurate picture of what the robot has been doing. We will maintain a simple model so that we can update it in real time. Communication from the real time monitor to the model update will be buffered to allow the update process to occasionally fall behind real time. This allows occasional complex updates without slowing down the monitor.

4.2. Sensor Handling

The whole notion of error recovery is very abstract. There are no physical laws couched in algebraic notation that will help isolate or repair errors. We have to contend with the fact that most of the sensors give us the wrong kind of information. We need qualitative facts, not quantitative measurements. We seldom care exactly where the arm is unless it has reached its destination. We don't care what force the arm is subjected to if it hasn't reached the force we're waiting for. The sensor handler process filters and interprets the sensor data so that the real time monitor only sees the significant sensor events.

In essence, these sensor events are signposts or critical moments in the progress of the robot's task. The sensor handler simply watches each sensor and waits for critical readings to appear. The sensor handler signals the monitor whenever a critical reading appears, reporting the sensor involved and the details of the reading.

The expected sensor readings change as the robot moves through its task. For example, we expect the touch sensor to signal if the arm grasps something, but not while the arm is simply moving from one place to another. We don't need the arm to signal when it passes a position it stopped at earlier if we aren't expecting it to stop there this time. These expectations change almost every time the robot moves. Whenever a move takes place the real time monitor encounters a new set of sensory expectations. The monitor transmits new expectations to the sensor handler when necessary, directing the handler's attention to the latest set of significant events.

4.3. The Dynamic Knowledge Base

To recover from an error, the system needs to know what objects are in the workspace and where the objects are. At the time of error the Recoverer can then find out where the AL program failed and what the values of the program's variables are. Unfortunately, we can't deduce the state of the workspace from the state of the program. The program doesn't keep the right kind of information; AL programs don't explicitly refer to objects anyway. But it is possible to deduce when and how the AL program manipulates objects and to place this information in the Augmented Program. This information is then available to the real time monitor.

To monitor objects in the workspace the real time monitor has to be told when in the robot task an object is acquired, grasped, moved, and discarded. For example, the robot *acquires* an object from a part dispenser, *moves* it somewhere, and maybe *discards* a part by placing it on a conveyor. This information is sufficient for keeping track of what objects are in the workspace and where they are. The workspace model update process can then follow objects by monitoring such activities in the event trace.

The workspace model must at least contain a catalog of objects and their locations. Along with the robot's most recent activities, this model should give enough information to determine what was going on at the time of an error. Other kinds of information about the objects may also be worthwhile if it can be easily maintained. For example, an object being manipulated could be tagged with some kind of object type. The recovery process could then use this object type to key into more information about the object stored in the offline world model. The catalog of objects is an appealing approach because it requires little time or space to maintain. It can also be used to construct more elaborate models in later phases of error interpretation.

4.4. The Qualitative Interpretation of Errors

The raw data indicating an error has very little meaning without some context. After detecting the error, the Recoverer has to decide what really occurred. For example, the sensor reading may indicate the misorientation of a part, absence of a tool, or an inaccuracy in arm position. We have developed a list of rules that produce this interpretation. These rules use raw sensor data, the context and semantics of the intended instruction, and knowledge about the effects of AL instructions. Here is the general form of the interpretation rules:

```
IF ((some raw data) and
   (some context) and
   (some semantics) and
   (knowledge about instruction effects))
THEN (a list of one or more Qualitative errors).
```

A variety of qualitative errors can occur during assembly. The following list enumerates typical errors that pertain to general assembly tasks. This list would be different for different tasks and sensors. Since we assume that the AL program is logically correct the list does not include programmer logic errors.

1. Misorientation of part/tool
2. Missing part/tool
3. Slippage of part/tool
4. Misfit or faulty part/tool
5. Inaccuracy in arm position (deflection)
6. Timeout
7. Motor oversaturation or servo error
8. A temporary wobble
9. Excessive force on wrist/hand
10. Cannot close fingers
11. Cannot open fingers
12. Wrist cannot rotate
13. Miscellaneous Gripper problems
14. Excessive speed
15. Collision
16. An attempt to CENTER with no object
17. Some other hardware errors

4.4.1. The Effects of an Instruction

To do precise reasoning, we must have a good causal model. In our context this means we must have a model about the actions. The basis of our approach is the *failure reason model* developed by Srinivas (Srinivas, 1976). Knowledge about the effects of instructions helps in cutting constraining the number of possibilities.

AL instructions can be used in a number of different ways to perform different actions. The table gives examples of the semantic and contextual meanings of the AL MOVE instruction and related kinds of qualitative errors.

Qualitative Errors in the AL MOVE Instruction

	Semantic and contextual meaning	Causes these qualitative errors
1	Move arm up with part (Lift the part)	Slippage, Timeout, Motor Saturation, Inaccuracy, Other system interrupts
2	Move the hand up	Timeout, Motor Saturation
3	Move arm down with part. (Bring down)	Slippage, Inaccuracy Other system interrupts
4	Move hand down	Inaccuracy, Other system interrupts
5	Screw or Unscrew or Rotation	Wobble, Timeout, Slippage, Missing part/tool, Misorientation, Excess force, Gripper problem
6	Carry an object (Transport)	Wobble, Slippage, Inaccuracy, Timeout, Motor Saturation, Excess force
7	Push or Slide	Wobble, Inaccuracy, Servo error, Excessive force or speed
8	Touch an object	Wobble, Collision, Inaccuracy

4.4.2. Error Interpretation using Rules

The error interpretation rules combine raw sensory data, context, and semantics to derive the qualitative meaning of the error situation. Here are two rules which interpret the same raw sensory data in different contexts, given first in an informal representation:

Rule 1. (Finger sensing rule)
IF the fingers were supposed to be open,
  but the fingers right now are actually closed,
  and the robot was trying to grasp part/tool.
THEN possible errors include:
  missing part/tool, or
  misoriented part/tool, or
  temporary hand wobble.

Rule 2. (Finger sensing rule)
IF the fingers were supposed to be open,
  but, right now the fingers are actually closed,
  and the robot was carrying a part or tool.
THEN possible errors include:
  total slippage of part/tool.

Here are the same rules given in a form closer to the actual representation to be used by the Recoverer. Let *Df* represent the measurement of the expected finger separation and let *Af* represent the runtime sensor reading of the finger separation.

```
Rule 1. (Finger sensing rule)
 IF (($Df > 0$) and ($Af = 0$) and
    (attempting grasp part/tool))
 THEN (missing part/tool or
       misoriented part/tool or
       temporary hand-wobble).

Rule 2. (Finger sensing rule)
 IF (($Df > 0$) and ($Af = 0$) and
     (attempting carry part/tool))
  THEN (total slippage part/tool).
```

4.5. Error Recovery

After the Interpreter has analyzed the error the Strategian examines the robot's task as described in the Local Knowledge Base and looks for a way to repair it. The repair could be very simple such as waiting for a temporary mechanical wobble to pass. The repair could also be very complex and involve patching the Augmented Program or involve producing an entirely new version of it. The Strategian performs the following steps:

1. Find the cause of the error from the Interpreter. To an extent the qualitative error description and the robot's execution trace will suggest where the error could have occurred. The error could be either very local or its effects may have propagated through several steps of the robot's task. The execution trace has a list of sensor readings that were explicitly tested by the Monitor. This tells the Strategian which robot actions have been explicitly verified as having happened. The Local Knowledge Base also has a list of the pre- and postconditions not explicitly verified by the Monitor. The unverified conditions indicate places where the Strategian will want to look for possible causes of the error. This combination of offline and online knowledge helps us constrain the list of possible error causes. This technique has been implemented in (Srinivas 1976).

   This technique alone does not guarantee that we will find a single instruction in error. Sometimes there will be ambiguities or conflicting causes of an error. A more powerful technique needs to be used to constrain the error set. The Strategian must know the precise effects of AL instructions.

2. Determine the initial feasibility of repairing or recovering from the error and, if feasible, develop a strategy for repair. To determine the initial feasibility the Strategian must evaluate the following:
   - is the error catastrophic?
   - have any vital tools been lost or put out of operation?
   - what is the qualitative nature of the error (motor burnt, wobble)?
   - are we now short of new materials for another reassembly?
   - what are the additional resources needed? How do we use them?
   - what is the nature of the repair (patch or complete overhaul)?
   - what repair strategies do we have? Some repairs could be faster in time, while others may be less expensive in other ways.
   - are there any any user defined heuristics or criteria to select (or override) our repair strategies?
3. Construct an Augmented Program that implements the recovery strategy.
4. Modify the original Augmented Program used by the Monitor. This may involve substituting new code for old or appending the new code to the existing code. In some cases the original code may need to be discarded completely so that the repair may be carried out.

There are several important issues in the application of planning to this problem. The planner must be able to reason about available resources, their usage characteristics, their purposes and their assignment. The planner must be able to reason about the available operators (applicable operations) at more than one level of abstraction (Wilkins 1982).

Another important problem is that of determining the *degree* of repair needed to recover from an error. More research is needed on recognizing the precise point in the task from which to continue, patch or discontinue the repair plan.

CONCLUSION AND ACKNOWLEDGEMENTS

We have presented the detailed design of a system that allow for a robust robot control. The main advantage of this system is that it allows the robot to recover from unexpected situations using knowledge bases that contain knowledge about what the robot itself is doing.

Special thanks go to Sharon Garber and Donald Stryker who participated in the early stages of this project.

BIBLIOGRAPHY

1. Albus, J., "*Brains, behavior and robotics,* " BYTE Publ., 1981.
2. Ambler, A.P., and others, "An experiment in the offline programming of robots," in *Proc. 12th International Symposium on Industrial Robots,* Paris, France, June 1982, pp 491-504.
3. Binford, T., "The AL language for intelligent robot," in *Languages et Methodes de programmation des robots industriels,* IRIA Press, France 1979, pp 73-88.
4. Bonner, S., and Shin, K., "A comparative study of robot languages," *Computer Magaz.*, December 1982, pp 82-96.
5. Fikes, R.E., and Nilsson, N.J., "STRIPS: a new approach to the application of theorem proving to problem solving," *Artificial Intelligence,* Vol 2, pp 189-208, 1971.

6. Finkel, R. and others, "Overview of AL, a programming system for automation", *Proc. 4th International Joint Conference on Artificial Intelligence,* Tbilisi, USSR, September 1975, pp. 758-765.

7. Friedman, L., "Robot learning and error correction," *Proc. 5th International Joint Conference on Artificial Intelligence,* pp 736, 1977.

8. Gini, G., and Gini, M., "Explicit Programming Languages in industrial robots," *Journal of Manufacturing Systems* Vol 2, N. 1, 1983, pp 53-60.

9. Gini, M., Gini, G., "Towards automatic error recovery in robot programs," *Proc. International Joint Conference on Artificial Intelligence 83,* August 1983, pp 821-823.

10. Lozano-Perez, T., "Robot programming", *Proc. of the IEEE,* Vol 71, N. 7, July 1983, pp 821-841.

11. Luh, J. Y. S., "An anatomy of industrial robots and their controls," *IEEE Trans. on Automatic Control,* Vol AC-28, N. 2, February 1983.

12. Mujtaba, M.S. and Goldman, A., "AL users' Manual", Stanford Artificial Intelligence Laboratory Memo AIM-323, Stanford, Ca, January 1979.

13. Nitzan, D., and Rosen, C.A., Programmable industrial automation, *IEEE Trans on Computers,* Vol C-25, N 12, pp 1259-1270, Dec 76.

14. Paul. R. P., "WAVE: a model based language for manipulator control," *The Industrial Robot,* Vol 4, N 1, pp 10-17, March 1977.

15. Paul, R. P., "*Robot manipulators: mathematics, programming and control,*" Boston, Mass: The MIT Press, 1981.

16. Rosen, C.A., and Nitzan, D., "Use of sensors in programmable automation," *Computer Magaz.,* pp 12-23, Dec 77.

17. Sacerdoti, E., "*A structure for plans and behavior,*" American Elsevier Publ. Company, 1977.

18. Shimano, B., "VAL: a versatile robot programming and control system," *Proc. IEEE Third Int. COMPSAC79,* Chicago, Ill, November 1979, pp 878-883.

19. Smith, R. G. and Nitzan, David, "A modular programmable assembly station," *Proc 13th International Symposium on Industrial Robots,* pp 5.53-5.75, Chicago, Ill, April 1983.

20. Srinivas, S., "Error recovery in a robot system," PhD Thesis, CIT, 1976.

21. Sussman, G. J., "*A computer model of skill acquisition,*" American Elsevier Publ. Company, 1975.

22. Takase, K., Paul, R.P. and Berg, J., "A structured approach to robot programming and teaching," *IEEE Trans. on Systems, Man, and Cybernetics,* Vol SMC-11, pp 274-289, April 1981.

23. Taylor, R.H., Summers, P. D., Meyer, J. M., "AML: a manufacturing language," *International Journal of Robotics Research,* Vol 1, N. 3, 1982.

24. Wilkins, D. E., "Representation in a domain-independent planner", *Proc. 8th IJCAI, Karlshrue, 1983.*

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

A PRECISE SCALING LENGTH FOR DEPLETED REGIONS

Microelectronic and Information Sciences Center

Technical Report #24

R. D. Schrimpf
Department of Electrical Engineering
University of Minnesota

R. M. Warner, Jr.
Department of Electrical Engineering
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

A PRECISE SCALING LENGTH FOR
DEPLETED REGIONS

MEIS Technical Report #24

R. D. Schrimpf
R. M. Warner, Jr.

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

ABSTRACT

A new scaling length is described that is useful in precise modeling of depletion layers. This scaling method also permits accurate empirical modeling of inversion layers associated with depleted regions.

INTRODUCTION

The extrinsic Debye length is an excellent choice for scaling accumulated regions in semiconductors, but it is less effective in scaling the thickness of depleted regions. Recognizing this, Jindal and Warner developed an approximate scaling length (L_{JW}) that is conveniently applied to the depleted regions near PN junctions and semiconductor surfaces [1]. Here we define more precise scaling methods that lead to very nearly universal curves of potential versus position throughout semiconductor samples that possess both depletion and inversion layers. These scaling techniques will be useful for modeling MOS structures and one-sided step junctions, as well as simple PN junctions.

ANALYSIS

We follow the approach and notation of the general solution that has recently been offered for step junctions at equilibrium and for the fully equivalent surface problem [2-5]. The general solution embodies sets of curves for position, electric field, volumetric charge density, and areal charge density, all as functions of electrostatic potential. Each set incorporates a doping-independent portion corresponding to depletion and a doping-dependent portion corresponding to inversion. Approximate-analytic expressions for the former portion have been proposed as a depletion-approximation replacement (DAR) [6], and more recently, another set of equations has been offered for the inversion regime [7]. Here we will concentrate on the relation between position and potential in the depleted region of the sample. Figure 1 shows normalized potential W versus normalized distance (x/L_D), where L_D is the general Debye length, with bulk potential U_B as a parameter. As each curve reaches its threshold of inversion, it can be seen to "peel off" from the common asymptote and rapidly to acquire a near-infinite slope. This threshold potential W_T is defined as occurring when $W = 2U_B$. By defining the position of the threshold plane as our spatial origin, we can obtain a potential-versus-distance curve, valid below threshold, that is universal in the sense that it is independent of bulk doping. A similar curve, valid above threshold, can be obtained using empirical scaling factors. A scaling length appropriate for use in the inversion region is described in the Appendix.

The spatial origin of the DAR occurs at the point where an extrapolation of the linear portion of the field profile intersects the spatial axis. Warner and Jindal used the distance from this origin to the threshold point as their scaling length [1],

$$L_{JW} = L_D \sqrt{2(W_T-1)} = (x/L_D)_T L_D . \qquad (1)$$

This distance is a measure of the maximum depletion-layer thickness that a semiconductor with bulk potential U_B can support at equilibrium. The maximum thickness of the inversion layer is taken to be the distance from the threshold point to the position at which the potential profile acquires a near-infinite slope [7]. This normalized thickness is

$$\left(\frac{\Delta x}{L_D}\right)_{inv} = 0.50455 \exp\left(-0.08344\ U_B\right) + 0.46838 \quad . \qquad (2)$$

The sum of the depletion thickness and the inversion thickness turns out to be a useful scaling length for the depleted portion of our sample:

$$L_{SW} = L_{JW} + [(\Delta x/L_D)_{inv}]\ L_D\ . \qquad (3)$$

Before plotting the results, we have further normalized W by dividing it by the threshold potential W_T. The result can be seen in Fig. 2, which is a plot of data obtained from the DAR, but normalized according to the procedure described above. The curve plotted is for $U_B = 15$, but it is applicable for all values of U_B between 10 and 20 (corresponding to doping levels between $2.2 \times 10^{14}/cm^3$ and $4.9 \times 10^{18}/cm^3$). The maximum error throughout this range is 0.6% of the threshold potential. Error increases gradually as one proceeds from threshold plane to bulk, peaks at about 0.7 L_{SW}, and then decreases again. To illustrate this, we can point out that the maximum error at the position 0.8 L_{SW} amounts to 0.47%. In view of these small error magnitudes, one can say that for practical values of doping, there is a single curve that describes the potential-versus-position relationship in the depletion region with excellent accuracy. This curve begins at the threshold point and asymptotically approaches the bulk potential with increasing distance.

Let us consider the following example to illustrate the use of this universal curve. Consider a MOSFET with bulk doping $N_D = 1.6 \times 10^{15}/cm^3$. The normalized bulk potential is thus $U_B = \ln (N_D/n_i) = 12$. Now apply a voltage to the sample such that the surface potential $U_S = -15$. Making the conversion from normalized potential U to normalized potential W gives us $W_S = U_B - U_S = 12-(-15) = 27$. For our sample with $U_B = 12$, we see that $W_T = 24$. Since $W_S > W_T$, an inversion layer is present. From Eq. (2) we find that the maximum normalized thickness of the inversion layer is $(\Delta x/L_D)_{inv} = 0.65376$. Equation (1) gives us the value of L_{JW}, which is a measure of the spatial extent of the depletion layer, and which we also need to calculate our scaling length. This distance is $L_{JW} = 6.78233\ L_D$. L_{SW} can be calculated from Eq. (3), and its value is $L_{SW} = 7.43609\ L_D$. Potential versus position in the depletion region can now be read off Fig. 2 using the values for W_T and L_{SW} calculated here to convert to the standard format of the DAR, the format of Fig. 1.

CONCLUSIONS

The scaling length L_{JW} offers simple algebraic form but relatively low scaling accuracy, while L_{SW} offers excellent accuracy, but a more complicated algebraic form. Either of these scaling lengths can be used to model depletion layers.

ACKNOWLEDGMENTS

For partial support of this work we are indebted to the Microelectronic and Information Sciences Center at the University of Minnesota, and to the Semiconductor Research Corporation.

APPENDIX

By using an empirical procedure, we will obtain a universal curve similar to Fig. 2, but valid in the inversion layer. For $W > W_T$, the semiconductor is inverted and W increases sharply as a function of x/L_D. In the inversion layer, there is an expression for W as a function of x/L_D that is valid for W up to $2U_B + 8$ [7]:

$$W = W_T + \frac{a}{2} \ln \left[\frac{(\Delta x/L_D)_{inv} + |x/L_D|}{(\Delta x/L_D)_{inv} - |x/L_D|} \right]^{1/b} . \qquad (A1)$$

The parameters a and b in this equation are given by [7]:

$$a = 2.21980 + 0.94357 \ln U_B \; ; \tag{A2}$$

and

$$b = 0.97167 + 0.01042 \ln U_B \; . \tag{A3}$$

A universal curve of potential versus position can be obtained in this regime by an appropriate choice of scaling factors. The required scaling length for distance is the unnormalized thickness of the inversion layer:

$$L_{inv} = (\Delta x/L_D)_{inv} L_D \; . \tag{A4}$$

The potential W must be normalized by the factor $U_B^{1/7}$. The result of this dual scaling is shown in Fig. A1. Again, the curve shown is for $U_B = 15$. For U_B between 10 and 20, the largest divergence in values occurs near $x/L_{inv} = -0.8$, where the maximum error is less than 2%. The actual distance from the origin to the surface can be obtained by substituting the surface potential W_S into this equation [7],

$$\frac{x}{L_D} = - (\frac{\Delta x}{L_D})_{inv} \tanh \frac{(W-W_T)^b}{a} \; , \tag{A5}$$

where a and b were defined above.

We will consider the same example as above to illustrate the use of this scaling technique. It has already been determined that $L_{inv} = 0.65376\ L_D$. To obtain the potential scaling factor, we note that $U_B^{1/7} = 1.42616$. To test our result, we can locate W_S on Fig. A1, and check whether it occurs at the surface position determined from Eq. (A5). This position is found to be -0.37614. The value of the ordinate is thus determined to be (27-24)/1.42616 = 2.1, and reading the abscissa from Fig. A1, we find that the surface is at approximately $x/L_{inv} = -0.58$. This value should agree with the value computed by dividing the surface position by L_{inv}. This ratio is (-0.37614)/(0.65376)

= -0.575 which agrees very well with our visual estimate. If the magnitude of the surface potential is increased, the only effect on the potential distribution is that the surface will now be located farther to the left on Fig. A1. For example, if W_S = 40, the surface will be located on the nearly vertical segment of the curve (the position at which inversion layer thickness is considered to saturate). The rest of the potential distribution remains the same.

REFERENCES

1. R. P. Jindal and R. M. Warner, Jr., "A New Scaling Length for Modeling Semiconductor Space-Charge Regions," IEEE Trans. Electron Devices, Vol. ED-29, 1944 (1982).

2. R. P. Jindal, "Bulk and Surface Effects on Noise and Signal Behaviour of Semiconductor Devices," Ph.D. Thesis, University of Minnesota, Minneapolis, MN, March 1981.

3. R. P. Jindal and R. M. Warner, Jr., "A General Solution for Step Junctions with Infinite Extrinsic End Regions at Equilibrium," IEEE Trans. Electron Devices, Vol. ED-28, 348 (1981).

4. R. P. Jindal and R. M. Warner, Jr., "An Extended and Unified Solution for the Semiconductor-Surface Problem at Equilibrium," J. Appl. Phys., Vol. 52, 7427 (1981).

5. R. M. Warner, Jr., R. P. Jindal and B. L. Grung, "Field and Related Semiconductor-Surface and Equilibrium-Step-Junction Variables in Terms of the General Solution," to be published in IEEE Trans. Electron Devices.

6. R. M. Warner, Jr. and R. P. Jindal, "Replacing the Depletion Approximation," Solid-St. Electron., Vol. 26, 335 (1983).

7. D.-H. Ju and R. M. Warner, Jr., "Modeling the Inversion Layer at Equilibrium," to be published in Solid-St. Electronics.

FIGURE CAPTIONS

1. Normalized potential W versus normalized position X/L_D with U_B as a parameter.

2. Doubly normalized potential W/W_T versus normalized position X/L_{SW} in the depletion region.

A1. Doubly normalized potential versus normalized position in the inversion region.

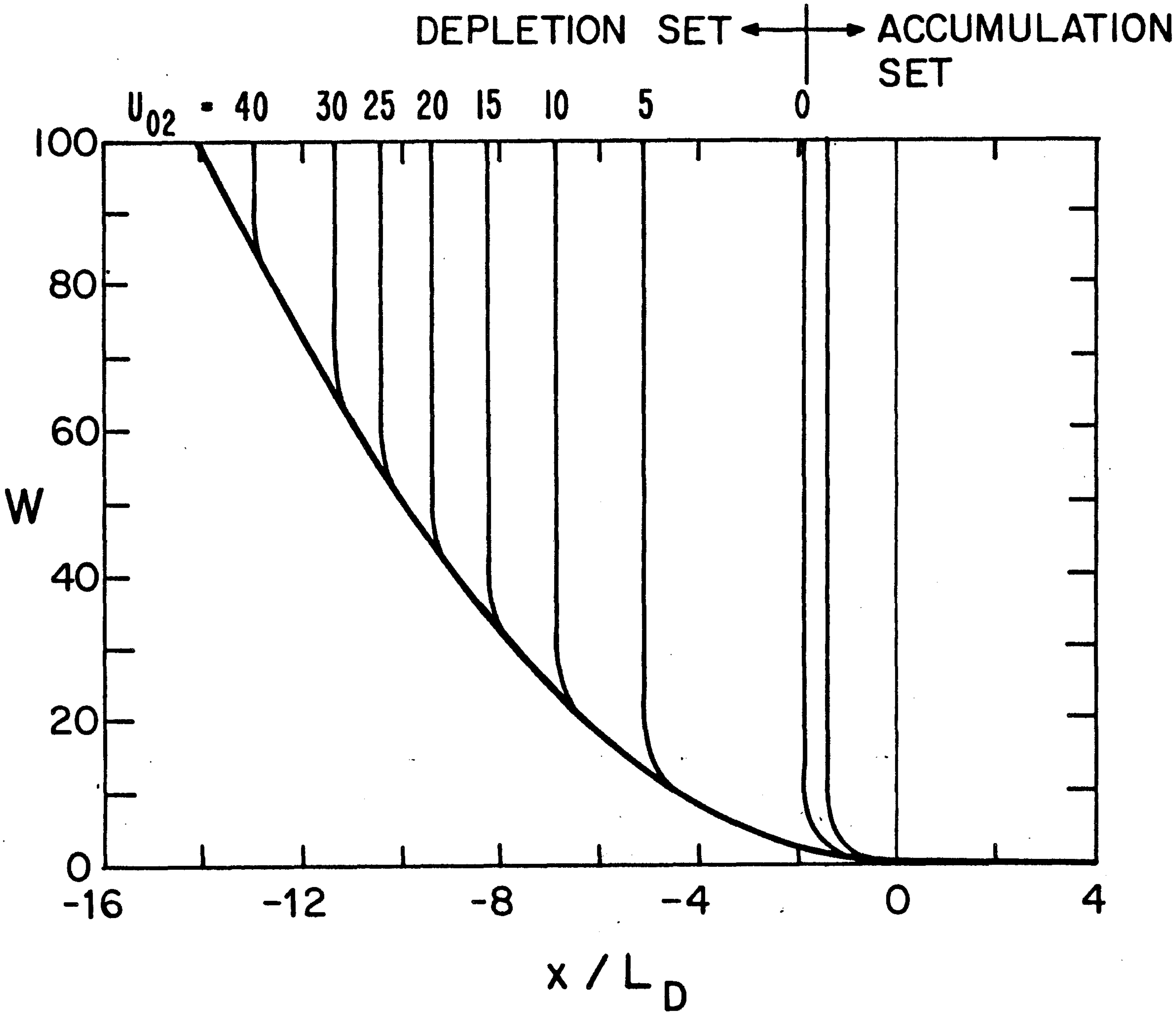

DEPLETION SET
ACCUMULATION SET
U_{02} = 40
30
25
20
15
10
5
0
100
80
60
40
20
0
W
-16
-12
-8
-4
0
4
x / L_D

Fig. 1

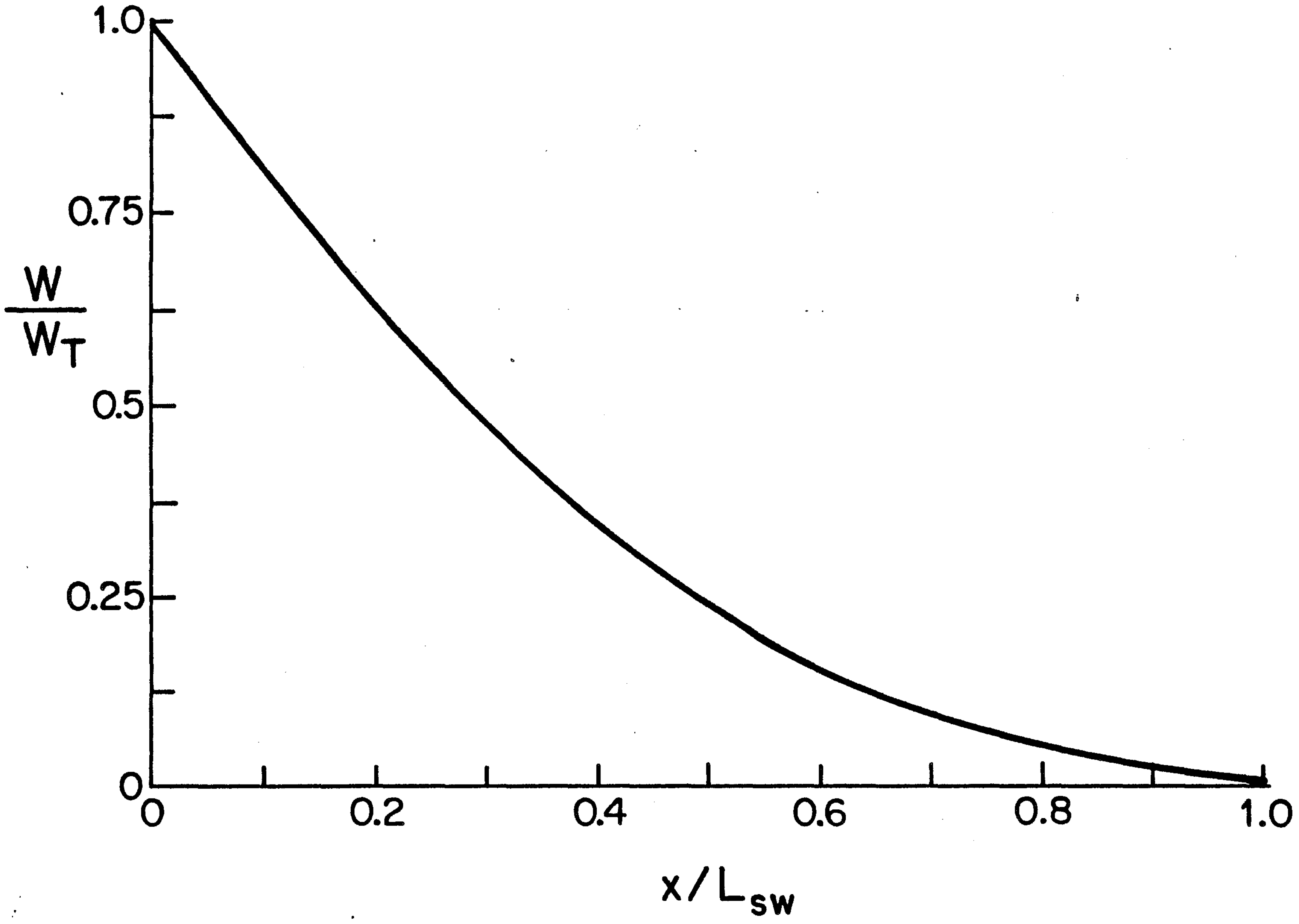


Fig. 2

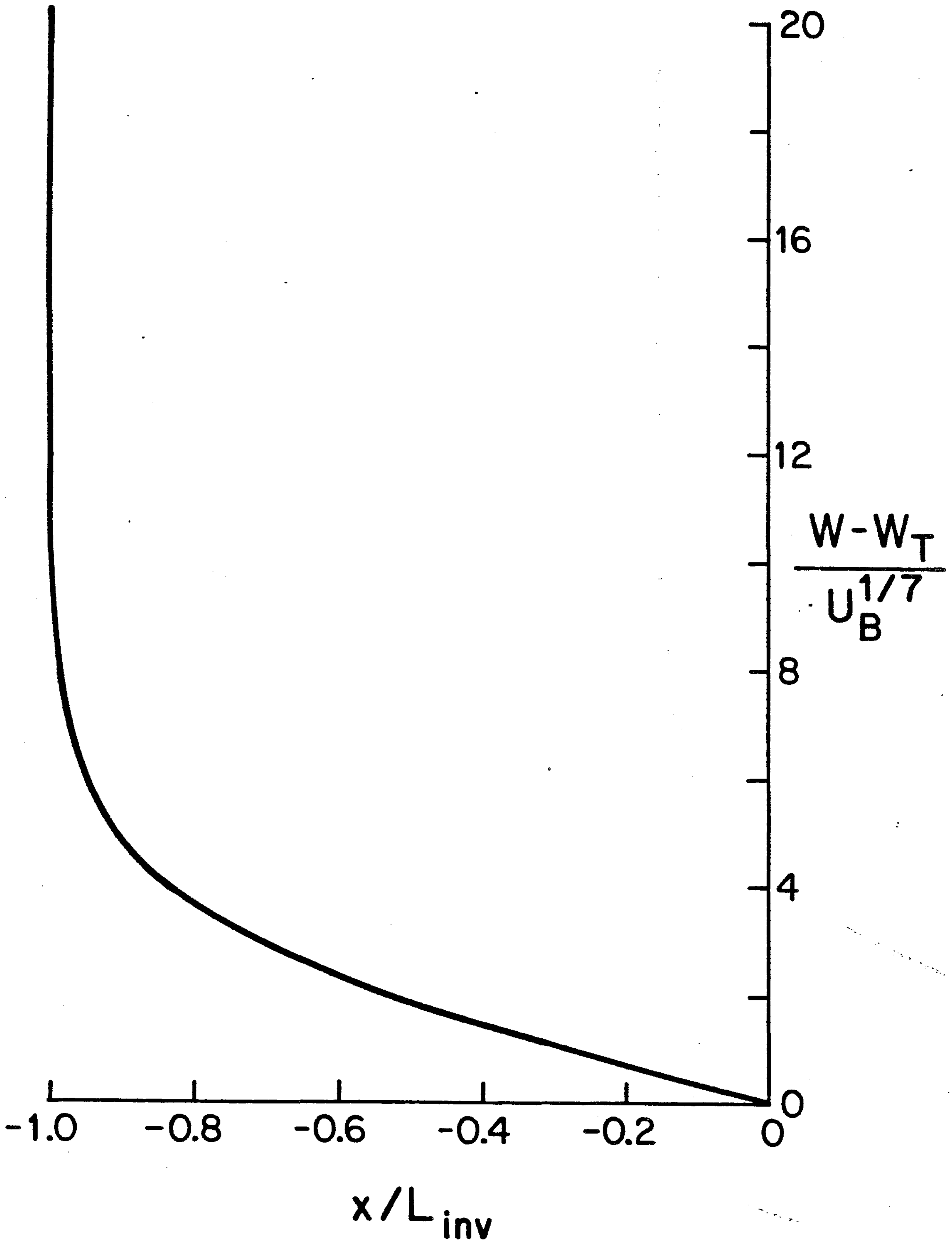


Fig. 3

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

TRANSITION FROM SINGLE-LAYER TO DOUBLE-LAYER STEPS ON GaAs(110) PREPARED BY MOLECULAR BEAM EPITAXY

Microelectronic and Information Sciences Center

Technical Report #25

J. Fuchs
Department of Electrical Engineering
University of Minnesota

J. M. Van Hove
Department of Electrical Engineering
University of Minnesota

P. R. Pukite
Department of Electrical Engineering
University of Minnesota

G. J. Whaley
Department of Electrical Engineering
University of Minnesota

P. I. Cohen
Department of Electrical Engineering
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

TRANSITION FROM SINGLE-LAYER TO
DOUBLE-LAYER STEPS ON GaAs(110) PREPARED
BY MOLECULAR BEAM EPITAXY

MEIS Technical Report #25

J. Fuchs
J. M. Van Hove
P. R. Pukite
G. J. Whaley
P. I. Cohen

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

[1986?]

ABSTRACT

Even though GaAs (110) is the only semiconductor whose surface structure is known with confidence, little is known about its microscopic growth mechanisms. We have used RHEED to study the role of steps in the MBE growth of GaAs on vicinal GaAs(110) surfaces which were misoriented by less than 2 mrad. After thermally desorbing the initial oxide, 20 layers of GaAs deposited at 700K produced a surface with single atomic-layer steps having an average terrace length of a few hundred Angstroms. Upon annealing to 800K, a slow mass migration occurred producing a surface with one thousand Angstrom average terrace lengths and predominantly double layer step heights. The RHEED pattern was nearly instrument limited at in-phase angles of incidence, with little background intensity and bright Kikuchi lines. Subsequent deposition showed only weak oscillations in the RHEED intensity, in contrast to growth on the (001) surface. The period of the observed oscillations indicates that the layer-by-layer growth involves single-layer steps. Growth of as little as 5 atomic layers on a surface with double steps could not be annealed to give a RHEED intensity as great as the first annealed surface. These measurements reconcile previous LEED results with the oxygen adsorption measurements of Ranke. The results clearly show the dominance of steps in the formation of RHEED streaks.

INTRODUCTION

Attempts to prepare the (110) surface of GaAs by molecular beam epitaxy (MBE) have met with varying degrees of success [1-5]. Most reports indicate that the surface morphology is much rougher than that of a (001) surface prepared under similar conditions. Low growth rates and low substrate temperatures are needed for optimum growth. This suggests that by going to low temperature to avoid Ga agglomeration, the surface diffusion of adsorbed species is also low and responsible for the defects produced. Petroff and co-workers [2] commented that smooth surfaces of the ternary GaAlAs could be prepared at more usual (001) growth conditions if the (110) surface were misoriented towards an As terminated (111)B surface, perhaps indicating that Ga clustering can be prevented by providing a sufficient density of As step sites. Despite these growth difficulties, the singular (110) is important since it connects the vast body of surface science performed on cleavage surfaces with the increasing literature of epitaxial growth on the (001). It is the only semiconductor whose surface structure is known with certainty [6], making it the best surface to test microscopic models of growth as well as, for example, dissolution processes [7]. The purpose of this work was to characterize the surface morphology of GaAs(110) prepared by MBE using reflection high-energy electron diffraction (RHEED). Especially because of the high mobility of Ga on the (110), we expect that surface steps are important in the growth and will use RHEED to study their formation.

There have been three previous low-energy diffraction (LEED) studies of steps on GaAs(110) surfaces prepared both by vacuum cleavage and by ion bombardment [8-10]. All found that surfaces with single-layer steps were formed and that the average terrace lengths would increase upon annealing. Clearfield and Lagally [10] (CL) examined the annealing rate and determined that two separate processes were involved. In addition Ranke and co-workers [5] modeled the adsorption of oxygen at defect sites on vicinal GaAs(110) surfaces prepared both by ion bombardment followed by annealing and by MBE. On the ion-bombarded surfaces, like the LEED studies, they found single layer steps. On the MBE prepared surfaces, which were inclined by less than a few degrees toward the (111)A (surfaces that Petroff et al found were rough), double layer steps were observed. For the latter surfaces prepared by MBE, the step distribution close to the (110) and toward the (111)B were less certain. To account for the difference between the ion-bombarded and MBE surfaces they, like Clearfield and Lagally, suggested that more than one ordering mechanism operated on the surface. In the next sections we will show that on nearly singular GaAs(110) surfaces either single- or double-layer steps could be prepared by MBE, depending only on the substrate temperature. Hence the differences in the processes might not be due to the damage introduced into the lattice by ion bombardment. Further we show that GaAs(110) grows by single-layer step propagation even though the formation of double-layer steps is preferred.

EXPERIMENTAL

Sn doped GaAs(110) wafers (Morgan, nominally 10^{18} cm^{-3} Si) were prepared by the methods normally used for the (001) [11] with the exception that small samples were cleaved from the wafers after etching. The samples were about 1 cm on edge. Upon introduction into the growth chamber of the MBE apparatus, the samples were heated, under an As_4 flux, to 900K for less than 5 min to drive off the surface oxide left by the chemical etch. The sample was then cooled to 700K where the initial depositions were performed. The sample temperature was measured with a thermocouple pressed against the back of the sample holder. Absolute temperatures were known to within ± 20°C. The apparatus and procedures for the 10 keV RHEED measurements have been described elsewhere [11,12]. Samples were found to be misoriented by less than 2 mrad by the method of x-ray goniometry and by the RHEED method of Pukite et al. [18]. The As_4 flux was maintained at 3×10^{14} $molec{\cdot}s^{-1}{\cdot}cm^{-2}$ throughout the experiments in order to prevent depletion of As from the surface.

DISCUSSION AND RESULTS

To characterize the step distribution on these GaAs(110) surfaces we will apply an analysis similar to that used by Henzler [8] and by Lagally and co-workers [9,10] but our measurements will be made with RHEED. The fundamental idea is that electrons scattered from the top and bottom terraces of surface steps will interfere constructively or destructively, depending upon the path length difference. In the LEED studies this path length difference is varied by changing the electron energy; in our RHEED measurements it is more convenient to vary the angle of incidence. At several points in the growth, we measure the intensity along the length of the specular RHEED streak at several angles of incidence. This is similar to LEED measurements of the intensity across a diffracted beam; the main differences are that in these RHEED measurements the component of momentum transfer perpendicular to the surface is not as constant and that RHEED is sensitive to order over much larger distances [13]. At angles of incidence where scattering is constructive (in-phase or Bragg angles) the diffraction is insensitive to steps and the diffracted beam is sharp. At angles where diffracted electrons from different surface levels are π out of phase the interference is destructive, and because of the range of terrace lengths the beam is broadened. Detailed discussions are given in refs. 14 and 15.

After the initial desorption of the surface oxide, the diffraction pattern was weak with a diffuse background evident. With the sample at 700K about 20 layers of GaAs were deposited. In agreement with Kroemer [1] we had found this temperature to give the strongest diffraction pattern during slow ($\lesssim$1/3 μm/h) steady-state growth on this nearly singular surface. At this temperature CL [10] found a noticeable change in the width of the diffracted beams on ion-damaged samples after annealing for about 10 min, indicating that surface species are mobile. After this initial short deposition the specular intensity doubled and the diffuse background became relatively weaker. Fig. 1 shows angular profiles of the specular streak from the resulting surface. These curves were measured with the incident electron beam directed along the $[\bar{1}10]$ axis. For single-layer (110) steps, constructive interference should occur at glancing angles of incidence of 31 and 62 mrad (no refraction correction is needed) and destructive interference at 46 mrad. In Fig. 1 the intensity along the specular streak is plotted vs. the angular deviation from the peak. Note that close to the out-of-phase angle the beam is broad (i.e. the streak is long) and at the Bragg angles the beam is sharp. To rule out double-layer steps one in principle could fit the data to a model and then check for the different angle of incidence dependencies of random double and single layer steps [14]. Though we have not yet done this, it is clear that single steps are present. At the Bragg angle, the beam is about 0.6 mrad wide, corresponding to the instrument limit. To determine the mean terrace size from the out-of-phase profile, one needs to make some assumptions about the distribution of steps. If the steps are noninteracting and the distribution of steps is the same on each level, then the mean terrace size is of the order of the reciprocal of the half-width [16] of the diffracted beam. Taking into account the low angle at which the Ewald sphere cuts the reciprocal lattice rod, this corresponds to about 200Å. The asymmetry of the profiles is similar to what is observed for two-level systems where one can fit the data by calculating the intersection of the Ewald sphere with a step-broadened reciprocal lattice rod [15]. Thus in contrast to the MBE experiments of Ranke et al., single-layer steps can be obtained on MBE prepared material when growth takes place at low temperature. Apparently ion damage is not needed to limit the surface mobility.

The annealing behavior of the step distribution is shown in Fig. 2. Keeping the As_4 flux constant, the sample temperature was raised to about 790K and the intensity along the streak scanned at increasing times. For these scans the glancing angle of incidence was fixed at 46 mrad which is the out-of-phase angle for a surface with steps that are a single atomic layer high. The time required to record a scan was 10 s. As shown, the specular beam sharpens to 1.7 mrad after 35 min at this temperature. A few points are worth noting. First, we could not reach steady state at 790K quickly enough to look for the break in the time dependence of the half-width observed by CL [10]. Second, though these data were measured with the incident electron beam along a symmetry axis, similar results are also obtained a few degrees away from symmetry. Hence we do not think that dynamic effects are important in analyzing the shape of the diffracted beam (they are clearly important in analyzing the intensity). Finally, the observed sharpening of the diffracted beam during annealing at the single-layer out-of-phase angle could mean either (1) the average terrace length of the existing steps becomes large or (2) the step height changes to two (110) layers. The latter possibility arises because the single-layer out-of-phase is also a double-layer Bragg angle.

To distinguish between these two different distributions, the angular profile of the specular streak was measured for a few angles of incidence as shown in Fig. 3. For double-layer steps the Bragg angles should be near 15, 31, and 46 mrad, with out-of-phase angles halfway in between. If the surface were single-layer stepped, then 46 mrad would be an out-of-phase angle and curves on either side of it would be sharper; instead, the data clearly indicates that double-layer steps predominate. Some single layer steps remain even after the annealing procedure since the angular profile at the double-layer Bragg (single-layer out-of-phase) is broader than the instrument response of the diffractometer. It is interesting that neither Henzler [8] nor Lagally and co-workers [9,10] saw this transition. A major difference is that the instrument response in these RHEED experiments was greater than that of their LEED instruments: the average terrace length of these double layer steps is of the order of $4/(k\cdot\Theta\cdot\delta\Theta)$~1000Å [14] which would have been unobservable within the several hundred Angstrom instrument limit of their experiments. Two other differences were that these samples were not damaged by ion-bombardment and that an external As_4 flux was present during the heat treatment so that the surface did not become As deficient. Our results are consistent with the more indirect measurements of Ranke et al. [5] who, in order to account for the orientation dependence of oxygen uptake, showed that double-layer steps were present on MBE prepared surfaces. From their data, though, the evidence on the singular surface was less conclusive than on the vicinal surfaces oriented towards the (111)A. On ion-damaged surfaces Ranke et al. had found only single-layer steps near the singular orientation; but because there was no external As flux they could not anneal to quite as high a temperature. In light of these measurements it is surprising that Lagally and co-workers did not see the formation of double-layer steps. They did observe a lengthening of the single-layer terraces.

Using RHEED one can also follow the steps during growth. Fig. 4 shows the intensity of the specular RHEED beam at the single-layer out-of-phase angle as a function of time after growth is initiated on the annealed (110). For comparison, data from (001) growth are also shown, though under slightly different growth conditions. These intensity oscillations result from the competition between nucleation and step-propagation in the layer-by-layer growth of GaAs on these surfaces [12, 14, 17]. The period of the oscillations corresponds to the time required to deposit a monolayer of GaAs. The striking features are that (1) the intensity oscillations are typically much weaker on the (110), dying out after about 5 periods, and (2) the period of the oscillations on the (110) corresponds to the deposition of single layers of GaAs, in this case 40s. These oscillations have been observed during the initial growth on annealed substrates held between 700K and 900K. Thus even at substrate temperatures where the surface prefers double-layer steps in steady state, the system tries to grow via the nucleation and propagation of single-layer steps. Even at the highest temperatures there is insufficient time for the double-layer steps to form.

If growth on this nearly singular surface is interrupted after only 5 layers are deposited and then allowed to anneal, the resulting pattern corresponds to a surface that is never as well ordered as the starting annealed surface. For example, after deposition and annealing at 850K, the 46 mrad angular profile is as sharp as the initial surface but the intensity is reduced by a factor of two, indicating that there is some random disorder over the surface other than atomic steps.

It is important to realize that there was a large difference in the quantitative scale of these measurements and those reported by Lagally and coworkers on ion-damaged GaAs(110). First, the average terrace length of the stepped surfaces prepared by deposition at low temperature began at several hundreds of Angstroms -- already above the instrument limit in the LEED experiments. The average terrace length of the annealed surface was an order of magnitude larger. Second, because of the greater sensitivity of RHEED to large distances, the annealing process could be followed for much longer times. Unfortunately, the sample heating and temperature measurements of our MBE apparatus are not yet suitable for making comparisons with either the early annealing behavior or the activation enthalpy measurements of CL. Third, the time variation of the width of the angular profiles in Fig. 2 can be seen to be far more rapid than the width variation reported by CL.

Last, we should mention two preliminary measurements of the kinetics of the surface migration processes. After the initial deposition of 20 layers at 700K the intensity increased with a $t^{1/2}$ time dependence. In addition, the average terrace length from the data of Fig. 2 increases according to $t^{0.4\pm0.1}$. Both indicate the role of surface diffusion.

CONCLUSION

The step distribution of very thin layers of GaAs deposited on nearly singular GaAs(110) surfaces by molecular beam epitaxy is shown to depend sensitively on the substrate temperature. At low temperatures random steps with several hundred Angstrom terrace lengths and single layer step heights are formed. The annealing behavior of these stepped surfaces is different than that observed by LEED measurements of ion-damaged, As deficient surfaces. At about 800K long average terrace lengths and double-layer steps are formed, directly corroborating the measurements of Ranke et al. on vicinal surfaces. Further growth on the very long terraces of these surfaces resulted in cyclic variations in the diffracted intensity, corresponding to the growth of single layer steps. After subsequent depositions disorder was present that could not be annealed, in contrast to the rapid anneal of the initial growth. The angular dependence of the shape of the RHEED streaks agrees with an analysis that emphasizes the role of steps. The shape and angular dependence are observed to change dramatically with temperature. The shape of the RHEED beams was not observed to depend on the azimuthal angle of incidence. Steps were found to be a major cause of RHEED streaks on these MBE surfaces.

ACKNOWLEDGEMENTS

This work was partially supported by the Corrosion Center of the University of MN, by the MN Microelectronics and Information Sciences Center, and by the NSF under grant DMR-8319821.

REFERENCES

1. H. Kroemer, K. J. Polasko, and S. C. Wright, Appl. Phys. Lett. 36 (1980) 763.
2. P. M. Petroff, A. Y. Cho, F. K. Reinhart, A. C. Gossard, and W. Weigmann, Phys. Rev. Lett. 48 (1982) 170.
3. J. M. Ballingall, and C. E. C. Wood, Appl. Phys. Let. 41 (1982) 947; J. Vac. Sci. Technol. B1 (1983) 162.
4. W. I. Wang, J. Vac. Sci. Technol. B1 (1983) 630.
5. W. Ranke, Physica Scripta. T4 (1983) 100; W. Ranke, Y. R. Xing, and G. D. Shen, J. Vac. Sci. Technol. 21 (1982) 426; W. Ranke, Y. R. Xing, and G. D. Shen, Surf. Sci. 120 1982) 67.
6. A. Kahm, Surf. Sci. Reports, 3 (1983).
7. H. Gerischer, J. Vac. Sci. Technol. 15 (1978) 1422.
8. M. Henzler, Surf. Sci. 22 (1970) 12.
9. D. G. Welkie and M. G. Lagally, J. Vac. Sci. Technol. 16 (1979) 784.
10. H. M. Clearfield and M. G. Lagally, J. Vac. Sci. Technol. A2 (1984) 844.
11. P. R. Pukite, J. M. Van Hove, and P. I. Cohen, J. Vac. Sci. Technol. B2 (1984) 243.
12. J. M. Van Hove, C. S. Lent, P. R. Pukite, and P. I. Cohen, J. Vac. Sci. Technol. B1 (1983) 741.
13. J. M. Van Hove, P. R. Pukite, P. I. Cohen, and C. S. Lent, J. Vac. Sci. Technol. A1 (1983) 609.
14. C. S. Lent and P. I. Cohen. Surf. Sci. 139 (1984) 121.
15. P. R. Pukite, C. S. Lent, and P. I. Cohen, to be submitted to Surf. Sci.
16. T.-M. Lu and M. G. Lagally, Surf. Sci. 120 (1982) 47.

17. J. M. Van Hove, P. R. Pukite, and P. I. Cohen, J. Vac. Sci. Technol. B (1985) in press.

18. P. R. Pukite, J. M. Van Hove, and P. I. Cohen, Appl. Phys. Lett. 44 (1984) 456.

Figure 1. Angular intensity profiles of (00) beam at several incident angles. Bragg angles are multiples of 31 mrad for single layer steps.

Figure 2. Intensity profiles of the (00) beam showing the transition from single- to double-layer steps at 790K.

Figure 3. Intensity profiles of (00) beam after annealing at 790K. Bragg angles for double-layer steps are multiples of 15 mrad.

Figure 4. Typical RHEED Intensity Oscillations vs. time for the (110) and (100) surfaces. The period corresponds to the single-layer deposition time (different time scales).

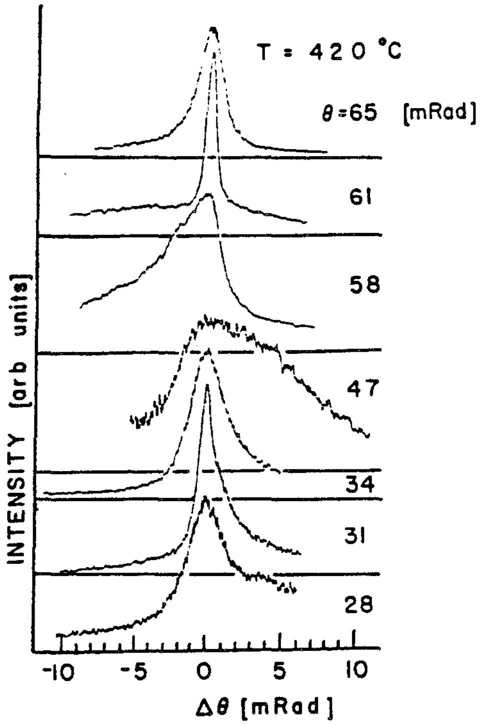


Figure 1. Angular intensity profiles of (00) beam at several incident angles. Bragg angles are multiples of 31 mrad for single layer steps.

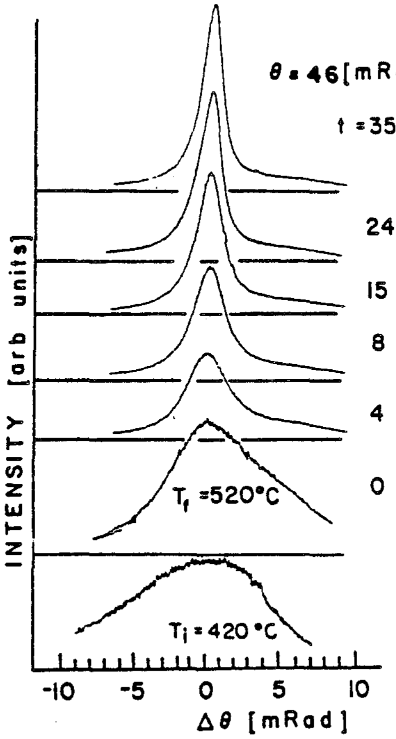


Figure 2. Intensity profiles of the (00) beam showing the transition from single- to double-layer steps at 790K.

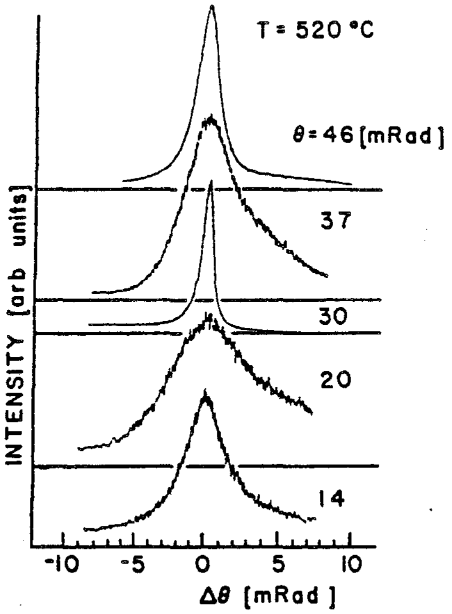


Figure 3. Intensity profiles of (00) beam after annealing at 790K. Bragg angles for double-layer steps are multiples of 15 mrad.

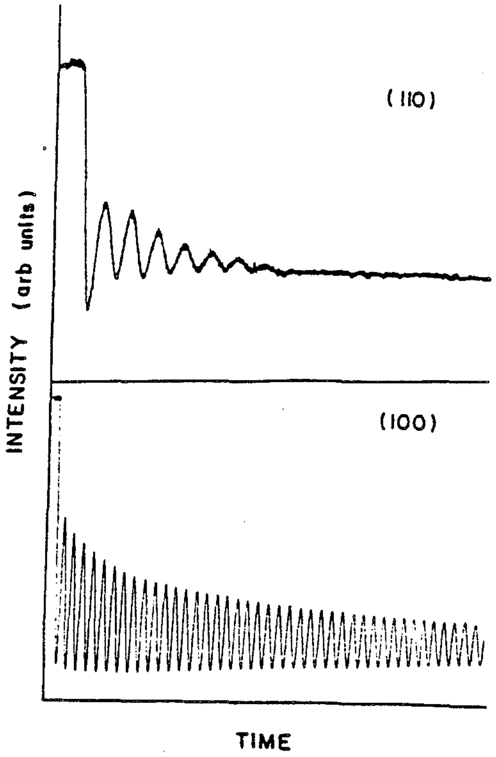


Figure 4. Typical RHEED Intensity Oscillations vs. time for the (110) and (100) surfaces. The period corresponds to the single-layer deposition time (different time scales).

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

SYNCHROTRON RADIATION PHOTOEMISSION STUDIES OF INTERFACES

Microelectronic and Information Sciences Center

Technical Report #26

J. H. Weaver
Department of Chemical Engineering
and Materials Science
University of Minnesota

Chapter 2 in Analysis and Characterization of Thin Films
ed. K.N. Tu and R. Rosenberg
Treatise on Materials Science and Technology, Vol. 28 (Academic Press)

SYNCHROTRON RADIATION PHOTOEMISSION
STUDIES OF INTERFACES

MEIS Technical Report #26

J. H. Weaver

This research was supported in part by the Army Research Office, the Office of Naval Research and the Microelectronic and Information Sciences Center

ABSTRACT

In this chapter we discuss synchrotron radiation photoemission as it is applied to studies of interfaces. The important scientific and technological issues of interface formation are defined. Case studies are used to show the contributions that can be made with the various techniques of photoemission. They emphasize ordered and disordered overlayer formation, cluster formation on surfaces, disruptive atomic intermixing, interface abruptness, diffusion barriers, and enhanced surface reactivities.

TABLE OF CONTENTS

Statement of Purpose.. 3

Introduction.. 4

Synchrotron Radiation Photoemission................................ 6

Case Studies..11

Ordered Chemisorption: Cl/Si..11

Abrupt Interfaces: Ag Overlayers and Ge/Ta(110)....................12

Cluster Formation and Weak Interaction: Ag/GaAs and Ag/Ge...........13

Cluster-Induced Reactions: Ce/Si(111)...............................14

Atomic Intermixing at Interfaces: Pd/Si and Cr/GaAs.................15

Atomic Mobility and Diffusion Barriers: Au/Al/GaAs and Au/Cr/Si.....20

Altering the Chemical Reactivity of Interfaces: Oxidation...........21

Concluding Remarks...22

Acknowledgments..23

References...24

Figures..31

STATEMENT OF PURPOSE

The purpose of this chapter is to describe the scientific and technological advances that can be made by using synchrotron radiation photoemission in studies of interfaces. We will show that synchrotron radiation photoemission is one of the most flexible probes of surface/interface properties and, although it is by no means the only probe, it is the technique of choice for many kinds of research. Much of the discussion will revolve around case studies and examples of metal/semiconductor and metal/metal interfaces. We will not attempt to review all the excellent work that has been done or be encyclopedic in our treatment of interfaces - the interested reader is referred to exhaustive and specific reviews by Brillson,[1] LeLay,[2] Poate/Tu/Mayer,[3] Rubloff,[4] and Ludeke[5] and others for that treatment. Instead, our focus here is on the techniques of synchrotron radiation photoemission as applied to interfaces.

The outline of this chapter is as follows: we will first define the challenges inherent in understanding interface formation. The discussion will consider the properties of atoms, thus defining the scale of our research to be atomic, and will investigate the evolution of interfaces from clean surfaces through thick overlayers or atomic intermixtures. The second section will review the techniques of synchrotron radiation photoemission, keeping in mind its use in interface modeling. Emphasis will therefore be placed on techniques which are best suited for that goal. The third section will address the persistent issues of modeling of any interface. In the closing paragraphs, we offer personal speculations as to the future of synchrotron radiation photoemission in interface research.

INTRODUCTION

Interfacial phenomena can be extraordinarily complex. However, since interfaces are everywhere, their study is both scientifically fundamental and technologically relevant. One need only consider the physical world to recognize that interfaces between dissimilar materials are unavoidable - no two distinguishable objects can be joined without an interface. Of course, the scale of the interface varies a great deal, depending on the objects (solids, liquids, vacuum, etc.) and the means by which they are joined.

In this chapter, we will focus on metal/metal and metal/semiconductor interfaces formed at room temperature. Ideally, we would take each side of the boundary to be infinite in extent, but photoemission is not able to study such buried interfaces because photo-excited electrons have finite scattering lengths. Instead, photoemission probes the upper 2-30 layers of a junction, depending on electron kinetic energy (probe depth ~ 4 times scattering length). Hence, we will be examining the properties of the surface region, starting with the vacuum/solid junction and modifying it as the nominal thickness of the overlayer changes. From these results, we will infer the character of the buried interface based on systematics regarding chemical reaction and possible atomic intermixing. We should point out, perhaps, that the technique of photoemission is nondestructive in its evaluation of interfaces and is able to determine the onset of reaction with very high surface sensitivity, thereby being complementary to more bulk-sensitive techniques like Rutherford back scattering or TEM.

In Fig. 1 we define an interface and, with it, some of the phenomena that are observed during its evolution. On the top left, we have sketched a perfect array of atoms representing a clean surface under the assumption that the surface structure has been adequately characterized (crystallographic orientation, surface reconstruction or relaxation, defect density, etc.). The wealth of excellent surface research of the last decade based on photoemission, LEED, Auger spectroscopy, ion scattering, electron energy loss spectroscopy and other techniques discussed elsewhere in this volume has lead to sophisticated understanding of many surfaces of importance. Progress in the next few years will doubtless bring other surfaces into the "understood" category, particularly in light of the major advances in modeling and theory.

When metal adatoms impinge on a metal or semiconductor surface, they can chemisorb and form hybrid bonds with the substrate. In Fig. 1 we show isolated adatoms with three-fold, two-fold, and single coordination on the surface representing (from left to right) hollow, bridge and atop site geometries. With increasing coverage, adatom-adatom interactions become possible, either involving the substrate (substrate modulated) or independent it. These overlayer interactions are the precursors for extended layer formation, with relevant chemical parameters being adatom-surface bonding and adatom-adatom bonding.

The growth of an adatom array is influenced by such variables as surface mobility, bonding energy, defect density, temperature, and method of deposition. Several growth patterns have been discussed in the literature of thin films.[6] One involves growth with the adatoms in registry with the substrate, i.e. commensurate epitaxial growth where the overlayer crystal structure matches that of the substrate,[6] as sketched in Fig. 1. This requires that the bulk lattice constants of the respective systems be close enough that lattice strain is small (the adatoms and substrate atoms are of equal size in our sketch of epitaxy). Also possible is incommensurate epitaxy in which the overlayer mesh can be generated by a single mathematical transformation of the substrate mesh.[6] Overlayer nucleation and lateral growth can still occur for systems where lattice mismatch is excessive ($\gtrsim$8%) or the electronic structures are unfavorable for epitaxy, but the result is an array of disoriented microcrystals, sketched as patches in Fig. 1. In both cases drawn in the second panel from the top of Fig. 1, the first layer is assumed to be complete before growth of the second begins. If, on the other hand, three dimensional growth begins immediately following nucleation on the surface, the surface is initially more heterogeneous with cluster formation. Ultimately these clusters coalesce must but the overlayer can be quite rough. A third pattern describes the formation of a full monolayer with subsequent island growth. Again, this represents a heterogeneous interface (not shown).

Each of these patterns describes an abrupt interface between the overlayer and the substrate. In contrast, adatom-induced disruption and atomic rearrangement across the original boundary can also occur. This is the most common behavior for metal/semiconductor interfaces, even at room temperature.[1] Disruption for these intermixed systems can occur spontaneously upon arrival of the first adatoms or can be delayed until well-defined or critical coverages are reached, as will be discussed later.

Once atomic intermixing has been initiated, the goal of interface research is to characterize the reaction products as a function of the physical and chemical parameters of the system. To do so requires a detailed assessment of local bonding configurations and the spatial distribution of constituent atoms. Three different distributions are sketched in Fig. 1 corresponding to a random distribution of atoms, an ordered array (compound), and an array with local order but no long range order.

In this chapter, we will consider examples of these interfaces, we will point out some of the exciting scientific challenges of this research discipline, and we will show the significance of synchrotron radiation for each study. In preparation for that, however, we first provide a general discussion of synchrotron radiation photoemission.

SYNCHROTRON RADIATION PHOTOEMISSION

The great power of photoemission for interface research rests in its ability to examine the electronic energy states of a system as they are modified by chemical processes. Photoemission does so with a variety of different techniques, including valence band and core level studies. Synchrotron radiation is the ideal source for such studies, as will be discussed here. In addition it should be noted that synchrotron radiation is extremely important for other types of interface research, including SEXAFS and NEXAFS/XANES, which are also photoemission techniques,[7] and photon stimulated desorption.[8]

During the ordered coalescing of an ensemble of atoms to form a solid, the outermost or more energetic electrons become significantly delocalized, as discussed in any introductory text on solid state or modern physics.[9] The resulting electronic energy states can be determined from the solution for the many-electron Hamiltonian with suitable approximations for electron-nucleus, electron-electron (Coulomb, exchange, correlation), and higher order terms in the potential energy. In turn, the equilibrium lattice constant and the crystal structure represent the stable or lowest energy arrangements of the atoms. In Fig. 2 we show a plot of the potential energy as a function of position in a plane cutting through the solid. As indicated, the outermost, loosely bound atomic energy levels overlap to form delocalized bands and the tightly bound core levels retain their atomic character. Solid state bonding is then the result of this redistribution of atomic states into valence states. In turn, this valence electron charge helps screen the core electrons from the nucleus and thereby influences their binding energies. Hence, an assessment of the chemical state is possible by probing the core states and detailed information about bonding comes from valence band studies (electron spectroscopy for chemical analysis or ESCA was the original acronym for x-ray photoelectron spectroscopy, appropriately developed for fingerprinting the chemical states of atoms). Upon chemical reactions at a surface/interface, there can be significant changes in the distribution of electronic states - related to adsorbates, clusters, epitaxial or non-epitaxial layers, or surface disruption and compound formation. These are shown very schematically by adatom modification of the potential near the surface. Photoemission is the most direct technique for studying electronic states and it has been widely used to examine interface properties, by itself or in conjunction with other probes.[1]

Photoemission can be most readily described by the three step model which stipulates the absorption of a photon by an electron, transport of the excited electron to the surface, and escape of the electron into vacuum where it can be detected.[10] Within this simple model, it is assumed that the energy difference between the initial and final states is equal to the photon energy. (See Ref. 11 for a discussion of final state or many body corrections to this assumption.) Although the photon skin depth or characteristic absorption depth can be 10's to 100's of Angstroms or more,[12] only electrons excited near the surface are likely to escape without inelastic scattering. Subsequent measurement in vacuum of the number of these primary electrons as a function of energy and direction of emission provides information about the energy and momentum distribution of initial states of the solid. If the photon energy is large so that its momentum is not negligible and the distribution of final states is sufficiently independent of crystal momentum, k, then the photoemission spectrum closely resembles the density of initial states (density of states regime). If the photon energy $\lesssim$30-50 eV, then variations in the dipole matrix elements involving initial and final states must be considered (band structure regime[13,14]) and the absorptive part of the dielectric function becomes

$$\varepsilon_2(\omega) = \frac{\hbar^2 e^2}{3\pi^2}\sum_{i,f}\int_k d^3k\,|\langle f|p|i\rangle|^2\delta(E_f-E_i-\hbar\omega)f(E_i)[1-f(E_f)] \qquad \text{Eq. (1)}$$

where the photon operator has been approximated by the first term in the series expansion of the vector potential A, the integral is over $\vec{k}$-space, and the sum is over all pairs of initial and final states, $|i\rangle$ and $|f\rangle$. The delta function assures conservation of energy, and only direct or k-conserving transitions are allowed. It is also required that the initial state is occupied and the final state is empty, consistent with the Fermi functions in Eq.(1). Derivation of Eq. (1) can be found in many intermediate discussions of the photoemission process and follows directly from Fermi's Golden Rule of time dependent perturbation theory,[15] starting with the Hamiltonian of a particle interacting with an electromagnetic field,

$$H=H_0+\frac{e}{2mc}\sum_j(\vec{p}_j\cdot\vec{A}_j+\vec{A}_j\cdot\vec{p}_j) \qquad \text{Eq. (2)}$$

with $\vec{p}$ and $\vec{A}$ being the momentum and the vector potential, respectively.

The photoemission process is sketched in Fig. 3. The inset shows the experimental arrangement with a photon incident upon a sample and an electron being ejected. Photoabsorption induces excitation from $|i\rangle$ to $|f\rangle$, drawn here without dependence on the matrix elements of Eq.(1). The distribution of photoelectrons measured in vacuum is given on the right panel. Structure in the experimental energy distribution curve (EDC) is due to primary electrons and the smoothly-varying background results from inelastically scattered secondary electrons. The energy extremes for the EDC correspond to the work function and the Fermi cutoff. Not shown are contributions from such processes as multi-electron events, characteristic loss satellites, or plasmon losses.[11,13,14]

To use synchrotron radiation photoemission fully requires a brief review of the experimental parameters (see Refs. 16-18 for greater details). The most obvious variable is the photon energy. Photoemission requires that the energy of the excited electron be sufficient that it can escape the solid such that $h\nu>\phi$ where ϕ is the work function. Further, an optimal measurement involves a full scan of the valence band region, which may be

10-15 eV wide. Hence, the minimum source is a gas discharge and the most frequently used gases are hydrogen (continuum below ~10 eV), helium (lines at 21.22 and 40.82 eV), and neon (16.8 and 26.9 eV). Their spectral output makes it possible to perform valence band studies and the acronym UPS for ultraviolet photoemission spectroscopy was coined. In contrast, the use of characteristic x-ray emission lines makes it possible to study core levels and the term XPS for x-ray photoemission was introduced (most commonly used sources are Mg and Al with k_α radiation at 1253.6 and 1486.6 eV). Although both discharge sources and x-ray sources are extremely useful for laboratory work, they suffer from their limited spectral range. Today, with synchrotron radiation sources available that span the range 10-1500 eV, the distinction between UPS and XPS is less important.

The optimal light source for photoemission (and many other photon techniques as well) is synchrotron radiation.[16-18] The advantages are many - indeed, the only disadvantage is that not all laboratories have their own synchrotrons and users must commute to the national light sources. The availability of such light sources is increasing in the US and abroad and such travel is now commonplace, if inconvenient.

An analogy that is sometimes used when synchrotron radiation photoemission is compared to line source photoemission is that of a motion picture recording versus a still picture recording of an athletic event. Attempts to understand the driving forces or the object of interest in a football game based on a few photographs taken at random are not likely to be successful. The likelihood that sense can be made out of apparent chaos is significantly improved through motion pictures (a continuum of stills). The odds are improved still further with high speed film, multiple angles, and object-specific cameras. When complemented by other sensing devices, it became possible to model the game and discern order, if football ever has any.

Synchrotron radiation is the electromagnetic radiation emitted by charged particles undergoing acceleration in instantaneously-curved paths.[12,17-20] Electrons (or positrons) are the particles of choice. Radiation is emitted when they traverse magnetic dipole fields oriented with field vector B perpendicular to their instantaneous velocity vector $\vec{v}$.[21] If the velocity is low, the radiation field has a dipole pattern, as shown in Fig. 4. At velocities approaching the speed of light, however, the radiation is concentrated in the forward direction by the relativistic transformation from the moving to the laboratory reference frame. As the particles sweep through the arc of the bending magnet they emit radiation tangent to their orbit. These emitted photons are confined to a relatively narrow angle close to the orbital plane; this opening angle depends on both the electron velocity and the photon energy and is smallest for high energies. The polarization of the beam is nearly 100% for photons emitted at the critical energy of the machine and nearly complete through the soft-x-ray and vacuum ultraviolet range (the $\vec{E}$ vector lies in the orbital plane). It then follows that optical systems for synchrotron radiation research have different constraints than those in the laboratory where the radiation is more isotropic and unpolarized.

The spectral distribution of synchrotron radiation is of the greatest interest for photoemission characterization of surfaces and interfaces. As shown in Fig. 5, the spectrum extends from the far infrared to the soft- or hard-x-ray range, with the short wavelength cutoff determined by the electron energy in the storage ring and the bending radius of the magnets [to convert from Å to eV, use $E(eV) = 12398/\lambda(Å)$]. Since the range $5 \leq h\nu \leq 2000$ eV makes it possible to thoroughly examine the valence bands and a great many of the core levels, there has been a push for light sources that optimize that range by operating with stored electron energies of 750-1000 MeV. Machines of higher energy are of great interest for x-ray research, but the presence of x-ray photons in the emission profile of a beamline compromises its efficiency and the ease of soft-x-ray and vuv operation.

With synchrotron radiation photoemission, it is possible to span the range of electronic interactions by selecting any photon energy for excitation. This tunability has led to the development of several photoemission techniques. Some involve the recording of energy distribution curves at a variety of photon energies with subsequent interpretations based of variations in the appearance of the initial state features (EDC mode). Others involve fixing the kinetic energy of the detected electron while sweeping photon energy (constant final states or partial yield mode[22,23]) or simultaneously sweeping photon energy and detected energy (constant initial state mode[23]). Each has advantages and since most synchrotron radiation beamline monochromators are under computer control they can be chosen at will as the goals of the experiment change. Of particular interest to us here are techniques which

1. Use of the variation in cross section of the initial state. In an atomic calculation of energy states, account must be taken of the $\ell(\ell+1)/r^2$ centrifugal barrier term in the Hamiltonian.[24] Since states of higher angular momentum have increasingly important centrifugal barrier terms, they are better seen in EDCs at high photon energy.[25] In particular, excitation of these states to the continuum (coupling of initial and final states) requires that the centrifugal barrier be overcome (delayed onset). Many studies use this photoionization cross section variation technique to identify the ℓ-character of initial states, though it should be clear that the ℓ-character of band states can be mixed through hybridization and is rarely pure.

2. Use of resonance phenomena for photoemission. These resonances result from quantum mechanical interference between equivalent paths leading from the ground state to the final state.[26] Using the Ce 4d-4f resonance as an example, we can see competition between Eqs. 3 and 4 and between Eqs. 5 and 6 where

$$4d^{10}4f^{n}(5d6s)^{3} + h\nu \rightarrow 4d^{10}4f^{n-1}(5d6s)^{3} + e \text{ (direct excitation)} \quad (3)$$

1

$$4d^{10}4f^{n}(5d6s)^{3} + h\nu \rightarrow 4d^{9}4f^{n+1}(5d6s)^{3} \rightarrow 4d^{10}4f^{n-1}(5d6s)^{3} + e \quad (4)$$

$$4d^{10}4f^{n}(5d6s)^{3} + h\nu \rightarrow 4d^{10}4f^{n}(5d6s)^{2} + e \text{ (direct excitation)} \quad (5)$$

$$4d^{10}4f^{n}(5d6s)^{3} + h\nu \rightarrow 4d^{9}4f^{n+1}(5d6s)^{3} \rightarrow 4d^{10}4f^{n}(5d6s)^{2} + e \quad (6)$$

Analogous expressions could be written for excitation from the closed p shells of the transition metals (np-nd where the d's constitute the valence states). As a result, features of d or f character can be suppressed or enhanced by scanning through photon energies corresponding the p or d core levels. The technique has been used for interface research by Weaver and coworkers[27] and Rossi et al.[28] exploiting p-d and d-f resonances in the photon energy range below ~150 eV.

3. Use of the Cooper minimum technique. The nodal character of the initial state wavefunctions makes it possible to vary the coupling to final states, analogously to what was observed in atomic spectroscopy by Cooper.[29] It has been used quite extensively by the Stanford/Milan group of Spicer/Lindau/Braicovich[30] to examine valence band features in interfaces involving 4d (one node) and 5d (two nodes) transition metal overlayers.

In studies of evolving interfaces where there is the need to examine the spatial distribution of different atomic species, the ability to vary the photon energy is extremely important because the scattering length of a photoelectron depends on its kinetic energy. Escape depth variation has been used effectively in a large number of interface experiments where core level intensities have been followed as a function of overlayer thickness.[31] Electron mean free paths can be maximized by selecting photon energies to give electron kinetic energies of $\lesssim 10$ eV or $\gtrsim 1000$ eV. In contrast, the mean free path goes through a minimum for electrons of ~50-100 eV energy. As shown by the "universal curve" of Fig. 6, escape depths are ~10-20 Å for low energy electrons, ~3-8 Å for minimum values, and then significantly higher for electrons of several hundred electron volts.[32]

A technique which takes advantage of the limited mean free path of the excited electron to nondestructively obtain atom depth profiles in the surface region involves the measurement of the core level emission as a function of emission angle. By varying the detection angle for the electrons from normal emission to grazing emission, $\Theta=0\rightarrow 90^\circ$, the thickness of the material that the electron transits increases as $(d/\cos\Theta)$ for an absorption site a distance d below the surface. Grazing incidence thereby increases surface sensitivity. This technique has recently been used in synchrotron radiation photoemission,[33] although it works equally well for XPS[34] or Auger electrons.[35]

Many elegant techniques have been developed for mapping the energy bands of solids,[36-40] i.e. determining the dependence of the electron energy on crystal momentum, E(k). What is needed is an oriented single crystal so that information obtained by varying the electron emission angle can be related to the initial state in the solid through conservation of crystal momentum parallel to the surface $k_\parallel$. These techniques also require a reasonable approximation to the band structure as a starting point when the bands are complex. For reacted interfaces which exhibit atomic intermixing without long range order, angle-resolved photoemission offers no significant advantage over angle-integrated techniques. For ordered overlayers, however, it is of greater importance. Since our use of angle resolved results will be limited to normal emission studies $(k_\parallel=0)$, we will not discuss the techniques of angle-resolved photoemission in any detail. Instead, the interested reader is referred to the excellent reviews of Refs. 36-40 and the papers cited therein.

CASE STUDIES

In previous sections, we defined the stages of interface growth and we indicated that synchrotron radiation photoemission has many properties which make it a nearly ideal tool for studying those interfaces. In the following, we reiterate the pressing questions of interface science and, through specific examples, examine the means by which synchrotron radiation photoemission can address them.

Ordered Chemisorption: Cl/Si

Photoemission is an excellent way to probe the details of the atomic interactions at a surface. Although it generally does not try to determine whether or not an overlayer is ordered without support from techniques like LEED and RHEED, it can critically address the validity of calculations of surface bondings and the treatment of electronic structure. (Even by itself, the technique of photoelectron diffraction is able to examine chemisorption geometries.[41]) We will examine the case of Cl chemisorption on Si and Ge to demonstrate two of the strengths of photoemission for characterization of an ordered overlayer, namely hν-variation and polarization dependences.

The Cl/Si system is representative of systems with well-defined ordering. It had been studied by a variety of experimental techniques, including LEED,[42] and had been modeled through elegant pseudopotential calculations.[43] Subsequently, Rowe et al.[44] performed a series of synchrotron radiation photoemission experiments of Cl chemisorption on Si(111)7x7 and 2x1 surfaces, as well as Ge(111). They found that Cl adsorption induced strong EDC structures 5-6 eV below the valence band maximum. This observation in itself was significant and offered a clue as to the nature of the chemical interactions. They also found that the binding energies of the Cl-induced states were relatively insensitive to the surface or its reconstruction. This was remarkable and appeared to contradict the LEED results which showed significant variation in overlayer/surface structure which resulted from bonding in the different surfaces, including formation of the primitive 1x1 pattern for Ge(111) and Si(111) but the 2x1 pattern for Si(100)2x1.

From simple electronegativity considerations, it might be expected that Cl would bond to three substrate atoms rather than one on the (111) surface, thus preferring the 3-fold site to the 1-fold site as shown in Fig. 7.[44] When attempts were made to calculate the differences in electronic structure for the two chemisorption geometries, however, it was found that both had similar overall densities of states.[43] Hence, it appeared that photoemission could not discriminate between the geometries, although LEED could.

The calculations for Cl chemisorbed onto Si had determined the binding energies of the Cl-induced states and had also determined their orbital character. For Cl/Si, it was found that the p_x, p_y, and p_z orbitals were almost degenerate for the 3-fold site but that the p_z states were well removed from the others for the 1-fold site. Hence, although the basic features of the DOS were similar, the character of the binding states varied. Rowe et al.[44] then conducted an experiment in which the electric field vector of the highly polarized photon beam was alternately parallel and almost perpendicular to the sample surface, i.e. s- and p-polarized photoemission. The result was that the states of p_z symmetry were suppressed in s polarization and were enhanced in p polarization because of dipole selection rules - these p_z states could then be identified as falling well below p_x and p_y states as shown in Fig. 7. Much weaker variation was observed for the Si(111)7x7 or Si(100)2x1 surfaces. Hence, Rowe et al.[44] concluded that the Si(111)2x1 surface stabilized the 1-fold covalent site for Cl, but that the chemisorption geometry was mixed for other surfaces - and showed that photoemission was able to examine the wavefunction symmetry of the bonding states.

Abrupt Interfaces: Ag overlayers and Ge/Ta(110)

Photoemission is ideally suited to examine the strength of adatom/substrate and adatom/adatom bonding. Indeed, depending on the strength of the substrate interaction, one can expect major changes in the valence band states or, in the limit, minimal changes beyond superposition of adatom and substrate states. Examples of systems which form abrupt interfaces are Ag/Si(111),[45] Ag/Cu(001)[46] Ag/Pd(100),[47] and Ge/Ta(110).[48] These examples allow us to highlight two more important properties of synchrotron radiation as a probe of surfaces and interfaces. First, the tunability of the photon energy makes it possible to select energies which emphasize either surface or bulk features and, second, the source intensity is great enough that high resolution studies can be performed.

Tobin et al.[46] used high resolution photoemission to characterize epitaxial growth of Ag on Cu(001). Capehart and coworkers[47] considered the Ag/Pd(100) interface and followed its evolution through the formation of a thick overlayer. They reported an abrupt interface with Ag epitaxy having a 5% lattice compression relative to bulk Ag. Both Tobin et al. and Capehart et al. showed the transition from two dimensional to three dimensional order. Two dimensional order at one monolayer was characterized by structure in the Ag normal emission photoemission spectra which had no dependence on the component of wave-vector perpendicular to the surface (a two-dimensional system should produce flat bands since $k_{\parallel}=0$). Upon formation of a second monolayer, however, a second initial state feature was observed. As shown in Fig. 8 for Ag/Pd(100), these two features had quite different photon energy dependences. Analysis showed the dependence to be related to the spatial variation of the wavevector in the z-direction. Since the final state momentum changed with hν, it was possible with synchrotron radiation to vary their cross sections by matching the initial and final state momentum. The number of initial states appeared to track the number of layers. For coverage greater than ~5 ML, the eigenstates were no longer distinguishable, converging to those of bulk Ag by 12 layers. Capehart et al.[47] also reported a blurring of the LEED pattern for Ag coverages of 8-16 layers which they interpreted as due to misfit dislocations generated by the 5% lattice mismatch of Ag(100) and Pd(100). Wachs, Miller and Chiang[45] reported angle-resolved photoemission studies for Ag(111) films grown on Si(111) that exhibited bulk electronic properties by ~7 ML. They also observed surface states and judged their 7 ML film to have long range order as perfect as that of bulk single crystals.[48]

Another example of a system which forms an abrupt, non-intermixing interface is Ge/Ta(110).[49] Both Ge and Ta have intense core emission and the intensity variations of these 3d and 4f cores can be followed as a function of coverage - with the result that interface abruptness can be modeled. Synchrotron radiation photoemission studies by Ruckman et al.[49] showed that the interface exhibited no intermixing for any coverage at 300°K and LEED results demonstrated an overlayer without long range order. In Fig. 9, we show the results of analysis of the Ta 4f core emission and the Ge 3d core emission as a function of thickness for the range 0-30 Å. The logarithm of the normalized Ta 4f core emission, $\ell n(I/I_o)$, shows linear behavior with characteristic scattering length of ~5 Å. Likewise, a plot of the growth of the Ge 3d emission shows a linear behavior when plotted in the form $[1-\ln(I/I_o)]$. These results are representative of the behavior to be expected for a uniform overlayer with no outdiffusion/indiffusion or island formation.[5] Furthermore, the continuously-shifting Ge 3d core level was indicative of an evolving overlayer in which the change in electronic configuration of Ge was slight, being reflected by a total core shift of only 0.3 eV for Θ=0.5 Å to Θ=42 Å.

Ruckman et al.[49] also examined the effect of temperature on the Ge/Ta junction. They found that Ta outdiffused into the Ge overlayer when heated, as shown by the reappearance of the Ta 4f core emission in the EDC's for 42 Å Ge overlayers on Ta. Furthermore, evidence for Ta/Ge intermixing and charge redistribution was found in the 0.7 eV chemical shift of the Ta 4f levels. Analysis of the relative amounts of Ge and Ta atoms suggested a stoichiometry of $TaGe_2$ for the interface reaction product.

Cluster Formation and Weak Interaction: Ag/GaAs and Ag/Ge

For the discussion here, we define a cluster as a three dimensional aggregate of atoms. It is then distinct from a monolayer patch of atoms, but the distinction fades when two-layer patches (rafts) are considered, and it may be academic to distinguish between a three-layer raft and a cluster. These aggregates or clusters are likely to form when adatom bonding to the surface is weaker than to other adatoms (see Fig. 1). Photoemission studies of clusters have shown a number of properties which facilitate their identification. First, it has consistently been observed that the measured binding energy of the valence band and core states is larger for clusters than for solids, e.g. the d states of Ag.[50-53] Second, the valence bandwidth is narrower for clusters than for bulk solids, corresponding to a measured reduction in nearest neighbor distances.[54] Third, in the limit of large clusters, the atomic orbital energy states broaden and overlap, converging to those of a band structure, although estimates for this convergence depend on the techniques for calculation and the s- or d-character of the solid.[55] Nonetheless, as the number of states per unit energy increases, the likelihood of observation of a distinct Fermi level cutoff increases. Results by Ludeke, Chiang and coworkers[52,53] have correlated the appearance of emission near E_F with the width of the d-bands of Ag showing, for example, bulk-like d-bands when the metallic Fermi cutoff is evident.

Zunger[56,57] recently considered cluster formation from the standpoint of total energy calculations for the cluster. He noted that one must consider the condensation energy of the adatoms, lateral diffusion across the surface, and the energy of clustering. In essence, he showed that clusters would be unlikely to form unless the process of removing an atom from a chemisorption site and adding it to an aggregate of other like atoms represented an energy gain.

Clear evidence for Ag nucleation to form 3-dimensional clusters which have negligible interaction with GaAs and Ge substrates has been reported by Ludeke et al.[52] and Miller et al.[53] Those authors reported high resolution core level studies coupled with an evaluation of the interface valence bands. By taking advantage of the tunable surface sensitity of synchrotron radiation photoemission, they examined the attenuation of the surface shifted core level, i.e. the component in the core EDC which is due to surface atoms having different configuration from the bulk (Fig. 10). The attenuation of this signal represented a sensitive measure of the extent to which the surface environment changes. For coverages of 5Å of Ag on Ge(100) they found that the surface-shifted component persisted and there was no evidence of changes in binding energy associated with Ag chemical bonding to the substrate. Valence band studies also showed that Ag nucleation occurred at low coverage and metallic clusters formed. From this they concluded that Ag was not homogeneously distributed on the surface. From LEED they argued that the clusters grew in crystalline form with small aspect ratio (height to diameter) and a Ag(110) surface.

Silver was also observed to form 3d islands for the GaAs(110) surface at room temperature. For that system Ludeke et al.[5,52] followed the behavior of the surface-shifted components to high coverage and demonstrated island formation. They subsequently modeled the formation of clusters on a surface and predicted the attenuation rate of the substrate as a function of aspect ratio of the clusters. As shown in Fig. 10, they predicted clear departure from a simple exponential behavior until coverages corresponding to the overlapping of clusters (complete surface coverage). In a recent review,[5] Ludeke expanded on this discussion and predicted that cluster formation is common event on semiconductor surfaces. As we will discuss in the following paragraph, however, the observation of clusters of small size on reactive interfaces is difficult to observe.

Cluster-Induced Reactions: Ce/Si(111)

The paragraphs above considered Ag cluster formation and growth. Those clusters were readily observed because the interface was nonreactive and they became quite large. Indeed, the study of the attenuation of the surface component by the gradually expanding clusters was an elegant use of synchrotron radiation photoemission.[5,52,53] Most systems are not that cooperative, however, and the range over which clusters might be important can be small, particularly when they induce reaction.

Much of the photoemission work with metallic clusters has involved deposition and growth on inert substrates, generally amorphous carbon.[51] These results intentionally minimized substrate interaction of the sort which might be expected for active metals on semiconductors. A study which seeks to discern clusters of three dimensional character on a semiconductor surface must then consider substrate interaction. Zunger[56,57] pointed out that this interaction varies and that clusters may become unstable as they grow. Upon reaching the instability limit, a disrupting cluster may provide the energy needed to induce chemical reaction with the substrate or pin the

Fermi level via defect formation. Zunger very convincingly considered the case of Al/GaAs.

Two recent synchrotron radiation photoemission experiments have reported the existence of clusters. Daniels et al.[58] investigated the Al/GaAs(110) system and reported chemisorption states for Al at ultralow coverage (<0.1 ML) which were consistent with island formation. Grioni et al.[59-60] examined the Ce/Si(111) system with synchrotron radiation photoemission, angle-resolved Auger spectroscopy, and LEED and unambiguously showed Ce cluster formation for Ce coverages of 0.1-0.6 ML. Ce/Si offered a special opportunity to demonstrate the formation of clusters and their role in triggering reaction because of the readily observed 4f emission features and the large core level shifts of Si which accompany intermixing.

In Fig. 11 we show the low-coverage photoemission results for Ce/Si. The EDCs on the left show the coverage dependence of the valence bands. By comparing the results for hν = 60 and 35 eV, it is possible to distinguish the states of f character, as discussed above, using the variation of the 4f photoionization cross section with photon energy. Difference curves created by subtracting the spectra on the left highlight the changing 4f character. As shown, the dominant 4f features appear well-separated from the Fermi level at low coverage and a weaker 4f structure, which occurs near E_F for bulk Ce, is absent. With increasing coverage, the binding energy of this structure does not change and the width of the Ce-derived d- and f-features also remained constant to within experimental uncertainty. Further, there was no detectable Fermi level for the Ce clusters at coverages less than 0.6 ML. The absence of a metallic Fermi cutoff suggested that the clusters were small because its presence would be expected at smaller cluster size for open-shelled d-band metals than for the noble metals - where 100 atom clusters have a metallic Fermi edge.[51] Finally, LEED I-V results showed the persistence of substrate 1x1 pattern and the angle-resolved Auger studies showed diffraction modulation characteristic of the uncovered Si(111) surface beyond monolayer coverage, indicative of a heterogeneous surface layer. At a critical coverage of 0.6 ML, disruptive Ce/Si intermixing occurred and a silicide was formed, as will be discussed at the end of the next section on atomic intermixing. At that point, the Ce 4f shifted in binding energy and appeared at the same binding energy as in $CeSi_2$. Likewise, a reacted Si-2p core feature was observed to grow, but with a growth/attenuation profile which demonstrated the heterogenous character of the Ce/Si surface.[59-60]

The results for Ce/Si point to the difficulties associated with demonstrating cluster formation, including the need for several analytical tools (LEED and angle-resolved Auger proved indispensable for the Ce/Si study), the limited coverage range over which clusters persisted, and the need for high quality data for results at very low coverage. The intensity and tunability of the synchrotron radiation source made the experiments possible.

Atomic Intermixing at Interfaces: Pd/Si and Cr/GaAs

Atomic intermixing lies at the very heart of most interface studies, particularly those involving metal overlayers on semiconductors, and there are a large number of interesting papers in the literature. (Brillson's review[1] includes 1050 citations.) Of interest is the spatial distribution of atoms and their chemical bonding across the interface region. One must consider the character of the developing intermixed system and determine whether a single compound forms, whether that compound is identical to one found in a bulk phase diagram and is thermodynamically stable, or whether a disordered system lacking distinct stoichiometry is prevalent. Underlying this, of course, are questions relating to the nature of the chemical bonds, the structural properties of the interface, and its electrical properties. We shall address these issues in the following paragraphs through examples which show how photoemission studies can offer valuable insights.

Pd/Si. Several important aspects of synchrotron radiation photoemission can be reviewed by examining the representative near-noble-metal/silicon interface Pd/Si, among them the importance of a tunable photon source, the application of p-d resonance photoemission, the Cooper minimum technique, and the important tie to theory.

In Fig. 12 we show the results of Rubloff et al.[4,61] for Pd overlayers on Si(111) for coverages 0-32 Å. The spectrum at the bottom of panel (a) is for clean Si(111). As can be seen, the overall emission increases with Pd coverage because of the higher photoionization cross section of the Pd-derived states relative to Si. The incremental difference curves for low coverages in panel (b) show the growth of Pd d-derived emission about 3.5 eV below E_F and weak emission nearer the Fermi level. These results demonstrate that reaction occurs at the lowest coverages and that there is no coverage threshold for reaction. As the coverage increases, Pd d-derived states move toward E_F and ultimately stabilize at -2.75 eV. From TEM measurements it has been shown[62] that the Pd_2Si lattice structure occurs for Pd coverages of $\gtrsim$5Å. Rubloff et al.[61-63] then concluded that the reacted region had a thickness of several unit cells for room temperature reactions and the boundary layer was abrupt on an atomic scale (~3 Å). A number of other experimental studies[64,65] confirmed that the Pd/Si interface configuration is the same as that of bulk Pd_2Si. SEXAFS[65] results showed that for a 1.5 ML film the Pd first and second nearest neighbor distances were the same as those of Pd_2Si to within 0.02 Å. Thick films of Pd_2Si can then be prepared on Si by increasing the temperature and hence the mass transport. (For a detailed discussion of the Pd/Si system, see the recent review by Rubloff.[4])

Theoretical insight into the character of the states observed in Fig. 12 for the reactive Pd/Si interface has been provided by several authors,[63,66-70] starting with Ho et al.[63] All show that there are hybrid Pd-d/Si-p bonds which constitute the dominant charge mixing for the silicide. Although both bonding and antibonding Pd-Si states develop, the antibonding states are largely empty. Significantly, a large number of d-states remain nonbonding with respect to Si. Comparison of photoemission results taken with synchrotron radiation in the photon energy range below about 180 eV[4,61,71-76] and those taken with x-ray energies[77-79] show that the lower energy results emphasize the nonbonding states ~2.5-3 eV below E_F whereas the higher energy results demonstrate more clearly the bonding and antibonding states.

An important issue in the characterization of a reacted surface region is whether there is surface segregation of one of the species. For the Pd/Si system, there is now consensus that Si segregates to the surface.[4,61,64,74-76,80] Synchrotron radiation photoemission studies which have added support to this conclusion come from comparisons of bulk samples with reacted samples[75,76] and examinations of the effect of Ar^+ sputtering of a thick reacted layer. Franciosi and Weaver[75] performed synchrotron radiation photoemission studies of fractured bulk samples over the spectral range 16-130 eV. The valence band results are shown in Fig. 13. As can be seen, feature "B" at 2.5 eV is dominant at all photon energies. Comparison to the interface results shows that "B" corresponds to what was identified in Fig. 12 as due to the Pd-derived nonbonding d states but that the binding energy is ~0.3 eV lower. The difference can be related to Si enrichment at the surface, as will be shown in the following paragraphs.

It is clear from Fig. 13 that the relative photoionization cross sections of initial state features A-E vary a great deal with photon energy. Two spectral ranges are important for us here in the context of the uses of synchrotron radiation and both highlight particular valence band states. Studies in the energy range 120-170 eV correspond to the Cooper minimum regime, as discussed by Abbati et al.[71] and Miller et al.[72] By analogy to atomic cross sections, these authors argued that atomic d-f transitions should go through a minimum for f states with kinetic energies in this range, corresponding to a matching of the wavefunction nodes for 4d and 4f states. The results of their work contributed to the identification of the origin of the initial states of Pd_2Si and the character of the chemical bond. Similarly, the p-d resonance photoemission results[75] in the energy range 40-70 eV suppressed and then enhanced the Pd-derived d states. As shown in Fig. 13, the result was that the relative cross sections of peaks C (p-d hybrid) and B (d nonbonding) changed. This hν-variation can be understood in terms of coupling of the Pd 4p core hole with the Pd 4d valence states. For states of increasingly hybrid character, the 4p-4d overlap is reduced relative to atomic-like 4d states.

The synchrotron radiation photoemission experiments which showed Si enrichment on the Pd/Si reacted surface involved comparisons of bulk Pd_2Si and Pd overlayers on Si reacted at temperatures of 200 and 700°C. The results shown at the top of Fig. 14 reveal emission for reacted Pd/Si samples in the regions indicated by the tic marks which is not present for cleaved Pd_2Si.[75-76] Further, the dominant Pd_2Si bulk peak at 2.5 eV was broadened and shifted, consistent with the results of Rubloff et al.[4,61] To demonstrate that the differences were related to surface segregation, the reacted interfaces were sputtered with low energy Ar ions. Comparison with results for cleaved Pd_2Si (Fig. 13) showed that the extra emission vanished and agreement with the bulk sample was then very good. Two cautionary comments should then be made. First, synchrotron radiation with its high surface sensitivity may encounter pitfalls associated with surface effects at reacting interfaces. Second, comparisons of bulk samples and reacted interfaces are of great value in detecting these artifacts.

Since one of the goals of the interface science is to characterize the reaction products, it is important to compare the results to first principles calculations of the electronic structure. Unfortunately, such calculations are hampered by the perversity of many of the interface compounds themselves. In particular, many exhibit complex crystal structures with low symmetry or large numbers of atoms per unit cell.[81] Nevertheless, theoretical guides can come from calculations based on the extended Huckel model[66-69,82,83] (which adequatedly treats the symmetry of the crystal but lacks some of the elegance of self consistent calculations) or the Augmented Spherical Wave method[63,84,85] (which is elegant but is presently limited to high symmetry crystal structures).

In Fig. 15 we compare synchrotron radiation photoemission results for the 3d transition metal disilicides with calculated schematic densities of states for those materials.[85] The calculations from which the DOS's on the right of Fig. 15 were derived revealed the metal-d and the Si-p character for the model compounds MSi_3, MSi, and M_3Si, where M denotes the metal. Based on those calculations, the energy location and width of the metal d- and Si p-bands was estimated - without attention to details but with particular interest in the energy location of the lowest d character (solid DOS line). What the results of Fig. 15 show is that there are always metal d states extending ~6 eV below the Fermi level and that they represent low-lying hybrid metal-d/Si-p bonding states. Most of the d states, on the other hand, are nonbonding with respect to Si. For Ca silicides, the nonbonding d states fall well above E_F but they sharpen and are drawn below E_F as the d occupancy increases across the transition metal row. By midrow, the Fermi level falls in the middle of this d manifold. For the noble metal silicides, the d character is well below E_F. It should also be noted that the energy location of the Si p states is relatively invariant for the 3d silicide series (dashed lines of Fig. 15).

Comparison of the schematic density of states results of Fig. 15 with synchrotron radiation photoemission data confirms the tendency of photoemission to emphasize d states and, even more, to emphasize the nonbonding d states. The bonding states 4-6 eV below E_F are less well observed in the experimetal results, although they do rise above the secondary electron background. For VSi_2, they can be clearly seen, and analysis of those results[84] shows the ℓ-character of the states responsible for the emission features.

Although comparison of experiment with theory is extremely important in interface science, it is not a trivial matter. As indicated above, phases with complex structure can form and, worse yet, there can be more than a single phase which forms. Likewise, there may be surface segregation or stoichiometry gradients. This makes a careful characterization of the evolving phase(s) all the more important through core level lineshape analysis and valence band studies, combined with fingerprinting based on bulk compounds and results from complementary studies.

Cr/GaAs. The Cr/GaAs[86] system is a second example which highlights several uses of synchrotron radiation photoemission for studies of interacting interfaces. Since there are three atomic components in this system, one might expect that Cr/GaAs might be more complicated than Pd/Si. As we will show, however, the results from synchrotron radiation photoemission and LEED studies are quite adequate for interface modeling.

In Fig. 16 we show valence band EDCs for Cr/GaAs(110) taken at a photon energy of 30 eV. At low coverage, the results can be described as a superposition of Cr-induced states and GaAs substrate emission such that difference curves show pronounced growth of emission within ~2 eV of E_F and attenuation of GaAs states. This regime is termed "weakly interacting" because no modification of the electronic structure of the substrate can be identified. At higher coverage, however, the valence bands indicate an intermixed or reacted region. Above ~20 Å coverage, the valence bands gradually converge to those of bulk Cr, indicative of the formation of a Cr overlayer. These three stages in the development of the Cr/GaAs interface are indicated at the right of Fig. 16.

The results of core level studies for the Cr/GaAs system make it possible to be more quantitative in identifying the onset of reaction and the extent of the reacted region. In Fig. 17 we show the core levels for Ga (left) and As (right) as a function of overlayer thickness, exploiting the tunability of synchrotron radiation to maximize surface sensitivity. These results show Fermi level pinning by ~1 Å coverages but the delay of reactive intermixing until ~2 Å (the spectra shown are corrected for band bending, a shift of 700 meV induced by band bending is indicated by the horizontal bar shown in Fig. 17). The onset of reaction is reflected by the appearance of a small shoulder on the low binding energy side of the Ga 3d core line. With coverage, the shoulder grows and its centroid moves steadily to lower binding energy. By 14 Å, no substrate emission is visible and the Ga cores have stabilized with a total shift of 1.25 eV, i.e. a greater shift relative to Ga in GaAs than observed for Ga droplets.[87] The core level results for As 3d core line show a more complicated behavior. At the onset of reaction, the single As peak (tic mark in Fig. 17) gives way to three peaks - one shifted 0.25 eV to greater binding energy and one shifted 0.40 eV to lower binding energy. Both shifted features are clearly visible as the Cr coverage increases, although they are ultimately lost when the overlapping Cr 3p core becomes dominant. (Analogous features have been followed to much higher coverage for the V/GaAs and Ti/GaAs systems). Weaver et al.[86] interpreted these Ga and As core level results in terms of the formation of (1) an intermixed Cr-Ga configuration with variable stoichiometry, (2) a stable Cr-As local configuration with distinct binding energy, (3) and surface segregated As coordinated primarily by metal atoms.

Interface Modeling. Studies of the attenuation of core level photoemission during reactive intermixing makes it possible to identify the moving species and examine the abruptness of the interface, as was discussed above for the Ge/Ta(110) example. In Fig. 18 we show the attenuation of the Ga core for the Cr/GaAs[86] system together with results for As in Ce/GaAs[88,89] and Si in Ce/Si[59,60] (see discussion of Fig. 11 above for an evaluation of the Ce/Si interface at low coverage where clustering was observed). The topmost curve in each panel gives a measure of the total attenuation for Ga, As, or Si atoms as a function of coverage. What is remarkable about the results of Fig. 18 is that it was also possible to identify the variation with coverage of each of the species resulting from reaction at the interface.[88] For Ce/Si, it can be seen that three Si species are present corresponding to substrate Si, reacted Si, and surface-segregated Si.[59,60] For Ce/GaAs, there are four As species corresponding to the substrate, two reacted species, and surface-segregated As.[89] For Cr/Ga, only two "states" were identified corresponding to substrate and reacted Ga since Ga did not stabilize a well-defined bonding configuration but exhibited a variable binding energy.[86] Clearly, such information makes possible detailed interface modeling according to the chemical environments and concentrations of the different species. Not surprisingly, this work requires synchrotron radiation so that the photon energies can be related, the surface sensitivity can be varied, and the necessary experimental resolution attained.

The results shown in Fig. 18 also address issues related to disorder at interfaces and the chemical species which form in ways which cannot be done with other techniques. In particular, by examining the core lineshapes[88] shown in Fig. 19 it is possible to determine whether the species are present in varying environments or stable environments. Further, the degree of disorder within each environment can be assessed by considering the full-width-at-half-maximum of the various core emission components. Since the interfacial reaction cannot, in general, be expected to stabilize extended perfect crystals at room temperature, one should expect not-quite-identical configurations for the atoms involved in those compounds, i.e. lineshape broadening.

Finally, the results of Figs. 18 and 19 show that interface fingerprinting is possible with high resolution synchrotron radiation. This can be seen, for example, from Fig. 19 where the experimental core lineshapes for a wide range of overlayer thicknesses were fit with well-defined or distinct components corresponding to substrate, reacted, and segregated species. For those systems the success of the fitting is clear evidence that local effects dictate the properties of the reaction product, even for very low coverage, and the local chemical environment is sufficient to stabilize distinct compounds on a microscopic scale. Subsequent analysis of the growth/attenuation of the respective features allows a detailed modeling of the interface.[60,88]

Atomic Mobility and Diffusion Barriers: Au/Al/GaAs and Au/Cr/Si

Synchrotron radiation photoemission can be used to great advantage to determine the atomic distribution across an interface within some depth corresponding to the probe depth (roughly 3-4 times the mean free path of the photoelectron). Of interest is the spatial profile of each specie. Brillson and coworkers[1,90-95] have recently used soft x-ray synchrotron radiation to great advantage in a number of pioneering studies. By considering the behavior of diverse metal overlayers on III-V and II-VI semiconductors, they correlated semiconductor constituent diffusivity to the strength of the metal-cation and metal-anion interface bonds and found a relationship between interface width and the heat of reaction between metal and semiconductor atoms.[90,91] For the III-V's, they argued that anion outdiffusion should be reduced when strong metal-anion bonds are formed. For less reactive metals, anion outdiffusion should be more important. The term chemical trapping was used to discuss reduced anion diffusion. This rule of thumb is followed, for example, at the Ce/GaAs interface described above.[89] For Ce/GaAs, the reaction is highly exothermic and predominantly ionic bonds are formed, producing an interfacial compound very similar to CeAs. Analysis of the relative diffusivity of Ga and As in Ce/GaAs, as measured with photoemission studies of the cores, shows that As is far less mobile than Ga.

These general statements regarding relative diffusion are valuable, but exceptions are not infrequent. For example, in the Cr/GaAs case discussed above[86] the p-d covalent bonding should place it among the class of reactive interfaces such that Ga should be mobile and As should be trapped. Photoemission results show that Ga is quite mobile, consistent with Brillson's observation[1] that Ga outdiffusion does not vary significantly for the reactive interfaces. They also show, however, that As constitutes a large fraction of the final surface, even for high coverage (~20% of the initial value for $\Theta \approx 50$ Å) and that a Cr-like overlayer starts to form only after ~20 Å coverage. Account must then be taken of the diffusivity of the different species through the reacted layer and the surface sensitivity of the experimental technique.

Brillson et al.[93-95] have also shown that synchrotron radiation photoemission is excellent for studying thin diffusion barriers, i.e. a few layers of adatoms placed between an overlayer and a substrate to reduce or enhance interdiffusion of the respective species. They recently reported a number of interesting studies with interlayers ranging in chemical activity from relatively nonreactive metals (Au, In, Zn) to highly reactive metals (Ti, Al). Their results for the Au/Al/GaAs system were particularly intriguing,[94] as shown in Fig. 20. First, for the Au/GaAs[96,97] interface they observed that both Ga and As outdiffuse into the Au overlayer and that Au indiffuses into the semiconductor. For the Al/GaAs interface, however, Al replaces Ga and Ga surface segregates.[93,98] Hence, it seems reasonable that Ga is released at the Al/GaAs interface and is free to diffuse into the Au overlayer. By increasing the thickness of the Al interlayer, they showed that the ratio of Ga to As changed markedly, ultimately reversing the trend of preferential As diffusion to the free Au surface

Brillson[1] has also shown that different interlayer materials influence diffusion according to the strength of the reaction with one or the other of the semiconductor atoms. For Ti, which reacted more strongly with As than did Al, the promotion of outdiffusion of Ga into a Au overlayer was stronger than for Al. For interlayers of nonreactive metals like In or Zn, on the other hand, there was no enhancement of outdiffusion.

In a recent interlayer study of Au/Cr/Si, Franciosi, O'Neill, and Weaver[99] showed that the amount of Si outdiffusion through a Cr interlayer into the Au overlayer can be correlated with the chemical state of Si at the Cr/Si junction. They chose the Au/Cr/Si system because Cr/Si was known to have three distinct stages of reaction[31] and because the Au/Si was well understood.[100,101] When a Cr layer of less than ~2 Å was sandwiched between Au and Si, the result was that Si outdiffusion into Au was dramatically reduced relative to a zero thickness Cr layer, as shown in Fig. 21. This was consistent with Cr/Si results[31] which showed that reactive intermixing did not occur until a critical coverage of ~2 Å. For interlayers between 2 and 9 Å, however, the amount of Si outdiffusion was greatly enhanced - because the Cr/Si interface exhibits intermixing for those thicknesses and Si outdiffusion was catalyzed by the Cr-induced disruption of the interface. At coverages greater than ~10 Å, the Cr interlayer again reduced Si diffusion into the Au overlayer because the Cr-Si room temperature reaction was complete and a Cr metal layer was forming over the Cr-Si reacted phase. Hence, the outdiffusion of Si could be enhanced or impeded according to the properties of the boundary region. These results clearly demonstrated that the details of the interlayer bonds (and hence morphology) must be understood if interlayer modeling is to be successful and diffusion control is to be predictable.

Altering the Chemical Reactivity of Interfaces: Oxidation

From the above discussions, it should be evident that reactions at interfaces can be altered. Controls for atomic diffusion can then be established by chemical means which enhance or reduce outdiffusion at the atomic level. Indeed, Nature established such surface control long ago by introducing passivating layers which protect reactive underlayers, e.g. oxides on Si and Al. Synchrotron radiation is now proving to be an effective technique for studying these reactions because it is atom specific, it can detect chemical changes in the constituents, and the surface sensitivity can be tuned.

Recent work has shown that the sensitivity of surfaces to oxygen can be varied by controlled addition of adatoms which disrupt or modify the substrate bonding. This can have a positive effect for systems which are slow to react. The technique for controlling the reaction is aptly termed interface catalysis or the catalytic effect. The results described above for Au/Cr/Si fall into that category.[99]

Intriguing possibilities can be considered when one thinks of oxidation processes catalyzed by these metal overlayers. The Auger and energy loss studies of Cros _et al._[102] examined the effect of thin layers of Au on Si for subsequent oxidation reactions. They found that SiO_4 tetrahedra were formed at room temperature, although the spatial extent of the reaction was small. By increasing the temperature to 400°C, they were able to form an extended oxide layer. The presence of Au clearly changed the kinetics of the oxidation reaction relative to the clean Si surface.

Katnani _et al._[103] used synchrotron radiation photoemission to examine the effect of a thin Al overlayer on the oxidation of Ge. They found that the oxidation rate was orders of magnitude faster than for untreated Ge. Further, they reported an oxide with different character than that found for Ge (presumably GeO_4). Abbati _et al._[104] investigated the effect of Cu, Ag, Au, and Pd on Si with respect to the oxidation of Si and also reported intriguing increases in reactivity. Franciosi[105] has recently undertaken studies of Si and GaAs oxidation catalyzed by thin Cr overlayers. Such variation in surface reactivity is not limited to semiconductor surfaces, of course. Latta and Ronway[106] reported XPS studies of catalytic oxidation of Nb by thin overlayers of Ce. Clearly, this is one of the directions that should be examined for possible application for microcircuit development and catalysis, and synchrotron radiation photoemission will doubtless play a major role.

CONCLUDING REMARKS

In this chapter we have examined many of the ways that synchrotron radiation can be used to study developing interfaces. The emphasis has been on photoemission studies - limitations in space have precluded discussions of all but a few select examples. These were chosen to represent ordered overlayers, abrupt interfaces, weakly interacting clusters, cluster-induced reactions, intermixing, interface modeling, diffusion barriers, and the control of surface reactivity. Since this is a very active research area, we can expect that the excellent work to be performed in the next few years will resolve many of the issues broached herein but left unresolved.

With the new synchrotron radiation sources and their extended spectral ranges and greater flux, it can be expected that high resolution studies will make possible better and better interface modeling by identifying the chemical state of the reacting species - the modeling studies presented here for Ce/Si represent only the beginning of such work. The studies of reactions at ultralow overlayer coverages should make it possible to better understand the triggering mechanisms for interface reaction. Recent results show that such research is possible albeit difficult because of the low count rates for ultralow coverages. At issue will be the importance of cluster formation and reactions which are triggered at coverages of less than 1-2 Å. Equally important will be studies which clearly identify the completion coverage[107] at which atomic intermixing ceases or is strongly reduced at reactive interfaces. The availability of the new light sources will also make it possible to vary the electron escape depth over a greater range than can now be done - by having tunable sources at kilovolt energies, it will be possible to study interfaces buried beneath greater thicknesses of overlayer. With the highly focussed photon beams expected with the new sources and optical systems, it should also become possible to study lateral variations on the surface, examining heterogenous interfaces parallel to the surface as well as perpendicular to it. We can also expect significant materials research dealing with interfaces in the coming years. In addition to the directions indicated above, we will doubtless see concentrated efforts for studying interlayers and using them to control diffusion and reactivity at surfaces and interfaces. Likewise, as has already been observed in the last few years, there will be intensified efforts to examine the reactivity of the transition metals and lanthanides on a greater variety of semiconductor interfaces.

It is the opinion of this author that the next decade will be the "decade of interfaces" during which many of the complex problems encountered in dynamic and multicomponent systems will be examined and better understood. As this is done, we will be in a better position to exploit the novel properties of interface-stabilized systems. This work will build on the excellent research base that now exists and on the availability of faster computers and better models for calculations of electronic phenomena.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge stimulating discussions with the researchers in the field of interface research. Particular thanks are due to O. Bisi, L. Braicovich, L.J. Brillson, T.-C. Chiang, A. Franciosi, F. Grunthaner, I. Lindau, R. Ludeke, G. Margaritondo, V.L. Moruzzi, G.W. Rubloff, K.N. Tu, W.E. Spicer, and to coworkers M. Grioni, M. del Giudice, M. Ruckman, S.A. Chambers, J.J. Joyce, D.G. O'Neill, and T.W. Capehart, to mention only a few.

References

1. L.J. Brillson, Surf. Sci. Rep. 2, 123 (1982) includes 1050 references, many of them involving studies of reacting interfaces.

2. G. LeLay, Surf. Sci. 132, 169 (1983) discusses noble-metal/elemental-semiconductor interface formation.

3. Thin Films - Interdiffusion and Reaction, ed. by J.M. Poate, K.N. Tu, and J.W. Mayer (Wiley - Interscience, NY 1978).

4. G. Rubloff, Surf. Sci. 132, 268 (1983) reviews the Pd/Si system in detail.

5. R. Ludeke, Surf. Sci. 132, 143 (1983) discusses interface formation on GaAs and examines the effects of clusters.

6. For discussions of epitaxial growth, see Epitaxial Growth, ed. by J.W. Matthews (Academic, N.Y. 1975) and E. Bauer, Applic. of Surf. Sci. 11/12, 479 (1982) and references therein.

7. For a discussion of SEXAFS, see Chapter xx of this volume.

8. M.L. Knotek, Physics Today 37, 24 (1984); see the Proceedings of the First International Workshop on Desorption Induced by Electronic Transitions, N.H. Tolk, M.M. Traum, J.C. Tully, T.E. Madey, eds. (Springer-Verlag, NY 1983).

9. See, for example, R.L. Sproull and W.A. Phillips, Modern Physics (Wiley 1980); W.A. Harrison, Electronic Structure and the Properties of Solids, (Freeman, San Francisco 1980); J.M. Ziman, Principles of the Theory of Solids (Cambridge 1964).

10. C.N. Berglund and W.E. Spicer, Phys. Rev. 136, 1030 (1964); 136, 1044 (1964); H.Y. Fan, Phys. Rev. 68, 43 (1945); W.E. Spicer, Phys. Rev. 112, 114 (1958); H. Mayer and H. Thomas, Z. Phys. 147, 149 (1959).

11. D.A. Shirley in Photoemission in Solids I, Vol. 26 of Topics in Applied Physics, ed. M. Cardona and L. Ley (Springer 1978).

12. J.D. Jackson, Classical Electrodynamics, 2nd Ed. (Wiley, NY 1975).

13. B. Feuerbacher, B. Fitton, and R.F. Willis, eds., Photoemission and the Electronic Properties of Surfaces, (Wiley Interscience, NY 1978).

14. M. Cardona and L. Ley, eds., Photoemission in Solids I and II, Vols. 26 and 27, Topics in Applied Physics, (Springer-Verlag, NY 1979).

15. L.I. Schiff, Quantum Mechanics, (McGraw Hill, NY 1955).

16. G. Margaritondo and J.H. Weaver, "Photoemission Studies of Valence States," Chapter 4 in Methods in Experimental Physics: Surfaces, M. Lagally and R.L. Park eds. (Academic Press) in press.

17. E.E. Koch, D.E. Eastman, and Y. Farge, Handbook on Synchrotron Radiation,

(North Holland, NY 1983).

18. H. Winick and S. Doniach, eds., Synchrotron Radiation Research, (Plenum, 1980).

19. N.G. Basov, Synchrotron Radiation, (Plenum, NY 1976).

20. J. Schwinger, Phys. Rev. 70, 798 (1946); 75, 1912 (1949).

21. In the new generation of synchrotron radiation sources, the bending magnets are being supplemented by wigglers which enhance the continuum radiation and undulators which produce coherent radiation. See J.E. Spencer in Ref. 18 for a discussion of these insertion devices.

22. G.J. Lapeyre, A.D. Baer, J.C. Hermanson, J. Anderson, J. Knapp, and P.L. Gobby, Solid State Commun. 15, 1601 (1974).

23. G.J. Lapeyre, A.D. Baer, J.C. Hermanson, J. Anderson, J. Knapp, and P.L. Gobby, Phys. Rev. Lett. 33, 1290 (1974).

24. S.T. Manson in Topics in Applied Physics, Vol. 26, Photoemission in Solids I, eds. M. Cardona and L. Ley (Springer-Verlag 1978).

25. The effect of cross sections on f-states (ℓ=3) can be seen through comparison of spectra taken at 30 and 60 eV for the Ce pnictides, discussed by A. Franciosi, J.H. Weaver, N. Martensson, and M. Croft, Phys. Rev. B 24, 3651 (1981) or results for CeAl-like compounds in M. Croft, J.H. Weaver, D.J. Peterman, and A. Franciosi, Phys. Rev. Lett. 46, 1104 (1981).

26. W. Lenth, F. Lutz, J. Barth, G. Kalkoffen, and C. Kunz, Phys. Rev. Lett. 41, 1185 (1978); L.J. Johansson, J.W. Allen, T. Gustafsson, I. Lindau, and S.B.M. Hagstrom, Solid State Commun. 28, 53 (1978); W. Gudat, S.F. Alvarado, and M. Campagna, Solid State Commun. 28, 943 (1978); D.J. Peterman, J.H. Weaver, and M. Croft, Phys. Rev. B 25, 553 (1982).

27. See, for example, the discussion of the Pd/Si interfaces in Ref. 75 or the VSi$_2$ results of Ref. 85.

28. G. Rossi, J. Nogami, I. Lindau, L. Braicovich, I. Abbati, U. del Pannino, and S. Nannarone, J. Vac. Sci. Technol. A1, 781 (1983); G. Rossi, J. Nogami, J.J. Yeh, and I. Lindau, J. Vac. Sci. Technol. B1, 530 (1983).

29. J.W. Cooper, Phys. Rev. 128, 681 (1962).

30. J.N. Miller, S.A. Schwarz, I. Lindau, W.E. Spicer, B. DeMichelis, I. Abbati, and L. Braicovich, J. Vac. Sci. Technol. 17, 920 (1980); G. Rossi, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, Solid State Commun. 39, 195 (1981); I. Abbati, G. Rossi, I. Lindau, and W.E. Spicer, J. Vac. Sci. Technol. 19, 636 (1981); E. Rossi, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, Surf. Sci. 112, L765 (1981); G. Rossi, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, Phys. Rev. B 25, 3619 (1982).

31. See, for example, A. Franciosi, D.J. Peterman, J.H. Weaver, and V.L. Moruzzi, Phys. Rev. B 25, 4981 (1982) and R. Ludeke, T.-C. Chiang, and D.E. Eastman,

J. Vac. Sci. Technol. 21, 599 (1982).

32. M.P. Seah and W.A. Dench, Surf. and Interface Analysis 1, 2 (1979).

33. N.G. Stoffel, M. Turowski, and G. Margaritondo, Phys. Rev. B 30, 3294 (1984) adapted the technique from C.S. Fadley, R.J. Baird, W. Siekhuns, T. Novakov and S.A.L. Bergstrom, J. Elect. Spect. & Rel. Phen. 4, 93 (1974).

34. M. Pijolat and G. Hollinger, Surf. Sci. 105, 114 (1981) discussed XPS depth profiling.

35. S.A. Chambers, T.R. Greenlee, G. Howell, and J.H. Weaver, Phys. Rev. B (1985).

36. J.A. Knapp and G.J. Lapeyre, J. Vac. Sci. Technol. 13, 757 (1976); R.J. Smith, J. Anderson, and G.J. Lapeyre, Solid State Commun. 21, 459 (1977); J. Hermanson and G.J. Lapeyre, Solid State Commun. 19, 975 (1976).

37. N.V. Smith, Chapter 6 in Photoemission in Solids II, Vol. 27 of Topics in Applied Physics, (Springer-Verlag, NY 1979); N.V. Smith and P.K. Larsen, Chapter 14 in Photoemission and the Electronic Properties of Surfaces, (Wiley, NY 1978).

38. F.J. Himpsel, Advances in Physics 32, 1 (1983).

39. N.V. Smith and F.J. Himpsel, "Photoelectron Spectroscopy," Chapter 10 in Handbook on Synchrotron Radiation, eds. E.E. Koch, D.E. Eastman, and Y. Farge (North Holland 1983).

40. E.W. Plummer, "Angle-Resolved Photoemission as a Tool for the Study of Surfaces," in Advances in Chemical Physics, eds. S. Prigodgine and S.A. Rice, (Wiley, NY 1982).

41. S.D. Kevan, D.H. Rosenblatt, D.R. Denley, B.-C. Lu, and D.A. Shirley, Phys. Rev. B 20, 4133 (1979).

42. T. Sakurai, J.E. Rowe, and H.D. Hagstrom (unpublished).

43. M. Schluter, J.E. Rowe, G. Margaritondo, K.M. Ho, and M.L. Cohen, Phys. Rev. B 37, 1632 (1976).

44. J.E. Rowe, G. Margaritondo, and S.B. Christman, Phys. Rev. B 16, 1581 (1977); M. Schluter, J.E. Rowe, G. Margaritondo, K.M. Ho, and M.L. Cohen, Phys. Rev. Lett. 37, 1632 (1976).

45. A.L. Wachs, T. Miller and T.-C. Chiang, Phys. Rev. B 29, 2286 (1984).

46. J.G. Tobin, S.W. Robey, L.E. Klebanoff, and D.A. Shirley, Phys. Rev. B 28, 6169 (1983).

47. T.W. Capehart, R. Richter, J.G. Gay, and J.R. Smith, Phys. Rev. B (in press). T.W. Capehart, D.G. O'Neill, J.J. Joyce, and J.H. Weaver, Phys. Rev. B (in press), D.G. O'Neill, J.J. Joyce, T.W. Capehart, and J.H. Weaver, J. Vac. Sci. Technol. (May/June 1985).

48. LeLay (Ref. 2) recently reviewed the Ag/Si system with particular attention

to the ordering of Ag at low coverage. See also references therein.

49. M. Ruckman, M. del Giudice, and J.H. Weaver, Phys. Rev. B (in press). For a discussion of the interesting Pd/Nb(110) system, see M. Sagurtin, M. Strongin, F. Jona, and J. Colbert, Phys. Rev. B 28, 4075 (1983).

50. Citrin and Wertheim recently reviewed the literature of clusters and raised fundamental questions about reference levels and final state effects. The reader would find their paper informative. P.H. Citrin and G.K. Wertheim, Phys. Rev. B 27, 3176 (1983). See also G.K. Wertheim, S.B. DiCenzo, and S.E. Youngquist, Phys. Rev. Lett. 51, 2300 (1983).

51. S.T. Lee, G. Apai, M.G. Mason, R. Benbow, and Z. Hurych, Phys. Rev. B 23, 505 (1981); M.G. Mason and R.C. Baetzold, J. Chem. Phys. 64, 271 (1976); M.G. Mason, L.J. Gerenser, and S.-T. Lee, Phys. Rev. Lett. 39, 288 (1977); Y. Takasu, R. Unwin, B. Tesche, and A.M. Bradshaw, Surf. Sci. 77, 219 (1978); R. Unwin and A.M. Bradshaw, Chem. Phys. Lett. 58, 58 (1978); M.G. Mason, Phys. Rev. B 27, 748 (1983);L. Oberli, R. Monot, H.J. Mathieu, O. Landolt, and J. Buttet, Surf. Sci. 106, 301 (1981); M.G. Mason, S.-T. Lee, G. Apai, R.F. Davis, D.A. Shirley, A. Franciosi, and J.H. Weaver, Phys. Rev. Lett. 47, 730 (1981).

52. R. Ludeke, T.-C. Chiang, and D.E. Eastman, J. Vac. Sci. Technol. 21, 599 (1982).

53. T. Miller, E. Rosenwinkel, and T.-C. Chiang, Phys. Rev. B 30, 570 (1984).

54. G. Apai, J.F. Hamilton, J. Stohr, and A. Thompson, Phys. Rev. Lett. 43, 185 (1979).

55. A.B. Anderson, J. Chem. Phys. 64, 4046 (1976); T. Tanabe, H. Adachi, and S. Imoto, Jpn. J. Appl. Phys. 16, 1097 (1977); A. Anderson, J. Chem. Phys. 68, 1744 (1977); C.F. Melius, J.H. Upton, and W.A. Goddard III, Solid State Commun. 28, 501 (1978); R.P. Messmer, S.K. Knudsen, K.H. Johnson, J.R. Diamond, and C.Y. Yung, Phys. Rev. B 13, 1396 (1976); P.C. Baetzold, J. Phys. Chem. 82, 738 (1978); B. Delley, D.E. Ellis, A.J. Freeman, E.J. Baerends, and D. Post, Phys. Rev. B 27, 2132 (1983).

56. A. Zunger, Phys. Rev. B 24, 4372 (1981).

57. A. Zunger, Thin Solid Films, 104, 301 (1983).

58. R.R. Daniels, A.D. Katnani, Te-Xiu Zhao, G. Margaritondo, and A. Zunger, Phys. Rev. Lett. 49, 895 (1982).

59. M. Grioni, J. Joyce, S.A. Chambers, D.G. O'Neill, M. del Giudice, and J.H. Weaver, Phys. Rev. Lett 53, 2331 (1984).

60. M. Grioni, J. Joyce, M. del Giudice, D.G. O'Neill, and J.H. Weaver, Phys. Rev. B 30, 7370 (1984).

61. G.W. Rubloff, P.S. Ho, J.L. Freeouf, and J.E. Lewis, Phys. Rev. B 23, 4183 (1981); J.L. Freeouf, G.W. Rubloff, P.S. Ho, and T.S. Kuan, Phys. Rev. Lett. 43, 1836 (1979)

62. P.E. Schmidt, P.S. Ho, H. Foll, and G.W. Rubloff, J. Vac. Sci. Technol. 18, 937 (1981); P.S. Ho, P.E. Schmidt, and H. Foll, Phys. Rev. Lett. 46, 782 (1981).

63. P.S. Ho, G.W. Rubloff, J.E. Lewis, V.L. Moruzzi, and A.R. Williams, Phys. Rev. B 22, 4784 (1980).

64. R. Tromp, E.J. van Loenen, M. Iwami, R. Smeek, and F.W Saris, Thin Solid Films, 93, 151 (1982).

65. J. Stohr and R. Jaeger, J. Vac. Sci. Technol. 21, 619 (1982). For Ni/Si, see F. Comin, J.E. Rowe, and P.H. Citrin, Phys. Rev. Lett. 51, 2402 (1983).

66. O. Bisi and K.N. Tu, Phys. Rev. Lett. 52, 1633 (1984).

67. I. Abbati, L. Braicovich, B. DeMichelis, O. Bisi and R. Rovetta, Solid State Commun. 37, 119 (1980).

68. O. Bisi, C. Calandra, L. Braicovich, I. Abbati, G. Rossi, I. Lindau, and W.E. Spicer, J. Phys. C 15 4707 (1982).

69. O. Bisi and C. Calandra, J. Phys. C. 14, 5479 (1981).

70. J. Ihm, M.L. Cohen and J.R. Chelikowsky, Phys. Rev. B 22, 4610 (1980).

71. I. Abbati, G. Rossi, I. Lindau, and W.E. Spicer, J. Vac. Sci. Technol. 19, 636 (1981).

72. J.N. Miller, S.A. Schwarz, I. Lindau, W.E. Spicer, B. DeMichelis, I. Abbati, and L. Braicovich, J. Vac. Sci. Technol. 17, 920 (1981).

73. G. Rossi, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, Solid State Commun. 39, 195 (1981).

74. I. Abbati, G. Rossi, L. Braicovich, I. Lindau, W.E. Spicer, and B. DeMichelis, J. Appl. Phys. 52, 6994 (1981).

75. A. Franciosi and J.H. Weaver, Phys. Rev. B 27, 3554 (1983).

76. Y.J. Chabal, J.E. Rowe, J.M. Poate, A. Franciosi, and J.H. Weaver, Phys. Rev. B 26, 2740 (1982).

77. P.J. Grunthaner, F.J. Grunthaner, and A. Madhukar, J. Vac. Sci. Technol. 21, 637 (1982).

78. P.J. Grunthaner, F.J. Grunthaner, and A. Madhukar, J. Vac. Sci. Technol. 20, 680 (1982).

79. P.J. Grunthaner, F.J. Grunthaner, A. Madhukar, and J.W. Mayer, J. Vac. Sci. Technol. 19, 649 (1981).

80. K. Oura, S. Okada, and T. Hanawa, Appl. Phys. Lett. 35, 705 (1979).

81. B. Aronsson, T. Lundstrom, and S. Rundqvist, Borides, Silicides, and Phosphides: A Critical Review of Thin Preparation, Properties, and Crystal Chemistry, (Wiley, NY 1965).

82. O. Bisi and C. Calandra have employed this technique with success in studies of several interfaces. See Refs. 66-69 and references therein.

83. A. Franciosi, J.H. Weaver, D.G. O'Neill, F.A.Schmidt, O. Bisi, and C. Calandra, Phys. Rev. B 28, 7009.

84. J.H. Weaver, V.L. Moruzzi, and F.A. Schmidt, Phys. Rev. B 23, 2916 (1981). A. Franciosi, D.J. Peterman, J.H. Weaver, and V.L. Moruzzi, Phys. Rev. B 25, 4981 (1982).

85. J.H Weaver, A. Franciosi, and V.L. Moruzzi, Phys. Rev. B 29, 3293 (1984) and references therein.

86. J.H. Weaver, M. Grioni, and J.J. Joyce, Phys. Rev. B 31, xxxx (1985) for Cr/GaAs. See also results for V/GaAs by M. Grioni, J.J. Joyce, and J.H. Weaver, J. Vac.Sci. Technol. May/June (1985) and for Ti/GaAs by M.W. Ruckman, M. del Giudice, and J.H. Weaver, Phys. Rev. B 31, xxxx (1985).

87. L.J. Brillson, R.Z. Bachrach, R.S. Bauer, and J. McMenamin, Phys. Rev. Lett. 42, 397 (1979).

88. M. Grioni, M. del Giudice, J.J. Joyce, and J.H. Weaver, J. Vac. Sci. Technol. May/June 1985.

89. J.H. Weaver, M. Grioni, J. Joyce, M. del Giudice, Phys. Rev. B 31 (1985).

91. L.J. Brillson, C.F. Brucker, N.G. Stoffel, A.D. Katnani, and G. Margaritondo, Phys. Rev. Lett. 46, 838 (1981).

92. L.J. Brillson, C.F. Brucker, N.G. Stoffel, A.D. Katnani, and G. Margaritondo, J. Vac. Sci. Technol. 19, 661 (1981).

93. L.J. Brillson, R.Z. Bachrach, R.S. Bauer, and J. McMenamin, Phys. Rev. Lett. 42, 397 (1979).

94. L.J. Brillson, G. Margaritondo, and N.G. Stoffel, Phys. Rev. Lett. 41, 667 (1980).

95. L.J. Brillson, Thin Solid Films, 89, 461 (1982).

96. I. Lindau, P.W. Chye, C.M. Garner, P. Pianetta, and W.E. Spicer, J. Vac. Sci. Technol. 15, 1332 (1978); P.W. Chye, I. Lindau, P. Pianetta, C.M. Garner, C.Y.Su, and W.E. Spicer. Phys. Rev. B 18, 5545 (1970).

97. L.J. Brillson, R.S. Bauer, R.Z. Bachrach, and G. Atanson, Appl. Phys. Lett. 36, 326 (1980).

98. R.Z. Bachrach, J. Vac. Sci. Technol. 15, 1340 (1978); R.Z. Bachrach and R.S. Bauer, J. Vac. Sci. Technol. 16, 1147 (1979); R.Z. Bachrach, R.S. Bauer, P. Chiaradi, and G.V. Hansson, J. Vac. Sci. Technol. 19, 335 (1981).

99. A. Franciosi, J.H. Weaver, D.G. O'Neill, Phys. Rev. B 28, 4889 (1983); A. Franciosi, D.G. O'Neill, and J.H. Weaver, J. Vac. Sci. Technol. B1,

524 (1983).

100. L. Braicovich, C.M. Garner, P.R. Skeath, C.Y. Su, P.W. Chye, I. Lindau, and W.E. Spicer, Phys. Rev. B 20, 5131 (1979).

101. T. Narusawa, K. Kinoshita, W.M. Gibson, and A. Hiraki, J. Vac. Sci. Technol. 18, 272 (1981); P. Perfetti, S. Nannarone, F. Patella, C. Quaresima, A. Savoia, F. Cerrina, and M. Capozi, Solid State Commun. 35, 151 (1980).

102. A. Cros, J. Derrien, and F. Salvan, Surf. Sci. 110, 471 (1981); A. Cros, F. Salvan, and J. Derrien, J. Appl. Phys. 52, 4757 (1981),

103. A.D. Katnani, P. Perfetti, Te-Xiu Zhao, and G. Margaritondo, J. Vac. Sci. Technol. A2, 650 (1983) A.D. Katnani, P. Perfetti, Te-Xiu Zhao, and G. Margaritondo, Appl. Phys. Lett. 40, 619 (1982).

104. I. Abbati, G. Rossi, L. Calliari, L. Braicovich, I. Lindau, and W.E. Spicer, J. Vac. Sci. Technol. 21, 409 (1982).

105. A. Franciosi, private communication and J. Vac. Sci. Tech. (to be published).

106. E.-E. Latta and M. Ronay, Phys. Rev. Lett. 53, 948 (1984).

107. M. del Giudice, J.J. Joyce, M.W. Ruckman, and J.H. Weaver, J. Vac. Sci. Technol. (1985).

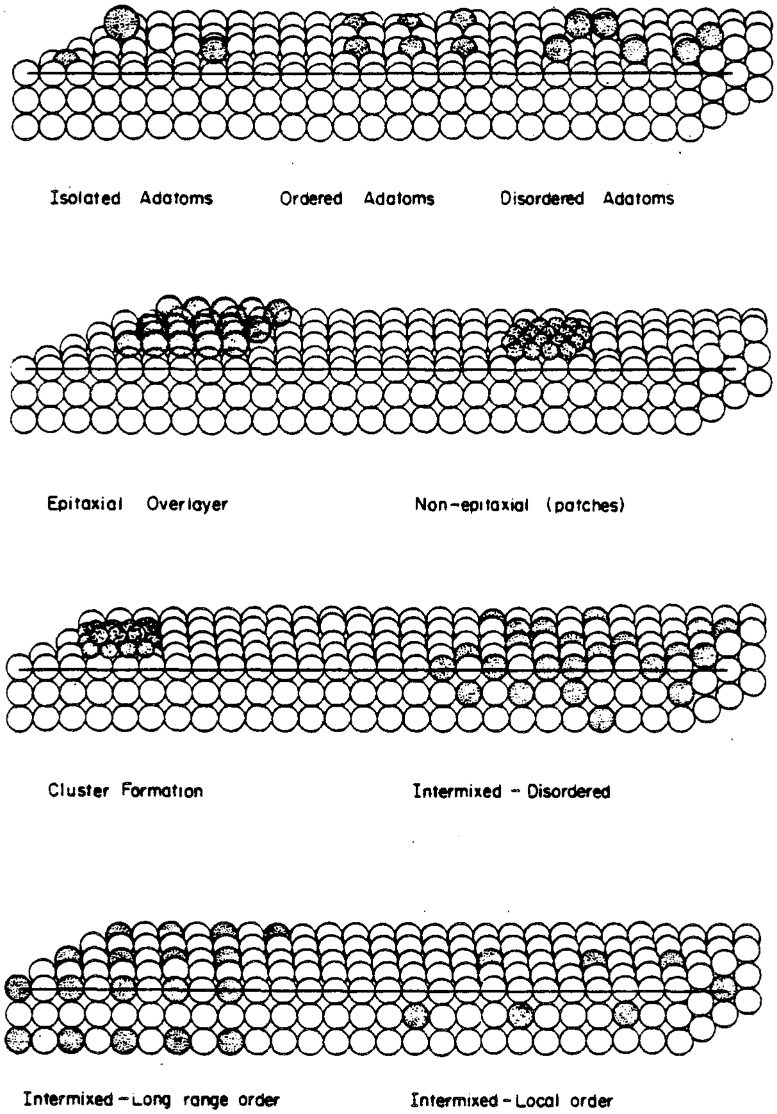


Fig. 1. Examples of possible interface morphologies starting with isolated adatoms (top left) and continuing through intermixed systems (bottom). Different isolated adatom geometries include hollow site and atop sites. Also shown are a "large" adatom on the surface and a vacancy. The ordered adatoms (top center) are all hollow-site-bonded. The disordered adatoms (top right) are schematically shown atop-bonded with the possibility of forming dimers or remaining isolated. The epitaxial overlayer has adatoms continuing the substrate array while the non-epitaxial array has adatoms bonded more strongly to each other than to the substrate, thereby establishing their own mesh. The transition from patch to cluster is depicted by the second row of atoms (3D array). Intermixing is shown disordered (completely random array of distinguishable atoms), ordered with long range order (a compound), and intermixed with only short range order (shaded atoms have only opposite-species nearest neighbors). These sketches become more complicated if one considered ternary systems like metals on compound semiconductors or surface segregation. No attempt has been made to show surface reconstruction.

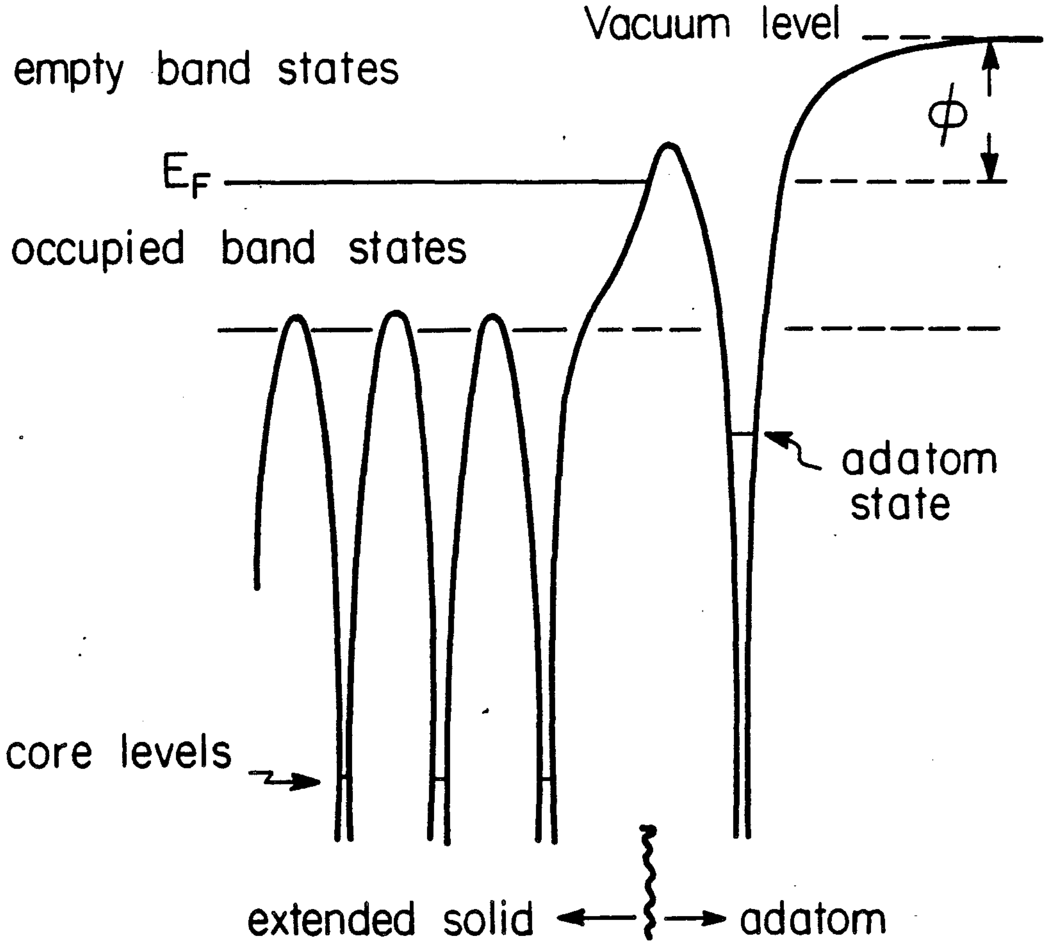


Fig. 2. A schematic representation of a slice through a crystal showing the potential energy variation as a function of position for bulk atoms and for an adatom on the surface. The band states result from overlap of the valence electrons. The Fermi level, E_F, separates occupied and empty states. Core levels are shown to be localized states. The presence of the adatom varies the potential energy and therefore the electronic states near the surface.

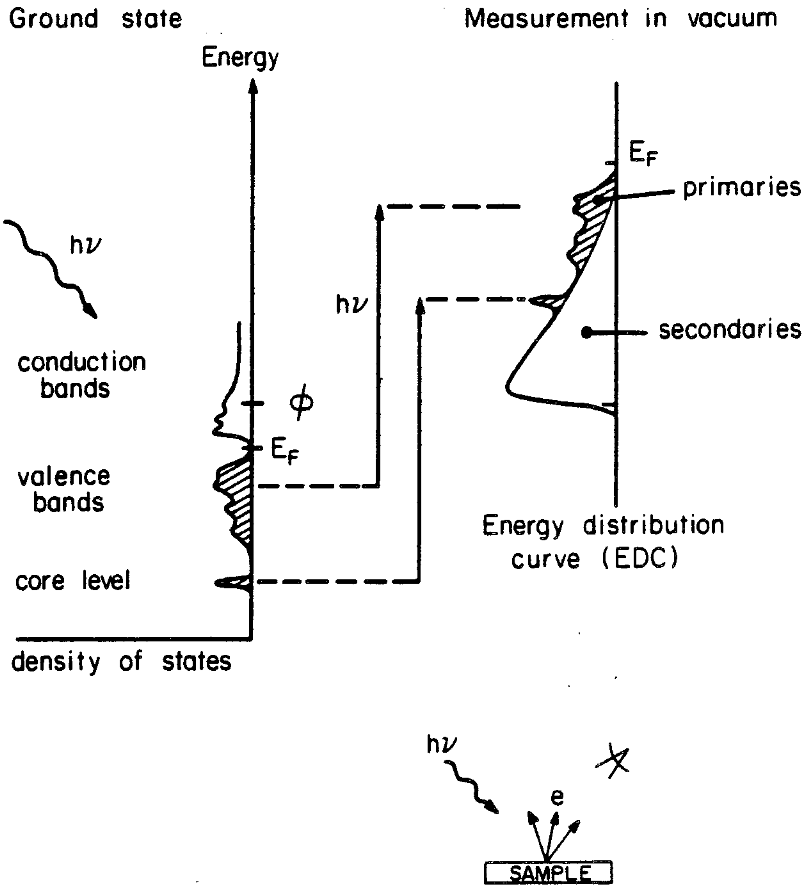


Fig. 3. The density of states of a fictitious semiconductor is shown on the left with structure in both the occupied and empty states and a shallow core level below the valence bands. Electrons are excited to higher-lying energy states upon absorption of a photon of energy $h\nu$. Measurement in vacuum of the energy distribution of photoemitted electrons (an EDC) provides information about the states of the electron in the solid. In this sketch, we assume that matrix elements are constant in the photoabsorption process described in Eq. 1. Secondary electrons arise from inelastic scattering before escape from the solid.

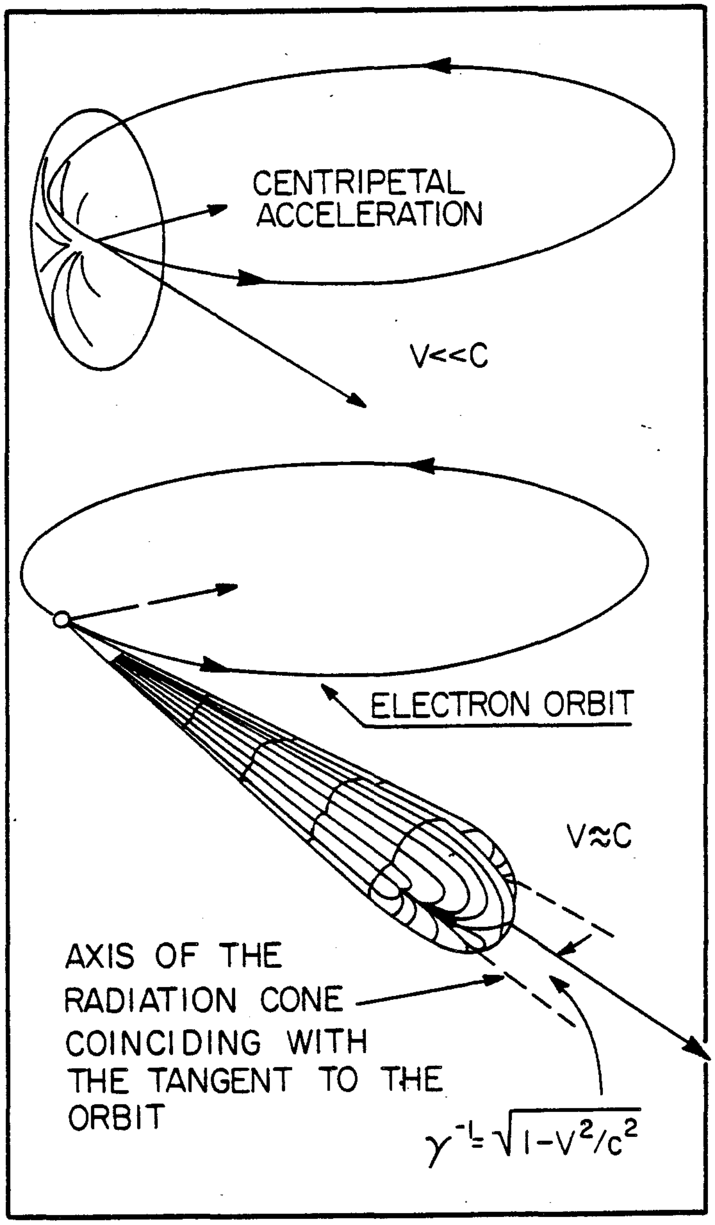

CENTRIPETAL ACCELERATION
V<<C
ELECTRON ORBIT
V≈C
AXIS OF THE RADIATION CONE COINCIDING WITH THE TANGENT TO THE ORBIT
$\gamma^{-1} = \sqrt{1-V^2/c^2}$

Fig. 4. The radiation pattern for electrons moving in circular orbits. The pattern for low velocity electrons is the donut-shaped Larmor distribution shown at the top. For velocities approaching the speed of light, the pattern is strongly peaked in the forward direction and is close to the orbital plane.

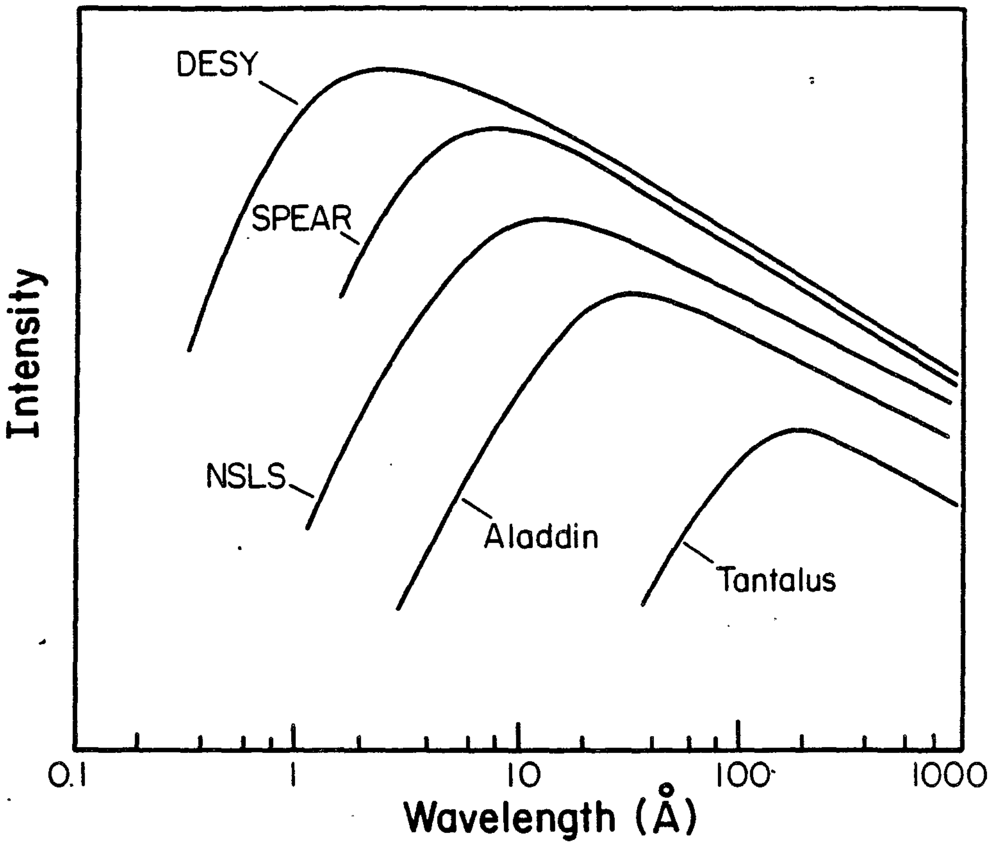


Fig. 5. A plot of the intensity of synchrotron radiation as a function of wavelength for representative synchrotron sources. Tantalus operates at 240 MeV, Aladdin will operate at 1 GeV, the x-ray ring at NSLS will operate at 2.5 GeV, SPEAR typically runs at 3 GeV for dedicated light source operation, and DESY in Hamburg operates at 6 GeV. Note that E(eV)=12398/λ(Å) and the vertical scale is logarithmic. See Refs. 17-20 for detailed discussions of intensities.

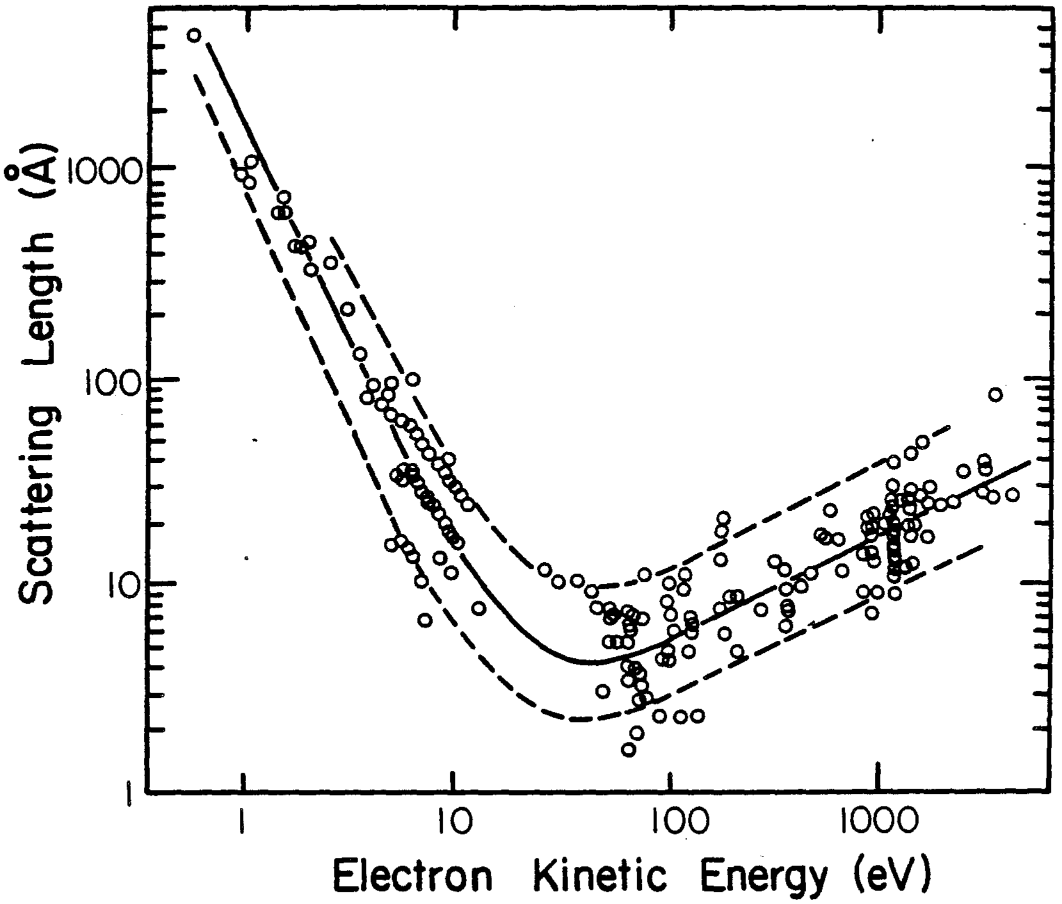


Fig. 6. Scattering length vs. electron kinetic energy showing the enhanced surface sensitivity to be gained at energies near the minimum. The different points represent different materials with most variations falling within the dashed lines. See Ref. 32.

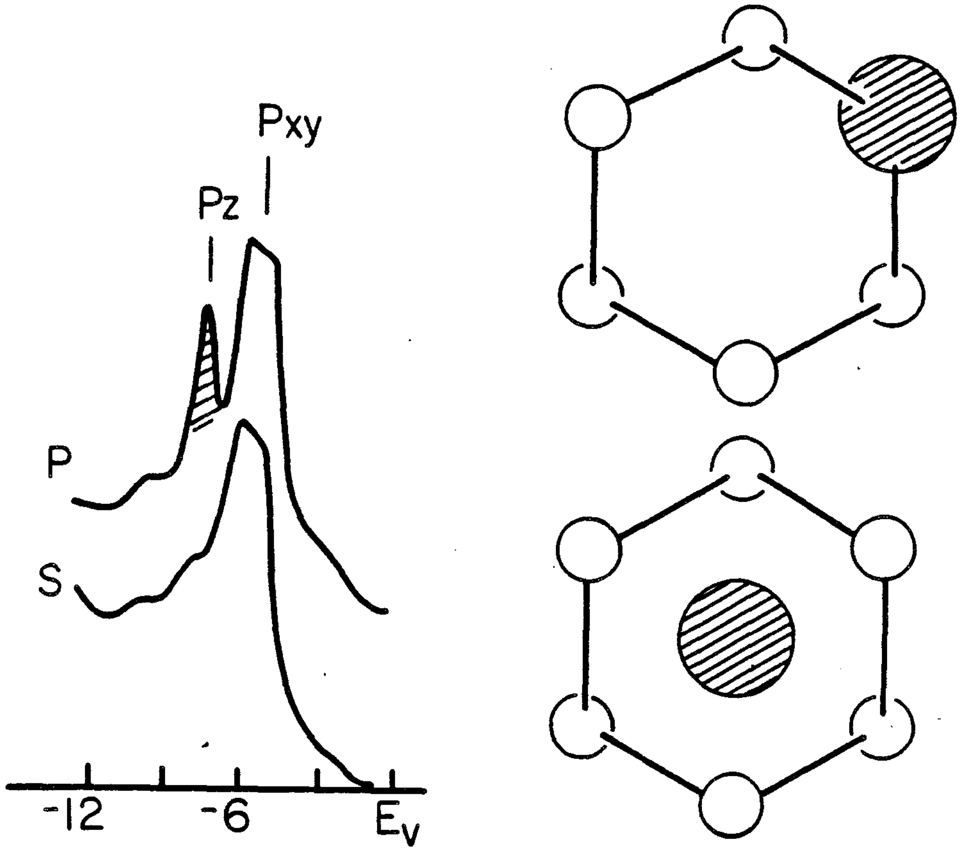


Fig. 7. Example of variation in the EDCs for Cl on Si(111) as a function of polarization as discussed in the text. Observation of the polarization dependence of the feature labelled p_z allowed identification of the orbital symmetry and subsequently the determination of the chemisorption site for Cl, as shown at the right (Refs. 43 and 44). Chlorine is drawn shaded, the surface Si atoms are drawn full, and the next layer Si atoms has "bonds" through circles. The upper configuration is preferred.

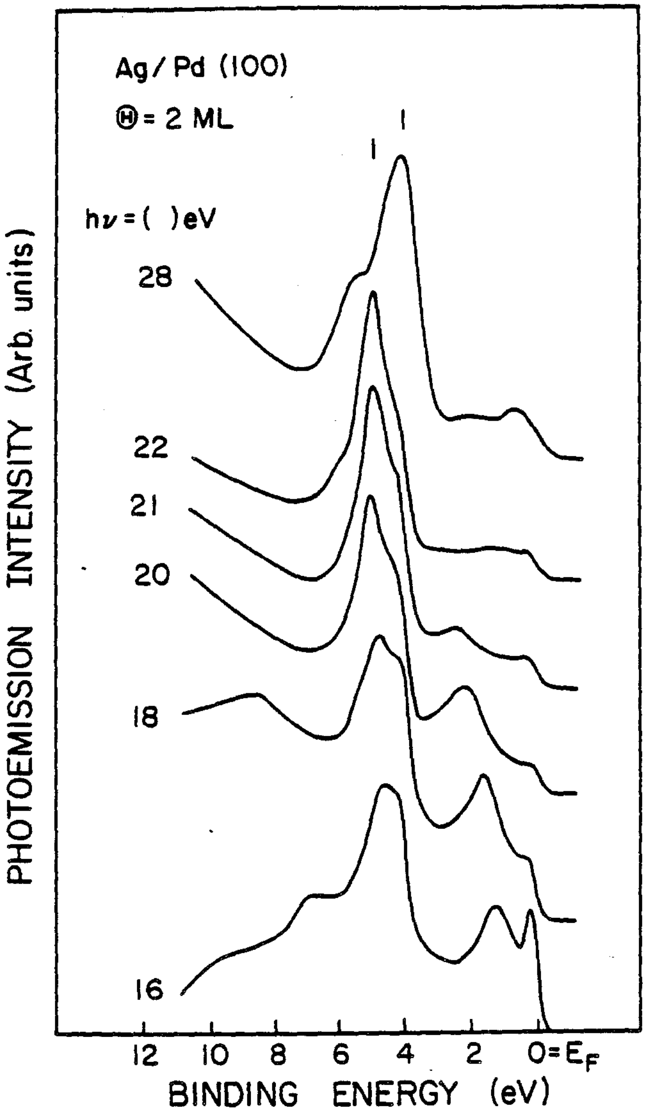


Fig. 8. Photoemission results for 2 ML of Ag on Pd(100) showing the hν-variation of the two Ag-derived states indicated by tic marks. For 1 ML only one states appears. With increasing coverage the EDCs converge to bulk Ag. These results were used in Ref. 47 to discuss the change from 2 to 3 dimensional epitaxial overlayers.

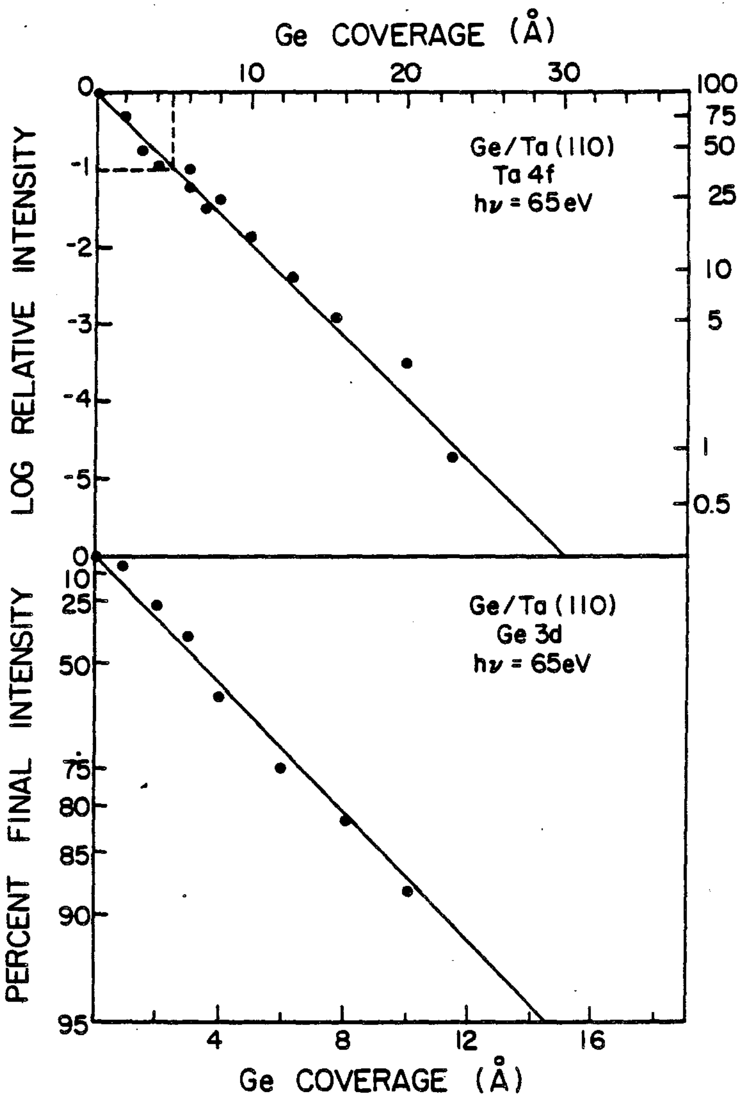


Fig. 9. Core level intensity results for the Ge/Ta(110) interface normalized to the emission of the clean Ge or Ta surface showing logarithmic Ta attenuation and Ge growth indicative of an abrupt interface. The mean free paths of the photoelectrons were shown to be ~5 Å after Ruckman *et al.* in Ref. 49.

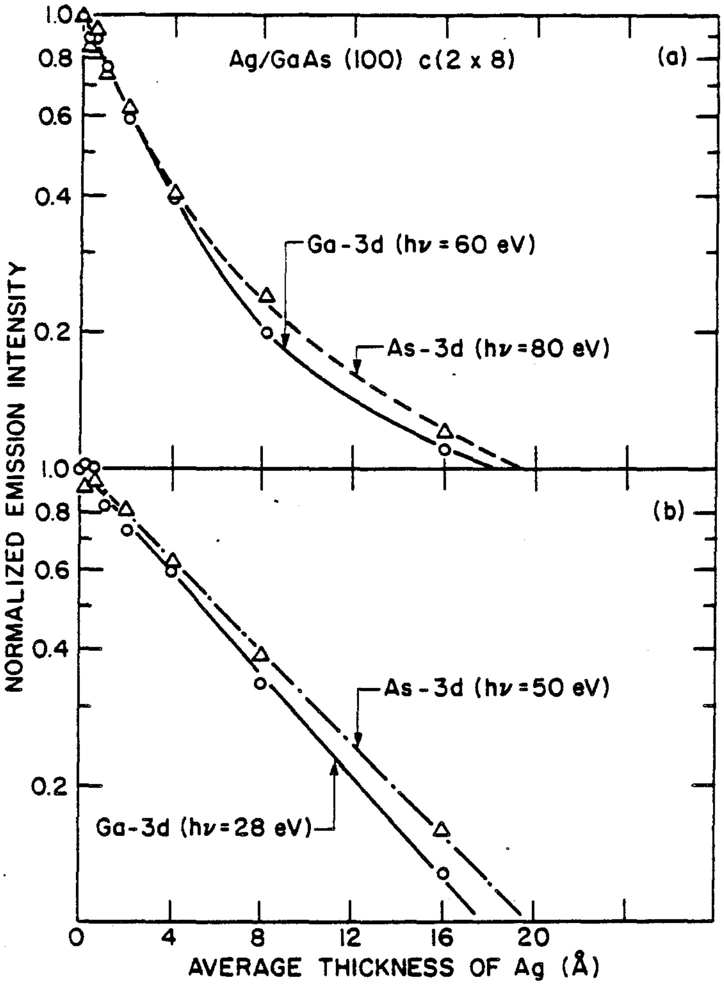


Fig. 10. Core level attenuation results for the Ga and As 3d core levels as a function of Ag overlayer thickness. The results at the top have significantly higher surface sensitivity than those at the bottom (~4.5 vs. 8 Å). Ludeke has used results such as this to argue for the formation of Ag clusters on the chemically inactive surface (Ref. 5).

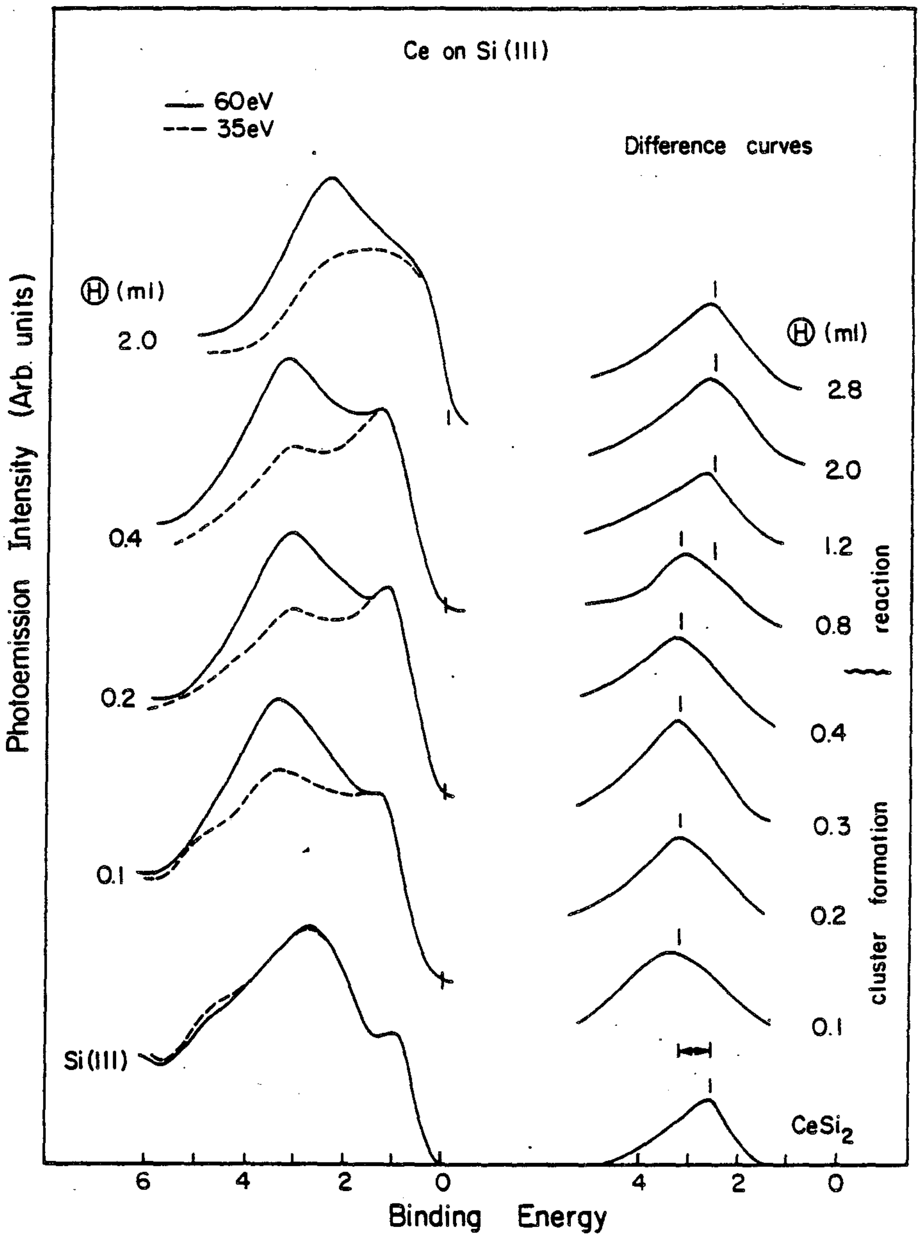


Fig. 11. Photoemission results for the Ce/Si(111) interface by Grioni et al. (Refs. 59 and 60) which show the growth of 4f-derived emission ~3 eV below E_F for ultralow coverages and the transition to results representative of CeSi$_2$ above 0.6 ML coverage. These results were used in conjunction with Si 2p core level studies, LEED, and angle-resolved Auger studies to demonstrate the formation of Ce clusters and their role in inducing disruption of the Si surface.

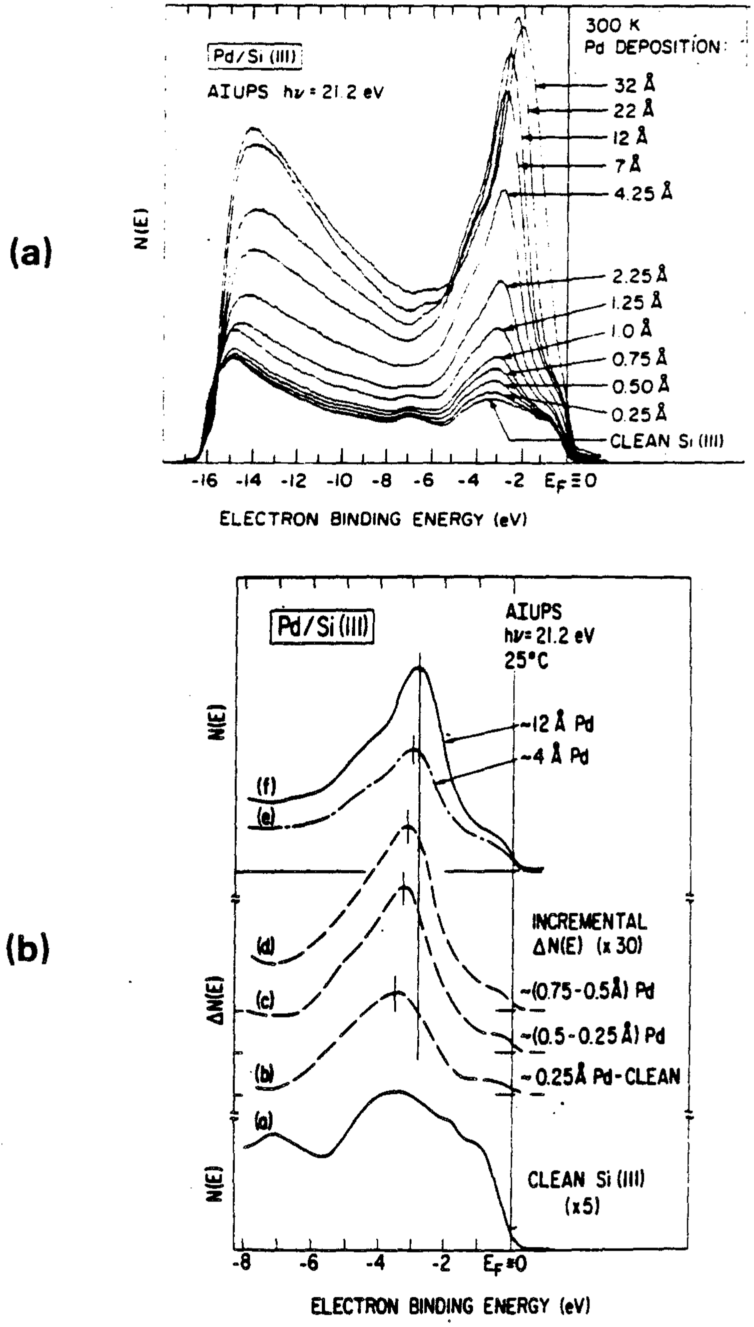


Fig. 12. Photoemission results from Rubloff et al. (Ref. 61) for the reactive Pd/Si(111) interface. EDCs are shown at the top and incremental difference curves are shown at the bottom. These results demonstrate the onset of reaction for this interface at lowest coverage, the formation of a Pd-Si phase, and the Pd enrichment of the surface layer at high coverages.

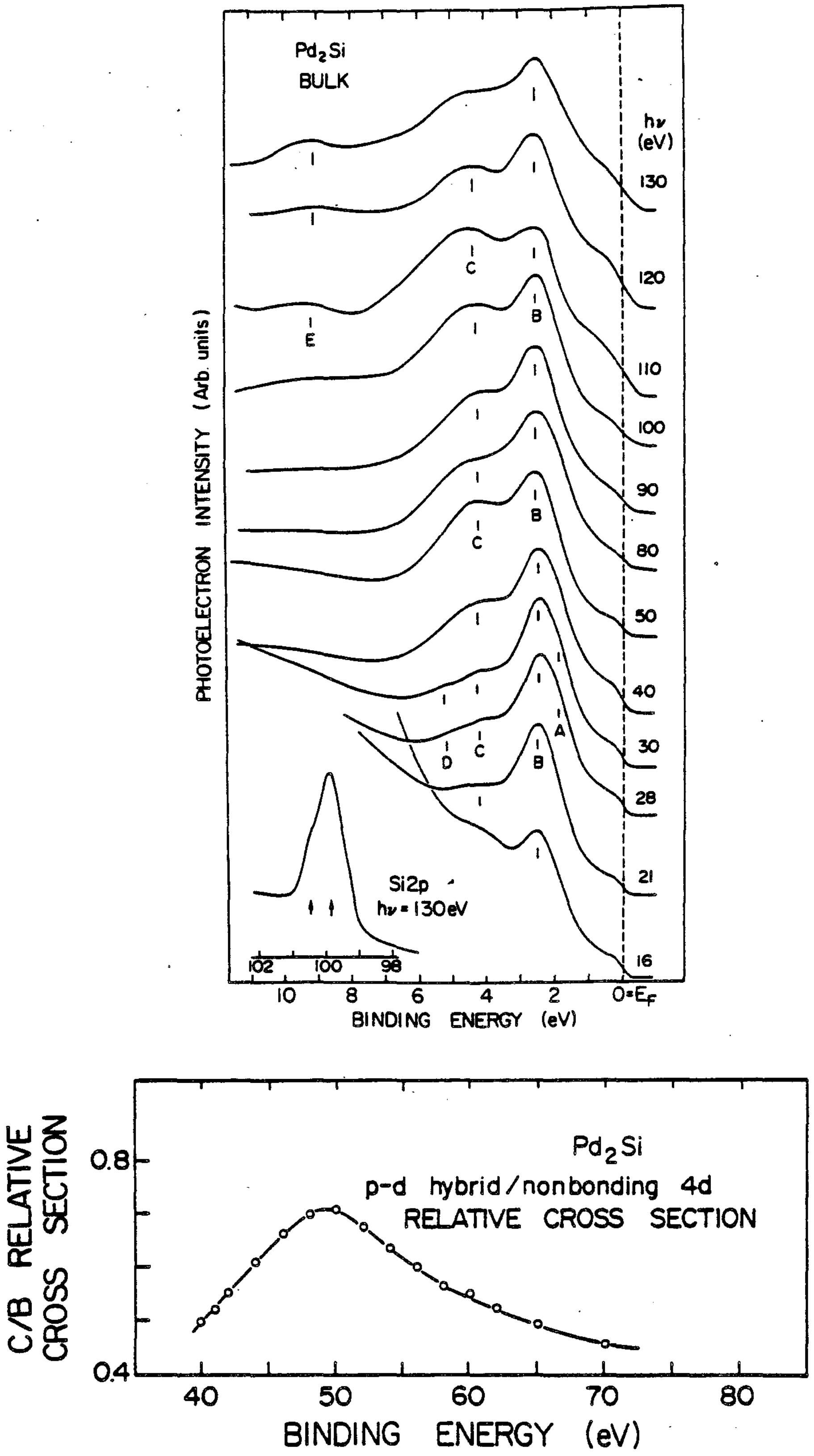


Fig. 13. Results of photoemission studies of bulk Pd$_2$Si as a function of photon energy showing the variation in cross section of the Pd states near 2 eV and the Pd-d/Si-p bonding states near 4 eV (peaks B and C, respectively). The relative cross sections of peaks B and C is given at the bottom of the figure. These results make it possible to identify the origin of different features in photoemission results and are an example of the p-d resonant photoemission technique. From Ref. 75.

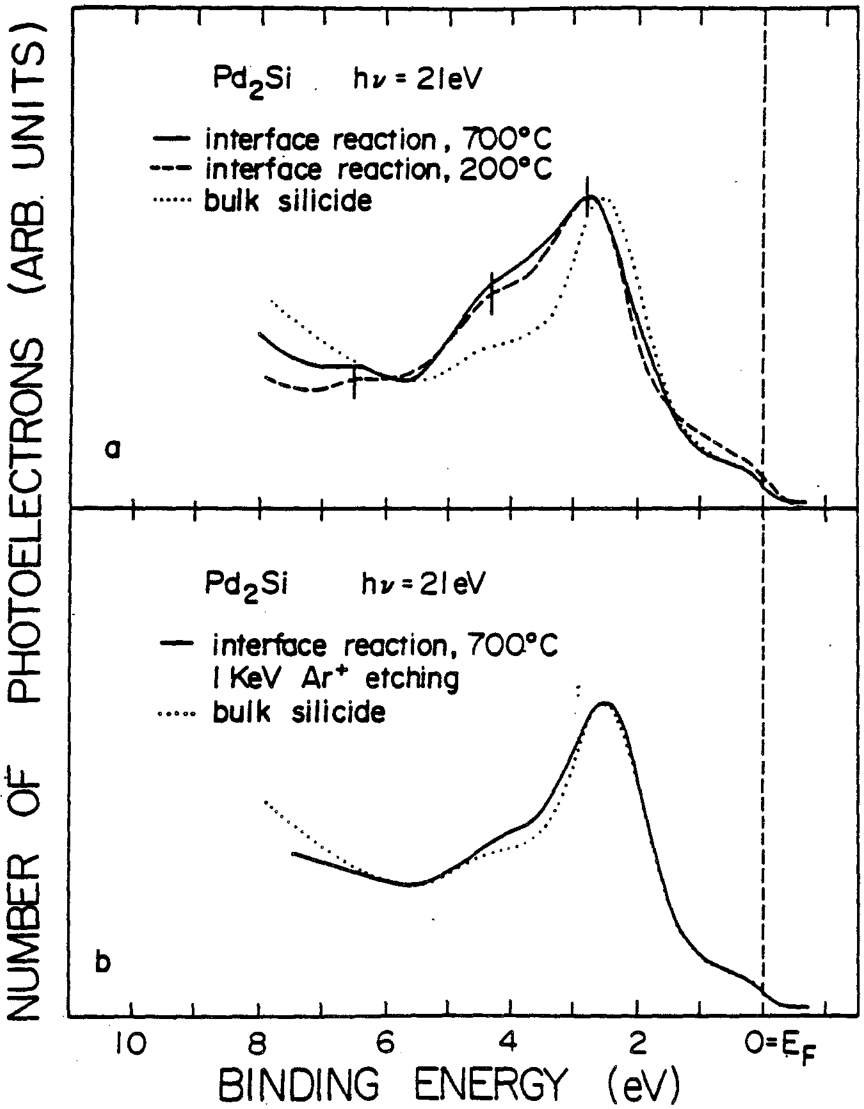


Fig. 14. Photoemission results which compare bulk Pd$_2$Si with interface reacted films of Pd on Si. The results at the top indicate that the surface of the reacted sample is Si-rich compared to the bulk sample. Light Ar sputtering removes this Si and produces results equivalent to the bulk. From Ref. 76.

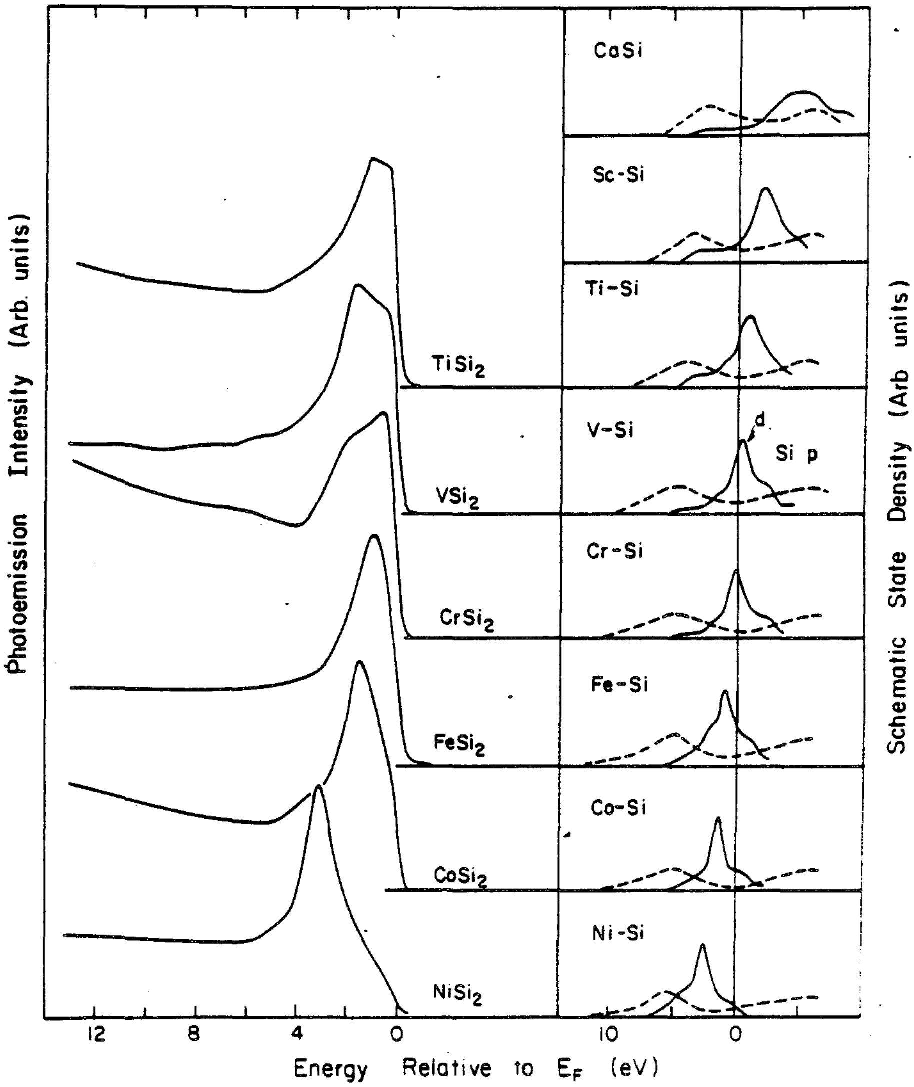


Fig. 15. Summary of photoemission results for the disilicides of the 3d transition metals. On the right, we show schematic densities of states for the metal-derived d states and the Si-derived p states. Comparison with experiment shows that the photoemission results emphasize the nonbonding metal d states but that metal/silicon hybrid states are formed to ~6 eV below E_F. Note that the scales for the Si and metal results are not the same. From Ref. 85.

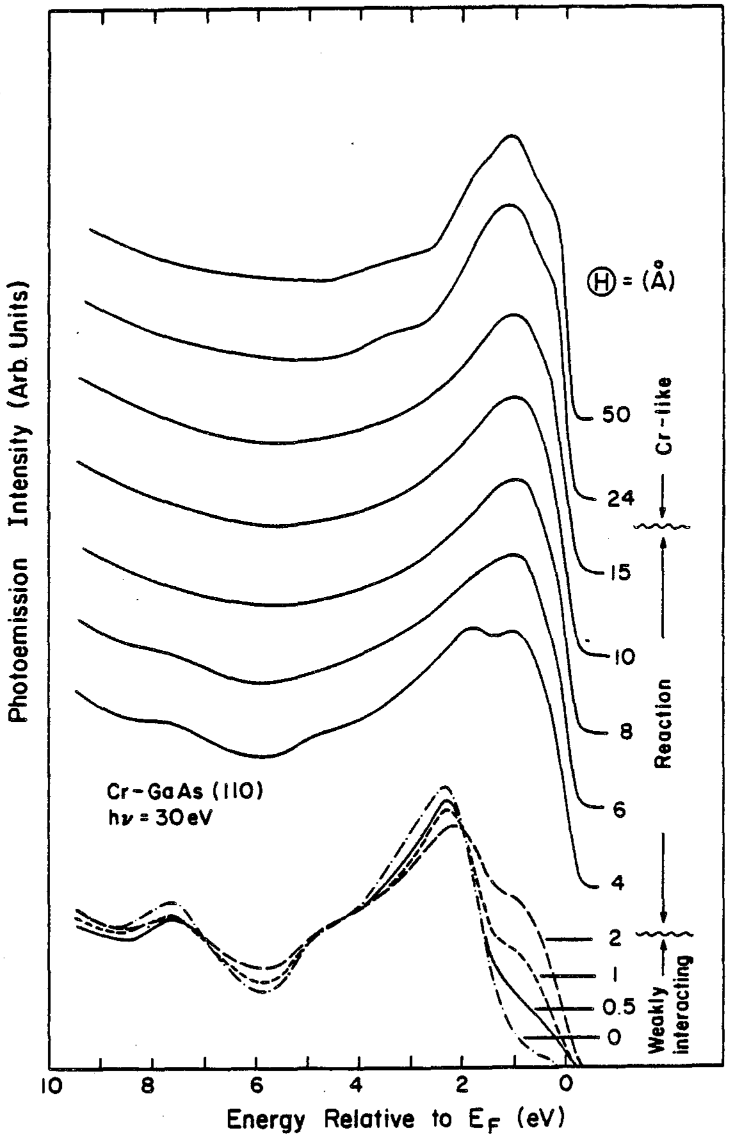


Fig. 16. Photoemission results for the Cr/GaAs(110) interface showing the evolution of the valence bands as a function of Cr coverage. At low coverage, Cr atoms form aggregates or patches which attenuate the substrate but are not metallic. Reactive intermixing is induced only after 2Å coverage and a Cr-rich film starts to grow on the reacted layer after ~20Å. From Ref. 86.

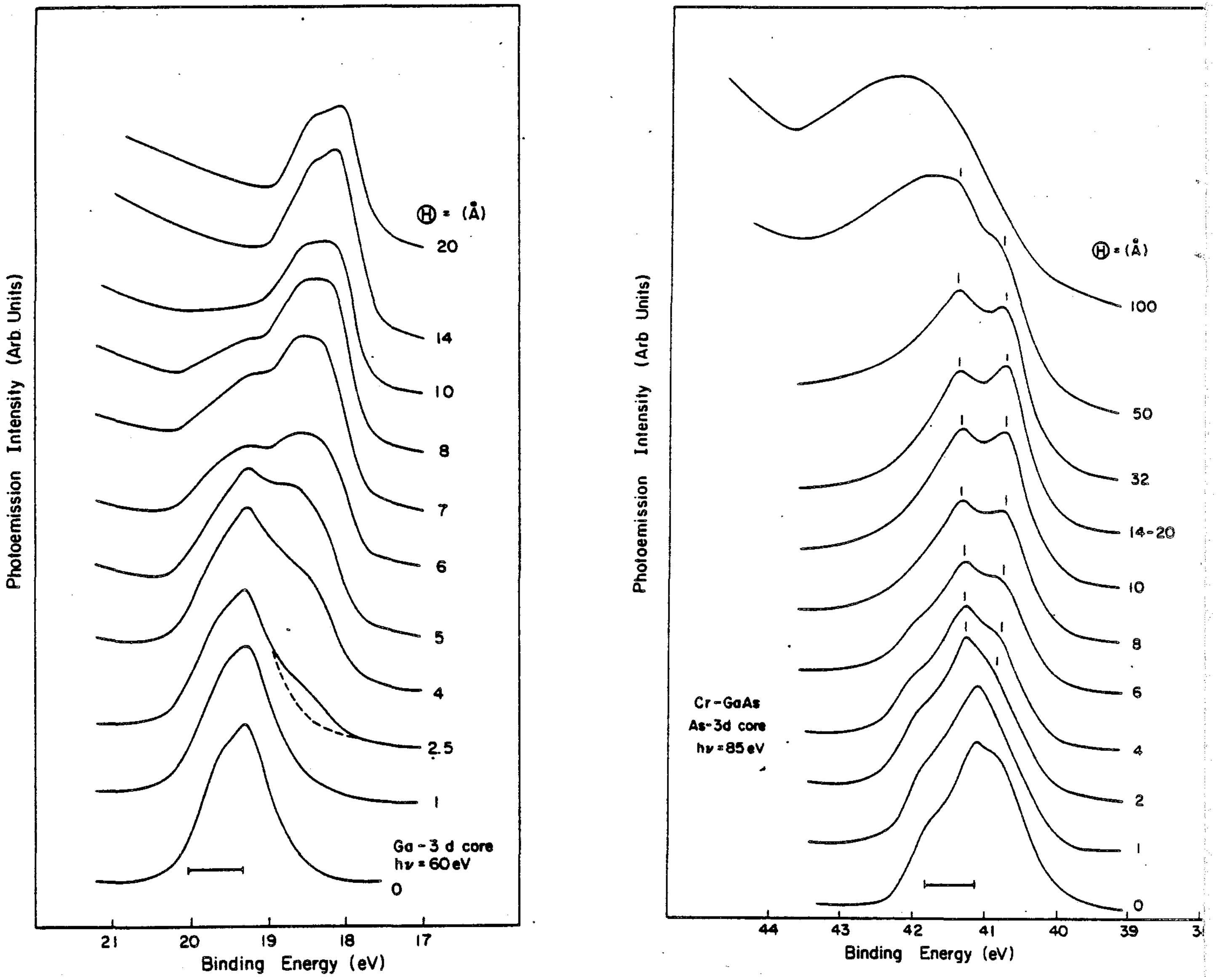


Fig. 17. Core level EDCs for the Cr/GaAs(110) interface showing significant chemical shifts of the Ga and As cores following the onset of reactive intermixing. The results for Ga on the left show steady shift with coverage and the absence of any well-defined chemical environment. Those for As on the right show the appearance of reacted-As and surface-segregated-As atoms which have fixed binding energies independent of coverage, i. e. well- defined chemical environments. From Ref. 89.

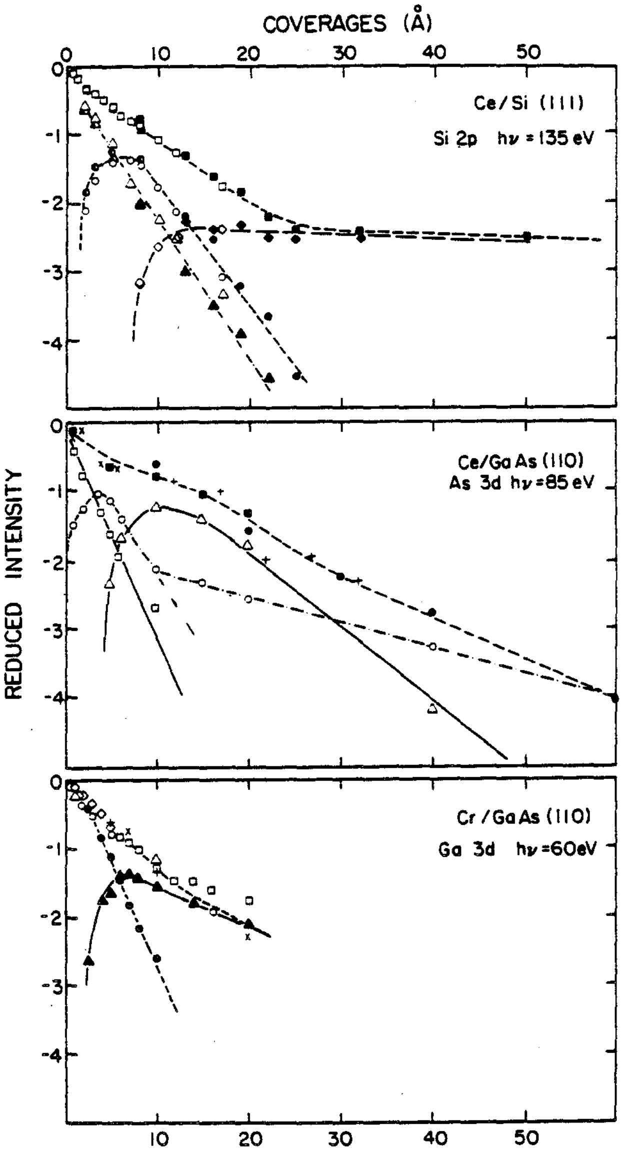


Fig. 18. Results of detailed analysis of the attenuation of the core emission for Si in Ce/Si, As in Ce/GaAs, and Ga in Cr/GaAs. The core EDCs for Si and As were deconvolved into components representing substrate, reacted, and surface segregated species. Such deconvolution was not possible for Ga and the results are distinguished only as substrate and reacted species. These component-specific results greatly facilitate modeling of the interface morpology, as discussed in the text. From Ref. 88.

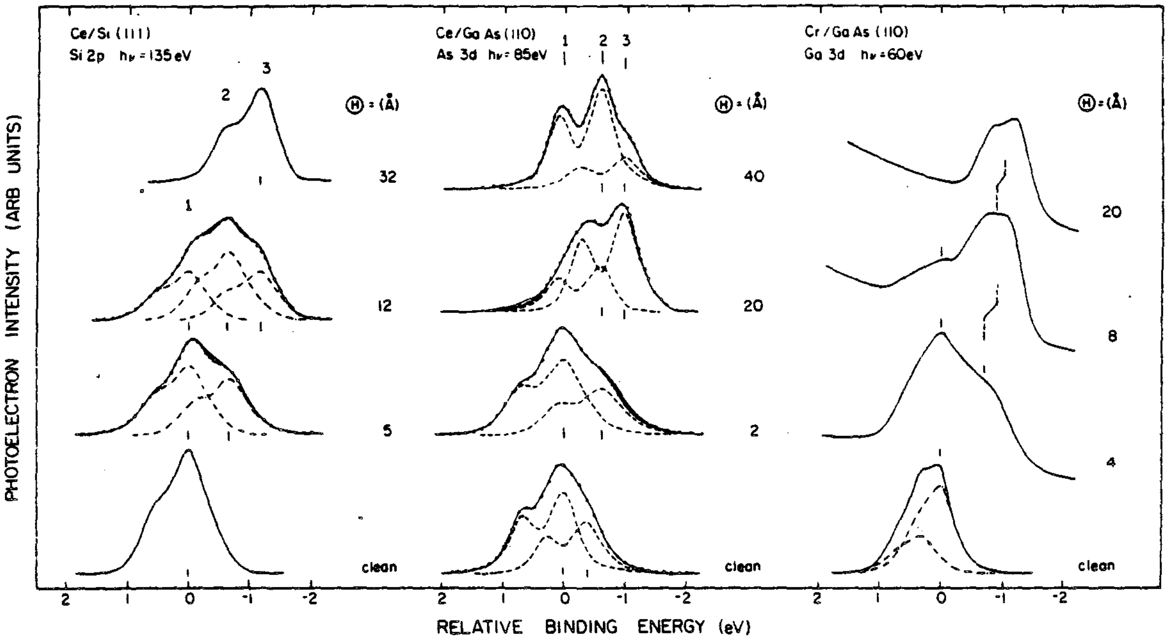


Fig. 19. Lineshape analysis results for core emission of Si in Ce/Si, As in Ce/GaAs, and Ga in Cr/GaAs from Ref. 88. Decomposition of the core EDCs makes it possible to examine the growth/attenuation of each component or chemical environment as shown in Fig. 18.

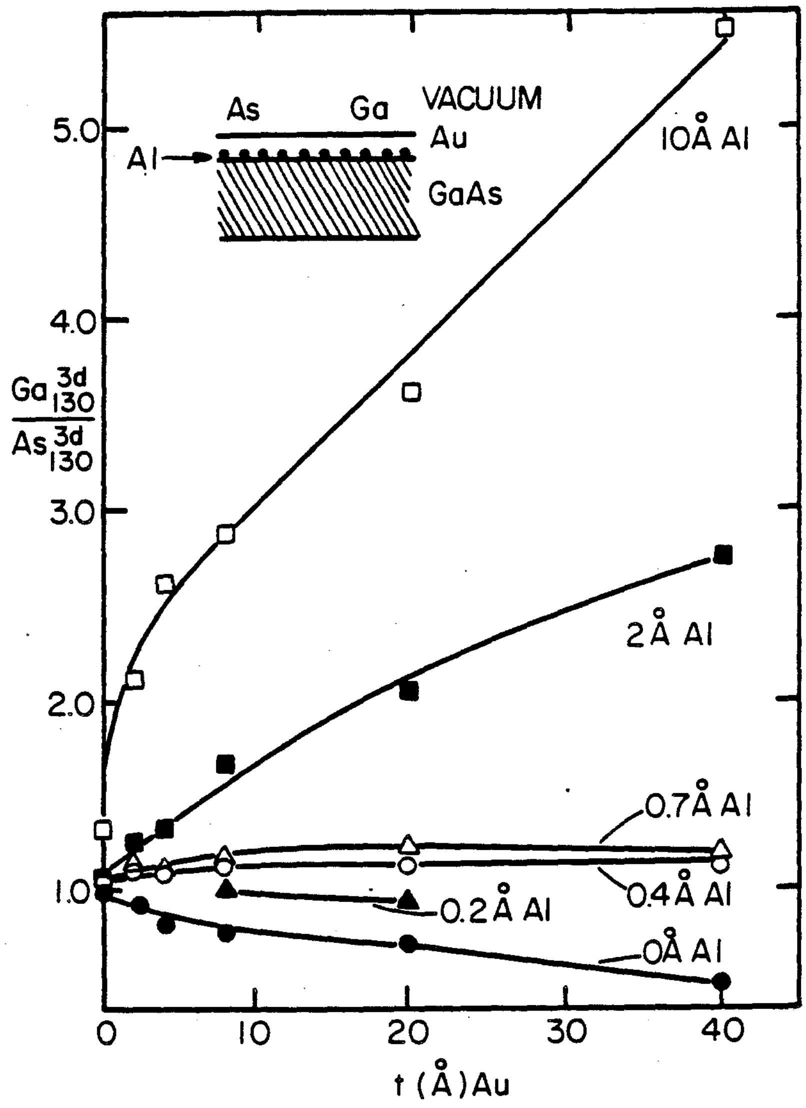


Fig. 20. The ratio of Ga to As for the Au/Al/GaAs system as a function of Al interlayer thickness. These results by Brillson et al. (Ref. 94) show that the thickness of the interlayer controls the outdiffusion of Ga and As into Au and has a significant impact on their relative abundance in the overlayer.

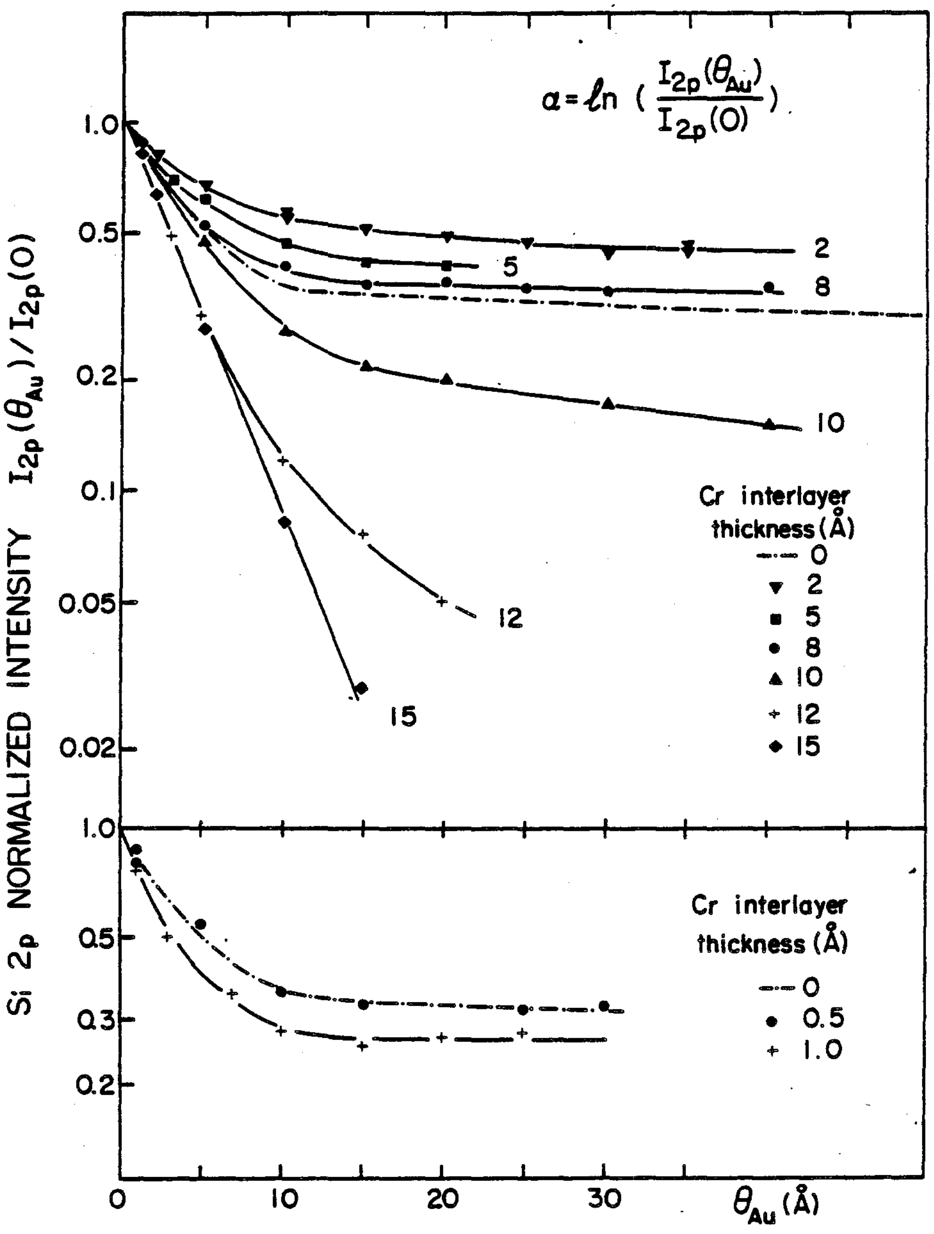


Fig. 21. The attenuation coefficient of the Si 2p core as a function of Au coverage and Cr interlayer thickness for the Au/Cr/Si system. As shown by Franciosi et al. (Ref. 99), the outdiffusion of Si can be controlled by varying the reactivity of the Cr/Si interface. For coverages below the onset of reaction, the outdiffusion of Si is reduced. It is greatly enhanced, however, following the Cr-induced disruption of the surface. At high coverage, the Cr layer which forms again acts as a diffusion barrier.

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

VLSI ARCHITECTURES FOR LU DECOMPOSITION

Microelectronic and Information Sciences Center

Technical Report #27

K. H. Cheng
Department of Computer Science
University of Minnesota

S. Sahni
Department of Computer Science
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

VLSI ARCHITECTURES FOR LU DECOMPOSITION

MEIS Technical Report #27

K. H. Cheng
S. Sahni

This research was supported in part by the National Science Foundation under grant MCS-83-05567 and the Microelectronic and Information Sciences Center.

ABSTRACT

VLSI architectures involving meshes, meshes with broadcast capability, and chains are examined for the LU decomposition problem. We develop designs with superior performance to earlier designs.

Keywords and Phrases

VLSI architectures, systolic systems, LU decomposition

1. Introduction

VLSI architectures for a variety of problems have been proposed by several authors. A bibliography of over 150 research papers dealing with this subject appears in [KUNG83]. In this paper, we are concerned solely with the LU decomposition problem. The input to this problem is an $n \times n$ matrix A whose LU decomposition can be performed by Gaussian elimination without pivoting. As observed in [KUNG79], pivoting is not required, for example, when A is positive definite, symmetric, or irreducibly diagonally dominant. The lower and upper triangular matrices $L = (l_{ij})$ and $U = (u_{ij})$ may be computed using the recurrences [KUNG79]:

$$a_{ij}^{1} = a_{ij}$$

$$a_{ij}^{k+1} = a_{ij}^{k} - l_{ik}\, u_{kj}$$

$$l_{ik} = \begin{cases} 0, & i < k \\ 1, & i = k \\ a_{ik}^{k}/u_{kk}, & i > k \end{cases} \tag{1}$$

$$u_{kj} = \begin{cases} 0, & k > j \\ a_{kj}^{k}, & k \le j \end{cases}$$

VLSI architectures for this problem have been proposed earlier in [HORO79], [KUNG78,79 and 84] and [LEIS83]. [HORO79] proposes a mesh architecture similar to that of Figure 1.1(c), while [KUNG78 and 79] and [LEIS83] propose a hexagonal architecture as in Figure 1.1(d). In [KUNG84], a ring architecture as in Figure 1.1(b) is used for the LU decomposition problem.

In this paper, we examine the mesh and chain (Figure 1.1(a)) architectures In addition, we consider broadcast lines as used in [HUAN82], and [CHEN84a and b] (among others) for the matrix multiplication and back substitution problems. To our knowledge, there has been no earlier research on the use of a broadcast mesh for LU decomposition. An example of a broadcast mesh is given in Figure 1.1(e). A broadcast line has the property that data put on this line becomes available at all PEs on the line in O(1) time.

In evaluating our designs, we assume that the VLSI system will be attached to the host processor using a bus as in Figure 1.2. The evaluation of a VLSI design should take the following into account:

1. Bus bandwidth — how much data is to be transmitted between the host and the VLSI system in any cycle? This figure is denoted by B.
2. Speed — how much time does the VLSI system need to complete its task? This time may be decomposed into the times T_C (time for computations) and T_D (time for data transmissions both within the VLSI system and between the host and the VLSI system).

One may expect that by using a very high bandwidth B and a large number of processors P, we can make T_C and T_D quite small. So, T_C and T_D are not in themselves a very good measure of the effectiveness with which the resources B and P have been used. Let D denote the total amount of data that needs to be transmitted between the host and VLSI system. The ratio

$$R_D = B * T_D / D$$

measures the effectiveness with which the bandwidth B has been used. Clearly, $R_D \geq 1$ for every VLSI design. As an example, consider the multiplication of two $n \times n$ matrices. The host needs to send $2n^2$ elements to the VLSI system and receive n^2 elements back. So, $D = 3n^2$. With a bandwidth of n, T_D must be at least $3n$. T_D will exceed $3n$ if the bandwidth is not used to capacity at all times.

Let C denote the time spent for computation by a single processor algorithm. The ratio

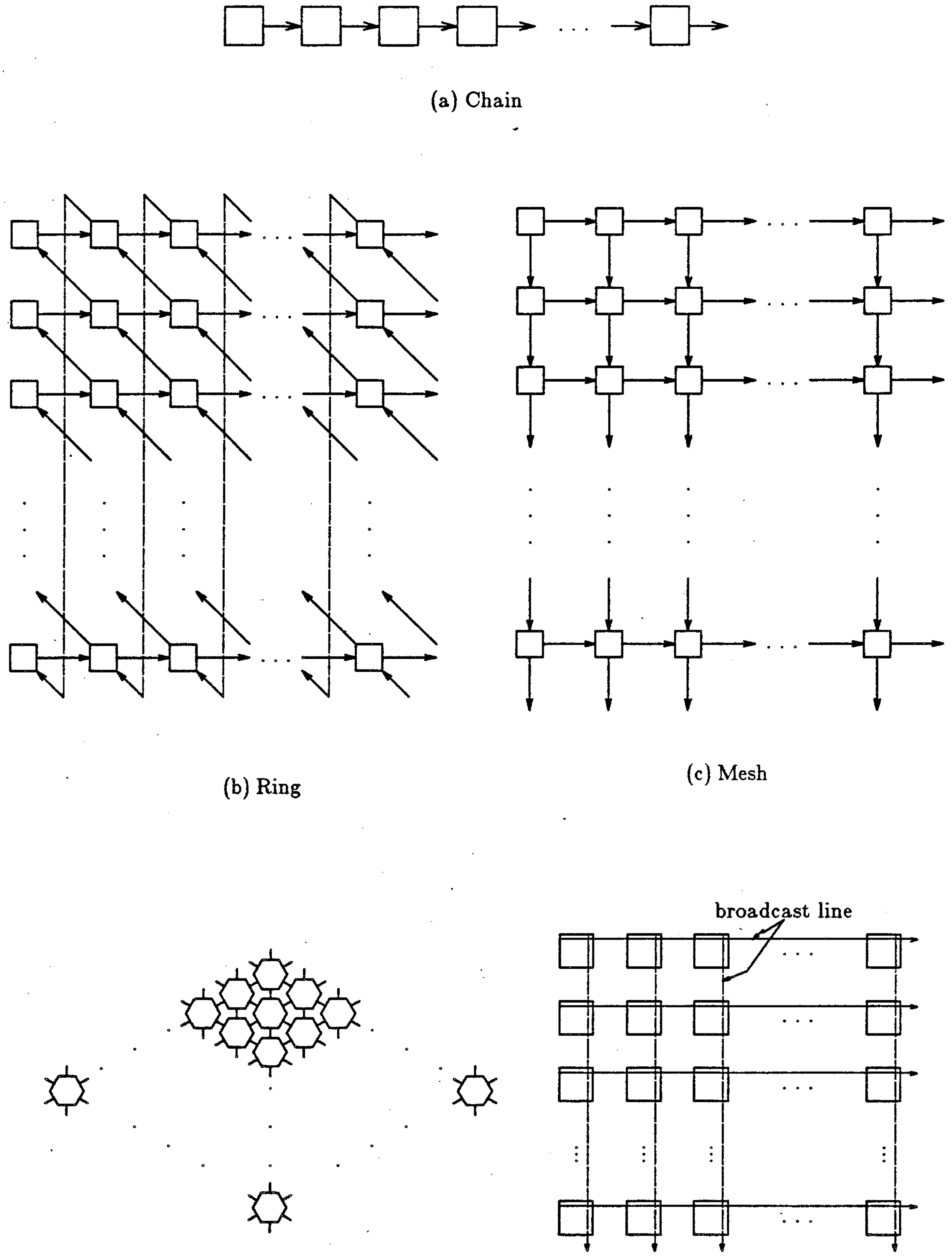
$$R_C = P * T_C / C$$

measures the effectiveness of processor utilization. Once again, we see that $R_C \geq 1$ for every VLSI design. Consider the problem of multiplying two $n \times n$ matrices A and B to get X. Each element of X is the sum of n products. We shall count one multiplication and addition as one arithmetic (or computation) step. Hence, $C = n^3$. If $P = n$, then $T_C \geq n^2$.

In evaluating a VLSI design, we shall be concerned with T_C and T_D and also with R_C and R_D. We would like R_C and R_D to be close to 1. Finally, we may combine the two efficiency ratios R_C and R_D into the single ratio $R = R_C * R_D$. A design that makes effective use of the available bandwidth and processors will have R close to 1.

The efficiency measure R as defined here is the same as that used in [CHEN84a and b] to evaluate VLSI designs for matrix multiplication and back substitution. This measure is also quite similar to that proposed in [HUAN82]. In fact, the two measures become identical when $T_C = T_D$.

For each of the designs considered in this paper, we compute R_C, R_D and R. In several cases, our designs have improved efficiency ratios than all earlier designs using the same model. In comparing different architectures for the same problem, one must be wary about over emphasizing the importance of R_C, R_D and R. Clearly, using $P = 1$ and $B = 1$, we can get $R_C = R_D = R = 1$ and no speed up at all. So, we are really interested in minimizing T_C and T_D while keeping R close to 1. Some of our designs reduce T_C at the expense of R.



(a) Chain

(b) Ring

(c) Mesh

(d) Hexagonal Array

(e) Broadcast Mesh

Figure 1.1

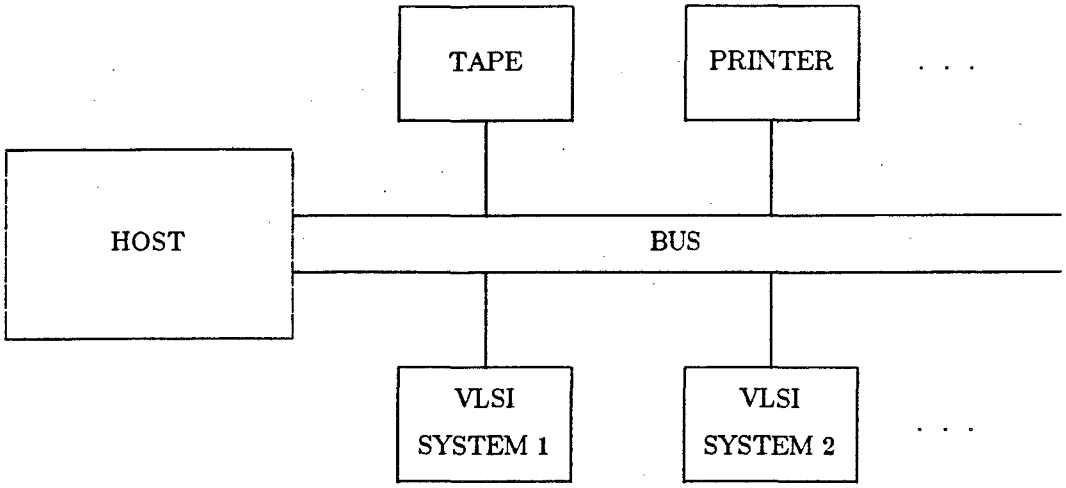


Figure 1.2

2. LU Decomposition

2.1. The Problem

Input: An $n \times n$ matrix A whose LU decomposition can be performed by Gaussian elimination without pivoting.

Output: A lower triangular matrix L and an upper triangular matrix U such that $A = LU$.

Parameters: $C \sim n^3/3$, $D = 2n^2$.

2.2. O(n) Bandwidth Mesh

Kung and Leiserson, [KUNG78], have proposed an n^2 PE hexagonal array to obtain the LU decomposition of an $n \times n$ matrix. Their design has $B = 4n/3$ and $T_C = T_D = 4n$. Consequently, $R_C \sim 12$, $R_D \sim 8/3$ and $R \sim 32$. Improved performance results if we use only four of the six connections that each PE has. In this case, the hexagonal array reduces to a mesh. We need $n(n-1)$ PEs.

Assume that the $n(n-1)$ PEs are arranged in $n-1$ rows as in Figure 2.1. The array is initially loaded with the a_{ij}s such that PE(i,j) contains a_{ij} in its A register, $A(i,j)$. This requires $n-1$ data move steps if the input is provided from the top by rows or n data move steps if the input is provided from the left by columns.

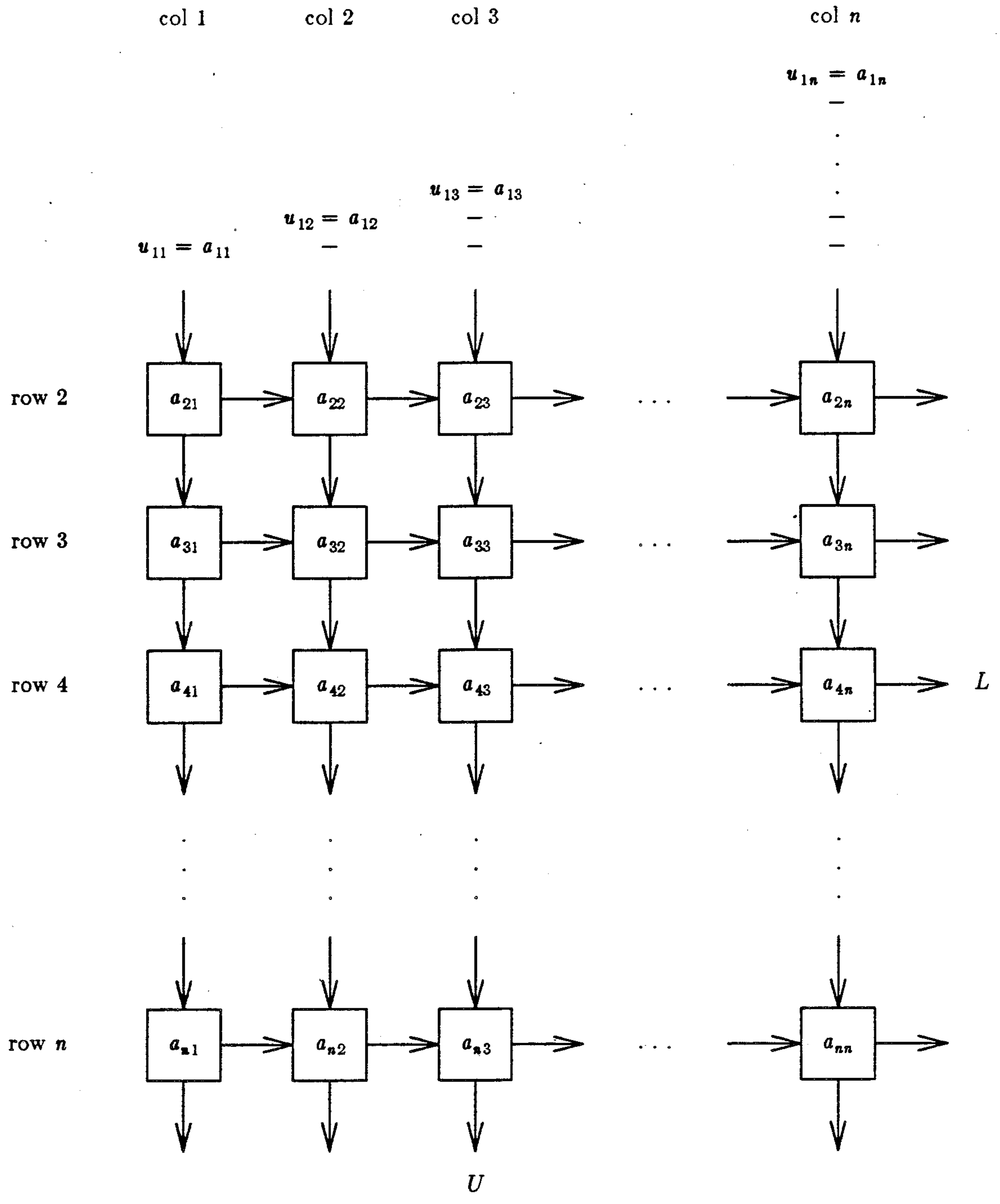

col 1
col 2
col 3
col n
$u_{1n} = a_{1n}$
$u_{13} = a_{13}$
$u_{12} = a_{12}$
$u_{11} = a_{11}$
row 2
a_{21}
a_{22}
a_{23}
a_{2n}
row 3
a_{31}
a_{32}
a_{33}
a_{3n}
row 4
a_{41}
a_{42}
a_{43}
a_{4n}
L
row n
a_{n1}
a_{n2}
a_{n3}
a_{nn}
U

Figure 2.1

The data movement pattern for each PE is as in Figure 2.2. PEs in column 1 receive no input from the left. Data move and computation steps alternate. The sequence begins with a data move step.

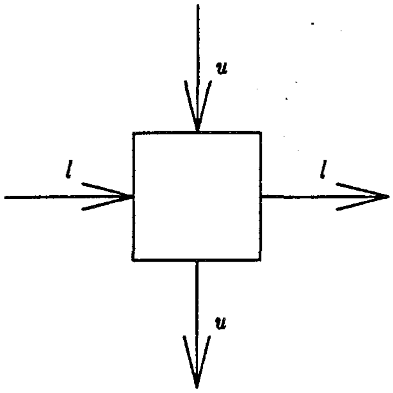


Figure 2.2

PE(i,j) computes l_{ij} if $i > j$ and u_{ij} if $i \leq j$. PE(i,j) performs its first computation after ($i + j - 2$) data move steps from the initial configuration of Figure 2.1. Thus, PE(2,1) starts after 1 data move. In this data move, $u_{11} = a_{11}$ is input from the top. In the first computation, PE(2,1) computes $l_{21} = a_{21}/u_{11}$. No other PE computes at this time. Next, there is a data move step. l_{21} is transmitted to PE(2,2) and u_{11} to PE(3,1). At the same time, PE(2,2) inputs $u_{12} = a_{12}$.

PEs (2,2) and (3,1) begin computing after 2 data move steps. At this time, PE(3,1) computes $l_{31} = a_{31}/u_{11}$ and PE(2,2) computes $u_{22} = a_{22} - l_{21}u_{12}$. PE(2,1) is idle at this time. This computation step is followed by a data move step in which PE(4,1) receives u_{11} from PE(3,1), PE(3,2) receives l_{31} from PE(3,1) and u_{12} from PE(2,2), and PE(2,3) inputs u_{13} and receives l_{21} from PE(2,2). PE(3,2) will receive u_{22} from PE(2,2) in the next data move step.

In the third computation step, PE(4,1) computes $l_{41} = a_{41}/u_{11}$; PE(3,2) computes $A(3,2) = a^2_{32} = a^1_{32} - l_{31}u_{12}$; and PE(2,3) computes $u_{23} = a^2_{23} = a^1_{23} - l_{21}u_{13}$.

Once a PE has computed an l or a u value, it stops computing. The l value computed is transmitted to its right in the next data move step, while the u value computed is transmitted down in the second data move step. The first data move step is used to transmit the u value used to compute the new u value. The algorithm is formally given in Algorithm 2.1.

The correctness of the above scheme may be established by induction. Essentially, we need to show that PE(i,j) receives l_{ik} and u_{kj} during the ($i + j + k - 3$)th data move step for

```
for $h \leftarrow n$ downto 2 do
    do in parallel    {set up initial configuration}
        $A(2,j) \leftarrow a_{kj}$,            $1 \leq j \leq n$
        $A(i,j) \leftarrow A(i-1,j)$,   $3 \leq i \leq n$,  $1 \leq j \leq n$
    end
end
for $h \leftarrow 1$ to $3n-4$ do
    {$m_1 = \min\{i,j\}$,    $m_2 = \min\{i-1,j\}$,    $p = i+j-2$}
    do in parallel    {$u(1,j)$ is input, $u(n+1,j)$ and $l(i,n+1)$ are output}
        $u(i,j) \leftarrow u(i-1,j)$,         $2 \leq i \leq n+1$,  $1 \leq j \leq n$,  $p \leq h \leq p+m_2-1$
        $l(i,j) \leftarrow l(i,j-1)$,         $2 \leq i \leq n$,  $2 \leq j \leq n+1$,  $p \leq h \leq p+m_1-2$
        $u(i,j) \leftarrow A(i,j)$,           $2 \leq i \leq j \leq n$,  $h = p+m_2$
    end
    do in parallel
        $l(i,j) \leftarrow A(i,j)/u(i,j)$,       $1 \leq j < i \leq n$,  $h = i+2j-3$
        $A(i,j) \leftarrow A(i,j) - l(i,j)u(i,j)$,  $2 \leq i,j \leq n$,  $p \leq h \leq p+m_1-2$
    end
end
do in parallel
    $u(n+1,n) \leftarrow u(n,n)$     {output $u_{n-1,n}$}
    $l(n,n+1) \leftarrow l(n,n)$     {output $l_{n,n-1}$}
    $u(n,n) \leftarrow A(n,n)$
end
$u(n+1,n) \leftarrow u(n,n)$     {output $u_{nn}$}
```

Algorithm 2.1

$k < \min\{i,j\}$ and it only receives u_{kj} in the $(i+j+k-3)$th data move step for $k = j$ where $i > j$.

Performance

The performance figures for the $(n-1) \times n$ mesh are $P = n(n-1)$, $B = n$, $T_C = 3n-4$, $T_D = 4n-3$, $R_C \sim 9$, $R_D \sim 2$ and $R \sim 18$. In comparing these figures with those for the design of [KUNG78], we note the following:

1. Our performance figures are better.
2. The mesh architecture is easier to realize.
3. Individual PEs are simpler for the design of [KUNG78].
4. Both designs can be pipelined to solve many LU decomposition problems back to back. The pipelining is easier for [KUNG78].

2.3. O(1) Bandwidth Mesh

The design is essentially the same as that for the O(n) bandwidth mesh. Computations proceed in exactly the same way. PE(2,1) has the single I/O port. The a_{ij}s are initially input through this PE. This takes $n(n-1)$ data moves. The u_{1i}s are also input before computation is initiated. This takes another n data moves. The l_{ij}s and u_{ij}s are output in the end. This takes $n(n-1)$ time if the u_{1i}s $= a_{1i}$ are not output and n^2 time if they are. Our assumption on D requires them to be included in the count for T_D. In addition, there is a total of $3n-4$ data move steps during the computation. So, $B = 1$ and $T_D = 2n^2 + 3n - 4 \sim 2n^2$. T_C and P are unchanged from the O(n) bandwidth case. Hence, $R_C \sim 9$, $R_D \sim 1$ and $R \sim 9$. By comparison, the O(1) bandwidth mesh design of [HORO79] has $P \sim (n+1)^2$, $B = 2$, $T_C \sim 5n$, $T_D \sim 2n^2$, $R_C \sim 15$, $R_D \sim 2$ and $R \sim 30$.

2.4. Broadcast Mesh with O(n) Bandwidth

The broadcast capability permits a very straight forward and efficient implementation of the recurrence (1) for LU decomposition. Our first design, Figure 2.3, employs $n(n-1)$ PE and $2n$ broadcast lines. The operation of the broadcast mesh is described in Algorithm 2.2.

In the first **for** loop, the a_{ij}s are input into the mesh row by row. The column broadcast lines are used for this purpose. The input time is the same as for the case when no broadcast lines are available (Section 2.2). The configuration following this input is shown in Figure 2.3. The next **for** loop simply goes through the necessary $n-1$ values of k of the recurrence (1). For each value of k, the u_{kk} value is broadcast to all PEs in column k and rows i, $i > k$. It is also output at this time. This broadcast takes 1 unit of time. Following the broadcast, column k can compute the l_{ik} values. The two broadcast steps in the **parallel do** are performed concurrently. The u and l values needed for the computation of the a^{k+1}s are broadcast to the PEs that need them. Also, one row of us and a column of ls are output. The parallel broadcast step is followed by the computation of the a^{k+1}s. The correctness of the algorithm follows from the correctness of the recurrence (1). The performance figures are $P = n(n-1)$, $B = n$, $T_C = 2n - 2$, $T_D = 3n - 2$, $R_C \sim 6$, $R_D \sim 1.5$ and $R \sim 9$.

```
for $i \leftarrow n$ downto 2 do    {input $A$}
    broadcast $a_{ij}$ to PE($i$,$j$),   $1 \le j \le n$
end
for $k \leftarrow 1$ to $n - 1$ do
    broadcast $a_{kk} = u_{kk}$ to PE($i$,$k$) and to output,   $i > k$
    $l_{ik} \leftarrow a_{ik} / u_{kk}$ ,   $i > k$
    do in parallel
        broadcast $a_{kj} = u_{kj}$ down and to output,   $j > k$
        broadcast $l_{ik}$ right and to output,   $i > k$
    end
    { compute $a_{ij}^{k+1}$ }
    $a_{ij} \leftarrow a_{ij} - l_{ik}\, u_{kj}$,   $i > k$ and $j > k$
end
output $u_{nn} = a_{nn}$
```

Algorithm 2.2

The same performance can be obtained by using the column broadcast lines alone. The horizontal broadcast lines are replaced by horizontal inter PE connections as in the case of a mesh. The PE interconnection pattern together with the configuration after the initial input stage are shown in Figure 2.4. The algorithm is given in Algorithm 2.3.

The broadcast capability may be further limited as in Figure 2.5. Now, each broadcast line spans at most k PEs for some k, $1 \le k \le n - 1$. As before, the broadcast lines are on columns only. Figure 2.5 shows the case for $k = 3$. The elements shown are the data inside each PE after the initial input and the two numbers in each PE give the times (after the initial input) at which that PE starts and finishes computation. Once a PE starts to compute, it will continue to compute in each succeeding time unit until it finishes the computation. When $k = 1$, the design becomes the mesh of Figure 2.1. When $k = n - 1$, it is the broadcast mesh of Figure 2.4. The algorithm used is given in Algorithm 2.4. The actual computation is similar to Algorithm 2.3 except that the computation is done in blocks of k or less processors and the same computation ripples down the column in blocks of k.

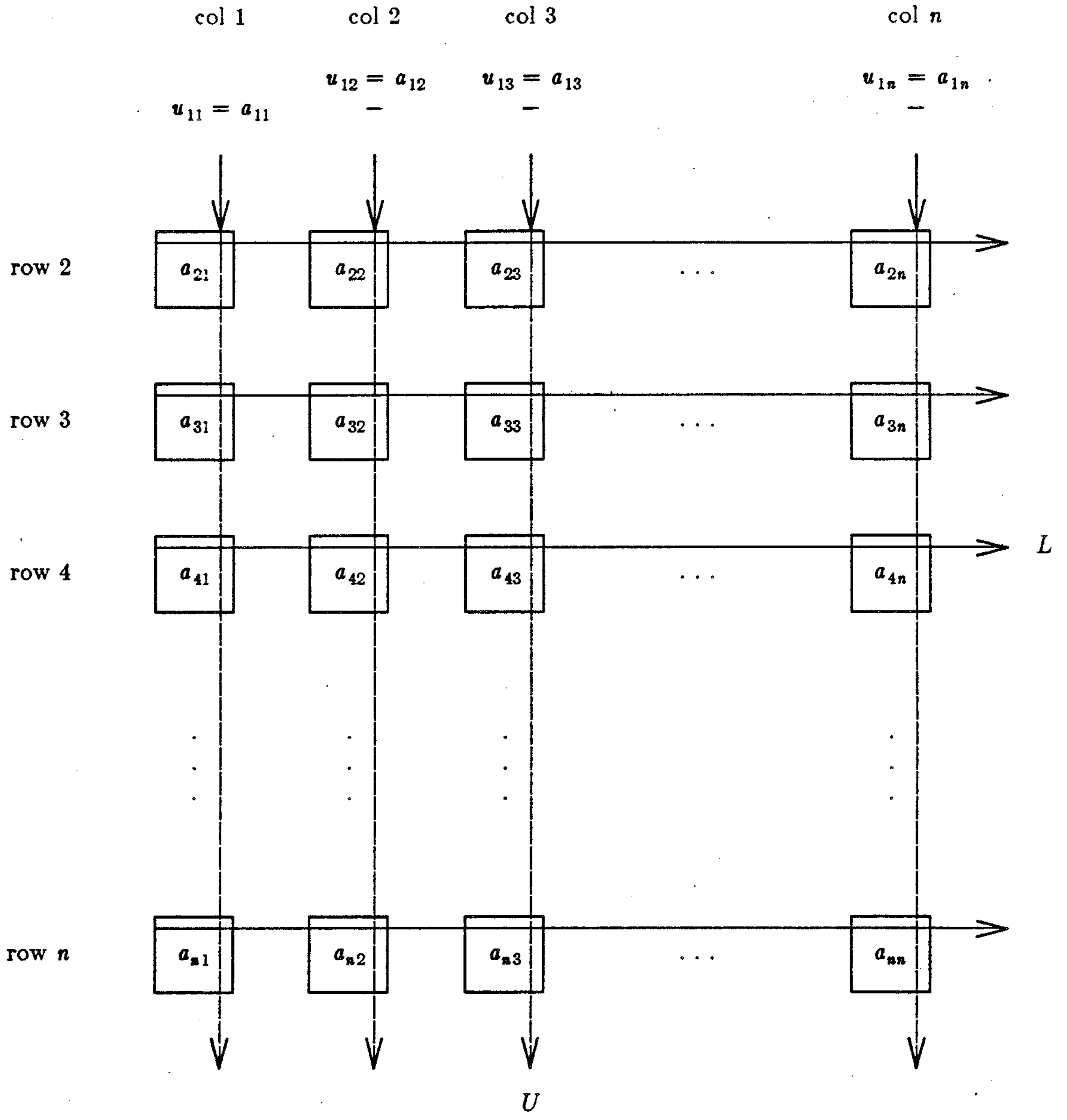


Figure 2.3

Performance

The performance figures are $P = n(n-1)$, $B = n$, $T_C = 2(n-1) + \frac{n-2}{k}$, $T_D = 3(n-1) + \frac{n-2}{k} + 1$, $R_C \sim 3(2k+1)/k$, $R_D \sim (3k+1)/(2k)$ and $R \sim 9 + \frac{15k+3}{2k^2}$.

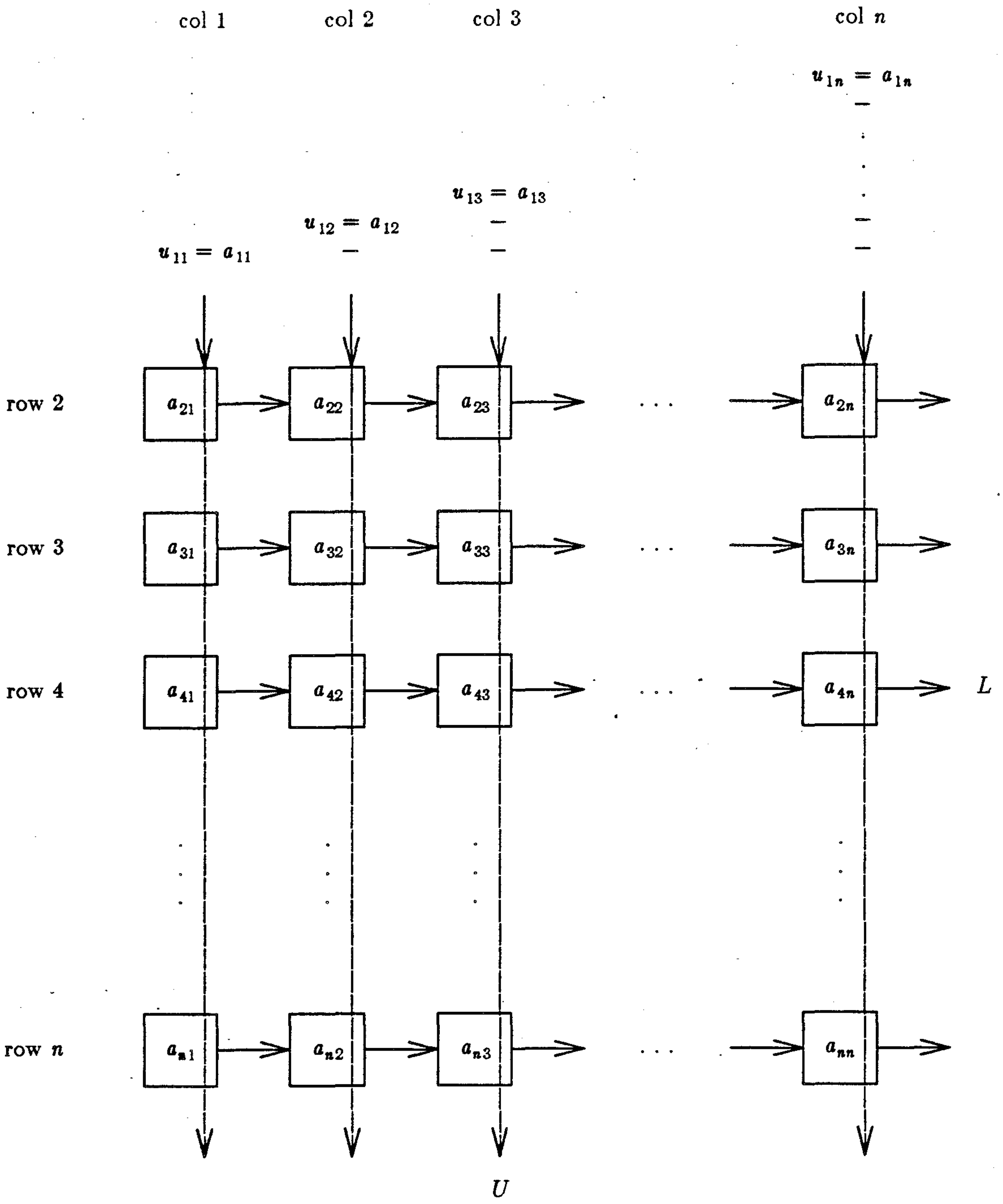

col 1
col 2
col 3
col n
$u_{1n} = a_{1n}$
$u_{13} = a_{13}$
$u_{12} = a_{12}$
$u_{11} = a_{11}$
row 2
a_{21}
a_{22}
a_{23}
a_{2n}
row 3
a_{31}
a_{32}
a_{33}
a_{3n}
row 4
a_{41}
a_{42}
a_{43}
a_{4n}
L
row n
a_{n1}
a_{n2}
a_{n3}
a_{nn}
U

Figure 2.4

```
for $i \leftarrow n$ downto 2 do    {input $A$}
    $A(i,j) \leftarrow a_{ij}, \quad 1 \le j \le n$
end
for $h \leftarrow 1$ to $2n - 2$ do
    do in parallel    {$A(1,j) = a_{1j} = u_{1j}$ is input, $u(n+1,j)$ and $l(i,n+1)$ are output}
        $u(i,j) \leftarrow A(h-j+1,j), \quad h-j+1 < i \le n+1, \quad 1 \le j \le n, \quad j \le h \le 2j-1$
        $l(i,j) \leftarrow l(i,j-1), \quad 2 \le i \le n, \quad 2 \le j \le n+1, \quad j \le h \le j + \min\{i,j\} - 2$
    end
    do in parallel
        $l(i,j) \leftarrow A(i,j)/u(i,j), \quad 1 \le j < i \le n, \quad h = 2j-1$
        $A(i,j) \leftarrow A(i,j) - l(i,j)u(i,j), \quad 2 \le i,j \le n, \quad j \le h \le j + \min\{i,j\} - 2$
    end
end
do in parallel
    $u(n+1,n) \leftarrow A(n,n)$    {output $u_{nn}$}
    $l(n,n+1) \leftarrow l(n,n)$    {output $l_{n,n-1}$}
end
```

Algorithm 2.3

2.5. O(1) Bandwidth Chain

Our design consists of $n-1$ PEs numbered 2 through n (Figure 2.6). Each PE is assumed to have enough memory to hold one row of the A matrix together with a few other values. This assumption is valid for the programmable systolic computers being designed by Kung, [KUNG83].

The working of this chain is given in Algorithm 2.5. The last $(n-1)$ rows of A are first input from the left. The input is done in such a way that PE(i) contains row i following the input. This is followed by row 1 of A. Note that $a_{1j} = u_{1j}$. Each PE retains the value u_{11} as well as one other u_{ij}. The configuration following this input is given in Figure 2.6. Specifically, PE(i), $i \ge 2$, contains row i of A, u_{11} and u_{1i}. The total time needed to complete the input is n^2.

At the start of each iteration of the outer **for** loop, PE(i), $i > k$, contains a_{ij}^{k}, $j \ge k$, u_{ki} and u_{kk}. Hence, column k of L may be computed in one unit of time. PE(i) computes l_{ik},

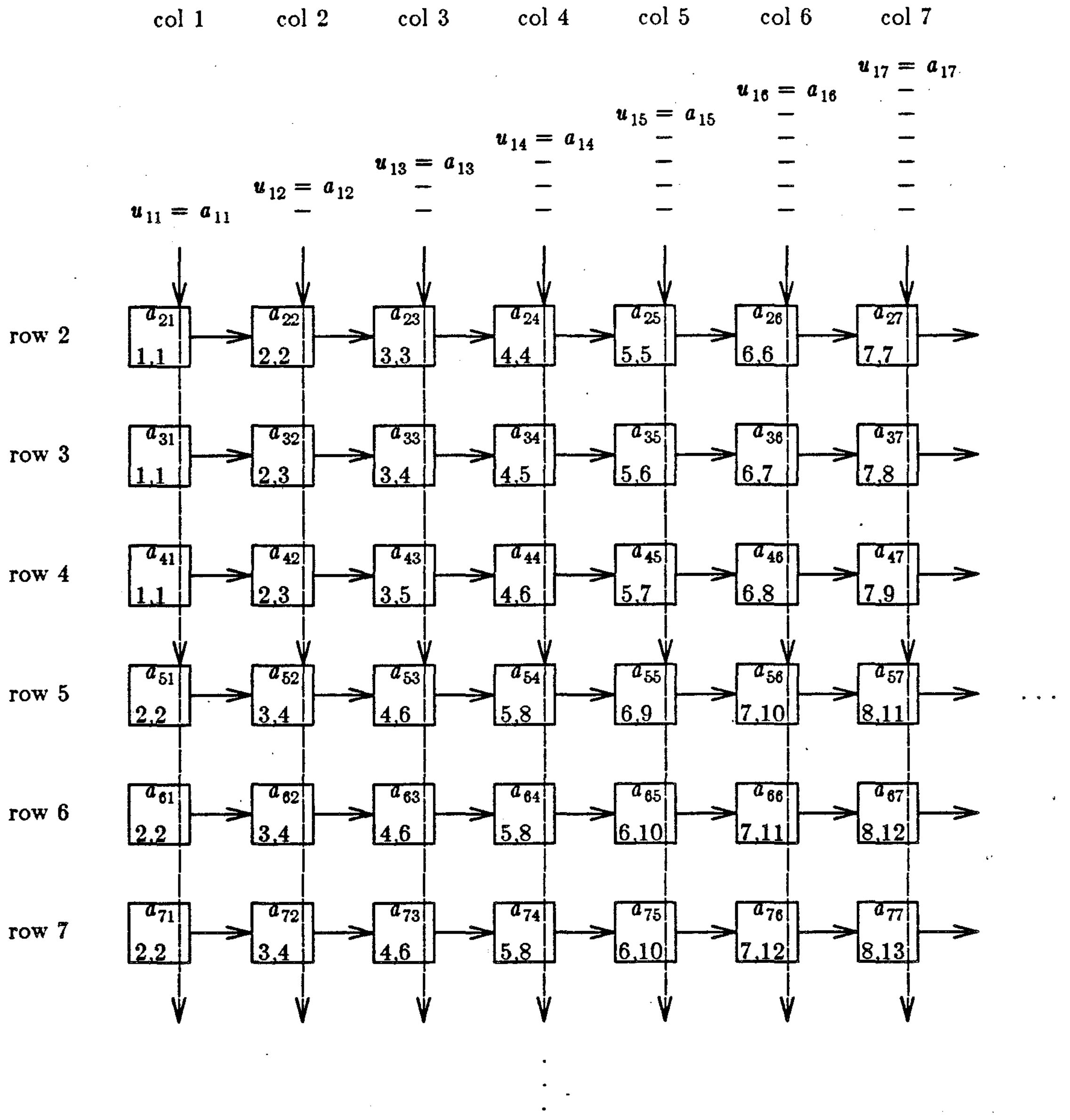


Figure 2.5

$i > k$. In the $n - k$ iterations of the inner **for** loop, PE(i), $i > k$, computes a_{ij}^{k+1}. These are computed in the order:

$$j = i,\ i - 1,\ \cdots\ ,\ k + 1,\ n,\ n - 1,\ \cdots\ ,\ i + 1.$$

To compute a_{ij}^{k+1}, we need the values a_{ij}^{k}, l_{ik} and u_{kj}. a_{ij}^{k} and l_{ik} are in PE(i). u_{kj} is in PE(i) only when $j = i$ (i.e., $p = 0$). For succeeding values of p, the right u_{kj} is in the PE on the left.

```
for $i \leftarrow n$ downto 2 do     $\{$input $A\}$
    $\{\ p = \left\lfloor \frac{i-2}{k-1} \right\rfloor,\quad q = \left\lceil \frac{n-1}{k} \right\rceil \}$
    do in parallel     $\{A(1,j) = a_{ij}$ is input$\}$
        $A((h+1)k+1,j) \leftarrow A(hk+1,j), \qquad 0 \le h < p$
        $A(i+h,j) \leftarrow A(hk+1,j), \qquad h = p < q$
    end
end
for $h \leftarrow 1$ to $2(n-1) + \left\lfloor \frac{n-2}{k} \right\rfloor$ do     $\{2 \le i \le n\}$
    $\{\ p = \left\lceil \frac{i-1}{k} \right\rceil - 1,\quad q = p + j,\quad r = (h-j)k+1,\quad s = h - j - p + 1\ \}$
    do in parallel
        $\{\ A(1,j)$ is input, $l(i,n+1)$ is output
          if $u(n,j)$ is involved, it is assumed that output is also involved $\}$
        $u(i,j) \leftarrow u(r,j), \qquad r < i \le r+k,\quad 1 \le j \le n,\quad q \le h < q + \min\{pk,j\}$
        $u(i,j) \leftarrow A(s,j), \qquad s < i \le (p+1)k+1,\quad 1 \le j \le n,\quad q + pk \le h \le q + j - 1$
        $l(i,j) \leftarrow l(i,j-1), \qquad 2 \le j \le n+1,\quad q \le h \le q + \min\{i,j\} - 2$
    end
    do in parallel
        $l(i,j) \leftarrow A(i,j)/u(i,j), \qquad 1 \le j < i \le n,\quad h = p + 2j - 1$
        $A(i,j) \leftarrow A(i,j) - l(i,j)u(i,j), \qquad 2 \le j \le n,\quad q \le h \le q + \min\{i,j\} - 2$
    end
end
do in parallel
    output $A(n,n)$     $\{$output $u_{nn}\}$
    output $l(n,n)$     $\{$output $l_{n,n-1}\}$
end
```

Algorithm 2.4

Note that PE(k) has all the u_{kj}s as $u_{kj} = a_{kj}^{k}$. Also, when $k = 1$, PE(1) denotes the input stream.

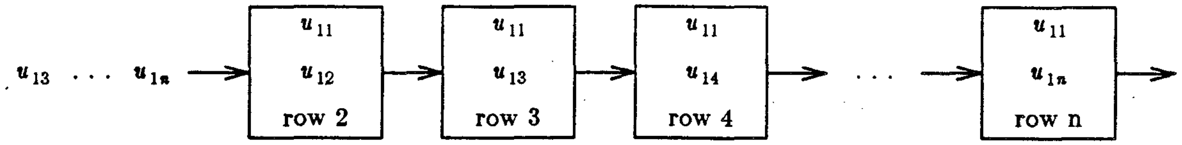


Figure 2.6

In order to satisfy the loop invariant for the outer **for** loop, it is necessary to set up the $u_{k+1,j}$s needed for the next k. These are computed in PE(k + 1) and transmitted rightwards. This requires each PE to have two additional registers to save the values of $u_{k+1,k+1}$ and a $u_{k+1,j}$.

Performance

When the outer loop terminates, the u and l values may be extracted in $\sim n^2$ time from the right. The performance figures are seen to be $P = n - 1$, $B = 1$, $T_C \sim n^2/2$, $T_D \sim 5n^2/2$, $R_C \sim 3/2$, $R_D \sim 5/4$ and $R \sim 15/8$.

The chain architecture can be used even when the amount of memory available per PE is O(1). In this case, the A matrix needs to be input n times and T_D becomes O(n^3). Hence, R is O(n).

2.6. Summary

The performance figures of the various VLSI architectures for the LU decomposition problem are summarized in Table 1. As can be seen, our designs represent an improvement over earlier designs for similar architectures. For instance, our mesh design has an R of 18 while the hexagonal design of [KUNG78] has an R of 32 and our O(1) bandwidth design has an R of 9 while that of [HORO79] has an R of 30.

```
input the last $(n - 1)$ rows of $A$ such that row $i$ is in PE($i$), $2 \leq i \leq n$
input $u_{11} = a_{11}$ such that $u_{11}$ is in each PE
input $u_{1j} = a_{1j}$ such that $u_{1j}$ is in PE($j$), $2 \leq j \leq n$
for $k \leftarrow 1$ to $n - 1$ do     {compute $l_{ik}$ and $a_{ij}^{k+1}$}
    {PE($i$) computes $l_{ik}$, $k < i \leq n$}
    $l_{ik} \leftarrow a_{ik}/u_{kk}$,   $k < i \leq n$
    {compute $a_{ij}^{k+1}$, $j \geq k + 1$ in PE($i$), $i \geq k + 1$}
    for $p \leftarrow 0$ to $n - k - 1$ do
        { PE($i$) computes $a_{i,f(i-p)}^{k+1}$, $i \geq k + 1$ where
          $f(i - p) = \begin{cases} i - p, & i - p > k \\ n - k + i - p, & \text{otherwise} \end{cases}$ }
        $a_{i,f(i-p)} = a_{i,f(i-p)} - l_{ik} * u_{k,f(i-p)}$
        do in parallel
            {shift $u$ values right}
            shift $u_{kj}$ right by one PE for PE($i$), $i > k$
            PE($k + 1$) gets $u_{k,f(i-p+1)}$ from PE($k$) or from input if $k = 1$
            {shift $u$ values for next $k$}
            PE($k + 1$) sends $u_{k+1,f(i-p)} = a_{k+1,f(i-p)}^{k+1}$ rightwards
        end
    end
end
output $U$ and $L$
```

Algorithm 2.5

Performance	Architecture				
	Hexagonal With O(n) Bandwidth	Mesh With O(n) Bandwidth	Mesh With O(1) Bandwidth		Ring With O(n) Bandwidth
	[KUNG78]		[HORO79]	Our	[KUNG84]
P	n^2	$n(n-1)$	$(n+1)^2$	$n(n-1)$	$n^2/3$
B	$4n/3$	n	2	1	$5n/3$
T_C	$4n$	$3n-4$	$5n$	$3n-4$	$3(n-1)$
T_D	$4n$	$4n-3$	$2n^2$	$2n^2$	$4(n-1)$
R_C	12	9	15	9	3
R_D	8/3	2	2	1	10/3
R	32	18	30	9	10

Performance	Architecture			
	Chain With O(1) Bandwidth		Broadcast Mesh With O(n) Bandwidth	
	With O(n) Memory	With O(1) Memory	With O(n) Span	With O(k) Span
P	$n-1$	$n-1$	$n(n-1)$	$n(n-1)$
B	1	1	n	n
T_C	$n^2/2$	$n^2/2$	$2n-2$	$2(n-1)+\frac{n-2}{k}$
T_D	$5n^2/2$	$O(n^3)$	$3n-2$	$3n-2+\frac{n-2}{k}$
R_C	3/2	3/2	6	$\frac{3(2k+1)}{k}$
R_D	5/4	$O(n)$	3/2	$\frac{3k+1}{2k}$
R	15/8	$O(n)$	9	$9+\frac{15k+3}{2k^2}$

$C \sim n^3/3,\ D = 2n^2$

Table 1

3. References

[CHEN84a] K.H. Cheng and S. Sahni, *VLSI Systems For Matrix Multiplication,* Department of Computer Science, University of Minnesota, **August 1984.**

[CHEN84b] K.H. Cheng and S. Sahni, *VLSI Architectures For Back Substitution,* Department of Computer Science, University of Minnesota, **November 1984.**

[HORO79] E. Horowitz, *VLSI architectures for matrix computations,* IEEE International Conference On Parallel Processing, **1979,** pp. 124-127.

[HUAN82] K.H. Huang and J.A. Abraham, *Efficient parallel algorithms for processor arrays,* IEEE International Conference On Parallel Processing, **1982,** pp. 271-279.

[KUNG78] H.T. Kung and C.E. Leiserson, *Systolic arrays for VLSI,* Department of Computer Science, Carnegie-Mellon University, **April 1978.**

[KUNG79] H.T. Kung, *Let's design algorithms for VLSI systems,* Proceedings CALTECH Conference on VLSI, **Jan. 1979,** pp. 65-90.

[KUNG83] H.T. Kung, *A Listing of Systolic Papers,* Department of Computer Science, Carnegie-Mellon University, **May 1984.**

[KUNG84] H.T. Kung and M. Lam, *Wafer scale integration and two level pipelined implementations of systolic arrays,* Journal of Parallel and Distributed Processing, Vol. 1, #1, **1984**

[LEIS83] C.E. Leiserson, *Area-Efficient VLSI Computation,* MIT Press, **1983.**

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

ELECTRONIC STRUCTURE OF TERNARY SEMIMAGNETIC SEMICONDUCTORS

Microelectronic and Information Sciences Center

Technical Report #28

A. Franciosi
Department of Chemical Engineering and
Materials Science
University of Minnesota

S. Chang
Department of Chemical Engineering and
Materials Science
University of Minnesota

C. Caprile
Department of Chemical Engineering and
Materials Science
University of Minnesota

R. Reifenberger
Department of Physics
Purdue University

U. Debska
Department of Physics
Purdue University

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

ELECTRONIC STRUCTURE OF TERNARY
SEMIMAGNETIC SEMICONDUCTORS

MEIS Technical Report #28

A. Franciosi
S. Chang
C. Caprile
R. Reifenberger
U. Debska

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

[illegible]

Abstract

The electronic structure of the ternary semiconductur alloys $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xSe$ was examined through synchrotron radiation photoemission studies of single crystals cleaved in situ for $0 \leq x \leq 20\%$. Comparison with the parent binary compounds HgSe, CdSe and ZnSe and resonant photoemission at the 3p-3d transition energy indicate that a Mn-derived density of states feature appears 3.5 ± 0.1 eV below the valence band maximum and exhibits elemental 3d character with no evidence of important hybridization effects. The constant binding energy of the Mn 3d states in all of the compounds explored forces a reevalution of existing models of bonding in ternary magnetic semiconductors.

Ternary magnetic semiconductors have potential application in a variety of devices[1-2] that exploit their composition-dependent transport parameters or their novel magneto-optical and magneto-transport properties.[3] However, far too little is known about the basic electronic and structural characteristics of these materials to fully exploit their potential. For example, binding energy and hybridization of the electronic states responsible for the magnetic properties are still subject to debate. In this letter we summarize systematic studies of the ternary semiconductor alloys $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$, and $Zn_{1-x}Mn_xSe$. We used synchrotron radiation photoemission spectroscopy to directly probe the density of electronic states of the materials and the different orbital contribution to the alloy bonding. To our knowledge, this is the first study of this kind ever reported for these semiconductor materials.[4] Our results show that the Mn-derived 3d states which are responsible for the semiconductor magneto-optical and magneto-transport properties lay 3.4-3.5 eV below the valence band maximum in all alloys, irrespective of composition. This is in sharp contrast with earlier estimates of the 3d binding energy based on simplified pictures of alloy bonding.[5] We also conclude, on the basis of Mn 3d photoemission cross section studies, that there is no evidence, in these compounds, of the important Se p-Mn d hybridization reported for $Cd_{1-x}Mn_xTe$.[6]

The single crystals used in this study were grown at Purdue University using a modified Bridgman method. Several posts (4x4x15mm) were cut and characterized through x-ray diffraction and standard x-ray microprobe analysis. The samples were loaded into a photoelectron spectrometer at operating pressure of about $5x10^{-11}$ torr by means of a fast-entry load lock and cleaved in situ with a variable degree of success, ranging from highly disordered stepped surfaces to flat natural cleavage surface. The data presented here appear independent of the cleave quality and provide information on the bulk electronic structure of these materials rather than on the surface morphology. The photoemission measurements were made at photon energies of

$40 \leq h\nu \leq 140$ eV using a grazing incidence "grasshopper" monochromator and the 240 MeV electron storage ring Tantalus at the Synchrotron Radiation Center of the University of Wisconsin-Madison. Photoelectron Energy Distribution Curves (EDC´s) were recorded by means of a double pass cylindrical mirror analyzer. The overall energy resolution (electron plus photons) was typically 0.4-0.6 eV for the data presented here.[7]

Representative EDC´s for $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xSe$ with x=0.15, 0.21 and 0.20 respectively, are shown in Fig. 1 at $h\nu=70$ eV. The binding energies are referred to the valence band maximum E_v, calculated by linearly extrapolating the valence band edge at low binding energies. The spectra have been approximately normalized to the main emission feature, and are given in arbitrary units. The main emission features in Fig. 1 correspond to the shallow Hg $5d_{5/2}$ and Hg $5d_{3/2}$ (top) and to the unresolved Cd 4d (mid-section) and Zn 3d (bottom) doublets. The valence states within 6 eV of E_v exhibit important differences as compared to the EDC´s of the parent binary compounds HgSe, CdSe and ZnSe. This can be seen in Fig. 2 where we directly compare the low binding energy valence states for the ternary alloys (solid line) and the binary compounds (dashed line). The Mn concentration in the alloys is the same than for the samples of Fig. 1. The EDC´s are shown at $h\nu=70$ eV after subtraction of a smooth secondary background and normalization to the Se-p derived emission feature at 1.5-2.0 eV. The spectra for the binary compounds are in good agreement with the HeI and HeII results of Shevchik et al.[8] and exhibit density of states features reflecting primarily Se-p character (within 3.5 eV of E_v) and metal s-character (3.5-6.0 eV of Ev). The spectra for the ternary alloys show a dramatically increased emission in the 2.5-4.5 eV binding energy range. To emphasize such differences, we subtracted the binary compound EDC from the ternary alloy spectrum. The results after smoothing are shown in Fig. 2 by a dot dashed line. In all cases the difference curve consists of a symmetric line centered 3.5±0.1 eV below E_v and with a Full Width at Half Maximum (FWHM) of 1.1-1.2 eV. This result at $h\nu=70$ eV is substantiated by difference curves at photon energy 40<hv<140 eV which all produce a similar line,[7] the only differences being a variation in the intensity of the 3.5 eV feature and an increase of FWHM at high photon energies that reflects the variation in the experimental energy resolution.[7] We therefore relate the observed 3.5 eV structure to a Mn-derived Density of States (DOS) feature. The intrinsic FWHM of this feature, after deconvolution of the experimental gaussian broadening, is estimated at 0.4-0.5 eV in all of the ternary alloys.[7]

Resonant photoemission at the Mn 3p-3d transition energy was used to identify the Mn 3d character in the valence DOS. Both for Mn metal and for atomic Mn emission from the 3d states is first reduced (antiresonance) and then enhanced (resonance) when the photon energy is swept through the 3p-3d transition energy.[7,9] This is shown in the top section of Fig. 3 where we plot the optical absorption coefficient for the Mn metal, from Ref. 10. At hv=40 eV the combination of the centrifugal barrier for d emission and of the Mn 3p-3d antiresonance greatly diminishes the Mn 3d emission. At resonance (hv=51 eV) the photoexcitation probability exhibits a three-fold enhancement. The resulting 3p-3d absorption line is consistent with a Fano resonance.[9] For the ternary alloys of Fig. 1-2 we calculated the integrated intensity of the 3.5 eV feature normalized to the monochromator output. The results for $Hg_{0.85}Mn_{0.15}Se$ and $Cd_{0.79}Mn_{0.21}Se$[11] are shown in the lower section of Fig. 3 as a function of photon energy. The three-fold enhancement observed demonstrates the Mn 3d orbital character of the electron states involved. Furthermore, since the 3.5 eV feature maintains the same lineshape throughout the photon energy range and no detectable resonant behavior of the other DOS features was found, we have no evidence that relevant Mn 3d hybridization takes place in any of the ternary alloys. This is in contrast with the reported behavior of Mn 3d levels in $Cd_{1-x}Mn_xTe$, where no evidence was found of a highly localized $3d^5$ ground state and strong hybridization was reported between the Mn^{++} d-levels and the primarily Te-derived valence bands.[6,12] This intriguing discrepancy, if confirmed, would indicate that the radial extension of the Te 5p orbitals, as opposed to the Se 4p orbitals, has an important role in determining the degree of p-d superposition and hybridization. Finally, we summarize in Table I the observed binding energies[13] for the Mn 3d levels in the alloys (column II) relative to E_v[14], an effective work function W (the sum of the electron affinity plus gap energy) from Ref. 8 appropriate for the binary parent compounds (column III), the estimate of the 3d binding energy proposed by Webb et al. in Ref. 5 (Column IV), and the difference between the actual binding energy and the estimated value (column V). The estimates proposed in Ref. 5 were obtained from an experimentally determined 3d binding energy of 3.5 eV for the Mn 3d in $Cd_{1-x}Mn_xTe_x$ and the work function values W, assuming that the ionization energy for the Mn $3d^5$ level remains constant in the series at 9.7 eV. We show in the rightmost column of Table I that this assumption leads to predicted binding energies in rather poor agreement with experiment. If the discrepancies were due to an erroneous starting value for the 3d ionization energy in $Cd_{1-x}Mn_xTe$ one would expect a systematic error of constant magnitude and sign in column. This is not the case, so that the discrepancies are due either to the value of W used in each case or to the assumption of constant 3d ionization energy in the alloy series. Since the experimental 3d binding energy does not appear to vary sensibly with x in the range of composition explored,[7] and relatively low Mn concentrations were used ($x \leq 20\%$), modifications of W are unlikely to explain the discrepancies. We conclude that the assumption of constant 3d ionization energy in the series is at the origin of the problem. We suggest that the detail of the local bonding situation and the difference in electronegativities between the cation and the Mn impurities have to be examined in each case to obtain reasonable estimates of the 3d ionization energy. Theoretical calculations of the kind briefly summarized in Ref. 14 appear to give more correct predictions of the 3d parameters, and we stress the need for further and more publicized theoretical effort in this direction.

Acknowledgments

This work was supported, in part, by the Graduate School of the University of Minnesota under grant 0100-4908-32. The crystal growth at Purdue University was supported by National Science Foundation - Material Research Laboratories Grant No. DMR80-22868 and we thank J.K. Furdyna for making the crystals available to us and for stimulating discussions. We thank D. Yoder-Short and F. Pool for the microprobe analysis. Finally we are grateful to the entire staff of the University of Wisconsin Synchrotron Radiation Center (supported by NSF Grant DMR80-20164) for their expert and cheerful support.

REFERENCES

1. A.E. Turner, R.T. Gunshor and S. Datta, Appl. Opt. 22, 3152 (1983).
2. J.K. Furdyna, Proc. Int. Soc. for Optical Eng. 409, 43 (1983).
3. J.K. Furdyna, J. Appl. Phys. 53, 7637 (1982).
4. HeI and HeII photoemission results for $Cd_{1-x}Mn_xTe$ and for $Cd_{0.60}Mn_{0.40}Se$ have been published by several groups. See references 5, 6, 12, 14 of this paper.
5. C. Webb, M. Kaminska, M. Lichetensteiger, and J. Lagowski, Sol. State Commun. 40, 609 (1981).
6. P. Oelhafen, M.P. Vecchi, J.L. Freeouf, and V.L. Moruzzi, Sol. State Commun. 12, 1547 (1982).
7. Further data, including EDC´s for the core emission of the Se 3d, Mn 3p, Hg 4f, Hg 5d, Cd 4d and Zn 3d cores as a function of composition, and details on the Mn 3d lineshape will be presented in a forthcoming paper.
8. N.J. Shevchik, J. Tejeda, M. Cardona, and D.W. Langer, Phys. Stat. Sol. B59, 87 (1973).
9. L.C. Davis and L.A. Feldkamp, Sol. State Commun. 19, 413 (1976) and Phys. Rev. A17, 2012 (1978).
10. B. Sonntag, R. Haenzel, and C. Kunz, Sol. State Commun. 7, 597 (1969).
11. Similar results for $Zn_{1-x}Mn_xSe$ have not been processed. Preliminary indications suggest an analogous resonant behavior.
12. However, previous photoemission results on sputter-cleaned surfaces were interpreted as evidence of localized $3d^5$ states. See B.A. Orlowski, Phys. Stat. Sol. B95, K31 (1979) and reference 5.
13. We emphasize, however, that photoemission-determined binding energy for highly localized orbitals (f or d) may reflect relevant correlation effects, and therefore differ from the calculated ground state Hartree-Fock energies.
14. HeI results for sputter-cleaned $Cd_{0.60}Mn_{0.40}Se$ reported by B.A.

Orlowski, K. Kopalko and W. Chab, Sol. State Commun. 50, 749 (1984), were interpreted as indicating a main Mn 3d-derived feature at 3.35 eV plus some degree of p-d hybridization.

Table Caption

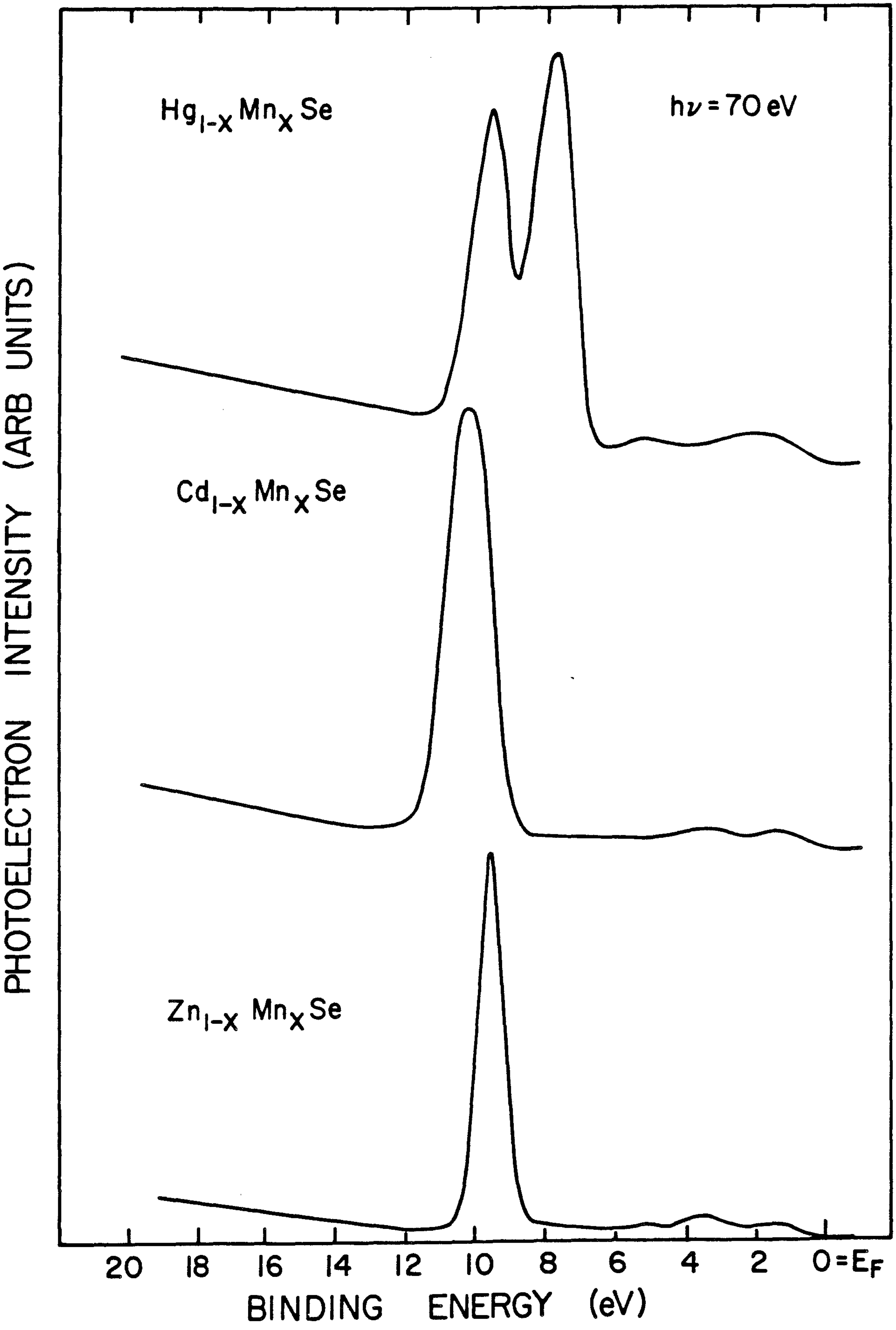
Table 1. Experimental binding energies (E_b column II) relative to the valence band maximum for $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xSe$ from this work. The value for $Cd_{1-x}Mn_xSe$ is in good agreement with the value reported in Ref. 14 for $Cd_{0.60}Mn_{0.40}Se$. The experimental values are compared with the values calculated by Webb et al[5] ($E_b{}^e$ column IV) assuming constant 3d ionization energy of 9.7 eV in the series and work function-electron affinity plus gap energy-W from ref. 8 (column III). The difference between the experimental value E_b and the estimated $E_b{}^e$ is shown in column V.

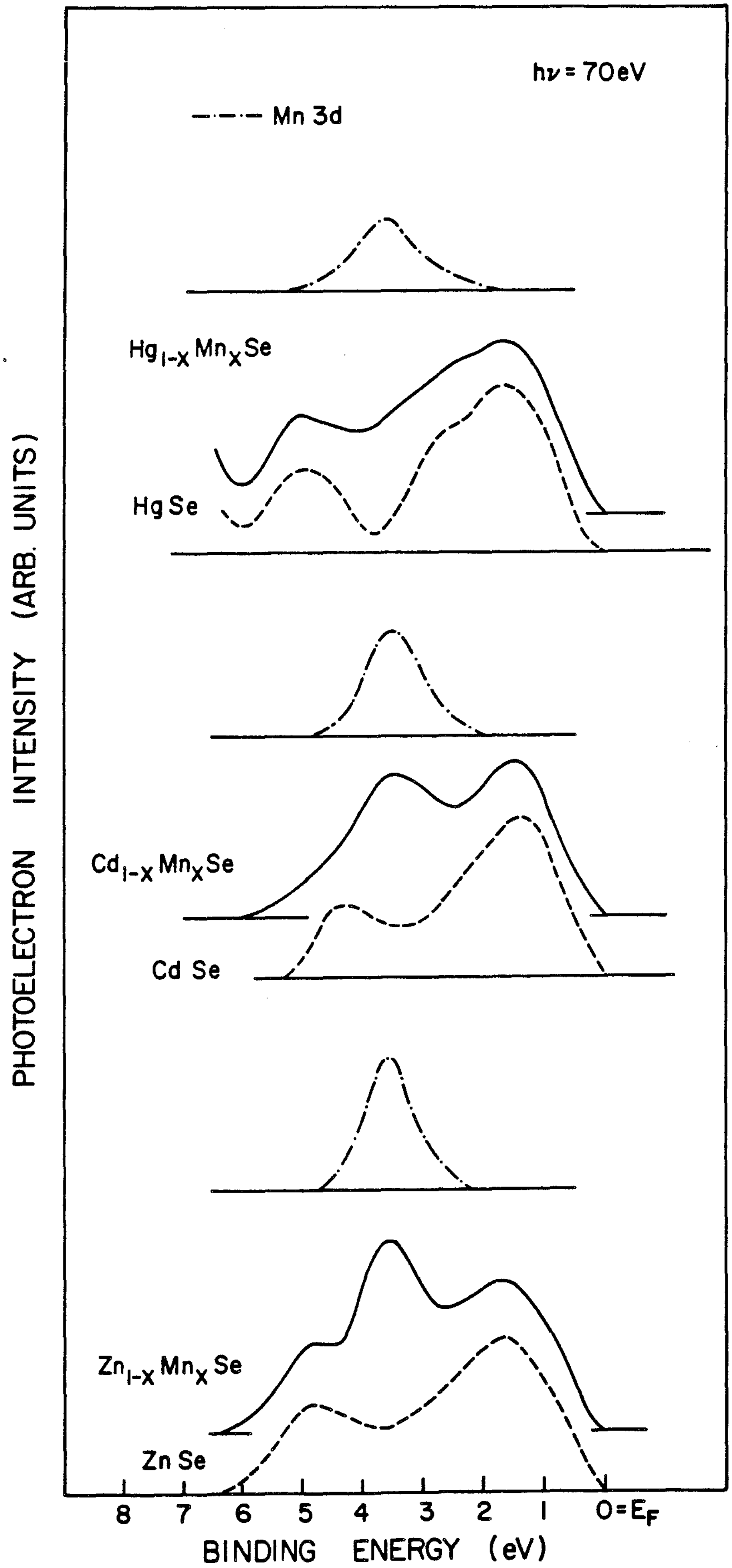
Figure Captions

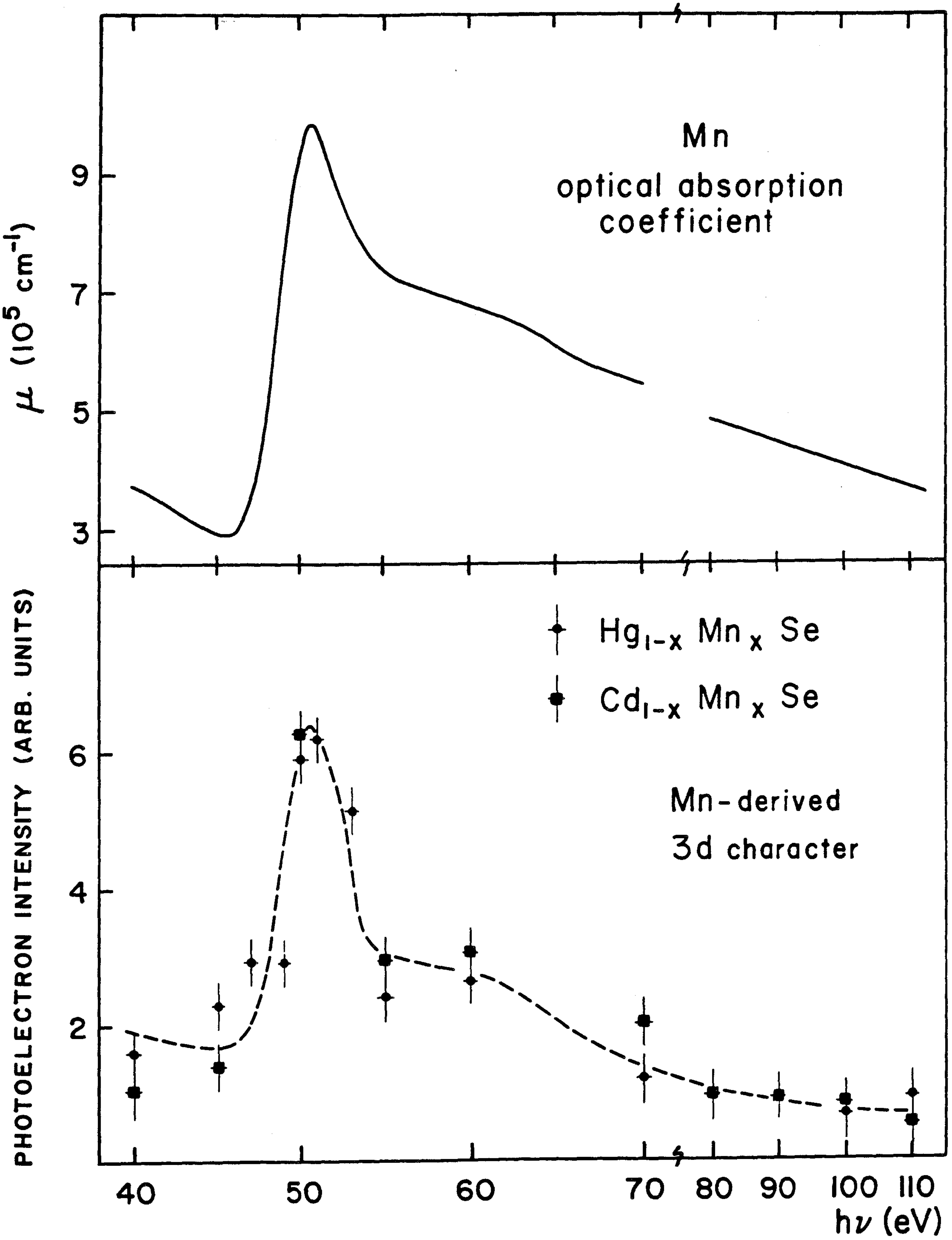
Fig. 1 Photoelectron energy distribution curves (EDC's) at hv=70 for $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xSe$ with x=0.15, 0.21, 0.20 respectively. The spectra have been normalized to the main emission features, i.e., the shallow Hg $5d_{5/2}$ and $5d_{3/2}$ cores, and the unresolved Cd 4d and Zn 3d spin-orbit doublets.

Fig. 2 Detail of the valence band emission of $Hg_{1-x}Mn_xSe$, $Cd_{1-x}Mn_xSe$ and $Zn_{1-x}Mn_xSe$ with x=0.15, 0.21 and 0.20, respectively. The EDC's after subtraction of a smooth secondary background have been normalized to the main Se-p derived emission feature of the corresponding binary compounds HgSe, CdSe and ZnSe (dashed line). The difference curves (dot-dashed line) emphasize the Mn 3d-derived contribution to the Density of Electron States (DOS).

Fig. 3 Top: optical absorption coefficient of Mn metal from Ref. 10. The sharp Fano lineshape derives from the 3p-3d resonant photoexcitation. Bottom: photon energy dependence of the Mn-derived emission feature identified in Fig. 2. We plot the integrated emission of the 3.5 eV DOS structure normalized to monochromator output. Results are given for $Hg_{1-x}Mn_xSe$ at x=0.15 (circles) and for $Cd_{1-x}Mn_xSe$ at x=0.21 (squares). The similar three-fold enhancement of the photoionization probability at resonance (hv=51 eV) as compared to anti-resonance (hv=40 eV) demonstrates the elemental Mn 3d character of the electron states involved.


$Hg_{1-x}Mn_xSe$
$h\nu = 70$ eV
$Cd_{1-x}Mn_xSe$
$Zn_{1-x}Mn_xSe$
PHOTOELECTRON INTENSITY (ARB UNITS)
20
18
16
14
12
10
8
6
4
2
0=E_F
BINDING ENERGY (eV)


hν = 70eV
Mn 3d
Hg1-x Mnx Se
Hg Se
Cd1-x Mnx Se
Cd Se
Zn1-x Mnx Se
Zn Se
PHOTOELECTRON INTENSITY (ARB. UNITS)
8
7
6
5
4
3
2
1
0=EF
BINDING ENERGY (eV)


Mn
optical absorption coefficient
μ (10^5 cm^{-1})
3
5
7
9
$Hg_{1-x}Mn_xSe$
$Cd_{1-x}Mn_xSe$
Mn-derived
3d character
PHOTOELECTRON INTENSITY (ARB. UNITS)
2
4
6
40
50
60
70
80
90
100
110
hν (eV)

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

MICROSCOPIC CONTROL OF SEMICONDUCTOR SURFACE OXIDATION

Microelectronic and Information Sciences Center

Technical Report #29

A. Franciosi
Department of Chemical Engineering
and Materials Science
University of Minnesota

S. Chang
Department of Chemical Engineering
and Materials Science
University of Minnesota

P. Philip
Department of Chemical Engineering
and Materials Science
University of Minnesota

C. Caprile
Department of Chemical Engineering
and Materials Science
University of Minnesota

J. Joyce
Materials Science Program
University of Wisconsin

MICROSCOPIC CONTROL OF
SEMICONDUCTOR SURFACE OXIDATION

MEIS Technical Report #29

A. Franciosi
S. Chang
P. Philip
C. Caprile
J. Joyce

The research reported in this paper was supported in part by the Microelectronic and Information Sciences Center.

Abstract

We explored the effect of ultrathin (0.1-10 Å) chromium overlayers on the reactivity with oxygen of Si(111) and GaAs(110) cleavage surfaces. Synchrotron radiation photoemission shows that for Cr coverages below a critical threshold coverage the overlayer does not affect substantially the oxygen adsorption rate. For chromium coverages above threshold the overlayer sharply enhances the oxygen adsorption kinetics so that most semiconductor atoms in the surface and near-surface region appear oxidized at activated oxygen exposures as low as 100 Langmuirs. The critical threshold coverage corresponds to the onset of reactive interdiffusion at the Si(111)-C and GaAs(110)-Cr interfaces. We suggest, therefore, that ultrathin Si-Cr and As-Cr reacted phases created at the surface act as activation layers for semiconductor oxidation.

Thin metal overlayers deposited on atomically clean semiconductor surfaces can change dramatically the reactivity of the surface for reactions with gas species[1-4] and metals.[5-6] We have recently shown, for example, that thin Cr overlayers on Si(111) surfaces can act both as passivating layers and as catalysts for Si(111)-Au interface reaction, so that one can control and modulate interdiffusion by varying the thickness of the Cr interlayer.[6] Similar effects have been observed by Brillson and co-workers for Al atoms at the GaAs(110)-Au interface.[5] As far as reactions with gaseous species are concerned, a few pioneering studies have addressed the effect of Ag and Au overlayers[1,3] on the oxidation of Silicon surfaces, while the presence of an Al overlayer has been shown to induce the formation of a potentially stable new oxide phase on Ge(111) surfaces.[2] An understanding of these phenomena requires a search for systematic correlations of the observed "catalytic" trends with the different local morphology of the metal/semiconductor surface layers. The goal is to establish a connection between the microscopic bonding situation of semiconductor and metal atoms at the surface and the observed specific catalytic activity.

We report here on the interaction of Si(111) and GaAs(110) surfaces with activated oxygen O_2^* in the presence of ultrathin (0.1-10Å) Cr overlayers deposited in situ. We selected the Si-Cr and GaAs-Cr systems because these are within the interfaces best characterized at room temperature.[7-8] Furthermore, chromium compounds show considerable catalytic activity for a number of chemical reactions.[9] Interaction with oxygen was selected as prototype reaction with oxidizing gas phases, and because of extensive literature existing on this subject.[10-13] Our results indicate a qualitatively similar effect of the Cr overlayer on the kinetics of Si and GaAs oxidation. While only limited oxygen adsorption is possible on the clean semiconductor surfaces at the highest activated oxygen exposures explored (10^4 Langmuirs), we see an enhancement of several orders of magnitude in the semiconductor oxidation rate for Cr coverages above a critical threshold coverage (1.3Å for Si). We relate this enhancement to the thin intermixed Si-Cr and As-Cr species that start to form at the surface at the critical threshold coverage. These intermixed species act as catalysts and dramatically increase the semiconductor oxidation kinetics. Most of the silicon, gallium and arsenic atoms in the surface and near surface layers appear oxidized at oxygen exposures as low as 100L. The major reaction product identified at the Si surface is a Si-oxide with average Si-oxygen coordination between 3 and 4. For GaAs, an As oxide similar to As_2O_3 is observed, together with a Ga-oxide phase that involves an effective Ga-oxygen coordination greater than for Ga_2O_3.

The experiments were performed on a clean Si(111)2x1 and GaAs(110)1x1 surfaces obtained through cleavage of n-type single crystals inside a photoelectron spectrometer, at operating pressure $<5x10^{-11}$ torr. Cr was deposited from a W coil at pressure $<3x10^{-10}$ torr, with overlayer thickness measured by a quartz thickness monitor. Since the clean semiconductor surfaces are relatively inert upon oxygen exposure, we elected to use activated oxygen in the pressure range 10^{-5} torr to enhance reaction kinetics.[14-15] A tungsten ionization filament was therefore positioned in line of sight of the substrate during oxygen exposure. The photoemission measurements were performed by positioning the sample at the focus of a synchrotron radiation beam and of a commercial double pass cylindrical mirror analyzer. The radiation from the 240 MeV electron storage ring Tantalus at the Synchrotron Radiation Center of the University of Wisconsin-Madison was monochromatized by means of a 3m toroidal grating monochromator in the 21-130 eV photon energy range. The overall experimental energy resolution (electrons and photons) was typically 0.3-0.4 eV for the valence band and Ga 3d core levels, and of about 0.6-0.8 eV for the As 3d and Si 2p core data.

Photoelectron Energy Distribution curves for the Ga 3d and As 3d core levels are shown in Figs. 1 and 2, respectively, as a function of O_2^* exposure ($10-10^4$L) of the clean Ga As(110) surface. The zero of the binding energy scale corresponds to initial flat-band core binding energy. Exposures up to 10^4 langmuir yield a rigid shift of the core levels that reflects the change in band bending, and attenuation of the As 3d surface contribution, visible as structure on the low binding energy side of the main As 3d line. These changes reflect the relatively low oxygen adsorption rate observed on cleaved GaAs(110) surfaces.[12-13] Even at the highest exposures explored here, the oxygen coverage is only a fraction of a monolayer.[12] The quasi-saturation value of band bending for 10^3L exposure (0.65 eV) corresponds to the value observed by Lindgren et al[12] at 10^4-10^5L exposure to ground state molecular oxygen, so that the use of activated oxygen in our case yields a 10 to 100-fold enhancement in adsorption.

Results for the Si(111) surface are summarized in Fig. 3. In the top section we show the Si 2p core emission for a clean Si(111)2x1 surface (dashed line) and for the same surface after exposure to 100L of activated oxygen (solid line). Weak oxygen-induced features appear on the high binding energy side of the main line. Vertical bars 0.9, 1.8, 2.6 and 3.5 eV below the main line mark the position of the chemically shifted Si 2p contributions associated by Hollinger and Himpsel[10] with silicon atoms bonded to 1,2,3 and 4 oxygen atoms, respectively. Further exposure to oxygen (10^3L) yields a 3 to 4-fold increase of the oxygen-induced features, that saturate in intensity and show little change[4] upon further oxygen exposure (10^4L), in agreement with the 1-1.5 monolayer oxygen saturation coverage observed in Ref. 10.

The spectra for the oxidation of the clean surface all show[4] the presence of one, two and three-fold silicon-oxygen coordination, as expected in the submonolayer and monolayer oxygen coverage range.[10] The effect of Cr overlayer on the oxygen adsorption kinetics is shown in the middle and bottom sections of Fig. 3. We distinguish two qualitatively different regions as a function of Cr coverage Θ. For Θ below a critical threshold coverage of 1.3±0.3 Å the Cr overlayer affects relatively little the silicon oxidation rate. For example, in the mid-section of Fig. 3 we show EDC's for the Si 2p emission at Θ=0.6 before (dashed line) and after exposure to 100L of activated oxygen (solid line). Deposition of 0.6Å of Cr onto the clean Si(111) cleavage surface[7] attenuates slightly the Si 2p emission with no visible lineshape changes. Exposure to 100L of activates oxygen gives rise to the same oxygen-induced feature observed for oxygen adsorption on the clean Si(111) surface.

The situation changes dramatically if Cr coverages above threshold are employed. This is shown in the bottom section of Fig. 3 for Θ=2Å. Exposure of the Cr-activated semiconductor surface to 100L of activated oxygen (solid line) yields a main oxide-induced feature centered 3.0 eV below the main line. Further oxygen exposure[4] increases this feature relative to the Si 2p substrate line, indicating that no saturation of oxygen adsorption is observed in this exposure range ($10-10^4$L). The width of the Si 2p oxide line suggests that several different oxidation states may coexist, and its binding energy, intermediate between those observed for Si-atoms locally bonded to 3 and 4 oxygen atoms,[10] shows that 3-fold and 4-fold coordination are likely to be dominant in the silicon oxidized layer.

The results of Fig. 3 indicate that Cr coverages above the critical threshold coverage change dramatically the reactivity of the semiconductor surface and enhance of several orders of magnitude the silicon oxidation rate at room temperature. The observation of enhanced oxygen adsorption kinetics for Au[1] and Ag[3] overlayers on silicon has been related, respectively, to the "metallic" state of Si atoms in amorphous Au-Si alloys, and to the disruption of the ordered Si(111) surface upon Ag deposition.[2] The effect of "amorphization" of the Si(111) surface on the Si oxidation rate can be estimated by the results of Riedel et al.[16] who recently studied amorphous Si and Ge layers upon exposure to activated oxygen. The vertical bar in the bottom section of Fig. 3 indicates the position of the dominant Si 2p oxide feature observed by Riedel et al.[16] The similarity with our results suggests that in both cases silicon atoms locally coordinated to 3 and 4 oxygen atoms are likely to coexist and give rise to the broad Si 2p oxide band observed experimentally. Comparison of the intensity of the Si 2p oxide feature relative to the main line, however, indicates that the oxygen adsorption rate on the Cr-activated Si surface is 20 to 30 times higher than on amorphous silicon layers.[4,16] We conclude that while the nature of the surface reaction products appears the same in both cases, the origin of enhanced oxidation has to be found in the Cr-induced modification of surface chemistry rather than in the amorphization of the semiconductor surface layer. In earlier studies of the Si(111)-Cr interface[7] we proposed that for Cr coverages below 1.5 monolayer the Cr atoms participate in weak chemisorption bonds that affect only slightly the stability of the Si-Si bonds in the surface and near surface region. At coverage above 1.5 monolayers a reactive interdiffusion interface formation stage is established, with formation of disordered Si-Cr intermixed species. The onset of reactive interdiffusion coincides within experimental uncertainty with the critical threshold coverage for oxidation determined in this work. The correspondence suggests that the silicide-like surface species formed above threshold act as catalyst for the oxidation of the semiconductor atoms in the surface and near-surface region. Since no saturation of oxygen adsorption is observed, it is conceivable that while Si atoms oxidize and segregate at the surface-vacuum interface, other Si atoms in the near surface region combine with Cr to maintain a steady state silicide-like bonding configuration. We have therefore started systematic studies of the Cr 3p core emission to ascertain if the silicide species remain pristine upon oxidation, or if mixed oxide species are formed.

The morphology and room-temperature revolution of the Si(111)-Cr and GaAs(110)-Cr interfaces present many similarities. Both interfaces react at room temperature only for coverages above 1.5 Å(Si[7]) and 2 Å (GaAs[8]). Reactive interdiffusion occurs in a limited coverage range of $1.5<\Theta<9$ (Si[7]) and $2<\Theta<20$ (GaAs[8]) and yields silicide-like species on Si and arsenide-like phases on GaAs. Further Cr deposition gives rise to an unreacted metal film on top of the reacted interface.[7,8] The chemical bonding for the main interface reaction products (silicide or arsenide-like) involves in both cases dominant coupling of the metal-d states with anion-p states, with similar modifications of the electronic density of states. These similarities between the two interfaces suggest that an oxidation promotion effect may be found also for GaAs-Cr above a critical threshold coverage value of $\Theta \approx 2$Å. In Fig. 4 and 5 we summarize, respectively, the effect of exposure to activated oxygen on the As 3d and Ga 3d core level emission. In the top-section of Fig. 4 and 5 we show the clean surface core emission before (dashed line) and after (solid line) oxygen exposure. In the mid-section we show the corresponding results for a Cr overlayer with $\Theta=1$, i.e. below the critical threshold coverage. In the bottom-most sections of Figs. 4 and 5 we present results for Θ=10. Again, the dashed line and solid line indicate, respectively, results before and after oxygen exposure. The zero of the binding energy scale corresponds to the flat-band initial core binding energy, and the spectra have been arbitrarily normalized to emphasize lineshape changes. As indicated in Fig. 1 and 2, and in the topmost section of Fig. 4 and 5, the clean GaAs surface is relatively inert and only low oxygen coverage can be obtained at room temperature. For Θ=1 the Cr overlayer yields only relatively small modifications in the oxygen uptake rate. The spectra at Θ=10, instead, show dramatic modification of the As 3d and Ga 3d lineshape upon oxidation. At Θ=10 the As 3d and Ga 3d lines before oxidation (dashed line) both include two distinct components.[8] For As (tic marks) a low binding energy As-Cr reacted 3d line appears above the initial clean surface emission, and a second line shifted to higher binding energy represents segregated arsenic and residual substrate emission.[8] For Ga (dashed line, bottom-most section) the main contribution corresponds to free Gallium atoms that are a byproduct of the As-Cr interface reaction and/or Gallium atoms dispersed in a Cr matrix.[8] Upon exposure to 100L of activated oxygen most of the As and Ga atoms within the experimental sampling depth are oxidized. We find an increase of several orders of magnitude in the overall surface oxygen uptake, and no evidence of saturation in the exposure range explored ($10-10^4$ L). The character of the oxidation reaction products can be examined by comparing the observed oxide-induced As 3d features with those reported by Landgren et al[12] 0.8, 2.3, 3.2 and 4.2 eV below the initial substrate As 3d line (vertical bars 1-4 in Fig. 4) and related to the presence of As coordinated, respectively, with one, two, three and four oxygen atoms, and with a 3.4-3.5 eV feature reported in Refs. 17 and 18 for As in As_2O_3 (vertical bar 5 in Fig. 4). The broad experimental oxide band suggests that several arsenic-oxygen bonding configurations must coexist in the surface and near surface region, with a dominant contribution coming from high oxidation states that are barely detectable on the oxidized GaAs surface[12] at coverages of 10^{14} L.

For the Ga 3d lines during oxidation of GaAs Landgren et al.[12] observed chemically shifted components 0.45 and 1.0 eV below the main line at low exposure (10^6L molecular oxygen), components at 0.8 and 1.4 eV at high exposure (10^{14}L). These are indicated by vertical bars 1-4 in Fig. 5. While Landgren et al.[12] suggest that the 1.4 eV component may correspond to Ga in Ga_2O_3, Su et al. report[19] a Ga 3d broad oxide feature centered some 2.2 eV below the main Ga $3d_{5/2}$ line for Ga_2O_3. This is marked by vertical bar 5 in Fig. 5. The results in the bottom-most section of Fig. 5 indicate that several non-equivalent oxidation states for Ga coexist within the sampling depth. Furthermore we note that a major spectral contribution derives from Ga 3d oxide features shifted 3 eV below the pinned Ga 3d position (mid-section of Figs. 2 and 5), i.e. from higher oxidation states than previously observed for Ga in Ga_2O_3. The nature of these new oxide species is not clear at present. It may involve mixed Cr-Ga oxide phases, but valence band results are consistent[20] with a main Cr_2O_3 oxidation state for Cr, with no evidence of mixed oxides. In analogy with the present case, we mention that Al overlayers on Ge[2] appear to stabilize a higher oxidation state for Ga atoms upon oxygen exposure at room temperature. Also in this case the morphology of the new potentially stable oxide phase remains unclear.

In summary, we have shown that thin Cr overlayers on Si and GaAs surfaces can dramatically enhance the semiconductor oxidation rate if Cr coverages above a critical threshold coverage are employed. This critical coverage corresponds to the onset of reactive interdiffusion of Cr and semiconductor atoms at the interface. The resulting enhancement of several orders of magnitude in the oxygen adsorption kinetics is presumably related to the catalytic activity of ultrathin silicide and arsenide-like overlayers formed for Cr coverages above threshold.[1,4] The end products of oxidation involve semiconductor atoms in several different coexisting oxidation states, with high oxidation states largely dominant. For silicon, for example, dominant 3 and 4-fold oxygen coordination was observed, with compelling analogies to α-Si oxidation processes.

Acknowledgements

This work was supported by the Office of Naval Research (G. Wright) under contract N00014-84-K-0545 and by the Center for Microelectronic and Information Science of the University of Minnesota. We thank G. Margaritondo and J.H. Weaver for communicating their results to us prior to publication and for generously providing equipment. The Synchrotron Radiation Center of the University of Wisconsin-Madison is supported by the National Science Foundation under Grant No. DMR-8020164 and we gratefully acknowledge the cheerful support of its staff.

References

1. A. Cros, J. Derrien, and F. Salvan, Surf. Sci. 110, 471 (1981); J. Derrien and F. Ringeisen, Surf. Sci. 124, L35 (1983).
2. A.D. Katnani, P. Perfetti, Te-Xiu-Zhao, and G. Margaritondo, Appl. Phys. Lett. 40, 619 (1982).
3. G. Rossi, L. Caliari, I. Abbati, L. Braicovich, I. Lindau, and W.E. Spicer, Surf. Sci. Lett. 116, L202 (1982).
4. A more detailed analysis of the results for silicon will be presented in a forthcoming paper: A. Franciosi, P. Philip and C. Caprile, to be published.
5. L.J. Brillson, G. Margaritondo and N.G. Stoffel, Phys. Rev. Lett. 44, 667 (1980).
6. A. Franciosi, D.G. O´Neill and J.H. Weaver, J. Vac. Sci. Technol. B1, 524 (1983); A. Franciosi, J.H. Weaver and D.G. O´Neill, Phys. Rev. B28, 4889 (1983).
7. Extensive results for Si(111)-Cr at room temperature have been presented in A. Franciosi, D.J. Peterman, J.H. Weaver and V.L. Moruzzi, Phys. Rev. B25, 4981 (1982)
8. Results for the GaAs(110)-Cr interface at room temperature were available to us in the form of a preprint: J.H. Weaver, M. Grioni, and J. Joyce to be published.
9. See, for example, C.N. Satterfield, "Heterogeneous Catalysis in Practice," McGraw Hill Book Co., New York (1980).
10. We will not try to summarize the extensive literature available on this subject. See for example, G. Hollinger and F.J. Himpsel, Phys. Rev. B28, 3651 (1983) and J. Vac. Sci. Technol. A1, 640 (1982) and references therein.
11. C.M. Garner, I. Lindau, C.Y. Su, P. Pianetta, and W.E. Spicer, Phys. Rev. B19, 3944 (1979); C.Y. Su, P.R. Skeath, I. Lindau, and W.E. Spicer, J. Vac. Sci. Technol. 19, 481 (1981) and references therein.
12. See for example G. Landgren, R. Ludeke, Y. Jugnet, J.F. Morar, and F.J. Himpsel, J. Vac. Sci. Technol. B2, 351 (1984); T. Miller and T.-C. Chiang, Phys. Rev. B29, 7034 (1984) and references therein.

13. C.Y. Su, I. Lindau, P.W. Chye, P.R. Skeath, and W.E. Spicer, Phys. Rev. B24, 4045 (1982) and references therein.

14. In these conditions the "activated" species consist mostly of excited molecular oxygen and atomic oxygen. See J.A. Silberman, D. Laser, I. Lindau, W.E. Spicer and A. Wilson, J. Vac. Sci. Technol. A1, 1706 (1983).

15. P. Pianetta, I. Lindau, C.M. Garner, and W.E. Spicer, Phys. Rev. B18, 2792 (1978).

16. R.A. Riedel, M. Turowski, G. Margaritondo, P. Perfetti, and C. Quaresima, unpublished.

17. C.Y. Su, I. Lindau, P.R. Skeath, I. Hino and W.E. Spicer, Surf. Sci. 118, 257 (1982).

18. R. Holm and S. Storp, Appl. Phys. 9, 217 (1976).

19. C.Y. Su, P.R. Skeath, I. Lindau, and W.E. Spicer, Surf. Sci. 118, 248 (1982).

20. S. Chang and A. Franciosi, unpublished.

Figure Captions

Fig. 1 Photoelectron Energy Distribution Curves (EDC´s) for the As 3d core emission from cleaved GaAs(110). Spectra displaced downward show the effect of exposure to increasing amount (100, 10^3, 10^4 L) of activated oxygen. The rigid shift of the core line reflect the variation in band bending from the initial flat-band situation.

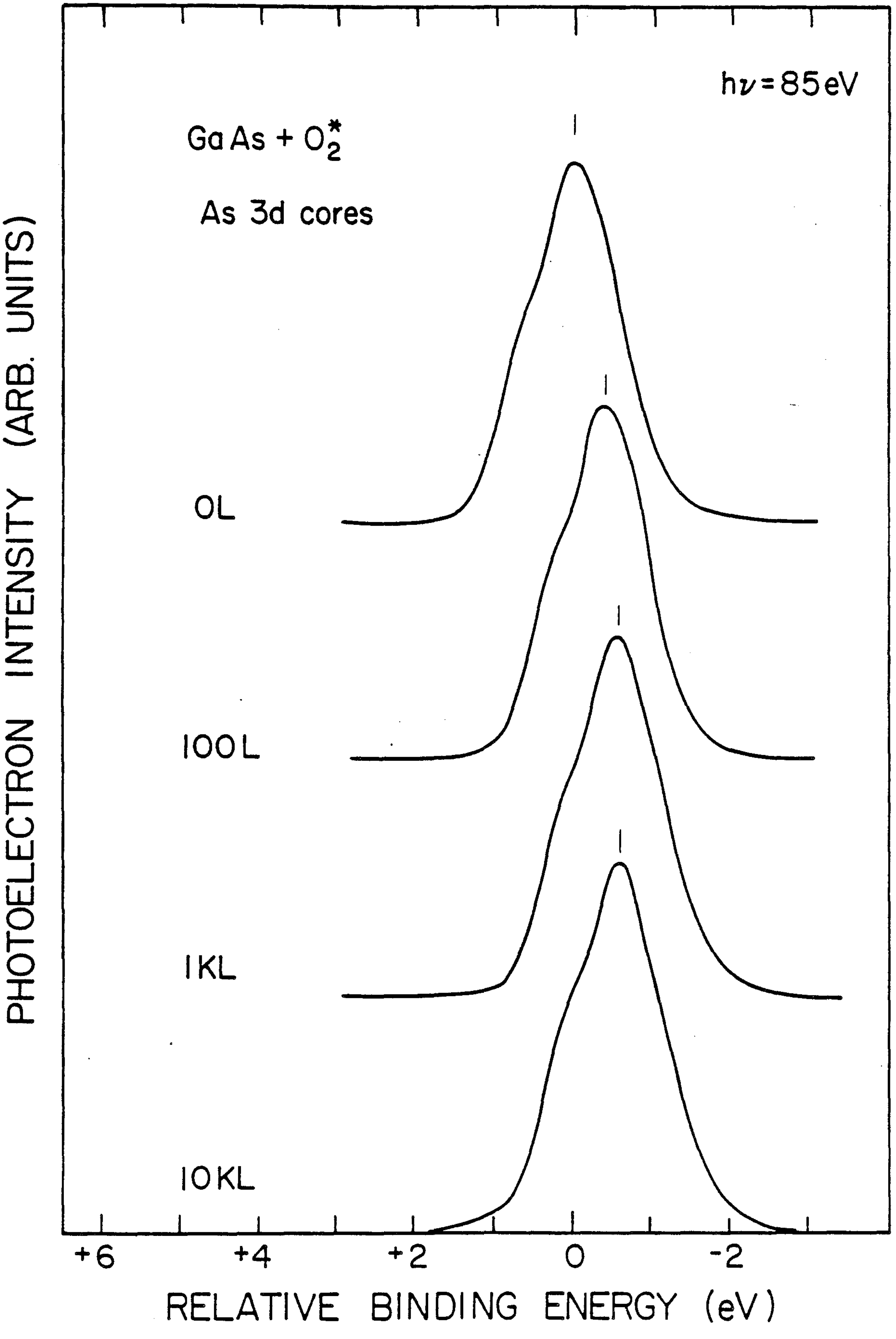
Fig. 2 EDC´s for the Ga 3d core emission from cleaved GaAs(110). Spectra displaced downward show the effect of exposure to increasing amounts of activated oxygen. The rigid shift of the core line reflect the variation in band bending.

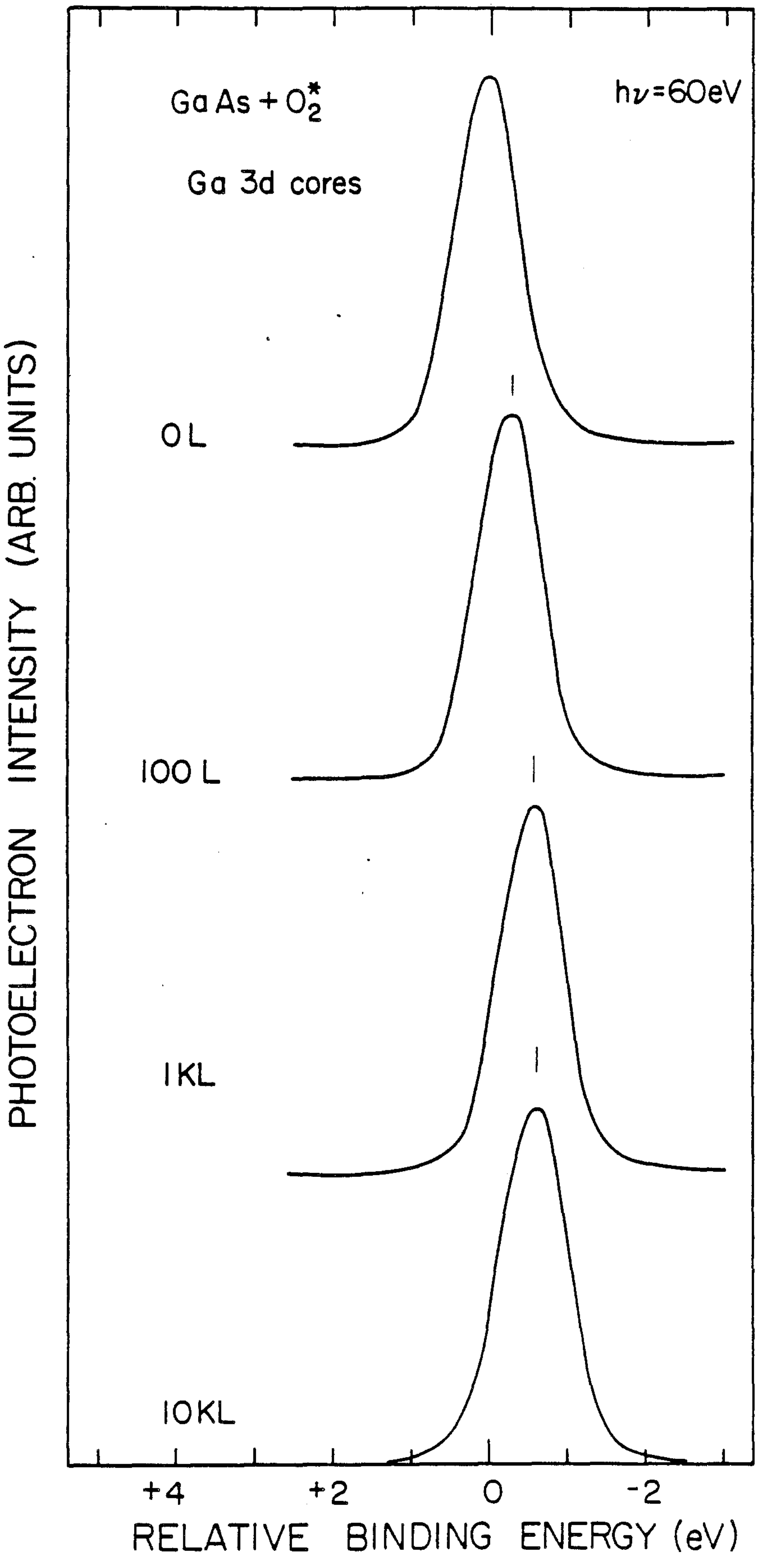
Fig. 3 Si 2p core emission from cleaved Si (111)2x1 surfaces. Top: clean surface emission before (dashed line) and after exposure (solid line) to 100L of activated oxygen. The vertical bars mark the position of Si 2p oxide features associated by Hollinger and Himpsel[10] with silicon atoms coordinated with one, two, three and four oxygen atoms. Mid-secion: A 0.6 Å Cr overlayer was deposited on a freshly cleaved Si(111) surface. The resulting Si 2p core emission is shown before (dashed line) and after (solid line) oxygen exposure. The vertical bars mark Hollinger and Himpsel´s Si 2p oxide features. Bottom: Effect of a 2Å Cr overlayer on the Si (111) surface oxidation. The Si 2p core lineshape before oxidation (dashed line) is similar to the initial Si 2p line. After exposure to 100L of activated oxygen (solid line) a major oxide band emerges. The vertical bar marks the position of a major Si 2p oxide features identified by Riedel et al.[16] during oxidation of amorphous silicon.

Fig. 4 Effect of thin Cr overlayers on the oxidation of As at the GaAs(110) surface. Top: Clean surface As 3d emission before (dashed line) and after (solid line) oxygen exposure. Mid-section: A 1Å Cr overlayer was deposited on a freshly

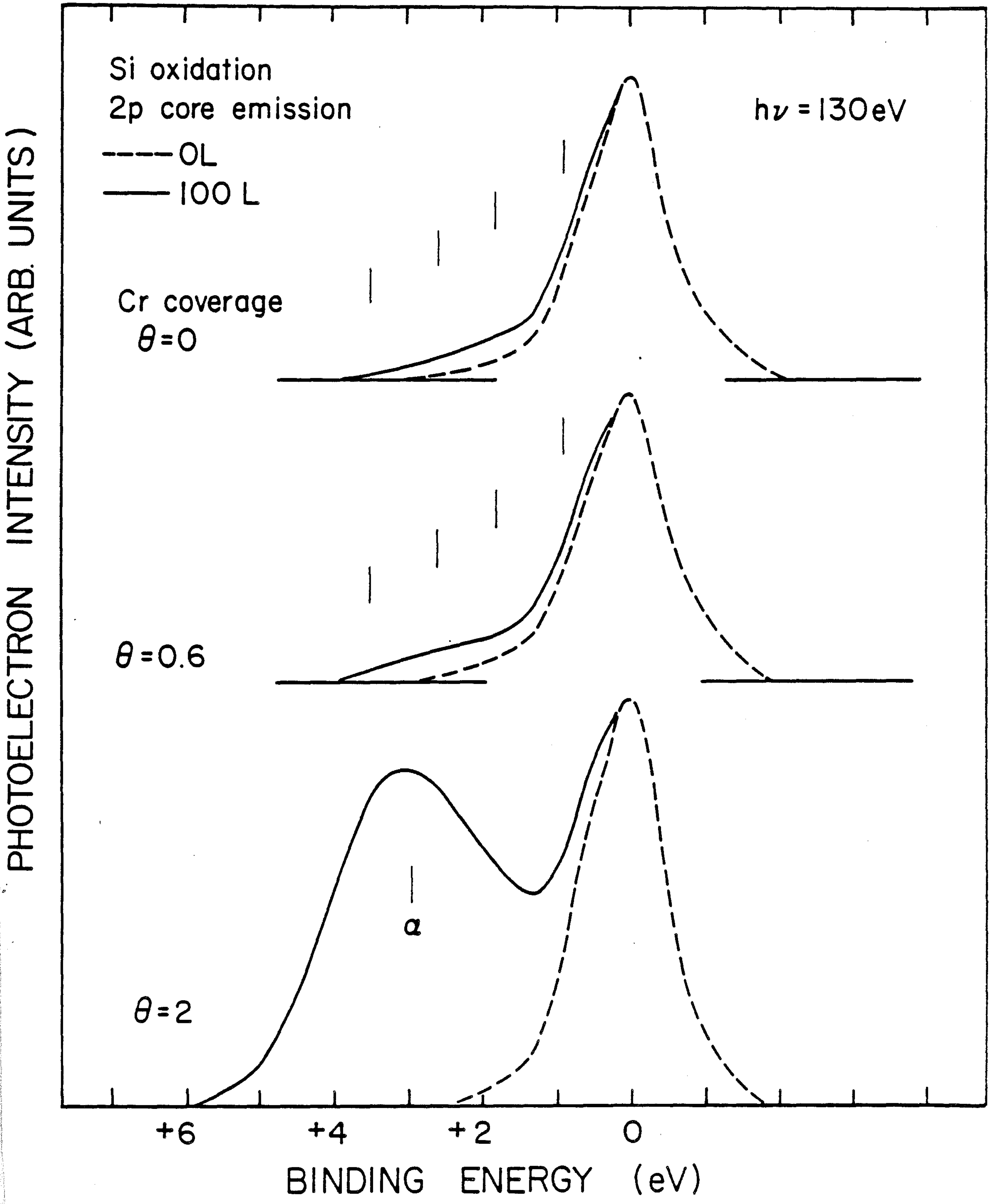
cleaved GaAs(110) surface. The resulting As 3d emission is shown before (dashed line) and after (solid line) oxygen exposure. Botton: Effect of a 10Å Cr overlayer on the GaAs(110) surface oxidation. The As 3d core lineshape before oxidation (dashed line) is composed of a low binding energy reacted As 3d feature from Cr-As interface species, and of a high binding energy segregated As/substrate contribution.[8] Upon oxidation (solid line) most of the As atoms appear oxidized. The vertical bars 1-4 mark the position of the oxidized As 3d features observed by Landgren et al.[12] for As coordinated with one to four oxygen atoms. The vertical bar 5 marks the position of the As 3d core level in As_2 O_3, from Su et al.[17]

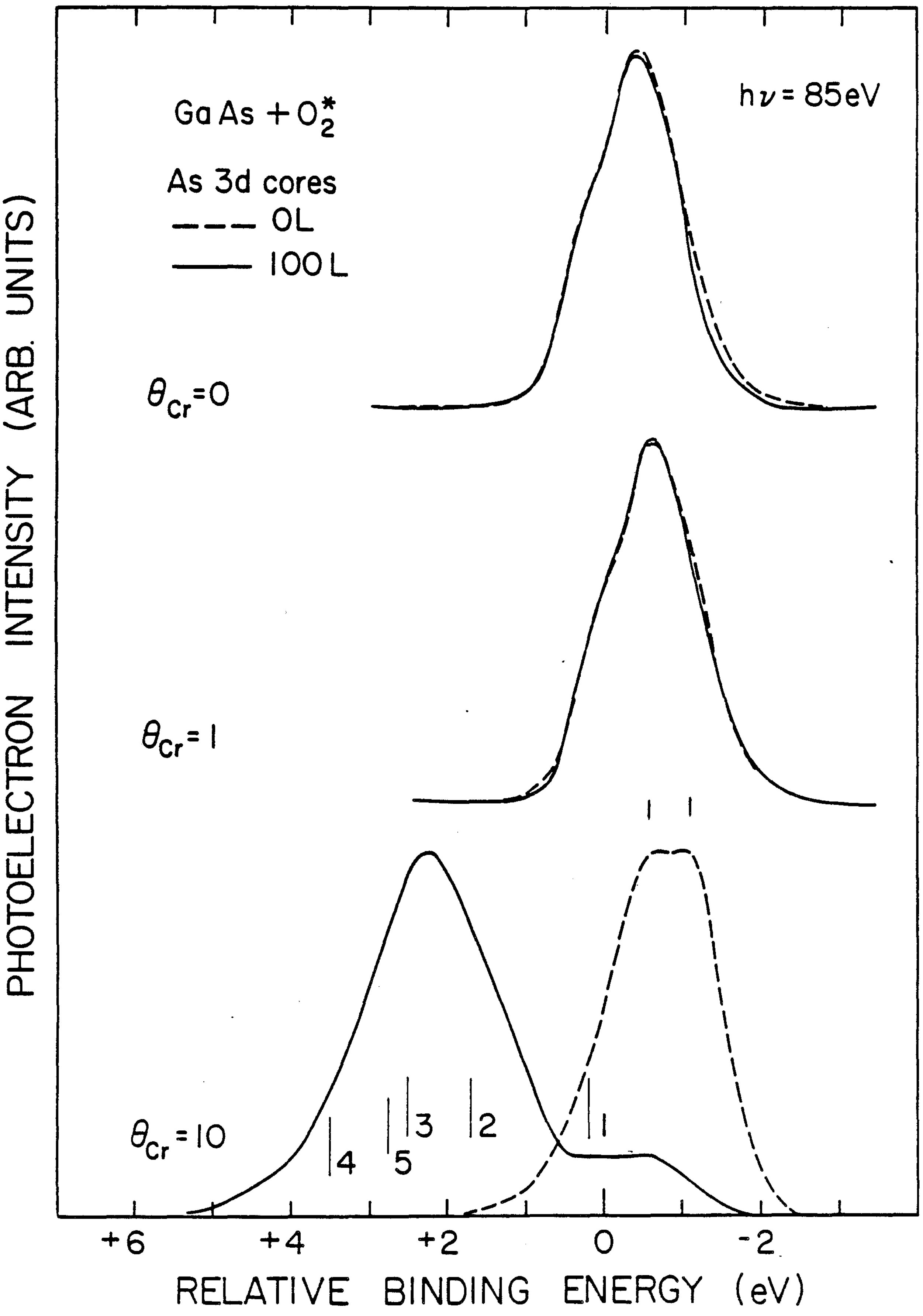
Fig. 5 Effect of thin Cr overlayers on the oxidation of Ga at the GaAs(110) surface. Top: clean surface Ga 3d emission before (dashed line) and after (solid line) oxygen exposure. Mid-section: A 1Å Cr overlayer was deposited on a freshly cleaved GaAs(110) surface. The resulting Ga 3d exposure is shown before (dashed line) and after (solid line) oxygen exposure Bottom: Effect of a 10 Å Cr overlayer.· The Ga 3d core line before oxidation (dashed line) includes a main contribution from dissociated Ga atoms or from Ga atoms in a Cr matrix.[8] Upon oxidation (solid line) most of the Ga atoms appear oxidized. The vertical bars 1-4 mark the position of oxidized Ga 3d features observed by Landgren et al.[12] The vertical bar 5 marks the position of the Ga 3d core level in Ga_2O_3, from Su et al.[19]

hν=85eV
GaAs + O_2^*
As 3d cores
0L
100L
1KL
10KL
PHOTOELECTRON INTENSITY (ARB. UNITS)
+6
+4
+2
0
-2
RELATIVE BINDING ENERGY (eV)

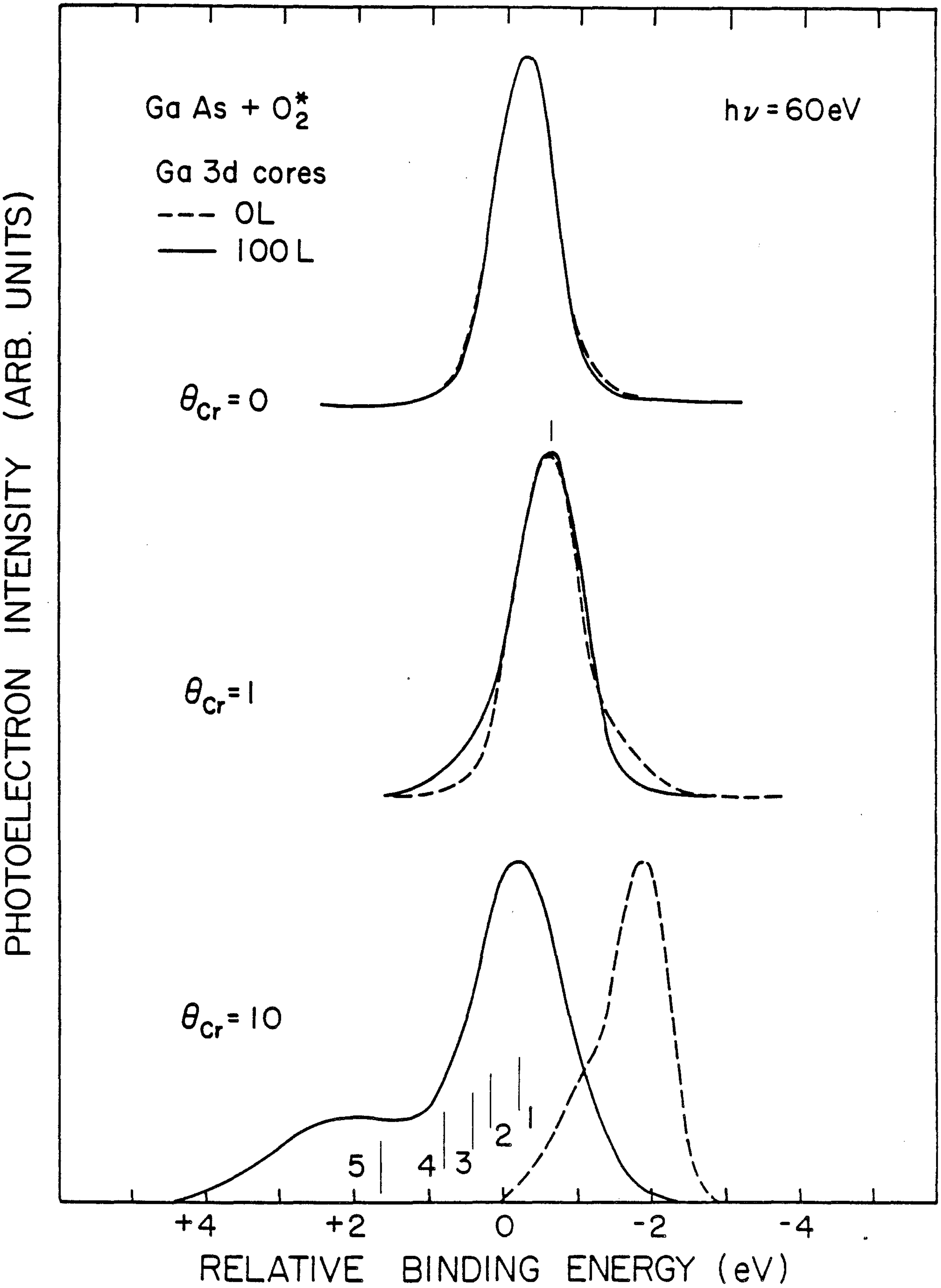


GaAs + O_2^*
Ga 3d cores
hν=60eV
0L
100L
1KL
10KL
PHOTOELECTRON INTENSITY (ARB. UNITS)
+4
+2
0
-2
RELATIVE BINDING ENERGY (eV)


Si oxidation
2p core emission
----0L
——100 L
hν = 130eV
Cr coverage
θ=0
θ=0.6
θ=2
a
+6
+4
+2
0
BINDING ENERGY (eV)
PHOTOELECTRON INTENSITY (ARB. UNITS)



GaAs + O_2^*
As 3d cores
--- 0L
— 100L
hν = 85eV
θ_{cr}=0
θ_{cr}=1
θ_{cr}=10
4
5
3
2
1
+6
+4
+2
0
-2
RELATIVE BINDING ENERGY (eV)
PHOTOELECTRON INTENSITY (ARB. UNITS)


Ga As + O_2^*
hν = 60eV
Ga 3d cores
--- 0L
— 100L
θ_{Cr} = 0
θ_{Cr} = 1
θ_{Cr} = 10
1
5
4
3
2
1
PHOTOELECTRON INTENSITY (ARB. UNITS)
+4
+2
0
-2
-4
RELATIVE BINDING ENERGY (eV)

MEIS

MICROELECTRONIC & INFORMATION SCIENCES CENTER

INSTITUTE OF TECHNOLOGY
UNIVERSITY OF MINNESOTA

227 Lind Hall / 207 Church Street S.E.
Minneapolis, Minnesota 55455
612/376-9122

EFFICIENT ALGORITHMS FOR LAYER ASSIGNMENT PROBLEM

Microelectronic and Information Sciences Center

Technical Report #30

K. C. Chang
Department of Computer Science
University of Minnesota

H. C. Du
Department of Computer Science
University of Minnesota

MEIS technical reports are published in order to provide early access to research results intended for publication. Manuscripts are reviewed and selected for MEIS technical reports by the MEIS Executive Committee.

EFFICIENT ALGORITHMS FOR LAYER
ASSIGNEMENT PROBLEM

MEIS Technical Report #30

K. C. Chang
H.C. Du

This work was supported in part by NSF Grant DCR-8405498, and by grants from CDC/Honeywell/Sperry, and the Microelectronic and Information Sciences Center.

Abstract

The layer assignment problem for interconnect is the problem of determining which layers should be used for wiring the signal nets. The objective of the layer assignment problem in general is to minimize the number of vias required. Thus, it is often also referred to as the via minimization problem. In a via minimization problem if the topology of the given layout is fixed, the problem is referred to as a Constrained Via Minimization (CVM) problem. On the other hand, if both the topology of the layout and the layer assignment are to be decided, it is referred to as an Unconstrained Via Minimization (UVM) problem. In this paper, both the CVM and UVM problems are studied. Efficient algorithms, which can be easily modified to take extra constraints into consideration and are intended for various objectives, for both problems are proposed. Experimental results show that the proposed algorithms for the CVM problem are time efficient compared with existing algorithms and generate better (near-optimal) results and the proposed algorithms for the UVM problem generate better results but may take more computational time. In the CVM problem, some vias are "essential" to the given layout. That is, they have to be selected and cannot be replaced by other possible vias. Efficient algorithms for identifying essential vias are also presented and discussed in this paper.

Indexed Terms : Layer Assignment, Via Minimization, Layout, Routing.

1. Introduction

The routing problem in VLSI/LSI layout is to realize a set of specified interconnections among modules in as small an area as possible. Most existing routing algorithms for two layers are based on the assumption that all the vertical wire segments are assigned on one layer while all the horizontal wire segments are assigned on the other. Therefore, a large number of vias are introduced to connect the wire segments on different layers.

Vias not only reduce the reliability and performance of the circuit, but also increase the manufacturing cost. Thus, it is desirable to reduce the number of vias. The objective of the layer assignment problem is to assign wire segments to the layers so that the number of vias is minimized.

To facilitate the understanding of this problem, the following definitions are introduced :

Net : A net is a collection of wire segments that electrically connect a set of terminals (pins).

Via : A via is a feed-through hole or a contact where wire segments on different layers are connected.

Split point :

A split point is a point in a net that connects two or more wire segments. Note that a via location is always considered as a split point.

Split number :

The split number of a split point is the number of different wire segments connecting to this split point. We shall say "a point is a k-way split point", if its split number is k.

Crossing :

A crossing is a point where two wire segments in different nets intersect. For simplicity, in this paper we shall say "two wire segments (or nets) cross each other", if they intersect or portions of the two wire segments (or nets) are overlapped. Note that no vias can be located at crossing points.

For example, there are three nets A, B, and C connecting terminals (a,a'), (b,b'), and (c,c', c") respectively in Figure 1. Points 1, 3, 4, 7, 9, 10 are vias and points 2, 5, 6, 8 are crossing points. The split number of all split points is 2, except point 9 which has a 3-way split. If we do not have the restriction that all the horizontal wire segments must be on one layer and all the vertical wire segments must be on the other, the number of vias can be reduced. As shown in Figure 2, only one via is required for the instance in Figure 1.

Since the objective of the layer assignment problem is to minimize the number of vias required, the problem is often also referred to as the via minimization problem. In this paper, we will use both terms interchangeably. In a via minimization problem if an interconnect layout including all the possible wire segments and all the possible vias is given as an initial condition, it is referred to as a Constrained Via Minimization (CVM) problem. That is, the topology of the layout is fixed except that the layers assigned to the wire segments are to be decided. On the other hand, the problem in which both topology of the layout and the layer assignment are to be decided is referred to as an Unconstrained Via Minimization (UVM) problem.

Hashimoto and Stevens first formulated the CVM problem [1]. However, only two-way splits were allowed in their model. Servit tried to reduce the number of vias by duplicating wire segments on the opposite layer [2]. By using a graphical model, Stevens and VanCleemput proposed an approximate method [3]. Kajitani presented a way in [4] to transform Hashimoto-

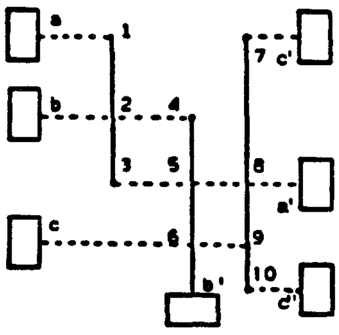


Figure 1. A 2-layer interconnect layout

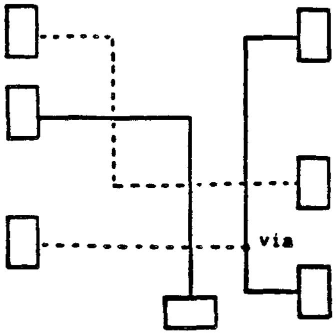


Figure 2. Only one via is required

Stevens type of via minimization problem into the problem of finding a maximum cut in a planar graph which subsequently can be solved optimally in polynomial time [5]. Ciesislski and Kinnen proposed an integer programming method to solve the general problem [6], but the time complexity of their algorithm is exponential. The CVM problem for two layers has been believed to be NP-Complete for quite some time until Chen et al [7-9] and Pinter [10] recently proposed optimal polynomial time algorithms based on the maximum cut algorithm for planar graphs. However, their algorithms can not take care of the cases where there exist vias with more than 3-way splits. Therefore, whether the general 2-layer CVM problem with possibly more than 3-way split points is NP-Complete is still an open question.

All the above algorithms are intended for minimizing the total number of vias. Due to the inherent properties of those algorithms, it is extremely hard, if is not impossible, to consider extra constraints. Pinter has pointed out the need for minimizing the largest number of vias in a net to avoid burdening one net with an excessive number of vias [10]. Also due to the fabrication technology, or performance and density considerations, more constraints should be associated with the CVM problem. For instance, power lines should always be assigned to the metal layer or large portions of several nets should be assigned to a particular layer. Lee, Hong and Wang discussed the problem of arranging vias which must be confined with certain neighborhood constraints [11]. Basically the problem becomes harder, if more constraints are considered. Recently, Du and Chang [14] proposed a heuristic algorithm based on the theory of bipartite graphs to cope with different practical constraints.

Hsu [12] first formalized a two dimensional routing problem and tried to find a topological routing solution and minimize the number of vias at the same time. He suspected this problem is NP-complete and proposed a heuristic algorithm for it. Later, Marek-Sadowska[13] extended the two dimensional routing problem to the Unconstrained Via Minimization (UVM) problem and proved that both of them are NP-complete. Since the topology of the layout is allowed to be changed in the UVM problem, the number of vias required can be smaller than the corresponding instance of the CVM problem. However, this improvement is often at the expense of a longer total wire length and irregular layout wiring.

All the existing optimal algorithms for the CVM problem with no more than 3-way split points [7-10] are based on the maximum cut algorithms for planar graphs and have time complexities of O($m^{2.5}$), where m is the number of clustered wire segments. An instance was given in [14] in which the number of clustered wire segments is in the order of n^2, where n is the number of nets. The heuristic algorithm proposed in [14] has time complexity of O($p.n^2$), where p is the number of vias selected and n is the number of wire segments formed in the final layout. However, in the worst case, p can be at least O(n^2). Therefore, it may take a long time to generate results for an instance involving a large number of nets. The objective of our study is to design and develop efficient and practical heuristic algorithms which hopefully run faster and generate better results than the existing algorithms for both the CVM and UVM problems.

In the next section, we will first discuss the foundations of the via minimization problem. The proposed approach and the basic algorithm for solving the CVM problem will be presented in Section 3.

In the CVM problem, some vias are essential to the given layout. That is, in order to achieve a feasible solution, they have to be selected and cannot be replaced by other possible vias. We shall call them "essential vias". Since essential vias have to be selected, if they can be identified by a fast pre-processing procedure the whole process of solving the CVM problem can be speeded up. In Section 4, we present fast algorithms to identify essential vias. In Section 5, we will show that a similar approach can be applied to solve the UVM problem and present some comparisons between the results generated by our approach and those generated by the method proposed in [12]. Some possible extensions of the proposed algorithms and conclusions will be discussed in Section 6.

2. Foundations of the Via Minimization Problem

In the following we discuss the foundations of both the CVM and UVM problems. Some terminologies will be introduced first.

A finite graph G=(V,E) consists of a finite set of vertices V=$\{v_1,v_2,\ldots,v_n\}$ and a finite set of edges E=$\{e_1,e_2,\ldots,e_m\}$. To each edge there corresponds a pair of vertices : if (v,w) corresponds to edge e, then e is said to be incident on vertices v and w, v and w are adjacent to each other with respect to e. The degree of a vertex v is the number of edges that are incident on v. A graph is undirected if the vertex pair (v,w) associated with each edge e is an unordered pair. A path from vertex v_p to vertex v_q in graph G is a sequence of vertices $v_p,v_{i_1},\ldots,v_{i_d},v_q$, such that (v_p,v_{i_1}), (v_{i_1},v_{i_2}), ..., (v_{i_d},v_q) are edges in E of G. The length of a path is the total number of edges on it. A simple path is a path in which all vertices except possibly the first and the last vertices are distinct.

A cycle is a simple path in which the first and the last vertices are the same. An undirected graph G is said to be connected if for every pair of distinct vertices v_i and v_j in V there is a path from v_i to v_j in G. An acyclic graph is a graph which does not have a cycle. A subgraph of a graph G=(V,E) is a graph whose vertices and edges are in G. The subgraph induced by S $\subseteq$ V is the subgraph of G that results when the vertices in V-S with all edges incident on them are removed from G. A connected, undirected acyclic graph is called a tree.

A spanning tree of a connected graph G is a tree that is a subgraph of the graph G and contains every vertex of G. A bipartite graph G=(V,E) is an undirected graph whose vertices can be partitioned into two disjoint sets V_1 and $V_2=V-V_1$ with the property that no two vertices in V_i are adjacent in G for i=1, 2. Consider a spanning tree (V,T) of a connected undirected graph G=(V,E). Any edge not in T, that is, any edge in E-T, will create exactly one cycle when added to T. Such a cycle is a member of the fundamental set of cycles of G with respect to T. In this paper, without possibly causing confusion, we will simply call cycles in the fundamental set of cycles with respect to a spanning tree "fundamental cycles". Since every spanning tree of a graph contains | V | -1 edges, there are | E | - | V | +1 cycles in the fundamental set of cycles with respect to any spanning tree of G, where | V | and | E | represent the number of vertices in V and the number of edges in E respectively.

Now we consider the CVM problem first. Since a layout including a set of possible vias is given in the CVM problem, a crossing graph G(V,E) for the layout can be constructed as follows : Each vertex $v_i \in V$ represents a wire segment n_i in the layout and an edge $(v_i, v_j) \in$ E if wire segments n_i and n_j cross each other. Initially, each net is considered as a single wire segment. A wire segment can be broken into two or more wire segments (depending on the split number of the via location) by choosing one via along that wire segment. In order to realize the layout in two layers, the corresponding crossing graph should be 2-colorable. If the original crossing graph is 2-colorable, no vias are required. The following two theorems related to 2-colorability of a graph are well known.

Theorem 1 [16] : A graph is 2-colorable if and only if the graph is bipartite.

Theorem 2 [16] : A graph with at least one edge is bipartite if and only if it has no cycle of odd length.

Therefore, the 2-colorability of a crossing graph can be easily verified by a depth first search algorithm in polynomial time. If a crossing graph is not 2-colorable, then a set of vias is required to allow the layout to be realized in two layers. Once a via is chosen along a wire segment, that segment is broken into two and the corresponding vertex in the crossing graph is split into two vertices (assuming that the via is a 2-way spilt point). This process is continued until the final crossing graph is bipartite so that the corresponding layout can be implemented in two layers. The objective of the CVM problem is to choose a set of vias as small as possible to achieve the 2-colorable crossing graph. For example, Figures 3(a) and (b) show a layout example and its corresponding crossing graph. From Theorem 2, the graph in Figure 3(b) is not a bipartite graph because it has a cycle of length 3 formed by nets 1, 4, and 5. By selecting a via in net 1, that net is broken into two subnets 1a and 1b. The corresponding crossing graph and the final layout are shown in Figures 4(a) and 4(b) respectively. Note that the crossing graph in Figure 4(a) is 2-colorable and that the layout can be implemented in two layers as shown in Figure 4(b), where the wire segments in dotted lines are on one layer while the wire segments in solid lines are on the other.

One big advantage of interpreting the CVM problem in this way is that many practical constraints can be easily incorporated into the proposed algorithm. For simplicity throughout the paper we concentrate on the algorithms intended for minimizing the total number vias. However, the via selection criterion can be easily modified to minimize the maximum number of vias in a net. Similar approaches as described in [14] can be taken to cope with other objectives and constraints.

By Theorem 2, all the cycles of odd length in the crossing graph should be broken before it can become a bipartite graph. Even though there are existing optimal algorithms for

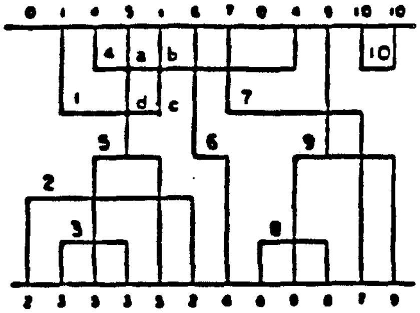


Figure 3(a): A layout example

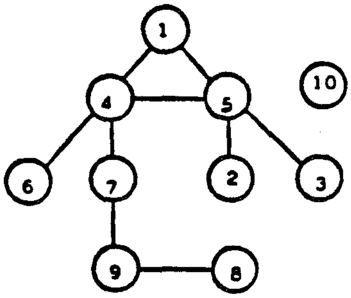


Figure 3(b): Crossing graph

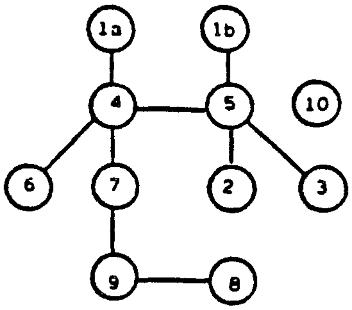


Figure 4(a): Final crossing graph

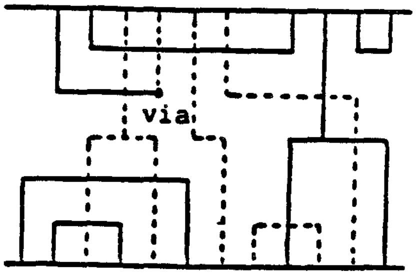


Figure 4(b): Final layout

minimizing the total number of vias when there are no more than 3-way split points, our objective is to develop efficient and practical algorithms for the general via minimization problem. We will therefore take a heuristic approach. In order to minimize the number of vias, intuitively we would like to introduce a via that will break as many odd cycles in the crossing graph as possible in each iteration. Therefore, a possible "greedy" approach is to select a via in a net whose corresponding vertex in the crossing graph involved in the maximum number of odd cycles. This implies that we might have to generate all the cycles of odd length in the crossing graph. Unfortunately, finding all the cycles of odd length in a graph is not practical because the number of cycles may be in the order of $2^{|V|}$. Due to the following theorem and the fact that a fundamental cycle set can be obtained in polynomial time [17], Du and Chang proposed an approximation method by generating a fundamental cycle set instead of generating all cycles of odd length.

Theorem 3 [14] : A graph is 2-colorable if and only if each fundamental cycle is of even length.

However, a fundamental cycle set corresponds to a particular spanning tree in the graph. It may not represent the global property of the graph. That is, a vertex involved in the largest number of fundamental cycles of odd length may not be the one involved in the largest number of cycles of odd length in the graph. Therefore, the quality of the results obtained may be affected. A new approach will be proposed in the next section.

In the UVM problem, only all terminal positions of each net are known and the topology of the layout is not fixed. From the terminal positions of each net, pairs of nets which must be crossed can be decided. Therefore, an initial crossing graph as described before can be constructed. Marek-Sadowska has shown that in an optimal solution for a UVM problem if a net is routed with vias, then it is routed with one via only [13]. A way of inserting a new net into a solution for a UVM problem such that it requires only one more via is given. Therefore, finding a solution for the UVM problem is equivalent to minimizing the number of vertices (nets) that have to be deleted from the initial crossing graph such that the reduced graph becomes bipartite. Let us call the above problem "Minimum Node Deletion Bipartite Subgraph Problem". For arbitrary graphs (not just for crossing graphs), this problem is NP-Complete [23]. Since the UVM has been shown to be NP-Complete [13] and the number of vias required is equivalent to the number of nodes deleted, the "Minimum Node Deletion Bipartite Subgraph Problem" has to be NP-Complete too. Therefore, we have to take a heuristic approach to solve the UVM problem. In order to minimize the number of vertices needed to be deleted, in a straightforward greedy approach we also need to identify and then delete the vertices which are involved in the maximum number of cycles of odd length. Therefore, a similar approach to the one for solving the CVM problem can be applied to the UVM problem too.

3. The CVM Problem

Since we decide to take a heuristic approach to solve the CVM problem and it takes exponential time to generate all the cycles of odd length, an efficient heuristic approach for selecting vias is required. For a given crossing graph, the number of cycles of odd length 3 will be much larger than the number of cycles of odd length greater than 3. This is especially true when the crossing graph corresponds to a layout in which the wiring is limited to grid lines (This is a very popular assumption in many layout models). Figures 5(a) and 5(b) show a cycle of length 3 and 5 respectively in a grid layout model. It is not hard to see that the chances of having a cycle of length 5 are certainly less than those of having a cycle of length 3. This claim has been verified by the experiments carried out by us. Since a cycle of odd length greater than 3 may share some wire segments with a cycle of length 5 as shown in Figure 5(c), it is also possible to break both cycles by simply intending to break the cycle of length 3. Thus, if we break all the cycles of length 3, the number of cycles of odd length will be reduced tremendously. In other words, after all the cycles of length 3 are broken, the crossing graph will be close to a bipartite graph. At this point, the number of vias required for the crossing graph to become 2-colorable is very small.

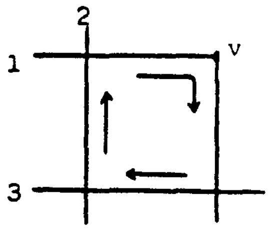


Figure 5(a) A cycle of length 3.

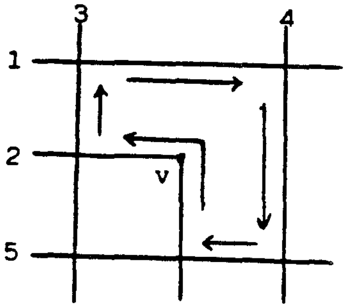


Figure 5(b) A cycle of length 5.

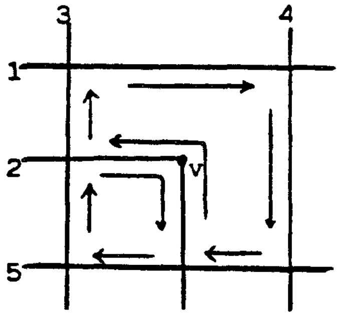


Figure 5(c) Two cycles share some wire segments.

Due to the above reasons, our approach basically consists of 2 phases. The first phase is to find the minimum number of vias to break all the cycles of length 3 in the crossing graph. In the second phase, we then try to break all the cycles of odd length by generating the fundamental cycle set. That is, the algorithms proposed in [14] can be applied in the second phase. Since only a few vias need to be selected in the second phase, the computational time will be fast and the quality of the results will not be greatly affected by the choice of a fundamental cycle set. Recall that the complexity of the algorithm proposed in [14] is $O(p.n^2)$, where p is the number of vias selected and n is the number of wire segments in the final crossing graph. In the following, we will mainly concentrate on the efficient algorithms for the first phase.

For a crossing graph, we first generate all the cycles of length 3 and then select the minimum number of vias to break them. To facilitate the explanation, we first assume that all the nets contain only 2-way split points. The case of having multi-way split points will be addressed in Section 6.

For the purpose of illustration, we describe the idea of the proposed algorithm with an example first. Figure 6 shows a layout example (borrowed from [10]). We label all the nets with numbers and all the crossing points with capital letters. We also label every wire segment (possibly containing a bend) which is long enough to place a via between two crossing points with a lower case letter. That is, every labeled wire segment contains a possible via.

A set of linked lists are created to represent the physical layout relationships among the nets, crossing points, and possible vias. For each net, a linked list is constructed to keep track of all the possible vias in it and all the other nets crossing with it. We can create a linked list by traversing a net from the beginning terminal to the ending terminal and link all the possible vias and all the crossing points as they are encountered. An extra bit is included in each node of the linked list to distinguish a possible via from a crossing point. The linked lists for Figure 6 are shown in Figure 7. For instance, when we traverse net 2 from module M3 to module M2 in Figure 6, we will encounter crossing point A, wire segment a, crossing point M, crossing point N, wire segment i, and crossing point P. This is exactly the sequence of the linked list for net 2 in Figure 7. Note that some wire segments are too short to have vias and they are not labeled. We assume that the terminal points are available in both layers. Therefore, the wire segments which have one end as terminal points are not labeled. If some terminals are only available in one of the two layers, we should label the wire segments which connect these terminals. Since we need to determine which two nets are crossing at a crossing point, a crossing point table is created as shown in Table 1. The corresponding crossing graph is shown in Figure 8.

Now we present an algorithm to find all the cycles of length 3 in the crossing graph. Potentially, any 3 vertices can form a cycle of length 3 and the total number the possible combinations of three different vertices will have the order of $|V|^3$, where $|V|$ is the number of vertices in the crossing graph. However, for any vertex v in V, it can only form a cycle of length 3 with other two vertices u and w that are both adjacent to v and adjacent to each other. After a vertex v has been considered (i.e., all the cycles of length 3 that contain v are generated), for a

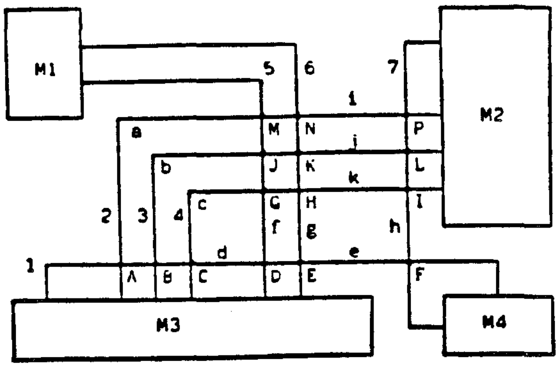


Figure 6: A layout example

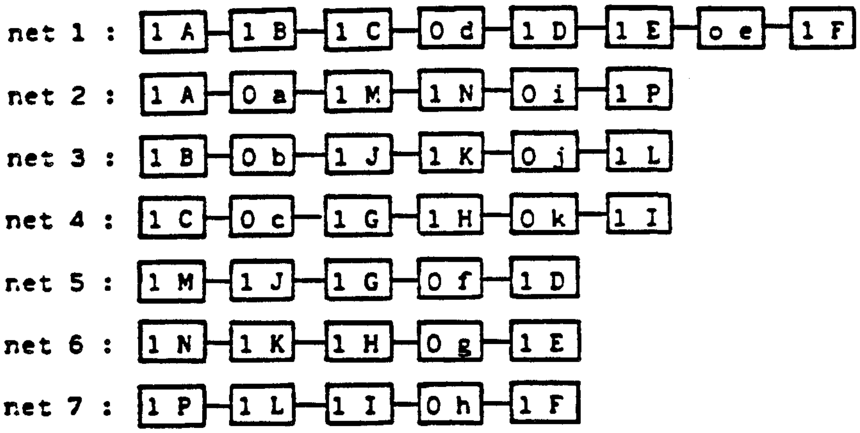

net 1 : 1 A 1 B 1 C 0 d 1 D 1 E 0 e 1 F
net 2 : 1 A 0 a 1 M 1 N 0 i 1 P
net 3 : 1 B 0 b 1 J 1 K 0 j 1 L
net 4 : 1 C 0 c 1 G 1 H 0 k 1 I
net 5 : 1 M 1 J 1 G 0 f 1 D
net 6 : 1 N 1 K 1 H 0 g 1 E
net 7 : 1 P 1 L 1 I 0 h 1 F

Figure 7: Linked lists for Figure 6

crossing point	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P
first net	1	1	1	1	1	1	4	4	4	3	3	3	2	2	2
second net	2	3	4	5	6	7	5	6	7	5	6	7	5	6	7

Table 1: Crossing point table

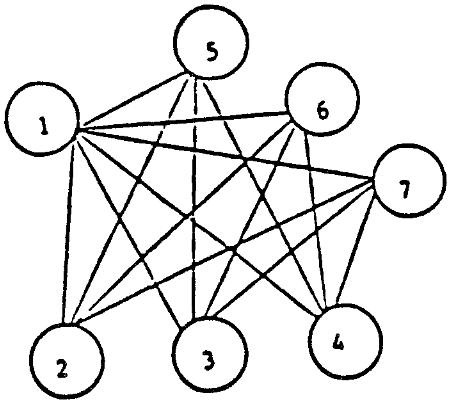


Figure 8: Crossing graph for Figure 6

vertex u which is adjacent to v it is not necessary to test if there is any cycle of length 3 containing both u and v when vertex u is considered. Let d_i represent the degree of vertex v_i in the crossing graph. For any vertex v_i, the number of pairs of vertices that can form a cycle of length 3 with vertex v_i is at most $d_i*(d_i-1)/2$. Let p_i be the number of pairs of vertices to be considered for vertex v_i. Suppose we consider all vertices in the sequence from v_1 to $v_{|V|}$, then $p_1=d_1 * (d_1-1) / 2$ and $p_2 = (d_2 - a_{2,1}) * (d_2 - a_{2,1}-1) / 2$, where $a_{i,j}$ is 1(0) if vertices v_i and v_j are (are not) adjacent. We subtract $a_{2,1}$ from d_2 because vertex v_1 has been tested. In general, for the ith vertex being considered, $p_i = f_i * (f_i-1) / 2$ where $f_i = d_i - (a_{i,i-1} + a_{i,i-2} +\ldots+ a_{i,1})$. This is the same as considering only the upper triangle of the adjacency matrix. In order to reduce the value of $\sum_{i=1}^{|V|-2} p_i$, the sequence of vertices being tested should be considered according to the ascending order of their degrees.

The algorithm immediately follows the above discussion and we will skip the details. The complexity of this algorithm is $O(|V| * d^2)$, where d is the maximum degree of the vertices.

Following the procedure described above, all the cycles of length 3 for Figure 8 can be found as shown in Table 2. Since we want to find all the vias that can be selected to break each cycle of length 3, we need to traverse every polygon formed by each cycle of length 3. For example, the cycle formed by nets 1, 2, and 5 will form a rectangle in Figure 6. If we start with the crossing point A and traverse the polygon clockwise, the crossing points and wire segments encountered are A, a, M, J, G, f, D, d, C, B, and A. Since we are interested in finding the positions of vias, we simply record the labeled wire segments which are traversed. In this example, we have wire segments a, d, and f that can place a via to break the cycle. From Figure 6, it is easy to see that a,d,and f are in nets 2, 1, and 5 respectively, a is located between crossing points A and M, d is between A and D, and f is between M and D. Nets 1, 2 and 5 cross each other. Since the linked lists represent the actual layout, we can get the same information by traversing the linked lists for nets 1, 2, and 5.

After traversing all the polygons formed by all the cycles of length 3, a cover table with a row for each cycle and a column for each labeled wire segment is generated as shown in Table 3. An X is placed at row i and column j if wire segment j is found when traversing the polygon formed by cycle i. In other words, by placing a via in wire segment j, cycle i can be broken. The last row of Table 3 shows the total number of X's in that column. To find minimum number of vias to break all the cycles of length 3 is equivalent to finding the minimum number of columns in Table 3 so that there exists one X in one of the selected columns for every row. This problem is equivalent to the subset cover problem which has been shown to be NP-hard [25]. A simple approximation method is to select the column that contains maximum number of X's repeatedly until every row is covered by an X in one of the columns selected. When a column is selected, the column and the rows that have an X in that column are deleted from the table. In this example, we only have to select column d and all the rows are covered. Therefore, we place a via in the wire segment d and all the cycles of length 3 are broken. Figure 9 shows the crossing graph after net 1 was split into subnets 1a and 1b and causing the crossing graph to become 2-colorable. Figure 10 shows the final layout.

Now we analyze the complexity of the algorithm. It is easy to see that traversing a polygon formed by 3 nets can be done in complexity of $O(|V|)$ since every linked list has at most $2*|V|$ nodes. Finding the minimum number of vias to break all the cycles of length 3 can be done in time complexity of O(c+p) where c is the total number of cycles generated and p is the total number of possible vias. Therefore, the total complexity of the first phase is $O(|V|^2 * d^2)$, since c is less than $|V| * d^2$. It is important to point out that the cover table can also be

Three nets formed a cycle		
1,2,5	1,2,6	1,2,7
1,3,5	1,3,6	1,3,7
1,4,5	1,4,6	1,4,7

Table 2: All the cycles of length 3 in Figure 8

wire segment	a	b	c	d	e	f	g	h	i	j	k
cycle: 1,2,5	x			x		x					
cycle: 1,3,5		x		x		x					
cycle: 1,4,5			x	x		x					
cycle: 1,2,6	x			x			x				
cycle: 1,3,6		x		x			x				
cycle: 1,4,6			x	x			x				
cycle: 1,2,7	x			x	x			x	x		
cycle: 1,3,7		x		x	x			x		x	
cycle: 1,4,7			x	x	x			x			x
Total count	3	3	3	9	3	3	3	3	1	1	1

Table 3: Cover table

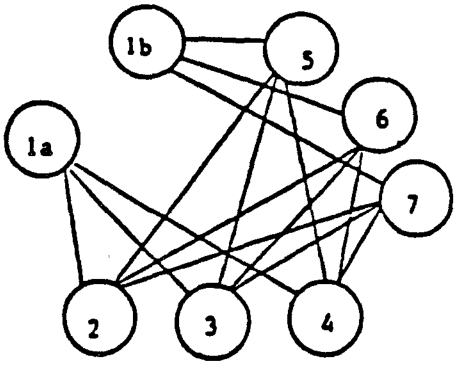


Figure 9: Final crossing graph

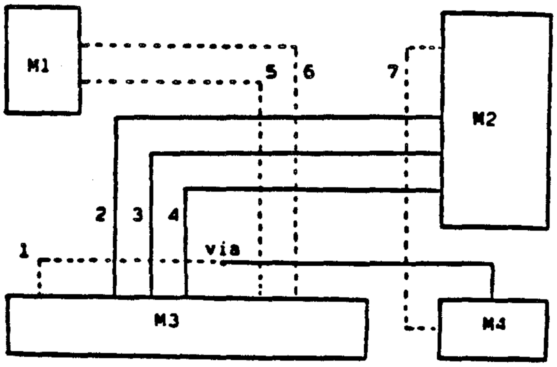


Figure 10: Final layout

implemented by a bipartite graph $G((V_1 \bigcup V_2),E)$, where V_1 and V_2 represent the set of labeled wire segments and the cycles of length 3 respectively. Let $v_1 \in V_1$ and $v_2 \in V_2$. (v_1,v_2) is an edge if and only if cycle v_2 contains wire segment v_1. Under this implementation the storage required for the cover table can be reduced from O(c*p) to O(c+p).

The proposed two phase algorithm has been implemented in Pascal on a VAX 11/780. The results obtained by this algorithm have been compared with those generated by the algorithm proposed in [14] for several instances borrowed from previously published papers in Table 4. The first 6 instances are channel routing instances and the last one is a general layout instance. For convenience we refer to the two phase algorithm proposed in this paper as Method 1 and the one proposed in [14] as Method 2. It can be seen in Table 2 that Method 1 generated optimal results for all seven instances and Method 2 generated optimal results only for those instances which require a small number of vias. The computation time required for Method 1 is always several times faster than that of Method 2. Although there is no guarantee that Method 1 will always generate optimal results, it indeed generates good results. The reason is that in the first phase we select a small set of vias to break all cycles of length 3 and by considering all cycles of length 3 the global property of the layout has been taken care of. After all cycles of length 3 have been broken, the number of vias required to break all cycles of odd length is usually very small which is exactly the case where the algorithm proposed in [14] has good performance.

4. Essential Vias

In the last section, we showed how to generate all the cycles of length 3 and how to traverse the polygons formed by those cycles. When traversing a polygon, if we do not

Example	method 1		method 2		Reference
	no. of vias generated	time in milli-seconds	no. of vias generated	time in milli-seconds	
1	1*	533	1*	766	[20] Fig. 2
2	2*	316	2*	683	[19] Fig. 23
3	9*	1,399	9*	6,916	[19] Fig. 22
4	16*	1,816	17	12,183	[19] Fig. 17
5	40*	10,682	45	57,433	[20] Fig. 25
6	72*	26,000	75	214,666	[20] Fig. 26
7	19*	8,866	21	38,100	[6] Fig. 3
*: optimal result					

Table 4. Comparisons of two methods.

encounter a wire segment in which a via can be placed to break the cycle of odd length, then the layout cannot be realized in two layers. Suppose we found only one wire segment in the polygon that contain a via, we must choose this wire segment. We shall call the via on this wire segment as an "essential via". In general, we define an essential via ev to be of type EV_k if ev is the only possible via in the polygon formed by a cycle of length k (k is an odd number and no less than 3). If we can identify all the essential vias in advance, the nets that contain an essential via can be considered as two (assume it is a two-way split via) subnets that are not crossing each other. By doing this, the number of cycles of odd length can be reduced.

Let's consider the essential vias of type EV_3 first. To find all the essential vias of the EV_3 type, obviously the algorithm described in the last section can be applied. We only have to check if only one wire segment is found when traversing the polygon formed by a cycle of length 3. Due to the fact that the layouts are usually very compacted, in many layout models the possible vias are restricted to the bend points only (e.g., channel routing).

Since we are interested in fast algorithms to identify essential vias such that they can be used as preprocessors for the via minimization algorithms, the above approach is not good enough. Therefore, our objective is set to develop fast heuristic algorithms to identify most of the essential vias instead of all the essential vias. Before we present an algorithm, let us first consider one example.

Consider the layout in Figure 11 which is similar to Figure 6 except that we only have 3 possible vias a, b, and c at the bend points of nets 2, 3, and 4. To identify c as an essential via of EV_3 type, we only have to traverse the smallest rectangle (with points C, D, G, and c as the corners) formed by nets 1, 4, and 5. It is not necessary to traverse any other larger rectangle once we can identify it is indeed an essential via.

Now we propose an efficient algorithm to identify most of the EV_3 type essential vias by restricting the number of polygons needed to be traversed. The major steps of the algorithm are listed below.

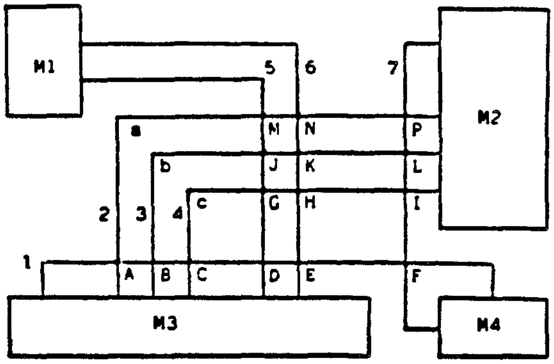


Figure 11: All the possible vias are at the bend points.

Algorithm 1 : Finding essential vias of EV_3 type.

Step 1 :

For each possible via v which has not been checked do Step 2 to Step 4.

Step 2 :

Traverse the linked list (net) t where v is located. Record the nodes traversed before and after visiting node v into an array A and B respectively.

Step 3 :

Let C_{A1} and C_{A2} (C_{B1} and C_{B2}) denote the last (first) two elements in the array A(B). Let N_{A1} and N_{A2} (N_{B1} and N_{B2}) denote the nets that cross net t in crossing points C_{A1} and C_{A2} (C_{B1} and C_{B2}) respectively.

Step 4 :

If nets N_{A2} and N_{B1} cross each other, nets t,N_{A2}, and N_{B1} form a cycle of length 3. Traverse the polygon formed by these three nets. If via v is the only via in the polygon, identify v as an EV_3 essential via and go to Step 1, otherwise do the same check for the nets N_{A2} and N_{B2}, nets N_{A1} and N_{B1}, and nets N_{A1} and N_{B2}.

In Step 3, we assume nodes C_{A1}, C_{A2}, C_{B1}, and C_{B2} are crossing points. If any one of them is a possible via point, the corresponding pair of nets are not tested in Step 4. In Step 4, we only consider the smallest 4 polygons that contain the possible via v. Once we can identify v as an EV_3 essential via, the rest of the pairs need not be tested. To make it clear, we present a layout example (borrowed from [19]) of channel routing[20] as shown in Figure 12.

In a grid routing model, we can embed the layout in a grid graph[18] G(V,E), where V is the set of all the possible vias and all the crossing points, and there is an edge between any two vertices v_i and v_j if and only if v_i and v_j are directly connected by a net. Grid graphs are a simple class of planar graphs for which vertices can be assigned to integer coordinates so that neighbors agree in one coordinate and differ in the other coordinate. Since terminals are available in both layers, they have nothing to do with essential vias and they are not included in the grid graph. Figure 13 shows the grid graph for Figure 12.

All the possible vias that have degree of 1 in the grid graph cannot be essential vias. Hence, possible vias a, b, c, h, m, p, and q are not essential vias. By algorithm 1, we can iden-

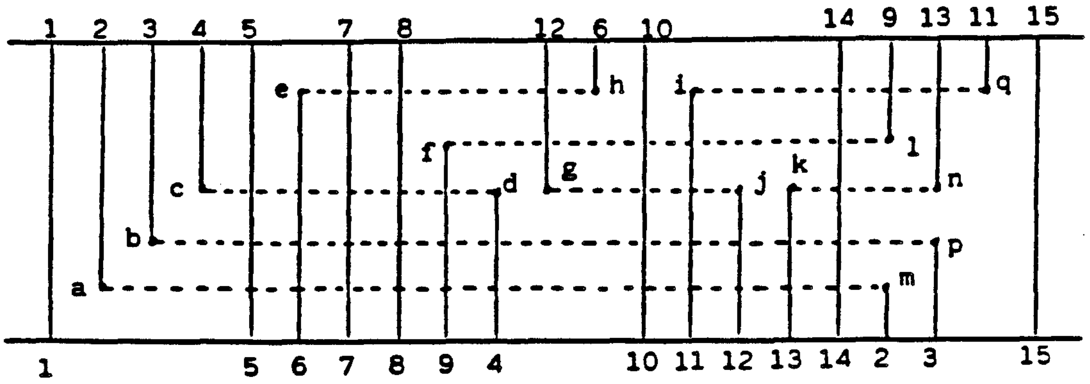


Figure 12: Channel routing example

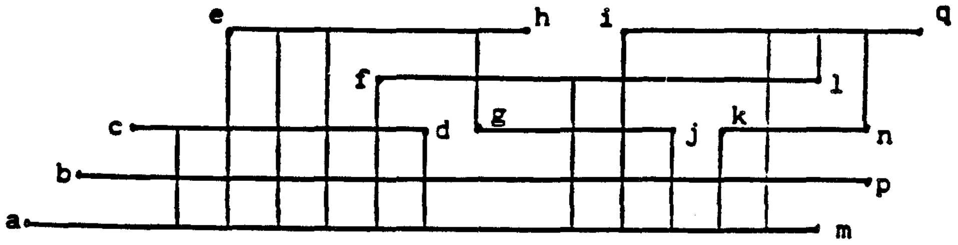


Figure 13: Grid graph for Figure 12

tify possible vias d, e, g, i, j, k, l, and n as EV_3 essential vias in the first test combination of C_{A2} and C_{B1}. The possible via f can be detected as an EV_3 essential via at the fourth test combination. In this particular example, after identifying all those EV_3 essential vias, the layer assignment and via minimization problems are already solved. Figure 14 shows the final layout.

The time complexity of algorithm 1 is $O(p^{*}|V|)$ because we have p possible vias and traversing a polygon can be done in linear time of $|V|$. It is possible to speed up algorithm 1 by using a pointer to every possible via in the linked list and using doubly linked lists so that we can directly traverse the polygon through both sides of a possible via.

Now let us discuss the case of identifying all essential vias of type EV_k where k is greater than 3. In order to do that, we need to find all the cycles of length k. There are two straightforward approaches to find all the cycles of length k in a graph. The first approach is to generate all the cycles in the graph and select the cycles of length k. Johnson[21] has developed an algorithm to generate all the cycles in a graph which has complexity of $O((|V| + |E|)(C+1))$ where C is the number of cycles generated. However, this approach is not practical because the number of cycles in a graph is in the order of $2^{|V|}$. The second approach is to select k vertices in the graph and check if they form a cycle of length k. The number of combinations of k different vertices has the order of $|V|^{min(k,|V|-k)}$, which is also a big number.

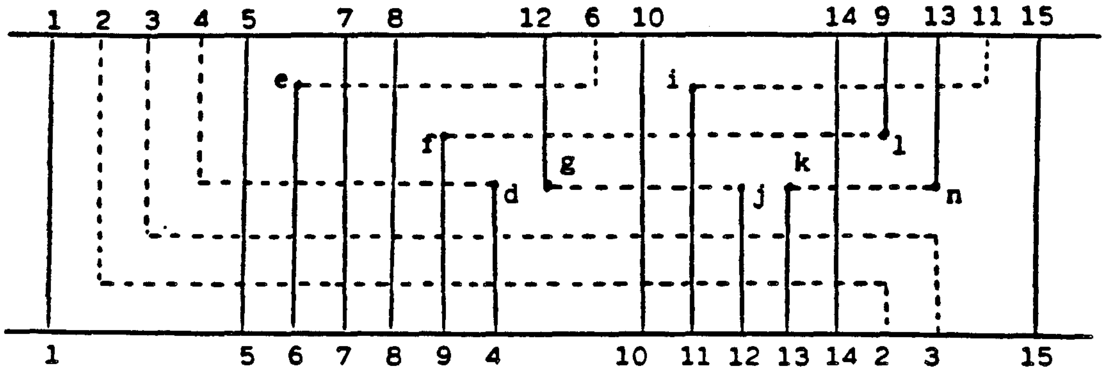


Figure 14: Final layout of Figure 12

The above discussions tell us that a similar approach to the one proposed by us for finding all the cycles of length 3 in the last section should be taken as it is unlikely that an efficient algorithm will be found to generate all the cycles of length k when k is large. Fortunately, the percentage of the number of cycles of length k (for k>3 and k is odd) is small compared with the number of cycles of length 3 and it will also be dramatically reduced if all the cycles of length 3 are broken.

The algorithms proposed in the last section for breaking all the cycles of length 3 and the algorithm proposed in this section for identifying essential vias of type EV_3 can be used as preprocessors for most of the existing via minimization and layer assignment algorithms. Since they are very efficient, the whole process might be speeded up.

Table 5 shows some results generated by algorithm 1 for the same examples used in Table 4. The first 6 layout examples are channel routing instances. 4 of them can be solved 100 percent and 87 percent of the vias required can be found in the other two examples. Example 7 is a general layout, over 84 percent of vias required can be found by Algorithm 1.

5. The UVM Problem

In the UVM problem, the topology of the layout is not fixed and the wire is considered as zero width. Basically both algorithms proposed in [12-13] for solving the UVM problem try to

Example	no. of nets	no. of possible vias	minimum no. of vias required	no. of vias of type EV_3 found by Algorithm 1	time in milli-second
1	10	20	1	1	100
2	5	6	2	2	50
3	15	16	9	9	183
4	15	26	16	16	300
5	21	57	40	35	666
6	45	91	72	70	1,367
7	45	55	19	16	1,150

Table 5: Experimental results of Algorithm 1

find a maximum bipartite subgraph of a circle graph[22]. A circle graph is a graph where the vertices are chords of a circle and 2 vertices are adjacent if and only if the corresponding chords intersect.

Since the UVM problem has been shown to be NP-Complete [13], a heuristic approach is required. Hsu [12] proposed an approximation technique by finding the maximum independent set(MIS) in a circle graph twice [22] while Mareck-Sadowska[13] applied the maximal planarization algorithm of a circle graph[24]. An Independent Set (IS) of a graph G(V,E) is a set of vertices I such that any two vertices in I are not adjacent. An independent set I is a Maximum Independent Set (MIS) if every vertex in I is adjacent to at least one vertex in V-I. However, in general a crossing graph may not be a circle graph. Based on the close relationship between cycles of odd length and bipartite graphs, we propose a new approximation algorithm for the Minimum Node Deletion Bipartite Subgraph problem.

The basic idea is similar to the one proposed for the CVM problem. Initially a crossing graph is constructed. In the first step, we generate all the cycles of length 3 using the algorithm presented in Section 3 and then create a similar cover table except now each column represents a vertex in the graph. In the second step, we use an approximation method to select a minimum number of columns that can cover all the rows and delete the corresponding vertices from the graph. If the crossing graph is not bipartite yet, in the last step, we generate all the fundamental cycles [17] and delete a vertex that is involved in as many fundamental cycles of odd length as possible. Repeat this process until all the fundamental cycles are of even length and the reduced graph is a bipartite graph.

The complexity of the above procedure can be analyzed as follows: We have shown that generating all the cycles of length 3 can be done in $O(|V| * d^2)$ where d is the maximum degree of the vertices in the graph. The second step has time complexity of $O(c+|V|)$, where c is the number of cycles of length 3. The last step has complexity of $O(p^*| V|^2)$ where p is the number of vertices that have to be deleted in the last step which is usually very small. Therefore the total complexity is max($|V| *d^2$, $p*|V|^2$).

Table 6 shows some experimental results of the above approximation algorithm. We test our algorithm by randomly generated graphs with different densities(0.25, 0.5, and 0.75). The density of an undirected graph is defined as the number of edges divided by the number of edges of a complete graph with the same number of vertices. Compared to the results generated by executing the maximum independent set algorithm twice, our algorithm is better for every example. For those cases where the number of vertices is not greater than 15, we were able to verify them manually and found that the results generated by the proposed algorithm are all optimal. Note that the number of nodes deleted by considering only the cycles of length 3 is very close to the total number of nodes that have to be deleted. It matches our previous claim that a graph will be very close to a bipartite graph if all cycles of length 3 are broken.

6. Discussions and Conclusions

We have proposed algorithms for solving both the CVM and UVM problems. For convenience, the algorithms are presented in a basic form. That is, only two-way split points are allowed in the layout and the objective is to minimize the total number of vias required. However, similar approaches as described in [14] can be taken to fit different minimization criteria and to take extra constraints into consideration.

The algorithm proposed in Section 3 for solving the CVM problem can be extended to handle the multi-way split nets by using tree structured linked lists. For a possible via of a multi-way split point, it may be contained in several polygons as shown in Figure 15. When traversing the polygon formed by nets 1, 2, and 4, the possible via v_1 is not an essential via because the polygon contains two other possible vias v_2 and v_3. However, v_1 is an essential via with respect to the polygon formed by nets 1, 2, and 3.

Every node in a linked list can be reduced to only one field if we use a positive or a negative number to indicate the content of the extra bit. Thus, the space required can be lesser.

density	A	our approach			MIS approach	
		B	C	time*	D	time*
0.25	8	0	0	17	1	16
	10	1	0	34	1	17
	12	1	1	34	2	33
	15	2	1	83	3	33
	18	3	2	134	6	34
	20	5	3	183	8	33
	25	7	6	200	10	33
	30	11	9	417	11	50
	35	15	12	716	17	67
0.50	8	1	1	33	1	17
	10	2	2	34	3	17
	12	3	2	84	5	17
	15	5	4	134	7	17
	15	5	4	134	7	17
	18	8	7	167	9	17
	20	9	8	217	11	34
	25	14	11	400	15	34
	30	18	16	550	19	50
	35	21	18	900	22	50
0.75	8	2	2	34	3	17
	10	3	3	50	6	17
	12	4	4	50	7	17
	15	8	7	134	8	17
	18	9	9	167	11	17
	20	11	11	233	13	33
	25	14	14	466	16	50
	30	20	19	984	21	33
	35	24	24	1,900	26	33

A - no. of vertices in the graph.
B - total no. of vertices deleted by our approach.
C - no. of vertices deleted by considering only the cycles of length 3.
D - total no. of vertices deleted by the MIS approach.
*time in milli-second.

Table 6: Experimental results

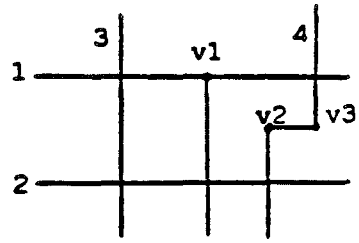


Figure 15: Possible vias with multi-way splits

In some layouts, there are many parallel running nets. Two nets are parallel running nets if they will not intersect each other and the nets they crossed are the same and the crossing points are in the same sequence. For example, nets 2, 3, and 4 (5, 6, and 7) are parallel running nets in Figure 6. We can treat a set of parallel running nets as one net and associate a weight for the equivalent net. The weight of the equivalent net can be defined as the number of parallel running nets since whenever a possible via in the wire segment of the equivalent net is selected, we need a via for every net in the set of the parallel running nets. When a wire segment in the equivalent net is traversed, the entry in the cover table is recorded as the reciprocal of the weight. For the ordinary wire segment 1 is used in the entry of the cover table. If we use a bipartite graph to implement the cover table, the numbers in the cover table are the weights of the edges in the bipartite graph. Our algorithm can be easily modified to handle this case.

The algorithms proposed in [9-10] have worst case complexities at least of $|V|^6$ and $|V|^5$ respectively which are higher than $|V|^{2*}d^2$ proposed in this paper. Therefore, the algorithms proposed in this paper are very good preprocessors for the general CVM algorithms to improve the worst case performance. We have shown that our approach can also be modified to solve the UVM problems. However, for the UVM problems, we may reduce the number of vias at the expense of more wire length and routing area since the wire width is not zero in the realistic case.

Experimental results show that most of the vias that are required can be selected by breaking all the cycles of length 3. We also introduced the idea of essential vias and proposed an efficient algorithm to identify most of the essential vias of type EV_3. For channel routing layouts, we have shown that after identifying the essential vias of the EV_3 type, the via minimization problem is almost done. In the Section 5, we apply the same idea to solve the UVM problem. Experimental results show that our approach generates better solutions than the maximum independent set approach. Intuitively, the approach presented in this paper can be

References

[1] Hashimoto, A. and Stevens, J., "Wire Routing by Optimizing Channel Assignment Within Large Apertures," Proc. 8th Design Automation Workshop, June 1971, pp. 155-169.

[2] Servit, M., "Minimizing the Number of Feedthroughs in Two-layer Printed Boards," Digital Processes, vol. 3, 1977, pp. 177-183.

[3] Stevens, K.R., and VanCleemput, W.M., "Global Via Elimination in Generalized Routing Environment," Proc. 1979 ISCAS, pp. 689-692.

[4] Kajitani, Y., "On Via Hole Minimization of Routing in a 2-layer Board," Proc. IEEE 1980 International Conference on Circuits and Computers, June 1980, pp. 295-298.

[5] Hadlock, F., "Finding a Maximum Cut of a Planar Graph in Polynomial Time," SIAM Journal on Computing, vol. 4, no. 3, Sept. 1975, pp. 221-225.

[6] Ciesielski, M.J., and Kinnen, E., "An Optimum Layer Assignment for Routing in ICs and PCBs", Proc. 18th Design Automation Conference, June 1981, pp. 733-737.

[7] Chen, R.W., Kajitani, Y. and Chan, S.P., "On the Via Minimization Problem for the Two-layer Printed Circuit Board," Conference Records of the 15th Asilomar Conference on Circuits, Systems and Computers, 1981, pp. 22-26.

[8] Chen, R.W., Kajitani, Y, and Chan, S.P., "Topological Considerations of the Via Minimization Problem for Two-layer PC Board," Proc. 1982, ISCAS, pp. 968-971.

[9] Chen, R.W., Kajitani, Y., and Chan, S.P., "A Graph-Theoretic Via Minimization Algorithm for Two Layer Printed Circuit Boards," IEEE Trans. on Circuits and Systems, vol. CAS-30, no. 5, May 1983, pp. 284-299.

[10] Pinter, R.Y., "Optimal layer Assignment for Interconnect," Proc. ISCAS, 1982, pp. 398-401.

[11] Lee, D.T., Hong, S.J., and Wong, C.K., "Number of Vias : A Control Parameter for Global Wiring of High-Density Chips," IBM J. Res. Develop., vol. 25, no. 4, July 1981, pp. 261-271.

[12] Hsu, C.P., Minimum-Via Topological Routing", IEEE Trans. on Computer Aided Design Vol. CAD-2, No.4, October 1983, pp. 235-246.

[13] Marek-Sadowska, M., "An Unconstrained Topological Via Minimization Problem for Two-Layer Routing", IEEE Trans. on Computer Aided Design, Vol. CAD-3, No.3, July 1984, pp. 184-190.

[14] Du, H.C., and Chang, K.C., "A New Approach for Layer Assignment Problem", Technical Report, TR-84-20, Computer Science Department, University of Minnesota, Minneapolis, Minnesota, 1984.

[15] Rengold, E.M., Nievergelt, J., and Deo, N. *Combinatorial Algorithms, Theory and practice*, Prentice-Hall, 1977.

[16] Deo, N., *Graph Theory with Application to Engineering and Computer Science*, Prentice-Hall, 1974.

[17] Paton, K., "An Algorithm for Finding a Fundamental Set of Cycles of a Graph," Comm. ACM, vol. 9, no. 9, Sept. 1969, pp. 514-518.

[18] Hadlock, F.O., "A Shortest Path Algorithm for Grid Graph", Networks, Vol.7, No.4, 1977, pp. 323-334.

[19] Gopal, I.S., Coppersmith, D., and Wong, C.K., "Optimal Wiring of Movable Terminals", IEEE Trans. on Computer Aided Design, Vol. CAD-1, No.1, January 1982, pp. 25-35.

[20] Yoshimura, T. and Kuh, E.S., "Efficient Algorithms for Channel Routing," IEEE Trans. CAD of Integrated Circuits and Systems, vol.CAD-1, no.1, Jan.1982, pp. 25-35.

[21] Johnson, D.B., "Finding All the Elementary Circuits of a Directed Graph", SIAM, J. Computing, Vol.4, No.1, March 1975, pp. 77-84.

[22] Gavril, F., "Algorithms for a Maximum Clique and a Maximum Independent Set of a Circle Graph", Networks, Vol.3, 1973, pp. 261-273.

[23] Garey, M.R., Johnson, D.S., and Stockmeyer, L., "Some Simplified NP-Complete Graph Problems", Theoretical Computer Science, Vol.1, 1976, pp. 237-267.

[24] Chiba, T., Nishioka, I., and Shirakawa, I., "An Algorithm of Maximal Planarization of Graphs", Proc. ISCAS, 1979, pp.649-652.

[25] Garey, M.R. and Johnson, D.S., "Computer and Intractability", San Fransisco, 1979.