QUASI-ELECTRIC FIELDS AND BAND OFFSETS: TEACHING ELECTRONS NEW TRICKS

Nobel Lecture, December 8, 2000

by

HERBERT KROEMER

ECE Department, University of California, Santa Barbara, CA 93106, USA.

I. INTRODUCTION

Heterostructures, as I use the word here, may be defined as heterogeneous semiconductor structures built from two or more different semiconductors, in such a way that the transition region or interface between the different materials plays an essential role in any device action. Often, it may be said that *the interface is the device.*

The participating semiconductors all involve elements from the central portion of the periodic table of the elements (Table I). In the center is silicon, the backbone of modern electronics. Below Si is germanium. Although Ge is rarely used by itself, Ge-Si alloys with a composition-dependent position play an increasingly important role in today's heterostructure technology. In fact, historically this was the first heterostructure device system proposed, although it was also the system that took longest to bring to practical maturity, largely because of the 4 % mismatch between the lattice constants of Si and Ge.

Table I. Central portion of the periodic table of the elements, showing the element from columns II through VI actively used in current heterostructure technology.

II	III	IV	V	VI
	Al	Si	P	S
Zn	Ga	Ge	As	Se
Cd	In		Sb	Te
Hg				

Silicon plays the same central role in electronic metallurgy that steel plays in structural metallurgy. But just as modern structural metallurgy draws on metals other than steel, electronics draws on semiconductors other than silicon, namely, the compound semiconductors. Every element in column III may be combined with every element in column V to form a so-called III–V compound. From the elements shown, twelve different discrete III–V compounds may be formed. The most widely used compound is GaAs – gallium arsenide – but all of them are used in heterostructures, the specific choice depending on the application. In fact, today the III-V compounds are almost always used in heterostructures, rather than in isolation.

Two or more discrete compounds may be used to form alloys. A common example is aluminum-gallium arsenide, $Al_xGa_{1-x}As$, where x is the fraction of column-III sites in the crystal occupied by Al atoms, $1-x$ is occupied by Ga atoms. Hence we have not just 12 discrete compounds, but a continuous range of materials. As a result, it becomes possible to make compositionally graded heterostructures, in which the composition varies continuously rather than abruptly throughout the device structure.

Similar to the III–V compounds, every element shown in column II may be used together with every element in column VI to create II–VI compounds, and again alloying is possible to create a continuous range of the latter.

II. BAND DIAGRAMS AND QUASI-ELECTRIC FORCES

Whenever I teach my semiconductor device physics course, one of the central messages I try to get across early is the importance of energy band diagrams. I often put this in the form of "Kroemer's Lemma of Proven Ignorance":

> If, in discussing a semiconductor problem, you cannot draw an **Energy Band Diagram**, this shows that you don't know what you are talking about,

with the corollary

> If you can draw one, but don't, then your audience won't know what you are talking about.

Nowhere is this more true than in the discussion of heterostructures, and much of the understanding of the latter is based on one's ability to draw their band diagrams – and knowing what they mean.

To illustrate the idea, consider first a homogenous piece of semiconductor, say, a piece of uniformly doped silicon, but with an electric field applied. The band diagram then looks like the top diagram in Fig. 1, consisting simply of two parallel tilted lines representing the conduction and valence band edges. The separation between the two lines is the energy gap of the semiconductor; the slope of the two band edges is the elementary charge e multiplied by the electric field E. When an electron or a hole is placed into this structure, a force $-eE$ is acting on the electron, $+eE$ on the hole; the two forces are equal in magnitude and opposite in direction, their magnitude is the slope of the bands, just the signs differ.

In a heterostructure, the energy gap becomes position-dependent, and the two band edge slopes are no longer equal, hence the two forces are no longer equal in magnitude. It would, for example, be possible to have a force acting only upon one kind of the carriers (Fig. 1b), or to have forces that act in the same direction for both types of carriers (Fig. 1c). Purely electrical forces in homogeneous crystals can never do this. This is why I call these forces "quasi-electric." ***They present a new degree of freedom for the device designer to enable him to obtain effects that are basically impossible to obtain using only "real" electric fields.***

This is the underlying **general design principle** of all heterostructure de-

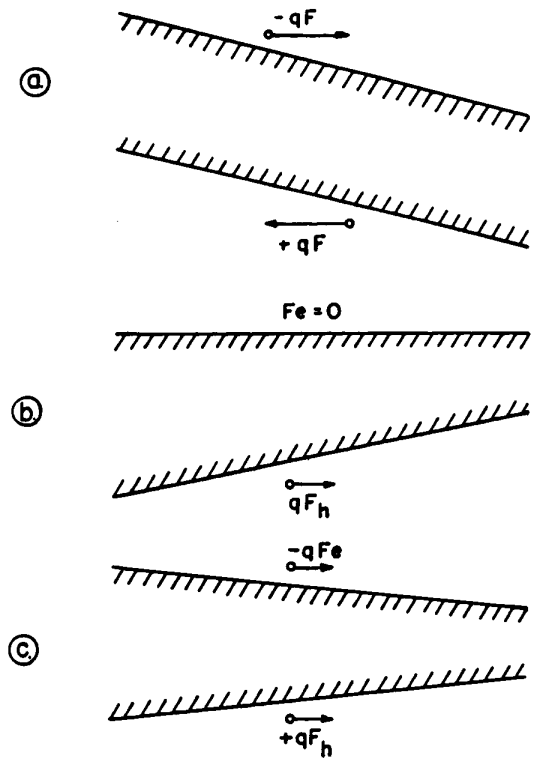

Figure 1. Quasi-Electric Fields: (a) A true electric field simply tilts the bands; (b) quasi-electric fields, with no force on electrons, but a force on holes; (c) quasi-electric fields forcing electrons and holes in the same direction. From Kroemer (1957a).

vices, first spelled out in a 1957 paper of mine (Kroemer, 1957a). In fact, the preceding paragraph is an only slightly edited version of a key paragraph in that paper.

When I wrote those lines, I did not know about Shockley's famous 1951 patent (Shockley), where the possibility of a bipolar transistor with an emitter of wider energy gap is explicitly mentioned. However, the wide-gap emitter idea appears to have been presented principally to cover alternative design possibilities, a procedure typical in patents. The patent gives no indication why such a design would have distinct advantages over a homostructure design, much less a general design principle extending to other kinds of devices. My own formulation might be viewed as a broad generalization of the idea in Shockley's patent. But my point of departure was different: not an *abrupt* energy gap change with accompanying band offset steps, but explicitly a *continuous* energy gap variation of "designable" width, of which the abrupt gap change is simply a limiting case.

Returning to Fig. 1b, it should be emphasized that the zero conduction band slope shown there does not imply a zero electric field. A true electric field is of course present, and it can in principle be determined by the integration of Poisson's equation, provided the local space charge densities are

known, often a non-trivial task. But this true field is not part of the band diagram. Nor do the electrons care: The band edge slopes are what matters, not the true electric field. The difference between the two becomes even more drastic in Fig. 1c, where we could not guess even the *direction* of the true field, much less its magnitude.

III. HETEROSTRUCTURE BIPOLAR TRANSISTORS

A. Graded-gap transistor

I had been led to the 1957 principle by a very practical question dating back to 1953/54, when I was working at the telecommunications research laboratory (Fernmeldetechnisches Zentralamt; FTZ) of the German Postal Service: The early bipolar junction transistors were far too slow for practical applications in telecommunications, and I set myself the task of understanding the frequency limitations theoretically – and what to do about them. One approach – not the only one – was to speed up the flow of the minority carriers from the emitter to the collector by incorporating an electric field into the base region. This could be done by using, not a uniform doping in the base, but one that decreased exponentially from the emitter end to the collector end – the so-called *drift transistor* (Krömer, 1953). While working out the details, I realized that

> "... a drift field may also be generated through a variation of the energy gap itself, by making the base region from a non-stoichiometric mixed crystal of different semiconductors with different energy gaps (for example, Ge-Si), with a composition that varies continuously through the base." [Translated from Krömer (1954)]

This was not yet the full general design principle, but it constituted the original conception of what has become known as the heterostructure bipolar transistor (HBT), and ultimately of the heterostructure device field in general.

The appropriate band diagram (Fig. 2) followed in the 1957 paper mentioned earlier, where I gave the 1954 idea as one example of the general design principle. Note that Fig. 2 shows a flat conduction band, as would be the case for a sufficiently heavy uniform doping; the band diagram of Fig. 1b represents essentially the base region of that early concept. The case of Fig. 1c illustrates the generality of the design principle.

Note that the original proposal explicitly gave the Ge-Si system as an example, rather than a III/V compound system. It was to take some four decades until Ge-Si HBTs were finally becoming commercially available, long after devices based on III/V compounds had done so.

B. Wide-gap emitter

The proposed graded-gap base structure was far beyond the technologies then available, a situation that was to remain unchanged for decades. The only possibility one of my colleagues – Mr. Alfons Hähnlein – could envisage

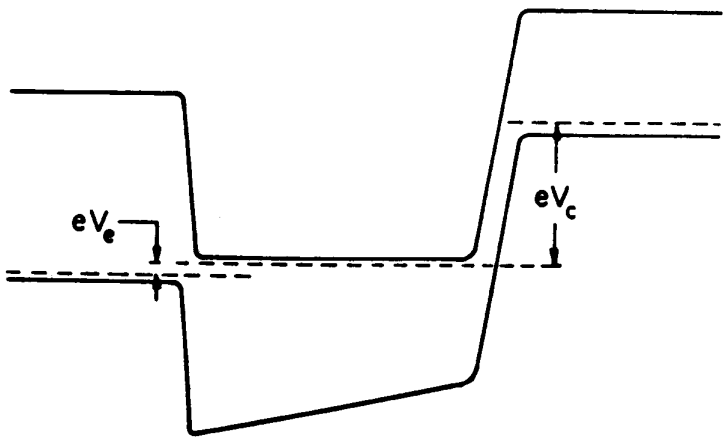

Figure 2. P-n-p transistor with a base region with a graded gap, to speed up minority carrier flow from emitter to collector [from Kroemer (1957a)]. P-n-p transistors were the preferred design for the Ge-based transistors of the mid-50's.

was a design in which the emitter was made from a wider-gap semiconductor than the base, with a quasi-abrupt transition at the interface between the two, leading to a band diagram as in Fig. 3, in essence – but unknowingly – re-inventing Shockley's design.

It was of course obvious that the objective of putting a drift field into the base of the transistor could not be achieved in this way. But on reflecting about what exactly might be the properties of such a structure, I realized that a wide-gap emitter has advantages of its own (Kroemer, 1957b; 1982): One of the problems with all bipolar transistors is minimizing the highly undesirable back-injection of majority carriers from the base (electrons in a p-n-p transis-

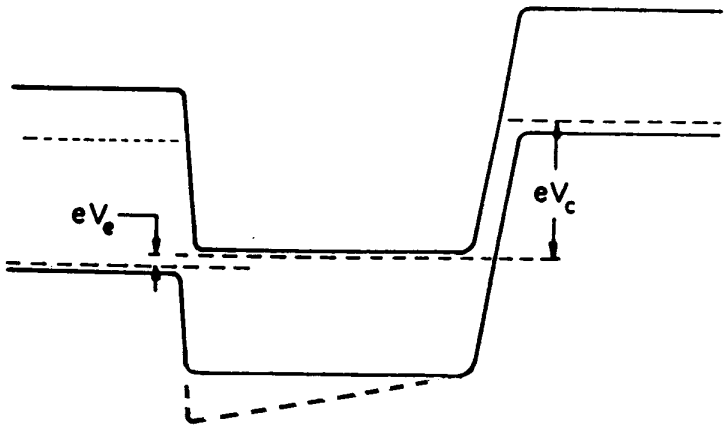

Figure 3. Wide-gap emitter. The energy gap variation has been compressed into a quasi-abrupt transition at the emitter-to-base interface. The base region still has a uniform energy gap without the transport-aiding quasi-field, but there is now a potential barrier for the escape of electrons from the base into the emitter that is larger than the barrier for holes entering the base from the emitter.

tor) into the emitter. In a homojunction transistor, this requirement sharply limits the base doping, which has other undesirable consequences, like a large base access resistance. A wide-gap emitter greatly suppresses this back-injection current: Expressed in terms, not of the quasi-electric forces, but of the associated potentials, any electrons escaping from the base into the emitter must overcome a higher potential barrier than the holes entering the base from the emitter. As a result, the electron escape current density is reduced roughly by a factor $\exp(-\Delta E_G/kT)$, where ΔE_G is the difference in energy gaps. This is very effective: An easily achieved energy gap difference of 0.2eV ($\approx 8kT$) implies a reduction by a factor $e^{-8} \approx 1/3000$.

Given this reduction, it now becomes possible to dope the base much more heavily, to reduce the base resistance. But in the presence of the inevitable junction capacitances, a reduction of base resistance reduces the *RC* time constants of the device, and thereby enhances its speed .

Because of the much greater technological simplicity of the wide-gap emitter design over the graded-base design, it was the wide-gap emitter design that dominated HBT technology until recently, but the highest-performance HBTs now use both approaches (Kroemer, 1983).

C. Follow-up

Because of the absence of any credible technology, I did not follow up the above 1954 ideas until three years later, after I had joined RCA Laboratories in Princeton, NJ. I realized the generality of the design principle outlined above, and wrote the *RCA Review* paper referred to earlier (Kroemer, 1957a). The paper was almost totally ignored, not only because the *RCA Review* was a somewhat obscure journal, but probably even more because I myself somehow never explicitly referred to the paper (nor to its 1954 precursor) in my own subsequent work until about 40 years later (Kroemer, 1996). The general design principle itself was extensively discussed in a 1982 HBT review (Kroemer, 1982), but without reference to the 1954 paper and the 1957 *RCA Review* paper.

The 1957 paper of mine that *is* widely cited was a second paper in that year, which gives a detailed analysis of the wide-gap emitter version of the HBT (Kroemer, 1957b). Having been published in a more visible journal, it drew considerable attention, and stimulated several attempts by others to realize the wide-gap emitter version of the HBT during the '60s. Unfortunately, technology was still not ready, and none of these early attempts led to anything useful. By 1970, people seemed to have largely given up.

While at RCA, I also made an unsuccessful attempt to build a Ge transistor with a Ge-Si alloy emitter, which might be sufficiently amusing (and characteristic of the primitive state of 1957 technology) to be told here (Kroemer, 1957c). The idea was to utilize the fact that the Au-Si phase diagram exhibits a low-melting (370 °C) eutectic. I prepared such a eutectic, smashed the fairly brittle material with a hammer into a coarse powder, placed small grains of the powder onto a Ge chip, and alloyed the combination at a temperature somewhere between 500 °C and 600 °C. The Au-Si alloy would then melt and

penetrate into the Ge chip, dissolving some Ge. Upon cooling, a Ge-Si alloy emitter would re-crystallize (Fig. 4). I actually got one or two transistors to work, but as a rule, the large thermal strains generated during the solidification of the eutectic caused the Ge chip to crack. The attempt was sufficiently unsuccessful that I never published the work. It was followed up by Diedrich and Jötten (1961), who knew about my work, but the technology clearly was unpromising, and Si-Ge HBTs had to wait several decades for their practical realization.

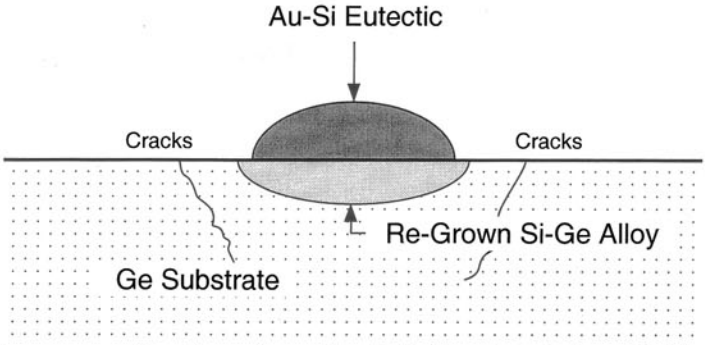

Figure 4. Attempt to realize a Ge transistor with a Ge-Si alloy emitter. A piece of Au-Si eutectic was alloyed into a Ge base, forming a Si-Ge alloy emitter upon cooling. From Kroemer (1957c).

IV. DOUBLE-HETEROSTRUCTURE LASER

Neither the graded-gap HBT nor the wide-gap emitter HBT draw on the full power of the idea expressed in the general design principle that the quasi-electric fields '***enable the device designer to obtain effects that are basically impossible to obtain using only "real" electric fields.***' They represent major improvements, alright, but do they represent something *basically impossible* otherwise?

An example of something that was indeed truly impossible to achieve otherwise emerged abruptly in March 1963. I was working at Varian Associates in Palo Alto at the time, and a colleague of mine – Dr. Sol Miller – had taken a strong interest in the new semiconductor junction lasers that had emerged in 1962, a topic then outside my own range of interests. In a colloquium on the topic he gave a beautiful review of what had been achieved, not failing to point out that successful laser action required either low temperatures or short low-duty-cycle pulses, usually both. Asked what the chances were to achieve continuous operation at room temperature, Miller replied that certain experts had concluded that this was fundamentally impossible.

It is instructive to review this argument here. Consider the (highly oversimplified) energy band diagram of a GaAs p-n junction, heavily doped on both sides, and forward-biased to the point that flatband conditions were reached (Fig. 5). Electrons then diffuse from the n-type side to the p-type side, and holes diffuse in the opposite direction, creating a certain concentration of electron-hole pairs in the junction region proper; their recombination would cause light emission. But in order to obtain *laser* action, a popula-

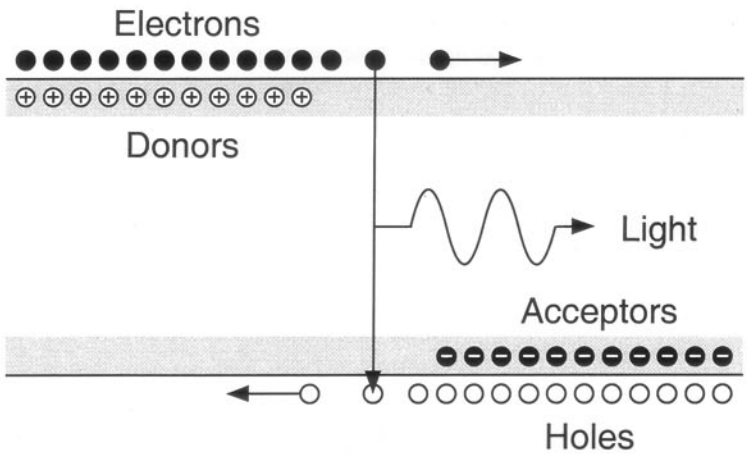

Figure 5. Schematic energy band diagram of a p-n homojunction forward-biased to flatband conditions, creating a high concentration of electron-hole pairs in the vicinity of the junction plane, leading to emission of recombination radiation.

tion *inversion* has to be achieved, which means that, in the active region, the occupation probability of the lowest states in the conduction band has to be higher than that of the highest states in the valence band. A *necessary* condition for such a population inversion is a forward bias larger than the energy gap. But even then, a population inversion is hard to achieve in an ordinary p-n junction. First of all, the electron concentration in the active region will always be lower than in the n-type doped region, with an analogous limitation for the holes. Inversion, therefore, requires degenerate doping on both sides. But even with degenerate doping, both the electrons and holes would diffuse out of the active region immediately into the adjacent oppositely doped region, preventing a population inversion from building up. Increasing the forward bias would not help much, because it would increase the rate of outflow just as much as the rate of injection.

I immediately protested against this argument with words somewhat like "but that is a pile of ..., all one has to do is give the injector regions a wider energy gap ." As is shown in Fig. 6, such a change would cause an electron-repelling quasi-electric field to be present on the p^+ side, and a similar hole-repelling barrier on the n^+ side. Carrier confinement would thus be achieved.

By increasing the forward bias further, potential wells develop for both the electrons and the holes (Fig. 7), with quasi-electric forces on *both* sides pushing *both* electrons and holes towards the active region. As a result, electron and hole concentrations can become much larger than the doping levels in the contact regions, and it becomes readily possible to create the population inversion necessary for laser action. This double-heterostructure (DH) laser finally represented a device truly impossible with only the real electric fields available in homostructures; note that the idea for it arose essentially at the instant I had been made aware that there was a problem.

I wrote up a paper describing the DH idea, along with a patent application.

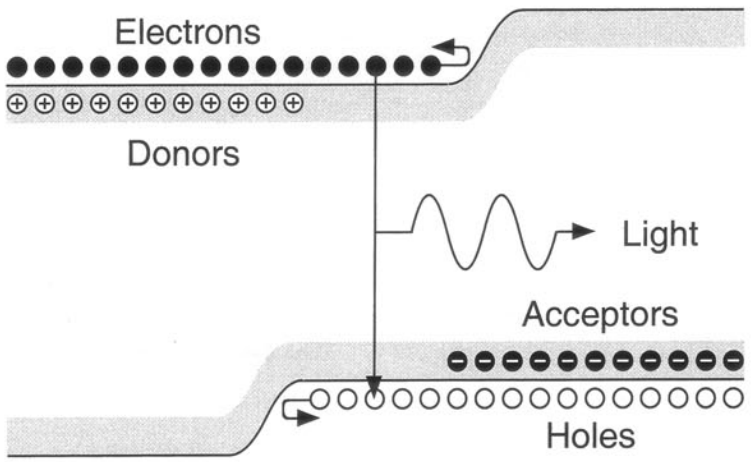

Figure 6. Carrier confinement in a double heterostructure, due to the presence of quasi-electric potential barriers at the ends of the light-emitting active region, preventing the outflow of injected electrons and holes, without interfering with the flow of majority carriers from the injector regions.

The paper was submitted to *Applied Physics Letters*, where it was rejected. I was persuaded not to fight the rejection, but to submit the paper to the *Proceedings of the IEEE* instead, where it was published (Kroemer, 1963) – but largely ignored. Fig. 8 shows the band diagram actually published.

The patent was issued in 1967 (Kroemer, 1967). It is probably a better paper than the *Proc. IEEE* letter. It expired in 1985.

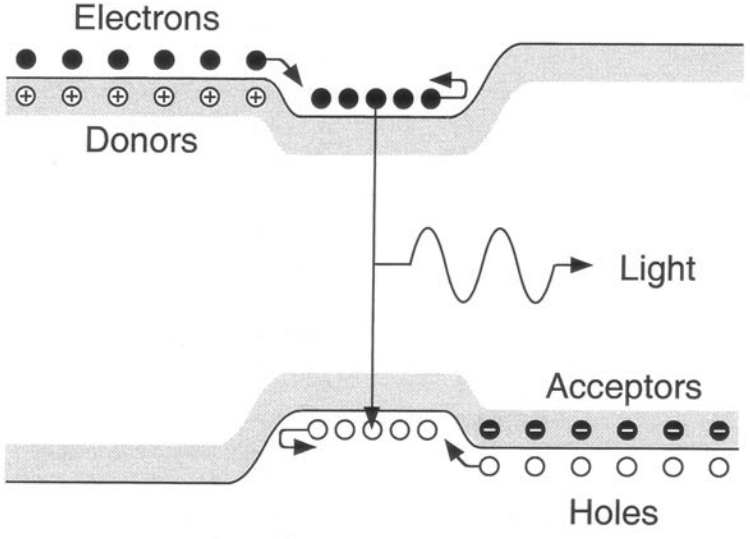

Figure 7. With a further increase of the forward bias, potential wells form for both electrons and holes, which permit the accumulation of the injected carriers to degenerate concentrations much higher than the values in the injector regions.

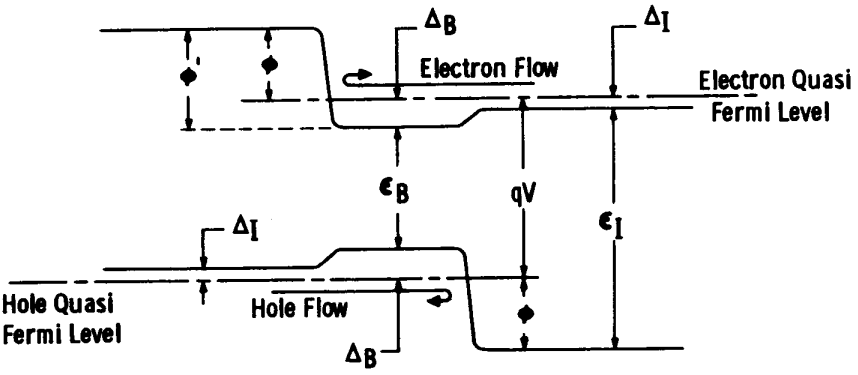

Figure 8. Band diagram of the double-heterostructure laser, as originally published (Kroemer, 1963).

Once again, here was an idea far ahead of any technology to realize it. DH lasers operating continuously at room temperature were finally demonstrated in 1970, first by Alferov et al. (1970), and shortly afterwards by Hayashi *et al.* (1970). For the history of the experimental work, see Alferov (2001); Alferov (1996); Casey and Panish (1978).

For reasons discussed below, I myself was not able to be a participant in the technological realization of the idea. For the next 10 years I worked on research on the Gunn effect, to return to heterostructures in the mid-70s.

V. ON HOW NOT TO JUDGE NEW TECHNOLOGY

When I proposed to develop the technology for the DH laser, I was refused the resources to do so, on the grounds that "this device could not possibly have any practical applications," or words to that effect. By hindsight, it is of course obvious just how wrong this assessment was.

It was really a classical case of judging a fundamentally new technology, not by what new applications it might *create*, but merely by what it might do for already-existing applications. This is extraordinarily short-sighted, but the problem is pervasive, as old as technology itself. The DH laser was simply another example in a long chain of similar examples. Nor will it be the last. I therefore believe it is worthwhile to say a few words about this kind of argument here.

Any detailed look at history provides staggering evidence for what I have called, on another occasion (Kroemer, 1995), the *Lemma of New Technology*:

> The principal applications of any
> sufficiently new and innovative technology always have been
> – and will continue to be –
> applications *created* by that technology.

As a rule, such applications have indeed arisen – the DH laser is just a good recent example – although usually not immediately.

But this means that we must take a long-term look when judging the applications potential of any new technology: It must *not* be judged simply by how it might fit into already existing applications, where the new discovery may have little chance to be used in the face of competition with already-entrenched technology. Dismissing it on the grounds that it has no known applications will only stifle progress towards those applications that *will* grow out of that technology.

I do not think we can realistically predict which new devices and applications may emerge, but I believe we can create an environment encouraging progress, by not always asking immediately what any new science might be good for (and cutting off the funds if no answer full of fanciful promises is forthcoming). In particular, we must educate our funding agencies about this historical fact. This may not be easy, but it is necessary. We must make it an acceptable answer to the quest for applications to defer that answer, and that at the very least a search for applications should be considered a part of the research itself, rather than a result to be promised in advance. Nobody has expressed this last point better than David Mermin in his recent put-down of so-called "strategic research" (Mermin, 1999):

"I am awaiting the day when people remember the fact that discovery does not work by deciding what you want and then discovering it."

What is *never* acceptable – and what we must refrain from doing – is an attempt to justify the research by promising credibility-stretching mythical improvements in *existing* applications. Most such claims are not likely to be realistic and are easily refuted; they only trigger criticism of just how unrealistic the promises are, thereby discrediting the whole work.

Ultimately, progress in applications is not *deterministic*, but *opportunistic*, exploiting for new applications whatever new science and technology happen to be coming along.

VI. CONSTRAINTS

1. Lattice Matching

Let me now turn to some of the problems in implementing heterostructures.

When two materials with significantly different lattice parameters are grown upon each other, whether graded or not, huge strains rapidly build up with increasing thickness, and eventually misfit dislocations will form, a defect without any redeeming features. As a result, the need for lattice matching is all but obvious. The problem is somewhat less severe in modern structures calling for very thin layers (see below); but even there, the lattice-matched case serves as the conceptual point of departure.

Historically, the importance of lattice matching was recognized almost from the beginning, especially for bipolar devices such as lasers. In my 1967 DH laser patent (Kroemer, 1967), I gave a table listing numerous semiconductors in the order of increasing lattice parameter (see Table II); the accompanying text in the patent called for semiconductor pairs with a lattice

mismatch below 0.01Å (≈ 0.2 %) as the most promising ones, indicating a recognition of the stringency of the lattice matching demand. The possibility to achieve lattice matching by alloying was explicitly recognized, though.

Table II. Partial copy of the 1963 table of semiconductors ordered by lattice constant (second column) from ref. (Kroemer, 1967). The third column gives the increase in lattice constant relative to the preceding material. Note that no distinction is made between column-IV elements, the III-V compounds, and the II-VI compounds. Also, the 1963 lattice constant of AlAs was significantly in error: The correct room-temperature value (5.661Å) is actually 0.02Å *larger* than the GaAs value, and the difference is much less at typical crystal growth temperatures. [Only the semiconductors up to ZnSe are shown here; the complete 1963 table can be found in Kroemer (1996)].

Semiconductor	a [Å]	Δa [Å]
ZnS	5.406	
Si	5.428	.022
GaP	5.450	.022
AlP	5.46	.01
AlAs	5.63	.17
GaAs	5.653	.02
Ge	5.658	.005
ZnSe	5.667	.009
...	...	...

Ironically, the 1963 literature value for the lattice constant of AlAs was incorrect. As a result, the GaAs-AlAs pair initially did not seem to meet the proposed stringent criterion, and the known poor stability of (binary) AlAs against oxygen did not help. It took some time to recognize its promise, not so much as a binary material, but as an alloy with GaAs, which greatly reduced the oxidation problems, and reduced the lattice mismatch to a completely negligible level.

A more instructive way to represent the information of Table II, including energy gaps as well, is in terms of what some of us call *The Map of the World*, a display of the energy gaps of semiconductors of interest vs. their lattice constants (Fig. 9), with interconnect lines shown to represent binary alloys.

Much of the reason for the continued dominance of the (Al,Ga)As alloy system in heterostructure studies is precisely the "Great Crystallographic Accident" that AlAs and GaAs have essentially the same lattice parameter. This natural lattice matching means, in particular, that an ideal substrate is readily available for the growth of such heterostructures, namely bulk GaAs, obtainable as high-quality single crystals with low dislocation densities, especially in semi-insulating form. If there remains *one* bad aspect to the (Al,Ga)As system, it is the obnoxious chemical affinity of aluminum to oxygen, the source of many residual defects in (Al,Ga)As. Following a 1983 suggestion by myself (Kroemer, 1983), the use of (Ga,In)P lattice-matched to GaAs has recently drawn some attention as an alternative to (Al,Ga)As, especially in HBTs, for which the band lineups at the (Ga,In)P-GaAs interface are more favorable than those of (Al,Ga)As-GaAs.

A second natural substrate is InP, widely used for both optoelectronic and

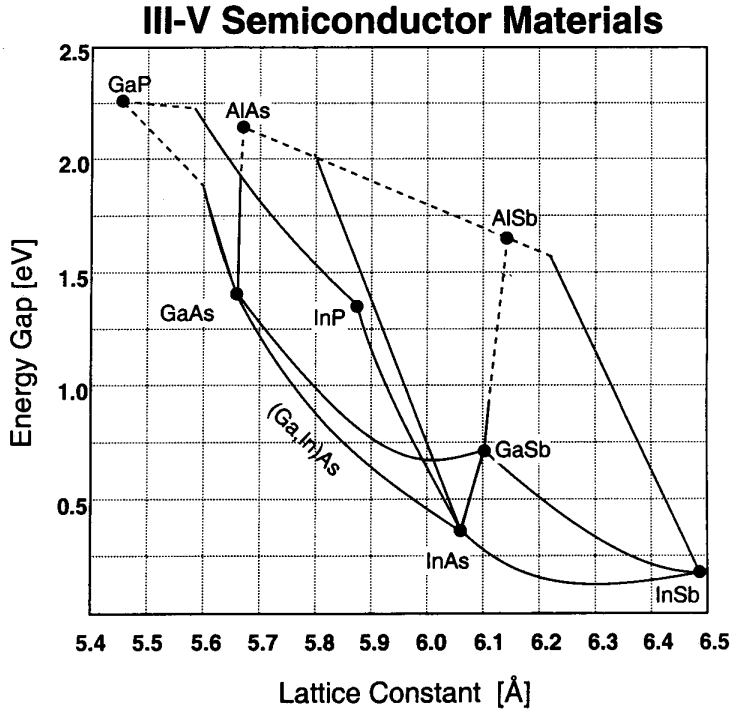

Figure 9. Partial "Map of the World," plotting the energy gap of various III-V compounds vs. lattice constant. The map omits the "Old-World Continents" of the column-IV and the II-VI semiconductors, and the "New World" of the nitrides.

high-speed device applications that call for energy gaps less than that of GaAs. There is no binary III-V compound lattice-matched to InP, but InP is widely used in devices, combined with a wide variety of alloys ranging from (Ga,In)As to Al(As,Sb).

With the emergence of quantum wells, superlattices, and other structures calling for very thin layers, the issue of strain induced by lattice mismatch has lost some of its tyrannical dominance. In sufficiently thin structures, remarkably large strains can be accommodated without dislocation formation, to the point that the modification of the energy band structure of a heterostructure by *deliberate* introduction of strain has become an important device design principle in its own right. The recent evolution of successful Si-Ge HBTs is perhaps the most dramatic triumph of this idea (see, for example, Abstreiter (1996); Kõnig (1996), but other examples are close behind, both in field-effect transistors (FETs) and in photonic devices. Some of the recent developments in self-assembling quantum dots are explicitly based on utilizing strain already during the crystal growth process.

2. Valence Matching

If lattice matching were the only constraint, the Ge-GaAs system would be the ideal hetero-system, as was in fact believed by some of us – including myself – in the early-'60s. At that time, the most successful heterojunctions that had

been demonstrated were the Ge-on-GaAs heterojunctions studied by Anderson (1960), suggesting a bright future for this system (the term *heterojunction* seems to have appeared first in Anderson's papers). Table II reflects this idea, in the form of combining III–V compounds, II–VI compounds, and group-IV semiconductors into a common table, making the GaAs-Ge system appear to be the most promising candidate It took a few years to realize that this was a blind alley – and why.

It is not a questions of chemical incompatibility, or even of cross-doping effects. Covalent bonds between Ge on the one hand, and Ga or As on the other are readily formed, but they are what I would like to call *valence-mismatched*, meaning that the number of electrons provided by the atoms is not equal to the canonical number of exactly two electrons per covalent bond. Hence the bonds themselves are not electrically neutral, as first pointed out in a 1978 "must-read paper" by Harrison *et al.* (1978).

Consider a hypothetical idealized (001)-oriented interface between Ge and GaAs, with Ge to the left of a mathematical plane, and GaAs to the right (Fig. 10). In GaAs, an As atom brings along 5 electrons (= 5/4 electrons per bond), and expects to be surrounded by 4 Ga atoms, each of which brings along 3 electrons (3/4 per bond), adding up to the correct number of 8/4 = 2 electrons per Ga-As covalent bond. But when, at a (001) interface, an As atom has two Ge atoms as bonding partners, each Ge atom brings along 1 electron per bond, which is one-half electron too many. Loosely speaking, the As atom "does not know" whether it is a constituent of GaAs, or a donor in Ge.

As a result, each Ge-As bond acts as a donor with a fractional charge, and each Ge-Ga bond as an acceptor with the opposite fractional charge. To be electrically neutral, a Ge-GaAs interface would have to have equal numbers of both charges, not only averaged over large distances, but locally. Given chemical bonding preferences, such an arrangement will not occur naturally during epitaxial growth. If only one kind of bonds were present, as in Fig. 10, the interface charge would support an electric field of 4×10^7 V/cm. Such a huge field would force atomic re-arrangements during growth, trying to equalize the number of Ge-As and Ge-Ga bonds. However, these re-arrangements will never go to completion, but will leave behind ill-defined locally fluctuating residual charges, with deleterious consequences for any device application. Interfaces with perfect bond charge cancellation are readily drawn on paper; but in practice there are always going to remain some local deviations from the perfect charge compensation, leading to performance-degrading random potential fluctuations along the interface.

Although Harrison *et al.* discuss only the GaAs-Ge interface, their argument applies to other interfaces combining semiconductors from different columns of the periodic table. In the specific case of compound semiconductor growth on a column-IV elemental semiconductor, the additional problem of antiphase domains on the compound side arises (see, for example, Kroemer (1987)).

The above discussion pertained to the most-widely used (001)-oriented interface. The interface charge at a valence-mismatched interface actually de-

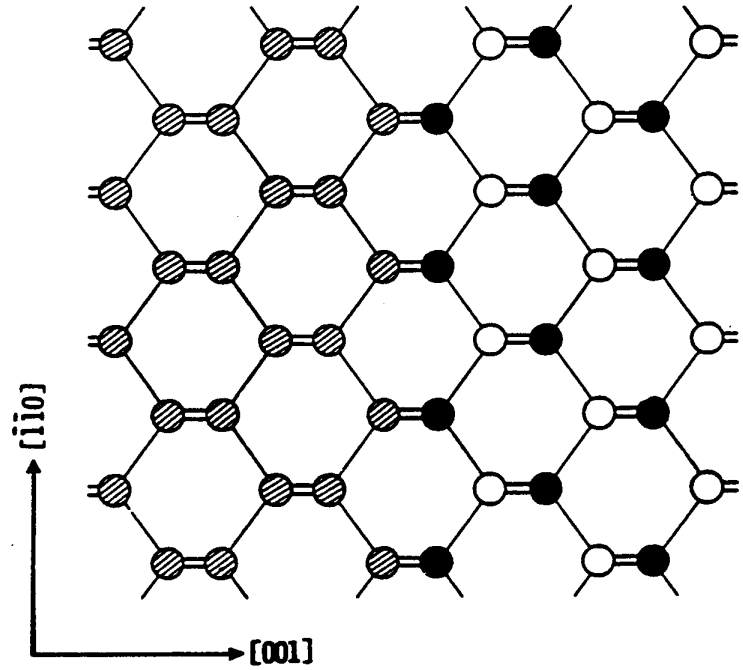

Figure 10. Departure from electrical neutrality at a "mathematically planar" (001)-oriented Ge/GaAs interface. The different atomic species – Ga or As atoms (white and black circles) and Ge atoms (shaded circles) – do not bring along the correct number of electrons to form electrically neutral Ga-Ge or As-Ge covalent bonds of 2 electrons per bond. From Harrison *et al.* (1978).

pends on the crystallographic orientation. It has been shown by Wright *et al.* that an ideal (112) interface exhibits neither an interface charge, nor antiphase domains, and it was in fact possible to demonstrate GaP-on-Si interfaces that had a sufficiently low defect density that they operated as emitters in a GaP-on-Si HBT (Wright *et al.*, 1982; 1984). However, the performance was still sufficiently poor that the approach was not pursued further.

VII. MOLECULAR BEAM EPITAXY AND ABRUPT HETEROSTRUCTURES

The 1970 DH laser demonstration was accomplished by liquid-phase epitaxy (LPE), a beautifully simple technology, but with severe limitations. The big technological breakthrough for heterostructures came only with the emergence of molecular beam epitaxy (MBE) as a practical crystal growth technology, largely pioneered by Al Cho (followed later by organometallic vapor phase epitaxy). In contrast to LPE, MBE permitted combining a wide range of semiconductors, even such hetero-valent combinations as GaP and GaAs on Si. Moreover, it offered a very high degree of control over the local composition, almost on an atomic layer scale. Suddenly, we could realize experimentally almost any band diagram we could draw, at least in the growth direction (lateral control on a similar scale remains an elusive goal to this day). By 1980, the progress in heterostructures had been so large, that I was able to

give an invited paper the provocative title "Heterostructures for Everything: Device Principle of the 1980's?" (Kroemer, 1981). It turned out to be an accurate prediction.

In particular, it had become possible to grow almost atomically abrupt heterojunctions. This also meant that two heterojunctions could be placed sufficiently closely together that quantum effects in the space between them became important, and could be utilized for new kinds of devices. The most obvious development was that of quantum wells (QWs), especially for laser applications, which soon became dominated by QW lasers. But we also saw an increasing use of heterostructures in non-bipolar applications, in effect applying the general quasi-electric field design principle outside its range of origin.

One such example is the use of pairs of tunneling barriers in resonant-tunneling diodes, for application as high-frequency sources up into the sub-terahertz frequency range. Another is the idea of Esaki and Tsu to use a periodic heterostructure superlattice as a quasi-bulk negative-resistance medium with an even higher frequency limit (Esaki and Tsu, 1970). It has so far remained an elusive goal, but it continues to be a very active field of research (including by myself).

I would like to single out here a less obvious new concept, that of *modulation doping*, due to Dingle *et al.* (1978). Consider a heterojunction in which only the side with the higher conduction band is doped (Fig. 11). The downward quasi-electric potential step at the interface will cause electrons to drain into the lower conduction band on the other side. Once they are past the range of the quasi-electric potential step associated with the abrupt hetero-interface itself, the electrons still see the ordinary electric field associate with

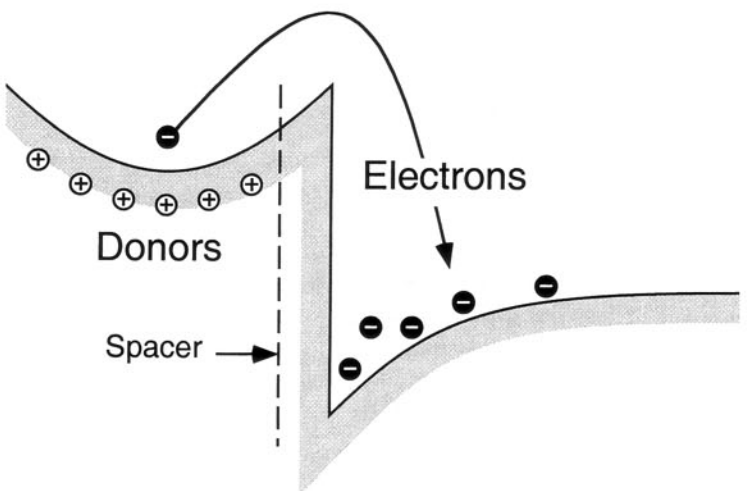

Figure 11. Modulation doping. At an abrupt heterojunction, electrons contributed by donors on the higher-energy side drain onto the lower-energy side, creating a quasi-two-dimensional electron gas there. Because the electrons are now spatially separated from the donors, impurity scattering is reduced, especially if an undoped spacer is inserted on the higher-energy side. The band curvature shown is due to the space charges on the two sides of the interface.

the Coulomb attraction by the donors left behind on the other side. It pulls the electrons towards the interface, creating a 2-dimensional electron gas (2DEG) inside a roughly triangular quantum well. Moreover – and most importantly – because the electrons have been spatially separated from "their" donors, impurity scattering is reduced, and the electron mobility is enhanced. To maximize these benefits, an undoped spacer region is left adjacent to the interface.

The idea had extremely far-reaching consequences, both for devices, and in basic solid-state physics. In devices, it formed the basis of a new class of field effect transistors (FETs), commonly referred to as HEMTs, meaning *H*igh-*E*lectron-*M*obility *T*ransistors (Mimura *et al.*, 1980; Delagebeaudeuf *et al.*, 1980). Their properties are superior to those of earlier classes of FETs. Because of their low noise, they are now used as the sensitive input stage in cellular phones, and thus have contributed to the explosive growth of this aspect of modern information technology.

In basic physics, the suppression of impurity scattering by modulation doping with optimized spacers has permitted the achievement of huge low-temperature mobilities. There is a direct path from the idea of modulation doping to the discovery of the fractional quantum Hall effect, by Tsui, Störmer, and Gossard (Tsui *et al.*, 1982; Stormer, 1999), in 2DEG samples of unprecedented structural perfection grown by Gossard. The subsequent theoretical interpretation of the effect by Laughlin (1999) revealed it as a true fundamental breakthrough in solid-state physics, for which Tsui, Störmer, and Laughlin received the 1998 Nobel Prize in Physics. Unfortunately, the Nobel statute prohibition against dividing the prize amongst more than three individuals excluded Gossard from sharing in the award.

VIII. BAND OFFSETS

In wake of the emergence of MBE technology in the early-70s, my own research returned to heterostructure problems, especially to the problem of band offsets at abrupt heterojunctions. In that limit, the energy band structure makes a discontinuous transition, and exactly how the bands on the two sides are lined up becomes a central question, both experimentally and theoretically. One of the reasons all my early device band diagrams show graded transitions was to sidestep this question of band lineups, of which I was actually well aware.

A. Offset Types

Given two semiconductors, there are evidently three different band lineups possible (Fig. 12)

1. Straddling Lineups

The most common lineup is the straddling one, with conduction and valence band offsets of opposite sign. It is, in essence, the abrupt limit of the graded band structure of Fig. 1c. In quantum wells and superlattices made from such

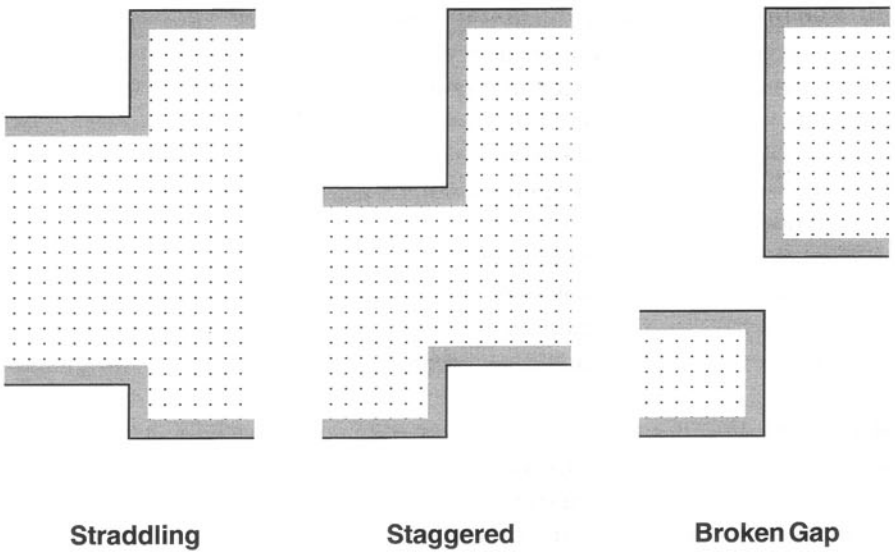

Figure 12. Straddling, staggered, and broken-gap band lineups.

pairs, the lowest conduction band states occur in the same part of the structure as the highest valence band states, which makes these pairs of particular interest for opto-electronic applications, like lasers, which are bipolar kinds of devices, with both electrons and holes involved in the device operation. The two kinds of carriers then occur in the same layers; hence such structures are sometimes referred to as *spatially direct.* Many of today's opto-electronic devices, such as quantum well lasers, are based on such a lineup. The most-widely studied heterojunction system, GaAs-(Al,Ga)As, is of this kind, as are a number of other systems, for example, (Ga,In)As lattice-matched to InP, and (Ga,In)P lattice-matched to GaAs.

2. Staggered Lineups

For some materials pairs, the two bands are shifted in the same direction, leading to a band structure in which the lowest conduction band minimum occurs on one of the sides, the highest valence band maximum on the other, with an energy separation between the two less than the lower of the two bulk gaps. The combination of AlAs-$Al_xGa_{1-x}As$ for $x > 0.3$ is of this kind, as is (Al,In)As lattice-matched to InP; there are several others. In bipolar structures with this lineup, the electrons and holes are confined to *different* layers, hence these structures are *spatially indirect.* Nevertheless, the wave functions overlap at the interface, making radiative recombination possible, with a photon energy less than the narrower of the two gaps (Kroemer and Griffiths, 1983; Caine *et al.*, 1984).

Staggered lineups imply large band offsets in either the conduction or the valence band, and for some applications this property is more important than the spatial indirectness. For example, the conduction band lineup at the InAs-AlSb interface, 1.35eV (Nakagawa *et al.*, 1989), is the highest that has

been reported for any III-V system, and several applications are based on this property, along with the low electron effective mass in InAs. The fastest resonant tunneling diode reported in the literature (Brown *et al.*, 1991), oscillating up to 712 GHz, was based on this system.

The high barriers also offer superb electron confinement in FETs, and the possibility of achieving extremely high levels of electron concentration (approaching $10^{13}cm^{-2}$) by modulation doping (i. e., putting the donors into the barriers rather than into the wells), while retaining high mobilities. This combination makes the InAs-AlSb system ideal for investigating the properties of quantum wells in the metallic limit, for example as coupling medium in a new class of superconducting weak links (Kroemer *et al.*, 1994).

3. Broken-Gap Lineup

If a staggered lineup is carried to its extreme, the result is a broken-gap lineup, in which the bottom of the conduction band on one side drops below the top of the valence band on the other. There exists at least one nearly-lattice-matched pair of this kind, InAs-GaSb, with a break in the forbidden gap at the interface on the order of 150 meV (Sakaki *et al.*, 1977).

The broken-gap InAs-GaSb lineup by itself is an exotic lineup, of interest especially to research physicist. To the theorist interested in understanding band offsets, the ability to predict such an offset, at least approximately, is one of the litmus tests of any lineup theory, and recent lineup theories pass this test with flying colors.

B. Theory

It should be self-evident from the above that the question as to the exact values of the band offsets at the various semiconductor pairs of interest is a central one, both theoretically and experimentally. I tried to contribute to both.

At the end of the '60s, the only rule for estimating band offsets theoretically was the *electron affinity rule* (Anderson, 1960), according to which the conduction band offset should be equal to the difference in electron affinity at the two free semiconductor surfaces. In a 1975 paper (Kroemer, 1975), I pointed out that this is an extraordinarily unsatisfactory rule. Even if good electron affinity data were available, the validity of the rule depended on hidden assumptions about the relations between the properties of the interface between two semiconductors, and those of the much more drastic vacuum-to-semiconductor interfaces, assumptions that almost certainly were invalid. Harrison aptly characterized the rule by saying that it "replaces one simple problem by two very difficult problems." (Harrison, 1977)

I called for a theory that would determine the band offsets from the *bulk* properties of the participating semiconductors, and I suggested it as a Ph. D. topic to Bill Frensley (now at the University of Texas in Dallas). One of the specific question I asked Bill to look into was whether broken-gap lineups might in fact occur. The resulting theory (Frensley and Kroemer, 1976; 1977), based on pseudopotentials, was the first to give a semi-quantitative derivation, from bulk properties, not only of band offsets that were already known, like

GaAs/AlAs; it also had a considerable predictive value. In particular, the theory predicted that the InAs/GaSb heterojunction either had a broken-gap lineup, or came very close to it.

The Frensley-Kroemer theory has since then been followed by the work of others based on different principles; see Harrison (1977) and Christensen (1988).

C. Band Offsets by C-V Profiling

Sometime in 1979, Jim Harris (then at the Rockwell Science Center, now at Stanford) showed me some capacitance-voltage (*C-V*) profiling data on an LPE-grown (Al,Ga)As/GaAs heterojunction. *C-V* profiling is a common technique to determine electron concentrations in semiconductors by measuring the capacitance of a reverse-biased Schottky barrier placed upon the surface of the semiconductor. By varying the bias, one can explore the depth distribution of the electrons over some distance. Near the hetero-interface, Harris' data showed a clear indication of an electron accumulation on the GaAs side, and an electron depletion on the (Al,Ga)As side, as one would expect from an appropriate band diagram. However, the apparent electron concentration was strongly smeared out by averaging over a Debye length. When I tried to understand the averaging process quantitatively, I realized that the dipole moment associated with the accumulation/depletion pair should be preserved during the averaging, and that its measurement should permit a determination of the conduction band offset (Kroemer *et al.*, 1980; Kroemer and Chien, 1981; Kroemer, 1985). The analysis yielded a band offset of approximately 66 % of the energy gap difference (Kroemer *et al.*, 1980), not far from today's generally accepted value of 62%.

The *C-V* technique has since then been used by many others and has provided some of the best data for band offsets for many heterojunction pairs.

IX. EPILOGUE

Throughout this paper, I have concentrated on my own work towards heterostructures, especially on the early parts of it, through 1963, which were dominated by bipolar device concepts. But today's heterostructure field would not be what it is without the subsequent contributions – technological or conceptual – by numerous others, especially on non-bipolar structures. It was only through this work of numerous others, on topics that went beyond my own contributions, that the significance of the latter eventually emerged. For this I owe all of them my thanks.

X. REFERENCES

Abstreiter, G., 1996, Physica Scripta T68, 68.

Alferov, Z. I., V. M. Andreev, D. Z. Garbuzov, Y. V. Zhilyaev, E. P. Morozov, E. L. Portnoi and V. G. Trofim, 1970, Fiz. Tekh. Poluprovodn. **4**, 1826. [Sov. Phys. - Semicond. 4, 1573-1575 (1971)].

Alferov, Z. I., 1996, Physica Scripta **T68**, 32.

Alferov, Z. I., 2001, this volume.
Anderson, R. L., 1960, IBM J. Res. Dev. **4**, 283.
Brown, E. R., J. R. Söderström, C. D. Parker, L. J. Mahoney, K. M. Molvar and T. C. McGill, 1991, Appl. Phys. Lett. **58**, 2291.
Caine, E. J., S. Subbanna, H. Kroemer, J. L. Merz and A. Y. Cho, 1984, Appl. Phys. Lett. **45**, 1123.
Casey, C. and M. Panish, 1978, *Heterostructure Lasers – Part A: Fundamental Principles* (Academic Press, New York). See Sec. 1.2.
Christensen, N. E., 1988, Phys. Rev. B **38**, 12687.
Delagebeaudeuf, D., P. Delescluse, P. Etienne, M. Laviron, J. Chaplart and N. T. Linh, 1980, Electron. Lett. **16**, 667.
Diedrich, H. and K. Jötten, 1961, Procs. of *Colloque international sur les dispositifs à semiconducteurs*, Paris (Editions Chiron, Paris) p. 330.
Dingle, R., H. L. Störmer, A. C. Gossard and W. Wiegmann, 1978, Appl. Phys. Lett. **33**, 665.
Esaki, L. and R. Tsu, 1970, IBM J. Res. Dev. **14**, 61.
Frensley, W. R. and H. Kroemer, 1976, J. Vac. Sci. Technol. **13**, 810.
Frensley, W. R. and H. Kroemer, 1977, Phys. Rev. B **16**, 2642.
Harrison, W. A., 1977, J. Vac. Sci. Technol. **14**, 1016.
Harrison, W. A., E. A. Kraut, J. R. Waldrop and R. W. Grant, 1978, Phys. Rev. B **18**, 4402.
Hayashi, I., M. B. Panish, P. W. Foyt and S. Sumski, 1970, Appl. Phys. Lett. **17**, 109.
König, U., 1996, Physica Scripta **T68**, 90.
Kroemer, H., 1957a, RCA Review **18**, 332. (Re-printed from the Proceedings of the Symposium *"The Role of Solid State Phenomena in Electric Circuits,"* Polytechnic Institute of Brooklyn, April 1957, p. 143)
Kroemer, H., 1957b, Proc. IRE **45**, 1535.
Kroemer, H., 1957c, unpublished.
Kroemer, H., 1963, Proc. IEEE **51**, 1782.
Kroemer, H., 1967, US patent 3,309,553 (filed Aug. 16, 1963).
Kroemer, H., 1975, Crit. Revs. Solid State Sci. **5**, 555.
Kroemer, H., W.-Y. Chien, J. S. Harris and D. D. Edwall, 1980, Appl. Phys. Lett. **36**, 295.
Kroemer, H., 1981, Jpn. J. Appl. Phys. Supplem. **20-1**, 9.
Kroemer, H. and W.-Y. Chien, 1981, Solid-State Electron. **24**, 655.
Kroemer, H., 1982, Proc. IEEE **70**, 13.
Kroemer, H., 1983, J. Vac. Sci. Technol. B **1**, 126.
Kroemer, H. and G. Griffiths, 1983, IEEE Elect. Dev. Lett. **4**, 20.
Kroemer, H., 1985, Appl. Phys. Lett. **46**, 494.
Kroemer, H., 1987, J. Cryst. Growth **81**, 193.
Kroemer, H., C. Nguyen, E. L. Hu, E. L. Yuh, M. Thomas and K. C. Wong, 1994, Physica B **203**, 298.
Kroemer, H., 1995, Procs. of *NATO Adv. Res. Wkshp. on Future Trends in Microelectronics*, Ile de Bendor, France, edited by S. Luryi et al., NATO ASI Series E **323** (Kluwer, Dordrecht) p. 1.
Kroemer, H., 1996, Physica Scripta **T68**, 10.
Krömer, H., 1953, Naturwissensch. **40**, 578.
Krömer, H., 1954, Archiv d. Elekt. Übertragung **8**, 499.
Laughlin, R. B., 1999, Revs. Mod. Phys. **71**, 863.
Mermin, D., 1999 (Aug.), Physics Today **52** (8), 11.
Mimura, T., S. Hiyamizu, T. Fujii and K. Nanbu, 1980, Jpn. J. Appl. Phys. **19**, L225.
Nakagawa, A., H. Kroemer and J. H. English, 1989, Appl. Phys. Lett. **54**, 1893.
Sakaki, H., L. L. Chang, R. Ludeke, C. A. Chang, G. A. Sai-Halasz and L. Esaki, 1977, Appl. Phys. Lett. **31**, 211.
Shockley, W., 1951, US patent 2,569,347 (filed 26 June 1948).
Stormer, H. L., 1999, Revs. Mod. Phys. **71**, 875.
Tsui, D. C., H. L. Störmer and A. C. Gossard, 1982, Phys. Rev. Lett. **48**, 1559.
Wright, S. L., M. Inada and H. Kroemer, 1982, J. Vac. Sci. Technol. **21**, 534.
Wright, S. L., H. Kroemer and M. Inada, 1984, J. Appl. Phys. **55**, 2916.