Characterisation and Performance of Optical Lithography Systems.

Thesis submitted by
Graeme Dunlop Maxwell
for the Degree of
Doctor of Philosophy.

Edinburgh Microfabrication Facility,
Department of Electrical Engineering,
University of Edinburgh.

November 1987.

Declaration

I declare that the work contained in this thesis is, except where otherwise indicated, entirely my own, and that the thesis has been composed entirely by myself.

Graeme Dunlop Maxwell,
November 1987.

Acknowledgements

Throughout the period in which the work which is described in this thesis was performed, I have benefited greatly from the advice and support of many people, and it is my pleasure to acknowledge their help here. I would like to thank my supervisor, Professor John Robertson, for his many helpful discussions, for guiding the direction in which the project should go, and for his detailed scrutiny of the material contained in this thesis. I would like to thank Mr. Tom Stevenson, for his patient help in passing on much of his knowledge of the lithographic process to me, and also for many technical discussions, from which a number of fruitful ideas came to light. I would also like to thank him for his help in proof-reading the thesis. In addition, software support on the E.M.F. VAX computer was greatly appreciated, from Dr. Tony Walton, Mr. Martin Fallon, and Dr. Jasbinder Singh. I would like to thank Mr. Alan Gundlach, Mr. John Fraser, Dr. Bob Holwill, and all the E.M.F. technicians for their support in helping me to make full use of the resources available in the Facility, and for teaching me how to use all types of semiconductor processing equipment. I would also like to thank Mrs. Liz Paterson for typing up a number of reports and other smaller documents.

Outside of the E.M.F., I would like to thank my industrial sponsors, Eaton Semiconductor Equipment Operations, for financial support, and in particular my industrial supervisor, Mr. Trevor Lewin, for his help with technical problems regarding the Optimetrix wafer stepper. For their support during visits to Eaton in California, both technical and social, I would like to thank Dr. Gordon MacBeth, Mr. Jim Dey, Mr. Gerry Alonso, and Mr. Don Harris. I wish also to thank Mr. René Vervoordeldonk and Dr. Rick Harwig, Philips NatLab in Eindhoven, The Netherlands, for their help in setting up a two month working visit to the NatLab, and particularly to René, for his support and technical advice both during this visit, and later in the course my project. Finally, I would like to thank the Science and Engineering Research Council for their financial support during the course of my studies.

Abstract

A detailed study of the performance of micro-lithographic systems used in the production of integrated circuits (IC's) is presented, with particular emphasis on the performance of optical wafer stepper alignment systems. An overview of the process used to manufacture IC's is presented, followed by a review of the likely future requirements for lithographic systems as IC processing enters the Ultra Large Scale Integration (ULSI) age. This review includes a literature survey, and covers both exposure tools and resist technology. The results are presented of a study into factors affecting the ability of light field alignment systems on wafer steppers to accurately perform the registration of one layer in the IC fabrication process to another. It is shown that the performance of such systems can be significantly improved by the application of appropriate signal processing, and also by the use of appropriate alignment targets. Particular emphasis is placed on the alignment system of the Optimetrix wafer stepper, which the University of Edinburgh owns and operates. Computer simulation of the lithographic process is also reviewed, and this tool is used to characterise the degradation of optical lithography processing, as the system resolution limit is reached.

Contents

Chapter 1

Review of Technology.

With the semiconductor industry about to enter the era of ULSI (Ultra Large Scale Integration), the semiconductor process engineer must prepare to meet the demands of 1μm design rules and below. This is a particularly arduous task for the photo-lithography engineer, whose function is to define in a photo-sensitive material (photo-resist) the pattern to be etched into an underlying material. To define patterns with ever decreasing dimensions requires the continuous improvement of both the photo-lithographic process and the equipment used by the lithographer. This continuous improvement must inevitably cost the industry a great deal in terms of money, and hence must have some compelling force behind it.

To answer the question of why it is desirable, or even necessary, to continue to reduce device dimensions in the manufacture of integrated circuits, many different factors must be addressed.

We begin by using the functional throughput rate (FTR)[†] as a figure of merit for a given chip. By definition, there are two methods by which we may increase the FTR of the circuit :

1. Increase the maximum clock rate of a gate, by either optimising the process for speed (eg. use low resistivity interconnect), or using innovative design methods (eg. the use of parallel processing).

2. Increase the number of gates/chip, by either increasing the chip size or decreasing the gate size (ie. reducing device dimensions). Figure 1.1(a) shows a graph of the growth of IC circuit complexity against time [1].

[†]The functional throughput rate (FTR) is defined as the product of equivalent gates per chip times the maximum clock rate of a chip.

The data illustrates a phenomenon which many people call 'Moore's Law', namely an exponential increase in the number of components/chip, with complexity doubling initially every year, and slowing down later to a doubling every two years. The reduction in rate after 1980 is explained by Moore [1] by the elimination at that time of nonfunctional chip areas.

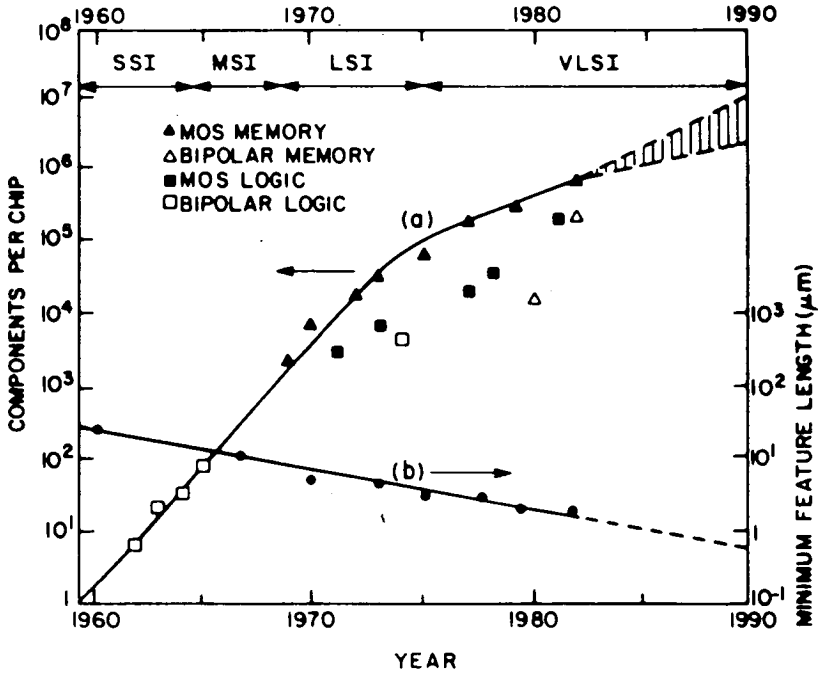

Figure 1.1: Exponential increase in circuit complexity (a) and decrease in device dimensions (b), with time. Taken from reference [2].

By far the strongest contributing factor towards this increase in circuit complexity has been the reduction of feature sizes, which has also been exponential in nature [2] (Figure 1.1(b)). For example, in the evolution of the dynamic RAM from 1K bits to 64K bits, it is estimated that circuit innovation has accounted for a factor of 3 increase in circuit complexity (due to the change over from the three transistor memory cell to the one transistor cell), device scaling for a factor of just under 10 increase in complexity, and increase in chip size for a factor of 1.65 increase in complexity [3].

Consideration of electrical characteristics of scaled devices also reveals advantages in going to smaller dimensions. Reducing dimensions (channel width, length, and oxide thickness) by a factor of $1/k$ ($k > 1$), while reducing the supply

voltages by the same factor, increases the number of gates/chip by a factor of k^2 and also decreases the delay time by a factor of k, thus increasing the FTR by a factor of k^3. Scaling has to be used with caution, however, since many of the physical parameters do not scale at all, or scale in an undesirable manner (eg. contact resistance of interconnect lines scales up by a factor of k).

For many IC process generations up to the VLSI level, the limiting factor in being able to shrink device dimensions has been the resolution of the lithographic process. This remains true at the ULSI level and beyond, although it becomes only one of a number of limiting factors, each of which has to be optimised in relation to the others. Indeed, this could be taken as a definition of a ULSI process, in which the individual process steps can no longer be regarded as completely isolated from one another, but have to be considered as a coherent whole. IC manufacturing can be regarded as a cyclic process in which the raw material (silicon, GaAs etc...) undergoes a variety of lithography, etch, and deposition stages, on its way to becoming a final product. This process is illustrated in Figure 1.2. The lithographic stage is repeated a number of times (from ~8 for a simple NMOS process to ~25 for an advanced CMOS process) with many stages requiring both high resolution and accurate alignment to the layers below.

If optical lithography is to be pushed to its limits, a full understanding of all it's aspects is required. It has been the intention both throughout the period that this work was performed, and also in the writing of this thesis, that a systematic and complete analysis of the main problems facing optical lithography should be made, in order to come closer to this understanding. Two particular aspects of imaging technology are dealt with in detail:

1. Increasing lithographic resolution.

2. Increasing layer-to-layer registration accuracy.

To begin the analysis, this chapter deals with the basic ideas of semiconductor processing, as well as introducing optical micro-lithography. Chapter 2 discusses in greater length the three main problems facing lithography (reduction of feature size, control of feature size, and alignment accuracy), and goes on to give an account, based on literature published in this area, of the solutions which have been proposed to deal with the control and reduction of feature size. These solutions are mainly resist processing ones; the only machine improvements which can be made in this area are lens modifications (either increasing the numerical aperture, or reducing the exposure wavelength). Chapter 3 then goes on to discuss the problem of alignment accuracy with a detailed account

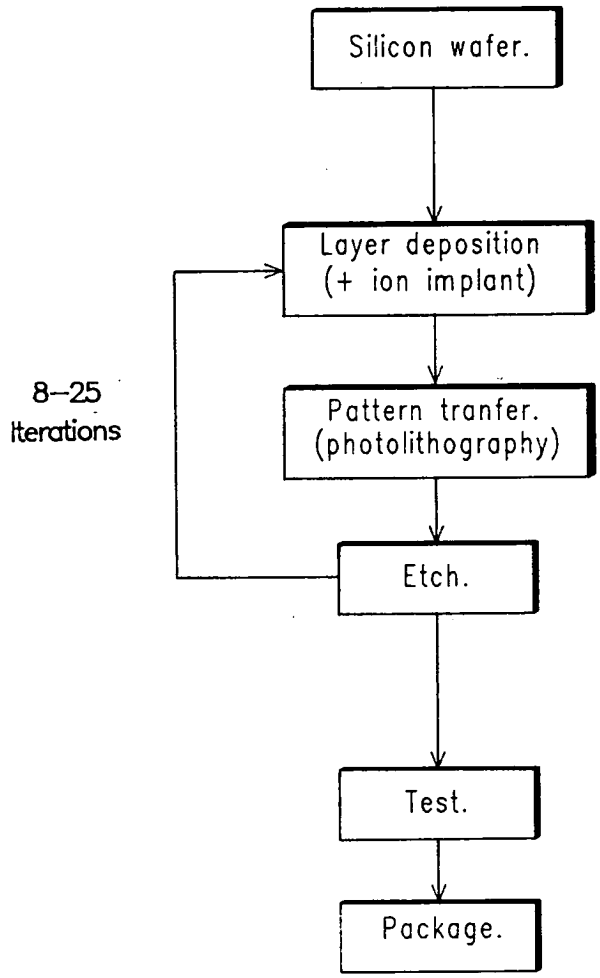

Figure 1.2: Generalised IC fabrication process.

of the alignment systems used by all wafer steppers currently on the market, as well as some lesser known systems which have been proposed in the literature.

Chapter 4 discusses at length the alignment system on the Optimetrix machine, and proposes several methods by which the alignment may be improved, in terms of speed and reliability. A software approach is adopted, using a program which was developed off-line, specially for the purpose of studying this system. Chapter 5 uses simulation as a tool, in order to study the effect of various optical parameters (layer thickness, sidewall profile, etc...) on the alignment signal on such a machine (although this work was performed with the Optimetrix machine in mind, the results apply equally well to any machine with the same type of alignment system). Chapter 6 goes on to study the effect of alternative marker structures on the alignment system, with a view to improving the alignment accuracy during a real process run.

Chapter 7 presents an overview of simulation in photo-lithography, including an in-depth study of the simulation program SAMPLE, as well as some other less well known programs. Chapter 8 goes on to use some of these programs to study the characteristics of aerial image profiles, and some useful deductions are made here from the point of view of reticle biasing. Finally in Chapter 9, a summary is made of the work which has been presented, along with ideas for future work.

A small amount of duplication of material has been used throughout the thesis, in order for certain chapters to exist on a stand-alone basis.

1.1 The Lithographic Process.

Lithography, as defined in the integrated circuit industry, is the transfer of features from a mask consisting of opaque and transparent areas, onto a semiconductor wafer which contains uncompleted devices. The opaque and transparent areas on the mask are reproduced on the wafer to define the various areas of the completed circuit (metal to connect devices, or oxide to isolate them, etc...). Figure 1.3 shows an example of the lithographic process, for both positive and negative working resists.

Commercially available negative resists consist of a combination of chemically active polymers and a photo-sensitive agent. On exposure, the photo-sensitive agent reacts with the polymer, causing cross-linking and hence a reduction in the resist solubility in organic developer. Exposed areas of the resist tend to swell, however, due to absorption of developer. Swelling can lead to bridging of gaps

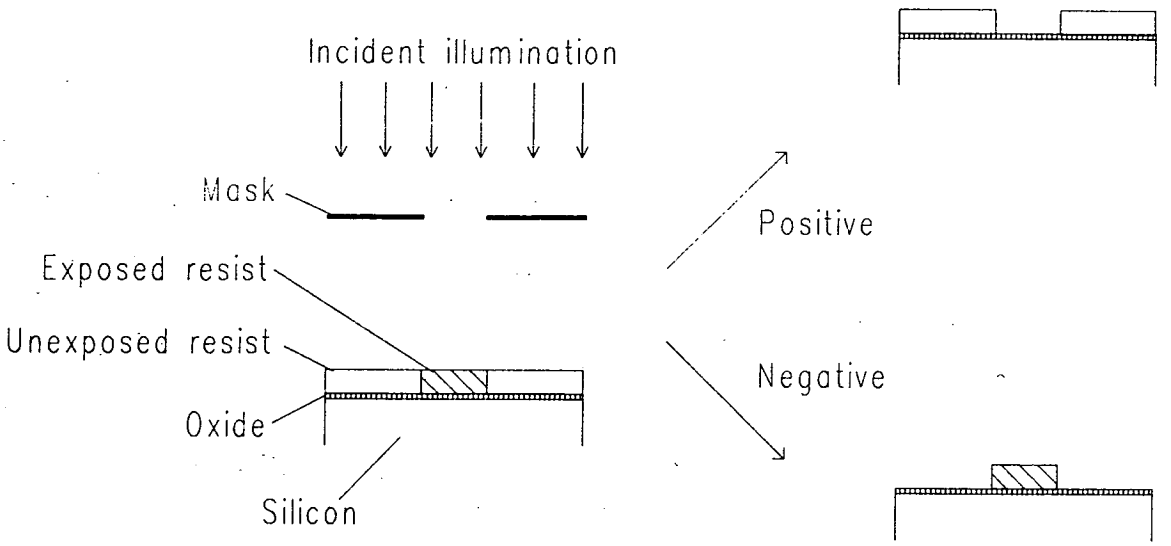

Figure 1.3: Generalised lithographic process.

in the resist, limiting the resolution of such systems to 2–3μm. An example of the chemistry of the negative resist process is illustrated in Figure 1.4.

Conventional positive resists, which have now become the industry standard, consist of a phenol-formaldehyde (NOVOLAC) type resin mixed with a diazo-quinone photo-active compound (PAC), with the latter acting as an inhibitor to prevent the dissolution of the resin in alkaline developer. When the PAC is exposed to radiation in the 350-500nm range, it is converted to a carboxylic acid, which acts as a dissolution enhancer when the resist is immersed in developer. The process is illustrated in Figure 1.5.

Good resist materials should possess high sensitivity to exposure wavelengths and resolution capabilities in the sub-micron range. In addition to this, coating properties must be good, exhibiting good adhesion and film thickness uniformity, as well as a low density of pinhole defects. Thermal stability should also be good, with no resist flow up to temperatures of 150°C, and the material should be able to resist erosion to both wet and dry etching.

Resists, both positive and negative, do not exhibit infinite contrast. In general, both types are conveniently characterised by plotting a graph of percentage of resist remaining after exposure against the log of the exposure energy [4]. Such a graph is shown in Figure 1.6 for positive resist. The resist contrast, or γ, is

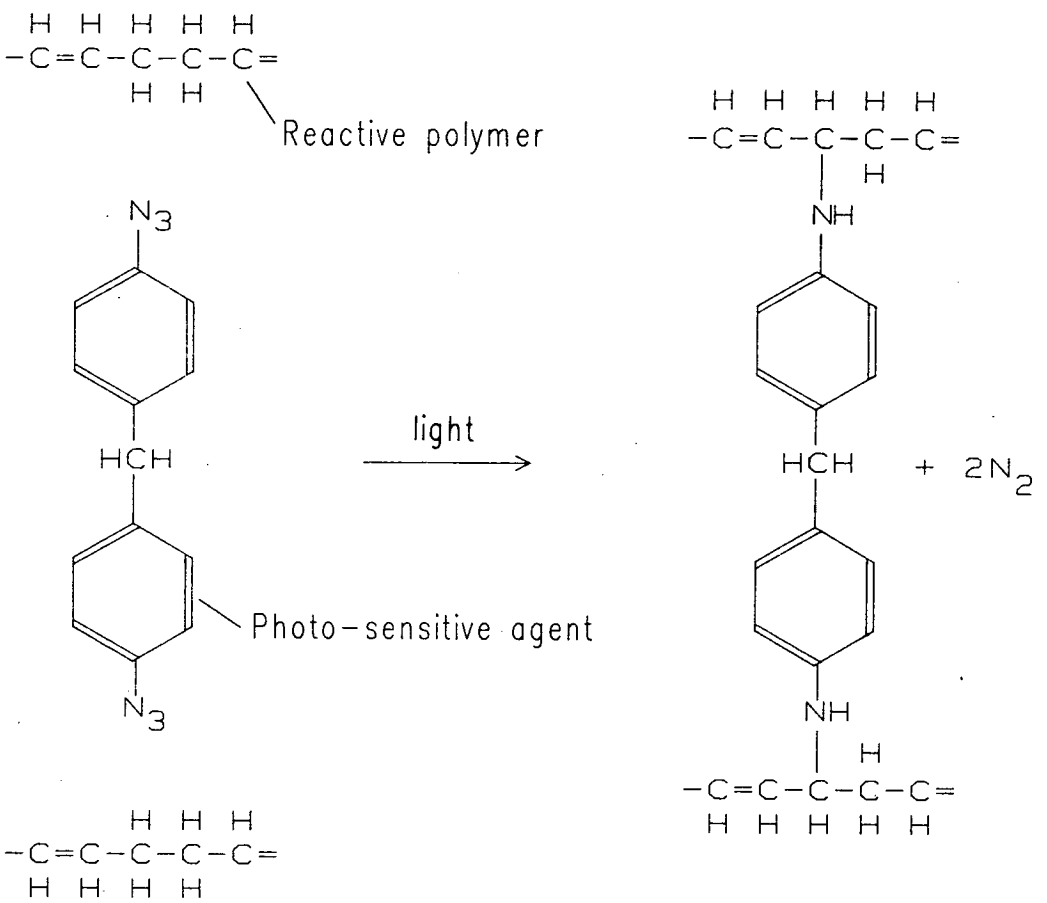

Figure 1.4: Example of negative resist chemistry.

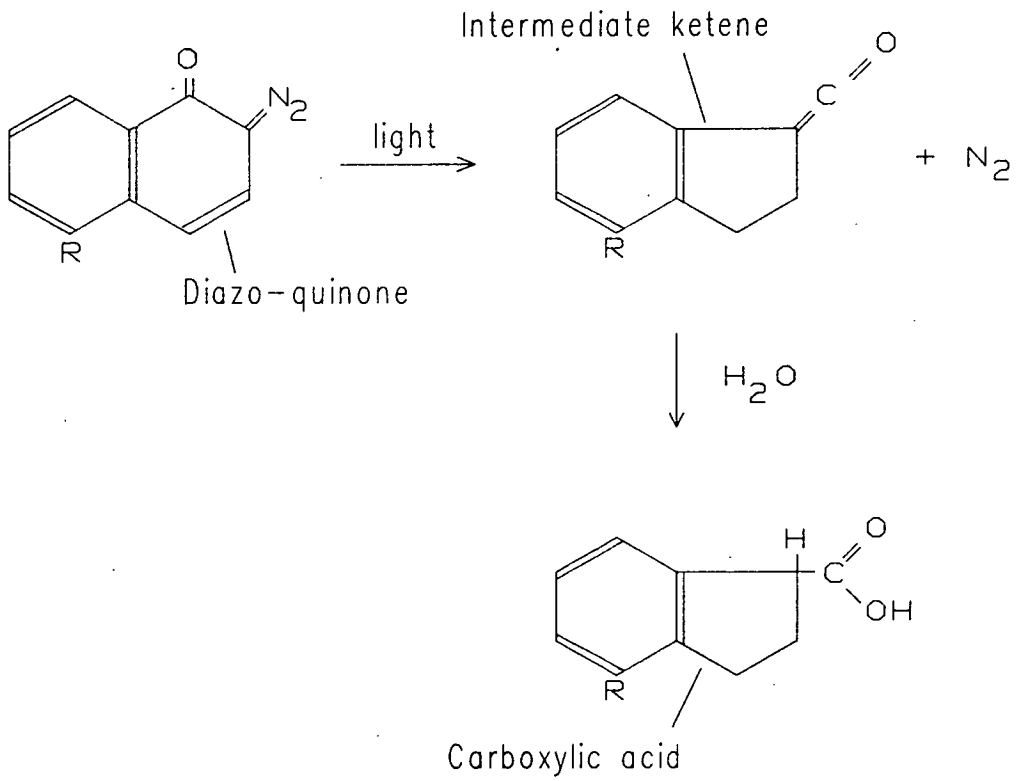

Figure 1.5: Example of positive resist chemistry.

defined by the following equation :

$$\gamma = \left[\log_{10}\left(\frac{E_o}{E_i}\right)\right]^{-1} \tag{1.1}$$

where E_o is the exposure energy required to completely remove the resist, and E_i is calculated by projecting back the tangent line which the curve makes with the energy axis to the zero resist removal line (see figure). E_i is defined to be the exposure energy at which significant resist removal begins.

By definition therefore, γ is simply the slope of the line at the intercept of the curve with the energy axis. γ is used as a figure of merit for a resist/developer system, with high values resulting in resist profiles with steep sidewalls. This can be explained with the help of Figure 1.7 which shows a sample aerial image (for a definition of the aerial image, see Section 1.2.4) which, due to the finite aperture of the imaging lens does not exhibit perfectly sharp edges. The position of I_o ($I_o = E_o/t$, where t is the exposure time) is marked, along with I_i ($I_i = E_i/t$) for both high and low γ systems. The positions P_{low} and P_{high} indicate the positions at which significant resist removal begins, for low and high contrast systems respectively, demonstrating that for the low contrast system resist removal begins further away from the desired edge of the resist line, thus generating a rounder profile.

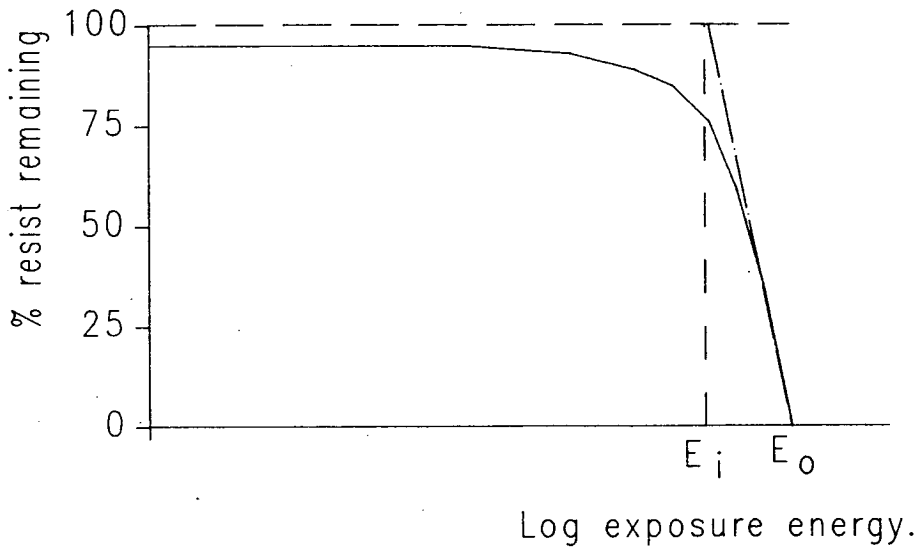

Figure 1.6: Characteristic resist exposure curve for positive resist.

Resist contrast and image contrast are to some extent interchangeable, and an increase in either of these factors can be used to extend the useful resolution of a lithographic process [5].

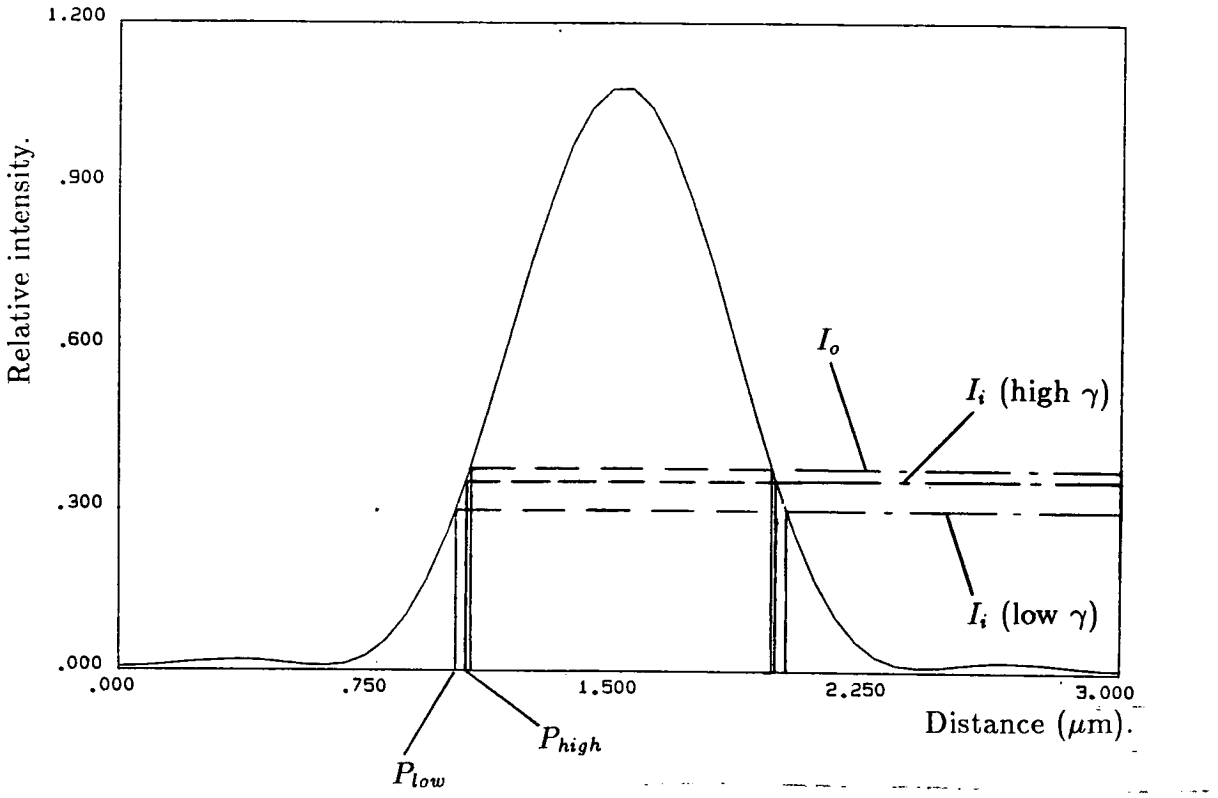

Figure 1.7: Aerial image of 1.0μm line and 2.0μm space, from the SAMPLE simulation program (NA=0.32, λ=0.436μm, coherence=0.5), showing positions of I_o, I_i, P_{low}, and P_{high}.

1.2 The Evolution of Lithographic Equipment.

Faced with the progression towards ever higher levels of integration, the semiconductor equipment industry has been forced to evolve its own products to cope with the demand for smaller feature sizes and tighter overlay requirements. Throughout the lifetime of micro-lithography in the electronics industry, there have been four predominant methods used for the reproduction of mask features on a wafer, each of which is explained below.

1.2.1 Contact Printing.

In contact printing, the mask is placed in contact with the wafer, and collimated light is shone through the mask onto the photo-resist (Figure 1.8). Resolution using this method is very good; it is limited simply by the resolution in the mask-making step and by diffraction within the resist film itself. The main problem with contact printing is one of defect density, which tends to be high because of the transfer of debris from wafer to mask and vice-versa. This transfer also severely limits mask lifetime, making this method uneconomical in the long term.

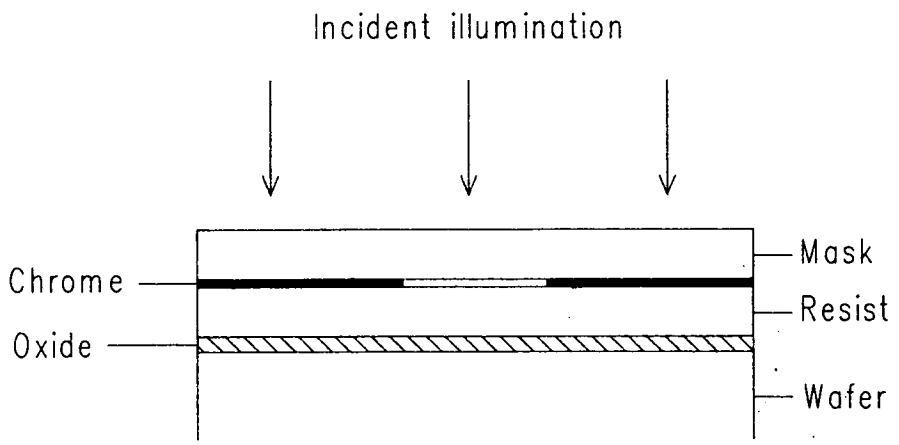

Figure 1.8: Contact print process.

Contact printing is economical in its use of exposure illumination, however, in that the whole spectrum of the mercury arc lamp can be used, without the need for filtering.

1.2.2 Proximity Printing.

Proximity printing was originally introduced as an evolutionary (as opposed to revolutionary) step up from contact printing. With proximity printing the wafer is held at a small distance from the mask, and collimated light is again used to illuminate the combination.

The separation of mask and wafer (generally around 15–20μm) serves to eliminate the transfer of debris from one to the other, but has negative consequences as well. Near-field (Fresnel) diffraction occurs at feature edges, resulting in an illuminated area on the wafer which is given by $W' = W(1+F)/F$ where $F = W^2/\lambda S$ (see Figure 1.9). The fact that the printed feature size depends upon the mask/wafer gap (S) implies a loss of dimensional control if either mask or wafer bow should occur. In addition to this, imperfect collimation of exposure illumination causes penumbral effects which also affect printed feature size.

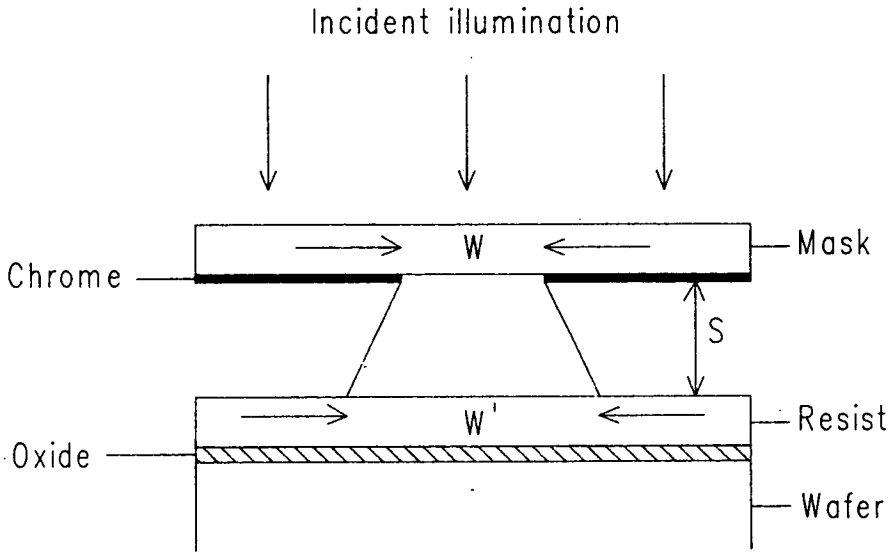

Figure 1.9: Proximity print process.

1.2.3 Whole Wafer Projection.

Figure 1.10 shows an example of a generalised projection system, in which an image of the mask is formed at the wafer by means of either reflective or refractive optics (although the figure shows a refractive system, the principles are the same for a reflective system). The optics consist of a condenser section, which collimates the source illumination and illuminates the mask, and a projection section, which forms the image of the mask (only one ray is traced through the

projection section in the figure for the purposes of simplification). The numerical aperture (NA) of the two sections, which is of importance in determining the working resolution of the system, is also defined in the figure.

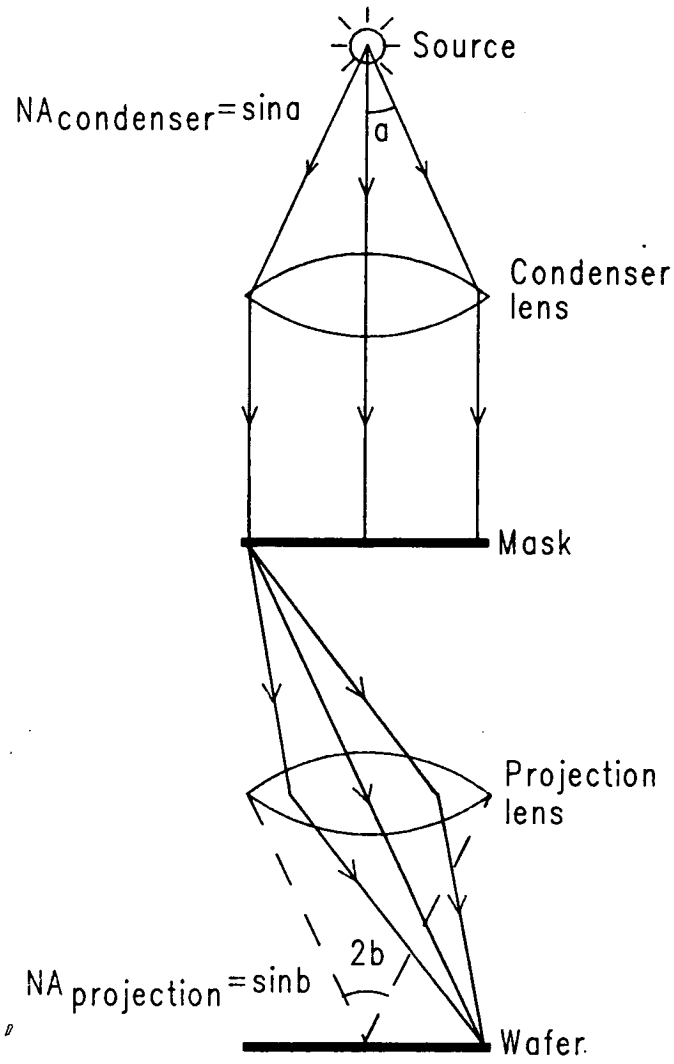

Figure 1.10: Generalised projection system.

With whole wafer projection, an image of the mask is formed at the wafer by means of a reflective optical system, and the resist is exposed by scanning a narrow slit across the mask and wafer simultaneously (Figure 1.11). Because the optical system is reflective, the focal length is independent of wavelength, and hence the full lamp spectrum may still be used, with the resolution being determined by the exposure wavelength and the numerical aperture of the optics,

according to the following equation :

$$\mathrm{SPF} = k\frac{\lambda}{\mathrm{NA}} \tag{1.2}$$

where SPF is the smallest printable feature, λ is the exposure wavelength, and k is a constant which depends on both the design of the system optics and the photo-resist process being used. For polychromatic exposure, a weighted sum of the SPF's for each exposing wavelength must be used.

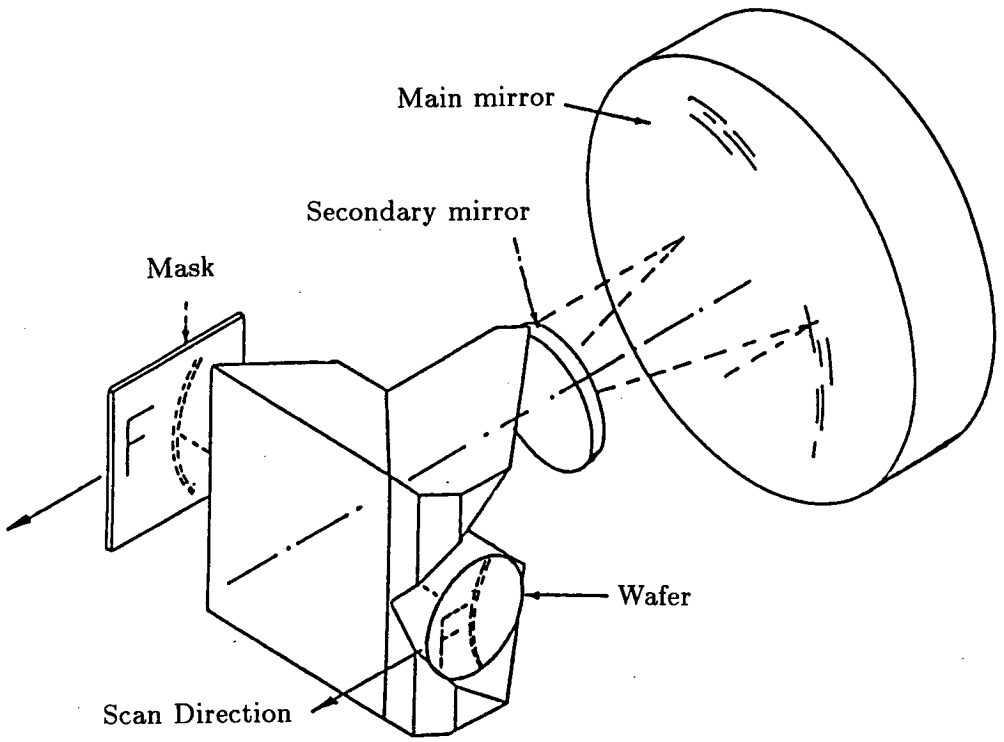

Figure 1.11: Whole wafer projection system.

The principal advantage of whole wafer projection over proximity printing is the increased dimensional control gained by the elimination of near field diffraction (at the expense of the introduction of far-field, or Fraunhofer, diffraction) and penumbral effects.

1.2.4 Step-and-Repeat Imaging.

A step-and-repeat system is a type of projection system in which an image of a reticle is formed on one small part of the wafer at a time. This image normally

consists of between one and four die, depending upon the size of the die and the size of the reticle. One area on the wafer is exposed, and the wafer is moved along to the next site. This process is repeated until the whole wafer has been exposed. Common reduction ratios for step-and-repeat systems (object size:image size) are 1:1, 5:1, and 10:1. In general the lower the reduction ratio the higher the throughput (number of wafers printed/hour), since more die are present per unit area on the mask. Lower reduction ratios also have the effect of increasing sensitivity to defects, however, since linewidths on the mask are smaller, hence reducing the size threshold of a fatal defect (a defect which causes a short or a break in the circuit). A fatal defect on the mask in step-and-repeat imaging is, of course, printed at every site on the wafer: thus great care has to be taken over mask cleanliness.

The advantages of this step-and-repeat imaging are two-fold. Firstly, due to the smaller image size than that used in whole wafer projection, it is possible to design lenses with a higher numerical aperture (at present 0.3–0.4NA as opposed to ~0.15NA in the whole wafer case), and thus to increase the resolution of the system. Secondly it is possible, using step-and-repeat exposure, to align at each die site on the wafer, thus enabling correction of runout (inelastic expansion or contraction) of the wafer which has occurred during high temperature processing. Alignment accuracy plays a vital role in the fabrication of integrated circuits, and the ability of step-and-repeat systems (wafer steppers) to perform alignment at each site allows these machines to take full advantage of their inherently high resolution.

Most steppers, however, suffer from one particular drawback in that they tend to be refractive optical systems (all except one, the Ultratech, use refractive optics). This results in the focal length being wavelength dependent, and hence to a requirement for filtering of the exposure illumination, to reduce the effect of chromatic aberration.

Projection systems are in general characterised by an optical transfer function, or OTF [6]. Figure 1.12 depicts the imaging of a mask with an equal line space pattern by a lens with finite aperture. The intensity distribution at the mask is a square wave with 100% contrast (contrast $= (I_{max} - I_{min})/(I_{max} + I_{min})$), which can be regarded as a sum of Fourier components (spatial frequency components). The fundamental spatial frequency is, in this case, equal to $1/(l + s)$, or $1/p$, where p is the periodicity. The intensity distribution at the mask[†] $O(x)$,

[†]Sometimes referred to as the input, or object signal.

is given by the equation :

$$O(x) = \frac{a_o}{2} + \sum_{n=1}^{n=\infty} a_n \cos(2\pi n x/p) \tag{1.3}$$

where the a_n are given by

$$a_n = \frac{2}{n\pi} \sin\left(\frac{n\pi}{2}\right) \quad \text{for } n \geq 0 \tag{1.4}$$

The intensity distribution in the image plane is then given by a linear combination of the Fourier components present in the object plane :

$$I(x) = \frac{b_o}{2} + \sum_{n=1}^{n=\infty} b_n \cos(2\pi n x/p) \tag{1.5}$$

The optical transfer function for a diffraction limited lens (a lens in which image degradation due to primary aberrations is negligible in comparison with image degradation due to diffraction [7]) is defined by the following equation :

$$\text{OTF}(\nu) = 1 - \frac{2}{\pi}\cos^{-1}\left[1 - \left(\frac{\nu}{\nu_o}\right)^2\right]^{\frac{1}{2}} - \frac{1}{\pi}\frac{\nu}{\nu_o}\left[1 - \left(\frac{\nu}{\nu_o}\right)^2\right]^{\frac{1}{2}} \tag{1.6}$$

where ν is the spatial frequency, and ν_o, the cut-off frequency, is given by:

$$\nu_o = 2\frac{\text{NA}}{\lambda} \tag{1.7}$$

The set of coefficients b_n in Equation 1.5 are constrained to lie on the curve defined by Equation 1.6 (ie. b_n are the values of the OTF at the discrete frequencies present in the object spectrum). The OTF is plotted in Figure 1.13 for a diffraction limited lens with NA = 0.32 and λ = 436nm, along with the corresponding modulation transfer function (MTF - see below). It can be deduced from the shape of the curve that the lens is acting as a low pass linear filter for the spatial frequencies present in the input signal, with a cut-off frequency at $\nu = \nu_o$.

The optical transfer function is, strictly speaking, defined only for incoherent illumination (the degree of coherence in a projection system is defined as $S = \text{NA}_{\text{condenser}}/\text{NA}_{\text{projection}}$, see Figure 1.10). Incoherence corresponds to $S = \infty$ (projection lens filled), with complete coherence corresponding to $S = 0$. For coherent illumination the modulation transfer function (MTF) is often used, which describes the transfer of amplitude (rather than intensity) information from object to image. The MTF is equal to 1 for spatial frequencies below $\nu = \text{NA}/\lambda$, and 0 for frequencies above $\nu = \text{NA}/\lambda$ (see Figure 1.13).

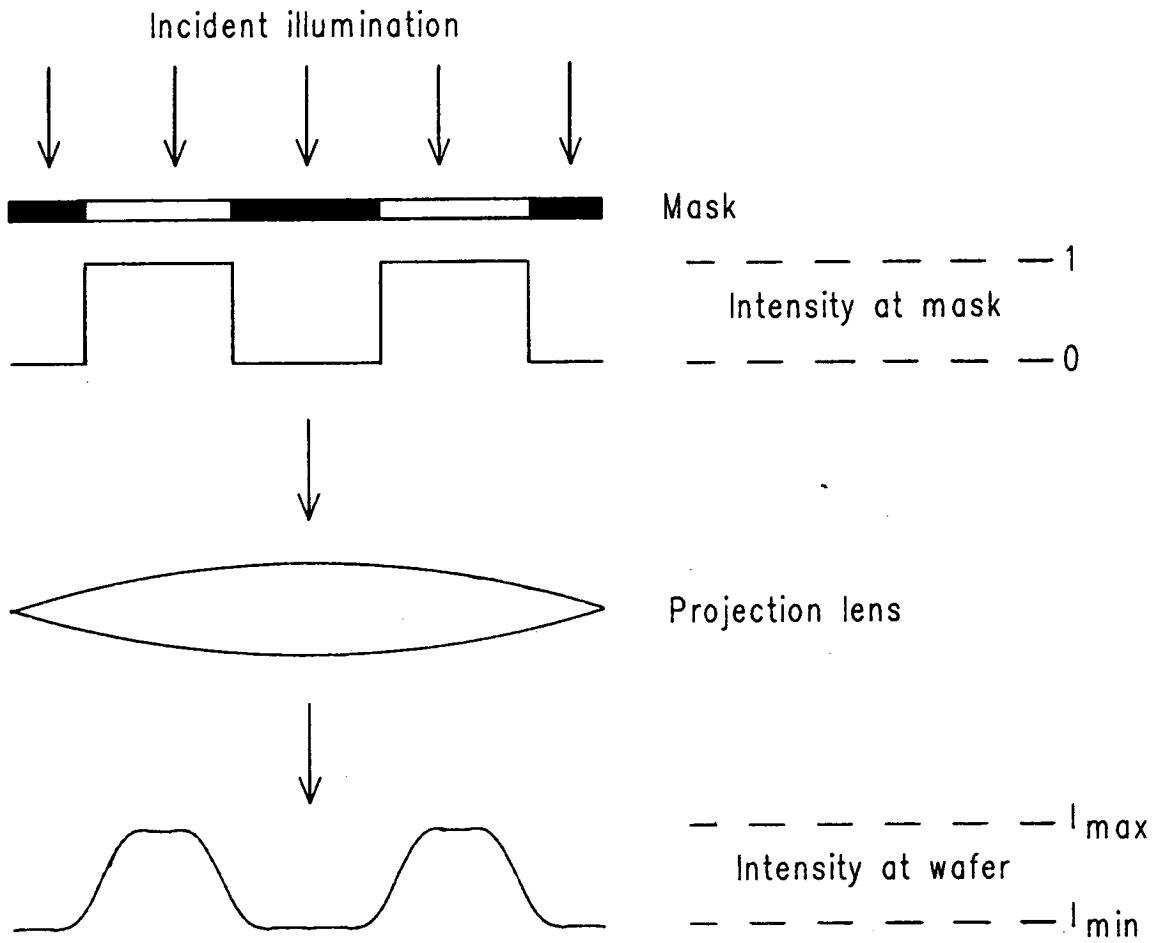

Figure 1.12: Imaging degradation due to finite aperture optics.

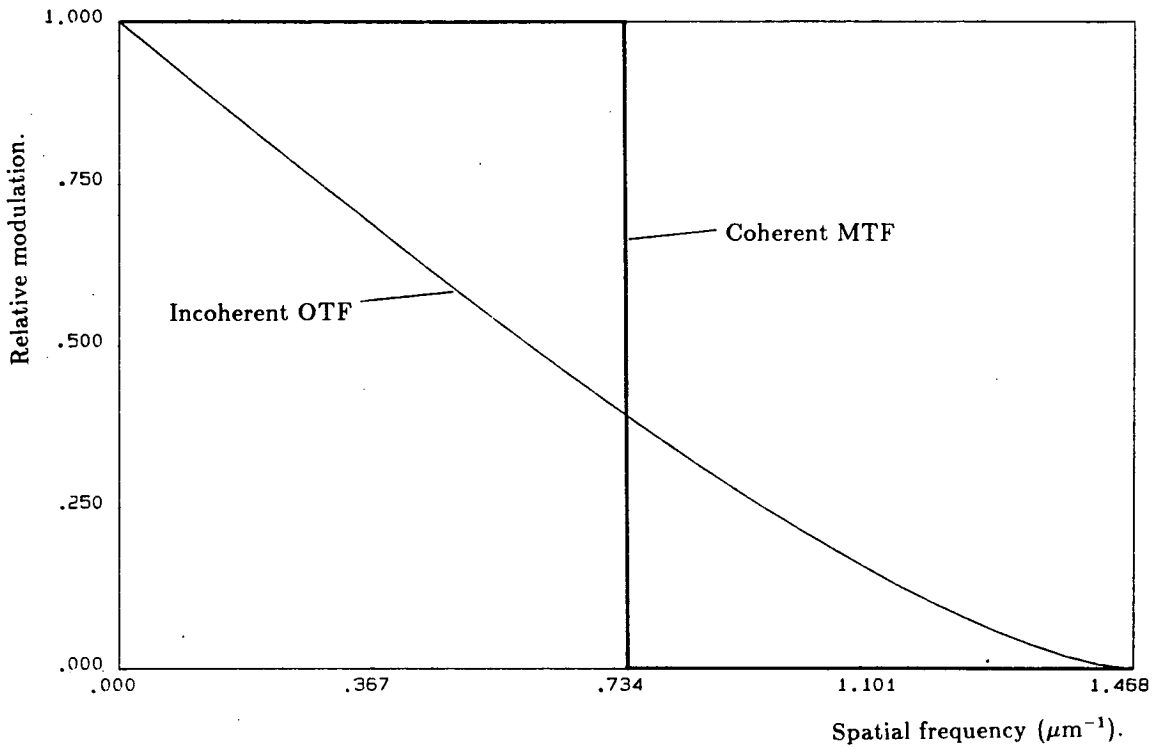

Figure 1.13: Incoherent optical transfer function (OTF) and coherent modulation transfer function (MTF), for a 0.32 NA lens operating at 436nm.

In both the coherent and incoherent regimes the imaging can be regarded as linear (with respect to intensity in the incoherent case, and amplitude in the coherent case). Between the two extremes the system is described as partially coherent, and the imaging is non-linear. In the partially coherent case there does not exist a transfer function which can be used to calculate the image spectrum directly from the object spectrum. Nevertheless people often plot image contrast as a function of spatial frequency for partially coherent systems [6] [8], and use the curves as figures of merit for a particular lens.

Most projection systems used in micro-lithography operate in the partially coherent regime (S =0.5–0.7), as a trade off between factors such as image contrast and depth of focus.

1.3 NMOS Technology.

Although the field effect principle of transistor operation was proposed as long ago as the early 1930's, the first functional MOSFET's were not demonstrated until 1960 [9]. These early MOS transistors were p-type, taking advantage of the fact that the surface of silicon is naturally n-type. Advances in ion implantation technology, however, allowed wafers to be doped very lightly p-type, paving the way for fabrication of n-type transistors which utilised the higher mobility of electrons over holes (by a factor of $\sim$3), which in turn led to higher circuit FTR. NMOS integrated circuits have continued to take a major share of the world market since the early 1970's. CMOS, bipolar, and GaAs circuits make use of identical lithographic systems, so for the purposes of this project, an NMOS process may be regarded as typical.

1.3.1 NMOS Fabrication Sequence.

Figure 1.14 shows the circuit diagram of a simple inverter, while Figure 1.15 shows a cross-sectional view of the same inverter which has been implemented in NMOS technology. It consists of an enhancement mode device (off when V_g = 0V) and a depletion device (on when V_g = 0V). The output V_{out} is low if and only if V_g is high and vice versa. The cross-sectional view of the inverter is best explained with reference to the individual process steps.

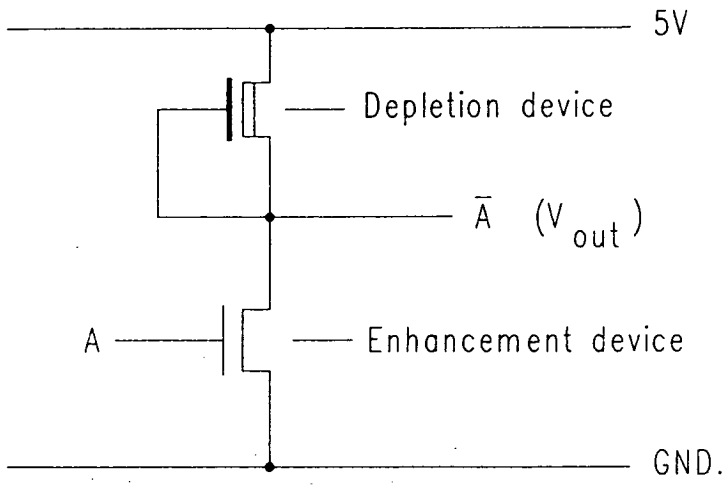

Figure 1.14: Circuit diagram of a simple NMOS inverter.

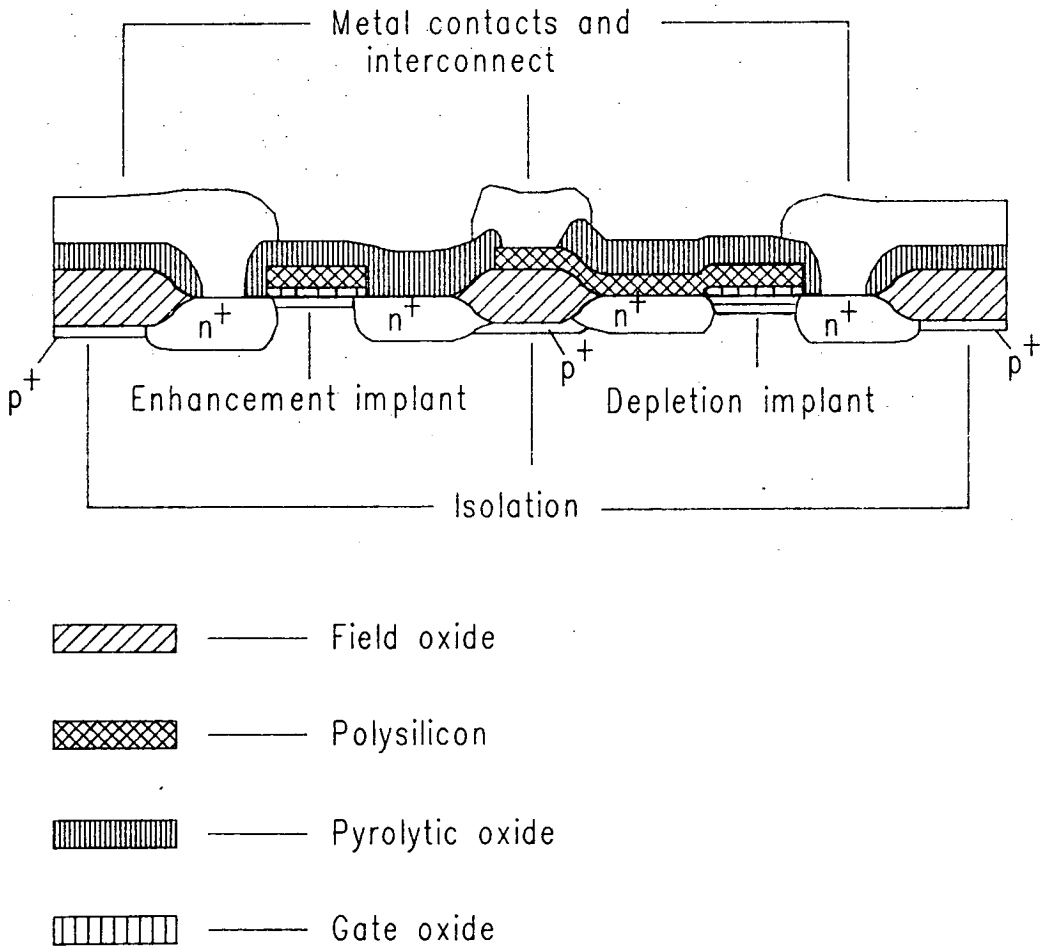

Figure 1.15: Cross-sectional view of NMOS inverter.

1.3.2 Starting Materials.

NMOS fabrication uses p-type silicon as a starting product, generally with a <100> crystal orientation, since this minimises charge build up at the $Si - SiO_2$ interface.

The doped wafers have a diameter anywhere between 3in and 8in, and a resistivity between 14Ω-cm and 20Ω-cm. Substrate doping is low ($\sim 10^{15}$ atoms/cm^3), in order to minimise source/drain to substrate parasitic capacitance. Too little substrate doping, however, can lead to problems with depletion regions from source and drain 'punching through' to each other.

1.3.3 Isolation and Field Stop Implant.

Figure 1.15 illustrates the existence of a parasitic $n^+p^+n^+$ transistor between adjacent devices. If the threshold voltage of this parasitic transistor is low enough, it can switch on, causing cross-talk between adjacent devices. In practice cross-talk is eliminated by providing isolation between the transistors, which normally consists of thick thermal oxide (field oxide) and a p^+ channel stop implant.

Field oxidation is normally accomplished by the LOCOS (**LO**Cal **O**xidation of **S**ilicon) technique [10], in which an initial thermal oxide of ~250Å thickness is grown over the whole wafer. A layer of silicon nitride is then deposited by chemical vapour deposition, using silane and ammonia as the reactant gasses. The purpose of the initial thermal oxide is to act as a stress relief barrier between the silicon and the nitride (the oxide having a thermal expansion coefficient which lies between the coefficients for the other two). Photo-resist is then applied to the wafer (Figure 1.16(a)), and the resist is selectively exposed in those areas where field oxide is desired. After development (Figure 1.16(b)), the nitride is plasma etched to remove it from the field regions, and the channel stop implant is performed (Figure 1.16(c)). This is normally a high energy (~100keV) implant, with a dosage between 10^{12} and 10^{13} atoms/cm^2. The wafers are then stripped of resist (Figure 1.16(d)), and thermally oxidised for 3–15 hours, depending on the thickness of oxide required (between 4000Å and 13000Å), with the nitride acting as a mask to prevent oxidation in the active areas (Figure 1.16(e)). The nitride and initial oxide are then stripped by wet etching in ortho-phosphoric acid and hydro-flouric acid respectively (Figure 1.16(f)). The LOCOS process has the advantages of providing a semi-recessed oxide (smooth surface topography), as well as a self-aligned channel stop.

Figure 1.16 also shows the encroachment of oxide under the masking nitride

(the 'bird's beak' effect). This encroachment results in a reduction of the effective transistor channel width, and lateral diffusion of the channel stop implant during subsequent thermal processing will compound this problem, resulting eventually in a limiting packing density for chips fabricated using this type of isolation. Recent developments in isolation technology have attempted to overcome this problem, by using techniques such as SWAMI (**Side**WA**ll** **M**asked **I**solation) [11] [12], SEPOX (**SE**lective **P**oly-silicon **OX**idation) [13], and trench isolation [14].

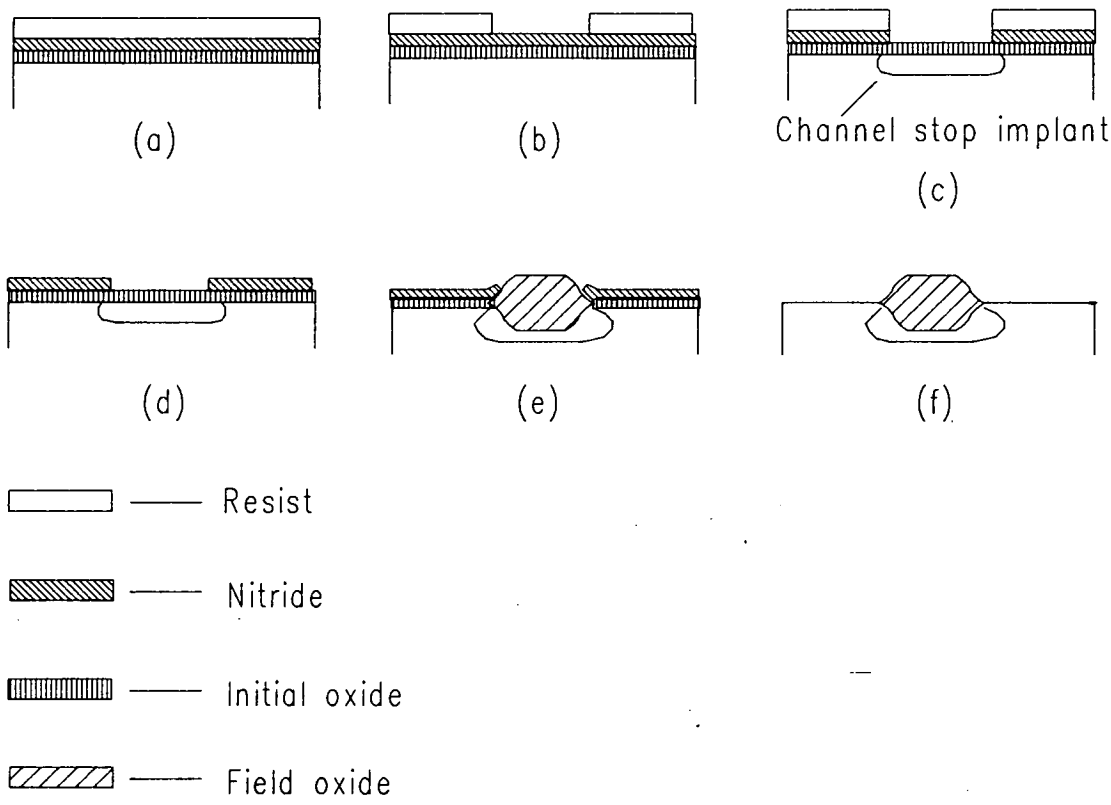

Figure 1.16: Cross-sectional view of LOCOS process with channel stop implant

1.3.4 Depletion Implant, Gate Oxidation and Threshold Adjustment.

After field oxidation, a second layer of photo-resist is applied and patterned, opening up holes for depletion implant (Figure 1.17). The resist acts as a barrier to the arsenic implant, which is applied to ensure that the depletion devices stay permanently turned on. This implant is again of quite high energy (~90keV),

with a dosage of around 10^{12} atoms/cm^2.

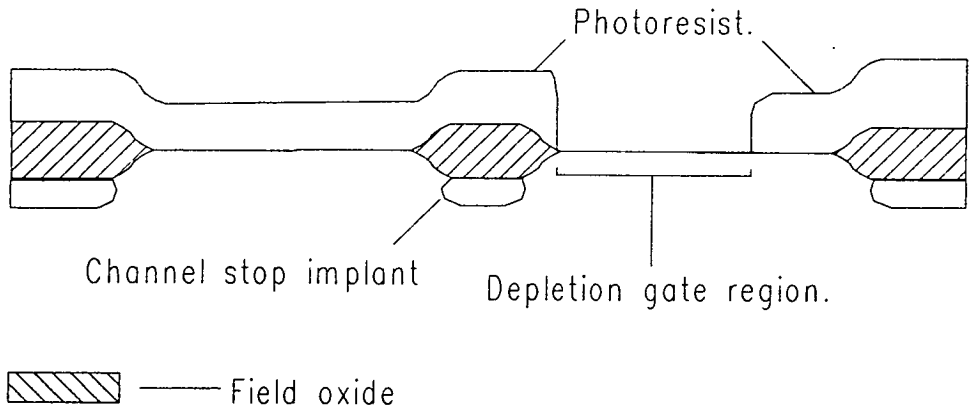

Figure 1.17: Depletion implant photo.

After resist removal, a gate oxide is grown to a thickness of approximately 500Å. It is critical that this oxidation is very well controlled, since properties of the gate oxide (eg. thickness and impurity concentration level) strongly influence device performance, in particular the threshold voltage V_t. A wet oxidation is performed (Figure 1.18) in the presence of hydrogen chloride, which fixes alkali metal contaminants.

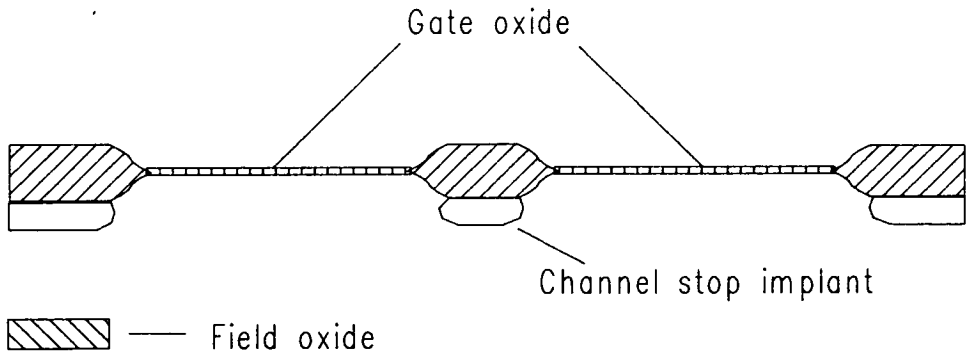

Figure 1.18: Cross-section of gate oxide.

Enhancement and depletion threshold voltages are finally set by a shallow boron implant (~25keV, 3×10^{11} atoms/cm^2). With only the shallow threshold adjustment performed it is still possible for the depletion regions of source

and drain to punch through to each other, causing sub-surface conduction with $V_g \ll V_t$, provided that $V_d > V_s$ (Figure 1.19). This problem is particularly acute at small geometries (gate length $\leq 2.5\mu$m), and is overcome by performing a deeper boron implant (140keV, 2×10^{12} atoms/cm^2), which peaks in concentration around 0.75μm beneath the surface. This is sufficiently deep to prevent punch through, while leaving the threshold adjust implant relatively undisturbed.

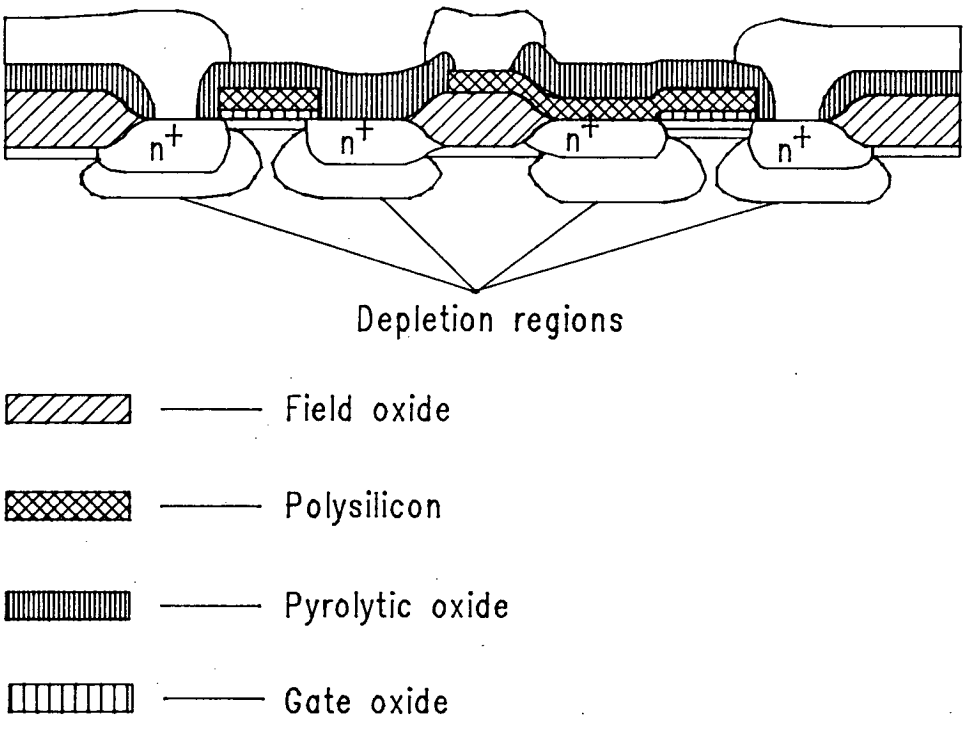

Figure 1.19: Encroachment of source/drain depletion regions under the gate oxide, causing likelihood of punch-through.

1.3.5 Buried Contact.

The third photo-stage defines those areas in which the poly-silicon gate material is connected directly to the silicon substrate, for example when the gate of a transistor is to be tied directly to the source, as in the depletion transistor in Figure 1.15. This stage is known as buried contact, and is accomplished simply by patterning the resist and either wet or reactive ion etching the underlying gate oxide in the areas where contact is to be made. After resist removal, a layer of poly-silicon (~4000Å) is deposited over the whole wafer, and the poly-silicon is doped with phosphorus. It is then oxidised to a depth of ~200Å to

promote adhesion of the next layer of photo-resist. The buried contact process is illustrated in Figure 1.20.

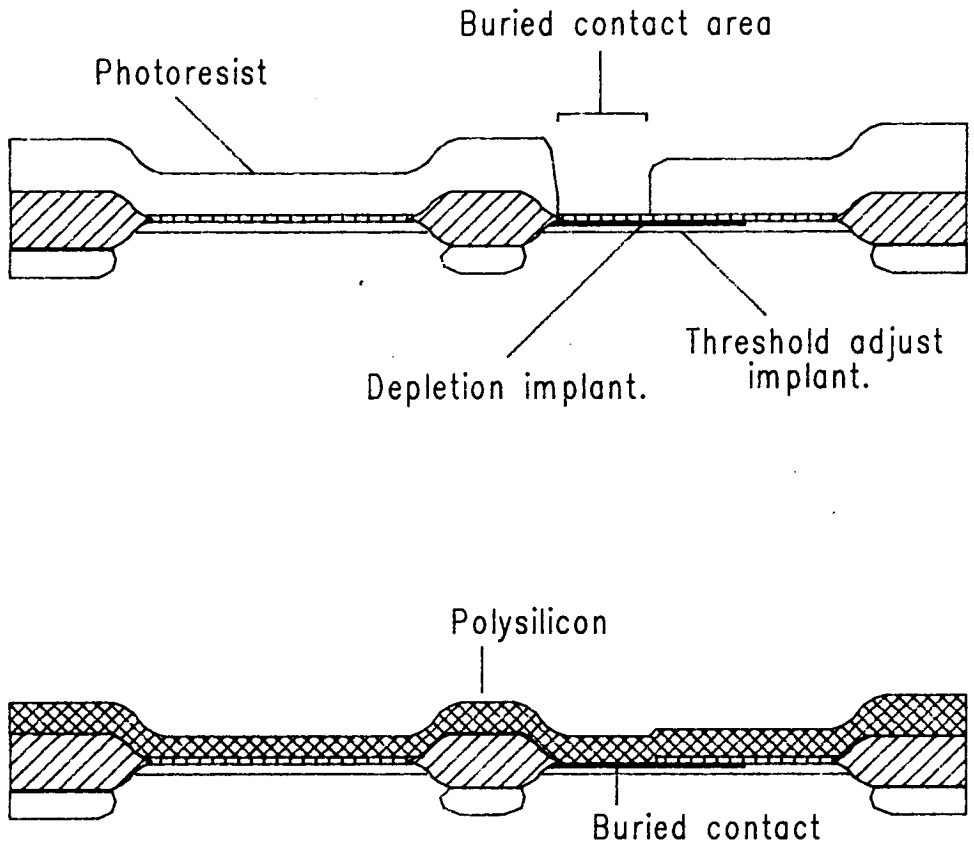

Figure 1.20: Cross-section of buried contact process.

1.3.6 Polysilicon Interconnect and Source/Drain Definition.

The fourth mask defines the poly-silicon gate and interconnect regions. After the poly-oxide has been grown, resist is spun on, exposed and developed. The poly-oxide is then wet etched by immersion in hydroflouric acid, and the poly-silicon is reactive ion etched using freon (CF_4) as the active species. The drain/source implant is performed after resist removal, and consists of high energy arsenic ($\sim$100keV) with a dosage of between 10^{15} and 10^{16} atoms/cm^2. This method of gate definition ensures that the source and drain are self-aligned to the gate itself, and is one of the principal advantages of the poly-silicon gate over the metal gate, eliminating as it does the need for large gate to source/drain overlap which results in a large parasitic capacitance. A small amount of overlap will

still occur, however, due to sideways diffusion of the source/drain implant during the subsequent anneal (Figure 1.21).

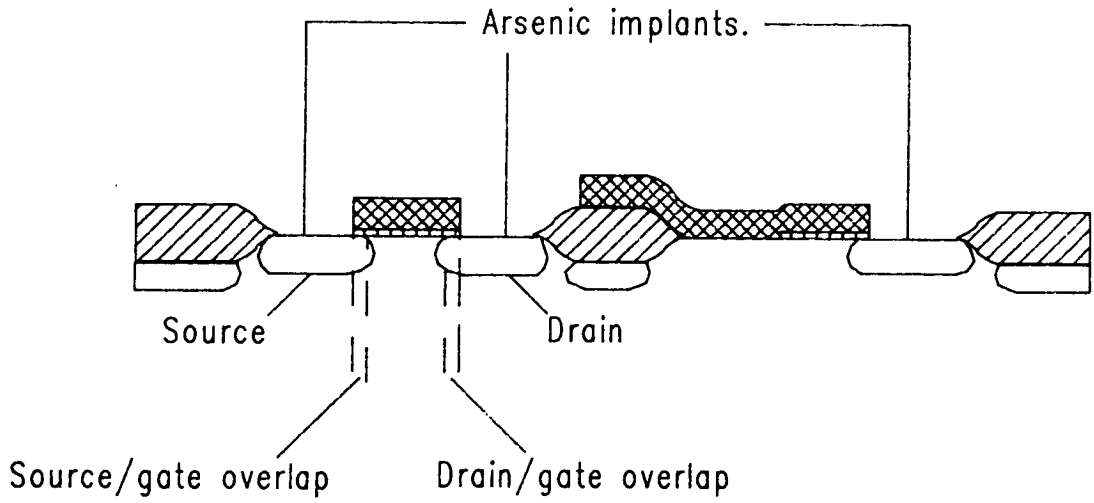

Figure 1.21: Sideways diffusion of source and drain under the gate region.

Recent research has concentrated on the use of refractory metal silicides ($TiSi_2$, WSi_2, $MoSi_2$ and $TaSi_2$) as replacement interconnect for both poly-silicon and aluminium layers [15] [16] [17] [18]. Typical sheet resistances are 1–3Ω/square for refractory metal silicides on top of doped poly-silicon, compared with upwards of 10 Ω/square for poly-silicon only. This reduction in resistance results in faster transistor switching times, and hence to an increase in circuit FTR.

After poly-silicon etch the wafers are oxidised, to a thickness of ~2000Å. The purpose of this oxide is to act as a base for pyrolytically deposited oxide, as well as a barrier against shorting caused by pinholes in the pyrolytic oxidation step. Heavily phosphorus doped silicon oxide is then deposited to a depth of about 5000Å by pyrolytic decomposition. This step is followed immediately by first reflow, which serves to smooth out the pyro-oxide over the poly-silicon steps. The phosphorus allows the material to flow and also acts as a getter for alkali metal impurities in the underlying layers. The purpose of the pyro-oxide is to act as an insulation layer between the poly-silicon gate level and the metal interconnect level (Figure 1.15).

1.3.7 Metal Contacts.

The fifth photo-mask is used to define areas of contact between metal and poly-silicon/diffusion. After the contact areas have been defined lithographically, the pyro-oxide and poly-silicon oxide are reactive ion etched, and the wafers then proceed to second reflow. The purpose of second reflow is to smooth off steps in the pyro-oxide to allow good metal step coverage. A layer of aluminium, or aluminium with ~2% silicon is then deposited onto the wafer by sputtering or evaporation, to a depth of ~1.0μm.

1.3.8 Metal Interconnect.

The sixth photo-mask is used to define the interconnect pattern in metal. This is the most difficult lithography stage in the whole process, because of the high reflectivity and granular structure of the material to be patterned. Both of these effects result in the loss of linewidth control, particularly over topography. After patterning the metal is either wet etched in ortho-phosphoric acid, or reactive ion etched, after which it is given a low temperature anneal (sinter) to provide good aluminium to silicon contacts.

1.3.9 Passivation.

A second layer of pyrolytic oxide is then deposited over the whole wafer. This layer acts as mechanical protection for the circuit, as well as reducing the risk of ionic contamination. A final photo-mask is used to expose the bonding and test pads, and the wafer is then ready for dicing, packaging and testing.

Chapter 2

Problems Facing Lithography.

Lithography faces three main problems as we enter the ULSI age, with each of these problems being critical in determining the ultimate limit of optical techniques in defining patterns in the IC industry:

1. Control of feature size. The relative error in IC feature sizes must be kept constant, or even reduced, as device dimensions decrease. Thus δl, the total variation in printed linewidth (including inter-wafer, inter-die and intra-die variations) must scale down as quickly as the linewidth itself.

2. Reduction of feature size. The requirement for increased resolution in the photo-lithographic process has been discussed in Chapter 1. Usable resolution is dependent upon exposure optics and the resist process being used.

3. Overlay accuracy. The total overlay budget in an IC fabrication process also scales down with device dimensions. Thus, alignment schemes must be developed which allow improved accuracy of pattern placement.

The above problems are common to all lithographic exposure systems, and must be addressed regardless of the nature of the equipment or resist being used. This chapter reviews each of these problems in turn, along with various techniques which have been proposed for their solution.

2.1 Control of Feature Size.

The control of critical dimensions is a complex problem involving the interaction of many factors, including variations in film thickness and exposure energy, as well as interference effects within the film itself. In this section some of these

factors will be discussed, along with some procedures which can be used to lessen their effect on linewidth control.

2.1.1 Standing Waves.

When light is incident onto a reflective surface, the reflected light will interfere with the incident light to produce standing waves with localised maxima and minima in intensity in the vicinity of the surface [19]. Figure 2.1 illustrates this interference phenomenon when a ray is incident onto a resist film coating a silicon substrate. The reflected ray R_3 is dominant in determining the intensity distribution within the photo-resist, since the ratio of intensities in successive reflected rays is less than 5%. Figure 2.2 illustrates a typical latent image produced in resist by photo-bleaching of the photo-active compound during exposure (from the SPESA simulation program, see Chapter 7). Because the solubility of the resist is dependent upon the localised intensity within the film [20], this results in the terraced sidewall profile seen in Figure 2.3, which shows a micrograph of a photo-resist line on a bare silicon substrate.

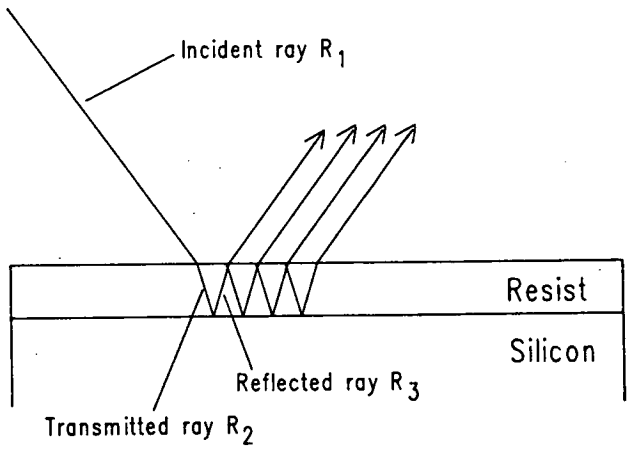

Figure 2.1: Interference in a thin film.

This interference phenomenon can lead to problems with linewidth control. In particular if a node occurs at the resist/substrate interface, the low intensity at this point may prevent the photo-resist from developing out. Minor variations in resist thickness (of the order of $\lambda/4n_{PR}$ where n_{PR} is the resist refractive index, ie. variations of $\sim 0.07\mu$m), can lead to large variations in the width of a developed profile.

In practice standing waves are smoothed out by performing a post-exposure

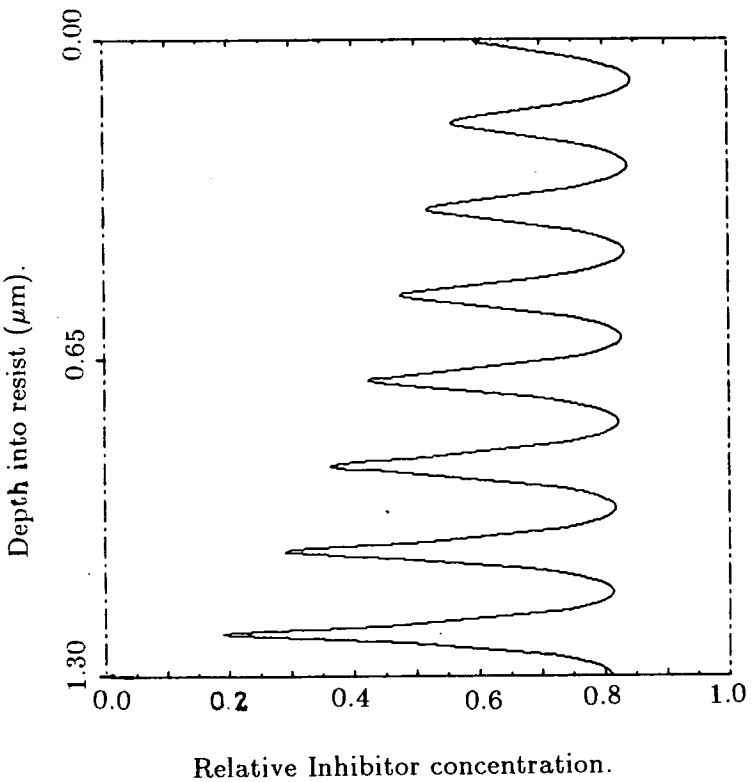

Figure 2.2: Latent image within photo-resist film due to standing wave effect. From the SPESA simulation program.

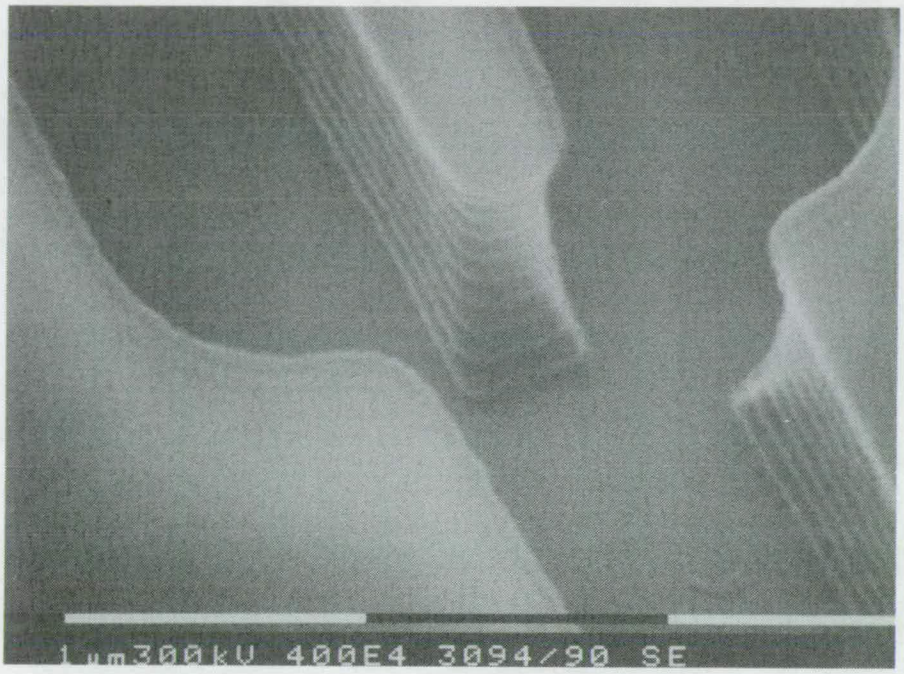

Figure 2.3: Micrograph of resist structure on a silicon substrate, showing terraced sidewall effect.

bake (PEB) after exposure, but prior to development of the resist (for example 90 seconds on a hot plate at 110°C). The baking process redistributes the exposed photo-active compound within the resist resin, thus smoothing out the terraced sidewall in Figure 2.3. This effect can be seen in Figure 2.4, which shows exactly the same type of structure as in Figure 2.3, this time with the post-exposure bake included during processing.

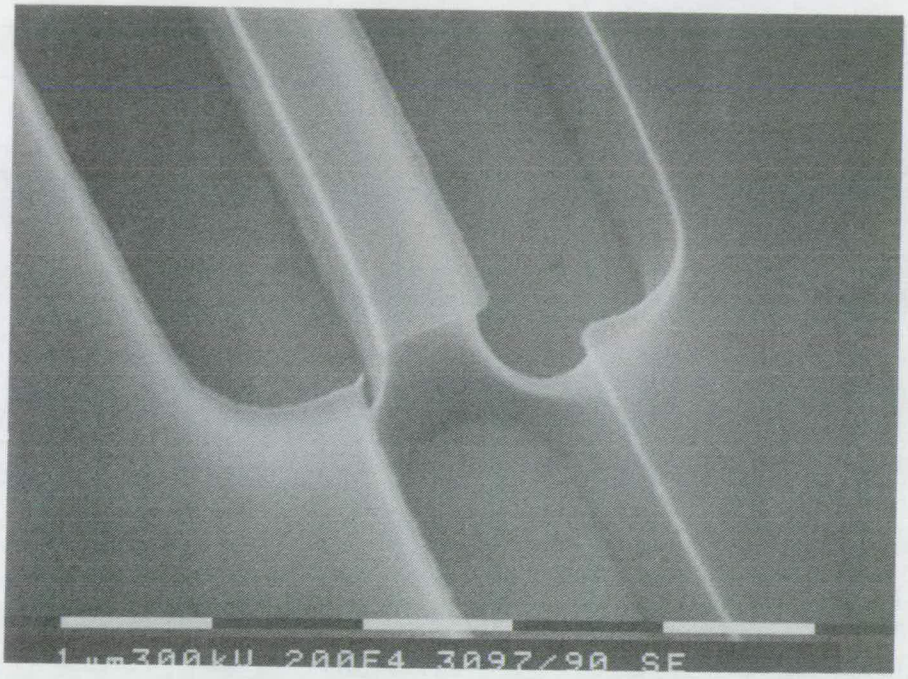

Figure 2.4: Micrograph of resist structure on a silicon substrate, showing effect of post-exposure bake.

Polychromatic exposure is also known to reduce the standing wave effect [8], and works well with reflective systems. The added complication of chromatic aberration, however, makes this approach undesirable when using refractive systems.

2.1.2 Variation of Feature Size Over Steps.

Surface topography itself poses a great problem in terms of linewidth control, since developed linewidth is strongly dependent on resist thickness. Variation in resist thickness over a step, as illustrated in Figure 2.5, results in necking and bulging of the resist line as it passes over topographical features [21] [22],

a phenomenon sometimes known as the 'integrated standing wave effect'. This effect is illustrated in the micrograph in Figure 2.6.

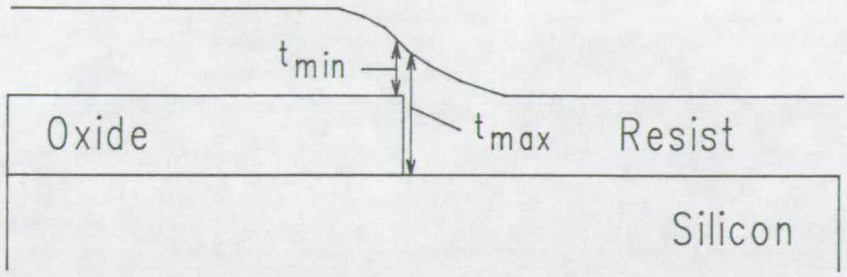

Figure 2.5: Variation of resist thickness from t_{min} to t_{max} over a step.

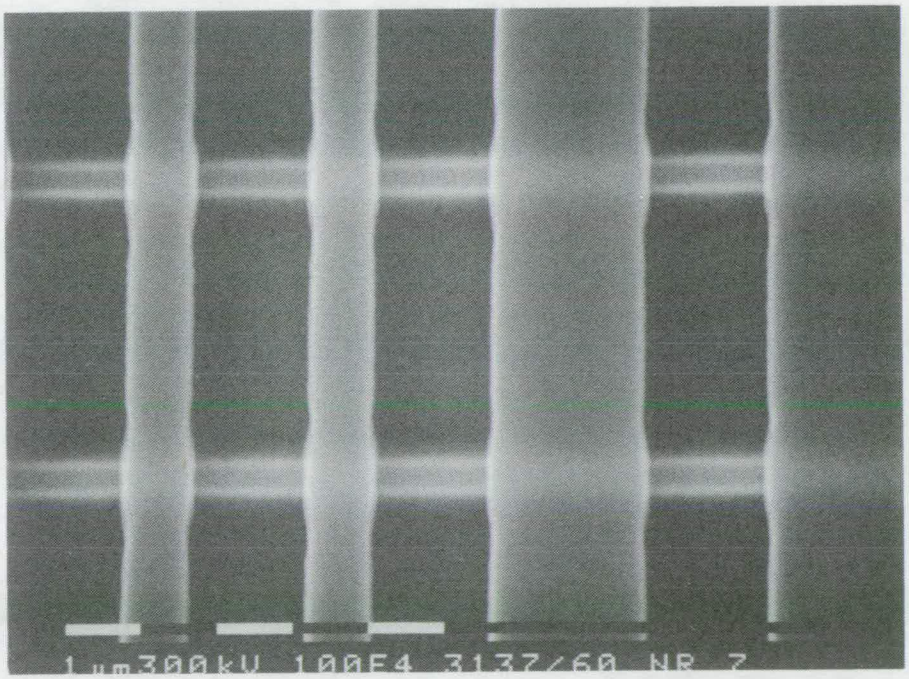

Figure 2.6: Micrograph showing resist necking over topography.

This problem can be successfully alleviated by the use of a suitable anti-reflective coating (ARC) [23], to reduce reflection from the substrate. This leads to a less thickness dependent exposure dose, and hence to a reduction in linewidth variation over steps. Alternatively a suitable dye can be added to the resist in order to reduce substrate reflection [24]. Both of these methods carry a penalty in terms of increased exposure time and hence a reduction in throughput.

Another source of linewidth variation over topography is the bulk effect [25], which results because of non-vertical sidewalls. Figure 2.7 shows a cross section of a resist line (sidewall angle θ) running over an oxide step (of height h) on a silicon substrate.

The width of the line at the silicon substrate is W_1, with the width on top of the oxide, W_2, given by :

$$W_2 = W_1 - \frac{2h}{\tan\theta} \tag{2.1}$$

This difference in linewidth on the silicon and on the oxide is a purely geometrical effect, and can only be reduced by increasing the sidewall angle by using advanced techniques such as image reversal or multi-layer processes (see Section 2.2).

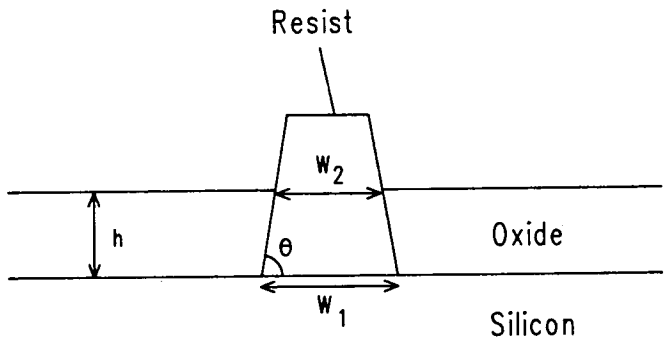

Figure 2.7: Cross-section of resist line over an oxide step, illustrating the bulk effect.

2.1.3 Variation of Linewidth due to Scattering.

Figure 2.8 illustrates the scattering of an incoming ray when patterning a granular substrate such as poly-silicon or aluminium. The non-specular reflection occurring at the resist/substrate interface leads to localised variations in intensity, causing resist line notching, as is evident in the micrograph in Figure 2.9. This notching can again be relieved by the use of ARC or an absorbing dye [26] [27] [28].

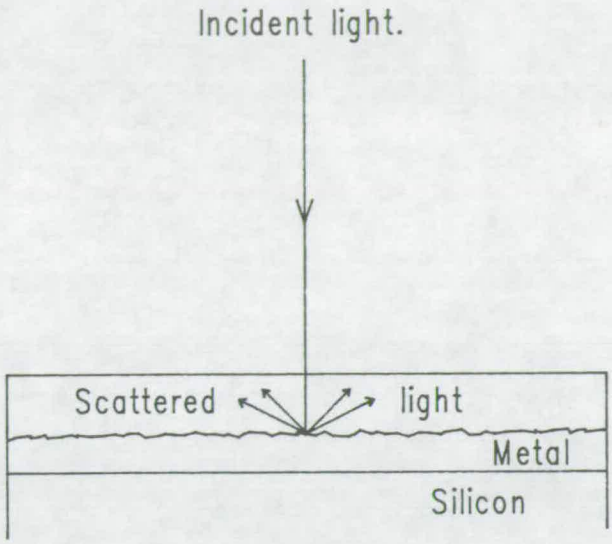

Figure 2.8: Light scattering due to a granular substrate.

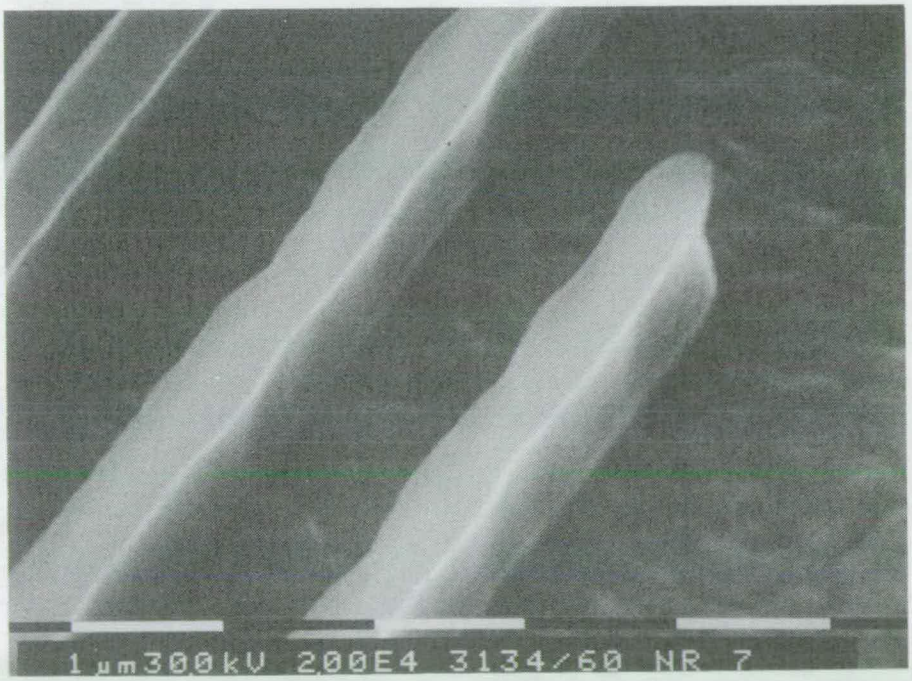

Figure 2.9: Micrograph showing notching due to non-specular reflection from a granular substrate.

2.1.4 Size Dependence.

It has long been known that a mask feature whose size approaches the limiting resolution of a system (feature size $\simeq \lambda/\mathrm{NA}$) requires a different exposure from a feature whose size is large ($\gg \lambda/\mathrm{NA}$), in order to attain the nominal (mask) feature size on the wafer after development [29] [30] [31].

Whether the required dose would be larger or smaller than that required for a large feature would depend upon the feature type, either dark line or clear space.

One possible solution to this problem would be to bias the feature sizes on the mask. Wherever a narrow feature occurred in the design data file, expert CAD software could adjust the width of this feature in order to allow for the effect. Obviously the software would have to be provided with information regarding the feature size at which the effect starts to become important, and the amount of correction which should be given at any particular feature size. This information would be strongly process dependent, and every effort would have to be made to characterise the effect fully and ensure adequate process control before attempting any reticle biasing. Nevertheless this problem will become increasingly important as required feature sizes approach current lens diffraction limits, and progress is required in this direction.

2.2 Reduction of Feature Size.

The smallest printable feature which can be defined by a projection lithographic process is given by :

$$\mathrm{SPF} = \frac{k\lambda}{\mathrm{NA}} \tag{2.2}$$

where k is a process parameter, λ the exposure wavelength, and NA the numerical aperture of the projection optics. Thus reduction of feature size divides itself into two broad categories; process related and optics related. Theoretically $k = 0.25$ and 0.5 for incoherent and coherent illumination respectively, corresponding to the cut-off frequencies of the transfer functions for each type of illumination. In production processes, using conventional positive resist, k is generally in the range 0.9–1.0. Recent advances in resist technology suggest that values approaching 0.6 are possible, using techniques which enhance either the contrast of the resist/developer system or the effective contrast in the aerial image.

2.2.1 Process Related Categories.

Many schemes have been proposed to enhance effective resist contrast. Among the most common extensions to the standard resist processes are multi-layer systems [32] [33] [23] [34], contrast enhancement layers [35] [36] [37], and image reversal [38] [39] [40]. These systems all provide the advantage of steep resist sidewalls, thus extending the useful resolution of the process. Normally this increased resolution has to be traded off against increased process induced defectivity, which results because of the larger number of process steps involved. Defectivity and resolution both have to be taken into account when assessing the economic viability of any new resist process.

Multi-Level Resists.

Multi-level resist schemes are divided broadly into two classes :

1. Bi-layer, or portable conformal mask (PCM) systems. In this system a planarising layer of UV-sensitive organic polymer (originally poly-methyl methacrylate, or PMMA) is spun on to a thickness of about twice the maximum step height on the substrate, in order to smooth over the surface topology. A thin (~0.5μm) layer of standard photo-resist is then spun on, exposed, and developed in the normal manner, and the resulting resist pattern is used as a mask (the PCM) for a subsequent deep-UV exposure. The standard resist is stripped and the PMMA is then developed in methyl ethyl ketone. The process is illustrated in Figure 2.10.

   PCM systems give good resolution (largely due to the planarising layer), and near vertical sidewalls. Disadvantages, however, include poor dry etch resistance of PMMA, poor sensitivity (exposure times of the order of one minute), and process complexity. In addition to this, layer intermixing can occur, which results in residues occurring in the final developed image.

   Improvements to the standard PCM process have included the addition of dyes to the PMMA layer to eliminate substrate reflection, the use of poly-methyl iso-propenyl ketone (PMIPK) as a replacement for PMMA to improve UV-sensitivity and etch resistance, and the inclusion of a layer of ARC between the PMMA and resist to eliminate both reflection and inter-layer mixing.

2. Tri-layer systems. Under this scheme, the pattern is defined in the top layer of photo-resist, then transferred through a barrier layer, and into the

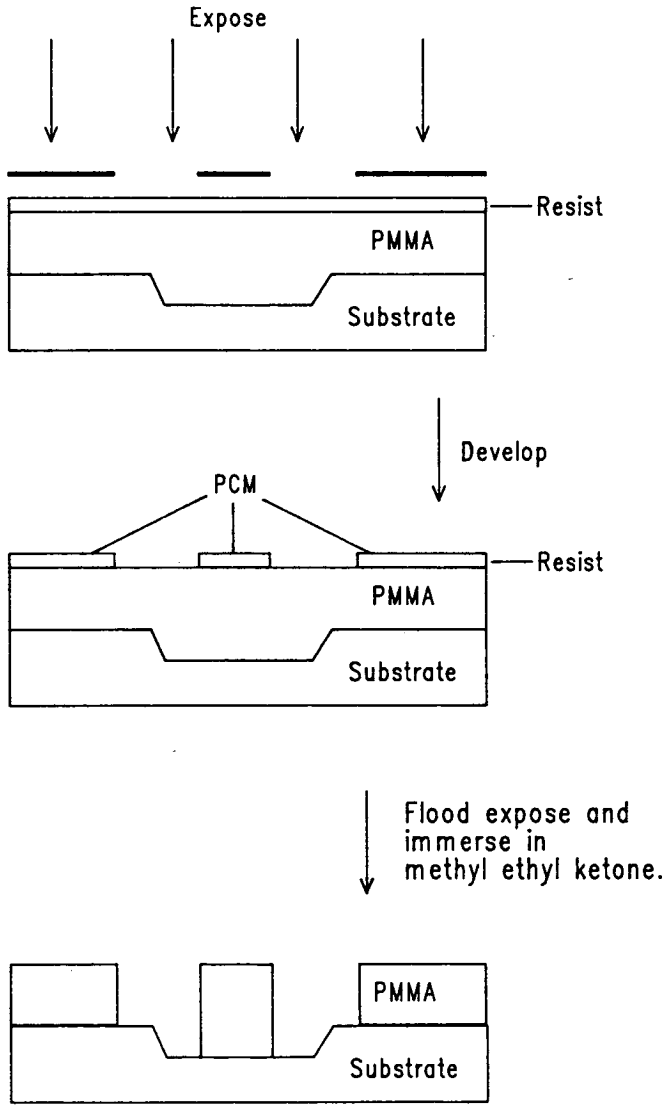

Figure 2.10: Portable-conformal mask exposure process.

planarising layer by RIE etching. The barrier layer prevents the inter-layer mixing common amongst bi-layer systems, and also acts as an etch barrier against oxygen plasma. A typical process uses positive resist for imaging and planarising layers, with a spin-on-glass for the barrier layer. The top layer of resist ($\sim 0.5\mu m$) is exposed and developed, the barrier layer RIE etched in CF_4, and the bottom resist layer RIE etched in O_2.

The principal advantages of tri-layer systems include excellent critical dimension (CD) uniformity and high resolution, as well as vertical sidewalls and good linewidth control over steps. These factors must, however, be traded against the increased process complexity and process induced defectivity involved.

Contrast Enhanced Lithography.

Contrast enhanced lithography (CEL) uses a thin layer of photo-bleachable dye on top of a layer of standard positive resist. The dye bleaches on exposure at a rate which is proportional to the incident intensity, thereby transferring the image to the underlying photo-resist, and simultaneously increasing the effective contrast in the image. CEL is believed to be capable of reducing the k factor in equation 2.2 by a factor of 30%, compared with 20% for bilayer and 25% for tri-layer schemes [23].

Image Reversal.

Image reversal, as illustrated in Figure 2.11, relies on the conversion of carboxylic acid to indene in the presence of a base. The resist is first of all exposed as normal, then the exposed photo-active compound (carboxylic acid) is converted to indene by the inclusion of a basic additive such as monazoline, or by baking in an ammonia atmosphere. The wafer is then flood exposed, converting the remaining (originally unexposed) photo-active compound into acid, and the resist is developed in the usual way.

The reduced dissolution rate of the NOVOLAC resin in the presence of indene increases the resist contrast, extending the resolution limit of the system. In addition to this, the flood exposure dose can be used as an extra degree of freedom to control the resist profile. It is possible using this method to go from normal tapered resist profiles, through vertical sidewalls to undercut profiles, simply by varying the flood exposure dose. Undercut profiles generated using this method have been found to be extremely useful in lift-off processing [41], in

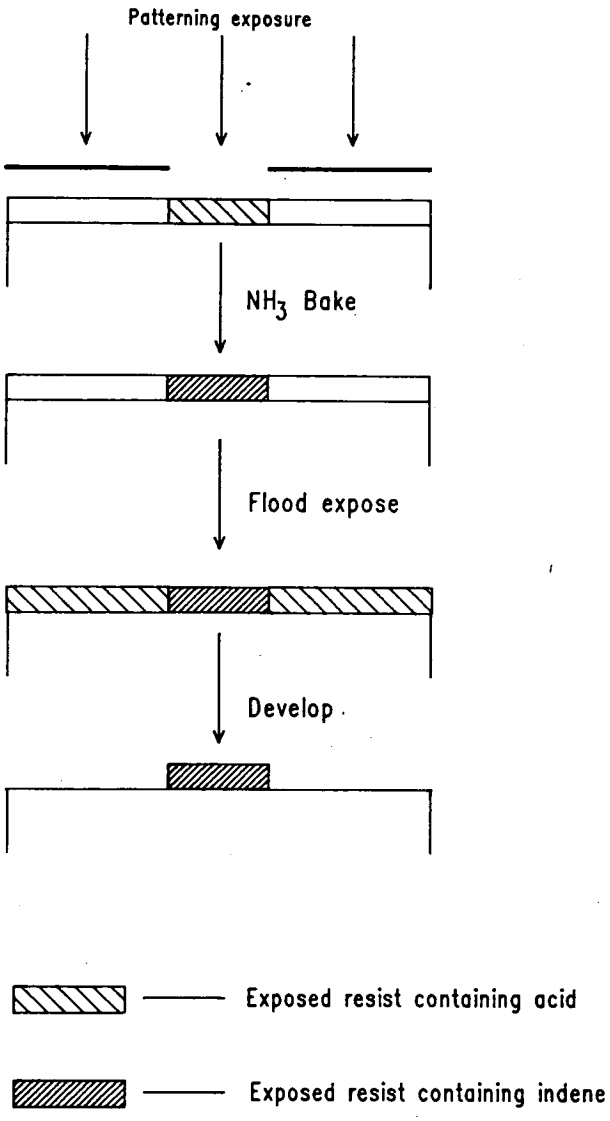

Figure 2.11: Image reversal process.

which metal is deposited on top of resist and selectively removed by developing.

2.2.2 Optics Related Categories.

From Equation 2.2 there are two obvious methods which can be used to increase the resolution lithographic equipment; either decrease the exposing wavelength or increase the system numerical aperture.

Usable exposure wavelengths in refractive systems are at present limited to 365nm and above, since conventional optical glasses begin to absorb strongly below this level [42]. It is possible that in future, quartz based lenses transmitting into the deep-UV will be developed. Materials of this type tend to be limited in terms of refractive index and dispersive power, however, which limits the lens designer greatly. Reflective systems can of course be used beyond 365nm, well into the deep-UV [43]. At present all reflective systems used in micro-lithography are 1:1 systems, however, which limits them in terms of sensitivity to defects.

Numerical aperture has no such limits imposed upon it by optical materials. Current production lenses with numerical apertures ~0.3 are likely to be replaced by 0.35NA and even 0.42NA lenses over the next ten years. However, depth of focus (the maximum allowable distance between the top surface of the resist and the focal plane while still maintaining good image fidelity) decreases with the square of the numerical aperture :

$$\delta z = \frac{\lambda}{2\mathrm{NA}^2} \tag{2.3}$$

This necessitates the use of planarising layers and thin resist when using high numerical aperture lenses on substrates with severe topography.

Numerical aperture also has to be traded off against field size. Objective lenses with numerical apertures around 0.8–0.9 are commonly found on optical microscopes, with image diameters around 100μm. For field sizes likely to be required in future IC fabrication (maximum 20mm in diameter), numerical apertures of around 0.5 are likely to be the limit, giving a limiting working resolution of $\sim 0.7\mu$m for single layer resist processes.

2.3 Overlay.

Overlay in IC fabrication is defined as the total pattern registration accuracy between one process level and another. As such it encompasses a number of sources of pattern placement error, including wafer distortion, die distortion,

alignment accuracy and stage accuracy, each of which must be minimised in order to ensure optimum registration. As a rule of thumb the total overlay accuracy of a process has to be roughly 20% of the minimum feature size; thus for a process with 1.0μm design rules, a total pattern placement error not exceeding 0.2μm is allowed. Some of the important sources of overlay error are discussed below.

2.3.1 Wafer Distortion.

Inelastic wafer distortion often occurs during processing steps in which a wafer is subject to thermal stress. This distortion is caused by a mismatch in the thermal expansion coefficients of the material being deposited and the underlying substrate. If the material is deposited at high temperature, wafer bowing occurs on cooling down to room temperature [44]. Bowing is an overall linear effect, resulting in a simple effective expansion or contraction of the wafer. Local variations in temperature and film thickness can result in non-linear, or random, in-plane distortion, however, and it is difficult to reduce this random component to below the 0.1μm level [6].

While it is still possible, using whole wafer projection, to correct for linear distortion by varying the magnification of the system, it is not possible to correct for non-linear distortion, or for die placement errors in maskmaking [45], using these systems. This again limits overlay accuracy in such a case.

Finally, optical distortion of whole-wafer systems will also affect overlay. In general, linear optical distortion will be well-corrected in a diffraction limited system (by definition a diffraction limited system possesses negligible amounts of first order aberrations, including distortion and curvature of field) [46]. Non-linear distortion can and will occur, however, and this distortion will vary from system to system. If more than one system is used for any particular process batch, mismatched lenses will contribute significantly to the total overlay error.

The cumulative effect of these errors is such that, if overlay accuracy to within 0.2μm or better is required, the use of step-and-repeat exposure systems is necessary.

2.3.2 Die Distortion

Non-linear optical distortion is important also for step-and-repeat systems, at the intra-die level [47]. Again the use of more than one stepper, or a 'mix and match' approach (typically, the use of a combination of whole wafer projection with step-and-repeat) can limit overlay accuracy because of imperfect matching of lenses.

Techniques have been proposed by which lenses whose distortions most closely match each other can be used on any particular process run [48], although the optimum overlay solution is still to dedicate each batch to a particular machine (a solution which does not necessarily lend itself to high wafer throughput).

2.3.3 Alignment Accuracy.

The total overlay accuracy of a system (whole wafer or step-and-repeat) will also depend upon the accuracy of the alignment sub-system being used. Many methods of achieving correct alignment have been proposed, all of which involve the measurement of the relative positions of a reference mark on the mask, and a mark on the wafer which has been printed at an earlier stage. This is achieved by shining light through a mark on the mask and picking up the light which undergoes either reflection, diffraction, or scattering at the wafer mark. Each of these methods of detection has its own particular merits, and they are discussed more fully in Chapter 3.

2.4 Conclusions.

A number of methods have been proposed for the improvement of processing and control of limiting feature sizes in micro-lithography, and some of these methods have been reviewed here. Control and reduction of feature size are to a large extent dependent on one another, since improving linewidth control allows smaller features to be printed with the same relative width variation. Each process must be assessed for factors such as improved process control, possible increase in defectivity due to complex processing, and cost of the complete process, before any decision is made with regard to the economic viability of such a process in production.

Chapter 3

Lithography Alignment Systems.

Many different approaches to the problem of mask-to-wafer alignment have been proposed over the years, with systems based on reflected, diffracted, and scattered light finding application in a number of different machines. Reference [49] gives a general overview of the specifications and options on the various machines.

3.1 Reflected Light Systems.

Reflected light schemes rely on a reflectivity change when an intensity scan is performed on a mark engraved on the wafer. Light is shone through a reference mark on the reticle, reflected from the wafer, and the super-imposed image of the wafer and reticle marks is picked up by a suitable detector. These systems are sometimes referred to as light field alignment systems, since the principle of operation is identical to that of a light field microscope, with the mask in the front focal plane of the objective.

3.1.1 Optimetrix.

Reticle-to-wafer alignment using the Optimetrix wafer stepper can be accomplished in a variety of ways, with each alignment mode performing automatically or under manual control [50]. The basic strategy can be summarised as follows:

1. On-axis reticle-to-wafer stage alignment, in x, y and θ.
2. Off-axis coarse global wafer alignment, in x, y and θ, using large wafer marks.
3. Off-axis fine global wafer alignment, in x, y and θ, using small wafer marks, *and/or*

on-axis fine global wafer alignment.

4. On-axis die-by-die alignment, in x and y, for maximum registration accuracy,

   or

   blind step, for maximum throughput,

   or

   on-axis zone alignment (alignment to selected die on the wafer) for high accuracy and increased throughput.

On and off-axis refers to the use of the on and off-axis microscopes for alignment (see Figure 3.1).

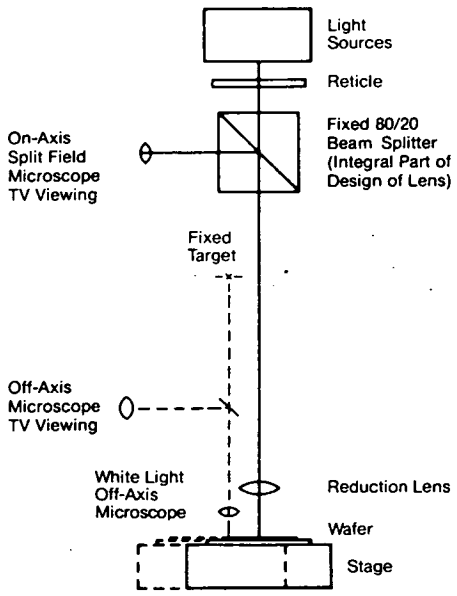

Figure 3.1: Optimetrix alignment scheme, taken from reference [50]

Figure 3.2 illustrates the structure used for fine reticle alignment, on-axis wafer alignment, and on-axis die-by-die alignment. The super-imposed images of the wafer mark (or wafer-stage mark, in the case of reticle alignment) and the reticle mark, are viewed through the on-axis microscope. A TV camera that is coupled to the microscope scans the super-imposed images and measures their relative alignment by looking for symmetry in each half of the chevron.

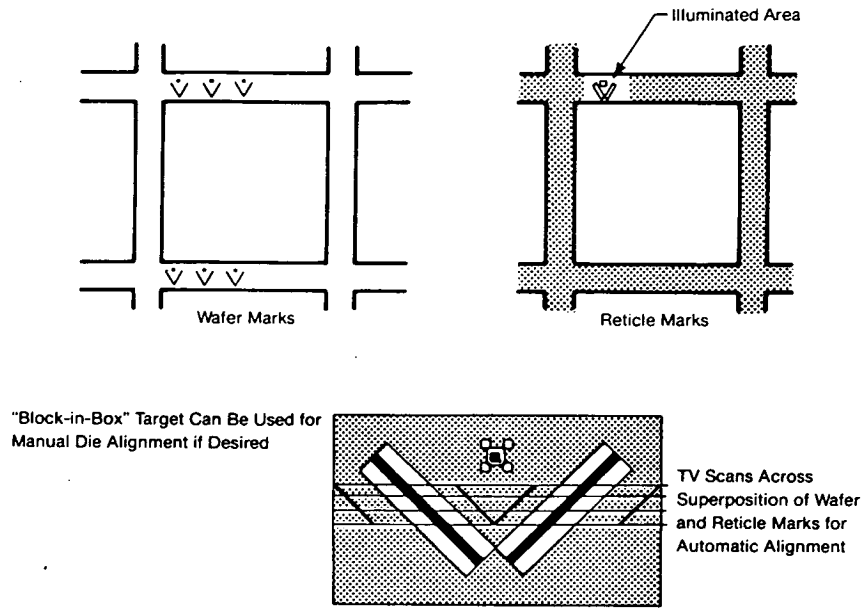

Figure 3.2: Optimetrix reticle/wafer alignment marks, showing possible placement of marks within the die, from reference [50]

Alignment corrections in x and y are accomplished by moving the wafer stage according to the flow diagram in Figure 3.3

Figure 3.4 shows an example of alignment data which can be obtained from the system under optimum conditions. The data exhibits good contrast, and the system software has no problem in aligning to such a mark. Figure 3.5, however, shows the type of mark which can be obtained under adverse conditions. In addition to the high noise level which is evident, the mark also shows evidence of diffraction fringes which can confuse the software to the point of giving erroneous alignment. This problem is inherent to all systems based on reflected light, and is the one major disadvantage of such a scheme.

The advantages of the Optimetrix system include the on-axis, real time viewing of both wafer and reticle. The on-axis microscope may be driven anywhere in the reduction lens field of view (field of view of the microscope is around $200\mu\text{m}\times100\mu\text{m}$), allowing the alignment marks to be placed anywhere within the die. This is found to be particularly useful in a research environment, where maximum flexibility in design is required.

In addition to this, the video analysis program is able to cope with a very wide range of contrasts in the alignment signal, allowing for considerable process

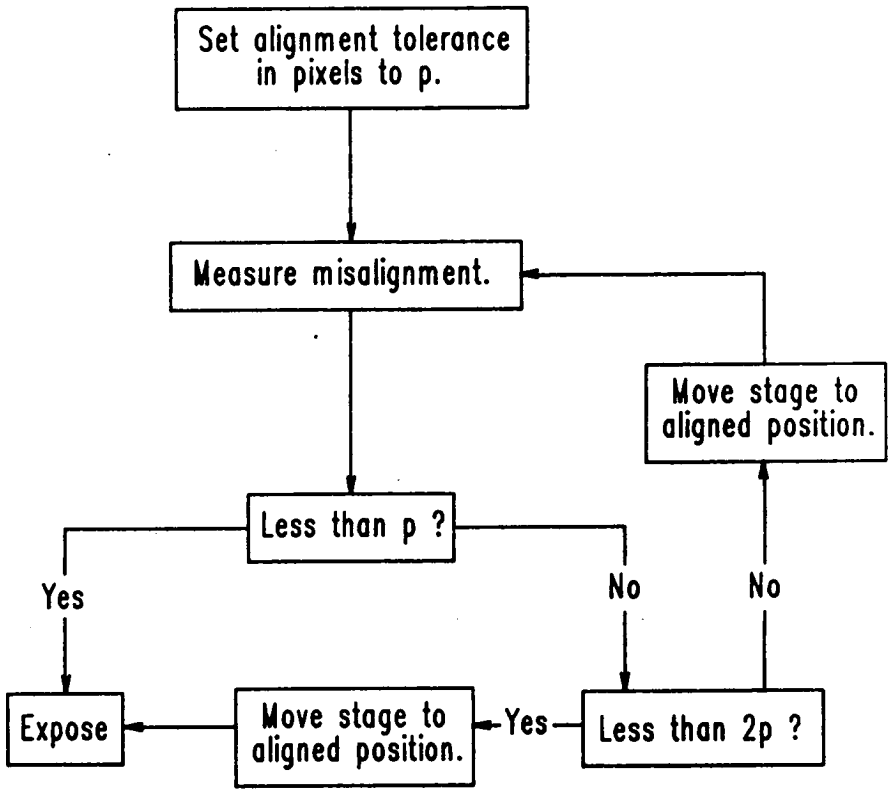

Figure 3.3: Optimetrix alignment flow diagram.

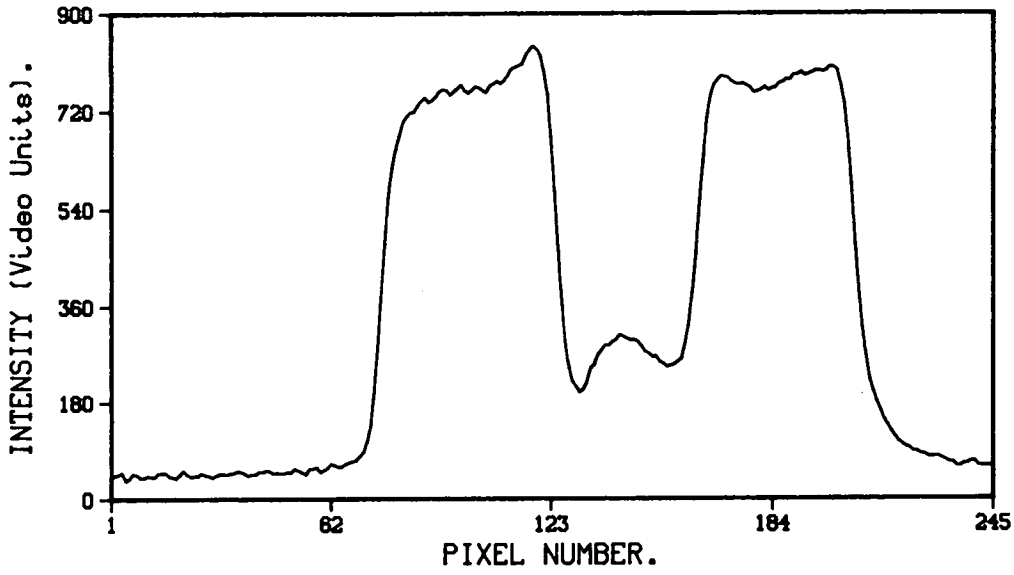

Figure 3.4: Optimetrix alignment data set, exhibiting good contrast (alignment of metal contact to active area).

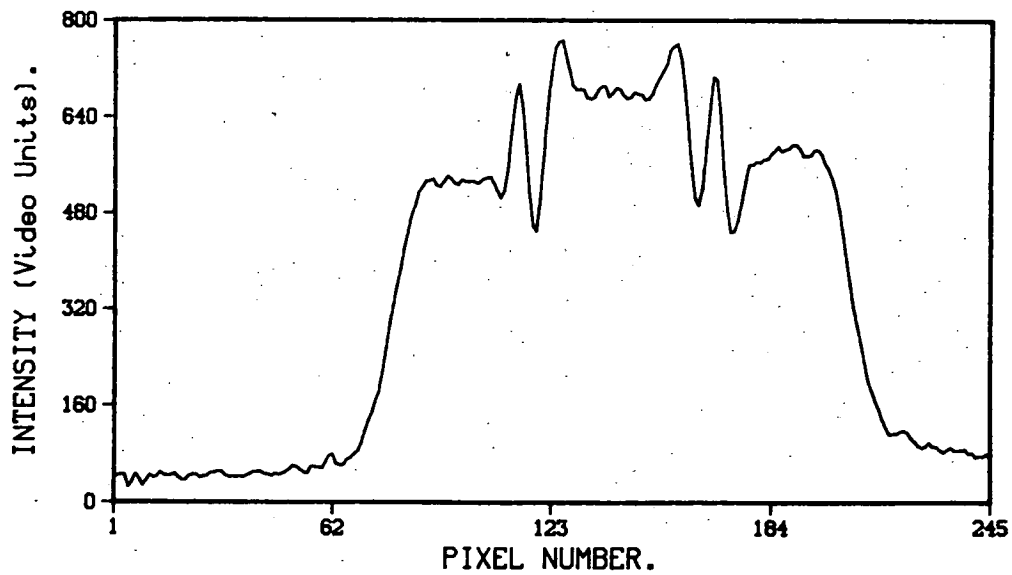

Figure 3.5: Optimetrix alignment data with low contrast, high noise level, and edge fringing (alignment of poly-silicon to active area).

latitude. In fact, because the system looks for symmetry, it is able to cope with very low mark contrast, provided that edge definition is good. Figure 3.6 shows an example of such a mark.

The alignment system is fairly slow, however, (due to the fact that it is based on signal processing software rather than dedicated hardware), with alignment on each die taking as long as 0.5 seconds. Accuracy is claimed to be $\pm 0.20\mu$m (3σ), giving a total overlay of ± 0.25 (3σ) [49].

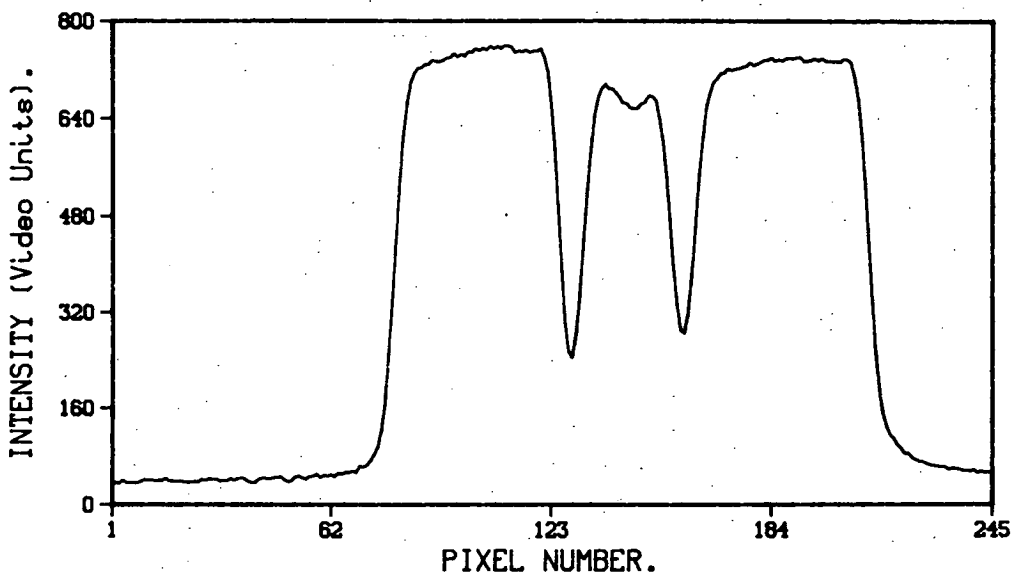

Figure 3.6: Optimetrix alignment data, exhibiting good edge definition, but low contrast (alignment of buried contact to active area).

3.1.2 Perkin-Elmer/Censor.

As was stated earlier, all reflected light systems use the same alignment principle, any differences being concerned with the structure of the marks and method of detection of the alignment signal [51]. The Censor wafer/reticle marks, and detector system are illustrated in Figures 3.7 and 3.8 respectively.

The alignment strategy can be summarised as follows:

1. Optical prealignment (flat-finding).

2. On-axis die-by-die alignment, in x, y and θ.

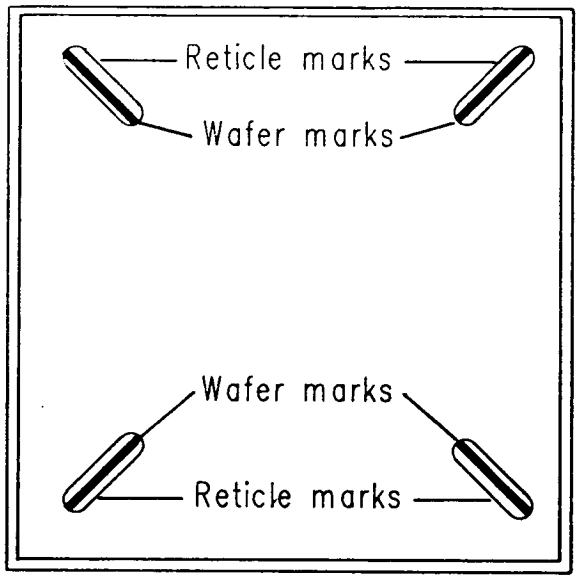

Figure 3.7: Structure of Censor wafer/reticle marks and placement within die.

Only one exposure mode is available - full die-by-die alignment. It uses four wafer marks arranged radially on the die, and these are viewed through corresponding radial reticle windows. The detector assembly used to scan the superimposed images consists of a rotating polygon mirror and a photo-diode. Figure 3.9 illustrates the intensity profiles obtained by the scan. Measuring the relative alignment of any three of the four reticle/wafer marks is sufficient to compute the error in x, y and θ. Normally all four marks are scanned, and a best fit is computed using all the signals. Alignment is performed by moving the reticle at each site in x, y and θ. The alignment marks are scanned continuously, with the reticle position being adjusted and rechecked only once.

Using this alignment method it is necessary that the wafer mark appears dark against a light background, with the mark contrast C (see Figure 3.9), equal to or greater than 20%. For this reason the Censor can use one of two alignment wavelengths, 547 nm or 578 nm. These two wavelengths were chosen to optimise the mark contrast over a wide range of resist thickness.

The use of different wavelengths for alignment and exposure is responsible for the folding of the optic axis at the detector assembly (due to the difference in focal lengths for the different wavelengths), and for the structure of the alignment marks (only radial lines can be imaged properly at the alignment wavelengths).

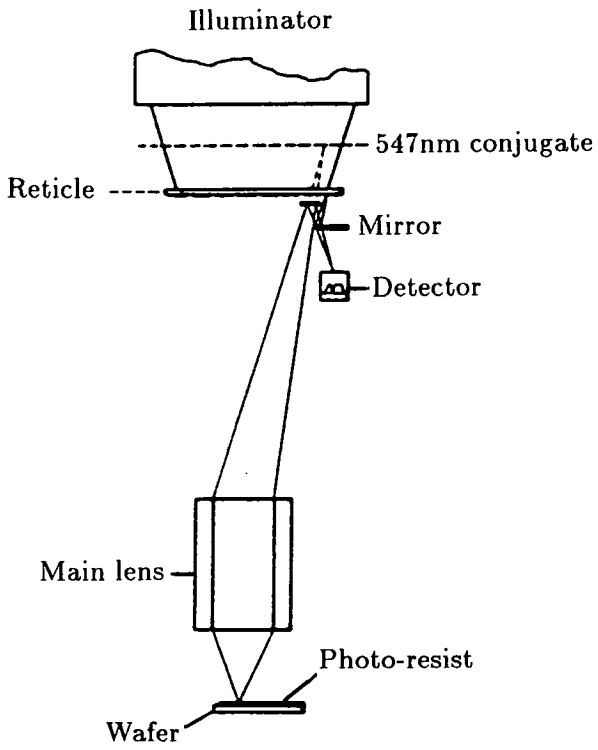

Figure 3.8: Schematic of Censor alignment system, from reference [51].

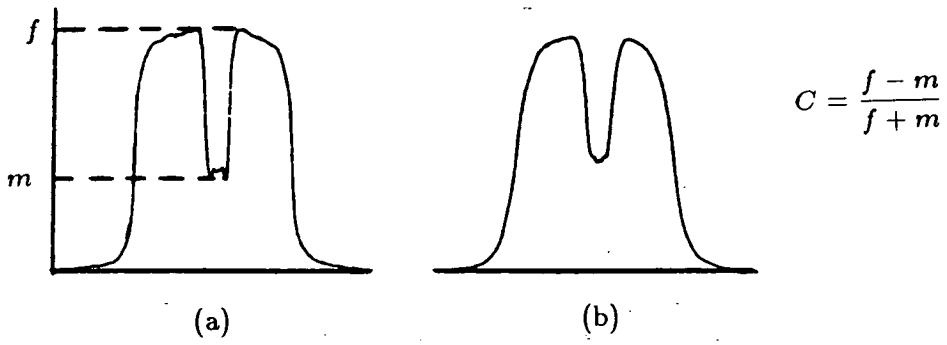

Figure 3.9: Censor alignment signals, in focus (a), and out of focus (b).

In addition to this, alignment and focus offsets occur which must be compensated during machine set-up.

The main advantage which the Censor possesses is high alignment speed. It achieves this by scanning the alignment marks continuously and aligning and rechecking only once. In order to achieve this speed, a consistent mark contrast is necessary, resulting in the various machine complications discussed above. In addition to this, adjusting the reticle rather than the stage position for alignment gives a gain in accuracy, due to the magnification factor of the lens (since smaller effective adjustments are possible). An alignment accuracy of $\pm 0.15\mu$m (3σ) is claimed for this system [49].

3.1.3 TRE.

The principle of alignment used by the TRE stepper is very similar to that used by the Optimetrix machine. Figure 3.10 shows a schematic of the TRE system, illustrating the through-the-lens viewing system [52]. The alignment strategy can be summarised as follows:

1. Reticle to optical column alignment (x, y and θ).
2. Optical prealignment (flat-finding).
3. Global alignment (x, y and θ).
4. Die-by-die alignment (x and y) for maximum accuracy,

   or

   blind-step for maximum throughput,

   or

   zone-alignment for high accuracy and increased throughput.

Firstly the reticle is automatically aligned to two marks which are engraved in the reticle stage, which is rigid with respect to the optical column. The wafer is then optically prealigned to establish the position of the flat. A global wafer alignment is then performed by making measurements in x and y on two marks (as illustrated in Figure 3.11) at opposite sides of the wafer. Die-by-die alignment is performed, again by viewing the superimposed images of wafer mark and reticle window through the lens, this time slightly off-axis. The alignment position is in this case slightly different from the exposure position, and overlay error will be a combination of alignment error and stage stepping error. The

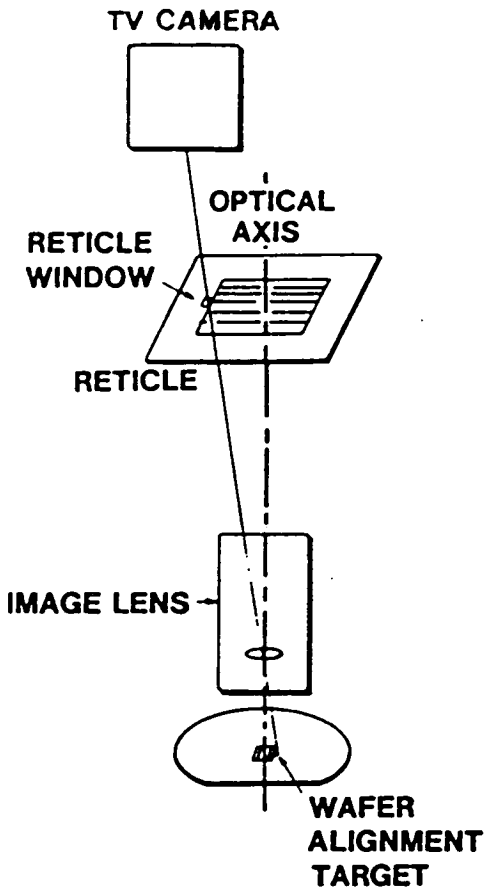

Figure 3.10: TRE alignment system.

stage stepping error should be small ($\leq 0.1\mu$m), since the stage is controlled by a laser interferometer.

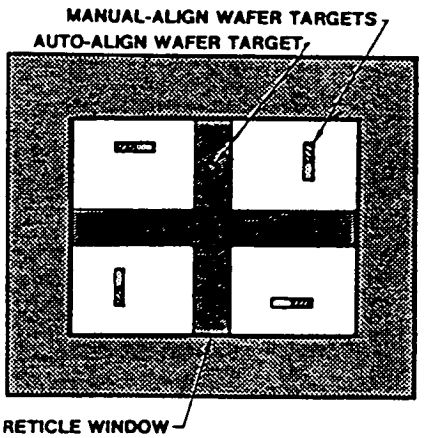

Figure 3.11: TRE global/die-by-die alignment mark.

Figure 3.12 illustrates an idealised intensity profile obtained from the TV screen. The screen is scanned 256 times in both x and y, with each pixel from a horizontal scan corresponding to 0.4μm and from a vertical scan to 0.3μm. Statistical theory is used to obtain an alignment precision significantly smaller than the pixel size [52]

The TRE system has the same advantages as the Optimetrix, real time through the lens viewing (of alignment marks only) and a software suite capable of coping with a wide range of mark contrasts. Blind step and zone-alignment, as in the Optimetrix case, allow throughput to be traded off against accuracy.

Because of slight off-axis viewing, the position of alignment marks is restricted to the edge of the reticle. The extra stage stepping involved requires a very precise stage, and the choice of a laser position monitor makes it desirable to have an environmental chamber with a consequent large footprint (floor area in the cleanroom). A total overlay performance of ± 0.25 (3σ) is quoted for this machine [49].

3.1.4 GCA with FAS (Field Alignment System).

Figure 3.13 illustrates the GCA FAS alignment system. The principle is very similar to that used in the Optimetrix, with a retractable reflective pellicle taking the place of the beam splitting cube, and produces similar reflected intensity

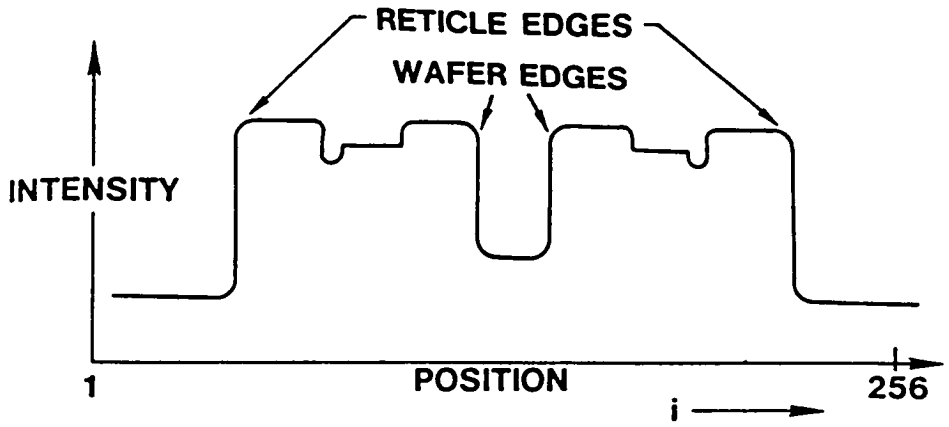

Figure 3.12: Idealised TRE intensity profile, from reference [52].

profiles. Again the use of advanced digital image processing allows correct alignment to marks with widely varying contrast. In this case the positions of zero crossings of the second derivative of the signal are taken to represent positions of wafer and reticle feature edges. The signal-to-noise ratio is kept as high as possible by averaging the signal over as many as 480 lines of data, as well as employing both a photo-multiplier and vidicon with internal gain [53]. The GCA alignment strategy can be summarised as follows :

1. Reticle alignment to optical column (x, y and θ).
2. Global (off-axis) wafer alignment (x, y and θ).
3. Die-by-die alignment for maximum accuracy (x and y),

   or

   blind-step for maximum throughput,

   or

   zone-alignment for high accuracy and increased throughput.

Zone-alignment on the GCA can take the form of either aligning to every nth die, or aligning to the first n die on a wafer, and applying an average correction factor to all subsequent die on the wafer. Initial tests on artifact wafers suggests an alignment accuracy of $\sim 0.15\mu$m (3σ) [53].

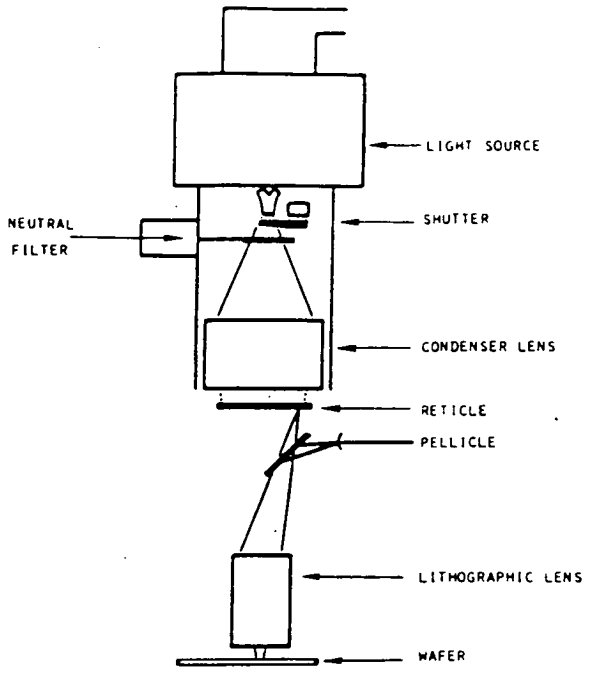

Figure 3.13: GCA FAS alignment system.

3.1.5 Additional Systems Based on Reflected Light.

Doemens and Mengel have reported an alignment system based on reflected contrast for use in an x-ray aligner [54]. The alignment mark has a cross-in-box type structure similar to that used in the TRE machine, repeated a total of nine times per TV screen to improve counting statistics (Figure 3.14). Accuracy of alignment is claimed to be a trade off between the total number of cross-in-box structures and the size of each individual structure. Nine is believed to be the optimum figure in this instance. Alignment repeatability is around 0.04μm. The speed of the system is fairly slow, however, taking a few seconds per alignment. Since this is a whole wafer projection system, the penalty in terms of throughput is not severe. Accuracy is good for this type of system, but it should be remembered that exposure is accomplished by x-rays, allowing optical parameters such as coherence, wavelength, and illumination bandwidth to be optimised for the reduction of diffraction fringes. This is in general not the case for an optical system where chromatic aberration and exposure of resist during alignment determine the desired wavelength and bandwidth, while image contrast and depth of focus considerations determine the required partial coherence.

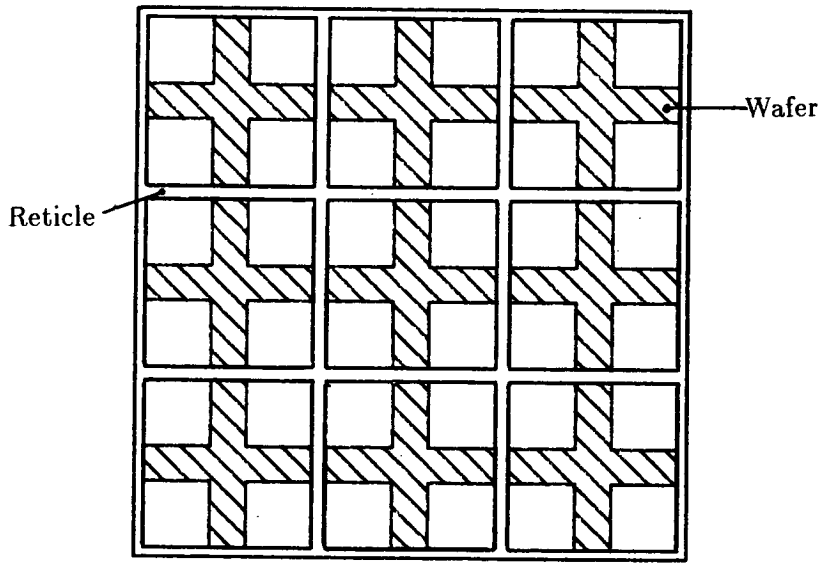

Figure 3.14: Nine cross-in-box structures to improve count statistics, from reference [54].

3.2 Diffracted Light Systems.

Diffracted light alignment systems rely on the well known properties of regular structures to diffract light in certain directions. The most common types of structure in use are Fresnel zone plates, which act as simple lenses, bringing collimated light to a focus above the wafer surface [55], and square wave gratings, which give reflected maxima and minima in intensity as a function of angle [56]. Both types of diffraction system have been studied extensively for the purposes of alignment, with circular [57] [58] [59] and linear [60] [61] Fresnel zone plates and linear gratings [62] [63] [64] finding wide application in proximity, projection, and x-ray lithography.

3.2.1 GCA with 5840 SXS (Site X Site) SiteAligner.

Figure 3.15 shows a schematic of the GCA 5840 SXS SiteAligner, which uses Fresnel Zone Targets (FZT's) imaged onto quadrant detectors to detect x-y misalignment. The position of the wafer stage is adjusted until the radiation from the zone plate is centered on the detector. The positions of the real and virtual images produced by the FZT are measured in order to allow for local tilt of the

wafer (see Figure 3.16) [59]. A summary of the GCA alignment strategy has already been given in Section 3.1.4.

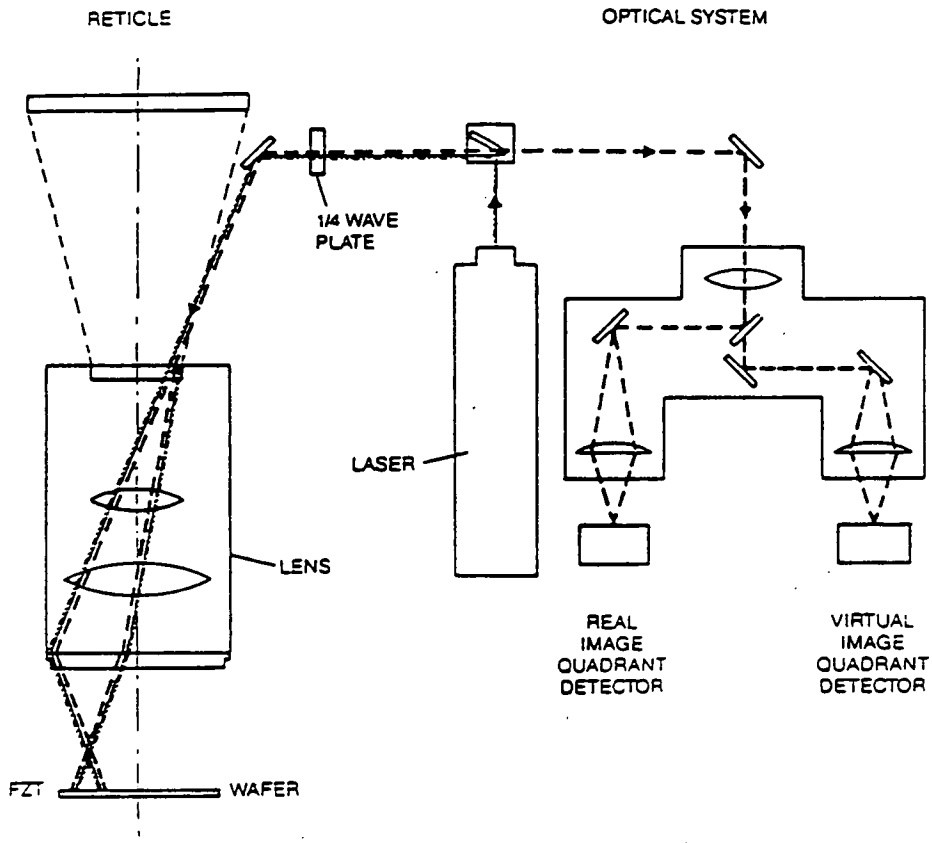

Figure 3.15: GCA FZT alignment system, from reference [59].

FZT's, and alignment signals obtained from them, have proved in the past to be extremely process-sensitive. One of the prices to pay when using such marks is that process steps have to be altered in order to optimise the marks for each particular level. Problems with baseline drift have also been shown in the past to contribute significantly to total overlay error, since the global alignment is performed using an off-axis microscope. Monitoring of baseline drift has been necessary, and in the past this has been a slow and tedious procedure. As with other hardware based systems, alignment time is fairly short (300 msec. maximum) [65]. Overlay performance is specified at $\pm 0.2\mu$m (95%).

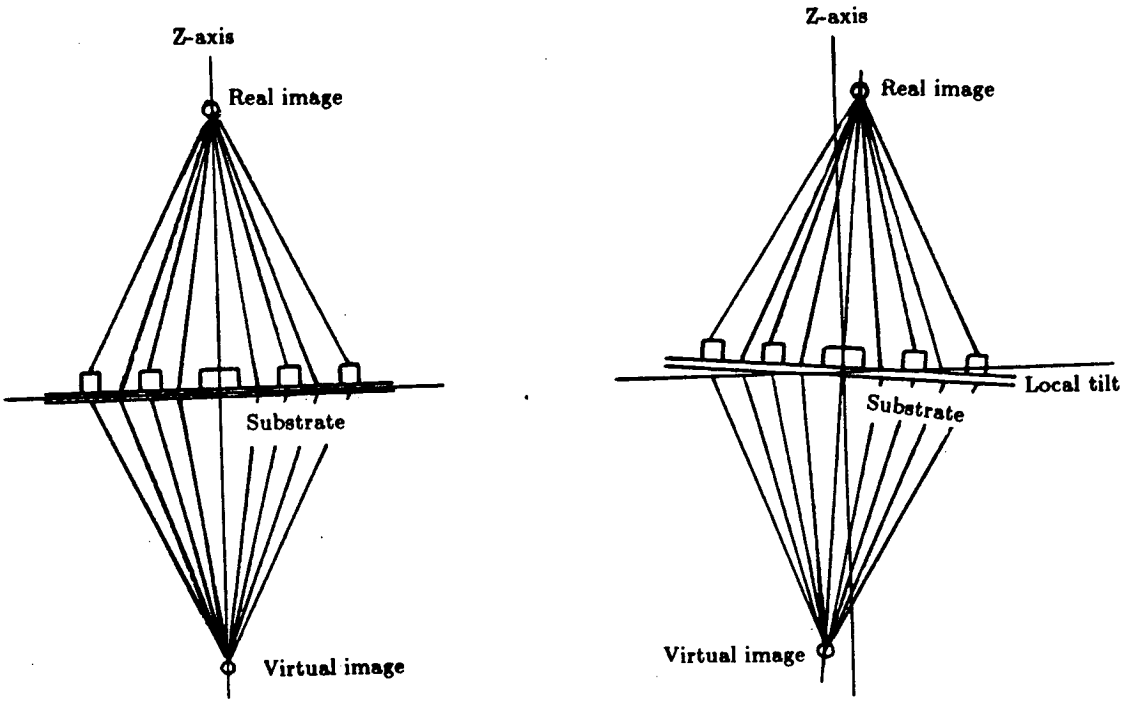

Figure 3.16: The use of real and virtual FZT images to eliminate the effect of local wafer tilt, from reference [59].

3.2.2 Additional Fresnel Zone Target Systems.

Feldman et al. [57] have proposed an alignment system based on circular zone plates which uses one mask target and one wafer target, whose focal lengths vary by the mask-to-wafer separation, so that both have foci in the same plane (Figure 3.17). The relative positions of the foci then give the mask-to-wafer misalignment.

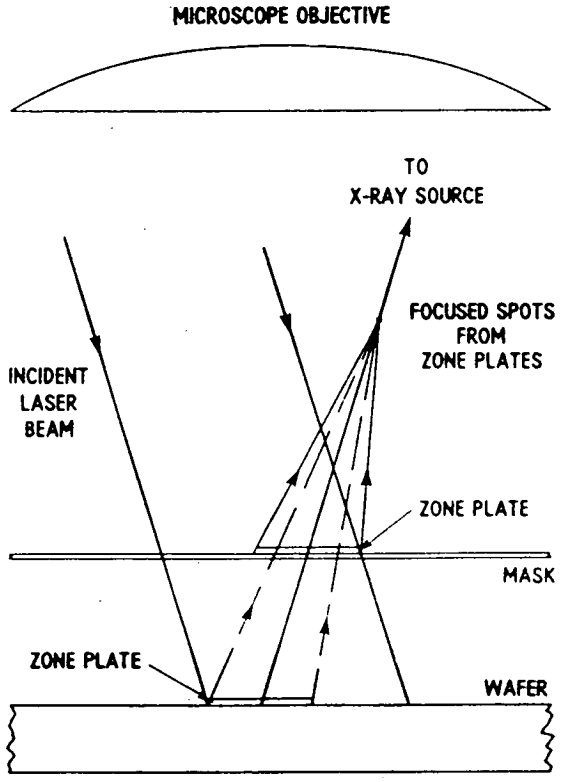

Figure 3.17: Mask and wafer alignment zone plates used in proximity x-ray printing, from reference [57].

Fay and Novak [58] have reported an extension to Feldman's system, in which three sets of wafer and mask targets are used, with each set consisting of two mask zone plates and one wafer zone plate, in an arrangement shown in Figure 3.18. Correct alignment in such a case is obtained when the focussed spot from the wafer target lies exactly half way between the focussed spots from the mask targets. By using three sets of marks, each of which yields an x and y misalignment value (ie. a total of six measured parameters) it is possible to evaluate the x, y and θ misalignments for the whole wafer, along with the Z_1, Z_2, and

Z_3 displacements which specify wafer or mask tilt and separation. Compensation for linear expansion is accomplished automatically using such a system by adjustment of the wafer/mask gap. An alignment accuracy of 0.25μm at 3σ is claimed using this method.

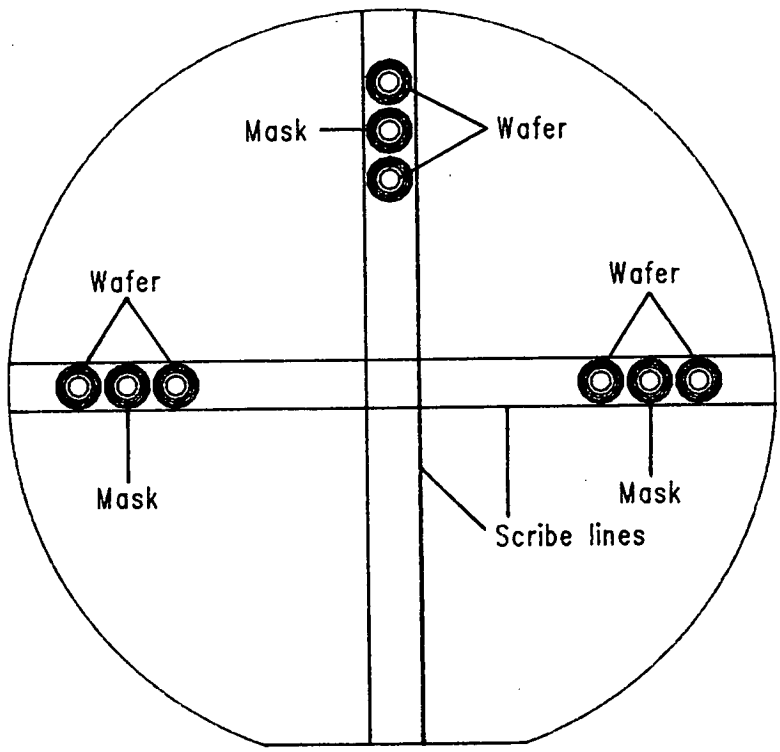

Figure 3.18: Three point alignment used to correct x, y, θ, wafer tilt, and mask/wafer gap, from reference [58].

Fay and Trotel proposed a scheme in which linear FZT's can be used in conjunction with linear gratings (Figure 3.19), giving a positioning accuracy of 0.1μm [61]. In its original form the linear zone plate produced a line image on the wafer, which was scanned across a simple line on the wafer using a rocking mirror, with the maximum in reflected intensity from the wafer indicating the relative misalignment. The structure of the wafer mark was later changed to a linear grating, with the the alignment signal being picked up in the first diffracted order from this grating. The purpose of this was to eliminate the contribution of the zero diffracted order from the FZT to the alignment signal.

Nelson et al. [60] have extended the method of Fay and Trotel to provide an alignment scheme with a precision of 0.01μm (precision is the smallest measurable

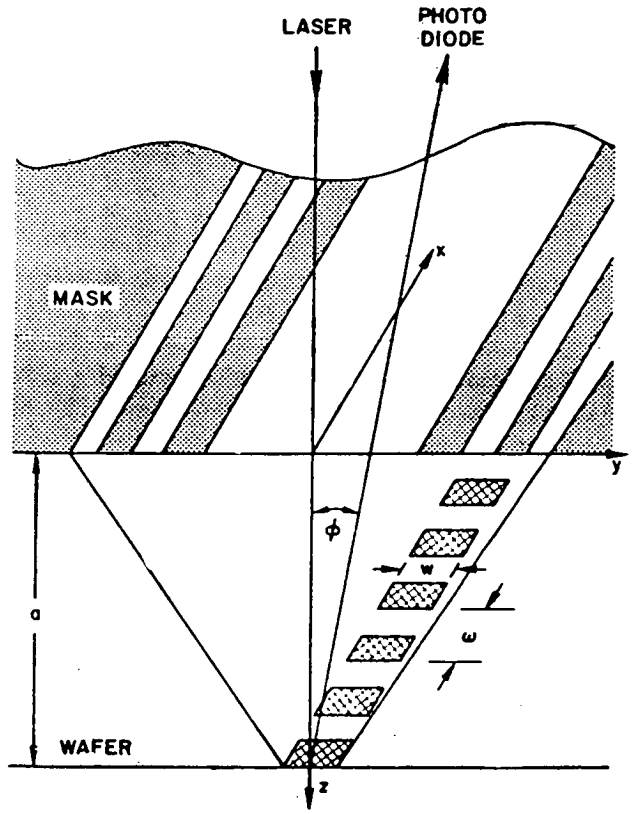

Figure 3.19: Combination of linear FZT and linear grating for alignment, from reference [61].

offset in this case, and is not the same as alignment accuracy). An extremely wide capture range is made possible by the use of a linear grating such as that illustrated in Figure 3.20. The alignment range depends largely on the width of the base triangle (b), while the positioning accuracy is determined by the width (w) of the upper section of the mark. In fact a non-linear array of such gratings is used to further increase the alignment range, making use of the fact that periodic structures give periodic alignment signals, while non-repeating structures give unique alignment positions.

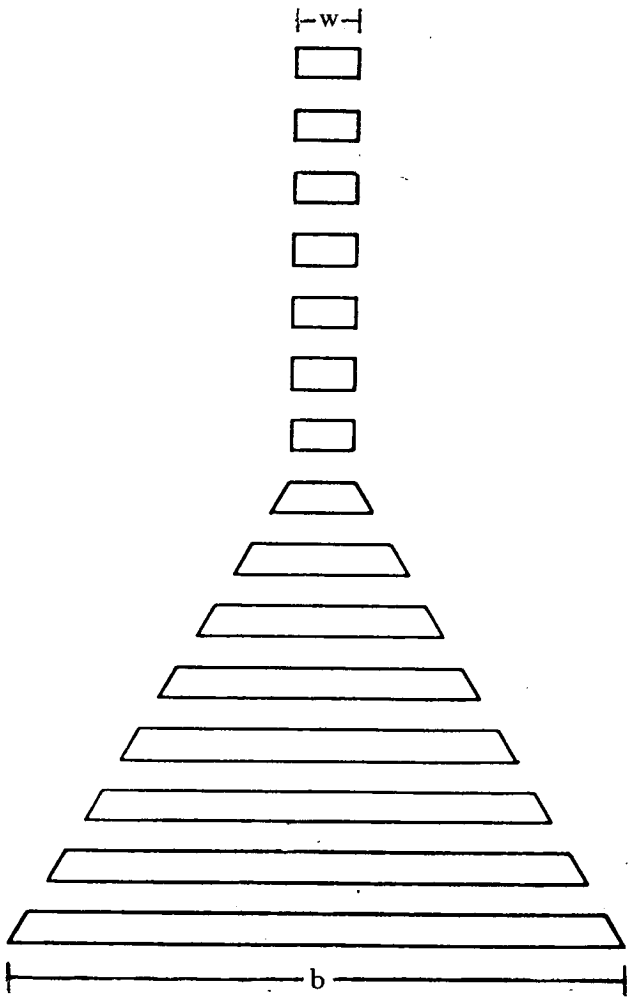

Figure 3.20: Linear grating for use in conjunction with linear FZT. The capture range is determined by b, and the accuracy by w, from reference [60].

3.2.3 Nikon.

Figure 3.21 shows a schematic of the alignment system used in the Nikon wafer stepper (NSR) [66]. A beam from a He-Ne laser is formed into an elongated spot on the wafer, with the stage being moved so that a grating on the wafer is scanned through the spot (Figure 3.22). The diffraction signal from the grating is then spatially filtered, to remove the zero diffracted order, and focussed onto a photo-detector. The stage position at which the maximum diffracted intensity is collected is recorded using a laser interferometer mounted beneath the stage. After alignment mark detection, the stage is blind stepped to the correct position, so that the total alignment accuracy σ_{TOT} is given by:

$$\sigma_{TOT} = \left(\sigma_A^2 + \sigma_S^2\right)^{\frac{1}{2}} \tag{3.1}$$

where σ_A is the alignment repeatability and σ_S is the stage position repeatability (this formula holds in general for any system in which the alignment and exposure positions differ). In practice a number of gratings are used where rough substrates are involved, to improve the signal-to-noise ratio.

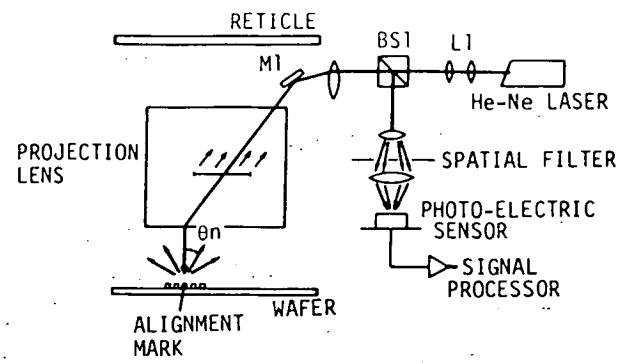

Figure 3.21: Schematic of Nikon NSR grating alignment system, from reference [66].

The alignment strategy on the Nikon can be summarised as follows :

1. Mechanical prealignment.
2. Off-axis global alignment (y and θ).
3. On-axis global alignment (x, y and θ).

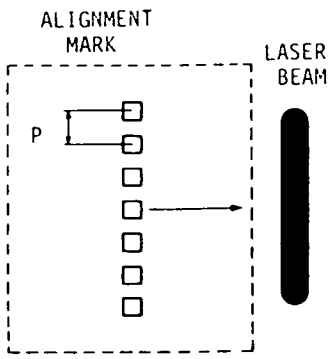

Figure 3.22: Nikon alignment mark and scanning beam, from reference [66].

4. Blind-step for maximum throughput.

   or

   Blind-step with scaling compensation (linear expansion is measured and compensated by the auto-align system).

   or

   Enhanced global alignment. The x and y co-ordinates of certain specified sites are measured before exposure. Blind step then follows, with localised compensation for measured data.

   or

   Full die-by-die alignment.

Enhanced global alignment appears to give optimised 'good die' throughput, having registration accuracy equal to that of the full die-by-die mode, with ~20% better throughput [66]. Alignment accuracy for die-by-die or enhanced global is quoted at $\pm 0.15\mu$m (2σ) [49].

3.2.4 Philips/ASM.

The Philips/ASM alignment system uses diffraction from a square wave grating as an alignment signal. Collimated light from a He-Ne laser is shone onto the wafer, and the diffracted light is then passed through a spatial filter which transmits only the +1 and -1 diffracted orders (see Figure 3.23). In this way the alignment signal is always sinusoidal in form, regardless of the shape of the

grating grooves themselves. The signal is then passed through a grating in the mask plane, which has the same period as the alignment signal at the mask. The correctly aligned position corresponds to maximum transmission through the mask grating [63].

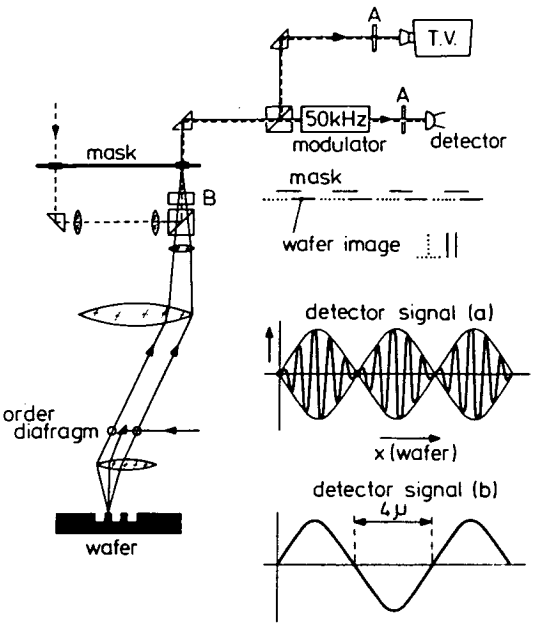

Figure 3.23: Philips diffraction alignment system, showing position of spatial filter (order diaphragm), from reference [63].

The alignment strategy can be summarised as follows :

1. Mechanical wafer prealignment.

2. Optical wafer prealignment.

3. Reticle alignment in x, y, and θ, to the x-axis motion of the stage, using one wafer mark as a reference. Magnification correction is also applied at this stage by measuring the separation of the reticle markers. The height of the reticle stage is then adjusted to make this distance exactly 96mm.

4. On-axis global wafer alignment (x, y, and θ). The distance between the wafer markers is measured using the interferometer stage, and a scaling factor introduced to the step size to compensate for this.

5. Repeat steps 3 and 4.

6. Blind step for maximum throughput.

   or

   Die-by-die alignment for local correction of wafer distortion.

The manufacturers claim an alignment accuracy of 0.15μm (mean + 3σ, in x and y) in blind step mode. Spatial filtering is claimed to improve image contrast, by elimination of the zero order of the wafer grating, as well as reducing the influence distortion from higher orders on the signal [67]. Perhaps surprisingly, the manufacturers claim slightly worse performance for die-by-die alignment than for blind stepping (the die-by-die specification is 0.2μm, mean + 3σ). This is due to the fact that the die-by-die marker is small compared with the global marker, and therefore gives a poorer alignment signal [67].

3.2.5 Additional Linear Grating Systems.

In 1977 Flanders, Smith, and Austin [68] proposed an alignment system based on diffraction from linear gratings, such as those depicted in Figure 3.24. Collimated coherent light is shone on to a regular mask grating, and diffracted onto a grating of the same period on the wafer. When perfect alignment is attained the intensity in the +1 and -1 diffracted orders will be equal. The alignment signal is simply the difference in the output of photo-detectors 1 and 2.

One major problem with this type of scheme is repetition of the alignment signal, which limits the capture range to around one half of the grating period. Lyszczarz et al. [62] have proposed a vernier type scheme in which gratings of slightly different period (P_1 and P_2) are used. The alignment capture range is thus extended to $P_1P_2/(P_1 - P_2)$

Kleinknecht [64] has also extended the method of Flanders, Smith, and Austin, designing a scheme whose alignment range is limited only by the size of the alignment target. The structure of the alignment target is shown in Figure 3.25 along with position of the diffracted spots produced by such a target. When an opaque square stop is positioned over the target, the relative alignment of the square and target can be determined from the relative intensities diffracted into the various spots. Using this system an accuracy of 0.1μm has been achieved, with a capture range of $\pm$450μm.

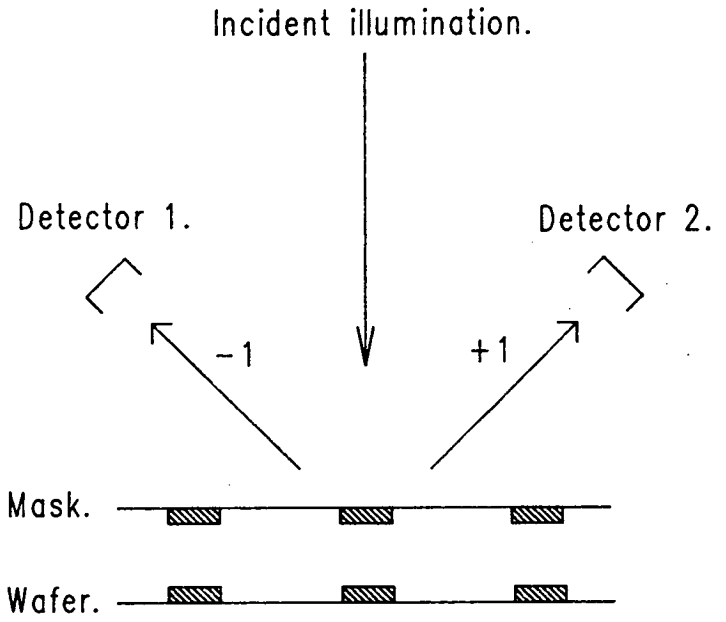

Figure 3.24: Basic grating alignment system due to Flanders, Smith, and Austin, from reference [68].

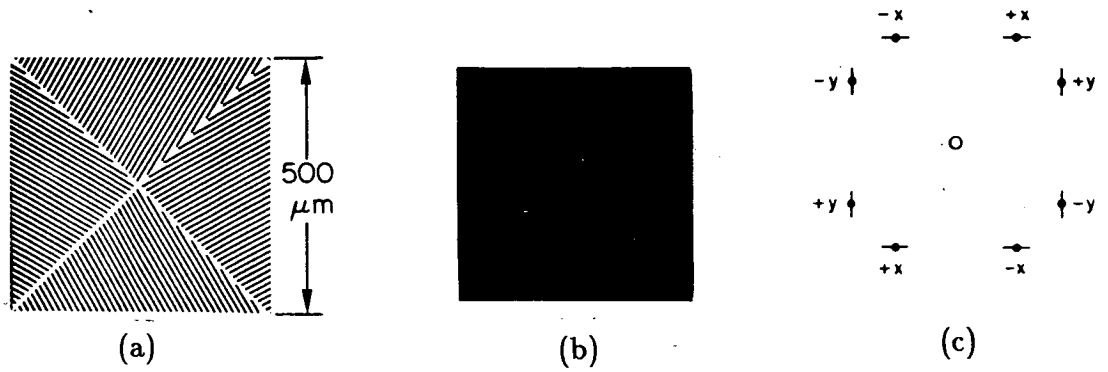

Figure 3.25: Alignment system due to Kleinknecht, from reference [64]. (a) wafer mark, (b) reticle mark (opaque square stop), and (c) position of first order diffracted spots from wafer target.

3.3 Scattered Light Systems.

Scattered light systems rely on the detection of the dark field image of a wafer feature for alignment. Figure 3.26 shows a simplified schematic of one such system, illustrating the principle of dark field alignment. Normal illumination of the wafer is obstructed by means of a stop placed in the aperture of the main imaging lens. If a detector is placed behind the stop, then light reflected specularly from a flat wafer surface is prevented from reaching it. When the incident light encounters a wafer feature, however, some light will be scattered in the direction of the detector, thereby locating the position of the edge of the feature. The accuracy of this method is limited essentially only by background noise caused by wafer surface roughness [69].

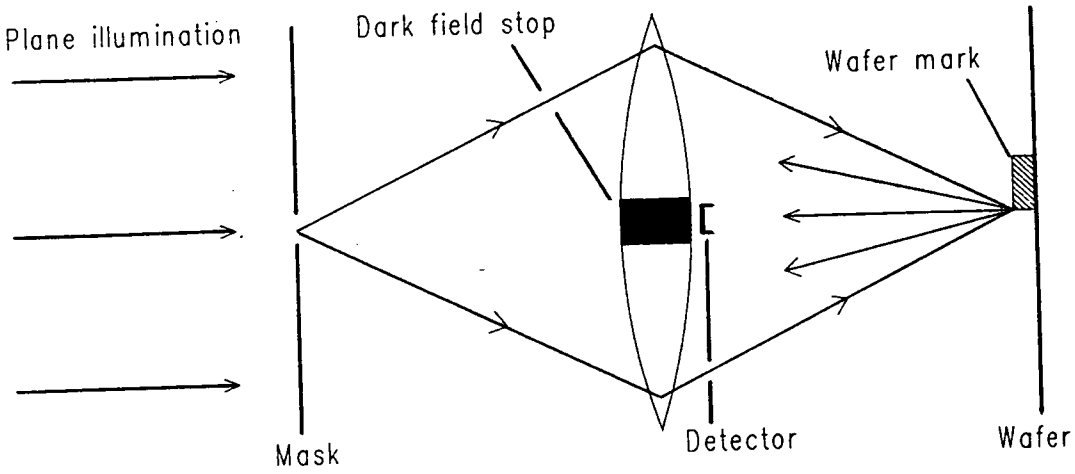

Figure 3.26: Generalised dark field alignment system.

3.3.1 Ultratech.

The Ultratech stepper uses precisely the type of dark field alignment system described above, although the optics are reflective in nature. Figure 3.27 depicts the main projection optics, as well as the alignment optics, with the centre hole in the main mirror taking the place of the stop in the centre of the lens in Figure 3.26. Only light scattered from the edges of the wafer mark is returned through the centre hole to be picked up by the photo-multiplier tube. Spatial filtering is employed to match the detector aperture to the alignment mark, thus minimising the contribution of scattering due to surface roughness [70].

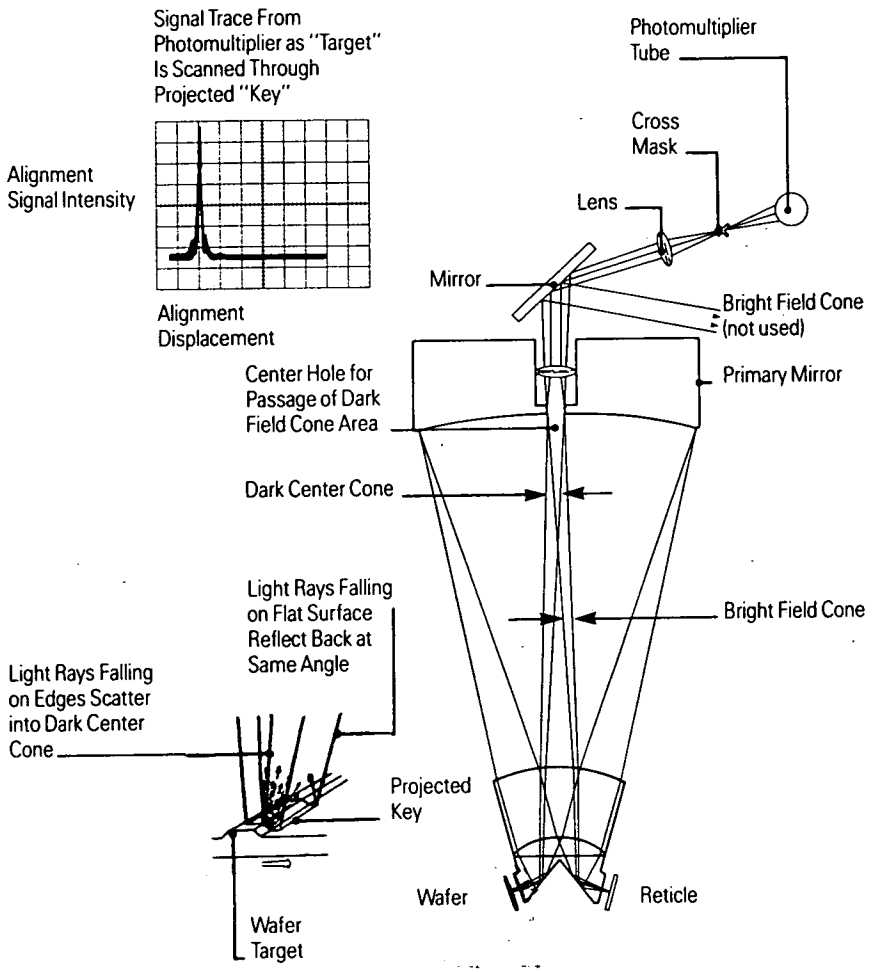

Figure 3.27: Ultratech projection/alignment system, from reference [70].

The alignment strategy can be summarised as follows :

1. Reticle alignment in x only, as well as measurement of reticle offsets in y and θ.

2. Mechanical prealignment (only used if blind stepping is required).

3. Coarse on-axis global alignment in x, y and θ, using large wafer alignment targets.

4. Fine θ alignment using small wafer alignment targets.

5. Die-by-die alignment in x and y.

The advantages of the Ultratech system include the fact that it is hardware based, resulting in high throughput (the manufacturers claim 40 4" wafers per hour, die-by-die). In addition, the dark field alignment system eliminates the need for high contrast (bright field) in the wafer targets, as well as susceptibility to diffraction fringing. Any asymmetry present in the wafer target profile, however, will lead to an error in the detected edge position, and hence to an alignment offset. A total overlay accuracy of $\pm 0.23\mu$m (2σ) is claimed, when matching machines.

3.3.2 Canon.

The Canon auto-align system is slightly different from the Ultratech in that the dark field stop is mounted in the detector optics, rather than in the projection optics, as illustrated in Figure 3.28. A 442nm wavelength He-Cd laser is used in conjunction with a rotating polygon mirror to continuously scan the combined mask and wafer mark, the structure of which is shown in Figure 3.29. The scanned beam is kept perpendicular to the wafer by having the point of reflection from the rotating mirror in the front focal plane of the main projection lens. With a projection lens designed to operate at 436nm, Canon claim identical lens performance at exposure and alignment wavelengths [71]. Two alignment marks are used per site to align in x, y and θ. The system is used on both reflective (whole wafer) and refractive systems, giving alignment accuracy of within 0.25μm (σ) on all levels of a MOS process [72], with an alignment speed of less than 0.4 seconds per site.

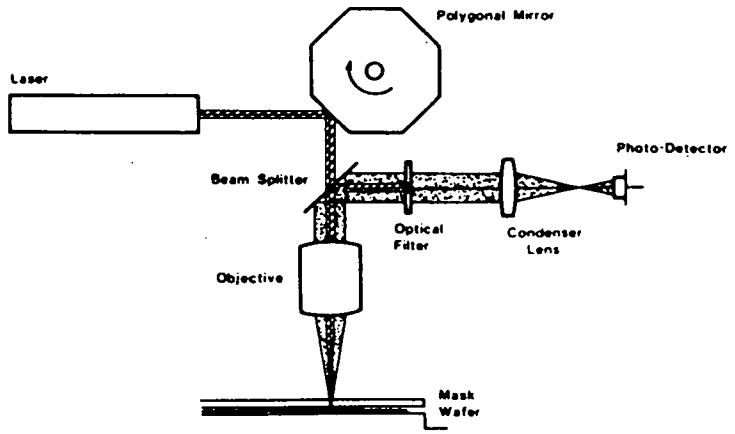

Figure 3.28: Canon alignment system, showing position of dark field stop (optical filter), from reference [72]

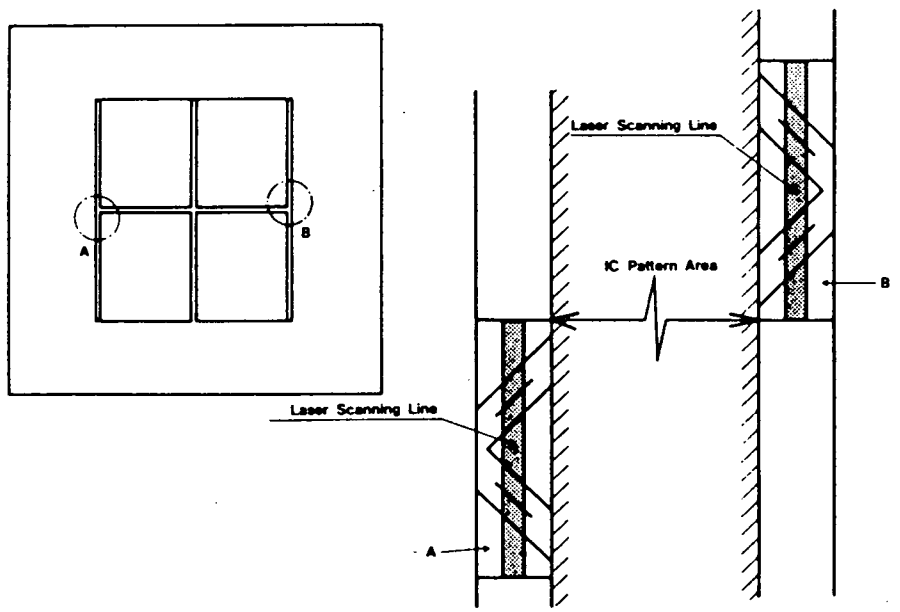

Figure 3.29: Structure of Canon alignment marks, from reference [72].

3.3.3 Additional Scattered Light Systems.

Bobroff et al. [73] have developed a similar alignment system for use in an x-ray aligner, in which both mask and wafer are independently aligned to a reference reticle (Figure 3.30). In this system the problem of scattering is alleviated by making the entrance aperture of the photo-detector smaller than the effective aperture of the dark field stop, thus reducing the contribution of surface roughness to the alignment signal.

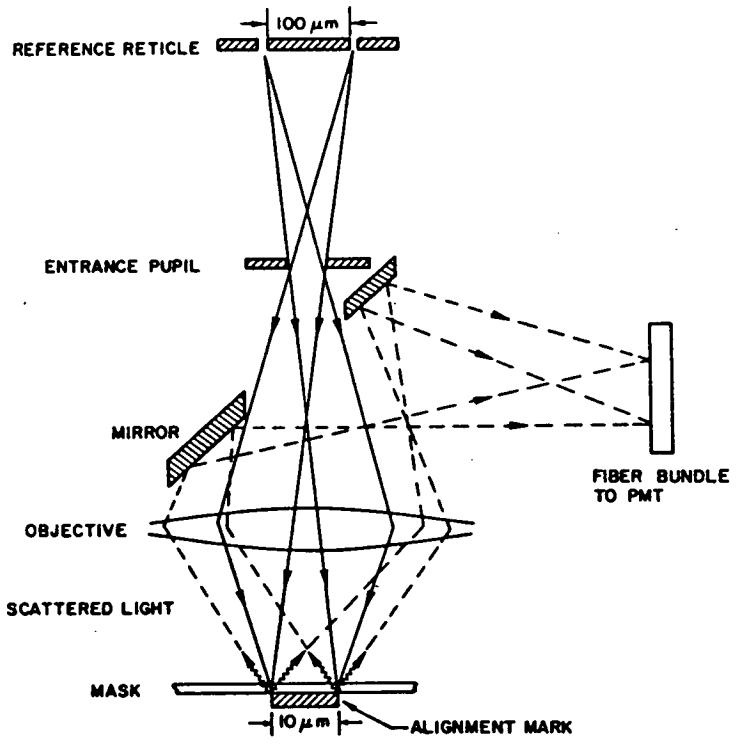

Figure 3.30: Dark field alignment system for x-ray exposure, from reference [73].

3.4 Conclusions.

In general all of the above systems have their good and bad points. There are two main criteria which should be considered when talking about the merits of these systems in production: accuracy and flexibility. It is quite possible, under certain circumstances, to depreciate the value of an accurate alignment system by, for example, not providing blind step option. In other words, the alignment

strategy, as well as the alignment accuracy, is important in determining the usefulness of any machine.

While in general the dark field and diffraction based systems are less susceptible to diffraction fringes, it is possible, using the flexibility of a software based light field system, to use special signal processing to reduce the effect of the fringes on the alignment performance, thus improving the overlay specification of light field systems by improving reliability considerably. This type of signal processing will be treated in more detail in the next Chapter 4.

Overlay accuracy will continue to be the dominant factor in optical lithography over the next ten years. While the way ahead for improving resolution by a factor of almost two in the next year or so is clear (mainly by reduction of wavelength using excimer laser sources, and increasing numerical aperture), the same cannot be said for overlay accuracy. While alignment accuracies of 10nm [74] and 5nm [75] have been reported, mask writing and stage positioning errors still make it difficult to reduce the total overlay error to less than 0.1μm 3σ.

Chapter 4

The Optimetrix Alignment System.

4.1 Optimetrix Software.

Before embarking on a discussion of the Optimetrix alignment system, it is important to understand a little about the operating system mounted on the machine itself.

In the highest level command mode, the parameters for basic commands such as BEG STEP (**BEG**in **STEP**ping wafer) and BEG RAL (**BEG**in **R**eticle **AL**ignment) are set up in job files such as that shown in Table 4.1. Job file parameters include the specification of items such as stepping pattern, size of die, exposure time, lamp spectral filtering, alignment mode (blind step or die-by-die) and alignment mark position.

In addition to this command mode there are two diagnostic modes which the operator should know about. These are COM DEB (**COM**mence **DEB**ugging) and DIA CAMERA (**DIA**gnostic **CAMERA**). In both of these modes the operator is able to alter the contents of specific system memory locations.

COM DEB is used to specify many machine-dependent parameters (parameters which do not vary from batch to batch), such as lens-to-stage separation for reticle alignment, or the automatic alignment correction limit (the maximum allowable alignment correction). In addition it can also be used to specify some job-dependent parameters, such as the alignment tolerance (how tight the alignment has to be on that particular layer) and alignment offset (used to correct any constant error in the alignment position). In fact COM DEB gives direct access to all user-definable memory locations.

DIA CAMERA is used to specify only job-dependent parameters, all of which

```
JOB = EXAMPLE      DIE X DIE ALIGN
WAFER MM DIAMETER =     76
Y U SHIFT, SIZE, STEPS & SYMMETRICAL =       0.00,   6000.00,  12, YES
X U SHIFT, SIZE, & STEPS =    0.0,   6000.00,  4@ 10,  8,  4,
EXPOSURE MILLISECONDS =    625  AT LINE = G+
FAST FOCUS CONTROL and FOCUS U SHIFT =  YES,    -1.0
EXPOSE STEPS
NORMAL   FRAMING U LOWER, UPPER, LEFT, RIGHT =  -2700,  2900, -2700,  2900
DROP OUT FRAMING U LOWER, UPPER, LEFT, RIGHT =      0,     0,     0,     0
DROP IN  FRAMING U LOWER, UPPER, LEFT, RIGHT =      0,     0,     0,     0
STEP IMAGE ALIGN LEFT SPLIT & CONTROL SINGLE & # =   NO, YES,    -1
STEP IMAGE MICROSCOPE U POSITION Y, XL & XR =  -1771, -6544,  2594
EPI Y, X SHIFT =       0.00,     0.00
LEVEL ON & # =  YES,     2
WAFER U Y, X SHIFT & SEPARATION  =      0.00,      0.00, 68000.00
WAFER ROTATION ASEC =      0.0
WAFER ALIGN: AUTO, AUXILIARY, plus ON-AXIS STEP, & RECHECK = NO YES YES  0
AUXILIARY ALIGN: TIME & plus ON-AXIS STEP =      0.0, YES
RETICLE U Y, X SHIFT & SEPARATION =    100.00,    250.00, 10300.00
RETICLE LOAD Y-LEFT, Y-RIGHT, X =      0.0,     0.0,     0.0
START with: RET #, WAFER LOAD, WAFER ALIGN, & STEP # =  0,  NO,  NO,    0
NEXT JOB =
 OK, Y-N ?
```

Table 4.1: Example job file for die-by-die alignment.

are concerned with the alignment process. Table 4.2 shows a list of all DIA CAMERA parameters. A complete description of the meaning of each parameter is given in reference [76]. A brief discussion of the relevant parameters is given below:

- CAM = -1 implies EXPERT mode. DIA CAMERA parameters (video thresholds, offset and gain) are automatically updated to optimise the alignment process.

  CAM = 0 implies no automatic update.

- REA : number of readings which are averaged before alignment begins.

- SPA : spatial analysis correlation coefficient (see Section 4.2).

- MWI : width of reticle window in pixels (1 pixel = 0.217μm).

- MTH : auto-correlation threshold for reticle feature, above which local gradient search begins (see Section 4.2).

- WTH : auto-correlation threshold for wafer feature, above which local gradient search begins (see Section 4.2).

- LNS : number of lines of video data which are averaged before alignment begins.

- WMI : minimum wafer mark width in pixels.

- WMX : maximum wafer mark width in pixels.

- O : video offset. Additive scaling constant used to bring video data within expected range (adjusted by EXPERT).

- G : video gain. Multiplicative scaling constant used to bring video data within expected range (adjusted by EXPERT).

- DIA : diagnostic mode. Useful for finding the cause of auto-align failures.

DIA CAMERA accesses memory locations which are also accessible via COM DEB. The DIA CAMERA command, however, allows quick and easy access to those parameters which are pertinent to the alignment process.

CAM =	0
REA =	1
SPA =	-21
MWI =	130
MTH =	500
WTH =	700
LNS =	16
WMI =	40
WMX =	42
0 =	128
G =	182
W =	8
DIA =	0
R =	627
M =	196
N =	16
T =	191
L =	255
B =	1

Table 4.2: List of DIA CAMERA parameters with typical values.

4.2 The Alignment Algorithm.

The machine software calculates the position of the centre of a wafer mark by measuring the amount of mirror symmetry in the video signal at any given point, using an algorithm which was designed to be sensitive to mirror symmetrical aspects of the data. While in the ideal case, reticle and wafer features would exhibit perfect mirror symmetry, in practice this is never achieved, and it is the job of the feature location algorithm to establish the point of maximum symmetry.

The measurement of the degree of mirror symmetry is accomplished by the use of an auto-correlation algorithm which is applied to the first derivative of the video data, according to the following equation:

$$A(p,w) = -\sum_{n=-\mathrm{SPA}/2}^{n=+\mathrm{SPA}/2} [V'(p-n+w/2) \times V'(p+n-w/2)] \qquad (4.1)$$

where p is the pixel number, w is a width parameter, SPA is the spatial analysis correlation coefficient, and $V'(p)$ is the first derivative of the pixel data at pixel p. The first derivative is used for two reasons:

- It eliminates monotonic variations in reflectivity due to changes in film thickness, thus making the algorithm insensitive to process variations.
- The first derivative correlates strongly with first order variations in intensity, and thus to feature presence.

Transfer of video data from the stepper to the VAX was achieved by connecting the output RS232 port of the stepper directly to an input line on the VAX. The appropriate alignment target could then be displayed on the TV screen of the stepper, and the pixel data stored by issuing the command "DIA VID SAV". This data could subsequently be dumped to the VAX via the RS232 port, and could then be analysed off-line, giving the possibility of plotting the raw data, and calculating and plotting the first derivative data and correlation data.

Figure 4.1 shows an example of video data taken from reticle alignment to the stage target (left hand side of chevron only), with the first derivative superimposed upon the raw data. A plan view of the target is included for reference. As can be seen from the figure, the derivative data of the wafer stage target contains one large positive and one large negative peak (the outer peaks correspond to the edges of the reticle window). The auto-correlation function folds the derivative data upon itself about p, and performs a convolution sum on an interval of length

SPA. For the data in Figure 4.1, this produces a peak in $A(p, w)$ at a position which corresponds to p half way between the positive and negative peaks, with w equal to the separation of the peaks. Thus the correlation function evaluates the position of the centre of the mark, as well as the mark width. It should be noted that nowhere are the positions of the edges of the mark themselves calculated.

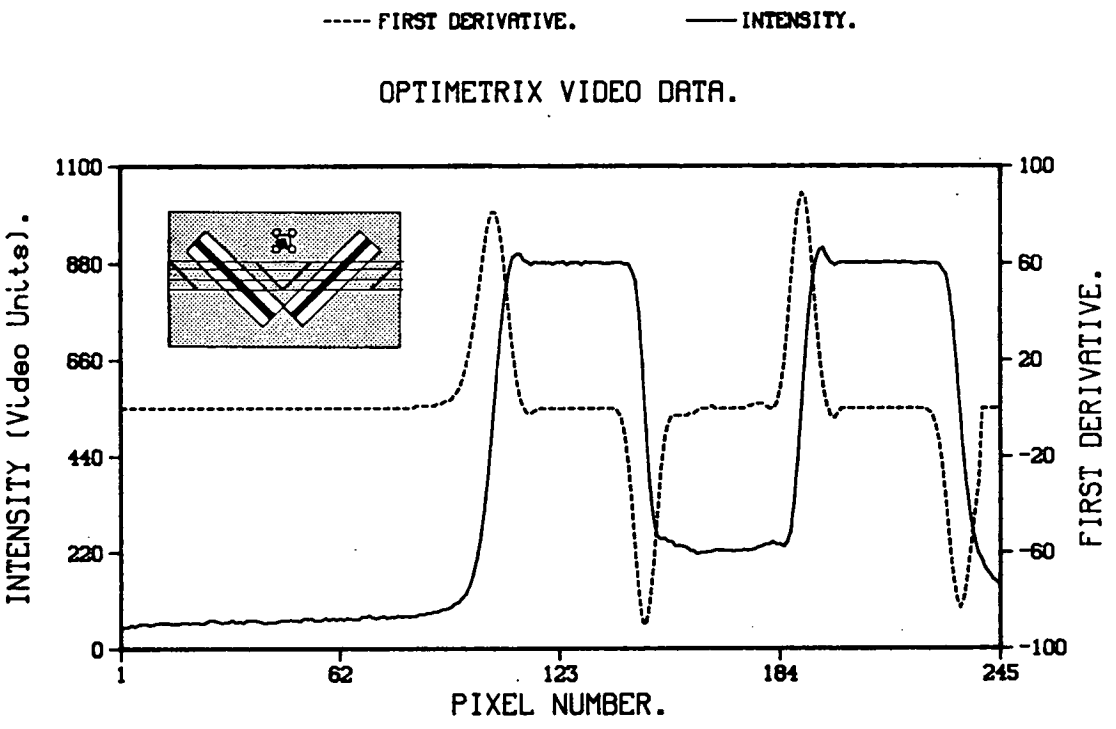

Figure 4.1: Ideal video data, with the first derivative superimposed. Also shown is a plan view of the target.

Figure 4.2 shows a surface plot of the auto-correlation function derived from the reticle alignment data of Figure 4.1 (the correlation function was computed off-line). In searching for a peak in the auto-correlation, the machine software explores all directions in (p, w) space, until the WTH (wafer threshold) parameter of DIA CAMERA is exceeded; it then follows a local gradient to a peak.

For the type of correlation function shown in Figure 4.2, the local gradient peak location algorithm will find the correct centre of the wafer mark every time (since the function has a single, well defined peak). Figure 4.3, however, shows the type of alignment signal which is typically found in a production process (in this case 1.1μm of resist covering a 6.0μm wide active area, with gate oxide, surrounded by field oxide). A contour plot of the corresponding correlation function for this data is shown in Figure 4.4. Instead of possessing a single peak as in the case of reticle alignment, the correlation function now has multiple

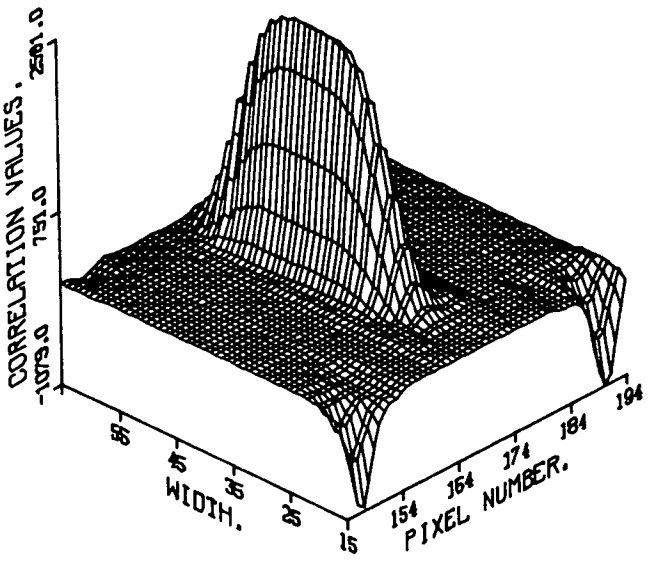

Figure 4.2: Surface plot of the auto-correlation function of the data in Figure 4.1

peaks (in fact, for every positive/negative pair in the first derivative data, there will be a corresponding positive peak in the auto-correlation). In this case a total of three peaks can be seen in the centre of the correlation function window.

This is the basis of the alignment mark fringe problem; the more fringes which are present in the mark, the more peaks will be present in the auto-correlation, hence increasing the likelihood that the local gradient search will latch on to the wrong peak.

The peaks themselves are in general separated by at least four pixels, leading to alignment errors of 0.9μm and larger. While it is true that there are other sources of error to contend with (stage accuracy, marker asymmetry etc...), these errors will be small by comparison (generally 0.25μm or less). Thus the fringe problem is the most important one to be addressed if reflected light alignment systems are to maintain a position in the market in the coming years (while it is true that other machines use different alignment algorithms, they are still confused by the presence of fringes). If this problem is not solved, these systems will probably never achieve the performance in production that their manufacturers claim for them.

Errors of up to 3.0μm were observed occasionally on the Optimetrix (due to

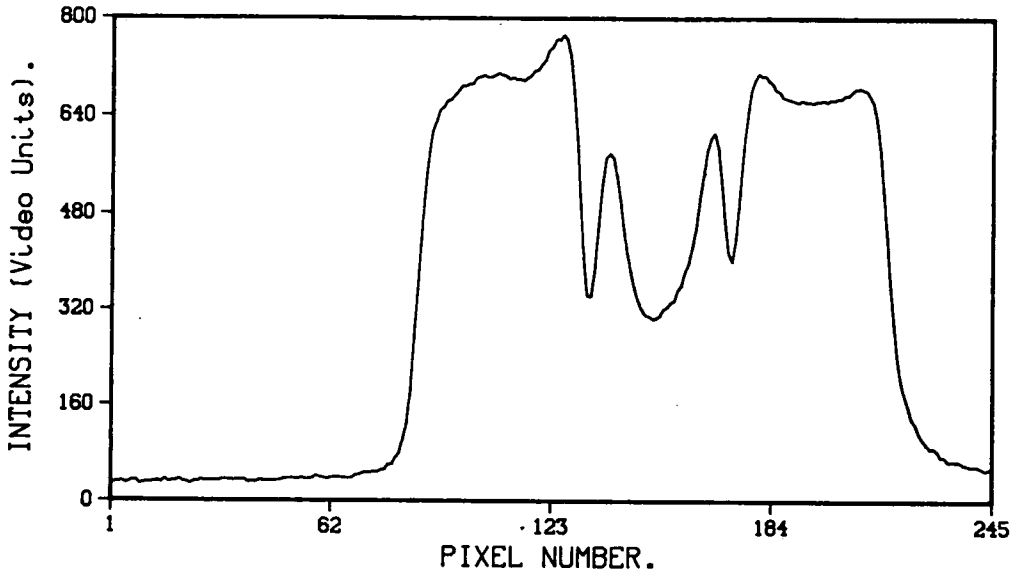

Figure 4.3: Typical alignment signal found in-process; aligning buried contacts to active area.

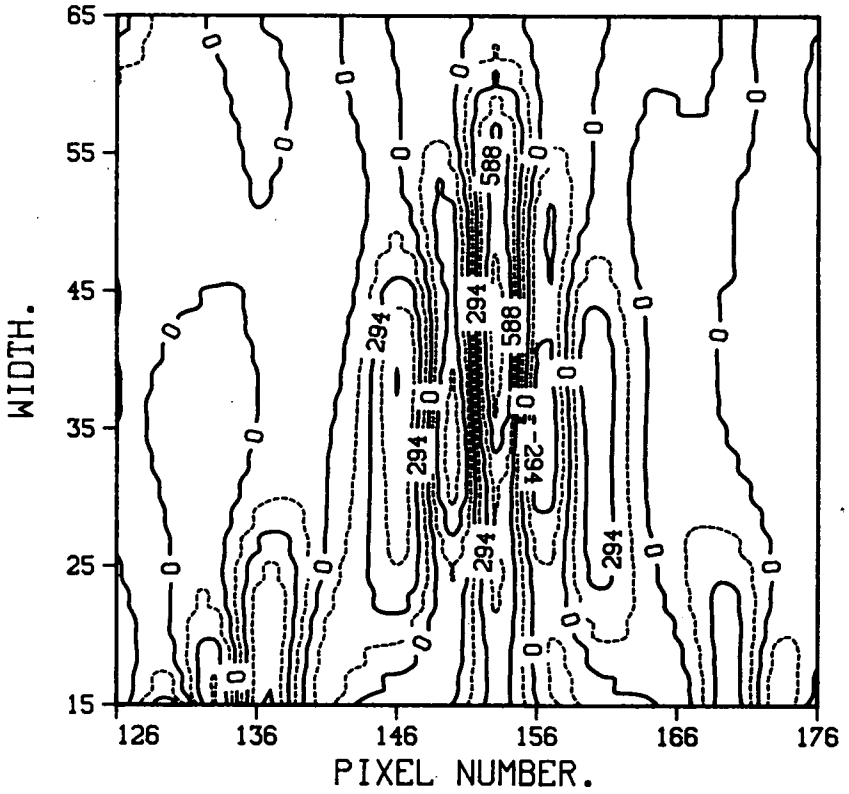

Figure 4.4: Contour plot of correlation function obtained from raw data in Figure 4.3

the machine thinking that one edge of the alignment mark was actually the centre of the mark). Clearly there is little point in trying to improve the performance of the machine by 0.05μm under ideal conditions, when under normal operating conditions the results can be out by an order of magnitude from the ideal case. While solving this problem will not guarantee perfect alignment, due to the profile asymmetry and stage position errors mentioned above, it will at least give these machines a chance to live up to their specification, something which has been very difficult to achieve until now.

4.3 Program OASIS.

In order to be able to study the alignment process more thoroughly, and to determine effective ways in which the problem of fringing could be reduced, a program was written on the EMF VAX 11/750 computer which simulated the steps performed by the machine software to perform alignment. The program, called OASIS (Optimetrix Alignment SImulation Software), was written in standard Fortran 77 for maximum portability. It was intended at the outset that OASIS would be able to calculate the auto-correlation function and look for peaks in the normal way (by following a local gradient), as well as having the capability to vary certain important DIA CAMERA parameters, and study the effect of these variations on the alignment process. In addition to this, it was hoped to assess the effect of various types of digital filtering on the video signal and auto-correlation. Filtering used by the program includes simple smoothing and low pass filtering (filtering out of any spatial frequencies above a certain specified level). It was also hoped that it would prove possible to assess the performance of other well known correlation algorithms, including the modulus difference, squared difference, and cross-correlations [77].

There are three distinct approaches to improving alignment performance and the program was written with these in mind :

- Improving accuracy and reliability. Improved reliability may be achieved through spatial filtering and reduction of the number of peaks in the correlation function. Reducing the chance of the alignment system latching onto the wrong peak will also improve the long term accuracy of the alignment system.

- Improving speed. In particular it was hoped that the modulus and squared difference algorithms could prove effective in increasing alignment speed.

- Optimisation with respect to existing DIA CAMERA parameters.

4.3.1 OASIS Structure.

The following is a brief outline of the structure of OASIS, showing some of its capabilities:

Input Section.

1. Terminal input (compulsory). Name of input data file, and whether analysis should be performed on left, right, or both sets of data.

2. Subroutine RDOPTN (compulsory). Reads a file named 'options', which specifies DIA CAMERA parameters, type of correlation processing required, and whether or not the smoothing and spatial filtering algorithms are implemented.

3. Subroutine RDDATA (compulsory). Reads in the video data from the input file.

Analysis Section.

1. Subroutine NEWSMO (optional). The data can be smoothed by fitting a polynomial (of degree 2, 3, 4 or 5) over a specific range. For example, a nine point, fourth order fit would take each pixel and evaluate the best fitting fourth order polynomial (using a least squares algorithm) over a range of nine points (pixel-4, ..., pixel, ..., pixel+4) [78]. The smoothed intensity would be the value of the polynomial at that point. This smoothing is useful for removing only high frequency noise ($\geq$1.0 cycles/μm). In general, the lower the order of the polynomial used, the larger the effect of the smoothing.

2. Subroutine LOPASS (optional). Computes an output data set $O(p)$, from the input data set $I(p)$, using the following algorithm:

$$O(p) = \frac{f_s}{2f_o} \sum_{i=-NT}^{i=+NT} I(i+p)\,\mathrm{sinc}\,(\Theta_o i) \tag{4.2}$$

where f_s =1/pixel size, and $\Theta_o = 2\pi f_o/f_s$. For a large number of taps, this approximates an ideal, low-pass filter, with a sharp cut-off at a frequency f_o. The parameter NT is defined in the 'options' file ($2NT + 1$ = number

of taps in the filter). The desired cut-off frequency f_o is prompted for at the terminal.

3. Subroutine FSTDRV (compulsory). Calculate the first derivative of the intensity data. The following algorithm is used:

$$V'(p) = \frac{1}{16} \sum_{i=-4}^{i=+4} i \times V(p) \tag{4.3}$$

At each pixel this algorithm fits a second order polynomial to a range of nine points ($\pm$ 4 from the centre pixel), and returns the value of the first derivative of the fitted polynomial at that point.

4. Subroutine HIPXAV (compulsory). Evaluates the average of the eight largest intensity values in the raw data, and stores this value in the variable AVMAX.

5. Subroutine LFTEDG (compulsory). Evaluates the position of the left hand reticle edge. This subroutine scans through the raw data, until a value exceeding AVMAX/2 is found. It then looks for a peak in the first derivative data within $\pm$ 10 pixels of this point. If no peak is found, it finds the first point at which the intensity exceeds AVMAX/4, and looks for a peak within $\pm$ 10 pixels of this point. If this also fails, the value $3\times$AVMAX/8 is used as a start point. If a peak is still not found, the program terminates. This subroutine has only one pixel accuracy.

6. Subroutine RHTEDG (compulsory). Same as LFTEDG, but for the right hand reticle edge.

7. Subroutine LFTDRV (compulsory). This subroutine interpolates around the position of the left hand reticle edge (as determined by LFTEDG), to find the position of the zero crossing of the second derivative, and uses this as the precise position of the left reticle edge ($\sim \frac{1}{5}$ pixel accuracy).

8. Subroutine RHTDRV. Same as LFTDRV, but for the right hand reticle edge ($\sim \frac{1}{5}$ pixel accuracy).

9. Perform correlation analysis. One of the four options (a)-(d) below must be specified:

   (a) Standard auto-correlation. Described earlier.

(b) Cross-correlation. At each position (p, w), the cross-correlation algorithm constructs a single positive Gaussian shaped peak, centred at $p-n+w/2$, and an identical negative peak at $p+n-w/2$. Figure 4.5 shows an example of the constructed array $G(p)$. The first derivative data is then convolved with this data to form the cross-correlation $C(p, w)$, as follows:

$$C(p,w) = \sum_{n=-\mathrm{SPA}/2}^{n=+\mathrm{SPA}/2} | V'(p-n+w/2) \times G(p-n+w/2) + V'(p+n-w/2) \times G(p+n-w/2)| \tag{4.4}$$

The cross-correlation was intended to be the correlation of the real alignment signal with the alignment signal in the ideal case, such as that shown in Figure 4.1. Since the constructed array $G(p)$ has only one positive and one negative peak, the cross-correlation should have fewer peaks than the standard auto-correlation. The modulus sign had to be included in the calculation, to ensure that only positive peaks would be produced by this algorithm (this has the adverse effect of doubling the number of peaks in the function, but is required, since the machine and OASIS only look for positive peaks). This algorithm is slow, due to the fact that the array $G(p)$ has to be recalculated at each position in (p, w) space.

(c) Modulus difference. Uses the following algorithm to construct the modulus difference array $M(p, w)$:

$$M(p,w) = \sum_{n=-\mathrm{SPA}/2}^{n=+\mathrm{SPA}/2} |V'(p-n+w/2) - V'(p+n-w/2)| \tag{4.5}$$

The modulus difference algorithm should be faster than the standard auto-correlation algorithm, since only SPA additions and SPA modulus operations have to be performed at each position in (p, w) space, rather than SPA multiplications. Although the difference in speed may not be apparent when the algorithm is implemented in a Fortran program, it was still studied, in order to determine whether or not the algorithm performed well enough to be used as a substitute for the auto-correlation process.

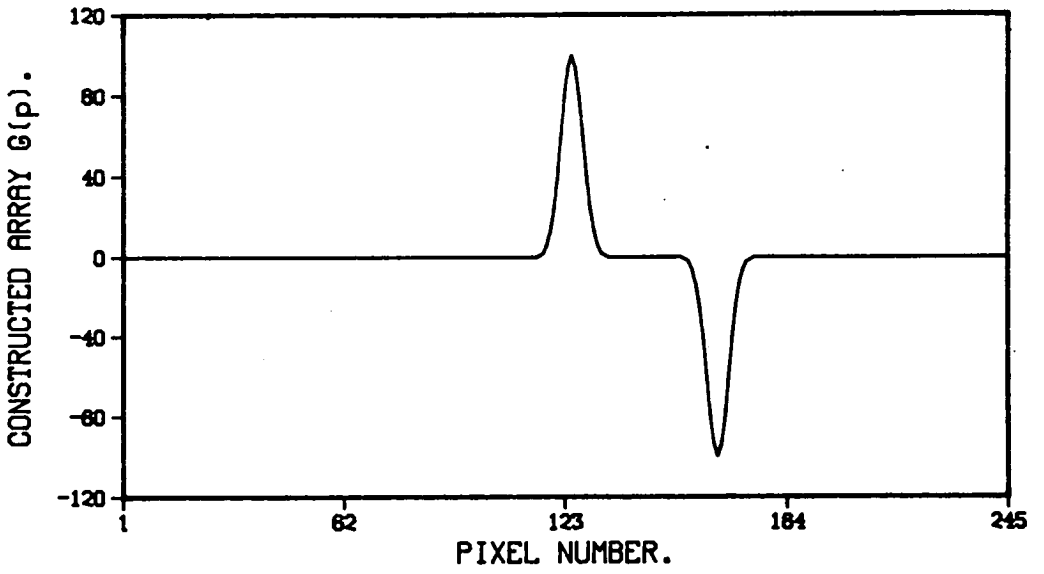

Figure 4.5: Constructed array $G(p)$ used in calculation of the cross-correlation.

(d) Squared difference. Uses the following algorithm to construct the squared difference array $S(p,w)$:

$$S(p,w) = \sum_{n=-\mathrm{SPA}/2}^{n=+\mathrm{SPA}/2} \left[V'\left(p-n+w/2\right) - V'\left(p+n-w/2\right)\right]^2 \qquad (4.6)$$

The squared difference algorithm was studied for the same reasons as the modulus difference.

Unless otherwise stated, a value for SPA of 13 was chosen when evaluating the correlation functions.

10. Subroutine CORSER (compulsory). This subroutine searches the correlation function ($A(p,w)$, $C(p,w)$, $M(p,w)$, or $S(p,w)$) for a peak. Uses a local gradient search, with the starting point of the search defined by the (p,w) window set up in the 'options' file (1 pixel resolution). With automatic windowing specified, the search begins in the p-direction at a distance half way between the left and right hand reticle edges.

11. Subroutine INTERP (compulsory). Evaluates the position of the zero crossing of the second derivative of the correlation function ($\frac{\partial^2 A(p,w)}{\partial^2 p}$). This value is used as the precise position of the correlation peak. Since the corre-

lation function varies much more slowly in w than it does in p, interpolation is not performed in the w-direction[†].

12. Subroutine MISALN (used only if the analysis has been performed on both halves of the chevron). Evaluates the misalignment in microns. This is given by the equations:

$$x_{err} = \frac{1}{2} \times 0.217 \times [(c_{wl} - c_{rl}) + (c_{wr} - c_{rr})] \tag{4.7}$$

$$y_{err} = \frac{1}{2} \times 0.217 \times [(c_{wl} - c_{rl}) - (c_{wr} - c_{rr})] \tag{4.8}$$

where x_{err} is the x-misalignment, y_{err} is the y-misalignment, 0.217 is the pixel size in microns, c_{rl} is the centre of the left hand reticle window, c_{wl} is the centre of the wafer target in the left hand window, c_{rr} is the centre of the right hand window, and c_{wr} is the centre of the wafer target in the right hand window.

Output Section.

1. Output positions of reticle window edges.

2. Output position of global maximum in the correlation function.

3. Output position of local maximum in the correlation function, as determined by the local gradient search.

4. Output position of the centre of the wafer mark, as determined by Subroutine INTERP.

5. If analysis has been performed on both halves of chevron, output the x and y-misalignments in microns.

6. Draw graphs of raw and derivative data if requested.

7. Draw contour plot of correlation function if requested.

[†] A rough estimate of the width of the mark (w position of the correlation peak) is used by the machine, however, in order to limit the size of the w window over which the correlation function must be calculated (the mark width is updated when in EXPERT mode).

4.4 Effect of Filtering and Smoothing on the Auto-correlation.

In this section the effect of the smoothing and filtering algorithms, NEWSMO, and LOPASS, is investigated, and their influence on the auto-correlation function is determined.

4.4.1 NEWSMO.

Figure 4.6 shows the resultant profile when the data set shown in Figure 4.3 is treated with subroutine NEWSMO, using 21 point, third order smoothing. The effect of the smoothing has been to reduce the size of fringes in the alignment mark signal. The correlation function obtained from the smoothed data still has three peaks within the region of interest as can be seen in Figure 4.7[†]. These peaks are more widely separated than in the unsmoothed case, however, which gives the peak search algorithm a better chance of identifying the central peak as the correct one, provided that the search begins in the correct place. As can be seen from the contour values, however, the magnitude of the central peak has been reduced with respect to the magnitude of the subsidiary peaks; this type of smoothing should therefore be used with caution.

Figure 4.8 shows the effect of using a third order, 25 point smooth on the same data, while Figure 4.9 shows the auto-correlation derived from this. Comparing the latter with Figure 4.7 shows that the magnitude of the central peak has again been lowered in comparison with the secondary peak on the left hand side. In fact OASIS found the local gradient peak in Figure 4.7 at p=153.3 and in Figure 4.9 at p=143.3, a difference of $\sim$2.0μm (1 pixel = 0.217μm). Again it is obvious that this type of filtering should be used with caution, since it can cause a large shift in the position of the captured peak.

Figures 4.10 and 4.11 show unsmoothed and smoothed (21 point, third order) video profiles, taken from the stepper, while aligning poly-silicon to active area. The alignment mark thus consists of a gap in field oxide, covered by poly-silicon and photo-resist. The granularity of the poly-silicon is evident in the high noise level present in the video data. The smoothing process has removed this noise from the signal, as well as reducing the size of the diffraction fringes. Figures

[†]The size of the window for the correlation plot is 50 $\times$ 50 pixels in (p, w) ($\sim$ 11μm$\times$11μm), which is large enough such that the local gradient search will never look for a peak outside this region.

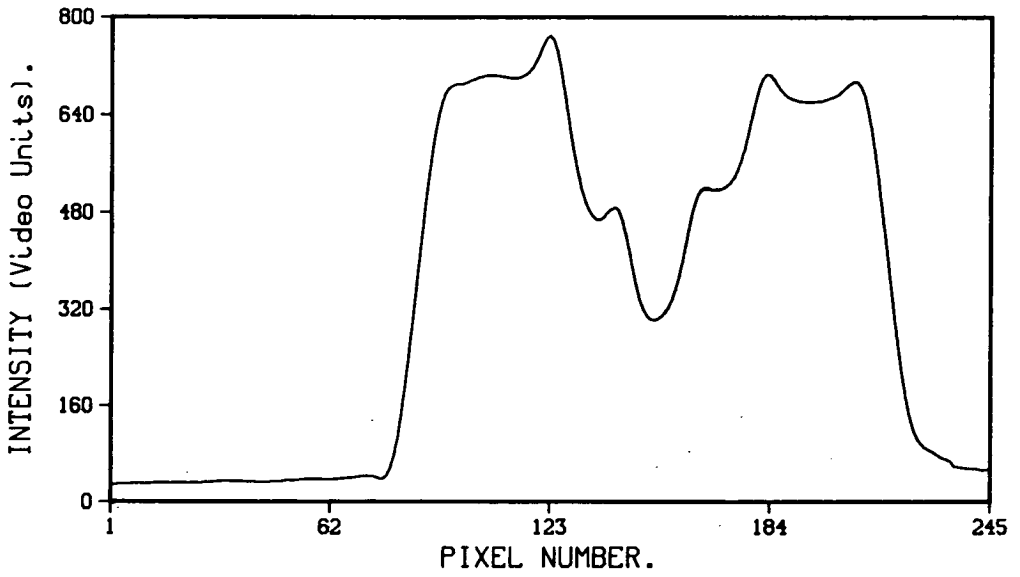

Figure 4.6: Effect of 21 point, third order smoothing on the data in Figure 4.3.

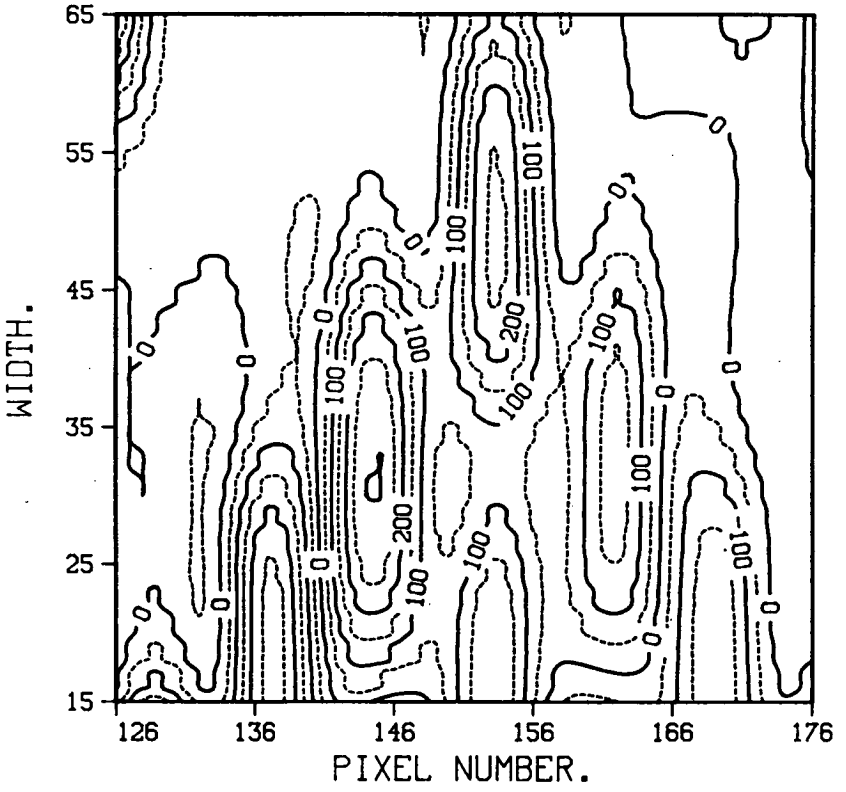

Figure 4.7: Correlation plot from data in Figure 4.6.

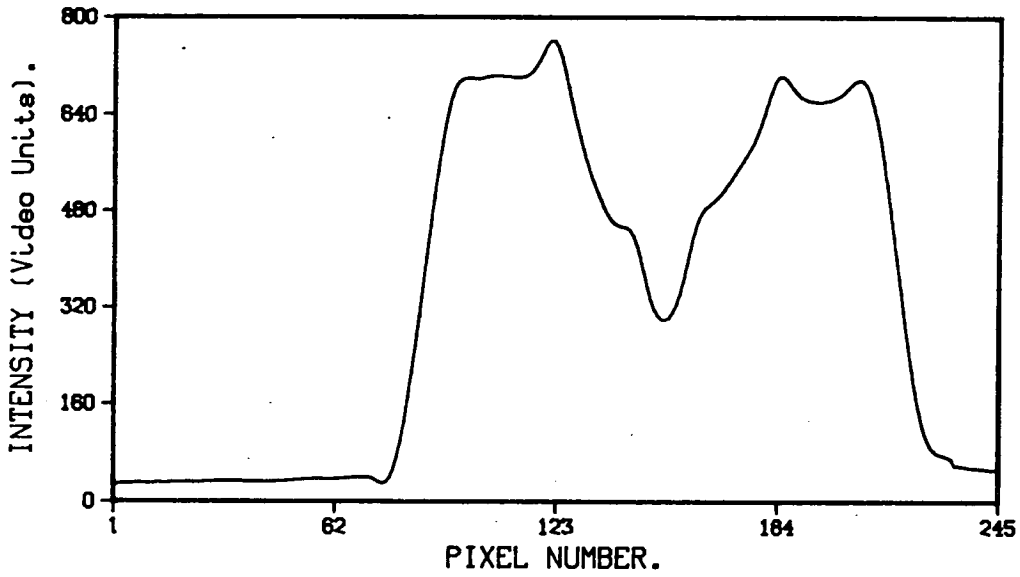

Figure 4.8: Effect of 25 point, third order smoothing on the data in Figure 4.3.

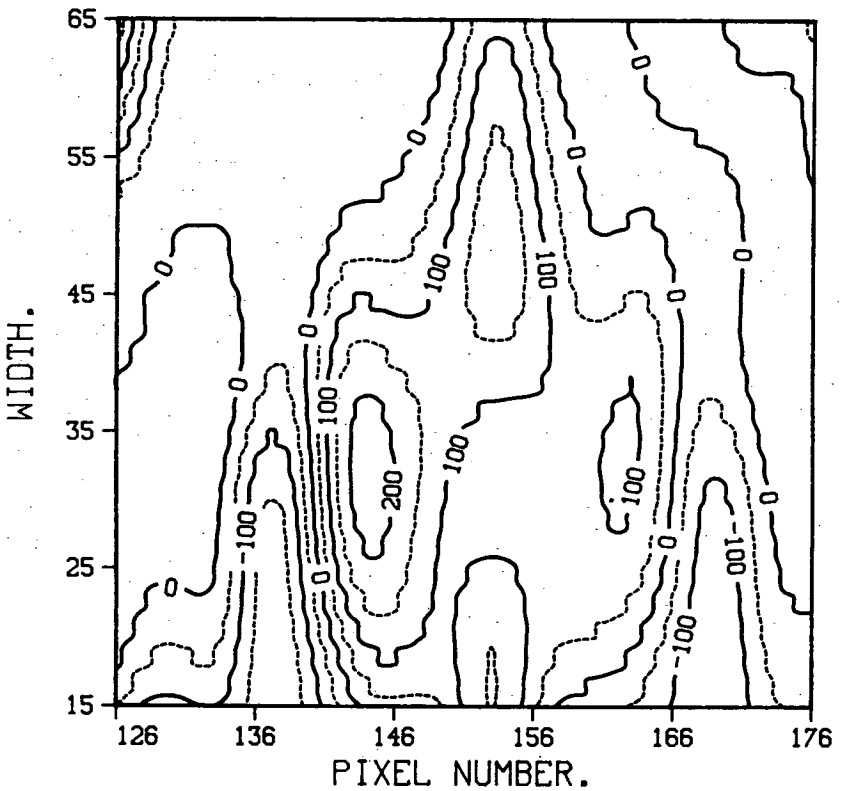

Figure 4.9: Correlation plot from data in Figure 4.8.

4.12 and 4.13 show correlation plots from the unsmoothed and smoothed data respectively. In this case the number of peaks has been reduced to one by the smoothing, allowing the peak location algorithm to unambiguously identify the correct peak. OASIS identified the central peak in Figure 4.12 at p=141.1, and in Figure 4.13 at p=141.9, a difference of less than 0.2μm. This distance is significant in terms of wafer alignment, but will result in a constant offset which can be adjusted for in the stepper software.

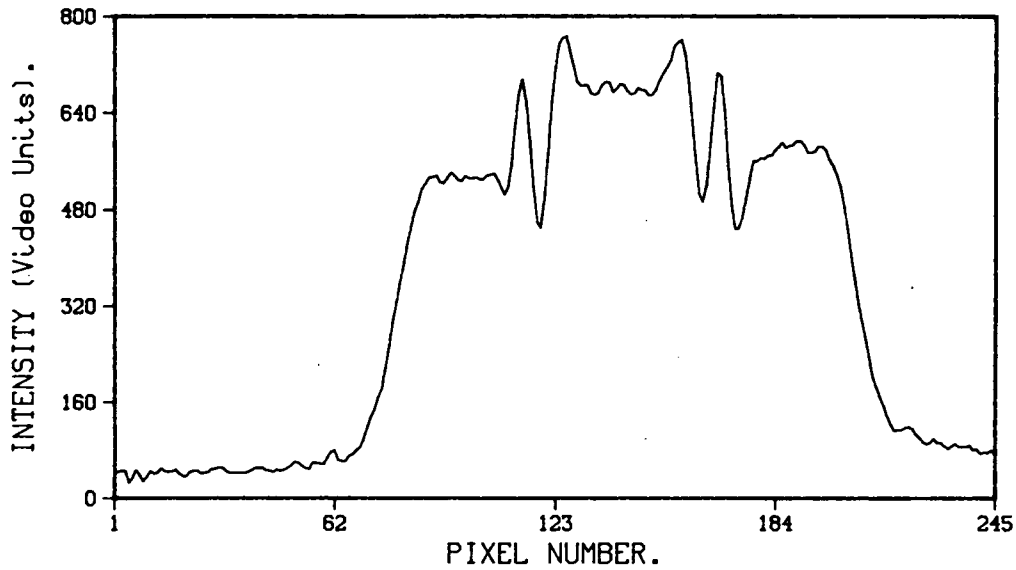

Figure 4.10: Alignment data taken while aligning poly-silicon to active area.

Figures 4.14 and 4.15 show the same data after a 25 point, third order smooth, and the auto-correlation function derived from this. In this case, since there is only one peak, the larger smooth will hardly affect the aligned position of the wafer (although the magnitude of the peak has again been reduced). OASIS identified the peak position at p=142.2, a difference of less than 0.07μm from the peak in Figure 4.13.

4.4.2 LOPASS.

Figure 4.16 shows the effect of a 99-point low-pass digital filter on the data set in Figure 4.3. The cut-off frequency was, in this case, chosen to be 0.167 cycles/μm, corresponding to a feature size of 3.0μm (this frequency is user selectable in the terminal input part of the program). This means that any information from

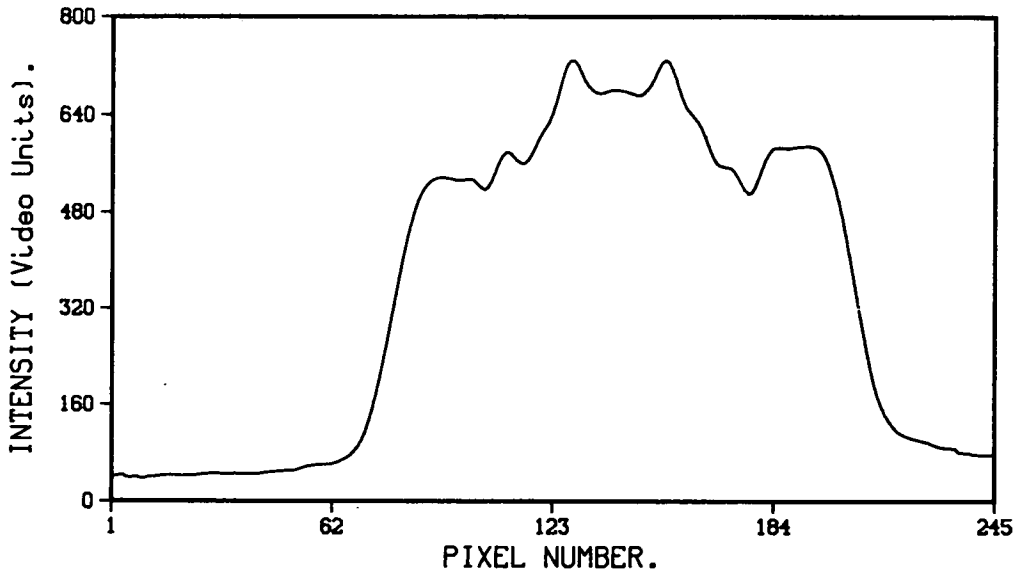

Figure 4.11: Data in Figure 4.10 after a 21 point, third order smooth.

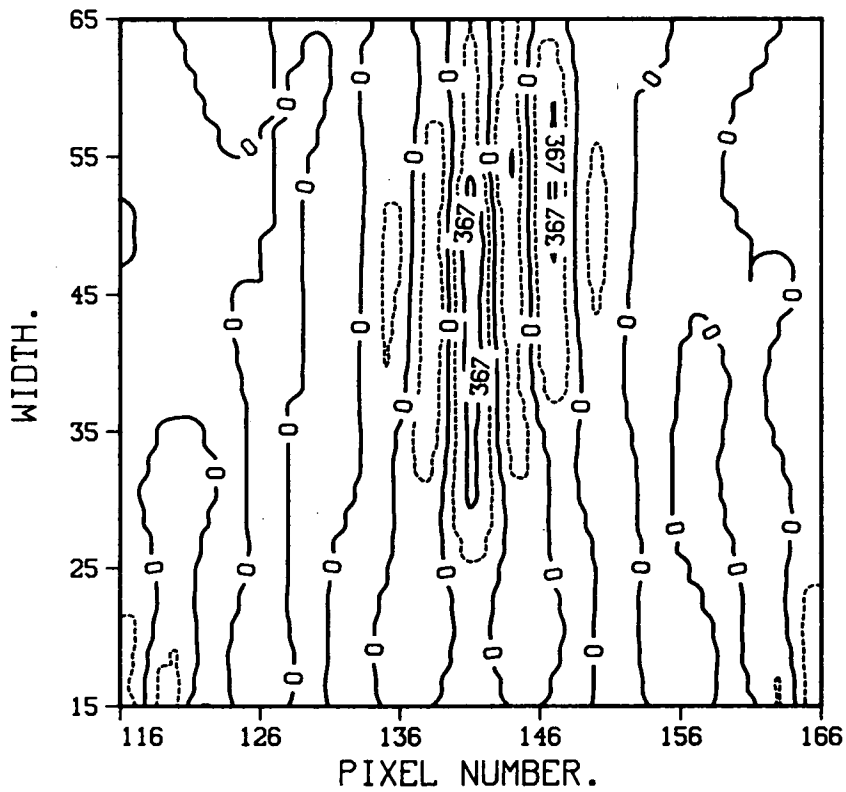

Figure 4.12: Correlation plot for data set in Figure 4.10.

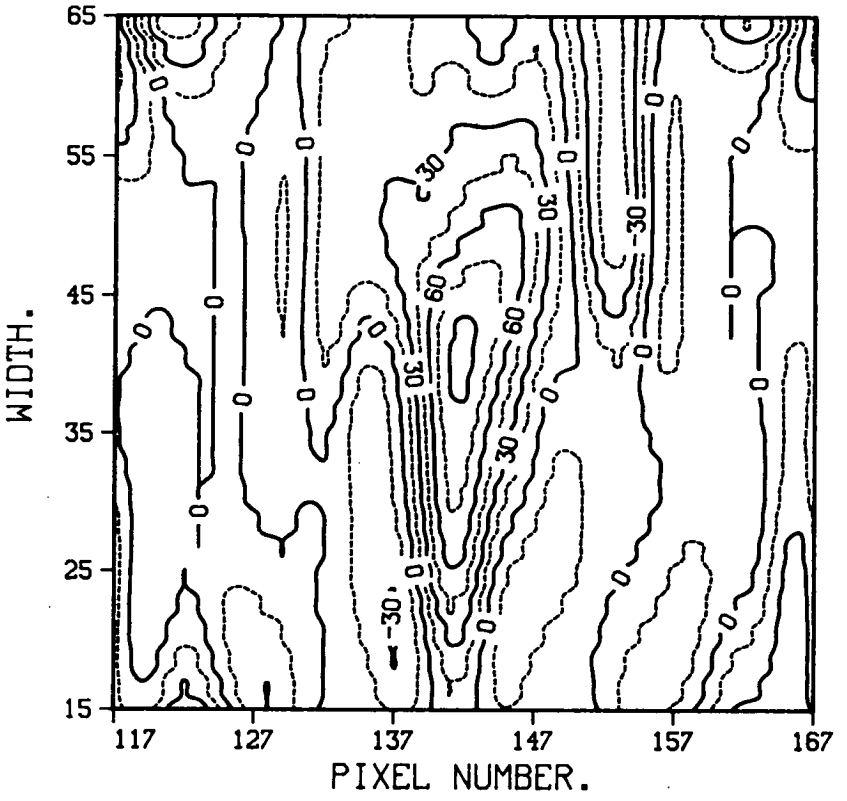

Figure 4.13: Correlation plot for data set in Figure 4.11.

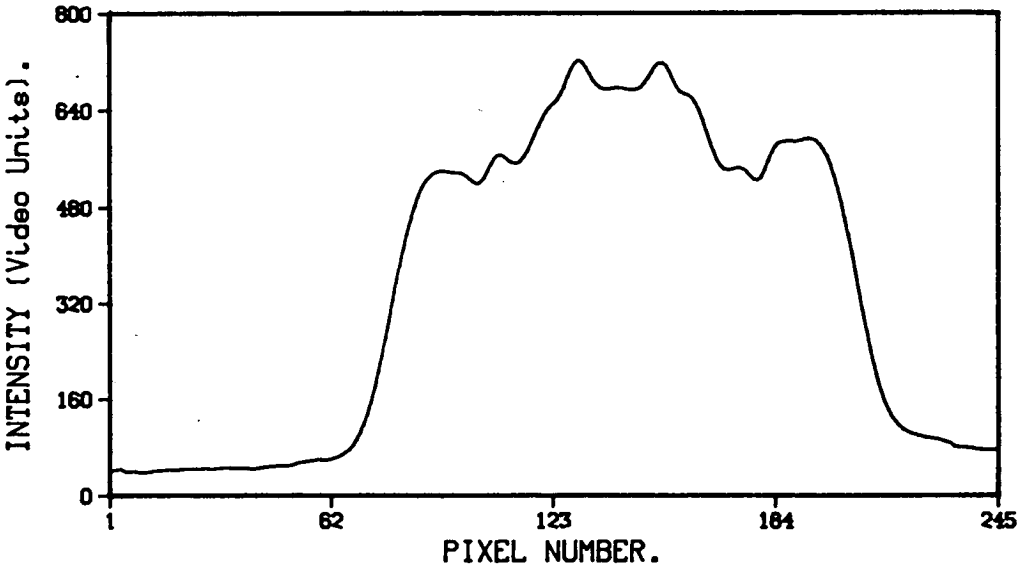

Figure 4.14: Raw data in Figure 4.10 after 25 point, third order smooth.

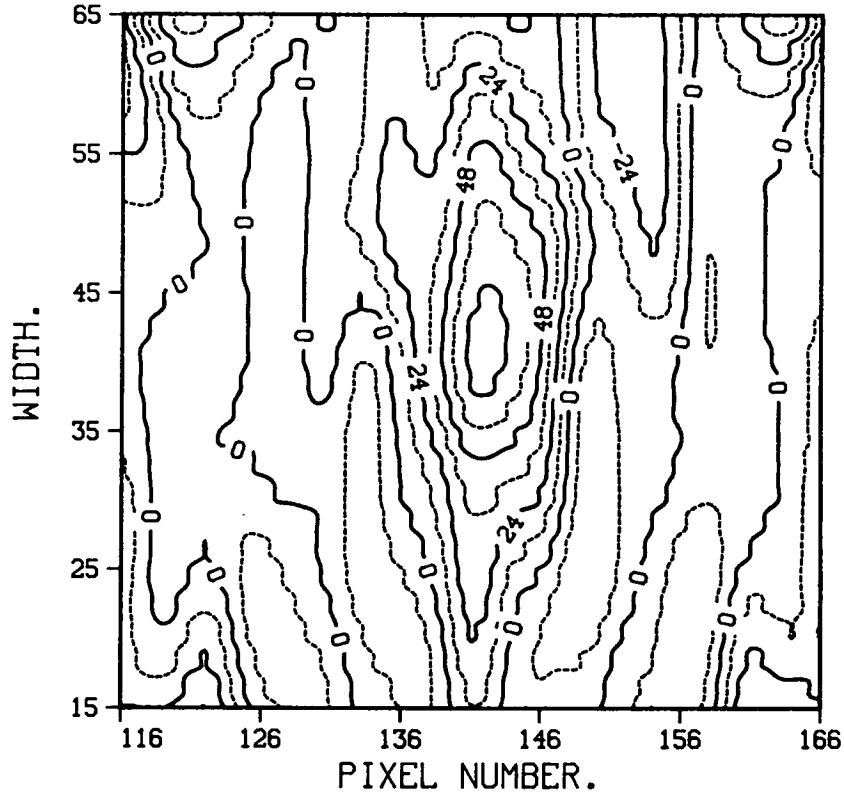

Figure 4.15: Correlation plot from data in Figure 4.14.

features smaller than 3.0μm should not appear in the output signal from the filter. This particular value was chosen since the width of a diffraction fringe tends to be around 1.0μm–1.5μm, while the width of the wafer mark is ~8.5μm[†]. A 3.0μm filter should, therefore, serve well to separate the diffraction fringes from the desired signal. As can be seen from the figure, the diffraction fringes have been completely eliminated from the signal. This filtering results in the auto-correlation function shown in Figure 4.17, which shows a single well defined peak at p=151.4. The peak position for the untreated data was p=153.2, a difference of ~0.4μm. This difference must again be compensated for in the machine software.

Identical filtering was carried out on the raw data shown in Figure 4.10, giving the resultant profile in Figure 4.18, and the auto-correlation in Figure 4.19. This again has a single well defined peak, this time at p=141.5 (cf. p=141.1 in the untreated case). Thus, even in the case of extreme fringing and high noise, the filtering process is able to produce a correlation signal which defines an unambiguous aligned position.

[†]Although the physical width of a wafer mark is 6.0μm, the 45°scan gives an effective mark width of $6.0\sqrt{2}$ μm.

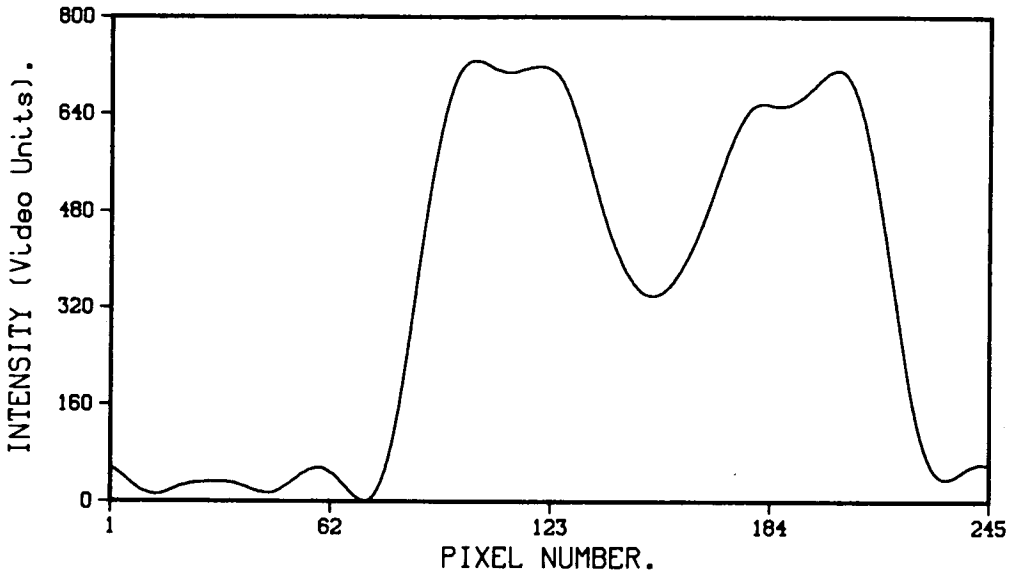

Figure 4.16: Effect of performing 99 point low-pass filtering on the data shown in figure 4.3.

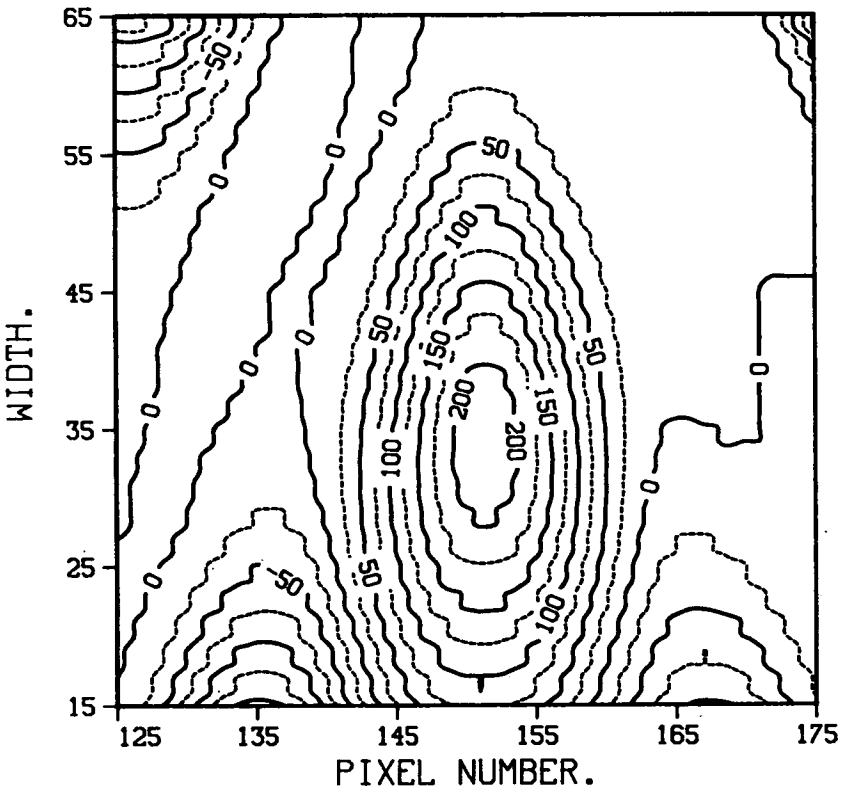

Figure 4.17: Auto-correlation function derived from data in Figure 4.16

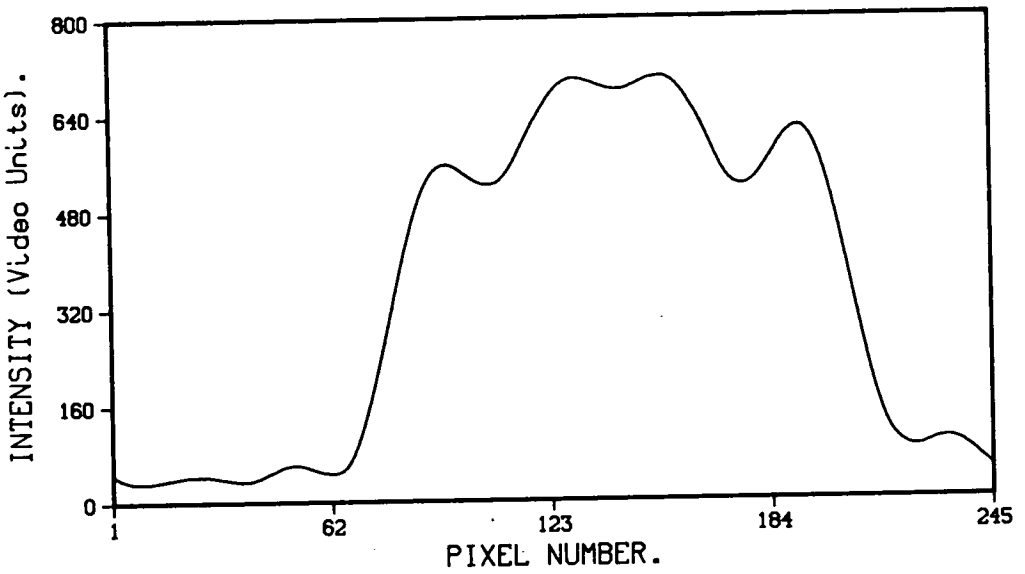

Figure 4.18: Effect of performing 99 point low-pass filtering on the data shown in figure 4.10.

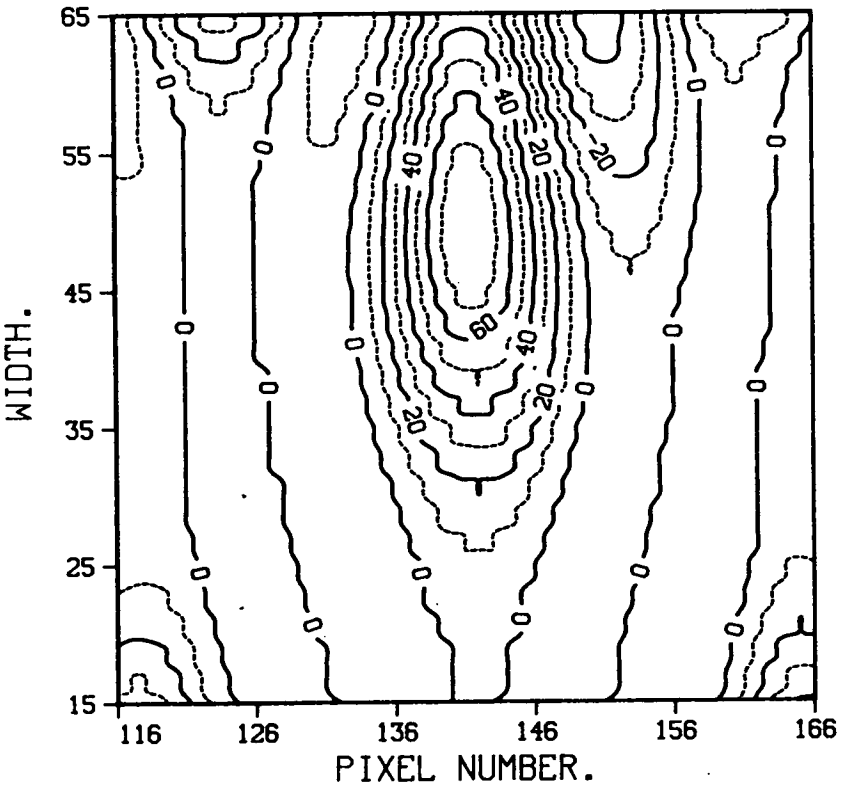

Figure 4.19: Auto-correlation function derived from data in Figure 4.18

4.5 Effect of Different Correlation Functions on the Alignment Signal.

In this section, the effect of different correlation algorithms (cross-correlation, modulus difference and squared difference), is investigated. Since the alignment data shown in Figure 4.10 is a fairly extreme test of the alignment system, due to the high noise level and fringing, the results of processing with these algorithms will be presented for this data only. The trends observed are consistent with the results obtained for other data sets which were tried out with these algorithms.

Figures 4.20–4.22 show the modulus difference, squared difference, and cross-correlation functions respectively, for this particular data set. While Figures 4.20 and 4.21 show a close correspondence to the auto-correlation function, with three peaks roughly centred in p on the correlation window, the shape of the peaks in the cross-correlation is somewhat different, with the main central peak being split into three separate peaks in the w-direction. This is due to the narrowness of the peaks in the constructed array $G(p)$. In general, narrow peaks in the first derivative, or a low SPA value (which limits the width of the peak which the algorithm uses to calculate the correlation function), give more resolution in the width direction. Since the correlation peaks produced by this increased resolution have the same position in the p-direction, this does not affect alignment accuracy.

The same three correlation functions are shown in Figures 4.23–4.25, this time after a 99-point digital filter had been applied to the data. All three plots are virtually identical. Table 4.3 shows the position, in p, in which OASIS detected the peaks in the correlation functions for this data, in both the filtered and non-filtered cases. The largest difference between different correlation types is 0.3 pixels, equivalent to $\sim$0.06μm, indicating that any of these algorithms would be a reasonable alternative to the standard auto-correlation.

The time taken to calculate the various correlation functions was estimated simply by evaluating the amount of time to execute the appropriate section of Fortran code. This is slightly dependent on the loading of the system, and was therefore performed at night when the system was relatively quiet. In addition to this, the program was run ten times for each type of correlation, and the average execution time was calculated from these runs. The results are shown in Table 4.4. The cross-correlation is by far the slowest of these algorithms, being around five times as slow as the others. The others are all approximately the same, with the modulus difference being 5% faster than the auto-correlation, and 6%

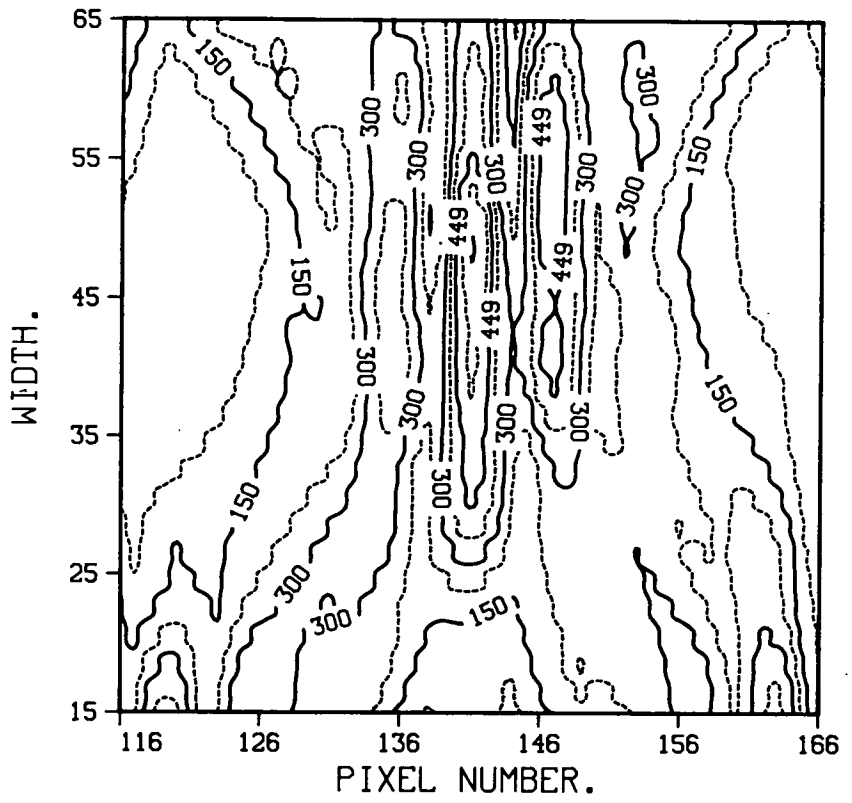

Figure 4.20: Modulus difference function derived from data in Figure 4.10.

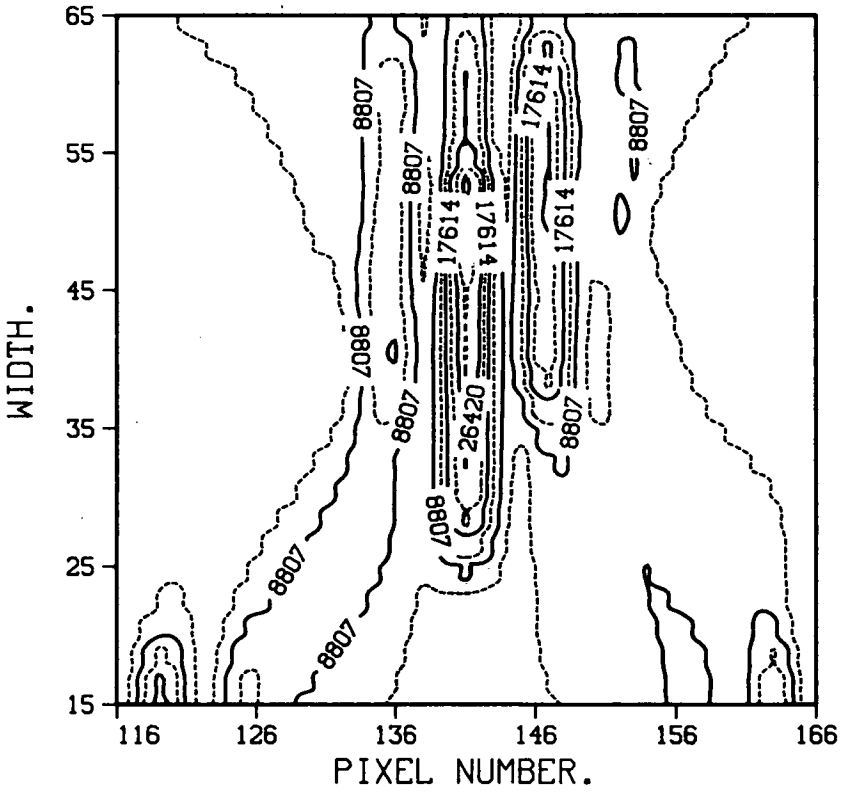

Figure 4.21: Squared difference function derived from data in Figure 4.10.

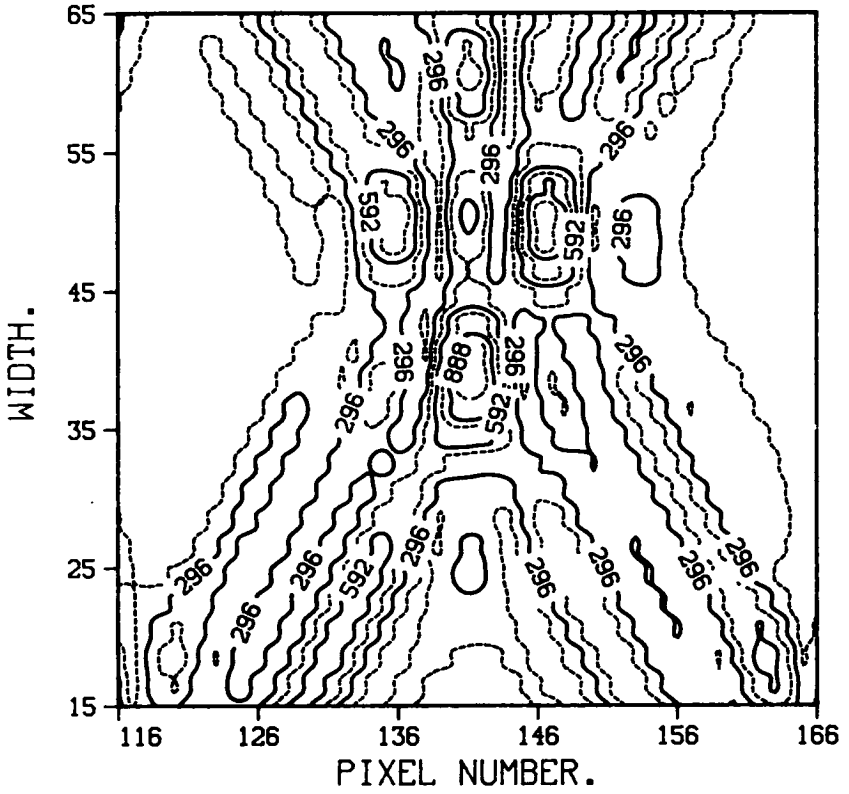

Figure 4.22: Cross-correlation function derived from data in Figure 4.10.

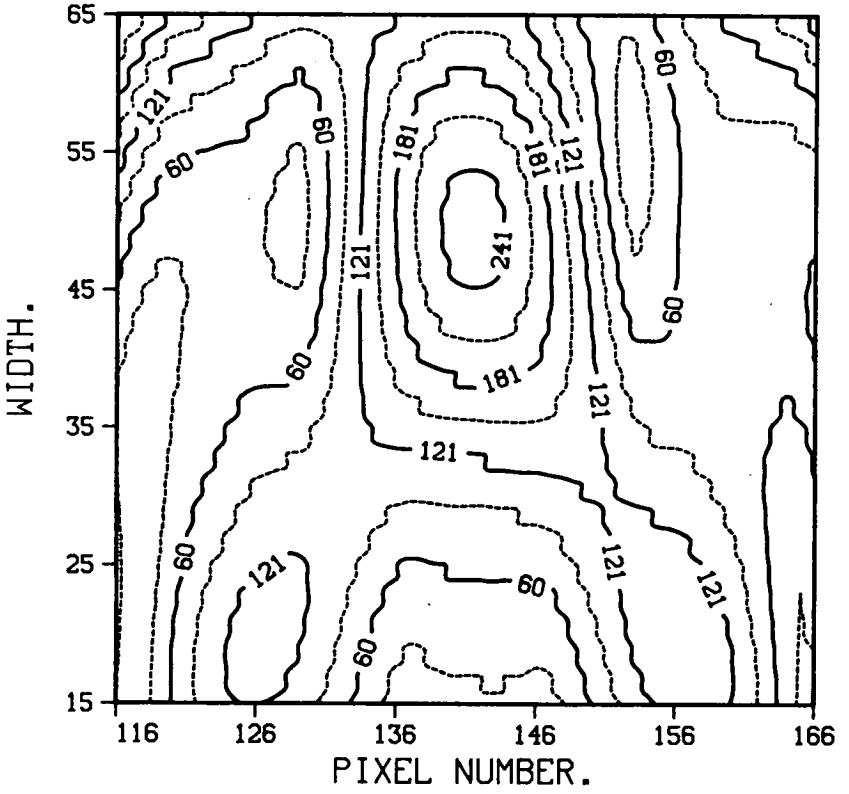

Figure 4.23: Modulus difference function derived from data in Figure 4.18.

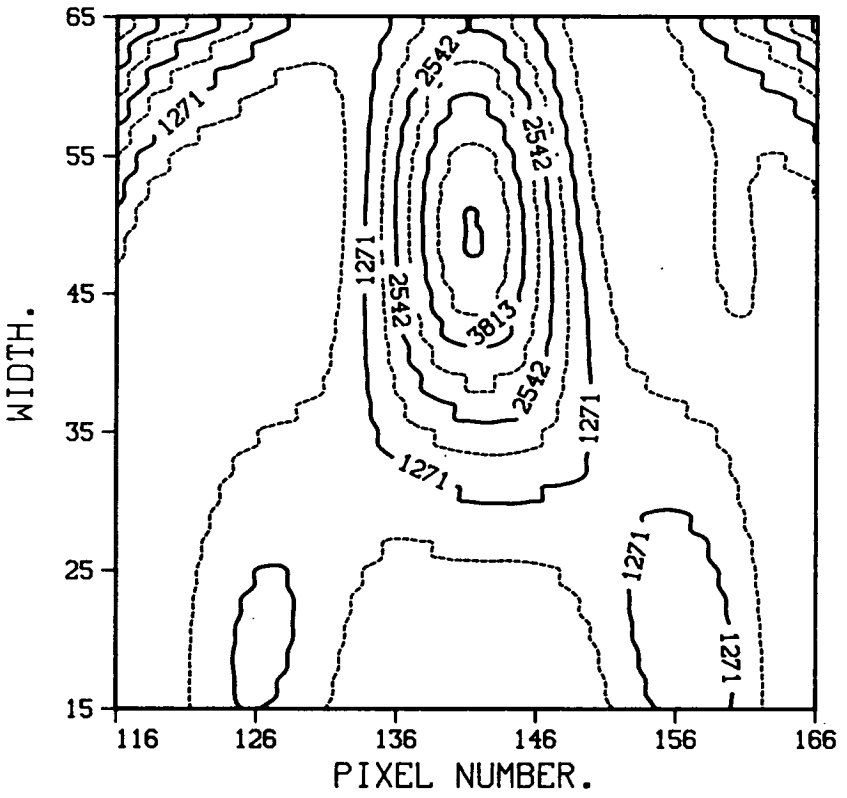

Figure 4.24: Squared difference function derived from data in Figure 4.18.

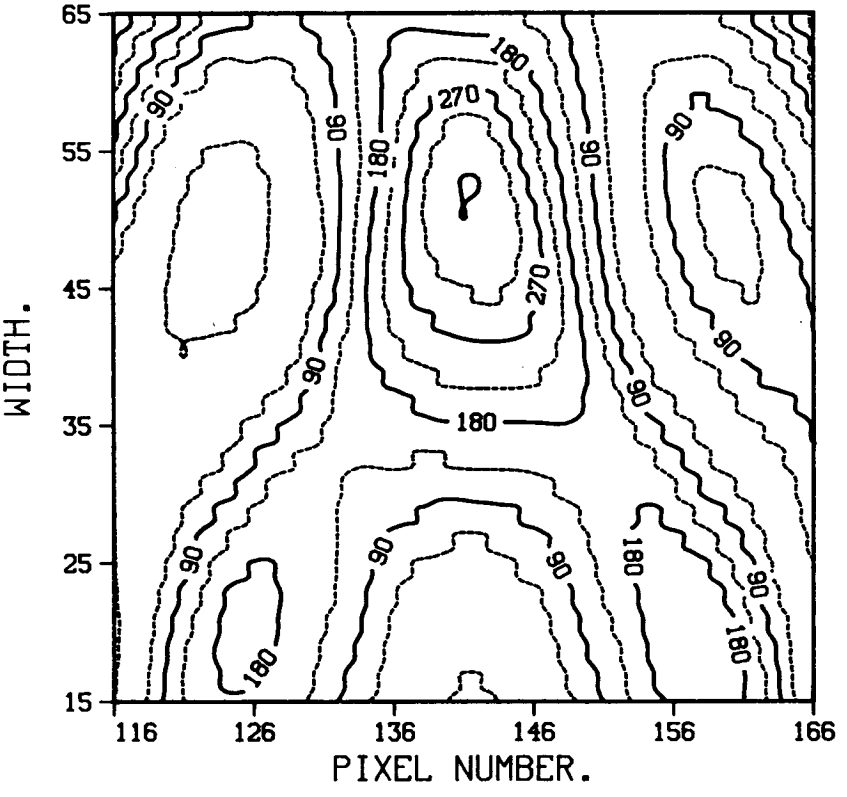

Figure 4.25: Cross-correlation function derived from data in Figure 4.18.

Correlation type.	Filtered (y/n)	Peak position (pixels)
Auto-correlation.	n	141.1
Auto-correlation.	y	141.5
Modulus difference.	n	141.2
Modulus difference.	y	141.6
Squared difference.	n	141.2
Squared difference.	y	141.5
Cross-correlation.	n	140.9
Cross-correlation.	y	141.5

Table 4.3: Position of correlation peak for different algorithms.

Correlation type.	Average execution time. (secs.)
Auto-correlation.	0.1680
Modulus difference.	0.1600
Squared difference.	0.1696
Cross-correlation.	0.8219

Table 4.4: Execution time for different algorithms.

faster than the square correlation. Alternative sources suggest that the modulus difference and squared difference algorithms, when implemented optimally in low level software, can result in a reduction of up to 50% in the calculation time [77].

4.6 Effect of Varying DIA CAMERA Parameters on the Correlation Function

There are three DIA CAMERA parameters which can be varied using OASIS:

1. The spatial analysis correlation coefficient, SPA.
2. The minimum permissible width of an alignment mark, WMI.
3. The maximum permissible width of an alignment mark, WMX.

The effect of the WMI and WMX parameters is relatively staightforward; the time taken to calculate the correlation function will be directly proportional to the difference WMX−WMI. The form of the correlation function will be unchanged. It is therefore advisable to make this difference as small as possible, while still allowing for some amount of linewidth variation across the wafer. Care should be taken not to make the difference too small, to prevent the correlation peak from falling outside the window. In the Optimetrix software, the parameter DIA in DIA CAMERA can be used to obtain an estimate of the mark width. In general this estimate ±2 pixels should be a wide enough window to include the correlation peak (this corresponds to a linewidth variation of $\pm 0.4\mu$m).

The effect of the SPA parameter was determined by evaluating the autocorrelation function for the data in Figure 4.3, for four different values of SPA (see Figure 4.4 for the correlation function when SPA=13). The correlation functions for SPA=1, SPA=5, and SPA=21 are plotted in Figures 4.26–4.28. From these plots it can be seen that lower values of SPA result in a succession of peaks in the w direction, and that these peaks merge into one another as SPA is increased. The positions of the peaks in these plots are tabulated in Table 4.5, along with the time taken to calculate the correlation functions. It should be noted that as SPA varies between 1 and 21, the position of the peak is constant to within 0.1 pixel ($\sim 0.02\mu$m), while the calculation time increases by a factor of 8. In fact, the same variation in SPA was tried out on a few different data sets, and the maximum offset found in the p-direction was 0.1 pixels. In over half the cases tried out, no difference in the peak position was detected, thus demonstrating that small values of SPA, while greatly reducing the calculation

time of the correlation function, have no adverse effect on the alignment accuracy.

SPA	Execution time (secs.).	Peak position. (pixels).
1	0.031	153.2
5	0.070	153.2
13	0.168	153.2
21	0.250	153.2

Table 4.5: Execution time and peak position for different SPA values.

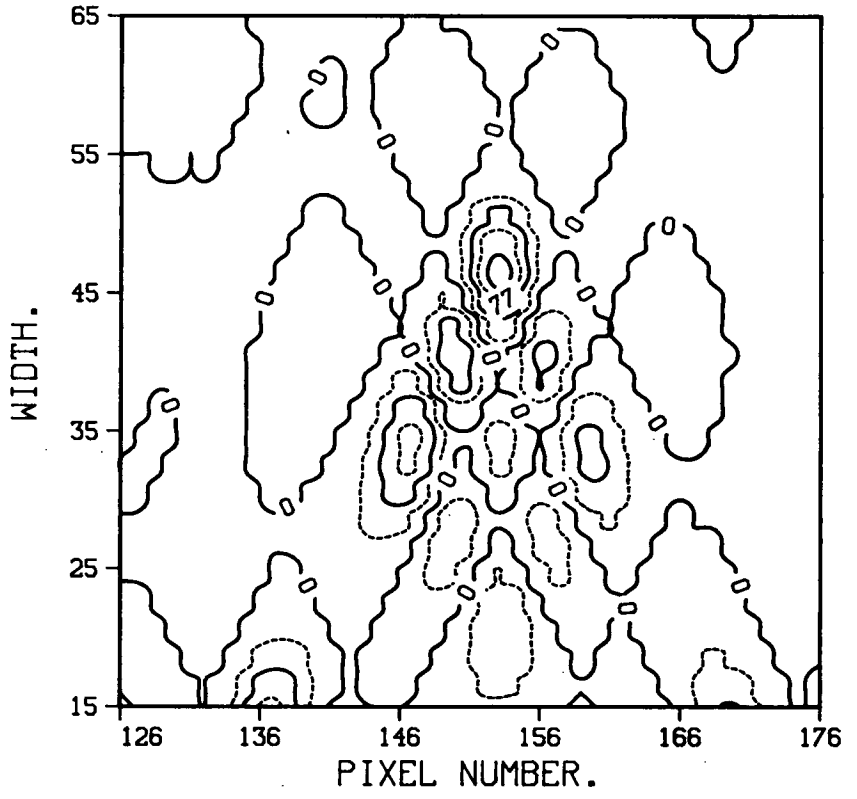

Figure 4.26: Auto-correlation function derived from data in Figure 4.3 (SPA=1).

4.7 Auto-align Limits.

As was stated in the last section, the parameters WMI and WMX specify the minimum and maximum allowable mark widths. There are no corresponding DIA CAMERA parameters which specify the maximum and minimum allowed

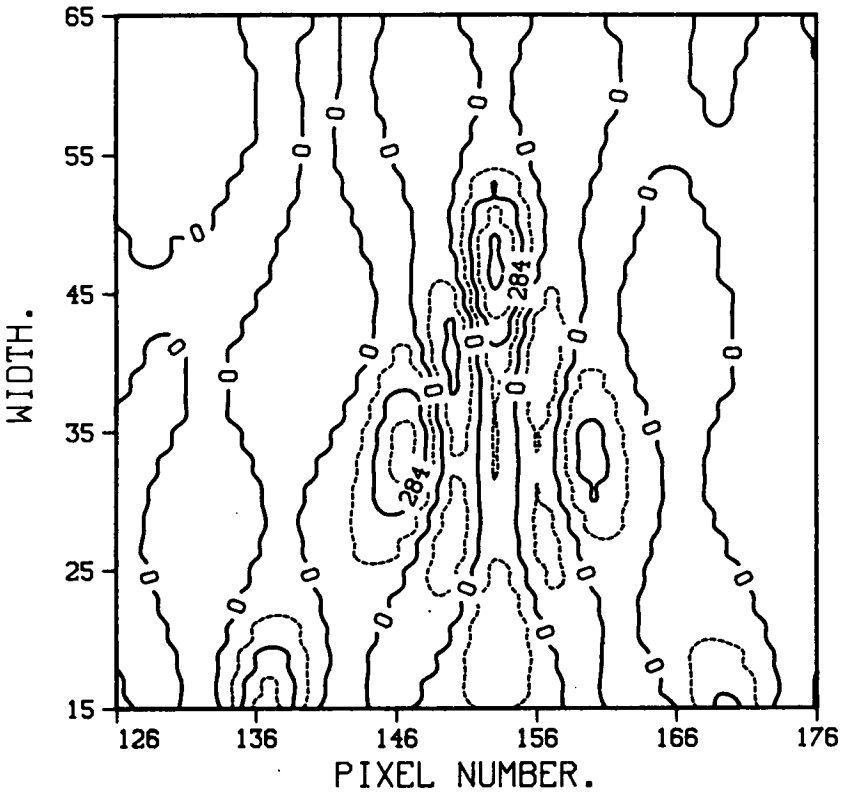

Figure 4.27: Auto-correlation function derived from data in Figure 4.3 (SPA=5).

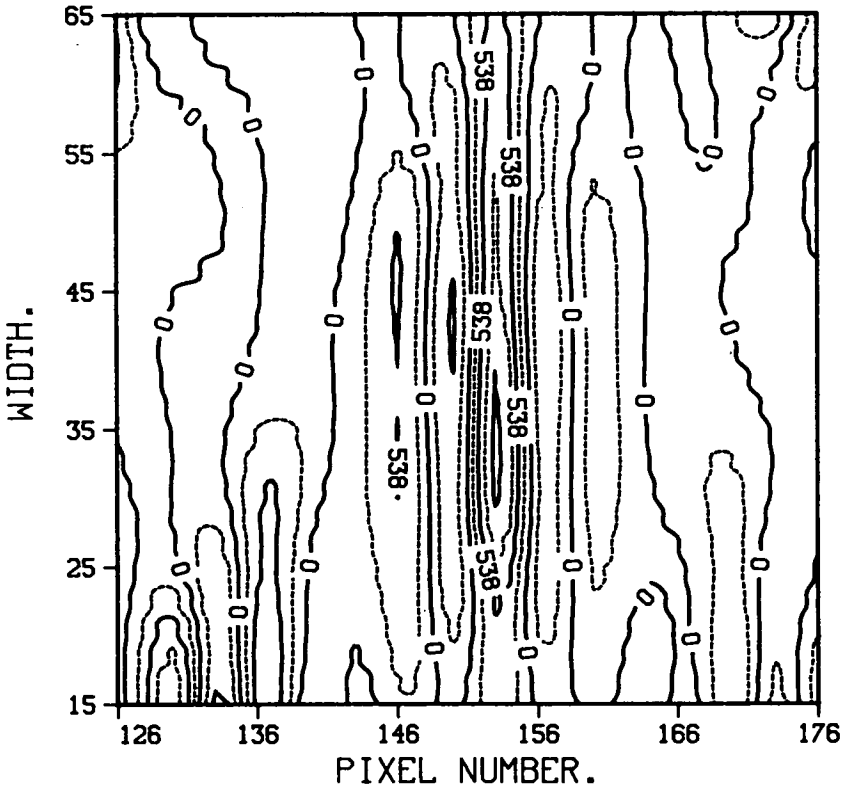

Figure 4.28: Auto-correlation function derived from data in Figure 4.3 (SPA=21).

positions for the centre of the mark (such parameters would in fact specify the limit of the correction that the stage would be allowed to make). There is a location in COM DEB, however (machine memory location S474), which specifies this limit of correction. The default value is 80 counts (1 count = 0.027μm), or $\sim$2.2μm. Stage stepping accuracy is claimed to be 0.3μ (3σ), resulting in a limit in the auto-align capability which far exceeds that which should be required. Assuming that the system has managed to align correctly to the global markers, and align to the first die on a wafer, 0.4μm is a conservative estimate of the largest step which should ever be required to bring the wafer back into correct registration (given a 0.3μm stage accuracy). If a larger step is ever performed, it is probably because the local gradient search has identified the wrong peak in the correlation function. A value of no larger than 20 for memory location S474 should be sufficient to prevent this from happening, and thus improve the alignment reliability. If no peak can be found within this window, the system should be allowed to auto-align fail, and default to the best known blind step position. The best blind step position should be calculated from the positions of the global markers, and should *not* be taken from any of the previous die-by-die alignment positions. If the latter method is used (as it is presently on the machine), one bad misalignment due to a damaged marker can cause the whole wafer to be exposed with the same misalignment.

4.8 Conclusions.

There are several possibilities for improving alignment performance on the Optimetrix stepper, using both capabilities which are currently available on the machine, and some which are not. In this chapter it has been demonstrated that smoothing and filtering algorithms can be applied which improve the reliability of the system, by eliminating multiple peaks in the correlation function. In particular, low-pass filtering was applied to eliminate the effect of high order diffracted components, which contribute strongly to fringes in the alignment signal. This type of filtering was particularly successful in providing single peaked, well behaved correlation functions. Although this work been presented with the Optimetrix stepper particularly in mind, there is no reason why this method of fringe reduction could not be used on any stepper whose alignment system is based on bright field detection.

The modulus difference and squared difference algorithms have been shown to be very similar in form to the standard auto-correlation, while they should

be very much faster to compute when implemented optimally in low level software. These alternative algorithms should, therefore, lead to an increase in wafer throughput, while maintaining the same alignment accuracy as the autocorrelation. Similarly, it has been shown that optimum use of DIA CAMERA parameters can also lead to a decrease in correlation function calculation time, and thus to an increase in wafer throughput.

Chapter 5

The Thick Film Problem.

As was stated in Chapter 3, the main drawback of an alignment system based on reflectivity variations is a susceptibility to fringes which occur in the reflected intensity profile. An example of this type of fringing is shown in Figure 3.5, which was taken from an alignment mark etched in poly-silicon and coated with photo-resist.

Many people have assumed that these fringes are caused by interference due to a variation in both resist thickness and poly thickness in the region of the etched mark [79] [80] [81], and have proceeded to calculate mark contrast ($I_{max} - I_{min}/I_{max} + I_{min}$) on this basis.

This approach is valid provided that the width of the etched mark is much greater than the wavelength used to illuminate the wafer. If, however, the mark is of similar dimension to the wavelength, or if an estimate of the reflected intensity at the edge of the mark is required, then the complete structure must be regarded as one in which the refractive index varies in two dimensions (in y and z). Under these conditions the Fresnel equations are no longer valid, and we must go back to Maxwell's equations to find a solution.

It is precisely this problem with which we must be concerned, if we wish to be able to predict the intensity profile which will be obtained from a video scan of an alignment target. By simulation of the profiles we should be able to gain an important insight into the properties of the target which affect the fringing most strongly (eg. variation of width of the target, variation of line edge slope etc...). This is potentially very important for *all* bright field, reflected light alignment systems, since a knowledge of the important factors could lead to the possibility of reducing mark fringing by tailoring of the appropriate property.

Some important work has been done in the solution of the appropriate equations by Burckhardt [82], Kaspar [83] [84], Nyyssonen [85], Streifer et al. [86],

and Kirk and Nyyssonen [87]. The purpose of this chapter is to review briefly some of this work, and to present some results produced by a modified version of a program written by Kirk for the purpose of modelling optical microscope images (although the program was applied to the simulation of alignment mark profiles). For completeness, the derivation of the relevant partial differential equation (PDE) from Maxwell's Equations is given.

5.1 The Derivation of the PDE from Maxwell's Equations.

We begin by considering a plane, time harmonic wave, incident upon a generalised optical structure whose refractive index varies in the y and z directions (see Figure 5.1). We consider only the component of the wave which is polarised with its electric vector, $\vec{E}$, perpendicular to the plane of incidence (the TE wave). Since any wave can be regarded as a sum of TE and TM waves (magnetic vector, $\vec{H}$, perpendicular to the plane of incidence), and the solution for a TM wave may be deduced from the solution for a TE wave [88], simply by interchanging ϵ and $-\mu$, the relative permittivity and permeability, it is in general sufficient to consider only this component.

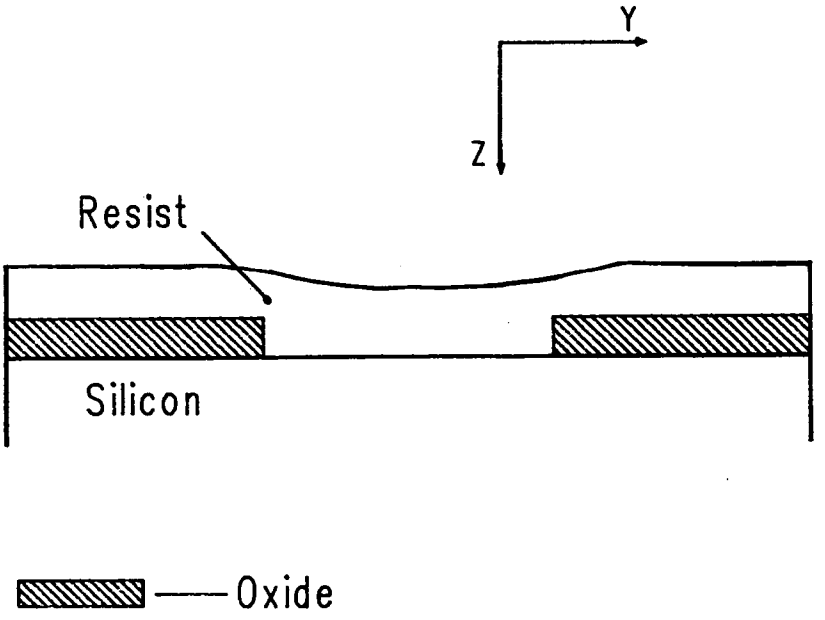

Figure 5.1: Optical structure with refractive index varying in y and z directions.

We take the plane of incidence to be the y-z plane. For a TE wave, E_y =

$E_z = H_z = 0$. The Maxwell relation

$$\vec{\nabla} \times \vec{H} = \frac{\varepsilon}{c}\dot{\vec{E}} + \frac{4\pi}{c}\sigma\vec{E} \tag{5.1}$$

where c is the speed of light in vaccuo, σ is the conductivity of the material, and $\dot{}$ denotes differentiation with respect to time ($e^{-i\omega t}$ dependence assumed), then reduces to the following three scalar equations:

$$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = -i\frac{\varepsilon\omega}{c}E_x + \frac{4\pi}{c}\sigma E_x \tag{5.2}$$

$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = 0 \tag{5.3}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = 0 \tag{5.4}$$

while the relation

$$\vec{\nabla} \times \vec{E} = -\frac{\mu}{c}\dot{\vec{H}} \tag{5.5}$$

reduces to the equations

$$\frac{\partial E_x}{\partial z} = i\frac{\mu\omega}{c}H_y \tag{5.6}$$

$$\frac{\partial E_x}{\partial y} = -i\frac{\mu\omega}{c}H_z \tag{5.7}$$

Substituting the expressions for H_y and H_z in 5.6 and 5.7 into 5.2, and assuming that all media present are non-magnetic (a good assumption in our case), leads to the second order partial differential equation:

$$\frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} + \hat{n}^2(y,z)k_o^2 E_x = 0 \tag{5.8}$$

where $\hat{n}^2 = \varepsilon + 4\pi i\sigma/\omega$ and $k_o = \omega/c$. The use of the complex refractive index $\hat{n}$ allows us to solve the equation for absorbing media (eg. dyed resist or aluminium). Substituting a solution of the form:

$$E_x(y,z) = Y(y)Z(z) \tag{5.9}$$

we arrive at:

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} + \frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} + \hat{n}^2(y,z)k_o^2 = 0 \tag{5.10}$$

In the case where $\hat{n}$ is a function of z only, this equation may be separated into two sides which are functions of one variable only, and solved analytically. Such an analysis is given in Born and Wolf [88], from an original theory by F. Abelès. In the general case, however, separation is not possible, and the equation must be solved using a numerical approach, or a series of approximations.

Burckhardt has solved the equation for a coherent beam incident onto a thick photographic plate, whose refractive index varies sinusoidally in y (Figure 5.2) [82]. Kaspar [83] [84] and Nyyssonen [85] have extended this to allow for the computation of the magnitude of the electric field within any periodic structure (Figure 5.3), by expressing the dielectric constant as a Fourier series:

$$\varepsilon(y) = \sum_{n=0}^{n=N} \varepsilon_n \cos(2\pi n b y) \tag{5.11}$$

where ε_n are the Fourier coefficients and b is the spatial period. Kaspar and Nyyssonen's analyses are limited, however, in that only one material may be patterned in the y-direction, and this material must have vertical edge walls.

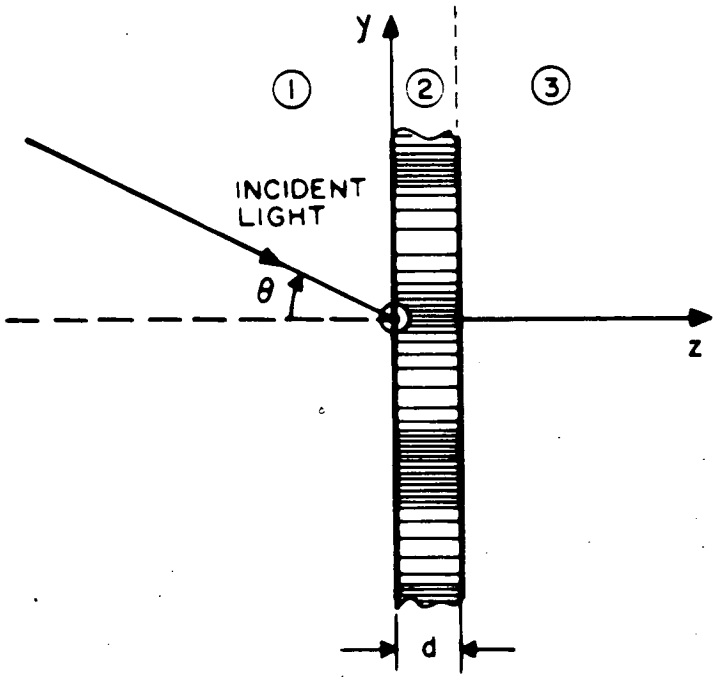

Figure 5.2: Photographic plate, with refractive index varying sinusoidally in y, taken from reference [82].

Streifer et al. [86], have included non-vertical edge walls of arbitrary shape (Figure 5.4), but are still limited to patterned layers of only one material (one

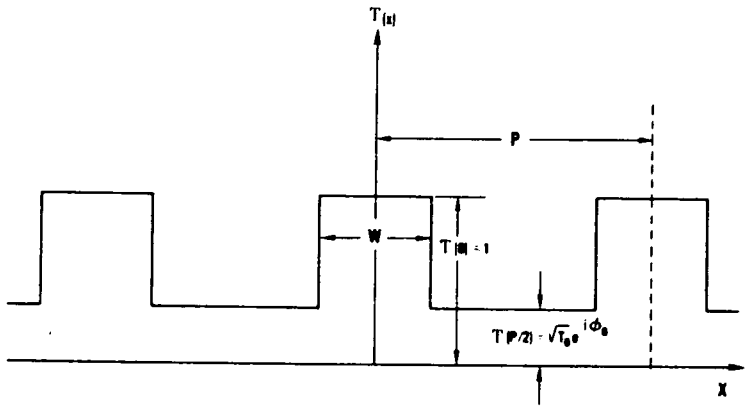

Figure 5.3: Periodic structure (single material) with vertical edge walls, taken from reference [85].

refractive index). Finally Kirk and Nyyssonen [87] have extended this model to enable the calculation of the electric field distribution within structures of arbitrary shape and arbitrary refractive index (Figure 5.5), by splitting the structure into thin layers in z.

Within the limitations set down by Kirk and Nyyssonen [89] this model is sufficient to predict image intensity profiles in the regime of linewidth measurement. These limitations include:

1. Assumption of normal incidence.

2. Assumption of monochromaticity.

3. Assumption of TE-mode polarisation (a solution for the TM-mode can be obtained by exchanging ϵ and $-\mu$ in the case of an analytical solution only).

4. The finite layer approximation. Strictly speaking the structure should be split into vanishingly thin layers in z ($\delta z \rightarrow 0$). However, computational considerations impose a lower limit on δz. Splitting the structure into 10 layers, for example, can result in CPU times of up to 30 minutes on a medium sized computer (VAX8600). In general, therefore, a δz of no less than 1/10th of the structure's thickness is recommended. In practice, for

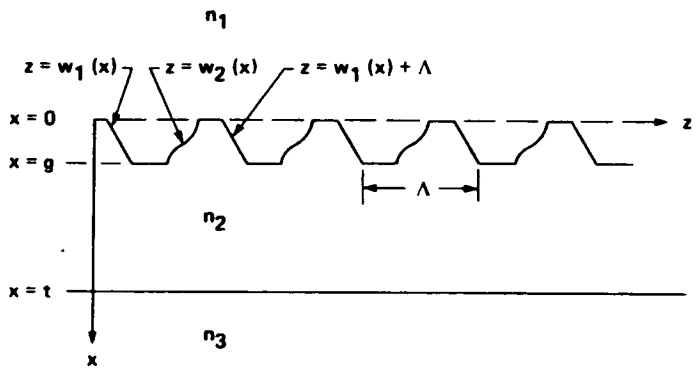

Figure 5.4: Periodic structure (single material) with non-vertical edge walls, taken from reference [86].

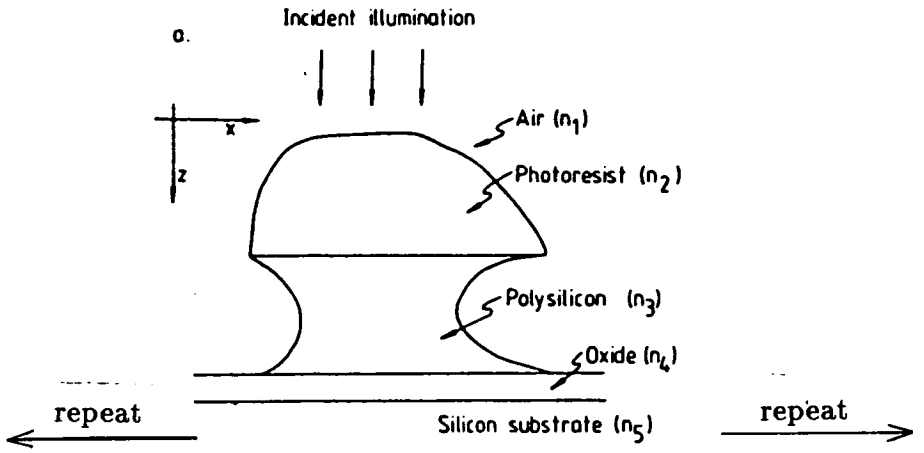

Figure 5.5: Periodic structure (multiple material) with non-vertical edge walls, taken from reference [87].

the results of the simulations presented here, splitting the structure into any more than 10 layers results in very little change in the output intensity profile.

5. Assumption of 1-dimensional object structures. The analysis assumes that ϵ does not vary in x, ie. that the structures are infinitely long.

6. Truncation of the field modes. Orders which are diffracted parallel to or into the substrate are discarded from the solution.

These approximations are discussed at some length in [89]. The best way to view such a structure on a wafer is to consider it as a waveguide, into which the incident radiation is coupled. A structure of a particular given width and height, and with a given sidewall profile, will be able to support various modes of vibration of the field, and these modes of vibration in turn give rise, through loss from the waveguide, to the reflected intensity profile.

5.2 Program NVEW.

A copy of the program written by Kirk to calculate the electric field distribution within such an arbitrary structure was obtained (the program is called NVEW – Non-Vertical Edge Wall), with the idea of gaining an insight into the parameters which affected the fringing most strongly. As it stood, however, the program was limited to modelling structures which lay on top of a substrate, and which were separated by air, as in Figure 5.5. Cases such as that shown in Figure 5.6 could not be modelled, since the structure consists of oxide islands, separated by photoresist. It was necessary first of all, therefore, to modify the program in order to be able to accommodate such structures before a study of fringing during alignment could be undertaken. This was accomplished by altering the input section such that, in addition to the refractive index of the islands, the refractive index of the surrounding material could also be specified.

The first step which was taken after the appropriate modifications had been made was to check that the new version was consistent with the old. This was accomplished by setting up a run identical to one of the runs in reference [89], with the complex refractive index of the surrounding material being given as $1.00+0.00i$ (air). The two outputs were visually checked by comparing graphs of the reflected intensity profiles, and appeared to be identical. Unfortunately the numerical output file is not given in the reference, so no direct comparison was possible. A numerical specification of the 'pseudo-Fourier series' which is

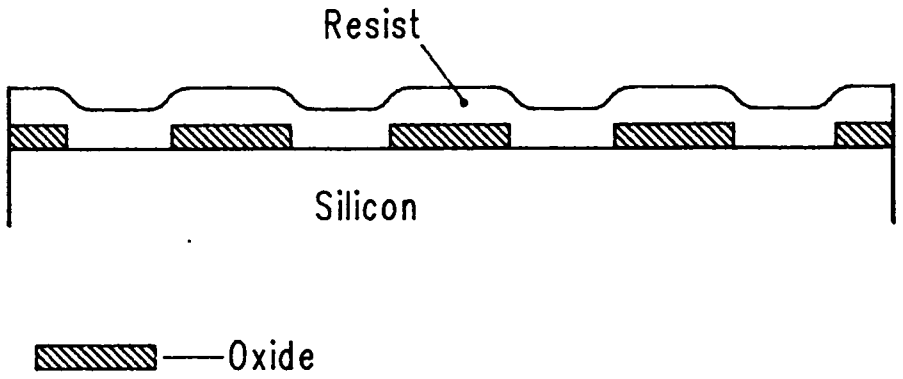

Figure 5.6: Periodic structure in which oxide is surrounded by resist.

used to form the image *is* given in [89], however, and this was compared with the same series from the modified program. The two were found to agree (in the worst case) to within 0.1%, so the modifications were assumed to be valid.

5.2.1 Results from the Program.

For a line which is fabricated in a given material (or materials), and on a given substrate, there are three parameters relating to the line which can affect the intensity profile.

1. The linewidth itself.
2. The thickness of the layers involved.
3. The line profile (eg. sidewall angle after etch, or width of bird's beak).

Several runs of NVEW were made in order to investigate the effect of these parameters on the output profile. An example output file from the program is given in Table 5.1. This output file, which is written in addition to the intensity profile, simply echoes the input, but gives a little more information about the input parameters. These parameters are described briefly below:

- Noise start - starting point for a random number generator to add noise to the profile (simulates a granular structure). Not used here.

- Wavelength - self-explanatory.
- Number of layers- total number of layers into which the structure is divided.
- Wave number - 2π/wavelength.
- Air layer - refractive index of air.
- Substrate - refractive index of silicon.
- Defocus - distance from the air/structure interface at which the electric field is calculated.
- Slit width in microns - desired slit width for simulation of a scanning slit microscope. Not used here.
- Camera width parameter - for simulation of TV imaging of the structure. The pixels are assumed to be Gaussian in shape; the camera width parameter is the 1σ value of the Gaussian curve. This Gaussian is convolved with the image intensity at the TV camera to yield the final profile. In the Optimetrix case, with each pixel representing 0.217μm, this leads to a 1σ value of no more than 0.036μm (adjacent pixels are assumed to be no closer than 6σ for good resolution). A value of 0.036μm for the camera width parameter was assumed throughout this chapter, although in practice, with features no smaller than 1.0μm being resolved by the microscope, this alters the final image very little.
- Layer - self-explanatory.
- Width - width in microns of that particular layer.
- Position - z-position of the interface between one layer and the next.
- RI - refractive index of the layer
- RI (surround) - refractive index of the surrounding material.
- Noise - multiplicative scaling factor for noise to simulate granularity. Not used here.
- Offset - offset in y of the layer (useful for making non-symmetrical patterns).

Within the code itself, the numerical aperture and coherence factor of the imaging optics must be specified; unless otherwise stated, the results which follow were obtained with NA=0.32 and σ=0.5 (this simulates as closely as possible the Optimetrix stepper). All results presented in the following discussion assume a periodicity of 12.0μm for the complete structure (periodicity is necessary in setting up the Fourier series for the refractive indices).

```
RUN PARAMETERS.
NOISE START =            7
WAVELENGTH =        0.4360000
NUMBER OF LAYERS=           10
WAVE NUMBER =    11.85506758717236
AIR LAYER = (1.000000,0.0000000E+00)
SUBSTRATE = (4.100000,-5.9999999E-02)
DEFOCUS =  0.0000000E+00
SLIT WIDTH IN MICRONS =  0.2000000
CAMERA WIDTH PARAMETER =  0.0000000E+00
```

LAYER	WIDTH	POSITION	RI	RI (SURROUND)	NOISE	OFFSET
1	6.030	0.060	1.460 0.000	1.000 0.000	0.000	0.015
2	6.090	0.120	1.460 0.000	1.000 0.000	0.000	0.045
3	6.150	0.180	1.460 0.000	1.000 0.000	0.000	0.075
4	6.210	0.240	1.460 0.000	1.000 0.000	0.000	0.105
5	6.270	0.300	1.460 0.000	1.000 0.000	0.000	0.135
6	6.330	0.360	1.460 0.000	1.000 0.000	0.000	0.165
7	6.390	0.420	1.460 0.000	1.000 0.000	0.000	0.195
8	6.450	0.480	1.460 0.000	1.000 0.000	0.000	0.225
9	6.510	0.540	1.460 0.000	1.000 0.000	0.000	0.255
10	6.570	0.600	1.460 0.000	1.000 0.000	0.000	0.285

Table 5.1: Output file from NVEW.

Variation in Line Profile and Layer Thickness.

Figures 5.7–5.10 show a series of output profiles from the program for a 6.0μm wide space in oxide (refractive index 1.47+0.00i), covered in resist (refractive index 1.68-0.01i). The resist was assumed to 'semi-planarise' the surface topography, ie. the step height on the surface of the resist was assumed to be half of the step height in the oxide layer. The oxide thicknesses and sidewall angles for these runs are given in Table 5.2.

Figure	Oxide thickness (μm)	Sidewall angle (degrees)
5.7	0.44	75
5.8	0.44	90
5.9	0.45	75
5.10	0.45	90

Table 5.2: Oxide thicknesses and sidewall angles for Figures 5.7–5.10.

As can be seen by comparing the figures, varying the sidewall angle by as much as 15° (a large variation for any given process step) changes the line profile hardly at all. This is comforting from the point of view of process control, since variations in sidewall slope due to process conditions (eg. defocus of the image in a lithography step, or changing gas mixtures during plasma etching) should then have little effect on the alignment signal during subsequent lithographic steps.

A variation in the oxide thickness of ~2%, however, has a large effect on the profile, as is seen by the increased intensity diffracted into the side fringe in Figures 5.9 and 5.10. Since control of most deposition steps of a process to within 2% is unlikely to be achieved in the near future in production environments, it appears that we must be prepared to accept large variations in alignment signal profiles, due to the effect of film thickness variations. This will certainly be the case for materials such as silicon dioxide, for which absorption within the film is negligible. For an aluminium layer, however, the energy density of the incident illumination will have fallen to $1/e$ of its original value at a distance d into the metal, given by:

$$d = \frac{\lambda_0}{4\pi n\kappa} \tag{5.12}$$

where λ_0 is the free space wavelength and $n\kappa$ is the imaginary part of the refractive index. For $\lambda_o = 589$nm, the value of $n\kappa$ is 5.23 [90]. This gives a value of d at this wavelength of ~9nm. Thus for any realistic thickness of metal, the reflectivity in the visible spectrum will be independent of the layer thickness. Care should be taken, therefore, in drawing conclusions about which layer thicknesses will affect the alignment profile. Dielectric thicknesses will only affect profiles as long as there is no metal on top of the non-absorbing medium.

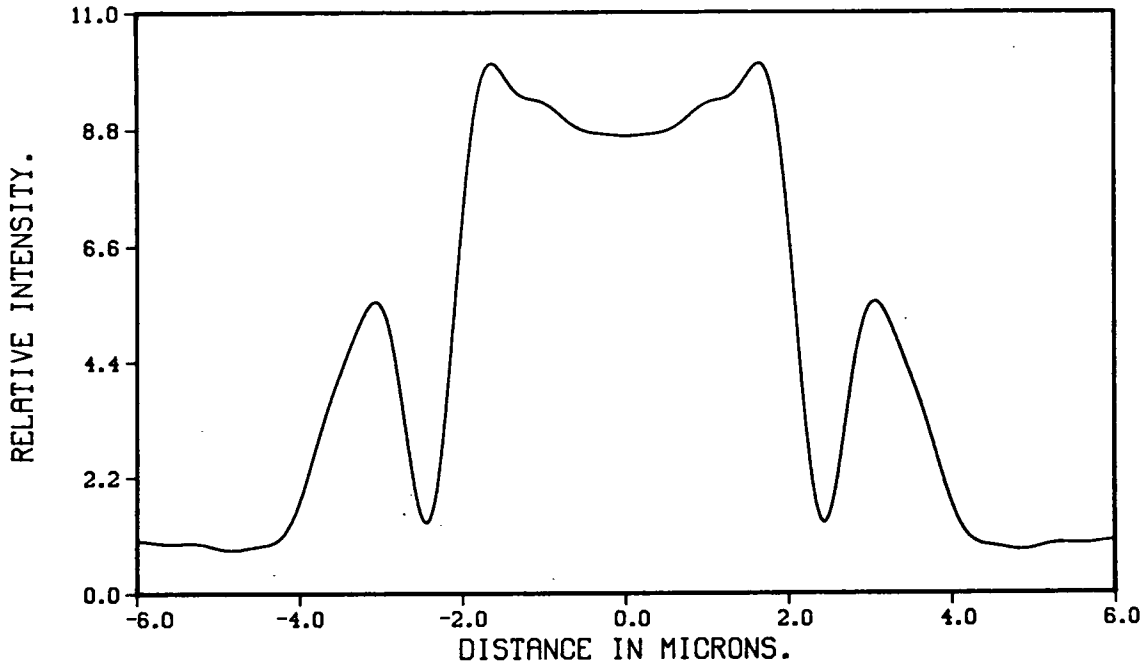

Figure 5.7: Output intensity profile from NVEW. See Table 5.2 for oxide thickness and sidewall slope.

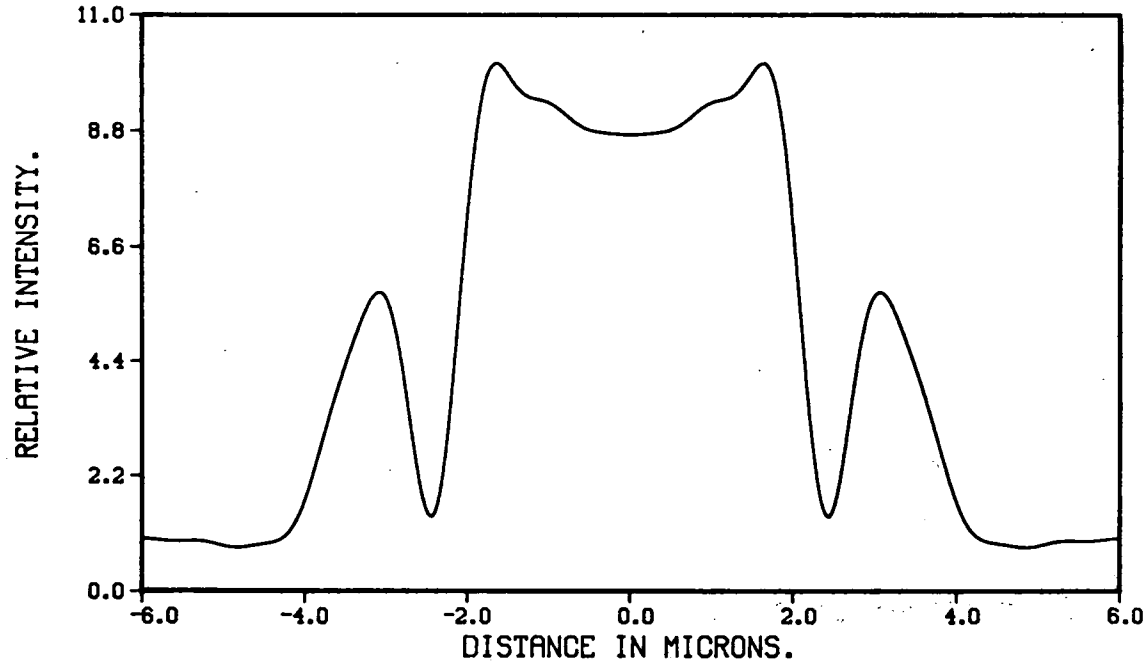

Figure 5.8: Output intensity profile from NVEW. See Table 5.2 for oxide thickness and sidewall slope.

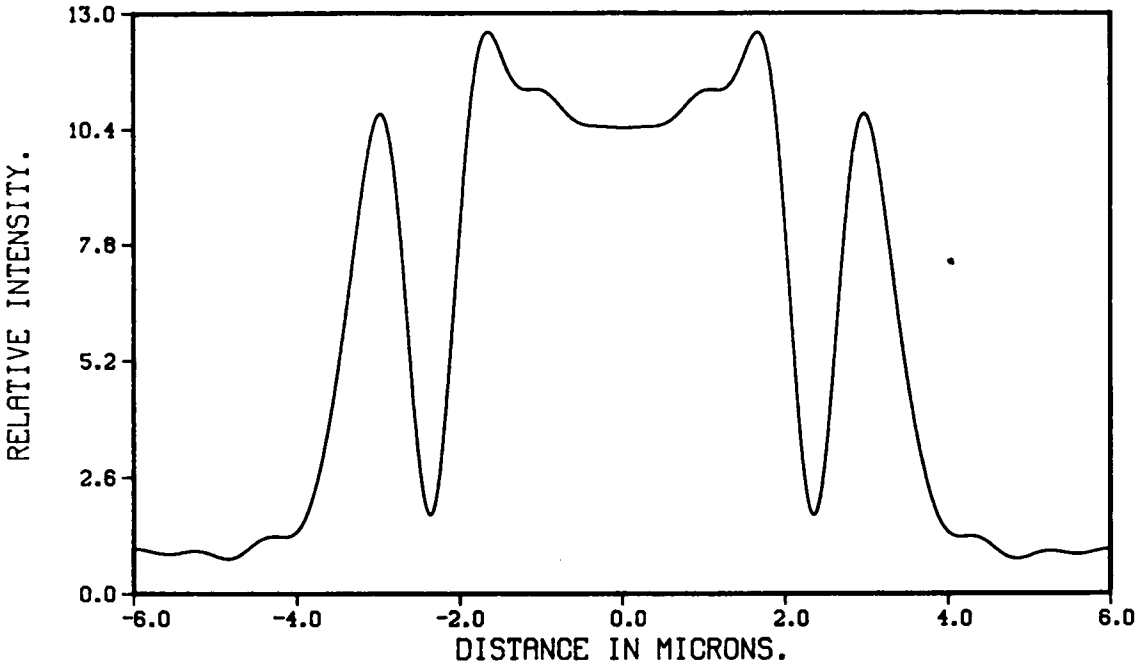

Figure 5.9: Output intensity profile from NVEW. See Table 5.2 for oxide thickness and sidewall slope.

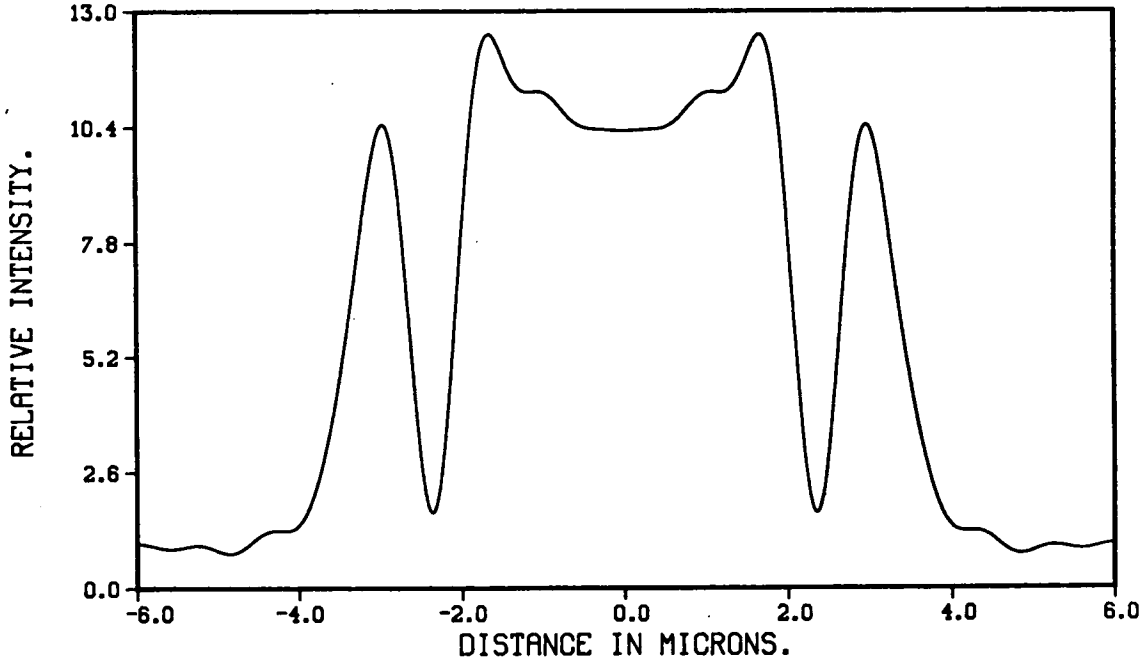

Figure 5.10: Output intensity profile from NVEW. See Table 5.2 for oxide thickness and sidewall slope.

Variation in Linewidth and Layer Thickness.

Figures 5.11–5.14 show line profiles obtained from similar structures to those in Figures 5.7–5.10, this time with the linewidth varying. The oxide thickness and gap width for these profiles is given in Table 5.3.

Figure	Oxide thickness (μm)	Oxide gap (μm)
5.11	0.44	4.0
5.12	0.44	8.0
5.13	0.45	4.0
5.14	0.45	8.0

Table 5.3: Oxide gap and thickness for Figures 5.11–5.14.

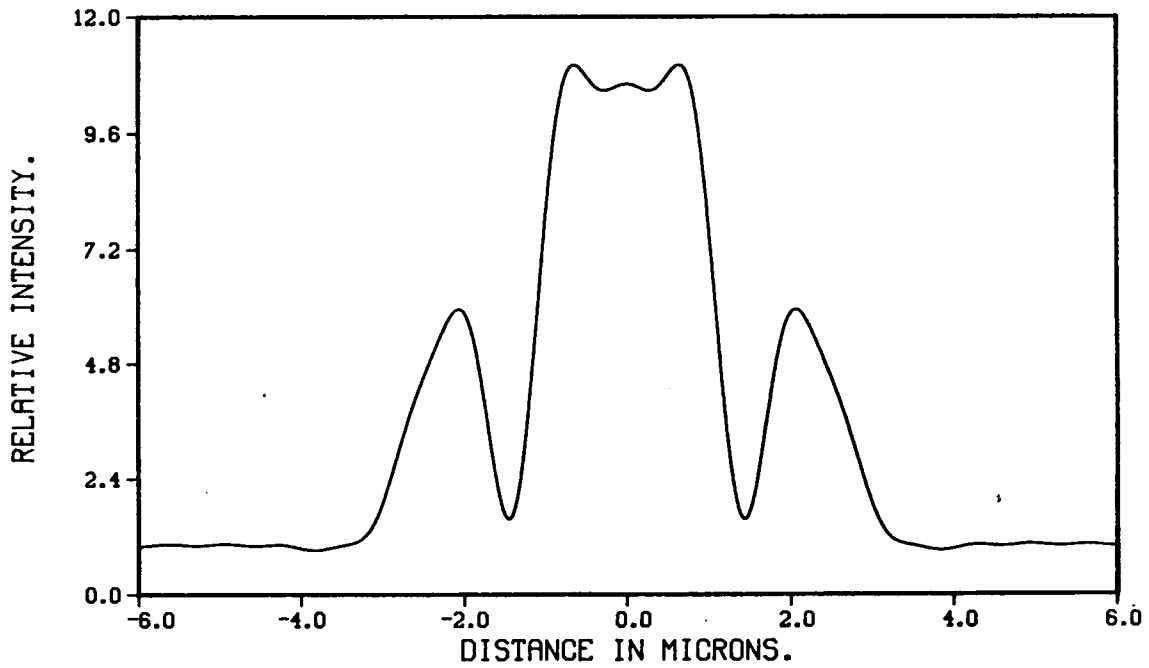

Figure 5.11: Output intensity profile from NVEW. See Table 5.3 for oxide thickness and oxide gap.

Comparison of the appropriate plots (compare 5.11 and 5.12 with 5.8, 5.13 and 5.14 with 5.10) shows that, for a given oxide thickness, changing the linewidth by up to 2μm does not significantly affect the edge fringing, it simply moves the fringes closer together or further apart. This is probably because,

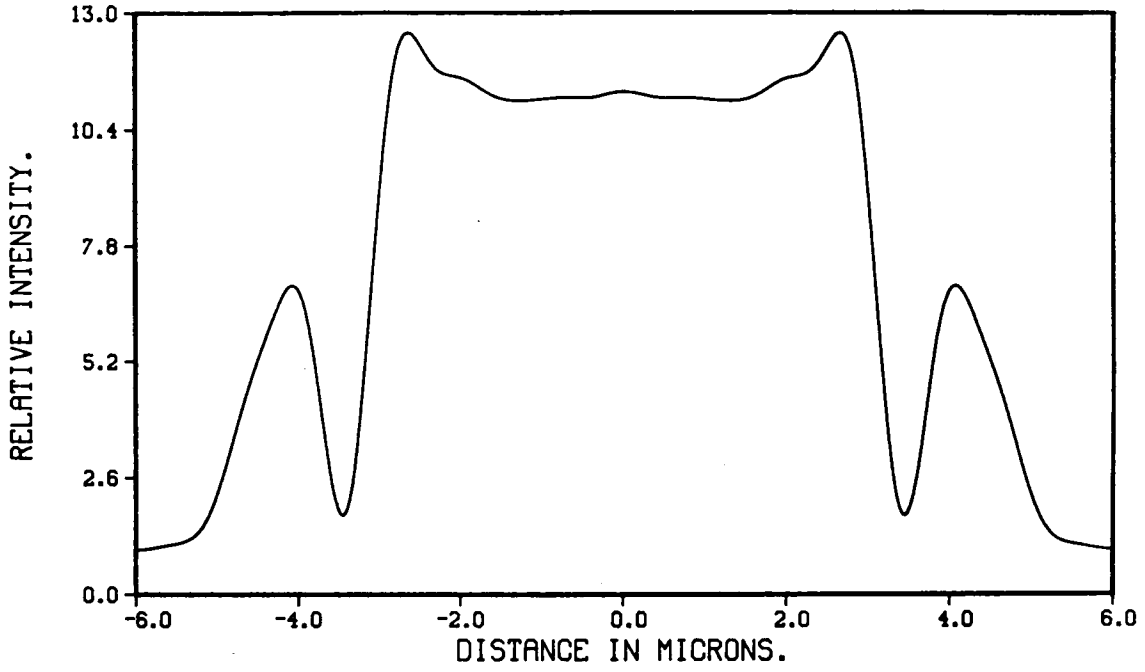

Figure 5.12: Output intensity profile from NVEW. See Table 5.3 for oxide thickness and oxide gap.

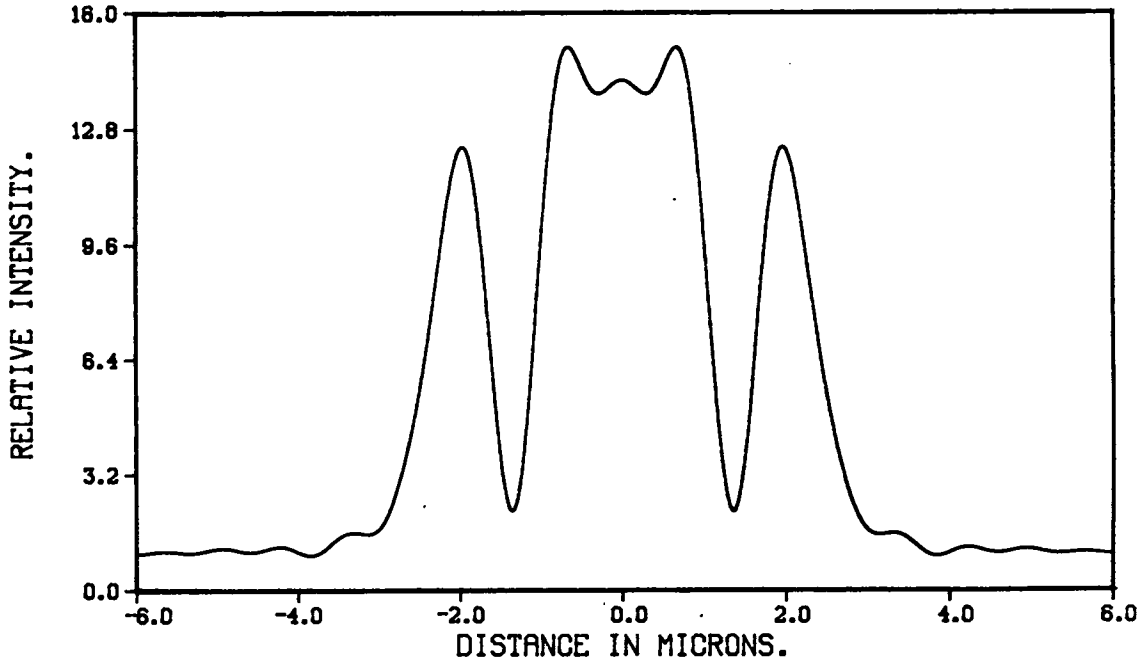

Figure 5.13: Output intensity profile from NVEW. See Table 5.3 for oxide thickness and oxide gap.

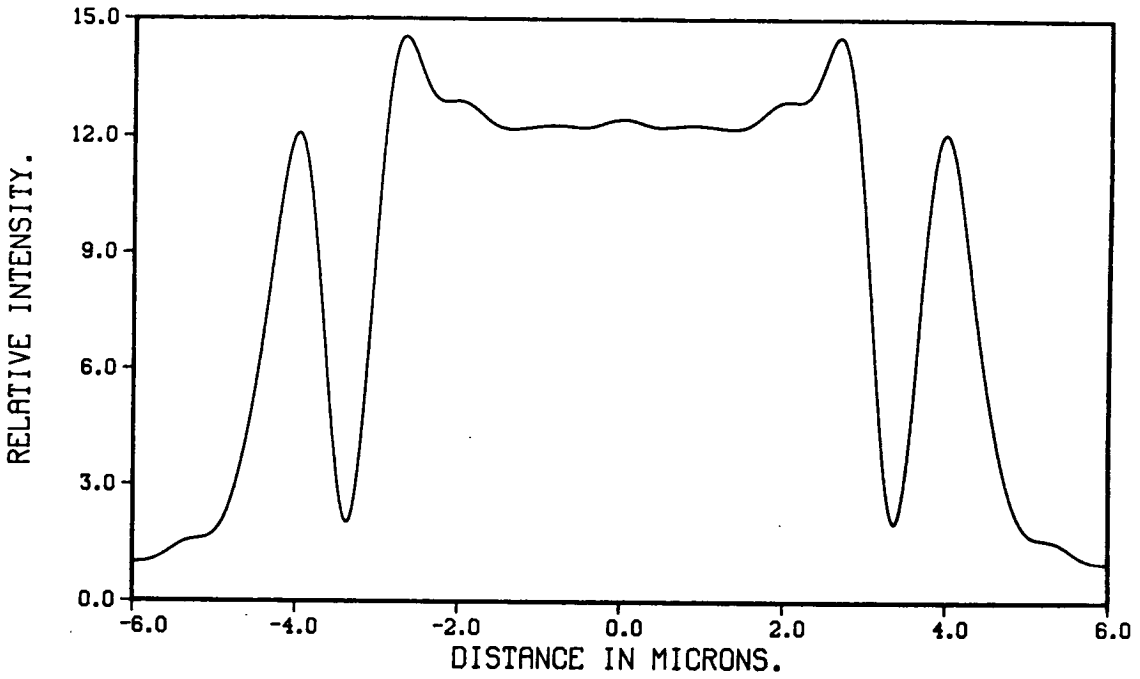

Figure 5.14: Output intensity profile from NVEW. See Table 5.3 for oxide thickness and oxide gap.

with an oxide gap of 4.0μm, the edges of the line are still separated by a distance l such that $l \gg \lambda$, where λ is the illumination wavelength. With an oxide gap of 8.0μm, the edges of the line are actually separated by a distance of 4.0μm (remembering that a periodicity of 12.0μm is assumed for all structures), resulting in very little cross-coupling of the radiation diffracted from each edge of the gap.

Figures 5.15–5.18 show the intensity profiles as the linewidths are varied further (the oxide gaps are given in Table 5.4, the oxide thickness is 0.45μm). Comparison of Figures 5.15 with 5.16 and 5.17 with 5.18 demonstrates that, as the edges of the oxide gap approach a separation which is comparable with the illumination wavelength, the edge fringing becomes strongly dependent on the width of the wafer feature, due to coupling of the radiation diffracted from each edge. This suggests that modification of reflected intensity profiles may be possible by controlling finely the width of a wafer target, but only as the width of the target approaches the resolution limit of the alignment optics.

As the Optimetrix is configured at present, the alignment optics and exposure optics are one and the same[†], thus targets approaching the resolution of the alignment system are also at the limits of the exposure system. As was mentioned in Chapter 2, linewidth control is non-optimal in this case, due to low contrast

[†]A spot lamp, rather than the main exposure lamp, is used to illuminate the alignment mark. The condenser and projection optics are shared by these two lamps, however.

Figure	Oxide gap (μm)
5.15	1.0
5.16	2.0
5.17	10.0
5.18	11.0

Table 5.4: Oxide gap for Figures 5.15–5.18.

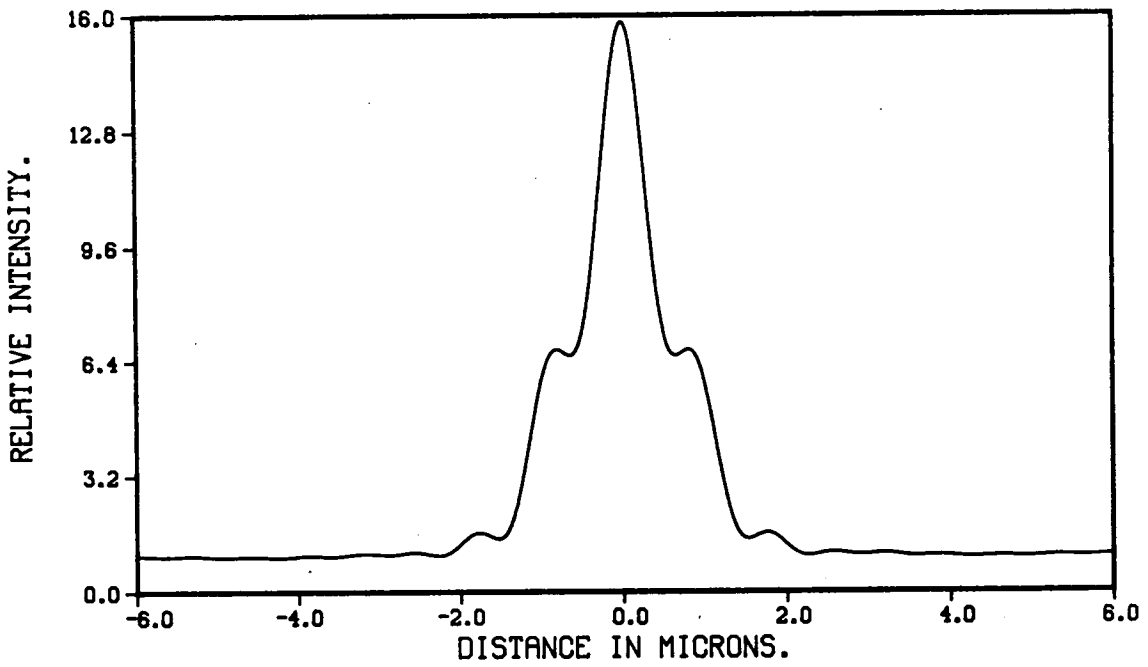

Figure 5.15: Output intensity profile from NVEW. See Table 5.4 for oxide gap.

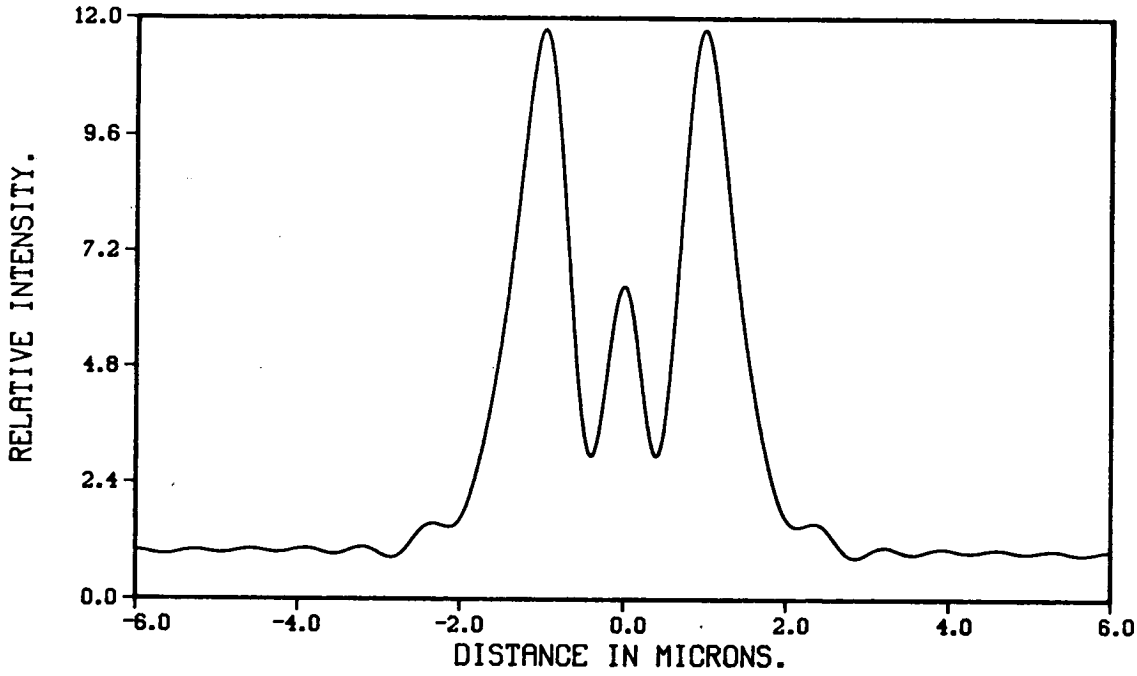

Figure 5.16: Output intensity profile from NVEW. See Table 5.4 for oxide gap.

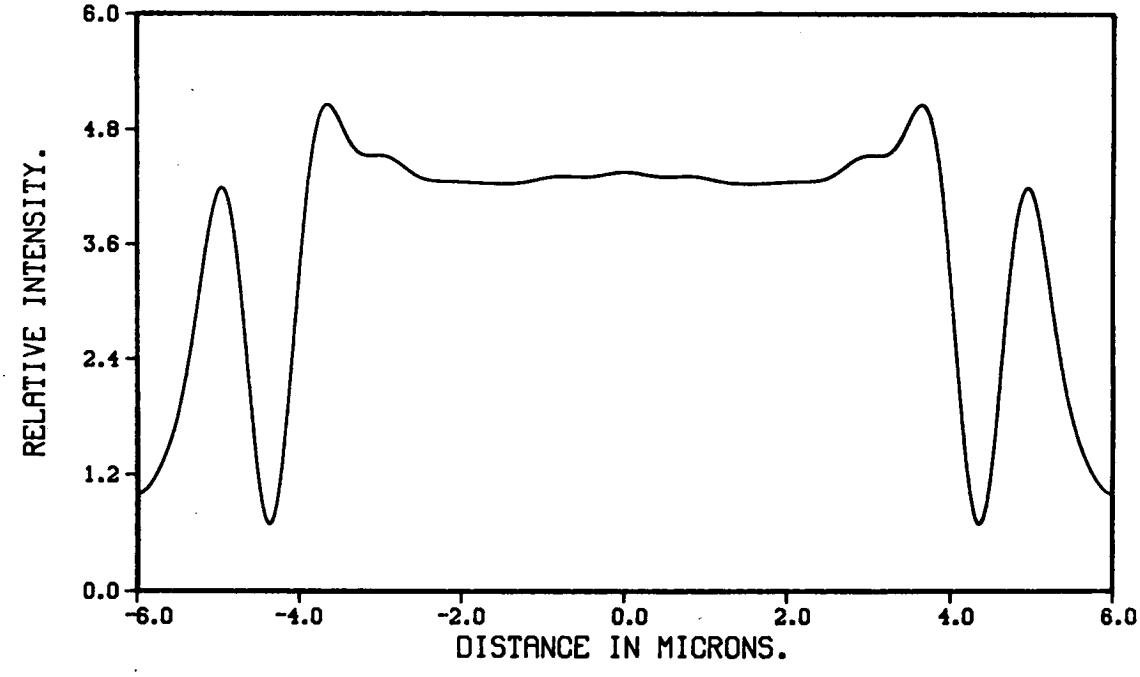

Figure 5.17: Output intensity profile from NVEW. See Table 5.4 for oxide gap.

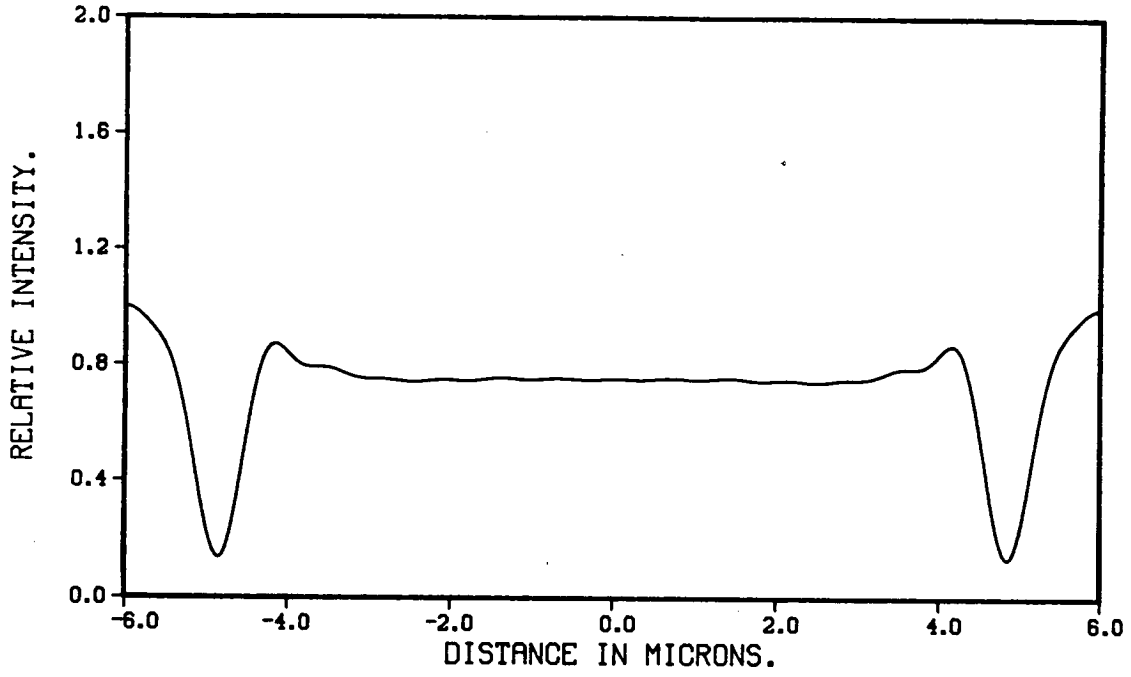

Figure 5.18: Output intensity profile from NVEW. See Table 5.4 for oxide gap.

and low edge slope in the aerial image. Since the reflected intensity profile is sensitive to linewidth variations in this regime (as we have just seen from Figures 5.15–5.18), this method of tailoring the alignment signal is to some extent self-defeating, since we must always expect changes in the signal due to variations in feature size at small linewidths. If, however, the alignment optics were to use a lower numerical aperture than the exposure optics, the fringes could be removed from the alignment signal, while still using targets which are well within the resolution capability of the main lens. There are two ways of accomplishing this:

1. Use of a separate lens to perform the fine alignment. An additional lens (such as the off-axis lens already present on the Optimetrix) could be used to perform the alignment. This option is not advisable, however, due to the added complication of base-line drift (instability of the position of the off-axis lens with respect to the main lens). One of the advantages of the Optimetrix is its ability to align on-axis, and expose without moving the stage, thus eliminating one source of overlay error. It would be unwise to sacrifice such an advantage.

2. Use of spatial filtering at an appropriate position in the lens to block out the higher diffracted orders. A simple circular diaphragm placed at an appropriate position in the lens during alignment would result in the desired reduction of the numerical aperture, thus reducing its resolution capability.

The diaphragm could then be removed during exposure.

Figures 5.19–5.22 show the output profiles from NVEW for 2.0μm and 10.0μm lines in oxide, using numerical apertures of 0.16 and 0.08 for the alignment optics (Table 5.5 shows the width of oxide gap and numerical aperture in each case). As can be seen by comparing the plots with Figures 5.16 and 5.17 (same wafer features, with NA=0.32), the lower numerical aperture lens partially filters out the fringes in the output profile. This approach is similar in principle to the software filtering used to reduce fringing in Chapter 4, although in this case the filtering is non-linear, due to the partially coherent nature of the optics.

For the 2.0μm gap case, there is an advantage in reducing the numerical aperture to 0.16, in that the central fringe evident in Figure 5.16 has disappeared in Figure 5.19. There is little extra advantage, however, in going to an numerical aperture of 0.08, since this results only in a reduction in image contrast (as is evident from the intensity scales on the plots), and also in a reduction of fringe edge slope. For the 10.0μm case, the 0.16 NA alignment optics have only reduced the size of the major fringe in the signal, while the 0.08 NA optics have removed this fringe altogether. Again in this case, the image contrast and edge slope have been substantially reduced by the 0.08 NA optics. This will not necessarily adversely affect an alignment system which is designed to be sensitive to mark symmetry, as is the Optimetrix.

Figure	Oxide gap (μm)	Numerical aperture
5.19	2.0	0.16
5.20	2.0	0.08
5.21	10.0	0.16
5.22	10.0	0.08

Table 5.5: Oxide gap and alignment optics numerical aperture for Figures 5.19–5.22.

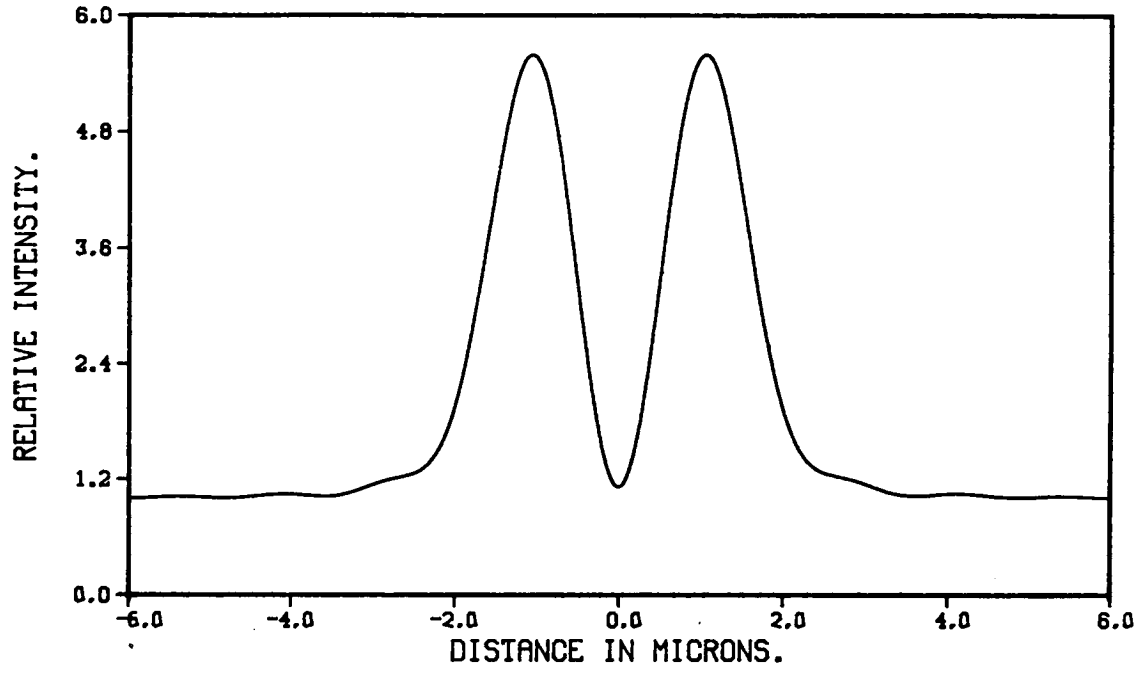

Figure 5.19: Output intensity profile from NVEW. See Table 5.5 for oxide gap and alignment numerical aperture.

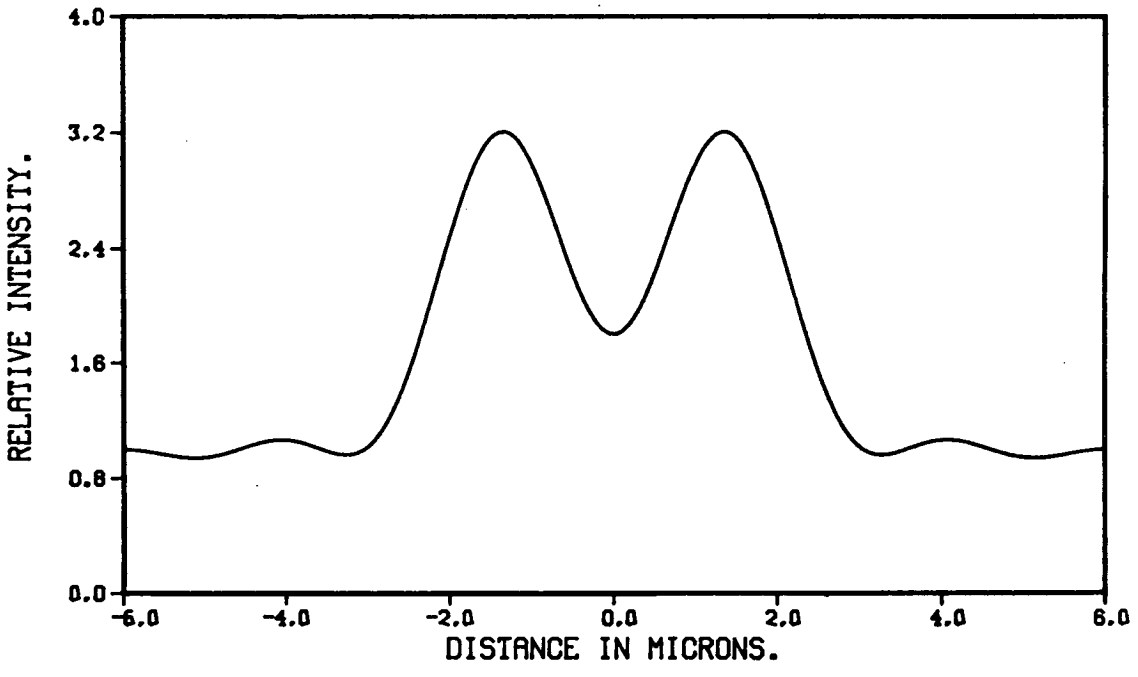

Figure 5.20: Output intensity profile from NVEW. See Table 5.5 for oxide gap and alignment numerical aperture.

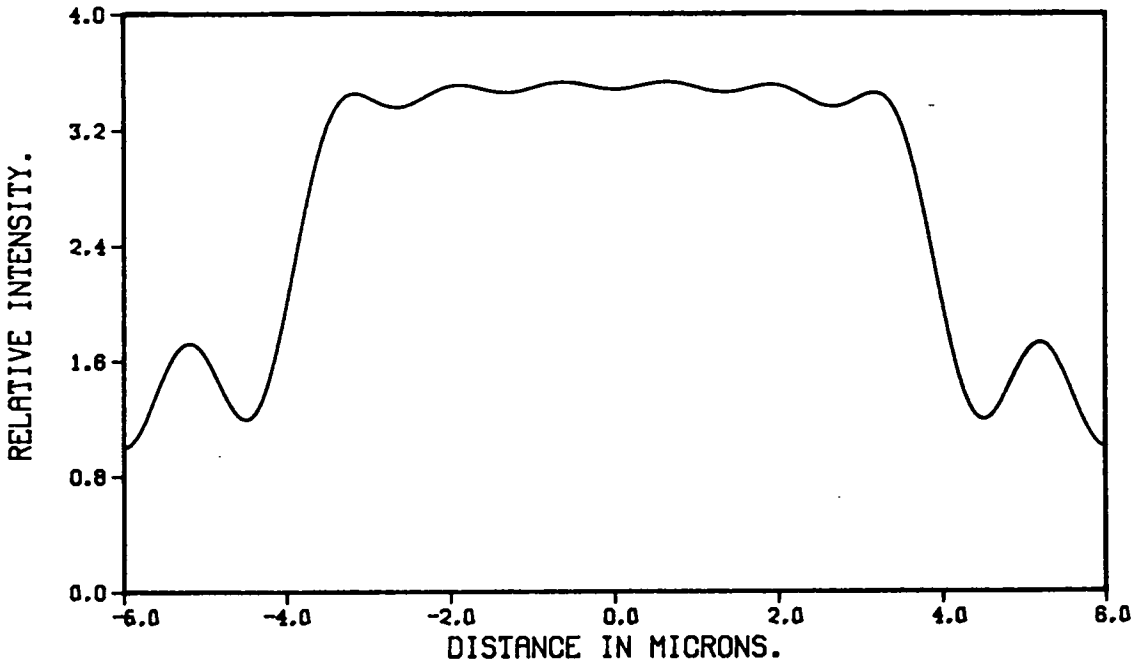

Figure 5.21: Output intensity profile from NVEW. See Table 5.5 for oxide gap and alignment numerical aperture.

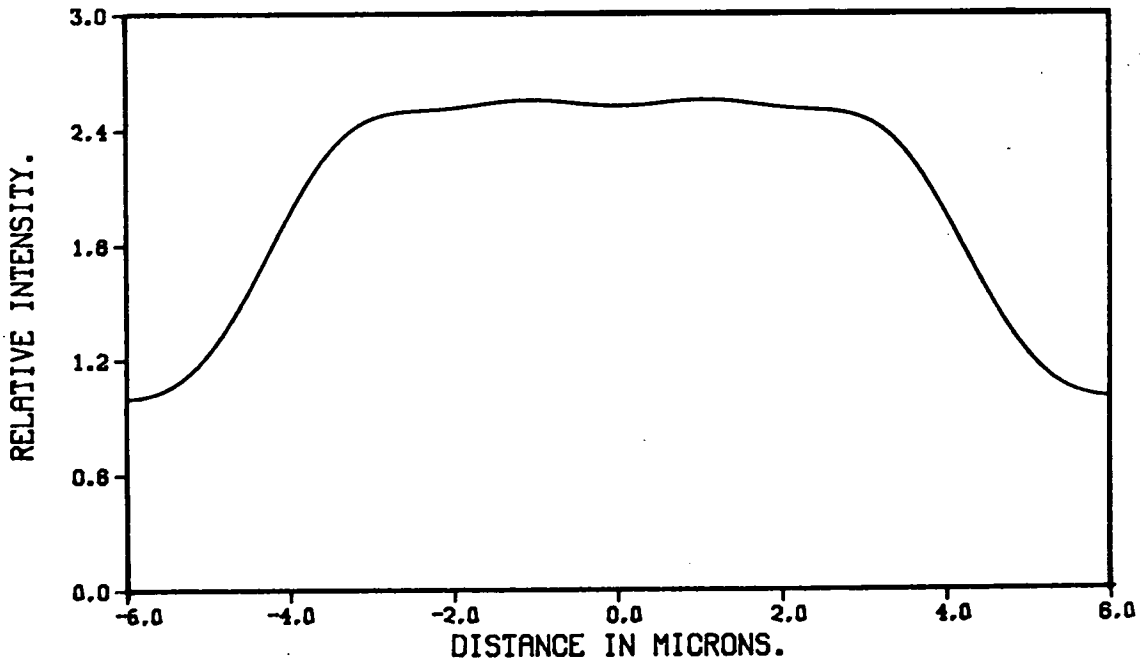

Figure 5.22: Output intensity profile from NVEW. See Table 5.5 for oxide gap and alignment numerical aperture.

5.3 Conclusions.

It appears from the calculations performed that there are two ways in which the fringe pattern may be modified by processing, to reduce edge fringing :

1. Modifying the layer thicknesses.

2. Reduction of the feature width to approach the resolution limit of the system.

Both of these methods have their drawbacks :

1. The desired electrical characteristics of a completed circuit (eg. threshold voltage or interconnect capacitance) normally determine the required thickness of any given device layer. Depending upon how critical the layer thickness is on any particular layer, there may be some room left for adjustment in order to optimise optical characteristics (for alignment or linewidth control). However, since we have seen that variations of only 2% can be responsible for large changes in the reflected signal, it is unlikely that layer thicknesses could be controlled finely enough to obtain the desired result. There is one important corollary to be considered here; using a reflective system we must always expect variations in the intensity profile across a wafer, from wafer to wafer, and from batch to batch, due to layer thickness variations. The alignment system must, therefore, be designed to allow for this fact.

2. Reducing the feature width of a mark also reduces the amount of fringing in the profile. Unfortunately, it also reduces the ability of the system to reproduce the feature faithfully, with good linewidth control. Since the fringe pattern depends heavily upon the feature size in the diffraction limit, and linewidth control is relatively poor in this area, we must again expect variations in the reflected intensity profile when using this method of tailoring. One way of getting around this problem would be to use alignment optics with low numerical aperture to filter out alignment fringes, while still using markers well within the resolution capability of the exposure system. This method can be thought of as a hardware equivalent of the software filtering demonstrated in Chapter 4, and has the advantage of being parallel in nature. The diaphragm accomplishes the filtering in effectively zero time, with the only throughput penalty being incurred by the repositioning of the stop into, and out of, the optical path.

Chapter 6

The Effect of Alternative Targets on the Optimetrix Alignment Accuracy.

It has already been shown in Chapter 5 that reducing the width of an alignment mark to a value close to the resolution limit of an optical system can reduce the amount of fringing in the reflected intensity profile. With this thought in mind, a reticle was designed to test different widths of alignment mark on the Optimetrix stepper. In addition to different mark widths, alternative configurations of target were also designed, while still retaining two vital ingredients:

1. The basic chevron structure.

2. Mirror symmetry in each arm of the chevron.

These marks included split structures and wedge structures of varying dimensions (see Figure 6.1 for details of these types of mark). The dimensions of the marks which were designed are shown in Table 6.1 (for a definition of the various sizes specified in the table, see Figure 6.1). Although these markers were designed specifically for the Optimetrix stepper, the same types of markers could equally well be tried out on other steppers using a similar alignment system. Bearing in mind the results from the simulation study in Chapter 5, it was hoped that a split structure mark (with the two halves of the structure being separated by 6.0μm) would have an intensity profile very similar to the one in Figure 3.6, with good edge definition and very little fringing. This is borne out in Figures 6.2–6.5, which show the reflected intensity profiles obtained when aligning poly-silicon to active area, for the 6.0μm standard, 3.0μm standard, 1.5μm standard, and the 1.0μm split mark. Figure 6.6, which is included for reference, shows a contour

Mark type.	st (μm)	sp (μm)	w1 (μm)	w2 (μm)
Standard.	6.0	n/a	n/a	n/a
	5.0	n/a	n/a	n/a
	4.0	n/a	n/a	n/a
	3.0	n/a	n/a	n/a
	2.0	n/a	n/a	n/a
	1.5	n/a	n/a	n/a
	1.0	n/a	n/a	n/a
Split.	n/a	2.0	n/a	n/a
	n/a	1.5	n/a	n/a
	n/a	1.0	n/a	n/a
Wedge.	n/a	n/a	18.0	0.0
	n/a	n/a	16.0	1.0
	n/a	n/a	14.0	2.0
	n/a	n/a	12.0	3.0

Table 6.1: Dimensions of mark details on test reticle (with reference to Figure 6.1).

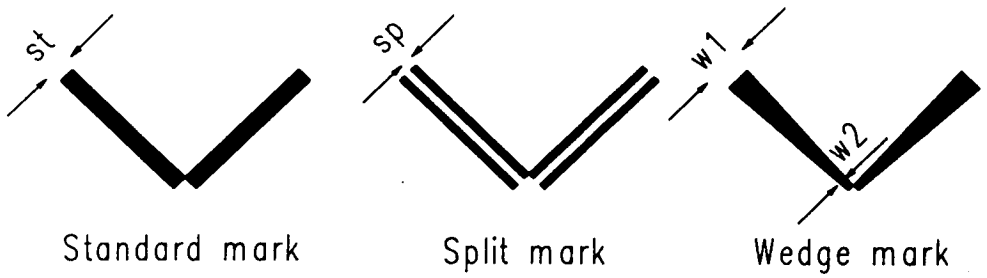

Figure 6.1: Different types of marker tried out on the Optimetrix stepper.

plot of the auto-correlation function obtained from Figure 6.5, illustrating a single peak in the region of interest[†].

The wedge mark presents a slightly different approach to the reduction of fringing. With reference to Figure 6.1, we can consider a number of scans performed in the x-direction, each one at a slightly different position in the y-direction. Each scan "sees" a slightly different linewidth, and hence sees fringes at a slightly different position. After integration over a number of scans, the fringes will to some extent cancel each other out, and hence should make the alignment more reliable. This integration may be achieved easily by the use of the LNS parameter in DIA CAMERA (LNS is the number of lines over which the data has to be integrated).

[†]There are in fact two additional peaks on the plot, at some ± 20 pixels from the centre of the window. This is equivalent to over $\pm 4.0\mu$m, which means that these peaks are outside the capture range of the system.

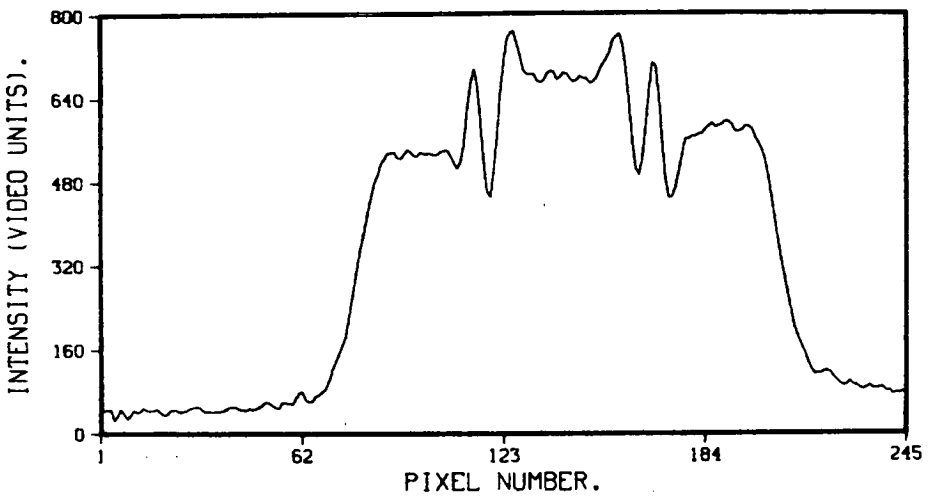

Figure 6.2: Video data plot of standard 6.0μm mark.

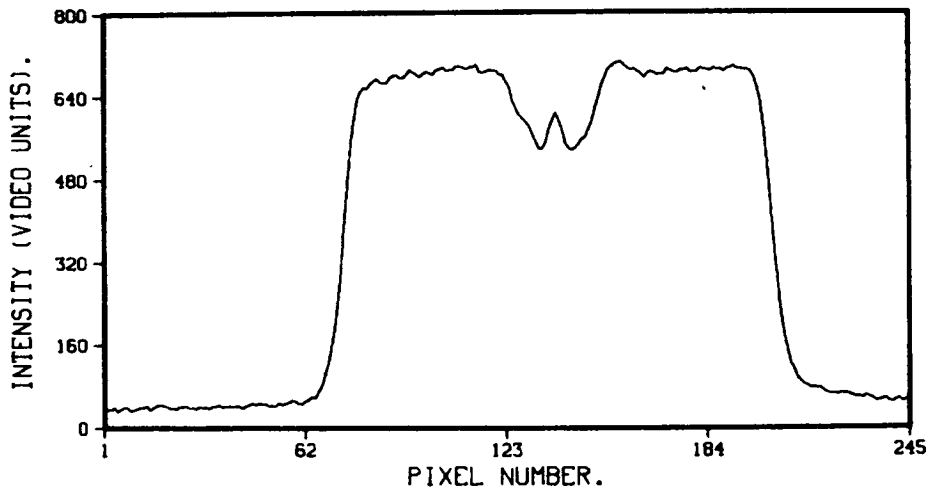

Figure 6.3: Video data plot of standard 3.0μm mark.

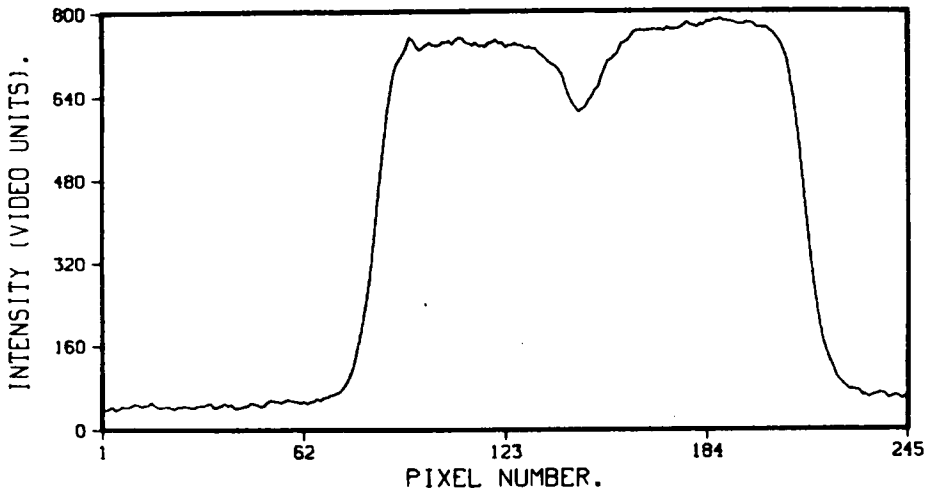

Figure 6.4: Video data plot of standard 1.5μm mark.

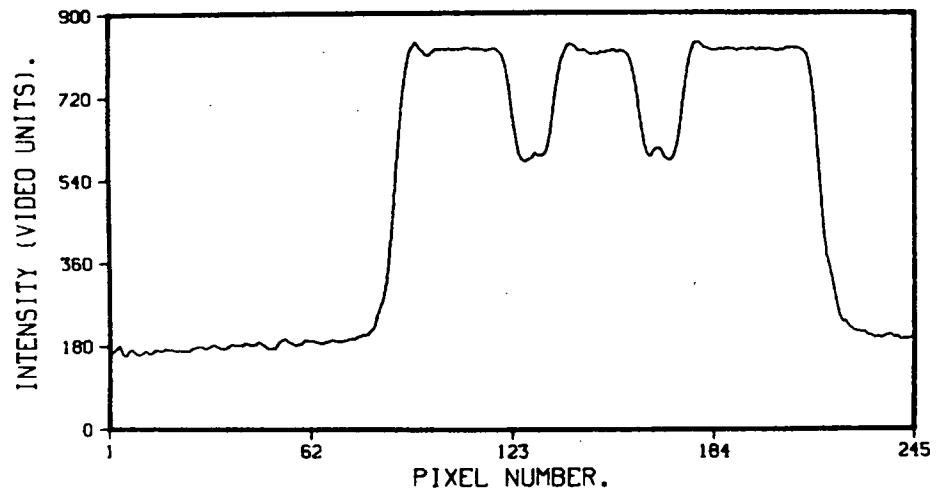

Figure 6.5: Video data plot of split 1.0μm mark.

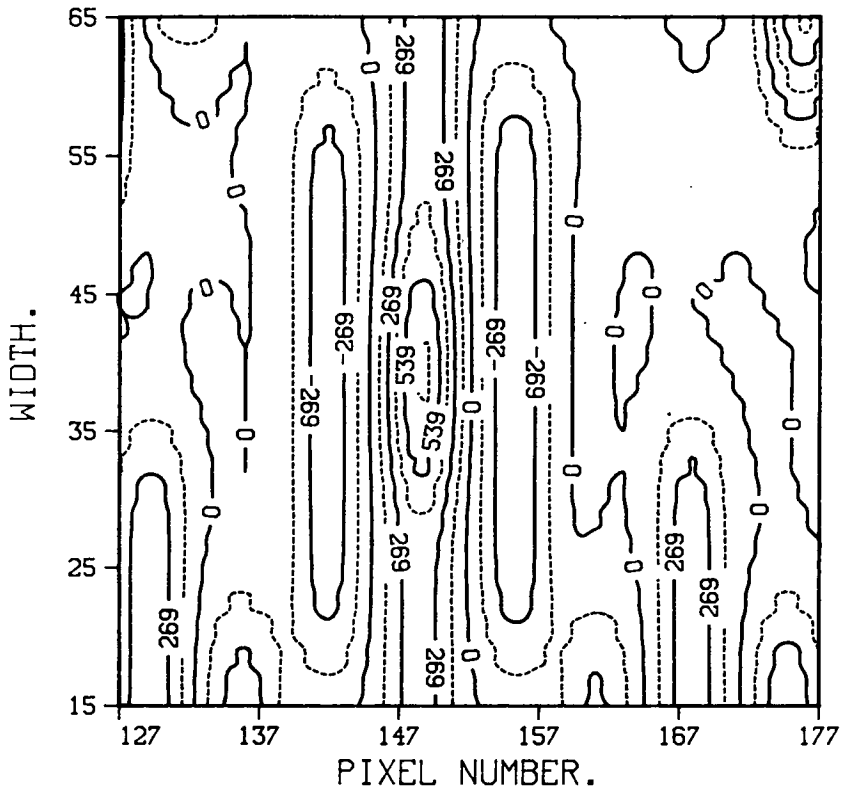

Figure 6.6: Contour plot of auto-correlation function, from the split 1.0μm mark.

6.1 Data Collection, Processing, and Experimental Procedure.

The reticle which was designed for this experiment consisted of a number of different markers, each offset from a reticle window by 500μm$\times$500μm. In addition to the markers to be tested, four extra (standard 6.0μm) markers were also included, along with four reticle markers, such that each combination of superimposed light and dark field wafer marks, and light and dark field reticle marks, would be produced when the mask was printed twice on a wafer, with an offset of 500μm$\times$500μm between each layer (see Figure 6.7 for the layout of these markers on the reticle). These four superimposed combinations shall be referred to as reference markers for the rest of this chapter.

The procedure which was used to evaluate the accuracy of the various markers followed closely an EMF 1.5μm NMOS process run, the first lithographic step being definition of the active area. This layer was printed on 50 wafers, with 144 exposures on each of 42 of the wafers, and 32 exposures on each of the remaining 8. After processing of this layer, the second photo-stage was performed (buried

contact), using the same mask and stepping file, with an offset of 500μm$\times$500μm incorporated into the stage position. In this way the buried contact layer could be aligned to active area, using any of the wafer targets defined at the previous layer. In fact, six wafers were each aligned to seven different types of marker (of the markers which had been made on the reticle, only those listed in Table 6.2 were chosen as suitable candidates for use during the experiment). The remaining eight wafers (those with only 32 exposures at the first layer) were blind stepped, after global alignment and alignment to the first die-by-die marker. Using this procedure, the four superimposed combinations (reference markers referred to above) were created automatically.

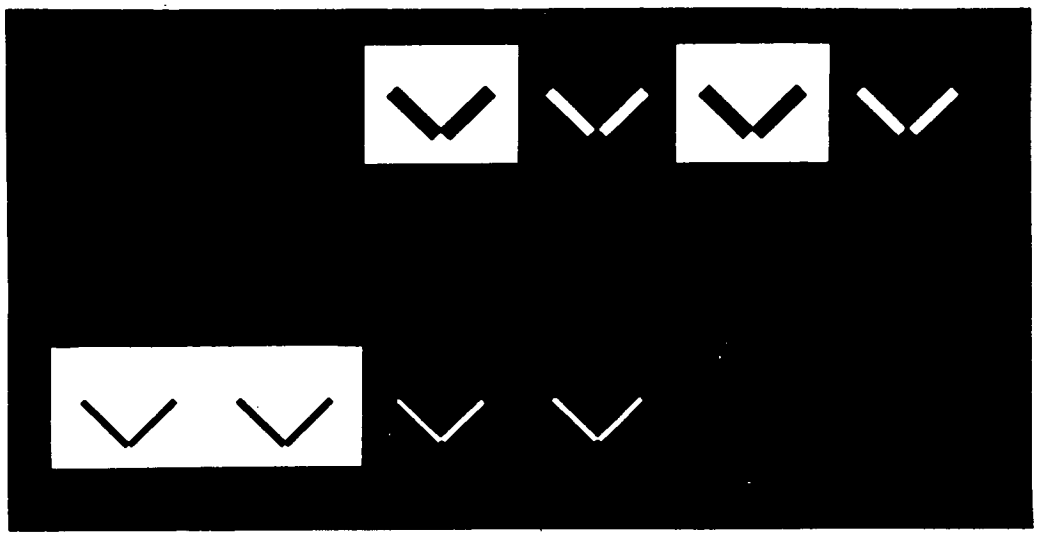

Figure 6.7: Layout of reference markers on reticle.

After development, the wafers were placed back on the stepper and the misalignment was read using one of the four reference markers. It was found that at least one reference marker always exhibited high contrast with very little fringing (a combination in which confidence could be placed in the machine to read the misalignment accurately). Measurement of misalignment using these types of marks has been shown to be considerably more reliable than alignment to marks in which fringing is present [91]. In fact, under these conditions, the performance of the system will be close to ideal. Optimetrix claim an accuracy of $\sim$0.2μm (3σ) for the alignment system, and since manufacturers tend to quote accuracy

Mark type.	st (μm)	sp (μm)	w1 (μm)	w2 (μm)
Standard.	6.0	n/a	n/a	n/a
	3.0	n/a	n/a	n/a
	1.5	n/a	n/a	n/a
Split.	n/a	1.5	n/a	n/a
	n/a	1.0	n/a	n/a
Wedge.	n/a	n/a	18.0	0.0
	n/a	n/a	12.0	3.0

Table 6.2: Mark types which were chosen for experimental evaluation.

under ideal conditions, the error in the readings in this experiment will be taken to be 0.07μm (1σ).

In addition to the wafer and reticle markers, optical verniers were printed on the eight wafers which had been blind stepped. On each of these wafers the optical verniers were inspected to determine the overlay at each of the 32 sites, and the results were compared with the measurements taken by the machine.

A program called VECPLO (**VEC**tor **PLO**t) was written on the EMF VAX, especially for the purpose of drawing vector maps, in order to present the results of the measurements in a concise and meaningful way. This program takes in data files which can be transferred from the stepper to the VAX using a software switch in COM DEB (machine software location S922 = 5). With this software switch set, the machine will evaluate the misalignment, and send the data to the RS232 port at the back of the machine, which had been connected up to the VAX to read in the data. No attempt is made to correct the misalignment by moving the stage when this switch is set. VECPLO reads the data in, and statistical analysis is performed, using a least squares fit to the data to obtain estimates of the translation, rotation and global wafer expansion [92]. The vector map may then be drawn with the aid of some standard Fortran graphics subroutines.

The program is quite general in that it may quickly be converted to plot misalignment maps from overlay data of any format. The user must provide two custom written subroutines, one to read in the data in the correct format, and one to scale this data, into millimetres for the position of a particular die, and into microns for the misalignment at that position. In this way overlay maps may be very quickly generated from data extracted from, for example, an automatic

prober which measures misalignment using electrical structures.

Figures 6.8 and 6.9 show vector maps, produced by the program, of the misalignment data taken from wafer number 47 at the depletion implant layer, for both the machine readout and the vernier readout. The maps show a good correspondence with one another, with both showing a significant clockwise rotation of the second photo-stage with respect to the first. Figures 6.10 and 6.11 show two maps (machine and vernier read data from wafer number 48, at the depletion implant layer) which do not, on first inspection, demonstrate particularly good correspondence (even after the difference in scale marker is taken into account). Inspection of the statistical data shows that the two are fairly close, however, with a difference in translation of 0.043μm in the y-direction, and 0.022μm in the x-direction. This difference is reasonable, given that the least count of the vernier readings is 0.1μm, compared with 0.027μm for machine readings. The expansion factors differ by no more than 0.04μm/cm, and the rotations by only 0.01μm/cm, both of which are again reasonable given the difference in least count. In fact the random appearance of the overlay errors in the vernier read case, compared with the more regular distribution of errors in the machine read case, suggest that the machine read errors are more accurate, and that the least count of 0.1μm is a factor which limits the usefulness of optical verniers for this purpose.

The whole process of printing, developing, and reading overlay, was repeated for each photo-stage of the process, except for the metal stage. Due to the high reflectivity of the metal it was impossible to obtain a combination of light/dark field reticle and wafer markers which gave a high contrast and zero fringing, therefore no confidence could be placed in the machine readings.

6.2 Results

After each photo-stage, the average x and y-translations and x and y-standard deviations were calculated for each type of alignment mark. This involved calculating these four parameters from 120 measurements in x and y on each wafer, then averaging over the seven wafers which were exposed using each type of alignment mark. (It was mentioned earlier that 144 exposures were made on each wafer, however only 120 of these were measured, since the machine consistently failed to align to many of the 24 die around the outer wafer edge. This was probably due to different film thicknesses in this region compared with the rest of the wafer.) The expansion and rotation were not evaluated for the wafers which were

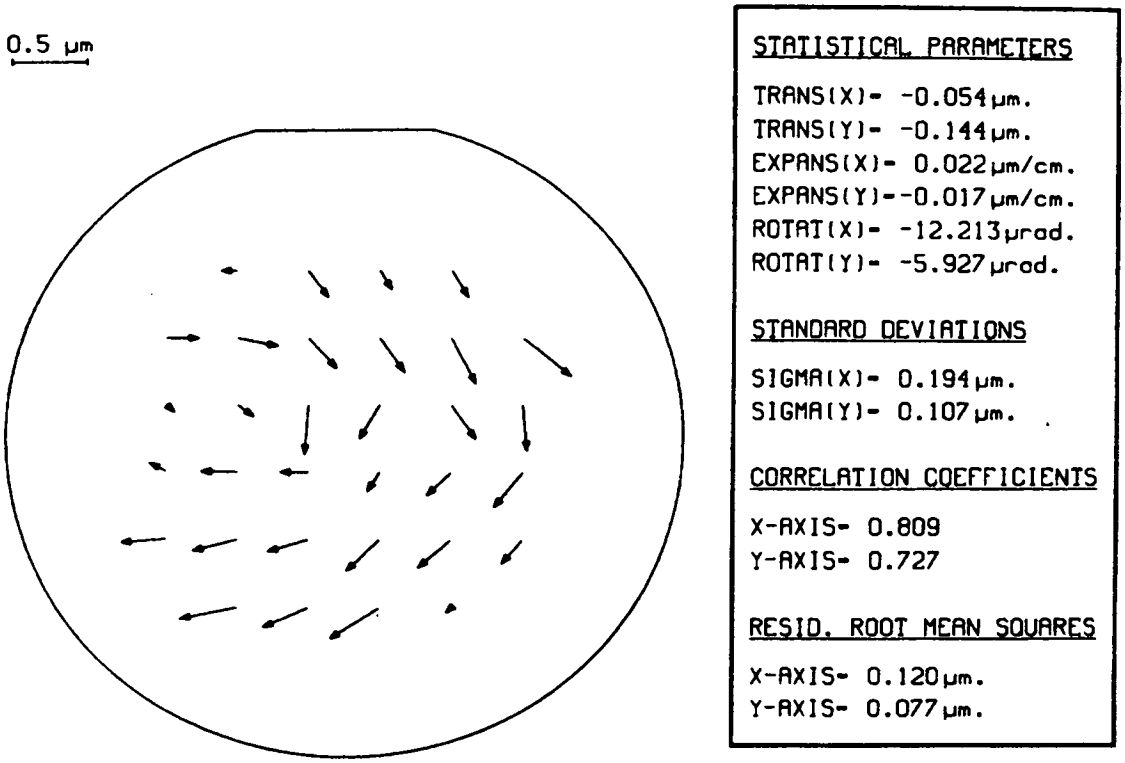

Figure 6.8: Vector map of machine read data from wafer 47.

aligned die-by-die, since these factors should not be significantly different from zero in this case. Since the measurement inaccuracy is approximately 0.07μm for each reading, averaging over 840 measurements gives a standard deviation of 0.002μm in the inaccuracy of the measurement of the x and y-translations (0.004μm in the case of the blind stepped wafers). This is negligible compared with the size of the misalignments being measured.

There are two distinct sources of overlay error to be considered when looking at the results from this experiment:

1. Misalignment due to multiple peaks in the auto-correlation. Peaks in the correlation function are generally separated by four pixels or more, leading to misalignments of $\sim$0.9μm or greater.

2. Misalignment due to other effects (asymmetry in the intensity profile, stage positioning accuracy, etc...). These errors are small compared with multiple peak errors, and are generally of a magnitude less than 0.25μm.

Thus any average errors of more than 0.25μm over a whole wafer are probably due to some die having multiple peak errors, and the marker used for alignment

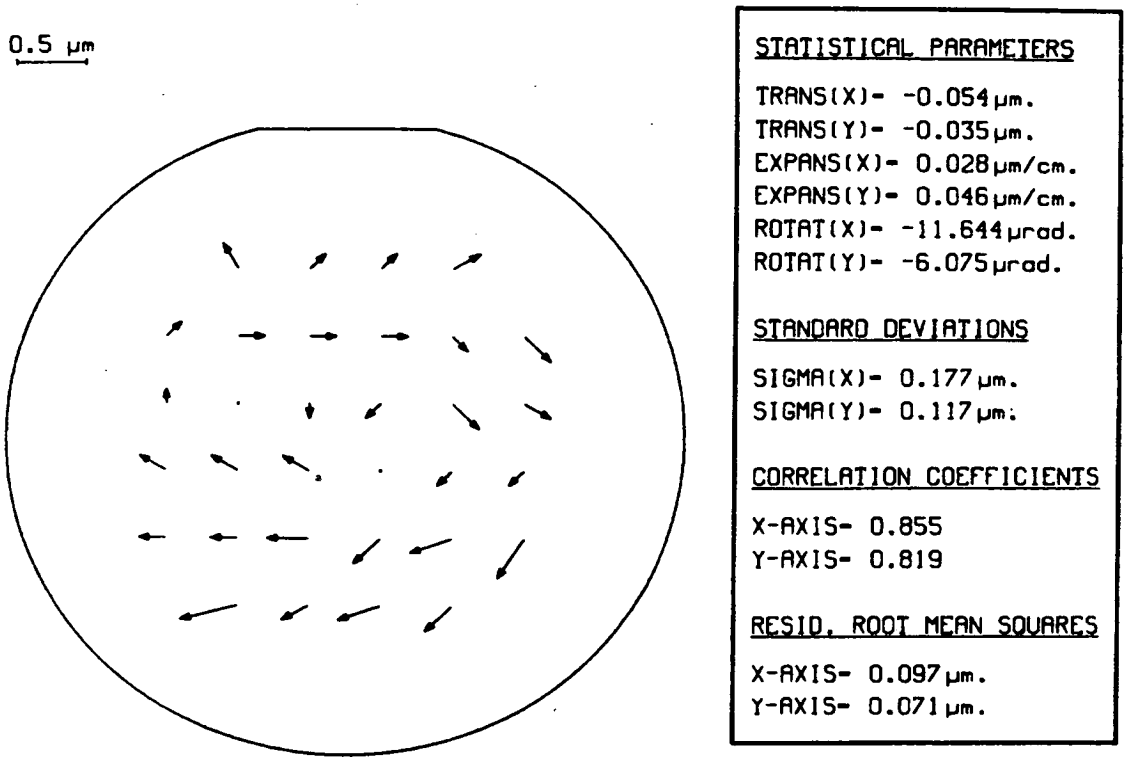

Figure 6.9: Vector map of vernier read data from wafer 47.

in that particular case should not be considered for that process layer. The situation when the mean misalignment is less than 0.25μm is more complicated. It may be that for such a marker, few, or no multiple peak errors have occurred. It could also be that a large number of multiple peak errors have occurred in one direction, with an approximately equal number of similar errors occurring in the opposite direction, thus cancelling each other out. If the latter is the case, the fact will be manifested in the form of a large standard deviation. In general, no σ value of much greater than 0.1μm can be tolerated for circuit features approaching 1.0μm. In summary, we define good alignment here as being when the x and y-translations are both less than 0.25μm, and the x and y-standard deviations are both less than 0.1μm.

6.2.1 Depletion Implant.

Figures 6.12–6.15 show the x-translation, y-translation, x-standard deviation and y-standard deviation for alignment of the depletion implant to active area. Only the 6.0μm standard mark, the 0–18μm wedge mark, and blind stepping fulfill our specification for average translations in x and y. In this case the 0–18μm mark

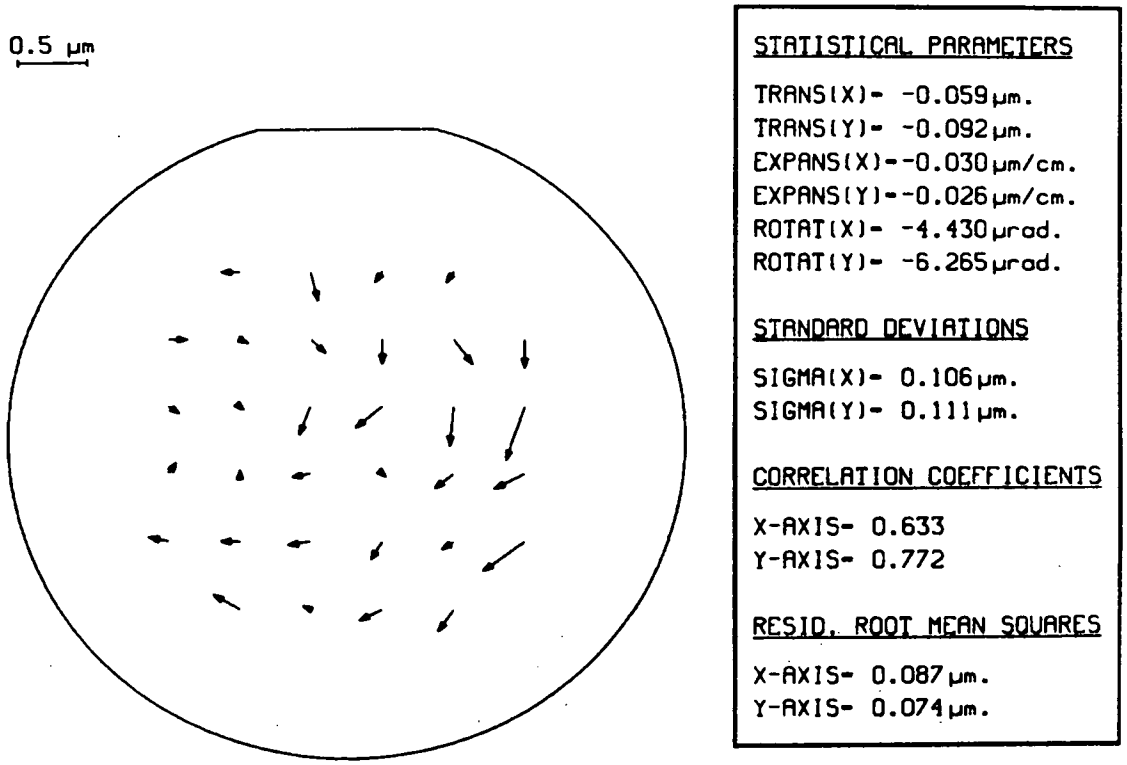

Figure 6.10: Vector map of machine read data from wafer 48.

can be ruled out due to a large σ value. The graphs of the standard deviations show that in this case, blind stepping actually out-performed all the markers which were used for die-by-die alignment. This can be explained by the fact that the auto-align limit for the alignment correction was set to its default value of 80 counts during exposure, thus allowing the system to align to the wrong peak (or peaks) in the auto-correlation function (see Section 4.7).

The two graphs showing the standard deviations illustrate very similar trends in x and y. This could be because the system is aligning one half of the target consistently using the same peak (although not necessarily the correct one) while aligning to different peaks on the other half. In the case where the magnitudes of the mean values are also similar in x and y, the consistently aligned chevron is probably being aligned to the correct peak. This can be seen by equating the magnitudes of Equations 4.7 and 4.8, (the x and y-misalignments[†]), which gives either $c_{wr} - c_{rr} = 0$, or $c_{wl} - c_{rl} = 0$. This implies that when the misalignment was read out, the wafer mark centre and reticle mark centre were coincident in one

[†]Although these were quoted for OASIS, the machine uses the same equations to calculate the misalignment.

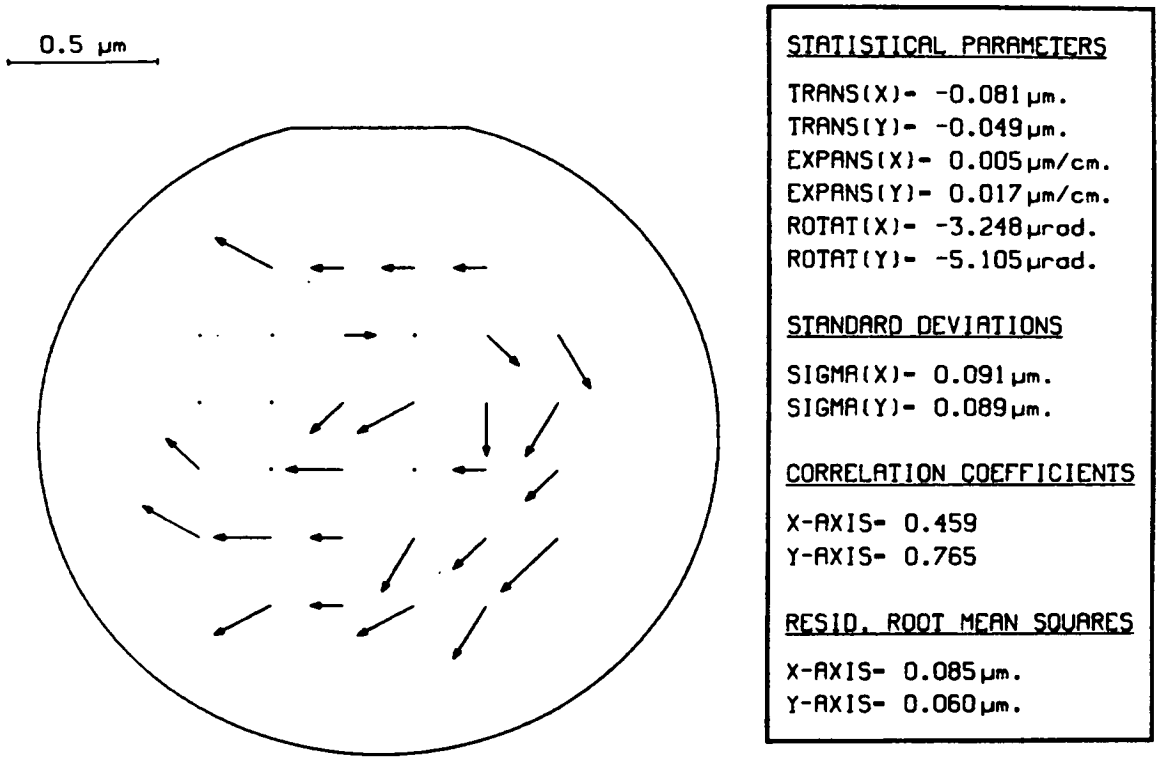

Figure 6.11: Vector map of vernier read data from wafer 48.

of the two halves, ie. that one half of the chevron was correctly aligned during exposure.

In the case where the x and y-translations are not the same, the consistent half is probably misaligned. One consistent and one inconsistent half is the only way in which similar standard deviations of this magnitude, and different mean values, can be explained. (Of course random errors in x and y could also give similar σ values and different means. If this were the case, however, the σ values would be much smaller than they are here.)

6.2.2 Buried Contact.

Figures 6.16–6.19 show the x-translation, y-translation, x-standard deviation and y-standard deviation for alignment of the buried contact layer to active area. From the translation data in this case, all the markers give acceptable results apart from the 3.0μm and 1.5μm standard marks. For these markers, the magnitudes of the x and y-translations have similar values, which implies the system has located the correct peak in one of the chevron halves, while locating the wrong peak in the other half. The large value of these translations suggests

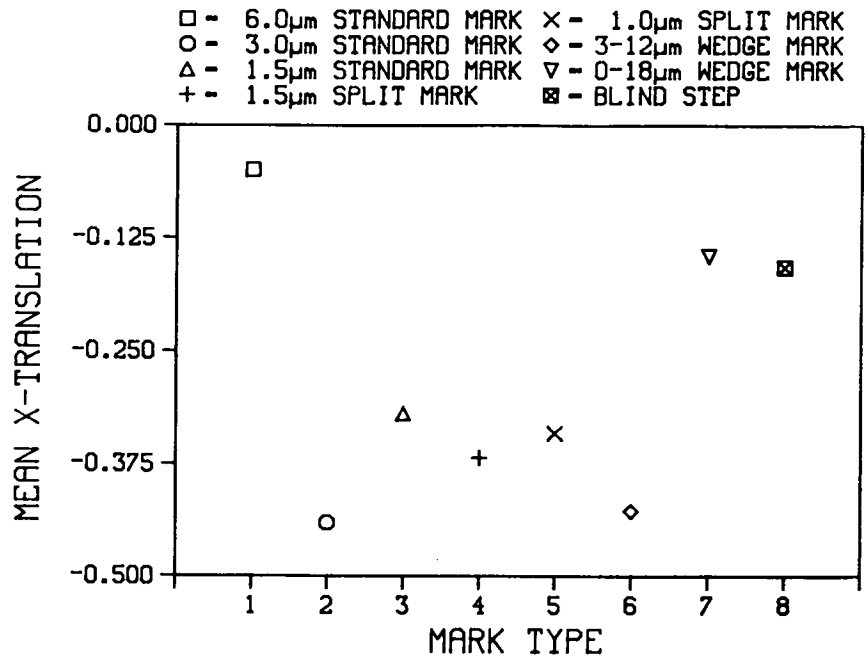

Figure 6.12: Average x-translation over the whole batch, for various types of marker, at the depletion implant level.

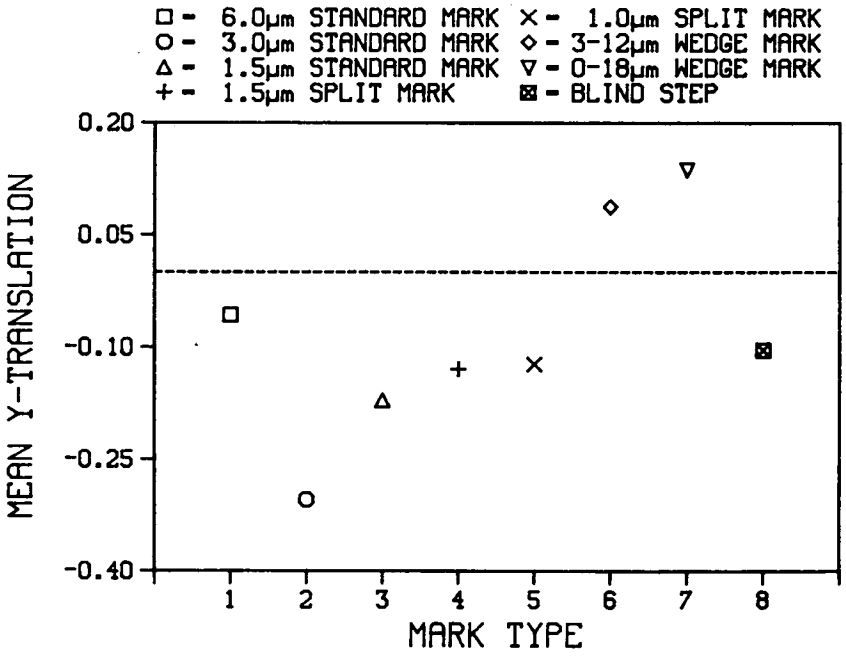

Figure 6.13: Average y-translation over the whole batch, at the depletion implant level. The dashed line represents zero misalignment.

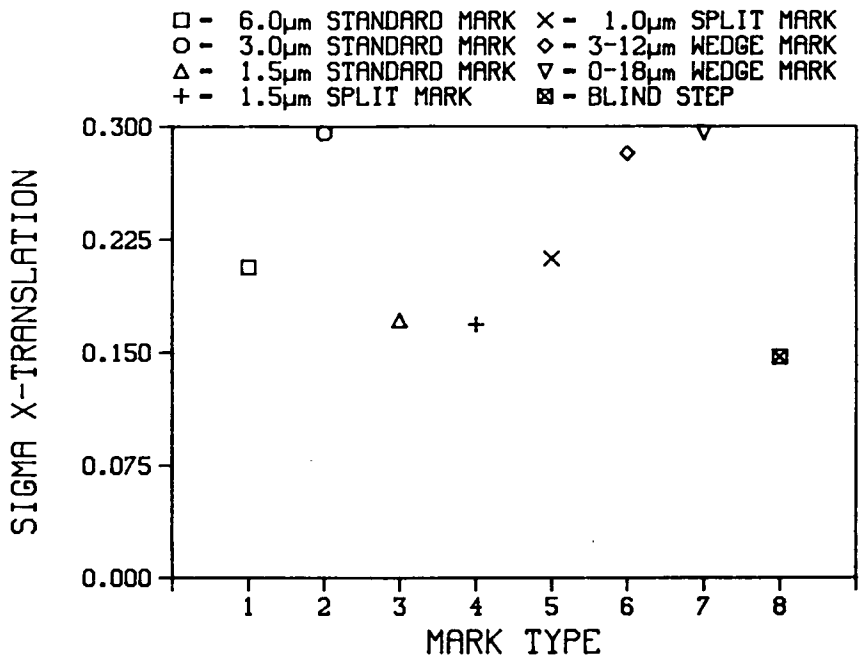

Figure 6.14: Average x-standard deviation over the whole batch, at the depletion implant level.

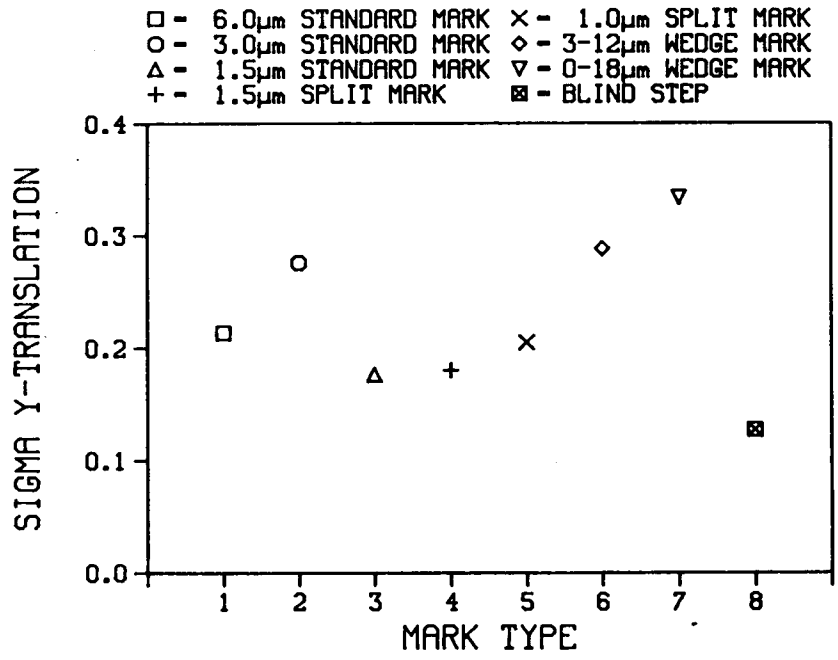

Figure 6.15: Average y-standard deviation over the whole batch, at the depletion implant level.

again that the value S474 = 80 is allowing the system to latch onto the wrong correlation peak.

From the standard deviation graphs, it can be seen that that the standard 6.0μm marker performs best on this layer, with the 3–12μm and 0–18μm markers close behind. All three types of marker fulfill the performance criterion of $\sigma \leq$ 0.1μm. These graphs also show that the x and y-standard deviations are similar in value for all markers, which suggests again that the system is misaligning by the same amount in x and y at some positions on the wafer.

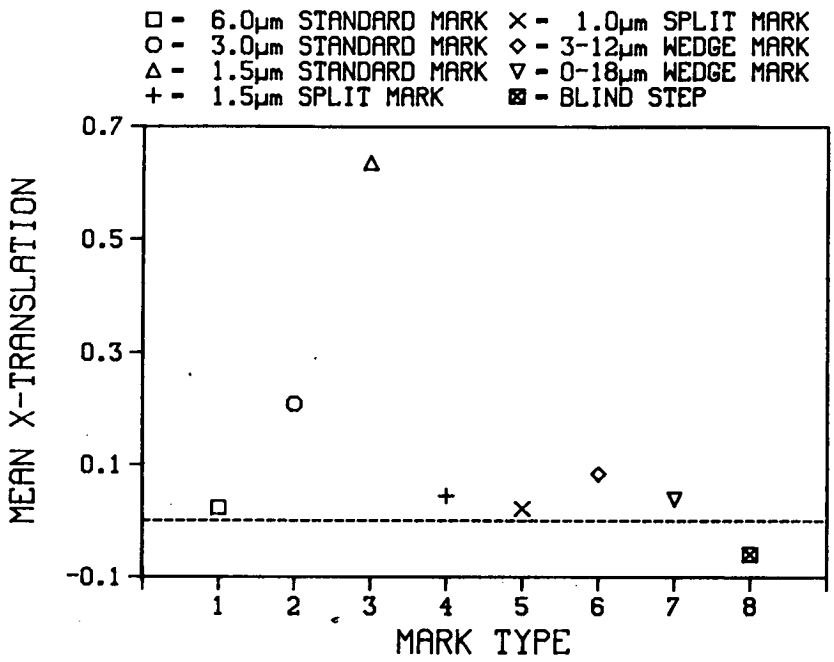

Figure 6.16: Average x-translation over the whole batch, at the buried contact level. The dashed line represents zero misalignment.

6.2.3 Poly-silicon.

Figures 6.20–6.23 show the x-translation, y-translation, x-standard deviation and y-standard deviation for alignment of the poly-silicon layer to active area. The average translations in this case show all the procedures to be acceptable except for blind step (the mean in x being too high in this case).

The standard deviation plots show that all markers performed with a similar degree of accuracy, with the 3.0μm standard mark being slightly better than the rest (in fact the 3.0μm marker is the only one which fulfills our 0.1μm 1σ criterion here). The reason for the improved alignment in this case, over the buried contact and depletion implant levels (particularly in terms of the standard deviations)

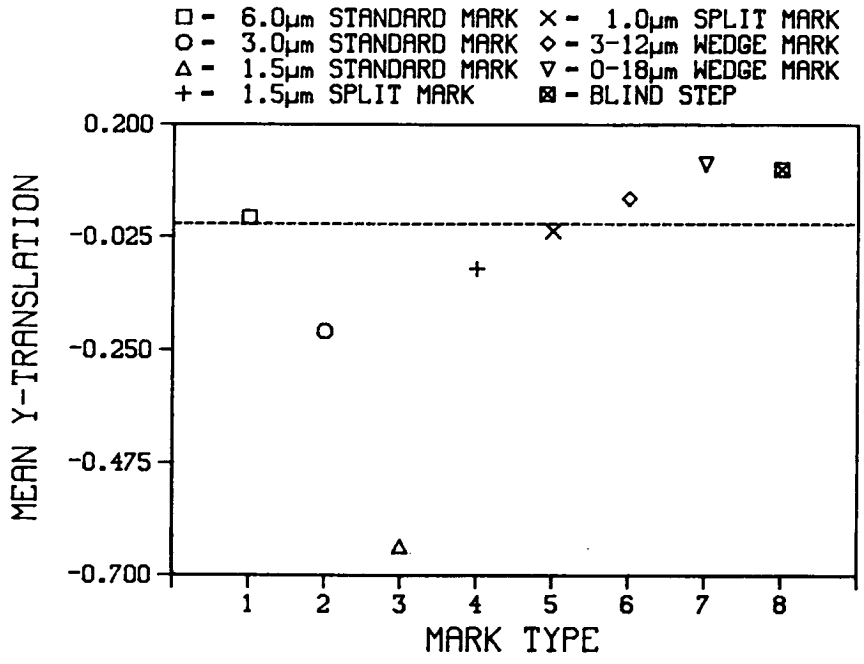

Figure 6.17: Average y-translation over the whole batch, at the buried contact level. The dashed line represents zero misalignment.

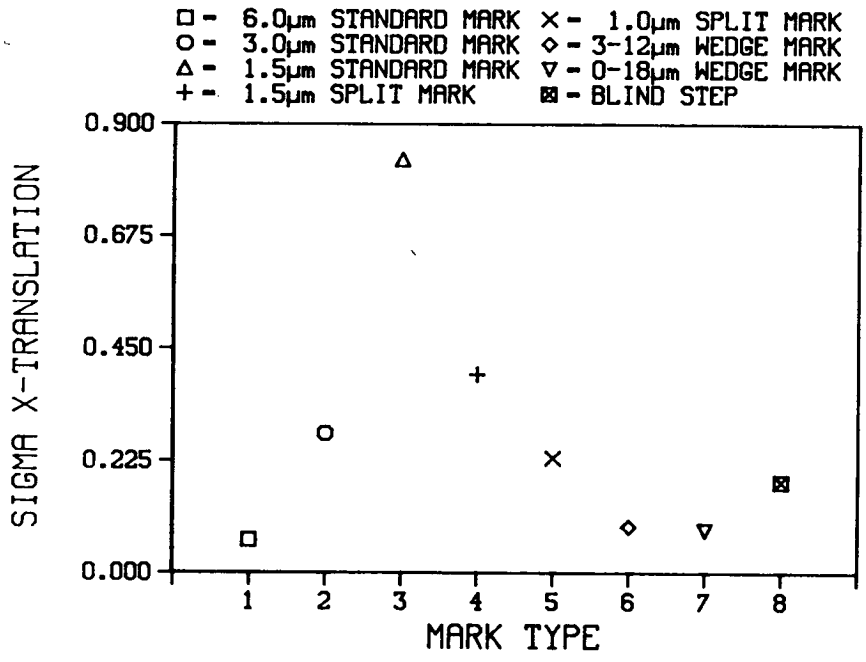

Figure 6.18: Average x-standard deviation over the whole batch, at the buried contact level.

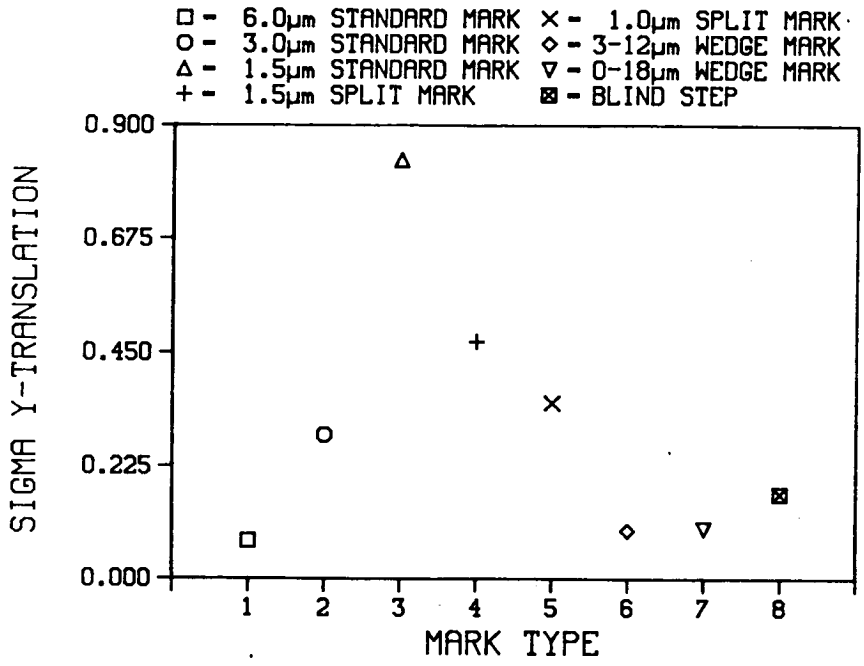

Figure 6.19: Average y-standard deviation over the whole batch, at the buried contact level.

is probably due to the fact that, at the poly-silicon level, the alignment was performed with machine location S474 set to 20, about 1/4 of the default auto-align limit (thus preventing the machine from latching onto the wrong peak in the correlation function). The σ values for all the auto-align marks are well below those of the blind stepped wafers in this case. Again the correlation between the x and y standard deviations is marked, suggesting that alignment on one half of the marker is better than on the other.

6.2.4 Metal Contacts.

Figures 6.24–6.27 show the x-translation, y-translation, x-standard deviation and y-standard deviation for alignment of metal contacts to poly-silicon. Again the average translation performance was satisfactory in all but the blind stepped case.

From the standard deviation data, the best performance was shown by the 1.5μm split mark, with the two wedge marks close behind (although only the 1.5μm split mark fulfills our 0.1μm 1σ criterion). This photo-stage was also performed with the auto-align limit set to 20 counts, with the die-by-die marks again easily out-performing the blind stepped wafers.

Again there is a marked correlation, for both the means and standard devia-

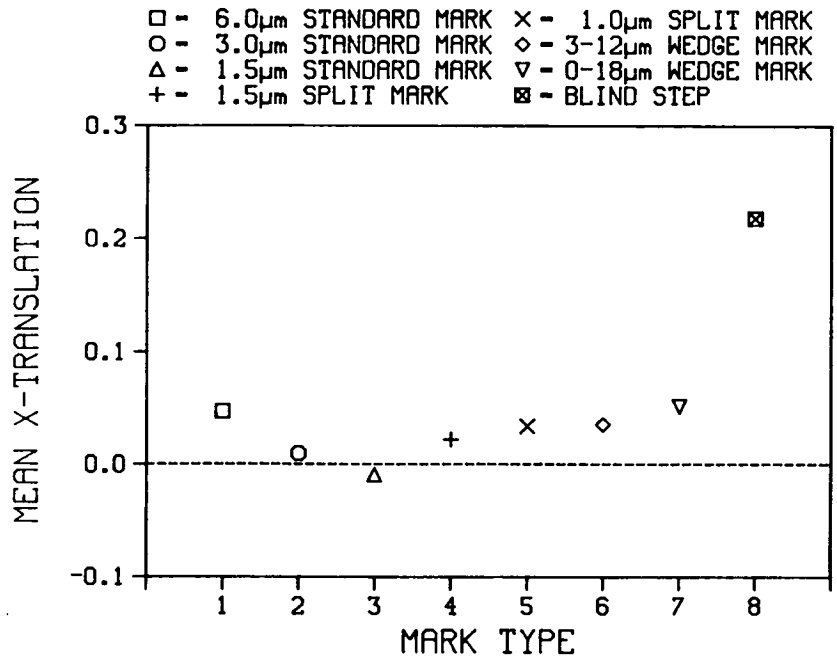

Figure 6.20: Average x-translation over the whole batch, at the poly-silicon level. The dashed line represents zero misalignment.

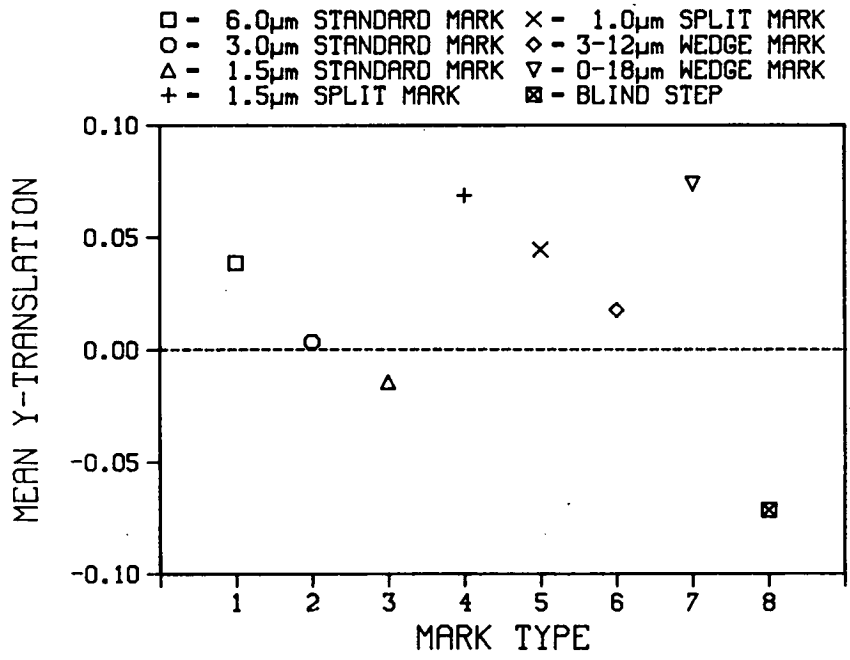

Figure 6.21: Average y-translation over the whole batch, at the poly-silicon level. The dashed line represents zero misalignment.

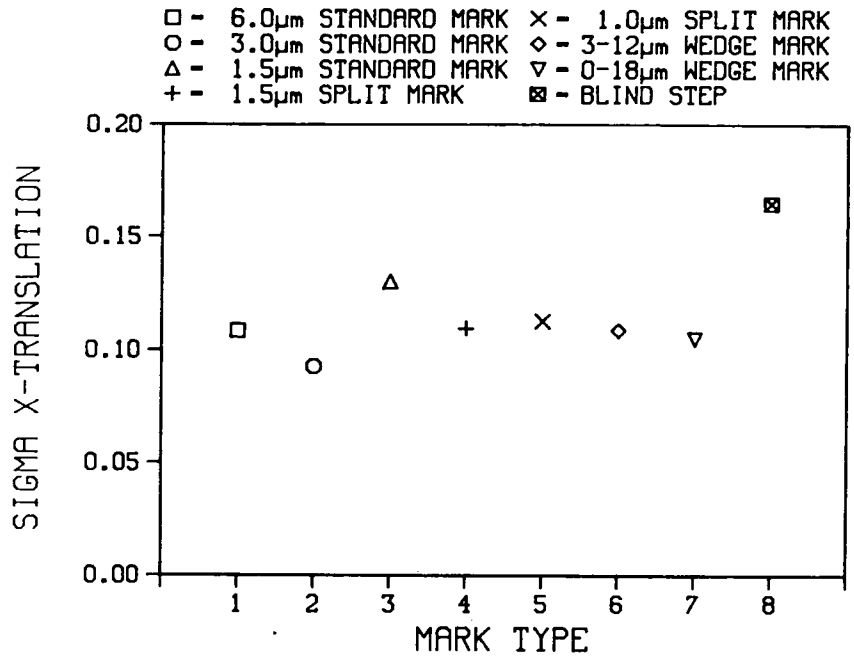

Figure 6.22: Average x-standard deviation over the whole batch, at the poly-silicon level.

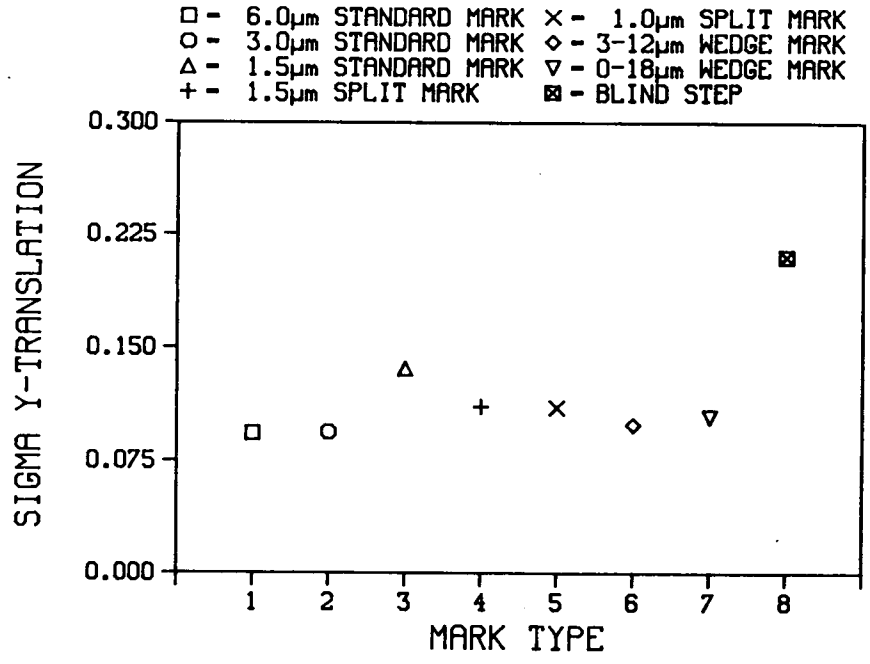

Figure 6.23: Average y-standard deviation over the whole batch, at the poly-silicon level.

tions, between the x and y results. This suggests that there is still an occasional large misalignment on one half of the chevron occurring here.

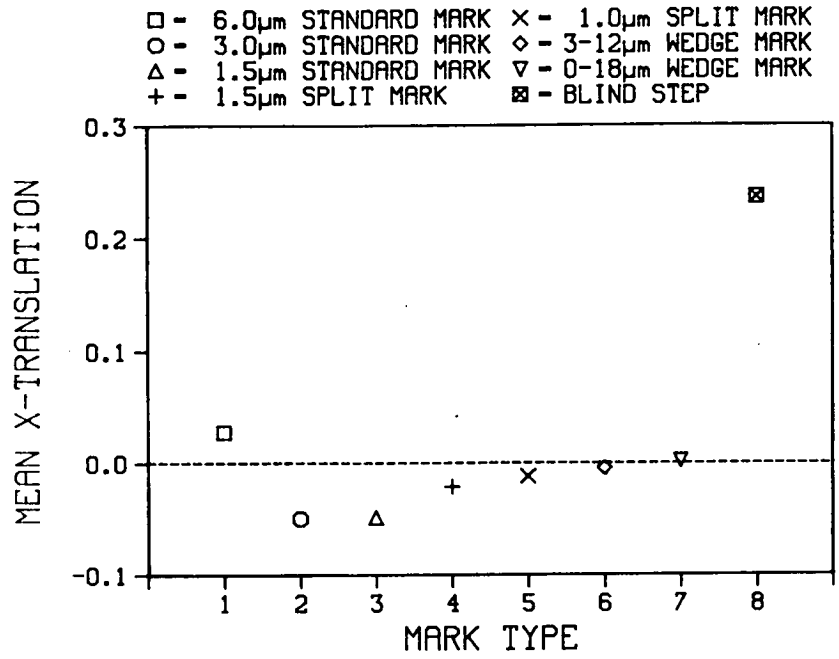

Figure 6.24: Average x-translation over the whole batch, at the metal contact level. The dashed line represents zero misalignment.

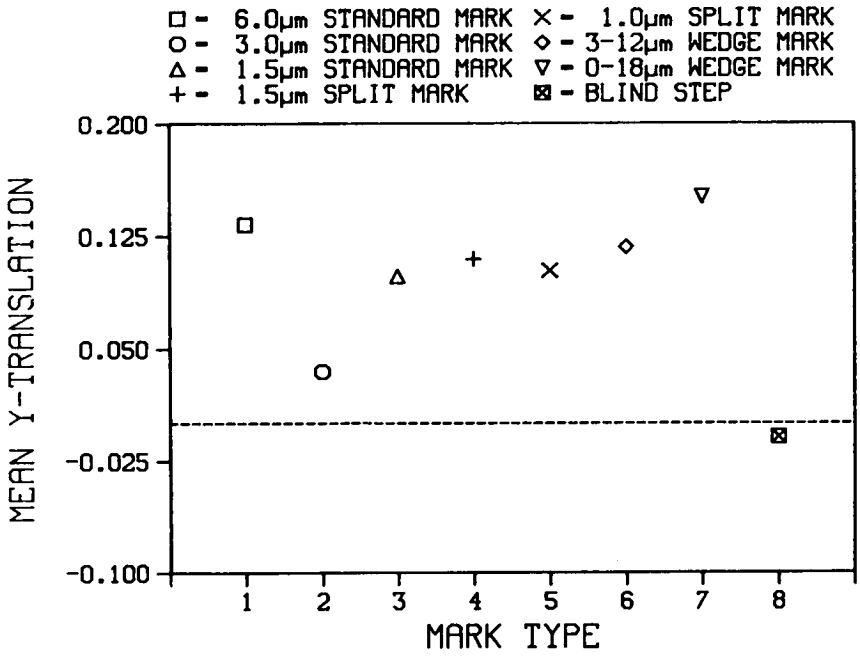

Figure 6.25: Average y-translation over the whole batch, at the metal contact level. The dashed line represents zero misalignment.

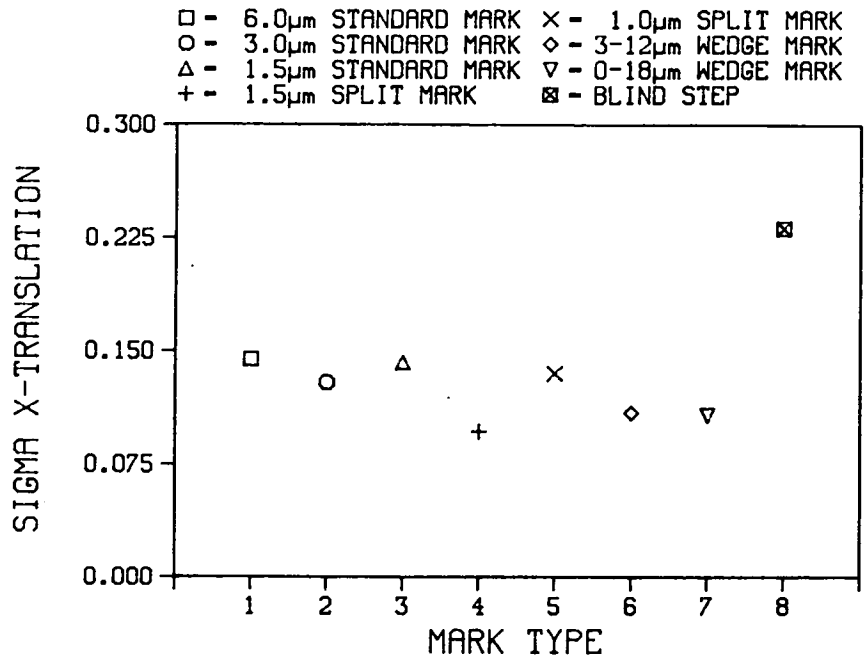

Figure 6.26: Average x-standard deviation over the whole batch, at the metal contact level.

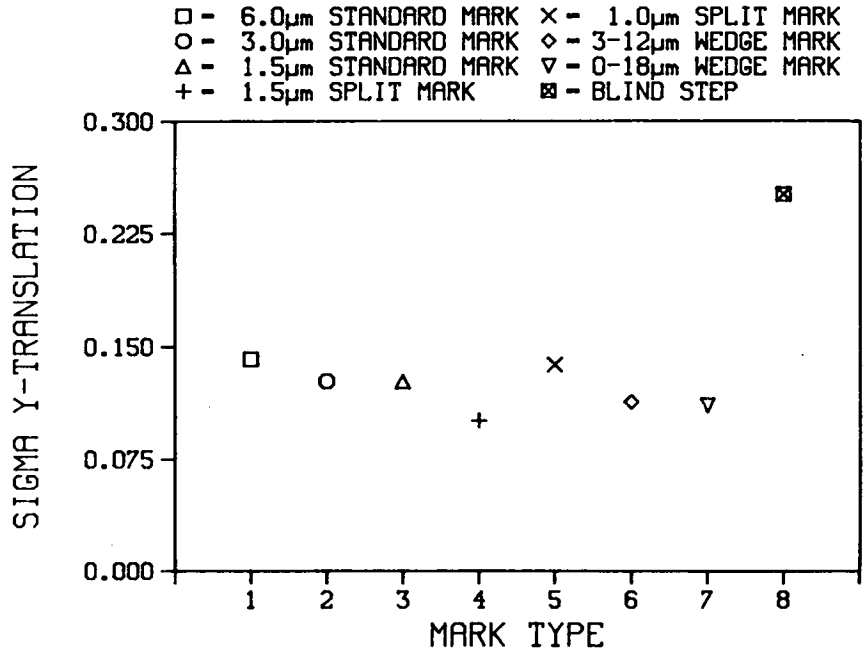

Figure 6.27: Average y-standard deviation over the whole batch, at the metal contact level.

6.3 Conclusions.

A few conclusions can be drawn from the experiment :

1. In order for die-by-die alignment to out-perform global alignment, it is necessary to specify a value for the auto-align limit which will not allow the die-by-die system to latch onto the wrong peak in the auto-correlation function. A value of 20 for machine location S474 should be used for this purpose (this will result in more auto-align failures, since the alignment system will sometimes find peaks which are outside its capture limits). It was noticed that, on layers which are generally accepted as difficult for alignment (poly-silicon and metal contact), the results were better with S474 = 20, than the results for the simpler layers (depletion implant and buried contact) had been with S474 = 80. For the poly-silicon and metal contact layers, all the observed σ values were under 0.15μm, while for depletion implant σ values of 0.3μm were seen. For buried contact, σ values as high as 0.8μm were observed.

2. The optimum system performance is not necessarily produced by the standard 6.0μm mark. In the metal contact case, the 1.5μm split mark gave a σ value of $\sim$0.10μm in both x and y, compared with $\sim$0.14μm for the 6.0μm mark, while the 3.0μm mark gave slightly better results for poly-silicon alignment. Although the wedge markers outperformed the standard 6.0μm marker at metal contact level, in no case did they prove to be optimal. In order to properly assess the effect of different markers in the buried contact and depletion implant stages, the experiment should be repeated, with S474 = 20 for these steps.

3. The direct correlation between the x and y results in this experiment suggest that misalignment is occurring predominantly on one half of the marker. In the case where the x and y translations are the same in magnitude and sign, the left hand side of the marker is being misaligned. When the magnitudes are the same, but the signs opposite, the right hand side is being misaligned.

   Where the standard deviations and mean values are large, this problem can be cured by reducing the auto-align limit of the system. The trend of misalignment by the same amount in x and y persists, although to a lesser extent, even when this is done. This may, in fact, be an artifact of the statistical processing used by VECPLO. Figure 6.28 shows a vector plot of

misalignment data, taken from a buried contact wafer which was aligned using a 3.0μm marker. Clearly, at some die the wafer has been misaligned by locating the wrong peak in the auto-correlation on half of the marker. From a statistical point of view, there are two distinct distributions of vector present on the wafer, which raises a question about the validity of the standard deviation as an indicator for the spread of the values. A relatively small number of large diagonal (same in x and y) misalignments can lead to similar standard deviations in x and y, even though the real standard deviations in the main part of the distribution are different.

The only alternative to this approach, however, is full inspection of all vector maps for all wafers (a total of 200 maps) which is obviously not desirable. In condensing the results we have necessarily lost some information content of the raw data; this has not prevented us, however, from drawing useful conclusions. It should be remembered that the results presented were extracted under real processing conditions. A low standard deviation in this case, while not comparable with a standard deviation in the pure sense, suggests that occasional large misalignments are less likely to occur, thus going some way towards our goal of increasing the number of good die/hour through the machine.

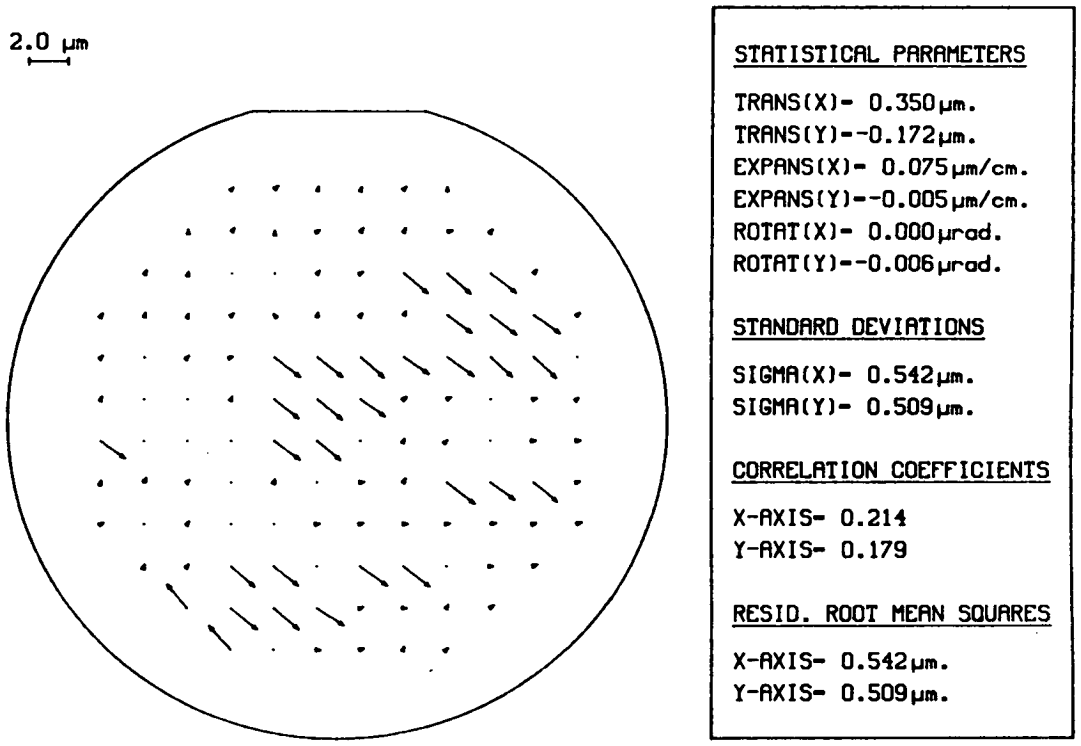

Figure 6.28: Vector map of data taken from buried contact wafer aligned using a standard 3.0μm marker.

Chapter 7

Simulation of Photo-lithography.

Process simulation has emerged in recent years as a major tool in developing new processes for circuit fabrication for two major reasons: increased speed and reduced cost when compared with experimental process development. While experimental evaluation may take weeks for any one critical process step in a typical R & D facility, several process steps may be studied in a matter of hours using a medium sized computer (eg. VAX 11/780).

At present, process development ultimately relies heavily on experimental evaluation. Simulation is used mainly for initial study in order that the process engineer may gain some idea of a starting point for subsequent experiment. Many programs have been written for the purpose of simulation of photo-lithography; in this chapter a review of one of these programs called SAMPLE will be presented in some detail. Some additional programs will also be outlined in brief.

7.1 SAMPLE

SAMPLE (Simulation And Modelling of Profiles for Lithography and Etching) is a 2-dimensional program designed to simulate line edge profiles and surface topography at various stages during an IC fabrication process [20] [93]. It has a wide range of capabilities, including the simulation of contact and projection lithography, wet and dry etching and material deposition, but only the lithography section will be covered here.

The program is written in standard FORTRAN IV to ensure optimum portability, and is essentially modular in design. It consists of an input interpreter to analyse input files, a process controller to pass the input cards to the appropriate processing machine, and finally the processing machines themselves which perform the various exposure and development stages.

Input to the program consists of various keywords which specify the type of action to be taken (expose, develop, ...), along with certain numerical parameters which assign physical values to the action taken (exposure dose, develop rate, ...).

Output from the program is of two types. Firstly, the standard output is sent to the terminal. This output simply echoes the input, and gives details of the progress of the run, as well as rough (line printer type) graphs of the final profiles. Secondly, full numerical output is sent to files which are used as input to a graphics post-processor.

The lithographic section of SAMPLE is divided into three main sections, each of which may be run independently, with the output from each run being used as the input for the next section. The three sections of the program are explained in turn below.

7.1.1 Calculation of the Aerial Image.

The aerial image is defined as the horizontal intensity distribution at the top surface of the photo-resist ($I(x)$), normalised by the intensity incident on the mask. The aerial image is dependent on several parameters of the imaging system, including the size and type of feature to be printed and the exposure wavelength or wavelengths. For a projection system, the numerical aperture (NA), partial coherence (σ) and defocus distance (distance between the image plane and the top surface of the resist) are also important. Figure 7.1 shows an example of the aerial image which SAMPLE computed for a pattern consisting of 3μm lines and 2μm spaces, using a projection system with NA = 0.32, σ =0.5, and a defocus distance of 1.0μm. The calculation is performed by making two assumptions: firstly that the imaging is 1-dimensional, which is valid for a long feature, and secondly that the lens is diffraction-limited, which is true of most modern lenses designed for microlithography.

7.1.2 Calculation of the Relative Inhibitor Concentration.

In a standard positive photo-resist process, the photo-active compound (PAC) acts as a dissolution inhibitor in alkaline developer when mixed with the base resin which forms the bulk of the resist. When PAC is exposed to radiation in the 300–450 nm range, the inhibitor is converted into carboxylic acid, which acts as a dissolution enhancer when the resist is immersed in developer. This

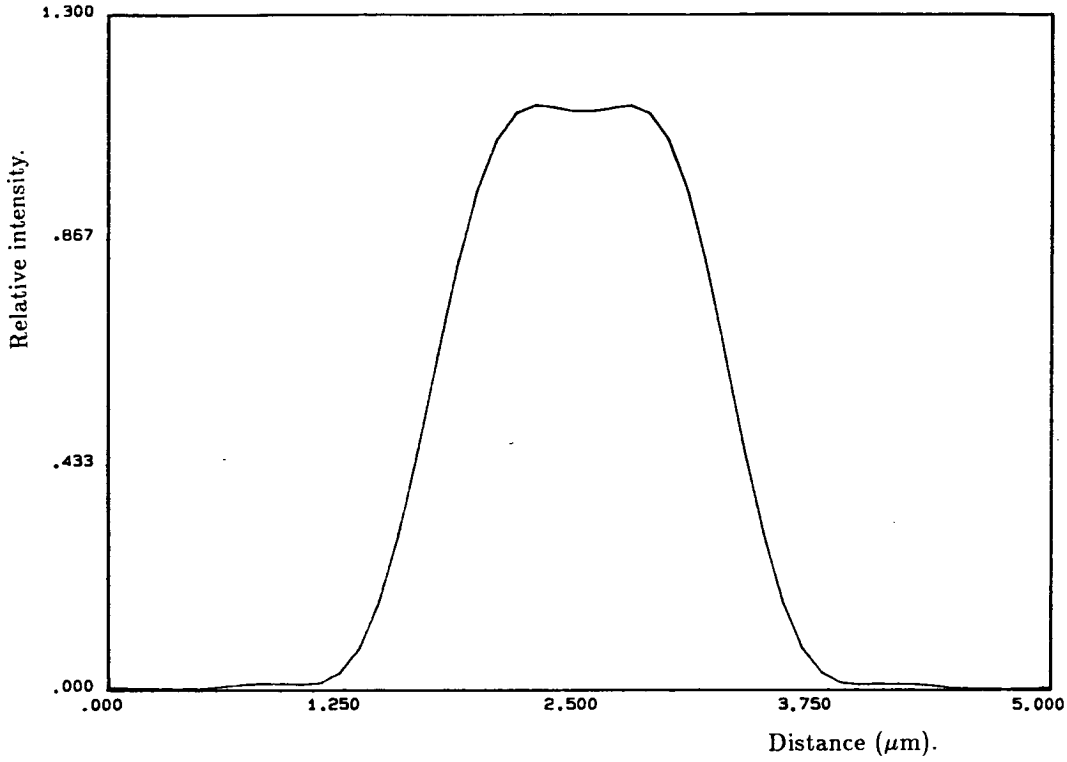

Figure 7.1: Aerial image of 3μm line, 2μm space.

photo-induced reaction forms the basis of all standard positive resist processes. SAMPLE uses the concentration of inhibitor after exposure, normalised by the concentration in unexposed resist, to calculate the dissolution rate of resist in a developer solution, as a function of exposure energy and depth into the resist.

The relative inhibitor concentration, $M(z,t)$, and localised intensity, $I(z,t)$, can be obtained at any point in the resist by solving the following equations subject to the appropriate boundary and initial conditions [94] :

$$\frac{\partial I(z,t)}{\partial z} = -I(z,t) \times (AM(z,t) + B) \tag{7.1}$$

and

$$\frac{\partial M(z,t)}{\partial t} = -I(z,t) \times M(z,t) \times C \tag{7.2}$$

The A, B, and C parameters in this equation represent, respectively, exposure dependent and independent absorption terms, and an optical sensitivity term. These parameters are empirical and may be obtained from estimates of the transmission of the resist film before and after exposure, and the initial rate of change of transmission. Solving (2) above for $M(z,t)$ leads in a straightforward manner to the relative inhibitor concentration as a function of depth and

exposure energy ($M(z,E)$) since the exposure energy, $E(z,t)$, is given by :

$$E(z,t) = \int_0^t I(z,t')dt' \tag{7.3}$$

From the calculation of the aerial image, we know the exposure energy as a function of x:

$$E(x) = I(x) \times t \tag{7.4}$$

This allows the inhibitor concentration to be calculated as a function of horizontal and vertical distance into the resist ($M(x,z)$).

7.1.3 Calculation of Developed Contours.

Once the $M(x,z)$ array has been calculated, SAMPLE uses the following equation to calculate the development rate $R(M)$:

$$R(M) = \exp\left(E_1 + E_2 M + E_3 M^2\right) \tag{7.5}$$

The E parameters are again purely empirical, and are adjusted according to a least squares algorithm in order to make the above equation fit the experimental data. There are drawbacks with this method in that it sometimes produces unphysical maxima in the development rate when M is small [94] (in fact SAMPLE will print a warning if M is lower than 0.4 anywhere in the resist). It also fails to take account of the 'surface induction effect' [95], the retardation of development rate near the surface of the resist, which is due to the formation of a skin at the resist/air interface. Both of these effects have recently been addressed by the inclusion of some additional E parameters [96].

Development itself is simulated using a string algorithm in which the boundary between developed and undeveloped regions consists of a series of points joined together by straight line segments (strings). Each point advances in a direction parallel to the angular bisector of the two adjacent strings, at a rate which depends upon the local value of $R(M)$. As the development front advances into the resist, extra points are added at positions where the front is expanding, and deleted where the front is contracting, thus keeping the density of points roughly constant in space and time.

Table 7.1 shows an example of a SAMPLE input file which can be used to simulate the exposure and development of a pattern consisting of 3μm lines and 2μm spaces printed by an exposure tool with a 0.32 NA lens operating at 436nm. Each line is explained by a comment card which begins with an asterisk. Figure 7.2 shows the developed profiles which were obtained from this run.

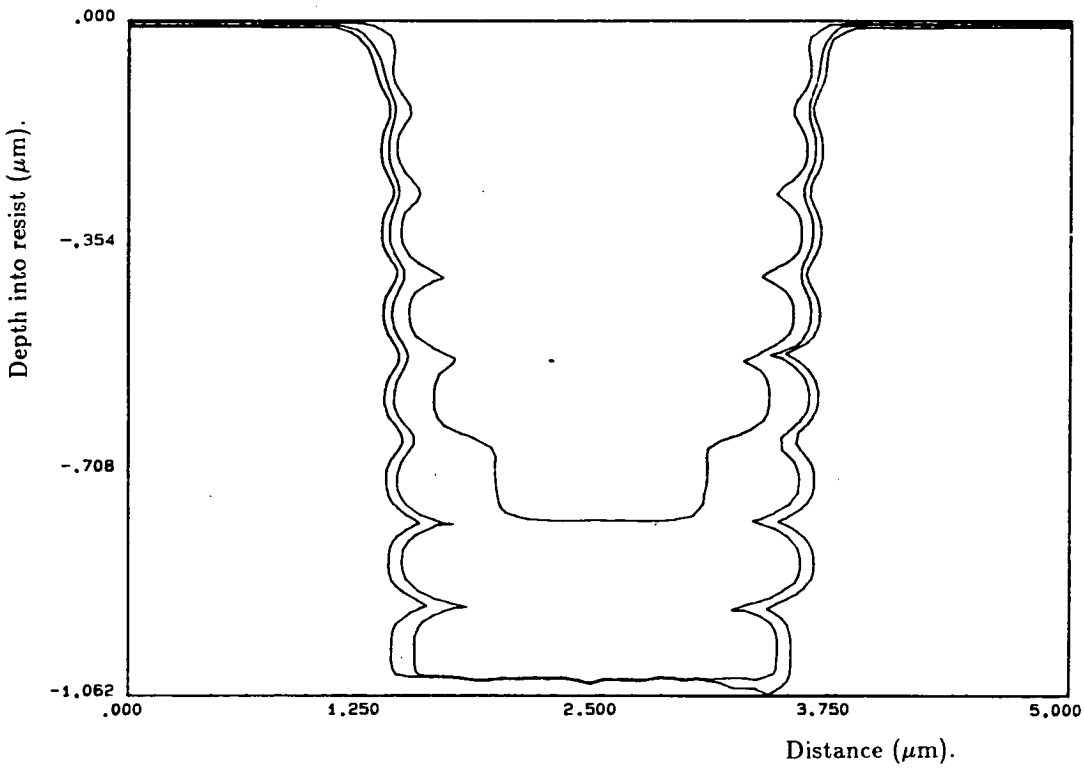

Figure 7.2: Developed contours for 3μm line, 2μm space, at intervals of 20 seconds.

```
*-------------------------------------------------------------
* lithography data.
*-------------------------------------------------------------
* set control functions - no diagnostics, output plotting data,
* high accuracy
trial 2 0 1 1
* request aerial image plot
trial 3 1 0 1
* projection system, numerical aperture = 0.32
proj 0.32
* exposure wavelength 436 nm
lambda 0.436
* periodic pattern of three micron line, two micron space.
linespace 3.0 2.0
* light source - partial coherence, sigma=0.5, defocus
* distance=1 micron.
trial 20 0 0.5 1
* horizontal window width=5 microns, edge location=2.0 micron.
trial 22 5.0 2.0
* refractive index of silicon and silicon dioxide(0.05 microns
* thick)
layers (4.73 -0.158) (1.47 -0.0 0.05)
* at 436nm resist is modelled with A=0.551, B=0.056, C=0.01
* resist refractive index=1.68, thickness=1.0 microns
resmodel (0.436) (0.551 0.058 0.01) (1.68 -0.0 1.0)
* exposure energy=150 mJ/(cm)**2
dose 150
* constants for etch rate: rate=exp(5.63+7.43m+(-12.6)m\^2)
* angstroms/sec
etchrate analytic (5.63 7.43 -12.6)
* develop from 20 to 60 secs (3 contours)
devtime 20 to 60 , 3
* run exposure machine
run
*-------------------------------------------------------------
```

Table 7.1: Input file for SAMPLE.

7.1.4 Additional Routines used in Photo-Lithography.

Post-Exposure Bake.

Post-exposure bake (PEB) is a commonly used method for the reduction of standing wave effects, which cause the terraced sidewall effect evident in Figure 7.2. PEB is modelled by SAMPLE by a simple diffusion in which the diffusion length (standard deviation) is specified. In practice a standard deviation of $\sim \lambda/4$ is sufficient to virtually eliminate any standing wave pattern.

Multiple Wavelength Input.

It is possible to simulate the effect of polychromatic exposure by entering multiple wavelengths in the input file. Each different wavelength must be specified along with its own A,B and C parameters, and a measure of its intensity, relative to the other wavelengths. This particular facility is useful in the modelling of reflective optical systems, such as the Perkin-Elmer whole wafer projection systems, or the Ultratech stepper.

Plasma Descum.

By specifying a development rate which is independent of exposure, a plasma descum may be performed once the normal development has been completed. The purpose of this is to remove the sharp nodes produced by the standing wave effect, and thus has a similar effect to the post-exposure bake. Table 7.2 shows the additional input required to simulate exposure of the previous pattern with a post-exposure bake, and also with a plasma descum. Figure 7.3 illustrates the output obtained from running the PEB only (no descum), while Figure 7.4 shows the effect of the descum only (no PEB).

7.1.5 Additional Options.

As well as being able to output developed contours and the aerial image, SAMPLE can also be used to plot the modulation transfer function of the lens [97]. In addition to this there are many TRIAL statements available for more complex options, such as irregular mask patterns, modelling of the surface induction effect, and light scattering during exposure.

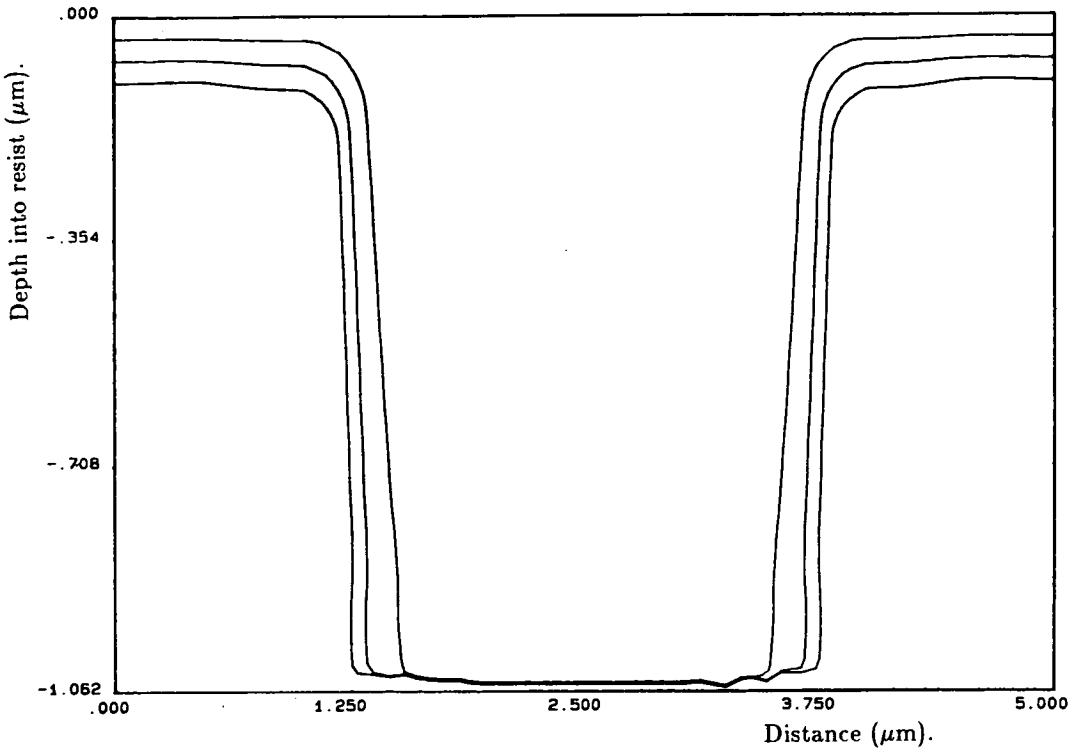

Figure 7.3: Developed profiles at 20 second intervals with post-exposure bake.

7.2 Other Simulation Programs.

Many other lithography modelling programs have been developed, some of which are closely related to SAMPLE, and some of which are totally different. A few of these are outlined below.

7.2.1 SPESA and VARYIM.

SPESA (Signetics Philips Enhanced SAMPLE Algorithm), is essentially identical to SAMPLE, with modified input/output stages. Instead of requiring an input file with command statements which are at best semi-mnemonic and at worst completely cryptic, SPESA is menu driven, and asks the user at each stage which parameter he would like to modify [98]. In addition to plotting the developed profiles the program can also output values for developed feature size, sidewall slope and process latitude (rate of change of feature size) as functions of development time. Figure 7.5 shows the evolution of process latitude with time; this is a particularly important characteristic here since it relates directly to the manufacturability of devices using a particular process. It should be noted that

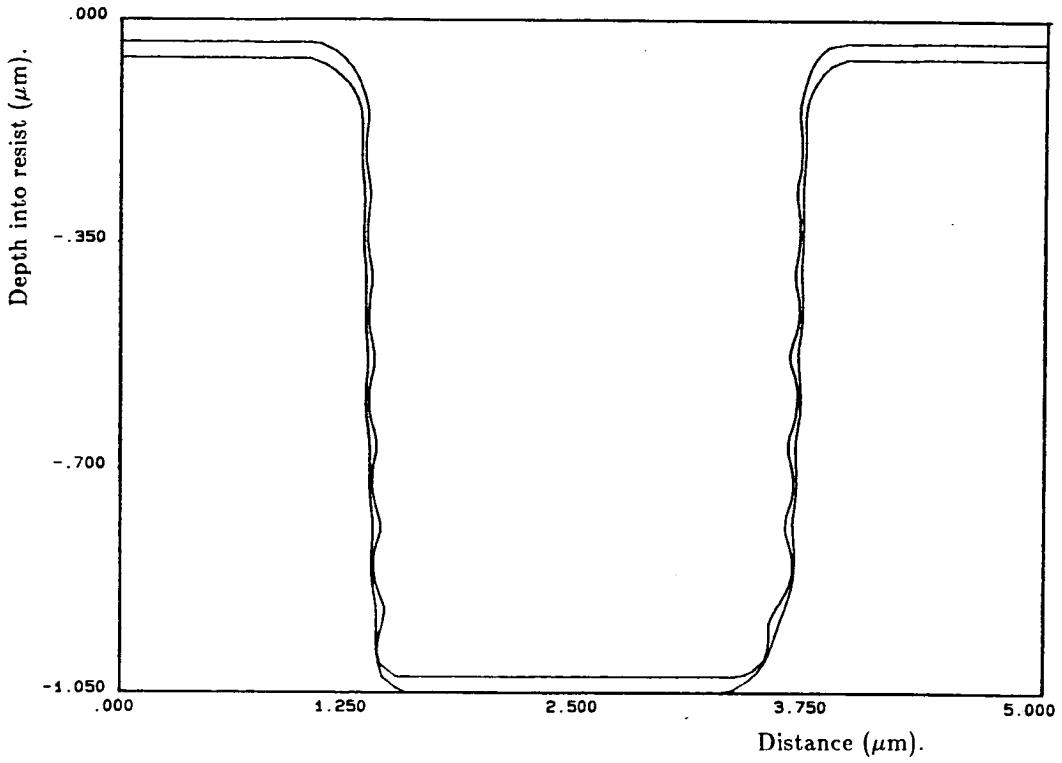

Figure 7.4: Final developed profile with 0.25μm and 0.5μm descum.

all the information available from SPESA is also available from SAMPLE, but in many cases access to the information from SPESA is simpler and more direct. Determination of process latitude is one such case.

VARYIM (**VARY IM**age) is simply the imaging module of SPESA which calculates and plots the aerial image of a mask pattern, while allowing the user to vary one imaging parameter (wavelength, defocus, numerical aperture, partial coherence or mask pattern) at a time. Figure 7.6 shows an example of VARYIM output where the numerical aperture of the lens was varied between 0.28 and 0.42, for a pattern with a 3.0μm linewidth and a 2.0μm spacewidth.

7.2.2 PROLITH

Mack has developed an optical lithography program called PROLITH (**P**ositive **R**esist **O**ptical **LITH**ography model) [99] which simulates the same lithography steps as SAMPLE, but differs in two important respects.

- Firstly, in addition to the calculation of projected images, it is also able to simulate proximity and soft contact printing (mask/wafer gap $\geq$ 1.0μm).

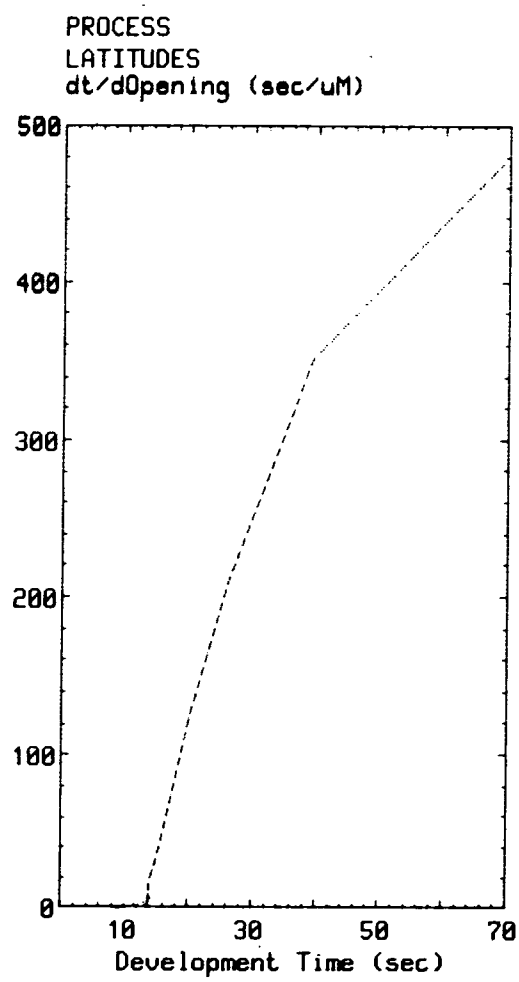

Figure 7.5: Development of process latitude with time, from SPESA.

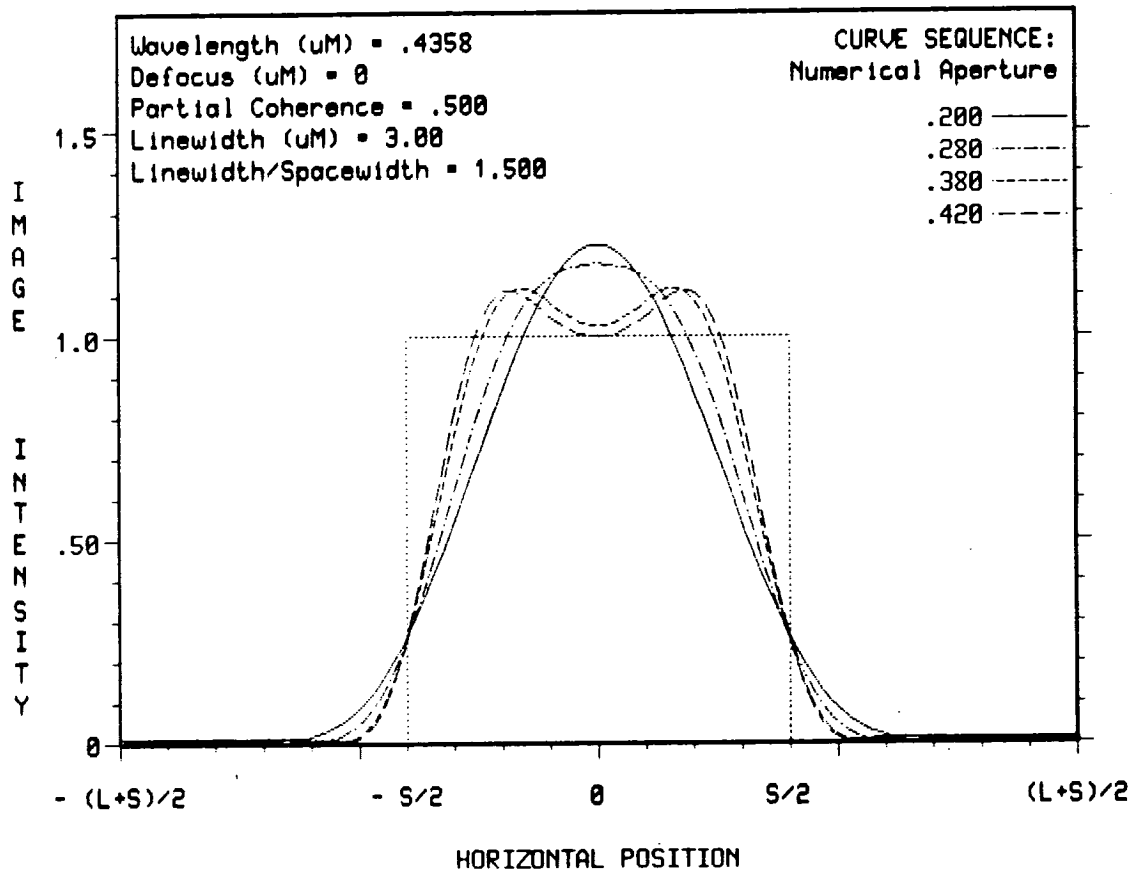

Figure 7.6: Aerial images of 3μm line, 2μm space, for different numerical apertures, from VARYIM.

```
*-----------------------------------------------------------
* request post-exposure bake, diffusion length = 0.1 um
*-----------------------------------------------------------
trial 1 0.1
* rerun develop machine after post-exposure bake.
run 4
* save output on iounit no 1
trial 92 1
*-----------------------------------------------------------
*    descum data
*-----------------------------------------------------------
* descum between 0.025 and 0.05 microns with two profiles
trial 60 0.025 0.05 2
* save output on iounit no 2
trial 92 2
*-----------------------------------------------------------
```

Table 7.2: Additional input for PEB and plasma descum.

This is accomplished by calculating the electric field in the resist, as described by the Kirchoff diffraction integral [100].

- Secondly, the development model is much less empirical, being based on mathematical modelling of diffusion of developer to the surface of the resist, reaction between the developer and resist, and diffusion of dissolved resist back into the bulk developer. Using the same exposure kinetics as SAMPLE, and in particular the same definition of relative inhibitor concentration M, this leads to a complete description of the resist/developer behaviour by the specification of four rate parameters:

  1. r_{min}, the development rate for unexposed resist (M=1).
  2. r_{max}, the development rate for completely exposed resist (M=0).
  3. n, the number of exposed PAC molecules required to remove one resin molecule.
  4. M_{TH}, the threshold inhibitor concentration, defined by a point of inflection in the $R(M)$ curve.

For a discussion of the relative merits of empirical/physical models, see the end of this chapter.

7.2.3 TRIPS-1

A 3-dimensional photo-resist simulator, TRIPS-1, (**T**hree-Dimensional **R**esist **I**maging **P**rocess **S**imulator-1) has been reported by Matsuzawa et al. [101] [102], which calculates the 2-dimensional light intensity distribution of a contact hole or similar feature using Lin's method [30], then uses an algorithm based on ray tracing in geometrical optics to perform development. The 'rays' here are defined by the direction of movement of the developing front in the resist, so that the surface perpendicular to the rays defines the developing front itself.

It is likely that 3-d simulation will become increasingly important in the future, with particular application to contact holes or track corners. References [101] and [102] discuss the application of TRIPS-1 to the simulation of contact holes on a weakly reflecting substrate.

7.2.4 ATLAS

Matsumoto et al. have developed a 3-dimensional simulation package [103], ATLAS, (**A**utomatic **T**heoretical **L**ithography **A**nalysis **S**ystem), for the purpose of studying the effects of aberrations on a 2-dimensional image intensity distribution. Using this program it is possible to study the effect of various primary aberrations of the optical system, as well as decentering of optical elements. Resist development is strictly 1-dimensional in this case however, which must limit the validity of the developed resist profiles.

7.3 Discussion.

In order for simulation programs to achieve strict quantitative accuracy, a great deal of experimental characterisation is required, particularly when the program is to a large extent empirical in nature. For example, SAMPLE requires a large amount of experimental data to evaluate E_1, E_2, and E_3. This characterisation has been simplified recently by the introduction of machines such as the Perkin-Elmer digital rate monitor (DRM) which measure resist removal as a function of time, thus greatly speeding up the evaluation of the E parameters (and also the A, B and C parameters). However, in many cases it could be argued that so much calibration would have to be done that the need for further simulation

would be eliminated. Even after thorough characterisation there are still many empirical relations applied (simulation of surface induction effect, application of post-exposure bake, etc...) which limit the numerical accuracy of such programs.

The development of physically-based models in lithography is therefore important, since this promotes a better understanding of the various stages involved in a process. In addition to this, the use of physically meaningful parameters is conceptually simpler than the use of empirical models, since it allows cause and effect to be related to a large extent. For example, if the same simulation is run using two different developers, with one resulting in a larger linewidth than the other, it is more helpful to be able to trace this back to the developers having different values of maximum development rate, than to having different values of 'E_3', when one has little conception of the physical basis of 'E_3' in the first place. Kim et al. have reported an extension to SAMPLE which uses a maximum and minimum development rate, along with a single parameter which describes the variation of development rate with exposure, in order to overcome this problem [96].

Against this, it must be said that the prime objective of simulation, from a process development point of view, is to be true to physical reality, and not to physical models which inevitably contain certain simplifying assumptions regarding the nature of the reactions involved.

Extensive calibration requirements of empirical programs, and physical approximations of physically based programs, tend to make these models most useful in a qualitative fashion, for analysing *trends* in surface topography, ie. what happens when we *vary* a certain parameter. The similarity of resist/developer chemistry from manufacturer to manufacturer is such that, by using the programs in this manner, it is possible to determine which parameters are important for further process development, as long as quantitative accuracy is not sought.

Chapter 8

Investigation of Aerial Image Characteristics and their Relationship to Developed Resist Profiles.

It has already been stated that advanced resist processing can lead to improvements in the working resolution of a lithographic process. It cannot, however, restore information lost by a lens operating beyond its resolution limit. It is important to understand the properties of the aerial image in this regime, and in particular the characteristics which most strongly affect printed feature size and linewidth control.

A simulation study of aerial image characteristics was undertaken for this purpose, and the various features of the aerial image were related back to properties of simulated developed profiles. The work was begun at Philips Research Labs in Eindhoven, using the SPESA and VARYIM simulation programs, and completed in Edinburgh using SAMPLE. Many of the results presented here have been published separately [104].

8.1 Developed Feature Simulations.

Two main characteristics of the developed image were studied using SPESA and SAMPLE :

1. D_{dev}, the deviation of the developed spacewidth on the wafer, from the mask spacewidth, as a function of mask linewidth and mask spacewidth. (Linewidth refers to the width of an opaque line on the mask, or the size of an undeveloped region on a wafer. Spacewidth refers to the width of a clear line on the mask, or to a developed region on the wafer.)

2. S_{dev}, the sidewall slope of the developed resist profile. Thus S_{dev} is a measure of the linewidth variation which we can expect when transferring a feature from mask to wafer.

Unless otherwise stated, all development simulations were performed using a lens operating at 436 nm with a numerical aperture of 0.38 and an illumination partial coherence (σ value) of 0.7. In each case the simulated exposure dose was 100 mJ/cm^2 onto AZ1350J positive resist, with a development of 60s in AZ developer [94].

8.2 Aerial Image Simulations.

A total of six characteristics of the aerial image were studied using VARYIM and SAMPLE (see Figure 8.1) :

1. I_{me}, the relative intensity at the nominal mask edge (the boundary between the line and space areas).

   It was expected that the relative intensity at the mask edge would show some correspondence with the developed feature size on the wafer. In general it would be expected that the higher the value of relative intensity at the mask edge, the larger would be the developed spacewidth.

2. G_{me}, the normalised gradient in the intensity profile at the mask edge. This parameter was chosen since the gradient in the aerial image at the mask edge will affect both the sidewall slope of the developed resist profile and the latitude of the development process.

3. I_{max}, the intensity maximum.

4. I_{min}, the intensity minimum.

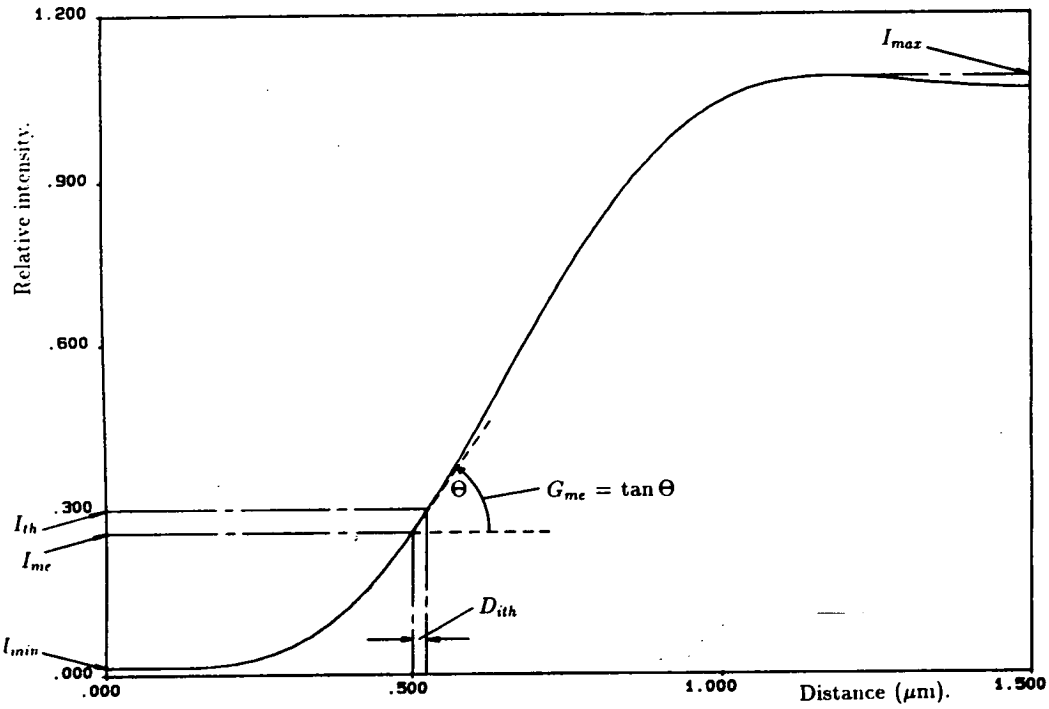

Figure 8.1: Example aerial image showing evaluated image characteristics (linewidth=1.0μm, spacewidth=2.0μm).

5. C_{im}, the image contrast, defined by the equation :

$$C_{im} = \frac{(I_{max} - I_{min})}{(I_{max} + I_{min})} \tag{8.1}$$

Image contrast was studied since this is the most frequently used image characteristic.

6. D_{ith}, the deviation of the x-position of the intensity threshold from the nominal mask edge. The intensity threshold (I_{th}) is defined simply as that value of relative intensity in the aerial image which will just result in the complete removal of the photo-resist from the substrate. This definition requires two assumptions :

   - That the aerial image is transferred vertically into the resist.
   - That the resist/developer interface advances vertically during development.

   If these two assumptions were strictly valid, the x-position of the intensity threshold would correspond precisely to the position at which the resist is just cleared from the substrate (ie. there would be a perfect correspondence between D_{ith} and the developed spacewidth). In practice neither the image transfer, nor the development are perfectly vertical. It was hoped, however, that there would still be some correspondence between this parameter and the developed feature size.

   Obviously in the real case the precise intensity threshold depends upon the exposure and development conditions chosen. Initial simulations indicated, however, that the relative intensity at the mask edge was around 0.3 in most cases, so this value was chosen for I_{th}. This value has been confirmed elsewhere [105].

The above six characteristics can be obtained relatively straightforwardly using VARYIM. SAMPLE, however, had to be modified in order to calculate these and output them to an appropriate file. By using a TRIAL 222 statement after each image simulation on the EMF VAX, this output may be placed in the file 'image.out'.

8.3 Presentation of Image Characteristics over the Full Line/Space Domain.

For completeness, any study of the above aerial image characteristics should encompass as large a matrix as possible of different linewidth and spacewidth values. A total of 144 different combinations of line and space were considered, corresponding to 12 different values of each parameter in a square matrix configuration. The sizes studied range from values which are too small to be realistically resolved by the lens, through to features which can be regarded as completely isolated.

All VARYIM simulations were performed using a 0.38 NA lens operating at 436 nm, after which both linewidth and spacewidth were normalised by λ/NA. This makes the results of the simulations independent of the lens being used. (Linewidth and spacewidth shall from now on be referred to as L and S respectively, while the normalised values shall be referred to as L_o ($=L\times$NA$/\lambda$) and S_o ($=S\times$NA$/\lambda$).) The values of line and space, both normalised and non-normalised, which were considered, are shown in Table 8.1.

Physical Feature Size. (μm.)	Normalised Feature Size.
0.50	0.436
0.60	0.523
0.70	0.610
0.80	0.698
0.90	0.785
1.00	0.872
1.25	1.090
1.50	1.308
2.00	1.744
3.00	2.616
4.00	3.488
5.00	4.360

Table 8.1: Non-normalised (physical) and normalised feature sizes for VARYIM simulations.

The simulation data is presented here as contour plots, each showing the behaviour of one of the six image characteristics (I_{me}, G_{me}, I_{max}, I_{min}, C_{im} or D_{ith}) over the 2-dimensional line/space domain. From these plots the value of the image characteristic can be read for any type of pattern (from isolated lines, via gratings, to isolated spaces). Figures 8.2 and 8.3 give an example of this method of data presentation, showing I_{max} and I_{min} respectively for partial coherence (σ value) =3.0, illustrating that I_{max} depends heavily on spacewidth but very little on linewidth, and that the reverse is true for I_{min}. In all of these plots, the region at the top represents isolated lines, and the region at the right, isolated spaces. The +45°diagonal represents gratings (line=space) of varying pitch.

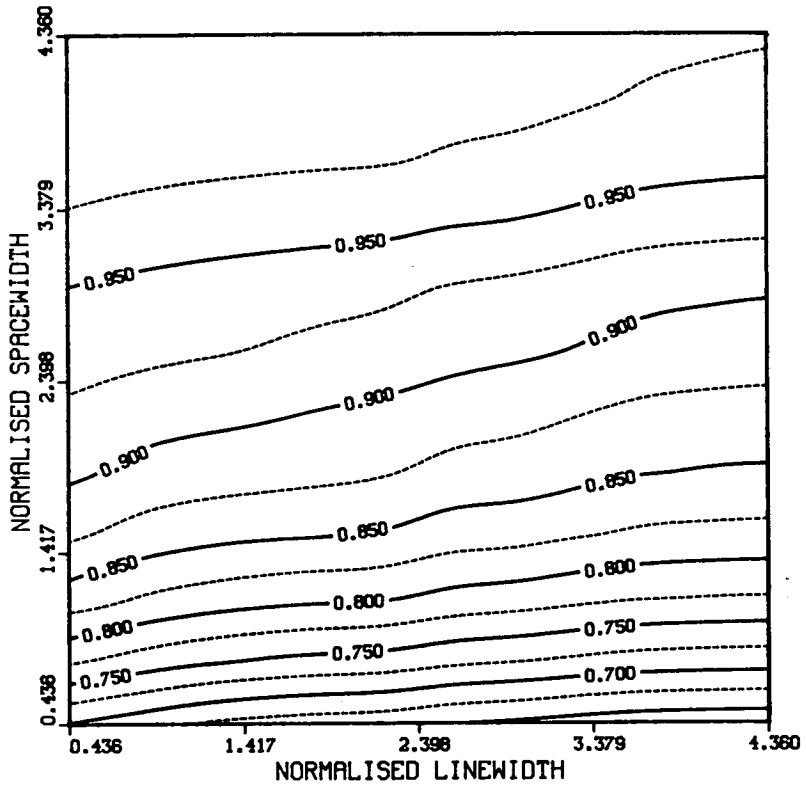

Figure 8.2: Contour plot of intensity maximum. σ=3.0, defocus=0.0 (Rayleigh units).

In addition to this matrix of linewidths and spacewidths, the image characteristics were studied over a range of values of partial coherence, including σ=0.01 (effectively coherent), σ=0.5 and σ=0.7 (typical working values in microlithography), and σ=3.0 (effectively incoherent).

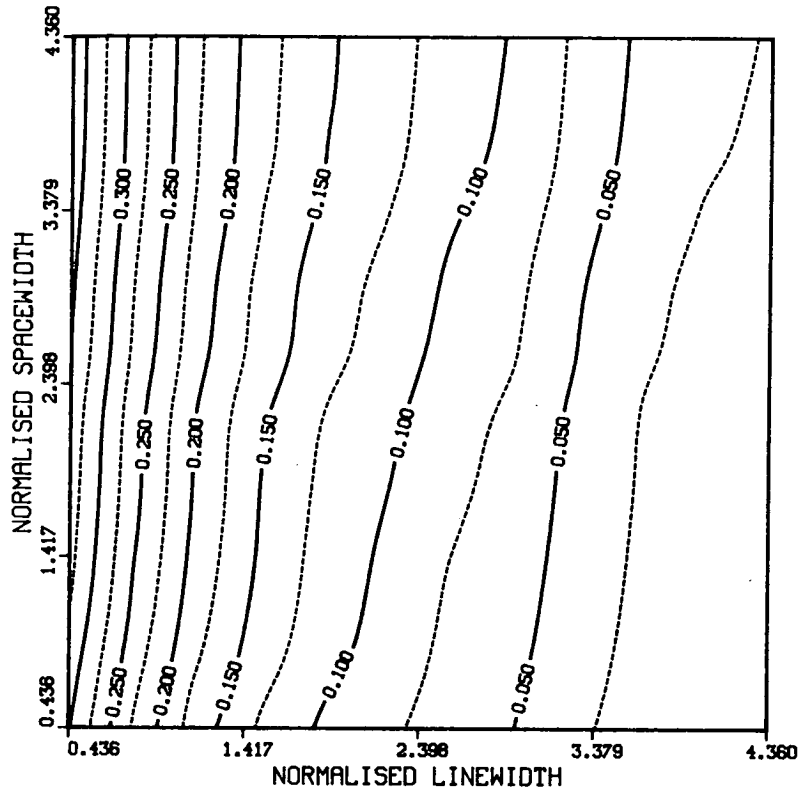

Figure 8.3: Contour plot of intensity minimum. σ=3.0, defocus=0.0 (Rayleigh units).

8.4 Factors Affecting the Deviation from Nominal Linewidth.

The results of SPESA simulations of D_{dev}, the deviation of developed spacewidth from mask spacewidth are plotted in Figures 8.4–8.6. Figure 8.4 shows this deviation as a function of linewidth, keeping spacewidth constant at 1.0μm (=0.87$\times\lambda$/NA for this lens), while 8.5 shows the deviation as a function of spacewidth, keeping the linewidth constant at 1.0μm. Figure 8.6 shows the variation as a function of grating size, with linewidth=spacewidth, for three different numerical apertures.

The graphs present some interesting features :

- Figure 8.4 (constant space, varying line) illustrates that there is little dependence of developed spacewidth on mask linewidth (less than 0.03μm variation down to a linewidth of 0.8μm, or 0.7$\times\lambda$/NA). In other words proximity (how close a space is to its nearest neighbour) seems to have very little effect on developed feature size, at least under the conditions of wavelength, partial coherence etc..., imposed here.

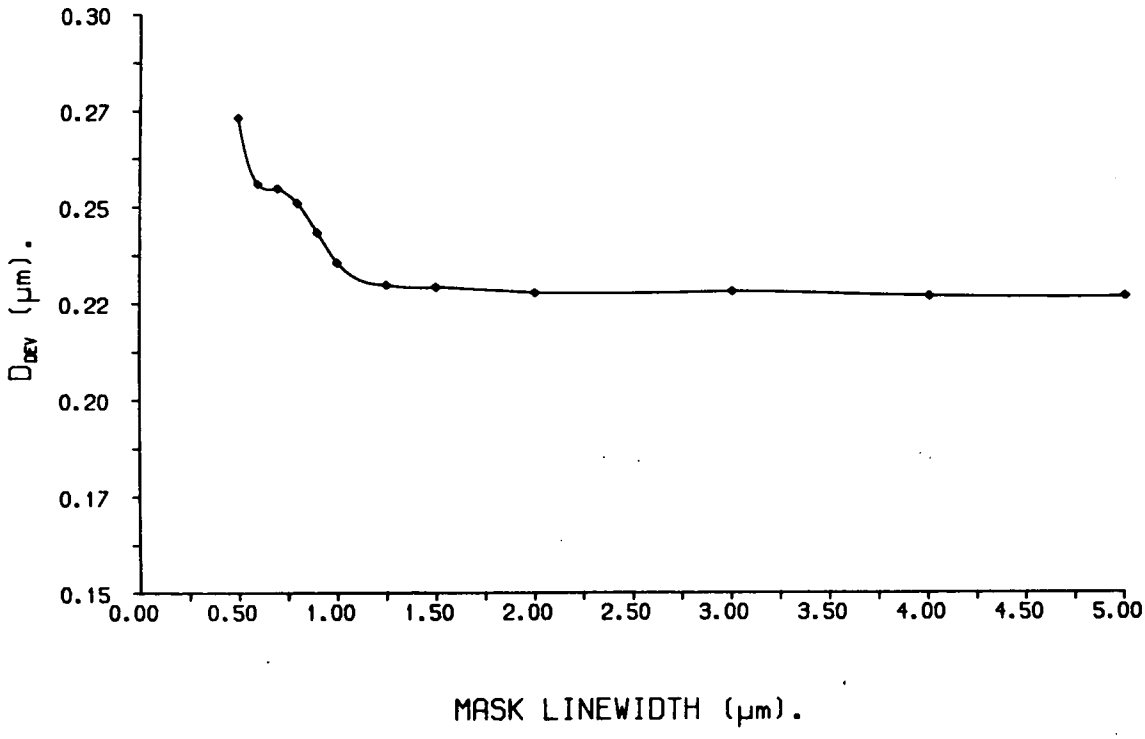

Figure 8.4: D_{dev} vs. mask linewidth (spacewidth=1.0μm).

- The curves corresponding to constant line, varying space and varying grating size show a much larger variation in D_{dev} than the curve for constant space, varying line (comparing Figures 8.5 and 8.6 with 8.4). Figures 8.5 and 8.6 show very similar features; in particular all the curves show a peak in developed spacewidth at fairly small feature sizes before the sharp fall off which would have been expected as the spacewidth was decreased. The feature size at which this peak in Figure 8.6 occurs, normalised by λ/NA, is given in Table 8.2. The fact that the peak position is roughly proportional to λ/NA suggests that this feature is dependent on the imaging process.

Figures 8.7 and 8.8 show the aerial image parameter D_{ith} as a function of mask linewidth and spacewidth respectively, as obtained from VARYIM. The existence of the peak in Figure 8.8 is probably related to a diffraction fringe from one edge of a mask space coinciding with the diffraction fringe from the other edge at this value of spacewidth. Comparison of these figures with Figures 8.4 and 8.5 illustrates the good functional relationship between the position of the intensity threshold and the final developed spacewidth.

Figures 8.9–8.12 show contour plots of D_{ith} and I_{me} respectively, over the full

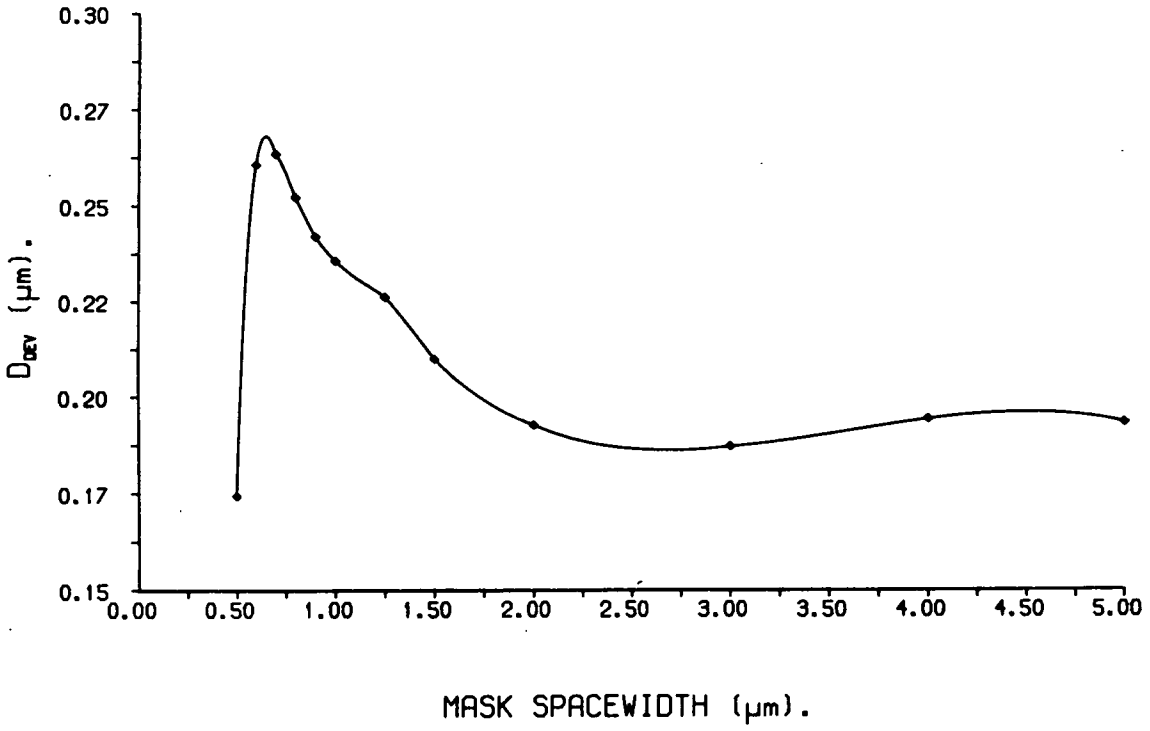

Figure 8.5: D_{dev} vs. mask spacewidth (linewidth=1.0μm).

Peak Position. (μm)	λ/NA. (μm)	Peak Position$\times$NA/λ.
0.54	1.04	0.520
0.61	1.15	0.532
0.81	1.56	0.520

Table 8.2: Position of peak in Figure 8.6 (normalised by λ/NA).

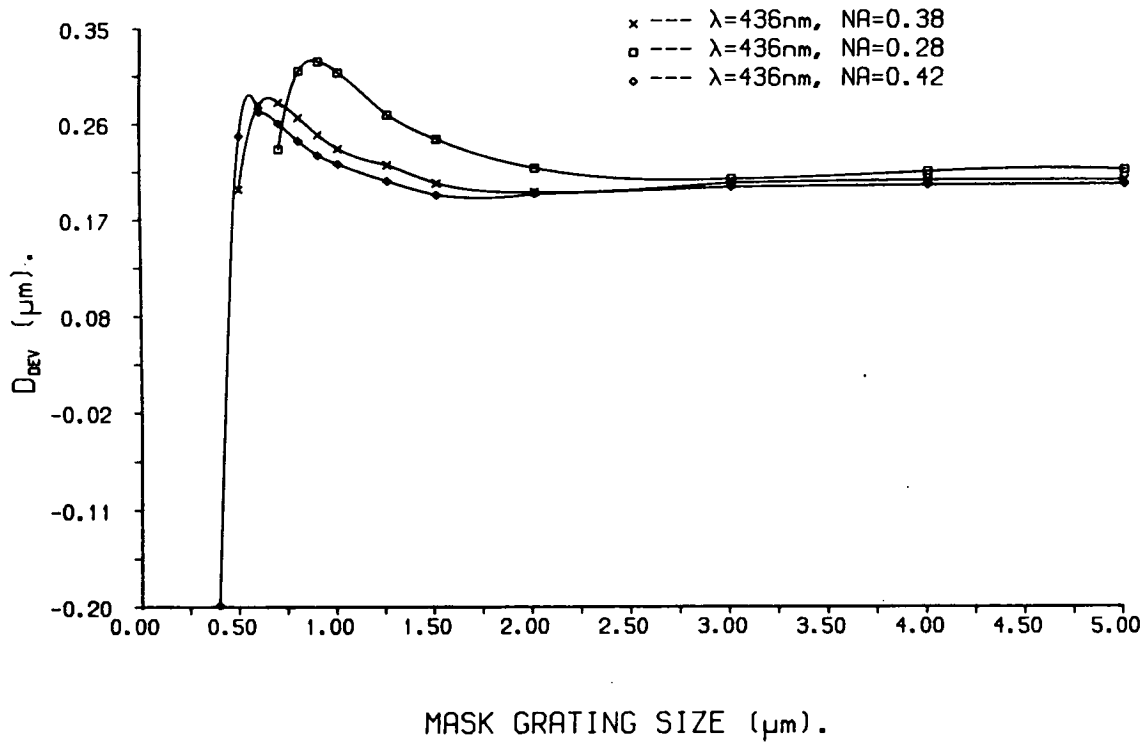

Figure 8.6: D_{dev} vs. grating size, for three different lenses.

12×12 array of L_o and S_o, for σ=0.5 and 0.7. The shape of the contour curves illustrate that, at least for low values of normalised spacewidth ($S_o \leq 2.4$) I_{me} and D_{ith} are stronger functions of spacewidth than they are of linewidth (since the contour lines are roughly parallel to the linewidth axis). Again this suggests that the developed feature size depends little on optical proximity effects, but depends very strongly on the spacewidth itself. Since it has already been seen that D_{ith} corresponds closely to the feature size on the wafer, the similarity between plots of I_{me} and D_{ith} suggests that the intensity at the mask edge may also be used as a good indicator for final developed feature size.

Figures 8.13 and 8.14 show the intensity at the mask edge for the coherent and incoherent cases respectively. Although these are of little importance in the practical lithographic sense, they are are interesting in that they display anti-symmetry around the line=space axis. This leads to the conclusion that line and space are only complementary when the imaging is linear, either in intensity or in amplitude. Of particular interest is the fact that, in the incoherent case, $I_{me} = 0.5$ when L and S are equal. Intuitively, this what we would expect.

The ill-behaved nature of the image characteristics in the coherent case can be ascribed to the discrete nature of the cut off in the coherent transfer function

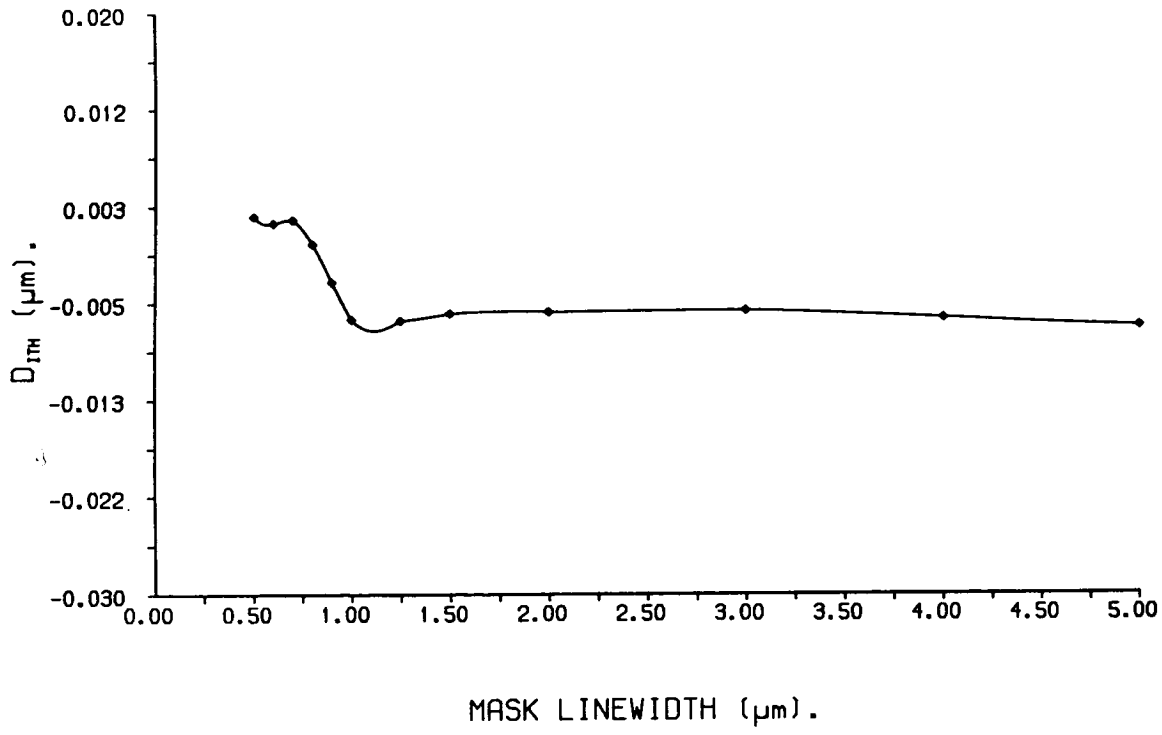

Figure 8.7: D_{ith} vs. mask linewidth (spacewidth=1.0μm).

[106], which leads to additional spatial frequencies being passed completely by the lens with only a very small change in either L_o or S_o. It is possible in this case that the imaging information as presented in the contour plot is incomplete due to the fact that the plotting algorithm will use approximately linear interpolation between the grid points, which may not be valid if the image characteristics vary over a very small range.

8.5 Factors Affecting the Process Latitude.

The developed profile characteristic which was assumed to be indicative of process latitude was the sidewall slope of the developed resist pattern (S_{dev}), since in general a steep sidewall slope corresponds to minimal sideways erosion of resist and hence to good process latitude. It has previously been predicted theoretically that the sidewall slope of the developed profile is related directly to the intensity gradient in the aerial image [25]. We would therefore expect to see some correspondence between the developed sidewall slope, as predicted by SPESA, and the intensity gradient at the mask edge, as predicted by VARYIM.

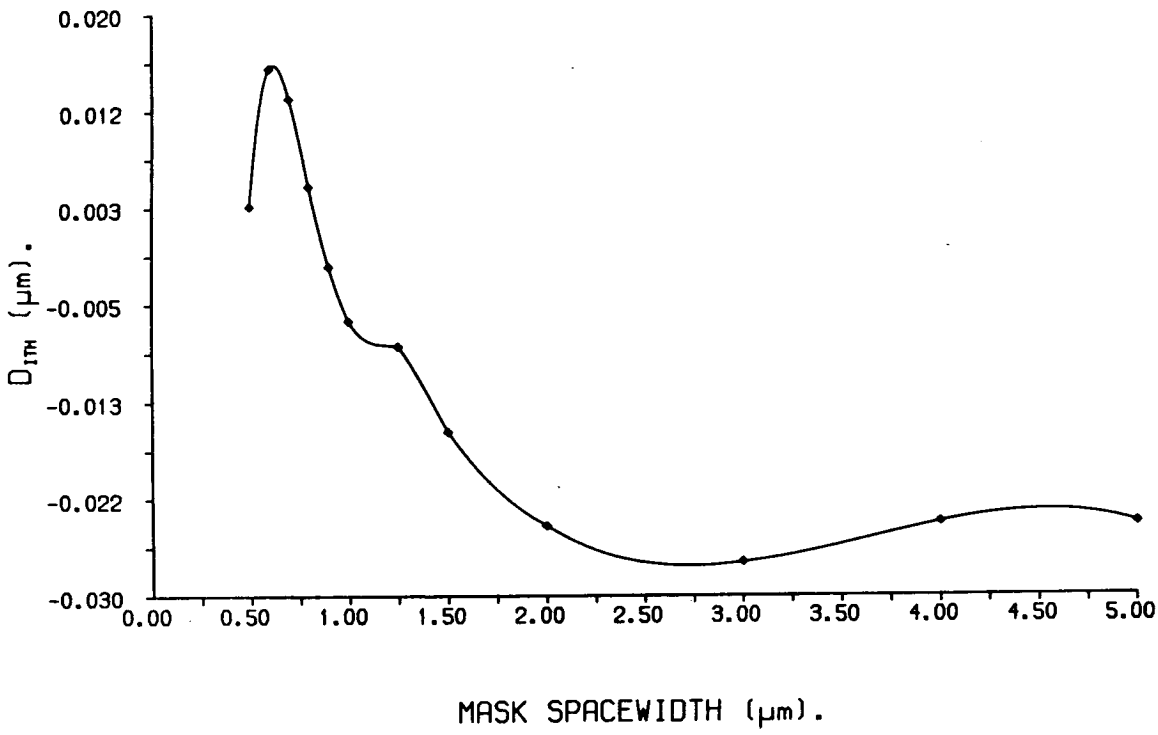

Figure 8.8: D_{ith} vs. mask spacewidth (linewidth=1.0μm).

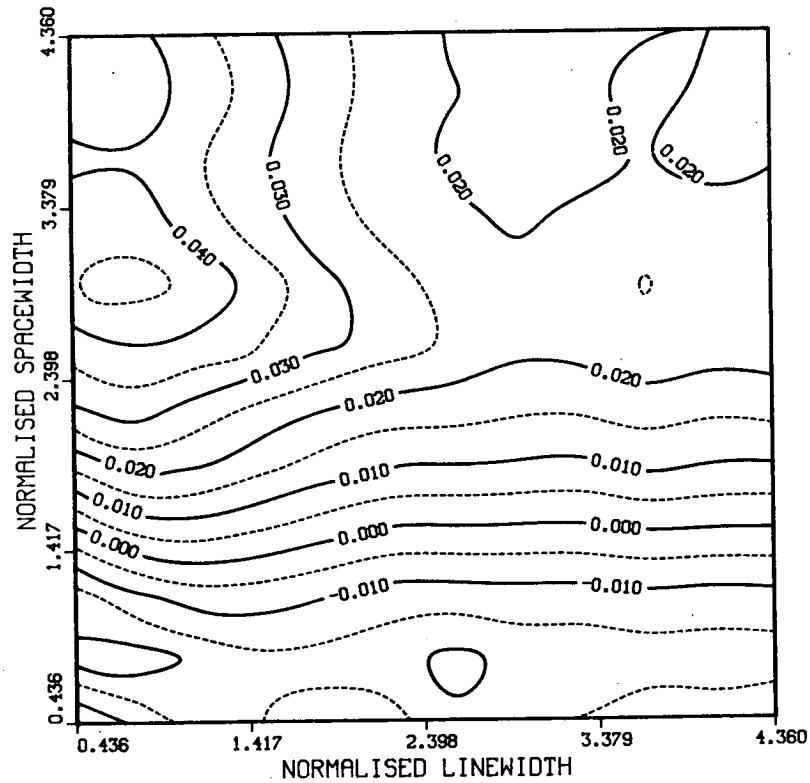

Figure 8.9: Contour plot of x-position of intensity threshold. σ=0.5, defocus=0.0 (Rayleigh units).

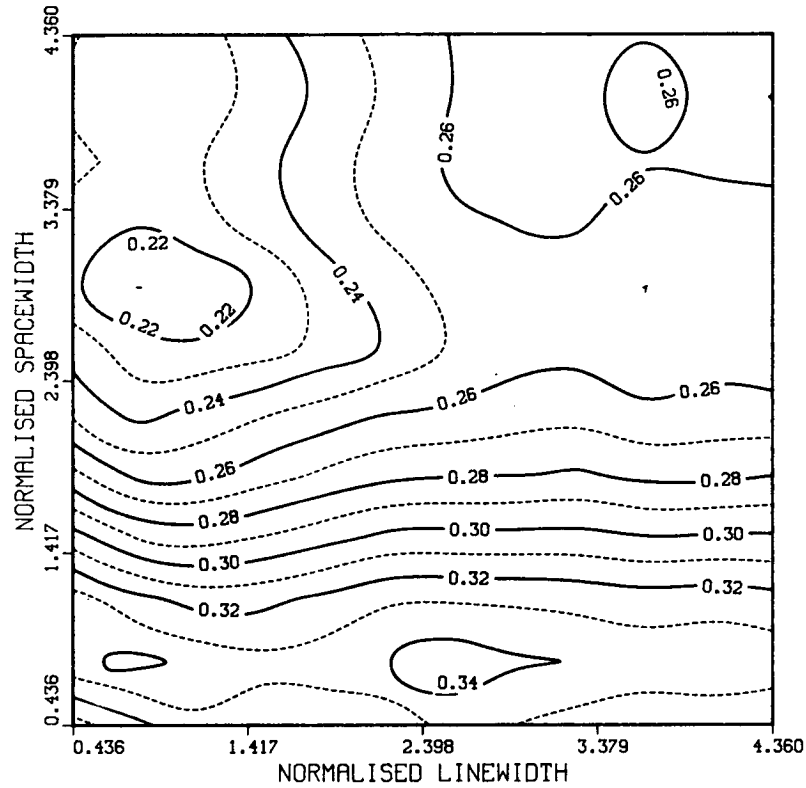

Figure 8.10: Contour plot of intensity at mask edge. σ=0.5, defocus=0.0 (Rayleigh units).

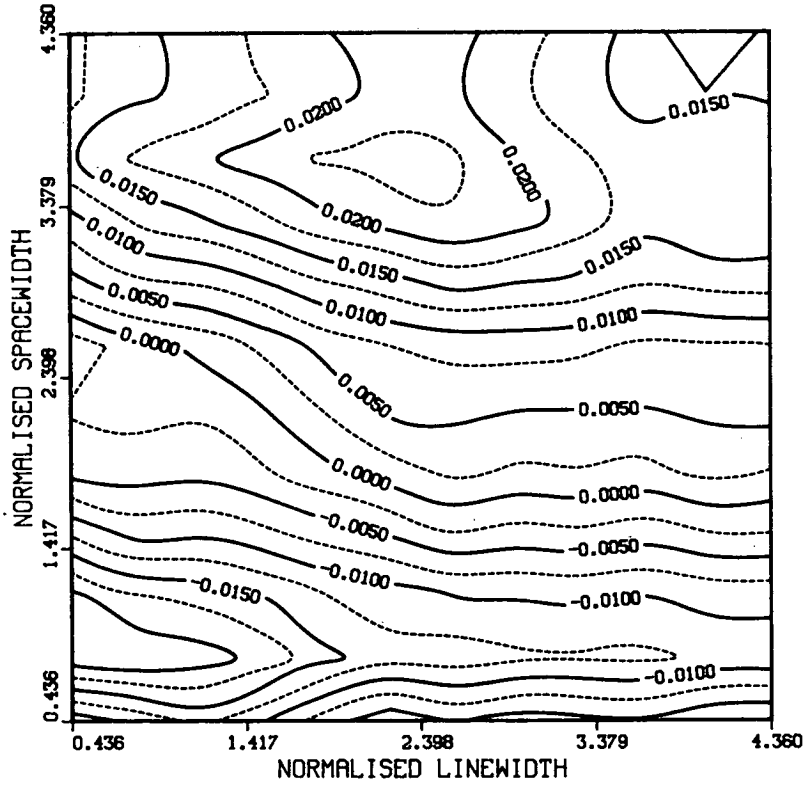

Figure 8.11: Contour plot of x-position of intensity threshold. σ=0.7, defocus=0.0 (Rayleigh units).

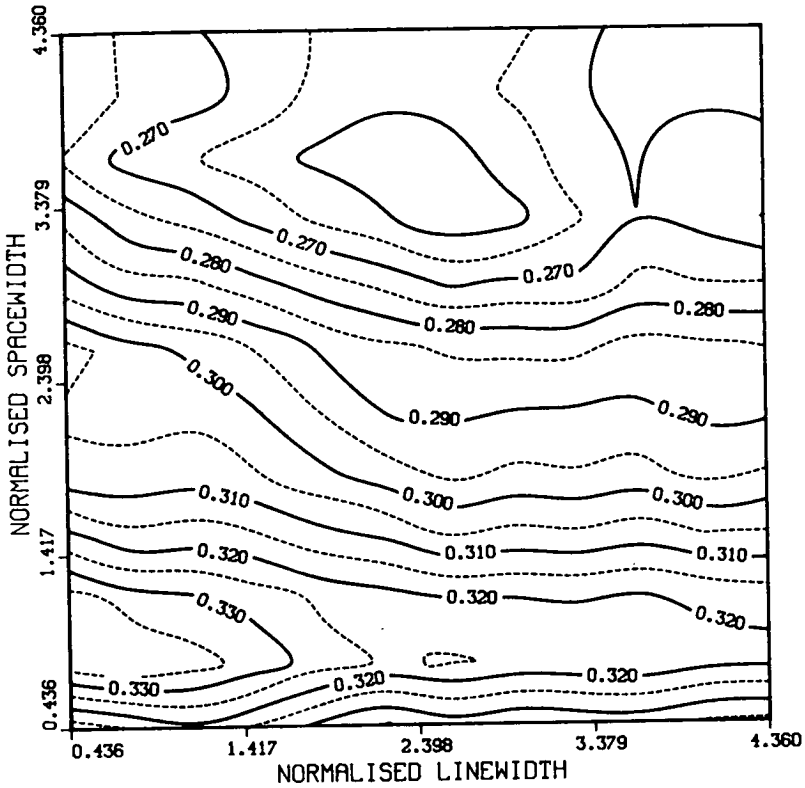

Figure 8.12: Contour plot of intensity at mask edge. σ=0.7, defocus=0.0 (Rayleigh units).

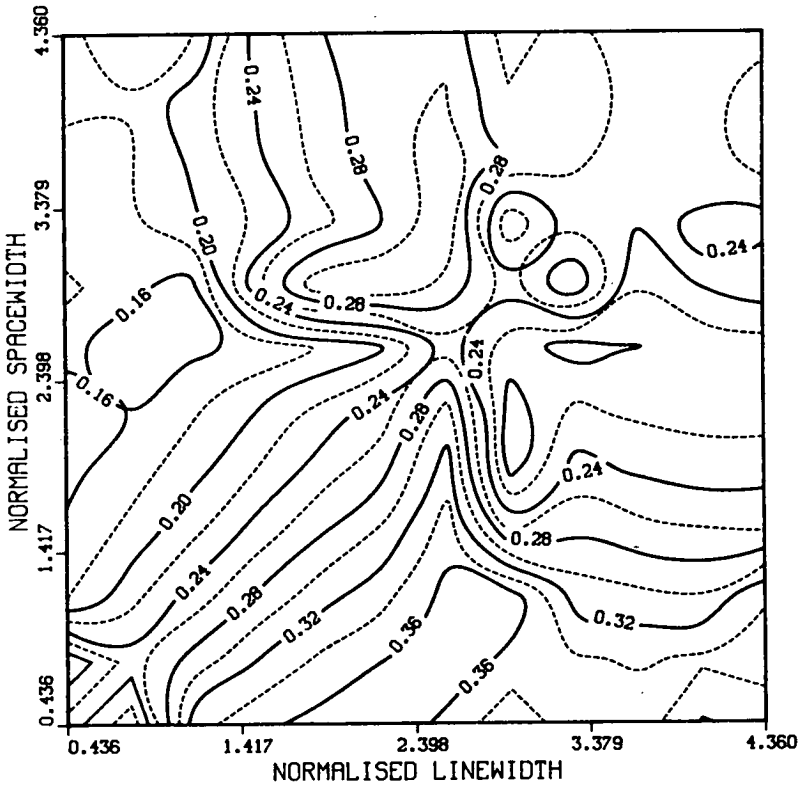

Figure 8.13: Contour plot of intensity at mask edge. σ=0.01, defocus=0.0 (Rayleigh units).

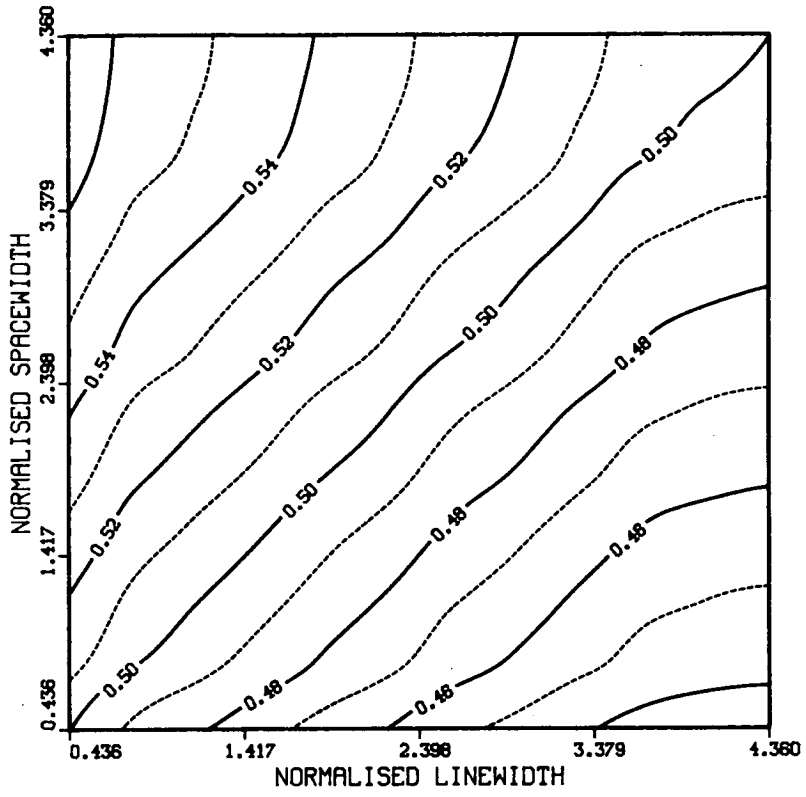

Figure 8.14: Contour plot of intensity at mask edge. σ=3.0, defocus=0.0 (Rayleigh units).

Figures 8.15 and 8.16 show the sidewall slope of the developed resist profile as a function of linewidth and spacewidth respectively, with an illumination partial coherence of 0.7, while Figures 8.17 and 8.18 show the intensity gradient at the mask edge (in normalised intensity units/μm) as a function of the same parameters. The curves in 8.17 and 8.18 follow the same basic trends as those in Figures 8.15 and 8.16, and so illustrate a good functional relationship between S_{dev} and G_{me}.

Figure 8.19 shows a contour plot of the gradient at the mask edge over the whole 12×12 array of L_o and S_o for a partial coherence of σ=0.7. The plot displays a peak at $S_o \simeq 1.4$ and $L_o \simeq 3.3$ and indicates that the intensity gradient is a stronger function of spacewidth than of linewidth (as indeed was the x-position of the intensity threshold). It should also be noted that G_{me} falls off rapidly at low values of spacewidth.

Figures 8.20–8.22 contain contour plots of I_{min}, I_{max} and C_{im} for σ=0.7. The shape of the curves suggests that the contrast is related more to intensity minimum than to intensity maximum. There is nothing particularly profound in this conclusion; it arises simply from the definition of image contrast (a given percentage change in I_{min} will cause a larger change in C_{im} than the same per-

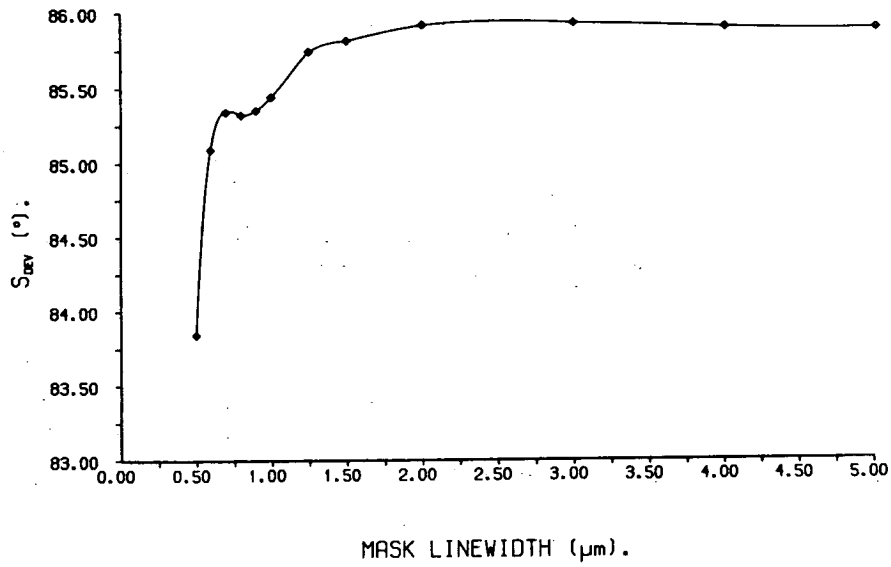

Figure 8.15: S_{dev} vs. mask linewidth (spacewidth=1.0μm).

centage change in I_{max}). This behaviour, plus the fact that contrast (Figure 8.22) shows little correlation with resist sidewall slope (Figure 8.15) leads to the conclusion that image contrast is a bad indicator for process latitude.

The introduction of defocus into the system results in a reduction of gradient at the mask edge, thus decreasing the process latitude for dose and development variations. In Figures 8.23 and 8.24 contour plots of G_{me} are shown for 1μm and 2μm defocus (corresponding to 0.66 and 1.33 Rayleigh units (RU)†, for this lens), and σ=0.7. The general shape of the plots is the same, but the gradient values are degraded upon defocusing (see Figure 8.19 for perfect focus case). For comparison, in Figures 8.25 and 8.26, a plot of intensity at the mask edge is shown for the same two values of defocus. Although the shape of the curves changes significantly at a defocus of 1.33RU, the I_{me} values (and thus the range of linewidth variations) change very little. This illustrates that defocusing does not prevent imaging of small patterns; it does, however, prevent us from doing this in a comfortable way, since it reduces process latitude.

†1 RU= $\lambda/2(NA)^2$

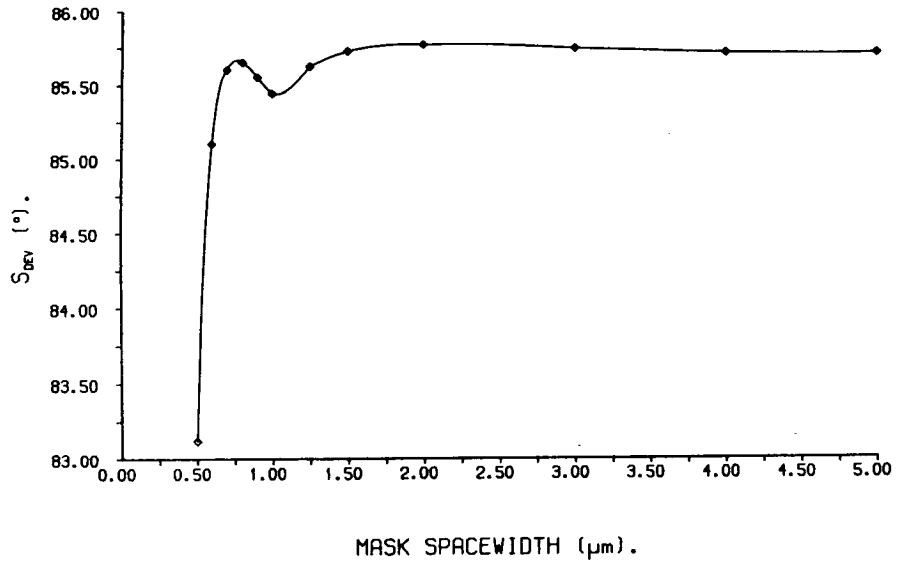

Figure 8.16: S_{dev} vs. mask spacewidth (linewidth=1.0μm).

8.6 Conclusions.

A novel method of presenting image characteristics over the full line/space domain has been introduced. Using this method it has been demonstrated that for the purposes of simulation, the x-position of the intensity threshold and the intensity at the nominal mask edge serve as good indicators for the developed feature size on the wafer. The correspondence between these parameters allows us to obtain information regarding feature size over a wide range of values very quickly, via these plots. These indicators have demonstrated that, under normal working conditions of partial coherence, deviation of developed feature size from nominal feature size depends more critically on mask spacewidth than on mask linewidth.

It has also been shown that the intensity gradient at the mask edge can be used as an indicator for process latitude since it is related to the sidewall slope of the developed resist profile. This is to be expected, since, if the gradient is small, dose variations will strongly influence the lateral position which just received a threshold dose. Small gradients also make the process more sensitive to development variations, since nominally unexposed areas may still receive a relatively

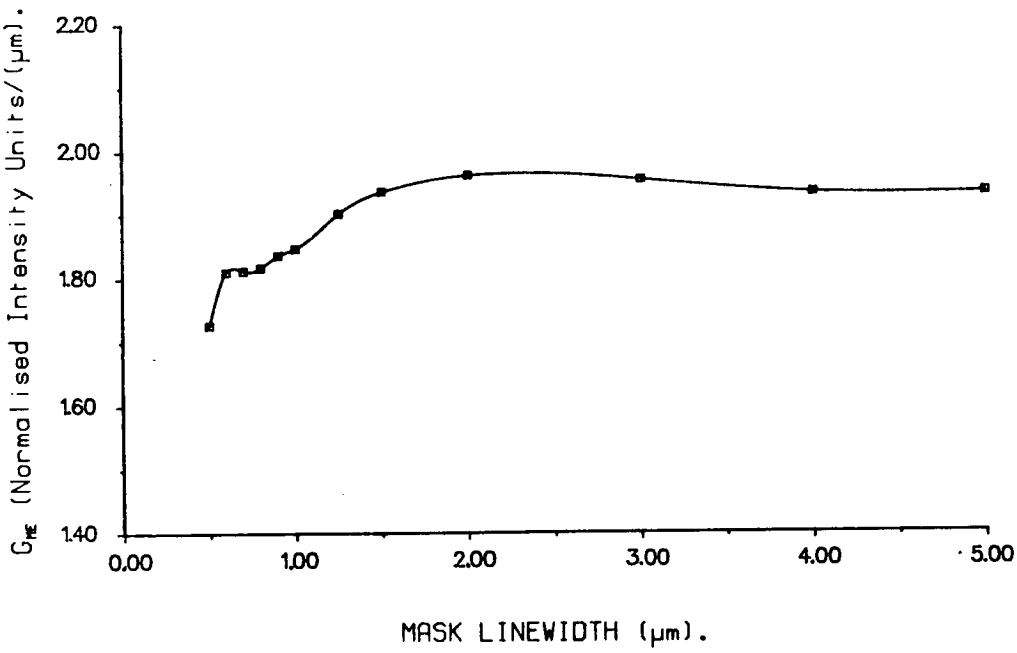

Figure 8.17: G_{me} vs. mask linewidth (spacewidth=1.0μm).

high exposure dose. We have seen that the gradient is also a stronger function of spacewidth than of linewidth. This implies that process latitude is degraded more quickly with decreasing spacewidth than with decreasing linewidth.

Image contrast has been shown to be a bad indicator for both developed feature size and also for process latitude, since it is by definition dominated by intensity minimum, which bears little relation to either of these characteristics. In addition it has been confirmed that defocus does not necessarily affect feature size, but strongly degrades process latitude.

One particular area of application for this particular type of plot is in the field of reticle biasing, which is likely to become increasingly important over the next few years. By using an indicator for developed linewidth (either D_{ith} or I_{me}), the amount of deviation from the mask linewidth can be estimated for gratings, isolated lines, and isolated spaces. From this information we can estimate for which type of features we require to bias the linewidth on the reticle, as well as estimating the amount of biasing which will be required.

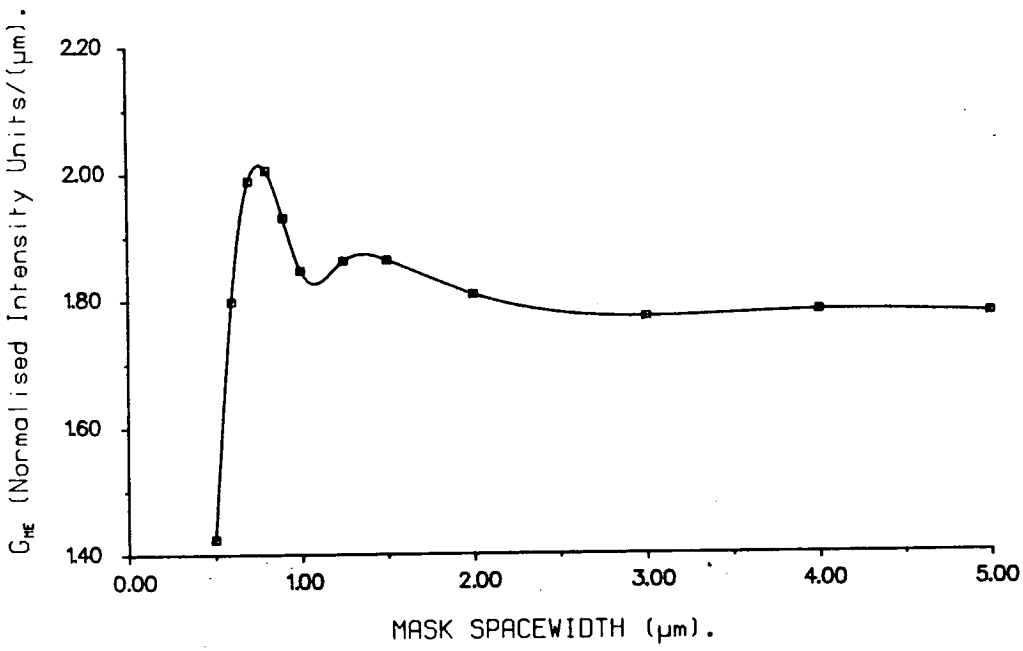

Figure 8.18: G_{me} vs. mask spacewidth (linewidth=1.0μm).

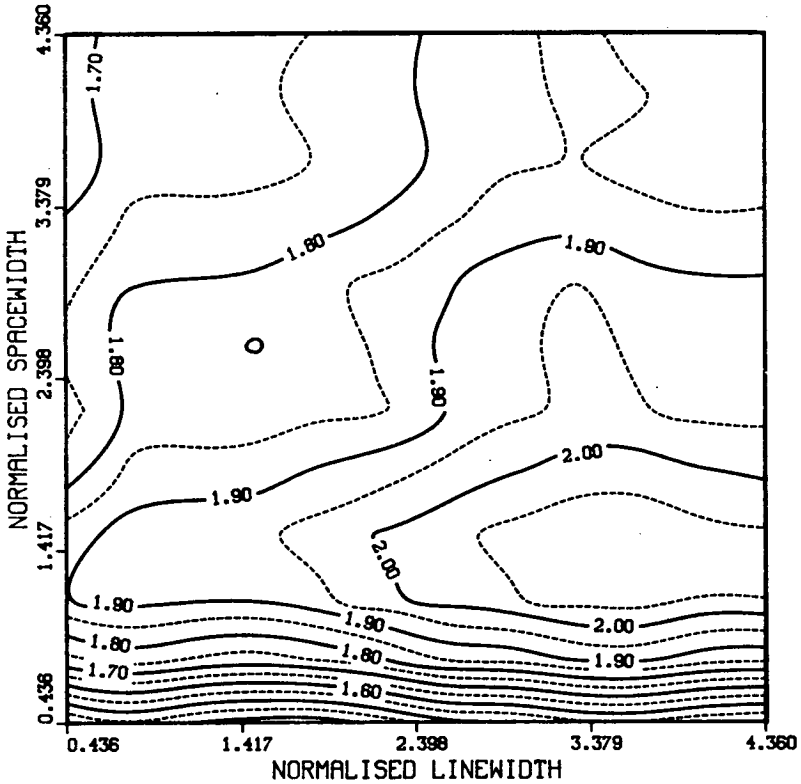

Figure 8.19: Contour plot of gradient at the mask edge. σ=0.7, defocus=0.0 (Rayleigh units).

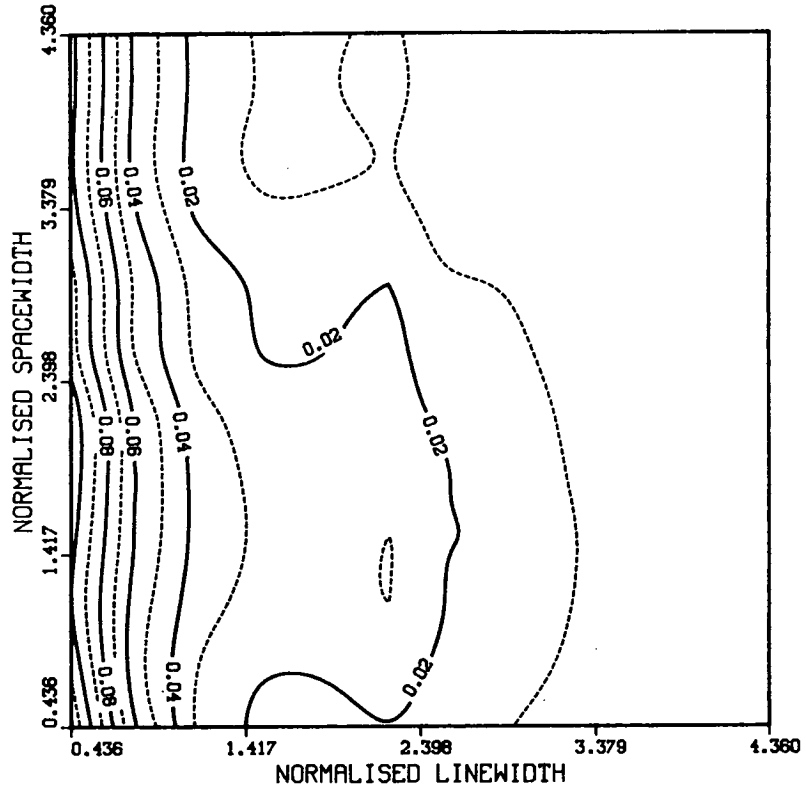

Figure 8.20: Contour plot of intensity minimum. σ=0.7, defocus=0.0 (Rayleigh units).

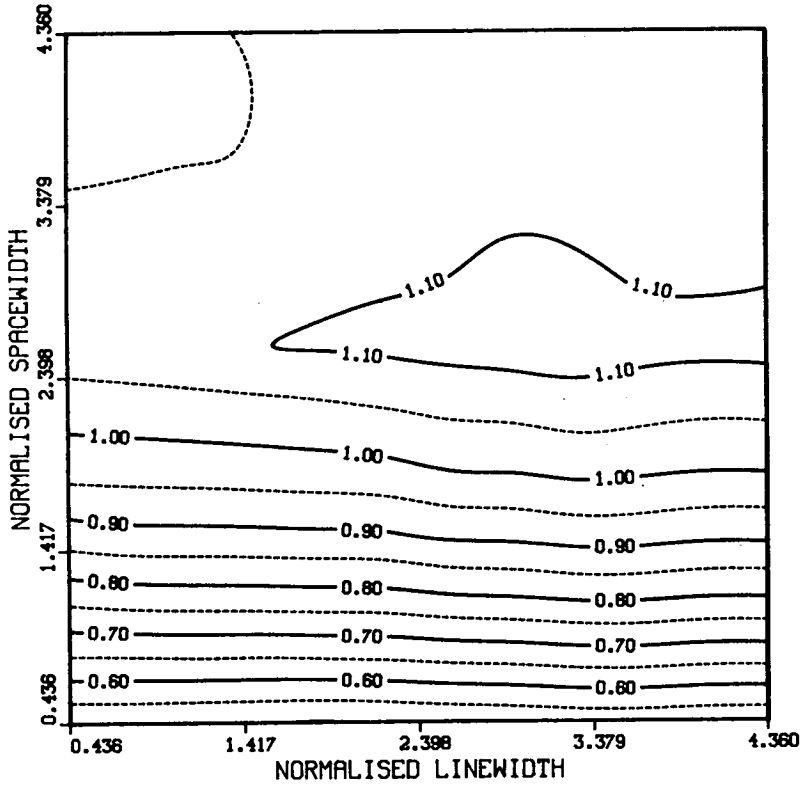

Figure 8.21: Contour plot of intensity maximum. σ=0.7, defocus=0.0 (Rayleigh units).

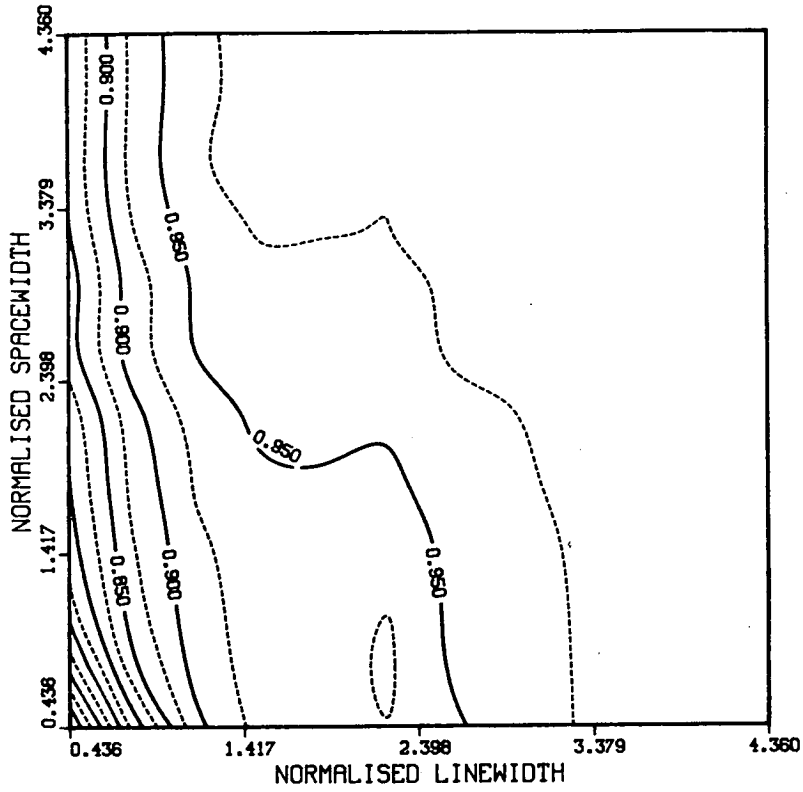

Figure 8.22: Contour plot of image contrast. σ=0.7, defocus=0.0 (Rayleigh units).

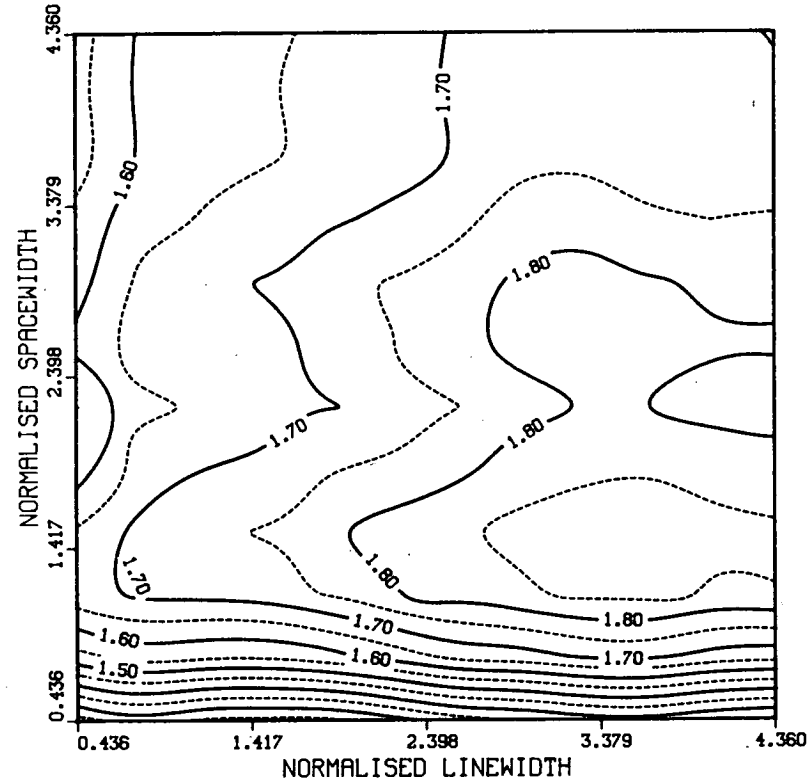

Figure 8.23: Contour plot of gradient at the mask edge. σ=0.7, defocus=0.66 (Rayleigh units).

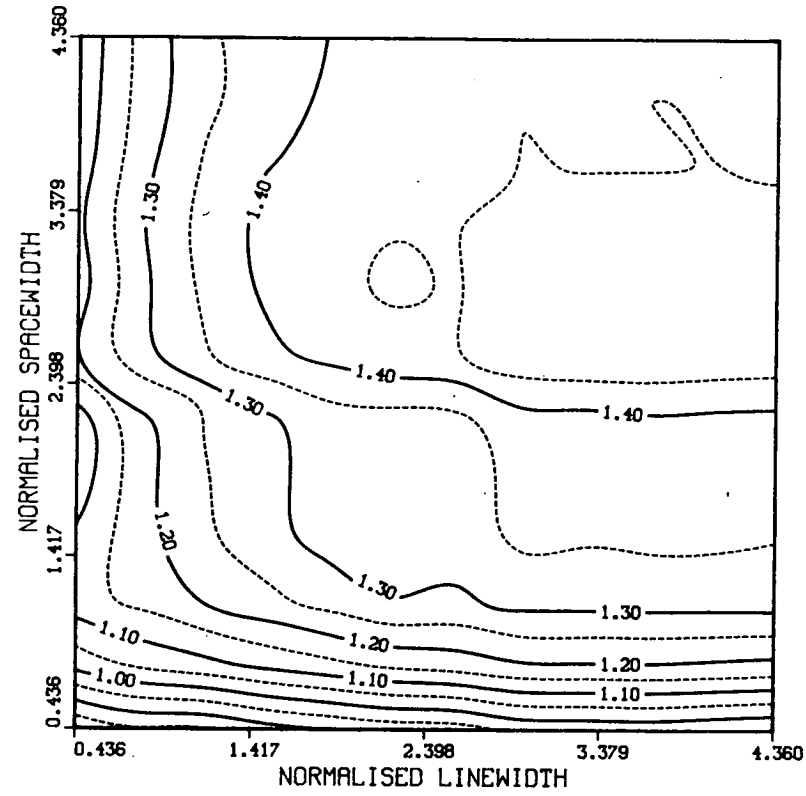

Figure 8.24: Contour plot of gradient at the mask edge. σ=0.7, defocus=1.33 (Rayleigh units).

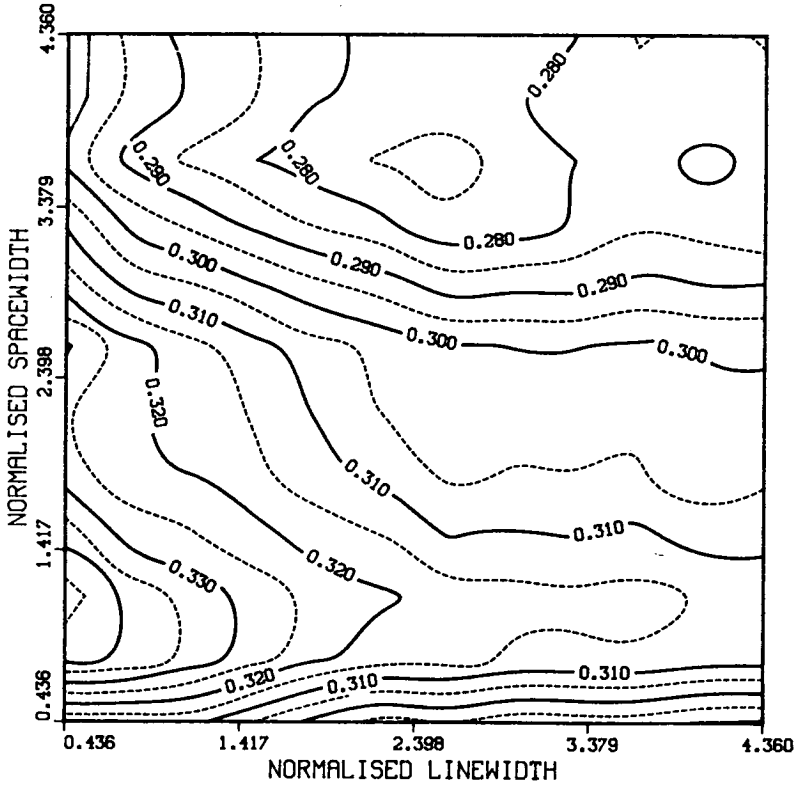

Figure 8.25: Contour plot of intensity at the mask edge. σ=0.7, defocus=0.66 (Rayleigh units).

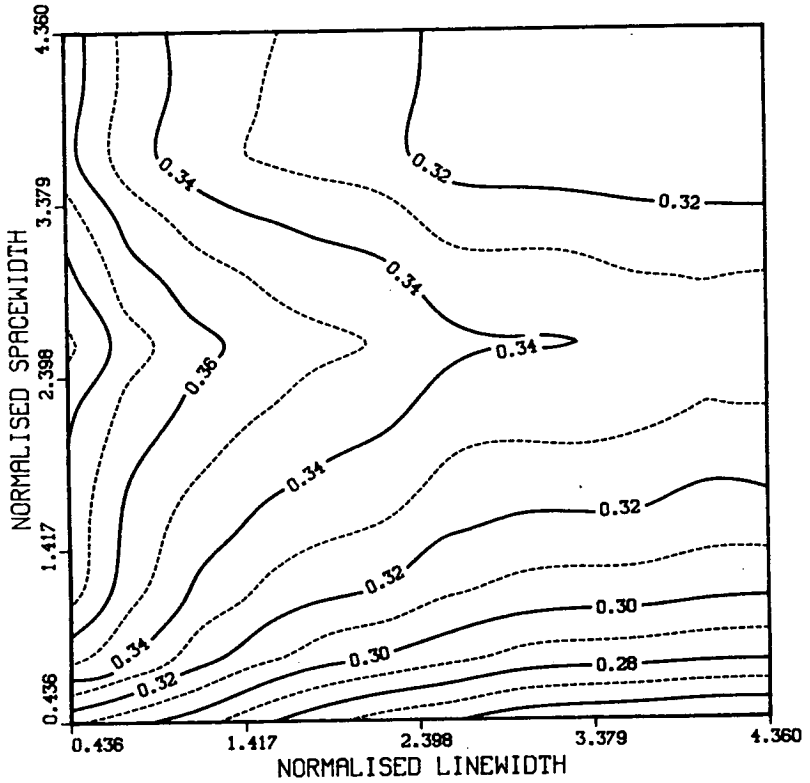

Figure 8.26: Contour plot of intensity at the mask edge. σ=0.7, defocus=1.33 (Rayleigh units).

Chapter 9

Conclusions and Further Work.

9.1 Conclusions.

Optical lithography will continue to be the dominant exposure technology in the microelectronics industry well into the next century. Already devices with a minimum feature size of 1.0μm are in volume production, with feature sizes as small as 0.7μm being produced by optical means in development environments. With the likelihood that optical lithography will eventually reach and probably exceed the 0.5μm barrier, it is extremely unlikely that rival technologies will be able to compete economically in production facilities for some time.

The continued success of this technology should not be taken for granted, however, and requires considerable effort in many areas to ensure that its potential is fully realised. Systems designers must continue with the development of alignment systems to improve overlay accuracy. It is likely also, that as alignment accuracy continues to improve, more emphasis will be put on stage control to ensure that the wafer is exposed at precisely that position which the alignment system has determined to be the correct one. (Previously this has not been necessary, since stage accuracy has been very much better than alignment accuracy.) Lens designers and optical researchers must continue to develop new generations of lenses, as well as to refine lens manufacturing techniques, as lens distortions become more influential in determining total pattern overlay accuracy, particularly in production facilities where it is impossible to dedicate batches to a single machine. In addition to this, quartz lenses must now be developed, which can operate well into the UV (~200–300nm), in anticipation of the arrival of excimer laser sources.

Reduced device dimensions create more challenges than merely those of reducing device dimensions and improving pattern overlay. As circuits continue

to scale more rapidly laterally than vertically, problems of step coverage and depth of focus become more severe. Process engineers must continue to develop advanced resist processes, perhaps eventually using a mixture of single-level, multi-level, and advanced techniques (image reversal, or contrast enhancement), depending upon the requirements of a particular process layer.

This thesis has looked at many of these problems. A review has been presented of the current status of lithographic processing, as well as the problems facing it, and possible solutions to the problems. An in-depth review of wafer stepper alignment systems has also been given, highlighting some of the advantages and disadvantages of the various systems on offer. A closer look at light field systems in general (Chapters 5 and 6 offer conclusions which may be applied to any light field alignment system), and at the Optimetrix in particular (Chapter 4), revealed many ways in which the performance of these systems may be improved. The solutions presented include the following:

1. Software spatial filtering of the alignment signal to remove high frequency components (diffraction fringes), in order to eliminate multiple peaks in the correlation function.

2. The use of hardware filtering to reduce the number of fringes in the alignment signal. Use of simulation has shown that a combination of narrow marks ($\sim$2.0μm) and low numerical aperture alignment optics produces profiles with less fringing than standard markers ($\sim$6.0μm) and high numerical aperture alignment optics. This can be thought of as a hardware version of case 1 above, although in this case the filtering is non-linear, due to the partially coherent nature of the illumination.

3. The use of different correlation functions, similar in performance, but faster computationally, than the standard auto-correlation.

4. Optimum use of software parameters on the machine to improve throughput (reducing the value of SPA in DIA CAMERA - see Chapter 4), and accuracy (reducing the value of COM DEB location S474 - see Chapter 6). Until now, no systematic study of the effect of these parameters has been made.

5. The use of novel alignment markers to reduce diffraction fringes. This has been shown to be effective on-line, during the process run of a batch through the E.M.F. 1.5μm NMOS process.

It should be noted that these solutions may be used in parallel, and that they have their own advantages and disadvantages. Spatial filtering (software or hardware) is likely to produce a poor alignment signal when used on markers with low contrast, but good edge definition, since the modulation for 6.0μm features is low for this type of mark. Such targets are relatively rare, but the ability to set a software switch to turn off the filtering should be included in the machine, in case of such an eventuality.

Optimising the machine software parameters can prevent the machine from latching on to the wrong peak in the auto-correlation function (S474=20), but this approach has the drawback of causing auto-align failures if a peak cannot be found within the given capture range. This reduces throughput, and is less accurate than successful alignment which can be achieved by filtering.

Tailoring of alignment markers has the disadvantage, mentioned in Chapter 5, that linewidth control is relatively poor when printing small features required to reduce fringing, and hence the alignment signal profiles will vary also. This problem can be overcome, however, by using low numerical aperture alignment optics.

A review of simulation in photo-lithography has also been given, concentrating on SAMPLE, but also including other less well known programs. A simulation study of aerial imaging, using the VARYIM and SPESA programs, has been presented, with particular emphasis on how the features are transferred from the aerial image to the developed feature. At small dimensions, this has particular application to the problem of reticle biassing, a topic which will become extremely important in the next few years.

9.2 Further Work.

A number of ideas are presented in this thesis which require some additional work:

1. The OASIS program has not been used exhaustively since its completion. More work can be done with this program, particularly in the area of optimising the low-pass filtering. All the results presented in Chapter 4 used a 99-point filter, which may be more than is required (thus taking longer than necessary). The number of taps should be kept as low as possible, while still maintaining a single-peaked correlation function, in order to maximise wafer throughput. The number of taps in the filter is easily adjusted, by editing one line in one file, and thus the effect of a

different number of taps can easily be determined.

The use of software filtering which passed only 6.0μm features could also be investigated (this approach would be similar to the ASM hardware filtering mentioned in Chapter 3). The use of such filtering would necessitate the use of more regular markers, for example, a 6.0μm dark area on either side of an 18.0μm reticle window, with the window containing a 6.0μm wafer marker in the center. This approach would again rely on having some image contrast, and not just on wafer symmetry.

The final test of this spatial filtering technique should, of course, be on-line. The implementation of these algorithms in hardware, and trial on the machine itself, would be one major extension of the work presented here.

2. Incorporation of a suitable stop in the lens during alignment to reduce the numerical aperture would be fairly simple to accomplish. Care should be taken to pick an optimum numerical aperture for the alignment system, such that fringes of less than 3.0μm width are filtered from the intensity profile (filtering out features of this size worked well for the software filtering in Chapter 4), while still maintaining sufficient slope in the profile so that the alignment accuracy is not degraded. The effect of this modification could be assessed manually before building automatic positioning of the stop (for alignment only) into the machine.

3. On-line implementation of the modulus difference algorithm should improve the throughput of the system in the die-by-die mode. This would involve only some reprogramming of the microprocessor control system on the machine, and could be accomplished fairly quickly given access to the correct information by the manufacturer.

4. Additional characterisation of alternative alignment structures should be undertaken, in order to fully assess the effect of the new markers on all layers. In addition, it would be an attractive addition to the experiment to include the metal layer (or layers) in the characterisation, and to determine the overlay errors by means of electrical measurement techniques [92] [45]. This would give an estimate of the performance of these markers which was completely independent of the stepper itself.

5. The practical determination of linewidth variation as a function of feature size and proximity should be thoroughly investigated. Since this effect will vary, depending upon the lens in use, with conditions of partial coherence

and resist processing, this is the sort of topic which would benefit greatly from process simulation. *Extreme* care should be taken, however, that the simulation uses realistic parameters, and that the results compare well (at least in a few known situations) with practical studies of the effect. Reticle biassing is a topic which needs to be addressed in some detail for the full-scale production of sub-micron devices.

Optical lithography, it can be concluded, is the most attractive method available for the exposure of integrated circuits in the semiconductor industry today, and will continue to be so well into the next century. This is due to a combination of many factors, such as the high degree of parallelism inherent in the transfer of information by optical means, the ready availability of suitable mask materials and light sources, and the large amount of experience which has been built up with the appropriate resist materials. Improvements in optical lithography are becoming harder to achieve every year, however, since the innovative stage of this technology (with the possible exception of holographic techniques) is over. Each subsequent improvement is now incremental, and has to be hard won. Assuming that progress continues at its present rate, optical lithography should be capable of sub-half micron resolution, and the production of 16 Mbit dynamic-RAMs, a situation which would have been regarded as unthinkable just a few years ago.

References

[1] G. E. Moore. VLSI: some fundamental challenges. *IEEE Spectrum*, 16(4):30–37, 1979.

[2] S. M. Sze, editor. *VLSI Technology*, page 2 et seq. McGraw-Hill, Singapore, 1983.

[3] D. F. Barbe, editor. *VLSI Fundamentals and Applications*, chapter 2. Springer-Verlag, Berlin, 1982.

[4] V. Marriot. High resolution positive resist developers: A technique for functional evaluation and process optimization. In *SPIE Optical Microlithography II: Technology for the 1980s*, 1983.

[5] J. S. Petersen and A. E. Kozlowski. Optical performance and process characterizations of several high contrast metal-ion-free developer processes. In *SPIE Advances in Resist Technology I*, pages 46–56, 1984.

[6] M. C. King. Future Developments for 1:1 Projection Photolithography. *IEEE Trans. Electron Devices.*, ED-26(4):711–716, 1979.

[7] D. A. Doane. Optical Lithography in the 1-μm Limit. *Solid State Technology*, 23(2):101–114, 1980.

[8] S. Wittekoek. Optical Microlithography for Microcircuits. In *Microcircuit Engineering '82*, pages 155–170, Delft, The Netherlands, 1982.

[9] S. M. Sze. *Physics of Semicoductor Devices.* Wiley, New York, 1981.

[10] J. A. Appels et al. Local Oxidation of Silicon and its Applications in Semiconductor Technology. *Philips Res. Rep.*, 25:118, 1970.

[11] J. L. Moll, K. Y. Chiu and J. Manolin. A Bird's Beak Free Local Oxidation Technology Feasible for VLSI Circuit's Fabrication. *IEEE Trans. Electron Devices.*, ED-29(4):536–540, 1982.

[12] K. Y. Chiu et al. The Sloped Wall SWAMI - A Defect Free Zero Bird's Beak Local Oxidation Process for Scaled VLSI Technology. *IEEE Trans. Electron Devices.*, ED-30(11):1506–1511, 1983.

[13] N. Matsukawa et al. Selective Poly-Silicon Oxidation Technology for VLSI Isolation. *IEEE Trans. Electron Devices.*, ED-29(4):561–567, 1982.

[14] R. S. L. Lutze, H. P. Vyas and J. S. T. Huang. A Trench-Isolated Submicrometer CMOS Technology. *IEEE Trans. Electron Devices.*, ED-32(5):926–931, 1985.

[15] A. K. Sinha. Refractory Metal Silicides for VLSI Applications. *J. Vac. Sci. Technol.*, 19:778–785, 1981.

[16] C. Korburger et al. Electrical Properties of Composite Evaporated Silicide/Poly-Silicon Electrodes. *J. Electrochem. Soc.*, 129:1307–1312, 1982.

[17] M. E. Alperin et al. Development of the Self-Aligned Titanium Silicide Process for VLSI Applications. *IEEE Trans. Electron Devices.*, ED-32(2):141–149, 1985.

[18] K. L. Wang et al. Composite $TiSi_2/n^+$ Poly-Si Low-Resistivity Gate Electrode and Interconnect for VLSI Device Technology. *IEEE Trans. Electron Devices.*, ED-29:547–553, 1982.

[19] Born and Wolf. *Principles of Optics (6th ed.)*, page 277 et seq. Pergamon Press, Oxford, 1980.

[20] A. R. Neureuther, W. G. Oldham, S. N. Nandgaonkar and M. O'Toole. A General Purpose Simulator for VLSI Lithography and Etching Processes: Part I-Application to Projection Lithography. *IEEE Trans. Electron Devices.*, ED-26(4):717–722, 1979.

[21] D. W. Widmann and H. Binder. Linewidth Variations in Photo-resist Patterns on Profiled Surfaces. *IEEE Trans. Electron Devices.*, ED-22(7):467–471, 1975.

[22] H. Binder and M. Lacombat. Step-and-Repeat Projection Printing for VLSI Circuit Fabrication. *IEEE Trans. Electron Devices.*, ED-26(4):698–704, 1979.

[23] J. Lee. Multilayer Resist Processing: Economic Considerations. *Solid State Technology*, 29(6):143–148, 1986.

[24] H. F. Sandford, J. F. Bohland and S. R. Fine. Dye Effects On Exposure and Development of Positive Photo-resists. *Microelectronic Manufacturing and Testing*, :18–20, Aug. 1985.

[25] H. Keller, W. Arden and L. Mader. Optical Projection Lithography in the Submicron Range. *Solid State Technology*, 26(1):143–150, 1983.

[26] I. Bol. High-resolution Optical Lithography Using Dyed Single-layer Resist. In *Kodak Microelectronics Seminar, Interface '84*, San Diego, 1984.

[27] M. Bolsen. To dye or not to dye - some aspects of today's resist technology. In *Microelectronic Engineering 3*, pages 321–328, Rotterdam, 1985.

[28] Y-C. Lin et al. Some aspects of anti-reflective coating for optical lithography. In *SPIE Advances in Resist Technology I*, page 30, 1984.

[29] F. Jones and M. Hatzakis. *Computer Aided Lithography Modelling Of The Effects Of Processing Parameters On Linewidth Variations When Fabricating Chrome Masks.* Technical Report RC 9136 (#39989), IBM, 1981.

[30] B. J. Lin. Partially Coherent Imaging in Two Dimensions and the Theoretical Limits of Projection Printing in Microfabrication. *IEEE Trans. Electron Devices.*, ED-27(5):9–16, 1980.

[31] M. Lacombat and G. Michel Dubroeucq. Coherent illumination improves step-and-repeat printing on wafers. In *SPIE Developments in Semiconductor Microlithography IV*, pages 28–36, 1979.

[32] M. Hatzakis. Multilayer Resist Systems for Lithography. *Solid State Technology*, 23(2):74–80, 1981.

[33] P. Burggraaf. Multilayer Resist Processing Update. *Semiconductor International*, Aug. 1985.

[34] L. V. Gregor, L. P. McDonnell Bushnell and C. F. Lyons. Multilayer Resist Lithography: Performance and Manufacturability. *Solid State Technology*, 29(6):133–138, 1986.

[35] P. R. West B. F. Griffing. Contrast Enhanced Lithography. *Solid State Technology*, 28(5):152–157, 1985.

[36] B. A. Heath, B. F. Griffing, P. R. West. 0.4-μm Gate-Length Devices Fabricated by Contrast-Enhanced Lithography. *IEEE Electron Device Lett.*, EDL-4(9):317–320, 1983.

[37] M. M. O'Toole. Simulated Performance of Contrast-Enhancement Material. *IEEE Electron Device Lett.*, EDL-6(6):282–284, 1985.

[38] M. L. Long and J. Newman. Image reversal techniques with standard positive photo-resist. In *SPIE Advances in Resist Technology I*, pages 189–193, 1984.

[39] S. A. MacDonald et al. Image reversal: The production of a negative image in a positive photo-resist. In *Kodak Microelectronics Seminar, Interface '82*, pages 114–117, 1982.

[40] R. Sigush, H. Klose and W. Arden. Image Reversal of Positive Photo-resist: Characterization and Modelling. *IEEE Trans. Electron Devices.*, ED-32(9):1654–1661, 1985.

[41] H. Moritz. Optical Single Layer Lift-Off Process. *IEEE Trans. Electron Devices.*, ED-32(3):672–676, 1985.

[42] V. Miller and H. L. Stover. Submicron Optical Lithography: I-Line Wafer Stepper and Photo-resist Technology. *Solid State Technology*, 28:127–136, 1985.

[43] A. Neureuther, D. C. Hofer, C. G. Wilson and M. Hakey. Characterization of the induction effect at mid-ultraviolet exposure: application to AZ2400 at 313 nm. In *SPIE Optical Microlithography I: Technology for the Mid-1980s*, pages 196–205, 1982.

[44] L. D. Yau. Process-Induced Distortion in Silicon Wafers. *IEEE Trans. Electron Devices.*, ED-26(9):1299–1305, 1979.

[45] K. H. Nicholas, I. J. Stemp and H. E. Brockman. Automatic Testing and Analysis of Misregistrations Found in Semiconductor Processing. *IEEE Trans. Electron Devices.*, ED-26(4):729–732, 1979.

[46] Born and Wolf. *Principles of Optics (6th ed.)*, page 459 et seq. Pergamon Press, Oxford, 1980.

[47] N. E. David and H. L. Stover. Optical Test Structures for Process Control Monitors, Using Wafer Stepper Metrology. *Solid State Technology*, 25:131–141, 1982.

[48] C. K. Van Peski. Minimising Pattern Registration Errors Through Wafer Stepper Matching Techniques. *Solid State Technology*, 25:111–115, 1982.

[49] P. Burggraaf. Wafer Steppers and Lens Options. *Semiconductor International*, :56–63, March 1986.

[50] J. Dey. *Advanced Step and Repeat Aligner for VLSI.* Technical Report, Eaton S.E.O., 1984.

[51] H. E. Mayer and E. W. Loebach. Design and Performance of a New Step-and-Repeat Aligner. In *Microcircuit Engineering '80*, page 191, Delft, The Netherlands, 1980.

[52] H. L. Stover. Stepping into the 80's with Die-by-Die Alignment. *Solid State Technology*, 24(5):112–120, 1981.

[53] A. J. Weger J. M. Lavine, R. B. Fish and C. Simpson. A direct-reticle-referenced alignment system. In *SPIE Optical Microlithography IV*, page 57, 1985.

[54] G. Doemens and P. Mengel. Automatic Mask Alignment for X-Ray Microlithography. *Siemens Forsch.-u. Entwickl.-Ber. Bd.*, 13(2):43, 1984.

[55] Hecht and Zajac. *Optics*, pages 375–376. Addison-Wesley, 1974.

[56] Born and Wolf. *Principles of Optics (6th ed.)*, pages 401–414. Pergamon Press, Oxford, 1980.

[57] A. D. White, M. Feldman and D. L. White. Application of zone plates to alignment in microlithography. *J. Vac. Sci. Technol.*, 19(4):1224, 1981.

[58] B. Fay and W. T. Novak. Automatic X-Ray Alignment System for Submicron VLSI Lithography. *Solid State Technology*, 28(5):175, May 1985.

[59] M. T. Mason, J. M. Lavine, C. Sparkes and S. A. Lis. Calibration Software Utilities To Ensure 0.3 Micron Registration: Autocal And Grid Standardization. In *Kodak Micro-Imaging Seminar '84*, 1984.

[60] V. Dimilia, D. A. Nelson and J. M. Warlaumont. A wide-range alignment system for x-ray lithography. *J. Vac. Sci. Technol.*, 19(4):1219, 1981.

[61] B. Fay and J. Trotel. Optical Alignment Systems for Submicron X-Ray Lithography. *J. Vac. Sci. Technol.*, 16:1954, 1979.

[62] T. M. Lyszczarz et al. Experimental evaluation of interferometric alignment techniques for multiple mask registration. *J. Vac. Sci. Technol.*, 19(4):1214, 1981.

[63] G. Bouwhuis and S. Wittekoek. Automatic Alignment System for Optical Projection Printing. *IEEE Trans. Electron Devices.*, ED-26(4):723–728, 1979.

[64] H. P. Kleinknecht. Diffraction gratings as keys for automatic alignment in proximity and projection printing. In *SPIE Developments in Semiconductor Microlithography IV*, pages 63–69, 1979.

[65] *Site by Site Alignment for DSW Wafer Stepper.* Technical Report, GCA, 1981.

[66] M. Ogawa, S. Murakami, T. Matsuura and M. Uehara. Laser step alignment for a wafer stepper. In *SPIE Optical Microlithography IV*, page 9, 1985.

[67] M. A. van den Brink et al. Performance of a wafer stepper with automatic intradie registration correction. In *SPIE Optical Microlithography VI*, 1987.

[68] H. I. Smith, D. C. Flanders and S. Austin. *Appl. Phys. Lett.*, 31:426, 1977.

[69] R. S. Hershel. Autoalignment in step-and-repeat wafer printing. In *SPIE Developments in Semiconductor Microlithography IV*, pages 54–62, 1979.

[70] *UltraStep 1000 Product Description Guide.*

[71] A. Suzuki. Double Telecentric Wafer Stepper Using Laser Scanning Method. In *SPIE Optical Microlithography IV*, page 2, 1985.

[72] A. Suzuki. Laser scanning autoalignment in projection system. In *SPIE Semiconductor Microlithography VI*, page 35, 1981.

[73] J. Wilsczynski, N. Bobroff, R. Tibbets and A. Wilson. An optical alignment microscope for x-ray lithography. *J. Vac. Sci. Technol.*, B 4(1):285–289, 1986.

[74] A. Une, H. Kinoshita and M. Iki. A Dual Grating Alignment Technique for X-Ray Lithography. *J. Vac. Sci. Technol.*, B 1(4):1276–1279, 1983.

[75] T. Kanayama, J. Itoh and N. Atoda. A New Mask-To-Wafer Alignment Technique for A-Quarter-Micron SR Lithography. In *18th (1986 International) Conference on Solid State Devices and Materials*, 1986.

[76] *Optimetrix User's Manual.*

[77] G. A. C. Jones, D. M. Holburn and H. Ahmed. A pattern recognition technique using sequences of marks for registration in electron beam lithography. *J. Vac. Sci. Technol.*, 19(4):1229, 1981.

[78] A. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[79] G. Kushibiki, H. Ohtsuka, T. Taguchi and T. Koikeda. Parameters Affecting the Ability to Align Aluminum Layers on an Optical Wafer Stepper. In *SPIE Optical Microlithography III*, page 70, 1984.

[80] J. Schaper, S. Cosentino and J. H. Peavy. Applications of Thin Film Reflectance Calculations to Linewidth Measurements for HMOS Circuit Fabrication. In *Kodak Microelectronics Seminar, Interface '83*, San Diego, 1982.

[81] S. Kuniyoshi et. al. Contrast improvement of alignment signals from resist coated patterns. *J. Vac. Sci. Technol.*, B 5(4):555–506, 1987.

[82] C. B. Burckhardt. Diffraction of a plane wave at a sinusoidally stratified dielectric grating. *J. Opt. Soc. Am.*, 56:1502–1509, 1966.

[83] F. G. Kaspar. Diffraction by thick, periodically stratified gratings with complex dielectric constant. *J. Opt. Soc. Am.*, 63:37–45, 1973.

[84] F. G. Kaspar. Computation of light transmitted by a thick grating, for application to contact printing. *J. Opt. Soc. Am.*, 64:1623–1630, 1974.

[85] D. Nyyssonen. Theory of optical edge detection and imaging of thick layers. *J. Opt. Soc. America*, 72:1425–1453, 1985.

[86] D. R. Scifres, W. Streifer and R. D. Burnham. Analysis of grating-coupled radiation in GaAs:GaAlAs lasers and waveguides. *IEEE J. Quantum Electron.*, QE-12:422–428, 1976.

[87] C. P. Kirk and D. Nyyssonen. Modelling the optical microscope images of thick layers for the purpose of linewidth measurement. In *SPIE Optical Microlithography IV*, 1985.

[88] Born and Wolf. *Principles of Optics (6th ed.)*, pages 51–70. Pergamon Press, Oxford, 1980.

[89] C. P. Kirk and D. Nyyssonen. 3-D Imaging of Line Structures in Semiconductor Microlithography. 1984. Leeds University report.

[90] Born and Wolf. *Principles of Optics (6th ed.)*, page 621. Pergamon Press, Oxford, 1980.

[91] G. Kushibiki, H. Ohtsuka, T. Taguchi and T. Koikeda. Double Structured Alignment Mark for Enhanced Automatic Alignmnent. In *SPIE Optical Microlithography IV*, page 102, 1985.

[92] D. S. Perloff. A Four-Point Electrical Measurement Technique for Characterizing Mask Superposition Errors on Semiconductor Wafers. *IEEE J. Solid-State Circuits*, SSC-13(4):436–444, 1978.

[93] A. R. Neureuther, W. G. Oldham, S. N. Nandgaonkar and M. O'Toole. A General Purpose Simulator for VLSI Lithography and Etching Processes: Part II-Application to Deposition and Etching. *IEEE Trans. Electron Devices.*, ED-27(8):1455–1459, 1980.

[94] P. S. Hauge, F. H. Dill, W. P. Hornberger and J. M. Shaw. Characterization of Positive Photo-resist. *IEEE Trans. Electron Devices.*, ED-22(7):445–452, 1975.

[95] F. H. Dill and J. M. Shaw. Thermal Effects on the Photo-resist AZ1350J. *IBM J. Res. Develop.*, 21:210, 1977.

[96] W. G. Oldham, D. J. Kim and A. R. Neureuther. Development of Positive Photo-resist. *IEEE Trans. Electron Devices.*, ED-31(12):1730–1735, 1984.

[97] J. A. Tuttle, F. H. Dill, A. R. Neureuther and E. J. Walker. Modelling Projection Printing of Positive Photo-resists. *IEEE Trans. Electron Devices.*, ED-22(7):456–464, 1975.

[98] D. A. Bernard. Optical lithography simulation - introduction to SPESA. In *Microelectronic Engineering III*, pages 379–386, 1985.

[99] C. A. Mack. PROLITH: a comprehensive optical lithography model. In *SPIE Optical Microlithography IV*, page 207, 1985.

[100] Born and Wolf. *Principles of Optics (6th ed.)*, page 375 et seq. Pergamon Press, Oxford, 1980.

[101] T. Matsuzawa et al. A Three-Dimensional Photo-resist Image Simulator: TRIPS-I. *IEEE Electron Device Lett.*, EDL-6(8):416–418, 1985.

[102] T. Ito, T. Matsuzawa and M. Tanuma. Three-Dimensional Photo-resist Image Simulation on Flat Surfaces. *IEEE Trans. Electron Devices.*, ED-32(9):1781–1783, 1985.

[103] K. Konno, K. Matsumoto and K. Ushida. Development and Application of Photolithography Simulation Program for Step-and-Repeat Projection Systems. In *Kodak Microelectronics Seminar, Interface '83*, San Diego, 1982.

[104] G. Maxwell and R. Vervoordeldonk. An analysis of the relevance to line-width control of various aerial image characteristics. In *SPIE Optical Microlithography V*, 1986.

[105] D. S. Goodman. Lithographic image simulations. In *Microelectronic Engineering 3*, pages 355–362, Rotterdam, 1985.

[106] M. C. King. The Characteristics of Optical Lithography. In *Kodak Microelectronics Seminar, Interface '80*, 1980.