Advanced MOSFETs and Novel Devices

Dr.-Ing. Josef Biba

1. Tutorial & Exercise

Yield Calculation

Exercise #1

1 Tutorial Yield
2 Shrink: Historic Example Lecture
3 Shrink: Example Fab Improvement
4 Shrink: Homework: Example Graphic Chips

Yield Calculation - Overview

1 Tutorial Yield

2 Shrink: Historic Example Lecture

3 Shrink: Example Fab Improvement

4 Shrink: Example Graphic Chips

What is yield?

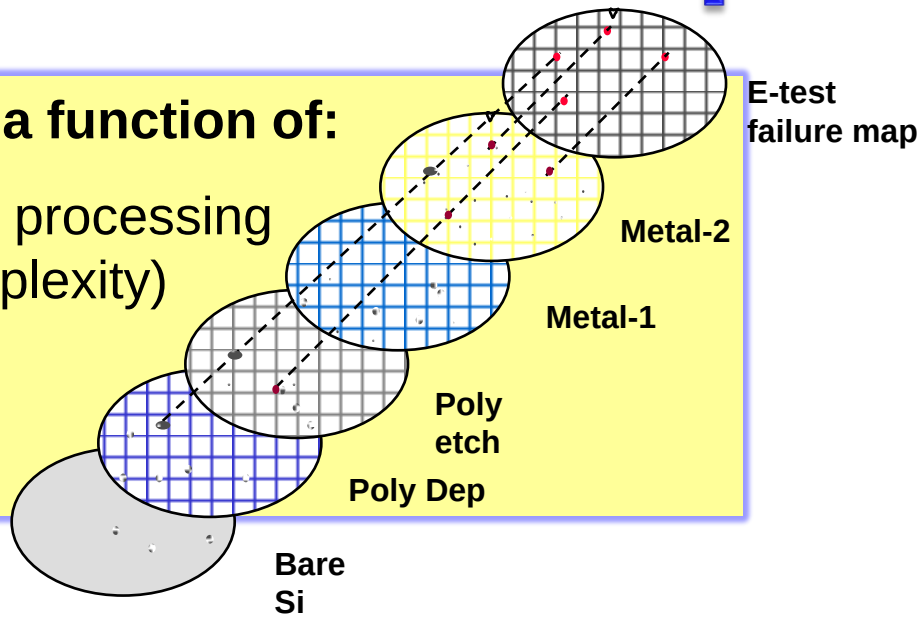
$$Y = \frac{N_{good}}{N_{total}}$$

N_{total} = number of dies per wafer
N_{good} = number of working dies per wafer

The yield of semiconductor fabrication is a function of:

- The defects per unit area introduced during processing (number of process steps and process complexity)
- Statistical defect distribution
- Die area (device size)

Packaging yield



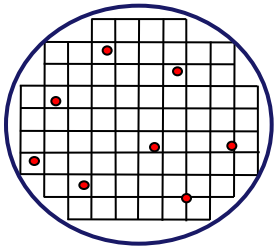
Calculate Y_R with Yield Y_i=0.99 and 50 steps.

$$Y_R = Y_{Si} * Y_{poly} * Y_{m0} * Y_{m1} * Y_{m2} * \ldots$$

Yield Calculation - Models

The Poisson Model:

$$Y = e^{-AD}$$

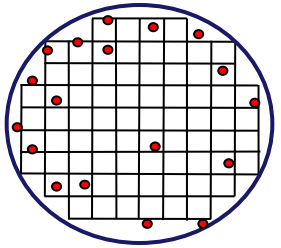


Defect density is constant

across each wafer and from wafer to wafer

The Murphy Model:

$$Y = \left[\frac{1-e^{-AD}}{AD}\right]^2$$

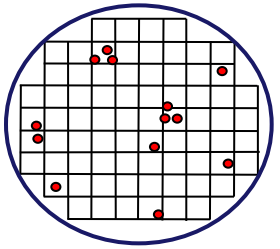


Defect density increases towards the edge of the wafer

Defects density varies across the wafer and from wafer to wafer.

The Seeds Model:

$$Y_{Seeds} = \frac{1}{1+AD}$$



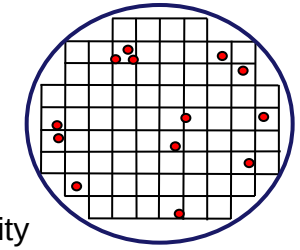
Defects tend to cluster

Defects density varies across the wafer and from wafer to wafer.

The Moore Model:

$$Y = e^{-\sqrt{AD}}$$

A = chip area
D = defect density



Empirical model,
Compromises for reality with both:

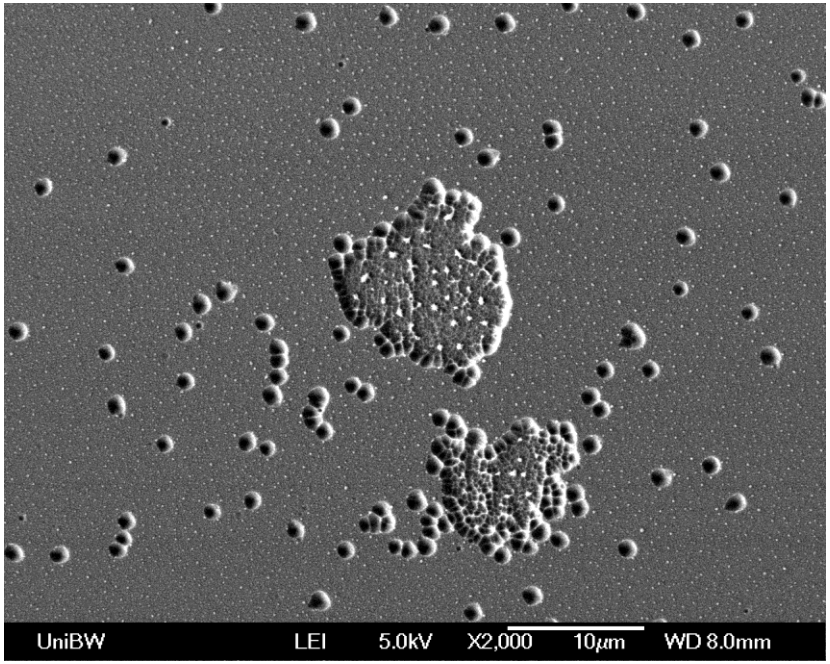
Clustering of defects (Seeds) and higher density at wafer edge (Murphy)

Yield Calculation – Distribution of Defects

Yield distribution of a processed Wafer

		I9 21%	I1 75%	SIMS	CV	H9 48%			
	B1 12%	C1 86%	D1 99%	E1	F1 99%	G1 96%	H1 71%		
A2 27%	B2 96%	C2 99%	D2 100%	E2	F2 36%	G2 100%	H2 100%	I2 93%	
Elli	A3 100%	B3 100%	C3 100%	D3 99%	E3	F3 36%	G3 100%	H3 100%	I3 98%	Elli
CV	A4	B4	C4	D4	E4	F4	G4	H4	I4	CV
Just	A5	B5	C5	D5	SIMS	F5	G5	H5	I5	Just
Elli	A6 98%	B6 99%	C6 98%	D6 98%	E7	F6 85%	G6 100%	H6 99%	I5 98%	Elli
SIMS	A7	B7	C7	D7	E7	F7	G7	H7	H7	SIMS
	A8 85%	B8 97%	C8 99%	D8 100%	E8	F8 100%	G8 100%	H8 99%	H8 82%	
		B9 87%	C9 98%	D9 98%	E9	F9 100%	G9 96%	H9 76%		
			CV	SIMS	CV					

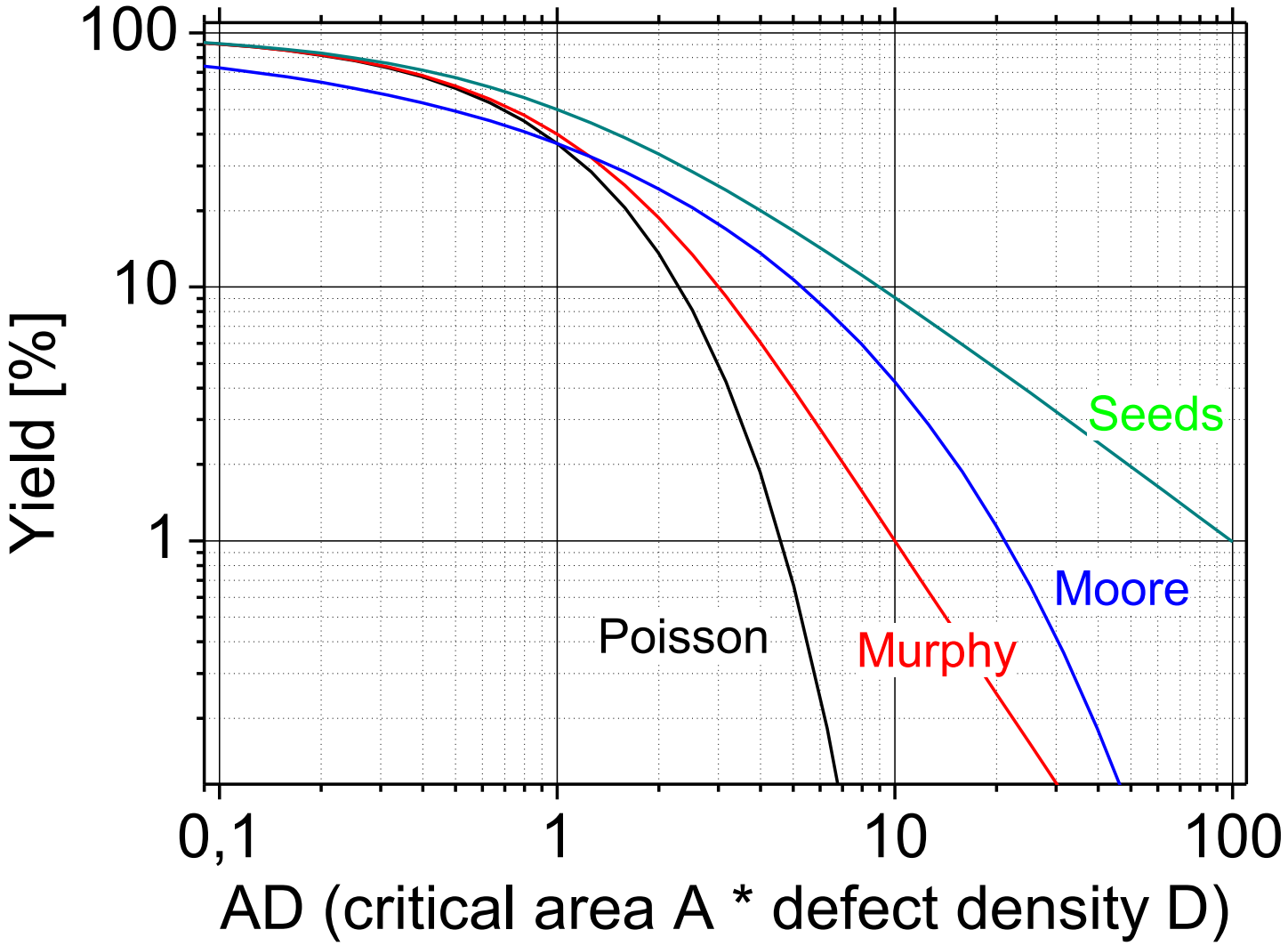
Clustering of epitaxial defects



Yield loss increases towards the wafer edge

Percentage = number of working devices

Yield Calculation - Models



Yield Calculation - Overview

1 Tutorial Yield

2 Shrink: Historic Example Lecture

3 Shrink: Example Fab Improvement

4 Shrink: Example Graphic Chips

Yield Calculation – Historical Example

Example: Fabrication costs per wafer

Technology	(~1990)	Application	3"	100mm	150mm
NMOS (array) + CMOS		1 Mb DRAM	75 $	90 $	130 $

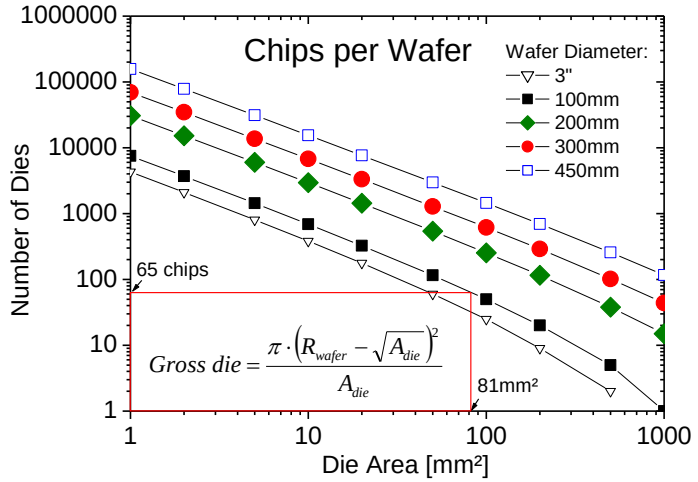
with:
wafer size: 100 mm
die size: 9x9 mm²
defect density: 3 / cm²

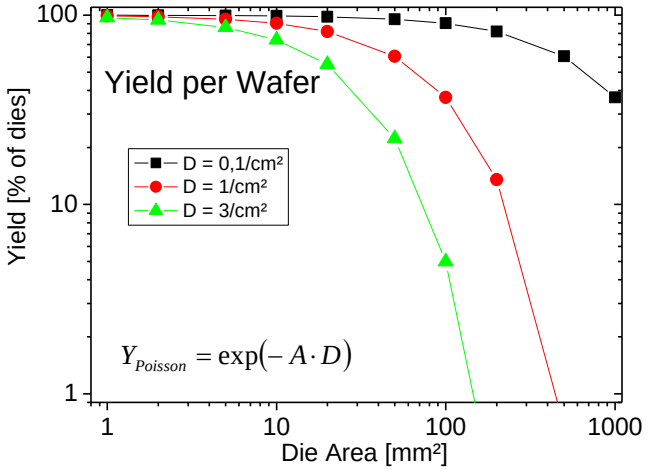
Production Costs per working chip

xx $

Redesign

Shrink by 15 %





Production Costs per working chip

xy $

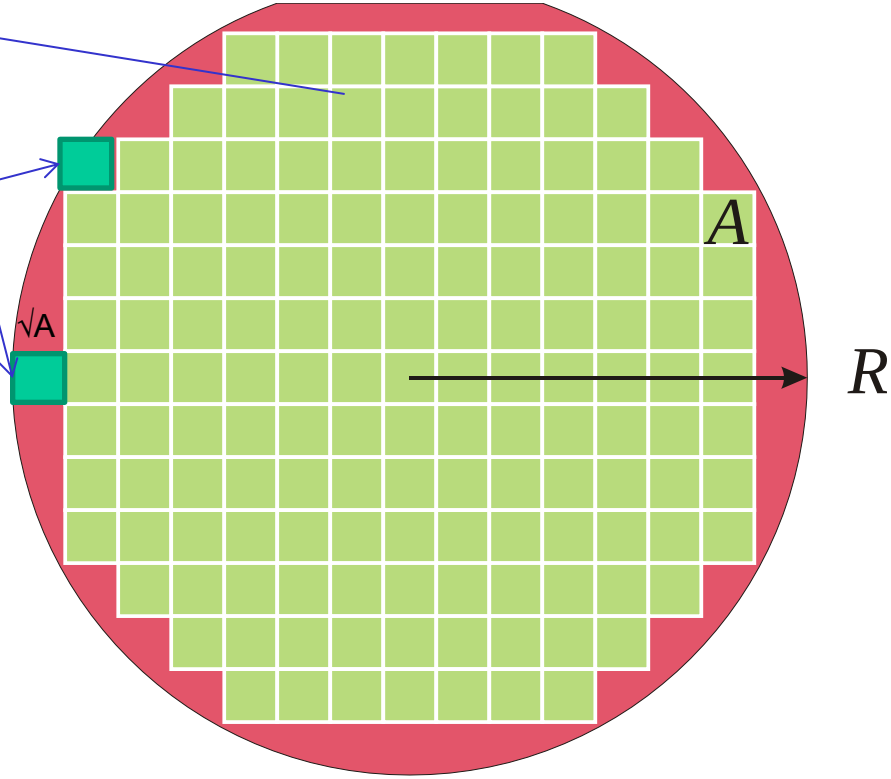
Small shrinking in device dimensions reduces fabrication costs / device drastically

Yield Calculation – Historical Example

1) First we calculate the the total number of dies on the wafer (gross die):

$$N_{\text{total}} = \frac{\pi \cdot \left(R - \sqrt{A}\right)^2}{A}$$

we have to subtract 1 chip size (√A for quadratic chip)
from the radius R to ensure not to calculate with incomplete chips



N_{total}	= number of dies per wafer
N_{good}	= number of working dies per wafer
A	= chip area
D	= defect density
R	= wafer radius

2) We use the Moore Model for the calculation of the Yield:

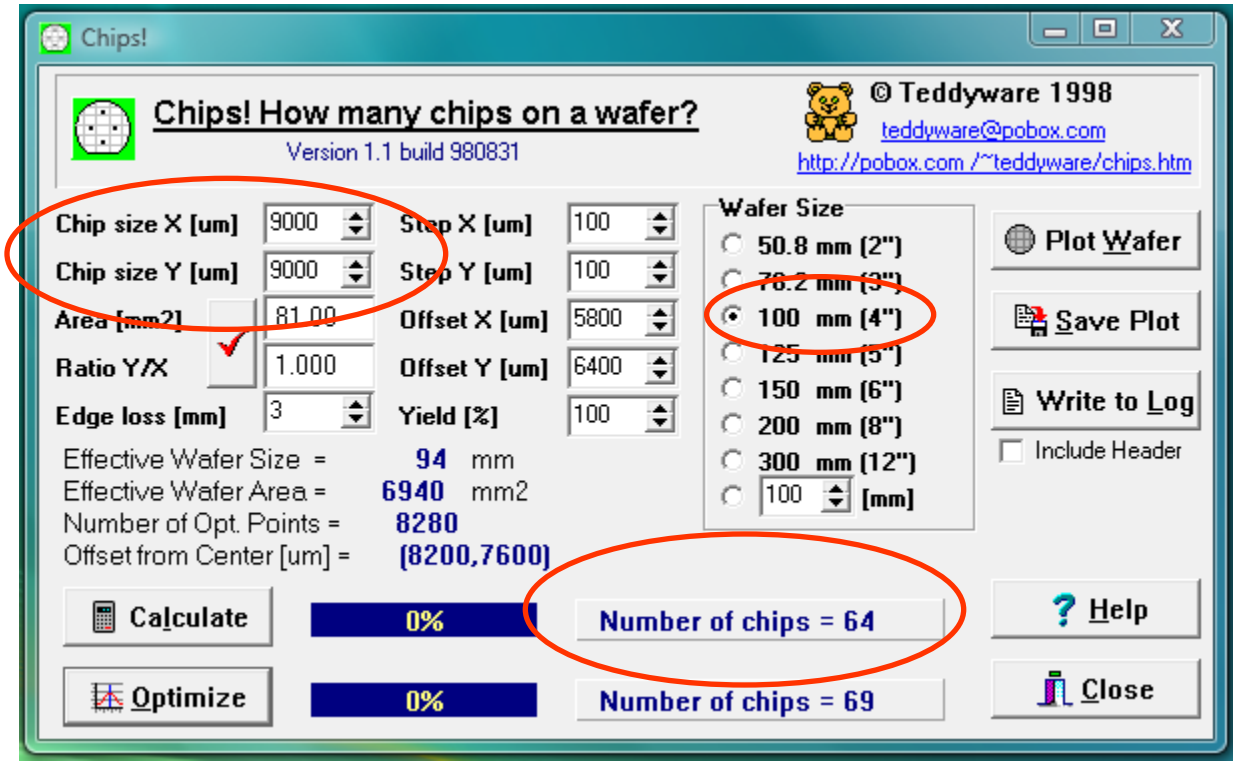
$$Y = e^{-\sqrt{AD}}$$

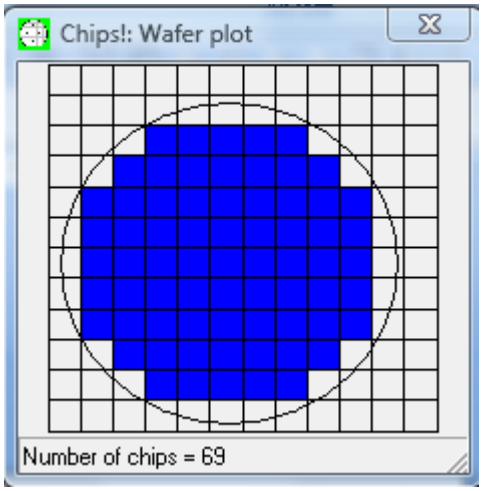
3) Finally we calculate the number of working dies:

$$Y = \frac{N_{\text{good}}}{N_{\text{total}}}$$

Yield Calculation – Historical Example

For many calculations we may use some free-ware





$$N_{\text{total}} = \frac{\pi \cdot \left(R - \sqrt{A}\right)^2}{A}$$

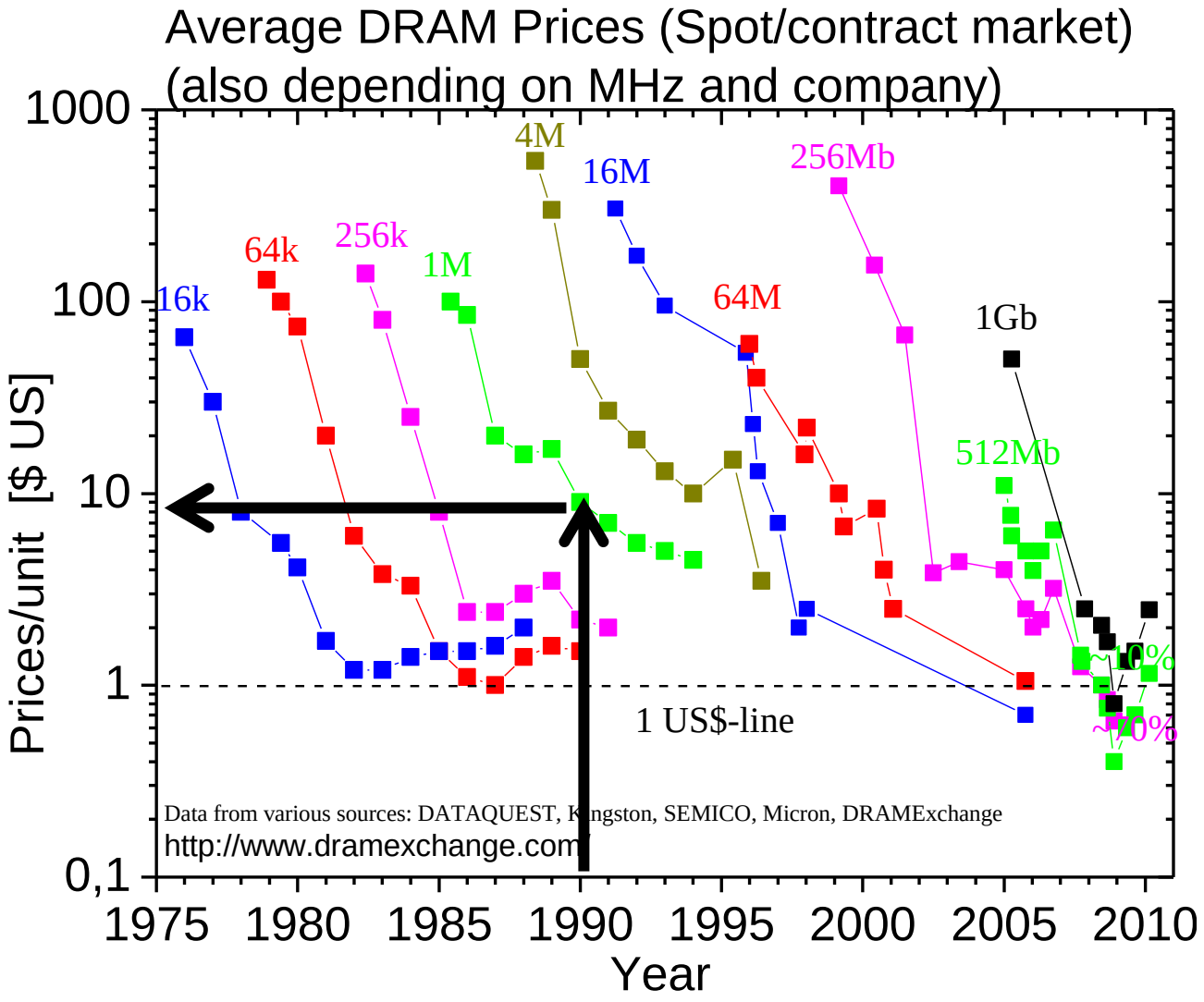
Yield Calculation – Historical Example

Let´s have a look to the selling prices before our calculation

In 1990 the 1 Mb-DRAM was around 9 USD

Now we do the cost calculation

Next page



Yield Calculation – Historical Example

Redesign: Shrinking the structure size 15%

Old technology

Defect Density: $D = 3\ \mathrm{cm}^{-2}$
Die Area: $A = 9 \times 9\ \mathrm{mm}^2$
Wafer Size: $2R = 100\ \mathrm{mm}$

Gross Die:

$$N = \frac{\pi \cdot \left(R - \sqrt{A}\right)^2}{A} = \frac{\pi \cdot (50\,\mathrm{mm} - 9\,\mathrm{mm})^2}{81\,\mathrm{mm}^2} = 65.2 \approx 65$$

Look: the software calculator calculated 64 dies

Yield:

$$Y = \exp\left(-\sqrt{A \cdot D}\right) = \exp\left(-\sqrt{81\,\mathrm{mm}^2 \cdot 3\,\mathrm{cm}^{-2}}\right) =$$
$$= \exp\left(-\sqrt{81\,\mathrm{mm}^2 \cdot 0.03\,\mathrm{mm}^{-2}}\right) = 0.21 = 21\%$$

Cost per Chip:

$$c = \frac{90\,\$}{65 \cdot 0.21} = 6.59\,\$$$

Look: the market price was ~ 9 USD

New technology

Defect Density: $D = 3\ \mathrm{cm}^{-2}$
Die Area: $A = 7.65 \times 7.65\ \mathrm{mm}^2$
Wafer Size: $2R = 100\ \mathrm{mm}$

Gross Die:

$$N = \frac{\pi \cdot \left(R - \sqrt{A}\right)^2}{A} = \frac{\pi \cdot (50\,\mathrm{mm} - 7.65\,\mathrm{mm})^2}{58.52\,\mathrm{mm}^2} = 96.28 \approx 96$$

Yield:

$$Y = \exp\left(-\sqrt{A \cdot D}\right) = \exp\left(-\sqrt{58.52\,\mathrm{mm}^2 \cdot 3\,\mathrm{cm}^{-2}}\right) =$$
$$= \exp\left(-\sqrt{58.52\,\mathrm{mm}^2 \cdot 0.03\,\mathrm{mm}^{-2}}\right) = 0.26 = 26\%$$

Cost per Chip:

$$c = \frac{90\,\$}{96 \cdot 0.26} = 3.60\,\$$$

Cost reduction by factor 1.83

Yield Calculation – Historical Example

Example: Fabrication costs per wafer

Technology	(~1990)	Application	3"	100mm	150mm
NMOS (array) + CMOS		1 Mb DRAM	75 USD	90 USD	130 USD

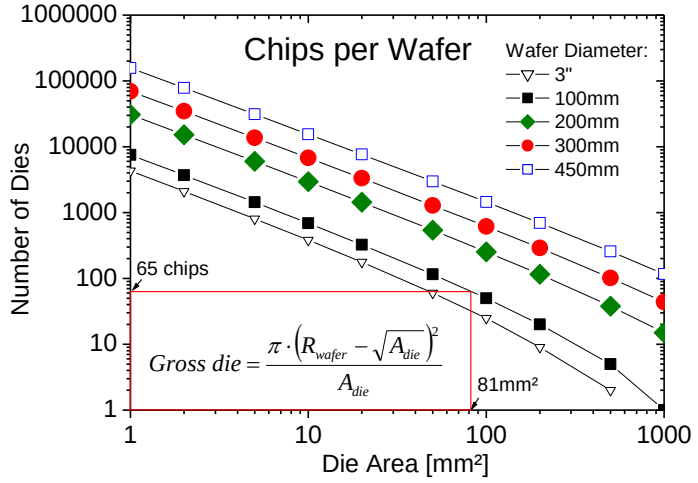
with:
wafer size: 100 mm
die size: 9x9 mm²
defect density: 3 / cm²

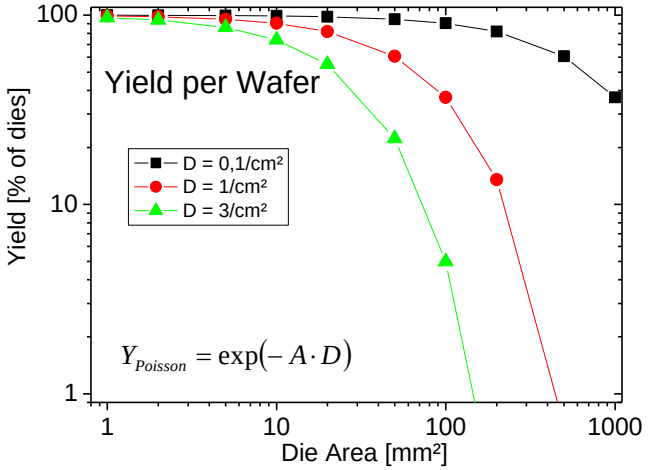
Production Costs per working chip

6.59 $

Redesign

Shrink by 15 %





Production Costs per working chip

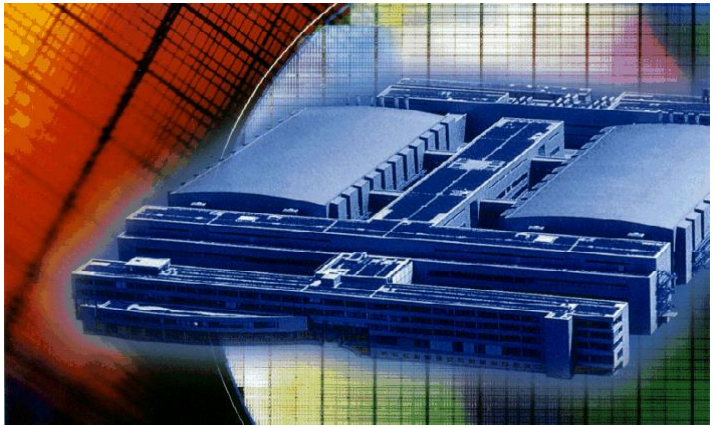
3.60 $

Small shrinking in device dimensions reduces fabrication costs / device drastically

Yield Calculation - Overview

1 Tutorial Yield
2 Shrink: Historic Example Lecture
3 Shrink: Example Fab Improvement
4 Shrink: Homework: Example Graphic Chips

Yield Calculation – Fab Improvement

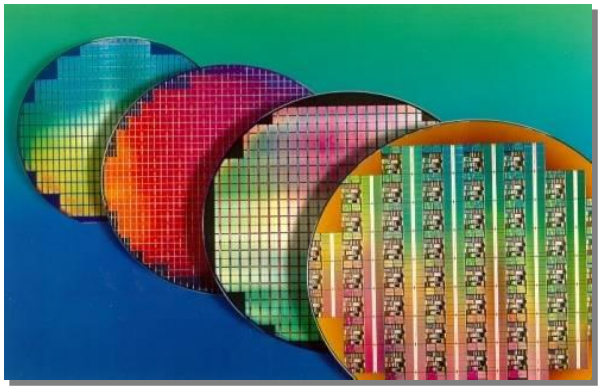


An IC manufacturer is faced with a decision concerning the manufacturing of a 10 mm × 10 mm die. This die has been produced on 200 mm wafers in Fab 1, which has a defect density of 2 defects/cm^2. The chips are offered for about 7.98 USD on the market.
Some competitors offer similar chips for about 2.98 USD recently.

Can we beat this price ?
Our IC manufacturer is calculating several scenarios to reduce his fabrication costs:

1) Continue manufacturing in Fab 1 with no changes until the price drops below the cost, then phase the product out.

2) Continue manufacturing the circuit in Fab 1 on 200 mm wafers,
but perform a 20% shrink (all sizes).
(we will neglect the development costs for the shrink)



3) After shrinking, the company decides to improve the defect density to 0.8 defects/cm^2.
(we will neglect the development costs for the reduced defect density)

4) The company decides to buy new equipment and transfer the production from Fab 1 to Fab 2. In Fab 2 the production is on 300 mm wafers.

Yield Calculation – Fab Improvement

	Fab 1	Fab 1	Fab 1	Fab 2
Line width	100 nm	80 nm	80 nm	80 nm
Wafer diameter	200 mm	200 mm	200 mm	300 mm
Process	Epi CMOS	Epi CMOS	Epi CMOS	Epi CMOS
Cost per finished wafer	$ 480.00	$ 480.00	$ 480.00	$ 870.00
Defect density	2 cm^{-2}	2 cm^{-2}	0.8 cm^{-2}	0.8 cm^{-2}
Chip size	10 x 10 mm^2	8 x 8 mm^2	8 x 8 mm^2	8 x 8 mm^2
Gross dies per wafer				
Yield				
Good dies per wafer				
Cost per die				
		Shrink 20%	Reduction D by 60%	Larger Wafer 33%
Complete Cost Reduction				
Additional Cost Reduction		-		

Yield Calculation - Results

1. The biggest effect on its own is resulting from line width shrinking.
2. The improved defect density produces more net good dies per wafer.
3. The larger wafer produces more gross and net good dies per wafer.
4. The biggest complete cost reduction per good die will be achieved by realizing Fab 2.

The IC manufacturer will decide to perform a 20% shrink (each direction) and move the circuit in Fab 2 with a improved defect density.

Yield Calculation - Overview

1 Tutorial Yield
2 Shrink: Historic Example Lecture
3 Shrink: Example Fab Improvement
4 Shrink: Homework: Example Graphic Chips

Clash in the High-End GPU Market 2008/2009

NVIDIA and AMD are batteling for the crown of the highest performing GPU in the market. Currently the Radeon 4870 is slightly ahead of NVIDIAs GeForce GTX 280. But it is also a fight between two different concepts: Monolithic GPU vs. Dual GPU
Lets look at this from a manufacturing perspective and the costs:

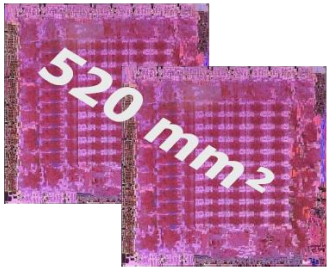
NVIDIA GeForce GTX 280 (launched 16/6/08)
Codename GT200, 65 nm @ TSMC
Size: 576 mm² (24 mm x 24 mm)

AMD ATI Radeon HD 4870 X2 (launched 12/8/08)
Codename RV700, 55 nm @ TSMC
Consisting of two RV770 Dice (Dual GPU)
Size (RV770): 256 mm² (16 mm x 16 mm)

NVIDIA GeForce GTX ??? (launch 4Q08)
Codename GT206, 55 nm @ TSMC
Die-Shrink of GT200
Size: 412 mm² (20,3 mm x 20,3 mm)

Technology Parameters TSMC (est.)
D= 0.2/cm²
Fabrication cost per 300 mm Wafer: 3000 US-$

ATI Radeon™ HD 4870 X2



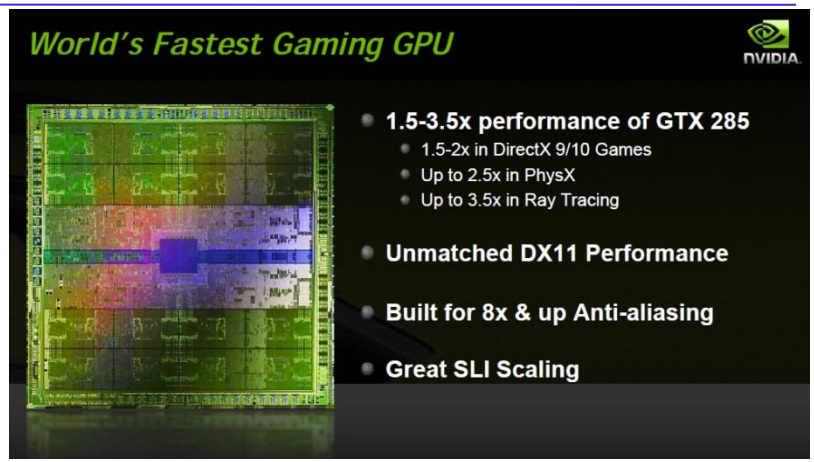
nVidia GeForce GTX 280

a) Calculate the cost of every working GT200.
b) Calculate the cost of every working RV770.
c) Calculate the cost difference between a GT200 and a RV700 (2x RV770)
d) Calculate the cost of every GT206
e) Interpret the results from the cost/performance perspective

Clash in the High-End GPU Market 2008/2009

Die GTX 480 ist die erste Grafikkarte mit der "Fermi" getauften Architektur, die Nvidia bereits Ende September 2009 vorgestellt hat - gerade mal eine Woche nachdem AMD mit der Radeon HD 5870 die schnellste GPU auf den Markt brachte. Die Fermi-GPU heißt bei Verwendung als Spielegrafikkarte GF100, sie ist bei GTX 470 und 480 identisch.
Nicht gleich ist die Anzahl der Rechenwerke. Ein Fermi-Chip besteht zwar aus 512 der "CUDA Cores", wie sie Nvidia nennt. Auf der GTX 480 sind jedoch nur 480 davon aktiviert, bei der GTX 470 sind es 448. Diese Differenzierung von Produkten durch das Abschalten von Einheiten pflegt Nvidia zwar schon länger, im aktuellen Fall dürften aber die beim Chiphersteller TSMC anhaltenden Fertigungsprobleme dafür verantwortlich sein.

Als Beleg ist ein kleiner Exkurs in die Welt der Halbleiterfertigung nötig. Die Defekte auf einem Wafer sind relativ gleichmäßig verteilt. Je größer ein einzelner Chip ist, um so mehr der Bausteine sind insgesamt von den Defekten betroffen. Bei kleineren Chips ist die Ausbeute nach Stückzahl höher.
Passen auf einen Wafer beispielsweise 100 Chips, und die Defektrate beträgt 20 Prozent - was schon als sehr hoch gilt -, so lassen sich aus einem Wafer 80 funktionierende Bausteine machen. Sind die Chips kleiner, so dass 500 auf den Wafer passen, ergeben sich schon 400 brauchbare Bausteine. Die Herstellung eines Wafers in gleicher Strukturbreite und mit gleichem Fertigungsprozess kostet stets gleich viel, und die Kapazität einer Fertigungstraße in Waferstarts pro Monat ist auch annähernd konstant, sobald ein bestimmter Prozess halbwegs rund läuft. Umstellungen daran verzögern die Serienfertigung.
TSMC kann für Nvidia also stets gleich viele Fermi-GPUs herstellen. Um dabei eine hinreichend hohe Stückzahl zu erreichen, können Chips verwendet werden, bei denen die Defekte auf leicht abschaltbaren Bereichen liegen. Da die 512 Rechenwerke eines Fermi den größten Teil der Die-Fläche einnehmen, bieten sie sich geradezu dafür an. Wie sich die Verbesserung eines Fertigungsprozesses auf die Produkte auswirken kann, hatte Nvidia selbst mit seiner letzten GPU-Generation vorgemacht: Die GTX 260 kam erst mit 192 Rechenwerken auf den Markt und erschien kurz darauf als "216 Core Edition". Die darauf folgenden GTX 285 und 275 konnte TSMC schon mit 55 Nanometern statt zuvor 65 Nanometern fertigen, ein baldiger Wechsel der 40 Nanometer, mit der die Fermis gebaut werden, ist nicht in Sicht. Womöglich behält sich Nvidia die wenigen Fermi-Chips, die 512 funktionierende Rechenwerke enthalten, auch für die Tesla-Produkte vor. Diese sind nicht für Grafik, sondern andere Anwendungen gedacht



479 €

Selling price: 661$
Let us estimate the fabrication costs

PC-Hardware / 27.03.2010 / 00:00
http://www.golem.de/1003/74109-2.html

Cost Comparison

vendor	product	technology	chip area	cost/300 mm wafer	defects	# chips	Y	cost/chip
NVIDIA	GT200	65nm	576 mm²	~ 3000 USD	~ 0.2 / cm²			
NVIDIA	GT206	55nm	412 mm²	~ 3000 USD	~ 0.2 / cm²			
AMD	RV770	55nm	2x 256 mm²	~ 3000 USD	~ 0.2 / cm²			

Basically: Same technology (55nm), same manufacturer (TSMC) -> same price

There maybe a design advantage to use single or dual chips:

- monolitic chip = 412 mm²
- two chips = 2 x 256 mm² = 512 mm²

-> larger area, so we think this way must be more expensive
but the additional area is needed for non critical peripheral devices

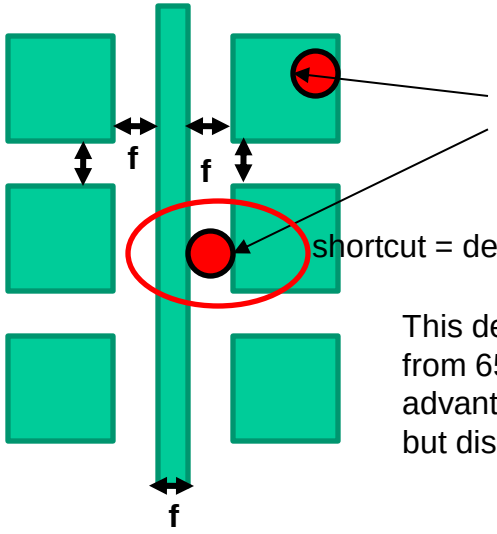
Our model of taking total chip area for yield is too simple

In reality only small fractions of the device footprint (e.g. gate area, interconnect space) are critical for yield

sometimes a more relaxed design consumes more area, but decreases critical area more drastic -> higher yield

Sometimes Shrink Models are too Simple

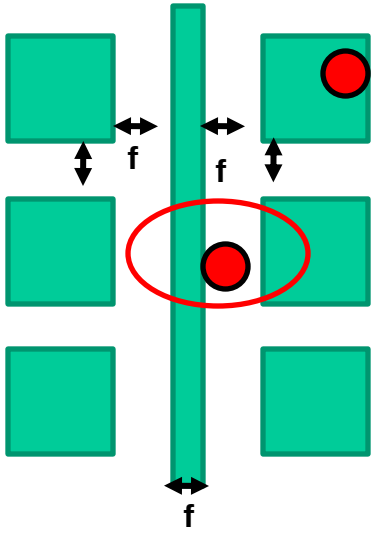
technology: 55nm -> minimum feature size f = 55nm



critical particle
average size ~60nm

shortcut = defect

This design changing all features
from 65 nm to 55 nm
advantage: little area benefit -> increase of # of chips + increase of yield due to area shrink
but disadvantage: critical particles may induce more defects -> lowering yield



keeping devices in 55 nm technology,
but relaxing some features (e.g. interconnect spacing) to 65 nm
- disadvantage: slighter larger area -> decrease of # of chips + decrease area yield
- advantage: reduced defects due to critical particles -> increasing yield

Technology (shrink) is the main driver for cost reduction,
but clever design also contributes to cost reduction

End of Exercise #1