# ECE 510, Lecture 13 Defect Models of Yield and Reliability

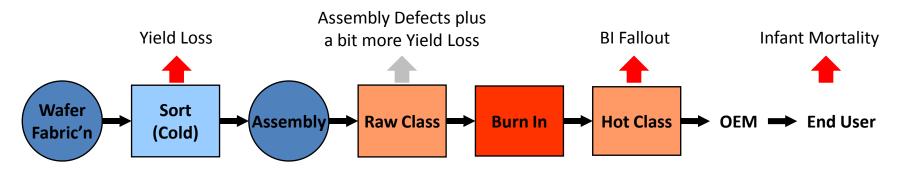
Glenn Shirley
Scott Johnson

#### Outline

- Introduction
- Models of Yield
- Models of Defect Reliability
- Analysis and Synthesis of Lifetest and Burn In

### Defect Yield and Reliability

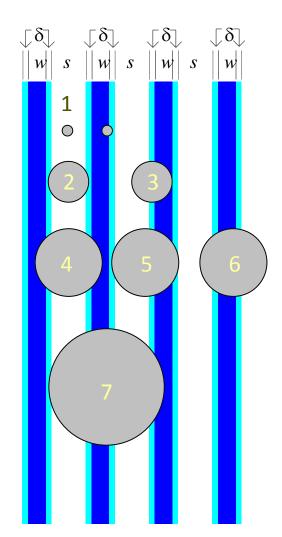
- Defects are inescapable.
  - The same kinds of defects that degrade yield perceived by the manufacturer, degrade "infant mortality" perceived by end users.
- Yield is measured at Sort initial wafer-level testing.
- Infant Mortality is measured by life-test, and controlled by burn in.
  - Life test is an extended burn in designed to acquire detailed reliability data.
- Burn in is a stress preceding final test which activates latent reliability defects (LRDs) so that they may be screened out at final test (Class).
- In these lectures we'll first cover models of Yield, and then cover Infant Mortality.
- Defect models of reliability describe only the left part of the bathtub curve; they don't describe wearout.



### Defect Model of Yield and Reliability

- Aspects of defects which affect yield and reliability are
  - Defect density. Number of defects per unit area on a wafer.
  - Spatial variation of defect density
    - Factory-to-factory
    - Lot-to-lot
    - Wafer-to-wafer
    - Across a wafer.
  - Size distribution of defects.
  - Sensitivity of circuits to defects.
- Models are used to
  - Plan for new products by predict yield and reliability figures of merit (FOMs) for hypothetical products and processes.
  - Compute the levels of fault tolerance required to meet yield goals.
  - Calculate burn in times needed to reach required levels of reliability.
  - Design life-test experiments that will provide sufficient data to build reliability models.

#### Killer vs Latent Reliability Defects

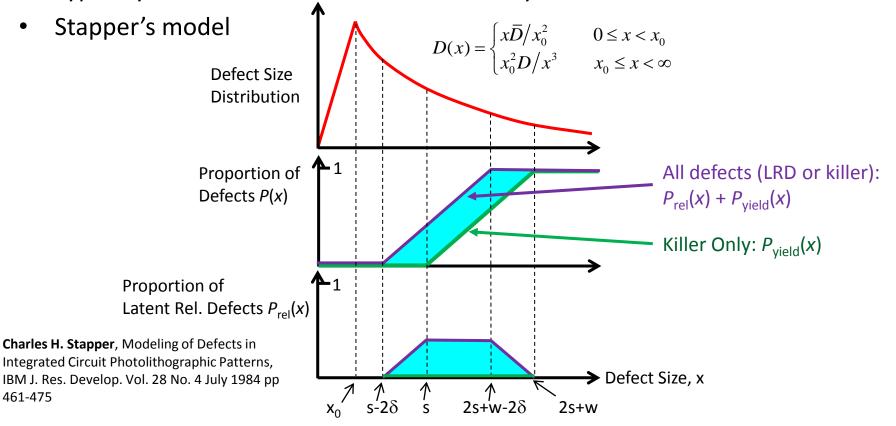


When defect is within  $\delta$  of line, failure is not immediate but will occur within the specified life of the device.

- Circuit design determines
  - Pattern pitch and space.
  - Different functional blocks have different characteristic pitch/spaces.
- Fab process determines
  - Spatial density of defects, D (defects/cm²)
  - Variation of spatial defect density.
  - Size distribution of defects.
- Ckt design plus size dist'n segregates defects into "killer" and latent reliability defects (LRD).
  - OK, never a yield or reliability defect (1).
  - Sometimes a latent reliability defect (2), sometimes OK (3).
  - Sometimes a killer defect (4), sometimes a latent reliability defect (5), sometimes OK(6).
  - Always a killer defect (7).

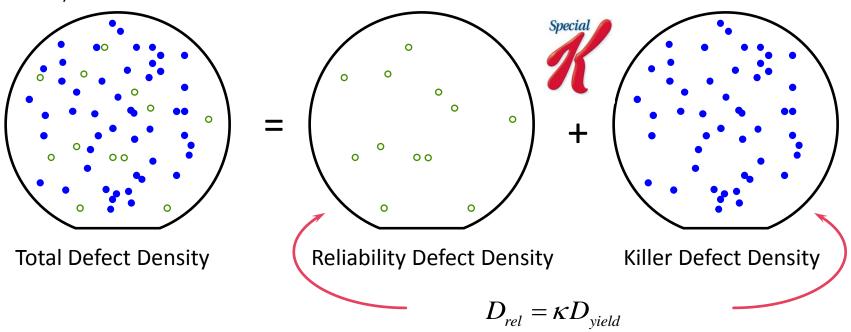
### Killer vs Latent Reliability Defects

- Defects much smaller or larger than circuit geometry are not latent reliability defects (LRD).
- Some defects with size commensurate with circuit geometry are latent reliability defects.
- Typically ~ 1% of defects are latent reliability defects.

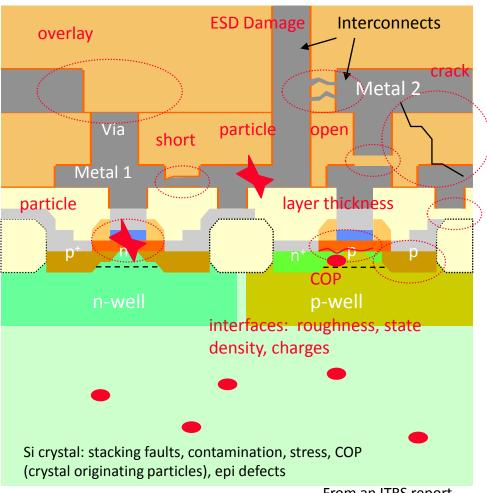


# Killer vs Latent Reliability Defects (LRD)

- Defects may be classified as "killer" defects which affect yield or LRD defects which affect reliability.
- Defects of either kind may be clustered. Described by defect density and defect density variance.
- Killer and LRD defects are from the same source, so Yield and Reliability defect densities are proportional:  $\kappa = D_{\rm rel}/D_{\rm vield} \approx {\rm constant}$  (typically ~ 1%).
- $D_{\text{yield}}$  is MUCH easier to measure and monitor in manufacturing than  $D_{\text{rel}}$ .

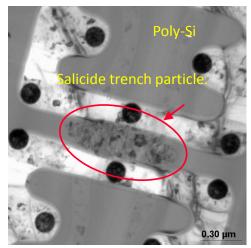


# **Ugly Reality**

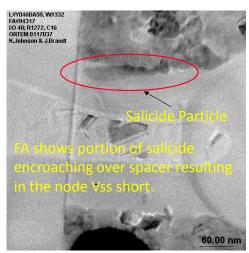


From an ITRS report.

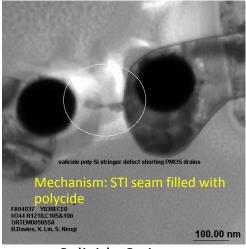
# Activated LRDs, Mainly Shorts



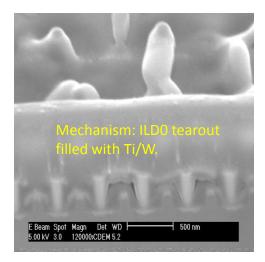
**STI Particle** 



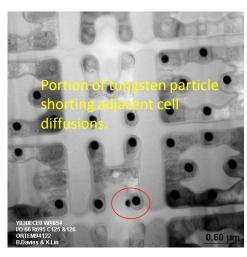
Salicide Encroachment



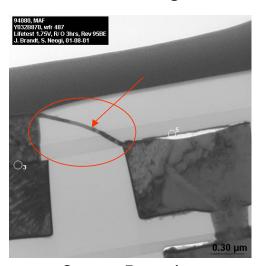
Salicide Stringer



Residual Ti

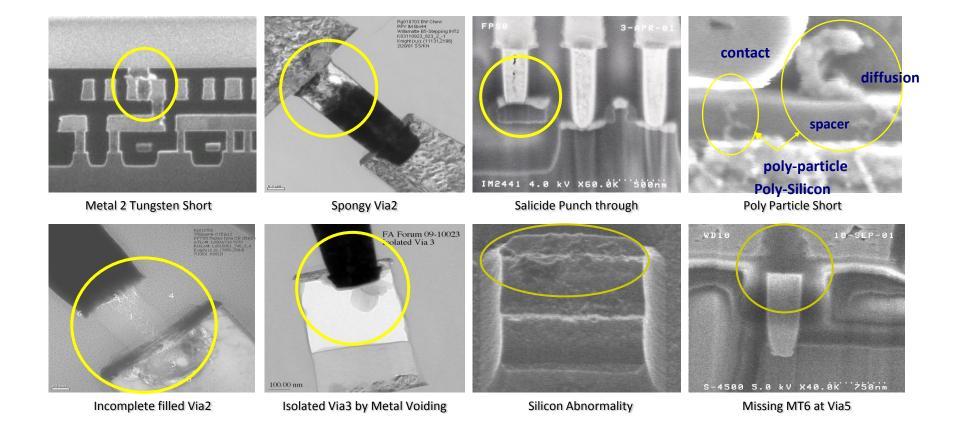


**Tungsten Particle** 



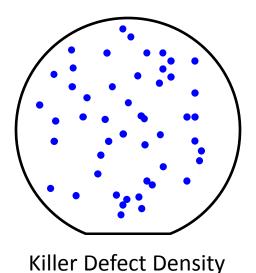
**Copper Extrusion** 

# Activated LRDs, Mainly Opens



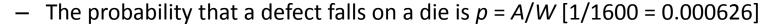
#### Outline

- Introduction
- Models of Yield
- Models of Defect Reliability
- Analysis and Synthesis of Lifetest and Burn In



#### **Defect Models of Yield**

- Assumptions for random Yield Model
  - N [1000] defects distributed spatially at random across wafers.
  - The silicon area is W [1600 cm<sup>2</sup>].
  - The die area is A [1 cm²]
  - The defect density is  $D = N/W [1000/1600 = 0.625 / cm^2]$



- The average defects per die is  $\lambda = Np = NA/W = AD$  [0.625 defects per die]

The probability that a die has exactly n defects is given by the binomial

theorem

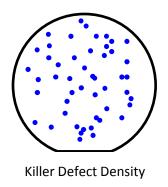
$$\frac{N!}{(N-n)!n!}p^n(1-p)^{N-n}$$

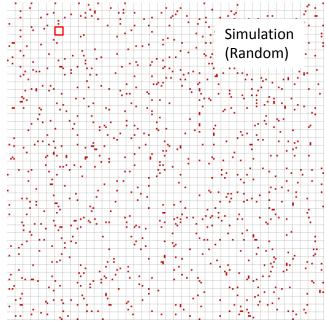
• Example: N = 1000 defects over 1600 dies

$$1600 \times \left\{ \left( 1 - \frac{1}{1600} \right)^{1000} \right\} = 856$$
 dies with 0 defects  

$$1600 \times \left\{ 1000 \left( 1 - \frac{1}{1600} \right)^{999} \left( \frac{1}{1600} \right) \right\} = 535$$
 dies with 1 defect  

$$1600 \times \left\{ \frac{1000 \times 999}{2} \left( 1 - \frac{1}{1600} \right)^{998} \left( \frac{1}{1600} \right)^{2} \right\} = 167$$
 dies with 2 defects





#### Poisson Limit of the Binomial Dist'n

• When  $p \to 0$  and  $N \to \infty$  in such a way that Np remains finite, it is much more convenient to use the Poisson limit.

$$\frac{N!}{(N-n)!n!} p^{n} (1-p)^{N-n} \xrightarrow{N \to \infty, p \to 0} \frac{\lambda^{n}}{n!} \exp(-\lambda), \quad \lambda = Np$$

$$\lambda = Np = 1000 \times \frac{1}{1600} = 0.625$$

$$1600 \times \left\{ \frac{\lambda^{0}}{0!} e^{-\lambda} = e^{-0.625} \right\} = 856 \qquad \text{dies with 0 defects}$$

$$1600 \times \left\{ \frac{\lambda^{1}}{1!} e^{-\lambda} = 0.625 \times e^{-0.625} \right\} = 535 \qquad \text{dies with 1 defect}$$

$$1600 \times \left\{ \frac{\lambda^{2}}{2!} e^{-\lambda} = \frac{0.625^{2} \times e^{-0.625}}{2} \right\} = 167 \qquad \text{dies with 2 defects}$$

- The Poisson limit is nearly always sufficient for yield models.
  - Works well for N ≥ 20 and p ≤ 0.05, or if N ≥ 100 and p ≤ 0.10
- Details

$$B(n|N,p) = \frac{N!}{(N-n)!n!} p(1-p)^{N-n}$$

$$= \frac{N(N-1)(N-2)..(N+1-n)}{n!} \left(\frac{\lambda}{N}\right)^n \left(1-\frac{\lambda}{N}\right)^{N-n} = N^{n} \frac{1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)..\left(1-\frac{n-1}{N}\right)}{n!} \times \frac{\lambda^n}{N^{n}} \left(1-\frac{\lambda}{N}\right)^{N} \left(1-\frac{\lambda}{N}\right)^{-n} = \frac{1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)..\left(1-\frac{n-1}{N}\right)}{\left(1-\frac{\lambda}{N}\right)^{n}} \times \frac{\lambda^n}{n!} \times \left(1-\frac{\lambda}{N}\right)^{N}$$

$$\xrightarrow{N\to\infty} 1 \times \frac{\lambda^n}{n!} \exp(-\lambda)$$
Note:  $\lim_{N\to\infty} \left(1-\frac{\lambda}{N}\right)^{N} = \exp(-\lambda)$ 

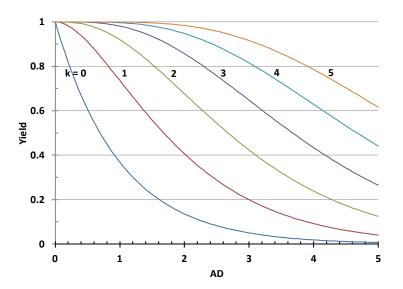
#### **Fault Tolerance**

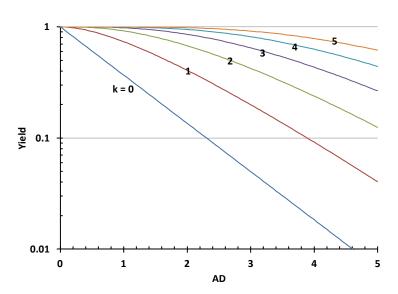
 If a die must be perfect to be "good" the yield is the probability that a die has 0 defects

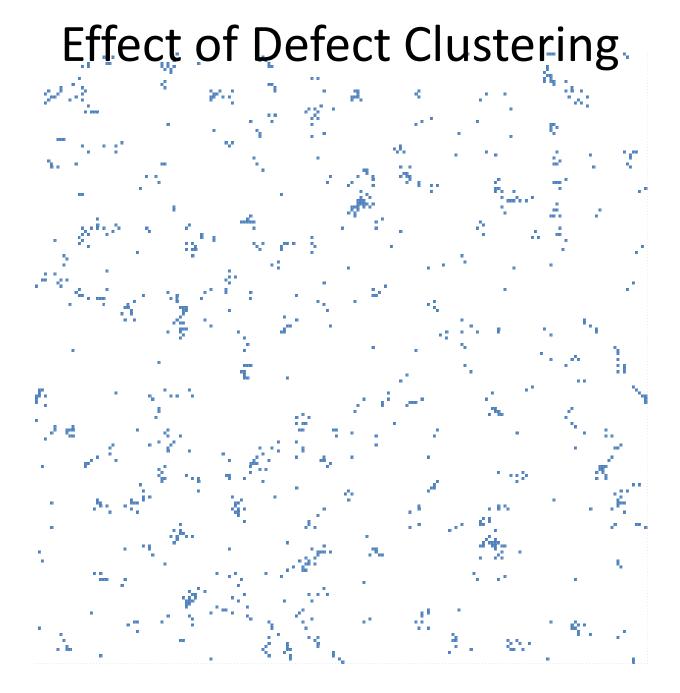
$$Y = \exp(-\lambda)$$
,  $\lambda = AD \equiv$  average number of defects per die

 If a die can be good with up to (and including) k defects, then the yield is the sum of probabilities of 0, 1, 2, .. k defects

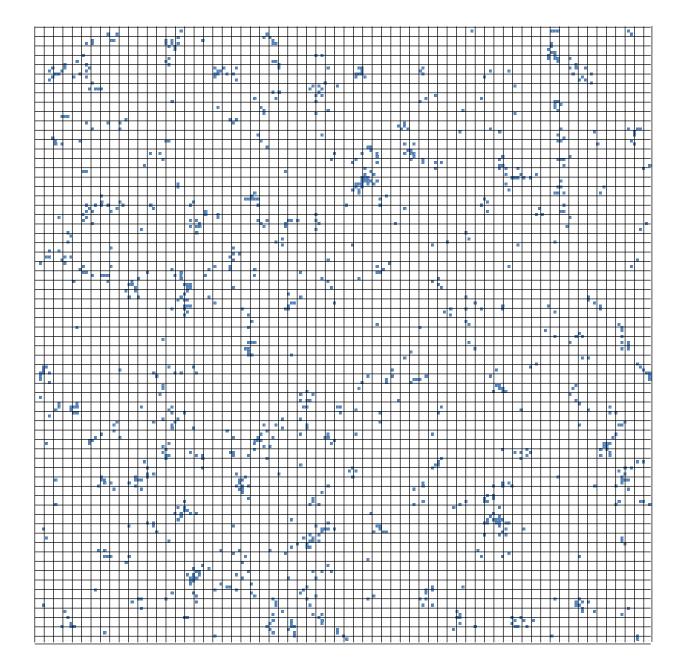
$$Y = \sum_{i=0}^{k} \frac{\lambda^{i}}{i!} \exp(-\lambda) = \text{POISSON}(k, \lambda, \text{TRUE})$$



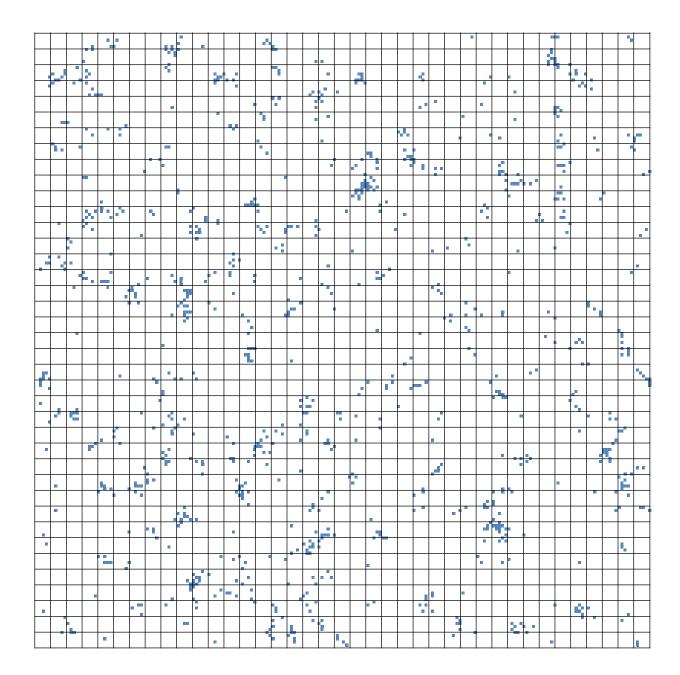




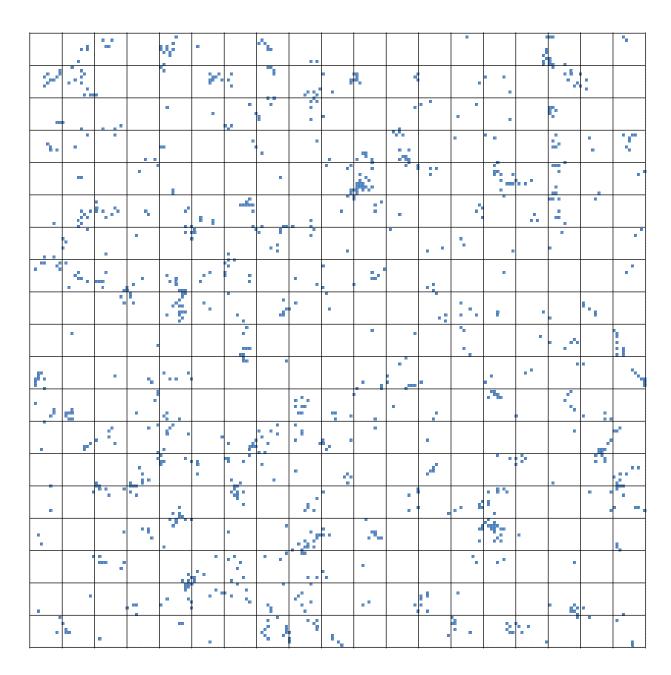
3x3



5x5

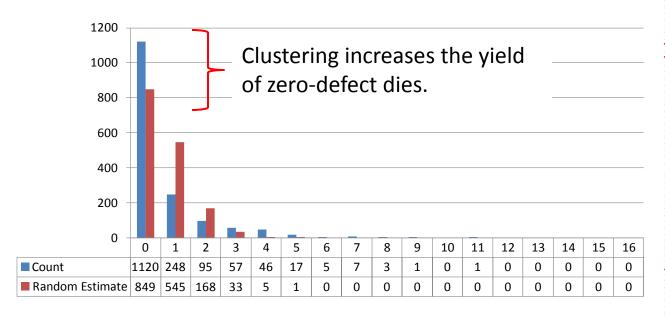


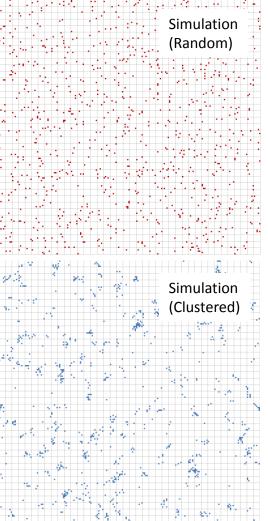
#### 10x10



# Effect of Defect Clustering

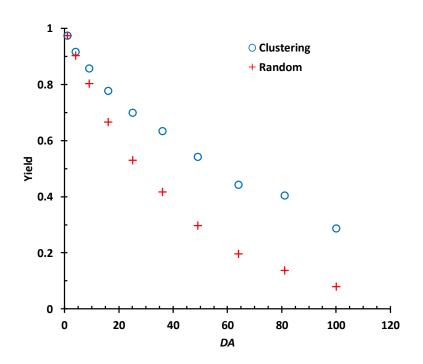
- 1000 defects were synthetically distributed on a 200x200 grid.
- 2 cases: 1) Random, 2) Clustered according to a special algorithm.
- 5x5 dies were superimposed on the grid.
- Counts of defect-free, 1-defect etc. dies were made.

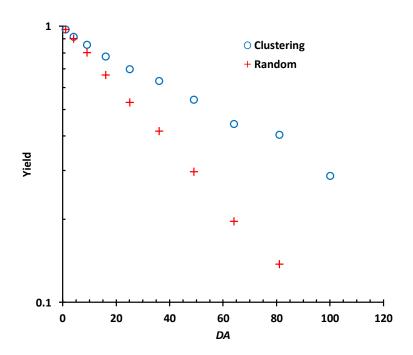




# **Effect of Clustering**

- Window Method: Overlay pattern with 1x1, 2x2, 3x3, .. non-overlapping windows and count defect-free cells.
- Clustered defect patterns have higher yield!
- Business opportunity: For any given yield, with defect clustering a die may be larger and have more functions, giving a more competitive product.



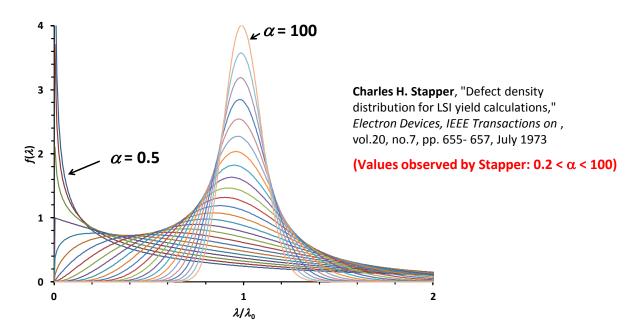


# Murphy/Stapper Yield Model

Murphy posited that defect density has a distribution so

$$Y = \exp(-AD)$$
 becomes  $Y = \int_{0}^{\infty} \exp(-AD) f(D) dD$ 

• Stapper proposed the Gamma distribution for f(D)



$$f(\lambda) = \frac{\alpha}{\Gamma(\alpha)\lambda_0} \left(\alpha \frac{\lambda}{\lambda_0}\right)^{\alpha-1} \exp\left(-\alpha \frac{\lambda}{\lambda_0}\right) = \text{GAMMADIST}(\lambda, \alpha, \lambda_0 / \alpha, \text{FALSE})$$

$$E(\lambda) = \lambda_0 \qquad Var(\lambda) = \frac{\lambda_0^2}{\alpha}$$

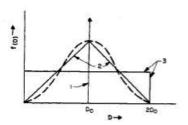


Fig. 1-Distribution functions.

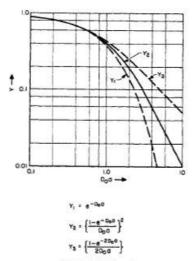


Fig. 2-Yield functions.

**B. T. Murphy**, "Cost-size optima of monolithic integrated circuits," *Proceedings of the IEEE*, vol.52, no.12, pp. 1537- 1545, Dec. 1964

#### Yield Model with Clustering

 Probability that a die has exactly n defects is the Poisson distribution compounded by the Gamma distribution

$$P(N = n \mid \lambda_0, \alpha) = \int_0^\infty \frac{\lambda^n \exp(-\lambda)}{n!} f(\lambda) d\lambda = \frac{\Gamma(n + \alpha)}{n! \Gamma(\alpha)} p^n (1 - p)^\alpha \text{ where } p = \frac{\lambda_0}{\alpha + \lambda_0}$$

$$\xrightarrow{\alpha \to \infty; n, \lambda_0 \text{ finite}} \frac{\lambda_0^n \exp(-\lambda_0)}{n!}$$

Special case: probability that a die has exactly 0 defects is the yield

$$Y = \left(1 + \frac{\lambda_0}{\alpha}\right)^{-\alpha} \xrightarrow[\alpha \to \infty; n, \lambda_0 \text{ finite}]{} \exp(-\lambda_0)$$

The cumulative negative binomial is yield if n defects are tolerated

$$P(N \le n \mid \lambda_0, \alpha) = \sum_{k=0}^{n} \frac{\Gamma(n+\alpha)}{n! \Gamma(\alpha)} p^n (1-p)^{\alpha} = 1 - B\left(\frac{\lambda_0}{\lambda_0 + \alpha}, n+1, \alpha\right)$$

- Special case: n = 0 (no fault tolerance)

$$P(N=0 \mid \lambda_0, \alpha) = 1 - B\left(\frac{\lambda_0}{\lambda_0 + \alpha}, 1, \alpha\right) = \left(1 + \frac{\lambda_0}{\alpha}\right)^{-\alpha}$$

# Fault Tolerant Clustering Model

#### Model parameters

$$E(\lambda) = \lambda_0 = D_0 A$$

$$Var(\lambda) = \frac{\lambda_0^2}{\alpha}$$

This is a one-page summary of fault-tolerance clustering yield model formulae.

Standard Error in 
$$\lambda = \frac{\sqrt{Var(\lambda)}}{E(\lambda)} = \frac{1}{\sqrt{\alpha}}$$

n = Number of defects tolerated.

• Yield formula; probability of  $\leq n$  failures.

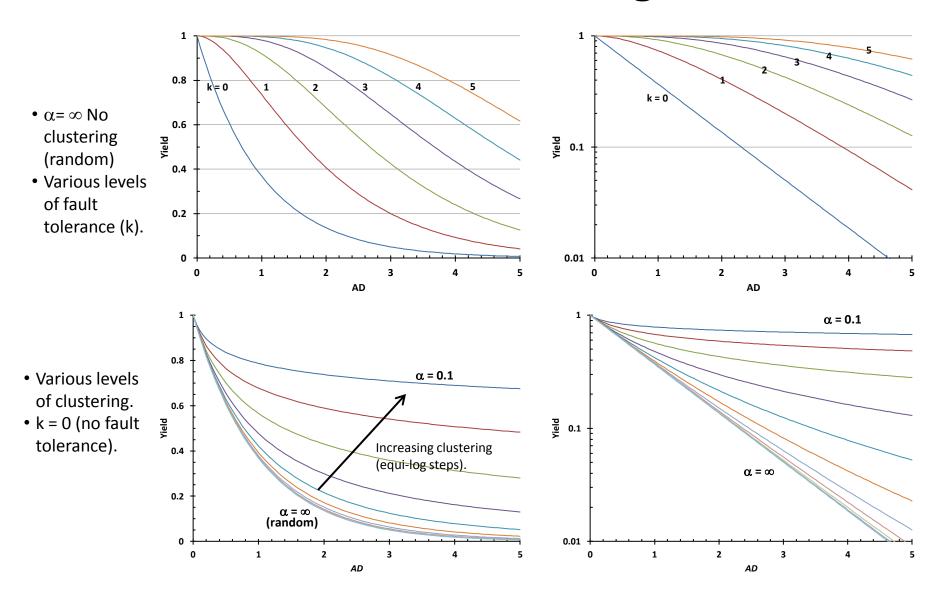
$$P(N \le n \mid \lambda_0, \alpha) = Y = 1 - B\left(\frac{\lambda_0}{\lambda_0 + \alpha}, n + 1, \alpha\right) = 1 - \text{BETADIST}\left(\frac{\lambda_0}{\lambda_0 + \alpha}, n + 1, \alpha\right)$$

Probability of exactly n failures

$$P(N = n \mid \lambda_0, \alpha) = \frac{\Gamma(n + \alpha)}{n! \Gamma(\alpha)} \left(\frac{\lambda_0}{\alpha + \lambda_0}\right)^n \left(\frac{\alpha}{\alpha + \lambda_0}\right)^{\alpha} = \text{NEGBINOMDIST}\left(n, \alpha, \frac{\alpha}{\alpha + \lambda_0}\right)$$

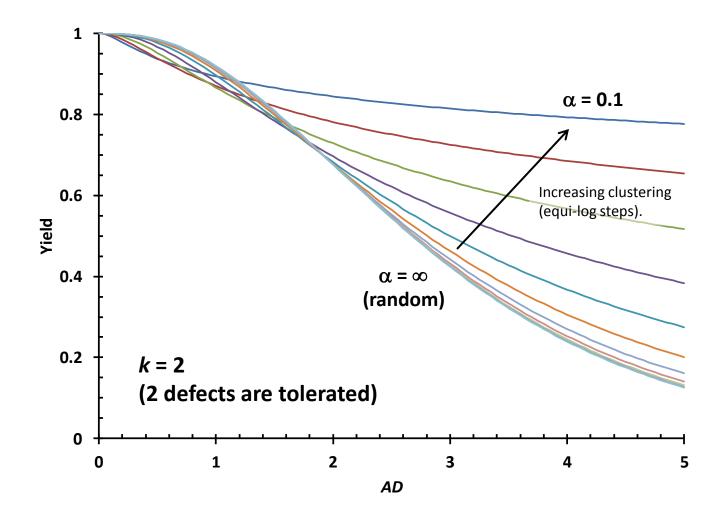
Excel's version of this function requires n and  $\alpha$  to be integers, but real  $\alpha \ge 0$  is meaningful in the theory. It is easy to write a user function for any real  $\alpha \ge 0$ , and integer n.

# Fault Tolerant Clustering Model



# Fault Tolerant Clustering Model

- What happens when fault tolerance and yield interact?
- With fault tolerance, yield can decrease with increasing clustering!



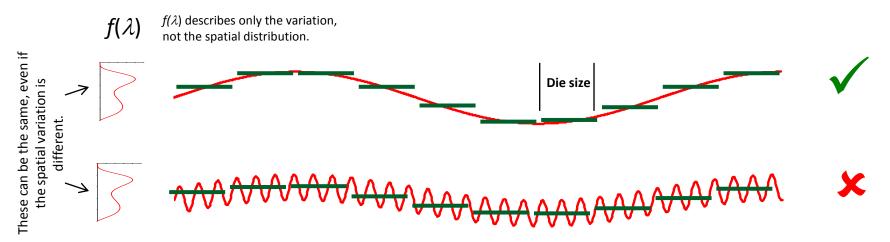
### Limitation of Compound Model

Defect density "compounding" models like this

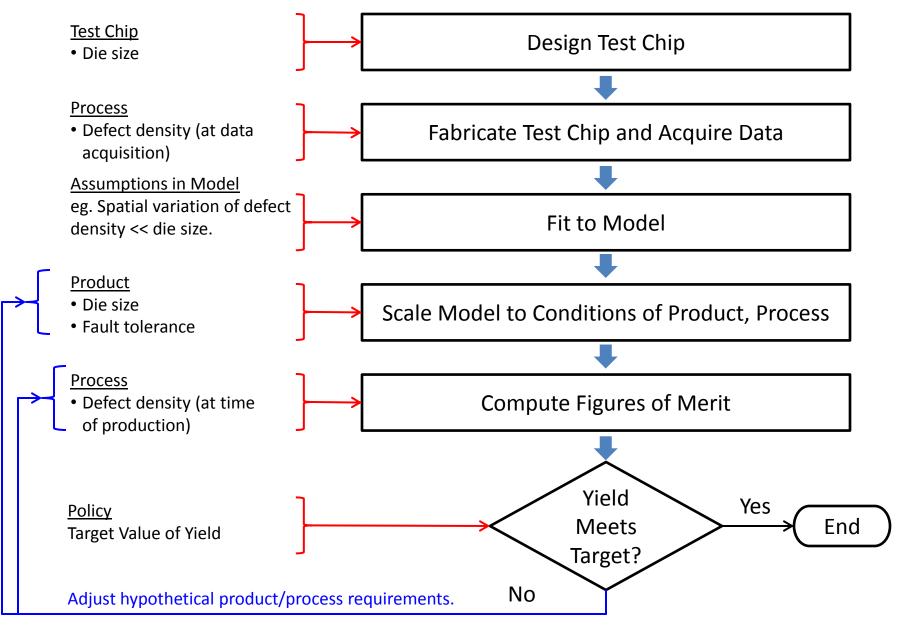
$$P(N = n \mid \lambda_0, \alpha) = \int_0^\infty \frac{\lambda^n \exp(-\lambda)}{n!} f(\lambda) d\lambda$$

are valid only when within-die spatial defect density variation is negligible.

- Compounding models are OK for spatial density variation from
  - Die-to-die
  - Wafer-to-wafer
  - Lot-to-lot
  - Factory-to-factory.



#### Motivation: What's it for?



#### Homework 13.1

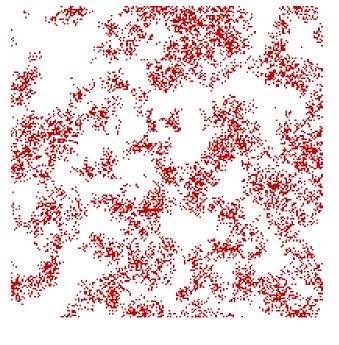
• Yield data was aquired for a test chip with 1 cm<sup>2</sup> area and no fault tolerance enabled. The following windowing data was acquired:

$$\lambda = AD$$
Clustering
$$Y = 1 - B\left(\frac{\lambda}{\lambda + \alpha}, n + 1, \alpha\right) = 1 - \text{BETADIST}\left(\frac{\lambda}{\lambda + \alpha}, n + 1, \alpha\right)$$

$$Y = 1 - B\left(\frac{\lambda}{\lambda + \alpha}, 1, \alpha\right) = \left(1 + \frac{\lambda}{\alpha}\right)^{-\alpha} \qquad (n = 0, \text{ no fault tolerance})$$
No Clustering
$$Y = \sum_{i=0}^{k} \frac{\lambda^{i}}{i!} \exp(-\lambda) = \text{POISSON}(\lambda, k, \text{TRUE})$$

$$Y = \exp(-\lambda) \qquad (n = 0, \text{ no fault tolerance})$$

Window	Yield (%)
1x1	80.00
2x2	51.56
3x3	37.31
4x4	28.80
5x5	24.38
6x6	19.98
7x7	16.78
8x8	11.52
9x9	11.34
10x10	6.75



- Fit the data to a cluster yield model and thereby determine alpha and defect density of the process.
- Use the fitted model to calculate yields for products that can tolerate 0, 1,2, 3, 4, and 5 defects with die sizes ranging from 0.5 cm<sup>2</sup> to 4 cm<sup>2</sup> on the same process as the test chip (same defect density).

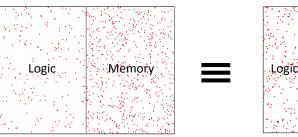
#### Critical Area Formulation

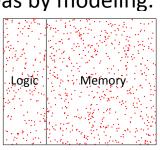
#### Physical picture:

- Defect densities for circuit blocks depend on the sensitivity of the block to defects (eg. memory vs logic).
- Circuit block areas are the physical areas of the blocks.
- Critical area picture:
  - Defect densities for all blocks are the same reference density,  $D_{Reference}$  determined by a standard measure of the fab process.
  - Areas of blocks are different from the physical areas of the circuit blocks.
- Benefit: Clear responsibility for parameters

$$A_{\text{Critical Area}} = \frac{D_{\text{Actual}}}{D_{\text{Reference}}} A_{\text{Physical}}$$

- Manufacturing owns measurement of defect density.
- Design owns determination of critical areas by modeling.





Defect densities greatly exaggerated!

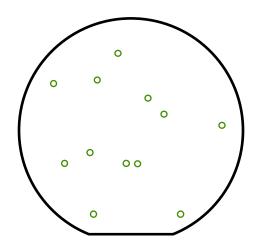
$$Y = Y_{\text{Logic}} Y_{\text{Memory}}$$

$$= \exp\left(-D_{\text{Logic}} \times A_{\text{Logic(physical)}} - D_{\text{Memory}} \times A_{\text{Memory(physical)}}\right)$$

$$\begin{split} Y &= Y_{\text{Logic}} Y_{\text{Memory}} \\ &= \exp \left[ -D_{\text{Reference}} \left( A_{\text{Logic}}' + A_{\text{Memory}}' \right) \right] \end{split}$$

#### Outline

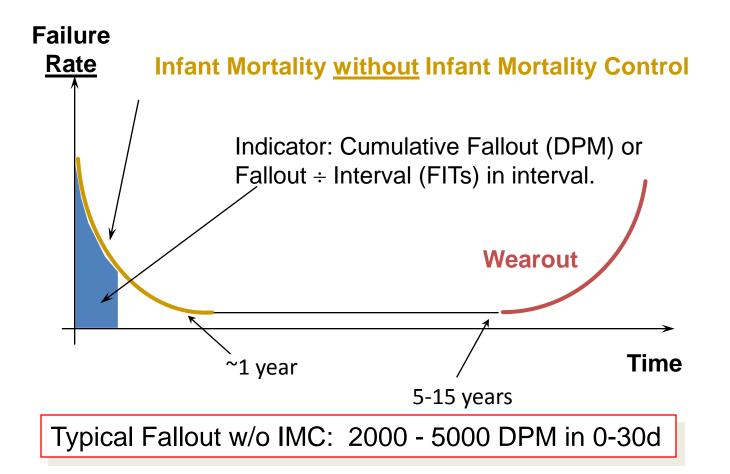
- Introduction
- Models of Yield
- Models of Defect Reliability
- Analysis and Synthesis of Lifetest and Burn In



Reliability Defect Density

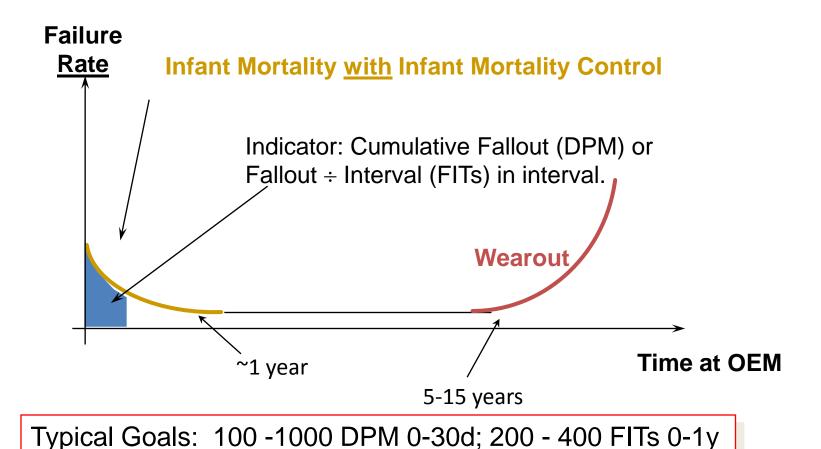
#### **Bathtub Curve**

- Defects: A declining failure rate, affects early life.
- Materials, Design: Wearout, increasing failure rate, affects late life.



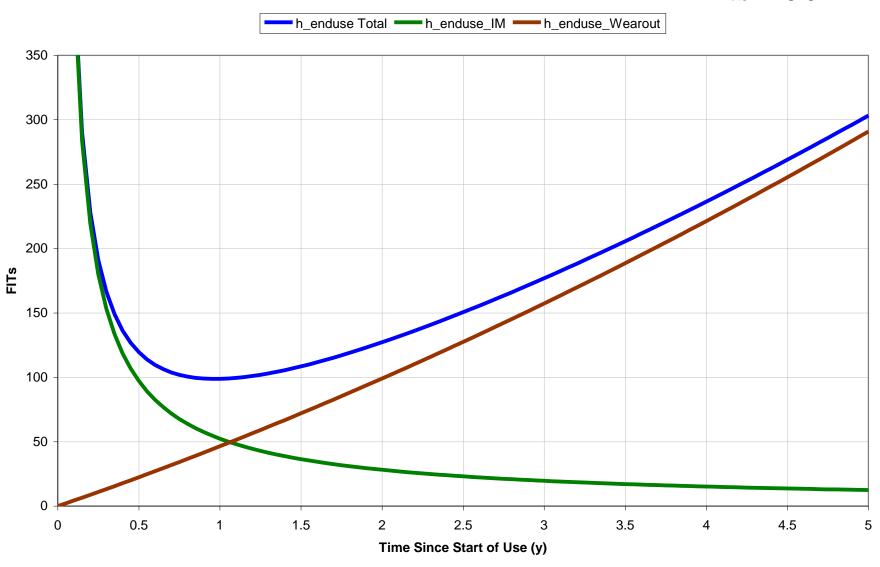
#### Customer-Perceived Bathtub Curve

 Use Infant Mortality Control (eg. Burn In) to reshape the bathtub fail rate curve as perceived by customers.



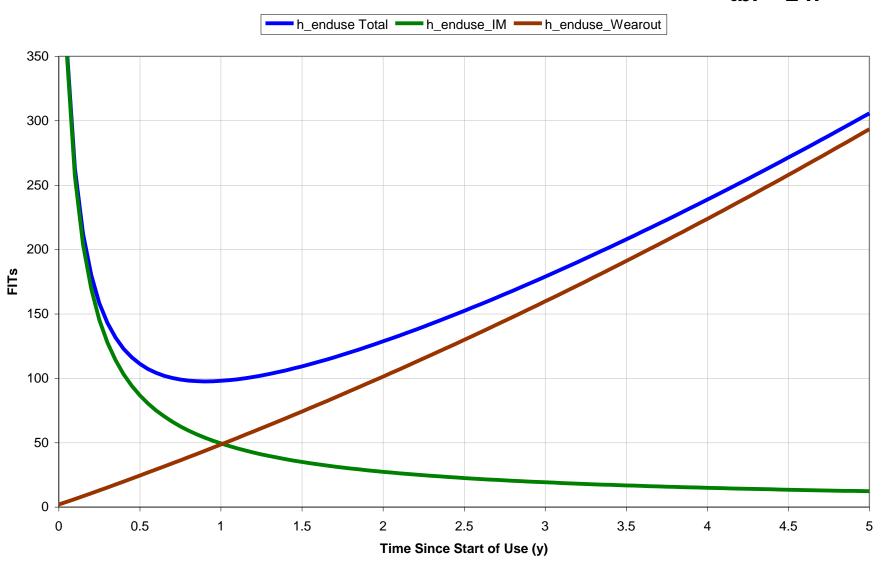
#### Instantaneous Failure Rates (FITs)

tbi = 0.01 h



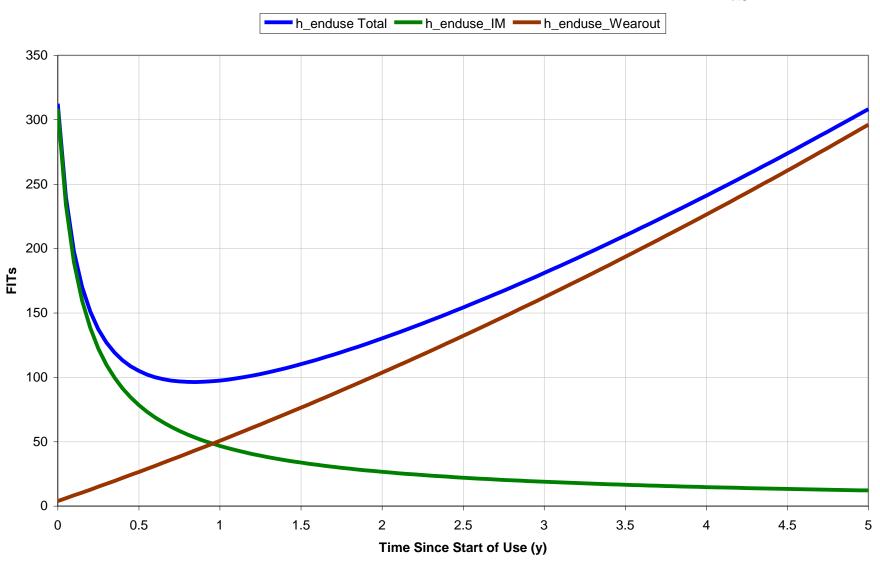


tbi = 2 h



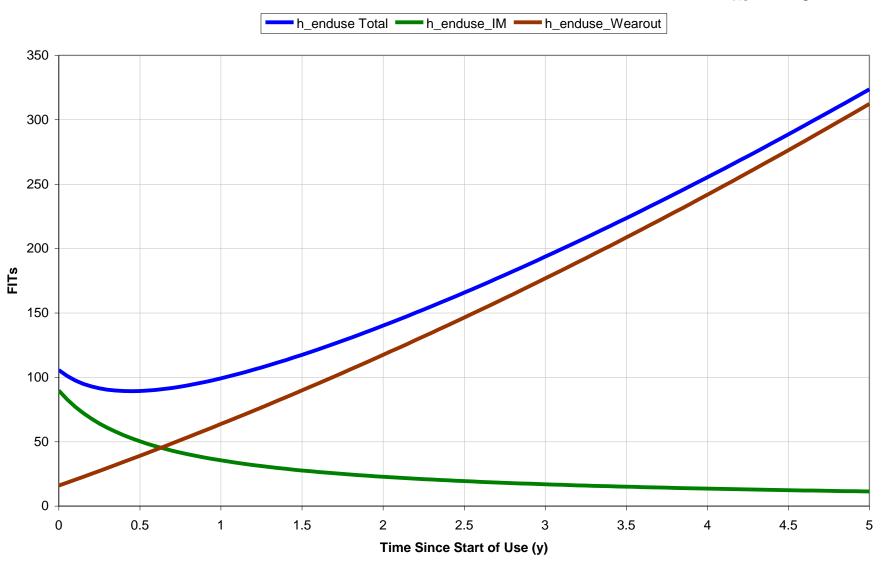


tbi = 4 h





tbi = 16 h



# Defect Model of Reliability

- Each defect has a "time-to-fail" the run time at a specific "reference condition" of temperature, voltage etc. before failure.
- The fraction of defects with time-to-fail less than t is the defect survival function, s(t).
- The number of reliability defects on a die is  $\lambda_{rel} = D_{rel} \times A$
- Survival function of chip is

$$S(t) = s(t)^{\lambda_{rel}} = \exp[\lambda_{rel} \ln s(t)] = \exp[-\lambda_{rel} H(t)]$$

where H(t) is the cumulative hazard of the rel defects.

Example. If defects have a Weibull life distribution

$$s(t) = \exp \left[ -\left(\frac{t}{\alpha}\right)^{\beta} \right], \qquad H(t) = \left(\frac{t}{\alpha}\right)^{\beta}$$

What if the defects have some other distribution?

See NIST gallery of distributions.

 $S(t)=s(t)\times s(t)\times ...$ 

Number of rel defects..

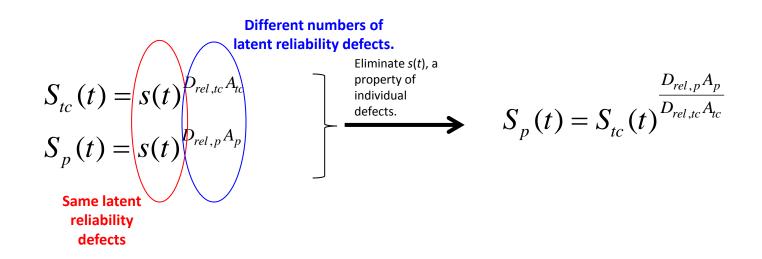
then the survival function of the chip is also Weibull but with different  $\alpha$ .

$$S(t) = \exp\left[-D_{rel}A\left(\frac{t}{\alpha}\right)^{\beta}\right] = \exp\left[-\left(\frac{t}{\alpha'}\right)^{\beta}\right] \qquad \alpha' = \frac{\alpha}{(D_{rel}A)^{1/\beta}}$$
 special to Weibull

**C. Glenn Shirley** "A Defect Model of Reliability," invited <u>Tutorial at 33rd Annual International Reliability Physics</u> <u>Symposium</u>, including <u>supplemental paper</u>. Las Vegas, Nevada, pp. 3.1 - 3.56, 1995.

# Scaling from Test Chip to Product

- Wanted: The survival function of a product from the survival function of a test chip, such as an SRAM, to avoid life-testing the product.
- Survival function of both product and SRAM (test chip, tc) is affected by the same latent reliability defects.
- But the numbers of defects are different because
  - Critical areas are different due to different "physical area" and design layout.
  - Latent reliability defect densities are different because of Fab process variation at time of production of test chip and product.



# Scaling from Test Chip to Product, ct'd

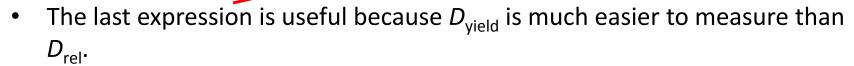
- Problem:  $D_{rel}$  is hard to measure.
- But remember...
  - The same kinds of defects that degrade yield, degrade "infant mortality" perceived by end users.
- So it is plausible that the latent reliability defect density varies in proportion to the "killer" defect density as the Fab yield varies.
  - This is the "special K" concept that Bill Roesch mentioned...

If 
$$\frac{D_{rel,p}}{D_{yield,p}} = \frac{D_{rel,tc}}{D_{yield,tc}} = \kappa$$
 then  $\frac{D_{rel,p}}{D_{rel,tc}} = \frac{D_{yield,p}}{D_{yield,tc}}$ 

- Typically  $\kappa \approx 0.01$  (1%)

So

$$S_p(t) = S_{tc}(t)^{\frac{D_{rel,p}A_p}{D_{rel,tc}A_{tc}}} = S_{tc}(t)^{\frac{D_{yield,p}A_p}{D_{yield,tc}A_{tc}}}$$



At this point, we can forget about  $D_{rel}$  because the models for relating test chip (aka SRAM, aka "ref") only involve  $D_{vield}$ 

Special

# Scaling from Test Chip to Product, ct'd

- So we now have a way to scale the SRAM (test chip) model to the product considering:
  - Different die areas between SRAM and product.
  - Yield (killer) defect densities different at the time of production of the SRAM and the time of production of the product.
  - The product operating at different temperature and voltage from the SRAM.

# Area/Defect Density Scaling

The scaling factor of a product (area A) produced on a process (with defect density D) to a reference such as a test vehicle is

Special

$$v_{p|ref} = \frac{D_p A_p}{D_{ref} A_{ref}}$$
  $D_p$  and  $D_{ref}$  are killer defect densities (ie  $D_{yield}$ ) measured at Sort.



 Example. Suppose test chip life data at a specific temperature and voltage is fitted to a Weibull distribution to produce a reference model. The product's survival function at the reference temperature and voltage is

$$S_{tc}(t) = \exp\left[-\left(\frac{t}{\alpha_{tc}}\right)^{\beta}\right]$$

$$S_{p}(t) = \exp\left[-v_{p|tc}\left(\frac{t}{\alpha_{tc}}\right)^{\beta}\right] = \exp\left[-\left(\frac{t}{\alpha_{p}}\right)^{\beta}\right] \text{ where } \alpha_{p} = \alpha_{tc}v_{p|tc}^{-1/\beta}$$

• That is, the product survival function is also Weibull with the same shape  $\beta$ , but with different  $\alpha$ .

$$\ln \alpha_p = \ln \alpha_{tc} - \frac{1}{\beta} \ln \nu_{p|tc}$$
$$\beta_p = \beta_{tc}$$

#### Acceleration

A commonly used burn in acceleration model is

$$A_{21} = \exp\left\{\frac{Q}{k} \left[ \frac{1}{T_1} - \frac{1}{T_2} \right] + C(V_2 - V_1) \right\}$$

- Where
  - $T_2$ ,  $V_2$ ,  $T_1$ ,  $V_1$  are operating temperatures (in deg K) and voltages at conditions 2 and 1, respectively.
  - $k = 8.61 \times 10^{-5} \text{ eV/K}$  is Boltzmann's constant.
  - Q (eV) is the thermal activation energy.
  - C (volts<sup>-1</sup>) is the voltage acceleration constant.
- Example: What is the acceleration of burn in relative to Use for the following conditions:

	T (C)	V
BI Condition	135	4.6
Use Condition	85	3.3

#### **Acceleration Model Parameters**

Q	0.3	eV
С	2.6	<b>/</b> V

Ans: 96.7

#### Burn In

- Now we put together defect density/area scaling and acceleration to write the complete model.
- The survival function of a product which has undergone  $t_{bi}$  hours at burn in conditions followed by  $t_{use}$  hours in use is

$$S_{p}(t_{use},t_{bi}) = \left[S_{ref}\left(AF_{bi|ref}t_{bi} + AF_{use|ref}t_{use}\right)\right]^{\nu_{p|ref}} \qquad \qquad \nu_{p|ref} = \frac{D_{p}A_{p}}{D_{ref}A_{ref}}$$

 The customer-perceived survival function is conditioned on surviving the test screen following burn in

$$S_{p}\left(t_{use} \mid t_{bi}\right) = \frac{S_{p}\left(AF_{bi|ref}t_{bi} + AF_{use|ref}t_{use}\right)}{S_{p}\left(AF_{bi|ref}t_{bi}\right)} = \left[\frac{S_{ref}\left(AF_{bi|ref}t_{bi} + AF_{use|ref}t_{use}\right)}{S_{ref}\left(AF_{bi|ref}t_{bi}\right)}\right]^{\nu_{p|ref}}$$

- From the product survival function any figure of merit can be calculated.
  - Fractions failing (DPM) between two times, expressed in DPPM.
  - Average failure rates (AFR) between two times, expressed in Fits.