Infant Mortality Control by Burn In and by Design

C. Glenn Shirley
Intel Corporation

With contributions from Ben Eapen.

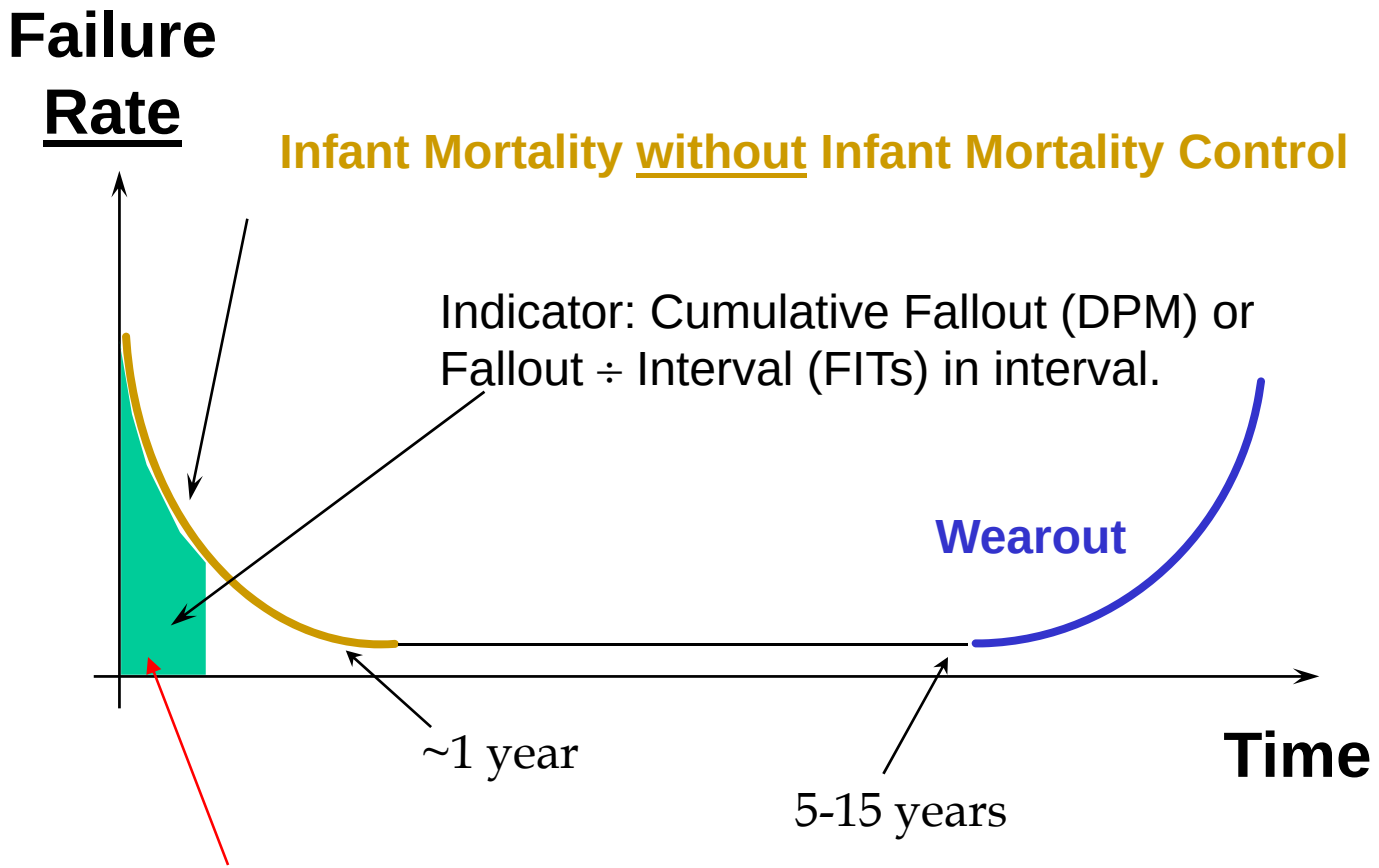
Outline

- Introduction
- Infant Mortality Control by Burn In (Manufacturing)
- Infant Mortality Control by Design
- Key Messages

Introduction

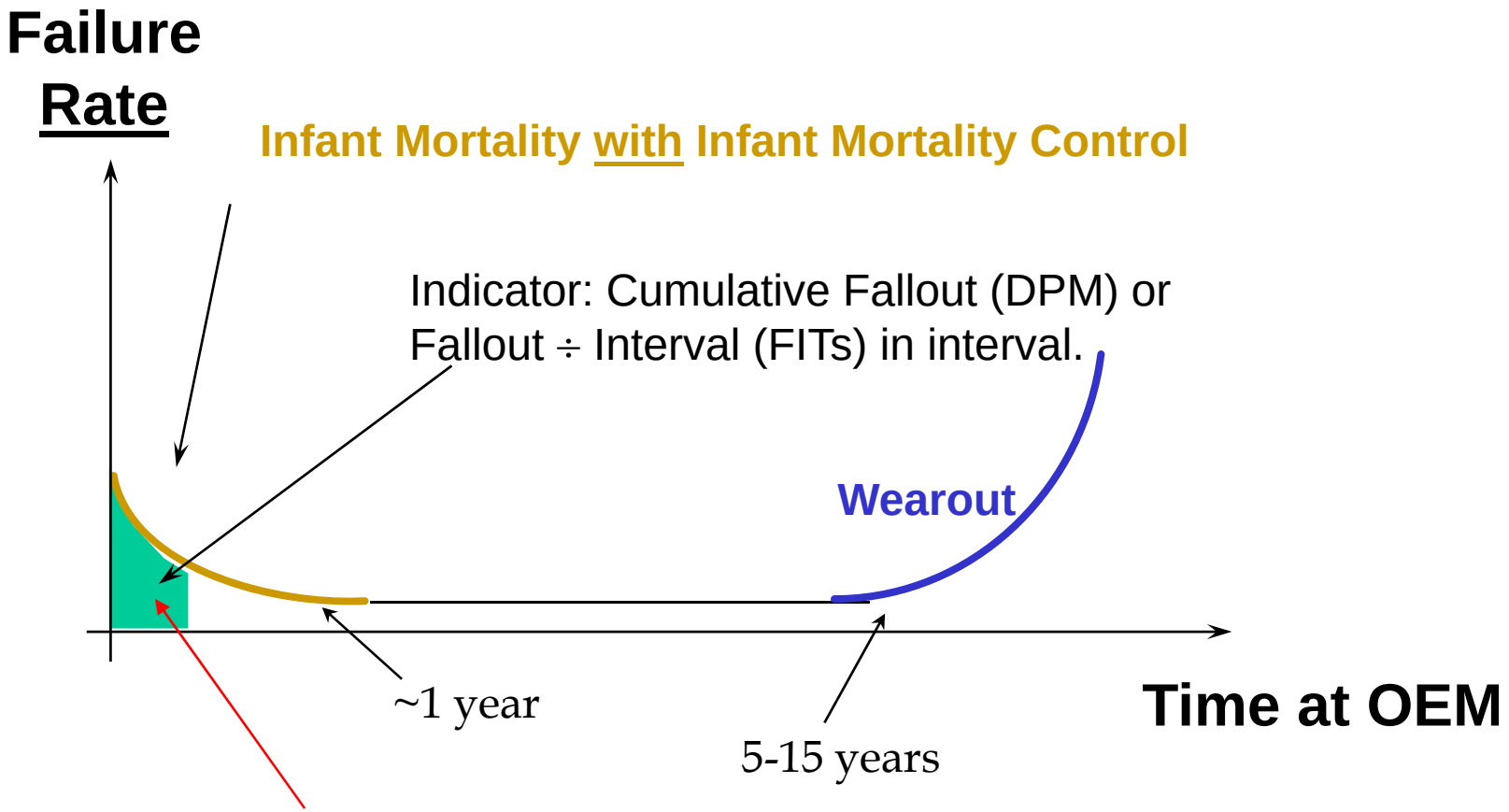
- Silicon fabrication introduces latent reliability defects which cause early-life failure - infant mortality (IM).
- Without IM control, IM DPM may be too high.
  - eg. Microprocessors need to have 0-30day IM DPM reduced from ~2000-5000 DPM to < 1000 DPM.
- Control the customer-perceived “bathtub curve” by
  - Applying stress (V, T) as part of manufacturing process flows.
    - Burn In to push weak units “over the edge” so that they can be screened in subsequent test.
  - Design for defect tolerance in “use”.
    - So that hard defects appearing after test will not affect performance.

Bathtub Curve



Typical Fallout w/o IMC: 2000 - 5000 DPM in 0-30d

Customer-Perceived Bathtub Curve



Typical Goals: 100 -1000 DPM 0-30d; 200 - 400 FITs 0-1y

Approaches to Infant Mortality Control

- Manufacturing (Burn In)
  - Burn In applies stress to activate latent reliability defects before final test.
  - Declining failure rate for latent reliability defects means that customer perceived IM is reduced.
  - Burn In conditions (time, temperature, voltage) are adjusted to meet IM and Wearout reliability goals, remain functional, and avoid thermal runaway.
- Design
  - Design devices, or parts of devices (cache), for tolerance to hard defects.
  - Hard-defect-tolerant areas of devices don't need burn in.
    - Reduced effective area means less (or no) burn in.
    - If hard-defect-tolerant areas can be electrically isolated in burn in, then burn in hardware power requirements are less.

Outline

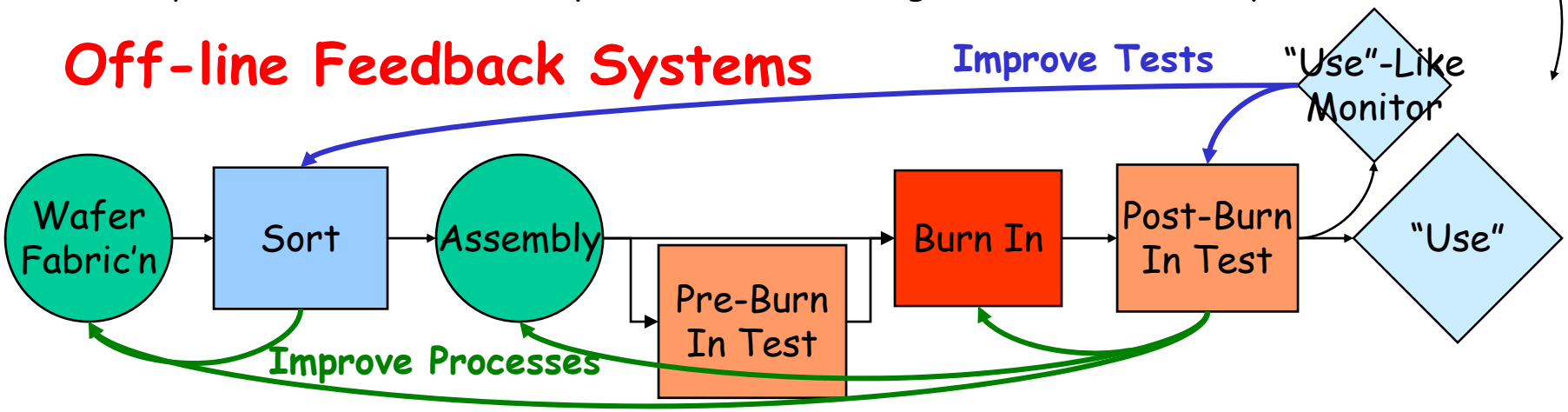
- Introduction
- Infant Mortality Control by Burn In (Manufacturing)
- Infant Mortality Control by Design
- Key Messages

Traditional Manufacturing Flow

Wafer Fabrication
- Source of Si Fabrication Defects
- Density = Dfab

Assembly
- Source of Package Assembly Defects
- Opens/Shorts/Leakage

Use DPM From
- Test Holes
- Additional latent reliability defects: Duse.



Sort
- Initial screen. Coarse electrical/thermal control, loose timings.
- Cold temperature to screen cold defects.
- Feedback to Fab.

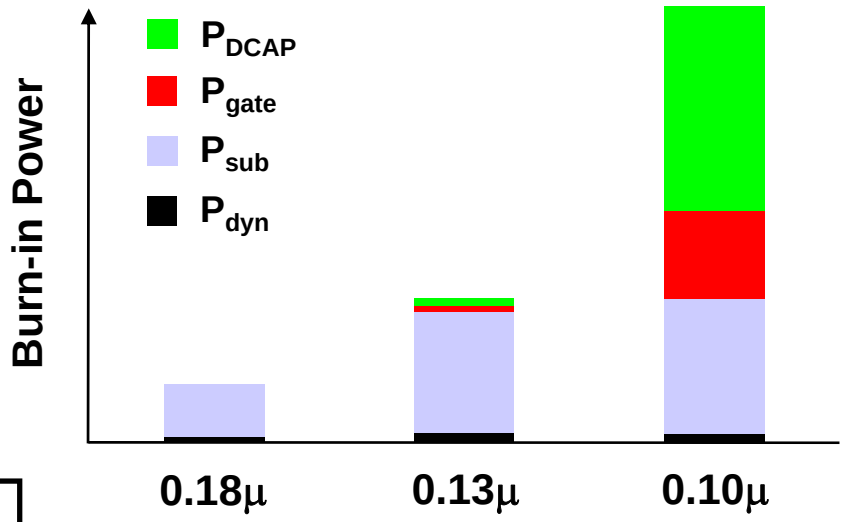
Burn In
- Exercise DUTs at Vcc and Tj > "use".
- Limited by DUT power, and intrinsic rel degradation.
- Induces additional Si defect density Dbi ("turns on" latent rel defects.)

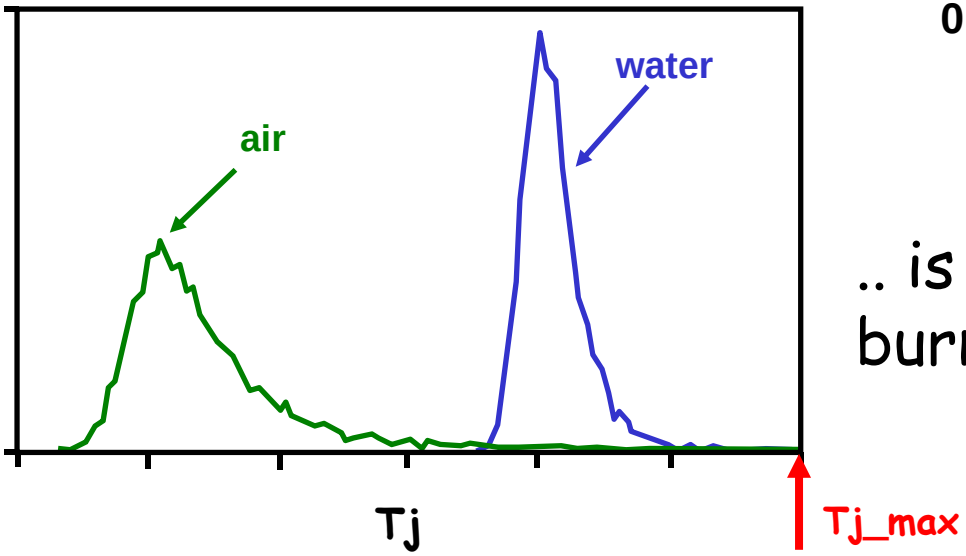
Post-Burn-In Test
- Final screen. Fine electrical/thermal control, tight timings.
- Hot temperature, low Vcc to guarantee spec.
- Infant mortality fallout is feedback to Fab and Assembly.

BI Mfg Trends: Power Management

- Burn in is done at high Tj and Vcc, but low frequency.
  - Under these conditions, static power dominates. (Idyn is small.)
- Power trends.
  - Itotal = Isub + Igate + Idcap + Idyn
  - Isub - subthreshold leakage current.
    - V-sensitive: increases 15-20% for a 0.1V increase
    - T-sensitive: increases 25-30% for a 10°C increase
    - Large (10X) within-wafer, -lot variation (sensitive to Le variation)
  - Oxide Leakage. Gate oxide leakage due to transistors (Igate) and decoupling capacitors (Idcap).
    - V-sensitive: increases 25-30% for a 0.1V increase
    - T-*in*sensitive: increases 30% for an increase from 0°C to 95°C
    - tox-sensitive: increases 2.5x for a 1Å decrease
    - Small statistical variation.

Increasing burn-in power..





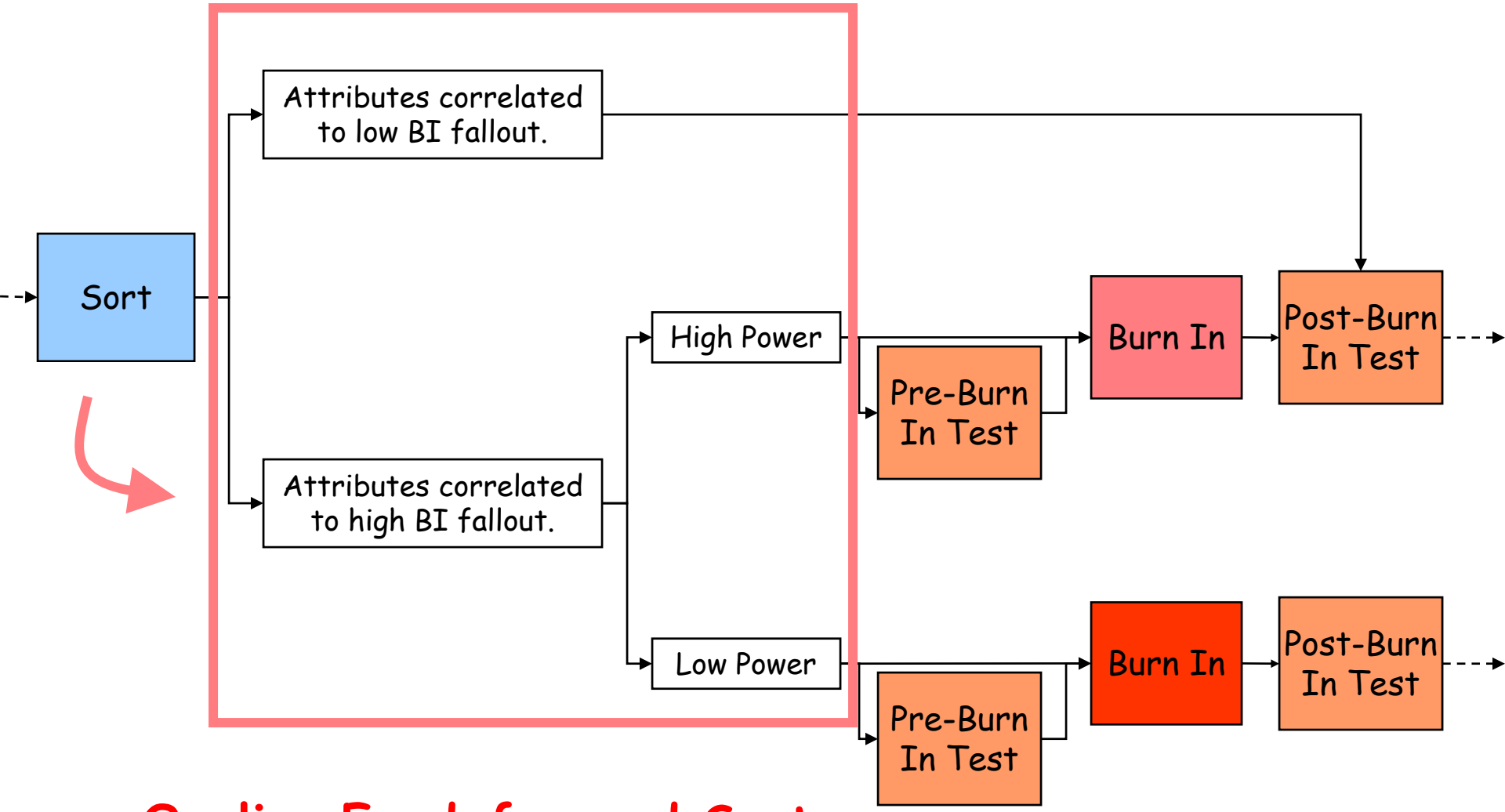
.. is managed by improved burn-in hard ware.

Improved thermal impedance gives shorter burn in times for the same Tj_max limit.

Wafer-Level Burn In

- Factors affecting feasibility:
  - Die size/defect density:
    - Small dies, $< 0.2\ cm^2$, low defect density, low probe count/die.
  - Available Stress
    - Hardware limitations for high power and temperature.
    - Maximum stress voltage. Limited by wearout.
  - Feasibility of parallel testing (especially wafer-level)
    - Requires small number of probes per die.
    - Available test time. (A few seconds per die.)
- Benefit: Convenient where feasible, but..
- Con: Limited envelope.
- Status:
  - Used by mature technologies, small dies, low power, etc.

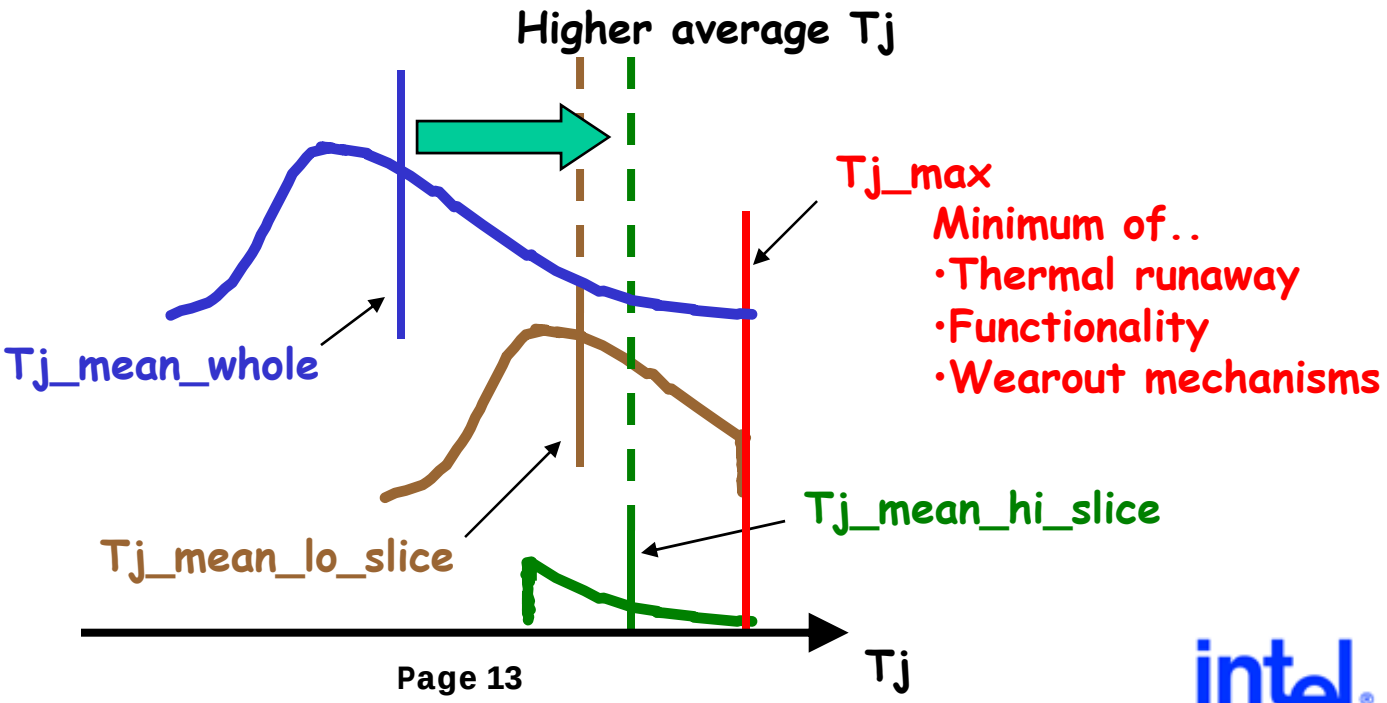
Adaptive Manufacturing Flows



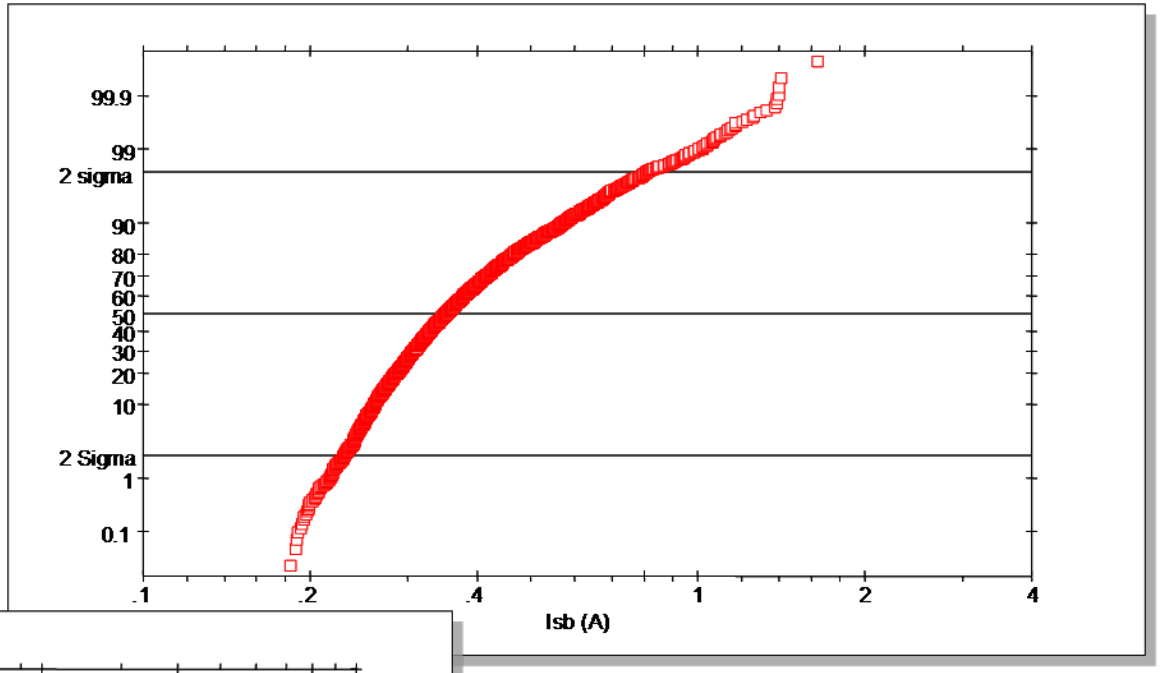
On-line Feed-forward Systems

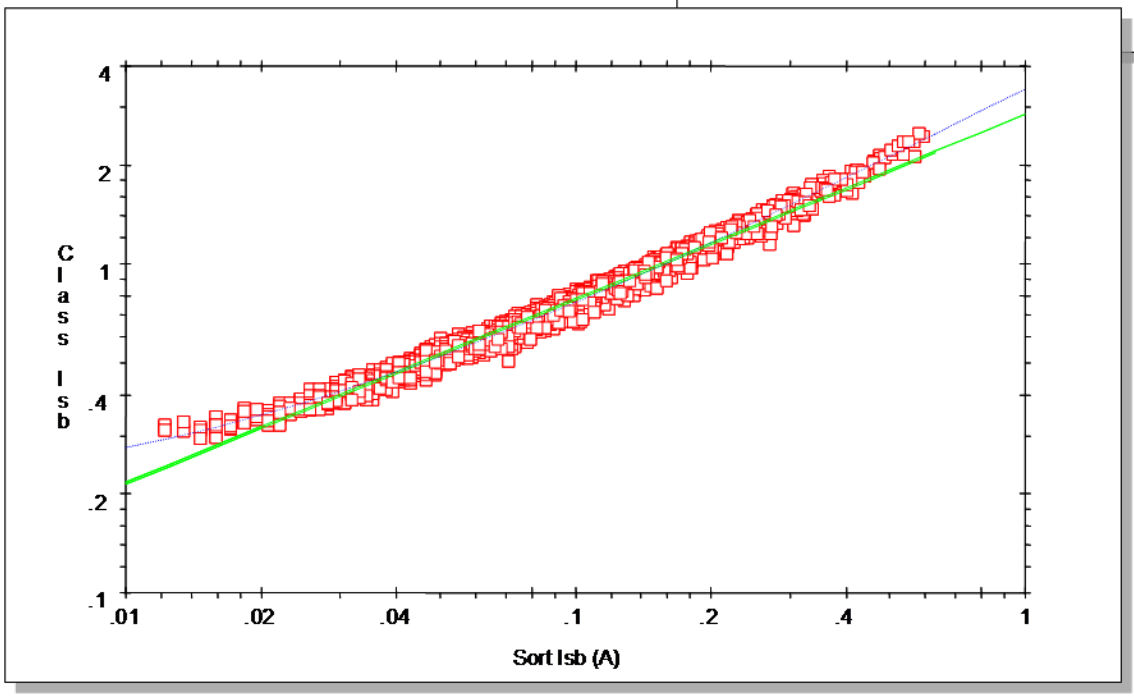
Adaptive Manufacturing Flows

- Manage Power by Sort Isb signal
  - Isb distributions are broad, but Sort/BI Isb correlation is good.
  - Benefit: Optimize utilization of burn in hardware.
  - Con: Complex manufacturing flows.
  - Status: Currently used in manufacturing.



Broad Isb Distribution, but...





..good Sort-Class Correlation

0.18μ technology.

Adaptive Manufacturing Flows

- Use Lot-level and wafer-level statistics at Sort.
  - Routing or screen depends on post-processed lot-level and/or wafer-level statistics.
    - Relies on quality of an established correlation between Sort “killer defect” density” and “latent reliability defect density”.
  - Benefit: Optimal utilization of BI hardware.
  - Cons:
    - Correlation is usually not good enough to enable a screen.
      - Screen must meet reliability goals without unacceptible overkill.
    - Can be used to optimize utilization of BI hardware (variable BI time)
      - But ROI depends on detailed defect distributions, and complex flows.
    - Depends on post-Si characterization of defects–late for planning.
    - Risk of reliability escapes for excursions.
  - Status: Used for mature processes.

W. C. Riordan, R. Miller, J. M. Sherman, J. Hicks (Intel), “Microprocessor Reliability Performance as a Function of Die Location for a 0.25μ, Five Layer Metal CMOS Logic Process”, IRPS 1999.

Adaptive Manufacturing Flows

- Defect-based die-level tests at Sort.
  - Kill or reroute dies at Sort by tests which detect defective dies. Combine..
    - Sensitive parametric measurements (eg. Iddq).
      - Possibly at several Vcc levels
    - Stress (brief) to activate defects.
  - Benefit: Optimal utilization of burn in hardware.
  - Cons:
    - Hard to find a signal buried in intrinsic Isb.
    - Tests/stress must be short.
    - Usually cannot be used as a screen (overkill).
      - Unless conditions for wafer-level burn in are satisfied.
    - Depends on post-Si characterization of defects–late for planning.
  - Status: Has not found wide application in high-volume manufacturing for logic. Memories (flash) may use this.

Outline

- Introduction
- Infant Mortality Control by Burn In (Manufacturing)
- Infant Mortality Control by Design
- Key Messages

Design for Infant Mortality Control

- Burn In reduces the number of latent reliability defects escaping final test.
- An alternative approach is to make dies tolerant to hard defects in “use”.
- We use a simple model which shows the infant mortality DPM benefit of “hard” fault tolerance.
- Manufacturing benefits derive from
  - Reduced burn in time.
  - Lower power requirements if areas of dies “immune” to hard defects don’t need to be powered in burn in.

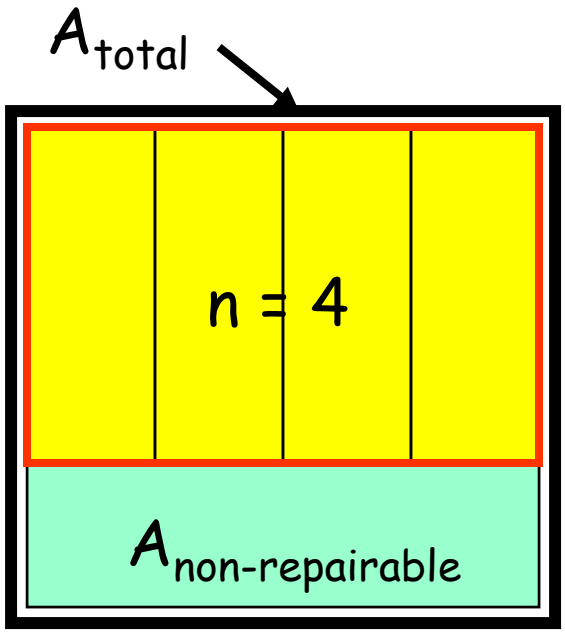
Redundancy Statistics

- Chip has repairable (usually cache) and non-repairable (usually random logic) areas.
  - Define $r = A_{repairable}/A_{total}$
- The repairable area of the chip is divided into a number "n" of repairable elements.
  - The larger n is, the more "survivable" is the chip, and the greater is the design/area overhead.
- Each repairable element is characterized by the number of defects it can "survive".
  - Assumption here: Repairable elements can survive up to 1 defect, and non-repairable cannot survive more than 0 defects.
  - There are different circuit/logic ways to realize this.

Note: This description is an approximation intended only to show the major sensivities.

Redundancy Statistics, cont'd

- Some *kinds* of defects are fatal even to repairable elements, depending on the redundancy scheme used.
  - f = fraction of all kinds of defects which can be repaired by repairable elements.

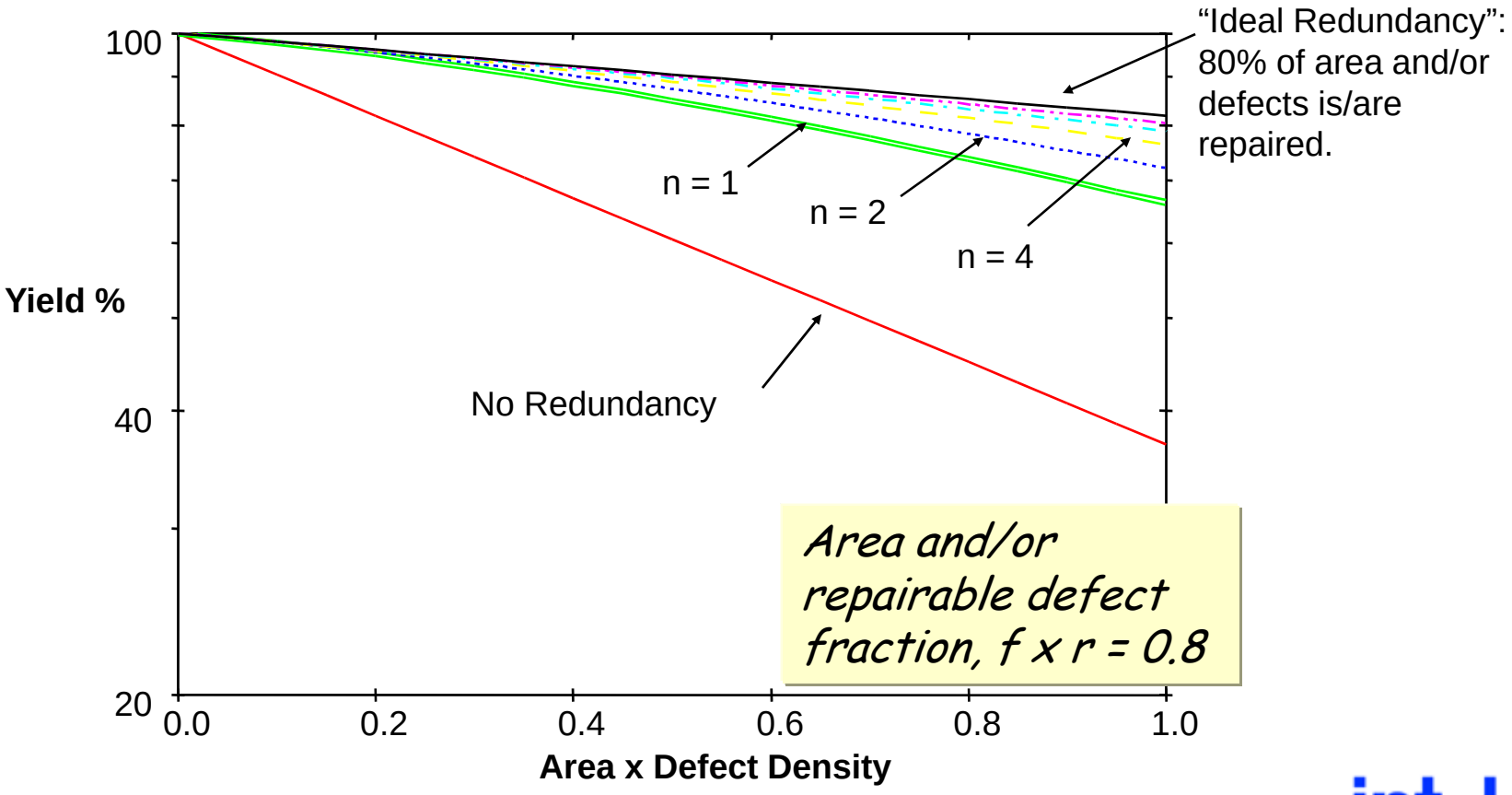


Two Limiting Special Cases

- No redundancy at all. (f x r = 0, irrespective of n).
  Yield and Infant Mortality for A_{total}.
- Ideal Redundancy. (n = very large).
  Yield and Infant Mortality for $A_{non\text{-}repairable}$

Yield Example

- Test programs at first test screen (eg. Sort) detect faults and connect “spare” elements (eg. by fusing).
  - Big yield gain for n = 1, diminishing return for n > 1.



Infant Mortality & Fault Tolerance

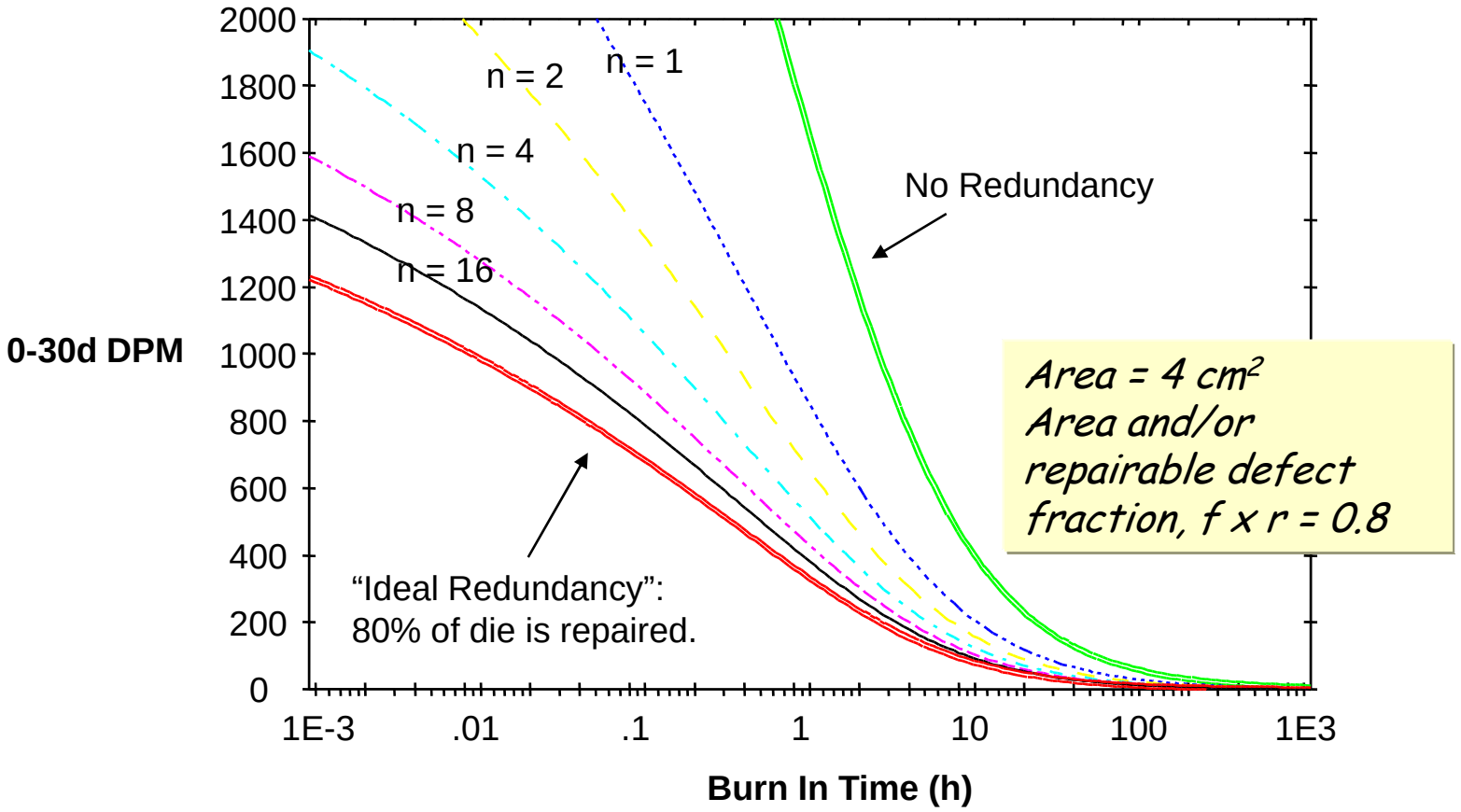
- Main opportunity is “in use” repair or tolerance of latent reliability defects escaping burn in - “Infant Mortality”.
  - Very little gain in *yield* for repair after burn in.
- Requires on-chip logic to detect and replace failing elements with “spares”, or correct data in failing elements.
- What is fraction of dies failing in 0-30d which have survived Sort, burn-in, and post burn-in test?
  - Account for repairs at Sort making redundant elements unavailable at burn in and in “use”.
  - As function of *f*, *r*, *n*, and burn in time (t_{bi})

Note: The following examples are not representative of Intel processes.

intel

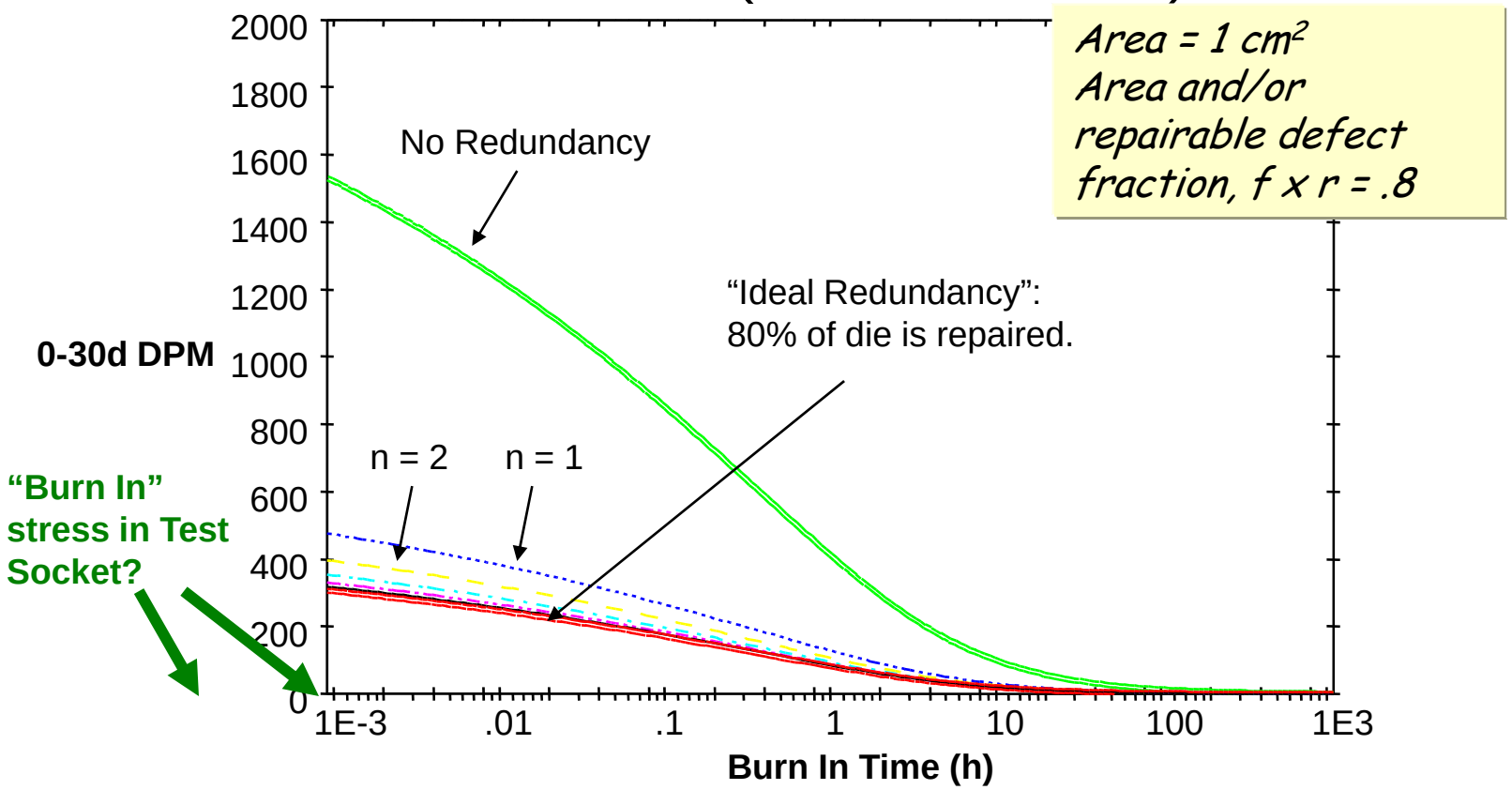
Infant Mortality Large Die Example

- 16-elements are needed to get most of available benefit.
- 10-20X burn in time reduction, depending on goal.



Infant Mortality Small Die Example

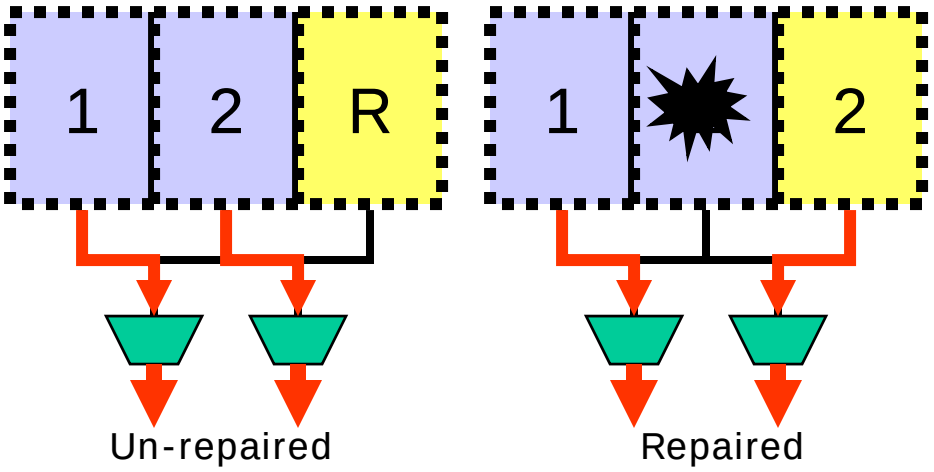
- 1 redundant element is sufficient for a large effect.
- Burn In stress time may be reduced enough to move the stress to a test socket. (10^{-3} h = 3.6 sec).



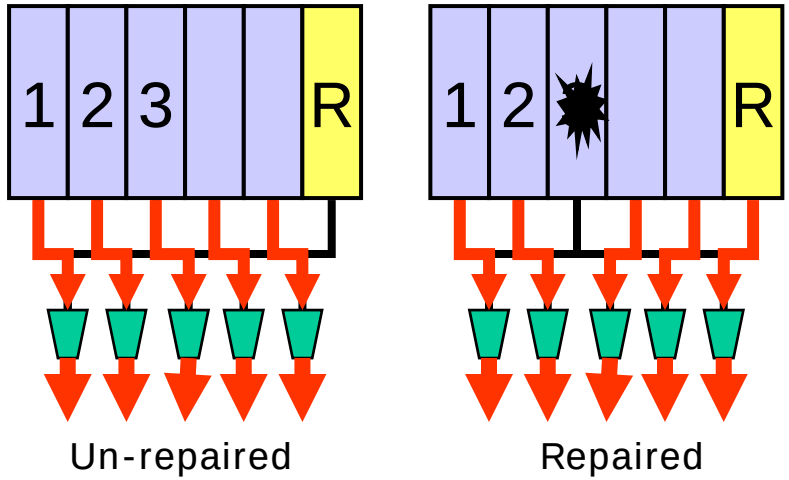
Fault-Tolerance Requirements

- Infant Mortality benefit requires “In Use” fault tolerance.
  - Mostly cache-oriented on-chip schemes, transparent to OEMs.
- Fault-tolerance requires:
  - Test to detect faults.
  - Logic to replace failing elements with “spares”, or to correct data.
- Kinds of In-Use Fault Tolerance
  - Test during POST, set up logic to avoid faults (redundancy).
    - Doesn’t reliably cover all spec conditions.
  - On-the-fly fault detection and repair/correction (ECC).
- Optimal implementation depends on
  - Effectiveness. Kind of scheme vs kind of defect vs defect pareto.
  - Cost: Area impact.
  - Performance impact.

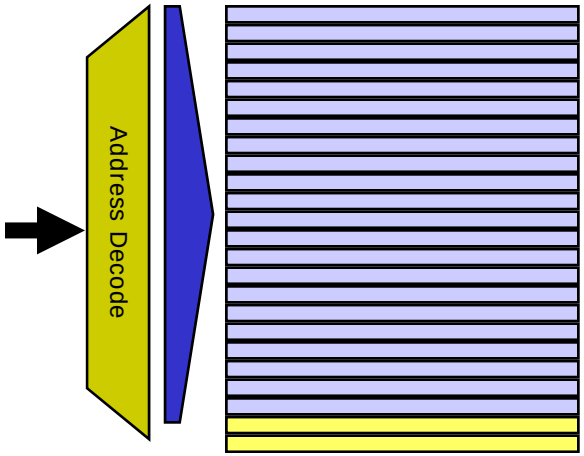
Kinds of Repair Schemes



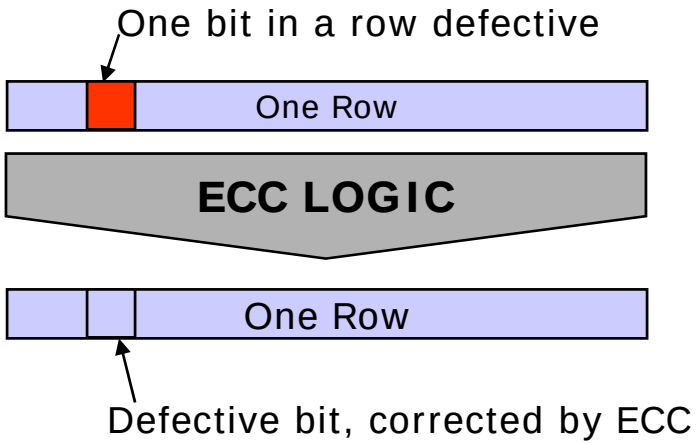
Block Repair



Column Repair



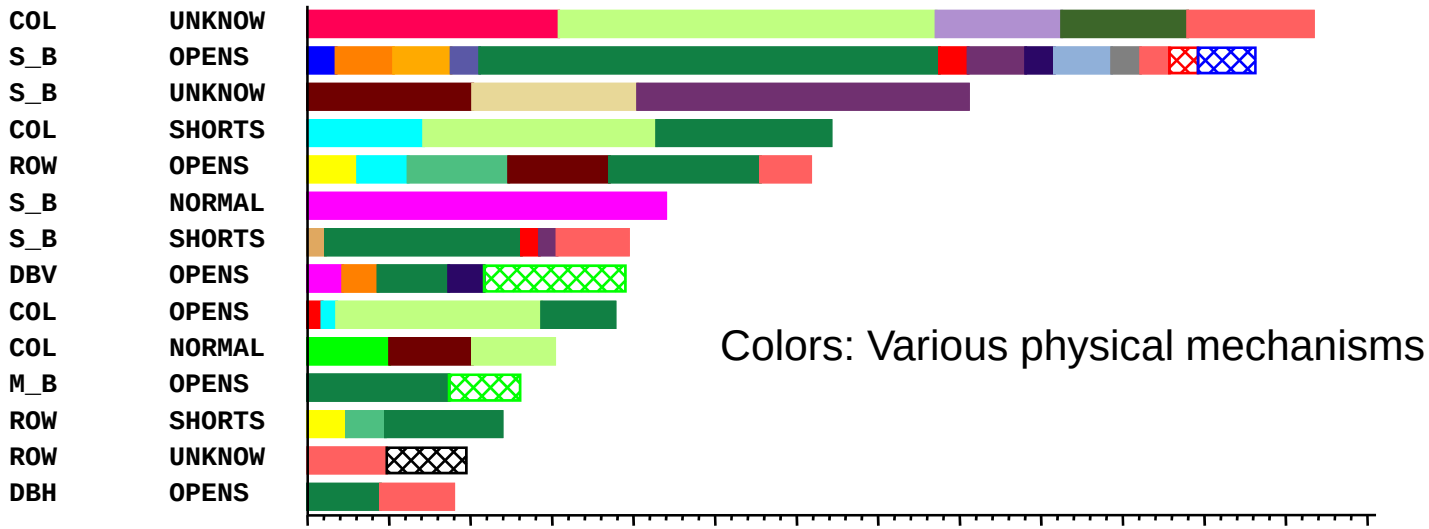
Row Repair



ECC Repair

Source: Ben Eapen

Failure Mode Pareto



- 4 Major failure modes in cache
  - Random Single-Bit Fails predominate.
  - Clustered (in Row/Column) Single Bit Fails
  - Column Fails
  - Row Fails
  - Array Fails

Source: Ben Eapen

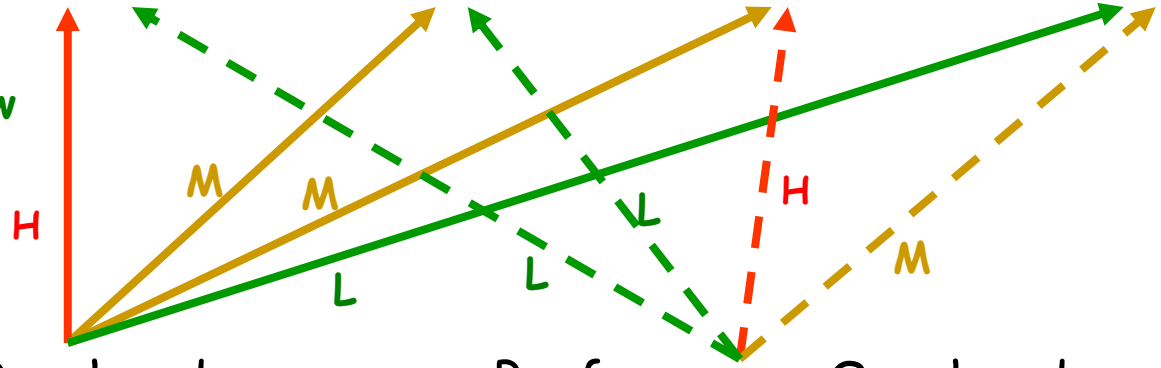
Repair Efficiency

Repair Scheme

Source: Ben Eapen

Fail Mode	Block	Column	Row	ECC
Random SB	✓	✓	✓	✓
Clustered SB	✓	–	–	–
Column	✓	✓	×	–
Row	✓	×	✓	×
Array	✓	×	×	×

H/M/L = High/Med/Low



✓ f is large (~1)
× f is small (~ 0)
- f depends in details of pareto & implementation

Outline

- Introduction
- Infant Mortality Control by Burn In (Manufacturing)
- Infant Mortality Control by Design
- Key Messages

Key Messages

- Traditional Burn In is running into limits.
  - Envelope is shrinking.
    - BI requirements limit end-use performance.
      - High-T, low freq, functionality impacts end-use performance.
    - Tradeoff between wearout and IM is problematic for large dies.
      - Long burn in times.
  - Hardware elect/thermal capability reqts are becoming expensive.
    - Largest part of test costs.
  - Adaptive manufacturing flows help, but are complex and have business risk.
    - Require real-time data automation.
    - Require elaborate simulation models to optimize.
    - Rely on post-Si characterization of defects.
    - Are more vulnerable to excursions.

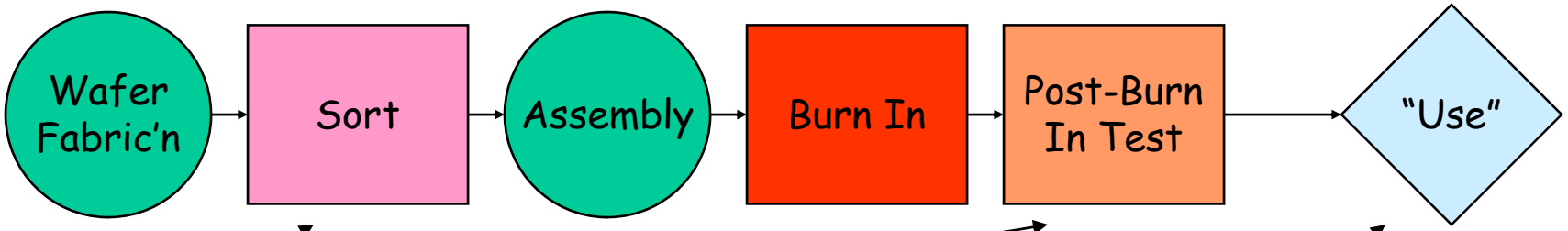
Key Messages, ct'd

- Hard-fault-tolerant cache designs reduce or eliminate burn in reqts.
  - Optimal (performance, area) fault tolerance schemes depend on nature of defect pareto.
  - Benefit:
    - Shorten burn in time (less wearout, reduced capacity).
    - Reduce power envelope if combined with power management.
    - Enable burn in for large dies.
  - Con:
    - Potentially impacts design costs, chip costs, and performance.
    - Only works for products which have appreciable memory.
- Next?
  - Cost-effective hard fault tolerance for logic.

Backup

Models of Defect Density

- Latent Reliability Defect Density vs Time & Stress
  - Lognormal time cumulative fraction failing distribution is used.
  - σ, μ, and AF are determined from test chip (SRAM) post-burn in test fallout vs burn in time and Tj, Vcc variation experiments.
  - Example values: $\sigma = 25$, $\mu = 70$, AF = 200.



D_{fab} from Yield at Sort

$$D_{bi} \cong 0.01 \times D_{fab} \frac{\ln\left[1-\Phi\left(\frac{\ln(t_{bi})-\mu}{\sigma}\right)\right]}{\ln\left[1-\Phi\left(\frac{\ln(t=1\,\mathrm{h})-\mu}{\sigma}\right)\right]}$$

$$D_{use} \cong 0.01 \times D_{fab} \frac{\ln\left[1-\Phi\left(\frac{\ln(t_{bi}+t_{use}/AF)-\mu}{\sigma}\right)\right]}{\ln\left[1-\Phi\left(\frac{\ln(t=1\,\mathrm{h})-\mu}{\sigma}\right)\right]}$$

(Assumes that the BI defect density is defined at 1h of BI.)

Redundancy Model for Yield

Backup

- Probability of a good die after Sort is given by

(Prob. of 0-defect redundant sub-element

or a 1-defect sub-element)$^{\text{Number of repairable sub-elements}}$

and Probability of 0 defects in the non-repairable portion of the die.

That is, $Y = [Y_r^0 + Y_r^1]^n Y_{nr}$

- Using Poisson expressions for probabilities in terms of defect density we get

$$Y = \left(1 + \frac{f \times r \times A_{tot} \times D}{n}\right)^n \times \exp(-A_{tot} \times D)$$

intel

Redundancy Model for Infant Mortality

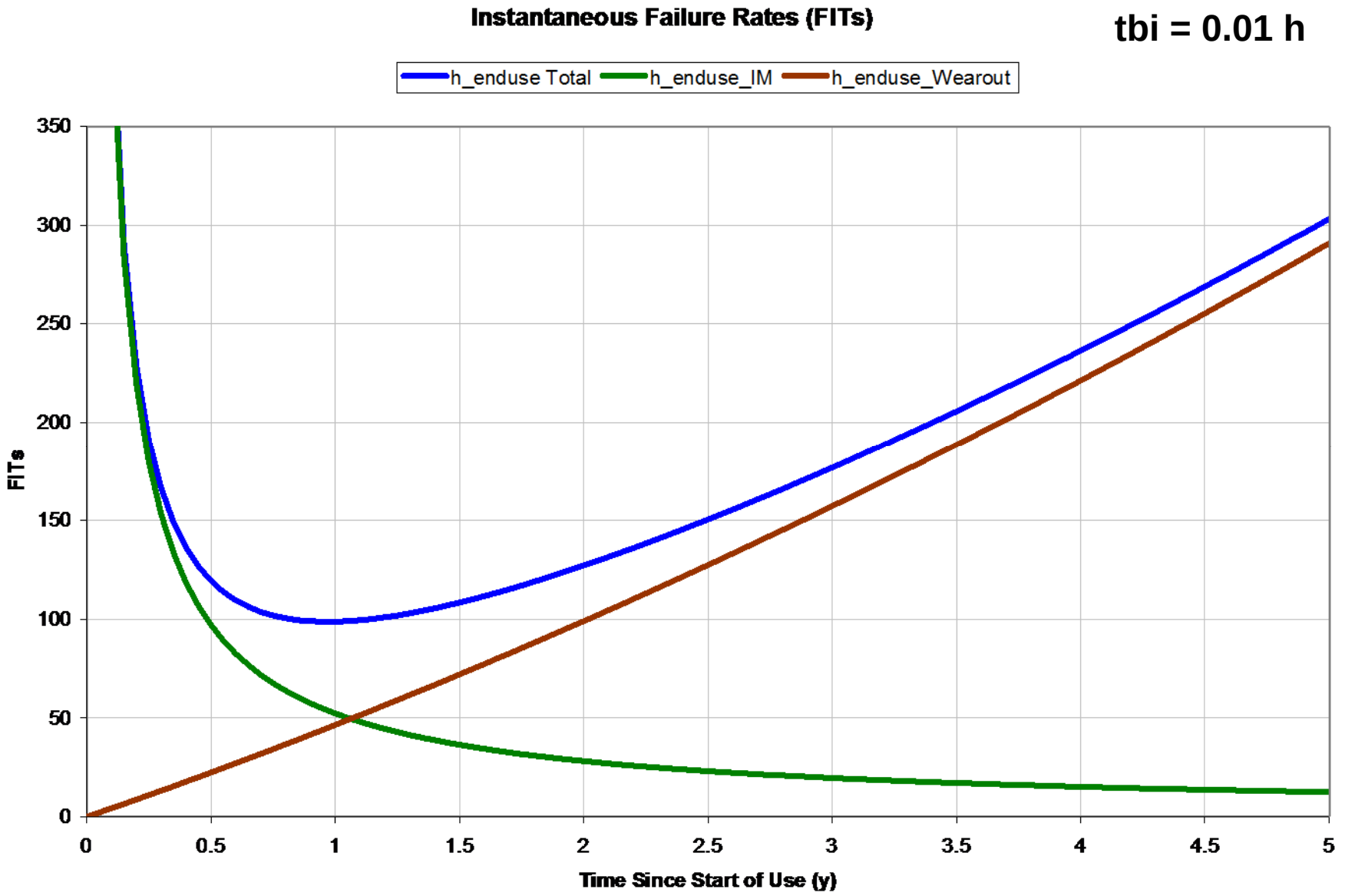
- The customer-observed fraction surviving burn in plus "use", is:

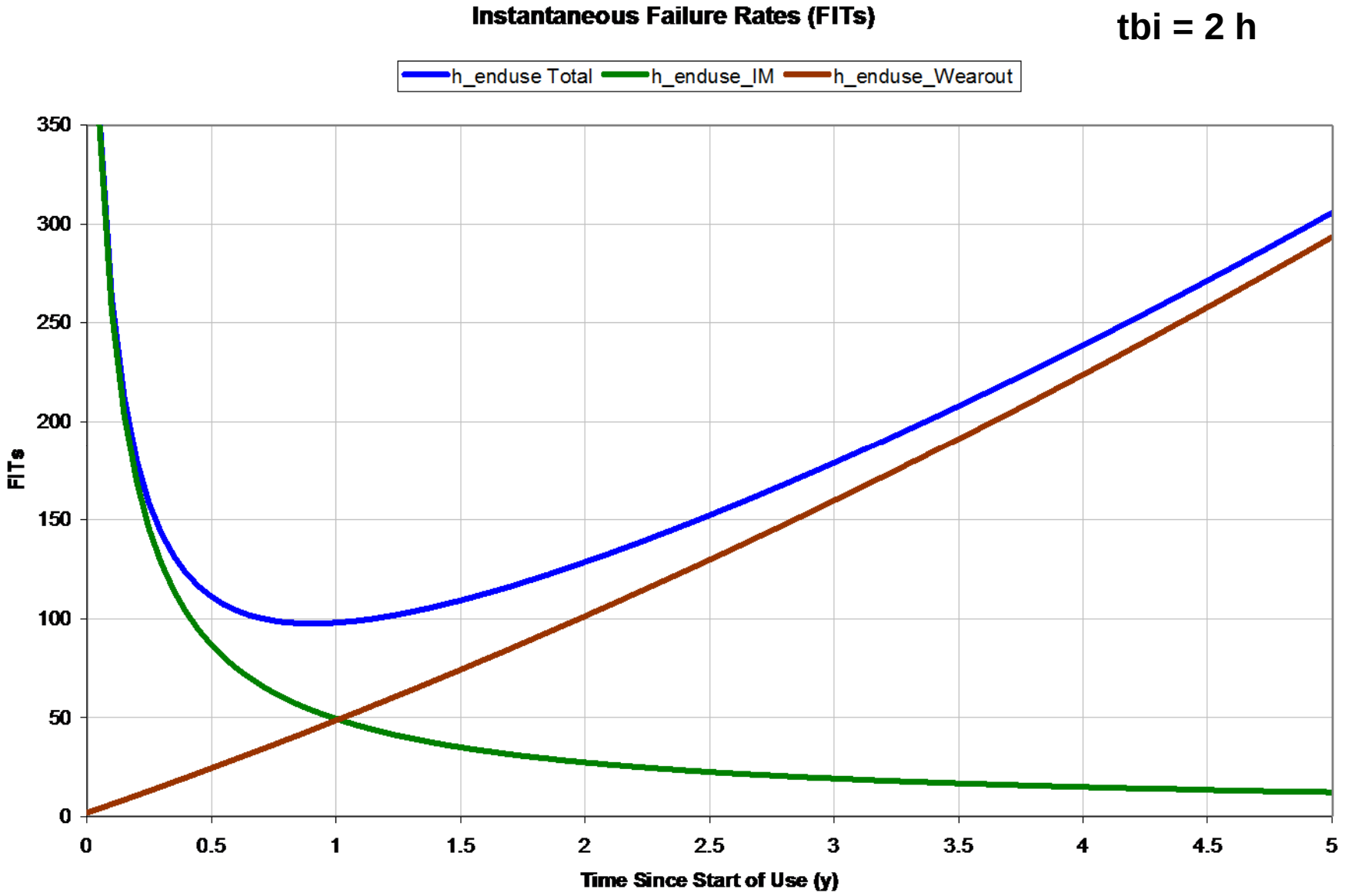
$$U = \left[\frac{1 + \frac{f \times r}{n} A_{total}(D + D_{use})}{1 + \frac{f \times r}{n} A_{total}(D + D_{bi})} \right]^{n} \times \exp\left[- A_{total} \times (D_{use} - D_{bi}) \right]$$

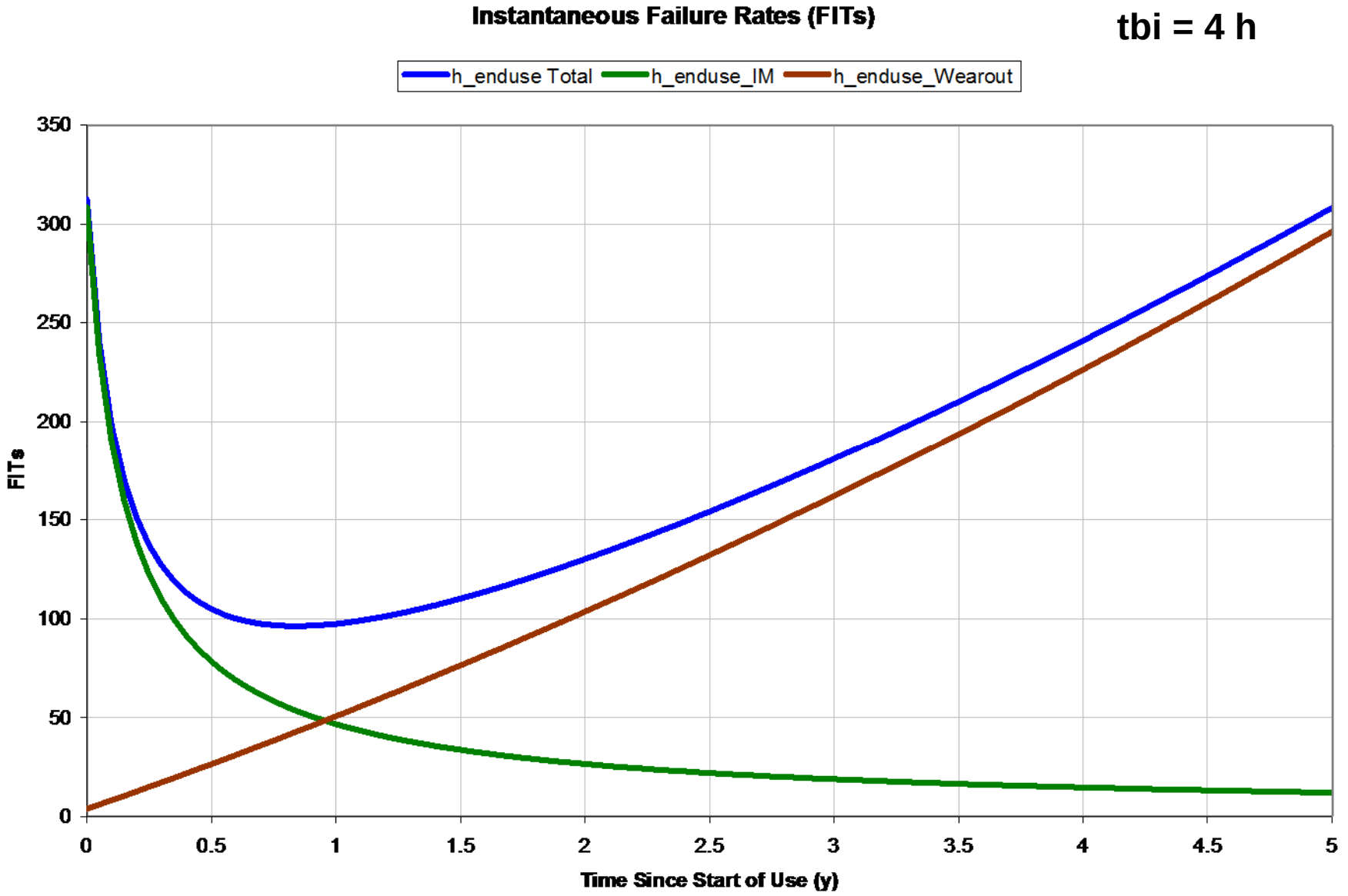
  where Poisson probability functions in terms of defect density were used.

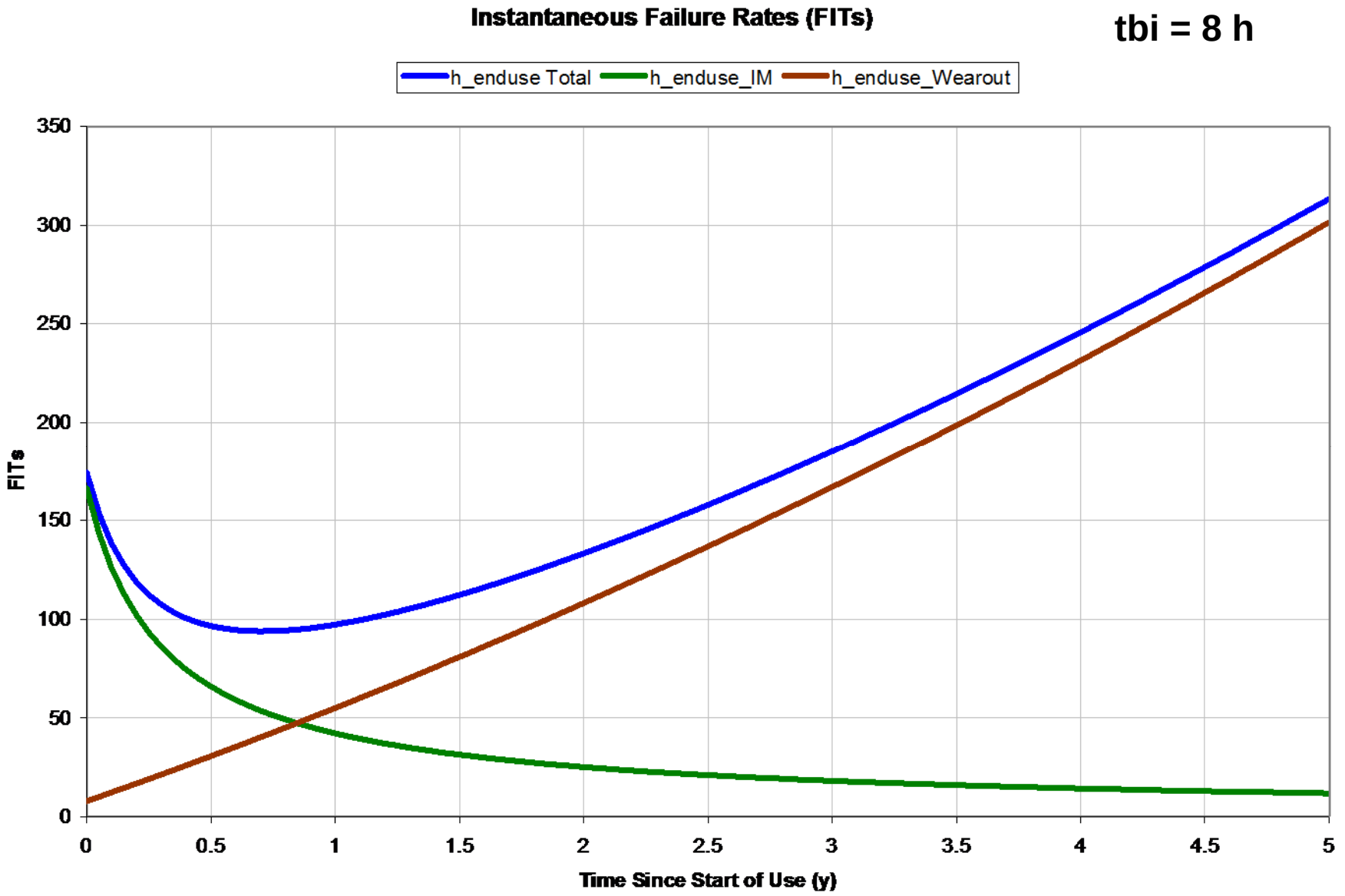
- So Infant Mortality DPM after t_{use} (= 720 h/30 d) and after t_{bi} of burn in is

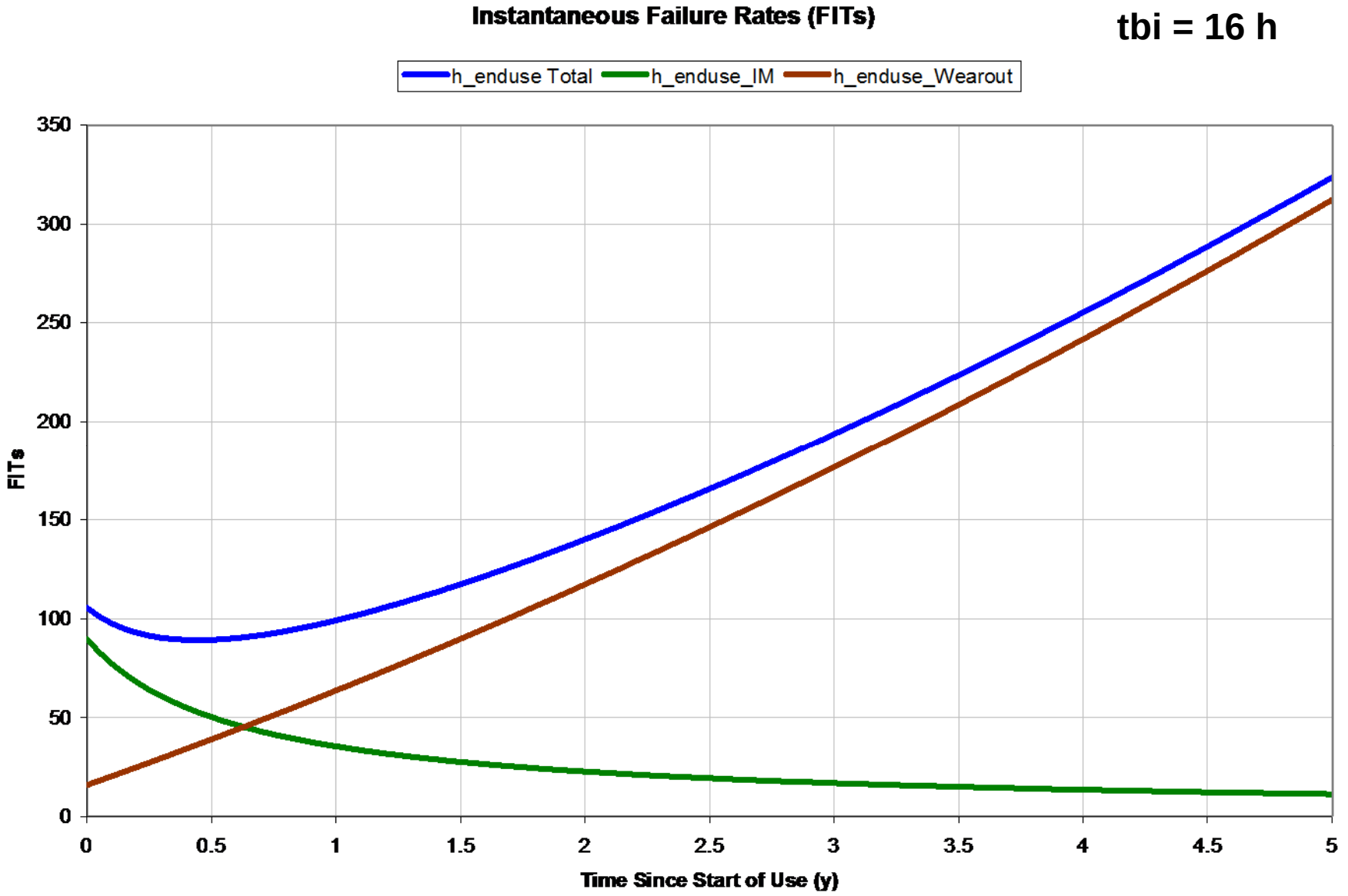
$$\text{Infant Mortality DPM} = 10^6 \text{ x } (1 - U)$$


Instantaneous Failure Rates (FITs)
tbi = 0.01 h
h_enduse Total
h_enduse_IM
h_enduse_Wearout
FITs
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time Since Start of Use (y)


Instantaneous Failure Rates (FITs)
tbi = 2 h
h_enduse Total
h_enduse_IM
h_enduse_Wearout
FITs
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time Since Start of Use (y)


Instantaneous Failure Rates (FITs)
tbi = 4 h
h_enduse Total
h_enduse_IM
h_enduse_Wearout
FITs
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time Since Start of Use (y)


Instantaneous Failure Rates (FITs)
tbi = 8 h
h_enduse Total
h_enduse_IM
h_enduse_Wearout
FITs
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time Since Start of Use (y)


Instantaneous Failure Rates (FITs)
tbi = 16 h
h_enduse Total
h_enduse_IM
h_enduse_Wearout
FITs
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time Since Start of Use (y)