Semiconductor Testing with AI/ML Application

Yi-Shing Chang
Principal Engineer, GEMS, Intel Corp;
Adjunct Prof, GSAT, NTU

Agenda

From Manufacturing to Testing

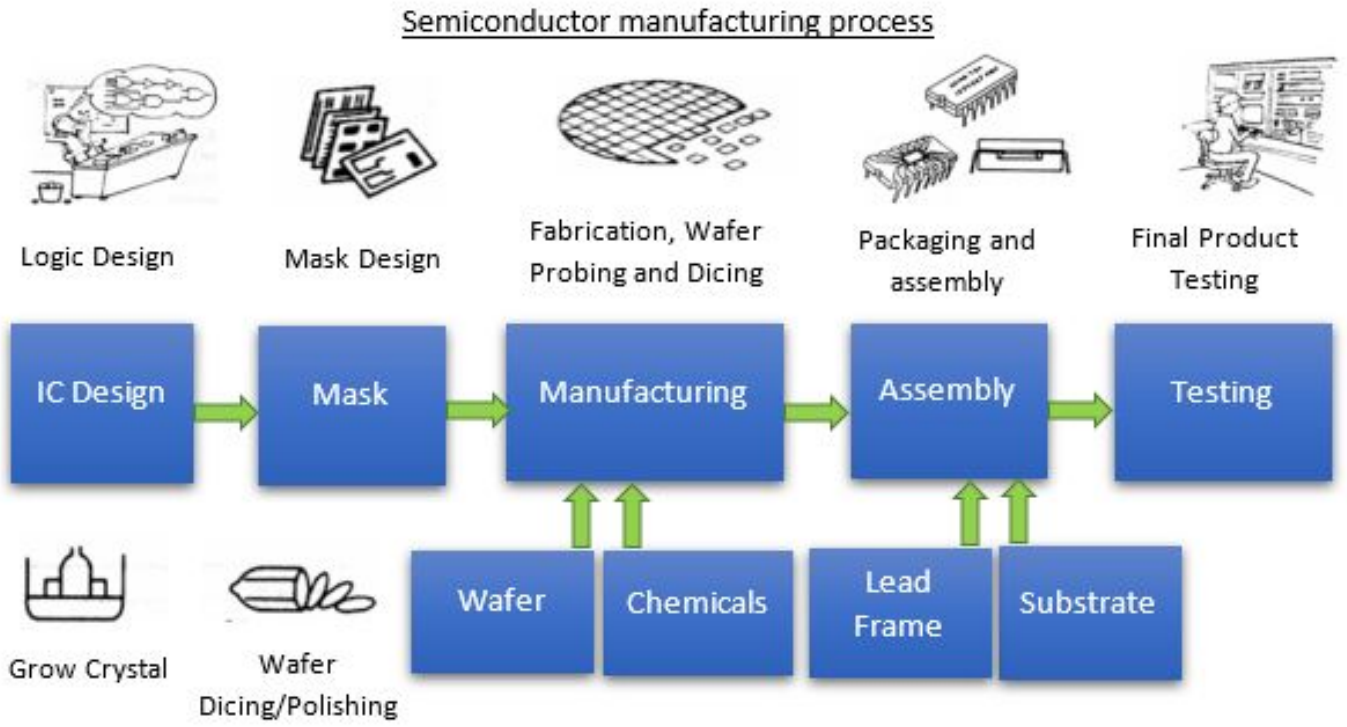
Testing Flow Overview

Test Method Overview

AI/ML Application

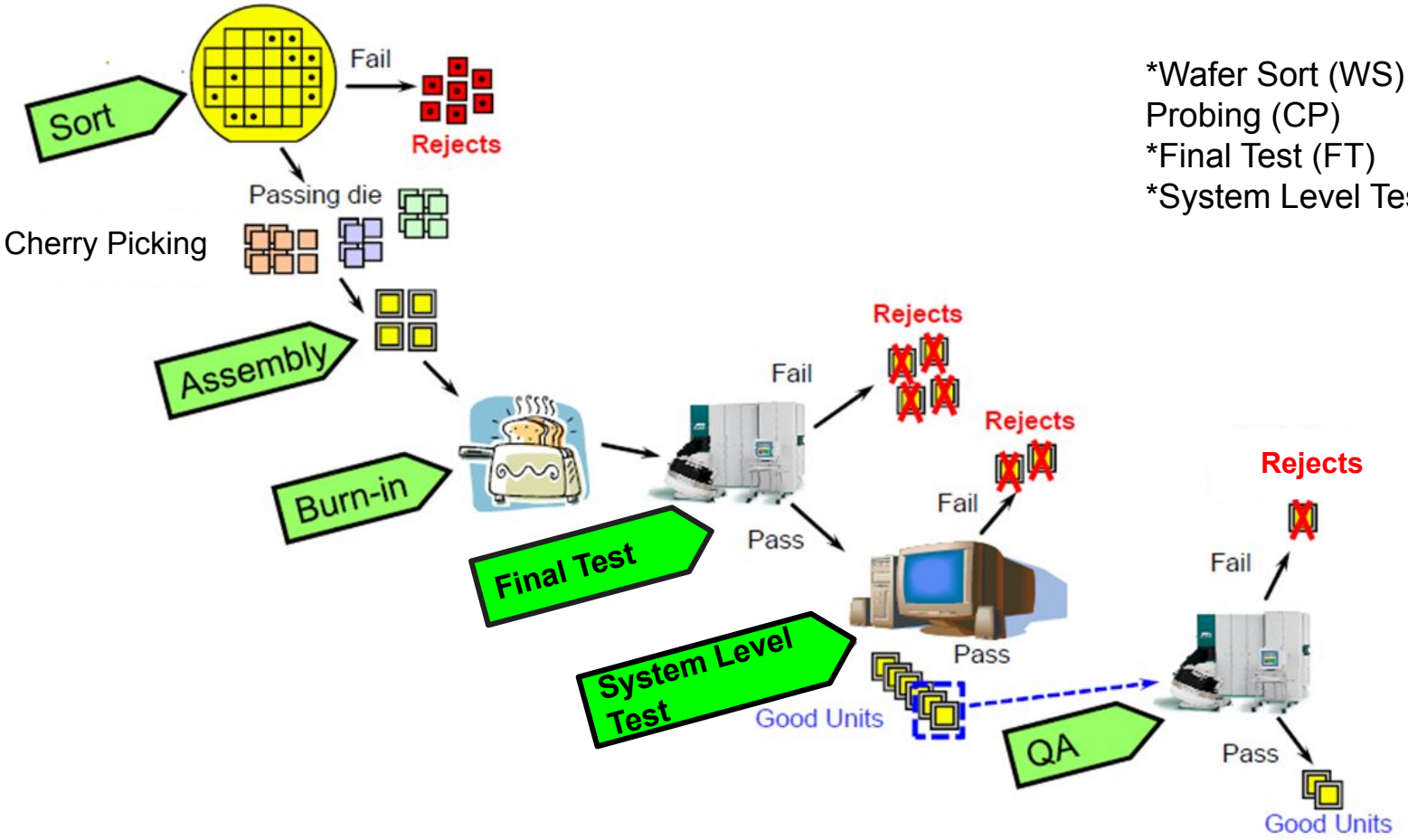
Summary

From Manufacturing to Testing



Source: Televisory's Research, Industrial Economics & Knowledge Center of Technology Research Institute (IEK/TRI)

Testing Flow Overview

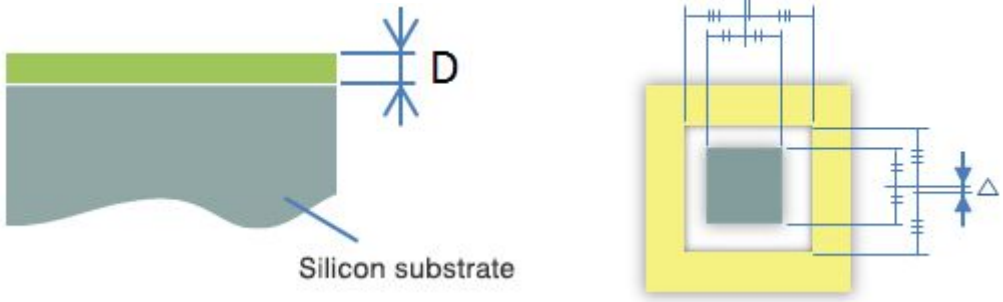


*Wafer Sort (WS) or Chip Probing (CP)
*Final Test (FT)
*System Level Test (SLT)

Before Fab-out

Metrology: a method of measuring numbers and volumes, mainly by using metrology equipment

- Not only to an act of measurement itself but to measurement performed by factoring in errors and accuracy, as well as the performance and mechanisms of metrology equipment
- Usually sampling



https://www.hitachi-hightech.com/global/en/knowledge/semiconductor/room/manufacturing/metrology-inspection.html

Before Fab-out

Wafer inspection

- It is a process for detecting any particles or defects in a wafer
- One of the causes of defects is the adhesion of dust or particles
- The inline findings may or may not have impact or become the “final” defects
- Many inline defects can be prevented to become real defects by using design rules or redundant Vias

Wafer Acceptance Testing

Wafer Acceptance Testing (WAT), also known as Process Control Monitoring (PCM) data, is data collected by the fab at the middle/end of manufacturing and generally made available to the fabless customer for every wafer

Scribe line is an area in a silicon wafer which is used to separate individual die at the end of wafer processing

Multiple test structures (e.g., ring oscillators,

leakage array) deployed in scribe lines

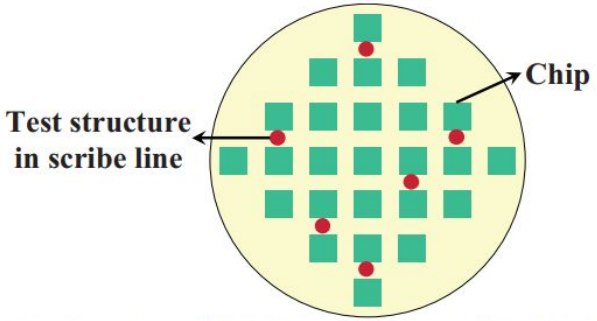


Fig 4. Test structures are deployed in wafer scribe lines to measure and characterize inter-die variations at wafer level.

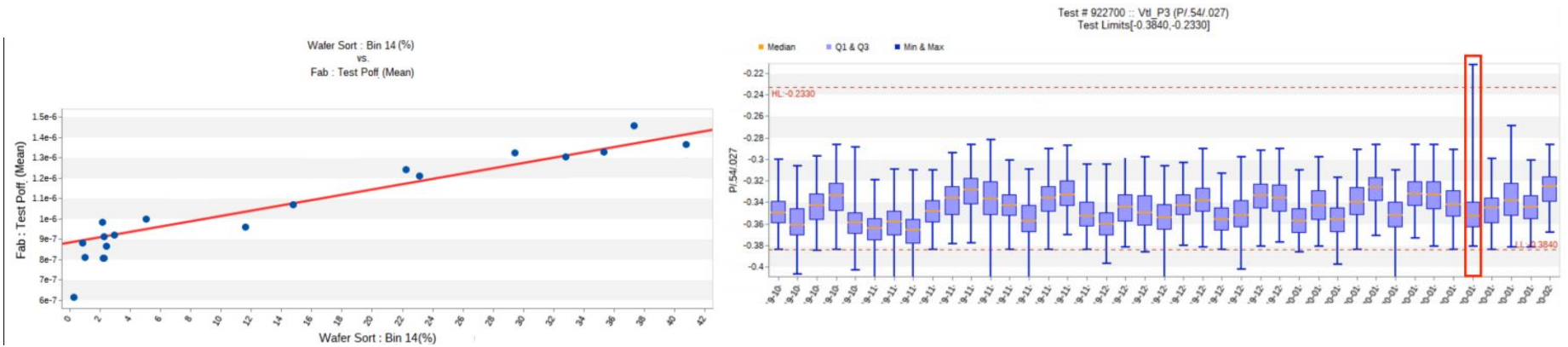
Xin Li, Rob R. Rutenbar and Ronald D. Blanton, “Cost Silicon Characterization of Nanoscale Integrated Circuits”, ICCAD 2009, November 2–5, 2009, San Jose, California, USA

How WAT is Used

Fab used WAT data to monitor the quality of wafers and manufacturing processes

- Skew wafers (making N/P devices faster or slower)

Foundry engineers use WAT to correlate the performance/leakage of the in-die ring oscillator (RO) built with different device type and monitor the variance



*https://semiengineering.com/whats-wat-testing-at-the-end-of-manufacturing/

Wafer Sort

Why is Wafer Sort Important?

- Screen as many defective die as possible to avoid Package Cost (known good die)
- Providing feedback to Fab on defects for yield improvement

Difference between Wafer Sort and Packaging Chip Test

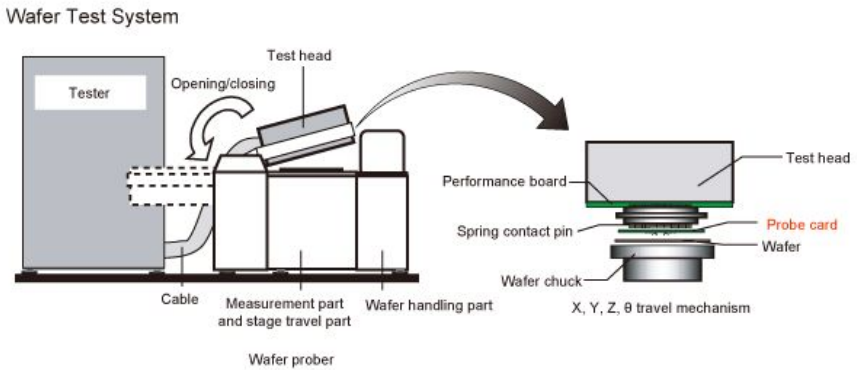
- Wafer need to be bumped vs. Chip has package with pins
- Formfactor, Thermal control, interface (probe card vs loadboard)
- Test conditions
- Mechanical Kits & handler
- Pattern types e.g. Functional and high speed IO tests may not be available for WS

Sort Test

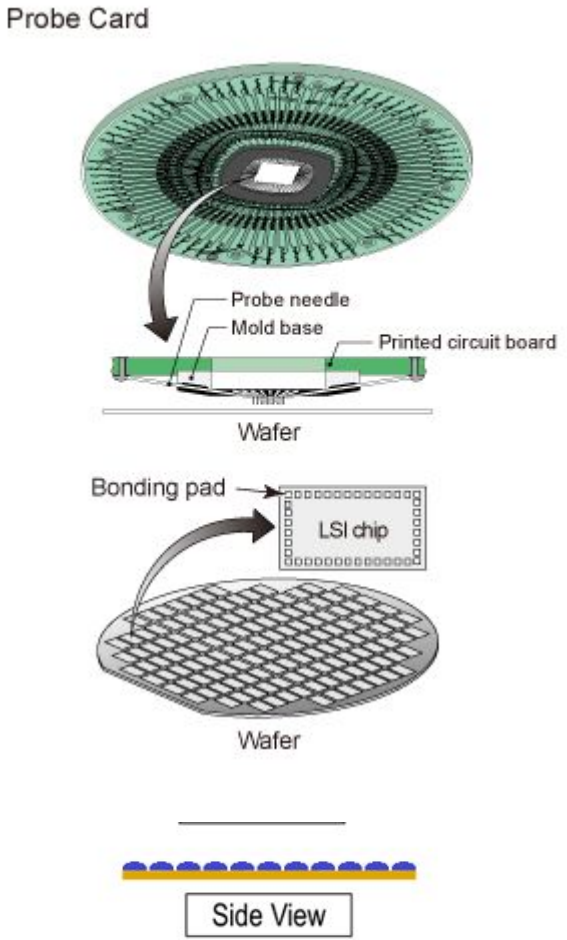
Wafer Sort Test is completed while die are still in wafer form

Each die has 1-20K's of pads/bumps that are used for power, ground, and I/O signal connections for Sort Test

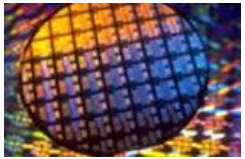
A probe card is used to make the connection between the wafer and tester

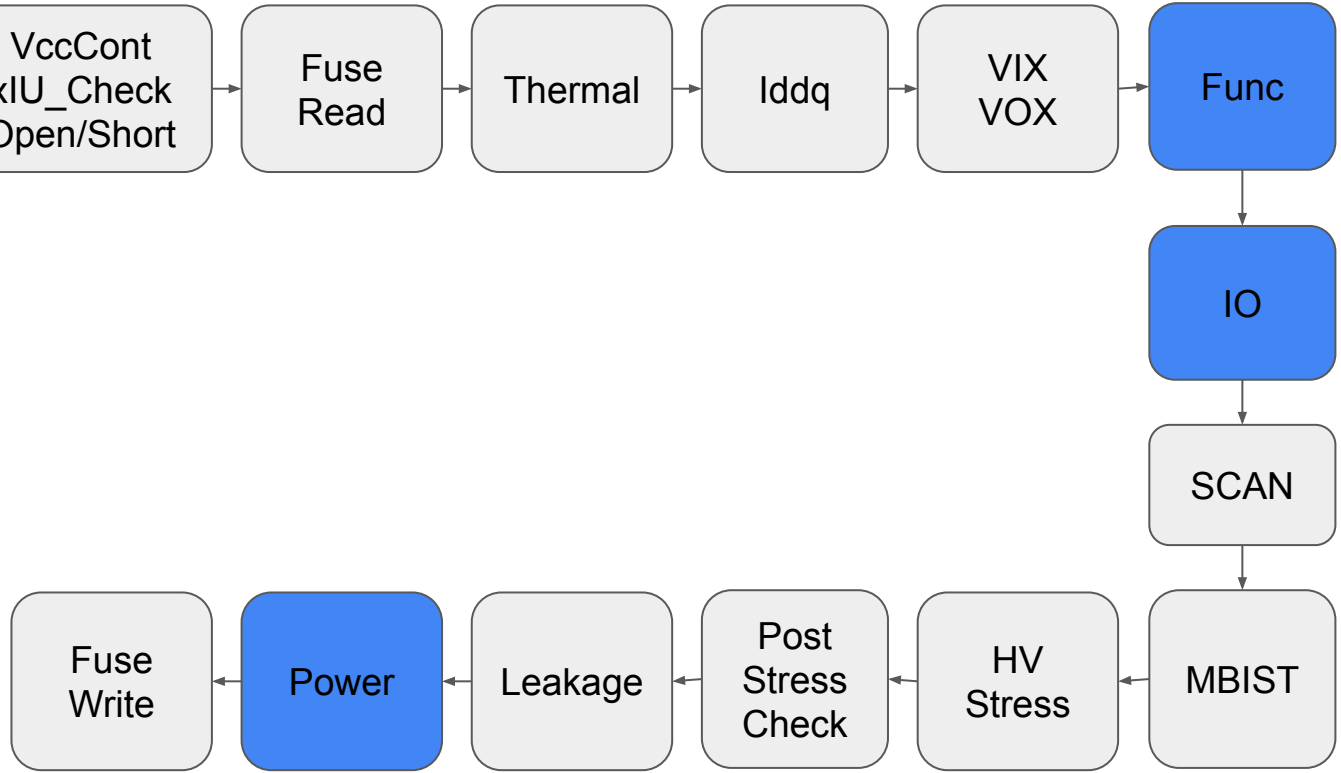


https://www.mjc.co.jp/en/technology/column/probe_card.html



Test Flow High-Level Overview





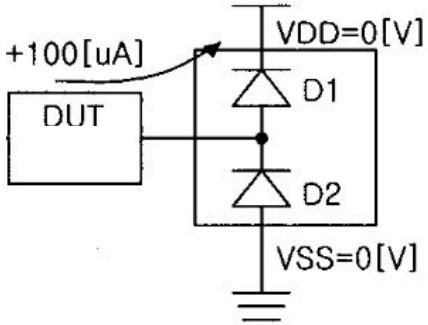
Test Program

A software developed based on tester OS and APIs which can be loaded into tester memory and executed to test wafer/chip automatically

The test program usually consists of many test modules of specific test methods such as DC characterization, SCAN, MBIST, and Functional test

The order of these test modules would be adjusted based on test time or product types

DC Characterization



DC measurement: VSIM/ISVM/VSVM(seldom used) against the limits

V:voltage, I:current, S:supply, M: Measure

VSIM(ISVM): Apply voltage(current), turn on the relay, then measure current(voltage), i.e. resistance(impedance) measurement

Vcc Continuity: Measure the resistance for each power domain to make sure the die/chip would be powered on properly using tester digital power supply (DPS) first . Modern chip/SOC can have many power domains. This test is to make sure no short among power domains

xIU_Check (Sort/Chip Interface Check): Check if tester channels/probes make proper connections to the die bumps or pins to prevent bad dice from damaging probe card or tester HW

Short/Open: Check the IO pins impedance by biasing the upper diode or the lower diode of pin electronics

Fuse Read

Data traceability is a must for testing. Without traceability, the test data collection is of very little use

Modern SOC contains fuse or OTP for unique ID, product configuration, and SKU management

Before starting real testing, make sure to read the ID, set the configuration, and record them in the datalog

Thermal

Modern SOC usually contains 1-2 thermal diodes, and many digital thermal sensors for die junction temperature monitoring and thermal protection to avoid overheat/burn out

Thermal diode usually design using BJT which has the following characteristic function and need to be calibrated. The calibration procedure is required measuring voltage values at two temperature setting.

$$\Delta V_{BE} = n_f \frac{KT}{q} ln(N) \quad (1)$$

where

ΔV_{BE} is the change in junction voltage when the diode is operated for to different current values.

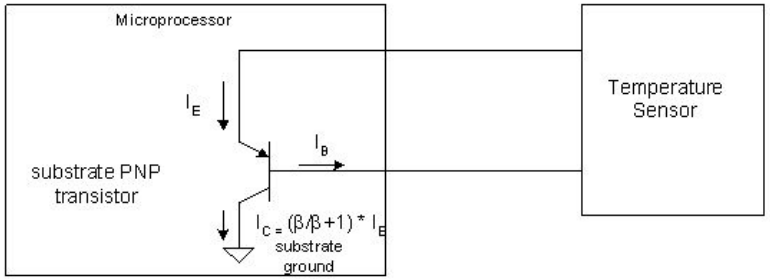
n_f is the ideality factor of diode

K is Boltzmann's constant

T is the absolute temperature in Kelvin

q is the electron charge and

N is the ratio of currents

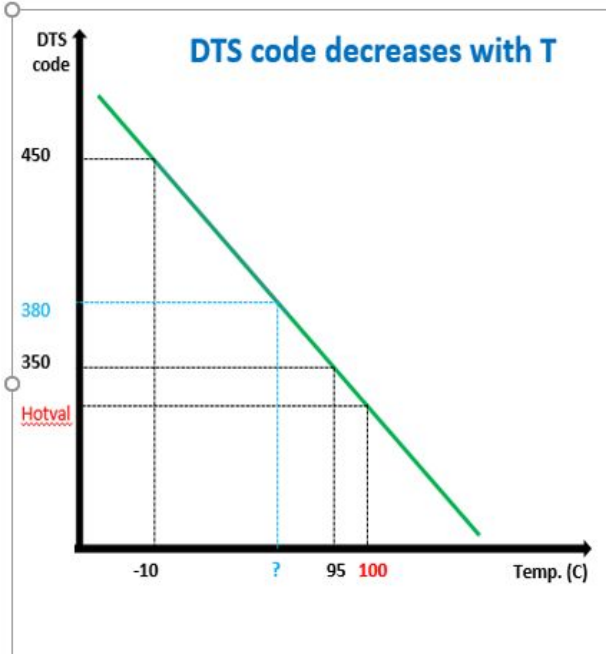


Thermal – Cont’d

The digital thermal sensor usually consists of thermistors and ADC

Trimming will be needed for each sensor

Many IC parametric and performance data are temperature dependent. Taking the measurement data without recording the measurement temperature would be troublesome or not useful



Iddq Test

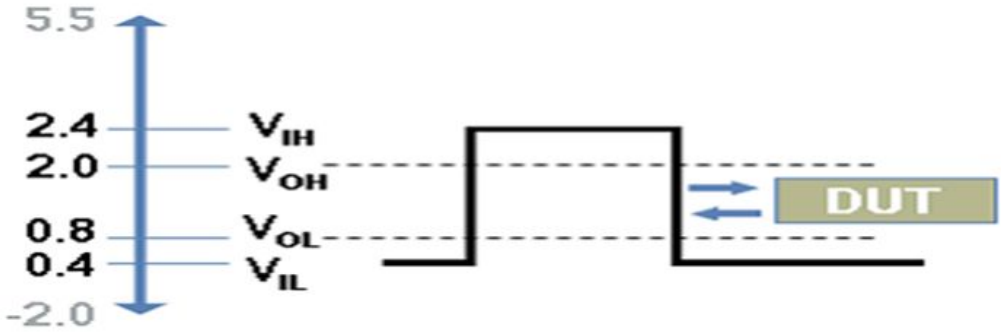
Iddq or static leakage current is an important parameter for process monitoring and defect screening

Multiple sources of leakage

- Subthreshold conduction
- Junction leakage
- Gate leakage

This test method usually required to turn the clocks on first, kill the clocks, set the triggering timing to take the static leakage current for each Vcc domain for consistency.

VIX and VOX



VIH / VIL (input high voltage and low voltage)

- Threshold voltage for digital inputs
- Force levels as a go-nogo test in HVM

VOH / VOL

- VOH is the minimum output voltage in the high state
- VOL is the maximum output voltage in the low state
- Tested using a go-nogo test in HVM

Func

The functional test is a basic test template for tester. It consists of setting pins/levels/timing, running a pattern list, and capturing the failing signature as failing cycles and failing pins

Because wafer does not have good I/O pins, most functional tests cannot be executed unless DFT circuits such as BIST are inserted

The alternative way to run functional tests in wafer sort is loading the assembly code to the cache through scan chain, execute the code, capture the result in the cache, shift the result out from scan chain, and then compare TDO response to determine the pass/fail

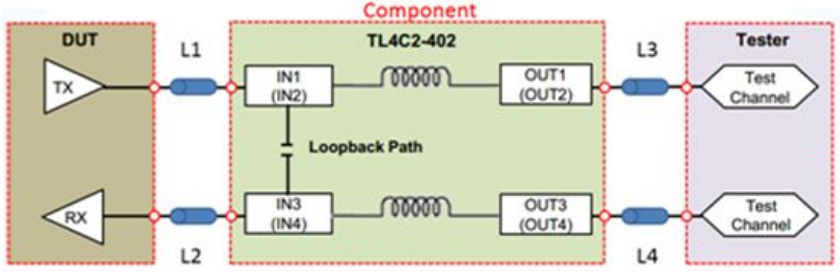
IO

IO tests are done with loop-back, either in-die loopback (DFT) or external loopback

For Chip Test, the component can be load board+substrate

Transmit to Receive through AC/DC connected space transformer used in Sort

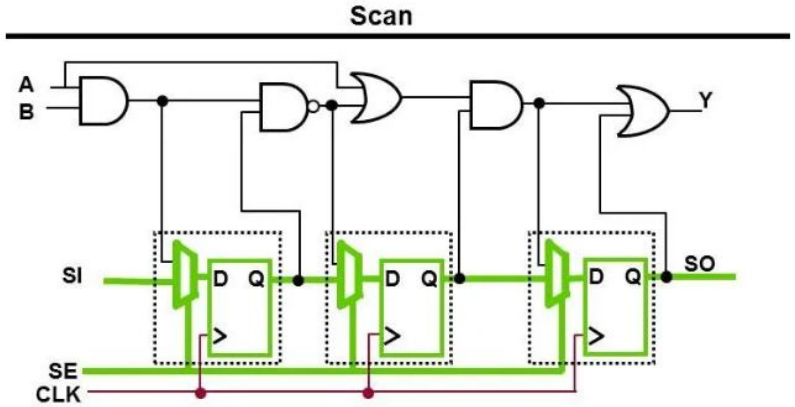
Eye margin mode for Eye height and Eye width to test Receiver



SCAN

The design’s flip-flops are modified (Design For Test or DFT)to allow them to function as stimulus and observation points, or “scan cells” during test, while performing their intended functional role during normal operation.

The modified flip-flops, or scan cells, allow the overall design to be viewed as many small segments of combinational logic that can be more easily tested, debug, and diagnosis.



SCAN Tests

The normal SCAN flow usually includes chain test, s@ ATPG, then transition ATPG to screen speed related defects

Once one of these tests failed, it will go to fail flow to collect failure data log for diagnosis

For compressed s@ pattern failure, bypass or one-hot patterns which only allow particular scan channel's response to be captured can be generated to improve the diagnosis resolution (remove the aliasing)

Layout Aware diagnosis is a requirement for diagnosis & pFA for advanced process nodes

Memory Test

Most memory design using memory compiler with programmable BIST engine, Redundancy, Built-In Self Repair (BISR), and Built-In Repair Analysis (BIRA)

Foundries will provide the redundancy requirements for each process node if design contains > certain size of memory to guarantee its yield

Various BIST sequences which can detect different memory failing mechanisms are required to be included in the BIST engine

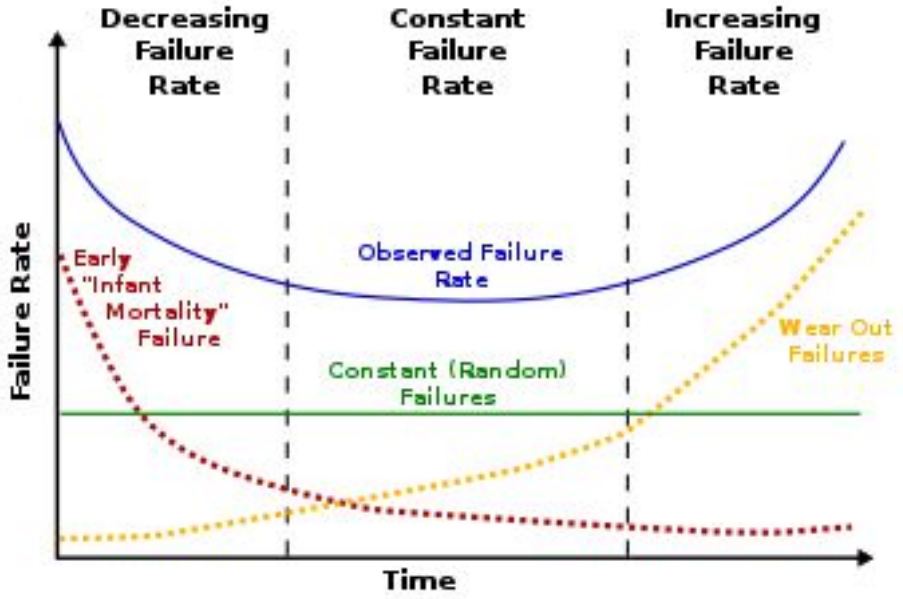
The memory compiler will create the controller circuits allowing to dump out the failure locations for diagnosis

Why Burn-In

The bathtub curve is widely used in reliability engineering.

- The first part is a decreasing failure rate, known as early failures “Infant Mortality”
- The second part is a constant failure rate, known as random failures
- The third part is an increasing failure rate, known as wear-out failures

The purposes of Burn-In Test are to (1) screen out Infant Mortality failure, and (2) accelerate the product wear-out for “guaranteed” product life or quality



https://en.wikipedia.org/wiki/Bathtub_curve

Traditional Burn-In Tests

The chips, PCBs, or semiconductor devices are tested under elevated temperature, voltage, and power cycling conditions. The test accelerates the appearance of the latent defects in the device by forcing it to undergo failure conditions under supervision

The burn-in test is an estimation method to find out the useful life of the semiconductor device

- The total lifespan of the device is shortened with a burn-in test.
- May take too long (hours or days)
- Sampling

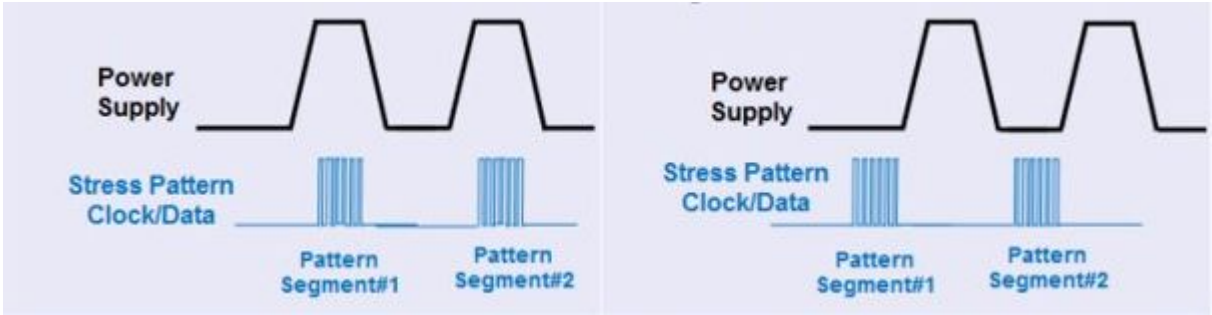
High Voltage Screen

The test method to reduce burn-in time while maintaining product quality & reliability. However, it may not remove traistional burn-in

Dynamic voltage stress (DVS): Foundry preferred method since the circuits are stressed during transition. Might suffer larger than expected failure due to timing closer or high leakage

Enhanced voltage stress or Electrical voltage stress (EVS): Better for higher static stress capability on screen, but less toggle converge

Quasi-Static Voltage Stress (QSVS): Multiple EVS to increase toggle coverage and may alternate stress and nominal voltages when finishing loading patterns



G. -G. Oh, M. -H. Ho, Y. -J. Shin, J. -W. Choi, J. -Y. Kim and Y. -D. Kim, "Dynamic Voltage Stress Sensing Circuits for Screening Out Early Device Reliability Issues in Advanced Technology Nodes," *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2021, pp. 1-3

Post Stress Check

High Voltage Stress is to minimize the burn-in requirements for product qualification and reliability with short test time.

During the stress, the DUT was not actually checked fully because design’s timing closure is not at the same high voltage level

The check was usually done after the stress either by comparison of Vmin delta before and after stress or running at-speed contents after stress

Leakage Test

The purpose of Leakage testing is to screen out unwanted current paths on high impedance I/O pins to VCC and ground

The primary requirement for Leakage test is the ability to tri-state the output buffers and force voltage onto the pins to measure the resultant current

In order to test Leakage, the device needs to be in a completely powered down state; all clock toggling is disabled, all DC paths are closed to enable lowest current measurements

It is usually done after high voltage stress

Power Test

Power envelope is the critical parameter for a CPU product and it sells

$P=Iddq*V+k*Cdyn*V^2*f$, where Cdyn is the dynamic capacitance when the chip operates at the frequency f

Cdyn is sensitive to Vcc, to Fmax, to Tj and also to resource types

It requires to have good “power virus” patterns to be have repeatable measurements of Cydn and Power

Fuse Write

A unique ID including wafer Lot, number, x- and y-coordinates with encoding would be written in the fuse for traceability

Repair info for array and memory are written

Binning and configuration based on test results would be fused

Offline fusing would be done considering future demand

AI/ML Application – Data Feedward Forward and Backward

It is very expensive and time consuming to raise temperature up & down during testing of a DUT -> test entire wafer or lot at the same temperature and then move the DUTs to another temperature

E.g. Thermal diode and thermal sensors required at least two temperature measurement for calibration/trimming-> measurement the reading at one temperature and then forward the data to another temperature measurement and complete the formula and use the formula to determine the temperature during test

The formulae might not be linear and required volume characterization data to model them -> machine learning

Data Feedback Backward: e.g. Monitoring test patterns vs defect detection -> if there is no any defects detected on said 1 million DUTs-> the pattern can be removed for TTR

AI/ML Application

Many AI/ML applications were used for test time reduction -> save the test cost

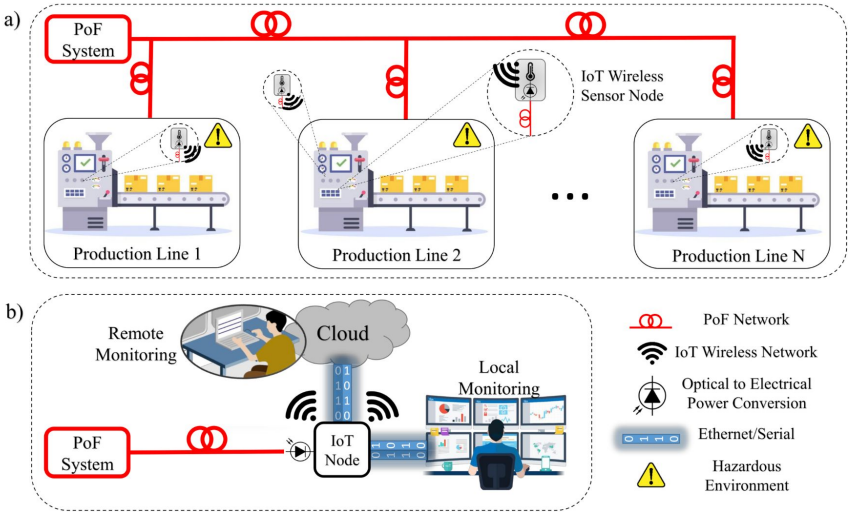
- Could data collected at Cold (Hot) temperature used to predict the measurement results at Hot (Cold) temperature, so test at Hot (Cold) temperature can be skipped?
- Or, only need to minimize the number of DUTs of False Positive and False Negative
  - A false positive is when something is true when it is actually false (also called a **type I** error). A false positive is a “false alarm.” -> yield loss
  - A false negative is saying something is false when it is actually true (also called a **type II** error). A false negative means something that is there was not detected; something was missed. -> test escape (Bad)
- Or, can re-measure these DUTs only -> Challenges
- Machine learning model build* (training data set v.s. test data set)
- Domain knowledge base (Measure a few parameters to calculate complex parameters with characterization data set)

“Machine learning application for silicon photonics transceiver testing,” Woosung Kim, Yeoh Hoe Seng, Yi-Shing Chang, Suohai Mei; SPIE 2018, San Diego, California, United States

AI/ML Application

Sensor networks (IOT) stream lots data real time to monitor the test and assembly lines

- Use unsupervised learning to identify patterns if any-> predictive analytics
- Build ML (CNN/DNN) model for Abnormality detection
- + labeling and reinforce learning



*Optically-Powered Wireless Sensor Nodes towards Industrial Internet of Things by Letícia C. Souza,et. Al. ORCID Laboratory WOCA, National Institute of Telecommunications, Brazil

Summary

This lecture covers an overview of semiconductor testing including wafer sort, burn-in, and packaging chip testing

The test flow and test methods are discussed

AI/ML learning applications discussed

It should help you understand how the modern CPUs are tested