Improvement of the volume ramp up process

Antoine Mauduit

Ecole d'ingénieurs eicnam

Improvement of the Volume Ramp Up process

Dossier de synthèse présenté par

Antoine, Serge, Pierre MAUDUIT

Pour obtenir

Le Diplôme d'ingénieur du CNAM Spécialité mesure-analyse, parcours Instrumentation qualité

Travaux présentés le 21 juin 2016 devant le jury composé de

M.	Patrick Juncar	Cnam PU Président
Mme	Annick Razet	Cnam PU
M.	Mark Plimmer	MCF , Cnam
Mme	Joanne Schel	NXP Semiconductor
M.	Giacomo Pecoraro	NXP Semiconductor

Résumé

La mise en production d'un circuit semi-conducteur est toujours une phase délicate. En effet, malgré toutes les étapes visant à qualifier le produit et les outils de test, le manque d'expérience sur le produit fait que les premières livraisons se font dans des conditions difficiles. C'est la capabilité des procédés de fabrication qui sont souvent en cause. Cela a pour conséquence que lorsque qu'une variation se produit sur un élément de la chaîne de fabrication, l'impact sur les performances peut s'avérer critique. Du fait que les rendements ne sont pas prévisibles, cela peut mener à des retards de livraisons. Ce qu'il faut savoir c'est que les procédés composant la chaîne de fabrication allant de la FAB à la livraison sont nombreux et les problèmes qui leur sont associés très spécifiques. Il est ainsi nécessaire d'optimiser au mieux la capabilité des procédés lors la phase précédant la mise en production du circuit afin de minimiser les risques sur les délais de livraisons par la suite. Typiquement plus tôt les points faibles sont révélés, plus élevées sont les chances de les résoudre avant la mise en production. Mieux exploiter cette phase, c'est aussi minimiser l'effort de l'équipe Opérations pour amener les performances des produits à maturité. Pour y parvenir, une approche structurée est primordiale.

Abstract

The release to production of a semiconductor device is a tough phase. Despite all stages of validation to qualify the product and its tools of test, the lack of experience about the product means that initial deliveries are made in difficult conditions. These difficulties seem to be mainly due to any process having a capability too limited. This has as consequence that if any slight variation on the production flow occurs, the impact on performances can be critical. Due to unpredictable production yields, this might delay deliveries. It is important to consider that many processes make up the whole supply chain and that the problems linked to each process is very specific. It is then necessary to optimize as much as possible the capability of each process during the phase before the release of the product to production. The goal is to minimize the risks of late deliveries. Typically, the sooner the limitations of the process are identified, the higher the chances to solve them before the release to production. By exploiting this phase better, one is also minimize the effort of Operations to bring performances of the product to their maturity. To achieve the goal, a structured approach is mandatory.

Mots Clés

Manufacture de semi-conducteurs, Industrialisation, Profitabilité, Phase de mise en production, Rendement, Satisfaction client

Key words

Semiconductor Manufacture, Industrialization, Profitability, Quality, Ramp Up Phase, Capability, Manufacturing, Yield, Customer satisfaction

Content

1. Introduction

NXP is a worldwide semi-conductors manufacturing company. Among its many sectors, NXP is the worldwide leader as an automotive supplier. This market has first of all the specificities not to tolerate any reject at the customer side. Its other specificity is that between the moment the product is released to production and its ramp up phase it might take 2 years. The problem encountered is that despite volumes remain low, delivery dates are hardly respected. The first reason is that orders are issued in a sporadic way. This means the production flow is restarted frequently which represents a risk. Indeed converting a production flow from a product to another requires some fine tuning until it reaches its normal performance. In addition, despite the fact that before its release to production the product passed any acceptance gates, it is not sufficient to figure out how the product will be sensitive to variations on manufacturing processes. It is to improve the Volume Ramp Up process that this project was undertaken with the aim of looking for solutions to fulfill deliveries during the ramp up phase.

In a first instance, the context will be introduced with a presentation of the problematic and also of the company. Afterwards, the processes composing the production flow will be explained with their associated difficulties. Then before we discuss the solutions applied, the issues will be quantified to ensure the solutions have properly addressed them. The final part gives the direction to be followed to reach the manufacturing excellence.

Before going into details with the detailing of processes composing the production flow, we explain the context of the problem. Then, still related to the context, a presentation of the company in which I work will be given.

1.1 Introduction to the problematic

Continuity and knowledge transfer are critical features in any production process but particularly so in the semiconductor industry. During the lifetime of a product, the ramp up

phase start coincides with the period for which the product is about to be qualified. This phase coincides with the issuing of early orders for its prototypes by the customer. It is also when the team who developed the product is re-allocated to another project. The work is then handed over to the Operations team which is in charge of industrializing the product, optimizing production yields and handling customer returns

It is thus the Operations team who will ensure that those first orders are delivered on time with the required level of quantity. The Ramp Up manager is in charge of ensuring correct progress during this phase. Briefly speaking, this phase will last until production yields reach both the required stability (i.e. they become predictable) and a level ensuring the profitability of the product according to the business plan. The reason why this phase requires a dedicated team is because it is such a delicate period. Despite lessons learnt during a pre-production, despite the product and its test process being qualified and released, this does not prevent the start of production from being chaotic. Several parameters contribute to this situation. The main ones are as follows.

1. For product qualification, data from three different sample lots are necessary.

This quantity of sample lots enables one to assess the reproducibility of the fabrication process. One difficulty faced during the ramp up phase comes from the fact that, statistically, three batches are not enough for one to detect all drifts faced during the whole product lifetime. Therefore certain performance issues only appear later on once the product has been released.

This aspect and also the ones listed below will not be detailed and analysed right now but left for later. In another section some solutions applied to minimize impact of the problem listed here will be also described.

2. Test process maturity

The problematic linked to the maturity of the test process is similar to the one described above. For product release, test process qualification is performed on a minimal number of test machines using a limited number of sample lots. Test process qualification is only

feasible at the end of the development phase for planning reasons. Upon qualification of the product, the test process is qualified but not totally mature. Despite acceptance criteria, when early orders arrive, the production often reveals unexpected yield losses frequently due to test methods or test limits optimization needs. Until this is solved, the yield remains unpredictable, a fact that might jeopardize future deliveries.

Another contributor to the difficulties faced during the Volume Ramp Up (VRU) phase is the low volume production with a discontinuous load. The fact that a production line is not continuously loaded leads to some interruptions of the production flow. Each production start requires a calibration phase with potential fine tuning, which causes yield losses or longer throughput time. During a ramp up phase, volumes are indeed low and irregular. The fact these quantities are low means it also takes longer to validate processes improvements. To give an idea, it takes roughly four months from the time a new sample lot to when its final performances are known.

3. One also has to consider a problematic specific to the automotive market, namely that it might take a couple of years until the volume ramp up really starts. During this phase, orders remain sporadic. For all the reasons listed previously, having a product with predictable test performances compliant with the datasheet represents a real challenge. The automotive market has also the particularity that the allowed reject rate on the customer side is nil.

4. To this list is to be added a resource constraint. Indeed, at the validation of the product the development team is no longer allocated to the product or else only to provide a very limited support. However it is this team that holds the key knowledge. This highlights the real need to invest effort in the ramp-up team during the development phase in order to benefit from existing resources and anticipate a maximum of potential issues during the ramp up phase.

For all the aforementioned reasons, this phase is delicate to handle as, on the one hand, customers expect deliveries on time with the quality level guaranteed while on the other, test and fabrication processes have not yet matured. The stake is thus to capitalize on resource allocated to the development to tackle what is detected during the first measures of samples. To detect weaknesses a structured process is a key aspect. It is to develop and improve this process that I have worked on and which is developed in the present dissertation.

1.2 Presentation of the Company

NXP Semiconductors N.V. is a global semiconductor manufacturer headquartered in Eindhoven, The Netherlands [1]. The company employs approximately 45,000 people in more than 35 countries, including 11,200 engineers in 23 countries. NXP reported a revenue of $6.1 billion in 2015, including one month of revenue contribution from the recently acquired firm Freescale Semiconductor.

NXP is currently the fifth-largest global non-memory semiconductor supplier globally, and the leading semiconductor supplier for the Secure Identification, Automotive and Digital Networking industries. The company was founded in 1953, with manufacturing and development in Nijmegen, Netherlands. Known then as Philips Semiconductors, the company was sold to a consortium of private equity investors in 2006, at which point the company's name was changed to NXP.

On August 6, 2010, NXP completed its IPO, with shares trading on NASDAQ under the ticker symbol NXPI. On December 23, 2013, NXP Semiconductors was added to the NASDAQ 100.Finally, on March 2, 2015, it was announced that NXP Semiconductors would merge with chip designer and manufacturer Freescale Semiconductor in a $40 billion US-dollar deal. The merger was closed on December 7, 2015.

NXP Semiconductors provides mixed signal and standard product solutions based on its security, identification, automotive, networking, RF, analogue, and power management expertise. With an emphasis on security of the connected vehicle and the growing Internet

of Things, the company's products are used in a wide range of "smart" automotive, identification, wired and wireless infrastructure, lighting, industrial, consumer, mobile and computing applications.

Along with Sony, NXP is the co-inventor of near field communication (NFC) technology and supplies NFC chip sets that enable mobile phones to be used to pay for goods, and store and exchange data securely. NXP manufactures chips for eGovernment applications such as electronic passports, RFID tags and labels, as well as transport and access management, with the chip set and contactless card for MIFARE used by many major public transit systems worldwide. In addition, NXP manufactures automotive chips for in-vehicle networking, passive keyless entry and immobilization, and car radios. NXP invented the I²C interface over 30 years ago and is a supplier of I²C solutions. It is also a volume supplier of standard logic devices, and in March 2012 celebrated its 50 years in logic (via its history as both Signetics and Philips Semiconductors). NXP currently owns more than 9,000 issued or pending patents.

In Gratkorn Austria where I work, products are developed for car access and immobilization. For this reason, the site is certified ISO/TS 16949. The norm ISO/TS16949 [2] can be applied throughout the supply chain in the automotive industry. Certification takes place on the basis of the certification rules issued by the International Automotive Task Force (IATF). The certificate, valid for three years, must be confirmed annually (as a minimum) by an IATF certified auditor (3rd Party Auditor) of an IATF recognized certification body. Re-certification is required at the expiry of the three-year period. Certification pursuant to ISO/TS 16949 is intended to build up or reinforce the confidence of a (potential) customer towards the system and process quality of a (potential) supplier. Today, a supplier without a valid certificate has little chance of supplying a Tier 1 supplier and certainly no chance of providing a car manufacturer with standard parts.

2.Description of the processes

In a project, every step of transformation is the object of a document that mainly describes the status of the product before and after it. This is the principle of processes constituting the full chain of transformation of a product: from the FAB to the delivery to the customer. It provides a reference point for every operation carried out during the manufacture of the product. Notably, it describes the initial status, the goal of the operation, the description of the operation, the expected result, resources involved. Each process can be composed of sub-processes. This notion is applicable to the supply chain e.g. which glue to use, which wire length, which test program to use, what wafer thickness with which to machine *etc.* It also applies to how to lead the project correctly: specification of the product, acceptance, validation, release. There exist tools to help structure the development of a project and the processes. Before we describe one such tool, we give a description of the APQP. After this we describe the main processes involved in the manufacture of semiconductors.

2.1 Advanced Product Quality Planning (APQP)

Advanced Product Quality Planning (APQP) is a quality framework used for developing new products in the automotive industry [3]. It can be applied to any industry and is similar in many respects to the concept of design for six sigma (DFSS). The APQP process is described in AIAG (Automotive industry Action Group) manual 810-358-3003. Its purpose is "to produce a product quality plan which will support development of a product or service that will satisfy the customer." It does this by focusing on:

- Up-front quality planning,
- Evaluating the output to determine if customers are satisfied & support continual improvement.

The Advanced Product Quality Planning process consists of four phases and five major activities along with ongoing feedback assessment and corrective action as shown below in figure 1.

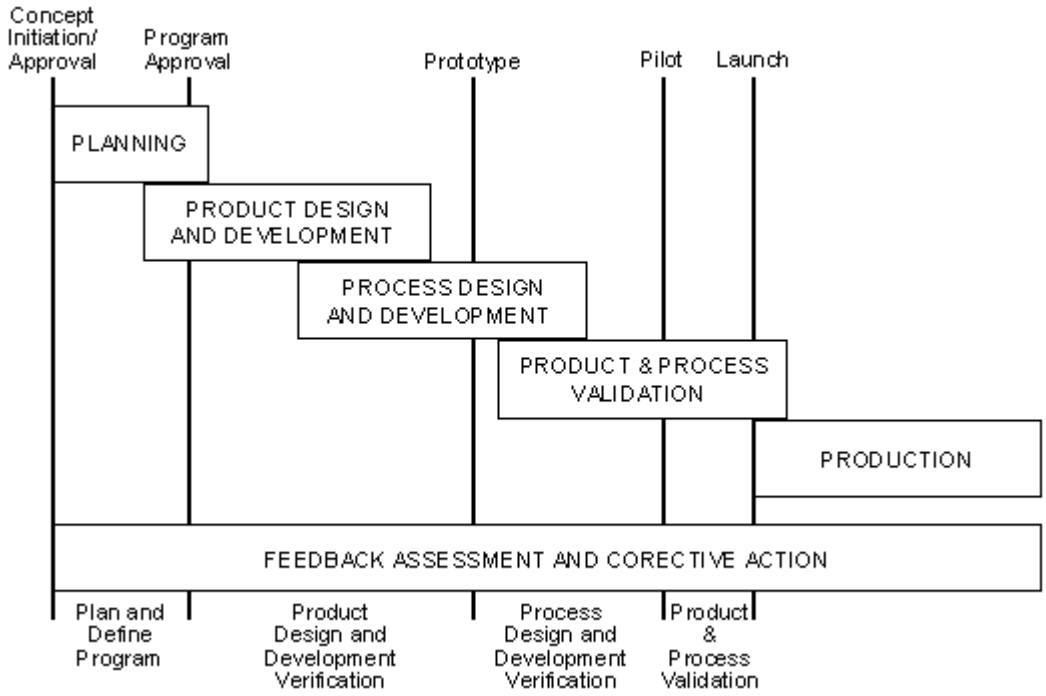


Figure 1: Flow chart of the development of project according to the APQP. For each step, a deliverable is required and leads to a level of maturity of produced samples.

A further indication of the APQP process is to examine the process outputs by phase. This is shown in table1.

Table 1: Overview of all main activities per phase of development.

Plan and Define Program	Product Design and Development Verification	Process Design and Development Verification	Product & Process Validation
▪ Design Goals ▪ Reliability & Quality Goals ▪ Preliminary Bill of Materials ▪ Preliminary Process Flow ▪ Preliminary Listing of Special Product & Process Characteristics ▪ Product Assurance Plan	▪ Design FMEA ▪ DFMA ▪ Design Verification ▪ Design Reviews ▪ Prototype Build ▪ Engineering Drawings ▪ Engineering Specifications ▪ Material Specifications ▪ Drawing & Specification Changes ▪ New Equip., Tooling & Facilities Reqmts. ▪ Special Product & Process Characteristics ▪ Prototype Control Plan ▪ Gages/Testing Equip. Requirements	▪ Packaging Standards ▪ Product/Process Quality System Review ▪ Process Flow Chart ▪ Floor Plan Layout ▪ Characteristics Matrix ▪ Process FMEA ▪ Pre-Launch Control Plan ▪ Process Instructions ▪ Measurement Systems Analysis Plan ▪ Preliminary Process Capability Study Plan ▪ Packaging Specifications	▪ Production Trial Run ▪ Measurement Systems Evaluation ▪ Preliminary Process Capability Study ▪ Production Part Approval ▪ Production Validation Testing ▪ Packaging Evaluation ▪ Production Control Plan ▪ Quality Planning Sign-Off

The APQP process involves the following main elements:

Understanding customer needs . This is done using voice of the customer techniques to determine customer needs and using quality function deployment to organize those needs and translate them into product characteristics/requirements.

Proactive feedback & corrective action. The advanced quality planning process provides feedback from other similar projects with the objective of developing counter-measures for the current project. Other mechanisms with verification and validation, design reviews, analysis of customer feedback and warranty data also satisfy this objective.

Designing within process capabilities. This objective assumes that the company has brought processes under statistical control, has determined its process capability and has communicated it process capability to its development personnel. Once this is done, development personnel need to determine formally that critical or special characteristics are within the enterprise's process capability or initiate action to improve the process or acquire more capable equipment.

Analysing & mitigating failure modes. This is done using techniques such as failure modes and effects analysis or anticipatory failure determination.

Verification & validation. Design verification means testing to assure that the design outputs meet design input requirements. Design verification may include activities such as: design reviews, performing alternate calculations, understanding tests and demonstrations, and review of design documents before release. Validation is the process of ensuring that the product conforms to defined user needs, requirements, and/or specifications under defined operating conditions. Design validation is performed on the final product design with parts that meet design intent. Production validation is performed on the final product design with parts that meet design intent produced production processes intended for normal production.

Design reviews . Design reviews are formal reviews conducted during the development of a product to assure that the requirements, concept, product or process satisfies the requirements of that stage of development, the issues are

understood, the risks are being managed, and there is a good business case for development. Typical design reviews include: requirements review, concept/preliminary design review, final design review, and a production readiness/launch review.

Control special/critical characteristics. Special/critical characteristics are identified through quality function deployment or other similar structured method. Once these characteristics are understood, and there is an assessment that the process is capable of meeting these characteristics (and their tolerances), the process must be controlled. A control plan is prepared to indicate how this will be achieved. Control Plans provide a written description of systems used in minimizing product and process variation including equipment, equipment set-up, processing, tooling, fixtures, material, preventative maintenance and methods.

The internal APQP to NXP is called BCaM (Business Creation and Management). This method of project management was put in place in Philips Semiconductors and kept by NXP. It was based on the project management process PMI (Project Management Institute). BCaM guides project life from conception to closure (figure 2). Its objective is to enable « Doing the right projects »and « Doing the projects right » i.e. launch only those projects of strategical and financial interest, and once project acceptance granted, to lead it to success in terms of time to market, cost, quality… According to BCaM, a project can be split into different phases. Between each phase, there exist milestones for which the project needs to reach a certain level of maturity. Some criteria are defined in consultation with the customer so that when they receive the part they already know the quality level.

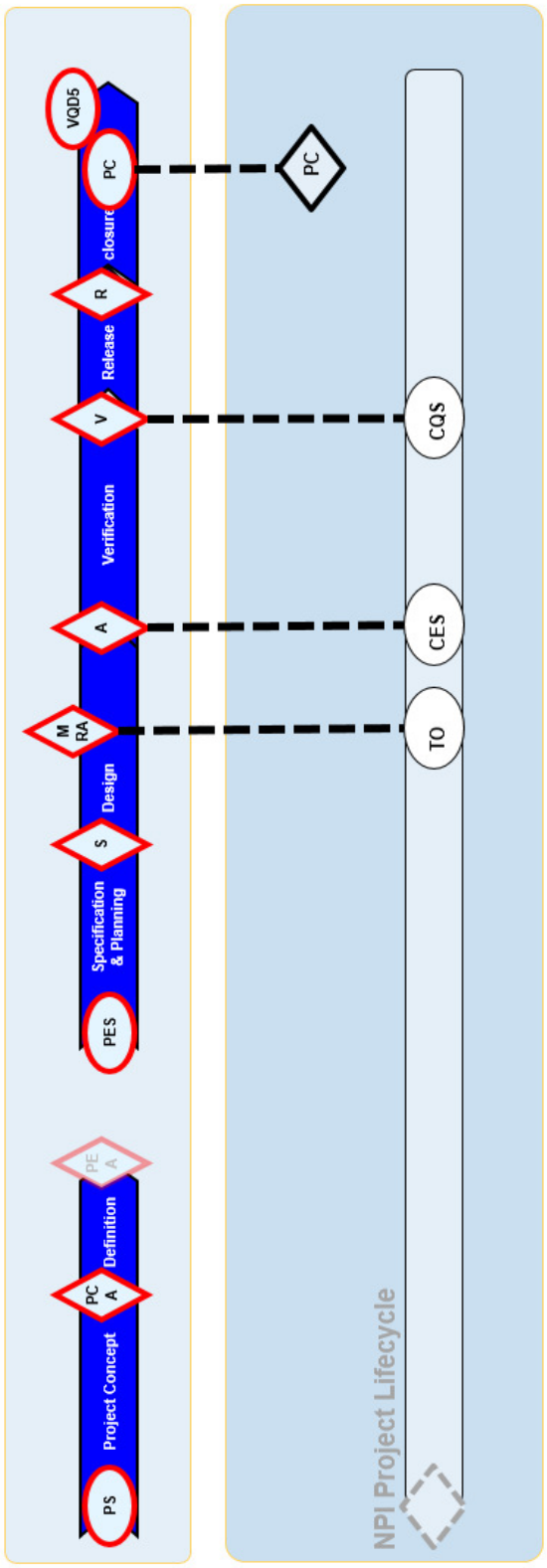


Figure 2: BCaM process : Project life cycle. => The product is officially specified at S-gate. The design of masks is then started at MRA. After the first samples are received the verification and validation starts and end at V-Gate. Between V-Gate and R-Gate the qualification is performed and the test process is prepared for production. PC stands for “Project Closure”. Once the PC gate is reached the development team is officially released and the product

handed over the product to Operations. The Ramp up phase officially ends at the VQD5 gate. VQD5 stands for Volume Quantity Defined 5. It corresponds to the 5% of the total quantity of units which will be produced over the life time of the product based on the business plan.

2.2 Flow of production

2.2.1 From Silicon wafers to active products

In the cycle of fabrication of a semiconductor there are several phases.

Wafer fabrication is a procedure composed of many repeated sequential processes to produce complete electrical or photonic circuits [4]. Examples include production of radio frequency (RF) amplifiers, LEDs, optical computer components, and CPUs for computers. Wafer fabrication is used to build components with the necessary electrical structures.

The main process begins with electrical engineers designing the circuit, defining its functions, and specifying the required signals, inputs, outputs and voltages. These electrical circuit specifications are fed into electrical circuit design software, such as SPICE, and then imported into circuit layout programs similar to those used for computer-aided design. It is necessary for the layers to be defined for photomask production. The resolution of circuits increases rapidly with each step in design: even at the outset of the design process the scale of the circuits is already measured in fractions of micrometres. Each step thus increases the number of circuits per square millimetre.

The silicon wafers start out blank and pure. The circuits are built up in layers in clean rooms. First, photoresist patterns are photo-masked in micrometric detail onto the wafer surfaces. The wafers are then exposed to short-wavelength ultraviolet light (λ=300 nm) and the unexposed areas are thus etched away and cleaned. Hot chemical vapours are deposited on to the desired zones and baked at high temperature (T=900°C), so as to permeate the vapours into the desired zones. In some cases, ions such as O^{+} or O^{2+} are implanted in precise patterns and at a specific depth using RF-driven ion sources.

These steps are often repeated many hundreds of times, depending on the complexity of the desired circuit and its connections. The clean room air quality is highly controlled as each particle is a source of contamination (refer to table 2)

Every year there emerge new processes to accomplish each of these steps with better resolution and in improved ways, with the result that technology in the wafer fabrication industry is forever changing. New technologies result in denser packing of minute surface features such as transistors and micro-electro-mechanical systems (MEMS). This increased density continues the trend often cited as Moore's Law. [4]

Table 2: According to the norm ISO 14644-1 these are the standards for the cleanrooms in terms of particles. Maximum number of particles/m^3

Class	maximum particles/m^3						FED STD 209E equivalent
	≥0.1 µm	≥0.2 µm	≥0.3 µm	≥0.5 µm	≥1 µm	≥5 µm	
ISO 1	10	2.37	1.02	0.35	0.083	0.0029	
ISO 2	100	23.7	10.2	3.5	0.83	0.029	
ISO 3	1,000	237	102	35	8.3	0.29	Class 1
ISO 4	10,000	2,370	1,020	352	83	2.9	Class 10
ISO 5	100,000	23,700	10,200	3,520	832	29	Class 100
ISO 6	1.0×10^6	237,000	102,000	35,200	8,320	293	Class 1,000
ISO 7	1.0×10^7	2.37×10^6	1,020,000	352,000	83,200	2,930	Class 10,000
ISO 8	1.0×10^8	2.37×10^7	1.02×10^7	3,520,000	832,000	29,300	Class 100,000
ISO 9	1.0×10^9	2.37×10^8	1.02×10^8	35,200,000	8,320,000	293,000	Room air

A “FAB” is a common term for the entity in which these processes are accomplished. Often the FAB is owned by the company that sells the chips, such as AMD, Intel, Texas Instruments, or Freescale. A foundry is a FAB at which semiconductor chips or wafers are fabricated to order for third party companies that sell the chip, such as FAB owned by the Taiwan Semiconductor Manufacturing Company (TSMC), United Microelectronics Corporation (UMC) and the Semiconductor Manufacturing International Corporation

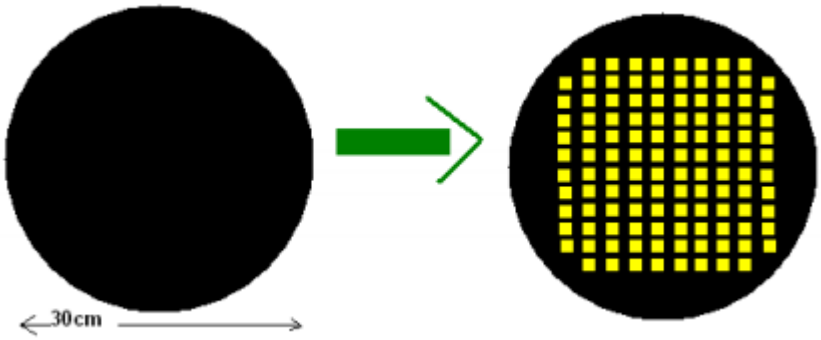


Figure 3: In the FAB, the silicon wafer is submitted to a series of treatments (chemical, ionization, and etching) in different process steps. Different layers are created by superposing masks. These layers form passive and active components composing the product die.

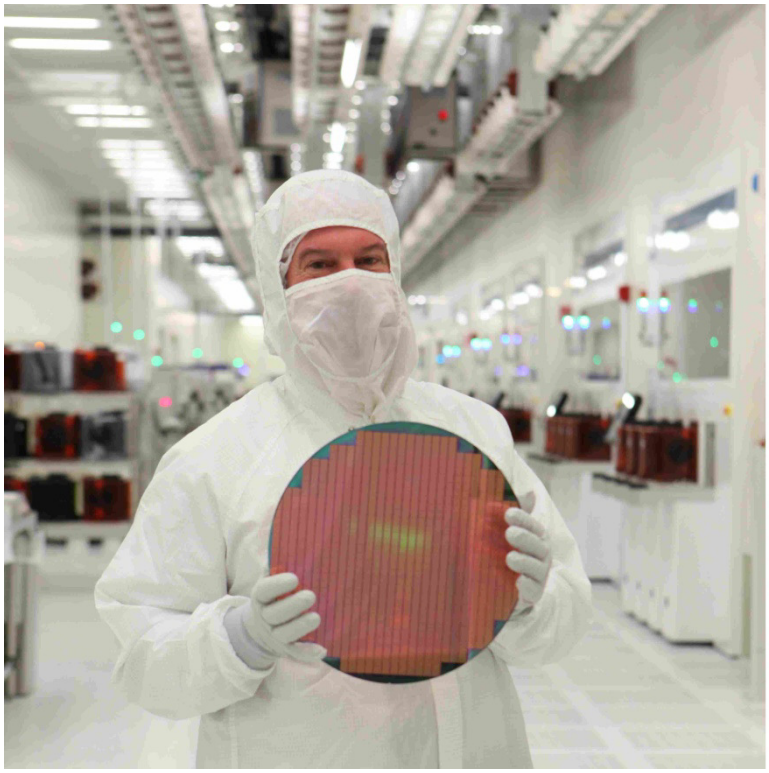


Figure 4: Picture of an engineer holding a 12 inch (30 cm) diameter wafer. He is wearing an integral suit to prevent spreading potential sources of particles

The wafer test is not performed in the FAB. To control the wafers, patterns commonly called PCM are etched along the saw lines. They consist of a set of capacitors, resistances, inductors, transistors which enable the FAB to have a preview of how centred the process is with regards to control limits. PCMs are probed in the FAB before being shipped for wafer

test. The check-up remains basic but potentially it helps the FAB to determine the drift of a parameter such as gate oxide, poly resistor, transistor speed *etc.*

Wafers are shipped from the FAB to the manufacturing centre where they undergo a series of tests and transformations to become a chip shippable to the customer. The first step is the wafer test.

2.2.2 The wafer test

Each die of the wafer is probed. A probe card makes the interface between the tester and the die using needles in contact with the die to carry electrical signals as symbolized in figure 5. The test coverage at wafer test is only partial. A first reason for this is a technological limitation. Needles in contact with the die under test cannot deliver high frequency or high power signals.

Another reason why the coverage is low is related to the efficiency: the wider the test coverage, the longer the test time. In wafer test, the test of one die should take less than two seconds. In addition to representing a first screen stage, the wafer test is an important source of information for the FAB to identify area on the wafer showing potential sensitivity area. Based on the data feedback, the FAB can trace back tools used and fine tune certain process steps. The goal of the wafer test remains to optimize final test yield in a minimum of time. Indeed, the assembly process is expensive. Yield losses in wafer test are “cheaper” than in final test.

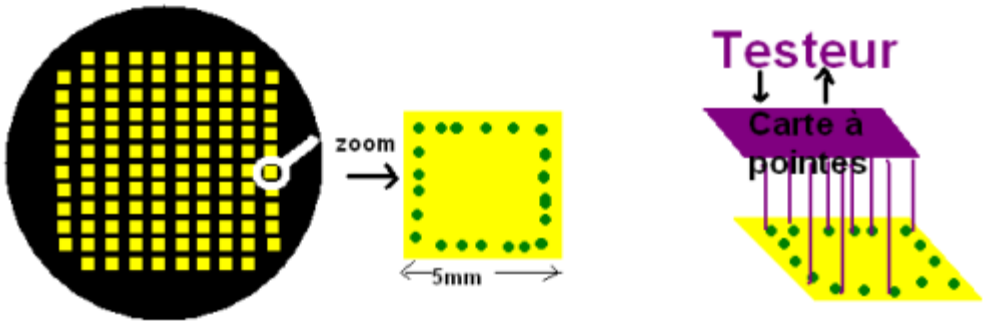


Figure 5: Each die is probed and is tested on wafer. Only good dies are assembled in package.

2.2.3 The assembly

Wafers are sawn along the saw lines (80 µm) and dies singled out as symbolized in figure 6. Based on wafer test results, good dies from wafer test will be assembled as a chip. Dies are pasted onto a lead frame. A wire bonding makes the electrical link between the die and the lead frame. The whole structure is covered by a moulding compound as a protection.

The complex process will not be detailed here. The important point to bear in mind is that the assembly yield is greater than 99.50% and the lead time roughly ten days. The assembly process reaches its maturity level very quickly in the product life span. Tin actual fact it reuses alread released technologies. During a ramp up phase, the assembly process represents only a low source of risk in terms of impact.

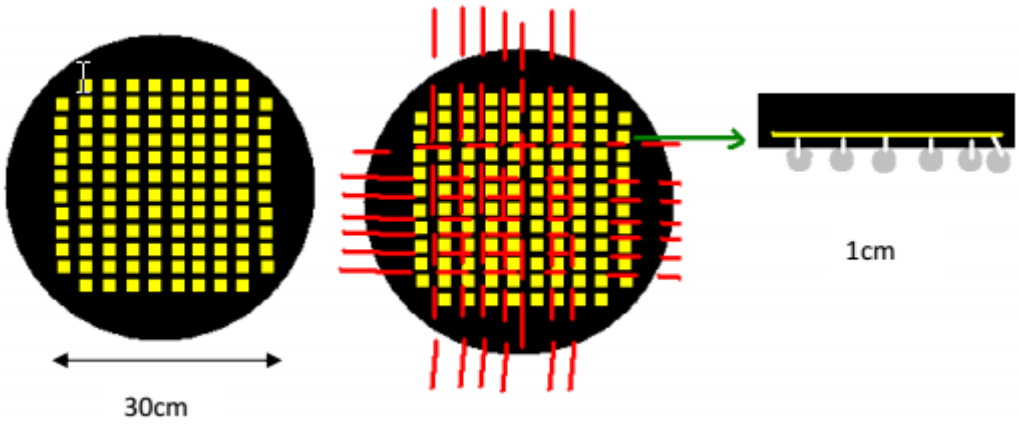


Figure 6: The wafers are sawn and good dies put into package for final test.

2.2.4 The final test.

Once assembled the finished product is submitted to the final test stage. The final test setup is inserted into a socket making the contact between tester resources and the unit under test as seen in figure 7. The test coverage is then exhaustive. The final test is the last control step before delivery to the customer. At this level, the unit will have to be controlled with the final goal being that is conform to the specification guaranteed to the customer. Good parts are then conditioned and packed. To give an idea, the test time per unit is roughly thirty seconds for the application type of the product line for which I am responsible.

If on the one hand the quality of the product is non-negociable, on the other hand excessive quality is a source of invalid yield losses. This optimization of the margin between acceptable quality and too high quality is one key element. The effort to distinguish between a valid reject and an invalid reject requires a lot of resources. And basically, optimizing one test limit once the product is released in production requires far greater effort of elaboration than making it before the product is released. The reason for this is that widening acceptance should not have any impact on the quality of the product, which is hard to prove. This fact was already mentioned earlier and it will be developed further later on. The main point is that effort has to be invested upstream.

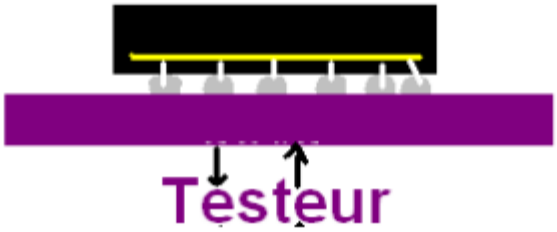


Figure 7: Assembled products are tested in a way they are compliant with the datasheet. Yields in final test are generally greater than 95%.

Figure 8 gives the flow chart summing up processes

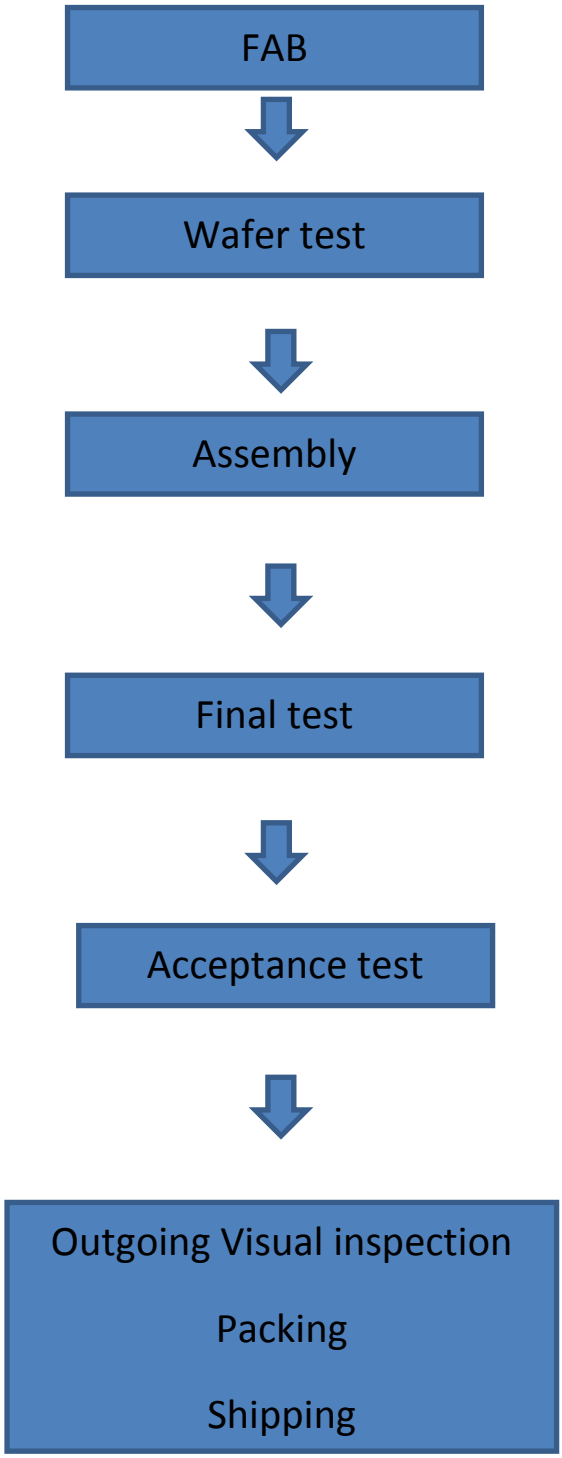


Figure 8: Manufacturing flow. Three months are necessary to obtain the wafers from the FAB. For the wafer test, the production line needs two weeks to perform all operations. Roughly ten days are needed to assemble parts and roughly one week to perform all steps at final test.

Before starting to quantify how large the problem is, let us define how indicators are defined and how Key Performances Indicators are settled.

3.Problem definition and definition of CTQ's

3.1 Yield-to-area calculations (d0)

A Chessboard Test Structure (CTS) provides a list of defect positions. Based on a known wafermap each chip can be marked as "pass" or "fail" depending on the absolute position of a defect. [5] A chip is marked as "fail", if at least one defect is detected inside the chip boundaries. For a specific wafermap a yield value *Y* can be calculated using the following equation (1):

$$Y=\frac{\text{number of pass chips}}{\text{total number of chips}} \quad (1)$$

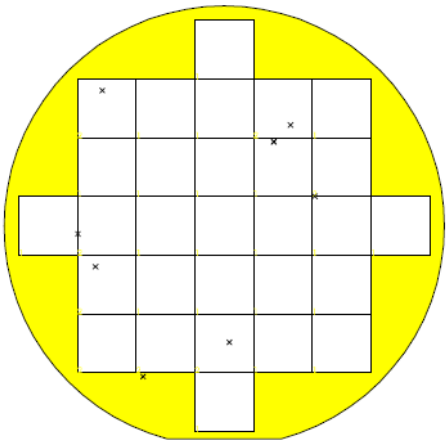


Figure 9: Wafermap containing 29 chips.

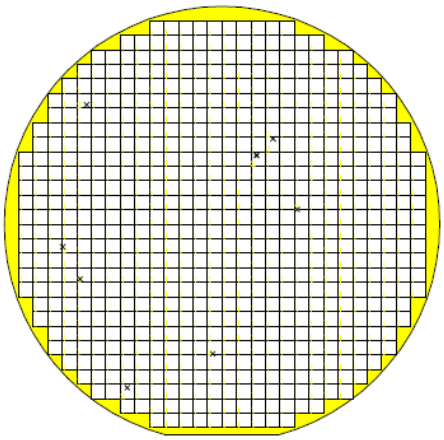


Figure 10: Wafermap containing 648 chips.

By generating a wafermap based on a given chip area *A* and projecting it on the original wafer, one can again calculate a yield value using the same defect list given by the data of the CTS. Figures 9 and 10 show two different imaginary wafermaps. In each wafermap, black symbols mark the defects detected inside the Chessboard Test Structures.

For a series of imaginary wafermaps, the yield will be determined dependent on the chip area. This results in a Yield-to-Area Curve as can be seen in the following figure 11.

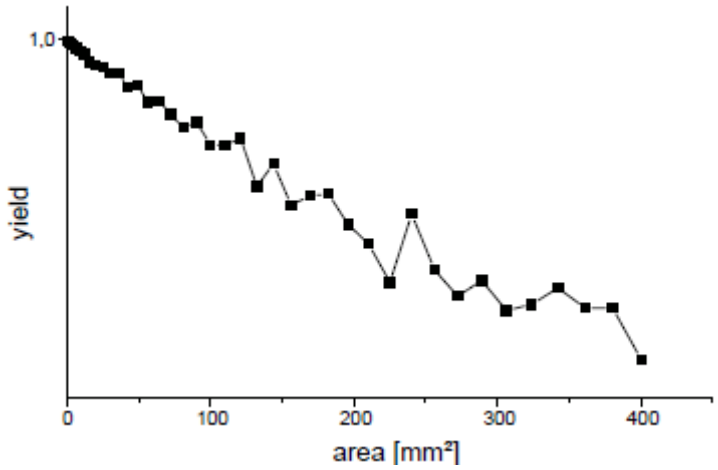


Figure 11: Illustration of a Yield-to-area curve of electrically detected defects on a wafer.

Based on a given imaginary chip area *A* a defect density value *D* can be calculated using the following equation:

$$D=\frac{1-Y}{A} \qquad \text{Equation 2}$$

Where *Y* represents the yield, *A* the area of a single imaginary chip and *D* the defect density. This equation assumes that the wafer area is completely covered by defect sensitive test structures. However, if test structures are combined with product chips or placed inside the sawing lines, they cover just a fraction of the complete wafer area. This fact needs to be taken into account in the calculation of the defect density. We now generate four different hypothetical wafer maps as can be seen in the following Figure 12.

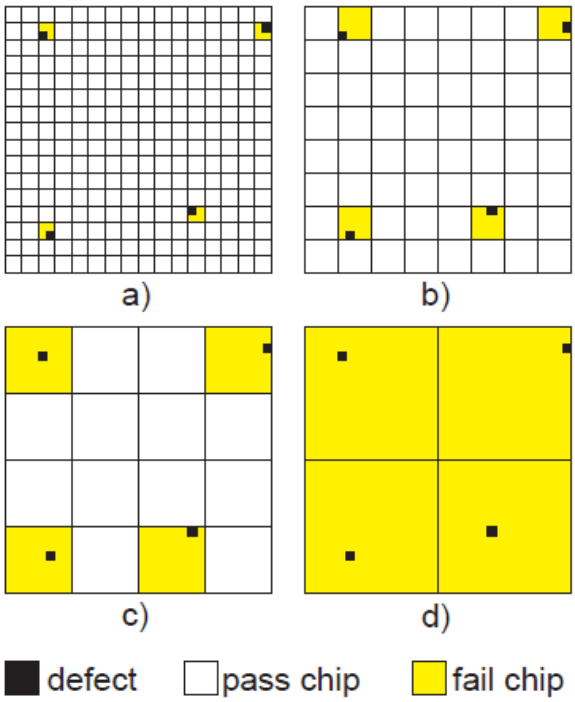


Figure 12: Here are four wafermaps with four different dies size. This is to illustrate that with a given density of defect, the yield loss is different. The bigger the die, the higher is the impact on the yield.

Based on these wafermaps, equation 2 results in the following four identical defect density values of D=0,0625 (unitless)

- (a) 256 chips with a chip area of 0.5 cm²
- (b) 64 chips with a chip area of 1.0 cm²
- (c) 16 chips with a chip area of 2.0 cm²
- (d) 4 chips with a chip area of 4.0 cm²

This illustrates the impact of the D0 on the yield based on the area of the die. The smaller the area, the better the yield for a given D0

This illustration with its series of examples remains purely theoretical.

Practically speaking, the D0 model will differ depending on :

the technology type (CMOS process, ABCD process, RF CMOS process, etc…),

the technology size (140 nm, 75 nm… corresponding to the oxide thickness),

the memory block proportion into the product,

conventions inside the FAB,

and will be customized to match the real situation.

Based on the observed performances and existing models, the D0 model is adapted.

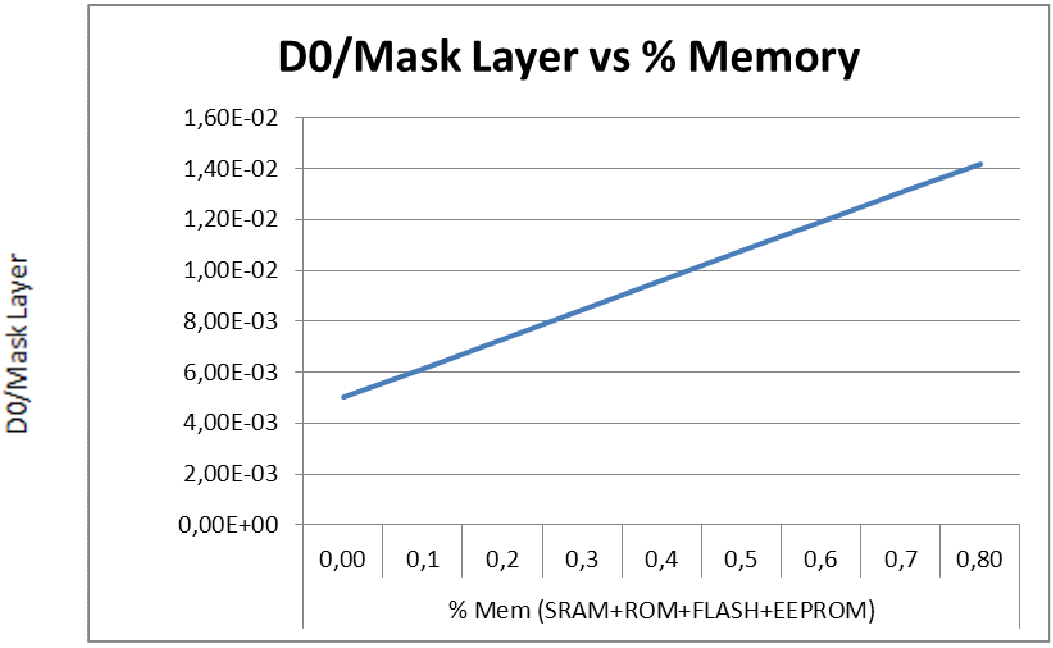


Figure 13: Impact of the memory percentage on the D0. The higher the memory percentage, the higher is the default density per mask layer.

The yield definition (Y) then matches the theoretical illustration given in the introduction to this section.

$$Y=0.985*EXP(-Area * D0) \quad \text{Equation \#3}$$

where

$$D0 = ((G \times T) + C) \times S \times L. \quad \text{Equation \#4}$$

To define its value, the D0 is indeed composed of several variables defined by the FAB to define it value such as the Model Gradient (G), the Total Memory percentage (T) , Model Constant (C) and the number of layers (L)

The values of these parameters for our current process (CMOS14) are given in table 3:

Table 3 : list of variables used for the D0 of the CMOS14

G	0.0087
C	0.00546
S	1.45
L	36

When they are inserted into the equation 2, we obtain

$Y_{(KL)}$= 0.985*EXP (-11.7/100*0.34)=94.6%

This yield is the target yield based on the FAB model. It corresponds to the theoretical yield the wafer test is supposed to reach to be in line with existing models for a given FAB process. This target yield is not necessarily in line with that defined during the business plan.

When a new defined product is accepted to go into development a profitability threshold is defined for a matrix of parameters such as

- the development budget (several tens of millions of Euros)
- a production plan for the next five years (volume in millions of units per year)
- the sale price
- a test time (will impact the machine usage and hence the test cost – roughly 20 s per tested product in final test stage)
- the area of the die (the larger the die, the more wafer silicon used – ~10 mm^2)
- the material used for assembly (copper or gold wire)

and the production yields target.

The production yield target is there to fix a limit that ensures the margin defined at the specification phase of the product is reached.

This yield is not necessarily aligned with the one defined with the D0. A frequent explanation for this is that during the feasibility assessment, the design of the product is made in such a way that the product might lie one the edge of the process capability, which will induce yield losses. These predicted yield losses are taken into account in the business plan, so the yield is expected to be lower than the baseline. The consequence is that there might be a gap between the yield target based on the D0 and the one from the business plan.

The yield target based on the D0 remains the initial engineering target yield once production has begun. However, it might turn out to be more worthwhile to stop yield improvement activities if the yield has already reached the financial target yield laid down in the business plan. Indeed yield improvement activities are expensive in terms of resources and materials. So before one invests in yield improvement activities, a solid business case has to be defined, especially if the financial target has already been reached.

3.2 Type of rejects

The term “yield improvement activity” refers most of the time to the wafer test yield. This yield reflects the FAB process performance. Although the wafer test is carried out in another plant, the FAB uses wafer test yield performances to monitor its own performance. To discuss this matter in greater depth, it is now important for us to distinguish between two types of rejects: functional and parametric.

Functional rejects are those rejects that by definition are not functional. A block inside the product does not respond correctly. For instance for a failing digital block, the response signal is “0” when a “1” is expected. For an analogue signal usually characterized by a normal distribution [11], the response of a functional reject will then be an outlier as symbolized in figure 14.

By contrast, parametric rejects are those rejects that remain functional but lie beyond specification limits. The term “marginal rejects” is another commonly used name for them. This only concerns analogue signal. This type of reject is more sensitive to any variation of processes such as FAB processes or test processes. Fine tuning these processes or test limits represents the main way of yield optimization.

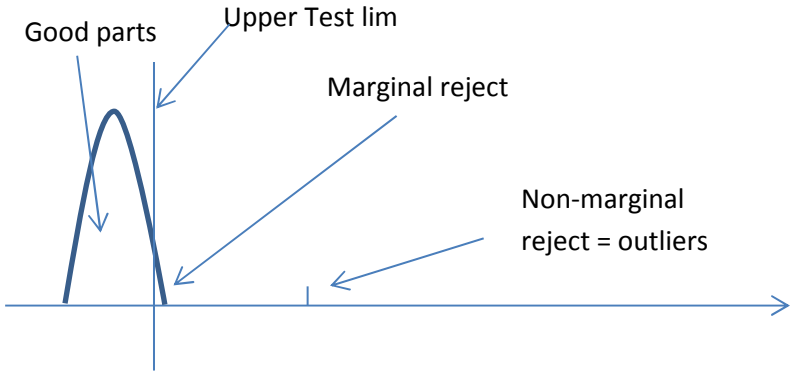


Figure 14: Distinction between an outlier reject and a marginal reject. Marginal rejects are sensitive to the variation of processes (test process repeatability and reproducibility, FAB process). They are contestable and are the usual targets for yield improvement activities by improving the stability of the test parameter or by re-centring the FAB process

3.3 Representativity of distribution

The central limit theorem states that if one has a population with mean μ and standard deviation σ and one takes sufficiently large random samples from it, the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually $n > 30$). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that $\min(np, n(1-p)) > 5$, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

For the random samples taken from the population, one can compute the mean of the sample means:

$$\mu_{\bar{x}} = \mu$$

Equation #5

and the standard deviation of the sample means:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Equation #6

Before illustrating the use of the Central Limit Theorem (CLT) we shall first illustrate the result. In order for the result of the CLT to hold, the sample must be sufficiently large ($n > 30$). Again, there are two exceptions to this. If the population is normal, then the result holds for samples of any size (i.e. the sampling distribution of the sample means will be approximately normal even for samples of size less than 30).

The figure 15 below illustrates a normally distributed characteristic, *X*, in a population in which the population mean is 75 with a standard deviation of 8. On the Y-axis, is the frequency of appearances.

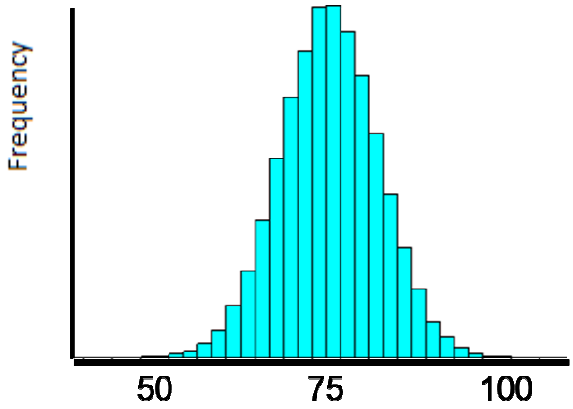


Figure #15 : Histogram representing a normal distribution

If we take simple random samples of size *n*=10 from the population and compute the mean for each of the samples, the distribution of sample means should be approximately normal according to the Central Limit Theorem. Note that although the sample size (*n*=10) is less than 30, the source population is normally distributed, so this is not a problem. The distribution of the sample means is illustrated below in the figure 16.

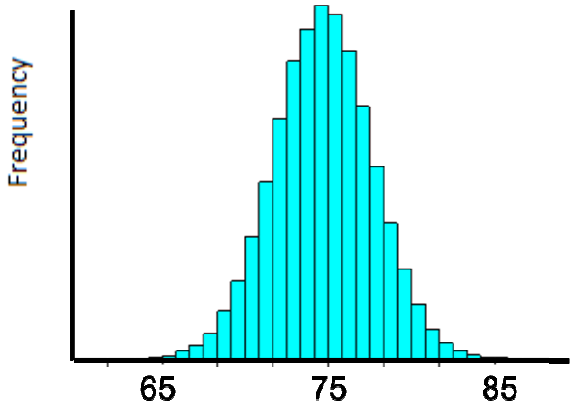


Figure #16 : Histogram representing the distribution composed of mean values

The mean of the sample means is 75 and its standard deviation sigmabar 2.5, with the standard deviation of the sample means computed as follows using the equation 6:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{10}}$$

If we were to take samples of *n*=5 instead of *n*=10, we would get a similar distribution albeit a broader one. In fact, when we did this we obtained a sample mean of 75 and a sample standard deviation of 3.6.

3.4 Origin of rejects

Having described these two kinds of rejects, we can now define their origin and how to tackle these types of yield losses. Theoretically functional rejects reveal a severe failure i.e. that something has not been well processed. This is typically a FAB matter. The origin of parametric rejects leaves more room for interpretation. During the D0 definition, the expression “process capability” was mentioned. Depending on how the device was designed and specified, its performances will lie more or less closer to the limit of what the selected FAB process is able to provide. Going beyond the capability of the FAB process will make the product performances more sensitive to any environmental variations. It will be at the origin of parametric yield losses due to FAB process variations but mainly due to the ultra-sensitivity of the product. Whereas a product designed with a six sigma capability approach would absorb minor variations, a product at the edge of the FAB process will show parametric yield losses due to its ultra-sensitivity. During the design phase, the objective is to make a robust product for which natural modulations in the FAB have no impact, which is not always the case. One consequence of this is that the FAB has to work on fine tuning activities to compensate somehow too sensitive a product.
This sensitivity level is revealed during wafer test.

The main role of the wafer test remains the screening out of rejects. As mentioned earlier in our description of the production flow (Section 2.2.2), there are several reasons why the test coverage cannot be total:

- the first is technical : the impedance of needles is a limitation

- the second one is because one has to optimize the throughput time. The wider the test coverage, the longer the test takes. Beyond the financial aspect, the game is to optimize the machine usage and the lead time. Considering this is not the final stage, a compromise test time test coverage has to be decided upon.

The wafer test is performed at room temperature at which the trimming is done, at high temperature and at low temperature. This is a requirement of the automotive market. Not surprisingly, most functional rejects occur during the test carried out at room temperature as it also corresponds to the first time the wafer is tested. By contrast, parametric rejects occur mainly at high and low temperatures which represent extreme test conditions.

3.5 The final test

The final test is the one performed after the assembly of the die into a package. It is above all the last test stage before shipment to the customer. This means that, beyond this stage, the delivered product has to be fully compliant with the datasheet of the product. It does not mean however, that all the parameters of the datasheet have to be tested in the final test stage. The goal of testing products is not so much to characterize the product but rather to ensure they are compliant with the datasheet and that no reject is sent to the customer. These two statements might look similar but the nuance makes a big difference to the design of the test process. It means that each datasheet parameter is not necessarily intended to be individually tested but satisfied. For instance, if two parameters are shown to be correlated, only one of them need be measured. Alternatively, it might turn out to be easier or quicker to test a parameter by an indirect method.

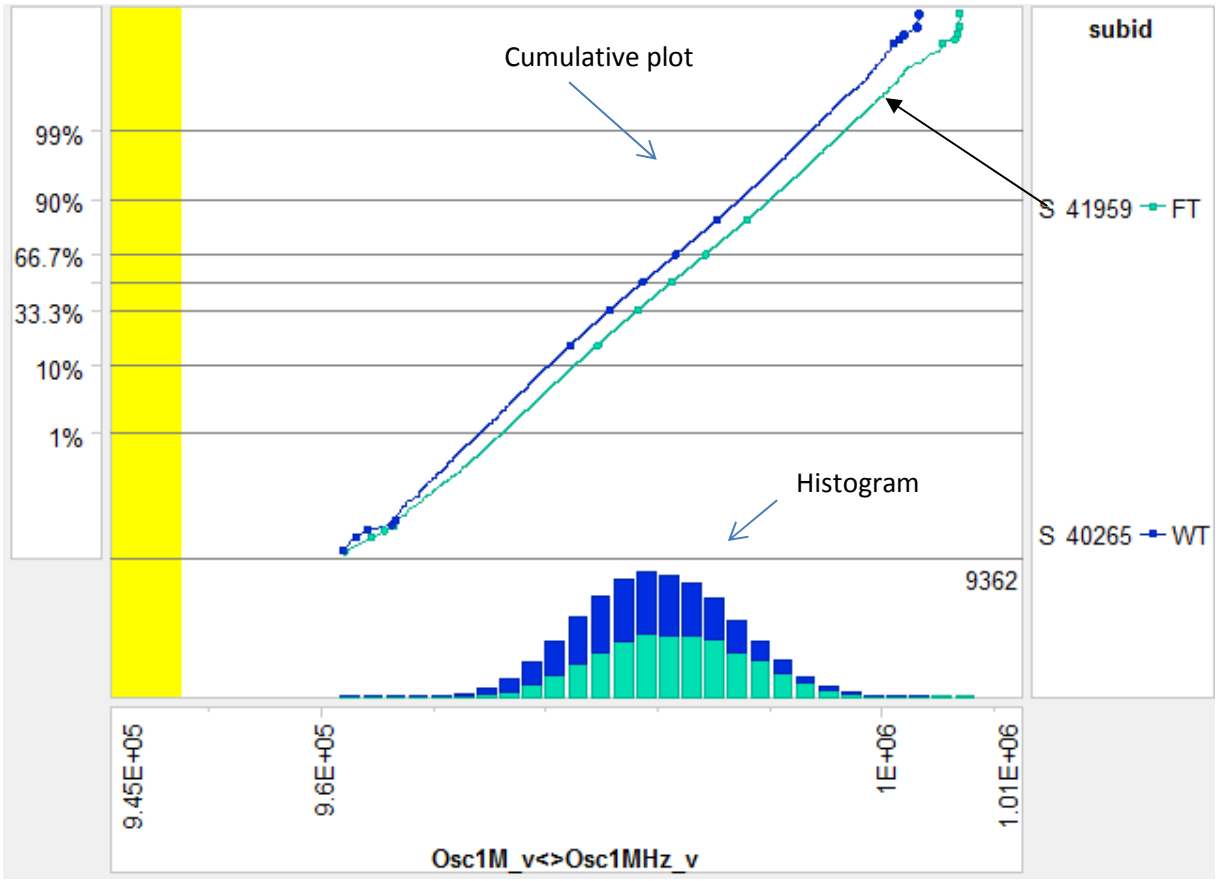
If most of parameters will indeed be tested during this last test stage, some of them will be guaranteed by design. Most of the time, parameters guaranteed by design are those that are not testable. For instance, they might not be testable because the test would require too complex an algorithm which itself would then take too long to be tested. Or they might not be testable because measured signals would be too weak to be correctly interpreted by the

tester. For these reasons, products will not be tested for parameters guaranteed by design. Of course to allow a datasheet parameter not to be tested, one has to verify it shows naturally enough margins (six sigma) with regard the datasheet limit. To prove this, a representative sampling population (thirty units drawn from three different mother batches) will be measured on the lab bench.

Apart from tests guaranteed by design, there is another category of tests not included in final test. To cut test time costs, a test done in wafer test might not be repeated in final test. To meet conditions for such a case, two conditions are necessary:

1. The wafer test conditions should be stricter or equal to those in final test
2. The correlation of the parameter between wafer test and final test should be sufficient to guarantee no risk of escape. Indeed, there are cases where the correlation between a test performed in wafer test and final test is not satisfactory because of external variables such as the effect of the package or the limitation on the impedance of the needles during wafer test.

Figure 17 shows an example of a test parameter with a good wafer test to final test correlation.



Group	Count	Mean	Std dev	Min	Max	Actual yield loss	Cpk
FT	41959	985037.532	5397.099	961657.375	1005564.000	0.00%	2.164
WT	40265	983007.576	5254.109	961557.688	1002645.000	0.00%	2.094

Figure 17: This distribution results from the measurement of a 1 MHz oscillator (X-axis). In dark blue is the distribution measured wafer test, in light green the same measure in FT. The offset on the mean between FT and WT is 20 kHz (2.0%). The offset is visible in the table above and on the cumulated plot. Typically, with so high a correlation [12], to optimize the throughput time, removing that test either in WT or in FT is to be considered. On the Y-axis, is the percentage of the cumulated plot.

The wafer test coverage is done in such a way that the yield in final test is greater than 99%. In theory, only rejects due to assembly failures should occur in final test. In practice there are tests not easily feasible in wafer test such as RF tests, once again due to the problems of needle impedance. There are technologies enabling RF tests in wafer test but they never reach performances met in final test. Performances would be sufficient to measure RF gain, output power but not to quantify parameters like RSSI. As the RF coverage is never total,

most of the time the choice is taken not to perform RF tests in wafer test. Making the choice not to have RF tests in wafer tests enables one

- to use cheaper needle technology and cheaper tester configuration;
- to have faster wafer test sequences since RF tests are time consuming
- to increase the level of parallelism (number of dies tested in parallel).

For these reasons, there are tests not done in wafer tests, which leaves the possibility of having yield losses in final test.

To recover the yield, rejects are retested once. The recovery rate is a good indicator of test process maturity. The lower the rate the more stable the test process and the higher the confidence level in the quality of the test process.

3.6 Acceptance test

An acceptance test is performed before the preparation of good parts to the delivery. It consists in retesting a sampling quantity of good parts are retested. This test has two goals:

- to ensure that good parts remain so after test and have not been damaged by it
- to ensure that good parts and rejects were not intermingled.

Test limits for acceptance tests are slightly wider than in final test. If for the final test a margin is taken towards specification limits to prevent a reject is accepted, for the acceptance specification limits are applied as test limits.

Of course the acceptance criterion is that all parts pass the acceptance test. An acceptance sampling scheme is a specific set of procedures which usually consists of acceptance sampling plans in which lot sizes, sample sizes, and acceptance criteria, or the amount of 100% inspection, are related. Such schemes typically contain rules for switching from one plan to another [7]. MIL-STD-105 E (1989) is an example of a sampling scheme. The sampling quantity is defined according the AQL (Acceptance Quality Limit) 4%. Given production batches are smaller than 10,000 units, this quantity is set to 315. Similarly to the retest recovery rate during final test which is a good indicator of the health of the test process, the acceptance test is also a good indicator : for instance if five runs are required to obtain 100% passing parts, some tests must be unstable.

3.7 Definition of test limits based on specification limits

Tests limits are defined in such a way that delivered samples are compliant with the product specification. Delivered samples have to be compliant to the product specification not only upon delivery of the products but also after several years in service.

3.7.1 Adapting test limits to include the aging of parts

If a drift is observed during the qualification phase and accelerated aging, then test limits have to be tightened to anticipate the drift over the years. Let us consider the example in which a distribution drifted after aging from A to B. For this parameter the upper test limit will be lessened by Δ to anticipate the drift over years as illustrated in figure 18.

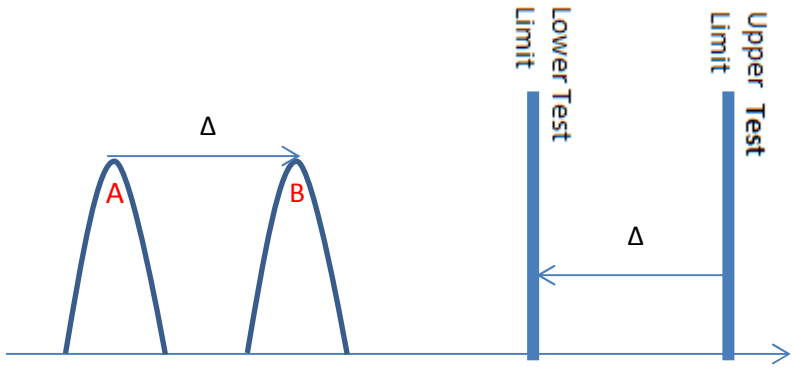


Figure 18: Illustration of a drift after life test (e.g. a consumption). If after a life test a drift has been observed (Δ), the upper test limit is tightened so parts remain compliant to the specification after years of usage.

3.7.2 Adapting test limits to include the repeatability of the test

Other parameters have to be incorporated in the calculation of the limit. The stability of the measurement is one of them. Each measure has its repeatability, its standard deviation. This means that if the measurement is repeated its result will vary. There are several parameters that influence the repeatability of a measurement. This one is mainly an interaction between the strength of the signal, the sensitivity of the tester, and the tester environment. Applying an averaging improve the repeatability but also increases the test time. In other words, instability is critical for the reason that it might allow through a reject as shown in figure 19.

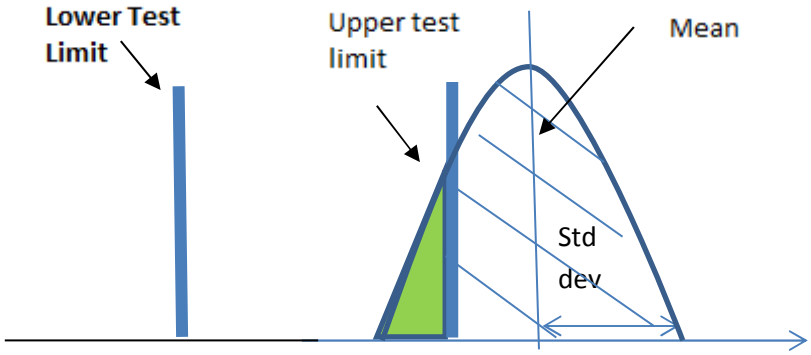


Figure 19: Distribution of a test parameter obtained from a repeated measurement on a single sample. The distribution is so close to the upper limit that from one run to another, the reading either passes or fails.

If a measurement is repeated, the overall results form a normal distribution. This graph in figure 19 is the distribution of a test parameter repeated in loop. The area shaded in green corresponds to the left side with regard to the upper limit. It then corresponds to the area in which results pass. In this example the measurement is repeated and it happens the result passes (dark area) but in most cases it will fail. In production, the measurement is performed only once. In cases such as the one given as an example, it may happen statistically that a single measurement result falls in the dark area. Intrinsically this test parameter fails so the product has to be rejected. But due to the repeatability of the test in that example it might pass. To prevent a sample from being delivered with the risk of a failing test parameter, the repeatability has to be taken into consideration. Therefore a margin on test limits will be adapted to prevent the risk any bad elements are let through. Typically, in our example, the upper test limit will be tightened by the value of the standard deviation.

3.7.3 Tester-to-tester variation

A testing machine has its own specification provided by the manufacturer of the tester. This includes the noise floor, the sensitivity, the stability and more generally speaking its capability. It also states the maximum difference observable between testers. Indeed, despite tester machines being calibrated using the same reference checker, the uncertainty

gives rise to a permanent that an offset remains. The maximum offset between testers is given in the specification of the machine. In a similar way the instability of the measurement might make a failing test parameter pass, an offset between testers might also make components with failing parameters pass.

Let us consider a parameter tested in a laboratory. In the example below, the result of the measurement fails in laboratory (L), on tester (B) but passes on tester (A). The difference between tester A and B symbolizes the reproducibility between testers as seen in figure 20.

With tester A there is a risk of over acceptance, potentially causing customer returns. With tester B there is a risk of over screening, leading to invalid yield losses with regard to the real value tester in laboratory.

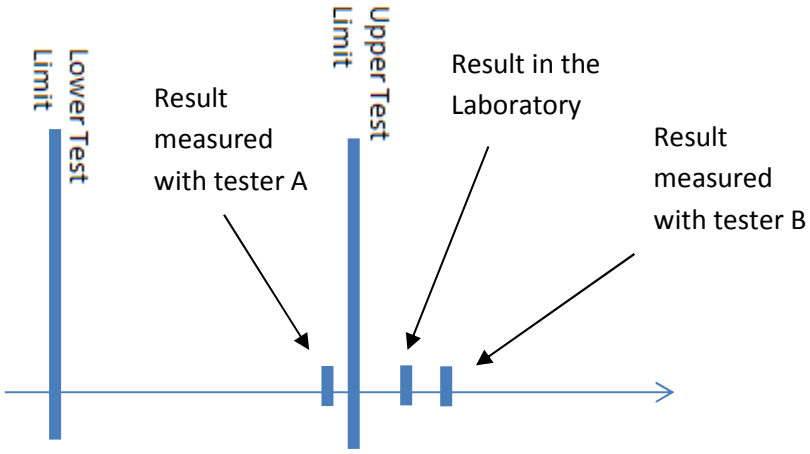


Figure 20: Parameter tested on three different benches. This risk of over-acceptance has to be prevented by tightening again the limit in order to include the reproducibility between testers.

The figure 21 sums up all the parameters which are taken into consideration to compute test limits with regard to specification limits.

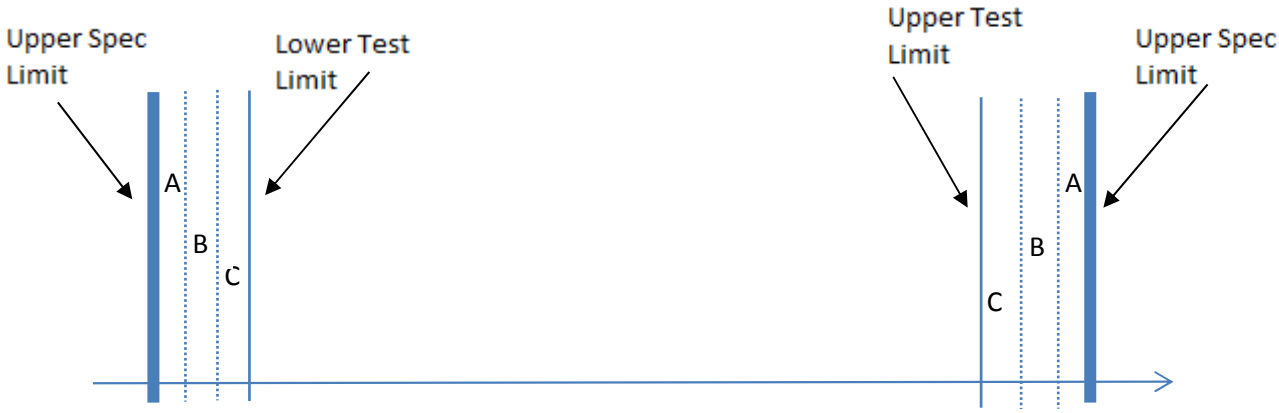


Figure 21: Ensemble of all the variables to be considered when computing test limits. To define test limits with regard to specification limits so there is no risk of any rejects reaching the customer as well as to prevent marginal parts drifting and failing over life time, test limits are tightened by the value of the test repeatability (for instance A), the tester-to-tester reproducibility (for instance B) and the drift over life time (C).

3.8 Hold lots criteria.

At each of these CTQ there are acceptance criteria. If any one of them is not satisfied, the production batch is put on hold for analysis.

- During wafer processing, inline monitoring guarantees that every process step lies within control limits.
- PCM must lie within control limits (generally within three sigma) before the wafer is allowed to leave the FAB.
- There are also visual inspections to check for any discoloration or scratches on the die.
- During wafer test and final test there are acceptance limits not only on the global yield but also on critical individual test parameters.
- During the assembly process there is an inline monitoring and acceptance criterion.
- And naturally during acceptance tests, the last test step before delivery, all retested good parts have to pass.

3.9 Impacts of CTQs and criticality during volume ramp up phase.

Naturally, production yields have a financial impact. The lower the yield the more material is needed to satisfy demand. In extreme cases, if yields are too poor the production batch might end up being scrapped. This is of course a non-negligible aspect but during a ramp up phase the priority is more delivery on time than the profitability. In addition, for low volumes, the financial consequences of having a low production yield remain acceptable.

The biggest impact of production yields during the ramp up phase is on the lead time. The poorer the yield the more material is required. This means that every process will lose in efficiency so to compensate low production yields more material has to be manufactured and tested. Here lies the main impact on the lead time. Especially during the ramp up phase there is no slack in the warehouse. An unpredictable lead time is a source of delay.

Another impact of having low yield is a risk of raw material shortage, the lead time or obtaining new wafers being roughly two months. Between the time the product is officially released and the time enough wafers are in the wafer bank to cover frequent manufacturing incidents, it can take months. During this period, having low production yield weakens delivery capability. A side effect of having low production yields is also that the quality level is at risk. Indeed, if for instance 50% or 60% of units in a sample batch fail, there must be something wrong with either the tester material or the test environment. Even if good parts of that batch passed the test sequence, the confidence level regarding the quality level is questionable.

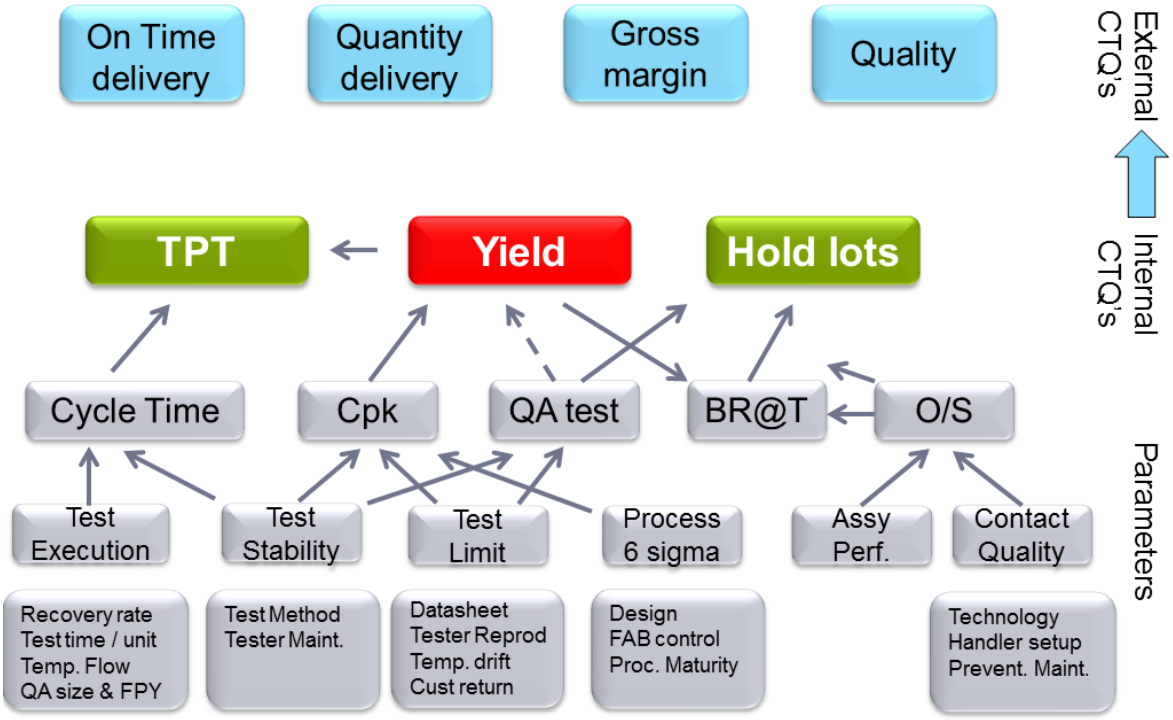


Figure 22: This figure gives the list of "Critical To Quality" factors, parameter contributing to them and how they interfere with each other. The production yield remains the biggest contributor. The yield coupled to the cycle time are the major contributor to the throughput time (TPT). The open/short rate (O/S) corresponds to the percentage of samples not functional at all. BR@T is the name of the system managing hold lot criteria

Low yields are mainly due to the fact that the test process and the manufacturing processes are not yet mature. Encountering low yields and going over the learning phase which is a cause of incidents increases the risk of hold lots at every process step. This slows down the throughput time. Figure 22 gives an illustration of how parameters impacting the manufacturing processes are cascaded leading to throughput time slow down. The length of this learning period will depend on the degree of innovation embedded into the new product.

The challenges faced during the volume ramp up are thus to provide delivery on time and with the correct quality level.

4.Quantification of the problem

4.1 Impact on yield performances

The problem which has been addressed by this study is to minimize the impact linked to the volume ramp up phase and secure the early deliveries: The aim is to make production yields and lead times stable and on target as soon as possible during the ramp up phase so as not to risk delivery shortage and customer returns. Before discussing the approach adopted to find solutions, let us first discuss the severity of the problem.

As an illustration let us consider just the production yields and hold lot rates of a few products released into production in 2010-2011 Here are yield figures for a high runner released in 2011 as shown in figure 23 and 24.

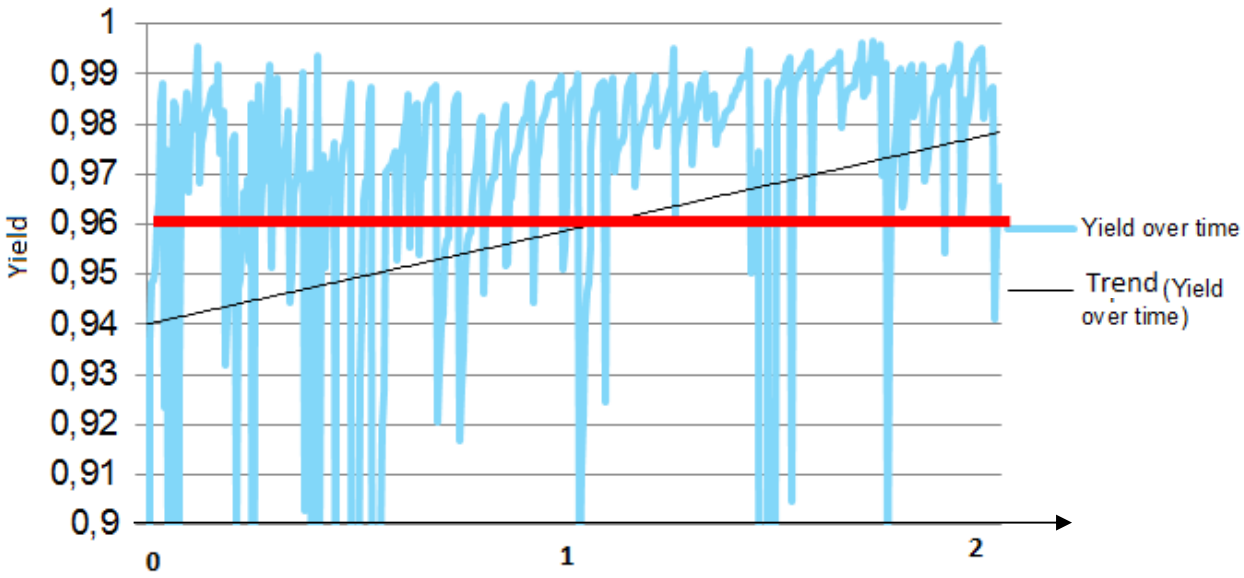


Figure 23: Product yield of one of our high runner products over a two year period. Each data represents the yield reached by production batches. To make our discussion more concrete, let us suppose that one production batch in final test is composed of roughly 4000 units which somehow matches with the quantity of the reel shipped to the customer. To make the graph easier to read, the scale was cut off at 90% although there were occurrences below the minimum shown on the graph.

The hold lot limit is symbolized by the thick line. If the yield of a production batch falls below 96% it will be put on hold and will require an analysis before being released. If the case is

known, the batch will be released in one or two days. If it is a new case, it might take several days to be released, which might lead to a rescreening of good parts. This has an impact on the throughput time. As a side effect, it has an impact on the workload on the engineering team.

What we also observe is the trend line increasing over the time thanks to yield improvement activities. The approach making a yield improvement will be detailed later. Basically, yield improvement activities are led to tackle invalid rejects. This is mainly to minimize the need to invest in yield improvement activities once the product is in production that the project has been led.

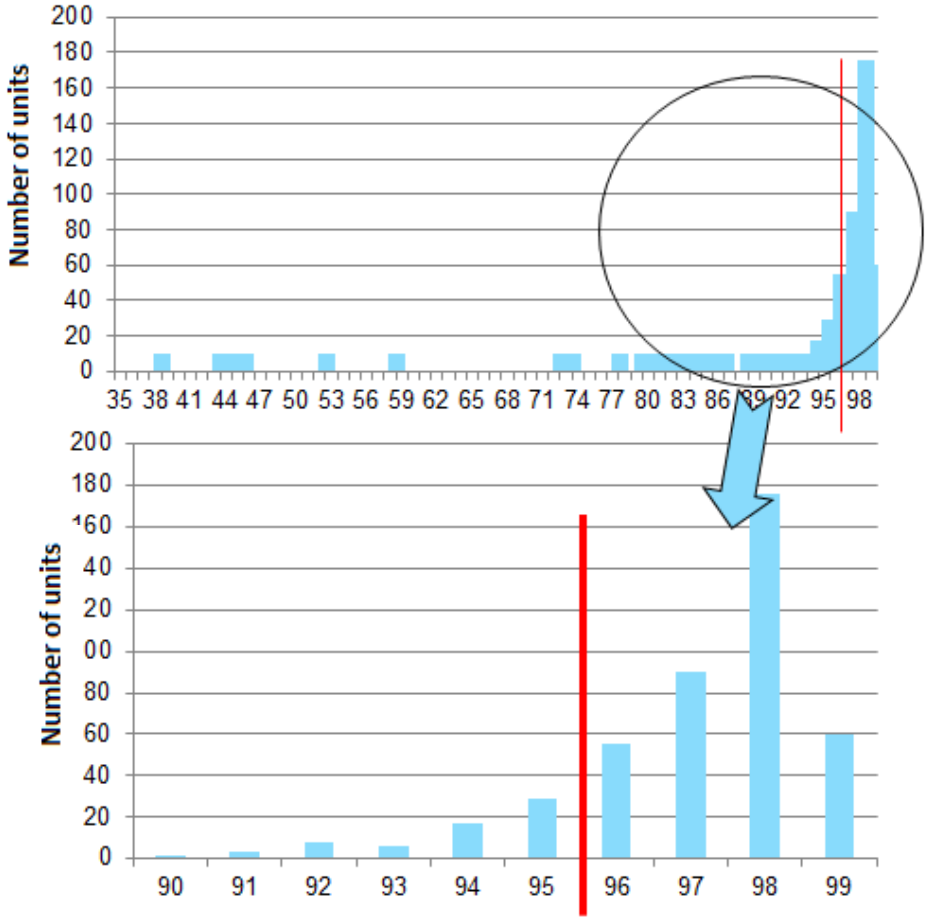
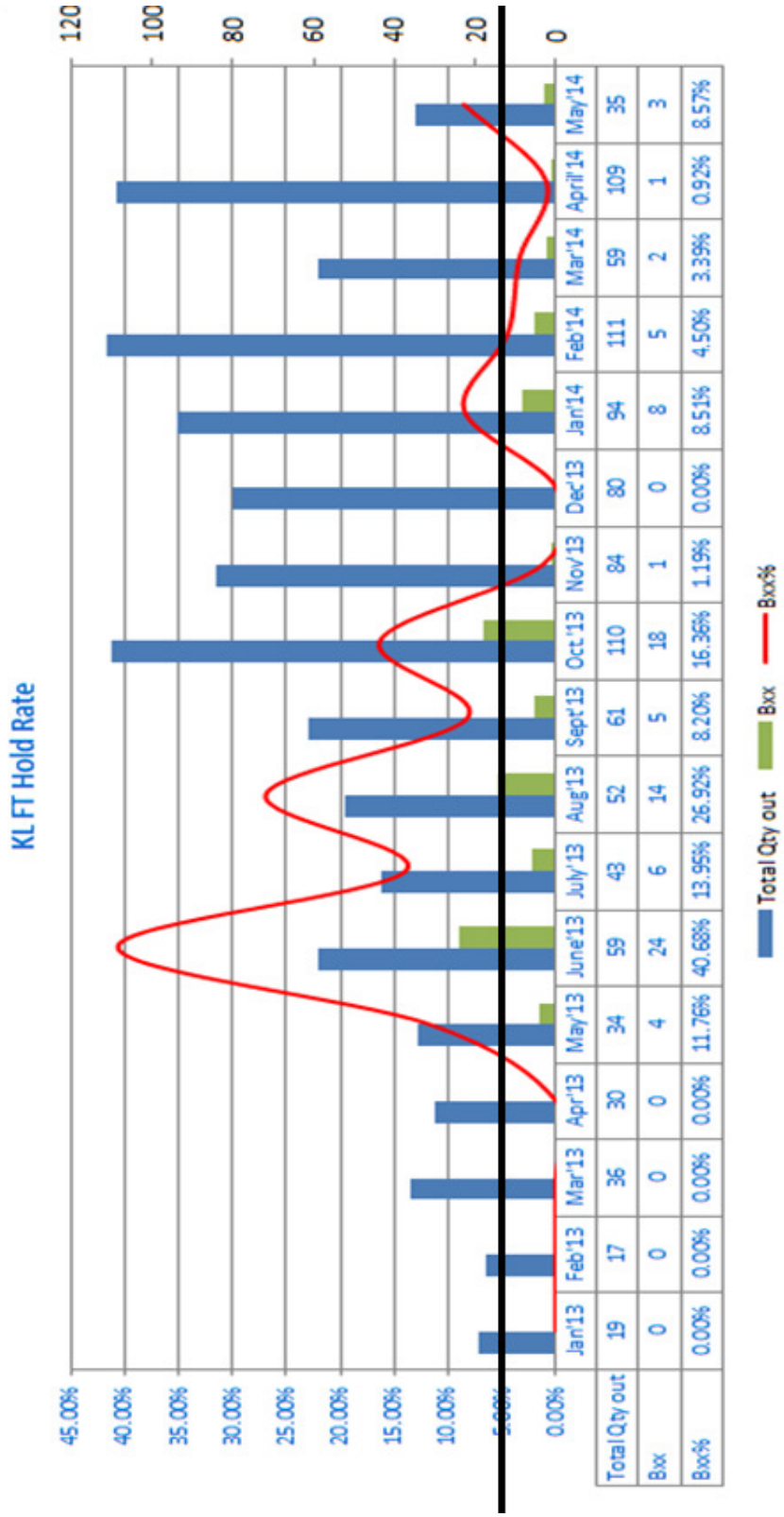


Figure 24: This graph displays data of Figure #23 this time in histogram form [10]. It helps one to figure out the frequency of occurrences.



	Jan'13	Feb'13	Mar'13	Apr'13	May'13	June'13	July'13	Aug'13	Sept'13	Oct'13	Nov'13	Dec'13	Jan'14	Feb'14	Mar'14	April'14	May'14
Total Qty out	19	17	36	30	34	59	43	52	61	110	84	80	94	111	59	109	35
Bxx	0	0	0	0	4	24	6	14	5	18	1	0	8	5	2	1	3
Bxx%	0.00%	0.00%	0.00%	0.00%	11.76%	40.68%	13.95%	26.92%	8.20%	16.36%	1.19%	0.00%	8.51%	4.50%	3.39%	0.92%	8.57%

Figure 25: Percentage of hold lots with regard to the number of tested batches. If the production well started in January 2013, the hold lot rate increased drastically in May leading to a crisis period. It took almost one year for the hold lot rate to be brought back to an acceptable level.

As mentioned earlier, a hold lot is a lot which does not pass acceptance criteria. The hold lot rate is the ratio between hold lots and passing lots. The usual limit of the hold lot rate accepted by manufacturing is 5.0%. Beyond this limit, guaranteeing deliveries, storing material on hold and managing analysis are no longer viable. Let us look at the origin of these hold lots.

While hold lots could be triggered intentionally when a potential risk is detected due to an abnormal behaviour, most of the time, they are triggered automatically when the performance of the batch is not normal, i.e. below the acceptance limit. Hold lot limits are defined by engineers to detect abnormalities. These abnormalities can arise not only from a problem encountered during manufacture or a drift in FAB or misprocessing during assembly but also from a test process issue such as a bad contacting of the unit under test or a problem with the tester machine. In the case of a test process issue, hold lots are caused by invalid rejects. In this case, rejects are not real; they fail because of the test environment.

There are two main approaches for tackling yield losses: either lowering the occurrence of valid rejects or preventing invalid rejects

4.2 Yield improvement on valid rejects.

By definition, valid rejects would fail on the customer side. So here there are only two options:

- Either working upstream with the FAB to fine tune the recipe and hence correct potential yield losses due to sensitivity to the FAB process. This is part of what is done during the learning phase to compensate design sensitivity.
- Or trying to negotiate with the customer a more tolerant datasheet relaxation to increase the margin. This is possible if the reason is justified and accepted by the customer. Whatever the case, this is a burdensome and sensitive process as it might also affect customer confidence. Indeed, it is tantamount to asking the customer to accept more than what was agreed to at the release of the product.

4.3 Yield improvement on invalid rejects.

There are many possible origins of invalid rejects.

They could be purely manufacturing related and might be caused by a tester calibration issue, a contact issue due to a handler setup or due to a dirty socket, a load board defect or a handler temperature setting. Most of the time, the impact of this category of invalid rejects is isolated and limited. Once the issue has been isolated and corrected, rejects are merely retested for recovery.

Invalid rejects could also be caused by a test process over screening or instability. Detecting and optimizing is the kind of activity that takes a lot of effort. A first significant step consists in identifying yield losses. Deciding what to do, how to improve yields is also an expensive activity as it requires DOE's (Design Of Experiments). Furthermore it is often challenging to convince Quality Engineers.

This situation makes the object of a business case to decide if the difference between the yield gain and the cost of resources and material for design of experiment is worth the investment (see table 4). It might happen that any invalid rejects are admitted since, even if the solution is known, its implementation would be too expensive with regard the investment.

Table 4: This table sums up manufacturing costs in cents per unit for one high runner. "Intermediates" is the cost linked to the silicone. "Raw material" is the raw material necessary for the assembly of the part (gold wire, moulding compound, etc…). All the costs are weighted with the production yield. This is to illustrate that it is more advantageous to stop part before assembly to optimize the profitability.

Direct Materiel	Intermediates	Raw Material	Machine Costs	WT	Pre-Assy	Assy	FT	Packing
37.55	32.56	4.99	9.45	4.82	0.26	1.72	2.44	0.20

4.4 Why did the problems seen during the VRU become more and more problematic?

Some of the products manufactured by our firm are contained in car keys. With the evolution of features, products have become more sophisticated. From a unidirectional low-frequency signal for simple information, the communication became radio frequency to contain much more information and as well as bi-directional between the car and the car key.

Most tests are performed to test the sensitivity of the chip, the power and the purity of the transmitted signal, and its consumption during both operation and sleep mode. Leakage tests are done as well to detect leaky transistors which might degrade more quickly than usual. These are generally analog signals. Though failures can occur with outlier readings, most failures are marginal. Examples of marginal rejects could be: a slightly too low a transmitted power, too low a sensitivity level or slightly too high a power.

In addition, products contain several different types of memory space such as ROM, RAM and flash. This is where the boot-up sequence, serialization and other firmware are stored by the customer. To test the functionality, a retention test and a memory margin test are carried out. These are mostly functional test or digital tests. For this category of tests there is no test limit but instead a positive or negative reply from the unit to a series of questions.

Eventually, the pressure of the market has direct consequences. For example it obliges one to shorten the development phase giving less time to industrialize the test process. Market forces also require that more and more advanced features be constantly developed with the same constraint in terms of quality level. All in all, the result is that any test processes becomes more complex but also less mature at the release of the product to the market.

4.5 Description of the existing VRU process

A process of sorts was in place even before the work I initiated to improve the VRU. An existing company document gives guidelines on the content of inputs and outputs regarding that process and requirements. The purpose of the procedure is not to provide the tools.

The existing VRU process is generic to the full organisation. It gives for each step during the development of the product required information, reports and deliverables needed to pass the gate and go to the next one.

A process is a sequence of interdependent and linked procedures which, at every stage, consume one or more resources (employee time, energy, machines, money) to convert inputs (data, material, parts, etc.) into outputs. These outputs then serve as inputs for the next stage until a given goal or end result is reached (see figure 28).

For the VRU process during the early development phase, deliverables mainly concern sample plans. The goal is to identify the state of the volume forecast, to establish if the supply chain is properly conditioned to enable the sourcing. Based on the sample delivery plan, one can ask whether the production should be done in a subcontractor FAB to enable a higher throughput or if the assembly and test plant are big enough. These are the kinds of questions that are addressed during this initial phase of the development of a new product. The business plan is still very hypothetical and based on the prediction of the market. But if a bottleneck is identified it has to be addressed upfront as it takes to time to release a new supply source. Therefore, even if the assumption is very hypothetical, it is important to carry out this exercise early during the development phase. This exercise is followed by a Failure Mode Effects Analysis (FMEA) to identify risks on delivery throughput due to innovation embedded into the new product. “Does the new package type have a potential impact on throughput?” “Does the process in FAB have a longer throughput? What is the maturity of the process? ”The risk of each item of the checklist is weighted. The degree of innovation will directly therefore impact the risk.

This VRU FMEA (see figure 26) comes during the specification phase of the project, i.e. considering a development phase of a product lasts roughly two years, this FMEA comes in the first six months. During the next twelve months, no real activity will be performed VRU wise except the regular update of the VRU FMEA to re-evaluate risks.

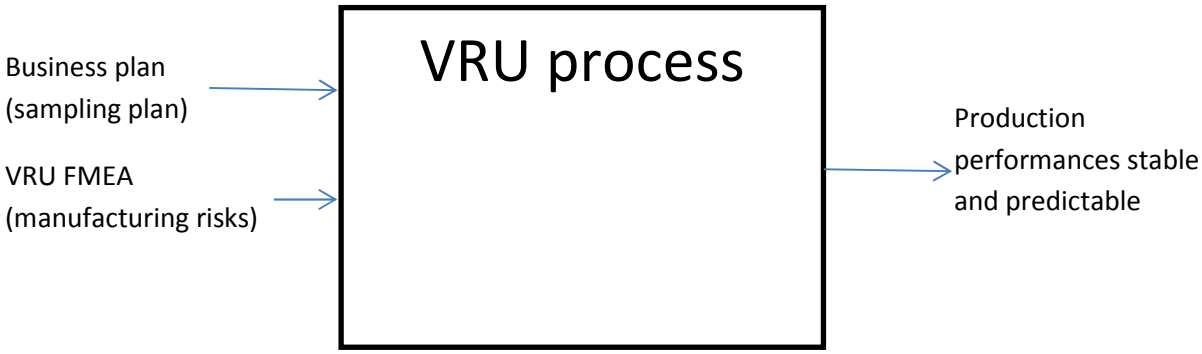


Figure 26: Illustration of the VRU process with its input and output parameters.

The next real step of the VRU preparation coincides with the start of the qualification of the product when the first samples are available. From this stage the first statistical test data generated on tester are then produced. First loop runs and test data from the initial sample delivery give a first view of the test maturity. The test coverage is then partial and will be developed week after week. This is one of the major difficulties faced during the VRU preparation: the test process will constantly be modified, improved and developed until the release of the product. This means that even when the product is released into production, the knowledge on the maturity of the test process in still incomplete. Despite this, more than 90% of the test coverage is frozen. A few test parameters might still be missing but the remaining 10% mainly concern limit optimization. It means stability and test process capability has already been given upper and lower bounds.

The real active part of the VRU comes after the product is released. The project is then officially closed almost immediately afterwards. The development team is then officially allocated to another project to hand over the product industrialisation to Operations. The product engineer who was in charge of the qualification during the development phase takes the ownership of the product. It means he or she is in charge of facilitating the release of the material with the guarantee that the quality level of the delivered material be constant and compliant with the datasheet.

His or her responsibility is also to set hold lot criteria. Too tight hold lot criteria means a high hold lot rate increase, which slows down the daily throughput and increases the engineering workload. With too loose hold lot criteria one could run the risk that an

abnormal sample batch is released undetected. The product engineer is also in charge of planning yield improvement activities.

The product life for the automotive market is greater than ten years. It takes so long for a car manufacturer to release a new module that, once the module is stable and has proved its reliability, it will continue to be used for years. Indeed some of our own products have been in use for twenty years. The other particularity already mentioned is that high volumes are not produced until two years after market release (figure 27). These two years are used by the customer to validate the new product into their platform before release it into a new car model. During this period volumes remain sporadic.

This period can be also called the industrialisation phase: transforming an immature test process into a stable and predictable one. It could also be dubbed the learning phase. Every new production batch reveals weaknesses undetected during the development phase. This represents a heavy load in terms of follow up. The VRU Manager's role is to make the link between logistics, marketing and the Operations during this phase. Based on the order book, he or she drives priorities in FAB, in the test development team and the product engineering team: based on the order book for the next quarters and the actual throughput he or she somehow steers activities to prevent there being any delay in deliveries. The "basic" VRU process provides overall guidelines on how to conduct the follow up of VRU types (PDCA approach).

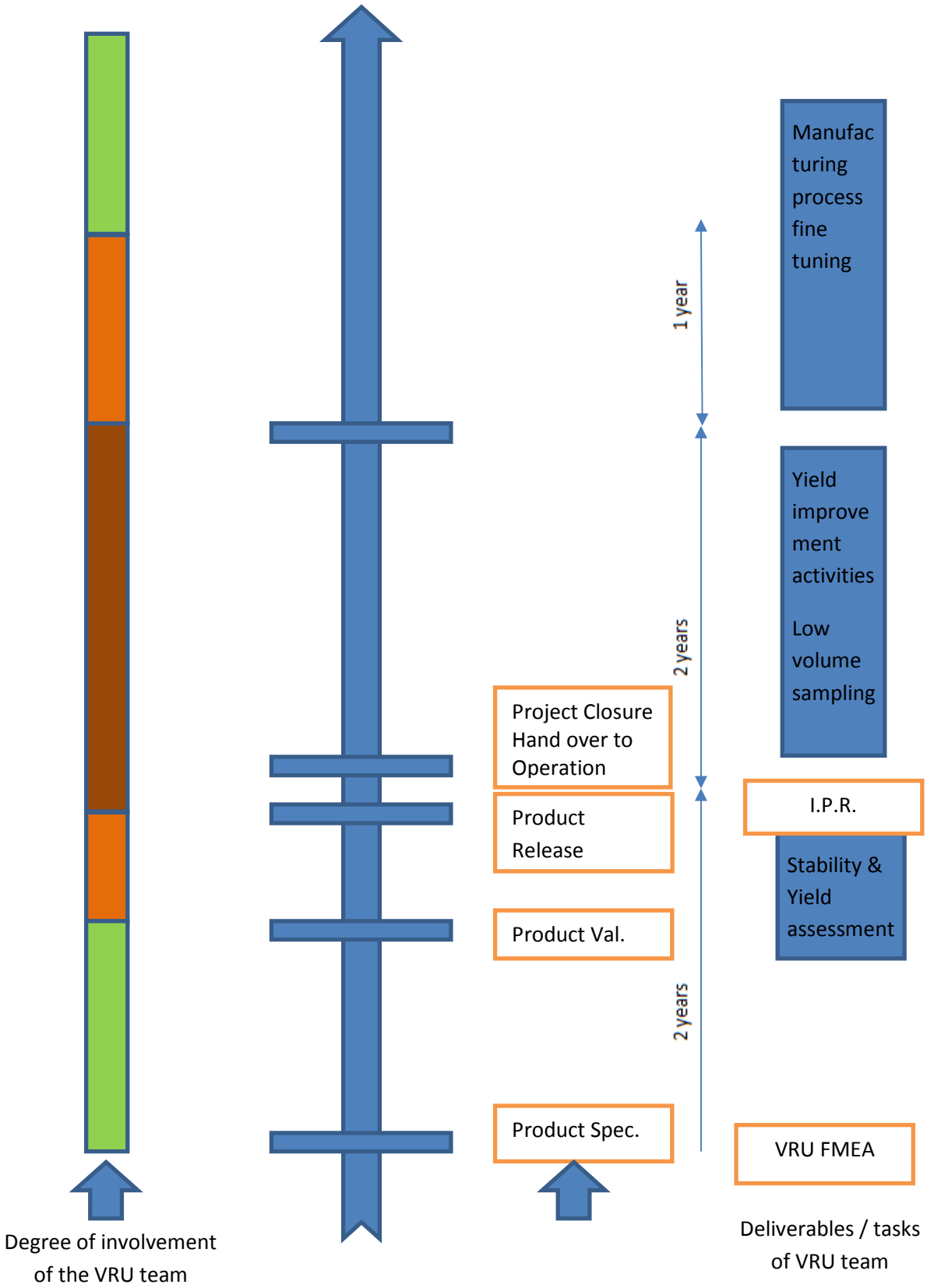


Figure 27: Representation of the level of involvement of the VRU team along the development of the project. The VRU team involvement starts mainly after the release of

the product which is somehow too late as the datasheet has then already been frozen and resources already lessened.

5.The solution

The elaboration of the solution to improve the overall VRU process started with the will to include before the release of the product what used to constitute yield improvement activities after release. The reasons for this have already been mentioned but are recalled here:

1 – To begin with better yields at the production start

2 – It is much easier to make change to a datasheet before it is released than to convince the customer and the quality department to change it several months later

3 – If a high risk of yield loss is identified before the product release, a datasheet limit change request is still feasible.

Now, let us see the usual approach to lead yield improvement activities.

5.1 Reminder of how to calculate a Cpk

The process capability is a measurable property of a process to the specification, expressed as a process capability index (*e.g. Cpk* or *Cpm*) or as a process performance index (*e.g. Ppk* or *Ppm*). The output of this measurement is usually illustrated by a histogram and calculations that predict how many parts will be produced out of specification (OOS).

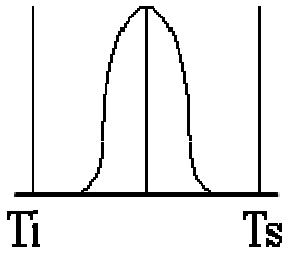
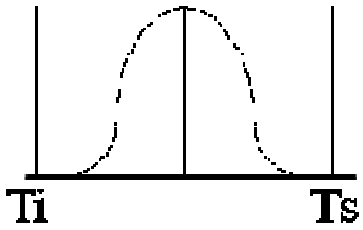
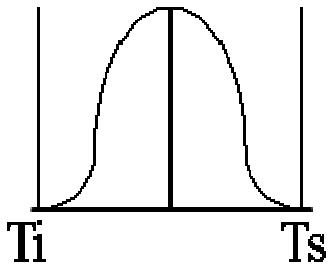
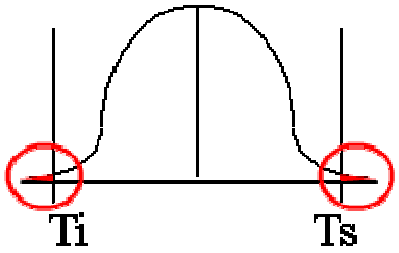
A process is deemed to be capable if the spread of its results is weak compared with the tolerance. The process capability is the ability, the capacity of a process to be compliant to specifications, to reach permanently the desired quality level. The indicator of the process capability (*Cp*) is a measure of the performance of a process with regards to acceptance limits. This dimensionless number shows the ratio between the spread (variability of the process) and the tolerance range. The higher the number, the more “capable” the process.

The capability (denoted by *Cp*) is calculated via the ratio: tolerance range / spread (figure 28)

$$\hat{C}_p = \frac{USL - LSL}{6\hat{\sigma}}$$ Equation # 7

$$\hat{C}_{pk} = \min\left[\frac{USL - \hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu} - LSL}{3\hat{\sigma}}\right]$$ Equation#8

5.2 Link between the *Cpk* and the yield performances [8]

1	Cp > 1.67		More than enough
2	1.67 > Cp > 1.33		Ideal situation. To be sustained
3	1.33 > Cp > 1.00		Requires attention. Any drift might be a source of reject
4	1.00 > Cp > 0.67		Existence of non-conformity. The process has to be analysed and improvements have to be found.

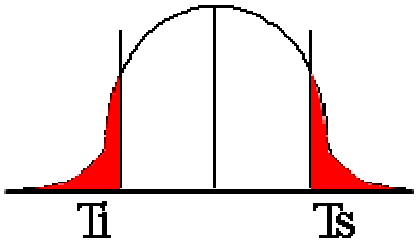
5	0.67 > Cp		It is urgent to define corrective actions.

Figure 28: For a Cpk above 1.67 (equivalent to 6σ) for a normal distribution, statistically a yield is 99.9999%. This is globally what is targeted.

At the start of production, the variability is high as the number of production batches is still limited. The longer production lasts, the more normal the distribution per test parameter becomes. Indeed many variables can influence the shape of the population: tester-to-tester variation, board-to-board variation, site-to-site (parallelism) variation, production batch to production batch variation. This might create multiple distributions. *Cpk* calculation on a non-normal distribution is not applicable. It takes time for all those sources of variability to be integrated and smoothed. For this reason, distributions at the start of production are rarely normal. Despite this, a *Cpk* of 1.67 ensures that yield loss per test parameter will be limited.

Physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have distributions that are nearly normal. Moreover, if the relevant variables happen to be normally distributed many results and methods (such as the propagation of uncertainty and least squares parameter fitting) can be derived analytically in an explicit form.

The heart of the problem during the VRU phase remains the production yield. A poor yield requires much more material to be used to compensate. An unstable yield could mislead the logistic if the output quantity is not at the level as given in their database. A yield not at the expected level is also source of hold lots.

To address the problem of the yield at the production start, the approach chosen to improve the VRU process is somehow to transfer activities performed as yield improvement, before production begins.

5.3 Activities for yield improvement

The key element is to build a representative picture of performances in production from a limited quantity of a sample batches. (Reminder: only three sample batches are tested at the release gate). The goal is to perform the exercise for every parameter that make up the test sequence. A final test sequence is made up of several thousand test parameters.

A frequent mistake is to address only the test showing yield losses (figure 29) rather than all test parameters with low *Cpk*. Focusing on tests showing yield losses looks efficient as it gives an immediate improvement but eventually tests with low *Cpk* not showing yield losses represent a potential yield loss. Statistically this will show up with the drift of external variables over the time. For this reason, it is important to make an exhaustive review of all test parameters. It gives also a vision per test block.

Description	Bin #	Device qty	%	Sublot qty
bin_1	B_1	84016	95.80%	800
bin_21	B_21	1934	2.21%	626
bin_24	B_24	1149	1.31%	596
bin_5	B_5	245	0.28%	206
bin_14	B_14	222	0.25%	195
bin_12	B_12	45	0.05%	27
bin_13	B_13	41	0.05%	40
bin_20	B_20	21	0.02%	21
bin_11	B_11	13	0.01%	13
bin_27	B_27	10	0.01%	10
bin_25	B_25	4	0.00%	4

Description	Test #	Device qty	%	Sublot qty
NV_AppSetPf_Min	T_2982	849	0.97%	118
Trim_AllNV	T_1094	818	0.93%	489
NV_AppSetUlp_Min	T_3012	251	0.29%	219
IDDQ_Mem1	T_2841	215	0.25%	190
ChkBrd	T_2143	175	0.20%	154
Trim_AllNV	T_1097	140	0.16%	131
MarginWr_CB_EE	T_9124	121	0.14%	113
RomSig	T_850	120	0.14%	110
IsAlive	T_359	119	0.14%	111
TrimCoreBias	T_1002	68	0.08%	67
MarginEr_PF	T_2044	63	0.07%	60

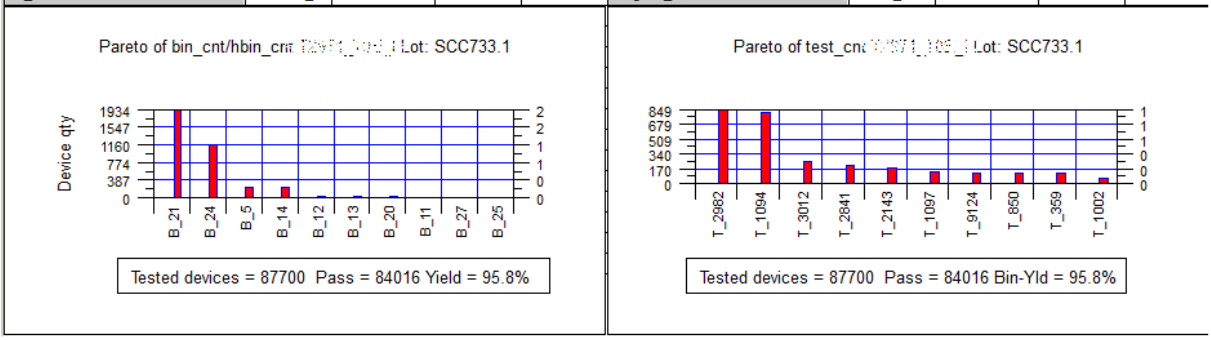


Figure 29: Failure Pareto diagram for a full wafer batch (87 700 tested dies from 25 wafers) The left side of the figure shows a failure Pareto per bin number. The test parameters are grouped per bin category. When a test parameter fails, its test number is recorded in addition of its bin number. By this way, it shows immediately the type of the reject. Bin#1 are for good dies. On the right side the failure Pareto per test parameter. “Description” is the test name; “Device Quantity” gives the quantity of failing dies per test parameter. The percentage is made out of the total input quantity (87700)

The first level analysis is done by gathering all test data, listing all test parameters and for each of them calculating the *Cpk* and potential yield losses. The *Cpk* is calculated for good parts only so as not to include outliers which would bias the mean value and the standard

deviation. Functional tests cannot be targeted by this approach as, by definition, populations for these tests are not normal but bimodal. This does not mean functional tests are not to be optimized but just that the approach is different.

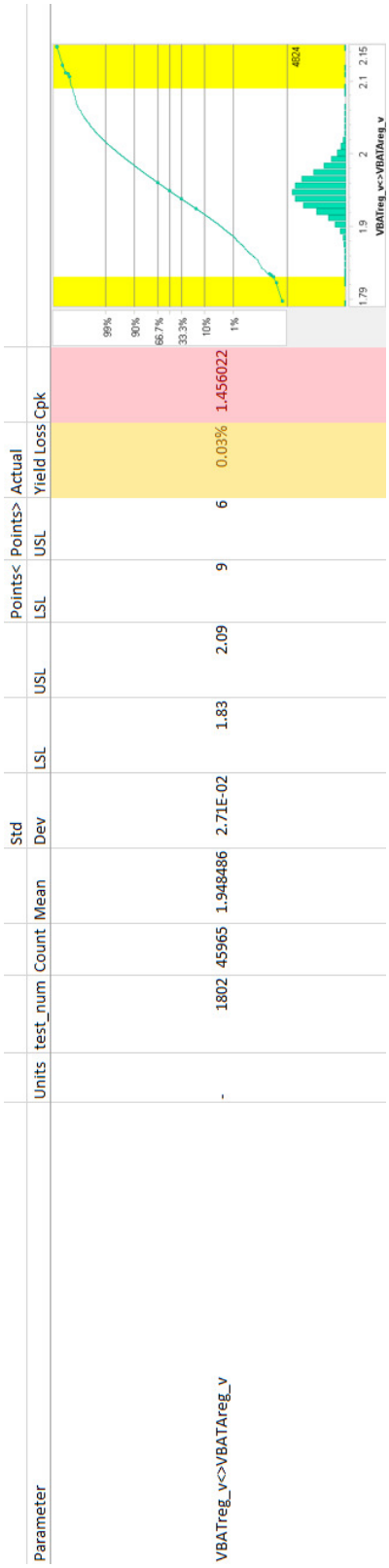


Figure 30: This is an extract of a Cpk report for a voltage regulator. “Count” gives the number of samples composing the population.

Tests with *Cpk* greater than 1.67 are then considered robust enough. Those tests with a *Cpk* lower than 1.67 deserve attention since, if no rejects were seen for the first three batches, statistically they represent a risk of yield loss.

The first thing to do is to check whether tests limits are correct and if it is a specification parameter. Indeed, it can happen that test limits were defined in an arbitrary way based on a very small sample size. If it is not a datasheet parameter but simply a test parameter for monitoring for instance, then a yield loss is not justified. In that case, whatever the shape of the distribution, if the limits can be widened in order that the *Cpk* reaches 1.67 then the solution is immediate. By chance this applies to roughly half of the tests with a low *Cpk*. Then is there a notion which cannot be quantified: engineering judgment. Particularly in cases where the population is not normal, the experience of the engineer makes a difference. It happens that sometimes *Cpk* are left intentionally low in order to screen potential outliers. This is also to illustrate that there is a case-by-case analysis which is hardly describable. But generally speaking, in the case of tests for which a limit widening is possible, limits are then adapted in order to produce a *Cpk* of 1.67.

For tests with low *Cpk* for which limits cannot be widened, the method consists in looking for a modulator is, a parameter that influences the population. This is also the method used to investigate how to improve functional tests showing yield losses. Let us look at some practical cases.

5.4 Main cases of distribution seen in production

Cpk can only be applied to normal distributions. Given distributions seen in production are not necessarily normal, let us see how they can be categorized and interpreted. This is the object of figures 31 to 34

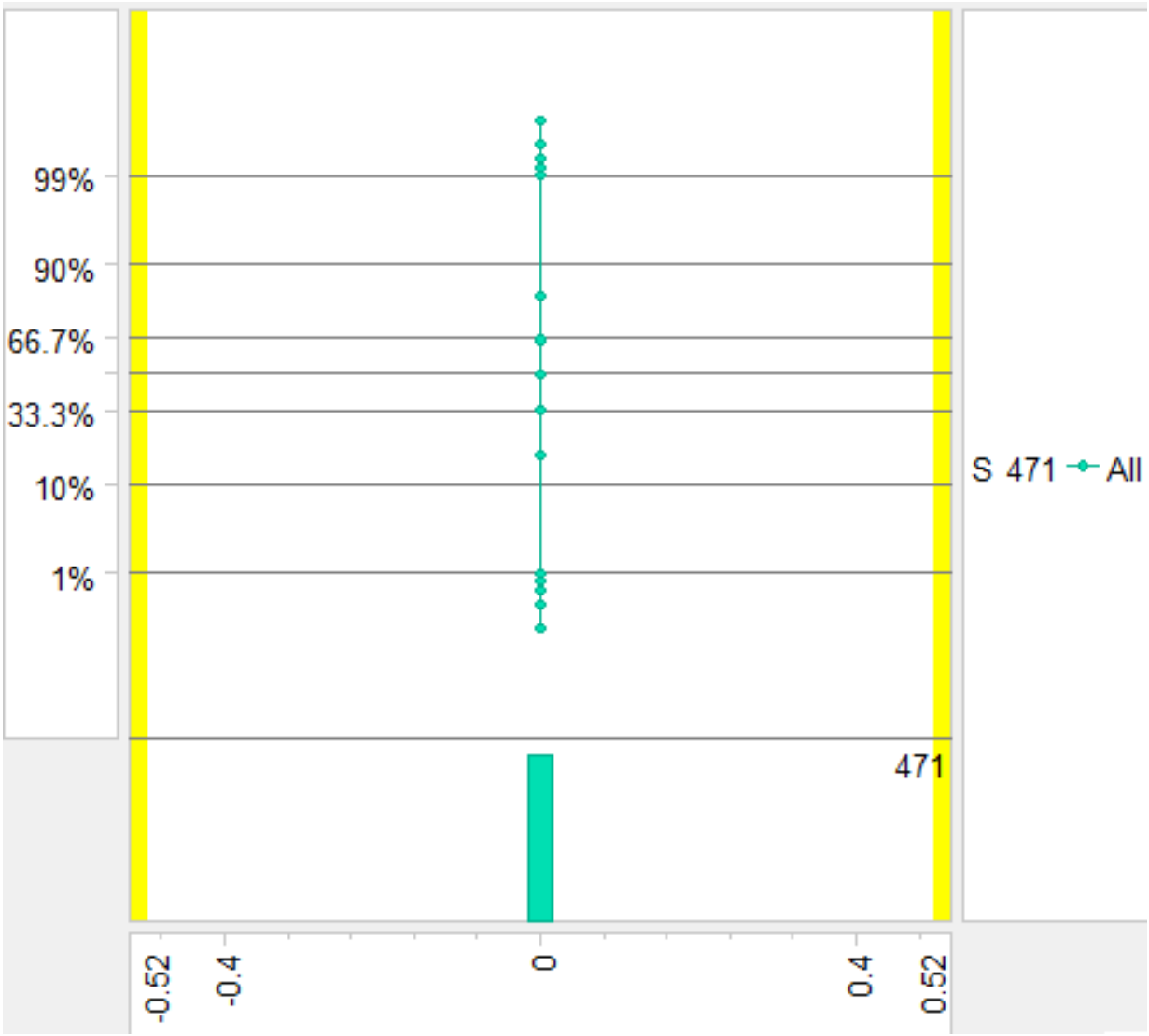


Figure 31: This is the distribution of a typical functional test parameter. Test limits are set to [-0.50;0.50] as the only passing value is null. When the test parameter fails, the reading is 1 or the failing vector. Limits are here not significant.

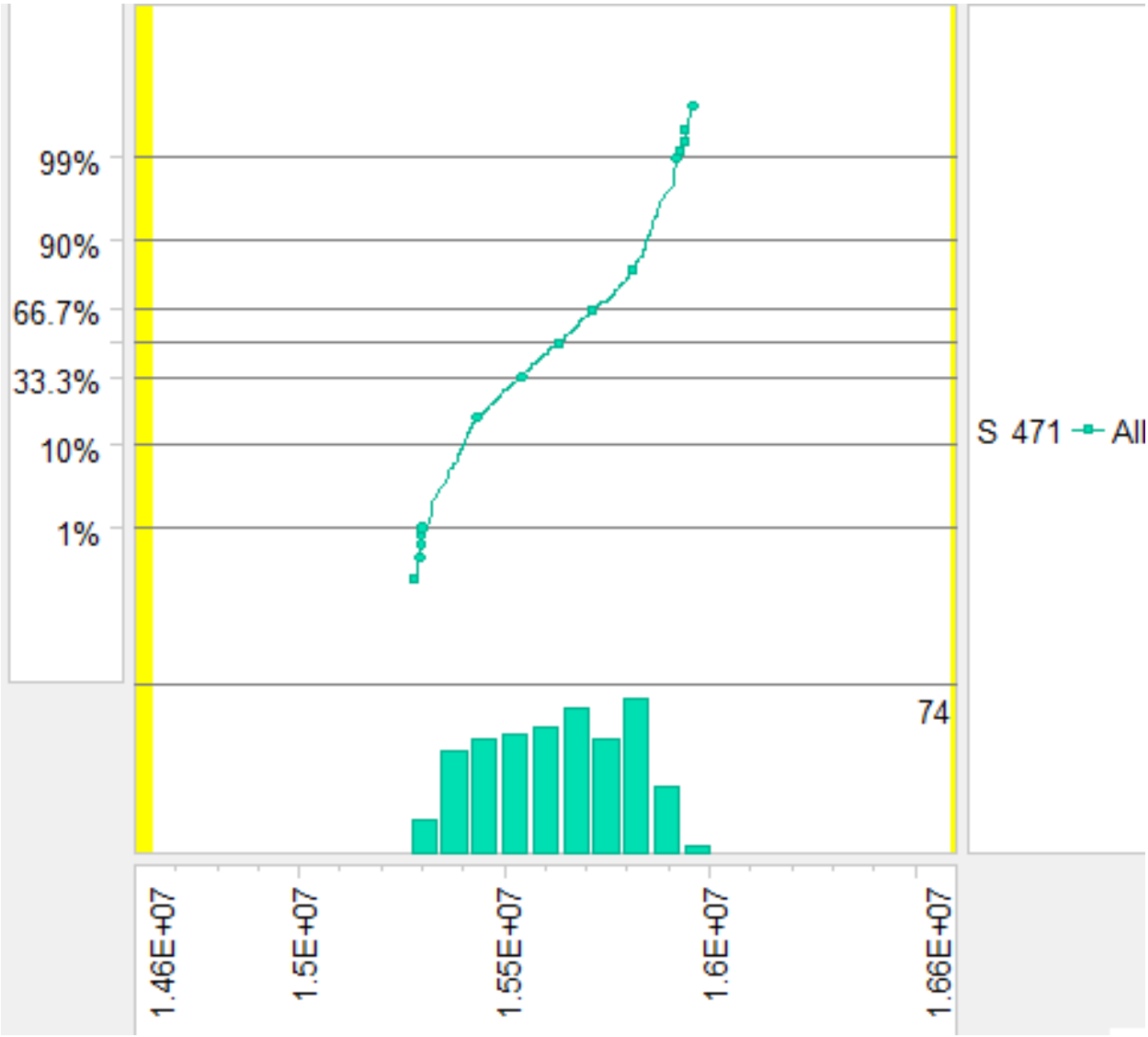


Figure 32: Type of distribution representative of a trimmed value. In this instance limits could be tightened without risk as this type of test parameter is insensitive to test environment variations or drift in FAB.

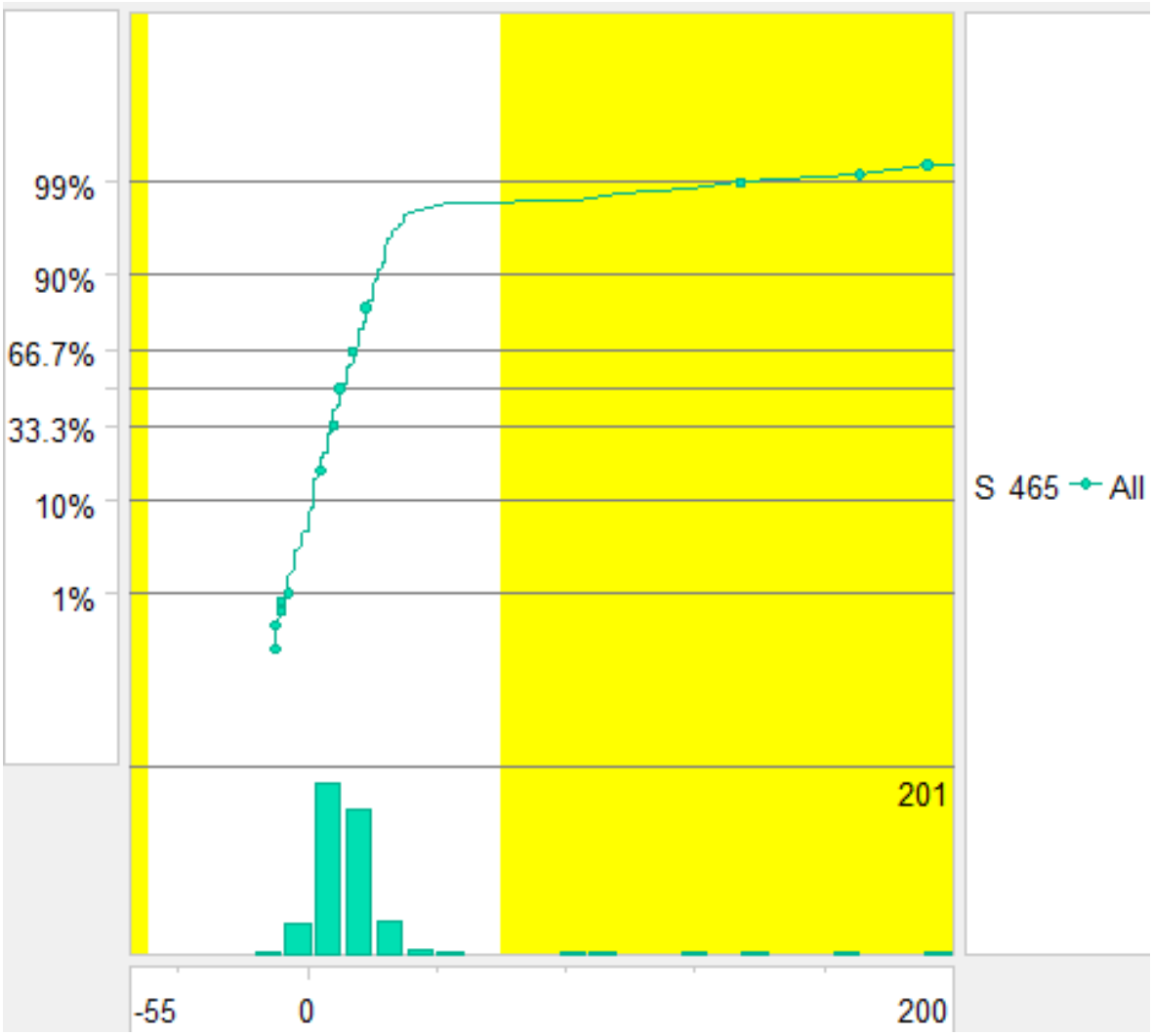


Figure 33: Typical normal distribution. Here the cumulative plot is rectilinear for a logarithmic Y-axis. The limit is well defined as it screened outliers.

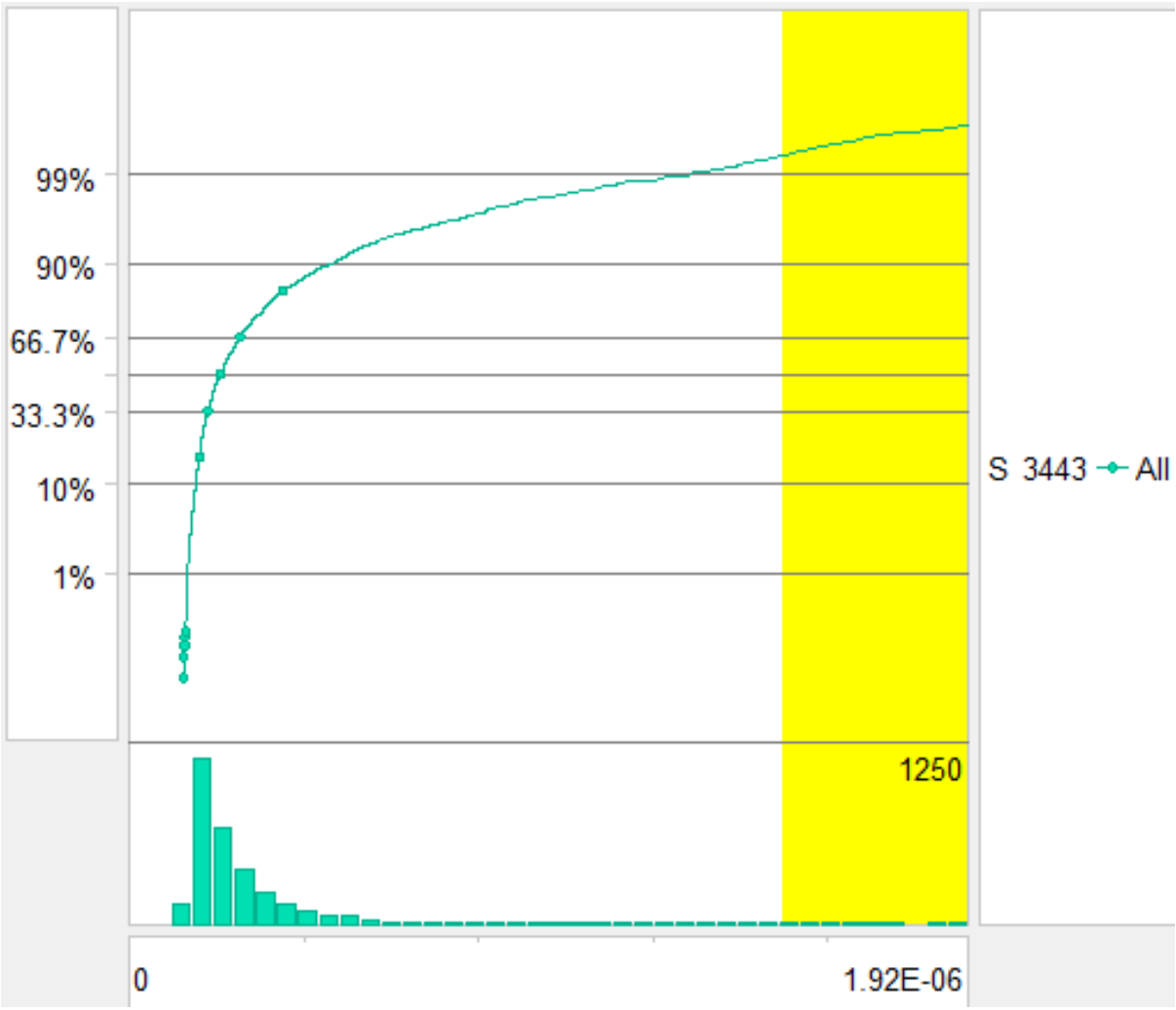


Figure 34: The particularity of this distribution is its long tail. All the difficulty lies in the definition of the upper limit to screen rejects properly and prevent over screening. Here the *Cpk* is not applicable. One method is to work with the logarithmic value recreating a normal distribution.

5.5 Detection of variables impacting the distributions

In fact, among the test sequences, the proportion of test parameters with a normal distribution lies in the minority. Despite this, *Cpk* analysis coupled with the test failure Pareto diagram remains a good indicator of the capability per test parameter. The diversity of cases means there is no systematic approach to improving the capability of a test process; especially at a production start while the impact of each variable is not smoothed. Statistically, thirty samples batches are necessary to provide a representative picture of the

variability. Nevertheless, before production was started in volume only three batches were released. Based on available data, tests showing a potential risk of yield loss have to be detected as early as possible. It induces that each variable has to be weighted and its impact contained (figure 35 to 38). The goal as already mentioned is to prevent invalid rejects.

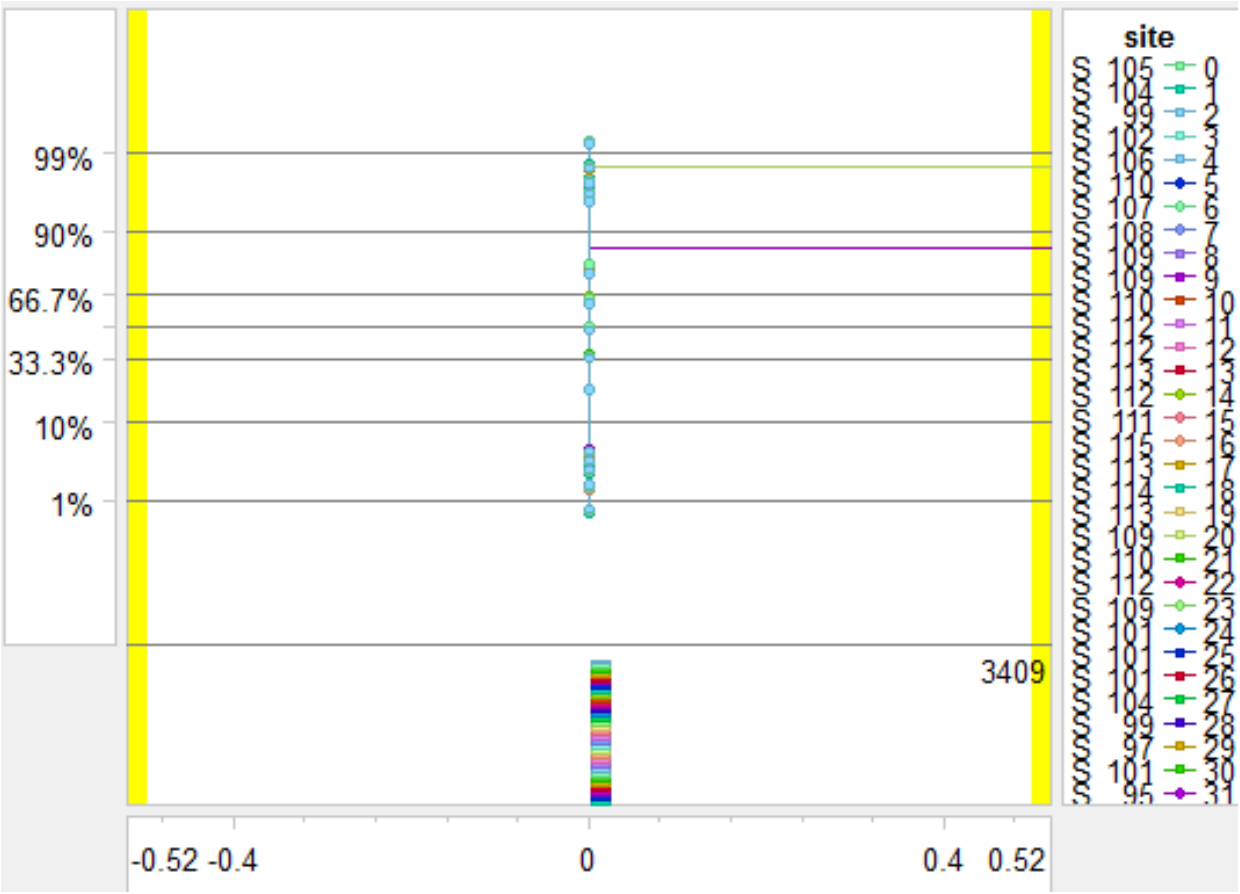


Figure 35: It was already stated earlier that functional rejects are valid rejects by opposition to marginal rejects. Nevertheless, there are functional tests sensitive to test environment. In the example above, are shown yields per test site. Test processes are made in a way several units are tested simultaneously to gain in throughput time. This is what is called the parallelism. In this example there are 32 tests sites per test flow. Here data are extracted from a wafer test flow. This means thirty two dies are tested in parallel. After investigation it as revealed that site #14 and #31 had a much higher yield loss than the others. As this was seen on several probe cards, it was concluded there was a fundamental problem with the test process and the test method was changed to correct the site to site discrepancy.

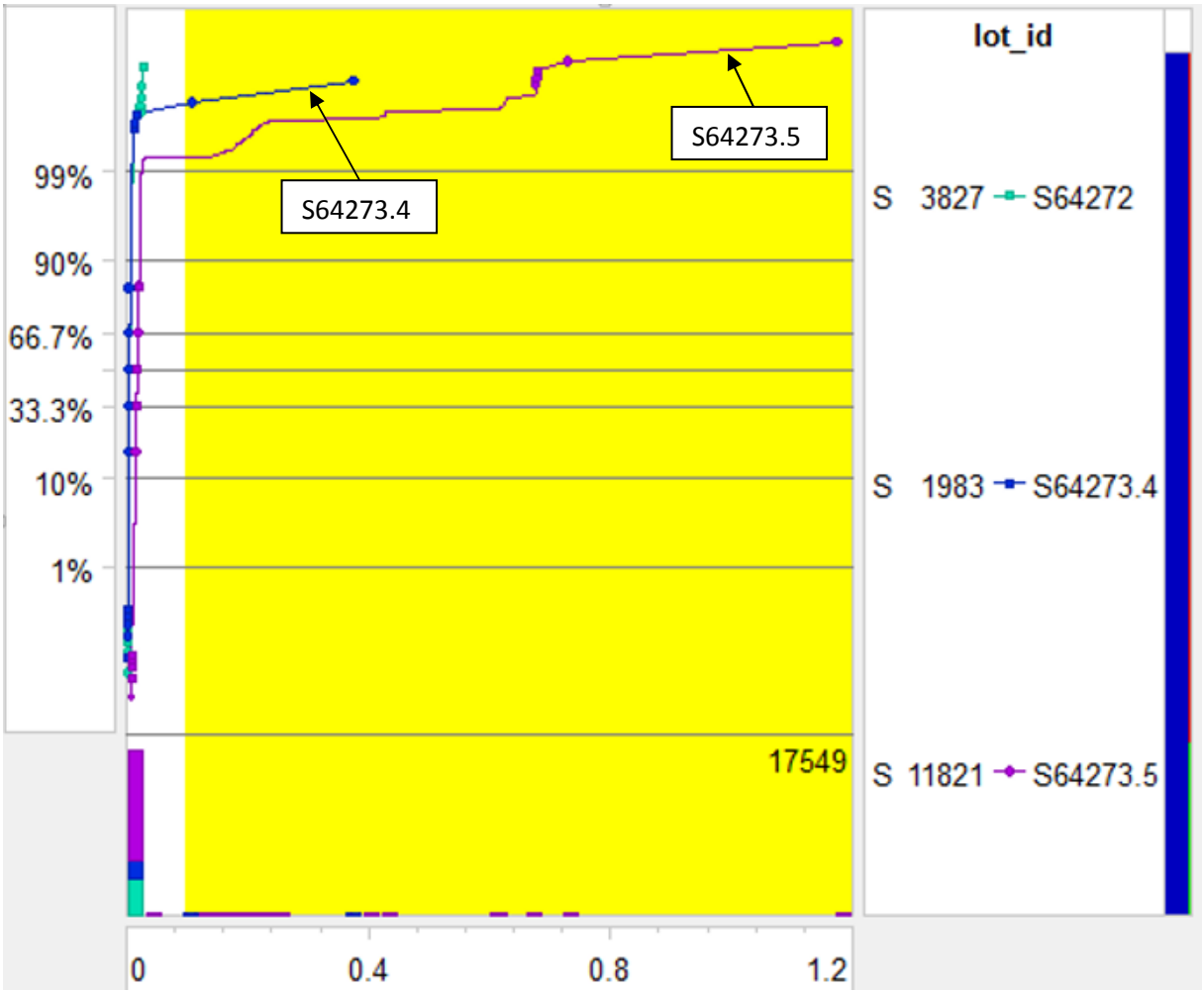


Figure 36: Here is a parameter showing a long distribution tail as well as outlier units. A commonality analysis was conducted to look for a modulator. By scanning all potential contributors, it appears that the yield loss is due to one particular wafer batch (lot_id) out of the three. The one is purple is the one responsible of yield losses (roughly 1%). This information was fed back to the FAB. After another commonality analysis in fAB, the defective machine in FAB was found and corrected.

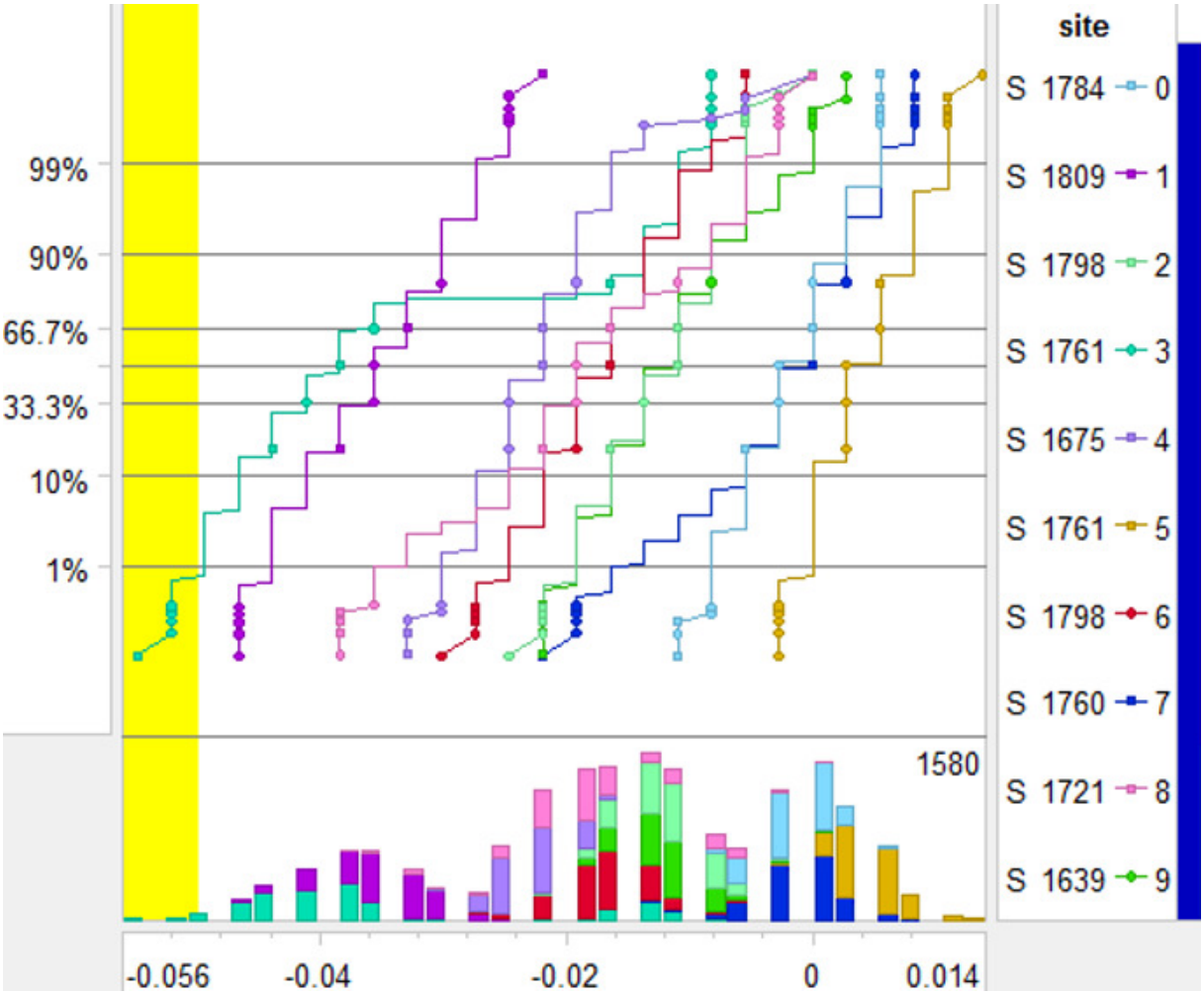


Figure 37: When that parameter was analysed for the first time it was showing a *Cpk*<1.0. After the extraction of data, it appears that the reason for this excessively large standard deviation is a test site-to-site dependency. The test method was subsequently reviewed before production started. In that case, yield losses were related to a bad test process variability.

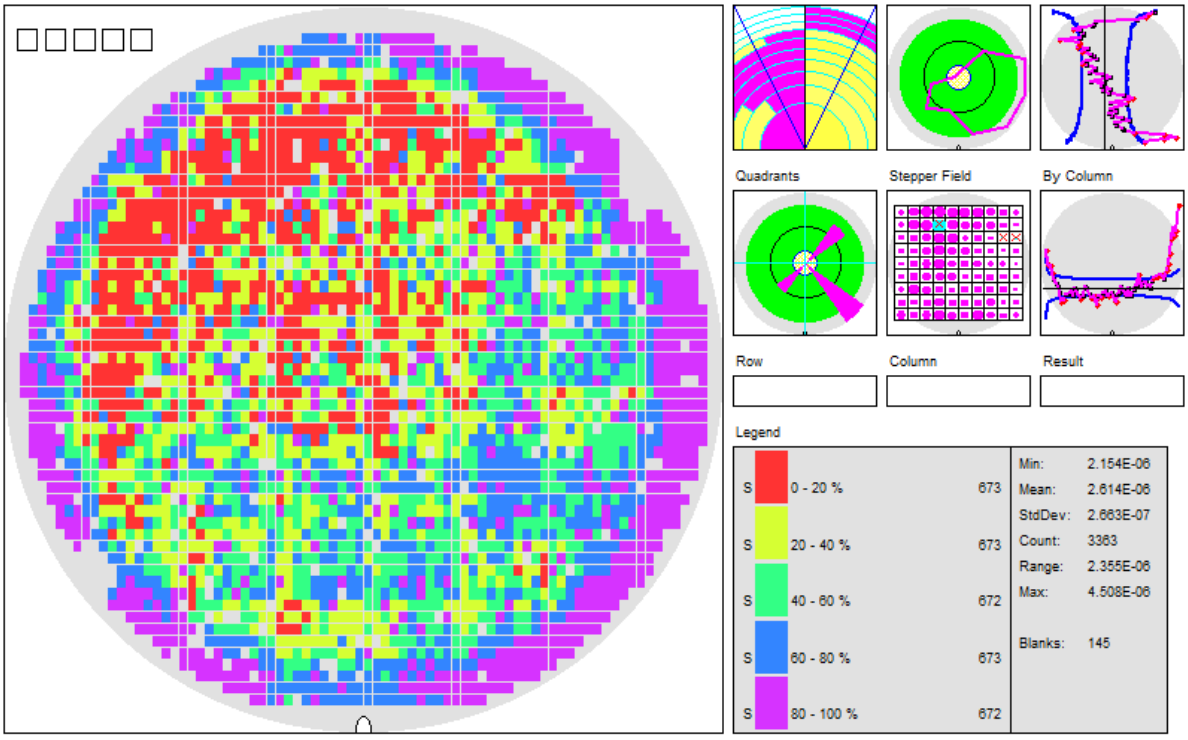


Figure 38: Another potential contributor to be considered is the variability of the parameter over the wafer. In this example the value of the parameter is higher on the edge of the wafer. That is an important information the FAB can use to better control their process.

A complementary indicator to the *Cpk* analysis is the recovery rate per test parameter. Generally the rejects are retested to improve the production yield and recover potential invalid rejects. This retest of rejects really affects machine efficiency. If the yield at the first pass is 80% and then 50% of rejects pass after retest, the final yield is 90%. The interest here is to ensure the first pass yield is as high as possible. In addition, as already mentioned previously in the section 3.9, even if there are rejects passing after retest, the quality of these recovered parts is questionable. Having a high recovery rate as well is also a potential alert for each good part. A high recovery rate is a sign of the instability of certain test parameters; and the instability can cause rejects to pass.

For all these reasons comparing, the test reject Pareto diagram before and after the retest of rejects is an important indicator (see figure 42).

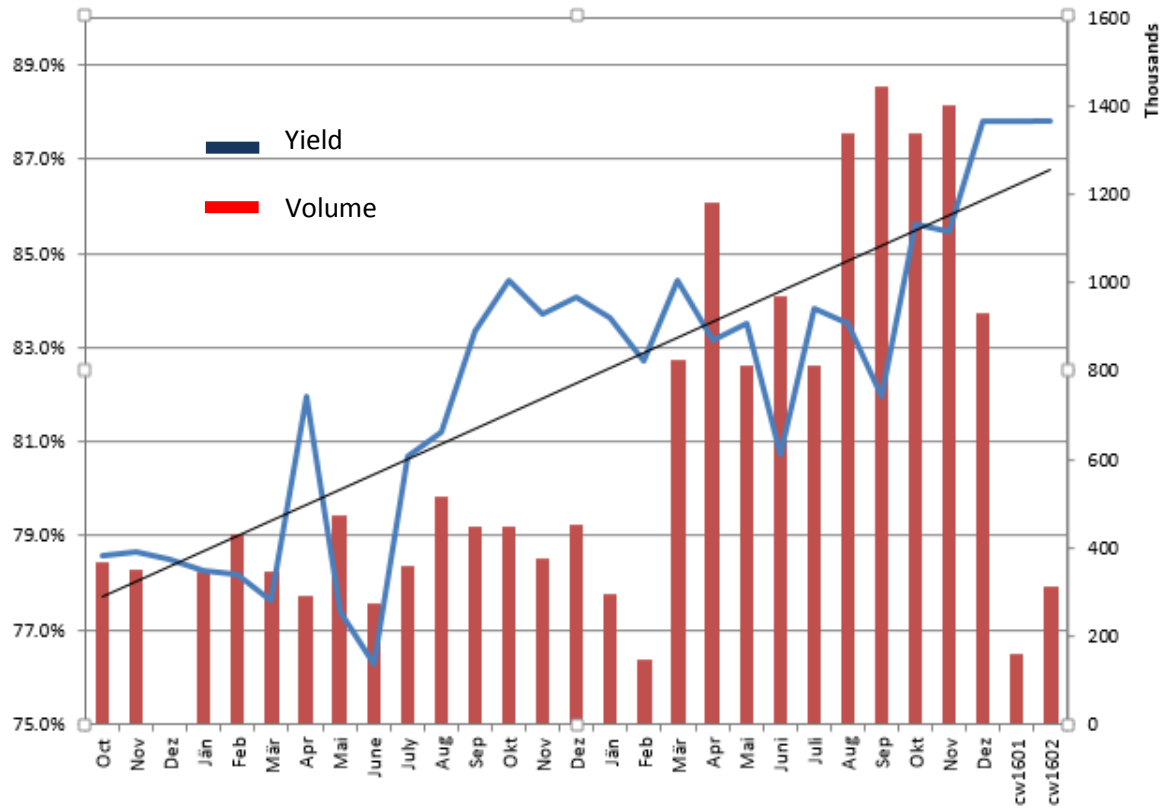


Figure 39 : Yield trend in wafer test for one of our higher runners. It is thanks to yield improvement activities that performances are becoming stable and acceptable.

5.6 Deming Wheel approach

The goal of the *Cpk* review is intended to improve finally the capability of the test process. By this approach, production yields are greater, more stable and more predictable. This does not mean the exercise is radical. The process has to be repeated (figure 40). Let us see why this is important.

5.6.1 The learning phase

At the production start, variables impacting performances such as wafer-batch-to-wafer-batch variability and/or the tester-to-tester variability are not smoothed enough and each sample lot might show a behaviour not seen during the validation. This requires a large sample batch quantity [9]. It will take several months until the variability becomes totally smoothed and distributions become normal. For this reason, each *Cpk* analysis revision will reveal an evolution compared with a previous review.

5.6.2 Non-normal distributions

It can be misleading to apply a *Cpk* to a non-normal distribution. Nevertheless, with a correct interpretation, the *Cpk* remains a good indicator of the capability of the test process. But due to the approximation, if one repeats the exercise in such a case with different sample batches distributions might show different shapes, different problematics and potentially fluctuating yield losses. To detect process drifts and fine tune the capability of the processes it is recommended one redo the exercise on a quarterly basis.

5.6.3 Side effects of only focusing on test failure pareto and not on the full process capability.

One notion has been omitted thus far: testing stop on fail. To save tester usage, once a test parameter fails, the test sequence stops. It might happen that, to learn more about our rejects upon request the test sequence is not stopped but this is not standard. The risk of testing in "Stop on Fail" mode is that if a test showing failures is improved, rejects are reported on following tests in the test sequence. This happens when two test parameters A and B are correlated and when only improvements are brought to parameter A. Then yield losses will be carried over to parameter B.

The test failure Pareto diagram might also evolve when parameter is tuned in FAB. In the test process, hen a new test parameter is added it impacts the test failure Pareto diagram as a side effect. Despite Design Of Experiments and Measure of System Analysis (DOEMSA) as preliminary validation, each peripheral parameter remains barely detectable.

5.6.4 The natural drift over time

Equipment used in production is regularly calibrated during preventive maintenance operations. This ensures that production batches remain within control limits with a limited spread and that test equipment induces neither over acceptance nor invalid rejects.

All parameters are not checkable during the preventive maintenance operation. However test results might drift over time due to the degradation of the machine. For this reason, regular monitoring of performances in production is necessary.

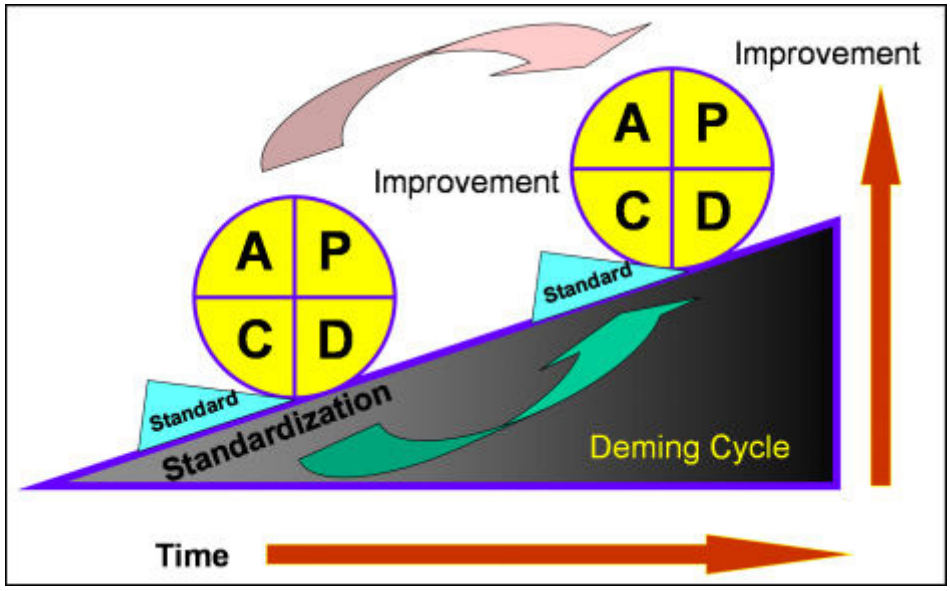


Figure 40: Plan, Do, Check, Act is the representation of the Deming cycle. Following this continuous improvement approach, performances of the process are globally improved.

5.7 Definition of indicators targets and rolls out to the team.

The goal of the research was to improve the VRU process so that early samples deliveries arrive on time with the required quality level. The production yield is the key element as it directly conditions

- the hold lot rate,
- the predictability of output quantities with regard to the material at the input of the production chain,
- the level of confidence in the delivered material,
- the profitability of the product,
- the throughput time.

Another parameter which strongly affects the throughput time is the test time per unit. Other parameters have an impact on the throughput time. Most of them were introduced earlier in this report. This gives a matrix of measurable indicators which somehow characterizes the producibility of the process. They are summarised in table 5.

Table 5: List of KPI necessary to properly monitor performances during the VRU phase and detect the origin of variations in the delivery throughput

KPI	Target	Description
Retest Recovery rate	<10%	Percentage of rejects from first pass, recovered after retest. It is an indicator of the stability of the test process. Recovering retest improves yield performances but directly impacts tester efficiency. It is also an indicator of the quality level of the production.
Comparison failure Pareto 1st pass vs retest		Recovery rate per test parameter between first pass and retest. Linked to the retest recovery rate, it pinpoints tests with high recovery. It identifies unstable test parameters requiring special attention.
Failure pareto acceptance test		When retesting good parts, only good parts should be obtained at the first attempt. If any parts require several trials to eventually pass it highlights potential unstable test parameters.
Cpk tracking table	1.67	This list tracks action items per test parameter showing low Cpk and/or yield loss. Even if Cpk and yield loss vary over time, it tracks decisions taken along the product life. The final goal is to identify test parameters impacting the capability of the production process. This is a document which as to be even regularly reviewed and more often during the ramp up phase when the variability on production is important.
Yield trend	Wafer T : D0 Final T : 99%	The yield is the ratio between the output quantity and the input quantity. The trend gives the indication of how efficient improvements are or conversely how any drifts of the process are degrading performances. It is the financial indicator. It drives the allocation of resources for yield improvements.

Daily output of good parts per machine		It is an indicator directly linked to the efficiency of the test machine. It is somehow a compilation of all other indicators. It is applicable to the wafer test and to the final test process. Main contributors are the yield, the first pass yield and the test time. But handling of parts by operators, the index time of the handler, the time needed for the maintenance (preventive and repair) also have to be considered
Hold lot rate	5%	Gives the success rate of lots passing acceptance criteria. Hold lot happens when performances are not compliant with a normal lot statistically speaking. Hold lot are mostly triggered by the interaction between the test process capability and drifts in FAB. This is why it is crucial to desensitize the production process from variations by optimizing the capability. Dealing hold lot highly consume engineering and tester resources.

6.Validation of the solution

6.1 A long incubation

The implementation of such a process is progressive until it reaches its maturity point. As of today, the VRU process is not yet as its maturity level. Firstly, as explained previously all parameters are not fully controllable. The reasons for this are multiple:

- A test sequence is composed of several thousands of test parameters with any of them correlating each other, some of them inversely proportional. Some other parameters are sensitive to the test environment. In this case, once in production where several machine and test hardware are used (i.e. digital scan test and transistor leakage), they are getting hard to be controlled. This is the difficulty brought by the test process requiring a high knowledge of the product and of the test methods. This is a condition to interpret the behaviours of test parameters.

The impact of what has been put in place by this project is to highlight these behaviours and lead the engineer to investigate them.

- Another reason why some elements are hard to control is related to the complexity of the FAB process. A single wafer is composed of roughly forty layers and its production involves a multitude of various tools and chemical products. Despite all efforts to enhance the capability of the manufacturing process, it is common to observe a sudden yield loss due to a drift in one of the tools. Isolating the root cause and correcting it can take months of engineering activity.

For both these reasons, having a VRU process able to integrate variations and reactive enough to solve issues on time requires an adjustment to the way of monitoring performances and improving the capability.

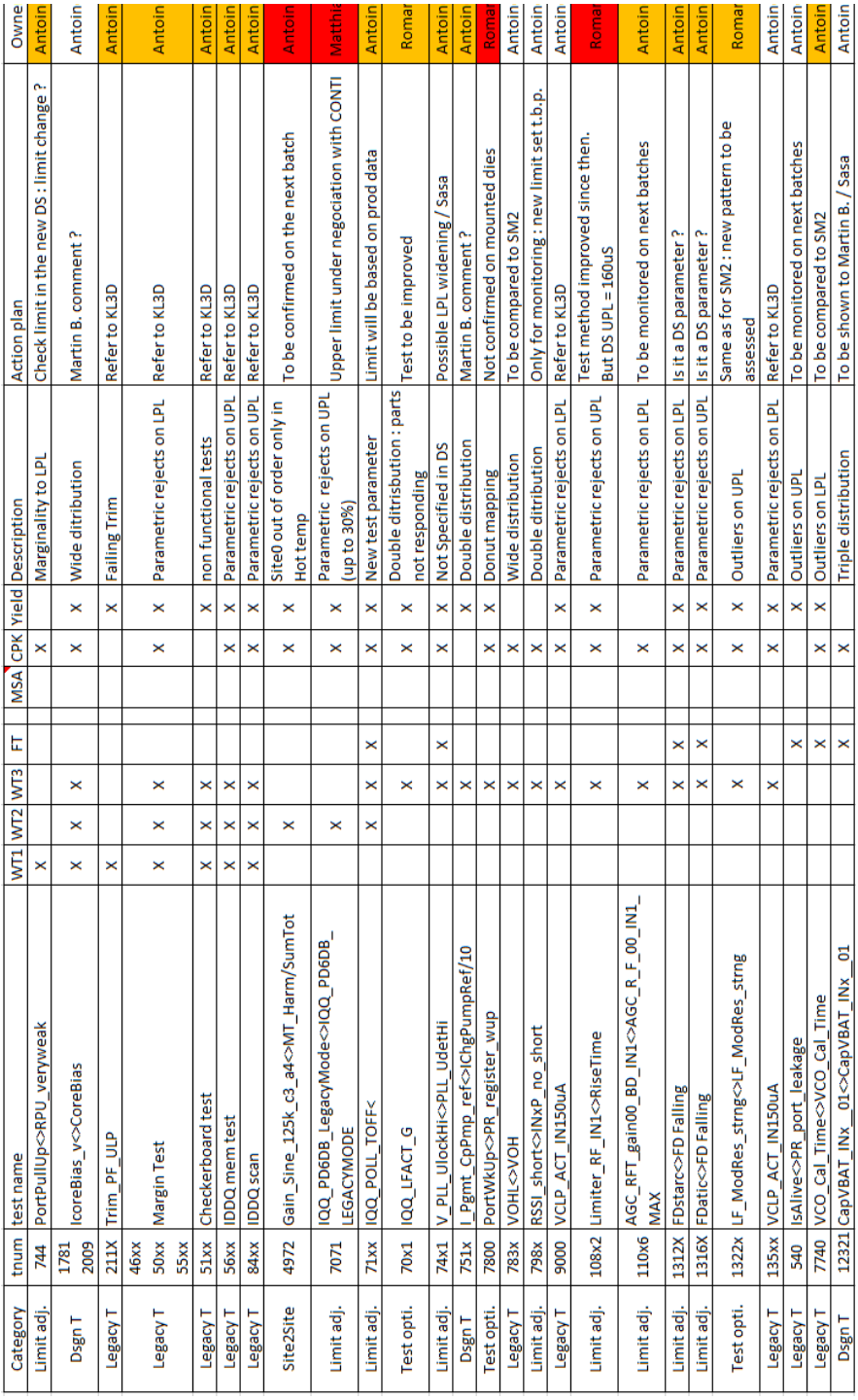
Another aspect explaining why, despite all improvements made to the project, the VRU process has still not reached its maturity level is the time it takes to change the working culture in a team. Once the structure is in place, it can take several months for the Product Engineer to change his or her way of working. In the present instance, it can take me several sessions of training to convince them that the structure is efficient. In addition to the Product Engineers, it is also all our partners in manufacturing centres who have to adopt this approach to achieve manufacturing excellence.

6.2 Results of any capability analysis

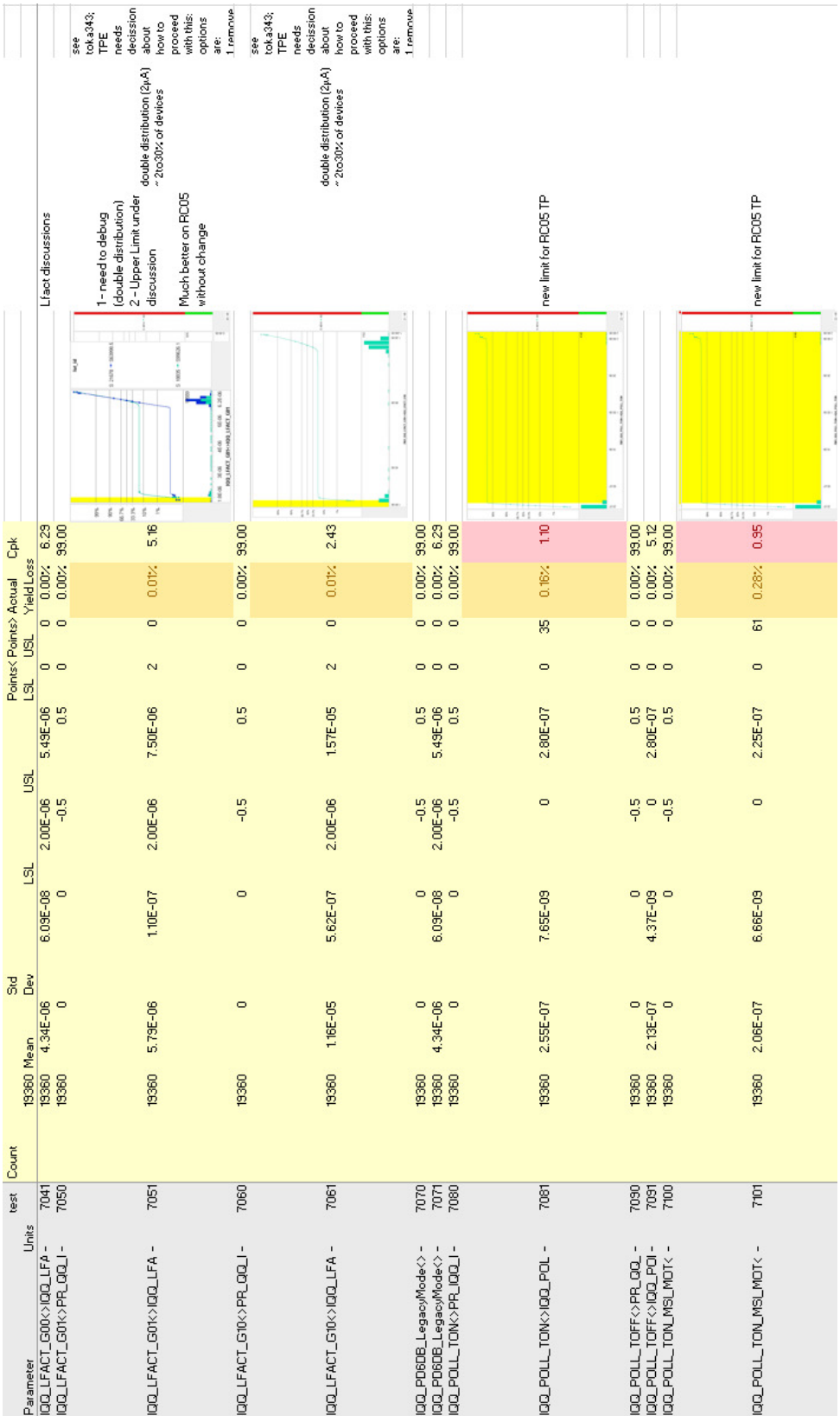
In the next pages some extracts from the Industrial Producibility Report are shown. This report is issued at the released of the product to give a picture of the production process before production starts. Above all it lists all actions to be undertaken in order to reach the maturity level. An action list is then decided upon in agreement with different stakeholders with a priority level based on the potential risks encountered by the process capability (see table 6).

Table 6: This table is an action list commonly approved with different stakeholders at the release of the product. Depending on the criticality of the highlighted issue (red, orange colours) actions will be more or less urgently executed. The important point is that not all the actions are taken immediately but are recorded. The goal of having an exhaustive list at a given instant is to provide a traceability of all parameters during the total product lifetime.

It helps to detect drifts during the lifetime of a product. In addition it facilitates the handover when a product owner change occurs.



Category	tnum	test name	WT1	WT2	WT3	FT	MSA	CPK	Yield	Description	Action plan	Owne
Limit adj.	744	PortPullUp<>RPU_veryweak	X					X		Marginality to LPL	Check limit in the new DS : limit change ?	Antoin
Dsgn T	1781 2009	IcoreBias_v<>CoreBias	X	X	X			X	X	Wide ditribution	Martin B. comment ?	Antoin
Legacy T	211X	Trim_PF_ULP	X						X	Failing Trim	Refer to KL3D	Antoin
Legacy T	46xx 50xx 55xx	Margin Test	X	X	X			X	X	Parametric rejects on LPL	Refer to KL3D	Antoin
Legacy T	51xx	Checkerboard test	X	X	X				X	non functional tests	Refer to KL3D	Antoin
Legacy T	56xx	IDDQ mem test	X	X	X			X	X	Parametric rejects on UPL	Refer to KL3D	Antoin
Legacy T	84xx	IDDQ scan	X	X	X			X	X	Parametric rejects on UPL	Refer to KL3D	Antoin
Site2Site	4972	Gain_Sine_125k_c3_a4<>MT_Harm/SumTot		X				X	X	Site0 out of order only in Hot temp	To be confirmed on the next batch	Antoin
Limit adj.	7071	IQQ_PD6DB_LegacyMode<>IQQ_PD6DB_LEGACYMODE		X				X	X	Parametric rejects on UPL (up to 30%)	Upper limit under negociation with CONTI	Matthia
Limit adj.	71xx	IQQ_POLL_TOFF<		X	X	X		X	X	New test parameter	Limit will be based on prod data	Antoin
Test opti.	70x1	IQQ_LFACT_G			X			X	X	Double ditrisbution : parts not responding	Test to be improved	Romar
Limit adj.	74x1	V_PLL_UlockHi<>PLL_UdetHi			X	X		X	X	Not Specified in DS	Possible LPL widening / Sasa	Antoin
Dsgn T	751x	I_Pgmt_CpPmp_ref<>IChgPumpRef/10			X				X	Double distribution	Martin B. comment ?	Antoin
Test opti.	7800	PortWkUp<>PR_register_wup			X			X	X	Donut mapping	Not confirmed on mounted dies	Romar
Legacy T	783x	VOHL<>VOH			X			X		Wide distribution	To be compared to SM2	Antoin
Limit adj.	798x	RSSI_short<>INxP_no_short			X			X		Double ditribution	Only for monitoring : new limit set t.b.p.	Antoin
Legacy T	9000	VCLP_ACT_IN150uA			X			X	X	Parametric rejects on LPL	Refer to KL3D	Antoin
Limit adj.	108x2	Limiter_RF_IN1<>RiseTime			X			X	X	Parametric rejects on UPL	Test method improved since then. But DS UPL = 160uS	Romar
Limit adj.	110x6	AGC_RFT_gain00_BD_IN1<>AGC_R_F_00_IN1_MAX			X			X		Parametric rejects on LPL	To be monitored on next batches	Antoin
Limit adj.	1312X	FDstarc<>FD Falling			X	X		X	X	Parametric rejects on LPL	Is it a DS parameter ?	Antoin
Limit adj.	1316X	FDatic<>FD Falling			X	X		X	X	Parametric rejects on UPL	Is it a DS parameter ?	Antoin
Test opti.	1322x	LF_ModRes_strng<>LF_ModRes_strng			X			X	X	Outliers on UPL	Same as for SM2 : new pattern to be assessed	Romar
Legacy T	135xx	VCLP_ACT_IN150uA			X			X	X	Parametric rejects on LPL	Refer to KL3D	Antoin
Legacy T	540	IsAlive<>PR_port_leakage				X			X	Outliers on UPL	To be monitored on next batches	Antoin
Legacy T	7740	VCO_Cal_Time<>VCO_Cal_Time				X		X	X	Outliers on LPL	To be compared to SM2	Antoin
Dsgn T	12321	CapVBAT_INx__01<>CapVBAT_INx__01				X		X		Triple distribution	To be shown to Martin B. / Sasa	Antoin

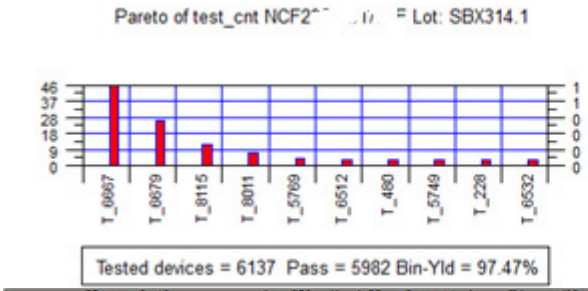


Parameter	Units	test	Count	Mean	Std Dev	LSL	USL	Points< LSL	Points> USL	Actual Yield Loss	Cpk	L1act discussions
IQQ_LFACT_G00<>IQQ_LFA	-	7041	19360	4.34E-06	0	6.09E-08	2.00E-06	5.49E-06	0	0	0.00%	6.29	
IQQ_LFACT_G01<>PR_QQ_I	-	7050	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_LFACT_G01<>IQQ_LFA	-	7051	19360	5.79E-06	0	1.10E-07	2.00E-06	7.50E-06	2	0	0.01%	5.16	1 - need to debug (double distribution) 2 - Upper Limit under discussion double distribution (2µA) ~ 2to30% of devices Much better on RC05 without change
IQQ_LFACT_G10<>PR_QQ_I	-	7060	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_LFACT_G10<>IQQ_LFA	-	7061	19360	1.16E-05	0	5.62E-07	2.00E-06	1.57E-05	2	0	0.01%	2.43	see tok343; TPE needs decission about how to proceed with this: options are: 1 remove double distribution (2µA) ~ 2to30% of devices
IQQ_PD60B_Legacy\Mode<>	-	7070	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_PD60B_Legacy\Mode<>	-	7071	19360	4.34E-06	0	6.09E-08	2.00E-06	5.49E-06	0	0	0.00%	6.23	
IQQ_POLL_TON<>PR_IQQ_I	-	7080	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_POLL_TON<>IQQ_POL	-	7081	19360	2.55E-07	7.65E-09		0	2.80E-07	0	35	0.16%	1.10	new limit for RC05 TP
IQQ_POLL_TOFF<>PR_QQ_	-	7090	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_POLL_TOFF<>IQQ_POL	-	7091	19360	2.13E-07	4.37E-09	0	0	2.80E-07	0	0	0.00%	5.12	
IQQ_POLL_TON_MSL_MOT<	-	7100	19360		0	0	-0.5	0.5	0	0	0.00%	99.00	
IQQ_POLL_TON_MSL_MOT<	-	7101	19360	2.06E-07	6.66E-09		0	2.25E-07	0	61	0.28%	0.95	new limit for RC05 TP

Table 7: This is an extract of the Cpk table made for a product already in high volume production. For every test with a low Cpk (red), and/or with yield losses (orange), a comment is given together with a distribution and potentially an action.

Description	Bin #	Device qty	%	Sublot qty
in_1	B_1	5982	97.47%	7
in_27	B_27	93	1.52%	7
in_17	B_17	21	0.34%	6
in_13	B_13	10	0.16%	6
in_29	B_29	9	0.15%	5
in_10	B_10	7	0.11%	4
in_12	B_12	6	0.10%	2
in_19	B_19	4	0.07%	3
in_18	B_18	2	0.03%	2
in_21	B_21	2	0.03%	2
in_25	B_25	1	0.02%	1

Description	Test #	Device qty	%
RSSI_AppOffset_ADC_3D	T_6667	46	0.75%
RSSI_App_ADC_3D	T_6679	26	0.42%
ADC_InMux	T_8115	12	0.20%
DAC_P16	T_8011	7	0.11%
MT_125k_c3_a3	T_5769	4	0.07%
Sine_125k_c1_a1_6dB_Offset	T_6512	3	0.05%
LeakPort	T_480	3	0.05%
MT_125k_c2_a3	T_5749	3	0.05%
P2P_Open	T_228	3	0.05%
Sine_125k_c3_a1_6dB_Offset	T_6532	3	0.05%
Sine_125k_c2_a1_6dB_Offset	T_6522	3	0.05%
IQQ_IT_APP_3V0	T_3741	3	0.05%
MT_125k_c1_a3	T_5729	2	0.03%
WriteAllULP	T_13100	2	0.03%
MT_125k_c2_a3_6dB	T_5931	2	0.03%
MT_BB_2k_b3	T_5583	2	0.03%
IQQ_FACT_VBAT_3V0	T_3711	2	0.03%
P2P_Open	T_213	1	0.02%
LeakPort	T_482	1	0.02%
P2P_Open	T_219	1	0.02%
P2P_Open	T_208	1	0.02%
MT_125k_c2_a3	T_5746	1	0.02%
LeakCoil	T_372	1	0.02%
StartupTime	T_4510	1	0.02%



Description	Bin #	Device qty	%	Sublot qty
bin_1	B_1	73	47.10%	4
bin_27	B_27	44	28.39%	4
bin_17	B_17	14	9.03%	4
bin_13	B_13	10	6.45%	4
bin_10	B_10	5	3.23%	3
bin_29	B_29	3	1.94%	2
bin_18	B_18	2	1.29%	2
bin_21	B_21	2	1.29%	1
bin_25	B_25	1	0.65%	1
bin_19	B_19	1	0.65%	1

Description	Test #	Device qty	%
RSSI_AppOffset_ADC_3D	T_6667	28	18.06%
RSSI_App_ADC_3D	T_6679	9	5.81%
ADC_InMux	T_8115	5	3.23%
IQQ_IT_APP_3V0	T_3741	3	1.94%
LeakPort	T_480	3	1.94%
IQQ_FACT_VBAT_3V0	T_3711	2	1.29%
MT_125k_c3_a3	T_5769	2	1.29%
RSSI_App_ADC_3D	T_6676	2	1.29%
WriteAllULP	T_13100	2	1.29%
DIDD	T_3980	1	0.65%
StartupTime	T_4510	1	0.65%
AGC_RFT_HI_BD_IN3	T_4653	1	0.65%
IQQ_FACT_APP_3V0	T_3721	1	0.65%
MT_125k_c1_a3	T_5729	1	0.65%
IQQ_FACT_APP_3V6	T_3831	1	0.65%
MT_125k_c2_a3	T_5746	1	0.65%
DIDD	T_3982	1	0.65%
MT_125k_c2_a3	T_5751	1	0.65%
MT_125k_c3_a3	T_5766	1	0.65%
MT_BB_2k_b3	T_5583	1	0.65%
MT_125k_c1_a3_6dB	T_5909	1	0.65%
MT_125k_c2_a3_6dB	T_5931	1	0.65%
MT_125k_c3_a3_6dB	T_5943	1	0.65%
MT_125k_c3_a3_6dB	T_5949	1	0.65%

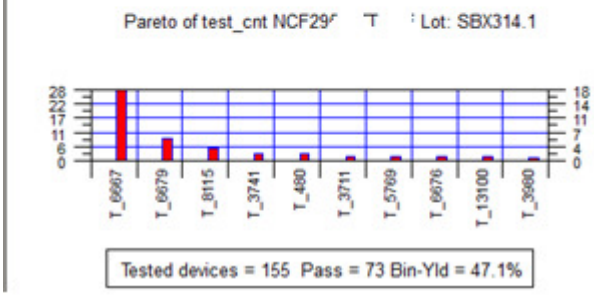


Figure 41: This is the comparison of the failure Pareto diagram between the first pass (upper part) and the retest (below) in final test. The initial quantity in is 6137. The total quantity of good parts is 5982 meaning the yield after first pass is 97,47%. Then the 155 rejects from the first pass have been retested and out of those 155 parts, 73 parts pass (yield =47.10%). The recovery rate is very high. By checking quantities of rejects per test parameter, there are 46 rejects on T_6667 at the first pass. After retest, only 28 rejects remain which means that the test parameter T_6667 has a very high recovery rate. This also means that if rejects had been retested once, statistically there would be again some yield recovery on this test. By improving the capability of the test parameter T_6667, the first pass yield would be greater, there would be fewer retested parts thus improving the tester efficiency. It would also improve the global yield. It may happen that any rejects are never recovered (e.g. due to a problem during the assembly process). In this case, to improve the tester efficiency, certain rejects are not retested. The final test yield of that assembly lot is 98.66%.

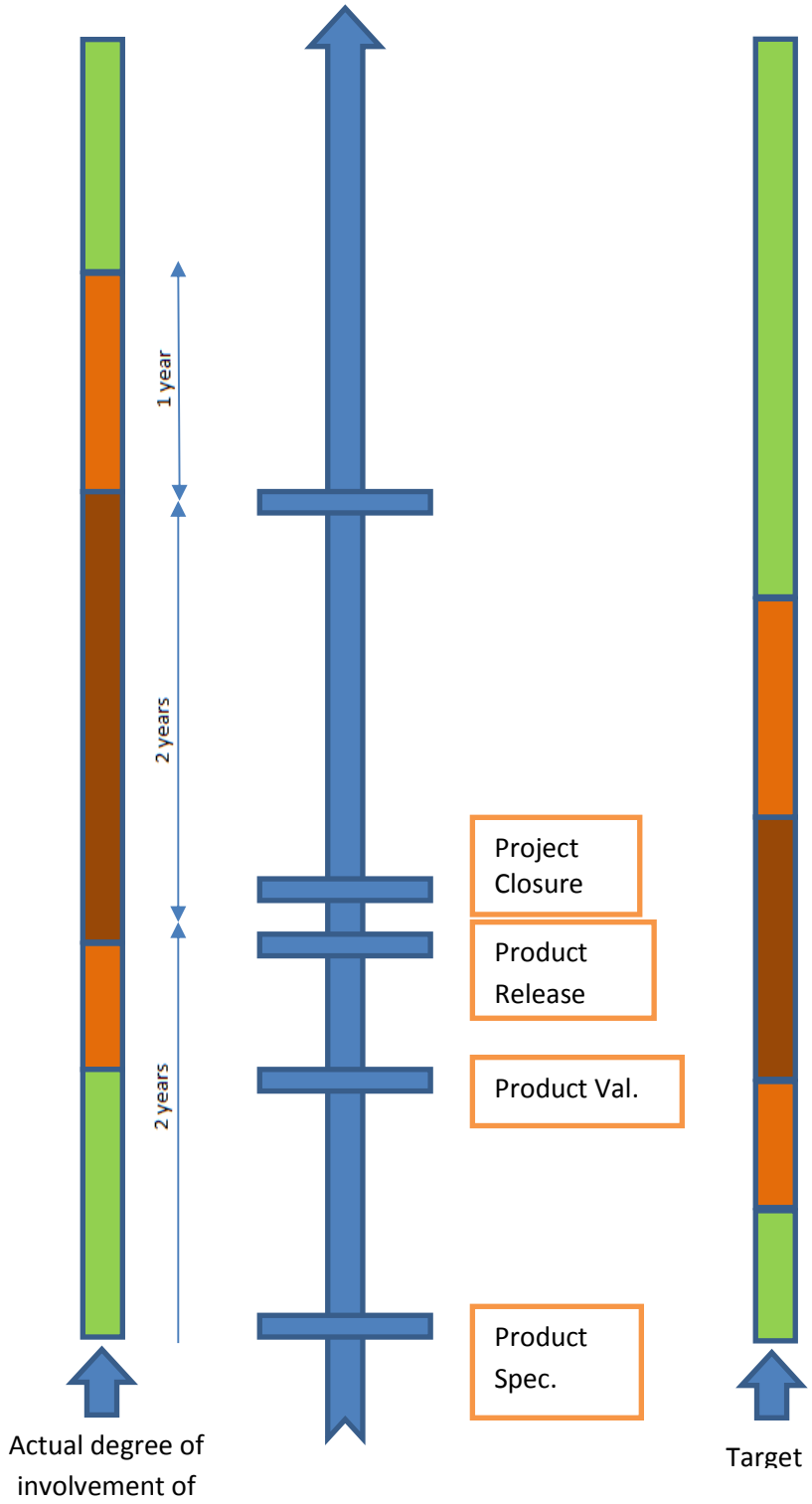


Figure 42: It refers to the figure #-- and gives the axis of the improvement of the VRU process. On the left side is the actual degree of involvement of the VRU team along the project development and after. On the right side, the VRU team involvement once the VRU process will have reached it maturity. We see here that the VRU process starts very late and will pursue on a long period. The idea is to start much earlier, to report issues as soon as

possible to take benefit of the development team and possibly adapt the datasheet before the product release. With that sequence, the length of involvement is shorter as the manufacturing process is matured faster.

6.3 Impact already visible

The majority of the work in that project was done on one of the higher runners of the division, a product released in 2011. As of April 2016, this product still requires a big effort from the engineering team to make it reach its maturity in terms of production performances and throughput time. Two other products released during this period were also taken as a reference and showed similar problems: some bad yields and some unpredictable ones. Different analysis revealed several tests parameters with low Cpk, most of which were linked to main yield losses. The capability of some of these parameters could be improved by optimizing the test limits. However this would not solve the entire problem with regard to deliveries on time. For this reason, these products have not yet reached their potential and require an active follow up.

Since this new approach was adopted, a new product has been released (Table 8). This product embedded major innovative features which makes it more complex in its architecture than its predecessor and many more tests parameters have to be covered (roughly 20% more). In terms of timing, we had enough time to make a proper capability analysis before product release and also before the datasheet was totally frozen. With hindsight it would have been profitable to start even earlier with interim data.

Product	Release date	Wafer Test yield as of June 2015	Final Test yield as of June 2015
#1	2010	70%	75%
#2	2011	75%	96%
#3	2011	70%	85%
#4	Dec-2014	85%	99%

Table 8: Most of the analysis for this project was made on the product identified here as product #2. It is the high runner of the division. A lot of work was put in to improve its performances in production to make it more profitable and above all improve the throughput time. Product #4 was developed to embed product #2 and product #3 and

several new features and upgrades. Thanks to the effort made during the development phase with the structured approach described in the report, the yield turned out to be very stable especially in Final Test and the batch-to-batch sensitivity low. It is the sign the capability of the test process is sufficient.

7.Conclusion

This dissertation has described aspects of Quality control in semiconductor manufacture. Among all the processes, the VRU process represents the link between the develop phase of the new product and its introduction to production. The goal of this process is to manage the initial deliveries to the customer on time and with the correct quality level, and also to follow up the product until it constantly reaches its target in terms of throughput and yield. The goal of this study was to enhance this process in order to improve its efficiency. To reach this goal it has been explained that it is necessary to optimize the capability of each process composing the supply chain. This one goes from the FAB until the delivery of the product. It is also explained that it is crucial that this work must be done before the product is released. Although the picture is then partial, it is a condition to address a maximum of potential risks before the datasheet is frozen and by this way to gain some time on the industrialisation phase.

More than having immediate results, the goal of the study was to structure and develop the process. For this aspect, the goal can be considered as reached. The proposed approach will definitely smooth the ramp up phase and lessen the involvement of the Operation team during the industrialization phase with the goal to bring the performances up to their optimized performances. The necessary condition is to start the process as soon as possible during the development phase since priorities during this phase, especially when the milestone approaches, could overshadow the importance of the preparation to production. Therefore it is important to always keep an eye on new products in development. As VRU manager I have a primordial role to play. I have to ensure every stakeholder fulfill its deliverables. It is also my job to ensure that Product Quality Engineers who will handle the production after the release of the product are fully aware of the importance of following the approach.

The new approach of the VRU process has already proved its efficiency. But it does not mean the deliveries during ramp up are already totally guaranteed. Further effort has to be made especially on the wafer test process. It still takes too much time to obtain stable yields, which also slows down the process of fine tuning of the acceptance criteria. The consequences are that

- The quantity of good dies per wafer is not predictable enough, which represents a problem for logistics to forecast the amount of material required.
- The hold lot rate is still too high slowing down the throughput time and requiring an intensive support from sustaining engineers and product engineers. However, test processes are qualified before the product release and test limits based on three different batch diffusions. That is not acceptable that the performances in production are so erratic. It means the test process is still too sensitive to variations in FAB.

The direction to be taken to improve it is to start the process as early as possible during the development phase. Another very crucial aspect is to capitalize on other released types. For instance:

- are the same phenomena observed on common blocks with other products ?
- was a metal fix efficient to improve the process capability ?

Globally speaking, we are talking about taking some distance from the project and widening the perimeter of the Product Quality Engineer: from a position where solutions to a problem are found, a proactive approach would be to prevent problems from happening by, for instance, better feedback to designers before a new release of design. This is an aspect requiring a change in the culture of the team. What is currently observed during an actual project being in development is encouraging. Another point still to be improved is the transfer of the test process in our production centres. The production start is sometimes difficult as performances in development centres are not identical to those in our production centres owing to multiple parameters. The improvement of this aspect passes by an improvement of the transfer of knowledge to our sustaining engineers who take care of products once in production.

All processes composing the production chain from the FAB to the delivery have not been described in this study. Only processes having the highest risk of yield losses and slow down of throughput were mentioned. Once performances have totally reached NXP standards then the other processes not described in this study will be analyzed. It is indeed an endless improvement process. However whatever the issue, the idea remains the same: look for solutions to improve the capability of process to ensure that whatever the applied drift, the process is insensitive. The interests are multiple. First, having performances with standard performances means that Production Centers can themselves sustain the product. This is a way to let engineers focus on higher level tasks. Another advantage is of course financial. Optimizing performances in terms of yield and throughput time lessen manufacturing costs and by consequence, improves profit margins.

List of abbreviations

AIAG	Automotive industry Action Group
APQP	Advanced Product Quality Planning
AQL	Acceptance Quality Limit
BCaM	Business Creation and Management
CLT	Central Limit Theorem
CPU	Central Processing Unit
CTQ	Cost To Quality
CTS	Checkerboard Test Structure
DFSS	Design For Six Sigma
DOE	Design of Experiment
FAB	Fabrication Unit
FMEA	Failure Mode Effects Analysis
FT	Final Test
IATF	International Automotive Task Force
IPR	Industrial Producibility Report LED light-emitting diode
KPI	Key Product Indicator
MEMS	Micro-electro-Mechanical Systems
MRA	Mask release acceptance
MSA	Measure System Analysis
NFC	Near Field Communication
PDCA	Plan DO Check Act
PMI	Project Management Institute
RF	Radio Frequency
RAM	Random Access Memory
ROM	Read Only Memory
VQD5	Volume Quantity Defined 5%
VRU	Volume Ramp Up

List of figures

List of tables

Bibliography

[1]Presentation of the NXP Semiconductors company: https://en.wikipedia.org/wiki/NXP_Semiconductors (April 2016)

[2] Presentation of the Technical Specification used for the Automotive sector: https://en.wikipedia.org/wiki/ISO/TS_16949 (April 2016)

[3]Explanation of the APQP: www.npd-solutions.com/apqp.html (April 2016)

[4]Documentation on the fabrication of th wafers: https://en.wikipedia.org/wiki/Wafer_fabrication (April 2016)

[5] Christopher Hess, Larg H. Weiland, Wafer Level Defect Density Distribution Using Checkerboard Test Structures, Conference on Microelectronic Test Structures, IEEE 1998 Int., Vol.11, Pages 101-106, March 1998

[6] C. Neil Berglung, Fellow, A Unified Yield Model Incorporating Both Defect and Parametric Effects, IEEE transactions on Semionductor manufacturing, vol. 9. No. 3. Pages 447-454, August 1996

[7]Reference on the AQL : http://brd4.braude.ac.il/~bashkansky/SQM/auxiliary/Inspection_sampling_standards.doc (April 2016)

[8] Explanation of the notion of capability : http://chohmann.free.fr/qualite/cp.htm (April 2016)

[9] Probability and Statistics for Engineering and the Sciences, Jay L. Devore, 9th edition, Measures of Variability, Page 14

[10] Devore, op. cit. , Histogram, Page 22

[11] Devore, op. cit. , Normal distribution, Pages 172-173

[12] Devore, op. cit. , Correlation, Page 430

Remerciements

Je tiens tout d'abord à remercier Mark PLIMMER pour sa pédagogie, sa disponibilité, sa patience ainsi que pour ses encouragements. Par Monsieur PLIMMER, c'est aussi toute l'équipe du CNAM que je tiens à remercier sincèrement. Cela fait plus de 15 ans que j'ai commencé cette formation. J'ai tout d'abord travaillé avec l'équipe du CNAM de Caen, puis celle de Grenoble pour finir avec celle de Paris. Ce long parcours atypique a fait qu'il a fallu réadapter mon cursus au fil des réformes. J'ai toujours trouvé des personnes prêtes à m'aider et trouver des solutions. Faire le choix d'une formation CNAM c'est bien souvent dans le cadre d'une démarche personnelle. Avoir le soutien auprès des enseignants et du personnel administratif est important pour garder la motivation. En effet, le CNAM c'est aussi une équipe d'enseignants de très grande qualité. La formation m'a apporté les notions techniques et notions générales nécessaires pour accéder au niveau d'ingénieur. C'est grâce à la qualité de cette équipe et de son enseignement que je peux aujourd'hui exercer ma fonction actuelle. C'est aussi grâce à toutes ces notions que j'ai pu mener à bien ce projet.

C'est un projet qui quelque part m'a permis de rassembler toutes les connaissances acquises « sur le terrain » et d'en faire bénéficier mon organisation. Pour y parvenir il m'a fallu prendre de la distance par rapport à mes habitudes de travail et synthétiser mes connaissances pour créer des nouvelles façons de travailler au sein de mon équipe. Ces compétences acquises dans le cadre de mon travail je les dois aux collègues que j'ai pu côtoyer. Ce sont des collègues que j'ai eu la chance de rencontrer au fil de mes trois changements de site. Je tiens ici à les remercier chacun d'entre eux. Je tiens à remercier plus particulièrement mon équipe actuelle qui m'accompagne dans cette dernière ligne droite.