THE MYTH ABOUT PROOF TEST COVERAGE AND MISSION TIME

There is a story going around that we can make up for limited proof test coverage of automated safety functions by reducing device mission time. It is a convincing story, but it is (almost) completely wrong.

Mission time has nothing to do with proof test coverage. The mission time for devices should be based on the useful operating life of the devices rather than on proof test coverage.

$$\xcancel{\frac{\lambda_{DN}.T_M}{2}}\ ?$$

Estimating probability of failure

Performance of safety functions that act on demand is measured by the average probability of failure on demand, PFD_{AVG}.

When safety functions are first put into service they are supposed to be fully tested. The tests should demonstrate that the functions perform as required. The probability of failure immediately after the test should effectively be zero if the testing was complete. Faults might occur as time progresses during operation.

Some faults are easily detected during operation by diagnostic functions. The detected faults can be repaired before they cause failure of the function on demand. Not all faults can be detected during operation, so as time progresses the probability of failure on demand gradually increases.

Faults are classed as 'dangerous' if they can cause failure of the safety function on demand.

The established theory behind estimating probability of failure on demand assumes that undetected dangerous failures accumulate at a reasonably constant average rate. Refer to IEC 61508, ISO/TR 12489, ISA-TR84.00.02 or to the SINTEF PDS Method Handbook for explanation of the theory.

The average **rate of undetected failures** is usually designated $\boldsymbol{\lambda_{DU}}$ (e.g. average number of failures per year). To a very good approximation, the probability of failure of a safety function device at time $\boldsymbol{t}$ since its last full test is simply λ_{DU} x t.

Safety function devices are regularly inspected and tested to limit the probability of failure. The interval at which inspections and tests are scheduled is called the **proof test interval.** It can be designated as T. Proof test intervals are typically in the range of 1 to 4 years.

The probability of failure increases linearly with time if the failure rate remains reasonably constant. The relationship is reasonably linear provided that $t < 0.1/\lambda_{DU}$. For example, if the average failure rate is one failure in 100 years then the increase in probability is close to linear for at least the first 10 years.

When the device is tested at time T, its probability of failure will be approximately λ_{DU} x T.

The average probability of failure on demand PFD_{AVG} over the whole period T can be estimated as:

$$PFD_{AVG} \approx \frac{\lambda_{DU}.T}{2}$$

The problem is in proof test coverage

The problem is that we cannot usually achieve perfect proof test and inspection.

The tests and inspections will fail to reveal some proportion of the undetected failures in the safety function.

We can define the term 'never-detected failures' to describe the failures that cannot be revealed by the planned tests and inspections. Some use the equivalent term 'not detectable by the planned testing'.

Those never-detected failures might persist for the remaining service life of the device or until a real demand occurs and the function fails to respond as intended.

An explanation of how to deal with imperfect test coverage can be found in section 14.2.4 of ISO/TR 12489 '*Petroleum, petrochemical and natural gas industries — Reliability modelling and calculation of safety systems*'.

If the never-detected failures were to occur at a reasonably constant rate, then the probability of failure would increase over time in proportion to the **rate of never detected failures, λ_{DN}**.

The **proof test coverage** PTC is defined as the proportion of dangerous undetected failures that can be revealed by inspection and testing.

$$PTC \quad = \quad \frac{\lambda_{DU} - \lambda_{DN}}{\lambda_{DU}} \quad = \quad 1 - \frac{\lambda_{DN}}{\lambda_{DU}}$$

and

$$\lambda_{DN} = (1 - PTC)\,.\,\lambda_{DU}$$

The **mission time** of a device can be defined as T_M, the **time in service before scheduled overhaul or replacement.** Note that the refurbishment or replacement process must include complete inspection and testing equivalent to the original validation process before the refurbished or replacement device is put back into service.

Mission times for safety function devices are typically in the range of 10 to 20 years. Devices in severe service might have relatively short mission times such as 5 years. In extreme cases, consumable devices such as furnace temperature probes might have mission times measured in months. Instead of testing them we simply replace them before they wear out.

If the never-detected failures accumulate at the average rate λ_{DN} over the mission time T_M, then the according to ISO/TR 12489 section 14.2.4.2, the average probability of failure can be estimated as:

$$PFD_{AVG} \approx \frac{(\lambda_{DU} - \lambda_{DN}).T}{2} + \frac{\lambda_{DN}.T_M}{2}$$

This relationship suggests that dangerous failures accumulate over time if they are not detected by regular testing. The probability of the safety function device failing on demand continues to increase until the device is taken out of service.

From this equation it appears that we could compensate for λ_{DN} by reducing the mission time T_M. **However, the assumption of a constant failure rate is unjustified. There is no reason why we should ever expect never-detected failures to accumulate at a fixed rate.**

ISO/TR 12489 section 14.2.1.3 clarifies the variability in failure rates and warns that *'even in the simplest cases, the reliability analyst should verify that the hypothesis of constant failure rates is realistic when he/she decides to implement formulae, Boolean or Markov calculations.'* ISO/TR 12489 sections 7.2 and B.3 also provide useful clarification.

The relationship $\lambda_{DN} = (1 - PTC)\,.\,\lambda_{DU}$ may be misleading. The PTC can be estimated from average measured failure rates, and the average failure rate measured over the whole life of a device can be estimated from PTC. The long-term averages are useful and can be used to estimate the average probability of failure over the whole life of the safety function. However, failure rates *during operation* cannot be assumed to be constant and cannot be calculated from the PTC.

Some of the popular calculation software packages are based on the simplifying assumption that the never-detected failures do accumulate at a constant average rate λ_{DN}. The assumption is useful because it allows the PFD_{AVG} over the life of the system to be estimated. Unfortunately some users make the mistake of manipulating the mission time in order to reduce PFD_{AVG}.

IEC 61508-6 section B.3.2.5 takes a slightly different approach. It also assumes constant average rates for λ_{DU} and λ_{DN}, but it uses the expected demand period T_2 in place of mission time T_M in the calculation of PFD_{AVG}. IEC 61508-6 does not suggest the strategy of reducing mission time.

The SINTEF PDS Method Handbook section 5.3.2 and ISA-TR84.00.02 section 6.2 provide similar guidance on the effect of partial test coverage. Instead of mission time or demand period they consider the time between full tests, assuming a full test can be executed to reveal failures not detected by the partial tests. Again, these standards do not suggest the strategy of reducing mission time.

The appropriate strategy is to reduce the time interval between tests with full coverage, as distinct from reducing the planned service life.

Overall average rates are measured by counting failures

We can always measure an overall average failure rate $\overline{\lambda_{DN}}$ simply by dividing the number of never-detected failures by the total time in service. The overall average failure rate (failures per unit of device service time) can be calculated as:

$$\overline{\lambda_{DN}} = \frac{n}{N\,.\,\overline{T_M}}$$

Where $\boldsymbol{n}$ is the number never-detected failures that occur in a population of $\boldsymbol{N}$ devices with an average mission time of $\overline{\boldsymbol{T_M}}$. The measured rate can be used to estimate the proof test coverage.

Failures rarely occur at constant average rates

The fact that we can measure an average rate does not mean that the failures occur at a fixed and predictable rate.

By definition, failures that occur as independent events with a constant average rate are purely random failures.

ISO/TR 12489 provides a good explanation of purely random ('catalectic') failures. Purely random failures are complete and sudden failures of a component. They usually occur only in electronic devices. A device may continue to operate with degraded performance if a component part fails.

The occurrence of purely random failures cannot be predicted by inspecting the devices. They can usually be revealed after they occur, either by diagnostics or by testing and inspection.

All measured failure rates (including purely random failures) vary within a wide uncertainty interval. The range of uncertainty usually spans at least an order of magnitude. No failure rates should be assumed to be fixed and constant.

The probability calculations described in ISO/TR 12489, IEC 61508-6 and ISA-TR84.00.02 are all based on the assumption of reasonably constant failure rate.

The calculations are still useful with variable failure rates, but the variability in failure rates and in the results needs to be understood. ISO/TR 12489 and ISA-TR84.00.02 include guidance clarifying the variability in failure rates and the causes of variability.

In practice, never-detected failures are never purely random failures. Failures that cannot be detected after they occur are more likely to be mechanical, electrical, pneumatic or chemical in nature. These failures are not purely random. They do not result from independent events or stochastic processes.

The illusion of random behaviour

In any sufficiently large population failures often appear to be random even if the failure mechanisms are not purely random.

Measured failure rates may appear to be reasonably constant, and this creates the illusion that the behaviour is random.

Consider a collection of 100 incandescent light bulbs as an example. The useful lifetime of each bulb could be approximately 1,000 hours. The useful lifetime might be shortened if the bulbs are placed under stress such as vibration or high temperatures.

If the bulbs are left in service and replaced only after they fail, the measured daily failure rate will eventually appear to be reasonably constant. We would expect an average of around 2.4 failures every day in a population of 100 bulbs.

The truth is that the bulb failures are almost all end-of-life failures. The random failure rate of each bulb during its useful life is lower. We could reduce the measured failure rate to perhaps 2 or 3 failures per month by replacing all the bulbs regularly at monthly intervals (i.e. after 700 hours of service).

Some failures result from pre-existing faults

Some failures are purely systematic. **Systematic failures** are caused by faults in design, manufacture, operation or maintenance.

Even though an average failure rate can always be measured, systematic failures cannot be characterised by a failure rate. They might instead be characterised by the probability of faults existing in the system.

The SINTEF PDS Method Handbook section 5.3.1 uses the term '**test-independent failure**' to describe failures that are not revealed by testing but revealed upon a true demand. These are failures that can be characterised by a fixed probability P_{TIF} rather than a fixed rate. Test-independent failures result from faults that are present from the outset.

ISO/TR 12489 section 7.3.1 introduces similar factors for failures characterised by fixed probabilities:

γ	probability of failure due to the test itself
ψ	probability of failure due to the demand of the safety action itself
ω	probability of human error.

Reducing the mission time will have little effect on the probability of purely systematic failures or test-independent failures. Systematic failures result from pre-existing faults or human actions and inactions. If a device is unreliable due its design, then simply replacing the device with a new one will not fix the problem.

The probability of test-independent failures can be reduced by improving quality management and by improving the coverage of inspection and testing.

Effective quality management techniques and measures reduce the likelihood of systematic faults. Inspection and testing are techniques that are used to reveal systematic faults.

Some failures occur at rates that increase with age or stress

Most failures are partially random and partially systematic. They might be characterised by a failure rate, but the rate will always vary over time and can vary significantly depending on effectiveness of equipment maintenance.

The failure rate will also depend on the level of stress and the duration of stress applied to a component (for instance, following an Arrhenius or Weibull-Arrhenius relationship).

Failures caused by age, wear and stress are partially random and partially systematic.

They can be characterised by an average failure rate, but that rate is not constant. The rate increases with time. These failures can be anticipated and prevented. The failure rate can be deliberately controlled, as in the example of light bulbs.

It is generally accepted that these failures can be modelled by constant failure rates, but the uncertainty intervals in the actual failure rates span at least an order of magnitude.

Mission time should be based on useful service life

Reducing the mission time of a device reduces the probability of failure caused by stress, deterioration and aging - up to a point.

Some failures caused by stress, age, wear and deterioration might be difficult to detect and might not be detectable by the planned inspection and testing. Even so, the rate of these failures cannot be characterised by a constant rate λ_{DN}. The rate of stress or age-related failures increases sharply when equipment reaches the end of its useful life.

Safety functions may also have some purely random failures that are not detectable by the planned inspection and testing. Theoretically, the mission time could be shortened to reduce probability of purely random failure. In practice purely random failures only occur in electronic components. Most of these failures are immediately revealed or easily detectable (if a test is actually done) because they are 'catalectic' failures. They cause a sudden and dramatic loss of function (refer to ISO/TR 12489). They do not generally make a significant contribution to the overall probability of safety functions failing on demand. The overall probability of failure is strongly dominated by systematic failures and by failures resulting from age, wear and deterioration.

Mission time of safety function devices should usually be based on the age beyond which age-related failures increase significantly, or beyond which the cost of maintenance exceeds the replacement cost.

The measured failure rate and the probability of failure are not directly proportional to mission time. Replacing the 100 light bulbs every fortnight instead of every month is not likely to reduce the measured failure rate significantly.

James Moubray provided a good explanation of how to determine useful service life in his book *Reliability-Centered Maintenance RCM-II* (Butterworth-Heinemann/Industrial Press Inc. 1997).

Conclusions about the myth

We can conclude that:

- Proof test coverage and mission time are both important, but they are not directly related to each other
- Failures are modelled using fixed failure rates for convenience in estimating the average probability of failure over the life of the system
- Never-detected (or not-detected) failures are not purely random in nature and do not occur at a fixed constant failure rate λ_{DN}
- Overall average rates of never-detected failure $\overline{\lambda_{DN}}$ can be measured over the lifetime of the devices
- The overall proof test coverage can be inferred from the overall measured average rate of never-detected failures $\overline{\lambda_{DN}}$
- A fixed constant never-detected failure rate λ_{DN} cannot be inferred from proof test coverage
- Never-detected failures may alternatively be characterised by a probability instead of a rate – such as the 'probability of test-independent failure' P_{TIF} described in the SINTEF PDS Method Handbook
- The probability of failure on demand due to never-detected failures may be estimated from the overall measured average rate and the mission time, $PFD_{AVG} \approx \overline{\lambda_{DN}}.T_M/2$
- The measured average failure rate depends on the mission time; reducing the mission time may increase or decrease the measured rate
- Mission time should be based on an analysis of the useful life of devices and their failure modes; it should not be based on $\overline{\lambda_{DN}}$
- The appropriate way to control the effect of reduced test coverage is to carry out full testing, and the overall PFD_{AVG} can be reduced by shortening the interval between full tests.

WHY PROOF TEST COVERAGE MATTERS

Proof test and inspection coverage must be accounted for when analysing safety function performance. Test-independent failures can cause a significant increase in the probability of safety functions failing on demand.

Test-independent failures are mostly systematic in nature; they result from pre-existing faults or from external causes. The potential causes of the test-independent failures need to be analysed and understood so that the probability of failure can be estimated.

Incomplete testing and inspection may be acceptable for SIL 1 or SIL 2 functions if the overall probability of failure meets the target.

By definition, a SIL 1 function has a target value for PFD_{AVG} in the range between 0.01 and 0.1. The PFD_{AVG} target for SIL 2 functions is in the range between 0.001 and 0.01. For SIL 3 functions the target is less than 0.001.

For the P_{TIF} to be negligible it would need to be an order of magnitude less than the target for PFD_{AVG}.

If we take the least onerous example possible, SIL 1 with a target of 0.1, we would need the P_{TIF} to be less than 0.01 in order to be comfortable that it does not significantly affect the performance of the safety function.

That might be readily achievable, but the challenge is in being able to quantify the probability of test-independent failure P_{TIF} with sufficient certainty. How can we be sure that the P_{TIF} < 0.01?

For SIL 3 we would usually need P_{TIF} < 0.0001, i.e. only a 1 in 10,000 chance of failure. Any test-independent failures will usually prevent SIL 3 levels of risk reduction from being achieved.

SIL 2 and SIL 3 safety functions should always be designed so that as far as is practicable they can be fully tested and inspected.

IMPROVING PROOF TEST COVERAGE

Proof test and inspection coverage is important

Proof test and inspection reduces the measured rate of never-detected failures. It reduces the proportion of failures that are never detected. Most of those failures result from pre-existing faults.

The purpose of inspection and testing is to find those faults before failure occurs.

Some manufacturers suggest values for proof test coverage that can be achieved for their devices, but proof test coverage actually depends on the system as a whole, rather than on the device alone.

Proof test coverage is limited by the extent to which the system has been designed to facilitate inspection and testing.

Hidden failures have known causes

Probability of failure can be reduced either by design or by inspection and testing.

It is essential to understand the causes of never-detected failures if we want to control their probability of failure.

Failures might remain hidden for any of these reasons:

- Proof testing is not possible on equipment that is damaged or destroyed by activation
- Faults in the design or installation of equipment may prevent performance as intended
- Faults in the design of equipment may preclude testing
- The equipment might not be accessible for inspection or testing
- The planned tests are incomplete or ineffective
- The equipment might never have been fully commissioned.

If the hidden failure modes are understood the system can be designed to eliminate the failures or to facilitate testing to reveal the failures.

Few of these hidden failures would occur at a constant average rate. They are either test-independent failures (mostly failures in design and installation) or else they are related to stress or age and deterioration of equipment.

Reducing the mission time of components may be useful for controlling failures caused by age and deterioration, or for controlling failures specifically due to random failure of electronic components. It is not useful for controlling failures caused by design faults.

One extreme: some devices can never be proof-tested

Devices that are destroyed during activation may have dangerous failures that can never be detected until the devices fail on demand.

The proof test coverage is zero; they cannot be tested with destroying the devices. The mission time is based on the useful service life of the devices.

Electrical fuses and automotive airbags are examples of devices that are destroyed during successful activation. Neither fuses nor airbags can be routinely tested, though they can be inspected.

Random sampling and destructive testing (i.e. type testing) can reveal:

- Whether the design performs as specified
- The proportion of devices that fail immediately due to faults in manufacturing
- The proportion of devices that fail due to deterioration with age
- The rate at which failures might be expected to accumulate in a population of devices.

The failures are not likely to accumulate at a constant average rate.

The probability of a correctly sized fuse failing on demand is not likely to increase with age. Fuses that have deteriorated with age are more likely to operate spuriously rather than fail to operate on demand.

A rupture disk for pressure relief is another example of equipment that cannot be routinely tested but can be inspected for deterioration and for correct installation. Failure rates of rupture disks can be expected to increase over time due to deterioration such as corrosion or contamination. Rupture disks may also fail if they are installed the wrong way around.

Even though routine proof testing is not possible some types of deterioration may be revealed by close inspection of each device. Destructive testing of aged samples may be useful in assessing the useful life of devices.

Some hidden failures may indeed be purely random

Automotive airbags rely on electronic components to detect rapid deceleration. These components are subject to purely random failure characterised by a constant failure rate. That failure rate does need to be considered in determining useful mission time. Mission time is not the only factor that can be used to limit failure probability. Though the failure rate remains reasonably constant, the rate depends on the design of the semiconductors. The rate can be reduced through radiation-hardening techniques.

The probability of an airbag failing on demand also increases with stress or age due to deterioration. Deterioration of the propellant can prevent successful activation. The failure rate due to deterioration increases with age.

The other extreme: full testability by design

Proof test and inspection coverage is limited by design and accessibility.

A pneumatically actuated shutdown valve is an example of a device for which full testing and inspection should be theoretically possible.

All dangerous failures of a valve can be revealed if the valve is readily accessible for regular full inspection and testing.

The testing and inspection would need to be designed specifically to find all types of failure that might prevent successful performance of the safety function.

Failures could remain undetected if the testing and inspection is not complete. Examples of failures that might never be detected include:

- Leakage through a tight-shut-off valve if no facilities are provided for in-situ leakage testing
- Sheared valve stem or coupling if the valve stroke time cannot be measured and the limit switches are mounted on the actuator.

For complete coverage of a valve, the inspection and testing would need to cover all specified performance requirements as well as the condition of the valve. It might include:

- Measurement of valve stroking time and/or stem torque for the full range of valve travel and with comparison against historical data and specified performance
- Measurement of leakage rates, or alternatively a review of trend recordings of changes in process flow or pressure showing successful operation
- Assessment of motive air or hydraulic supply quality
- Inspection for corrosion, contamination, deterioration, damage.

The compromised position: testability limited by design

If testability has not been specifically considered in the design it might not even be possible to carry out full validation testing (proving that it works as specified, before putting it into service), let alone regular proof testing.

Some safety functions fail on demand because they have never functioned properly since they were installed. They have never been fully commissioned and validated because there is no practicable way of fully testing them.

The measurement of liquid level in a pressure vessel provides some good examples of measurement techniques with failure modes that might never be found by testing:

- Level measurements relying on density may have hidden failures if the fluid density deviates beyond the expected range (e.g. Texas City Refinery Explosion 2005)
- Level measurements relying on guided wave radar may detect either the top or the bottom of a layer of foam or emulsion, and it might not be practicable to create an emulsion during validation testing
- The level in a stilling well or in an external bridle can be significantly different to the level inside a vessel because of variations in density or due to contamination
- Standing waves caused by high velocity flow can lead to a hazardous level that is not detected at the position of the level sensor
- An increase in wiring loop impedance due to corrosion of connections may reduce the voltage available at the sensor terminals, causing calibration drift at high signal levels (easily detected, but only if it is included in the test plan or in diagnostic functions).

Some types of level sensor can only be fully tested by filling the vessel right up to the trip point under normal process operating conditions. The risk associated with such a test might not be acceptable to the operators.

Some devices fail by design

Some safety functions fail on demand because the design is not suitable for the service, cannot be fully tested in routine tests, and has not been fully type-tested.

The failure modes of sensors and final elements depend on the physical principles employed, and on the environment and service conditions.

The practicality and feasibility of inspection and testing needs to be considered very early in the conceptual design of the process.

If full testing (either routine testing or type testing) is not likely to be acceptable or feasible then different measurement techniques should be considered.

Failure mode, effects and diagnostic analysis

The best way to develop requirements for proof test coverage and for system design is through FMEDA: an analysis of the failure modes, failure effects, failure rates and diagnostic coverage for each component in a sub-system.

FMEDA aims to identify all failure modes (including hidden failures) and to assess the criticality of each failure mode.

FMEDA compares the relative contribution from different failure modes by using measured average failure rates.

FMEDA allows the proof test coverage to be estimated from the relationship $1 - \lambda_{DN}/\lambda_{DU}$.

It does not matter that the failures do not occur at fixed rates. It does not matter that the uncertainty in failure rates typically spans an order of magnitude. The estimate of proof test coverage is still useful even though it is never precise.

The rates λ_{DN} and λ_{DU} can usually both be estimated to the nearest half order of magnitude (a factor of 3). Realistically, the ratio $\lambda_{DN}/\lambda_{DU}$ can only be estimated to the nearest value of 1, 0.3, 0.1, 0.03, 0. The corresponding values of PTC are 0, 70%, 90%, 97% and 100%.

If FMEDA reveals that some failures might never be detected, the design can be modified to eliminate failure modes or to improve testability. Proof test coverage or diagnostic coverage can be improved.

Proof testing and inspection procedures need to be based on a complete understanding of the potential failure modes of the system as whole.

Many of the failures result from the selection of equipment type and from the design of the installation. For example, level sensors can fail if variations in fluid density are not considered. We should not rely on the sensor manufacturers alone to define proof testing and inspection without considering the design of the overall system.

FMEDA provides an objective and rational basis for proof testing and inspection.

Engineers at *exida* developed the FMEDA method by extending FMEA with reliability engineering techniques, including the addition of diagnostic analysis.

For detailed guidance on FMEA refer to the standard IEC 60812 *Analysis techniques for system reliability – Procedure for failure mode and effects analysis (FMEA).*

The probability of errors in design, installation, operation and maintenance may also be estimated through techniques such as HEART and THERP. Refer to the HSE UK publication '*Review of human reliability assessment methods*', prepared by the Health and Safety Laboratory for the Health and Safety Executive 2009.

Techniques such as FMEDA, HEART and THERP can only produce rough estimates of failure probability. Even though the estimates include significant uncertainty they are very useful because they reveal the factors that influence failure performance. They enable the design of systems and procedures to reduce the probability of failure.

Conclusions about improving proof test coverage

Undetectable failures may make a significant contribution to the probability of failure of safety functions.

Some undetectable dangerous failures might be tolerated in SIL 1 functions – if the probability of undetected failure can be estimated and is low enough to meet the target.

Any never-detectable dangerous failures in a safety function are likely to prevent SIL 3 performance from being achieved.

All safety functions must be designed so that they can be fully tested at least when they are first put into service.

In SIL 2 or SIL 3 service the sensors and final elements must be designed so that full testing is possible at regular intervals.

FMEDA is essential as a tool for understanding proof test coverage. The FMEDA should consider the entire safety function from end to end. The process interfaces must be analysed together with the safety function devices.

The objective should not be to calculate a precise value for PTC or for λ_{DN}, the objective should be to ensure that safety functions can be fully tested.

Prepared by Mirek Generowicz, I&E Systems Pty Ltd www.iesystems.com.au

I&E Systems Pty Ltd gratefully acknowledges contributions to this paper from William Goble and Sam Holden of *exida*, and from Harvey Dearden of SISSuite.