Text Proposal for a sub-Section of Section 12. and/or a new Annex on Redundancy in IEEE 802.1DG/D1.3

Max Turner
Ethernovia
Jan. 2021

Redundancy vs. Duplication

This text will refer to Duplication, if two path of equivalent functionality are available. Equivalent meaning the two may not be fully in agreement with each other in terms of the data delivered. This is usually the case if the paths use a different ground truth to derive the data from. An example of this may be two clocks, where one uses a GPS or DCF77 signal as its source, while the other relies on the AC power mains frequency as an underlying beat source. In such a constellation, there needs to be a priori agreement on which system is used under which conditions and if there is a need to monitor the difference between the two, as there may or may not be a limit towards the permissible difference. If not stated differently, this text will assume that duplication has no limit and no monitoring for difference, only for lack of input on one path (failure of the path).
Redundancy on the other hand is defined for this text as two paths which are in full agreement with each other during all times of normal operation. A disagreement beyond a pre-defined limit or lack of signal on any one path represents a failure and requires corrective action. This means there is constant monitoring of the well defined maximum difference required.
For redundancy and duplication it is true, that the overall system can still give the desired information after an initial failure with the same quality as the duplicate or redundant systems can.

Security vs. Safety

While security here is to describe a malicious and deliberate manipulation of software, configuration or communication data, safety (as e.g. per ISO 26262) is usually concerned with random (non deliberate) or accidental changes in software, configuration or communication data.
This document will use safety in the sense of ISO 26262 as well as in the sense of anything that may bodily harm the inhabitants. This would be any loss of control of the vehicle or harmful actions by any vehicle system.
Security is used when deliberate changes to the vehicles software, configuration, or communication data aiming to circumvent limitations imposed by the current user, owner or manufacturer are to be prevented. Examples may be activation of features which are not paid for, using a rental/shared car without payment, or gaining access to steal items from the interior.
Obviously the two intersect where deliberate interference intentionally or unintentionally leads to bodily harm. One could therefore argue, there is no Safety without Security, but there may well be a certain level of Security without Safety.
From an implementation perspective cryptographic functionality used for security is usually difficult to validate against the requirements of safety due to their complexity. This often leads to separate implementations.

Definition of Path

As this section is part of a communications standard, it talks about paths as the elements of a system. The reader should keep in mind, that a system using communication paths, will also entail sources and sinks for the communicated information. Each of these sources and sinks may have internal components which may or may not be considered elements of the communication system. In general the same logic applies to such components inside sources and sinks as we discuss them here for communication paths.

Secondary Failure

This text often speaks about a failure, this is always meant at the very first failure in a previously fully operational system. If nothing else is stated, it is assumed that a secondary failure will terminate the functionality of the system. Taking the example of an emergency exit sign. These are often supplied by mains power as well as a battery. If the mains power fails (first failure), the sign is still illuminated using the battery power. As soon as the battery is drained (secondary failure) and mains is not restored, the sign will no longer be illuminated, meaning the illumination system has failed completely at this point.

Use-Case Lifetime Extension

This text refers to lifetime extension if a system, which is expected to degrade over its life time is equipped with spare resources so the desired functionality is available to the customer for a longer time of use. This may for example be achieved by duplication, where only one path is operational when the product is shipped and a second one lies dormant, only to be activated after the first has either failed or no longer fulfills certain performance criteria. The first path will be deactivated permanently after the second one has taken over and will/can never be used to take over when the second path fails as well. When the last spare path fails, the system fails and will need repair.

Use-Case Fail Safe

A system is called fail safe, if after a failure in can go into a safe state by itself (i.e. without intervention or further input). Depending on the safety assessment the system may be able and allowed to recover after the failure has been removed. One may think of an emergency exit door, which defaults to the open state in case of a loss of electrical power, as this loss of electrical power may be cause by the same event that requires people to exit the building.
It can easily be imagined how safety and security can conflict with each other in many applications.

Use-Case Fail Secure

A fail secure system will assume a secure state by itself (i.e. without intervention or further input) after a failure. The secured door to a bank vault for example will likely not default to the open position upon a power failure.
It can easily be imagined how safety and security can conflict with each other in many applications.

Use-Case Hold-Over

The term hold-over is used in this document to describe a system, which usually processes an input, but is capable (at least for a limited time) of generating a useful output even if this input is lost. The pre-condition being that good input has been received beforehand (for a sufficiently long time) and the output may degrade over time after the loss of input. Depending on the safety assessment the system may allow for a recovery in case the input is re-established without intervention. As an example one may think of a GPS receiver in a navigation system which might assume constant heading and speed for a while after satellite coverage has been lost and recovers after coverage has been re-established, e.g. after a tunnel or overpass.

Do not confuse this with a fail operational system, as the original functionality is actually not available during hold-over. This is also usually not fail safe, as the quality of the available data keeps deteriorating.

Use-Case Fail operational

A fail operational system can fully execute its desired functionality for a limited time after an initial failure. It is set apart from the hold-over capability by the fact that no degradation is expected after the initial failure event if there is no secondary failure occurring. This means a fail operational system will go into a lower reliability state after the first failure. A secondary failure is not covered and will lead to the system failing completely. Depending on the safety assessment and the reliability of the system, there may be a need to limit the time of operation after the initial failure in order to limit the likelihood of a secondary failure. Consider again the example of the emergency exit lighting system.

Use-Case Limp-Home

A system described as limp-home capable extends the time a system can execute its desired functionality after an initial failure as there is a specification and means in place for the system to handle a secondary failure gracefully, e.g. by going into a fail safe state.

Definition Availability

According to [1] “Availability refers to the probability that a system performs correctly at a specific time instance (not duration).” For this discussion the instance in question is the start of the particular system.
For a redundant system it is assumed, that all paths of the redundancy need to operational and agree with each other before the system can be called available. Meaning all paths have to be considered in series connected elements when calculating the overall system availability. This will lead to a reduced overall system availability if one just adds a second path of the same availability as the first.
Keep in mind this limitation may be viewed completely differently of only duplication is desired/required.

Definition Reliability

“Reliability is the probability that a system performs correctly during a specific time duration.” [1]
For the purpose of this document this is interpreted as the inverse probability for a system or a path inside a system to fail from a state of normal operation.
For combining paths into a system, here we may assume them to be in parallel connected elements when calculating the overall system reliability. This will lead to a higher reliability when combining multiple paths of equal reliability, than using just a single one and obviously represents the motivation to have multiple paths in the first place.

Definition Recovery

In this document the term recovery is used to describe a system (with multiple paths or not) capable to go back to normal operation after a failure, if certain conditions are re-established without intervention after the failure. Consider for example the loss of data for a single picture-frame in a TV transmission system due to EMI. While for some time a scrambled image may be displayed, the system will go back to normal operation without intervention from the outside.
Depending on the safety assessment of a system, this may not be allowed or there may need to be a counter (monitoring) of such recovery events and if they occur too often or too frequently, a repair may be required, meaning the system will not ultimately recover.

Definition Repair

A repair in this context means there needs to be some intervention, usually by a person, to restore the functionality of a system after a failure. Before this repair has not happened, the system can not go back to normal operation. Depending on the safety assessment and the type of failure, the system may either be limited in its operation between failure and repair or it may go into a dysfunctional state to prevent secondary failures.

Retransmission vs. Redundancy

Redundancy can be viewed as a very special case of retransmission. While in systems with retransmission (e.g. TCP) information is only repeated (retransmitted) if an issue (loss) has been reported by the receiver, which limits the response time to such losses, in redundant systems all information is always transmitted multiple (at least two) times and the receiver is responsible to discard any duplicated information while in normal operation no loss is occurring. While redundancy requires the full bandwidth on each redundant path, retransmission may allocate a certain amount of added bandwidth, depending on the reliability of the elements in a system and the likelihood of certain failure cases.

Seamless Redundancy

This document refers to a system as first level seamless redundant if a single path can fail and the underlying functionality is retained without any interruption.

Single Point of Failure

Technically any element in the system that lies within at least two paths of the redundancy solution has to be considered a single point of failure. Depending on the reliability and safety assessment of such common elements in at least two paths vs. the reliability and safety assessment of the independent elements and the statistical probability of certain failure cases in the elements, a quantitate analysis of the criticality of such common points may be required. While some system designs avoid single points of failure painstakingly as the quantitative analysis is often difficult, this document wants to leave such design decisions up to the implementer and will therefore refer to relevant single points of failure. Relevant single points of failure are defined as common elements in at least two paths, where there is an unacceptably high likelihood, their failure will lead to a failure of all paths connected. It is up to the implementer to make such judgement calls and decide on the assessment and validation procedures required to ensure sufficient reliability of any single point of failure.

Mechanical aspects of redundancy

As discussed in the concept of the single point of failure, it is important to ensure that a system entailing redundant elements avoids relevant single points of failure. As an automobile is still very much a mechanical system, which is also exposed to the elements, it is important to ensure all paths of a redundant system are not easily subject to the same mechanical influences. This could mean to not attach them to a housing with too close proximity or routing them trough the vehicle on two different sides of the body.

Redundant Electrical Power

All communication systems require (electrical) power to operate. If all paths of a redundant communication system are powered from a single source, it should be taken into account, that this power source may well become a relevant single point of failure.
The same arguments about mechanical aspects of the actual communication need to be applied to the power distribution system.

Safety Decomposition

ISO 26262 allows for certain decompositions of safety relevant systems (ASIL x(y), read as “ASIL x of y”), where two elements of lower ASIL level x are combined to form a system of the higher ASIL level y. It is important to point out, that such a decomposition can not be represented as redundancy under the definitions of this document, as the failure of any one element/path will immediately lead to the failure of the system, as it can no longer operate at the desired higher ASIL level y.

Monitoring requirement

In order to determine if one of the redundant paths has failed, a constant monitoring is required. There are multiple ways of achieving such monitoring. If the path includes not only communication links, but also the data source and sink, it is usually easiest to send cyclic messages across the system, even if there is no change in the data at the source.
A hop by hop monitoring of the link-up status may seem a more bandwidth efficient solution, but requires for each hop to ensure notification of every system affected by a link loss, which may lead to even more communication.

Application to IEEE 802.1CB

IEEE 802.1CB delivers redundancy in the sense of this document. The ground truth here is the data being fed into the splitting element, which is then multiplexed onto the communication paths. A comparison of the data received is not part of the standard, as there is an assumption, the underlying mechanisms ensure there will be no (accidental) alteration of the transported data. Depending on the safety assessment of the application this may or may not be a valid assumption.
As discussed in the section on single point of failure it is important to note, that IEEE 802.1CB relies on a splitting element as well as on a merging element. This concept is likely derived from the telecommunications industry, where a loss of a communication link is very often due to mechanical damage to a cable buried in the ground. Think of a digger on land or a fishing net for undersea cables. The building and area where the two paths (cables) start is usually considered better protected against such mechanical interference. As is pointed out in the section on mechanical considerations, it is important to ensure these assumptions apply to the application inside the automobile, where IEEE 802.1CB is to be used.
Assuming a system with two paths used for redundant data communication, where one has a significantly higher latency than the other (a bulk stream per IEEE 802.1CB), there is an implicit requirement to only process the information when it has arrived via the longer path at the earliest if the system is to allow healing. This is due to the fact, that when the shorter (faster) path fails and later recovers, there would be a gap in the data available for processing and seamless redundancy would otherwise not be feasible. This is outlined in [2] and may be in conflict with a requirement of minimum latency data transmission in the system.
It also requires both paths to be operational for the system to start up, potentially impacting the availability.
Using IEEE 802.1CB for safety decomposition may be possible in cases where the safety level ogf

Bibliography

[1] “System Reliability and Availability Calculations – BMC Blogs.” https://www.bmc.com/blogs/system-reliability-availability-calculations/ (accessed Sep. 18, 2020).

[2] N. Finn, “Seamless Reduncancy Issues - Supporting comments on P802.1CB Draft 2.0,” Dallas, TX, USA, Oct. 22, 2015, Accessed: Jan. 15, 2021. [Online]. Available: https://www.ieee802.org/1/files/public/docs2015/cb-nfinn-seamless-issues-1015-v02.pdf.