Implementing SCADA Redundancy

Walid Ali – GE Vernova, Canada
Justin Turner – GE Vernova, USA

Abstract

While system redundancy is a must have on a protection scheme especially on a transmission substation, redundancy is also highly desirable for substation automation. Redundancy is a balance between the criticality of a substation and equipment reliability and availability. Redundancy can enable the automation system to run flawlessly even through issues like equipment failure and cyber security concerns. This paper will explore the various types of redundancy including system redundancy with emphasis on independent dual redundancy and various types of standby redundancy and combination of both. The paper will describe different considerations when implementing redundancy in various scenarios.

Introduction

Redundancy improves system reliability, maintainability, and availability, particularly through its pivotal role in mitigating failures and minimizing downtime. Redundancy bolsters system reliability by reducing system failure rates. SCADA redundancy involves duplicating critical components within a SCADA system, ensuring seamless operation even in the event of a component failure. Different system redundancy modes, with a focus on the substation gateway or RTU), provide different levels of system availability from the substation to the central energy management center. Consideration of redundancy extends beyond individual systems or devices and encompasses network infrastructures, increasing fault tolerance and ensuring uninterrupted data acquisition and transmission.

Reliability, Availability, and Maintainability

Redundancy plays an important role in improving reliability, availability, and maintainability of systems.

Reliability

Reliability denotes the probability of a system or device operating without failure for a specified duration under defined environmental conditions. It is often quantified by the equation $R=1-F$, where F represents the failure rate of a system. Failure rate of a system is the inverse of mean time between failure (MTBF) or $F=1/MTBF$. This equation simplifies the concept, illustrating the impact of redundancy on system reliability. By considering redundancy, we can refine this equation to $R=1-(F)^2$.

Reliability Example

Consider a scenario where a device exhibits a reliability of 80%. Upon introducing redundancy, the reliability increases to 96%. This increase is evident when solving the equation $R=1-(.2)^2$. However, it is crucial to note that this approach offers a simplified perspective of identical systems.

Maintainability

Maintainability is the probability that a failed component or system will be restored or repaired to a specified condition within a specified period. A failed system that has a low probability of being repaired has low maintainability. A failed system that takes a significant amount of time to repair also has low maintainability. Redundancy may allow for a system to continue to operate while an individual component is out of service for repair or replacement with minimal system downtime.

- Redundancy allows for more bug updates and security patch updates of redundant components in a timely manner.
- Redundancy allows for more troubleshooting by isolating the faulty components or system so root cause analysis can be done at the source.
- Redundancy reduces or eliminates mean time to repair (MTTR) and in turn improves the availability of the system.

Availability

Availability is the probability of the system or device to perform its function and not in a failed or repair state. While there are several classifications of availability, we are going to focus on two in particular:

- Ai (Inherent Availability) = MTBF/(MTBF+MTTR), where MTBF is Mean Time between Failure, MTTR is the Mean Time to repair. Note this equation does not take into account preventive maintenance down time, logistics delay.
- Ao (Operational Availability) = Uptime / Operating Cycle. Ao takes into account all downtime including preventative maintenance and logistics and better reflects the real time experience the system is available.

Availability Example

Redundancy bringing reliability from 99% to 99.99% ($A = 1-(.01)^2 = 99.99\%$) means reducing unavailability from 3.65 days per year to 50 minutes per year and conversely increasing availability.

Reliability	Maintainability	Availability
Constant	Decrease	Decrease
Constant	Increase	Increase
Increase	Constant	Increase
Decrease	Constant	Decrease

Table 1 Relationship Between Reliability, Maintainability and Availability

System Redundancy vs Network Redundancy

System redundancy refers to the duplication of one or more critical components in a system. System redundancy typically applies to hardware and software within a device. In a SCADA system, system redundancy can refer to (but not limited to) the duplication of Substation Gateways, HMIs, or Remote Controlling Stations. System redundancy ensures that if a component fails, another redundant component can take over, minimizing downtime and maintaining system functionality.

Network redundancy involves duplicating or diversifying communications network infrastructure components and paths to ensure continuous connectivity and data transmission in the event of network failures. Network redundancy applies to the networking layer of a communications infrastructure, including (but not limited to) routers, switches, and cables. Network redundancy aims to eliminate single points of failure in the communications network by providing duplicate or alternate routes.

System Redundancy

No Redundancy

In some applications, it may not make sense to have a redundant system. This would be in a system with no criticality or low criticality. A single purpose, low impact system may not need redundancy if it can be easily repaired or replaced without significant impact on operations. In a system with low risk of failure, a redundant system may not be required, and there may not be justification for the additional resources. Additionally, a legacy system may not support redundancy and implementing redundancy may not be feasible without replacing the legacy system and sometimes surrounding infrastructure.

Cold Redundancy

Cold redundancy refers to system redundancy configurations where a secondary system or components are powered off until needed. In a cold redundancy setup, the secondary system is not synchronized with the primary system and require manual intervention to bring the secondary system online and start processing tasks. Cold redundancy is often used in situations where maintaining a powered secondary system is prohibitive, and the downtime associated with transitioning to a secondary system is acceptable. Cold redundancy may be appropriate for a system where the likelihood of failure is low, and the impact of downtime is minimal.

Warm Standby Redundancy

Warm standby is a type of redundancy configuration where the secondary system is powered on and ready to take over in case of the primary system failure. Only one system, typically the primary system, is active at a time. Data synchronization between the primary system and secondary system is minimal. Upon primary system failure, the secondary system will automatically detect the failure and perform the failover operation to the secondary system. A warm standby system allows for minimal, but acceptable, downtime while transitioning from the primary system to the secondary system with only critical data, if any, synchronized between the two systems.

Hot Standby Redundancy

In hot standby redundancy, both the primary and secondary systems are powered on, and the secondary system is ready to take over in case of primary system failure. Only the primary system is active during normal operation, but the secondary system is in constant data synchronization with the primary system, ensuring it can quickly take over operations with minimal downtime and minimal data loss. Upon primary system failure, the secondary system will automatically detect the failure and perform the failover operation to the secondary system. Hot redundancy systems are commonly used in critical applications where minimal operational downtime and minimal data loss is essential.

Hot-Hot Redundancy

A hot-hot redundant system has a primary and secondary system both powered one with the secondary system ready to take over in case of primary system failure. Communications between the primary system and external devices and communications between the secondary system and external devices are constantly active. The primary system will process data as received when online. The secondary system may either ignore the incoming data or process the data simultaneously. If the secondary system is not processing data, the secondary system will be in constant synchronization with the primary system. If the primary system fails, the secondary system will failover seamlessly as the two databases are constantly synchronized and communication links between the secondary system and external devices are already established. Hot-hot redundancy configurations are typically used in very critical applications were any downtime and data loss cannot be tolerated and high availability and fault tolerance are critical requirements.

Hybrid Redundancy

Some redundant systems may employ multiple types of redundancy in the same system. A redundant system may have multiple applications which support different levels of redundancy. Additionally, redundant systems may support different external devices that support different levels of redundancy. For example, a redundant system may have a highly critical application, App1, that supports hot-hot redundancy that is connected to modern external devices that also support hot-hot redundancy. The same redundant system may have a less critical application, App2, that only supports warm standby redundancy and connects to different devices that support warm standby redundancy. This redundant system would be support both hot-hot redundancy and warm standby redundancy depending on the application and capabilities of the external devices.

Serial Network Redundancy

Serial Ring

In a serial ring topology, each device is connected to two devices forming a ring. Data travels around the ring. If one section of the ring fails, all devices are still connected on the communications bus. The communications bus is now a linear bus, instead of a ring. Serial ring topologies are common in RS-485 communications.

Dual Homing or Dual Attachment Concentrator (DAC)

Dual homing in serial network redundancy refers to a configuration where a network device is connected to two separate network paths. Dual homing can be achieved using any device with multiple serial interfaces that can manage failover. Each network interface connects to a separate network path,

providing redundancy in case one path fails. The device itself manages the failover between the network interfaces. Dual homing may be implemented using standard serial interfaces, with standard serial protocols, without specialized redundancy protocols.

Parallel Serial Redundancy Protocol (PSRP)

PSRP is a protocol specifically designed for redundant serial communication links. PSRP duplicates data packets and transmits them simultaneously over redundant parallel communication paths. PSRP includes mechanisms for automatic failover between redundant paths and packet comparison at the receiving end to ensure data integrity. PSRP provides for superior fault tolerance as packet duplication and transmission happens on both communication path simultaneously. As both network paths are active, failover time is minimal due to no failover between serial interfaces.

Ethernet-based Network Redundancy

Single LAN

A single LAN network is not a redundant network. A single LAN network consists of two or more ethernet-based devices typically communicating to each other over copper ethernet cables, fiber optic cables, or various wireless transmission technologies. These devices are all on the same subnet with unique IP addresses or MAC addresses. These devices are often connected through an Ethernet switch. Single LAN networks may be used in smaller, simple networks for non-critical applications.

Dual LAN

A dual LAN network is communicating the same information between the same devices on two independent networks simultaneously. These networks should have at least two networking paths for device communications. These two paths can be completely independent of each other with separate hardware, e.g., separate switches, separate cabling, etc. The paths can be interconnected but special attention needs to be paid to networking loops, and protocols such as rapid spanning tree protocol (RTSP) may need to be used. Each device will need to have two independent ethernet interfaces with unique IP addresses (or have an external dual LAN device), and the two networks should be on different subnets. The dual LAN ports are both active at the same time. Dual LAN networks are two independent single LAN networks that connect the same devices together through different hardware and different subnets. Dual LAN networks are redundant at the device application level do not require any special redundancy protocol. The failover from the primary network to the secondary network is handled at the device application. The client (or controlling device) application needs to support dual LAN, i.e., the application needs to support a primary and secondary IP address. Downstream devices need to have two independent ethernet ports but do not necessarily have to support dual LAN.

Redundant LAN

Redundant LAN network interfaces work as a redundant pair. Each device on the redundant LAN will need to have two ethernet interfaces that supports redundant pair. The ethernet interfaces will share a common IP address and subnet. Redundant pairs are supported at the device ethernet interface level, not the application. Only one ethernet interface will be active at a time. If the primary ethernet interface loses the communication link, communications will switch over to the secondary ethernet interface. Each ethernet interface should be connected to distinct networking hardware and should not

share common networking hardware. Non-redundant devices can also be connected on the primary LAN.

Parallel Redundancy Protocol (PRP)

PRP is a redundancy protocol. PRP network interfaces work as a PRP pair. Each device on the PRP LAN will need to have two ethernet interfaces that support PRP. The ethernet interfaces will share a common IP address and subnet. Both ethernet interfaces will be active at the same time. Ethernet frames are duplicated and sent over both network paths. The two network paths cannot be connected to each other. At the receiving end redundant frames are compared and the primary packet is selected based on predefined criteria. Upon failure of the primary ethernet communication link, communications will switch over to the secondary ethernet interface seamlessly. Since both links are active and communicating duplicate data simultaneously, failover will be immediate with no data loss. Each ethernet interface should be connected to distinct networking hardware and should not share common networking hardware. PRP does not require a specific topology and can handle complex topologies. Failover times in PRP networks are deterministic which makes PRP suitable for networks that require precise time synchronization. PRP networks can also support non-PRP devices.

High-availability Seamless Redundancy (HSR)

HSR is also a redundancy protocol that requires two ethernet interfaces per device that support HSR. HSR creates a redundant ethernet ring topology where data is transmitted in both directions, a primary direction and secondary direction, around the ring. Each device in an HSR ring replicates and forwards ethernet frames not intended for that device from one port to the next creating a ring. If a link or device in the ring fails, HSR detects the failure and reroutes traffic in the opposite direction around the ring. HSR may require less networking resources since it employs a ring networking topology and does not require the duplication of networking hardware. HSR does require a ring topology and all the devices in the ring must support HSR. HSR networks may face bandwidth constraints since networking infrastructure is not duplicated, each device must support redundant network traffic.

Redundancy Implementation Considerations

One of the main considerations in implementing redundancy is the criticality of the system. Systems that require high availability with low loss of current and historically in which downtime will have a significant impact on operations should consider more advanced redundant systems. While less critical, low impact system may require less redundancy or no redundancy. Other considerations when implementing redundancy is the amount of resources required to implement redundancy and the ability of legacy systems to be implemented into a redundancy scheme.

High Criticality Systems

In a highly critical substation, such as a remote, high voltage, interconnection substation, SCADA redundancy is very important. The impact to operations plus the low maintainability due to its remote location could lead to a significant and time-consuming impact in service to a large area. At such a site multiple levels of redundancy should be considered. Some level of system redundancy, hot-hot, hot standby, or warm standby, should be considered at the RTU or Gateway level and a highly available

network redundancy such as PRP or HSR may be implemented at the LAN level. A redundant SCADA network leaving the substation should also be considered.

Medium Criticality System

Medium criticality substations, for instance a transmission or large distribution substation, may still have a need for some SCADA redundancy but the impact of failure may not justify the same level of redundancy as a highly critical site. A medium criticality site with a highly reliable substation gateway may only employ highly available network redundancy locally such as PRP or HSR. A redundant SCADA network leaving the substation can also be considered.

Lower Criticality System

Less critical substations, for example a medium distribution substation, may still employ some level of redundancy. A less critical substation may only have a single substation gateway, a redundant LAN with slower failover (seconds as opposed to milliseconds), and only a single external SCADA connection.

Non-Critical System

Sites with little or no criticality to the, such as a small distribution station serving a small number of customers who may have a backup connection to the grid, may need little to no SCADA redundancy. There may be no redundancy or a mix of redundant and non-redundant devices. Only a single gateway and single external SCADA connection may be needed.

Conclusion

Redundancy, both at the system and network levels, emerges as a cornerstone in SCADA system reliability, availability, and maintainability. Warm standby, hot standby, or hot-hot redundancy configurations offer distinct advantages in mitigating risks and sustaining operational resilience. Network redundancies such as PRP, HSR, redundant LAN, and dual LAN further enhance fault tolerance and ensure seamless data acquisition and transmission, increasing overall SCADA system robustness. By embracing the strategic deployment of redundancy strategies, engineers can navigate the complexities of modern SCADA systems, safeguarding against disruptions and implementing sustained operational excellence.