NERC
NORTH AMERICAN ELECTRIC RELIABILITY CORPORATION

Lesson Learned

Loss of Energy Management System Functionality due to Server Resource Deadlock

Primary Interest Groups
Reliability Coordinators (RC)
Transmission Operators (TOP)
Balancing Authorities (BA)

Problem Statement
An antivirus software engine installed on energy management system (EMS) production servers had a flaw that caused affected servers to deadlock server resources and become unresponsive, effectively making the EMS unavailable to operators. This flaw was latent in the engine and activated by certain malware signatures applied to the EMS production environment. The flaw was not recognized in test environments due to the difference in input/output (I/O) workload on test servers versus those seen on production servers. Deadlocks only appeared on live production servers that have extremely high file I/O, causing the server deadlock flaw to manifest more quickly. The high file I/O on these servers is due to routine EMS processes and file backup services that resulted in more opportunity for the antivirus engine to deadlock on files.

Details
Two separate events over the span of two weekends led to a period of 31 consecutive minutes of complete loss of EMS functionality; this occurred again on the following Saturday for a period of 81 consecutive minutes.

These performance degradation events removed the ability to control Bulk Electric System (BES) elements at the impacted substations, and the entity was unable to calculate Reporting Area Control Error (ACE), control performance standards, or implement automatic generation control. The entities' state estimator (SE) and real-time contingency analysis (RTCA) were not solving, and real-time monitoring and alarming was not functioning on the EMS.

The impacted entities' System Operators were able to implement a loss of primary tools process for the duration of the events. The process allowed for separate and independent supervisory control and data acquisition (SCADA) systems to maintain monitoring of all BES substations by operating personnel at regional dispatch centers who would provide notification to the impacted entity of any abnormal conditions or operations. The entities RC was notified of the failure and verified that their RTCA was still solving accurately.

Deadlock:
In an operating system, a deadlock occurs when a process or thread enters a waiting state because a requested system resource is held by another waiting process, which in turn is waiting for another resource held by another waiting process. If a process remains indefinitely unable to change its state because resources requested by it are being used by another process that itself is waiting, then the system is said to be in a deadlock.
Silberschatz, Abraham (2006). Operating System Principles (7th ed.). Wiley-India. p. 237. ISBN 9788126509621

The entity received calculated ACE from the RC every 10 minutes to assess load, generation, and interchange and dispatched units manually. The System Operator, with the assistance of the RC and assisting operating personnel at remote sites, was able to assess the system status for pre and post contingent system operating limit violations or encroachments, system frequency, and maintain situational awareness sufficient to perform real-time assessments.

Due to the first instability of the EMS event, the entity declared a Conservative Operations Alert which suspends all work on critical infrastructure systems such as SCADA/EMS, ICCP maintenance, telecommunication equipment, and relaying, unless such maintenance was emergency support work that would result in improved BES monitoring, control, and reliability. The alert also initiated a review of staffing levels to ensure appropriate staffing was available to manage conditions.

The entity had an emergency callout process that was utilized for each event to engage the incident response team (IRT). The EMS system was restored to a functional state following each event. The deadlocked servers did not trigger the EMS automatic failover process and all performance counters (CPU, memory, network, disk, etc.) on the servers remained within limits. EMS personnel had to perform manual failovers to available servers via script and manually reboot affected systems in some instances. The performance issue was resolved from recurring on the first event by the proactive step of the IRT to disable all non-critical services on the active EMS servers. The antivirus software vendor provided a proposed identification of a root cause being a flawed signature. Following this potential root cause identification, the IRT applied a new signature and re-enabled non-critical services on the EMS servers.

Five days later, the incident recurred and the manual soft re-boot process did not successfully restart the EMS processes. Additional support personnel were brought into the incident response to quarantine the impacted server for forensic analysis and to perform a hard reboot of the servers. The performance issue was again resolved from recurring by the proactive step of the IRT to disable all non-critical services on the active EMS servers. Through IRT testing, most servers were determined to be safe and were re-enabled. Through forensic analysis, the IRT identified that the deadlock could be duplicated when the EMS processes, backup service process, and anti-malware processes ran at the same time. The antivirus vendor confirmed that the root cause was a malware engine flaw that was exercised by certain signatures.

Corrective Actions

Several immediate corrective actions were implemented to resolve the EMS server instability and return the system to a stable and reliable state. These included the disabling of select services, uninstallation of the flawed malware engine, and a recovery process for restoring services after application of patches containing a fixed malware engine.

An event analysis team identified areas of further investment in the operational technology value stream to prevent recurrence of a similar event. The areas assessed in event analysis covered the following components of the operational technology value stream:

- Prevention: How could the issue be prevented from happening?
  - Implement holistic architecture changes to the EMS to enhance system resiliency

  - Work with the EMS vendor to seek improvements to the system failover processes that allow failovers during non-responsive server performance issues
  - Revise the patching and patch testing processes to harden the process and catch challenging-to-discover issues like this patch issue ahead of time
- Detection: How could the issue be detected more quickly?
  - Develop a central incident response toolkit to avoid spending valuable time in the initial response doing low-value data collection activities. This includes a dashboard for EMS failover tools, Windows logs, EMS logs, patch reports, performance tools, and process status to assist IRT in identifying issues
  - Develop proactive alerts based on logs and performance counters to notify the IRT when performance issues are detected
  - Develop test environments that have the same or similar level of I/O workload as production
- Incident Response: How could the response process be improved?
  - Develop production-like test environments that can be used to simulate incident response
  - Implement a training program for incident response using scenario-based response activities
  - Implement a more rigorous testing process during incident response to verify the root cause prior to concluding the incident response

Lesson Learned

- **Verify Vendor Root Cause Assertions**
  During the initial event, the antivirus software vendor initially misdiagnosed the root cause as a signature flaw. Following the second event, the vendor provided an updated root cause that the flaw was in the malware engine and that certain signatures exercise the flaw in the engine. The lesson learned is that vendor assertions should be tested in a more rigorous way before concluding they are the correct root cause. A more rigorous testing process should be implemented during incident response to verify the root cause prior to concluding the incident response.
- **Shared Physical Space Enabled Complex Troubleshooting**
  The multi-day activities to diagnose and resolve the underlying root cause were all performed in a central "war room" where work was coordinated, information shared, sub-teams chartered, and tasks given. Pulling all incident responders into a shared physical space enabled speedy, low-waste troubleshooting for a complex problem to diagnose. Sub-teams would break off into separate meeting rooms near the central war room and come back at defined times during the day to share learnings and determine next steps. This room also became a place where the team's work was made visible with active hypothesis, tests, and sub-team tasks and learnings posted on the wall for all members to see and update together. The shared physical space allowed the team to decompose a complex problem, manage many cross-discipline contributors, and maintain momentum and energy during the long response.

- **Maintain an Operations Presence**
  The IRT staffed an office in the System Operations Control Center during all incident response activities. IRT members assigned to the shared Operations office maintained a continuous video feed to the primary IRT "War Room" acting as a direct link for the IRT to System Operators. Having staff in physical proximity to Operations during the event allowed the IRT to be notified immediately of any situation changes, provided a visible sign of the incident response to operators, and allowed the IRT to coordinate activities quickly with operators. The lesson learned is a positive one, that maintaining staff in physical co-location to Operations during the event was a valuable action that should be part of an incident response playbook.
- **Remote Connectivity Sped Initial Response**
  A 2018 EMS upgrade enabled the use of secure remote connectivity to support an incident response. Prior to the 2018 upgrade, incident responders had to drive onsite to gain access to the EMS and perform response activities. Secure remote connectivity shaved significant time off the response. In two cases, the system was brought back on-line by using the secure remote connectivity, avoiding the time required for a responder to drive onsite. In addition, during these remote responses, the team used video conferencing tools to pull all incident responders into a single video conference to coordinate decisions and work.
- **Have Central Incident Response Tooling and Training**
  An early finding from the event analysis team is that modern EMS technology environments make incidents more complex to respond to. The instability events required several individuals across multiple teams to perform actions to bring the system back to stability. The events also required the responders to use multiple disparate tools (EMS failover tools, Windows logs, EMS logs, patch reports, performance tools) to build a picture of the situation. The lesson learned is that responders should have a central incident response toolkit to avoid spending valuable time in the initial response doing low-value data collection activities. Those responders also need to have scenario-based training on events so they are ready to use these tools when an incident occurs.
- **Forensic Analysis in Incident Response**
  When the incident recurred, the IRT quarantined the impacted server for forensic analysis. Having a quarantined server allowed the IRT to test theories and determine which services could be restored without causing the issue to reoccur. Future incident response will include quarantine of an impacted server for forensic analysis. The capability to perform forensics to quarantine a server, snapshot its state, and clone multiple instances of the failed server to perform forensic tests was enabled by the virtual server infrastructure.

NERC's goal with publishing lessons learned is to provide industry with technical and understandable information that assists them with maintaining the reliability of the bulk power system. NERC is asking entities who have taken action on this lesson learned to respond to the short survey provided in the link below.

Click here for: Lesson Learned Comment Form

For more Information please contact:

NERC – Lessons Learned (via email)

Lesson Learned #: LL20220901

Date Published: September 28, 2022

Category: Communications

This document is designed to convey lessons learned from NERC's various activities. It is not intended to establish new requirements under NERC's Reliability Standards or to modify the requirements in any existing Reliability Standards. Compliance will continue to be determined based on language in the NERC Reliability Standards as they may be amended from time to time. Implementation of this lesson learned is not a substitute for compliance with requirements in NERC's Reliability Standards.