UNOS
UNITED NETWORK FOR ORGAN SHARING

Saving lives together

700 North 4th Street, Richmond, VA 23219
tel: 804-782-4800
fax: 804-782-4816

unos.org

Maureen McBride, Ph.D.
Executive Director & CEO

June 2, 2023

The Honorable Ron Wyden
Chairman
Senate Committee on Finance
219 Dirksen Senate Office Building
Washington, D.C. 20510

The Honorable Charles Grassley
Member
Senate Committee on Finance
135 Hart Senate Office Building
Washington, D.C. 20510

VIA ELECTRONIC MAIL

Dear Chairman Wyden and Senator Grassley,

I appreciate your continued engagement as we collectively work to improve the organ donation and transplant system. I am writing to provide information about UNOS' ongoing actions to strengthen the Organ Procurement and Transplantation Network's (OPTN) IT infrastructure to prevent future UNet^{SM} outages. Our Senior Director of Enterprise Information Technology, Tiwan Nicholson, who participated in the April 20, 2023, briefing with your staff, worked with members of his team to develop the specific actions outlined below, which are aimed at further securing and improving the OPTN system and safeguarding it against future disruptions. UNOS IT staff identified the root cause of the February outage on May 10, 2023, and the root cause analysis is attached for reference. In addition to the incident-specific corrective and preventive actions detailed in the root cause analysis, more generalized short-term and long-term improvement actions are summarized below.

I. Short-Term Action to Improve the System: Enhanced Monitoring and Alerting (Expected Completion by the End of August 2023)

Based on the complex series of events that occur during an automatic database failover[1], our database administrators incorporated additional monitoring and alerting in the weeks following the outage, providing full visibility of all the database conditions that would lead to an automatic failover attempt. This enables the team to engage more quickly and take remedial actions should any type of cluster error[2] occur in the future. In addition, we are expanding our full Gigamon network packet monitoring[3] throughout the UNet application stack for real-time packet visibility should a network-based issue occur. This work is in progress with expected completion by the end of August 2023.

[1] An automatic database failover is the process by which an application is moved to standby servers during a system failure to minimize downtime.
[2] A cluster error refers to the loss of connection or interaction between servers in a clustered group of computers designed to provide functional redundancy.
[3] Gigamon network packet monitoring is technology to continuously secure, monitor, and manage the network infrastructure.

With this enhanced monitoring and alerting in place, a similar event to the one that took place on February 15 would be caught and resolved within a matter of minutes. This will help further ensure system availability and improve our ability to identify and resolve potential issues before they grow in scope.

II. Long-Term Action to Improve the System: Database Hosting Evaluation (Expected Completion by Mid-2024)

While our database tier is currently hosted in our private cloud environment on Nutanix, this incident, as well as the ongoing search for a root cause, have encouraged us to speed up our plans to transition the UNet database into one of the Azure public cloud database platforms as a service (PaaS) offering. We are currently considering two options: the SQL Managed Instance and Azure SQL solutions. Both tools comply with industry best practices and will provide an additional layer of security.

The transition to a public cloud database platform was already in our organizational plans as part of efforts to obtain Federal Risk and Authorization Management Program Authority to Operate (FedRAMP ATO) certification[4]. Considering this incident, we are now working with Microsoft Tech for Social Impact to review the database configuration to determine compatibility for migrating to Azure. This is a first step in migrating UNet into Azure, which is expected to be complete by mid-2024.

We share your concerns about any disruption to system access. UNOS will undertake the short-term and long-term activities described in this letter in addition to our regular and ongoing system improvements, including our annual HRSA audit. Taken together, these actions continue to drive progress within the system, aligning with our efforts to maintain a safe, efficient and effective IT infrastructure.

Sincerely,

Maureen McBride

Maureen McBride, Ph.D.
Chief Executive Officer, United Network for Organ Sharing

Enclosure:
IT Operations Incident Report and Root Cause Analysis

[4] FedRAMP ATO certification is the gold standard for secure and modern cloud-based services for government contractors. Obtaining a FedRAMP certification must either be sponsored by a federal government agency or approved by the FedRAMP program management office (PMO).

UNOS UNITED NETWORK FOR ORGAN SHARING

IT Operations Incident Report and Root Cause Analysis (Final)

Incident Type: UNet full system outage (Category 1 incident, as per PWS Task 3.4.1)

Date & Duration of Incident: 2023-02-15 from 07:05AM ET - 07:56AM ET (51 minutes)

Incident Summary

On the date and time specified above, UNOS Technical Operations Center began receiving UNet automated monitoring alerts indicating a UNet availability issue. Shortly thereafter the Organ Center and Customer Service Desk began receiving calls from users reporting UNet being unavailable. Initial troubleshooting indicated the Microsoft SQL Server Always On database cluster was in an unhealthy state rendering the UNet database inaccessible. The active primary database node (the Active Node) in the SQL database cluster was restarted, restoring the entire cluster to a healthy state. Other conditions noted just prior to the incident and following the restore of the database cluster include UNet System Message Block (SMB) file export time-outs followed by failures and the inability of the Active Node to establish proper connectivity with the UNet Resource Server (Resource Server), which functions as a file repository for the UNet application.

Investigation Details

The root cause investigation examined all activities occurring in the UNet computing environment leading up to, and during the time of the incident. Initially cases were opened with Microsoft (including both Windows and SQL Support teams) followed by other key vendors including Veeam, Nutanix, Tenable, and Fortinet. Each product vendor reviewed available logs, diagnostic data, and configuration information to help determine their product's potential contribution to the outage. The investigation ultimately centered on the initial stress being experienced on the Active Node, and the subsequent behavior of the SQL Server Always On Availability Group (SQL AG), where instead of failing over to another healthy node, that process failed resulting in the entire SQL AG going offline.

The investigation initially focused on the interplay between Microsoft Windows Server Failover Cluster (WSFC) and SQL AG. The scope quickly expanded to include other vendors as Microsoft was unable to provide any initial actionable insight based on the logs. Early on, UNOS attempted to recreate the 2/15 conditions in a test environment with limited success; however, testing did indicate that a distressed Resource Server could impact the performance of the Active Node. Over the course of the investigation, UNOS worked with both Microsoft support teams and initiated a Microsoft Risk Assessment Program (RAP) as a Service Complex for SQL Server in order to review our SQL configuration. The assessment was performed resulting in no critical or significant findings as well as no insight for a root cause of the 2/15 incident.

UNOS also engaged Nutanix, our Hyperconverged Infrastructure (HCI) platform vendor, to help with determining a root cause. UNOS initiated a Nutanix Ultimate FitCheck Assessment to help determine if our configuration could have contributed to the 2/15 incident. Similarly, this assessment yielded no significant insight to root cause or issues noted with the UNOS Nutanix configuration.

While all other vendor support cases and configuration reviews also yielded no additional insight into root cause, UNOS continued its attempts to recreate the conditions of 2/15. A final test conducted on 5/10 provided the most definitive results and indicated the most probable root cause of the 2/15 incident. During this test, the Resource Server was subjected to the same operating conditions as occurred leading up to the 2/15 incident, and both the Resource Server and the Active Node rapidly began to exhibit the same degradation behaviors that were observed prior to the outage.

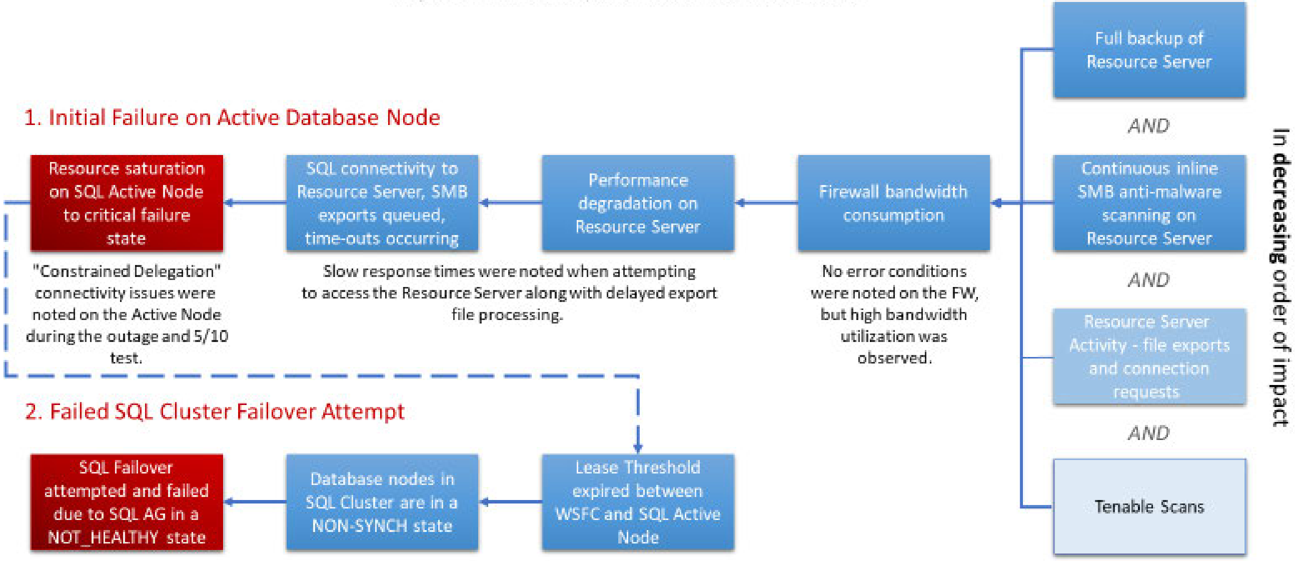
Root Cause Determination

Based on our testing results as well as the information provided through our vendor support cases and assessments, UNOS concludes that under a unique set of conditions, the Resource Server became in a distressed state, which led to process queuing and severe resource contention on the Active Node, which in turn led to the SQL AG attempting but failing to perform a clean automated failover *because of* the degradation (but not complete failure) of the Active Node. This is a crucial detail; had the Active Node fully failed and crashed, the SQL AG would have successfully executed an automated failover. However, the degraded state that the Active Node was experiencing actually *prevented* the failover from executing successfully. The detailed sequence of events is described below.

1. A confluence of events attributed to the degraded performance of the Resource Server which prevented the Active Node from properly communicating to it. These events included:
   - A Veeam full backup running against the Resource Server;
   - Fortinet firewall anti-malware scanning of inbound Resource Server file traffic;
   - SQL Server-to-Resource Server activity including file exports and connectivity requests; and
   - Tenable security vulnerability scans running against the SQL Server and Resource Server, but to a lesser degree based on our 5/10 testing.
2. The Active Node, unable to reach the Resource Server, began queuing SMB-based file export activity until resource utilization on the Active Node reached a critical failure state.
3. The SQL AG detected the distress on the Active Node and attempted to execute a failover to another healthy node. However, the cluster itself was not in a healthy state as a lease time-out condition occurred between the WSFC and the Active Node, resulting in a critical non-synchronous state among the database nodes (see Action Item #4 in next section for more detail on this condition). This prevented a successful failover, and instead caused the entire SQL AG to go offline.
4. The distressed Active Node, while it remained powered on, was in a completely disabled state due to resource saturation until a forced reboot was initiated. The reboot enabled the cluster failover to complete successfully.

See **Figure 1** for a visual representation of the root cause.

Figure 1. Visual Representation of Root Cause


1. Initial Failure on Active Database Node
Resource saturation on SQL Active Node to critical failure state
SQL connectivity to Resource Server, SMB exports queued, time-outs occurring
Performance degradation on Resource Server
Firewall bandwidth consumption
Full backup of Resource Server
AND
Continuous inline SMB anti-malware scanning on Resource Server
AND
Resource Server Activity - file exports and connection requests
AND
Tenable Scans
In decreasing order of impact
2. Failed SQL Cluster Failover Attempt
SQL Failover attempted and failed due to SQL AG in a NOT_HEALTHY state
Database nodes in SQL Cluster are in a NON-SYNCH state
Lease Threshold expired between WSFC and SQL Active Node

"Constrained Delegation" connectivity issues were noted on the Active Node during the outage and 5/10 test.

Slow response times were noted when attempting to access the Resource Server along with delayed export file processing.

No error conditions were noted on the FW, but high bandwidth utilization was observed.

Corrective Actions and Future Risk Mitigations

As a result of this incident, UNOS has identified the following action items specific to this incident along with improvements to our preventive and reactive capabilities for critical incidents.

1. **<u>Veeam backup configuration changes:</u>** Under normal processing, an incremental backup (i.e., changes only) of the Resource server is performed every 90 minutes rather than a full backup. UNOS determined there were more full backups being performed than what normally should occur. Working with Veeam, UNOS performed configuration changes to reduce the number of backup retries along with the disabling of a disk utility on the Veeam backup repository server. These changes have resulted in no full backups being initiated since 2/22/23 when the changes were made, thus reducing the possibility of a full backup being initiated. **This activity is complete.**
2. **<u>Resource Server backup reconfiguration:</u>** Due to the criticality of backing up the Resource Server, there still exists the potential of future full backups being initiated under certain conditions even with the changes UNOS has implemented. As a result of our 5/10 testing, UNOS has configured a second virtual Network Interface Connection (NIC) on the Resource Server and has routed this backup traffic to [redacted]. This configuration will greatly reduce firewall bandwidth consumption and other traffic contention on the Resource Server. Although the firewall does not scan backup traffic for malware and the firewall itself showed no signs of session or bandwidth issues, moving this backup traffic off the firewall is best practice and will mitigate any future bandwidth concerns. **This activity is complete.**
3. **<u>Additional SQL cluster error alerting:</u>** Based on the complex series of events which occur before and during a database failover, UNOS DBAs have incorporated additional monitoring and alerting recommended by Microsoft in order to have more visibility of all the precursors leading up to a potential lease timeout condition and cluster failure. This would enable UNOS to engage

more quickly and take remedial actions should any type of cluster error occur in the future. **This activity is complete.**

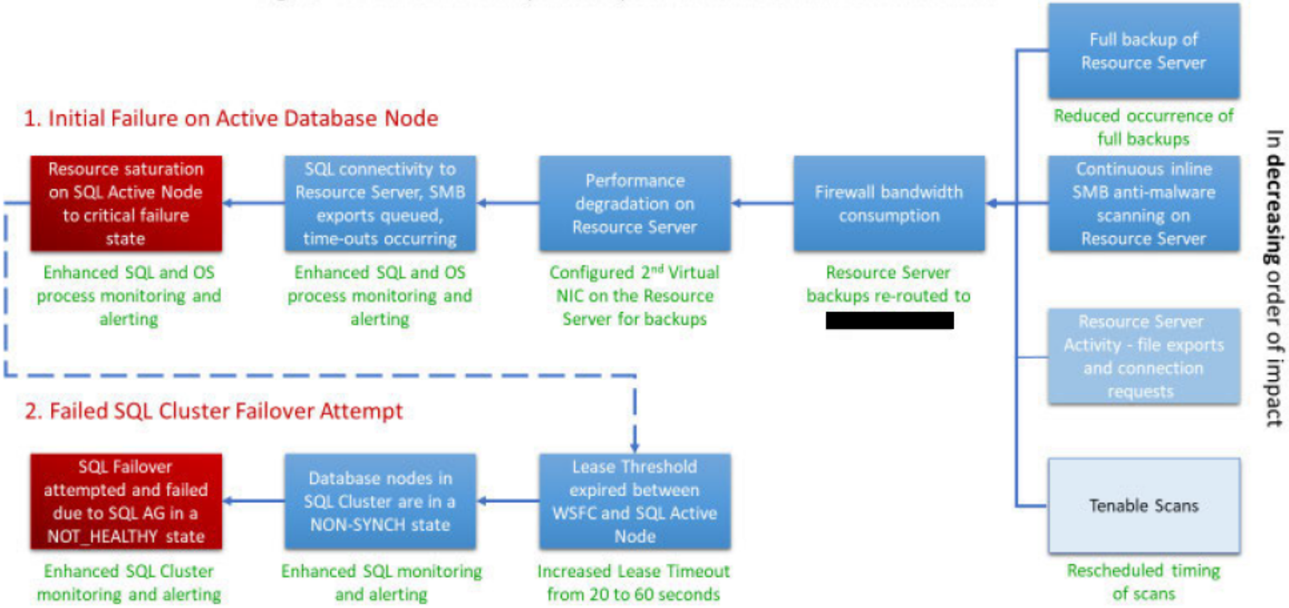
4. **SQL Lease timeout increased:** The lease mechanism (released with SQL Server 2017) is a heartbeat-type counter used by SQL Server to determine whether the Active Node "looks alive". If the Active Node does not reply within a set timeframe (default is 20 seconds), the lease is declared "expired", and to prevent a "split-brain" scenario (that is, multiple nodes thinking they are the primary), the entire SQL AG takes itself offline, ceasing all replication between nodes and preventing further data modification. (See How It Works: SQL Server AlwaysOn Lease Timeout - Microsoft Community Hub for a detailed explanation of this feature.) Microsoft recommended increasing our timeout to 60 seconds. Microsoft also confirmed that given the duration of the 2/15 outage, the restarting of the Active Node to force the failover was the appropriate intervention. **This activity is complete.**
5. **Microsoft RAP as a Service for Complex SQL recommendations:** UNOS is currently reviewing the recommendations from this assessment and is in the process of addressing them. **This activity is in progress, estimated completion Q4 FY2023.**
6. **Nutanix Ultimate FitCheck recommendations:** UNOS is currently reviewing and addressing the recommendations from this assessment and is in the process of addressing them. **This activity is in progress, estimated completion Q4 FY2023.**
7. **Process Schedule Review:** UNOS has reviewed the timing and frequency of known processes that run against the UNet components. In addition to the improvements to the Veeam backup configuration, particular focus was given to nightly SQL Server backups (performed by SQL Server itself to a separate staging location before it is picked up by Veeam), as well as Tenable security scans. UNOS enabled compression (enhanced in SQL Server 2019) on SQL backup jobs, resulting in an improvement of about 50% on backup creation times, 70% on restore times, and 85% less storage utilization. We are also in the process of modifying the Tenable SSP compliance scans which currently run for about [redacted], to a schedule [redacted]. **This activity is in progress, estimated completion Q3 FY2023.**
8. **Automated remediation scripts:** Our DevOps automation team is reviewing available logs and monitoring diagnostics to determine if there are any opportunities to develop Ansible-based automation scripts to potentially intervene should the SQL AG encounter a similar scenario to what occurred on 2/15. Such scripts could be designed to resolve any Active Node conflict by taking the offending node offline and enabling the cluster to successfully failover to the healthy node. **This activity is in progress, estimated completion Q4 FY2023.**
9. **Full network packet monitoring:** UNOS is currently reviewing the costs, performance impact, and operational complexity to extend Gigamon packet monitoring throughout the UNet application stack so that real-time packet visibility is available should an inter-component network issue occur. **This activity is in progress, estimated completion is Q4 FY2023.**
10. **Expansion of IT Customer Service Desk coverage:** Our IT Customer Service Desk (CSD) currently operates from 7am–7pm Monday through Friday, and calls roll over to the Organ Center (which is primarily a clinical support function) overnight and during the weekends. While we have On-Call escalation with our engineering teams, we also recognize the need for more effective Level 1 technical triage on a 24x7 basis in support of our contractual 99.9% availability target. To this end, work is underway to assess requirements for expanding the CSD to 24x7 coverage and equipping the team with monitoring dashboards and automated scripts to aid in early detection

and resolution of issues before users call in. **This activity is in progress, estimated completion Q4 FY2023.**

11. <u>Incident Response Orchestration:</u> Complementing the 24x7 expansion of the CSD, we are evaluating top Incident Response Orchestration and Automation tools, which would provide a curated and enriched view of all real-time logs and diagnostics during an incident, along with pre-configured escalations, automated remediations, and communications to aid the incident commander and support teams managing incidents. This solution would greatly reduce the manual steps and drive consistency in response actions during time-critical issues. **This activity is in progress, estimated completion Q4 FY2023.**
12. <u>Database hosting evaluation:</u> While our database tier is currently hosted in our private cloud environment on Nutanix and has been confirmed by the vendor to be configured properly, this incident has caused us to consider moving up our plans to transition the UNet database into one of the Azure public cloud database platforms as a service (PaaS), such as SQL Managed Instance or Azure SQL. This activity was already in our plans as part of our actions to obtain FedRAMP Authority to Operate (ATO), but considering this incident, we are also evaluating the implications of initiating that migration project in the shorter term. **This activity is in progress, estimated completion Q1 FY2024.**

See **Figure 2** for a visual representation of the improvements that have been made specifically to the various components of the root cause.

Figure 2. Overview of Component-Specific Actions to Address Root Cause


Full backup of Resource Server
Reduced occurrence of full backups
1. Initial Failure on Active Database Node
Resource saturation on SQL Active Node to critical failure state
Enhanced SQL and OS process monitoring and alerting
SQL connectivity to Resource Server, SMB exports queued, time-outs occurring
Enhanced SQL and OS process monitoring and alerting
Performance degradation on Resource Server
Configured 2nd Virtual NIC on the Resource Server for backups
Firewall bandwidth consumption
Resource Server backups re-routed to [redacted]
Continuous inline SMB anti-malware scanning on Resource Server
Resource Server Activity - file exports and connection requests
In decreasing order of impact
2. Failed SQL Cluster Failover Attempt
SQL Failover attempted and failed due to SQL AG in a NOT_HEALTHY state
Enhanced SQL Cluster monitoring and alerting
Database nodes in SQL Cluster are in a NON-SYNCH state
Enhanced SQL monitoring and alerting
Lease Threshold expired between WSFC and SQL Active Node
Increased Lease Timeout from 20 to 60 seconds
Tenable Scans
Rescheduled timing of scans