Root Cause Analysis for PRB000000001088

(SeaQuest Data Taking Interrupt)

Report submitted 2015-Feb-19

(CS-doc-5508)

(v2, amended to change perspective back to time when investigation concluded/first draft report was written and minor corrections)

Summary of Incident

On Friday November 7th, 2014 node e906-gat3.fnal.gov started to contact the Fermilab Kerberos key distribution center, KDC, for unknown clients/servers at an excessive rate. Within days the rate increased to around 400k requests a day. On Wednesday December 3rd the rate more than doubled and Computer Security issued a network block for the machine. The data acquisition, DAQ, application on the machine could no longer access target data information and data taking of the SeaQuest experiment came to a stop. People on the experiment traced the stop to the network block and notified Computer Security. The block was lifted 41 minutes after being applied and data taking resumed. The excessive KDC contacts were debugged and resolved in the evening of that day.

While the technical issue causing the excessive KDC contacts is well understood, the fact that data taking of an active experiment was interrupted triggered the investigation.

List of related Incident tickets:

- INC000000471681 Tu, 2014-Nov-18 Close to 400,000 failed Kerberos requests

Background Concepts

The following points are useful in understanding the nature and impact of the problem:

- SeaQuest is one of the smaller Fermilab experiments. Stage 1 started in 2001 with small Fermilab participation and little support from the Computing Sector. In 2012 SeaQuest started to use more of Fermilab computing/computing services and a liaison was assigned.
- The SeaQuest data acquisition setup has evolved over the years. The SeaQuest DAQ machines are all in a private network with node e906-gat3 being one of the gateways to reach them. The different targets used by the experiment are controlled/monitored via a machine of the Particle Physics

Division, PPD, which is on the standard Fermilab network and not in the private DAQ network.

- SeaQuest's online and DAQ machines are system managed by the DAQ team of the collaboration. This is common in many experiments. The DAQ systems include machines recycled from the excess pool for backup systems.
- The experiment and control room are located in the old KTeV building. It is currently the only experiment in this location.
- The Scientific Computing Division, SCD, liaison to SeaQuest is not a member of the collaboration. The liaison and SeaQuest computing coordinator usually meet every two weeks to discuss open issues and tickets but had not met for over a month due to scheduling conflicts.
- Authentication at Fermilab is based on Kerberos. The key distribution center, KDC, consists of a master and about a dozen slave servers. Linux desktops and servers are installed preconfigured for Kerberos.
- Computers, on-site and off-site, are regularly misconfigured and flood the Fermilab KDC with false requests.
- The Fermilab Kerberos administrators and support staff are not necessarily trained system administrators and/or have online/DAQ experience.
- The Computer Security Team, CST, uses a commercial product to monitor log files. The license of the product restricts the amount of data that can be analyzed.
- November 27th was Thanksgiving and thus followed by a long weekend in the US.
- SeaQuest management followed up with the DAQ team on the issue via email during the long Thanksgiving weekend.
- While searching for the source of the excessive false KDC requests the network file system, NFS, mounts were considered a possible source and investigated for almost two weeks. This turned out to not be the source.
- The excessive false KDC requests were the result of a secure shell, ssh, misconfiguration on the DAQ machines in the private network. Ssh and scp commands from those machines to e906-gat3 triggered the secure shell daemon on e906-gat3 to contact the KDC for unknown clients/servers.
- The ssh misconfiguration is a known culprit for excessive false KDC requests. ServiceNow KnowledgeBase articles exist on the topic.
- Fermilab has a system administrator group that manages control room and DAQ machines for other experiments.
- Fermilab has a Linux user support email list.
- A small section in the 2008 memorandum of understanding, MoU, between SeaQuest and Fermilab covers computing. It lists networking in the old KTeV building as Computing Sector responsibility. Service and/or support level, etc. are not specified. Experiment and SCD are working on a Technical Statement of Work to reflect the additional computing resources used/provided.
- There is no critical systems document for the SeaQuest DAQ system.
- Node e906-gat3 is not exempt from the auto-blocker.

Timeline

This is a partial timeline for the service disruption and restoration.

Fr, 2014-Nov-07	node e906-gat3.fnal.gov starts to contact the Fermilab KDC for non-existing clients/servers
Tu, 2014-Nov-11 11:20	email from Fermilab Kerberos administrator to the SeaQuest e906-gat3 system administrator about excessive failed Kerberos requests from node e906-gat3.fnal.gov
We, 2014-Nov-12 16:30	Fermilab Kerberos administrator points SeaQuest e906-gat3 system administrator and DAQ team to the two existing secure shell configuration KnowledgeBase articles on the subject, KB0011329 and KB0011330.
Tu, 2014-Nov-18 15:30	ServiceNow ticket for e906-gat3 failed Kerberos requests opened by Fermilab Kerberos administrator
15:40	ServiceDesk adds SeaQuest e906-gat3 system administrator to ticket
We, 2014-Nov-19 9:59	ServiceDesk assigns ticket to "SeaQuest Support". (The SCD liaison, the SeaQuest computing coordinator, and the SeaQuest spokesperson are on this support list.)
:	
Fr, 2014-Nov-21	Fermilab Kerberos administrator and SeaQuest e906-gat3 system administrator meet at KTeV to look at the setup of node e906-gat3 and to identify the source of the excessive false KDC requests. The NFS configuration of the node was wrongly believed to be the cause of the excessive false KDC requests.
:	
We, 2014-Nov-26 8:45	SeaQuest computing coordinator informs Fermilab Kerberos administrator that the experiment is still looking into/debugging the issue
We, 2014-Dec-03 pm	The Authentication Services group asks the Computer Security Team to block the node as the rate of failed requests have reached over 580k in the first 15 hours of the day and is causing a service degradation to the Fermilab KDC. CST notices also that the license of the log analyzer software is at risk to be overrun.

We, 2014-Dec-03 14:59	node e906-gat3 is manually blocked
We, 2014-Dec-03 15:03	Fermilab Kerberos administrator informs SeaQuest e906-gat3 system administrator and SeaQuest Support that errors have increased and node e906-gat3 is being blocked on the network
We, 2014-Dec-03 15:31	SeaQuest spokesperson informs Fermilab Kerberos administrator that experiment cannot take data anymore with machine being blocked
We, 2014-Dec-03 15:40	node e906-gat3 unblocked
We, 2014-Dec-03 18:36	a member of the SeaQuest DAQ team changes the ssh configuration on node e906-gat3 and later on all unkerberized DAQ machines on the private network connecting to e906-gat3
We, 2014-Dec-03 19:25	failed Kerberos requests from e906-gat3 stop

Analysis

The SeaQuest data acquisition system was designed to take data without Fermilab network. However, the DAQ setup evolved and at the time acquired target information from a PPD machine via the Fermilab network. A secure shell misconfiguration of unkerberized machines in the private network triggered the DAQ gateway machine to contact the Fermilab KDC excessively. The issue lingered for almost a month as the urgency to resolve the issue was not understood. On December 3rd the issue became so severe that it degraded the Fermilab Kerberos service and threatened to blind the Computer Security Team. The machine was thus blocked on the network since it was not known to be a critical system. The SeaQuest experiment was informed promptly about the block through the ServiceNow ticket. When data taking halted, SeaQuest collaborators quickly traced it to the network block and pointed out that the machine is critical for data taking. The network block was lifted promptly and data taking resumed. The DAQ team corrected the ssh misconfiguration a few hours afterwards.

About 45 minutes of beam time was lost.

Direct Cause

The direct cause of the data loss is the network block of node e906-gat3.fnal.gov by the Computer Security Team.

The direct cause of the incident is the secure shell misconfiguration on the DAQ machines in the private network.

Contributing Factors

SeaQuest people did not realize the severity of the excessive false KDC requests and thus handled the issue at lower priority.

Fermilab Kerberos and Computer Security people did not know that the node being blocked was a DAQ machine critical for data taking.

The incident ticket was created with Impact: Minor/Localized and Urgency: Medium. The ticket was not updated when the false KDC requests became more severe.

ServiceNow swamps the SCD liaison with tickets (for FermiGrid) and he does not look at most of the ticket emails. (The SeaQuest ServiceNow group was setup to track account requests.)

The system administrator of e906-gat3 was unfamiliar with Kerberos (and NFS) and was not aware of his responsibility in resolving the issue.

The SeaQuest DAQ system evolved over the years and picked up unnoticed a dependency on the Fermilab network inside the KTeV control room.

The SeaQuest DAQ team assumed that the person they communicated with on the ticket was an expert assigned to resolve the issue with them and did not understand that they were communicating with one of the Fermilab Kerberos administrators.

The SeaQuest DAQ team knows about the sysadmin mailing list and round table meeting. They were not aware that they can ask the Fermilab system administrators for help.

The false identification of NFS and miscommunication of who will contact more NFS experts stalled resolving the false KDC requests by over a week.

There was no communication on the ticket for a week before the jump in false KDC requests.

Root Cause

The criticality of node e906-gat3 was not fully realized by the SeaQuest collaboration and not known to Fermilab Kerberos administrators and the Computer Security team.

The severity of the false KDC requests was not properly communicated to SeaQuest.

Training/experience of the e906-gat3 system administrator was not at the level to allow him to diagnose the source of the false KDC requests and resolve them in a timely manner.

Observations & Comments

Based on the problem investigation and root cause analysis, the following observations and comments are made:

1. The beam time lost by the SeaQuest experiment was brief and does not disadvantage/hurt the experiment in the long term.

2. After a first attempt via email to resolve the issue failed a ticket in ServiceNow was opened. The ticket was quickly routed to the right support group. Kerberos administrators followed up regularly.

3. The ServiceNow ticket is effectively a reverse ticket, where the service provider requests action from a user group. User groups receive no ServiceNow training and are not familiar with all the functionality.

4. The rate of false KDC requests coming from e906-gat3 jumped on December 3rd requiring prompt action.

5. Fermilab Cyber Security Services has put process improvements in place to:

    - Engage SCD liaisons/SCD management in the event of an issue or delay in tickets, either assigned to or created by the department;

    - Make a best effort notification to SCD liaison/SCD management when performing manual network blocks;

    - Tickets opened regarding the Fermilab KDC or logging overruns will be opened with high priority;

6. The network node database of the Computing Sector does not contain fields to indicate critical systems or the purpose/task of a machine.

7. The responsibilities of liaisons are defined in CS-doc-4924. Effort for the task can vary greatly based on size/scope of the experiment, the state it is in, and how acquainted the liaison is with the experiment/collaboration. Priorities and amount of time should be agreed upon with line management so that tickets can be followed timely and be handled proactively in case an important system, service, and/or software is involved.

Recommendations

Based on the problem investigation and root cause analysis, the following recommendations are made for preventing future occurrences of this nature:

1. The SeaQuest DAQ team should consider outsourcing system administration of their control room/DAQ systems to the Fermilab system administration group. [done]
2. The Computer Security Team should review the log analyzer setup for single point of failure (like a rogue user or misconfigured system both inside/outside of Fermilab). [done]

3. SeaQuest plans to decouple the data acquisition system from the Fermilab network. This should be made high priority and tested regularly to verify no unknown dependencies have been picked up. [done]
4. Node e906-gat3 should be exempt from auto-blocking until the SeaQuest DAQ is decoupled from the Fermilab network and this tested. SeaQuest should review if auto-blocker exemptions are needed/appropriate for other DAQ machines and request those. [done]
5. Fermilab Scientific Computing Division should inform new experiments about its system administration resources. [done]
6. SCD should consider critical systems documents also in the case of smaller experiments. [done]

Root Cause Analysis Committee

The following people participated in the problem investigation and root cause analysis:

Andrew Chen

Markus Diefenthaler

Gabriele Garzoglio

Darnell Green

Joe Klemencic

Paul Reimer

Patrick Riehecky

Olga Terlyga

Brian Tice

Stephan Lammel