Root Cause Analysis Report

Amazon S3 Service Disruption

sologic root cause analysis

Problem Statement

Report Number	RCA-07-03-2017-083	**RCA Owner**	Brian Hughes
Report Date	3/6/2017	**RCA Facilitator**	Brian Hughes

Focal Point: Service Disruption Northern VA, US EAST-1 Region > 4 hours

When

Start Date: 2/28/2017	End Date: 2/28/2017
Start Time: 9:37AM	End Time: 1:54PM
Unique Timing	While debugging an issue causing the S3 billing system to run at sub-optimal speeds.

Where

System	Amazon Simple Storage Service (S3)
	Amazon.com
Component	Placement Subsystem
Component	Index Subsystem
Site	Northern Virginia (US-EAST-1) Region
Business Unit	Amazon Web Services (AWS)

Actual Impact

Customer Service	Multiple system outages for several hours	$0.00
Customer Service	Unreported impact to customer businesses	$1,000,000.00
		Actual Impact Total: $1,000,000.00
Frequency	2 times Year	
Frequency Note	Unreported frequency	

Potential Impact

Customer Service	All impacts potentially could have been worse.	$0.00
		Potential Impact Total: $0.00

Report Summaries

Executive Summary

READ THIS FIRST:

We need to disclose that this EXAMPLE RCA is based upon publicly available information published in a single report by Amazon and not from any independent investigation conducted by Sologic. Sologic has not investigated this incident in any official capacity, and we do not want to imply that we were in any way associated with this event. The only purpose of this root cause analysis report is for it to be used as an example for our students and other interested parties.

A root cause analysis has two primary goals: 1) Organize a wide array of information from disparate sources in a way that makes it easier to understand, and 2) Identify a set of evidence-based solutions to present to decision makers. IT outage reports are often vague and peppered with tech-heavy terms. This style makes them a bit opaque to those outside the industry. But it does not have to be that way – a cause and effect chart provides a nice visual reference to go along with the report. The chart puts the causal interactions into context with respect to time, allowing the reader to see how the event unfolded.

A few thoughts about IT and root cause analysis in general, not necessarily associated with this particular event. It's been our experience that IT professionals are often extremely intelligent. But at the macro level, IT is relatively new to the world of structured problem solving. Many ITSM efforts focus first on Incident Management, with the intent of standing up Problem Management at some later point. When a large problem like the one detailed in this example occurs, IT professionals are under extreme pressure to complete the investigation as quickly as possible. Often, new problems have cropped up that need their attention. And their customers are demanding answers. Couple this environment with the fact that these systems are complex and the investigation team is often inexperienced with root cause analysis, and you get the right conditions for a sub-optimal investigation.

The problem with this is the continued exposure to risk, even when steps are taken to formally solve the problem. An investment in a formal root cause investigation is supposed to finance a reduction in risk. The risk of problem recurrence is directly related to the quality of solutions implemented by the team, and solution-quality depends on a logical, thorough, and evidence-based root cause analysis. When the consequences of failure are high, an investment in RCA capability pays off in a big way. This investment includes training, software, and consulting (all the things Sologic provides). But equally important is the investment leadership makes in change management. Building capability requires the structure of an RCA program, and this requires recognition by leadership that their success is incumbent upon the collective problem-solving capability of the organization. This is particularly true in IT.

If possible, consider printing the following summary report and follow along with the cause and effect chart as you read the report. Notice the solutions Amazon has put in place, along with which causes they control. What do you think?

Cause and Effect Summary

Orignal Amazon Report

Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region:

We'd like to give you some additional information about the service disruption that occurred in the Northern Virginia (US-EAST-1) Region on the morning of February 28th. The Amazon Simple Storage Service (S3) team was debugging an issue causing the S3 billing system to progress more slowly than expected. At 9:37AM PST, an authorized S3 team member using an established playbook executed a command which was intended to remove a small number of servers for one of the S3 subsystems that is used by the S3 billing process. Unfortunately, one of the inputs to the command was entered incorrectly and a larger set of servers was removed than intended. The servers that were inadvertently removed supported two other S3 subsystems. One of these subsystems, the index subsystem, manages the metadata and location information of all S3 objects in the region. This subsystem is necessary to serve all GET, LIST, PUT, and DELETE requests. The second subsystem, the placement subsystem, manages allocation of new storage and requires the index subsystem to be functioning properly to correctly operate. The placement subsystem is used during PUT requests to allocate storage for new objects. Removing a significant portion of the capacity caused each of these systems to require a full restart. While these subsystems were being restarted, S3 was unable to service requests. Other AWS services in the US-EAST-1 Region that rely on S3 for storage, including the S3 console, Amazon Elastic Compute Cloud (EC2) new instance launches, Amazon Elastic Block Store (EBS) volumes (when data was needed from a S3 snapshot), and AWS Lambda were also impacted while the S3 APIs were unavailable.

S3 subsystems are designed to support the removal or failure of significant capacity with little or no customer impact. We build our systems with the assumption that things will occasionally fail, and we rely on the ability to remove and replace capacity as one of our core operational processes. While this is an operation that we have relied on to maintain our systems since the launch of S3, we have not completely restarted the index subsystem or the placement subsystem in our larger regions for many years. S3 has experienced massive growth over the last several years and the process of restarting these services and running the necessary safety checks to validate the integrity of the metadata took longer than expected. The index subsystem was the first of the two affected subsystems that needed to be restarted. By 12:26PM PST, the index subsystem had activated enough capacity to begin servicing S3 GET, LIST, and DELETE requests. By 1:18PM PST, the index subsystem was fully recovered and GET, LIST, and DELETE APIs were functioning normally. The S3 PUT API also required the placement subsystem. The placement subsystem began recovery when the index subsystem was functional and finished recovery at 1:54PM PST. At this point, S3 was operating normally. Other AWS services that were impacted by this event began recovering. Some of these services had accumulated a backlog of work during the S3 disruption and required additional time to fully recover.

We are making several changes as a result of this operational event. While removal of capacity is a key operational practice, in this instance, the tool used allowed too much capacity to be removed too quickly. We have modified this tool to remove capacity more slowly and added safeguards to prevent capacity from being removed when it will take any subsystem below its minimum required capacity level. This will prevent an incorrect input from triggering a similar event in the future. We are also auditing our other operational tools to ensure we have similar safety checks. We will also make changes to improve the recovery time of key S3 subsystems. We employ multiple techniques to allow our services to recover from any failure quickly. One of the most important involves breaking services into small partitions which we call cells. By factoring services into cells, engineering teams can assess and thoroughly test recovery processes of even the largest service or subsystem. As S3 has scaled, the team has done considerable work to refactor parts of the service into smaller cells to reduce blast radius and improve recovery. During this event, the recovery time of the index subsystem still took longer than we expected. The S3 team had planned further partitioning of the index subsystem later this year. We are reprioritizing that work to begin immediately.

From the beginning of this event until 11:37AM PST, we were unable to update the individual services' status on the AWS Service Health Dashboard (SHD) because of a dependency the SHD administration console has on Amazon S3. Instead, we used the AWS Twitter feed (@AWSCloud) and SHD banner text to communicate status until we were able to update the individual services' status on the SHD. We understand that the SHD provides important visibility to our customers during operational events and we have changed the SHD administration console to run across multiple AWS regions.

Finally, we want to apologize for the impact this event caused for our customers. While we are proud of our long track record of availability with Amazon S3, we know how critical this service is to our customers, their applications and end users, and their businesses. We will do everything we can to learn from this event and use it to improve our availability even further.

Solutions

SO-0001	**Solution**	From Amazon: Make changes to improve recovery time.		
	Cause(s)			
	Note	Factor services into "cells" (small partitions). This will allow engineering teams to assess and throroughly test recovery processes of even the largest service or subsystem.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	3/6/2017	**Status**	Approved
	Term	medium	**Cost**	

SO-0002	**Solution**	From Amazon: Audit other operational tools to ensure similar safety checks are in place.		
	Cause(s)			
	Note	This is a preventive action and will help identify other at-risk areas.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	3/10/2017	**Status**	Completed
	Term	short	**Cost**	

SO-0003	**Solution**	From Amazon: Change AWS Service Health Dashboard (SHD) administration console to run across multiple AWS regions.		
	Cause(s)			
	Note	This will provide better visibility during future outages.		
	Assigned	Jon Boisoneau	**Criteria**	Not Checked
	Due	3/10/2017	**Status**	Approved
	Term		**Cost**	

SO-0004	**Solution**	From Amazon: Modify the "trouble shooting tool."		
	Cause(s)	Restart took longer than expected		
	Note	Modify trouble shooting tool to remove capacity more slowly and to prevent capacity from being removed when it will take any substyem below its minimum required capacity level.		
	Assigned	Jon Boisoneau	**Criteria**	Passed
	Due	3/6/2017	**Status**	Completed
	Term	short	**Cost**	

Team

Facilitator	Owner
Brian Hughes	Brian Hughes
brian.hughes@sologic.com	brian.hughes@sologic.com

Participants

Cory Boisoneau	Chris Eckert
cory.boisoneau@sologic.com	chris.eckert@sologic.com

Notes

NO-0001	**Note**	"NGR" is short for "No Good Reason Not To Act." It is presumed, but not confirmed that the employee did not know that his actions would cause such a huge outage.
	Cause(s)	NGR: Team member unaware of consequences?

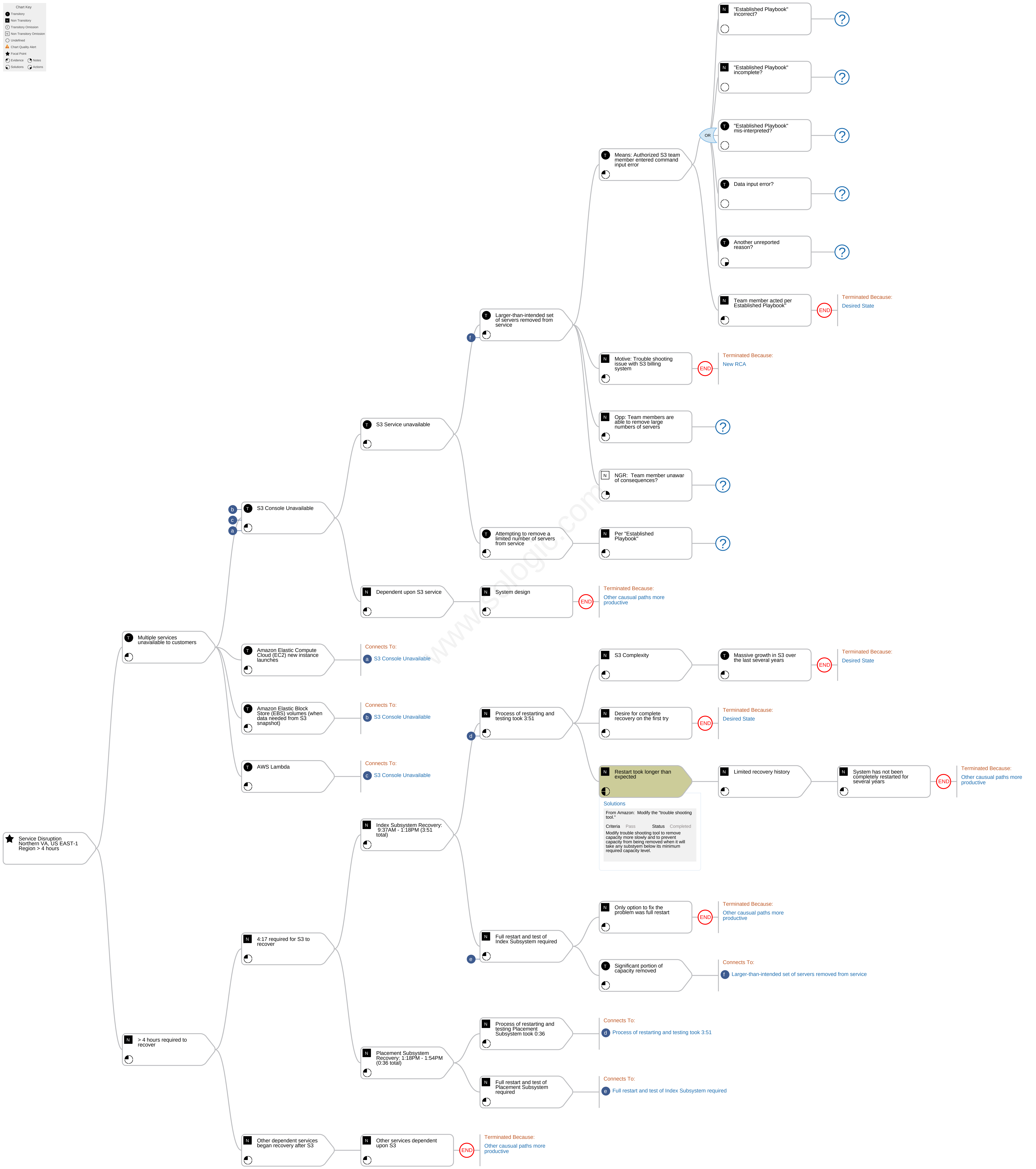
Chart Key
T Transitory
N Non Transitory
T Transitory Omission
N Non Transitory Omission
Undefined
Chart Quality Alert
Focal Point
Evidence
Notes
Solutions
Actions
Service Disruption Northern VA, US EAST-1 Region > 4 hours
T Multiple services unavailable to customers
N > 4 hours required to recover
T S3 Console Unavailable
b
c
a
T Amazon Elastic Compute Cloud (EC2) new instance launches
Connects To:
a S3 Console Unavailable
T Amazon Elastic Block Store (EBS) volumes (when data needed from S3 snapshot)
Connects To:
b S3 Console Unavailable
T AWS Lambda
Connects To:
c S3 Console Unavailable
T S3 Service unavailable
N Dependent upon S3 service
N System design
END
Terminated Because:
Other causual paths more productive
f
T Larger-than-intended set of servers removed from service
T Attempting to remove a limited number of servers from service
N Per "Established Playbook"
?
T Means: Authorized S3 team member entered command input error
OR
N "Established Playbook" incorrect?
?
N "Established Playbook" incomplete?
?
T "Established Playbook" mis-interpreted?
?
T Data input error?
?
T Another unreported reason?
?
N Team member acted per Established Playbook"
END
Terminated Because:
Desired State
N Motive: Trouble shooting issue with S3 billing system
END
Terminated Because:
New RCA
N Opp: Team members are able to remove large numbers of servers
?
N NGR: Team member unawar of consequences?
?
N 4:17 required for S3 to recover
N Index Subsystem Recovery: 9:37AM - 1:18PM (3:51 total)
d
N Process of restarting and testing took 3:51
N S3 Complexity
T Massive growth in S3 over the last several years
END
Terminated Because:
Desired State
N Desire for complete recovery on the first try
END
Terminated Because:
Desired State
N Restart took longer than expected
N Limited recovery history
N System has not been completely restarted for several years
END
Terminated Because:
Other causual paths more productive
e
N Full restart and test of Index Subsystem required
N Only option to fix the problem was full restart
END
Terminated Because:
Other causual paths more productive
T Significant portion of capacity removed
Connects To:
f Larger-than-intended set of servers removed from service
N Placement Subsystem Recovery: 1:18PM - 1:54PM (0:36 total)
N Process of restarting and testing Placement Subsystem took 0:36
Connects To:
d Process of restarting and testing took 3:51
N Full restart and test of Placement Subsystem required
Connects To:
e Full restart and test of Index Subsystem required
N Other dependent services began recovery after S3
N Other services dependent upon S3
END
Terminated Because:
Other causual paths more productive

Solutions

From Amazon: Modify the "trouble shooting tool."

Criteria Pass Status Completed

Modify trouble shooting tool to remove capacity more slowly and to prevent capacity from being removed when it will take any substyem below its minimum required capacity level.