

Incident Report: RabbitMQ Node Failures and Network Connectivity Issues

1. Incident Summary

Duration: May 12, 2024, to May 13, 2024

Exact: May 12, 2024, at 06:41:17 IST to May 13, 2024, at 10:35:06 IST

Approximately 1 day 5 hours, and 16 minutes.

Issue: Network connectivity between cluster nodes result in RabbitMQ node failure. XYZ-prod-celery-rabbitmq-node2 connection to the cluster failed.

Affected Systems: RabbitMQ cluster nodes ('rabbit@XYZ-prod-celery-rabbitmq-node1' and 'rabbit@XYZ-prod-celery-rabbitmq-node2') connections.

2. Timeline of Event

- Several warnings were logged indicating unexpected node down or inaccessible in backend logs with CRITICAL severity. Time- on 13 May-2024 09:48 AM IST
- Node 'rabbit@XYZ-prod-celery-rabbitmq-node1' was reported as not responding, resulting in the removal of its listeners.
- Additional warnings and errors occurred related to AMQP channel exceptions and database inconsistencies.
- The issue persisted, with repeated warnings about client TCP connection closures and AMQP channel errors.
- Restarted the rabbitmq-service in XYZ-prod-celery-rabbitmq-node2 at May 13, 2024, at 10:35:06 IST.
- System back to normal.

Alerts: The first alert received in slack channel [#XYZ_prod_debug](#) on 13 May-2024 09:48 AM IST with the below error log -

```
reply_code, reply_text, (class_id, method_id), ChannelError,  
amqp.exceptions.NotFound: Queue.declare: (404) NOT_FOUND - home node 'rabbit@██████-prod-celery-  
rabbitmq-node1' of durable queue 'celery' in vhost '/' is down or inaccessible
```

3. Impact Analysis

Service Disruptions:

- Queue service affected. Backend servers experienced connectivity issues with Rabbitmq cluster and were unable to establish stable connections to RabbitMQ node2. During time May 12, 2024, at 06:41:17 IST to May 13, 2024, at 10:35:06 IST.
- After research concluded that there were no existing tasks in the queue.
- Task creation attempt made during partial downtime. But they failed after multiple retries. (Ref: 4)
 - List of the affected tasks
 - Update Profile n Times
 - Add Candidate n times

Features affected:

- Add Candidate & Update Profile from [admin](#).

Supportive Logs:

Time	request	user_agent	remote_addr	status	referrer	host.name	referrer_type.device
> May 13, 2024 @ 10:15:59.000	/api/v2/admin/candidate/27975718-3c94-4014-a866-d51b7ca1ea5/gpdate.profile/	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36	172.31.29.198	500	https://	prod-backend-asp-e c2-172-31-33-239	Other
> May 13, 2024 @ 10:06:30.000	/api/v2/admin/candidate/add_candidate/	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36	172.31.29.198	500	https://	prod-backend-asp-e c2-172-31-33-239	Other
> May 13, 2024 @ 10:06:27.000	/api/v2/admin/candidate/add_candidate/	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36	172.31.29.198	500	https://	prod-backend-asp-e c2-172-31-33-239	Other
> May 13, 2024 @ 09:55:52.000	/api/v2/admin/candidate/add_candidate/	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36	172.31.29.198	500	https://	prod-backend-asp-e c2-172-31-33-239	Mac
> May 13, 2024 @ 09:48:41.000	/api/v2/admin/candidate/add_candidate/	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/124.0.0.0 Safari/537.36	172.31.11.99	500	https://	prod-backend-asp-e c2-172-31-33-239	Other

4. Root Cause Analysis

Network Issues: Investigations revealed network connectivity problems between RabbitMQ nodes and client connections, leading to TCP connection failures.(Ref: 1, 2)

Supportive Logs:

```

root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log# cat crash.log
2024-05-13 05:01:15 =SUPERVISOR REPORT=====
Supervisor: {<0.4066.402>,rabbit_channel_sup_sup}
Context: shutdown_error
Reason: shutdown
Offender: [{nb_children,1},{name,channel_sup},{mfargs,{rabbit_channel_sup,start_link,[]}}, {restart_type,temporary},{shutdown,infinity},{child_type,supervisor}]

2024-05-13 05:01:24 =ERROR REPORT=====
Mnesia('rabbit@prod-celery-rabbitmq-node2'): ** ERROR ** mnesia_event got {inconsistent_database, starting_partitioned_network, 'rabbit@prod-celery-rabbitmq-node1'}
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log# cat crash.log
crash.log crash.log.0 crash.log.1 crash.log.2 crash.log.3 crash.log.4
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log# cat crash.log.0
2024-05-12 01:11:17 =ERROR REPORT=====
** Node 'rabbit@prod-celery-rabbitmq-node1' not responding **
** Removing (timedout) connection **
2024-05-12 01:11:31 =ERROR REPORT=====
Mnesia('rabbit@prod-celery-rabbitmq-node2'): ** ERROR ** mnesia_event got {inconsistent_database, running_partitioned_network, 'rabbit@prod-celery-rabbitmq-node1'}
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log# cat crash.log.1
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#
root@prod-celery-rabbitmq-node2:/var/log/rabbitmq/log#

```

5. Remediation

Node Recovery:

- Restarted the rabbitmq-service in XYZ-prod-celery-rabbitmq-node2.
- Re-enabled high availability settings in Rabbitmq cluster nodes.(Ref: 3)

Client Communication: Informed via email itdev@xyz.com

6. Follow-ups Actions

Monitoring:

- Systems were kept under observation for 3 days after the incident recovery to ensure completeness of the recovery.

Additional Tasks Identified:

- Enhancing monitoring systems to proactively monitored and detect Crash logs under centralized logging from all nodes.
- Address node failures by Crash alerts.

7. Conclusion

The incident involving RabbitMQ node failures and network connectivity issues resulted in service disruptions and connectivity problems. Prompt response, thorough investigation, and resolution measures helped mitigate the impact and restore service stability. Enhancing monitoring systems to proactively detect and address node failures will be essential to minimize the risk of similar incidents in the future.

8. References

1. <https://www.rabbitmq.com/docs/partitions>
2. <https://stackoverflow.com/questions/8654053/rabbitmq-cluster-is-not-reconnecting-after-network-failure>
3. <https://www.rabbitmq.com/docs/ha>
4. <https://docs.celeryq.dev/en/stable/userguide/tasks.html>

Reported by: Backend and DevOps Teams

Date of Report: 16 May 2024