An automated Site Reliability Engineering (SRE) - driven performance engineering

Prashant Mane

Abstract— Much of our day to day technology is powered by SRE - An automated Site Reliability Engineering (SRE) - driven performance engineering is a solution implemented in one of the largest Tier-1 Telecomm Company in Canada, North America, the first of the big three Telcos in Canada. The solution was targeted to overcome the performance, site reliability & availability issues in its Production environment. In this paper, we will specifically discuss "Methodology and work undertake". The knowledge base constantly helps in storing and building pattern, which in turn helps self-learning and responding to actions.

Index Terms— Site Reliability Engineering, Performance testing, Software Quality, Predictive Analytics, Availability Testing, Change management, Incident management, Monitoring and alerting.

———————————— ◆ ————————————

1 INTRODUCTION

A Couple of years back the company had launched a new Digital platform to cater to its wide range of business segments. During Covid lockdown & various events in various times of the year, significant spikes were observed in online sales. This additional traffic in the Sales channel specially for new plans, devices and offers impacted the environment performance. It was difficult to meet the desired performance SLAs during such aberrations. The unprecedented spikes on online sites and mobile apps posed a challenge to Telcos in its quest towards quality supply services and last mile operations.

However, due to the most significant increase was seen in online purchase of new plans during COVID-19 lockdown & Special events, the platform had performance issues and was not able to meet the performance SLA's in production. This indicator was the overall increase in online purchase of new deals, plans, TV packages deals and mobile devices when compared with the previous year. The unprecedented spikes on online sites and mobile apps had caused delays in the deliveries while Telecomm Company adapt their supply services and last mile operations.

The SRE - driven performance engineering team was created and called upon with the objective of ensuring a new version of the environment that was going to be deployed in production, free of any performance issues and a consistent availability and reliability of 99%.

• *Prashant Mane is currently Technical Delivery Manager in Tech Mahindra Ltd, India, PH-+919766597848. E-mail: pm00116667@techmahindra.com*

2 BACKGROUND AND SIGNIFICANCE

This platform solution was developed with a view to assure the end user experience across different business sectors comprising of 10 million consumers, which was the anticipated business forecast.

Canada is a highly mobile-connected market. 70 percent of people have access to at least one smartphone and more than half of these devices were purchased within the last few years. Telecomm Company is a diversified Canadian communications and media company. The new Digital platform in Telcos was built with state-of-the-art technology that provided to its customer - round-the-clock availability, easy accessibility, a wide range of services and international reach. It was a smart solution that could Build, Run and Scale its online business with customer friendly features related to Account management in Mobile Apps, Sale via websites, Payment Gateways, Efficient Shipping and a world class Call Center agent interface.

That is when Digital platform were incentivising customers to purchase online, and alleviating congestion from the stores; and address all the performance issues. In order to ensure a reliable solution was delivered, it was pertinent that a robust QA solution was established. One that could validate the performance of the solution comprehensively and could adapt to translate the rising risks and online demands into long lasting customers, and to manage online applications for peaks in Special events traffic by using SRE- driven performance engineering to enhance site reliability in production environment.

3 PROJECT AIMS/GOALS

The main objective of this project was to optimize the current infrastructure, and to provide confidence during COVID-19 lockdown that the architecture is scalable and is able to

handle the optimal performance of current and near future 15 million volumes without impacting the performance.

The key objective of overhauling the entire platform was to understand the factors impacting the end user experience and address the same in the upcoming version:

• Address the overall performance, reliability & availability issues of the platform and make it more robust.

• Improve the support from an existing 12 million consumers to 15 million consumers.

• The platform was expected to be connected to 20+ external components at any given point of time with 99% availability

4 METHODOLOGY/WORK UNDERTAKEN

4.1 AN AUTOMATED SRE-DRIVEN PERFORMANCE ENGINEERING:

Objective of setting up SRE - driven performance engineering team was to create a bridge between development and operations by applying a software engineering mind-set to system administration topics. SRE - driven performance engineering team was a prescriptive way to do DevOps, which focused on tenets of operating a service through enabling developers and operations by lowering delivery and incident management lifecycles time and reducing toil in production environment.

TechMs automated SRE-driven performance engineering strategy achieved the following goals:
• Maintain uptime
• Establish visibility into the application performance by using SRE Engineering
• Set SLI's and SLO's and error budgets
• Increase speed by assuming calculated risks
• Eliminate toil
• Reduce the cost of failure to lower new feature cycle time.

4.2 TOOLS AND TECHNIQUES:

We utilised a mix of in-house developed frameworks like "An automated Site Reliability Engineering (SRE) - driven performance engineering platform", non-functional test simulator along with a suite of cost effective on demand commercial tools from our tool partners.

4.3 EXECUTION STRATEGY:

The hourly performance test execution would not affect end user customers in production environment:
a. Hourly performance test: performed after every one hour with a very minimum vusers in production environment for performance, reliability and availability
b. Live stats of Key Digital APIs (updated every 15 minutes)
c. Health status of multiple database servers (updated every 15 minutes)
d. Health status of multiple Middleware and Application servers (updated every 15 minutes)

CRITICAL BUSINESS FLOWS: - As depicted below -

The daily tasks of an SRE driven performance engineering team was consist of:
a. Availability
b. Latency
c. Performance
d. Efficiency
e. Change management
f. Monitoring and alerting
g. Emergency response
h. Incident response

4.4 APPLICATION PERFORMANCE ANALYSIS, MONITORING AND ALERTING FOR SRE:

The SRE - driven performance engineering team had the responsibility for maintaining and establishing service level indicators (SLIs), objectives (SLOs), agreements (SLAs), and error budgets for systems and ensured these are met.

A. Application Architecture:

The application architecture can be viewed by clicking on the respective application name.

The color of each item in the above topology represents the health status of respective application (which gets refreshed every 15 minutes). The detailed health status of the respective application can be viewed by clicking on the same. After clicking on ESB DG in topology we will see the Image screen:
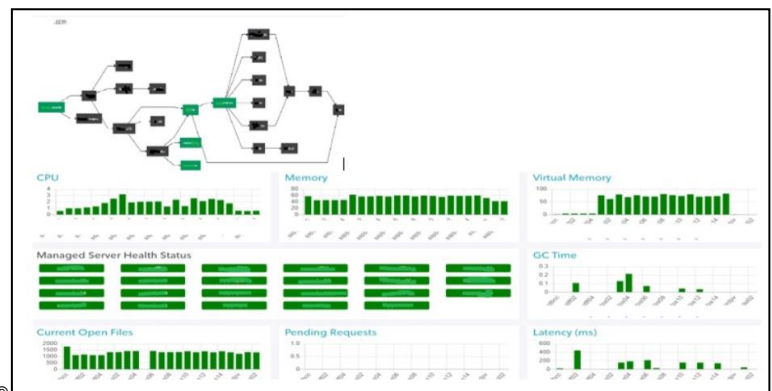
B. API to Middleware mapping/ MW Service Sequence:

Each of the critical flow on dashboard page is a clickable item which shows the list of APIs invoked by it. The APIs are mapped to the corresponding Middleware calls. After clicking on XX.Com – HardwareUpgrade flow on dashboard page, below view gets displayed.

Each of the MW Service calls in the above view is a clickable item which shows the sequence diagram of the same. Sequence diagram of UTE/ActServicesPS: readxxtSummary call.

This item can be pulled by clicking on options button on homepage. Live stats of Key Digital APIs (updated every 15mins) were procured by using Dynatrace queries. This enables us to view the main attributes of those APIs like Request count, Avg latency, Server Side Failure Rate and HTTP 4xx/5xx errors.

Each of the API in above page is a clickable item which shows the trending of the tabulated attributes: (The trending time-range can be altered from top-right section of the page)

C. Availability Status/Average Latency:

The values under Availability Status/Average Latency columns are clickable items which shows the trending (graphical view) of the corresponding values. The time range for trending can be altered in the top-right section of the page. After clicking on the response time for UTE/AccountServices: readAccountSummary (.594 under Average Latency column) in the Image 3:

As depicted in Image 3
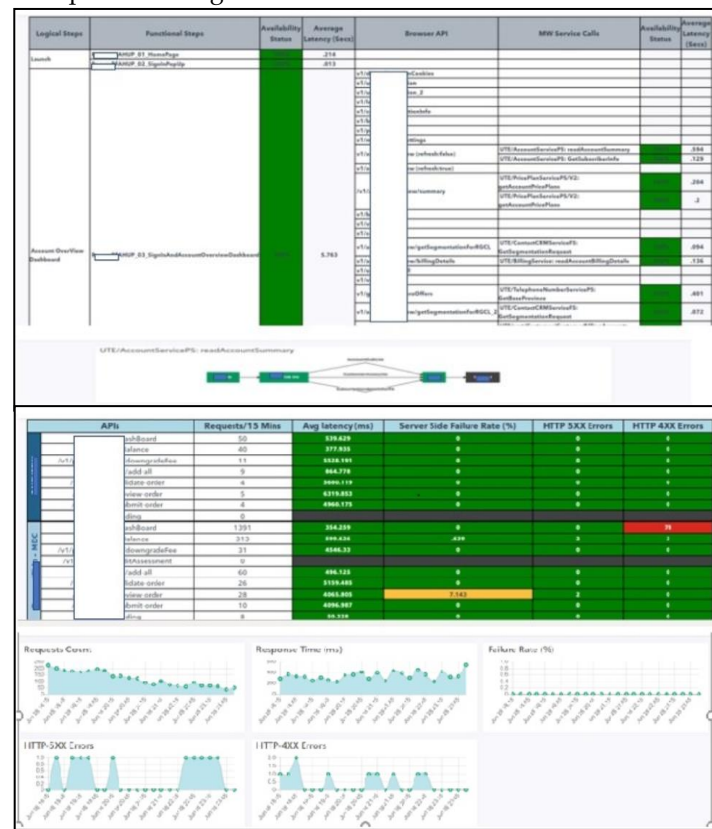
Health status of multiple database servers, multiple Middleware and Application servers (updated every 15 minutes)

The health status for most of the Digital application servers were being captured from different tools like Sitescope, Dynatrace, Diagnostics and were being pubished on PAT Sanity Dashboard (which gets refreshed every 15mins). Health status for few of the servers which doesn't have a monitoring tool configured is captured by using CODA queries.

This item can be pulled by clicking on options button on homepage Portal dashboard. The data on this page is being pulled from a feed generated by DBA team (for every 15mins).

Each database name in Database column is a clickable item which leads us to trending of above tabulated attributes: Each metric in which shows the trending for a customizable time range: (After clicking on CPU link in image page, we can view containing the CPU usage for all the servers)

As per below Image

D. Alert:

Automated alerts have been configured to report the issues identified in hourly sanity. In case of any issues (YELLOW/RED status) observed for a flow, at 5th minute of that hour an auto-

mated alert will get triggered to corresponding teams specifying the flow name and pass % of the flow.

At 8th minute of the hour, the errors found in the above flows will get mailed to same DL depicted in Image 5:

7 CHALLENGES AND OBSTACLES OVERCOME

- The reliability of end user internet connection played a crucial role in determining the performance of the digital platform, so it was essential to verify that data will still be detected and properly stored in the event of a disruption in service. This was overcome by simulating the network packet drops and error handling techniques within the test scripts. This allowed the teams to re-send any failed request once normal service had resumed.
- Business were reporting increases of nearly 100% in their segments, and of up to 5 times more volume in the physical stores. We ensured to use advanced (all-in-one) solution to automate route optimisation, had real-time visibility over the tasks and potential issues as well as to control customer experience.
- The unusual volumes were disrupted business operations when delivery volumes were not managed by a flexible approach and capable of integrating variables. It also collected all data real-time, increasing the user view with daily reports on all activities.

8 CHALLENGES AND OBSTACLES OVERCOME

- Response time of Communication Engine for online applications, Mobile apps and Shipping Providers - Search Engine Devices came down from few seconds to 1 second allowing it to handle a load of more than 20 million.
- Micro services and APIs were tuned which was more than the current capacity
- Any degradation of performance and consumer experience would have result in a loss of money, revenue, and opportunities and cause a big issue with traditional operations not able to cope with the observability, including monitoring and alerting in real-time.
- Due to the daily involvement of an SRE driven performance engineering team, application's efficiency, availability, reliability, and performance were improved and reduced operational costs
- Automated SLA validation by SRE team optimized pro-

cesses, workflow, automation, and removed unnecessary manual effort and rework.

9 CONCLUSION

- We gained a good amount of experience from this project which helped us to further extend this solution to other customers who was faced similar situation. The benefits realised by this customer are as below:
- Provided value add to the customer by setting up a robust and comprehensive "An automated Site Reliability Engineering (SRE) - driven performance engineering platform" which is being used as an enterprise solution for multiple projects.
- E2E response time for processing the order was brought down
- New coronavirus pandemic lockdown was imposing on Company a novel situation that consumers demanded a new level of expectations, not only faster, but more flexible, more transparent and greener deliveries.
- The digital transformation requires new and more expeditious ways for delivering software solutions and SRE - driven performance engineering is a model adaptable to company's requirements. Therefore, because the SRE mind-set and processes can be applied to IT operations at any scale, companies should become SRE - driven performance engineering within core I&O teams.

REFERENCES

[1] Site Reliability Engineering (SRE) by google (https://cloud.google.com/sre)
[2] Guide to automated SRE by Dynatrace (https://www.dynatrace.com/news/blog/guide-to-automated-sre-driven-performance-engineering-analysis/)