SITE RELIABILITY ENGINEERING

in Services Projects

Abstract

This document intends to provide the concepts and guide to implement Site Reliability Engineering in services projects

Varun Kumar

varkum@microsoft.com

Table of Content

SRE in Services Projects

How SRE is different for Services projects?

Firstly, Services projects means the custom application/system/solution developed for the customer based on their requirements, to meet business goals.
These systems are developed over certain platform or services provided by some product companies and are deployed on the Infrastructure provided by cloud providers or on-premises (own or rented infrastructure).

Most of the companies do implement SRE on their products like google, facebook, Adobe and even Microsoft. However, it doesn't mean that it can't be implemented in services projects because ultimately the reliability challenges in software projects are similar. It can be implemented in a product or a Customer's service that has operational challenges.

SRE for Services project is different in two ways:

1. The customer should be educated and convinced to adopt SRE principles because, there is no one other than the customer can define the critical flows and components of the system i.e. the things important for customer from business point of view. Also, it is important to work together and create a blameless environment with common goal to increase reliability.
2. Right reliability target should be set with the customer. Since, no cloud provider can provide 100% reliability of the services they offer. Therefore, it’s not possible for system using those services guarantee 100% reliability. The reliability can be increased to near-100% based on the design and implementation of the system and its various components. Example, the components can be designed to scale up or out based on load to cater all requests, geo replication of the storage and databases for high availability and fault tolerance etc.

Project Maturity Levels

The services projects are generally at one of the following maturity levels:

- A greenfield development, with nothing currently deployed
- A system in production with some monitoring to notify failures, but no formal objectives, no concept of an error budget, and an unspoken goal of 100% uptime.
- A running system with a defined SLO, but without an understanding of its importance or how to leverage it to make continuous improvement decisions.

Customer Reliability Engineering (CRE)

CRE is helping customers to implement SRE practices so that they use services in the same way as cloud provider and align their targets with them. It breaks organization barriers between cloud provider and customer and makes sure Failure is accepted as normal.

1. **Share Cloud provider’s (Azure’s) SLO and SLI with customers.**
   This way helps to know if the break happened is at customer end or at cloud. Also helps customer define the accepted level of reliability and error budget so that they don’t panic when down.

2. **Using the power of cloud to achieve higher reliability**
   Provide measures to achieve higher reliability with the help of the scalability rules, geo-replication, sharding, distribution, vertical partitioning etc.

3. **Understand what's critical for customer's business.**
   Understanding Customer's business, critical journeys and components, and pain areas and the reliability targets.

Implementing CRE in projects

Onboard

1. Identify and create the SRE team for the project and define roles and responsibilities of each individual in the team.
2. Production Readiness Review with development team.
3. Explain SRE to Customer and convincing Customer to use SRE.
4. Understand Customer's business and Identify critical features and journeys from Customer's business point of view.
5. Define SLO with customer budget and need.
   a. Explain why 100% is not possible.
   b. Explain cost for higher SLO. (10x cost with each 9).
6. Define Error budget policy with action and consequences of breaching various level of thresholds.
7. Create Central Postmortem Repository.
8. Chose and Implement monitoring system that has features as defined in SRE Guide.

SRE Governance

9. Create SRE Governance and review board.
10. Define Postmortem review and service improvement process.

Measure and Analyze

11. Implement target specific monitoring and alerting to track and measure SLIs.
12. Analyze metrices data and Identify gaps.

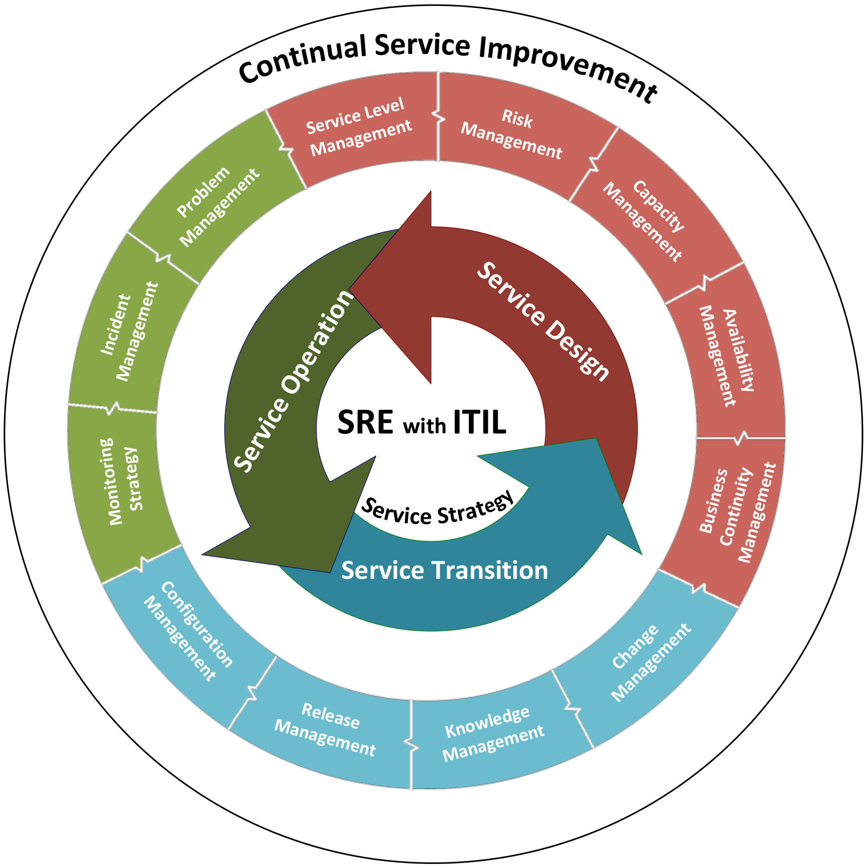
Incident Management

13. Respond to Incidents.
14. Follow the runbooks defined for the category of the Incident, If available.
15. Identify the duration, scope, and frequency of Incident.
16. Fix the Issue.

Improve

17. Postmortem of Incident and define action plan to prevent Incident from re-occurring.
18. Prioritize reliability work with customer.
19. Reduce the Toil.
20. Define Runbooks for persons responsible to respond to incident.
21. Review and re-define SLOs based on the data and metrices.

SRE with IT Service Management (ITSM)



Service Strategy

SRE requires to develop a service strategy with the customer to define service levels based on the reliability requirements of the service from customer's business point of view and considering 100% is a wrong reliability target. This includes understanding of customer's business, identifying critical components or flows that are most important for the customer, the demand or utilization of the critical components and required capacity and budget to meet demands to maintain the positive relationship between the Customer and Customer's customer.

Service Design

- **Service Level Management**
  SRE uses Service Level Indicators (SLIs) and Service Level objectives (SLOs) for Service Level Management and to meet SLAs.

- **Risk Management**
  SRE, manages service reliability largely by managing risk. SRE gives equal importance to figuring out how to engineer greater reliability and identifying the appropriate level of tolerance for the services. The goal is to explicitly align the risk taken by a given service with the risk the business is willing to bear. This is achieved through measuring the risk and define risk tolerance of a service.

- **Capacity Management**
  Capacity Management in SRE ensures that the service infrastructure and design is able to deliver the agreed SLOs and meet reliability targets.

- **Availability Management**
  SRE provides SLI and SLO to define, analyze, plan, measure and improve all aspects of the availability for the services that are critical for business. Availability Management in SRE is responsible for ensuring that all IT infrastructure, processes, tools, roles etc. are appropriate for the agreed availability targets.

- **Business Continuity Management**
  SRE defines the error budget policy to ensure proper actions are taken on various level of thresholds to maintain business continuity.

Service Transition

- **Change Management**
  SRE take a different approach to change management. In SRE, error budget is defined that represents the gap between the present reliability and agreed service level objectives (SLOs). While the team is allowed to regulate its own workload, there is an Error Budget policy defined, that consists of the actions and consequences when the error budget is blown or service levels breached. Since error budgets are meant to be spent, the team can make autonomous decisions to increase flow.

- **Shared Knowledge (Knowledge Management)**
  Postmortems that are recorded and stored in a Central repository, are accessible by all the SREs helps to share the knowledge and experience of all other Site Reliability Engineers. This knowledge is based on various Incidents in different projects. Lessons learnt and action Items explained in the postmortem can be re-used in the similar category and severity of Incidents, thus, improving efficiency by reducing the need to rediscover knowledge.

- **Release Management**
  Release management in SRE consists of the best practices and tools that covers all elements of the release process to make sure projects are released using consistent and repeatable methodologies. The process is automated and well documented to prevent spending time reinventing the wheel and to achieve higher release velocity. Also, it defines the strategies for Canary releases and rolling back the releases in case of a failure.

- **Configuration Management**
  Configuration changes are one of the most common cause of the incidents in production environment. Configuration management requires an effective collaboration between release engineers and SREs. Google's models for distributing configuration files:
  - Use the mainline for configuration -modify configuration files in the main branch. The changes are reviewed and then applied to the running system.

- Include configuration files and binaries in the same package - for projects with few configuration files or projects where the files (or a subset of files) change with each release cycle, the configuration files can be included in the release package along with the binaries.
- Package configuration files into "configuration packages." - release configuration files alongside binaries.

Service Operation

- **Monitoring (SLI and Error Budget)**

  SRE is particularly applicable in vast and distributed IT environments—including cloud, on-site, and hybrid infrastructure environments. The SRE role is focused on maximizing the opportunities and mitigating risks associated in these infrastructures by monitoring the Key Indicators and keeping the track of the error budget spent.

- **Incident Management**

  SRE defines a well-designed Incident Management is that clearly defines the role and responsibilities of everyone in the team, involved in the Incident.

  The “three Cs” (3Cs) of incident management in SRE:

  - Coordinate - response effort.
  - Communicate - between incident responders, within the organization, and the customer.
  - Control - maintain control over the incident response.

  Both ITIL4 and SRE focuses on fast engagement to reduce the time and impact of an incident. SRE defines the monitoring systems and dashboards to be shared with key stakeholders to measure the current state. SREs have the technical ability to diagnose and fix incidents independently, so the ability to capture knowledge at the source is shortened and enables breakdown of silos by recording incident activities.

- **Problem Management through postmortems**

  Postmortem enables learning from the incidents and preventing repeat outages because of the same cause. Postmortem in SRE, involves deep diving of the problem and actions to prevent repetition.

Continual Service Improvement

- **Foster continuous learning**

  SRE encourages a culture where "fail fast and learn fast" is the key to improvement similar to ITIL. In SRE, failure is an opportunity to improve.
- **Other areas for service improvement in SRE are:**
  a. Review and revise objectives (SLOs).
  b. Reduce the toil.
  c. Prioritizes Reliability work.

SRE Adoption and Change management (Prosci® ADKAR® model approach)

A

AWARENESS OF THE NEED FOR SRE

Why we are adopting SRE and what is the risk of not using SRE.

D

DESIRE TO ADOPT SRE

What SRE has for me and my organization.

Why should I embrace the change?

K

KNOWLEDGE OF HOW TO ADOPT SRE

Build Knowledge and ability (training, workshops etc.) to participate.

A

ABILITY TO DEMONSTRATE SKILLS & BEHAVIORS

Define Tools, process and method and ensue team’s readiness.

R

REINFORCEMENT TO USE SRE

Measure and Sustain.

Continuous learning and Improvement.

Production Readiness Review (PRR) Model

Production Readiness Review (PRR) is the most typical initial step of SRE engagement. PRR is a process that identifies the reliability needs of a service based on its specific details. A PRR is a prerequisite for an SRE team to accept responsibility for managing the production aspects of a service.

When a development team requests that SRE take over production management of a service, SRE measures the service from various aspects and the availability of skills in SRE team post which SRE initiates a Production Readiness Review with the development team.

The objectives of the Production Readiness Review are as follows:

- Verify that a service meets accepted standards of production setup and operational readiness, and that service owners are prepared to work with SRE.
- Improve the reliability of the service in production, and minimize the number and severity of incidents that might be expected.

PRR success criteria include affirmative answers to the following key exit questions:

- Has the system product baseline been established and documented to enable the system to be produced?
- Has a change control process been established?
- Are adequate processes and metrics in place?
- Are the risks known and manageable?
- Is the system producible within the production budget?

A PRR follows several phases, that may proceed independently in parallel with the development lifecycle. Please go through the PRR Life cycle in Google SRE Book.

SRE Engagement Model

SREs owns production responsibility for the reliability of the important services. SRE is concerned with several aspects of a service such as:

- System architecture and interservice dependencies
- Instrumentation, metrics, and monitoring
- Emergency response
- Capacity planning
- Change management
- Performance: availability, latency, and efficiency

SRE targets to improve the services from all of these axes to make managing service easier.

SRE Governance

The first priority in SRE Governance is to assure alignment of services with organizational goals and objectives. This priority is accomplished through the implementation of IT Service Management processes. IT Service management is a close looped system that continually provides feedback for process and service improvement.

SRE Governance is about the customer and meeting the needs and requirements of the organization.

1. Operations should not operate in a vacuum but should be aligned with the business strategy, needs and objectives
2. It is equally important for the business to understand the principles SRE and how it can help manage risks and assure effective service delivery and customer satisfaction.

SRE recognizes the unique needs of every organization and helps design processes and metrics that will ensure Continual Service Improvement throughout the lifecycle of a service.

References

1. What is ‘Site Reliability Engineering’? [video]
   https://landing.google.com/sre/interview/ben-treynor-sloss/

2. Site Reliability Engineering [book]
   Edited by Betsy Beyer, Chris Jones, Jennifer Petoff and Niall Richard Murphy
   https://landing.google.com/sre/books/

3. The Site Reliability Workbook [book]
   Edited by Betsy Beyer, Niall Richard Murphy, David K. Rensin, Kent Kawahara and Stephen
   https://landing.google.com/sre/books/

4. Business Monitoring: If You Can’t Measure It, You Can’t Improve It [blog]
   https://www.anodot.com/blog/business-monitoring-incidents-cycle/

5. Fundamentals of an error budget policy [video]
   https://www.coursera.org/lecture/site-reliability-engineering-slos/fundamentals-of-an-error-budget-policy-EMoEZ

6. How ITIL4 and SRE align with DevOps [blog]
   https://techbeacon.com/enterprise-it/how-itil4-sre-align-devops

7. SRE vs ITOps: Are SRE and IT Operations the Same? [blog]
   https://www.bmc.com/blogs/sre-vs-itops/

8. Production Readiness Review (PRR) [pdf]
   https://myclass.dau.edu/bbcswebdav/institution/Courses/Deployed/TST/TST303/Student_Materials/Student%20CD%20Jan17%20%28Obsolete%20Update%20Pending%29/Tech%20Reviews%2C%20M%26S%2C%20DoDAF%20Fact%20Sheets/PRR%20Fact%20Sheet.pdf