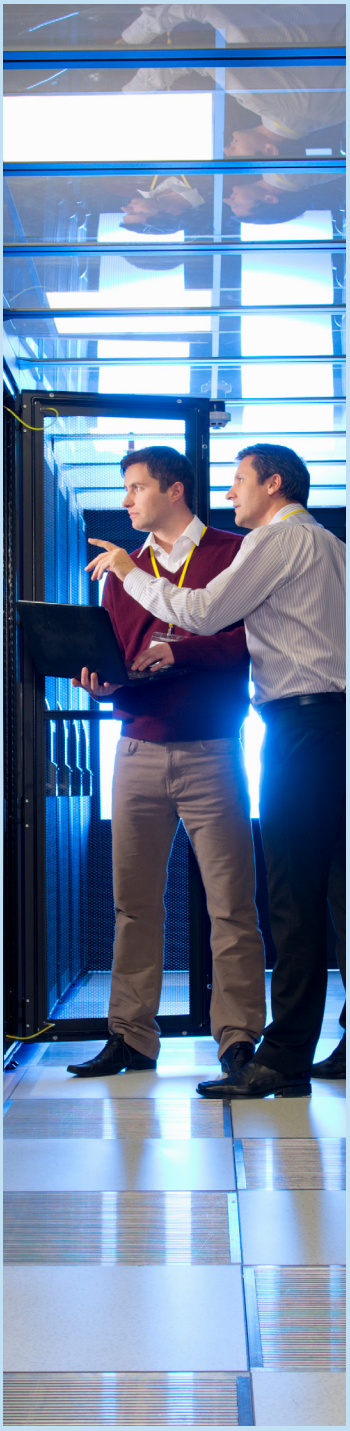
VIEW POINT



MATURING SITE RELIABILITY ENGINEERING – A PRAGMATIC APPROACH

Infosys®
Navigate your next

Last year, in December, there were three AWS outages within three weeks that affected end-users of services like Xbox Live, Ring, Disney+, T-Mobile, Fortnite, Slack and more. AWS was not the only one; in 2021, there were several major outages from similar large providers like Google Cloud, Fastly and Facebook. This brings home the following key points:

- No service is robust enough to fail. In the complex world of the internet, there are too many unknowns to guarantee a never-fail service being used by millions of users. How quickly the service recovers is key in such scenarios.
- Companies like Google, Facebook, Microsoft, Apple, Amazon and Netflix generate almost 57% of the internet traffic (source: IEEE Communications Society). Compared to these numbers, the one-off outages are still a drop in the ocean. Therefore, we must recognize the ever-evolving and well-engineered practices of these technology pioneers, which also predict and prevent wide-scale failures from occurring every second.

Site Reliability Engineering (SRE) is a key engineering practice that organization are adopting rapidly to ensure that their popular services are working optimally round the clock in a secure, stable and resilient manner. As a result, SRE is becoming one of the main enablers to help maintain the equilibrium between stability, innovation and agility of software services. In addition, its adoption is becoming widespread beyond the hyperscalers of the world

- SRE adoption grew to 22% in 2021 from 15% in 2020 and is increasingly becoming a must-have skill *(Upskilling 2021: Enterprise DevOps Skills Report)*.
- SRE is one of the top tech jobs with growing global demand *(LinkedIn)*

Some important drivers fueling SRE adoption include digital transformation, global shift to SaaS based products, accelerated adoption of Agile, highly distributed and loosely coupled system architecture and talent turnover. Another key catalyst in the adoption of SRE is the shift to remote working in the 2020s – the ability to connect and consume heavy-duty services from anywhere, anytime, any device and not necessarily from a secure firewall in your office has stressed the need to have strong SRE practices to ensure your software works seamlessly. That is also why in the last two years, we saw many vulnerabilities and their underlying infra (on-prem or cloud) exposed via scenarios for which these services were not designed or tested. As a result, SRE is now one of the main strategic engineering bets for most software product and services organizations.

At Infosys Engineering, we offer a pragmatic approach and an industry standard capability maturity model for organizations across industries to adopt SRE. Our approach has been built and influenced by several years of industry experiences especially from services provided in mission-critical areas like medical and aero industries, software product pioneers like Google and fault tolerant computing leaders like Tandem. In addition, the approach caters to all reliability requirements starting from 3 9s to even 6 9s.

This SRE Maturity Model can be harnessed to assess any client application or product to understand its current SRE capability and maturity level. This model covers assessment across seven key dimensions - continuous observability, seamless upgrade, design for resiliency, automated incident and service management, SLO & OKR management, compliance management, continuous feedback and knowledge sharing. These dimensions are further assessed based on over 60 parameters to understand the current state of SRE adoption. These seven dimensions are:

1. Continuous Observability:

This dimension assesses the system capability to collect all kinds of **telemetry data and** have **automated alerts** based on thresholds. It also checks whether the data is captured at the **right data aggregation and granularity level** and if the information is **correlated** for decision making to filter out noise. Overall, this can help reduce Mean Time To Detect (MTTD) and Meant Time To Identify (MTTI) issues. Comprehensive observability encompassing your technical services and dependent components from third-party providers (CDN, ISPs, etc.) is a must.

2. Seamless upgrades:

Automated, frequent and robust deployments with **quick feedback loops** enable "fail forward" and avoid process overhead. Seamless upgrades utilize strategies like **blue/green, automate canary releases, and feature toggle** to ensure that deployment time and success criteria are under control.

3. Design for resiliency:

This dimension tries to address **unavoidability and unpredictability of failures.** Systems should be designed with a **capability to predict failure patterns** and **seamlessly recover from them** without manual intervention to ensure dependable services. It ensures that strategies like **graceful degradations, chaos testing** and **auto-scalability** are applied to achieve a resilient design.

4. Automated incident and service management request:

With the complex systems of today, issues are bound to arise. But it is important to ensure that **critical events are detected, addressed** and **resolved quickly and efficiently.** As a result, **user service requests and housekeeping tasks are automated** and self-serve mode (as much as possible) is enabled. This dimension looks at different angles like automated triaging of issues, self-service abilities and even recommendation and auto-healing aspects

5. SLO and OKR management:

Objectives and key results (OKR) is a goal setting framework used by individuals, teams and organizations to define measurable goals and track their outcomes. **Right SLOs (Service Level Objectives) ensure right balance between operations, development velocity and release frequency.** This, in turn, helps define the right SLIs (Service Level Indicators) and ensures acceptable reliability for end-users in terms of availability, latency, quality for request response kind of scenarios and coverage, correctness, freshness and throughput for data processing. Regular **automated measuring of SLIs to make sure SLO goals are met, and error budgets are utilized optimally** in making the system more reliable is a critical step in ensuring happy users

6. Compliance Management:

Continuous compliance via automation of compliance processes helps **speed up release cycles** and free the team's bandwidth to **focus on innovation and reliability** of existing services. It also explores the possibility of implementing **compliance as code** where **policies, rules and approvals are enforced and tracked via automated controls.** Another key aspect of compliance management is to have comprehensive traceability and auditing mechanisms

7. Continuous Feedback and Knowledge Sharing

This practice has **blameless postmortems** at its core, which if practiced in its true spirit **improves the processes collaboratively.** Sharing best practices via the SRE Community of Practice (COP) helps in continuous innovation and automation, leading to a reduction in toil. It is important that **SRE teams are part of PI planning and smooth communication channels exist for this to happen**

This maturity model has evolved with usage and experiences with many of our global clientele. In addition, we have also developed and utilized various other SRE practices, tools and accelerators. These include structured frameworks for toil management, chaos engineering practices, a comprehensive framework for blameless postmortem, an error budget calculator, reference operating models, industry standard SRE metrics and more.

Here are some stories on how we enabled our clients on next-generation SRE models that harness our maturity model, associated tools and frameworks.

- At one of the world's largest software product companies based out of the US, one of their flagship products used by millions worldwide was experiencing multiple challenges on the continuous support and customer experience fronts. It was a reactive support model with an MTTD (Mean time to Detection) of issues running into hours. We implemented a solution with always "on" telemetry feature usage and user interaction backed by actionable production diagnostics and a common logging framework leading to 'smart' observability across the product. The cultural transformation happened via a new operating model based on an SRE DRI (Directly Responsible Individual) role from dev leading to proactive identification and resolution of customer issues. New tools were introduced to make automation even more reliable. '2-factor feature toggle' was another novel solution aspect that helped with exposure control of new feature issues to end-user. As a result of all these solution elements, which were a combination of SRE patterns like observability, cultural transformation, and automation, average MTTD was reduced to under five minutes and MTTR to less than an hour with a highly improved customer experience.
- A media client, a video streaming and mass marketing pioneer, faced challenges in introducing and monitoring new features. Monitoring of the application and infrastructure hosted on multiple clouds required considerable effort. In addition, there were no value additions, improvements and automation to the existing platform and the company struggled to introduce innovative features in a timely manner. From a monitoring perspective, we introduced SLI/SLO Objectives for different services to provide better uptime and SLAs for end users. Upgrades were identified and implemented to the existing deployment solutions, cross-platform connectivity like different AWS accounts and GCP for seamless deployment and monitoring. Disaster recovery strategies with automated backups and service key rotations for infra were implemented, making the platform more robust. Another key dimension of SRE maturity - knowledge sharing – was improved by setting up a process to document tickets and alerts, thereby reducing toil considerably, especially for repeat occurrences. A combination of these strategies led to considerably reduced deployment time, easy maintenance, fewer resources to maintain infra, lesser onboarding time for new apps and hence faster delivery of new value-added features to end-users.

As digitization is on the rise, especially to support **'Work from Wherever' or 'Work from Anywhere'** models, SRE adoption will only increase and cause changes across organizations' processes, tools and team cultures. We are already looking ahead to be at the forefront of these changes by bringing in solutions in areas of AIOps, self-healing, sustainability (green code) and predictive insights to augment current SRE practices.

About the Authors

Amit Sadana
Industry Principal, Infosys

Uday Kumar Gupta
AVP , Senior Industry Principal, Infosys

For more information, contact askus@infosys.com

Infosys®
Navigate your next