Site Reliability Engineering for CSPs: operational excellence in a hybrid world

How adapting Site Reliability Engineering (SRE) to the needs of communication and media service providers is driving a new era of innovation and operational excellence

amdocs

# launching faster and optimizing costs in the hybrid IT reality

As communication services providers are increasingly focused on providing more innovative services and digitalizing the customer experience, many are crafting strategies to expand the business into new domains, whether by investing in next-generation communications (such as 5G and IoT), expanding into media, IT, and security services, or via global expansion.

Executing these strategies requires the introduction of new technologies such as cloud computing, microservices, and the agile mode of operation with DevOps. Yet, at the same time, the majority of their existing IT systems, which are not digitalized, will not be going away any time soon.

Thus, operators must manage IT operations in a hybrid mode, running and managing the old and the new in parallel.

Hybrid IT crosses many dimensions of change, and it's here to stay
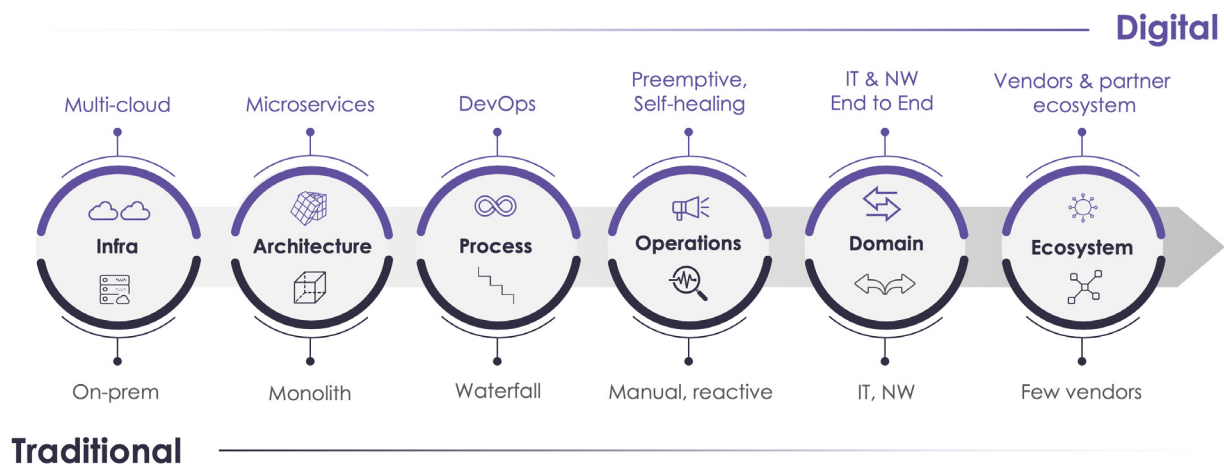
Operators' hybrid IT crosses multiple dimensions of change. As service providers strive to achieve a faster time-to-market, while providing a seamless customer experience and reducing operational costs, they must manage a broad variety of technologies and environments, skills, processes and organizational structures effectively, and bridge many internal silos.

In fact, they must not only digitalize customer channels and services but also transform their IT operations to become a **digital organization**.

At Amdocs, we believe that the most efficient means to success in managing the hybrid environment is to create integrated operations that mask the underlying technological complexities, bridge organizational silos, and manage end-to-end business processes.

The key to achieving these goals is applying a data-driven approach and leveraging artificial intelligence (AI) and machine learning. In doing so, IT operations can drive efficiencies like never before – via smart automation, optimized decisioning, and improved service reliability.

An exciting innovation and enabler of operational effectiveness and agility in a digital world is Site Reliability Engineering (SRE), which is implemented by many digital-natives for ensuring a superb customer experience.

what is SRE and why is it crucial for digital service providers?

> *"SRE is what happens when a software engineer is tasked with what used to be called operations."*
> Ben Treynor, VP Engineering, Google

Traditional IT operations systems, tools, and methodologies were not designed for the demands of the lightning-fast, cloud-native, complex digital world. To overcome the limitations of these systems and to strike the optimal balance between speed of innovation and reliability at ultra-scale, **Google created and introduced the SRE concept and approach in 2003.** SRE calls for applying a software engineering approach to running IT operations. This way the SRE team, established by Google's then VP of Engineering, Bob Treynor, was able to break the linear connection between scale and the amount of resources required to be "always available, performant, and efficient.". By implementing SRE methodologies **companies scale faster, improve reliability, and optimize operational processes.**

SRE constitutes a proactive, systematic, and continuous approach that embraces risk and promotes innovation. This approach, together with DevOps tools and methodologies, **helps automate and digitalize IT operations** so the team can focus on building and improving the customer experience.

the main principles of Site Reliability Engineering

The SRE approach is grounded in four main principles:

1. Using "error budget" to maintain the right balance between innovation and reliability

SREs achieve this objective by leveraging two main metrics that were created by the Google SRE team: (1) **Service Level Objective (SLO)**, i.e. the target level of service reliability that is considered the threshold for customers to be satisfied with the service, and (2) **Service Level Indicator (SLI)**, the actual (measured) level of service performance.

The margin between the SLO and 100% reliability is called the **Error Budget**. For example, an SLO of 99.9% reliability leaves an error budget of 43.2 minutes each month. SREs leverage the error budget to improve operations, and whenever SLIs fail to meet the SLOs, they consume part of the error budget.

Note, the high availability required in the communications industry means that SLOs that are defined for 99.999% reliability would leave approximately half a minute of error budget each month. In such an environment, clearly it would be very difficult to operate strictly within the error budget.

2. Automation

The second **key objective of SREs is to "eliminate toil", primarily through automation**. Toil is defined by Google's SRE organization as "the kind of work tied to running a production service that tends to be manual, repetitive, automatable, tactical, devoid of enduring value, and that scales linearly as a service grows."

Eliminating toil provides value to the employee, who can then focus on more engaging and challenging tasks. And, when toil is automated the whole team can then focus on what the company needs most to compete effectively and stay ahead, i.e. to innovate.

3. Blameless postmortem

Learning from failure via 'blameless post mortem' is the third main principle of SRE culture. When failure does happen – and at some point, it will – it is critical to be able to extract the requisite insights and prevent issues from becoming recurring events.

The main tool for learning from incidents is the post-mortem, the official report on what happened, what was its impact, what actions were taken to resolve it, what is believed to be the root cause/s, and what is being done to prevent the failure from happening again.
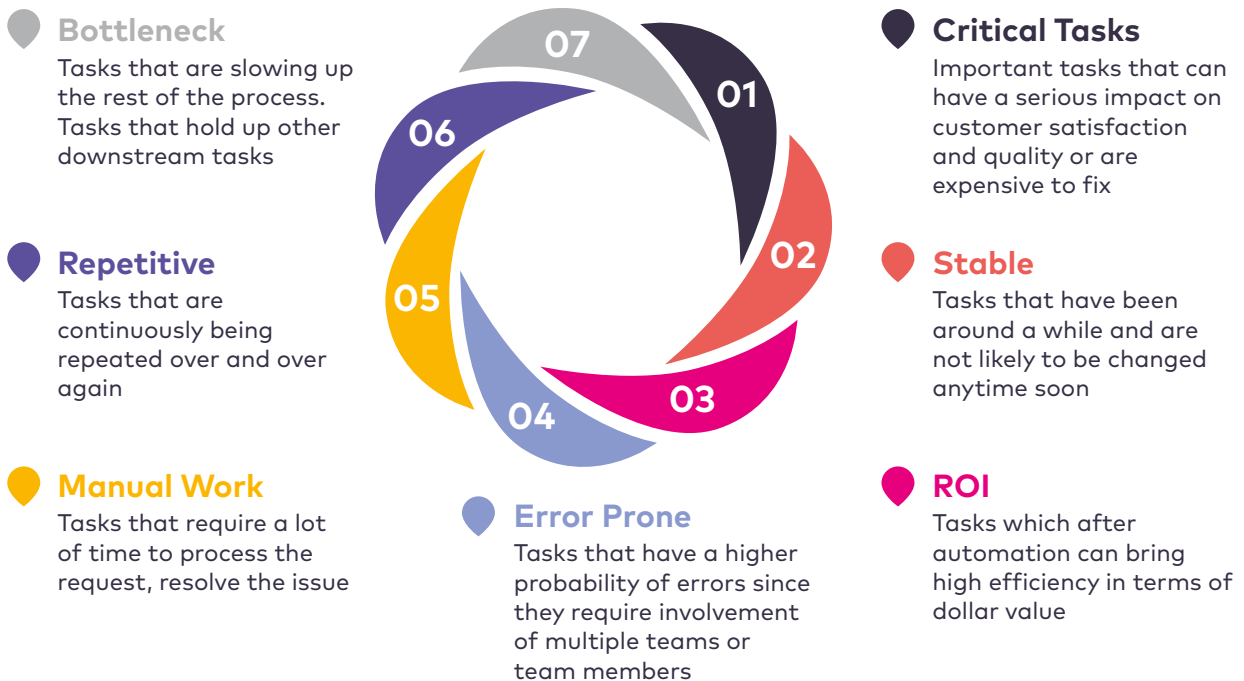
Blameless postmortem is different from other post-mortem approaches, in that it does not include finger pointing. When an issue occurs, the blame is placed on the process that enabled mistakes to happen, and not on the individual who made the mistake: in case of a human mistake the process should have been able to accommodate or prevent the mistake.

4. Measuring and maintaining reliability

As noted earlier, the main metrics used for optimizing service reliability decisions are SLOs and SLIs, which help to improve service reliability, nurture customer focus, and ensure data-driven decision making.

Any changes made to improve reliability are validated against the SLI curve. This curve helps the SRE who is asked: "are we now at the expected SLO level?" If the answer is "no," the SRE will analyze the curve and then seek out other areas that require improvement in order to bring the SLI to the acceptable level.

Identifying the right automation candidate

Bottleneck
Tasks that are slowing up the rest of the process. Tasks that hold up other downstream tasks

Repetitive
Tasks that are continuously being repeated over and over again

Manual Work
Tasks that require a lot of time to process the request, resolve the issue

Error Prone
Tasks that have a higher probability of errors since they require involvement of multiple teams or team members

Critical Tasks
Important tasks that can have a serious impact on customer satisfaction and quality or are expensive to fix

Stable
Tasks that have been around a while and are not likely to be changed anytime soon

ROI
Tasks which after automation can bring high efficiency in terms of dollar value

07 06 05 04 03 02 01

how amdocs is bringing SRE to the hybrid environment of service providers

SRE was originally developed at Google to help software developers improve the production-related operations of the applications that they develop (e.g. Gmail developers working on Gmail production reliability).

This context is well suited for many SaaS companies, who deliver their own products as a cloud service to users. However, the hybrid IT complexity of communications and media providers constitutes a very different reality.

In our industry, applications are developed by multiple vendors and with a broad variety of technologies, infrastructures, and business processes. Availability is expected to be much higher than in other sectors. Namely, "telco-grade" is expected to deliver five-9's (99.999%) uptime (i.e. where there are only five minutes of downtime annually!). As a result, it is not practical to use the error budget as a tool – since the margin for error is virtually non-existent.

Moreover, SREs in our industry require domain expertise in software engineering for both new technologies as well as legacy, operations, and in the specific business processes of our domain. This combination of knowledge and skill-set is rare.

That's why, despite all the benefits that SRE brings to the world of IT operations, we have yet to see its proliferation in the communications and media industry. Until today.

Amdocs Global SmartOps is proud to introduce SRE methodologies that we adapted to the specific and unique needs of the hybrid reality of communication and media service providers.

The main principles of Amdocs SRE mirror those of Google, but with the following necessary adaptation:

- **Support for hybrid environments** with a variety of services, as well as both Amdocs and third-party systems and products;

- Amdocs SREs operate as **integrated members of existing account** teams, and not a as a separate, elite team as at Google;

- **SLOs (Service Level Objectives)**, rather than error budgets, are used **to measure** and improve reliability;

- The introduction of an **SRE automation role**, which does not exist in the Google framework, and which scales-up automations development through inner-sourcing – together with dedicated automation tools

---

SRE in SmartOps – what are the differences?

- All services provided over the cloud

- SRE is a separate elite unit

- Error budget is used as a tool to improve reliability

- Hybrid environment, need to support a variety of services, Amdocs and 3rd party systems, products, etc. sometimes without access to original code

- Scaling-up SRE adoption through inner-sourcing

- Focus on SLOs to improve reliability (error budget not feasible)

Along with the above noted adaptations, the original Google SRE concepts do remain, i.e.:

- Operations is run by **engineers**;

- SREs spend up to **50% of their time** performing ops-related tasks with the rest spent on agile development for **automating** and improving the operations;

- Nurturing a **blameless culture** through blameless postmortems;

- Using **SLIs (Service Level Indicators) and SLOs (Service Level Objectives)** to measure and improve service reliability

Change Management

Adapting SRE at Amdocs required a great effort of change management.

Evolving the mindset and work practices of an organization comprised of thousands of employees who serve dozens of customers with varying needs is no easy task. This required a strong commitment from top management as well as access to specific tools that would help individuals across the organization recognize the benefits of this change and participate in bringing about the paradigm shift. To support the implementation process we allocated dedicated SRE coaches who served as change agents aiming to facilitate the mindset transition and promote SRE practice implementation, by ensuring SRE principles, processes and routines are being followed.

At the same time, we also needed to closely monitor progress at the account and overall level.

Moreover, further to our view of SRE implementation as a change in mindset, we also decided to adopt two new SRE roles:

Global SRE
Classic SRE role, global team working on proactive SRE missions for different accounts.

Automation SRE
Focusing on automation development and spending roughly 50% of their time on ongoing operations and 50% to identify toil and develop new automations.

An Amdocs SRE innovation: the atomIQ platform enabling SRE automation at scale

While the original SRE concept holds automation and DevOps tools as major focus areas, in practice it does not dictate SRE-specific tools nor a dedicated SRE platform.

SREs are expected to use existing tools and methodologies for solving root-cause issues in the application code, among other tasks.

In a telco environment, it is nearly impossible for the operational workforce to fix application code as it is written by completely different teams or external third-parties. Therefore, solving operational issues requires a different approach, that may include the involvement of both the original developers as well as various automation tools.

This requirement drove us to develop a unique automation platform that harnesses AI and inner-sourcing to provide scalability specifically for SRE work.

This platform, atomIQ – provides not only a safe environment for creating automations using "light code" (i.e. simple python scripts), but also serves as a global repository of automations that can be used and re-used for development of new ones across customers, while ensuring customer privacy and confidentiality.

In addition, it enables scaling up automation development as it allows a large number of operations engineers to implement automations in "inner-sourcing" mode.

atomIQ SRE Platform:
Re-inventing operations

6
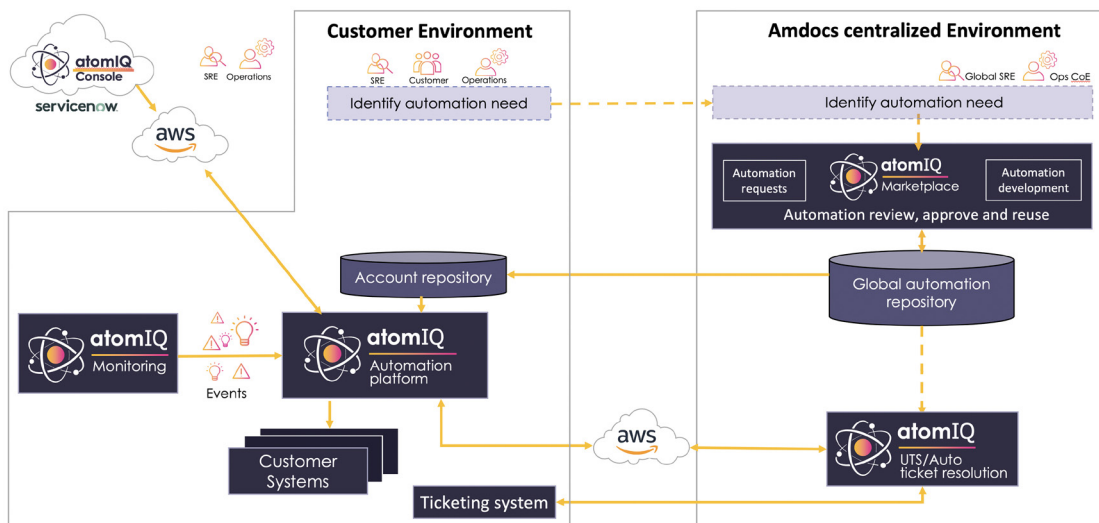
an amdocs SRE case study

A large, multi-play service provider contracted Amdocs SmartOps to run its IT operations as a managed service. The customer's IT environment was extremely complex, a result of multiple mergers and acquisitions.

The Amdocs SRE team analyzed various operational domains in order to identify which ones can be improved, while focusing on driving the greatest value with the fastest results.

Since the operator runs a large pre-paid business, failed USSD (Unstructured Supplementary Service Data) transactions were a major issue (pre-paid customers frequently use USSD for balance checks, top-ups, plan changes, etc.).

The SRE team measured the SLI for USSD transaction success and identified an up to 5% failed USSD transaction rate at peak hours. At the same time, they decided that the objective, i.e. the SLO, should be a rate of only 0.5% failures (i.e. 99.5% success).

The first step in the root-cause analysis was to map the full transaction process, which involved approximately 40 components. In addition, data flow and the system resources that were used were closely analyzed. This data was then used to re-build the process for robustness, where configurations were changed to accommodate the discrepancy between off-peak and peak hours, and databases were tuned as resource allocation was re-mapped.

Within just a few weeks the impact was profound. Peak-hour transaction failures decreased from 5% to 0.3%, outperforming even the SLO.

summary: revolutionize your operations

At Amdocs, we are excited to bring a new paradigm in how the IT operations of service providers can ensure reliability and efficiency while empowering innovation for digital leadership.

We believe this new paradigm is vital both for our success as well as for the success of our customers on their way to becoming digital service providers.

To get started today on the revolution with Service Reliability Engineering from Amdocs SmartOps, we invite you to read more about Amdocs SmartOps and get in touch with us at SmartOps@amdocs.com

about amdocs

Amdocs is a leading software and services provider to communications and media companies of all sizes, accelerating the industry's dynamic and continuous digital transformation. With a rich set of innovative solutions, long-term business relationships with 350 communications and media providers, and technology and distribution ties to 600 content creators, Amdocs delivers business improvements to drive growth.

Amdocs and its 25,000 employees serve customers in over 85 countries. Listed on the NASDAQ Global Select Market, Amdocs had revenue of $4.1 billion in fiscal 2019.

For more information, visit Amdocs at www.amdocs.com

www.amdocs.com

amdocs