Site Reliability Engineering

Nov 2022

Reliability Engineering at EY

Need for Reliability Engineering

EY technology products are increasingly designed and dependent on diverse Microservices cloud ecosystems. Supporting EY's increasing portfolio of complex cloud technology products requires a new approach founded on industry-proven reliability engineering principles.

Reliability Engineering at EY

Reliability Engineering improves product availability by combining highly skilled reliability engineers with software and system engineering mindsets, who apply intelligent monitoring providing holistic observability across critical product architectures, driving continuous product improvement processes through automation and self-healing capabilities while embedding with product teams helping design the next generation of resilient and highly available EY products.

It's finally all about delivering business value to the customer through reliable services.

What Site Reliability Engineering(SRE)?

Site Reliability Engineering(SRE) is a practice dedicated to helping organizations sustainably achieve the appropriate level of reliability in their system, services, and products through intelligent monitoring and observability, continuous improvement processes, and automation. SRE capability also includes DR as a service based on Business Impact assessments.

Value Proposition

- Reliability focus on system, service, and products to achieve the client requirements with SLI/SLO/SLA targets.
- SRE tools for all the SRE practice implementations
- Implement intelligent Monitoring, observability and alerting
- Continuous product and platform reliability improvements
- SRE Squad with deep practice knowledge
- Automation and toil reductions
- 24 * 7 on-call availability for critical incidents and Blameless RCA
- Disaster recovery service design and implementations
- Monthly reliability reports and self-service tools
- Product Performance consulting engineering

Benefits

- Availability targets based on Nines (99. XXX)
- Measure reliability with SLO/SLI/Error Budget
- Proactive Monitoring to reduce incidents
- improve on lower MTTR and increased MTTB
- Reduction of Toils
- Automated tools for Alerts, Keys
- Organization: Stable, reliable services, improved customer experience, culture of collaboration between development and operations

BigDash

Dashboarding
- Automate deploying infrastructure level dashboard.

Monitoring
- SSL, SPN, DNS, Key vault monitoring, alerting expiry.
- Mega dashboard based on availability.

Reporting
- SLO/SLI document automation
- Monthly report on reliability

Alert Management
- Deploy, enable, disable, and delete alert rules with a single click.
- Set up a maintenance plan to disable alerts.

Integration/API and many self-service utilities

Bigeye

- Single stack view of Business, Application, and Infrastructure metrics
- Integrate with diverse data sources.
- End-to-end observability with hierarchal level dashboard (N1,N2,N3,N4)

N1- Availability SLI/SLO
Dashboard for measuring workflow and user journey

N2- Performance
Dashboard for measuring detail-level API performance matrices

N3- Infrastructure
Dashboard to capture all the Azure resources and Services Metrics.

N4- Custom/Business needs
Custom Dashboard based on product and business requirements. E.g., alerts dashboard, SSL&SPN, User Level, NFR, etc

Custom Tools

- Alertsite for Synthetic transactions
- Custom functions based on requirements
- Scripts developments based on requirements.
- Performance engineering
- Automated configuration drifts
- Pipelines for resource deployments
- SharePoint for collaboration and documentation

Contacts: Bijay.nayak@gds.ey.com/pankaj.vishwakarma@gds.ey.com

SRE Services

Owns Reliability

Ensure the availability of applications according to business requirements.

Observability

Implement robust observability solutions to collect and analyse metrics, logs, traces, and other relevant data.

Measure Reliability with SRE practices

Define SLI/SLO/Error Budget based on the User Journey of Products. SLI- Service Level Indicator, SLO-Service Level Objective

Synthetic Monitoring

Track the performance of applications by mimicking user actions and directing the flow within the application.

Incident/Alerts Management

Effectively managing unforeseen incidents or disruptions and efficiently collaborating with stakeholders can be achieved through automation of alert management.

On Call

24*7 on-call support for critical incidents impacting service reliability.

Blameless Postmortem

Help in blameless Root Cause Analysis for critical incidents. Response and actions to avoid re-occurrence without emphasizing individuals.

Efficiency and Performance Engineering

Optimization of application components to avoid slowdowns and have minimum reaction time under production loads for great user experience

Production Readiness Review

Review Application readiness in productions ensuring reliability.

Automation

Automation aims to improve efficiency by eliminating toil and reducing manual, repetitive tasks.

Self Service Tools

Product teams will have easy access to self-service tools.

Reliability View

Monthly or quarterly reports on product reliability.

Costing

Reports include Azure cost, which helps the team stay on top of consumption.

DR Implementation (DRaas Service)

DR Solutioning and designing for High availability and Disaster recovery capabilities.

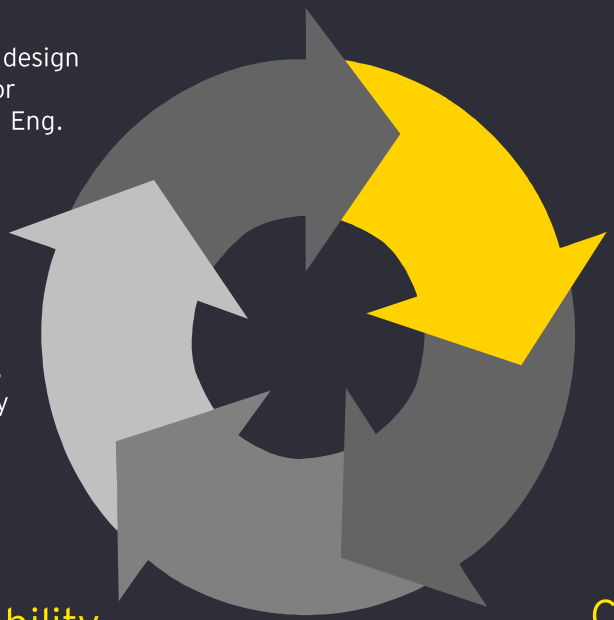
SRE Practice

Design Reliable Products

Assess and Identify current product reliability design gaps. Product design and architecture, built for reliability @ scale, leveraging Software/system Eng. approach for continuous improvement

Establish Product SLO/SLIs

Agree to product service level indicators and service level objectives, defining acceptable product error budget measuring product reliability.



Minimize Toil by Automation

Eliminate/Minimize Toil and improve reliability through automation and auto-remediation. Reduces manual processes while improving product reliability through automated response.

Create Monitoring/Observability

Ensure services and components are intelligently monitored for metrics, events, logs and traces. Drives error budget, measures reliability, enables alerting for rapid incident response

Culture and Continuous Reliability Improvement

Reliability backlog drives continuous product improvement process.

Critical Incident Response

Rapidly respond and resolve critical Product reliability incidents. MTTR (mean time to respond) and Blameless Retrospectives.

SRE Keys

SLI [Metric identifier] [Operator] [Metric]	SLO [Objective] [SLI] [Period]	ERROR BUDGETS [Error Budget] [SLI]
Home page request served in < 100 ms	95% of home page requests served in < 100ms over past 24 hours	Allow 5% failure of home page requests served in < 100ms over past 24 hours
95th percentile of Home page latency over 5 mins < 200ms	99% of 95th percentile of home page latency over 5 mins < 200ms for the past month	Allow 1% failure of 95% percentile home latency over 5 minutes < 200ms for the past month
Requests should be completed within 250 ms	95% of requests should be completed within 250 ms over 24 hours	Allow 5% failure of requests should be completed within 250 ms over 24 hours
Services should be available for 99.99% of time (based on heartbeat events from bounded system)	95% of Services should be available for 99.99% of time over 30 days	Allow 5% failure of services availability over 30 days
Book page request response code != 5xx	99% of book page request response code !=5xx over the past 7 days	Allow for 1% failure of book page request response code != 5xx over the last 7 days

SRE Practice and Product team responsibilities

SRE Team Responsibilities

- Ensuring Product Availability/Safeguarding the Error Budget
- Establish Observability real-time Monitoring, SLO/SLIs, Alerting
- Responding to and Resolving critical Product availability incidents
- Ongoing Error budget reviews with Product Teams
- Manages reliability backlog tracking recommendations and continuously implementing reliability improvements
- Prioritizes and manages reliability Backlog. priorities based on current Error Budget
- Controls new product feature release based on reliability and error budgets.

Shared Responsibilities

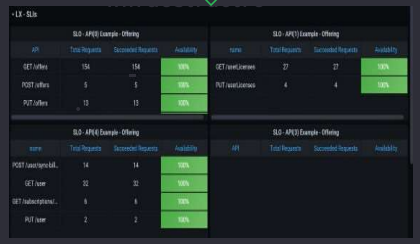
- Partners ensuring Available and Reliable Products

Product Team Responsibilities

- Creating Reliable Products from inception
- Mandatory reliability Improvements when Error Budget exceeded
- Partners with SRE team making reliability a priority

SRE Capabilities : Big Eye

N1 – Availability-SLI-SLO



N2 – Performance

N3 - Infrastructure

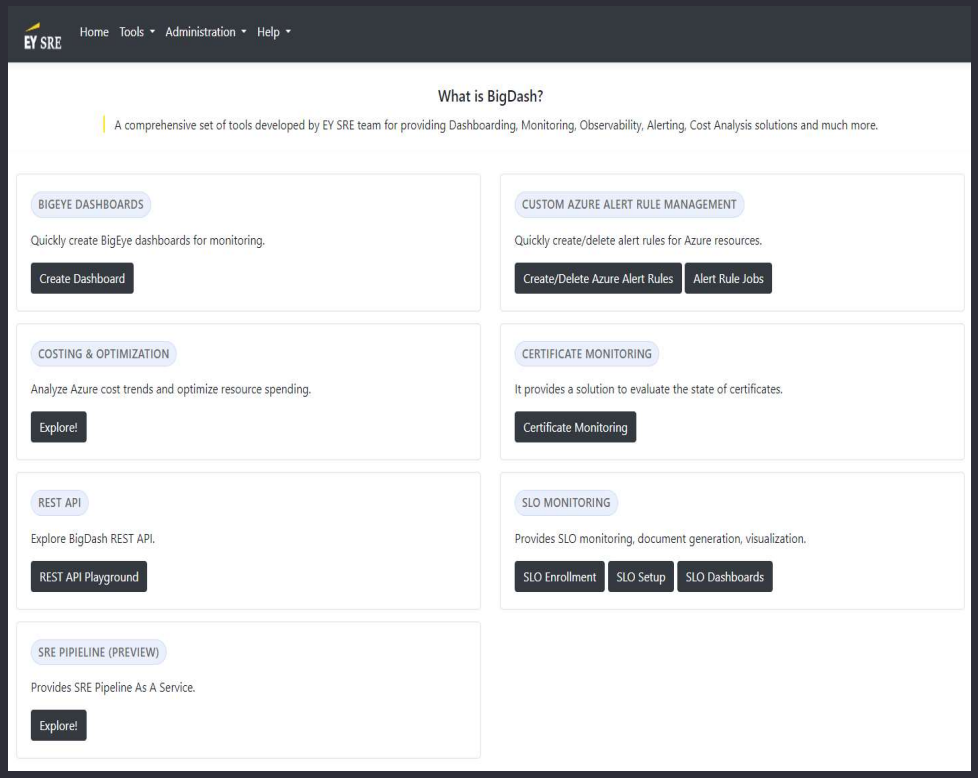
- N4- Custom Dashboard/Business needs

- Single Full stack view of Business, Application, and Infrastructure metrics
- Meets all monitoring needs of operations, development, and business teams
- Access and login through EY active directory
- Integrate with diverse data sources Azure log analytics, Appin sight, Prometheus, SQL, Elasticsearch, Azure graph explorer, simple JSON, Postgres, etc.
- End-to-end observability with hierarchal level dashboard (N1,N2,N3,N4)

- N1- Availability SLI/SLO
  - Dashboard for measuring workflow and user journey with latency, traffic, errors, and saturation indicators.
- N2- Performance
  - Dashboard for measuring detail-level API performance matrices.(Request, response, exception, Latency)
- N3- Infrastructure
  - Dashboard to capture all the Azure resources and Services Metrics.
- N4- Custom/Business needs
  - Custom Dashboard based on product and business requirements. E.g., alerts dashboard, SSL&SPN, User Level, NFR, etc.

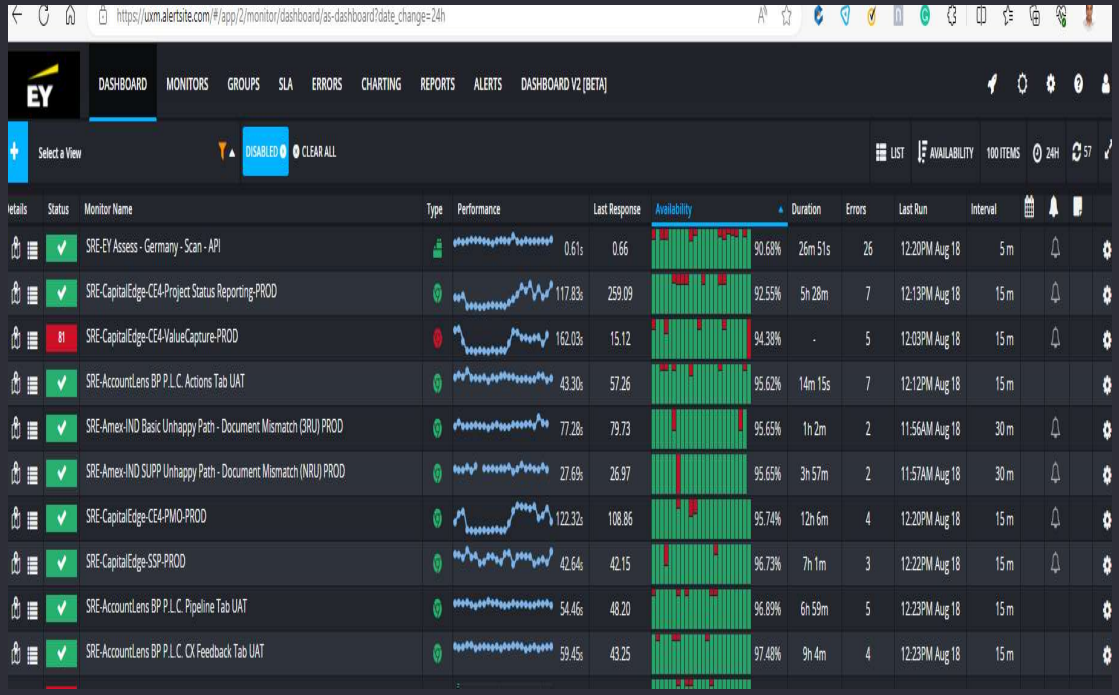
- SLI - SLI stands for Service Level Indicator and refers to a measurable metric used to assess the performance or quality of a service
- SLO- SLO stands for Service Level Objective and represents a specific target or goal for a service's performance or reliability.
- ErrorBudget- Error Budget refers to the allowable amount of service degradation or downtime within a specific timeframe that can be tolerated while still meeting the Service Level Objective (SLO).

SRE Capabilities : Big Dash



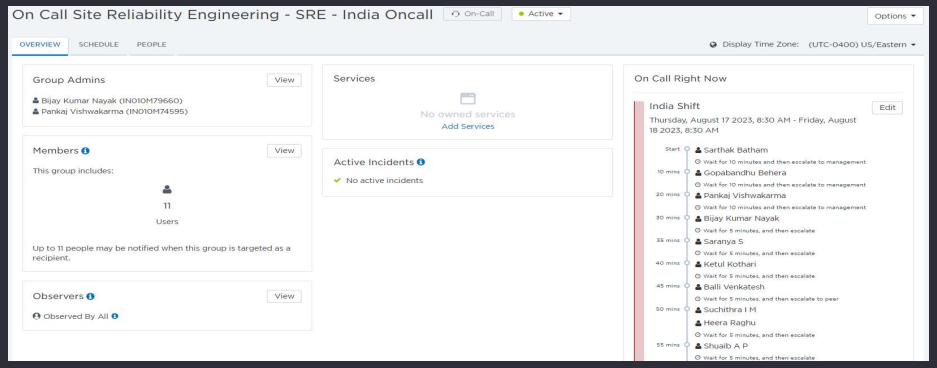
- A comprehensive set of tools the EY SRE team developed for providing Automation, Dashboarding, Monitoring, Observability, Alerting, Reporting, and Cost Analysis solutions.
- Integrated with Azure active directory to support login.
- Team management features to share monitoring setup with teams.
- Dashboarding
  - Automate deploying infrastructure level dashboard.
- Monitoring
  - SSL certificate monitoring.
  - Service principal and application registration.
  - DNS monitoring.
  - SLO/ SLI and Error budget monitoring.
- Reporting
  - Mega dashboard: High-level executive report with SLO details about SRE onboarded products.
  - Product monthly report: Overview of SRE offering - monitoring (SLO/SLI, Service principals, Certificates, and DNS), alert details (triggered and deployed), and Costing
- Alert Management
  - Manage Alert rules - deploy, enable, disable, and delete alert rules at scale with single click.
  - Set up a maintenance plan to disable alert rules to reduce noise during deployment/maintenance.
- Integration
  - Integrated with SNOW to create tickets for SSL, SP expiration.
  - Integrated with SendGrid, Teams API to send notifications via emails and on teams.
  - Integrated with Azure monitor (log analytics and application insights) used for SLO/SLI monitoring.
- REST API
  - Offers Rest API to perform operations related to managing dashboards deployment, alerts, certificates, service principals, etc.
- Utilities
  - Ability to view and close SRE-specific incidents.
  - Ability to enable the diagnostic setting for resources at scale to send logs.
  - Allows raising Service requests for any SRE offering.
  - Easy usability with inbuilt documentation.

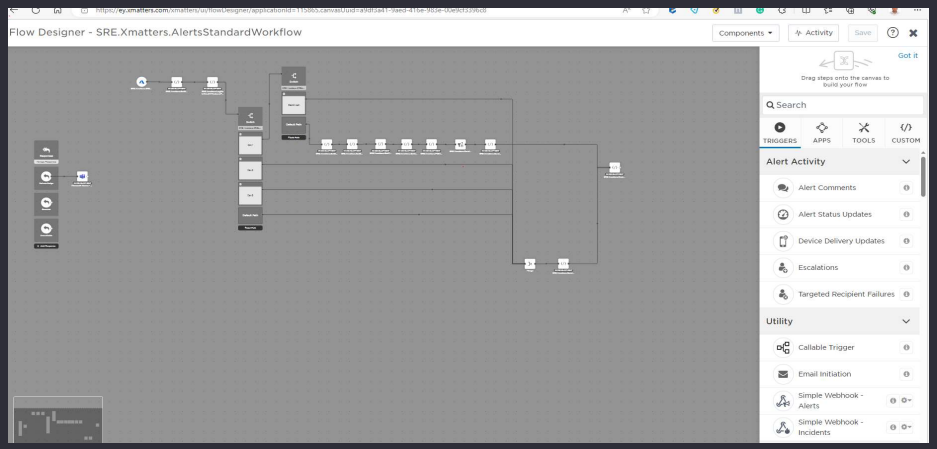
SRE Capabilities : Alert site(Synthetic)



- Synthetic monitoring to simulate transactions or steps to identify issues related to availability and performances.
- Capability
  - Web monitoring
    - Create single URL or multi-step web transaction monitoring.
  - API monitoring
    - Import existing ready API or postman test cases
  - Network monitoring
    - DNS, Ping, and TCP monitoring
  - FTP monitoring
    - FTP and FTPS server monitoring.
- Ability to invoke monitor from different locations including EY private network
- Generate availability and performance alerts

SRE Capabilities: XMatters

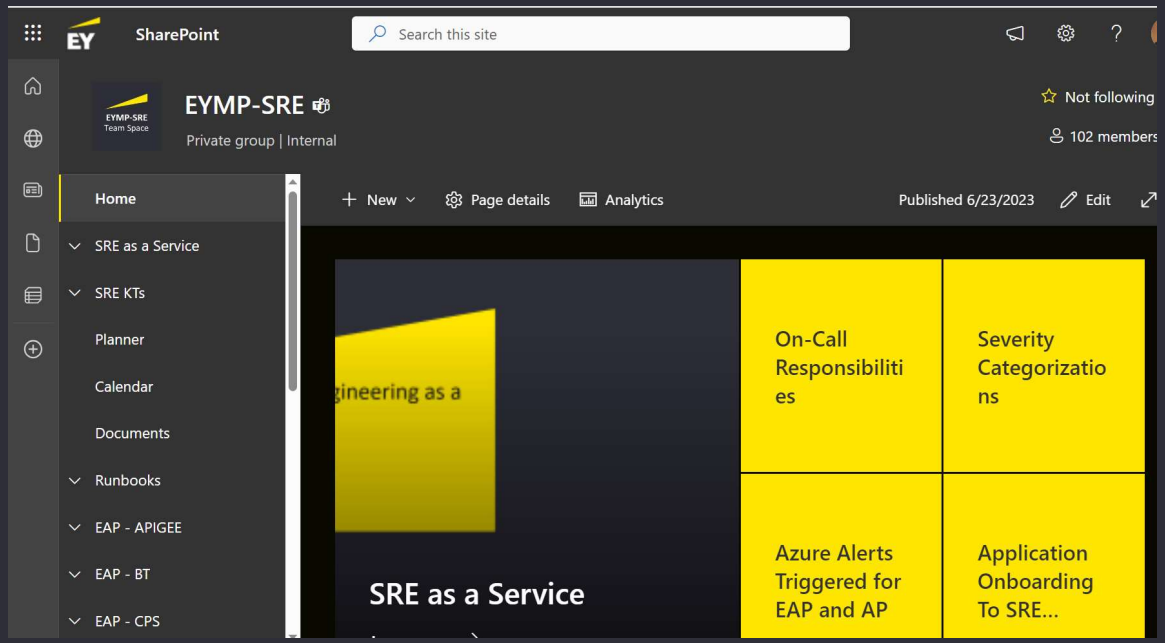




- Tool to automate Incident workflows.
- Integration with Azure monitoring.
- Send alerts using email and phone calls.
- Integration with ServiceNow for auto ticket generation for critical alerts.
- Ability to log all triggered alerts
- Custom workflow as per clients' requirements.
- 24 * 7 SRE on-call for Critical alerts
- Manage on-call Rota

SRE Capabilities : SharePoint Portal

- Alert and Monitoring metadata
- Runbook to document processes
- Root cause analysis
- Collaborations for all the stakeholders in one place
- Planner to visualize the work and tasks



SRE capability: DR as Service

What is Disaster Recovery

Resources and activities to re-establish information technology services (including components such as infrastructure, networks, systems, applications and data) at an alternate site following a disruption of IT services.

- Disaster recovery is a critical requirement to maintain business continuity.
- Overall business dependency on compute systems increasing continuously.
- Increased adoption of cloud technologies has added complexity as a lot of services are abstracted.
- In EY context, client has been specifically for DR Processes and DR Plans.

Prerequisites for DR

- BIA tier or RTO/RPO are identified
- Critical Data is continuously replicated to the DR environment
- Infrastructure in the DR region is either pre-created or should be easily deployable in the case of a DR event.
- DR plan should be kept up to date based on the changes in the Primary environment.
- Regular testing should be done to conducted to confirm if the solution is working.
- Contact details of the relevant team-members/Vendors should be kept up to date.

SRE capability: DR as Service Process

1. Identify business sponsors and kick-off
2. Identify process staffing and plan
3. Identify SLO (RTO/RPO), define disaster, legal obligations
4. Business impact assessment (BIA)
5. System assessment and cost analysis
6. DR Test execution and Iteration
7. DR Playbook and Checklist
8. Change management and periodic refresh

SRE Onboarding delivery

Phase 1 - (1 to 2 Weeks)*

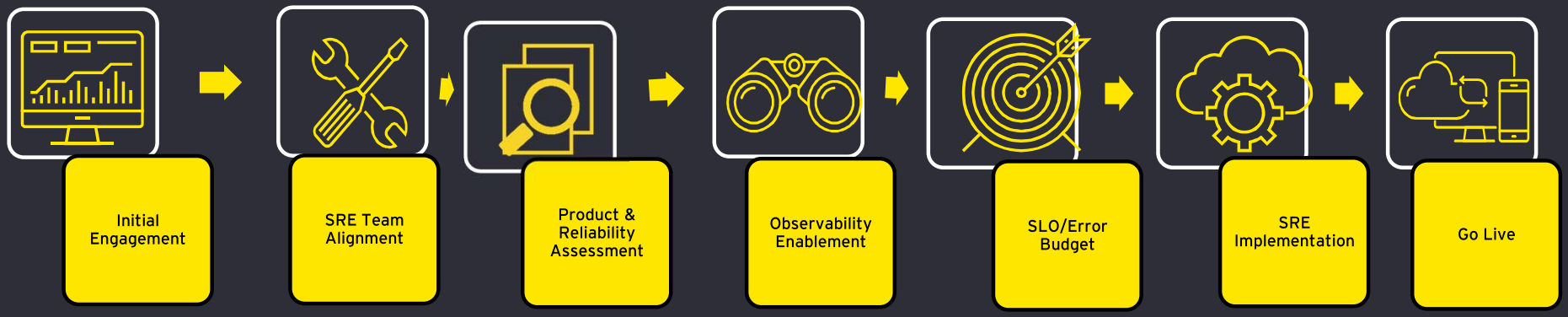
Fast Track Monitoring

Product Reliability Assessment

Phase 2(2-3 Weeks)*

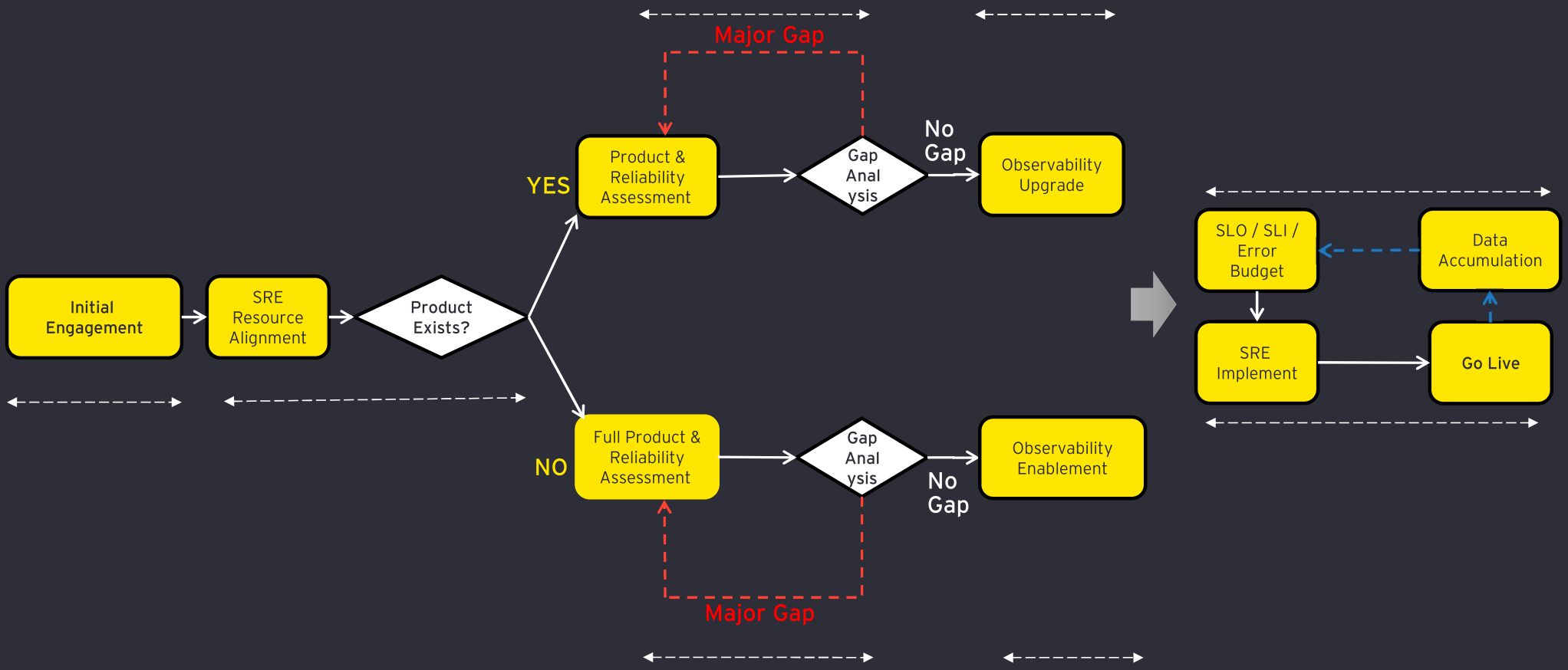
Error Budget Agreement

SRE (React, Respond, Resolve) Go-live



*varies based on product architecture complexity

SRE Onboarding Workflow



SRE Onboarding

SRE Resource Alignment

- Determine size and skillsets of the SRE team.
- Assign team if capacity is available.
- Get additional staff and assign if capacity is not there.

Product and Reliability Assessment

- Review Prd to understand the objective function Of product.
- Knowledge of Product Functional flow.
- Identify critical user paths.
- Understand Deployment view, Data models, Historical Data, Change Mgmt process & Lower-Level User Interactions.
- Production Readiness Review.
- High-Level Code Walkthrough.
- Instrumentation, metrics, and monitoring of all user-level interactions.

Observability Enablement

- Production environment user access.
- GIT Repo read access.
- Azure Subscription access.
  Observability data exploration.
- Identify minor and major gaps.
- Baseline reliability of a product.

SLO/Error Budget

- SLI Indicators
- Agree of an SLO number (Reviewed every 6 weeks)
- Error Budget Policy

SRE Implementation

- SLO/Business driven Observability, Visualization, Monitoring, Alerting
- Performance Driven.
- Infrastructure Driven.
- DR Evaluations
- Capacity Planning
- Critical Incident Management

Go Live

- On-Call 24 x7
- Incident Management (P1)
- Blameless Postmortem
- Runbook Auto Remediation.
- Automation

Pre-requisite

- Product knowledge transfer (Technical IAD and Functional) and documentation.
- Azure Subscription, Resource group Information.
- Existing logging information like Application Insights, Log Analytics,
- Azure SPN (Client ID, Tenant ID, and Client Secret) with "Monitoring Contributor" access at the Subscription Level for all Product environments. (owner should be SRE to track the expiry and able to renew)
- Please provide our AD App CTP-Platform-SRE_Monitoring-P01 "Monitoring Contributor"
- Please provide IT-CTPSRE_MSP01-Team with "Contributor" access to Product environments at the subscription level. (if require support for resolutions)
- Service Account for Synthetic monitoring Setup.
- Development team Collaboration for enabling tracing at the code level. (Opentelemetry)
- Need to enable diagnostic setting if required.
- Read-only access to the code repository.
- Product charge code for billing.
- Information about existing BCP/DR setup (If present).
- Product group email id (If present).

SRE Cost Small Product

SMALL

- < 5 Region
- < 3 dependent Clients Implementation
- Business Critical
- Non-Complex architecture(Web/App)
- Less Frequent releases
- Less customization on observability requirements
- Product Infra Cost Less

Standard

Cost **$28K/Per Year (Quarterly 7K)**

Observability Set Up on Bigeye

- N1- Availability Dashboard (SLI/SLO)
- N2- Application Performance API level
- N3- Infrastructure /Resources Level
- Synthetic tests(UMX Alertsites)
- Intelligent Monitoring, Observability, and Alerting
- All the Sev1 alerts are integrated into matter/ServiceNow and assigned to the Support team queue.
- Alerts Svr2/svr3 are set up and routed to your Team Channel/Email.
- Enabled SPN and DNS, SSL monitoring with Bigdash
- Monthly Product reports.
- Support for any future requirements and modifications based on the requirements.
- 18*5

Premium

Cost $46K /Per year (Quarterly 11.5K)

- All Services in Standard included +
- 24 * 7 On-call SRE Support
- Manpower Intervention during Critical alerts and Incidents
- Blameless RCA
- Continuous Automation and Toil reduction
- Work closely with Dev and Ops team on Reliability and fix the gaps
- All the Sev1 alerts integrated to xmatter ServiceNow and page to SRE on-call
- Svr2/Svr3 alerts are actionable.
- Help in Performance Engineering
- SRE Squad, deep in product architecture/Technical knowledge
- DR Services (additional cost)

Small	Standard	Premium
Q1	7K	11.5K
Q2	7K	11.5K
Q2	7K	11.5K
Q4	7K	11.5K
Total	28K	46K

SRE Cost Product

LARGE PRODUCT (anyone)

- >= 5 regions (< 10)
- >=3 Dependent clients Implementation (<10)
- Business Critical and Mission-Critical
- Complex Architecture (Bigdata/analytics/Micro services/Container/Custo m web jobs/Functions)
- Frequent releases
- Gold/Silver Applications (High availability)
- Custom Monitoring requirements
- Product Infra cost high

Standard

Cost $42K/Per Year (Quarterly 10.5K)

Observability Set Up on Bigeye

- N1- Availability Dashboard (SLI/SLO)
- N2- Application Performance API level
- N3- Infrastructure /Resources Level
- Synthetic tests(UMX Alertsites)
- Intelligent Monitoring, Observability, and Alerting
- All the Sev1 alerts are integrated into matter/ServiceNow and assigned to the Support team queue.
- Alerts Svr2/svr3 are set up and routed to your Team Channel/Email.
- Enabled SPN and DNS, SSL monitoring with Bigdash
- Monthly Product reports.
- Support for any future requirements and modifications based on the requirements.

Premium

Cost $92.5K

All Services in Standard included +

- 24 * 7 On-call SRE Support
- Manpower Intervention during Critical alerts and Incidents
- Blameless RCA
- Continuous Automation and Toil reduction
- Work closely with Dev and Ops team on Reliability and fix the gaps
- All the Sev1 alerts integrated to xmatter ServiceNow and page to SRE on-call
- Svr2/Svr3 alerts are actionable.
- Help in Performance Engineering
- SRE Squad, deep in product architecture/Technical knowledge
- DR Services Solutioning and enablement (one time) and handover

Large	Standard	Premium
Q1	10.5K	23.25K
Q2	10.5K	23.25K
Q3	10.5K	23.25K
Q4	10.5K	23.25K
Total	42K	92.5K

Size (Small or Large)

- **Quantitative Criteria:**

1. **Throughput and Traffic:** Measure the product's incoming requests or transactions per minute/hour. For example:
   - Small Product: Up to 1000 requests per minute.
   - Large Product: More than 1000 requests per minute.
2. **Data Volume:** Measure the total volume of data processed or stored by the product. For example:
   - Small Product: Up to 1 TB of data.
   - Large Product: More than 1 TB of data.
3. **User Base:** Count the number of active users or customers. For example:
   - Small Product: Up to 10,000 active users.
   - Large Product: More than 10,000 active users.
4. **Service Components:** Count the number of microservices or services that comprise the product. For example:
   - Small Product: Up to 5 microservices.
   - Large Product: More than 5 microservices.
5. **Availability Requirements:** Measure the required uptime percentage (availability) for the product. For example:
   - Small Product: 99.9% availability.
   - Large Product: 99.99% availability.

- **Qualitative Criteria:**

1. **Criticality to Business:** Assess the product is importance to the core business functions. For example:
   - Small Product: Non-core, auxiliary service.
   - Large Product: Core, revenue-generating service.
2. **Complexity:** Evaluate the architectural complexity of the product. Consider the number of integrations, third-party dependencies, and the product's internal complexity.
3. **Change Frequency:** Assess how frequently changes, updates or deployments occur. Frequent changes may indicate a larger product.
4. **Resource Utilization:** Consider resource consumption, including CPU, RAM, storage, and network bandwidth.
5. **Geographic Reach:** Evaluate the geographic distribution of users or customers. A global user base might indicate a larger product.
6. **Regulatory Requirements:** Examine the level of compliance and regulatory obligations the product must meet. Highly regulated products may be considered more significant due to compliance efforts.
7. **Customer Impact:** Assess the potential impact on customers or users in case of incidents or downtime.
8. **SLOs and Error Budgets:** Consider the stringency of Service Level Objectives (SLOs) and the size of the error budgets assigned to the product.