Xebia

Site Reliability Engineering Fundamentals (1/2)

MAERSK

Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2

- Error Budgets
- Monitoring distributed systems
- Toil

Part 3

- Incident Management
- Postmortems

Part 4

- SRE Culture
- Onboarding SRE in an Organization
- Ask me anything & Reflection

© Xebia IT Architects - 2022

Ground rules

- All participants’ inputs are equally valued
- Only one conversation will go on at once (unless subgroups are working on a topic)
- Respect other opinions
- Turn on your camera as much as possible!
- The more effort you put in the training, the more insights you will get out of the training
- Be on time – if you can’t join let me know
- Have fun!

© Xebia IT Architects - 2022

Agenda

Part 1
- **Introduction to SRE**
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2
- Error Budgets
- Monitoring distributed systems
- Toil

Part 3
- Incident Management
- Postmortems

Part 4
- SRE Culture
- Onboarding SRE in an Organization
- Ask me anything & Reflection

© Xebia IT Architects - 2022

EXERCISE

Please go to the Miro board.

Take 10 mins to define what is SRE?

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

“Historically, companies have employed Systems Administrators to run complex computing systems”

Google – Site Reliability Engineering

© Xebia IT Architects - 2022

The Systems Administration approach to Service Management

- Running the service
- Often “operation” teams
- Responding to events
- On-call
- Requires a different skillset than software development

© Xebia IT Architects - 2022

Advantages and Disadvantages of the SysAdmin Approach

Advantages

- Easy to implement
- Tools and software widely available
- Clear roles and separation of concern

Disadvantages

- The team grows as the system grows
- Gap between Development & Operations
- They often require a different skillset & tools than developers
- Developers don't feel the pain of low performing code – risk on quality

© Xebia IT Architects - 2022

The Wall of Confusion

© Xebia IT Architects - 2022

The history of SRE

© Xebia IT Architects - 2022

The history of SRE

- Founded in 2003 by Ben Treynor
- The way Google runs operations
- Bridges the gap between software development and operations

"Fundamentally, it's what happens when you ask a software engineer to design an operations function."

Ben Treynor – VP of engineering at Google

© Xebia IT Architects - 2022

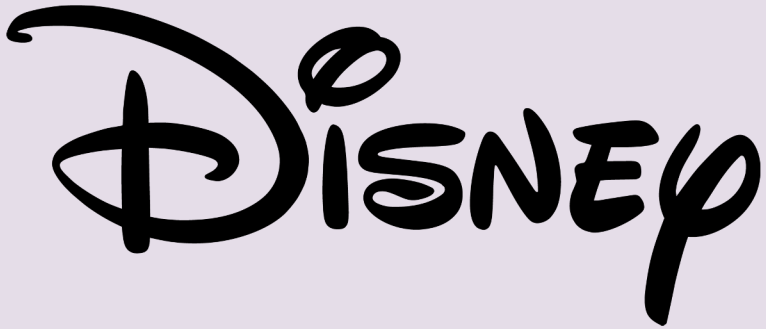
How does SRE help teams and organizations?

- Increases system reliability through applicable principles and practices
- Improves collaboration between development and operations teams
- Creates an environment of continuous improvement
- Enables an effective incidents management process
- Enables teams to operate systems at a massive scale

© Xebia IT Architects - 2022

It’s not just Google

NETFLIX



© Xebia IT Architects - 2022

You are not Google...

© Xebia IT Architects - 2022

Successful SRE in your organization

- Impacts your whole organization
- Changes the way you think about:
  - Product Development
  - Product Management
- Implements principles and practices in the aspects of:
  - Culture
  - Organization
  - Technology
- Implements data driven decision making

© Xebia IT Architects - 2022

Agenda

Part 1
- Introduction to SRE
- **The principles of SRE**
- SLIs, SLOs, SLAs

Part 2
- Error Budgets
- Monitoring distributed systems
- Toil

Part 3
- Incident Management
- Postmortems

Part 4
- SRE Culture
- Onboarding SRE in an Organization
- SRE Game
- Ask me anything & Reflection

© Xebia IT Architects - 2022

EXERCISE

Please go to the Miro board.

Take 10 mins to identify the principles of SRE.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Principles of Site Reliability Engineering

1. Embrace Risk
2. Define Service Level Objectives
3. Eliminate Toil
4. Monitor Distributed Systems
5. Embrace Release Engineering
6. Strive for Simplicity

© Xebia IT Architects - 2022

What is the difference between DevOps and SRE?

© Xebia IT Architects - 2022

EXERCISE

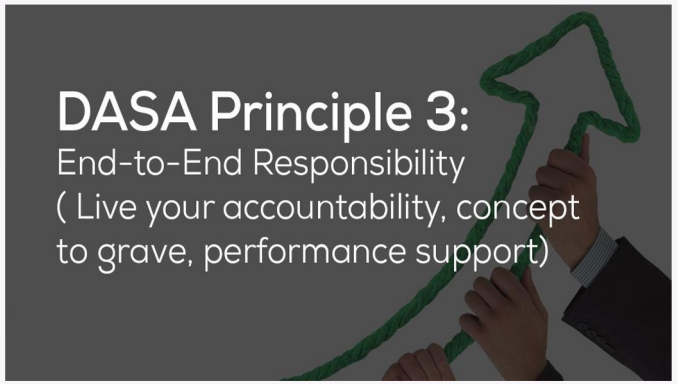
Please go to the Miro board.

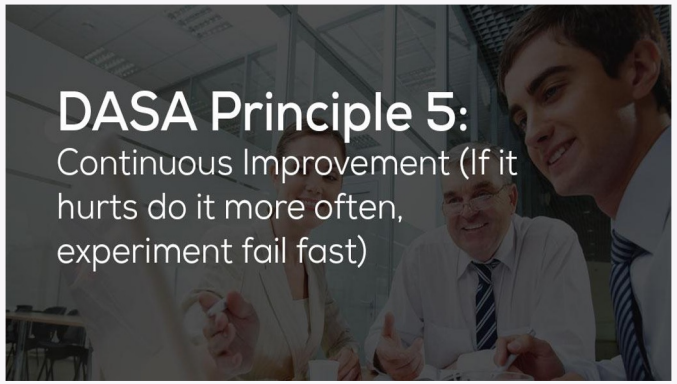
Take 7 mins to define DevOps.

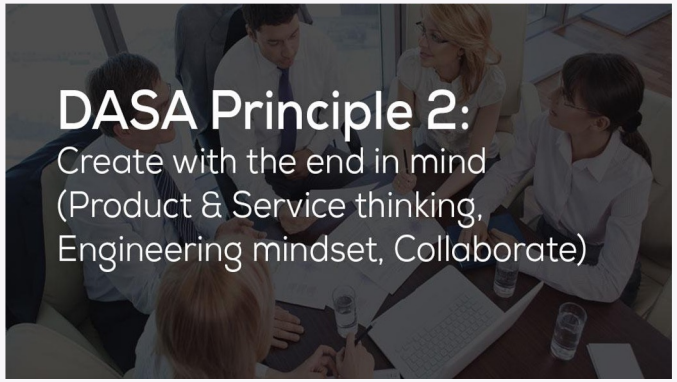
Place stickies on the board and prepare a summary for in the main room.

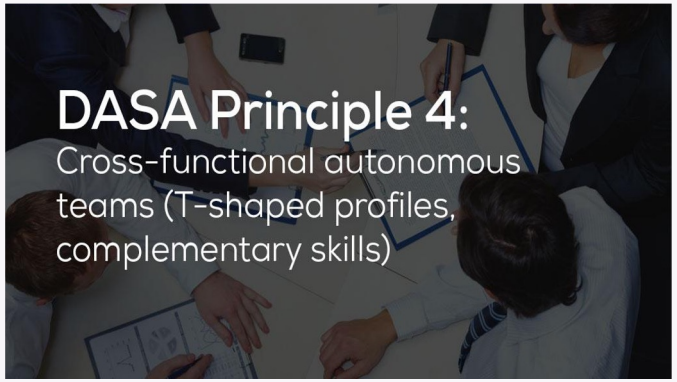
Discuss the result with the group.

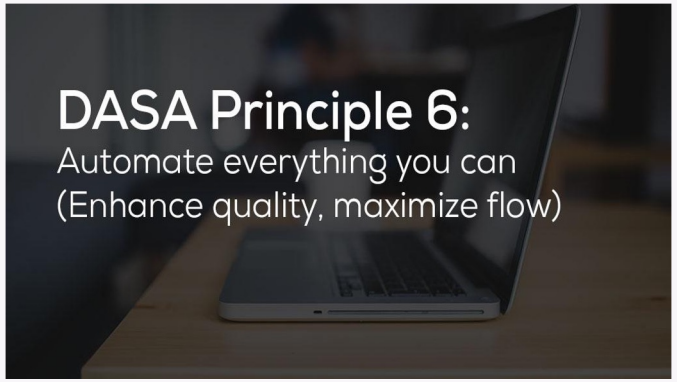
DASA DevOps











Source: DASA. For more information on DASA: https://www.devopsagileskills.org/

How does SRE relate to DevOps?

What is reliabilty?

An example on reliability and availability

- It takes a 30 min startup-time to produce quality paper
- Consider a downtime of 1 minute / hour
- The machine is:
  - Operational for 59/60 minutes
  - Produces the desired quality of paper for 29/60 minutes

© Xebia IT Architects - 2022

Reliability is Service Oriented

- Delivery to (internal) customers
- Service reliability is relevant to stakeholders
- Difference between service reliability and system availability:
  - Systems = part of greater whole & delivers a service
  - Low system availability = low service reliability, high availability ≠ high reliability
  - Customers care about high reliability = serves their purpose

© Xebia IT Architects - 2022

Expectations about reliability can be Quantified and Measured

- Expectations = Service Level Objectives
- Measurements = Service Level Indicators

© Xebia IT Architects - 2022

Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- **SLIs, SLOs, SLAs**

Part 2

- Error Budgets
- Monitoring distributed systems
- Toil

Part 3

- Incident Management
- Postmortems

Part 4

- SRE Culture
- Onboarding SRE in an Organization
- SRE Game
- Ask me anything & Reflection

© Xebia IT Architects - 2022

SLIs drive SLOs which inform SLAs

SLI – Service Level Indicator: How good is your service performing?

SLO – Service Level Objective: How good must your service perform over time?

SLA – Service Level Agreements: What have you promised your service consumers?

© Xebia IT Architects - 2022

Source: https://landing.google.com/sre/resources/practicesandprocesses/art-of-slos/

Service Level Indicators

© Xebia IT Architects - 2022
© Xebia IT Architects - 2021

The elements of a Service Level Indicator

- Type of indication: What is the indication? latency, error rate, traffic, availability?
- Service / endpoint: What service are you monitoring?
- Point of Measurement: What do you take as the source of information?
- Calculation: How do you measure your indication?

© Xebia IT Architects - 2022

Types of Service Level Indicators

- **Availability**
- **Latency**
- Quality
- Freshness
- Coverage
- Correctness

© Xebia IT Architects - 2022

Types of SLI - Availability

- Time a system is serving interactive request
- System not responding = not working properly

© Xebia IT Architects - 2022

Types of SLI - Latency

- Time it takes to serve a request initiated by user
- Loading time = system not interactive and waiting users

© Xebia IT Architects - 2022

Types of SLI

- Quality
  - Service able to run without full functionality (degraded performance)
- Freshness
  - Is data still relevant
- Coverage
  - Expect data available & in time?
- Correctness
  - Data correct?

© Xebia IT Architects - 2022

Example: Defining a Service Level Indicator

"SLI's tell us how our services or systems are preforming. If not implemented correctly, we could end up thinking that our systems are healthy while under the hood everything is on fire..."

Example of a bad SLI to measure availability:
Number of successful web-requests on the server.

Example of a good SLI to measure availability:
Proportion of HTTP GET requests (for /service/{id}/content) that have a successful (2xx status-code)
divided by all the HTTP GET requests measured from the load balancer

© Xebia IT Architects - 2022

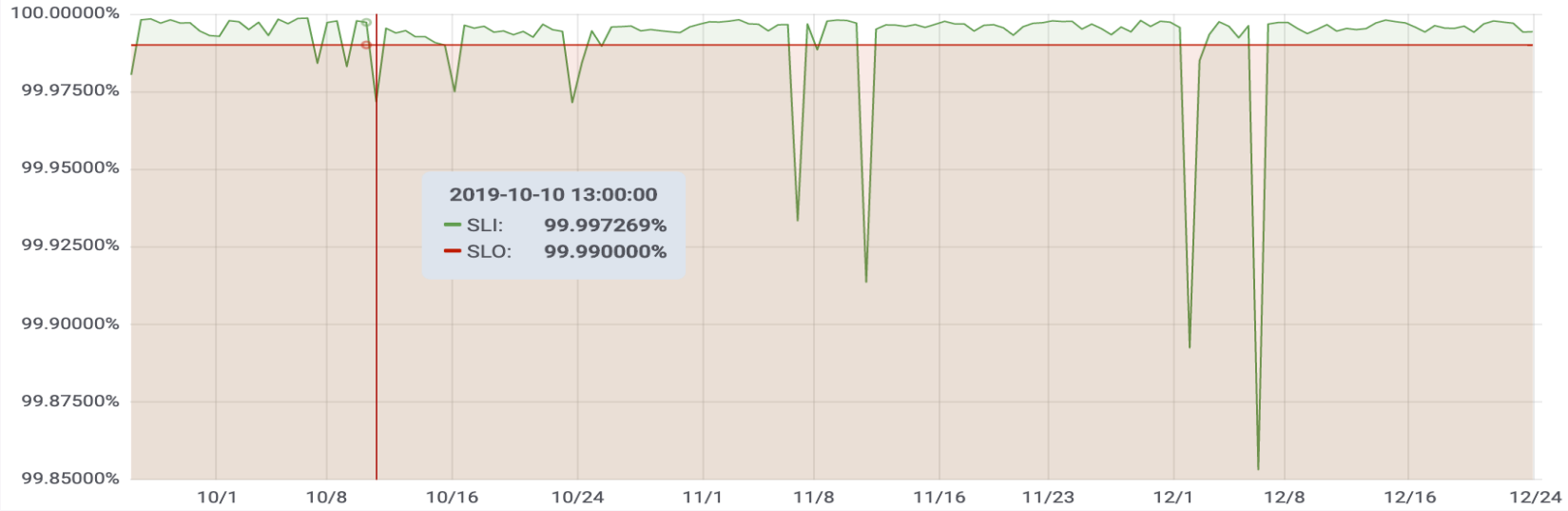
Service Level Objectives

© Xebia IT Architects - 2022

Example – Defining Service Level Objectives

SLOs contain agreements about an expected service level target value for a SLI over an x period.

A very simple example based on the SLI in the previous slide would be:
SLO: Over a monthly period for 24/7 the LB_web_requests_status_2xx > 99,99% 99 %



Service Level Agreements

© Xebia IT Architects - 2022

Service Level Agreements

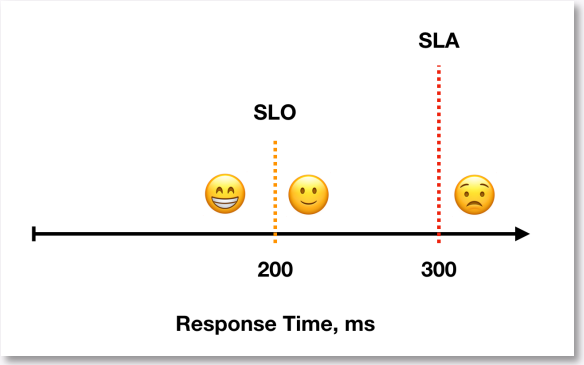
- Customer contracts
- What happens when you do not deliver the promised service?
- Should all SLO’s be in the SLA?

© Xebia IT Architects - 2022

SLA – Service Level Agreements

What are the differences between SLO's and SLA's and how to recognize them?

- SLA's contain concrete actions when agreements regarding reliability targets are not met.
  - Penalties: Refund / Rebate
  - Additional subscription time



Company	SLA	Penalty
Amazon WS	Service Uptime Less than 95.0%	100% Refund
GCP DataProc	Service Uptime Less than 95.0%	50% Refund
Azure DevOps	Service Uptime Less than 99.0%	25% Refund
AliBaba ECS	Service Uptime Less than 95.0%	100% Refund

Source: https://landing.google.com/sre/

© Xebia IT Architects - 2022

The case: amazingsportshop.com

© Xebia IT Architects - 2022

EXERCISE

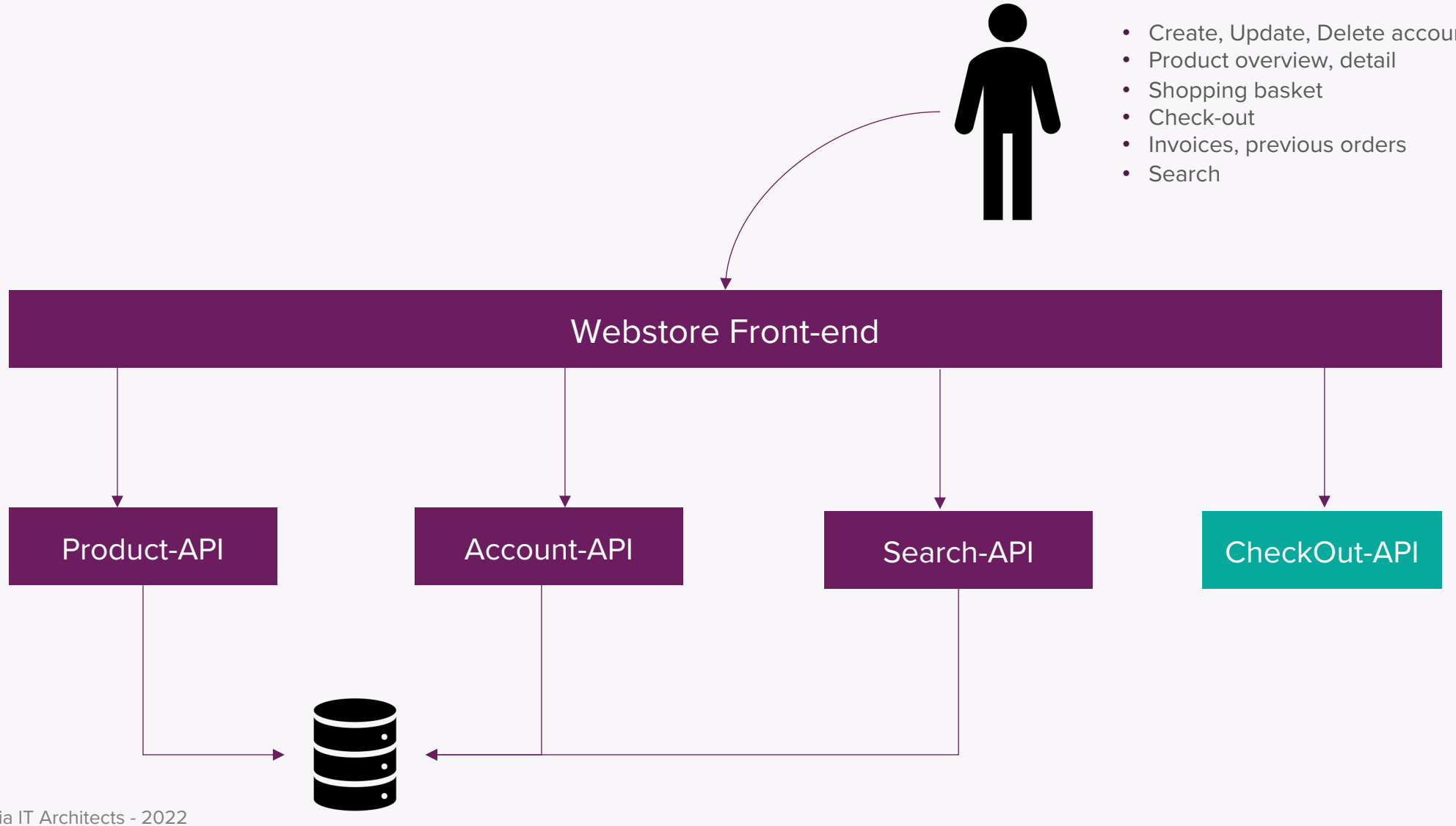
Please go to the Miro board.

Take 15 mins to define SLOs and SLIs.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Architecture - Amazingsportshop.com



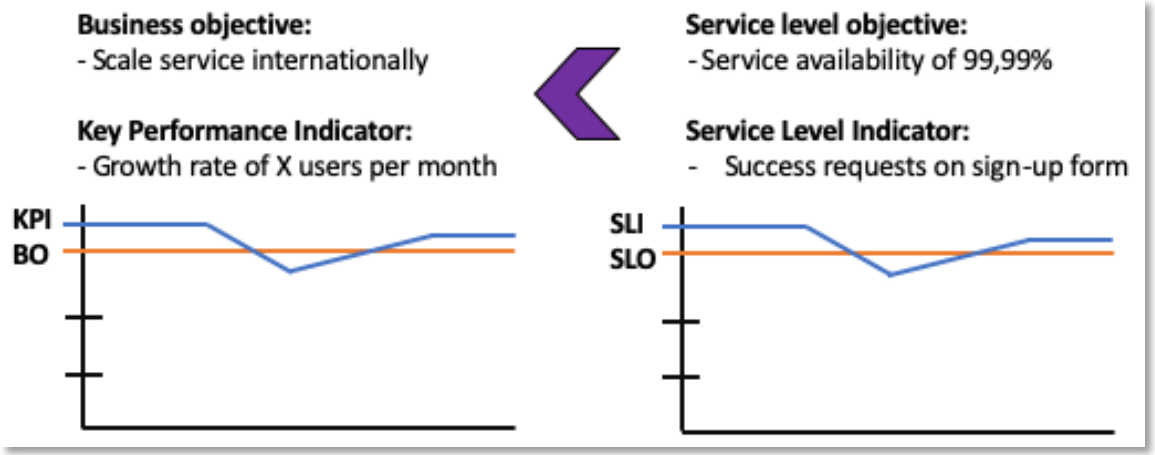
- Create, Update, Delete account
- Product overview, detail
- Shopping basket
- Check-out
- Invoices, previous orders
- Search

© Xebia IT Architects - 2022

SLAs, SLOs and SLIs – Key takeaways

- Clear SLOs help you to build the right stuff, measure what is important and improve what is important
- SLIs and SLOs should be defined around system capabilities
- SLOs should tie explicitly back to the business objectives

If done right, impacts on your SLI's should reflect the impacts on your KPI's for your business objectives.



Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2

- **Error Budgets**
- Monitoring distributed systems
- Toil

Part 3

- Incident Management
- Postmortems

Part 4

- SRE Culture
- Onboarding SRE in an Organization
- SRE Game
- Ask me anything & Reflection

© Xebia IT Architects - 2022

Once the SLO’s are set we can start using them for smart descision making

© Xebia IT Architects - 2022

The SLO is a target percentage and the **Error Budget is 100% minus the SLO**

For example:
Service receives 3 million requests over a four-week period
The Success ratio SLO is set to 99.9%
The Error Budget then is 3,000 (0.1%) errors for that time period
If a single outage is responsible for 1,500 errors, then the outage costs 50% of the error budget

© Xebia IT Architects - 2022

Error budgets can help us break the wall of confusion

EXERCISE

Please go to the Miro board.

Take 15 mins to identify the benefits & challenges of error budgets.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Agenda

Part 1
- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2
- Error Budgets
- **Monitoring distributed systems**
- Toil

Part 3
- Incident Management
- Postmortems

Part 4
- SRE Culture
- Onboarding SRE in an Organization
- SRE Game
- Ask me anything & Reflection

© Xebia IT Architects - 2022

Why monitoring

- Building dashboards
- Ad-hoc retrospective analysis (debugging)
- Analysis of long-term trends
- Alerting and paging

Api_res	Api_in	Api_out	Api_acc	Api_jdo	Api_tra
Api_fnc	Api_trd	Api_mem	Api_cpu	Api_dsk	Api_web
Api_mrk	Api_clp	Api_blk	Api_hep	Api_sys	Api_nds
Api_srv	Api_epo	Api_crt	Api_dlr	Api_elb	Api_dod

© Xebia IT Architects - 2022

Expectations for monitoring

- Keep it simple
- Create understandable Alerts
- Low noise and high signal
- Re-think monitoring rules
- Prioritize monitoring work for your services

Monitoring should be a mindset not a job!

"It's important that monitoring systems — especially the critical path from the onset of a production problem, through a page to a human, through basic triage and deep debugging—be kept simple and comprehensible by everyone on the team."

Rob Ewaschuk, Google SRE

© Xebia IT Architects - 2022

The four golden signals for Websites

- Latency
  - The time it takes to for a service to respond to a request
- Traffic
  - The amount of activity in your application
  - what number of requests/connections/bandwidth
- Errors
  - Rate of failing requests
- Saturation
  - The usage level of your service
  - CPU and Memory, Disk I/O rates

© Xebia IT Architects - 2022

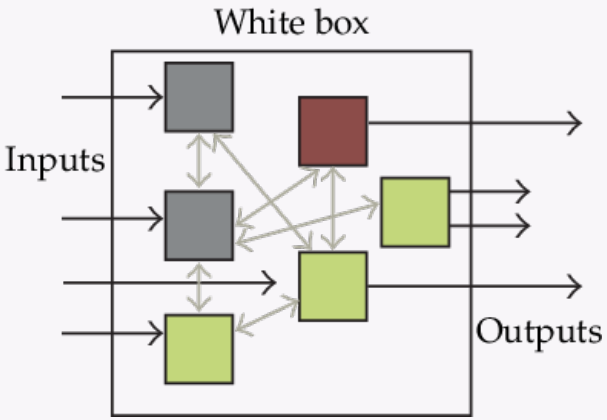
Whitebox versus Blackbox monitoring

© Xebia IT Architects - 2022

USE – RED Method

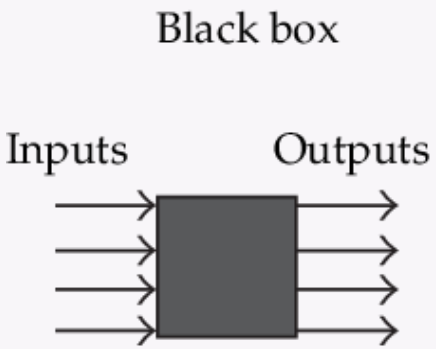
Whitebox - you know the internal workings of the system
You can measure:

- USE - Utilization, Saturation & Error rate



Blackbox - you only know what goes in and comes out of the system
You can measure:

- RED – Request throughput, Error rate & Duration (latency)



Simplicity in monitoring

- Simple, Predictable, Reliable
- Data collection, aggregation, and alerting configuration not used? **Remove!**
- Collected Signals not exposed nor used for alerts? **Remove!**

© Xebia IT Architects - 2022

Observability

© Xebia IT Architects - 2022

What is observability?

- Understanding the System
- Faster issue detection
- Improve service quality (QOS)
- Transparency
- Proactive instead of Reactive

© Xebia IT Architects - 2022

Symptoms vs. Causes

- Traditional monitoring → Symptoms
- Observability → Symptoms and Causes
- What vs Why

  Maximum signal and minimum noise

Symptom	Cause
System is serving **HTTP 500s**	Backend component is refusing connections
Responses are slow	CPU's are overloaded
Users in country x do not see the website logo	CDN (accidently) blocked some IP-ranges

© Xebia IT Architects - 2022

Pillars of Observability

- Logs

  A log is a text line that describes the event that happened at a certain time

- Metrics

  A metric is a numeric value measured over a period of time

- Traces

  A trace is a span that represent an execution of code

© Xebia IT Architects - 2022

Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2

- Error Budgets
- Monitoring distributed systems
- **Toil**

Part 3

- Incident Management
- Postmortems

Part 4

- SRE Culture
- Onboarding SRE in an Organization
- SRE Game
- Ask me anything & Reflection

© Xebia IT Architects - 2022

“In SRE, we want to spend time on long-term engineering project work instead of operational work. Because the term operational work may be misinterpreted, we use a specific word: toil.”

Vivek Rau - Google

What is E in SRE?

Engineering is

- Software development
- Systems engineering and operations
- Overhead
- Toil

© Xebia IT Architects - 2022

What is Toil?

- Manual
- Repetitive
- Automatable
- Tactical
- Has no enduring value
- Grows linearly with the service

© Xebia IT Architects - 2022

EXERCISE

Please go to the Miro board.

Take 10 mins to write down examples for different types of toil.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Is toil always a bad thing?

© Xebia IT Architects - 2022

Toil is not always bad

- Meetings are Toil
- Bug fixing
- Not (cost) effective to automate
- Training juniors

How much time should you spent on toil?

© Xebia IT Architects - 2022

Disadvantages of toil

- Career stagnation
- Low morale
- Creates confusion
- Slows progress
- Sets precedent
- Promotes employee attrition
- Causes breach of faith

© Xebia IT Architects - 2022

Where is the toil?

© Xebia IT Architects - 2022

Most reported reasons of toil

Top three sources of toil

1. Interrupts (emails and non-urgent-service-related messages)
2. On-call responses
3. Releases and pushes

© Xebia IT Architects - 2022

Toil can be measured!

- Surveys
- Project management tooling
- Hour logging tools

"The amount of time spent on toil should not exceed 50% of the available time of an engineer"

Google – Site Reliability Engineering

© Xebia IT Architects - 2022

How can we remove toil from our work?

© Xebia IT Architects - 2022

Remove toil

Automate (Consistency, Platform, Faster Action & Repair)

- Rollouts of new software versions
- (Runtime) configuration changes
- Software / hardware installation
- Turnup / turndown of services
- User account creations

© Xebia IT Architects - 2022

Color your backlog!

	Visible	Invisible
Positive Value	Feature	Architecture
Negative Value	Bug	Technical Debt

Source: Kruchten, P. 2009. What colour is your backlog?
Agile Vancouver Conference. http://pkruchten.wordpress.com/Talks.

End of Day 1