

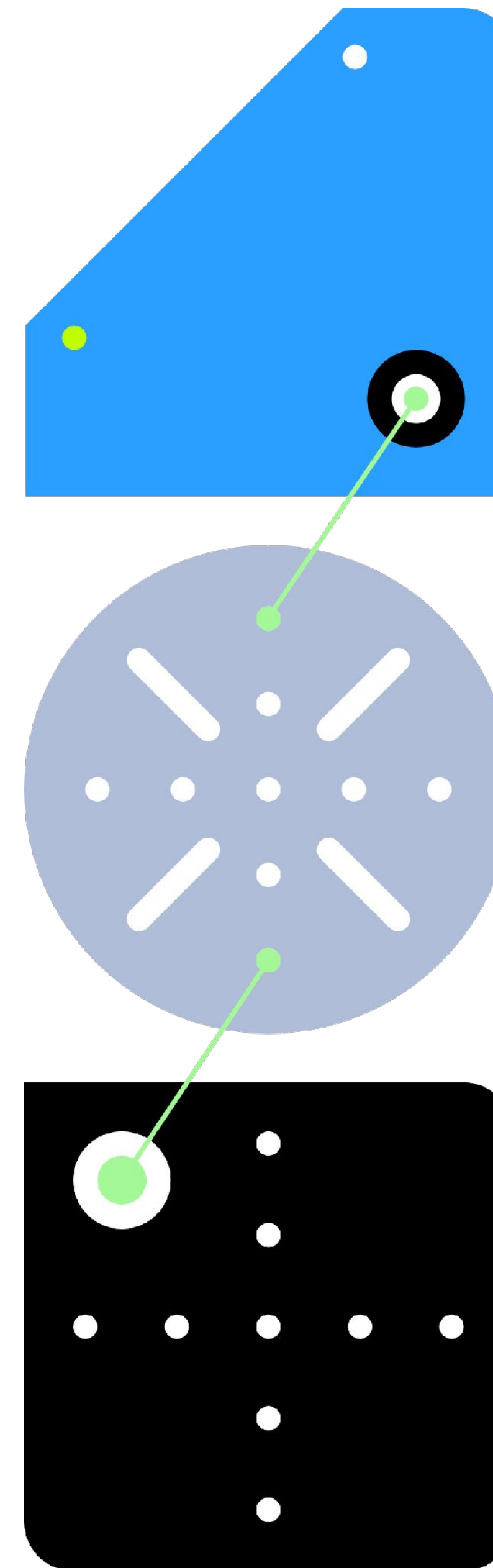
Reliability map

Куда она ведет? Как выбрать следующий шаг в надёжности.



Антон Егорушков

Руководитель группы Customer
Reliability Engineering, Yandex Cloud



Утро

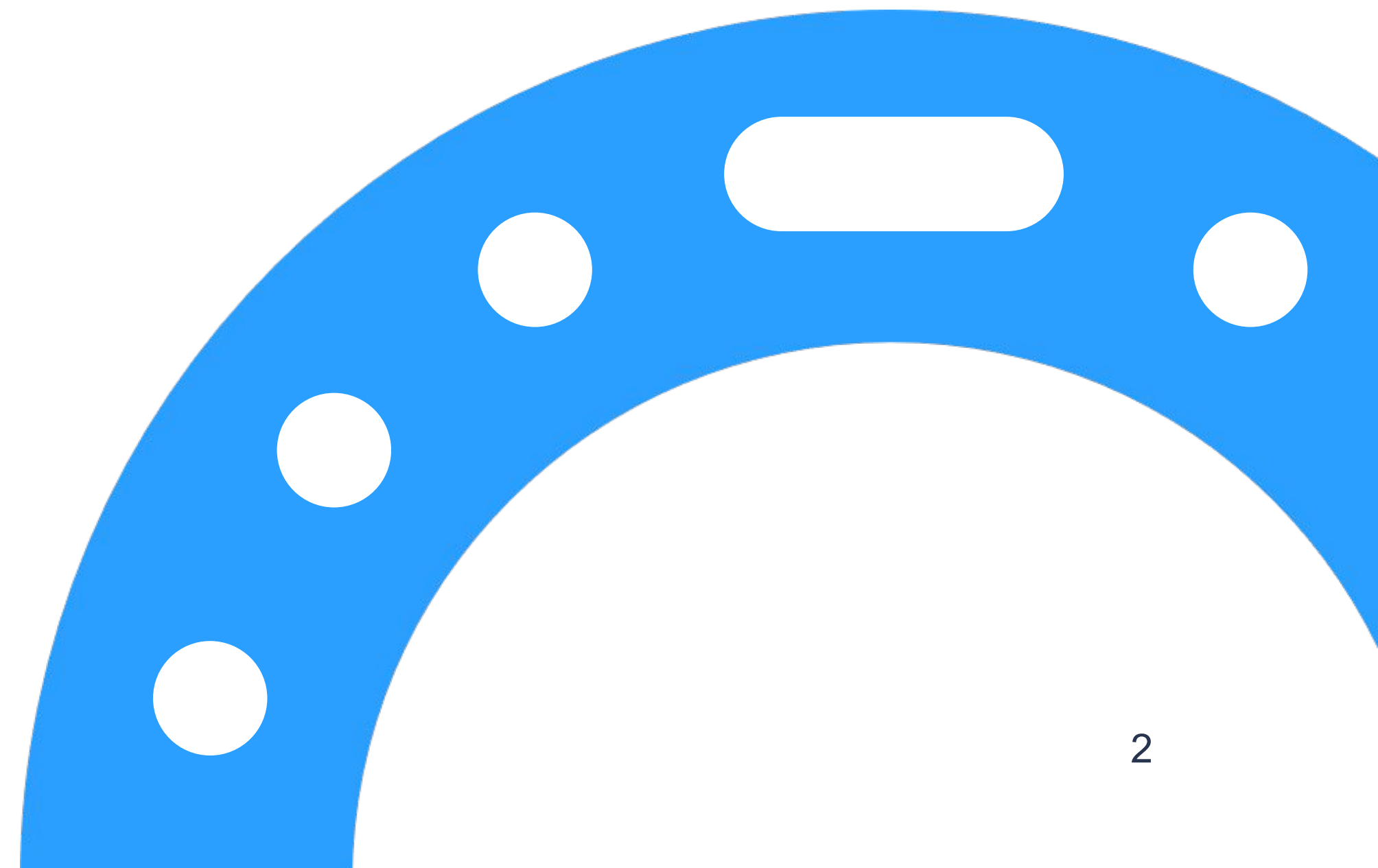
06:00 — меня позвали

INCIDENT

MAP

PATH

CRE



Утро

INCIDENT

MAP

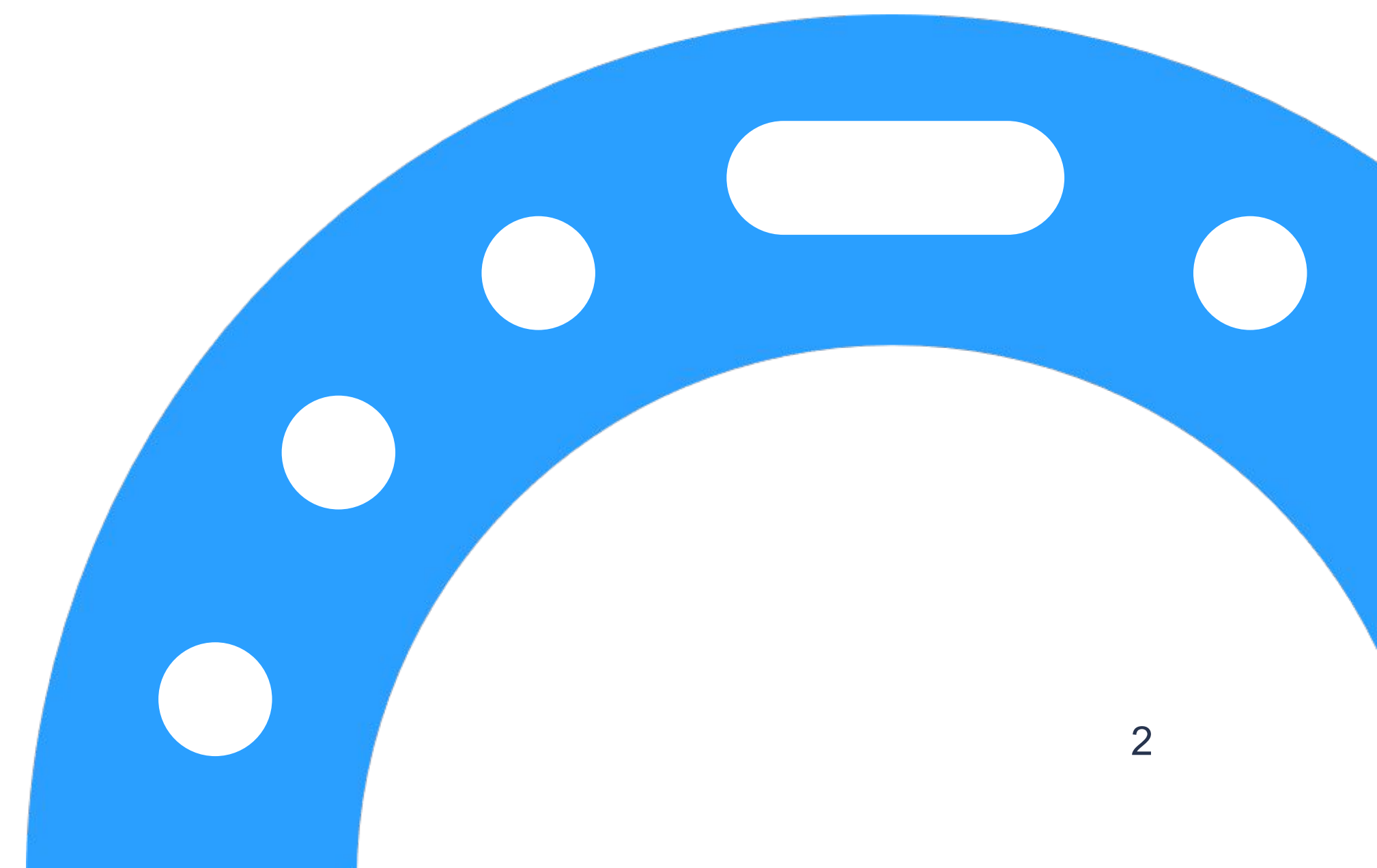
PATH

CRE

06:00 — **меня позвали**

05:53 — начало деградации → 06:00 — вызов / ask

7 минут до подключения: обнаружение (MTTD) + подтверждение (MTTA)



Утро

INCIDENT

MAP

PATH

CRE

06:00 — **меня позвали**

05:53 — начало деградации → 06:00 — вызов / ask

7 минут до подключения: обнаружение (MTTD) + подтверждение (MTTA)

Сейчас всё починю
и как раз успею
на конференцию

Утро

INCIDENT

MAP

PATH

CRE

06:00 — **меня позвали**

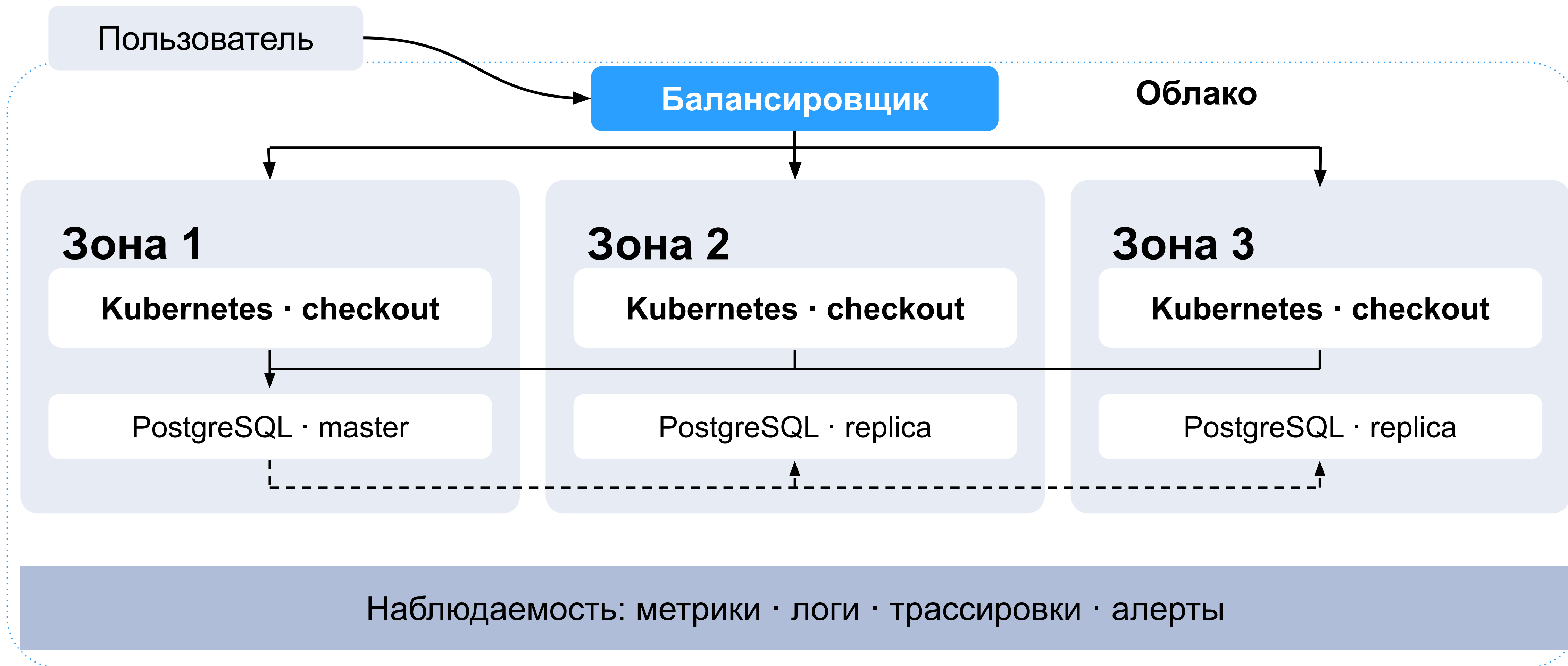
05:53 — начало деградации → 06:00 — вызов / ask

7 минут до подключения: обнаружение (MTTD) + подтверждение (MTTA)

Сейчас всё починю
и как раз успею
на конференцию



Checkout в облаке



06:03 — факты

INCIDENT

MAP

PATH

CRE

То есть всё прекрасно

Мониторинг есть

Деградация с 05:53.
Меня позвали в 06:00.
Пользователь видит
сбой раньше
инфраструктурных
алертов

06:03 — факты

INCIDENT

MAP

PATH

CRE

То есть всё прекрасно

Мониторинг есть

Деградация с 05:53.
Меня позвали в 06:00.
Пользователь видит
сбой раньше
инфраструктурных
алертов

Multi-AZ есть

Но потерю зоны
и реальный restore
никто не проверял

06:03 — факты

INCIDENT

MAP

PATH

CRE

То есть всё прекрасно

Мониторинг есть

Деградация с 05:53.
Меня позвали в 06:00.
Пользователь видит
сбой раньше
инфраструктурных
алертов

Multi-AZ есть

Но потерю зоны
и реальный restore
никто не проверял

RCA тоже есть

И всё же похожий
сюжет вернулся
через три месяца

06:03 — факты

INCIDENT

MAP

PATH

CRE

То есть всё прекрасно

Мониторинг есть

Деградация с 05:53.
Меня позвали в 06:00.
Пользователь видит
сбой раньше
инфраструктурных
алертов

Multi-AZ есть

Но потерю зоны
и реальный restore
никто не проверял

RCA тоже есть

И всё же похожий
сюжет вернулся
через три месяца

Только сервис не работает

06:03 — impact есть, root cause пока нет

INCIDENT

MAP

PATH

CRE

Сначала снимаем impact.

СИМПТОМ

Checkout errors / latency.
Это факт, не диагноз.

1

06:03 — impact есть, root cause пока нет

INCIDENT

MAP

PATH

CRE

Сначала снимаем impact.

Симптом

Checkout errors / latency.
Это факт, не диагноз.

1

Диагностика

Logs, traces, DB latency,
dependencies.
Ищем failure mode.

2

06:03 — impact есть, root cause пока нет

INCIDENT

MAP

PATH

CRE

Сначала снимаем impact.

Симптом

Checkout errors / latency.
Это факт, не диагноз.

1

Диагностика

Logs, traces, DB latency,
dependencies.
Ищем failure mode.

2

Mitigation

Цель — вернуть critical
flow.

3

06:17 → 06:22 → 06:43 · сбой и mitigation

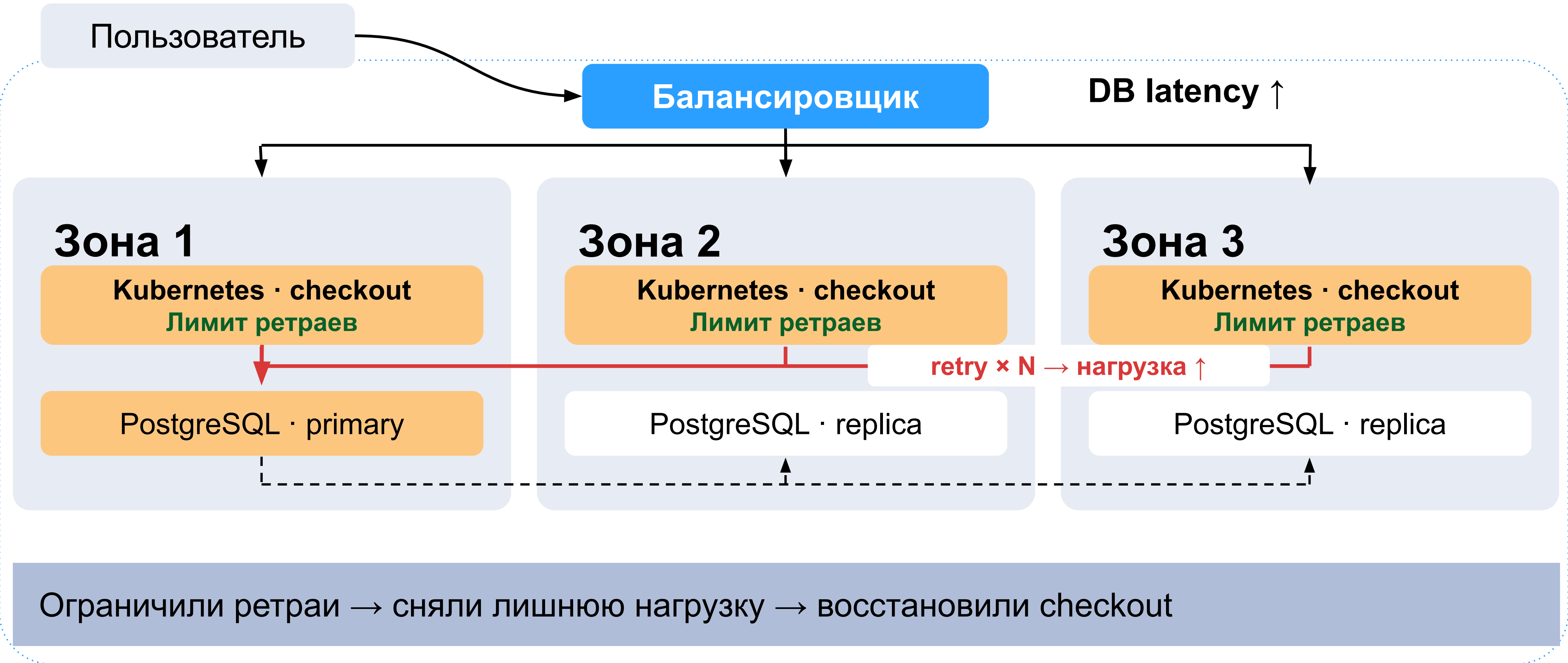
INCIDENT

MAP

PATH

CRE

БД замедлилась. Ретраи усилили отказ



06:17 — ещё немного информации

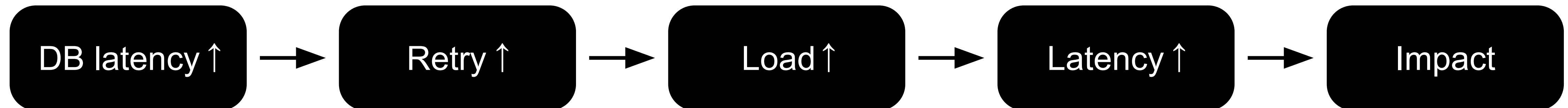
INCIDENT

MAP

PATH

CRE

Мы начали убивать вполне живую базу ретраями



06:22

Ограничиваем retry и снимаем пользовательский impact

06:43

Инцидент потушен. Но системно ничего ещё не исправлено

До mitigation: 29 мин = 7 мин до ask + 22 мин после

06:45 — mitigation закончен

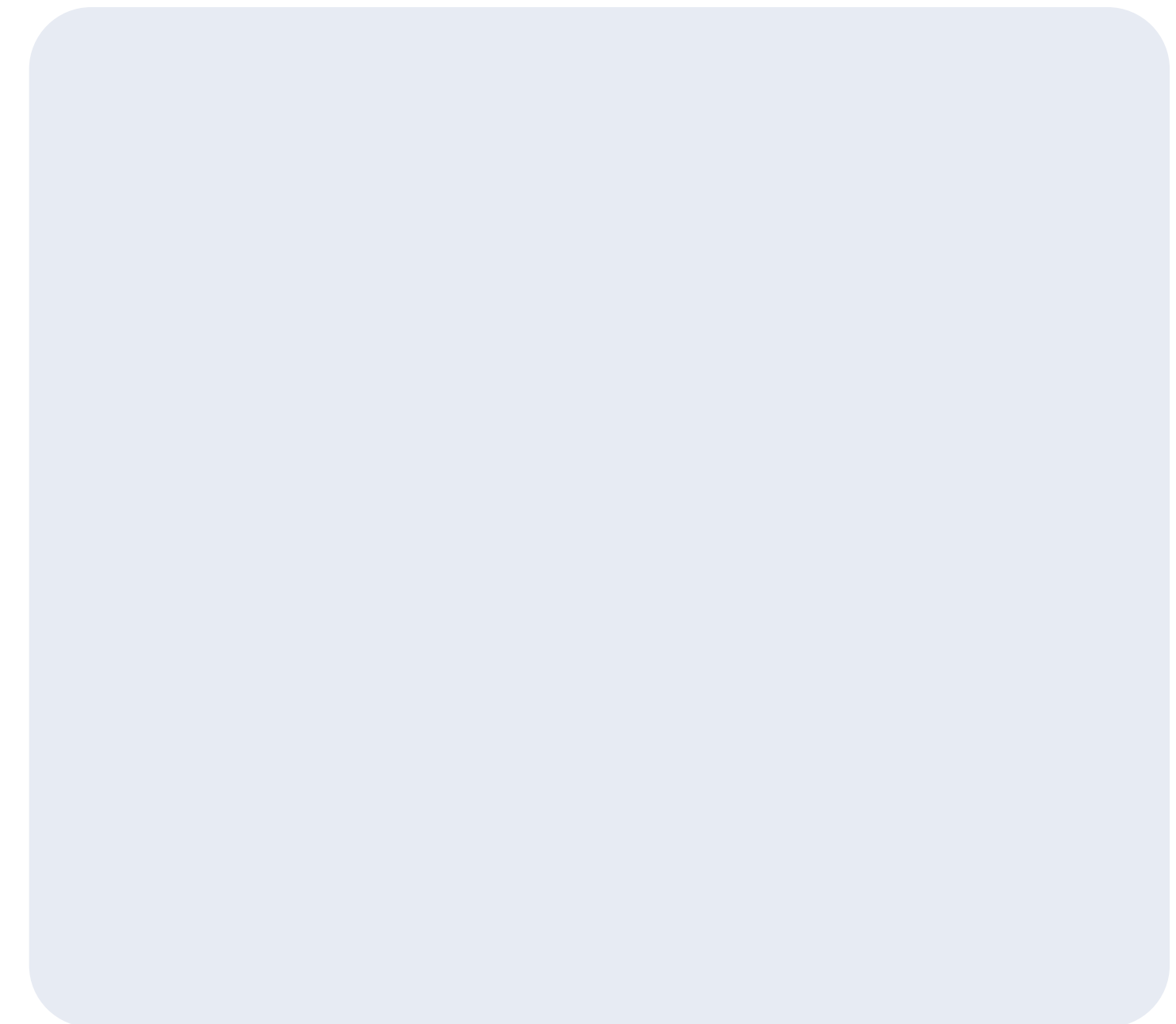
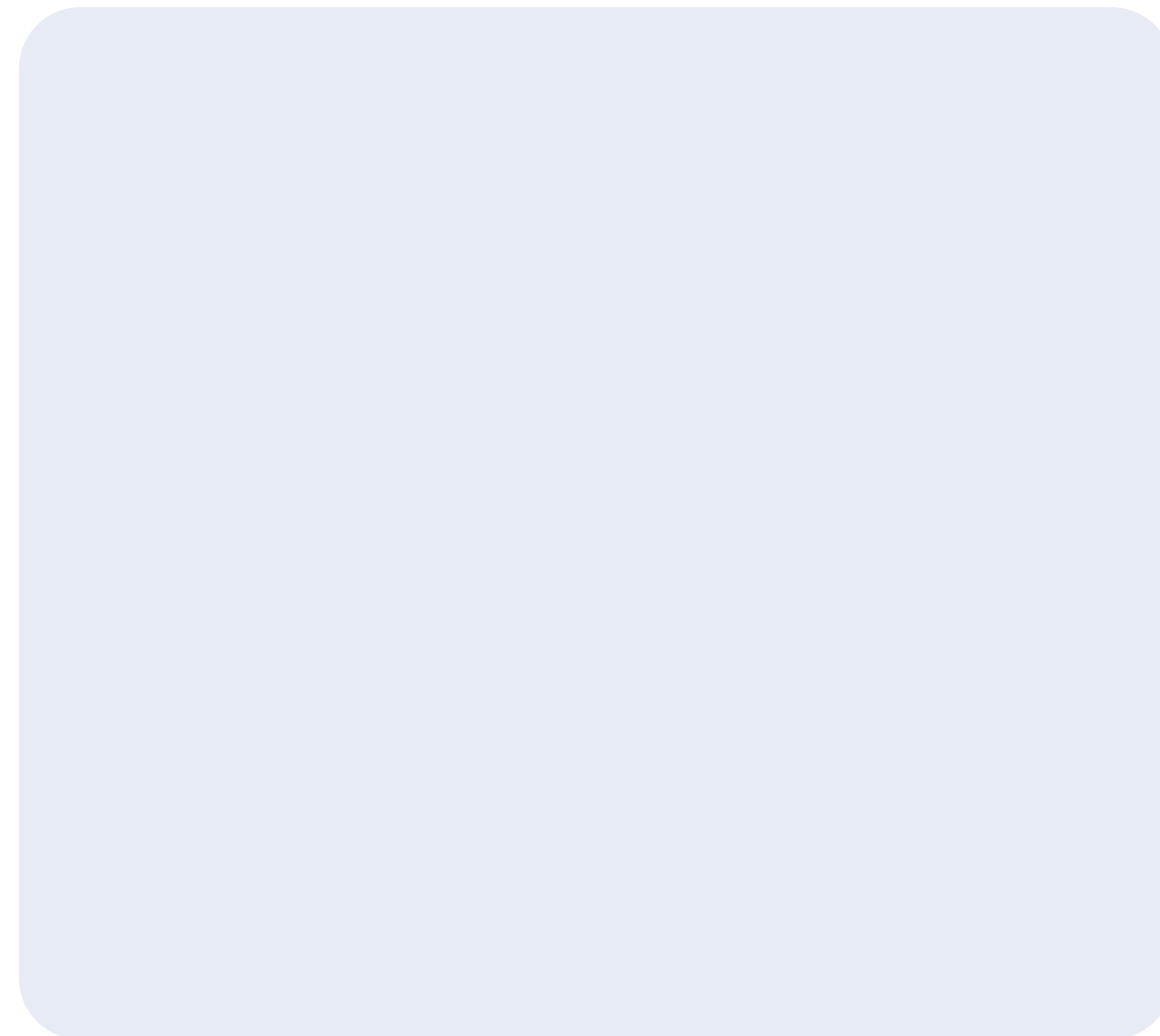
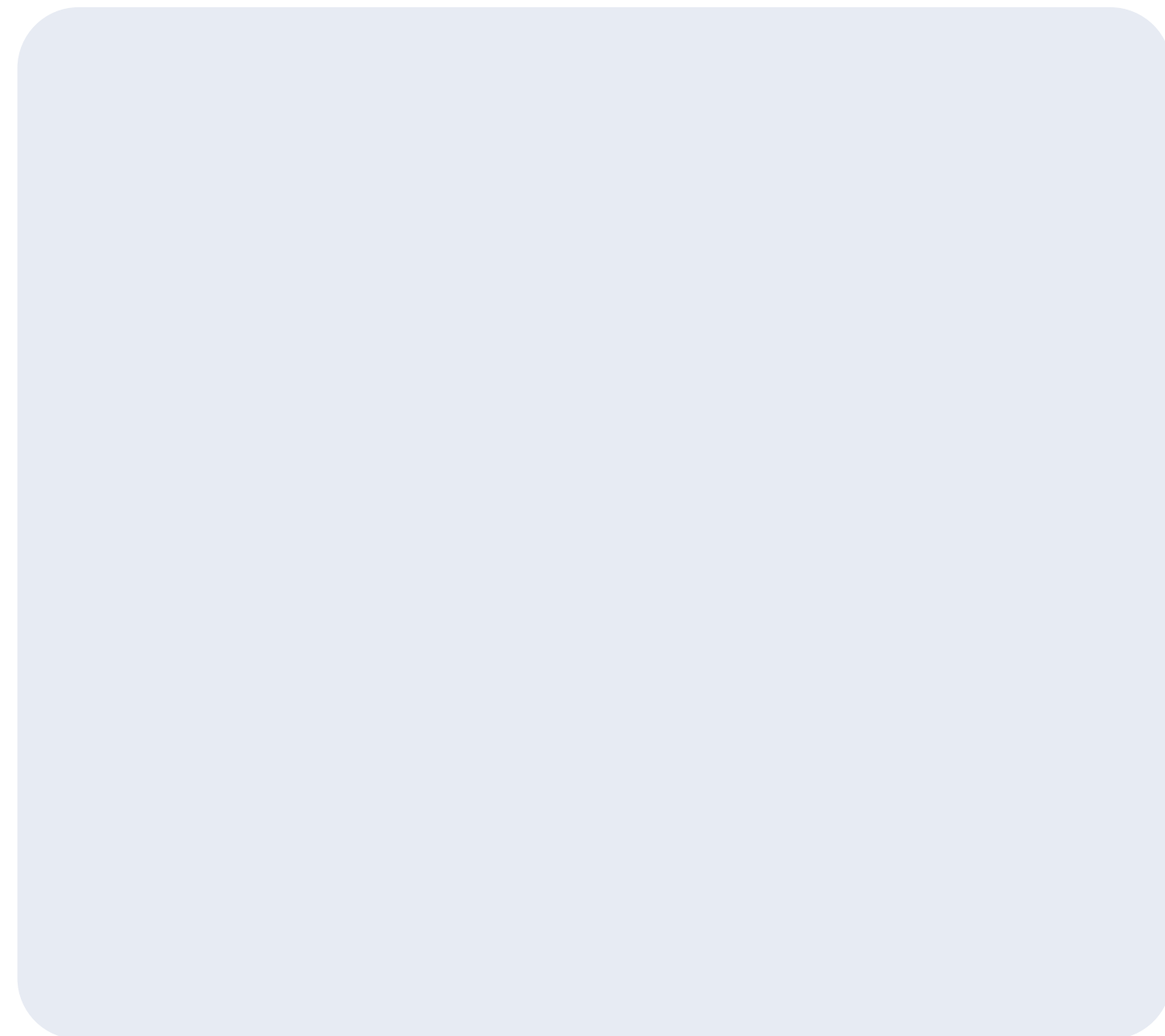
INCIDENT

MAP

PATH

CRE

У вас **один** инженер.
Что исправляем первым?



06:45 — mitigation закончен

INCIDENT

MAP

PATH

CRE

У вас **один** инженер.
Что исправляем первым?

**Видеть раньше
пользователя**

Observability →
SLI → SLO

1

06:45 — mitigation закончен

INCIDENT

MAP

PATH

CRE

У вас **один** инженер. Что исправляем первым?

**Видеть раньше
пользователя**

Observability →
SLI → SLO

1

**Доказать
отказоустойчивость**

Resilience →
recovery → DR test

2

06:45 — mitigation закончен

INCIDENT

MAP

PATH

CRE

У вас **один** инженер. Что исправляем первым?

**Видеть раньше
пользователя**

Observability →
SLI → SLO

1

**Доказать
отказоустойчивость**

Resilience →
recovery → DR test

2

**Не повторить
этот инцидент**

Incident → RCA →
problem

3

Видеть раньше пользователя

Synthetic monitoring

Service level indicators (SLI)

Service level objectives (SLO)

Видеть раньше пользователя

Synthetic monitoring

Service level indicators (SLI)

Service level objectives (SLO)



Observability → Synthetic monitoring → SLI/SLO

Сначала делаем пользовательский симптом измеримым. Затем связываем его с SLO и error budget.

Продолжить доклад →

Доказать отказоустойчивость

RPO / RTO

Recovery path

DR-тест

Доказать отказоустойчивость

RPO / RTO

Recovery path

DR-тест



Infrastructure / Operations
→ **recovery** → **DR test**

Проверяем recovery path и снимаем крупнейшую неизвестность, прежде чем добавлять ещё одну зону.

Продолжить доклад →

Не повторить ЭТОТ инцидент

RCA / 5 whys

Postmortem reviews / actions

Problem management

Не повторить ЭТОТ инцидент

RCA / 5 whys

Postmortem reviews / actions

Problem management

■ ITIL STYLE NOC	■ DR SITE EXISTS	■ MANUAL REMEDIATION PLAYBOOKS	■ FORMAL INCIDENT RESPONSE ROLES	■ FORMAL INCIDENT RESPONSE PROCESSES	■ ROLLBACKS/ROLLFORWARD TESTED
■ SCHEDULED DOWNTIME	■ BASIC INCIDENT MANAGEMENT	■ REPEATABLE DEPLOYMENTS	■ AUTOMATION OF TOIL	■ PROBLEM MANAGEMENT FUNCTION	■ DEDICATED OPERATIONS TOOLING
■ PATCHING WINDOWS	■ GOLD IMAGE AUTOMATION	■ CENTRAL CERTIFICATE ROTATION	■ BREAKGLASS SECRET ACCESS		
■ HOST PING TESTS	■ SYNTHETIC MONITORING		■ APM METRICS AND TRACES	■ INTERNAL SLAS	■ ERROR BUDGETS
■ SSH TO GREP LOGS	■ CENTRALIZED LOG COLLECTION	■ REALTIME CENTRALIZED LOG ANALYTICS	■ AUTOMATED TOPOLOGY VIEW	■ SERVICE LEVEL INDICATORS (SLI)	
			■ SERVICE LEVEL OBJECTIVES (SLO)		
■ RCA/5 WHYS	■ INCENTIVISE TRUST/SAFETY	■ UNDERSTAND BUSINESS IMPACTS	■ BLAMELESS POSTMORTEMS	■ POSTMORTEM REVIEWS/ACTIONS	■ SINGLE CENTRAL CAB
■ MEASURE EVERYTHING	■ DATA DRIVEN DECISIONS	■ SERVICE OWNERSHIP	■ INCENTIVISE CROSS SILO COLLABORATION	■ DEDICATED R9Y STAFFING	■ CHANGE FREEZES

Incident → RCA → Problem Management

Action item получает владельца и критерий готовности. Проверенный результат предотвращает повторение инцидента.

Продолжить доклад →

После mitigation — системный выбор

Мы только что выбрали путь по Reliability Map



INCIDENT

MAP

PATH

CRE

The diagram is a matrix titled "The SRE Maturity Model". It is organized into four main rows representing different dimensions of SRE: Development, Infrastructure, Operations, and SRE. Each row is further divided into four maturity levels: Demo (9.0), Well-Engaged Ops (9.99), Well-Engaged Ops (9.99), and Autonomous (9.999). The matrix cells contain various SRE practices and concepts, which are interconnected by lines, indicating a holistic and evolving approach. The practices are grouped into four main categories: Development, Infrastructure, Operations, and SRE, each with its own set of sub-categories and specific practices.

Dimension	Development	Infrastructure	Operations	SRE
Development	DEMO 9.0 LOCAL DEVELOPMENT MONOLITH SMOKE TESTS MANUAL TESTS DEPENDENCY DIFFERENTIATION	DETERMINISTIC 9.0 CODE REVIEW AUTOMATED UNIT TESTS CODE VERSION CONTROL MULTI-SERVICE DEVELOPMENT FUNCTIONAL TESTS MANUAL INTEGRATION TESTS SEMI-AUTOMATED INTEGRATION REGULAR RELEASE CALENDA	REACTIVE 9.9 - WELL-ENGAGED OPS PRE-MERGE HOOKS DISTRIBUTED SYSTEMS AWARENESS DATA VERSIONING ACTIVE-PASSIVE CLUSTERS DEPLOYMENTS IN PLACE TRAFFIC SHIFTING MICROSERVICES FEATURE FLAGS INSTRUMENTATION FOR IN-PROCESS TRACES CONTAINERS	PROACTIVE 9.99 - WELL-ENGAGED OPS LEFT-SHIFT RELIABILITY DESIGN BACKWARDS VERSION COMPATIBILITY BY DEFAULT BLUE-GREEN DEPLOYMENTS GRACEFUL SERVICE DECOMMISSION (INDIVIDUAL CLUSTERS) ACTIVE-ACTIVE MULTI-CLUSTER CANARY DEPLOYMENTS LEFT-SHIFT QA TESTING (SOFT) FUZZ TESTING LEFT-SHIFT PERFORMANCE TESTING BASIC CHAOS TESTING CQC TESTING DISTRIBUTED SYSTEMS (NO ACTIVE-MEASURE)
Infrastructure	INFRA LOCAL DATA STORAGE PET HOST SINGLE ZONE INFRA	INFRA DNS + SIMPLE LB >1 COMPUTER DISTRIBUTED STORAGE BASIC LINEAR CAPACITY PROJECTION ADVANCED LOAD-BALANCING ALTERNATE SITE REPLICATION BASIC LOAD TESTING HIGH WATER MARK PREDICTION	INFRASTRUCTURE-REACTIVE IAC CATTLE INFRASTRUCTURE MULTI-ZONE HIGH WATER MARK PREDICTION UNDERSTAND INFRASTRUCTURE SHAPES DOMAINS CONTAINER ORCHESTRATOR HOLI-WINTER CAPACITY PROJECTIONS MULTIPLE INFRASTRUCTURE	INFRASTRUCTURE-PROACTIVE AUTO FALLOVER AUTO SCALING FAILURE INJECTION ASSURED CAPACITY LOAD TESTING PRODUCTION LAUNCH AUTOMATOR FAILURE TESTING IN PROD ELIMINATE SPOFS (HARDWARE & SOFTWARE) N+1 REGIONAL PLANNING REAL-WORLD TRAFFIC LOAD TESTING N+1 AS STANDARD SERVICE DISCOVERY L7 GLOBAL LB L4 REGIONAL LOAD BALANCING
Operations	OPERATIONS OFF-HOST BACKUP MANUALLY CREATED MACHINES REPORTS DEFINED DR PLAN CUSTOM VMs VIA SEMI-AUTOMATION MANUAL REMEDIATION	OPERATIONS REPORTS REFINED DR PLAN SIMULATED/TABLETOP DR SITE EXISTS SCHEDULED DOWNTIME PATCHING WINDOWS BASIC INCIDENT MANAGEMENT GOLD IMAGE AUTOMATION REPEATABLE DEPLOYMENTS CENTRAL CERTIFICATE ROTATION	OPERATIONS-REACTIVE CONTINUOUS INTEGRATION FORMAL INCIDENT RESPONSE RULES AUTOMATION OF TOOL RIPPAKAR'S SECRET ACCESS CONTINUOUS DELIVERY FORMAL INCIDENT RESPONSE PROCESSES PROBLEM MANAGEMENT FUNCTION REGULAR BOP TESTING (BURN, PROX, ALTERNATE SITE) HOLBACKS/HOLLO-COW-TESTED DEDICATED OPERATIONS TOOLING	OPERATIONS-PROACTIVE % BASED TRAFFIC STEERING CONTINUOUS DEPLOYMENT AUTOMATED SERVICE DISCOVERY GLOBAL POLICY ENFORCEMENT ACTIVE-ACTIVE DATASTORES EXTERNAL RATE LIMITING DATA COLLECTION AUTOMATION VANILLA CDPS PROTECTION INTERNAL RATE LIMITING CENTRALIZED PRODUCTION CHANGES MOSTLY AUTOMATED REMEDIATION DIRT TESTING PRODUCT-SPECIFIC CDPS PROTECTION (E.G. VMs)
SRE	HOST METRICS AND LOGGING PER-HOST ALARMS ON-HOST LOG GREP	HOST PING TESTS SSH TO GREP LOGS CENTRALIZED LOG COLLECTION REAL-TIME CENTRALIZED LOG ANALYTICS	APM METRICS AND TRACES AUTOMATED TOPOLOGY VIEW INTERNAL SLAs SERVICE-LEVEL INDICATORS (SLI) SERVICE-LEVEL OBJECTIVES (SLO) ERROR BUDGETS	CUSTOM IN-PROCESS TRACING RECORD AND REPLAY TRAFFIC EVENT CORRELATION CROSS-SERVICE TRANSACTION TESTING ADVANCED VISUALIZATIONS (HEATMAPS, FLAMEGRAPHS) MULTI-MACHINE DEBUGGING, HOTSPOT LFC
Product	HIGH CONTEXT BEHAVIOURS MANAGING PET CONFIGURATION DRIFT TODO LISTS	INCENTIVISE TRUST/SAFETY DATA-DRIVEN DECISIONS SMART GOALS UNDERSTAND BUSINESS IMPACTS SERVICE OWNERSHIP YOBIBALL PROJECTS/PW	BLAMELESS POSTMORTEMS POSTMORTEM REVIEW SESSIONS DEDICATED RBY STAFFING SCALES > OBJECTIVES (OKRS) ARCHITECTURE REVIEWS SINGLE CENTRAL CAB CHANGE FREEZES	HOLISTIC VIEW OF RBY AS HIGH VALUE VERTICAL SCALE IS AN ANTI-PATTERN HIGH-PERFORMING STAFF PROMOTION AND HIRING INTRODUCING DEPLOYMENT SPIES SELF-DRIVEN QUICK-TEST LAUNCHER RELIABILITY EXECUTIVE SPONSOR EXISTS SRE-SWE ROLES INTRODUCED REACTION RISK ANALYSIS TOOL BUDGETS DECREASED RELIANCE ON RBY PARTY GUNS

2 · Переживать

Recovery
HA / DR-тесты

3 · УЧИТЬСЯ

Incident / RCA

Problem management

Карта задаёт координаты
следующего шага.

Надёжность растёт не от количества практик

INCIDENT

MAP

PATH

CRE

Она растёт от правильно выбранного
следующего шага

Надёжность растёт не от количества практик

INCIDENT

MAP

PATH

CRE

А теперь можно познакомиться

INCIDENT

MAP

PATH

CRE

Антон — Customer Reliability Engineering TeamLead

Yandex Cloud

Работаю с продакшн-сервисами клиентов.

Руковожу CRE-командой: надёжность на границе клиентской системы и облака

Зачем я здесь

Помочь выбрать практику, которая закрывает риск продакшна и помогает выполнять согласованные SLA

Одна система — четыре языка

INCIDENT

MAP

PATH

CRE

Надёжность выглядит по-разному

Что такое надёжность?

Надёжность выглядит по-разному

Бизнес

Какой customer journey нельзя потерять? Сколько стоит простой?

Разработка

Что приложение делает при timeout, retry и полной или частичной деградации?

SRE

Как измеряем? SLI/SLO, error budget, MTTD/MTTR, варианты деградаций и падений.

Провайдер

SLA компонентов, домены отказов, ограничения платформы.

Надёжность выглядит по-разному

Бизнес

Что с заказами?
Сколько заказов потеряли?
Что об этом сказал клиент?

Разработка

Какой нужен лимит ретраев?
Какой компонент сбоит?

SRE

Что с **успешностью** заказов?
Как быстро увидели и восстановили?

Провайдер

Пострадал ли SLA сервиса DB?
Повлияли ли ограничения bouncer, balancer, pooler.

r9y reliability map: где мы и что обычно появляется дальше

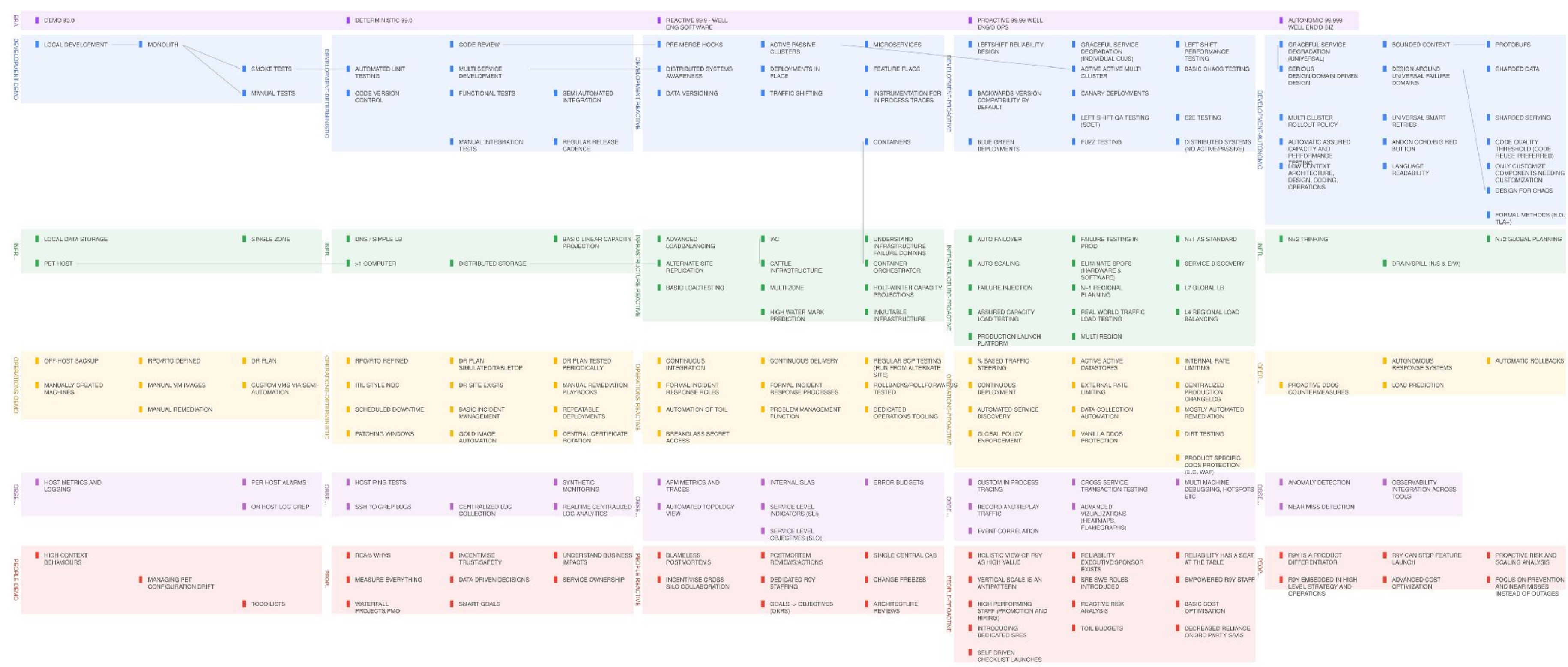
DEMO
90.0

DETERMINISTIC
99.0

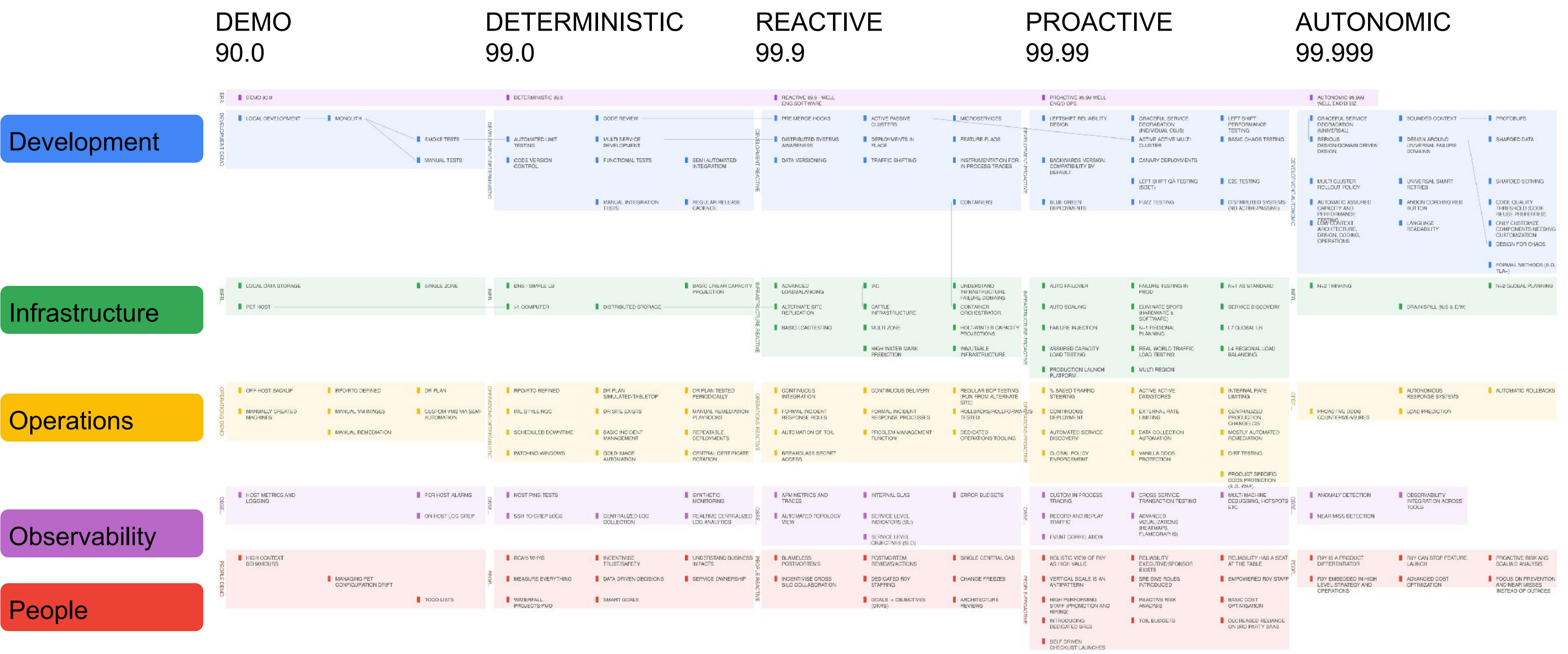
REACTIVE
99.9

PROACTIVE
99.99

AUTONOMOUS
99.999



r9y reliability map: где мы и что обычно появляется дальше



Первая проблема maturity model

INCIDENT

MAP

PATH

CRE

У реальной системы нет одного уровня зрелости

	Demo	Deterministic	Reactive	Proactive	Autonomic
Observability				✓	
Incident mgmt					
DR					
Release					
Capacity					

Первая проблема maturity model

INCIDENT

MAP

PATH

CRE

У реальной системы нет одного уровня зрелости

	Demo	Deterministic	Reactive	Proactive	Autonomic
Observability				✓	
Incident mgmt			✓		
DR					
Release					
Capacity					

Первая проблема maturity model

INCIDENT

MAP

PATH

CRE

У реальной системы нет одного уровня зрелости

	Demo	Deterministic	Reactive	Proactive	Autonomic
Observability				✓	
Incident mgmt			✓		
DR	✓				
Release					
Capacity					

Первая проблема maturity model

INCIDENT

MAP

PATH

CRE

У реальной системы нет одного уровня зрелости

	Demo	Deterministic	Reactive	Proactive	Autonomic
Observability				✓	
Incident mgmt			✓		
DR	✓				
Release			✓		
Capacity					

Первая проблема maturity model

INCIDENT

MAP

PATH

CRE

У реальной системы нет одного уровня зрелости

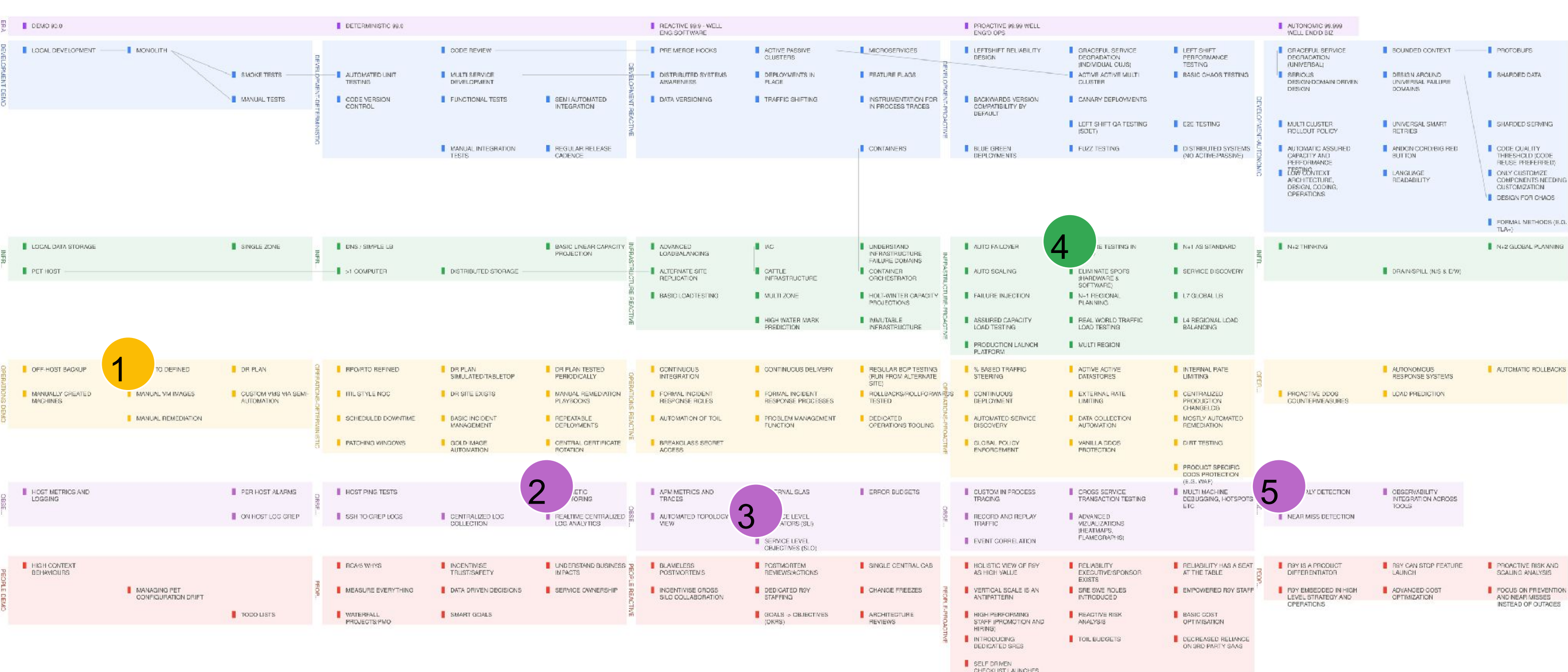
	Demo	Deterministic	Reactive	Proactive	Autonomic
Observability				✓	
Incident mgmt			✓		
DR	✓				
Release			✓		
Capacity		✓			

**Поэтому нам нужна не оценка компании «на 3 из 5».
Нам нужна карта областей и зависимостей**

Один checkout · пять точек на реальной карте

Система та же.

Мы каждый раз зумимся в другой узел r9y



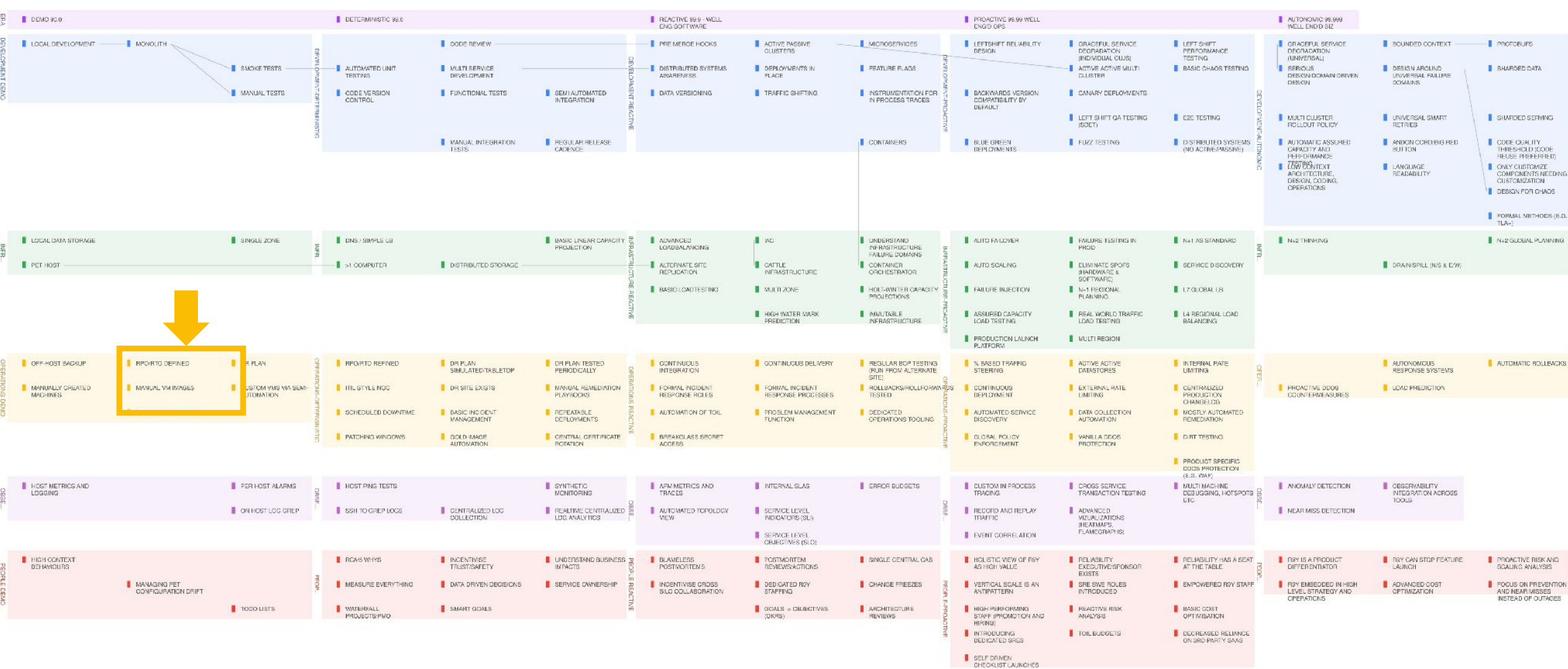
- 1 RPO/RTO defined
- 2 Synthetic monitoring
- 3 SLI / SLO / error budget
- 4 Failure testing in prod
- 5 Anomaly detection + ARS

Маршрут checkout. Порядок шагов зависит от рисков системы.

MAP → Operations / Demo

1/5 · RPO/RTO defined

Operations



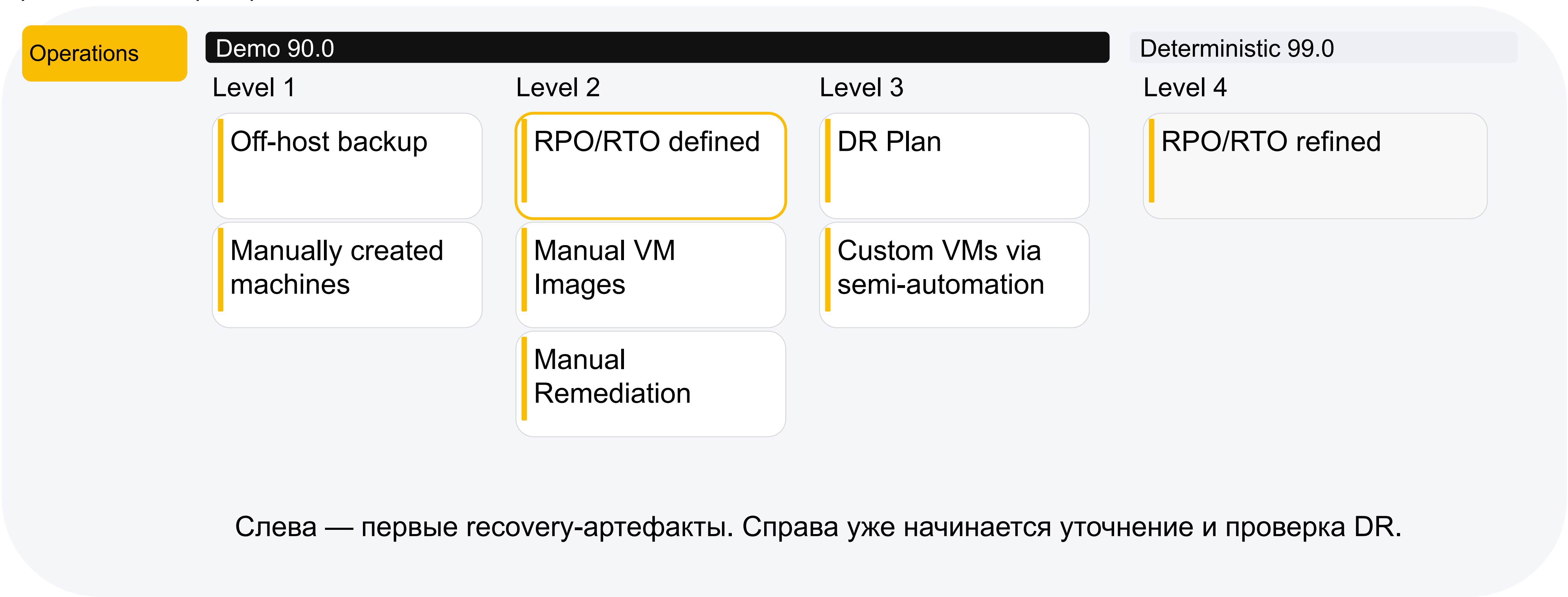
map.r9y.dev

RPO/RTO defined Demo · Level 2

Допустимый простой и потеря данных становятся явными.

RPO/RTO defined

Фрагмент live map · Operations-Demo · Level 2



1. RPO/RTO defined

Наш checkout · 05:53

Инцидент сняли лимитом ретраев.
Следующий риск — потеря данных БД.
Как восстановим тот же checkout?

Вопрос Простой checkout / потеря заказов

Артефакт RTO/RPO + тест восстановления

Checkout Ресурсы K8s + данные заказов



Velero
Restore ресурсов
checkout в Kubernetes



pgBackRest
Restore заказов
в PostgreSQL

1. RPO/RTO defined

Наш checkout · 05:53

Инцидент сняли лимитом ретраев.
Следующий риск — потеря данных БД.
Как восстановим тот же checkout?

Вопрос Простой checkout / потеря заказов

Артефакт RTO/RPO + тест восстановления

Checkout Ресурсы K8s + данные заказов



Velero
Restore ресурсов
checkout в Kubernetes



pgBackRest
Restore заказов
в PostgreSQL

Restore checkout

RTO, мин

30

Тестовый restore, мин

75

RPO, мин

5

Backup interval, мин

15

Результат

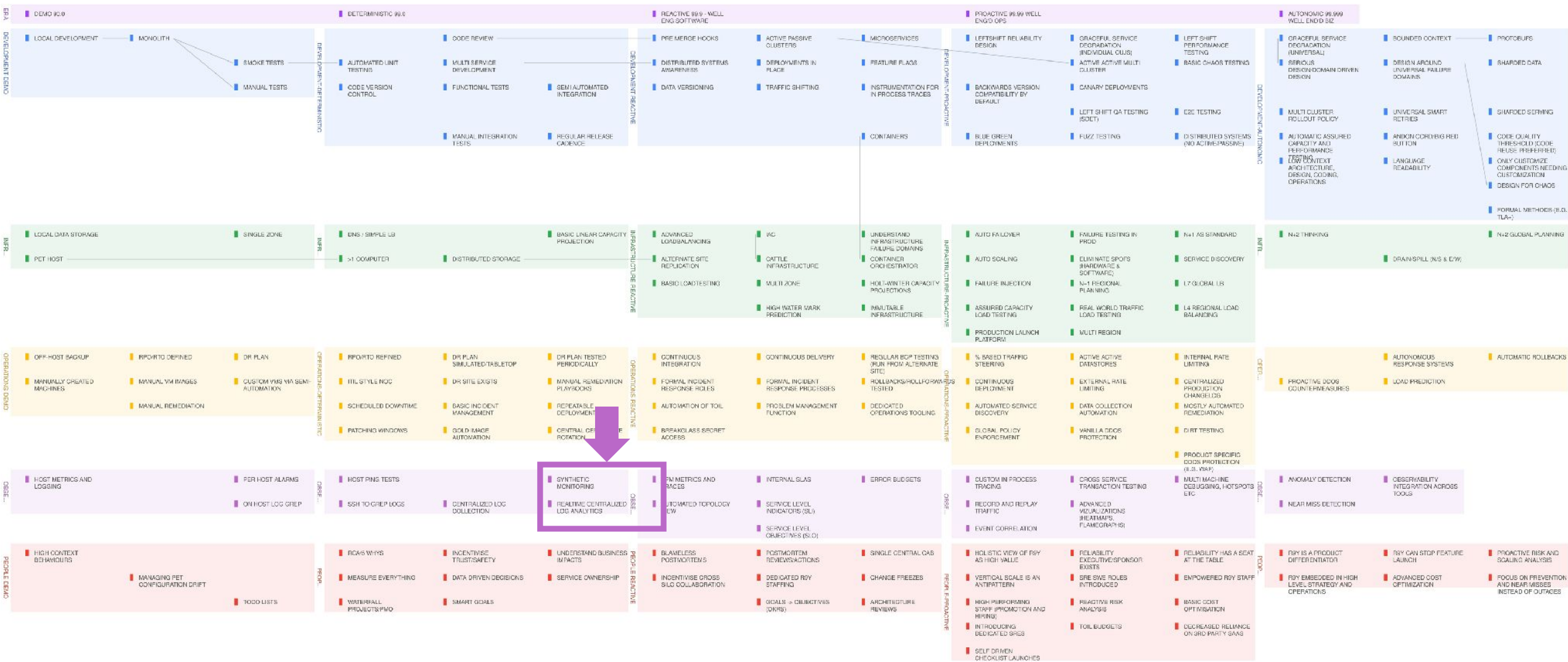
Restore 75 мин > RTO 30 мин.
Checkout ещё **не укладывается** в цель.



MAP → Observability / Deterministic

2/5 · Synthetic Monitoring

Observability



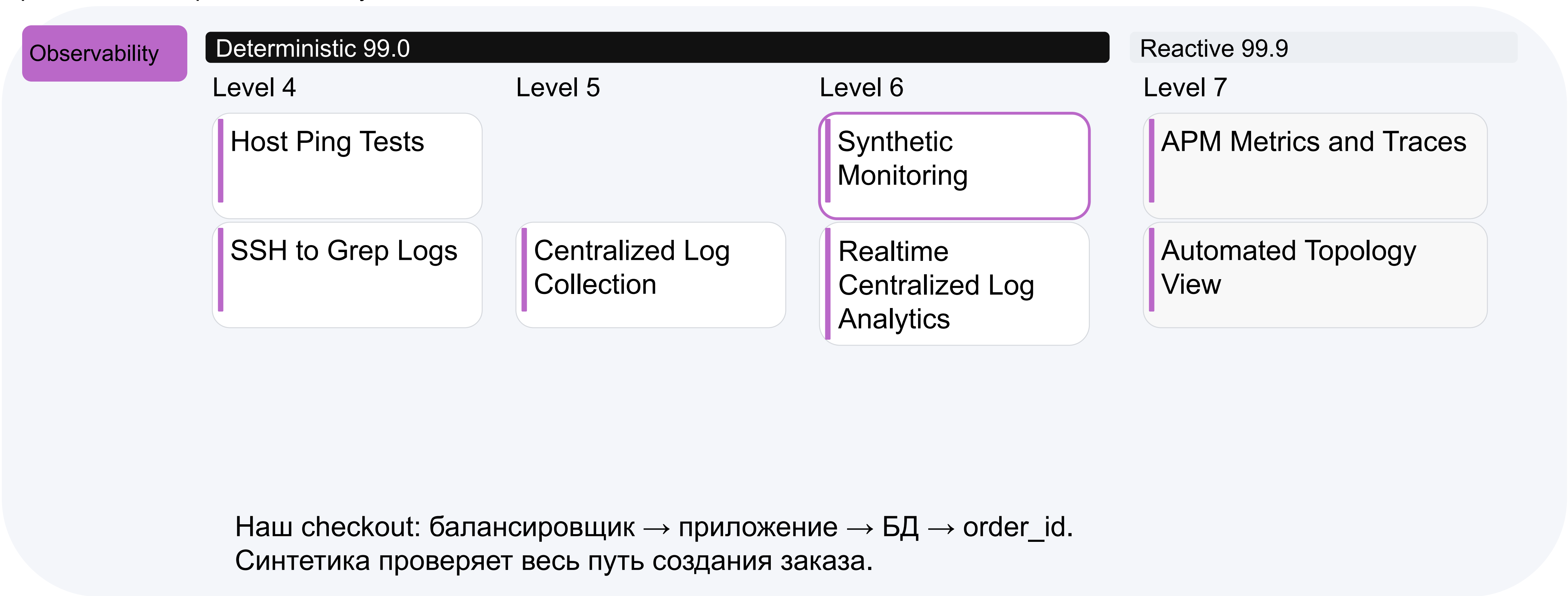
Synthetic monitoring

Deterministic · Level 6

С 05:53 checkout отдавал ошибки.
Проверяем заказ через балансировщик.

Synthetic Monitoring

Фрагмент live map · Observability-Deterministic · Level 6



2. Synthetic monitoring

Наш checkout · 05:53

С 05:53 checkout отдавал ошибки.
Проверяем путь через балансировщик:
создать заказ → получить order_id.

Сбой 05:53 Ошибки и задержка создания заказа

Практика Внешний POST /checkout каждые 60 с

Успех HTTP 2xx + order_id + ответ < 2 с



blackbox_exporter
HTTP-проба на
наш балансировщик



Grafana_k6
POST /checkout
+ проверка order_id

2. Synthetic monitoring

Наш checkout · 05:53

С 05:53 checkout отдавал ошибки.
Проверяем путь через балансировщик:
создать заказ → получить order_id.

Сбой 05:53 Ошибки и задержка создания заказа

Практика Внешний POST /checkout каждые 60 с

Успех HTTP 2xx + order_id + ответ < 2 с



blackbox_exporter
HTTP-проба на
наш балансировщик



Grafana_k6
POST /checkout
+ проверка order_id

Проверка checkout

Проверка

POST/checkout →
order_id

Интервал

60 с

Latency guardrail

Ответ < 2 с

Точка запуска

Из внешней сети

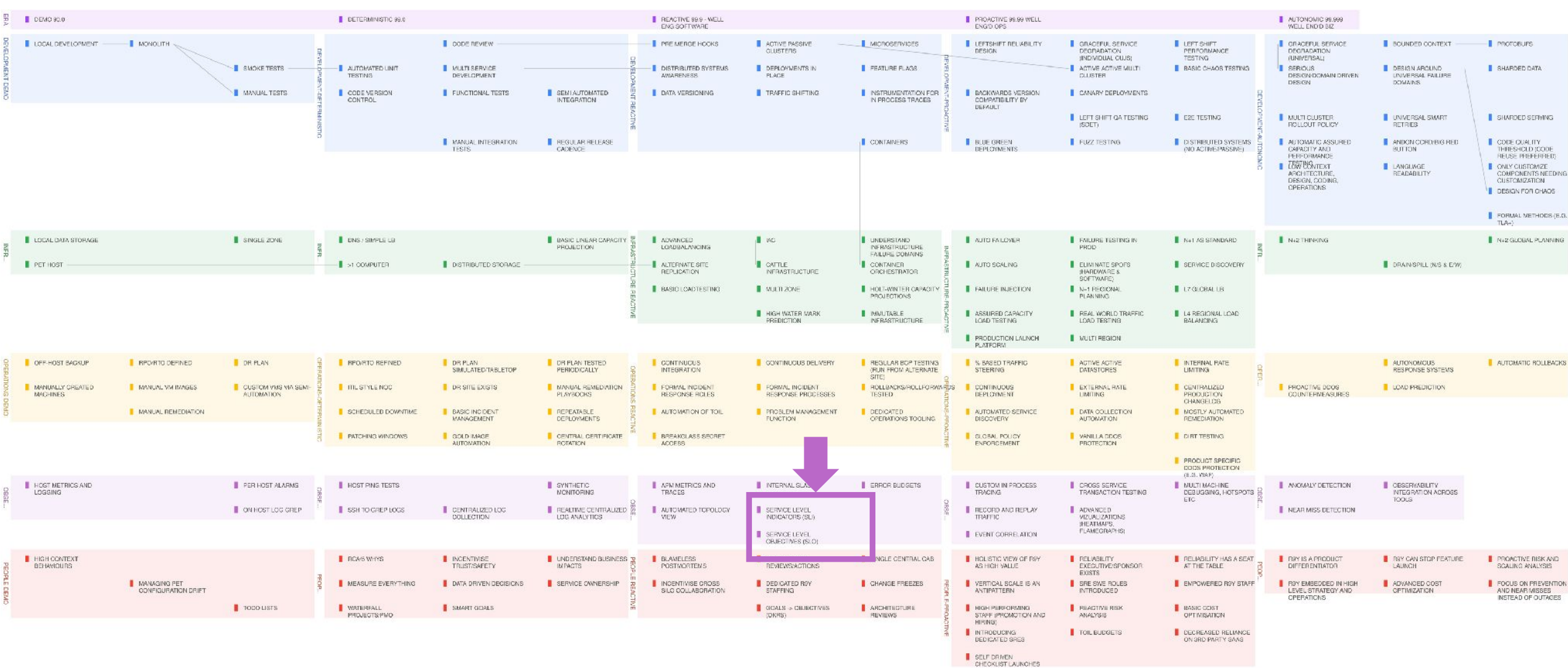
Результат

Заказ проходит через LB → app → БД.
order_id получен за < 2 с.

MAP → Observability / Reactive

3/5 · SLI → SLO → Error Budgets

Observability



SLI / SLO / error budget

Reactive · Levels 8–9

С 05:53 checkout теряет заказы.
SLI измеряет ошибки checkout;
SLO задаёт допустимые потери.

MAP → Observability / Reactive

INCIDENT

MAP

PATH

CRE

SLI → SLO → Error Budgets

Фрагмент live map · Observability-Reactive · Levels 8–9

Observability

Reactive 99.9

Proactive 99.99

Level 7

Level 8

Level 9

Level 10

APM Metrics and
Traces

Internal SLAs

Error Budgets

Record and Replay
Traffic

Automated
Topology View

Service Level
Indicators (SLI)

Service Level
Objectives (SLO)

Тот же инцидент: ошибки checkout расходуют бюджет. SLI → SLO → burn rate показывают масштаб ущерба.

3. SLI → SLO → error budget

Наш checkout · 05:53

В инциденте ретраи усилили сбой БД.
Ущерб пользователю — ошибки
при создании заказа.

SLI Заказы с order_id / все попытки

SLO Например, 99,9% за 30 дней

Решение Burn rate ошибок → приоритет работ



Sloth

Burn alerts по
ошибкам checkout



Pyrra

SLO и бюджет
нашего checkout

3. SLI → SLO → error budget

Наш checkout · 05:53

В инциденте ретраи усилили сбой БД.
Ущерб пользователю — ошибки
при создании заказа.

SLI Заказы с order_id / все попытки

SLO Например, 99,9% за 30 дней

Решение Burn rate ошибок → приоритет работ



Sloth

Burn alerts по
ошибкам checkout



Pyrra

SLO и бюджет
нашего checkout

Бюджет checkout

SLO, %

99,9

Период, дней

30

Checkout attempts в день

100 000

Ошибок сегодня

1200

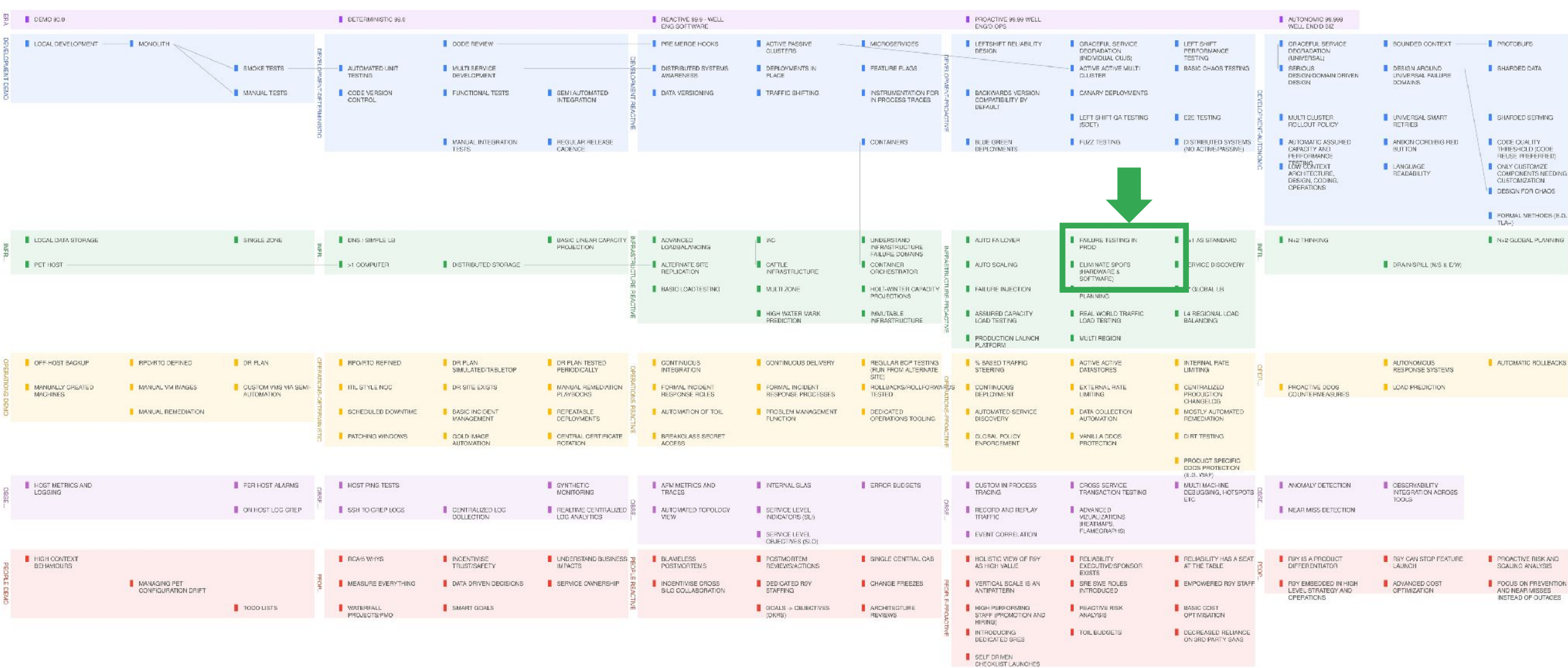
Результат

Бюджет: 3 000 ошибок за 30 дней.
1 200 за день: burn rate 12×.

MAP → Infrastructure / Proactive

4/5 · Failure Testing in Prod

Infrastructure



Failure testing in prod

Proactive · Level 11

Реактивно: ограничили ретраи и восстановили checkout.

Проактивно: задерживаем БД. Проверяем лимит ретраев до следующего инцидента. Нужны **оба** подхода.

Failure Testing in Prod

Фрагмент live map · Infrastructure-Proactive · Level 11

Infrastructure

Proactive 99.99

Level 10

Auto Failover

Auto Scaling

Failure Injection

Assured Capacity Load
Testing

Production Launch Platform

Level 11

Failure Testing in Prod

Eliminate SPOFs (hardware
& software)

N+1 Regional Planning

Real World Traffic Load
Testing

Multi Region

Level 12

N+1 as standard

Service Discovery

L7 Global LB

L4 Regional Load
Balancing

Наш тест: задержка от выбранных checkout к primary PostgreSQL. Проверяем лимит ретраев и пользовательский SLI.

4. Failure testing in prod

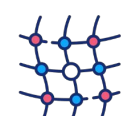
Наш checkout · 05:53

В инциденте: задержка БД → retry storm.
Повторяем этот сценарий и проверяем
добавленный лимит ретраев.

Гипотеза Задержка к primary — без retry storm

Blast radius Выбранные checkout: ≤1% трафика

Stop condition Checkout SLI ниже заданного порога



Chaos_mesh

Задержка пути
checkout → primary



LitmusChaos

Сценарий сбоя БД
+ проверка SLI

4. Failure testing in prod

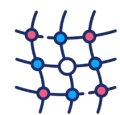
Наш checkout · 05:53

В инциденте: задержка БД → retry storm.
Повторяем этот сценарий и проверяем
добавленный лимит ретраев.

Гипотеза Задержка к primary — без retry storm

Blast radius Выбранные checkout: ≤1% трафика

Stop condition Checkout SLI ниже заданного порога



Chaos_mesh

Задержка пути
checkout → primary



LitmusChaos

Сценарий сбоя БД
+ проверка SLI

Тест checkout

Гипотеза сформулирована
Blast radius ограничен
Пользовательский SLI наблюдаем
Stop/rollback известен

Результат

Ретраи не растут лавиной; SLI в норме.
Старт — после 4/4 guardrails.

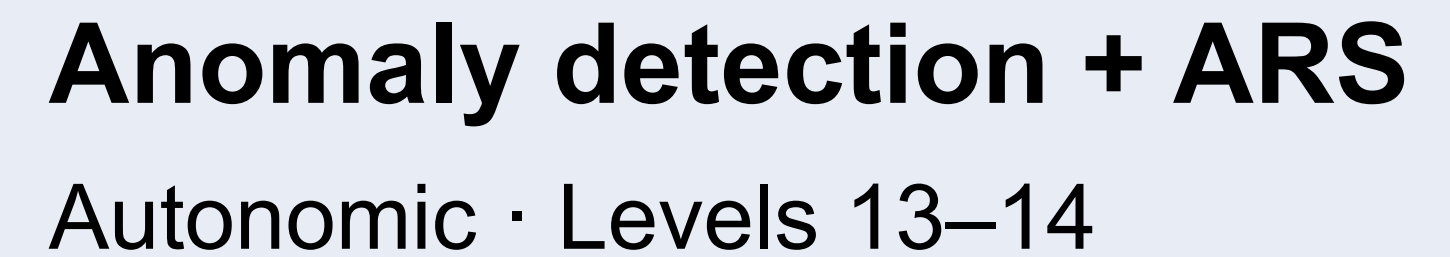
INCIDENT

MAP

PATH

CRE

Observability + Operations

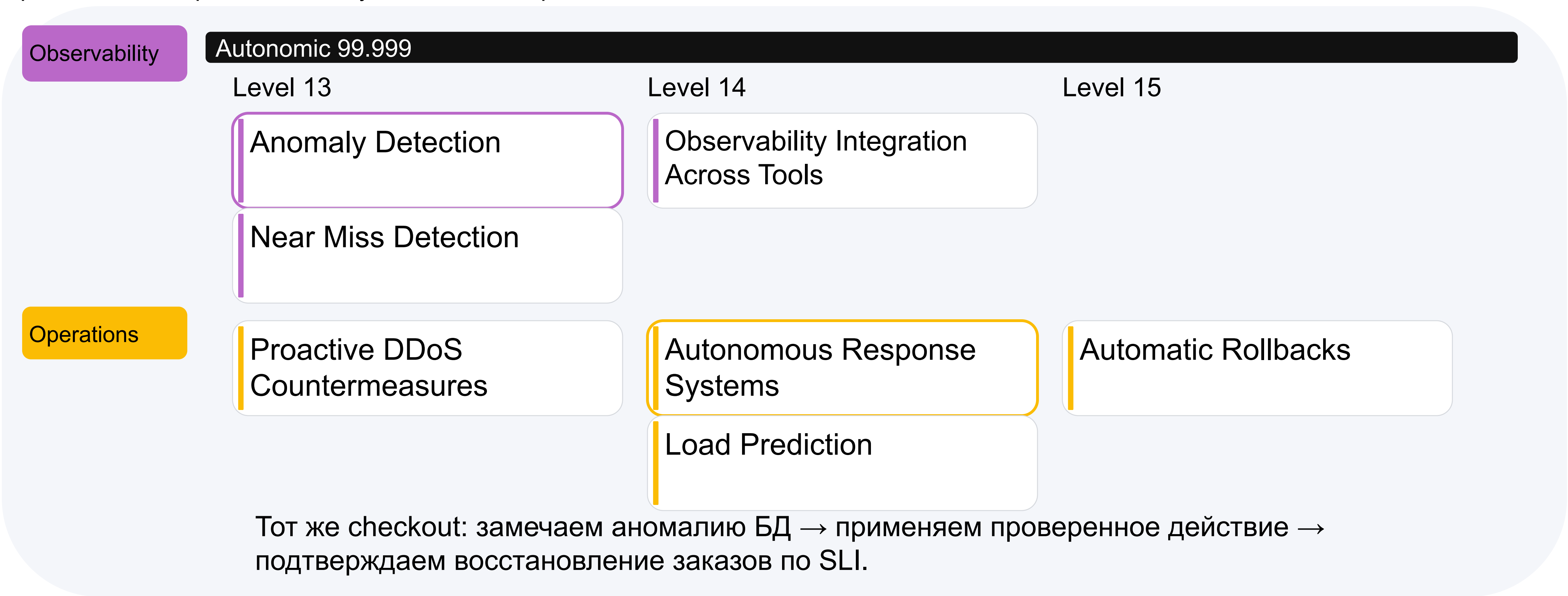


Следующий шаг: сигнал → ответ
→ проверка SLI нашего checkout.

map.r9y.dev

Anomaly Detection + Autonomous Response

Фрагмент live map · Observability-Autonomic + Operations-Autonomic · Levels 13–15



5. Anomaly detection + autonomous response

Наш checkout · 06:22

В 06:22 ретраи ограничили вручную.
Автоматизируем проверенное действие
и проверяем восстановление заказов.

Detect Рост задержки БД + ретраев checkout

Act Runbook: включить лимит ретраев

Verify Доля заказов с order_id восстановлена



OpenSearch AD
Аномалии задержки
БД нашего checkout



Flagger
Если виноват релиз:
rollback по SLI

5. Anomaly detection + autonomous response

Наш checkout · 06:22

В 06:22 ретраи ограничили вручную.
Автоматизируем проверенное действие
и проверяем восстановление заказов.

Detect Рост задержки БД + ретраев checkout

Act Runbook: включить лимит ретраев

Verify Доля заказов с order_id восстановлена



OpenSearch AD
Аномалии задержки
БД нашего checkout



Flagger
Если виноват релиз:
rollback по SLI

Ответ для checkout

Сигнал

DB latency anomaly

Действие

Retry limit

Guardrail

Max 10% traffic

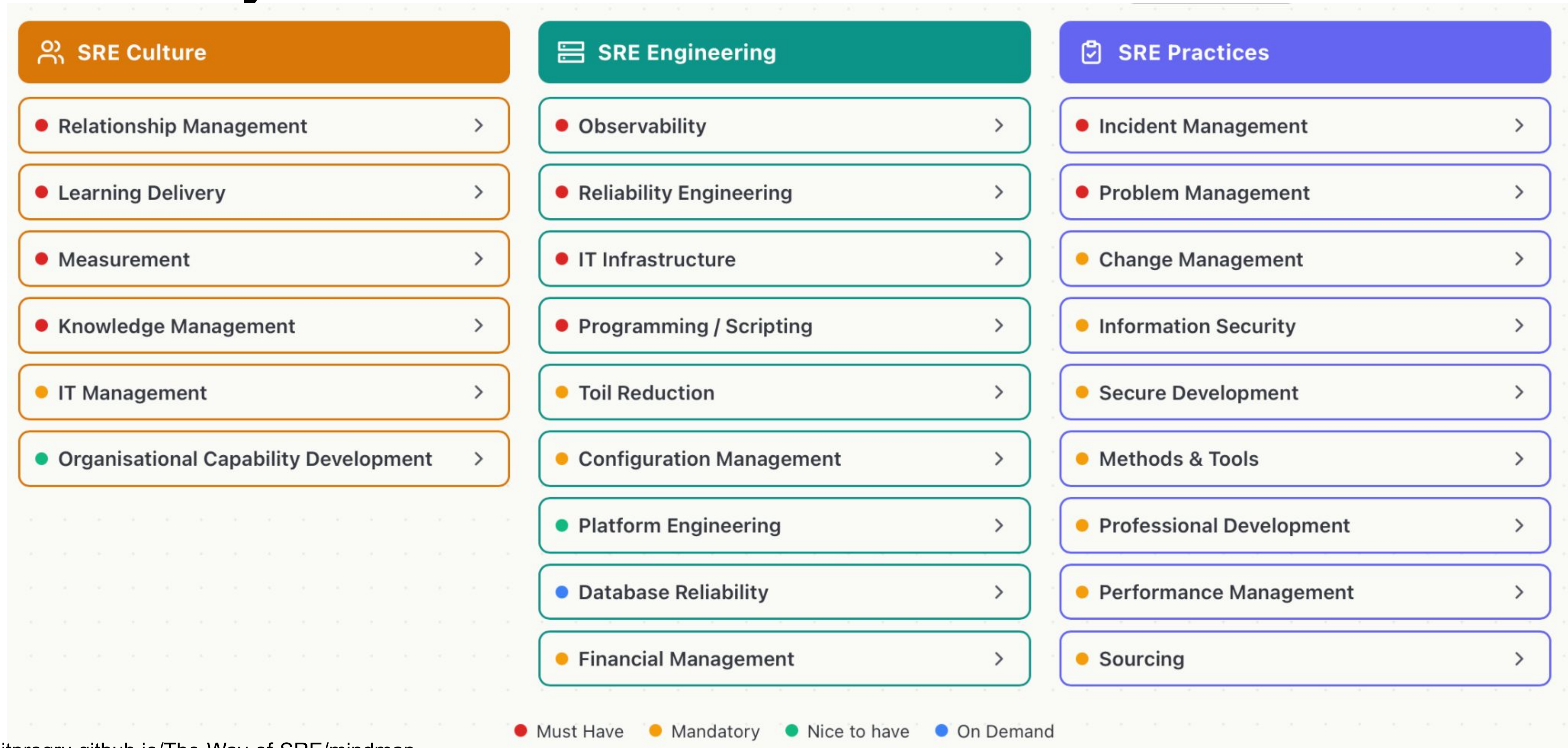
Verify

Заказы снова
проходят

Результат

Сигнал БД → retry limit → SLI заказов.
Действие — по проверенному runbook.

The Way of SRE



The Way of SRE: чем вообще можно влиять?

SRE culture

SRE engineering

SRE practices

The Way of SRE: чем вообще можно влиять?

SRE culture

Ownership, learning,
stakeholders,
collaboration

SRE engineering

Observability, SLO,
resilience, capacity,
performance

SRE practices

Incident, problem,
change, runbooks,
postmortems

**Прекрасно. Теперь мы знаем ещё больше
вещей, которыми можно заняться**

Замысел автора · Миша Савин

INCIDENT

MAP

PATH

CRE

The Way of SRE: карта компетенций SRE

Для русскоязычного сообщества: от junior до principal

Для кого

Есть сейчас

Куда дальше

Замысел автора · Миша Савин

INCIDENT

MAP

PATH

CRE

The Way of SRE: карта компетенций SRE

Для русскоязычного сообщества: от junior до principal

Для кого

Инженеру — рост
от junior до principal.
Тимлиду — грейды,
онбординг и план
развития.

Есть сейчас

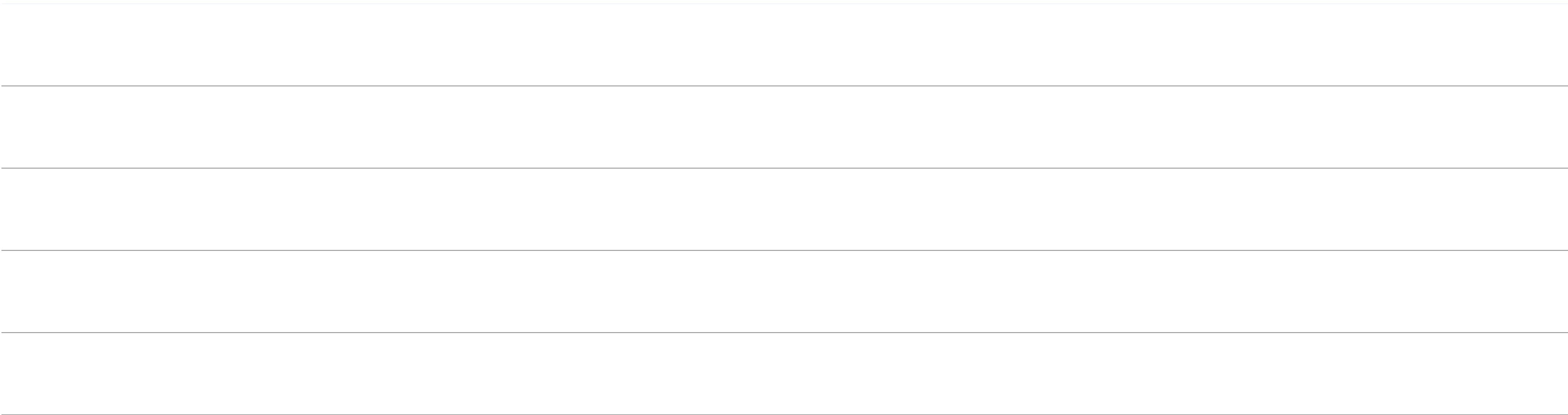
3 ветви, ≈70
компетенций.
Приоритеты для ролей,
уровни SFIA,
материалы
и best practices.

Куда дальше

v1.0: закрыть пробелы.
Критерии готовности.
Каталог шаблонов.
Кейсы и соавторы
из сообщества.

**Референс для своего контекста, не стандарт.
Личный open source · bus factor = 1.**

Что означают уровни reliability hierarchy



Foundation

Основа

Видеть состояние сервиса и проблемы пользователя

Foundation	Основа	Видеть состояние сервиса и проблемы пользователя
------------	--------	--

Что означают уровни reliability hierarchy

Learning	Обучение на сбоях	Разбирать причины и доводить действия до результата
Response	Реагирование	Ограничить ущерб и восстановить сервис
Foundation	Основа	Видеть состояние сервиса и проблемы пользователя

Что означают уровни reliability hierarchy

Prevention	Предотвращение	Снижать риск через тесты и безопасные изменения
Learning	Обучение на сбоях	Разбирать причины и доводить действия до результата
Response	Реагирование	Ограничить ущерб и восстановить сервис
Foundation	Основа	Видеть состояние сервиса и проблемы пользователя

Что означают уровни reliability hierarchy

Planning	Планирование	Учитывать рост нагрузки и нехватку ресурсов заранее
Prevention	Предотвращение	Снижать риск через тесты и безопасные изменения
Learning	Обучение на сбоях	Разбирать причины и доводить действия до результата
Response	Реагирование	Ограничить ущерб и восстановить сервис
Foundation	Основа	Видеть состояние сервиса и проблемы пользователя

Что означают уровни reliability hierarchy

Strategy	Стратегия	Закладывать надёжность в продукт и архитектуру
Planning	Планирование	Учитывать рост нагрузки и нехватку ресурсов заранее
Prevention	Предотвращение	Снижать риск через тесты и безопасные изменения
Learning	Обучение на сбоях	Разбирать причины и доводить действия до результата
Response	Реагирование	Ограничить ущерб и восстановить сервис
Foundation	Основа	Видеть состояние сервиса и проблемы пользователя

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента



Monitoring

Видим пользовательский симптом

Foundation

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Incident response

Умеем снять impact

Response

Monitoring

Видим пользовательский симптом

Foundation

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Postmortem/RCA	Понимаем, почему это произошло	Learning
Incident response	Умеем снять impact	Response
Monitoring	Видим пользовательский симптом	Foundation

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Testing & release	Не вносим риск бесконтрольно	Prevention
Postmortem/RCA	Понимаем, почему это произошло	Learning
Incident response	Умеем снять impact	Response
Monitoring	Видим пользовательский симптом	Foundation

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Capacity	Понимаем пределы	Planning
Testing & release	Не вносим риск бесконтрольно	Prevention
Postmortem/RCA	Понимаем, почему это произошло	Learning
Incident response	Умеем снять impact	Response
Monitoring	Видим пользовательский симптом	Foundation

Что должно работать раньше?

INCIDENT

MAP

PATH

CRE

Reliability hierarchy: не лестница, а проверка фундамента

Development/product	Надёжность встроена в решения	Strategy
Capacity	Понимаем пределы	Planning
Testing & release	Не вносим риск бесконтрольно	Prevention
Postmortem/RCA	Понимаем, почему это произошло	Learning
Incident response	Умеем снять impact	Response
Monitoring	Видим пользовательский симптом	Foundation

Next best reliability step

INCIDENT

MAP

PATH

CRE

Чтобы выбрать следующий шаг, задайте четыре вопроса

Где мы?

Какая область
реально
ограничивает
систему?

Что можно сделать?

Какие практики
закрывают
этот риск?

Что раньше?

Что является
prerequisite/
foundation?

Зачем?

Какой business
impact, цена
и сложность?



Следующий шаг = пересечение maturity, риска, зависимостей и экономики

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя
2. Retry guardrail	Закрыть конкретный failure amplification
3. DR test	Проверить RTO/RPO и recovery path
4. Chaos program	Расширить набор проверенных failure modes
5. Auto remediation	Замыкать только безопасные повторяемые действия

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя	1–2 недели
2. Retry guardrail	Закрыть конкретный failure amplification	
3. DR test	Проверить RTO/RPO и recovery path	
4. Chaos program	Расширить набор проверенных failure modes	
5. Auto remediation	Замыкать только безопасные повторяемые действия	

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя	1–2 недели
2. Retry guardrail	Закрыть конкретный failure amplification	1 спринт
3. DR test	Проверить RTO/RPO и recovery path	
4. Chaos program	Расширить набор проверенных failure modes	
5. Auto remediation	Замыкать только безопасные повторяемые действия	

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя	1–2 недели
2. Retry guardrail	Закрыть конкретный failure amplification	1 спринт
3. DR test	Проверить RTO/RPO и recovery path	1 game day
4. Chaos program	Расширить набор проверенных failure modes	
5. Auto remediation	Замыкать только безопасные повторяемые действия	

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя	1–2 недели
2. Retry guardrail	Закрыть конкретный failure amplification	1 спринт
3. DR test	Проверить RTO/RPO и recovery path	1 game day
4. Chaos program	Расширить набор проверенных failure modes	Потом
5. Auto remediation	Замыкать только безопасные повторяемые действия	

Тот же checkout, теперь roadmap

INCIDENT

MAP

PATH

CRE

Не «сделать всё». Снять крупнейшие неизвестные по очереди

1. Synthetic + SLI	Перестать узнавать о проблеме от пользователя	1–2 недели
2. Retry guardrail	Закрыть конкретный failure amplification	1 спринт
3. DR test	Проверить RTO/RPO и recovery path	1 game day
4. Chaos program	Расширить набор проверенных failure modes	Потом
5. Auto remediation	Замыкать только безопасные повторяемые действия	Ещё позже

У нас всё ещё один инженер. Service mesh пока положите обратно

Карта — не backlog

He skill tree

Autonomic —
не финальный босс,
которого обязаны
победить все



He procurement list

Наличие
инструмента
не означает
наличие практики



He maturity score

Разным областям
системы нужен
разный уровень
зрелости



Хорошая карта помогает осознанно сказать «Не сейчас»

SLA — это вход, а не цель

INCIDENT

MAP

PATH

CRE

User availability может быть выше SLA отдельной зависимости



SLA — это вход, а не цель

INCIDENT

MAP

PATH

CRE

User availability может быть выше SLA отдельной зависимости

Компонент провайдера

SLA 99,9% относится к отдельной зависимости.

Доступность сервиса измеряем отдельно.

SLA — это вход, а не цель

INCIDENT

MAP

PATH

CRE

User availability может быть выше SLA отдельной зависимости

Компонент провайдера

SLA 99,9% относится к отдельной зависимости.

Доступность сервиса измеряем отдельно.

Система

Multi-AZ, резервирование, ограниченные ретраи, graceful degradation и быстрое восстановление.

SLA — это вход, а не цель

INCIDENT

MAP

PATH

CRE

User availability может быть выше SLA отдельной зависимости

Компонент провайдера

SLA 99,9% относится к отдельной зависимости.
Доступность сервиса измеряем отдельно.

Система

Multi-AZ, резервирование, ограниченные ретраи, graceful degradation и быстрое восстановление.

User journey

За наблюдаемый период checkout может завершить 100% пользовательских сценариев успешно.

99,99% или 100% — чего именно и за какой период?

SLA — это вход, а не цель

INCIDENT

MAP

PATH

CRE

User availability может быть выше SLA отдельной зависимости

Компонент провайдера

SLA 99,9% относится к отдельной зависимости.
Доступность сервиса измеряем отдельно.

Система

Multi-AZ, резервирование, ограниченные ретраи, graceful degradation и быстрое восстановление.

User journey

За наблюдаемый период checkout может завершить 100% пользовательских сценариев успешно.

SLA клиента — обязательства; SLO/SLI — цели и измерения.

Критичный сценарий → допустимый ущерб → проверка результата.

100% за период — измеренный **результат**.

Организационная граница

INCIDENT

MAP

PATH

CRE

Тот самый checkout — продакшн-сервис клиента

Клиент знает

Приложение, бизнес-сценарии,
зависимости, историю решений,
допустимый impact

Client model

Тот самый checkout — продакшн-сервис клиента

Клиент знает

Приложение, бизнес-сценарии,
зависимости, историю решений,
допустимый impact

Облако знает

Failure modes сервисов, ограничения
платформы, внутреннюю эксплуатацию,
эскалационный контекст

Client model

+

Cloud model

Тот самый checkout — продакшн-сервис клиента

Клиент знает

Приложение, бизнес-сценарии,
зависимости, историю решений,
допустимый impact

Облако знает

Failure modes сервисов, ограничения
платформы, внутреннюю эксплуатацию,
эскалационный контекст

Client model

+

Cloud model



CRE



Reliability roadmap

Практики CRE — для продакшна клиентов с чёткими SLA

Чекап ИС + ИБ

Риски продакшна и план улучшений под SLA/SLO
Best practices

Инженерная синхронизация

Риски, изменения, архитектура и контекст системы

Клиентский incident+problem management

Инженерный трек, RCA и предотвращение повторения

Эскалации

Факты, гипотезы, дополнительная диагностика и контекст

Hands-on

Совместная работа: архитектура, лимиты и особенности сервисов

Maintenance / upgrade

Окно работ, контроль SLI, rollback и помощь при сбое

Паспорт ИС + runbooks

SLA/SLO, RTO/RPO, контакты и runbooks для продакшна

HA/DR-тесты

Учения, отчёт, фактические RTO/RPO и ограничения

Практики CRE — для продакшна клиентов с чёткими SLA

Чекап ИС + ИБ

Риски продакшна и план улучшений под SLA/SLO
Best practices

По запросу

Инженерная синхронизация

Риски, изменения, архитектура и контекст системы

Регулярно

Клиентский incident+problem management

Инженерный трек, RCA и предотвращение повторения

Регулярно

Эскалации

Факты, гипотезы, дополнительная диагностика и контекст

По запросу

Hands-on

Совместная работа: архитектура, лимиты и особенности сервисов

По запросу

Maintenance / upgrade

Окно работ, контроль SLI, rollback и помощь при сбое

По запросу

Паспорт ИС + runbooks

SLA/SLO, RTO/RPO, контакты и runbooks для продакшна

Регулярно

HA/DR-тесты

Учения, отчёт, фактические RTO/RPO и ограничения

По запросу

Два режима работы

INCIDENT

MAP

PATH

CRE

Регулярный контур помнит
систему.
Запросный — делает конкретную
работу

Регулярные

Инженерная синхронизация,
клиентский incident management,
паспорт ИС / контекст эксплуатации

По запросу

Чекап, эскалации, hands-on,
maintenance/upgrade, HA/DR-тесты

Эти режимы не конкурируют. Один снижает стоимость
контекста, второй — стоимость конкретной инженерной задачи

Возвращаемся в 06:00

INCIDENT

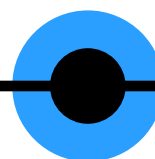
MAP

PATH

CRE

В 06:00 дежури́м рядом с клиентом

06:00



Дежурство CRE
+ эскалация

Инженерный трек,
контекст и контроль SLI

Возвращаемся в 06:00

INCIDENT

MAP

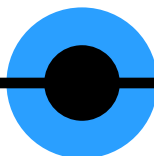
PATH

CRE

В 06:00 дежури́м рядом с клиентом

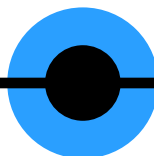
05:53 — деградация → 06:00 — вызов / ask · 7 минут до подключения

До 05:53



Деградация

06:00



**Дежурство CRE
+ эскалация**

Инженерный трек,
контекст и контроль SLI

Возвращаемся в 06:00

INCIDENT

MAP

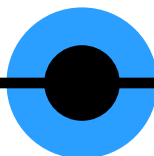
PATH

CRE

В 06:00 дежури́м рядом с клиентом

05:53 — деградация → 06:00 — вызов / ask · 7 минут до подключения

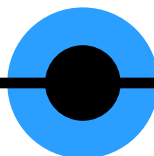
До 05:53



**Чекап + паспорт
+ синхронизация**

Знаем critical flow,
SLA/SLO, зависимости и риски

06:00



**Дежурство CRE
+ эскалация**

Инженерный трек,
контекст и контроль SLI

Возвращаемся в 06:00

INCIDENT

MAP

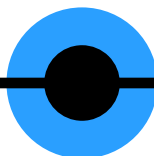
PATH

CRE

В 06:00 дежури́м рядом с клиентом

05:53 — деградация → 06:00 — вызов / ask · 7 минут до подключения

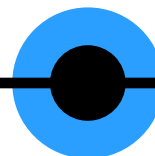
До 05:53



**Чекап + паспорт
+ синхронизация**

Знаем critical flow,
SLA/SLO, зависимости и риски

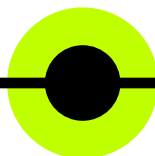
06:00



**Дежурство CRE
+ эскалация**

Инженерный трек,
контекст и контроль SLI

После



**RCA + problem
management + DR**

Вывод превращается
в проверяемое изменение

Возвращаемся в 06:00

INCIDENT

MAP

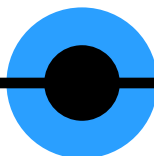
PATH

CRE

В 06:00 дежури́м рядом с клиентом

05:53 — деградация → 06:00 — вызов / ask · 7 минут до подключения

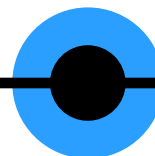
До 05:53



**Чекап + паспорт
+ синхронизация**

Знаем critical flow,
SLA/SLO, зависимости и риски

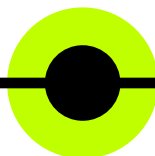
06:00



**Дежурство CRE
+ эскалация**

Инженерный трек,
контекст и контроль SLI

После



**RCA + problem
management + DR**

Вывод превращается
в проверяемое изменение

Ценность CRE не в том, что в 06:00 появляется ещё один инженер.
Ценность в том, что он не начинает узнавать систему в 06:00

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик,
которые **уже** работают у вас

1 · Что уже есть?

2 · Где пробел?

3 · Что сначала?

4 · Как
проверить?

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик, которые **уже** работают у вас

1 · Что уже есть?

Отметьте на карте практики, которые команда действительно применяет.

2 · Где пробел?

3 · Что сначала?

4 · Как проверить?

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик, которые **уже** работают у вас

1 · Что уже есть?

Отметьте на карте практики, которые команда действительно применяет.

2 · Где пробел?

Сопоставьте результат этих практик с рисками и ограничениями системы.

3 · Что сначала?

4 · Как проверить?

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик, которые **уже** работают у вас

1 · Что уже есть?

Отметьте на карте практики, которые команда действительно применяет.

2 · Где пробел?

Сопоставьте результат этих практик с рисками и ограничениями системы.

3 · Что сначала?

Выберите соседнюю практику.
Проверьте, готов ли необходимый фундамент.

4 · Как проверить?

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик, которые **уже** работают у вас

1 · Что уже есть?

Отметьте на карте практики, которые команда действительно применяет.

2 · Где пробел?

Сопоставьте результат этих практик с рисками и ограничениями системы.

3 · Что сначала?

Выберите соседнюю практику.
Проверьте, готов ли необходимый фундамент.

4 · Как проверить?

Задайте владельца, критерий готовности и способ измерить результат.

Что делать теперь

INCIDENT

MAP

PATH

CRE

Начните с практик, которые **уже** работают у вас

1 · Что уже есть?

Отметьте на карте практики, которые команда действительно применяет.

2 · Где пробел?

Сопоставьте результат этих практик с рисками и ограничениями системы.

3 · Что сначала?

Выберите соседнюю практику.
Проверьте, готов ли необходимый фундамент.

4 · Как проверить?

Задайте владельца, критерий готовности и способ измерить результат.

Пример нашего сервиса checkout

1 · Что есть

Мониторинг хостов.

2 · Где пробел

О сбое узнаём от пользователя.

3 · Что сначала

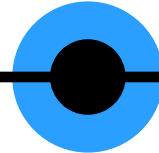
Синтетика заказа, затем SLI → SLO.

4 · Как проверить

SRE/CRE, алерт и order_id за < 2 с.

Что делать завтра. Завтра, в 06:00

06:00



Что делать завтра. Завтра, в 06:00

INCIDENT

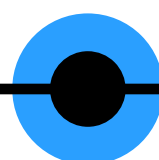
MAP

PATH

CRE

Пусть это хотя бы будет НОВЫЙ ИНЦИДЕНТ

А не тот же самый, который мы уже три раза красиво описали в RCA
06:00



Что делать завтра. Завтра, в 06:00

INCIDENT

MAP

PATH

CRE

Пусть это хотя бы будет НОВЫЙ ИНЦИДЕНТ

А не тот же самый, который мы уже три раза красиво описали в RCA

Где мы?

Состояние
конкретной области



Какой риск важнее?

Impact, частота,
RTO/RPO,
error budget



Что следующее?

Практика
→ инструмент
→ проверяемый
результат



Что забрать из доклада

Что понять

Что взять

Что сделать

Что забрать из доклада

Что понять

Надёжность — это успешный сценарий пользователя.
SLI и тесты это проверяют.

Что взять

Что сделать

Что забрать из доклада

Что понять

Надёжность — это успешный сценарий пользователя.
SLI и тесты это проверяют.

Что взять

r9y — практики и зависимости.
TWO SRE — компетенции и развитие команды.

Что сделать

Что забрать из доклада

Что понять

Надёжность — это успешный сценарий пользователя.
SLI и тесты это проверяют.

Что взять

r9y — практики и зависимости.
TWO SRE — компетенции и развитие команды.

Что сделать

Выберите одну практику.
Назначьте владельца и критерий готовности.
Проверьте результат.

Один сервис → один риск → одна практика → проверяемый результат.

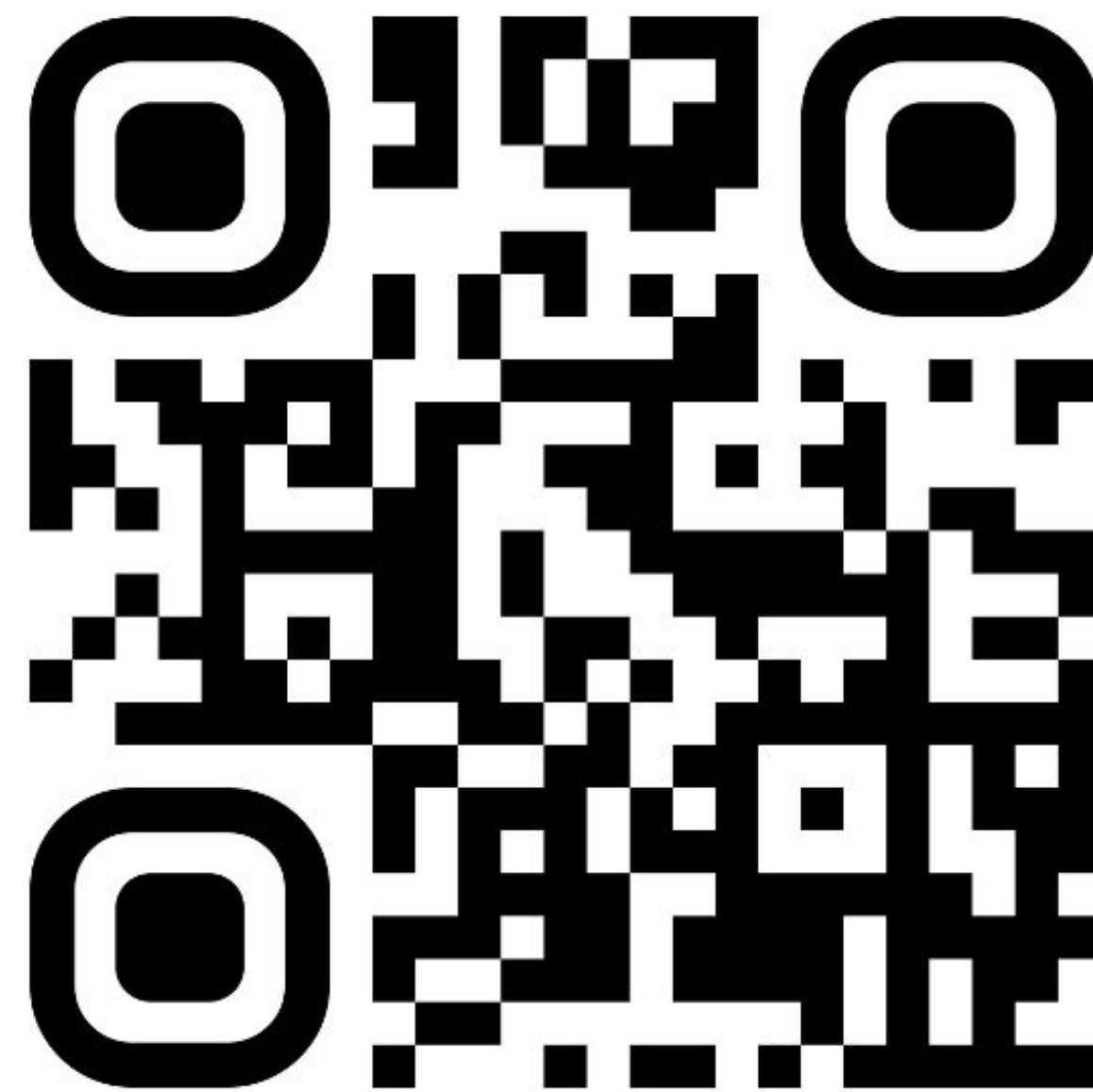
Сделайте вклад в обе карты

Кейсы, уточнения, шаблоны — ваш опыт поможет следующему инженеру.



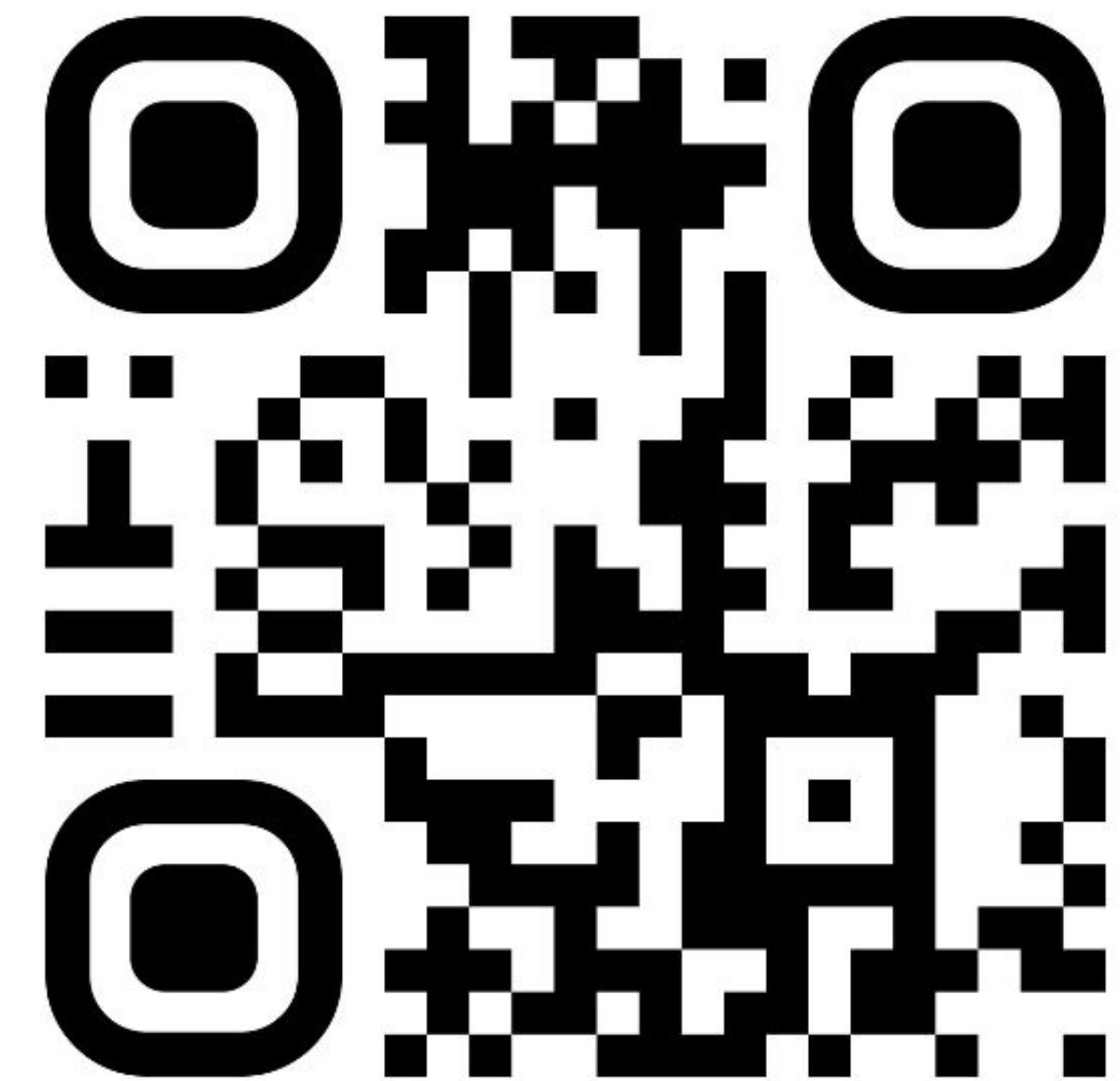
r9y reliability map:

clck.ru/3VcRme



The way of SRE:

clck.ru/3VcSQd



Антон:

t.me/owl_tech

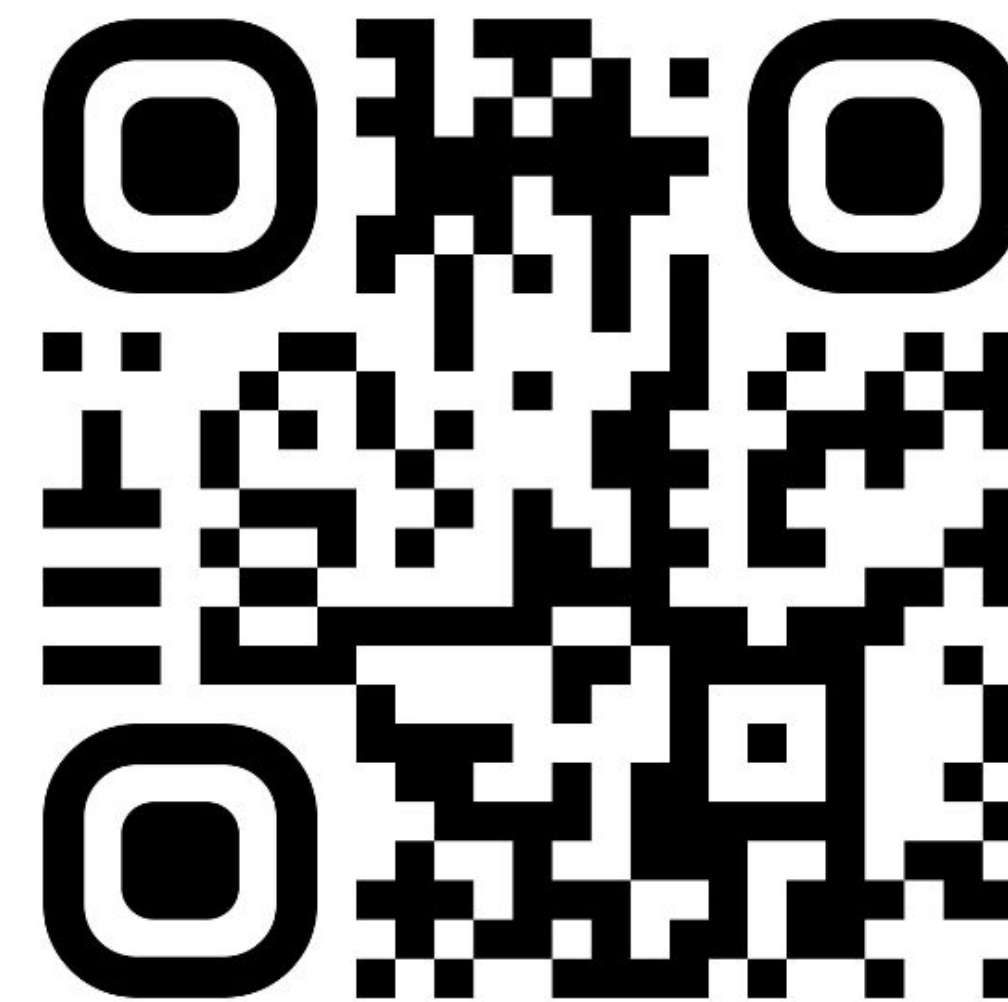
Надежность требует
совместных усилий

Спасибо за внимание!



Антон Егорушков

Руководитель группы Customer
Reliability Engineering, Yandex Cloud



t.me/owl_tech



Appendix: источник

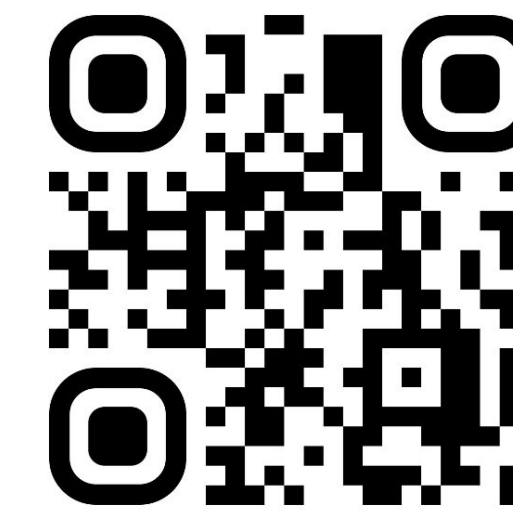
r9y reliability map: оригинал



Оригинальная интерактивная tech-tree-карта

Эры, области, зависимости, карточки практик:

clck.ru/3VcTRf

clck.ru/3VcTZD

Appendix: источник

The way of SRE: оригинал

Три ветви, 24 компетенции верхнего уровня, десятки подветвей

