DEVIEW
2019

Fail Fast, Learn Faster SRE
(실패에서 배워나가는 SRE)

김재헌, 유호균
Service System Reliability
NAVER

CONTENTS

DEVIEW
2019

1. 왜 SRE를 도입해야 하는가?
2. 우리들의 SRE
3. Metric + Meta Data = Insight
4. Fail Fast, Learn Faster
5. Lessons Learned

DEVIEW
2019

1. 왜 SRE를 도입해야 하는가?

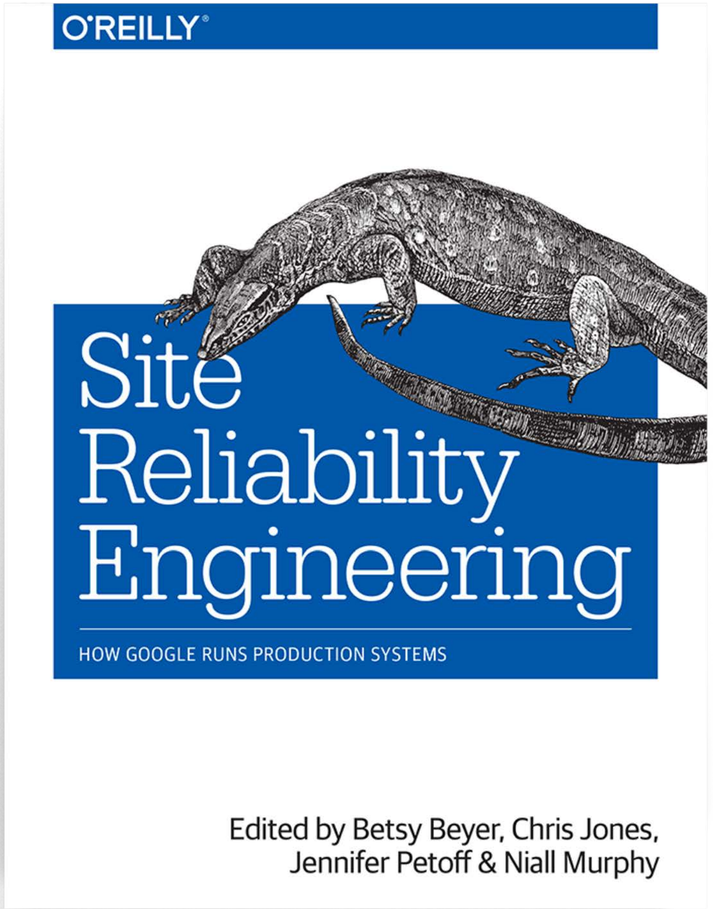
1. 왜 SRE를 도입해야 하는가

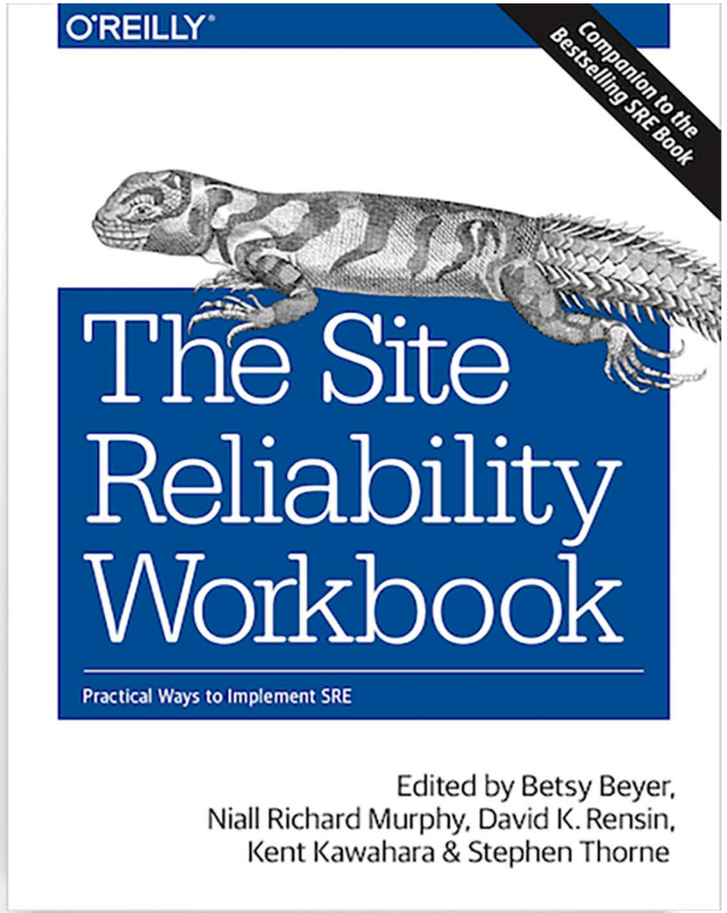
DEVIEW
2019

SRE: Site Reliability Engineering
사이트 신뢰성 엔지니어링

글로벌 스케일, 또는 대규모의 인터넷 서비스를 제공하면서 어떻게 하면
시스템의 신뢰성을 보장할 수 있을지 고민하는 기술 분야이자 방법론, 문화

1. 왜 SRE를 도입해야 하는가





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

1. 왜 SRE를 도입해야 하는가

DEVIEW
2019

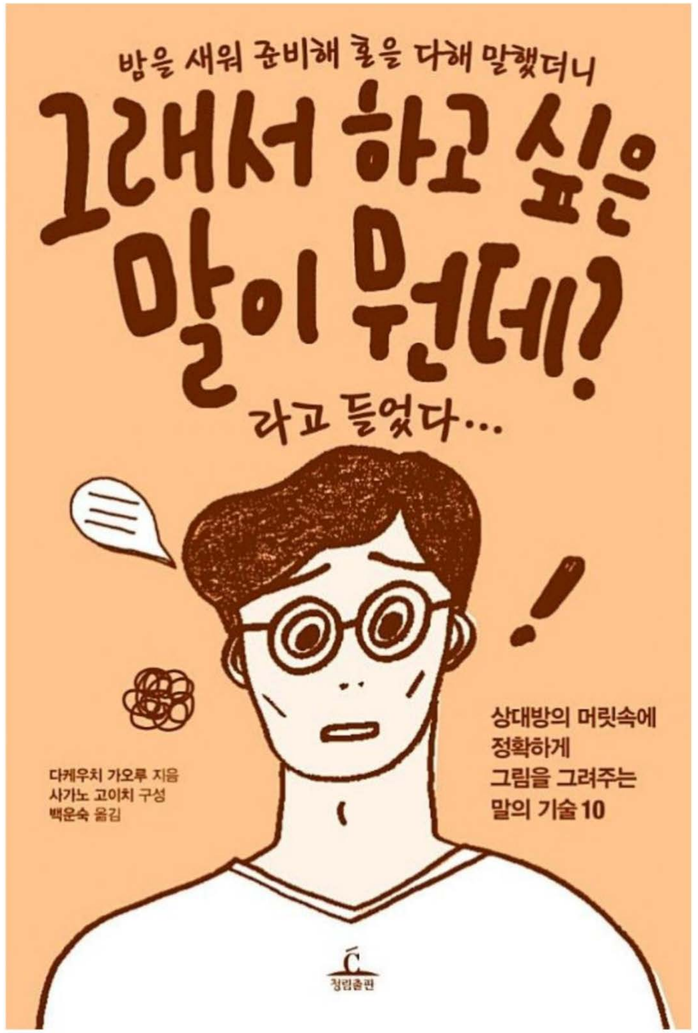
SRE: Site Reliability Engineering

사이트 신뢰성 엔지니어링

SLI, SLO, SLA, Error Budget, DevOps, Post-Mortem, Blameless,
Toil management, Silo, Availability, Emergency Response,
Capacity Planning, Change Management, On-Call, Monitoring

~~**네... 좋은 말씀 잘 들었습니다...**~~

1. 왜 SRE를 도입해야 하는가



1. 왜 SRE를 도입해야 하는가

SRE: Site Reliability Engineering
사이트 신뢰성 엔지니어링

글로벌 스케일, 또는 대규모의 인터넷 서비스를 제공하면서 어떻게 하면
시스템의 **신뢰성**을 보장할 수 있을지 고민하는 기술 분야이자 **방법론**, **문화**

네이버 검색 SRE팀의 정의

신뢰성: 네이버 검색 서비스가 정상적으로 제공되는 것

방법론: Availability, Capacity Planning, Monitoring, Contingency Plan

문화 : Post-Mortem, Blameless, Fact-Based

1. 왜 SRE를 도입해야 하는가

DEVIEW 2019

SRE: Site Reliability Engineering

사이트 신뢰성 엔지니어링

글로벌 스케일, 또는 대규모의 인터넷 서비스를 제공하면서 어떻게 하면
시스템의 **신뢰성**을 보장할 수 있을지 고민하는 기술 분야이자 **방법론**, **문화**

그런데 SRE, 꼭 필요한가요?

네이버 검색 SRE팀의 정의

신뢰성: 네이버 검색 서비스가 정상적으로 제공되는 것

방법론: Availability, Capacity Planning, Monitoring, Contingency Plan

문화 : Post-Mortem, Blameless, Fact-Based

1. 왜 SRE를 도입해야 하는가

DEVIEW 2019

SRE: Site Reliability Engineering

사이트 신뢰성 엔지니어링

1.

네이버 검색은 365일 / 24시간 무중단 제공되어야 한다

글로벌 스케일, 또는 대규모의 인터넷 서비스를 제공하면서 어떻게 하면 시스템의 **신뢰성**을 보장할 수 있을지 고민하는 기술 분야이자 **방법론**, **문화**

네이버 검색 SRE팀의 정의

신뢰성: 네이버 검색 서비스가 정상적으로 제공되는 것

방법론: Availability, Capacity Planning, Monitoring, Contingency Plan

문화 : Post-Mortem, Blameless, Fact-Based

1. 왜 SRE를 도입해야 하는가

DEVIEW
2019

만약...

1. 사용자가 검색을 했는데..

2. 검색 불가

1. 왜 SRE를 도입해야 하는가

검색은 **절대** 서비스 제공이

중단되면 안됩니다.

온 국민이 사용하는 **공공재**니까요!

1. 왜 SRE를 도입해야 하는가

DEVIEW 2019

SRE: Site Reliability Engineering

사이트 신뢰성 엔지니어링

2.

서비스 규모가 커지는 만큼

엔지니어는 늘어나지 않는다

글로벌 스케일, 또는 네이버의 인터넷 서비스를 제공하면서 어떻게 하면
시스템의 **신뢰성**을 보장할 수 있을지 고민하는 기술 분야인 **방법론**, **문화**

네이버 검색 SRE팀의 정의

신뢰성: 네이버 검색 서비스가 정상적으로 제공되는 것

방법론: Availability, Capacity Planning, Monitoring, Contingency Engineering

문화 : Post-Mortem, Blameless, Fact-Based

1. 왜 SRE를 도입해야 하는가

DEVIEW
2019

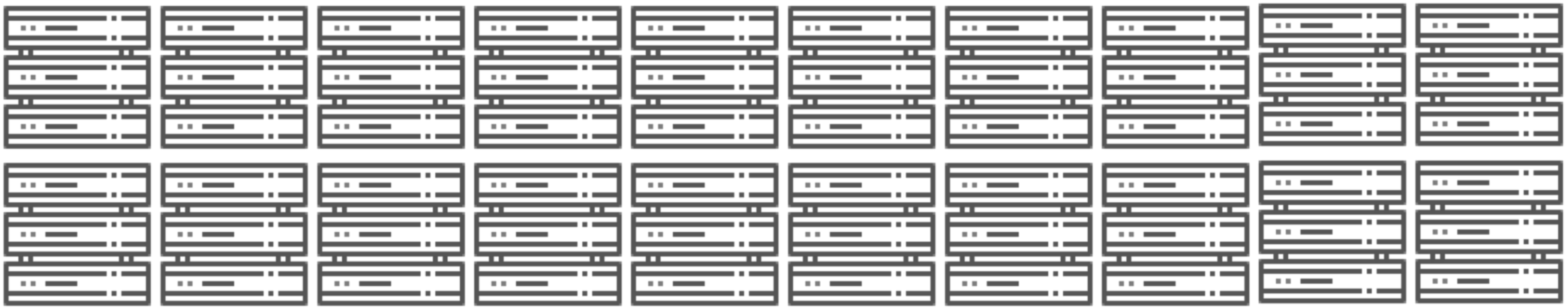
몇대의 서버는 1명이 관리 가능

1. 왜 SRE를 도입해야 하는가

DEVIEW
2019

몇대의 서버는 1명이 관리 가능

수십 ~ 수백대의 서버는 몇명이 관리 가능

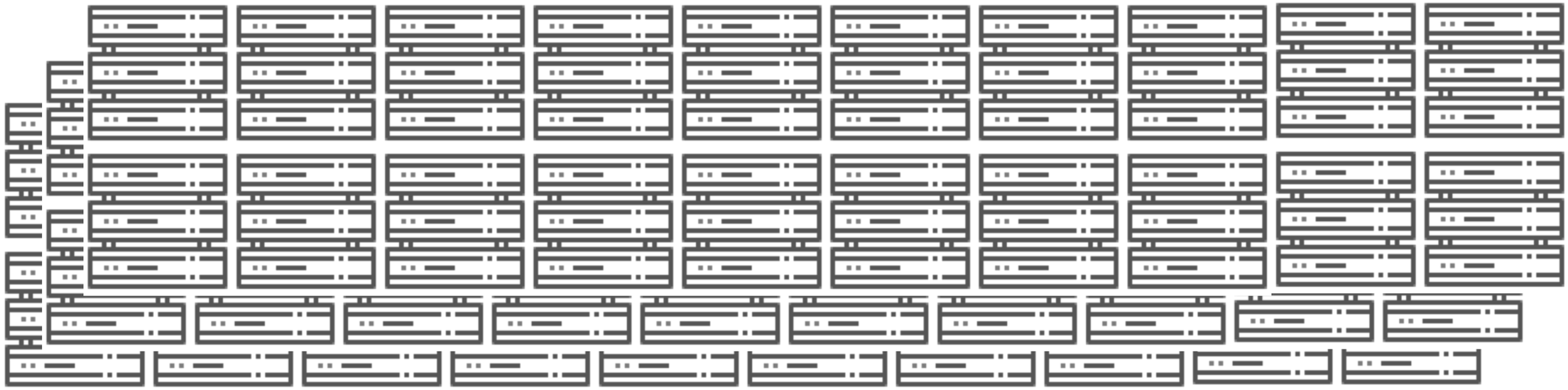


1. 왜 SRE를 도입해야 하는가

몇대의 서버는 1명이 관리 가능

수십 ~ 수백대의 서버는 몇명이 관리 가능

그렇다면, 수천 ~ 수만대의 서버는?



1. 왜 SRE를 도입해야 하는가

서버는 무한히 늘어나도,

사람은 무한히 늘어날 수 없다.

1. 왜 SRE를 도입해야 하는가

DEVIEW 2019

SRE: Site Reliability Engineering

사이트 신뢰성 엔지니어링

3.

지진과 같은 재난상황에서는
기존의 대응방안이 동작하지 않는다

글로벌 스케일, 또는 데이터의 인터넷 서비스를 제공하면서 어떻게 하면
시스템의 **신뢰성**을 보장할 수 있을지 고민하는 기술 분야입니다. **방법론**, **문화**

네이버 검색 SRE팀의 정의

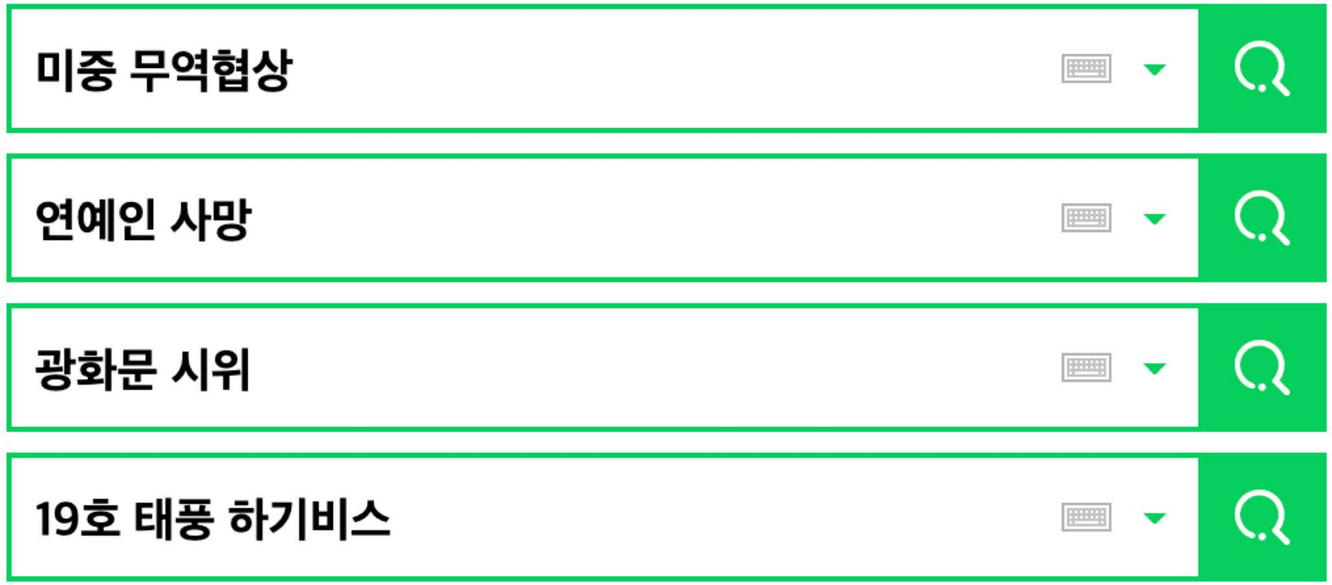
신뢰성: 네이버 검색 서비스가 정상적으로 제공되는 것

방법론: Availability, Capacity Planning, Monitoring, Contingency Engineering

문화 : Post-Mortem, Blameless, Fact-Based

1. 왜 SRE를 도입해야 하는가

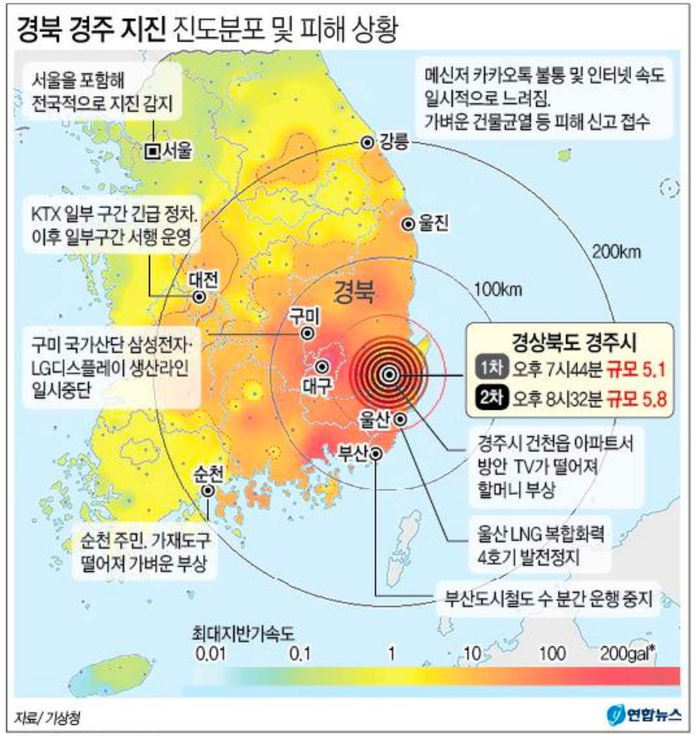
DEVIEW
2019



대한민국의 모든 이슈가 트래픽에 반영

1. 왜 SRE를 도입해야 하는가

DEVIEW 2019



장예진 기자 / 20160912
트위터 @yonhap_graphics, 페이스북 tuney.kr/LeYN1

2016년 9월 경주 지진

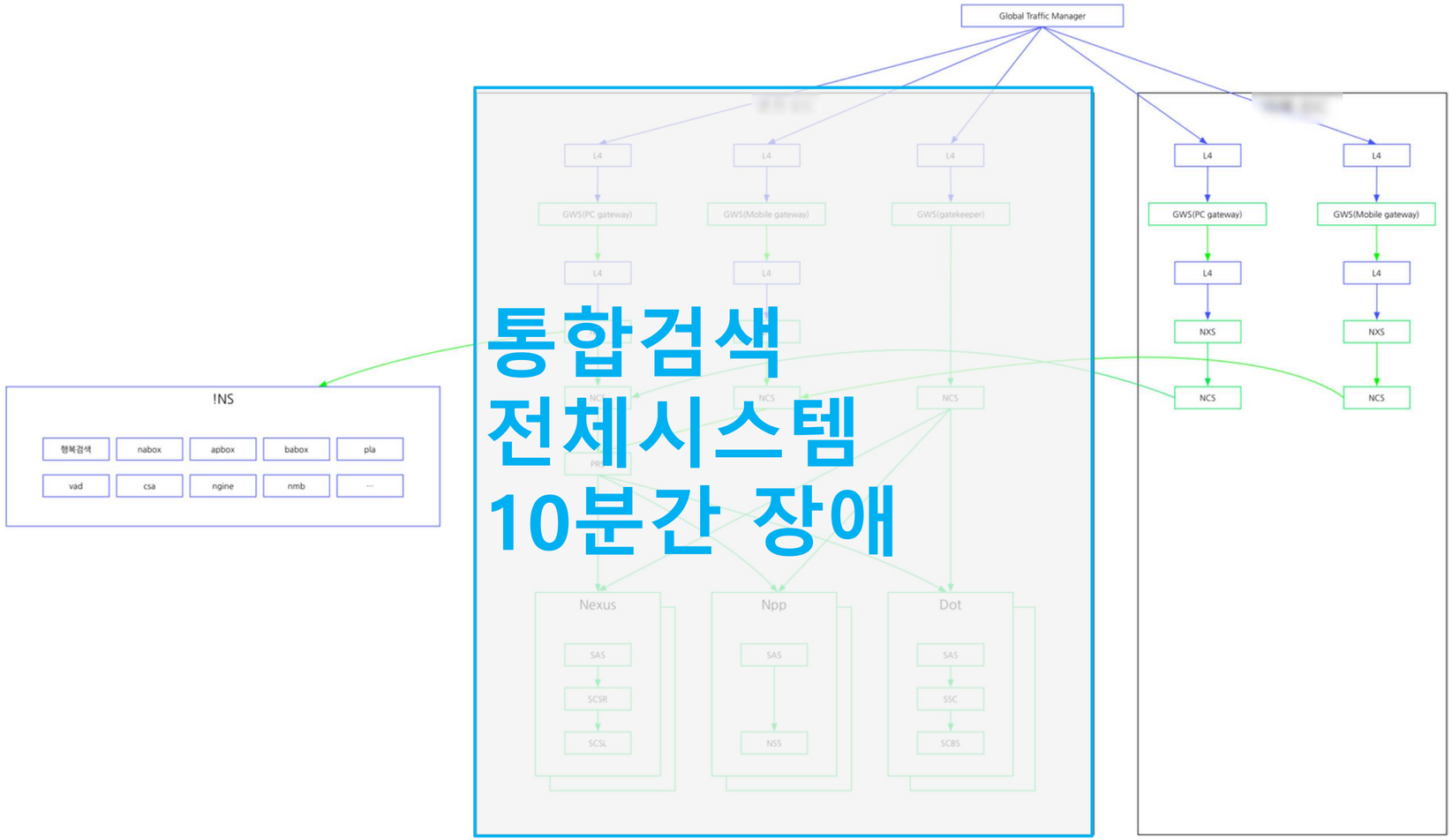
지진

경주 지진

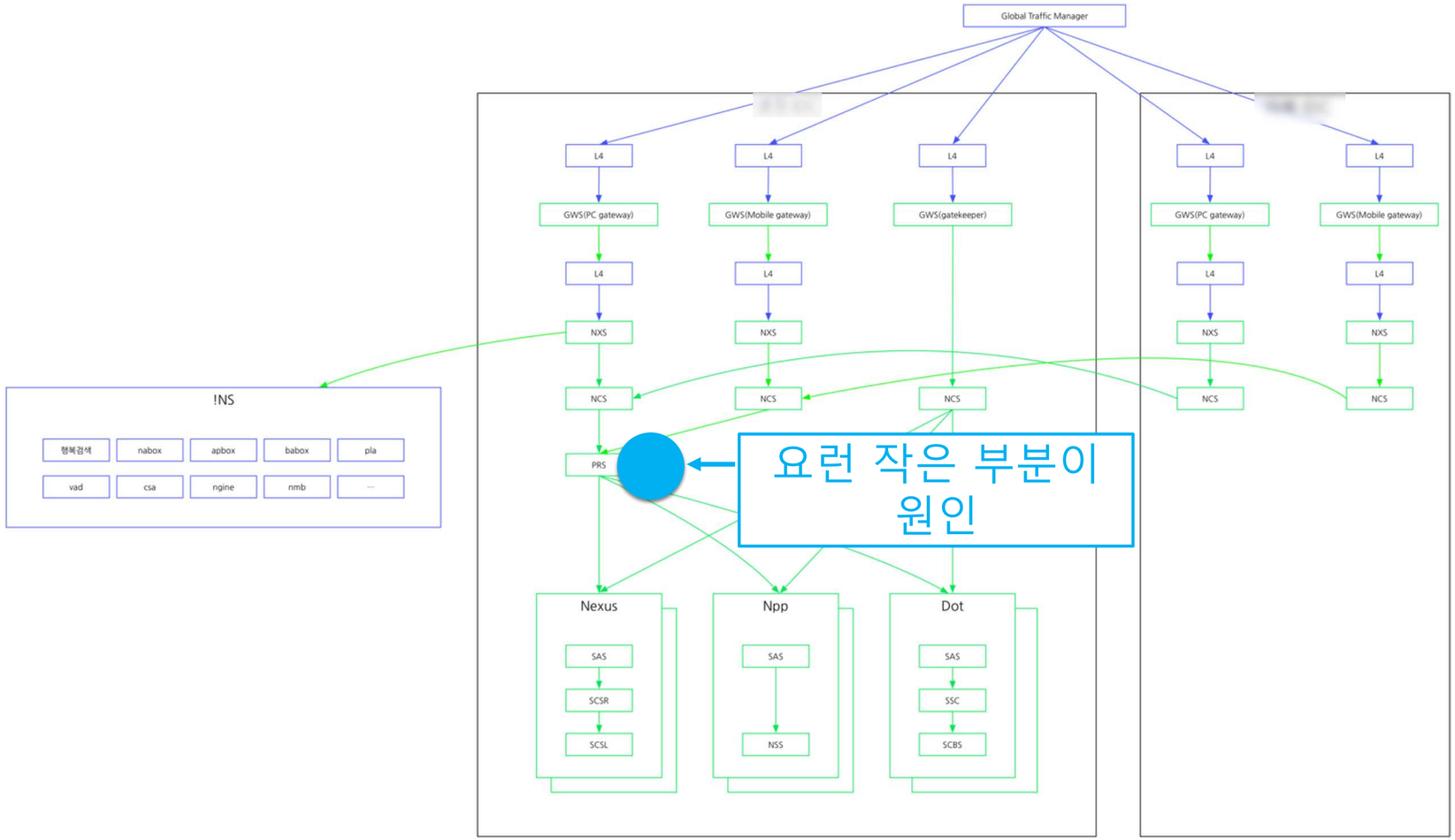
재난문자

1. 왜 SRE를 도입해야 하는가

DEVIEW
2019



1. 왜 SRE를 도입해야 하는가



1. 왜 SRE를 도입해야 하는가

DEVIEW
2019

장애 탐지	:	2분
증상 완화	:	20분
원인 파악	:	40분
장애 영향도 파악	:	1시간
장애 복구 완료	:	수시간
포스트모텀 작성	:	48시간

1. 왜 SRE를 도입해야 하는가

기존의 DevOps 방법론으로는

규모도, 재난도 감당할 수 없는 현실

SRE의 필요성을 절실히 느끼고

2016년 말 도입을 결정

DEVIEW
2019

2. 우리들의 SRE

2. 우리들의 SRE


kubernetes


ZIPKIN


Prometheus


Stack


pagerduty


Grafana


TensorFlow

2. 우리들의 SRE

우리에게 중요한 건

~~Tool~~이 아니라

현실에 닥친 문제를 해결하는 것

2. 우리들의 SRE

문제점

1. 비상상황 **대응 체계**의 부재
2. 장애 발생 **Detection**의 어려움
3. **Post-Mortem 문화** 부재
4. 전체 시스템 **현황** 확인 불가

2. 우리들의 SRE

DEVIEW
2019

문제점

1. 비상상황 대응 체계의 부재

- 비상 상황(Contingency)이 무엇인지 정의가 없음
- 비상 상황에 어떤 행동을 해야 하는지 대응 체계가 없음

검색 contingency plan

개요

- 2가지 관점의 contingency plan 운영 (이하 CP)
  - 경쟁사 서비스 불능에 대비한 CP_1
  - 국가적 이슈 발생에 대비한 CP_2

CP_1

- 정의
  - daum, google 등의 경쟁사 서비스 마비로 naver 검색 유입이 증가
  - 검색 순유입량(unique query) 폭증
  - 장시간 지속 가능
  - 검색시스템 전체적인 부하 증가로 이어짐
- 대비
  - N배 가용량 확보를 위한 scale out 정책 수립
  - layer 간 cache server 도입
  - IDC 이중화
  - 통합검색 일간 peak traffic 기준 [redacted]배의 가용률 준비
    - 모든 경쟁사 outage에 대응 가능한 수준
- CP_1 발동
  - peak 기준 [redacted]배 이상의 traffic 유입시 CP_1 발동
  - 성능 병목이 예상되는 검색서비스 판단하여 통검 결과에서 제외 준비
  - 단계적 결과 제외로 검색시스템 신뢰성 유지

CP_2

- 정의
  - 자연 재해를 비롯한 국가적 이슈 발생으로 naver 검색 유입이 증가
  - 동일한 검색어 유입량(repetitive query) 폭증
  - 이슈 발생 시점의 burst한 traffic 유입
  - 검색시스템 Front-end layer에 부하 집중
- 대비
  - 9/12 경주지진 계기로 CP_2 대응 시작
  - 서버의 역할(role) 별로 peak 기준 최대가용률 파악
  - 병목 지점의 scale out 정책 수립
  - 통합검색 일간 peak traffic 기준 [redacted]배의 가용률 준비
    - 경주지진 발생 시점의 traffic 유입에 대응 가능
- CP_2 발동
  - 국가 차원의 이슈 발생시 CP_2 발동
  - 성능 병목이 예상되는 role의 scale out 수행하여 peak 기준 [redacted]배 가용률 확보
  - CP_2 발동 중 CP_1 상황이 동시 발생할 경우 통검 결과 선별적 노출

DEVIEW 2019

부재

경쟁사 서비스 장애로 인한

트래픽 유입 집중상황 CP_1

을 해야 하는지 대응 체계가 없음

검색 contingency plan

개요

- 2가지 관점의 contingency plan 운영 (이하 CP)
  - 경쟁사 서비스 불능에 대비한 CP_1
  - 국가적 이슈 발생에 대비한 CP_2

CP_1

- 정의
  - daum, google 등의 경쟁사 서비스 마비로 naver 검색 유입이 증가
  - 검색 순유입량(unique query) 폭증
  - 장시간 지속 가능
  - 검색시스템 전체적인 부하 증가로 이어짐
- 대비
  - N배 가용량 확보를 위한 scale out 정책 수립
  - layer 간 cache server 도입
  - IDC 이중화
  - 통합검색 일간 peak traffic 기준 [redacted]배의 가용률 준비
    - 모든 경쟁사 outage에 대응 가능한 수준
- CP_1 발동
  - peak 기준 [redacted]배 이상의 traffic 유입시 CP_1 발동
  - 성능 병목이 예상되는 검색서비스 판단하여 통검 결과에서 제외 준비
  - 단계적 결과 제외로 검색시스템 신뢰성 유지

CP_2

- 정의
  - 자연 재해를 비롯한 국가적 이슈 발생으로 naver 검색 유입이 증가
  - 동일한 검색어 유입량(repetitive query) 폭증
  - 이슈 발생 시점의 burst한 traffic 유입
  - 검색시스템 Front-end layer에 부하 집중
- 대비
  - 9/12 경주지진 계기로 CP_2 대응 시작
  - 서버의 역할(role) 별로 peak 기준 최대가용률 파악
  - 병목 지점의 scale out 정책 수립
  - 통합검색 일간 peak traffic 기준 [redacted]배의 가용률 준비
    - 경주지진 발생 시점의 traffic 유입에 대응 가능
- CP_2 발동
  - 국가 차원의 이슈 발생시 CP_2 발동
  - 성능 병목이 예상되는 role의 scale out 수행하여 peak 기준 [redacted]배 가용률 확보
  - CP_2 발동 중 CP_1 상황이 동시 발생할 경우 통검 결과 선별적 노출

DEVIEW 2019

부재

국가적 관심사로 인한

폭발적인 트래픽 급증 CP_2

을 해야 하는지 대응 체계가 없음

2. 우리들의 SRE

DEVIEW
2019

문제점

2. 장애 발생 Detection의 어려움

- 장애를 detection하는 기준이 필요함
- 아주 쉬운 기준으로!

2. 우리들의 SRE

DEVIEW 2019

가용량 지표 개발

새로운 방식

부하증가배수

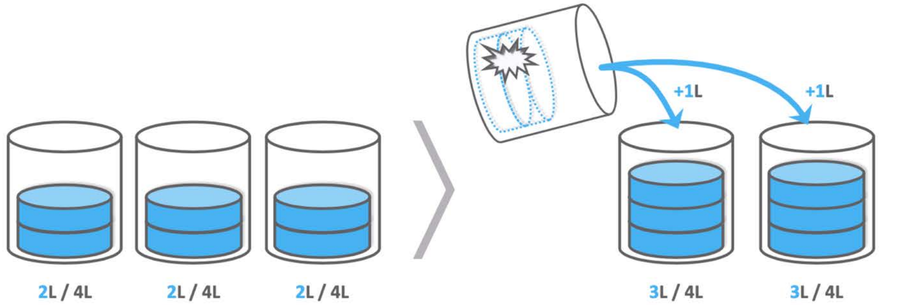
한 친구가 죽으면 나머지 친구들은 몇 배를 받나?

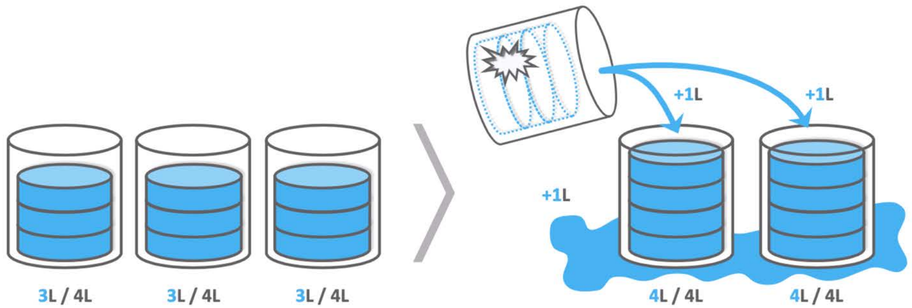
최대가용배수

한 명이 현재 몇 배까지 받을 수 있나?

"임계 상황" 판단

부하증가배수 > 최대가용배수





아주 쉬운 공식

가용량 지표를 정의, 특허 출원

가용량을 기반으로 한 **Alerting**

2. 우리들의 SRE

DEVIEW
2019

문제점

3. Post-Mortem 문화 부재

- 장애 복구 후 사후 처리 Report가 누락
- 비슷한 장애의 재발 시 대응 대책이 미비

NAVER-SEARCH / Defense-System-Outage / Defense-System-Out...

Unwatch 84 | Star 15 | Fork 0

Code | Issues 161 | Pull requests 0 | ZenHub | Projects 0 | Wiki | Insights | Settings

Filters | is:issue is:open | Labels 12 | Milestones 2 | New issue

161 Open | 49 Closed | Author | Labels | Projects | Milestones | Assignee | Sort

- 20191014 Palantir service alerter 메일 발송 누락
  #210 opened 3 days ago by minchul-kang | New Issues
- 20191010 오픈캐스트 서비스 코어 발생
  #209 opened 6 days ago by keyolk | New Issues
- 20191010 책검색 신규문서 지연 업데이트 이슈
  #208 opened 8 days ago by jiayou | New Issues
- 20191007 펜션검색 미노출 이슈 (TTS 5급) TTS — 3
  #207 opened 11 days ago by jinoh-kim | New Issues
- 20191001 동영상검색 nb장비 다운 이후 11번 파티션 장비 전체 다운 이슈 — 2
  #206 opened 11 days ago by dongsu-an | New Issues
- 20190927 네이버쇼핑 증감 데이터 미 반영 — 1
  #205 opened 14 days ago by ljh1156 | New Issues
- 20190925 국어/중국어사전 통검 api 전환 중 가용량 부족 — 1
  #204 opened 22 days ago by kj-choi | New Issues
- 20190906 LINE Japan Cuve SQL 추출 장애 fault
  #203 opened on 6 Sep by nowol | New Issues
- 20190905 LineOA Profile thumbnail 미노출 장애
  #202 opened on 5 Sep by kluster86 | New Issues
- 20190903 검색게이트웨이 블로그 nxcache 서버 메모리 사용량 증가로 인한 일부 요청 응답 성능 저하 — 1
  #201 opened on 3 Sep by jihoonc | New Issues
- 20190827 뉴스/리뷰 외 여러 서비스 타임아웃 발생 이슈 — 13
  #200 opened on 30 Aug by wj-lee | New Issues
- 20190828 카페 스키마 변경중 NSS 다운(TTS 3급) TTS — 2
  #199 opened on 29 Aug by hyeonjun-jo | New Issues
- 20190820 뮤직 배포중 core 다량 발생
  #198 opened on 21 Aug by mmii-kim | New Issues
- 20190802 연관어 인코딩 깨짐 장애 — 1
  #197 opened on 6 Aug by hannuriha | New Issues
- 20190724 지식인 서비스 모듈 검색 결과 미노출 (TTS 5급)
  #196 opened on 1 Aug by hangyoung-kim | New Issues
- 20190731 뮤직 PR PC 타이틀 링크 오류
  #195 opened on 1 Aug by mmii-kim | New Issues

DEVIEW 2019

2016년 말

Post-Mortem 작성 포맷 공유

처음 2–3 개월은 SRE가 작성

사후 분석을 습관화하는 문화를 전파

현재는 자율적 && 상시 운영

2. 우리들의 SRE

DEVIEW
2019

문제점

4. 전체 시스템 현황 확인 불가

- 다양한 서비스의 연계 장애 및 전체 현황 파악 불가
- Detection을 통해서 대응 필요시 Alerting이 필요함

2. 우리들의 SRE

문제점

4. 전체 시스템 현황 확인 불가

- 다양한 서비스의 연계 장애 및 전체 현황 파악 불가
- Detection을 통해서 대응할 필요성 Alerting이 필요함

전체 현황을 파악하는

시스템은 어찌 만들어야 하지?

3. Metric + Meta Data = Insight

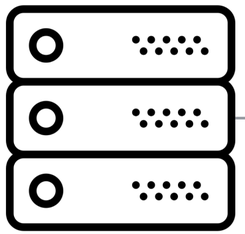
기존에도,

Monitoring은 하고 있었다.

Metric도 잘 쌓고 있었다.

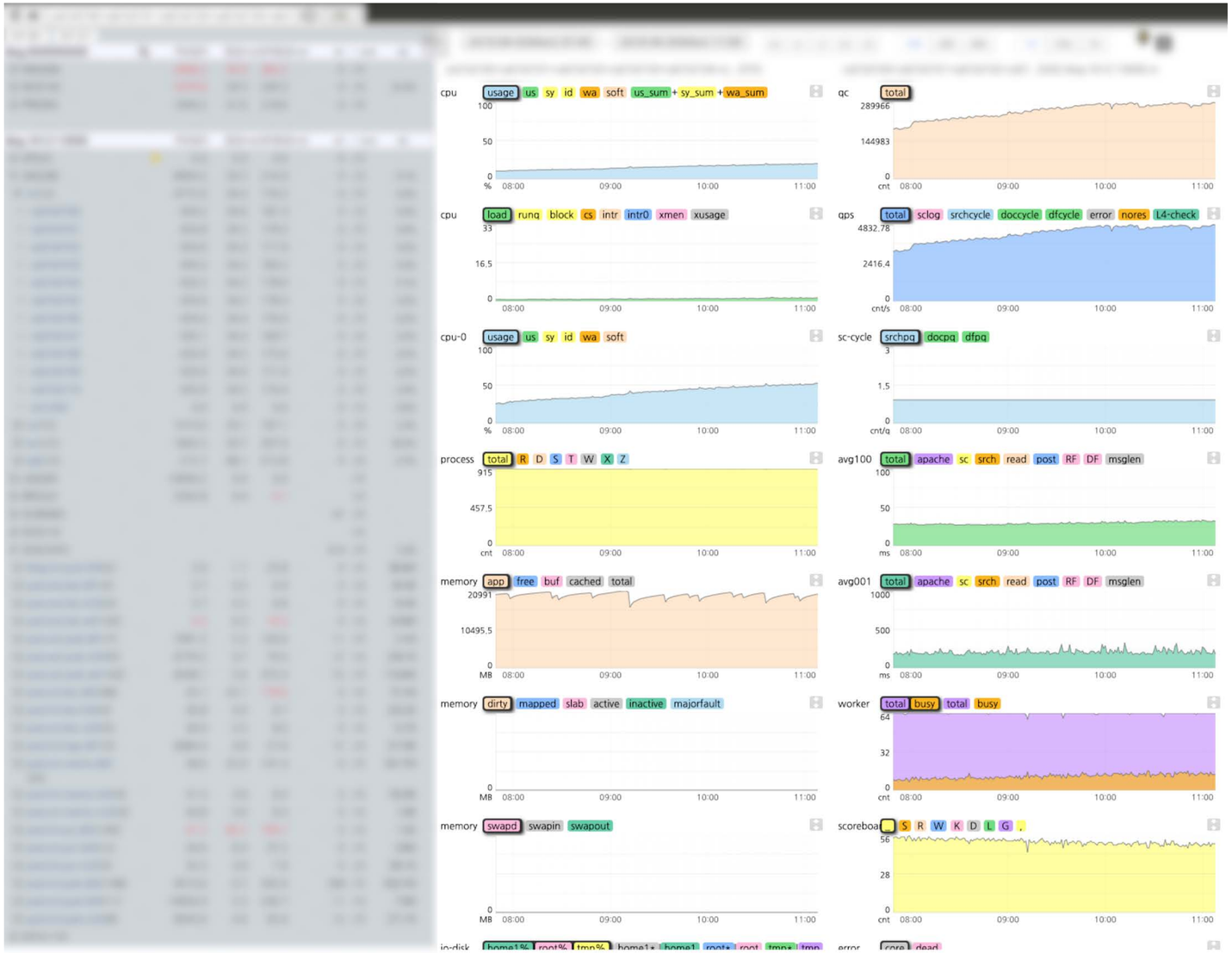
3. Metric + Meta Data = Insight

DEVIEW 2019



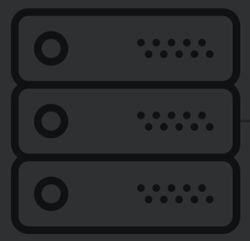
By Node Metric

- CPU Usage
- CPU-0
- MEMORY Usage
- DISK
- IO-RW
- Network
- Error
- Status Code
- Cache Hit Ratio
- Worker
- Core
- Document Count
- .. 그 외 수십가지 지표들



3. Metric + Meta Data = Insight

DEVIEW 2019



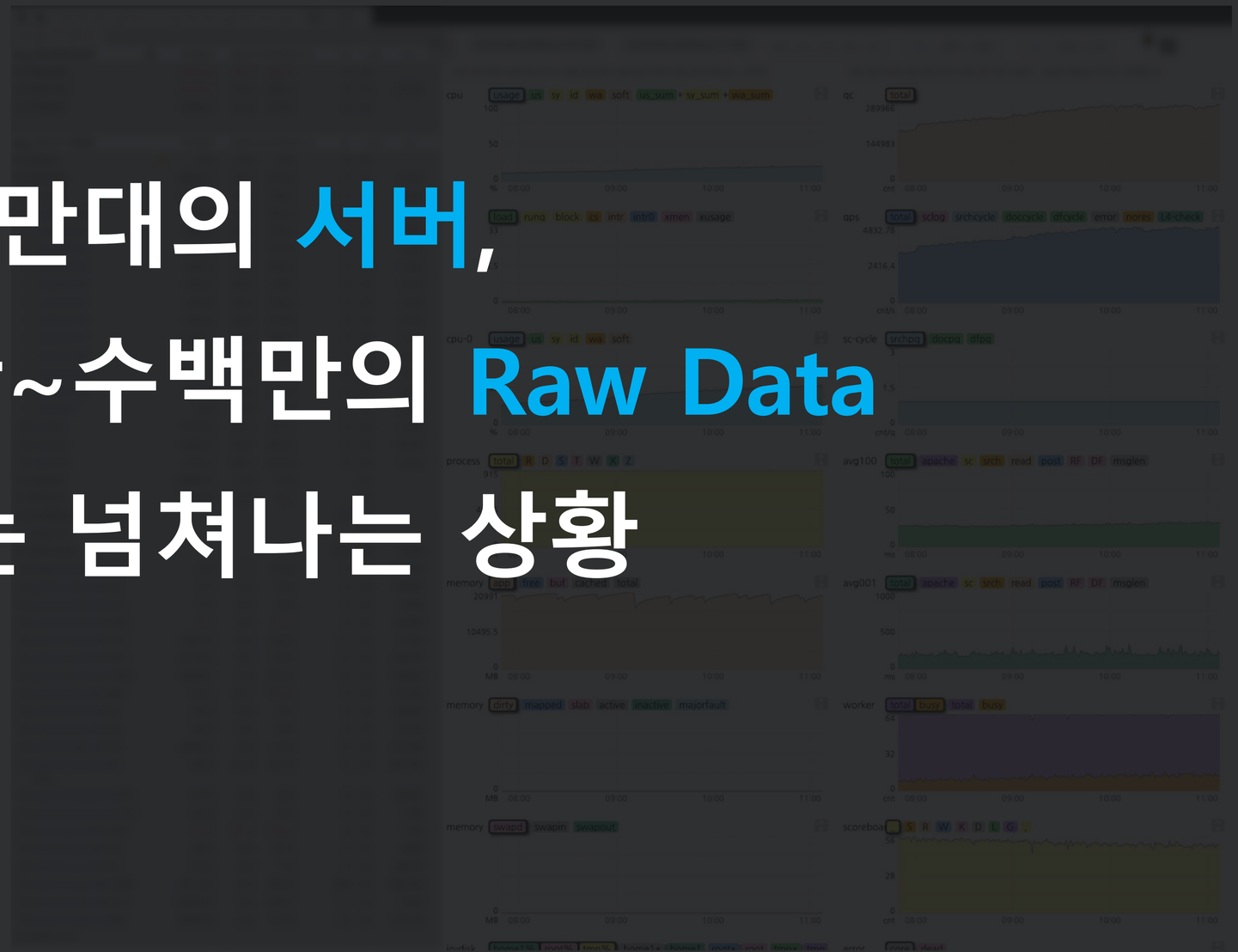
By Node Metric

- CPU Usage
- CPU-0
- MEMORY Usage
- DISK
- IO-RW
- Network
- TCP
- Status Code
- Cache Hit Ratio
- Worker
- Core
- Document Count
- .. 그 외 수십가지 지표들

수만대의 **서버**,

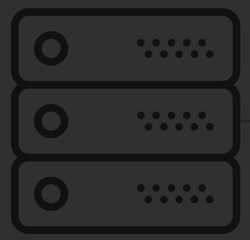
분당 수십만~수백만의 **Raw Data**

정보는 넘쳐나는 상황



3. Metric + Meta Data = Insight

DEVIEW 2019



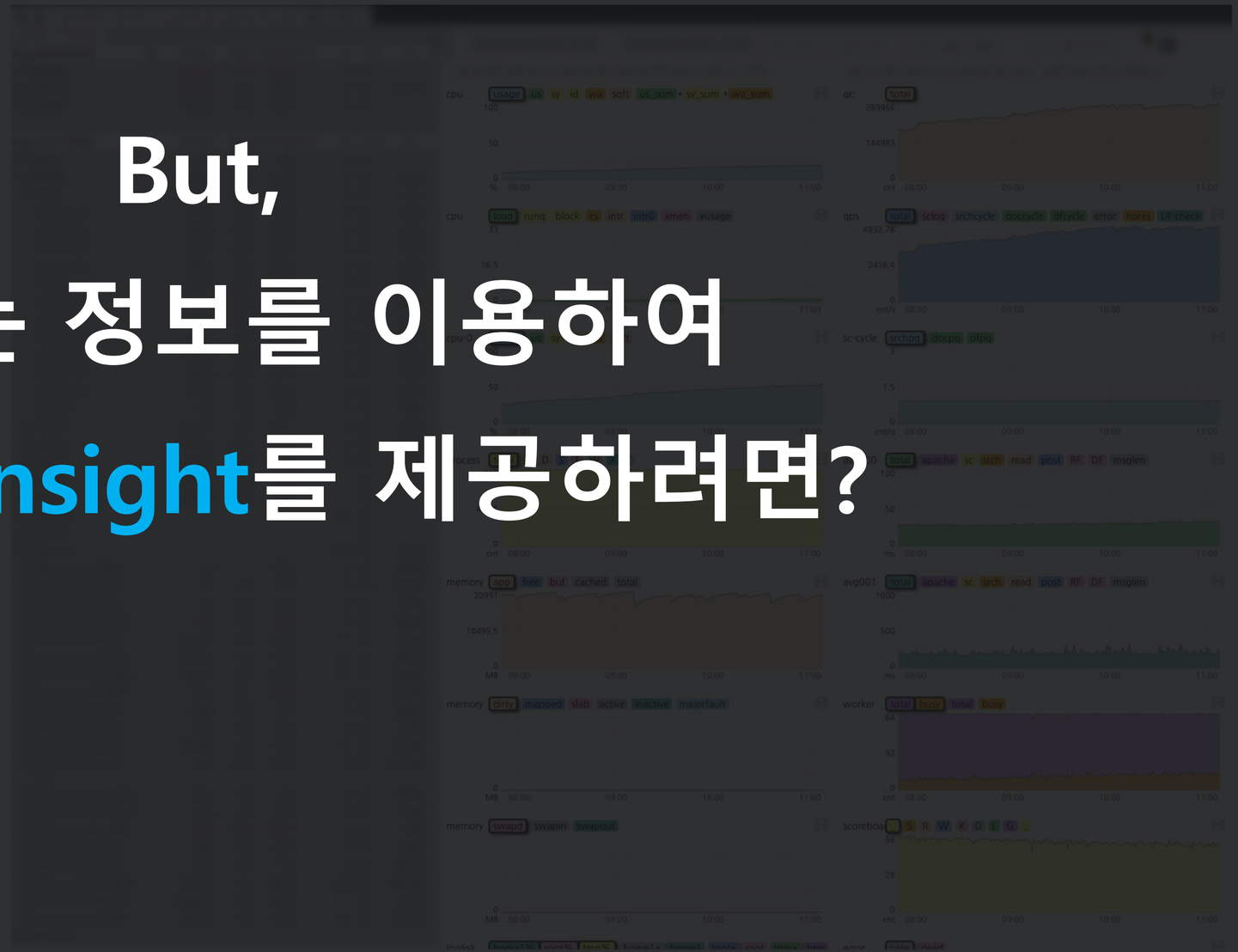
By Node Metric

- CPU Usage
- CPU-0
- MEMORY Usage
- DISK
- IO-R/W
- Network
- Error
- Status Code
- Cache Hit Ratio
- Worker
- Core
- Document Count
- .. 그 외 수십가지 지표들

But,

넘쳐나는 정보를 이용하여

꼭 필요한 Insight를 제공하려면?



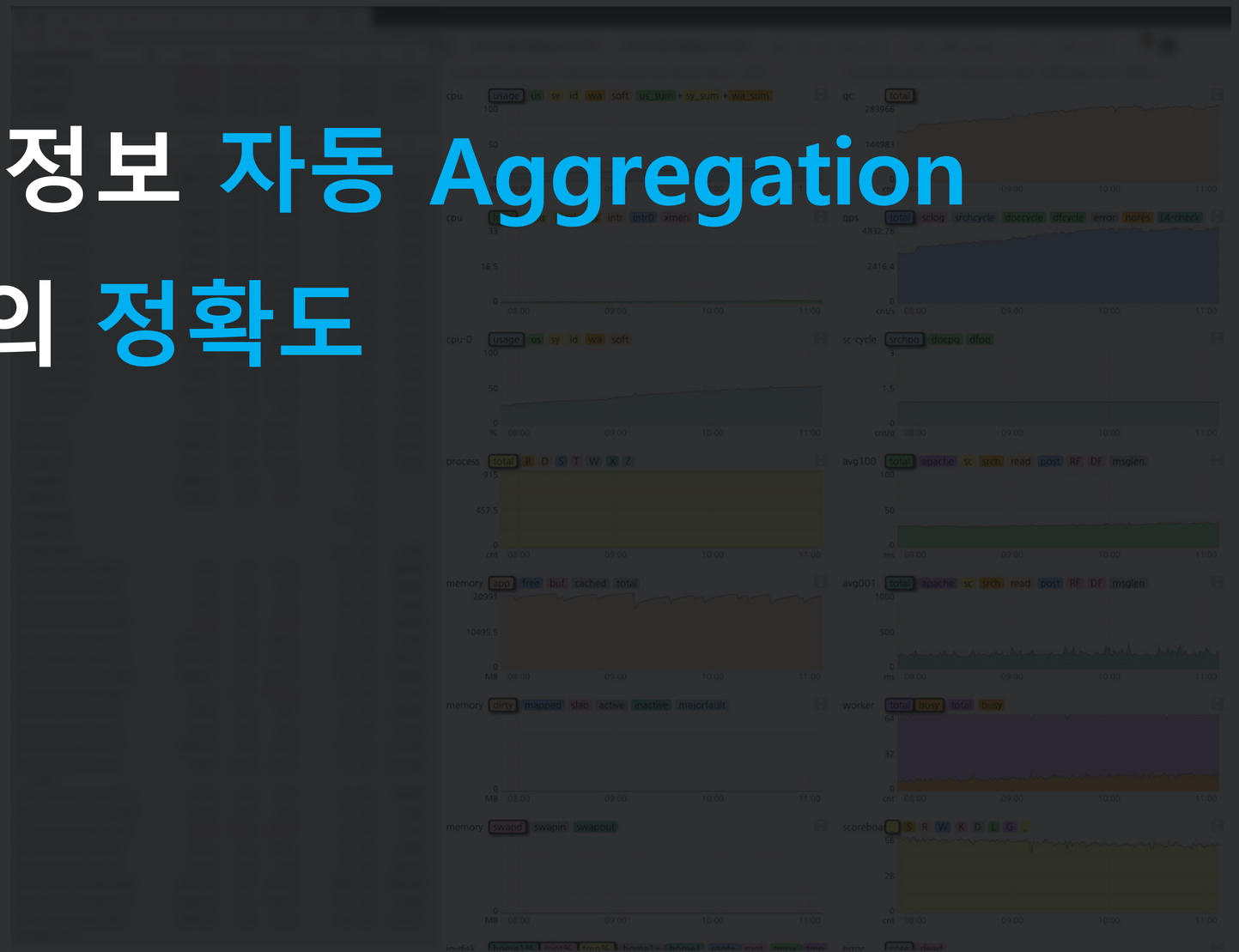
3. Metric + Meta Data = Insight

DEVIEW 2019

1. 서비스 단위 정보 자동 Aggregation
2. 담당자 정보의 정확도

By Node Metric

- CPU Usage
- CPU IO
- MEMORY Usage
- DISK
- IO-BW
- Network
- Error
- Status Code
- Cache Hit Ratio
- Worker
- Core
- Document Count
- .. 그 외 수십가지 지표들



3. Metric + Meta Data = Insight

(1). SSUID (서비스 유니크 아이디) 발급

Needs: 복잡한 구성의 서비스를 **묶어서** 가시화 하고싶다.

Solution : 서비스에 **유니크한 아이디**를 부여, 모든 계층을 묶음으로 가시화

지역 검색

- Cache Layer : local, local_openapi
- View Layer : locallink, localnx
- Search Layer : m_local, nx_local, openapi_local, svc_local
- Etc : local_xxx

[Aggregation] →

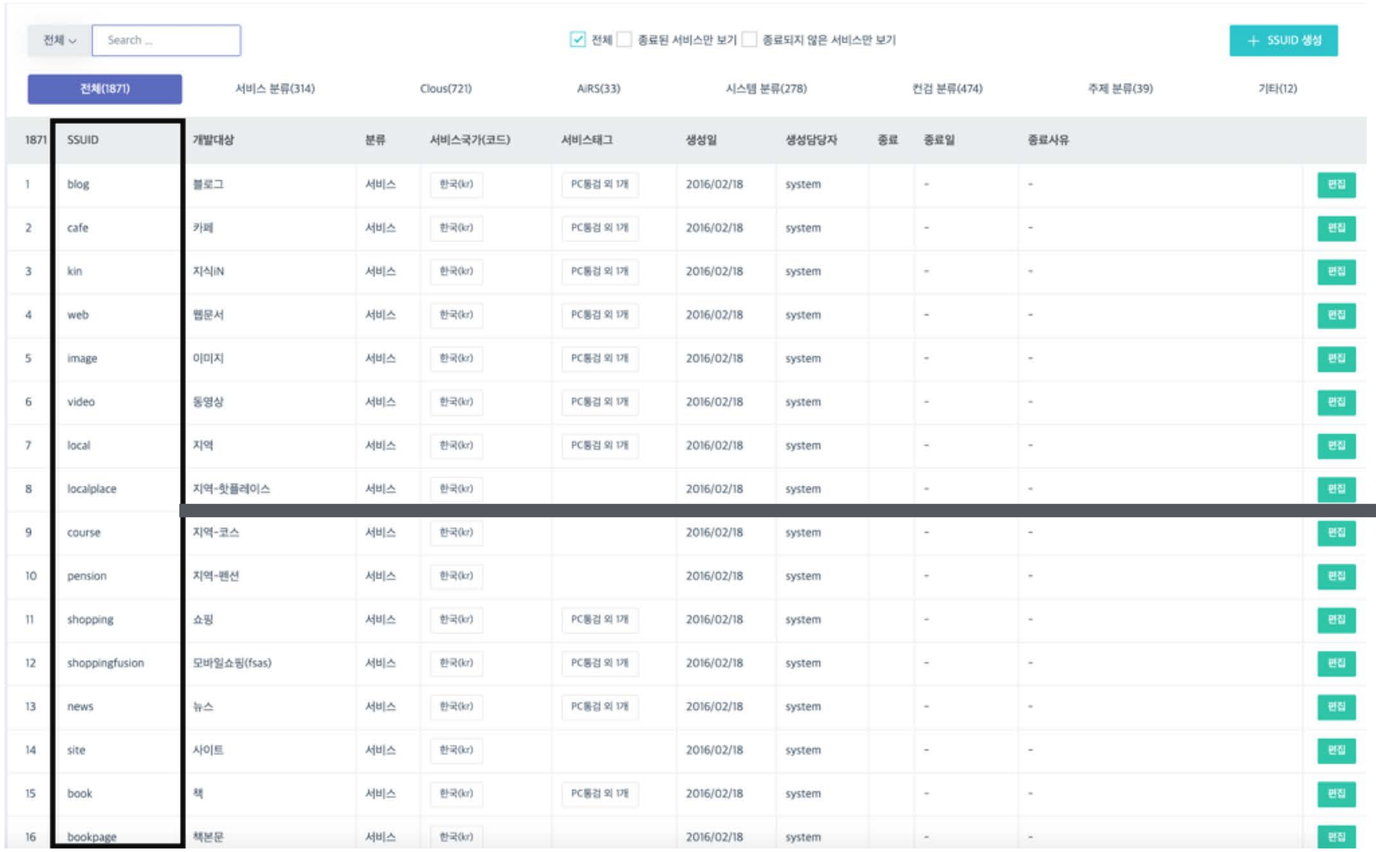
SSUID: local

(서비스 유니크 아이디)

3. Metric + Meta Data = Insight

DEVIEW 2019

(1). SSUID (서비스 유니크 아이디) 발급



1871	SSUID	개발대상	분류	서비스국가(코드)	서비스태그	생성일	생성담당자	종료	종료일	종료사유	
1	blog	블로그	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
2	cafe	카페	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
3	kin	지식iN	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
4	web	웹문서	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
5	image	이미지	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
6	video	동영상	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
7	local	지역	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
8	localplace	지역-핫플레이스	서비스	한국(kr)		2016/02/18	system		-	-	편집
9	course	지역-코스	서비스	한국(kr)		2016/02/18	system		-	-	편집
10	pension	지역-펜션	서비스	한국(kr)		2016/02/18	system		-	-	편집
11	shopping	쇼핑	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
12	shoppingfusion	모바일쇼핑(fsas)	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
13	news	뉴스	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
14	site	사이트	서비스	한국(kr)		2016/02/18	system		-	-	편집
15	book	책	서비스	한국(kr)	PC통검 외 1개	2016/02/18	system		-	-	편집
16	bookpage	책본문	서비스	한국(kr)		2016/02/18	system		-	-	편집

SSUID

1900	SSUID	개발대상
1	blog	블로그
2	cafe	카페
3	kin	지식iN
4	web	웹문서

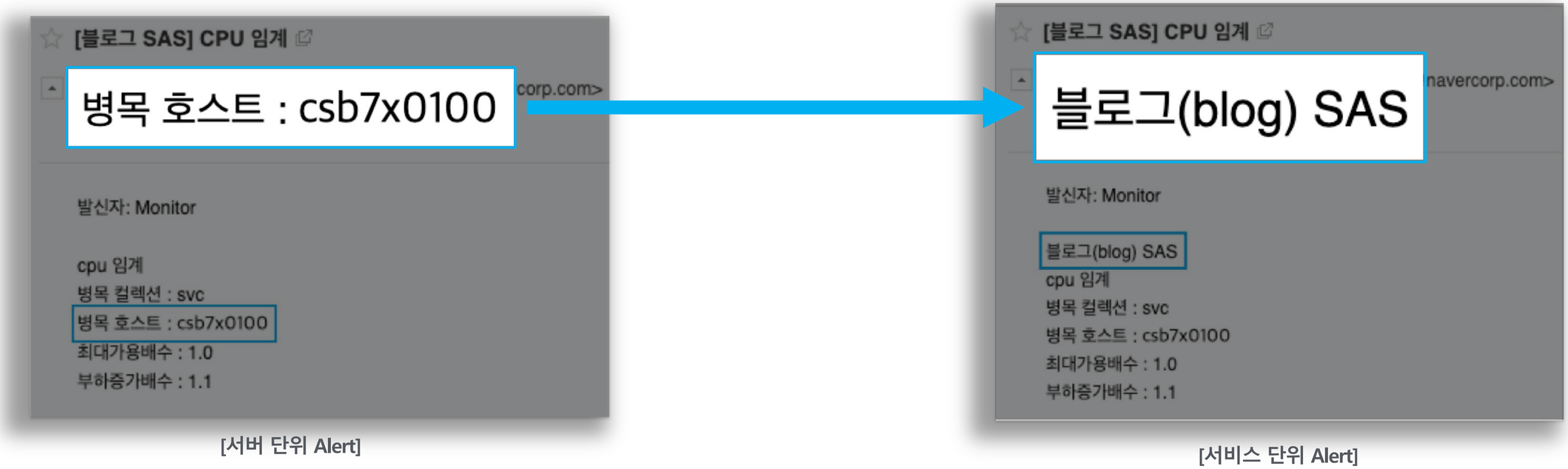
3. Metric + Meta Data = Insight

(1). SSUID (서비스 유니크 아이디) 발급

서비스 유니크 아이디를 정리하니,

[서버] 단위의 Alerting이 아니라

[서비스] 단위의 Alerting이 가능!



[서버 단위 Alert]

[서비스 단위 Alert]

3. Metric + Meta Data = Insight

(2). DRI (서비스 담당자 정보)

Needs: 1. 경보를 **서비스 담당자**에게 직접 보내야 함

2. 그런데 담당자 정보는 언제나 **outdated**

Solution : 경보에 특화된 **Meta Data API**를 만들고, **강제성 부여**

3. Metric + Meta Data = Insight

DEVIEW 2019

(2). DRI (서비스 담당자 정보)

local

비고에 Auto 표시가 있는 DRI는 장비정보에 의해 자동 생성된 DRI 입니다. 장비정보에 따라 삭제될 수 있습니다. DRI가 삭제되길 원하지 않으시다면 Auto를 한번 눌러주세요.

20	SSUID	개발대상	담당자	조직	
1	local	지역	조지선		
2	local	지역	구동현		
3	local	지역	최지훈	AiRSPACE	
4	local	지역	강영길	Local Search	
5	local	지역	정승재	POI Refinement	
6	local	지역	최지훈	AiRSPACE	
7	local	지역	이해웅	POI Refinement	
8	local	지역	조영재		
9	local	지역	이상서 장비담당	System&Solution	
10	local	지역	박연주		운영
11	local	지역	유준현	Local Search	모델링
12	local	지역	조한성	System&Solution	PRS
13	local	지역	이해웅 장비담당	POI Refinement	DCB
14	local	지역	박일순		운영
15	local	지역	주승규	System&Solution	SAS
16	local	지역	윤동빈		PRS
17	local	지역	이원진 장비담당	System&Solution	PRS

20	SSUID	개발대상	담당자	조직
1	local	지역	조지선	
2	local	지역	구동현	
3	local	지역	최지훈	AiRSPACE

DRI

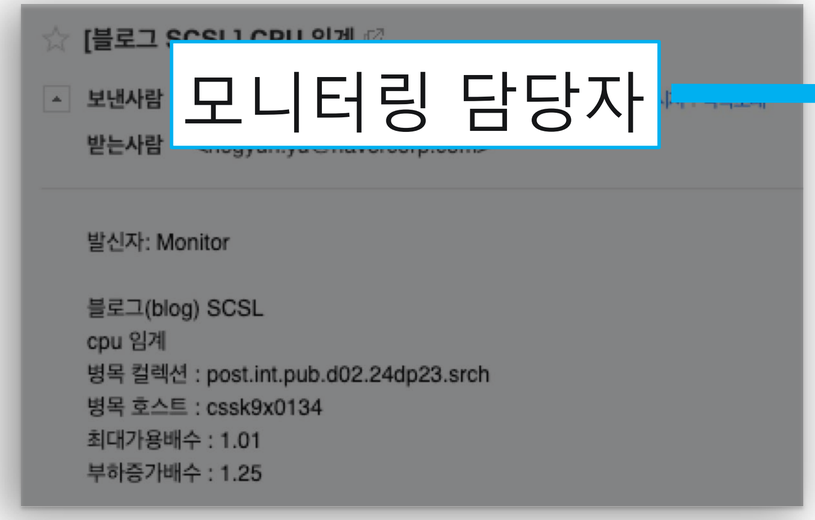
3. Metric + Meta Data = Insight

(2). DRI (서비스 담당자 정보)

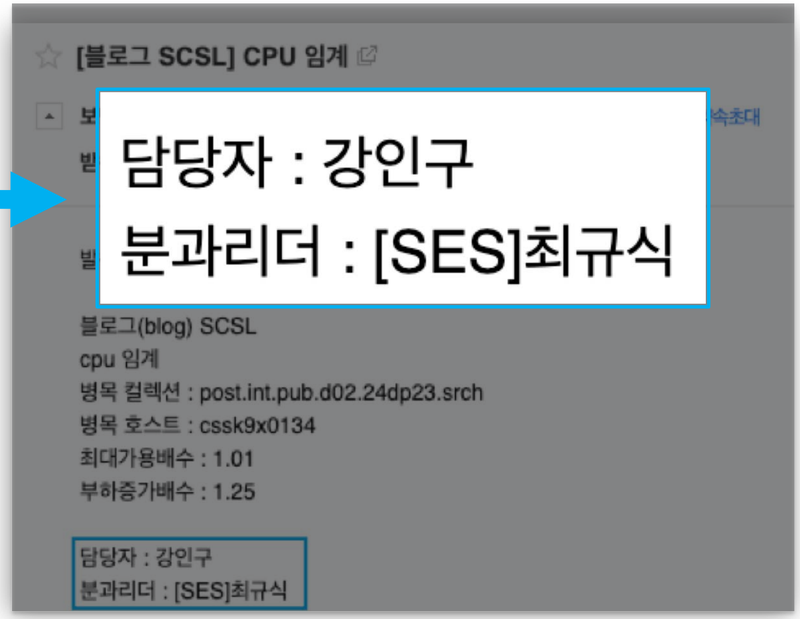
경보를 위한 담당자 정보를 정리하니,

[모니터링 담당자] 가 아니라

[서비스 담당자] 가 경보를 수신



[모니터링 담당자 Alert]



[서비스 담당자 Alert]

3. Metric + Meta Data = Insight

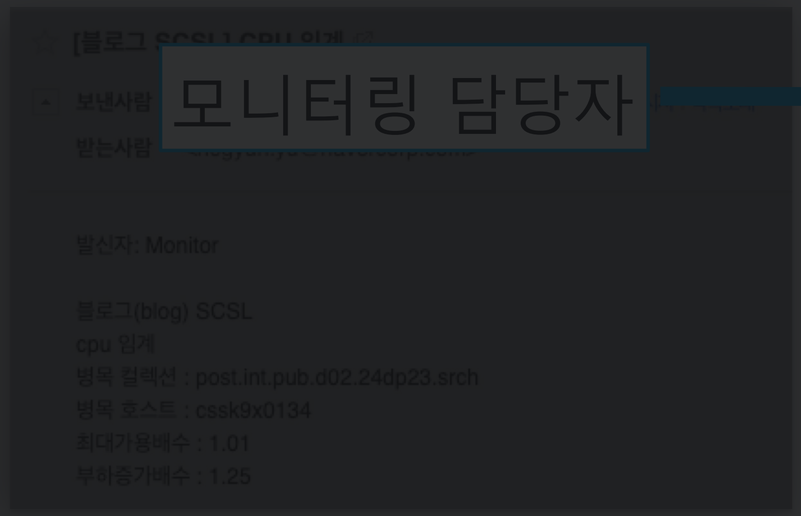
(2). DRI (서비스 담당자 정보)

경보를 위한 담당자 정보를 정리하니,

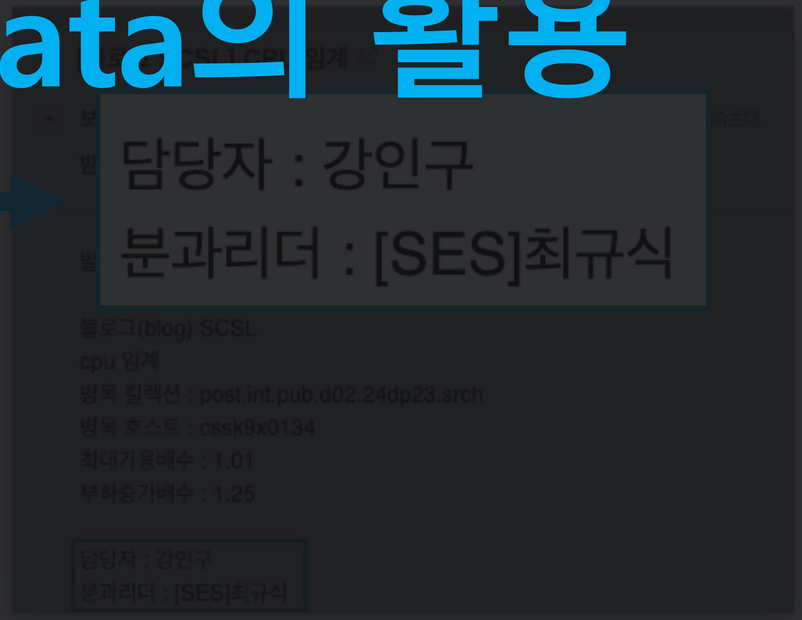
[모니터링 담당자] 가 아니라

[서비스 담당자] 에게 경보를 보내야 함을 알게 됨

다양한 Meta-Data의 활용



[모니터링 담당자 Alert]



[서비스 담당자 Alert]

3. Metric + Meta Data = Insight

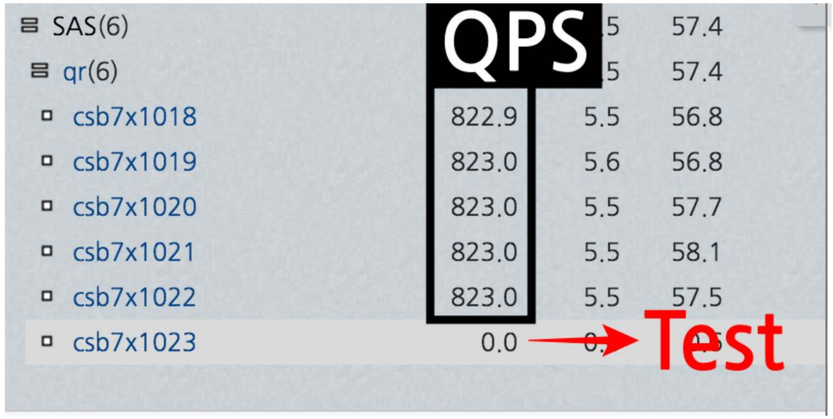
(3). Environment (장비 환경)

Problem: 수만대의 서버 중 테스트 서버가 섞여 있으면 **합산 지표**에 왜곡이 발생

Solution: dev장비, stage장비, test장비, **prod**장비 분류

Production Server

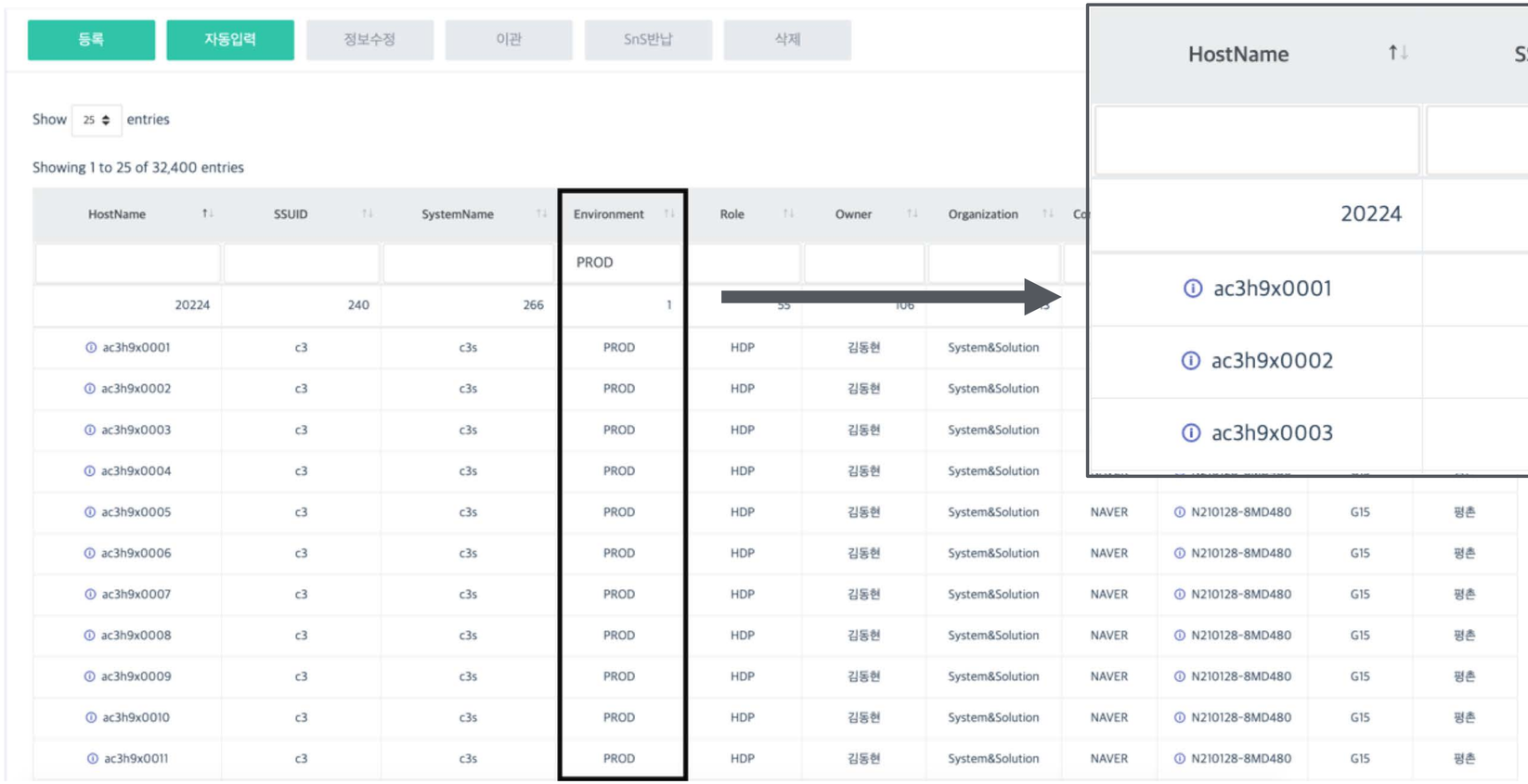
Test Server

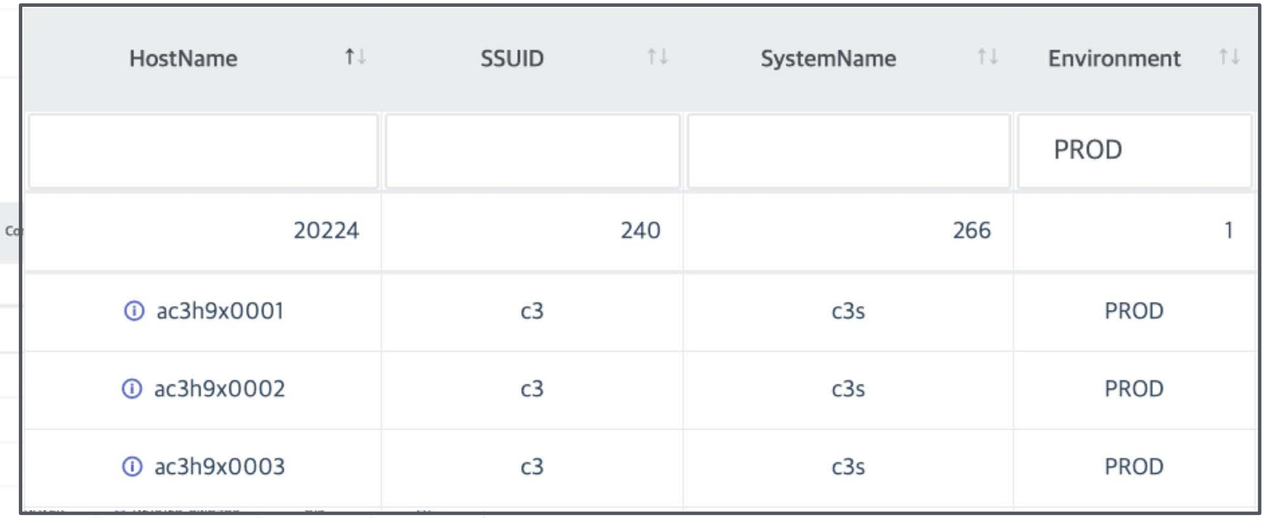


3. Metric + Meta Data = Insight

DEVIEW 2019

(3). Environment (장비 환경)





Environment

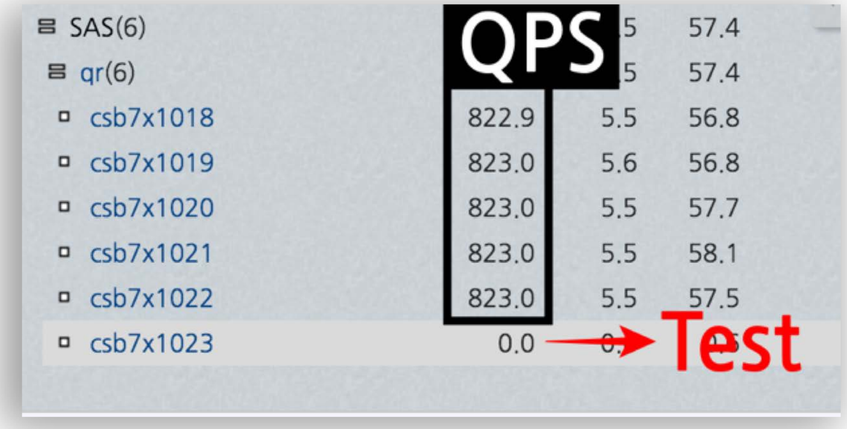
3. Metric + Meta Data = Insight

(3). Environment (장비 환경)

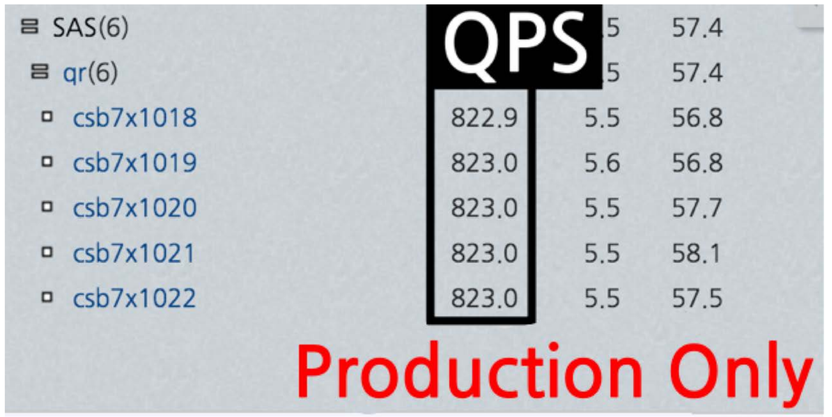
장비 Environment를 정리하니

[서비스 전체 합산 지표] 가 아니라

[Production 서비스 합산 지표] 를 알 수 있게 되었음



[Test를 포함한 합산 지표]



[Production 서비스 합산 지표]

3. Metric + Meta Data = Insight

DEVIEW 2019

(3). Environment (장비 환경)

장비 Environment를 정리하다

[서비스 전체 합산 지표] 가 아니라

[Production 서비스 합산 지표] 를 알 수 있게 되었음

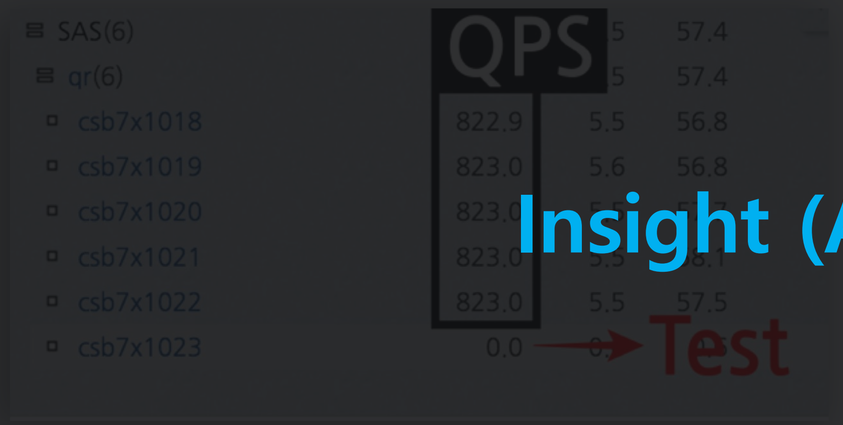
Metric (Raw Data)

+

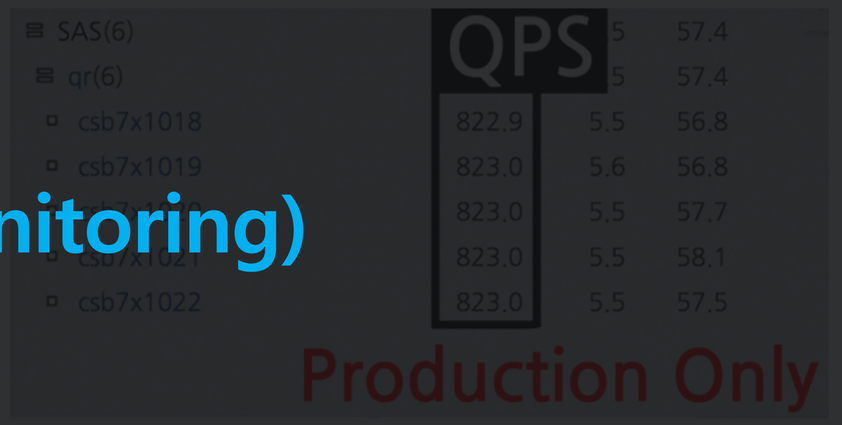
Meta Data (SSUID, DRI, Environment)

=

Insight (Advanced Monitoring)

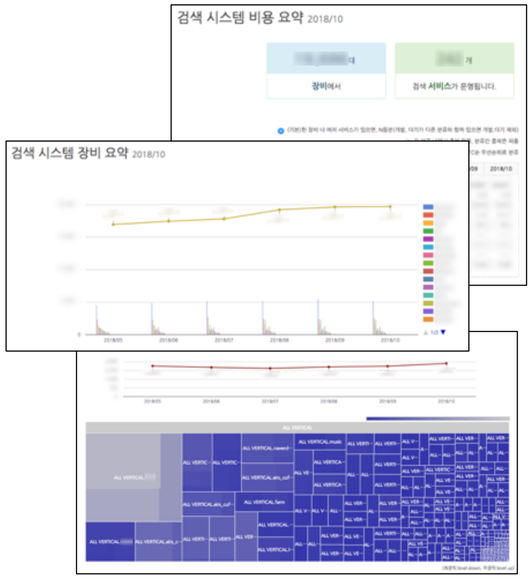


[Test를 포함한 합산 지표]

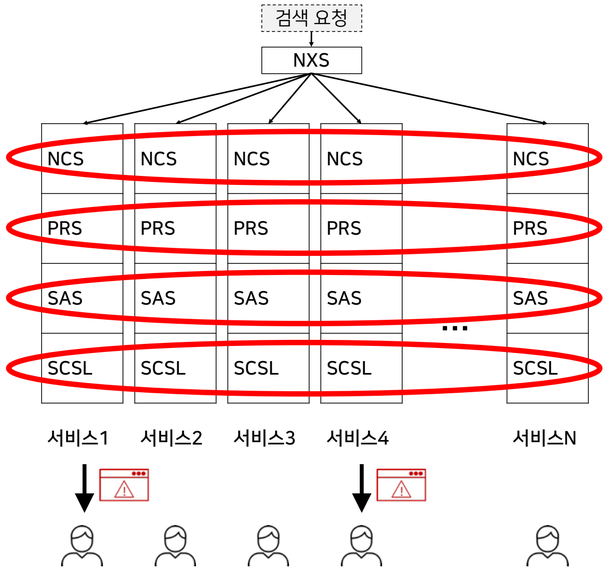


[Production 서비스 합산 지표]

3. Metric + Meta Data = Insight



Meta Data
Tagging, Mapping



서비스/계층 단위
Aggregation

3. Metric + Meta Data = Insight

DEVIEW 2019

365일 / 24시간

+

검색서비스 전체 관점

=

이상탐지 / 가시화 / 경보발송

3. Metric + Meta Data = Insight

새로운 도구

+

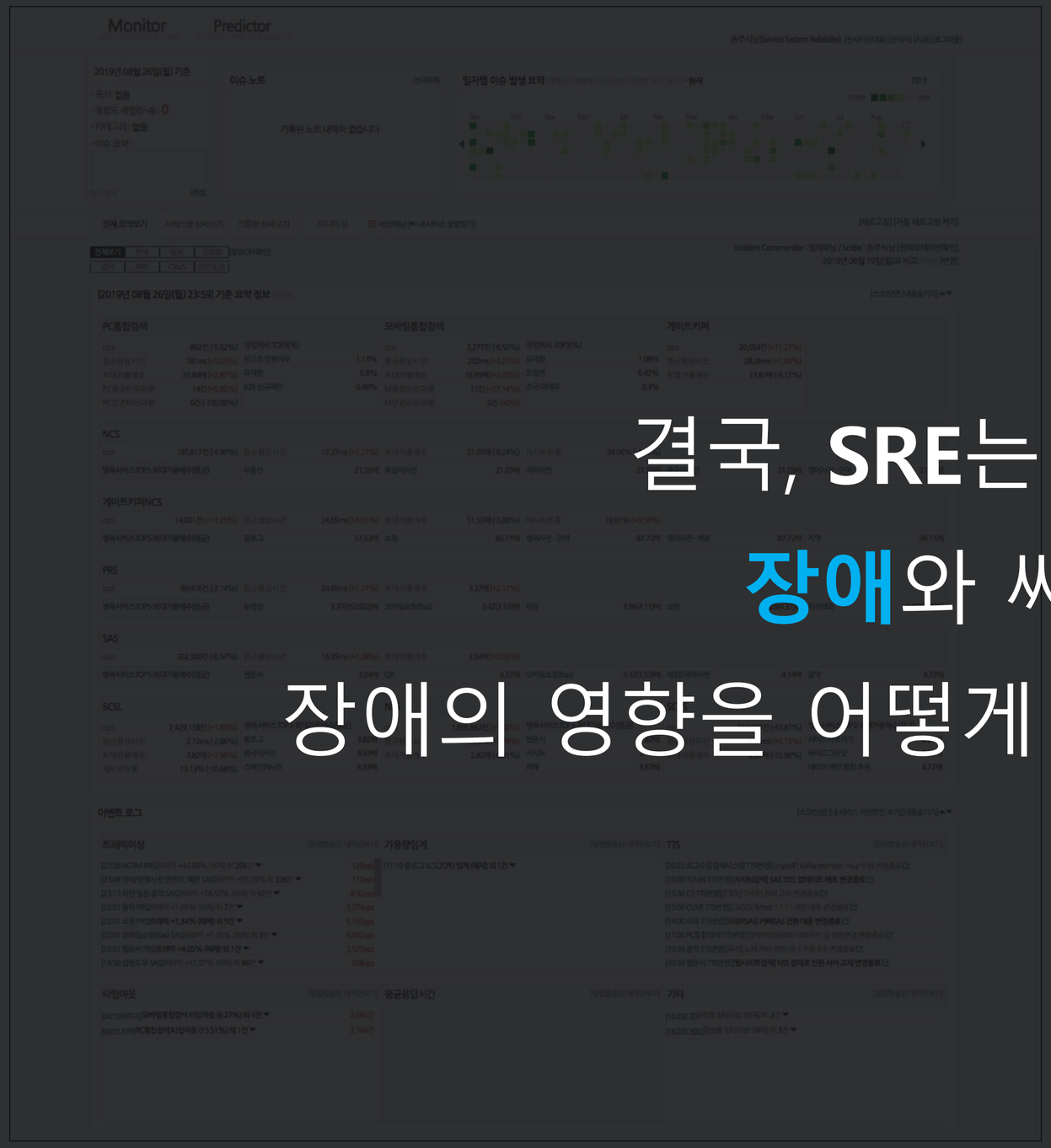
관찰

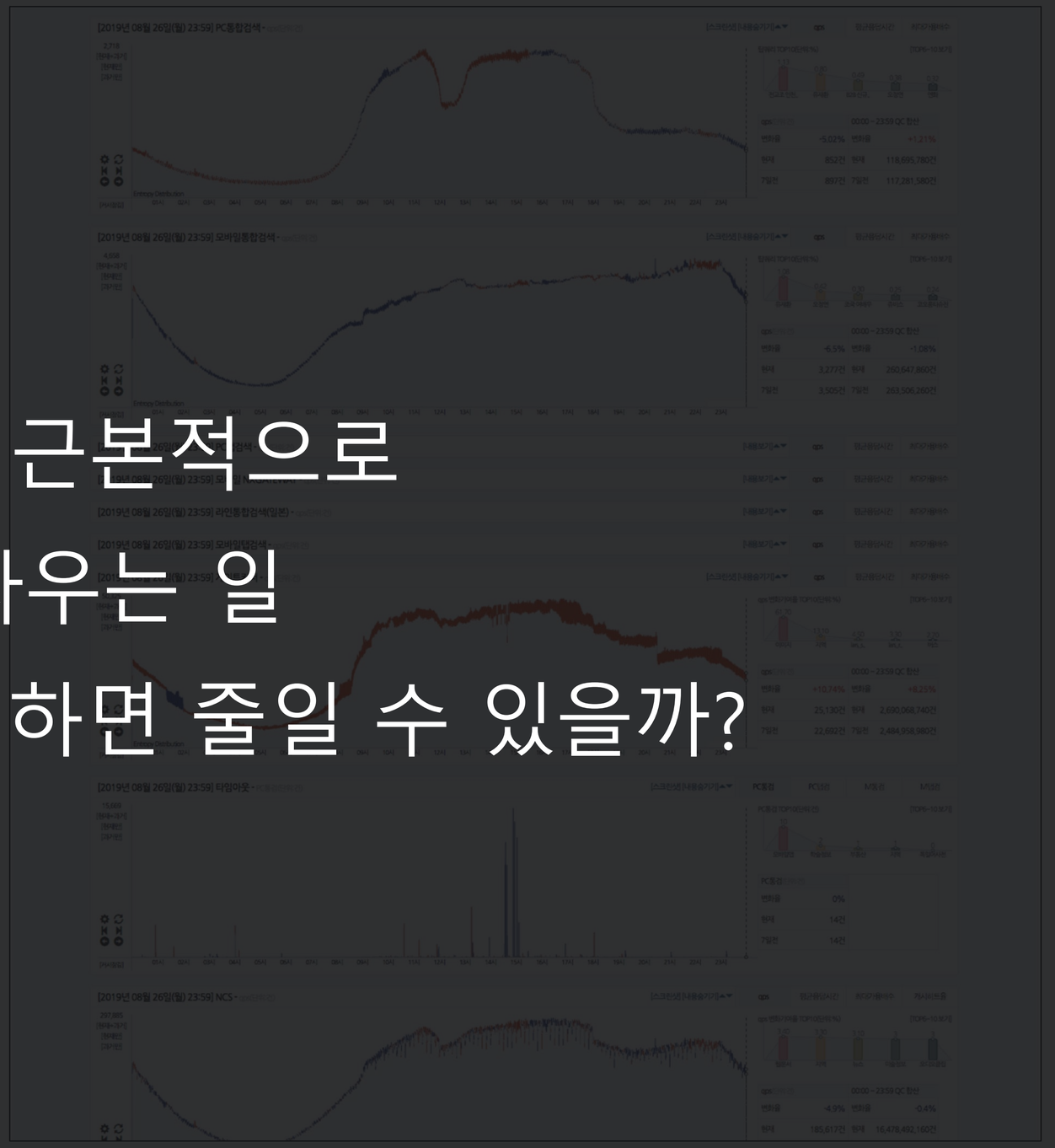
=

실패를 통한 개선

DEVIEW
2019

4. Fail Fast, Learn Faster



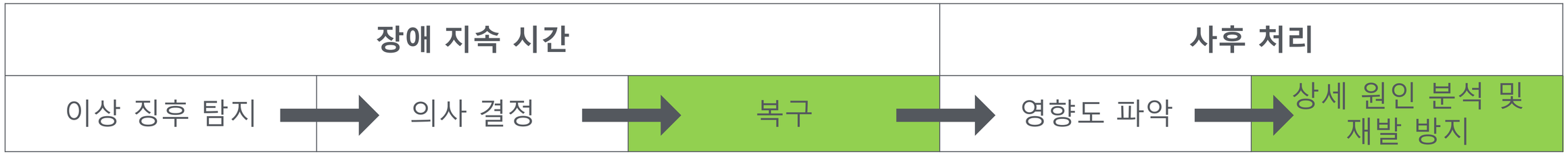


결국, **SRE**는 근본적으로

장애와 싸우는 일

장애의 영향을 어떻게 하면 줄일 수 있을까?

4. Fail Fast, Learn Faster

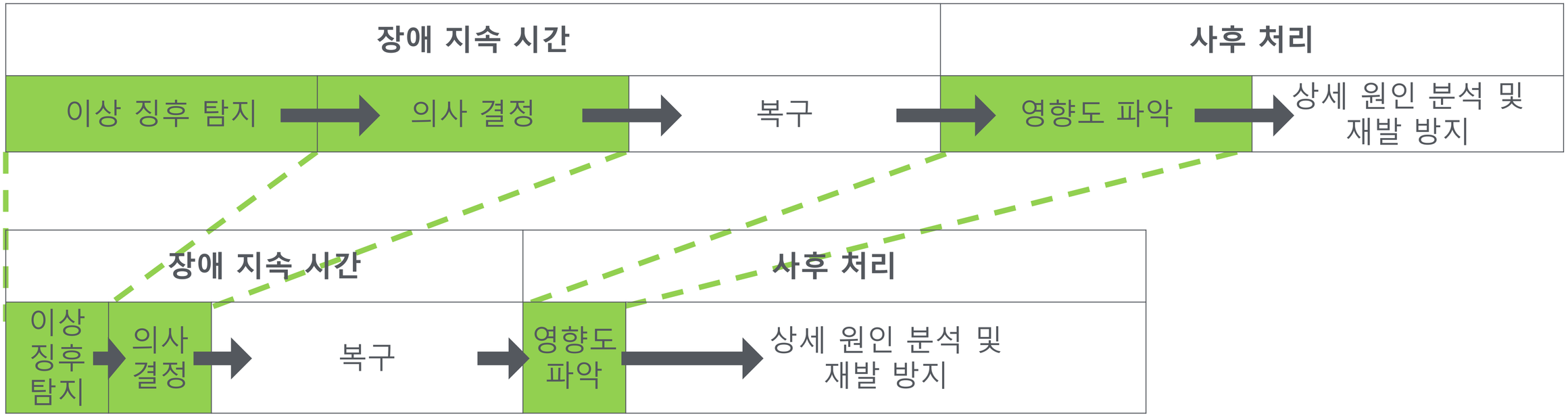


[장애 복구, 원인 분석, 재발 방지 대책]에

소요되는 시간은 줄이기 어려움

4. Fail Fast, Learn Faster

DEVIEW 2019



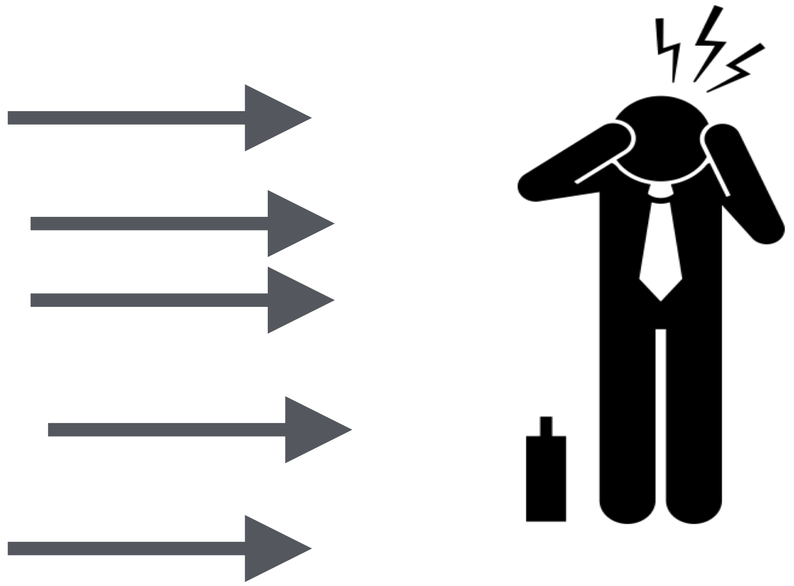
그렇지만 [이상 징후 탐지, 의사 결정, 영향도 파악]

시간은 줄일 수 있음

4. Fail Fast, Learn Faster

DEVIEW 2019

[검색서버팜(태국어) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(번체) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(영어) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(인도네시아어) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(간체) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(일본어) SAS] 가용성 임계상황 해제	2017-12-06 11:08
[검색서버팜(태국어) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(한국어) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(번체) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(간체) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(영어) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(인도네시아어) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(일본어) SAS] 임계상황 1.3<=1.3 farm c3a057	2017-12-06 11:07
[검색서버팜(태국어) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(일본어) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(번체) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(영어) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(인도네시아어) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(한국어) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(간체) SAS] 가용성 임계상황 해제	2017-12-06 11:04
[검색서버팜(인도네시아어) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(일본어) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(번체) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(간체) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(한국어) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(태국어) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(영어) SAS] 임계상황 1.2<=1.3 farm c3a057	2017-12-06 11:03
[검색서버팜(영어) SAS] 가용성 임계상황 해제	2017-12-06 11:00
[검색서버팜(간체) SAS] 가용성 임계상황 해제	2017-12-06 11:00



[이상 징후 탐지 시간]을 줄이기 위해 **[경보 빈도]**를 늘리고

하루에 **200통 이상 경보**를 받기도 하면서

경보 피로가 발생

4. Fail Fast, Learn Faster

경보를 줄이기 위한 **개발**이 아니라,

원칙이 필요

[이상 징후 탐지 시간]을 줄이기 위해 **[경보 빈도]**를 늘리고

하루에 **200통 이상 경보**를 받기도 하면서

경보 피로가 발생

4. Fail Fast, Learn Faster

원칙 1 [간단함]

오캄의 면도날 (Occam's Razor)

- **가능하면 간단한 추론**을 하라.
- 조건이 같다면 **간단한 것**이 더 강력하다.

4. Fail Fast, Learn Faster

원칙 1 [간단함]

가용량 규칙

부하증가배수

한 서버가 죽으면 나머지 서버들은
몇 배를 받나?

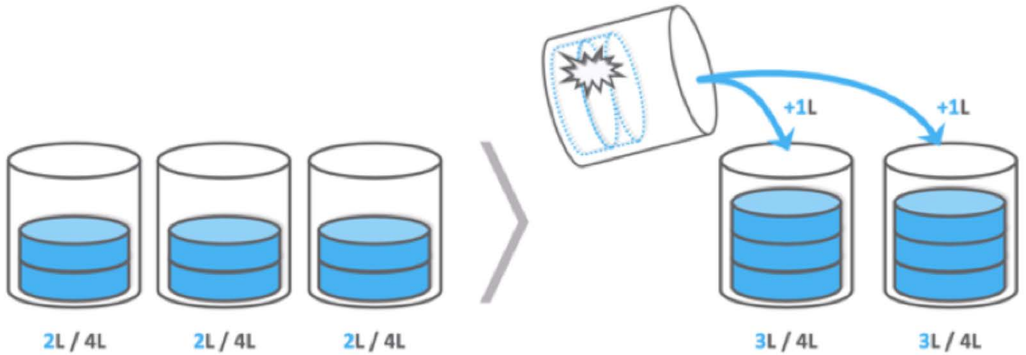
최대가용배수

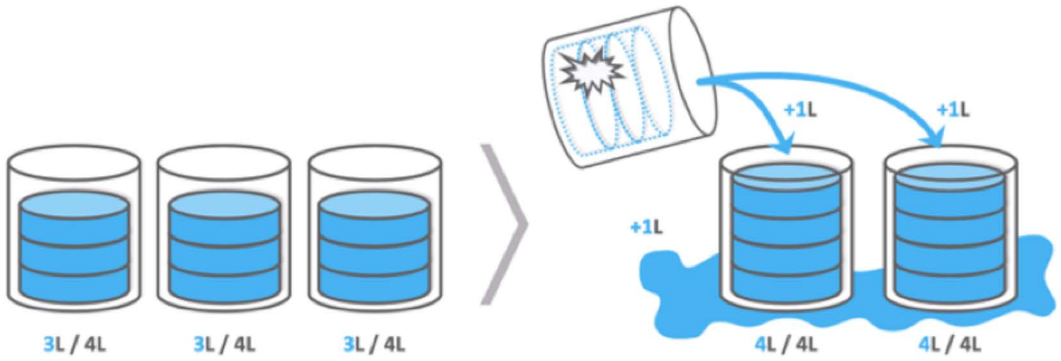
한 서버가 현재 몇 배까지 받을 수 있나?

So Simple!

"임계 상황 판단 공식"

부하증가배수 > 최대가용배수





4. Fail Fast, Learn Faster

원칙 1 [간단함]

트래픽 경보

- 사람들의 생체 주기를 따라가는 트래픽 (PC 통합검색)

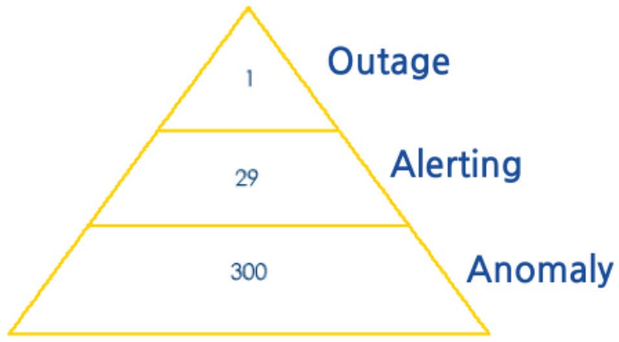
“정상 트래픽”에서 벗어나면 **“경보 발송”**

4. Fail Fast, Learn Faster

원칙 2 [1:29:300 법칙]

하인리히 법칙 (Heinrich's Law)

- 어떤 **대형 사고**가 일어나기 전에는
  수십 차례의 경미한 사고, **수백 번의 징후**가 반드시 일어난다.



[SRE 시선의 하인리히 법칙]

4. Fail Fast, Learn Faster

원칙 2 [1:29:300 법칙]

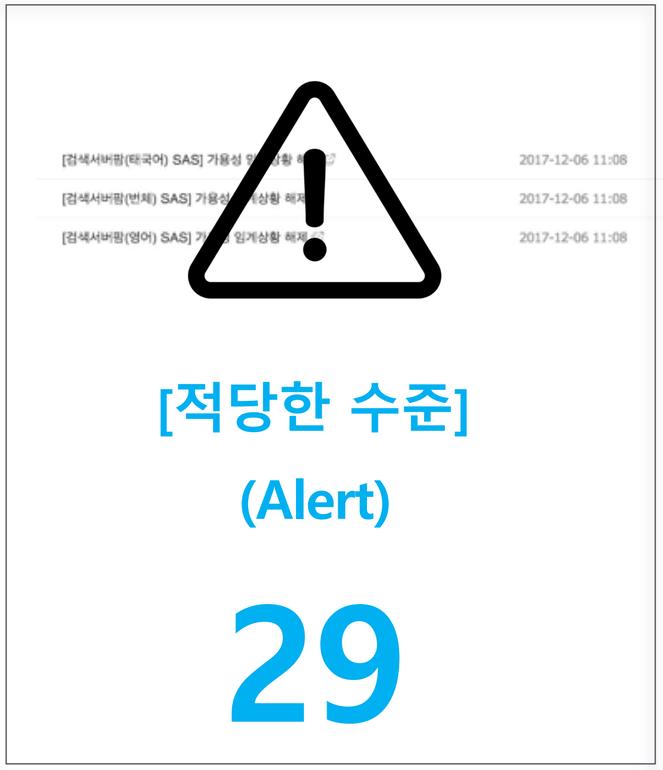
적당한 수준의 경보 빈도

[적음]

(Outage)

1

:



[적당한 수준]

(Alert)

29

:

[많음]

(Anomaly)

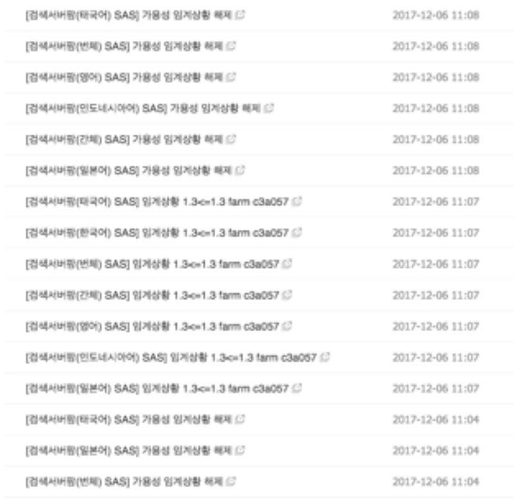
300

4. Fail Fast, Learn Faster

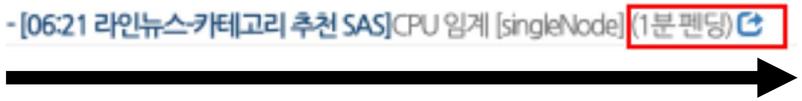
원칙 2 [1:29:300 법칙]

적당한 수준의 경보 빈도

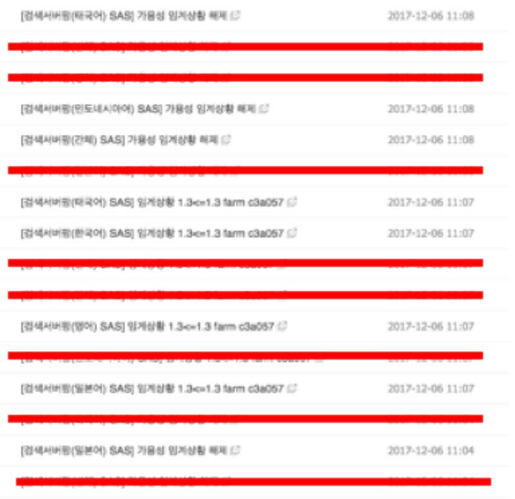
- **Pending:**
  즉각적으로 대응해야 하는 경우가 아니라면 경보를 **잠시 보류**



[Threshold 넘으면 바로 경보]



[Pending]



[Pending으로 경보피로 완화]

4. Fail Fast, Learn Faster

원칙 2 [1:29:300 법칙]

적당한 수준의 경보 빈도

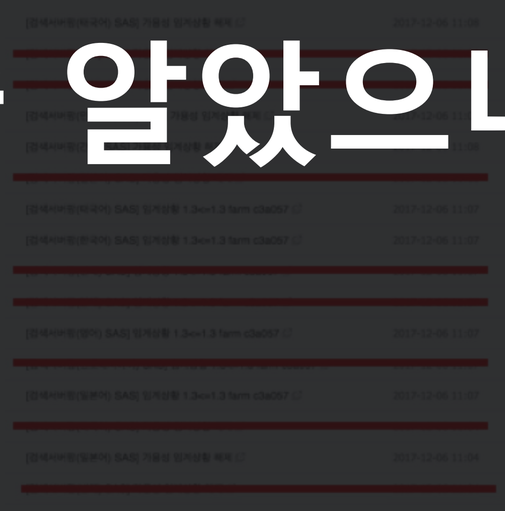
- Pending:
  즉각적으로 대응해야 하는 경우가 아니라면 경보를 잠시 보류

경보가 상당수 줄어들고

행복한 SRE가 시작되는 줄 알았으나...

[Threshold 넘으면 바로 경보]

[Pending]



[Pending으로 경보피로 완화]

4. Fail Fast, Learn Faster

DEVIEW 2019

평일 공휴일

그런데... 평화로운 **크리스마스**에 경보가 대량 발생

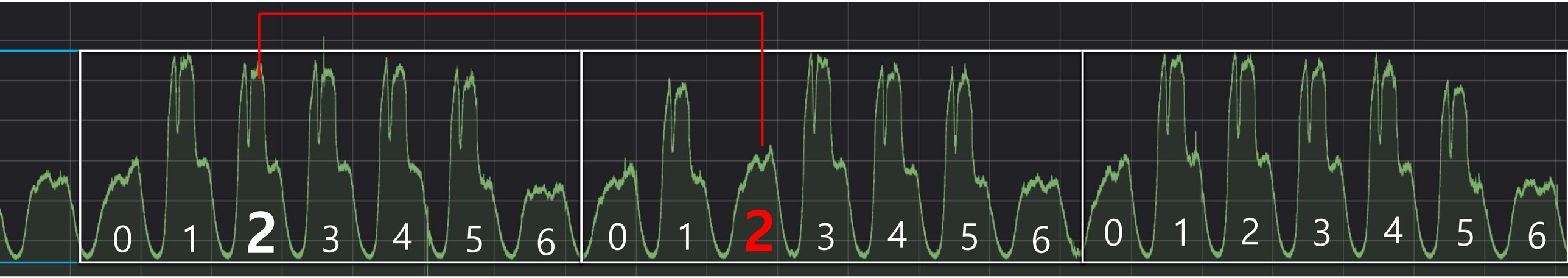
SRE가 긴급 대응 시작

4. Fail Fast, Learn Faster

DEVIEW
2019

평일 공휴일 (ex. 크리스마스 화요일)

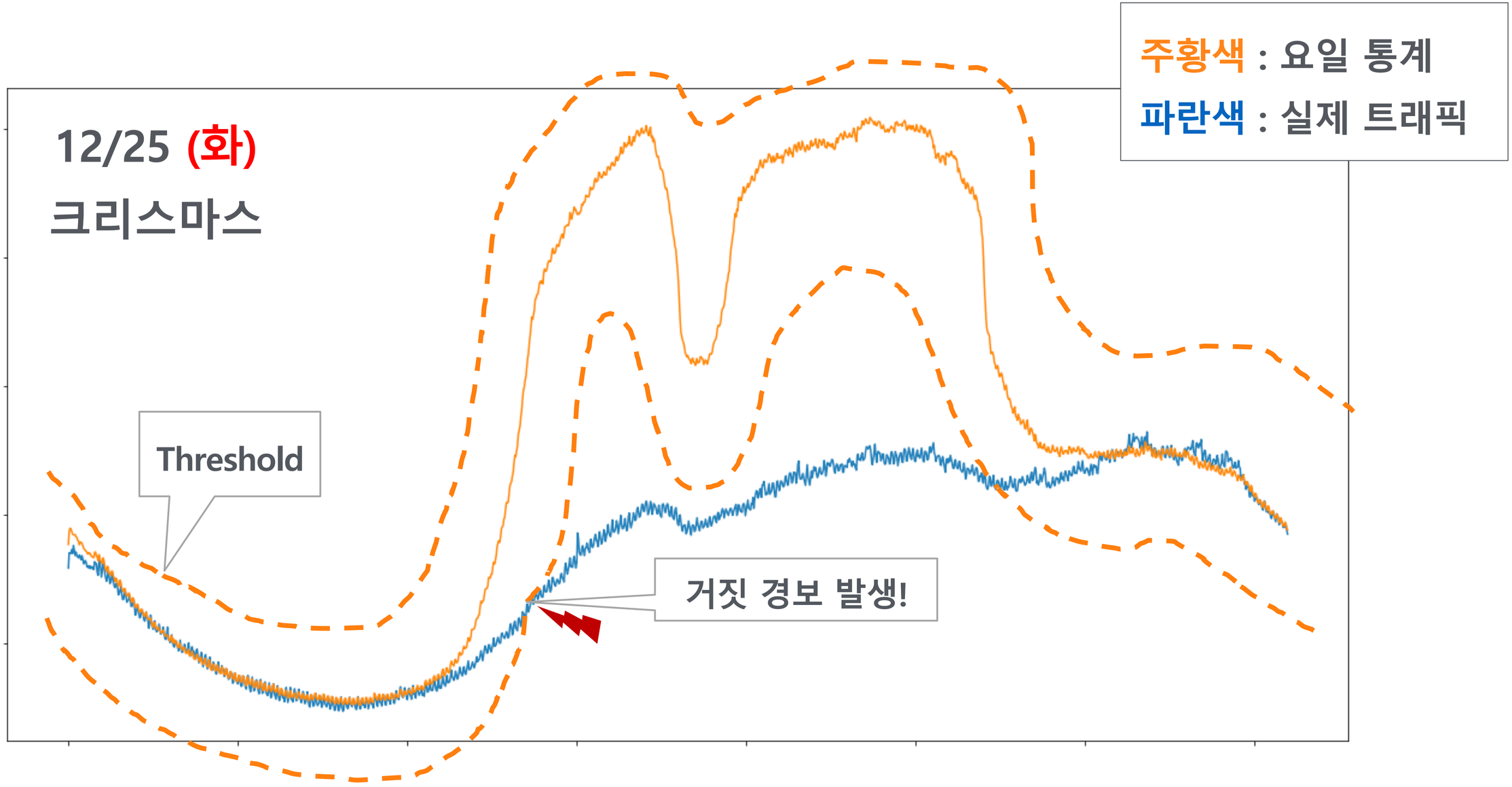
대량의 **거짓 경보** 발생



일	월	**화**	수	목	금	토	일	월	**화**	수	목	금	토	일	월	화	수	목	금

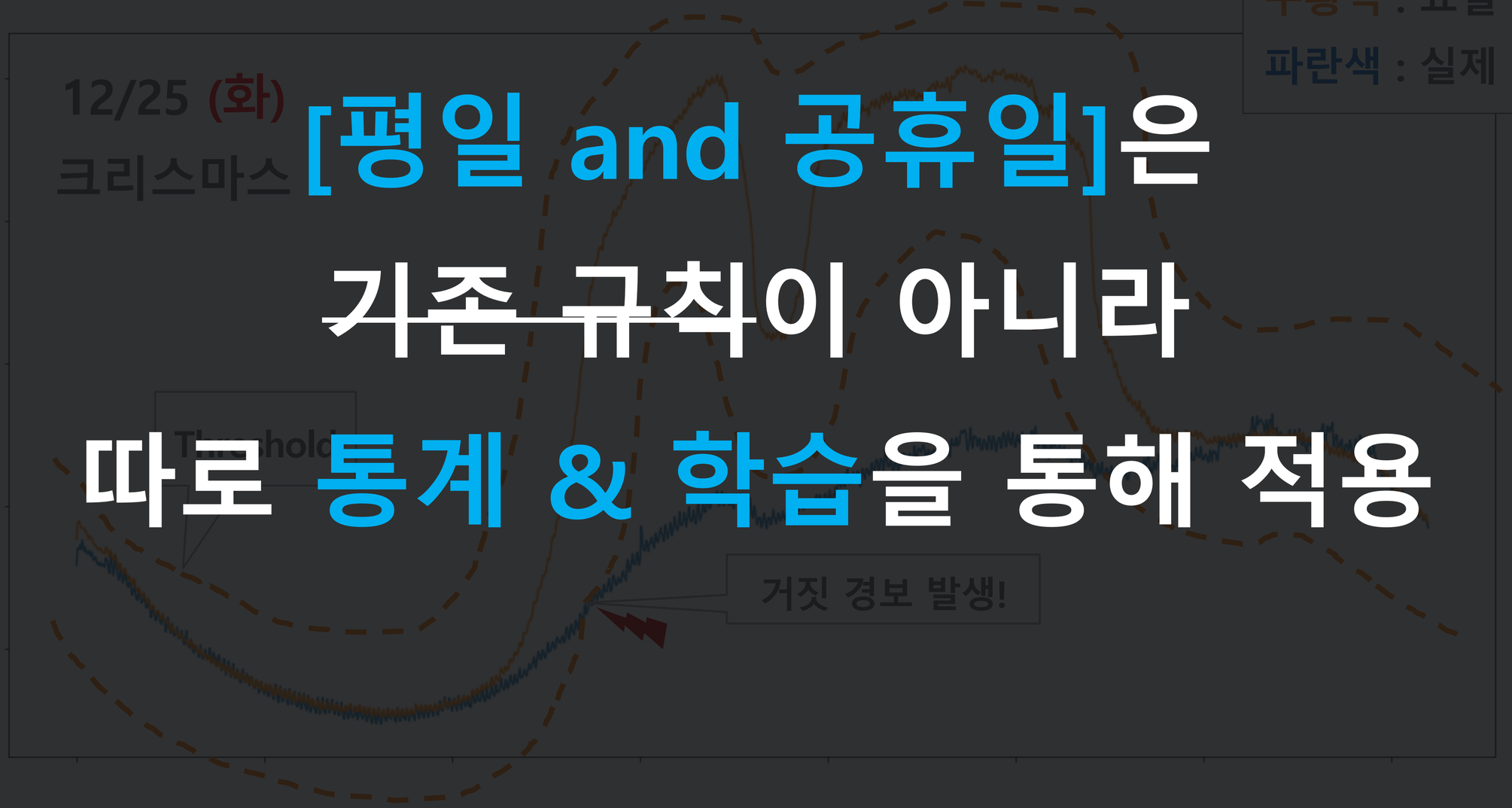
12/25
(크리스마스)

4. Fail Fast, Learn Faster



4. Fail Fast, Learn Faster

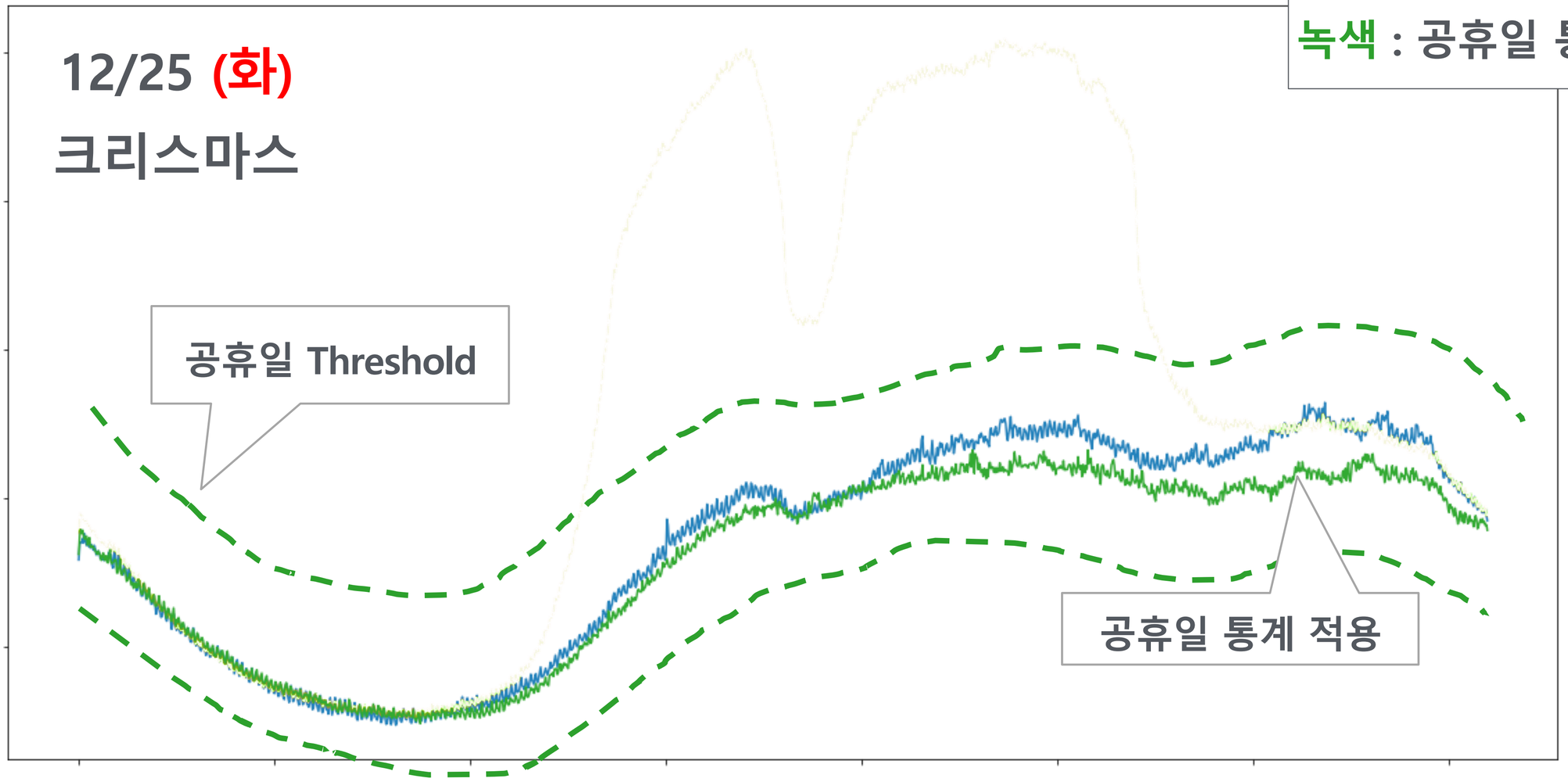
주황색 : 요일 통계
파란색 : 실제 트래픽



[평일 and 공휴일]은
~~기존 규칙~~이 아니라
따로 통계 & 학습을 통해 적용

4. Fail Fast, Learn Faster

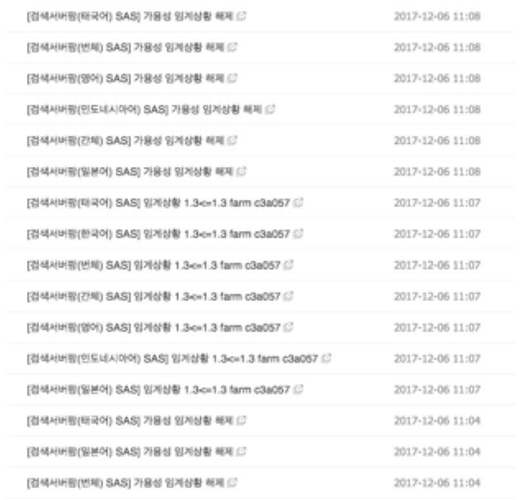
파란색 : 실제 트래픽
녹색 : 공휴일 통계



4. Fail Fast, Learn Faster

DEVIEW 2019

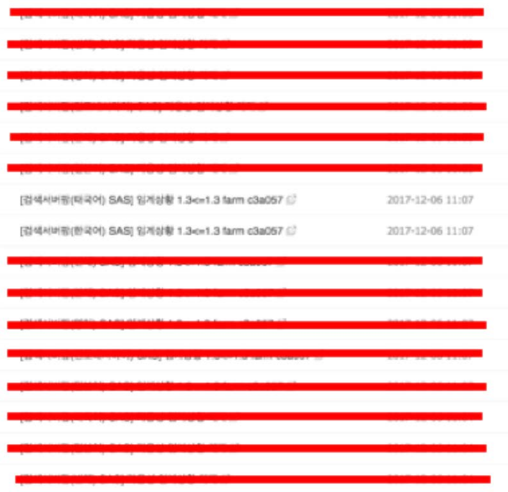
[Pending] + [평일 공휴일 통계, 예측]



[기존 경보]

[Pendng] + [평일 공휴일 통계, 학습 값] 적용 후

(거짓 경보 **90%** 감소)



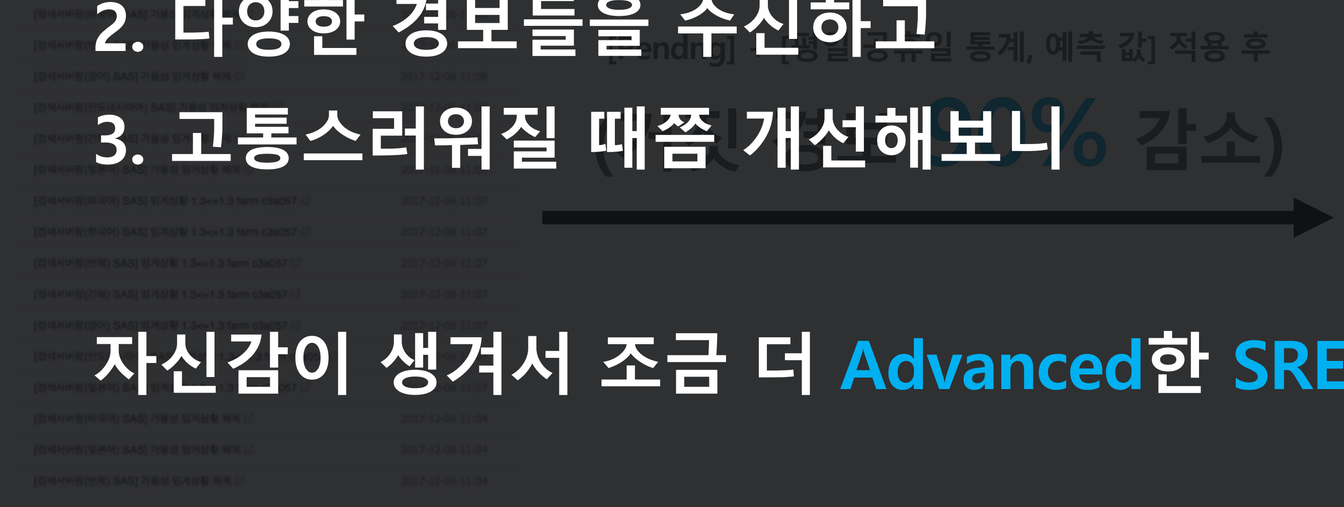
[Pending, 평일 공휴일 통계 적용]

4. Fail Fast, Learn Faster

DEVIEW 2019

Pending + 평일 공휴일 통계, 예측

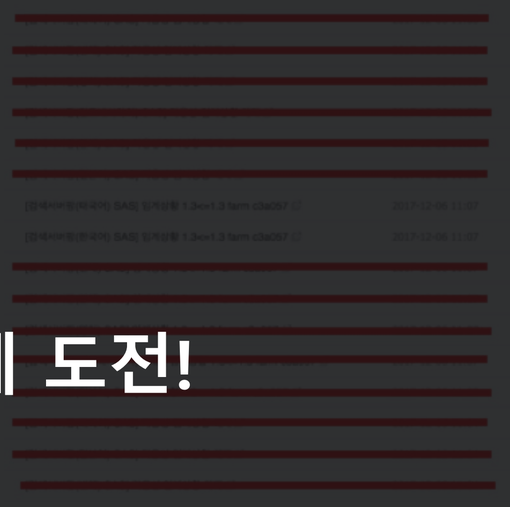
1. 트래픽을 자세히 관찰하고
2. 다양한 경보들을 수신하고
3. 고통스러워질 때쯤 개선해보니



[기존 알람]

[Pending] 평일 공휴일 통계, 예측 값] 적용 후

(약 99% 감소)



[Pending, 평일 공휴일 통계 적용]

자신감이 생겨서 조금 더 Advanced한 SRE에 도전!

4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced 1. 트래픽 패턴 분석

- 현재도 발생하는 거짓 경보는 **어떤 이유**로 발생하는 것일까?

[Pending, 평일 공휴일 통계 적용]

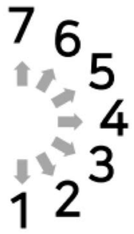
도대체 이런 거짓 경보는 왜 발생하는 걸까?

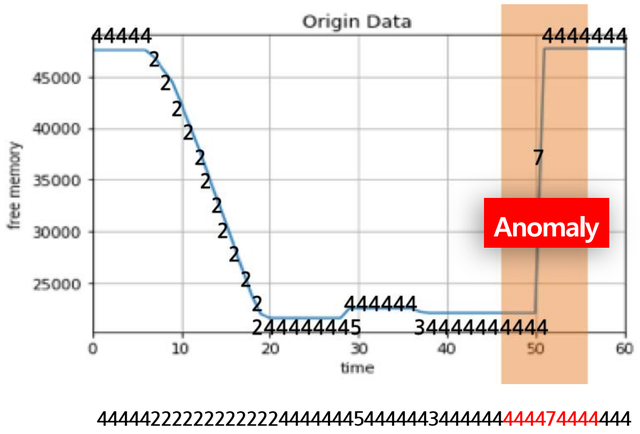
4. Fail Fast, Learn Faster

DEVIEW 2019

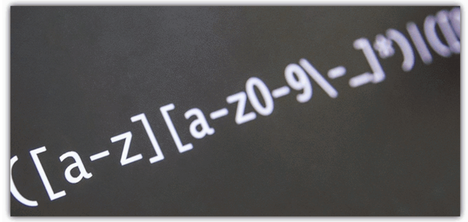
Advanced 1. 트래픽 패턴 분석

- 그래프의 **기울기**만으로도 경보가 발생한 원인 파악





미리 정의한 패턴과 매칭



[기울기] [그래프 인코딩] [정규 표현식]

4. Fail Fast, Learn Faster

Advanced 1. 트래픽 패턴 분석

- **패턴 예제: 트래픽(Traffic)**이 다른 곳으로 **이전(Migration)** 되었는지 판단

4. Fail Fast, Learn Faster

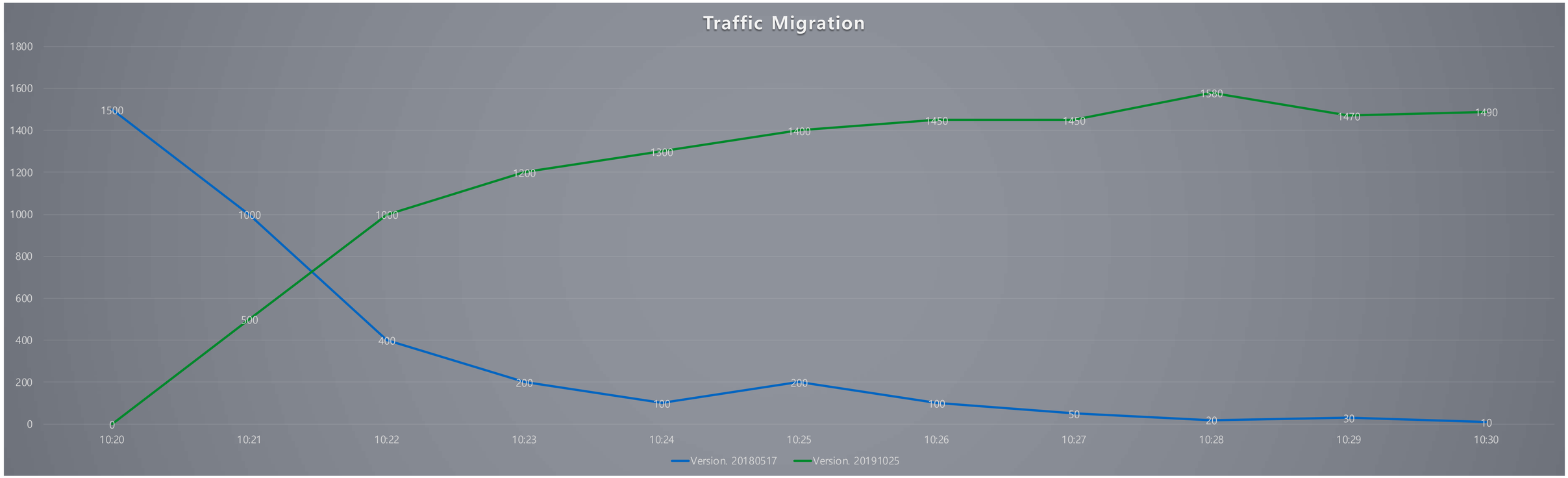
DEVIEW 2019

Advanced 1. 트래픽 패턴 분석

- **Traffic Migration: 트래픽(Traffic)**이 다른 곳으로 **이전(Migration)** 되었는지 판단

Old Service Version
(Production Service Version)

New Service Version



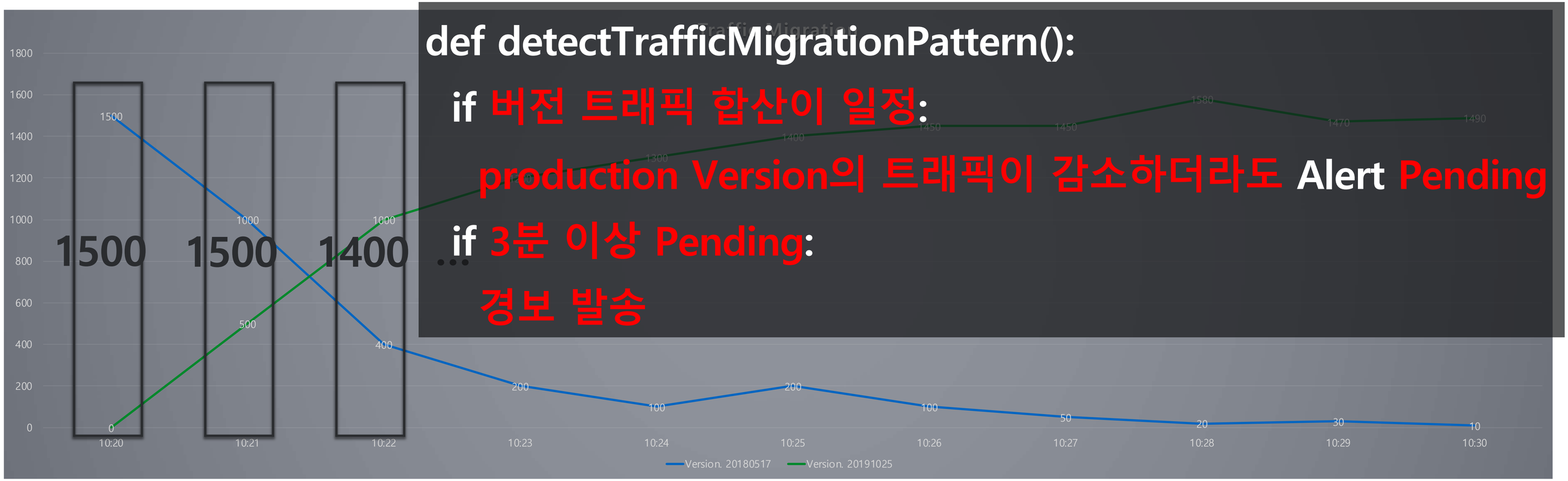
4. Fail Fast, Learn Faster

Advanced 1. 트래픽 패턴 분석

- **Traffic Migration: 트래픽(Traffic)**이 다른 곳으로 **이전(Migration)** 되었는지 판단

Old Service Version
(Production Service Version)

New Service Version



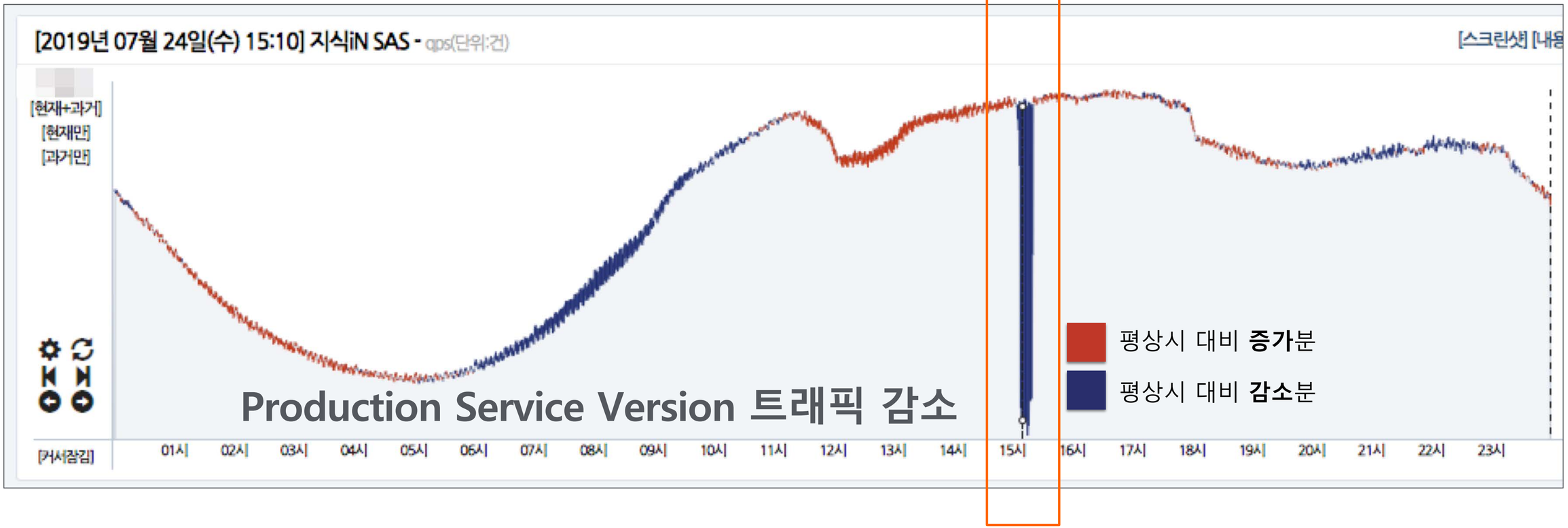

```
def detectTrafficMigrationPattern():
  if 버전 트래픽 합산이 일정:
    production Version의 트래픽이 감소하더라도 Alert Pending
  if 3분 이상 Pending:
    경보 발송
```

4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced 1. 트래픽 패턴 분석 (사례)

Traffic Migration Pattern



4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced 1. 트래픽 패턴 분석 (사례)

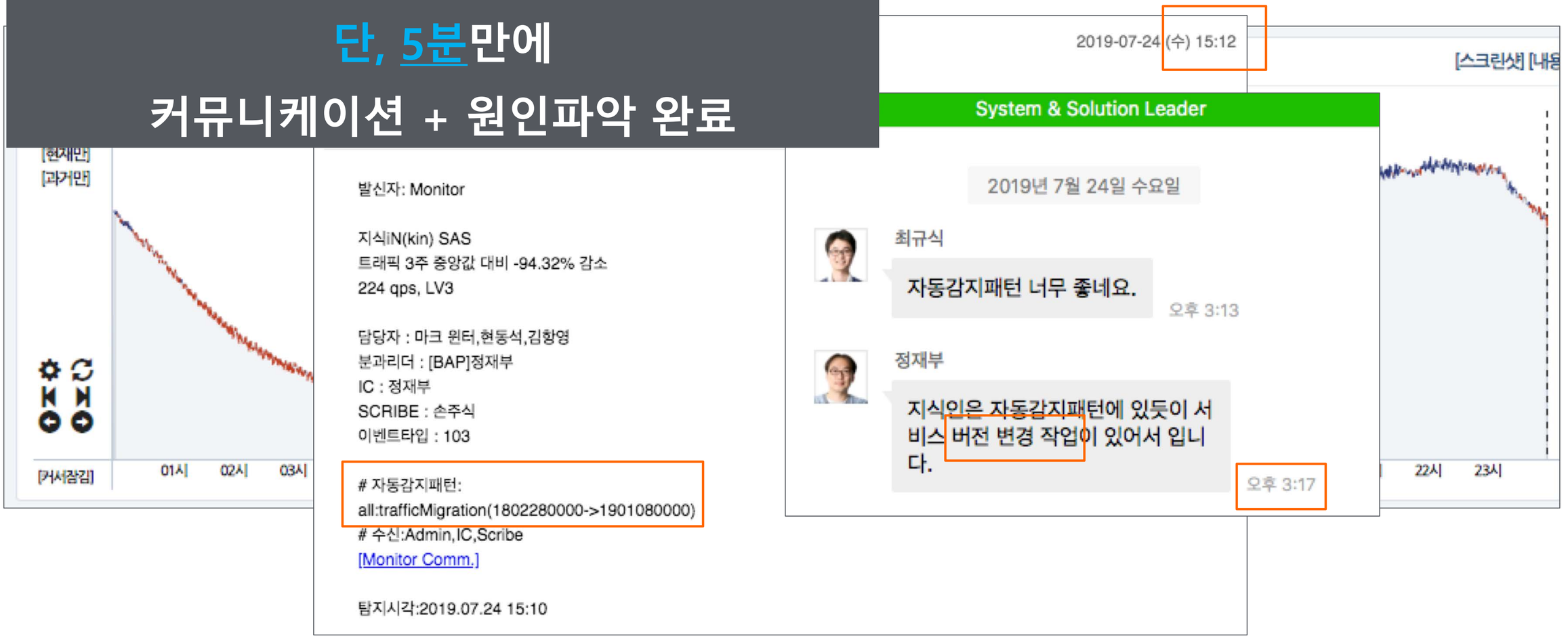
Traffic Migration Pattern

4. Fail Fast, Learn Faster

경보 발송 후

단, 5분만에

커뮤니케이션 + 원인파악 완료



4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced 2. 자동 비상 대응

- 2017년 11월 15일 포항 지진

4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced 2. 자동 비상 대응

- 2017년 11월 15일 포항 지진

2017년 말

자동 비상 대응 시스템

개발 시작

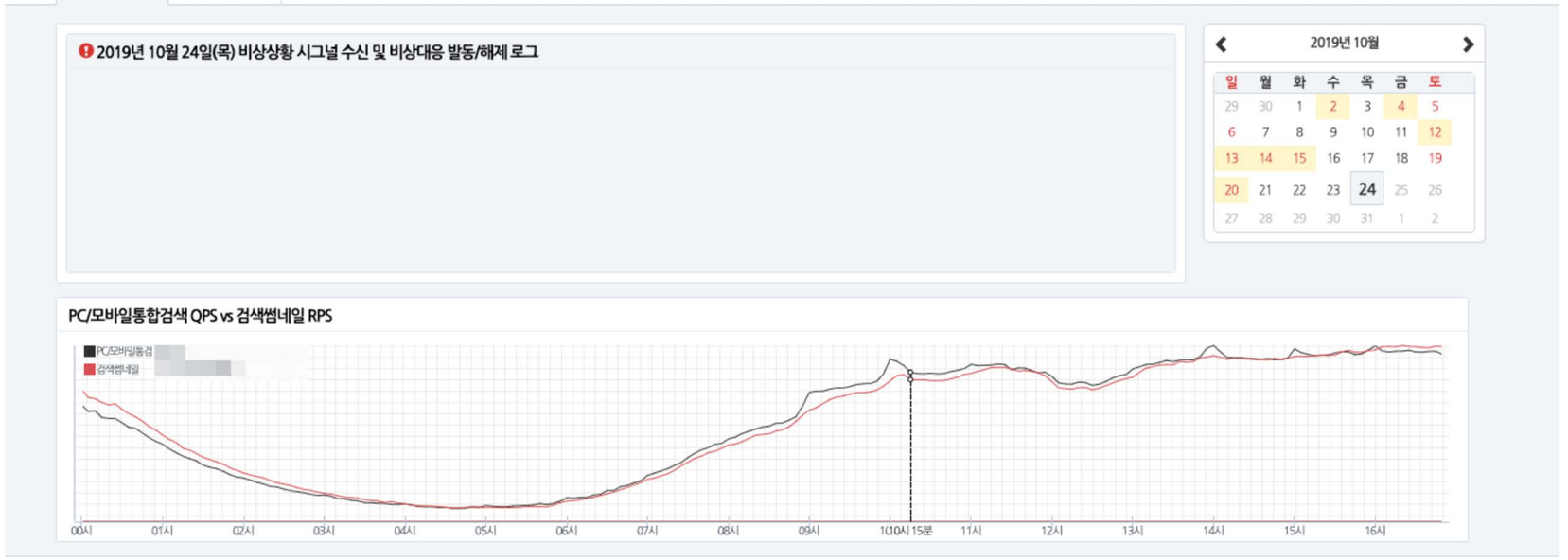
4. Fail Fast, Learn Faster

Advanced 2. 자동 비상 대응 시스템 개발 완료 (2018년)

if (대한민국 4.0 이상 지진 발생 OR 통합검색 트래픽의 급격한 증가) {

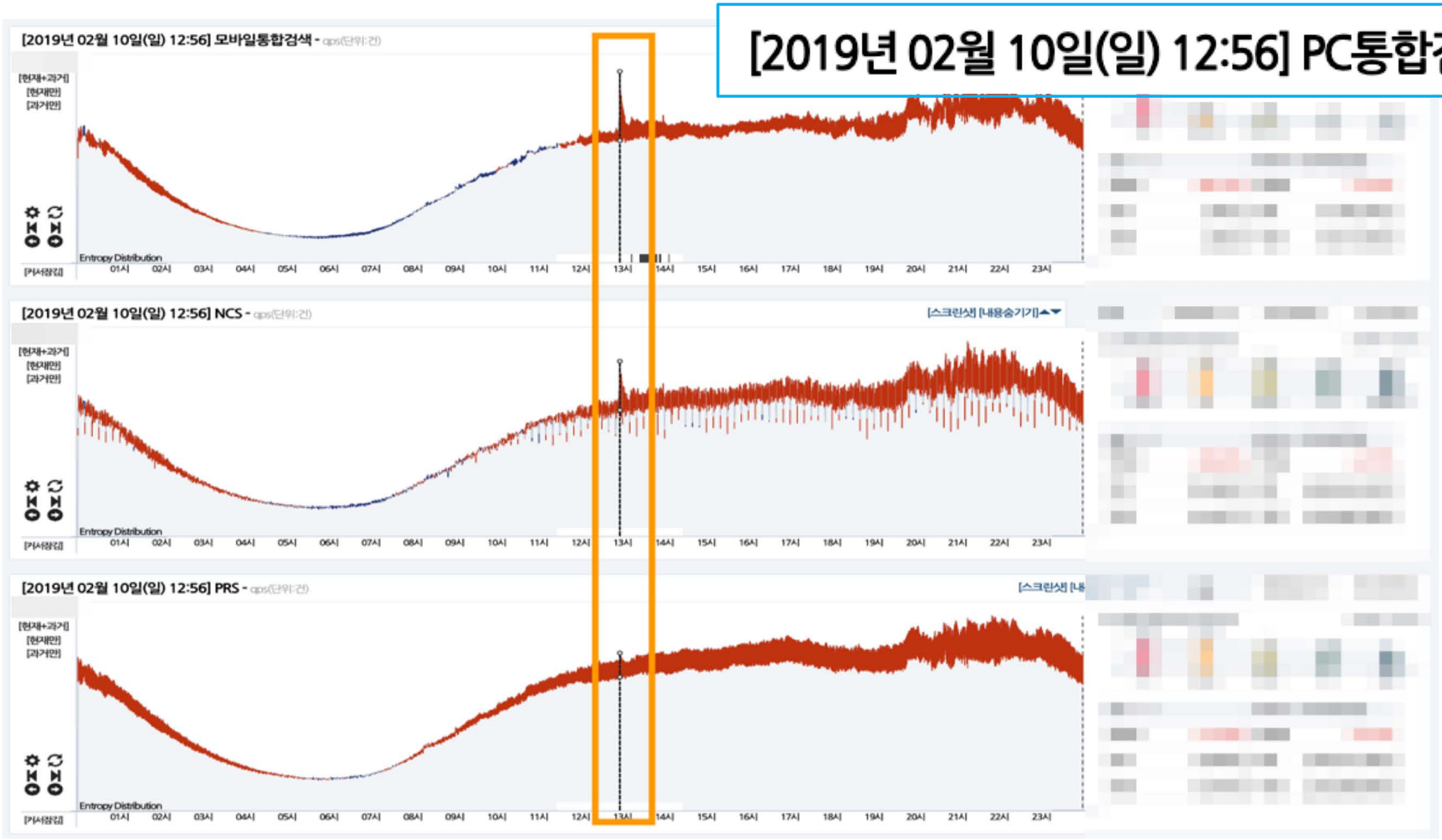
Cache Expire Time 연장 -> 서버 하단 부하를 방지

}



4. Fail Fast, Learn Faster

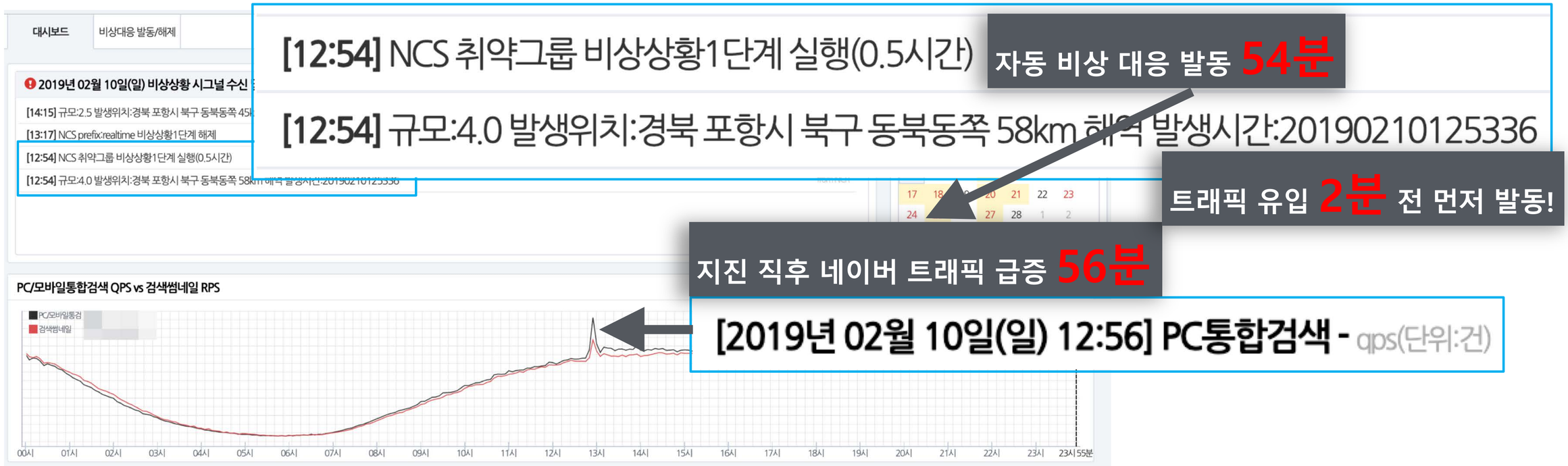
Advanced 2. 자동 비상 대응: 2019년 2차 포항 지진



4. Fail Fast, Learn Faster

Advanced 2. 자동 비상 대응: 2차 포항 지진

트래픽 급증하여 유입되기 전 **자동 방어 장치 가동** SRE의 궁극적 목적!



4. Fail Fast, Learn Faster

DEVIEW 2019

Advanced Result. 빨라진 영향도 분석

장애 Detection	2분
증상 완화	20분
원인 파악	40분
장애 영향도 파악	1시간
장애 복구 완료	수시간
포스트모텀 작성	48시간

[SRE Before]

전체 현황, 영향도 분석

[SRE After]

DEVIEW
2019

오캄의 면도날

하인리히 법칙

트래픽 경보

1:29:300

거짓 경보

최대가용배수

부하증가배수

평일 공휴일

Summary

가용량

임계상황

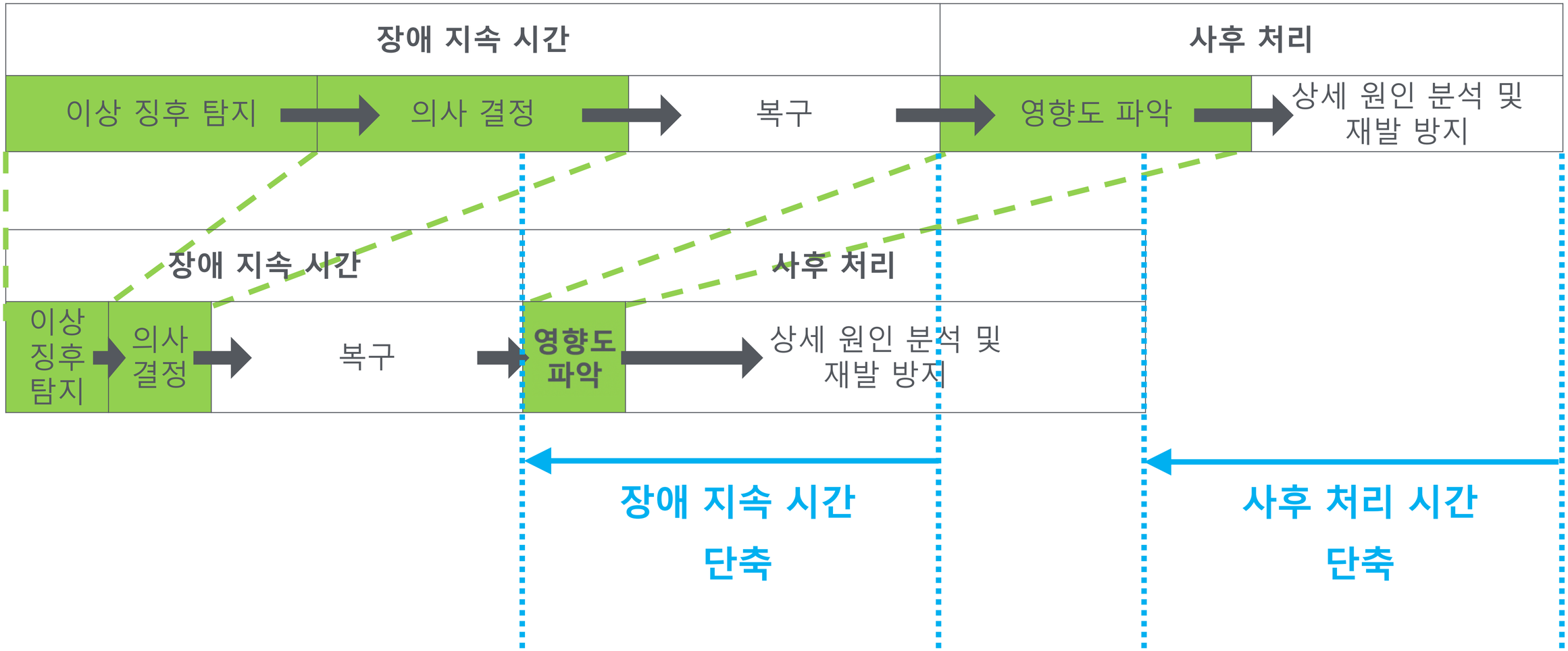
트래픽 분석

영향도 분석

비상 대응 시스템

4. Fail Fast, Learn Faster

DEVIEW 2019



4. Fail Fast, Learn Faster

단순하고, 효과적인 방법으로

[가용량, Treshold기반의 경보]

4. Fail Fast, Learn Faster

DEVIEW
2019

단순하고, 효과적인 방법으로

빠르게 실패하고

[많은 거짓 경보, 평일 공휴일 에러, 지진]

4. Fail Fast, Learn Faster

단순하고, 효과적인 방법으로

빠르게 실패하고, 빠르게 배워서

[Pending + 평일 공휴일 통계 + 트래픽 패턴 분석 + 자동 비상 대응]

4. Fail Fast, Learn Faster

단순하고, 효과적인 방법으로

빠르게 실패하고, 빠르게 배워서 [개선]한다.

[업데이트 대비 장애 비율 감소]

4. Fail Fast, Learn Faster

많아진 변경 작업에도, 줄어든 장애 (19.09.03 기준)

	2017년	2018년	2019년
업데이트 (건)	1311	2110	1566
장애 (건)	57	49	18
변경 대비 장애 비율 (%)	4.35	**2.32**	**1.15**
증감 YoY (%)	-	**-46.67**	**-50.43**

Fail Fast, Learn Faster

5. Lessons Learned

5. Lessons Learned

DEVIEW 2019

(1) SRE ≒ 재난 엔지니어링

본질적인 **SRE**에 집중하다 보면,
결국 **재난 엔지니어링**으로 귀결된다.

[이상탐지(anomaly detection)의 자동화 + 경보압축(alert compression)] 추구
= 시스템 문제점 해결 + 안정적인 시스템 구축
= 재난 엔지니어링

5. Lessons Learned

(2) SRE를 하기 위해서 필요한 Attitude

- **이상적이고 멋들어진 도구**가 아닌 **실질적인 문제 해결**에 주력하기
  (Realist)
- 무식한 방법이더라도 힘겹더라도 **포기하지 않고** 끊임없이 **개선**하기
  (Grit)
- 본인이 속한 **도메인**에 적절한 방법론을 적용하기
  (Domain Specific)

5. Lessons Learned

DEVIEW 2019

(2) SRE를 하기 위해서 필요한 Attitude

- **이상적이고 멋들어진 것**인 것이 아닌 **실질적인 문제 해결**에 주력하기 (Realistic)
- 무식한 방법이더라도 힘겹더라도 **포기하지 않고** 끊임없이 **개선**하기 (Greedy)
- 본인이 속한 **도메인**에 적절한 방법론을 적용하기 (Domain Specific)

이런 Attitude를 장착하고

우리와 재미있는 일을 함께할

당신을 찾습니다.

(채용, 엄격, 근엄, 진지)

5. Lessons Learned

DEVIEW 2019

(2) SRE를 하기 위해서 필요한 Attitude

- **이상적이고 멋들어진 것**이 전부가 아닌 **실질적인 문제해결**에 주력하기 (Realist)
- 무식한 방법이더라도 **일단 시작**하고 **두려워하지 않고** 끊임없이 **개선**하기 (Greet)
- 본인이 속한 **도메인**에 적절한 방[illegible]하기 (Domain Specific)

이렇게만 하면 아무도 연락 안 할 테니,

yoohoogun114으로 ~~**[카톡]**~~ 주세요!

그린팩토리에서 커피 사드립니다

[채용공고 링크 Click]

:smile:

5. Lessons Learned

DEVIEW
2019

Q & A

DEVIEW
2019

Thank You