New Relic.

{FUTURE}STACK 16

Own Your Own Impact: Incident Response at Airbnb

Cameron Tuckerman-Lee, Site Reliability, Airbnb

CAMERON TUCKERMAN-LEE / 2016 NOVEMBER 17 / FUTURESTACK

Own Your Own Impact: Incident Response at Airbnb

@TUCKERMAN

Who am I?

Owning Your Own Impact

Our next 45 Minutes Together

DevOps & Sysops

Convincing 50 people go on-call for all of Airbnb voluntarily.

People First On-Call

The future of what it means to take the pager at Airbnb.

Tooling

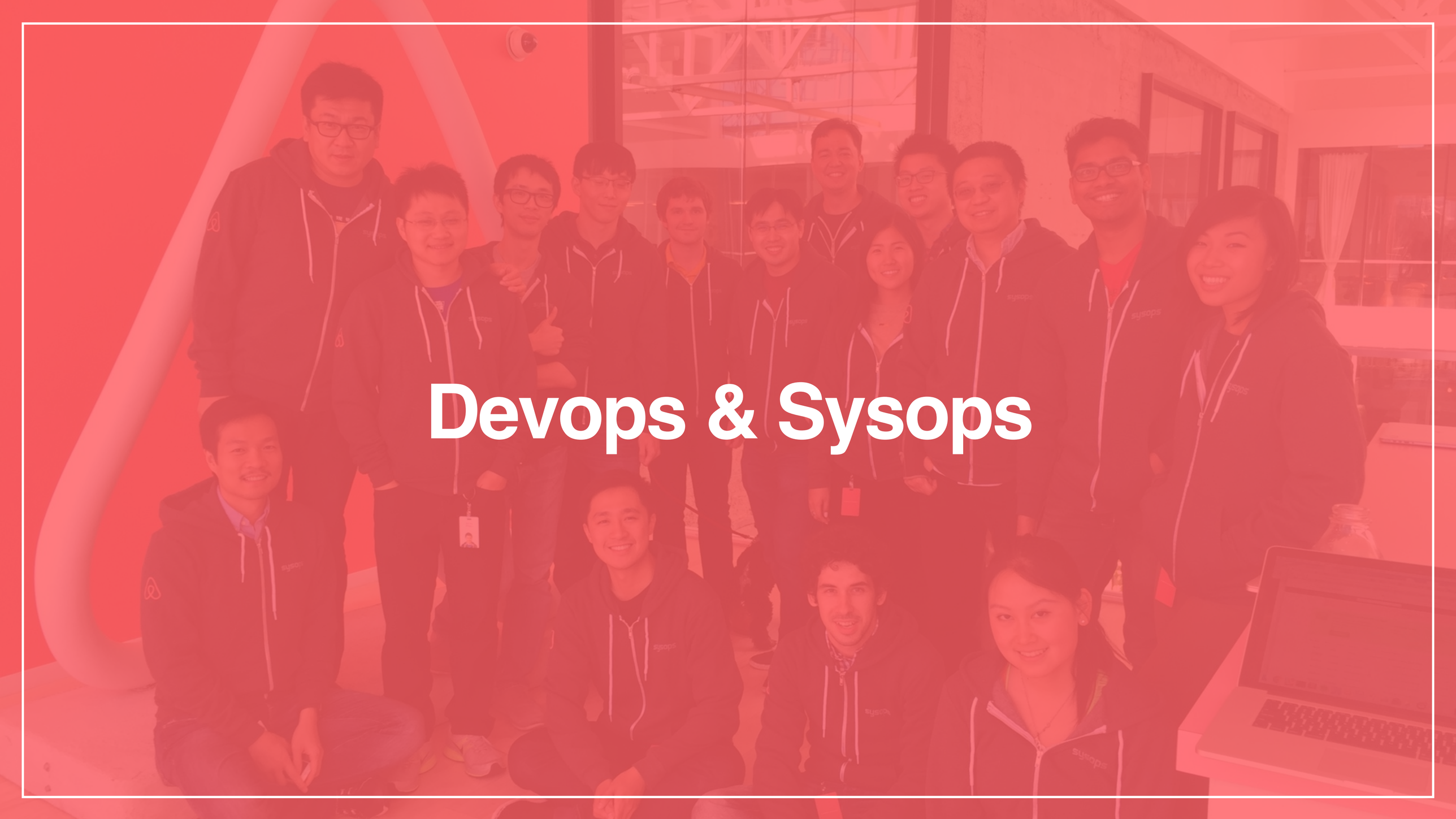
The tools and services we use that make this all possible.

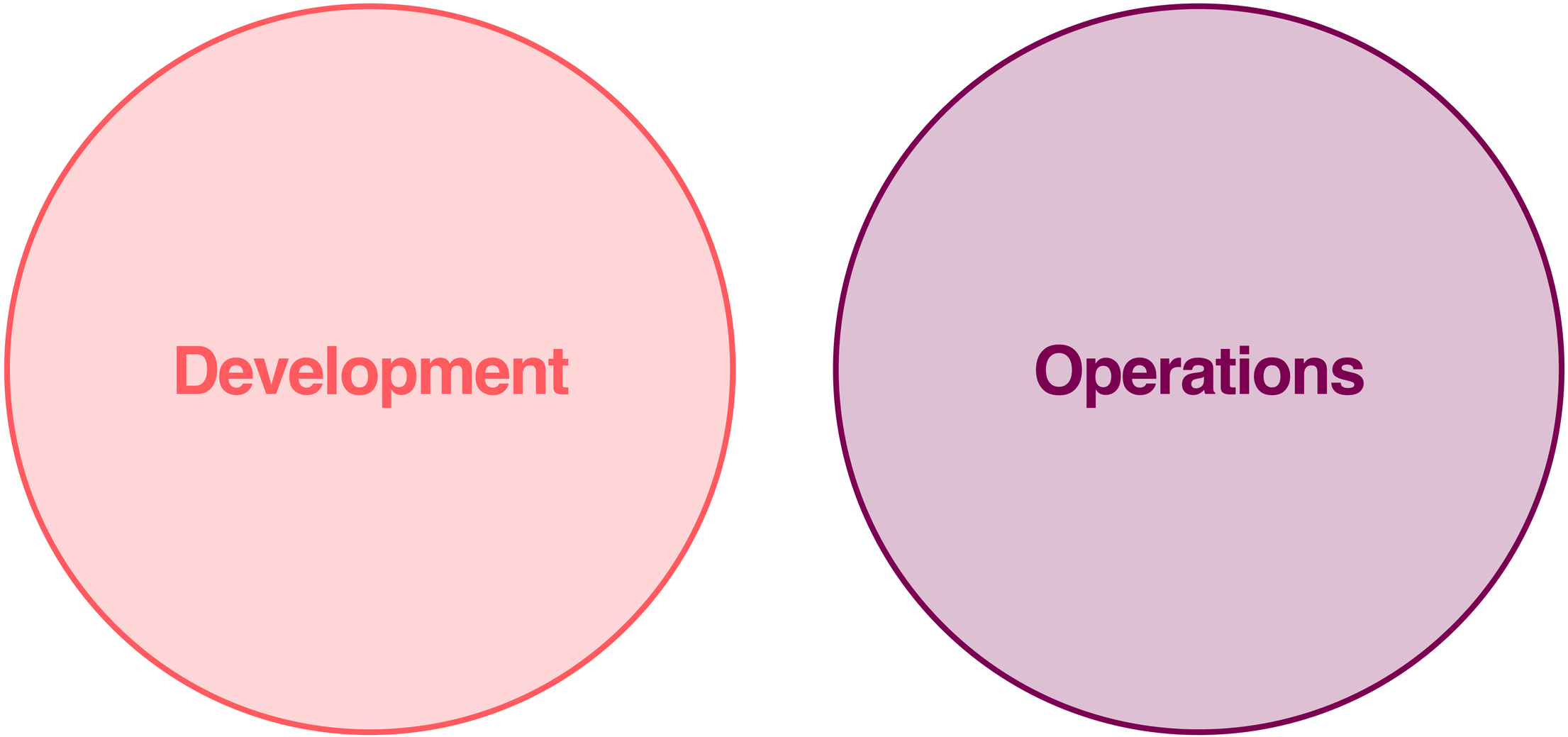
Incidents Walk-Thru

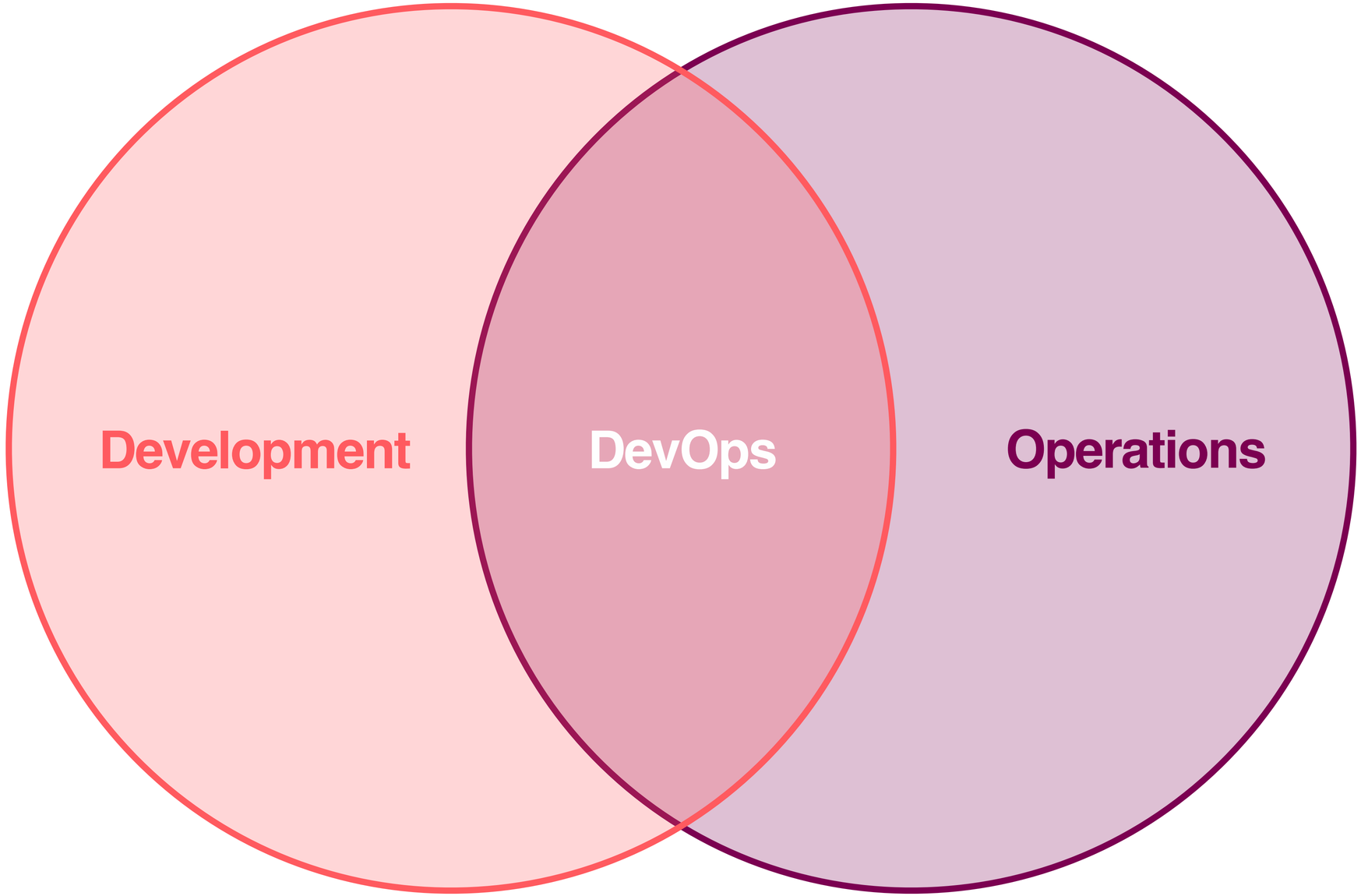
Real past incidents at Airbnb.

Airbnb Open & New Relic

Feeling so comfortable about our biggest launch ever that I’m here instead of in a war room.


Devops & Sysops


Development
Operations


Development
DevOps
Operations

DevOps

Usually the Best of Both Worlds

- **Velocity:** Teams are able to move quickly as they own both the development and release of their project.
- **Ownership:** Smaller teams owning a product area or feature. In depth knowledge about its architecture and how to operate it.
- **Automation:** Incentivized to automate as much of the operations as possible.

DevOps

Sometimes the Worst of Both Worlds

- **Upstream Outages:** Cloud provider outages
- **Not So Isolated:** Features often have to come together in the view layer, where a problem with one can cause a problem with another.
- **Incident Management:** Managing the lifecycle of an incident, gauging customer impact, communicating to stake holders, prioritizing remediations are hard work and specialized knowledge.
- **Coordination:** The same incident is happening to multiple teams and they are all investigating from square one.

Sysops

History

SRE

Airbnb hires the first SRE tasked with ensuring the site doesn't go down. He is on-call 24/7 responding to all incidents.

Volunteers

Airbnb starts to get a bit bigger. Incidents are more frequent and the stakes are higher. Engineers volunteer to help out the SRE one day at the weekly meeting.

Sysops

The volunteer group grows, is formalized, and officially takes over owning incident response. Hats were made.

Sysops

History

SRE

Airbnb hires the first SRE tasked with ensuring the site doesn't go down. He is on-call 24/7 responding to all incidents.

Volunteers

Airbnb starts to get a bit bigger. Incidents are more frequent and the stakes are higher. Engineers volunteer to help out the SRE one day at the weekly meeting.

Sysops

The volunteer group grows, is formalized, and officially takes over owning incident response. Hats are made.

Sysops

History

SRE

Airbnb hires the first SRE tasked with ensuring the site doesn't go down. He is on-call 24/7 responding to all incidents.

Volunteers

Airbnb starts to get a bit bigger. Incidents are more frequent and the stakes are higher. Engineers volunteer to help out the SRE one day at the weekly meeting.

Sysops

The volunteer group grows, is formalized, and officially takes over owning incident response. Hats are made. Sysops is born.

Sysops

History

SRE

Airbnb hires the first SRE tasked with ensuring the site doesn't go down. He is on-call 24/7 responding to all incidents.

Volunteers

Airbnb starts to get a bit bigger. Incidents are more frequent and the stakes are higher. Engineers volunteer to help out the SRE one day at the weekly meeting.

Sysops

The volunteer group grows, is formalized, and officially takes over owning incident response. Hats are made.

Sysops at Airbnb

Today

- **Volunteer Group:** Frontend and backend. Product and infrastructure. Individual contributor and manager.
- **Specialized Training:** Biennial training covering infrastructure, incident response, communication, and pizza.
- **Ownership:** Escalation point of last resort. Primary on-call has the authority to make important decisions in the moment.

Sysops at Airbnb

Today

25

Engineers in Primary Rotation

50

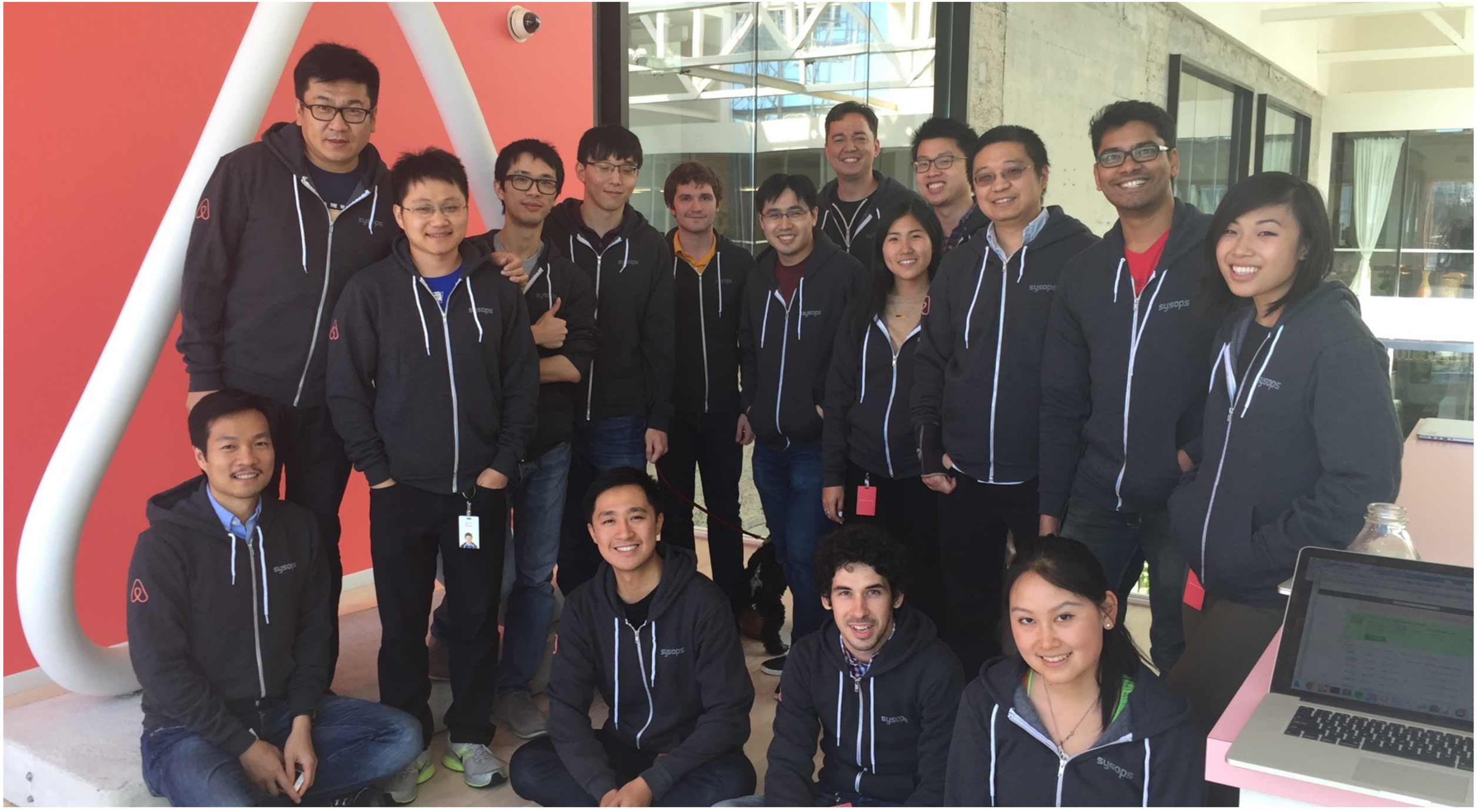
Engineers Currently On Sysops

33%

Engineers Attended Sysops Training

How does this work?

- **Reliability Matters:** The uptime of our site isn’t just important for conversions, it’s important for the entire end-to-end experience.
- **Making Mistakes:** Blameless postmortem. Blameless remediation.
- **Learning Opportunity:** Training, collaboration, truly “fullstack".
- **Tooling:** Tooling is built with Sysops in mind.
- **SRE:** SRE builds tools that make on-calls more productive.
- **Culture:** Sysops grew out of our core values; cereal entrepreneurship.


sysops

People First On-Call

People First On-Call

The Future of On-Call at Airbnb

- **Pager-Life Balance:** Ensure that more involved, tenured engineers aren't always the ones waking up at 3 AM to put out fires.
- **Learning/Growth Focused:** Continuing education and learning opportunities for on-call engineers.
- **Evaluation Metrics:** Engineers should know where they can improve and should be recognized for excellent work.
- **Intelligent Scheduling:** In DevOps when every team has at least two on-call rotations, how can we schedule around lives outside of work (and responsibilities inside of work)?

Tooling

STATSD Protocol

Measure Everything

- **Metrics Live in Code:** When adding new features in the codebase, engineers can instrument them at the same time in line with the feature.
- **System/Business Metrics:** Allows single dashboards where you can correlate low and high level metrics together, e.g. packets-in and number of nights booked.
- **Simple:** Engineers don't need to ask for permission, submit PRs to the SRE team, or use other tools. Adding metrics is as easy as thinking of a metric name.

https://codeascraft.com/2011/02/15/measure-anything-measure-everything/

Alerts

Interferon

- **Configuration as Code:** Exposes best practices and allows re-syncing metrics in the case of emergencies.
- **Autogeneration of Alerts:** System-CPU alerts get created for every host, and alerts are automatically routed to the owner based on “Host Sources”.
- **Extensible:** Only used for our STATSD metrics, but are working on adding New Relic. Open Source so you can add your own!

https://github.com/airbnb/interferon

Example Alert

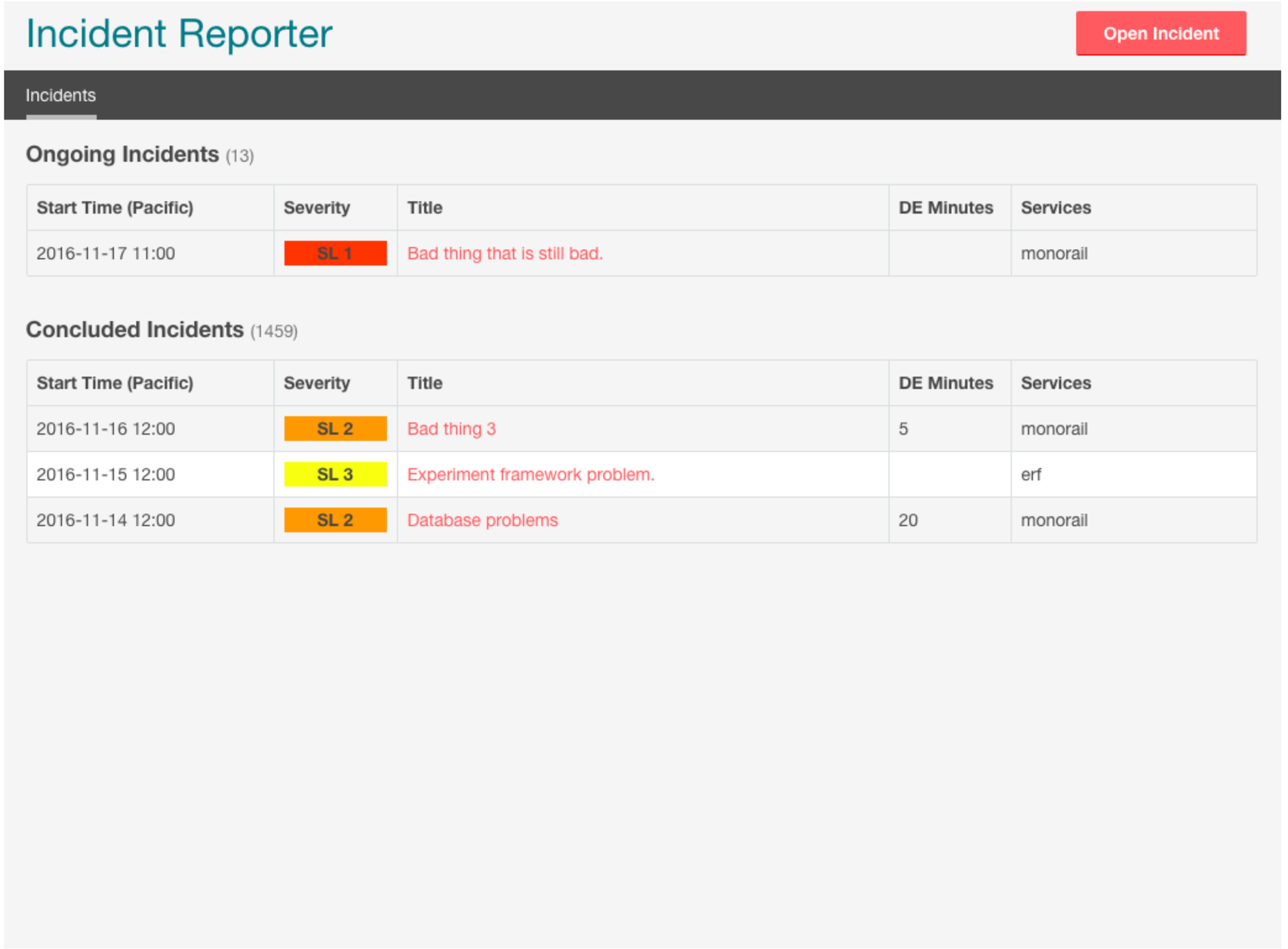
High CPU

```
name “#{@hostinfo[:role]}: load spiking”

message <<<EOM
You can put the body of your alert here if the
destination supports messages.

Context. Context. Context!
EOM

applies !@hostinfo[:role].end_with?(‘-test’)

notify.groups [
  ‘sre’,
  @hostinfo[:owner]
]

metric.query <<<EOQ
min(last_5m):avg:
  system.load.norm.1{role:#{@hostinfo[:role]}}
  > 0.95
EOQ
```

Postmortems

Incidents Reporter

- **Blameless Postmortems:** Goal is to learn how to prevent/ mitigate it in the future.
- **Tracking Impact:** Allows resource allocation to take into account the needs of bettering reliability.
- **Follow Up:** Integration with JIRA for bug-fix/improvement tracking.



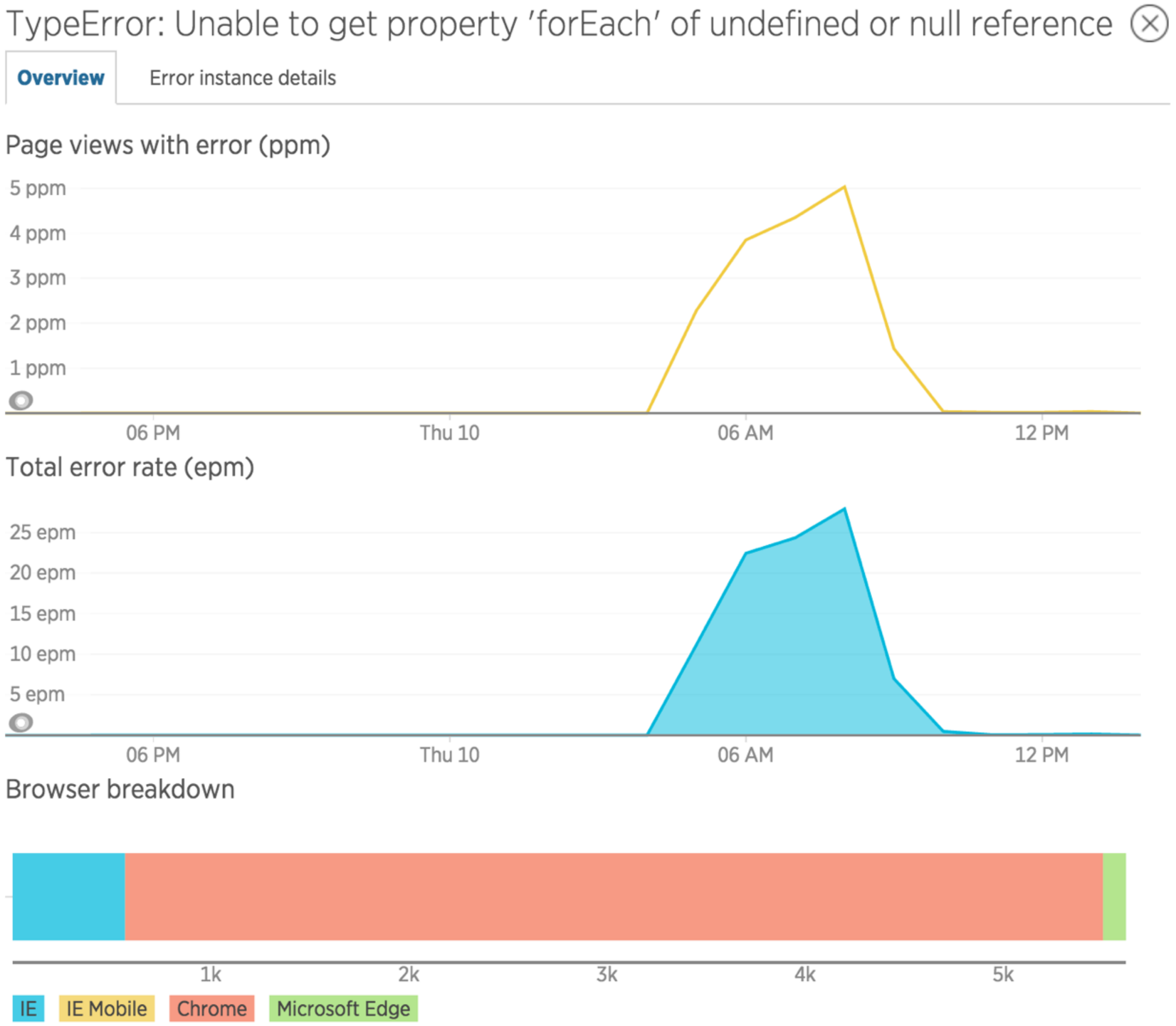
New Relic at Airbnb

Browser

Not every engineer is a frontend engineer…

- **Our Product is in the Browser:** Backend changes by backend engineers can have profound (i.e. terrible) implications for front-end code.
- **Alerts:** Before Browser, errors/metrics would have to propagate to the backend before engineers would be alerted. Now we can be alerted about regressions.
- **Browsers Your Engineers Don’t Us:** Engineers at Airbnb only use Chrome — that isn’t true about our users.

Browser Error Reporting



APM

Every application, every deploy

- **Out of the Box:** Chef auto-{installs, configures} the New Relic agent for all services without the engineer having to know about it.
- **Consistent Monitoring Across Platforms:** At Airbnb, we have java, node, and ruby services — APM gives all engineers a familiar monitoring tool without having to be familiar with the underlying framework.
- **Deploying:** Engineers know that they can monitor the health of a code deployment for any service (and downstream services) in APM. We also embed relevant APM information in our deploy tool, Deployboard.

Deployboard

Merges and Deploys Unlocked

Monorail

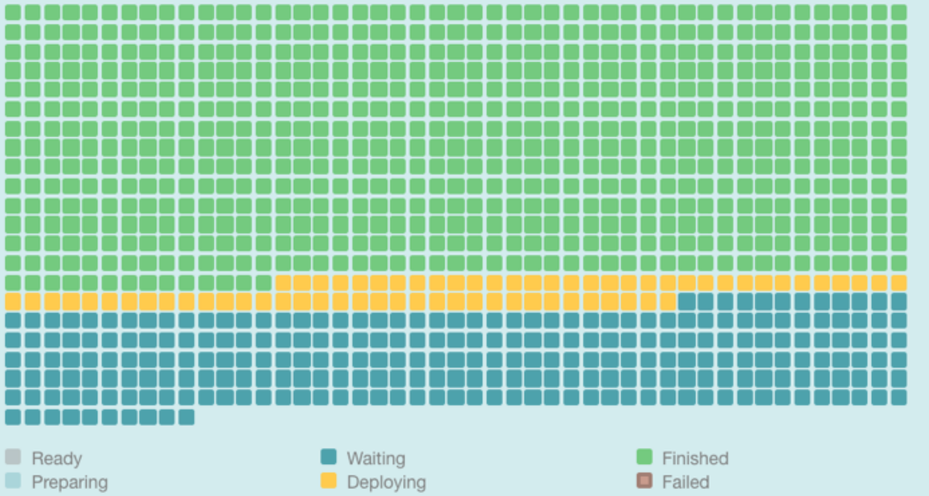
Snapshots | Deploys | Targets | Silhouettes | master

Deploy 123456 of *Application* to *Production*

672 / 997

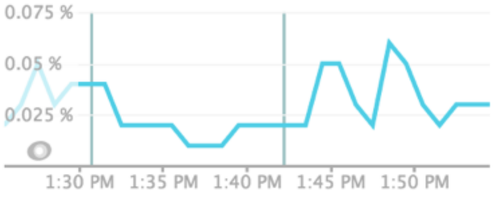
Started by user_1 at 1:42 pm.
Deploying snapshot #12345 (abc1234)
Deploy is running! Please watch metrics!

Roll Back | **Abort**

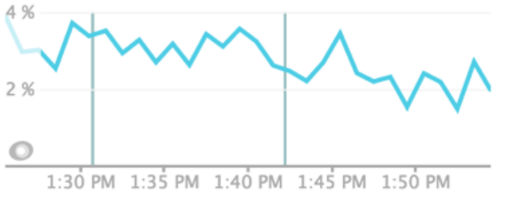


Application: Error rate (last 30 minutes)

Resque: Error Rate (last 30 minutes)



JS: Error Rate (last 30 minutes)



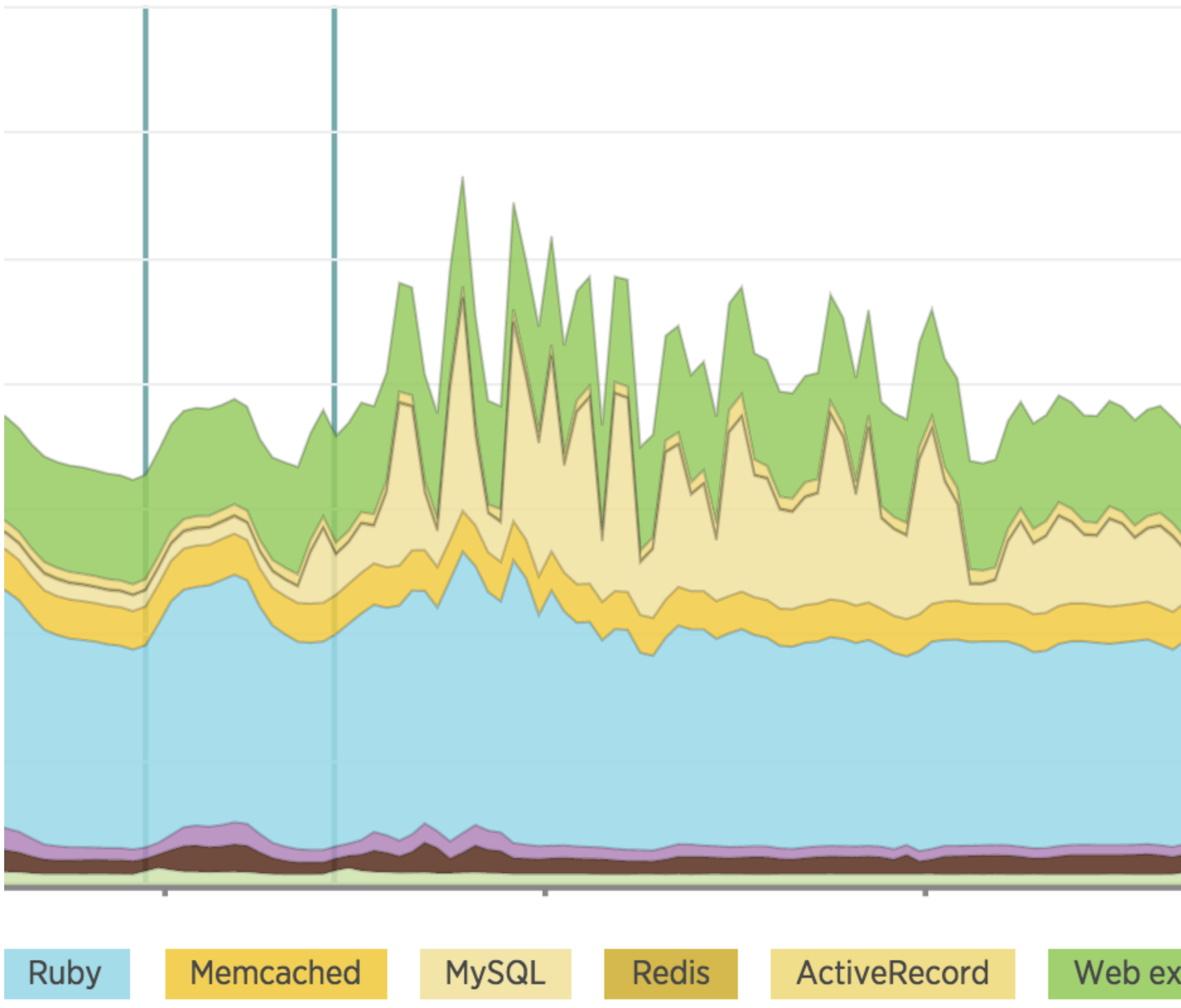
New Relic | Sentry | Resque | Synapse | Dashboards | Datadog Dashboards

Hide graphs

		Snapshot	SHA	Author	Pushed At (PDT)	Deployed	Compare To	Notes	
☐	☐	#12345	abc1234	user_1	Thu, Nov 17, 2016 1:23 PM		PREV HEAD C N P	☐	Deploy
☐	☐	#12344	xyz5678	user_2	Thu, Nov 17, 2016 11:59 AM		PREV HEAD C N P	☐	Deploy
☐	☐	#12343	a1b2c3d	user_3	Thu, Nov 17, 2016 11:58 AM		PREV HEAD C N P	☐	Deploy

Pale Yellow

Uh oh…



Synthetics

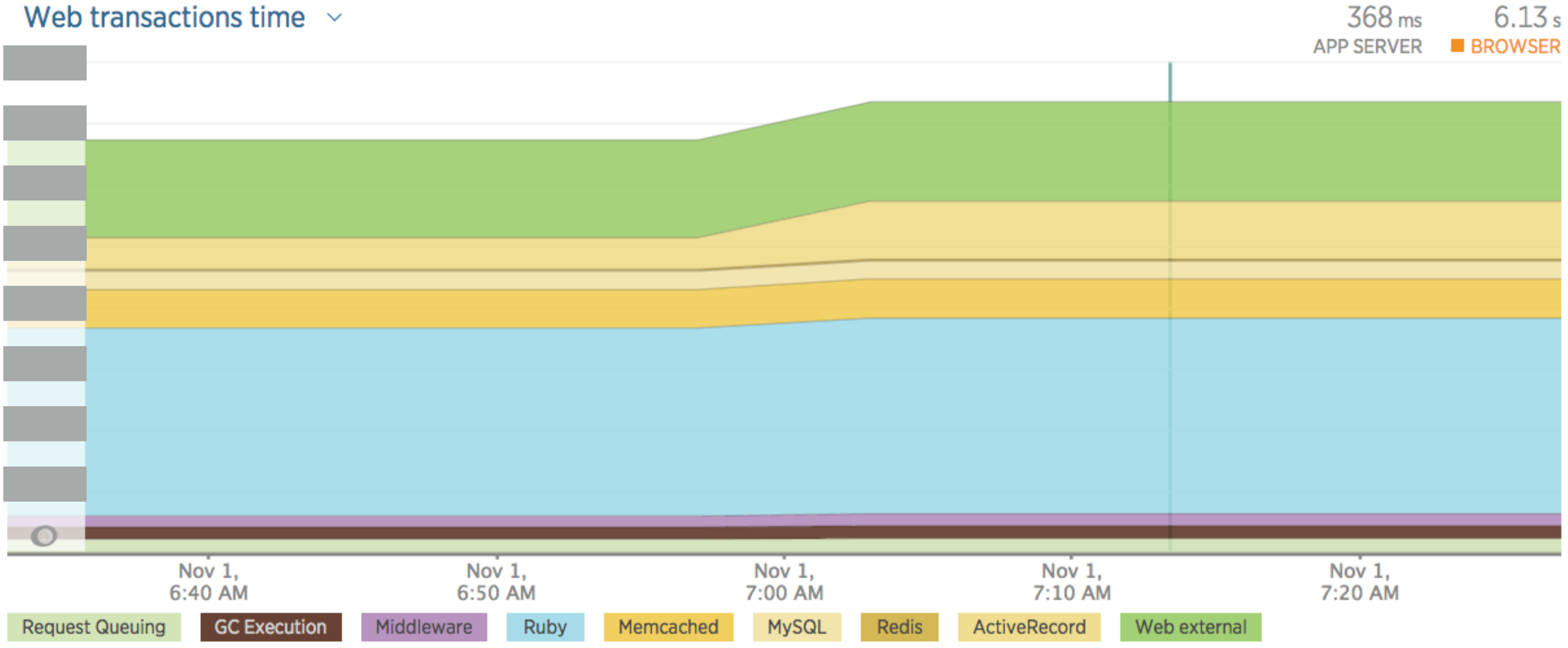
Last Line of Defense

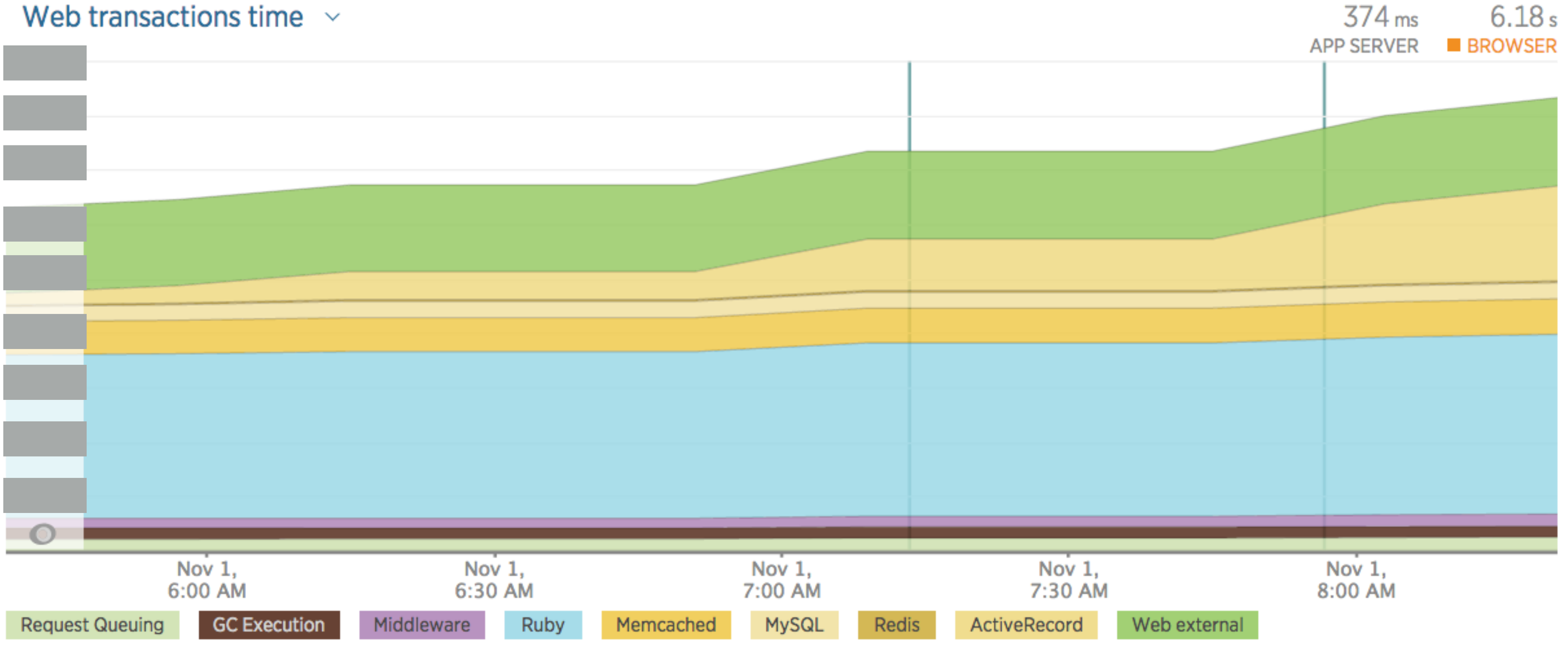
Can sleep comfortably at night knowing that, if other safeguards fail, any user-facing downtime will get caught.

Name	Success rate (24 hours)
Apex Ping *https://www.airbnb.com*	100%
API Health Ping *https://api.airbnb.com/healt...*	100%
Health Ping *https://www.airbnb.com/healt...*	100%

Real World Walk-Throughs

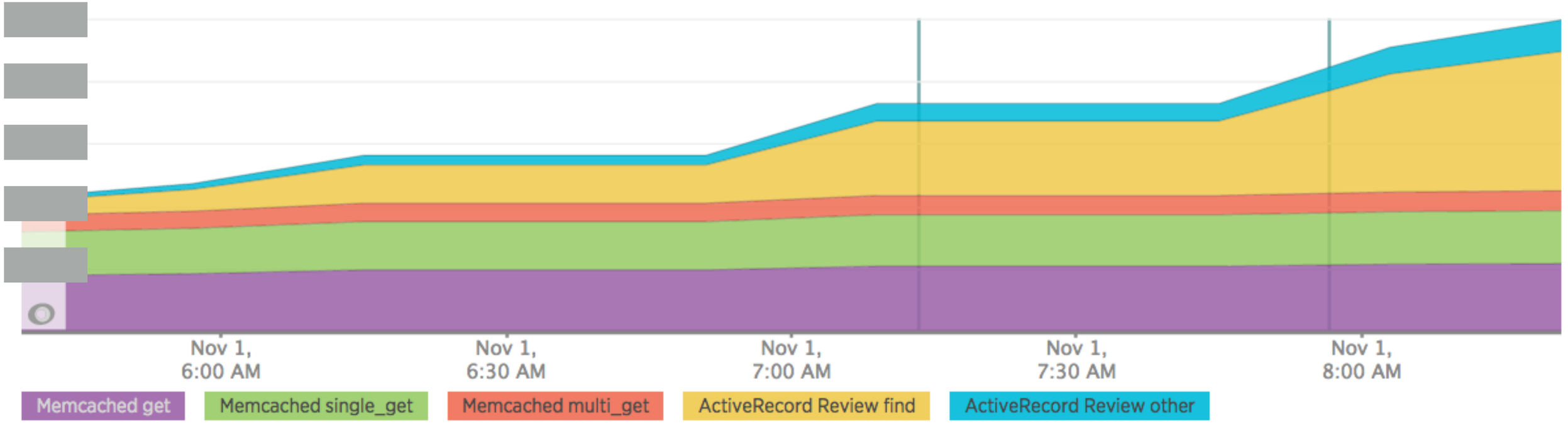
Frontend Response Time Degradation


Web transactions time
368 ms
APP SERVER
6.13 s
BROWSER
Nov 1,
6:40 AM
Nov 1,
6:50 AM
Nov 1,
7:00 AM
Nov 1,
7:10 AM
Nov 1,
7:20 AM
Request Queuing
GC Execution
Middleware
Ruby
Memcached
MySQL
Redis
ActiveRecord
Web external

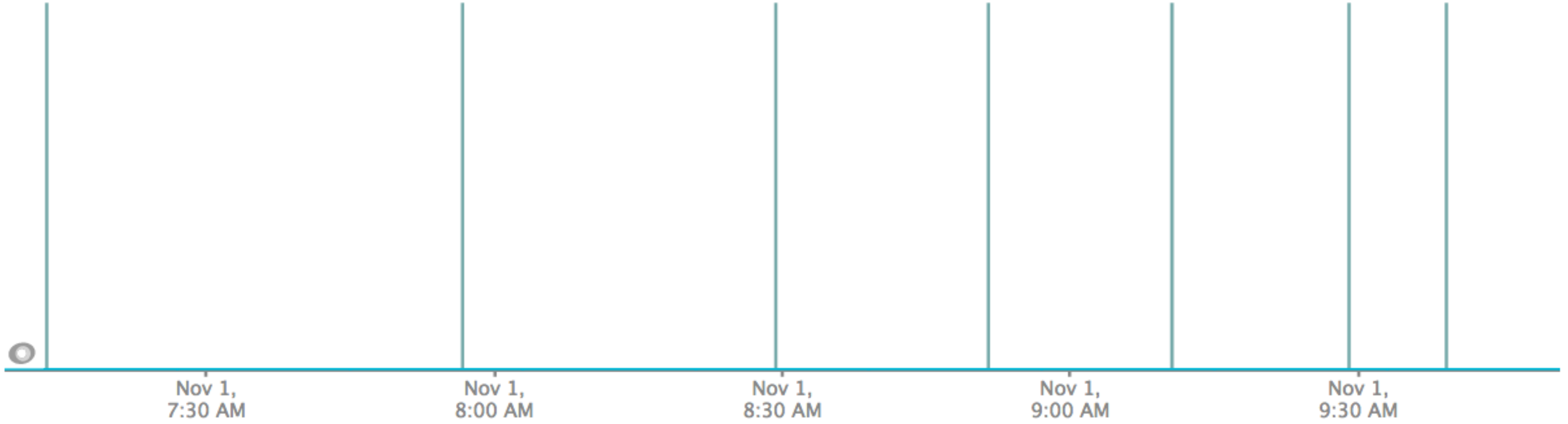

Web transactions time
374 ms
APP SERVER
6.18 s
BROWSER
Nov 1,
6:00 AM
Nov 1,
6:30 AM
Nov 1,
7:00 AM
Nov 1,
7:30 AM
Nov 1,
8:00 AM
Request Queuing
GC Execution
Middleware
Ruby
Memcached
MySQL
Redis
ActiveRecord
Web external

All databases overview

Top database operations by time consumed



Error rate (errors per request)

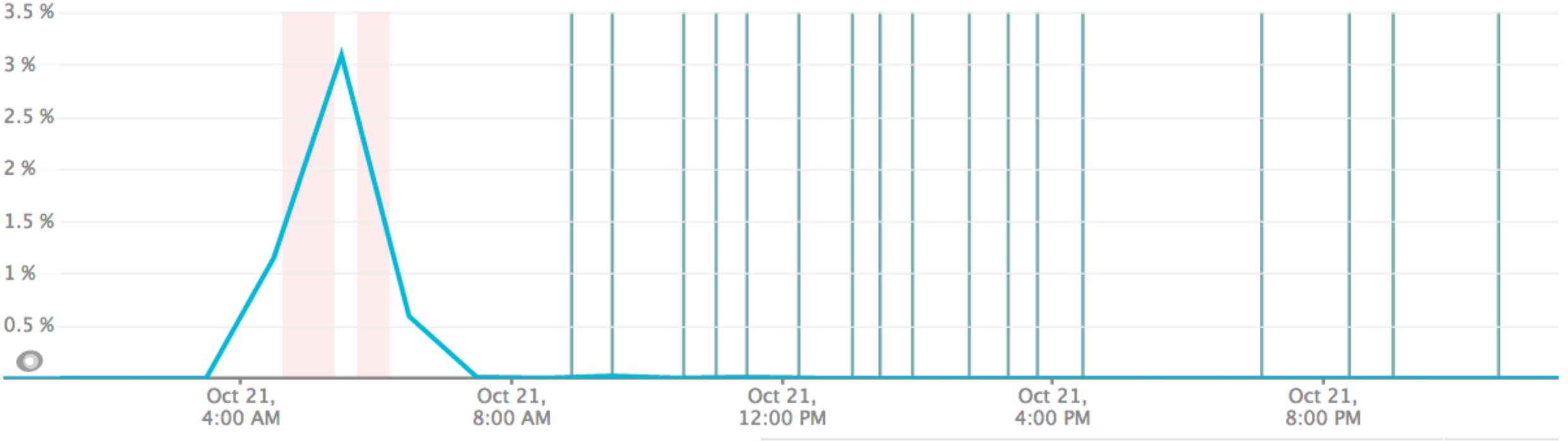


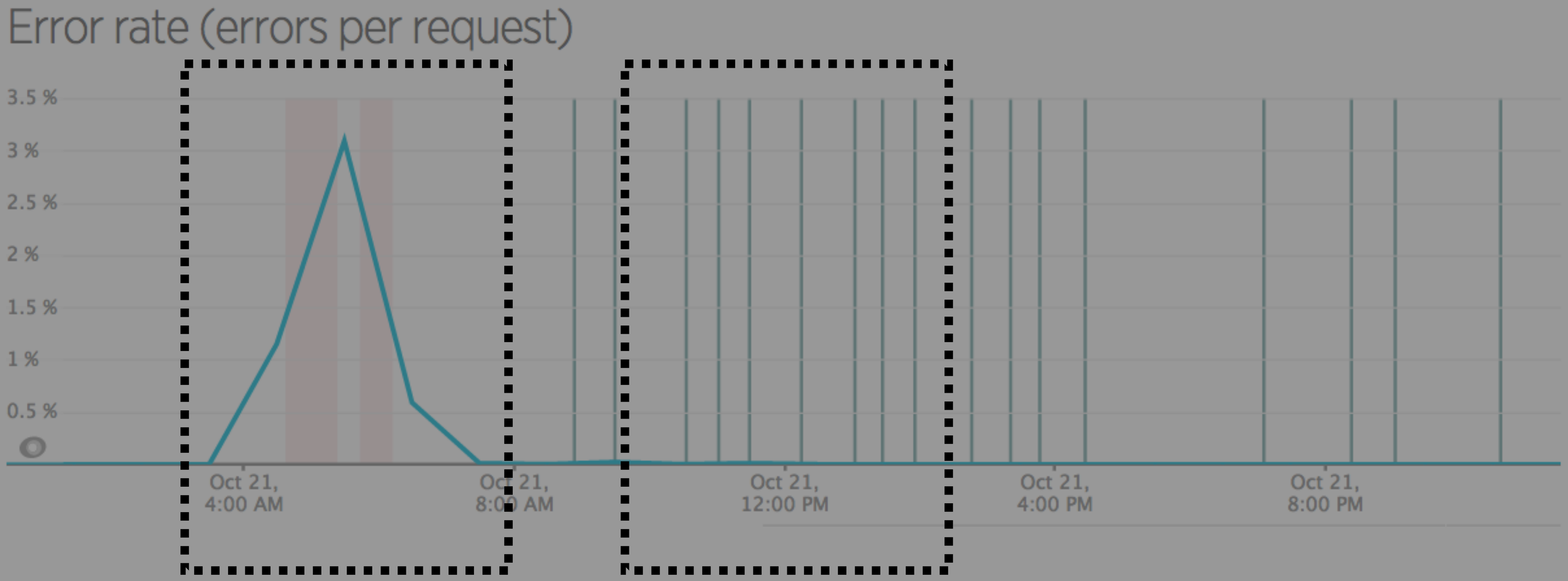
Questions?

- **Which database is having the problem?** Use the “Databases” tab to dig into which query is adding the most to the response time.
- **What is the end user impact?** The query will show which transactions/controllers it is part of.
- **How can I recreate the regression?** New Relic automatically samples slow queries — allowing you to dig into them. Includes information such as which index is being used for the query.

Remember that one time that toasters and webcams tried to take down the internet?

Error rate (errors per request)

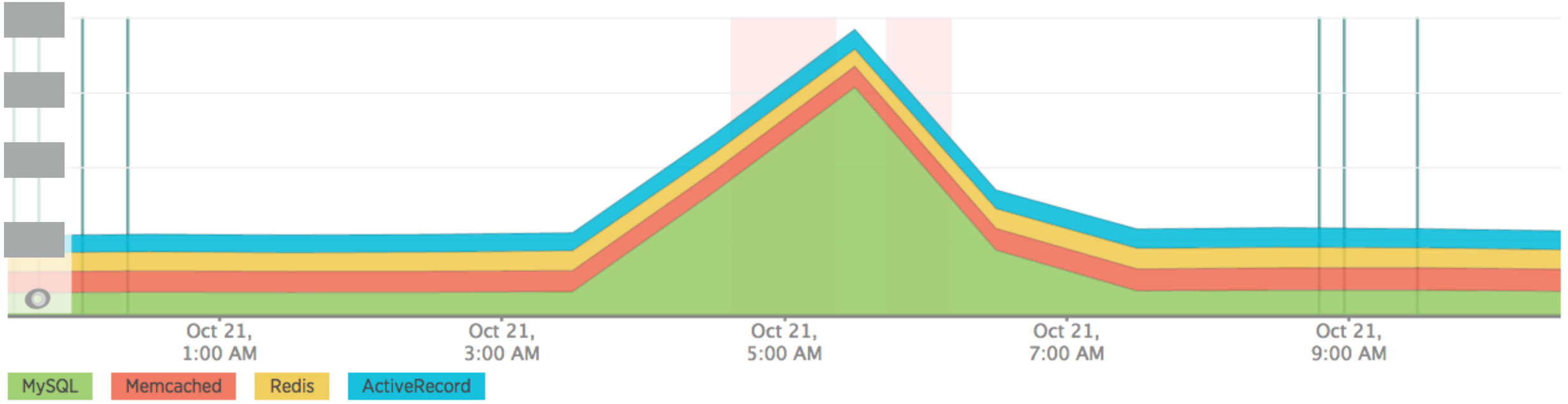




First Attack

Second Attack

Top database operations by query time



Questions?

- **Where are the errors coming from?** Both database queries timing and out from clients being unable to resolve DNS names.
- **Which database is having the problem?** All of them.
- **Which external providers are having the problem?** Some of them.

Airbnb Open

Airbnb Open

Using New Relic for a Live Event Launch

- **Development:** All applications developed at Airbnb come with New Relic “out of the box” without additional configuration. Developers are using APM from day one.
- **Deployment:** As you make changes to dark code, you can ensure that your deploy doesn’t impact live code paths.
- **Load Testing:** Understand how increased load affects your service and its external service dependencies.
- **Launch:** The day of the big keynote presentation, you are able to monitor for performance and regressions and triangulate their cause and understand their impact on the end user.

Inspiration and Citations

- Allspaw, John. *Blameless PostMortems and a Just Culture*. Code as Craft, 2012. < https://codeascraft.com/2012/05/22/blameless-postmortems/ >
- Beyer, Betsy, Chris Jones, and Jennifer Petoff. S*ite Reliability Engineering: How Google Runs Production Systems*. O'Reilly Media, 2016.
- Fowler, Susan. *Who's On Call?* 2016. < http://www.susanjfowler.com/blog/2016/9/6/whos-on-call >
- Malpass, Ian. *Measure Anything, Measure Everything.* Code as Craft, 2011. < https://codeascraft.com/2011/02/15/measure-anything-measure-everything/ >
- Underwood, Todd. *PostOps: A Non-Surgical Tale of Software, Fragility, and Reliability*. Lisa, 2013. < https://www.usenix.org/conference/lisa13/technical-sessions/plenary/underwood >

Thanks!