Building a Culture of Reliability

SRECON EMEA 2017

Arup Chakrabarti
Director of Engineering, PagerDuty

@arupchak

pagerduty

Disclaimers

I work with ~~smrt~~ smart people

You are not PagerDuty

We get this wrong too

Definitions

Reliability

Probability that your software works*

What every CTO claims they want because numbers

Culture

Social behavior and norms for a group of people

A way to get your colleagues to behave the way you want them to without staring at them all the time

Metrics

“Show me the business impact”

-Your Pointy Haired Manager

“Here is a graph of open File Descriptors going through the roof”

-Frustrated Engineer

“What the $%#! is a File Descriptor?”

-Your Pointy Haired Manager

Business Metrics Managers Care About

Metrics Your Customers Care About

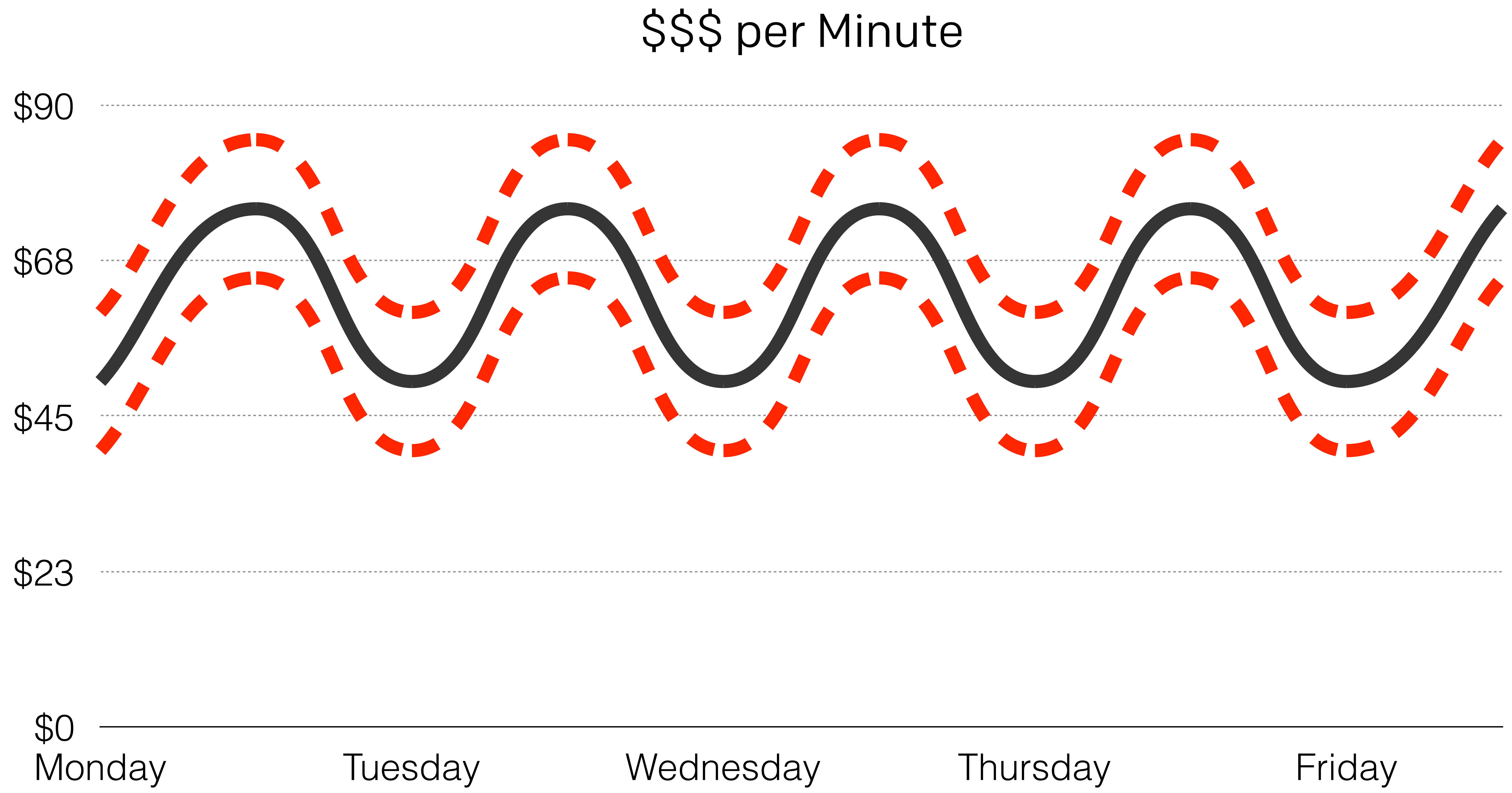
Two Types of Online Businesses

- Individual Transaction Businesses
- Subscription Businesses

Individual Transaction Business

$$$ per Minute
$90
$68
$45
$23
$0
Monday
Tuesday
Wednesday
Thursday
Friday

Individual Transaction Business



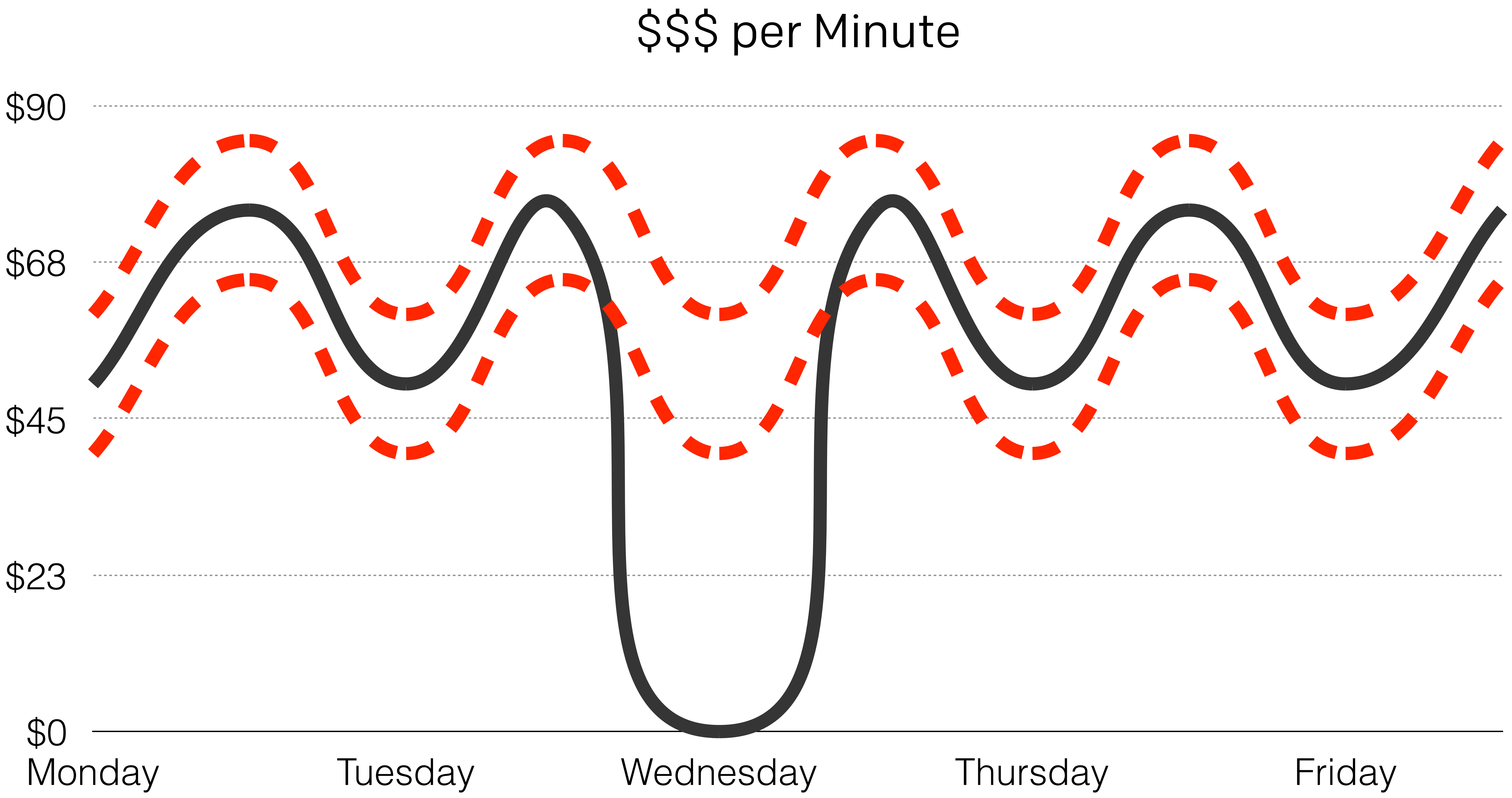
@arupchak

Individual Transaction Business

@arupchak

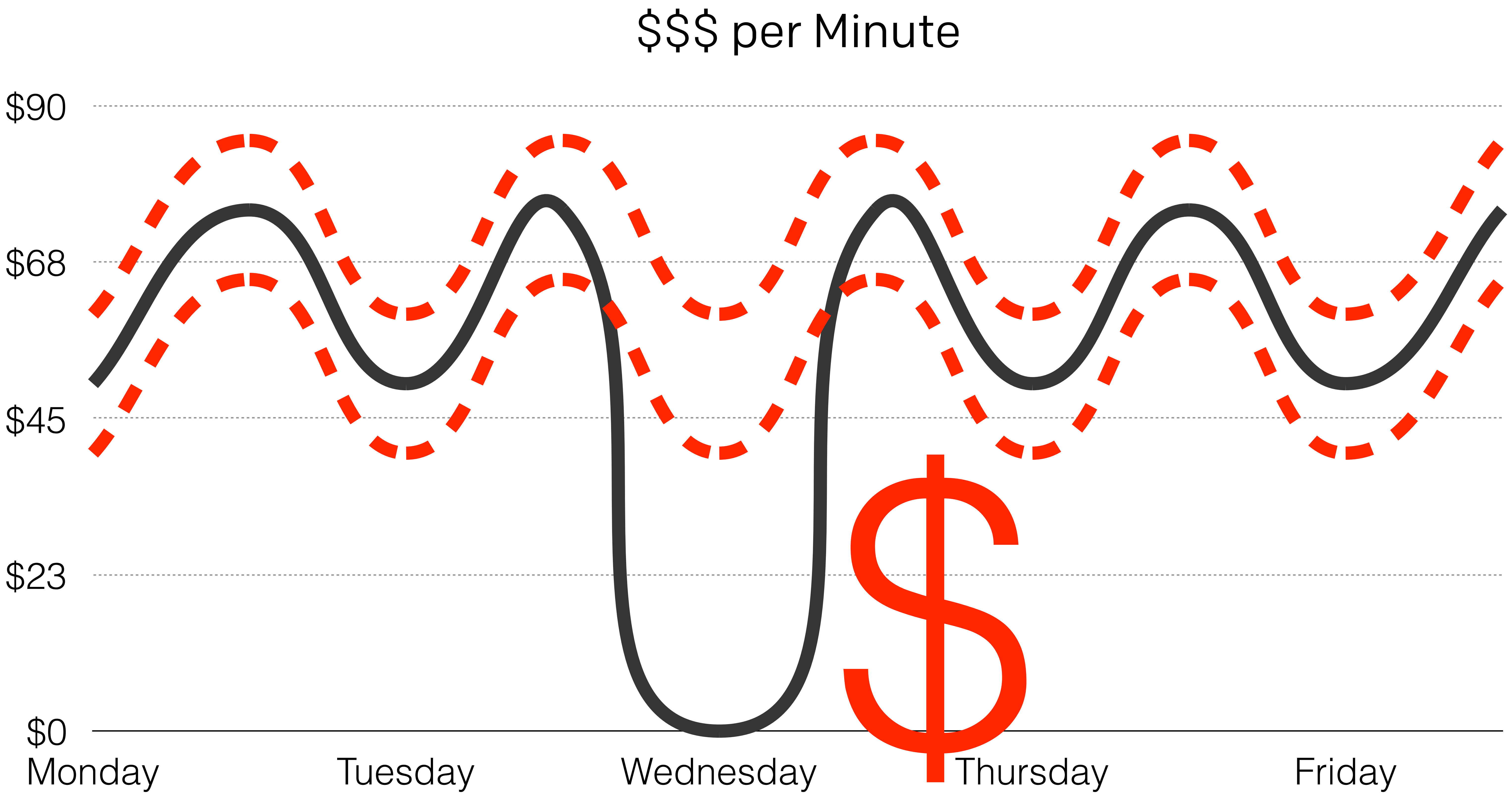
pagerduty

Individual Transaction Business



@arupchak

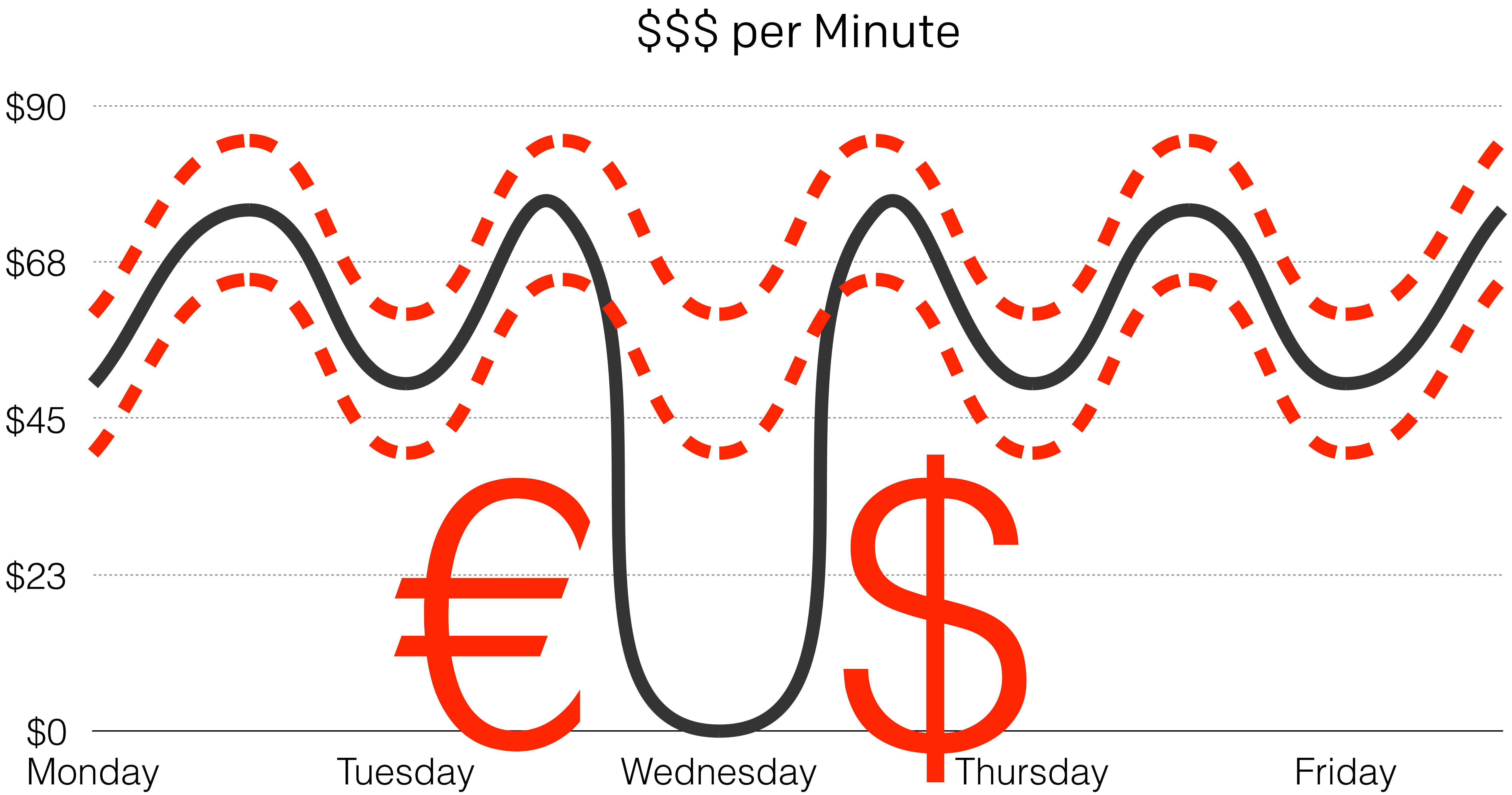
Individual Transaction Business



@arupchak

pagerduty

Individual Transaction Business



@arupchak

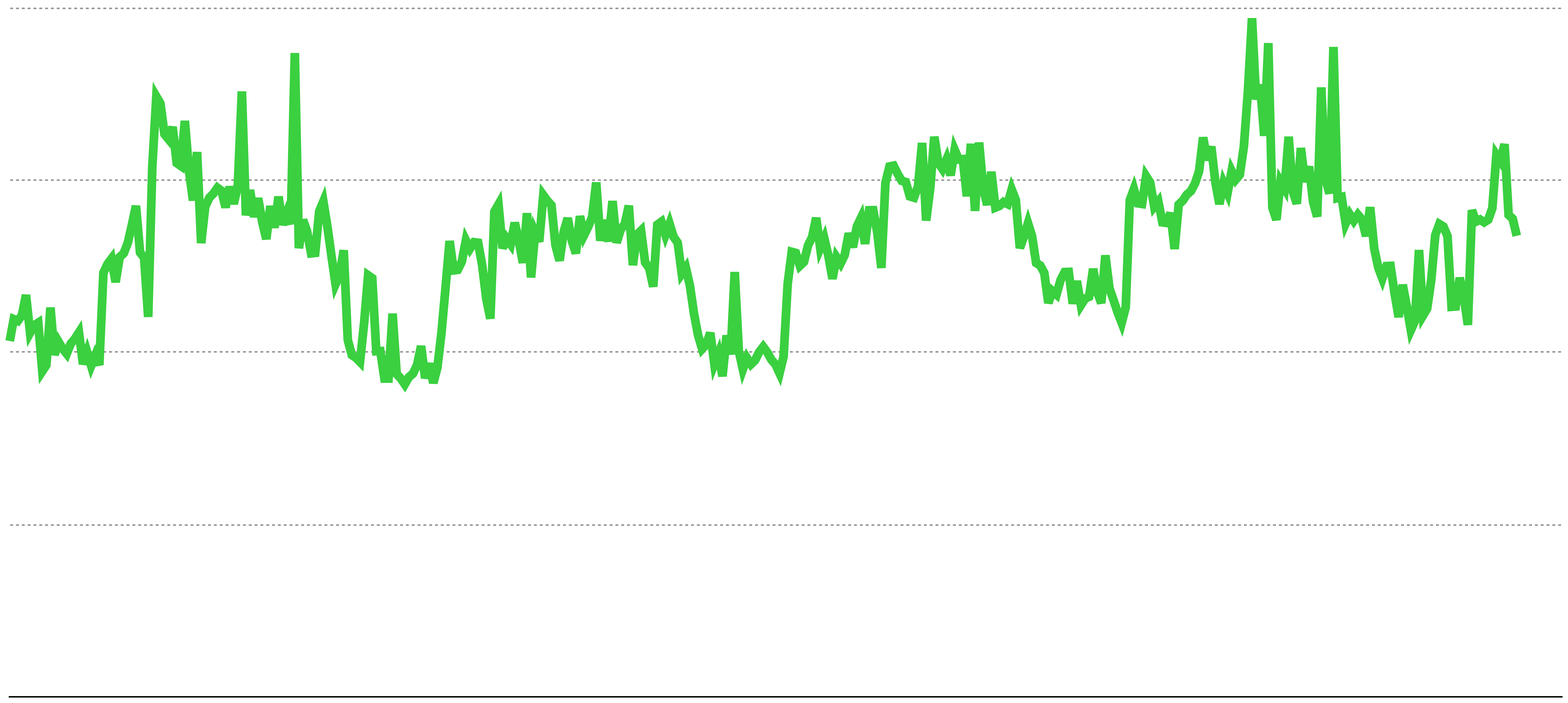
pagerduty

Subscription Businesses

- Cannot solely measure when you make money
- Poor Reliability erodes trust and will cause you lose revenue
- Need to find something between how money is made and what customers care about

Subscription Businesses

Incidents Resolved per Hour - July 2017



@arupchak

pagerduty

Finding the right metrics is hard

But still worth looking for

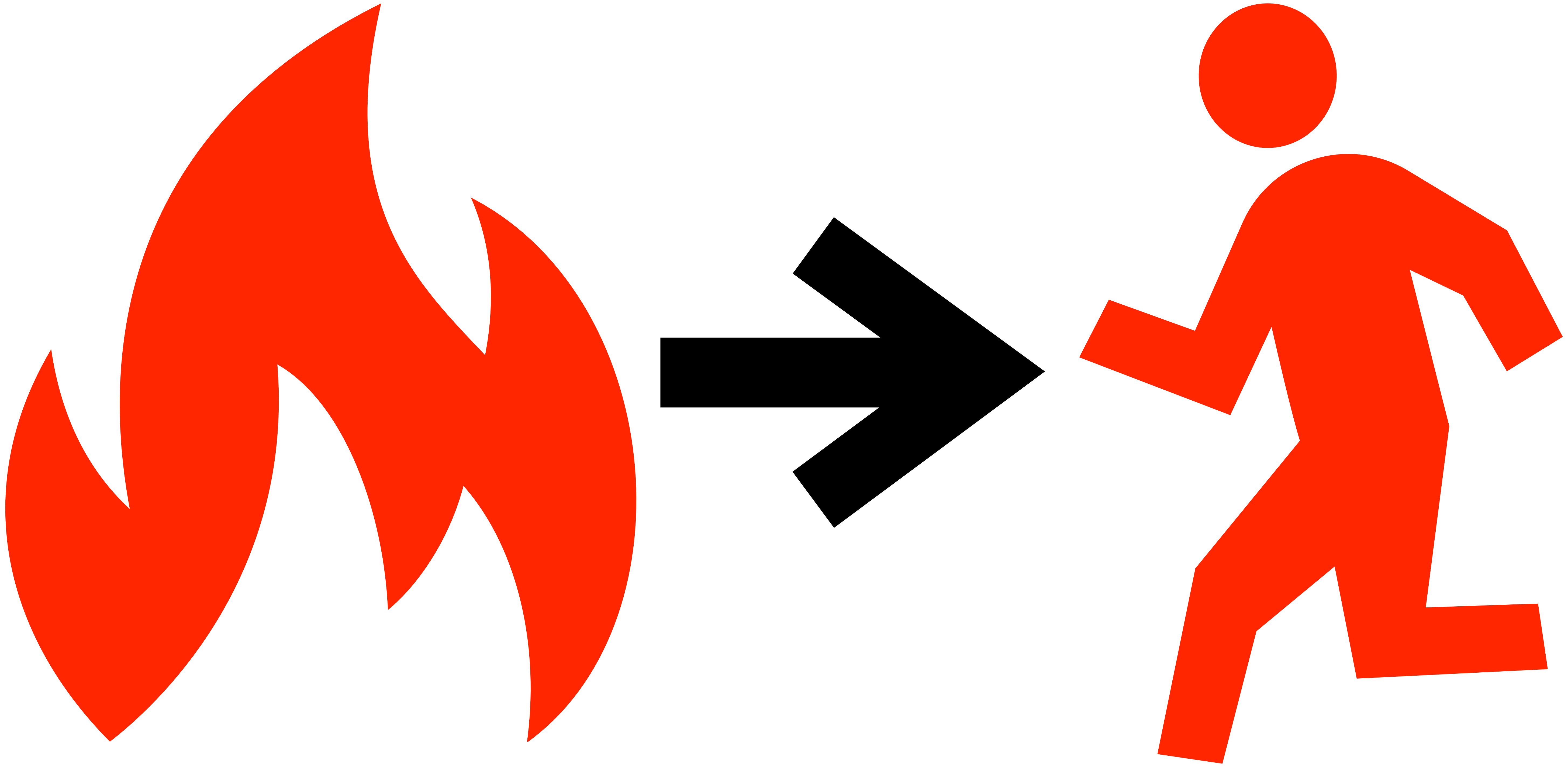
More People On-Call

Customers do not care who gets paged

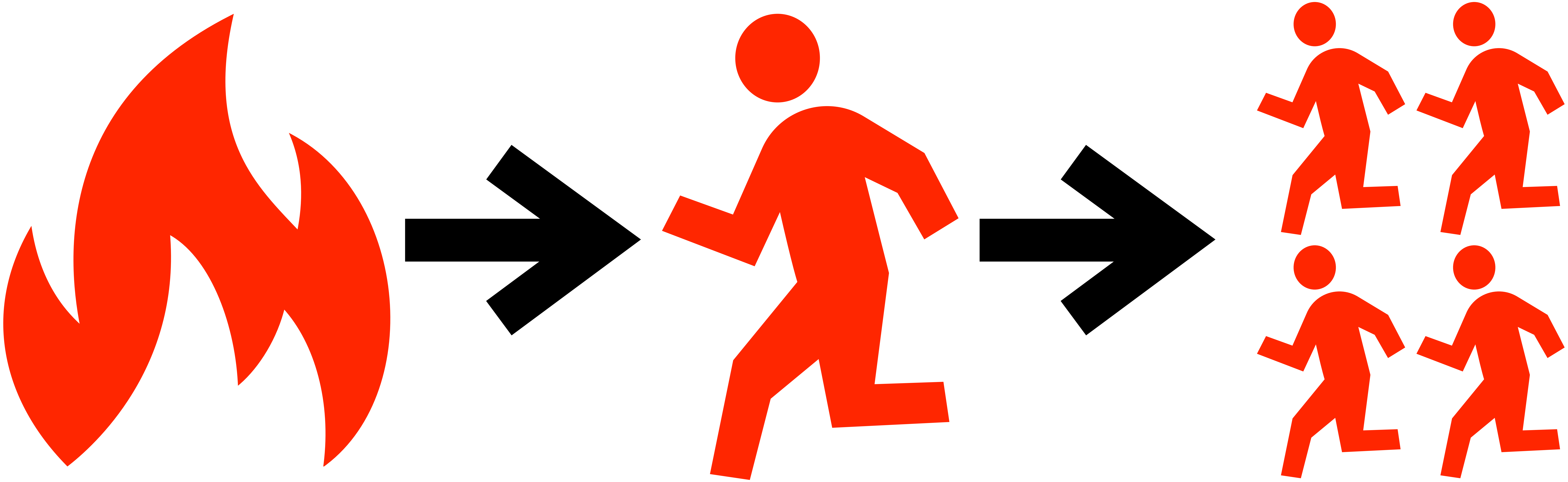
Customers just want to use your service

Centralized Operations Engineering Org

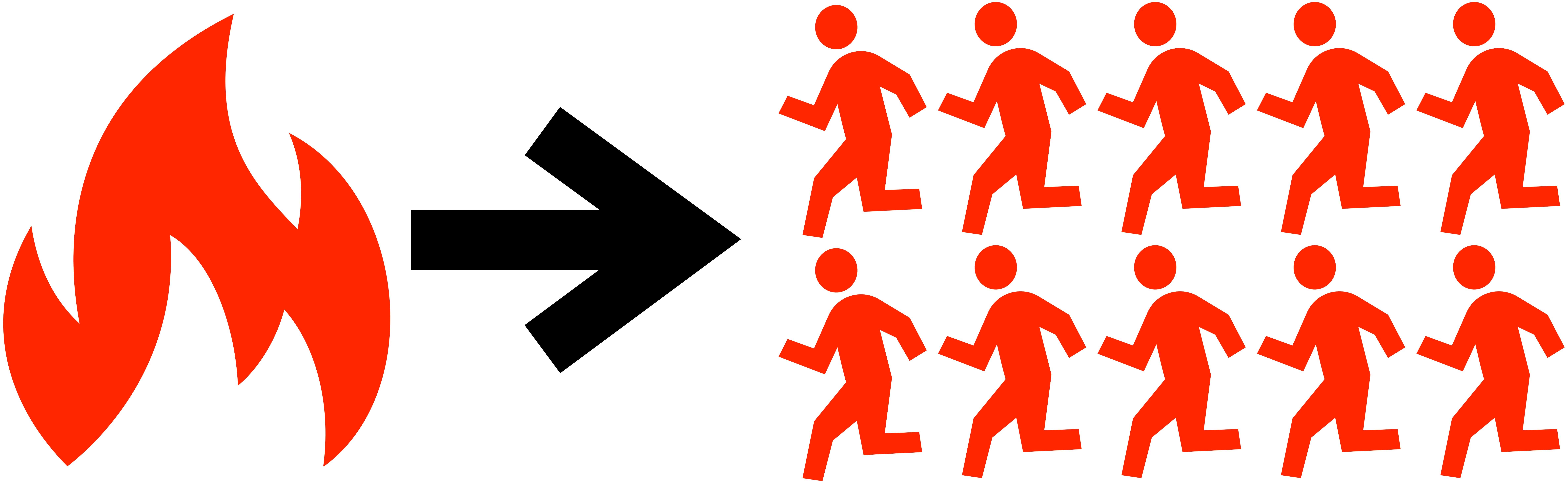
Centralized Operations Engineering Org



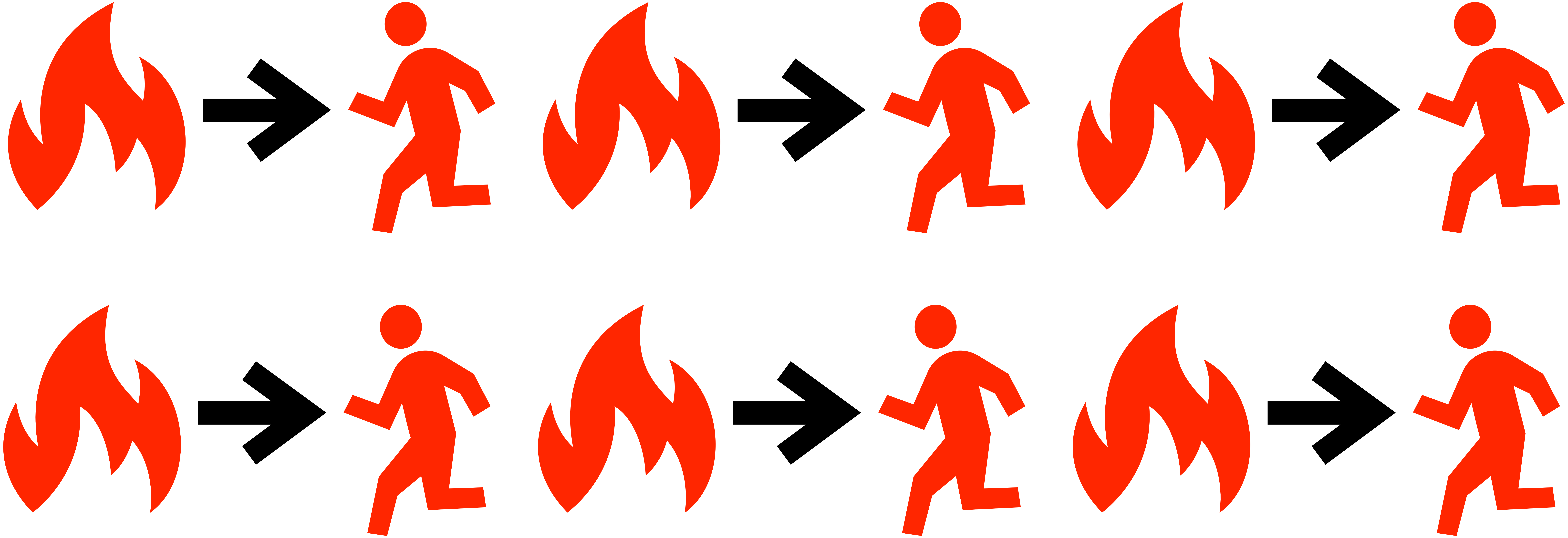
Centralized Operations Engineering Org



Distributed Operations Engineering Org



Distributed Operations Engineering Org



Distributed Operations ~~Engineering~~ Org

Eng

Product

HR

UX

Marketing

Execs

Distributed Operations Org

- Sets expectations around availability of people
- More small incidents over single major incident
- Builds empathy and why Reliability is hard

Tooling and Processes

“If we just install Nagios, everything will be fine and all of our problems will be solved”

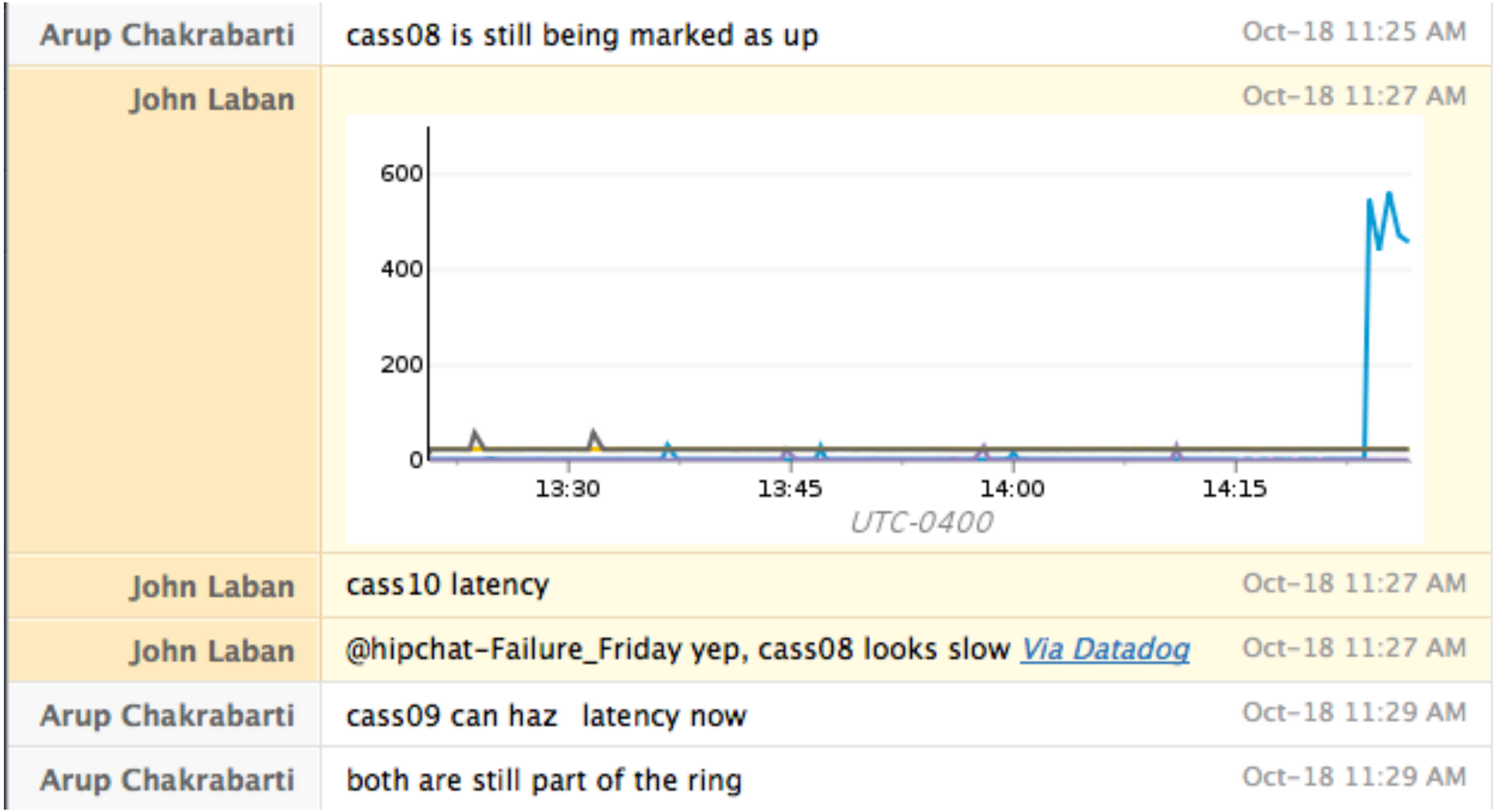
-Arup in 2002

“We humans co-evolve with our tools. We change the tools, and the tools change us, and that cycle repeats.”

-Jeff Bezos

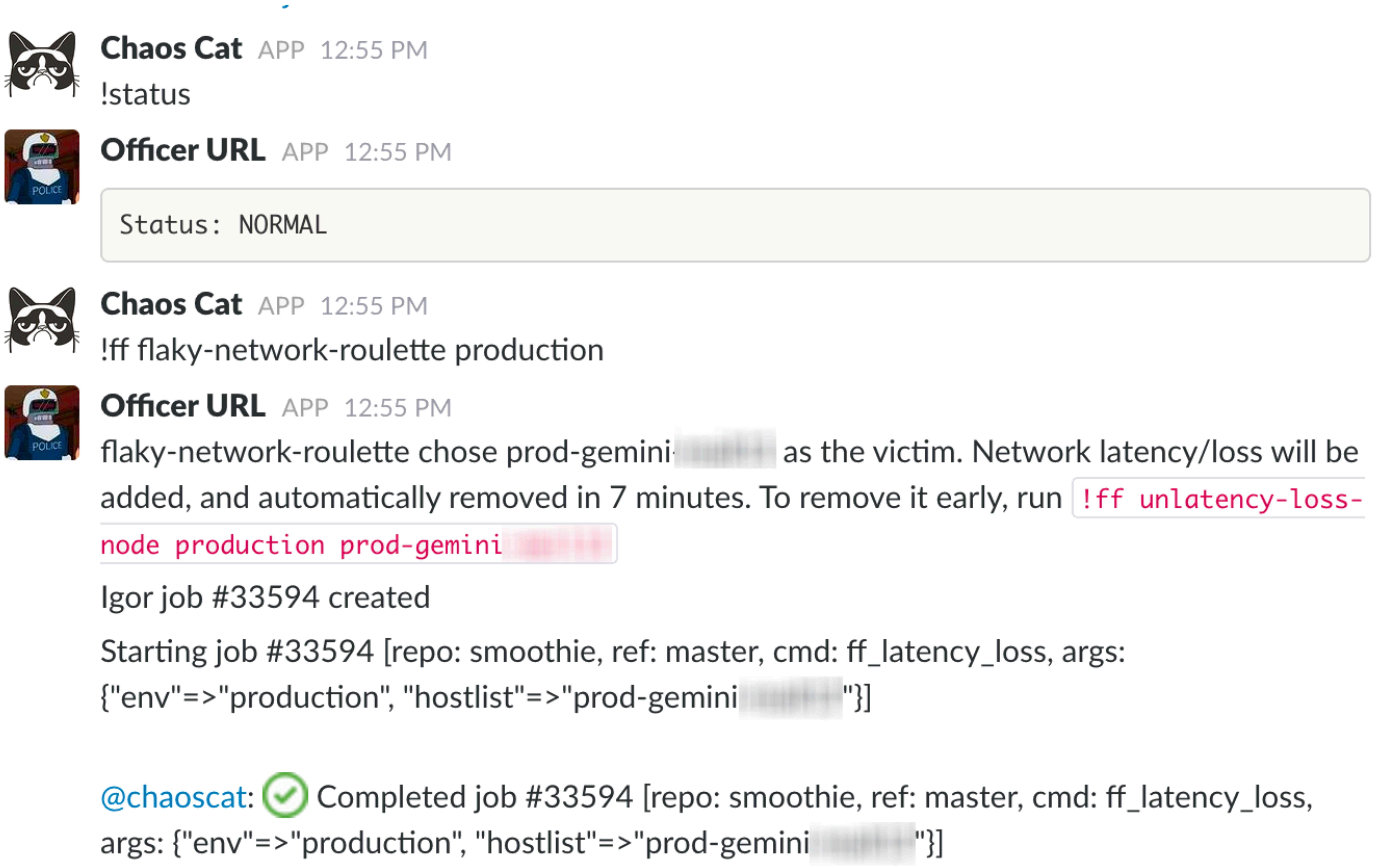
Failure Friday (Process)

Started Small



Arup Chakrabarti	cass08 is still being marked as up	Oct-18 11:25 AM
John Laban		Oct-18 11:27 AM
John Laban	cass10 latency	Oct-18 11:27 AM
John Laban	@hipchat-Failure_Friday yep, cass08 looks slow *Via Datadog*	Oct-18 11:27 AM
Arup Chakrabarti	cass09 can haz latency now	Oct-18 11:29 AM
Arup Chakrabarti	both are still part of the ring	Oct-18 11:29 AM

Got Bigger and Smarter



Chaos Cat APP 12:55 PM

!status

Officer URL APP 12:55 PM

```
Status: NORMAL
```

Chaos Cat APP 12:55 PM

!ff flaky-network-roulette production

Officer URL APP 12:55 PM

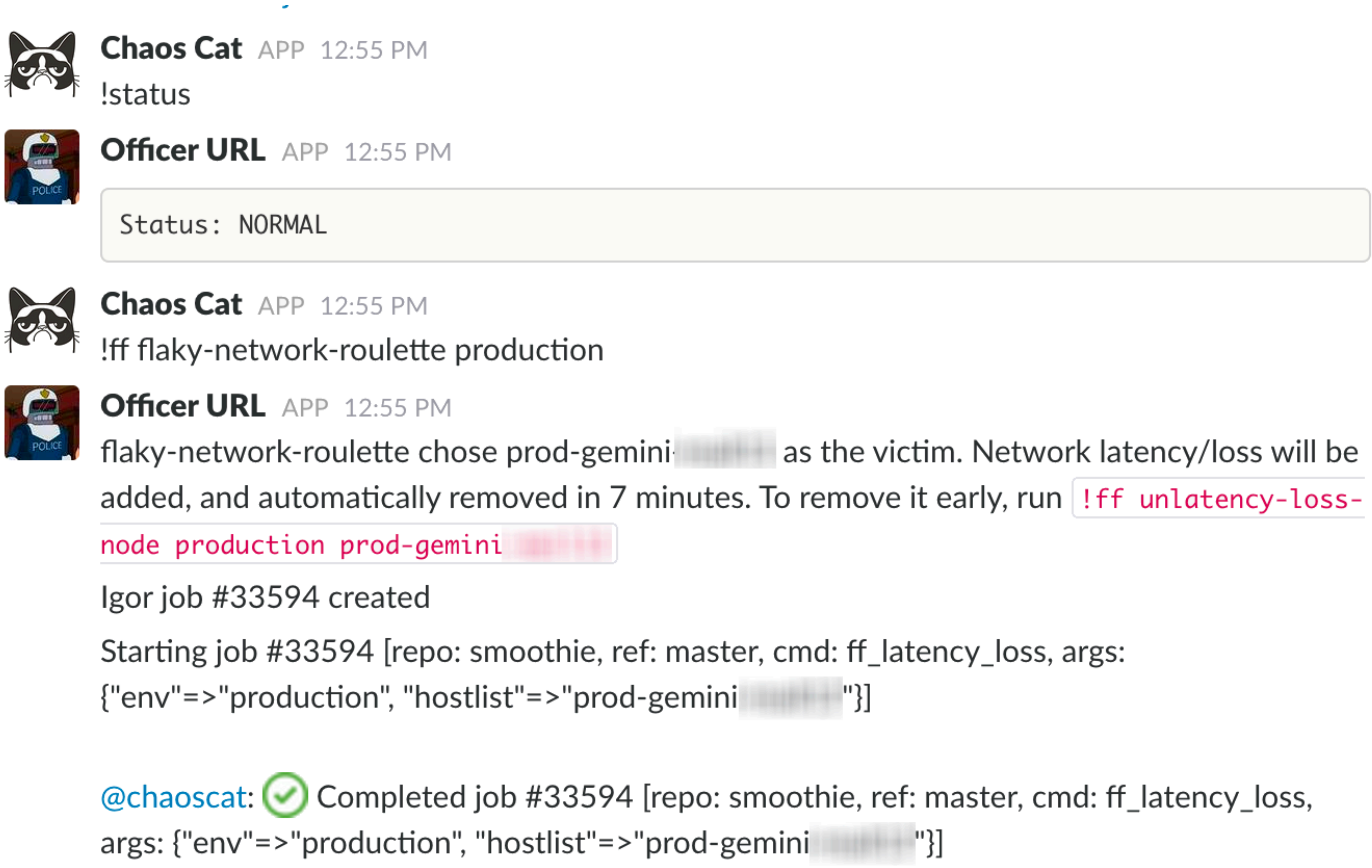
flaky-network-roulette chose prod-gemini[redacted] as the victim. Network latency/loss will be added, and automatically removed in 7 minutes. To remove it early, run `!ff unlatency-loss-node production prod-gemini[redacted]`

Igor job #33594 created

Starting job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

@chaoscat: ✓ Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

?



Chaos Cat APP 12:55 PM

!status

Officer URL APP 12:55 PM

```
Status: NORMAL
```

Chaos Cat APP 12:55 PM

!ff flaky-network-roulette production

Officer URL APP 12:55 PM

flaky-network-roulette chose prod-gemini[redacted] as the victim. Network latency/loss will be added, and automatically removed in 7 minutes. To remove it early, run `!ff unlatency-loss-node production prod-gemini[redacted]`

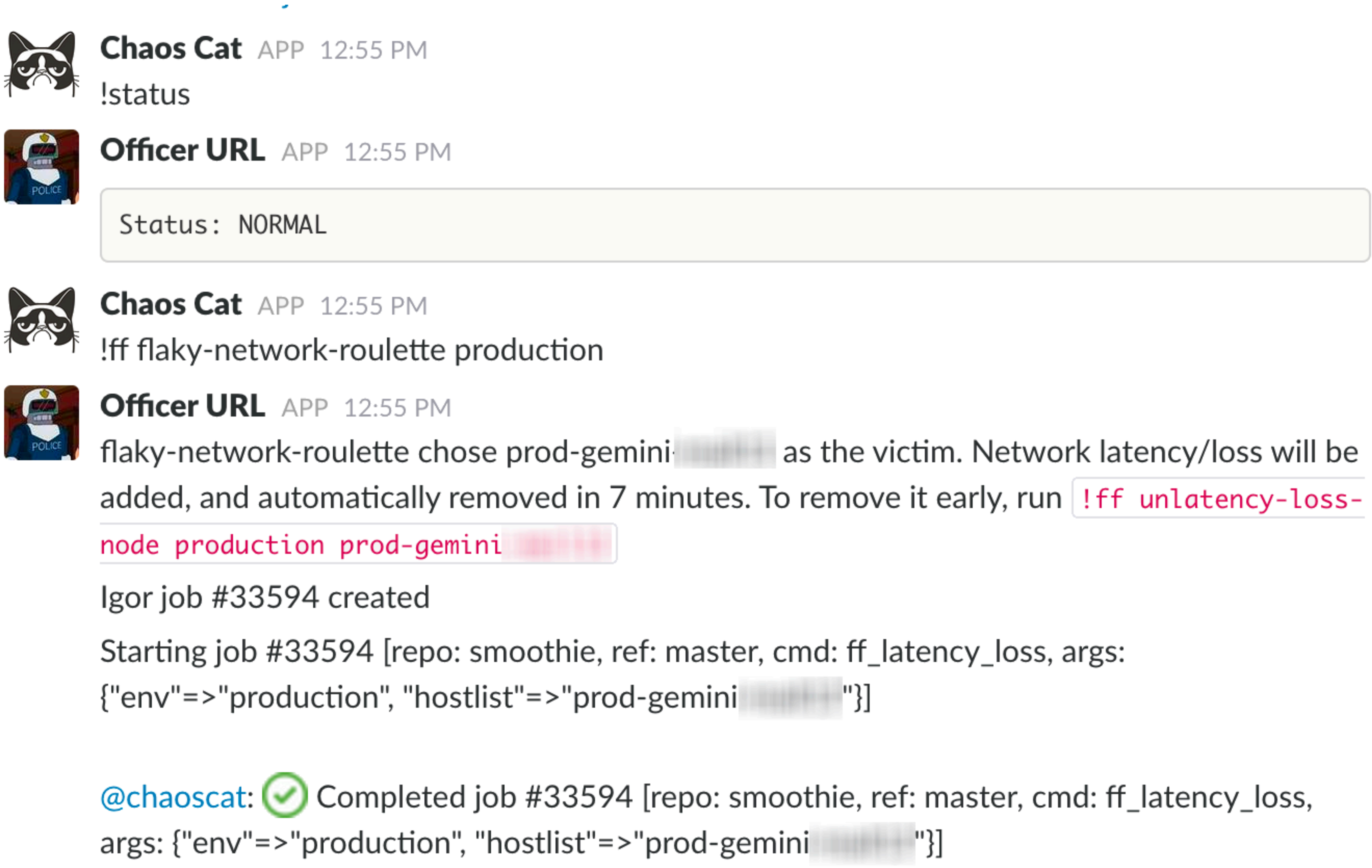
Igor job #33594 created

Starting job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

@chaoscat: ✓ Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

?

✓



Chaos Cat APP 12:55 PM

!status

Officer URL APP 12:55 PM

```
Status: NORMAL
```

Chaos Cat APP 12:55 PM

!ff flaky-network-roulette production

Officer URL APP 12:55 PM

flaky-network-roulette chose prod-gemini[redacted] as the victim. Network latency/loss will be added, and automatically removed in 7 minutes. To remove it early, run `!ff unlatency-loss-node production prod-gemini[redacted]`

Igor job #33594 created

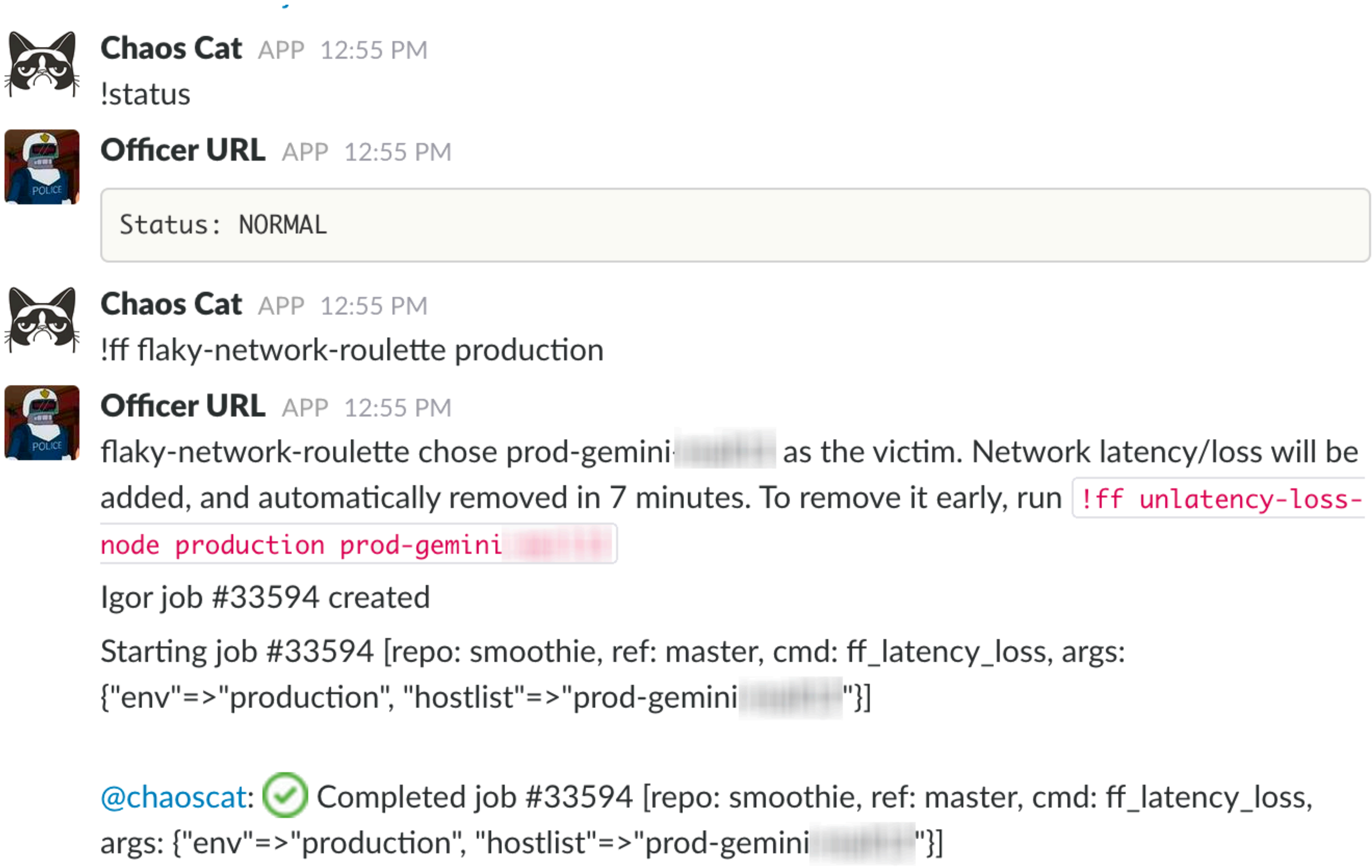
Starting job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

@chaoscat: ✓ Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

?

✓

🔥



Chaos Cat APP 12:55 PM
!status

Officer URL APP 12:55 PM

```
Status: NORMAL
```

Chaos Cat APP 12:55 PM
!ff flaky-network-roulette production

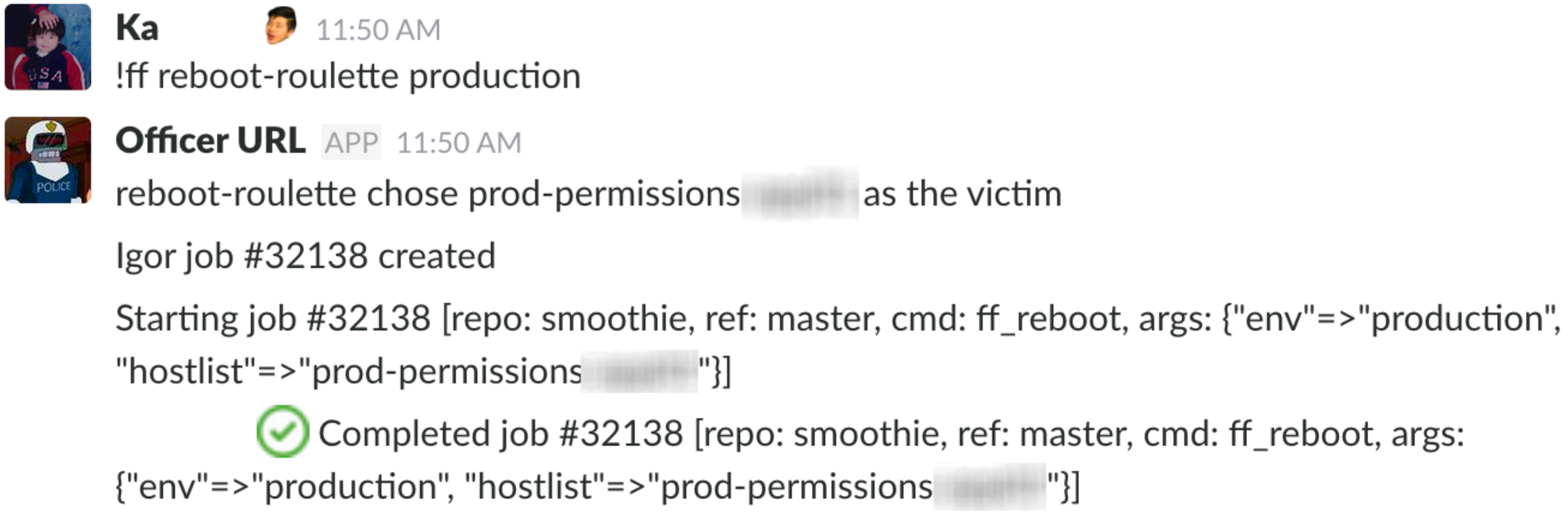
Officer URL APP 12:55 PM
flaky-network-roulette chose prod-gemini[redacted] as the victim. Network latency/loss will be added, and automatically removed in 7 minutes. To remove it early, run `!ff unlatency-loss-node production prod-gemini[redacted]`

Igor job #33594 created

Starting job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

@chaoscat: ✓ Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env"=>"production", "hostlist"=>"prod-gemini[redacted]"}]

Reboot Roulette (Tool)



Ka 11:50 AM

!ff reboot-roulette production

Officer URL APP 11:50 AM

reboot-roulette chose prod-permissions[redacted] as the victim

Igor job #32138 created

Starting job #32138 [repo: smoothie, ref: master, cmd: ff_reboot, args: {"env"=>"production", "hostlist"=>"prod-permissions[redacted]"}]

✓ Completed job #32138 [repo: smoothie, ref: master, cmd: ff_reboot, args: {"env"=>"production", "hostlist"=>"prod-permissions[redacted]"}]

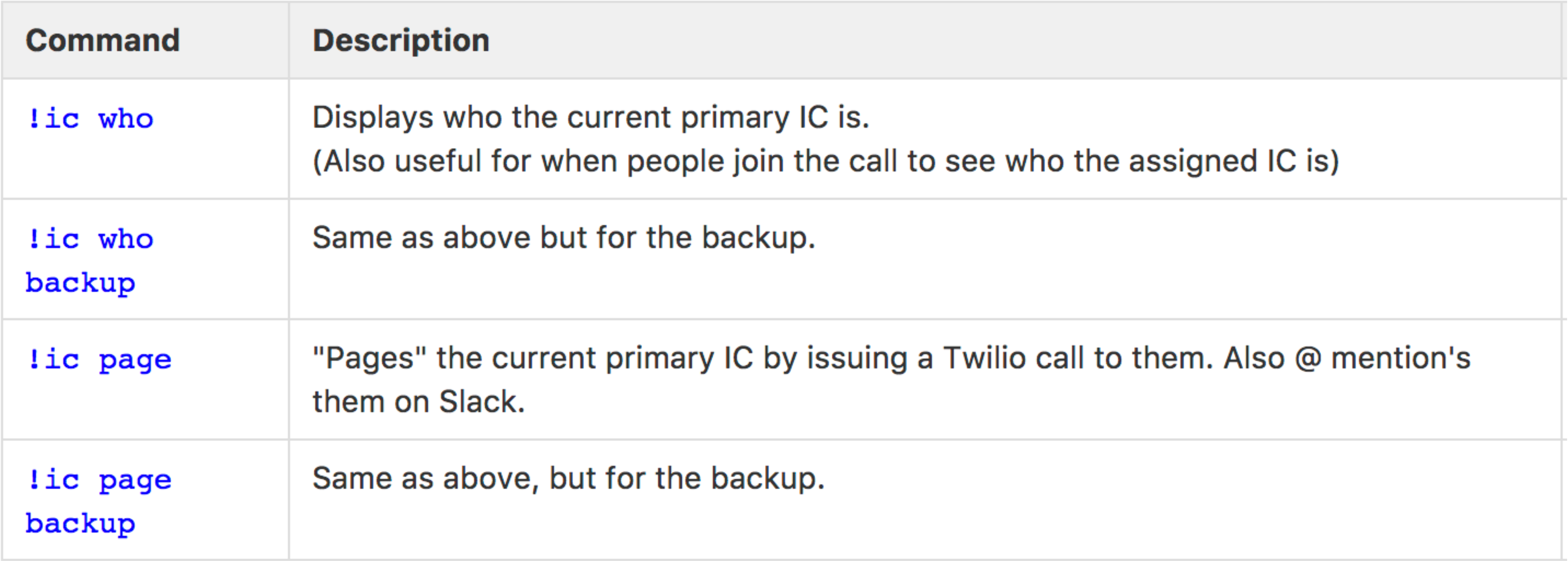
Major Incident Response (Process and Tooling)

@arupchak

pagerduty

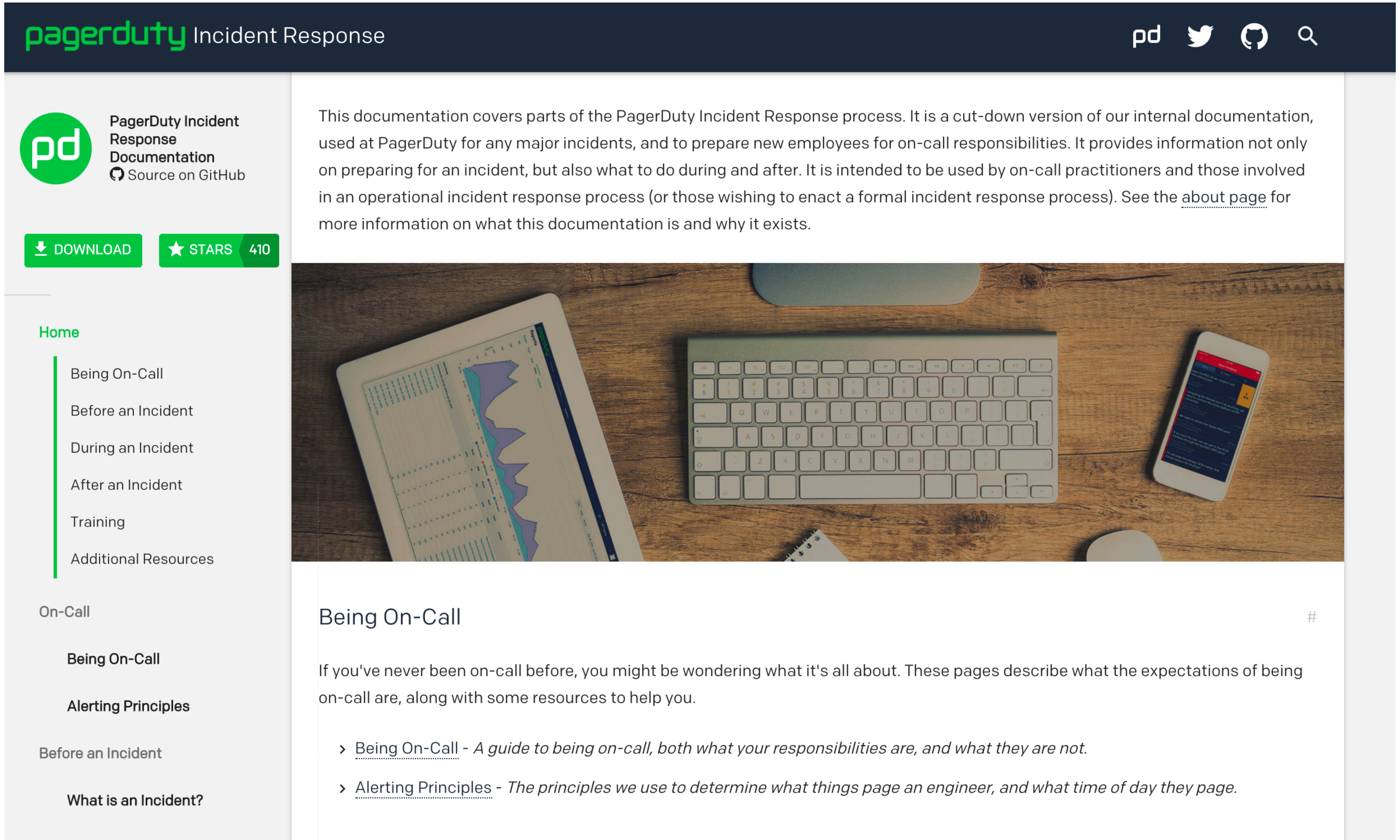
Started Really Poorly

Got A Little Better Each Time



Command	Description
`!ic who`	Displays who the current primary IC is. (Also useful for when people join the call to see who the assigned IC is)
`!ic who backup`	Same as above but for the backup.
`!ic page`	"Pages" the current primary IC by issuing a Twilio call to them. Also @ mention's them on Slack.
`!ic page backup`	Same as above, but for the backup.

Still Not Perfect



This documentation covers parts of the PagerDuty Incident Response process. It is a cut-down version of our internal documentation, used at PagerDuty for any major incidents, and to prepare new employees for on-call responsibilities. It provides information not only on preparing for an incident, but also what to do during and after. It is intended to be used by on-call practitioners and those involved in an operational incident response process (or those wishing to enact a formal incident response process). See the about page for more information on what this documentation is and why it exists.

Being On-Call

If you've never been on-call before, you might be wondering what it's all about. These pages describe what the expectations of being on-call are, along with some resources to help you.

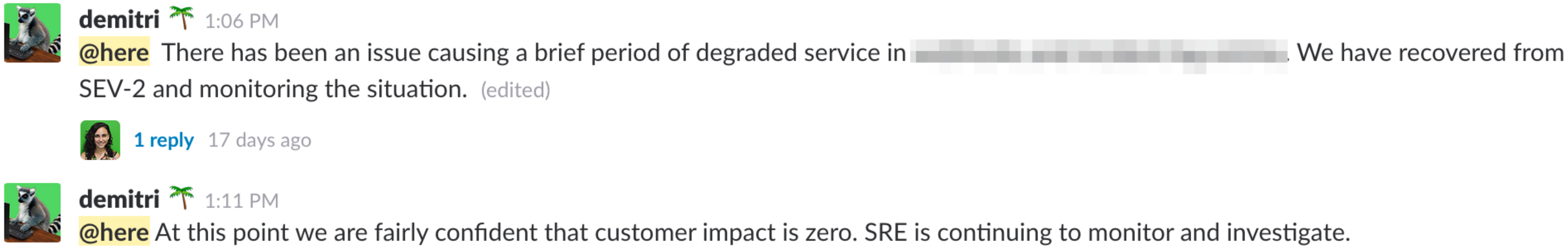
- Being On-Call - *A guide to being on-call, both what your responsibilities are, and what they are not.*
- Alerting Principles - *The principles we use to determine what things page an engineer, and what time of day they page.*

Internal Liaison Role (Process)

Over-communicate during Major Incidents

@arupchak

pagerduty



demitri 🌴 1:06 PM
@here There has been an issue causing a brief period of degraded service in [redacted]. We have recovered from SEV-2 and monitoring the situation. (edited)

1 reply 17 days ago

demitri 🌴 1:11 PM
@here At this point we are fairly confident that customer impact is zero. SRE is continuing to monitor and investigate.

Improving Reliability means constantly failing, constantly recovering, and constantly learning

Yes, it can be exhausting,
but it is worth it

Improving Culture means constantly failing, constantly recovering, and constantly learning

Yes, it can be even more exhausting, but it is really really really worth it

Thank You

WE ARE HIRING PAGERDUTY.COM/CAREERS
ARUP@PAGERDUTY.COM

Arup Chakrabarti
Director of Engineering, PagerDuty

@arupchak

pagerduty

Related Reading

- https://response.pagerduty.com/
- https://www.pagerduty.com/blog/intern-insights-on-call-experience/
- https://www.pagerduty.com/blog/failure-fridays-four-years/
- https://speakerdeck.com/arupchak