Building a Culture of Reliability
SRECON EMEA 2017

@arupchak
Disclaimers
I work with 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

@arupchak
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

$0  $23  $45  $68  $90

Monday  Tuesday  Wednesday  Thursday  Friday

$0

@arupchak
Individual Transaction Business

$$ per Minute

$90

$68

$45

$23

$0

Monday  Tuesday  Wednesday  Thursday  Friday
Individual Transaction Business

$$$$ per Minute

Monday Tuesday Wednesday Thursday Friday

$90 $68 $45 $23 $0
Individual Transaction Business

$0  $23  $45  $68  $90

Monday  Tuesday  Wednesday  Thursday  Friday

$ per Minute
Individual Transaction Business

$\$$ per Minute

$90
$68
$45
$23

$0 Monday Tuesday Wednesday Thursday Friday

@arupchak
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
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

@arupchak
Centralized Operations Engineering Org

@arupchak
Centralized Operations Engineering Org

@arupchak
Distributed Operations Engineering Org

@arupchak
Distributed Operations Engineering Org

@arupchak
Distributed Operations Engineering Org

Eng ➔ Product ➔ HR ➔ Execs

UX ➔ Marketing ➔
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
<table>
<thead>
<tr>
<th>User</th>
<th>Message</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arup Chakrabarti</td>
<td>cass08 is still being marked as up</td>
<td>Oct-18 11:25 AM</td>
</tr>
<tr>
<td>John Laban</td>
<td>cass10 latency</td>
<td>Oct-18 11:27 AM</td>
</tr>
<tr>
<td>John Laban</td>
<td>@hipchat—Failure_Friday yep, cass08 looks slow [Via Datadog]</td>
<td>Oct-18 11:27 AM</td>
</tr>
<tr>
<td>Arup Chakrabarti</td>
<td>cass09 can haz latency now</td>
<td>Oct-18 11:29 AM</td>
</tr>
<tr>
<td>Arup Chakrabarti</td>
<td>both are still part of the ring</td>
<td>Oct-18 11:29 AM</td>
</tr>
</tbody>
</table>
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-geminian 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-geminian`

Igor job #33594 created

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

@chaoscat: ✔ Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: "env"=>"production", "hostlist"=>"prod-geminian"]
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 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`

Igor job #33594 created
Starting job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args:
{"env":"production", "hostlist":"prod-gemini"]

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

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

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

@arupchak
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 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`

Igor job #33594 created

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

@chaoscat: Completed job #33594 [repo: smoothie, ref: master, cmd: ff_latency_loss, args: {"env":"production", "hostlist":"prod-gemini"}]
Reboot Roulette (Tool)
K: !ff reboot-roulette production

Officer URL: reboot-roulette chose prod-permissions... as the victim

Igor job #32138 created

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

☑ Completed job #32138 [repo: smoothie, ref: master, cmd: ff_reboot, args: {"env":"production", "hostlist":"prod-permissions..."}]
Major Incident Response (Process and Tooling)
Started Really Poorly
Got A Little Better Each Time
<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>!ic who</td>
<td>Displays who the current primary IC is. (Also useful for when people join the call to see who the assigned IC is)</td>
</tr>
<tr>
<td>!ic who backup</td>
<td>Same as above but for the backup.</td>
</tr>
<tr>
<td>!ic page</td>
<td>&quot;Pages&quot; the current primary IC by issuing a Twilio call to them. Also @ mention's them on Slack.</td>
</tr>
<tr>
<td>!ic page backup</td>
<td>Same as above, but for the backup.</td>
</tr>
</tbody>
</table>
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
@here There has been an issue causing a brief period of degraded service in SEV-2 and monitoring the situation. (edited)

@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 really worth it.
Thank You

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

@arupchak
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

@arupchak